PCIe Gen 4/5/6/CXL/NVMe, NAND, DDR5 测试技术和工具白皮书 Ver 10

随着 PCIe Gen 6 产品开发的推进以及 CXL3.0 相关产品的逐渐浮现，包括各类芯片，模块，插卡，系统等，针对 PCIe 6.0 总线的问题诊断分析和测试工具也需要提前提到议事日程。同时，符合新的 ONFI 5.0 规范的 2.4GT NAND, DDR5 在固态盘和主机中的广泛发布使用也对于测试带来挑战。

本文主要针对 PCIe Gen 5.0/6.0 协议层及以上各类研发测试使用的分析、诊断和测试工具进行了系统阐述和图解剖析。本文推荐的产品都是在全球业内各大知名芯片设计、系统集成设计公司获得普遍使用的针对 PCIe Gen 5/6 总线进行问题定位、诊断分析以及相关测试的工具，这些产品广泛适用于当前国内从事计算、网络、存储、SSD、AI、ML/DL、大数据、CPU/GPU/DPU/MaPU 以及基于 FPGA 的各类数据加速产品，SmartNIC，移动通讯，嵌入式系统设计，汽车电子等诸多领域的 PCIe Gen 5 高速总线的开发和测试过程中。

本文涉及的产品包括并不限于 PCIe Gen 5 /6 协议分析，底层故障注入，热插拔测试，电压拉偏&功耗测试，电压/电流/功耗/sideband 等信号的长时间监测和回溯分析，掉电测试，性能/功能/协议兼容性/InterOp 测试，高低温测试，以及如何构建 PCIe Gen 5 测试环境，从主板和 Host Card 选型开始，涉及各类 PCIe 接口（AIC, M.2, U.2, U.3, E1.S, E1.L, E3.S, E3.L, Cable 接口如 MCIO, MINI-SAS-HD, SLIM-SAS, OCULINK 等）的端口扩展，各种常用的主机卡，转接卡，盘柜，延长线的选择等，以及考虑到测试便利性使用的主板托架和实验室批量测试机架等解决方案。

该文附录章节也提供了针对 PCIe/CXL/NVMe 等相关技术协议从基础概念，到协议层及上层协议的介绍，方便在校学生或从事研发、测试工作的工程师速查使用。

Saniffer Co., Limited
sales@saniffer.com
021-50807071 / 13127856862
Ver 2024.3

目录

本次版本相对于 ver9.0 版修订内容一览

章节	修订内容
封面	将 Gen5x16 analyzer 替换为 Gen6x16 analyzer + exerciesr + host/device smart，fixture，增加 Gen5 JBOF
1.1	将之前的 1.1 推到 1.2，增加的 1.1.1 章节主要介绍了 PCIe Gen5 和 Gen6 带来的挑战，同时，之前的保留了 2 年前 Gen5 的当时的现状的回溯和总结在 1.1.2
1.3.9	删除了 UFS/eMMC analyzer 的样机说明；同时删除了白皮书内大部分关于 UFS 协议培训相关的内容和说明
1.4.1	增加了 25-30 的公众号视频链接
2	增加 2.4 章节关于 PCIe Gen6 analyzer/exerciser 产品介绍，其它章节相应后移
2.1	调整原来的 2.1.1, 2.1.2 章节，增加了“2.1.2 PCIe Gen6 和 CXL3.0 新增的特性”章节
2.1.1	替换原有 gen4 u.2 连接图到 gen5 x16 连接图 - 带试验台杆 + 夹具
2.3.3	增加：支持外连 10GE 接口基于 NVMe SSD 的 NAS 存储
2.3.12	调整 widget 的顺序和标题
2.4.1	增加了 host smart fixture 和 device smart fixture 如何连接 analyzer 的连接拓扑图；增加了 PCIe 协议训练器（PCIe protocol exerciser）的功能基本说明
3.6.10	增加了 Flexible Data Placement (FDP)的功能测试描述
3.8	去掉了之前的“独立式 SMBus 仿真和协议分析工具”，增加了 SanBlaze PCIe Gen5 iRiser 故障注入卡技术介绍 - SANBlaze 用于精密信号控制和测量的新型专利 iRiser5
4.1.6	增加：通过 Quarch M.2 或者插卡类控制模块实现针对 M.2 SSD 和各类插卡的热插拔自动化测试
4.2.1	去掉了之前的 PPM U.2 连接图，增加了 PPM U.2 和 PPM M.2 连接图
4.2.5	新增：TechPowerUP Labs 采用 Quarch PPM 测试 Acer Predator GM7 1 TB SSD w/Maxio 主控+128 层 YMTC NAND 闪存
4.3	增加 QPS 图片和 Quarch 在 2023 年参加的一些 tradeshow 的图片
4.3.1	增加：两张针对 PCIe Gen5 U.2 SSD 和 M.2 SSD 的实际连接图片
4.3.5.1	增加了三相电源 PAM 分析工具在电动车充电上面的应用
4.3.6	增加了针对 IEC 220V 单相 AC 供电 PAM 分析模块
4.3.6.1	增加了 Quarch QTL2843 IEC Mains Power Analysis Module Review
4.3.7	增加：使用 PAM 分析 GPU/AI 卡/FPGA 加速器功耗
4.3.7	增加了简易图标比较 - PAM 和 PPM 的主要功能区别
4.3.8	增加了很多工程师困扰的”功耗测量：我需要什么采样率？“的疑问的解答
4.5.1	增加：你需要什么工具测试 CXL?
4.5.3	增加：PCIe Gen5 U.2 SSD 经济型掉电/上电/功耗计量/边带信号拉高/拉低测试工具
5	增加新的 PCIe Gen5 各类转接卡，延长线的型号
5.1.2	增加 Gen5 x16 retimer 卡，带 2 个 x8 QSFP-DD connector，有两款，芯片分别是 Astera 以及 Parade

5.2.1	在 MCIO 技术英文解释的下面，增加了“构建 PCIe Gen5 dual port SSD 仅找到一块 ssd 的解决办法小贴士”
6.1	增加了 intel 和 amd 两家公司关于 Gen6 CPU 的一些情况介绍
6.1.1	Intel Xeon PCIe Gen6/CXL 3.0 support
6.1.2	AMD PCIe Gen6/CXL 3.0 support
6.1.3	PCIe Gen6 over Fibre demo
6.1.3.2.3	增加”Phison E26 Max14um 2TB SSD 测试报告_2024“
7.2	增加：新型闪存开发电参数测量、特性测试和分析平台 TESTMESH TMA-100
7.3	增加了 NAND 协议分析仪
7.3.2	删除“7.3.2 ONFI Electrical Timing Analysis Software”
7.4	NAND 颗粒筛选设备删除了所有的设备图片
7.5	标题去掉“- Visual NAND Reconstructor”
7.12	修改部分格式
8.6.1	增加了“PCIe 协议分析仪 interposer 托架+夹具”
9	修订第 9 章节为新的 Protocol Insight UFS 4.0 analyzer 和 Introspect I3C analyzer/exerciser
10.2	增加了 Legacy 和 UEFI BIOS（Basic Input/Output System）执行过程的解释，UEFI 比 legacy 传统 BIOS 好的地方， PCIe 设备初始化中 cold reset 和 hot reset 的区别，以及 PCIe LTSSM recovery 发生的原因以及过程
11	速查部分增加新的 PCIe Gen5 各类转接卡，延长线的型号
11.4	增加了 Gen5 x16 parade Retimer 卡和 Phison Redirver 卡
12	增加 quarch 关于电压、电流、功耗、边带信号等测试的产品速查手册
16	增加了四篇关于 mini cooling device 的主动散热方案的介绍，针对 Gen5 M.2 SSD，以及 macbook air 等笔记本；以及其工作原理

搜索，添加并关注我们的公众号：Saniffer，认准下面的 LOGO，第一时间了解全球 PCIe Gen5/6, CXL 3.0, NVMe 2.0 测试、SSD 测试工具最新动态，相关信息等。

有任何问题请添加下面的微信号二维码，请注明：通过该白皮书扫描获得。或者致电：**021-50807071 / 13127856862, 发送邮件：sales@saniffer.com**

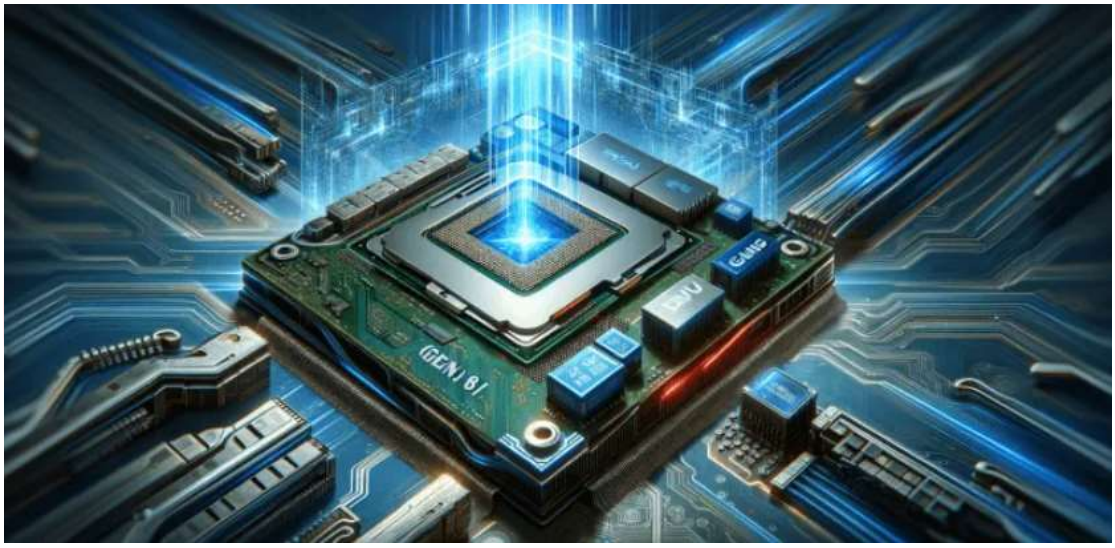
1. 前言

1.1 PCIe 6.0 和 PCIe 5.0 设计和测试带给业内的挑战

1.1.1 PCIe Gen5 与 Gen6：您需要了解什么？

发布日期：2024 年 1 月 25 日

PCI Express (PCIe) 的发展对于推动现代计算的性能发挥了重要作用。在比较 Gen5 与 Gen6 PCIe 时，您应该注意下面一些重大变化。



Gen5 与 Gen6 PCIe：人工智能是 Gen6 采用的一大推动力

1.1.1.1 是什么推动了 PCIe 向 Gen6 发展？

在过去的 PCIe 几代发展过程中，PCIe 性能是由多种设备的发展所驱动的，包括：

- **GPU**
- **NVMe 存储设备 (SSD)**
- **高速网络接口卡 (NIC)**
- **其他高速互连外设，例如 USB-C 和 Thunderbolt**
- **然而，随着 Gen5 PCIe 的出现，我们看到较小范围的设备受益于速度的提升。**

由于控制器和 NAND 闪存芯片的限制，大多数 SSD 不需要 Gen5 速度。Gen5 的优点是可以在使用一半数量的 PCIe 通道的情况下实现相同的速度。

同样，大多数 NIC 和其他外设可以从少量 Gen5 通道获得所需的带宽。

Gen5 和更高速度的主要驱动力现在来自游戏 GPU 和（最近）人工智能 AI 加速器的要求。在这两种情况下，通常使用 PCIe 插槽的完整 16 个通道，因此进一步提高性能需要升级到下一代。

到 2023 年下半年，我们看到了引入 Gen6 系统路线图的巨大推动力，多家公司计划在 2024 年发布重大开发版本。对 GPU 计算能力的需求几乎完全由人工智能发展驱动。总线速度加倍将给人工智能公司带来重大好处。

1.1.1.2 速度变化

Gen5 PCIe 的推出每通道的最大理论带宽为 32 GT/s（每秒千兆传输）。这相当于每通道大约 4 GB/s（每个 lane）或 x16 连接的 64 GB/s。

PCIe Gen6 更进一步，将带宽再次加倍至每通道 64 GT/s。这意味着每个方向上每个通道的速度约为 8 GB/s，对于 x16 设备而言，速度可达惊人的 128 GB/s。

请注意，这是“理论”带宽，由于协议的开销，实际性能会较低。

1.1.1.3 信号变化

- **PAM4**

  Gen5 PCIe 使用具有 2 个电压电平的 NRZ 信号，表示二进制 0 和 1 。Gen6 使用更复杂的 PAM4 系统，有 4 个电压电平。Gen5 和 Gen6 具有相同的 16 GHz 基频，但由于采用 PAM4 编码，Gen6 的抗噪能力仅为 Gen5 的三分之一，因此需要更严格的设计容差。了解 NRZ 和 PAM4 信令

- **前向纠错或 FEC**

  这是协议的新增内容，用于纠正传输中发生的较小错误，而无需重新传输数据。与第 5 代相比，这些额外的错误恢复数据增加了额外的开销。

- **流量控制单元 (FLIT)**

Gen 6 PCIe 引入了新的数据传输结构以及 FEC，以减少开销并实现更快的数据传输。这将需要新的解码系统，并且与第五代相比是一个巨大的变化。

1.1.1.4 电源效率

新的节能状态允许一些 PCIe Lane 通道关闭，而另一些 PCIe Lane 通道继续运行。这样可以在负载变化时实现可扩展的性能，同时最大限度地减少功耗。新的电源状态称为 L0p

1.1.1.5 连接器变化

市场上已经有适用于 AIC（插槽）设备的 Gen6 连接器。

适用于 E1、E3 和 CXL 的现代 EDSFF 连接器预计也将在 Gen6 版本中提供。

鉴于该连接器较旧的性质以及 SSD 上 Gen6 的优势较低，SFF-8639（用于 U.2 和 U.3 驱动器）很可能不会进入 Gen6。

1.1.1.6 板级布线设计带来的挑战

Gen6 信号对于板级走线和布线将是一个重大挑战。FEC 将有助于恢复较小的错误，但转向 PAM4 将显着减少系统中的 SI（信号完整性）开销。这将使数据容易因丢失和串扰而出错

- **损耗**

  Gen6 的总插入损耗预算为 32dB，低于 Gen5 规范中的 36dB。这是一个很小但很重要的变化，将限制走线的长度和转换的数量（连接器和类似的）
- **串扰**

  这是从一个通道到另一个通道的干扰（串扰）。使用 PAM4，干扰更改数据位的可能性显着增加。这使得 PCIe Gen6 的串扰风险更高，需要更复杂的设计来缓解。

1.1.1.7 Gen5 与 Gen6 PCIe 兼容性

与前几代 PCIe 一样，保持了向后和向前兼容性。旧设备应在新的 Gen6 插槽中运行，而 Gen6 设备在放入旧插槽时应降低速度。向后兼容性的要求显着增加了 Gen6 硬件的复杂性，特别是 SerDes，它必须支持 NRZ 和 PAM4 编码并在运行中在它们之间进行切换。

1.1.1.8 Gen6 测试工具和测试环境搭建

- **SerialTek 作为业内领先的协议分析仪厂家，已经推出了 PCIe Gen6 协议分析仪，训练器，可以模拟 PCIe Gen6 RC（CPU 端）以及 EP（各类 device controller 插卡），以及针对协议的兼容性测试套件 CTS**

- **Quarch 正在努力开发 Gen6 产品，因此当您拥有原型设备时，我们应该准备好用于功耗分析和热插拔/故障注入的测试解决方案**
- **SanBlaze 将于 2024 年下半年将推出支持 PCIe Gen6 SSD 和 SSD 卡的测试设备**
- **SerialCable 将于 2024 年将陆续推出支持 PCIe Gen6 SSD 和插卡的各种 Switch 卡，retimer 卡，测试盘柜，转接，延长线，转接延长线等等服务 PCIe Gen6 信号品质的搭建 PCIe Gen6 测试环境所需的基础组件**

1.1.2 PCIe Gen5 在过去 2 年在国内的发展回溯和总结

注意：下面是 2021 年底/2022 年初针对 PCIe Gen5 在国内和国内发展的一个回溯

随着高速计算机设计，GPU，DPU，MaPU，SSD，SmartNIC, 高性能网络，存储，以及 AI 人工智能等技术的发展，系统内部或者外联基本都是通过高速 PCIe 总线互联， 本文档专注于探讨美国、欧洲以及中国业内公司广泛使用的针对 PCIe Gen 4/5 协议诊断、分析、测试工具以及测试环境搭建过程中用到各种工具，尤其是针对基于 PCIe Gen 4/5 高速总线开发测试所用到的各种工具的图解剖析，相信读过本文档可以帮助你解决 PCIe Gen 4/5 测试工作中可能遇到的大部分疑问。

首先，我们从和全球知名的 PCIe 和 NVMe SSD 测试工具厂商的交流以及他们的销售统计来看一下全球 PCIe Gen 5 研发面临的挑战和趋势。

图 1-1

Sniffer 合作的全球知名的四家 PCIe/CXL/NVMe 测试工具厂商，即 SerialTek, SanBlaze, SerialCables, Quarch 大概从 3 年前开始都陆续发布 PCIe Gen5 相关的产品，通过最近和这些厂商的密切沟通交流，以及和国内客户的沟通，针对 Gen5 产品开发我们发现存在下面几个挑战和趋势：

1）PCIe Gen5 诊断分析工具存在很大挑战

其实这部分调整不仅仅体现在测试工具上面，同时也体现在下述涉及的搭建 Gen5 测试环境。在研发阶段，不仅需要高性能的示波器，对于大部分芯片开发来讲也需要 PCIe Gen5 协议分析仪。众所周知，协议分析仪必须将对应接口的 interposer 串接在链路中间，由于 Gen5 对于信号的高要求，目前这种机制已经带来了大量的问题，包括建链不成功，

无法复现故障，抓不到数据，抓到的全是错误，等等。据 SerialTek 公司 VP 介绍，在美国，自从 2021 年下半年起遇到大量客户涌入要求试用 Gen5 分析仪，这其中包括不少前期已经购买了 Gen5 分析仪的公司，这从侧面间接也说明了信号问题的严重性。在国内，我们从客户端也发现了一样的问题。

2）PCIe Gen5 测试环境搭建成为目前的突出问题

目前 Intel 仅提供桌面级的 Gen5 CPU，几个主要的主板厂商的 Gen5 x16 插槽信号质量都不大好。我们大概两个月前购买并测试了 A 厂商的产品，Gen5 x16 host card 插入其唯一的 Gen 5 x16 slot 只能和 Intel Gen 5 CPU 协商成 Gen 3，大概 2 周前 A 公司发布新的 BIOS，升级后总算可以协商成 Gen 4。B 公司产品到目前为止也仅能协商成 Gen 4。总体来看，两台设备的 Gen5 插槽信号质量堪忧。

一些前期的 Gen 5 用户，采用转接卡进行转接的时候信号质量出现严重衰减。Gen 5 对于 PCB 板材，过孔，接插件的要求非常高，并且目前在全球物料短缺的情况下很多物料很难买到。PCIe Gen5 的信号质量成为一个突出问题，尤其是涉及到转接，延长等问题的时候。这个在 Gen 5 问题诊断的时候给用户带来了极大的挑战。国外一些专家甚至建议在采用 PCIe Gen 5 系统设计的时候每隔一个比较短的距离就增加一个 Gen5 retimer。

3）PCIe Gen5 SSD 的接口选择的趋势

下面是 Saniffer 公司和 Quarch CTO 在 2021 年 12 月底沟通的总结：

Quarch 在 2021 年全年各类产品出货中，各种接口的测试模块的顺序如下：

Gen 4 NVMe SSD 测试模块出货量最大的是 U.2, 其次是 U.3 和 EDSFF (E1)。

Gen 5 NVMe SSD 测试模块，目前出货量最大的仍然是 U.2, 其次是 EDSFF (E3)，后面才是 U.3。Gen 5 插卡模块也很受欢迎，主要用于芯片等验证卡的早期产品测试，例如芯片早期 bring-up 等测试。根据目前的趋势来看，2022 年 U.2 和 EDSFF E3 将增长非常迅速。EDSFF 接口总体上将非常快的受到市场的欢迎。

另外，根据我们和 SanBlaze VP 的沟通，目前 SanBlaze 大量出货的 PCIe Gen5 RM5 默认配置就是 16 个 Gen5 E3.S 插槽（对比 Gen4 RM4 默认配置是 16 个 U.2 插槽）。同时，我们也看到一些存储系统业界主流厂商基于 Gen 5 NVMe SSD 的首发全闪存阵列也多采用 E3.S。还有，国内目前开发企业级 NVMe SSD 的公司很多都从我们公司购买了 SerialCables 公司的 EDSFF 各种转接卡和线缆。这些都从侧面印证了 EDSFF 会在 2022 年会获得不错的进展。

针对电源拉偏测试设备 PPM（Programmable Power Module），销量按照 U.2/U.3, EDSFF, AIC 和 M.2 依次递减。

当然，如果考虑到 M.2 在低功耗状况下各种问题的分析，针对电压/电流/Sideband Signal 的实时和回溯分析监测产品 PAM（Power Analysis Module）对于客户的吸引力显

然更大，使用 PAM 进行问题分析的效率大大超过了使用示波器，Keysight 后者吉时利 Keithley 传统电源类产品的效率，成为目前业内分析 M.2 低功耗问题的首选设备。

国内主流的 M.2 SSD 厂商、系统商（如笔记本厂商），以及第三方政府评测机构都大量采用这 PAM 设备做问题故障分析和横向对比评测。

4）国内领先的公司都在 2022 年开始 PCIe Gen 5 项目

这个不仅仅体现在 SSD controller 的 drive 盘领域，同时也涉及到其它各种需要使用高带宽的芯片和板卡设计领域。区别仅是 2022 年上半年还是下半年而已。

下面是 Saniffer 公司更新的针对 PCIe Gen 5 总线以及 NVMe SSD 的测试工具汇总白皮书 10.0 版本，该版本和之前 9.0 版本做了很多改动，可以扫描下面的微信或者访问 Saniffer 官网链接下载。

提示：访问 saniffer 官方网站（https://www.saniffer.com/cn/downloads/）下载"PCIe Gen5&CXL 总线协议分析和 NVMe SSD 测试工具白皮书"文档获得更多信息，也可以扫描下面的微信获取。

如果你还有其它方面的技术问题，或者针对 PCIe Gen5&6，UFS 4.0, DDR5 等最新技术的进展情况感兴趣，请扫描下面的微信二维码咨询。

1.2 关于 Saniffer 公司

Saniffer 公司位于上海张江高科技园区，是国内专注于计算、网络、存储以及移动领域的测试工具综合服务提供商，产品涉及各种总线协议从协议层到应用层测试的各类研发测试工具。

随着 2012 年 NVMe SSD 技术的起步，Saniffer 成为中国 SSD 测试工具领域的知名服务提供商，我们提供的测试工具涉及了 SSD 研发、测试过程中常用的各种工具，包括 SAS/SATA/PCIe/NVMe 协议分析仪，SSD 性能/功能/协议兼容性/IOT 测试，热插拔自动化测试，故障导入，电压拉偏&功耗测试，电压/电流/功耗/sideband 信号被动监测，掉电测试，高低温测试，以及如何构建 PCIe Gen 4/5/6 测试环境，从主板和 Host Card 选型开始，涉及 NVMe SSD 各类接口的端口扩展和转换，各种常用的主机卡，转接卡，盘柜，延长线的选择等，以及考虑到测试便利性使用的主板托架和实验室批量测试机架等解决方案。

我们提供的测试工具产品和相关技术不仅适合从事 NVMe SSD 的 firmware 开发/测试(FW, FTE - Firmware Testing Engineering)部门学习，也适合于 NVMe SSD 架构设计(Architect)，产品验证(PVE - Product Validation Engineering)，产品工程/测试工程（PE/TE - Product Engineering/Test Engineering)，客户方案工程（CSE - Customer Solution Engineering），应用工程（AE），技术支持（FAE - Field Application Engineer），失效分析（FAE - Failure Analysis Engineer）等部门。同时，这些产品和技术也适用于研发/测试各类通用 PCIe Gen 4/5 控制器芯片、板卡工程师学习，例如显卡，RAID 卡，HBA 卡，网卡（含有线以及 M.2 WIFI 网卡）各种用途板卡，以及主机、网络、存储系统类工程师使用。

我们提供的测试产品和工具覆盖了 SSD 存储生态的各个环节，从芯片开发，底层固件和驱动开发/验证，测试工程，应用工程，RDT 可靠性测试，一直到生产测试。

Saniffer 提供的测试产品涉及下面的相关技术：

1.2.1 计算/网络/存储相关总线技术

- **PCIe Gen 4/5/6**
- **NVMe 1.4, 2.0**
- **SAS 12G, 24G**
- **SATA 6G**
- **NAND 400MT, 800MT/1.2GT/1.6GT, 2.0/2.4GT**
- **LPDDR/DDR 4/5**
- **FC 32/64G**
- **FCoE 10GE**
- **iSCSI 10GE/40GE/25GE/100GE**

- **NVMoF (NVMe over Fabric)　100GE**
- **FC-NVMe (NVMe over FC)　32/64G**

1.2.2 消费类/移动/汽车电子相关总线技术

- **UFS 3.0/3.1 & 4.0**
- **eMMC 5.1**
- **SD/SDIO 3.0, SD 4.0**
- **USB 3.0/3.1/3.2/4.0**
- **CAN**
- **LIN**
- **FlexRay**
- **Display Port**
- **100Base-T1**
- **TTE**
- **TSN**
- **I2C/SPI, QSPI**
- **I3C**
- **SPMI**
- **RFFE**
- **WIFI**
- **… …**

我们提供下述各种常见的各种常用总线的协议分析和仿真工具，以及某些总线的基于 Tek 示波器的协议解码和 Electrical Validation 验证。

Protocol /Interface	Protocol Analyzer	Exerciser	Protocol Decode (Scope Based)	Electrical Validation (Scope Based)
I2C	✓	✓	✓	✓
SMBus	✓	✓		
PMBus	✓	✓		
I3C	✓	✓	✓	✓
UART	✓	✓	✓	✓
SPI	✓	✓	✓	✓
QSPI	✓	✓	✓	✓
eSPI	✓	✓	✓	
RFFE	✓	✓	✓	
SPMI	✓	✓	✓	TBA
SMI (MDIO)	✓	✓		
JTAG	✓	✓	✓	
100BaseT1 (Automotive)	✓		✓	
eMMC	✓		✓	✓
SD	✓		✓	✓
SDIO	✓		✓	✓
UFS 4.0 (backward Compatible)	✓		✓	
UniPro			✓	
LLI			✓	
I2S	✓	✓	✓	✓
HSIC			✓	
DigRF v4			✓	
SSIC			✓	
ONFI v4				✓
USB PD, 3.1, 3.0, 2.0			✓	
HDMI, MHL			✓	
FlexRay			✓	

图 1-2

1.2.3 UNH IOL 官方认证的 SerialTek, SanBlaze, Quarch 中国独家合作伙伴

随着近几年 PCIe Gen 4/5/6 技术及 NVMe SSD 在国内的快速发展，Sniffer 迅速成为国内在该领域的知名供应商，成为 ***UNH IOL 认证的 SerialTek, SanBlaze, Quarch 在中国的独家合作伙伴。***

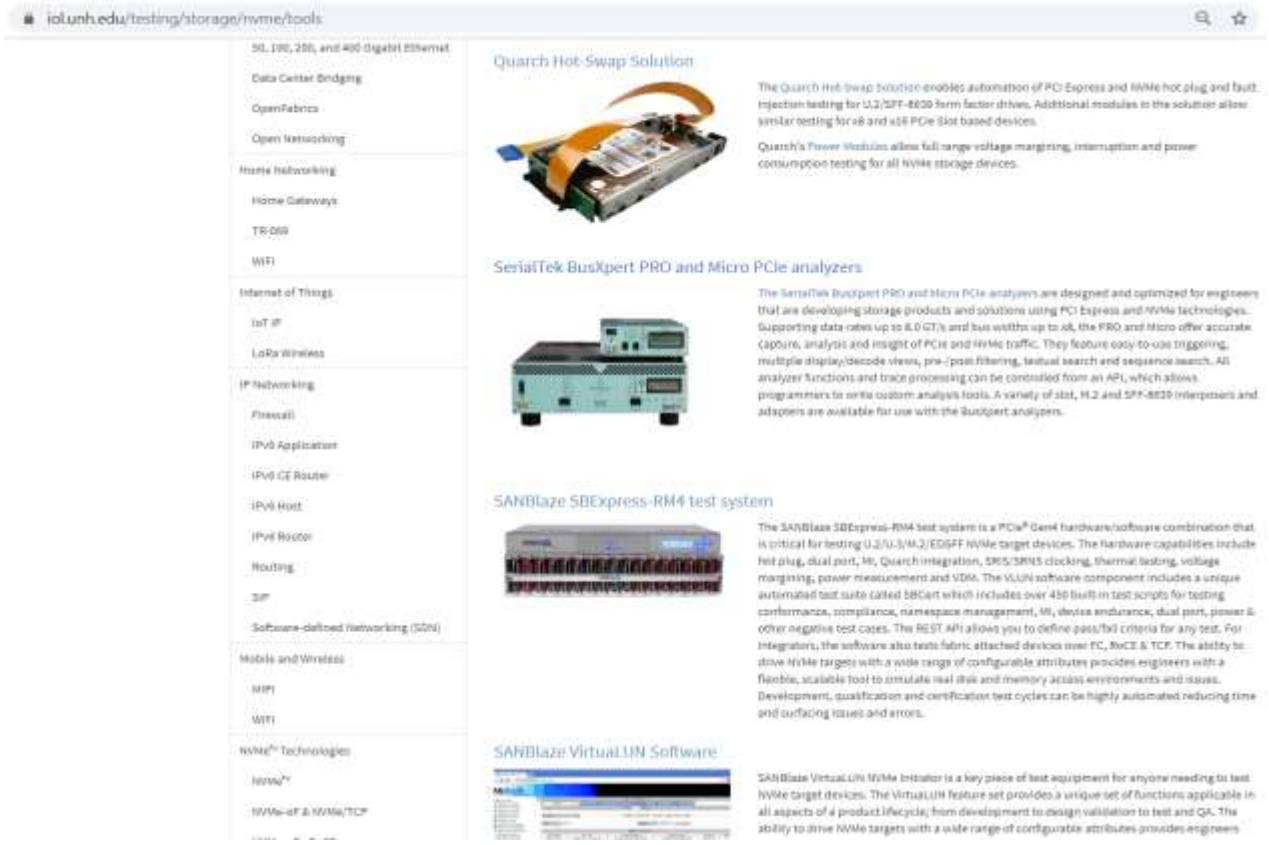


图 1-3 UNH IOL 官网认证推荐产品

1.3 关于 Sniffer 开放实验室

为了更好地服务于国内众多的中小芯片设计公司，Sniffer 公司发挥其在总线协议分析业内累积近 18 年的测试工具相关经验，在上海张江高科技园区设立开放实验室，为涉及 PCIe, NVMe, SAS/SATA, NAND, DDR5, UFS 等总线的客户提供免费和付费的协议分析、问题诊断和测试等相关服务，有需求的用户可以通过下面的方式提前两周预约：021-50807071 / 13127856862，sales@sniffer.com。

注：Sniffer 开放实验室提供各种常用的总线接口的协议分析和测试工具，不仅仅支持各类 NVMe SSD 接口类型；其 PCIe 协议分析和测试工具也支持各类插卡类应用，包括各种计算，网络，存储系统开发和测试等领域。

目前 Sniffer 开放实验室库存提供如下测试工具：

1.3.1 PCIe 协议分析仪

1.3.1.1 PCIe Gen 6 x16 协议分析仪和训练器

SerialTek/Ellisys，该款分析仪适用于各种 PCIe Gen6 X16 插卡诊断分析和协议训练，例如 CPU, GPU, DPU, AI，加速卡等；支持 CXL 协议分析，最高配置，288G BUFFER, 4TB trace 内置闪存 SSD，2 个 10GEtrace 文件导出端口，该设备目前提供 Gen6 x16 插卡，以及 M.2, EDSFF 等常用接口 interposer。

1.3.1.2 PCIe Gen 5 x16 协议分析仪

SerialTek/Ellisys，该款分析仪适用于各种 PCIe Gen5 X16 插卡诊断分析，例如 CPU, GPU, DPU, AI，加速卡等；支持 CXL 协议分析，最高配置，144G BUFFER, 2TB trace 内置闪存 SSD，2 个 10GEtrace 文件导出端口，该设备目前提供 Gen5 x16 插卡，以及 M.2, U.2, U.3, E1.S, E1.L, E3,S E3.L 等常用接口 interposer。

1.3.1.3 PCIe Gen 5 x4 协议分析仪

SerialTek/Ellisys，该款分析仪适用于各种 PCIe Gen5 NVMe SSD 诊断分析，支持 CXL 协议分析，最高配置，144G BUFFER, 2TB trace 内置闪存 SSD，2 个 10GEtrace 文件导出端口，目前含 Gen5 x4 插卡，以及 M.2, U.2, U.3, E1.S, E1.L, E3,S E3.L 等常用接口 interposer。

1.3.1.4 PCIe Gen 4 协议分析仪

SerialTek/Ellisys，最高配置，144G BUFFER, 2TB 内置闪存盘，无需设置过滤条件，秒级解码，大大提高用户到开发实验室进行问题诊断的效率。提供针对各种 SSD 接口 U.2/U.3 Single Port/Dual Port, M.2, AIC, EDSFF, 以及用于存储盘柜互联的 Cable 等 5 种协议分析 Interposer

1.3.1.5 PCIe Gen 3 协议分析仪

SerialTek 公司产品，最高配置，36G BUFFER。提供 U.2, M.2, AIC 三种最常用的 SSD 接口协议分析 Interposer

1.3.2 SAS/SATA 协议分析仪

1.3.2.1 12G SAS/SATA 协议分析仪

SerialTek 12G SAS 协议分析仪，兼容 6G SAS/SATA

1.3.2.2 6G SAS/SATA 协议分析仪

SerialTek 6G SAS/SATA 协议分析仪，提供如下协议测试功能：

- **analyzer 协议分析**
- **generator 发包**
- **jammer 故障注入**
- **initiator emulator （6G SAS 主机仿真）**
- **host emulator（6G SATA 主机仿真）**
- **target emulator（6G SAS 硬盘仿真）**
- **device emulator（6G SATA 硬盘仿真）**

1.3.3 SSD 性能/功能测试设备（研发测试）

1.3.3.1 PCIe Gen 4 NVMe SSD 测试设备

SanBlaze DT4 研发/测试设备，该设备提供 700+测试用例，支持测试各种消费类和企业级 NVMe SSD 的性能，功能（例如：各种 admin 命令，I/O,reset, namespace 管理，MI(SMBus and MCTP over PCIe), dual-port, hotplug, link testing, signal glitch, ZNS, SRIS, TCG, VDM,数据中心 NVMe DSSD 2.0 特性测试, SR-IOV, UNH IOL NVMe v1.4 和部分 NVMe 2.0 certification, UNH IOL NVMe-MI v1.4 certification, JEDEC endurance test 等），支持各种 SSD 接口 PCIe Gen 4 NVMe SSD：U.2, U.3, M.2, EDSFF E1.S, E1.L, AIC 等

1.3.3.2 12G SAS SSD 测试设备

SanBlaze, 支持测试 6G SAS/SATA SSD

1.3.4 SSD 热插拔自动化测试设备

1.3.4.1 PCIe Gen 5 插卡控制模块

Quarch 公司产品，支持 PCIe Gen 5 x16，提供模拟插卡通/断，底层信号问题导入 bit error/CRC error，模拟某些针脚虚焊断掉，接触不好等各种异常。

1.3.4.2 PCIe Gen 4 热插拔模块

Quarch 公司产品，提供 Gen 4 U.2, U.3, EDSFF E1.S, E1.L 等各种支持热插拔的盘的自动化测试模块，也提供针对 PCIe Gen 4 x16 插卡和 M.2 SSD 的控制模块，提供模拟插卡通/断，模拟底层信号问题导入 bit error/CRC error，模拟某些针脚虚焊断掉，接触不好等各种异常。

1.3.4.3 PCIe Gen 3 热插拔模块

Quarch 公司产品，其它同 Gen 4 热插拔模块

1.3.4.4 12G SAS 硬盘热插拔模块

Quarch 公司产品, 兼容 6G SAS/SATA

1.3.4.5 6G SAS/SATA 热插拔模块

Quarch 公司产品

1.3.4.6 12G SAS 线缆自动化切换模块

Quarch 公司产品

1.3.4.7 6G SATA 自动化切换模块

Quarch 公司产品

1.3.4.8 12G SAS 线缆热插拔模块

Quarch 公司产品

1.3.4.9 6G SAS/SATA 线缆热插拔模块

Quarch 公司产品

1.3.5 SSD 电压拉偏，功耗测试，电压/电流检测自动化设备

1.3.5.1 电压拉偏和功耗主动测试工具 PPM (Programmable Power Module)

Quarch 公司产品，提供针对各种接口的 SSD 的电压拉偏，最高可超过标准 20%输出，最低可以降到 0V, 可以方便模拟各种供电异常和故障，电压输出精度 4mV，同时提供电压/电流/功耗的实时采样（最高 250K/s，测量精准度 4Mv/25uA）和分析、回溯分析等功能，支持 PCIe Gen 4 U.2, U.3, M.2, AIC, EDSFF E1.2/E1.L 等，以及 12G SAS, 6G SAS, 6G SATA SSD 等

1.3.5.2 电压/电流/功耗被动分析工具 PAM (Programmable Analysis Module)

Quarch 公司产品，支持 PCIe Gen AIC, U.2/U.3, M.2 等主流 SSD 接口以及 SAS/SATA SSD，PAM 模块串接在盘和背板/主板或插卡和 PCIe 插槽之间，可以实时分析或者长时间记录监测电压/电流/功耗用于分析，尤其对于分析应用于笔记本电脑的 PCIe M.2 SSD 的低功耗模式非常有帮助。

1.3.6 PCIe Gen 4/5 SSD 测试环境

1.3.6.1 PCIe Gen5 主机

Intel Xeon Server, Gigabyte, Asus, AsRock, MSI 等基于 Intel 和 AMD Gen5 CPU 主机

1.3.6.2 PCIe Gen5 Host Card

SerialCables 公司基于 BCM 的 Gen5 switch card

1.3.6.3 PCIe Gen5 Retimer Card

Astera Gen5 retimer card

1.3.6.4 PCIe Gen5 转接卡

SerialCables 公司 U.2/AIC 转接卡

1.3.6.5 PCIe Gen5 延长线

Gen5 x16 qualified 延长线

1.3.6.6 PCIe Gen4 主机

Gigabyte, Asus, AsRock, MSI 等基于 Intel 和 AMD Gen5 CPU 主机

1.3.6.7 PCIe Gen 4 转接卡

SerialCables 公司，包括 M.2/AIC, M.2/U.2, U.2/AIC 等

1.3.6.8 PCIe Gen 4 Host Card

SerialCables 公司基于 BCM, MicroChip 的 Gen 4 Host Card

1.3.6.9 PCIe Gen 4 SSD 盘柜

SerialCables 公司的 Active/Passive 盘柜，支持 8 盘位 SSD

1.3.6.10 PCIe Gen 4 Retimer Card

SerialCables 基于 Astera Gen4 retimer card

1.3.6.11 PCIe Gen 4 延长线

SerialCables U.2 延长线，M.2 延长线，PCIe slot 延长线等

1.3.6.12 PCIe Gen 4 M.2 掉电卡

SerialCables M2U2, M2/AIC 掉电卡

1.3.7 NAND 特性测试和分析设备

NplusT 公司产品，我们目前维护两台 NAND 特性分析和测试工具，支持标准版 STD 测试模块（800MT）和 HS 测试模块（1.6GT）两种速度，最新的 2.4GT 测试模块在协调过程中。

1.3.8 DDR5 协议分析仪

JKI 公司产品，目前提供 DDR5 RDIMM interposer，兼容 Samsung, SK Hynix，Micron 等各种主流 DDR5 内存。

1.3.9 上海开放实验室 PCIe Gen5 测试环境部分清单汇总

欢迎各位到我们位于上海张江高科技园区的 Saniffer 实验室进行测试，地址：秋月路 26 号 1 号楼，来之前扫描本文底部的二维码联系。

存储/计算/网络协议测试开放实验室

1.3.9.1 PCIe Gen5 测试主机

- **Intel Gen5 Xeon 服务器和 AMD Gen5 Genoa 服务器**
- **AMD Gen5 Genoa 服务器，Super Micro**
- **Intel 12 代酷睿 Gen5 CPU + Z690 工作站，包括 Asus，Gigabyte，MSI，AsRock**
- **Intel 13 代酷睿 Gen5 CPU + Z790 工作站，Asus**
- **Intel 12 代酷睿 Gen5 CPU + B660 工作站，Asus**
- **AMD Gen5 CPU + X670E 工作站，包括 Asus，Gigabyte，MSI，AsRock 四台**

1.3.9.2 PCIe Gen5 测试外设

- **SerialCables PCIe Gen5 Switch 卡 qty: 2**
- **Gen5 x16 GPU 卡，qty: 2**
- **Gen5 x16 Retimer 卡，qty: 1**
- **Gen5 x4 Kioxia CD7 E3.S SSD**
- **Gen5 x4 Samsung PM1743 U.2 SSD**
- **Gen5 x4 Inland M.2 SSD**
- **Gen5 x4 Gigabyte AORUS 10000 M.2 SSD**

1.3.9.3 PCIe Gen5 测试环境

- **SerialCables Gen5 x4 SSD 8 盘位盘柜 qty: 1**
- **SerialCables Gen5 x4 SSD U.3/E3.S 转接卡 qty: 2**
- **SerialCables Gen5 x4 SSD U.2/E3.S 转接卡**

- SerialCables Gen5 x4 SSD M.2/E3.S 转接卡
- SerialCables Gen5 x4 SSD U.2/AIC 转接卡
- SerialCables Gen5 x4 SSD E3.S/AIC 转接卡
- SerialCables Gen5 x4 SSD E1.S/AIC 转接卡
- SerialCables Gen5 x4 SSD M.2/AIC 转接卡 qty: 2
- SerialCables Gen5 x4 SSD M.2/U.2 转接卡 qty: 2
- SerialCables Gen5 x4 MCIO x4/U.2 1x4 线缆 qty: 4
- SerialCables Gen5 x4 MCIO x4/U.2 2x2 线缆 qty: 4
- SerialCables Gen5 x4 MCIO x4/E.3 1x4 线缆 qty: 4
- SerialCables Gen5 x4 MCIO x4/E.3 2x2 线缆 qty: 4
- SerialCables Gen5 x4 MCIO x4/U.3 1x4 线缆 qty: 4
- SerialCables Gen5 x4 MCIO x4/ U.3 2x2 线缆 qty: 4
- SerialCables Gen5 x4 MCIO x4 / MCIO x4 延长线 0.5m qty: 4
- SerialCables Gen5 x4 MCIO x4 / MCIO x4 延长线 1m qty: 4
- SerialCables Gen5 x16/x4 reducer 卡 qty: 2
- SerialCables Gen5 x16 lane swizzling card qty: 3
- SerialCables Gen5 x8 lane swizzling card qty: 1
- SerialCables 2* Gen5 x2/U.2 延长转接线 qty: 1
- SerialCables Gen5 x4 E3.S/U.2 转接卡 qty: 1
- SerialCables Gen5 x4 MCIO/x8 MCIO 适配卡
- Astera Gen5 x16 Retimer 卡
- 硅谷定制 Gen5 x16 延长线 1m Qty: 1
- 定制 Gen5 x16 延长线 A 工厂 0.15m, 0.3m, 0.5m, 1m Qty: 1
- 定制 Gen5 x16 延长线 B 工厂 0.15m, 0.3m, 0.5m, 1m Qty: 1

1.3.9.4 PCIe Gen5 协议分析仪

- SerialTek PCIe Gen5 x16 协议分析仪
- SerialTek PCIe Gen5 x4 协议分析仪

1.3.9.5 PCIe Gen5 SSD 测试设备

- SanBlaze DT4, DT5

1.3.9.6 PCIe Gen5 热插拔、故障注入、电压拉偏等设备

- **Quarch Gen5 各种企业级 SSD，包括 U.2, U.3 E1.S, E3.S 热插拔和故障注入工具**
- **Quarch Gen5 x16 插卡、Gen5 x4 M.2 SSD 故障注入工具**
- **Quarch Gen5 Programmable Power Module (PPM)**
- **Quarch Gen5 Power Analysis Module (PAM)**

如果你对 Saniffer 上海开放实验室感兴趣，或者希望获得更多测试工具的产品信息和经验分享，请扫描下面的二维码联系我们，或者搜索微信公众号关注“Saniffer”。

1.4 Sniffer 技术讲座和培训视频录像

Sniffer 公司在 2022 年 2~4 月份上海疫情封城期间连续组织了 7 次针对 NAND, SSD, PCIe Gen5，数据中心架构、虚拟化、企业级 NVMe SSD 特性的讲座，下面是讲座的视频回放连接，需要 PPT 可以扫描本文档首页的二维码，或拨打我们电话获取。

Sniffer 在 2025 年上半年也拍摄了大量培训视频供读者参考。由于一台 PCIe Gen5 协议分析动辄几百万人民币，普通工程师无法接触到，所以对于 PCIe Gen5 协议分析仪有一种神秘感。我们从 2023/5 月底开始通过大概两周时间，以视频讲解的方式来依次解密平时使用最多的两类 PCIe Gen5 协议分析仪：

- **Gen5 x4 协议分析仪**
- **Gen5 x16 协议分析仪**

其中，Gen5 x4 协议分析仪分为下面几期讲解（参见下面 1.3.1 的 15 开始的部分）：

- **1. SerialTek PCIe Gen5x4 协议分析仪开箱介绍**
- **2. SerialTek PCIe Gen5 协议分析仪组成和架构简介**
- **3. SerialTek PCIe Gen5 协议分析仪 Gen5 Pod + AIC Interposer 连接演示**
- **4. SerialTek PCIe Gen5 协议分析仪 Gen5 Pod + U.2&U.3 Interposer 连接演示**
- **5. SerialTek PCIe Gen5 协议分析仪 Gen5 Pod + M.2 Interposer 连接演示**
- **6. SerialTek PCIe Gen5 协议分析仪 Gen5 Pod + EDSFF (E1.S, E1.L, E3.S, E3.L) Interposer 连接演示**
- **7. SerialTek PCIe Gen5 x16 WebGUI 软件界面演示(Gen5 x4 分析仪操作类同）**

Gen5 x16 协议分析仪分为下面几期讲解：

- **1. SerialTek PCIe Gen5 x16 协议分析仪开箱视频**
- **2. SerialTek PCIe Gen5 x16 协议分析仪架构组成和 x16 slot interposer 连接实操演示**
- **3. SerialTek PCIe Gen5 x16 协议分析仪开机演示**
- **4. SerialTek PCIe Gen5 x16 协议分析仪拆机过程演示**
- **5. SerialTek PCIe Gen5 x16 WebGUI 软件界面演示(Gen5 x4 分析仪操作类同）**

感兴趣的朋友可以将 Saniffe 公众号链接转发给同事参考。

1.4.1 2023 年上半年 PCIe Gen5 演示部分视频汇总

1. 目前实测唯一可达 PCIe Gen5 x16 速度的 1 米 PCIe 插槽延长线！！！

https://mp.weixin.qq.com/s/K91s5yqyw-tmE8tSmtYnbw

2. 双端口（Dual Port) NVMe SSD 技术简介和演示

https://mp.weixin.qq.com/s/mHiygKNtbwJnwz-5qqmb4Q

3. DDR5/LPDDR5 协议分析仪简介

https://mp.weixin.qq.com/s/Aa32jcfwhr5kmXwyOGEM4Q

4. Kioxia Gen5 x4 E3.S SSD CD7 在华硕 AMD 和 Intel Gen5 工作站的初始化过程

https://mp.weixin.qq.com/s/EmywOS9G_4JhsjpXbVZ2cQ

5. PCIe Gen5 E3.S SSD 测试盘柜，switch 卡，转接卡手把手演示

https://mp.weixin.qq.com/s/hj9wXDWbNMwBhQwkpmewyw

6. SerialTek PCIe Gen5 分析仪和 Gigabyte AORUS Gen5 M.2 SSD 抓包分析演示

https://mp.weixin.qq.com/s/VXbFLdIT5GKfNoNwB9sRlw

7. 在 Gen5 CPU 和 Gen5 SSD 中间串接 Gen5 switch 后对于性能的影响演示

https://mp.weixin.qq.com/s/wB0IpOZ9B8UA9uZ6kQYDWg

8. PCIe Gen5 switch+retimer+1m extension cable+SerialTek Gen5 分析仪演示

https://mp.weixin.qq.com/s/6y3ZcEwWJBN-bVMI4hOq0Q

9. PCIe 协议分析仪 debug 诊断分析 Samsung PM1743

https://mp.weixin.qq.com/s/ma9mCmx4WICWPatz9oMbHg

10. 国产 CPU 和国产企业级 NVMe SSD 碰到的问题分析

https://mp.weixin.qq.com/s/AoxACsIJcIi3GQlfnqAPpQ

11. 国产 100GE DPU 碰到的问题分析

https://mp.weixin.qq.com/s/2p_4U6jmqV6v-TsBztKGyg

12. 国产高端企业级 Gen5 SSD 和消费类 Gen5 M.2 SSD 的接收端信号对比 https://mp.weixin.qq.com/s/zzbM9CcvbV3EhdGxRCuJog

13. 售价仅几万美元的 PCIe Gen5 协议分析仪值得买吗？

https://mp.weixin.qq.com/s/ieMLV0bk6uaSwQsAx3PRFw

14. PCIe Gen5 协议分析仪和协议仿真卡需要买一家的吗?

https://mp.weixin.qq.com/s/SIC_Q5C-pFYZmbkms-Pidw

15. PCIe Gen5 协议分析仪抓取 Astera Gen5 x16 retimer 卡和不同 device 之间的初始化分析视频

https://mp.weixin.qq.com/s/kKE4Asckrey37XOr7xlAfQ

16. 两块 Broadcom PCIe Gen5 switch 卡对接加电初始化简析

https://mp.weixin.qq.com/s/DtVZ6ETyGCjj5geUl7mKyg

17. PCIe Gen5x4 协议分析仪解密系列 - 1. SerialTek 协议分析仪开箱介绍

https://mp.weixin.qq.com/s/0j0WwKjCeeOj-Ln0h2UIpw

18. PCIe Gen5x4 协议分析仪解密系列 - 2. SerialTek PCIe Gen5 协议分析仪组成和架构简介

https://mp.weixin.qq.com/s/RZBMdxEWqedmqawkj9TSMw

19. PCIe Gen5x4 协议分析仪解密系列 - 3. PCIe Gen5 协议分析仪 Gen5 Pod + AIC Interposer 连接演示

https://mp.weixin.qq.com/s/xw4F0HBTiXuKk3RoWVXLhQ

20. PCIe Gen5x16 协议分析仪解密系列 - 4. 协议分析仪 Gen5 Pod + U.2&U.3 Interposer 连接

https://mp.weixin.qq.com/s/8bRLdAcEJtK3whoPRL2N8w

21. PCIe Gen5x4 协议分析仪解密系列 - 5. Gen5 Pod+M.2 Interposer 连接演示

https://mp.weixin.qq.com/s/I-3qRTAz8pg8fTzIgkO-5w

22. PCIe Gen5x4 协议分析仪解密系列 - 6. Gen5 Pod+EDSFF (E1,E3) Interposer 连接演示

https://mp.weixin.qq.com/s/TUOk5fkwnul_xQrXJl5g4w

23. PCIe Gen5x4 协议分析仪解密系列 - 7. SerialTek PCIe Gen5 x16 WebGUI 软件界面演示(Gen5 x4 分析仪操作类同）

https://mp.weixin.qq.com/s/4YS_iIBxqeIobjDxbI2kIQ

24. PCIe Gen5 x16 协议分析仪解密系列 - 1. SerialTek PCIe Gen5 x16 协议分析仪开箱视频

https://mp.weixin.qq.com/s/YG0Q14UQ1X8OyGtH25yOmw

25. PCIe Gen5x16 协议分析仪解密系列 - 2. PCIe x16 分析仪架构和 x16 slot interposer 连接演示

https://mp.weixin.qq.com/s/zXe-SH-e2FPzP_u0Vpmfww

26. PCIe Gen5x16 协议分析仪解密系列 - 3. SerialTek PCIe Gen5 x16 协议分析仪开机演示

https://mp.weixin.qq.com/s/rpFkeIE5SbMZ7EbjkvtdZg

27. PCIe Gen5x16 协议分析仪解密系列 - 4. SerialTek PCIe Gen5 x16 协议分析仪拆机过程演示

https://mp.weixin.qq.com/s/d_oIpu7xX-62sOK0IF9qnA

28. PCIe Gen5x16 协议分析仪解密系列-5. Gen5 x16 WebGUI 软件界面演示

https://mp.weixin.qq.com/s/ZVP7YoWm_zTnsjiu250Ozg

29. 益企研究院使用 Quarch PAM 测试 Solidigm D5 企业级 SSD

https://mp.weixin.qq.com/s/nwu3rDoNEYNfvzF-ToBIgQ

30. Quarch 热插拔_故障注入模块 + PAM 联合测试演示

https://mp.weixin.qq.com/s/xhEPqqwKkRPlwgAeIGSsPQ

1.4.2 2022 年 3/4 月技术讲座汇总

1. 面向 SSD 开发的 NAND 特性分析技术讲座 - 3.19.2022 by Tamas Kerekes

https://mp.weixin.qq.com/s/3q40uXw_x_DkXbb90aVIKA

2. PCIe Gen5 诊断测试经验分享讲座 - 3.26.2022 by Michael Wang

https://mp.weixin.qq.com/s/aNON2X7ekarN9F4rD7-ppA

3. PCIe Gen5 热插拔以及功耗测试和分析讲座 - 4.2.2022 by Andy Norrie

https://mp.weixin.qq.com/s/y4H9fVigFK89JWtE7zZ77w

4. 企业级 NVMe SSD 新特性简介以及测试讲座 - 4.9.2022 by Michael Wang

https://mp.weixin.qq.com/s/ejQelBklyL-G-Q70b2y63g

5. 图解数据中心基础设施, PCIe 总线和 NVMe SSD 测试讲座 - 4.16.2022 by Michael Wang

https://mp.weixin.qq.com/s/QdXtTJ-NKvlCCd2LS3e6sQ

6. 图解数据中心基础架构、虚拟化、主机/存储/SSD 相关技术及测试技术讲座 - 4.23.2022 by Michael Wang

https://mp.weixin.qq.com/s/DYy2gE6BVDGAUibeX6gaMA

7. 图解企业级 dual-port SSD 和 SR-IOV 技术和相关测试技术讲座 - 4.30.2022 by Michael Wang

https://mp.weixin.qq.com/s/oMz0ifRCwAt5sOIFD5gHLg

1.4.3 2019 年底和 2020 年初技术讲座

1. SSD 测试技术讲座第一期: NAND 特性分析技术分享 2019.11.16 by Tamas Kerekes

https://v.qq.com/x/page/o30228mhtx0.html

2. SSD 测试技术讲座第二期: PCIe Gen3/4 NVMe SSD 测试技术分享 2019.11.23 by Michael Wang

https://v.qq.com/x/page/t30722rfj17.html

3. SSD 测试技术讲座第三期:PCIe Gen3/4 NVMe SSD 热插拔和电压拉偏测试技术 2019.11.30 by Andy Norrie

https://v.qq.com/x/page/k30286z9iwz.html

4. SSD 测试技术讲座第四期: PCIe/NVMe Gen4/5 协议分析诊断测试技术分享 2019.12.7 by Michael Wang

https://v.qq.com/x/page/p3031ezscox.html

5. SSD 测试技术讲座第六期: UNH IOL NVMe/NVMe-oF 兼容性测试技术讲座 2020.4.11 by David Woolf

https://v.qq.com/x/page/i09488bygl7.html

1.4.4 日常技术培训，产品演示视频

1. 解决 NAND 相关的 SSD 问题 by Nicola Campanelli

https://mp.weixin.qq.com/s/s1Gr6Tl3M7TzYa2zoLgwLw

2. Quarch PAM (Power Analysis Module)操作培训视频 by Michael Wang

https://mp.weixin.qq.com/s/q8O4HX7jPE3ZIYKlCpFBEQ

3. SerialTek PCIe Gen5 协议分析仪演示 by Michael Wang

https://mp.weixin.qq.com/s/rOCZhrHQQZuMjDJzSWlUTg

4. SerialTek PCIe Gen4 协议分析仪开箱、连接、使用演示视频 by Michael Wang

https://mp.weixin.qq.com/s/QgJWHoZWWVSroVwtlg-E8Q

5. 没想到 PCIe Gen4 协议分析仪原来功能这么强大!

https://mp.weixin.qq.com/s/7n_8J0q5L0GqlWYJyb8qCg

1.4.5 技术文章，产品演示视频

1. PCIe Gen4 NVMe SSD 测试环境搭建和常用工具视频演示

https://mp.weixin.qq.com/s/J_b8o1OCRCt8FgOlLVc_XA

2. 业界主流 SSD 研发中心选择 PCIe Gen4/5/6 分析仪最关心哪几个点？

https://mp.weixin.qq.com/s/mbffr8r3D70mbMY0ZQSHjw

3. PCIe Gen4/5 协议故障注入，热插拔，电压拉偏和功耗测试视频演示

https://mp.weixin.qq.com/s/UvKdI79xmU31V2hh3AaTUg

4. PCIe Gen4 NVMe SSD 测试视频演示

https://mp.weixin.qq.com/s/YmM1gfJfU6Ugk-RzBKm-6A

5. 分析解决 M.2 NVMe SSD 低功耗问题的利器

https://mp.weixin.qq.com/s/Qlj3Z6-rToEUAys0Mb0yWw

6. PCIe Gen4 协议分析、诊断、测试常用工具白皮书介绍

https://mp.weixin.qq.com/s/u4Xai2_kkNk1r1f5Ywe1Uw

7. 业内能用的 PCIe Gen5 协议分析仪 VS 最经济的 Gen5 协议分析卡

https://mp.weixin.qq.com/s/iWIv1YCmBXia_-VuDj5fBg

8. PCIe 和 NVMe SSD 初始化过程简介

https://mp.weixin.qq.com/s/vkLwlVAODNCHGWj3l2PUpA

9. PCIe Gen 5 业内常用测试工具解读

https://mp.weixin.qq.com/s/qfwSRvUPSvu3a8rQ-cReHg

10. PCIe M.2 SSD 低功耗 L1.2 全新分析方法以及 Gen5 测试新产品介绍

https://mp.weixin.qq.com/s/b-OwSOMXp1JD1LF592J_Ww

11. 业内第一款带免费 PCIe Gen 5 x16 协议抓包功能的 Host Card 可以试用了!

https://mp.weixin.qq.com/s/yFMiThcOTkozOKD4wKiOPg

12. PCIe Gen 4 NVMe SSD 测试环境搭建和常用工具视频演示

https://mp.weixin.qq.com/s/J_b8o1OCRCt8FgOlLVc_XA

13. 针对 PCIe Gen5 NVMe SSD 的研发测试工具来了！

https://mp.weixin.qq.com/s/majMU8RDScWKQdj3Tl7RiA

14. SERIALTEK 公司正式发布 PCIE GEN5 X16 协议分析仪和 WEB 应用程序

https://mp.weixin.qq.com/s/4WxYj-YxG75YLJHaGEiDHQ

15. 目前实测唯一可达 PCIe Gen5 x16 速度的 1 米 PCIe 插槽延长线！！！

https://mp.weixin.qq.com/s/K91s5yqyw-tmE8tSmtYnbw

16. 双端口（Dual Port) NVMe SSD 技术简介和演示

https://mp.weixin.qq.com/s/mHiygKNtbwJnwz-5qqmb4Q

17. PCIe Gen5 Switch 卡测试环境构建演示

https://mp.weixin.qq.com/s/ixKJxaDrTjcCKsB4dmBBWQ

1.5 联系 Saniffer 公司

我们欢迎业内公司参观访问 Saniffer 公司，请添加下面的微信提前联系，或者预约访问我们在张江高科技园区的办公室。

Saniffer 上海公司地址：

上海市浦东新区秋月路 26 号矽岸国际 1 号楼

021-50807071 / 13127856862, ***sales@saniffer.com***

2. PCIe/CXL Gen 4/5/6 协议分析

图 2-1 PCIe Gen6 x16 analyzer

图 2-2 PCIe Gen5 x16 analyzer

图 2-3 PCIe Gen4 analyzer

** *下面的介绍大部分以 SerialTek Kodiak 系列 PCIe 协议分析仪为例，限于篇幅关系，除非特别指出外，同样适用于 SerialTek PCIe Gen 5 x16/x8/x4 协议分析仪。*

图 2-4 Panda PCIe PCIe analyzer

*** SerialTek 最新推出的基于 Broadcom 内嵌 SerialTek 抓包分析逻辑的 Gen5 switch 芯片的“经济”协议分析解决方案，具体说明参见 2.8.2 章详细介绍*

参见下图的 PCIe 协议分析仪解码界面，PCIe 分析仪可以抓取链路上的所有的 ordered set, DLLP 和 TLP Packet 等，然后按照时间戳的顺序依次解析出 TS1/TS2, DLLP, TLP, NVMe 命令等，使研发、测试工程师在遇到问题的时候可以快速抓包、分析定位各种问题。图中可见 NVMe Read 和 Write 命令解码。

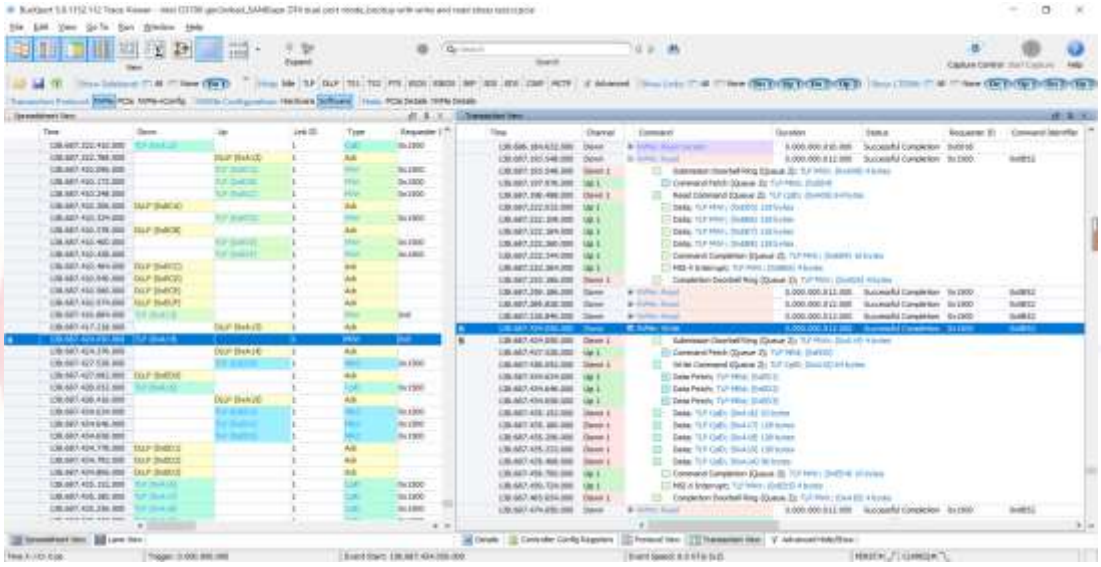


图 2-5 PCIe Gen4 analyzer 软件界面

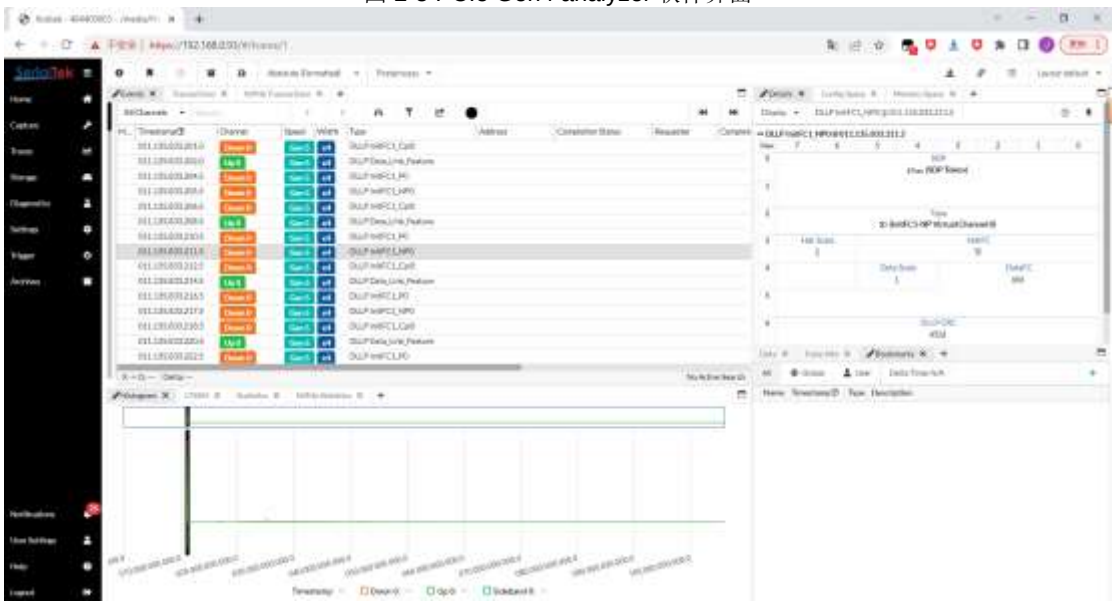


图 2-6 PCIe Gen5/6 analyzer Web 软件 Event View 界面

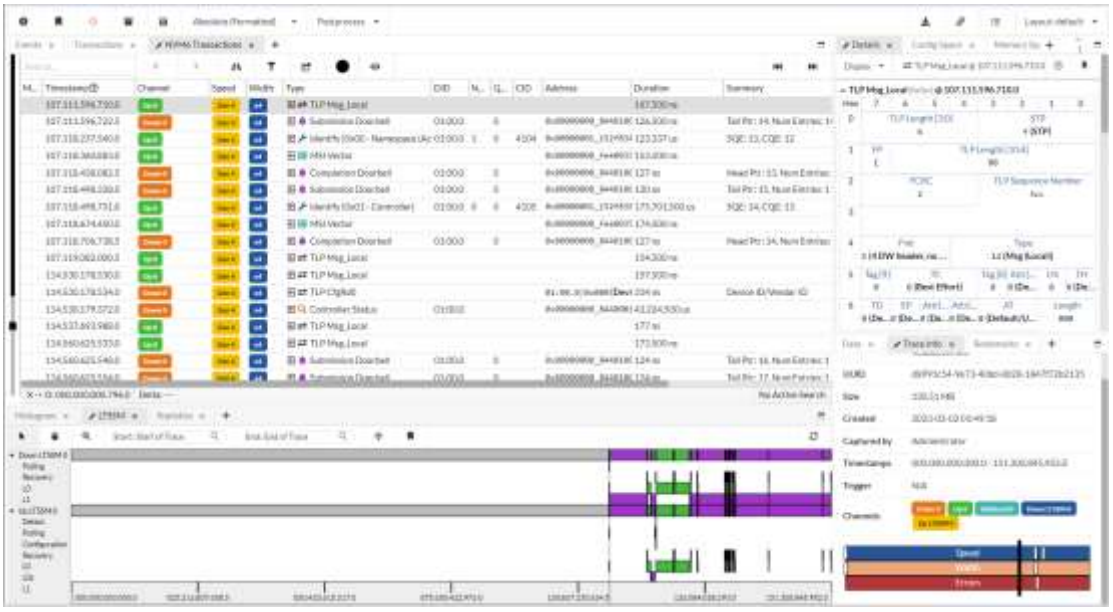


图 2-7 PCIe Gen5/6 analyzer Web 软件 NVMe Transaction View 界面

2.1 PCIe Gen 4/5/6 协议分析面临的技术挑战

在介绍 PCIe Gen4/5/6 协议分析面临的挑战之前，我们先来简单了解一下 PCIe 总线的起源和历史。

PCI Express，简称 PCI-E，官方简称 PCIe，是计算机总线的一个重要分支，它沿用既有的 PCI 编程概念及信号标准，并且构建了更加高速的串行通信系统标准。目前这一标准由 PCI-SIG 组织制定和维护。PCIe 初期主要应用于内部互连，但是后来也支持外连使用，例如用于连接 PCIe Switch 卡和 JBOF 盘柜，或者存储系统机头和盘柜之间的互联，这些连接通过各种标准接口的 PCIe cable 连接使用。由于 PCIe 是基于既有的 PCI 系统，所以只需修改物理层而无须修改软件就可将现有 PCI 系统转换为 PCIe。

PCIe 拥有更快的速率，所以几乎取代了以往所有的内部总线（包括 AGP 和 PCI）。除此之外，PCIe 设备能够支持热拔插以及热交换特性，目前支持的三种电压分别为+3.3V、3.3Vaux 以及+12V。同时，考虑到现在显卡功耗的日益增加，PCIe 而后在规范中改善了直接从插槽中取电的功率限制，x16 的最大提供功率一度达到了 75W。

PCIe 保证了兼容性，支持 PCI 的操作系统无需进行任何更改即可支持 PCIe 总线。这也给用户的升级带来方便。由此可见，PCIe 最大的意义在于它的通用性，不仅可以让它用于南桥和其他设备的连接，也可以延伸到芯片组间的连接，甚至也可以用于连接图形处理器，这样，整个 I/O 系统重新统一起来，将更进一步简化计算机系统，增加计算机的可移植性和模块化。

2.1.1 PCIe 协议发展的历史

在 2001 年的春季英特尔开发者论坛（IDF）上 Intel 公布取代 PCI 总线的第三代 I/O 技术，被称为“3GIO”。该总线的规范由 Intel 支持的 AWG（Arapahoe Work Group）负责制定。2002 年 4 月 17 日，AWG 正式宣布 3GIO 1.0 规范草稿制定完毕，移交 PCI 特殊兴趣组织（PCI-SIG）进行审核，2002 年 7 月 23 日经过审核后正式公布，改名为“PCI Express”，并根据开发蓝图在 2006 年正式推出 Spec 2.0（2.0 规范）。

2017 年 11 月 29 日 PCI SIG 在发布了 PCIe Gen4 规范，支持速率 16GT，时隔大概一年半，在 2019 年 5 月 28 日，PCI-SIG 官方又发布了 PCIe 5.0 的 1.0 版基础规范，规范主要定义了 PCIe5.0 的架构（architecture）、互联属性规范（interconnect attributes）、网络结构管理（fabric management）以及编程接口（programming interface）等内容。

PCIe 总线从 Gen4 开始对于信号质量的要求相较于 Gen3 有了很大的提高，PCIe Gen5 的推出使得信号质量的问题变得更加严重，这对于基于 PCIe Gen4/5/6 研发产品的公司提出了更多的挑战，包括底层示波器、误码仪，以及 PCIe Gen4/5 协议分析仪等方面，具体可以参考后续仪器章节关于协议分析仪的介绍。

2022 年 1 月 11 日，PCI-SIG 官方正式发布了 PCIe 6.0 规范，PCIe 6.0 采用了和之前全然不同的编码格式，即 PAM4 FEC 编码格式，这个和之前 PCIe Gen 1~5 采用 NRZ 编码格式有了很大的不同，同时也对于 Gen6 产品的实现带来相当大的挑战。

2022 年 6 月 22 日，发布和维护 PCIe 标准的联盟 PCI-SIG 宣布推出最新一代 PCIe 规范 PCIe 7.0 或 PCIe Gen 7 。该公告是在 2022 年 PCI-SIG 开发者大会上宣布的，该组织正在庆祝其成立 30 周年。

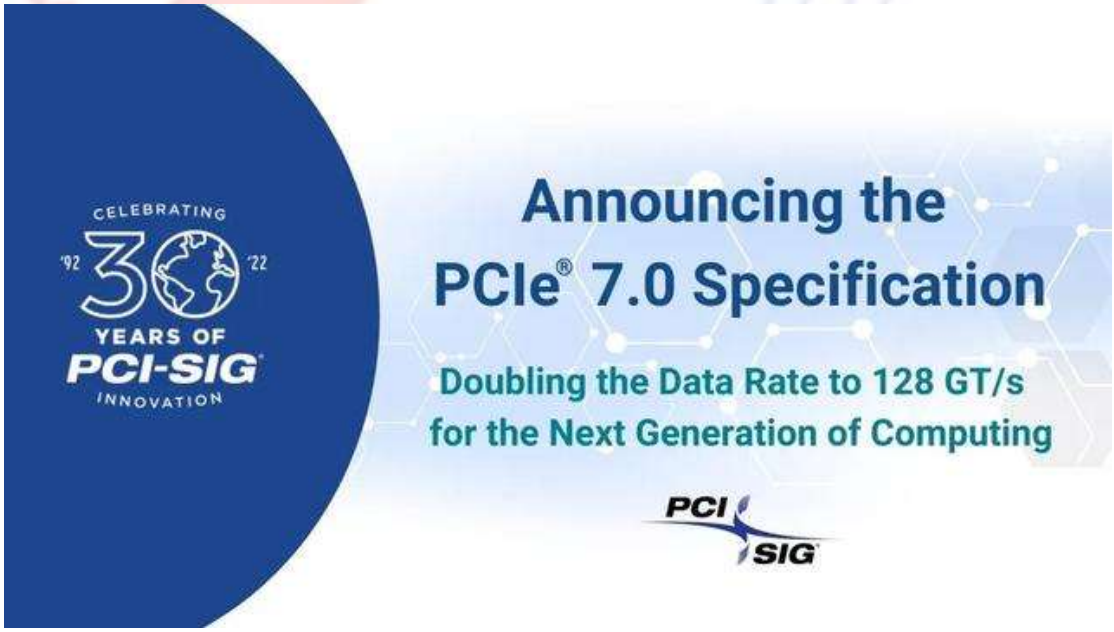


图 2-8

最新一代 PCIe 带宽翻了一番，在一条通道 (x1) 上单向实现 128GT / s 或 128Gbps 总吞吐量。综上所述，在 PCIe x16 插槽上，与独立显卡一样，双向总理论吞吐量为 512GB /

s。同时，通常与 x4 PCIe 插槽配对的 NVMe SSD 可提供高达 64GB / s 的单向速度。最终规格将于 2025 年发布。

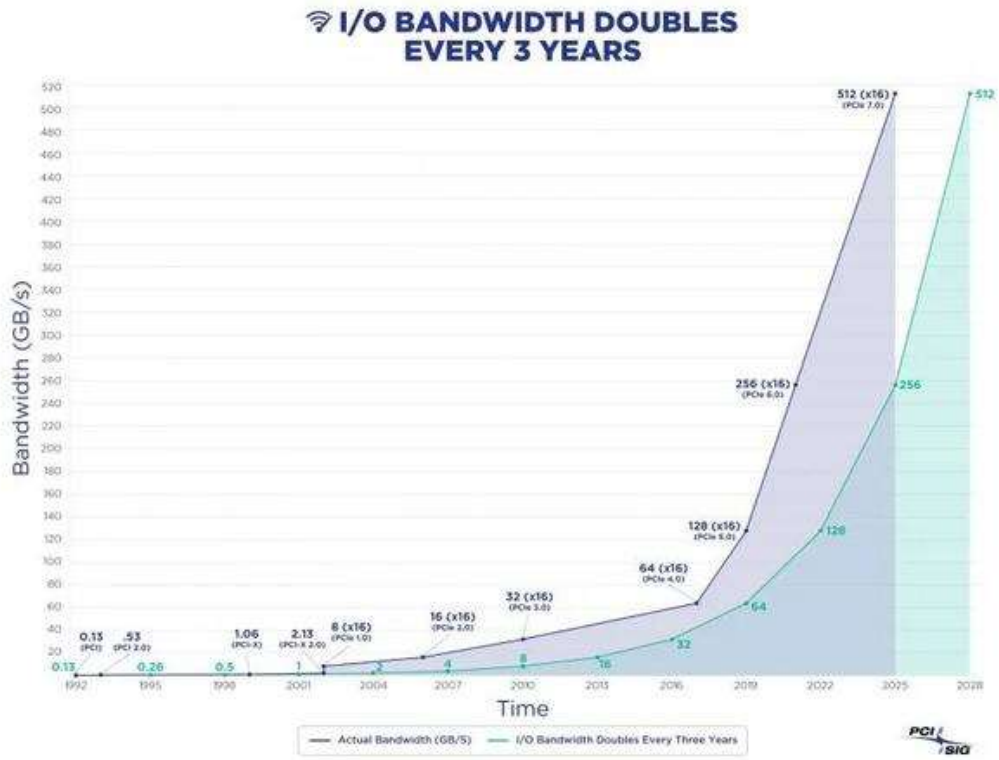


图 2-9

以下是新 PCIe 7.0 规范的亮点：

- **通过 x16 配置提供 128 GT / s 的原始比特率和高达 512 GB / s 的双向传输速率**
- **使用 PAM4（4 级脉冲幅度调制）信令**
- **聚焦渠道参数和覆盖范围**
- **继续交付低延迟和高可靠性的目标**
- **提高电源效率**
- **保持与所有前几代 PCIe 技术的向后兼容性**

图 2-10

规划其路线图的公司可以包括下一代 PCIe 技术，并保证它将满足对可靠、高速、低延迟 I / O 互连的需求。PCIe 7.0 技术将扩大 PCI-SIG 的路线图，包括数据密集型应用和市场，如 800 Gig 以太网、人工智能和机器学习（AI / ML）、高性能计算（HPC）、量子计算、超大规模数据中心和云。

PCI Express 总线性能

PCIe 版本	推出时间	Line 编码	原始传输率	带宽（单个方向）				
				×1	×2	×4	×8	×16
1.0	2003	8b/10b	2.5 GT/s	250 MB/s	0.50 GB/s	1.0 GB/s	2.0 GB/s	4.0 GB/s
2.0	2007	8b/10b	5.0 GT/s	500 MB/s	1.0 GB/s	2.0 GB/s	4.0 GB/s	8.0 GB/s
3.0	2010	128b/130b	8.0 GT/s	984.6 MB/s	1.97 GB/s	3.94 GB/s	7.88 GB/s	15.8 GB/s
4.0	2017	128b/130b	16.0 GT/s	1969 MB/s	3.94 GB/s	7.88 GB/s	15.75 GB/s	31.5 GB/s
5.0	2019	NRZ 128b/130b	32.0 GT/s	3938 MB/s	7.88 GB/s	15.75 GB/s	31.51 GB/s	63.0 GB/s
6.0	2022	PAM4 & FEC 128b/130b	64.0 GT/s	7877 MB/s	15.75 GB/s	31.51 GB/s	63.02 GB/s	126.03 GB/s

单向带宽计算公式：PCI-E 串行总线带宽（MB/s） = 串行总线时钟频率（MHz） 编码方式* 串行总线位宽（bit/8 = B），例：双工 PCIe Gen4 x4 NVMe SSD 理论单向带宽，其带宽 = 16*128/130*4/8 =7.88 GB/s。*

综上，PCIe 4/5/6 协议分析相较于 PCIe 3.0 有很多技术挑战，最主要的有两点：信号问题和解码速度。

2.1.2 PCIe Gen6 和 CXL3.0 新增的特性

2.1.2.1 What Disruptive Changes to Expect from PCI Express Gen 6.0

28 Apr 2021 • 3 minute read

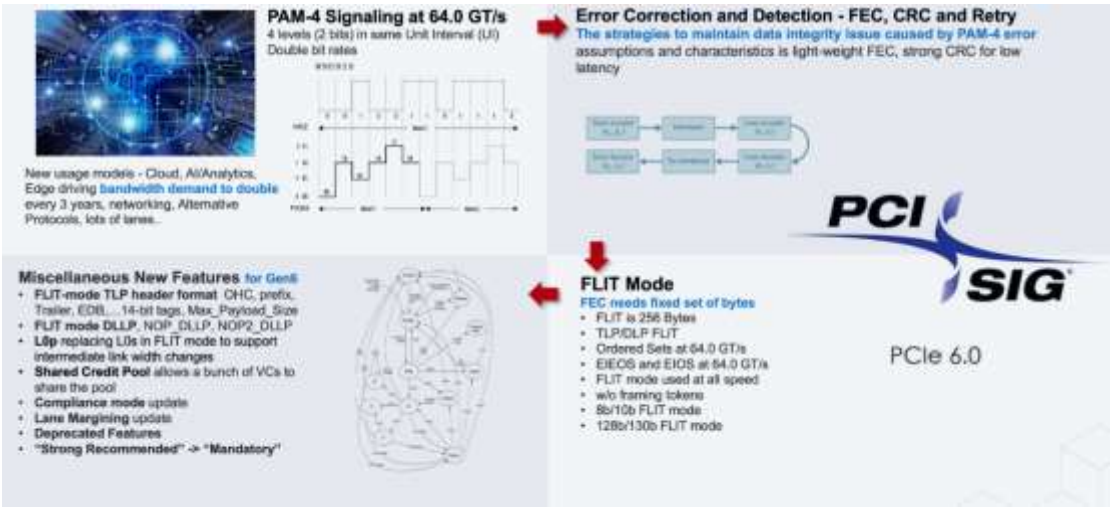
PCIe (Peripheral Component Interconnect Express) has long been the backbone of complex systems, and provides a high-bandwidth, high-performance link for interconnecting devices imposed by cloud-based computing power, storage capacity network bandwidth, artificial intelligence automotive platforms. PCIe 6.0, in turn, is the most important and most disruptive update to the PCIe standard since PCIe 3.0 almost a decade ago.

The PCIe 6.0 introduces a new physical layer change, with PAM4 (Pulse Amplitude Modulation with 4 levels) signaling to replace NRZ (Non-Return to Zero), a key ingredient in the generational bandwidth doubling effort. Rather than traditional 0/1 high/low signaling, PAM4 uses 4 signal levels so that a signal can encode for four possible two-bit patterns: 00, 01, 10 and 11 per Unit Interval (UI) without increasing the transmission frequency. It doubles the data rate to 64 GT/s, adequate bandwidth over PCIe 5.0 to 256 GB/s of throughput and retains the same maximum x16 lanes.

The additional signal states of PAM4 results more fragile signal than an NRZ. The PCIe 6.0 counters this by incorporate a combination of light-weight correction through FEC (Forward Error Correction) and strong detection and retry through CRC (Cyclic Redundancy Check) scheme to maintain the data integrity and low latency being a load-store protocol. Through these two methods, the result is a correlation between errors on a Lane and across Lanes: FEC operates on the principle of sending redundant data that can be deployed to correct some errors at the Receiver. CRC is an error detection code used to authenticate packet transmission between the sender and the receiving end, resulting in the eventual correction through Link Layer Retry.

The error correction needs to operate on fixed-sized packets, hence the adoption of FLIT (Flow Control Unit) for PCIe 6.0. The TLP header changes TLP Header Base followed by 0 to 7 additional DW of OHC (Orthogonal Header Content), end-to-end TLP prefixes integrated into the header, along with other changes to improve the robustness and extensibility for TLP content and structure. DLLPs are fixed based upon FLIT mode. Since error correction happens on FLIT, we have the CRC check as well as Retry at the FLIT level. Once the Link operates in FLIT mode, any speed change to lower data rates will also have to use the same FLIT mode. Thus, once enabled, FLIT mode is followed in the Link, irrespective of the speed or 8b/10b and 128/130 block encoding.

Beyond PAM4, FEC, and FLIT, PCIe 6.0 improved various improved power consumption with L0p (L0 Partial) that replaces L0s for FLIT Mode. The new state L0p is symmetric and maintains at least one active Lane that supports scalable power consumption and ensures uninterrupted traffic flow, even during width transitions. The Link always trains in the highest possible width and subsequently can modulate its width depending on the bandwidth need in FLIT mode. The Shared Credit Pooling allows multiple VCs to share the pool and reduce the cost. It is optional for a receiver to implement but mandatory for a PCIe 6.0 device to support as a transmitter.



One important that has not change is that , PCIe 6.0 retains backward compatibility with all five older generations of PCIe, which could mean the PCIe slot on

motherboards do not look any different. The PCIe protocol's intricate nature requires verifying interoperability and backward compatibility in multi-layer functionalities, previous spec generations, diverse topologies, and configurations. On top of that, the new PCIe 6.0 introduces disruptive changes at all layers, which requires require a comprehensive functional verification and testing approach.

Furthermore, in the last few years, their efforts have taken on an increased level of importance, as other major interconnect standards are building off PCIe. CCIX (Cache Coherent Interconnect for Accelerators), Intel's CXL (Compute Express Link), and other interfaces have all extended PCIe and benefit from PCIe improvements. So PCIe speed boosts serve as the core of building ever-faster and more interconnected systems, which requires require advanced system-level verification.

To be sure, taking advantage of all these new features and enhancements with the latest generation of PCIe while maintaining backward compatibility requires advanced system-level verification to ensure at a first order that things did not break and then to verify the intricate new changes.

2.1.2.2 Insights Into the Evolutions and Optimizations of PCIe 6.0

16 Nov 2023 • 4 minute read

The PCIe protocol (Peripheral Component Interconnect Express) had its first generation in 2003, being a huge breakthrough in the industry by allowing up to 2.5 GT/s per lane in a serial computer expansion bus. The protocol has since evolved many times, always doubling its transfer rate compared to the previous generation and bringing new features and optimizations whenever needed.

The latest release was announced in 2022, in which PCIe 6.0 was introduced with up to 64.0 GT/s speed per lane. As was announced at the PCI-SIG Developers

Conference in San Jose, 2023, PCIe 6.0 not only again doubled the speed but also prepared the grounds for many generations to come. The changes were made considering many necessary optimizations to the existing rules, considering the industry usage and experience of 20 years. New concepts and technologies were introduced, such as 1b/1b encoding, PAM4 modulation, and Flit Mode operation.

For more information on Flit Mode, see Unraveling PCIe 6.0 Flit Mode Challenges.

2.1.2.2.1 Understanding PCIe 6.0 Optimizations

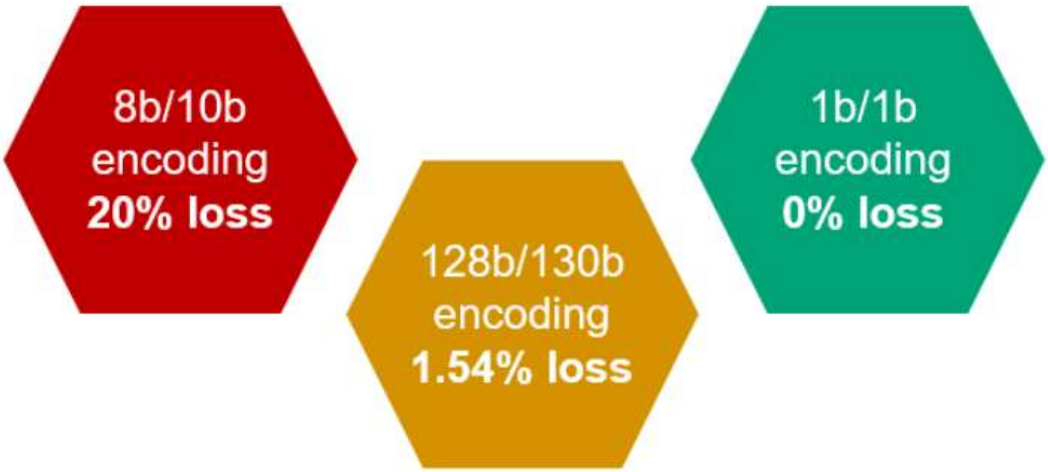
The changes in all features in PCIe 6.0 specification were done considering that they needed to be optimized to keep up with the higher throughput rate. Hence, the following guidelines were followed:

- **Reduce loss: By avoiding unnecessary encoding;**
- **Make assumptions: Based on established patterns;**
- **Avoid transmitting unnecessary information: That could be inferred by the other side;**
- **Avoid reconfiguration: If it was already configured before.**

2.1.2.2.2 1b/1b Encoding and Loss Reduction

The new encoding introduced in PCIe 6.0 is the biggest example of loss reduction by avoiding unnecessary encoding. Previously, instances used 128b/130b encoding when operating at 8.0 GT/s or higher speed. This means that every 128 bits of data required 2 extra bits in order to be correctly decoded by the other side. This caused an inefficiency in the serial link, in which 1.54% of the bandwidth was lost in the bit level simply due to encoding.

1b/1b addresses this problem by guaranteeing that every bit transmitted can be used as actual information by the other side. This is done by implementing internal counters in each side of what kind of data to expect. As long as the designs are correctly verified to respect those counters, it is guaranteed that they will be able to communicate without needing to send any extra unnecessary information in the link.

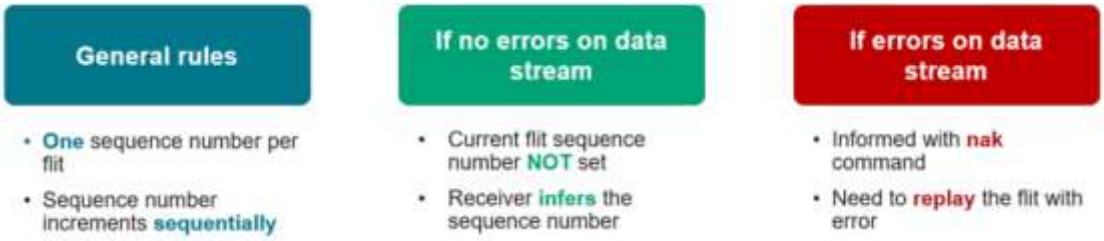


2.1.2.2.3 Flit Sequence Number and Optimizing Transmission of Information

The flit sequence number is a new concept introduced in PCIe 6.0, which was added together with the flit mode of operation. It replaces the old sequence number present in the Transaction Layer Packet (TLP), together with their acknowledgments or replay mechanisms.

Previously, sequence numbers were always attached to every TLP transmitted. Although it added robustness in the link, it turned out to be a waste of resources, considering that TLP had sequential sequence numbers. Therefore, knowing the sequence number of one TLP implied knowing the number of the TLP next to it, and so on.

The flit sequence number protocol optimized this by implementing the implicit sequence numbers, in which the sequence number is inferred by the other side. Not only that, but also the sequence number is in the flit level, which can accommodate many TLPs at once. Therefore, the space previously used for always transmitted sequence number information can be used to increase the bandwidth with useful information.



2.1.2.2.4 L0p and Optimizing Unnecessary Reconfiguration

Previously, the procedure to change the link width dynamically after link-up was costly for the devices since it required going through all the Configuration states of Link Training and Status State Machine (LTSSM). It meant reconfiguring all the details of the lanes, even though the only variable that required change was the link width being used.

This was enhanced in PCIe 6.0 with the introduction of L0 partial (L0p) feature, only present in Flit Mode. When the L0p sequence is performed, the link width can be changed during active data transfer, with no need to bring the link down. It means that performing power savings by changing the link width is much more effective and also, that devices can easily keep a smaller width in case they are having thermal throttling issues.

To know more about L0p, check Unraveling New Introduced PCIe 6.0 L0p.

In summary, PCIe 6.0 brought many changes, which were all optimizations to guarantee that all layers of PCIe protocol can keep up with the higher transfer rates. PCIe 7.0, which is currently in progress, continues the PCIe support and optimizations on top of all these changes. Therefore, it is very important to verify that devices follow the functional behavior of all those features to ensure they can take benefit from the advantages offered by the new protocol generation and also the versions that are yet to be announced.

2.1.2.3 Unraveling New Introduced PCIe 6.0 L0p

17 Oct 2022 • 4 minute read

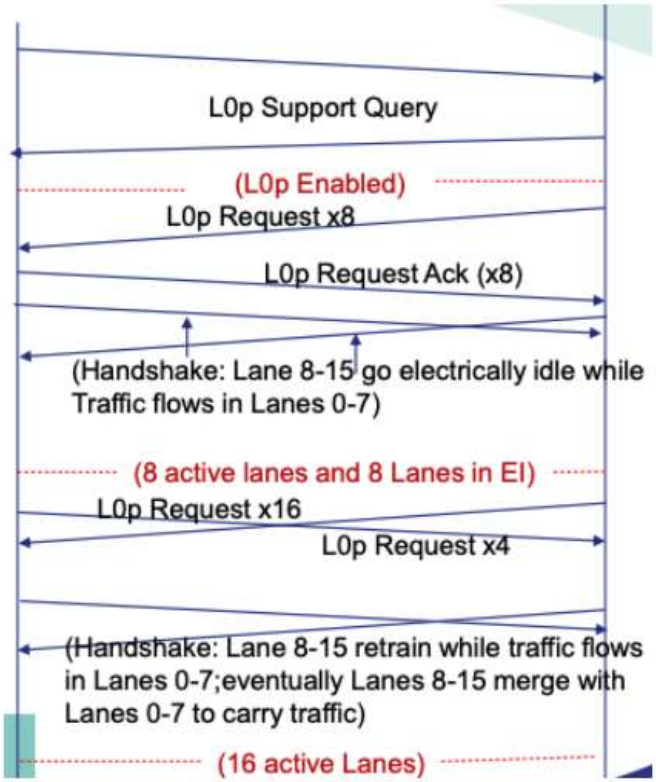
The PCIe 6.0 Specification released in 2021 doubles the performance to 64GT/s transfer rate with PAM4 (Pulse Amplitude Modulation with 4 levels) modulation and uses FLIT (Flow Control Unit) as the unit of communication for efficiency. In 'What Disruptive Changes to Expect from PCI Express Gen 6.0? we covered what significant features PCIe 6.0 evolved to embrace.

Amongst many new features and changes in PCIe 6.0, we will talk about one major significant new feature: L0p. The following mainly touches upon challenges and corresponding solutions based on our design and verification experiences. For more relevant PCIe 6.0 verification challenges, see Unraveling PCIe 6.0 FLIT Mode Challenges and Unraveling PCIe 6.0 Training Sequences Update and Verification Challenges.

L0p in PCIe 6.0

With the increase of demand for power consumption scaling with bandwidth usage without impacting traffic flow, the new L0p state is introduced in PCIe 6.0. Meanwhile, L0s is not supported in FLIT mode (L0s is less robust and less effective and does not support retimes, etc.). L0p is optional for link width resizing and is used in FLIT mode only.

The existing dynamic link width change can change the link, but it will lose several microseconds to do the state transition. L0p is symmetric in terms of the same width in both directions. It maintains at least one active lane during the width change to ensure uninterrupted traffic flow.



Spec introduced the negotiation details about how to enable L0p from Configuration.Complete state along with FLIT mode with TS2 exchange. Once both sides support it, how to request down or up link size, how to ack or reject, and the detailed handshake mechanism are explained in Chapter 4.2.6.7. Moreover, the new mechanism uses a new Link Management DLLP to carry the expected value.

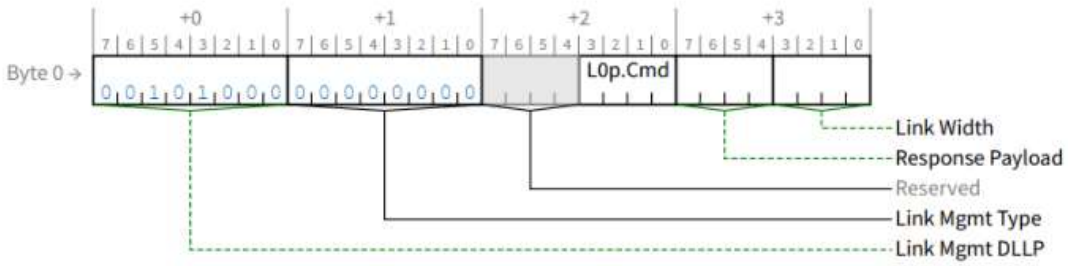


Figure 6-18 Link Management DLLP

Below are the verification challenges for L0p mode:

- Spec introduced the negotiation details about how to enable L0p from Configuration.Complete state along with FLIT mode with TS2 exchange. It is used with FLIT mode only, so we also need to consider backward compatibility to all lower speeds.

- According to the Spec, L0p width transition can be initiated by either side, and the link partner needs to either ACK or NAK in 4 us in 8b/10b mode and 1us in the other modes. The negotiation is done by using Link Management DLLP with new DLLP fields defined, such as link management type, L0p command (Request, ACK, NAK), L0p priority, link width, and response payload. All the combinations and rules need to be covered.

- During the width change, Spec defined the sequence of OS for downsizing and upsizing. For the downsizing, the lanes turned off will send an EIOSQ on the next SKP interval and then go to electrical idle. The active lanes remain to send SKP and FLITs. For the upsizing, while the data stream continues the active lanes, the lanes to be activated will follow the EIEOS rule if the data rate is Gen2 or above, followed by TS1/TS2. After the TS2 exchange, the ports send SDS followed by Control SKP. Then, eventually, data stream on all lanes. The challenge is not only implementing the logic to handle up/down-sizing but also handling different scenarios on different lanes, besides implementing checkers on Pre-lane based scenarios.

- Like link dynamic up/down configuration in previous versions, we need to verify all the valid up/down-sizing combinations in L0p or invalid requests through error injection test cases.

In summary, PCIe 6.0 is a complex protocol with many verification challenges. You must understand many new Spec changes and think about the robust verification plan for the new features and backward compatible tests impacted by new features.

2.1.2.4 Unraveling PCIe 6.0 Training Sequences Update and Verification Challenges

14 Oct 2022 • 4 minute read

The PCIe 6.0 Specification released in 2021 doubles the performance to 64GT/s transfer rate with PAM4 (Pulse Amplitude Modulation with 4 levels) modulation and uses FLIT (Flow Control Unit) as the unit of communication for efficiency. In What Disruptive Changes to Expect from PCI Express Gen 6.0? we covered what significant features PCIe 6.0 evolved to embrace.

Amongst many new features and changes in PCIe 6.0, we will talk about one major significant new feature: Training Sequences. The following mainly touches upon challenges and corresponding solutions based on our design and verification experiences. For more relevant PCIe 6.0 verification challenges, see Unraveling PCIe 6.0 FLIT Mode Challenges and Unraveling New Introduced PCIe 6.0 L0p.

Updates on Training Sequences in PCIe 6.0

TS1 and TS2 were used in every previous version but had their symbols positions redefined in 64GT/s. The symbol positions are entirely different from previous generations, and some symbols like link number and lane number maintain duplicate meanings, which is based on different LTSSM states.

Symbol Numbers	Description
0, 8	**TS1/TS2 identifier** - Unscrambled
1, 9	**Link Number** in Configuration or Hot Reset - Scrambled Equalization Byte 0 in Recovery and Loopback for TS1 - Scrambled Equalization Byte 0 in Recovery for TS2 – Scrambled RSVD in other states
2, 10	**Lane Number** in Configuration or Hot Reset - Scrambled **Equalization Byte 1** in Recovery – Scrambled RSVD in other states
3, 11	**Equalization Byte 2** in Recovery for TS1, Reserved for TS2 – Scrambled RSVD in other states
4, 12	**Equalization Byte 3** in Recovery for TS1, Reserved for TS2 – Scrambled RSVD in other states
5, 13	**Date Rate Identifier** - Scrambled
6, 14	**Training Control** Used in Recovery.RcvrCfg to Disable , Hot Reset , or Loopback . Reserved for TS2 Ordered Sets in Configuration - Scrambled
7, 15	If DC Balance needs adjustment at the start of the TS1 or TS2 : **DC Balance Symbol** - Unscrambled else: Byte level even parity over Symbols 0-6 (or 8-14) - Scrambled

In addition, an essential concept of valid halves was introduced for TS1/TS2 in 1b/1b. is the 16 bytes of each TS are composed of two halves of 8 bytes. For the receiving rules, if any half is valid, we consider the TS valid. Besides that, the scrambling rule for TS1/2 in 64GT/s speed slightly differs from other speeds.

When verifying TS1/TS2, we found several parts that might be interesting or challenging:

- With the new concept of halves, both halves are to be validated with mirrored values from the transmitter side, but for the receiver side, a valid TS can be considered when either half is valid.

- Some bytes like link number and lane number maintain dual meanings, for example, in Recovery.RcvrCfg, symbol 0 can be used as link number or Equalization byte0 (6:3, transmitter preset in recovery) for TS2 based on the certain scenario (normal or EQ TS2). Spec has ambiguous about it which our designer must have our interpretations. It will be resolved in future errata

gradually, but for now, we must be careful and may need to validate existing LTSSM state transition test cases in 1b/1b.

- Once we have implemented the new TS logic, we highly recommend checking TS1/TS2 in non-64GT/s speed under FLIT mode. Although there is no direct relationship between the old TS format and FLIT mode, the new logic may have bugs and be accidentally triggered in other speeds.

- It is difficult to predict if Symbol 7 and 15 reflects 'DC-balance' or 'Parity' because it used 2 different symbols in the previous format.

TS0 is the new TS used for Equalization in 64GT/s. It is used to communicate equalization information in specific symbols, which is just like TS1 does in previous versions. We need that because we consider that there may be many bit errors on the first entry to 64GT/s, so it has the same halves concept as TS1/TS2 in 64GT/s and even bits of all symbols are identical to odd bits for implementing NRZ encoding.

Symbol Numbers	Description
0, 8	TS0 identifer – Unscrambled 33h
1, 9	Bit 3,1 – Equalization Control (EC) Bit 5 – Reset EIEOS Interval Count Bit7 – Use Preset
2, 10	Bit 7,5,3,1 - Phase 0,1: FS[3:0] Phase 2: |C-1| [3:0]
3, 11	Bit 3,1 - Phase 0,1: FS[5:4] Phase 2: |C+1| [1:0] Bit 7,5 - Phase 0,1: LF[1:0] Phase 2: |C+1| [3:2]
4, 12	Bit 1 - Phase 0,1: LF[2] Phase 2: |C+1| [4] Bit 7,5,3 - Phase 0,1: LF[5:3] Phase 2: |C-2| [2:0]
5, 13	Bit 7,5,3,1 - Phase 0,1: Preset [3:0] Phase 2: if Use Preset is 1b: Preset[3:0] Else: |C0| [3:0]
6, 14	Bit 3,1 - Phase 2: |C0| [5:4] Bit 5 - Retimer Equalization Extend
7, 15	If DC Balance adjustment is needed at start of TS0: 00h,02h,22h (Unscrambled) Else: Bit 1,3,5,7 Half Scrambled Byte Level Even Parity

The first challenge here is that we must ensure that once we move to the Equalization state for 64GT/s, the model can transmit/receive the TS0 and decode it.

So, implement necessary checkers to check all new symbols and make sure halve concept is verified, which we have discussed in TS1/TS2.

The second challenge is that TS0 will be used in phase1/2 for Downstream Port and all phases for the upstream port.

There is a requirement that TS0 needs to be sent in the beginning of the phase2 for DP and the beginning of the phase3 for UP, then followed by TS1. We need to verify the TS0 behavior regarding phase transitions, timing requirement and the corner cases in Recovery.Equalization state.

Current Data Rate / Port	Phase 0 / Phase 1	Phase 2	Phase 3
8.0 GT/s, 16.0 GT/s, or 32.0 GT/s; Upstream/Downstream Lanes	TS1	TS1	TS1
64.0 GT/s Downstream Lanes	TS0	TS0 followed by TS1	TS1
64.0 GT/s Upstream Lanes	TS0	TS0	TS0 followed by TS1

The last one is that TS0 cannot be used in the non-Equalization state or phase3 of DP.

To make sure TS0 will not be sent or received at the wrong time and that DUT can handle the incorrect TS0 receiver, some EI test cases can be considered. Also, implementing a monitor to track all phase transitions and LTSSM state transitions and reporting errors if there is any protocol violation is highly recommended.

In summary, PCIe 6.0 is a complex protocol with many verification challenges. You must understand many new Spec changes and think about the robust verification plan for the new features and backward compatible tests impacted by new features.

2.1.2.5 Unraveling PCIe 6.0 FLIT Mode Challenges

12 Oct 2022 • 6 minute read

The PCIe 6.0 Specification released in 2021 doubles the performance to 64GT/s transfer rate with PAM4 (Pulse Amplitude Modulation with 4 levels) modulation and uses FLIT (Flow Control Unit) as the unit of communication for efficiency. In ‘What Disruptive

Changes to Expect from PCI Express Gen 6.0? we covered what significant features PCIe 6.0 evolved to embrace.

Amongst many new features and changes in PCIe 6.0, we will talk about one major significant new feature: FLIT. The following mainly touches upon challenges and corresponding solutions based on our design and verification experiences. For more relevant PCIe 6.0 verification challenges, see Unraveling PCIe 6.0 Training Sequences Update and Verification Challenges and Unravelling New Introduced PCIe 6.0 L0p.

What is FLIT in PCIe 6.0?

The transactions in previous versions had a variable length of size, known as TLPs. They may have a fixed header size but had a different length of data payload. No matter how long the TLP is, it is protected by 32-bit CRC. In PCIe 6.0, the additional signal states of PAM4 result in a more fragile signal than an NRZ. The new modulation requires FEC to compensate for PAM4's higher bit-error rate, and the error correction needs to operate on fixed-sized packets, hence the adoption of FLIT (Flow Control Unit) for PCIe 6.0.

FLIT has a fixed 256-byte length of size which consists of 236-byte of TLP, 6-byte of DLP, 8-byte of CRC and 6-byte of FEC. It removes Sync Header in 1b/1b encoding, framing token, etc. FLIT also has a similar sequence number concept, in which the first 2 bytes of the DLP carry information dedicated to FLIT level sequence number, Ack/Nak, Retry mechanism, etc.

FEC (Forwarded Error Correction) is designed for latency and complexity increases exponentially with the number of symbols corrected. 6 bytes of FEC are responsible for 3 interleave groups and each group has 2 FEC bytes. This is to prevent burst errors if it is smaller than 3 bytes.

The first challenge is regarding new FLIT format and encoding changes.

The FLIT-enabled mechanism and negotiation are happening by the beginning of the link training, Polling and Configuration, using the FLIT Mode Supported bit in the ‘Data Rate Identifier’ field (Symbol 4, Bit 0) in the TS1. Once it is negotiated, it applies to all data rates, implying FLIT is also supported 8b/10b and 128b/130b (Hybrid mode).

In FLIT mode, we are using a completely new TLP Header format. Previous TLP Headers had many limitations, like no room for increasing tag size. PCI-SIG redesigned header to suites suited FLIT mode. The challenge will be to test all the new combinations.

TLP Header is composed of a 3 to 7 DW TLP header Base, followed by 0 to 7 additional DWs of OHC (Orthogonal Header Content).

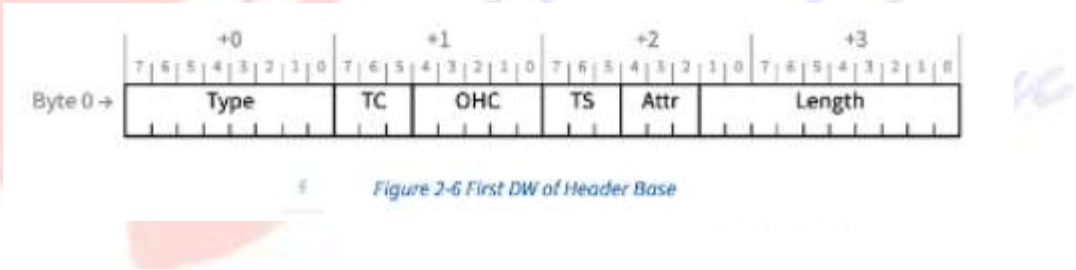


Figure 2-6 First DW of Header Base

The new type of field has a fully decoded 8b Packet Type field. This means that all the 256 Type values are defined or reserved for a certain group to permit proper framing and forwarding.

Also, there are new completion rules designed for FLIT mode. The completion for non-posed TLPs also has major updates including a 14-bit tag, error report using OHC-A5, etc.

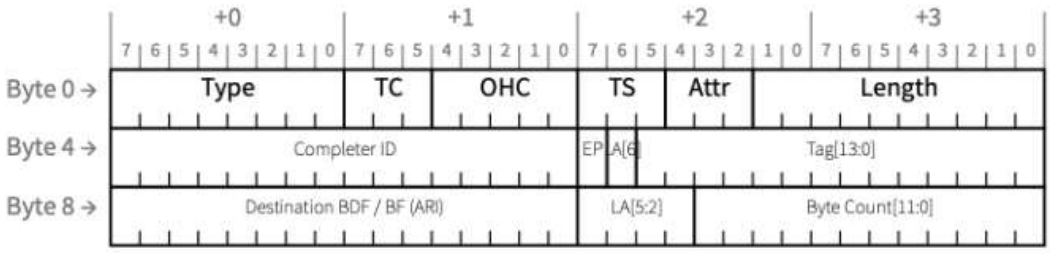


Figure 2-76 Completion Header Base Format - Flit Mode

As we know, 236 bytes can accommodate 1 TLP or many TLPs, or if you have a long TLP, it can be split into multiple FLITs. Also, between each TLP, it may have a NOP TLP if no further TLPs are scheduled. There are rules like no more than 4 TLPs (non-NOP) per FC/VC in the 32 DW boundary of the FLIT, which are also new to the PCIe 6.0 spec.

The DLP is a 6 bytes sequence, and the first 2-bytes are dedicated to FLIT level Ack/Nak, Retry, etc. so there is a new format for it. For instance, the “FLIT usage” decides between IDLE FLIT, NOP FLIT, or payload FLIT; “Prior FLIT” had non-NOP or NOP which is used to avoid retry error; and “replay command” along with “sequence number” decide Ack/Nak/Retry.

Moreover, multiple FLIT transmission and exchange rules are newly defined in the FLIT sequence number and retry mechanism. For example, the CONSECTIVE_TX_EXPLICIT_SEQ_NUM_FLITS and CONSECUTIVE_TX_NAK_FLITS counters and it is relevant rules.

To test all major encoding and format changes, we must ensure all new packet fields are covered. The generic solution for the above new features will be to define and exercise good coverage to test the new feature well. A good coverage model assists to test new features well.

Other than the FLIT format, another big challenge for FLIT is the new sequence number and the Retry mechanism.

One of the most difficult verification challenges is the IMPLICIT_RX_FLIT_SEQ_NUM rules. This counter is essential for the replay mechanism. As the implicit FLIT sequence number is not carried in the FLIT, it all depends on internal logic to handle it. The internal logic/counter needs to handle multiple scenarios to make sure the IMPLICIT_EX_FLIT_SEQ_NUM calculation is correct.

It is essential to make sure that the TX retry buffer is correct as it needs to be stored in all FLITs before receiving Ack or Nak. As multiple TLPs can be in one FLIT or one large TLP can be split into various FLITs, we need to guarantee that the retried FLIT should not skip or add an extra TLP to the original FLIT. It is essential for Posted TLPs as there is no completion for it. Lost TLPs will cause uncorrectable errors.

The new Standard Nak/selective Nak can let the transmitter replay a certain FLIT or multiple FLITs. The relevant rules are impacting both TX and RX retry buffer buffers. Also, sending Standard or selective NAK is implementation-specific, so sometimes it hard to predict and make checking if there are protocol violations.

The FEC algorithm is a new feature that we need to ensure there are no bugs in the calculation on both TX and RX sides.

Based on the above pointers that we have discussed, below are the recommended solutions we have tried to verify our design:

- In addition to following Spec to implement the IMPLICIT_RX_FLIT_SEQ_NUM counter correctly, we found it is necessary and valuable to make implicit FLIT sequence number available in the status register or make it visible in the debug log for easier comparing and debugging.

- Make sure your monitor can store all FLITs and can compare saved FLIT with retired FLIT on every symbol and report the error accordingly.

- You must make both RTL and monitor have a big enough retry buffer in case standard NAK is received, and multiple FLITs will be resent. Also, implement a checker to predict standard and selective NAK transmit conditions based on Spec. Third, to examine if the replayed FLIT's sequence number matches with the selective NAK received.

- To make sure FEC algorithm logic is implementation is correct, use an Error Injection test case with randomized symbol location needed.

In summary, PCIe 6.0 is a complex protocol with many verification challenges. You must understand many new Spec changes and think about the robust verification plan for the new features and backward compatible tests impacted by new features.

2.1.2.6 CXL 3.0 Scales the Future Data Center

17 Oct 2022 • 4 minute read

CXL is emerging as the industry focal point for coherent I/O with Open CAPI and Gen-Z transfer specification and assets to CXL Consortium. In August, the next full version of the CXL 3.0 standard was announced. With the continued proliferation of cloud computing, AI and analytics, increasing need for system-level optimization among high performance accelerators, system memory, smart NICs and leading edge networking. The new standard version introduced memory-centric fabric architectures and expanded capabilities for improving scale and optimizing resource utilization, which could change how some of the world's largest data centers and fastest supercomputers are built.

Double the bandwidth and zero added latency

CXL 3.0 builds on top of PCIe 6.0 and inherits the full bandwidth improvements of PCIe 6.0 along with Pulse Amplitude Modulation 4-level (PAM4) and Forward Error Correction (FEC), doubling total bandwidth to 64 GT/s. Notably, CXL 3.0 takes one step

further in reducing latency, resulting in CXL 3.0 having the same latency as CXL 1.x and CXL 2.0.

CXL 3.0 bumps 68-byte FLIT in CXL 1.x/2.0 up to 256 bytes. The standard CXL 3.0 FLIT is very similar to the PCIe 6.0 FLIT layout, with a 2-byte FLIT header, to indicate the protocol stack CXL.io, CXL.cachemem. The larger 256-byte FLIT size is one of the critical communications changes with more bits in the header FLIT, which enables the complex topologies and fabrics in CXL 3.0 standard.

CXL 3.0 also offers a Latency-Optimized (LOpt) version of FLIT mode that breaks up the Cyclic Redundancy Check (CRC) into 128-byte half FLIT granular transfers to mitigate store-and-forward overheads in the physical layer. A 6-byte CRC independently protects each half, but the FEC is across the whole 256 bytes. It allows consuming 128 bytes with good CRC for much lower latency. Notably, it cannot go back and forth between standard and LOpt modes.

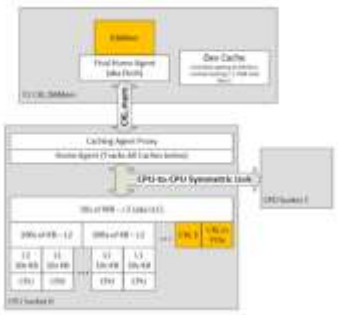
New protocols (UIO and BI) for enhanced coherency and peer-to-peer communication

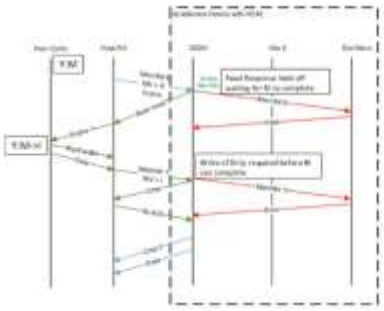
CXL 3.0 enables non-tree topologies and peer-to-peer communication (P2P) within a virtual hierarchy of devices that associates with devices that maintain a coherency domain. It disaggregates to allow the device-to-device connectivity directly rather than go through the host for every communication to overcome the host bottleneck in the tree

topology. These two major powerful enhancements are Unordered IO(UIO) and Back-invalidation (BI) protocols.

The producer-consumer ordering semantics are enforced at every entity, whether a switch, an endpoint, or a root port. These all enforce the same ordering semantics. UIO is a way to break what unordered IO does. It moves the producer-consumer enforcement to the source, avoids unnecessary traffics, and enables parallel paths to deliver better bandwidth and latency. And all of those are the factors to allow peer-to-peer.

BI protocol maps large memory in Type 2 devices to Host-managed Device Memory (HDM) with back invalidation. In CXL 1.x/2.0, the bias flip mechanism needs HDM to be tracked fully since the device could not back-snoop the host. Back invalidation with CXL 3.0 enables snoop filter implementation in large memory mapped to HDM.



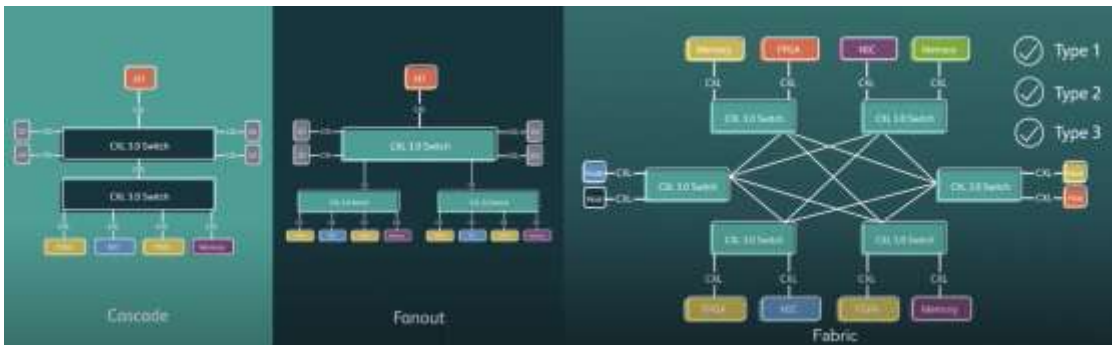


Multi-level Switching and Fabric Management

A significant addition to CXL 3.0 is multi-tiered switching and switch-based fabrics. CXL 2.0 allows for a single switching layer, with switches connecting vertically to upstream hosts and downstream devices but not supporting connections to other switches. The scale is limited to the available ports on a switch. With CXL 3.0, switch fabrics are enabled, where switches can connect to other switches, and each root port can connect to more than one device type, vastly increasing the scaling possibilities.

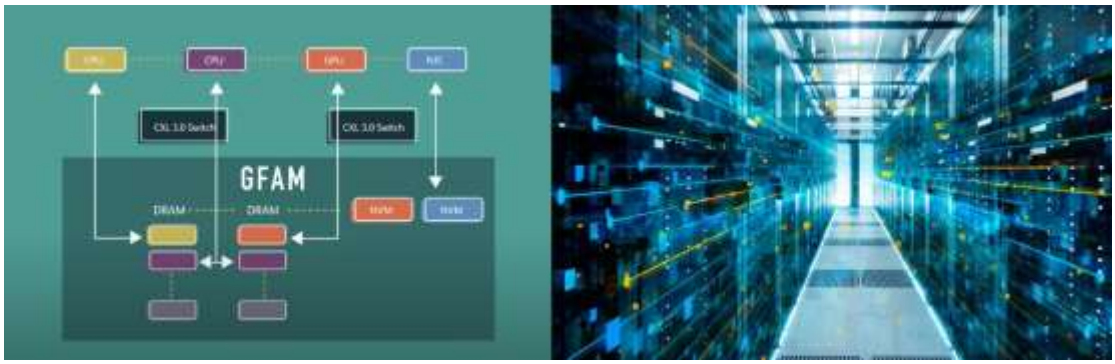
CXL fabric is available in virtually any configuration needed for a system. This composability combines heterogeneous compute elements, Type 1, 2 or 3, into the

overall system with no restrictions on architecture. These fabric-enabled systems can be dynamic, flexible, and intelligent, allowing system design around any application and right-sizing resources instead of over-provisioning. Multi-headed and fabric-attached devices enhance fabric management and composable disaggregated infrastructure.



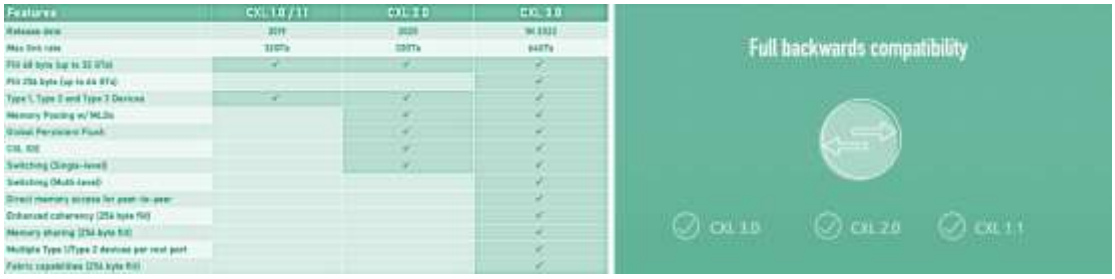
Improved Memory Sharing and Pooling

CXL 3.0 enables Global Fabric Attached Memory (GFAM) by disaggregating the memory from the processing unit and implementing a large shared memory pool memory can be of many different types, e.g. mixture of DRAM, and NAND flash, which can be accessed by multiple processors directly connected to GFAM or through a CXL switch. Even if the disaggregated memory is spread around the rack, the access time is still fast. Rack-scale memory fabric is a step on the journey to realizable memory-centric computing.



Fully Backwards Compatibility with CXL 1.x and CXL 2.0

Finally, CXL 3.0 ensures full backward compatibility with CXL 1.x and CXL 2.0, devices and hosts can downgrade as needed to match the rest of the hardware chain, albeit losing newer features and speeds in the process.



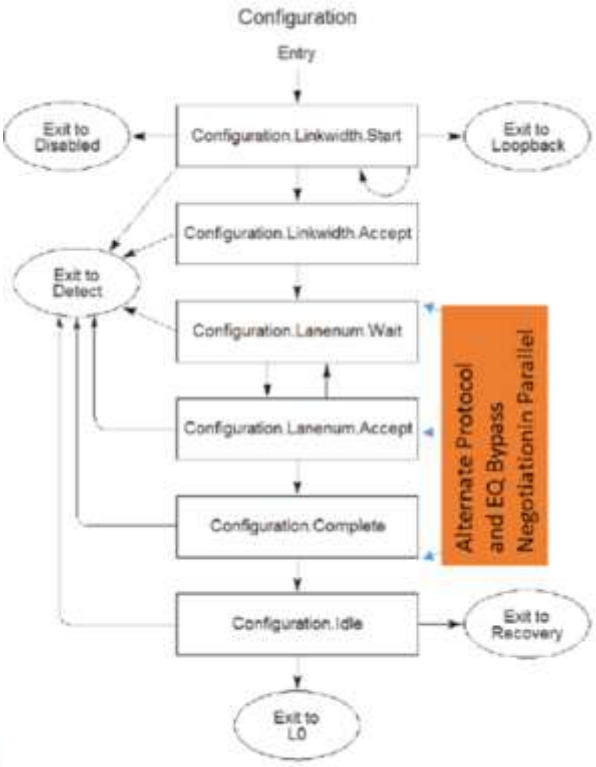
Summary

CXL 3.0 features facilitate the move to distributed, composable architectures and higher performance levels for AI/ML and other compute-intensive or memory-intensive workloads. The CXL 3.0 protocol can support up to 4,096 nodes, go beyond rack. Composable server architectures are when servers are broken apart into their various components and placed in groups where these resources can be dynamically assigned to workloads on the fly. CXL technology continues to enable game-changing innovations for the modern data center at scale.

2.1.2.7 Leveraging the PCIe for CXL Mode Link Up Using Alternate Protocol Negotiation Technique

19 Oct 2022 • 3 minute read

An Alternate Protocol negotiation (APN) can be understood as a non-PCIe protocol that makes use of the PCIe PHY layer. It may be chosen to run the PCIe protocol in addition to one or multiple alternate protocols in the alternate protocol mode. This is negotiated by the link partners during Configuration LTSSM states while communicating their own capabilities with each other. For the CXL protocol, the same outline is utilized to bring the link up in CXL mode.



(Modified Training Sets in Config State of LTSSM to negotiate Protocol)

A Downstream Port (DSP) that supports Alternate Protocol Negotiation will start the negotiation process when it first enters Configuration.Lanenum.Wait, LinkUp = 0, and Modified TS Usage Mode Selected field is 010b. Modified TS1/TS2 Ordered Sets are exchanged during Alternate Protocol negotiation with Modified TS Usage = 010b. The DSP is responsible for ensuring that they arrive at a consensus on the Alternate Protocol Negotiation prior to transitioning to Configuration.Complete substate. It is permitted to fall back to PCIe protocol if the Alternate Protocol Negotiation does not arrive at a conclusion. On a successful negotiation to alternate protocol, the Link moves to L0 at 2.5 GT/s, switches the data rate to the higher data rates, performs equalization, if needed and enters L0 at the desired highest data rate.

7.7.6.2 32.0 GT/s Capabilities Register (Offset 04h)

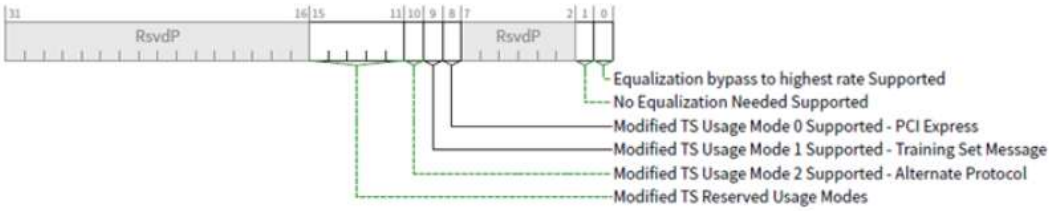


Figure 7-89 32.0 GT/s Capabilities Register

Bit 10 Modified TS Usage Mode 2 Supported - Alternate Protocol - This bit indicates that this Port supports negotiating to use alternate protocols (Modified TS Usage 010b).

7.9.21.2 Alternate Protocol Capabilities Register (Offset 04h)

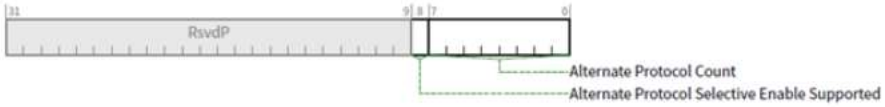


Figure 7-291 Alternate Protocol Capabilities Register

Table 7-235 Alternate Protocol Capabilities Register

Bit Location	Register Description	Attributes
7:0	***Alternate Protocol Count*** - Indicates the number of Alternate Protocols supported by one or more Lanes of this Link. Since support for PCI Express is mandatory, the value of this field must be greater than or equal to 1.	HwInit
8	***Alternate Protocol Selective Enable Supported*** - If Set, the Alternate Protocol Selective Enable Mask Register is present. If Clear, the Alternate Protocol Selective Enable Mask Register is not present and Alternate Protocol Negotiation is controlled soley by the Alternate Protocol Negotiation Global Enable bit. In Upstream Ports, this bit is hardwired to 0b. In Downstream Ports, this bit is HwInit with an implementation specific default value.	RO/HwInit

It is a two-phase process that occurs while in Configuration.Lanenum.Wait, Configuration.Lanenum.Accept, and Configuration.Complete before entering L0 at Gen1 speed. The capabilities between Host and Device is communicated using modified ordered sets, modTS1 and modTS2.

In Phase 1, Downstream Port (DSP) and Upstream Port (USP) advertise their Flex Bus capabilities by sending stream of modified TS1 ordered sets. During this time, it is

important for both the link partners to support the alternate protocol. In Phase 2, the DSP sends a stream of modified TS2 Ordered Sets to the USP to indicate the negotiated capabilities which are nothing but whether the link should operate in PCIe mode or in CXL mode; for CXL mode, it also specifies which CXL protocols and features to enable and whether to operate in 1.1 or 2.0 mode. The USP acknowledges by sending modified TS2 Ordered Sets with the same Flex Bus enable bits set. This exchange occurs during Configuration.Complete.

CXL alternate protocol negotiation successfully completes only after the Downstream Port has confirmed that the Flex Bus enable bits reflected in the eight consecutive modified TS2 Ordered Sets it receives that causes the transition to Configuration.Idle match what it transmitted; otherwise, the Downstream Port logs an error in the Flex Bus Port Status register and the physical layer LTSSM returns to Detect.

After the link is successfully up in CXL mode with the common negotiated capabilities, it reaches L0 state at Gen1 speed, for it to function it should achieve speed of atleast 8GT/s. Because if the CXL link up happens but speed is less, then link will fail and go to Detect state. The capabilities logged in Flex Bus Port Control register and the capabilities communicated by exchange of modified TS1/TS2 ordered sets must be consistent.

2.1.3 PCIe Gen4/5/6 协议分析和诊断碰到的难点

2.1.3.1 信号问题

由于 PCIe 3.0 链路速度为 8Gbps，信号质量问题还不是特别突出。但是，PCIe 4.0 (16Gbps)，PCIe 5.0 (32Gbps)，PCIe 6.0 (64Gbps)对于分析仪的 interposer 设计提出了很大挑战。PCIe 协议分析仪的架构决定了如果要抓取链路的双向数据，必须依靠相应接口的 interposer 串接在链路中间。例如，参见下图，如果要分析 U.2 NVMe SSD，那么就需要一个 U.2 interposer 串接在 U.2 背板和 U.2 SSD 中间，interposer 的设计目标是不影响双向数据交互，同时将双向信号各分出一路信号送到“旁路”的 PCIe 协议分析仪前面板的

upstream 和 downstream 端口，upstream 和 downstream 端口在机箱内部分别连接到一块抓包分析板，将收到的 ordered set 或者 packet 打好时间戳、格式化好以后写到内部的高速 buffer 缓存。

图 2-11

但是，真实的 interposer 产品，由于设计、架构等方面的原因，对于 16Gbps、32Gbps、64Gbps 的高速信号往往非常容易在串接 interposer 后导致各种问题。

2.1.3.2 解码速度

PCIe Gen3 时代时候的协议分析仪一般的抓包 buffer 在 4.5G, 9G, 18G, 36G 等，但是 PCIe Gen4/5/6 由于速度提高了 2 倍，4 倍，相应的分析仪的 buffer 最低都在 36G, 一般配置 72G, 144G, 288G。如果分析一些 Gen4 或者 Gen5 x16，一般都是建议配置最低 144GB buffer，因为即便这样在双向流量较大的情况也仅能抓取很短时间的数据。

我们以 PCIe Gen 4 x4 为例，即 16G x4 = 64G，如果双向打满流量（Read/Write: 50% / 50%）的情况下的理论吞吐量大概为 64G * 2 = 128Gbps（约 12GB/S, 128/130bit 编码）。由于协议分析仪抓包的时候需要加上时间戳（timestamp）以及很多其它格式化信息（例如标识 Packet 是否有 CRC Error），所以实际占用的 buffer 远比我们链路上传输的数据要大，基本上读/写压力同时加上的话，几秒钟几十个 GB 字节。但是这对于 PCIe Gen 4/5/6 协议分析仪的 Trace 解码分析速度和文件保存速度带来了挑战。

市场上见到的传统的 PCIe 协议分析仪，包括所有除 SerialTek 之外的 Gen 4/5/6 分析仪，都是一种嵌入式架构，可以简单理解成和一台传统“打印机”或者“投影仪”架构类似，分析仪硬件的主要功能只是抓取双向数据，之后需要通过两个步骤才能实现协议解码。

第一步，通过分析仪内部嵌入式 CPU（一般都是几百 Mhz）将数据传输到协议分析仪软件，一般使用网口或者 USB，这个过程由于受制于分析仪内部较弱的 CPU 限制往往传输速度非常慢；

第二步，协议分析仪软件进行解码显示，这个过程由于都是通过程序单线程进行解码分析会非常慢，同时由于处理上百 GB 的数据的时候受制于电脑内存（例如笔记本的内存通常为 16GB）的限制会在内存和硬盘之间来回“倒腾”数据，所以很慢。总之，用户的体验很不好。实际测试结果：传输 32GB 数据大概需要 4 小时，解码还需要 4 小时，需要 8 个小时才能看到完整解码。如果解决一些读/写不一致的问题，那么可能要抓取尽量多的数据，例如 100GB 数据，即便电脑和分析仪软件不崩溃的话，可能需要 48 小时以上才能看到解码。这对于好不容易复现一个问题，然后再进行问题分析来讲，效率非常低。

对于 PCIe Gen 5 x16 而言，双向打满数据，大概为 32Gbps * 16 lane *2 = 1TBbps（约 100GB/s），所以，buffer 大小为 128GB 的分析仪还无法抓到 1 秒钟数据，所以处理这些大的数据的能力就成为考虑协议分析仪的一个非常重要的方面。

SerialTek 公司的创新设计的 PCIe Gen4/5/6 协议分析仪采用高性能 server 架构，所有解码等工作都在分析仪内部完成，解码完毕后直接将通过 Web 界面展示到客户端的 Chrome 等浏览器，有点类似于视频会议 Webex/Zoom/腾讯会议等，占用带宽只有 20KBps，所以解码显示非常快，早期的 Gen4 协议分析仪解码 144GB 字节基本 1 秒钟即可完成，当然 Gen5/6 协议分析仪停止抓数据后会有一个 post-process 预处理过程稍微需要一点时间，后处理完毕后工程师可以在解码界面上直接拖动到最后一行查看解码，如果需要重新抓取数据，下一秒即可开始。

2.2 SerialTek PCIe Gen 4/5/6 协议分析仪的革命性设计

SerialTek 总公司为位于瑞士的 Ellisys 公司（Made in Switzerland），但是其 Kodiak 系列 PCIe Gen 4/5/6 协议分析仪硬件（包括分析仪主机，以及所有的 Interposer）的研发，设计以及测试完全在英国伦敦完成。

SerialTek 从 PCIe/NVMe Gen 4 分析仪开始首度开始采用其业内首创的全新架构设计，开创了协议分析仪采用高端服务器架构的先河，其分析仪内置 12 核高性能 CPU，区别于传统的协议分析仪软件的“胖客户端”模式*（该模式下，所有的分析等功能全部依赖工程师的电脑的性能）*，该 Gen 4 分析仪采用 Client/Server 架构，即“瘦客户端”+高性能 server 的架构，客

户端协议分析软件只是负责产品设置，管理以及显示，所有的需要处理的内容都放在 server 端进行处理，这样工程师的电脑将获得很大的解放，也不会成为协议分析时的瓶颈。

SerialTek Gen 4 分析仪提供千兆以太网管理端口，以及 2 x 10GE 管理端口，提供 36/72/144G Trace Buffer 用于抓取 PCIe/NVMe 流量，内置 2TB 本地闪存用于快速保存 Trace 文件，同时也可以直接保存到连接在分析仪前面板的 USB 盘或者 PCIe 盘柜，是全球目前最快的 Gen 4 分析仪。

PCIe 协议分析仪作为 PCIe 总线分析的基本工具，不仅仅用于主机，网络，存储系统等各种 IT 和通讯设备针对 PCIe 插卡的问题分析，同时也是 PCIe/NVMe SSD 分析的必备工具。

作为 PCIe 协议分析革命性创新的领导者，SerialTek 公司的 PCIe Gen 4/5/6 协议分析仪不仅颠覆了传统的 PCIe 协议分析仪架构设计，大大提高了协议分析仪的性能以及用户的测试效率，改变了用户使用 PCIe 协议分析仪的习惯，同时，它也提供了超高的灵活性和业内最高的性价比，让更多的公司买得起 PCIe Gen 4/5/6 协议分析仪。下面我们简要介绍一下这些创新功能。

图 2-12 SerialTek PCIe Gen5 x16 协议分析仪前面板

图 2-13 SerialTek PCIe Gen5 x4 协议分析仪前面板（后面板同 Gen5 x16）

图 2-14 SerialTek PCIe Gen5 x16 协议分析仪后面板

2.3 SerialTek PCIe Gen 4/5/6 协议分析仪创新功能

2.3.1 信号高保真

参见下图所示，该图装置为 **PCI SIG Workshop 测试 PCIe Gen5 Lane Margining 的测试环境**所使用设备，不同客户的 Gen 5 DUT 插卡插入 SerialTek PCIe Gen 5 x16 slot interposer 上面进行测试。参见下图倒数第二行，2021/9/27 一周针对全球的 PCIe Gen 5 的 Interop 以及 2021/12/6 针对 PCIe Gen 4 的 Interop 测试即采用 SerialTek Kodiak Gen 5 协议分析仪。

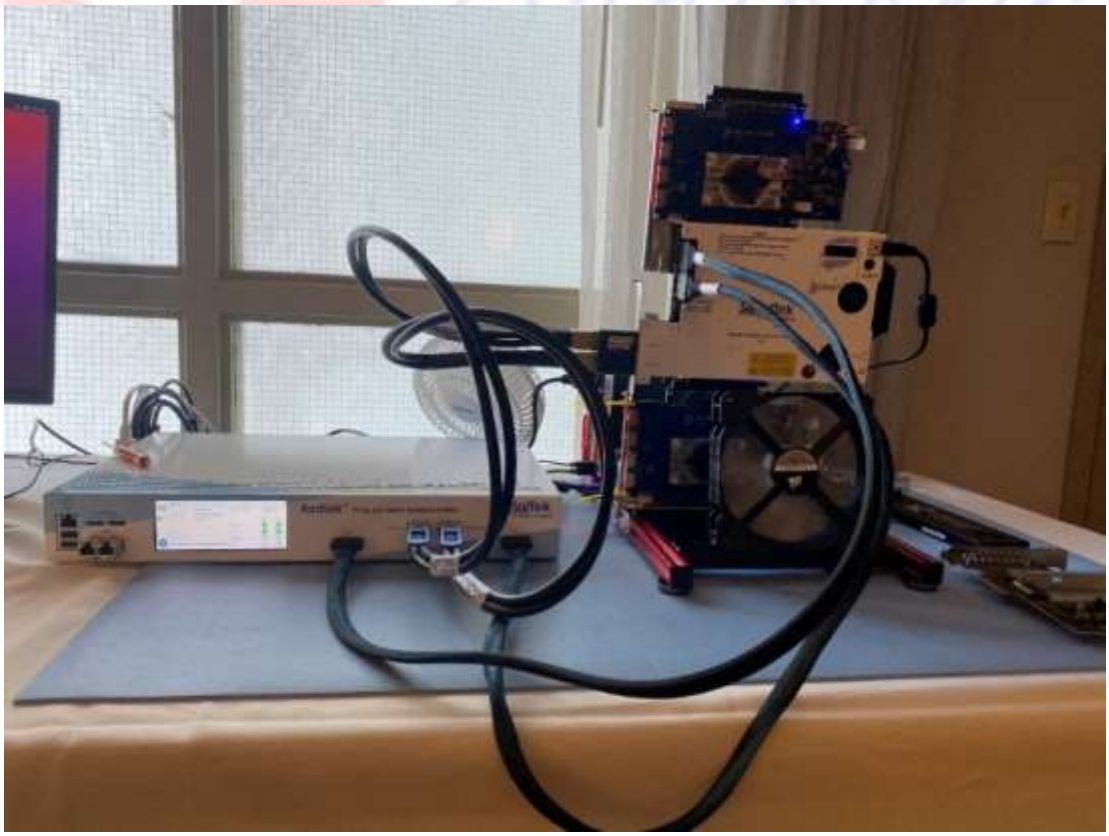


图 2-15

图 2-16

分析仪必须配合各种接口的 Interposer 串接在 PCIe 链路中一起使用，由此导入的信号问题是判断一个分析仪是否可用的一个基本问题，业内估计没有公司愿意花上百万或者几百万人民币购买一台 PCIe 分析仪后发现在很多场景下无法使用。不幸的是，我们发现传统架构的 PCIe Gen4/5/6 分析仪普遍存在这种问题甚至更严重。如果公司在做最终决定前可以选择在尽量多的真实环境中试用 PCIe 分析仪，这样可以大大避免购买以后出现问题。下面的问题是传统 PCIe 分析仪在真实环境中经常出现的问题

- **完全抓不到任何数据**
- **待测系统无法启动**
- **待分析问题症状消失**
- **信号不好，抓到各种错误**

碰到上述问题以后，传统 PCIe 分析仪需要非常复杂的 calibration，用户工程师一般无法搞定，其实，即便原厂 R&D 设计工程师使用内部专用工具软件进行信号的 calibration 校准也不一定搞好，因为这些问题大多由于其 interposer 内部设计造成的。

SerialTek 公司的 PCIe Gen 4/5/6 协议分析仪设计具备自适应的 EQ 能力，并且当 PCIe 链路特性发生变化的时候（例如 Hotplug 或者 NSSR）分析仪可以动态调整，其 interposer 采用昂贵的高端 analog passthrough 的模拟芯片将 upstream 和 downstream 信号导入分析仪，避免了上述这些问题的出现，也无需用户进行信号校准（Calibration-free）。

我们来看一下业内知名的芯片公司 Phison 的首席工程师是如何来评价 SerialTek 的这一创新性的专利技术- 信号高保真 SI-FI （Signal Fidelity）。

I've been using protocol analyzers for 31 years and PCIe analyzers and Interposers extensively for the past 5 years. We use them for important assignments that affect revenue and customer satisfaction," said John Wehman, Principal Engineer at Phison Technology. "With other analyzers I have had to abandon my testing many times, because I could not find a good quality signal lock. SerialTek's Kodiak

analyzer and SI-Fi Interposers have changed all that. I have 100% confidence in Kodiak's ability to achieve lock and give me the trace I need to do my job. Kudos to Ellisys and SerialTek for creating not only an electrically reliable platform, but the actual mechanical hardware itself is beautiful.

SerialTek 的所有 AIC, U.2/U.3, M.2, EDSFF, Cable Interposer 等分析板卡采用其专利技术的 SIFI 信号高保真（Signal Fidelity）设计，Interposer 除了两端接插件部分有极低的的信号衰减外，板内几乎没有信号衰减，主要原因在于其设计采用了高成本的"宇航级"分路器件将 PCIe Upstream 和 Downstream 双向数据导出到协议分析仪主机，进/出 Interposer 的信号眼图几乎一样。

对比：传统分析仪在处理 Gen 4/5/6 的信号的时候为了减少 Interposer 带来的衰减，全部采用对于信号进行增强（采用 Gen5 retimer 或者 redriver）的方式进行处理，导致 Interposer 入口的信号和出口的信号的眼图差距较大。这就是导致很多用户工程师看到很奇怪的现象：1）接入分析仪 Interposer 以后"原来的问题不见了"；或者 2）接入分析仪 Interposer 以后"出现了新问题"，因为 interposer 在 PCIe 链路中间将信号完全改变了。

2.3.2 "超快"解码

在开发/测试的不同阶段使用分析仪的时候可能存在不同的应用场景。例如，非常早期的阶段可能在使用分析仪的时候需要设置触发条件抓取少量的数据分析即可，但是在产品的后期测试阶段，AE 支持，客户方案支持，或者产品发布以后的技术支持阶段，很多不容易复现的问题往往需要抓取大量数据，例如，NVMe SSD 运行很长时间以后出现读/写不匹配(Read/Write Miscompare)，这种情况下往往采用大压力并发读/写， Write 和 Read 同一个 Sector 的时间会间隔几秒有时甚至更长，这个时候需要抓取所有读/写数据然后进行对比分析。SeriaTek 支持最大配置 144G Buffer，但是这么大的 Buffer 抓到以后解码就成为影响测试效率多个一个严重问题。SerialTek 的创新设计使得抓取 144G Buffer 以后可以在 1 秒钟之内全部解码所有的 PCIe 层（DLLP, TLP）以及 NVMe 层命令。如果工程师简单分析解码后如果发现不是所需要的 Trace 可以立即重新开始抓取。

对比：根据前面的概述，传统分析仪抓取数据以后必须经过两个步骤解码：1）将 Trace 从分析仪 Buffer 读取到电脑；2）通过电脑的 CPU/内存进行解码。实测：导出+解码 32G 需要 8 小时，导出+解码 144G 字节数将花费超过 48 小时，并且很可能工程师的笔记本电脑会死机。

2.3.3 "极速"存储

抓到 144G buffer 分析解码以后如果觉得需要保存下来供其他部门或者同事协同分析，那么需要多少时间呢？SerialTek 提供多种方式保存 Trace 文件：

- **保存到分析仪内置的闪存盘**

  SerialTek PCIe Gen 4 协议分析仪内置最大 2TB Gen 3 x4 NVMe SSD（Samsung EVO970），写入速度大概在 350MB/S，保存 144G 大概需要 6.5 分钟。这是目前使用 SerialTek 最推荐的方式。

- **保存到用户电脑（通过网络传输）**

SerialTek 提供 1 个千兆，外加 2 个万兆 10GE 端口用于管理和导出数据，工程师通过千兆端口导出 Trace 的速度大概在 90MB/s (千兆理论速度是 1.25Gbps)；单端口 10GE 提供 1GB/s 速度，双端口 10GE 提供 2GB/s 速度。如果需要使用 10GE，建议使用台式机或者服务器配置 Intel 双 10GE 端口网卡。

- **保存到外置 PCIe 闪存盘或者阵列 ****

SerialTek 支持 2 个 Oculink，可以通过 Oculink to U.2 线缆接入 NVMe SSD 或者直接接入 Oculink 的盘柜。

- **保存到 USB 3.0 移动硬盘**

SerialTek 支持 2 个 USB 3.0 端口，可以直接将 U 盘插入，然后直接将 trace 文件保存到 U 盘。

- **支持外连 10GE 接口基于 NVMe SSD 的 NAS 存储**

建议使用类似于 QNAP 公司的 10GE 网口的 NAS 存储，内置 NVMe SSD 固态盘，提供高速访问。

对比：SerialTek 分析仪采用高性能服务器设计，内置标准 Linux 系统，大大提高了对比：SerialTek 分析仪采用高性能服务器设计，内置标准 Linux 系统，大大提高了 Trace 文件保存的速度和保存方式的多样性。传统的 PCIe 分析仪架构可以认为类似一台终端“打印机”，本身仅提供抓包功能，不提供处理和分析，所有操作全部依靠工程师的电脑。所以，电脑性能配置强一些可能体验稍好一些。但是，传统分析仪内部较低频率的嵌入式 CPU 及其精简 OS 系统严重束缚了 USB 或者千兆以太网导出数据的速度，平均导出速度在 3~5MB/s。实测仅导出 32GB trace 需要 4 小时。

2.3.4 无需抓取“上电过程”

SerialTek 分析仪设计采用其专利技术，无需抓取 PCIe 上电初始化过程即可实现正确解码。这一点非常重要，因为很多问题的复现需要反复重启电脑或者插拔 NVMe SSD，SerialTek PCIe 分析仪只要处于加电状态，不论工程师打开/连接协议分析软件与否，分析仪内部都会时刻监控每次上电 PCIe 初始化过程中任何 config space, controller register, 协商的速度和位宽（如：Gen 4 x4），以及 admin 和 I/O 队列的创建和拆除时间等等，所以任何时候连接分析仪开始抓取数据（即便已经错过了上电初始化过程），然后分析仪都会按照正确的信息进行解码。

对比：传统 PCIe 分析仪由于设计的缺陷，必须严格按照如下顺序抓取数据：1）协议分析仪上电；打开软件，连接并且锁定协议分析仪；2）设置分析仪参数，开始启动抓包；3）待测环境上电。如果主机上电以后再连接分析仪抓取数据很可能会解码错误。这样设计的原因在于分析仪必须先“经历”PCIe 初始话过程学习到这些参数，作为后面解码的依据。但是这在工程师需要反复重启复现问题的时候成为严重缺陷，即，某一次重启可能导致 PCIe 协商的某些参数变化，结果导致解码全部错误。该问题在传统分析仪上无法解决，反复重启解码错误的问题 100%会出现。

2.3.5 创新性的基于时间轴的 LTSSM 分析

下面是传统的 LTSSM 分析图，工程师随着时间分析这些状态之间的跳转的时候非常费劲。

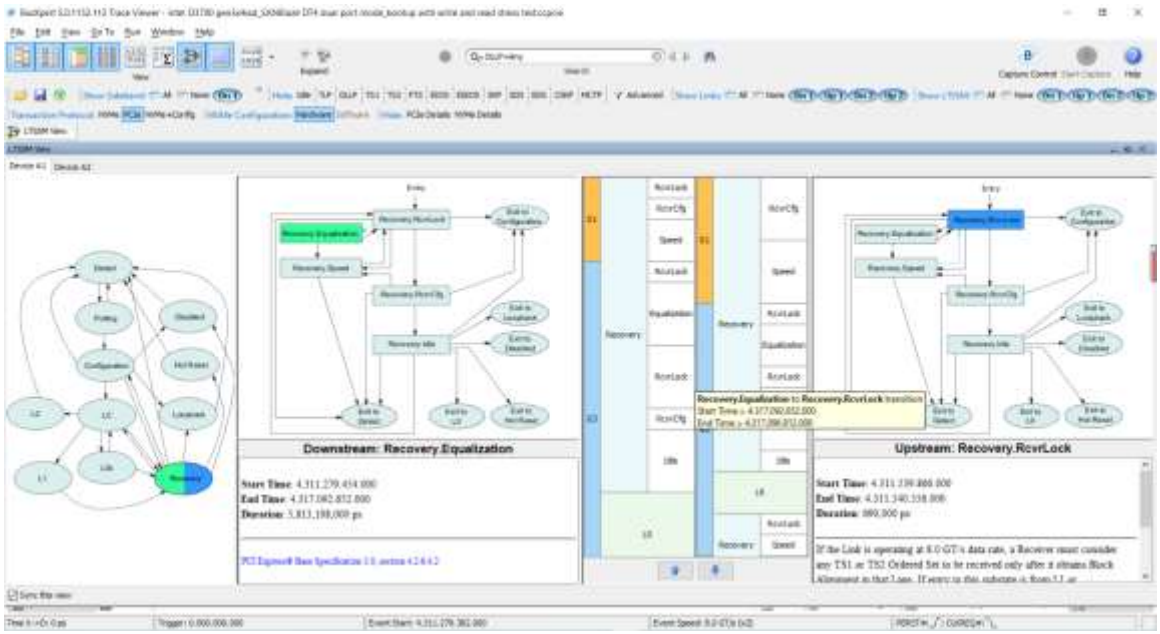


图 2-17

下面是 SerialTek 创新的 LTSSM 分析图，横轴是时间轴，纵轴是 upstream 和 downstream 多个 LTSSM 之间的状态。

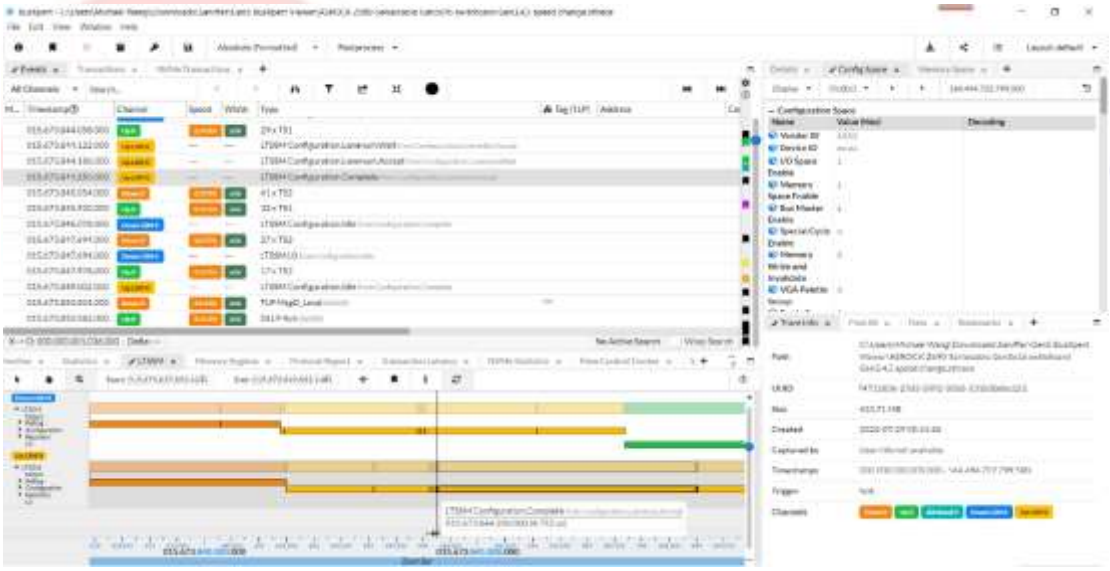


图 2-18

工程师可以通过鼠标拖动时间轴快速查找到问题点。首先，从全局预览方式找到大致的问题点，然后通过 zoom in 放大到局部，然后将鼠标放在具体的位置即可显示 LTSSM 的当前状态，这大大提高了问题分析的效率。

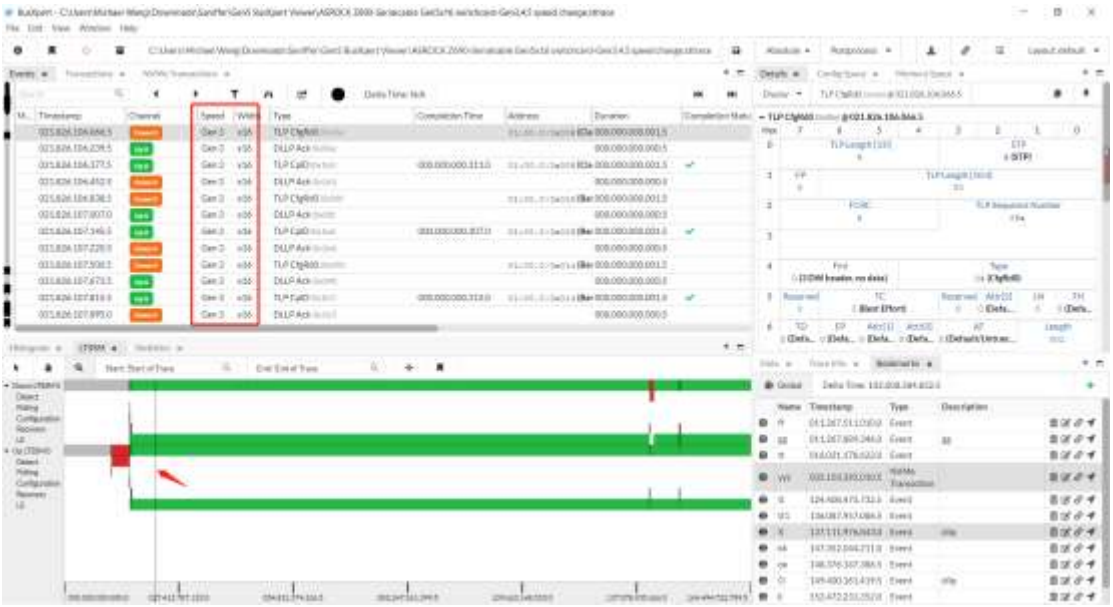


图 2-19

** 上图的 LTSSM 展示页面任何有竖线的地方表示有状态机变化，横轴为抓取的 trace 的时间轴，所以，在抓取的时间段内任何地方有 LTSSM 状态机变化都一目了然。

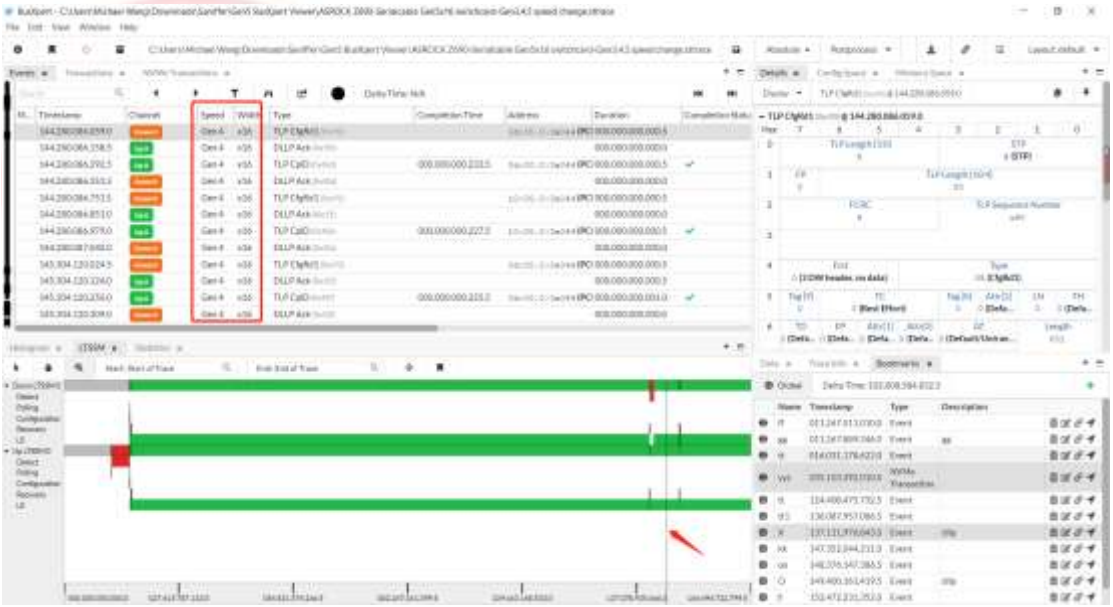


图 2-20

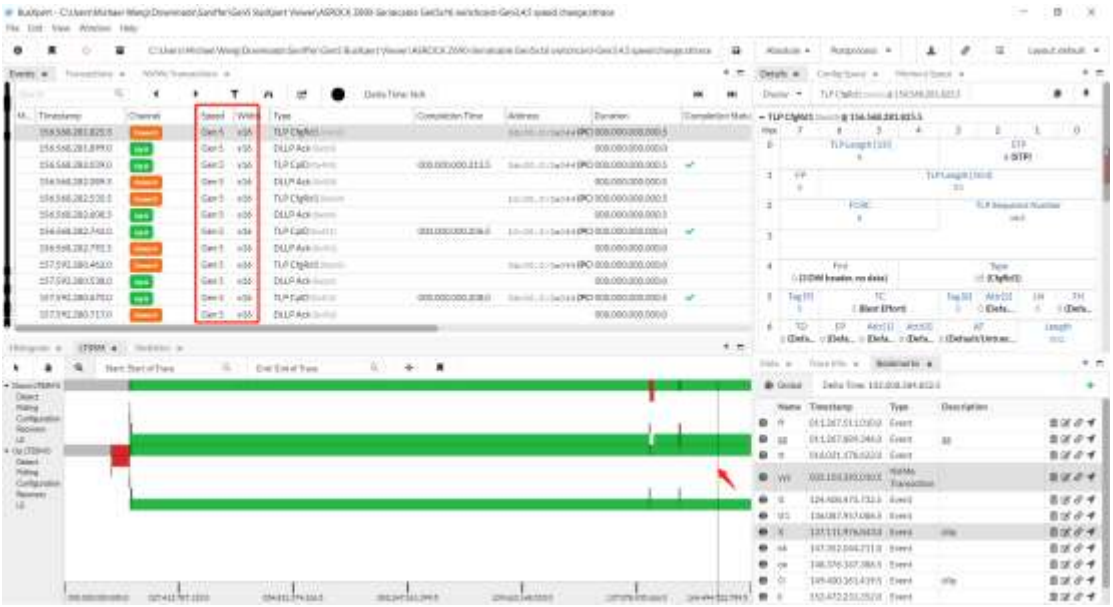


图 2-21

上面三张图是整个 trace 文件 LTSSM 展示的图片。我们看到经过 3 次 recovery 后 PCIe Gen5 主机和 PCIe Gen5 switch card 协商依次进入 Gen3 x16, Gen4 x16 和 Gen5 x16。下图是第一次 recovery 之前的放大。

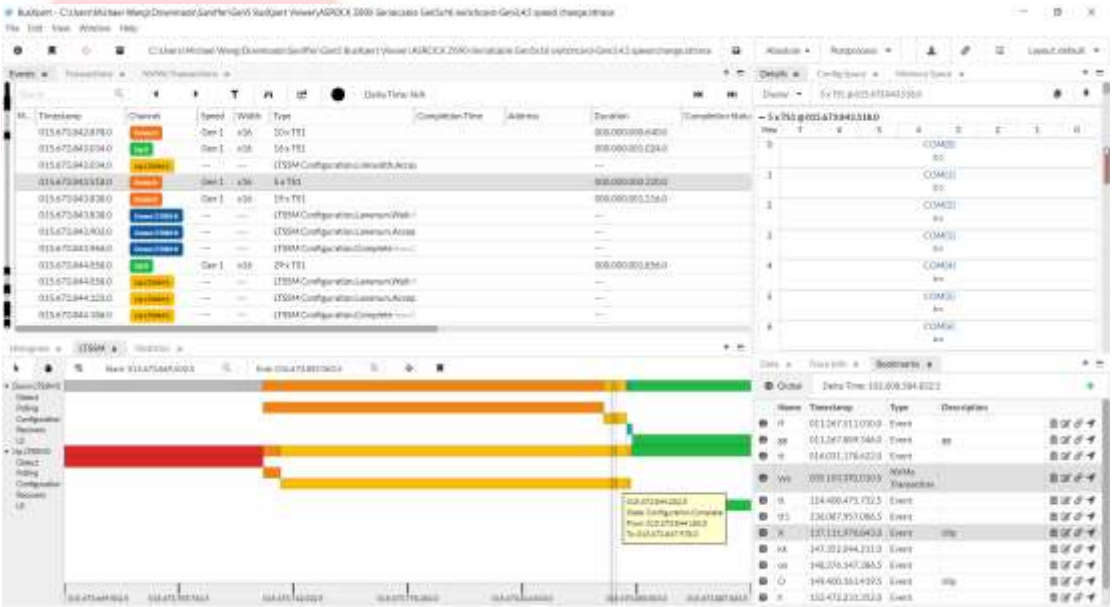


图 2-22

下图是第一次 recovery。

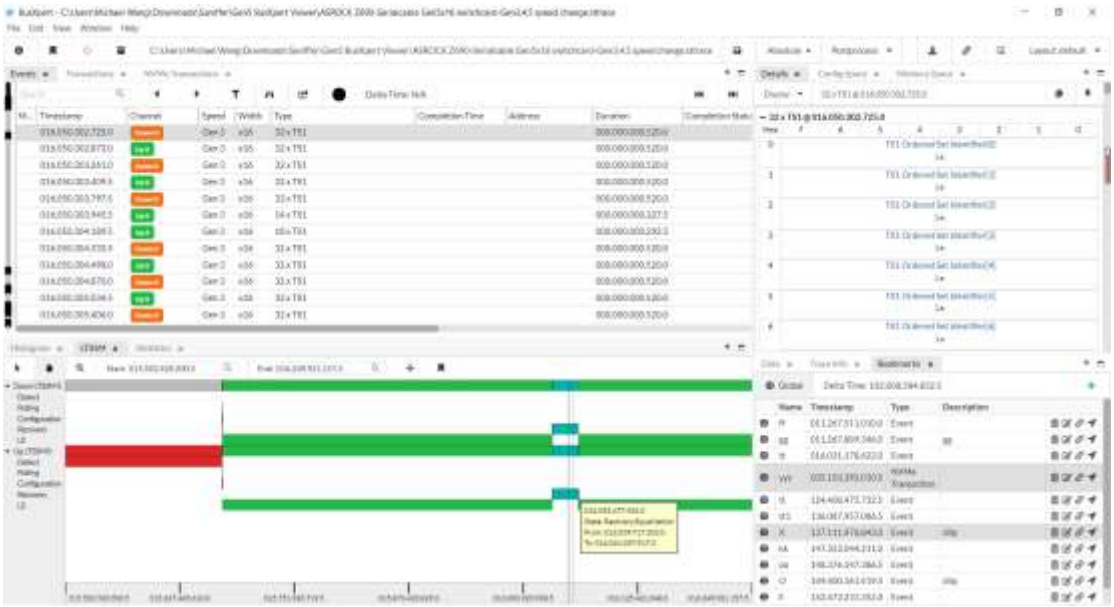
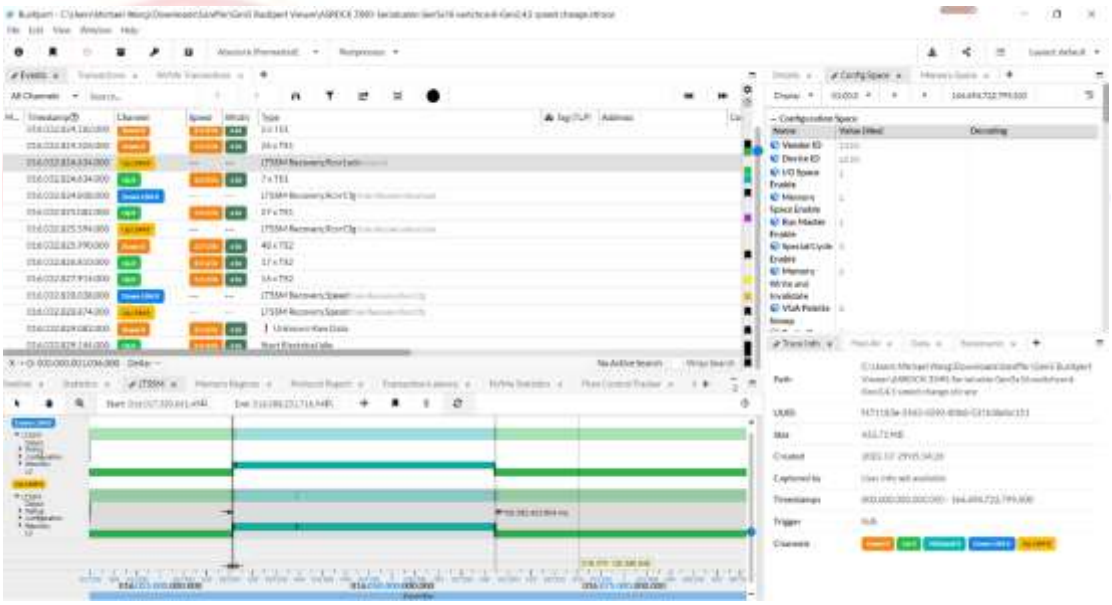


图 2-23

下图是 recovery 持续的时间。从图上我们看到 第一次 recovery 的时间大概持续了 28ms。



下图是第一次 recovery 从 Gen1 -> Gen3 快结束时候的放大展示。我们看到 downstream 最后先进入 L0, 但之前双方曾经短暂进入 L0。

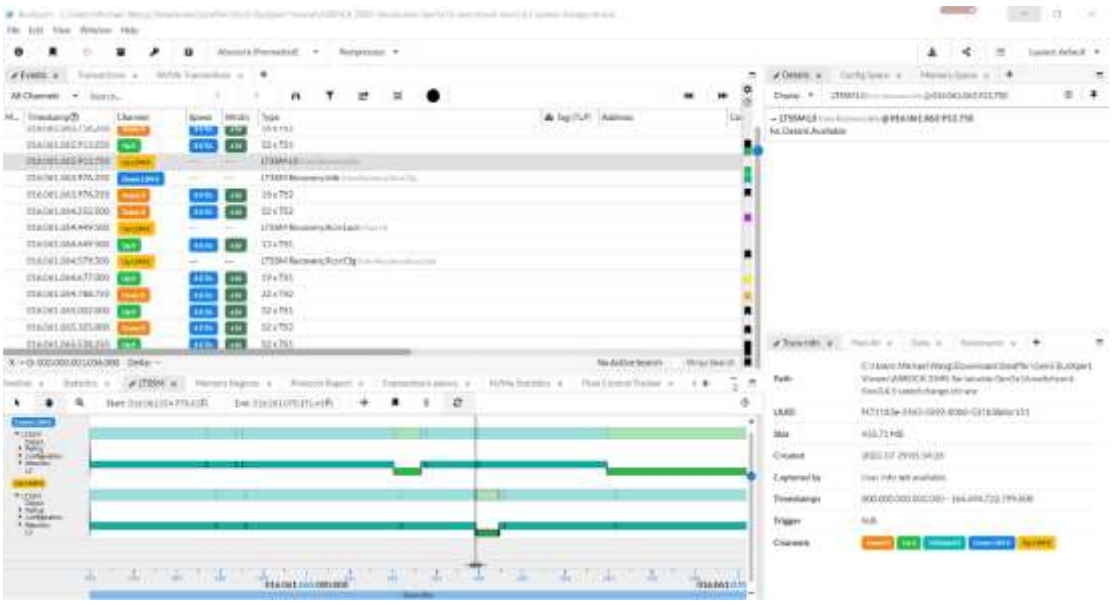


图 2-24

下图是最后一次从 Gen4 经过 recovery 协商到 Gen5 的细部放大。

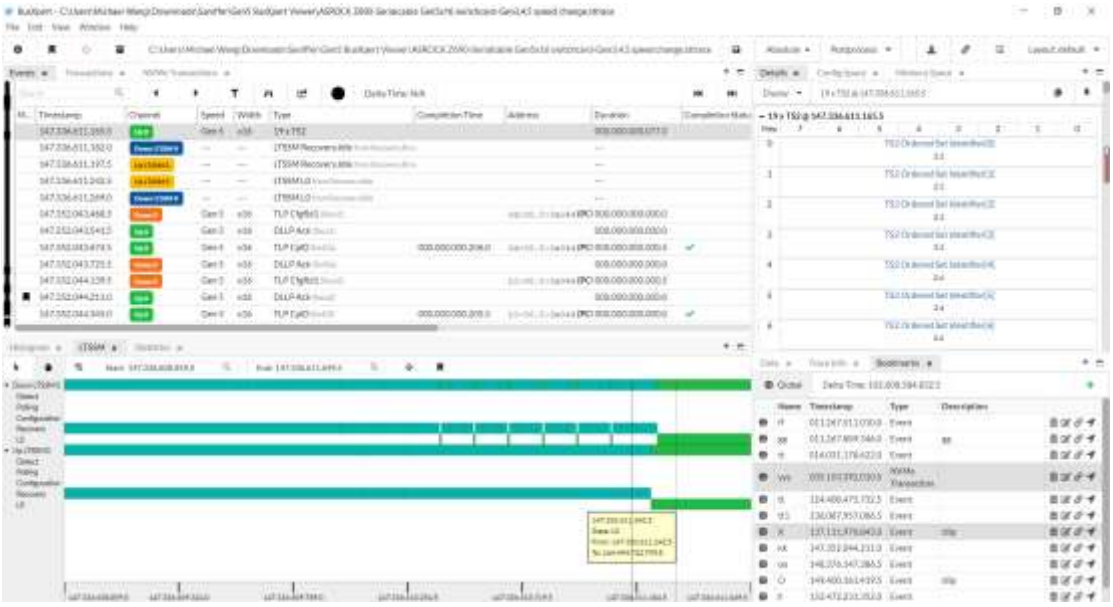
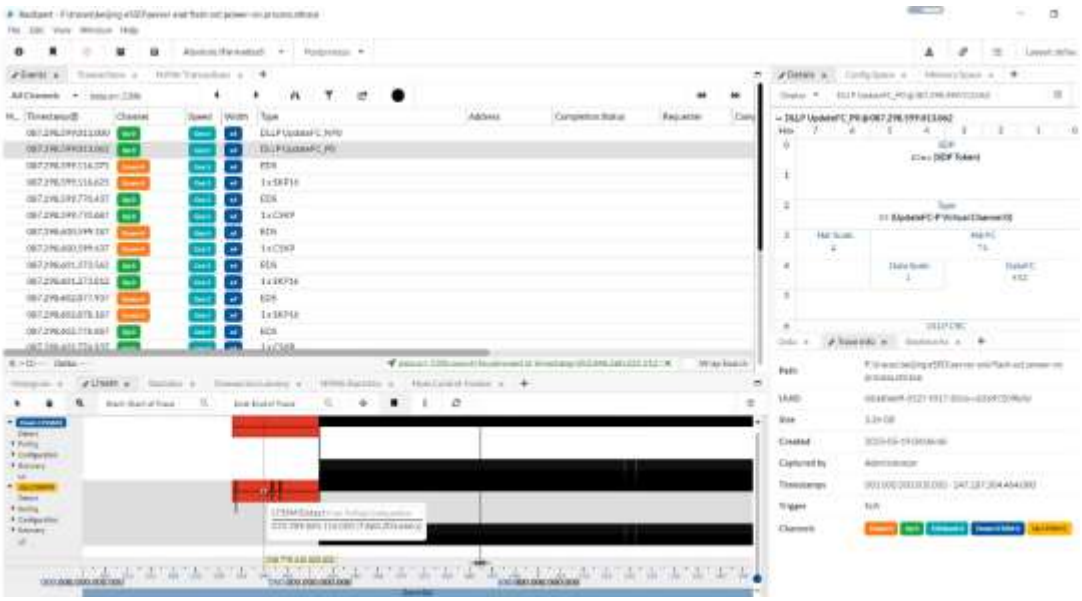
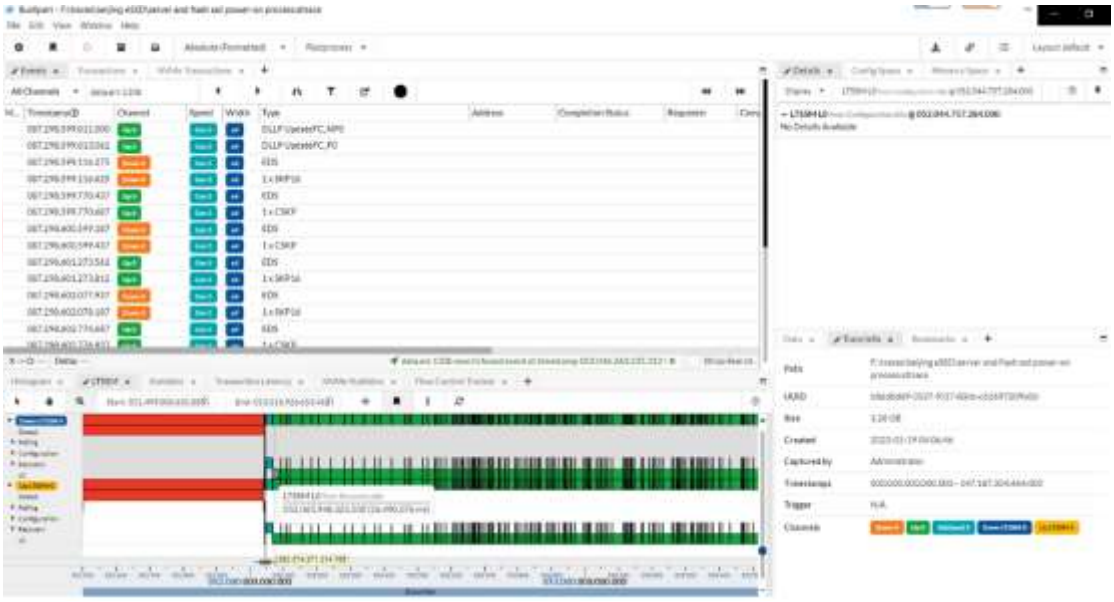


图 2-25

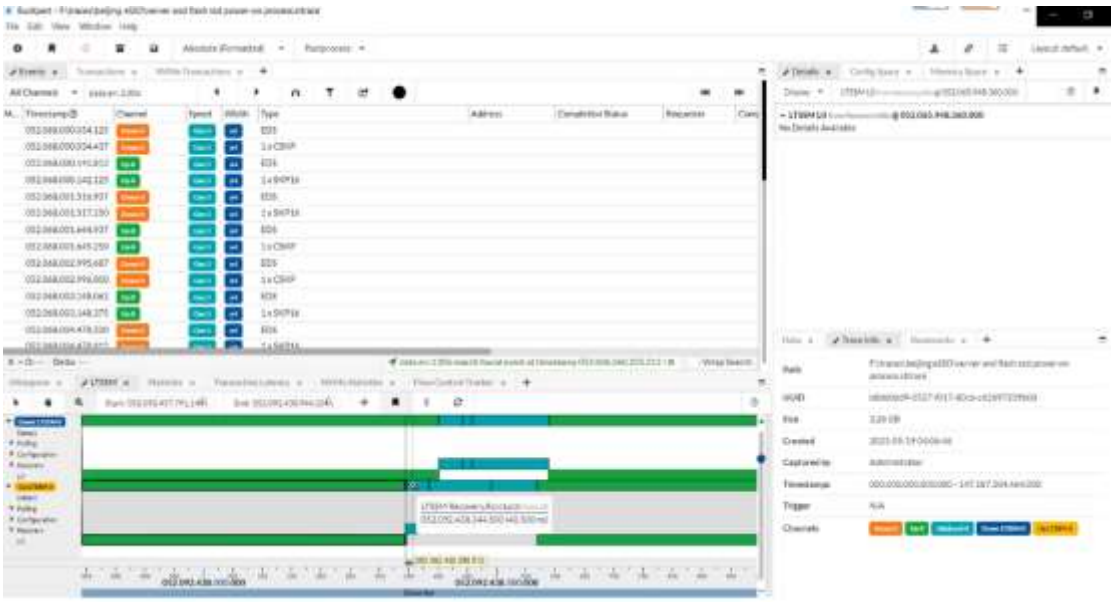
我们有的时候也会发现链路质量不好的时候会有大量的 recovery，类似于下面的显示，recovery 相当多，黑乎乎一大片。



你如果去放大到一定程度，就会发现展示如下。



上面这些竖线都状态机变化，我们再来进一步放大看没一条竖线，发现都是由 upstream 初始的 recovery：

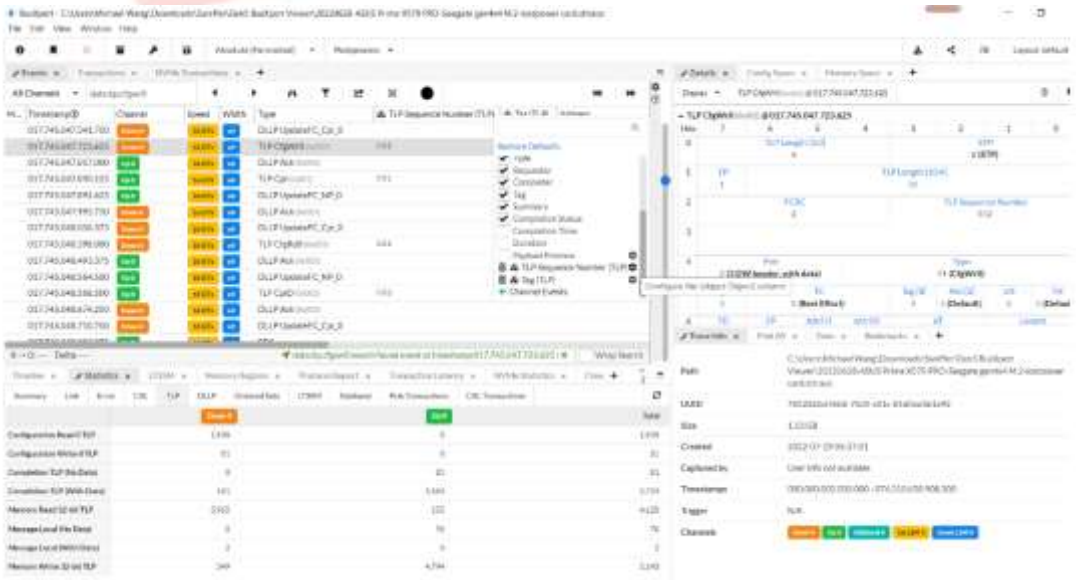


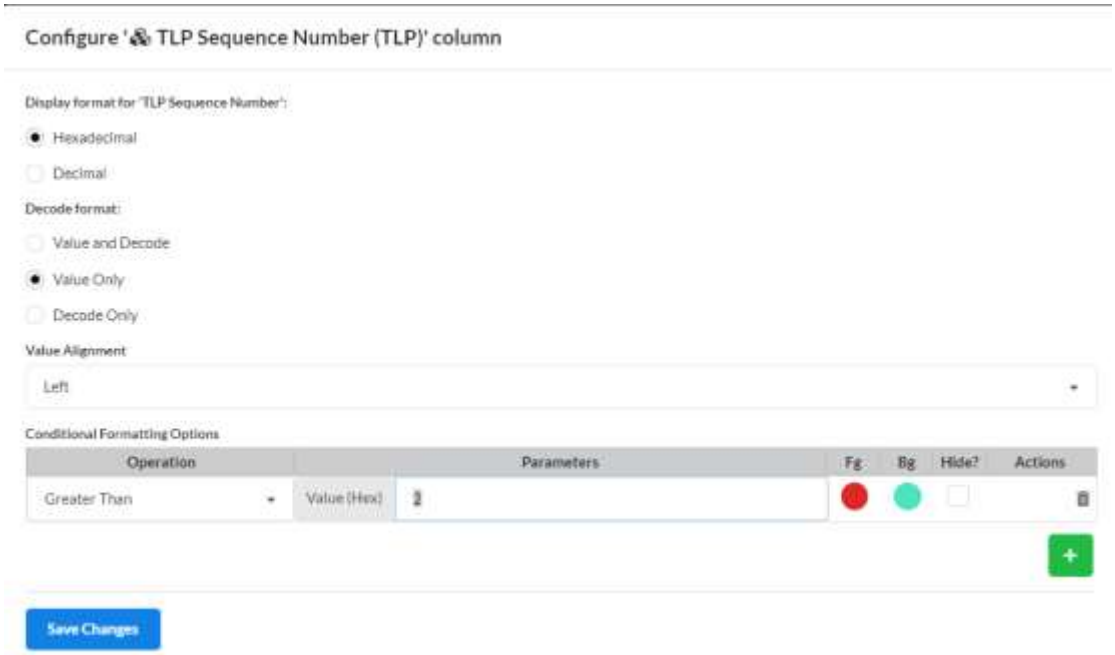
PCIe 信号不好的时候我们经常发现大量的 recovery，可能每秒几十个到几千个，即便没有掉速、掉 lane，也没有错误，应该也要引起注意，因为这表示信号不好。

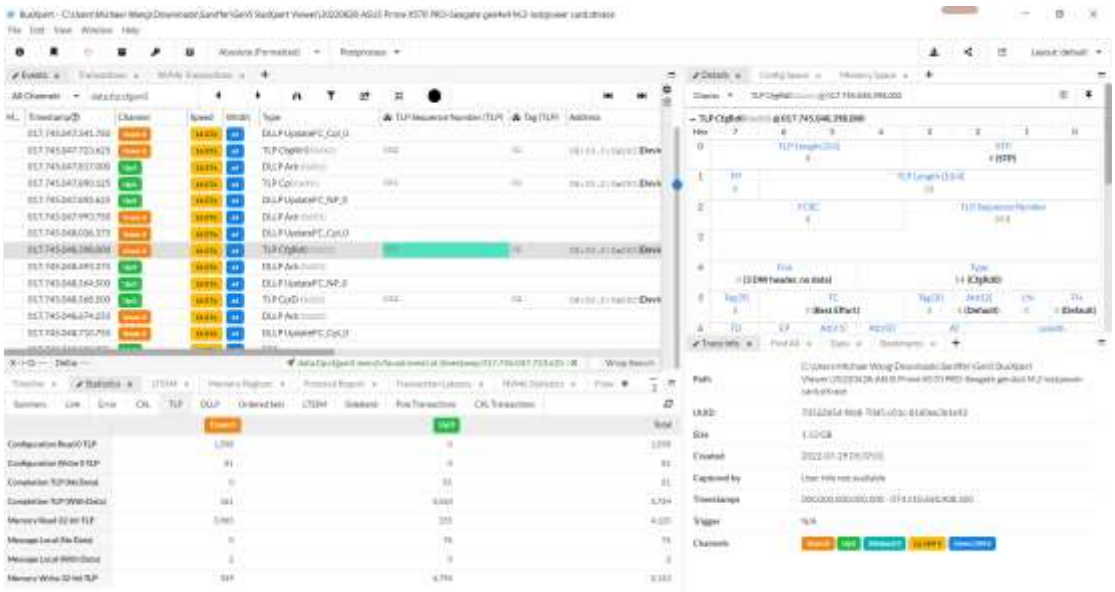
需要参考 L1.2 低功耗的 LTSSM 展示可以参考 2.3.7 章节。
需要具体了解该功能，可以访问下面的链接看实际演示。

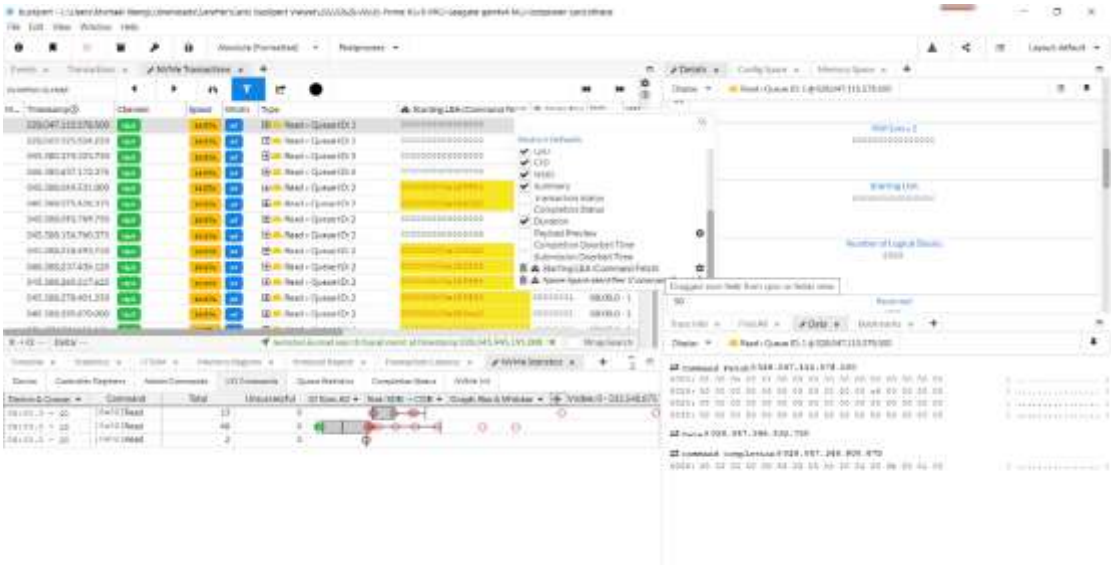
业内能用的 PCIe Gen5 协议分析仪 VS 最经济的 Gen5 协议分析卡
https://mp.weixin.qq.com/s/iWIv1YCmBXia_-VuDj5fBg

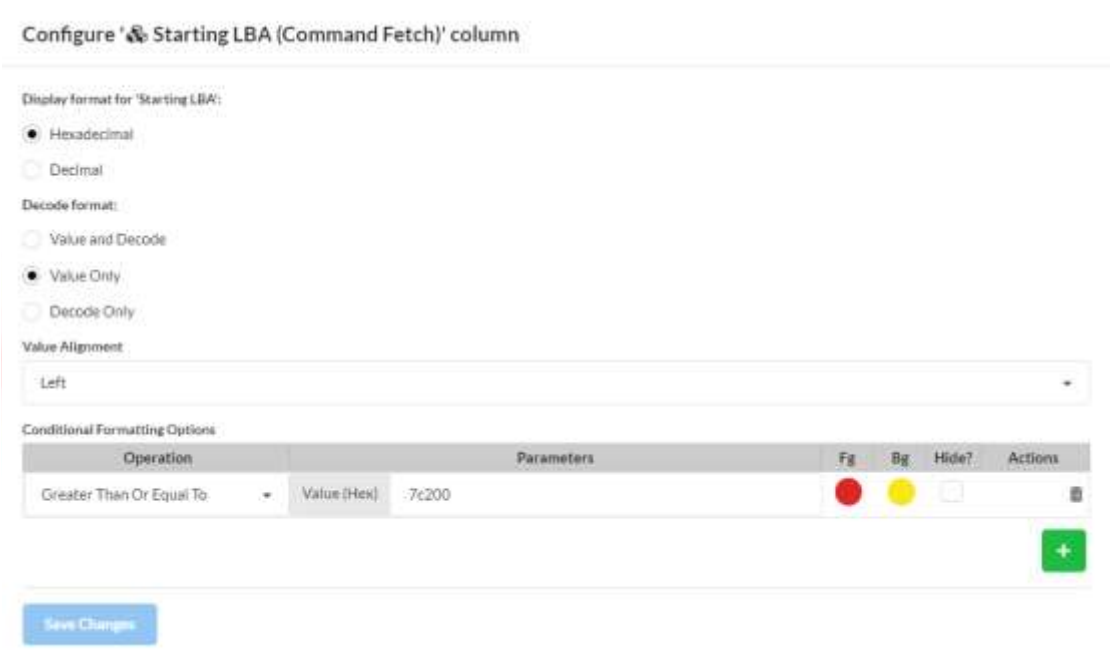
2.3.6 任意定制解码窗口的显示列

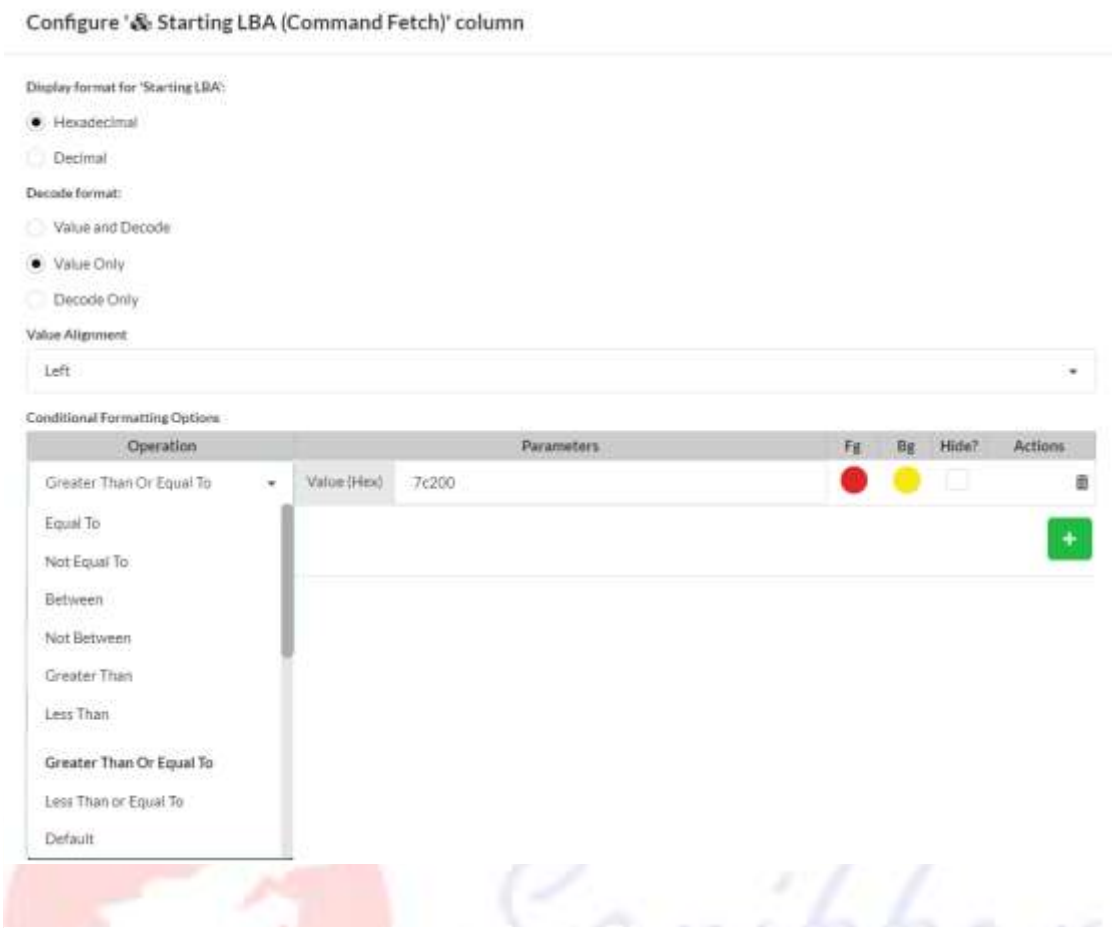



Configure '🔊 TLP Sequence Number (TLP)' column
Display format for 'TLP Sequence Number':
Hexadecimal
Decimal
Decode format:
Value and Decode
Value Only
Decode Only
Value Alignment
Left
Conditional Formatting Options
Operation
Parameters
Fg
Bg
Hide?
Actions
Greater Than
Value (Hex)
2
Save Changes






Configure '🖧 Starting LBA (Command Fetch)' column
Display format for 'Starting LBA':
Hexadecimal
Decimal
Decode format:
Value and Decode
Value Only
Decode Only
Value Alignment
Left
Conditional Formatting Options
Operation
Parameters
Fg
Bg
Hide?
Actions
Greater Than Or Equal To
Value (Hex)
7c200
Save Changes


Configure '🖧 Starting LBA (Command Fetch)' column
Display format for 'Starting LBA':
Hexadecimal
Decimal
Decode format:
Value and Decode
Value Only
Decode Only
Value Alignment
Left
Conditional Formatting Options
Operation
Parameters
Fg
Bg
Hide?
Actions
Greater Than Or Equal To
Value (Hex)
7c200
Equal To
Not Equal To
Between
Not Between
Greater Than
Less Than
Greater Than Or Equal To
Less Than or Equal To
Default


Configure '🖧 Starting LBA (Command Fetch)' column
Display format for 'Starting LBA':
Hexadecimal
Decimal
Decode format:
Value and Decode
Value Only
Decode Only
Value Alignment
Left
Conditional Formatting Options
Operation
Parameters
Fg
Bg
Hide?
Actions
Greater Than Or Equal To
Value (Hex)
7c200
Equal To
Not Equal To
Between
Not Between
Greater Than
Less Than
Greater Than Or Equal To
Less Than or Equal To
Default

2.3.7 “完美”M.2 低功耗支持

SerialTek 的 PCIe Gen 4/5/6 协议分析仪在 Asus Z390，以及 Thinkpad, Dell, Huawei 等笔记本电脑，以及 Gen 4 M.2 （例如：如最新的微软 Surface Book）的 ASPM L1.2 低功耗模式下工作非常良好，进出 TS1/TS2，抓取数据，不丢包，解码正常。

下面是 2.3.5 介绍的最新的 LTSSM 状态机分析 L1.2 低功耗的截图供参考。

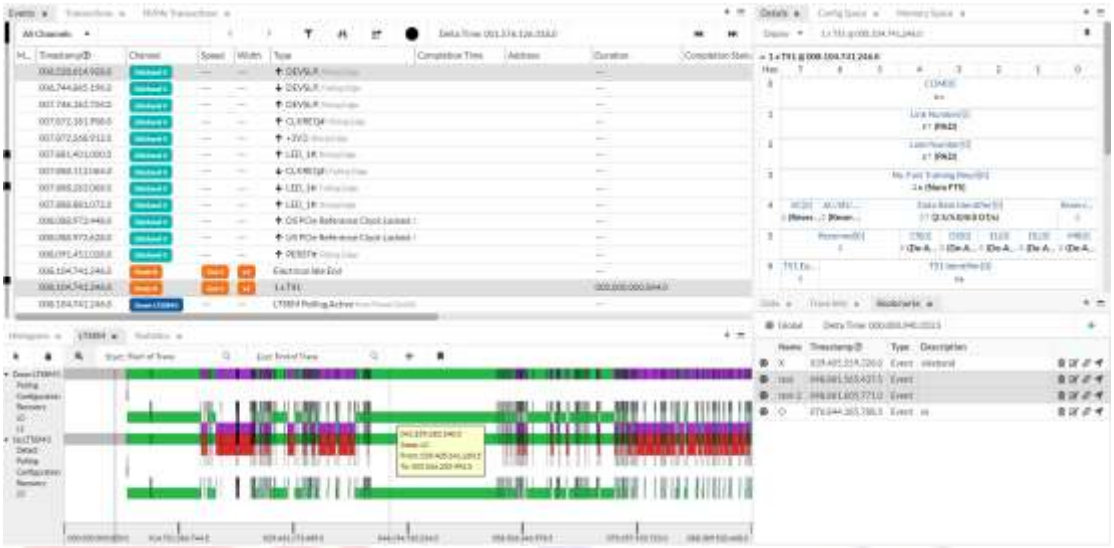


图 2-26

上图中左下角的 LTSSM 视图中的横轴是抓取的整个 trace 文件的时间跨度，左边的 downstream 和 upstream 的各个子状态依次在时间轴上展示。其中每一根竖线表示一个状态变化，不同的颜色标识不同的状态，例如红色为 DETECT，绿色为 L0，紫色为 L1.X。通过工具栏中的“放大镜”功能可以快速对于任何觉得有疑点的时间轴上面的 LTSSM 状态进行放大，然后通过旁边的“手状”拖曳模式图标左、右拖动迅速找到问题点，然后通过“右键”快速“同步”定位到左上窗口的 Events 解码界面进行问题分析。下图是上图中的密密麻麻的频繁进出 L1.2 低功耗的一个局部放大。

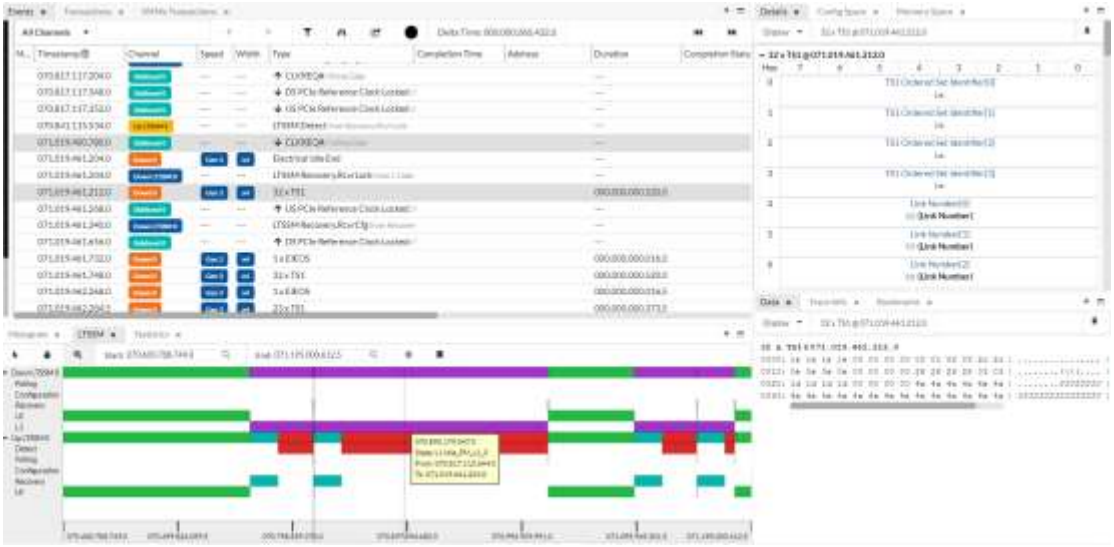


图 2-27

对比：传统分析仪由于 Interposer 以及内部芯片设计问题，可能在 M.2 SSD 进/出低功耗的过程中出现错包，丢包等一些异常问题。

2.3.8 Gen4 “四盘”分析合一以及 Gen5 “八盘”分析合一

SerialTek PCIe4 分析仪提供的 U.X Interposer 实现了 Single Port/Dual Port, U.2/U.3 SSD 四种组合的 NVMe SSD 通过一个 Interposer 即可实现支持，大大方便了企业级客户分析当前以及未来 2.5’ NVMe SSD 的需要，也降低了采用第三方外接卡转换对于 Gen 4 信号带来的风险，同时也间接降低了产品的拥有成本。而且，SerialTek 支持在一个软件界面上同时显示 Dual Port 两个端口的解码（支持在界面上显示/隐藏某一个端口）。参见下图为 dual port 的前面板展示，可以清楚地看到 Link 1 和 Link 2。

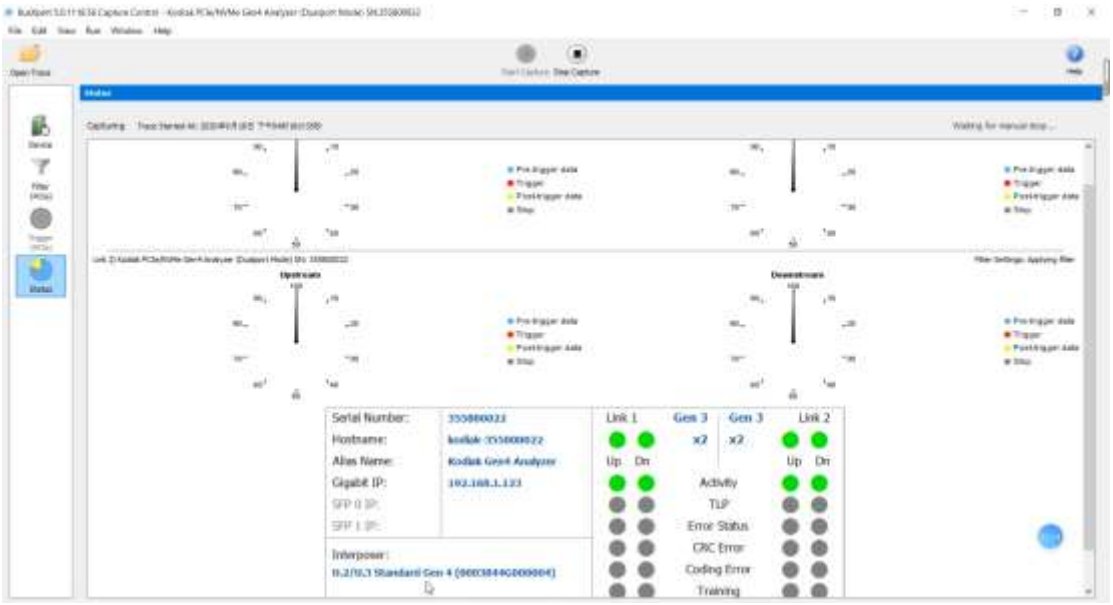


图 2-28

下图是 dual port NVMe SSD 解码界面。

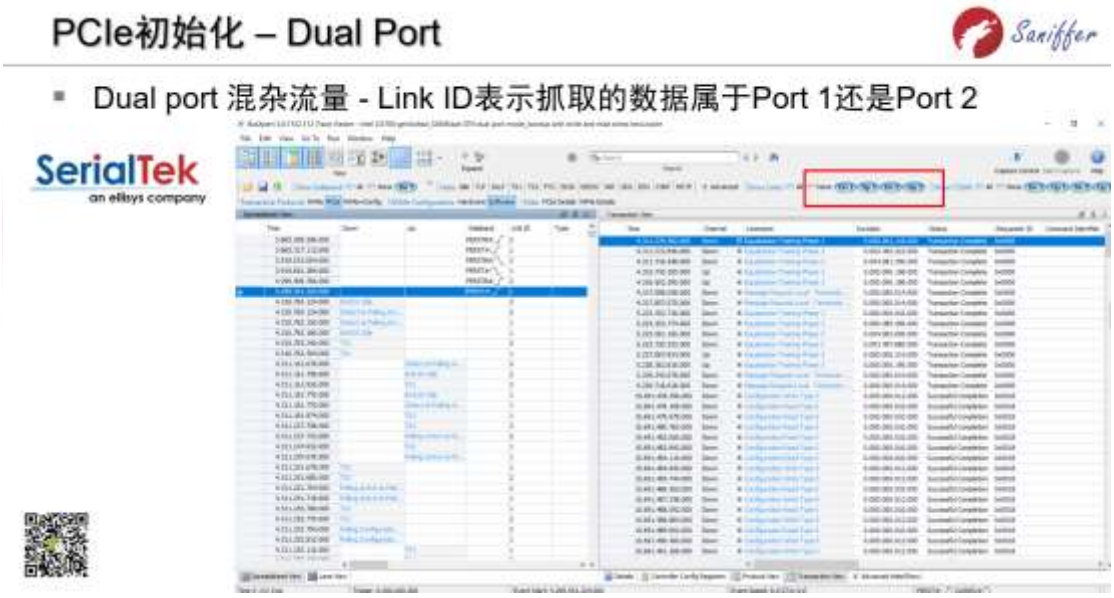


图 2-29

下面是显示过滤掉 dual port 2 (Dn2, Up2 灰色表示 disable display)，仅显示 port 1 的截图。当然也可以反过来通过过滤到 port 1 而仅查看 port 2。

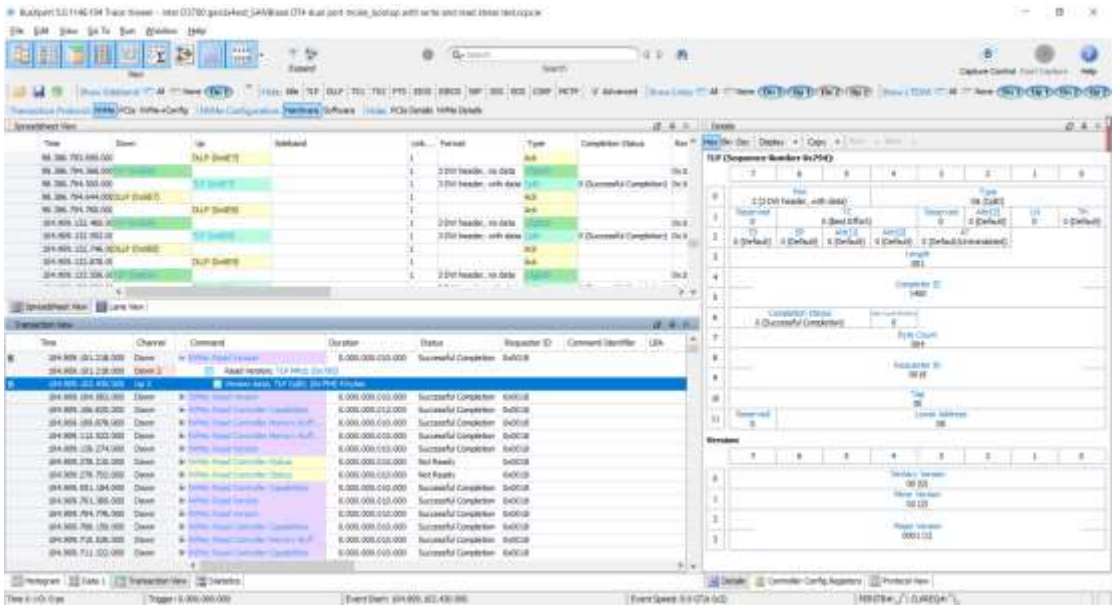


图 2-30

同理，参考下图，如果仅需显示 port 1 的 LTSSM，只要在 LTSSM 视图下面选择 port 1 即可。

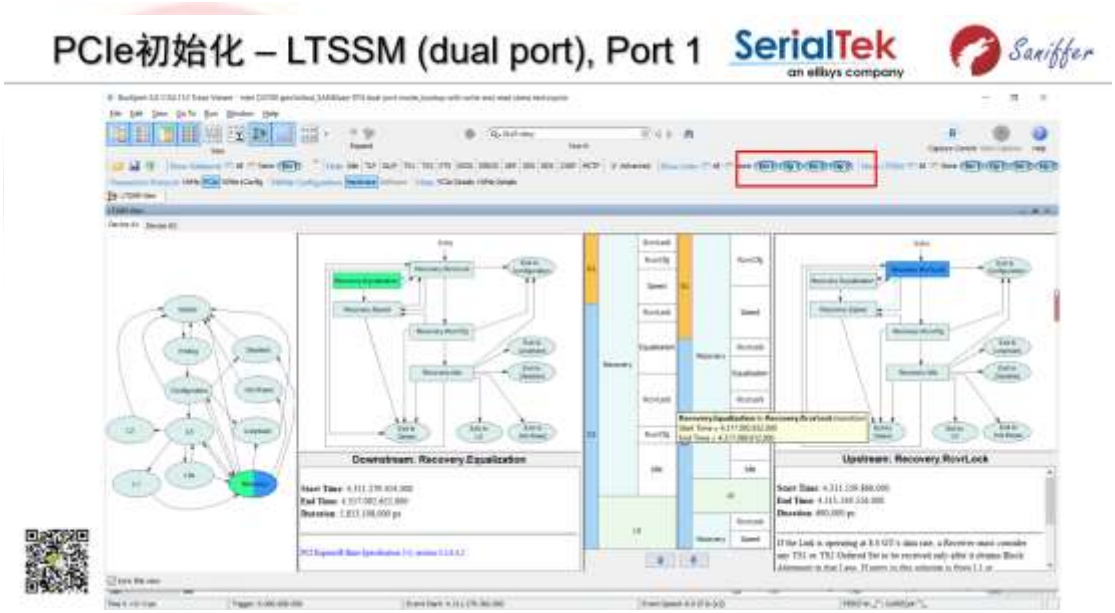


图 2-31

SerialTek Gen5 协议分析仪推出的“八合一 interposer”非常方便各类从事 NVMe SSD 研发的客户使用，这个“八合一 interposer”包括了下面 8 种接口支持：

- **AIC 插卡**
- **M.2**
- **U.2**
- **U.3**
- **E1.S**
- **E1.L**
- **E3.S**

- **E3.L**

上述8种interposer的套装使用非常方便，性价比非常高，对于开发PCIe Gen5 NVMe SSD controller从FPGA仿真验证阶段，到流片后做成PCIe Gen5插卡或者各种接口形态的NVMe SSD进行验证，甚至对比分析其它厂家的不同接口的NVMe SSD，购买该套装PCIe Gen5分析仪将一步到位。

而且，SerialTek支持在一个软件界面上同时显示Dual Port两个端口的解码（支持在界面上显示/隐藏某一个端口）。

对比：对于传统Gen4分析仪无法作到下面两点中的至少一点：1）无法通过一台设备分析Dual Port，需要购买两台PCIe Gen 4/5/6分析仪通过复杂的堆叠技术实现每台分析仪抓取其中一个Port；2） 不支持U.3，需要外配第三方U.2/U.3转换卡。
对于传统Gen5分析仪，如果要购买7种最常见的Gen5 SSD interposer的成本非常昂贵并且不方便，那么采用SerialTek Gen5 Pod的interposer一次性解决问题。

2.3.9 “远程分析”和“远程协作”

SerialTek分析仪通过内部的CPU进行解码分析，用户可以认为客户端协议分析软件只是接收分析仪传过来的画面，类似于使用Teamviewer或者微软的远程桌面一样。当前，美国受“疫情”影响下的员工远程使用SerialTek PCIe Gen 4/5/6分析仪已经成为常态，只要让实验室的同事搭建好测试环境，测试工程师在家通过Cisco VPN或者其它类似软件登陆公司内网后，可以直接在家里电脑上连接，锁定，配置分析仪，然后抓取数据，解码分析的速度和在办公室本地操作一样，几乎没有任何影响。

另外，抓取到数据后，工程师可以邀请其它站点的同事直接连接到该分析仪打开buffer或者存储在分析仪内部的Trace文件系统分析，该功能对于跨国公司分析PCIe问题非常便利，无需再在不同站点之间来回搬运Trace文件。

对比：传统的PCIe协议分析仪必须在本地使用，因为它必须通过USB/ETH将Trace文件导出到本地电脑进行分析，效率非常低下。

2.3.10 支持Web管理，免除升级带来的混乱

使用过协议分析仪的工程师知道，由于协议分析仪经常需要升级内部的FPGA来接解决一些bug或者增加某些功能，这会导致一些使用上很大的混乱状况。举例，某工程师了解到最新的软件解决了某个问题或者增加了他希望使用的功能，他可能下载了“协议分析仪客户端软件”安装包，然后通过网络将分析仪内部的FPGA固件升级到了匹配的版本，但是如果公司里面还有其它工程师后面也在使用该分析仪，那么当该工程师打开软连接分析仪的时候就提示分析仪内部FPGA firmware版本和他当前的客户端版本不匹配，是否需要刷新到匹配版本，这样就相当于降级了。如果再换回到前面的工程师使用，那么他有会无法使用， 必须再次刷新到新版本。反复降级、升级FPGA firmware是隐藏着很大的风险的，因为万一将firmware删除后在刷新过程中断电分析仪就需要回到原厂返修。

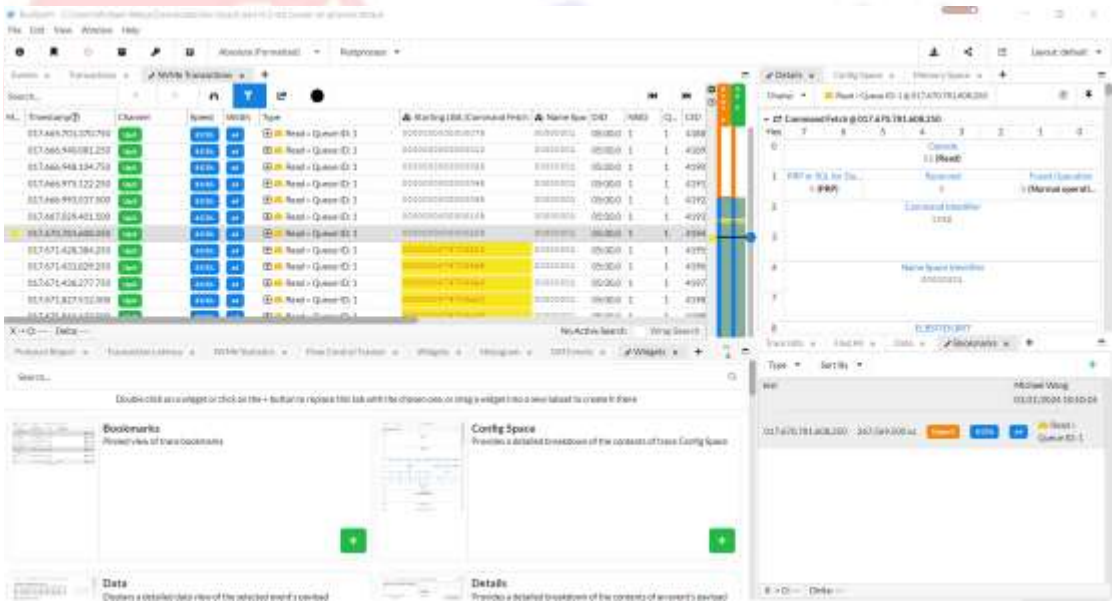
SerialTek 创新性的使用 Web 管理界面就避免了上面的问题，即，一个工程师刷新了最新的 FPGA firmware 以后，因为其他工程师都是使用 Web 来管理、配置、使用该分析仪就无所谓反复降级、升级的事情发生了。

2.3.11 “随时断网”

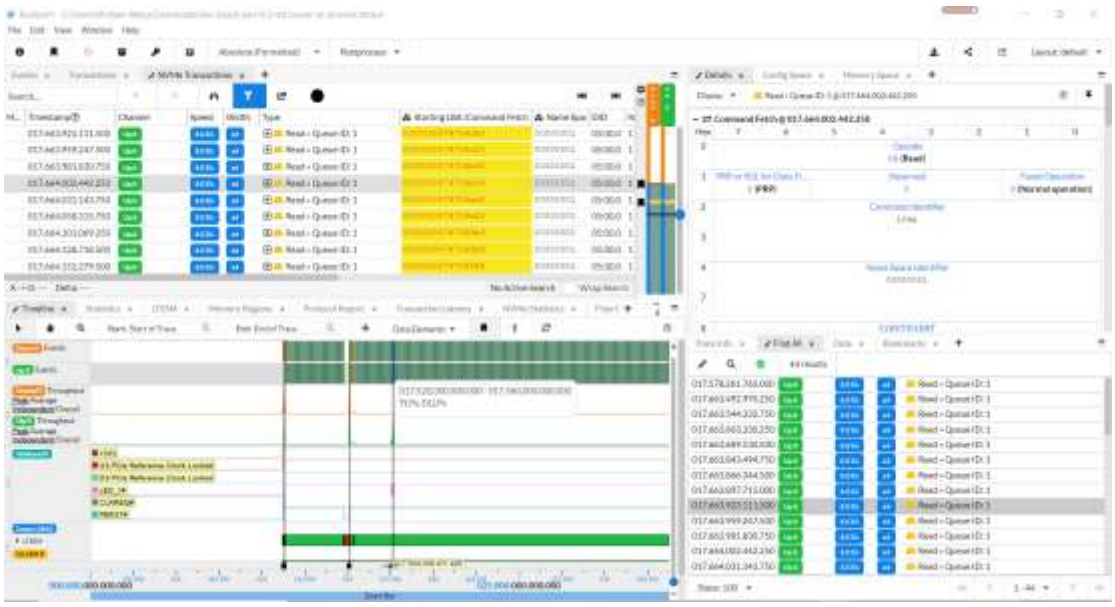
由于 SerialTek 采用高性能主机设计，里面使用标准 Linux 系统而不是精简 Linux，所以可以完整保持网络连接的状态。工程师在使用 SerialTek Gen 4 分析仪的时候再也不用担心万一网线碰掉导致抓取的数据无法读取的问题了。工程师可以随时断开网线，然后接上网线。这对于很多负责客户支持的工程师非常有帮助，在客户现场设置好分析仪开始抓取数据以后，即可合上笔记本离开而不用担心第二天过来无法连接分析仪。

对比：传统的 PCIe 分析仪通过 USB/ETH 不能断开连接，一旦断开（例如笔记本休眠后重启打开），那么即便分析仪已经抓到需要的数据了，该数据再也无法读取。工程师必须重新启动分析仪重新连接电脑进行抓取，这个抓取过程必须保证电脑和分析仪实时在线。

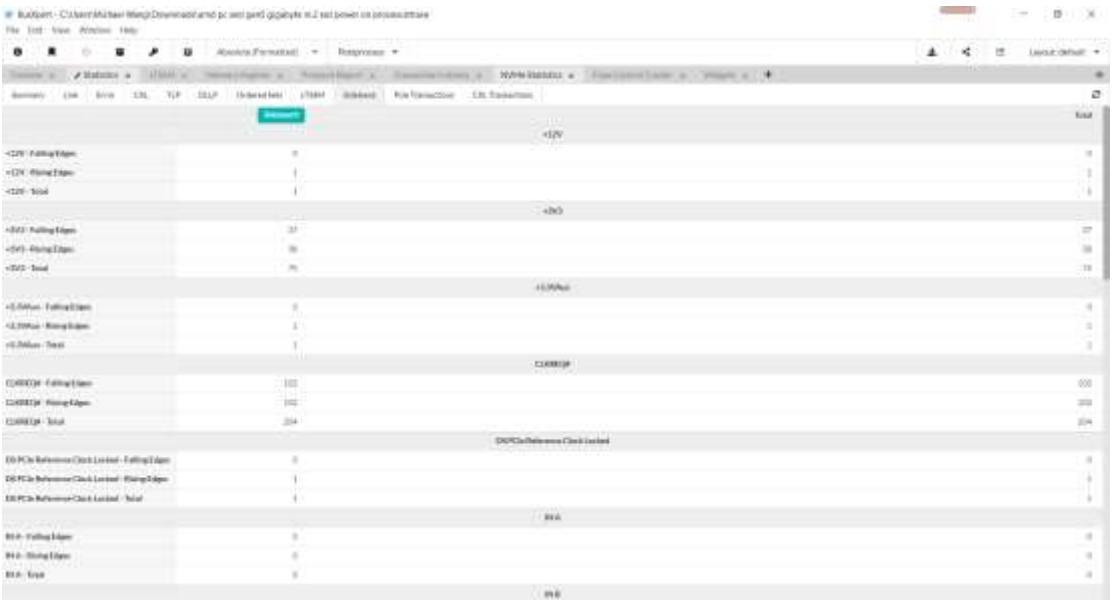
2.3.12 基于 Widget 小工具提供的高级分析功能

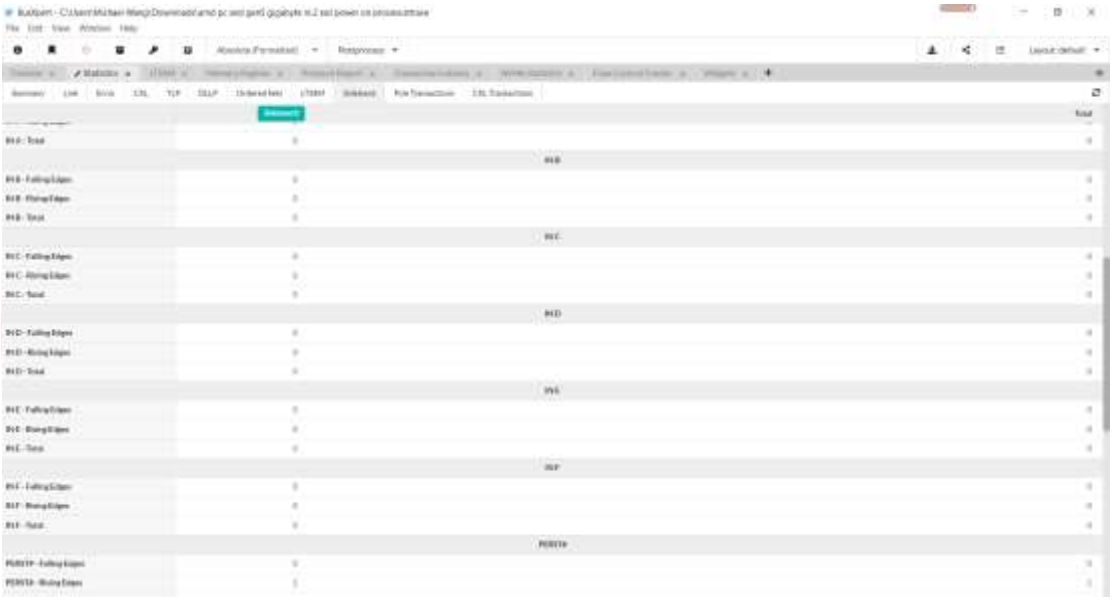


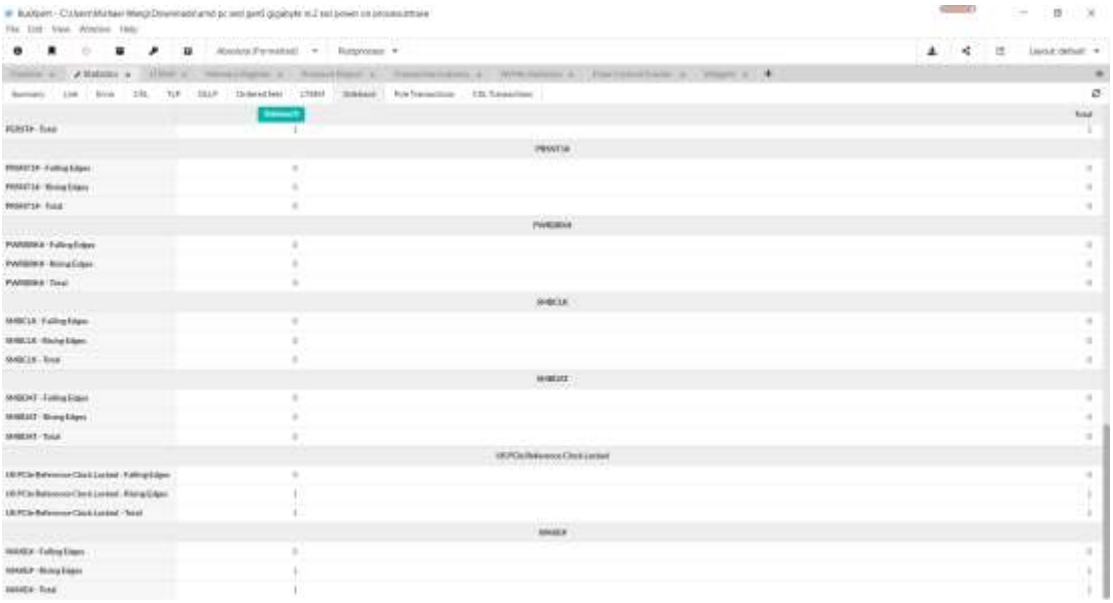
2.3.12.1 基于时间轴的总线活动一览



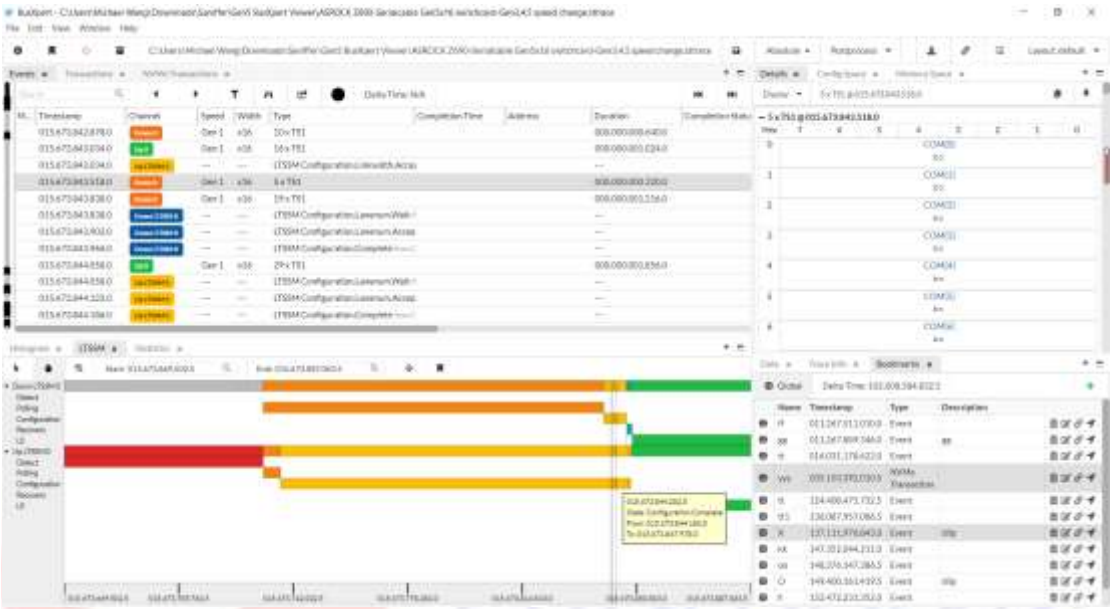
2.3.12.2 PCIe/CXL/LTSSM 统计分析功能



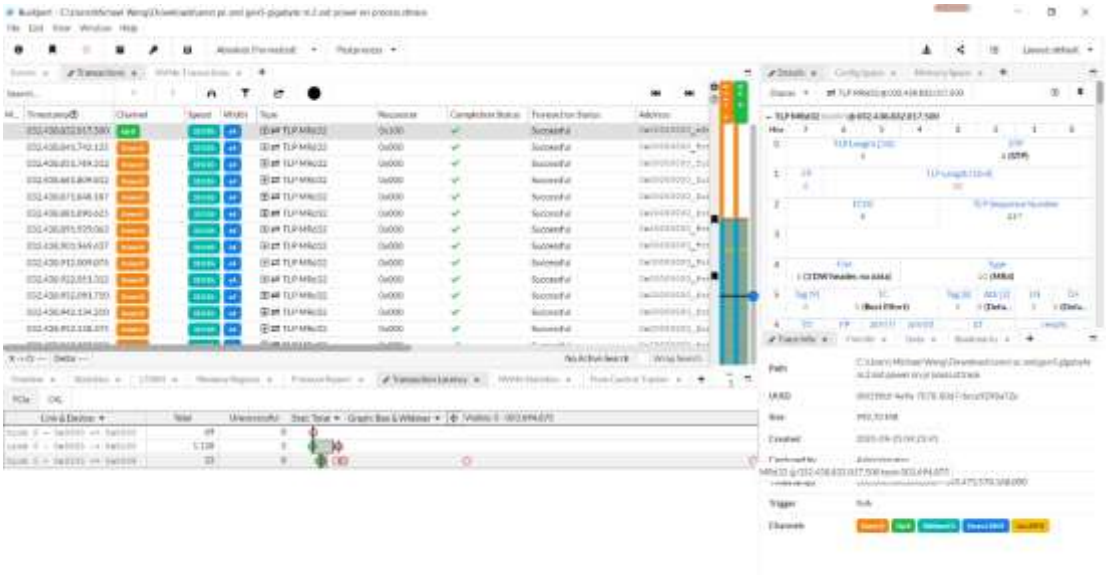


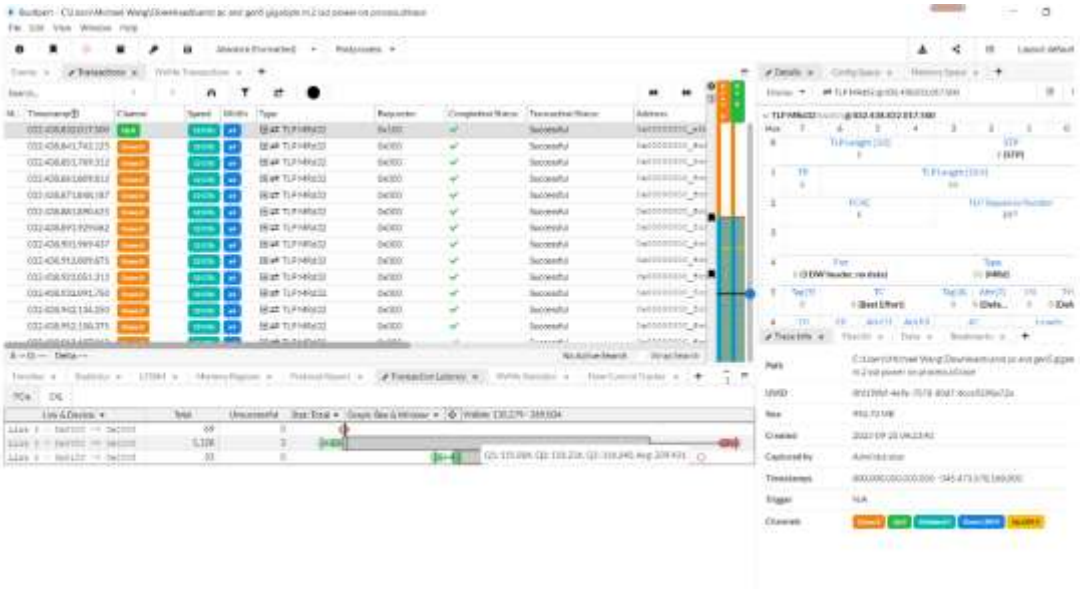


2.3.12.3 基于时间轴的 LTSSM 链路状态机分析

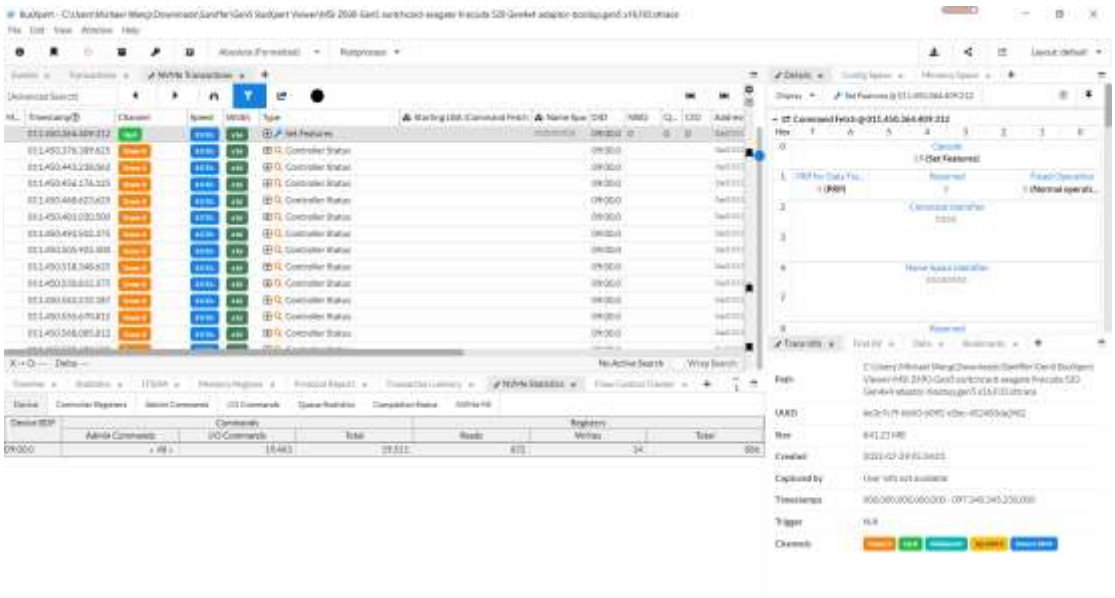


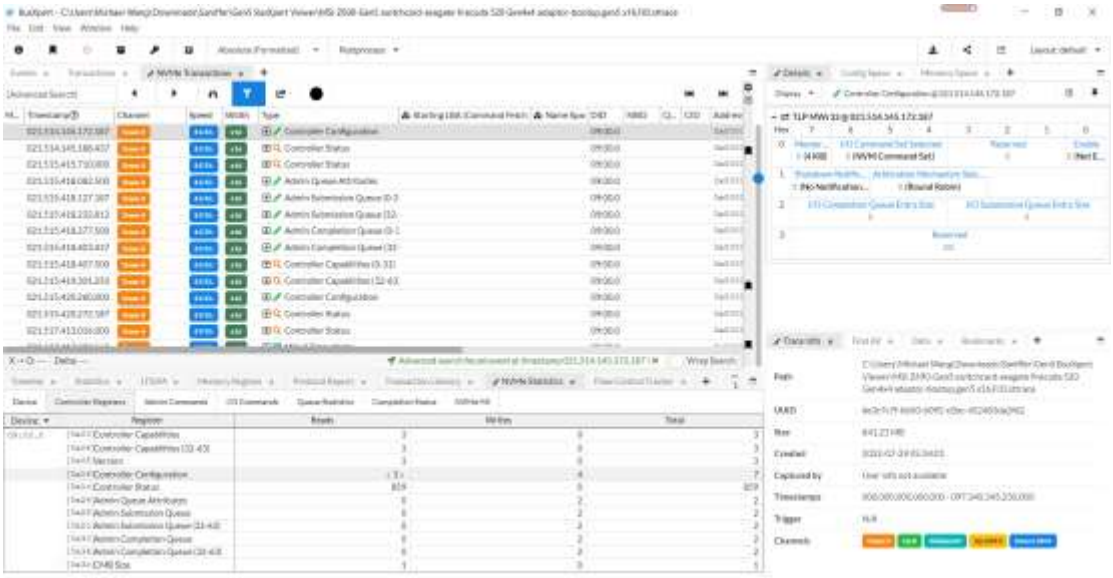
2.3.12.4 TLP 响应延迟分析

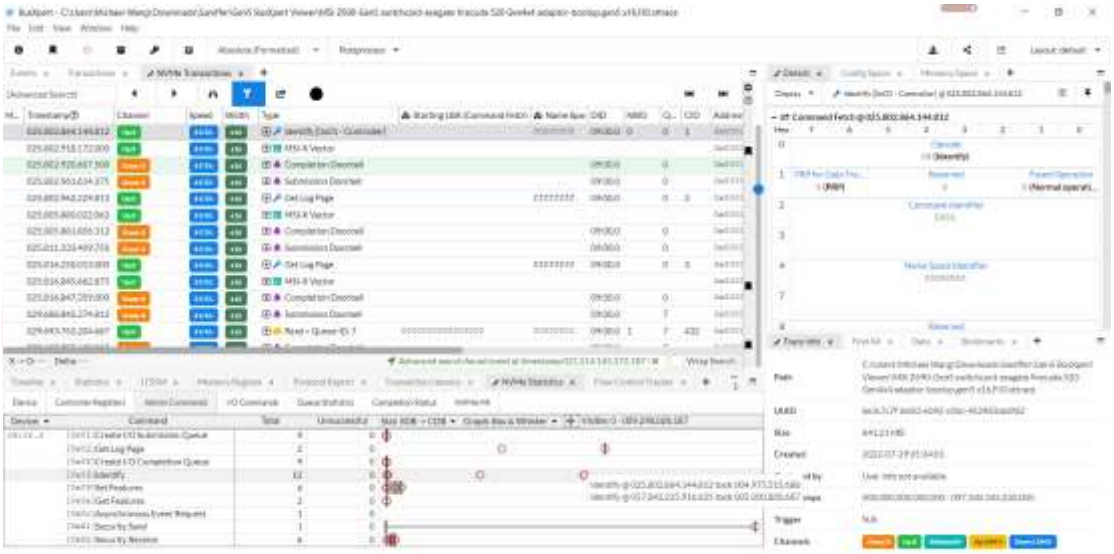


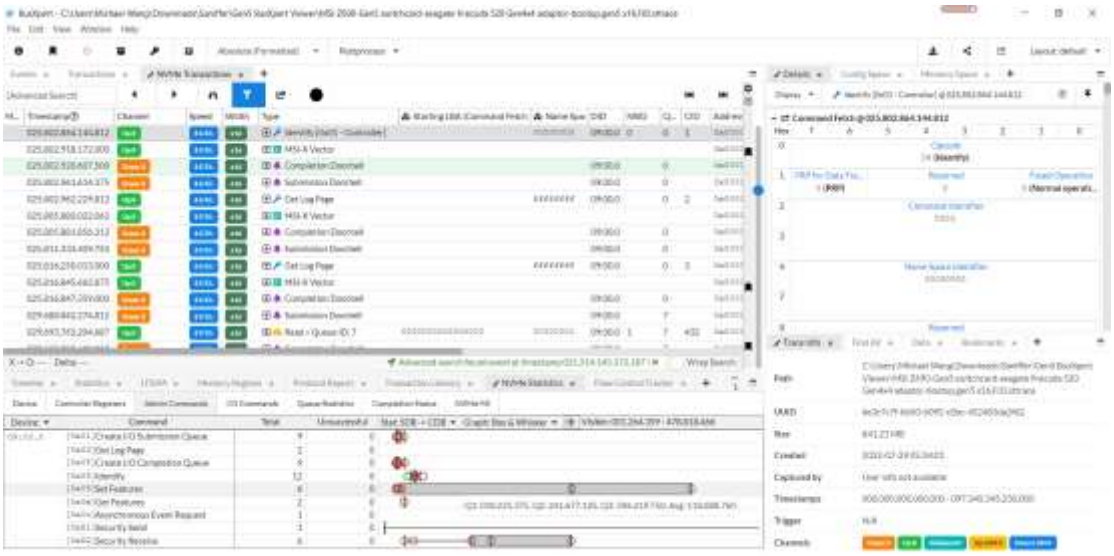


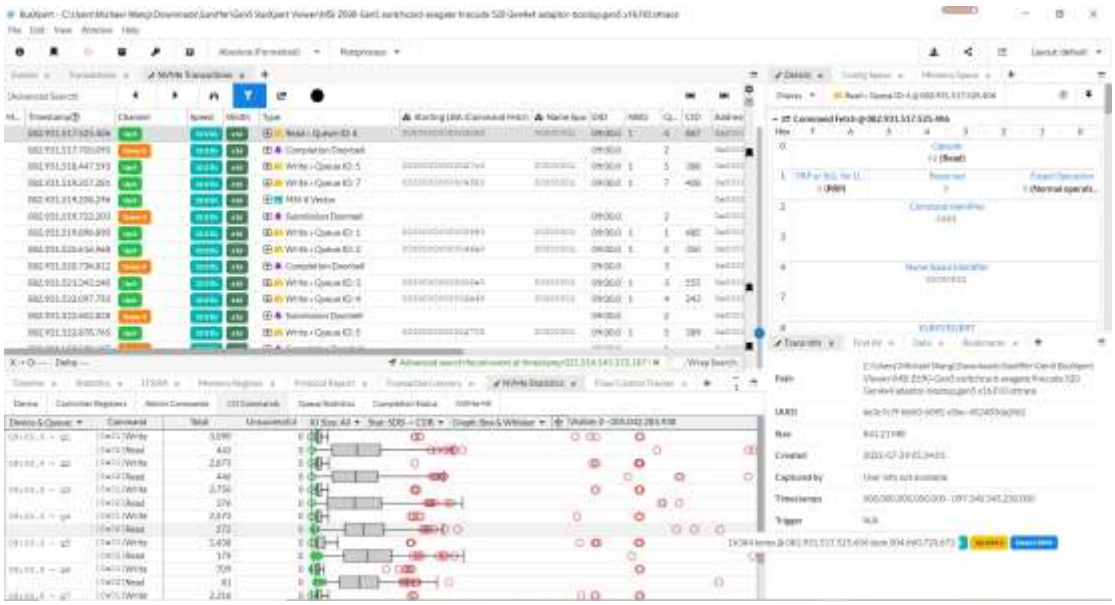
2.3.12.5 NVMe 统计和延迟分析

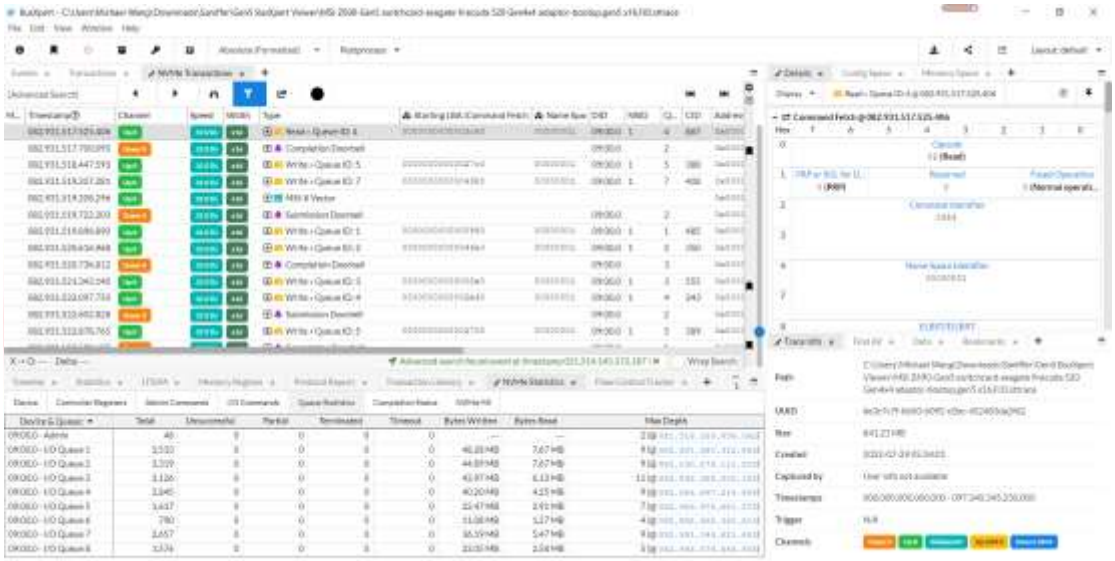


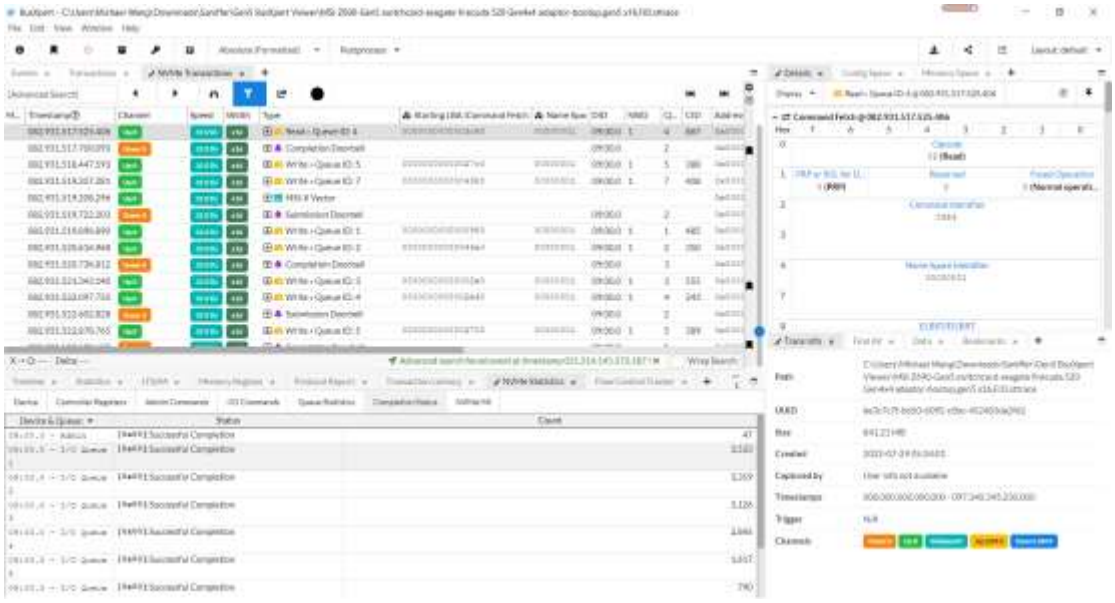


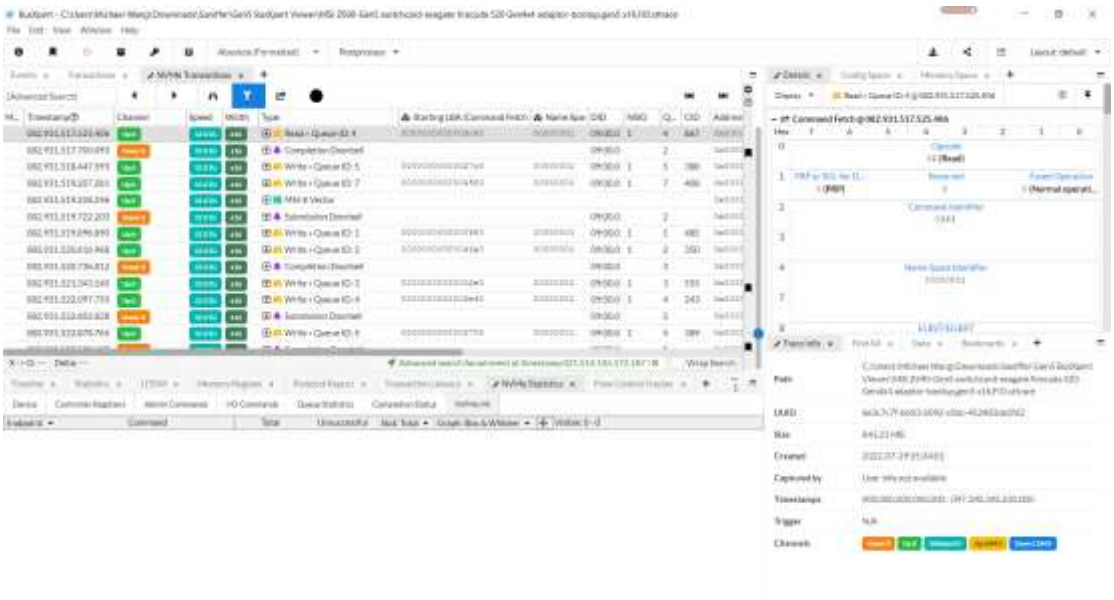




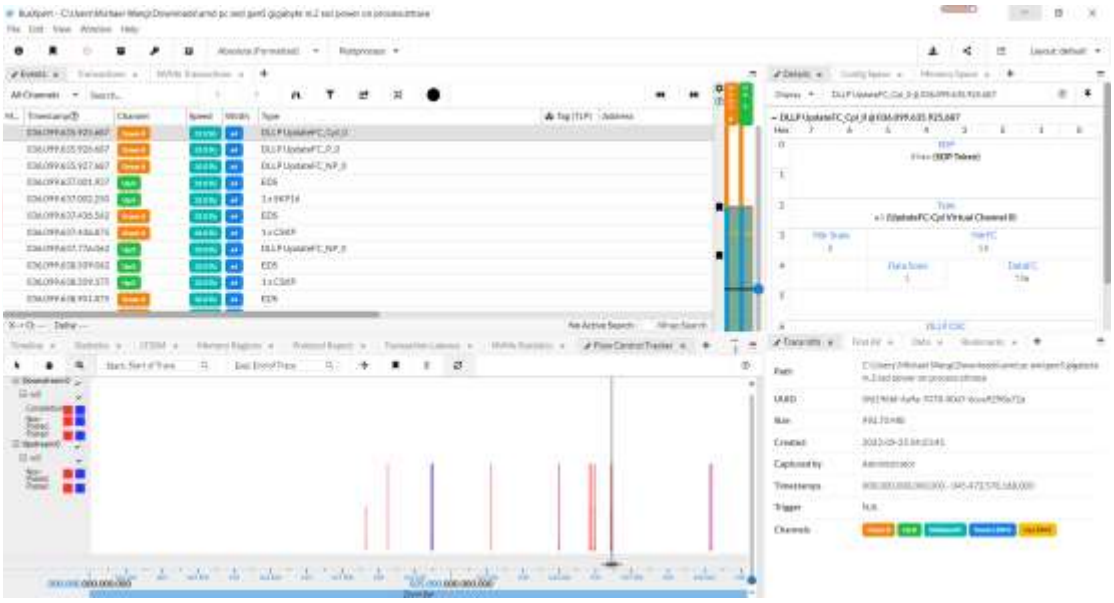




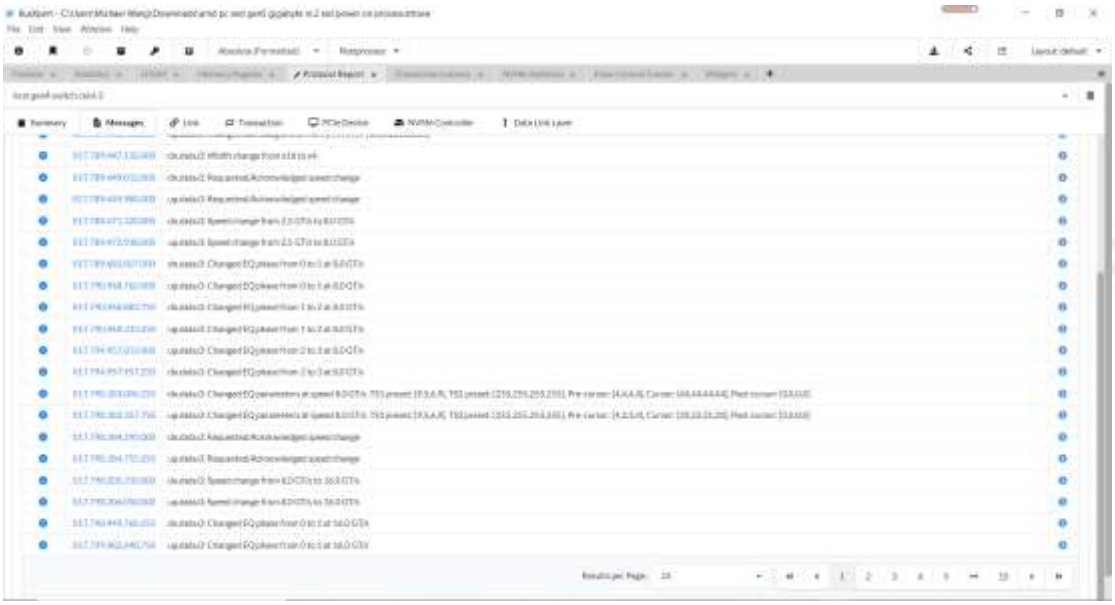


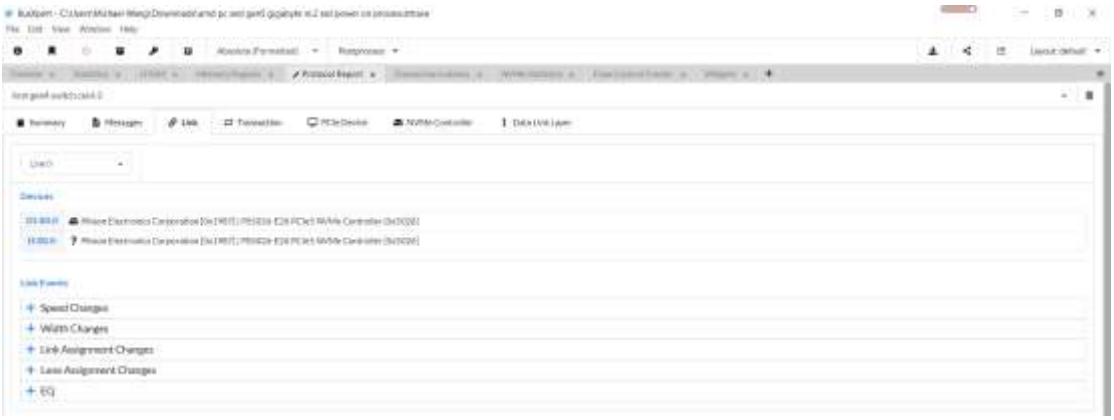


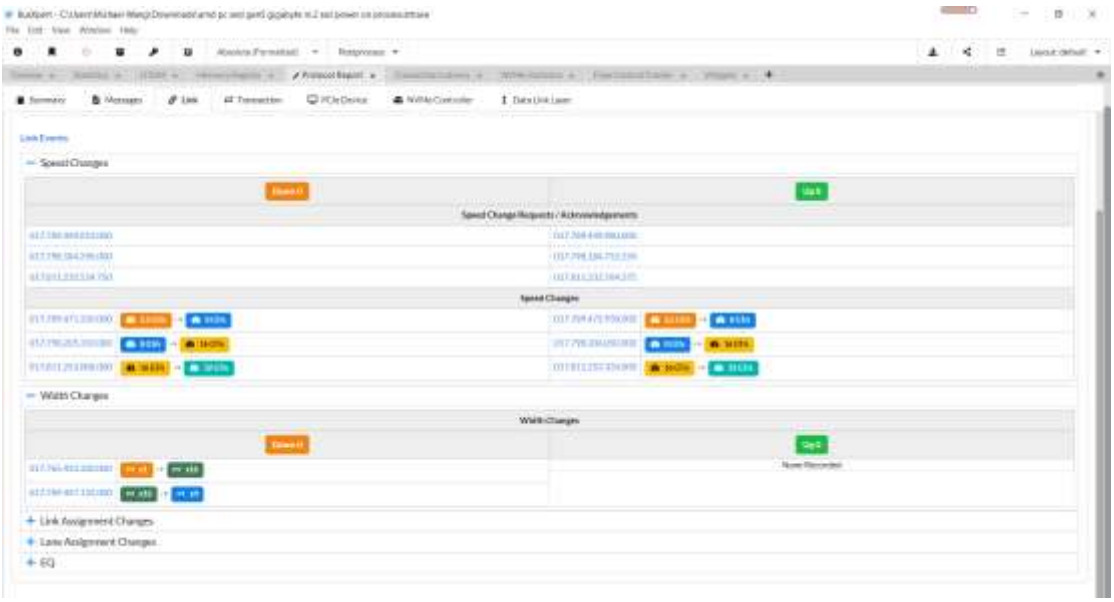
2.3.12.6 Flow Control 流控分析

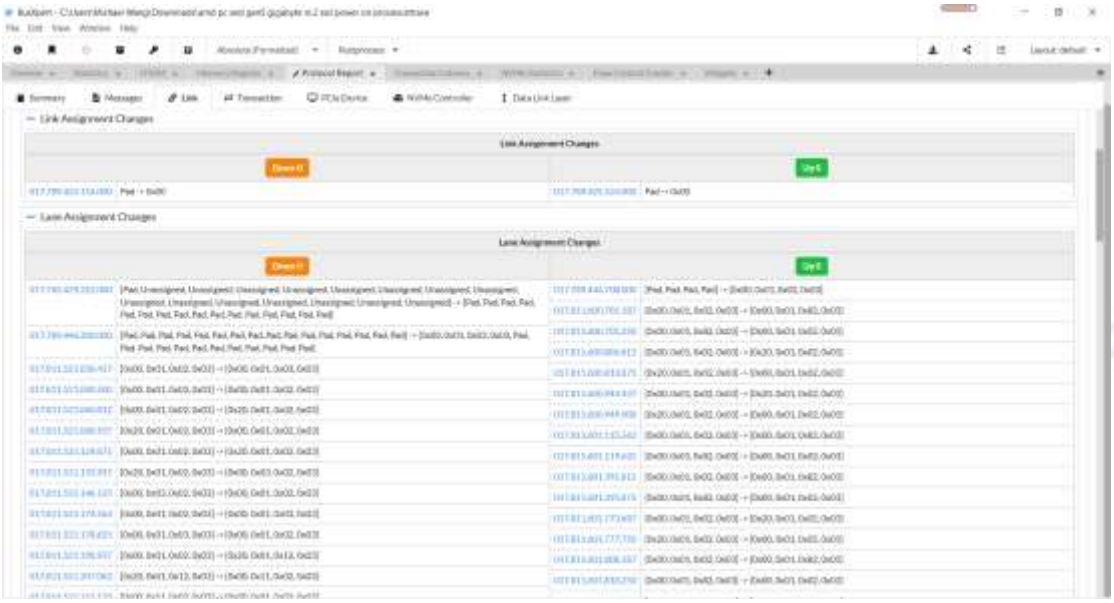


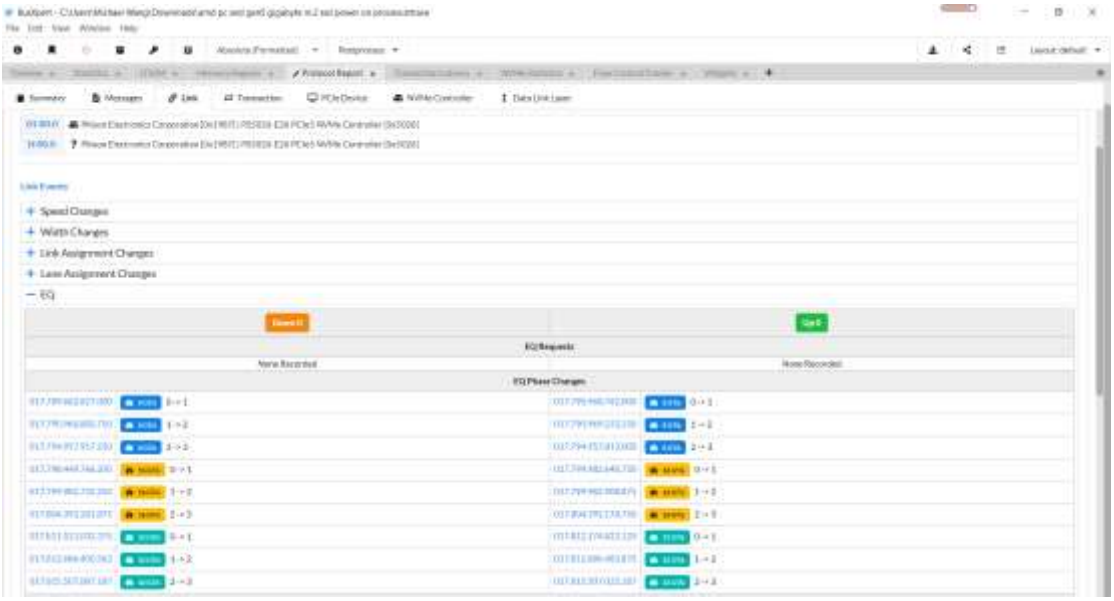
2.3.12.7 协议报告 Protocol Report

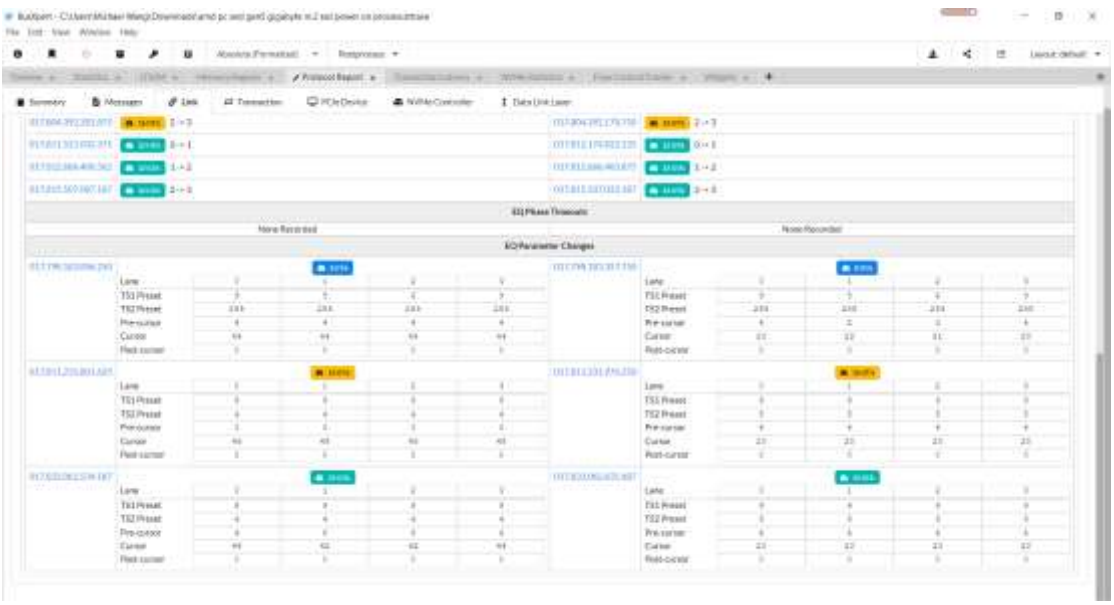




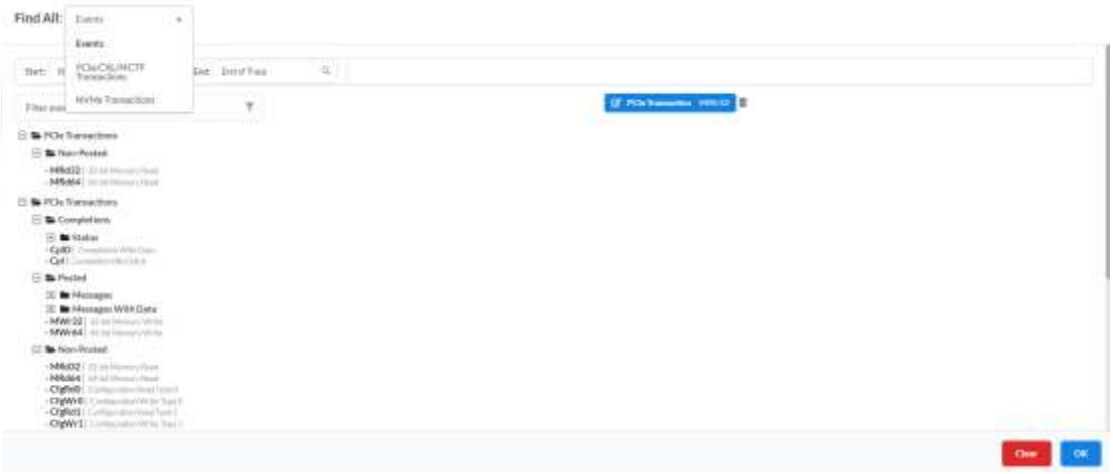

Link Events
Speed Changes
Speed Change Requests / Acknowledgements
Speed Changes
Width Changes
Width Changes
Link Assignment Changes
Lane Assignment Changes
EQ

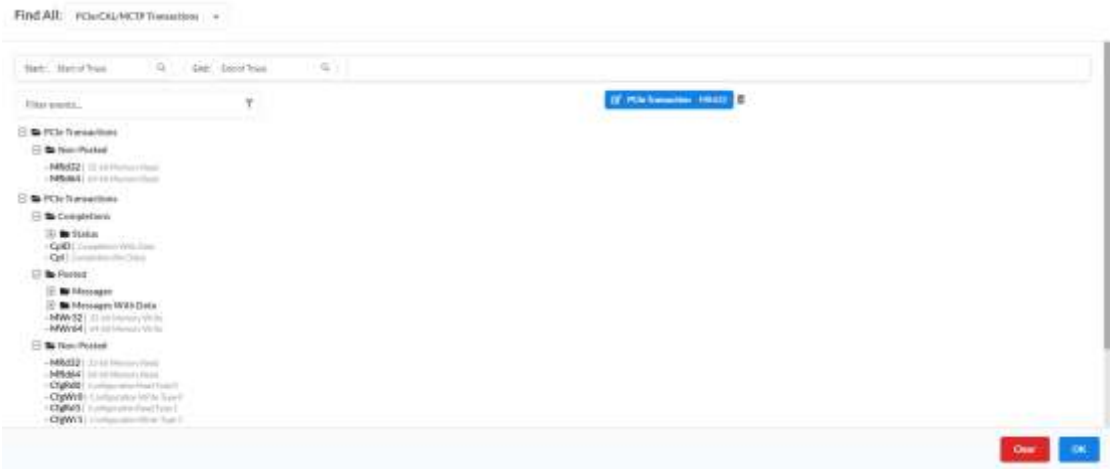

Link Assignment Changes
Link Assignment Changes
Lane Assignment Changes
Lane Assignment Changes

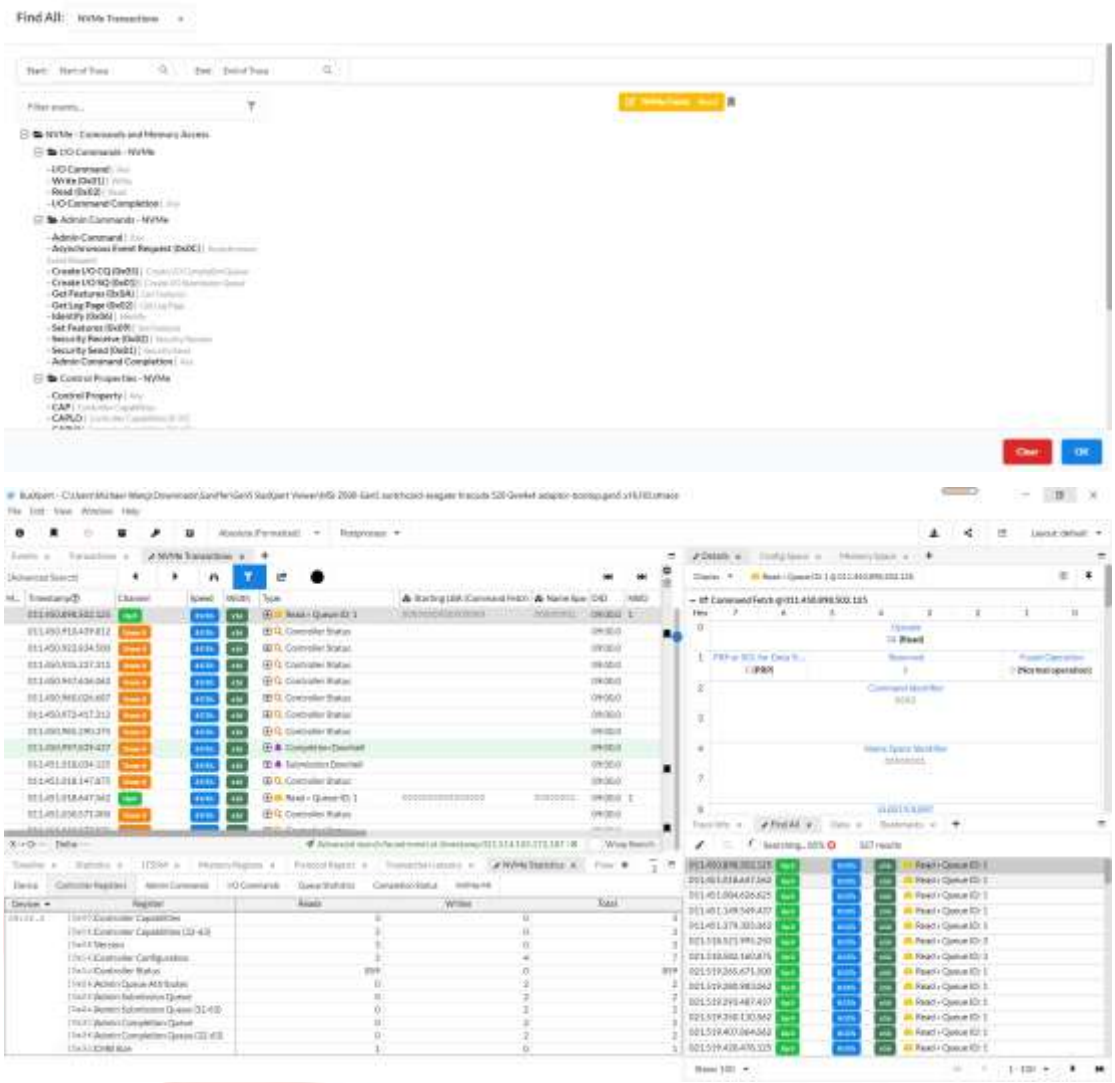

Link Events
Speed Changes
Width Changes
Link Assignment Changes
Lane Assignment Changes
EQ
EQ Requests
EQ Phase Changes



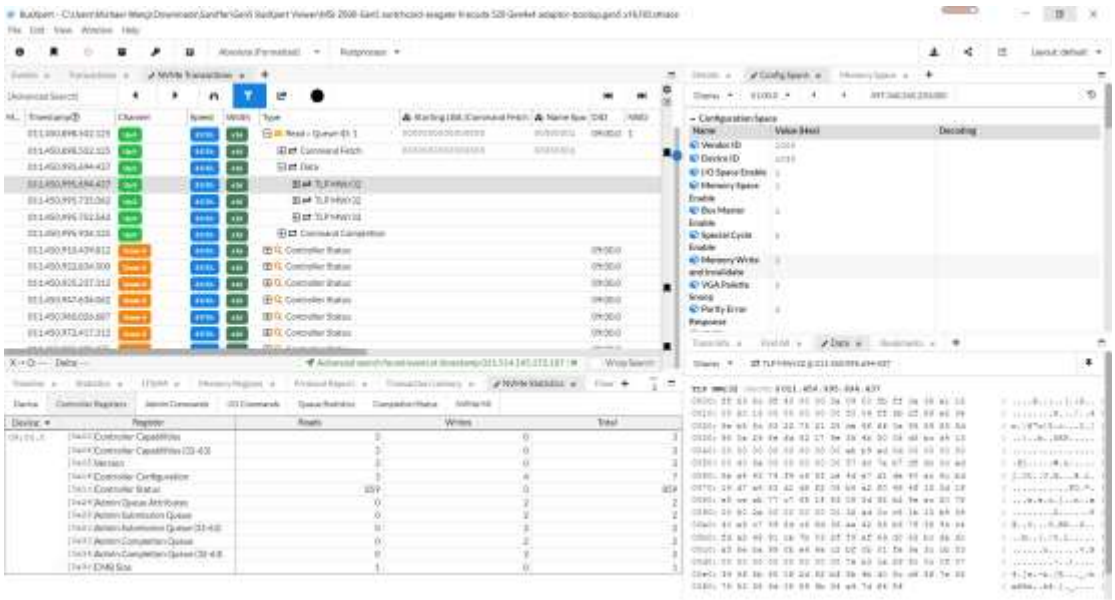
2.3.12.8 查找所有特定的 Packet 功能

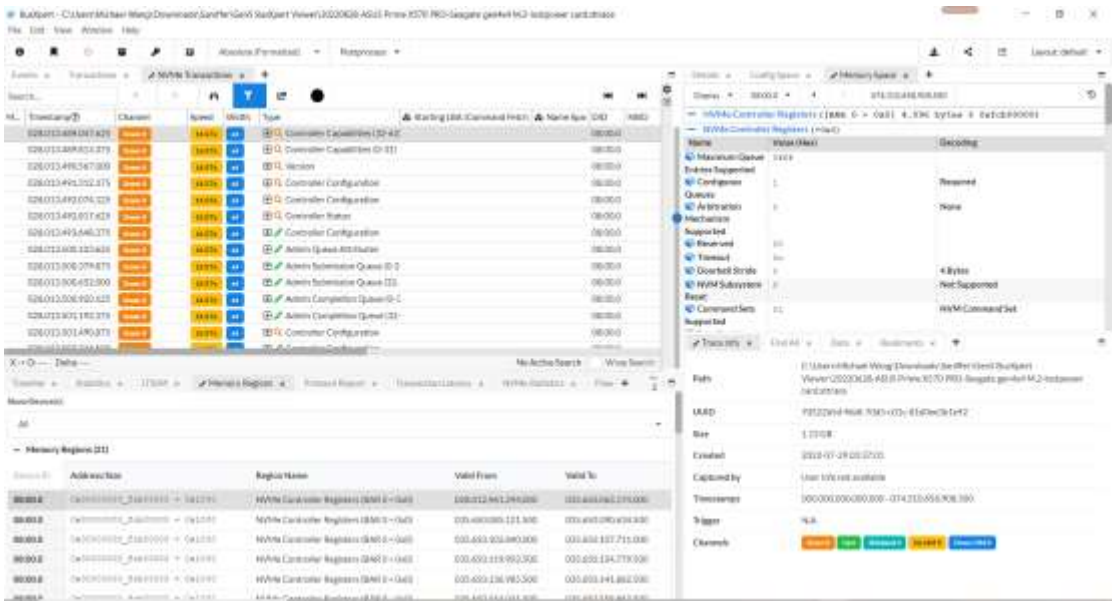




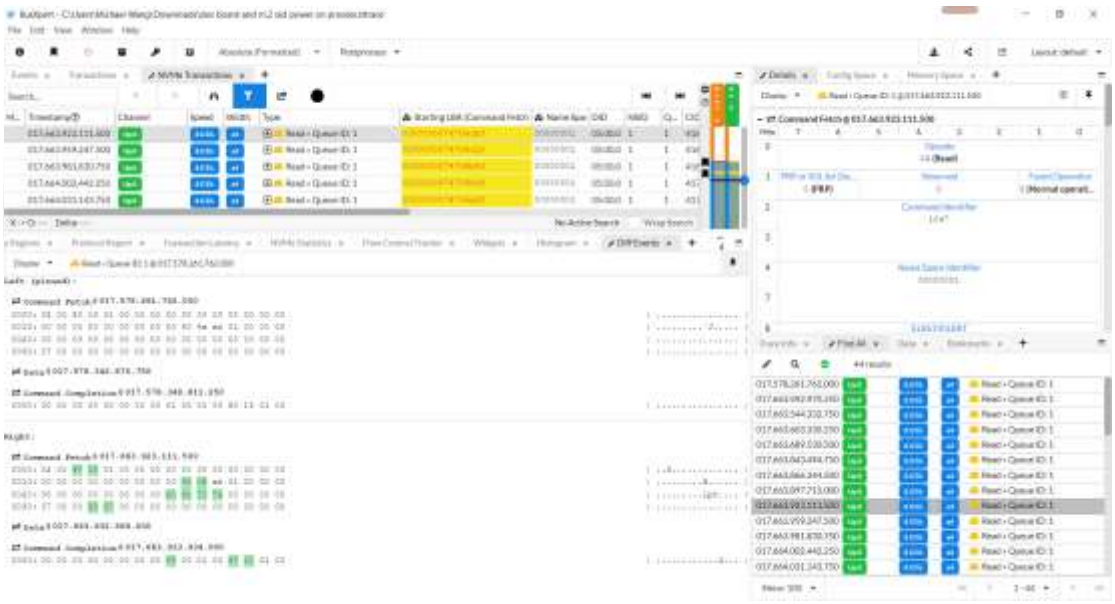


2.3.12.9 Config space/Memory space/Memory Region

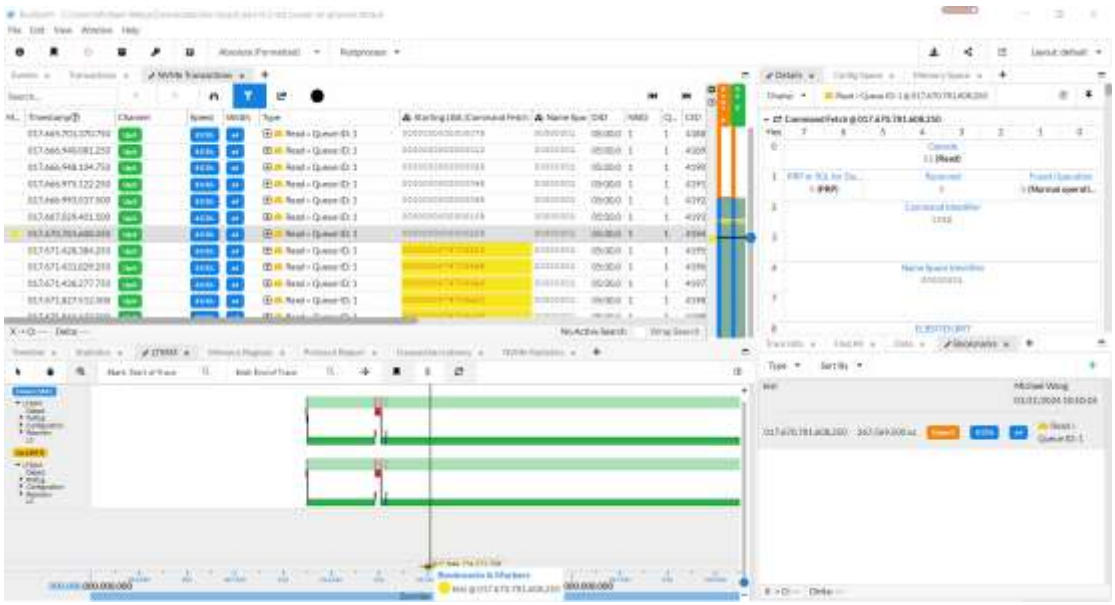


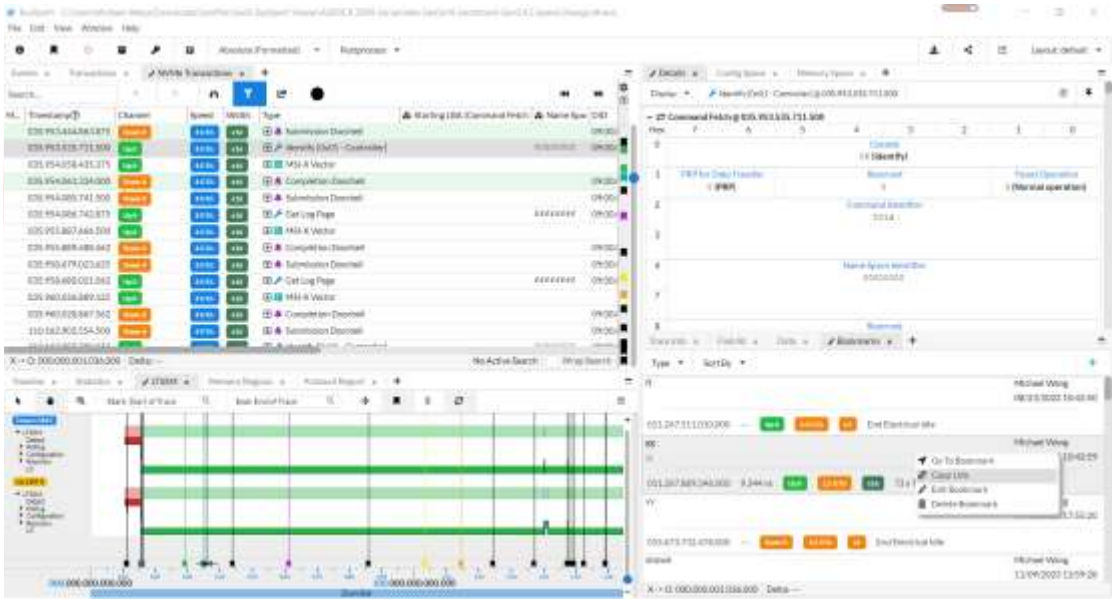


2.3.12.10 比较两个位置的异同

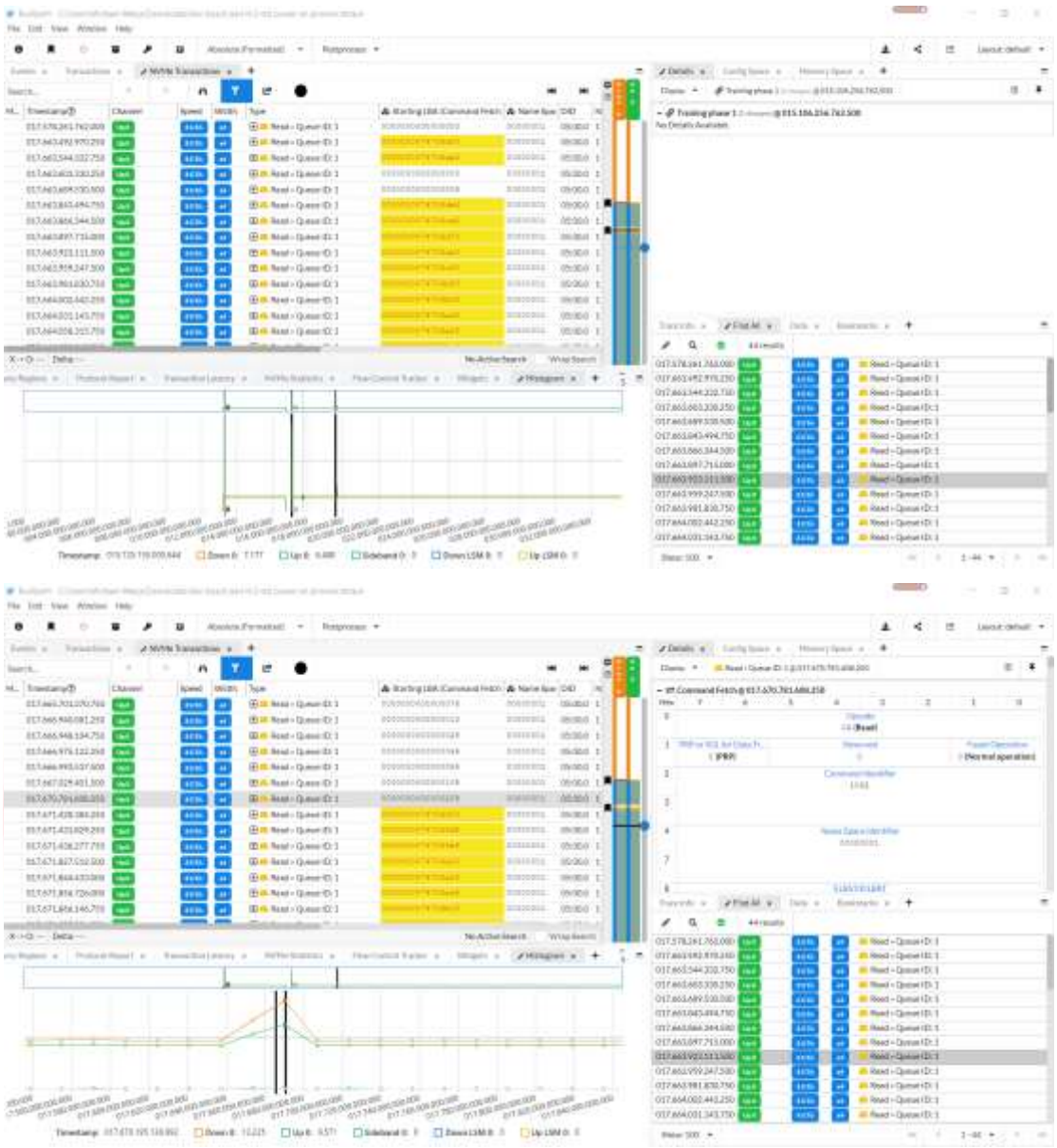


2.3.12.1 Bookmark 书签管理



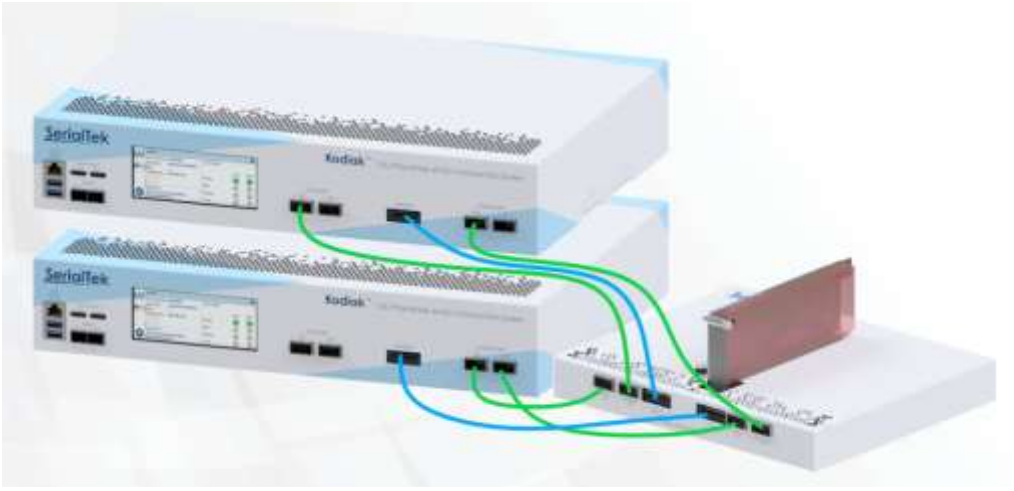


2.3.12.1 流量直方图一览

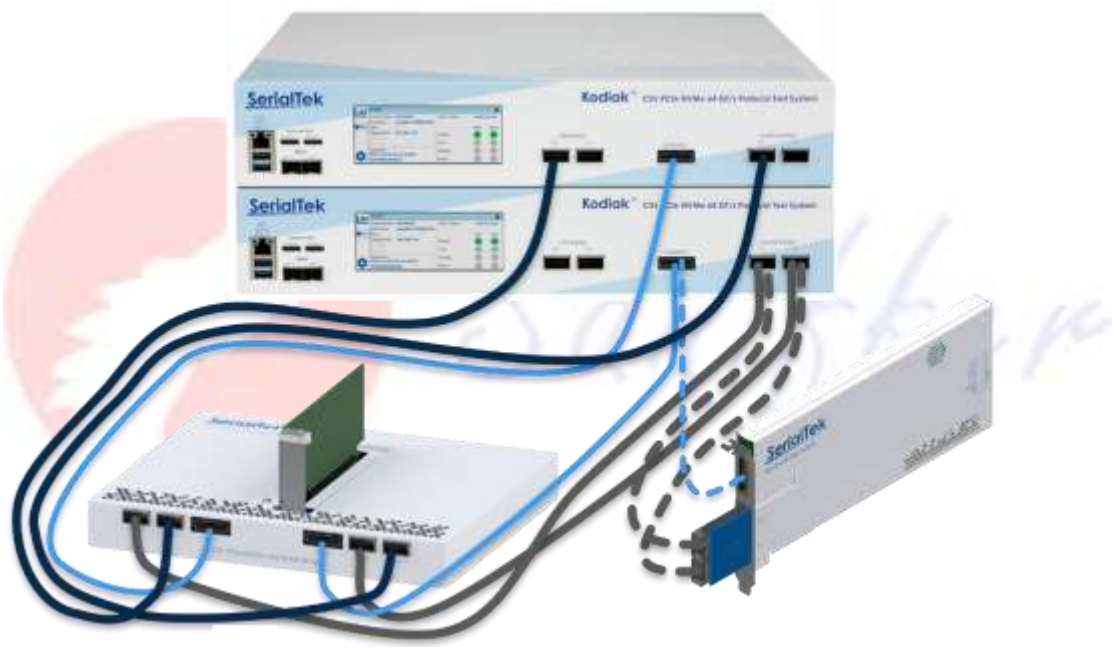


2.4 SerialTek PCIe Gen6/CXL 3.0 协议测试系统

2.4.1 SerialTek 最先进的 Kodiak 系列 PCIe Gen6/CXL 3.0 协议分析和训练器架构



下图是 host smart fixture 和 device smart fixture 如何连接 analyzer 的连接拓扑图。



PCIe 协议训练器（PCIe protocol exerciser）是一种用于测试和验证 PCI Express（PCIe）总线协议的设备。它可以实现以下功能：

1. 仿真测试：模拟 PCIe 设备与主机之间的通信，以验证硬件和软件的兼容性。

2. 错误注入：通过注入错误来测试系统的容错能力，例如丢失数据包、错误的传输速率等。

3. 性能评估：测量 PCIe 总线的吞吐量、延迟等性能指标，以优化系统设计。

4. 流量生成：生成不同类型的数据流量，以测试系统对各种负载的响应能力。

5. 协议分析：捕获和分析 PCIe 总线上的数据传输，以便排查问题和优化性能。

6. 配置寄存器访问：模拟配置空间中的寄存器访问，以测试设备的配置和控制功能。

总的来说，PCIe 协议训练器是一种用于测试、验证和优化 PCIe 系统的强大工具，可以帮助开发人员和工程师确保其系统在实际部署中的稳定性和性能。

2.4.1.1 体验无与伦比的协议分析功能

SerialTek 公司推出的 Kodiak 系列 PCIe 协议分析和训练设备拥有业内最先进的架构，彻底改变了协议分析仪的标准。凭借其创新设计和全面的功能集，Kodiak 使 CXL、PCIe 和 NVMe 开发人员能够对复杂的 PCI Express 设计拥有无与伦比的可视性，并能够对 PCIe、CXL 和 NVMe 协议进行高效分析。

2.4.1.2 基于 Web 浏览器的高级 BusXpert™ 应用程序

通过利用网络浏览器的熟悉性和可访问性，Kodiak 简化了用户体验并提高了工作效率。Kodiak 配备了先进的 BusXpert™ 应用程序，可通过 Web 浏览器界面或相同的独立软件应用程序进行访问。这种直观的应用程序允许用户访问强大的分析工具、配置分析仪设置并方便地监控设备状态。

2.4.1.3 强大的触发器、过滤器和 Trace 处理套件

Kodiak 提供一整套 PCIe、CXL 和 NVMe 触发器、过滤器和强大的 Trace 处理功能。这些先进的工具为用户提供了对数据采集的精确控制，使他们能够捕获和分析感兴趣的特定事件。

2.4.1.4 灵活的 Trace 存储

Kodiak 提供高度灵活的 Trace 存储解决方案，以满足用户的多样化需求。我们的 Trace 存储旨在提供 Trace 文件的轻松访问和共享。用户可以将 Trace 直接保存到 Kodiak 的内部 SSD 存储中，无需将大型 Trace 文件下载到本地 PC。保存后，可以在 Kodiak 中轻松打开 Trace 或将其下载到客户端以进行进一步分析。为了促进高效的数据管理，在下载之前可以在 Kodiak 中方便地压缩该 trace 文件。

凭借高达 8TB 的内部 SSD 存储，用户可以存储大量 Trace 数据以供分析和参考。通过两个 USB 3.1 端口和两个 PCIe OCuLink 端口可选择外部存储，存储容量可进一步扩展。

2.4.1.5 多用户访问和深度 Trace Buffer

Kodiak 支持多用户访问，允许多人同时协作和处理 Trace。这种协作工作流程简化了团队合作并加快了项目进度。此外，Kodiak 的深度 Trace Buffer 功能可确保保留重要的 Trace 数据，即使在高速数据捕获场景中也是如此。开发人员可以高效地访问和查看 Trace 信息，确保准确的分析、全面的故障排除和无错误的根本原因分析。

2.4.1.6 紧凑便携的设计

Kodiak 的设计考虑到了便利性，采用紧凑便携的设计，并配有硬壳旅行箱，可保护设备投资。其时尚的外形和结构使其便于携带，适合现场测试，使工程师能够随时随地处理协议分析任务。借助 Kodiak，开发人员可以在不同的环境中高效地进行测试和验证，确保 PCIe、CXL 和 NVMe 协议的顺利运行。

2.4.1.7 凭借 Kodiak 系列最先进的架构保持领先地位

投资 Kodiak 使开发人员能够在 PCIe、CXL 和 NVMe 协议的动态世界中保持领先地位。其最先进的架构与用户友好的界面和强大的分析功能相结合，为协议设计的复杂性提供了无与伦比的可见性。在 Kodiak 的陪伴下，释放 PCIe 开发的真正潜力、简化流程、缩短开发周期并加快上市时间，推动您的项目取得成功。

2.4.1.8 强大的 Kodiak 系列功能

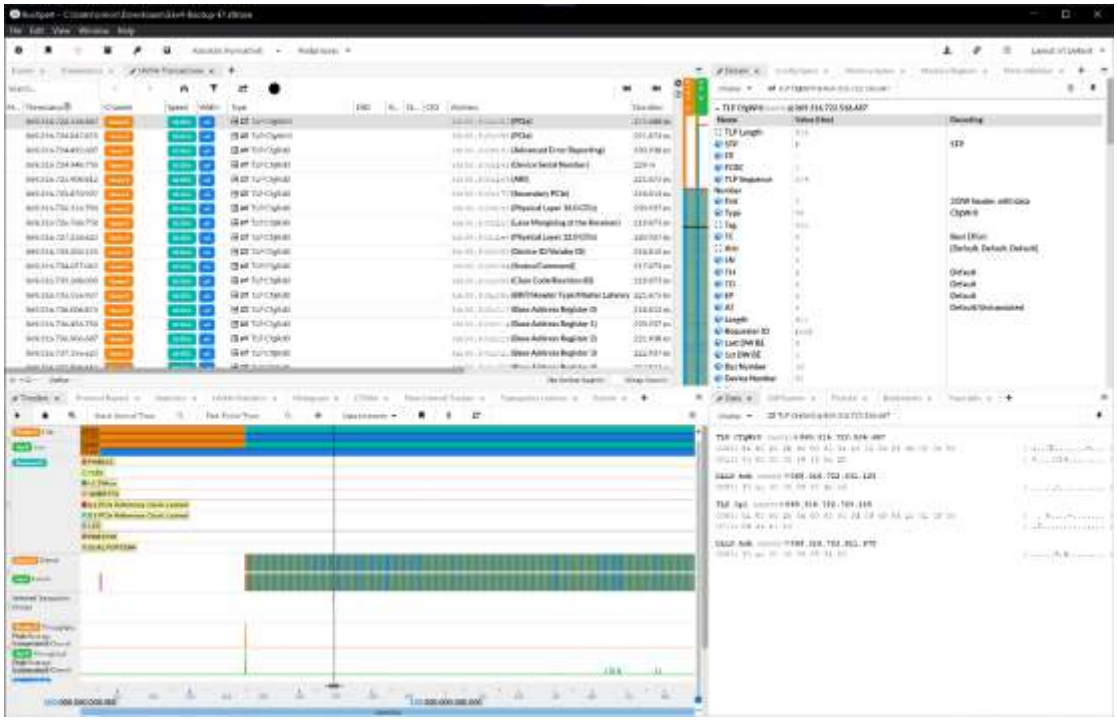
- 支持 CXL、PCIe、NVMe、DOE/IDE、SMBUS 和所有边带信号
- 无需调整（即无需校准）。Kodiak 的 Rx 自动快速均衡 (EQ) 所有数据速率下的 PCIe 信号
- 嵌入式 Trace 处理架构和加速性能
- 深度 Trace Buffer 区
  - 72GB、144GB 或 288GB
- 内部 Trace 存储 (SSD)
  - 2、4 或 8TB
- 直连存储
  - 两个 OCuLink (PCIe) 端口
  - 两个 USB 3.1 端口
- 网络和直接连接
  - 两个 10GbE SFP+（光纤/铜质）
  - 1 个 1GbE RJ-45
- 在一个平台上支持单端口 (1×4) 和双端口 (2×2) 分析

▪内存中的实时 Trace（保存之前）。用户可以查看和分析捕获的 Trace，而无需将 Trace 下载到客户端 PC

▪用于分析仪设置、控制和状态的触摸屏 LCD

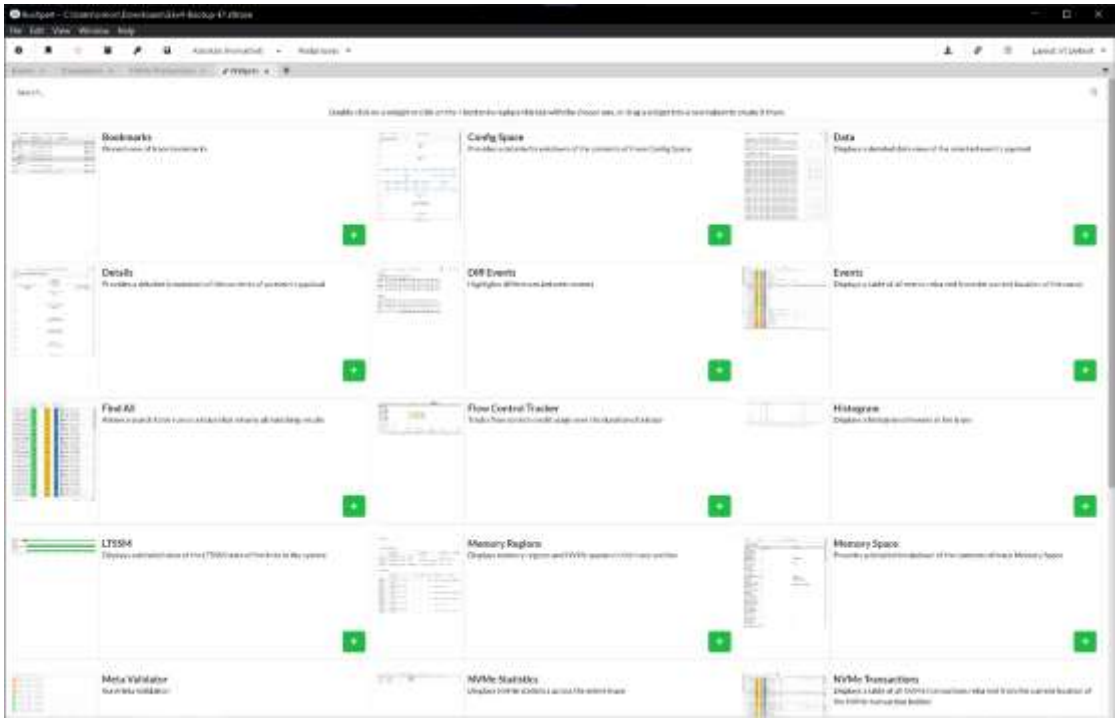
2.4.1.9 BusXpert 软件

方便的用户界面 - Web 浏览器和独立应用程序

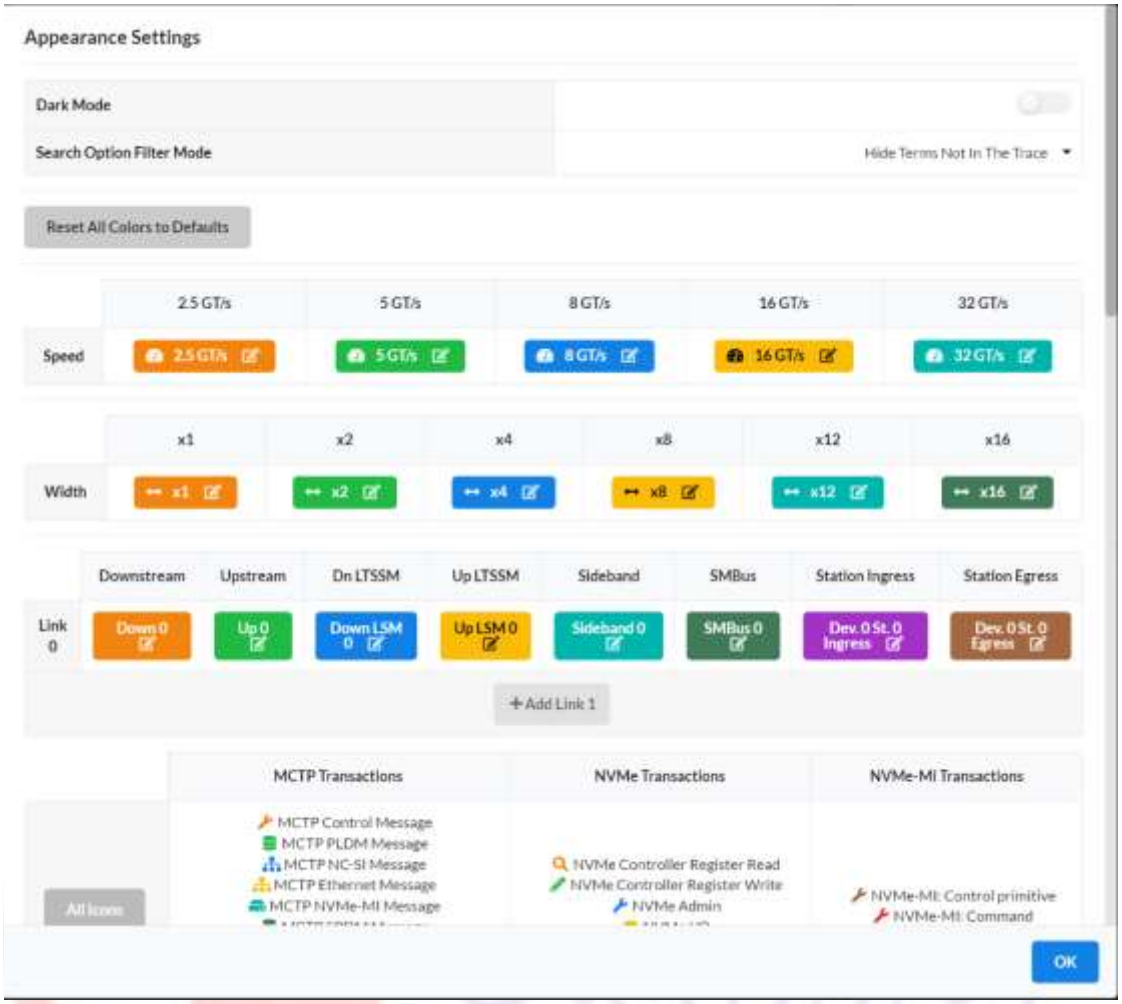


Kodiak 分析仪的一大特点是其方便的用户界面，可以通过 Web 浏览器或 SerialTek 基于 Electron® 的应用程序进行访问。这个用户友好的界面提供了一套强大的触发器、过滤器和 Trace 处理功能，使用户能够轻松解码和分析他们的数据。

该界面具有可定制的视图和 Widget，允许用户分析各种格式的 Trace 数据。这些 Widget 具有适合其功能的特定控件，并且还有一个适用于所有 Widget 的全局工具栏。



为了进一步增强易用性，可以使用布局管理器，使用户能够自定义其主页、捕获和 Trace 查看屏幕。这可以实现个性化且高效的工作流程。

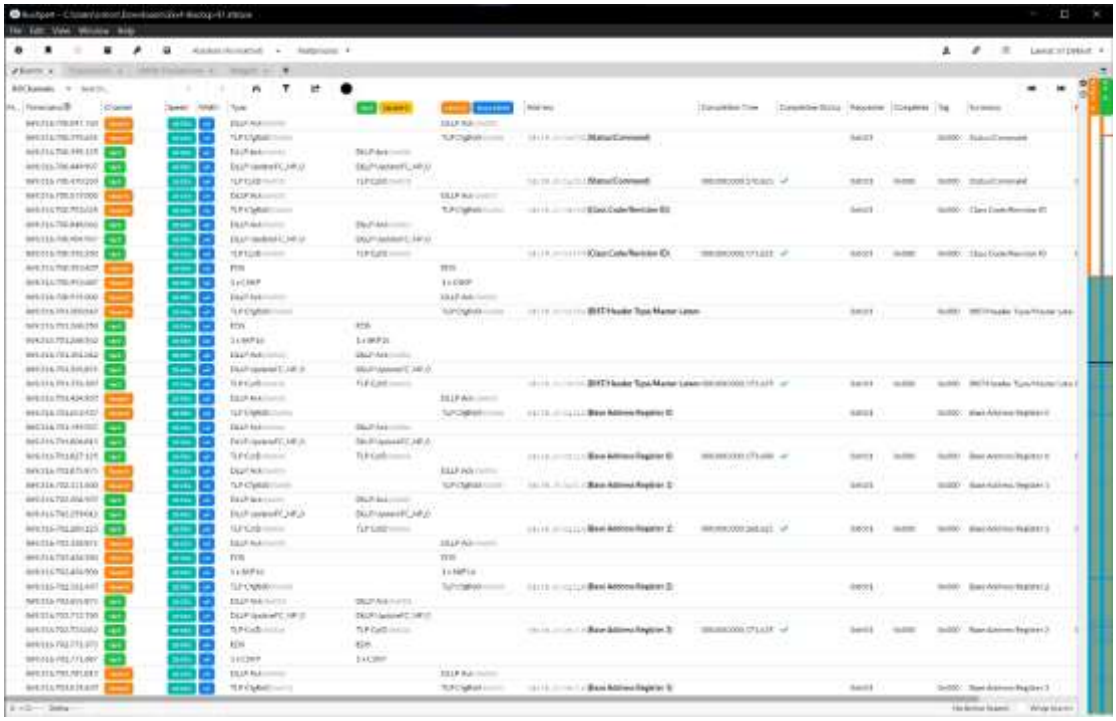


此外，Kodiak PCIe 6.0 分析仪通过其 REST API 提供轻松的自动化。该 API 提供了用于监控和捕获流量、执行统计分析以及进行详细搜索的编程工具。通过利用 REST API，用户可以简化自动化流程并提高效率。

总体而言，Kodiak PCIe 6.0 分析仪方便的用户界面可通过网络浏览器或独立应用程序访问，确保对 Trace 数据进行快速、可靠且用户友好的解码和分析。

2.4.2 BusXpert Widget 高效分析小工具

Event 事件视图



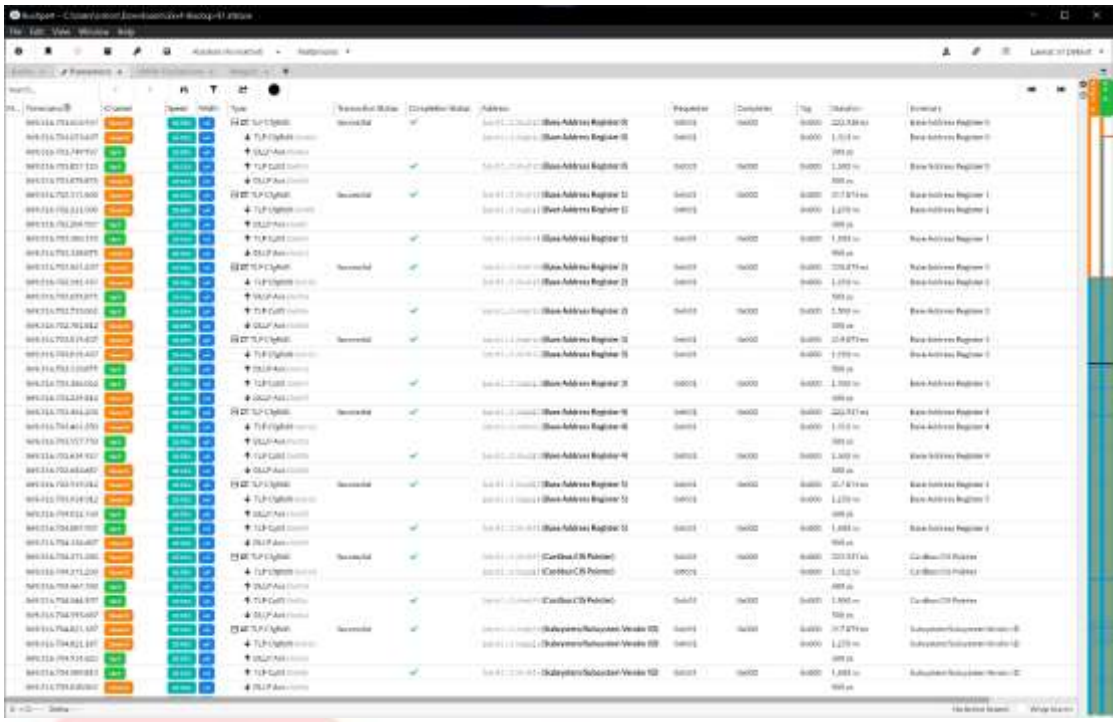
事件视图 Widget 是一个功能强大的 Kodiak PCIe 6.0 分析器视图，用于全面分析 Trace 数据。该 Widget 提供了总线上发生的事件的方便且有组织的表示，例如操作系统、DLLP、TLP、LTSSM 等。

在 Widget 中，每个事件都显示在单独的行中，并附有指示事件在总线上发生时间的时间戳。每个事件的相应数据填充在相应的列中，以便于分析和解释。

为了增强用户体验，事件视图 Widget 包括一个带有多个控件的工具栏，用于浏览数据，从而能够将 Widget 移动到 Trace 中的特定位置。此外，用户可以通过删除不感兴趣的事件来自定义视图，从而确保分析清晰且重点突出。此外，右键单击菜单提供了进一步细化分析的附加功能。

总体而言，Kodiak PCIe 6.0 分析仪中的事件视图 Widget 提供了一个全面且用户友好的界面，用于探索、分析 Trace 数据并从 Trace 数据中获得有价值的见解。其直观的控制、可定制的显示和有组织的事件表示提高了分析过程的效率和有效性。

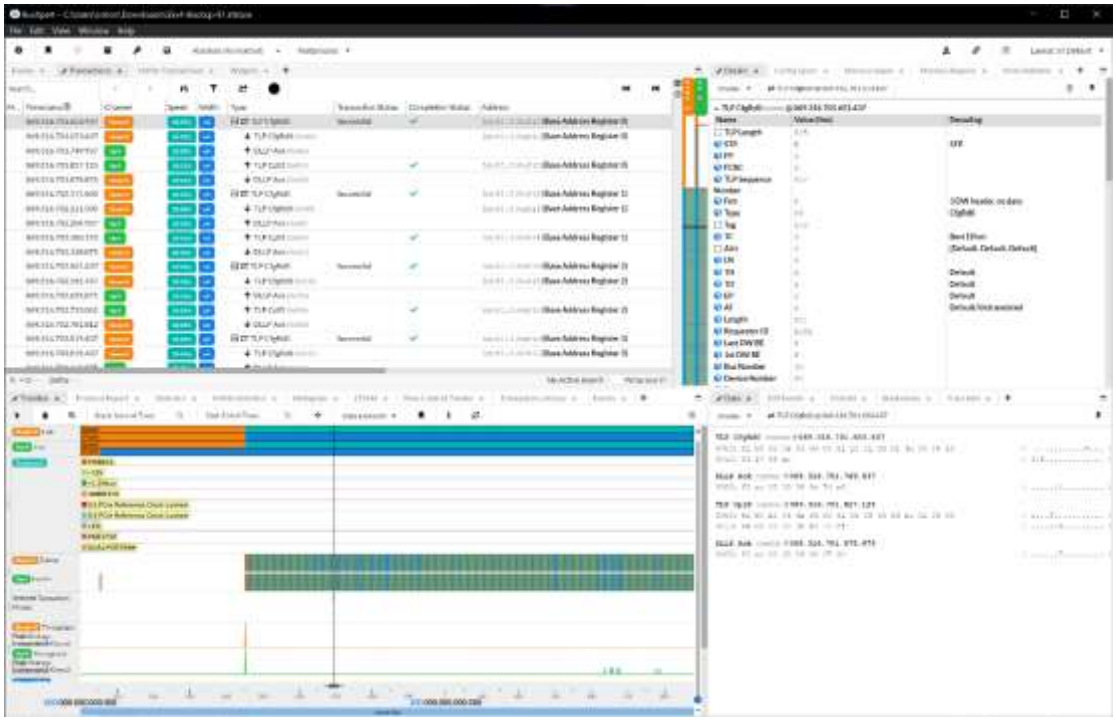
2.4.2.1 PCIe/CXL 和 NVMe 事务



这些 Widget 将各个事件组织成连贯的序列，从而在总线上形成完整的事务。

事务通常由从流的一侧发送的命令和从另一侧接收的相应响应组成。为了提供清晰简洁的交易视图，Widget 提供了扩展和折叠功能。展开后，事务中的所有事件都是可见的，允许用户单独检查每个事件。相反，折叠时仅显示交易摘要，使用户能够快速掌握整体流程，而不会被过多的事件信息淹没。

事务视图 Widget 中的每一行代表一个特定的数据包，提供有关它所代表的数据包的基本详细信息。这些列填充了从数据包中提取的信息，使用户可以轻松访问和分析相关数据。这些详细信息可能包括数据包类型、操作码值、地址、数据有效负载等，具体取决于所分析的特定数据包类型和协议。

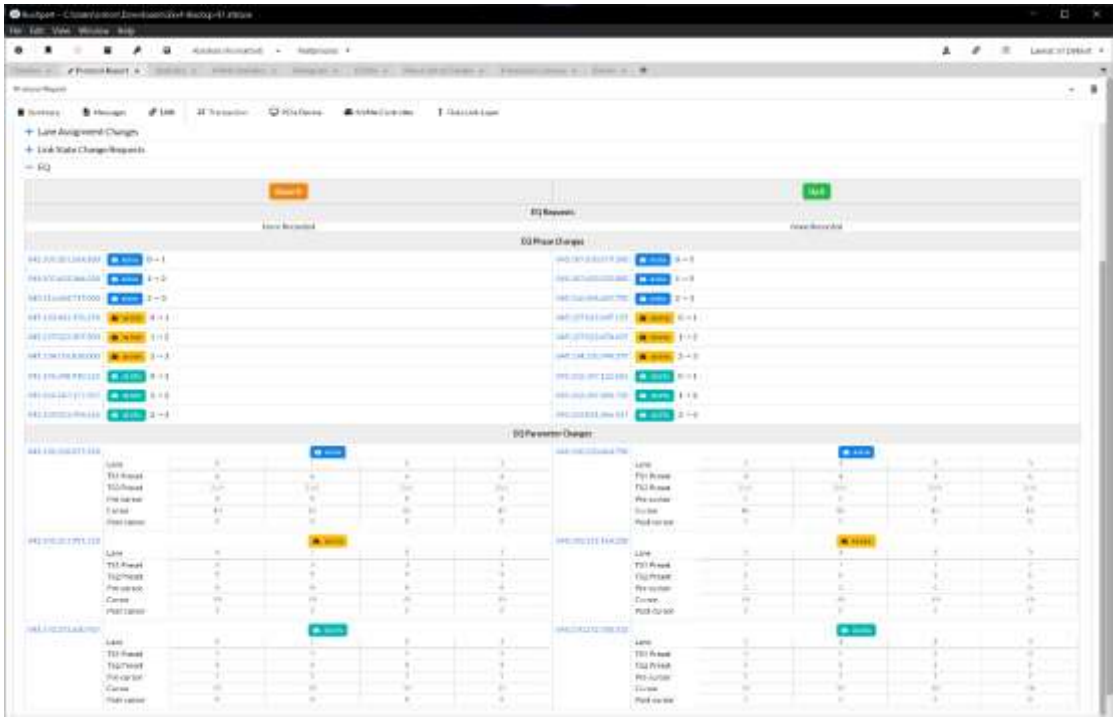


通过利用 Transaction 交易视图 Widget，用户可以更深入地了解每个交易中事件的顺序和结构。这可以实现高效的分析和故障排除，因为用户可以识别与命令和响应交互相关的任何异常、错误或关键见解。

交易视图 Widget 与 Kodiak 提供的其他分析工具和功能相结合，使用户能够彻底评估 PCIe 总线交易的复杂性。这种增强的可视性和控制可以更有效地分析、调试和优化 PCIe 通信。

2.4.2.2 协议报告

协议报告是一种全面的、可定制的报告工具，旨在突出显示和总结 Trace 中发生的变化以及捕获的协议信息。凭借其直观的界面和各种选项卡（包括摘要、消息、链接、事务、PCIe 设备、NVMe 控制器和数据链路层），协议报告 Widget 为用户提供了对其数据的深入洞察。



摘要选项卡

摘要选项卡提供了生成的报告的简明概述，显示了关键信息，例如创建日期、用户属性、消息总数和可用部分。此外，组织良好的表格显示了报告的配置设置，使其易于 Trace 和参考。

消息选项卡

在“消息”选项卡中，用户可以查看分析器捕获的所有更改的综合表格列表。为了增强可用性，可以使用三个选项根据严重性显示消息：信息、警告和错误。

信息性：呈现捕获期间发生的更改，例如链接 分配、速度、宽度、EQ 相位和参数更改。

警告：显示捕获期间检测到的警告，允许用户解决 潜在问题。警告的示例包括意外的 TLP 序列号、流 控制更新延迟和耗尽的流控制信用。

错误：突出显示捕获期间在链路上检测到的错误，包括 BAR 不一致、DLLP CRC 错误、意外的流控制状态转换和链路 CRC 错误。

链接选项卡

“链接”选项卡提供了对通过分析仪连接到主机系统的设备及其分配的总线设备功能 (BDF) 标识符的宝贵见解。通过提供变化的快速摘要，例如速度、宽度、链路和车道分配、链路状态以及均衡请求和相位变化，链路事件部分提供了系统连接的全面视图。

交易选项卡

“事务”选项卡为用户提供了一个方便的列表，其中包含协议报告标记的所有事务层事件的时间戳。此功能使用户能够轻松导航和分析关键交易事件，增强他们对捕获数据的理解。

PCIe 设备选项卡

与“存档管理”弹出菜单一样，“PCIe 设备”选项卡提供了所连接设备的完整概述。用户可以访问 BDF、供应商、设备、子系统、类代码和 BAR 等信息。还包括枚举、BAR 分配、内存启用、I/O 启用、总线主控启用和其他功能等基本事件的时间戳，提供主机启动期间设备行为的详细历史记录。

NVMe 控制器选项卡

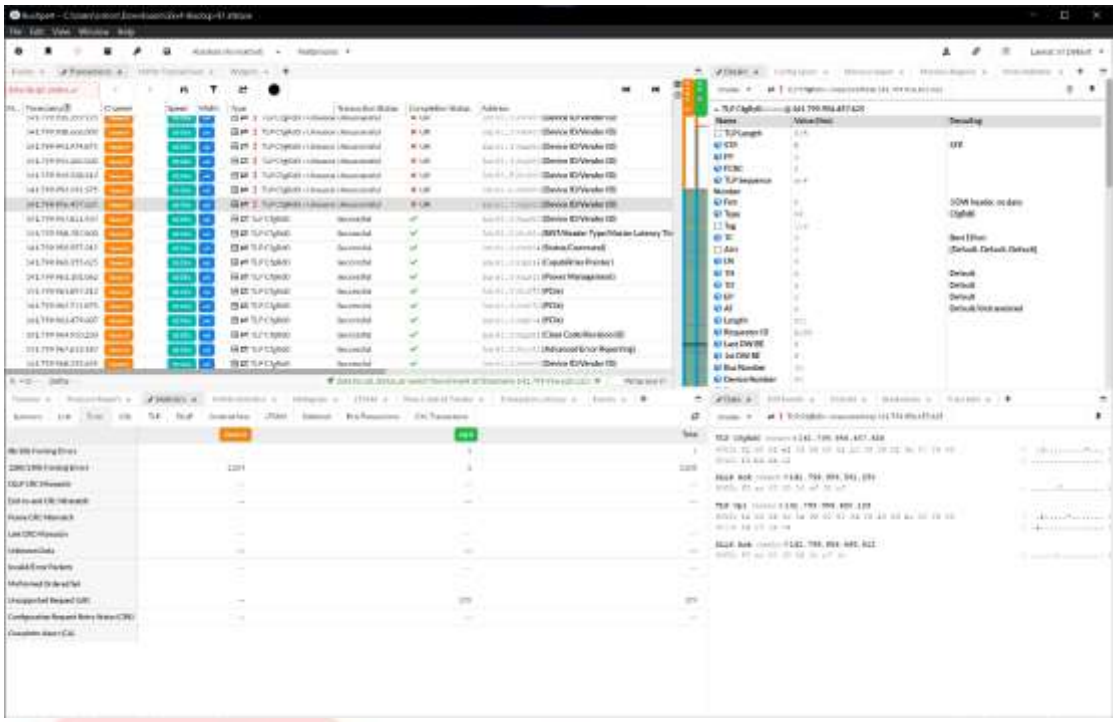
NVMe 控制器选项卡提供专门针对 Trace 中检测到的 NVMe 设备的见解。用户可以访问 BDF、供应商、子系统供应商以及设备启用和就绪时间戳等信息。通过提供此目标视图，用户可以获得对 Trace 中捕获的 NVMe 设备行为的宝贵可见性。

数据链路层选项卡

数据链路层选项卡侧重于捕获和呈现与 Trace 捕获期间发生的链路事件有关的信息。捕获事件的示例包括序列号翻转、意外的流控制状态、过期的流控制更新、耗尽的流控制信用和流控制信用警告。本节提供了识别和解决链接相关问题的重要信息。

凭借其全面的报告结构和用户友好的界面，协议报告 Widget 使用户能够分析和理解 Trace 期间捕获的数据。该 Widget 的可定制功能和相关选项卡选项使其成为高效、准确分析的宝贵工具。

2.4.2.3 统计数据



BusXpert 软件中的统计 Widget 提供了有关 PCIe 总线的性能和行为的宝贵见解。该 Widget 为用户提供与总线事务相关的各种参数和指标的全面统计分析。

使用统计 Widget，用户可以访问关键信息，例如带宽利用率、错误率、延迟和吞吐量。该 Widget 以结构化且易于理解的格式呈现此数据，使用户能够快速识别可能影响系统的任何性能瓶颈或异常情况。

此外，统计 Widget 提供了可视化表示，例如图表和图形，以促进更好的数据解释。这些视觉效果可以帮助用户可视化性能指标随时间变化的趋势、模式和变化，使他们能够做出明智的决策并采取适当的行动来优化 PCIe 总线。

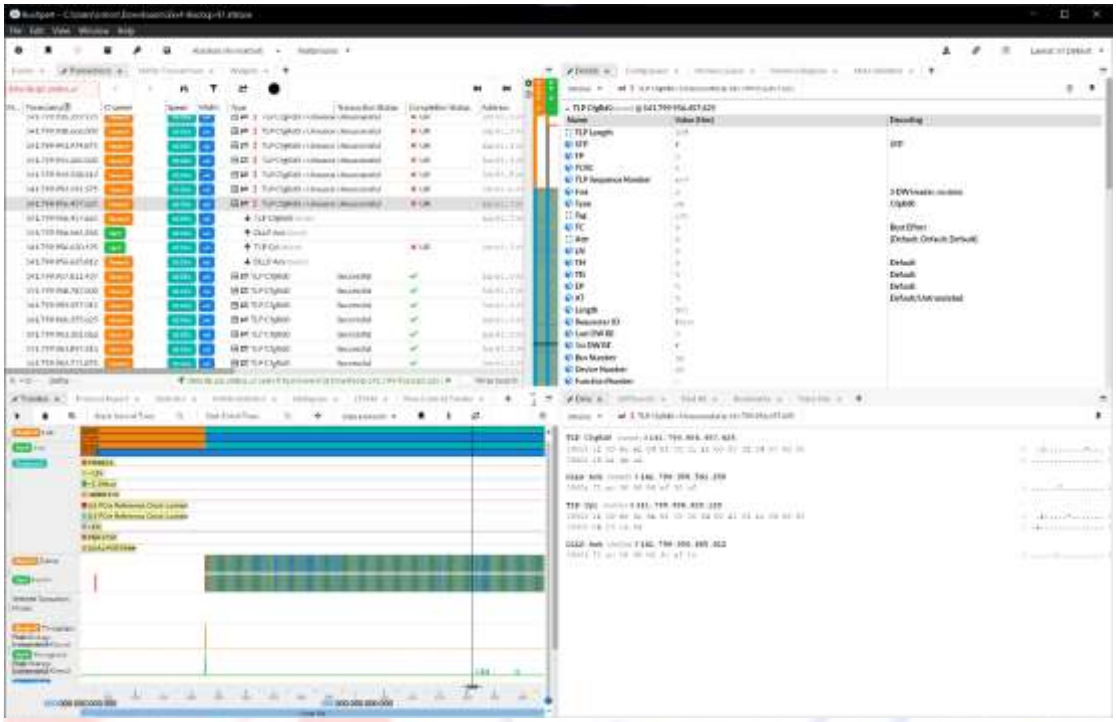
通过利用统计 Widget，用户可以全面了解总线性能并主动解决潜在问题，确保 PCIe 基础设施的最佳运行和效率。

2.4.2.4 数据包详情

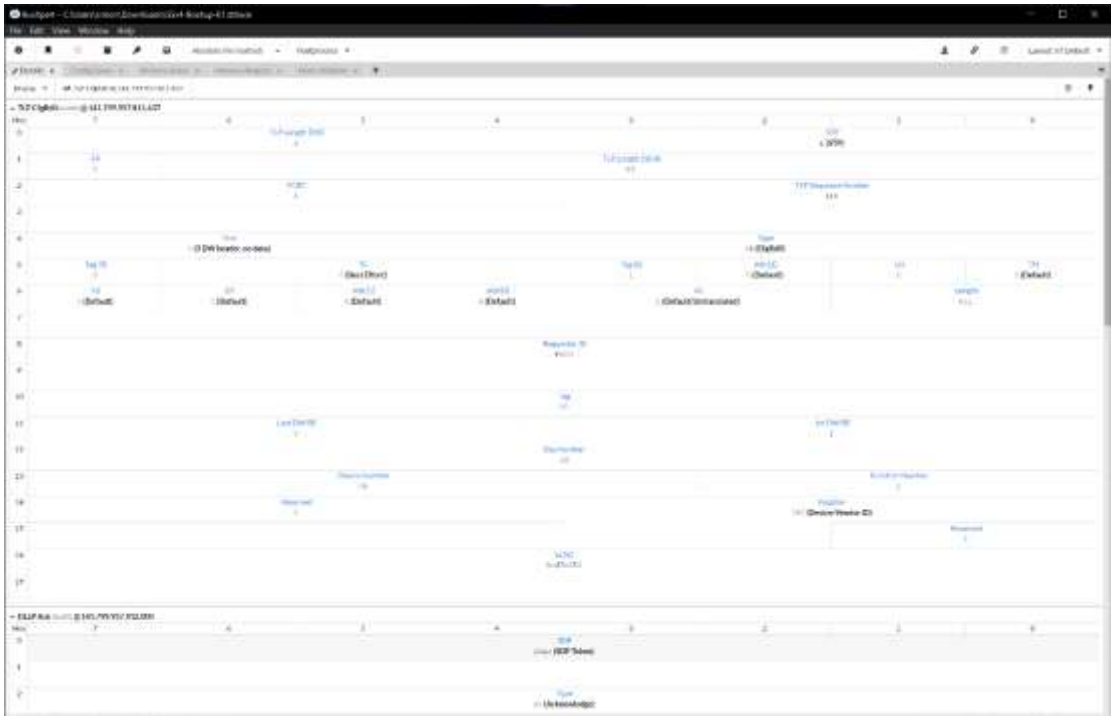
数据包详细信息 Widget 为用户提供 PCIe 总线内各个数据包的具体详细信息和内容的全面视图。

该 Widget 为用户提供了两种查看选项：

列表视图：列表视图专为快速、轻松的可读性而设计，以用户友好的格式呈现重要信息。用户可以快速访问与数据包类型、数据有效负载、地址和其他相关详细信息相关的重要信息。



Field View： Field View 格式源自 PCIe 规范，根据特定的数据包类型将数据包数据分解为多个字段。这使用户能够根据这些字段中的值分析文本解码，从而更详细地了解数据包内容。

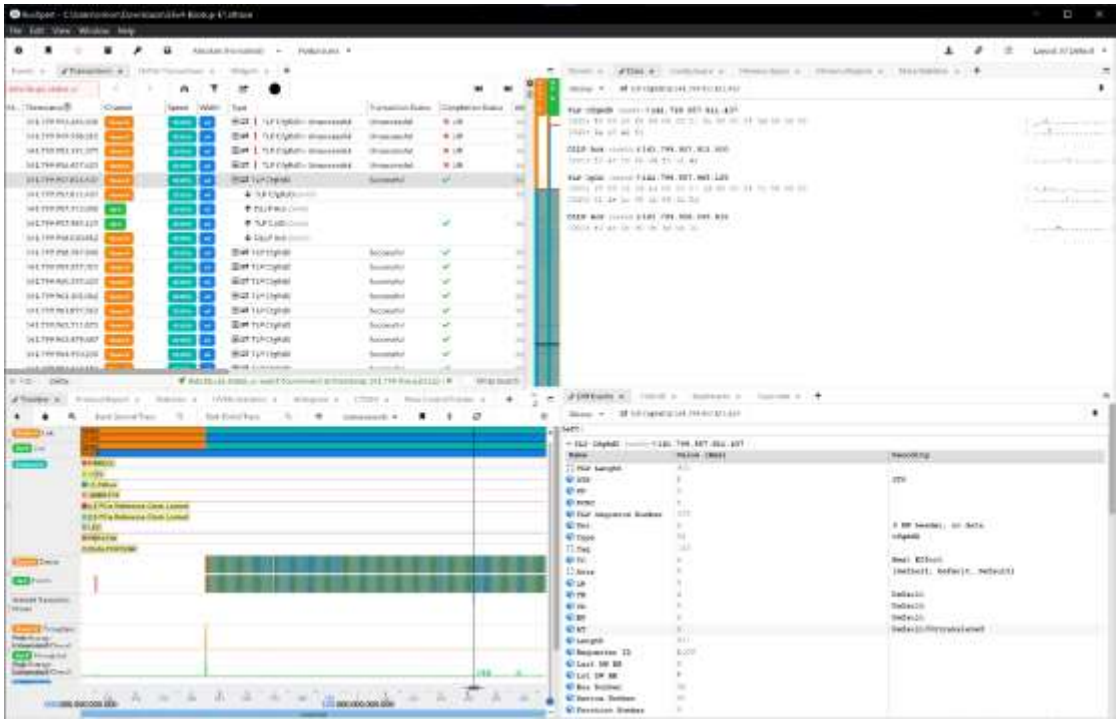


通过数据包详细信息 Widget，用户可以浏览 Trace 数据并沉浸在每个数据包的复杂详细信息中。这种级别的可见性和分析对于诊断问题、验证协议合规性以及深入了解 PCIe 总线的行为和性能非常有价值。

2.4.2.5 数据包数据

数据包数据 Widget 是一个显示数据包原始字节的组件。该 Widget 使用户能够以多种方式查看和操作数据包数据，以帮助调试和分析。

使用“数据包数据”Widget，用户可以选择复制数据、格式化/调整列宽以及更改字节序。这些功能使用户可以更轻松地根据自己的具体分析需求检查和修改数据包数据。



2.4.2.6 基于时间轴的 Activity 和 LTSSM 展示

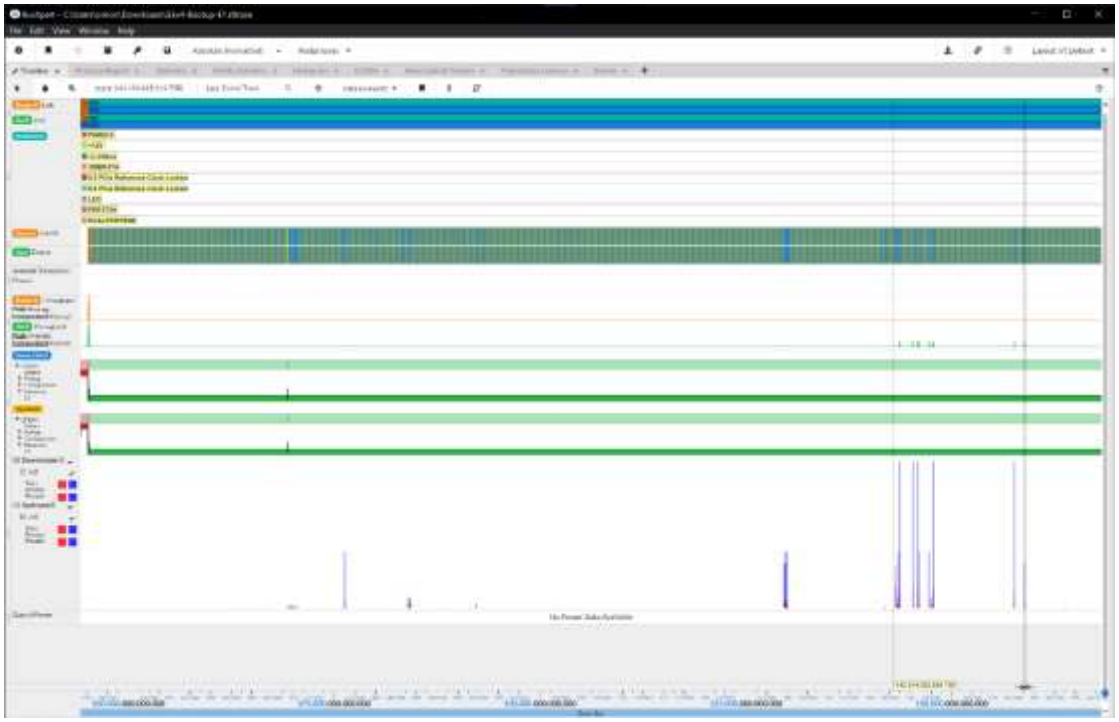
时间轴 Widget 为用户提供了 Trace 数据中发生的事件和活动的可视化表示。它提供了时间轴的全面视图，允许用户浏览 Trace 数据并深入了解事件的时间顺序。

在时间轴 Widget 中，用户可以通过平移和缩放以交互方式探索 Trace 数据，以关注特定的感兴趣区域。他们还可以启用同步，以确保在浏览时间轴时当前选定的事件仍保留在视图中。

时间轴 Widget 允许用户通过选择要显示的数据元素（例如边带、数据、吞吐量和其他时间轴元素）来自定义其视图。这种灵活性确保用户可以专注于与其分析最相关的特定元素。

总体而言，Kodiak PCIe Gen6 分析系统中的时间轴 Widget 为用户提供了 Trace 数据的强大可视化表示，使他们能够有效地分析值得注意的事件并轻松浏览时间轴以进行详细检查。

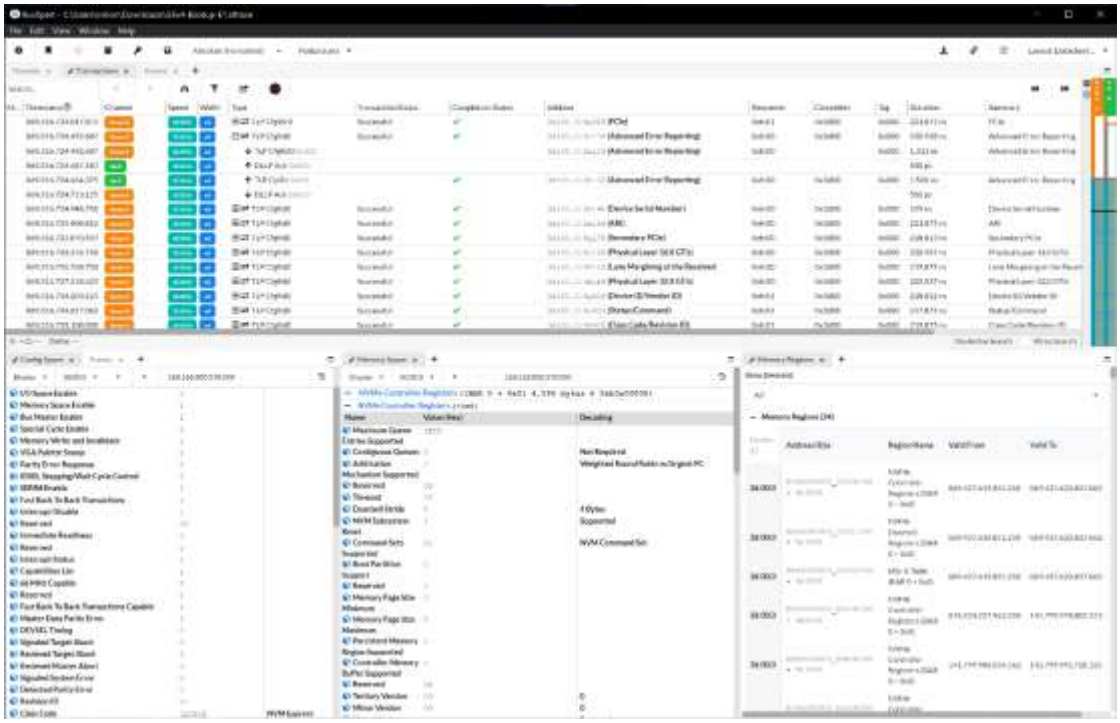
LTSSM 视图是 Trace 数据中所示的“链路训练和状态机 LTSSM”的时间轴。该 Widget 按链接的上下游排列。数据显示为展开的树，并且可以放大以更细粒度地查看数据。



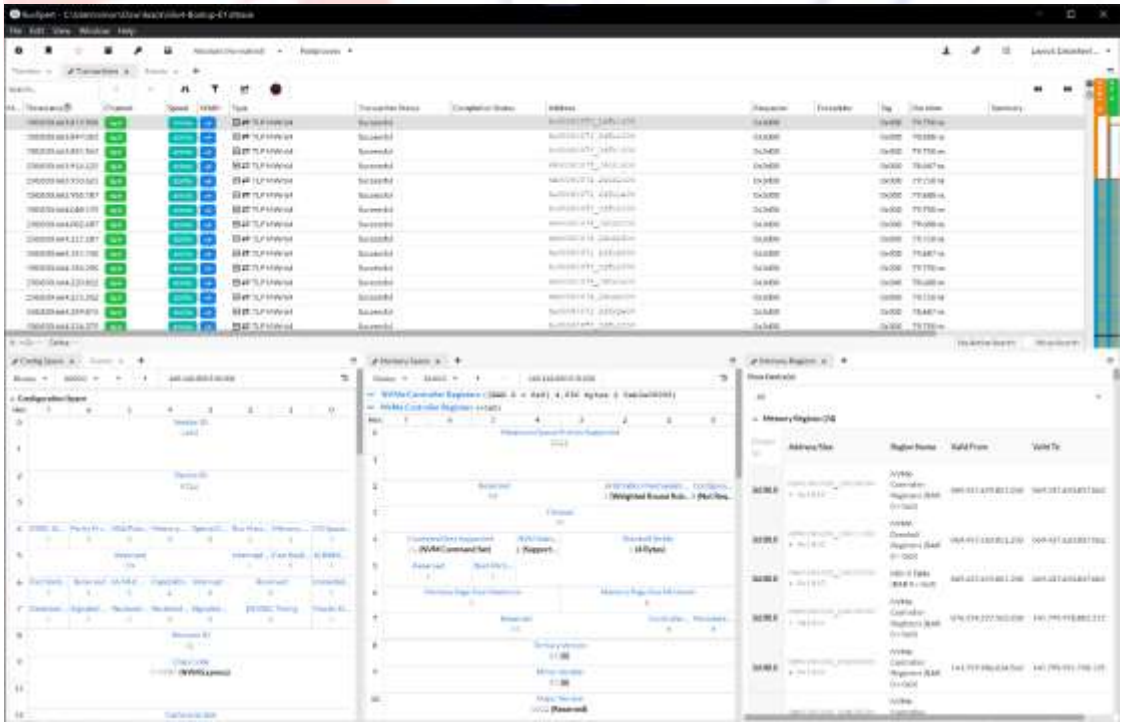
2.4.2.7 配置空间、内存空间和内存区域

配置空间 Widget 是 Kodiak PCIe Gen6 分析系统的关键组件，为用户提供两个视图：字段视图和规格视图。该 Widget 还为用户提供了配置和内存空间历史/更改的详细视图，只需触摸按钮即可。

在“字段”视图中，“配置空间”Widget 以列表格式显示数据，允许用户查看重要信息。用户可以复制数据、更改数字格式并将数据可视化为字节或双字。



另一方面，*Spec 视图根据 PCIe 规范呈现数据。*它根据数据包类型将数据分解为字段，并根据这些字段中的值提供文本解码。



同样，*“内存空间”Widget* 允许用户探索所分析的 PCIe 设备内的内存区域。它提供了内存空间的全面视图，使用户能够检查和修改不同内存区域内的数据。

此外，内存空间 Widget 提供了各种功能来增强分析。用户可以通过选择仅显示特定内存区域或固定重要数据字段以供快速参考来自定义视图。该 Widget 还可以简化诸如复制数据、修改数字格式以及以字节或双字形式可视化数据等任务。

内存*区域 Widget* 是 Kodiak PCIe 分析系统的另一个组件，允许用户检查设备内的特定内存区域。用户可以选择并探索所需的内存区域，以更深入地了解其中包含的数据。

总之，Kodiak PCIe Gen6 分析系统中的配置空间 Widget、内存空间 Widget 和内存区域 Widget 为用户提供了强大的工具来分析 PCIe 设备的配置和内存空间内的数据。

附加 Widget/功能

- 实时链接统计
- 差异事件：比较两个 PCIe 数据包
- 全局和用户书签
- Trace 摘要信息
- 缓冲区状态
- LED

Kodiak CXL PCIe 6.0 测试仪

Kodiak PCIe 6.0 / CXL 3.0 协议测试仪是一款功能强大的工具，专为测试和验证 PCI Express 和 CXL 技术而设计。它提供了广泛的功能来确保合规性、验证符合规范的行为以及优化设备和系统性能。以下是 Kodiak 测试仪的主要优点和功能：

1. 确保合规并验证质量：

- 验证 PCIe 认证测试套件 (CTS) 和 SerialTek 特定质量测试套件的一致性。
- 根据行业标准验证实施的稳健性和质量。

2、主机和设备不可用时的接入测试：

- 使用测试仪进行早期测试和故障排除，克服对主机和设备的访问受限的挑战。

3. 解决具体问题：

- 将有问题的 Trace 转换为可重现的测试以进行故障排除。

4. 验证具体功能：

- 测试和验证数据对象启用 (DOE)、中断、显示启用 (IDE) 以及链路训练和状态机 (LTSSM) 状态转换等功能。

5. 手动测试和定制：

- 允许用户修改功能并执行有限的手动测试，例如发送特定的 TLP 或测试对格式错误的 TLP 的响应。

6. 早期性能测试：

- 在开发的各个阶段执行性能测试，以优化设备和系统性能。

7. 测试仪支持多种模式，包括一致性测试模式、

- 手动测试模式、环回模式、功能测试模式、Trace 重放模式、性能测试模式和码型生成器模式。

对于 CXL 合规性测试，测试仪遵循 CXL 规范中概述的指南，涵盖各个层，例如 CXL.io 和 CXL.cache 应用层/事务层测试、链路层测试、ARB/MUX、交换机测试和配置寄存器测试。

此外，Kodiak 测试仪还提供全面的 PCIe 合规性套件，测试涵盖链路层、事务层和协议功能。

手动模式允许用户模拟主机或设备环境、修改配置空间、更改 LTSSM 状态以及强制边带信号进行全面验证。此模式可以测试协议规范中定义的特定功能和行为。

该测试仪提供用户友好的软件界面和脚本构建功能，允许用户创建定制的测试场景。

为了优化性能，Kodiak 测试仪通过基于 Web 的界面和灵活的 REST API 提供轻松的性能测试和自动化。它还提供真实世界的模拟功能来模拟各种场景和条件。

该测试仪可确保在开发的每个阶段进行彻底的测试，帮助识别性能瓶颈并优化数据传输速率。它提供详细的性能指标、测试结果分析以及数据驱动的决策工具。

使用 Kodiak PCIe 6.0 / CXL 3.0 协议测试系统体验您的技术的全部潜力，并为您的设备或系统实现卓越的性能和合规性。

Kodiak 系列外壳

- 尺寸：443 x 81 x 331 毫米（17 x 3.2 x 13 英寸）
- 重量：8 千克（17.6 磅）
- 安装：19” 机架安装选项、倾斜脚选项
- 环境工作温度：海拔 2133m（7000 英尺）以下时为 5-35°C

显示屏和指示器

- 前面板 LCD：800×320 4.6” WCGA，触摸屏
- 系统状态：RGB LED

前面板连接器

- 内插器连接：4x QSFP-DD 和 1x MCIO
- 以太网 (10 GbE)：2 个 SFP+ 端口
- 以太网 (1 GbE)：RJ45
- PCIe 接口：2x OCuLink

- USB 接口：2 个 USB 3.2 A 型

后面板连接器

- 电源：IEC C13，100-240 伏交流电，50-60 赫兹
- 时钟输出：SMA、50 Ω、3.3 Vdc、10 MHz
- 时钟输入 (10 MHz)：SMA、50 Ω、3.3 Vdc、10 MHz
- 触发输出：SMA、50 Ω、3.3 Vdc
- 触发输入：SMA、50 Ω、3.3 Vdc
- 维护：RJ45、USB-C（不供客户使用）

维护和许可

- 终身免费软件更新 – 无维护费
- 免费的全功能查看器软件 – 在计算机和同事之间轻松共享带注释的 Trace 并重播捕获的流量
- 在任何计算机上使用 SerialTek 硬件 – 无需额外许可证

保修单

- 一年有限保修，标准版
- 两年有限保修，专业版
- 三年有限保修，企业版
- 六个月有限保修，Interposer

最低要求

- Intel Core、2 GHz 或兼容处理器
- 4 GB 内存
- 1280 x 1024 显示分辨率，至少 65,536 色
- 仅 64 位操作系统（Windows 7、Ubuntu 14、Centos7 或更高版本）
- 1GbE 控制器

Kodiak PCIe 6.0/CXL 3.0 x16 协议分析仪	
Kodiak PCIe 6.0 x16 协议分析仪和测试仪 - 企业版。 288GB 缓冲区、8TB SSD、2x 10GE PCIe 6.0、CXL、SMBUS、DOE、IDE、PAM ***测试仪许可证 CXL/PCIe 6.0、PCIe 6.0（包括 5.0）CTS 模式、手动模式***	PK3-A-G6-16-ENT
Kodiak PCIe 6.0 x16 协议分析仪和测试仪 - 专业版。 获得 144GB、4TB SSD、2x 10GE、2 年保修 分析仪许可证 PCIe 6.0、CXL、SMBUS、DOE、IDE、PAM ***测试仪许可证 PCIe 6.0、CTS 模式、手动模式***	PK3-A-G6-16-PRO
Kodiak PCIe 6.0 x16 协议分析仪 - 标准版。 获得 72GB、2TB SSD、10GE x1 PCIe 6.0 分析仪许可证、 PCIe 6.0 PAM 测试仪许可证、PCIe CTS、手动模式许可	PK3-A-G6-16-STD

2.5 SerialTek PCIe 协议分析仪的连接方式

注意：协议分析仪主机单独是无法使用的，必须配合一个 interposer。Interposer 的种类多种多样，主要是根据不同的接口来设计。目前 SerialTek PCIe Gen5 analzer 提供如下主流接口的 interposer。

图 2-32

下面，我们看一下几种最常使用的 PCIe 接口 interposer 是如何连接到分析仪的，熟悉了这几种连接方式后对于其它 interposer 连接方式类推即可。

2.5.1 U.2/U.3 NVMe SSD 协议分析连接图

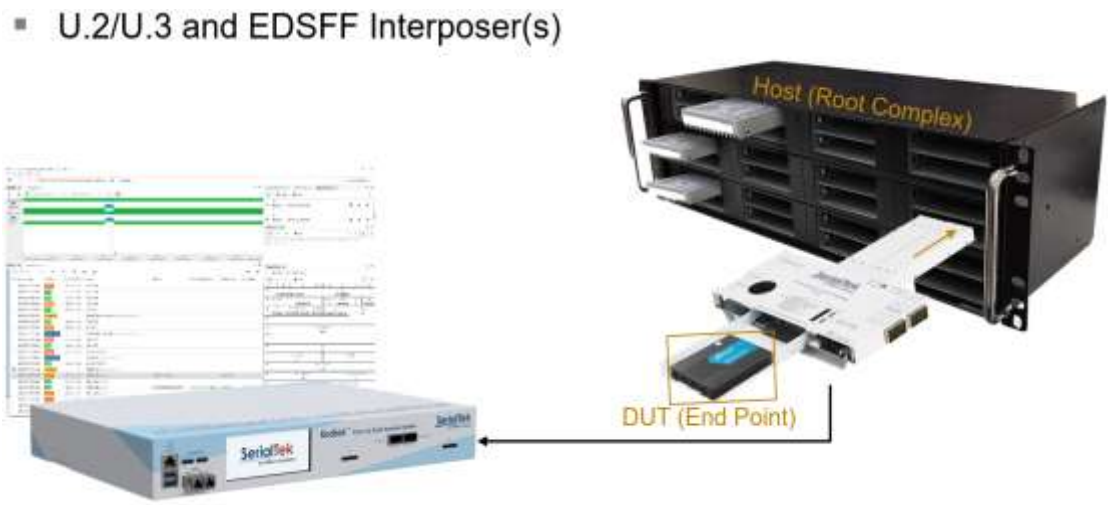


图 2-33

参见上图，U.2/U.3 interposer 串接在 U.2/U.3 NVMe SSD 和背板之间，不影响双向交互，双向数据（upstream, downstream）通过 interposer 侧面的线缆接口直接旁路到 PCIe 分析仪。下图是一个实际连接图供参考，图中的 Intel NVMe SSD 可以是 single port，也可以是 dual port。

图 2-34 Gen 4 U.2 NVMe SSD 协议分析仪连接实拍图

2.5.2 M.2 NVMe SSD 协议分析实际连接图

图 2-35

参见上图，M.2 interposer 串接在 M.2 NVMe SSD 和背板之间，不影响双向交互，双向数据（upstream, downstream）通过 interposer 侧面的线缆接口直接旁路到 PCIe 分析仪。

注意：图片右上角列出了 4 种长度的 M.2 HSA (host side adapter)， 即 2240, 2260, 2280, 22110。一般如果是连接到笔记本或者台式机主板普遍采用 2280 的 HSA adapter。即，先将 2280 HSA 插入到笔记本或者台式机主板 M.2 slot，锁紧螺丝，然后将 2280 HSA 上面的两跟高速线缆连接到 M.2 interposer（两根高速线缆为两个方向流量），然后接入 M.2 NVMe SSD。这样，M.2 interposer 通过侧面的 upstream, downstream 线缆连接到分析仪前面板。

图 2-36 Gen 4 M.2 NVMe SSD 协议分析仪连接实拍图

图 2-37 M.2 interposer 细部连接实拍图

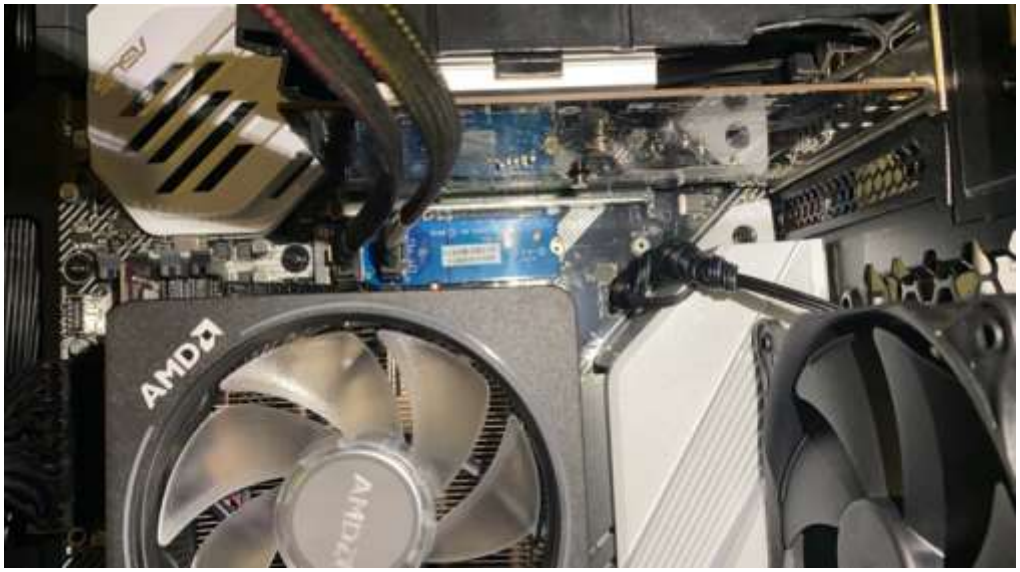


图 2-38 M.2 interposer 的 Host Adapter 细部连接实拍图

2.5.3 PCIe AIC 插卡协议分析实际连接图

- AIC Interposers

图 2-39

参见上图，slot interposer (有时也称为插卡 interposer，或者 AIC (add-in-card) interposer, CEM (Card Electromechanical) interposer 等)串接在待测插卡（例如 SSD 卡，GPU 卡，DPU 卡，网卡，RAID 卡，加速卡等）和主板插槽之间，不影响双向交互，双向数据（upstream, downstream）通过 interposer 侧面的线缆接口直接旁路到 PCIe 分析仪。

类似于上述的针对插卡的协议分析，有的时候受制于机箱插槽位置和空间的限制，Slot interposer 无法直接插入 PCIe 插槽，这个时候就需要 PCIe Gen 4/5/6 x4 延长线，将 PCIe Slot 延伸出来方便 Slot interposer 插入。参见下图：

图 2-40 Slot interposer 协议分析仪连接实拍图（带延长线）

2.5.4 PCIe EDSFF E1.S/E1.L/E3.S/E3.L Interposer

参见下面的 EDSFF interposer 的图片说明，该 interposer 适用于上述 4 种接口的 EDSFF SSD 开发、验证、测试，已经使用这些接口 SSD 的高密度大容量服务器、存储系统设计。

使用这些 EDSFF interposer 的时候，只需要将 EDSFF interposer 串接在 EDSFF 槽位和 EDSFF SSD 中间，然后两根标识 upstream 和 downstream 的线缆连接到 PCIe 协议分析仪即可。

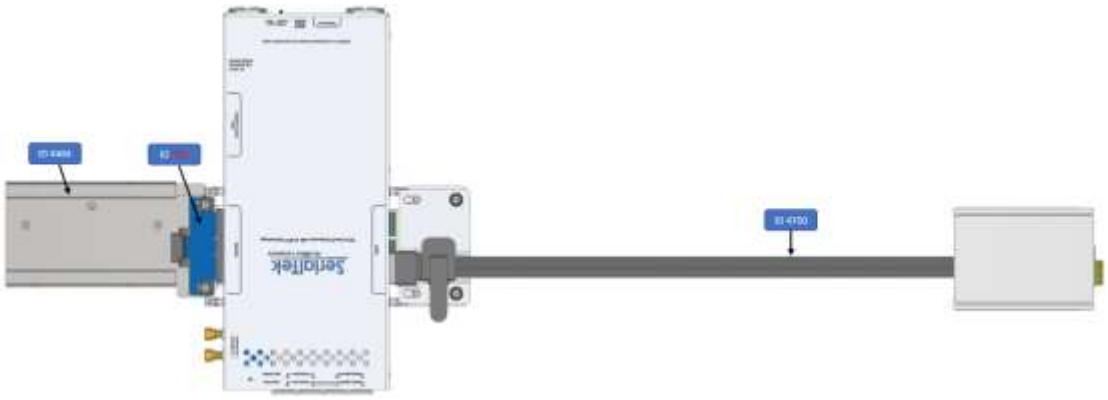


图 2-41

上图是 Gen5 E3.S interposer，用户只要将待分析的 E3.S SSD 插在左边的 tray 盘，右边的类似于 E3.S 的 HAS (Host Side Adapter) 插入主机或存储系统盘柜的 E3.S 背板插槽即可。

2.5.5 PCIe Cable Interposer

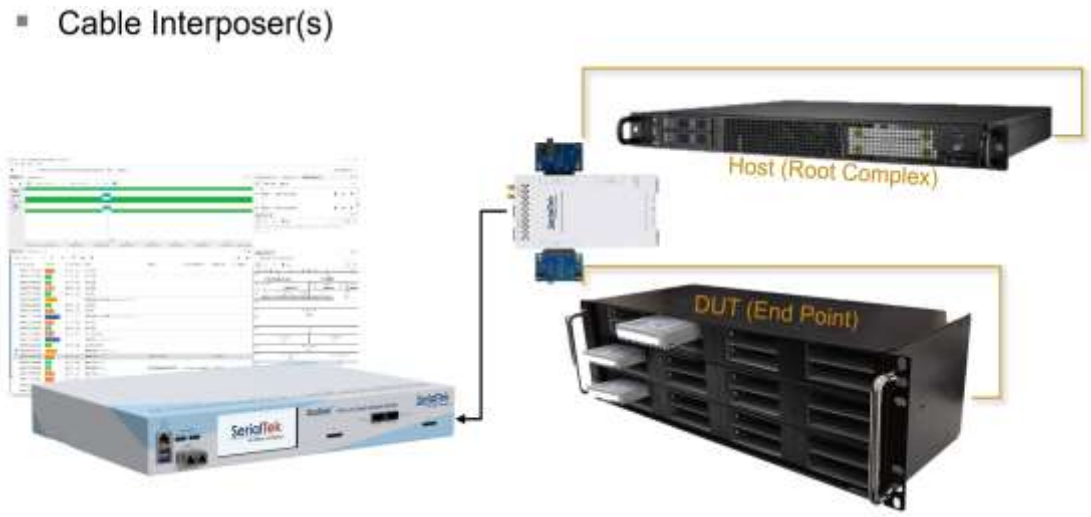


图 2-42

参见上图，cable interposer 串接在 PCIe Cable 中间，一般是主机和 JBOF 盘柜之间，或者存储系统机头和盘柜（或者盘柜和盘柜级联链路）之间。Cables interposer 根据不同给的接口，又分为不同的 cable interposer，常见下面几种：

- **MCIO cable interposer**
- **HD-MINI-SAS cable interposer**
- **Slim-SAS cable interposer**
- **Oculink cable interposer**

使用这些 Interposer 的时候，只要将 cable interposer 串接在 PCIe cable 中间，然后将两根标识 upstream 和 downstream 的线缆连接到 PCIe 协议分析仪即可。

2.6 SerialTek PCIe 协议分析仪产品硬件

2.6.1 Gen5 协议分析仪 Interposer 展示

下面以 SerialTek 公司的 PCIe Gen 5 analyzer 为例，介绍一下业内最常用的各种接口的 interposer，目前提供至少 13 种接口（如果将 3 种 slot interposer 以及 6 种企业级盘的接口的 dual port 单列的话，那么总数将达到 21 种），使用户可以轻松应对分析各种接口的 PCIe 链路问题。

- **AIC 插卡，支持 x16, x8, x4**
- **U.2**
- **U.3**
- **E1.S**
- **E1.L**
- **E3.S**
- **E3.L**

- **M.2**
- **OCP**
- **MCIO cable**
- **HD Mini SAS cable**
- **Slim SAS cable**
- **Oculink cable**

除了单独的 interposer 之外，针对 SSD 用户，SerialTek 还提供一个打包配置，即，通过一个客户可自由组合的 PCIe Gen5 Interposer 设计轻松解决 U.2, U.3, M.2, E1.S, E1.L, E3.S, E3.L 等 7 种 SSD 接口的问题分析，这大大降低了分次、分别购买各个接口的 interposer 带来的不便和成本，对于专注于开发、使用 NVMe SSD 的公司大大提高了便利性。

下面的各种 Interposer 的工作机制前面章节有介绍，Interposer 的功能主要是“串接”在各类接口的 PCIe 链路中间，将 upstream (SSD -> CPU)和 downstream (CPU -> SSD)的双向流量导出到“旁路”的 PCIe 协议分析仪进行处理和分析。

下面只展示各种 interposer 图片，大家可以重点看一下 interposer 两侧的 female 和 male 接头即可很容易了解每种 interposer 是如何串接在链路中间的，具体功能不再单独赘述。

2.6.1.1 Gen 5 Slot Interposer

图 2-43

下面是配合 Gen5 x16 slot interposer 使用的各种 adapter 的图片展示。

图 2-44

2.6.1.2 Gen 5 U.2/U.3 Interposer

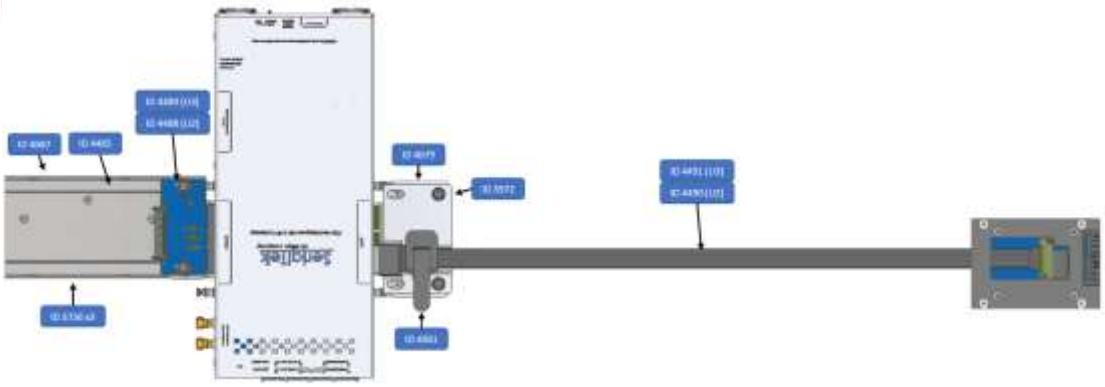
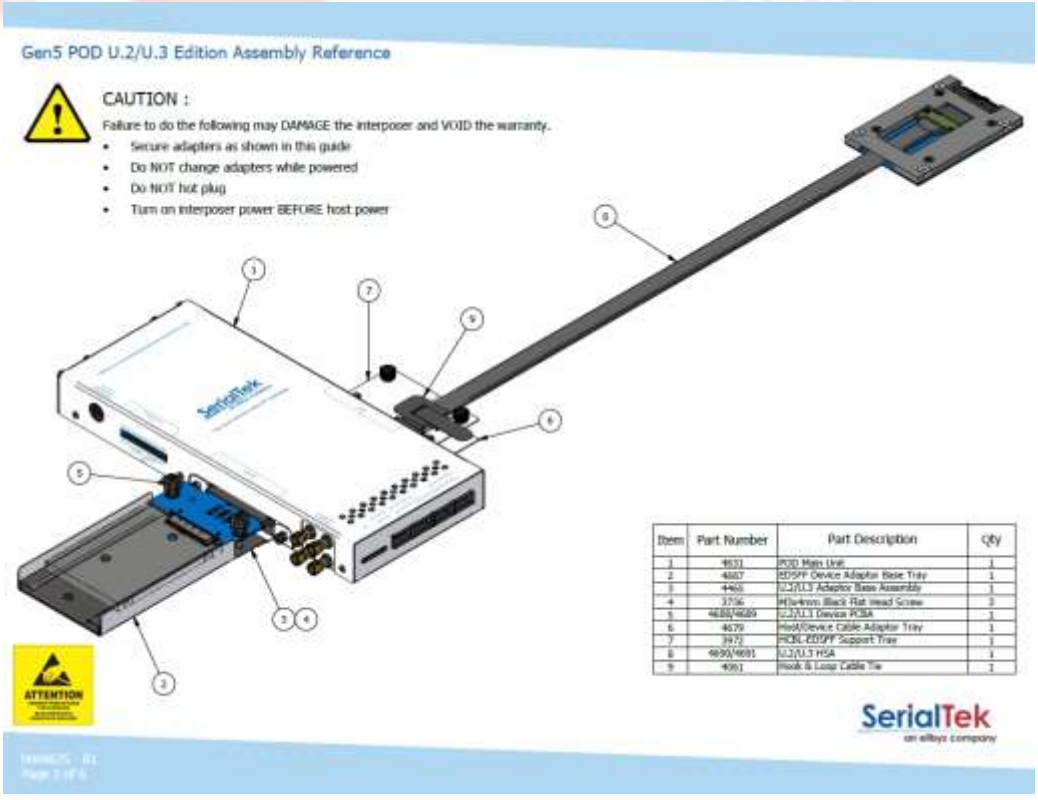


图 2-45

下图为 U.2 或者 U.3 interposer 的实拍照片。

图 2-46

2.6.1.2.1 Gen5 Pod 组装示意图 – U.2 & U.3



Item	Part Number	Part Description	Qty
1	4631	POD Main Unit	1
2	4687	EDSFF Device Adaptor Base Tray	1
3	4466	U.2/U.3 Adaptor Base Assembly	1
4	3736	M3x4mm Black Flat Head Screw	3
5	4688/4689	U.2/U.3 Device PCBA	1
6	4679	Host/Device Cable Adaptor Tray	1
7	3972	PCIE-EDSFF Support Tray	1
8	4690/4691	U.2/U.3 HSA	1
9	4061	Hook & Loop Cable Tie	1

图 2-47

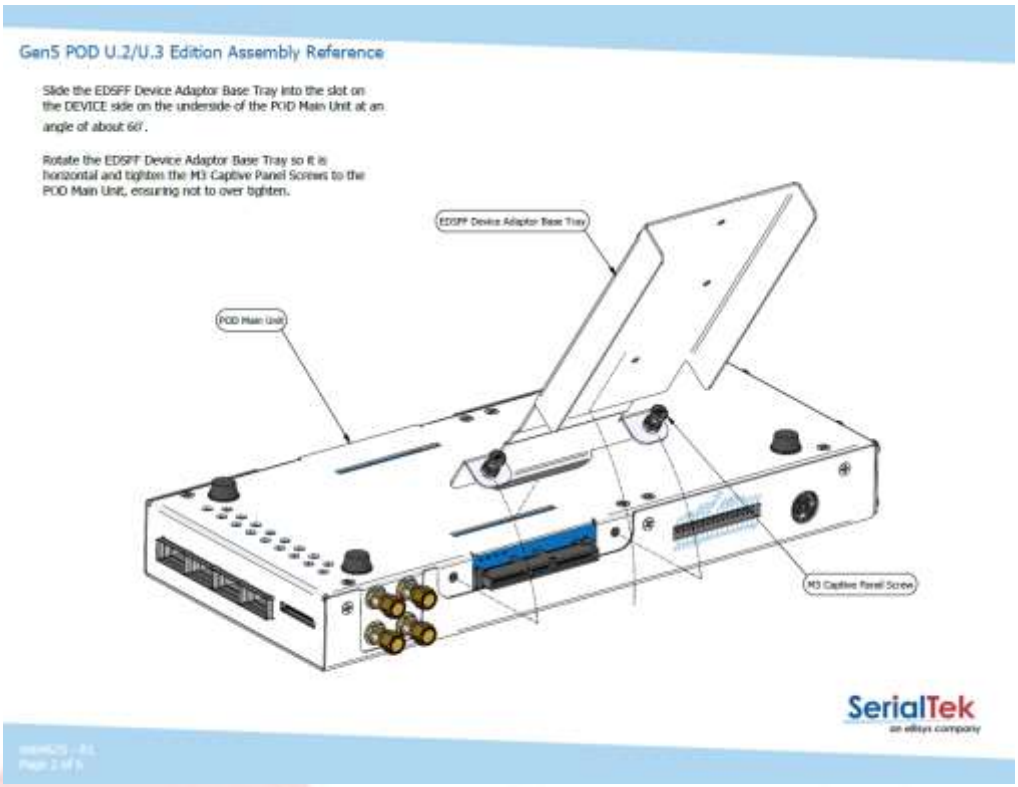

Gen5 POD U.2/U.3 Edition Assembly Reference
Slide the EDSFF Device Adaptor Base Tray into the slot on the DEVICE side on the underside of the POD Main Unit at an angle of about 60°.
Rotate the EDSFF Device Adaptor Base Tray so it is horizontal and tighten the M3 Captive Panel Screws to the POD Main Unit, ensuring not to over tighten.
EDSFF Device Adaptor Base Tray
POD Main Unit
M3 Captive Panel Screw
SerialTek

图 2-48

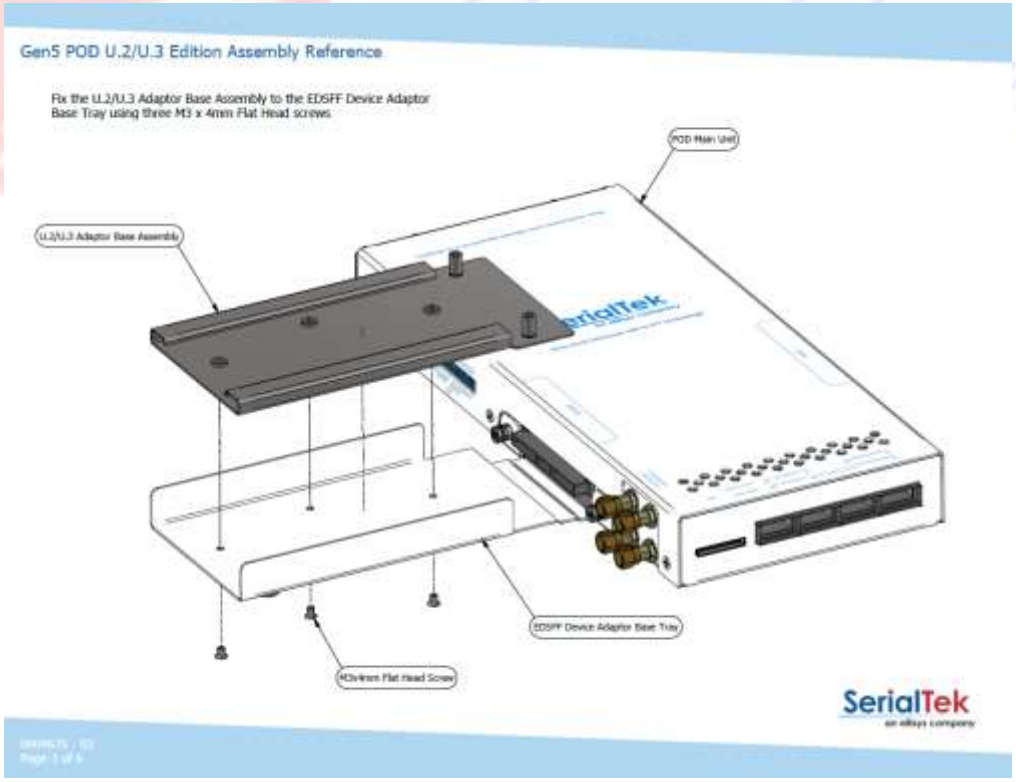

Gen5 POD U.2/U.3 Edition Assembly Reference
Fix the U.2/U.3 Adaptor Base Assembly to the EDSFF Device Adaptor Base Tray using three M3 x 4mm Flat Head screws
POD Main Unit
U.2/U.3 Adaptor Base Assembly
EDSFF Device Adaptor Base Tray
M3x4mm Flat Head Screw
SerialTek

图 2-49

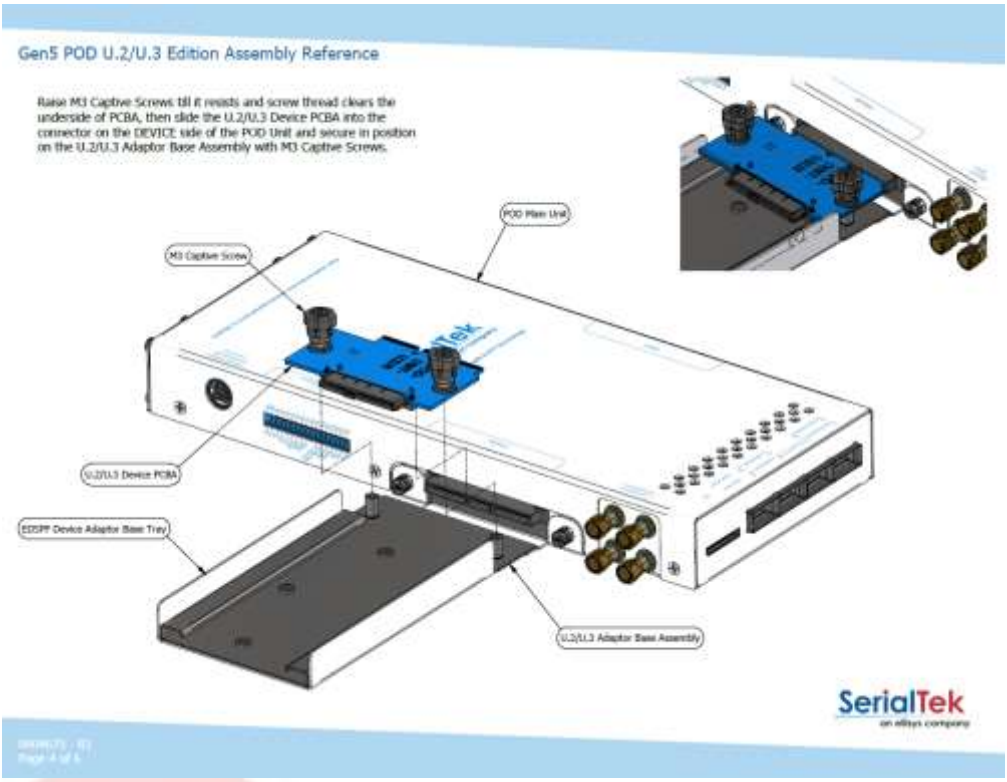

Gen5 POD U.2/U.3 Edition Assembly Reference
Raise M3 Captive Screws till it resists and screw thread clears the underside of PCBA, then slide the U.2/U.3 Device PCBA into the connector on the DEVICE side of the POD Unit and secure in position on the U.2/U.3 Adaptor Base Assembly with M3 Captive Screws.
POD Main Unit
M3 Captive Screw
U.2/U.3 Device PCBA
EDSFF Device Adaptor Base Tray
U.2/U.3 Adaptor Base Assembly
SerialTek

图 2-50

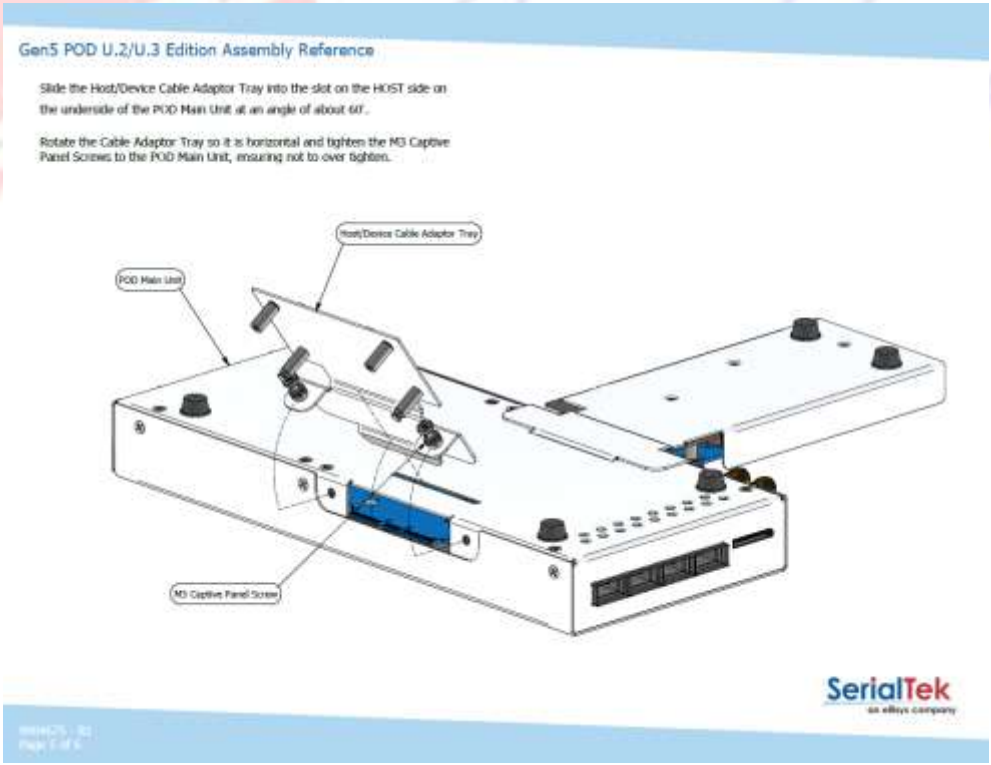

Gen5 POD U.2/U.3 Edition Assembly Reference
Slide the Host/Device Cable Adaptor Tray into the slot on the HOST side on the underside of the POD Main Unit at an angle of about 60°.
Rotate the Cable Adaptor Tray so it is horizontal and tighten the M3 Captive Panel Screws to the POD Main Unit, ensuring not to over tighten.
Host/Device Cable Adaptor Tray
POD Main Unit
M3 Captive Panel Screw
SerialTek

图 2-51

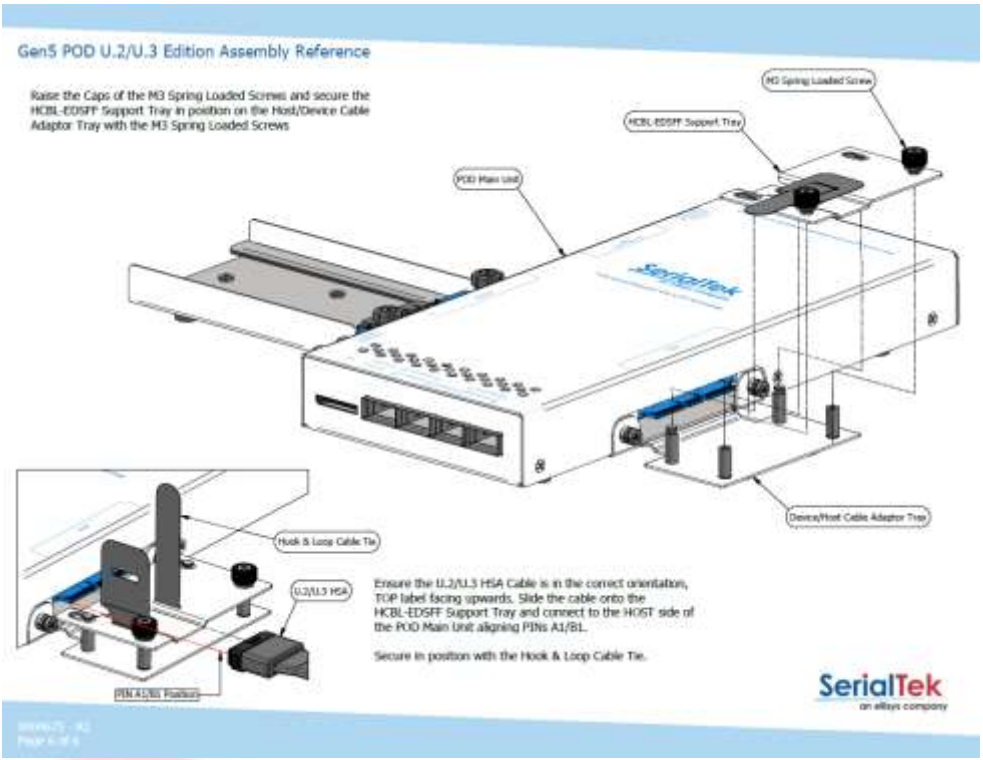

Gen5 POD U.2/U.3 Edition Assembly Reference
Raise the Caps of the M3 Spring Loaded Screws and secure the HCBL-EDSFF Support Tray in position on the Host/Device Cable Adaptor Tray with the M3 Spring Loaded Screws
M3 Spring Loaded Screw
HCBL-EDSFF Support Tray
POD Main Unit
Device/Host Cable Adaptor Tray
Hook & Loop Cable Tie
U.2/U.3 HSA
PIN A1/B1 Position
Ensure the U.2/U.3 HSA Cable is in the correct orientation, TOP label facing upwards. Slide the cable onto the HCBL-EDSFF Support Tray and connect to the HOST side of the POD Main Unit aligning PINs A1/B1.
Secure in position with the Hook & Loop Cable Tie.
SerialTek
an eBays company

图 2-52

Kyle McRobert
Hardware Engineer
Posted: 30th April 2019

With Big Data getting bigger all the time, there's undoubtedly a growing need for higher performance data storage solutions. Storage host controllers, specifically, need to be way more reliable under such intensive workloads.

U.3 is a 'Tri-mode' standard, building on the U.2 spec and using the same SFF-8639 connector. It combines SAS, SATA and NVMe support into a single controller. Where firmware support is available, U.3 can also support hot-swap between the different drives.

Here's a hardware engineer's perspective on how this controller is new, challenging and full of potential.

Key changes from U.2

Bringing it all together…

With U.2, you'd need a separate connector pinout/backplane, a separate mid-plane and controller for each protocol. U.3 only requires 1 backplane, 1 mid-plane and 1 controller, supporting all these drives in the same slot. This could be a great advantage, with SAS and NVMe forecasted to increase over the coming years—and SATA to decrease (according to OpenCompute).

Pinout changes

To configure itself to work with any of the above drives, U.3's extra pins determine which one is inserted. U.3 has two IfDet pins whereas U.2 only has one: U.3 needs two to allow for sufficient different combinations to identify the different drives. The host uses different combinations of PRESENT and the two IfDet pins to identify which drive is present.

This brings us to the main difference between U.3 NVMe drives compared to U.2 NVMe

drives: the pin-out for the data lanes are different, with all protocols using a common set of data pins. Then there's the Host Port Type pins (HPT0 and HPT1). These pins are mainly used by the U.3 Gen-Z and U.3 NVMe drives. Different combinations of these pins tell the drive which host is it inserted into.

U.3 drives are still backward compatible with U.2, though—the U.3 drives determine which host type it is inserted into (by using different combinations of HPT0 and HTP1 being open or ground), and configure the data lanes on power-up. U.3 drives can still act as dual-port NVMe drives in supporting systems.

U.2 drives are not compatible with U.3 hosts.

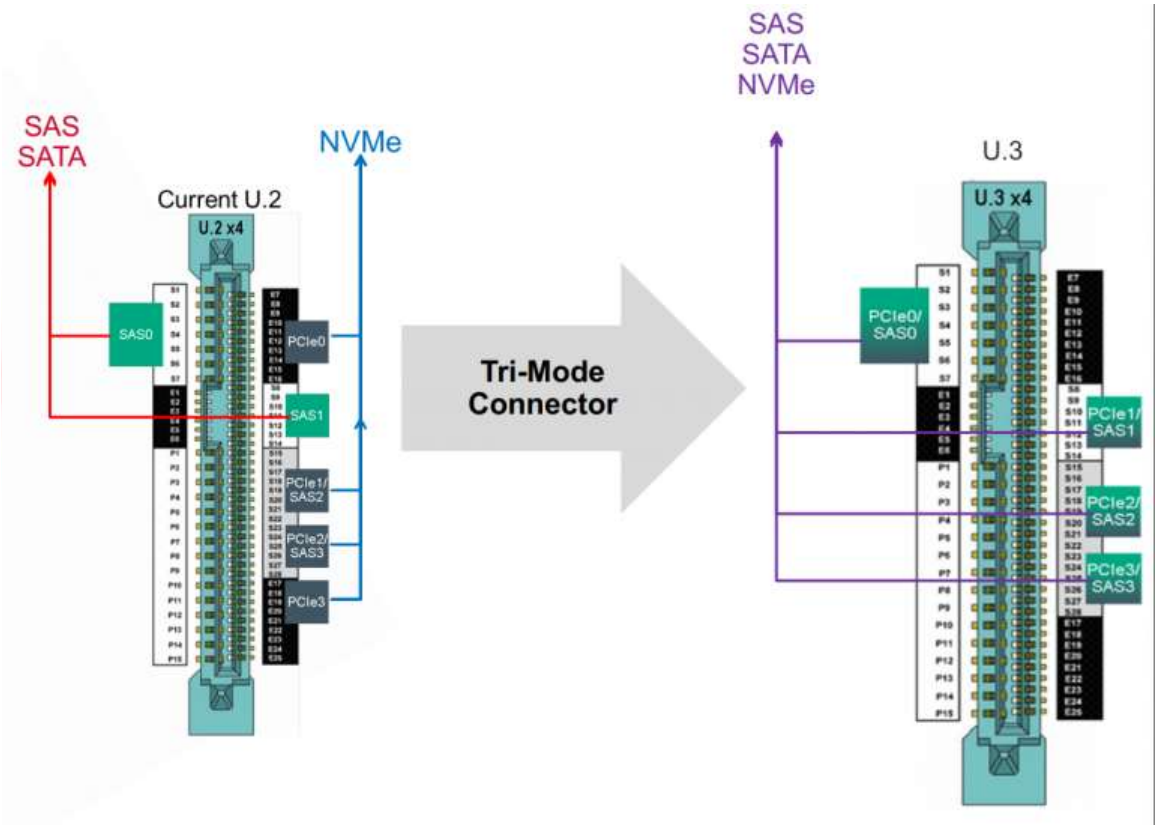


图 2-53

Image © OpenCompute

Challenges still remain—how do we overcome them?

While working with one of the companies leading the way in developing the U.3 host, we encountered a hurdle: compatible NVMe drives for U.3 are difficult to get hold of, with U.3 being so new to the scene. U.2 NVMe drives, unfortunately, don't work in U.3 because the data lane pin-out doesn't match. However, SFF-TA-1001 seems to be the key—it sets out what is needed for U.3 drives to work in U.3 while keeping the backward compatibility, meaning U.3 drives work in both U.2 and U.3.

At Quarch we're always adapting, and this is no exception! We worked to develop a small adapter which fits between the U.3 host and a U.2 NVMe drive, allowing the U.2 drive to work in the U.3 host without a problem. This allows hosts to have initial testing performed with existing U.2 drives.

We didn't stop there. We then added a flexible cable that allows the user to tap into—or break—the SM Bus for testing purposes.

What's next for U.3?

It'll be crucial to have reliable testing solutions for U.3 hosts and their compatible devices. Excitingly, I'm closely involved in developing this. We're now making a U.3 breaker module, which provides full hot-swap automation by allowing all data lines, sideband and power connections to be connected or disconnected either individually or in groups. It's also one of the first modules to support external triggering, and will soon be upgraded to include sideband monitoring!

For me, though, a challenge still remains: while reading the standard, the reader is told to refer to either the PCIe or SAS standards depending on which drive type you're interested in. The issue here is that the data lanes in PCIe have an impedance of 85 Ohms, while SAS is 100 Ohms. So, when making a U.3 host, which impedance is used?

It'll be interesting to see, in the future, whether manufacturers counter this problem by changing the impedance of these devices to bring them in line with one another.

2.6.1.3 Gen 5 M.2 Interposer

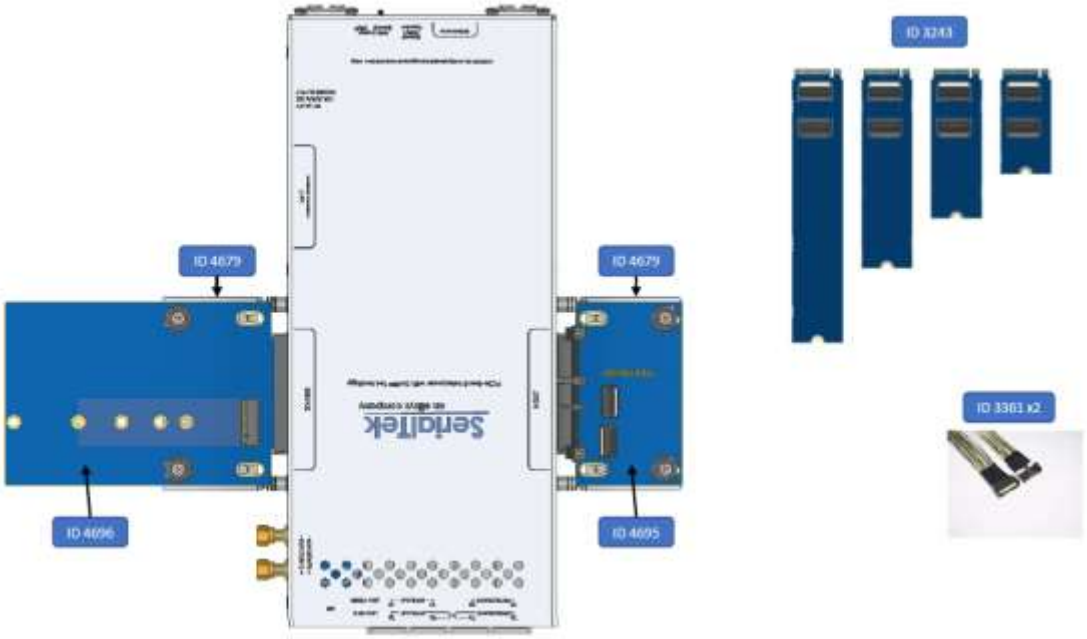


图 2-54

下图为实拍照片。

图 2-55

2.6.1.3.1 Gen5 Pod 组装示意图 – M.2

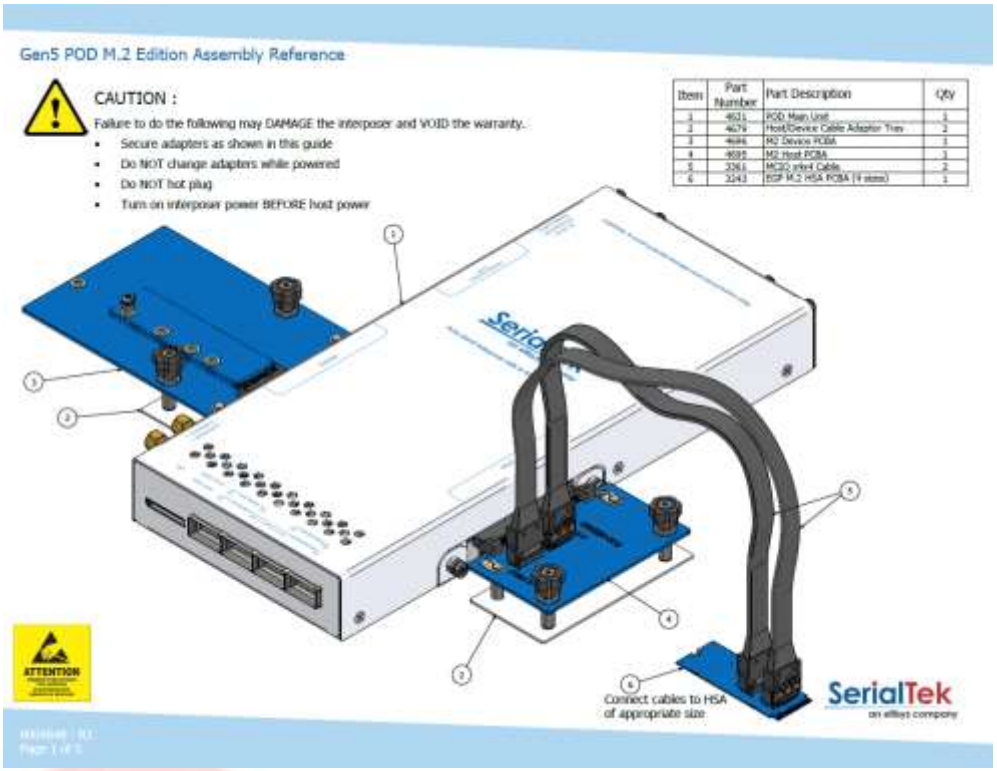

Gen5 POD M.2 Edition Assembly Reference
CAUTION :
Failure to do the following may DAMAGE the interposer and VOID the warranty.
Secure adapters as shown in this guide
Do NOT change adapters while powered
Do NOT hot plug
Turn on interposer power BEFORE host power
Item
Part Number
Part Description
Qty
1 4631 POD Main Unit 1
2 4679 Host/Device Cable Adaptor Tray 1
3 4686 M2 Device PCBA 1
4 4695 M2 Host PCBA 1
5 3261 MCIO x8H Cable 2
6 3243 EGF M.2 HSA PCBA (4 slots) 1
ATTENTION
Connect cables to HSA of appropriate size
SerialTek
an elitys company

图 2-56

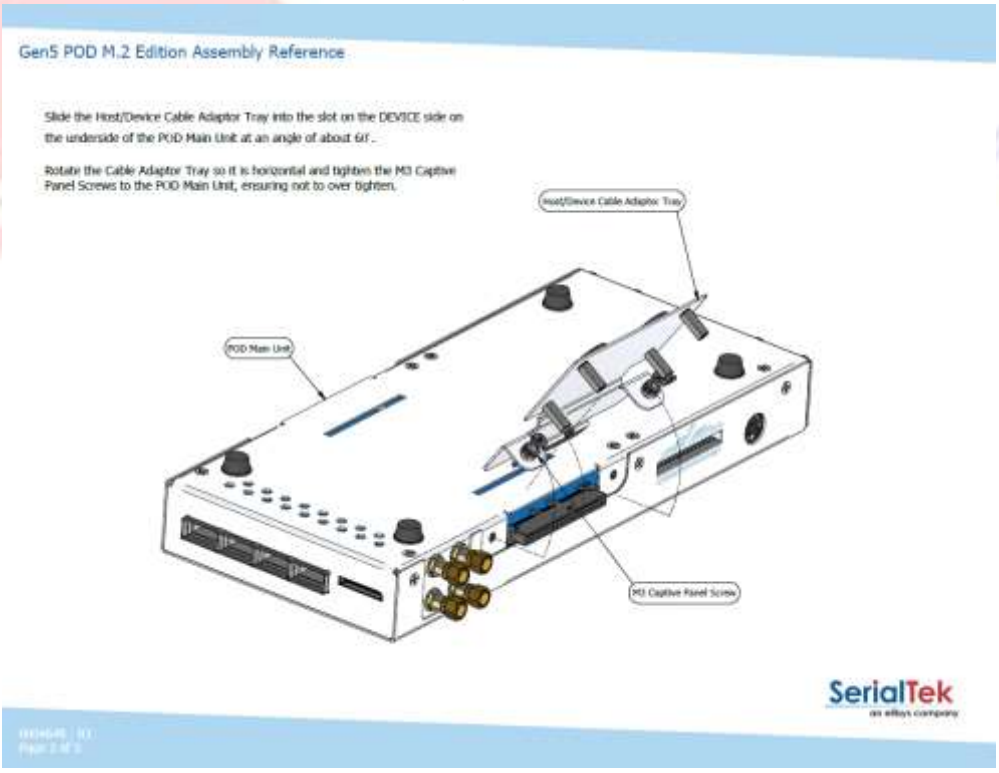

Gen5 POD M.2 Edition Assembly Reference
Slide the Host/Device Cable Adaptor Tray into the slot on the DEVICE side on the underside of the POD Main Unit at an angle of about 60°.
Rotate the Cable Adaptor Tray so it is horizontal and tighten the M3 Captive Panel Screws to the POD Main Unit, ensuring not to over tighten.
Host/Device Cable Adaptor Tray
POD Main Unit
M3 Captive Panel Screw
SerialTek
an elitys company

图 2-57

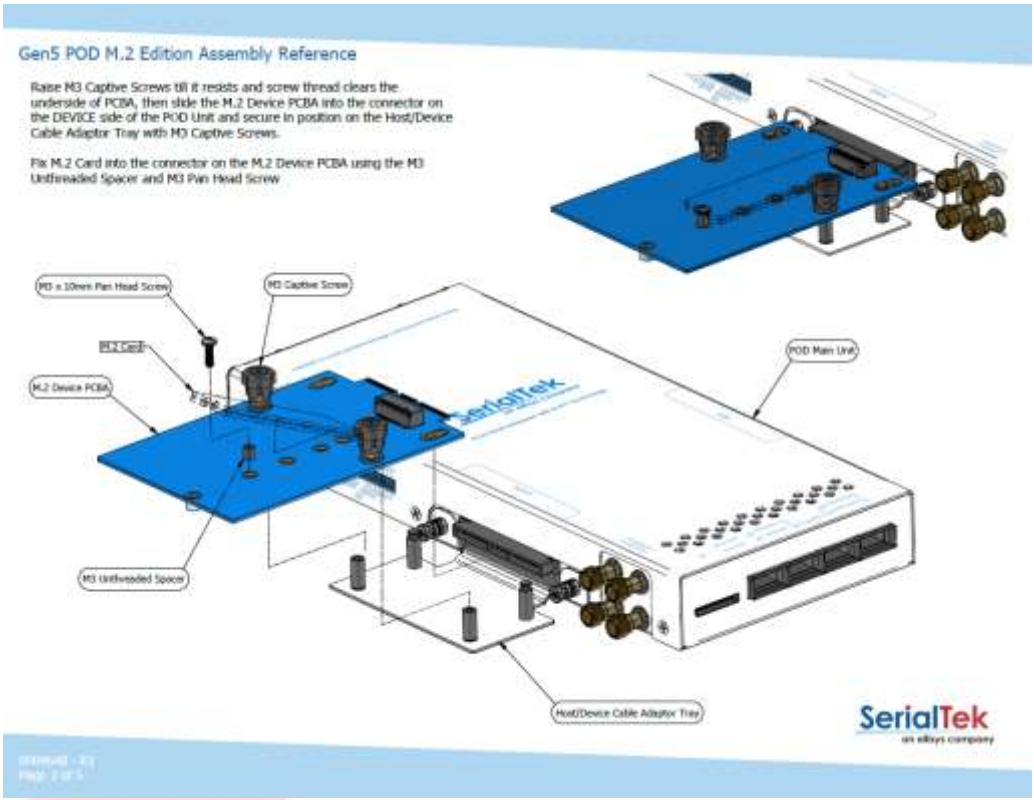

Gen5 POD M.2 Edition Assembly Reference
Raise M3 Captive Screws till it resists and screw thread clears the underside of PCBA, then slide the M.2 Device PCBA into the connector on the DEVICE side of the POD Unit and secure in position on the Host/Device Cable Adaptor Tray with M3 Captive Screws.
Fix M.2 Card into the connector on the M.2 Device PCBA using the M3 Unthreaded Spacer and M3 Pan Head Screw
M3 x 10mm Pan Head Screw
M3 Captive Screw
M.2 Card
M.2 Device PCBA
POD Main Unit
M3 Unthreaded Spacer
Host/Device Cable Adaptor Tray
SerialTek

图 2-58

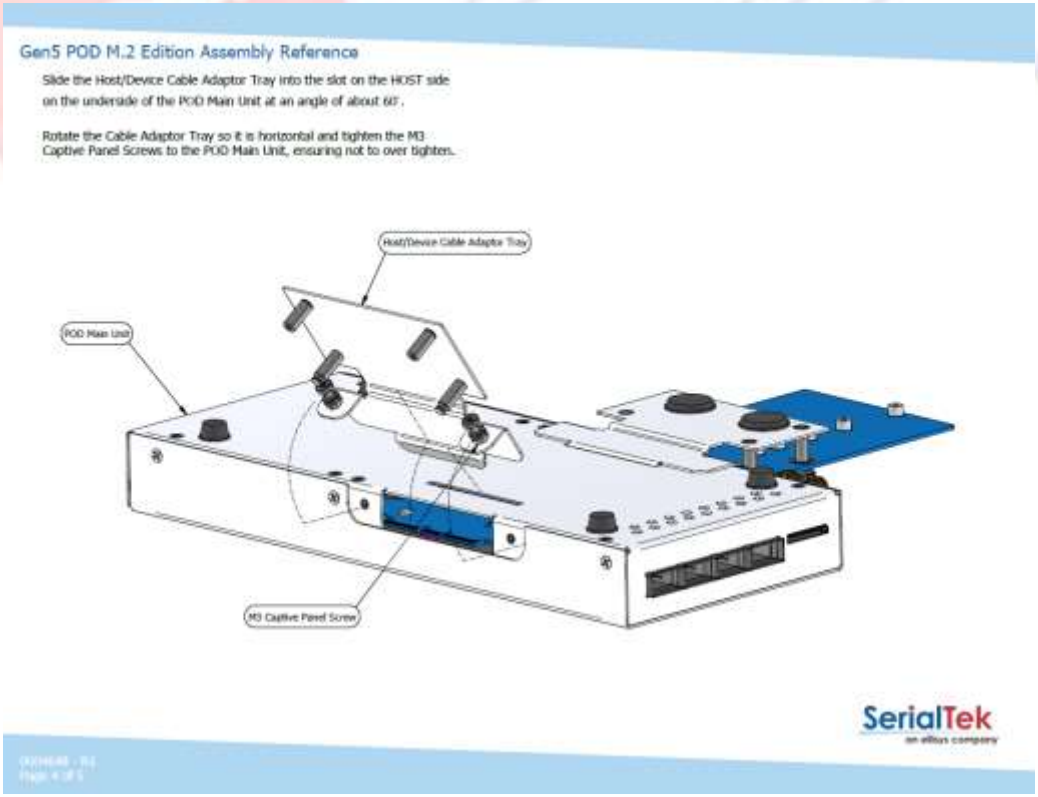

Gen5 POD M.2 Edition Assembly Reference
Slide the Host/Device Cable Adaptor Tray into the slot on the HOST side on the underside of the POD Main Unit at an angle of about 60°.
Rotate the Cable Adaptor Tray so it is horizontal and tighten the M3 Captive Panel Screws to the POD Main Unit, ensuring not to over tighten.
Host/Device Cable Adaptor Tray
POD Main Unit
M3 Captive Panel Screw
SerialTek

图 2-59

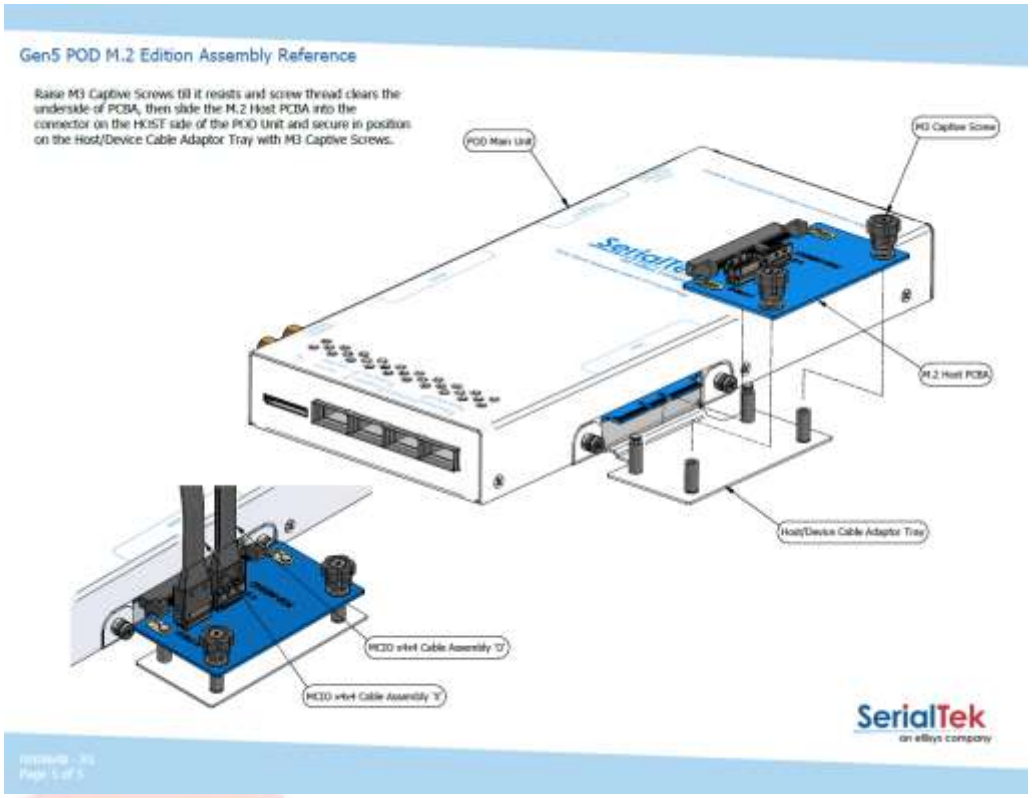


图 2-60

2.6.1.4 Gen 5 E1.S Interposer

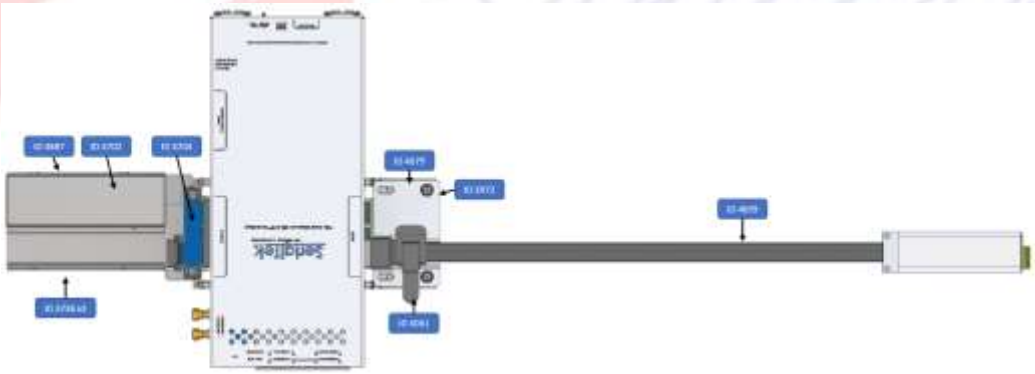


图 2-61

2.6.1.4.1 Gen5 Pod 组装示意图 – EDSFF 所有 form factor

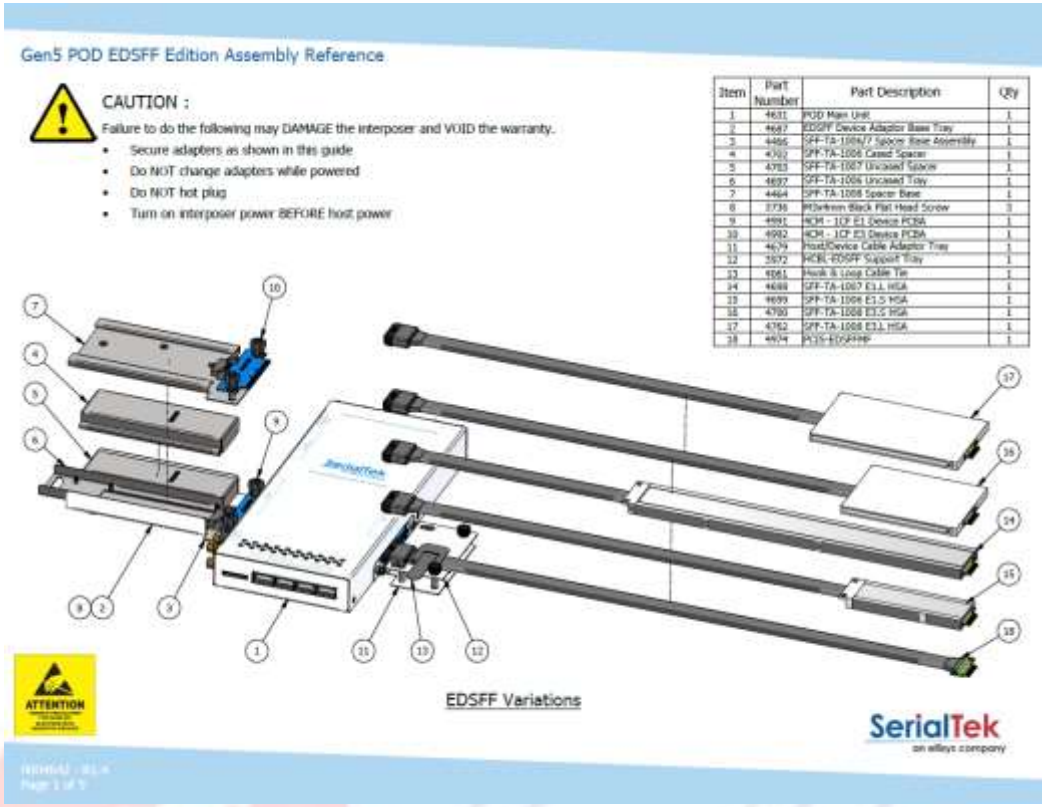


图 2-62

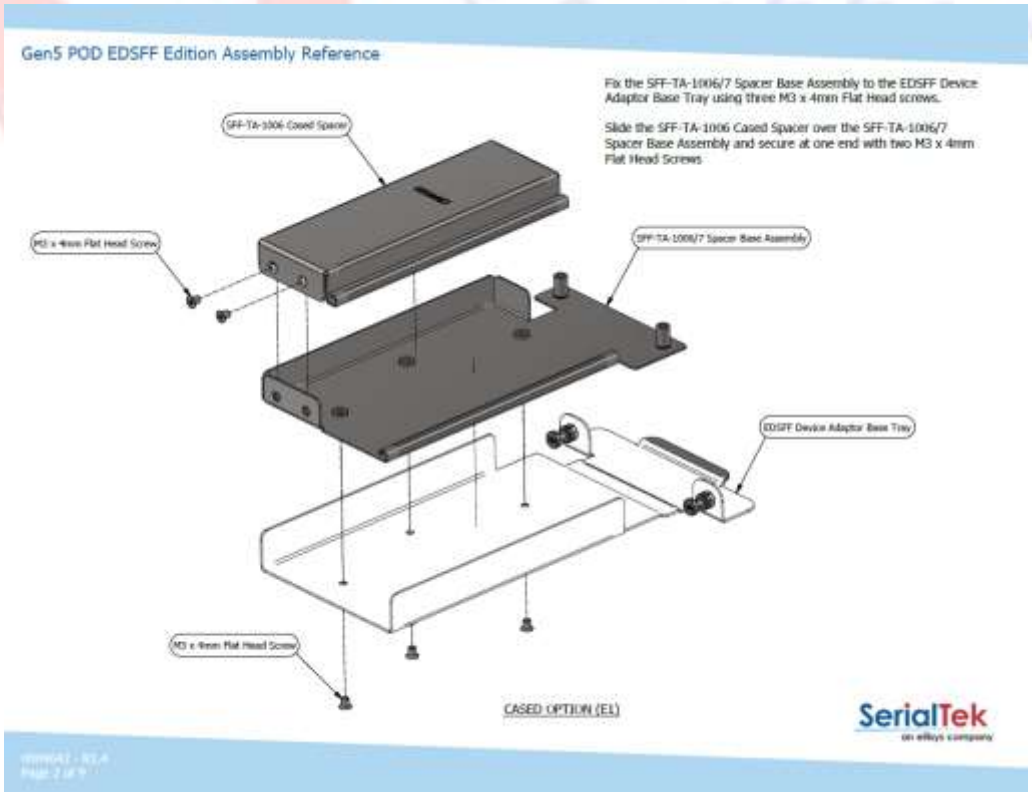


图 2-63

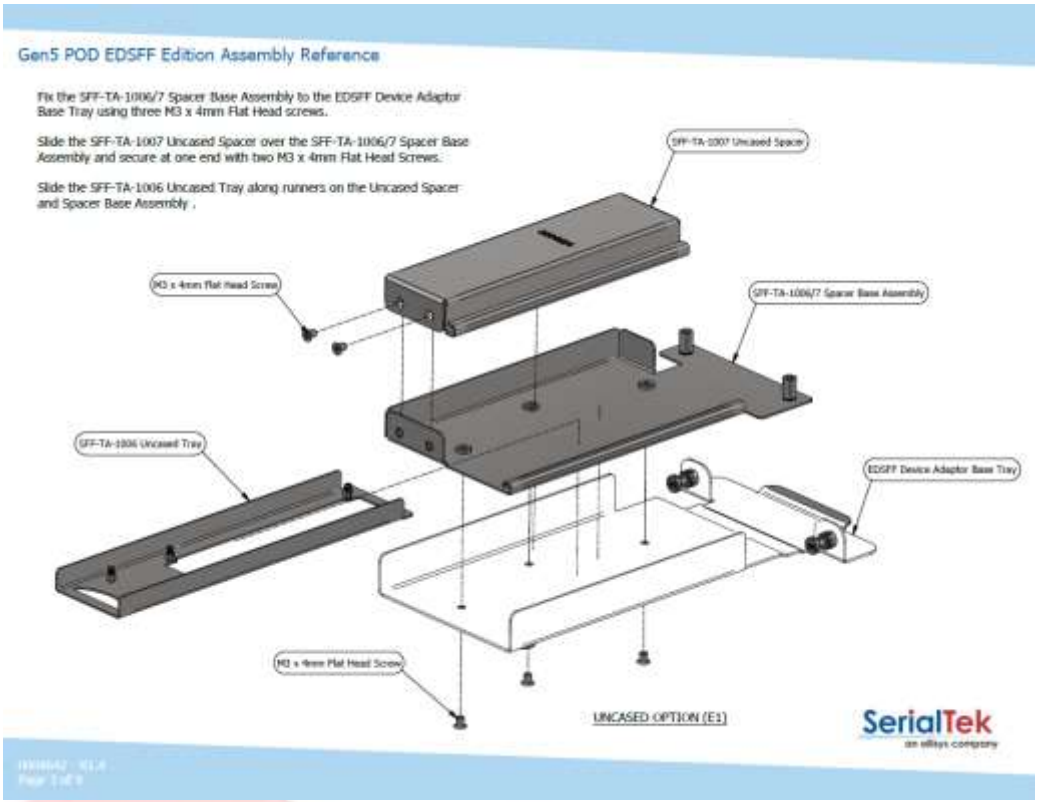

Gen5 POD EDSFF Edition Assembly Reference
Fix the SFF-TA-1006/7 Spacer Base Assembly to the EDSFF Device Adaptor Base Tray using three M3 x 4mm Flat Head screws.
Slide the SFF-TA-1007 Uncased Spacer over the SFF-TA-1006/7 Spacer Base Assembly and secure at one end with two M3 x 4mm Flat Head Screws.
Slide the SFF-TA-1006 Uncased Tray along runners on the Uncased Spacer and Spacer Base Assembly .
SFF-TA-1007 Uncased Spacer
M3 x 4mm Flat Head Screw
SFF-TA-1006/7 Spacer Base Assembly
SFF-TA-1006 Uncased Tray
EDSFF Device Adaptor Base Tray
M3 x 4mm Flat Head Screw
UNCASED OPTION (E1)
SerialTek

图 2-64

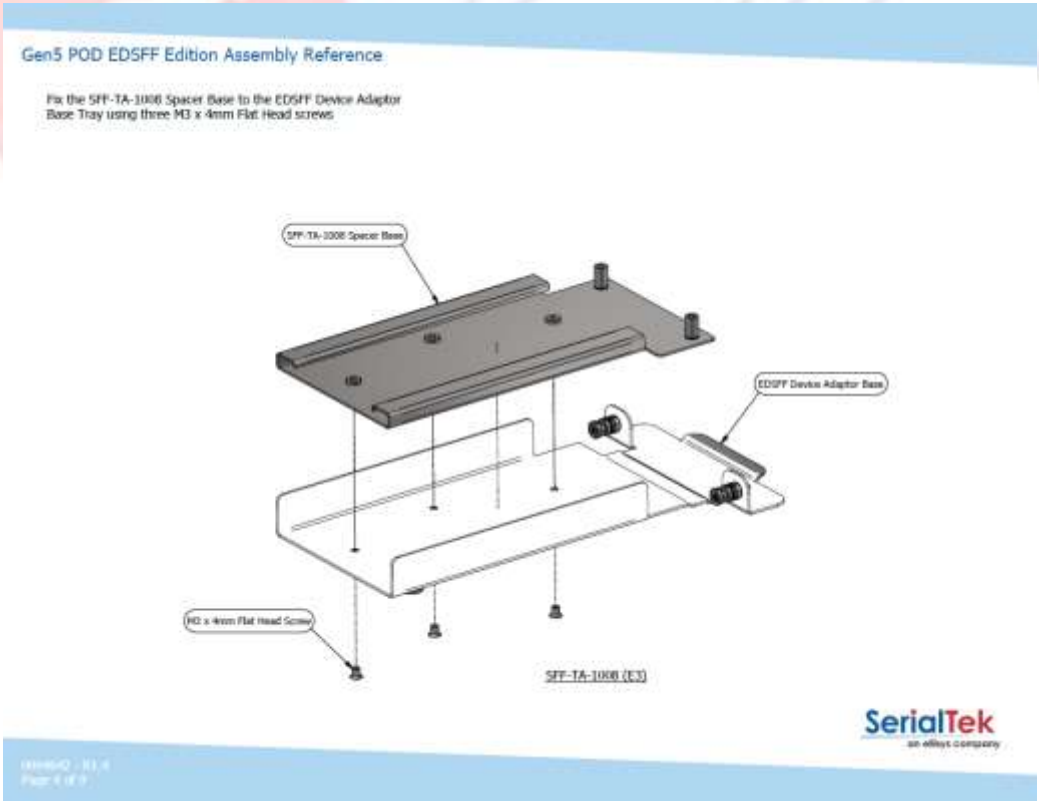

Gen5 POD EDSFF Edition Assembly Reference
Fix the SFF-TA-1008 Spacer Base to the EDSFF Device Adaptor Base Tray using three M3 x 4mm Flat Head screws
SFF-TA-1008 Spacer Base
EDSFF Device Adaptor Base
M3 x 4mm Flat Head Screw
SFF-TA-1008 (E3)
SerialTek

图 2-65


Gen5 POD EDSFF Edition Assembly Reference
Slide the EDSFF Device Adaptor Base Tray into the slot on the DEVICE side on the underside of the POD Main Unit at an angle of about 60°.
Rotate the EDSFF Device Adaptor Base Tray so it is horizontal and tighten the M3 Captive Panel Screws to the POD Main Unit, ensuring not to over tighten.
EDSFF Device Adaptor Base Tray
POD Main Unit
M3 Captive Panel Screw
SerialTek

图 2-66

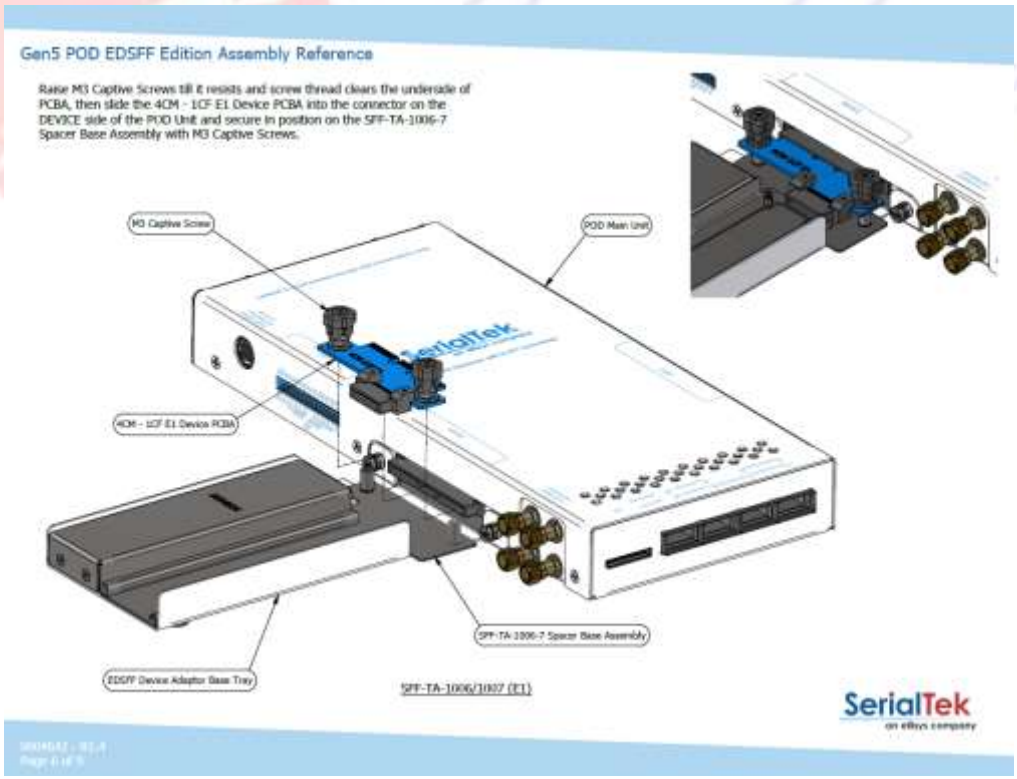

Gen5 POD EDSFF Edition Assembly Reference
Raise M3 Captive Screws till it resists and screw thread clears the underside of PCBA, then slide the 4CM - 1CF E1 Device PCBA into the connector on the DEVICE side of the POD Unit and secure in position on the SFF-TA-1006-7 Spacer Base Assembly with M3 Captive Screws.
M3 Captive Screw
POD Main Unit
4CM - 1CF E1 Device PCBA
SFF-TA-1006-7 Spacer Base Assembly
EDSFF Device Adaptor Base Tray
SFF-TA-1006/1007 (E1)
SerialTek

图 2-67

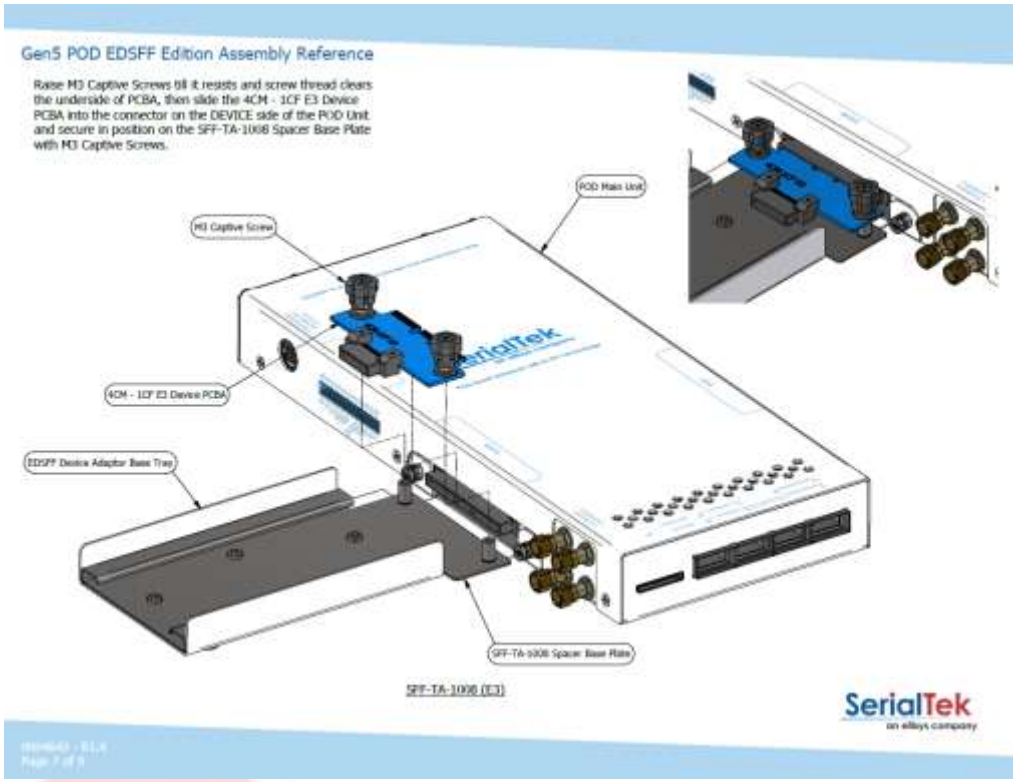

Gen5 POD EDSFF Edition Assembly Reference
Raise M3 Captive Screws till it resists and screw thread clears the underside of PCBA, then slide the 4CM - 1CF E3 Device PCBA into the connector on the DEVICE side of the POD Unit and secure in position on the SFF-TA-1008 Spacer Base Plate with M3 Captive Screws.
POD Main Unit
M3 Captive Screw
4CM - 1CF E3 Device PCBA
EDSFF Device Adaptor Base Tray
SFF-TA-1008 Spacer Base Plate
SFF-TA-1008 (E3)
SerialTek
an eBays company

图 2-68

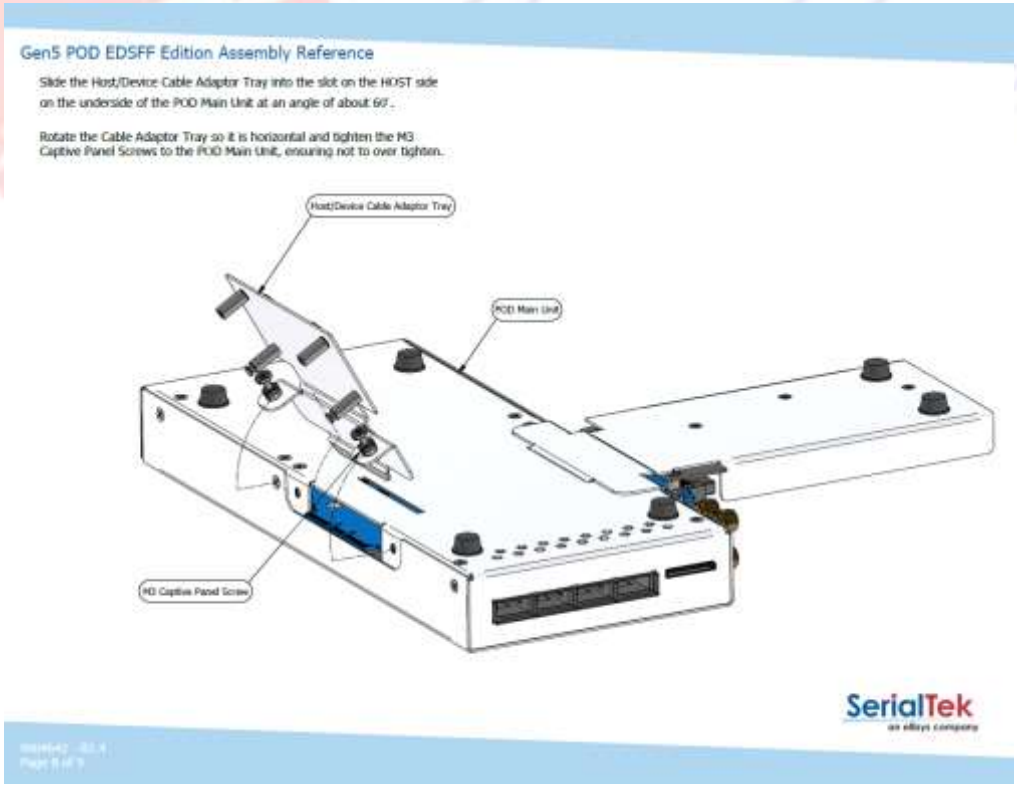

Gen5 POD EDSFF Edition Assembly Reference
Slide the Host/Device Cable Adaptor Tray into the slot on the HOST side on the underside of the POD Main Unit at an angle of about 60°.
Rotate the Cable Adaptor Tray so it is horizontal and tighten the M3 Captive Panel Screws to the POD Main Unit, ensuring not to over tighten.
Host/Device Cable Adaptor Tray
POD Main Unit
M3 Captive Panel Screw
SerialTek
an eBays company

图 2-69

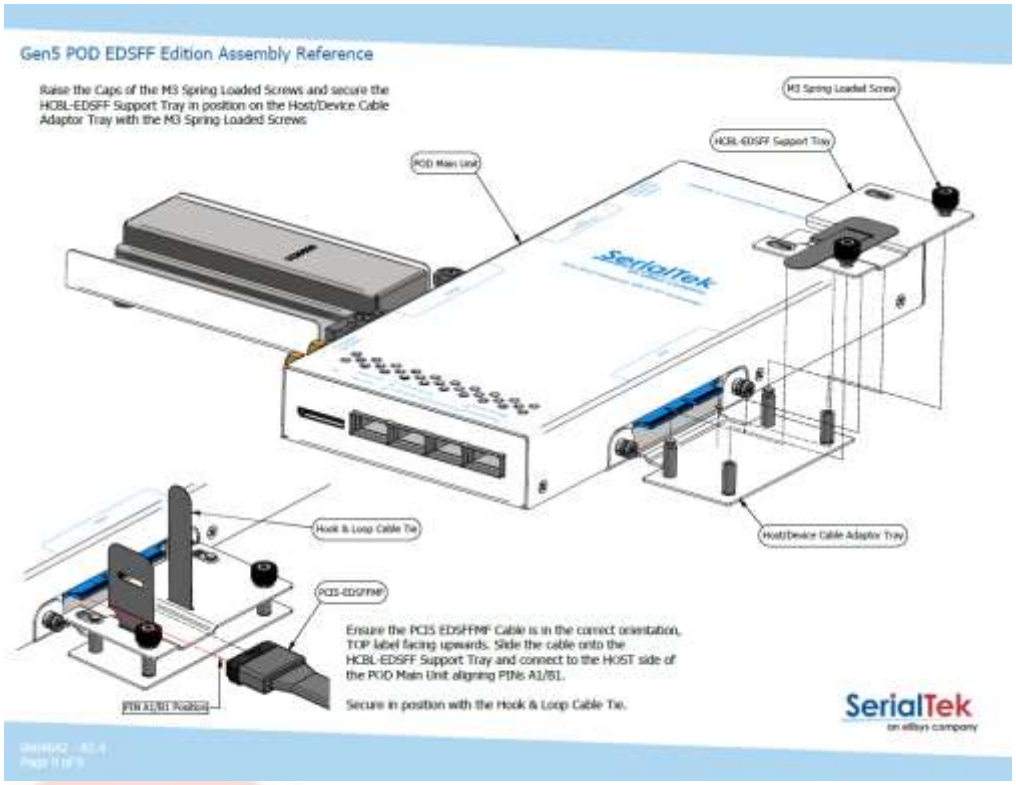


图 2-70

2.6.1.5 Gen 5 E1.L Interposer

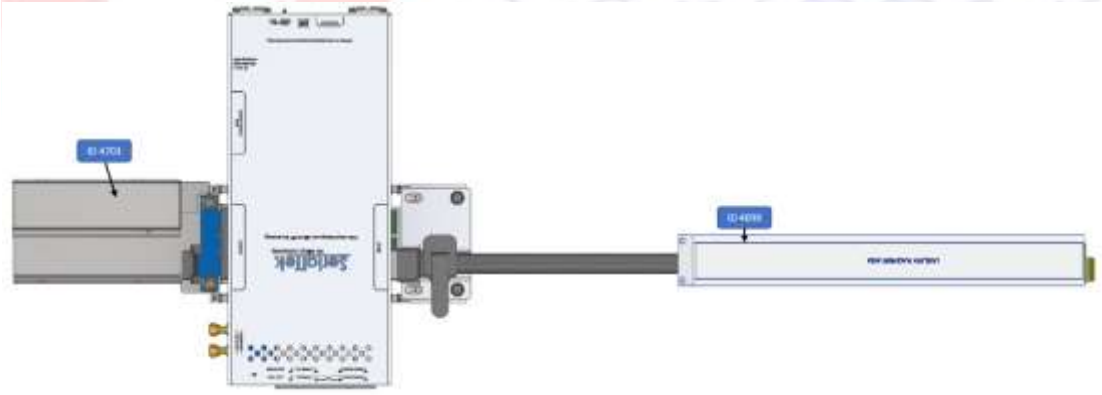


图 2-71

下图为实拍照片。

图 2-72

2.6.1.6 Gen 5 E3.S Interposer

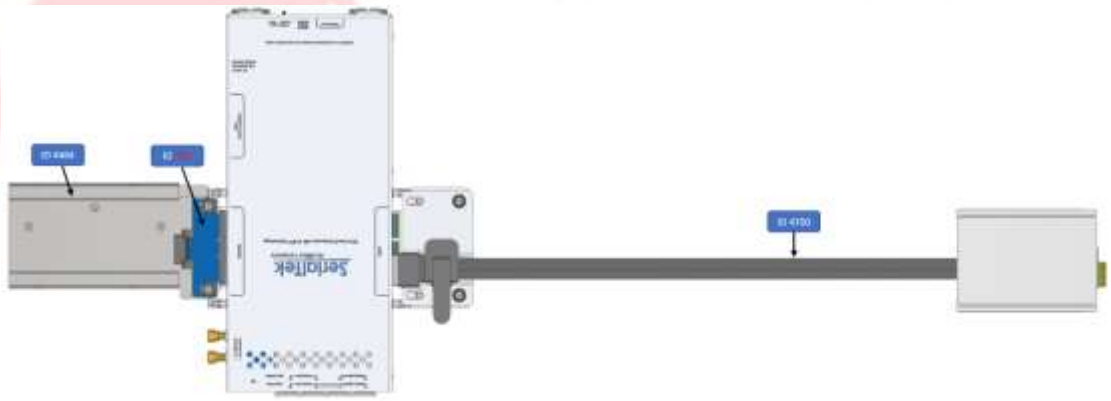


图 2-73

下图是实拍照片。

图 2-74

2.6.1.7 Gen 5 E3.L Interposer

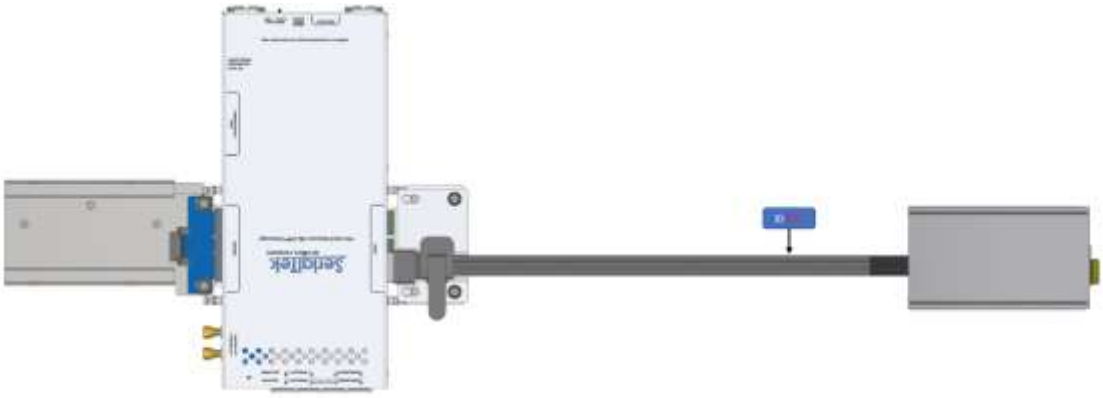
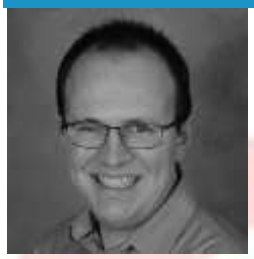


图 2-75

Why EDSFF could be a game changer for SSDs



Tom Pope

Hardware Engineer

Posted: 22nd April 2019

EDSFF is a new form factor developed for Solid State Drives. It stands for Enterprise and Datacenter Small Form Factor. Currently, it utilizes the PCIe NVMe protocol with potential plans to add in SAS and SATA functionality. So, what is different about EDSFF and why could this change a lot about datacenter information storage?

What is EDSFF?

EDSFF is a new form factor that aims to fix some of the shortcomings of existing interfaces that apply to SSD usage—and it is promoted by some of the biggest names in the industry.

SSDs, as opposed to conventional magnetic disk hard drives, require large PCB surface area and very little height. Trying to fit SSD storage into form factors designed for magnetic hard disks isn't very efficient: the two storage technologies have vastly different height requirements. EDSFF drives, by comparison to a 2.5-inch drive, are long and thin. This is much more space-efficient for SSD media, allowing greater memory density per rack while having excellent cooling Performance.

图 2-76

Example System Pictured in SFF-TA-1007 (Copyright SNIA. Reproduced with permission of SNIA.)

Why should you be interested in EDSFF?

There is one simple reason why you should be very excited by what EDSFF can offer. In a single 1u enclosure, you can store 32 drives. Placing an intel ruler in each bay, that gives you 983TB (using Intel 30.72TB ruler) of storage. In a 42U rack, that total is 41.3PB of storage in 1344 drives. Compare that to a 2U enclosure capable of fitting 24 U.2 drives from supermicro. A 42U rack can fit 504 drives, allowing 3.87PB (using intel 7.68TB U.2). EDSFF allows for a 10x increase in storage density while still using the Gen3 NVMe (Non-Volatile Memory express) standard for communication—which is faster than SAS3 (Serial Attached SCSI). Higher data densities can be achieved with 3.5" SAS/SATA drives but that higher data density is not as useful with the lower data rate compared to the Gen3 NVMe protocol.

Technical specifications of EDSFF

The EDSFF form factor is an up to 16-lane interface along a single edge finger. It is like a PCIe slot but smaller, being only 51mm wide for a x16 interface compared to 84mm on a x16 PCIe slot.

The current specification for EDSFF supports up to Gen4 PCIe using the NVMe protocol. The specification allows for three broad types of EDSFF:

- **E1.S, which utilizes a x84 interface and is designed to fit vertically in a 1u enclosure.**
- **E1.L, which utilizes up to a x8 interface, fits vertically in a 1u enclosure but is approximately 3 times the length of an E1.S.**

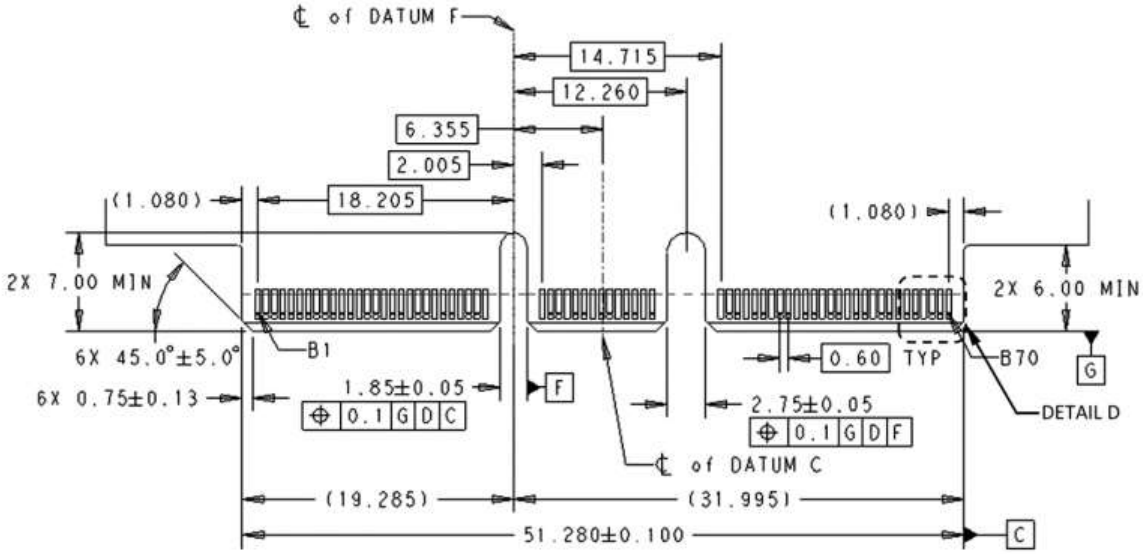


图 2-77

x16 Edge Finger Drawing Used in SFF-TA-1008 (Copyright SNIA. Reproduced with permission of SNIA.)

- **E3, which can use up to a x16 interface, fits vertically in a 2u enclosure or horizontally in a 1u enclosure and comes in four variants. It comes as either short (142mm) or long (105mm); then 7.5mm or 16.8mm thick.**

Each of these major types can support dual port and can also utilize a smaller width than the available maximum number of lanes. For example, an E3 drive can utilize a x4 interface.

EDSFF utilizes various sidebands to support PCIe features such as:

- **PERST# that acts as a functional reset for the drive.**
- **CLKREQ# allows the drive to request a reference clock.**
- **REFCLK, a reference clock that is provided to the drive. When in dual port mode, two independent reference clocks are provided for each port.**
- **PRSNT pins for lane width determination.**
- **SMBUS for low-speed communication.**
- **PWRDIS for notifying the drive to turn off everything reliant on the 12V supply.**
- **LED/ACTIVITY which is either an output to drive an LED on the drive or an input to signal the status of data transfer.**

The maximum power consumption of a drive is limited to 70W using only a 12V supply. The interface also includes a 3V3 supply designed for powering the sideband signalling. However, the 70W limit can be reduced depending on which type of EDSFF drive is used. The lowest power device is the E1.S at 12W while the highest power device is the E3.L 16.8mm thick being the only type allowed 70W.

What products do we provide?

Currently, we have 6 modules compatible with the EDSFF standards. These are in two different product ranges; our power injection fixtures and our breaker modules.

EDSFF Standard	Lane Width	Power injection	Breaker	Breaker with triggering
E1.S	X4	QTL2330	QTL2334	QTL2351
E1.L	X8	QTL2191	QTL2161	QTL2272

2.6.1.8 Gen 5 OCP Interposer

图 2-78

2.6.1.9 Cable Interposer – MCIO, HD Mini SAS, Slim SAS, Oculink connector

图 2-79

说明：上述 interposer 图片实际上代表了 4 个不同型号的 cable interposer，因为每种接口定义都不一样，但是基本结构都是一样的，所以采用同一张图片示例。

下图为 Gen5 MCIO cable interposer 实拍图。

图 2-80

下图为实拍 HD-MINI-SAS (SFF 8674) Cable interposer 的照片。

图 2-81

2.6.1.9.1 Gen5 Pod 组装示意图 – MCIO

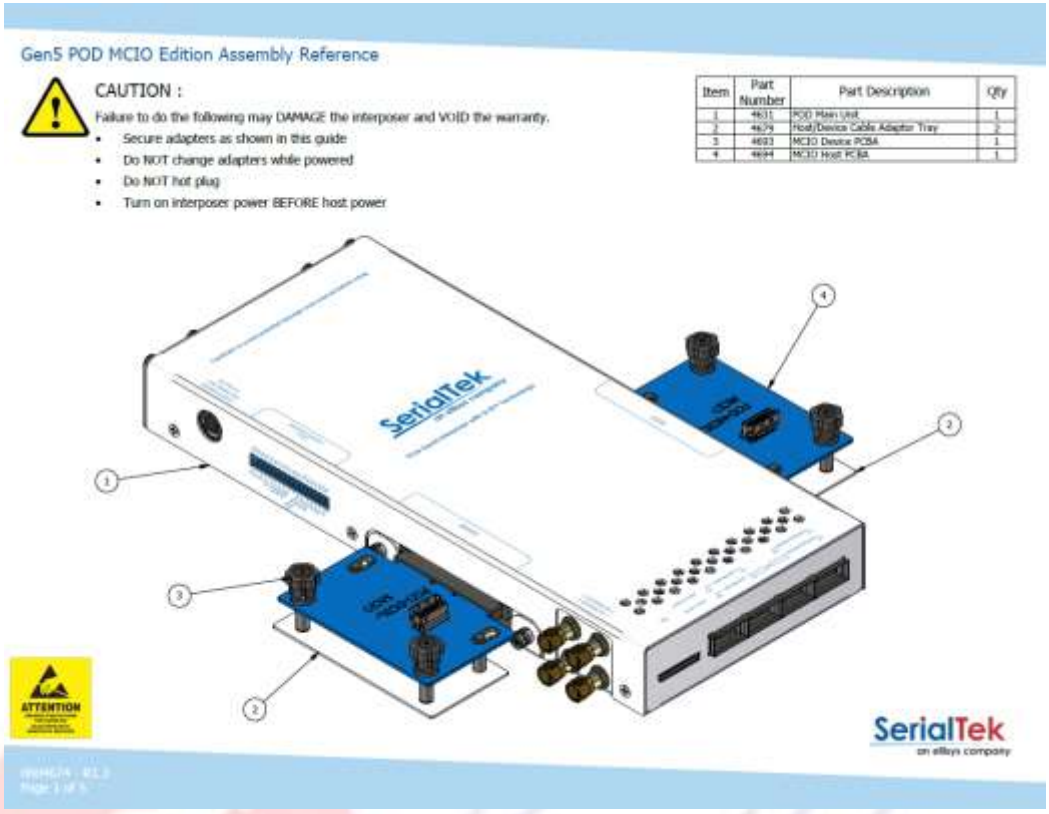


图 2-82

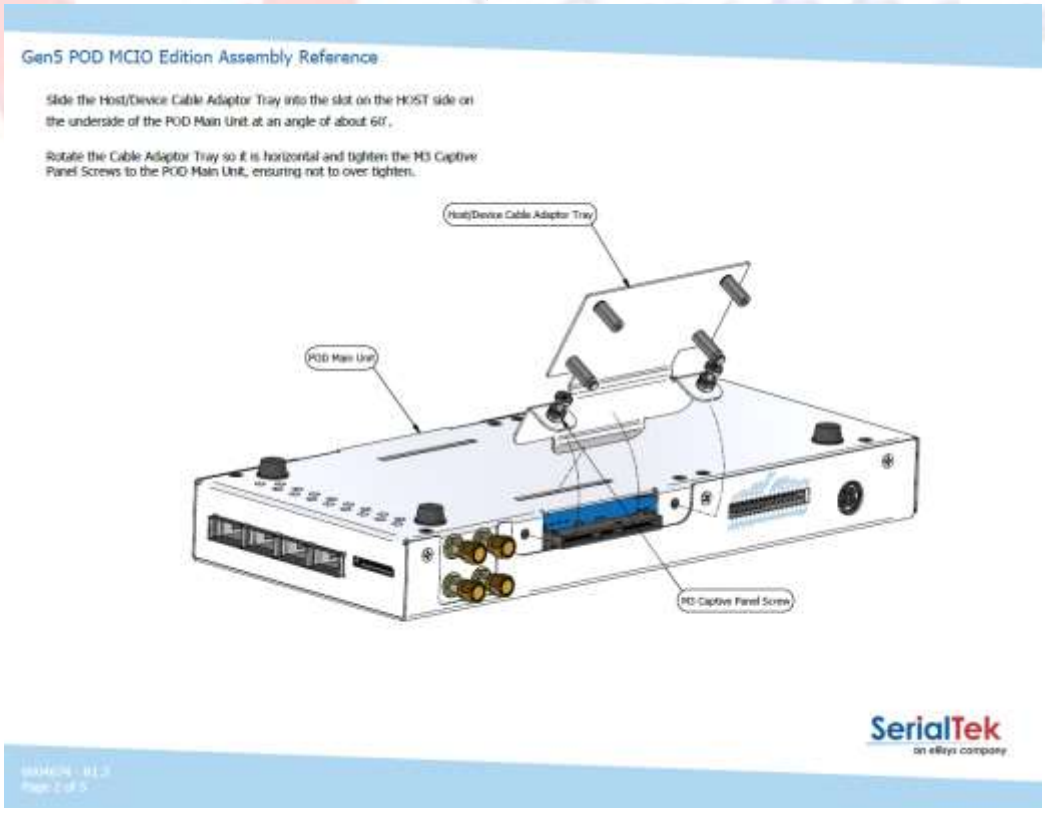


图 2-83

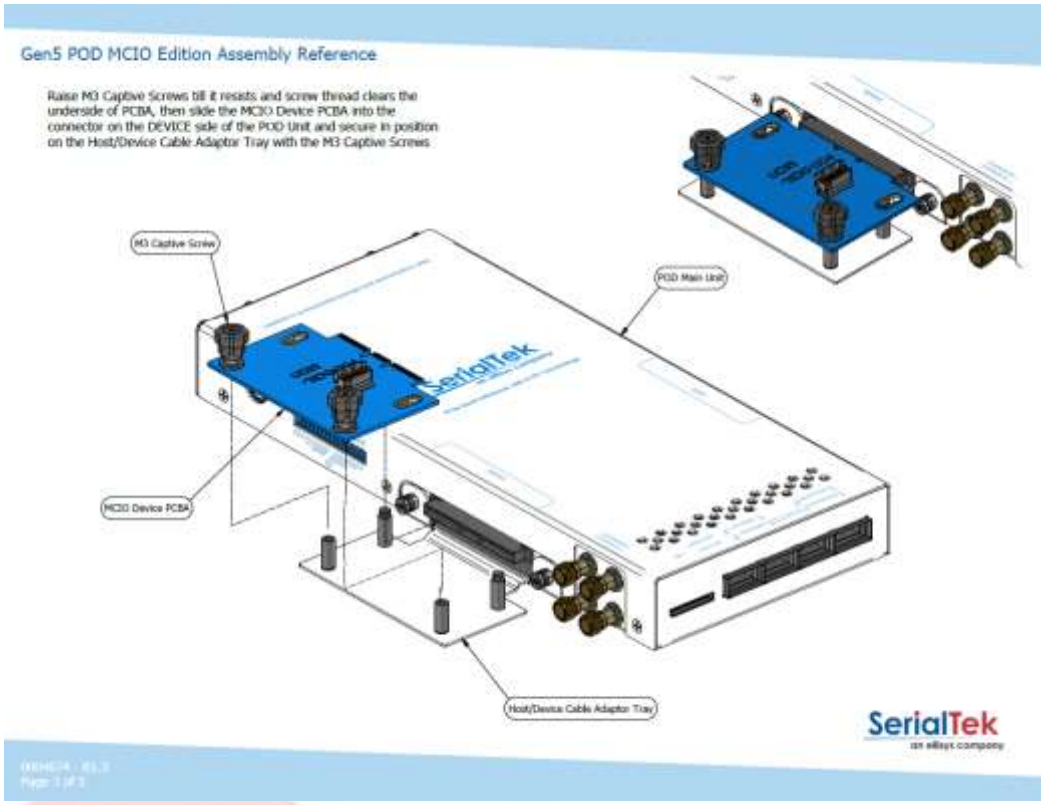

Gen5 POD MCIO Edition Assembly Reference
Raise M3 Captive Screws till it resists and screw thread clears the underside of PCBA, then slide the MCIO Device PCBA into the connector on the DEVICE side of the POD Unit and secure in position on the Host/Device Cable Adaptor Tray with the M3 Captive Screws
M3 Captive Screw
POD Main Unit
MCIO Device PCBA
Host/Device Cable Adaptor Tray
SerialTek

图 2-84

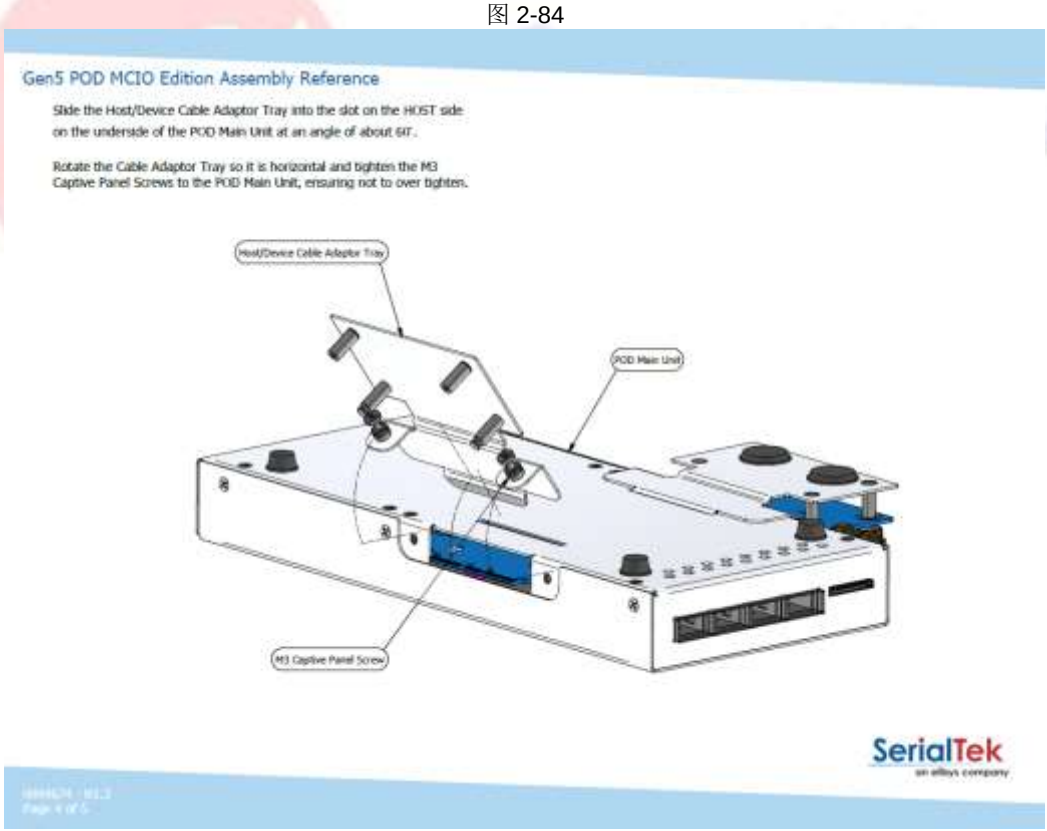

Gen5 POD MCIO Edition Assembly Reference
Slide the Host/Device Cable Adaptor Tray into the slot on the HOST side on the underside of the POD Main Unit at an angle of about 60°.
Rotate the Cable Adaptor Tray so it is horizontal and tighten the M3 Captive Panel Screws to the POD Main Unit, ensuring not to over tighten.
Host/Device Cable Adaptor Tray
POD Main Unit
M3 Captive Panel Screw
SerialTek

图 2-85

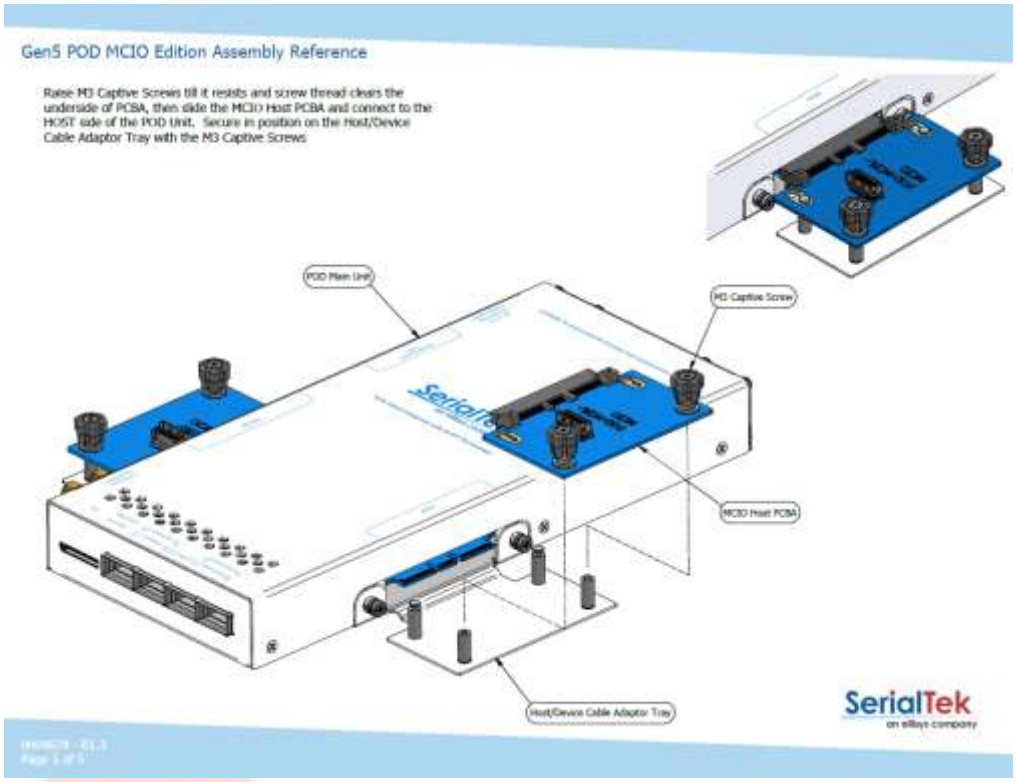


图 2-86

2.6.2 Gen4 协议分析仪 Interposer 展示

SerialTek PCIe Gen 4 analyzer 提供业内最常用的各种接口的 interposer，包括 AIC, U.2, U.3, M.2, E1.S, E1.L, Cable (HD-MINI-SAS)等接口的问题分析。

2.6.2.1 Gen 4 Slot Interposer

图 2-87

2.6.2.2 Gen 4 U.2/U.3 Interposers

U.2 & U.3 support in one interposer*

Access to all side-bands, including SMBus

图 2-88

2.6.2.3 Gen 4 M.2 Interposer

- X4 M.2 slot interposer
  - Uses dual x4 MCIO cables from HSA
- Access to all side-bands
  - Supports all form factors up to 110mm

图 2-89

2.6.2.4 Gen 4 EDSFF Interposer

In addition to being highly accurate electrically and simple to use like all SerialTek SI-Fi interposers, the new EDSFF interposer is mechanically modular and easily converts from E1.S to E1.L to E3.S form factors in a single unit, an industry first for E3.X. The included EDSFF host adapters are easy to change and plug into a host system. High-quality cabling (instead of lossy PCB material) from the interposer to the storage enclosure preserves signal quality while adding flexibility, saving customers money and providing for safe placement of the device under test (DUT) on a bench or in a test rack. Additionally, SerialTek's patent-pending tool-less EDSFF interposer tray easily converts to securely hold E1.S, E1.L, or E3.S SSD's.

图 2-90

2.6.2.5 Gen 4 Cable Interposer

图 2-91

2.6.3 Gen3 协议分析仪 Interposer 展示

图 2-92

对于一些嵌入式系统设计，或者要求传输速度不是很高的场景，PCIe Gen 3 analyzer 还有一定的应用场景，尤其是用户希望获得一台高性价的 PCIe Gen 1/2/3 协议分析仪的时候。下面是 Gen 3 analyzer 最常用的 interposer 展示。

2.6.3.1 Gen 3 Slot Interposer

图 2-93

2.6.3.2 Gen 3 U.2 Interposer

图 2-94

2.6.3.3 Gen 3 M.2 Interposer

图 2-95

2.6.4 顶级专业拉杆箱方便外携和快递（Gen4/5/6）

图 2-96

2.7 SerialTek PCIe Gen5 x16 协议分析仪简介

新的 Kodiak™平台将 SerialTek 的优势带入更多的计算和数据存储市场

作为 PCI Express®，NVM Express®，SAS/SATA 协议测试解决方案的领先提供商，SerialTek/Ellisys 正式发布 Kodiak 系列 PCIe Gen5 x16 协议分析仪，以及业界首个免校准的 PCIe Gen 5 x16 插卡（AIC）分析板卡（Interposer)和新的基于 Web 的 BusXpert™用户界面，这允许用户比以往更有效地管理和分析所抓取的 trace 数据。在 Kodiak 协议分析仪和 SI-Fi™分析板卡（Interposer)系列中增加了 PCIe Gen5 x16，为计算，数据中心，网络，存储，AI 和其它 PCIe Gen5 x16 应用程序带来了传统 PCIe 分析仪无法企及的分析功能和诊断效率。借助 SerialTek 久经考验的免校准 SI-Fi 分析板卡（Interposer)技术，Kodiak 创新的最先进设计，用户可以更容易地搭建协议分析仪硬件连接，更精确地捕获 PCIe 信号，更有效地分析抓取的 trace 数据。

图 2-97 Kodiak PCIe Gen5 x16 协议分析仪

Kodiak PCIe Gen 5 x16 分析仪的核心是其高性能硬件体系结构，该体系结构在捕获，搜索和处理加速方面提供了无与伦比的进步。接口响应能力显着提高，涉及大量数据的搜索速度更快，并且硬件筛选功能灵活而强大。

“一旦将其安装在客户测试环境中，Kodiak 的功能和优势将立即显现出来，” SerialTek 首席执行官 Paul Mutschler 说，“用户界面现代，易于使用且非常灵活。除易于设置和节省时间外，Interposer 免校准设计还支持PCIe 主机与Device 之间的“真实世界”中的各种PCIe 链路training 过程，从而使其更加准确。”

该分析仪支持 144GB trace buffer，并且可以将 trace 文件直接保存到分析仪的内部存储（最大 4TB），附加存储（USB3.2 或 PCIe OCuLink）中，或者可以通过两个 10GE（SFP +）端口或 1GE（ RJ-45）网络连接。**作为 PCIe Gen 5 switch 芯片的领先公司，Broadcom 和业内主流 CPU 厂商以及使用他们产品的早期服务器/存储系统厂商（early adopter），都在使用 SerialTek PCIe Gen 5 协议分析仪。**

“Broadcom 非常高兴 SerialTek 通过其创新的协议分析仪将其测试和分析解决方案扩展到 Gen 5 x16 PCIe 市场，为需要更多选择以进行更好的硬件分析的数据中心客户提供高级 PCIe 诊断功能，” Broadcom 数据中心解决方案组 IC 研发副总裁 Dan Roehrich 说道，“ PCIe Gen5 生态系统继续蓬勃发展，与 SerialTek 等公司合作，提供了使客户能够进行计算和更快接收对数据访问的尖端技术。”

PCIe Gen5 x16 SI-Fi 分析板卡（Interposer)大大增加了使用便利并简化了相关设置

SI-Fi 分析板卡（Interposer)设计扩展了产品的使用的覆盖范围，可以应用于各种应用场景测试，包括链路训练（LTSSM），电源管理，热插拔，重置以及物理通道特性可能会

更改的其它情况。新型 PCIe Gen5 x16 SI-Fi 分析板卡（Interposer)具有极高的电气精确性，并且易于使用。分析板卡（Interposer)旨在轻松安全地保护客户的设备，并通过高质量 QSFP-DD 电缆连接到 Kodiak 分析仪。使用时无需进行优化调整或者校准。Host 和 Device 信号通过分析板卡（Interposer)互联，透传 PCIe 链路训练并大大简化了相关的设置。市场上其它厂家的 PCIe 分析仪和分析板卡（Interposer）通常需要调整优化相关参数或校准，这可能会导致可靠性问题，因为 PCIe Gen5 链路训练序列可以动态发生，而不仅仅是在启动时发生。借助 SI-Fi 技术和 Kodiak 的自适应 EQ 功能，用户可以节省设置时间。如果链路特性发生变化（例如，热插拔或 NSSR），则 Kodiak 可以动态地跟踪这些变化，最终大大节省了用户测试的时间。

通过 Web 浏览器或新的 BusXpert Electron® App 进行 PCIe 协议分析

结合新的 BusXpert 用户界面，用户可轻松访问 Kodiak PCIe Gen5 x16 分析仪的所有功能。基于新的高性能软件框架和 RESTFUL API，BusXpert 与 Kodiak 硬件无缝集成。用户可以通过网络浏览器或基于 SerialTek 的 Electron®的新应用程序访问，同时 BusXpert 包括一套功能强大的触发器，过滤器和 trace 处理功能，以及一个新的用户界面，可实现快速，轻松和可靠的解码。用户可以实时协作处理 trace 文件，并远程验证分析仪和分析板卡（Interposer)的配置是否正确，包括 Interposer 电缆连接正确与否的识别，链接状态，记录状态等等。新的 RESTFUL API 使自动化变得简单高效，并提供了用于监控和捕获流量，统计分析和详细的搜索功能。Kodiak 的高级硬件设计还意味着，用户可以开始查看分析之前无需下载 GB 量级的 trace 文件，因为数据立即准备就绪。

2.7.1 KODIAK™ PCIE GEN 5 X16 协议分析仪

内置 12 核高性能 CPU+144G Capture Buffer 硬件设计，无需校准的 SI-Fi™信号捕获和自动均衡，内部 SSD 存储，触摸屏 LCD 和标准 PCIe 电缆的 PCIe 分析平台。

强大的分析功能

- **无需调整（校准），Kodiak 的 Rx 在所有数据速率下自动均衡输入 PCIe 信号（EQ）**
- **高性能 trace 处理架构和业内最快的性能，一秒钟实现所有 DLLP/TLP/NVMe 解码**
- **144GB trace 抓包缓存**
- **内部 4TB trace 闪存存储（SSD），允许多用户访问**
- **支持直接外置存储用于存储 trace，包括两个 PCIe OCuLink 端口和两个 USB 3.2 端口**
- **2 个 10GbE SFP +（光纤/铜缆）和 1 个 1GbE RJ-45 连接，可实现快速可靠的连接**
- **下载之前可以实时访问内存中抓取的 trace**
- **触摸屏 LCD 用于分析仪设置和状态查询**

最新的体系结构

Kodiak PCIe Gen5 协议分析仪代表了协议分析仪设计的最新技术。Kodiak 平台包括一系列高性能创新，而这些创新是通过先进的设计得以实现的，该设计摆脱了繁琐的旧有的需要将 trace 数据上传电脑然后通过本地电脑再解析的做法，而采用了超响应式高性能数据处理，由分析仪内部的 12 核高性能 CPU 直接解码完毕后将界面传给电脑客户端显示。接口响应能力显着提高，涉及大量数据的搜索速度很快，并且硬件筛选功能灵活而强大。

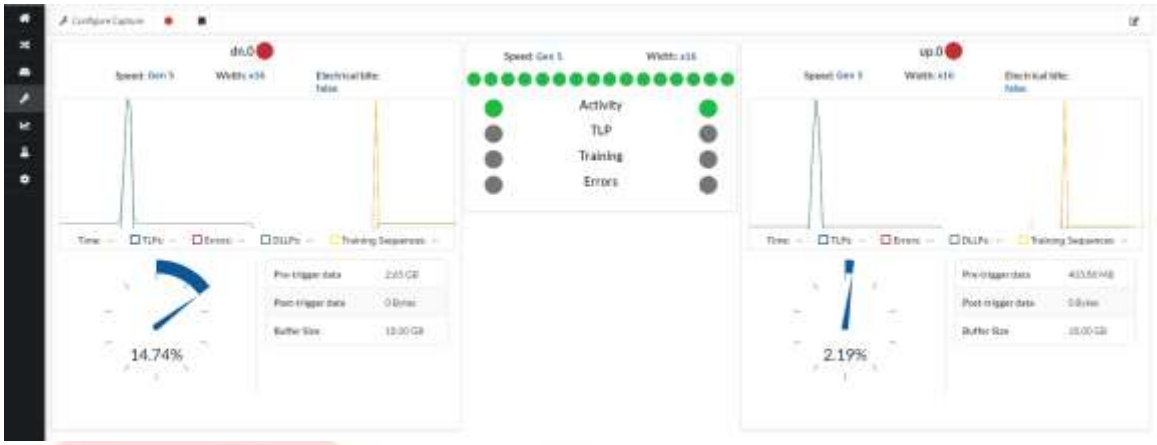


图 2-98

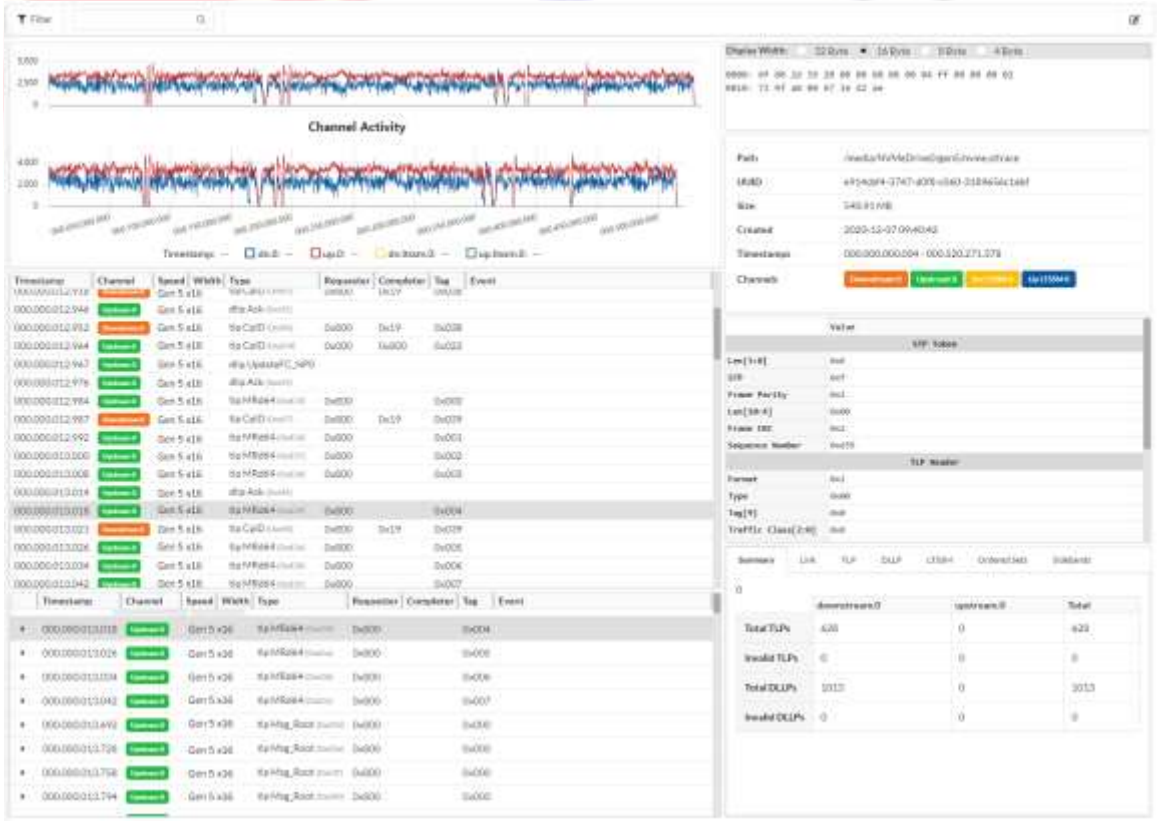


图 2-99

Interposer 设计的透明性是不断增长的数据传输需求的关键，PCI Express 信号的设计异常复杂且难以监视。借助 SerialTek 的 SI-Fi™分析板卡（Interposer)技术，来自一个

连接伙伴的发送器阈值和预加重到达另一个连接伙伴的接收器，因此链路可以正确地训练到最佳条件，从而使分析板卡（Interposer)尽可能透明。

这项技术的核心是高度专业化的线性放大器设计，在该设计中，PCIe 模拟信号在差分输入端接收，并分配给两个独立的相位匹配差分输出，标称理想化增益为 0dB。

这种方法可以简化分析仪和被测产品搭建环境时候的相关设置，并避免了其它厂商老的 interposer 设计带来的固有的各种限制，在那些老旧 interposer 设计方法中，链接训练序列（link training sequence）无法通过分析板卡（Interposer)。

SerialTek 的 SI-Fi™分析板卡（Interposer)技术可以扩展并覆盖常见的各种关键测试场景，包括链路训练（LTSSM），电源管理，热插拔，重置以及物理链路/通道特性可能会改变的情况。

灵活的 trace 存储和检索

Kodiak 包括两个 10GbE SFP +端口和一个 GbE 端口，如果需要的话，可以用于将 trace 上传到主机端保存，以及高达 4TB 的内部 SSD trace 存储（其它用户具有只读访问权限）。直连存储选择包括两个 USB 3.2 端口和两个 PCIe 3.0 OCuLink 端口。

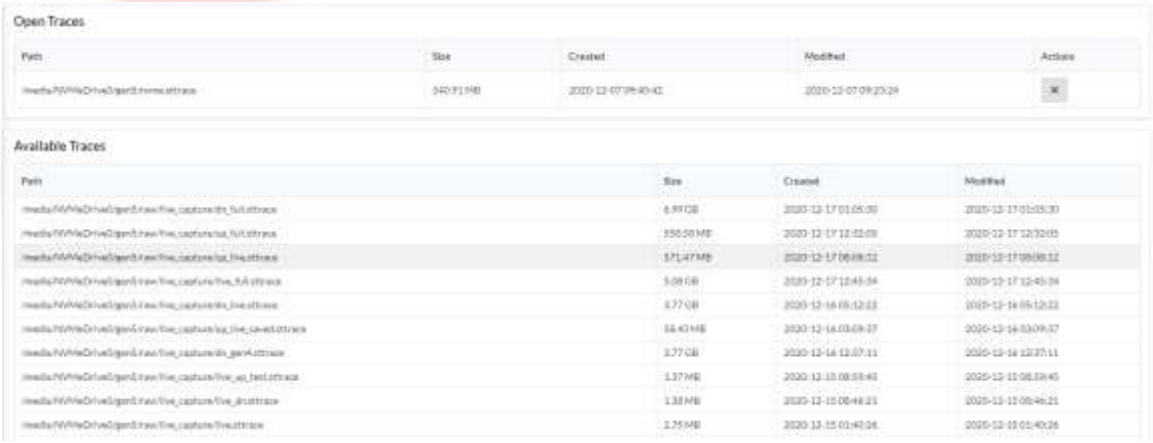


图 2-100

无需校准

业内其它厂商的 PCIe Gen5 分析仪和 Interposer 需要调整或校准，这会导致可靠性问题，因为现在的 PCIe 链路训练序列可以动态发生，而不仅仅是在启动时发生。PCIe Gen5 x16 SI-Fi 分析板卡（Interposer)具有高度的电气精确性，易于使用且可安全地固定到客户的设备上。无需调整（校准）。Host 和 Device 信号通过分析板卡（Interposer)，从而可以进行实际的 PCIe 链路训练并简化设置。借助 SI-Fi™技术和 Kodiak 的自适应 EQ 功能，用户可以大大节省设置的时间。而且，如果链路特性发生变化（例如，热插拔或 NSSR），Kodiak 可以动态地跟踪到这些变化，从而最终大大节省了测试时间。

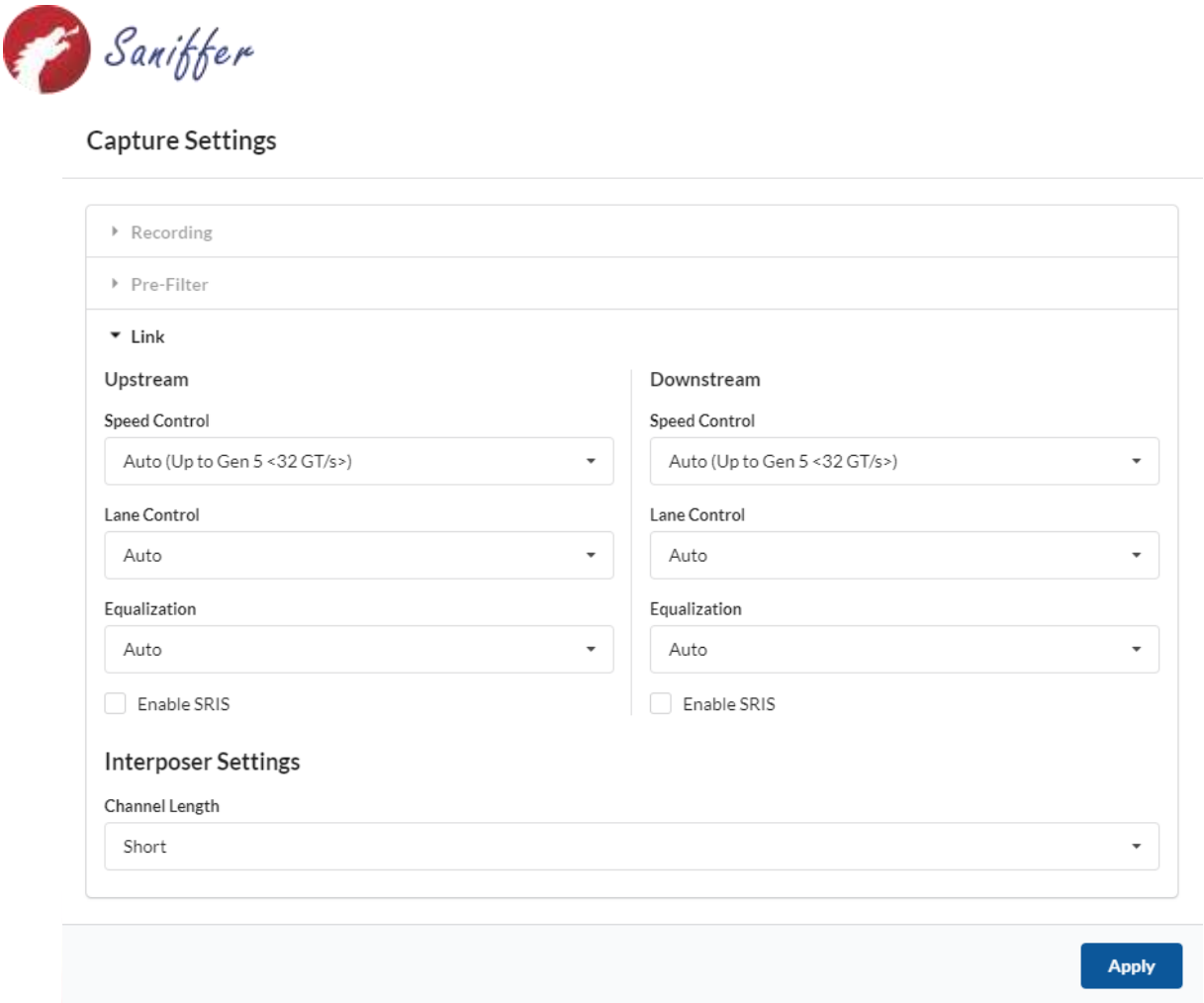


图 2-101

2.7.2 SI-FI™PCIe Gen5 分析板卡(Interposer)

带有 SI-Fi 技术的 SerialTek PCIe Gen5 Slot Interposer 以及其它各种接口的 Interposer 是专门设计的测试用适配器，物理上插在 PCIe Host 和 PCIe Device 插卡之间，将高速差分信号线以及其它一些 sideband signal 信号旁实时旁路到 Kodiak PCIe 协议分析仪。与 Kodiak™一起，SerialTek 的 Gen5（32.0 GT / s）PCI Express®（PCIe®）分析板卡（Interposer)与 SI-Fi™一起使用，使用户能够以无与伦比的强大功能和便捷的易用性分析各种 PCIe 总线流量。

通过 SerialTek 专有的 SI-Fi™技术，用户可以比使用老旧的 interposer 校准的传统方法大大节省了时间。

即使在变化的条件下，例如链路训练（LTSSM），电源管理，热插拔，重置以及其它可能改变物理链路/通道特性的测试，该技术也可通过提供高信号完整性来提高关键测试的覆盖范围。

SI-Fi™ PCIe Gen5 分析板卡（Interposer)继续了 SerialTek 的 TCO 方法。基于信号完整性，基于 QSFP-DD 的电缆将每个分析板卡（Interposer)连接到分析仪。这些电缆从市场上很容易获得，并且额定频率超过 32GHz，从而在所有 PCIe 传输速率下均具有不受影

响的 SI。所有边带 sideband 信号都通过分析板卡（Interposer)从 Root Complex（主机）传递到 End Point 外设，所有信号都提供给分析仪用于触发，解码和分析。

图 2-102

2.7.3 分析板卡（Interposer)主要功能

- **无需调整（校准）。主机和设备信号通过分析板卡（Interposer)，从而可以进行实际的 PCIe 链路训练并简化设置**
- **支持 PCI Express Gen 1.0、2.0、3.0、4.0 和 5.0**
- **访问包括 SMBus 在内的所有边带 sideband signal**
- **以 32.0 GT / s（Gen5）到 2.5 GT / s（Gen1）的线速准确捕获 PCIe 流量**
- **Interposer 的 Passive 信号获取方式可以避免掩盖，隐藏或清除电气和/或链接问题**
- **低成本，灵活，高性能的电缆，可实现可靠的分析仪与 Interposer 的连接**

2.7.4 AIC 分析板卡（Interposer)

PCI Express 插槽在计算，存储，网络和通信设备应用中普遍存在于 ATX 或基于 ATX 的主板上。SerialTek 的 PCIe Gen5 插槽分析板卡（Interposer)支持对 x1，x2，x4，x8 和 x16 链接宽度的分析。支持所有相关的边带 sideband signal，包括来自主机或来自外部/第三方注入或生成工具的 SMBus。

PCIe Gen 5 其它各种接口的特性和AIC 基本一致，这里以AIC 或者说Slot Interposer 为例简单讲述一下。

2.7.4.1 AIC 分析板卡（Interposer)概述

- 尺寸：309 x 167 x 31 毫米（12 x 6.5 x 1 英寸）
- Power connector: 4 Pin Mini DIN
- 分析仪连接器：QSFP-DD
- 设备连接器：PCIe CEM 插槽 x16 跨装接口
- 主机模块连接器：PCIe CEM x16 边缘指状件
- SMBUS 注入连接器：2×5 针 0.1 英寸接头，3.3 Vdc
- REFCLK 输出连接器：2 个 U.FL，交流耦合 LPHCSL
- REFCLK 输出控制连接器：2 针 0.1 英寸接头连接器
- REFCLK 缓冲器控制连接器：3 针 0.1 英寸接头连接器
- 边带信号访问连接器：2×9 针

2.7.4.2 AIC 边带(sideband)信号

- JTAG_TRST＃
- JTAG_TMS
- JTAG_TDO
- JTAG_TDI
- JTAG_TCK
- PRSNT2＃_2
- PRSNT2＃_1
- PRSNT2＃_0
- PRSNT＃
- GND
- RSVD3（仅 x8）
- RSVD2
- RSVD1
- SMBDAT
- SMCLK
- PWRBRK＃
- CLKREQ＃
- WAKE#
- GND

技术指标

Kodiak 机箱

- 尺寸：443 x 67 x 305 毫米（17 x 2.6 x 12 英寸）
- 重量：7 公斤（15 磅）
- 安装：19 英寸机架安装选件，倾斜脚选件
- 环境工作温度：5 - 35°C，最高海拔 2133m（7000 英尺）

显示和指示灯

- 前面板 LCD：800×320 4.6 英寸 WCGA，触摸屏
- 系统状态：RGB LED

前面板连接器

- Interposer 连接：4 个 QSFP-DD
- 以太网（10 GbE）：2 个 SFP +（10 GbE）
- 以太网（1 GbE）：RJ45
- PCIe 接口：2 个 OCuLink
- USB 接口：2 个 USB 3.2 Type A

后面板连接器

- 电源：IEC C13，90-264 Vac，47-63 Hz
- 时钟输出：SMA，50Ω，3.3 Vdc，10 MHz
- 时钟输入：SMA，50Ω，3.3 Vdc，10 MHz
- 触发输出：SMA，50Ω，3.3 Vdc
- 触发输入：SMA，50Ω，3.3 Vdc
- 维护：RJ45，USB Micro-B

2.8 Broadcom Gen 5 switch 芯片内嵌 Serialtek 协议分析功能

图 2-103

2.8.1 SerialTek 的 BusXpert iTAP 框架介绍

Broadcom 最新的 PCIe Gen 5 switch 芯片内置了 SerialTek PCIe 抓包分析功能，可以大大方便服务器、存储系统厂商调试 PCIe Gen 5 在初始化等各环节可能遇到的各种疑难问题，工程师只需要免费下载 SerialTek PCIe 协议分析仪软件 BusXpert 即可直接配置 PCIe Gen 5 switch 进行抓包分析，可以分析 upstream 以及 downstream 的流量；当然，用户也可以购买专业的 SerialTek PCIe Gen 5 x16 或者 x8/x4 协议分析仪实现使用非 Broadcom PCIe Gen 5 switch 或者其它各种应用场景的抓包分析。

SerialTek 的 BusXpert iTAP 框架是 Broadcom Gen5 PEX89000 PCI Express 嵌入式分析仪技术的基石，可提供片上针对 PCIe 协议的更深入的分析和诊断功能

SerialTek 是 PCIe，NVMe 和 SAS/SATA 协议测试解决方案的全球领先提供商，今天推出的 PCIe®测试和分析市场的最新技术和产品 BusXpert™iTAP™，支持 Broadcom 的 PCIe 嵌入式分析仪技术，这是 Broadcom 新的 PEX89000 Gen5 PCIe 交换芯片的一项突

破性功能。 BusXpert iTAP 基于一种方便的“TAP”即 PCIe Packet 侦听架构，与用于 Kodiak™ PCIe Gen5 x16 协议分析仪的 SerialTek BusXpert 分析仪软件集成在一起，使 PCIe 开发和测试工程师能够设置，捕获，解码和查看 Broadcom Gen 5 Switch 芯片上捕获的 PCIe packet。

使用 BusXpert iTAP 对 Broadcom PCIe 嵌入式分析仪进行更深入的分析

借助 SerialTek 的 Kodiak PCIe 分析仪和 BusXpert 协议分析软件，用户可以非常高效地分析 PCIe 协议层的各种问题，可以“秒级”解码 144GB buffer 所捕获的所有 DLLP/TLP/NVMe 层 Packet，同时实现了信号的高保真以及无需信号校准功能。但是，在当今高度集成的系统中，存在交换芯片和许多芯片间互连的系统，从源头访问所有可能的 PCIe 互连的物理限制可能会给 PCIe 的问题分析带来挑战。 BusXpert iTAP 通过利用并直接连接到 Broadcom Gen 5 switch 内部的 PCIe 嵌入式分析仪来解决此问题，从而实现了无需专业的硬件分析仪即可实现问题分析的功能。

“ Broadcom 有先见之明，他们知道需要与协议分析仪公司合作，以使这种嵌入式分析仪尽可能高效有用。” SerialTek 首席执行官 Paul J. Mutschler 说，“凭借我们在 PCIe Gen 5 协议分析和在 WebUI 界面方面的最新技术，我们显然是 Broadcom 芯片内置协议抓包工具的首选合作伙伴。” Broadcom 将这项技术集成到其 PCIe Gen5 PEX89000 交换芯片中，并支持以下强大的功能：

- **BusXpert iTAP –与 BusXpert 集成在一起，用于常见的 PCIe trace 分析仪**
- **嵌入式分析仪–无需外部硬件，消除了 interposer 分析板卡插入带来的挑战**
- **高级触发功能，可编程过滤，数据压缩–功能齐全的嵌入式解决方案**
- **并行分析–嵌入式分析仪能够触发其他实例以允许同步并行分析。**
- **带内（PCIe）和带外连接（JTAG）**

“ Broadcom 致力于满足市场对军工和 AI 人工智能应用的需求，提供先进的 PCIe 协议抓包解决方案来捕获芯片内部的数据流量，例如下一代 Broadcom PEX89000 系列交换芯片上提供的我们的新型 PCIe 嵌入式分析仪技术，” Broadcom 数据中心解决方案事业部副总裁兼总经理 Jas Tremblay 说， “与 SerialTek 合作推出这种新的 PCIe 协议诊断功能，使我们能够进一步提高收集捕获数据包的能力，并大大增强当前市场上可用的 PCIe Gen 5 协议分析工具。”

2.8.2 Broadcom/SerialTek 内嵌协议分析仪 PEA 功能简介

目前，Broadcom 在全球范围内要求使用其 PCIe Gen5 switch 的系统集成商，例如服务器、存储系统以及网络设备等厂商，他们的 PCIe Gen5 交换芯片嵌入了 Viking 功能，可

用作内嵌 PCIe 协议分析仪使用，我们一般称之为 PEA (PCIe Embedded Analyzer)，这需要配合 SerialTek GUI 一起提供 PCIe 总线抓包分析解决方案。

Broadcom 要求使用其 Gen5 PCIe switch 的厂商下载 SerialTek PCIe analyzer 软件管理内置于其 Gen5 switch 内部的 iTAP 协议分析仪进行协议抓包、解码、分析，方便解决系统初始化阶段出现的各种 PCIe 总线问题。

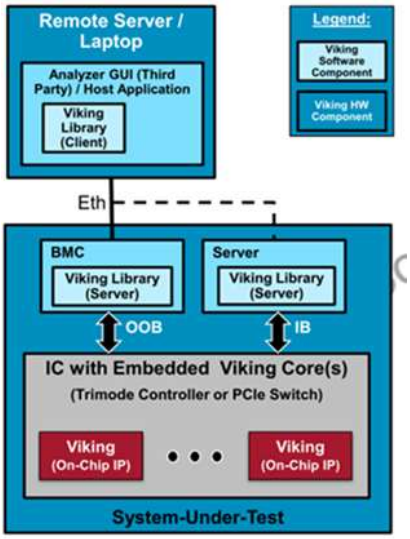


图 2-104

目前使用该基于 iTAP 的 PCIe Embedded Analyzer 主要是通过购买 SerialTek 公司 2022 年上半年推出的 “熊猫” Panda iTAP PCIe Gen5 协议分析仪，该分析仪非常便携，尺寸只有 **154 x 100 x 38 mm**，该分析仪内部具有计算和存储能力，以及 SerialTek 定制的嵌入式 Linux 内核、iTAP 软件和 Broadcom API，所有这些都被加密并用来保护 Broadcom IP。SerialTek 还将提供适用于 3.3v 和 1.8v Broadcom SDB（system debug） 连接/实现的 iTAP 探头。通过 USB Type-C 将 iTAP 探头物理连接到 iTAP 盒并连接到主机接口板上的 SDB 端口或自定义硬件上的调试端口后，您只需将 iTAP 设备插入网络，然后打开浏览器并运行 iTAP 用户界面即可。这样做允许多个用户连接到设备并共享 trace 文件、在同一 trace 文件中协同工作等。SerialTek 希望用户体验与成熟的 Kodiak Gen5 硬件协议分析仪相同。

下面是一个连接示意图。

图 2-105

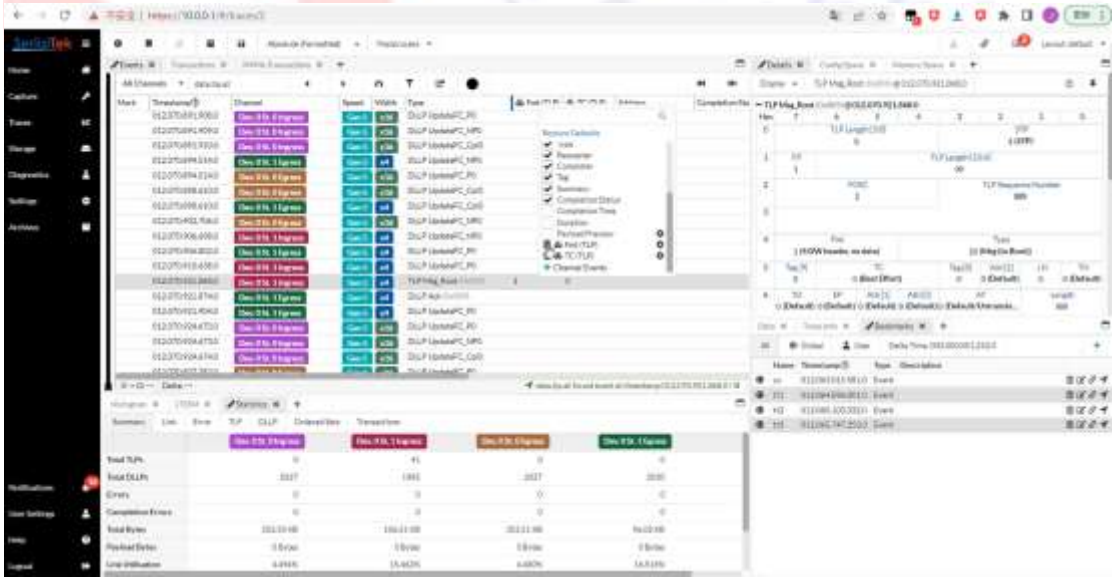


图 2-106

Convenient Access to Embedded PCIe Applications

The Panda appliance conveniently brings together SerialTek's iTAP™ technology and BusXpert™ software with Broadcom PEA technology to smartly solve the problem of analyzing embedded PCIe links in ICs such as the Broadcom PEX89000 PCI Express Gen 5.0 ExpressFabric Platform. The iTAP software is a specialized "tap" framework that is fully integrated with BusXpert, the feature-rich analysis software developed for the revolutionary Kodiak™ PCIe 5.0 Protocol Analysis System.

With the increasing ubiquity of sophisticated, highly integrated systems employing switches and chip-to-chip interconnects, physical limitations to accessing PCIe traffic can prove quite challenging. With the easy access provided by the Panda appliance and SerialTek's PCIe software packages, developers are able to quickly set up, capture, decode, and view PCIe trace data captured on-chip by Broadcom's PCIe Embedded Analyzer (PEA) technology. When used in conjunction with the Kodiak Protocol Analysis System, developers can seamlessly transition between Kodiak's extremely deep traces and iTAP's embedded traces from all instances of PEA stations / ports.

Powerful Features

- Advanced BusXpert analysis software with modern collaborative web UI
- Use in conjunction with full-sized Kodiak PCIe 5.0 protocol analyzers with the same look & feel
- Captures TLPs, DLLPs, and Ordered Sets
- 288K capture buffers per station / port
- Synchronized multi-station capture
- Filtering capabilities
  - Discard TLPs, DLLPs and Ordered Sets
  - Maximize capture buffers efficiency
  - Repeated bytes compression for further buffers optimization
- Triggering capabilities
  - Two-level triggering on input trace data
  - Trigger in: pin-based or routed from internal signals
  - Trigger Out: routed to other trace buffer instances and to a chip output pin

Technical Specifications

Panda Appliance	
Part Number	PP1A-8PEA-ENT
Dimensions	154 x 100 x 38 mm (6 x 4 x 1.5 in)
iTAP PEA Probe	- Attaches via 1m cable - Supports 1.8V and 3.3V operation - Individual receptacles for Rx Tx and GND to suit any target board 0.1in pin header
RAM size	8GB
Storage	480GB
Warranty	2 years - Hardware, including software maintenance

下图连接 Panda 协议分析仪的一些治具，以及 SerialTek BusXpert GUI 的一些配置界面。

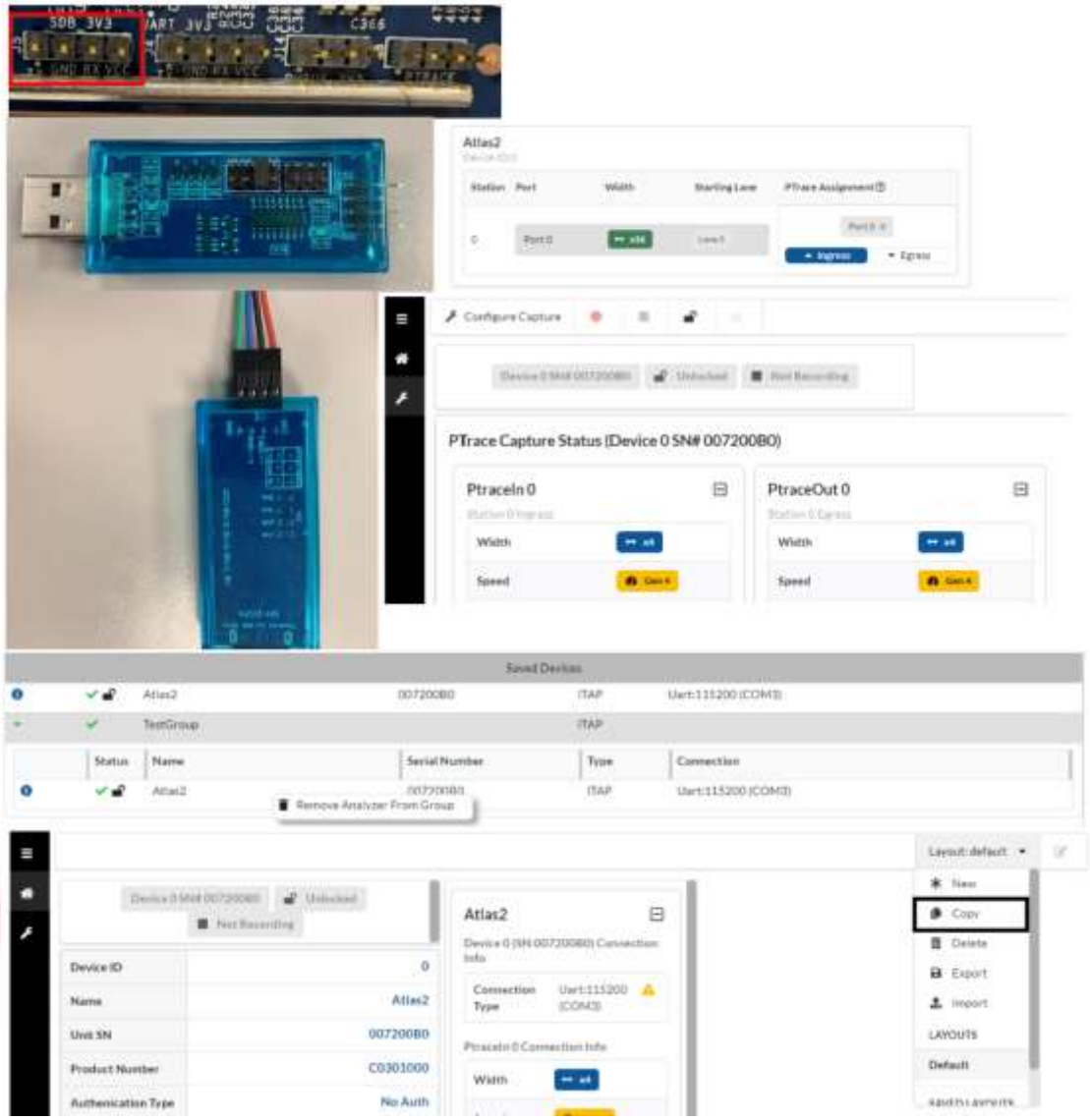


图 2-107

将 USB 串口 dongle 的三根线连接 switch 旁边的 SDB 口的 TX，RX，GND 三根线，注意，dongle 这边的 TX->SDB 口的 RX，dongle 这边的 RX->SDB 口的 TX,两边 GND 互连。

图 2-108

图 2-109

图 2-110

下面是针对该协议分析仪产品配置功能的一些截图介绍。

已分组但尚未配置的设备：

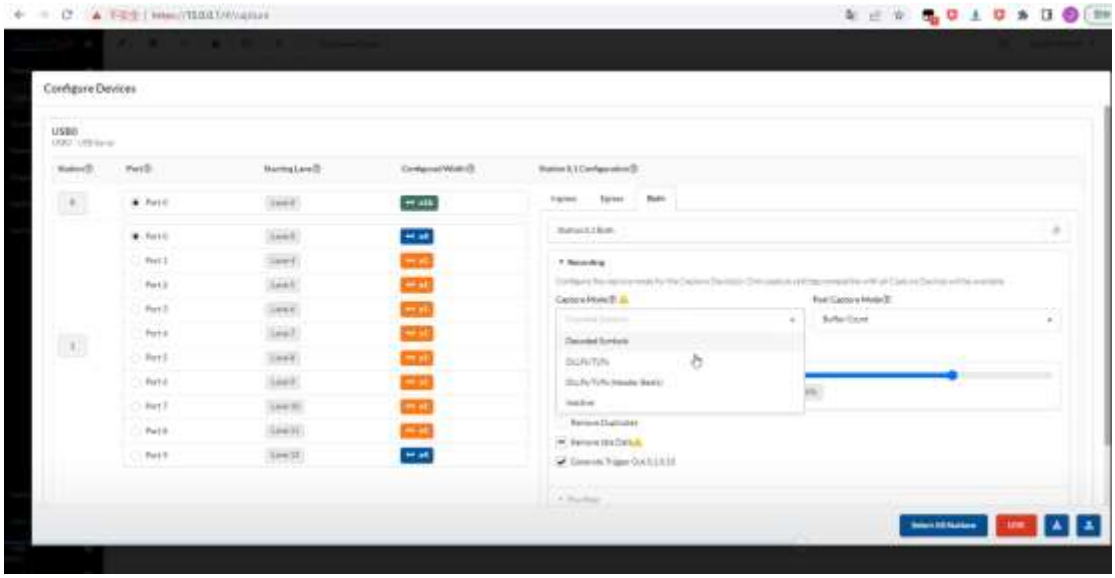


图 2-111

下面是设置好后的准备抓包界面，点击左上角的红色按钮即可开始抓包。

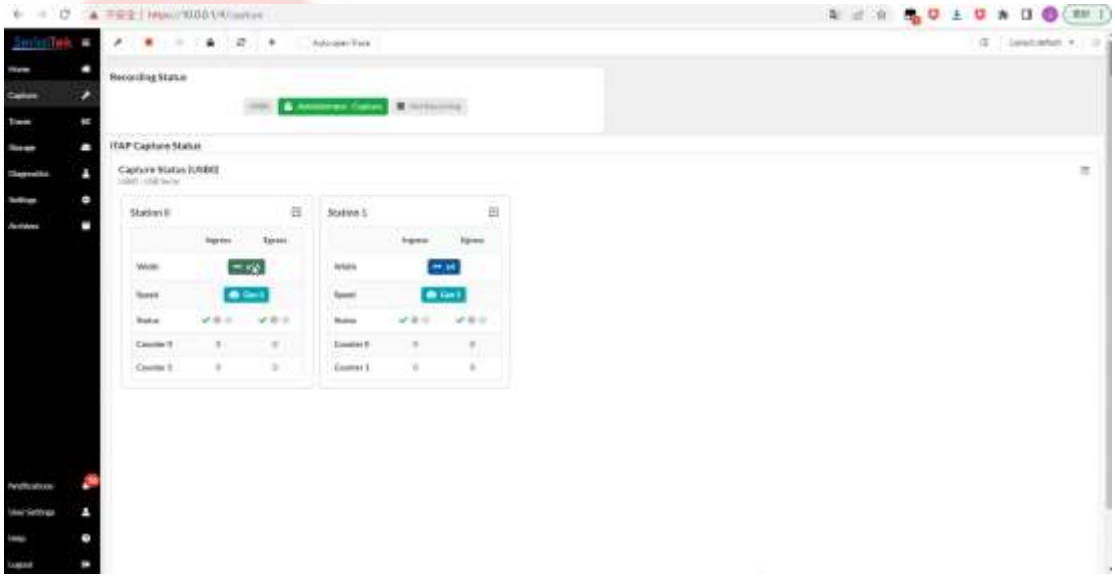


图 2-112

将某些端口分配给 analyzer，开始抓取数据。抓取数据后停止下来的 buffer loading 界面：

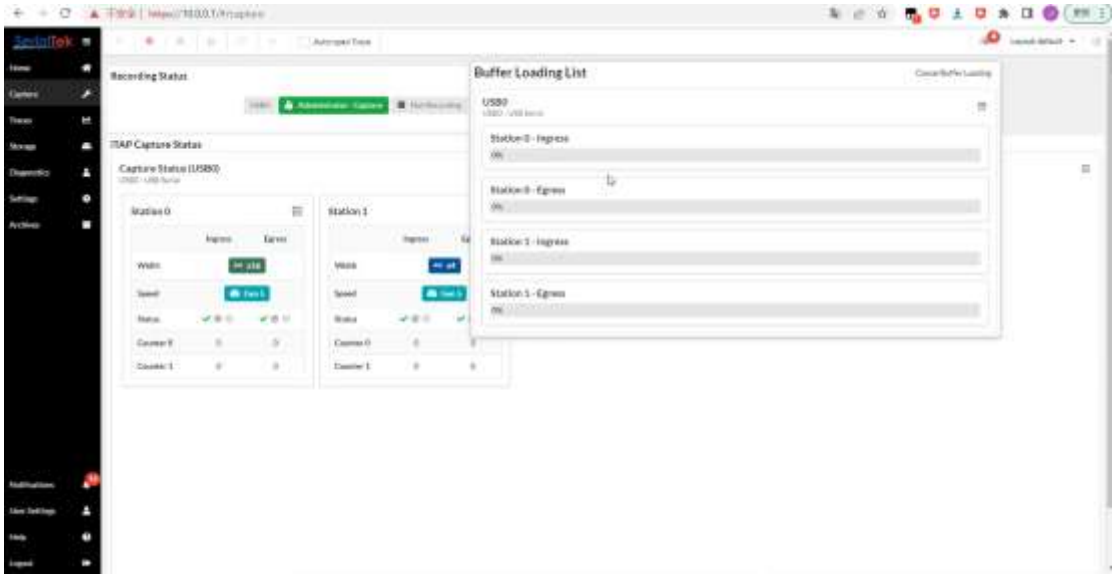


图 2-113

抓到数据后的解码界面：

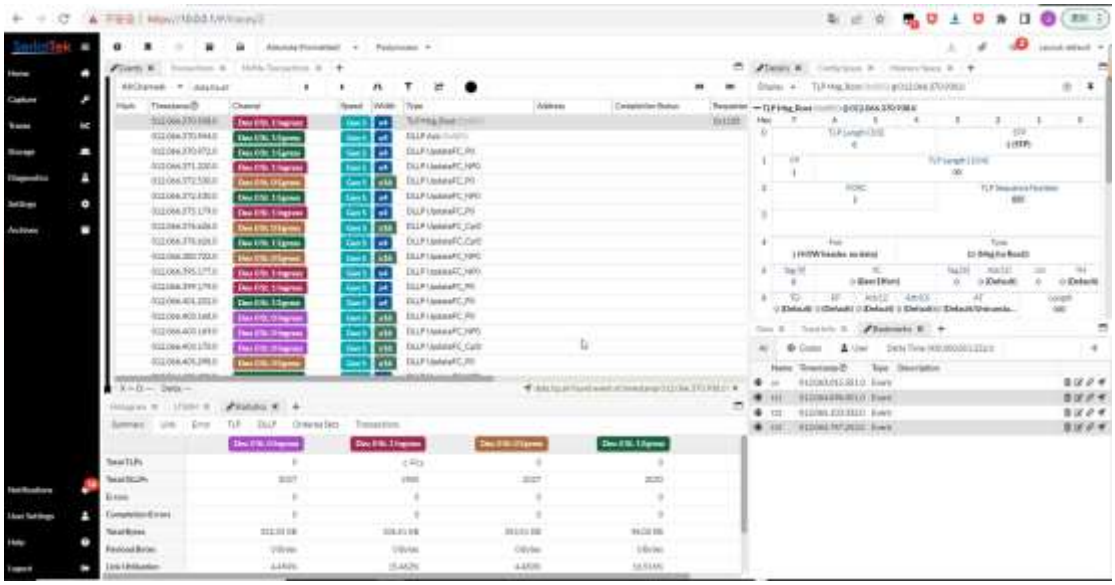


图 2-114

下面是显示过滤界面，和传统独立式 pcie gen5 analyzer 一致。

图 2-115

配置 Pre-filter 过滤功能：

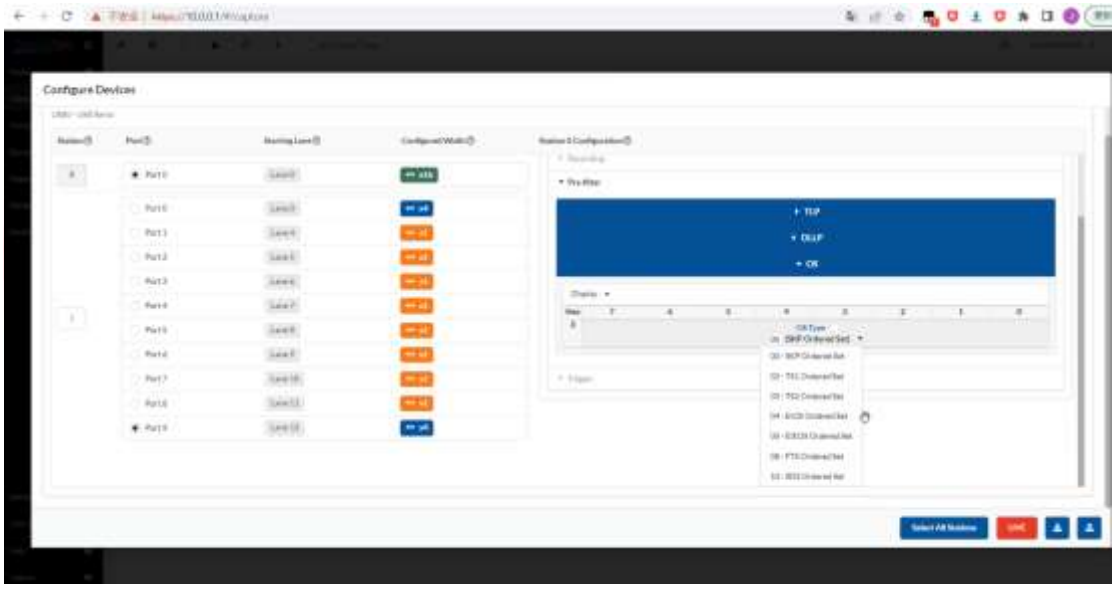


图 2-116

配置 trigger 触发条件：

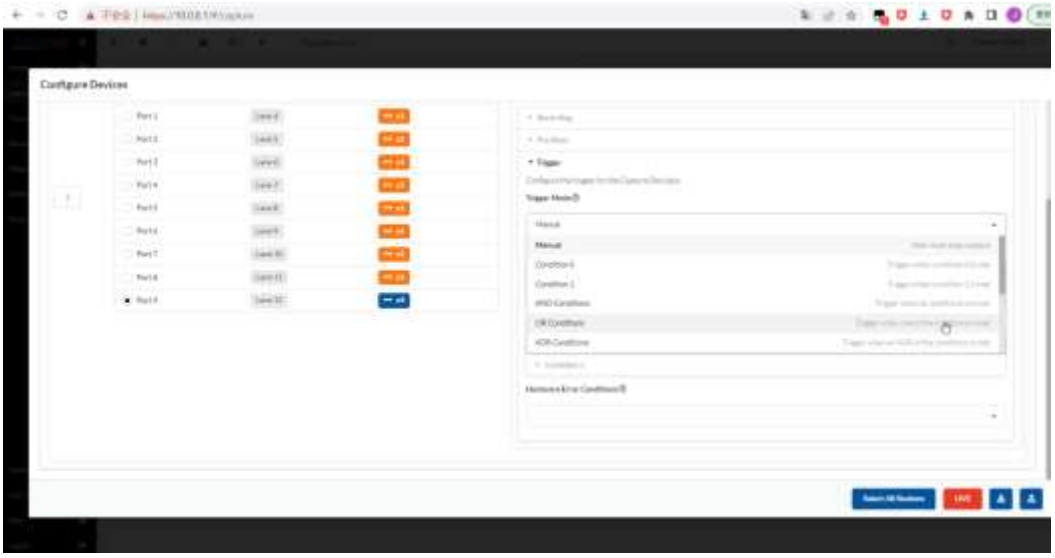


图 2-117

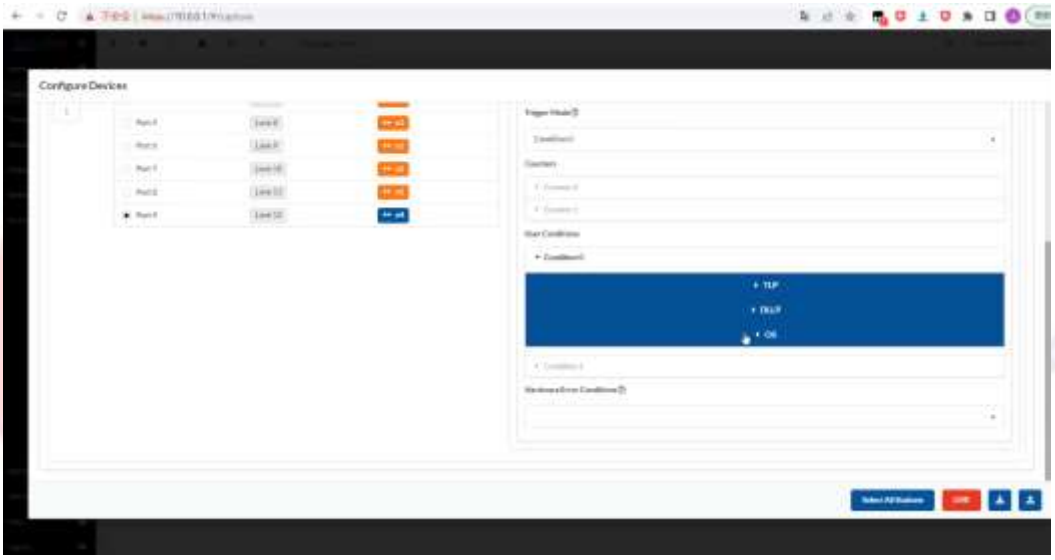


图 2-118

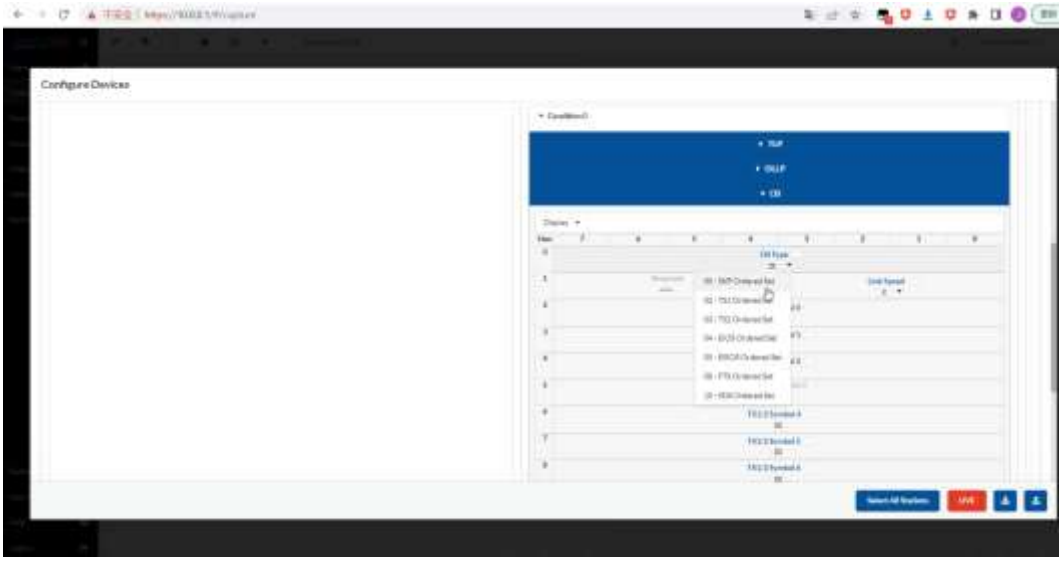


图 2-119

2.8.3 Broadcom 的 PCIe Gen 5.0 产品组合为下一代服务器奠定了基础

2022 年 3 月 4 日米歇尔·弗罗斯

Broadcom Inc.是 PCIe、SerDes 和交换芯片技术的全球供应商，宣布其广泛的 PCIe Gen 5.0 产品组合正在为构建高性能下一代服务器的生态系统奠定基础。

图 2-120

Broadcom PCIe Gen 5.0 SerDes、交换芯片和定制硅产品（现在可供 OEM、ODM 和云服务提供商使用）正在积极地进行测试和演示，以实现大规模的互操作性。

“在过去的 20 年里，PCIe 一直是互连 CPU、GPGPU、FPGA 和存储的最关键和最广泛采用的协议，”博通数据中心解决方案事业部副总裁兼总经理 Jas Tremblay 说。“博通市场领先的 PCIe 传统和新的 PCIe Gen 5.0 产品组合使我们能够无缝统一高速计算、低延迟和数据处理至关重要的数据中心。”

优势

PCIe Gen 5.0 的速度是 PCIe 4.0 的两倍，这对于支持采用 AI/ML 的数据中心和存储服务器以及需要高性能、低延迟、可扩展和经济高效的 PCIe 结构的超大规模计算系统至关重要。

Broadcom 的 PCIe Gen 5.0 产品连接产品组合包括业界可用的一些最低功耗、最高通道数和最低延迟。从 ASIC 到 SerDes 再到交换芯片和嵌入式 PCIe 协议分析仪，Broadcom 正在通过各种产品在整个行业范围内使能 PCIe Gen 5.0：

- **定制 ASIC 开发支持，采用 7ns、5nm 和 3nm 工艺技术的 32GT/s PCIe 5.0 SerDes，以及对 PCIe 5.0 标准产品的支持。**
- **其新的 PEX89000 PCIe Gen 5.0 系列交换芯片采用 32GT/s PCIe SerDes，与前几代产品相比，每 GB 数据传输功率大幅提升。**
- **PEX89000 交换芯片的功耗不到 5.0 代交换芯片替代品的一半，并提供高于标准的最佳 5.0 代 SerDes 电气性能余量。PEX89000 系列还包括新的增强安全功能，包括证**

明和 SPDM1.1，以及遥测和诊断功能。

- **全球第一款在 PCIe Switch 交换芯片中集成嵌入式 PCIe Gen5 协议分析仪，通过 SerialTek 的专有 iTAP 协议抓包 IP 以及 BusXpert 协议分析仪软件实现 在 PEX89000 交换芯片系列实现针对 PCIe 协议的实时抓包分析功能。**
- **用于 Gen 5.0 互操作活动的灵活 PEX89000 评估板，包括多个扩展槽，用于评估具有主机、GPGPU、NIC、retimer 和 NVMe 设备连接的复杂系统。**
- **领先的服务质量功能来管理拥塞。**

“三星一直与博通密切合作，以确保 PCIe Gen 5.0 互操作性和测试是无缝的，”三星解决方案产品与开发企业副总裁 Soonjae Won 分享道。“我们全新的 PCIe Gen 5.0 NVMe 驱动器与 Broadcom PEX89000 交换芯片一起，通过下一代 PCIe 端到端解决方案在数据中心实现大规模部署，在生态系统中提供业界首创。”

端到端生态系统

Broadcom 致力于并投资于 PCIe Gen 5.0 的行业互操作性。从行业活动中的端到端演示到数百个评估板的发货，随着技术的迅速成熟和对其评估板的依赖性不断增长，Broadcom 正在促进生态系统向 PCIe Gen 5.0 的过渡。

NVIDIA DGX 系统副总裁兼总经理 Charlie Boyle 表示：“随着尖端 AI 继续在各行各业中传播，对我们的多 GPU 系统的需求从未如此强烈。“NVIDIA 正在与 Broadcom 密切合作，将他们的 PCIe 第 5 代交换芯片整合到我们的 DGX 系统系列中，以帮助我们的客户提供最佳工具来加速他们的工作。”

加速计算平台、处理器、GPGPU、NVMe、重定时器和测试设备产品的所有供应商都收到了用于互操作性和演示的 Broadcom PCIe Gen 5.0 评估板：

- **加速计算提供商：NVIDIA**
- **GPGPU 供应商：英特尔和 AMD**
- **NVMe 供应商：三星、铠侠、英特尔等 NVMe 供应商**
- **重定时器供应商：Astera Labs 和其他重定时器供应商**
- **测试设备供应商：SerialTek 等测试设备供应商**
- **50 多家 OEM 和 ODM**

AMD 数据中心生态系统和解决方案公司副总裁 Raghu Nambiar 表示：“AMD 和 Broadcom 延续了我们长期的互操作性测试关系。“作为下一代技术的领先创新者，我们致力于提供深入的测试，以帮助确保在现代数据中心运行关键业务应用程序的 PCIe Gen 5.0 系统能够成功运行并符合市场实施的要求。”

“英特尔和博通继续合作，将先进的 PCIe 技术推向市场，而 PCIe 5.0 技术延续了我们为生态系统提供构建高性能服务器解决方案所需的强大构建块的悠久历史，”Debendra Das Sharma 博士（英特尔 IO 技术与标准高级研究员、首席架构师）补充道。“作为 PCIe

5.0 标准的主要贡献者，英特尔现在正在对业界首个基于 PCIe 5.0 的服务器平台进行采样。我们很高兴博通还通过 PEX89000 交换芯片、SerDes 和 PHY 技术为生态系统做出贡献。”

2.9 SerialTek PCIe Gen5 x4 协议分析仪第三方评测

Kodiak PCIe Gen5 协议分析仪评测

作者：莱尔·史密斯 2023 年 5 月 19 日

https://www.storagereview.com/review/kodiak-pcie-gen5-analysis-system-review

Kodiak PCIe Gen5 协议分析仪是一款先进的协议分析仪，旨在解决复杂且快速发展的存储和数据中心 I/O 技术带来的挑战。该 Kodiak 系统的亮点在于其高度响应的基于高端服务器 CPU 的数据处理架构，该架构可以增强界面响应能力、涉及大数据量的快速搜索功能以及灵活、强大的硬件过滤功能。该平台还具有基于网络浏览器的 BusXpert 应用程序，以进一步促进分析过程。

一般来说，PCIe/NVMe 分析仪是测试、测量调试工具，旨在分析和验证利用 PCIe 技术和工艺的设备的性能、合规性和互操作性。它们主要是为使用 PCIe NVMe 技术的设备的硬件/软件开发人员和工程师而设计的；在这种情况下，最高支持到 Gen5 接口（也兼容支持 1/2/3/4 代）。

它通过提供更快、更准确的测试和分析、改进的验证和验证流程以及更有效地识别和解决问题的能力来帮助用户。此外，它还可以通过在开发周期的早期识别和解决问题来减少开发时间和成本。

2.9.1 Kodiak PCIe Gen5 协议分析仪功能

Kodiak 系统的关键功能之一是其专有的实时协议处理器 (RTPP)，无论分析仪正在主动记录还是空闲，它都会动态自动查询和保存 PCI 配置空间、主机控制器寄存器和 NVMe 队列。这项创新消除了耗时的重新启动的需要，并能够使用当前值进行精确解码、触发和过滤。

该系统还具有广泛的外形规格支持，例如 AIC (x4)、EDSFF、M.2、U.2 和 U.3，具有单端口 (1x4) 和双端口 (2x2) 分析组合成一个 Gen5 Pod。这意味着在使用所有 form factor 的企业环境中可能会节省成本，因为它无需购买额外的设备。此外，SI-Fi Interposer 支持所有相关边带，包括来自主机或来自外部/第三方注入或生成工具的 SMBus（例如 NVMe-MI）。

Kodiak 灵活的 Trace 存储和检索选项包括两个 10GbE SFP+ 端口、一个 GbE 端口、高达 2TB 的内部 SSD 存储、两个 USB 3.1 端口和两个 PCIe OCuLink 端口。这些功能允许用户将 Trace 文件下载到主机或网络，为其他用户提供只读访问并启用各种直接连接存储选项。

SerialTek 专有的 SI-Fi Interposer 技术是系统设计中的关键要素，可确保探测过程的透明度。该技术的专用线性放大器设计在差分输入处接收 PCIe 模拟信号，并将其分配到两个独立的相位匹配差分输出，标称理想增益为 0dB。这种方法简化了分析仪和被测产品的设置，并避免了其他探测方法中固有的限制，即链路训练序列不通过 Interposer。

SI-Fi 技术和 Kodiak 的自适应 EQ 功能消除了校准的需要，而其它 PCIe Gen5 分析仪和 Interposer 则需要校准，而校准通常会导致可靠性问题。借助这些功能，用户可以节省数小时的设置时间并动态 Trace 文件链路特性变化，例如热插拔或 NVM 子系统重置 (NSSR)，最终简化测试过程。

2.9.2 Kodiak PCIe Gen5 协议分析仪设计和架构

Kodiak Gen5 协议分析仪采用全金属外壳，采用漂亮的浅蓝色和灰色设计。前面板上有双 1GbE 和 USB 3.0 端口、两个 PCIe OCuLink 端口（用于连接外部 PCIe 设备的高速接口连接器）和双 10GbE SFP+ 端口，可用于将系统连接到网络进行数据传输和远程控制目的。

前面板中间是触摸屏 LCD，它允许用户对分析仪进行配置、控制和接收来自分析仪的状态更新。

右侧是系统总线的 sideband 边带信号（SB0、SB1），它们是分析仪上的主要接口端口，用于从被测系统捕获数据； Upstream PCIe 端口，用于连接被分析设备的上行端口；Downstream PCIe 端口，用于连接被分析设备的下游端口。

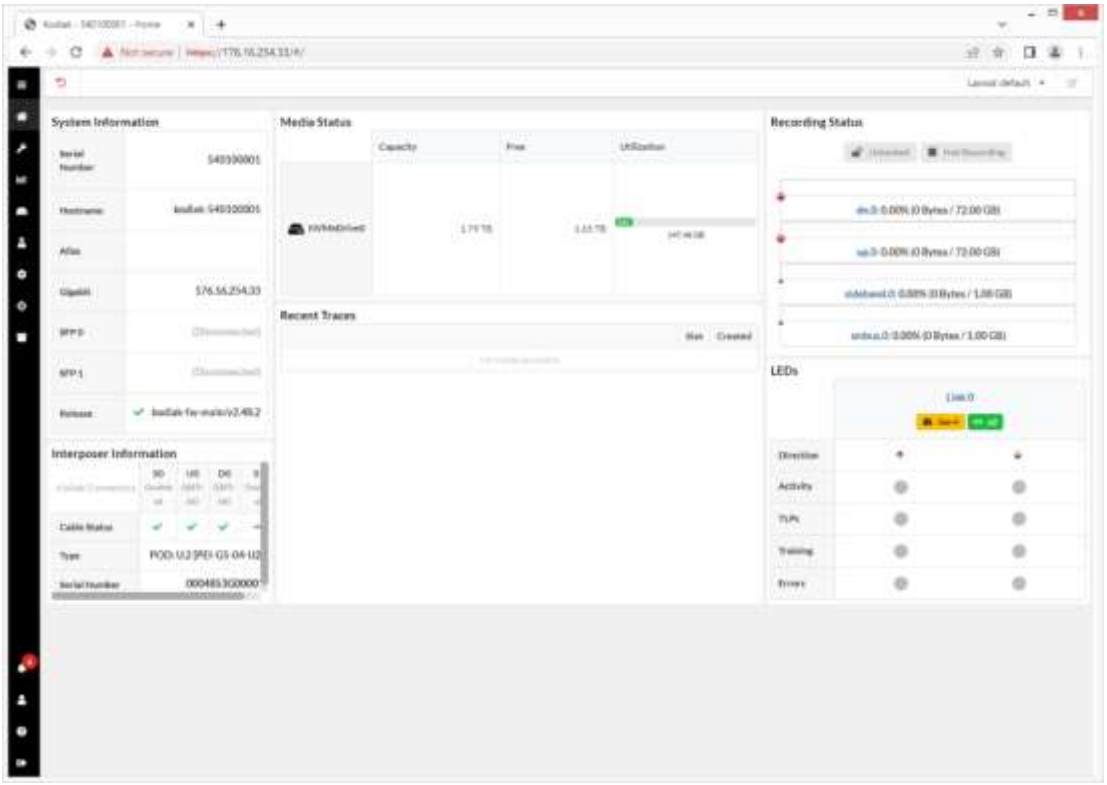
六个系统风扇分布在后面板上。风扇之间是时钟输入/输出和触发/定时输入/输出连接器，它们同步并触发 PCI 流量的分析。时钟输入/输出连接器提供精确的时钟信号，确保 PCI 流量捕获和分析的精确同步，而触发/定时输入/输出连接器提供启动或终止 PCI 流量捕获和分析的信号。

下面是维护端口，最右侧是电源连接器和开关。

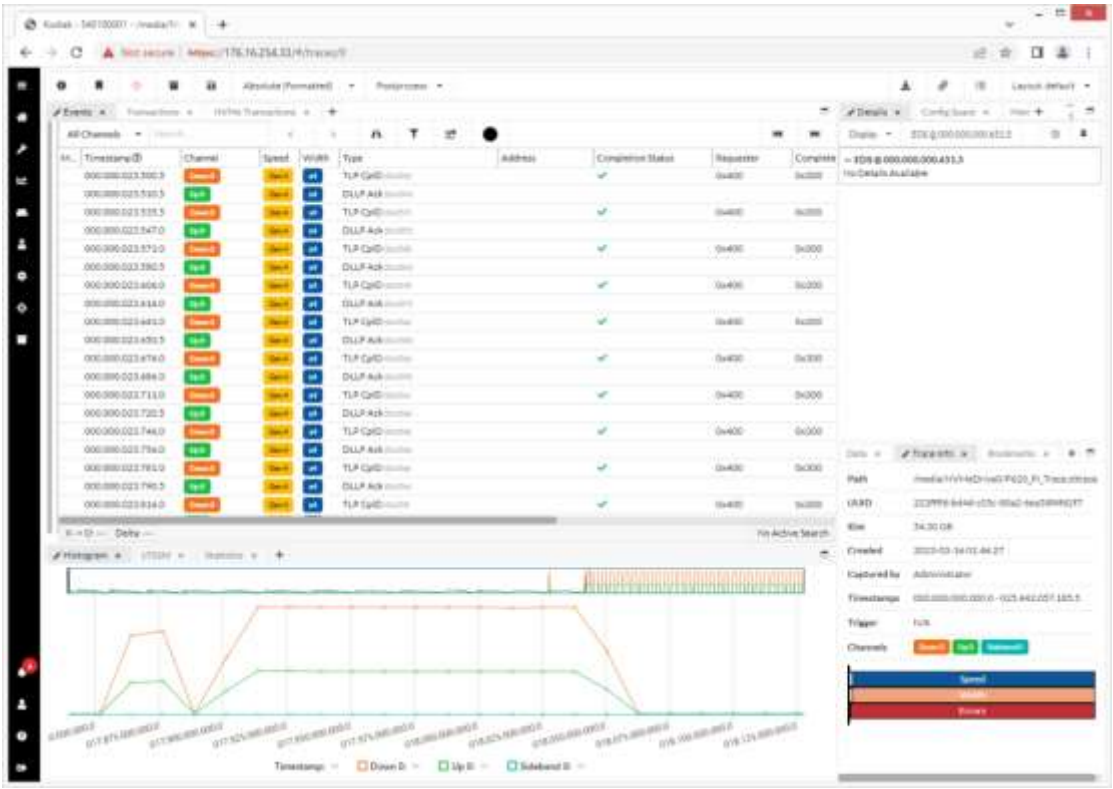
2.9.3 Kodiak PCIe Gen5 协议分析仪管理

“ BusXpert ”与 Kodiak PCIe x16 Gen5 分析仪硬件集成，提供用于访问和解释捕获数据的用户界面。 BusXpert 软件基于嵌入式软件框架和 REST API，使其能够与 Kodiak 硬件无缝集成。

主页仪表板包含系统信息（序列号、主机名、别名、千兆位、版本）、媒体状态（SSD 类型、容量信息和利用率）、有关系统 LED 的信息和录制状态等部分。在我们的例子中，我们在 Solidigm QLC SSD 上运行了 Trace 文件。

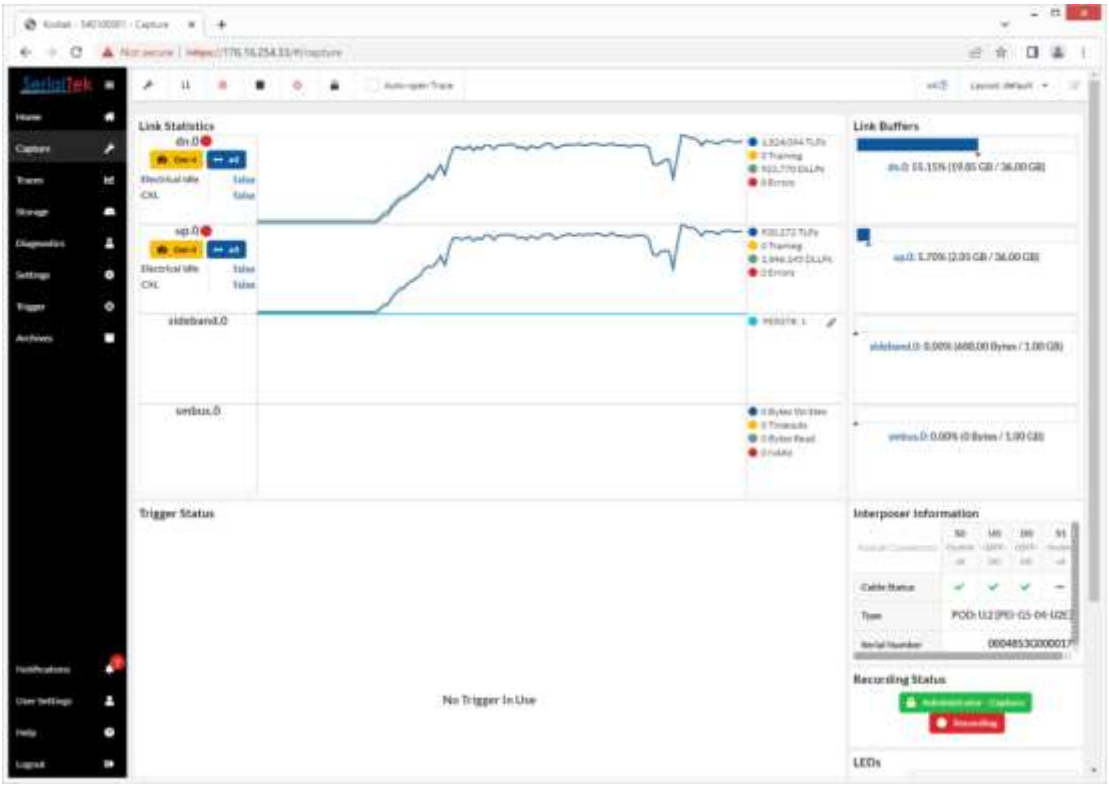


“事件”选项卡是一个显示表格的部分，其中的行包含与 PCIe 事件相关的各种信息，包括发生时间 (Timestamp)、事件发生的通道 (Channel)、事件的速度和宽度。 PCIe 链路（速度和宽度）、事件类型 (Type)、与事件关联的地址 (Address)、完成状态 (Completion)、请求者信息 (Requestor) 和完成详细信息 (Complete)。



在“事件”选项卡下方，用户可以查看其他信息，例如“直方图”选项卡，它提供事件随时间分布的图形表示。这有助于用户在 PCIe Gen5 分析中可视化不同类型事件的频率或发生情况。

BusXpert 中的“链路统计”部分为用户提供有关 PCIe 链路或通道的性能和状态的信息。 “dn.0”和“up.0”分别指 PCIe 总线的下游 (dn) 和上游 (up) 链路，其中“0”表示第一个链路或通道。



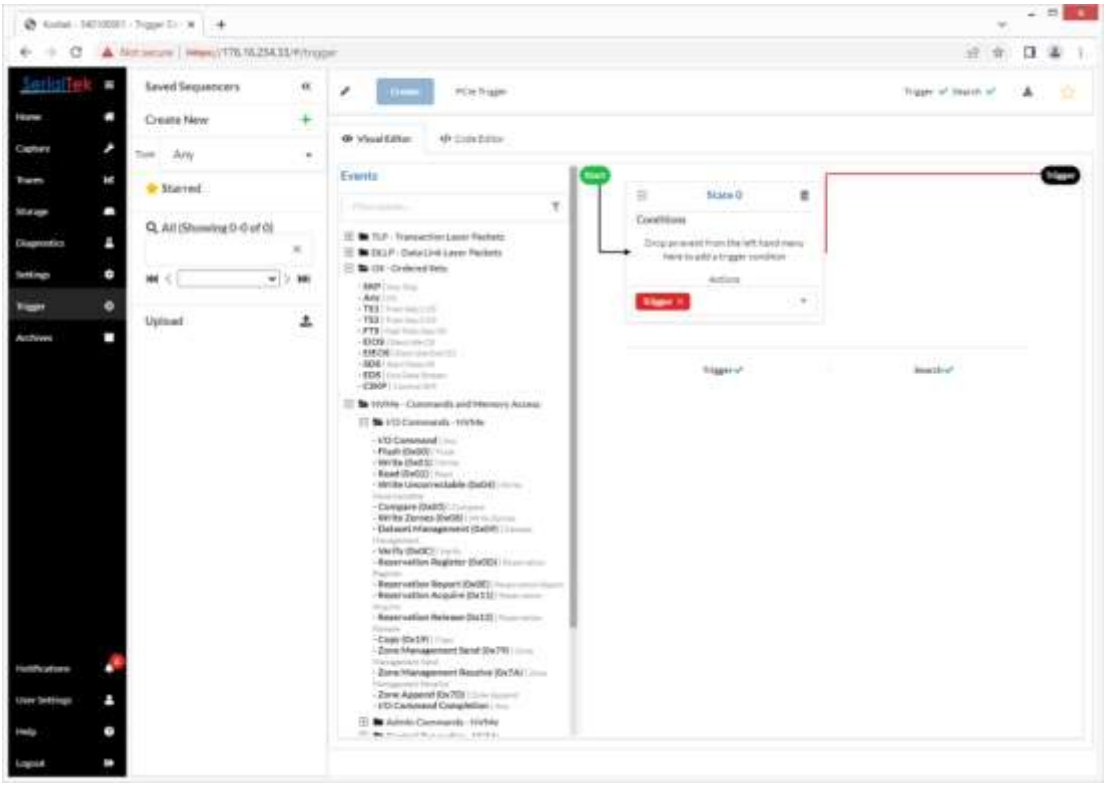
它还包括链接速度（例如，Gen3、Gen4 或 Gen5）和链接宽度（例如，x2、x4、x8、x16）等详细信息。

折线图的右侧是颜色代码：

- TLP（事务层数据包）由蓝色表示，表示在 PCIe 链路或通道上捕获或观察到的 TLP 数量。
- 训练是指 PCIe 链路训练或链路训练和状态机 (LTSSM) 事件，用黄色表示，表示检测到或捕获的训练事件的数量。
- DLLP（数据链路层数据包）由绿色表示，表示在 PCIe 链路或通道上捕获或观察到的 DLLP 数量。
- 错误可能包括各种类型的 PCIe 链路错误或其他异常，由红色表示，表示检测到的错误数量。

每个颜色代码旁边的数字可能表示监控或分析期间每种类型事件的计数或频率。这为用户提供了对这些事件的深入了解。

在“触发器”部分中，您可以加载、创建或上传“已保存的序列”。这些是预定义或定制的测试用例或场景集，可以保存并重用以供以后分析或测试之用。



右侧是“可视化编辑器”和“代码编辑器”。这些允许用户以可视化或编程方式定义或修改与 PCIe 事务、事件或场景相关的特定参数、设置或配置。

2.9.4 Kodiak PCIe Gen5 协议分析仪兼容性

Kodiak PCIe Gen5 协议分析仪的主要优势在于它能够连接到几乎任何外形尺寸的 NVMe 驱动器。这种多功能性对于当今快速发展的存储环境至关重要，其中使用了许多不同的外形尺寸和接口。

该系统与各种 PCIe Gen5 接口兼容，包括 x16 和 x8 插槽 Interposer，以及适用于 EDSFF 和 M.2 等特定外形尺寸的 Interposer。此外，该系统支持的软件包包括用于 x4 AIC、EDSFF、M.2、U2 和 U3 的智能接口适配器，从而可以轻松测试和分析各种设备。

其他支持的插入器包括 Slim-SAS、 OCuLink 和 MCIO 电缆插入器，以及适用于 Gen4 HD MiniSAS 电缆的插入器。

支持的 PCIe Gen5 SI-FI Interposer

PCIe Gen5 x16 插槽 Interposer

PCIe Gen5 x8 插槽 Interposer

PCIe Gen5 x16 OCP Interposer（不包括 Quarch PAM）

PCIe Gen5 x8 EDSFF E3 Interposer（不包括 Quarch PAM）

PCIe Gen5 x4 高级套件。包括 SI-Fi Interposer POD 和用于 x4 AIC、EDSFF（E1.S、E1.L、E3.S、E3.L）、M.2、U2、U3 的智能接口适配器； 6MB

PCIe Gen5 x4 插槽 Interposer

PCIe Gen5 EDSFF（E1.S、E1.L、E3.S、E3.L）

PCIe Gen5 M.2 Interposer

PCIe Gen5 U2 Interposer

PCIe Gen5 U3 Interposer

PCIe Gen4 超薄 SAS 电缆 Interposer

PCIe Gen4 OCuLink 电缆 Interposer

PCIe Gen5 MCIO 电缆 Interposer

PCIe Gen4 HD MiniSAS 电缆 Interposer

结论

Kodiak PCIe Gen5 协议分析仪对于参与测试和分析 PCIe Gen5 和 NVMe 设备的任何人来说都是一个关键工具。它的先进功能和多功能接口选项使其成为硬件和软件开发人员和工程师的必备工具，因为它使他们能够快速有效地识别和解决问题。该系统强大的实时协议处理器 (RTPP)、灵活的 Trace 存储和检索选项以及 SI-Fi Interposer 技术可确保准确可靠的测试结果，而无需耗时的校准。

此外，其广泛的外形规格支持（例如 AIC、EDSFF、M.2、U.2 和 U.3）使其成为使用多种外形规格的企业环境的绝佳选择。

也就是说，对于某些用例来说，它可能有点矫枉过正，因为它的高级特性和功能对于较小规模的测试和分析可能不是必需的，而且对于那些需求更有限的人来说，它的高价位可能不合理。然而，对于在测试和分析过程中需要最高水平的准确性和可靠性的实验室来说，Kodiak PCIe Gen5 协议分析仪是一个绝佳的选择。

莱尔·史密斯
StorageReview 的特约撰稿人，涵盖广泛的最终用户和企业 IT 主题。

2.10 SerialTek PCIe Gen5/CXL2.0 协议分析仪单页

SerialTek
an ellisys company

Kodiak™

Next-Generation Gen5 PCIe®/ NVMe® and CXL Analyzer

Innovative ▪ Cutting-Edge ▪ Integrated

www.serialtek.com/kodiakgen5

PCIe/NVMe and CXL Analysis Platform with Embedded Hardware, Real-Time Protocol Processor™, Calibration-Free SI-Fi™ Interposer Probing and Automatic Equalization, Internal SSD Storage, Touchscreen LCD, and Standard PCIe Cabling.

State-of-the-Art Architecture

The Kodiak PCIe Gen5 Analysis System represents the state-of-the-art in protocol analyzer design. The Kodiak platform includes an array of high-performance innovations, made possible by an advanced design that breaks free from cumbersome legacy data upload practices in favor of ultra-responsive embedded data processing.

Interface responsiveness is markedly advanced, searches involving massive amounts of data are fast, and hardware filtering is flexible and powerful.

The Kodiak platform, with its new web browser based BusXpert (TM) application, is built to tackle the challenges presented by the complexities of rapidly advancing storage and datacenter I/O technologies.

Real-Time Protocol Processor

Kodiak employs an innovative system register processing concept called Real-Time Protocol Processor (RTPP™). This proprietary feature dynamically and automatically queries and saves PCI configuration space, host controller registers, and NVMe queues, whether the analyzer is actively recording or idle. This alleviates the need for time-consuming and highly impractical reboots, and provides the ability to precisely decode, trigger, and filter using current values.

Multiple Form Factor Support

SI-Fi™ interposer form factors include AIC (x4), EDSFF, M.2, U.2, and U.3,. Additionally, U.2, U.3, single-port (1x4), and dual-port (2x2) analysis is combined into one interposer unit, providing significant cost savings in enterprise environments where all form factors are re- quired. SI-Fi™ interposers also support all relevant side- bands, including SMBus (e.g., NVMe-MI) from the host or from external / third-party injection or generation tools.

Flexible Trace Storage and Retrieval

Kodiak includes two 10GbE SFP+ ports and a GbE port to offload traces to a host computer or network and internal SSD trace storage of up to 2TB (with read-only access for other users). Direct attach storage choices include two USB 3.1 ports and two PCIe OCuLink ports.

Transparency in Probe Design is Key

Driven by the need for ever-faster data transfers, PCI Express signaling has become exceptionally complex in design and difficult to monitor unobtrusively. Signal conditioning methods used for earlier PCIe generations/speeds now seem primitive compared to the complex approaches used for PCIe Gen5. Further challenges are presented by NVMe, which adds critical requirements like hot-plug and NVM Subsystem Reset (NSSR), where the PCIe signals are renegotiated. SerialTek's proprietary SI-Fi™ technology meets and overcomes these challenges with the features and capabilities needed to work efficiently.

With SerialTek's SI-Fi™ interposer technology, the transmitter threshold and pre-emphasis from one link partner reaches the receiver of the other link partner, so the link properly trains to optimum conditions, making the interposer as transparent as possible.

At the core of this technology is a highly specialized linear amplifier design where PCIe analog signals are received at a differential input and distributed to two separate phase-matched differential outputs with a nominal, idealized gain of 0dB. This approach results in easier set up of the analyzer and product under test and avoids a variety of limitations inherent to other probing approaches where link training sequences don't pass through the interposer.

SerialTek's SI-Fi™ interposer technology expands and enables coverage in critical test areas, including link training (LTSSM), Power Management, Hot Plug, Reset, and other situations where the physical link/lane characteristics may change.

No Need for Calibration

Competing PCIe Gen5 analyzers and interposers require tuning, or calibration, which leads to reliability issues as modern PCIe link training sequences can occur dynamically, not just at boot-up.

With SI-Fi™ technology and Kodiak's adaptive EQ capabili- ties, users can save hours in setup time. And if the link characteristics change (e.g., Hot Plug or NSSR), Kodiak can follow those changes dynamically, ultimately saving your test.

Powerful SerialTek Features

- No tuning (calibration) required
  - Kodiak's Rx automatically equalizes (EQs) the PCIe signals at all data rates
- Embedded trace processing architecture and fastest performance
- Real-Time Protocol Processor
  - No boot trace needed
  - Automatically captures PCI Config Space, Controller Registers, and NVMe Queues
  - Native NVMe triggers by device (BDF), Queues, and Packet/Event
  - Native NVMe filters by device (BDF), Controller Registers, Queues, and Packet/Event
- Deep Trace Buffers
  - 72GB, 144GB
- Internal Trace Storage (SSD)
  - 2TB
  - Read-only access for non-primary users
- Direct Attach Storage
  - Two OCuLink (PCIe) ports
  - Two USB 3.1 ports
- Network and Direct Connectivity
  - Two 10GbE SFP+ (optical/copper)
  - One 1GbE RJ-45
- Single-port (1x4) and dual-port (2x2) analysis in one platform
- Real-time access to traces in memory (prior to downloading)
  - Users can review and analyze captured traces without downloading the trace
- Touchscreen LCD for analyzer setup and status

Interposers with SI-Fi™ Technology

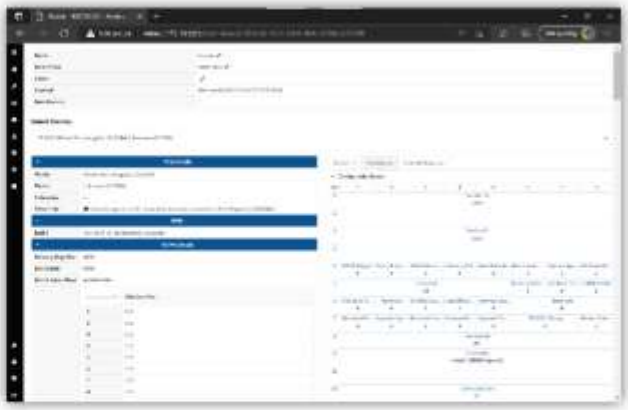
- No tuning (calibration) required
  - Host and Device signals pass through the interposer, allowing for real-world PCIe link training and easier setup
- SI-Fi™ interposer probes expand coverage to enable testing in critical areas, including link training (LTSSM), Power Management, Hot Plug, Reset, and other situations where the physical link/lane characteristics may change
- AIC (x4), M.2 (x4), U.2 (x4), and U.3 (x4)
  - U.2, U.3, single-port (1x4), and dual-port (2x2) in one interposer
- Access to all sidebands, including SMBus

SerialTek Kodiak™

Next-Generation Gen5 PCIe/NVMe and CXL Analyzer

SerialTek
an eBays company

Real-Time Protocol Processor™



Automatically identi ies & updates:

- PCIe configuration space
- Controller data structures (queue attributes, etc.)
- NVMe queue creation and deletion

Uses

- Capture and decode PCIe and NVMe protocols without a boot trace
- Easy analyzer set up
- Correctly decode trace if any of the above attributes change
- Native NVMe triggering: by event (packet), device (BDF), and queue - eliminates false triggers
- Native NVMe Filtering: by device (BDF), controller registers, and queue

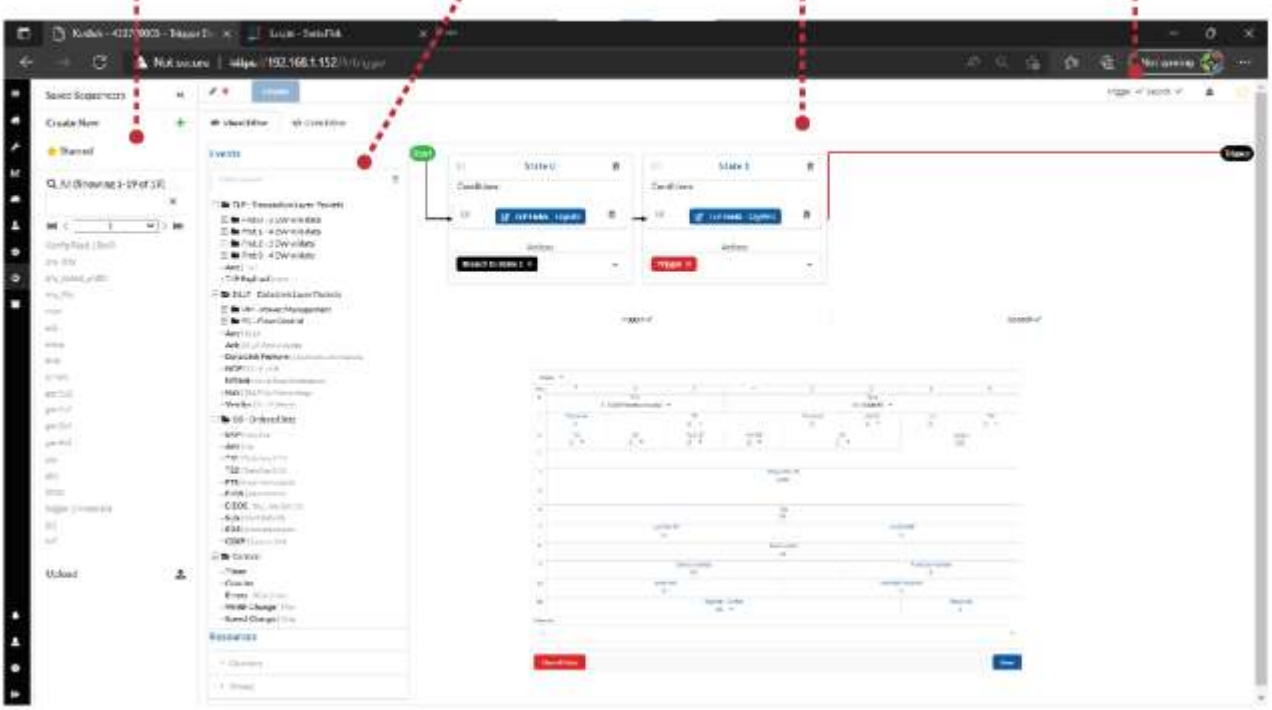
PCIe Hardware Triggers

Savable Trigger and Search conditions

Create triggers easily with searchable drag/drop interface

Simple, advanced multi-state and multi-sequencer triggering

Interchangeable Trigger and Search conditions

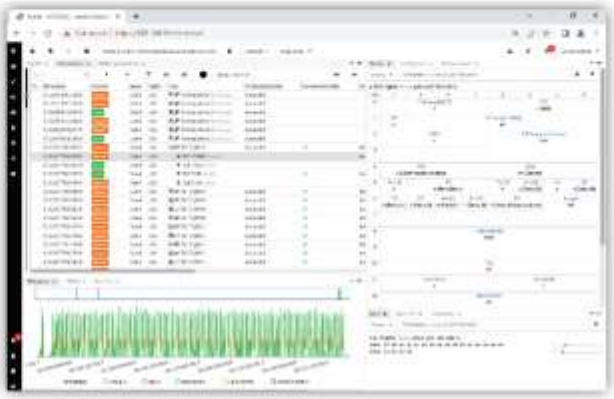


4

BusXpert Software

Web Browser and Standalone Application - Two NEW User Interfaces

Based on an embedded software framework and REST API, The BusXpert software integrates with Kodiak hardware seamlessly. Accessed via a web browser or SerialTek's Electron®-based app, BusXpert includes a suite of powerful triggers, filters, and trace processing capabilities coupled with a new user interface for fast, easy, and reliable decoding. Users can work with trace files collaboratively in real-time and even remotely verify proper configuration of the analyzer and interposers, including visual identification of cables, link status, recording status, and much more.



Customizable Views – Widgets

One major aspect of the new GUI are the widgets and how the user interacts with the trace data through them. The widgets contain controls that are specific to them and there is a global toolbar that applies to all widgets.

There are a collection of widgets that a user can user to analyze the trace data in several different formats, easily accessed via a layout manager used to customize your Home, Capture, and Trace Viewing screens.

Easy Automation - REST API

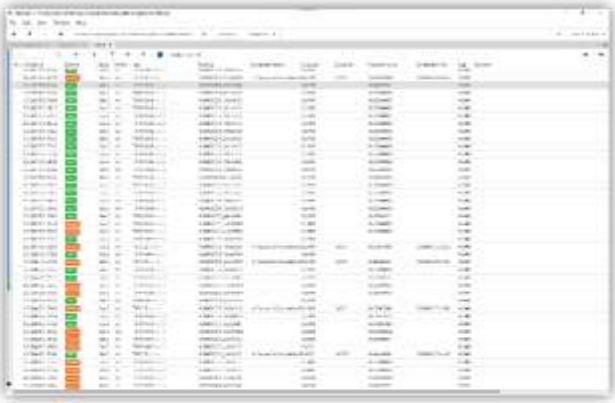
The Kodiak REST API makes automation straightforward and efficient, providing programmatic facilities for monitoring and capturing traffic, statistical analyses, and detailed searching. Kodiak's advanced hardware design also means there is no need to download a multi-gigabyte trace before the user can begin to review the analysis – data is ready immediately.

Trace Widgets

Trace widgets are a collection of views used to analyze the trace data in several different formats, including Events (i.e. spreadsheet), PCIe Transaction, and NVMe Transaction widgets.

Events

Displays a table of all events returned from the current location of the curser.

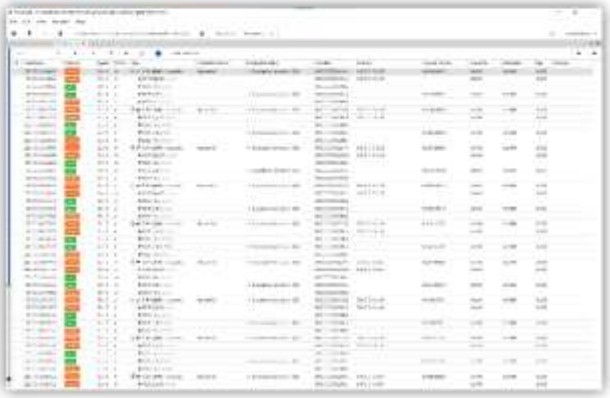


Statistics

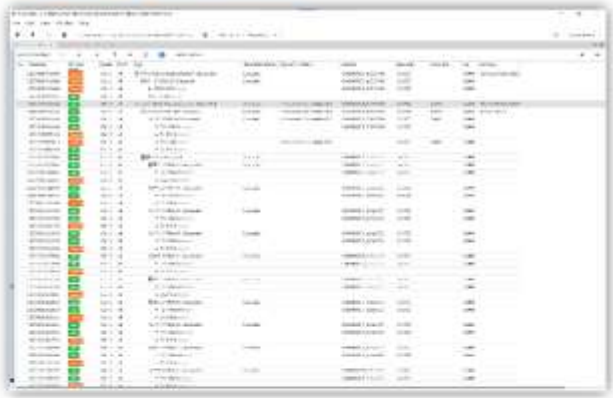
Transaction View Widgets

Events are collated into sequences that make up transactions on the bus. The transactions consist mainly of commands coming from one side of the stream and responses coming from the other. The transactions can be expanded (all events visible) or collapsed (only the transaction summary visible). Each row contains information about the packet it represents, and the columns are filled with the information from that packet.

PCIe Transactions

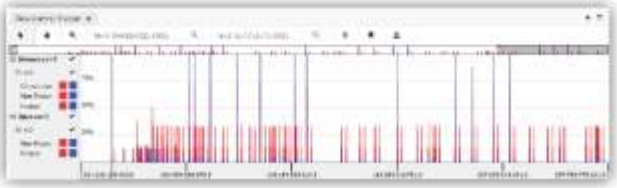


NVMe Transactions



Flow Control

Tracks flow control credit usage over the duration of the trace.

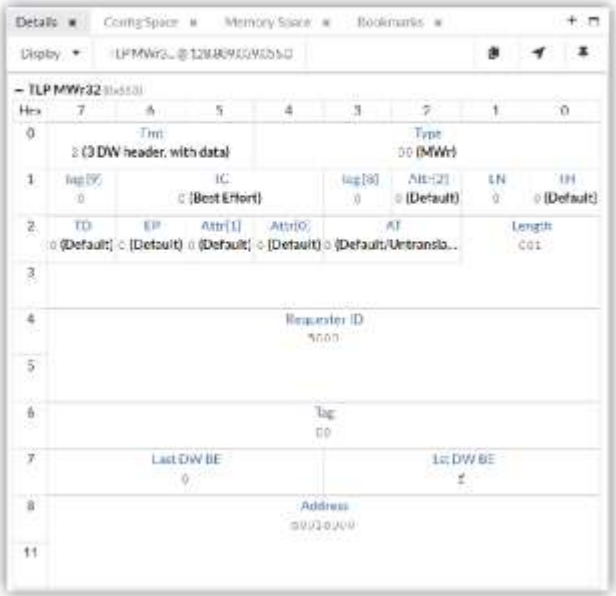


LTSSM

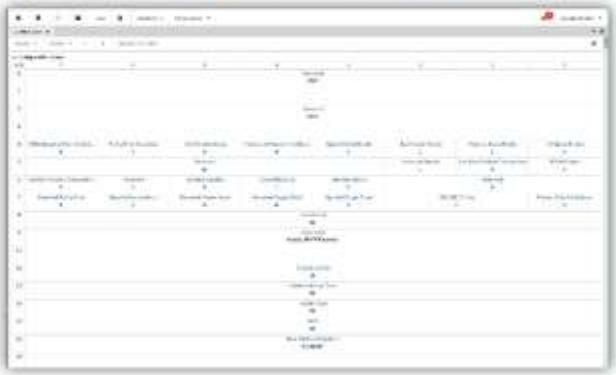
A timeline view of the "Link Training and Status State Machine" as seen in the trace data. The widget is arranged by up & down streams of the link. The data is shown as a tree that expands out and can be zoomed in to view data with finer granularity.

Packet Details

Provides a detailed breakdown of contents of an events payload.

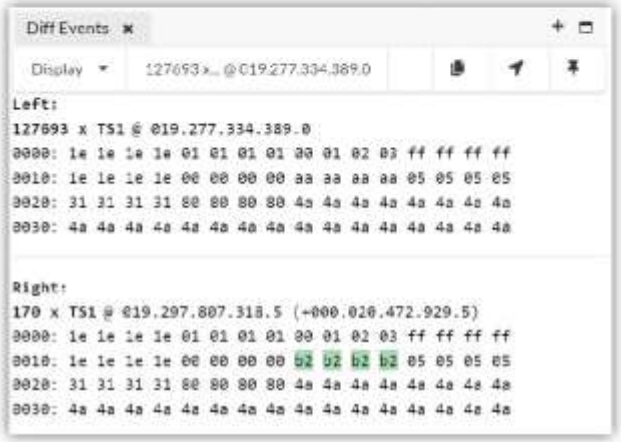


Config Space & Memory Space



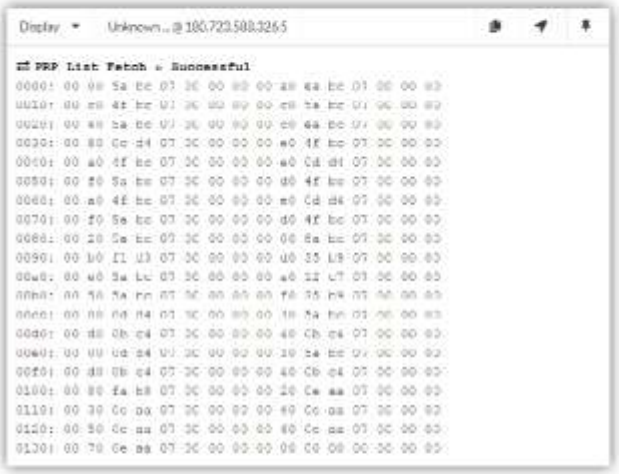
Diff Events

Highlights differences between two selected trace events.



Packet Data

Provides a detailed breakdown of the data/payload of an event.



Additional Widgets / Functions

- Real-time Link Statistics
- Global & User Bookmarks
- Quick Search: Contextual type search field
- Trace Summary Information
- Buffer Status
- LED's
- Concurrent user trace access (Events & Transactions Views)
- User, User Groups, User Permissions, & LDAP support

SerialTek Kodiak™

Next-Generation Gen5 PCIe/NVMe and CXL Analyzer

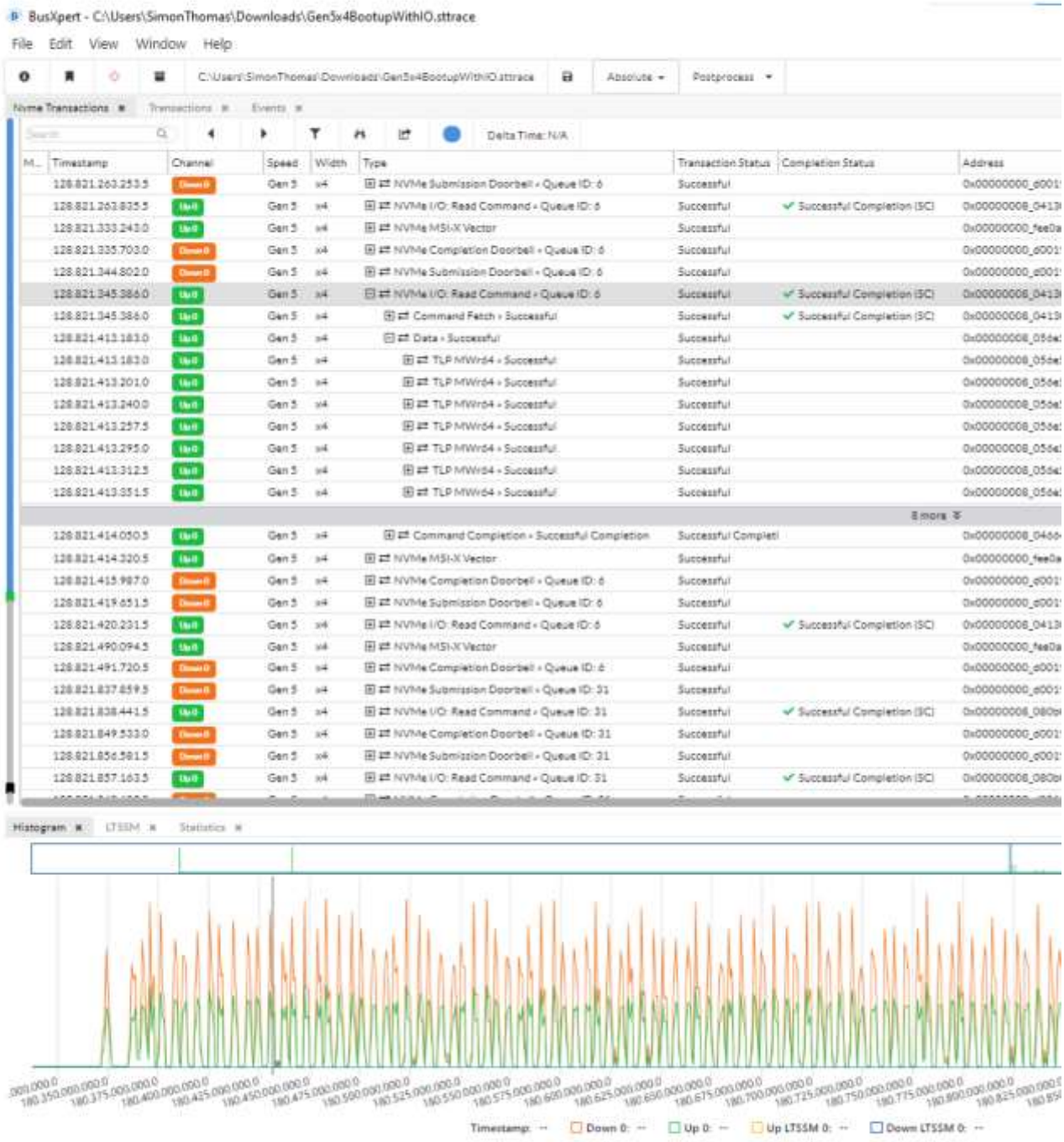
SerialTek
an ellsys company

Protocol Trace Widgets

Low-level and stacked protocol elements are hierarchically and chronologically displayed in easily configurable views.

Fast & Advanced Search

Quickly find events using a contextual search field. Includes multi-state search, and copy/paste from the trace views.



Precision Timestamping

Every event is given a precise timestamp and synchronized across all views. Measuring delta time is easy via ctl+ selecting any two events in the trace.

Whole Trace Timing Views

Every event, synchronized across all views, can be displayed in a custom graphing widget. Measuring is easily set via right-clicking the mouse.

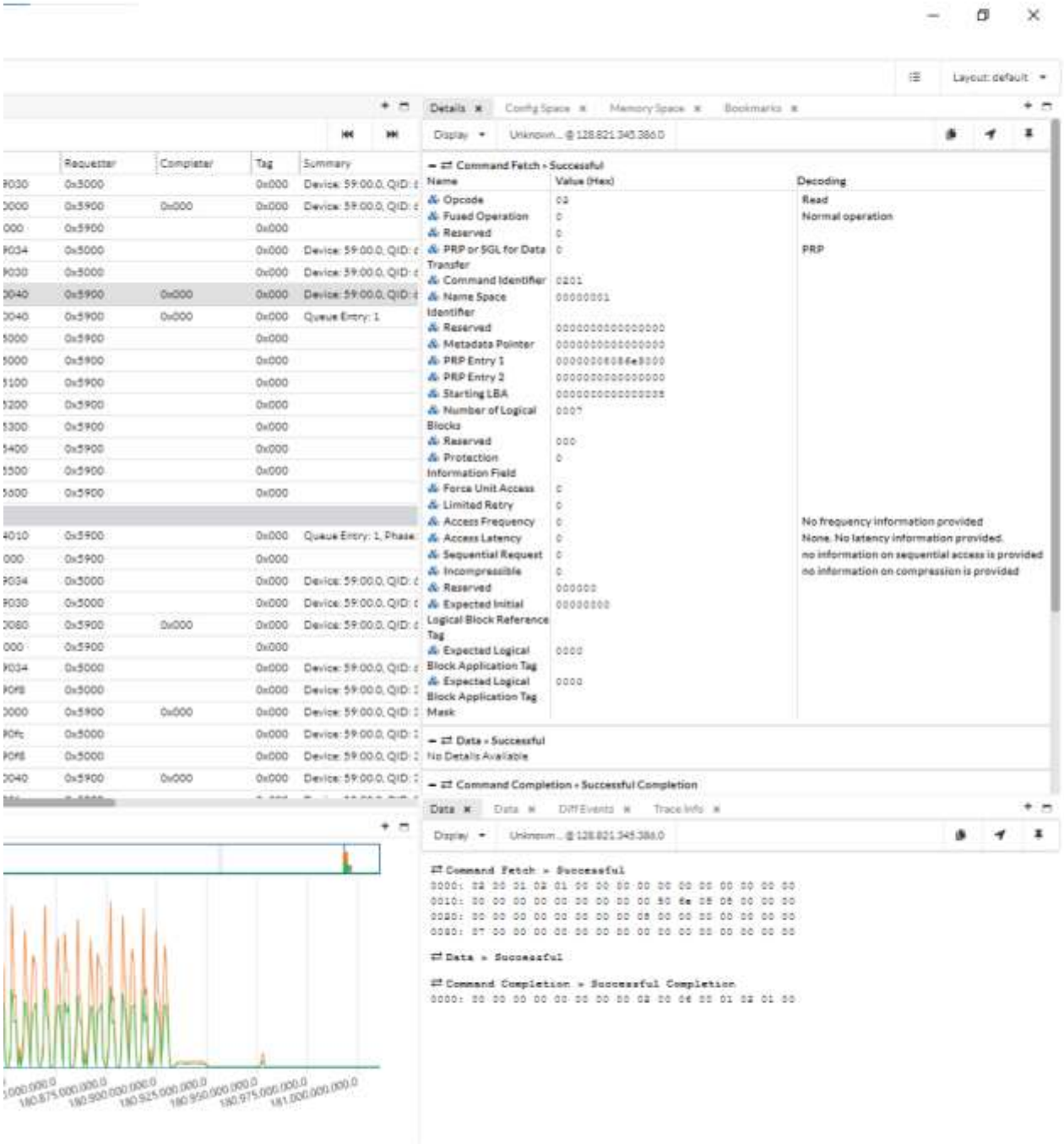
SerialTek Kodiak™

Next-Generation Gen5 PCIe/NVMe and CXL Analyzer

SerialTek
an eSays company

Fast & Advanced Hide/Show
Quickly show/hide links, sidebands, LTSSM, and protocol events.

Real-Time Protocol Processor
Automatically queries/saves the Configuration space, controller registers, and NVMe queues, whether recording or idle.



User-Configurable Views & Layouts
Easily modify tab sets and protocol views by adding or re- moving widgets, columns, and more.

Export Everything
All data is exportable to JSON or CSV. Exports are customizable.

SI-Fi™ Interposers

SerialTek's Gen5 (32.0 GT/s) PCI Express® (PCIe®) and Non-volatile Memory Express® (NVMe®) interposers with SI-Fi™ allow users to
monitor an unprecedented variety of PCIe and NVMe bus traffic with unparalleled power and ease.

Enabled by SerialTek's proprietary SI-Fi™ technology, users can save hours over legacy approaches requiring interposer calibration. This technology improves critical test coverage by providing high signal integrity, even over changing conditions, such as link training (LTSSM), power management, hot plug, reset, and other tests where the physical link/lane characteristics may change.

Each lane's analog signal is received at the probe's differential input and distributed to two separate phase matched differential outputs with a nominal gain of 0dB, allowing the host and device signals to pass through the interposer, allowing for real-world PCIe link training and easier set-up of the analyzer and DUT.

SI-Fi™ PCIe Gen5 Interposers continue SerialTek's TCO approach. With the focus on signal integrity, flexible, low-cost, SFF-8644-based cables connect each interposer to the analyzer. These cables are readily available and rated greater than 20GHz, resulting in uncompromised SI at all PCIe transfer rates.

All sideband signals are passed through the interposer from root complex (host) to controller (device), and all are made available to the analyzer for trigger, decode, and analysis.

Key Features

- SI-Fi™ Interposers require no calibration
- Supports PCI Express Gen 1.0, 2.0, 3.0, and 4.0
- Accurate capture of PCIe data traffic at line rates including:
  32.0GT/s (Gen5), 16.0GT/s (Gen4), 8.0 GT/s (Gen3), 5.0 GT/s (Gen2), and 2.5 GT/s (Gen1)
- Single U.2 / U.3 interposer supports single-port and dualport capture (only one analyzer is needed for dual-port)
- "Passive" tapping to avoid masking, hiding, or "cleaning up" electrical and/or link issues
- Low-cost, flexible, high-performance cabling for reliable analyzer to interposer connections

SerialTek Kodiak™
Next-Generation Gen5 PCIe/NVMe and CXL Analyzer

SerialTek
an ellsys company

Gen5 Slot/AIC Interposer

PCI Express slots are ubiquitous in ATX or ATX-based form factors in computing, storage, networking, and communication equipment applications. SerialTek's PCIe Gen5 slot interposers supports analysis of x1, x2, x4, x8, and x16 link-widths. SerialTek's PCIe Gen5 Slot (AIC) Interposers with SI-Fi technology are specially designed test adapters that are physically placed in between the PCIe host and a PCIe endpoint to intercept and relay a copy of the high-speed signaling and discrete data lines to the Kodiak PCIe Analysis system in real-time. All sideband signals are passed through the interposer from root complex (host) to controller (device), and all are made available to the analyzer for trigger, decode, and analysis. All relevant sidebands, including SMBus (e.g., NVMe-MI) from the host or from external/ third-party injection or generation tools are supported.

Overview

- Dimensions: 25 x 116 x 248 mm (1 x 4.5 x 9.7")
- Power connector: Molex 87427-0602
- Analyzer connectors: QSFP-DD
- Device connector: PCIe CEM slot x16 straddle mount connector
- Host module connectors: PCIe CEM x16 Edge fingers
- SMBUS injection connector: 2×5 pin 0.1" header, 3.3 Vdc
- REFCLK output connectors: 2x U.FL, AC coupled LPHCSL
- REFCLK output control connector: 2 pin 0.1" header
- REFCLK buffer control connector: 3 pin 0.1" header
- Sideband signal access connector: 2×9 pin 0.1" header, 3.3 Vdc

M.2 Interposer

M.2 is a specification supporting specifically keyed modules of different lengths that facilitate the addition or expansion of functions via a small form factor. The PCIe M.2 form factor is typically used for PCIe adaptation and small form factor NVMe SSD's. SerialTek's M.2 interposer supports all relevant keys, link-widths, and sidebands, including SMBus (e.g., NVMe-MI) from the host or from external / third-party injection or generation tools. SerialTek's PCIe Gen4 M.2 Interposers with SI-Fi technology are specially designed test adapters that are physically placed in between the M.2 port and an M.2 endpoint to intercept and relay a copy of the high-speed signaling and discrete data lines to the Kodiak PCIe Analysis system in real-time.

Overview

- Dimensions: 154 x 34 x 232 mm (6 x 1.3 x 9")
- Power connector: Molex 87427-0602
- Analyzer connectors: 2x SFF-8644
- Device connector: M.2 Socket 3, Key M, 22110, 2280, 2260, 2242, 2230
- Host module connectors: 2x MCIO 38 pin
- SMBUS injection connector: 2×5 pin 0.1" header, 3.3 Vdc
- REFCLK output connectors: 2x U.FL, AC coupled LPHCSL
- REFCLK output control connector: 2 pin 0.1" header, 3.3 Vdc
- REFCLK buffer control connector: 3 pin 0.1" header, 3.3 Vdc
- Sideband signal access connector: 2×9 pin 0.1" header, 3.3 Vdc

SerialTek Kodiak™
Next-Generation Gen5 PCIe/NVMe and CXL Analyzer

SerialTek
an elloys company

EDSFF Interposer

SerialTek's EDSFF interposer is mechanically modular and easily converts from E1.S to E1.L to E3.S form factors. The included EDSFF host adapters are easy to change and plug into a host system. High-quality cabling (instead of lossy PCB material) from the interposer to the storage enclosure preserves signal quality while adding flexibility, saving customers money, and providing for safe placement of the device under test (DUT) on a bench or in a test rack. SerialTek's EDSFF Interposer supports all relevant sidebands, including SMBus (e.g., NVMe-MI) from the host or from external / third-party injection or generation tools. SerialTek's PCIe Gen5 EDSFF Interposers with SI-Fi technology are specially designed test adapters that are physically placed in between the EDSFF port and an E1.x or E3.x EDSFF target to intercept and relay a copy of the high-speed signaling and discrete data lines to the Kodiak PCIe Analysis system in real-time.

Overview

- Dimensions: 154 x 34 x 232 (6 x 1.3 x 9")
- Power connector: Molex 87427-0602
- Analyzer connectors: 2x SFF-8644
- Device connectors: SFF-TA-1009
- Host connectors: SFF-TA-1009
- SMBUS injection connector: 2×5 pin 0.1" header, 3.3 Vdc
- REFCLKA output connectors: 2x U.FL, AC coupled LPHCSL
- REFCLKA output control connector: 2 pin 0.1" header
- REFCLKA buffer control connector: 3 pin 0.1" header
- REFCLKB output connectors: 2x U.FL,AC coupled LPHCSL
- REFCLKB output control connector: 2 pin 0.1" header
- REFCLKB buffer control connector: 3 pin 0.1" header
- Sideband signal access connector: 2×9 pin 0.1" header, 3.3 Vdc

PCIe Cable Interposer

PCIe External Cabling provides an electrically efficient channel to connect external or internal PCIe components directly to a root port, daughter card, backplane, adapter, or other PCIe based ports. Cable interposers support OCuLink (SFF-8611) and SlimSAS (SFF-8654) form factors and supports all relevant sidebands, including SMBus (e.g., NVMe-MI) from the host or from external / third-party injection or generation tools. SerialTek's PCIe Gen5 Cable Interposers with SI-Fi technology are specially designed test adapters that are physically placed in between the OCuLink (SFF-8611) or SlimSAS (SFF-8654) host port and its endpoint to intercept and relay a copy of the high-speed signaling and discrete data lines to the Kodiak PCIe Analysis system in real-time.

Overview

- Dimensions: 154 x 34 x 232 mm (6 x 1.3 x 9")
- Power connector: Molex 87427-0602
- Analyzer connectors: 2x SFF-8644
- Device connectors: SFF-8611 (OCuLink), SFF-8654 (SlimSAS)
- Host connectors: SFF-8611 (OCuLink), SFF-8654 (SlimSAS)
- SMBUS injection connector: 2×5 pin 0.1" header, 3.3 Vdc
- REFCLKA output connectors: 2x U.FL, AC coupled LPHCSL
- REFCLKA output control connector: 2 pin 0.1" header
- REFCLKA buffer control connector: 3 pin 0.1" header
- REFCLKB output connectors: 2x U.FL,AC coupled LPHCSL
- REFCLKB output control connector: 2 pin 0.1" header
- REFCLKB buffer control connector: 3 pin 0.1" header
- Sideband signal access connector: 2×9 pin 0.1" header, 3.3 Vdc

U.2/U.3 Interposer

SerialTek's U.2 and U.3 interposers support standard and extended length storage bays, and all relevant sidebands, including SMBus (e.g., NVMe-MI) from the host or from external / third-party injection or generation tools. SerialTek's PCIe Gen5 U.2 and U.3 Interposers with SI-Fi technology are specially designed test adapters that are physically placed in between the U.2/U.3 port and an U.2/U.3 target to intercept and relay a copy of the high-speed signaling and discrete data lines to the Kodiak PCIe Analysis system in real-time.

Overview

- Dimensions: 194 x 29 x 337 mm (7.6 x 1 x 13")
- Power connector: Molex 87427-0602
- Analyzer connectors: 4x SFF-8644
- Device connector: SFF-8639 receptacle
- Host connectors: SFF-8639 plug
- SMBUS injection connector: 2×5 pin 0.1" header, 3.3 Vdc
- REFCLKA output connectors: 2x U.FL, AC coupled LPHCSL
- REFCLKA output control connector: 2 pin 0.1" header
- REFCLKA buffer control connector: 3 pin 0.1" header
- REFCLKB output connectors: 2x U.FL,AC coupled LPHCSL
- REFCLKB output control connector: 2 pin 0.1" header
- REFCLKB buffer control connector: 3 pin 0.1" header
- Sideband signal access connector:
  2×9 pin 0.1" header, 3.3 Vdc

"I've been using protocol analyzers for 31 years and PCIe analyzers and interposers extensively for the past 5 years. We use them for important assignments that affect revenue and customer satisfaction," said ***John Wehman, Principal Applications Engineer at Phison Technology****. "With other analyzers I have had to abandon my testing many times, because I could not find a good quality signal lock. SerialTek's Kodiak analyzer and SI-Fi interposers have changed all that. I have 100% confidence in Kodiak's ability to achieve lock and give me the trace I need to do my job. Kudos to Ellisys and SerialTek for creating not only an electrically reliable platform, but the actual mechanical hardware itself is beautiful."*

Kodiak PCIe WebUI – Remote Access!

Access from Anywhere

- Easily connect to Kodiak via your web browser
- Real-time online system and recording status and I/O graphs
- Real-time Analyzer and Interposer configuration information, including cable status
- Online Trace file management

Interposers

- Real-time information for the PCIe Gen5 interposer and its status; including type, serial number, and connectivity state
- Kodiak automatically detects any good, bad, and unnecessary cable connections with easy to identify colors (green, orange, red)

System Settings

- Linux, easy for IT to manage
- Supports User Names, User Groups, and LDAP
- 1G and 10G ethernet ports (DHCP, Static IP settings)
- Update and/or verify Kodiak firmware remotely
- Update and/or verify Kodiak licenses remotely
- Remote system restart or full reset

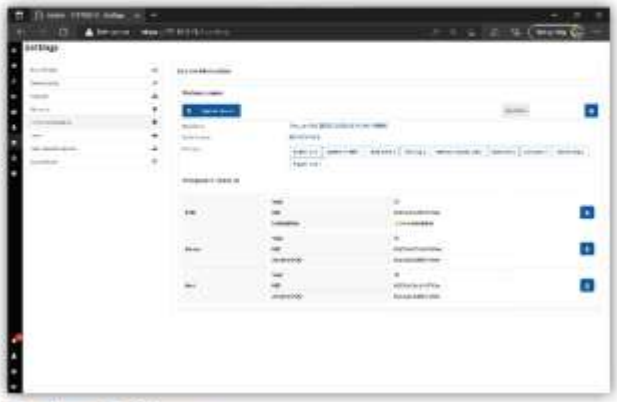
Management and System Reset

- Secure and manage Kodiak remotely
- Configure user permissions (read, read/write, update, admin, ...)
- Reboot and recover the analyzer
- Reset analyzer to factory settings (with or without user data)
- Remote factory reset

Trace Storage and Management

- Save traces to Kodiak's internal NVMe SSD storage
- Open traces while saved in Kodiak or download them to a client
- Zip traces in Kodiak and then download

Capture Settings



System Settings

System Users

Configurations and Purchase Information

Edition	Speed	Width	NVMe	CXL (future)	Buffer	Internal Storage	Multi-State Triggering	Multi-State Search	10GE	Dual-Port
Enterprise	Gen5	x16	Yes	Yes	144	2TB	Yes	Yes	2x	Yes
Enterprise	Gen5	x8	Yes	Yes	144	2TB	Yes	Yes	2x	Yes
Enterprise	Gen5	x4	Yes	Yes	144	2TB	Yes	Yes	2x	Yes
Professional	Gen5	x4	Yes	Yes	144	2TB	Yes	Yes	2x	No
Standard	Gen5	x4	Yes	n/a	72	2TB	Yes	Yes	1x	No

Kodiak PCIe Analyzers

Description	Code
Kodiak Gen5 PCIe x16 Protocol Analyzer Enterprise Edition	PK2A-G5-16-ENT
Kodiak Gen5 PCIe x8 Protocol Analyzer Enterprise Edition	PK2A-G5-08-ENT
Kodiak Gen5 PCIe x4 Protocol Analyzer Enterprise Edition	PK2A-G5-04-ENT
Kodiak Gen5 PCIe x4 Protocol Analyzer Professional Edition	PK2A-G5-04-PRO
Kodiak Gen5 PCIe x4 Protocol Analyzer Standard Edition	PK2A-G5-04-STD
Kodiak Gen4 PCIe x16 Protocol Analyzer Enterprise Edition	PK2A-G4-16-ENT
Kodiak Gen4 PCIe x8 Protocol Analyzer Enterprise Edition	PK2A-G4-08-ENT
Kodiak Gen4 PCIe x4 Protocol Analyzer Enterprise Edition	PK2A-G4-04-ENT
Kodiak Gen4 PCIe x4 Protocol Analyzer Professional Edition	PK2A-G4-04-PRO

SI-Fi Interposers

Description	Code
PCIe Gen5 x16 slot interposer	PEI-G5-16-AIC
PCIe Gen5 x8 slot interposer	PEI-G5-08-AIC
PCIe Gen5 x4 slot interposer	PEI-G5-04-AIC
PCIe Gen5 U2 interposer	PEI-G5-04-U2E
PCIe Gen5 U3 interposer	PEI-G5-04-U3E
PCIe Gen5 EDSFF interposer	PEI-G5-04-EDS
PCIe Gen5 M.2 interposer	PEI-G5-04-M2S
PCIe Gen5 MCIO Cable interposer	PEI-G5-04-MCS
PCIe Gen4 Slim-SAS Cable Interposer	PEI-G5-04-SCS
PCIe Gen5 x4 Premium Package; U2, U3, EDSFF, M.2	PEI-G5-04-PRE

SerialTek Kodiak™

Next-Generation Gen5 PCIe/NVMe Analyzer

SerialTek
an ellsys company

Technical Specifications

Kodiak Enclosure

- Dimensions: 443 x 67 x 305 mm (17 x 2.6 x 12")
- Weight: 7 kg (15 lbs)
- Mounting: 19" Rack Mount Option, Tilt Feet Option
- Ambient Operating Temperature: 5-35ºC at up to 2133m (7000 feet) altitude

Displays and Indicators

- Front Panel LCD: 800x320 4.6" WCGA, Touchscreen
- System Status: RGB LED

Front-Panel Connectors

- Interposer Connection: 4x QSFP-DD
- Ethernet (10 GbE): 2x SFP+ (10 GbE)
- Ethernet (1 GbE): RJ45
- PCIe Interface: 2x OCuLink
- USB Interface: 2x USB 3.1 Type A

Rear-Panel Connectors

- Power: IEC C13, 90-264 Vac, 47-63 Hz
- Clock Out: SMA, 50 Ω, 3.3 Vdc, 10 MHz
- Clock In (10 MHz): SMA, 50 Ω, 3.3 Vdc, 10 MHz
- Trigger Out: SMA, 50 Ω, 3.3 Vdc
- Trigger In: SMA, 50 Ω, 3.3 Vdc
- Maintenance: RJ45, USB Micro-B (Not for customer use)

Interposer Power Unit (Common)

- Input: 100-240 Vac/50-60 Hz
- Output: 5 Vdc
- Power: 50 W
- Plug: Molex 039-01-2060
- Safety: UL, CUL, CE, TUV-GS, PSE
- EMI: CE, FCC
- Environmental: ROHS, WEEE, VI

M.2 Interposer

- Dimensions: 154 mm(W) x 34 mm(H) x 232 mm(L) (6 x 1.3 x 9")
- Power connector: Molex 87427-0602
- Analyzer connectors: 2x SFF-8644
- Device connector: M.2 Socket 3, Key M, 22110, 2280, 2260, 2242, 2230
- Host module connectors: 2x MCIO 38 pin
- SMBUS injection connector: 2x5 pin 0.1" header, 3.3 Vdc
- REFCLK output connectors: 2x U.FL, AC coupled LPHCSL
- REFCLK output control connector: 2 pin 0.1" header, 3.3 Vdc
- REFCLK buffer control connector: 3 pin 0.1" header, 3.3 Vdc
- Sideband signal access connector: 2x9 pin 0.1" header, 3.3 Vdc

U.2/3 Interposer

- Dimensions: 194 x 29 x 337 mm (7.6 x 1 x 13")
- Power connector: Molex 87427-0602
- Analyzer connectors: 4x SFF-8644
- Device connector: SFF-8639 receptacle
- Host connectors: SFF-8639 plug
- SMBUS injection connector: 2x5 pin 0.1" header, 3.3 Vdc
- REFCLKA output connectors: 2x U.FL, AC coupled LPHCSL
- REFCLKA output control connector: 2 pin 0.1" header
- REFCLKA buffer control connector: 3 pin 0.1" header
- REFCLKB output connectors: 2x U.FL,AC coupled LPHCSL
- REFCLKB output control connector: 2 pin 0.1" header
- REFCLKB buffer control connector: 3 pin 0.1" header
- Sideband signal access connector: 2x9 pin 0.1" header, 3.3 Vdc

X4 Slot Interposer

- Dimensions: 25 x 116 x 248 mm (1 x 4.5 x 9.7")
- Power connector: Molex 87427-0602
- Analyzer connectors: 2x SFF-8644
- Device connector: PCIe CEM slot x16 straddle mount connector
- Host module connectors: PCIe CEM x4 Edge fingers
- SMBUS injection connector: 2x5 pin 0.1" header, 3.3 Vdc
- REFCLK output connectors: 2x U.FL, AC coupled LPHCSL
- REFCLK output control connector: 2 pin 0.1" header
- REFCLK buffer control connector: 3 pin 0.1" header
- Sideband signal access connector: 2x9 pin 0.1" header, 3.3 Vdc

Maintenance and Licensing

- Free lifetime software updates – no maintenance fees
- Free full-featured web browser and standalone software – easily share traces between computers and colleagues and replay captured traffic
- Use SerialTek hardware on any computer – no additional licenses needed

Warranty

- 1, 2, and 3 year limited warranties available, Basic and Standard Editions
- Six-month limited warranty, Interposers

Minimum Requirements

- Intel Core, 2 GHz or compatible processor
- 4 GBytes of RAM
- 1280 x 1024 display resolution with at least 65,536 colors
- 64-bit OS only (Windows 7, Ubuntu 14, Centos7 or higher)
- 1GbE controller

More information at: www.serialtek.com/kodiakgen5

rev06082022

3. PCIe Gen 4/5/6 NVMe SSD 性能/功能测试

图 3-1

The SANBlaze VirtuaLUN has been an excellent tool for traffic generation and performance testing at UNH-IOL plugfests for Fibre Channel and NVMe. We're happy to use the VirtuaLUN as a resource in our lab for stressing systems when performing interoperability testing."

— David Woolf

Engineering Manager – Datacenter, UNH-IOL

3.1 SanBlaze RM5 & DT5 Gen 5 测试设备

图 3-2 PCIe Gen 5 RM5

图 3-3 PCIe Gen 5 DT5

3.1.1 Sanblaze Gen5 测试设备规格说明

3.1.1.1 产品端口配置

- **RM5 规格**

提供 16 个端口，支持 Gen5 EDSFF, U.2, U.3, M.2 等接口类型

- **DT5 规格**

4 个插槽

– 插槽 0、1、2 内置 PCIe Gen5 Riser；提供 U.2 / U.3 / M.2/EDSFF（短）等接口类型

– 插槽 3，PCIe Gen5 x16 插槽

图 3-4

3.1.1.2 软件可控的硬件特性

- **双/单端口可选**
- **风扇速度控制**
- **上电/断电测试每个设备**
- **每个 device 支持 PERST（reset）和 HotPlug**
- **电压拉偏 +/- 15%**
- **SRIS 支持和内置测试（可选）意外/优雅 SSD 拔插测试**
- **VDM（可选）**
- **SMBus 和带内 MI 测试**
- **通过 VDM、SMBus 和带内 下载固件升级**
- **Gen5 TLP/DLLP/接收错误监控以进行 PCIe Gen5 验证**
- **L0、L1、L1.1、L1.2 sub-state 低功耗支持**
- **** RM5 支持 38 个温度传感器，以及在 70 摄氏度下面提供每槽位 25W 功耗支持**

SanBlaze 正式发布了 2021/9 月份发布了针对 PCIe Gen 5 NVMe SSD 的研发测试设备，2021/12 正式出货，可以大大加速针对 PCIe Gen 5 SSD 的各项性能，功能，协议层兼容性的测试。参考下面的新闻稿。

美国东部标准时间 2021 年 12 月 16 日

马萨诸塞州利特尔顿--(BUSINESS WIRE)--全球领先的高级存储测试和验证技术提供商 SANBlaze Technology Inc. 今天宣布推出业界首个支持 NVMe® over PCIe® Gen5 验证和合规性测试的平台. SBExpress-RM5™ 平台为开发、质量保证、验证和制造团队提供广泛的测试功能，并包括公司的 Certified by SANBlaze (SBCert™) 一致性测试套件，该套件被公认为行业基准。

SBExpress-RM5 是一款 16 托架企业级 NVMe 测试设备，支持从 Gen1 到 Gen5 的热插拔和 PCIe 速度。该系统采用独特的模块化“Riser”设计，可实现用户可配置的可变插槽支持，以及对所有 Gen5 SSD 接口规格的现场升级支持，包括 U.2、M.2、EDSFF 和新的 E3/EDSFF 等接口。

在通用和 SRIS/SRNS 模式下拉偏并且测量电压、电流和功率、毛刺信号注入和测试扩频时钟 (SSC) 或传统时钟的能力使 SBExpress-RM5 在 NVMe SSD 测试领域脱颖而出。数据完整性通过一套全面的读/写/比较测试进行验证，运行 IO 时出现电源故障等例外情况，以及作为 SANBlaze 认证测试套件一部分的内置“Write Atomicity”测试。可以通过设备的 Web 界面或通过系统标配的 Python、XML 和 REST API 访问测试。 SBExpress™ Gen5 软件包括超过九百个测试脚本，可在客户实验室进行 IOL 测试，然后再进行正式测试，以及 ZNS、VDM 和 TCG Opal 验证。

和 SerialTek 以及 Ellisys 的战略合作加快了 RM5 的上市时间

SBExpress-RM5 的发布标志着 SANBlaze 与 SerialTek 和 Ellisys 合作推出的第一个主要产品，这两家公司专门从事存储、数据中心、消费电子和物联网技术的高级协议测试解决方案。这种合作帮助 SANBlaze 通过额外的资源和专业知识缩短了其 SBExpress-RM5 的上市时间，显著缩短了这一行业首创交付的开发周期。这种公司间共生的合作关系将加速创新并带来新的独特能力和协同效应，也使 SANBlaze 客户受益匪浅。

“SANBlaze 很高兴宣布推出业界首款 NVMe over PCIe Gen5 测试系统，将我们的 Certified by SANBlaze 自动化测试套件的领先地位扩展到 Gen5 架构，”SANBlaze 首席执行官 Vince Asbridge 表示。 “利用 SerialTek 的 Gen5 专业知识显着缩短了我们的上市时间，并能够及早使用最先进的工具，例如 SerialTek 的创新 Kodiak™ Gen5 PCIe 分析仪，我们用它来加快测试以确保 SBExpress-RM5 系统满足 PCIe Gen5 的协议和严格的信号完整性要求。”

SerialTek 首席执行官 Paul Mutschler 表示：“SerialTek 和 SANBlaze 之间的合作对两家公司都带来了巨大的好处，将 SANBlaze 团队的 NVMe 专业知识与 SerialTek 团队的

PCIe 和信号完整性领导能力相结合。”“结合这两个世界级工程团队的人才，加快了 SBExpress-RM5 NVMe Gen5 测试系统和 Kodiak Gen5 PCIe 分析仪的推出，从而为 Gen5 及更高版本提供了一个紧密集成的分析和认证平台。”

“我们很高兴有机会与 SANBlaze 这样的高品质的组织合作，并与 SerialTek 建立互补且已经富有成效的关系，”Ellisys 总裁兼首席执行官 Mario Pasquali 说。“我们的整个组织期待延续 SANBlaze 提供优质产品和与其客户群专家合作的传统，以及将极大地使我们所有客户受益的奇妙协同效应。”

交期、配置和定价

模块化“riser 转接卡”设计允许每个插槽支持单端口或双端口 NVMe 驱动器。每个插槽都可以直接从软件在双端口和单端口之间切换，以便于在任何配置下测试两种驱动器类型。所有转接卡都支持使用 NVMe 设备配置进行功率控制和测量，包括 U.2、M.2 和 EDSFF，包括新的 E3 EDSFF 接口。

有关可用性、配置和定价的更多信息，请联系 sales@saniffer.com。

3.2 SanBlaze RM4 & DT4 Gen 4 测试设备

图 3-5 SanBlaze 机架式 RM（Rack Mounted）测试主机和盘柜实拍图（支持 16 个 Gen 4 U.2 SSD）

图 3-6 SanBlaze 桌面型 DT（Desk Top）测试主机和盘柜实拍图（支持 6 个 Gen 4 SSD）

美国 SanBlaze 针对 PCIe Gen 4/5/6 NVMe SSD 的测试设备是 UNH IOL 官方认证的测试工具，支持 NVMe 1.4 conformance test，以及相关 feature 例如 ZNS 等，该设备的的测试盘柜可以直接放入温箱进行测试，支持 75~85 度。

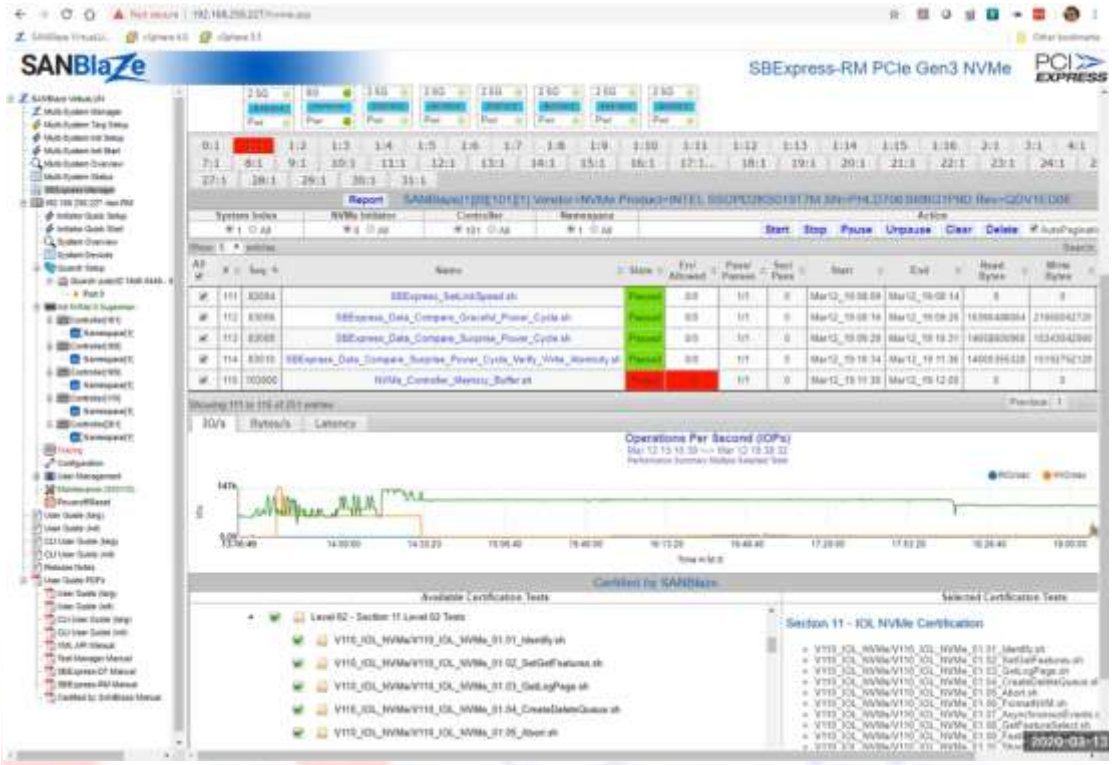


图 3-7 UNH IOL CTS 测试界面

3.3 SanBlaze 重点特性

- **负载压力测试，包括带宽，IOPS，延迟测试**
- **数据读写一致性校验测试**
- **900+个自动化定制脚本 NVMe 测试脚本支持 UNH/IOL**
- **电压/单流/功耗测试**
- **掉电测试，支持数据校验/Atomicity 校验**
- **完整 Quarch 热插拔集成，导入信号毛刺测似乎**
- **多盘位机箱支持**
- **NVMe 1.4 规范完整支持**
- **SGL/PRP 支持可变 block size/metadata/Bit Buck**
- **针对 VF 的完整 SR-IOV 支持**
- **支持 VDM/PCI/MTCP, ZNS, SRIS, TCP Opal**
- **MI 1.0 Conformance 脚本支持**
- **Gen 4/5/6 完整解决方案**
- **针对 NVMe SSD 和盘柜的环境测试支持（0C to 85C）** （RM4/5 功能）**

- 纯软件方案/基于定制的硬件环境
- 制动选择 Single 或者 Dual Port NVMe SSD
- 基于 Web 的 GUI 管理

3.4 SanBlaze 系统功能

- PCIe® Gen 4/5/6 支持
- 16 个盘位，通过前面板方便插拔
- 通过软件控制进行热插拔，插槽功耗以及检测 SSD 盘是否在位
- 通过 U.2 Riser 卡实现配置 16 个盘位为 Single Port 或者 Dual Port NVMe SSD，也支持 U.3 Single Port Riser 卡或者 U.3 Dual Port Riser 卡，以及 E1, E3, AIC 等多种接口
- 可现场更换 Riser 卡支持各种接口 SSD
- 针对每个盘可以进行电压、电流和功耗测量
- 对于分体式 RM 系列设备，采用优化的散热设计，气流无阻塞
- 由温度触发速度自动调节的 6 组风扇设计（用户可以通过 config 文件控制）
- 15mm 盘的 LFM 为 1,373
- 38 个温度监控点
- 75 摄氏度工作时候的功耗为每插槽 25W
- 电压拉偏范围+/- 15%

下图是 dual-port 测试场景，配合 SerialTek PCIe Gen4 analyzer 分析 dual-port 流量。从分析仪前面板的 LCD 显示屏可以看到 port 1, 2 两个链路。

图 3-8

3.5 SanBlaze 软件功能

- 测试范围涉及 NVMe 规范的所有方面
- 支持 UNH IOL Conformance 测试
- 支持基于 SMBus 对于 NVMe-MI (Management Interface)进行测试
- 支持 SGL, SR-IOV, 完整 namespace 的控制和预留
- 支持同时测试多个 NVMe SSD 盘
- 支持通过脚本格式实现发送特定或者定制的 op code
- 读/写/比较测试
- 故障注入
- Vendor-unique 命令支持
- 驱动和测试单个或者多个 NVMe 目标设备

3.5.1 NVMe SSD 测试基本功能介绍

- NVMe spec 1.3/1.4 feature support
  - Namespace Management/Dual Port
  - PRPs + SGLs (including Bit Bucket SGL)
  - SR-IOV
  - Reservations
  - Controller Memory Buffer
  - Mutli Streams
  - MI (SMBus and NVMe)
  - TCG
- Customizable controller/namespace parameters
  - All namespace formats supported
    - i.e. all combinations of LBAF, MSET, PI, PIL
  - Modification of number of IO queues/IO queue depth/queue priority (for Weighted Round Robin arbitration)
- Error injection/negative testing
  - NVMe protocol errors
  - T10 DIF/DIX guard/ref tag errors
  - NVMe resets
  - Quarch integration/power glitch/data integrity
  - Dual Port Failover
  - Variable Block Sizes
- Conformance
  - UNH-IOL conformance spec V13

图 3-9

3.5.2 NVMe SSD 测试 New Feature 介绍

- 8.1 Software Release
  - SRIS/SNRS clock support
    - Script base Clock Validation testing
  - Zoned Namespace (ZNS)
    - Script base test structure for ZNS
    - ZNS command set support
    - ZNS validation tests (capabilities)
    - DATA Center applications
      - Supported by Microsoft and Facebook
  - Vendor Defined Messaging (VDM)
    - MTCP messaging over PCI Bus
    - MI (SMBus and NVMe)
  - TCG
    - Script Based test structure for TCG
  - Python Based API for test infrastructure support
- Conformance
  - UNH-IOL conformance spec V13 (June plugfest)
- Available in Software only deployment for Gen 4
  - Supporting AMD 7702
  - Supporting AMD 570X

图 3-10

3.5.3 NVMe 预封装测试脚本（涵盖 18 大类测试, 1000+个测试用例）

该脚本测试完成可以生成完整的测试报告，针对 failed 的条目会给出详细的测试 log 用于用户进行问题分析。支持 Shell 和 Python，所有的脚本均提供源代码，用户可以在之上进行二次定制开发，参见下面的测试大类简介。

- **NVMe Commands test**
- **NVMe I/O Tests**
- **NVMe Resets (all supported reset methods)**
- **NVMe Namespace Management**
- **NVMe Basic Management Commands**
- **NVMe-MI Full Command Set**
- **NVMe Dual Port Drive Tests**
- **NVMe SBExpress Hotplug and Link Testing**

- NVMe Quarch Testing Pull/Plug Glitch
- NVMe Miscellaneous Commands (e.g. SR-IOV)
- NVMe ZNS test
- NVMe VDM test
- NVMe Clocking Mode Test (SRIS)
- NVMe TCG Opal/Ruby test
- NVMe DSSD conformance test
- IOL NVMe Certification
- IOL NVMe-MI Certification
- SSD Endurance JEDEC Spec. (long runtime)

图 3-11

3.5.4 SanBlaze Certified 测试用例集

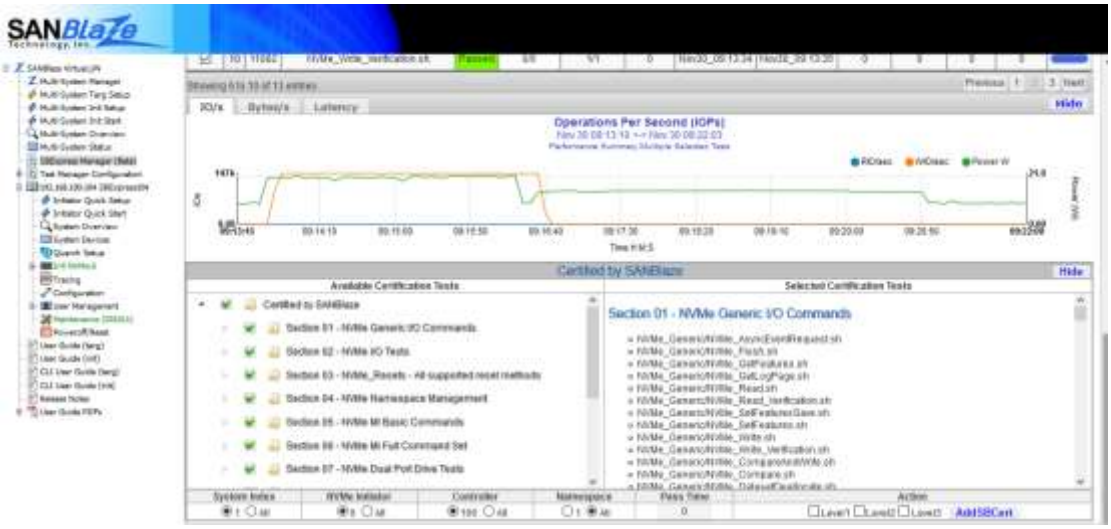


图 3-12

3.5.5 SanBlaze Certified 测试过程

图 3-13

3.5.6 SanBlaze Certified 总结和 Log

图 3-14

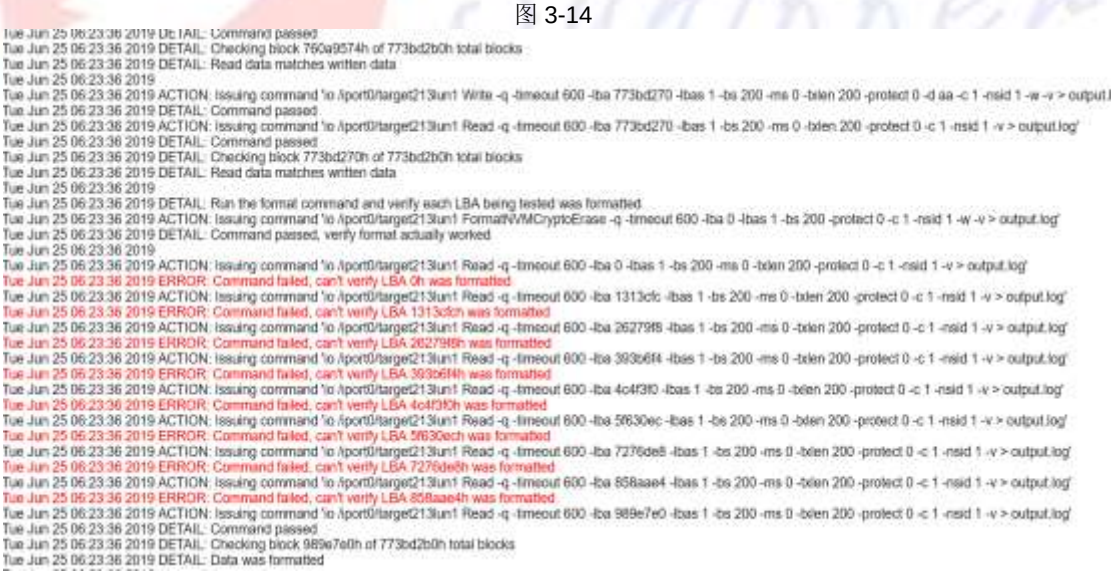


图 3-15

3.5.7 SanBlaze Certified 测试报告

测试完毕后可以将测试结果生成 pdf 报告，提供非常详细的信息供测试部门进行问题分析，一般超过 500 页的报告，不仅在前面几页给你一个测试结果综述（executive

summary）,之后的几页会详细到预装测试脚本每个章节的 pass, fail, skipped 的分析，然后深入到每个脚本的测试细节，含详尽的 log。

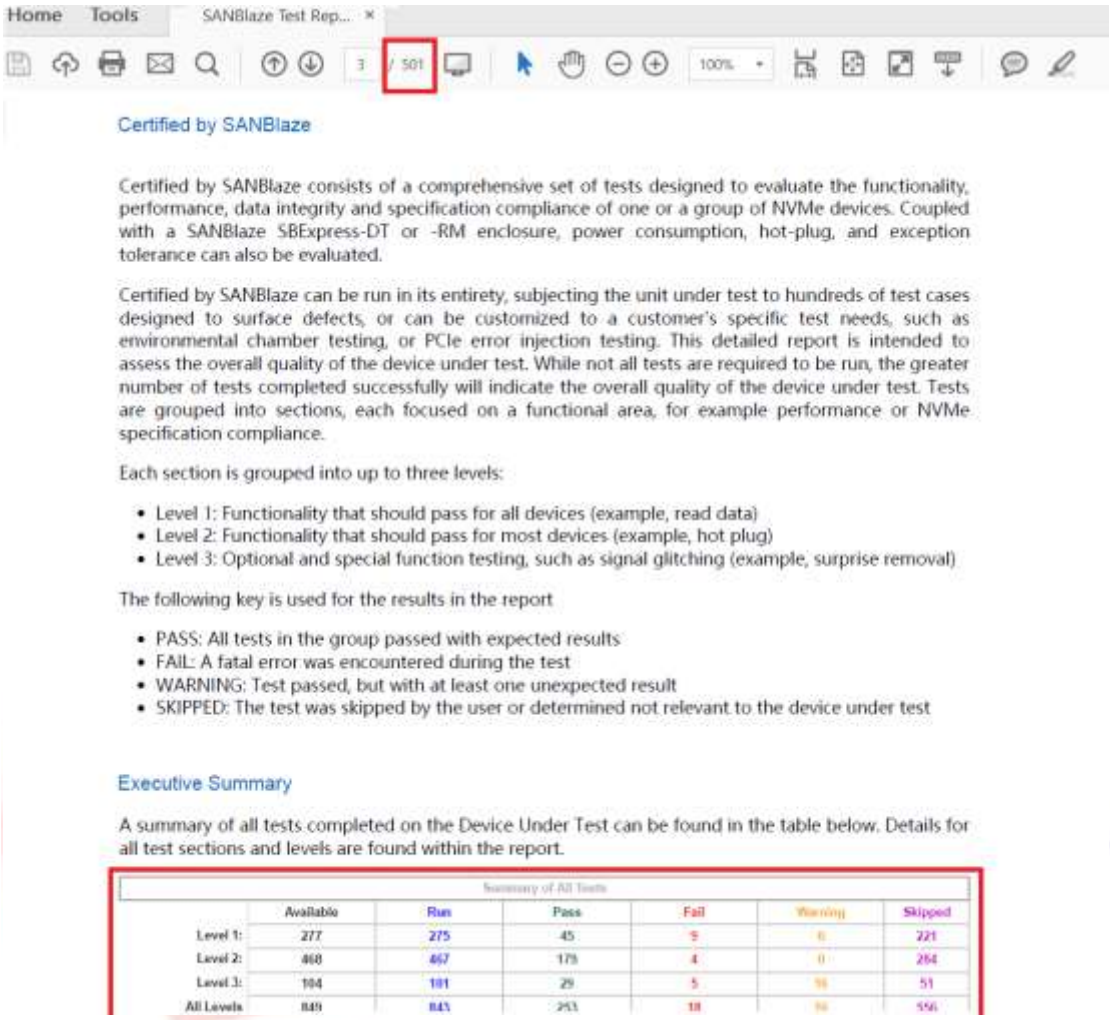


图 3-16

下面是测试脚本各个部分 pass, fail, skipped 的例子。

Summary of Test Results by Section

A summary of all test sections run against the Device Under Test is found below. Details for each section and detailed test results follow.

Section 1:	NVMe_Generic Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	20	20	19	1	0	0
Level 2:	24	24	10	0	0	14
Level 3:	0	0	0	0	0	0
All Levels	44	44	29	1	0	14

Section 2:	IO_Tests Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	26	26	26	0	0	0
Level 2:	156	156	156	0	0	0
Level 3:	26	26	26	0	0	0
All Levels	208	208	208	0	0	0

Section 3:	NVMe_Beasts Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	0	0	0	0	0	0
Level 2:	4	4	3	1	0	0
Level 3:	0	0	0	0	0	0
All Levels	4	4	3	1	0	0

Section 4:	NVMe_Namespace Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	0	0	0	0	0	0
Level 2:	1	1	0	0	0	1
Level 3:	32	32	3	0	16	13
All Levels	33	33	3	0	16	14

Section 5:	NVMe_MI_Basic Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	0	0	0	0	0	0
Level 2:	3	3	0	0	0	3
Level 3:	0	0	0	0	0	0
All Levels	3	3	0	0	0	3

Section 6:	NVMe_MI Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	43	43	0	0	0	43
Level 2:	43	43	0	0	0	43
Level 3:	0	0	0	0	0	0
All Levels	86	86	0	0	0	86

Section 7:	NVMe_DualPort Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	0	0	0	0	0	0
Level 2:	1	1	0	0	0	1
Level 3:	5	5	0	0	0	5
All Levels	6	6	0	0	0	6

Section 9:	NVMe_Quarch Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	0	0	0	0	0	0
Level 2:	0	0	0	0	0	0
Level 3:	5	5	0	5	0	0
All Levels	5	5	0	5	0	0

Section 10:	NVMe_Misc Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	0	0	0	0	0	0
Level 2:	0	0	0	0	0	0
Level 3:	3	3	0	0	0	3
All Levels	3	3	0	0	0	3

Section 11:	V14_IOL_NVMe Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	0	0	0	0	0	0
Level 2:	99	99	10	3	0	86
Level 3:	0	0	0	0	0	0
All Levels	99	99	10	3	0	86
Section 12:	V14_IOL_NVMe_MI Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	68	68	0	0	0	68
Level 2:	68	68	0	0	0	68
Level 3:	0	0	0	0	0	0
All Levels	136	136	0	0	0	136
Section 13:	NVMe_2165_Certification Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	102	102	0	0	0	102
Level 2:	68	68	0	0	0	68
Level 3:	28	28	0	0	0	28
All Levels	198	198	0	0	0	198
Section 14:	NVMe_TCG Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	14	14	0	8	0	6
Level 2:	0	0	0	0	0	0
Level 3:	0	0	0	0	0	0
All Levels	14	14	0	8	0	6
Section 15:	SSD_Endurance Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	0	0	0	0	0	0
Level 2:	0	0	0	0	0	0
Level 3:	2	2	0	0	0	2
All Levels	2	2	0	0	0	2
Section 16:	TestManager Loop 1 of 1 loops					
	Available	Run	Pass	Fail	Warning	Skipped
Level 1:	2	2	0	0	0	2
Level 2:	0	0	0	0	0	0
Level 3:	0	0	0	0	0	0
All Levels	2	2	0	0	0	2

图 3-17

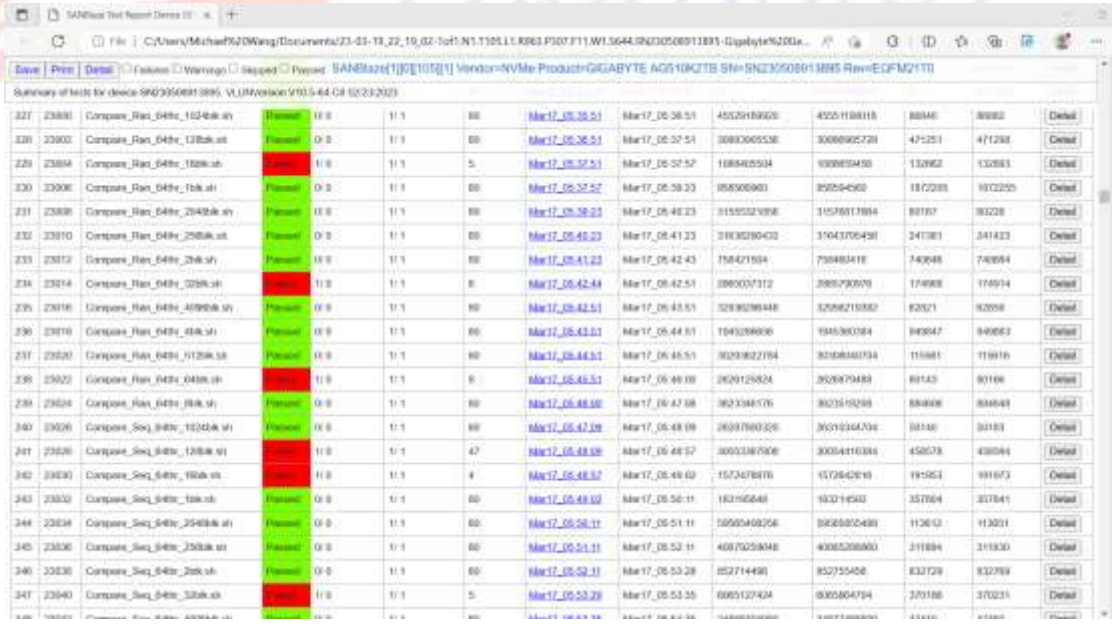


图 3-18

3.5.8 NVMe 测试 - SanBlaze 前面板模式切换视图

除了在前面板通过物理的开关进行 single port / dual port 切换外，也可以很容易地在 web GUI 上面切换 single port / dual port 模式。

- **Sanblaze DT4 Web GUI，提供 6 个测试端口**

图 3-19

- **Sanblaze RM4 Web GUI，提供 16 个测试端口**

图 3-20

3.6 VDM, ZNS, SRIS,TCG,双端口,DSSD,CMB/HMB/T10 DIF_DIX 测试

关于使用 SanBlaze 测试很多企业级 NVMe SSD 的特性，例如 VDM, ZNS, SRIS,TCG, CMB/HMB/T10 DIF_DIX 等等功能，请到下面的链接下载压缩包。里面含有如图所示的白皮书供参考。

https://www.saniffer.com/中文/文档下载/

CMB_HMB_White_Paper-V3
T10 DIF_DIX Verification with SANBlaze
Triggering a PCIe Analyzer from your SBExpress System
NVMe Power and Reset Testing with the SBExpress-RM4
Python Support for Verifying NVMe SSDs
Verification of SRIS_SRNS Clocking
FEATURE validation using NVM compare
Verification of TCG Opal SSC
Verification of ZNS
SanBlaze Testing NVMe-MI Over PCIe VDM

图 3-21

3.6.1 NVMe-MI Over PCIe VDM 测试

SANBlaze 引入了使用 PCIe MCTP 在 PCIe VDM)上测试管理接口 (MI) 命令的能力。

NVM Express 组织 (https://nvmexpress.org/) 定义的 Non-Volatile Memory Express 规范定义了一个寄存器级接口，允许带内主机软件与 NVMe 子系统进行通信。

NVMe 规范进一步定义了多种机制来管理 NVMe 存储设备或 NVMe 机箱。一种这样的机制，NVMe-MI（管理接口）允许应用程序与 NVMe 存储设备进行带外通信。 MI 规范的详细信息可从 nvmexpress.org 的以下链接 (NVM-Express-Management-Interface) 获得。

SANBlaze SBExpress-RM4 和 SBExpress-DT4 将在 VDM 上测试 MI 的能力引入到 SANBlaze NVMe 测试系统的现有功能中，现在允许使用 NVMe-MI 发送和接收在 PCIe 上对 MI 进行带内测试，以及带外测试使用 MI over SMBus 或 PCIe VDM，完善 NVMe-MI 的完整测试系统。

本白皮书介绍了 SANBlaze SBExpress VDM 功能，并为您使用 SBExpress 系统进行 MI over VDM 测试提供了一个起点。

NVMe-MI 规范为 NVMe 存储设备的带外和带内管理定义了架构和命令集。SanBlaze 支持下图中的两种实现方式。

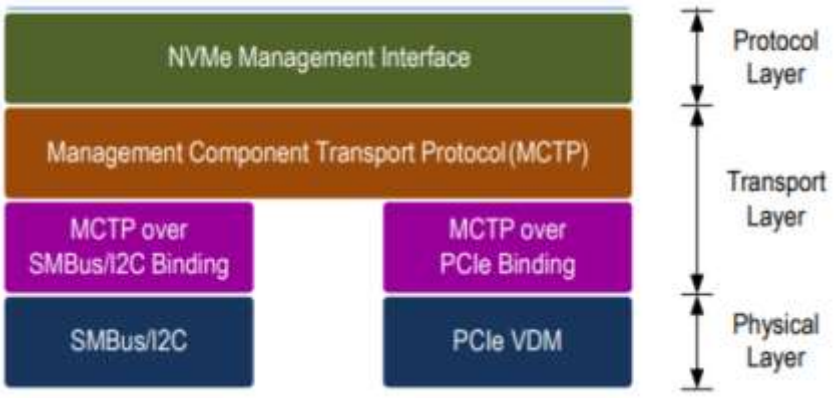


图 3-22

NVMe-MI 定义了 NVMe 存储设备的关键功能：

- **发现存在的 NVMe 存储设备并了解每个 NVMe 存储设备的功能**
- **存储有关主机环境的数据，以便管理控制器或其他实体稍后查询数据**
- **健康和温度监测**
- **多个并发命令可防止长延迟命令阻塞监控操作**
- **带外机制与主机处理器和操作系统无关**
- **VPD 的标准格式和定义的读/写 VPD 内容的机制**
- **保持静态数据的安全性**

下图是 MCTP over PCI Express Vendor Defined Message (VDM) 的 packet 格式。

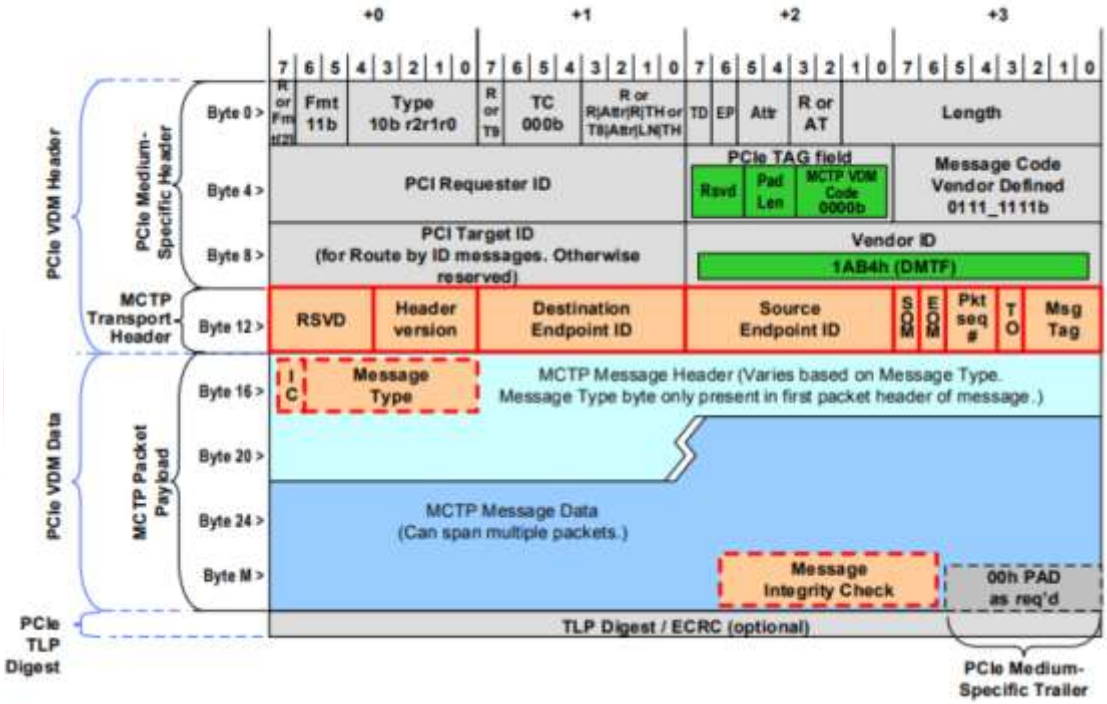


图 3-23

获取测试白皮书详情，请登录 https://www.saniffer.com/中文/文档下载/ ，下载 SanBlaze White Papers，然后查看 SanBlaze Testing NVMe-MI Over PCIe VDM.pdf 文档。

3.6.2 ZNS(Zoned Name Space)测试

SANBlaze 宣布推出 ZNS（分区命名空间）验证，让您可以快速有效地测试和验证固态驱动器 (SSD) 的 ZNS 实施。本白皮书介绍了 ZNS 并描述了如何使用 SANBlaze SBExpress 测试和验证系统验证您的 SSD 是否正确实现了所有 ZNS 功能。

理解 ZNS

NVMe™ 分区命名空间 (ZNS) 是 NVM Express™ 组织标准化的技术提案。它将命名空间的逻辑地址空间划分为区域。每个区域都提供一个逻辑块地址 (LBA) 范围，该范围必须按顺序写入，如果再次写入，则必须明确重置。此操作原理允许创建的命名空间暴露设备的自然边界，并将内部映射表的卸载管理提供给主机。

为什么要为 SSD 使用 ZNS？

由于闪存特性，SSD 本质上是分区设备。页是 NAND 闪存中支持写操作的最小区域，由同一字线上的所有存储单元组成。擦除块是闪存中可以在单个操作中擦除的最小区域。页和块大小因制造商和闪存代而异。例如，19nm 64Gb MLC 闪存包含 16KB 页面大小和 4MB 块大小。 16KB 页面大小对应于专用于数据的 16,384 字节和可用于控制和纠错码 (ECC) 信息的 1,280 字节。

NAND 闪存技术已经从 SLC（Single-Level Cell，1bit per cell）发展到 MLC（Multi-Level Cell，2 bits per cell），再到 TLC（3 bits per cell）和现在的 QLC（4 bits per cell） ）。 SLC NAND 提供更快的写入速度和更长的写入耐久性（大约 30,000 – 50,000 个编程/擦除周期），但更昂贵。 MLC NAND 提供更大的容量，是 SLC 密度的两倍，但耐用性更低（大约 3,000 次编程/擦除周期）。 TLC 和 QLC 显着增加了容量，但代价是耐用性大大降低（可能大约 300 个编程/擦除周期）、性能降低以及需要更多的 DRAM 来映射更高的容量。 DRAM 是典型 SSD 中仅次于 NAND 的最高成本。

ZNS 引入了一种新型 NVMe 驱动器，与传统 SSD 相比具有多项优势。它将一个命名空间划分为多个区域，并且每个区域只允许顺序写入。

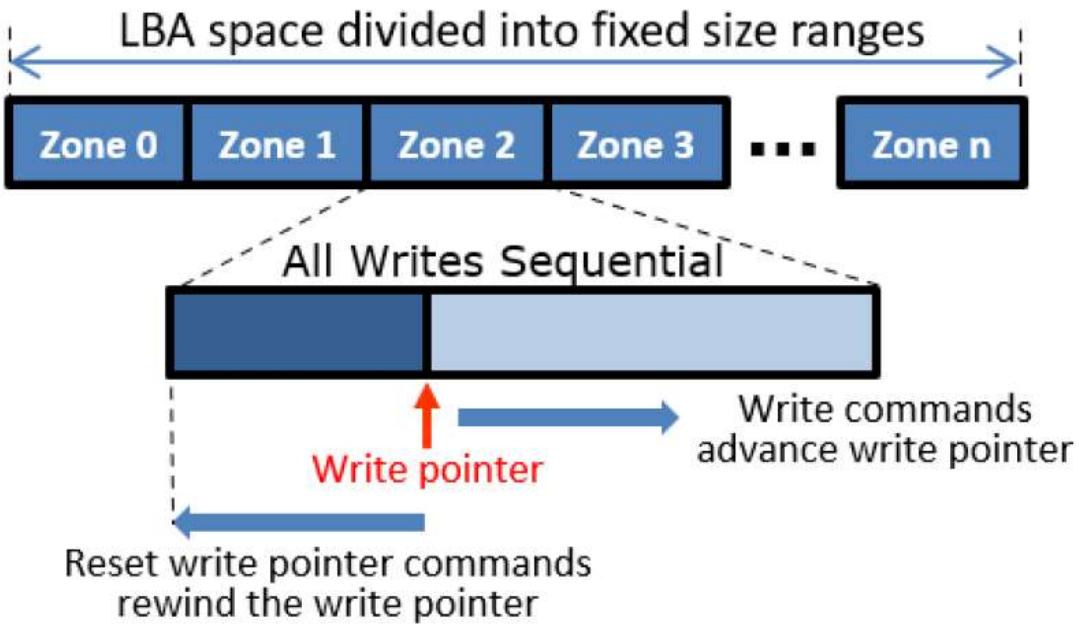


图 3-24 分区命名空间

SSD 协作使用分布式 FTL 进行顺序访问并消除多层间接访问。 不需要复杂的拓扑配置，因为区域是合乎逻辑的。 ZNS 减少了写入放大，改善了内部数据移动，改善了磨损减少，改善了延迟异常值和吞吐量，减少了 SSD 中的 DRAM（较小的 L2P）并减少了对媒体过度配置的需求。 通过将区域与 NAND 闪存的内部物理特性对齐，可以消除数据放置中的一些低效问题。 尤其是俗称的 log-on-log 挑战的问题自然解决了。

获取测试白皮书详情，请登录 https://www.sniffer.com/中文/文档下载/ ，下载 SanBlaze White Papers，然后查看 Verification of ZNS.pdf 文档。

3.6.3 SRIS Clocking Mode 测试

SANBlaze 推出业内首个 NVMe Gen4 测试平台，该平台能够测试 PCIe NVMe 设备所需的各种时钟模式。本文详细介绍了现代 NVMe 设备所需的时钟模式，以及如何验证您的 NVMe 设备是否可以在每种可能的时钟方案下正确运行。

了解 PCIe/NVMe 时钟模式

PCI Express (PCIe) 是从 PCI/PCI-X（外围组件互连）发展而来的，它首先作为个人计算机上的并行互连总线进入市场。 PCIe 是由内置于端点设备（Root Complex 或 Peripherals）中的 SER/DES（串行器/解串器）启用的串行总线。随着 PCIe 的发展，时钟速度以及总线的数据速率和带宽从 Gen1 到 Gen4 增加了六倍。 PCIe 被设计为点对点拓扑，其中每个端点由多个串行通道连接。

SSC 定义

随着 PCIe 总线扩展到主机计算机之外（例如在 NVMe 驱动器机架的情况下），并且时钟频率增加，降低来自互连的电磁干扰 (EMI) 的需求成为当务之急。 PCIe 通过使用“扩频”调制来调制参考时钟来解决 EMI 问题。这种技术称为扩频时钟或 SSC。 SSC 时钟通过在一系列频率上传播辐射能量来降低 EMI 水平，从而降低 PCIe 时钟中心频率的峰值发射。

虽然 SSC 时钟减少了 EMI 干扰，但这样做的代价是将时钟抖动引入 PCIe 子系统。当 SSC 启用时，通常在 PCIe 连接的两端使用公共参考时钟，例如在根联合体和外围设备上。

PCIe 时钟架构

PCIe 支持各种时钟架构，如下所述。 NVMe 设备供应商可能需要测试这些时钟方案中的一个、全部或组合，这给测试工程师带来了重大问题。主机系统通常实现一种时钟方案，而不是多种方案，并且可能支持也可能不支持 SSC 单独或与独立或公共时钟架构相结合。

SANBlaze SBExpress-RM4 支持 PCIe 时钟方案的所有组合，因此提供了一个理想的测试平台来验证您的 NVMe 设备的正确实施和稳定性。

支持以下时钟操作模式。

公共参考时钟

公共参考时钟架构是指将公共时钟提供给上游设备的配置。对于 SBExpress-RM4，这是一个 Gen4 PCIe 桥接器，在这种情况下，端点外围设备是一个 NVMe 驱动器。

没有 SSC 的普通时钟是最基本的时钟方案。它具有最高的性能、最低的延迟，并且最不可能产生错误。两个端点共享一个具有低抖动的稳定时钟。

无 SSC (SRNS) 的独立参考时钟

第二种最常见的时钟方案称为 SRNS（无 SSC 的独立参考时钟），该方案为 PCIe 链路的每一端提供独立时钟。例如，时钟“A”提供给上游桥接器，时钟“B”提供给端点设备，即 NVMe 驱动器。

SRNS 的性能和稳定性应该与普通时钟架构相同，因为即使时钟是独立的，它们也是相同的频率。 PCIe 链路每一端的缓冲将补偿两个独立时钟（每个时钟 +/-300ppm）之间高达 600ppm 的抖动。

具有 SSC (SRIS) 的独立参考时钟

将扩频时钟 (SSC) 技术与独立时钟的使用相结合，给设计和测试工程师带来了最大的挑战。如果 SSC 允许的抖动为 5000 ppm，独立时钟允许的抖动为 600 ppm，则端点之间的总抖动必须为 5600 ppm。

PCIe 协议通过使用 SKP 有序集 (OS) 来补偿两个端点时钟的频率差异。每个设备都必须实施缓冲区来补偿端点时钟频率的差异，同时协议发送 SKP OS 事务层协议 (TLP) 以同步端点，速率与端点频率的差异成正比。

由于需要传输 SKP TLP，设计人员必须在每个端点的缓冲区设计中提供足够的弹性，并且还预计会增加延迟。

SANBlaze 提供了一种方法来启用每个 PCIe 时钟方案，同时提供数据完整性、异常测试和性能监控，以确保 NVMe 设备已正确实施支持的时钟方案。

获取测试白皮书详情，请登录 https://www.saniffer.com/中文/文档下载/ ，下载 SanBlaze White Papers，然后查看 Verification of SRIS_SRNS Clocking.pdf 文档。

3.6.4 TCG Opal 测试

SANBlaze 宣布推出 TCG Opal SSC 验证功能，使客户能够快速有效地测试和验证其 SSD 的 Opal SSC 实施。本白皮书介绍了 TCG Opal SSC 实施并描述了如何使用 SANBlaze 平台验证您的 SSD 是否正确实施了 Opal SSC 规范。

了解 TCG Opal SSC

TCG（Trusted Computing Group）存储工作组（SWG）制定了 Core Specification，正式名称为 TCG Storage Architecture Core Specification，为 TCG 存储设备提供了 TCG 相关功能的全面定义。核心规范可以进一步分解为称为安全子系统类 (SSC) 的多个功能子集。 SSC 明确定义了特定“类”中存储设备的最低可接受核心规范功能，并可能扩展超出核心规范中定义的功能。

Opal SSC 规范以易于实施和集成为前提，也称为“Opal SSC”或“Opal”，是一种存储设备的安全管理协议。 Opal 系列 SSC 包括 Opal、Opalite、Pyrite 和 Ruby，并定义了在存储设备上实现核心规范的功能，例如存储设备上的文件管理。 Opal 系列 SSC 还定义了用于存储和检索文件的类级别权限，从而保护用户数据。符合 Opal SSC 规范的设备可称为 TCG Opal 设备。

Opal 存储规范是一组用于将基于硬件的加密应用于存储设备的安全规范。换句话说，它是一种自加密驱动器 (SED) 规范，驱动器上的所有数据始终加密，无需使用第三方加密解决方案。一些 NVMe SSD 可能会采用纤薄的 Opal，Opalite 或 Pyrite。

为什么要为 SSD 使用 TCG Opal?

固态驱动器 (SSD) 的数据安全性仍然是一个热门问题，因为大多数电子存储设备都是为了将数据直接存储到闪存空间中，而且它们的存储速度更快、效率更高，并且可以在苛刻的环境中使用。然而，SSD 设备并非天生就具有抵御外部数据威胁的能力。一些旨在防止潜在盗窃的软件实用程序可用于主机计算机系统，但由于计算资源的消耗，软件加密程序会减慢对数据的访问时间和对存储设备的总带宽。

随着 SSD 在存储敏感数据方面变得越来越流行，越来越需要强大的数据加密来降低数据被盗的风险。最推荐的方法之一是在驱动器内实施加密。意识到数据安全正迅速成为信息技术行业面临的最紧迫问题之一，大多数 SSD 制造商以 Opal 安全子系统类 (SSC) 的形式将基于硬件的全驱动加密 (FDE) 功能引入固态驱动器中- 加密驱动器 (SED)。

自加密驱动器无需任何用户交互即可自动加密数据，因此在发生数据泄露时对保护敏感和机密信息具有巨大而积极的影响。在本白皮书中，我们将研究 SED 基础知识、SED 和硬件加密的优势以及 Opal 兼容存储设备的功能。

获取测试白皮书详情，请登录 https://www.saniffer.com/中文/文档下载/ ，下载 SanBlaze White Papers，然后查看 Verification of TCG Opal SSC.pdf 文档。

3.6.5 Dual Port NVMe SSD 测试

请参考 Sanblaze 针对 dual port NVMe SSD 在冗余、预留、namespace 命名空间、SGL 工作负载和跨控制器数据比较的双端口 SSD 的测试用例，如下图。

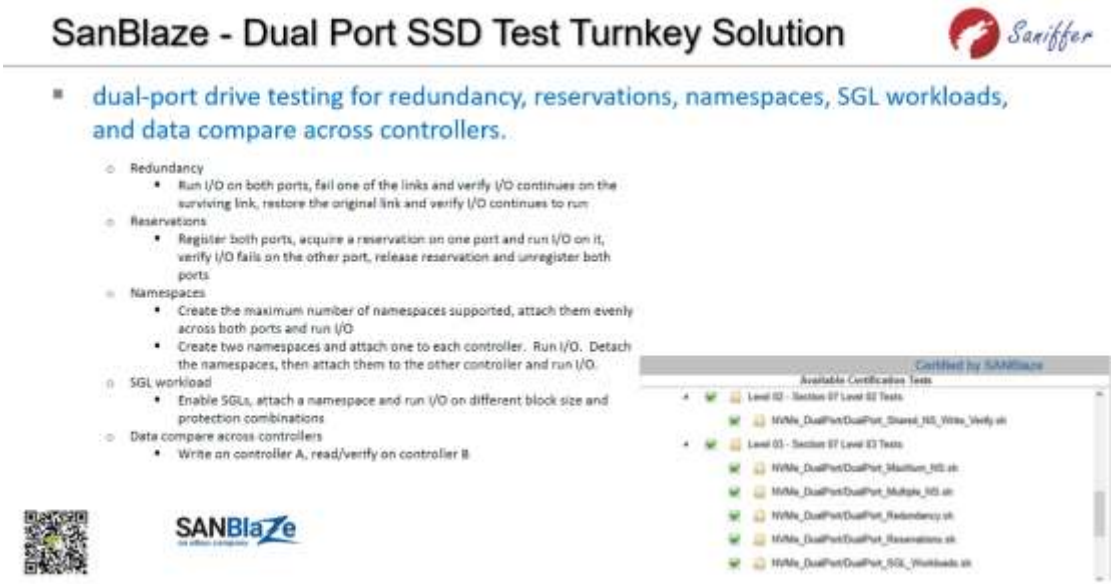


图 3-25

3.6.6 OCP 2.0 Enterprise Data Center NVMe SSD 功能验证测试

关于这部分的技术讲座，请参考下面的链接:

[企业级 NVMe SSD 新特性简介以及测试讲座 - 4.9.2022 by Michael Wang]

https://mp.weixin.qq.com/s/ejQelBklyL-G-Q70b2y63g

Datacenter NVMe SSD Specification v2.0r21 测试用例

- **SanBlaze 将于 2022 年 5 月底发布所有数据中心 NVMe SSD 认证测试用例。**
- **Microsoft 使用 SanBlaze DCSSD 认证工具来认证所有想要部署到 MS Azure 云中的 SSD**
- **DatacenterSSD 规范是由一组超大规模数据中心公司与 SSD 供应商和企业集成商合作创建的。 本规范中有什么内容？ 它如何扩展 NVMe 规范？ 设备如何证明合规性？ 我们将回顾 DatacenterSSD 规范中的重要项目，以了解它如何扩展 NVMe 系列规范以用于数据中心环境中的特定用例。 由于 DatacenterSSD 规范不仅仅是一个接口规范，我们还将展示可用于证明合规性的测试设置。**

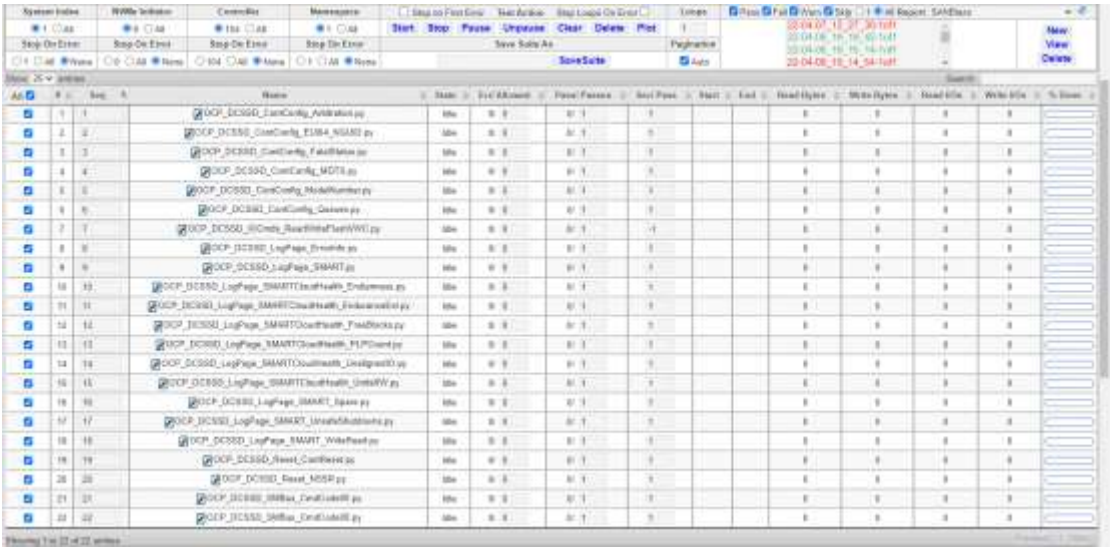


图 3-26

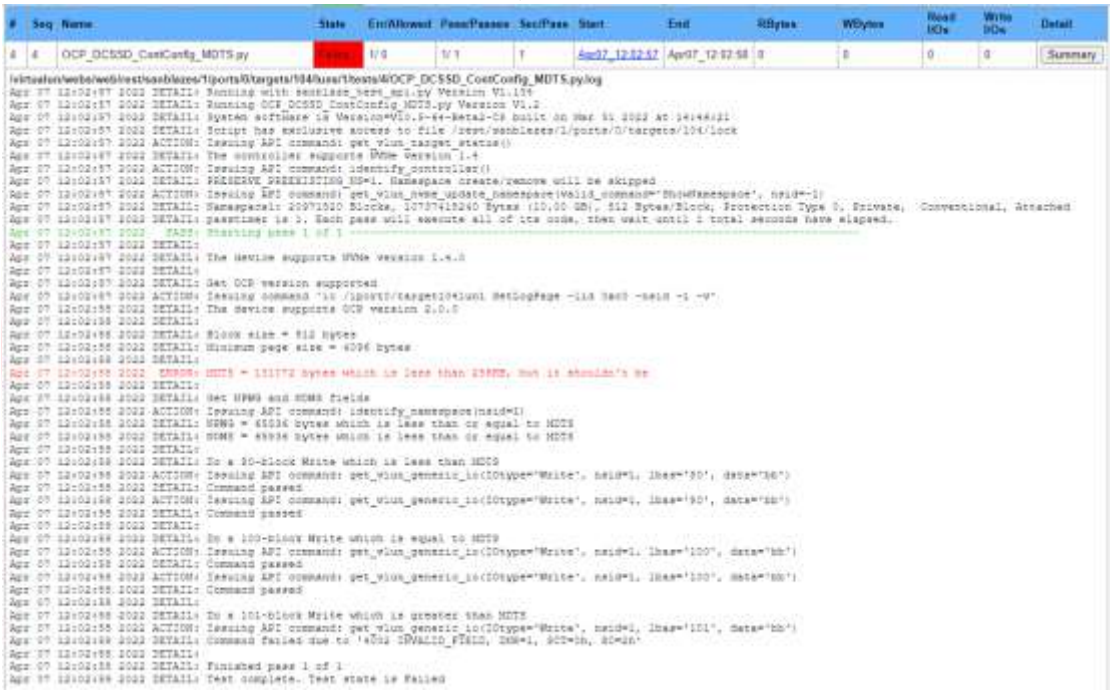


图 3-27

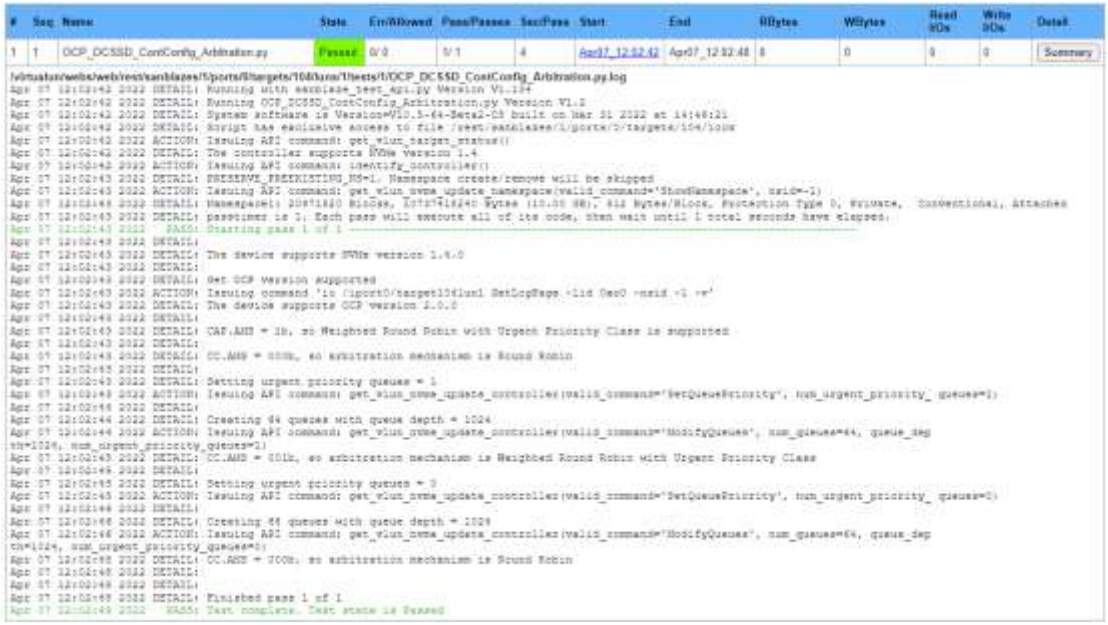


图 3-28

图 3-29

3.6.7 NVMe Power and Reset 测试

PCIe 重置 (PERST) 与电源 ASSERTED 时间相关的时序因平台而异，并且是 NVMe 设备在执行初始代码和固件初始化时潜在问题的根源。

虽然将所有 NVMe 设备暴露给各种硬件服务器供应商和各种 BIOS 供应商的所有可能实现是不切实际的，但可以使用 SANBlaze SBExpress Gen4 硬件上的综合测试来模拟电源和重置时序。

从 SANBlaze CLI 访问的工具 sb_sdb 可以控制电源和 PERST 的断言，并且可以改变彼此相关的两个信号的时序，以模拟复杂的电源和复位时序组合。 sb_sdb 工具在 8.1 Beta 6 或更高版本中可用。

本文档演示了控制电源和复位的语法，并给出了典型时序场景的示例。

来自 CLI 的电源和复位信号的控制

还可以从 CLI 控制电源、重置和时钟，以便在脚本中使用从而创建自动化测试脚本。为了证明 NVMe 设备能够经受住现场可能发生的不太可预测的事件，必须针对可能在现场

发生但在实验室中难以重现的事件对 NVMe 设备进行测试，这一点至关重要。 这些场景包括：

- **意外断电**
- **发生 IO（读取和写入）时意外断电**
- **PCIe 重置 (PERST)**
- **超出规范的 PERST 时序（多个 ASSERTION）**
- **与功率相关的不规范 PERST**

参见下图，一目了然，GUI 显示插槽 3 中的 NVMe 设备位于 PCIe 地址 0e:00.0，采用 U.2 提升板，链接 x4、Gen4 (16G) 速度，并且有电源。

选择移除按钮将触发插槽上的 HotPlug 事件，这将导致设备从系统中正常移除。

选择绿色 Pwr LED 将从插槽中移除电源，从而创建“意外”移除情况，即设备在没有通知的情况下从系统中移除。

图 3-30

注意：

Web GUI 将反映 NVMe 设备的新状态（离线），如下所示：

黄色背景表示插槽 1 中的设备仍然存在，但未链接到 PCIe，这也由现在报告为 x0 的通道宽度指示。

绿色 Pwr LED 按钮提供了仅使用软件从插槽中断开电源的最简单方法。

下面截图是 SanBlaze 提供的超过 30 种各种各样的 Reset 测试脚本。

图 3-31

获取测试白皮书详情，请登录 https://www.saniffer.com/中文/文档下载/ ，下载 SanBlaze White Papers，然后查看 NVMe Power and Reset Testing with the SBExpress-RM4.pdf 文档。

3.6.8 CMB_HMB 测试

本白皮书通过几个示例概述了使用 SANBlaze 系统的 CMB/HMB、状态检查、设置和配置。

CMB 简介

CMB（控制器内存缓冲区）是控制器上的通用读/写内存区域。在 PCIe 初始化期间，控制器将请求它需要的一定数量的内存。主机内存管理器将通过写入控制器上的寄存器来分配内存大小和基地址。所有 NVMe 控制器都有自己的 BAR0 和 BAR1，供主机写入控制器内存的起始位置。每个 NVMe 控制器在 BAR0/BAR1 中都有不同的基地址。控制器内存地址高于系统通常使用的任何地址。

下面的图显示了主机可用的 64 位内存地址范围。

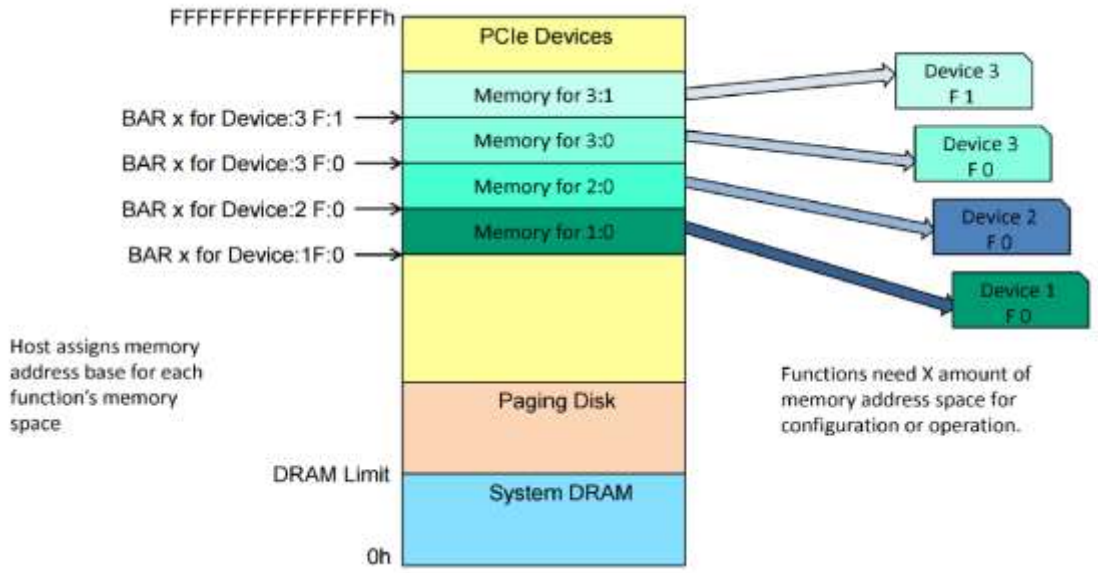


图 3-32 PCIe 内存位置

系统物理内存限制标记为“DRAM 限制”，在系统 DRAM 之上是系统分页内存，然后是可用于不同 PCIe 设备/功能的系统内存。 CMB 物理上位于设备上，但像主机内存一样寻址和处理。这样，通过添加更多功能，每个功能提供它需要的内存而不占用任何系统内存。

在 NVMe 1.2 之前，读/写命令、SQ、CQ、PRP 和 SGL 列表的数据和元数据存储在主机物理内存中。从 NVMe 1.2 开始，可选的 CMB 功能允许主机在控制器中定义一个区域缓冲区来保存上面提到的部分或全部数据。如果主机将此数据放在设备上的缓冲区中，则设备不必使用 PCIe 设施来获取数据或将其发送到主机。存储设施的寻址是主机地址范围的一部分，由设备配置头空间中的 BAR (BAR0-BAR5) 标识。

CMB 相关寄存器和设置

在最新的 NVMe 规范（“NVMe Base 1.4_NEXT 2020.10.12a.docx”）中定义了 6 个与 CMB 相关的 NVMe 控制器寄存器，如下所示。支持 NVMe v1.4（“2019.06.10-Ratified”）的 NVMe 驱动器固件仅定义了下面的前 4 个 CMB 相关寄存器。 NVMe v1.3 或更早版本仅定义了以下前 2 个 CMB 相关寄存器。

- **1. CMBLOC——控制器内存缓冲区位置，偏移 38h**
- **2. CMBSZ – 控制器内存缓冲区大小，偏移 3Ch**
- **3. CMBMSC——控制器内存缓冲区内存空间控制，偏移 50h**
- **4. CMBSTS – 控制器内存缓冲区状态，偏移 58h**
- **5. CMBEBS – 控制器内存缓冲区弹性缓冲区大小，偏移 5Ch**
- **6. CMBSWTP——控制器内存缓冲区持续写入吞吐量，偏移 60h**

在 NVMe v1.4 中，NVMe 控制器 CAP（控制器功能）寄存器具有用于 CMBS（支持控制器内存缓冲区）的 RO（只读）位 57。如果将 CAP.CMBS 设置为“1”，则控制器支

持 CMB。主机通过将 CMBMSC.CRE 设置为“1”来指示使用 CMB 的意图。一旦该位设置为“1”，控制器将通过 CMBLOC 和 CMBSZ 寄存器指示 CMB 的属性。在 NVMe 1.3 或更早版本中，没有定义 CAP.CMBS 位。 SANBlaze 驱动程序可以支持不同的 NVMe 版本，它也基于 SSD 支持的 NVMe 版本支持 CMB。例如，如果 SSD 支持 NVMe 1.4，则将检查 CAP.CMBS，但对于支持 NVMe 版本低于 1.4 的 SSD 驱动器，则不会检查 CAP.CMBS。

获取测试白皮书详情，请登录 https://www.saniffer.com/中文/文档下载/ ，下载 SanBlaze White Papers，然后查看 CMB_HMB_White_Paper-V3.pdf 文档。

3.6.9 T10 DIF/DIX 测试

T10 DIF/DIX 介绍

DIF（数据完整性字段）和 DIX（数据完整性扩展）是保护计算机数据存储中的数据完整性免受数据损坏的方法。它是由国际信息技术标准委员会 T10 小组委员会于 2003 年提出的。 2012 年引入了这项技术的演变，称为 PI（保护信息）。 2016 年，NVMe 1.2.1 规范中添加了类似的数据完整性方法。

DIF 和 DIX 之间的主要区别在于保护信息的位置。在 DIF 中，PI（保护信息）与逻辑块数据相邻并创建一个扩展的逻辑块，而在 DIX 中，PI（保护信息）存储在单独的缓冲区中。本规范定义的端到端数据保护机制在功能上兼容 DIF 和 DIX。 DIF 功能是通过将元数据配置为与逻辑块数据连续来实现的，如下所示：

图 3-33 DIF 元数据 – 与 LBA 数据相邻，形成扩展的 LBA

DIX 功能是通过将元数据和数据配置在单独的缓冲区中来实现的，如下所示：

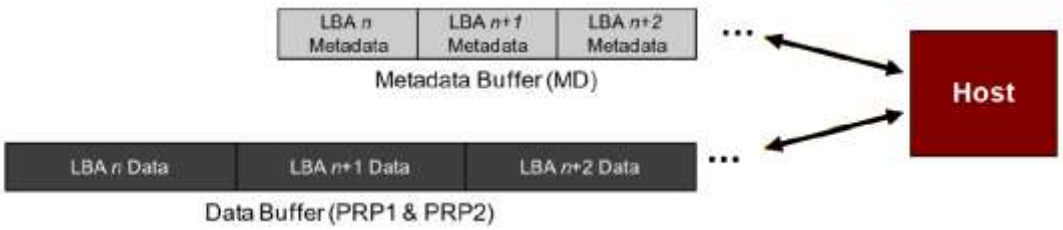


图 3-34 DIX 元数据 – 作为单独的缓冲区传输

本白皮书通过几个示例概述了使用 SANBlaze 系统的 PI 字段、PI 类型、PI 格式、PI 状态检查和 PI 错误注入。

3.6.10 FDP 功能测试

通过 SANBlaze 测试套件认证有助于开发、验证和调试 FDP 技术

全球领先的先进存储测试和验证技术提供商 SANBlaze Technology Inc.宣布推出 Flexible Data Placement (FDP)测试套件，允许 SSD 制造商在以下设备上测试 FDP：他们的驱动时间是使用自研或传统工具的一半。

“Flexible Data Placement (FDP) 是一种创新的下一代 SSD 技术，可提高性能、提高服务质量并减少 SSD 磨损，从而使 SSD 使用寿命更长。使用测试套件来开发、验证和调试该技术是重要的一部分生态系统，以实现高质量产品的部署。”

Ross Stenfort，Meta 硬件系统工程师

“测试 FDP 的能力可帮助组织优化其数据存储策略，确保每个数据集根据其独特特征存储在最合适的位置”。“测试 FDP 的能力可以帮助组织优化其数据存储策略，确保每个数据集根据其独特的特征存储在最合适的位置，”销售和营销副总裁 Rick Walsh 说。“SANBlaze 的一流工程团队一直在熟练地将 FDP 验证和测试编码到我们的 *Certified by SANBlaze* 软件中，我们很自豪于 2023 年 8 月 8 日至 10 日在 FMS 613 号展位上宣布这一消息。”

*由 SANBlaze 认证的*测试 套件在 SANBlaze 硬件上运行，包括 SBExpress-RM5 PCIe 5.0 NVMe 机架式测试系统和 SBExpress-DT5 PCIe 5.0 NVMe 桌面测试系统。

硬件包括高级功能，与 SerialTek 协议分析仪集成

第六代 SBExpress-RM5 和 SBExpress-DT5 都是进化型产品，源自广泛部署的前代产品的成功系列，并且具有创新性，具有先进的测试功能，包括灵活数据放置 (FDP)、针对 NVMe 的供应商定义消息传递 (VDM) 测试支持- PCIe 上的 MI，以及带内和 SMBus 测试功能。PCIe 5.0 速度完全支持 *SANBlaze 企业测试套件认证*的所有功能。该系统与 SerialTek 的 Kodiak™ PCIe 5.0 协议分析系统无缝集成，为用户提供全面的测试、调试和分析功能。

Riser 技术

SBExpress-RM5（机架式）16 盘测试系统和 SBExpress-DT5（台式）测试系统均采用 SANBlaze 专有的 Riser 转接卡技术，并方便地使用同一组 NVMe PCIe 5.0 转接卡。所有 SANBlaze 转接卡都能够进行单端口和双端口操作，并且可以在软件控制下进行动态切换。转接卡适用于 PCIe 5.0 速度的所有 NVMe 外形规格。

特性和功能

SBExpress-RM5 和 SBExpress-DT5 支持以下特性和功能：

- U.2/E3 Single/Dual Port
- All EDSFF form factors
- *Certified by SANBlaze* test suite, with automated report generation and performance plotting
- Flexible Data Placement (FDP)
- L1.1 and L1.2 Low Power State testing with CLKREQ monitor
- Vendor Defined Message (VDM) testing
- SR-IOV and multi-root testing
- SMBus testing up to 1MHz
- MI/MCTP testing over SMBus, VDM, and in-band transports
- Power On/Off testing under SW control, all devices
- Clock disable testing to all devices
- Hardware PERST testing to all devices
- Single-/dual-port testing on-the-fly under software control
- SRIS/SRNS and SSC advanced clock mode testing (see our whitepaper)
- Power monitoring of each device under test
- Full PCIe 5.0 bandwidth (PCIe 5.0 x4) to each riser and PCIe 5.0 x16 to root complex
- Remote system control for power up/down/reset
- PCIe 5.0 x16 "Top Slot" for PCIe Add-In Card (AIC) testing (e.g., FPGA or PCIe Analyzer)
- Python APIs for full system control and integration into corporate test infrastructure

3.6.11 SR-IOV 功能测试

SR-IOV 基本概念介绍

SR-IOV （single root I/O virtualization）规范可以用来在虚拟环境中共享单个物理 PCI Express 总线设备。SR-IOV 为物理服务器机器上的不同虚拟组件（例如网络适配器，

NVMe SSD 等）提供不同的虚拟功能。 SR-IOV 使用物理和虚拟功能来控制或配置 PCIe 设备。 物理功能能够将数据移入和移出设备，而虚拟功能是轻量级的 PCIe 功能，支持数据流动但也具有一组受限的配置资源。 虚拟机管理程序（VMM – Virtual Machine Manager）或客户操作系统 （Guest OS）可用的虚拟或物理功能取决于 PCIe 设备。

SR-IOV 允许虚拟环境中的不同虚拟机 (VM) 共享单个 PCI Express 硬件接口。 相比之下，MR-IOV 允许 I/O PCI Express 在不同物理机上的不同 VM 之间共享资源。

下图是一个 SR-IOV 的使用环境的框图，左边是安装有多个虚拟机的物理机，右边是插在某个 PCIe slot 里面的的支持 SR-IOV 的网卡，其中网卡上面 2 个网口为两个 function。

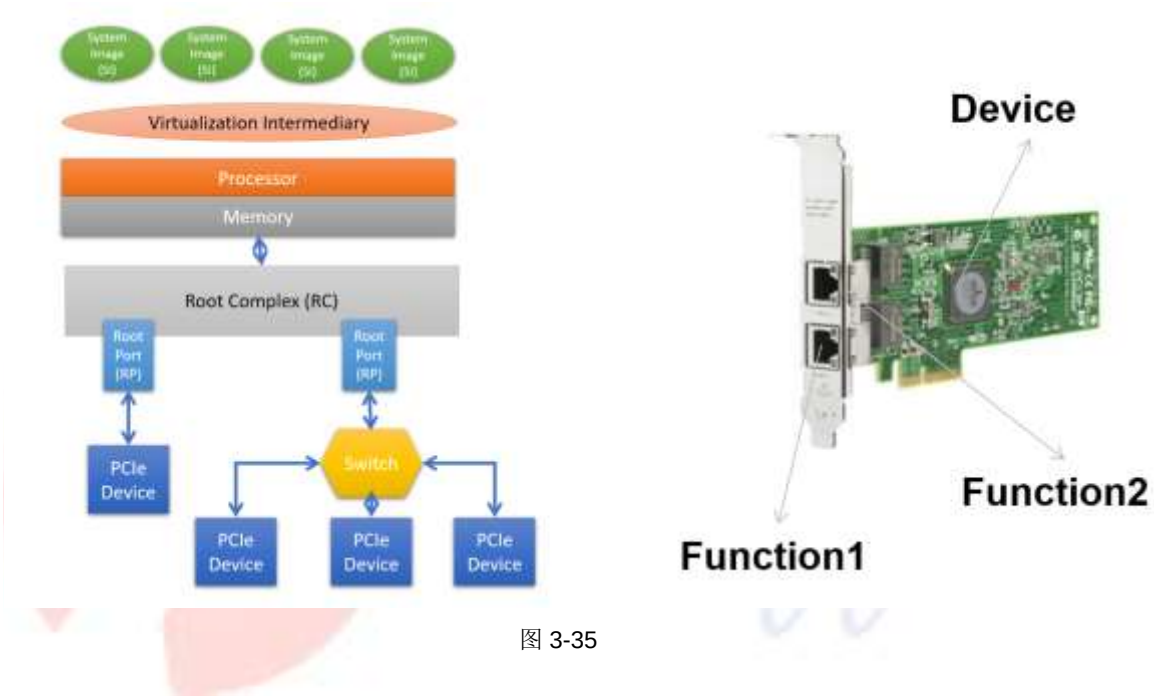


图 3-35

熟悉 Linun 操作的话，很容易通过 lspci -v 命令下面获得该张网卡是否支持 SR-IOV，参见下图。

图 3-36

SR-IOV 工作原理（以为 NVMe SSD 插卡为例）

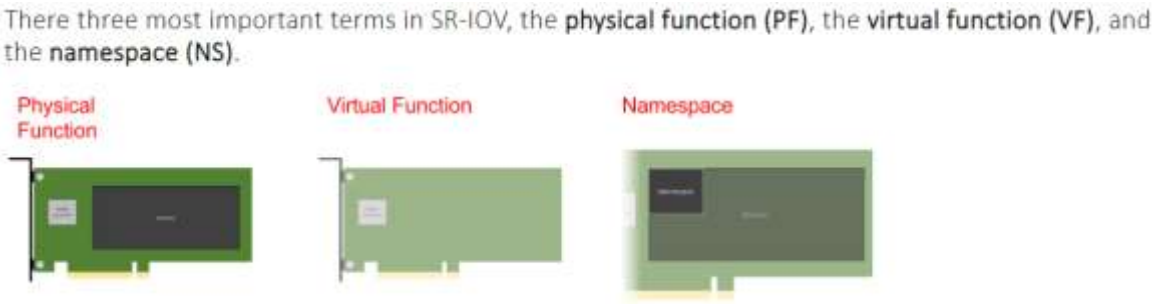


图 3-37

上图中，PF 指的是物理设备，或 NVMe SSD 本身。 SR-IOV 中的所有内容（例如创建 VF 和命名空间）都是从 PF 开始的。

一个 VF 是从一个 PF 扩展而来的，多个 VF 共享设备的底层硬件和 PCI Express 链路。

Namespace 命名空间是主机软件可寻址的逻辑块的隔离。

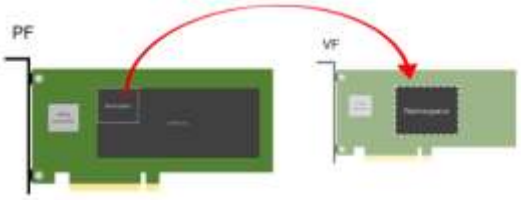


图 3-38

用户可以在创建命名空间时分配命名空间的共享属性。 namespace 是实际的存储资源，但是需要一个 VF 来处理虚拟机和 namespace 之间的 I/O。 将命名空间附加到 VF，以便 VM 可以访问存储容量。 在实践中，命名空间附加到 VF，然后将 VF 提供给主机。

SR-IOV 的架构

我们先来看一下 SR-IOV 部署的环境中虚拟化相关的概念。

- **虚拟化中间层**

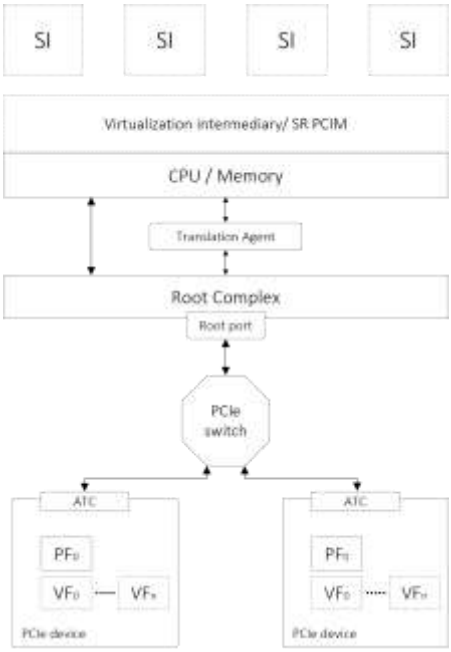


图 3-39

虚拟化中间层 (VI - Virtual Intermediary) 是管理虚拟机和物理计算/存储资源之间所有交互的软件层，通常是管理程序或 VMM。 在经典 VM 应用程序中，VI 处理所有中断，映射哪些数据应该发送到哪个虚拟机。 结果，系统的 IO 性能下降，因为它消耗大量 CPU 资源来处理所有中断，并且随着 VM 数量的增加而变得更糟。

* 系统映像 (SI)，例如客户操作系统，可以为其分配虚拟和物理设备

** 地址转换缓存 (ATC - Address Translation Cache)

- **绕过管理程序或 VMM 以提高性能**

使用 SR-IOV 设备和不使用该功能的 NVMe SSD 在虚拟化话环境中经过的环节参见下图。

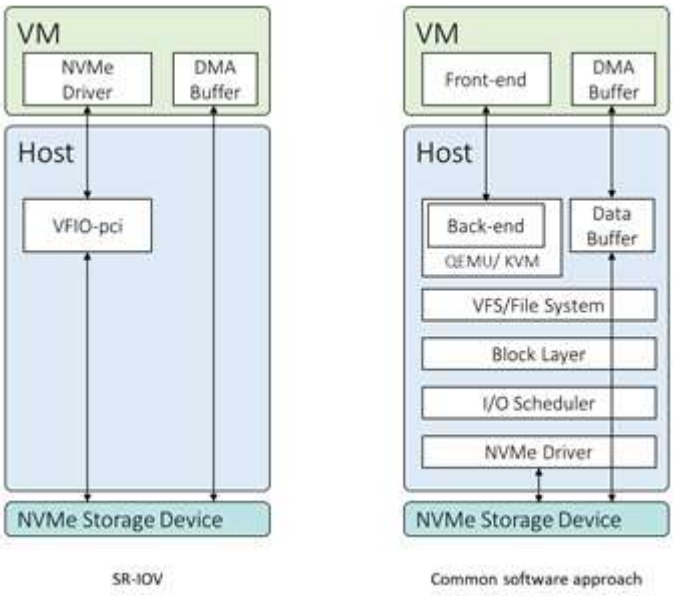


图 3-40

解决此性能问题的直接方法是降低 CPU 开销。也就是说，让其他人处理 IO 中断，而不是 VI，让 CPU 资源用于更有意义的任务。这就是 SR-IOV 发挥作用的地方。当 NVMe SSD VF 成功配置到虚拟机时，它会在虚拟机和物理设备之间引入直接 IO 路径，更准确地说，是与附加到 VF 的命名空间相关联的物理资源。此直接 IO 路径绕过管理程序，允许 VM 自行处理 IO 中断。

现在虚拟机可以直接访问 NVMe SSD，就像物理服务器访问安装在其上的物理 SSD 设备一样，虚拟系统可以利用物理 SSD 提供的所有带宽，而不受其间任何软件层的影响，因此性能接近裸机。

虚拟化看起来很棒，SSD 可以充分利用，如果数据不必传输到任何其他计算设备，性能不会受到网络延迟的影响。这种方法背后的问题是 NVMe SSD 仍然不灵活，只能在安装它的服务器内访问，并且当您需要创建许多虚拟机以匹配 SSD 性能时，CPU 要求变得非常高。

- **多主机 NVMe SR-IOV（也称为 MR-IOV / Multi Root I/O virtualization)**

通过支持 MR-IOV 的 PCIe switch 的支持可以实现将 SSD 与服务器分离，将其虚拟功能提供给不同的计算机，即多个根(multi-roo)，参见下图。

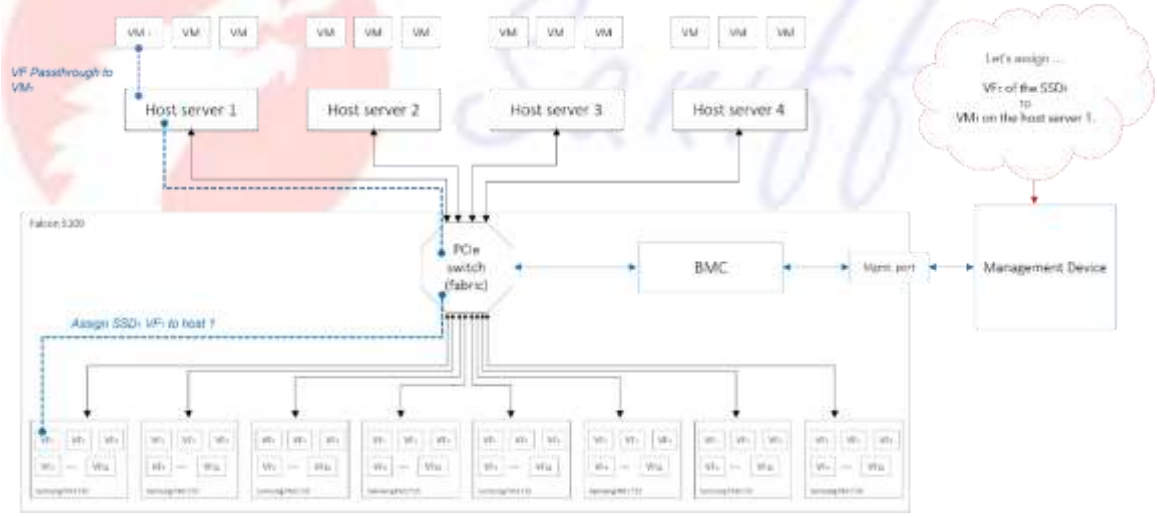


图 3-41

目前业内支持 SR-IOV 的 NVMe SSD 不是特别多，这些 SSD 里面通过 VMWare 兼容性认证的产品目前全球还仅有 Huawei 一家的产品（查询时间：2022/5），参见下图。

图 3-42

图 3-43

目前，云数据中心越来越倾向于部署支持SR-IOV的部件，从传统的网卡需要支持SR-IOV，现在延伸到 NVMe SSD 也需要支持 SR-IOV。

例如，拥有超过 400 万台主机的微软公司在 2021/9 提出需求需要对于支持 SR-IOV 的 NVMe SSD 进行兼容性和协议等功能测试，委托 SanBlaze 开发了大量相应的测试脚本。参加下图。

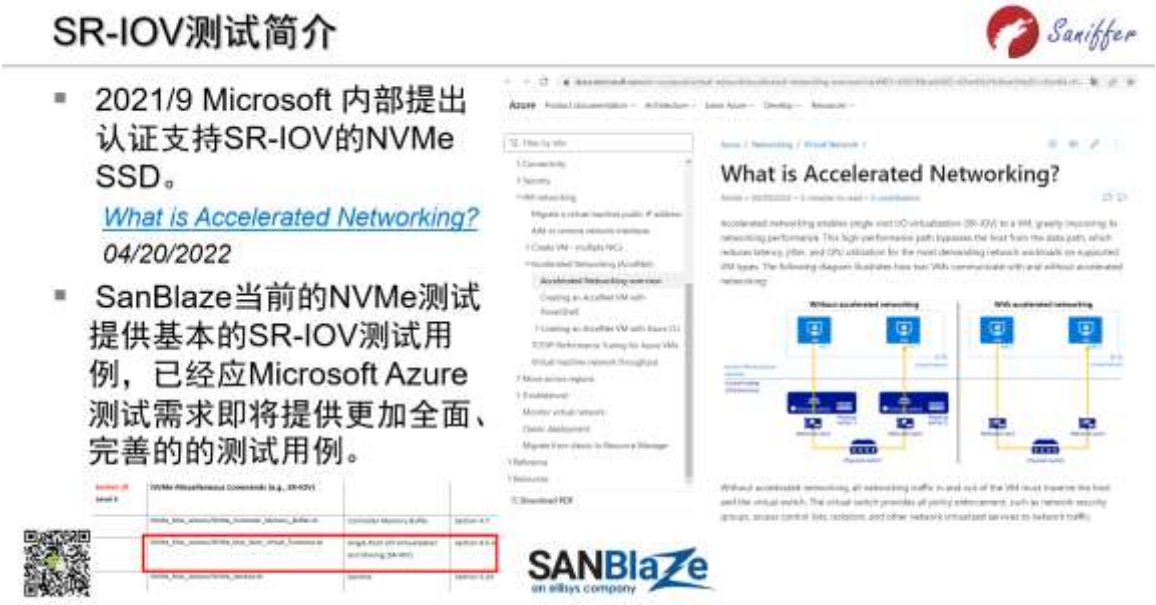


图 3-44

3.6.12 NVMe 功能验证测试

在测试 NVMe SSD 时，一些客户喜欢执行复杂的测试，调整多个变量以创建一个可以在现场复制的各种现实生活场景。对于这些客户来说，重要的是他们能够控制单个选项，以完成复杂的测试来验证他们的 SSD。反过来，其他客户希望运行基本的 NVMe 测试，以确保他们符合各种规范。对于这些客户，使用 SANBlaze 的预制脚本大大简化了测试过程。

SANBlaze 为 NVMe 测试提供了一个完整的系统，包括完整的 Gen 4 和 Gen 5 NVMe SSD 硬件平台。

无论您是要执行复杂的测试还是需要最少调整参数和选项的测试，它都是有利于学习和了解使用 SANBlaze 时“幕后”的内容。知道背后发生了什么简化脚本的场景可以帮助您实现测试的潜力，并可能为您节省许多质量时间进行中。

本白皮书以 NVM 比较为例，带您完成 SANBlaze 验证过程。

了解此过程将使您能够以相同的方式使用任何 SANBlaze NVMe 命令 - 确定在您的驱动器中支持命令，运行命令，检查输出，决定是否需要重新运行，并查看日志文件中的输出。了解“幕后”发生的事情有助于了解 SANBlaze 以及我们如何帮助您以最快、最准确的方式测试您的 NVMe 驱动器。

3.6.13 从 SanBlaze 触发 SerialTek PCIe Gen4/5 协议分析仪进行问题分析

SANBlaze SBExpress 和 Certified by SANBlaze 测试方法提供了一种创建复杂测试套件、验证规范合规性、数据完整性、电源和重置测试以及 MI 合规性的简单方法，大大简化了 PCIe NVMe 设备的认证。

借助 SANBlaze 内置事务 trace、包括链路训练和状态机 (LTSSM) 在内的低级错误计数器、使设备经受电压变化和复杂复位、链路训练和数据完整性测试并记录其行为的能力，人们可以问，“为什么我需要 PCIe 分析仪？”。

最重要的是，当 PCIe 总线上发生意外情况时，例如对一个永远不会完成的设备的事务，没有比 CPU Root Complex 和端点设备之间的 PCIe 协议分析仪更好的工具来确定发生故障的根本原因。问题。

大海捞针

由于高达 8GB/s 的数据在 PCIe Root Complex 和您的被测设备之间来回传输，找到 SBExpress 错误的来源并将其与协议分析仪中的数据进行协调就像查找事务层数据包 (TLP) 一样协议分析仪中众所周知的大海捞针的大小。

为了进一步增加将 SBExpress 系统标记的错误与分析仪中的数据协调起来的难度，您正在寻找的数据实际上可能不再存在于分析仪中。以具有 64GB 数据的相当大的缓冲区存储容量和 4GB/秒的典型 PCIe 顺序读取速率的分析仪为例，您的分析仪最多可以捕获 16 秒的数据，直到缓冲区开始包装和大海捞针针已被装载到平板上，当您意识到它们不见了时，已经到了巴黎的一半。

你需要一个触发器

你需要的是一个触发器！一种在错误发生时停止分析仪的方法，以便错误本身和导致错误的一些事件（包括错误原因）以及错误发生后的一些事件显示您的设备如何处理错误。

但是触发什么

您打电话给固件团队中最聪明的工程师，并向她描述问题。 “我的 SANBlaze 系统在测试我们的最新固件时发生错误，SANBlaze 错误日志显示错误。您能否在我的分析仪上设置触发器以查找在 12 小时测试中发生一次的事件？触发器需要与未知事件协调发生，且需要在事件发生后 16 秒内发生”。

当她笑完后，您的工程师会告诉您您的要求几乎是不可能的，您正在大海捞针。

你需要一块巨大的磁铁

用众所周知的“干草堆中的针”来比喻，你需要的是一块巨大的磁铁，它将从海量数据干草堆中拉出细小的针，这样你就可以更仔细地观察它。您需要 SANBlaze SBExpress 触发器。

介绍 SANBlaze SBExpress 触发器

SANBlaze 团队感受到了您的痛苦，事实上，我们的实验室和客户也面临同样的问题。有问题，客户上传系统日志到 JIRA 系统。我们查看日志并得出结论，该设备在 PCIe 总线上行为不当，请向我们发送 trace 信息。

触发什么以便捕获的 trace 与 SANBlaze 系统协调的问题是一个难题。在看到错误周围的 trace 之前，您不会知道如何设置协议分析仪来捕获它，但是您需要捕获 trace 以查看需要触发的内容。经典的 Catch22 情况。

输入 SANBlaze SBExpress 触发器。我们在 SanBlaze 软件 V8.2 软件中内置了一个触发机制，以便我们在确定发生故障的确切点触发您的协议分析仪，因此现在当我们说“向我们发送 trace”时，您只需将协议分析仪的 trace 文件上传到 SANBlaze 触发器。

本文的其余部分将介绍如何在您的 SBExpress 系统和分析仪上设置 SANBlaze 触发器。

关于 SerialTek PCIe Gen4 分析仪的一句话

有许多高质量、最先进的工具可用于捕获和分析 PCIe Gen4（现在是 Gen5） trace。我们选择使用 SerialTek 的 Kodiak Gen4 PCIe 分析仪来配合 SBExpress 触发器。

下面的示例将直接转换为您的 PCIe Gen4 或 Gen5 协议分析仪，并且会以类似的方式运行。请联系您的 PCIe 协议分析仪供应商或 SANBlaze 以获取有关配置 SANBlaze 触发器的帮助。

当坏事发生时

当 SBExpress 系统确定在测试 NVMe 设备时发生了意外情况时，将生成特殊模式并通过 PCIe 发送。正确配置的协议分析仪将检测特殊的 SANBlaze 触发器并捕获 trace 以供进一步分析。

除了固定的触发模式之外，SANBlaze Trigger 还将包括 16 位的诊断信息，SANBlaze 支持团队可以使用这些信息直接在我们的代码中生成触发器的位置，从而大大简化您捕获 trace 的工作和我们的工作确定我们检测到错误的原因。

获取测试白皮书详情，请登录 https://www.saniffer.com/中文/文档下载/ ，下载 SanBlaze White Papers，然后查看 Triggering a PCIe Analyzer from your SBExpress System.pdf 文档。

3.7 SanBlaze Gen5 测试设备 DT5/RM5 产品单页

SANBlaZe
an ellisys company

SBExpress-DT5 Gen5
Data Sheet

SANBlaze SBExpress-DT5 Gen5 NVMe SSD Test System

Highlights

- Client/Enterprise class NVMe qualification in a desktop enclosure
- 100% compatible with SBExpress-RM5 for "Prototype to Production" scaling
- Plug and Play, testing in under five minutes
- Quiet enough for home, capable enough for the Enterprise

Overview

The SANBlaze SBExpress-DT5 (Desktop, Gen5) is a complete turnkey PCIe® Gen5 NVMe SSD validation test system. With industry leading *Certified by SANBlaze* automated testing, the SBExpress-DT5 brings Enterprise Class NVMe validation to the developer's desktop.

The DT5 is 100% Software Compatible with the SBExpress-RM5 sixteen device system. The DT5 provides features — formerly the exclusive domain of the qualification and incoming inspection teams — right to the engineer's desk.

Using the same Single/Dual Port U.2, M.2, EDSFF, and Riser5 risers as the 16-bay rack mount system, the desk top DT5 packs the same enterprise class features including including OCP, TCG, SRIS, L1.1/L1.2 substates, ZNS and power control. All attributes available through the SANBlaze automated environment support Python, REST and CLI/XML API control.

The "out-of-the-box" *Certified by SANBlaze* test suite, or tests highly customized by your qualification team or a customer's inbound inspection team, can be run on the engineer's desk providing a huge time-to-market advantage in finding issues before they leave your workstation.

Quiet enough for remote locations and in an enclosure designed for portability using the optional high impact travel case, the DT5 is perfect for your "work at home" environment or to take directly to your customer.

A complete system, fully self-contained, the DT5 can be set up, configured, and testing NVMe devices in under five minutes.

System Features

Six PCIe Gen4 slots

- Slots 0, 1, 2 Gen5 Riser; options for U.2 / U.3 /M.2/EDSFF (short)
- Slot 3, PCIe x16

Hardware Features under SW control

- Dual/Single Port selectable
- Fan speed control
- Power up/down testing each device
- PERST (reset) and HotPlug on each device
- Power measuring all slots
- Voltage margining +/- 15%
- SRIS support and built in tests (optional)
- Surprise/Graceful removal testing
- VDM (optional)
- SMBus and In-Band MI testing
- FW download via VDM, SMBus and In-Band
- Gen5 TLP/DLLP/Receive Error monitoring for PCIe Gen5 validation
- L0's, L1, L1.1, L1.2 substate support

Dimensions

Length: 17" | **Width:** 13.5" | **Height**: 5.5"

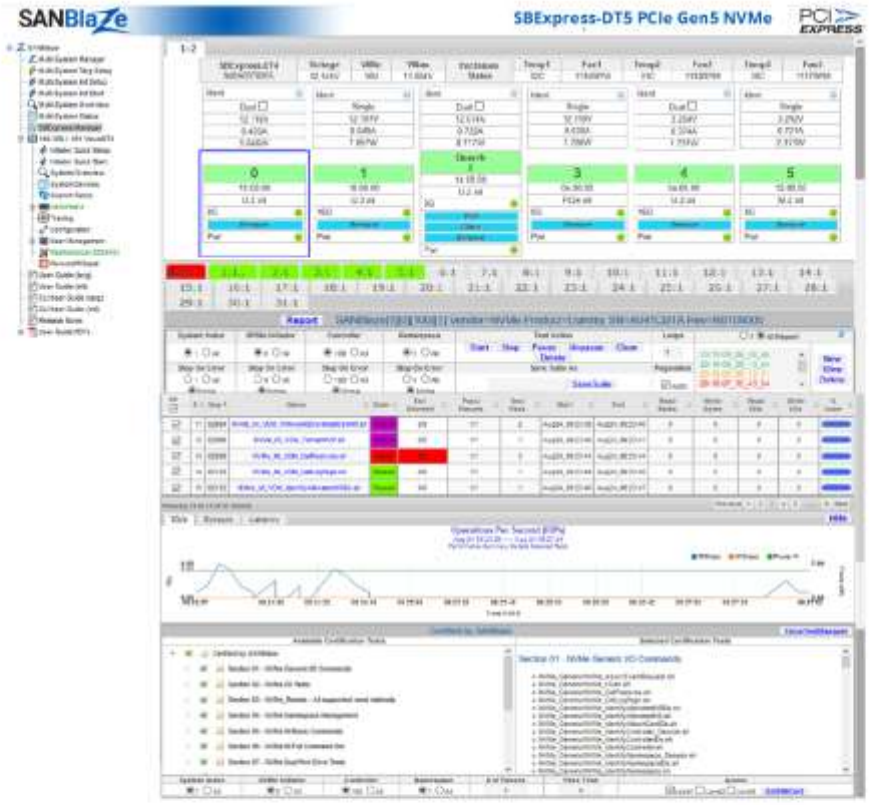
Power

AC Input 100-127 Vac/200-240 Vac
Max 7A at 100V

Certified by SANBlaze

Using the SBExpress Manager software on the SBExpress-DT5 hardware, SANBlaze enables you to test your NVMe drives for compliance and performance. The ability to test your drives from prototype to production provides your end-customers the assurance of 3rd party testing validation - a must-have as NVMe takes its place as the leader in enterprise storage.

Contact us today for a demo of OCP, TCG Opal, VDM, SRIS and other major features of the SANBlaze-DT5 NVMe SSD Test System.



SANBlaze SBExpress Manager with Certified by SANBlaze Automated Test Suite

Sniffer

SANBlaze
an ellisys company

SBExpress-RM5 Gen 5
Data sheet

SANBlaze SBExpress-RM5 Dual Port NVMe Drive Test System

Overview

The SANBlaze SBExpress-RM5 is a complete turnkey PCIe® Gen 5 NVMe SSD validation test system. The SBExpress-RM5 feature set provides unique functions applicable to all aspects of a product lifecycle - from engineering development, through validation and QA, to manufacturing test environments.

The ability to drive NVMe SSDs with a wide range of configurable attributes provides engineers with a flexible, multi-controller supported validation test platform.

Development, qualification, and certification test cycles can be highly automated, thus reducing overall test time, and rapidly surfacing errors and non-conformance.

The SANBlaze SBExpress-RM5 hardware provides a rackmount chasis with sixteen dual or single port front-accessible drives. The hardware provides excellent air flow for thermal chamber testing.

Dimensions

Height: 3.5"
Width w/out Ears: 17 1/8"
Width w/ Ears: 19"

Depth w/out handles: 21 3/4
Depth with handles: 22 3/4

Thermal Characteristics

Temperature Rating	-5C to 70C
Humidity	95% RH Non-condensing

Power

AC input 100-127 Vac/200-240 Vac
Max 14A at 100V

System Features

- PCIe® Gen 5
- Multiple clocking architecture support (SRIS/SRNS)
- VDM
- Gen5 PCIe NVMe testing on a Gen4 or Gen5 host
- Sixteen drives, all front accessible
- Hot plug, slot power, and drive presence under software control
- U.2 single/dual port risers (switchable)
- U.3 single/dual port risers (switchable)
- EDSFF
- M.2 Adapter support
- Field-installable riser cards support both single or dual port drives
- Measure voltage and power at each drive
- Optimal thermal design with unobstructed air flow
- Six fans speed-controlled through temperature sensing on each fan (user controllable through config file)
- LFM for 15mm drive is 1,373
- 38 temperature monitoring points
- 25W per slot at 70C
- Voltage margining +/- 15%

Software Features

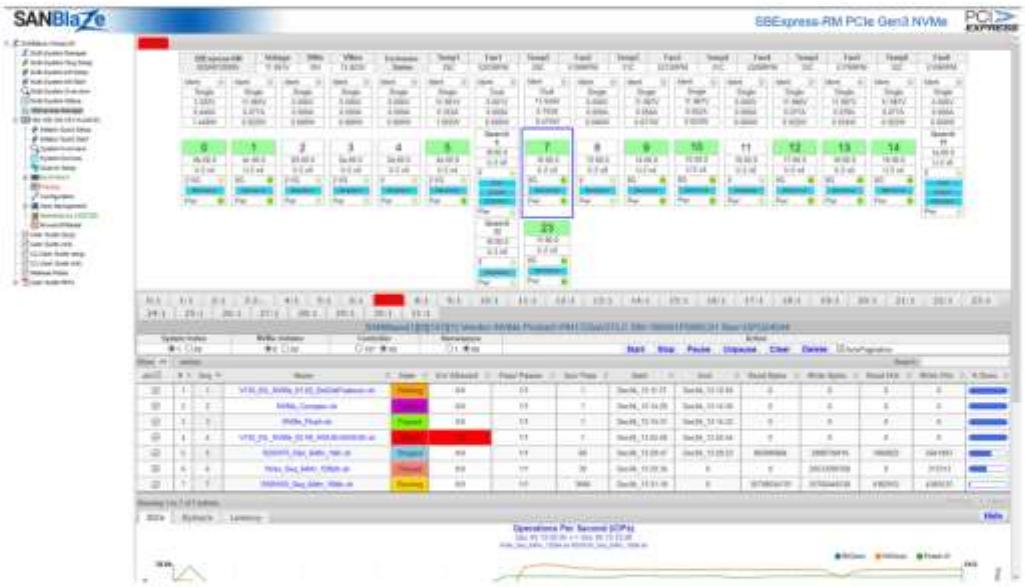
- Test coverage for all aspects of the NVMe specification
- UNH Conformance Testing supported
- NVMe-MI (Management Interface) testing over SMBus supported
- SGL, SR-IOV, full namespace control and reservations
- Drive multiple ports of traffic simultaneously
- Send specific or custom op codes in an easy to use scriptable format
- Read / write / compare testing
- Error injection
- Vendor-unique commands supported
- Drive and test single or multiple NVMe target devices
- SBCert (Certified by SANBlaze) test suite
- ZNS

Sniffer

SANBlaZe an ellisys company

SBExpress-RM5 Gen 5 Data sheet

SANBlaze VLF Gen5 Server and SBExpress-RM5



SANBlaze SBExpress Manager with Certified by SANBlaze tests

Certified by SANBlaze

- Over 900 out-of-the-box tests
- Enables IOL testing in the customer's lab, before undergoing official testing
- Widely recognized industry benchmark
- ZNS, VDM, and TCG Opal verification is included

All information in this document is subject to change. Configurations rely on host system BIOS support and may be restricted by the host. Contact SANBlaze for additional information.

3.8 SANBlaze 用于精密信号控制和测量的新型专利 iRiser5

SSD 测试团队现在可以设计专门定制的测试场景

美国马萨诸塞州利特尔顿 — 2024 年 2 月 28 日 — 全球领先的先进存储测试和验证技术提供商 SANBlaze Technology Inc.很高兴地宣布推出 iRiser5 设备，该设备通过 SANBlaze 为 Gen5 PCIe NVMe 测试带来精确的信号控制和测量 SBexpress-DT5 和 SBExpress-RM5 PCIe NVMe 测试系统。

SANBlaze iRiser5 设备

“iRiser5 能够精确控制从 NVMe 设备到主机的数据路径中的 PCIe 通道以及复位 (PERST) 和电源等带外信号，使用户能够设计专门针对其测试需求的测试场景， ”SANBlaze 总裁文斯·阿斯布里奇 (Vince Asbridge) 说道。“获得专利的 SANBlaze iRiser5 能够以每秒高达 百万个采样的速度监控 SSD 盘功率，在 SSD 盘上线并处理重置或电源周期时提供近乎实时的功率响应。”

受控制的信号彼此之间的间隔为 80nS，从而使用户能够最终控制复杂的测试场景，例如断言后在特定时间进行 SSD 盘重置。

“SANBlaze FPGA 控制的 iRiser5 器件能够以非常精细的间隔实现背靠背信号毛刺，目前市场上其他地方无法实现，”SANBlaze 高级副总裁 Rick Walsh 说道。“这项最先进的技术是 SANBlaze 独家专利的，旨在满足客户对测试控制和精确功率测量的需求。”

iRiser5 与 SANBlaze 标准转接卡功能（例如 SRIS/SRNS 和电源状态测试）无缝运行，在首次购买或购买后作为硬件升级添加 iRiser5 时不会丢失任何功能。

iRiser5 故障

Gen5 PCIe 通道受 iRiser5 控制，“故障”速度可达 10nS，允许在 PCIe 上故意注入错误。PCIe 通道可以在软件控制下启用和禁用，以便可以设计测试来禁用连接到 NVMe 设备的四个通道中的任何一个或全部，以测试对 PCIe 子系统实际故障的响应。

SBExpress-DT5 和 SBExpress-RM5 PCIe NVMe 测试系统

SANBlaze SBExpress-RM5 Gen5 PCIe NVMe Test System

SANBlaze SBExpress-DT5 Gen5 PCIe NVMe Test System

SANBlaze DT5 和 RM5 测试系统采用模块化设计，具有适用于 EDSFF 设备的提升板。iRiser5 占用 DT5 或 RM5 中的转接卡插槽，为 EDSFF 设备提供额外的测试功能。您可以在 SBExpress-RM5 系统中最多安装四个 iRiser5 设备，在 SBExpress-DT5 测试系统中最多安装三个 iRiser5 设备。从 10.7 版本开始，SANBlaze 提供支持 iRiser5 的软件。

SANBlaze 软件 IP

SANBlaze 的 V10.7 软件包包含多个新的 SANBlaze 认证测试套件，包括 FDP 和 OCP、其他重要更新以及客户要求的增强功能。Certified by SANBlaze 测试套件在 SANBlaze 硬件上运行，包括 SBExpress-RM5 PCIe 5.0 NVMe 机架式测试系统和 SBExpress-DT5 PCIe 5.0 NVMe 桌面测试系统。

Riser 卡技术

SBExpress-RM5（机架式）16 盘测试系统和 SBExpress-DT5（台式）测试系统均采用 SANBlaze 专有的转接卡技术，并方便地使用同一组 NVMe PCIe 5.0 转接卡。所有 SANBlaze 转接卡都能够进行单端口和双端口操作，并且可以在软件控制下进行动态切换。转接卡适用于 PCIe 5.0 速度的所有 NVMe 外形尺寸。iRiser5 无缝融入现有的立管配置，不会损失功能。

iRiser5 的特点

除了支持上述功能外，iRiser 5 还包括以下附加功能：

- 精确控制 PCIe/NVMe 电源和控制信号，同时持续监控每个被测设备 (DUT) 的电源
- 可以在每条信号线上安排一系列事件，毛刺精度高达 10 纳秒 (nS)，每个事件操作以 80nS 间隔至数小时加载。
- 可以定义简单或复杂的序列并将其从主机系统加载到 iRiser。

功能非常适合各种团队

事实证明，SANBlaze 认证软件、SBExpress-RM5 和 SBExpress-DT5 对于各种团队都很有用，包括固件、制造、开发工程师、系统工程师和现场应用工程师。通过内置的错误检测触发，可以轻松实现分析任务的自动化，从而快速、准确地解决问题。请参阅白皮书《从 SBExpress 系统触发 PCIe 分析仪》了解更多相关信息。

4. PCIe Gen 4/5/6 NVMe SSD 故障注入/热插拔和电压拉偏/功耗测试

NVMe SSD 热插拔、电压拉偏和功耗测试是 PCIe Gen 4/5/6 NVMe SSD 的一个必测项目。掉电测试是加速复现 firmware 问题的一种手段。下面是业内常用的诊断方面的测试工具和测试方法。

Quarch 在 2022 年上半年新增了不少常用的 PCIe Gen5 相关的热插拔和物理层、链路层故障注入测试模块。参见下图，图中第一个为 PCIe Gen5 MCIO to U.2 磁盘控制模块（热插拔、故障注入），该模块可以连接我们常见的 SerialCables 公司的 PCIe Gen5 switch 卡侧面的 MCIO 接口。图中的 Gen5 E3 x4 是非常常用的针对 Gen5 E3.S/E3.L 的磁盘控制模块（热插拔）。Gen5 U.3 模块则是针对传统数据中心里面的 U.3 NVMe SSD 磁盘控制模块（热插拔）。当然，Quach 也推出了一款 Gen4 EDSFF x8 的热插拔模块，之前推出的都是 Gen4 EDSFF x4 模块。

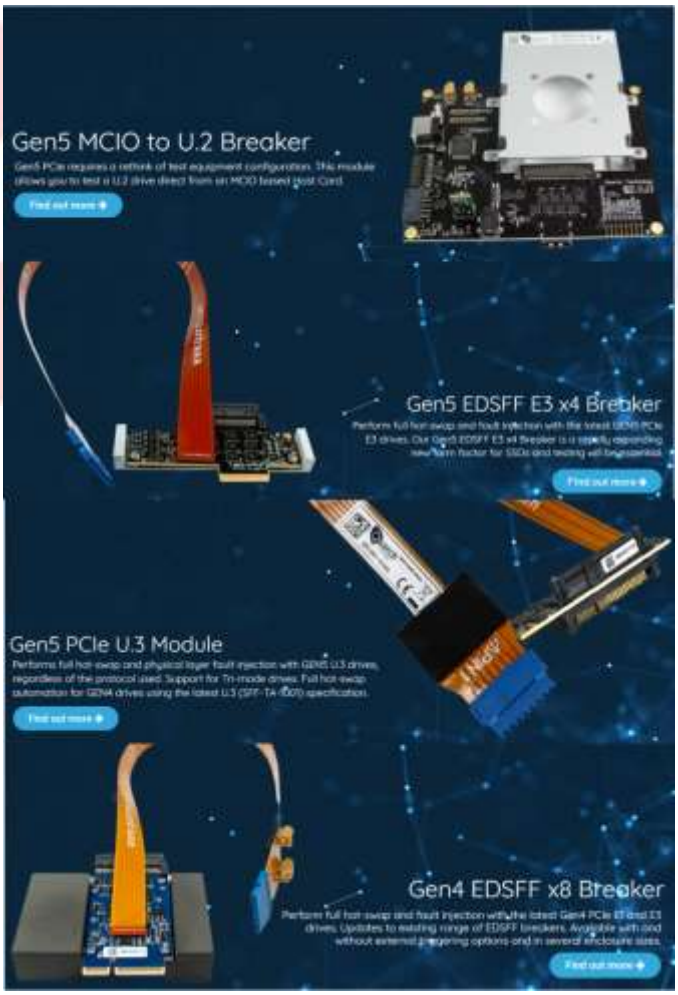


图 4-1

参见下图，如果使用该 MCIO/U.2 热插拔模块连接 SerialCables Gen5 switch 卡，需要购买 MCIO x4 – MCIO x4 cable，一端连接下图 Gen5 switch 的任意一个 MCIO 口，另外一端连接 MCIO/U.2 热插拔模块的 MCIO 口即可。当然，还有一种对于 Gen 5 U.2 盘进行热插拔和故障注入测试的方法：MCIO-U.2 cable 然后连接到 Gen5 U.2 热插拔模块。这里补充一点，SerialTek Panda PCIe Gen5 嵌入式协议分析仪可以连接 SerialCables PCIe Gen5 switch 上面的 sdb 端口实现管理抓包分析（黄色框中的三根针脚）

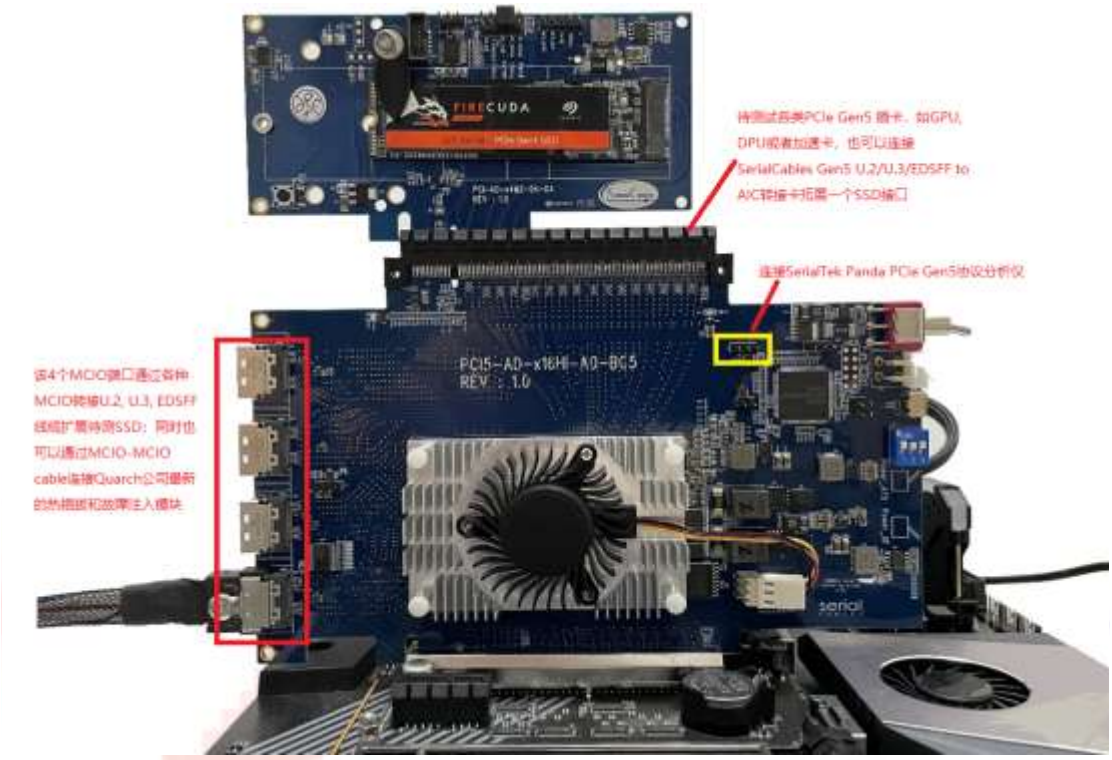


图 4-2

当然，除了新推出的这些 Gen5 模块，Quarch 之前针对各个接口也都有完整方案，包括 Gen5 U.2, U.3, EDSFF x4 E3.S, Gen 5 x16 插卡；另外转接类热插拔以及故障注入模块还有一个 U.2/AIC 模块，参见下图；

图 4-3

同时，Quarch 电源拉偏模块 PPM（programmable power module）之前也推出了常用的 Gen5 U.2/U.3, E3 x8，E1 x4 等各种 fixture 治具，Gen5 AIC 插卡和 M.2 fixture 也会马上推出。Quarch 电源和 sideband 边带信号监控分析模块 PAM（power analysis module）也推出了 Gen5 U.2/U.3, E3 x4，E1 x4 等多种型号 fixture，M.2 治具等都已经推出到市场。

4.1 PCIe Gen 4/5/6 热插拔和底层故障注入测试

该测试套件不仅为 UNH IOL 实验室使用，同时 SSD 测试业界 SAS/SATA SSD 到 PCIe Gen 3/4 NVMe SSD 的主流公司都在使用该热插拔套件进行测试。

说明：1）针对 Hot Plug 热插拔的基本概感兴趣的话，可以参考附录第 11 章的 11.5.7 部分。

2）注意：

“热插拔”仅对企业级 SSD 试用，一般是 U.2/U.3/E1.S/E1.L/E3.S/E3.L 或者 SAS 以及企业级 SATA SSD，一般称为 Drive Control Module（磁盘控制模块）。

针对 M.2, AIC 插卡我们一般称呼该模块为 Card Control Module （卡类控制模块），一般使用这类模块主要用来导入物理层和链路层问题，从而测试待测 DUT 碰到各类错误的异常处理能力。

热插拔是 PCIe Gen 4/5/6 NVMe SSD 以及 SAS/SATA HDD/SSD 的一个必测项目。热插拔测试和掉电测试是完全不同的两个测试，一般来讲，掉电测试相对简单，主要用来加速复现一些 SSD 或者 firmware 的问题。但是热插拔测试的几个关键点是单纯的掉电测试无法仿真模拟的，如下：

- **热插拔有针脚接入顺序的先后差异；**
- **热插拔过程中可能会出现针脚接触槽位时出现信号时通时断类似于接触不好的情况；**
- **热插拔过程中以及插到位置稳定以后可能会产生信号毛刺的问题**

当然，热插拔测试带来的一个附属功能就是掉电测试，只是通过热插拔模块实现掉电功能需要选择合适的型号，例如有些信号可以去除信号毛刺注入功能，这些产品价格就会经济一些。下面以业内目前最常测试的 NVMe SSD 热插拔为例，SAS/SATA 测试工具和测试方法基本一致，这里不在赘述。

- Mandatory for NVMe promoters group certification
  - UNH-IOL plugfest events
  - Free test provided in QCS (other tests are a paid subscription)
- Essential for enterprise drives
- Useful in many other test cases
  - Power cycles are useful in many tests
  - Power cycle affects the entire storage system (SSD, host, BIOS, drivers, applications)
  - Many related problems can be found through this automation

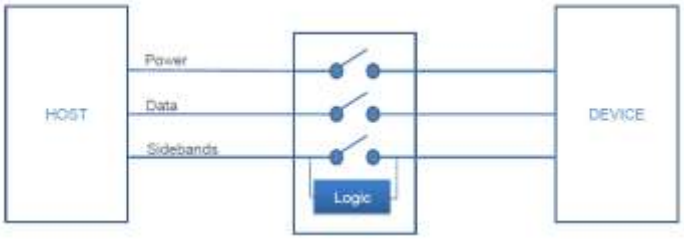


图 4-4

下面介绍的 Quarch 测试套件不仅为 UNH IOL 实验室使用，同时 SSD 测试业界 PCIe Gen 3/4/5 NVMe SSD 和 SAS/SATA SSD 的主流公司都在使用该热插拔套件进行测试。

如上所述，SSD 热插拔测试套件的主要功能是实现对于盘的热插拔的测试，尤其是 U.2/U.3/EDSFF NVMe SSD，对于 M.2 NVMe SSD 或者基于 PCIe 插卡的 SSD 一般情况下不进行热插拔，但是通过该套件对于某些针脚（包括电源）模拟断掉，接触不好，x4 的差分信号线中某个方向，甚至某个方向的两个差分信号中的一根断掉不通，或者对于该差分信号在 SSD 盘的方向导入一些信号毛刺来模拟一些故障场景看 SSD 在这些特殊情况下是否还可以可靠稳定地工作等。简单列举一下通过 Quarch 这些热插拔模块可以实现的测试。

- **模拟 SSD 在热插拔过程中如下问题和故障**
- **模拟盘的热插拔**
- **模拟盘热插拔过程中导致的 pin bounce 接触不好的情况**
- **模拟某些针脚断掉**
- **模拟某些针脚长通**
- **模拟某些针脚上面有信号毛刺**
- **物理毛刺的多少？注入一次毛刺，还是一直有毛刺？间隔时间多长？**
- **毛刺的高低，疏密，持续的时间长短**
- **模拟非常快速的通/断测试**

目前这些热插拔套件提供各种接口类型，包括 Gen 4 U.2/U.3, M.2, PCIe Slot 以及 EDSFF x4/x8 等，这些热插拔模块需要串接在 SSD 和主板/背板之间，通过 Python 脚本或者图形化 GUI 界面进行测试。

- Modules available for U.2, U.3, E1, E3 and more
- Also relevant to fixed interfaces (AIC and M.2)
  - Lane width reduction
  - Power cycle / reset (Driving PERST)
- Fault injection tests
  - Simulate bad cables, damaged connectors and data corruption

Real world capture of pin-bounce

图 4-5

4.1.1 Quarch Gen 5 热插拔模块

下面是针对 PCIe Gen 5 常见 SSD 和插卡的控制模块，其它接口控制模块陆续增加过程中。

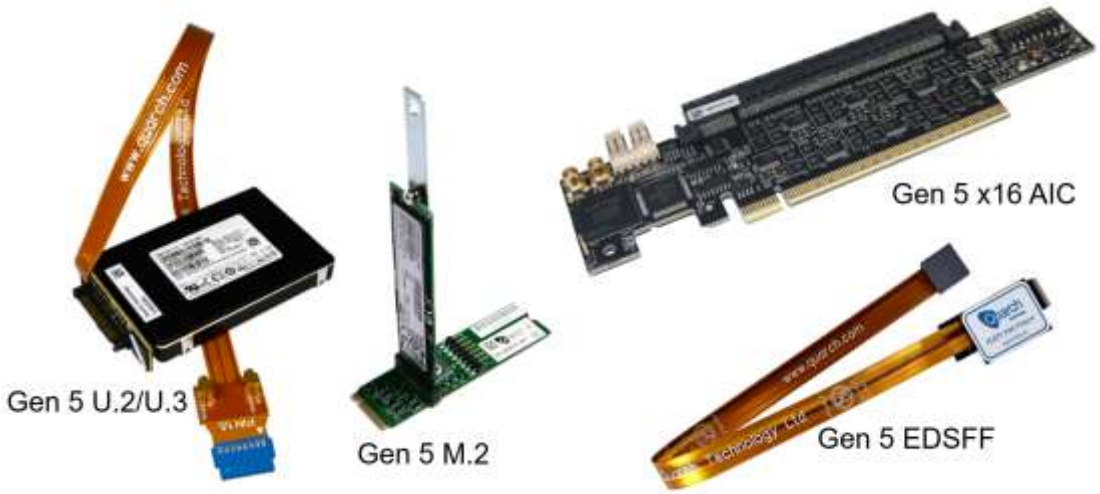


图 4-6

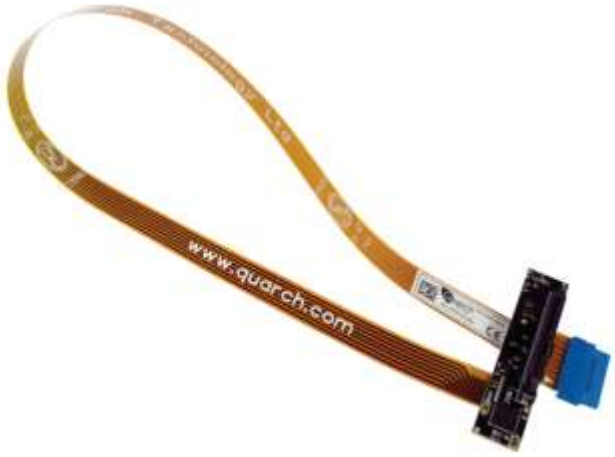
4.1.1.1 PCIe Gen5 U.2/U.3 热插拔/故障注入模块

Optional triggering versions have MCX trigger in/out.

U.2 and U.3 versions, bothsupporting up to Gen5 speeds.

SFF BREAKERS: SUPPORTS SFF-8639 DEVICES UP TO GEN5 SPEEDS. DUALPORT DRIVES SUPPORTED.
U.2 BREAKER: DRIVE+MONITOR: PERST, CLKREQ/PERSTB, SMCLK, SMDAT, DUALPORTEN, IF_DET, PWR_DIS, PRSNT, HPT0, HPT1. MONITOR: ACTIVITY, WAKE. U.3 BREAKER: DRIVE+MONITOR: PRSNT, PWRDIS, DUALPORTEN, IFDET, IFDET2, HPT0, HPT1, PERST, PERSTB. MONITOR: SMBCLK, SMBDAT, WAKE 8

4.1.1.2 PCIe Gen5 经济型热插拔模块



Supports SFF-8639 drives: U.2, U.3, SAS and SATA.

Supports speeds up to Gen5 PCIe and SAS4.

LITE BREAKER: SUPPORTS SFF-8639 U.2 DEVICES UP TO GEN5 SPEEDS. DUALPORT DRIVES SUPPORTED. SUPPORTS ALL SFF-8639 DEVICES INCLUDING U.2, U.3 SAS AND SATA. CONTROLLED SIGNALS: 12V_POWER, 12V_CHARGE, 5V_POWER, 5V_CHARGE, 3V3_AUX, PERST_A, PERST_B, SIDEBAND (ALL OTHER SIDEBANDS).

4.1.1.3 24G SAS SSD 热插拔模块

Optional triggering version has MCX trigger in/out.

Supports SAS/SATA drives up to SAS4 speeds

4.1.1.4 12G SAS SSD 热插拔模块

Supports SAS/SATA drives up to SAS3 speeds.

SAS LITE BREAKER: SUPPORTS SAS/SATA UP TO SAS3 SPEEDS. CONTROLLED SIGNALS: 5V_CHARGE, 5V_POWER, 12V_CHARGE, 12V_POWER, MATED, POWER_DISABLE.

4.1.1.5 Gen5 E1 SSD 热插拔模块

Optional triggering version has MCX trigger in/out.

Supports E1.S and E1.L drives up to Gen5 speeds

4.1.1.6 Gen5 E3 SSD 热插拔模块

Optional triggering version has MCX trigger in/out.

Supports E3 drives up to Gen5 speeds.

E3 x4, E3 x8 and E3-2T x8 versions available.

E3BREAKERS:SUPPORTSEDSFFE3DEVICESUPTOGEN5SPEEDS,VERSIONSAVAILABLEDIFFERBYLANEWIDTHANDENCLOSURESIZE.
E1 SIGNALS: DRIVE+MONITOR: PERST0, PERST1, PRSNT0, LED, SMBRST, PWRDIS, DUALPORTEN. MONITOR: SMBCLK, SMBDAT.
E3 SIGNALS: DRIVE+MONITOR: PERST0, PERST1, PRSNT0, PRSNT1, LED, SMBRST, PWRDIS, DUALPORTEN. MONITOR: SMBCLK, SMBDAT, RFU_A42, RFU_B8.

4.1.1.7 Gen5 M.2 故障注入模块

Optional triggering version has MCX trigger in/out.

Supports M.2 M-Key devices up to Gen5 speeds.

Horizontal form factor.

M.2 BREAKERS: SUPPORTS M.2 M-KEY DEVICES UP TO GEN5 SPEEDS. DRIVE+MONITOR: PEWAKE, CLKREQ, LED1, PERST, SUSCLK, ALERT

4.1.1.8 Gen5 x16 插卡故障注入模块

Optional triggering version has MCX trigger in/out.

Optional inrush limit version reduces inrush current load on hosts during SSD hot-plug

Power injection port for Programmable Power Module, allowing measurement and margining

AIC BREAKERS: SUPPORTS AIC/SLOT DEVICES WITH UP TO 16 LANES, AND UP TO GEN5 SPEEDS, TRIGGERING AND INRUSHLIMITED VERSIONS ARE AVAILABLE. DRIVE+MONITOR: PERST, WAKE, CLKREQ, PWRBRK. MONITOR ONLY: SMCLK, SMDAT

4.1.1.9 Gen5 x16 经济型热插拔模块

Supports AIC/slot devices up to Gen5 speeds.

Optional inrush limit version reduces inrush current load on hosts during SSD hot-plug.

Direct USB header to power/ control direct from a host PC.

AIC LITE BREAKERS: SUPPORTS PCIE AIC/SLOT DEVICES UP TO GEN5 SPEEDS. CONTROLLED SIGNALS: 12V_POWER, 3V3_POWER, 3V3_AUX, PRSNT, WAKE, PERST, CLKREQ, REFCLK, JTAG, SMCLK, SMDAT, PWRBRK.

4.1.1.10 其它非对称（带接口转接）热插拔和故障注入模块

Reduce SI losses for Gen5 testing by removing additional adaptors/cables and test direct in a slot

Combine breaker and power analysis features onto a single device for all-in-one testing.

Larger form factor allows additional features on some devices, such as sideband probe points and PAM (Power Analysis Module) functionality with power, sideband and REFCLK monitoring.

4.1.1.10.1 GEN5 AIC TO U.2 转接热插拔模块

Supports PCIe U.2 drives up to Gen5 speeds in a PCIe x4 slot.

Optional triggering version with in/out trigger points.

PPM power injection port for Quarch Power Analysis Module.

Probe points for selected sidebands

Direct USB control port in addition to Torridon controller.

AIC TO U.2 BREAKER: SUPPORTS U.2 X4 DRIVES UP TO GEN5 SPEEDS, TRIGGERING VERSIONS ARE AVAILABLE. DIRECT USB OR TORRIDON CONTROL PORT. DRIVE+MONITOR: PERST, PERSTB, DUALPORTEN, IF_DET, WAKE, PWR_DIS, HPT0, HPT1. MONITOR: SMCLK, SMDAT, ACTIVITY, PRSNT, IFDET_2, PRSNT1, PWRBRK, RSVD_A19, RSVD_A32, RSVD_P2

4.1.1.10.2 GEN5 MCIO TO U.2 转接热插拔模块

Connect to a compatible MCIO host card, removing the need for adaptors or a U.2 host.

PPM power injection.

Probe points for selected sidebands.

Direct USB control port in addition to Torridon controller.

External 12V PSU for drive supply.

4.1.1.10.3 GEN5 AIC/EDSFF 转接热插拔模块

Supports EDSFF devices up to x16. Versions available with lower lane widths.

Optional triggering version with in/out trigger points.

Optional PAM connector to add power and sideband analysis

Direct USB control port in addition to Torridon controller.

AIC TO EDSFF BREAKER: COMBINED BREAKER AND POWER ANALYSIS PRODUCT, DUE Q1 2024

4.1.2 Quarch Gen 4 热插拔模块

下图为使用销售较多的常见的 PCIe Gen 4 SSD 盘和插卡（针对芯片验证卡）的控制模块。

图 4-7 各种接口 Gen 4/5 插卡以及 SSD 热插拔模块

下图为最精简配置 Gen 4 U.2 NVMe SSD 热插拔模块连接图实拍，采用 QTL1260 单端口控制器，支持 USB 或者串口管理，直接连接控制电脑。

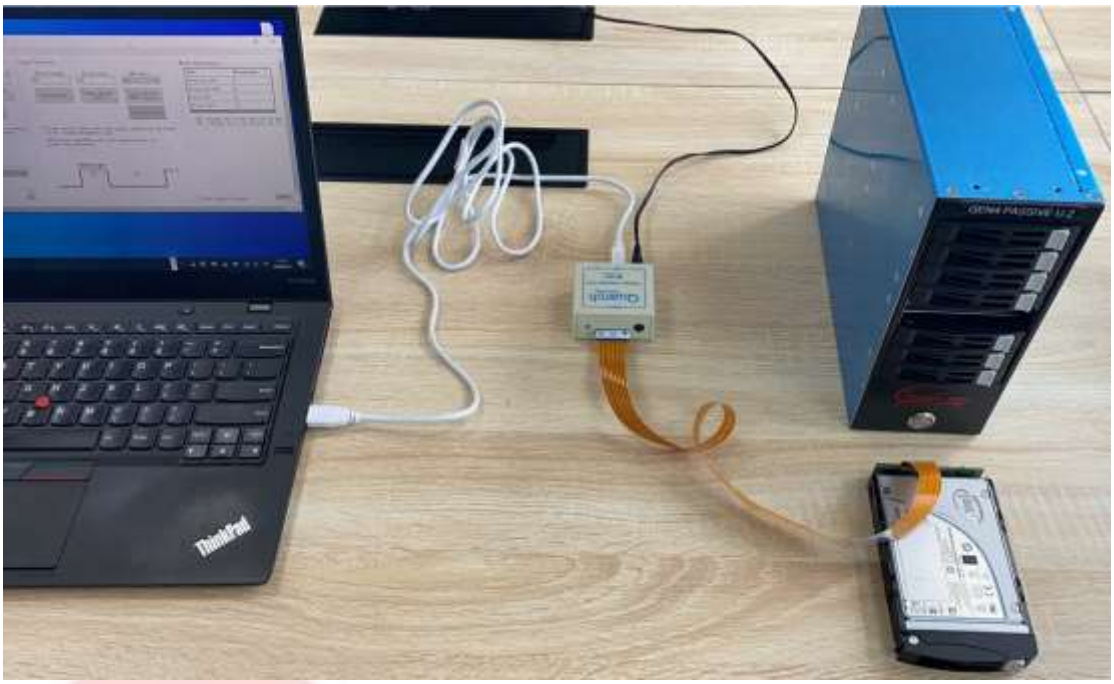


图 4-8 最精简配置 Gen 4 U.2 NVMe SSD 热插拔模块连接图实拍

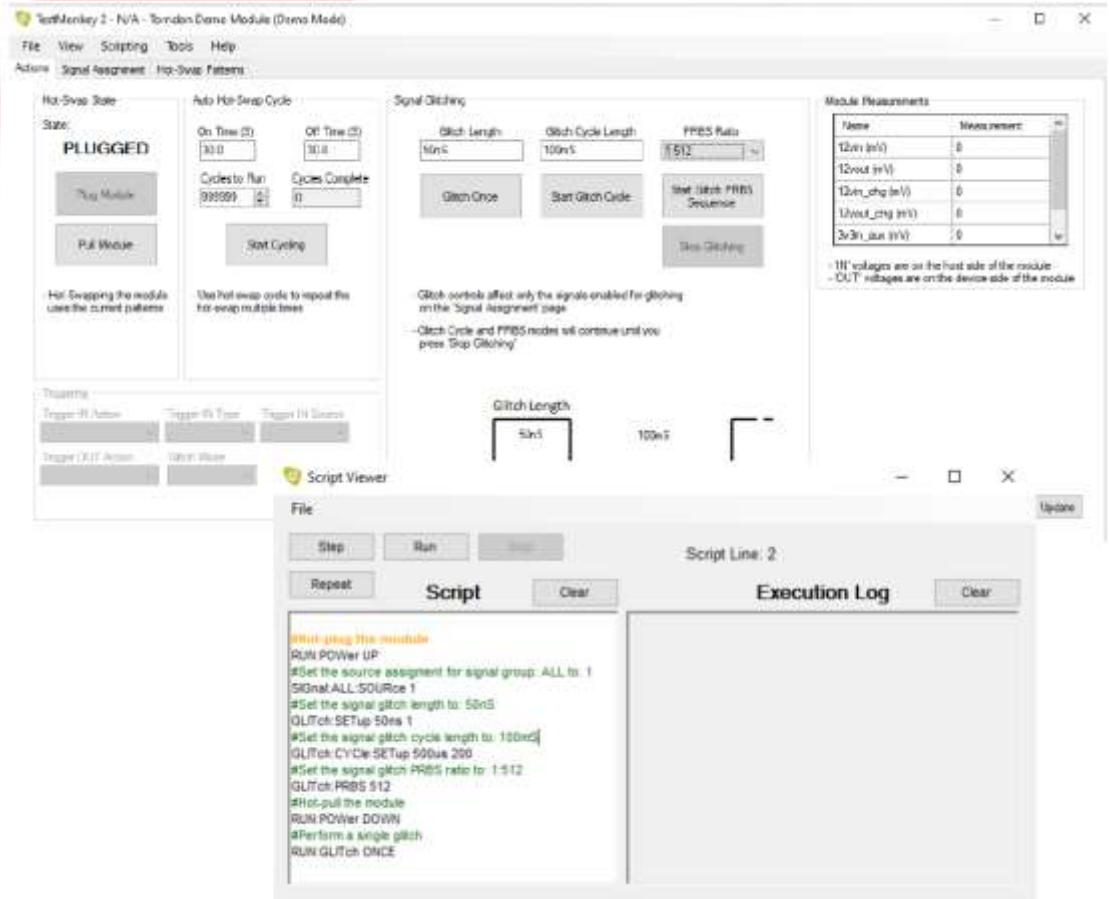


图 4-9

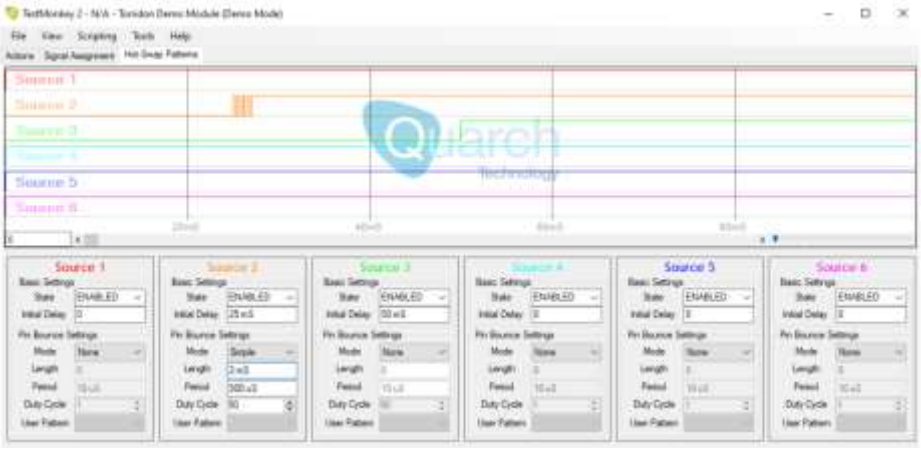


图 4-10

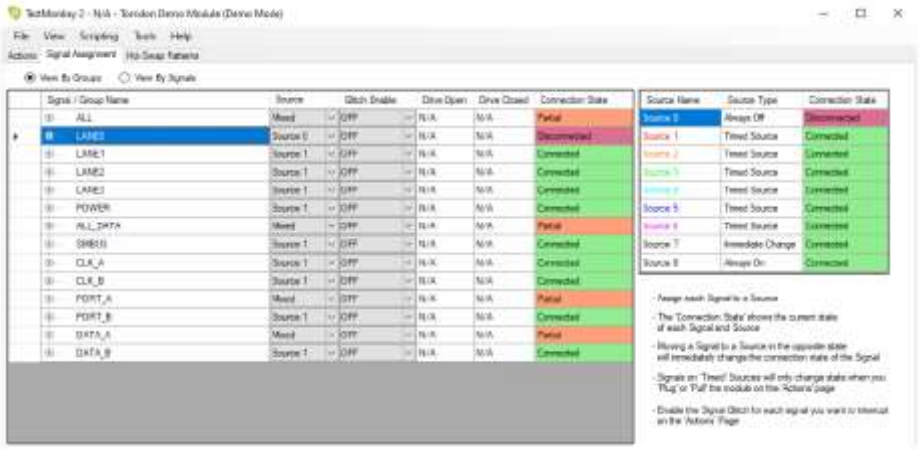


图 4-11

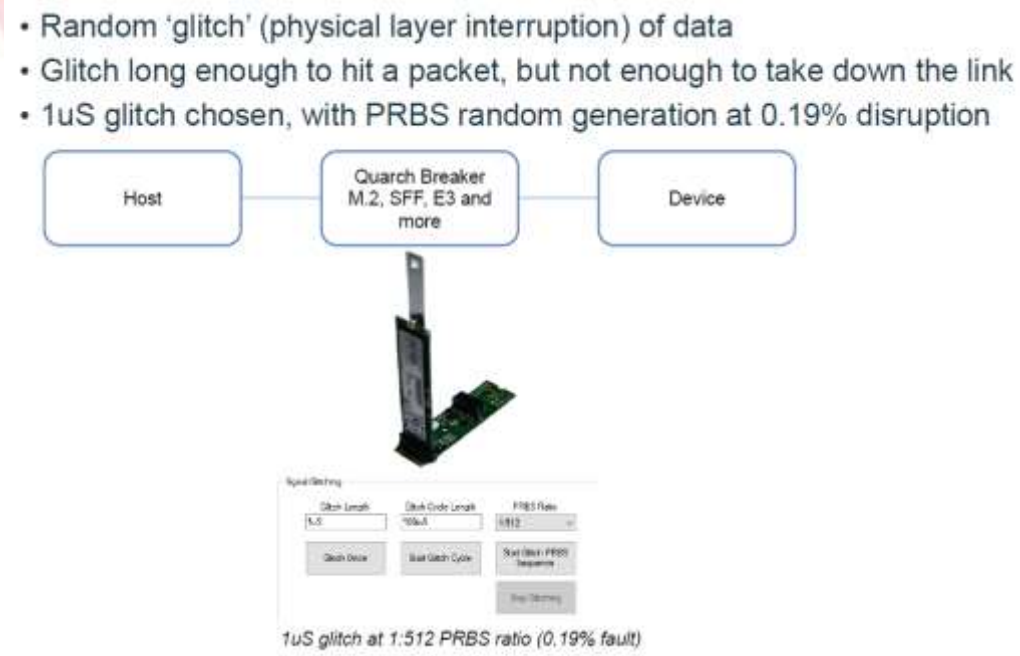


图 4-12

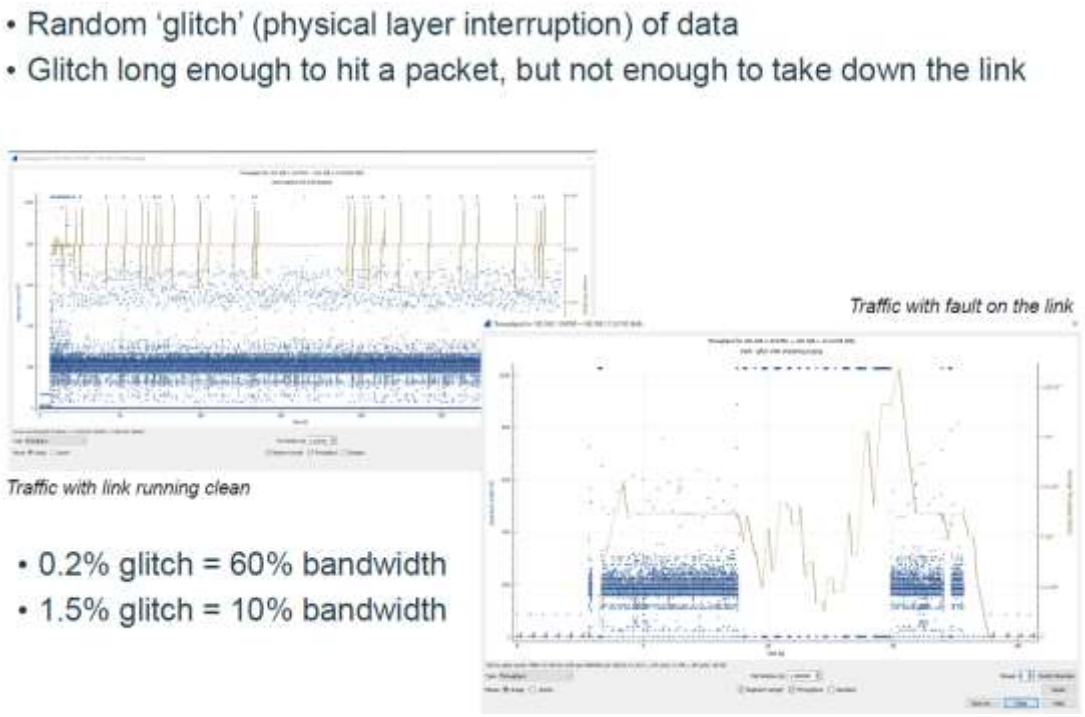


图 4-13

4.1.3 Torridon 系列管理模块

如果需要通过脚本同时管理很多 SSD 热插拔模块，建议采用 4-port 或者 28-port 管理控制器，支持网络管理，可以同时控制 4 个或者 28 个热插拔模块；同时这些 28 端口模块可以支持级联 4 台，达到最大管理 112 端口的热插拔/故障注入模块。参见下图左边两图。

图 4-14

4.1.3.1 单端口控制模块 – interface kit，支持串口+USB 管理

Button switches between direct USB and USB VCOM mode.

12v PSU powers the Breaker, so it is independent of host power.

Rear USB-2 and RJ-45 serial port for control.

下面是一个连接示意图。

Control a single breaker module via USB or Serial.

Ideal for simple bench tests and small scale automation.

INTERFACE KIT: CONTROLS A SINGLE BREAKER ACROSS EITHER: DIRECT USB, USB VIRTUAL COM PORT OR DIRECT RS-232 VIA RS232-D WIRED RJ-45 PORT. 12V 15W SUPPLY.

4.1.3.2 4 端口控制模块 – 支持网络+USB 管理

Control up to 4 Breaker modules atthe same time.

1U high and can be front or rear rack mounted with an additional kit.

USB, LAN and RS232 Serial control ports.

12v PSU powers the Breakers, so they are independent of host power.

Link port allows two controllers to be chained for 8 device control.

4.1.3.3 28 端口控制模块 – 支持网络+USB 管理

下图是一个多台 28 端口控制器连接的图片

USB, LAN and RS232 Serial control ports.

12v PSU powers the Breakers, so they are independent of host power.

Link port allows two controllers to be chained for 48 device control.

Rear ports are ideal for cable breaker modules.

4.1.4 多协议低速协议通断测试模块

- Aimed at automotive, aviation and similar
- Supports a wide range of serial links
  - SPI
  - I2C
  - RS232
  - 1000BaseT1
- Controls up to 4 lines + power
- Full 'breaker' features set for fault injection
- PAM monitoring for digital signals
  - Also for 'power good' on the voltage rail

图 4-15

4.1.5 Quarch Compliance Suite 软件

通过上述热插拔模块，配合 Quarch Compliance Suite 可以很方便地在实验室实现 UNH IOL 的 hotplug 认证测试。

- Required test for NVMe promoters group certification
- Run at all UNH Plugfest events
  - Multiple different hot-plug speeds from 10mS to 500mS
  - Drive must enumerate at the same link speed/lane width every time

Common Failures

- Server blue screen / crash
- Unexpected power off
- Drive does not enumerate
- Drive fails completely after a number of cycles

图 4-16

- Helps customers get basic tests done quickly
  - Hot-swap tests (including pin-bounce and timing sweeps)
  - Power analysis
  - PCIe lane reduction

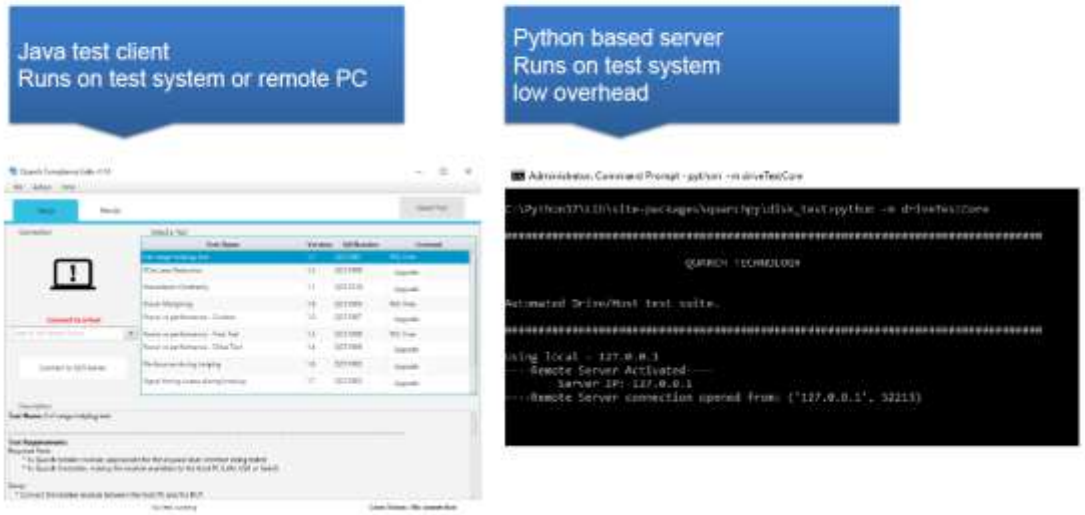


图 4-17

Sniffer

- Real-time log of progress
- 'Check-points' which must pass
  - Drive enumerating (link speed and lane width verified)
  - FIO job completed
- Information points
  - IOPS and MB/s during workload
  - Idle Power consumption
- QPS analysis files
  - For tests which contain measurement
- PDF report
  - Summarizing check-points and information points in one document

图 4-18

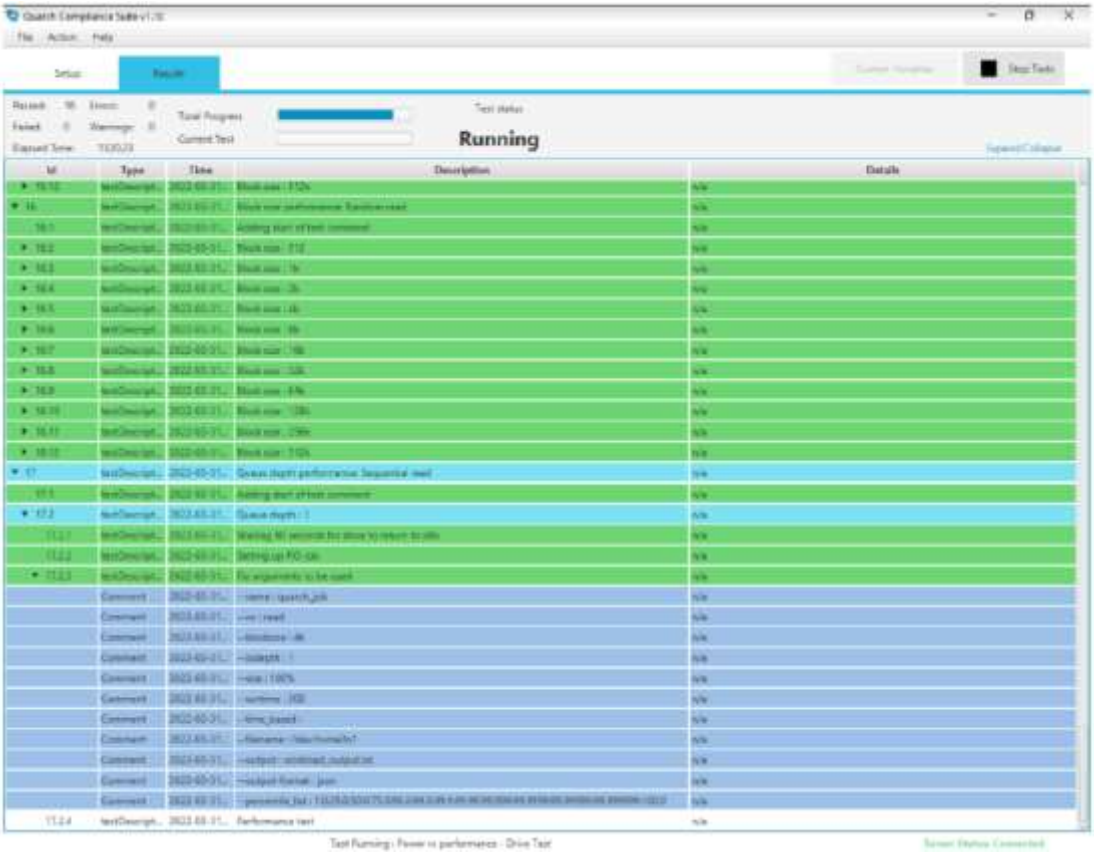


图 4-19

Enhanced margining test showing drive running at idle, but failing under workload

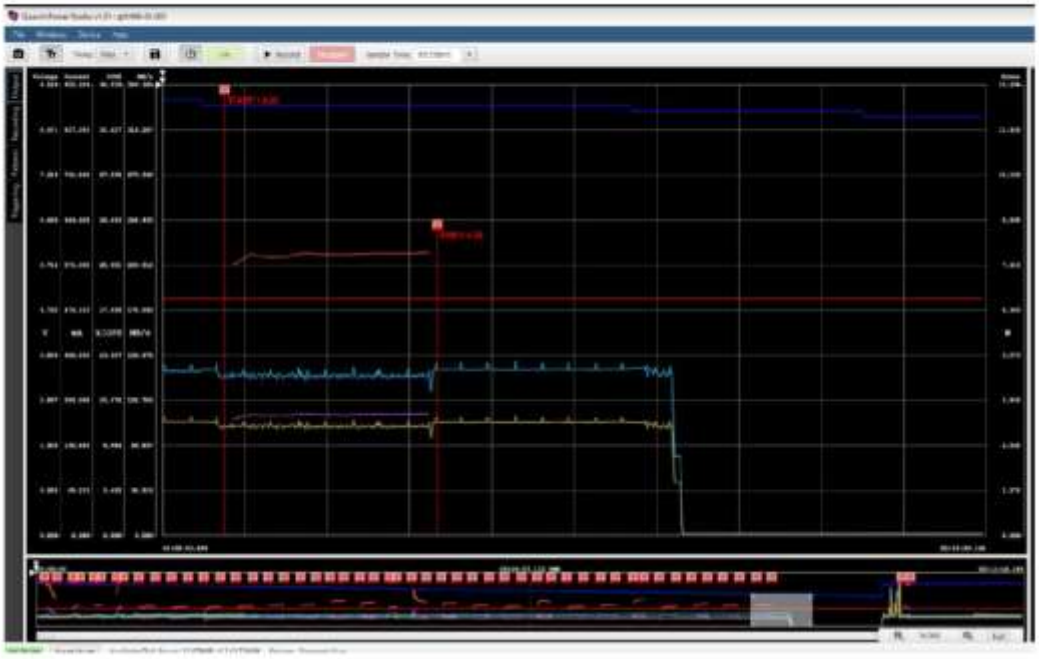
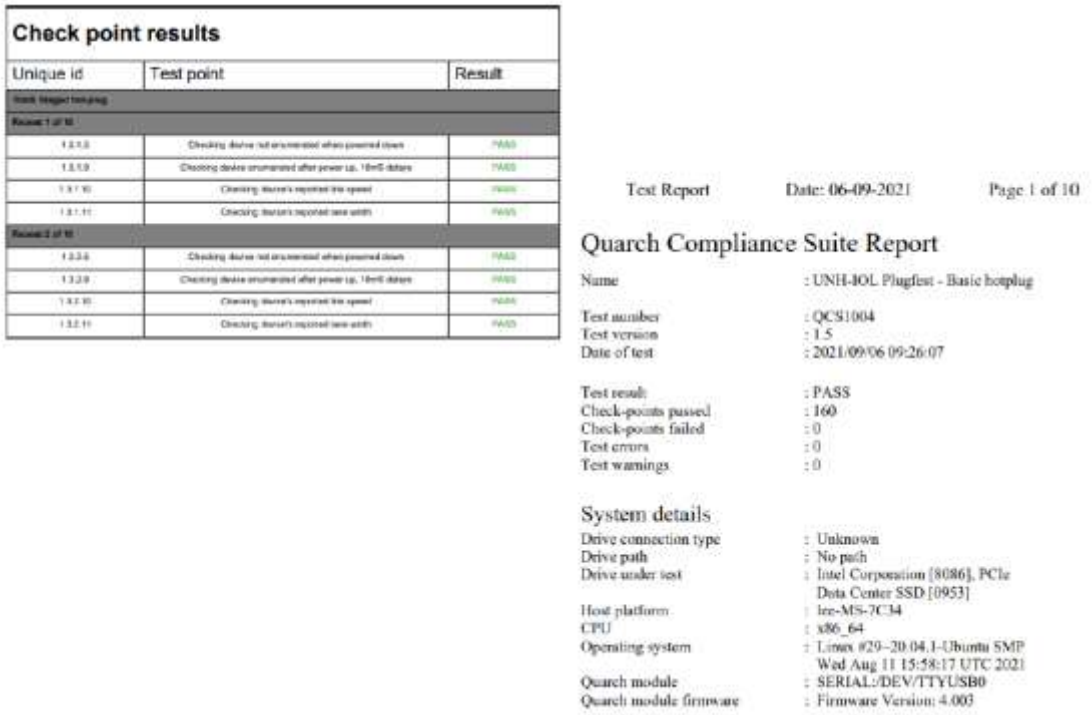


图 4-20

- 27 hours of recording, 50+ workloads

图 4-21



Check point results

Unique id	Test point	Result
Hotplug test group		
Round 1 of 10		
1.3.1.3	Checking device not enumerated when powered down	PASS
1.3.1.9	Checking device enumerated after power up, 10mS delays	PASS
1.3.1.10	Checking device's reported link speed	PASS
1.3.1.11	Checking device's reported lane width	PASS
Round 2 of 10		
1.3.2.3	Checking device not enumerated when powered down	PASS
1.3.2.9	Checking device enumerated after power up, 10mS delays	PASS
1.3.2.10	Checking device's reported link speed	PASS
1.3.2.11	Checking device's reported lane width	PASS

Test Report Date: 06-09-2021 Page 1 of 10

Quarch Compliance Suite Report

Name	: UNH-IOL Plugfest - Basic hotplug
Test number	: QCS1004
Test version	: 1.5
Date of test	: 2021/09/06 09:26:07
Test result:	: PASS
Check-points passed	: 160
Check-points failed	: 0
Test errors	: 0
Test warnings	: 0

System details

Drive connection type	: Unknown
Drive path	: No path
Drive under test	: Intel Corporation [8086], PCIe Data Center SSD [0953]
Host platform	: Iee-MS-7C34
CPU	: x86_64
Operating system	: Linux #29~20.04.1-Ubuntu SMP Wed Aug 11 15:58:17 UTC 2021
Quarch module	: SERIAL:/DEV/TTYUSB0
Quarch module firmware	: Firmware Version: 4.003

图 4-22

- Python package (2.x and 3.x support)
- Simple commands for easy automation

```
from quarchpy.device import *

# Scan for quarch devices on the system
deviceList = scanDevices ()
# Ask user to select a device
moduleStr = userSelectDevice (deviceList)
# Connect to the device
myDevice = quarchDevice(moduleStr)

myDevice.sendCommand("run:power up")
sleep(5)
myDevice.sendCommand("run:power down")

myDevice.closeConnection()
```

图 4-23

通过 Quarch 各种 SSD 磁盘控制模块或者插卡类模块，配合 Quarch Compliance Suite 可以很方便地在实验室实现 UNH IOL 的 hotplug 认证测试，以及很多其它测试项目，例如针对 M.2 低功耗测试的 power state test 测试等。同时配合 PPM, PAM 等可以实现更多、更复杂的测试功能。参加下图。

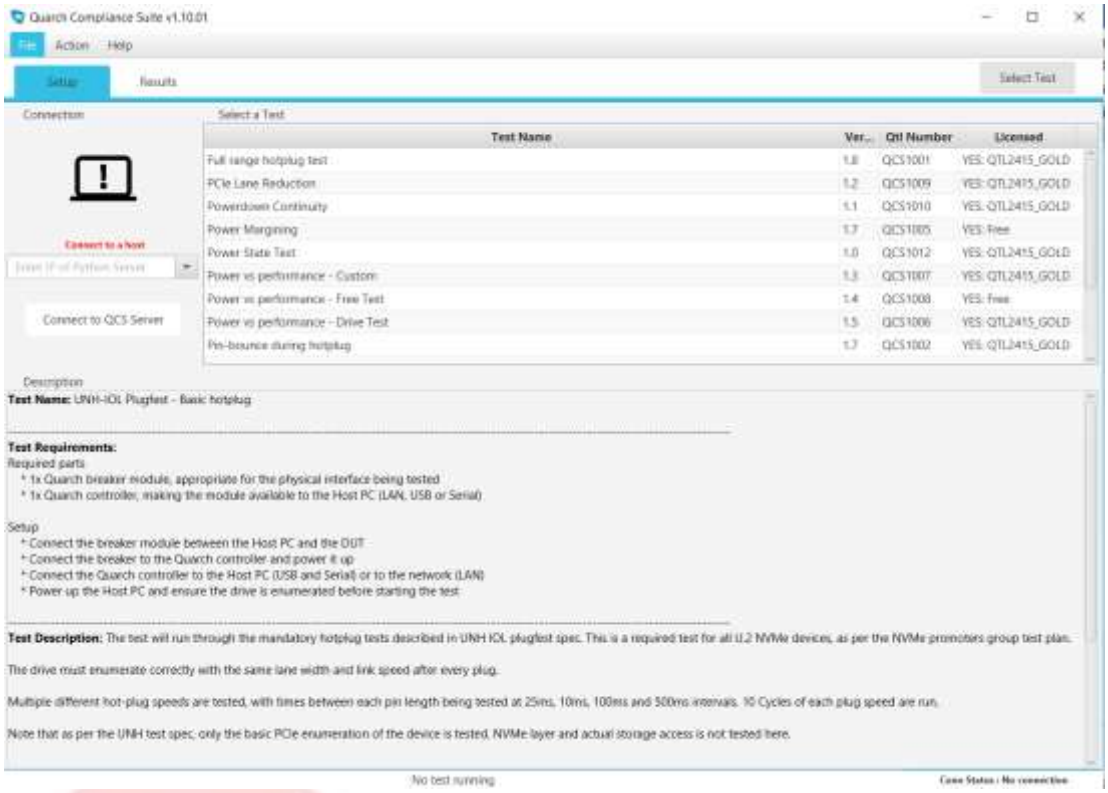

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Full range hotplug test
PCIe Lane Reduction
Powerdown Continuity
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Connect to a host
Connect to QCS Server
Description
Test Name: UNH-IOL Plugfest - Basic hotplug
Test Requirements:
Required parts
* 1x Quarch breaker module, appropriate for the physical interface being tested
* 1x Quarch controller, making the module available to the Host PC (LAN, USB or Serial)
Setup
* Connect the breaker module between the Host PC and the DUT
* Connect the breaker to the Quarch controller and power it up
* Connect the Quarch controller to the Host PC (USB and Serial) or to the network (LAN)
* Power up the Host PC and ensure the drive is enumerated before starting the test
Test Description: The test will run through the mandatory hotplug tests described in UNH IOL plugfest spec. This is a required test for all U.2 NVMe devices, as per the NVMe promoters group test plan.
The drive must enumerate correctly with the same lane width and link speed after every plug.
Multiple different hot-plug speeds are tested, with times between each pin length being tested at 25ms, 10ms, 100ms and 500ms intervals. 10 Cycles of each plug speed are run.
Note that as per the UNH test spec, only the basic PCIe enumeration of the device is tested. NVMe layer and actual storage access is not tested here.
No test running

图 4-24

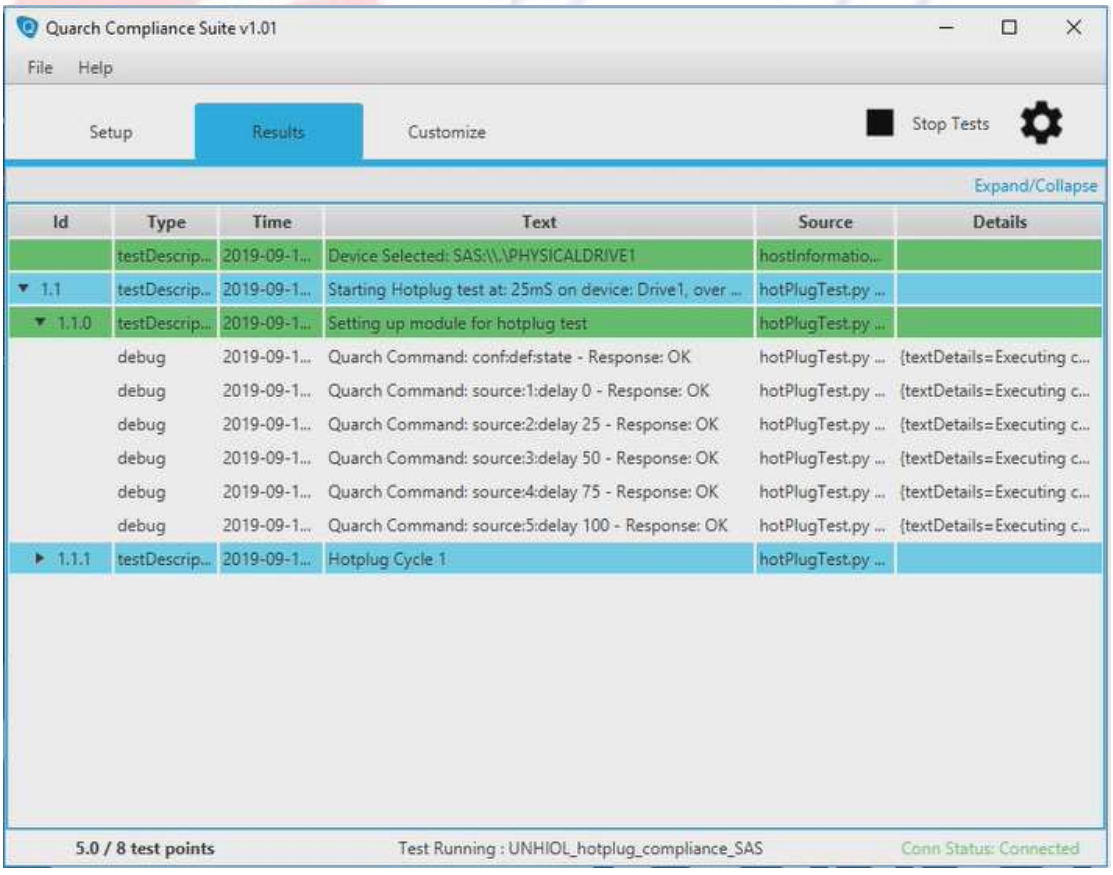

Quarch Compliance Suite v1.01
File
Help
Setup
Results
Customize
Stop Tests
Expand/Collapse
Id
Type
Time
Text
Source
Details
testDescrip...
2019-09-1...
Device Selected: SAS:\\.\PHYSICALDRIVE1
hostInformatio...
1.1
Starting Hotplug test at: 25mS on device: Drive1, over ...
hotPlugTest.py ...
1.1.0
Setting up module for hotplug test
debug
Quarch Command: conf:def:state - Response: OK
Quarch Command: source:1:delay 0 - Response: OK
Quarch Command: source:2:delay 25 - Response: OK
Quarch Command: source:3:delay 50 - Response: OK
Quarch Command: source:4:delay 75 - Response: OK
Quarch Command: source:5:delay 100 - Response: OK
{textDetails=Executing c...
1.1.1
Hotplug Cycle 1
5.0 / 8 test points
Test Running : UNHIOL_hotplug_compliance_SAS
Conn Status: Connected

图 4-25

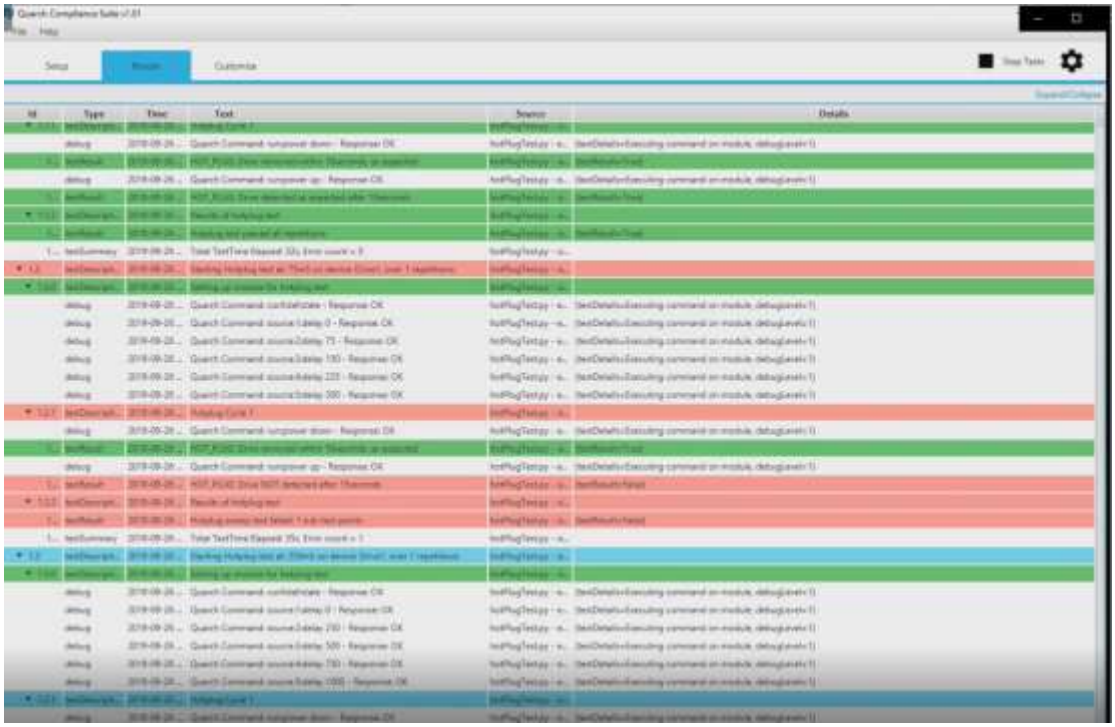


图 4-26 针对 UNH IOL 实验室的 hotplug 兼容性测试软件

4.1.5.1 Quarch Compliance Suite v1.10.01 测试用例介绍

下面是最新的 Quarch Compliance Suite v1.10.01 的主要测试用例的截图，供参考，大家可以看到每个测试用例需要配置的 Quarch 模块的要求，以及主要的测试内容。

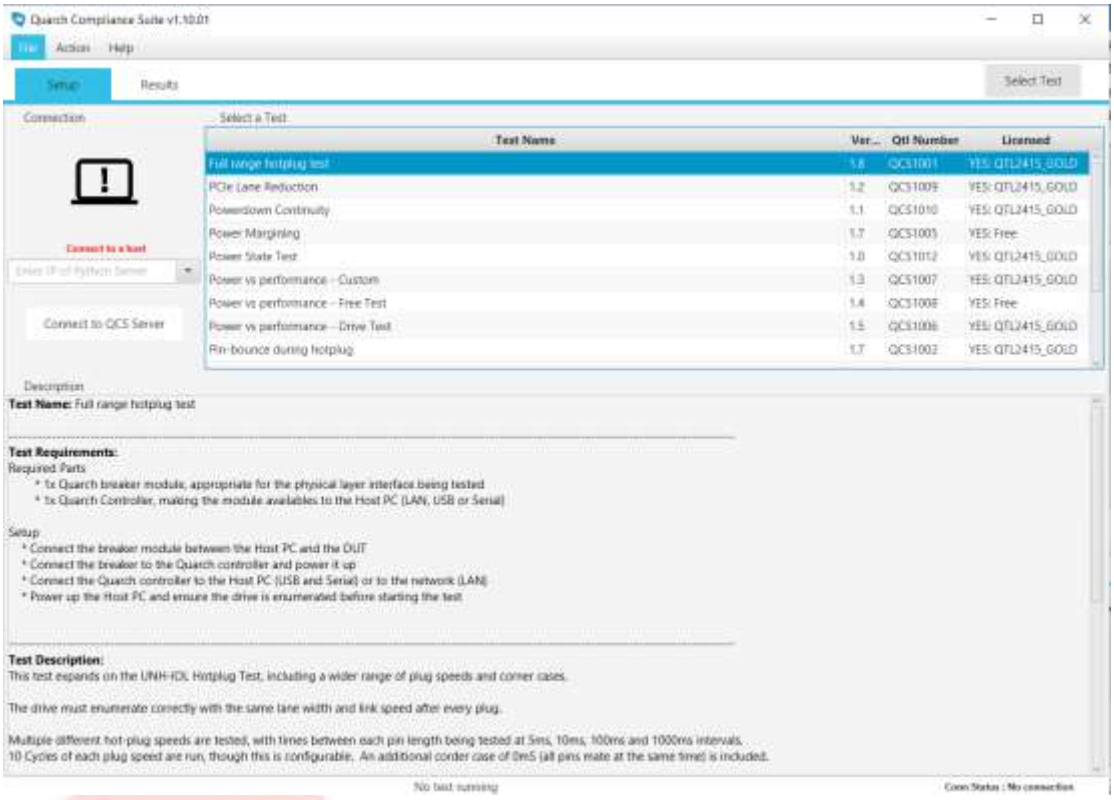

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Full range hotplug test 1.8 QCS1001 YES: QTL2415_GOLD
PCIe Lane Reduction 1.2 QCS1009 YES: QTL2415_GOLD
Powerdown Continuity 1.1 QCS1010 YES: QTL2415_GOLD
Power Margining 1.7 QCS1005 YES: Free
Power State Test 1.0 QCS1012 YES: QTL2415_GOLD
Power vs performance - Custom 1.3 QCS1007 YES: QTL2415_GOLD
Power vs performance - Free Test 1.4 QCS1008 YES: Free
Power vs performance - Drive Test 1.5 QCS1006 YES: QTL2415_GOLD
Pin-bounce during hotplug 1.7 QCS1002 YES: QTL2415_GOLD
Connect to a host
Connect to QCS Server
Description
Test Name: Full range hotplug test
Test Requirements:
Required Parts
* 1x Quarch breaker module, appropriate for the physical layer interface being tested
* 1x Quarch Controller, making the module available to the Host PC (LAN, USB or Serial)
Setup
* Connect the breaker module between the Host PC and the DUT
* Connect the breaker to the Quarch controller and power it up
* Connect the Quarch controller to the Host PC (USB and Serial) or to the network (LAN)
* Power up the Host PC and ensure the drive is enumerated before starting the test
Test Description:
This test expands on the UNH-IOL Hotplug Test, including a wider range of plug speeds and corner cases.
The drive must enumerate correctly with the same lane width and link speed after every plug.
Multiple different hot-plug speeds are tested, with times between each pin length being tested at 5ms, 10ms, 100ms and 1000ms intervals.
10 Cycles of each plug speed are run, though this is configurable. An additional corder case of 0mS (all pins mate at the same time) is included.
No test running

图 4-27

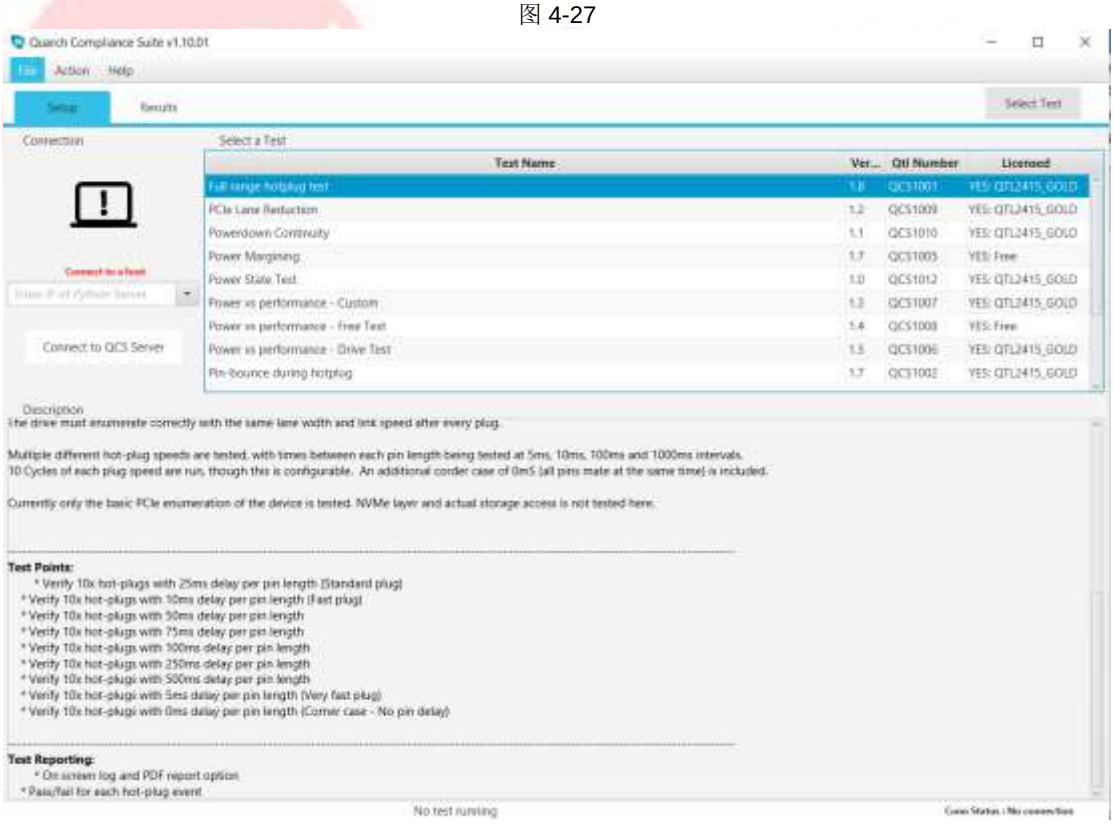

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Full range hotplug test 1.8 QCS1001 YES: QTL2415_GOLD
PCIe Lane Reduction 1.2 QCS1009 YES: QTL2415_GOLD
Powerdown Continuity 1.1 QCS1010 YES: QTL2415_GOLD
Power Margining 1.7 QCS1005 YES: Free
Power State Test 1.0 QCS1012 YES: QTL2415_GOLD
Power vs performance - Custom 1.3 QCS1007 YES: QTL2415_GOLD
Power vs performance - Free Test 1.4 QCS1008 YES: Free
Power vs performance - Drive Test 1.5 QCS1006 YES: QTL2415_GOLD
Pin-bounce during hotplug 1.7 QCS1002 YES: QTL2415_GOLD
Connect to a host
Connect to QCS Server
Description
the drive must enumerate correctly with the same lane width and link speed after every plug.
Multiple different hot-plug speeds are tested, with times between each pin length being tested at 5ms, 10ms, 100ms and 1000ms intervals.
10 Cycles of each plug speed are run, though this is configurable. An additional corder case of 0mS (all pins mate at the same time) is included.
Currently only the basic PCIe enumeration of the device is tested. NVMe layer and actual storage access is not tested here.
Test Points:
* Verify 10x hot-plugs with 25ms delay per pin length (Standard plug)
* Verify 10x hot-plugs with 10ms delay per pin length (Fast plug)
* Verify 10x hot-plugs with 50ms delay per pin length
* Verify 10x hot-plugs with 75ms delay per pin length
* Verify 10x hot-plugs with 100ms delay per pin length
* Verify 10x hot-plugs with 250ms delay per pin length
* Verify 10x hot-plugs with 500ms delay per pin length
* Verify 10x hot-plugs with 5ms delay per pin length (Very fast plug)
* Verify 10x hot-plugs with 0ms delay per pin length (Corner case - No pin delay)
Test Reporting:
* On screen log and PDF report option
* Pass/fail for each hot-plug event
No test running

图 4-28

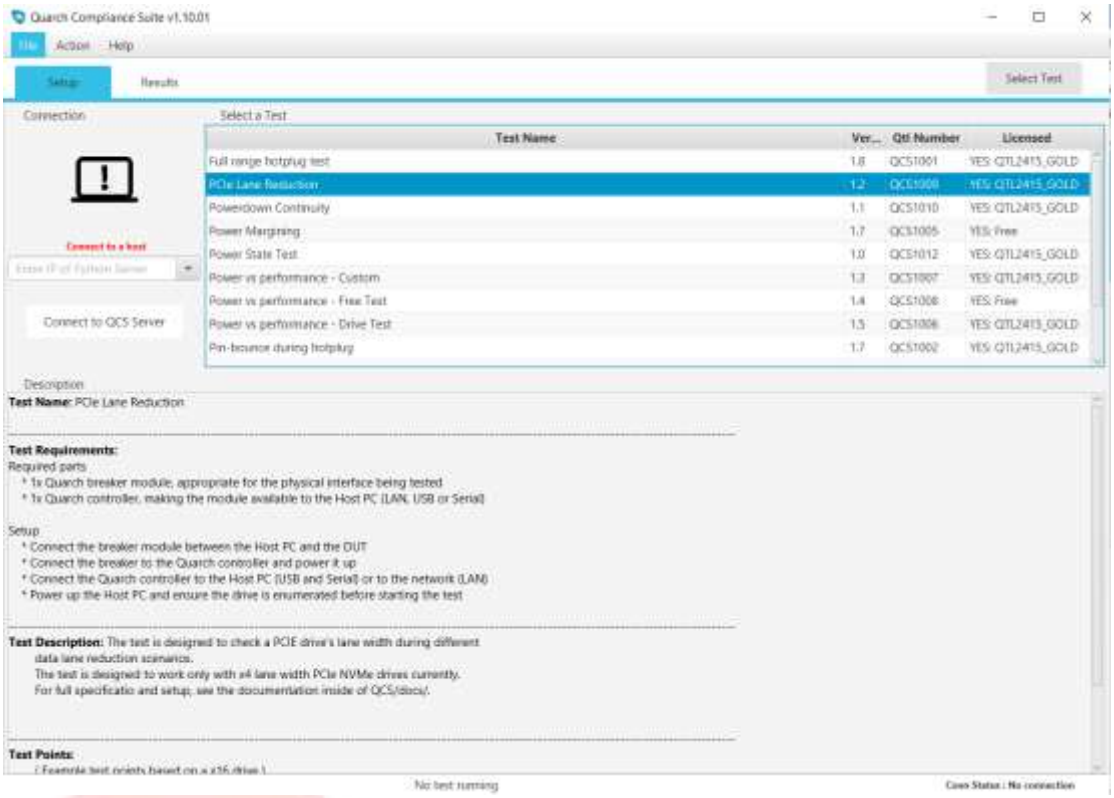

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Full range hotplug test
PCIe Lane Reduction
Powerdown Continuity
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Connect to QCS Server
Description
Test Name: PCIe Lane Reduction
Test Requirements:
Required parts
* 1x Quarch breaker module, appropriate for the physical interface being tested
* 1x Quarch controller, making the module available to the Host PC (LAN, USB or Serial)
Setup
* Connect the breaker module between the Host PC and the DUT
* Connect the breaker to the Quarch controller and power it up
* Connect the Quarch controller to the Host PC (USB and Serial) or to the network (LAN)
* Power up the Host PC and ensure the drive is enumerated before starting the test
Test Description: The test is designed to check a PCIE drive's lane width during different
data lane reduction scenarios.
The test is designed to work only with x4 lane width PCIe NVMe drives currently.
For full specificatio and setup, see the documentation inside of QCS/docs/.
Test Points:
No test running

图 4-29

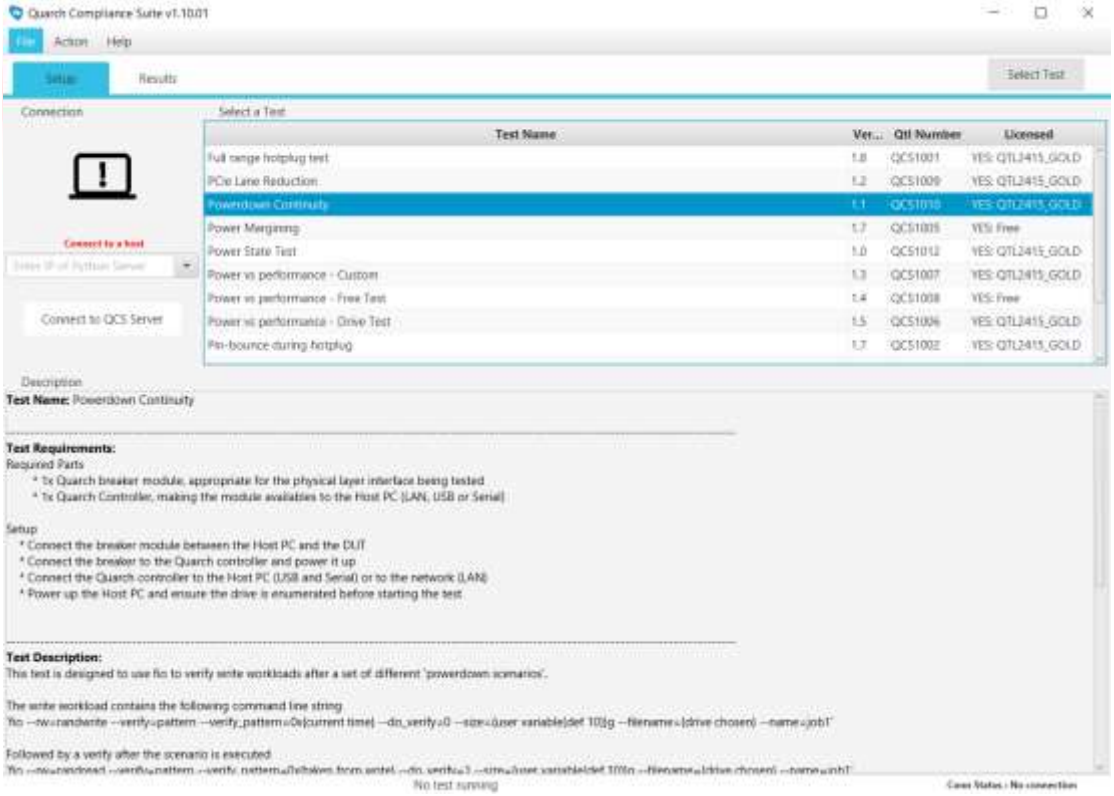

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Full range hotplug test
PCIe Lane Reduction
Powerdown Continuity
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Connect to QCS Server
Description
Test Name: Powerdown Continuity
Test Requirements:
Required Parts
* 1x Quarch breaker module, appropriate for the physical layer interface being tested
* 1x Quarch Controller, making the module available to the Host PC (LAN, USB or Serial)
Setup
* Connect the breaker module between the Host PC and the DUT
* Connect the breaker to the Quarch controller and power it up
* Connect the Quarch controller to the Host PC (USB and Serial) or to the network (LAN)
* Power up the Host PC and ensure the drive is enumerated before starting the test
Test Description:
This test is designed to use fio to verify write workloads after a set of different 'powerdown scenarios'.
The write workload contains the following command line string
fio --rw=randwrite --verify=pattern --verify_pattern=0x[current time] --do_verify=0 --size=[user variable|def 10]g --filename=[drive chosen] --name=job1
Followed by a verify after the scenario is executed
No test running

图 4-30

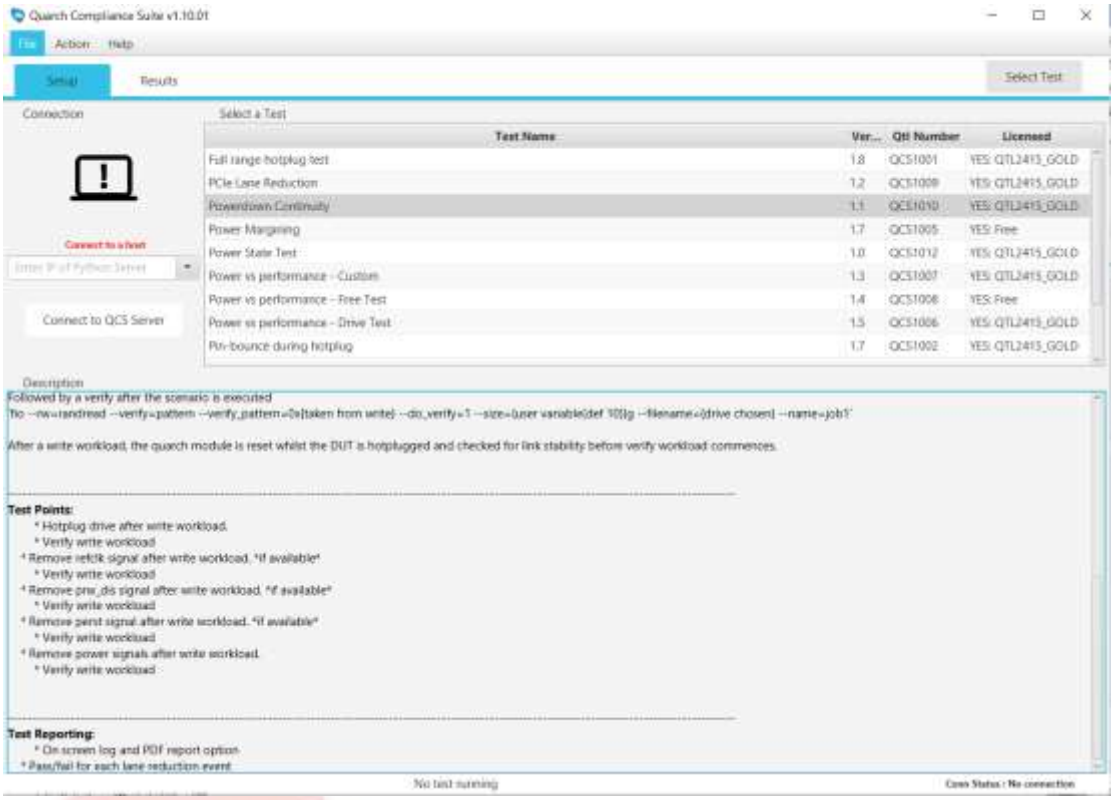

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Full range hotplug test
PCIe Lane Reduction
Powerdown Continuity
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Connect to a host
Connect to QCS Server
Description
Followed by a verify after the scenario is executed
No --rw=randread --verify=pattern --verify_pattern=0x[taken from write] --do_verify=1 --size=[user variable(def 10)]g --filename=[drive chosen] --name=job1'
After a write workload, the quarch module is reset whilst the DUT is hotplugged and checked for link stability before verify workload commences.
Test Points:
* Hotplug drive after write workload.
* Verify write workload
* Remove refclk signal after write workload. *If available*
* Verify write workload
* Remove pnw_dis signal after write workload. *if available*
* Verify write workload
* Remove perst signal after write workload. *If available*
* Verify write workload
* Remove power signals after write workload.
* Verify write workload
Test Reporting:
* On screen log and PDF report option
* Pass/fail for each lane reduction event
No test running

图 4-31

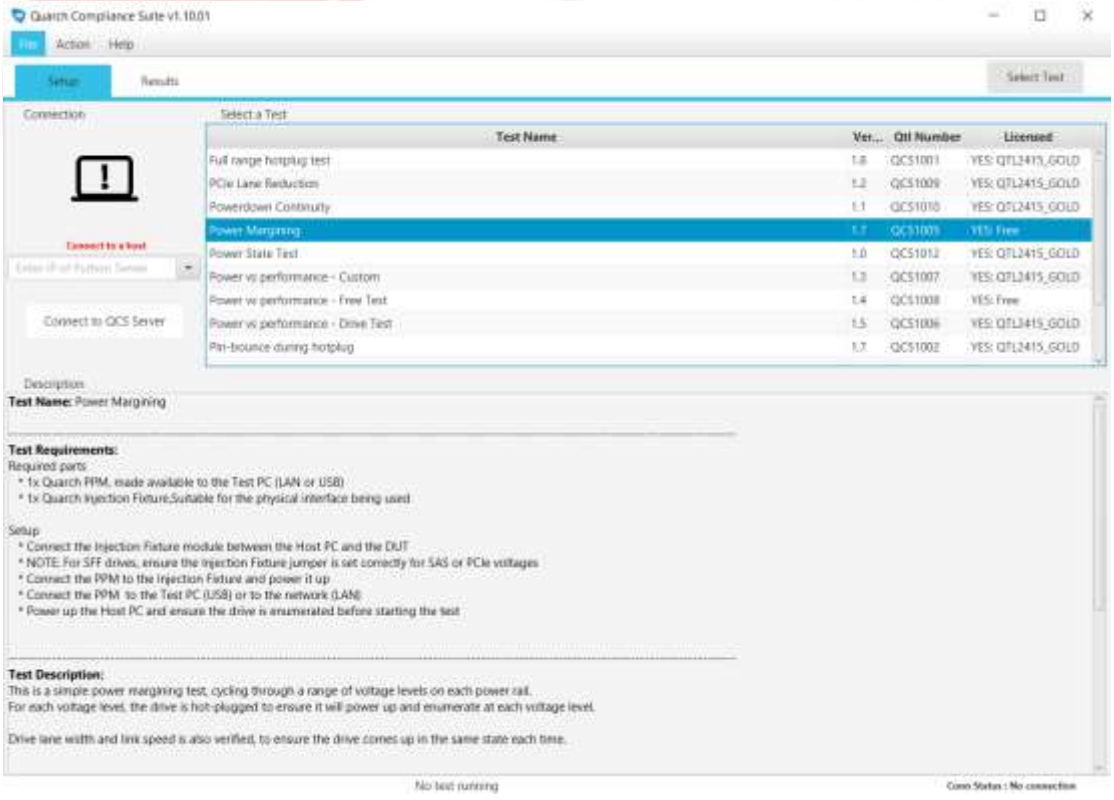

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Full range hotplug test
PCIe Lane Reduction
Powerdown Continuity
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Connect to a host
Connect to QCS Server
Description
Test Name: Power Margining
Test Requirements:
Required parts
* 1x Quarch PPM, made available to the Test PC (LAN or USB)
* 1x Quarch Injection Fixture,Suitable for the physical interface being used
Setup
* Connect the Injection Fixture module between the Host PC and the DUT
* NOTE: For SFF drives, ensure the Injection Fixture jumper is set correctly for SAS or PCIe voltages
* Connect the PPM to the Injection Fixture and power it up
* Connect the PPM to the Test PC (USB) or to the network (LAN)
* Power up the Host PC and ensure the drive is enumerated before starting the test
Test Description:
This is a simple power margining test, cycling through a range of voltage levels on each power rail.
For each voltage level, the drive is hot-plugged to ensure it will power up and enumerate at each voltage level.
Drive lane width and link speed is also verified, to ensure the drive comes up in the same state each time.
No test running

图 4-32

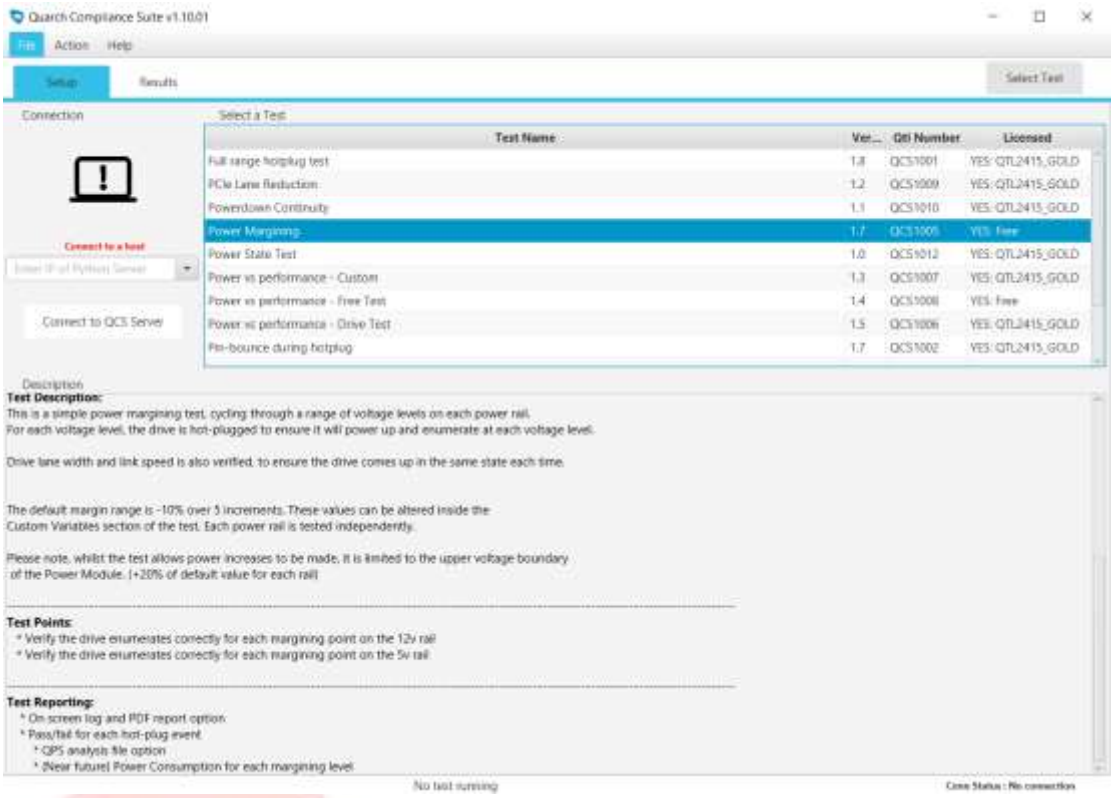

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qti Number
Licensed
Full range hotplug test
PCIe Lane Reduction
Powerdown Continuity
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Connect to QCS Server
Description
Test Description:
This is a simple power margining test, cycling through a range of voltage levels on each power rail.
For each voltage level, the drive is hot-plugged to ensure it will power up and enumerate at each voltage level.
Drive lane width and link speed is also verified, to ensure the drive comes up in the same state each time.
The default margin range is -10% over 5 increments. These values can be altered inside the
Custom Variables section of the test. Each power rail is tested independently.
Please note, whilst the test allows power increases to be made, it is limited to the upper voltage boundary
of the Power Module. (+20% of default value for each rail)
Test Points:
* Verify the drive enumerates correctly for each margining point on the 12v rail
* Verify the drive enumerates correctly for each margining point on the 5v rail
Test Reporting:
* On screen log and PDF report option
* Pass/fail for each hot-plug event
* QPS analysis file option
* (Near future) Power Consumption for each margining level
No test running

图 4-33

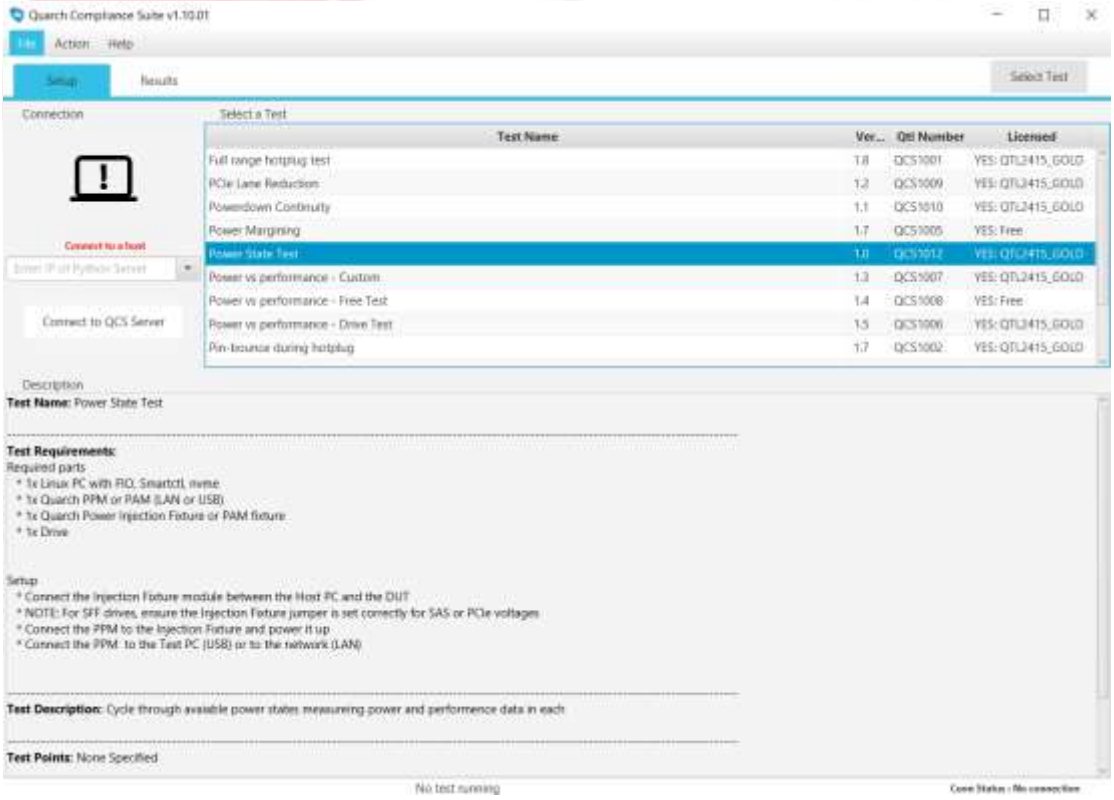

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qti Number
Licensed
Full range hotplug test
PCIe Lane Reduction
Powerdown Continuity
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Connect to QCS Server
Description
Test Name: Power State Test
Test Requirements:
Required parts
* 1x Linux PC with FIO, Smartctl, nvme
* 1x Quarch PPM or PAM (LAN or USB)
* 1x Quarch Power Injection Fixture or PAM fixture
* 1x Drive
Setup
* Connect the Injection Fixture module between the Host PC and the DUT
* NOTE: For SFF drives, ensure the Injection Fixture jumper is set correctly for SAS or PCIe voltages
* Connect the PPM to the Injection Fixture and power it up
* Connect the PPM to the Test PC (USB) or to the network (LAN)
Test Description: Cycle through available power states measuring power and performance data in each
Test Points: None Specified
No test running

图 4-34

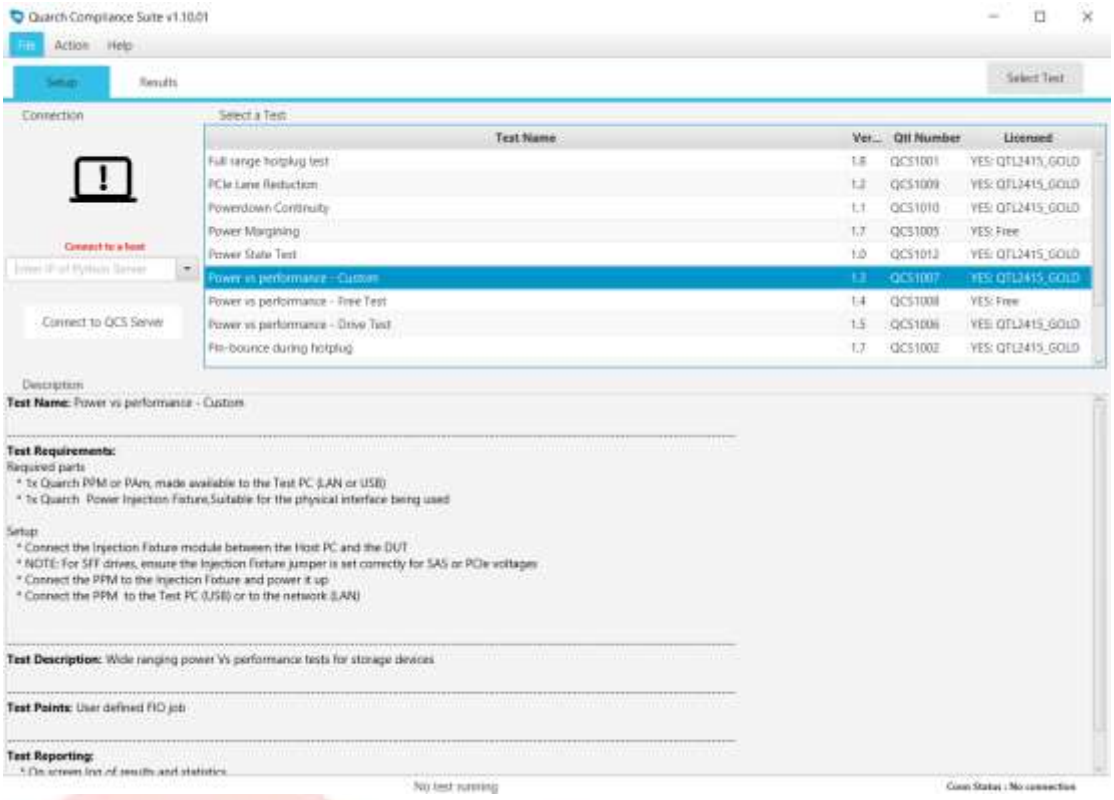

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Full range hotplug test 1.8 QCS1001 YES: QTL2415_GOLD
PCIe Lane Reduction 1.2 QCS1009 YES: QTL2415_GOLD
Powerdown Continuity 1.1 QCS1010 YES: QTL2415_GOLD
Power Margining 1.7 QCS1005 YES: Free
Power State Test 1.0 QCS1012 YES: QTL2415_GOLD
Power vs performance - Custom 1.3 QCS1007 YES: QTL2415_GOLD
Power vs performance - Free Test 1.4 QCS1008 YES: Free
Power vs performance - Drive Test 1.5 QCS1006 YES: QTL2415_GOLD
Re-bounce during hotplug 1.7 QCS1002 YES: QTL2415_GOLD
Connect to a host
Connect to QCS Server
Description
Test Name: Power vs performance - Custom
Test Requirements:
Required parts
* 1x Quarch PPM or PAm, made available to the Test PC (LAN or USB)
* 1x Quarch Power Injection Fixture,Suitable for the physical interface being used
Setup
* Connect the Injection Fixture module between the Host PC and the DUT
* NOTE: For SFF drives, ensure the Injection Fixture jumper is set correctly for SAS or PCIe voltages
* Connect the PPM to the Injection Fixture and power it up
* Connect the PPM to the Test PC (USB) or to the network (LAN)
Test Description: Wide ranging power Vs performance tests for storage devices
Test Points: User defined FIO job
Test Reporting:
No test running

图 4-35

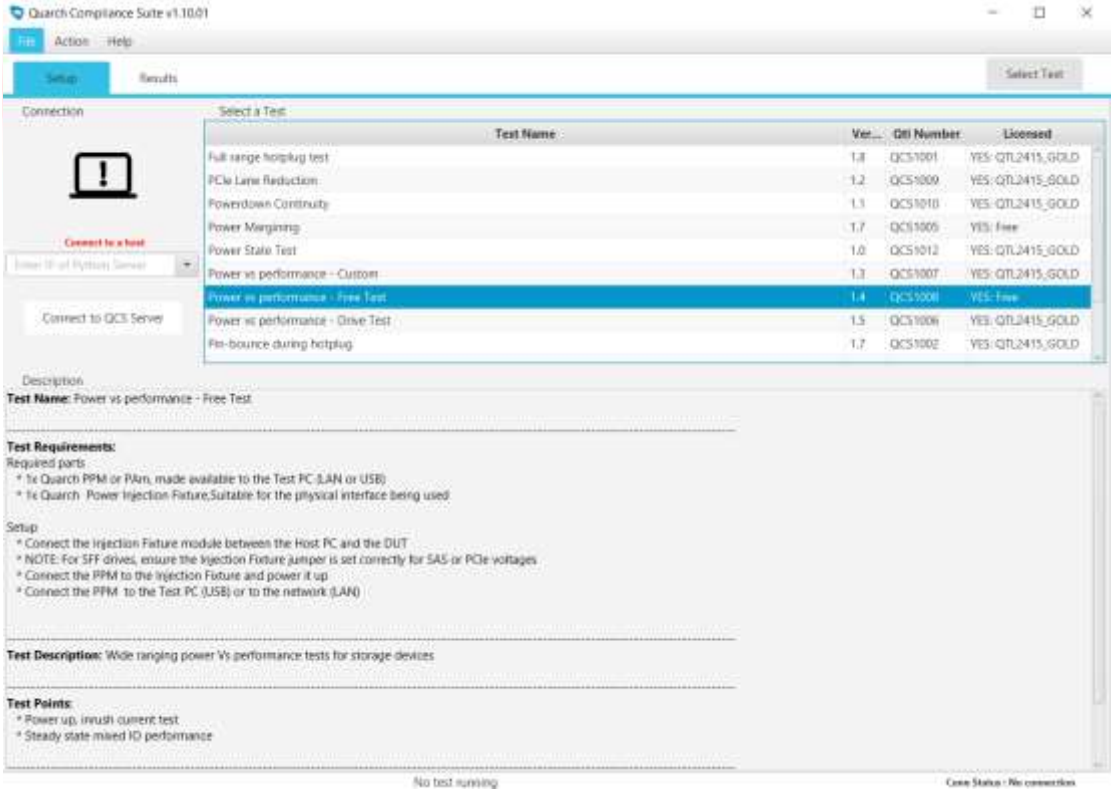

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Full range hotplug test 1.8 QCS1001 YES: QTL2415_GOLD
PCIe Lane Reduction 1.2 QCS1009 YES: QTL2415_GOLD
Powerdown Continuity 1.1 QCS1010 YES: QTL2415_GOLD
Power Margining 1.7 QCS1005 YES: Free
Power State Test 1.0 QCS1012 YES: QTL2415_GOLD
Power vs performance - Custom 1.3 QCS1007 YES: QTL2415_GOLD
Power vs performance - Free Test 1.4 QCS1008 YES: Free
Power vs performance - Drive Test 1.5 QCS1006 YES: QTL2415_GOLD
Re-bounce during hotplug 1.7 QCS1002 YES: QTL2415_GOLD
Connect to a host
Connect to QCS Server
Description
Test Name: Power vs performance - Free Test
Test Requirements:
Required parts
* 1x Quarch PPM or PAm, made available to the Test PC (LAN or USB)
* 1x Quarch Power Injection Fixture,Suitable for the physical interface being used
Setup
* Connect the Injection Fixture module between the Host PC and the DUT
* NOTE: For SFF drives, ensure the Injection Fixture jumper is set correctly for SAS or PCIe voltages
* Connect the PPM to the Injection Fixture and power it up
* Connect the PPM to the Test PC (USB) or to the network (LAN)
Test Description: Wide ranging power Vs performance tests for storage devices
Test Points:
* Power up, inrush current test
* Steady state mixed IO performance
No test running

图 4-36

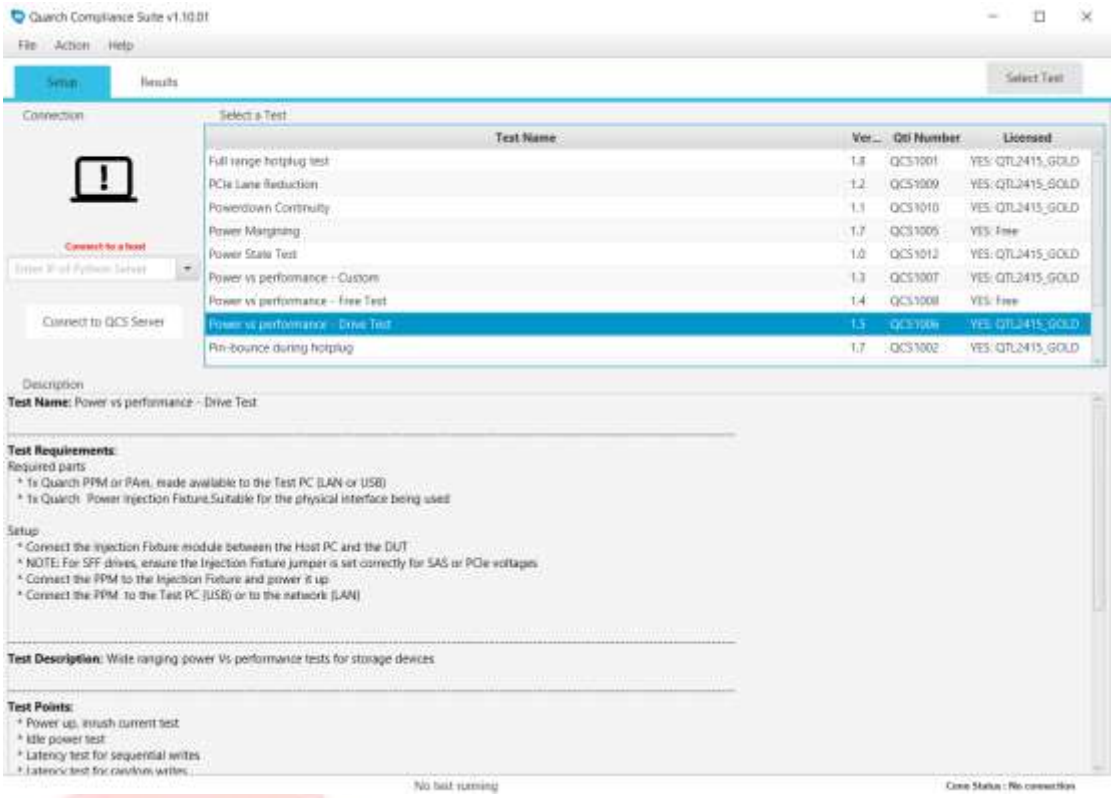

Quarch Compliance Suite v1.10.01
File Action Help
Setup Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Full range hotplug test 1.8 QCS1001 YES: QTL2415_GOLD
PCIe Lane Reduction 1.2 QCS1009 YES: QTL2415_GOLD
Powerdown Continuity 1.1 QCS1010 YES: QTL2415_GOLD
Power Margining 1.7 QCS1005 YES: Free
Power State Test 1.0 QCS1012 YES: QTL2415_GOLD
Power vs performance - Custom 1.3 QCS1007 YES: QTL2415_GOLD
Power vs performance - Free Test 1.4 QCS1008 YES: Free
Power vs performance - Drive Test 1.5 QCS1006 YES: QTL2415_GOLD
Pin-bounce during hotplug 1.7 QCS1002 YES: QTL2415_GOLD
Connect to QCS Server
Description
Test Name: Power vs performance - Drive Test
Test Requirements:
Required parts
* 1x Quarch PPM or PAm, made available to the Test PC (LAN or USB)
* 1x Quarch Power Injection Fixture,Suitable for the physical interface being used
Setup
* Connect the Injection Fixture module between the Host PC and the DUT
* NOTE: For SFF drives, ensure the Injection Fixture jumper is set correctly for SAS or PCIe voltages
* Connect the PPM to the Injection Fixture and power it up
* Connect the PPM to the Test PC (USB) or to the network (LAN)
Test Description: Wide ranging power Vs performance tests for storage devices
Test Points:
* Power up, inrush current test
* Idle power test
* Latency test for sequential writes
No test running

图 4-37

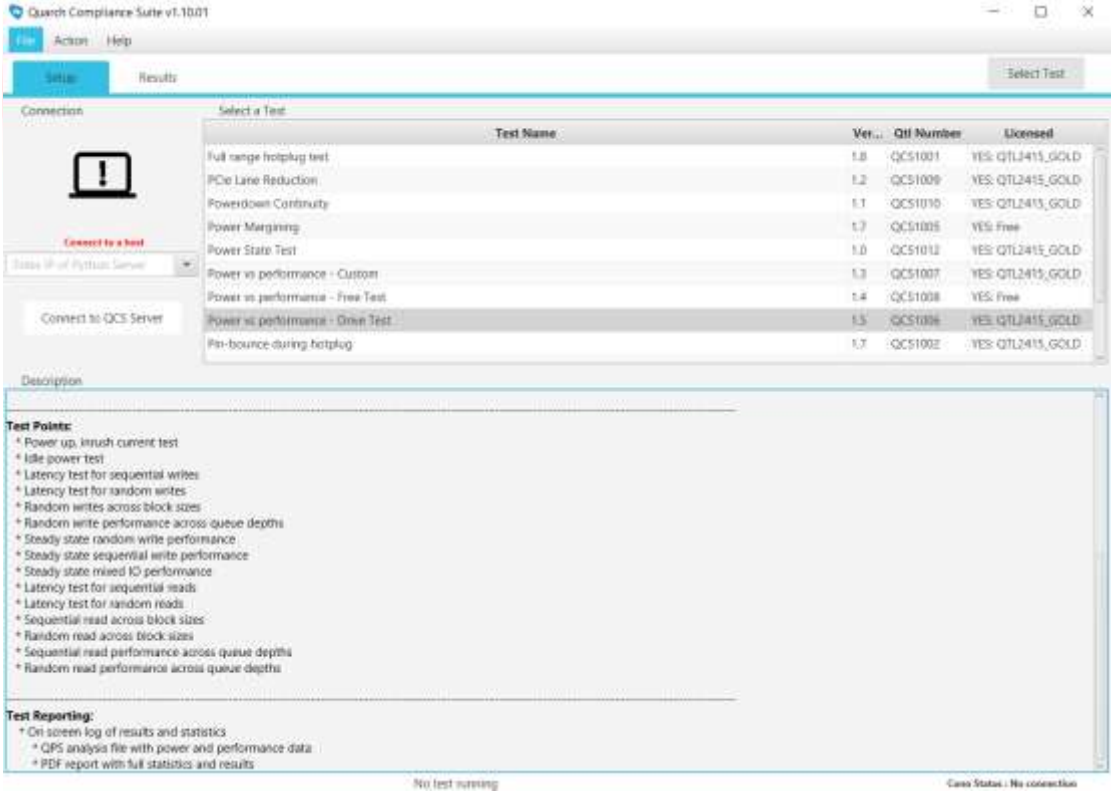

Quarch Compliance Suite v1.10.01
File Action Help
Setup Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Full range hotplug test 1.8 QCS1001 YES: QTL2415_GOLD
PCIe Lane Reduction 1.2 QCS1009 YES: QTL2415_GOLD
Powerdown Continuity 1.1 QCS1010 YES: QTL2415_GOLD
Power Margining 1.7 QCS1005 YES: Free
Power State Test 1.0 QCS1012 YES: QTL2415_GOLD
Power vs performance - Custom 1.3 QCS1007 YES: QTL2415_GOLD
Power vs performance - Free Test 1.4 QCS1008 YES: Free
Power vs performance - Drive Test 1.5 QCS1006 YES: QTL2415_GOLD
Pin-bounce during hotplug 1.7 QCS1002 YES: QTL2415_GOLD
Connect to QCS Server
Description
Test Points:
* Power up, inrush current test
* Idle power test
* Latency test for sequential writes
* Latency test for random writes
* Random writes across block sizes
* Random write performance across queue depths
* Steady state random write performance
* Steady state sequential write performance
* Steady state mixed IO performance
* Latency test for sequential reads
* Latency test for random reads
* Sequential read across block sizes
* Random read across block sizes
* Sequential read performance across queue depths
* Random read performance across queue depths
Test Reporting:
* On screen log of results and statistics
* QPS analysis file with power and performance data
* PDF report with full statistics and results
No test running

图 4-38

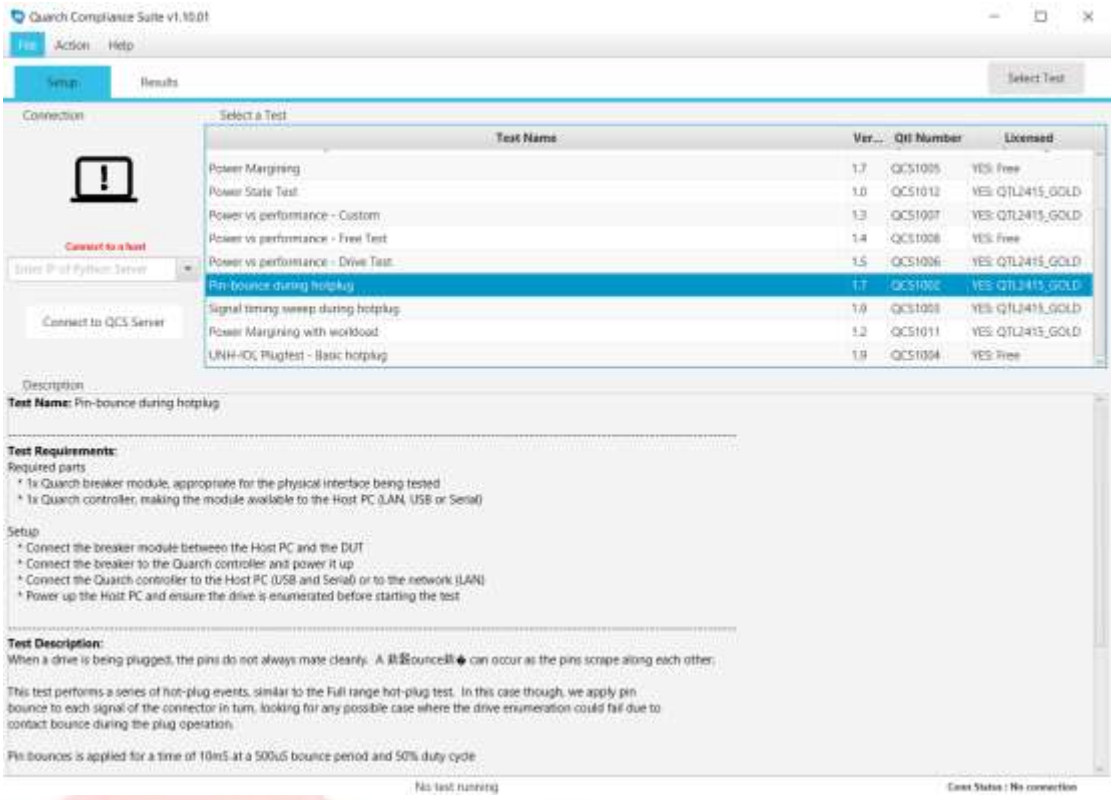

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Power Margining 1.7 QCS1005 YES: Free
Power State Test 1.0 QCS1012 YES: QTL2415_GOLD
Power vs performance - Custom 1.3 QCS1007 YES: QTL2415_GOLD
Power vs performance - Free Test 1.4 QCS1008 YES: Free
Power vs performance - Drive Test 1.5 QCS1006 YES: QTL2415_GOLD
Pin-bounce during hotplug 1.7 QCS1002 YES: QTL2415_GOLD
Signal timing sweep during hotplug 1.9 QCS1003 YES: QTL2415_GOLD
Power Margining with workload 1.2 QCS1011 YES: QTL2415_GOLD
UNH-IOL Plugfest - Basic hotplug 1.9 QCS1004 YES: Free
Connect to a host
Connect to QCS Server
Description
Test Name: Pin-bounce during hotplug
Test Requirements:
Required parts
* 1x Quarch breaker module, appropriate for the physical interface being tested
* 1x Quarch controller, making the module available to the Host PC (LAN, USB or Serial)
Setup
* Connect the breaker module between the Host PC and the DUT
* Connect the breaker to the Quarch controller and power it up
* Connect the Quarch controller to the Host PC (USB and Serial) or to the network (LAN)
* Power up the Host PC and ensure the drive is enumerated before starting the test
Test Description:
When a drive is being plugged, the pins do not always mate cleanly. A 跳跳ounce跳跳 can occur as the pins scrape along each other.
This test performs a series of hot-plug events, similar to the Full range hot-plug test. In this case though, we apply pin bounce to each signal of the connector in turn, looking for any possible case where the drive enumeration could fail due to contact bounce during the plug operation.
Pin bounces is applied for a time of 10mS at a 500uS bounce period and 50% duty cycle
No test running
Conn Status : No connection

图 4-39

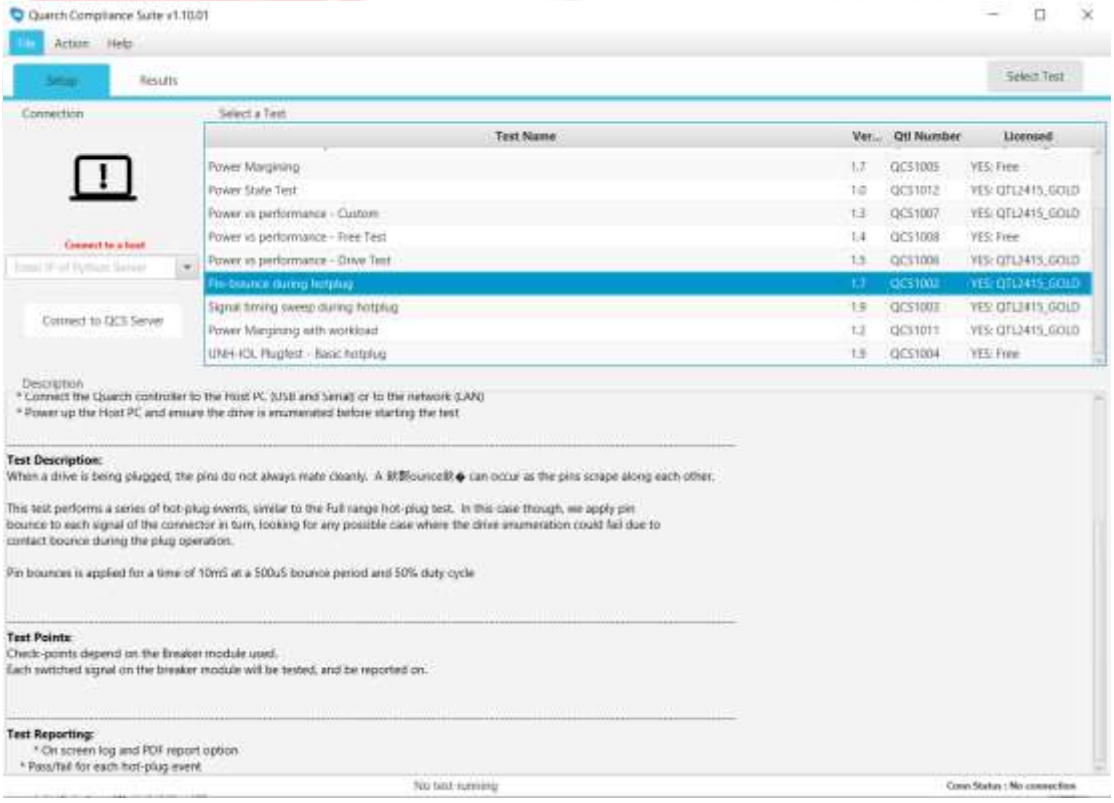

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Power Margining 1.7 QCS1005 YES: Free
Power State Test 1.0 QCS1012 YES: QTL2415_GOLD
Power vs performance - Custom 1.3 QCS1007 YES: QTL2415_GOLD
Power vs performance - Free Test 1.4 QCS1008 YES: Free
Power vs performance - Drive Test 1.5 QCS1006 YES: QTL2415_GOLD
Pin-bounce during hotplug 1.7 QCS1002 YES: QTL2415_GOLD
Signal timing sweep during hotplug 1.9 QCS1003 YES: QTL2415_GOLD
Power Margining with workload 1.2 QCS1011 YES: QTL2415_GOLD
UNH-IOL Plugfest - Basic hotplug 1.9 QCS1004 YES: Free
Connect to a host
Connect to QCS Server
Description
* Connect the Quarch controller to the Host PC (USB and Serial) or to the network (LAN)
* Power up the Host PC and ensure the drive is enumerated before starting the test
Test Description:
When a drive is being plugged, the pins do not always mate cleanly. A 跳跳ounce跳跳 can occur as the pins scrape along each other.
This test performs a series of hot-plug events, similar to the Full range hot-plug test. In this case though, we apply pin bounce to each signal of the connector in turn, looking for any possible case where the drive enumeration could fail due to contact bounce during the plug operation.
Pin bounces is applied for a time of 10mS at a 500uS bounce period and 50% duty cycle
Test Points:
Check-points depend on the Breaker module used.
Each switched signal on the breaker module will be tested, and be reported on.
Test Reporting:
* On screen log and PDF report option
* Pass/fail for each hot-plug event
No test running
Conn Status : No connection

图 4-40

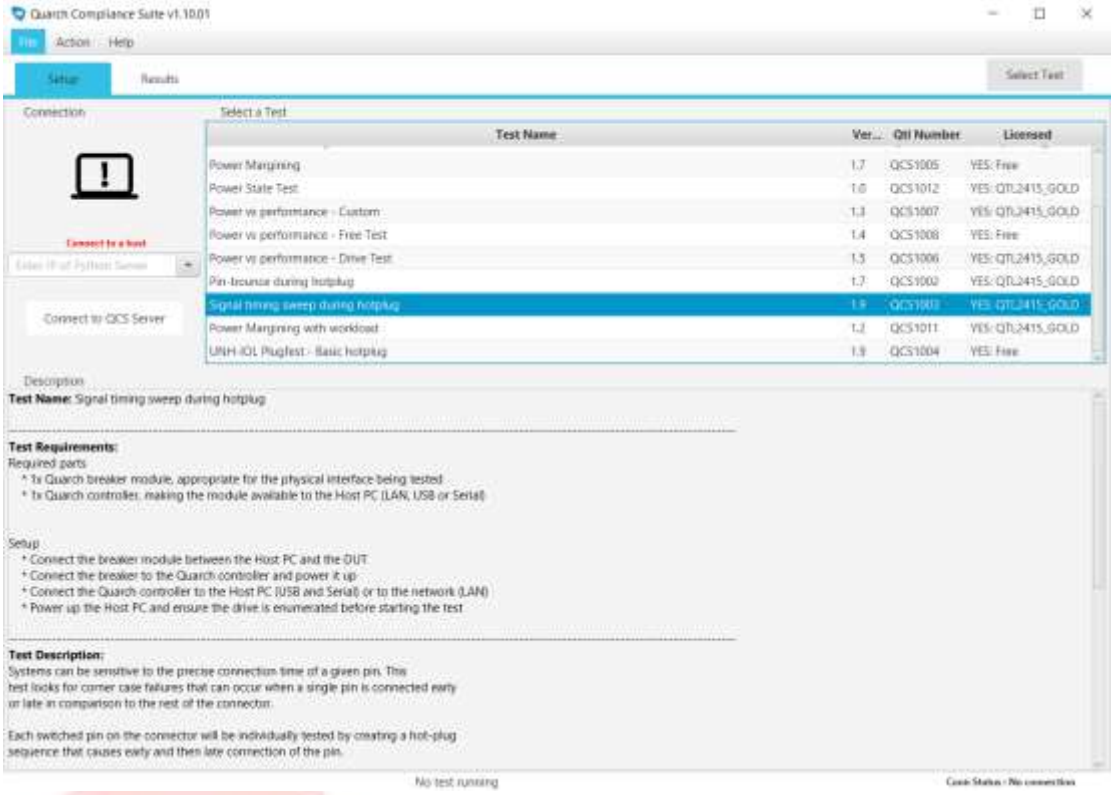

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Signal timing sweep during hotplug
Power Margining with workload
UNH-IOL Plugfest - Basic hotplug
Connect to QCS Server
Description
Test Name: Signal timing sweep during hotplug
Test Requirements:
Required parts
* 1x Quarch breaker module, appropriate for the physical interface being tested
* 1x Quarch controller, making the module available to the Host PC (LAN, USB or Serial)
Setup
* Connect the breaker module between the Host PC and the DUT
* Connect the breaker to the Quarch controller and power it up
* Connect the Quarch controller to the Host PC (USB and Serial) or to the network (LAN)
* Power up the Host PC and ensure the drive is enumerated before starting the test
Test Description:
Systems can be sensitive to the precise connection time of a given pin. This
test looks for corner case failures that can occur when a single pin is connected early
or late in comparison to the rest of the connector.
Each switched pin on the connector will be individually tested by creating a hot-plug
sequence that causes early and then late connection of the pin.
No test running

图 4-41

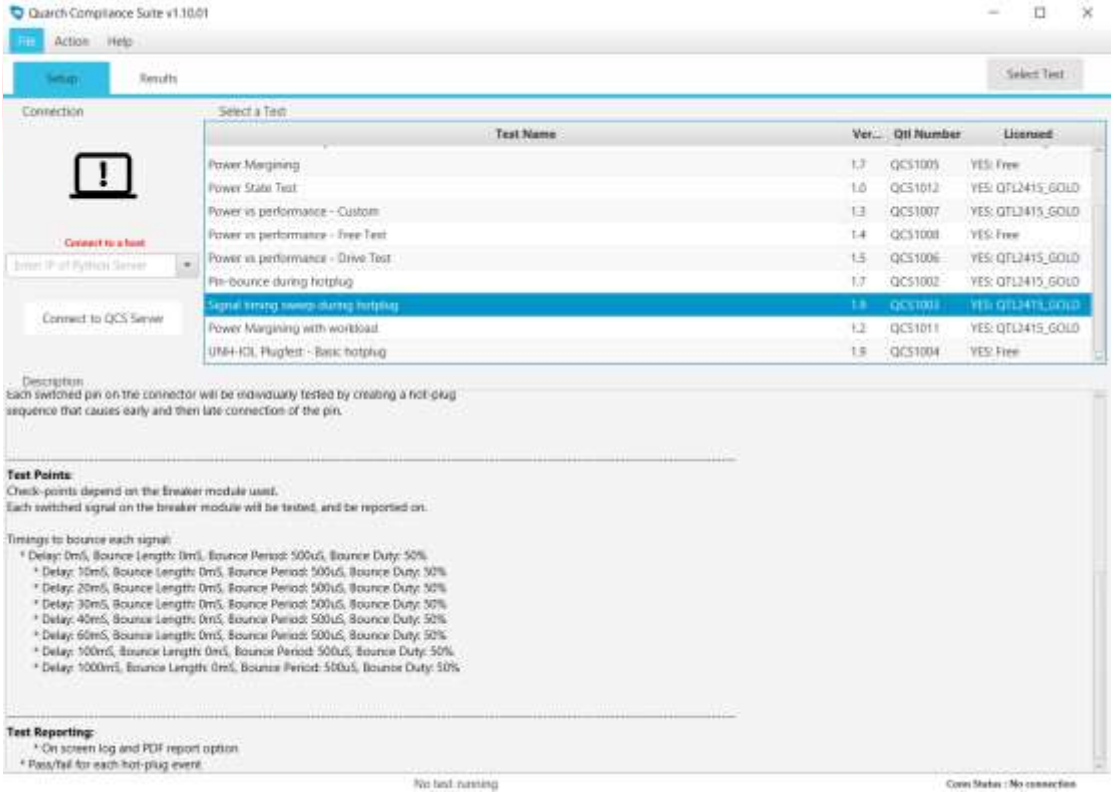

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Signal timing sweep during hotplug
Power Margining with workload
UNH-IOL Plugfest - Basic hotplug
Connect to QCS Server
Description
each switched pin on the connector will be individually tested by creating a hot-plug
sequence that causes early and then late connection of the pin.
Test Points:
Check-points depend on the Breaker module used.
Each switched signal on the breaker module will be tested, and be reported on.
Timings to bounce each signal:
* Delay: 0mS, Bounce Length: 0mS, Bounce Period: 500uS, Bounce Duty: 50%
* Delay: 10mS, Bounce Length: 0mS, Bounce Period: 500uS, Bounce Duty: 50%
* Delay: 20mS, Bounce Length: 0mS, Bounce Period: 500uS, Bounce Duty: 50%
* Delay: 30mS, Bounce Length: 0mS, Bounce Period: 500uS, Bounce Duty: 50%
* Delay: 40mS, Bounce Length: 0mS, Bounce Period: 500uS, Bounce Duty: 50%
* Delay: 60mS, Bounce Length: 0mS, Bounce Period: 500uS, Bounce Duty: 50%
* Delay: 100mS, Bounce Length: 0mS, Bounce Period: 500uS, Bounce Duty: 50%
* Delay: 1000mS, Bounce Length: 0mS, Bounce Period: 500uS, Bounce Duty: 50%
Test Reporting:
* On screen log and PDF report option
* Pass/fail for each hot-plug event
No test running

图 4-42

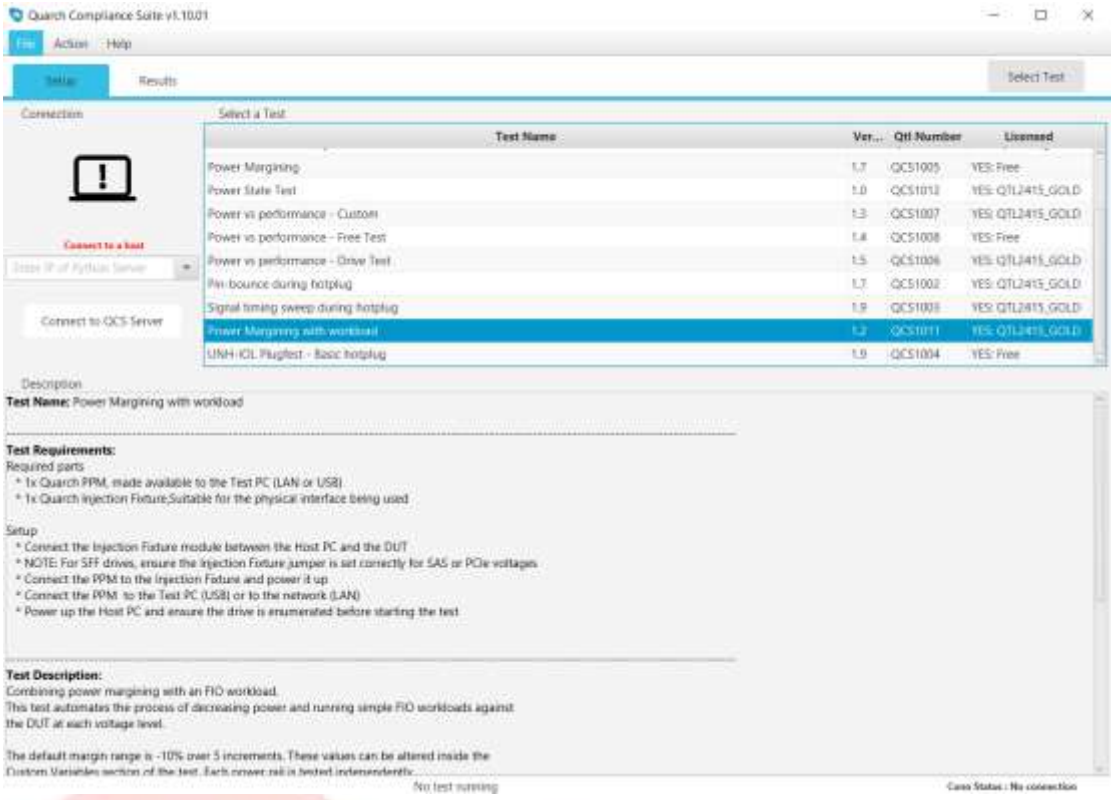

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Signal timing sweep during hotplug
Power Margining with workload
UNH-IOL Plugfest - Basic hotplug
Connect to QCS Server
Description
Test Name: Power Margining with workload
Test Requirements:
Required parts
* 1x Quarch PPM, made available to the Test PC (LAN or USB)
* 1x Quarch Injection Fixture,Suitable for the physical interface being used
Setup
* Connect the Injection Fixture module between the Host PC and the DUT
* NOTE: For SFF drives, ensure the injection Fixture jumper is set correctly for SAS or PCIe voltages
* Connect the PPM to the Injection Fixture and power it up
* Connect the PPM to the Test PC (USB) or to the network (LAN)
* Power up the Host PC and ensure the drive is enumerated before starting the test
Test Description:
Combining power margining with an FIO workload.
This test automates the process of decreasing power and running simple FIO workloads against
the DUT at each voltage level.
The default margin range is -10% over 5 increments. These values can be altered inside the
No test running

图 4-43

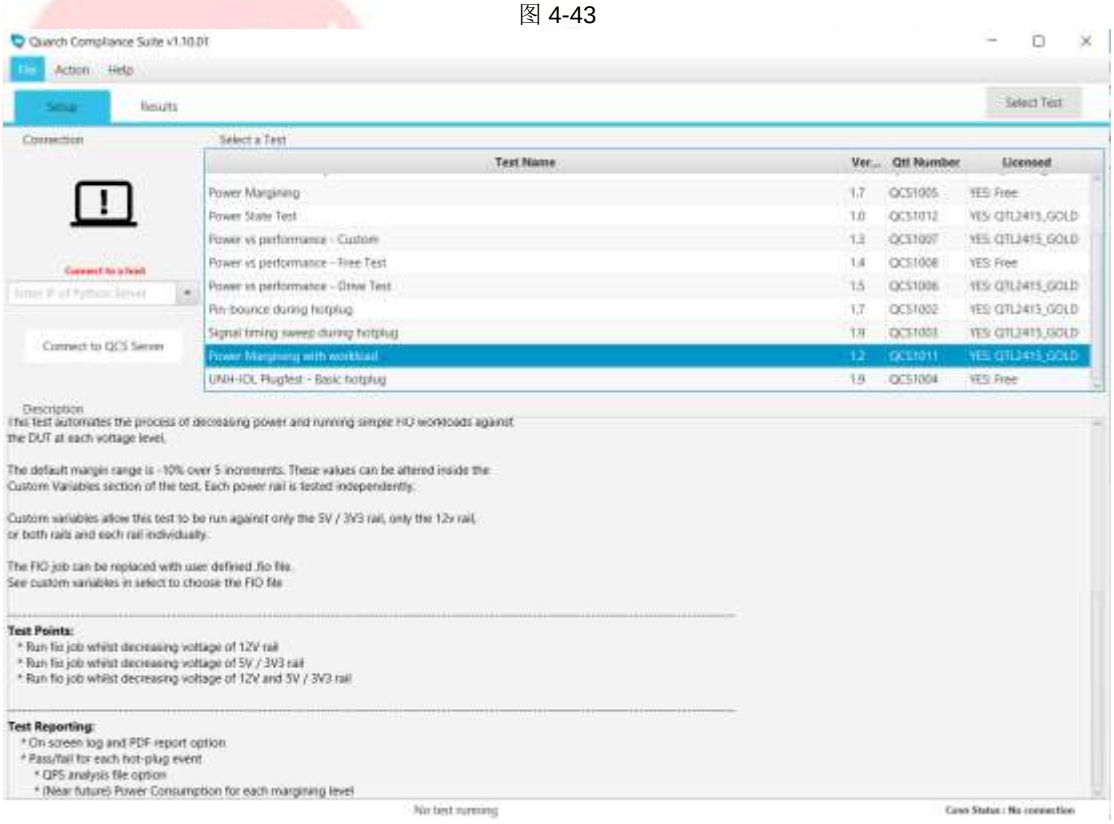

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qtl Number
Licensed
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Signal timing sweep during hotplug
Power Margining with workload
UNH-IOL Plugfest - Basic hotplug
Connect to QCS Server
Description
This test automates the process of decreasing power and running simple FIO workloads against
the DUT at each voltage level.
The default margin range is -10% over 5 increments. These values can be altered inside the
Custom Variables section of the test. Each power rail is tested independently.
Custom variables allow this test to be run against only the 5V / 3V3 rail, only the 12v rail,
or both rails and each rail individually.
The FIO job can be replaced with user defined .fio file.
See custom variables in select to choose the FIO file
Test Points:
* Run fio job whilst decreasing voltage of 12V rail
* Run fio job whilst decreasing voltage of 5V / 3V3 rail
* Run fio job whilst decreasing voltage of 12V and 5V / 3V3 rail
Test Reporting:
* On screen log and PDF report option
* Pass/fail for each hot-plug event
* QPS analysis file option
* (Near future) Power Consumption for each margining level
No test running

图 4-44

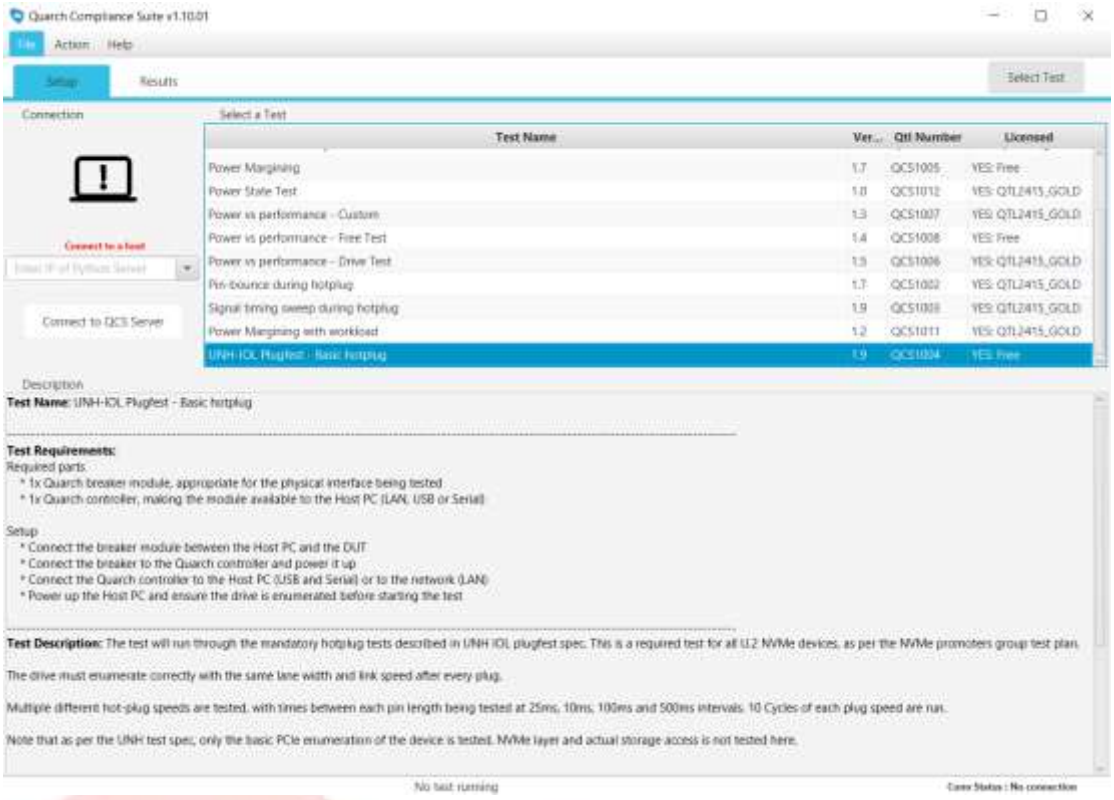

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qti Number
Licensed
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Signal timing sweep during hotplug
Power Margining with workload
UNH-IOL Plugfest - Basic hotplug
Connect to a host
Connect to QCS Server
Description
Test Name: UNH-IOL Plugfest - Basic hotplug
Test Requirements:
Required parts
* 1x Quarch breaker module, appropriate for the physical interface being tested
* 1x Quarch controller, making the module available to the Host PC (LAN, USB or Serial)
Setup
* Connect the breaker module between the Host PC and the DUT
* Connect the breaker to the Quarch controller and power it up
* Connect the Quarch controller to the Host PC (USB and Serial) or to the network (LAN)
* Power up the Host PC and ensure the drive is enumerated before starting the test
Test Description: The test will run through the mandatory hotplug tests described in UNH IOL plugfest spec. This is a required test for all U.2 NVMe devices, as per the NVMe promoters group test plan.
The drive must enumerate correctly with the same lane width and link speed after every plug.
Multiple different hot-plug speeds are tested, with times between each pin length being tested at 25ms, 10ms, 100ms and 500ms intervals. 10 Cycles of each plug speed are run.
Note that as per the UNH test spec, only the basic PCIe enumeration of the device is tested. NVMe layer and actual storage access is not tested here.
No test running

图 4-45

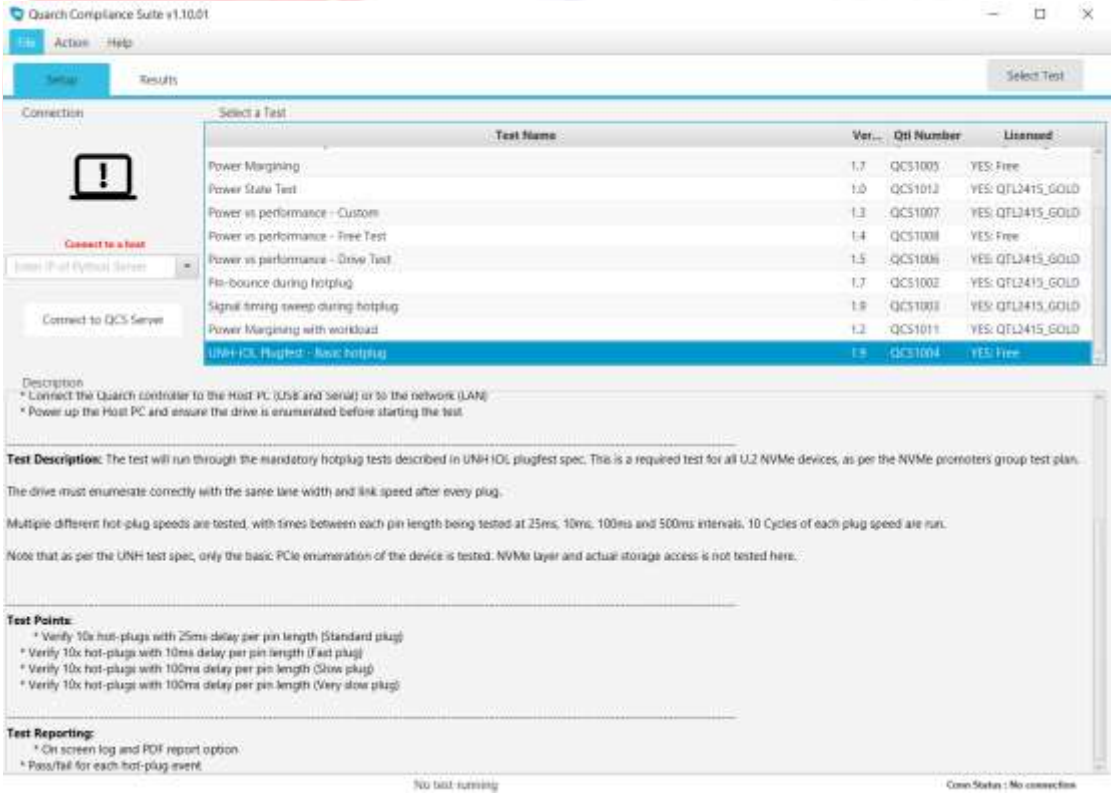

Quarch Compliance Suite v1.10.01
File
Action
Help
Setup
Results
Select Test
Connection
Select a Test
Test Name
Ver...
Qti Number
Licensed
Power Margining
Power State Test
Power vs performance - Custom
Power vs performance - Free Test
Power vs performance - Drive Test
Pin-bounce during hotplug
Signal timing sweep during hotplug
Power Margining with workload
UNH-IOL Plugfest - Basic hotplug
Connect to a host
Connect to QCS Server
Description
* Connect the Quarch controller to the Host PC (USB and Serial) or to the network (LAN)
* Power up the Host PC and ensure the drive is enumerated before starting the test
Test Description: The test will run through the mandatory hotplug tests described in UNH IOL plugfest spec. This is a required test for all U.2 NVMe devices, as per the NVMe promoters group test plan.
The drive must enumerate correctly with the same lane width and link speed after every plug.
Multiple different hot-plug speeds are tested, with times between each pin length being tested at 25ms, 10ms, 100ms and 500ms intervals. 10 Cycles of each plug speed are run.
Note that as per the UNH test spec, only the basic PCIe enumeration of the device is tested. NVMe layer and actual storage access is not tested here.
Test Points:
* Verify 10x hot-plugs with 25ms delay per pin length (Standard plug)
* Verify 10x hot-plugs with 10ms delay per pin length (Fast plug)
* Verify 10x hot-plugs with 100ms delay per pin length (Slow plug)
* Verify 10x hot-plugs with 100ms delay per pin length (Very slow plug)
Test Reporting:
* On screen log and PDF report option
* Pass/fail for each hot-plug event
No test running

图 4-46

4.1.5.2 如何获得 QCS license 许可证？

使用 QCS 软件需要获得 license 许可，并且锁定在运行该 suite 客户端 (Java GUI) 端的 PC 硬件上。这意味着您首先需要确定一台将控制测试的单台 PC。

请按照以下步骤操作：

- **1. 转到帮助->设置->许可，或单击测试上的“升级”按钮**

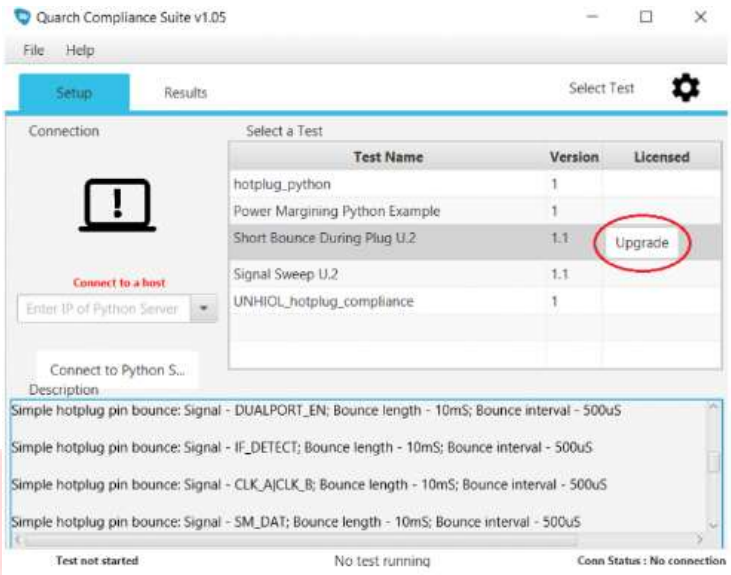


图 4-47

- **2. 您需要向我们提供硬件 ID 文件。首先，单击“生成”按钮。**

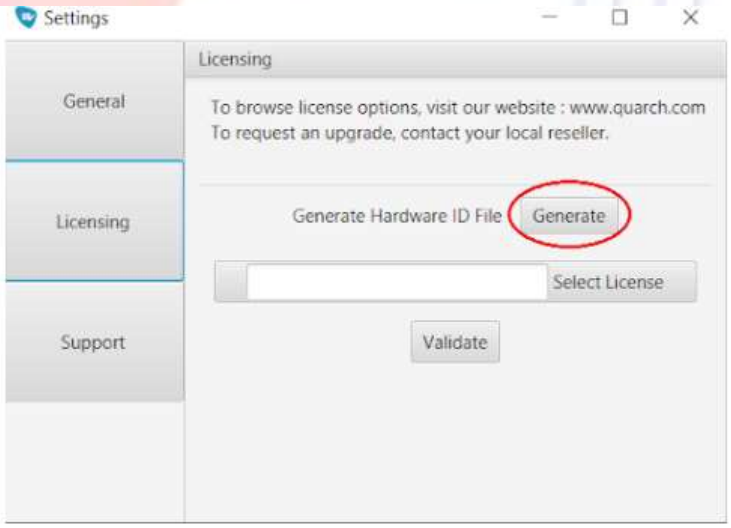


图 4-48

- **3. 在您的 QCS 启动目录中，找到您的硬件 ID 文件。**
- **4. 为了让我们生成您的许可证，请发送电子邮件至 sales@saniffer.com 或联系 Quarch 合作伙伴 Saniffer (021-50807071)，并包括以下内容：**
  - 您上面生成的硬件 ID 文件作为附件

- 所需的许可证类型，由于该 license 是 annual subscription 年度预定形式，所以默认是一年，也可以参考你们的购买合同选择 3 年
- 本许可证将绑定到的公司名称、详细地址和电子邮件地址。

- **5. 等待我们的电子邮件回复：这将包含您的许可证文件。**
- **6. 获得此许可证文件后，重新打开 QCS 并转到帮助 > 支持 > 许可。**

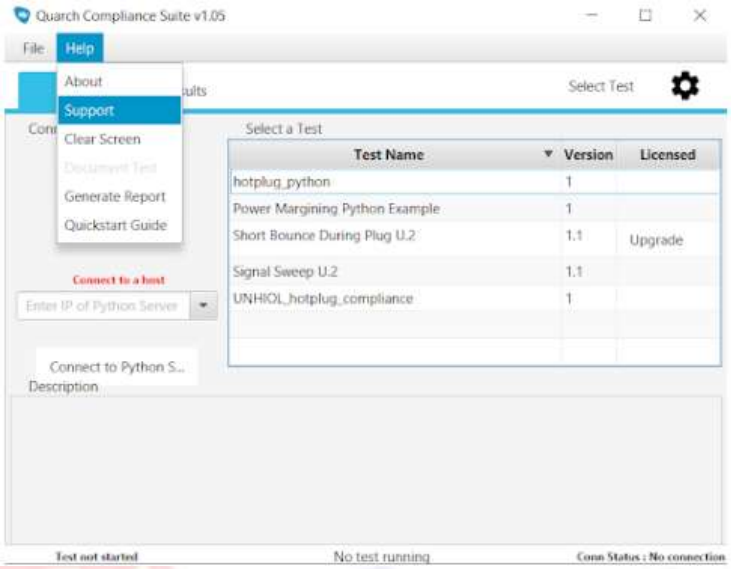


图 4-49

- **7. 要打开许可证文件，请单击“选择新许可证”**

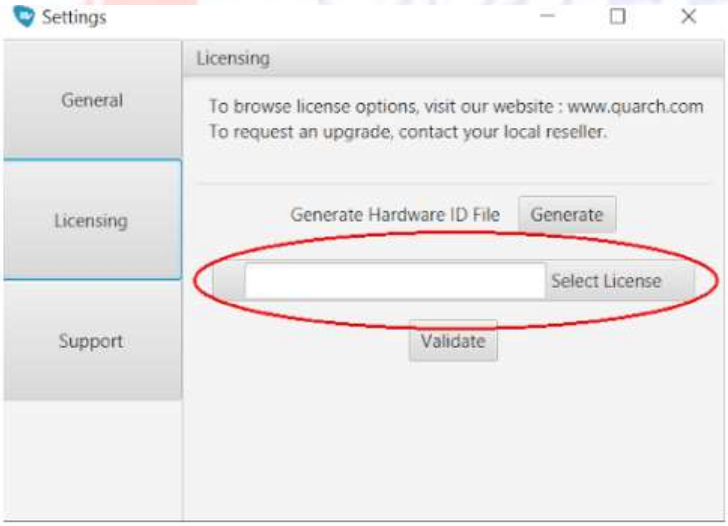


图 4-50

点击“验证”按钮。

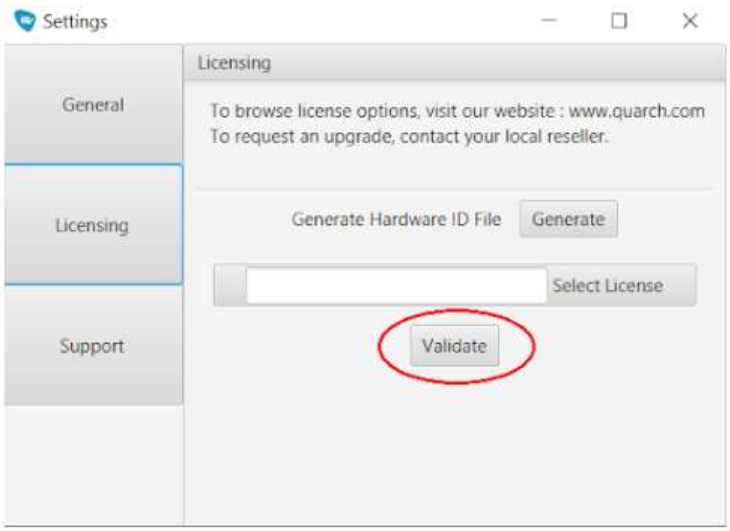


图 4-51

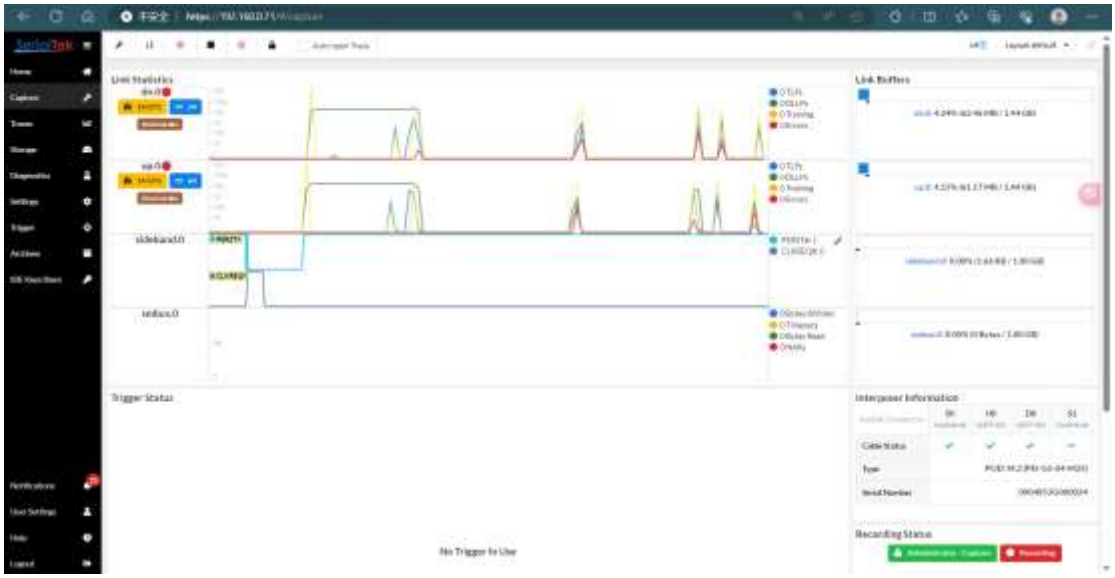
如果验证成功，应该会提示您。如果是这样，重新启动 QCS，现在可以使用新的许可证。

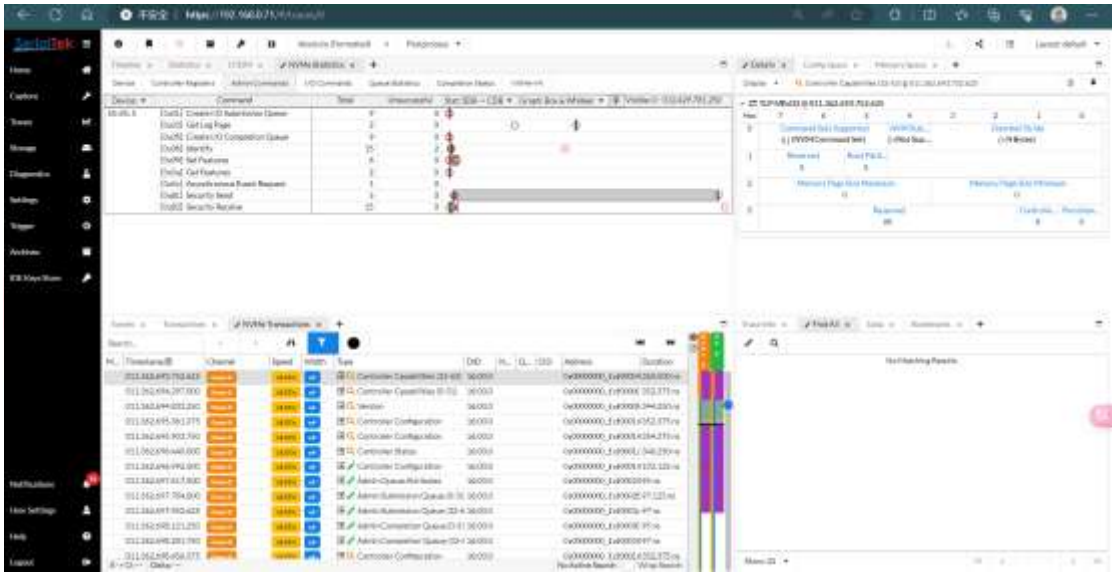
4.1.6 通过 Quarch M.2 或者插卡类控制模块实现针对 M.2 SSD 和各类插卡的热插拔自动化测试

对于不支持热插拔功能的 PCIe 设备，研发测试中，可能需要频繁热插拔设备，但是我们往往不愿意重启操作系统，故需要对设备(本文中为 m.2 接口 NVMe 硬盘)进行 rescan 以测试。

下面的示例为采用 Quarch QTL1260 管理模块 + M.2 card control module，配合 serialte PCIe analyzer 更直观地显示该测试过程。

1. 开机启动主机，M.2 SSD 初始化正常，通过分析仪可以看到正确解码，参见下图

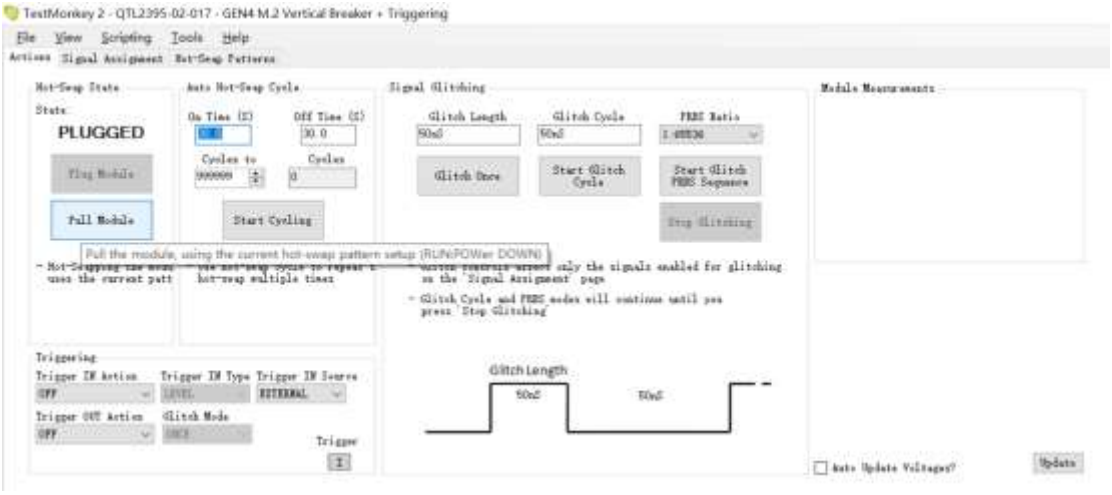


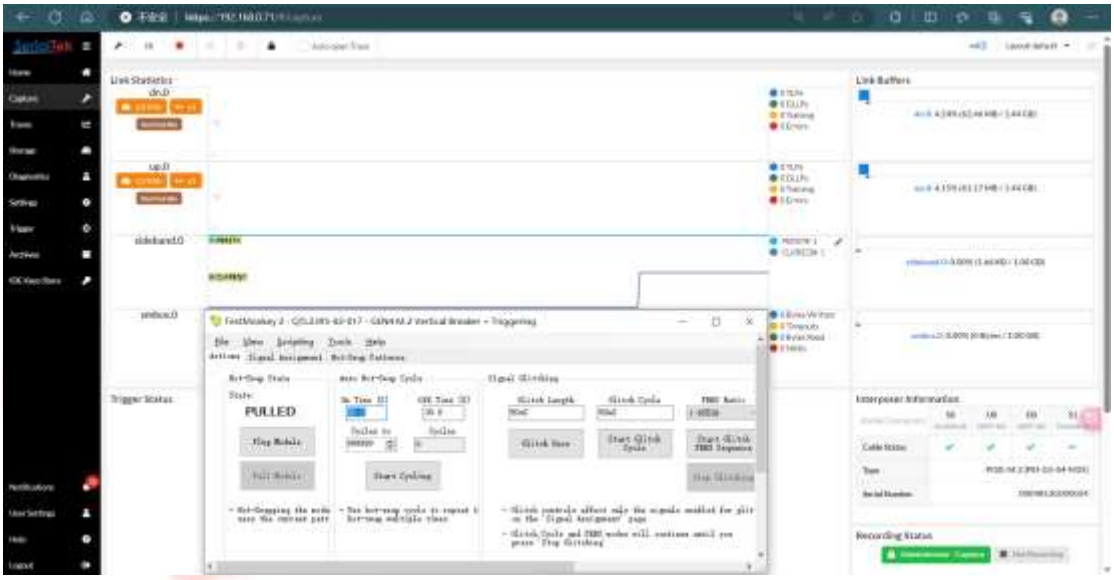


可以找到 PCIe device,可以显示 NVMe SSD

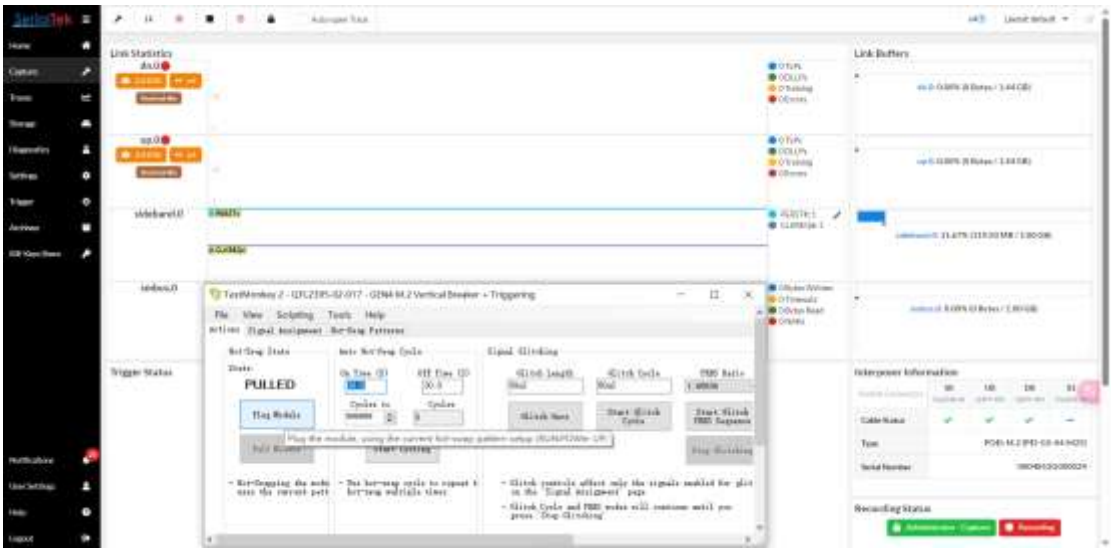
```
[root@localhost admin]# lspci | grep NVMe
16:00.0 Non-Volatile memory controller: Phison Electronics Corporation E16 PCIe4 NVMe Controller (rev 01)
[root@localhost admin]# nvme list
Node                  Generic               SN                   Model                                    Namespace
Usage                      Format           FW Rev
--------------------- --------------------- -------------------- ---------------------------------------- ----------
-------------------------- ----------------- --------
/dev/nvme0n1          /dev/ng0n1            SN20180970903        GIGABYTE GP-AG4500G                      0x1
500.11  GB / 500.11  GB    512   B +  0 B   EGFM11.2
[root@localhost admin]#
```

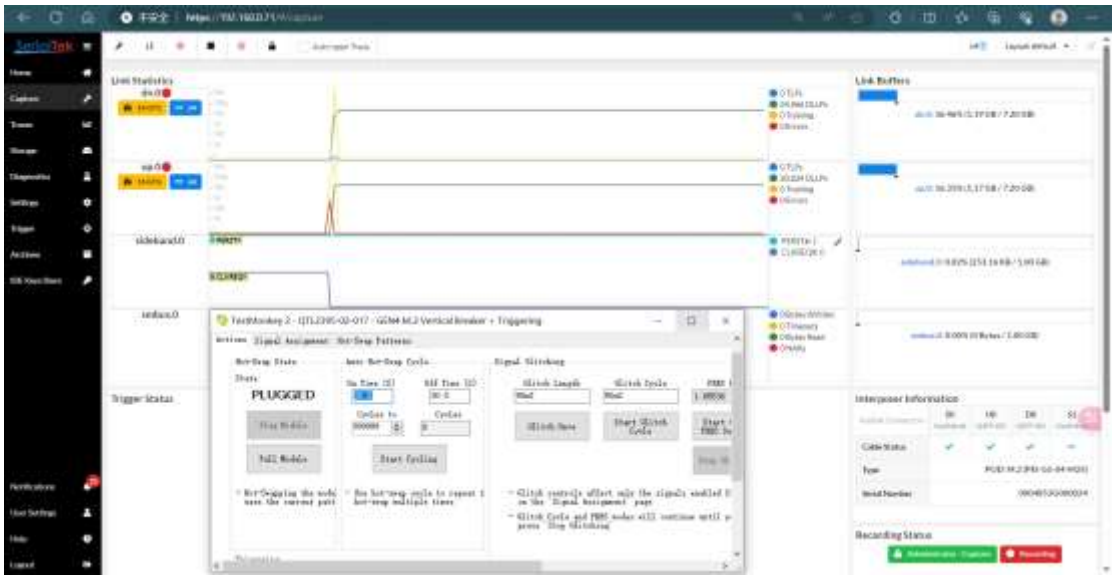
2. 通过 quarch m.2 drive control module 的操作界面 test monkey 拔出 m.2 ssd，参见下图。





3. 通过 quarch test monkey 软件插入 m.2 ssd，参见下图。





注意：在拔出 nvme 设备后，通常设备 pcie 信息仍然保存在系统中，参见下面的 demesg 内核信息，lspci 设备信息截图。

```
[root@localhost admin]# dmesg | grep nvme
[    0.000000] Command line: BOOT_IMAGE=(hd0,gpt2)/vmlinuz-5.14.0-362.el9.x86_64 root=/dev/mapper/cs-root ro crashker
nel=1G-4G:192M,4G-64G:256M,64G-:512M resume=/dev/mapper/cs-swap rd.lvm.lv=cs/root rd.lvm.lv=cs/swap rhgb quiet nvme_c
ore.multipath=N
[    0.008485] Kernel command line: BOOT_IMAGE=(hd0,gpt2)/vmlinuz-5.14.0-362.el9.x86_64 root=/dev/mapper/cs-root ro c
rashkernel=1G-4G:192M,4G-64G:256M,64G-:512M resume=/dev/mapper/cs-swap rd.lvm.lv=cs/root rd.lvm.lv=cs/swap rhgb quiet
 nvme_core.multipath=N
[    6.571920] nvme nvme0: pci function 0000:16:00.0
[    6.576469] nvme nvme0: Shutdown timeout set to 10 seconds
[    6.577233] nvme nvme0: 8/0/0 default/read/poll queues
[    6.577933] nvme nvme0: Ignoring bogus Namespace Identifiers
[root@localhost admin]#
```

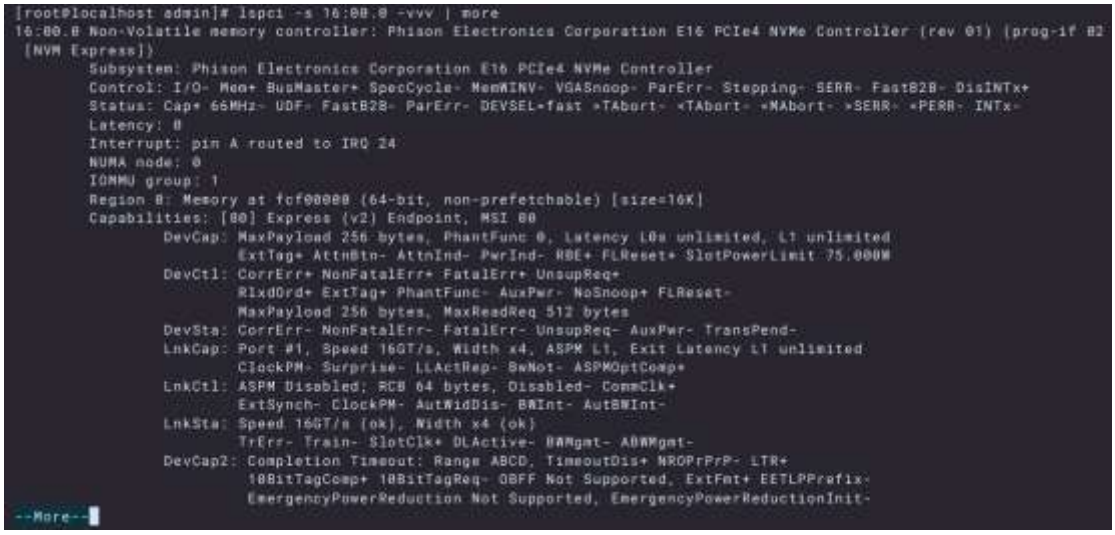
```
[root@localhost admin]# lspci | grep NVMe
16:00.0 Non-Volatile memory controller: Phison Electronics Corporation E16 PCIe4 NVMe Controller (rev ff)
[root@localhost admin]#
```

如果插入设备后不做任何处理直接 rescan，将无法找到 nvme ssd。需要在 rescan 之前先 remove 系统中残留的 pcie 信息，参见下面的图片。

```
[root@localhost admin]# echo 1 > /sys/bus/pci/devices/0000:16:00.0/remove
[root@localhost admin]# lspci | grep NVMe
[root@localhost admin]#
```

remove 命令移除对应设备的 PCIe 信息，再次 rescan，成功识别到设备

```
[root@localhost admin]# lspci | grep NVMe
16:00.0 Non-Volatile memory controller: Phison Electronics Corporation E16 PCIe4 NVMe Controller (rev 01)
[root@localhost admin]# nvme list
Node                  Generic               SN                   Model                                    Namespace
Usage                      Format           FW Rev
--------------------- --------------------- -------------------- ---------------------------------------- ----------
---------------------------- ---------------- --------
/dev/nvme0n1          /dev/ng0n1            SN20310897098S       GIGABYTE GP-AG4500G                      0x1
500.11  GB / 500.11  GB    512   B +  0 B   EGFM11.3
[root@localhost admin]#
```



综上，通过 Quarch Torridon drive control module 或者 card control module 注入模块，可以实现对于不支持热插拔的各类 m.2 ssd，插卡类产品实现自动化测试。

```
总结：
1.通过 quarch 热插拔或 M.2/插卡类从之模块对于 ssd 或者插卡进行掉电操作
2.删除 PCIe EP
echo 1 > /sys/bus/pci/devices/0000:b:d.f /remove        //* 0000 为 domain ID
3.重新枚举 PCIe 端点
echo 1 > /sys/bus/pci/rescan
```

4.1.7 Serial Cables Lane Reversal 测试 （U.2）

在实际测试过程中，有些用户希望进行 lane reversal 测试验证，例如将 lane 0,1,2,3 reverse 到 lane 3,2,1,0，这就需要使用 SerialCable 的 PCI4-AD-39M39F-KIT 套件（PCIe Gen4 U.2 – U.2 dual port, lane swapping adapters. Set of 4），如果需要这一类转接卡可以参照 5.3.2 Gen 4 U.2 转接卡。

该 KIT 内含个 adapter，支持 4 种不同的 lane swap 固定配置，每个 adapter 也提供 test points 将信号引出给示波器或者逻辑分析进行分析。

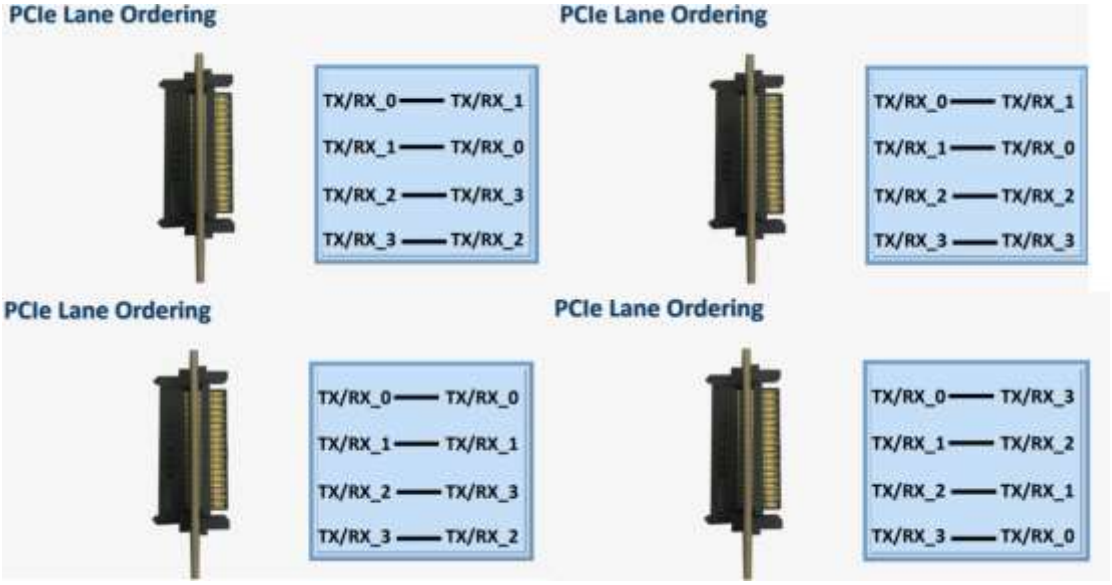
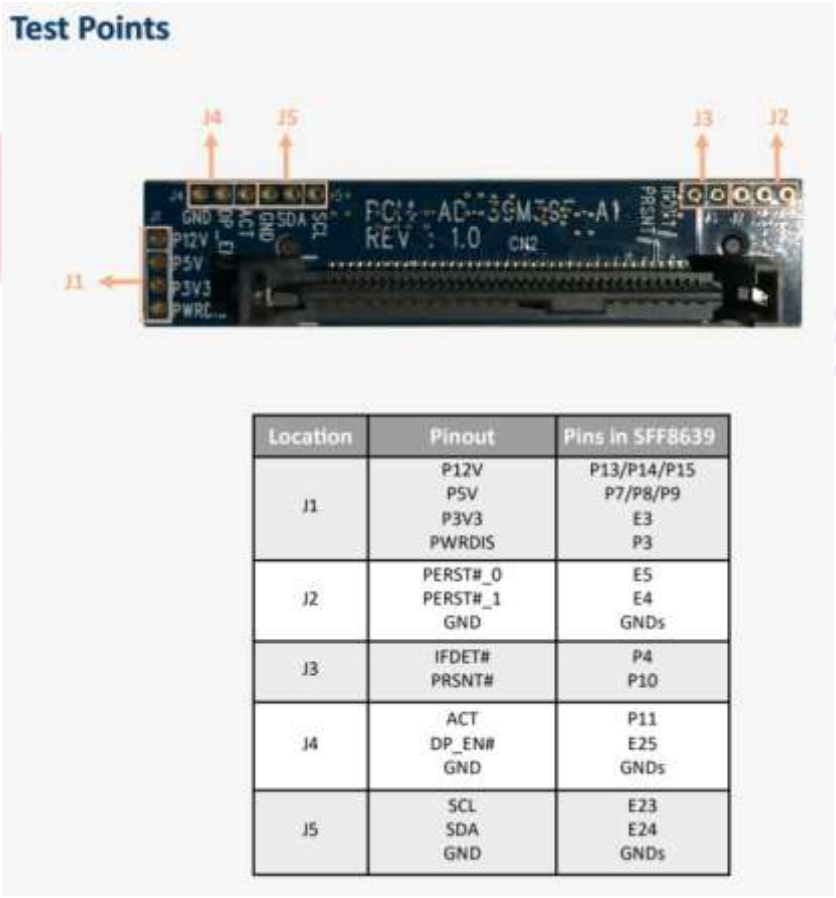


图 4-52



Location	Pinout	Pins in SFF8639
J1	P12V P5V P3V3 PWRDIS	P13/P14/P15 P7/P8/P9 E3 P3
J2	PERST#_0 PERST#_1 GND	E5 E4 GNDs
J3	IFDET# PRSNT#	P4 P10
J4	ACT DP_EN# GND	P11 E25 GNDs
J5	SCL SDA GND	E23 E24 GNDs

图 4-53

4.1.8 Serial Cables Lane Reversal 测试（插卡）

Serial Cables 公司新发布的 Lane Reversal 测试 adapter 套件，非常方便模拟各类插卡如果碰到主板 slot 不是按照标准定义的 lane 0~15 排列时候可能出现的问题。参见下图。

图 4-54

4.1.9 PCIe Gen4 或者 24G SAS cable 插拔测试

图 4-55

4.2 可编程电源 PPM – 电压拉偏和功耗测量

4.2.1 Quarch PPM 产品功能和配置介绍

该测试工具不仅为 UNH IOL 实验室在 Plugfest 使用，业内主流公司 SSD 基本也都在使用该可编程电源（简称 PPM, Programmable Power Module）进行电压拉偏和功耗测试。

该可编程电源可以通过 API 或者 GUI 界面实现突然将电压输出将为 0，也可以模拟各种各样的电压异常和波动进行电压拉偏（同时也间接实现了电流的波动）。程序控制的最低粒度为 1us，即你可以设置这 1us 内的电压输出为一个数值，然后设置下一个 1us 输出另外一个数值。下面是该可编程电源的一些主要技术参数。

- **12V/5V or 12V/3.3V mode software selectable**
- **Custom Pattern Generator**
- **250 KHz max sample rate**
- **Output Resolution:**
- **4mV**
- **Measurement Resolution:**
- **4mV, 25 uA**
- **Measurement Accuracy:**
- **±(2uA + 2%) @ 100uA-1mA**
- **±(2mA + 1%) @ 1mA-3000mA**
- **External trigger in/out**
- **Output Capacitance**
- **Pull Down**

可编程电源前面板右下角的输出口可以输出 12V/5V 电压，通过转接线缆，另外一端串接在各种接口的 SSD 和主板插槽之间实现对于 SSD 的供电，参见下图的右下角的各种转接线缆。最常见的 M.2 SSD 连接笔记本或者台式机 PCIe 插槽，U.2 SSD 以及 PCIe 插卡的转接线。

图 4-56 可编程电源，管理软件和转接线缆

下图的两张图分别是 U.2 和 M.2 治具连接 PPM 和测试环境的连接图， power fixture 夹具分别接入测试盘柜的背板或者主板 M.2 插槽进行测试，该夹具将阻隔背板给 SSD 供电，真正的供电将有软件控制 PPM 给 SSD 供电。

图 4-57 PPM 的 U.2 和 M.2

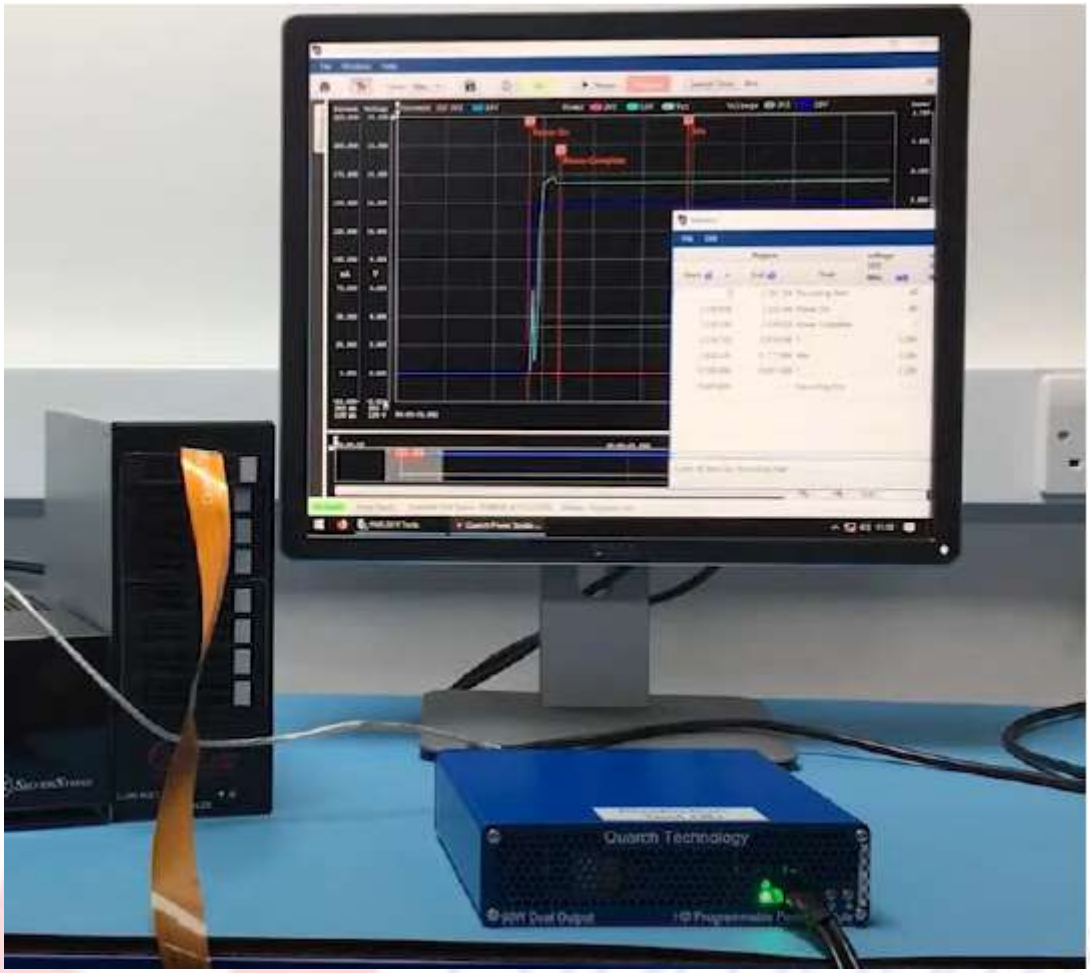


图 4-58 可编程电源，管理软件和转接线缆

下面是使用 Test Monkey 管理该电源的主界面，你可以在右下角点击 Edit Pattern 设置电压输出，然后即可 Run Pattern 输出你需要的电压。

图 4-59 Test Monkey 管理电源的主界面

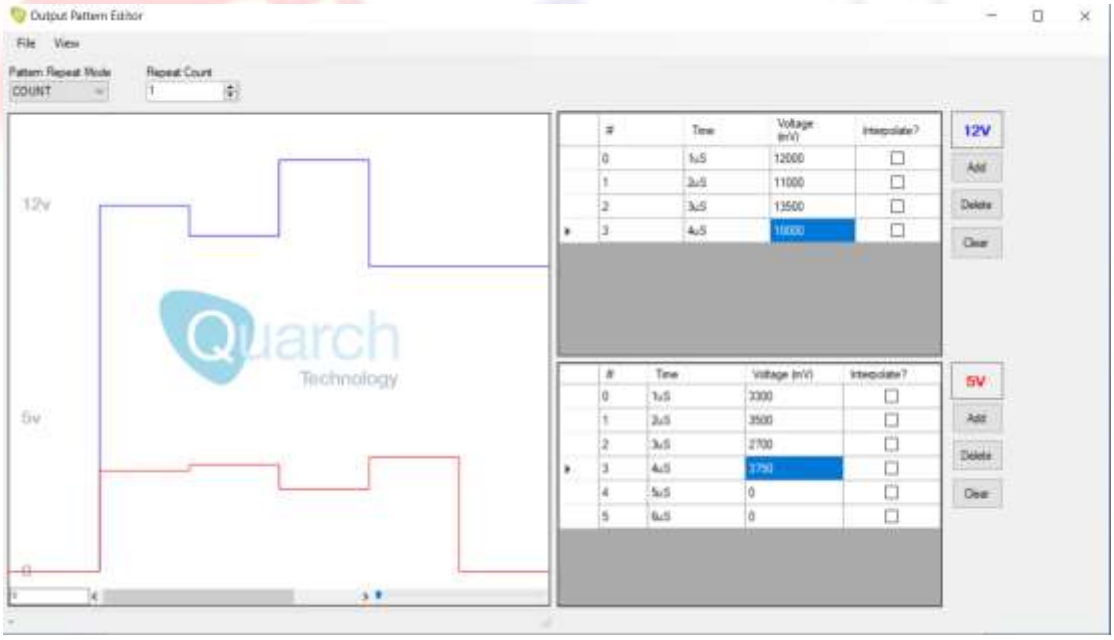


图 4-60 Edit Pattern 设置电压界面

下图为通过可编程电源可以很容易实现模拟一些故障或者问题电压。

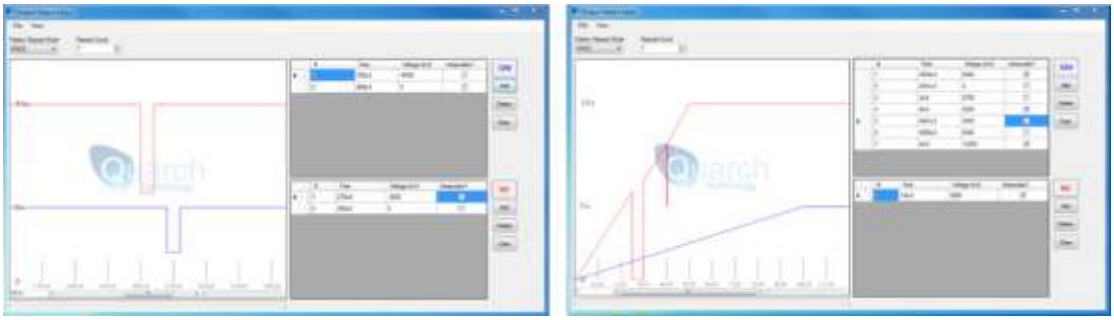


图 4-61 通过 Edit Pattern 注入异常波动电压

4.2.2 Quarch PPM 校准 Python 脚本包

https://quarch.com/products/quarchcalibration-python-package/

Quarch Calibration 是一个 Python 包，允许校准 Quarch 模块。

这个包需要 QuarchPy 才能与 Quarch 模块通信。更多信息在这里：https://quarch.com/products/quarchpy-python-package/

该软件包包含一个命令行工具，可引导您完成设备和仪器的选择和设置，或者可以传递命令行参数以完全自动化该过程。

QuarchCalibration 安装到 Windows 和 Linux 上的 Python 3.x 中。

要安装，请使用 PIP “pip install quarchCalibration”

校准过程的详细帮助可以在 Quarch 网站上的应用说明 19 中找到：https://quarch.com/file/an-019-hd-ppm-calibration/

注意：1）觉得大多数情况下无需校验；

2）上述的校准需要 Keithley 2460 source meter 和 calibration switch kit (QTL 2294)配合

4.2.3 Quarch: 使用 PPM&PAM 相对于传统使用示波器的优势分析

Andy Norrie

Posted: 23rd January 2018

Power consumption is a critical factor in both the design and the purchase of storage devices. Yet it can be hard to measure, especially on an individual device.

Traditional methods using an oscilloscope and current probes can be effective but are expensive and hard to implement. A tool that's specifically designed for drive power testing will give you a much wider range of test options and is much easier to set up. The cost savings of using a purpose-built tool are significant too.

4.2.3.1 Comparison tests

To quantify the difference between using one popular traditional approach and using a purpose-built tool, Quarch set up two key test scenarios in the lab, using:

- **A Quarch XLC Programmable Power Module (PPM), and**
- **A combination of a Tektronix DPO 3032 Oscilloscope and TCP0030 Current Probes.**

图 4-62

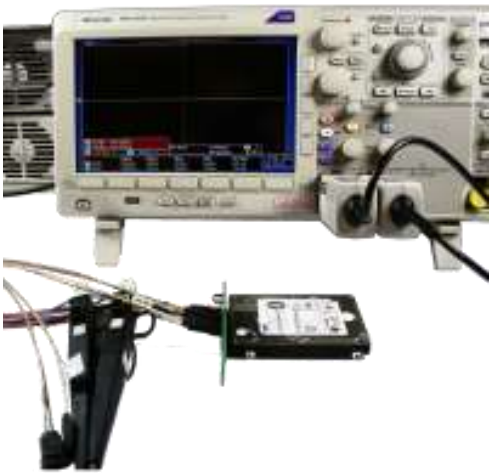


图 4-63

Each of the above was used to measure:

- **The average power use of an idle drive over 400 seconds**
- **Start-up current (the mean and RMS current for 20 seconds, beginning 2**

seconds before drive start-up).

4.2.3.2 Comparison results

The results give a clear comparison of the accuracy and usability of each method.

Oscilloscopes excel at measuring fast signals but aren't specifically designed for measuring the currents relevant to storage devices. Set-up is complicated and it's difficult to obtain accurate results or record results for long durations of time.

Quarch power modules solve these problems. In addition, they supply power, allowing you to run a range of extra tests.

The availability of injection fixtures and USB and Ethernet control, with dedicated software, makes running tests both simple and cost-effective.

- **Download the Programmable Power Module Vs Scope application note for the** full technical details **of the comparison tests and results.**
- **See below for a summary of the advantages of using Quarch power modules.**

4.2.3.3 Why use a purpose-built tool?

One of the main benefits of using Quarch power modules is that they are specifically designed for testing storage devices. They are purpose-built to eliminate the problems associated with traditional testing methods – and therefore perform better.

Some of the various traditional multi-unit testing methods may have a few of the following features; the Quarch power module has all of these:

- Quick, easy set-up **– specific power injection fixtures, for 2.5" drives, PCIe cards and M.2 devices, remove the hassle of clamping current probes in awkward places**

图 4-64

- **The ability to** run tests for long periods of time and continuously record power use **– ideal when running a drive workload simulation which may last for several hours or longer**
- Simple automation **– saving engineer time and making testing reliable, repeatable and fast**
- Low-current accuracy **– Quarch specify measurement down to 100uA, which is accurate at a far lower current than many other devices, and is ideal for measuring low power states**
- Fast sampling **– up to 250,000 samples per second, faster than a multi-meter or even many expensive Source Measure Units (SMUs)**
- **Full range, dual rail, power margining – create** fast slew-rate custom patterns**, from 0V to nominal +20%, allowing for ramps, glitches, brown-out and more, at 1uS resolution**
- External triggering **– allowing synchronization with third-party equipment (analyzers and similar).**

4.2.3.4 Cost-saving implications

Multiple financial benefits are achievable using the Quarch power module set-up:

- **Savings on the initial purchase price, compared to a scope**
- **Time-savings during the testing process via increased automation**
- **More accessible power data, allowing you to solve problems earlier.**

4.2.3.5 Initial purchase price

The total cost of an oscilloscope plus probes set-up can be easily three times that of a

Quarch XLC PPM. The single-output HD PPM achieves a similar cost-saving, while the multi-port, rack-mounted HD PPM offers further economic advantages if you're planning automated multi-device power testing.

4.2.3.6 Time-saving in the testing process

Significant time-savings can be made using an automated PPM-based system:

- **Save on set-up time**
- **Run unattended/overnight tests**
- **Get accurate results faster**
- **Test up to six devices simultaneously with the multi-port, rack-mounted HD PPM.**

As a result, you get your new storage products to market quicker and save on engineering time, leaving your engineers free to focus on designing the next generation of your company's products.

"Our readers need accurate, comprehensive information about the properties of the drives they buy, so we use Quarch's power modules to test drive power performance for our SSD reviews. Our testing process is much easier – and even more accurate – since we introduced the Quarch modules, making our SSD reviews even more relevant for our readers."
Gustav Gager, Nordic Hardware (the largest test lab in Sweden)

4.2.3.7 Solve problems faster

A scope can capture an image easily, but getting access to power data can be difficult. Using a Quarch PPM gives you easy access to raw data, so you can solve potential problems earlier in your design and testing processes.

4.2.3.8 What Quarch Partners say about our product

Over the years, Quarch have built up a great network of industry partners, including:

- Test labs **– helping product designers to test and validate their products**
- Reviewers **– testing and comparing the performance of the latest storage devices**
- Test equipment manufacturers **– providing complementary test products.**

图 4-65

The University of New Hampshire's InterOperability Laboratory (UNH-IOL) is the home of the SAS and NVMe Plugfest events. The UNH-IOL plays a critical role in ensuring the quality and interoperability of new storage products. Quarch modules are used at the Plugfests, with Quarch engineers normally on hand to assist with testing.

"The support that Quarch is providing to the UNH-IOL and the NVMe Integrators List program is enabling us to implement new types of testing that help prove the reliability and robustness of NVMe. The Quarch Torridon tool allows us to test the hot-plug features that are so important for enterprise SSD implementations."

David Woolf, Research and Development of Storage and Mobile Technologies, UNH-IOL

图 4-66

Tom's IT PRO is a major reviewer of storage products. Their test lab uses Quarch Programmable Power Modules as part of the review process for new drives.

"Our search for the one device that measures the power consumption of every type of storage device ends with the Quarch XLC Programmable Power Module… One of the most attractive features of the XLC PPM is its ability to displace multiple pieces of equipment, such as two bench power supplies, two scopes and two current probes, with one small device."

Paul Alcorn, Tom's IT PRO

图 4-67

Myce are also major reviewers of storage devices; they too use Quarch Power Modules as part of their storage devices review process.

"Myce has now secured a piece of state-of-the-art test equipment, which takes a sample every four micro-seconds, that I will be using to measure the power consumption of consumer grade SSDs and HDDs. Myce.com, in partnership with Quarch Technology, now aims to bring our readers the most comprehensive, and accurate, power consumption tests ever carried out on consumer grade storage devices, to be found anywhere on the Internet."

Wendy Robertson, Myce

图 4-68

Major drive reviewer, The SSD Review, also use Quarch Power Modules.

"We came across Quarch Technology at Flash Memory Summit and secured an amazing piece of equipment to use in our future reviews. To show our readers the power consumption of different drives, their XLC Programmable Power Module is quite honestly perfect. The Quarch XLC PPM enables us to easily and accurately analyze and log the power consumption of a device."

Sean Webster, The SSD Review

图 4-69

Nordic Hardware – the biggest test lab in Sweden – is another user of Quarch Power Modules.

“Our readers need accurate, comprehensive information about the properties of the drives they buy, so we use Quarch’s power modules to test drive power performance for our SSD reviews. Our testing process is much easier – and even more accurate – since we introduced the Quarch modules, making our SSD reviews even more relevant for our readers.”

Gustav Gager, Nordic Hardware

图 4-70

SANBlaze are leading specialists in storage emulation. Their VirtualLUN emulation solution includes automated control over Quarch modules. This allows hot-swap and physical layer fault injection to be easily combined with the rest of the SANBlaze test suite.

“The integration of the SANBlaze Emulation Tools along with the Quarch Signal Margining & Power Measurement Modules allows an easy way to control the test environment. All tests for both SAS & PCIe/NVMe tools can be run through a single GUI or automated through scripts.”

Marc Catanese, Director of Sales, SANBlaze

图 4-71

E8 Storage (now acquired by Amazon) developed a next-generation flash storage architecture that can deliver performance levels dramatically higher than existing products and with a lower TCO. The E8 team has a clear focus on testing and quality, and Quarch

products are integrated into their comprehensive automated testing plan.

“E8 is committed to building solutions of the highest quality. Quarch’s unique test tools are a simple and effective way to help test our new products.”

Ziv Serlin, Director of Systems Architecture, E8 Storage

SerialTek
an ellisys company

图 4-72

SerialTek LLC has been a provider of innovative data storage test tools and solutions since 2007. Many leading storage manufacturers depend on SerialTek analyzers to improve product quality and fulfill time-to-market requirements.

“The Quarch testing solutions have allowed us to quickly improve our test coverage and free up resources for other projects. With the simple scripting language and ease of system control, we are now capable of injecting power faults and measurements that were previously unavailable to our automated testing system. Thanks to the Quarch products we have increased our test coverage, removed manual tests, and feel more confident in our products’ designs and abilities.”

Greg Brown, Director of Software Engineering, SerialTek LLC

xyratex

图 4-73

Xyratex, later bought by Seagate, was one of Quarch’s first customers – and a key partner in the adoption of Quarch products as an industry standard for hot-swap testing.

“The Quarch Torridon systems are helping us maintain our world-leading reputation for quality. The automation of our hot-swap testing is being used in our Design Verification Testing (DVT).”

Paul Gregory, Development Test, Xyratex

4.2.4 Quarch: 功耗 VS 性能测试比拼，哪家企业级 SSD 更占优？

在我们的电源 VS 性能测试比拼中，哪个公司的企业级 NVMe SSD 的性能优于其它 SSD？

https://quarch.com/news/tests-reveal-which-enterprise-grade-drive-outperform-rest/

我们想与实验室中的几个 NVMe SSD 进行比较，以了解它们之间的差异。所有 NVMe SSD 都安装在我们的标准测试机之一上。这些是小型 PC，使用优质的消费类部件搭建而成。我们使用 SerialCables 主机卡、电缆和 JBOD 来实现简单、灵活的设置。

谁将是赢家？

为了确保公平测试，我们使用 Power Vs Performance 测试套件 (v1.0) 运行了 Quarch 合规套件。每个 NVMe SSD 都以空的、格式化的状态启动。

测试脚本从一组测试开始，以衡量“最佳情况”写入性能；然后在运行稳态测试之前运行 200% 写入过程以使 NVMe SSD 饱和。最后，我们执行读取测试。唯一的主要区别是 AIC 卡不支持热插拔，因此该设备跳过了热插拔/上电测试。

结论 – 您的企业是否愿意迎接挑战？

这对我们来说是一个迷人的过程，这是我们第一次能够以如此少的努力捕获如此多的功率和性能数据。测试运行确实需要很长时间，特别是对于三星 8TB SSD，但是一旦测试开始就不需要用户干预，因此可以放置一夜。

虽然许多结果在意料之中，但功耗和性能在单个工作负载中的变化方式非常有趣。

在我们之前的测试中，在 Gen3 主板上使用 SerialCables Gen4 主机卡似乎效果很好，我们使用了单一的、高度优化的 FIO 工作负载。我们为此测试打开的更详细的报告似乎会导致更低的 IO 率，因此我们需要研究 FIO 中可用的更复杂的选项。还需要调查缺乏 Optane 就绪的驱动程序。

好消息是，如果您不喜欢我们选择的 FIO 工作负载，您很快就可以选择自己的了！下一个 QCS 版本计划进行自定义 Power Vs Performance 测试，您可以在其中选择一个或多个工作负载文件来运行。然后，您可以在使用所需的确切工作负载的同时从我们的功率测量和报告中受益。

下面是英文版测试相关步骤和说明，仅供参考。

We wanted to do a comparison with several drives we had in our labs to see what the differences were between them. Our tests reveal which enterprise-grade drive outperforms the rest.

All Enterprise grade drives were set up on one of our standard test stations. These are small form factor PC’s, using good quality consumer components. We use SerialCables

host cards, cables, and JBOD for a simple, flexible setup.

图 4-74

Test Bed Setup:

- **Intel I7-8700 CPU with 16GB Ram on a ROG STRIX Z390-I motherboard**
- **Windows Server 2016**
- **SerialCables Gen4 Host Card (PCI-AD-x16HE-BG4)**
- **SerialCables 8 Bay JBOD**
- **Quarch HD PPM for power measurement**

图 4-75

HD PROGRAMMABLE POWER MODULE

A separate PC is used to run the QCS client and power capture, so there is no additional load on the CPU of the test PC Quarch Compliance Suite

The SSD drives

We chose some very different drives for the test:

TOSHIBA PX02SM SAS3 SFF SSD 400GB

INTEL SSDPED1D280GA Optane PCIe Gen3 x16 AIC SSD 280GB

The Toshiba SSD was connected directly to the host via SATA, as our SAS HBA was in use elsewhere. The Samsung SSD uses the U.2 2.5" form factor which can run through the SerialCables HBA and JBOD enclosure. The Intel drive was directly installed in the PCIe slot. These drives were chosen as a spread of enterprise-grade devices that we had on hand. The Toshiba is an older, mid-range SAS drive while the Samsung is a new high-cost, high-capacity Gen4 NVMe drive. The Intel Optane drive is a different storage technology and makes an interesting comparison.

Our initial expectations were that the Toshiba SSD would fall well behind the other two, due to its age and SAS protocol; SAS (Serial Attached SCSI) uses the SCSI command set which was designed for spinning drives, whereas NVME (Non Volatile Memory Express) is designed for flash and should be more optimal. Connecting it to a motherboard SATA port would further limit its performance.

We expected the Samsung drive to use more power due to its high capacity and Gen4 speed, but with no idea 'how much more it would be. While our test PC is Gen3, we are using the SerialCables x16 Gen4 host card. In theory, the 16 lanes of Gen3 to the host card will be enough bandwidth to fill the 4 lanes of Gen4 from the host card to the SSD. The SSD does indeed link up and report it is running at Gen4 speeds, but the overall performance may still be below a native Gen4 system.

Who will be the winner?

A comparison with several drives we had in our labs; fair testing was essential. We ran Quarch Compliance Suite, using the Power Vs Performance test suite (v1.0). Each drive started in an empty, formatted state.

The test script begins with a set of tests to measure the 'best case' write performance; it then runs a 200% write process to saturate the drive before running steady state tests. Finally, we perform read tests. The only major difference is that the AIC card does not support hot-plug, so the hot-plug/power-up test is skipped for this device.

Results of interest

The first thing to note is just how different the performance traces look. The Intel drive looks very 'clean', with very predictable, steady power consumption.

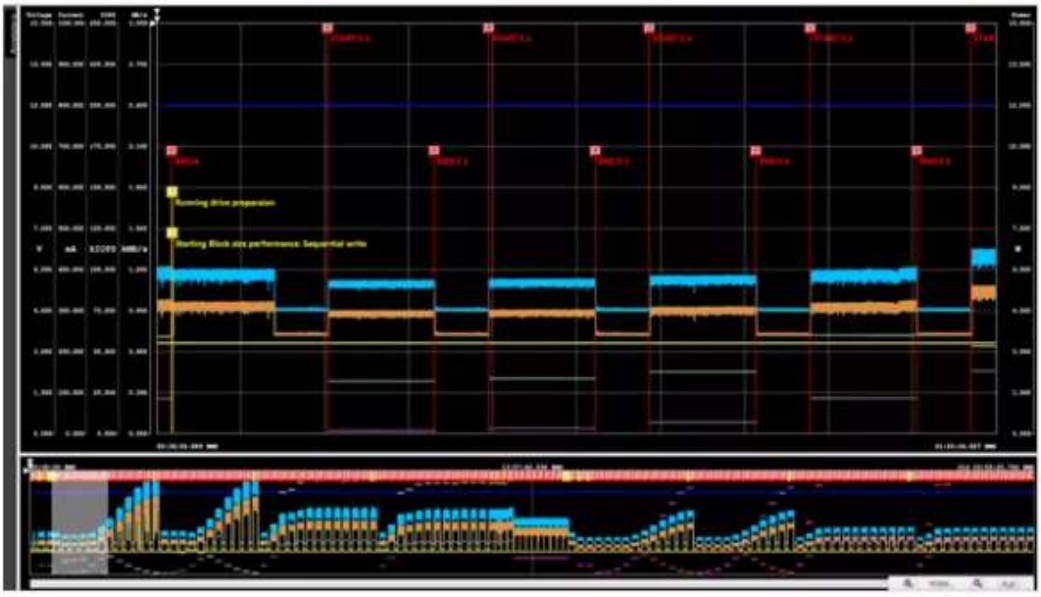


图 4-76

Intel Optane drive – Steady power and performance in all tests

The Toshiba drive trace is a lot more cluttered, due to the greater variation in power consumption through the workloads. The Toshiba drive is oscillating between 3-6 watts while the Intel drive moves over a much smaller range: 5.3-5.6 watts for the same test.

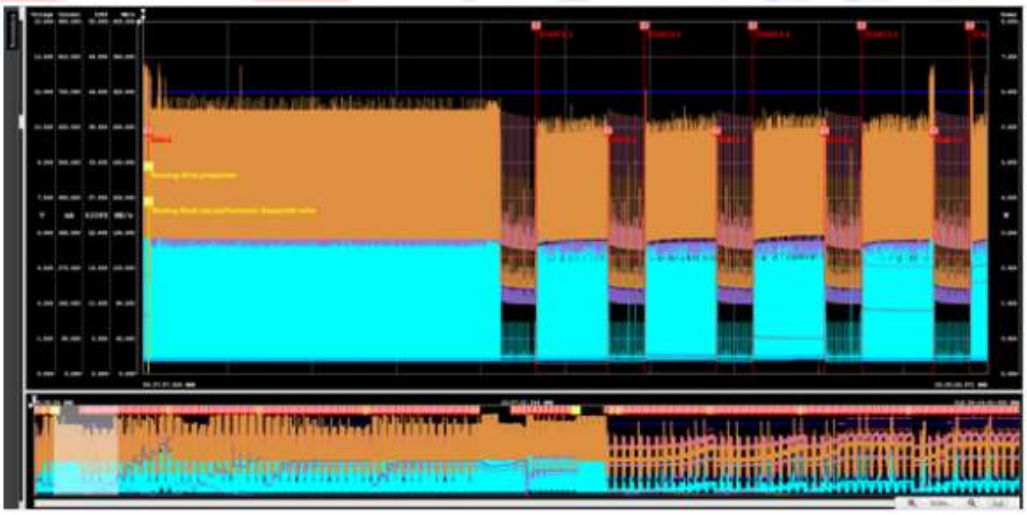


图 4-77

Toshiba SAS drive – Much noisier on the power rails

The Samsung drive moves vary between 6.8-7.7 watts but peak up to 8.7 watts around once every 600mS in a regular pattern.

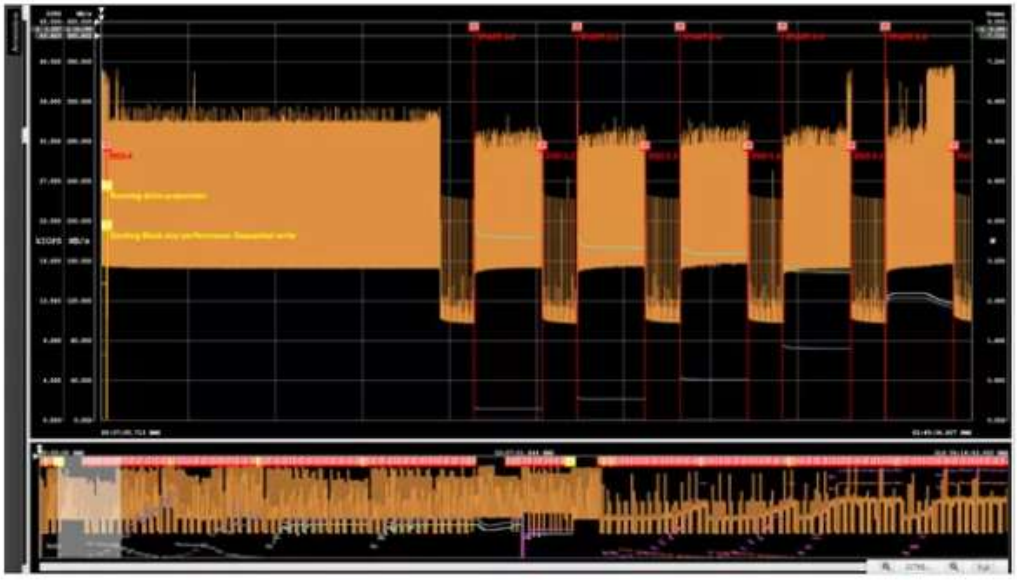


图 4-78

Samsung U.2 drive – The highest power consumption of all drives

In QPS we can hide the voltage, current, and sidebands, and show just the total power and IO performance for a trace that is easier to view.

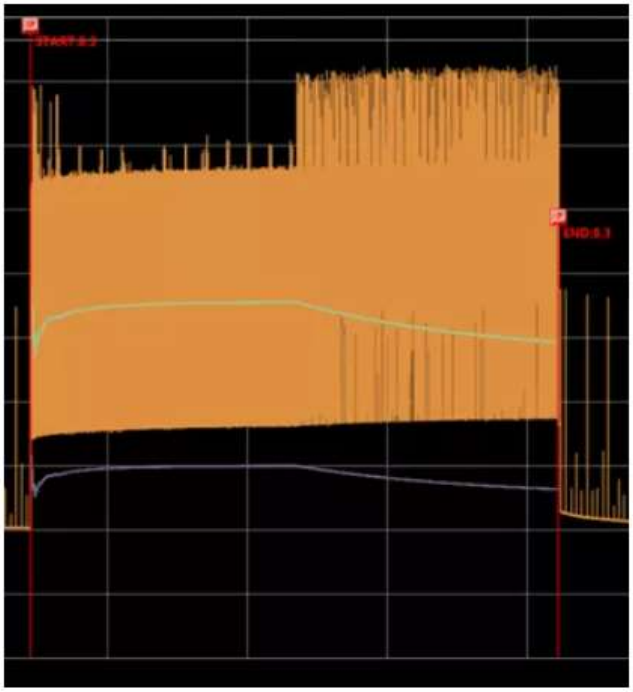


图 4-79

Toshiba SAS drive – Now with only total power and IO performance shown

Example: Queue depth test

This example is a random write test, working through a range of queue depths, from 1 to

256. Each workload runs for 10 minutes, with a 5 minute idle before each it, to give the drive time to perform housekeeping tasks. Let's look at all three drives during the QD=2 workload.

We can see the differences in performance stability. The Intel drive is almost completely flat across this workload (and almost every workload we tried)

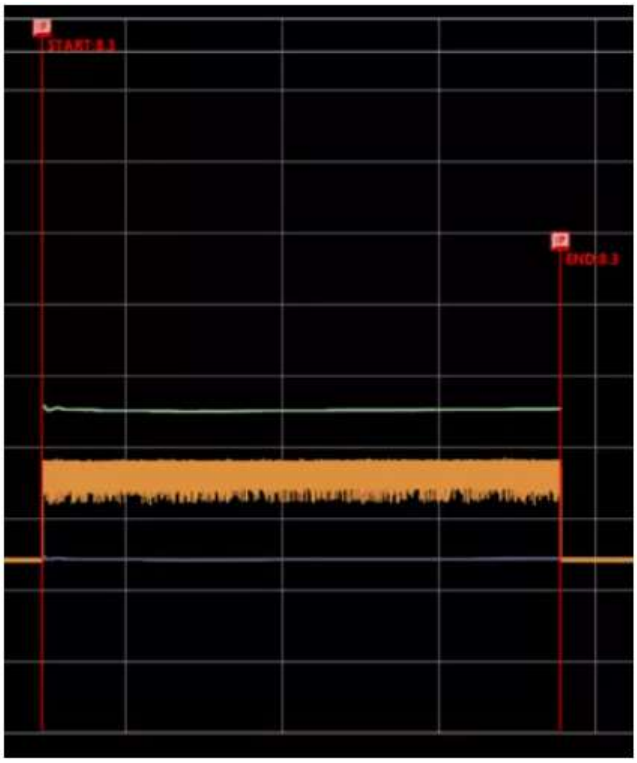


图 4-80

Intel Optane drive – Very predictable performance

Toshiba changes significantly over time. There is a big dip then recovery in performance over the first 20 seconds, then a steady slow down after the 6-minute mark. Power use also increases at the same time.

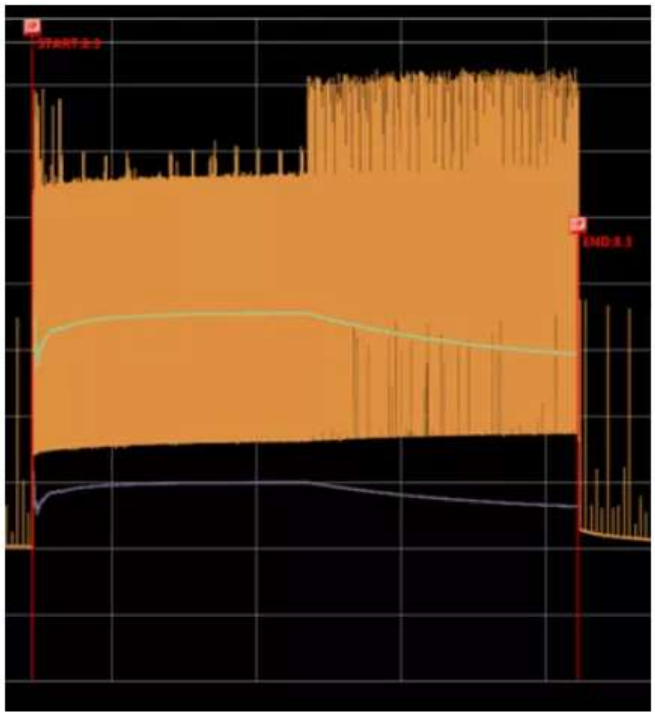


图 4-81

Toshiba SAS drive – Performance drops and power increases

The Samsung drive has much less of a performance dip, but the power consumption change is even more significant than on the Toshiba drive. The Samsung device shows this ‘double hump’ of power consumption across many of the initial workloads

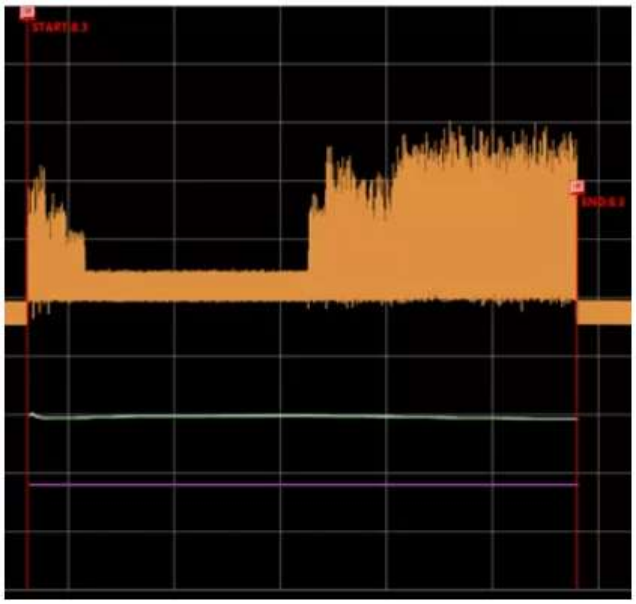


图 4-82

Samsung U.2 drive- Power increases, but performance is flat

Remaining tests

This is just a quick look into a couple of test points. The Power Vs Performance test has

18 sections, each of which looks at one specific workload type. The entire test takes several hours to run (Over 24 hours in the case of the larger drives).

Each part of the test shows fascinating detail into the operation of the drive. For example, the Samsung drive performs quite differently after the 200% write pre-conditioning has been done.

After this, it has a far more steady power consumption:

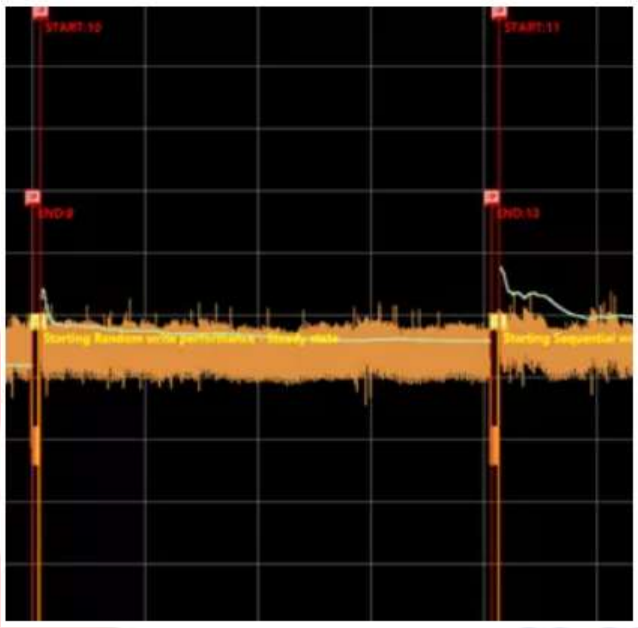


图 4-83

Samsung U.2 drive- Operating in a steady state after pre-conditioning

Of course, these specific workloads may not pick up on some specific case you are interested in. The next version of QCS is coming with a custom power test, so you can select your own FIO job files to run!

PDF Report

Now let's take a look at the PDF report output, which is generated at the end of the test

Idle power

The Optane drive uses more power at idle but has a lower peak consumption. The big Samsung drive uses greatly more power than the other two.

Toshiba SAS:

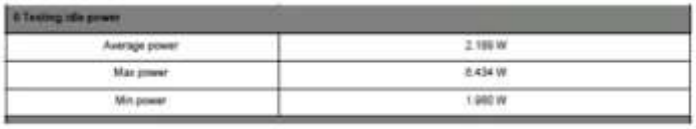


图 4-84

Intel Optane:

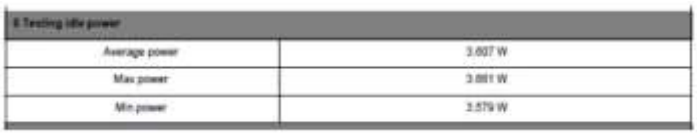


图 4-85

Samsung 8TB

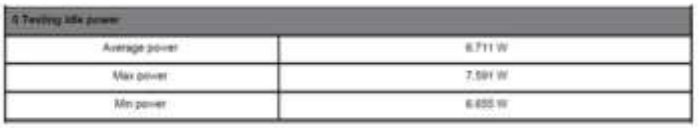


图 4-86

Latency

Unsurprisingly, the Optane drive has very good latency during random writes when compared to the older Toshiba drive. While it uses a bit more power, it is almost 3 times better in terms of power efficiency per MB written.

The newer Samsung drive is very interesting though. Its latency figures match the Optane drive closely, (though the actual throughput and power consumption are poorer). This may be due to our test system not be fast enough to get the best out of the Optane technology. It is also possible that the regular NVME driver is an issue. We later learned that Intel recommends the use of a custom driver for Optane products. One data point to note is that the Optane drive spec says it can manage 10uS latency, but we are only seeing around 75uS.

Toshiba SAS:

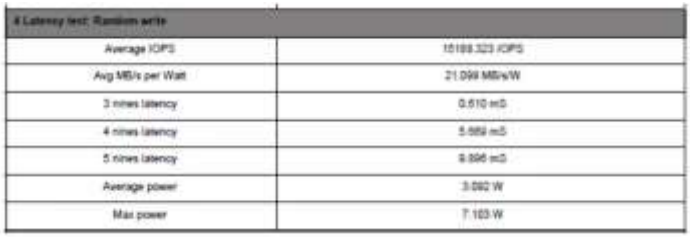
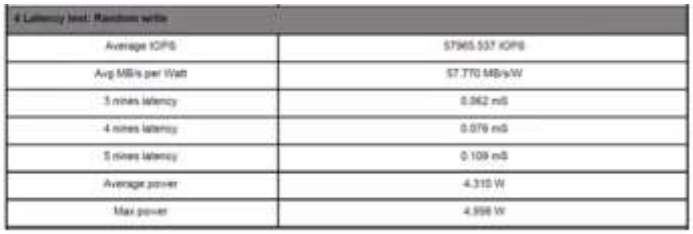


图 4-87

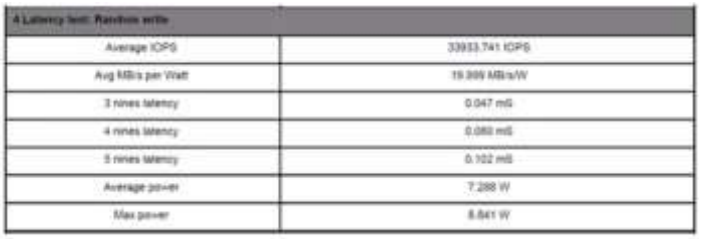
Intel Optane:



4 Latency test: Random write	
Average IOPS	57965.537 IOPS
Avg MB/s per Watt	57.770 MB/s/W
3 nines latency	0.062 mS
4 nines latency	0.076 mS
5 nines latency	0.109 mS
Average power	4.310 W
Max power	4.998 W

图 4-88

Samsung 8TB:



4 Latency test: Random write	
Average IOPS	33933.741 IOPS
Avg MB/s per Watt	19.999 MB/s/W
3 nines latency	0.047 mS
4 nines latency	0.060 mS
5 nines latency	0.102 mS
Average power	7.288 W
Max power	8.841 W

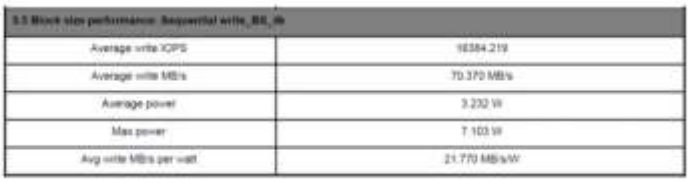
图 4-89

Write performance

Here's one more example. A Sequential write test with a 4k block size. As we expected, the Optane drive is significantly faster. It also gets much better performance in terms of MB/s/watt (The amount of data written for the power consumed). Again though, it looks like the drives are being limited by our test PC and not achieving the full performance they could manage in a modern server.

The large Samsung SSD does draw the most power, both on average and at peak, as we expected.

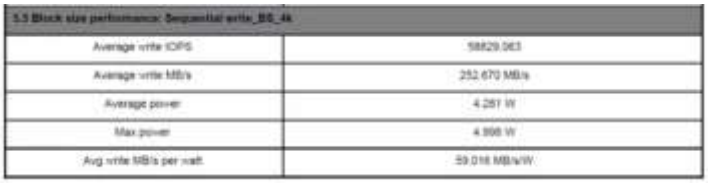
Toshiba SAS:



5.5 Block size performance: Sequential write_BS_4k	
Average write IOPS	16384.219
Average write MB/s	70.370 MB/s
Average power	3.232 W
Max power	7.103 W
Avg write MB/s per watt	21.770 MB/s/W

图 4-90

Intel Optane:



5.5 Block size performance: Sequential write_BS_4k	
Average write IOPS	58829.063
Average write MB/s	252.670 MB/s
Average power	4.267 W
Max power	4.998 W
Avg write MB/s per watt	59.016 MB/s/W

图 4-91

Samsung 8TB:

5.5 Block size performance: Sequential write_BS_4k	
Average write IOPS	34225.826
Average write MB/s	146.999 MB/s
Average power	7.570 W
Max power	8.311 W
Avg write MB/s per watt	19.419 MB/s/W

图 4-92

Conclusions – Is your enterprise drive up to the challenge?

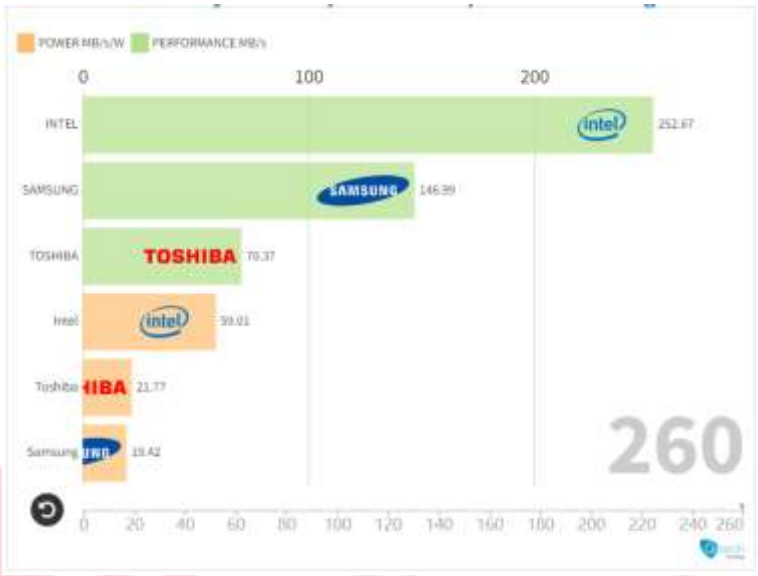


图 4-93

This was a fascinating process for us, being the first time we have been able to capture so much power and performance data and with so little effort. The test runs did take a long time, especially for the Samsung 8TB SSD, but it needed no user intervention once the test was started and so could just be left overnight.

While many of the results were expected, the way that power consumption and performance can vary within a single workload is very interesting.

The use of the SerialCables Gen4 host card on a Gen3 motherboard seemed to work very well in our previous testing where we used a single, heavily optimized FIO workload. The more detailed reporting we have turned on for this test seems to result in lower IO rates, so we need to look into the more complex options available in FIO. The lack of an Optane ready driver also needs to be investigated.

The good news is that if you do not like the FIO workloads we have chosen, you'll soon be able to select your own! The next release of QCS is planned to have a custom Power Vs Performance test, where you can select one or more of your workload files to run. You can then benefit from our power measurement and reporting while using the exact workloads you want.

4.2.5 TechPowerUP Labs 采用 Quarch PPM 测试 Acer Predator GM7 1 TB SSD w/Maxio 主控+128 层 YMTC NAND 闪存

Acer Predator GM7 1 TB Review - Impressive Performance

by W1zzard, on Apr 25th, 2023; SSD Manufacturer: Acer

Acer is a world-leading manufacturer of computer hardware. The company was founded in 1976, in Taiwan and are mostly known for their laptops, desktop PCs, and monitors. BIWIN Storage has been granted an official license by Acer to produce, market and sell solid-state-drives using their name.

图 4-94

With this review we're introducing our new 2023 SSD Test Bench, which comes with upgraded hardware, refined tests and new power consumption measurements. The first SSD that we're testing is the Acer Predator GM7, which is the company's newest cost-efficient drive. It is based on the new Maxiotech MAP1602 controller, paired with 128-layer YMTC TLC NAND flash.

The Acer Predator GM7 is available in capacities of 512 GB GB ($50), 1 TB ($66) and 2 TB ($100). Endurance for these models is set to 300 TBW, 600 TBW and 1200 TBW, respectively. Acer includes a five-year warranty with the GM7.

Specifications: Acer Predator GM7 1 TB SSD	
Brand:	Acer
Model:	GM7-1TB / BL.9BWWR.118
Capacity:	1024 GB (953 GB usable) No additional overprovisioning
Controller:	Maxiotech MAP1602A
Flash:	YMTC 128-Layer 3D TLC Rebranded as BIWIN BWN09TC1B1RCAD
DRAM:	N/A but 32 MB HMB
Endurance:	600 TBW
Form Factor:	M.2 2280
Interface:	PCIe Gen 4 x4, NVMe 1.4
Device ID:	Predator SSD GM7 M 2 1TB
Firmware:	SN08560
Warranty:	5 years
Price at Time of Review:	$66 / $64 per TB

图 4-95

4.2.5.1 Power Consumption

图 4-96

For the SSD power consumption tests, a Quarch QTL1999 programmable power module is used, paired with in-house TPU software. This method allows for the monitoring of a drive's power usage profile with microsecond precision in our custom workloads. Data is pulled into our processing pipeline over Ethernet, on a separate machine, ensuring that measurements do not affect the SSD being tested. All measurements are drive power only, not full system.

4.2.5.2 Idle Power

We present two results for SSD Idle Power Consumption. The first one is called "Desktop" and represents the usage in a typical desktop, which by default does not have the advanced PCIe sub-states enabled, and drives cannot enter their lowest power state.

The second result, named "Mobile," represents power consumption with L1 ASPM enabled, as found on most modern laptops.

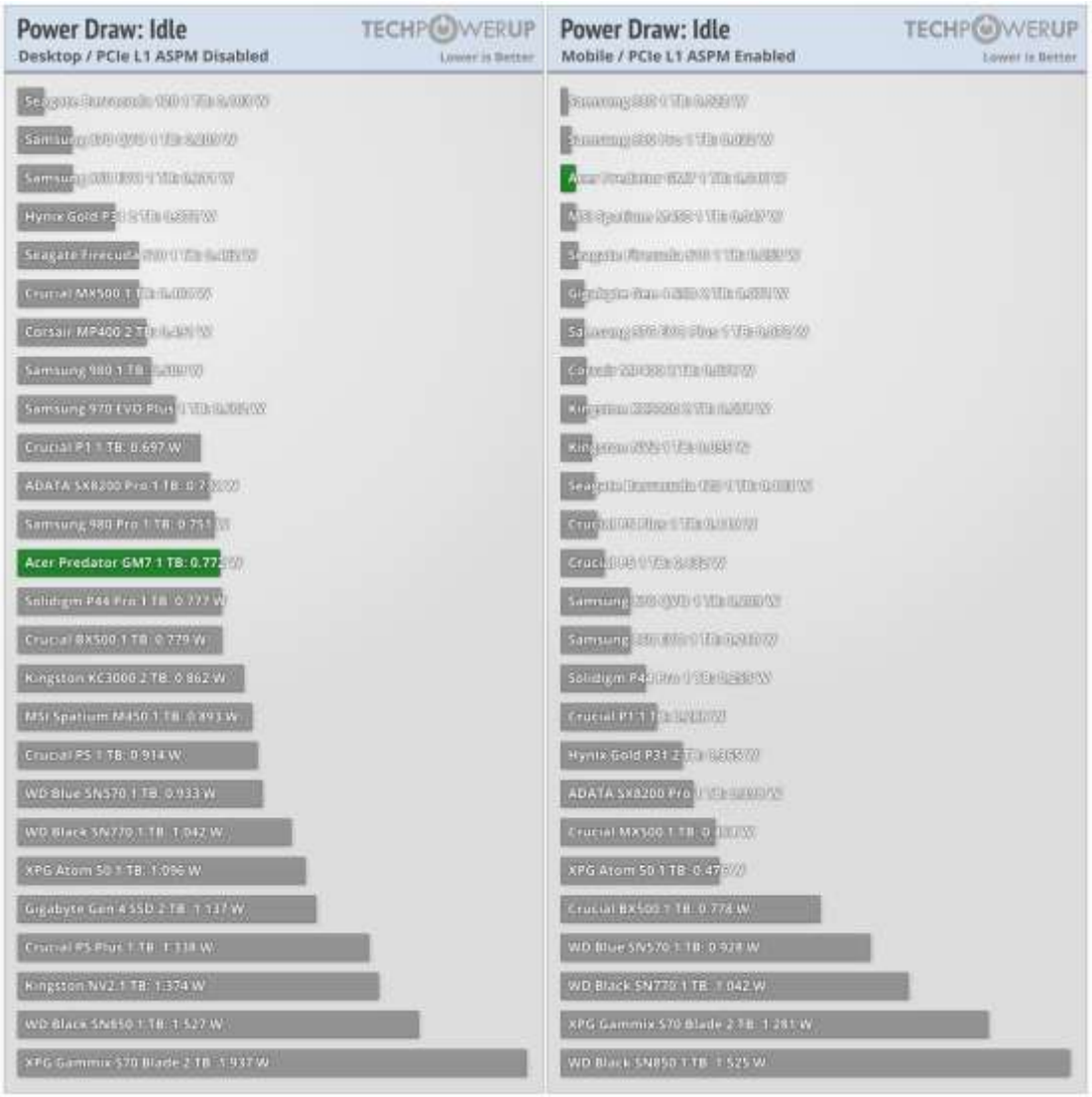


图 4-97

4.2.5.3 Power Consumption under Load

In order to obtain a comprehensive understanding of the drive's power consumption when not idle, we subject it to both random and sequential load patterns at different queue depths. The "Weighted" result in the comparison charts takes into account the typical behavior of today's client workloads, which mainly operate at low queue depths.


Acer Predator GM7 1 TB
4K Random, Power Consumption
TECHPOWERUP
Read
Write
3.0 W
2.5 W
2.0 W
1.5 W
1.0 W
0.5 W
0.0 W
QD1
QD2
QD4
QD8
QD16
QD32
QD64
QD128
Acer Predator GM7 1 TB
512K Sequential, Power Consumption
TECHPOWERUP
Read
Write
4.0 W
3.5 W
3.0 W
2.5 W
2.0 W
1.5 W
1.0 W
0.5 W
0.0 W
QD1
QD2
QD4
QD8
QD16
QD32
QD64
QD128

图 4-98

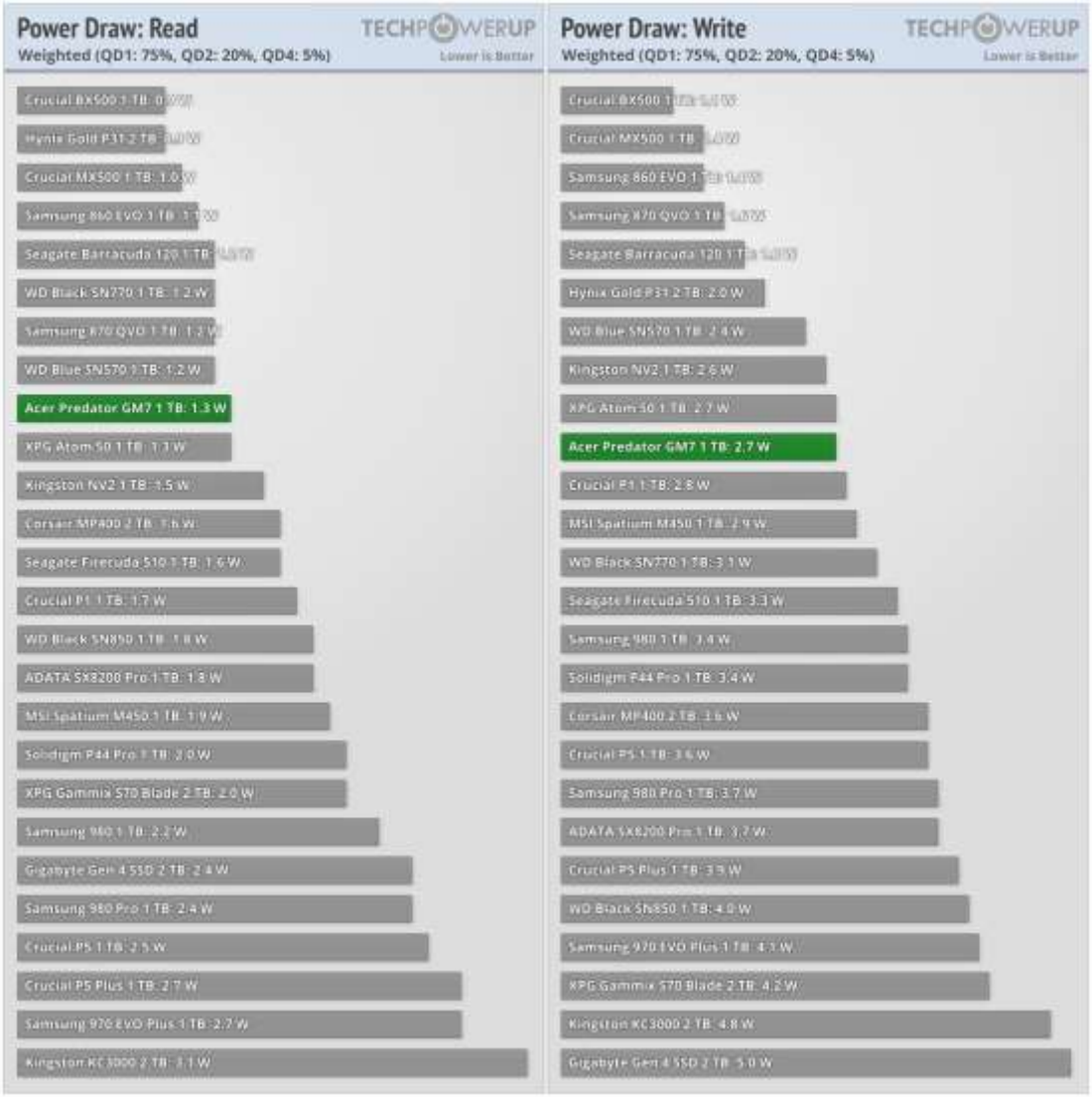


图 4-99

4.2.5.4 Gaming Power Draw

For our gaming power draw test we chose to measure Red Dead Redemption 2 while loading a savegame from the main menu.

图 4-100

4.2.5.5 Maximum Power Consumption

The highest observed value of all testing on this page is reported as maximum power consumption.

图 4-101

4.2.5.6 Power at Fixed Speed

Most of the time an SSD will not run at its highest transfer rates. Testing in this section accounts for that and tests sequential read and write at specific transfer rates, so you can get an idea of power consumption when only lightly loaded.

图 4-102

4.2.5.7 Energy Efficiency

Last but not least, we have energy efficiency, which is calculated from the "Power Consumption under Load" results, at both read and write.

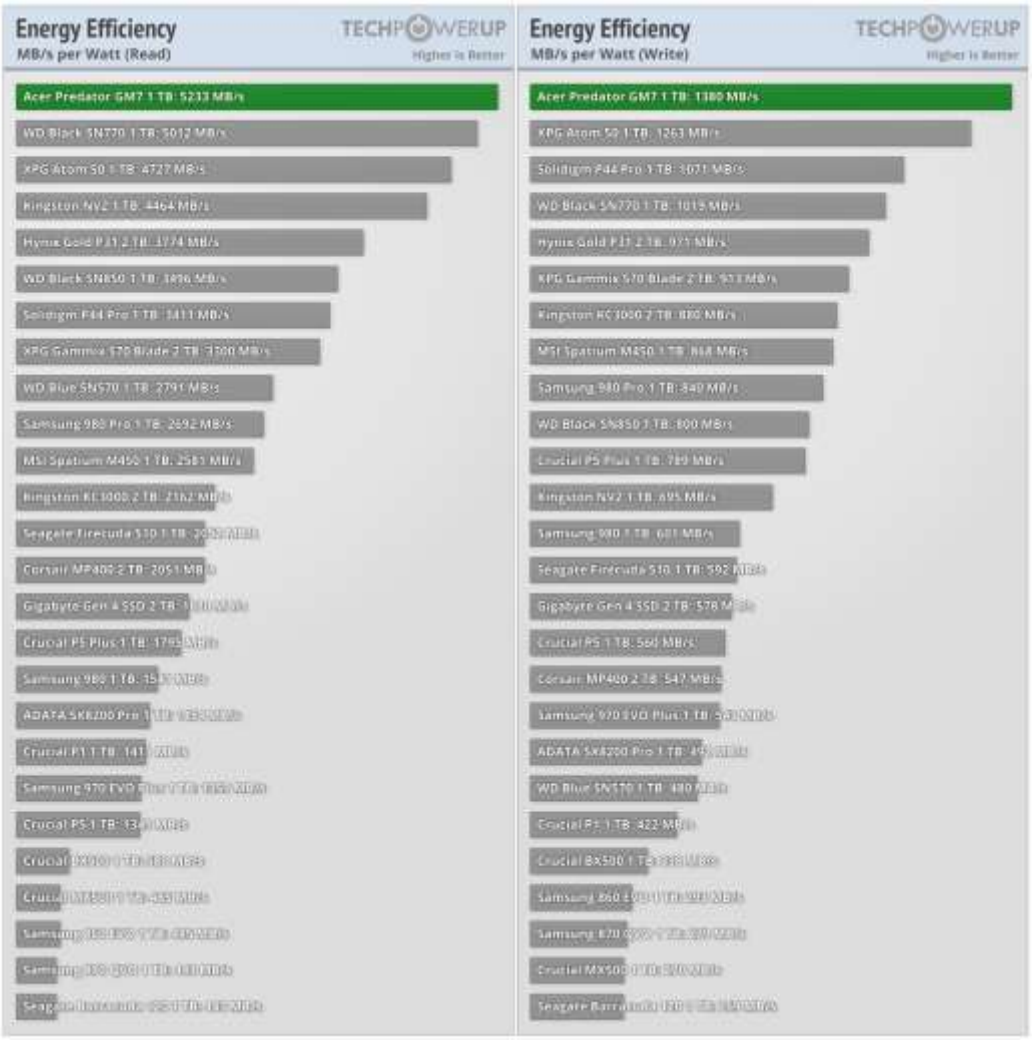


图 4-103

4.3 电源分析模块 PAM - 电压/电流/Sideband

4.3.1 Quarch PAM 产品功能和配置介绍

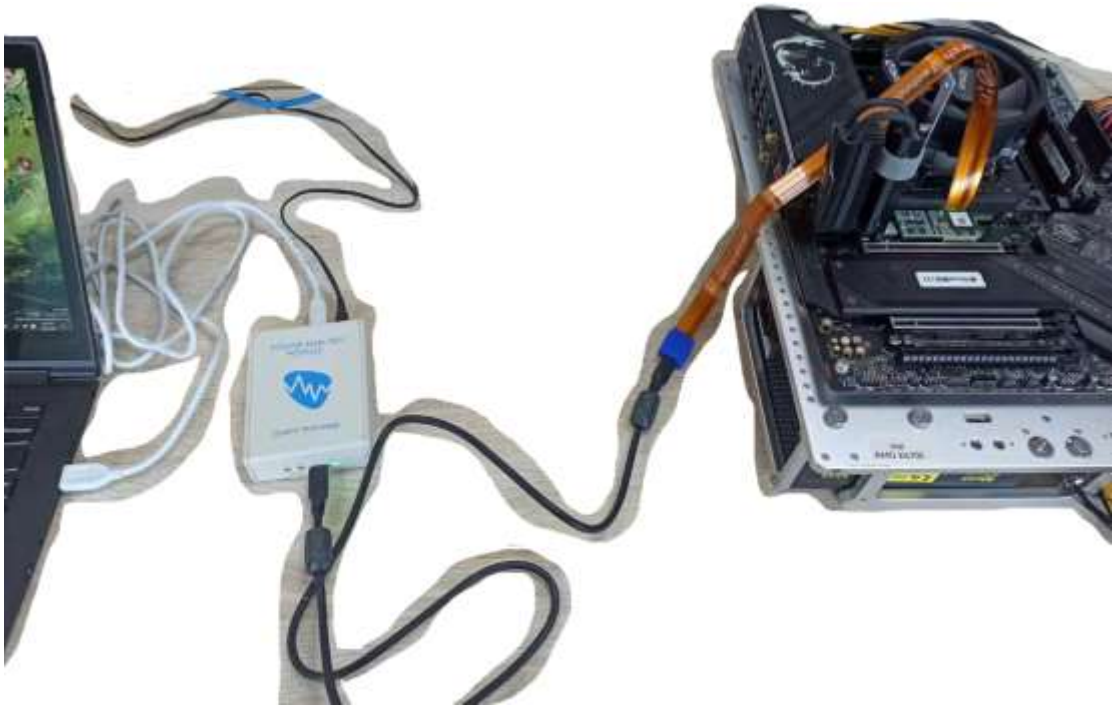
随着众多的消费类 M.2 NVMe SSD 应用于各种各样的场景，例如笔记本，GPS，汽车电子等，研发/测试工程师发现诊断、分析、排除关于低功耗的问题，传统的手段，例如使用万用比或者示波器等捕获波形等越来越难以应付出现的问题。

图 4-104

Quarch 新发布的 PAM (Power Analysis Module) 使得用户分析这些低功耗问题变得易如反掌，Quarch M.2 NVMe SSD PAM 模块串接在 M.2 SSD 和 M.2 socket 之间，可以长时间、高精度地记录电压，电流，功耗，以及各个 sideband signal，例如 PERST#, SMDAT, SMCLK, CLKREQ#, WAKE#等，这样，当笔记本进入低功耗的时候，通过 PAM 的管理软件 QPS (Quarch Power Studio)可以实时地、清晰地获得所有你想获得的信号信息，也可以事后回溯分析，同时抓取的数据也可以生成 CSV 等表格用于后处理分析。

图 4-105

下面是针对 PCIe Gen5 M.2 SSD 实际连接图片。



下面是另外一张常用的针对 PCIe Gen5 U.2 SSD 的实际连接图片。

下面是 PAM 管理模块的前、后面板，以及 Gen5 M2. PAM Fixture 治具的图片细节。

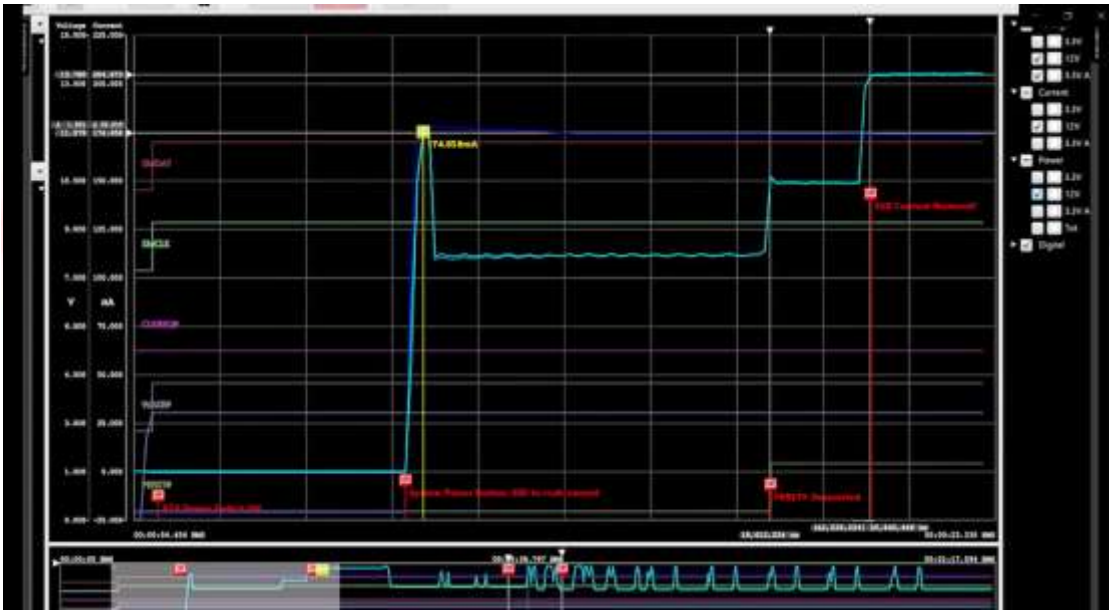


图 4-106

除了 M.2 PAM 之外，Quarch 也提供针对 Gen 4 U.2/U.3 SSD 的 PAM, 可以满足分析 2.5’ Gen 4 NVMe SSD 的电压，电流和功耗等，同时，Quarch 也提供针对 AIC 插卡的 PAM, 允许工程师分析各种 PCIe 插卡，包括 PCIe Gen 4/5/6 x8 NVMe SSD 卡，或者任意其它的网卡，显卡，GPU 卡，HBA 卡，RAID 卡的电压，电流和功耗信息。

图 4-107

上面的图片是测量一张 Intel Optane SSD 卡的连接示意图。插卡，PAM 可以分析各种各样的插卡，例如下图的网卡，GPU 卡等。

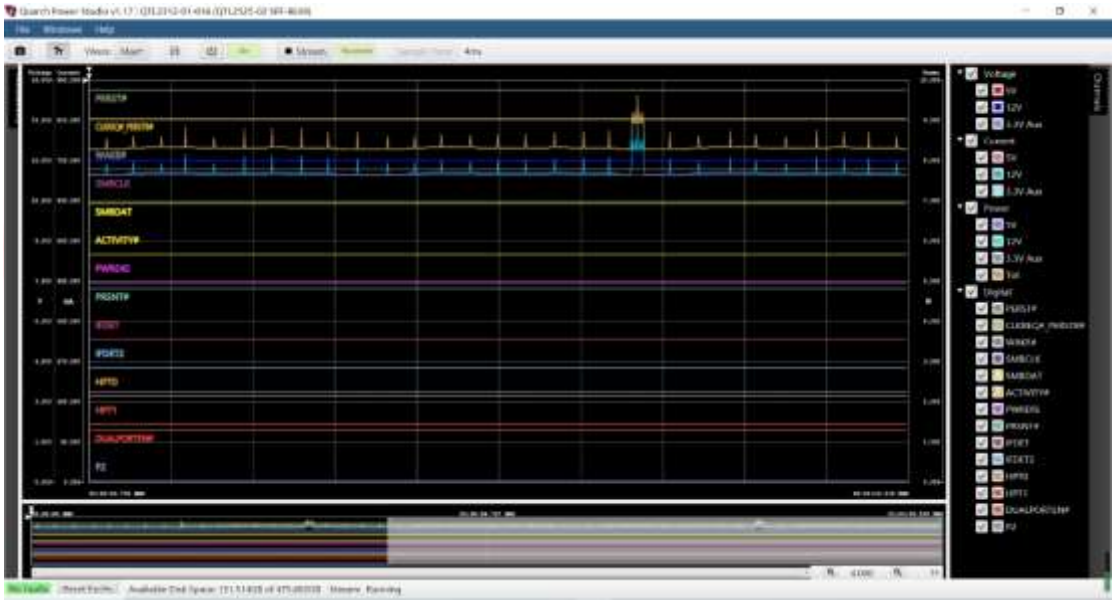


图 4-108

该 PCIe x16 插卡夹具可以测试任意的卡类产品。如果需要测试各种 SSD 盘，U.2, U.3, M.2, EDSFF 等只要选择相应的夹具即可。下面是一些测量的截图。

- AIC PAM
  - PSU is chirping during power loss (spiking up in voltage briefly)
  - WAKE signal is asserting on each jump in voltage

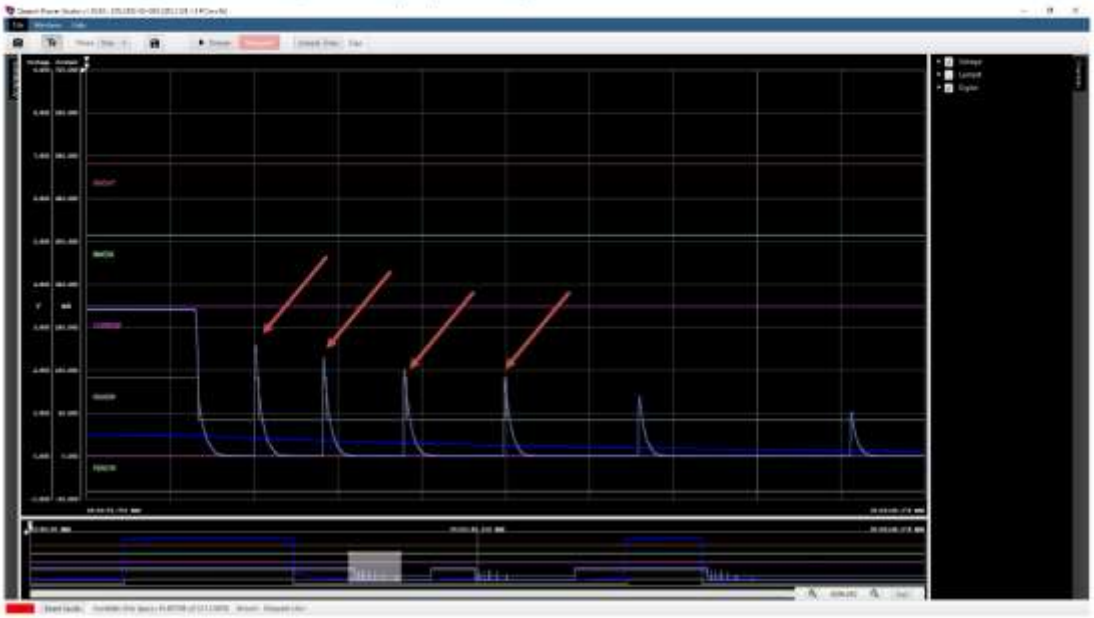


图 4-109

- All major sidebands are recorded
- See when your host is asserting PERST in relation to hot-plug

图 4-110

- Power level seen moving with CLKREQ# signal

图 4-111

- Best case was 4.4 mW on this drive

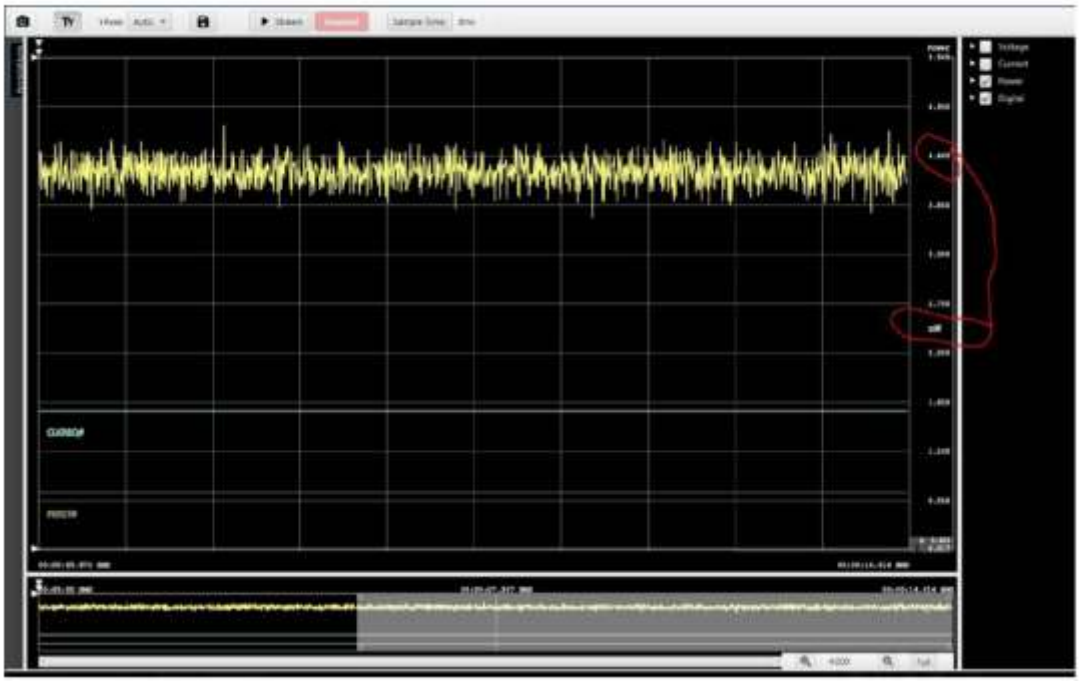


图 4-112

4.3.1.1 M.2 - POWER STATE 0 示例

参见下面的截图 –

- 11.5 watts average
- 100k IOPS average
- 440 MB/s average

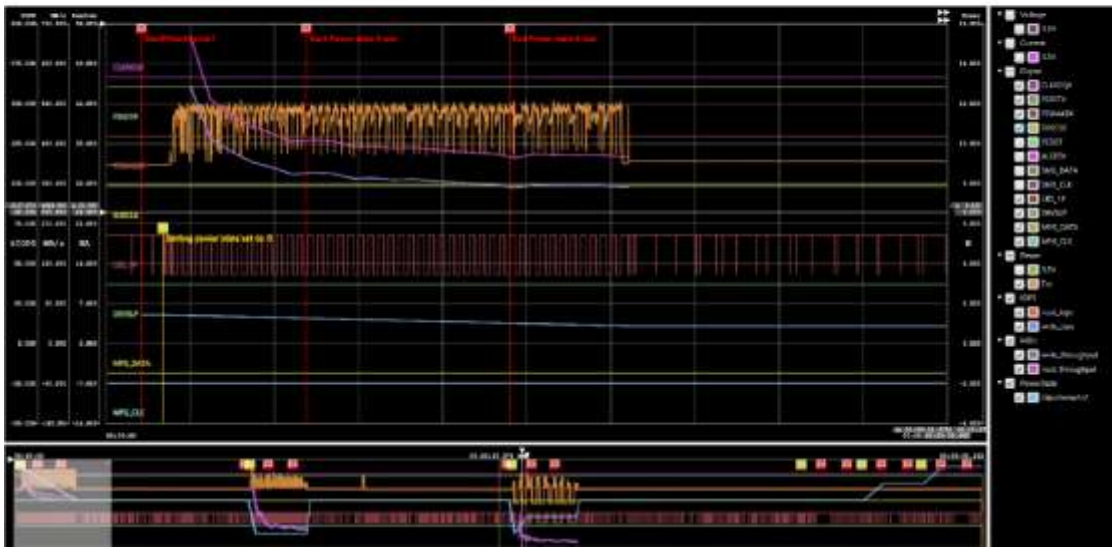


图 4-113

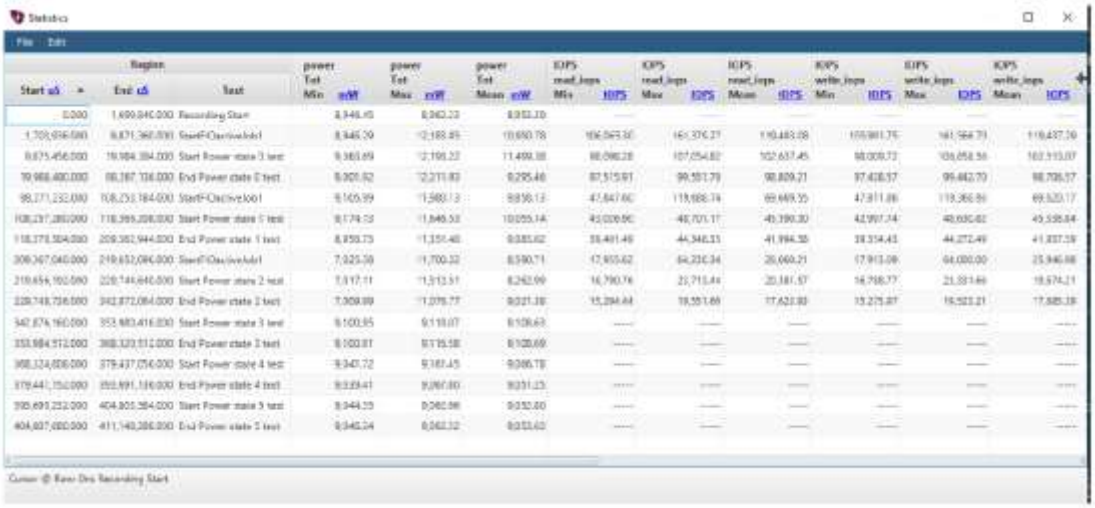


图 4-114

4.3.1.2 M.2 - POWER STATE 1 示例

参见下面的截图 –

- 10 watts average
- 45k IOPS average
- 195 MB/s average

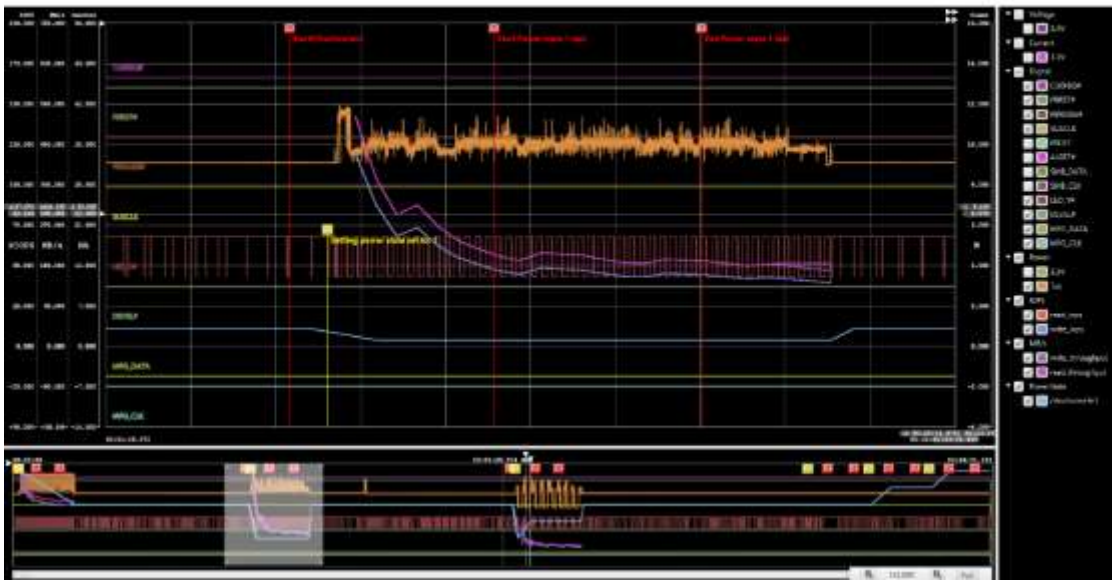


图 4-115

4.3.1.3 M.2 - POWER STATE 2 示例

参见下面的截图 –

- 8.2 watts average
- 20k IOPS average
- 84 MB/s average

- Power drops by: 28%
- Performance drops by: 80%

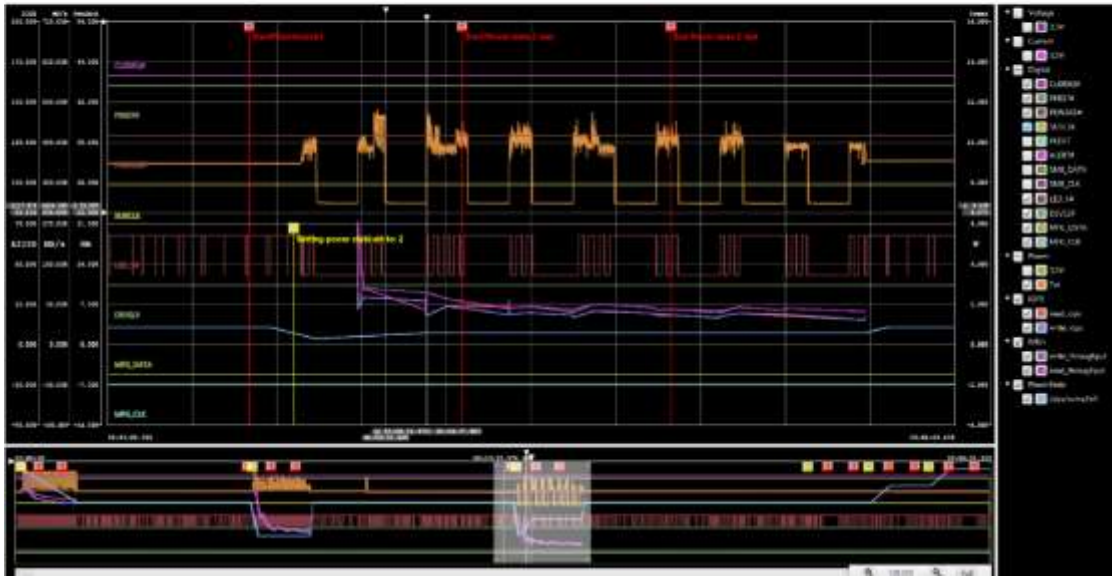


图 4-116

4.3.2 Quarch PAM 针对非标准接口的信号监测夹具

有些非标的电压、电流或者信号监测，可以采用 Quarch 2-channel PAM 夹具，该夹具也需要连接到 PAM，参见下图。采用该夹具的时候，需要将被监测信号引出到该夹具，然后再绕回到设备。例如，如果用户想监测 DDR4 DIMM 条的某些信号就可以采用这个方式。

4.3.2.1 TECHNICAL SPEC

- **2 high resolution power rails, up to 12Amps at 15V**
- **16 digital rails from 1.8v to 5v**
- **250KHz base sampling rate**
- **Fully calibrated for high accuracy measurements**
- **Simple pluggable terminal blocks for fast connection**

4.3.2.2 PRODUCT FEATURES

- **Capture long, high-resolution power traces**
- **Diagnose issues with power up timings**
- **Test power performance**
- **Capture digital traces**
- **Connect easily in to custom wiring looms**

4.3.2.3 PRODUCT DETAILS

This fixture provides a simple analysis for custom wiring harnesses.

The ideal measurement of ICs and similar. 2 power rails and 16 digital channels can be connected via the pluggable terminal blocks. This makes way for a very simple setup.

Power Analysis Module Fixtures (PAM Fixtures) are calibrated measurement devices with digital sideband capture. Developed and used for detailed investigation of the interaction between the host and device under test.

Accurate voltage and current measurements are taken on all power rails. Selected sidebands have digital capture, giving a full picture of host/device interaction.

This fixture can only be used with a Power Analysis Module.

图 4-117 PAM + 2 channel fixture 连接示意图

图 4-118 2 channel fixture 图片

图 4-119 连接好电源线缆和信号线的 2 channel fixture

4.3.3 Quarch: 如何测量 M.2 NVMe SSD 低功耗

Lee Hayashida

Posted: 12th November 2020

Quarch - Where's my sleep state? Measuring NVMe SSD low power

Laptop manufacturers are in competition to provide a combination of the best performance and the best battery life, and as part of this requirement, the manufacturers must select consumer SSDs which advertise great performance and support low power states. Benchmarks for NVMe SSD read and write performance are fairly well understood, but what exactly are *low power states* and how can these states be measured?

With low power SATA devices, HIPM and DIPM (Host Initiated Link Power Management and Device Initiated Link Power Management respectively), if supported, allow the device to go to a low power state if a certain amount of idleness occurs. If the physical pin DEVSLP is supported, an even lower power state can be achieved.

NVMe SSDs have similar mechanisms, however the SSD must support APST (autonomous power state transition) at the NVMe layer and ASPM (PCIe Active State Power Management) at the PCI Express layer. In order to achieve the lowest power state, a physical pin analogous to SATA's DEVSLP pin is the PCI Express CLKREQ# pin.

Let's examine a couple of case studies using a common platform, a common SSD but two different operating systems. The watermark that we are seeking is sub 10mW power consumption from an NVMe SSD and we can actually see this watermark achieved using either a Windows based or Linux based operating system. We will use Quarch's Power Analysis Module (PAM), Quarch's Programmable Power Module (PPM) and Quarch Power Studio (QPS) software. PAM not only measures power but also provides a useful scope representation of the SSD's sideband signals, of importance here is CLKREQ#. The PPM acts as the actual power source for the DUT and can be used to perform voltage margining, slew rate ramping and rail noise injection.

1. Windows 10 and the steps to achieve this low watermark using the PPM tool.

Notice how frequently the OS appears to wake the SSD. The SSD is unable to reside in its lowest sleep state, L1.2 for extended periods of time.

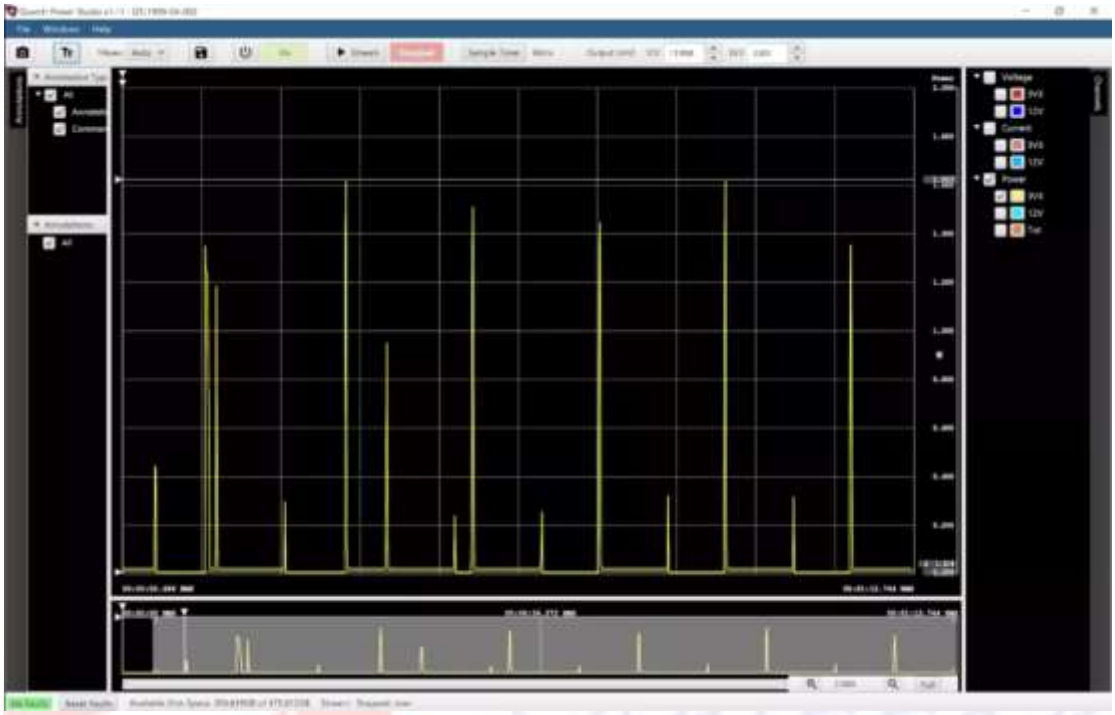


图 4-120

2. Linux (Ubuntu) and the steps to achieve this low watermark using the PAM tool.

PAM has the capability to also capture some of the sideband signals such as CLKREQ#. You can see that the same SSD in the same laptop PC in the example above, is allowed to reside for longer periods of time in its deepest sleep state under Ubuntu.

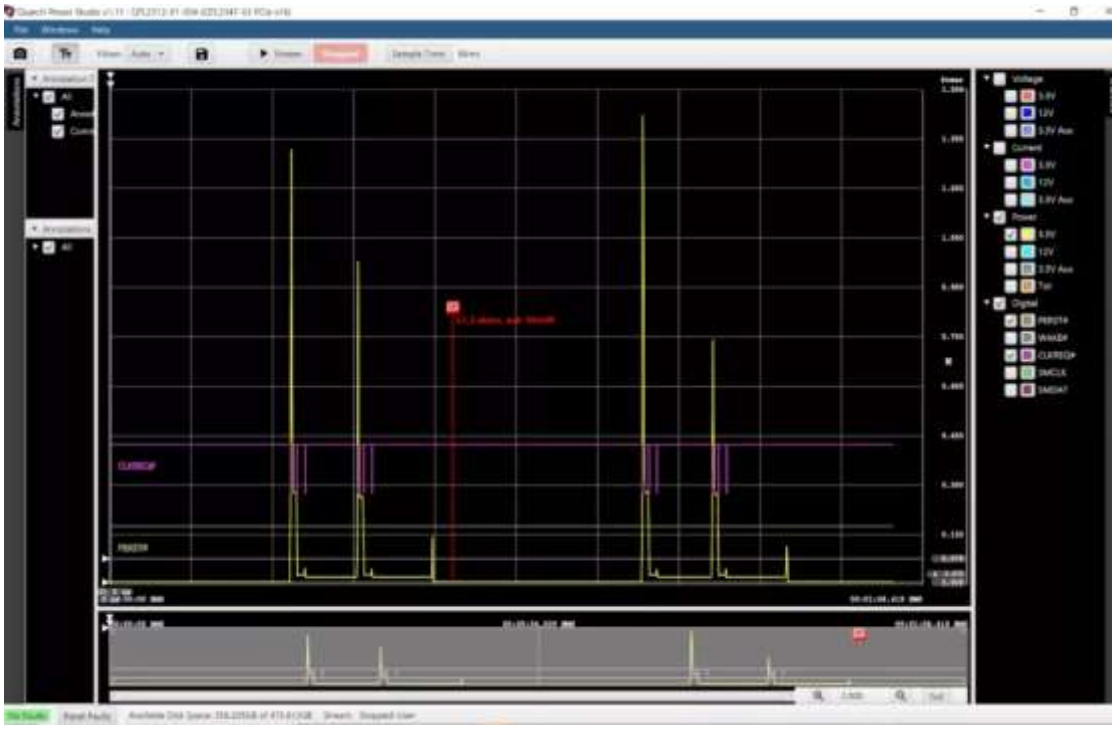


图 4-121

We have shared two examples of an NVMe SSD's ability to achieve the deepest level of sleep. A detailed explanation of the exact mechanisms of the NVMe and PCI Express software and hardware sleep transitions is best left in the capable hands of the SSD manufacturers. It was a surprise for us to see such a difference in the actual time that an SSD is allowed to reside in sleep between the two operating systems… Again, we defer to the industry experts to explain the differences.

What Quarch can deliver are tools that are easy to use, accurate, affordable and practical for nearly every firmware or hardware engineer to have at their home office or lab bench at work. The compact footprint of either the PPM or PAM allows for collecting measurements in a confined space. With Quarch's PAM or PPM, the user can make the determination if the advertised sleep state and power consumption has been achieved, it is up to the user to decide which tool is the best fit for the given job.

4.3.3.1 PCI SIG - Making the Most of PCIe® Low Power Features

https://pcisig.com/making-most-pcie%C2%AE-low-power-features

Scott Knowlton, Marketing Work Group Co-chair, PCI-SIG

PCI Express® – also known as PCIe – used to get a bad rap for being power hungry on servers and PCs. But I'm happy to say that this is no longer the case. Are you aware that PCIe today is extremely power efficient with built-in low power features? By delivering an

I/O technology that delivers high performance, low cost AND low power, PCI-SIG has ensured that PCIe is the interconnect of choice – across multiple devices, including smartphones, tablets, IoT, laptops, and more. On a mobile phone for instance, low power is a stringent requirement, and running PCIe delivers the best of both worlds – a high performance solution with low power options.

Let's take a closer look at how PCIe supports low power devices and applications.

PCI-SIG continually evolves the PCIe specification to improve performance, increase efficiency, and lower power consumption to satisfy the very divergent needs of many different applications. Since the PCIe 3.0 spec, PCI-SIG has focused on reducing power consumption while the PCIe interface is active to enable better platform power management. I'll give you a quick overview of some definitions and how they work.

The Latency Tolerance Reporting (LTR) mechanism allows the host to decide how long to wait before servicing the interrupt from the device in order to coordinate multiple devices and achieve the maximum power optimizations for the system.

- **L0 – a link which is operating normally**
- **L1 – a link state where no data is being transferred so key portions of the PCIe transceiver logic can be turned off**
- **L2 – a link state identical to L3 but in which power has not (yet) been removed**
- **L3 – when the device is powered off**
- **L0s – a link state where data may be being transferred in one direction but not the other, so the two devices on a link can each independently idle their transmitter**

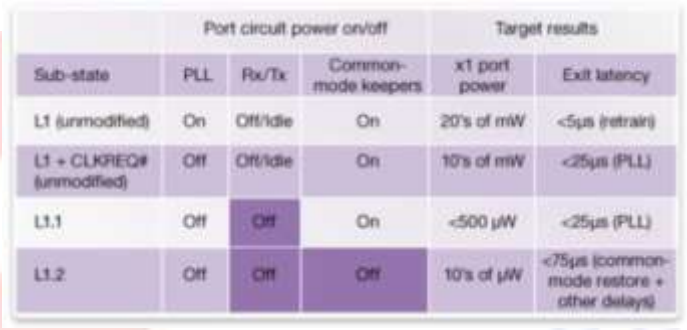
And then there are the sub-states…

As the industry evolved to more battery-powered devices such as mobile phones and other handheld/mobile devices that need to power on quickly, the focus of power management shifted from gross on-vs-off to finer-grained, moment by moment switching. It became clear to PCI-SIG that for these applications, L2 resume latencies were too high to allow its use for this rapid and frequent state switching, while L1 power savings were too low to meet the device power consumption goals. An innovative solution to this conundrum came in the form of what we call the L1 sub-states.

The fundamental idea behind L1 sub-states is to use something other than the high-speed logic inside the PCIe transceivers to wake the devices. The goal is to achieve near zero power consumption with an active state.

That is done by adding additional functionality to an existing PCIe pin (CLKREQ#) to provide a very simple signaling protocol. This allows the PCIe transceivers to turn off their high-speed circuits and rely on the new signaling to wake them up again. In fact, two of these new sub-states were defined: L1.1 and L1.2 providing their own power vs. exit latency trade-off choices. Both L1.1 and L1.2 permit the PCIe transceivers to turn off their PLLs along with their receivers and transmitters, while L1.2 even allows turning off the common mode keeper circuits.

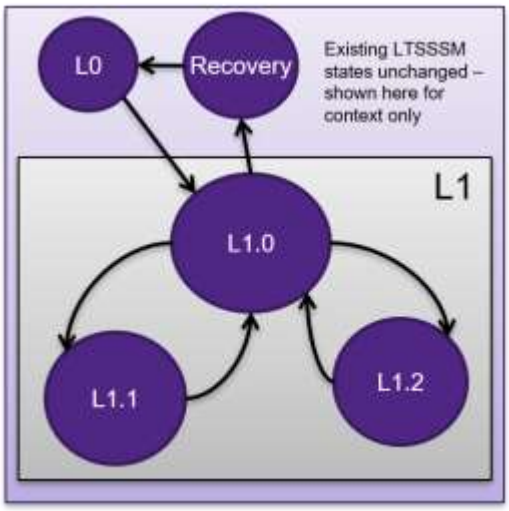
The results are dramatic. Efficient circuit design and modern silicon processes mean that a representative PCI Express 4.0 x4 PHY (4 transceivers plus related digital logic for four lanes) running at the full 16GT/s data rate in L0 consumes somewhere in the range of 400-500mW. Utilizing L1.1, the same PHY's power consumption drops by a factor of around 20x to consume only 20-30mW. Accepting the slightly longer exit latency of L1.2 permits power consumption to fall by another 10x to a mere 2-3mW.



	Port circuit power on/off			Target results	
Sub-state	PLL	Rx/Tx	Common-mode keepers	x1 port power	Exit latency
L1 (unmodified)	On	Off/Idle	On	20's of mW	<5µs (retrain)
L1 + CLKREQ# (unmodified)	Off	Off/Idle	On	10's of mW	<25µs (PLL)
L1.1	Off	Off	On	<500 µW	<25µs (PLL)
L1.2	Off	Off	Off	10's of µW	<75µs (common-mode restore + other delays)

图 4-122

The figure below shows the low power solutions available with the existing L1 state compared to using L1 sub-states. It is expected that the power savings scale linearly for multi-lane links and implementing the L1 sub-states feature reduces power consumption at the increase of the L1 exit latency. Implementing L1 sub-states is key to reducing power consumption for mobile designs using PCIe.



	Port Circuit Power On/Off		
Sub-State	PLL	Rx/Tx	Common-Mode Keepers
L1.0	On	Off/Idle	On
L1.0+CLKREQ#	Off	Off/Idle	On
L1.1	Off	Off	On
L1.2	Off	Off	Off

图 4-123

Table 1: Comparison of proposed solutions

The low exit latencies and tremendous power savings of the L1 sub-states feature, combined with PCI Express' load/store architecture and upcoming 32GT/s speed provide the optimal interface for use in mobile devices, storage, compute acceleration, networking and other high-speed devices well into the future.

Learn more about Low Power Features through our new video series on the *PCI-SIG YouTube channel.* I also encourage you to learn more about PCIe at *www.PCISIG.com*, and to stay up to date on all the latest PCI-SIG developments by following us on *Twitter* and *LinkedIn*.

4.3.4 Quarch: 为什么 PAM 量测的数值和你自己量测的好像不同？

https://quarch.com/news/accurate-power-measurement/

有时候用户可能觉得好像使用 PAM 测试的数值和自己采用 Keysight，Keithley 或者示波器测量的数字不一致。根据我们之前的经验，一定**要确保测量的时候将两种仪器同时接入链路而不是分开接入，否则测试提交不一致将导致无法比较**。另外，采用不同工具测试的时候由于**不同的通风散热条件也可能会导致测试的数值不一致**。下面是 Quarch 公司的一篇文章，考虑到专业性，保留英文未翻译。

4.3.4.1 Why does my power trace look different from yours?

We've recently had a few interesting requests from customers who have seen measurements taken on their scope that did not look the same as those taken via the

Quarch tools.

The two traces of an HDD spin-up do show some similarities in terms of timing, but the shape of the main region clearly looks very different.

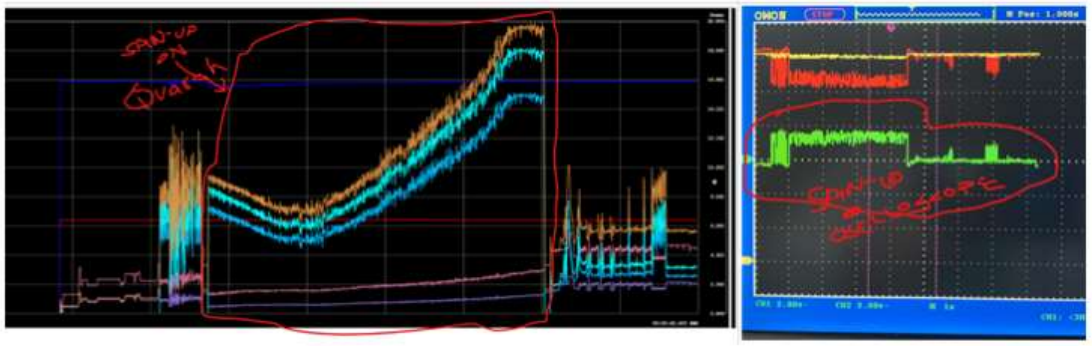


图 4-124 Images sent from our customer

The first thing to do is to understand how the two traces have been measured. In this case we confirmed:

CUSTOMER'S QUARCH SETUP

Quarch SFF injection fixture attached behind drive

4uS sampling rate, with 1024 samples being averaged to give one power sample every 4.096mS

Recorded in Quarch Power Studio

CUSTOMER'S SCOPE SETUP

Power resistor attached in series with the +12V signal

20 MHz Bandwidth probe set

2x voltage probes measuring the voltage drop to calculate current

(A second setup using a current probe was used later, showing similar results)

Quarch Power Analysis

Quarch power tools come in two main ranges, the PPM (which supplies power) and the PAM which monitors the host power. Both use simple plug-and-play fixtures for fast setup and allow capture from both Quarch Power Studio or via simple scripted automation:

https://quarch.com/solutions/power-analysis/

Sampling and Averaging

This immediately shows an issue: The Quarch trace is averaging many samples together, producing a smoother waveform. When comparing power traces, it is important to choose an appropriate averaging rate and apply them to both traces.

OCP Example

The Open Compute Platform (OCP) has specific requirements when it comes to power analysis of storage devices

One requirement is for device peak power to be measured at:

4uS base sampling rate

Analysed with an averaging window of 100uS

Many Quarch customers have similar requirements, and while the exact rates and averaging windows may differ, the idea is the same: To produce a realistic value of power consumption for an SSD or HDD, such that different devices and workloads can be compared.

Making the comparison

We set up a Quarch Power Analysis Module and an Oscilloscope and Current probe up in parallel, so we could take both measurements at the same time for comparison, setting both to similar sampling rates (as close as we could get). We used a 2-channel PAM fixture and external ATX PSU to supply the drive, to be as close as possible to the customer setup.

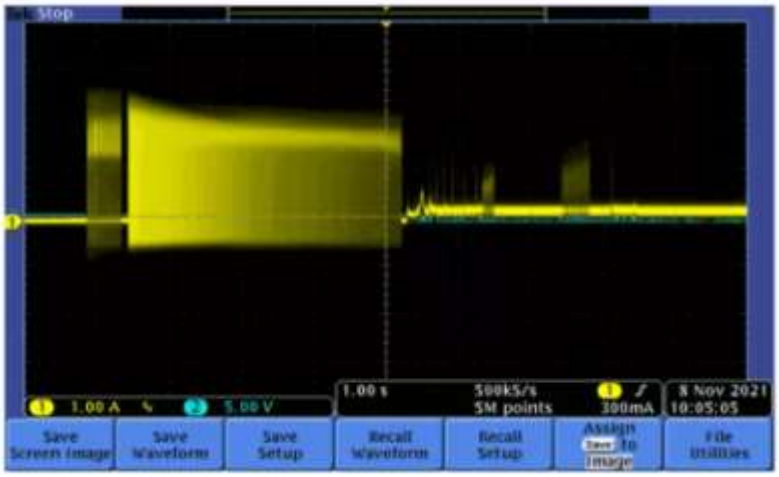


图 4-125 Scope trace capture at 2uS sampling rate

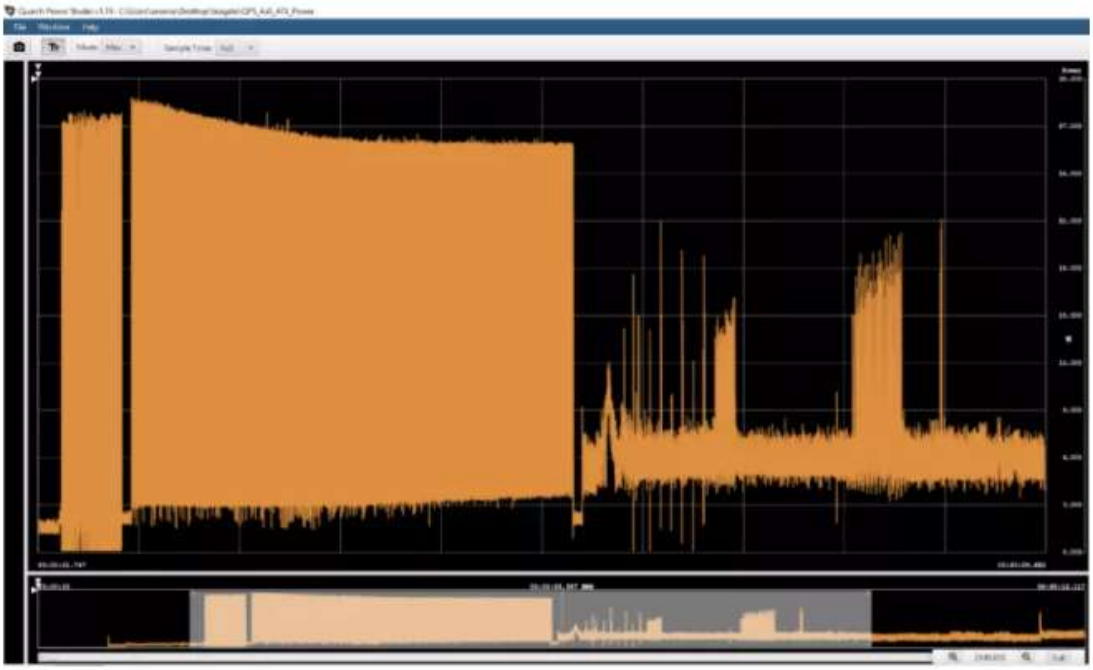


图 4-126 Quarch 2-channel PAM capture at 4uS sampling rate

图 4-127 Quarch 2-channel PAM capture at 4uS then averaged to 10mS

The issue was the shape of the spin-up profile. If we run the Quarch tools at a similar sampling rate, we see the same profile BUT when we average this out, we can see the underlying trend is different.

Zooming in

When we zoom in on the trace in Power Studio, we can see the power trace (orange) is very noisy during spin-up and jumps between around 1 and 20 watts

Let’s zoom in on the higher rate capture and look more closely:

At the start of the spin-up, we see a lower power consumption, with a few spikes up to the higher level

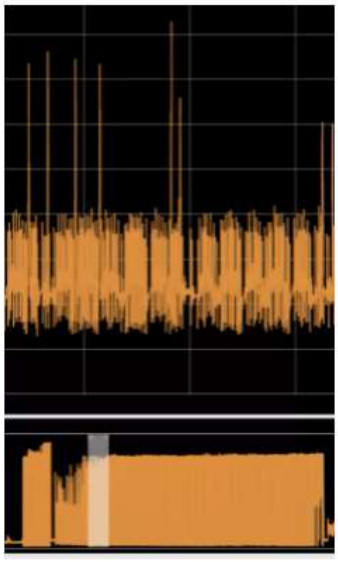


图 4-128

Zoom in at the start of the spin-up

In the middle of the spin-up, we can now see larger and longer spikes

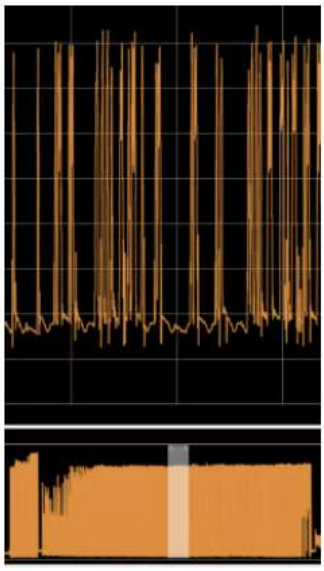


图 4-129

Zooming in to the middle of the spin-up

Now at the end of the spin-up period, we see that the power is mainly in the high state and only dropping intermittently to the lower level.

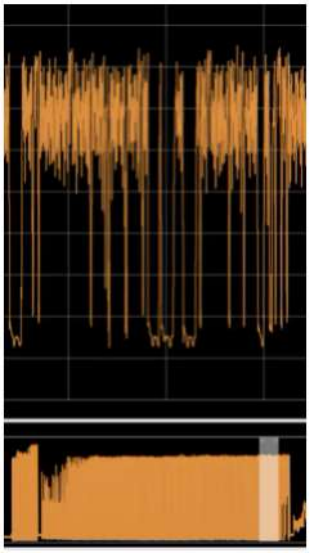


图 4-130

Now showing the end of the spin-up period

This explains the main difference. The high resolution one at 4uS is showing the rapid transitions of power draw during the drive spin-up, while the 10mS capture is giving a general trend, showing that power consumption increases during the process.

Resampling the scope

To prove the data is comparable, we downloaded the scope data to CSV (a painful process requiring a legacy USB stick and additional tools)

We then wrote a python function to resample the data to the same 10mS averaging rate as used on the Quarch tools. This was necessary, as the scope did not support sufficient internal averaging to do it internally.

This is current data, not power, but we see the same clear trends across the full spin-up period.

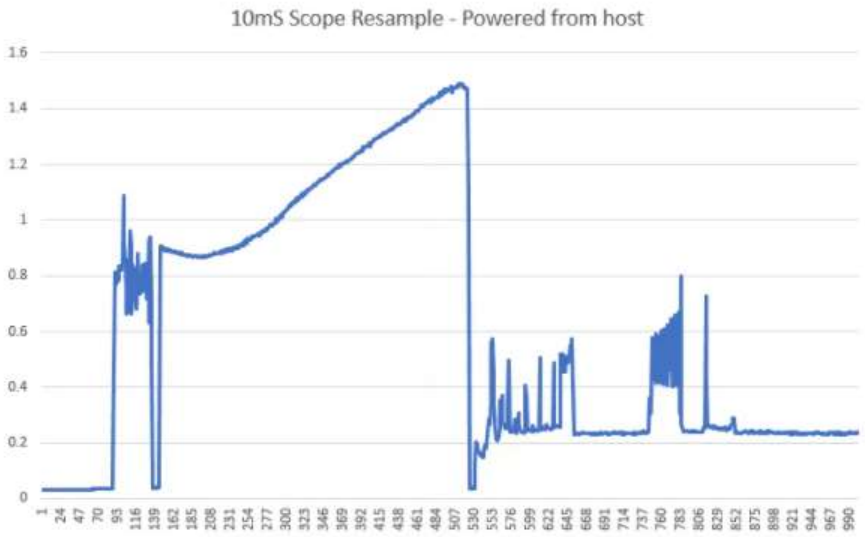


图 4-131

Conclusions

It is easy to look at a trace (especially a static capture) and not be aware of the underlying detail. In this case, the drive is drawing power in a modulated fashion, similar to a PWM waveform. This was not clear from a zoomed out scope trace.

The ability to perform windowed averaging and easy zooming is essential to understand what the device is doing.

Ensuring all capture sources use the same averaging window is essential if you want to make a valid comparison

Setting up and capturing with the scope was ‘much’ harder than with the Quarch tools. It required a custom cable that the current probes could attach to and while saving traces to a USB stick was possible, it was slow and prone to error.We ended up using a phone to take pictures of the scope screen as it was easier! With over 20 comparisons during testing, this was a pain to track.

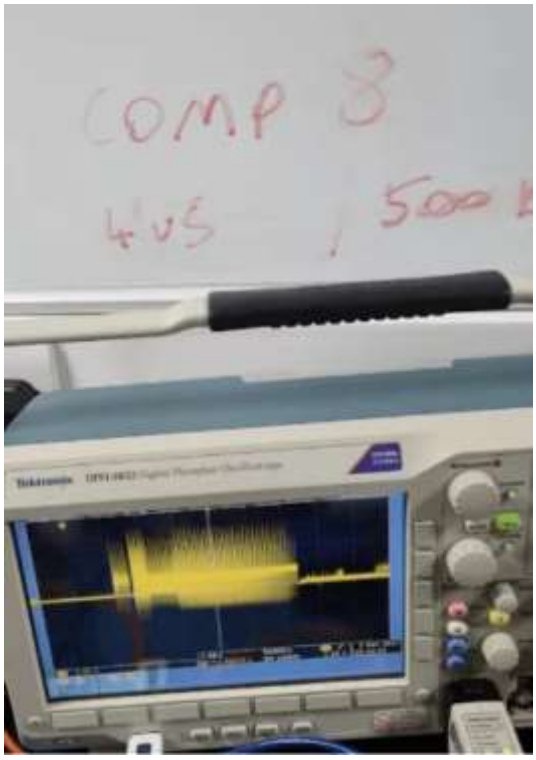


图 4-132 Test #8 – Tracked with a whiteboard!

4.3.5 针对主机等三相 AC 交流 PAM 分析模块

- AC PAM takes us into the mains world
  - AC measurement up to 600V and 63A (3-phase + netural)
  - Long term, high-resolution capture
  - Single phase IEC version coming later in 2022

图 4-133

Making a coffee, RMS power at 125uS resolution

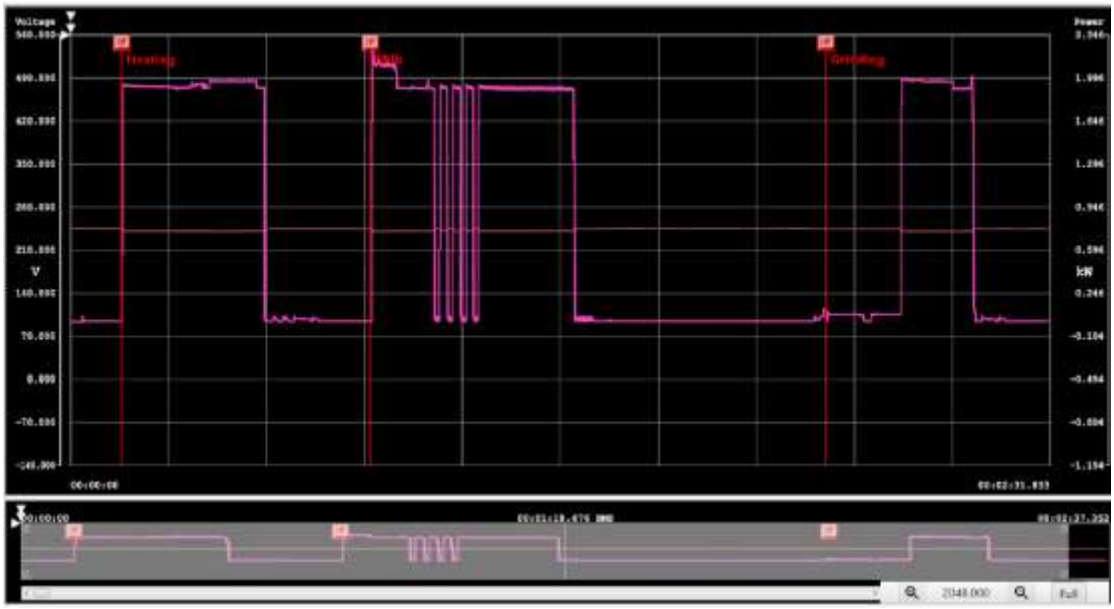
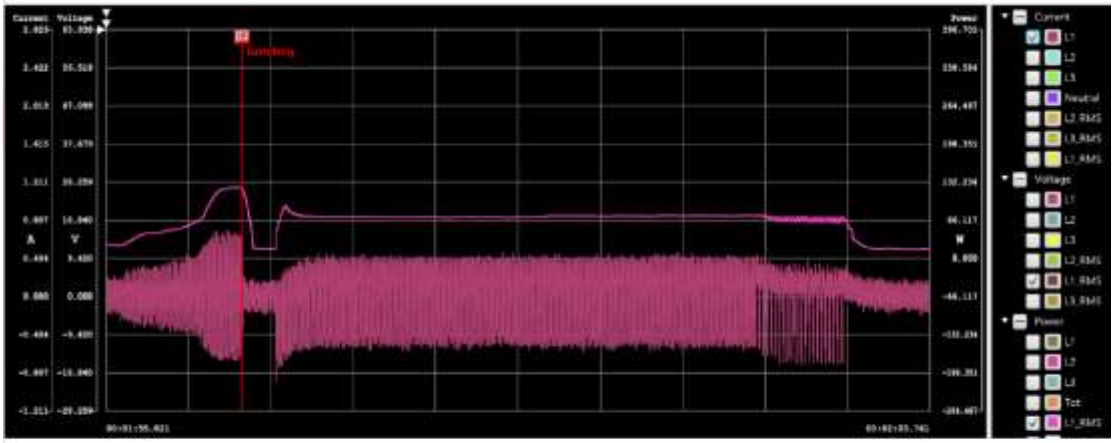


图 4-134

Zooming in to the coffee grinding process



- High resolution 8k samples/second
- Detailed view of power consumption changes
- Same Quarch tools and automation is available

图 4-135

- Power loss on a compressor
  - Voltage trace (purple) ramps down and reduced in frequency (compressor back EMF)
  - A poorer design could result in a large voltage spike here, but would be hard so see without this tool

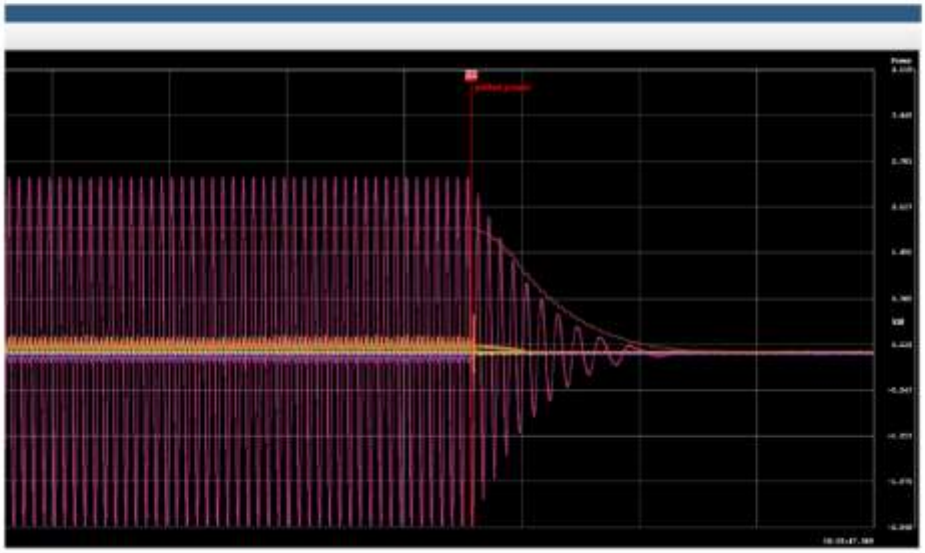


图 4-136

4.3.5.1 How to analyze an EV charger

Kyle McRobert

Senior Hardware Engineer

After graduating from Heriot-Watt University in Edinburgh with a BEng(Hons) in Electrical and Electronic Engineering. I joined the Royal Air Force which took me across the world for two years to places like Bahrain, the Falklands and Germany (to name a few).

Keen to put down roots, I joined Quarch in 2018 as a Hardware Engineer. I enjoy working closely with our customers to help create custom setups. My goal is to help them get the most out of our products.

As if I'm not kept busy enough, I'm still in the Royal Air Force reserves as a Police Dog Handler. Oh, and I've successfully trained as a firefighter.

Posted: 21st September 2023

4.3.5.1.1 Setting up

EV charging is a complex process. With many different vehicles and chargers, we would expect to see a range of compatibility issues and charging speeds.

Understanding and solving problems requires data, and this is where the 3-phase AC PAM (Power Analysis Module) comes in. The 3-phase AC PAM (Power Analysis Module) plays a crucial role in understanding and solving these problems. This module is designed to collect and analyze data related to the three-phase alternating current (AC) power, providing valuable insights into the performance and efficiency of the power system.

This device can capture high-resolution AC traces for long periods of time.

-
- 8,000 Samples per second
- 16, 32 and 63-amp versions
- Plug-and-play setup
- Manual and fully automated capture options

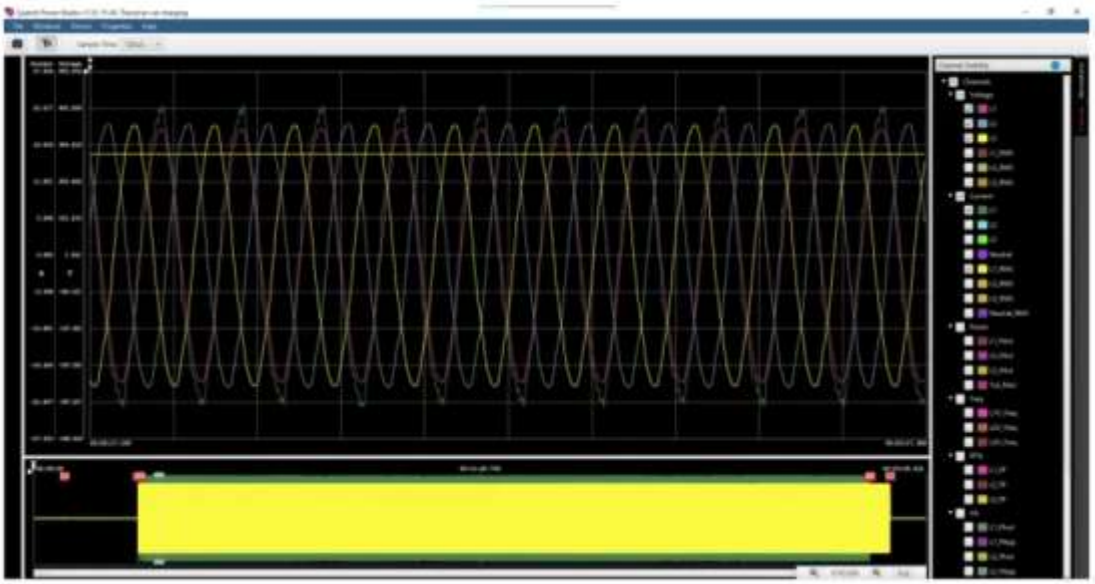
In this instance, we used an EV and plugged the AC PAM in series with the charging cable.

We’re using the 16Amp version of the product here, which matches the rating of the charger (An evolt charger connected to a 16A, 3-phase supply)

This takes just a few seconds to set up, though I did use a custom that bypasses the PP/CP as without this, the EV could not communicate with the charger.

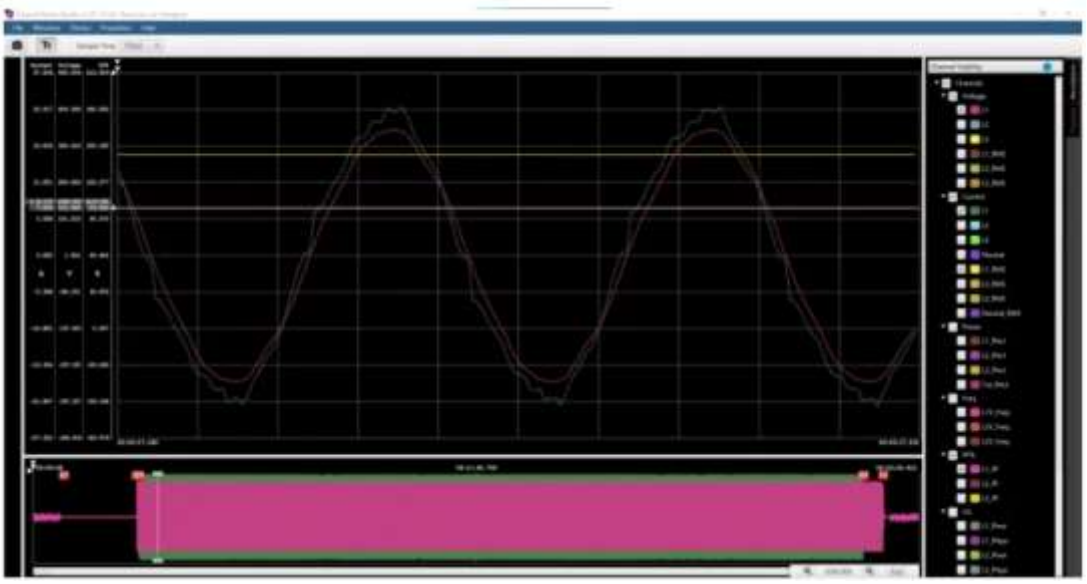
The laptop is running Quarch Power Studio, our capture/visualisation software

4.3.5.1.2 Output voltage and current



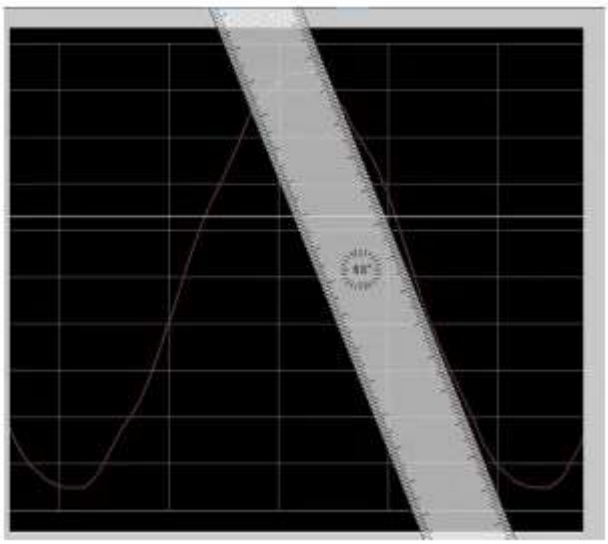
The first thing we see is that the charger has the ability to switch its output on and off; it does not output a voltage until it communicates with the car and begins the charge cycle.

We also see above that voltage is applied on all 3 phases, but the EV only takes current on phase 1.



Zooming in further, we can see the current waveform is pretty good, and adding the ‘power factor’ channel gives us a 99.4% power factor during this point in the charge.

Interestingly, the AC waveform from the charger is not perfect.



There is a clear distortion, and this is seen all the time, even when no significant current is drawn. Additional testing would be needed to confirm if this is an artifact of the charger or if the AC supply to the entire site shows the same effect!

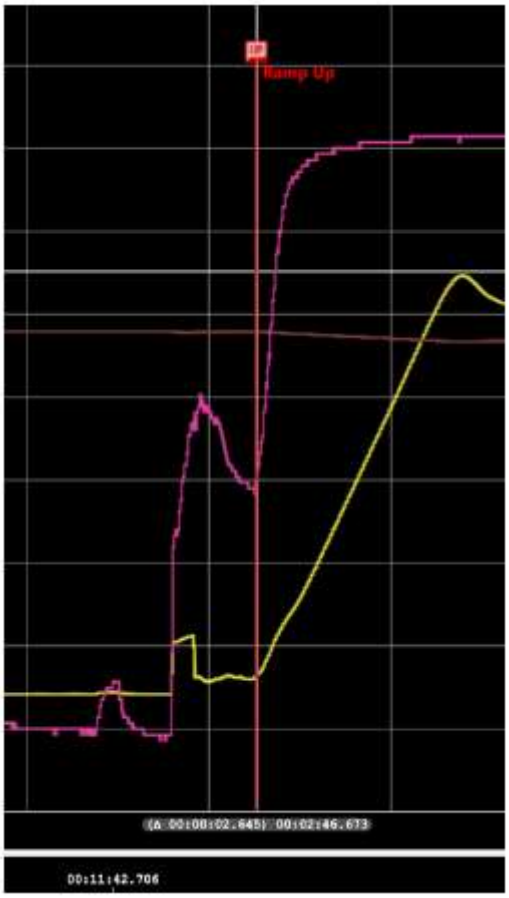
4.3.5.1.3 Starting the charge cycle

Moving on to the charge cycle, lets look at the start-up

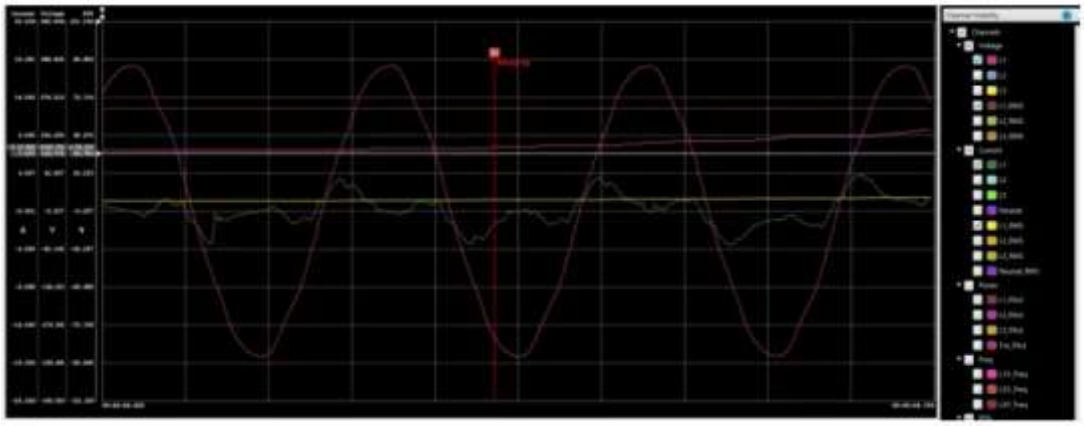
The brown trace is the RMS voltage, which rises rapidly at the start of the charge and remains steady until the very end. We do see a small voltage drop as the current draw ramps up, though (annotated in yellow). This is just over 6 volts drop.

The yellow trace is RMS current. We see an inrush spike to 6.5A followed by a couple of seconds of delay before we ramp up to the 15.1 Amp charging current that remains steady for the full test.

The purple trace is Power Factor. This is invalid at the start, as no voltage is applied, but we do see an interesting step as the current ramps up:

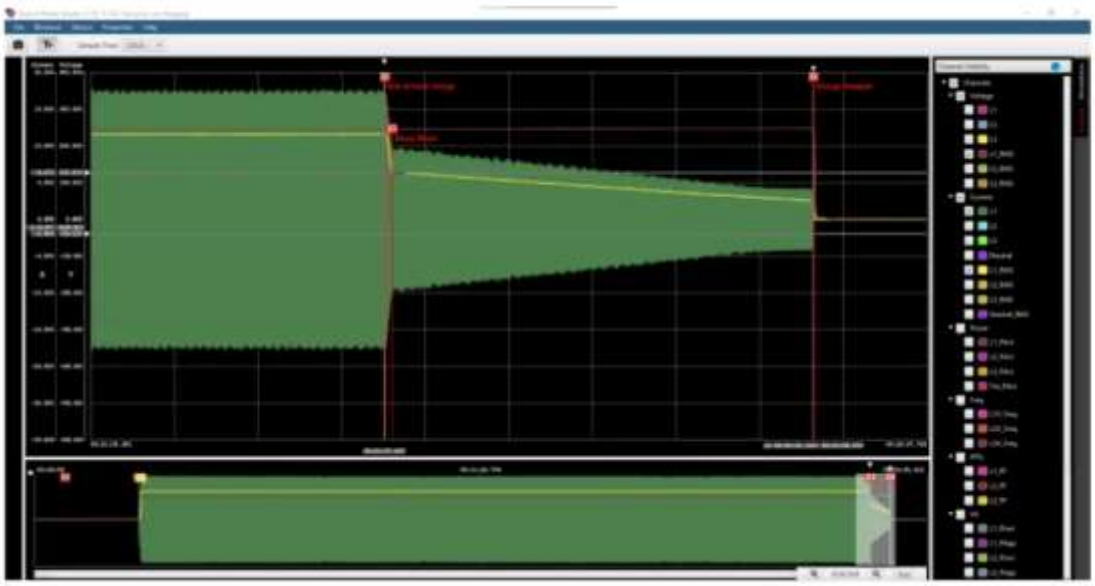


The (green) current trace is not at all clean at this point, which would explain the poor power factor



4.3.5.1.4 End of charging

The battery was already well charged, so the cycle only ran for about 20 minutes. Towards the end of this time, the charger slows the charge across a 32-second period:



The current waveform is in green, and the RMS current is in yellow. We can see it first drops from 14.9A to 8.2A over 486mS. Next, it steps down slowly in 150mA increments until it reaches 3.3A. It then turns off completely.

4.3.6 针对 IEC 220V 单相 AC 供电 PAM 分析模块

4.3.6.1 Quarch QTL2843 IEC Mains Power Analysis Module Review

written by Brian Beeler October 31, 2023

The Quarch QTL2843 IEC Mains Power Analysis Module allows for extreme visibility into the power consumption of just about any Ac-powered device. In the context of our lab, that means we can look at a range of devices from workstations to servers, to get a more precise view of not just overall power consumption, but the impact of key devices like GPUs and accelerator cards, under load.

This data gives our reviews more depth and provides a more complete picture of the ongoing cost to operate these high-power machines. The Quarch QTL2843 supports all major US/EU/ROW voltages and frequencies and is essentially plug-and-play. The device sits in between the system under test and your PDU, capturing all of the consumption data along the way with a refresh time of up to 125 microseconds.

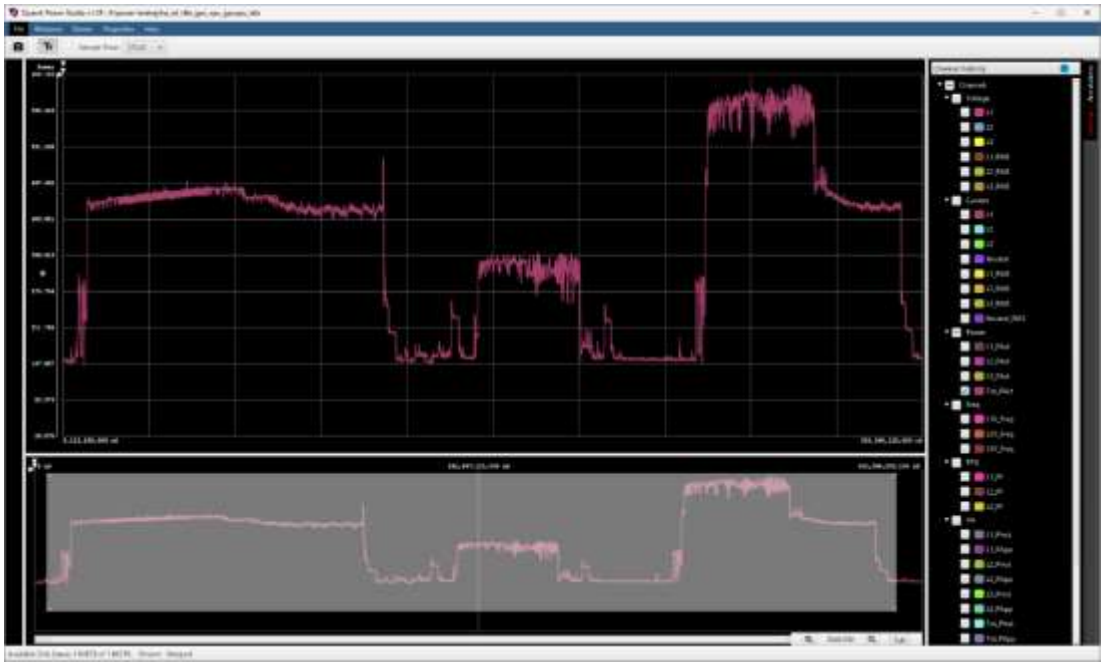
4.3.6.1.1 Quarch QTL2843 IEC Mains Power Analysis Module Specifications

- Support up to 400v AC
- Supports 40-60Hz (and beyond)
- Fused at 10 Amps
- PoE powered or via 12v PSU
- USB and LAN connectivity
- 8KHz sampling rate
- Fully calibrated

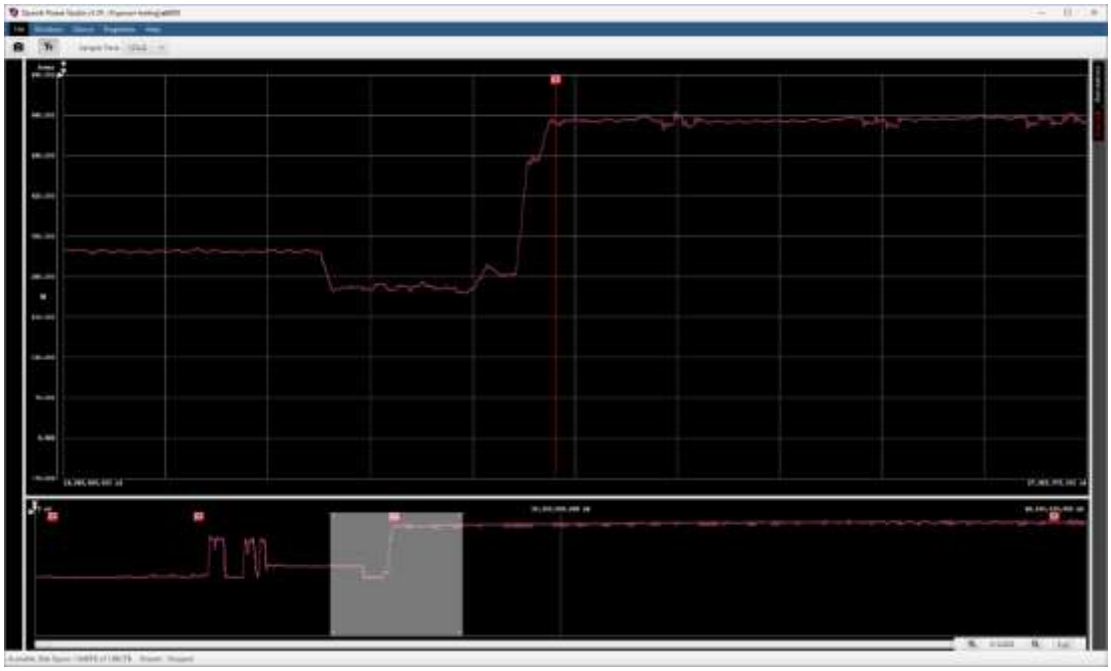
4.3.6.1.2 Power Testing Scenarios

With the analysis module comes the Quarch Power Studio, which provides visual insight into power consumption. The software can capture data over extended periods of time and can easily be mapped to system activity. We've used Power Studio in a few reviews already to help us understand workstation power consumption. Let's explore a few of those use cases to better articulate the value of the Quarch QTL2843.

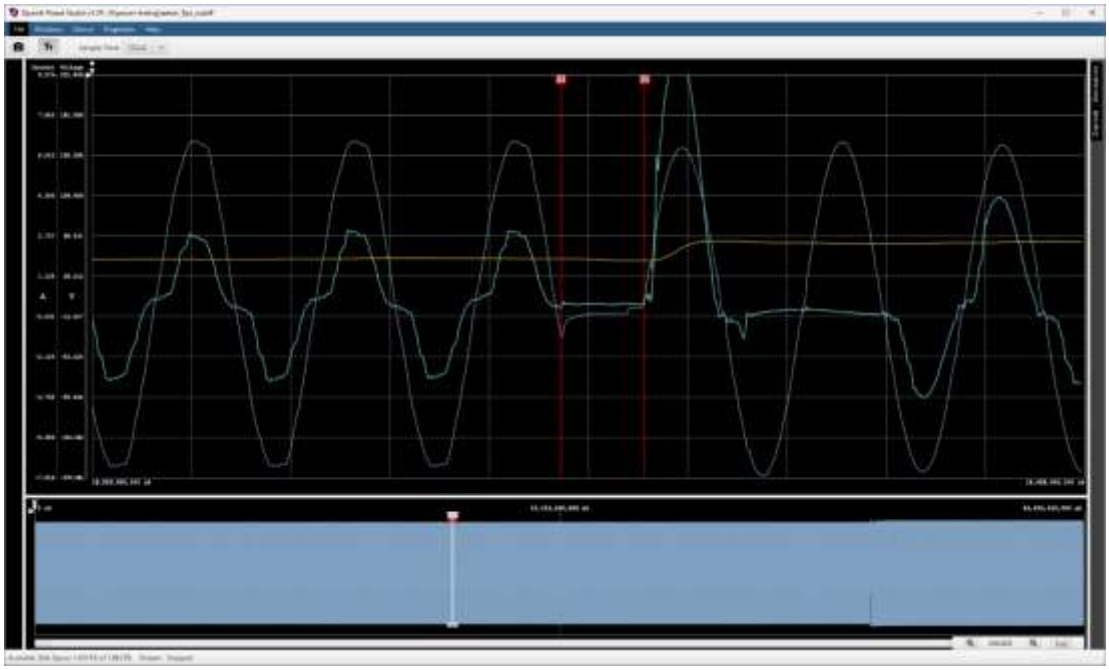
The Quarch Mains Power Analysis Module has been leveraged in our recent review of the HP Z4 Rack G5 and TYAN Transport HX FT65T-B8050. The software allowed us to capture the baseline idle power, GPU load, CPU load, and GPU and CPU combined.



With a precise measurement offered, we were also able to leverage it in that same system when we compared the NVIDIA A6000 head-to-head against the NVIDIA RTX6000 Ada graphics card.



Another area beyond power draw testing that the Quarch unit has proven itself to be quite useful, is in how we test data center UPS hardware along with emerging portable power stations in AC-output quality and response times during power failure situations. Here the 125-microsecond sample time helps out quite a bit to measure how fast devices are able to switch between line current and internal battery power.



4.3.6.1.3 Conclusion

Expect to see continued use of the Quarch QTL2843 IEC Mains Power Analysis Module in upcoming reviews. Power usage plays a huge role in data center operation as it has a direct impact on operational costs and cooling impact. The QTL2843 will help us

measure the impact at a system level across a wide range of systems, as well as how individual components can play a role in shaping system power usage.

While this is still early days for our power testing capabilities, we're now able to at least answer some of the basic questions around system power consumption. As we scale up to larger systems, we will need a bigger boat, but for now, the Quarch QTL2843 is a reasonable and capable starting point.

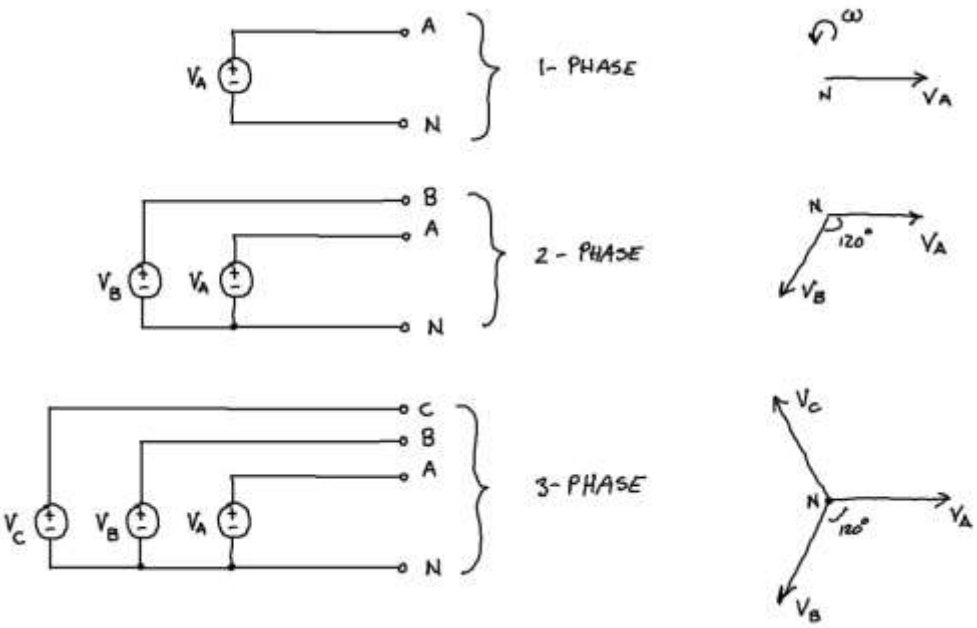
IEC / DATA CENTER POWER CORDS

IEC 60320 is the international standard set by the International Electrotechnical Commission (IEC) that is used by most countries. The standards sets the specifications for power cords up to 250 volts.

SIGNAL+POWER has a wide variety of IEC data center power cords in various colors, lengths, and connectors. IEC 60320 power cords uses even numbers for plugs and odd numbers for receptacles.

Single phase power has two wires; an active and a neutral. It supplies power at around 240 volts and is used in homes and businesses for most appliances and lighting. 3 phase power has four wires; three actives and one neutral, and supplies power at both 240V and 415V.

What is the difference between single-phase two-phase and three-phase?



Single phase is just one active phase (1 active conductor) & neutral, double phase is 2 phases (2 active conductors) & neutral & 3 phase (3 active conductors) can be 3 phases with or without a neutral conductor.

Three-phase power is a three-wire ac power circuit with each phase ac signal 120 electrical degrees apart. Residential homes are usually served by a single-phase power supply, while commercial and industrial facilities usually use a three-phase supply.

4.3.7 使用 PAM 分析 GPU/AI 卡/FPGA 加速器功耗

GPU/AI 卡功耗分析：PCIe 插卡的功耗不断增加，许多现代设备需要辅助电源。高功耗意味着高运营成本、高热量输出以及对大电源的需求。了解和优化功耗可以对生命周期运营成本产生巨大影响。最新一代的 GPU、AI 加速器、FPGA 加速器等可以受益于 Quarch 高质量 PAM 功耗分析测试，而该设备使一切变得简单。

功率分析模块夹具（PAM 夹具）是具有数字边带捕获功能的校准测量设备。允许对主机和被测设备之间的交互进行详细调查。

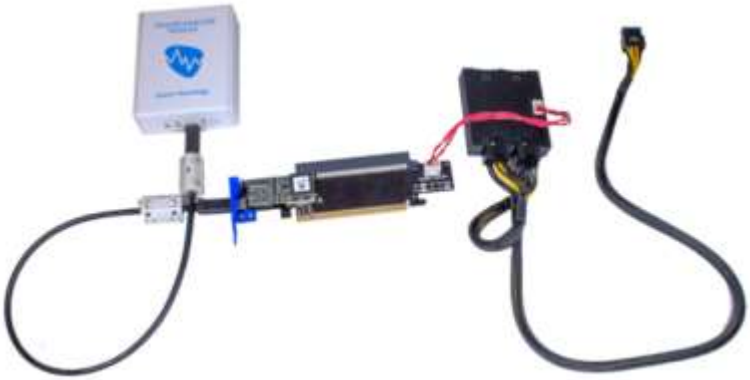
该夹具可轻松对 x16 PCIe 设备进行功率性能和边带时序分析，运行速度高达 Gen5。PCIe 设备上的不同 AUX 电源布局将提供多种选项。

在所有电源轨上进行精确的电压和电流测量，并且选定的边带具有数字捕获功能，从而全面了解主机/设备交互。（现在包括 REFCLK_LOSS）

高分辨率、长时间记录以及可视化分析软件和简单的 Python 自动化的选择使这成为一个非常强大的产品

该装置需要 PAM 控制器才能运行。

下图是 PAM 和外供电源的连接方式，以及产品的一些细节照片和安装环境照片。



4.3.8 PAM 和 PPM 的主要功能区别

4.3.9 功耗测量：我应该选择什么采样率设置 PPM/PAM？

Power measurement: What sample rate do I need?

Andy Norrie

I graduated from Heriot-Watt University with a BEg in Computing/Electronics. Since then I have worked across multiple industries from test engineering at Agilent to oil well analysis at General Electric.

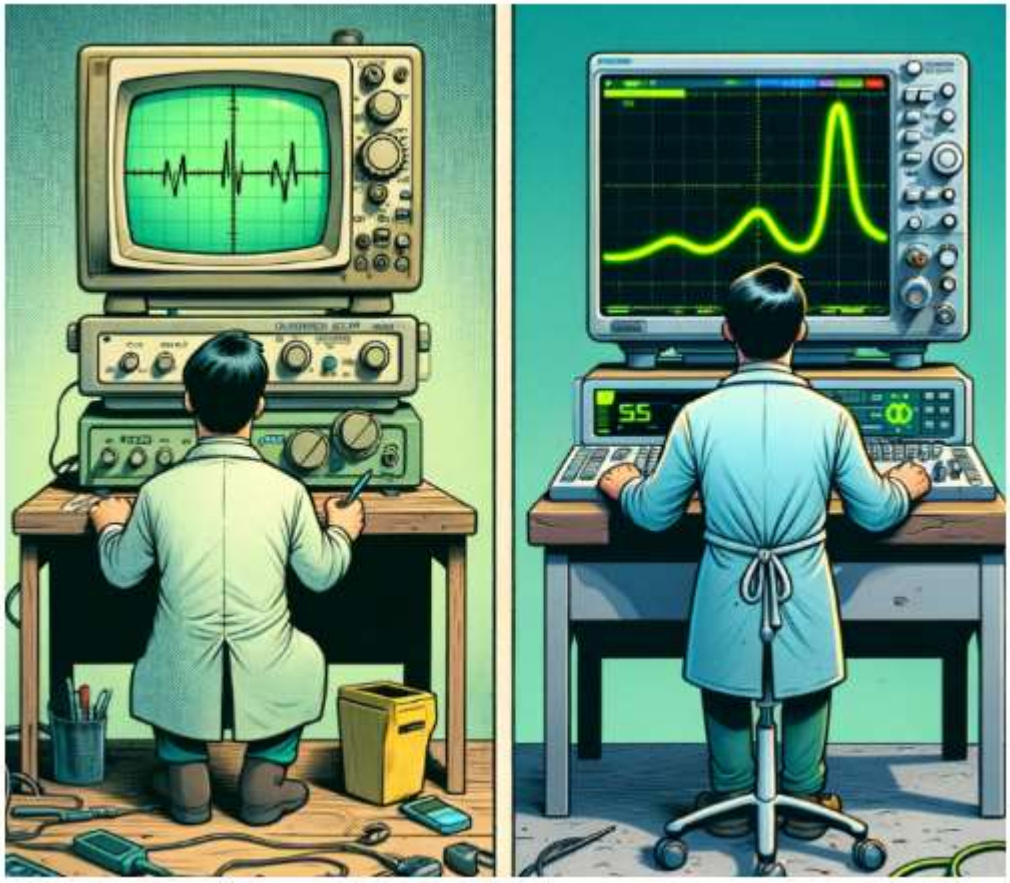
I joined Quarch in 2008, as the first employee and co-director. Initially, I was responsible for the early firmware and software, much of which still underlies our products today. Now the Operations Director, I split my time between general management, customer relations and software architecture.

Posted: 2nd February 2024

It's very easy to get caught up with how 'fast' a scope or power analyser can go, but how important is the sample rate? Is faster always better? Will newer/faster equipment always give you a better result?

Running faster allows us to resolve smaller details and more accurately measure brief spikes. The downsides include a more noisy capture, which can hide the general trend, and a much larger amount of data.

Quarch has been contributing to industry specifications on power measurement, and we thought it would be interesting to look at why certain sample rates are chosen.



How fast do you need to go?

Sample rate comparison

Let's start by looking at industry specifications for power measurement. The OCP (Open Compute Platform) among others, specify a 'moving window average' for some power measurements. This is used to find the maximum power a device consumes within a workload.

The OCP spec gives two window widths, 100uS and 100mS. In each case, we move the window over the data, averaging each point within the 'window', so we give a single value. We look for the window with the highest value, and this is the worst case that we report.

Averaging over a window acts as a filter, reducing the effect of noise and small spikes. It gives a value that is meaningful in terms of the power supply needed for the device.

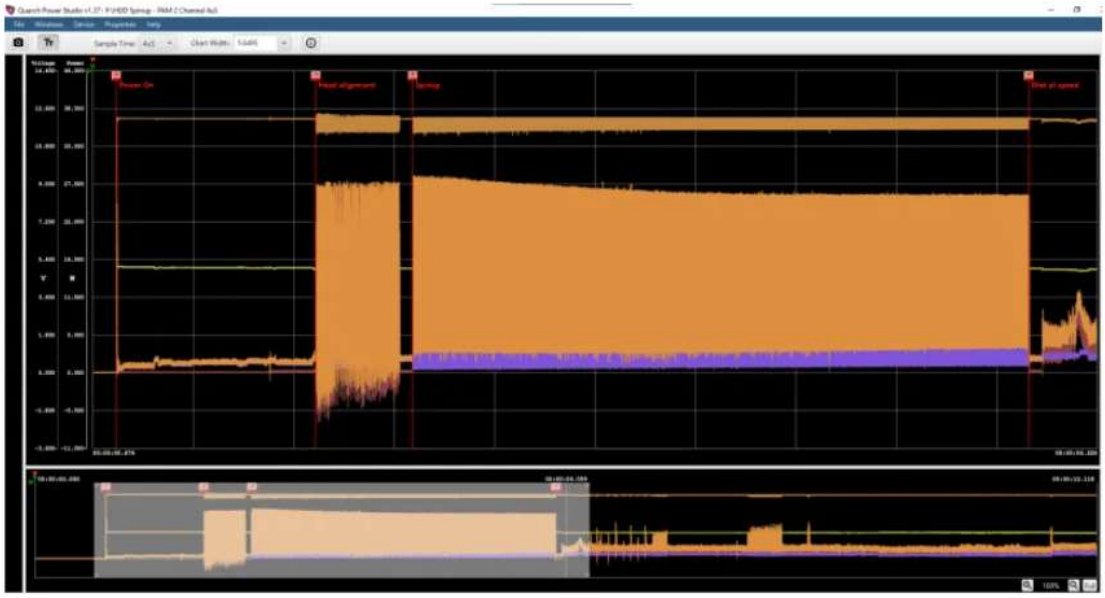
For a 100mS averaging window, we need to sample multiple times within the 100mS window, but how many times?

Introducing our data

For the test, we need something that changes in power consumption a lot over the trace. If the load is steady, almost any sample rate will give the same result, so we need to make a realistic test.

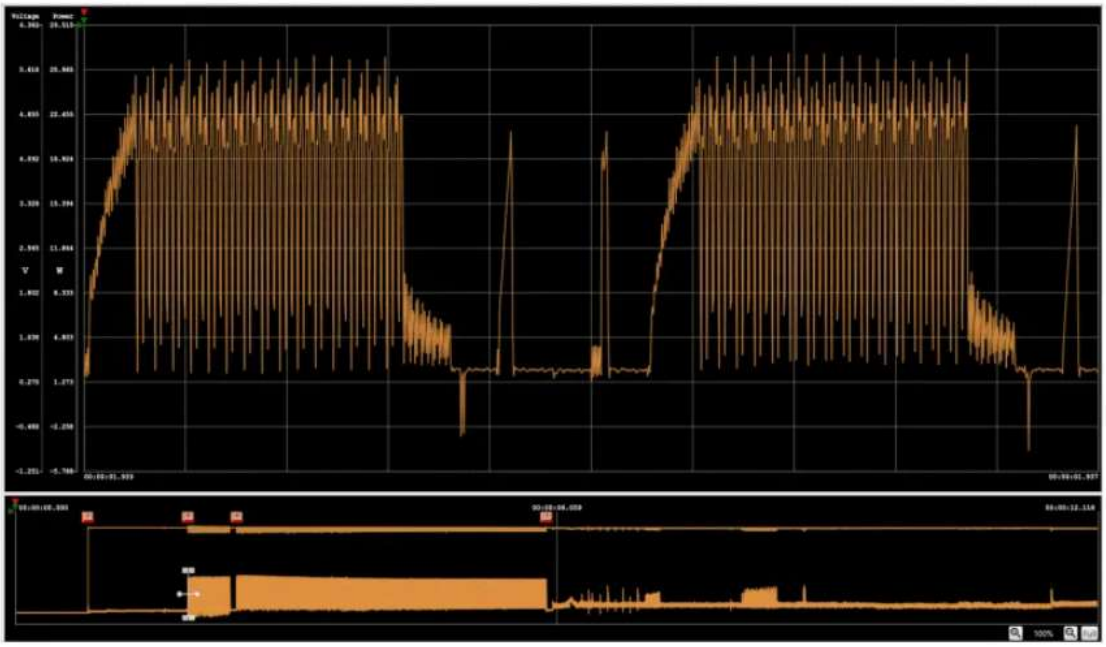
This is a power up capture of a modern enterprise HDD. The full trace is about 12 seconds long and includes a small time with the device off, then the full spin-up sequence until it becomes ready on the host. The trace was captured using our PAM range of power analysis tools and Quarch Power Studio (QPS).

The data was captured at 250 KS/s (4uS between samples) which is the maximum sample rate our device can handle and the rate set in the OCP spec.



Enterprise HDD Spin-up Capture in Quarch Power Studio

The 12V channel is the one we are interested in, as it drives the motor. During spin-up, the current profile jumps rapidly between several different levels. This means we risk an inaccurate measurement if we do not sample fast enough



4mS Slice, showing the motor current profile during spin-up

Processing the data

I used the export feature of QPS to dump the entire trace into a large CSV file of around 275MB.

Next I wrote a quick Python script to remove every second sample from a CSV file. This is a crude but effective way of halving the sample rate. By doing this multiple times, we end up with a set of files, each at half the sample rate of the previous.

This is similar to capturing the spin-up at different sample rates, but has the advantage that we remove the uncertainty that comes from each power-up being different. Here we are working with the same raw data set, which will give a much clearer comparison.

OCP 100mS peak power test

Starting at 4uS sampling (250 KS/s), I worked down to 4096uS (~244 S/s). I then used a second script to run a 100mS averaging window over the data and report the worst case for each sample rate

	A	B	C	H	I	J	K	L	N
1									
2					OCP 100mS windowed peak				
3	Sample Rate	Frequency Hz			Peak uW	Peak mW	Error mW	Error %	
4	4us	250,000.00			18968537.12	18,968.54	0	0	
5	8us	125,000.00			18958599.07	18,958.60	9.94	-0.052%	
6	16us	62,500.00			19008356.52	19,008.36	39.82	0.210%	
7	32us	31,250.00			19031249.7	19,031.25	62.71	0.331%	
8	64us	15,625.00			19099941.56	19,099.94	131.40	0.693%	
9	128us	7,812.50			19242000.03	19,242.00	273.46	1.442%	
10	256us	3,906.25			19403352.93	19,403.35	434.82	2.292%	
11	512us	1,953.13			19464960.49	19,464.96	496.42	2.617%	
12	1024us	976.56			19494460.99	19,494.46	525.92	2.773%	
13	2048us	488.28			20025597.54	20,025.60	1,057.06	5.573%	
14	4096us	244.14			20832846.46	20,832.85	1,864.31	9.828%	
15									

Sample rate comparison for OCP 100mS windowed averaging test

For each sample rate, we have the worst-case (peak) power measured across the 100mS averaging window and can immediately see a problem.

At 4uS and 8uS sampling, there is minimal difference (around 0.05% error), but this rapidly increases and the sample rate drops.

By 128uS sampling (~8 KS/s) we are well over 1% error, and this gets steadily worse. While the magnitude and trend of the error will depend on the data set, this clearly shows that we need at least 10 KS/s if we want to avoid sample rate contributing significantly to the final error. Remember that errors can stack, calibration accuracy, and similar factors have to be accounted for in the total potential error estimate.

OCP 100uS peak power test

The OCP test also require a 100uS averaging window. As this is a smaller window, it should follow that we have to sample faster to avoid errors.

	A	B	C	M	N	O	P	Q
1								
2					OCP 100uS windowed peak			
3	Sample Rate	Frequency Hz			Peak uW	Peak mW	Error mW	Error %
4	4us	250,000.00			22163363.04	22,163.36	0	0
5	8us	125,000.00			22177736	22,177.74	14.37	0.065%
6	16us	62,500.00			23816743.33	23,816.74	1,653.38	7.460%
7	32us	31,250.00			24778574.67	24,778.57	2,615.21	11.800%
8	64us	15,625.00			26184825	26,184.83	4,021.46	18.145%
9	128us	7,812.50			26044407	26,044.41	3,881.04	17.511%
10	256us	3,906.25			26032206	26,032.21	3,868.84	17.456%
11	512us	1,953.13			25968436	25,968.44	3,805.07	17.168%
12	1024us	976.56			25549635	25,549.64	3,386.27	15.279%
13	2048us	488.28			25549635	25,549.64	3,386.27	15.279%
14	4096us	244.14			25529028	25,529.03	3,365.66	15.186%
15								

Sample rate comparison for OCP 100uS windowed averaging test

And that turns out to be correct! Again, the 4uS and 8uS sampling rates are very similar, but the errors increase faster with the smaller window. Here, we would not recommend running slower than 125 KS/s to get an accurate result.

The OCP specification requires a base sample rate of 4uS for these measurements, and we can see here that this is a sensible requirement.

Workload-average power tests

In other standards, total energy efficiency is the focus, rather than peak power. Here we are more likely to look at the average power over a full workload.

A new standard we are contributing to specifies 4uS sampling, averaged down to 1 sample per second. If a workload takes 100 seconds, then the 100 readings would be further averaged to get the final efficiency result. This allows users to get a basic trend-over-time graph of power usage and also the single average power value for the workload.

	A	B	C	D	E	F	G
1							
2				Full Trace Mean			
3	Sample Rate	Frequency Hz		Mean Power uW	Mean Power mW	Error mW	Error %
4	4us	250,000.00		6767417.029	6,767.42	0.000	0.000%
5	8us	125,000.00		6766914.986	6,766.91	-0.502	-0.007%
6	16us	62,500.00		6769165.104	6,769.17	1.748	0.026%
7	32us	31,250.00		6769615.209	6,769.62	2.198	0.032%
8	64us	15,625.00		6764108.93	6,764.11	-3.308	-0.049%
9	128us	7,812.50		6773580.331	6,773.58	6.163	0.091%
10	256us	3,906.25		6770460.385	6,770.46	3.043	0.045%
11	512us	1,953.13		6762499.66	6,762.50	-4.917	-0.073%
12	1024us	976.56		6779879.666	6,779.88	12.463	0.184%
13	2048us	488.28		6793972.982	6,793.97	26.556	0.392%
14	4096us	244.14		6696411.077	6,696.41	-71.006	-1.049%
15							

Sample rate comparison for average power

Unsurprisingly, we see that taking an average of the whole trace makes for a more stable result as we reduce the sample rate. We effectively have a single ‘window’ that is the size of the whole trace here, and so the error increases more slowly.

In this example, we could have gone as low as ~2KS/s before getting a significant loss of accuracy. This means that the sample rate is important, but not nearly as much as with the peak power tests.

Note that 1 second averaging is NOT the same as 1 second sampling. In the workload averaging test, we will be capturing data at 4uS (250KS/s) and averaging all 250,000 values into a single figure for the second. While this loses detail (as we only have one measurement per second), it perfectly maintains the average power consumption.

Quarch tools always average rather than skip samples, so ensure the most accurate power reading possible.

Do we need to go faster?

All the Quarch tools run at 250 KS/s for analog sampling. We saw here that this looked good enough for the above test cases, but that does not prove that we would not see better results at higher rates.

To do this, we grabbed the scope from our lab, which has current probes. This is an older one, but it was a decent unit.

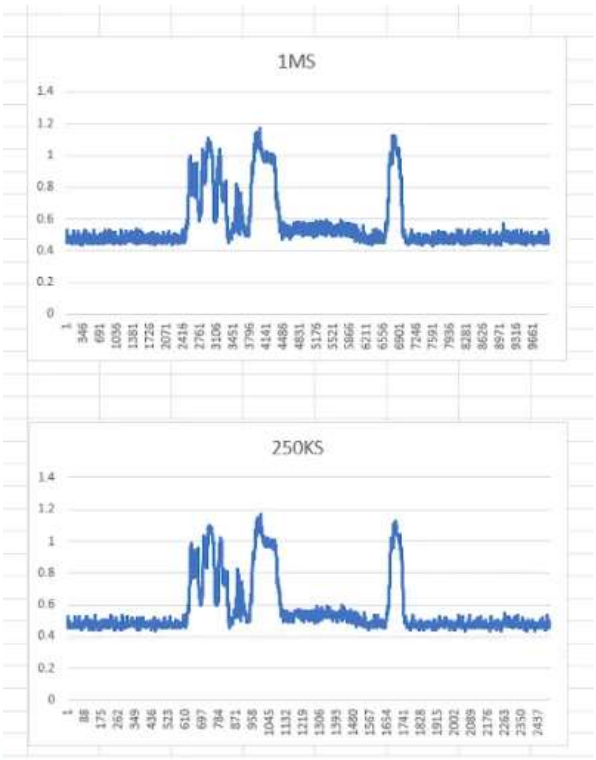
Tektronix DPO 3032 with TCP0030 current probes.

It has a vastly faster sample rate at 2.5 GS/s, though its bandwidth is limited to 300 MHz which makes the higher end less useful.

We hooked up a Gen5 SSD this time to ensure we looked at both drive technologies. We ran the scope at 1 MS/s to. We hooked up the PAM in series so we could capture both tools at the same time.

Gen5 U.2 SSD with Quarch PAM and current probe

Let’s zoom in on a small section of the SSD trace and compare at 1 MS/s to 250 KS/s.



Scope data export from SSD trace

We can see that both traces follow the exact same trend; the only difference is that the fast trace has slightly more noise.

	A	B	C	D	E
1					
2		Sample rate	mean	mean error %	
3		1 MS/s	0.564	0.0000%	
4		500 KS/s	0.563976	-0.0043%	
5		250 KS/s	0.564003	0.0005%	
6					
7					

Scope data processed statistics

The calculated statistics confirm this; the difference between the 3 sample rates shows extremely small differences to the mean.

This demonstrates that sampling faster than 250 KS/s has no benefit for most tests. By testing both HDD and SDD, we have also ensured our conclusions are accurate for both.

Conclusions: 4 costs of speed

When we tried to run the scope comparison in the lab, we quickly hit problems:

This scope has a memory depth of 5 megapoints, so while it can run happily at 50 MS/s, it can only capture at this rate for a fraction of a second. Even at 1MS/s it could not capture the full power-up cycle.

Now our scope is pretty old, but looking at more modern scopes, memory depth is often rather limited.

- InfiniiVision 4000 X Series Digital Bench Oscilloscope, MSOX4154A
    - $31k USD – 4 Mega points (matching our scope)
- Keysight DSO91304A Infiniium
    - $41k USD – 1000 Mega points

1. The Keysight 9000 series has a lot of memory, but it has a price to match! Trying to sample at a faster rate than you need can be very costly and yet give no benefit.
2. A fast ADC sampling rate is of no use unless you are recording and actually using the data from every single sample. If any part of your analysis chain is discarding samples, your effective sample rate is lower.
3. At high sample rates, the amount of data produced is very significant. A drive generally has 2 power rails, and we need to sample both voltage and current. That is 4 measurement channels and can turn into gigabytes of data quickly at high sample rates.
4. At high sample rates, you begin to sample increasing levels of ‘noise’ which is not actually part of the true power consumption of the device and can actually make your result less useful. We’ll have more coming on this in a later blog!

4.4 各类线缆热插拔/故障注入模块

4.4.1 24G MINISAS HD 线缆热插拔模块

24G MINISAS HD CABLE BREAKER

Supports SFF-8674 cables for external HD SAS interconnects up to 24G speeds.

44cm removable cable to connect to Torridon Controller.

Up to 4 units can be rack mounted in 1U.

SAS BREAKER: INDIVIDUAL RF SWITCHES ON EACH OF THE 4 SAS LANES. SIDEBAND SWITCHING FOR VMAN, VACT_0, VACT_1, MODPRSL, SDA, SCL, INTL. ACCESS TO EEPROM REGISTERS ON ACTIVE CABLES.

4.4.2 PCIe Gen4 MINISAS HD 线缆热插拔模块

Supports SFF-8644 cables for external PCIe interconnects up to Gen5 speeds.

44cm removable cable to connect to Torridon Controller.

Up to 4 units can be rack mounted in 1U.

PCIE BREAKER: INDIVIDUAL RF SWITCHES ON EACH OF THE 4 PCIE LANES. SIDEBAND SWITCHING FOR .PWR_B1, PWR_D1, CMI_SCL, CMI_SDA, CBL, PRSNT, MGT, PWR, CADDR, CINT. FULLY WIRED CUSTOM CABLE PROVIDED TO ENSURE THE BREAKER IS FULLY TRANS-PARENT /WHEN IN USE. DRIVE+MONITOR: CADDR, CINT. MONITOR: CMI_SCL, CMI_SDA, CBL_PRSNT, MGT_PWR.

4.4.3 PCIe Gen4 OCULINK 线缆热插拔模块

Sniffer

GEN4 OCULINK CABLE BREAKER

Supports SFF-8611 cables for external interconnects up to Gen4 speeds.

44cm removable cable to connect to Torridon Controller.

Up to 4 units can be rack mounted in 1U.

OCULINK BREAKER: INDIVIDUAL RF SWITCHES ON EACH OF THE 4 LANES. SIDEBAND SWITCHING FOR: VACT_1, VACT_2, VSP_PL, VSP_MN,CWAKE, SMDAT, SMCLK, PERST, CPRSNT, RSVD_A9.

4.4.4 SFP28 25GE/32G FC 线缆热插拔模块

Supports SFP cables up to SFP28 speeds. Designed for compatibility with SFP56.

44cm cable to connect to Torridon Controller.

'Inline' form factor, plugging into the host
and extending the receptacle.

SFP28 BREAKER: INDIVIDUAL RF SWITCHES FOR THE DATA LANE. SUPPORTS ALL PROTOCOLS THAT FOLLOW THE SFP28 ELECTRICAL SPECIFICATION. SIDEBAND SWITCHING FOR: VCC_TX, VCC_RX, MOD_ABS, SDA, SCL, TX_FAULT, TX_DISABLE, RX_LOS, RS0, RS1

4.4.5 QSFP28 100GE/128G FC 线缆热插拔模块

Supports QSFP cables up to QSFP28 speeds. Designed for compatibility with QSFP56.

44cm cable to connect to Torridon Controller.

'Inline' form factor, plugging into the host and extending the receptacle.

QSFP28 BREAKER: INDIVIDUAL RF SWITCHES FOR ALL 4 DATA LANES. SUPPORTS ALL PROTOCOLS THAT FOLLOW THE QSFP28 ELECTRICAL SPECIFICATION. SIDEBAND SWITCHING FOR: VCC_TX, VCC_RX, VCC_1, MOD_PRSL, SDA, SCL, INTL, RESETL, MODSELL, LPMODE

4.4.6 RJ-45 1000M 以太网线缆热插拔模块

Supports 10/100/1000 Base-T devices.

44cm cable to connect to Torridon Controller.

Triggering option available

RJ-45 BREAKER: INDIVIDUAL RF SWITCHES FOR THE DATA LANES. SUPPORTS 10/100/1000 BASE-T. DATA SWITCHING ONLY, POE WILL NOT PASS THROUGH. TRIGGERING VERSION AVAILABLE FOR SYNC WITH EXTERNAL DEVICES.

4.4.7 USB 3.0 线缆热插拔模块 A/B 口

Supports USB devices up to USB 3.0 speeds.

44cm cable to connect to Torridon Controller.

Individual switching for USB-3, USB-2 and VBUS.

USB 3.0 BREAKER: INDIVIDUAL RF SWITCHES FOR ALL USB LANES. FET SWITCHING FOR POWER

4.4.8 USB 3.1 线缆热插拔模块 Type-C

Supports Low Speed, Full Speed, High Speed, Super Speed and Super Speed Plus devices.

44cm cable to connect to Torridon Controller.

Individual switching for USB-3, USB-2 and VBUS.

USB TYPE-C BREAKER: INDIVIDUAL RF SWITCHES FOR ALL USB LANES. FET SWITCHING FOR POWER

4.4.9 -48V DC 电信供电热插拔模块

Supports power switching of -48V power to telecoms devices.

44cm cable to connect to Torridon Controller. Triggering option available.

Fitted with ELCON Mini 2-pin connectors.

-48V BREAKER: FET SWITCHING FOR 48V_POWER, SIDEBANDS: DETECT_1, DETECT 2. SUPPORTS UP TO 20A AT -48V. VOLTAGE TOLERANCE FROM -25 TO -72V.

4.4.10 多协议汽车电子总线热插拔模块

Supports USB-2, RS-232, RS-422, RS-485, CAN/LIN, I2C and 1000Base-T1.

Switches 4 data lanes and one power rail.

Pluggable screw terminals for quick connection

External triggering via MCX

Direct USB and LAN ports (no Torridon controller required).

Power from 12v or PoE.

Capture bus transactions with up to 1MHz sampling

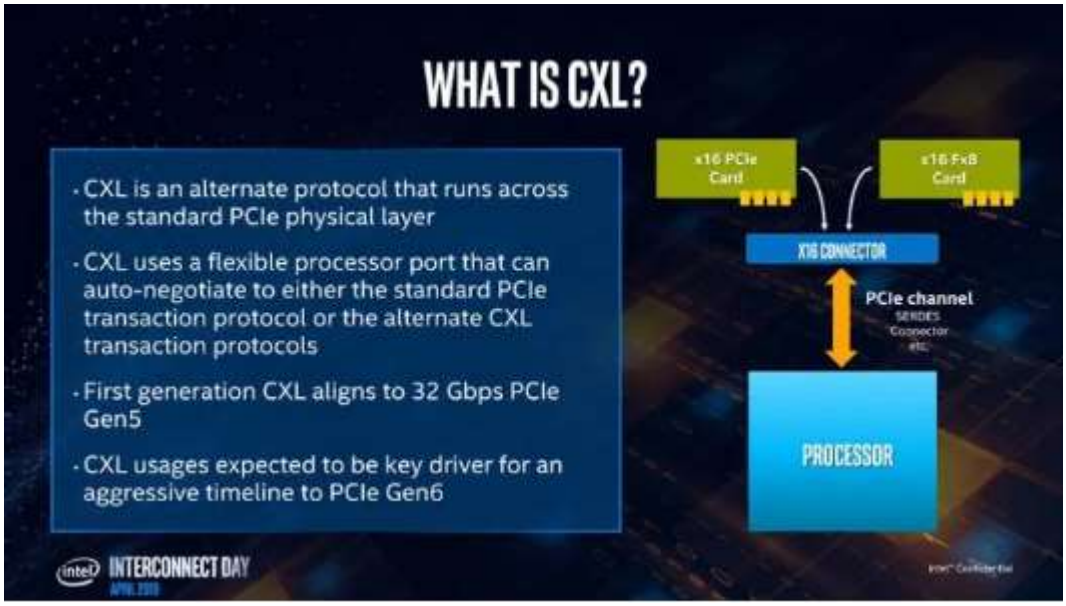
MULTI-PROTOCOL BREAKER: ANALOG SWITCHING SUPPORTS MANY PROTOCOLS THAT HAVE A SIGNAL RANGE WITHIN -15V TO +15V ANDBANDWIDTH OF LESS THAN 460 MHZ. FOR CAPTURE, BUS TRANSACTIONS ARE DIGITALLY SAMPLED AT 1MHZ MAX WITH CONFIG-URABLE

4.5 你需要什么工具测试 CXL?

发布日期：2023 年 5 月 31 日

Andy Norrie

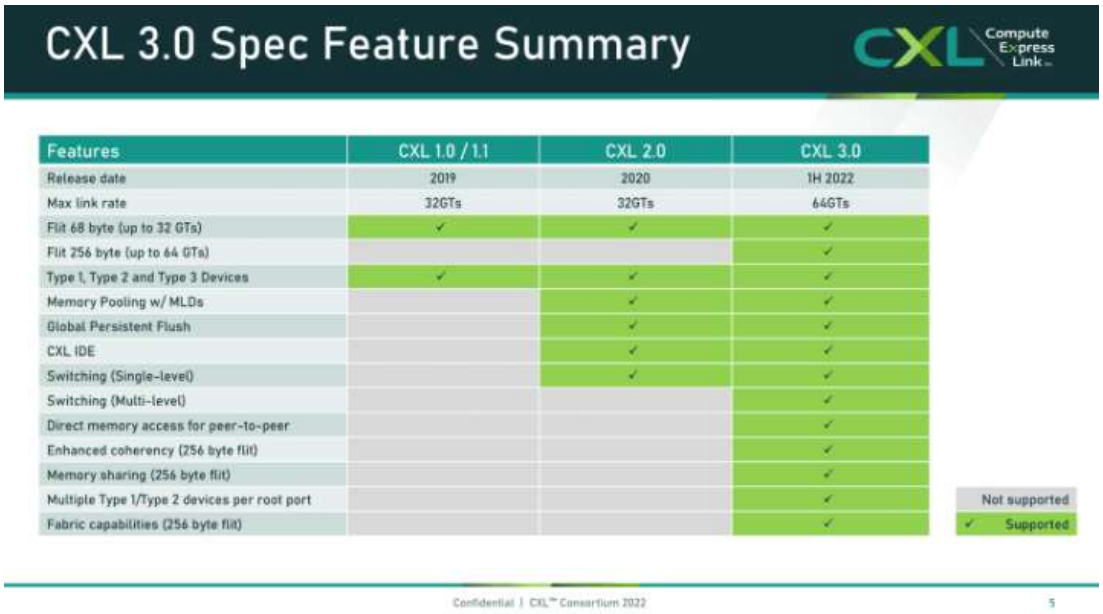
什么是 CXL?



Compute Express Link (CXL) 是一种高速 CPU 互连，旨在加速下一代数据中心性能。它是一种开放标准技术，可在 CPU 和工作负载加速器（例如 GPU、FPGA 和网络设备 (NIC)）之间实现高带宽、低延迟数据通信。CXL 的一项关键创新在于其能够保持 CPU 和内存附加设备之间的内存一致性，从而允许资源共享以实现更高的性能、降低软件堆栈复杂性并降低总体系统成本。

与 PCIe 热插拔和故障注入影响 PCIe 的方式相同，CXL 在 PCIe 总线上运行（如 NVMe）。我们 Quarch 的工程师将指导您应对以下挑战：测试 CXL 需要什么？

测试 CXL 需要什么？



好消息是，专为 PCIe 设计的 Quarch 产品与协议无关，可用于测试任何 PCIe 兼容设备。这意味着我们现有的一系列获奖产品可用于 CXL。如果您的实验室中有现有的 Quarch 产品，那么您今天就可以重新利用它们！

CXL 的复杂性对任何组织来说都存在一些重大的测试挑战，因此如果您需要帮助，请给我们发送电子邮件并直接与我们的工程师联系以寻求帮助：sales@saniffer.com

热插拔和故障注入

Quarch“断路器”模块在整个行业中广泛使用，可自动执行各种测试，从简单的驱动器拆卸到复杂的热插拔时序、通道限制等。这对于证明您的设备可以处理现实世界中遇到的各种配置和时序至关重要。

Gen5 AIC 断路器

EDSFF E3 断路器

功率分析

功耗是任何现代设备的一个重要指标。在设备的整个生命周期中，能源（和冷却）成本很高。由于数据中心内运行着数千台设备，即使是很小的效率提升也会产生巨大的影响。

制造商数据表可以提供有用的指导，但验证系统中的实际功耗至关重要。

Quarch 工具可以通过几种主要方式帮助进行功率测试

我们的 PPM（可编程电源模块）系列可以为设备供电并改变电压以进行各种测试：电压拉

偏、掉电、毛刺等。我们的 PPM 和 PAM（功率分析模块）系列都可以为您提供高分辨率和长期记录的校准功率测量。当查看设备如何响应不同的工作负载时，这非常有用。

两个设备的量程均可测量至 100uA，以验证设备的睡眠状态并准确了解其空闲时使用的电量。通常，如果空闲功率较低，则设备是“有能力”的，但在实践中可能无法进入该状态。

测试实例

如果您没有 Quarch 产品的经验，您会发现我们提供的 application notes 实现了各种有用的测试：

- UNH-IOL Plugfest 测试 – 专为 NVMe SSD 热插拔而设计，但与 CXL 非常相关。验证设备每次都正确枚举，具有一系列快速和慢速的插入速度。

- 功耗与性能 – 在驱动器上运行许多工作负载并分析设备的功耗与性能。

如需详细阅读，请查看我之前关于 SSD 故障注入测试建议的博客 - SSD 自动化测试计划 – Quarch Technology

预构建测试

并非每个团队都有时间或能力编写自动化测试。Quarch 合规性套件包含一组自动化测试，涵盖从热插拔到电源性能等各个方面。可免费尝试基本测试并包含详细报告：

Quarch 合规套件 – Quarch 技术

4.6 PCIe Gen 4/5/6 NVMe SSD 掉电测试工具

4.6.1 SerialCables 标准可管理掉电卡

4.6.1.1 PCIe Gen 4 M.2/AIC SSD 掉电卡

该 M.2/AIC 掉电卡非常适合将 M.2 插入台式机主板进行测试，通过自带的 microprocessor 实现上、下电功能。

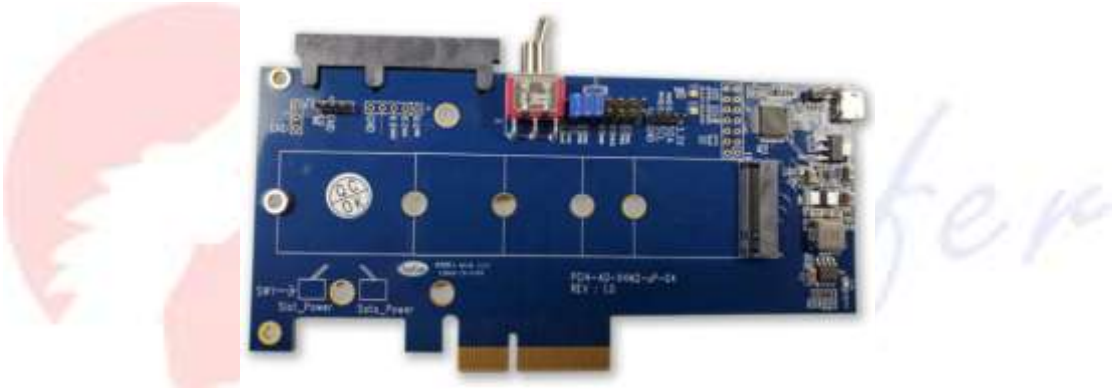


图 4-137

PCIe Gen4 x4 slot – M.2 with secondary power (15 pin SATA) and microUSB based microprocessor CLI

4.6.1.2 PCIe Gen 4 M.2/U2 SSD 掉电卡

该 M2/U.2 带掉电功能的转接卡，可以非常方便用户将 M.2 转接到 U.2，然后插入服务器或者其它盘柜的前面板的 24-bay 个 U.2 插槽中实现批量测试，通过其自带的 microprocessor 实现掉电、上电等管理功能。

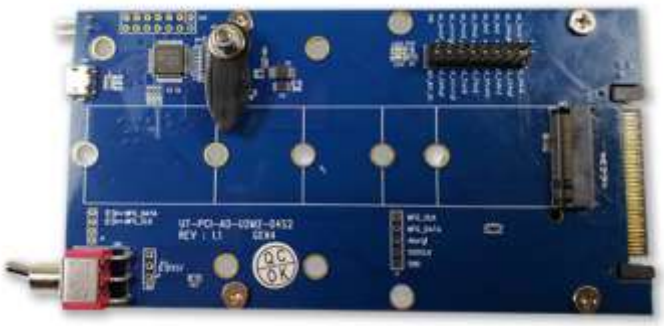


图 4-138

PCIe Gen4 ready U2M2 adapter w/microprocessor. Made with Thunderclad 3+ low-loss dielectric and the newest PCIe Gen4 connectors. Independently control power on/off to DUT. Measure current/voltage and temperature as well as access SMBus through CLI. (see user guide for more specifics). For all M.2 lengths up through 110mm w/quick latching mechanism for easy changing of SSD's on the adapter.

4.6.1.3 PCIe Gen 4 U2/AIC SSD 掉电卡

下面的 U.2/AIC 转接卡本身不带 microprocessor，但是可以通过接入 Quarch 公司的 PPM 实现掉电、上电功能。

图 4-139

4.6.2 PCIe Gen 4 M.2 SSD 定制可管理掉电卡

4.6.2.1 单盘位 M.2 SSD 掉电卡

如果希望不仅仅针对单个 PCIe Gen 4/5/6 M.2 NVMe SSD 实现掉电测试，还可以实现对于某些信号，例如 CLKREQ 或者 PERST 等实现拉高、拉低等操作，那么采用下面的

HHHL 的设计的掉电卡可以很方便地进行测试，支持通过 USB serial 进行脚本控制，支持 inband 固件升级。

同时，该卡也可以结合用户的需求进行二次定制开发。

图 4-140

4.6.2.2 四盘位 M.2 SSD 掉电卡

如果需要提供 4 个 M.2 SSD，并且希望放入温箱进行测试，推荐使用下面的 Gen 4 M.2 掉电板实现。同时，该卡也可以结合用户的需求进行二次定制开发。

图 4-141

4.6.2.3 四盘位 M.2 SSD 掉电卡

根据 OCP 规范标准设计的 FHFL 全高的 PCIe Gen 4/5/6 x16 插卡，提供 4 个 M.2 NVMe SSD 接入，通过设置主板 BIOS 将 x16 分叉成 4x4 即可实现对于 4 个 M.2 NVMe

SSD 的测试。支持通过 USB 串口实现脚本控制。另外， 由于 4 个 M.2 socket 设置伸出主板挡板之外，方便插拔。参见下图。

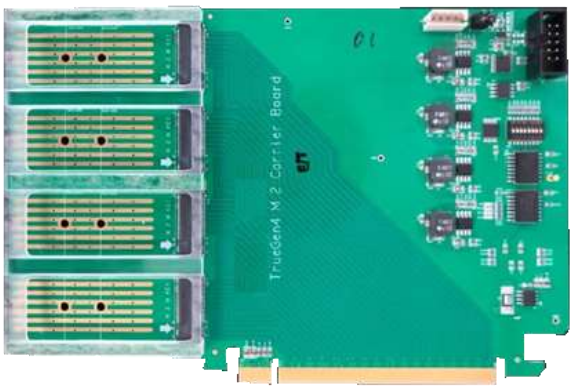


图 4-142

4.6.2.4 八盘位 U.2 SSD 掉电背板

参见下图，该 8 盘位 Gen 4 U.2 NVMe SSD 背板主要用户常温批量测试环境，也可以去掉一些管理组件用于高、低温温箱环境。 该背板可以结合用户的需求进行二次定制开发。

图 4-143

一般情况下，需要温箱中通过 SerialCables Internal Host Card 实现和该 4-BAY enclosure 的连接，参见下图。

图 4-144

4.6.3 PCIe Gen 4 盘柜背板掉电

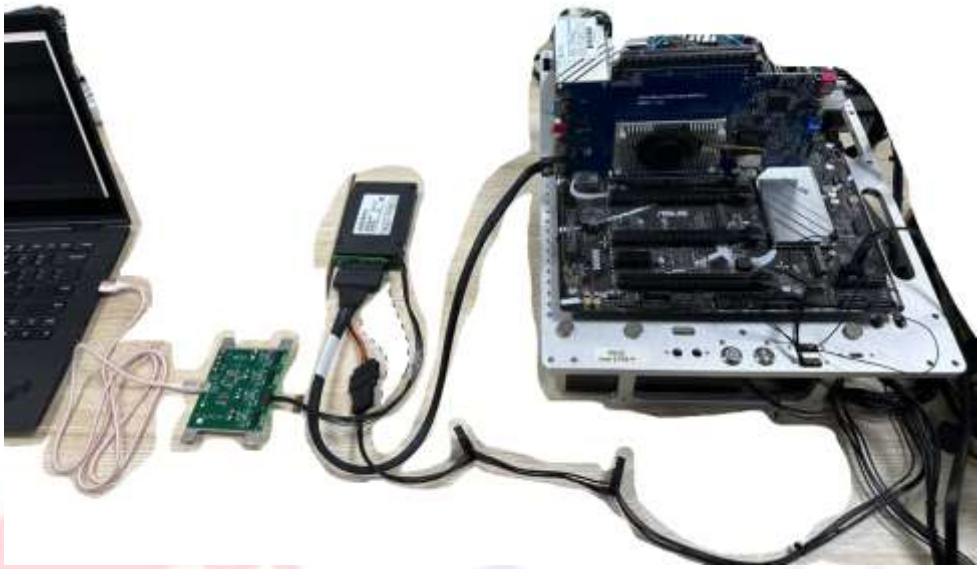
参见 1.2.1.4 PCIe Gen 4/5/6 盘柜部分，通过盘柜内置的 API 很容易实现在测试过程中通过 python 脚本调用对任意盘位进行掉电/上电操作。

图 4-145 通过命令控制每个 Gen 4 U.2 SSD 盘位掉电/上电

4.6.4 PCIe Gen5 U.2 SSD 经济型掉电/上电/功耗计量/边带信号拉高/拉低测试工具

如果在产线上进行掉电/上电/功耗计量/边带信号拉高/拉低的自动化测试，那么就需要一种经济型的测试治具，下面的测试治具可以满足绝大部分 PCIe Gen5 U.2 SSD 的产线自动化测试需求。

4.6.4.1 硬件连接示例



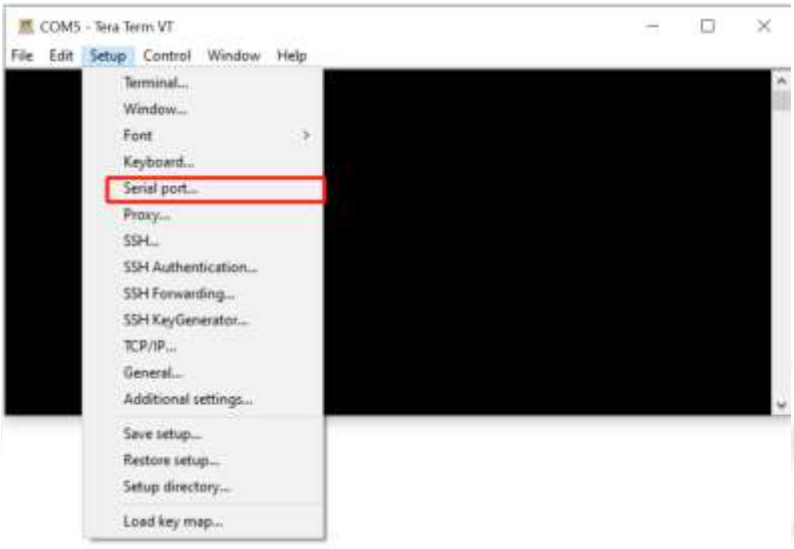
图中的绿色板卡为控制模块，可以控制 4 块串接在 Gen5 U.2 SSD 和背板之间的治具。

4.6.4.2 串口连接设置

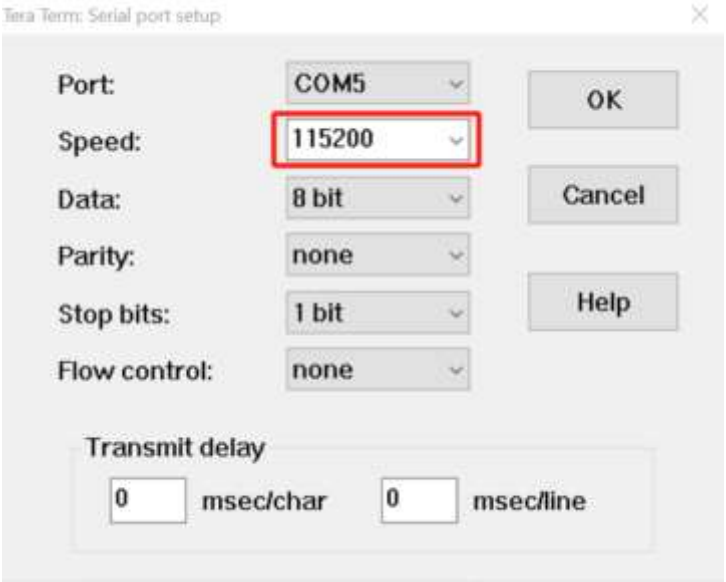
1. 打开 Teraterm 软件

2. 选中正确的端口，点击‘确认‘连接

3. 进入终端界面后，选择用户 **bar** 里的 **setup**，下拉选项中选择串口设置



4. 在串口设置界面将串口速率改为 **115200**（一般默认串口速率为 **9600**），设置后选择‘确认’完成设置



5. 回到终端窗口界面，回车显示 CMD>字样，表明设备已经正确连接了

4.6.4.3 CLI 命令行操作

1. help：命令查询

（1）直接输入'help'回车，显示所有支持的命令类

（2）输入'help [命令类]'回车，显示某个命令类的描述、参数

2. iap：通过 Y-modem 的方式更新应用

注意：客户请勿使用该命令，会无法退出导致系统崩溃，要重新刷 firmware 才能修复

3. monitor：监控系统状态，实时监控系统的电压、电流、功耗和数字信号（CLKREQ，ALERT，WAKE，3V3P，PRE）信号的变化

4. set：设置系统参数并且立即执行

```
CMD>help set
set - set system variables.
Usage:
set name [value]
   - set variable 'name' value to 'value'
        - set vc on      - set 3v3 vc power on.
        - set vc off     - set 3v3 vc power off.
        - set vp on      - set 3v3 vp power on.
        - set vp off     - set 3v3 vp power off.
        - set rom on     - set SSD ROM mode sigle to low.
        - set rom off    - set SSD ROM mode sigle to high.
        - set rst low    - set SSD reset signal to low.
        - set rst high   - set SSD reset signal to high.
        - set pre low    - set SSD present signal to low.
        - set pre high   - set SSD present signal to high.
        - set pln low    - set SSD pln signal to low.
        - set pln high   - set SSD pln signal to high.
        - set fd low     - set fd signal to low.
        - set fd high    - set fd signal to high.
        - set clk low    - set clkeq signal to low.
        - set clk high   - set clkeq signal to high.
        - set pwr low    - set SSD Powerdis signal to low.
        - set pwr high   - set SSD Powerdis signal to high.
        - set wake low   - set SSD wake signal to low.
        - set wake high - set SSD wake signal to high.
```

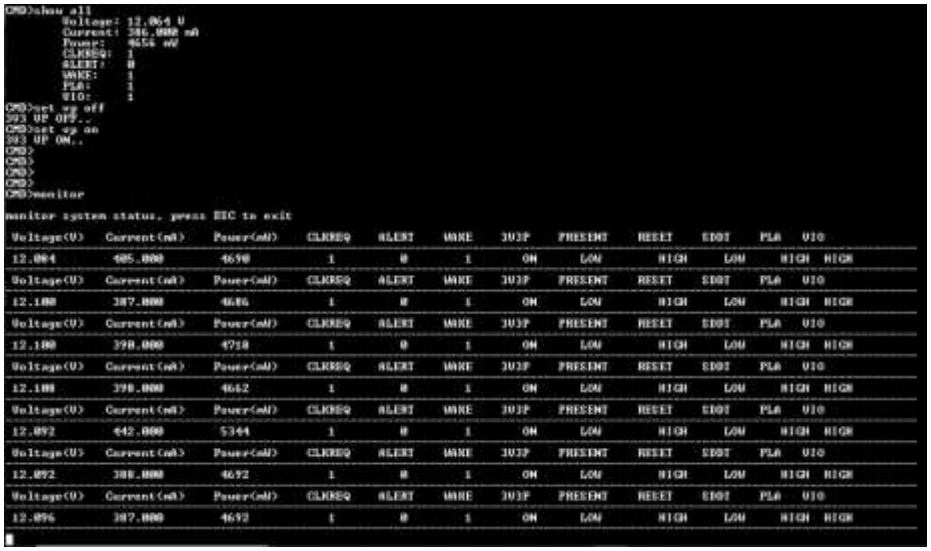
5.show：显示系统变量

```
CMD>help show
show - show system variables.
Usage:
show name
   - show system variables. etc.
   - show voltage    - show  voltage value.
   - show current    - show  current value.
   - show power      - show  power value.
   - show clkreq     - show  CLKREQ status.
   - show alert      - show  ALERT status.
   - show wake       - show  WAKE status.
   - show pla        - show  PLA status.
   - show vio        - show  VIO status.
   - show all        - show  voltage\current\power value.
```

6. version：查看系统版本信息

```
CMD>help version
version - show system vesion infomation
Usage:
version
```

7. 查看实时的电压/电流/功耗



4.7 针对主机进行异常掉电的自动化工具

一般情况下，对于企业级 NVMe SSD 等测试过程中，会在 I/O 读写过程中做异常掉电，来检验 atomic write 是否正常。参见下图 SanBlaze 测试工具手工测试 atomic write 的设置。

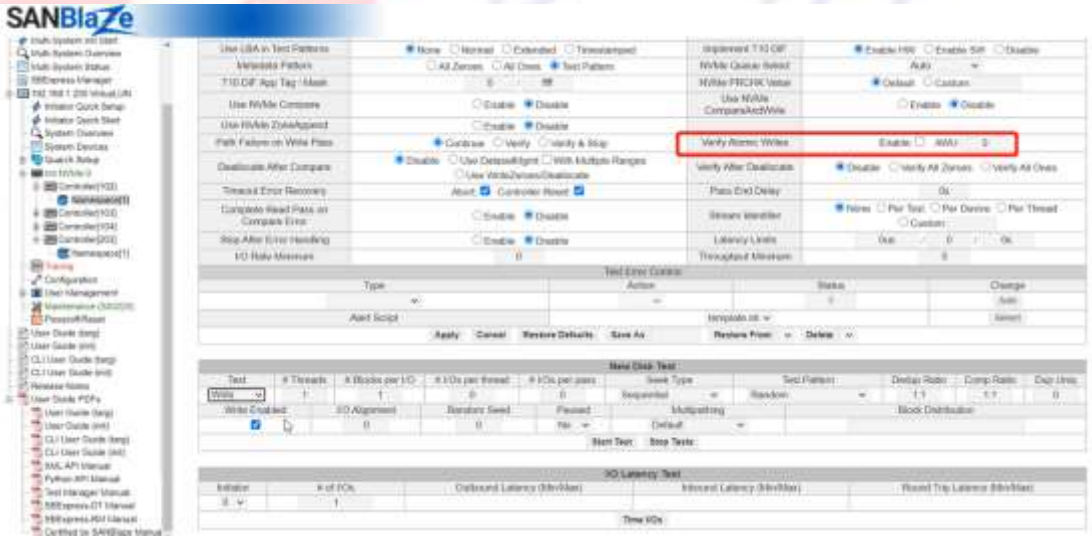


图 4-146

当然，如果采用脚本自动化测试，一般需要配合 Quarch 的热插拔模块然后采用 SanBlaze 脚本如下图所示。

异常掉电的时候，SSD 内部有电容（参见下图的 Kioxia 企业级 NVMe SSD 的 6 颗大电容）可以工作 1~2 秒钟，所以，我们一般会异常掉电，然后重新上电后再将最后掉电时刻前面的所有的写操作对应的扇区内容重新读取回来进行 verify 校验，从而验证“掉电保护”功能是否正常。

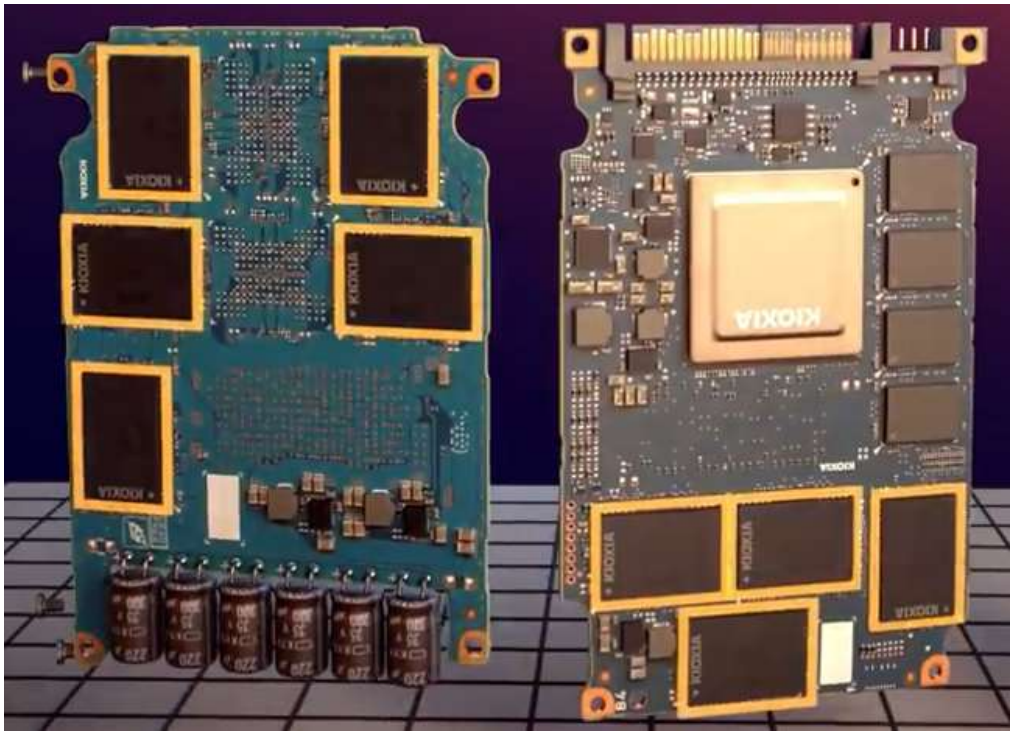


图 4-147

有的时候，可能有的客户也有想法是，主机在 I/O 过程中突然掉电，其实这种情况非常少见，尤其是数据中心，因为有 UPS 等的保护，但是如果还是想做这方面的测试，我们推荐采用可编程智能 PDU，支持 web GUI 管理，或者脚本控制自动化测试，参见下图。

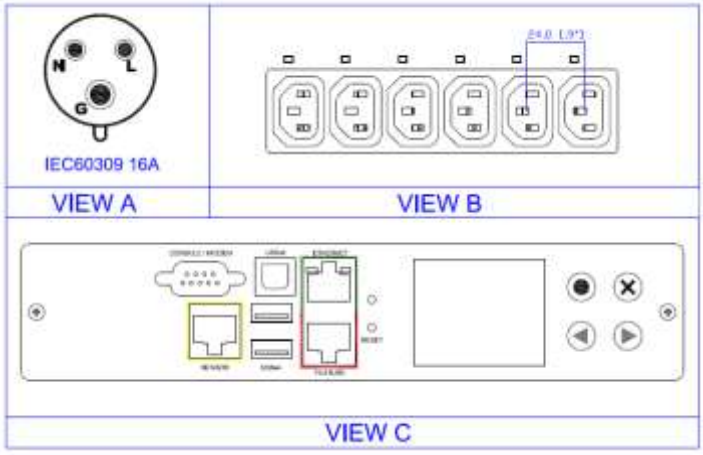


图 4-148

注意，这种产品都部署在机房机柜里面，所以电源输入采用 IEC60308。

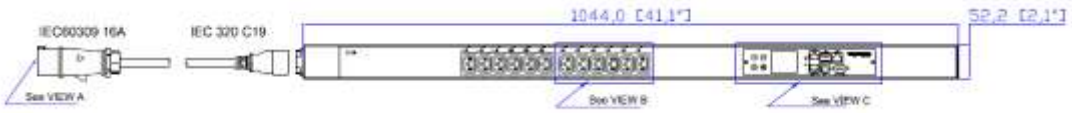


图 4-149

输出到主机的电缆采用 C13 转 C14 电源线，参见下图。

图 4-150

下面是 Web GUI 的 power on/off 功能。

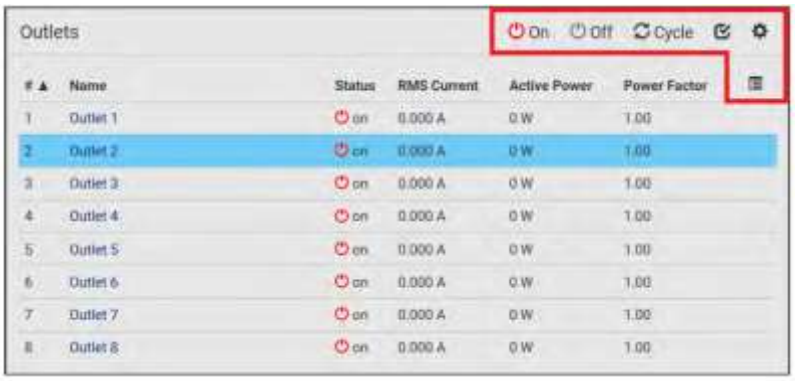


图 4-151

下面是全局监控功能。

图 4-152

说明：这类进口产品价格也比较经济，大概人民币几千元。

4.8 USB over Network 远程管理小工具

本章节涉及的 Quarch 等很多产品都需要 USB 或者串口管理，有的时候，工程师希望在自己工位上管理机房里面的 Quarch 产品，那么下面的工具可以助你一臂之力。

FlexiHub’s key task is to set up a reliable USB over network link for you to access any remote USB devices* or share those from your local PC through all the network - https://www.flexihub.com/

下图是支持各种 USB 设备列表，Quarch 属于计算机类。

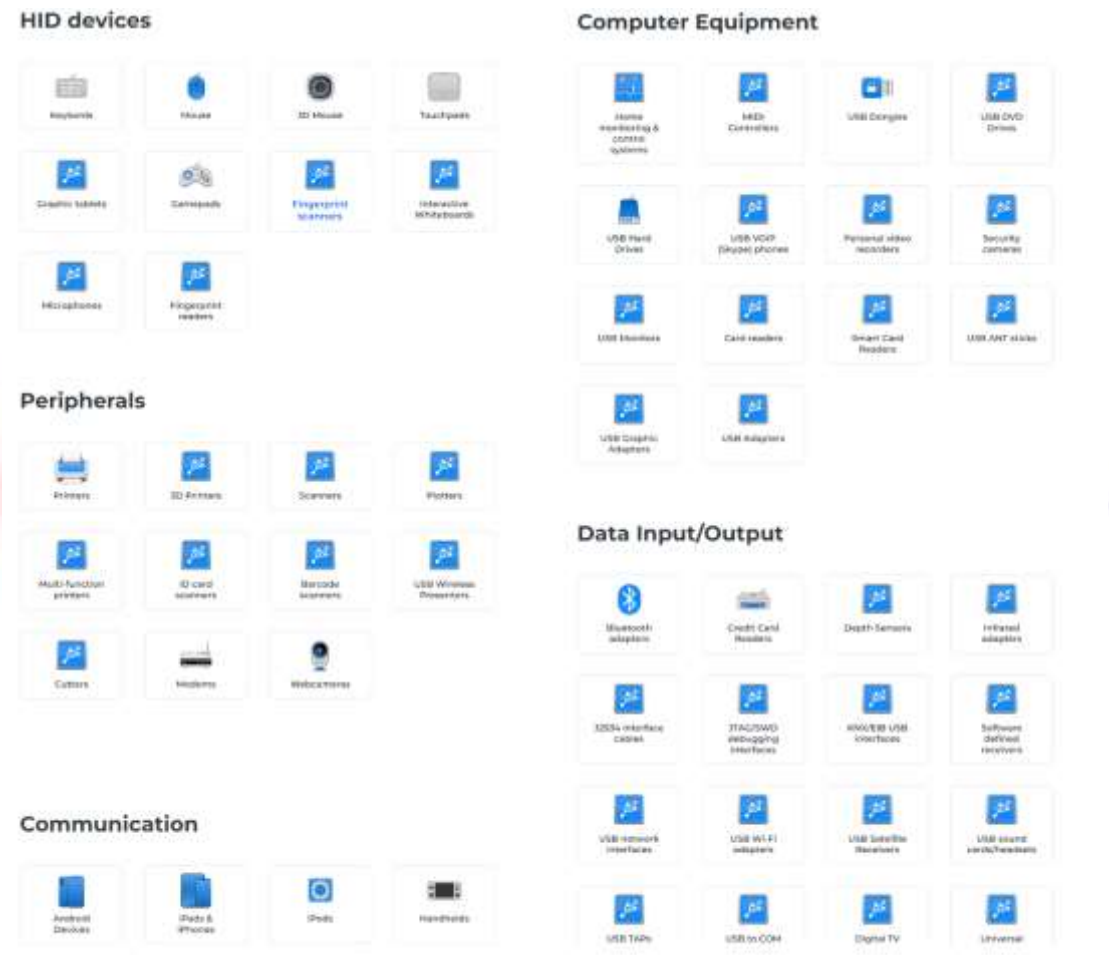


图 4-153

5. PCIe Gen4/5/6 NVMe SSD 测试环境搭建一: 主机卡/盘柜/转接/延长线

5.1 构建 PCIe Gen5 企业级 NVMe SSD 和各类插卡测试环境必备的各类产品

在研发实验室构建针对 PCIe Gen5 x4 企业级 NVMe SSD 的测试环境，或者针对 Gen5 x16 接口的企业级 GPU, DPU/SmartNIC，AI 板卡，或加速卡的测试环境，需要综合考虑这些产品在企业级数据中心或者 Internet 网数据中心的实际使用环境。

企业级 SSD 和各类板卡除了可能直连 PCIe Gen5 CPU 外，大多数都需要连接 PCIe Gen5 switch 或者 Retimer 卡，所以构建测试环境的时候就必须要考虑这些治具。

下面我们简要介绍一下构建 Gen5 测试环境涉及的相关的产品和技术。

5.1.1 PCIe Gen5 Switch 卡

在目前市场上还几乎买不到任何量产发布的 PCIe Gen5 服务器和 endpoint 卡的情况下，SerialCables 公司的 PCIe Gen5 switch 卡成为用户搭建 Gen5 测试环境的唯一选择。

该卡对于构建 PCIe Gen5 测试环境具备两重属性：

- **对于测试 RC 端，例如 CPU 来讲，它可以作为可靠的“EndPoint”，用来训练 RC 端的 PCIe Gen5 链路。**
- **对于测试 EP 端，例如上述的 SSD 和各类板卡，它作为可靠的“RC”，用来训练 EP 端的 PCIe Gen5 链路。**

下图为 SerialCables 最新版本的 Rev1.6 的 Gen5 switch 卡图片，目前国内主流芯片公司基本都购买该卡进行 Gen5 CPU 或者板卡/SSD 盘的测试。

图 5-1

该 Gen5 switch 卡提供上行 Gen5 x16 金手指，下行分成两部分：插槽和 MCIO 接口。

- **顶部提供 PCIe Gen 5x16 插槽，提供非常好的信号质量，可以用来测试各种 Gen5 板卡，如 GPU/DPU。**

下图是采用该 PCIe Gen5 x16 switch 卡测试图形卡的示意图。

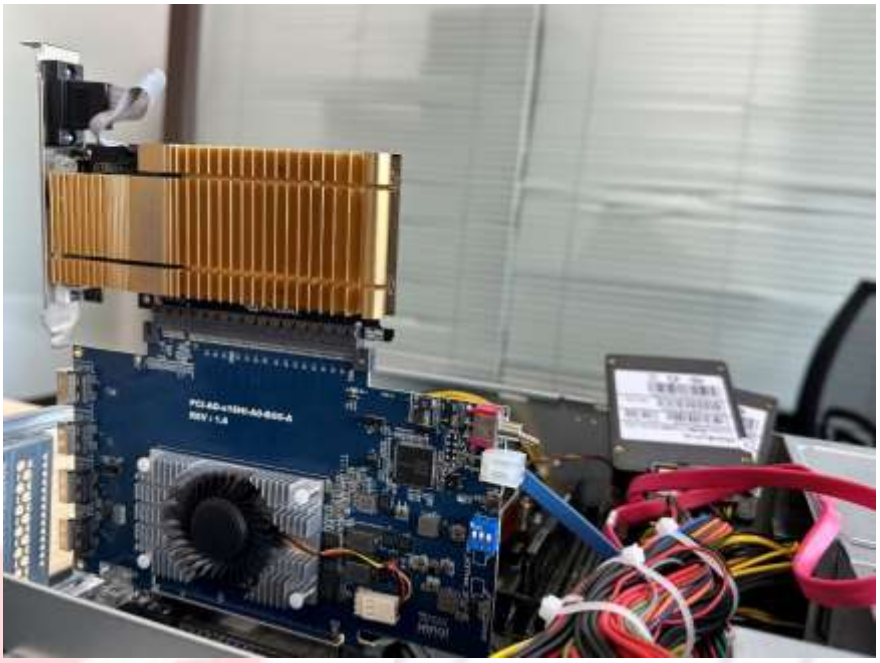


图 5-2

– 该 Gen5 Switch 卡的顶部插槽信号输出的眼图质量如下图所示。

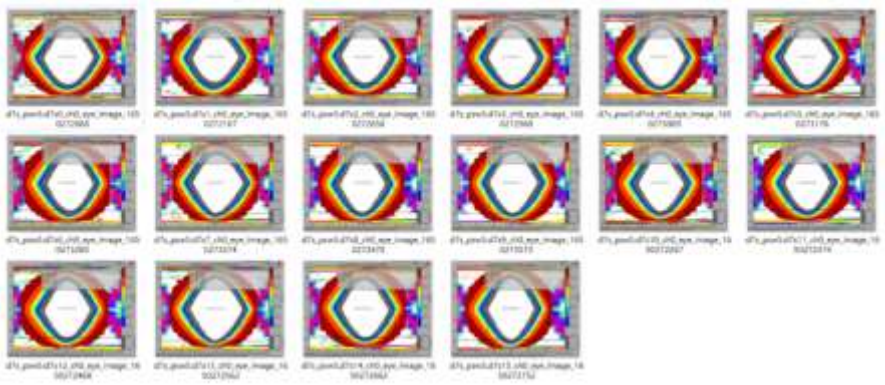


图 5-3

– 相比较来讲，目前不少客户在 x86 Gen5 CPU 服务器量产之前采用将插卡插入工作站/台式机的 PCIe Gen5 x16 插槽的方式进行测试，可能碰到很多问题，除了 CPU 本身的问题外，主板信号质量也是一个很重要的因素，下面的眼图是在号称业内最好的工作站主板厂商提供的 Intel Z690 主板芯片组的 PCIe Gen5 x16 插槽（直连 CPU）获得的眼图。

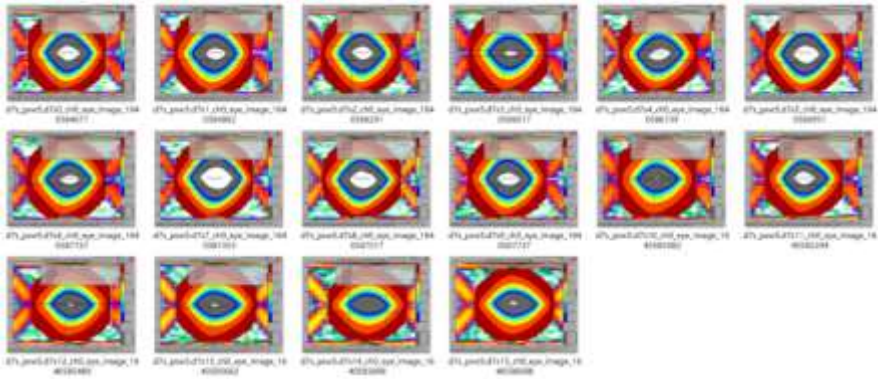


图 5-4

- **Gen5 Switch 卡左边提供 4 个 Gen5 x4 MCIO 接口。通过各种 MCIO 转接 U.2, U.3, EDSFF 的 Gen5 转接线缆可以连接测试各类常见接口的 PCIe Gen5 NVMe SSD。当然这 4 个端口是动态配置的，也就是说，如果用户自己定制或者购买 2*MCIO 转接 1*x8 插槽的转接治具，或者 4*MCIO 转接 1*x16 插槽，也可以用来分别测试 Gen5 x8 或者 Gen5 x16 的各种产品。**

下图展示了左侧通过 Gen5 MCIO/U.2 1x4 线缆连接 single port 盘，以及通过 Gen5 MCIO/U.2 2x2 线缆连接 dual port 的图片。顶部的插槽通过 SerialCables Gen5 U.2/AIC 转接卡测试 Gen5 U.2 single port 盘。

图 5-5

2023/5 月份 SerialCables 推出了仅支持连接 4 块 Gen5 SSD 的 Gen5 x16 switch 卡，去掉了顶部的 x16 插槽，但是 Broadcom Gen5 switch 为 B0 版本，支持 PCIe 协议抓包，配合 SerialTek 协议分析仪软件使用。

SerialCables 也推出了下行方向为 2 个 QSFP-DD 的 Gen5 x16 switch 卡， 主要是连接下行的 Gen5 SSD 扩展测试盘柜使用，当然也可以连接扩展 GPU 等板卡的扩展板。

图 5-6

图 5-7

综上，SerialCables 的 PCIe Gen5 卡板既可以插在主机 PCIe Gen5 x16 插槽测试 CPU 的 Gen5 建链能力，也可以测试插在顶部的 Gen5 x16 插槽的客户的 endpoint 板卡，例如 GPU, DPU, AI 卡，加速卡等，或者左边 4 个 Gen5 x4 MCIO 接口通过 MCIO 转接 U.2, U.3, EDSFF 等线缆实现对于各类 Gen5 x4 single port 盘和 dual port SSD 的测试。

Saniffer 公司本周刚拍摄并且处理了 Gen5 switch 卡的演示视频，展示了两张 Gen5 switch card 对接协商成 gen5 x16, 同时展示了 switch card 连接 dual port SSD 和 single port ssd 的端口以及热插拔演示。演示环境参见下图。

图 5-8

演示视频包括下面几个部分：

- **PCIe Gen5 x16 switch card 实物介绍**
- **实物连接主机、对接、连接 dual port ssd 和 single port SSD**
- **switch 卡内置的管理 MCU 支持的 CLI 命令行介绍，具体命令列表参见下图。**

MCU Commands List

Commands	Description
fdl	Update the configuration file or firmware for Atlas2 PCIe switch.
lsd	Shows switch temperature, host card consumed current, FAN speed, voltages and Side-band mode.
mw	Write 32bits data into any register as defined in Atlas2 switch
dr	Dump the values of Atlas2 switch for any register with specified address.
dp	Dump the values of Atlas2 switch for any register with specified port number.
df	Dump the values of Atlas2 flash with specified address.
ssdrst	Issue 300ms duration PERST# to attached devices in MCIO ports or straddlePCIe connector.
pwrdis	Set PWRDIS to H state (disable SSD power), or L state (enable SSD power)
hled	Turn ON/OFF the host LED inside EDSFF drive
showport	Show link status for USP in golden finger, DSP for MCIO ports and Straddle port.
setmode	Set MCIO ports bifurcation mode.
showmode	Show MCIO ports bifurcation mode in operating.
bist	On-board I2C devices diagnostic.
spread	Show spread information or set -0.5% SSC in PCIe reference clock to Atlas2 switch.
clk	Show the clock output status or disable/enable the clock output for all MCIO ports.
iicwr	SMBus data read from drive attached in MCIO port.
iicw	SMBus data write to drive attached in MCIO port.
ver	Shows card information, MCU FW and Atlas2 FW version.
reset	MCU FW reset (not including Atlas2 PCIe switch)

图 5-9

上述命令里面，对于测试 SSD 来讲 ssdrst 可以用来发送 300ms 的 PERST#让待测 SSD 或者插卡强行复位，spread 和 clk 配合可以用来测试盘的 SRIS 功能支持，iicwr 和 iicw 分别用来针对 SSD 通过 iic/smbus 总线进行读取和写入操作。Showport 可以用来检查上、下行的链路训练状态是否符合预期。

微信查找公众号"saniffer"在线查看演示视频

另外，SerialCables 的 Gen5 switch 卡使用 Broadcom PEX89000 Gen5 交换芯片，年底即将量产发布的基于 Broadcom B0 版本芯片的 switch 卡内置了 SerialTek 的 PCIe Gen5 协议抓包功能，可以实现对于初始化阶段碰到的各种 PCIe 底层问题进行抓包分析，具体介绍可以联系 Saniffer 公司或者访问 www.saniffer.com 官网下载的"PCIe Gen 4/5/6 总线协议和 NVMe SSD 测试技术和工具白皮书"，参照 2.8 章节获得更多信息。

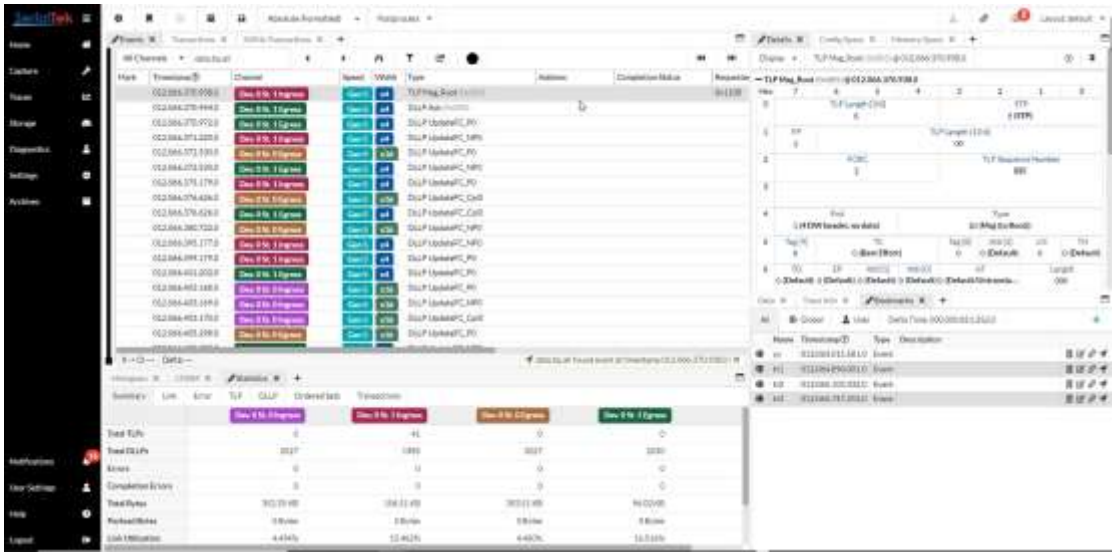


图 5-10

5.1.2 PCIe Gen5 Retimer 卡

目前业界正在开发中的 PCIe Gen5 服务器出于成本考虑会使用 Retimer 卡，所以，对于 Gen5 SSD 或者各种板卡在实验室测试的时候也需要构建这类测试环境进行提前测试，主要是要测试信号以及兼容性问题等方面是否有问题，以便可以提前进行问题修复。

注意：

- **Retimer 处理到协议层，其原理是串接在 PCIe Gen5 链路中间，将从 CPU 过来的差分信号串并转化后，进入 retimer 然后重新生成这些信号再转发出去到 endpoint 插卡或者 PCIe 背板，也就是说 retimer 通过这种方式实现了对于较弱的信号增强。**
- **Retimer 除了上下行分别连接 CPU 和 endpoint 板卡或者盘之外，有的场景出于 PCIe endpoint 扩展的需要，也会在下行方向先连接 PCIe Gen5 switch 然后即可连接多个 endpoint。我们使用 SerialTek Gen5 x16 协议分析仪抓取的 Gen5 CPU -> Gen5 retimer 卡 -> Gen5 x16 switch 以及 Gen5 CPU-> Gen5 retimer 卡 -> Gen5 x16 400GE 网卡的 trace 文件，对于 PCIe Gen5 协议底层 Retimer 和 switch 或者网卡加电 PCIe 初始化流程感兴趣的朋友可以联系我们索取该 trace 文件和解码软件。**

图 5-11

上图的 Gen5 retimer 是最常用的 Retimer，比较适合用各类 Gen5 x16 板卡测试，金手指上行连接 CPU，顶部插槽用来插入待测的 Gen 5x16 板卡。

下面是 Gen5 x16 retimer 卡，带 2 个 x8 QSFP-DD connector，有两款，芯片分别是 Astera 以及 Parade，基于 Montage 澜起的卡在开发中。

当然，如果是 Gen5 x4 SSD 或者 x8 EDSFF SSD，需要借助 SerialCables 公司的 Gen5 U.2/AIC 转接卡，或者 E3.S/AIC，E1/AIC 转接卡实现插入顶部插槽的目标。参见下图。

图 5-12

当然， SerialCable 也计划推出类似于上述第二种 Gen5 switch 卡的 Retimer 型号，提供 2 个 QSFP-DD 接口，用来将 Gen5 CPU 信号扩展到 Gen5 盘柜或者扩展板使用。

5.1.3 PCIe Gen5 各类转接卡和延长线

受制于现实测试环境的限制，在研发、测试阶段，经常需要在各类接口之间互相转接，或者将某些接口，例如 Gen5 x16 CEM 插槽，Gen5 x4 U.2 等进行延长，以便连接比较大的 Gen5 验证板，或者将 EP 板卡/SSD 放入温箱测试。在 PCIe Gen5 时代，这些转接或者延长的需求变得问题重重。我们平时看到实验室碰到的大量问题都和采用了劣质的转接卡和延长线有关。对于 Gen4 或者 Gen5 信号质量，国内常用电商平台买到的基本都不合格，这浪费了研发、测试工程师大量的时间。

限于篇幅，我们本文仅介绍一下常用的 Gen5 x16 延长线的一些基本信息，其它各种接口的转接卡和延长线，请直接联系 Saniffer 或者到官方网站 www.saniffer.com 下载 Saniffer PCIe Gen5 adapter, cable, switch and retimer cards quick guide_rev1.0 文档，下面是一个简要目录供参考，***具体产品速查请参考Chapter 11 部分，有详细的产品图片，型号和描述。***

- **GEN5 ADAPTERS 转接卡**
  - PCIe GEN5 U.2 ADAPTERS
  - PCIe GEN5 U.3 ADAPTERS
  - PCIe GEN5 EDSFF ADAPTERS
  - PCIe GEN5 OTHER ADAPTERS
- **GEN5 CABLES 转接线和延长线**
  - GEN5 MCIO CABLES
  - GEN5 EDSFF CABLES
  - GEN5 U.2 CABLES
  - GEN5 SlimSAS CABLES
  - GEN5 PCIE CEM CABLES
- **GEN5 SWITCH 卡**
- **GEN5 RETIMER 卡**

下面是开发各类 PCIe Gen5 x16 芯片和控制器经常要用到的延长线，尤其是早期开发阶段原型卡尺寸较大无法插入主机，那么这个时候必须采用 Gen5 延长线延长出来。

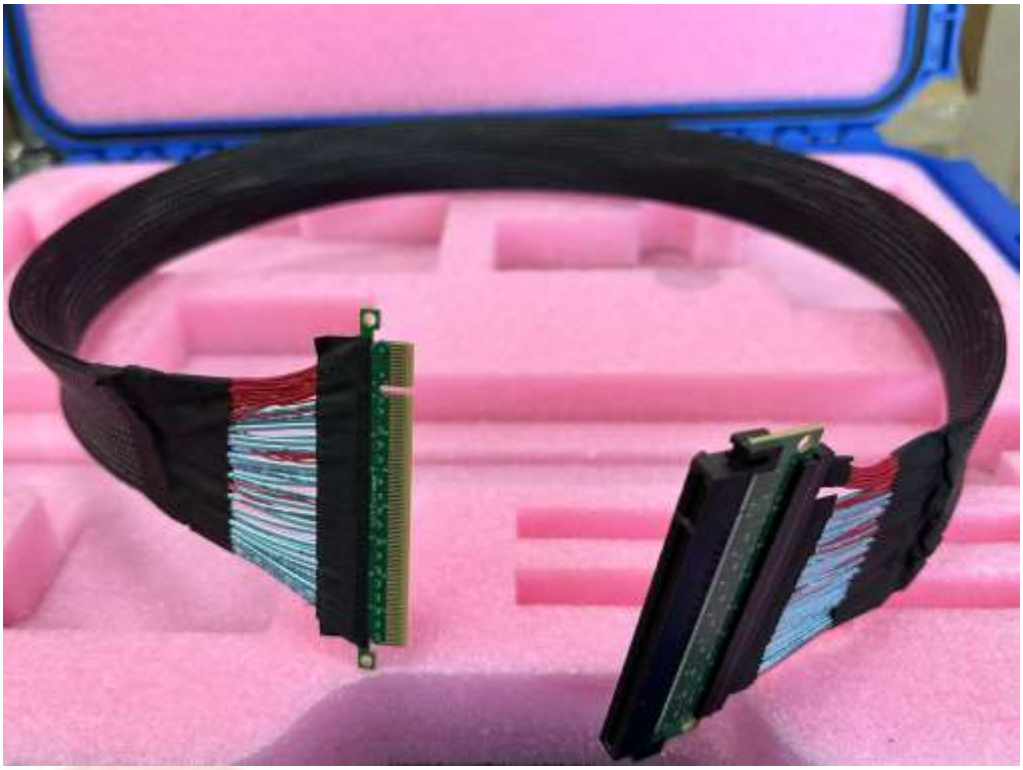


图 5-13

PCI-E X16 Gen 5 164P 延长电缆

- **小型 PCB 公连接器**
- **卓越的信号完整性性能**
- **阻抗：85+/-10% 欧姆**
- **FEXT 和 NEXT 功率总和：40dB 高达 25GHz**
- **符合 PCIe 至 CEM**
- **坚固的机械结构**
- **弯曲支撑**
- **提供灵活的版本**
- **电缆长度：支持延长线长度 0.4M, 0.5M, 0.7M, 1M**

Gen5 延长电缆的典型性能

- **先进的 crosstalk 串扰抑制技术支持 xtalk 的功率总和小于 40dB 到 25GHz**
- **阻抗控制在 85ohm+/-7ohm，反射<-10dB 最高 20GHz**
- **先进的 Twinax 电缆和 PCB 设计技术支持 4.5dB/m 的 IL**
- **在 16GHz 时最大损耗 6dB，以支持高达 1.0m/40 英寸的延伸范围 Max**

5.2 常用 PCIe Gen 4/5/6 Host 和 Retimer 卡

5.2.1 基于 Gen 5 BCM switch 的 Host Card

该卡基于 BCM PCIe Gen5 PEX89000 交换芯片实现，非常方便测试各种 Gen 5 device, 例如 Gen 5 SSD 控制器或者 SSD。

图 5-14

产品介绍

该 PCIe Gen 5 x16 主机卡基于 BCM PCIe Gen5 PEX89000 交换芯片实现，该 ASIC 芯片内置免费的 SerialTek PCIe Gen 5 协议抓包分析功能，可以帮助用户解决绝大部分芯片在 bring-up 过程中可能碰到的各种初始化相关问题。

该 Host Card 上行 upstream 为 Gen 5 x16 金手指，下行 downstream 提供两种接入方式：1）Gen 5 x16 插槽，可以插入各种芯片验证卡，例如，AI, NVMe SSD controller 等插卡；2）4 个 Gen 5 x4 external MCIO 接口，通过 MCIO 线缆可以转接成 U.2, U.3, EDSFF 等，参见下图。

图 5-15

该产品插入主机插槽，目前适用于下述测试场景：

- **Gen 5 CPU (Root Complex 端)**

如果用户在开发 Gen 5 的 CPU 或者 retimer, switch 等芯片，可以将该插卡理解成 End Point，这样可以测试 CPU, retimer, switch 是否和该插卡协商成 Gen 5 的速率，初始化过程是否有问题，兼容性和互操作性等，以及性能测试。

- **Gen 5 Device (End Point 端)**

大部分公司从事 Gen 5 芯片设计的都是 device 端，例如 PCIe Gen 5 NVMe SSD controller，加速卡，GPU 卡，AI 卡，高性能计算芯片等，这类芯片验证板一般都是直接插入该卡顶部的 Gen 5 x16 插槽来验证速度协商，bring-up 过程中各种问题，兼容性和互操作性等，以及性能测试。

对于 Gen 5 NVMe SSD 测试来讲，只要配置相应接口的转接线缆或者转接卡即可实现 Gen 5 测试。目前我们提供如下的转接线缆和转接卡。

- MCIO to U.2 (1x4)
- MCIO to U.2 (2X2)
- MCIO to U.3 (1x4)
- MCIO to U.3 (2X2)
- MCIO to EDSFF(1x4)
- MCIO to EDSFF (2X2)
- U.2 to AIC adapter
- U.3 to AIC adapter
- E.3 to AIC adapter

图 5-16

What is mini cool edge Io (MCIO)?

MCIO cables are designed for data center, networking and telecommunications markets that use SAS, PCIe, Ethernet and othersignal protocols. The solution can support cable to board and card to board ap plications in system, which include chip to chip, chiptomodule, chip to board and card edge option.
Mini Cool Edge IO(MCIO) is a flexible, robust and high performance connector and cable assembly solut ion that helps server and networking equipment design flexibility, reduces overall space, and extends th e reach for high data rate signals. MCIO cable assemblies are provided with both discrete and ribbon ra w cable 34AWG to 30AWG.

构建 PCIe Gen5 dual port SSD 仅找到一块 ssd 的解决办法小贴士

有的时候我们发现安装 redhat linux 或者 CentOS 后，通过 Gen5 switch card + MCIO to U.2 2X2 cable 连接支持 dualport 的 SSD，例如，Samsung PM1743。通过 lspci 可以找到两个 samsung nvme ssd，或者 nvmelist-subsys 是两个，但是 nvme list 显示只有一个。

一般通过在 CLI 下面键入下面的命令：
grubby --update-kernel=ALL -args="nvme core.multipath=N'
然后 reboot 后重新进入系统#提示符后，再次通过 nvme list 即可找到 2 块 gen5 x2 nvme ssd。

5.2.1.1 构建用于 PCIe Gen5 SSD 常温和温箱测试的批量测试可靠硬件平台

随着各大 SSD 厂商的 PCIe Gen5 SSD 的发布，以及这些产品即将进行市场发货，如何构建用于 PCIe Gen5 SSD 常温以及高低温批量测试的可靠硬件平台的任务逐渐摆在了 SSD 厂商研发测试部门以及即将使用 Gen5 SSD 的厂商测试部门的面前。

本文将带大家探讨如何构建一套可靠、稳定、良好 Gen5 信号的通用测试硬件，从而使得测试部门开发的各种针对 SSD 的性能、功能测试用例可以稳定地运行在上面。我们在 Sniffer 上海实验室搭建了用于本文的测试环境，Gen5 台式机主板通过 PCIe Gen5 switch 卡的 MCIO 接口延长 1 米后连接到*拟放入温箱的* Gen5 测试背板背面的 MCIO 接口，然后经过背板到正面的 Gen5 U.2 接口，然后再经过 U.2/E3.S 转接卡接到 Gen 5 E3.S SSD，我们可以看到整个链路还是可以非常稳定地工作在 Gen5 x4 状态。

具体视频讲解可以参考本文底部的视频链接。下图是我们本次搭建的环境一览。

图 1：构建 Gen5 SSD 测试环境组成一览

下面我们来逐一介绍一下 Gen5 SSD 测试环境硬件的每一个环节。

❶ Gen5 Switch Card

首先，为什么要使用 PCIe Gen5 switch 卡？这主要有下面几个方面的考虑：

❑ **提供批量测试的多个 Gen5 SSD 接入**

参见下图，SerialCables 公司的 Gen5 switch 卡提供 4 个标准的外接 MCIO 接口，工程师通过 Gen5 MCIO 延长线很容易拓展 4 个 gen5 SSD。

图 2：SerialCables Gen5 Switch 卡正面图

❑ **提供最好的 Gen5 信号质量**

参考《PCIe5.0, CXL, NVMe, NAND, DDR5, UFS4 测试技术和工具白皮书 Ver 10.0》的第 5.1 章节，我们平时实验室使用的常见品牌的 PC 或者工作站主机，无论是 Intel Gen5 CPU 还是 AMD Gen5 CPU，从 CPU 经过主板的 PCIe 差分信号线“走到”Gen5 x16 插槽的时候，信号就不好了。参见下图，我们看到 16 个 lane 的眼图大概总归有 4~5 个 lane 的眼图很不好。这样就导致如果直接通过转接卡连接盘的话很可能会导致“掉速”或“掉 lane”。

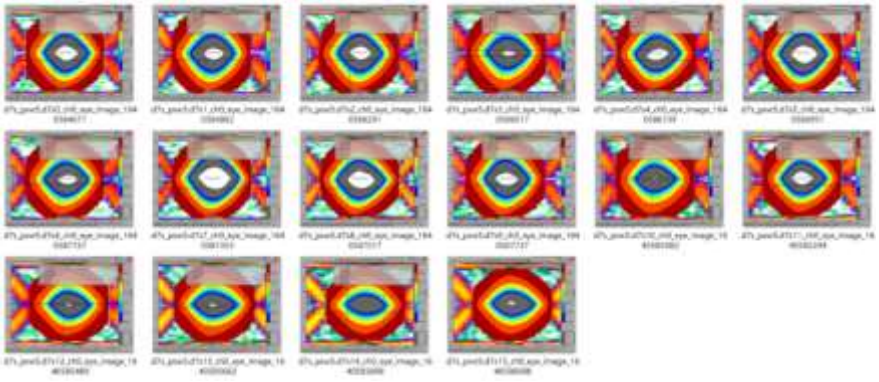


图 3：PCIe Gen5 PC 工作站的 x16 插槽输出信号眼图

参照下图，经过 Gen5 switch 后的信号质量，该图是上面的 Gen5 switch 顶部的 gen5 x16 插槽的输出眼图，或者是外接的 4xMCIO 的输出眼图（4 个 MCIO 可以自适应成 x16, 2x8 或者 4x4 使用）。

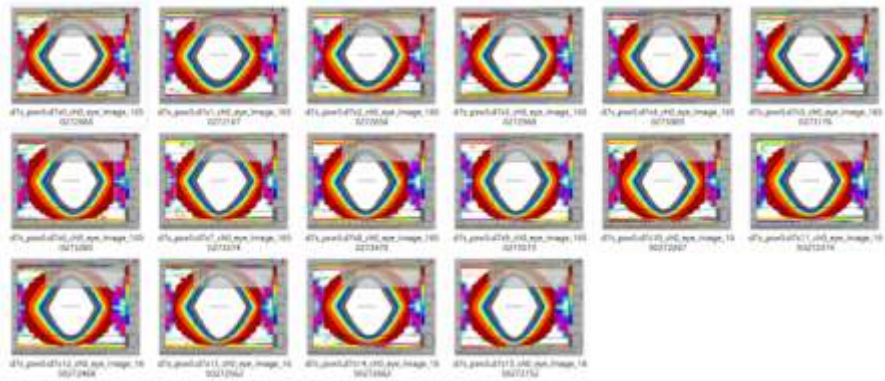


图 4：经过 Gen5 Switch 卡后的 x16（或者 4x4）的 16 个 lane 的输出信号眼图

这样的话，就可以提供待测的 SSD 相对较好的信号质量。另外要注意点的一点：尤其是企业级 Gen5 SSD，大部分情况下都是连接 PCIe Gen5 switch 背板。所以，使用 Gen5 switch 连接待测 Gen5 SSD 更加贴近真实应用场景。

- **提供热插拔能力****

这也是一个很重要的因素，即支持热插拔测试。在测试过程种如果某个或盘出现问题，那么需要将该盘在主机仍然工作的情况下拔出来，然后重新换一个新盘，然后接着测试。如果不使用 Gen5 switch 卡，例如将盘直接转接后连接 CPU，一般对于热插拔支持不好。当然，热插拔除了硬件支持外，也需要 OS, BIOS 和盘的支持。

为什么一般不使用 Gen5 retimer 或者 redriver 芯片来构建 Gen5 测试硬件呢？一个方面是这些芯片一般对于 PC 主板支持不好或者说不支持，即便采用服务器主机，也会面临诸多兼容性问题，简单来讲，你即便可以调通一个型号的 Gen5 SSD，但是可能换另外一个信号的 Gen5 SSD 可能就会出现问题，随意非常不适合搭建测试环境使用。

❷ Gen5 MCIO-MICO 1M cable

我们使用 4 根 1 米长的 MCIO to MCIO 的线缆连接 Gen5 SSD 测试背板，使用 1 米的线缆比较容易将测试背板放入到高低温箱里面。当然，如果搭建常温测试环境，也可以使用 0.5 米的 MCIO to MCIO 的线缆。参见下图为 SerialCables Gen5 MCIO to MCIO 延长线。

图 5：SerialCables 1 米 Gen5 MCIO to MCIO 线缆

❸ Gen5 Backplane

该 Gen5 SSD 测试背板采用符合 Gen5 信号质量的板材设计，所有 connector，包括 MCIO 和 U.2 均采用最新的大厂 Gen5 connector。参见下图。

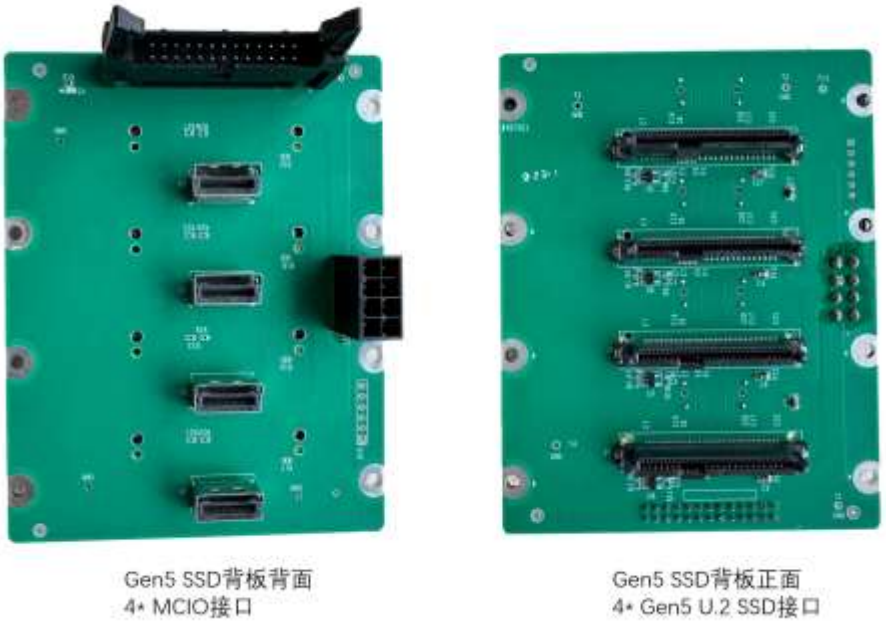


图 6：Gen5 U.2 背板背面和正面

为了方便将 Gen5 U.2 SSD 插入到❸ Gen5 Backplane 背板，尤其是该背板在温箱里面的时候，一般需要用户单独定制机构件，参见下面的示意图。

图 7：配合 Gen5 U.2 背板定制开发的机构件示例

前面的❶ Gen5 Switch Card 通过❷ Gen5 MCIO-MICO 1M cable（总计 4 根）分别从 Switch 卡的 4 个 MCIO 连接到该❸ Gen5 Backplane 背面的 4 个 MCIO 接口，然后通过背板到达正面的对应的 4 个 Gen5 U.2 SSD 接口，如果是 Gen5 E3.S SSD 或者 Gen5 M.2 SSD 则需要通过转接卡转接到 U.2。本文示例即使用 Gen5 E3.S SSD 进行演示。

下图是 MCIO cable 连接到背板背面的实际图片。其中左侧的线束为给四张 SSD 的供电线缆，右侧的扁平线缆为给每张盘的 sideband signal 进行拉高、拉低的控制线缆。中间的网纹线缆为 Gen5 MCIO to MCIO 线缆，另外一端为 Gen5 switch 卡的 MCIO 接口。示例图中为连接的端口 3（从上面到下面以为为端口 0, 1, 2, 3）。

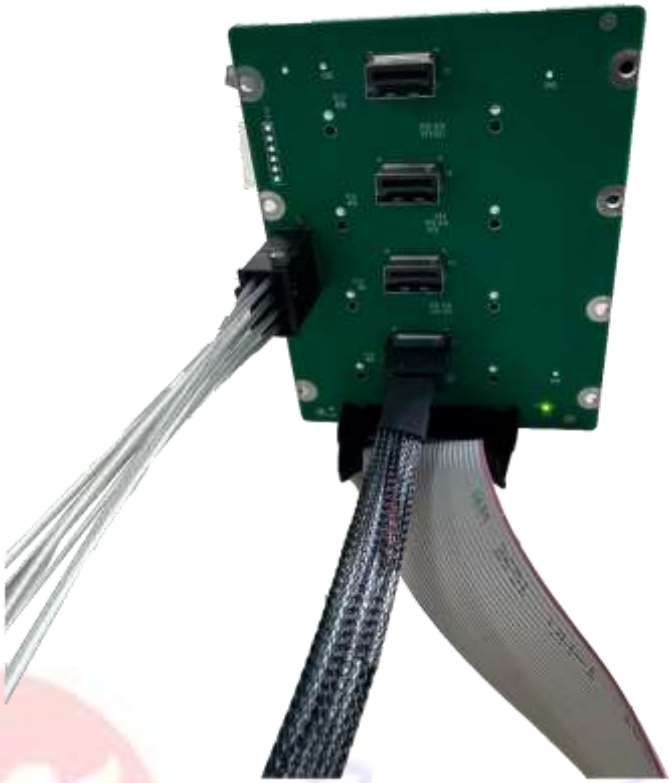


图 8：PCIe Gen5 背板的背面（4 个 MCIO 接口连接 Switch 卡）

下图为 Gen5 背板的正面，我们在最下面的 Gen5 U.2 接口连接了一块 Gen5 SSD。但是该 Dell EMC (OEM Kioxia) SSD 是 Gen5 E3.S 接口，所以我们通过一块 Gen5 U.2 to EDSFF 转接卡将该盘转成 U.2 以后插入到背板最底下的 3 号口 U.2 插槽。

图 9：PCIe Gen5 背板的正面（Gen5 E3.S SSD 经过 E3/U.2 转接卡插在 slot 3）

❹ Gen5 E.3/U.2 adapter

正常情况下测试 Gen5 U.2 SSD 无需此转接卡，直接将 U.2 SSD 插入测试背板即可。本次由于我们实验室暂时没有 Gen5 U.2 SSD，所以不得已采用 Gen5 E3.S SSD，所以我们必须采用此转接卡将 E3.S male 转结成 U.2 male 从而插入测试背板，下图是我们的实拍图。

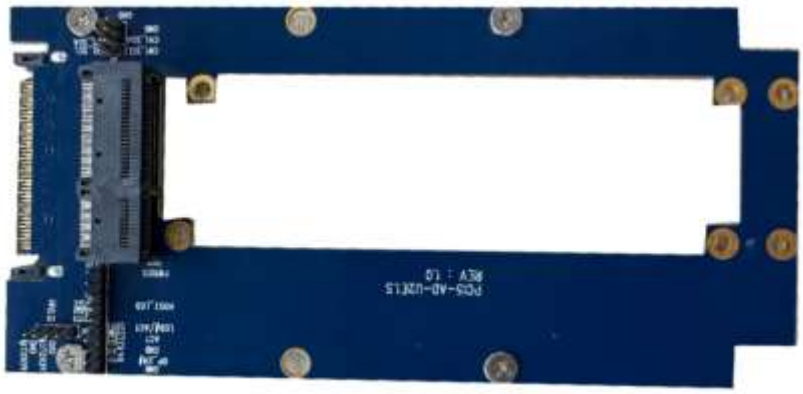


图 10：SerialCables Gen5 U.2 to EDSFF 转接卡

当然，如果测试的 Gen5 M.2 SSD，那么需要下面的 Gen M.2 to U.2 adapter

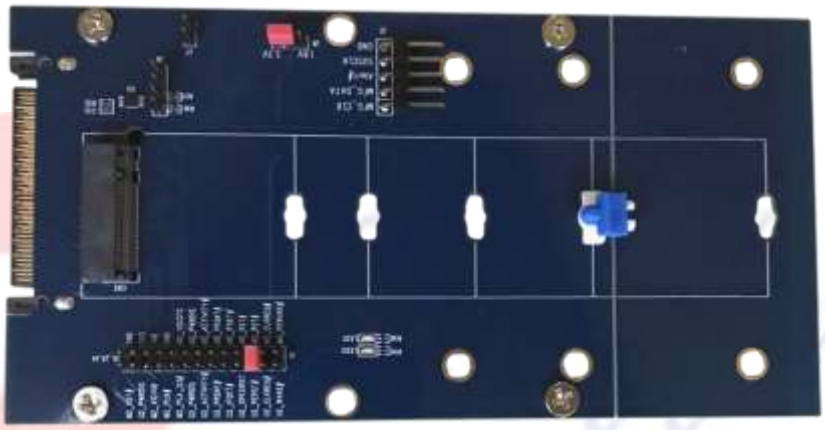


图 11：SerialCables Gen5 U.2 to M.2 转接卡

❺ Gen5 E.3 NVMe SSD

下图是我们本次演示使用的 Dell EMC (OEM Kioxia) Gen5 E3.S SSD。参见下图。

图 12：Dell EMC (OEM Kioxia) Gen5 E3.S SSD

❻ Power Control Module

下图是针对 Gen5 SSD 进行掉电/上电、电压拉偏、电压/电流/功耗量测，sideband 信号控制的控制模块。参见下图。左边的图片为接入了 USB TYPE-C 管理线缆；右边的图片展示了连接 Gen5 SSD 背板的电源线和 sideband 边带信号控制线缆。

图 13：电源控制模块的侧视图

❼ Signal Control Ribbon Cable

❽ Power Supply Cable

❾ Type-C USB management

上述三个部分请参见❸ Gen5 Backplane 背板背面图中对应的三根线缆的介绍。

❿ Micro-USB management

我们通过该 Micro-USB 接口使用 Micro-USB to USB Cable 连接到控制电脑（例如笔记本），实现对于 Gen5 switch 的管理和配置，例如：查看上下行的速率和 lane 等建链情况；对于 SSD 进行掉电、上电；发送 300ms PERST#拉高、拉低信号到 SSD；通过 SMBus 发送相关的命令给 SSD；测试 SRIS 等时钟支持模式；对于 switch 进行重置复位等操作。具体常用 CLI 命令参见下图。

Commands	Description
fdl	Update the configuration file or firmware for Atlas2 PCIe switch.
lsd	Shows switch temperature, host card consumed current, FAN speed, voltages and Side-band mode.
mw	Write 32bits data into any register as defined in Atlas2 switch
dr	Dump the values of Atlas2 switch for any register with specified address.
dp	Dump the values of Atlas2 switch for any register with specified port number.
df	Dump the values of Atlas2 flash with specified address.
ssdrst	Issue 300ms duration PERST# to attached devices in MCIO ports or straddlePCIe connector.
pwrdis	Set PWRDIS to H state (disable SSD power), or L state (enable SSD power)
hled	Turn ON/OFF the host LED inside EDSFF drive
showport	Show link status for USP in golden finger, DSP for MCIO ports and Straddle port.
setmode	Set MCIO ports bifurcation mode.
showmode	Show MCIO ports bifurcation mode in operating.
bist	On-board I2C devices diagnostic.
spread	Show spread information or set -0.5% SSC in PCIe reference clock to Atlas2 switch.
clk	Show the clock output status or disable/enable the clock output for all MCIO ports.
iicwr	SMBus data read from drive attached in MCIO port.
iicw	SMBus data write to drive attached in MCIO port.
ver	Shows card information, MCU FW and Atlas2 FW version.
reset	MCU FW reset (not including Atlas2 PCIe switch)

图 14：Gen5 x16 Switch 卡的管理命令行 CLI

** 如果不考虑将 Gen5 SSD 放入温箱，那么在常温下面，也可以考虑将上文的 Gen5 U.2 背板换成 8 槽位 E3.S Gen5 SSD 背板的测试盘柜，参见下图。

图 15：SerialCables Gen5 8-bay JBOF 测试盘柜（E3.S 背板）

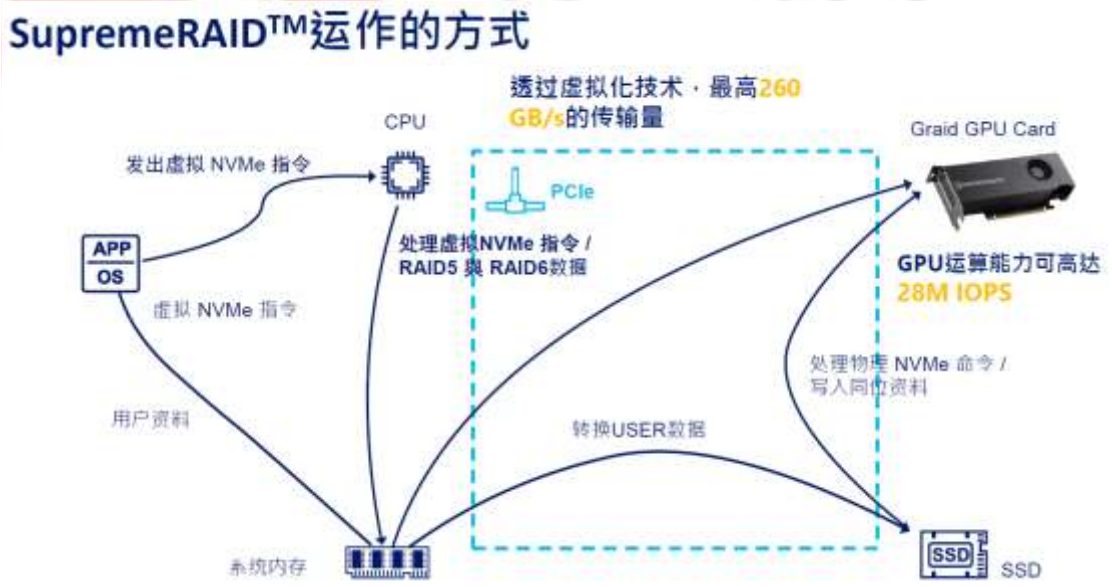
注意：由于一张 Gen5 switch 卡提供 4 个 MCIO，如果希望盘柜的 8 个 SSD 接口都可以使用，那么需要配置 2 块 Gen5 switch 卡。

5.2.1.2 如何使用 Gen5 switch 卡（和 Gen5 JBOF）构建 RAID 5/6 高性能生产或测试环境

考虑到业内常用的硬件 RAID 方案和软件 RAID 的各种短板，我们推荐采用 GRAID 的硬件 RAID 实现快速，简易，高性能、高可靠性的生产或者测试环境。除了将 GRAID 卡插入存储服务器实现针对 4-24 块 PCIe Gen5 NVMe SSD 的 RAID 方案外，也可以通过 1）普通服务器或者 PC + SerialCables Gen5 switch 卡 + 4 块 Gen5 SSD +GRAID 卡，或者 2）普通服务器或者 PC +2* SerialCables Gen5 switch 卡 + 8 块 Gen5 SSD （这里也可以使用 8 槽位 SerialCables Gen5 JBOF 盘柜） +GRAID 卡的实现，下面是方案 1）的一张测试图片，以及 GRAID 卡的一些产品信息供参考。

注意：RAID 可以管理的 NVMe SSD 可以是直接连接 CPU 的 SSD，或者经过 PCIe Switch 的 SSD，或者是远端映射的 SSD，例如 NVMoF 映射到主机的 SSD。

下面是 GRAID 的工作原理：



AI 训练通常对存储的访问性能有更高的需求，因为在训练过程中需要频繁读取和写入大量的数据，例如训练数据集、模型参数等。这需要高吞吐量和低延迟的存储系统来支持训练任务的高效执行。相比之下，推理阶段对存储的访问需求较小，因为推理通常涉及加载已经训练好的模型并对输入数据进行预测，因此需要的数据量相对较小且访问次数较少。

下图是 GPU server 通过网络访问一个高性能 NAS，NAS 内部使用了 GRAID 卡提供读写吞吐量和降低延迟。

SupremeRAID™ Use Case

深度学习

SupremeRAID™提升
深度学习训练效率

挑战：

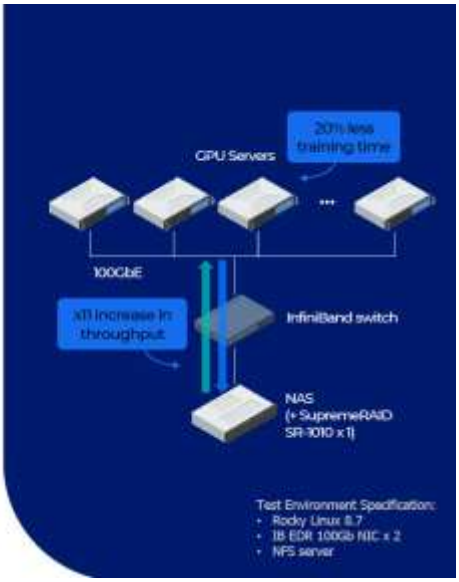
由于目前储存的瓶颈导致了高度延迟性

方案：

使用20台GPU Server进行深度学习，SupremeRAID™作为Targer端使用

结果：

- 提升了**11倍**的性能提升（从 8Gbps 到 88Gbps）



下面是 GRAID 目前提供的几种卡的型号以及支持的最大的 NVMe SSD 数量。

下面是使用 GRAID SR-1000 创建一个包含 5 个 NVMe SSD 的 RAID 的卷。

To create a RAID-5 virtual drive with 5 NVMe SSDs:

Step 1 Create a physical drive.

```
$ sudo graidctl create physical_drive /dev/nvme0-4
```

Step 2 Create a drive group.

```
$ sudo graidctl create drive_group raid5 0-4
```

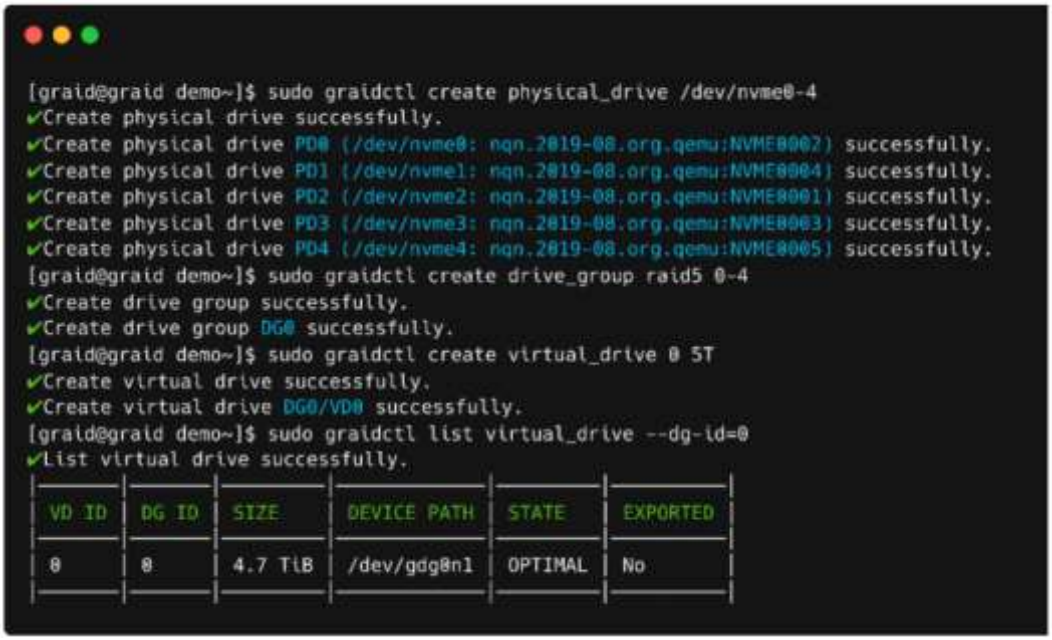
Step 3 Create a virtual drive with a 5TB volume size.

```
$ sudo graidctl create virtual_drive 0 5T
```

Step 4 Check the device path of the new virtual drive.

```
$ sudo graidctl list virtual_drive --dg-id=0
```

下面是上述创建过程的 CLI 命令行输出截图。



创建好 virtual drive 以后，就可以和使用正常的 NVMe SSD 一样，通过下面的步骤，就可以对于 SSD 进行读写操作了。

通常在使用 SSD 之前，需要进行以下步骤：

1. 分区： 将 SSD 分成一个或多个逻辑分区。可以使用工具如`fdisk`、`parted`或者图形化分区工具来进行分区操作。

2. 创建文件系统： 在每个分区上创建文件系统，以便 Linux 可以在其上进行文件存储和管理。常用的文件系统包括 Ext4、XFS、Btrfs 等。可以使用命令如`mkfs.ext4`、`mkfs.xfs`等来创建文件系统。

3. 挂载文件系统： 将创建好的文件系统挂载（mount）到 Linux 的目录结构中的某个位置，以便可以访问其中的文件。可以使用`mount`命令手动挂载，也可以在`/etc/fstab`中配置自动挂载。

这些步骤确保了 SSD 被正确地配置和准备，可以在 Linux 系统中进行使用。文件系统挂载（mount）的下面这些步骤有助于确保在 SSD 上进行文件拷贝时数据的正确性和稳定性。

1.文件系统检查：可以使用诸如`fsck`命令之类的工具对 SSD 上的文件系统进行检查和修复，确保文件系统的一致性和完整性。

2.文件拷贝：使用命令行工具（如`cp`、`rsync`等）或图形界面文件管理器，在 Linux 中进行文件拷贝操作。

3.同步缓存：在完成文件拷贝后，可以使用`sync`命令强制将文件系统缓存中的数据写入 SSD，以确保数据的完整性和一致性。

5.2.2 基于 Gen 4 BCM Switch 的 Host Card

请参见前面章节的描述，下图汇总的三张 Gen 4 Host Card 采用 Broadcom 交换芯片：

- **图 2 PCIe Gen 4/5/6 x16 Host 卡（带 2 个 Gen 4 x8 internal port），**
- **图 5 PCIe Gen 4/5/6 x16 Host 卡（带 4 个 Gen 4 x4 external cable port）**
- **图 6 PCIe Gen 4/5/6 x16 Host 卡（带 M.2 和 Slot）**

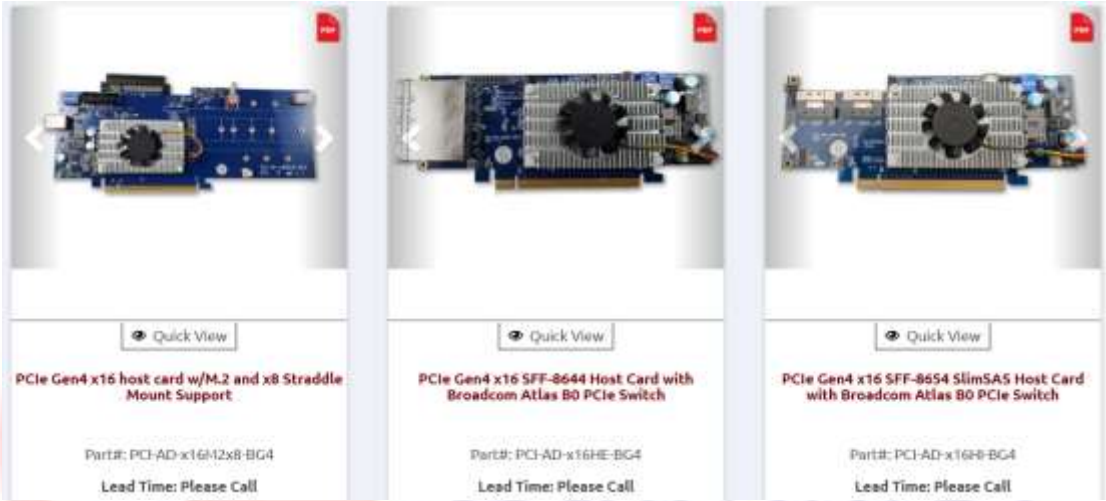


图 5-17

HD-MINI-SAS 外连和 Slim-SAS 内连 Gen4 switch 卡的连接方式如下。

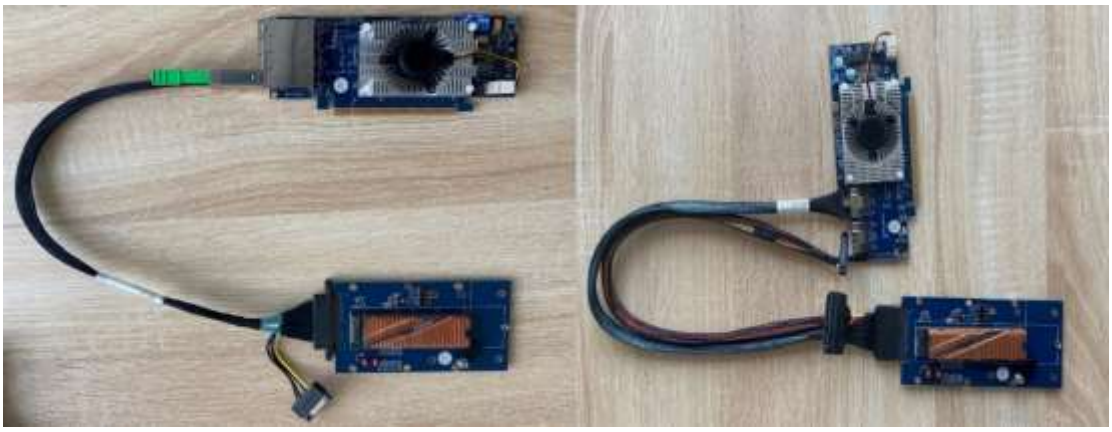


图 5-18

5.2.3 基于 Gen 4 Microchip Switch 的 Host Card

该基于 Microchip switch 芯片的 Host Card 提供了更高的灵活性和性价比，目前有 4 种规格参见下图：

图 5-19

5.2.4 基于 Gen 4/5 Retimer 芯片的插卡

SerialCables 公司新发布的 PCIe Gen 4/5/6 Retimer 卡(Part #:PCI4-AD-x16HE-RT-A)，支持 SRIS/SRNS，热插拔，以及 PCIe 分叉功能。参见下面的描述和图片。Retimer 提供比 Host Card 更加经济的 PCIe Gen 4/5/6 NVMe SSD 测试方案，可以直接接入 4 根 HD-MINI-SAS to U.2 线缆连接 4 个 Gen 4 NVMe SSD。

*Description: PCIe Gen4 x16 Retimer Host Board. Based on the PT4161L x16 PCIe 4.0 Smart Retimers. ***Supports SRIS/SRNS and common clock, hot-plug and bifurcation. ***Can only bifurcate down to the minimum lane count as the host system BIOS.*

图 5-20

5.3 常用 PCIe Gen 4/5/6 JBOF 测试盘柜

我们提供 active 和 passive 两种用于实验室测试盘柜，有的时候称为 JBOF (just bunch of flash)或者 enclosure，又细分为 active enclosure (带 pcie switch 芯片)和 passive enclosure（不带 pcie switch 芯片）。 每种又提供 U.2, U.3 等不同接口的 NVMe SSD 支持。该盘柜具体通过 USB 串口或者千兆以太网实现管理功能，可以远程通过 python 实现

针对每个盘位进行掉电，或者切换 single port/dual port 盘位配置，以及查看上、下行 PCIe 链路状态等功能。

5.3.1 PCIe Gen5 Passive 盘柜

下面是 SerialCables 公司的 PCIe Gen5 passive enclosure 图片。

图 5-21

5.3.1.1 Gen5 Passive 盘柜前、后面板接口介绍

图 5-22

5.3.1.2 Gen5 SSD 各种接口转接后连接背板示意图

下图是标配的盘盒的图片，支持标准的 Gen5 E3.S SSD，例如 Kioxia CM7 E3.S SSD。

E3.S SSD 通过 4 颗螺丝固定在盘盒上，然后推入盘柜背板，盘柜背板默认是 EDSFF 插槽。

图 5-23 Gen5 SSD 盘盒

如果需要测试 U.2, M.2，那么需要另购的的转接卡，即 E3/U.2, E3/M.2 转接卡，如下。

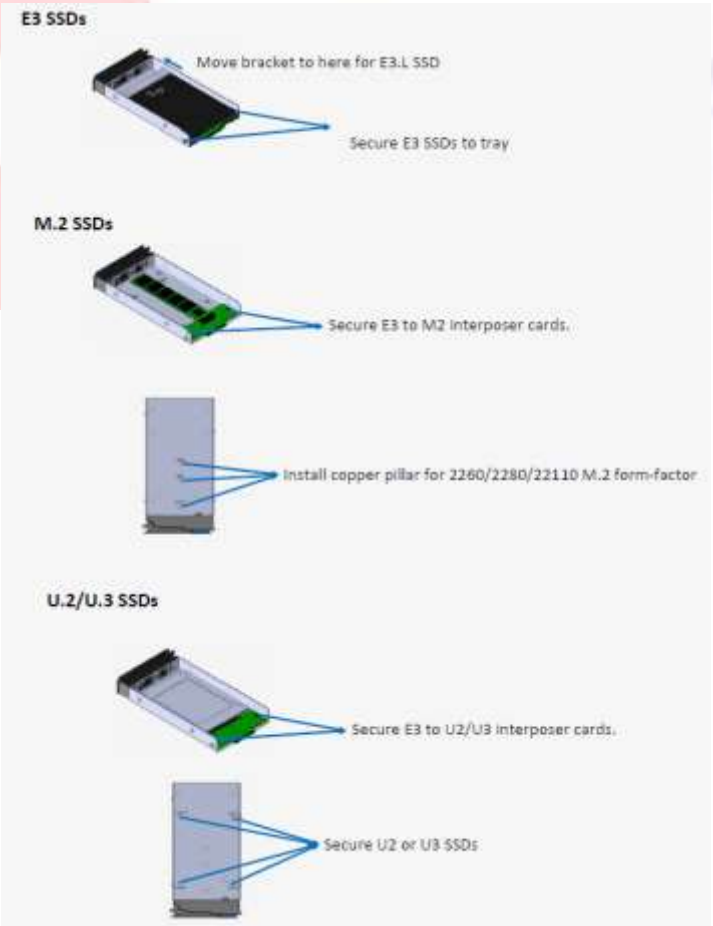


图 5-24

5.3.1.3 Gen5 Switch 卡连接 Passive 盘柜示意图

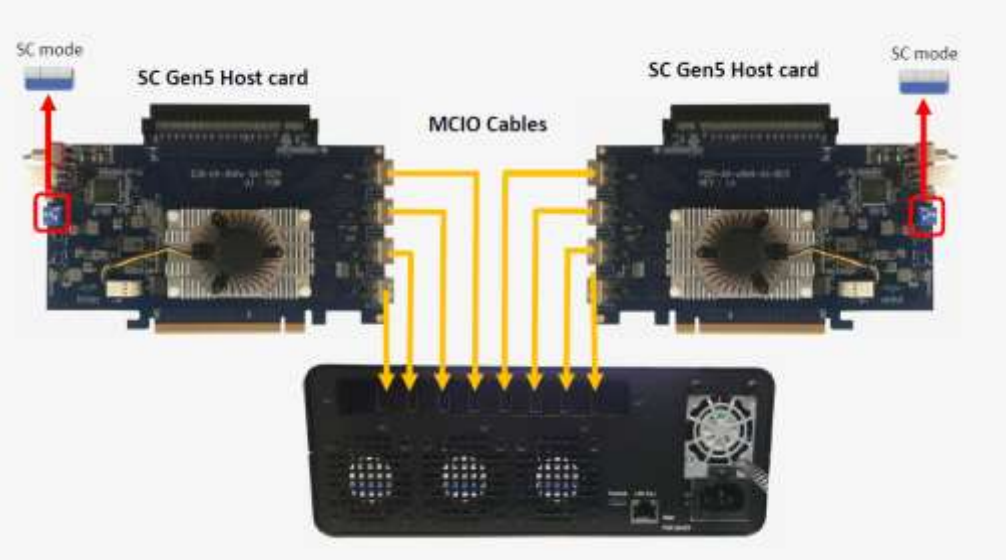


图 5-25

5.3.1.4 Gen5 Enclosure CLI 管理行管理接口

Commands	Description
help	Show list of commands
syspwr	NVMe JBOF enclosure power ON/OFF control
eth	Ethernet IP configuration
dhcp	Ethernet DHCP function control
setmac	Set Ethernet MAC address
fdl	Update PCIe switch config/FW or MCU FW
lsd	Show environmental info, including temperatures, FANs, PSUs, voltages.
ssdpwr	Control the power of each EDSFF slot.
showslot	Show slot information
ssdrst	To reset SSDs which install in slots
setmode	Set enclosure mode
showmode	Show enclosure mode
dual	Set dual channel enable on/off
pwrdis	Set pwrdis in slot pin3 level to high/low
hled	EDSFF drives HLED control
buz	buzzer control
bist	On-board devices diagnostic
iicwr	I2C read/write
iicw	I2C write
ver	Show on-board mcu F/W information
sysinfo	Show system information
reset	Reset JBOF

图 5-26

5.3.2 PCIe Gen4 Active 盘柜

下面是 SerialCables 公司的 PCIe Gen4 两种 enclosure 图片。

图 5-27

下面是两种盘的连接方式示意图。

5.3.2.1 Active enclosure 和 Host Card 的连接示意图

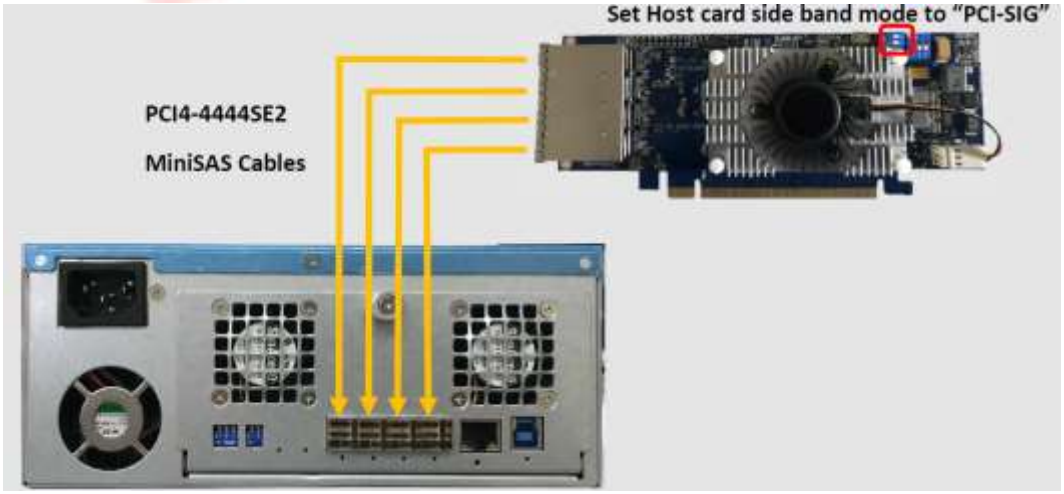


图 5-28

5.3.2.2 Active enclosure 和 Host Card 的实拍连接图

5.3.2.2.1 盘柜实际连接拓扑

图 5-29 Active 盘柜和 Host Card 的连接方式

5.3.2.2.2 主机、SerialTek PCIe 分析仪实际连接拓扑

下面是上述测试环境配合 SerialTek PCIe Gen4 分析仪实际连接拓扑图。

注意：该分析仪支持分析 single port 和 dual port，可以同时抓取两个 port 的数据。

图 5-30

5.3.2.2.3 Gen4 Enclosure CLI 命令行管理接口

这些命令展示了 JBOF 的主要管理功能，包括掉电、上电 ssdpwr，dual port 设置，PERST 拉低、拉高 ssdrst，SMBUS 管理 IICWRITE 等可以写 VPD 信息，整机掉电测试 reset，以及 showport 可以随时查看上、下行 PCIe 链路的协商状态等。

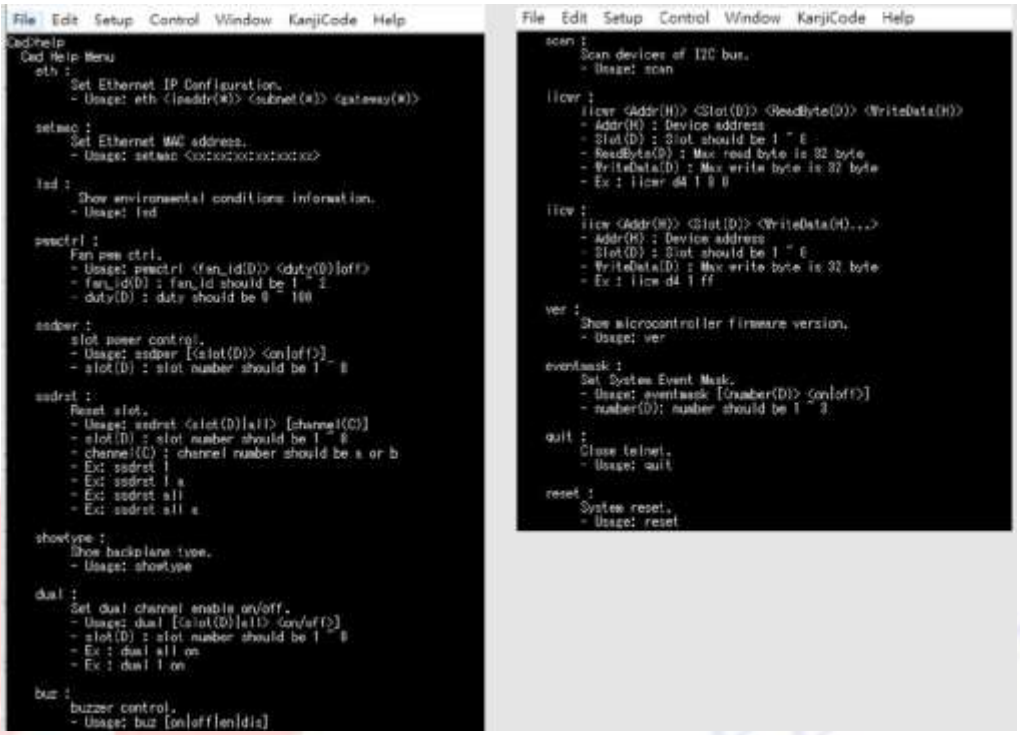


图 5-31

5.3.3 PCIe Gen4 Passive 盘柜

5.3.3.1 Passive enclosure 和 Host Card 的连接示意图

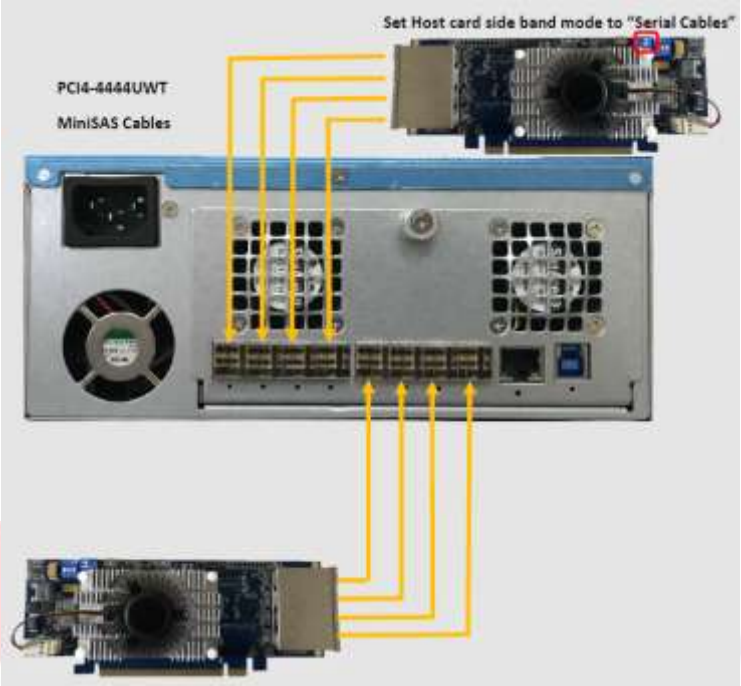


图 5-32

说明：一张 Host Card 只能连接 4 张 SSD，如果 8 个盘位都需要接盘，那么需要 2 张 Host Card。

下面是两种盘柜的实际连接的照片，注意两者使用的线缆是不一样的，都是从 Host 卡连接到盘柜后面板的接口。

5.3.3.2 Passive enclosure 和 Host Card 的实拍连接图

图 5-33 Passive 盘柜连接方式（使用特殊定制线缆）

5.4 常用 PCIe Gen 4/5/6 转接卡

请参见前述下图：

- **图 1 PCIe Gen 4 x4 M.2 to AIC 转接卡**
- **图 4 Gen 4 x4 M.2 to U.2 转接卡**
- **图 9 Gen 4 U.2 to AIC 转接卡（竖插）**
- **图 10 Gen 4 U.2 to AIC 转接卡（横插）**

5.4.1 Gen 5 转接卡

5.4.1.1 Gen5 U.2 转接卡

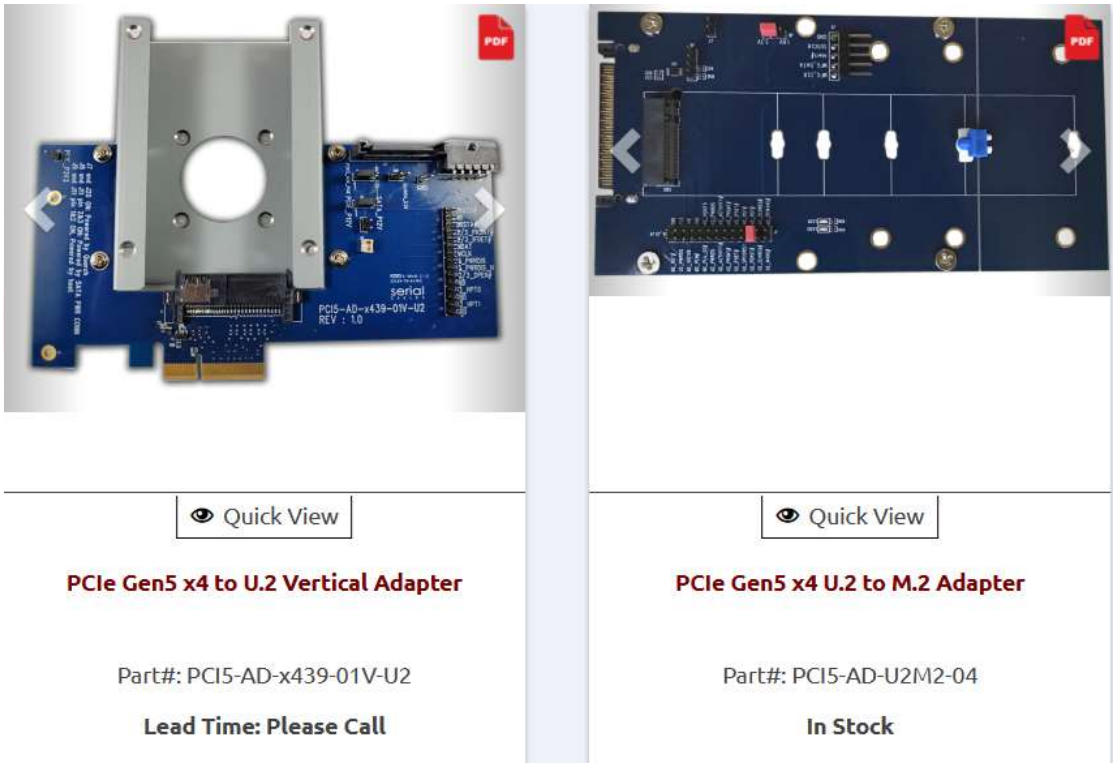


图 5-34

5.4.1.1.1 Intel Demos Lightning Fast 13.8 GB/s PCIe 5.0 SSD with Alder Lake PLUS SerialCables Gen5 U.2/AIC adapter

By Anton Shilov

published December 31, 2021

https://www.tomshardware.com/news/intel-demos-138-gbps-pcie-50-ssd-with-alder-lake

Intel shows off lighting-fast PCIe 5.0 SSD with Alder Lake

图 5-35

(Image credit: Intel/Ryan Shrout)

Intel has demonstrated how its Core i9-12900K Alder Lake processor can work with Samsung's recently announced PM1743 PCIe 5.0 x4 SSD. The result is as astonishing as it is predictable: the platform demonstrated approximately 13.8 GBps throughput in the IOMeter benchmark.

Intel planned to show the demo at CES, however, the company is no longer going in person. So, Ryan Shrout, Intel's chief performance strategist, decided to share the demo publicly via Twitter.

The system used for the demonstration included a Core i9-12900K processor, an Asus Z690 motherboard and an EVGA GeForce RTX 3080 graphics board. Intel hooked up Samsung's PM1743 SSD using a special PCIe 5.0 interposer card and the drive certainly did not disappoint. From a practical standpoint, 13.8 GBps may be overkill for regular desktop users, but for those who need to load huge games, work with large 8K video files or ultra-high-resolution images will appreciate the added performance. However, there is a small catch with this demo.

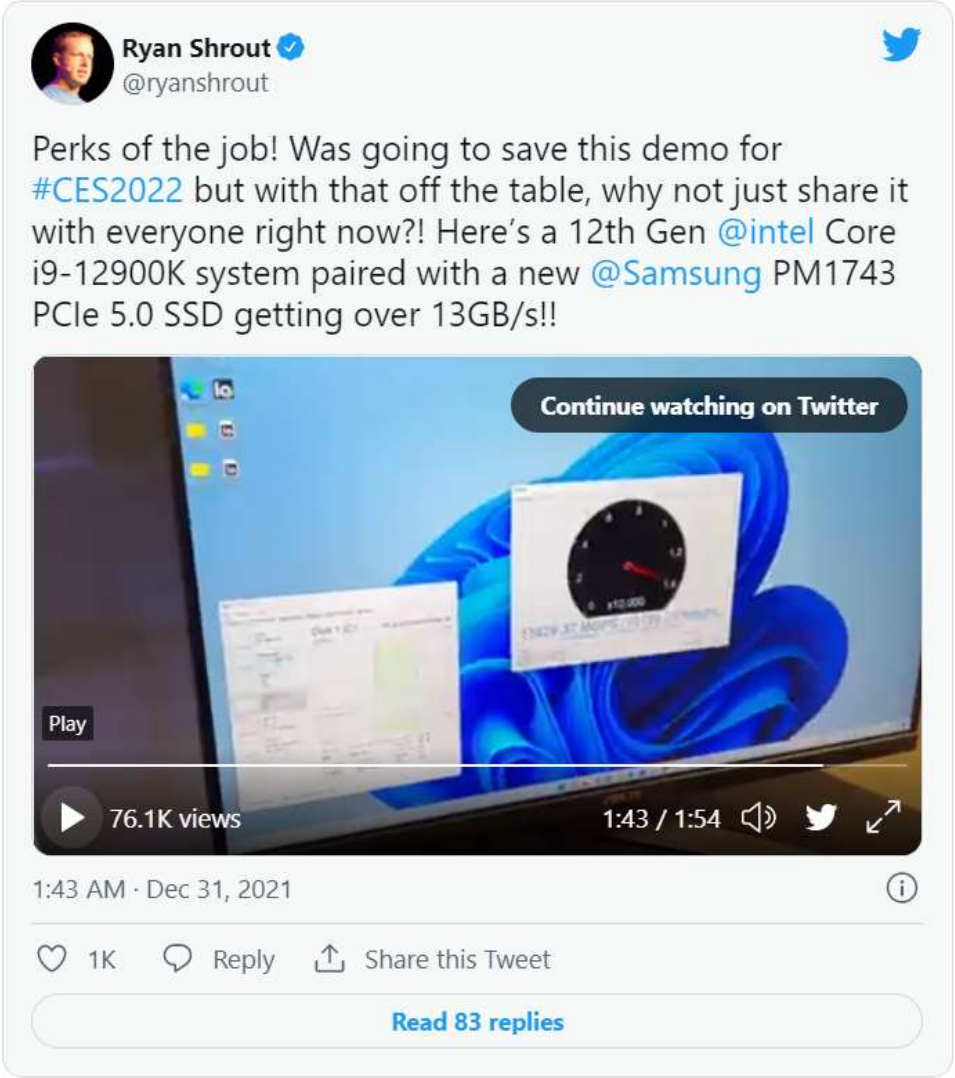


图 5-36

Apparently, Samsung will be among the first to ship its PM1743 PCIe 5.0 drives, which is why Intel decided to use this SSD for the demonstration. But Samsung's PM1743-series is aimed at enterprises, so it will be available in a 2.5-inch/15mm with dual-port support and new-generation E3.S (76 × 112.75 × 7.5mm) form-factors, so it is not aimed at desktops (and Intel admits that).

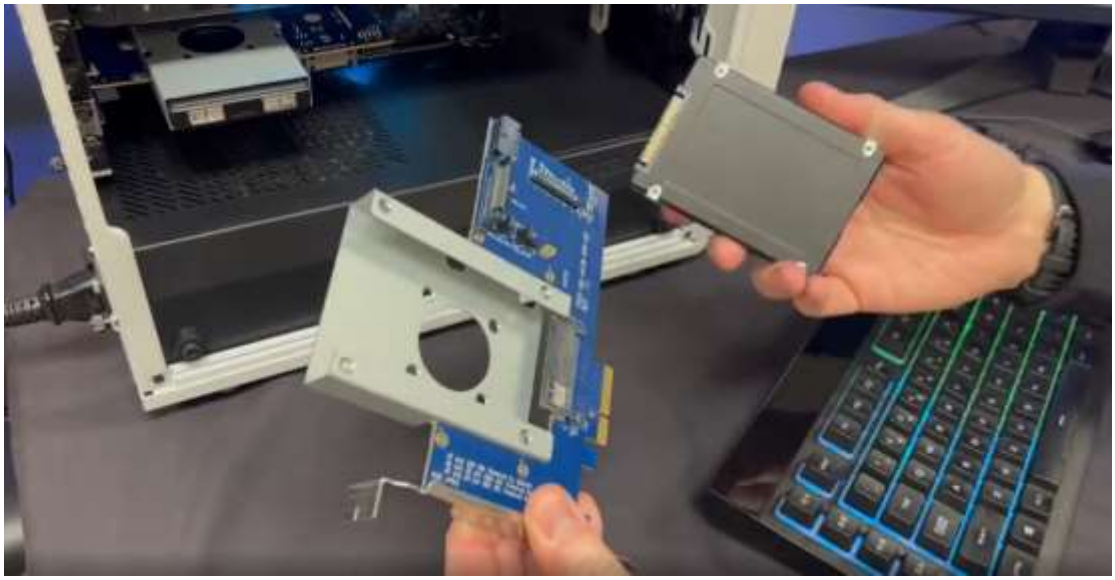


图 5-37

Meanwhile, there are several companies preparing PCIe 5.0 SSDs and SSD controllers for client PCs (Adata,Phison, Silicon Motion), so we are probably going to learn details about their hardware or even see it in action next week. Eventually, these drives will join our ranks of the best SSDs, though we are not sure when.

One of the main selling points of Intel's 12th Alder Lake processors is undoubtedly PCIe 5.0 support. Unfortunately, Intel could not demonstrate any benefits of the next-generation interface back in November during the platform's launch since there were no PCIe 5.0-capable graphics cards and SSDs on the market at the time.

5.4.1.2 Gen5 U.3 转接卡

Quick View

PCIe Gen5 x4 to U.3 Vertical Adapter

Part#: PCI5-AD-x439-01V-U3

Lead Time: Please Call

图 5-38

5.4.1.3 Gen5 EDSFF 转接卡

Quick View

PCIe Gen5 x8 slot to E3 EDSFF Vertical Adapter

Lead Time: Please Call

Choose an option

Choose an option
E3.S/L & E3.S.2T/E3.L.2T
E1.S 5.9mm
E1.S 8.01mm & 9.5mm

图 5-39

上图的 PCIe Gen5x8 slot 转接 E3 适配卡的正视图和平视图高清图如下。

E3

图 5-40

5.4.1.4 Gen 5 其它转接卡

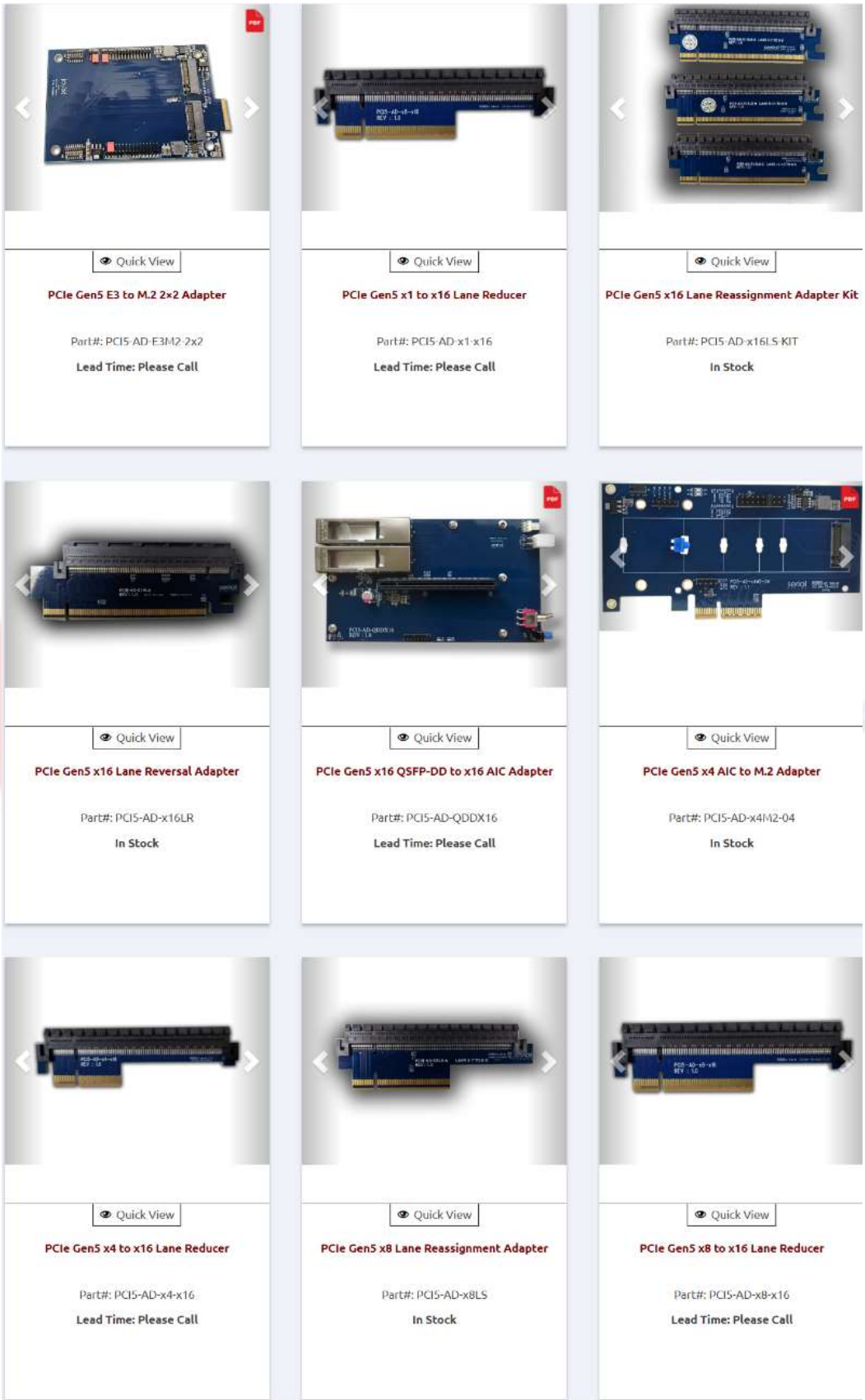


图 5-41

5.4.2 Gen 4 转接卡

5.4.2.1 Gen4 U.2 转接卡

图 5-42

5.4.2.2 Gen4 U.3 转接卡

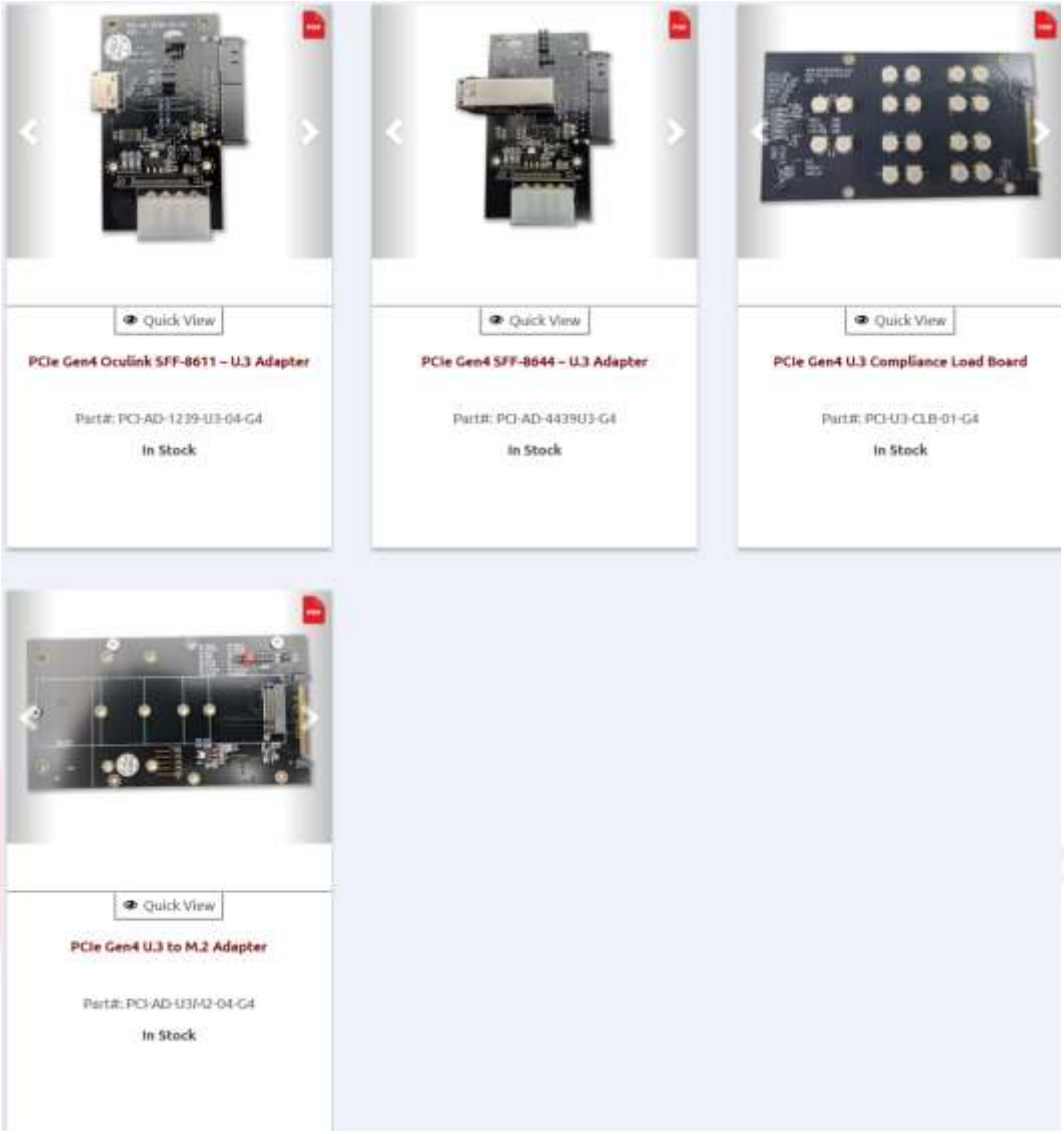


图 5-43

5.4.2.3 Gen4 其它转接卡

图 5-44

5.5 常用 PCIe Gen 4/5/6 转接线

5.5.1 Gen 5 转接线缆

5.5.1.1 Gen5 MCIO 线缆

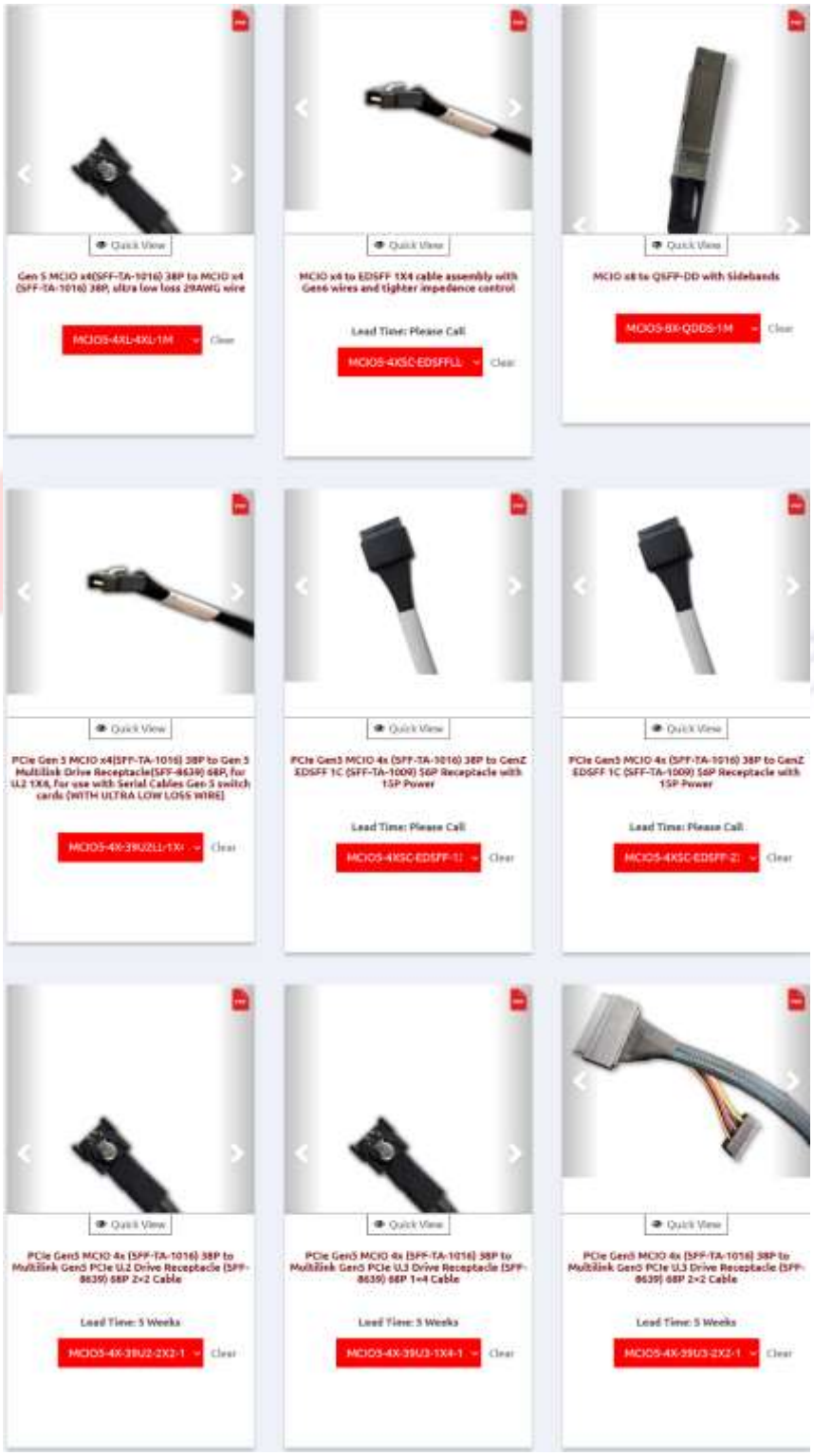


图 5-45

5.5.1.2 Gen5 EDSFF 线缆

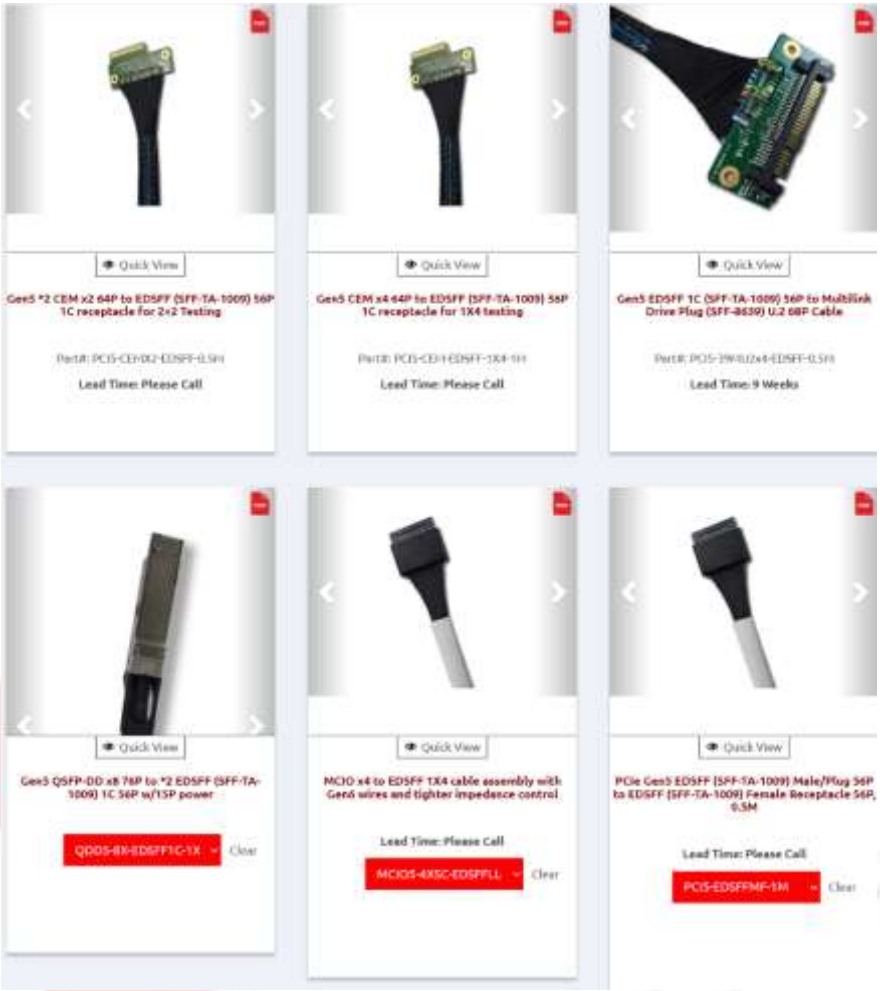


图 5-46

5.5.1.3 Gen5 U.2 线缆

图 5-47

5.5.1.4 Gen5 SlimSAS 线缆

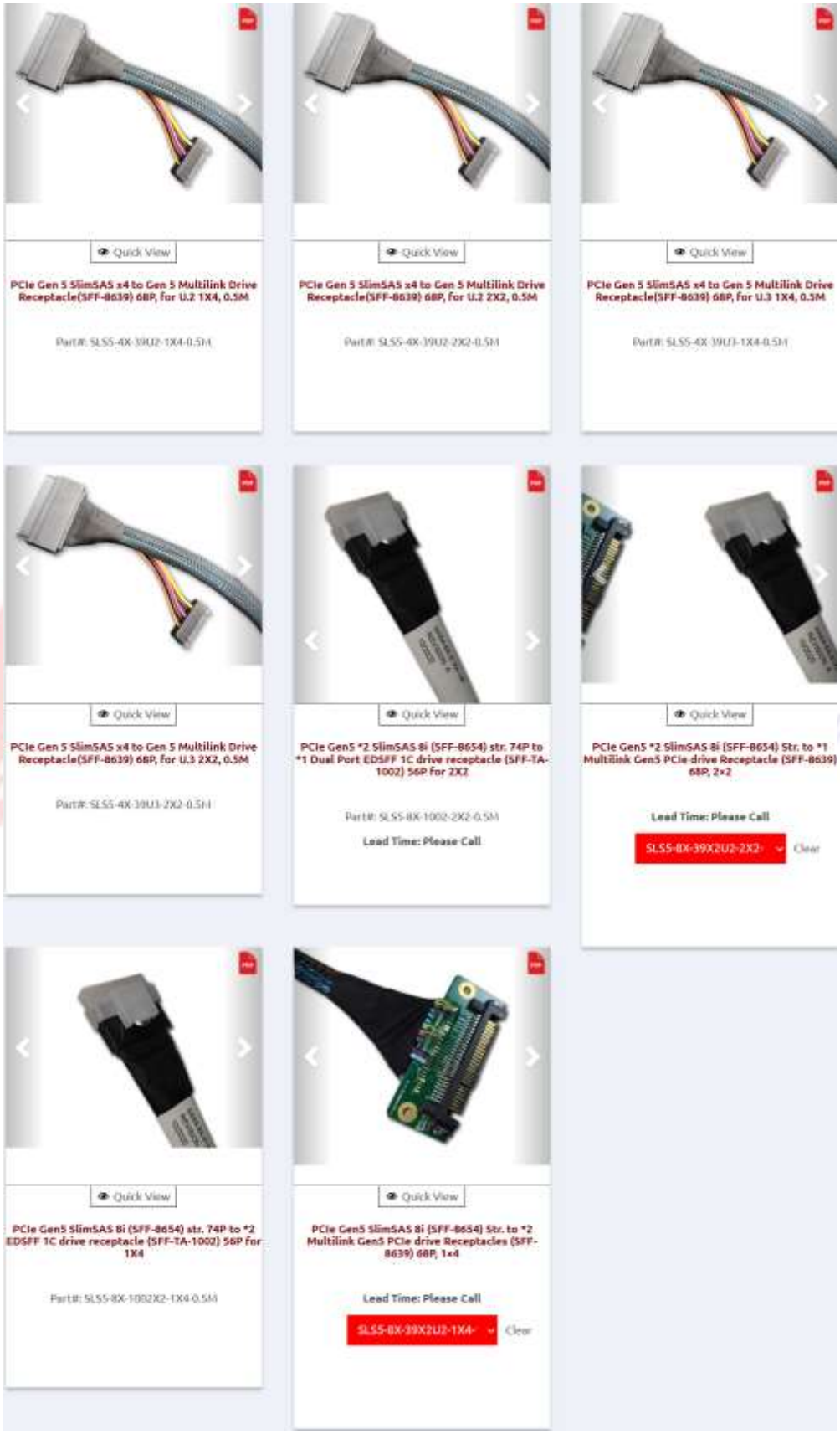


图 5-48

5.5.2 Gen4 转接线缆

5.5.2.1 Gen4 Oculink 线缆

Gen 4 Oculink 接口设备国内使用不多，但是国外在连接盘阵等设备的时候使用较多。

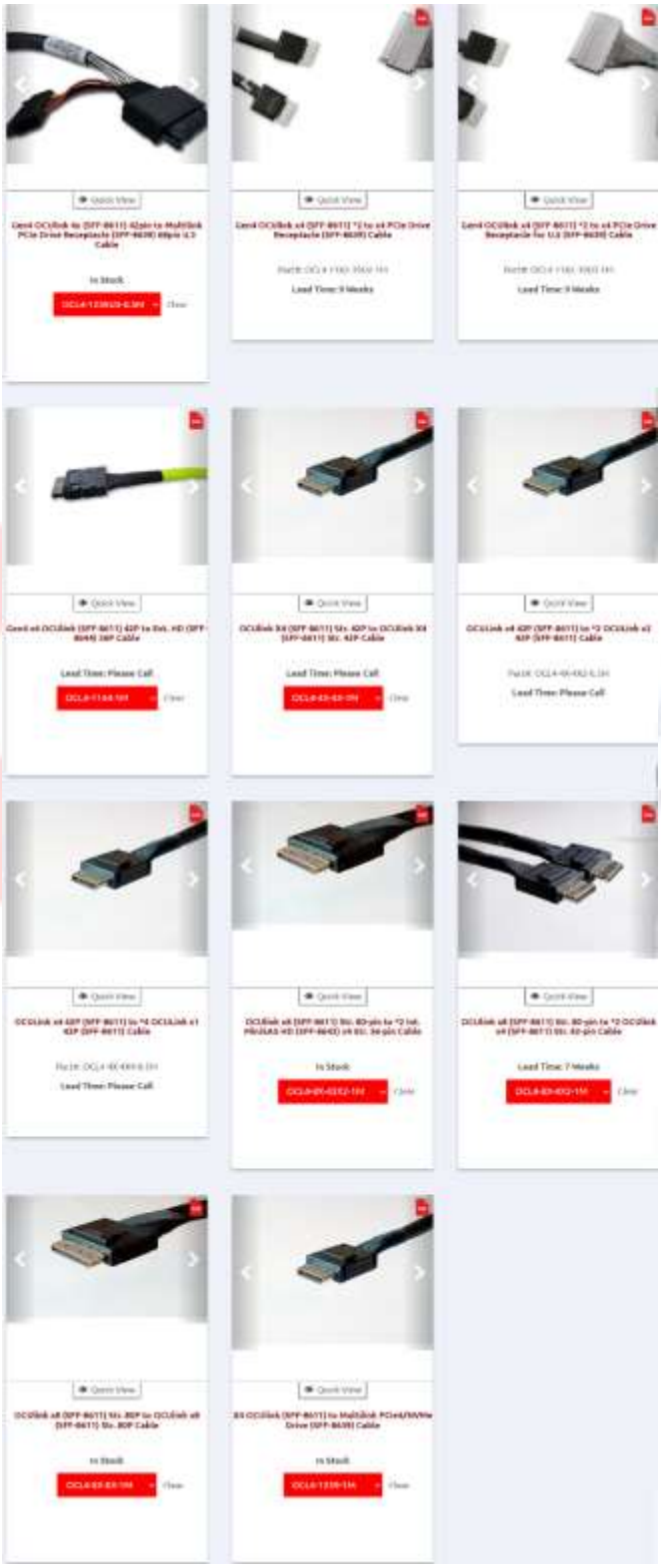


图 5-49

5.5.2.2 Gen4 SlimSAS 线缆

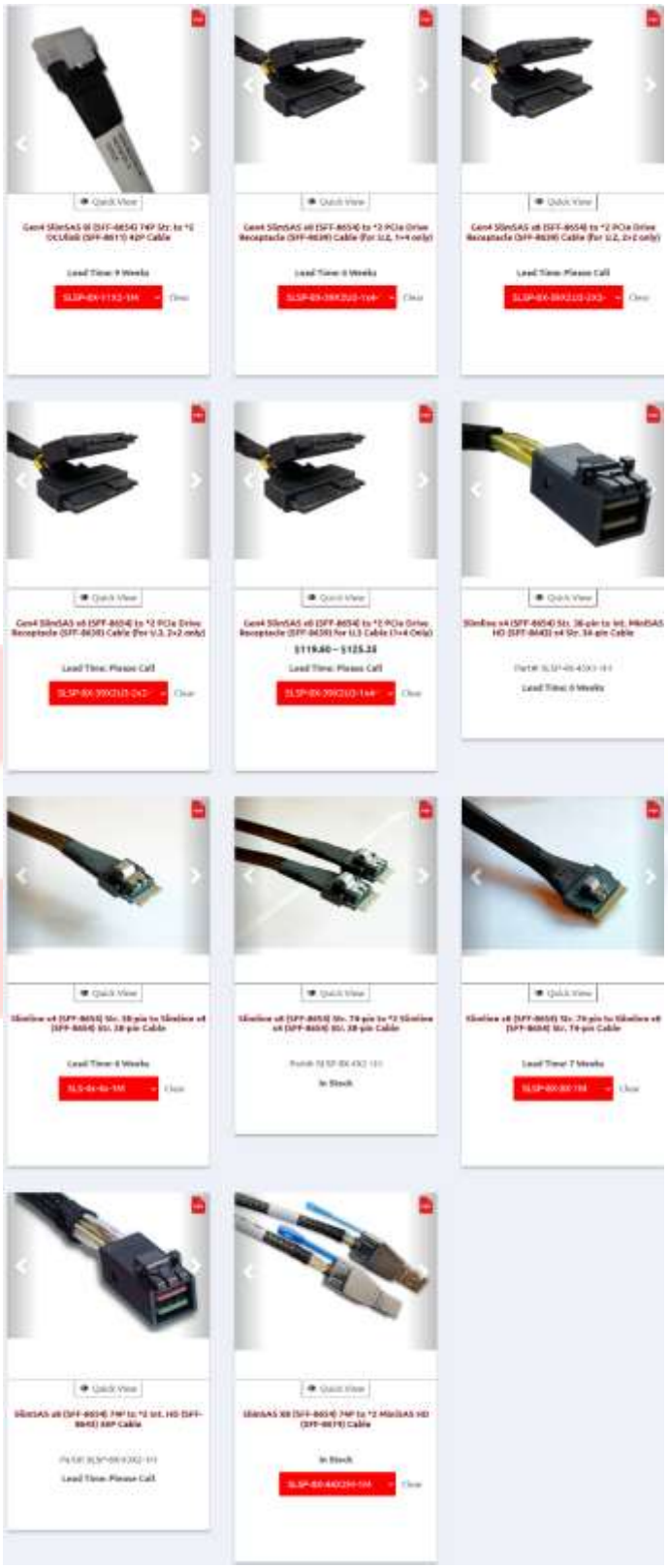


图 5-50

** Gen 4 SlimSAS /Slimline 除了连接盘柜之外，也提供可以直接转接成 Gen 4 single port （1x4）或者 dual port (2x2)的 U.2 接口的线缆，直接连接 NVMe SSD 盘。

5.5.2.3 Gen4 SFF-8644 线缆

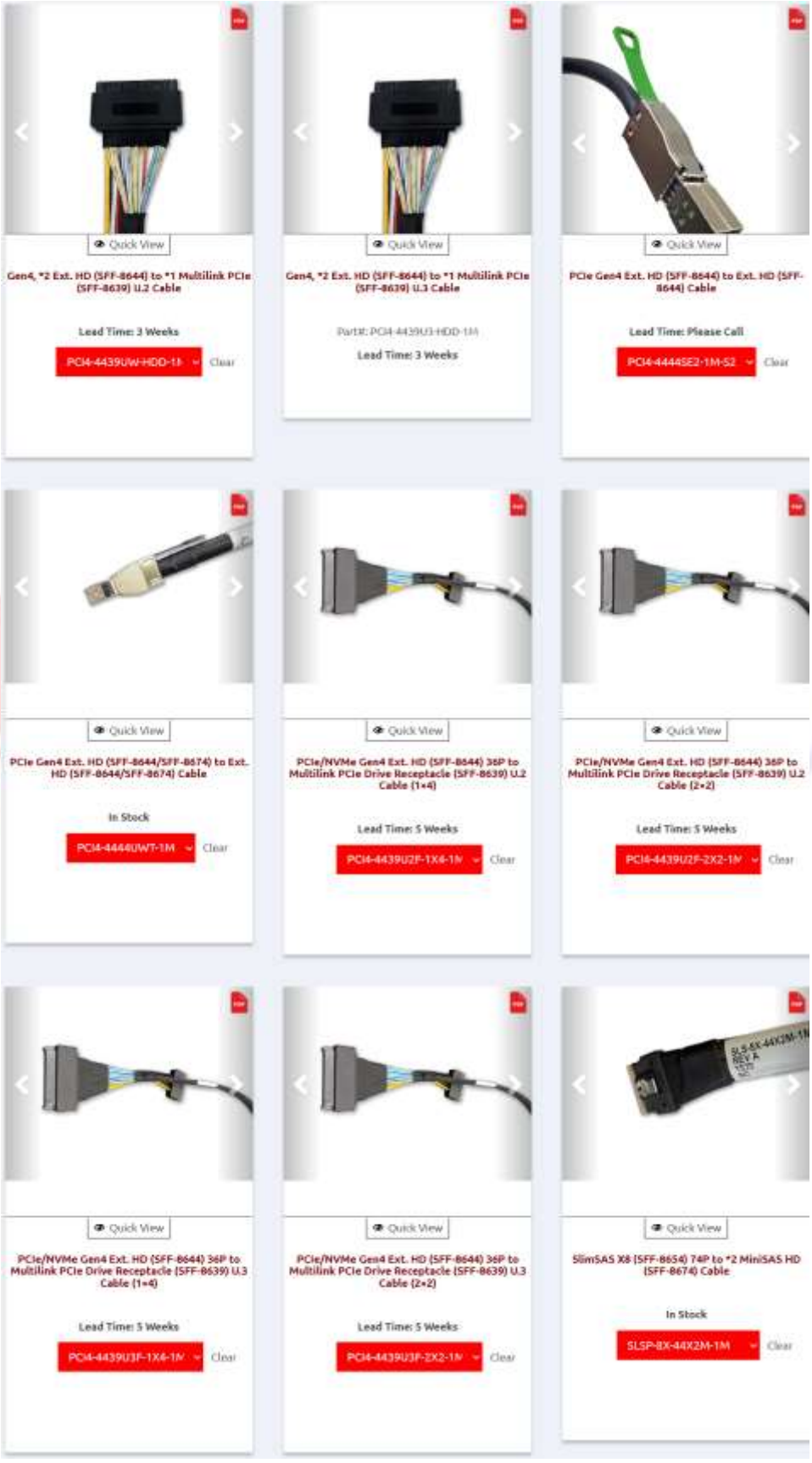


图 5-51

5.5.2.4 Gen4 EDSFF/GENZ 线缆

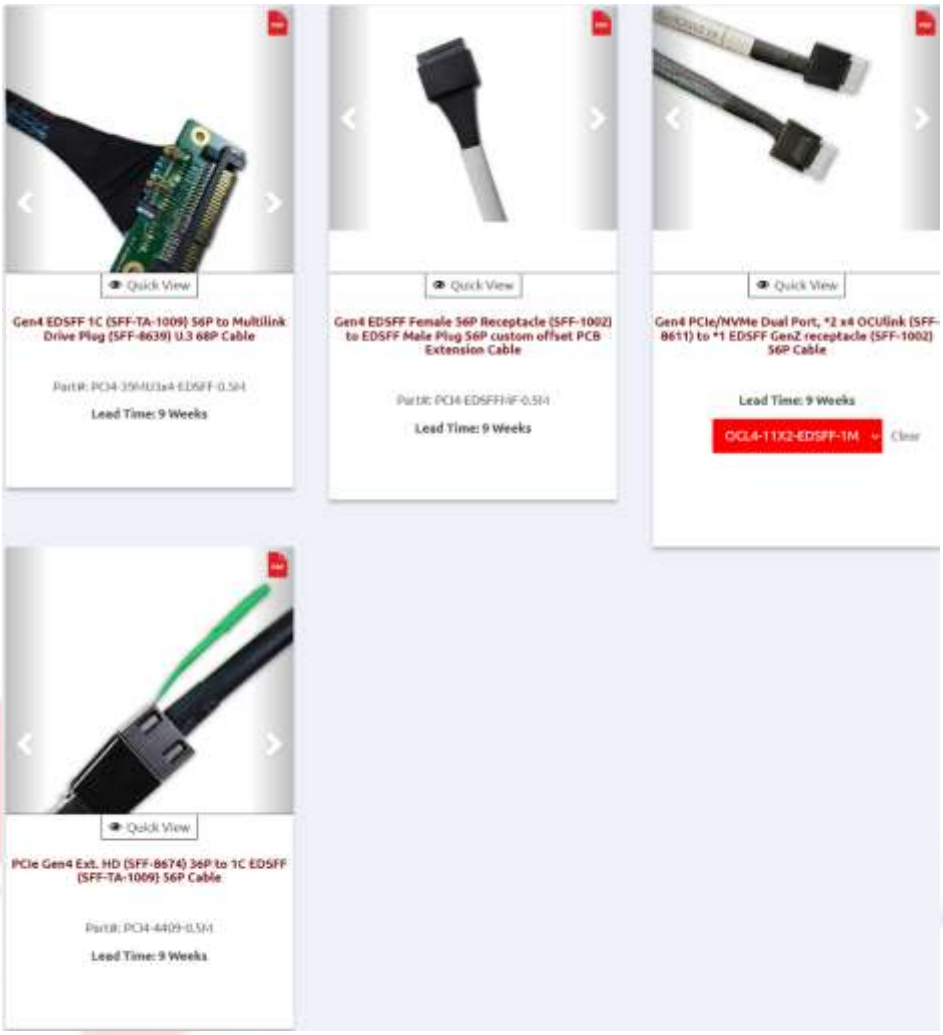


图 5-52

5.5.2.5 Gen4 其它线缆

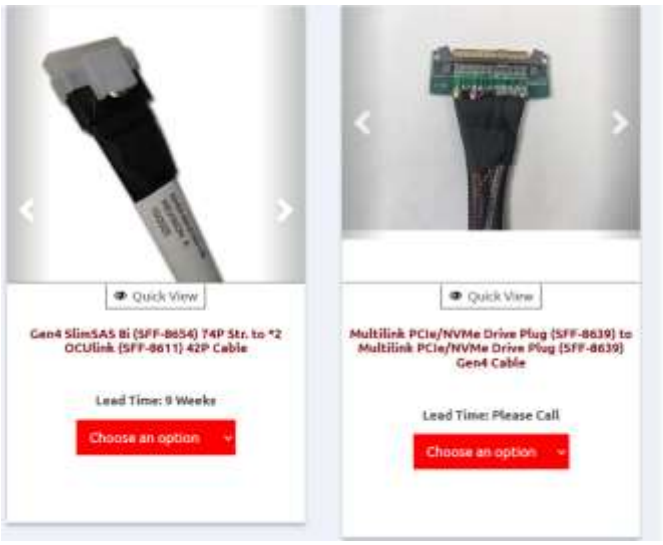


图 5-53

5.6 常用 PCIe Gen 4/5/6 延长线

各种 PCIe Gen 4 延长线，支持高低温测试（-25 ~ +85 摄氏度）的 PCIe Gen 4/5/6 x4 以及 PCIe Gen 4/5/6 x16 等。下面这两种线缆在各个主要客户处均做过实际测试，信号质量非常好，25cm 以及 50cm 延长线接入以后，NVMe SSD 测试的性能/延迟和吞吐量等信息和没有接入前保持一致，中间串接 PCIe Gen 4/5/6 analyzer 以后实际长时间追踪未发现有 bit error。

5.6.1 PCIe Gen 5 Slot 延长线

下面是 SerialCables 公司开发的 PCIe Gen5 x16 延长线，有 0.3m, 0.45m 两种规格，提供非常好的信号质量。

下面的 Gen5 x16 延长线是硅谷的厂商定制开发，信号也不错，但是价格较贵。

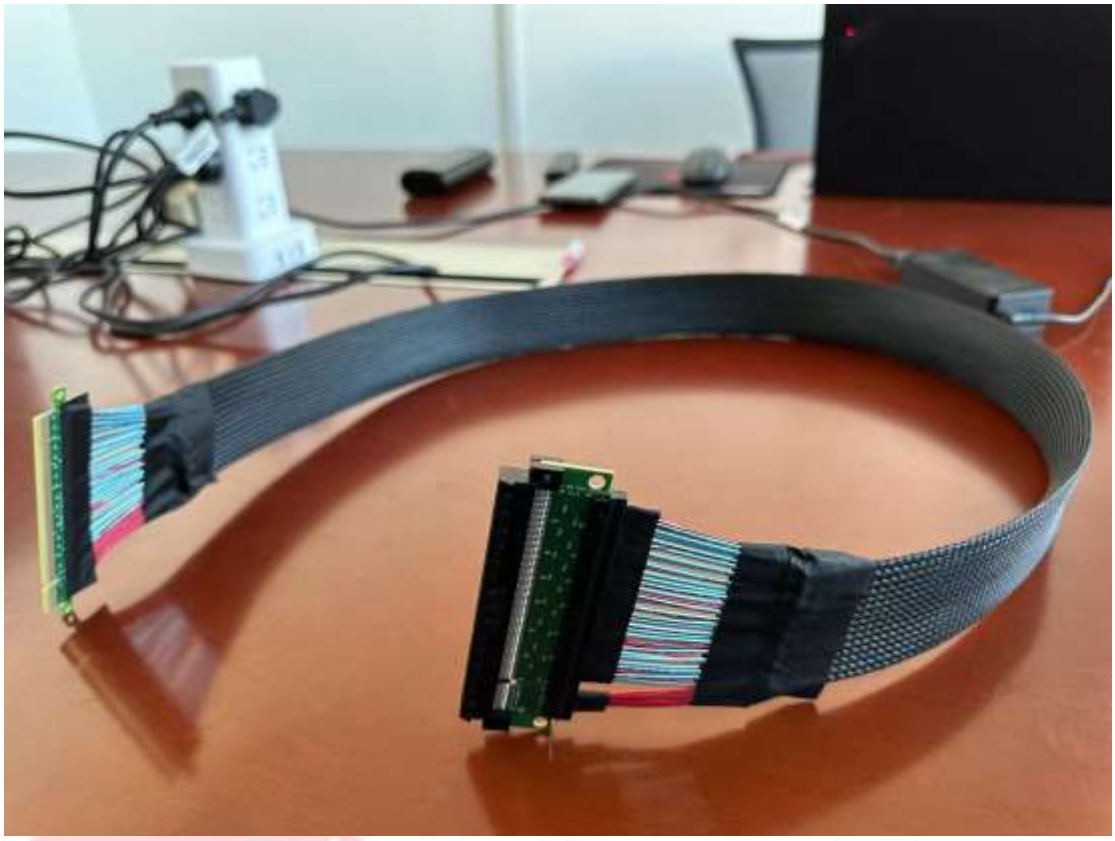


图 5-54

PCI-E X16 Gen 5 164P 转接电缆

- 小型 **PCB** 公连接器
- 卓越的信号完整性性能
- 阻抗：**85+/-10%** 欧姆
- **FEXT** 和 **NEXT** 功率总和：**40dB** 高达 **25GHz**
- 符合 **PCIe** 至 **CEM**
- 坚固的机械结构
- 弯曲支撑
- 提供灵活的版本

加工特点：

- 自动化焊接工艺
- 可用于生产的分立电缆和扁平电缆选项
- 电缆长度：提供多种长度，例如 **0.4M, 0.5 M, 0.7 M, 1M**
- 可定制长度（联系我们了解详情，需要最低起订量 **MoQ**）

Gen5 立管电缆典型性能

- **先进的串扰抑制技术支持 xtalk 的功率总和小于 40dB 到 25GHz**
- **阻抗控制在 85ohm+/-7ohm，反射<-10dB 最高 20GHz**
- **先进的 Twinax 电缆和 PCB 设计技术支持 4.5dB/m 的 IL**
- **在 16GHz 时最大损耗 6dB，以支持高达 1.0m/40 英寸的延伸范围 Max**

5.6.2 PCIe Gen4 Slot 延长线

下面的延长线都是一些进口线缆，我们测试在接入实际环境后，通过连接 PCIe Gen 4 analyzer 分析未发现任何错包。

图 5-55 PCIe Gen 4 金手指延长线和跳线（25CM）

图 5-56 20CM 和 50CM 两种延长线图

5.6.3 PCIe Gen 4 M.2 Socket 延长线

图 5-57 Gen 4 M.2 NVMe SSD M-key 延长线

如上图所示，有的时候测试 M.2 NVMe SSD，由于 M.2 slot 在台式机主板上面的位置不方便插拔 M.2，或者用户需要将 M.2 NVMe SSD 连接到笔记本主板上面，尤其是需要在 M.2 NVMe SSD 和主板 M.2 slot 中间串接 M.2 interposer（连接协议分析仪），或者 M.2 power fixture（用于电压拉偏或者功耗测量）时候，由于主板 M.2 slot 的 M-key 反向设计到这些 interposer 或者 fixture 无法接入的时候，可能需要使用上述的 M.2 NVMe SSD M-key 延长线。该延长线接入实际环境后，通过连接 PCIe Gen 4/5/6 analyzer 分析未发现任何错包。

5.6.4 PCIe Gen 4/5/6 U.2 Socket 延长线

图 5-58 Gen 4 U.2 NVMe SSD 延长线

上述延长线接入实际环境后，通过连接 PCIe Gen 4/5/6 analyzer 分析未发现任何错包。

5.7 常用 PCIe Dual Port NVMe SSD 测试环境搭建

在可以购买到昂贵的支持 PCIe Gen5 dual port NVMe SSD 的服务器，或者存储盘柜和盘阵之前如何进行针对企业级 dual port SSD 的测试是经常碰到的问题。目前主要有三种方案：

- **主机 + SerialCables PCIe switch 卡 + dual port 2x2 U.2 转接线缆**
- **主机 + SerialCables PCIe switch 卡 + SerialCables 8-bay 盘柜（通过管理接口的 CLI 设置 U.2 端口为 dual port）**
- **2*主机 + 2*HD-MINI-SAS/AIC adapter + 1* Y cable (2*HD-MINI-SAS to a U.2 female)，这种方式一般是需要测试 dual port 和 multi-host 交互的问题；如果 不考虑 multi-host 仿真的话，可以使用一台主机的两个不同的 PCIe slot，当然，一些 Intel Xeon server 支持一个 PCIe x16 slot 可以分叉成 8 个 x2，这样通过一个主机 + 一个 U.2/AIC adapter 即可达到目标。**

其它转接卡信息请直接联系我们获得更多些信息，邮箱：sales@saniffer.com，或者访问 https://www.saniffer.com/cn/downloads/ 下载下面的资料：

- **PCIe Gen 4/5/6 and NVMe SSD 测试工具速查手册**
- **SerialCables 存储测试环境连接产品速查手册**
- **SSD 测试工具白皮书和速查手册**

6. PCIe Gen4/5/6 NVMe SSD 测试环境搭建二: 主机和端口扩展

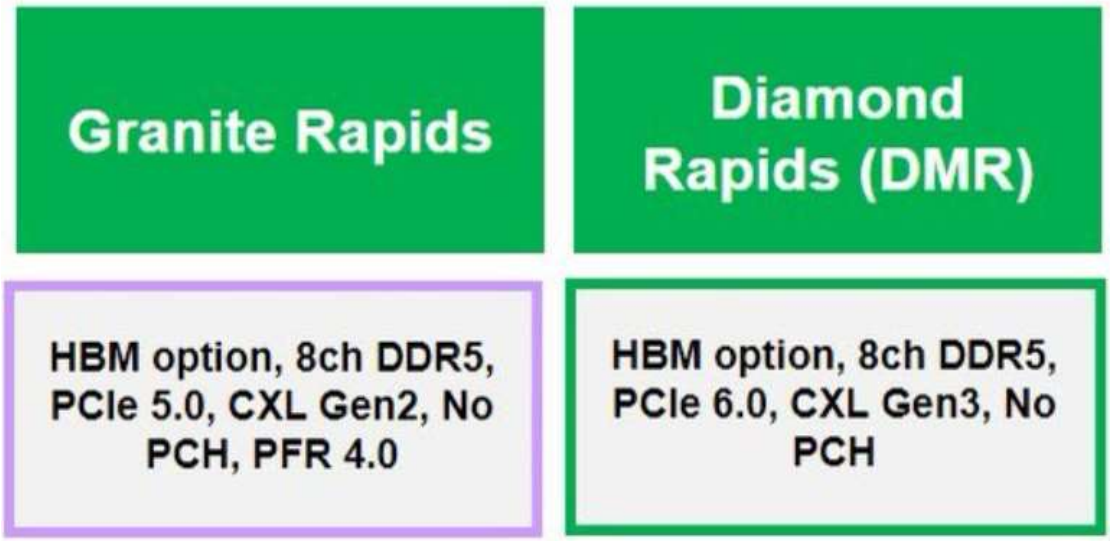
6.1 PCIe Gen6 CPU 和相关技术进展

6.1.1 Intel Xeon “Diamond Rapids” to support PCIe Gen6 and CXL Gen3

Published: Oct 13th 2022

Hardware leaker YuuKi_Ans reveals first details on future Xeon architecture

Recently, the same leaker revealed the first specifications of Intel Emerald Rapids, a successor to Sapphire Rapids and Granite Rapids, both product series seemingly pushed back due to never-ending delays of Intel HPC products. It looks like the newest update confirms that four generations from now, Intel will adopt PCI Express 6.0 technology.



Intel Granite and Diamond Rapids, Source: YuuKi_AnS

Intel has not confirmed when Diamond Rapids will be launching, in fact, the company never even confirmed the codename. However, the product was listed as 'Future Gen' featuring only big core (Performance).

The new details confirm that Diamond will support PCIe Gen6, which would be an upgrade over Granite Rapids with Gen5. Furthermore, 7th Gen Xeon Scalable architecture would also support CXL 3.0 (Compute Express Link), which is based on PCIe Gen6 implementation. The CXL 3.0 interface specifications were only released in August this year, but it will take another 3 years at least until we see it implemented in Intel Xeon products.

Unfortunately, there are no details on core count or CPU cache. The leaked slide may confirm, though, that Intel will continue using HBM memory in their HPC processors in the foreseeable future.

Intel Xeon Series Roadmap

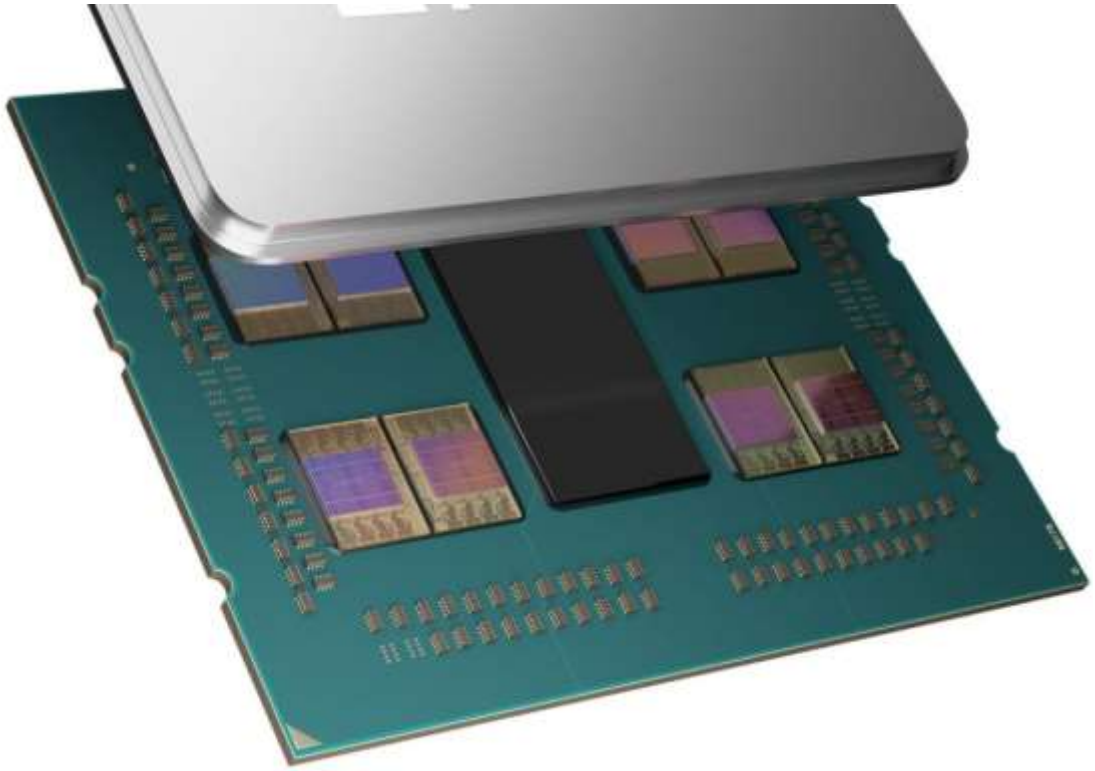
VideoCardz	Sapphire Rapids	Emeralds Rapids	Granite Rapids	Diamond Rapids
Series	4th Gen Xeon	5th Gen Xeon	6th Gen Xeon	**7th Gen Xeon**
Socket	Socket E	Socket E	TBC	TBC
Release Year	2022	2023	2024	**2025+**
Platform	Eagle Stream	Eagle Stream	Birch Stream	TBC
Core µArch	Golden Cove	Redwood Cove	Next Performance-Core	**Next Performance-Core**
Fabrication Node	Intel 7	Intel 7	Intel 3	TBC
Max Cores	56	~64	TBC	TBC
Max TDP	350W	~370W	TBC	TBC
Max L3 Cache	112MB	120MB	TBC	TBC
Memory Support	8x DDR5-4800	8x DDR5-5600	8x DDR5	**8x DDR5**
HBM Support	up to 64GB HBM2e	Yes	Yes	**Yes**
PCI Express	PCIe 5/4, 80 lanes	PCIe 5.0, 80 lanes	PCIe 5.0	**PCIe 6.0**
CXL Support	Gen1	Gen1	Gen2	**Gen3**

6.1.2 AMD Zen 6 document leak: More cores, PCIe 6.0 and 2.5D packaging

News, Processors | 5. 12. 2023 | Jan Olšan

Some time ago, youtuber Moore's Law Is Dead leaked the first information with AMD's Zen 6 architecture coming after the yet-to-be-released Zen 5. Now he has another juicy rumor regarding these CPUs, which could come to market in 2026 (late 2025 at best), as he has received documents showing a server and embedded version of them. These designs may however reveal quite a bit about desktop Ryzens with this architecture as well.

Moore's Law Is Dead showed documents and diagrams related to the next generation Embedded Epyc with Zen 6 architecture. Embedded Epyc is a derivative of AMD server processors, but it exists not only in a socket version (which is basically a server Epyc CPU), but also in a BGA package soldered to the board. However, it uses shared silicon – IO dies shared with regular server dies and the CPU core dies were previously shared with desktop CPUs as well.

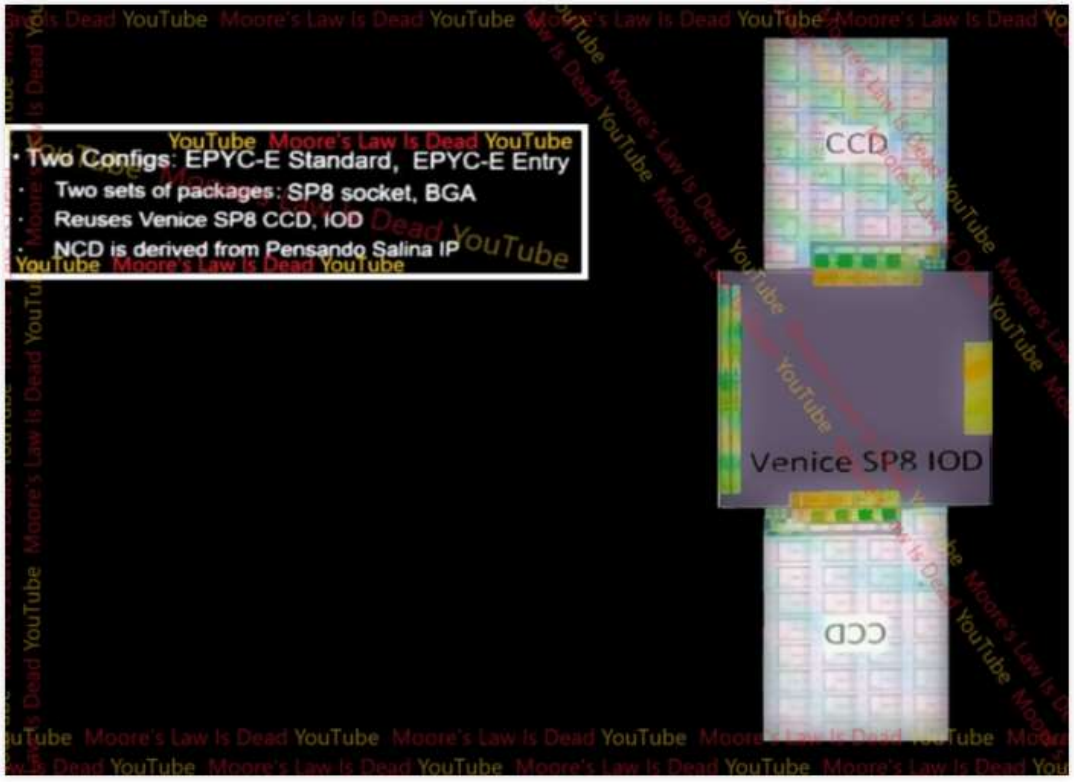


6.1.2.1 32-core chiplet (for servers), PCI Express 6.0

Perhaps the most interesting information is that Zen 6 should bring a new chiplet design and part of that will be an increase in the number of cores in a single CPU chiplet to up to 32 cores – in the server variant, at least. The embedded version of the Zen 6-based Epyc is to be codenamed Venice (like the successful 90nm Athlon 64 processors 20 years earlier), and according to AMD documents, it uses the same IO die as well as the same CPU die (CCD) as the standard server version, which will have the new SP8 socket (this is also a pretty important piece of information – using new socket means servers won't be upgradeable from Zen 4 or 5 to Zen 6).

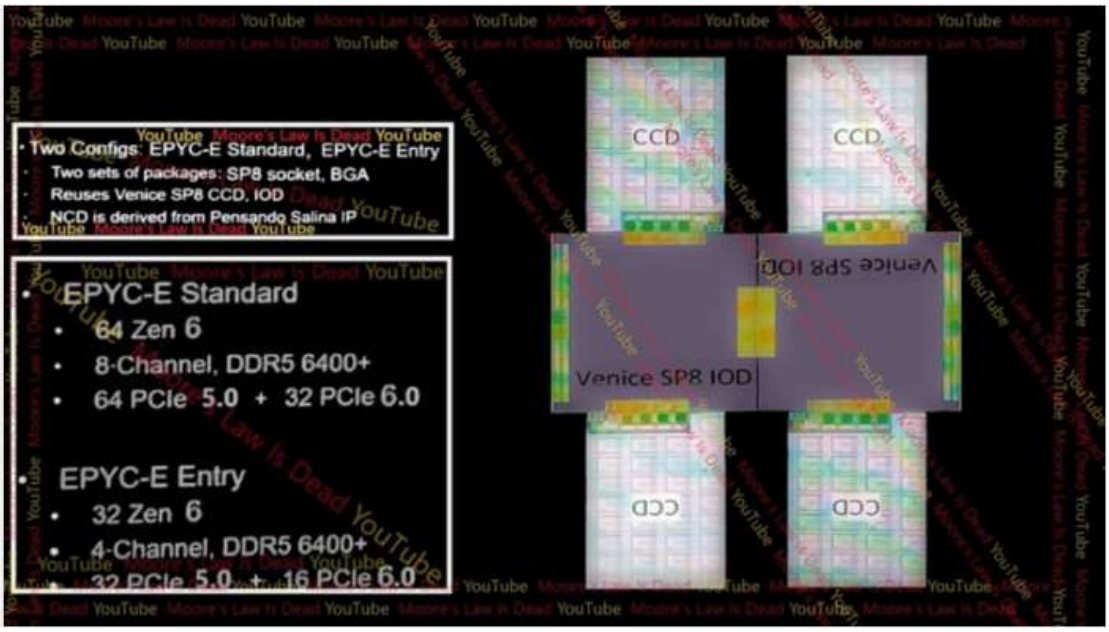
The embedded version will also not be compatible with the previous ones. It's supposed to have two versions. The less powerful version, the **Epyc-E Entry** in BGA package will contain one IO die and a maximum of two CCDs (chiplets with Zen 6 CPU cores). One CCD has 32 cores, so this version of the processor is expected to have a maximum of 64 cores in a BGA package. One IO die is designed to provide 32 PCIe 5.0 lanes and 16 **PCI Express 6.0** lanes. This platform will be the first from AMD to support PCIe 6.0 and it will be interesting to see if this carries over into the desktop version of processors with the Zen 6 architecture.

The memory controller provides four channels and supports DDR5-6400 memory (there is probably some chance of higher speeds, as it says “6400+”) – so the socket and platform will change, even though the time for DDR6 memory has not yet come.



AMD Epyc-E Entry source: Moore’s Law Is Dead)

The more powerful **Epyc-E Standard** version for a socket will use the same SP8 implementation as the standard servers. This version is supposed to have an eight-channel DDR5-6400+ controller and apparently two interconnected IO dies(physically consisting of the same silicon as in the cheaper version). Also, the connectivity will be doubled – 64 PCIe 5.0 lanes and 32 PCIe 6.0 lanes, so the connectivity of the two IO dies will be adding up.



Epyc-E Standard for the SP8 socket (source: Moore's Law Is Dead)

This eight-channel SP8 platform is apparently a lower-cost server platform that is the successor and analogue of the recently released SP6 platform with Epyc 8004 "Siena" processors. So alongside the SP8, there should probably be a more powerful SP7 platform, which will have Epyc 9006 processors, which could have a 12-channel or 16-channel memory controller, more PCIe lanes and support for 2S configurations. This in turn would be the successor to today's SP5/Epyc 9004 "Genoa" platform. It is quite possible that this SP7 platform will use the same IO dies, but in a higher number of three or four (and almost certainly the same CPU dies). If four IO diesand eight CCDs were used, the resulting processor could have up to 256 cores / 512 threads. However, the sharing of the BGA and SP8 version IO dies with the high-end SP7 platform is not yet confirmed.

6.1.2.2 Alternative accelerators instead of CPU cores

However, according to the piece of documentation shown by Moore's Law Is Dead, other sorts of chiplets can be connected to IO dies, not just CPU dies. Instead of one (or more?) CCDs, it should be possible to use alternative silicon. This could theoretically be an integrated FPGA (produced by Xilinx), but the document so far mentions another alternative: the integration of a "NCD" (Network Compute Die) chiplet with SmartNIC or DPU functionality.

AMD bought the Pensando company producing these products some time ago and this NCD chiplet is to use Pensando's "Salina" IP. A processor so equipped would have

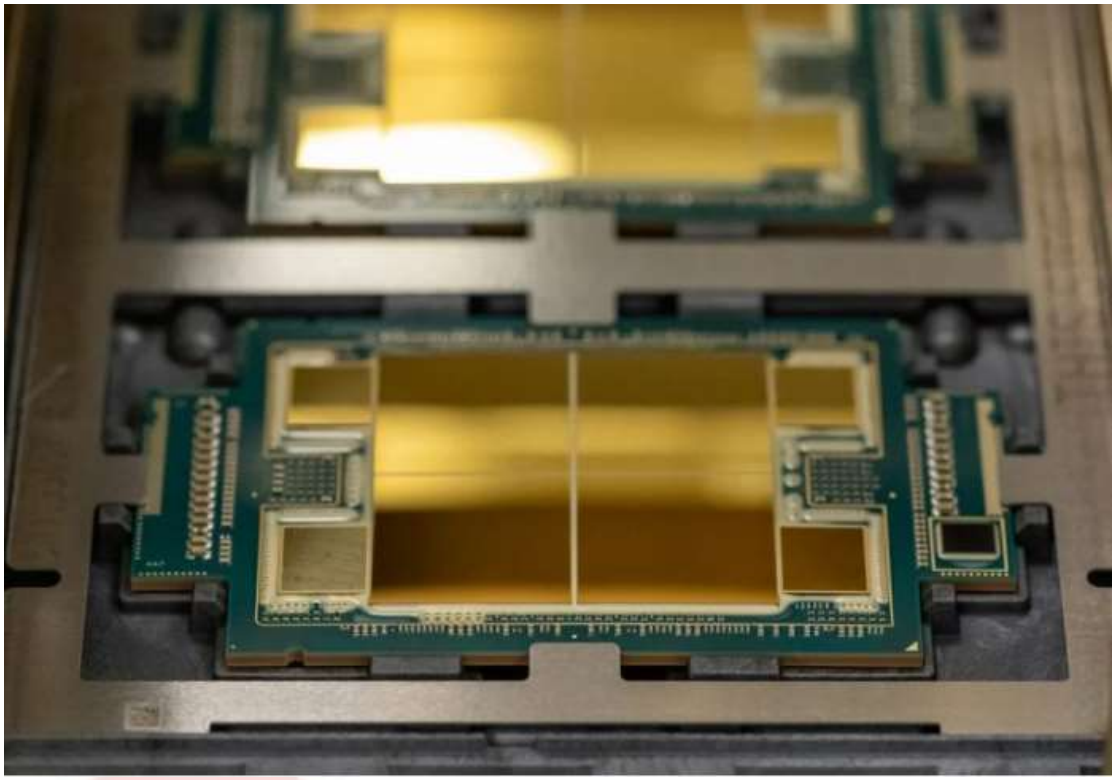
integrated acceleration for network tasks, network traffic analysis and data processing, and probably also Ethernet network interfaces. The BGA version of the processor would thus have a maximum of 32 cores in addition to the network part.

Possibility of using a NCD or hypothetically a FPGA die within the Venice design (source: Moore's Law Is Dead)

6.1.2.3 Advanced 2.5D packaging?

According to Moore's Law Is Dead schematics, the chiplets in this generation are finally connected by advanced silicon-bridge packaging, which could reduce the negative effects on performance and power draw that are inflicted by AMD's current chiplet solution that is based on a common type of packaging, where all chiplet communication goes through thick (and long) wires in a the underlying ordinary glass or organic substrate.

This advanced packaging should then be used to connect the two IO dies as well as to connect the IO dies and CPU dies. Hopefully this will make it to the desktop version as well, where it could again be a major innovation for power efficiency of processors, the biggest since the introduction of the MCM chiplet design in the Ryzen 3000 generation. This would also bring AMD in line with the chiplet/tile design of Sapphire Rapids, Meteor Lake a Arrow Lake from Intel.



Intel Sapphire Rapids processors currently have a more advanced 2.5D CPU chiplet interconnection, using Foveros silicon bridges. Photo shows Sapphire Rapids processor with HBM2E (source: CNET)

6.1.2.4 Zen 6 or Zen 6c?

The cores in the 32-core CPU die mentioned in these documents are likely the compact Zen 6c versions, similar to AMD's 16-core CCD with Zen 4c cores used today. Interestingly, though, the documents don't use the Zen 6c designation anywhere and only ever talk about Zen 6. Theoretically, this could be because this Embedded series will not use any cores other than compact cores and the specced clock speeds will simply reflect this – thus there will be no reason to make the distinction, even though technically it will be Zen 6c. It is also possible that the compact variant will simply be called Zen 6 and conversely the more powerful version, which is now considered the "vanilla" version, will be given a special name instead (say, Zen 6p as P-Core/"Phat"?).

6.1.2.5 Will desktop finally get more cores?

Anyway, it is likely that the 32-core dies are compact versions of the core reaching lower clock speeds and not the standard ones. Why? Because otherwise it would be a jump in the number of cores per chiplet from eight to 32, which doesn't sound likely. That's why we think Zen 6 could have this compact-core 32-core chiplet for servers

and embedded marktet, which would only represnet a doubling compared to the current 16-core Zen 4c-based chiplet.

In parallel, there should be a classic CCD with big Zen 6 cores (and offering the 3D V-Cache options), which will be used in desktop and some server processors. The fact that the compact version is 32-core probably opens up the possibility that the fat version will be 16-core, if the current 2:1 ratio is maintained. So finally, there could be Ryzen processors for mainstream desktop with 32 cores (and also probably 24 cores in a cut-down lower model). As far as we know, Zen 5 won't bring such an upgrade and will still max out at 16 cores, but in the Zen 6 generation there's finally a chance that the core and thread count will move higher (for the first time since 2019 and the Zen 2-based Ryzen 9 3950X). However, this is not yet guaranteed and we have no confirmation yet.

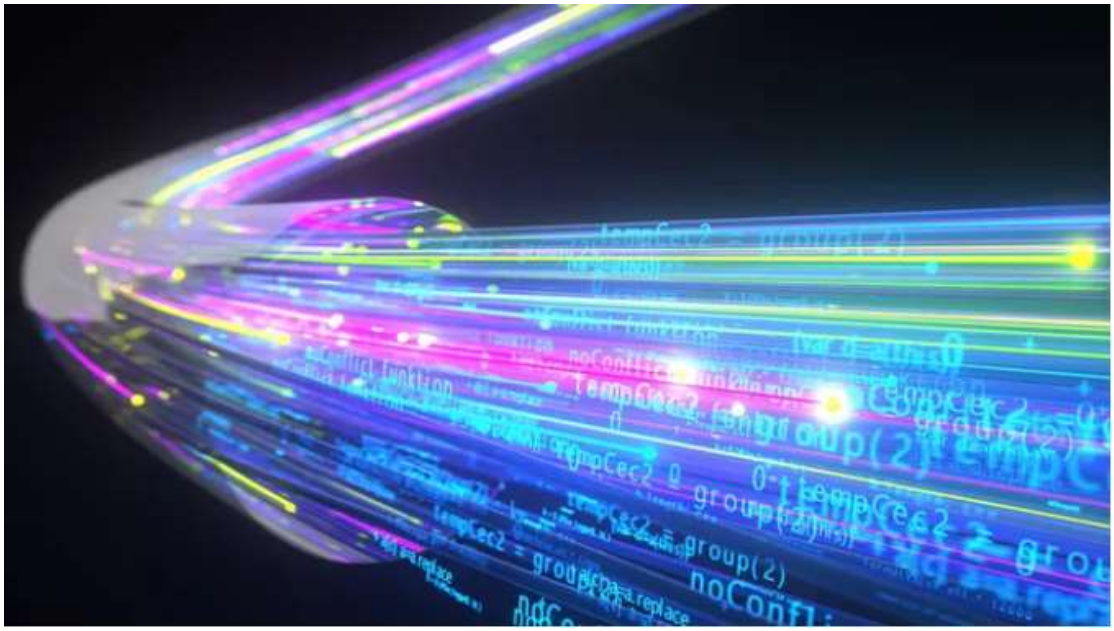
Desktop will probably use a different IO chiplet with two channels of DDR5 memory, but since the CPU chiplets will be designed for advanced 2.5D packaging, perhaps silicon bridges with better power efficiency and performance could be used here as well. The question is whether it will still be on the AM5 socket and whether this new more advanced chiplet technology concept will be able to be installed as an upgrade to today's boards. According to Moore's Law Is Dead, it's still possible, especially since DDR5 memory keeps being used – there's reportedly no indication yet that a new AM6 socket will be used. Again, though, it's not out of the question at all.

As usual, we must add the caveat that this is an unofficial leak and it is possible that some of the data and information in this source may be an error, misinformation, or some other form of noise. AMD may also still change its plans, so even if the leak is currently legitimate, the prediction may not come true. So for now, don't take this as a completely definitive thing and count with the uncertainty that is always inherent in this kind of leaked information.

6.1.3 PCIe 6.0 over optical cables demo in custom data center solution

published February 06, 2024

Nubis and Alphawave demo PCIe 6.0 x16 link without retimers.



Nubis Communications and Alphawave Semi have teamed up to showcase a PCIe 6.0 interconnection using an optical link. The demonstration used Alphawave's PCIe 6.0 controller and a Nubis linear optical engine, primarily meant to showcase ability of the companies to enable next-generation data center connectivity at 64 GT/s.

The demonstration features an Alphawave Semi PCIe subsystem based on the PiCORE Controller IP and PipeCORE PHY that drives and receives PCIe 6.0 traffic through a Nubis XT1600 linear optical engine. ***This setup achieves PCIe 6.0 x8 optical link at 64 GT/s per fiber, without retimers.*** The demonstration showcases the technical viability and high-speed capabilities of this custom PCIe over Optics solution.

"Our high level of integration with 16 lanes full-duplex in a single low-power, low-latency optical engine is a great match to the maximum bandwidth of PCIe x16 for next-generation compute and storage deployments," said Scott Schube, VP of Marketing at Nubis Communications. "Our demonstration of the Nubis XT1600 linear optical engine and Alphawave Semi's PCIe 6.0 Controller and PHY IP showcases the viability of a PCIe 6.0 x8 link over optical fiber at 64 GT/s."

Optical PCIe technology can significantly extend link distances without sacrificing bandwidth, compared to traditional copper cables. This capability is crucial for supporting larger AI/ML server clusters, distributed over multiple nodes, and paves the way for innovation in new disaggregated network architectures.

Sampling for the Nubis XT1600 linear optical engine has started and interested parties can contact the company. Meanwhile, it should be noted that the solution by

Nubis and Alphawave is a custom solution that has nothing to do with PCI-SIG's optical PCIe initiative.

"AI applications are reshaping data center networks, with hyperscalers deploying increasingly large clusters of disaggregated servers distributed over longer distances," said Tony Chan Carusone, CTO at Alphawave Semi. "This shift has generated heightened interest in PCIe over Optics among several of our customers. Through our collaboration with Nubis, we're pleased to demonstrate how we're leveraging Alphawave Semi's leadership in connectivity IP and silicon to enable PCIe optical connectivity solutions that accelerate high-performance AI computing and data infrastructure."

6.2 PCIe Gen5 测试主机和 Gen5 SSD 选择

目前 PCIe Gen 4 测试主机主要采用 AMD Gen 4 芯片组的主板搭建，例如 X570 芯片组，市场上面可以看到技嘉，华硕，华擎，超微，微星等主板；也可以采用 Intel Z590 芯片组的主板，不过由于 Intel 推出 Gen4 方案较晚，实测结果感觉没有早 2 年发布的 AMD X570 芯片组的主板来的稳定。

目前，针对 PCIe Gen 5 测试主机也是技嘉，华硕，华擎，微星等工作站级别的消费类主板，采用 Intel Gen5 桌面级 CPU 以及 Z690 芯片组，以及更加经济、低端配置的 H670, B660, H610 也在 Q1 2022 发布出货。另外，Intel NUC 12 extreme 在 2022/3 月份也开始发售，价格在 US$3,000 以上。

注意：基于 Intel Gen5 CPU 的 server 已经在 2023 年 Q1 已经开始发货；基于 AMD Gen5 CPU Genoa 的服务器主板或者服务器在 2023 年 2 月份开始发货（高 CPU 核主机）。

Intel 600 Series Desktop Chipsets

	Z690	H670	B660	H610
Launch Date	Q4 2021	Q1 2022	Q1 2022	Q1 2022
CPU Overclocking	Yes	No	No	No
PCIe 5.0 Slots via Processor	One x16 or Two x8	One x16 or Two x8	One x16	One x16
PCIe 4.0 Slots via Processor	One x4	One x4	One x4	None
PCIe 4.0 Lanes via Chipset	Up to 12	Up to 12	Up to 6	None
PCIe 3.0 Lanes via Chipset	Up to 16	Up to 12	Up to 8	8
Memory Overclocking	Yes	Yes	Yes	Yes
Memory Channels	2	2	2	1
Integrated Wireless	Wi-Fi 6E AX211 (Gig+)	Wi-Fi 6E AX211 (Gig+)	Wi-Fi 6E AX211 (Gig+)	Wi-Fi 6E AX211 (Gig+)
DMI 4.0 Lanes	8	8	4	4
USB 3.2 Gen 2x2 (20 Gbps)	Up to 4	Up to 2	Up to 2	None
USB 3.2 Gen 2x1 (10 Gbps)	Up to 10	Up to 4	Up to 4	Up to 2
USB 3.2 Gen 1x1 (5 Gbps)	Up to 10	Up to 8	Up to 6	Up to 4
USB 2.0 Ports	14	14	12	10
Wi-Fi 6E	Yes	Yes	Yes	Yes
SATA 3.0 Ports	Up to 8	Up to 8	Up to 4	4
PCIe RAID	0, 1, 5, 10	0, 1, 5, 10	None	None
SATA RAID	0, 1, 5, 10	0, 1, 5, 10	0, 1, 5, 10	None

图 6-1

另外，基于 AMD PCIe Gen5 CPU 的工作站主板也在 2022/9 月底已开始发货，基于其 PCIe Gen5 Server CPU 的服务器在 H1/2023 发货，低核数服务器 CPU 较晚。

测试环境搭建的时候需要根据需求具体定制，包括 PCIe Gen4 还是 Gen5, PCIe 插槽数量，扩展性，待测试的 NVMe SSD 盘接口和数量（主要和 CPU 核数以及 DDR4 内存大小配置有关，尤其是通过 PCIe Gen 4/5/6 Host Card 外接盘柜的测试场景）。常见的消费类主板一般含有 1~2 个 Gen 4 M.2 socket，以及 1~3 个 PCIe Gen 4/5/6 插槽，但是往往这些插槽都是共享例如 x16 lane 的，这个要特别注意，例如如果通过外置 PCIe Gen 4/5/6 Host 插卡外联测试盘柜，一般该 Host 插卡需要 Gen 4 x16，那么就不要在相邻的插槽再插入显卡（AMD CPU 不含显示控制器，必须配置外置显卡），否则可能会导致显卡占用了 x8 从而影响到 Host 插卡只能工作在 x8 状态。

注意：Gen 4 主板的选择不能仅考虑上述这些因素，因为目前市场上的 AMD Gen 4 主机选择存在一些坑，很多人只是考虑性价比，结果导致选择的主板存在很多兼容性问题。这是因为虽然各个主板厂商都使用 AMD CPU 或者同样的芯片组，但是各家主板的 BIOS 更新的频度，对于 PCIe 外设的兼容性，是否热插拔支持等很多方面都可能存在这样那样的问题。

由于我们专注于 PCIe Gen 4/5/6 NVMe SSD 的相关测试工具，我们在上海的实验室为了搭建测试环境已经测试了多家厂商的 PCIe Gen 4 和 Gen5 主板，同时结合我们美国合作伙伴提供的美国各大 SSD 公司选择 Gen 4 和 Gen5 主板遇到的坑，这些经验可以有效地帮助用户避免选择陷阱。我们已经通过 SerialTek PCIe Gen5 x16 analyzer 以及 Serial Cables Gen5 switch card 测试了 Intel Gen5 PDK 原型机和国内多家厂商基于 Intel Gen5 Server CPU 设计的服务器。

下面是我们摘选的针对目前市场上常见的 Gen 5 CPU 的的简单介绍，供大家参考。

6.2.1 Intel 架构平台

6.2.1.1 Intel Gen5 Xeon CPU 服务器

我们从 2022 年 1 月份测试的一款基于 Intel Gen5 Xeon 服务器 CPU 主板，该主板支持 3* Gen5 x16 slot, 2* Gen5 x8 slot, 1* Gen5 x4 slot, 1* Gen5 x1 slot，所有插槽都直连 CPU，并且支持 1x16, 2x8, 4x4, 8x2 等多种 bifurcation 设置，非常灵活，参见下图。

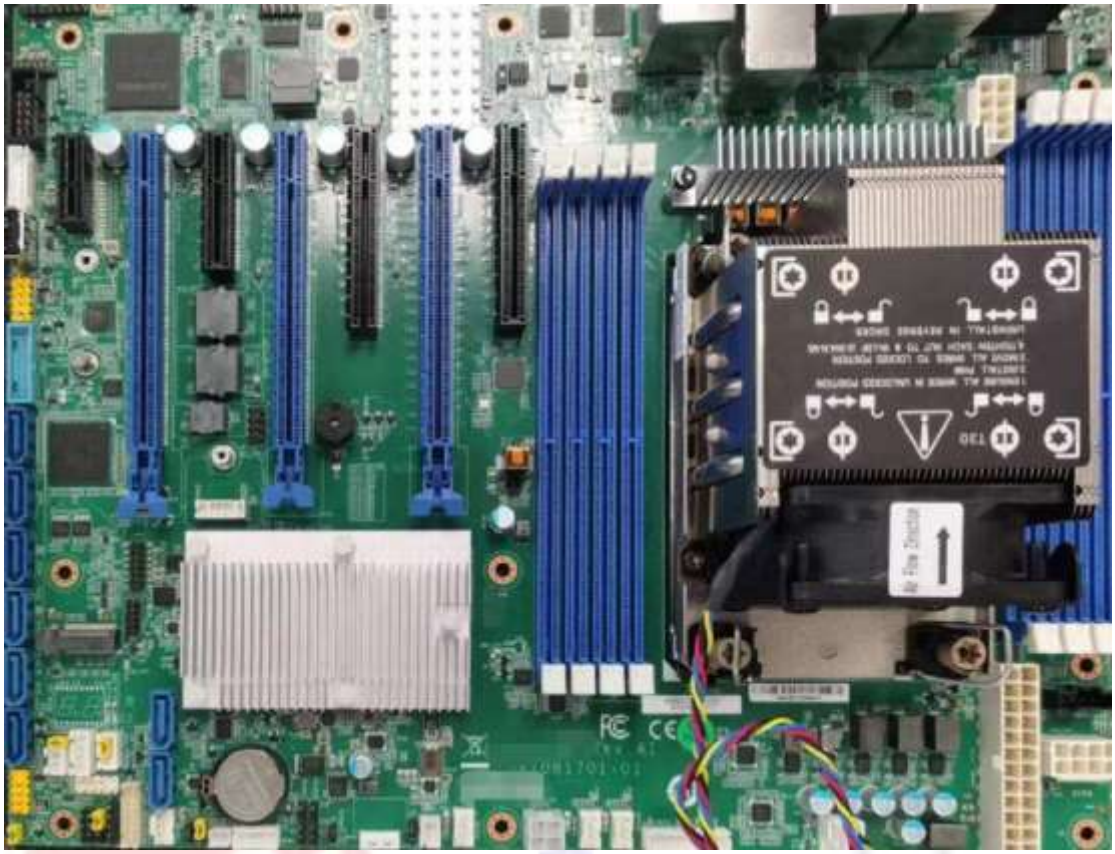


图 6-2

注意：如果需要支持 CXL，需要购买 Intel Saphhire Rapids 和 Emerald Rapids。

6.2.1.2 Intel Gen5 Core CPU 工作站

2022 年基于 Intel Z690/790 芯片组 PCIe Gen5 主板

Intel has launched their 12th Generation Core CPUs codenamed Alder Lake, but choosing the **best Z690 motherboards** for your shiny new CPU might not be a straightforward task. The three new CPU SKUs that have been unveiled as of the time of writing include the flagship **Intel Core i9-12900K and 12900KF**, along with the **Core i7-12700K and 12700KF.** The Core i5-12600K and 12600KF make up the midrange offerings from Intel on the new platform. The socket has been changed to LGA 1700 from last gen's LGA 1200, so that removes any backward or forwards compatibility for the **Z690 motherboards** and 11th Gen Core processors.

You might be wondering what makes the new platform so exciting. Well, after constant pressure from AMD and their brilliant Ryzen CPUs, Intel has come back swinging with a host of brilliant features in the Alder Lake CPUs and the **Z690 chipset**. For the first time, we have DDR5 RAM memory support on a consumer desktop platform, along with the standard DDR4 support. The new Intel platform also introduces a significant performance improvement, and also offers much better functionality. Unfortunately, it will be paid at

much higher prices. People planning to modernize a computer will feel the most – they will have to buy a processor and a new motherboard for the LGA 1700 and new DDR5 RAMs, though some Z690 moth. However, also support DDR4 ram. Well, I guess we have to get used to the fact that the new generation of equipment brings improved performance and higher prices.

You will have to choose which memory platform you want since you can't use both simultaneously. Furthermore, Intel has also included PCIe Gen 5 support on **Z690 motherboards**, although there is currently no device out there that can take advantage of this technology right now. Still, Intel has provided a pretty solid upgrade here with the Z690 and Alder Lake platform. You can also check out different types of Motherboards For I9-12900k.

Table of Contents

- **DDR4 vs. DDR5**
  - Frequently Asked Questions

6.2.1.2.1 Best Z690 Motherboards

If you have decided to get onboard the Alder Lake train, you will need a Z690 motherboard that can perfectly suit your needs. This is why you will be needing our reviews today, primarily because picking the best Z690 motherboard out of the dozens of options that are available on the market is not an easy task.

The high-end CPUs in the Alder Lake lineup, especially the Core i9 parts, need a lot of stable power so the VRM performance is absolutely key here. We should also take into consideration whether the board is DDR4 or DDR5 since there are different variants of each motherboard depending on the memory type. We have already done a DDR5 and DDR5 Rams for Intel 12th Generation – Alder Lake, so must check that out.

With that out of the way, as per our reviews – here are all the recommended Z690 motherboards for your next upgrade.

6.2.1.2.1.1 ASUS ROG Maximus Z690 Hero

Best Overall Z690 Motherboard

图 6-3

ASUS ROG Maximus Z690 Hero

Specifications="**Chipset**: Z690 | **Memory**: 4x DIMM, 128GB, DDR5-6400 | **Video Outputs**: HDMI | WiFi | **USB Ports**: 11x rear IO, 9x internal | **Network**: 1x 2.5 GbE LAN, 1x Wi-Fi 6E | **Storage:** 5x M.2, 6x SATA"]

Pros

- **Top Of The Line Z690 Motherboard**
- **Excellent Connectivity Options**
- **2x Thunderbolt 4 Ports**
- **20Gbps USB-C port**
- **Nice Aesthetics**
- **Excellent Power Delivery**

Cons

- **Very Expensive Z690 Motherboard**

Any enthusiast-grade motherboard tier list would be incomplete without mentioning the ROG Maximus series from ASUS. This is the high-end motherboard category from ASUS that falls under the ROG banner, and for the Z690 versions, ASUS has done away with confusing Roman numerals. Instead, they opted for the name of the chipset in the name of the product, what an ingenious thought. Still, the ROG Maximus Z690 lineup of boards is among the best you will ever find on this chipset, and the Hero variant is the one we are particularly interested in.

Packing a monstrous 20+1 VRM design with 90A power stages, the Maximus Z690 Hero is more than enough for any Alder Lake CPU you can throw at it, even at overclocked settings. In fact, we would recommend the Maximus Z690 Hero for any potential buyers of the Core i9 12900K and the 12900KF, since those CPUs have lots of overclocking potential that can be achieved using this fantastic z690 motherboard. The VRM is also cooled effectively by large heatsinks, so temperatures should be perfectly sound. The board also supports DDR5 memory at up to 6400MHz speeds, but time will tell if investing in DDR5 right now is a good idea.

Aesthetically, the board looks really impressive as you would expect from a Maximus series board. ASUS has gone for a sort of modernized look with the pixelated ROG script on the I/O cover and the ROG eye on the chipset heatsink which is also pixelated. These surfaces are finished in a glossy material so fingerprints would be quite visible. Other than that, there is not a lot to comment on in terms of aesthetics since the board is mostly matte black and covered with heat spreaders, so there is barely any PCB visible. The I/O shield is built-in, which is an excellent feature that is being standardized now.

In terms of features, there is not a lot to complain about with the Maximus Z690 Hero. The board brings absolutely everything you could want from a premium board on the bleeding edge. The 2x Thunderbolt 4 ports on the rear I/O stand out immediately, and they are an excellent addition to the I/O of this board. The board also supports 5 M.2 devices at the same time, if you install the included PCIe M.2 expansion card as well. PCIe Gen 5 is also supported, of course, but that is limited to the PCIe 16x slots which is not really important right now. The board also has a 2.5 GbE LAN port and WiFi 6E for connectivity.

All in all, you can’t really go wrong with the Maximus Z690 Hero for the brand new Alder Lake platform and that’s why it is our pick for the **best overall Z690 motherboard**. It has one of the best power delivery systems of any motherboard available on the market and combines that with a premium feature-set that doesn’t really lack anything of note. The price is hefty, understandably, but it might be worth it in the long run if you plan to keep a premium Alder Lake system with you for years down the line.

6.2.1.2.1.2 Gigabyte Z690 AORUS Master

Best High-end Z690 Motherboard

图 6-4

Gigabyte Z690 AORUS Master

Specifications=”**Chipset**: Z690 | **Memory**: 4x DIMM, 128GB, DDR5-6400 | **Video Outputs**: DisplayPort | WiFi | **USB Ports**: 11x rear IO, 9x internal | **Network**: 1x 10 GbE LAN, 1x Wi-Fi 6E | **Storage:** 4x M.2, 6x SATA”]

Pros

- **Z690 Motherboard With Unrivalled VRM Design**
- **10 GbE LAN**
- **Several Storage Options**
- **20Gbps USB-C port**
- **Attractive Design**
- **High-End Performance**

Cons

- **Quite Pricey**

Gigabyte's high-end AORUS motherboards recently have been absolutely fantastic for both platforms, so it is no surprise that the Gigabyte Z690 AORUS Master ends up as our pick for the **best high-end Z690 motherboard**. Not only does it have an excellent power delivery system which is almost a necessity for the new Alder Lake CPUs, but it also has an excellent feature set that is comparable with some of the most expensive Z690 motherboards out there. The AORUS Master itself isn't cheap by any stretch of the imagination, but its pricing is not outrageous in view of what it brings to the table.

Starting off with the highlight of this board, the extremely overbuilt VRM. It consists of a staggering 22 phases, out of which 19 supply the CPU with 105A of power. This is an insane VRM setup that will almost never be completely utilized, even by a fully overclocked 12900K running 24/7. This VRM design is truly ahead of its time due to the fact that any Alder Lake CPU and even possibly the next upcoming Intel 13th Gen flagship is probably not enough to saturate the VRM of this motherboard. The VRM heatsinks are also quite sizeable so the board should have no problems dissipating the heat that is coming out of those VRMs.

As far as looks are concerned, the AORUS Master is certainly one of the most over-the-top boards out there. It doesn't have any fancy gold accents or anything, but it does have massive heatsinks and heat shields that cover up around 95% of the PCB itself. The I/O cover has a large RGB AORUS logo and some lighting accents which look cool, and a similar style continues down to the chipset heatsink which has a "glitched" AORUS logo. This modern design language seems to be rampant in the Z690 motherboards. Speaking of the heatsink, the chipset heatsink sort of extends over the entire bottom half of the PCB and serves as the heatsink for the M.2 drives as well. This is a very ingenious implementation.

When it comes to features, we really have no reason to doubt the AORUS Master Z690 motherboard. It is absolutely jam-packed with top-of-the-line features such as the DDR5 memory support for up to 6400MHz DIMMs. There are 11 USB ports on the I/O including two USB-C ones. The main feature, however, is the 10 GbE LAN port on the back from Aquantia. This, paired with the two WiFi 6E antennas adjacent to it, means that you will be absolutely spoiled by choice when it comes to connectivity. PCIe Gen 5 is also supported, and the board has four M.2 drives as well for your storage needs. Clearly, AORUS understands what kind of user is looking to buy this Z690 motherboard at this price point. Speaking of M.2 drives, you might also be interested in our review of the XPG Gammix S50 Lite PCIe Gen 4 SSD.

Conclusively, the AORUS Master Z690 is one of the best premium motherboards for the

Z690 chipset out there right now. It has more features than you can shake a stick at, and it pairs them well with a top-of-the-line VRM setup that is clearly more than you could ever need. All of this does come at a price though, but just like the Maximus Z690 from ASUS, it could prove to be worth it in the long run.

6.2.1.2.1.3 MSI MPG Z690 Carbon WiFi

Best Overclocking Z690 Motherboard

图 6-5

MSI MPG Z690 Carbon WiFi

Specifications="**Chipset**: Z690 | **Memory**: 4x DIMM, 128GB, DDR5-6666 | **Video Outputs**: HDMI and DisplayPort | WiFi | **USB Ports**: 9x rear IO, 7x internal | **Network**: 1x 2.5 GbE LAN, 1x Wi-Fi 6E | **Storage:** 5x M.2, 6x SATA"]

Pros

- **Great VRM For Overclocking on Z690**
- **5 M.2 Slots For Storage**
- **Attractive Aesthetics**
- **Attractive Design**
- **High-End Performance**

Cons

- **Not Affordable For Regular Gamers**

MSI has made some excellent boards over the past few years, especially for the AMD chipsets, but they have not exactly ignored Intel either. MSI's MPG series, which stands for MSI Performance Gaming, packs some of the most overbuilt and premium motherboards out there for any chipset, be it Intel or AMD. Continuing this trend for the new Alder Lake platform, we have the MSI MPG Z690 Carbon WiFi, a fantastic high-end motherboard on the Z690 platform that is quite possibly the **best overclocking Z690 motherboard** on our list. It certainly rivals our two top picks in terms of VRM design and features.

Speaking of the power delivery, MSI has hit the nail on the head with an 18+1+1 phase VRM design with 75A power stages which is excellent for overclocking pretty much any modern CPU you can put in it. Particularly, the Core i9 12900K would be a great match for this Z690 motherboard since you can completely max out the overclock on the flagship Alder Lake CPU without taxing the VRMs too much. In terms of heat dissipation, MSI has provided massive VRM heatsinks that are actually finned in order to increase the surface area of the metal, aiding in dissipation. The board supports DDR5 memory up to a whopping 6666 MHz, which sounds really fast by today's standards but time will tell if speeds like this are normalized.

The MSI MPG Z690 Carbon WiFi motherboard is also one of the finest-looking motherboards you can find on this platform. It is a fairly expensive board, sure, but the looks are a huge selling point especially in today's market, and MSI has nailed that aspect as well. The board is covered with huge heatsinks that are basically the extension of the chipset heatsink when it comes to the lower half of the PCB. The I/O cover has a really cool MSI dragon that illuminates in every color of the rainbow as one would expect. The whole board has this really unique, angular pattern that is definitely in tune with the 2022 design language. The looks of the board will certainly not disappoint you unless you have a particular issue with dragons.

When it comes to features, there is not a lot to say really. All of these high-end Z690 motherboards are absolutely jam-packed with fantastic features such as PCIe Gen 5 support and DDR5 memory support, so the Z690 Carbon is no different in this regard. Connectivity is handled by a 2.5 GbE LAN port or WiFi 6E, both of which are excellent options. The Z690 Carbon also has 4 M.2 ports running at PCIe Gen 4 speeds, which is more than enough for any normal user today. Realistically, this board has everything you

will ever need for a few years to come, and maybe even more.

Overall, the MSI MPG Z690 Carbon is an extremely competent Z690 motherboard from MSI for the Z690 platform. It trades blows with the premium AORUS Master and Maximus Z690 Hero motherboards but is slightly cheaper while offering similar VRM performance, therefore it is our preferred choice for high-end overclocking. The price is still not cheap by any means, but it is competitive in the context of the features that it offers.

6.2.1.2.1.4 ASUS ROG Strix Z690-A

Best Looking Z690 Motherboard

图 6-6

ASUS ROG Strix Z690-A Specifications

Chipset: Z690 | **Memory**: 4x DIMM, 128GB, DDR5-5333 | **Video Outputs**: HDMI and DisplayPort | WiFi | **USB Ports**: 10x rear IO, 7x internal | **Network**: 1x 2.5 GbE LAN, 1x Wi-Fi 6 | **Storage:** 4x M.2, 6x SATA”]

Pros

- **Decent Power Delivery System**
- **Best Looking Z690 Motherboard**
- **An Affordable Z690 Motherboard Option**

Cons

- **Not Suitable For Core i9 Overclocking**
- **Slightly Lower DDR5 Memory Compatibility**

Since the Maximus series is not something that everyone can or should buy, ASUS has also released several mid-range and entry-level motherboards on the Z690 platform. The ASUS ROG Strix Z690-A is one of the more mid-range boards that also has lots of great features and a pretty decent VRM design. The standard ROG series also comes with a choice among DDR5 and DDR4 memory configurations, so be sure to purchase the board that supports the specific memory type that you plan to run.

The power delivery system of the ASUS ROG Strix Z690-A is certainly not on the same level as the ones mentioned before, but that does not mean that it is bad by any stretch. ASUS has packed a 16+1 phase VRM for the Z690-A, which is pretty decent by modern standards. This VRM should be capable of overclocking an i7 12700K comfortably, while it should also be able to achieve some degree of overclocking on the i9 12900K. The VRM cooling is also adequate as ASUS has provided quite large VRM heatsinks with diagonal fins to maximize heat dissipation.

Perhaps the most attractive feature of the ASUS ROG Strix Z690-A is its design and appearance. The board is clad beautifully in white thanks to its sizeable white heatsinks and I/O cover. The RGB design and the Strix script on the I/O cover are absolutely sublime, giving a glassy appearance that would go perfectly in many RGB systems. The chipset heatsink is also white and it extends over the PCIe Gen 4 M.2 drive slots as well to a certain extent. You can see a bit of the black PCB, but that still doesn't take anything away from the looks of this board. The Z690-A from ASUS is our pick for the **best looking Z690 motherboard** out there.

The Z690-A is not all form over function, however, as it packs a pretty serious feature set as well. The PCIe Gen 5 support is standard across all Z690 boards so that also makes an appearance here, along with DDR5 memory support of up to 5333MHz. There are four M.2 slots that support PCIe Gen 4 functionality, which is pretty standard stuff in Z690 motherboards. Furthermore, connectivity is handled by a 2.5 GbE LAN port and WiFi 6. Several high-speed USB ports are also present on the rear I/O with different speed

characteristics.

All in all, you can't really go wrong with the midrange pick of the ASUS ROG Strix Z690-A if you are building a midrange modern gaming PC. Pair this motherboard with an i7 12700K or an i5 12600K and you will have yourself a really competent gaming platform that will easily serve you for several years to come. Sure, it is not as jam-packed with features as the really high-end boards, but it gets the job done while being a bit lighter on your wallet as well.

6.2.1.2.1.5 MSI Pro Z690-A WiFi

Best Value Z690 Motherboard

图 6-7

MSI Pro Z690-A WiFi Specifications

Chipset: Z690 | **Memory**: 4x DIMM, 128GB, DDR5-6400 | **Video Outputs**: HDMI and DisplayPort | WiFi | **USB Ports**: 6x rear IO, 9x internal | **Network**: 1x 2.5 GbE LAN, 1x Wi-Fi 6E | **Storage:** 4x M.2, 6x SATA"]

Pros

- **Relatively Affordable**

- **Solid Connectivity**
- **Great Storage Options**

Cons

- **Poor Design**
- **Mediocre VRM Design**

MSI's Pro series of motherboards has been synonymous with value for a while now, and the trend carries on with the MSI Pro Z690-A motherboards. There are actually 4 different motherboards under the same name, but all 4 of them are the same PCB differentiated only by the two variables of WiFi and DDR generation. In this particular roundup, we recommend the MSI Pro Z690-A DDR5 WiFi motherboard as it still offers a great price-to-performance ratio despite having a higher price than the DDR4 and non-WiFi variants.

When we talk about value-oriented motherboards, we have to be reasonable with our expectations when it comes to power delivery and VRM design. The MSI Pro Z690-A motherboard packs a decent 8+4 phase VRM that will be great for most of the regular users that will be buying a motherboard in this price bracket. The VRM should be able to handle a Core i5 12600K at its highest overclock, but it should also be able to handle some light overclocking on its Core i7 brother. We wouldn't recommend putting a Core i9 part in this Z690 motherboard as those CPUs need a lot of clean, stable power when they are operating even under stock conditions, so that is one pairing we would advise against.

When it comes to looks, the MSI Pro Z690-A is as basic as they come. It is just a massive chunk of black PCB with black heatsinks in various places. This is one area where MSI has disappointed us since there are several boards that are cheaper than this one, that look way better than the Pro Z690-A. MSI has also not included any RGB lighting on the board, which is a bit of a bummer if you are building a PC in 2022. The plus side to this design language, however, is that you don't have to install any RGB bloatware for the motherboard, and this board will go nicely in a stealthy black-themed PC.

The feature-set of the MSI Pro Z690-A is nothing to scoff at. It is not as feature-rich as some of the other boards on this list, understandably, but it has everything a regular gamer might want from a Z690 motherboard. The support for DDR5 and PCIe Gen 5 comes standard with the Z690 chipset, and the Pro Z690-A also has 4 M.2 slots with PCIe Gen 4 capability. There is no built-in I/O shield, but the rear I/O itself is decent with a nice selection of USB ports and a 2.5 GbE LAN port along with WiFi capability if you choose that particular motherboard variant.

Conclusively, the MSI Pro Z690-A is certainly one of the more value-oriented motherboards on the market and one that puts the price-to-performance ratio as first priority. It is for this

reason that the MSI Pro Z690-A is our pick for the **best value Z690 motherboard** out there. It just does everything you would want from a Z690 motherboard on the Alder Lake platform, and it does it at a cheaper price than most competitors. It does have its flaws, but that is to be expected given the price point.

6.2.1.2.1.6 ASUS ROG Strix Z690-I

Best Mini-ITX Z690 Motherboard

图 6-8

ASUS ROG Strix Z690-I

Specifications=”**Chipset**: Z690 | **Memory**: 2x DIMM, 128GB, DDR5-6400 | **Video Outputs**: HDMI | WiFi | **USB Ports**: 9x rear IO, 5x internal | **Network**: 1x 2.5 GbE LAN, 1x Wi-Fi 6E | **Storage:** 2x M.2, 4x SATA”]

Pros

- **Solid Power Delivery**
- **Great For Compact Builds**
- **Impressive Connectivity**

Cons

- **Only 2 DIMM Slots**
- **Expensive**

The market for compact gaming PCs has become really popular over the past few years as more and more enthusiasts are jumping ship over to smaller computers. Not only does a small form factor PC save a lot of space on your setup, but it is also easier to carry around from one place to another should you fancy doing so. Consequently, the mini-ITX motherboards are becoming better and better as well, and the ASUS ROG Strix Z690-I is an example of one of the best mini-ITX motherboards around. The Z690-I is the answer from ASUS for all the small form factor enthusiasts that want to enjoy the features of the full-sized Z690 motherboards in a smaller package.

Starting off with the VRM design and power delivery, there is not a lot to complain about here. Obviously, ASUS has downgraded the VRM a little in order to adapt to the size constraints, but the 10+1 phase VRM with 105 amp power stages should not have any major problems overclocking the 12700K or even the 12900K. The VRM is also cooled by a sizeable finned heatsink, so temperatures should be reasonable. The Z690-I also supports DDR5 memory at up to 6400 MHz speeds, which is a welcome feature as always.

The design of the ASUS ROG Strix Z690-I is quite interesting and also quite daring. The PCB itself is quite small as you would expect from a mini-ITX motherboard, but it is absolutely dwarfed by these gigantic heatsinks and I/O cover. The main I/O cover also extends laterally over the M.2 drive slots and serves as a heatsink by making an L-shape over the PCB. Of course, being a ROG board, the Z690-I also has a healthy share of RGB on the lower-left corner of the I/O cover in the form of the ROG eye that glows up in every color of the rainbow. All in all, a very nice-looking board for its compact size.

ASUS has definitely not skimped on the feature-set despite it being a mini-ITX Z690 motherboard. The board’s connectivity is impressive thanks to a 2.5 GbE LAN port and WiFi 6E capability. There are also two M.2 slots wired for PCIe Gen 4 connectivity, while the main PCIe Gen 5 slot is the full-sized PCIe 16x slot. There are also plenty of other cool features such as BIOS flashback, Clear CMOS button, RGB and aRGB headers, 4 SATA ports, and two additional PCBs for certain components that come with the package. A solid feature-set for a mini-ITX motherboard, no doubt.

Therefore, it can be said that the ASUS ROG Strix Z690-I is the absolute **best mini-ITX Z690 motherboard** on the market right now. Its combination of a great power delivery system and a robust feature set makes it an excellent choice for enthusiasts building a

compact gaming PC on the Alder Lake platform. It does not come cheap, but for the specific use cases that it's designed for, it seems to be worth it.

6.2.1.2.1.7 ASUS PRIME Z690-A

Best Budget Z690 Motherboard

图 6-9

ASUS PRIME Z690-A

Specifications="**Chipset**: Z690 | **Memory**: 4x DIMM, 128GB, DDR4-6000 | **Video Outputs**: HDMI and DisplayPort | WiFi | **USB Ports**: 8x rear IO, 7x internal | **Network**: 1x 2.5 GbE LAN | **Storage:** 4x M.2, 4x SATA"]

Pros

- **Budget Z690 Motherboard**
- **Impressive Looks**

Cons

- **No WiFi**
- **Mediocre VRM Design**

- **Few SATA Ports**

6.2.1.2.1.8 ASROCK Z690 TAICHI

If you're looking for the best Z690 motherboard for a high-end PC build, then we think the ASRock Z690 Taichi is as good as it gets. The Z690 Taichi represents the best of what ASRock has to offer right now. It's packed with excellent features and delivers reliable performance, building a solid platform for a high-end build. It features the Z690 chipset for the new Alder Lake CPUs and it also carries the new LGA 1700 CPU socket for the new chips. You can also buy the Taichi motherboard with chipsets compatible with other CPUs too.

The ASRock Z690 Taichi is a premium motherboard and it looks very unique. You'll know it's a Taichi motherboard when you see one. This particular motherboard features a 20 phase SPS Dr.MOS power design. It's capable of handling even the most demanding CPUs in the Alder Lake series like the Core i9-12900K. The Z690 Taichi also supports overclocking and the VRM heatsink ensures everything stays cool at all times. In addition to the RGB lights, the VRM heatsink cover and the chipset heatsink cover, both feature a very unique bronze-colored gear element. This is exclusive to the Taichi motherboard and it looks great.

As one of the new high-end Z690 motherboards, the Z690 Taichi comes with four DIMM slots that let you install up to 128GB of DDR5 memory modules. ASRock says it supports memory speeds of up to DDR5-6400, which is excellent. It's, however, worth pointing out that you will not be able to install your older DDR4 memory modules on this board. We think it's worth making a note of because DDR5 memory modules are both expensive and super hard to come by, at least for now.

You also get support for PCIe 5.0 peripherals, so you'll be able to install those PCIe Gen 5 peripherals whenever they come out. You get three full-sized reinforced x16 slots on the board in addition to an x1 slot. You only get three M.2 slots for the drives, but you can install additional SATA drives with support for RAID 0,1,5, and 10. The M.2 slots are hidden under the heatsink. This means you don't have to buy M.2 drives with a built-in heatsink as the motherboard will handle the thermals for you.

The ASRock Z690 Taichi comes with a pre-installed IO shield at the back and it covers a ton of different ports. You get as many as eight USB ports at the back including two Thunderbolt 4 USB Type-C ports. There's also an HDMI port for those who want to boot without a discrete GPU. You also get two ethernet ports along with a full stack of audio ports. As a premium board, you also get plenty of headers for additional USB ports, RGB

lights, fans, pumps, and more. Overall, we think the Z690 Taichi is one of the best Z690 motherboards you can buy right now. It’s an expensive motherboard, but it’ll pair nicely with a high-performance CPU like the Intel Core i9-12900K.

ASROCK Z690 TAICHI

图 6-10

The ASRock Z690 Taichi represents the best of what ASRock has to offer in the mainstream PC space. This board comes with all the bells and whistles that you'd expect from a premium motherboard.

Pros

- **Reliable performance**
- **Support extreme overclocking**
- **Supports DDR5 & PCIe 5.0**

Cons

- **Expensive**

6.2.1.2.1.9 ASRock Z690 Extreme WiFi 6E

图 6-11

The new compatible motherboards for Intel Alder Lake CPUs are still relatively new to the market. That's one of the reasons why they're so expensive and hard to come by. However, ASRock has some reliable options which we think are worth considering. We think the Z690 Extreme WiFi 6E is arguably one of the best Z690 motherboards you can buy right now from ASRock. For less than $300, this motherboard offers a lot of value and it doesn't skimp on any important features. In fact, we think this board is powerful enough to handle even the most demanding builds with a high-performance CPU.

The ASRock Z690 Extreme WiFi 6E motherboard, as the name suggests, features Intel's new Z690 chipset for the Alder Lake CPUs. It also carries the LGA 1700 CPU socket that's necessary for the new chips. This ATX motherboard will fit into most mid-tower and full tower PC cases without a hitch. The motherboard has a black-colored PCB that is covered with matching heatsinks and shrouds with RGB lights. A lot of premium ASRock motherboards have some RGB bling and this one's no different.

The VRM heatsink with ASRock branding and the chipset heatsink with 'Extreme' branding, both light up when the system's powered on. There's also a glowing RGB LED strip on the right edge of the motherboard. You can always choose to turn them off and have a muted look, but we think it looks good and adds to the overall aesthetics of the build. The VRM heatsink is separate from the chipset heatsink, but you do get a heat-spreader for the M.2 slots.

The ASRock Z690 Extreme WiFi 6E sports four DIMM slots with support for up to 128GB of DDR4 memory. That's right, this is one of those Z690 motherboards that'll let you carry your older DDR4 memory modules as opposed to making you buy the new DDR5 ones. We think DDR4 memory is still the way to consider how difficult it really is to get your hands on DDR5 memory modules right now. It also makes the overall platform entry cost low, so you save more money for other core components.

You still get support for PCIe 5.0 with the Z690 Extreme WiFi 6E motherboard, though. This means you'll be able to install the new PCIe Gen 5 peripherals whenever they come out. You get a single reinforced PCIe 5.0 x16 slot on the top, a PCIe 4.0 x16 slot under that, followed by another PCIe 3.0 x16 and a PCIe 3.0 x1 slot. For storage, you get a total of three M.2 slots along with support for a bunch of SATA drives with RAID support. The motherboard also offers a decent selection of ports and the IO shield comes pre-installed out of the box, which is great.

The ASRock Z690 Extreme WiFi 6E isn't the most premium Z690 motherboard out there on the market. However, we think it brings a great set of features for a relatively affordable price. It also lets you use your older DDR4 memory modules and comes with support for PCIe 5.0.

ASROCK Z690 EXTREME WIFI 6E MOTHERBOARD

图 6-12

The ASRock Z690 Extreme WiFi 6E may not be the most premium Z690 board out there, but we think it offers a good set of features and reliable performance for the price.

Pros

- **Supports DDR4 memory**
- **Reliable power delivery**
- **Supports PCI 5.0**

Cons

- **Limited M.2 slots**

6.2.1.2.1.10 ASRock Steel Legend WiFi 6E

图 6-13

Intel’s new Alder Lake CPU, as you probably already know supports the new DDR5 memory standard. This means you’ll now be able to buy the new DDR5 memory modules and use them for your PC build with one of the Alder Lake CPUs. DDR5 memory modules are better than outgoing DRR4 modules, but they’ve yet to become mainstream. This is why we think the older DDR4 modules are perfectly viable, even for a high-end build. Well, if you’re in the market to buy a new motherboard with support for DDR4 modules, then we think the ASRock Z690 Steel Legend WiFi 6E is worth considering.

The ASRock Z690 Steel Legends motherboard supports the new Alder Lake CPUs thanks to the Z690 chipset and the LGA 1700 CPU socket. But the best thing about this motherboard is that it lets you bring your older DDR4 memory modules. You can install up to 128GB of DDR4 memory with memory speeds of up to 5000MHz. Sure, the new DDR5 modules bring higher speeds to the table, but they’re equally expensive and very hard to come by. Being able to buy the older modules will also reduce the overall platform entry cost.

The ASRock Z690 Steel Legend motherboard also supports PCIe 5.0. This makes it a futureproof board as you can install the new PCIe Gen 5 peripherals when they come out. You get an x16 PCIe 5.0, a PCIe 4.0 x16, a PCIe 3.0 x16 slot, and two PCIe 3.0 x1 slots on the board. You also get three M.2 slots for storage with the inclusion of additional SATA

connectors. The M.2 slots are hidden under the heat-spreader that extends from the chipset heatsink. The VRM module on this board also has a sophisticated VRM heatsink with a white-colored rear panel cover carrying RGB lights. This motherboard has white aesthetics and we think it'll blend nicely with a white PC case.

This motherboard features a 13 phase Dr.MOS power design for reliable power delivery. It also supports overclocking, which means it's a great option for pushing the new Alder Lake CPUs to their limits. Pair with it a nice LGA 1700 CPU cooler and you should have a solid PC. The IO shield comes pre-installed on the board and it covers a variety of ports including 11 USB ports, a 2.5G LAN, a WiFi 6E antenna socket, an HDMI, and a Displayport for integrated GPUs, and more. You'll also find plenty of headers for USB, fans, pumps, and more for your build.

All in all, the ASRock Z690 Steel Legend is a fantastic motherboard if you're planning a new PC build involving the new Alder Lake CPUs. Just remember that you won't be able to install the new DDR5 modules on this motherboard. You might want to step up either the Z690 Taichi or some other DDR5 compatible ASRock board. The Steel Legend Z690 motherboard is a solid board for the price otherwise, and you can't possibly go wrong with this one.

ASROCK Z690 STEEL LEGEND WIFI 6E

图 6-14

The ASRock Z690 Steel Legend motherboard will let you carry your older DDR4 memory modules with you for the new build, thereby reducing the overall platform entry cost for Alder Lake chips.

Pros

- **Stunning all-white aesthetics**
- **Support overclocking**
- **Reliable performance**

Cons

- **Limited M.2 slots**

6.2.1.2.1.11 ASRock Z690 Phantom Gaming 4

图 6-15

One of the main concerns of building a new Alder Lake PC is the high platform entry cost. Not only do you need to buy the new CPUs, but you'll also need a new motherboard and the new DDR5 RAM kit to get the best performance out of your CPU. Both Z690 motherboards and the DDR5 memory modules are expensive and are mostly out of stock. Well, if you're strapped for cash, then we recommend you check out the ASRock Z690 Phantom Gaming 4 motherboard. This is a sub $180 motherboard that supports the new Alder Lake CPUs. It comes with a Z690 chipset and carries the new LGA 1700 socket too.

The ASRock Z690 Phantom Gaming 4 also supports DDR4 memory modules, which means you don't have to worry about buying the new DDR5 modules. An affordable motherboard that also supports DDR4 memory? It sounds too good to be true, but the ASRock Z690 Phantom Gaming 4 is exactly that. It's a fairly basic motherboard that comes with only the basic features. That being said, it's perfectly serviceable.

Looking at the motherboard itself, the first thing you'll notice is that it looks very basic. There are no sophisticated heatsinks or metal shrouds covering the components. In fact, it doesn't even have any RGB lights. It almost looks like a barebones motherboard. But as we mentioned earlier, it comes with all the essentials to build a reliable Alder Lake PC. It carries the Z690 chipset and the LGA 1700 CPU socket for the Alder Lake chips. You also

get four DIMM slots with support for up to 128GB DDR4 memory. ASRock says it can handle memory speeds of up to 5000MHz, which is pretty good.

The Phantom Gaming 4 motherboard comes with a 9 phase Dr.MOS PD for CPU power. The VRM has a very basic heatsink, but it should be able to handle the new chips. Overclocking performance, however, isn't going to be as reliable as it would be on other ASRock motherboards that we've mentioned in this collection. You might want to keep your expectations in check with that. In terms of the PCIe support, you get one PCIe 5.0 x16 slot, a PCIe 4.0 x16, and a PCIe 3.0 x1 slot at the bottom. For storage, you get a total of three M.2 slots, which isn't too bad. Besides that, you can also add SATA drives to your build.

The IO shield isn't pre-installed on the board out of the box and will instead be bundled inside the box. That's an additional step to take care of while building the PC, but it's not really a surprise considering the price. You do get a decent selection of ports at the back, though. You get as many as 8 USB ports at the back including a USB Type C port. There's also an HDMI port for booting with an integrated GPU and an ethernet port. Overall, we think the ASRock Z690 Phantom Gaming 4 is a very basic motherboard. This is a good option for those looking to build an Alder Lake chip-based PC on a budget, but there's not much to expect in terms of features.

ASROCK Z690 PHANTOM GAMING 4

图 6-16

The ASRock Z690 Phantom Gaming 4 is a very basic budget motherboard for those looking to build a PC with Intel's new Alder Lake CPUs for cheap.

Pros

- **Supports PCIe 5.0**
- **Reliable performance**

Cons

- **No pre-installed IO shield**
- **Barebones design**

Sure, Z690 is the flagship chipset for the Alder Lake platform but there are certainly budget-oriented options in this lineup as well. One of them is the ASUS PRIME Z690-A, which is a motherboard that aims to provide the basic features that you can expect from the Z690 platform at a reasonable price. It is nothing fancy, but it should be perfect if you just want a decent value z690 motherboard to get your Alder Lake system up and running.

One of the areas where you definitely compromise when buying a budget motherboard is the power delivery system. Thankfully, ASUS has not downgraded the VRM design too much and provided a totally acceptable VRM for the price point. The board has a 16+1 phase VRM design with decent cooling that is provided by large heatsinks that are also finned. One should be able to overclock the Core i5 12600K fairly comfortably on this board,

however, you should avoid overclocking any Core i7 or Core i9 parts on this board. DDR5 memory is also supported with speeds up to 6000 MHz.

Aesthetically, the PRIME Z690-A is actually one of the better-looking Z690 boards out there. It looks extremely similar to the Z690-A from the ROG Strix lineup that we mentioned earlier. It has the same white accents as that board, although it is a little bit downgraded in terms of the heatsink area. There is a nice RGB implementation over the I/O cover with the word PRIME written in big block letters as well. M.2 heatsinks are independent of the chipset heatsinks, which is something not common in the Z690 boards we have already seen. All in all, a very attractive motherboard.

The feature-set of the PRIME Z690-A is acceptable as well, although it is a little bit lacking when compared to the high-end boards on this list. The board does support PCIe Gen 5 as standard through the main 16x PCIe slot, and there is Thunderbolt 4 support on this board as well. There is also a 2.5 GbE LAN port for connectivity, although WiFi is missing from the board which can be a bit of a downer for some users. ASUS has included 4 PCIe Gen 4 M.2 slots on the board, which make for a great storage setup on a budget.

The PRIME Z690-A might not be the most premium board out there in terms of VRM performance or features, but it is one of the **best Z690 motherboards** in terms of value. It might just be the **best budget Z690 motherboard** on the market thanks to its decent power delivery system and a versatile feature-set. Quite frankly, this is all a typical gamer needs to get up and running with their new Alder Lake CPU.

6.2.1.2.2 How We Choose The Best Z690 Motherboard

Since the motherboard is the basic building block of any computer, we are very careful in our recommendations when we select Z690 motherboards for our roundups. A particular board has to satisfy several standards for it to be considered in our tier lists. The VRM design and power delivery of the board take the utmost priority for us as enthusiasts, and it is also the parameter that can overpower other characteristics of the board when it comes to recommendations. A board can have a great design and excellent features, but if the VRM is below par then it is hard for us to recommend that board.

Furthermore, we also pay close attention to the feature set of the board. Many Z690 motherboards skimp on some particular features such as WiFi or Ethernet connectivity, which are carefully examined and mentioned by us in these roundups. The general aesthetics and design of the motherboard are also something that we take into account, although not to an extent where it might seem more important than the actual features and specs of the board. Finally, the price tag and value proposition are what really drive our

recommendations. In the modern-day, there are no bad products, only bad prices. This mantra also applies to the products in these roundups, so our selection is driven strongly by the value that these products offer.

6.2.1.2.3 PCIe Gen 5

The new Z690 platform brings with it support for the latest and greatest in PCIe protocols, the PCIe Gen 5. Now this generation does bring huge advantages in terms of bandwidth and link speeds, but only on paper for the time being. Currently, as of the time of writing, there are no devices out there for consumers that make use of PCIe Gen 5 technology. It is also important to note that the PCIe Gen 5 link is only supported by the full-size PCIe 16x GPU slot(s) in Z690 motherboards, not the PCIe x4 slots that the M.2 devices use. The SSDs are still limited to PCIe Gen 4 operation.

PCIe Gen 4 is still far from being saturated when it comes to storage media such as the superfast PCIe Gen 4 NVMe SSDs that are becoming more and more mainstream. Graphics cards that use PCIe Gen 4 technology such as the RTX 3000 series are not bothered at all by the difference between PCIe Gen 4 and PCIe Gen 3. So it is hard to see PCIe Gen 5 making a substantial difference to anyone's day-to-day experience on Z690. However, it is a welcome and forward-looking addition by Intel and the motherboard manufacturers so there's nothing particularly negative to say about it.

6.2.1.2.4 DDR4 vs. DDR5

The choice between DDR4 and DDR5 is an important one that you have to consider when purchasing the **best Z690 motherboard** for you. Firstly, you have to choose one of the two memory generations to run, since there are different motherboard variants that support different memory types. You can make the choice between two different RAM generations, but you can't run both of them on the same motherboard. Since DDR4 and DDR5 modules are also physically different, it would be impossible to insert one RAM stick into the DIMM slot of the other, unless you go for some slightly unconventional methods.

The DDR4 variants of the Z690 motherboards are far cheaper than the DDR5 variants of the same motherboards. It is also to be noted that DDR5 memory is extremely expensive right now and offers little-to-no benefit in terms of performance over DDR4, so you should evaluate your choice carefully. If you choose a DDR4 motherboard now and plan to upgrade to DDR5 later, you will have to change out your entire motherboard as well which can be a bit of a hassle. The bottom line is that DDR5 variants of the Z690 motherboards are better for future-proofing, while DDR4 variants should be your choice if you want to maximize the value proposition.

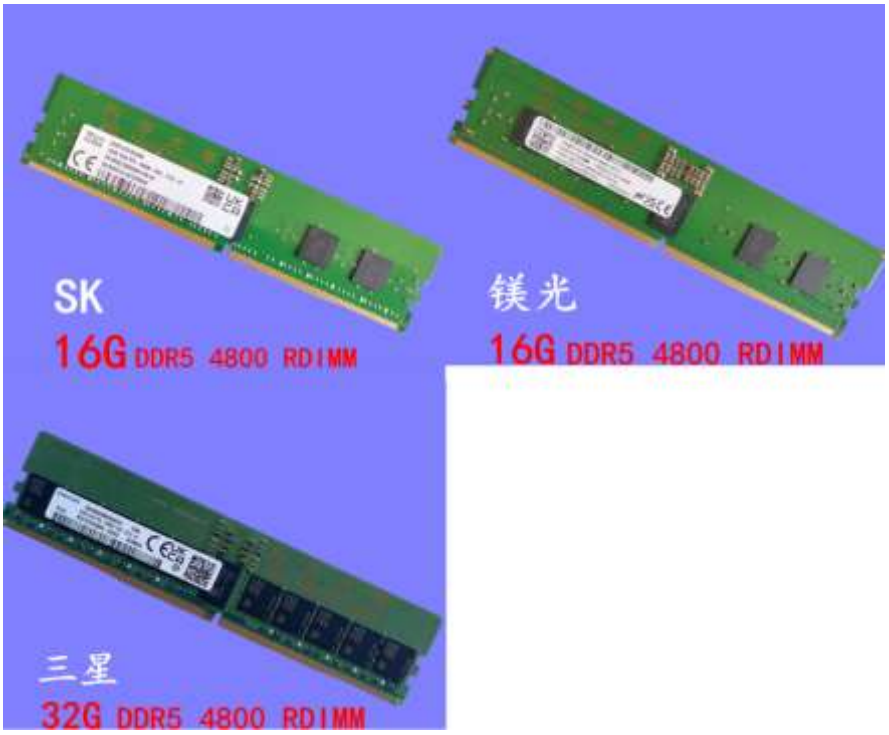


图 6-17

6.2.1.2.5 Frequently Asked Questions

Is Z690 compatible with i9 12900K?

Yes, the Z690 motherboards are compatible with all the new Intel 12th Gen CPUs on the Alder Lake architecture including the top range Intel Core i9 12900K. The i9 12900K is a premium CPU with 8 Performance Cores and 8 Efficiency Cores with a total of 24 threads, so it would need a pretty premium power delivery system to be functioning at peak capacity. Fortunately, most Z690 motherboards do have overbuilt robust VRM designs, so they should be able to handle a 12900K fairly easily. Other CPUs that support the Z690 chipset as of the time of writing include the i9 12900KF, the i7 12700K and 12700KF, and the i5 12600K and 12600KF.

What socket is Z690?

The Z690 motherboards have the LGA 1700 socket which supports the new Intel Core 12th Gen processors on the Alder Lake architecture. The socket is physically different from the older Z590 motherboards that supported the LGA 1200 socket compatible with 11th Gen Rocket Lake processors. This means that you cannot install an 11th Gen CPU on a 12th Gen motherboard or vice versa, since they are both physically different sockets. Forcefully doing so can lead to some unwanted consequences.

Does Z690 support PCIe 5.0?

Yes indeed, the Z690 motherboards do support PCIe Gen 5 capability. However, this new protocol does not have any real-life advantages as of the time of writing since there is no consumer PCIe Gen 5 devices available to us right now. The PCIe lanes are also all linked to the first 16x PCIe slot, so the only devices that can take the benefit of the PCIe Gen 5 technology would be the graphics cards. There are no M.2 slots with PCIe Gen 5 capability yet in the Z690 motherboards.

Can I use 11900K with Z690?

No, the Z690 platform is an entirely new platform that is not compatible with any previous Intel CPUs. The Core i9 11900K is a Rocket Lake 11th Gen CPU that is compatible with an LGA 1200 socket on the Z590 platform, among other chipsets. The new Z690 motherboards use the LGA 1700 socket which is physically different from the older LGA sockets, so the CPU would not even fit in the new socket. You would need either a Z590 motherboard or a B560 or similar board for the older 11th Gen CPU.

Does Z690 support DDR5?

Yes, the Z690 platform does officially support DDR5 memory. To use DDR5 RAM, you would need to have a motherboard that is compatible with DDR5 memory, since the memory sticks are physically different from DDR4. Z690 motherboards are offered in both DDR5 and DDR4 memory configurations, so you have to be careful about which memory type you want to buy when it comes to motherboard selection. The Alder Lake 12th Gen CPUs also support both DDR5 and DDR4 memory.

6.2.2 AMD 架构平台

6.2.2.1 AMD Gen5 Genoa CPU 服务器

下面是一款典型的基于 AMD Gen5 CPU 的服务器主板，平时开放式测试比较方便。

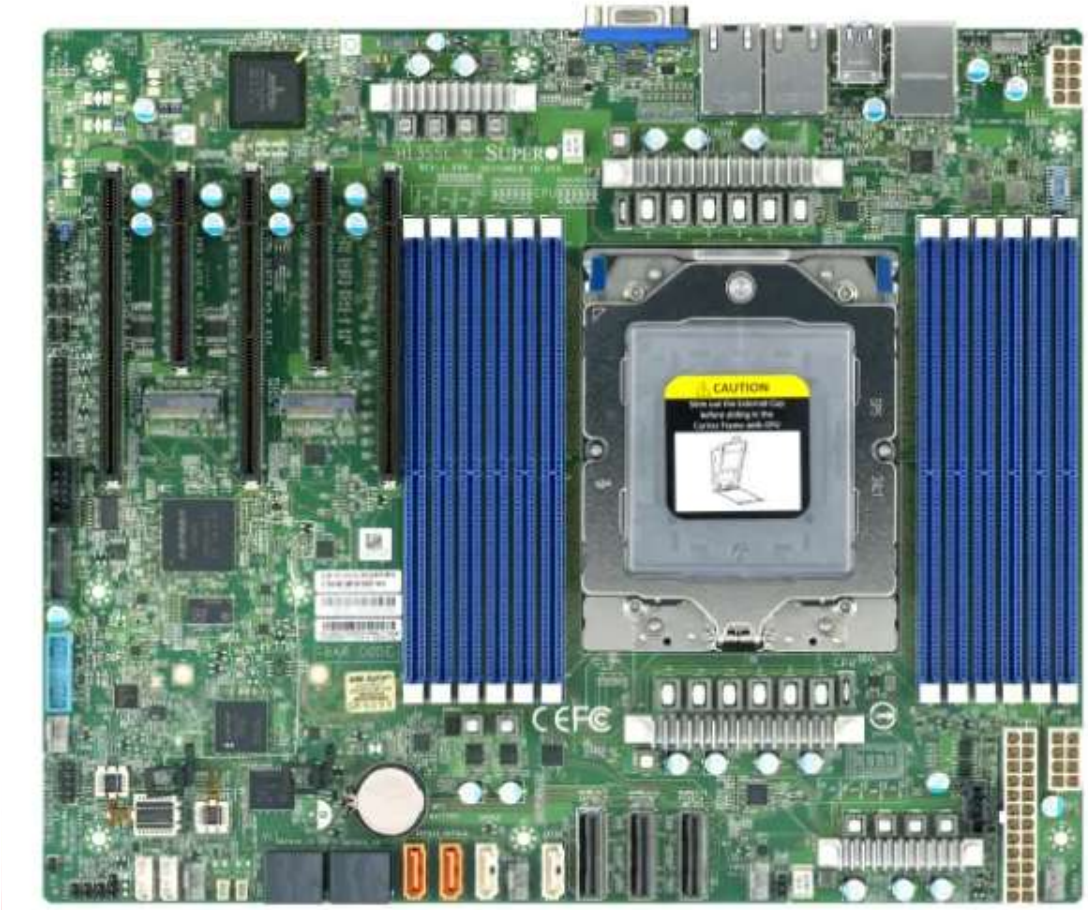


图 6-18

主要特征

- **AMD EPYC（霄龙）9004 系列处理器**
- **片上系统**
- **高达 3TB 3DS ECC RDIMM，12 个 DIMM 插槽中的 DDR5-4800MHz**
- **3 个 PCI-E 5.0 x16 2 个 PCI-E 5.0 x8**
- **多达 6 个 USB 3.0 端口（4 个后置 + 2 个通过接头连接器）**
- **8 个 SATA3**
- **6 个带转速计状态监控的 PWM 4 针风扇**
- **M.2 接口： 2 PCI-E 4.0 x4 M.2 规格： 2280、22110**

注意：AMD 的 Genoa CPU 都支持 CXL

当然，如果不考虑开放式测试的便利性，也可以直接购买 Super Micro 等公司的 1RU, 2RU 整机使用更方便。参见下图。

图 6-19

6.2.2.2 AMD Gen5 CPU 工作站

2022 年基于 AMD X670 芯片组 PCIe Gen5 主板

6.2.2.2.1 AMD X670E VS X670 Motherboards – Key Differences

图 6-20

6.2.2.2.1.1 Introduction

With the announcement of AMD's newest Ryzen 7000 CPU launch at Computex 2022, we were given an insight into the latest generation of AMD CPUs set to rival Intel's 12th Generation chips.

AMD teased us with titbits of information about the new chipsets, motherboard choices, and gave us some benchmarks. We're all very excited at the GeekaWhat office and can't wait for AMD to drop these new CPUs, but a burning question arises from this new product launch. What is the difference between the X670 chipsets?

Today we'll be answering that question and covering the major differences between the brand new X670E chipset and X670 with all of the info we know so far. We'll be looking at some of the technical information from the new motherboard releases, and covering all the various specs and prices that come with each chipset.

Table Of Contents

6.2.2.2.1.2 What is a Chipset?

First and foremost lets briefly cover what a chipset is. Although a chipset and motherboard go hand-in-hand, they are not the same thing. A motherboard is the physical circuit board that allows you to slot in your individual components. A chipset is a data communication centre and traffic controller that sits on top of your motherboard. Your chipset also determines the features you should expect on your motherboard such as: PCI-E expansion lanes, overclocking support, internal and rear IO etc.

AMD's chipset names and numbers are somewhat different to Intel. AMD's flagship chipset has always been 'X#70', whereas Intel's current high-end chipset is Z690. AMD essentially jumps the number up by one every time a new series of CPU is released. Ryzen 5000 CPUs were X570, and the previous generation was X470, making the new 7000 CPUs X670. Where these new Ryzen 7000-series differ is that AMD have never had an 'X#70**E**'chipset before. The E designation is supposed to allude to the chipset being more 'extreme' and enthusiast in its makeup.

Originally the 'X#70' chipset was geared towards the highest-end market, but it appears that may have changed with the introduction of X670E. Both chipsets are set to have a

wide range of features and by contrast to Intel's Z690 motherboards, will have more PCI-E Gen4 SSD support, PCI-E 5.0 support across the board, and plenty of rear IO.

6.2.2.2.1.3 Major Differences

With the release of the new Ryzen 7000 range, AMD kindly sent us two of their brand new CPUs, along with one of MSI's top-end motherboards to test performance. We've been able to take a look at all of the new features on this board, along with technical information that AMD has released leading up to this launch. We've been working day and night to review all of the new features, and as part of this, we've updated this article to reflect all of the new technical information about the new chipsets.

6.2.2.2.1.3.1 VRM Power Phases

As a general rule with AMD, overclocking is not limited to the motherboard but instead the CPU. AMD CPUs that have all cores unlocked have an 'X' designation, and provided AMD is sticking with the same philosophy, all of the new chipsets should support overclocking. Where they will differ is the amount of power phases and VRM cooling on the board.

图 6-21

For the X670E chipset, these motherboards are designed to push the overclocking capability of Ryzen 7000 CPUs to the max. The ASUS ROG Crosshair X670E motherboard has 20+2 power stages designed for optimal stability and to push performance to the highest it can go. If you're picking up a next-gen Ryzen 9 CPU, you will likely be buying an X670E motherboard to pair with it for the best performance.

On the other side of the coin, X670 motherboards are not far behind when it comes to overclocking. The Gigabyte X670 AORUS Pro AX board features a 16+2+2 power delivery system. This again is great for pushing your clock speed to the max, and with optimal cooling you'll have a powerful system.

6.2.2.2.1.3.2 PCI-E Lanes, Graphics Cards & SSDs

With concerns to PCI-E lanes, the X670E chipset will have the most PCI-E 5.0 designated lanes out of all of the options. AMD announced that the new AM5 socket will have 24 PCI-E 5.0 lanes to play around with, but ultimately the full utilisation of these lanes will depend on the manufacturer and motherboard.

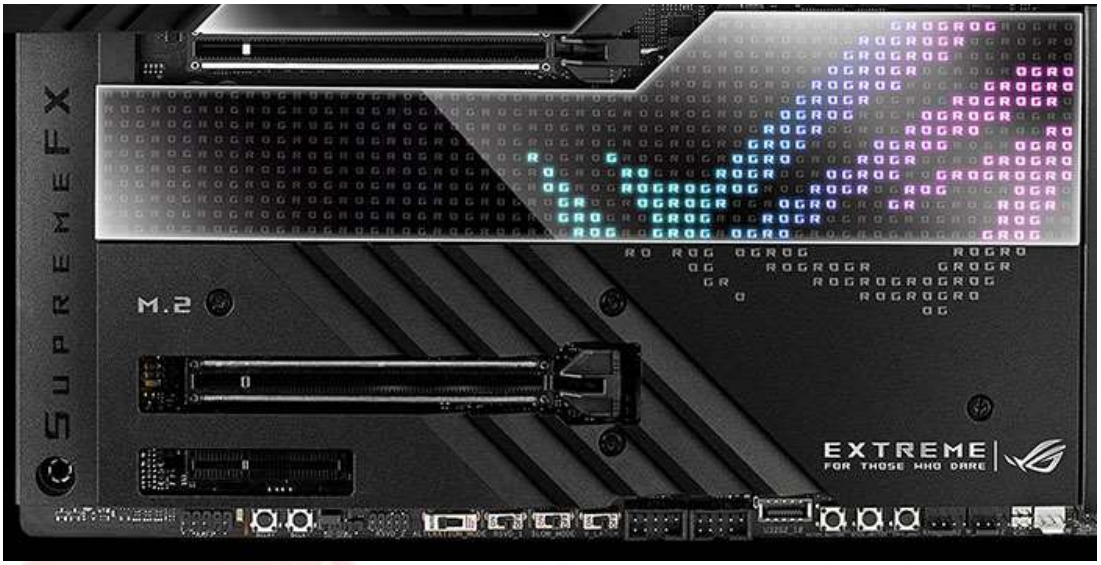


图 6-22

As X670E is the top-end chipset we believe that these motherboards will have full access to all of the 24 PCI-E 5.0 lanes. For the standard X670 chipset, these boards will only have four PCI-E 5.0 lanes, and 44 usable lanes in total (of which the rest will be PCI-E 4.0 and 3.0). This means you'll unfortunately only have one usable PCI-E 5.0 slot which will in this case be an M.2 drive.

To summarize, if you're wanting to pick up the next generation of graphics cards and maximize your next-gen storage options, then the X670E chipset will be your go-to choice. But if you're not bothered about all of the new component choices, then you can get away with picking up an X670 board which will still have plenty of options, but much less of the next-gen features.

6.2.2.2.1.3.3 CPU Overclocking

This might sound a little dire, but it isn't all bad. The new chipsets are set to have out of the box support for even higher clock speeds than Intel's 12th-gen processors, meaning for those that want to overclock, you'll definitely be in luck here. The Ryzen 9 7900X has a whopping boost clock speed of 5.6GHz, with their mid-range option offering up 5.3GHz on the boost. This is very impressive, and for overclockers out there, we expect these CPUs to be pushed over the 6GHz threshold.

图 6-23

We briefly touched upon this above in the VRM power phases section, but although every Ryzen CPU (so far) is unlocked for overclocking, whether you decide to overclock will depend on a few key factors. The X670E chipsets will have the most amount of VRM power phases, and for this reason they are best positioned for those that want to overclock their CPUs. But the major caveat to picking up one of these boards is that they're pretty expensive.

X670 motherboards seem to have less power phases overall compared to the X670E options, but these boards will still be perfectly fine to get some overclocking underway. Ultimately we'd recommend checking your manufacturer's board product page before deciding whether you should overclock, as the amount of power phases and cooling can hinder how much performance you can get out of your CPU.

6.2.2.2.1.3.4 Memory Overclocking

The biggest point to note here with memory support is that the newest Ryzen 7000 CPUs will not support DDR4 RAM and this is a standard across all of the available chipsets meaning an upgrade to DDR5 is unfortunately a requirement.

This could go one way or another. AMD forcing the upgrade to DDR5 might make the sticks become a lot cheaper making it easier for consumers to pick up. Or on the other hand, DDR5 memory DIMMs may still stay at their exorbitant price making the new CPUs inaccessible for budget consumers.

图 6-24

Although the choice of DDR5-only on the AM5 platform could be an issue, AMD has introduced a new XMP profile that can see performance increases. AMD's EXPO memory technology increases the clock speed of your DIMMs, while reducing the latency of the kit. Doing this could see performance improvements, but we're yet to see a huge difference in performance when this new profile is enabled. All of the new AMD chipsets support this technology, but you'll need to pick up a memory kit that supports the new tech.

6.2.2.2.1.4 Pricing Expectation & Comparison

Firstly, you'll notice that both of the chipsets are immediately much higher than the price range we witnessed with the release of the X570 chipset. This is ultimately down to new tech in the motherboards and the cost associated. PCI-E 5.0 components aren't available yet for consumers, and because the tech doubles the bandwidth of the previous generation we'll be seeing some significant performance boosts. DDR5 being a requirement also immediately boosts the price as the DIMMs themselves are still rather expensive.

X570 motherboards were generally geared towards those that wanted the best performance that they could get, but there were also some options for mid-range and potentially even budget builds. For a budget experience with Ryzen 7000 CPUs, you'll need to look out for B650 boards, as these motherboards will have less features, making them cheaper overall. The X670 chipset is positioned in the middle of the road versus the other chipsets. We have seen one motherboard priced below the $300 mark in the form of MSI's Pro X670-P WiFi, but you're still having to pay in excess for the other components.

For X670E, this is the cream of the crop when it comes to performance and features. These new motherboards will have all of the latest tech available with a wide range of features across the board – hence the higher price. If you're saddened by our estimation of the price then we'd recommend holding out for the B650 boards, but if you're looking to build the most kitted out system with no price restrictions, then X670E will be for you.

X670E	X670	X570
$500 – $1000	$300 – $600	$200 – $500

Note- This is a rough guide on the average price of available motherboards on AMD chipsets. X670E and X670 motherboards do not have an official price and will be subject to change.

6.2.2.2.1.4.1 Further Pricing Updates

As we're nearing the alleged release date of Ryzen 7000, more leaks and information circulates the internet relating to the motherboards. One of these pieces of information is pricing updates for many of the X670 and X670E motherboards. With the image we've provided you'll be able to see that these motherboards are priced very high. Note, that this likely isn't the final price of these Ryzen 7000 motherboards, but if we're remotely close to these prices upon release, the top-end ASUS motherboard will release for around $1500. On the other side of this, ASUS' lowest price motherboard will release for a more reasonable $500.

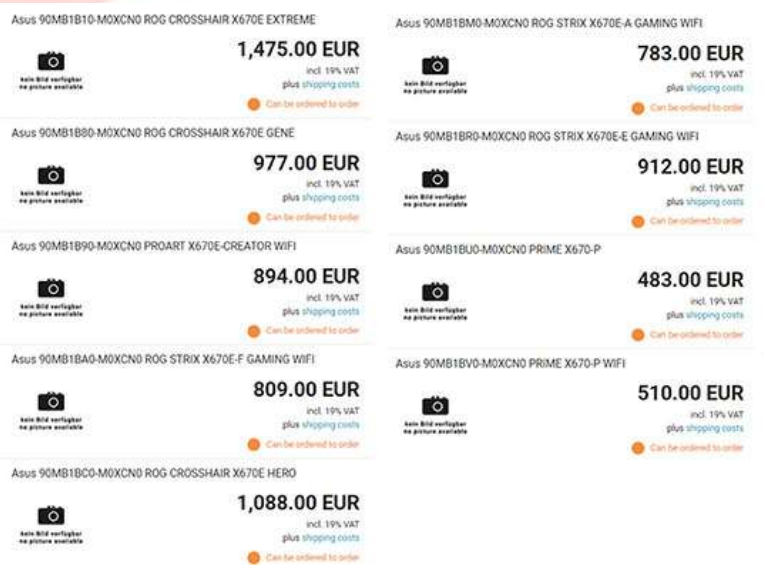


图 6-25

Leaked from an Italian retail store, we're seeing some other reasonable prices from MSI. MSI's Pro range of boards tend to be more budget oriented, but we're seeing this X670-P price at 418 euros, which roughly converts to $416. Although we know that X670 is more

of a middle-of-the-road chipset, over $400 does seem expensive. Presumably this is due to the DDR5 upgrade, but even then, this does make the boards less appealing based on the price.

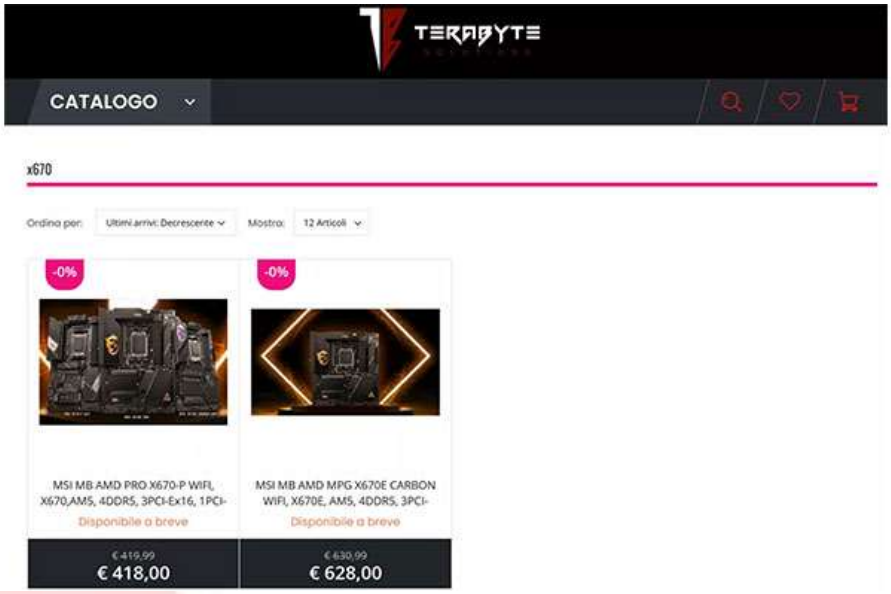


图 6-26

Recently we've also seen MSI confirm the prices for four of their motherboards. We've got the prices for their flagship and GodLike motherboards along with an MPG Carbon board and a Pro X670 motherboard. We're pleasantly surprised with these prices, obviously these boards are more expensive due to them supporting DDR5 only. The more budget oriented Pro X670-P board is nearing around $300. This is what we expect for a cheaper X670 board, and this is a very fair offering considering the DDR5 and PCI-E 5.0 utilisation.

图 6-27

As a final point with pricing expectations, although these boards are definitely going to be more expensive than your average DDR4 board. It is unlikely that manufacturers will price them ridiculously high, or else they could face a paper launch. As much as these new

boards are DDR5 only, it wouldn't be sensible to price them out of reach for all types of consumers. So despite the fact that these ASUS boards are looking to expensive, it is unlikely these new motherboards will be priced that high other than variations at the top-end for enthusiasts (as an example, MSI's Z690 Godlike board). Either way, don't worry about not being able to afford one of these new boards.

6.2.2.2.1.5 Overall Differences

Features	X670E	X670
CPU Overclocking Support	Yes	Yes
CPU PCI Express 5.0 Lanes	24	4
PCI-E 5.0 Slots	1 x16 Slot	1 x16 Slot
	1 x8 Slot	1 x4 M.2 Slot
	1 x4 M.2 Slot	
VRM Power Phases	24+	20+
Chipset PCI Express 4.0 Lanes	20	40
PCI-E 4.0 M.2 Slots	2 x4 M.2 Slots	2 x4 M.2 Slots
Chipset PCI Express 3.0 Lanes	0	8
Max Number of USB Ports	16	14
Max USB 4.0 Ports	2	0
Max USB 3.2 Gen 2×2 Ports (20Gbps)	2	2

Table of Expected Differences between Chipsets (Based off of ASRock X670E Taichi Specs)

6.2.2.2.1.5.1 Overall Differences Breakdown

With any new CPU release, there are a huge amount of new specifications that comes with each chipset. First and foremost we'll talk about PCI-E lanes. AMD has told us that the AM5 socket can support up to 24 PCI-E 5.0 lanes. As X670E is the top-end chipset of the line-up we know that new motherboards will likely utilise all of these both on the x16 slots and x4. As X670 is the middle of the road, you'll have plenty of PCI-E lanes, but only four of them are PCI-E 5.0 compatible.

The other guesses we've made come from technical information we have about the motherboards. For example, we know that the X670E ROG Extreme motherboard has two PCI-E x16 Gen 5 slots, four PCI-E 5.0 x4 slots, so we know that there is going to be very little focus on PCI-E 4.0 for this particular chipset. As a general rule of thumb, the X670E chipset is going to maximise all of the available features (of which there are many), but at a more price conscious level. This positions the X670 motherboard in a more mid-range market, but at a slightly lower cost.

6.2.2.2.1.6 Where to Buy

The Ryzen 7000 CPUs are dropping on September 26th, and we're giddy just talking about it. The motherboards will release at the same time that the CPUs drop, so for the time being you'll want to check out all of our reviews and coverage on the new motherboards to

see what features are available, and where you can buy them.

6.2.2.2.2 Best X670E Motherboards

6.2.2.2.2.1 Asus X670E 主板

ASUS X670E / X670 Motherboards

Brand/Series		ROG CROSSHAIR			ROG Strix			
Model		ROG CROSSHAIR X670E EXTREME	ROG CROSSHAIR X670E HERO	ROG CROSSHAIR X670E GENE	ROG STRIX X670E-E GAMING WIFI	ROG STRIX X670E-F GAMING WIFI	ROG STRIX X670E-A GAMING WIFI	ROG STRIX X670E-I GAMING WIFI
Photo								
CPU socket		AM5						
Chipset		X670						
Form factor		EATX	ATX	mATX	ATX	ATX	ATX	Mini-ITX
Power architecture		20+2 (110A)	18+2 (110A)	16+2 (110A)	18+2 (110A)	16+2 (90A)	16+2 (70A)	10+2 (110A)
Memory	# Slots, Maximum capacity	4 x DIMM, Max. 128GB, DDR5	4 x DIMM, Max. 128GB, DDR5	2 x DIMM, Max. 64GB, DDR5	4 x DIMM, Max. 128GB, DDR5	4 x DIMM, Max. 128GB, DDR5	4 x DIMM, Max. 128GB, DDR5	2 x DIMM, Max. 64GB, DDR5
	OptiMem	OptiMem II	OptiMem II	OptiMem II	OptiMem II	OptiMem II	OptiMem II	OptiMem II
Graphics output		USB4	HDMI 2.1 / USB4	USB4	HDMI 2.1 / DP 1.4	HDMI 2.1 / DP 1.4	HDMI 2.1 / DP 1.4	HDMI 2.1 / USB4
Expansion slots	PCIe 5.0 x16	2 x PCIe 5.0 x16 (@x16 or x8/x8)	2 x PCIe 5.0 x16 (@x16 or x8/x8)	1 x PCIe 5.0 x16 (@x16)	2 x PCIe 5.0 x16 (@x16 or x8/x4)	1 x PCIe 5.0 x16	1 x PCIe 5.0 x16 (@x16)	1 x PCIe 5.0 x16 (@x16)
	PCIe 4.0 x16	—	—	—	1 x PCIe 4.0 x16 (@x4)	1 x PCIe 4.0 x16 (@x4)	1 x PCIe 4.0 x16 (@x4)	—
	PCIe 4.0 x4	1 x PCIe 4.0 x4	—	—	—	—	—	—
	PCIe x1 4.0 or 3.0	—	1 x PCIe 4.0 x1	1 x PCIe 4.0 x1	—	1 x PCIe 3.0 x1	1 x PCIe 3.0 x1	—
Storage & Connectivity	SATA 6Gb/s	6	6	4	4	4	4	2 from FPS-II
	M.2	1 x 22110 (PCIe 5.0 x4) from ROG GEN-Z.2 1 x 22110 (PCIe 4.0 x4) from ROG GEN-Z.2 1 x 22110 (PCIe 5.0 x4) from PCIE 5.0 M.2 Card 2 x 2280 (PCIe 5.0 x4)	2 x 2280 (PCIe 5.0 x4) 1 x 22110 (PCIe 5.0 x4) from PCIE 5.0 M.2 Card 2 x 2280 (PCIe 4.0 x4)	1 x 22110 (PCIe 5.0 x4) from ROG GEN-Z.2 1 x 22110 (PCIe 4.0 x4) from ROG GEN-Z.2 1 x 2280 (PCIe 5.0 x4)	1 x 22110 (PCIe 5.0 x4) 2 x 2280 (PCIe 5.0 x4) 1 x 2280 (PCIe 4.0 x4)	2 x 2280 (PCIe 5.0 x4) 1 x 22110 (PCIe 4.0 x4) 1 x 2280(PCIe 4.0 x4)	2 x 2280(PCIe 5.0 x4) 1 x 22110 (PCIe 4.0 x4) 1 x 2280(PCIe 4.0 x4)	1 x 2280(PCIe 5.0 x4) 1 x 2280(PCIe 4.0 x4)
	Front Panel Type-C Connector	1 x USB 3.2 Gen 2x2 (with Quick Charge 4+) 1 x USB 3.2 Gen 2	1 x USB 3.2 Gen 2x2 (with Quick Charge 4+)	1 x USB 3.2 Gen 2x2 (with Quick Charge 4+)	1 x USB 3.2 Gen 2x2	1 x USB 3.2 Gen 2x2	1 x USB 3.2 Gen 2x2	1 x USB 3.2 Gen 2
	USB4®	2	2	2	—	—	—	2
	USB 3.2 Gen 2x2	2 (1C@B, 1C@F)	2 (1C@B, 1C@F)	1 (1C@F)	2 (1C@B, 1C@F)	2 (1C@B,1C@F)	2 (1C@B,1C@F)	—
	USB 3.2 Gen 2	10 (8A1C@B, 1C@F)	9 (8A1C@B)	6 (5A1C@B)	12 (10A+2C @B)	9 (7A2C@B)	8 (7A1C@B)	6(5A@B, 1C@F, 1C@ROG STRIX HIVE)

	USB 3.2 Gen 1	4 (4@F)	4 (4@F)	2 (2@F)	2 (2A@F)	2 (2A@F)	3 (1A@B, 2A@F)	2 (2A@F)
	USB 2.0	4 (4@F)	6 (6@F)	6 (2@B,4@F)	6 (6A@F)	6 (2A@B, 4A@F)	6 (2A@B, 4A@F)	6 (3A@B,3@FPS-II, 1A@ROG STRIX HIVE)
Wireless		Wi-Fi 6E	Wi-Fi 6E	Wi-Fi 6E	Wi-Fi 6E	Wi-Fi 6E	Wi-Fi 6E	Wi-Fi 6E
Networking	Ethernet	1 x Intel® 2.5 Gb 1 x Marvell® 10 Gb	1 x Intel® 2.5 Gb	1 x Intel® 2.5 Gb	1 x Intel® 2.5 Gb	1 x Intel® 2.5 Gb	1 x Intel® 2.5 Gb	1 x Intel® 2.5 Gb
Audio	Audio codec	ALC 4082	ALC 4082	ALC 4080	ALC 4080	ALC 4080	ALC 4080	ALC 4050
	Impedence sense	V	V	V	V	V	V	—
	DAC/AMP	ESS® ES9218 QUAD DAC	ESS® ES9218 QUAD DAC	Savitech SV3H712 AMP	Savitech SV3H712 AMP	Savitech SV3H712 AMP	Savitech SV3H712 AMP	ESS® ES9260Q QUAD DAC
	Audio cover	V	V	V	V	V	V	—
	LED illuminated audio jacks	V	—	—	—	—	—	—
	Premium capacitors	V	V	V	V	V	V	—
	Audio effects	Sonic Studio III Sonic Studio Virtual Mixer Sonic Radar III DTS® Sound Unbound	Sonic Studio III Sonic Studio Virtual Mixer Sonic Radar III DTS® Sound Unbound	Sonic Studio III Sonic Studio Virtual Mixer Sonic Radar III DTS® Sound Unbound	Sonic Studio III Sonic Studio Virtual Mixer Sonic Radar III DTS® Sound Unbound	Sonic Studio III Sonic Studio Virtual Mixer Sonic Radar III DTS® Sound Unbound	Sonic Studio III Sonic Studio Virtual Mixer Sonic Radar III DTS® Sound Unbound	Sonic Studio III Sonic Studio Virtual Mixer Sonic Radar III DTS® Sound Unbound
Cooling	Onboard fan headers	8 ROG FAN CONTROLLER (Bundled)	8	7	8	8	8	3
	Heatsinked M.2	5	5	3	4	4	4	2
	M.2 backplate	2	3	1	1	1	1	—
Dynamic OC Switcher		V	V	V	V	V	V	V
AI Overclocking		V	V	V	V	V	V	V
AI Cooling II		V	V	V	V	V	V	V
AI Networking (GameFirst VI)		V	V	V	V	V	V	V
Two-Way AI Noise-Cancelation		V	V	V	V	V	V	V
DIY Friendly	M.2 Q-Latch	V	V	V	V	V	V	V
	PCIe Slot Q-Release	V	V	V	V	V	V	—
	PCIe 5.0 SafeSlot	V	V	V	V	—	V	V
	SafeDIMM	V	V	V	V	V	V	V
	Pre-mount I/O shield	V	V	V	V	V	V	V
BIOS FlashBack™		V	V	V	V	V	V	V @ ROG STRIX HIVE
Aura Sync		V	V	V	V	V	V	V
Aura lighting control		3 x ARGB + 1 x RGB	3 x ARGB + 1 x RGB	2 x ARGB + 1 x RGB	3 x ARGB + 1 x RGB	3 x ARGB + 1 x RGB	3 x ARGB + 1 x RGB	1 x ARGB+1 x RGB
CPU power connector		8+8 Pin ProCool II	8+8 Pin ProCool II	8+8 Pin ProCool II	8+8 Pin ProCool II	8+8 Pin ProCool II	8+8 Pin ProCool II	8 Pin ProCool II
COM port/ header		—	—	—	—	—	—	—
TPM		Firmware TPM	Firmware TPM	Firmware TPM	Firmware TPM	Firmware TPM	Firmware TPM	Firmware TPM

Brand/Series		ProArt	TUF Gaming		Prime		
Model		ProArt X670E-CREATOR WIFI	TUF GAMING X670E-PLUS WIFI	TUF GAMING X670E-PLUS	PRIME X670E-PRO WIFI	PRIME X670-P WIFI	PRIME X670-P
Photo							
CPU socket							
Chipset							
Form factor		ATX	ATX	ATX	ATX	ATX	ATX
Power architecture		16+2(70A)	14+2(70A)	14+2(70A)	14+2(70A)	12+2 (60A)	12+2 (60A)
Memory	# Slots, Maximum capacity	4 x DIMM, Max. 128GB, DDR5	4 x DIMM, Max. 128GB, DDR5	4 x DIMM, Max. 128GB, DDR5	4 x DIMM, Max. 128GB, DDR5	4 x DIMM, Max. 128GB, DDR5	4 x DIMM, Max. 128GB, DDR5
	OptiMem	OptiMem II	OptiMem II	OptiMem II	OptiMem II	OptiMem II	OptiMem II
Graphics output		HDMI 2.1 / USB4	HDMI 2.1 / DP 1.4	HDMI 2.1 / DP 1.4	HDMI 2.1 / DP 1.4	HDMI 2.1 / DP 1.4	HDMI 2.1 / DP 1.4
Expansion slots	PCIe 5.0 x16	2 x PCIe 5.0 x16 (@x16 or x8/x8)	1 x PCIe 5.0 x16	1 x PCIe 5.0 x16	1 x PCIe 5.0 x16	—	—
	PCIe 4.0 x16	— 1 x PCIe 4.0 x16 (@x2)	1 x PCIe 4.0 x16 (@x4)	1 x PCIe 4.0 x16 (@x4)	1 x PCIe 4.0 x16 (@x4)	1 x PCIe 4.0 x16 2 x PCIe 4.0 x16 (@x4)	1 x PCIe 4.0 x16 2 x PCIe 4.0 x16 (@x4)
	PCIe 4.0 x4	— —	1 x PCIe 4.0 x4	1 x PCIe 4.0 x4	1 x PCIe 4.0 x4	—	—
	PCIe x1 4.0 or 3.0	— —	—	—	—	1 x PCIe 3.0 x1	1 x PCIe 3.0 x1
Storage & Connectivity	SATA 6Gb/s	4	4	4	4	6	6
	M.2	2 x 2280 (PCIe 5.0 x4) 1 x 22110 (PCIe 4.0 x4) 1 x 2280(PCIe 4.0 x4)	1 x 2280 (PCIe 5.0 x4) 2 x 2280 (PCIe 4.0 x4) 1 x 22110 (PCIe 3.0 x4 & SATA)	1 x 2280 (PCIe 5.0 x4) 2 x 2280 (PCIe 4.0 x4) 1 x 22110 (PCIe 3.0 x4 & SATA)	1 x 2280 (PCIe 5.0 x4) 2 x 2280 (PCIe 4.0 x4) 1 x 22110 (PCIe 3.0 x4 & SATA)	1 x 2280 (PCIe 5.0 x4) 1 x 22110 (PCIe 4.0 x4) 1 x 2280 (PCIe 4.0 x4)	1 x 2280 (PCIe 5.0 x4) 1 x 22110 (PCIe 4.0 x4) 1 x 2280 (PCIe 4.0 x4)
	Front Panel Type-C Connector	1 x USB 3.2 Gen 2x2 (with Quick Charge 4+)	1 x USB 3.2 Gen 2	1 x USB 3.2 Gen 2	1 x USB 3.2 Gen 2	1 x USB 3.2 Gen 1	1 x USB 3.2 Gen 1
	USB4®	— 2	—	—	—	—	—
	USB 3.2 Gen 2x2	— 2 (1C@B,1C@F)	1 (1C@B)	1 (1C@B)	1 (1C@B)	1 (1C@B)	1 (1C@B)
	USB 3.2 Gen 2	7 (7A@B)	5 (1C3A@B, 1C@F)	5 (1C3A@B, 1C@F)	5 (1C3A@B, 1C@F)	3 (3A@B)	3 (3A@B)
	USB 3.2 Gen 1	2 (2A@F)	7 (5A@B, 2A@F)	7 (5A@B, 2A@F)	7 (4A1C@B, 2A@F)	9 (4A@B, 4A+1C@F)	9 (4A@B, 4A+1C@F)
	USB 2.0	7 (1A@B, 6A@F)	6 (6A@F)	6 (6A@F)	6 (6A@F)	6 (2@B,4@F)	6 (2@B,4@F)
Wireless		Wi-Fi 6E	Wi-Fi 6E	—	Wi-Fi 6E	Wi-Fi 6	1 x V-M.2 slot (Key E)
Networking	Ethernet	1 x Intel® 2.5 Gb 1 x Marvell® 10 Gb	1 x Realtek 2.5 Gb	1 x Realtek 2.5 Gb	1 x Realtek 2.5 Gb	1 x Realtek 2.5 Gb	1 x Realtek 2.5 Gb

Audio	Audio codec	Realtek S1220A	Realtek S1220A	Realtek S1220A	Realtek S1220A	Realtek ALC 897, 3jack	Realtek ALC 897, 3jack
	Impedence sense	— / V	V	V	V	—	—
	DAC/AMP	Internal AMP	Internal AMP	Internal AMP	Internal AMP	—	—
	Audio cover	— / V	V	V	V	—	—
	LED illuminated audio jacks	— —	—	—	—	—	—
	Premium capacitors	— / V	V	V	V	V	V
	Audio effects	—	DTS Audio Processing	DTS Audio Processing	DTS X®:Ultra	—	—
Cooling	Onboard fan headers	8	7	7	7	6	6
	Heatsinked M.2	4	3	3	4	1	1
	M.2 backplate	— / 1	—	—	—	—	—
Dynamic OC Switcher		V	V	V	V	—	—
AI Overclocking		V	V	V	V	—	—
AI Cooling II		V	V	V	V	V	V
AI Networking (GameFirst VI)		CreationFirst	—	—	—	—	—
Two-Way AI Noise-Cancelation		V	V	V	V	—	—
DIY Friendly	M.2 Q-Latch	V	V	V	V	V	V
	PCIe Slot Q-Release	— / V	—	—	V	—	—
	PCIe 5.0 SafeSlot	— / V	—	—	—	—	—
	SafeDIMM	V	V	V	V	—	—
	Pre-mount I/O shield	V	V	V	V	—	—
BIOS FlashBack™		V	V	V	V	V	V
Aura Sync		V	V	V	V	V	V
Aura lighting control		3 x ARGB + 1 x RGB	3 x ARGB + 1 x RGB	3 x ARGB + 1 x RGB	3 x ARGB + 1 x RGB	3 x ARGB + 2 x RGB	3 x ARGB + 2 x RGB
CPU power connector		8+8 Pin ProCool	8+8 Pin ProCool	8+8 Pin ProCool	8+8 Pin ProCool	8+4 Pin ProCool	8+4 Pin ProCool
COM port/ header		— / 1	1	1	1	1	1
TPM		Firmware TPM	Firmware TPM	Firmware TPM	TPM header	TPM header	TPM header

图 6-28

6.2.2.2.2.2 Gigabyte X670E 主板

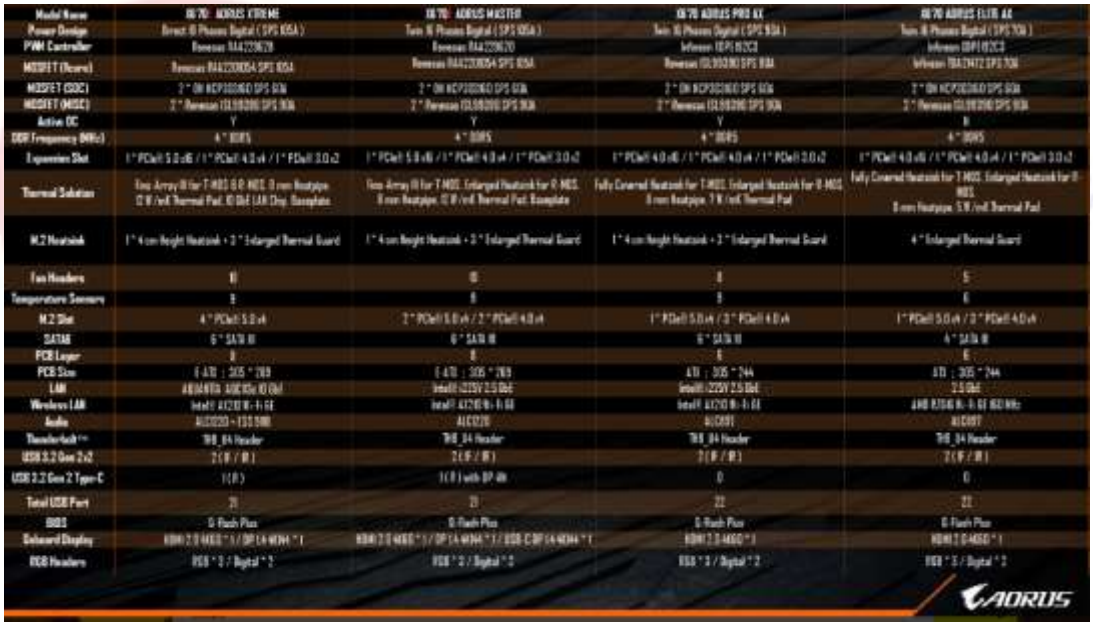


图 6-29


X670E AORUS XTREME
AORUS
X670E AORUS XTREME
X670E
• AMD Socket AM5 : Supports AMD Ryzen™ 7000 Series Processors
• Unparalleled Performance : Direct 18+2+2 Phases Digital VRM Solution
• Dual Channel DDR5 : 4*SMD DIMMs with XMP Memory Module Support
• Next Generation Storage : 4*NVMe SMD PCIe 5.0 x4 M.2 Connectors
• Fins-Array III & M.2 Thermal Guard III : To Ensure VRM Power Stability & 25110 PCIe 5.0 M.2 SSD Performance
• EZ-Latch Plus : SMD PCIe 5.0 x16 Slot & M.2 Connectors with Quick Release & Screwless Design
• Hi-Fi Audio with DTS:X® Ultra : ALC1220 CODEC and Rear ESS SABRE Hi-Fi 9118 DAC with Front ESSential USB DAC Bundled
• Fast Networks : AQUANTIA® 10GbE LAN & Intel® Wi-Fi 6E 802.11ax
• Extended Connectivity : DP, HDMI, USB-C® 10Gbps, Dual USB-C® 20Gbps and Upcoming GIGABYTE USB4 AIC Support
• Q-Flash Plus : Update BIOS Without Installing the CPU, Memory and Graphics Card
AORUS

图 6-30


AORUS
AORUS DESIGNLAB
GIGABYTE
X670E AORUS XTREME
AORUS XTREME SERIES
USB3.2
USB3.2

图 6-31

图 6-32

Index	X670E AORUS XTREME (rev.1.0)	X670E AORUS MASTER (rev.1.0)	X670 AORUS ELITE AX (rev.1.0)	X670 GAMING X AX (rev. 1.0)
CPU	1. AMD Socket AM5, support for : AMD Ryzen™ 7000 Series Processors (Please refer "CPU Support List" for more information.)	1. AMD Socket AM5, support for : AMD Ryzen™ 7000 Series Processors (Please refer "CPU Support List" for more information.)	1. AMD Socket AM5, support for: AMD Ryzen™ 7000 Series Processors (Please refer "CPU Support List" for more information.)	1. AMD Socket AM5, support for: AMD Ryzen™ 7000 Series Processors (Please refer "CPU Support List" for more information.)
Socket	Socket AM5	Socket AM5	Socket AM5	Socket AM5
Chipset	1. AMD X670	1. AMD X670	1. AMD X670	1. AMD X670
	1. Support for DDR5 6600(OC) / 6400(OC) /	1. Support for DDR5 6600(OC) / 6400(OC) /	1. Support for DDR5 6600(OC)/6400(OC)/6200(OC)/6000(OC)/5600(O	1. Support for DDR5

Memory	6200(OC) / 6000(OC) / 5600(OC) / 5200 / 4800 / 4400 MHz memory modules 2. 4 x DDR5 DIMM sockets supporting up to 128 GB (32 GB single DIMM capacity) of systemmemory 3. Dual channel memory architecture 4. Support for non-ECC Un-buffered DIMM 1Rx8/2Rx8/1Rx16 memory modules 5. Support for AMD EXtended Profiles for Overclocking (AMD EXPO™) and Extreme Memory Profile (XMP) memory modules (Please refer "Memory Support List" for more information.)	6200(OC) / 6000(OC) / 5600(OC) / 5200 / 4800 / 4400 MHz memory modules 2. 4 x DDR5 DIMM sockets supporting up to 128 GB (32 GB single DIMM capacity) of systemmemory 3. Dual channel memory architecture 4. Support for non-ECC Un-buffered DIMM 1Rx8/2Rx8/1Rx16 memory modules 5. Support for AMD EXtended Profiles for Overclocking (AMD EXPO™) and Extreme Memory Profile (XMP) memory module (Please refer "Memory Support List" for more information.)	C)/5200/4800/4400 MHz memory modules 2. 4 x DDR5 DIMM sockets supporting up to 128 GB (32 GB single DIMM capacity) of systemmemory 3. Dual channel memory architecture 4. Support for non- ECC Un-buffered DIMM 1Rx8/2Rx8/1Rx16 memory modules 5. Support for AMD EXtended Profiles for Overclocking (AMD EXPO™) and Extreme Memory Profile (XMP) memory modules (Please refer "Memory Support List" for more information.)	6400(OC)/6200(OC)/6000(OC)/5600(OC)/5200/4800/4400 MHz memory modules 2. 4 x DDR5 DIMM sockets supporting up to 128 GB (32 GB single DIMM capacity) of systemmemory 3. Dual channel memory architecture 4. Support for non-ECC Un-buffered DIMM 1Rx8/2Rx8/1Rx16 memory modules 5. Support for AMD EXtended Profiles for Overclocking (AMD EXPO™) and Extreme Memory Profile (XMP) memory modules (Please refer "Memory Support List" for more information.)
	Integrated Graphics	Integrated Graphics Processor: 1. 1 x DisplayPort, supporting a maximum resolution of 3840x2160@144 Hz	Integrated Graphics Processor: 1. 1 x HDMI port, supporting a maximum resolution of 4096x2160@60 Hz * Support for HDMI 2.0 version and HDCP 2.3.	Integrated Graphics Processor: 1. 1 x HDMI port, supporting a maximum resolution of 4096x2160@60 Hz

Onboard Graphics	Processor: 1. 1 x HDMI port, supporting a maximum resolution of 4096x2160@60 Hz * Support for HDMI 2.0 version and HDCP 2.3. 2. 1 x DisplayPort, supporting a maximum resolution of 3840x2160@144 Hz * Support for DisplayPort 1.4 version and HDR. (Graphics specifications may vary depending on CPU support.)	* Support for DisplayPort 1.4 version and HDR. 2. 1 x USB Type- C® port, supporting USB 3.2 Gen 2 and DisplayPort video outputs and a maximum resolution of 3840x2160@144 Hz * Support for DisplayPort 1.4 version and HDR. 3. 1 x HDMI port, supporting a maximum resolution of 4096x2160@60 Hz * Support for HDMI 2.0 version and HDCP 2.3. (Graphics specifications may vary depending on CPU support.)	(Graphics specifications may vary depending on CPU support.)	* Support for HDMI 2.0 version and HDCP 2.3. (Graphics specifications may vary depending on CPU support.)
Audio	1. Realtek® ALC1220-VB CODEC * The front panel line out jack supports DSD audio. 2. ESS ES9118 DAC chip 3. Support for DTS:X® Ultra 4. High Definition Audio 5. 2/4/5.1-channel * You can change the functionality of an audio jack using the audio software. To configure 5.1- channel	1. Realtek® ALC1220-VB CODEC * The back panel line out jack supports DSD audio. 2. Support for DTS:X® Ultra 3. High Definition Audio 4. 2/4/5.1/7.1-channel output * You can change the functionality of an audio jack using the audio software. To configure	1. Realtek® Audio CODEC 2. High Definition Audio 3. 2/4/5.1/7.1-channel * You can change the functionality of an audio jack using the audio software. To configure 7.1- channel audio, access the audio software for audio settings.	1. Realtek® Audio CODEC 2. High Definition Audio 3. 2/4/5.1/7.1-channel * You can change the functionality of an audio jack using the audio software. To configure 7.1- channel audio, access the audio software for audio settings.

	audio, access the audio software for audio settings. 6. Support for S/PDIF Out	7.1-channel audio, access the audio software for audio settings. 5. Support for S/PDIF Out		
LAN	1. Marvel® AQtion AQC113C 10GbE LAN chip (10 Gbps/5 Gbps/2.5 Gbps/1 Gbps/100 Mbps)	1. Intel® 2.5GbE LAN chip (2.5 Gbps/1 Gbps/100 Mbps)	1. Realtek® 2.5GbE LAN chip (2.5 Gbps/1 Gbps/100 Mbps)	1. Realtek® 2.5GbE LAN chip (2.5 Gbps/1 Gbps/100 Mbps)
Expansion Slots	CPU: 1. 1 x PCI Express x16 slot, supporting PCIe 5.0* and running at x16 (PCIEX16) * Actual support may vary by CPU. * The M2B_CPU and M2C_CPU slots share bandwidth with the PCIEX16 slot. When the M2B_CPU or M2C_CPU slot is populated, the PCIEX16 slot operates at up to x8 mode. * For optimum performance, if only one PCI Express graphics card is to be installed, be sure to install it in the PCIEX16 slot.	CPU: 1. 1 x PCI Express x16 slot, supporting PCIe 5.0* and running at x16 (PCIEX16) * Actual support may vary by CPU. * For optimum performance, if only one PCI Express graphics card is to be installed, be sure to install it in the PCIEX16 slot. Chipset: 1. 1 x PCI Express x16 slot, supporting PCIe 4.0 and running at x4 (PCIEX4) 2. 1 x PCI Express x16 slot, supporting PCIe	CPU: 1. 1 x PCI Express x16 slot, supporting PCIe 4.0 and running at x16 (PCIEX16) * For optimum performance, if only one PCI Express graphics card is to be installed, be sure to install it in the PCIEX16 slot. Chipset: 1. 1 x PCI Express x16 slot, supporting PCIe 4.0 and running at x4 (PCIEX4) 2. 1 x PCI Express x16 slot, supporting PCIe 3.0 and running at x2 (PCIEX2) Support for AMD CrossFire™ technology (PCIEX16 and PCIEX4)	CPU: 1. 1 x PCI Express x16 slot, supporting PCIe 4.0 and running at x16 (PCIEX16) * For optimum performance, if only one PCI Express graphics card is to be installed, be sure to install it in the PCIEX16 slot. Chipset: 1. 1 x PCI Express x16 slot, supporting PCIe 4.0 and running at x4 (PCIEX4) 2. 1 x PCI Express x16 slot, supporting PCIe 3.0 and running at x2 (PCIEX2)

	Chipset: 1. 1 x PCI Express x16 slot, supporting PCIe 4.0 and running at x4 (PCIEX4) 2. 1 x PCI Express x16 slot, supporting PCIe 3.0 and running at x2 (PCIEX2) Support for AMD CrossFire™ technology (PCIEX16 and PCIEX4)	3.0 and running at x2 (PCIEX2) * The PCIEX2 slot shares bandwidth with the SATA3 4/5 connectors. The SATA3 4/5 connectors wil become unavailable when a device is insta led in the PCIEX2 slot. Support for AMD CrossFire™ technology (PCIEX16 and PCIEX4)		Support for AMD CrossFire™ technology (PCIEX16 and PCIEX4)
Wireless Communication module	Intel® Wi-Fi 6E AX210 1. WIFI a, b, g, n, ac, ax, supporting 2.4/5/6 GHz carrier frequency bands 2. BLUETOOTH 5.3 3. Support for 11ax 160MHz wireless standard and up to 2.4 Gbps data rate (Actual data rate may vary depending on environment and equipment.)	Intel® Wi-Fi 6E AX210 1. WIFI a, b, g, n, ac, ax, supporting 2.4/5/6 GHz carrier frequency bands 2. BLUETOOTH 5.3 3. Support for 11ax 160MHz wireless standard and up to 2.4 Gbps data rate (Actual data rate may vary depending on environment and equipment.)	AMD Wi-Fi 6E RZ616 (MT7922A22M) 1. WIFI a, b, g, n, ac, ax, supporting 2.4/5/6 GHz carrier frequency bands 2. BLUETOOTH 5.2 3. Support for 11ax 160MHz wireless standard and up to 2.4 Gbps data rate (Actual data rate may vary depending on environment and equipment.)	AMD Wi-Fi 6E RZ616 (MT7922A22M) 1. WIFI a, b, g, n, ac, ax, supporting 2.4/5/6 GHz carrier frequency bands 2. BLUETOOTH 5.2 3. Support for 11ax 160MHz wireless standard and up to 2.4 Gbps data rate (Actual data rate may vary depending on environment and equipment.)
	CPU: 1. 1 x M.2 connector (Socket 3, M key, type 25110/2280 PCIe 5.0* x4/x2 SSD support) (M2A_CPU) * Actual support may vary by	CPU: 1. 1 x M.2 connector (Socket 3, M key, type 25110/2280 PCIe 5.0* x4/x2 SSD support) (M2A_CPU) * Actual support may vary by CPU. 2. 1 x M.2 connector (Socket 3, M key, type 22110/2280 PCIe 5.0* x4/x2 SSD support) (M2B_CPU)	CPU: 1. 1 x M.2 connector (Socket 3, M key, type 25110/2280 PCIe 5.0(Note) x4/x2 SSD support) (M2A_CPU) (Note) Actual support may vary by CPU. 2. 1 x M.2 connector (Socket 3, M key, type 22110/2280 PCIe 4.0 x4/x2 SSD support) (M2B_CPU) Chipset:	CPU: 1. 1 x M.2 connector (Socket 3, M key, type 25110/2280 PCIe 5.0(Note) x4/x2 SSD support) (M2A_CPU) (Note) Actual support may vary by CPU. 2. 1 x M.2 connector (Socket 3, M key, type 22110/2280 PCIe 4.0 x4/x2 SSD support) (M2B_CPU)

Storage Interface	CPU. 2. 3 x M.2 connectors (Socket 3, M key, type 22110/2280 PCIe 5.0* x4/x2 SSD support (M2B_CPU/M2C_CPU/M2D_CPU) * Actual support may vary by CPU. Chipset: 1. 6 x SATA 6Gb/s connectors RAID 0, RAID 1, and RAID 10 support for NVMe SSD storage devices RAID 0, RAID 1, and RAID 10 support for SATA storage devices	* Actual support may vary by CPU. Chipset: 1. 2 x M.2 connectors (Socket 3, M key, type 22110/2280 PCIe 4.0 x4/x2 SSD support) (M2C_SB, M2D_SB) 2. 6 x SATA 6Gb/s connectors RAID 0, RAID 1, and RAID 10 support for NVMe SSD storage devices RAID 0, RAID 1, and RAID 10 support for SATA storage devices	1. 2 x M.2 connectors (Socket 3, M key, type 22110/2280 PCIe 4.0 x4/x2 SSD support) (M2C_SB, M2D_SB) 2. 4 x SATA 6Gb/s connectors RAID 0, RAID 1, and RAID 10 support for NVMe SSD storage devices RAID 0, RAID 1, and RAID 10 support for SATA storage devices	Chipset: 1. 2 x M.2 connectors (Socket 3, M key, type 22110/2280 PCIe 4.0 x4/x2 SSD support) (M2C_SB, M2D_SB) 2. 4 x SATA 6Gb/s connectors RAID 0, RAID 1, and RAID 10 support for NVMe SSD storage devices RAID 0, RAID 1, and RAID 10 support for SATA storage devices
	CPU: 1. 1 x USB Type- C® port on the back panel, with USB 3.2 Gen 2 support 2. 2 x USB 3.2 Gen 2 Type-A ports (red) on the back panel CPU + USB 2.0 Hub: 1. 4 x USB 2.0/1.1 ports on the back panel Chipset:	CPU: 1. 1 x USB Type- C® port on the back panel, with USB 3.2 Gen 2 support 2. 2 x USB 3.2 Gen 2 Type-A ports (red) on the back panel CPU + USB 2.0 Hub: 1. 2 x USB 2.0/1.1 ports on the back panel Chipset: 1. 2 x USB Type- C® ports, with USB 3.2 Gen 2x2 support (1 port on the back panel, 1 port	CPU: 1. 2 x USB 3.2 Gen 2 Type-A ports (red) on the back panel 2. 2 x USB 3.2 Gen 1 ports on the back panel CPU + USB 2.0 Hub: 1. 4 x USB 2.0/1.1 ports on the back panel Chipset:	CPU: 1. 2 x USB 3.2 Gen 2 Type-A ports (red) on the back panel 2. 2 x USB 3.2 Gen 1 ports on the back panel CPU + USB 2.0 Hub: 1. 4 x USB 2.0/1.1 ports on the back panel Chipset:

USB	1. 2 x USB Type- C® ports, with USB 3.2 Gen 2x2 support (1 port on the back panel, 1 port available through the internal USB header) 2. 4 x USB 3.2 Gen 2 Type-A ports (red) on the back panel 3. 4 x USB 3.2 Gen 1 ports available through the internal USB headers 4. 4 x USB 2.0/1.1 ports available through the internal USB headers	available through the internal USB header) 2. 2 x USB 3.2 Gen 2 Type-A ports (red) on the back panel 3. 4 x USB 3.2 Gen 1 ports available through the internal USB headers 4. 4 x USB 2.0/1.1 ports available through the internal USB headers Chipset+USB 3.2 Gen 1 Hub: 1. 4 x USB 3.2 Gen 1 ports on the back panel	1. 2 x USB Type- C® ports, with USB 3.2 Gen 2x2 support (1 port on the back panel, 1 port available through the internal USB header) 2. 8 x USB 3.2 Gen 1 ports (4 ports on the back panel, 4 ports available through the internal USB headers) 3. 4 x USB 2.0/1.1 ports available through the internal USB headers	1. 2 x USB Type- C® ports, with USB 3.2 Gen 2x2 support (1 port on the back panel, 1 port available through the internal USB header) 2. 8 x USB 3.2 Gen 1 ports (4 ports on the back panel, 4 ports available through the internal USB headers) 3. 4 x USB 2.0/1.1 ports available through the internal USB headers

Internal I/O Connectors	1. 1 x 24-pin ATX main power connector 2. 2 x 8-pin ATX 12V power connectors 3. 1 x CPU fan header 4. 1 x water cooling CPU fan header 5. 4 x systemfan headers 6. 4 x system fan/water cooling pump headers 7. 2 x addressable LED strip headers 8. 2 x RGB LED strip headers 9. 1 x CPU cooler LED strip/RGB LED strip header 10. 4 x M.2 Socket 3 connectors 11. 6 x SATA 6Gb/s connectors 12. 1 x front panel header 13. 1 x front panel audio header 14. 1 x USB Type-C® header, with USB 3.2 Gen 2x2 support 15. 2 x USB 3.2 Gen 1 headers 16. 2 x USB 2.0/1.1 headers 17. 1 x noise detection header 18. 1 x THB_U4 add-in card connector 19. 1 x Trusted Platform Module header (For the GC-TPM2.0 SPI/GC-TPM2.0 SPI 2.0 module only) 20. 1 x power button	1. 1 x 24-pin ATX main power connector 2. 2 x 8-pin ATX 12V power connectors 3. 1 x CPU fan header 4. 1 x water cooling CPU fan header 5. 4 x systemfan headers 6. 4 x system fan/water cooling pump headers 7. 2 x addressable LED strip headers 8. 2 x RGB LED strip headers 9. 1 x CPU cooler LED strip/RGB LED strip header 10. 4 x M.2 Socket 3 connectors 11. 6 x SATA 6Gb/s connectors 12. 1 x front panel header 13. 1 x front panel audio header 14. 1 x USB Type-C® header, with USB 3.2 Gen 2x2 support 15. 2 x USB 3.2 Gen 1 headers 16. 2 x USB 2.0/1.1 headers 17. 1 x noise detection header 18. 1 x THB_U4 add-in card connector 19. 1 x Trusted Platform Module header (For the GC-TPM2.0 SPI/GC-TPM2.0 SPI 2.0 module only)	1. 1 x 24-pin ATX main power connector 2. 2 x 8-pin ATX 12V power connectors 3. 1 x CPU fan header 4. 1 x water cooling CPU fan header 5. 3 x systemfan headers 6. 1 x CPU cooler LED strip/RGB LED strip header 7. 2 x addressable LED strip headers 8. 2 x RGB LED strip headers 9. 4 x SATA 6Gb/s connectors 10. 4 x M.2 Socket 3 connectors 11. 1 x front panel header 12. 1 x front panel audio header 13. 1 x USB Type- C® header, with USB 3.2 Gen 2x2 support 14. 2 x USB 3.2 Gen 1 headers 15. 2 x USB 2.0/1.1 headers 16. 1 x THB_U4 add-in card connector 17. 1 x Trusted Platform Module header (For the GC-TPM2.0 SPI/GC-TPM2.0 SPI 2.0 module only) 18. 1 x power button 19. 1 x reset button 20. 1 x Clear CMOS button 21. 1 x reset jumper 22. 1 x Clear CMOS jumper	1. 1 x 24-pin ATX main power connector 2. 2 x 8-pin ATX 12V power connectors 3. 1 x CPU fan header 4. 1 x water cooling CPU fan header 5. 3 x systemfan headers 6. 1 x CPU cooler LED strip/RGB LED strip header 7. 2 x addressable LED strip headers 8. 2 x RGB LED strip headers 9. 4 x SATA 6Gb/s connectors 10. 4 x M.2 Socket 3 connectors 11. 1 x front panel header 12. 1 x front panel audio header 13. 1 x USB Type-C® header, with USB 3.2 Gen 2x2 support 14. 2 x USB 3.2 Gen 1 headers 15. 2 x USB 2.0/1.1 headers 16. 1 x THB_U4 add-in card connector 17. 1 x Trusted Platform Module header (For the GC-TPM2.0 SPI/GC-TPM2.0 SPI 2.0 module only) 18. 1 x power button 19. 1 x reset button 20. 1 x Clear CMOS button 21. 1 x reset jumper

	21. 1 x reset button 22. 2 x temperature sensor headers 23. 1 x reset jumper 24. 1 x Clear CMOS jumper 25. Voltage Measurement Points	20. 1 x power button 21. 1 x reset button 22. 1 x reset jumper 23. 1 x Clear CMOS jumper 24. 2 x temperature sensor headers 25. Voltage Measurement Points		22. 1 x Clear CMOS jumper
Back Panel Connectors	1. 1 x Q-Flash Plus button 2. 1 x Clear CMOS button 3. 2 x SMA antenna connectors (2T2R) 4. 1 x DisplayPort 5. 1 x HDMI 2.0 port 6. 1 x USB Type- C® port, with USB 3.2 Gen 2x2 support 7. 1 x USB Type- C® port, with USB 3.2 Gen 2 support 8. 6 x USB 3.2 Gen 2 Type-A ports (red) 9. 4 x USB 2.0/1.1	1. 1 x Q-Flash Plus button 2. 2 x SMA antenna connectors (2T2R) 3. 1 x DisplayPort 4. 1 x HDMI 2.0 port 5. 1 x USB Type- C® port (DisplayPort), with USB 3.2 Gen 2 support 6. 1 x USB Type- C® port, with USB 3.2 Gen 2x2 support 7. 4 x USB 3.2 Gen 2 Type-A ports (red) 8. 4 x USB 3.2 Gen 1 ports	1. 1 x Q-Flash Plus button 2. 2 x SMA antenna connectors (2T2R) 3. 1 x HDMI 2.0 port 4. 6 x USB 3.2 Gen 1 ports 5. 4 x USB 2.0/1.1 ports 6. 2 x USB 3.2 Gen 2 Type-A ports (red) 7. 1 x USB Type- C® port, with USB 3.2 Gen 2x2 support 8. 1 x RJ-45 port 9. 3 x audio jacks	1. 1 x Q-Flash Plus button 2. 2 x SMA antenna connectors (2T2R) 3. 1 x HDMI 2.0 port 4. 6 x USB 3.2 Gen 1 ports 5. 4 x USB 2.0/1.1 ports 6. 2 x USB 3.2 Gen 2 Type-A ports (red) 7. 1 x USB Type- C® port, with USB 3.2 Gen 2x2 support 8. 1 x RJ-45 port 9. 3 x audio jacks

	ports 10. 1 x RJ-45 port 11. 1 x optical S/PDIF Out connector 12. 2 x audio jacks	9. 2 x USB 2.0/1.1 ports 10. 1 x RJ-45 port 11. 1 x optical S/PDIF Out connector 12. 2 x audio jacks		
Operating System	1. Support for Windows 11 64- bit 2. Support for Windows 10 64- bit	1. Support for Windows 11 64- bit 2. Support for Windows 10 64- bit	1. Support for Windows 11 64- bit 2. Support for Windows 10 64- bit	1. Support for Windows 11 64- bit 2. Support for Windows 10 64- bit
Form Factor	1. E-ATX Form Factor; 30.5cmx 26.9cm	1. E-ATX Form Factor; 30.5cmx 26.9cm	1. ATX Form Factor; 30.5cmx 24.4cm	1. ATX Form Factor; 30.5cmx 24.4cm

6.2.2.2.2.3 **MSI X670E 主板**

图 6-33


WORK SMARTER
PRO X670-P WIFI combines stable functionality and high quality assembly to solve professional workflows.
• Extended Heatsink Design
• 14 + 2 phases / 80A SPS
• Lightning Gen5 M.2 support
• 1x Double-side M.2 Shield Frozr
• Onboard 2.5G LAN with Wi-Fi 6E
• USB Type-C supports up to DP1.4 HBR3

图 6-34

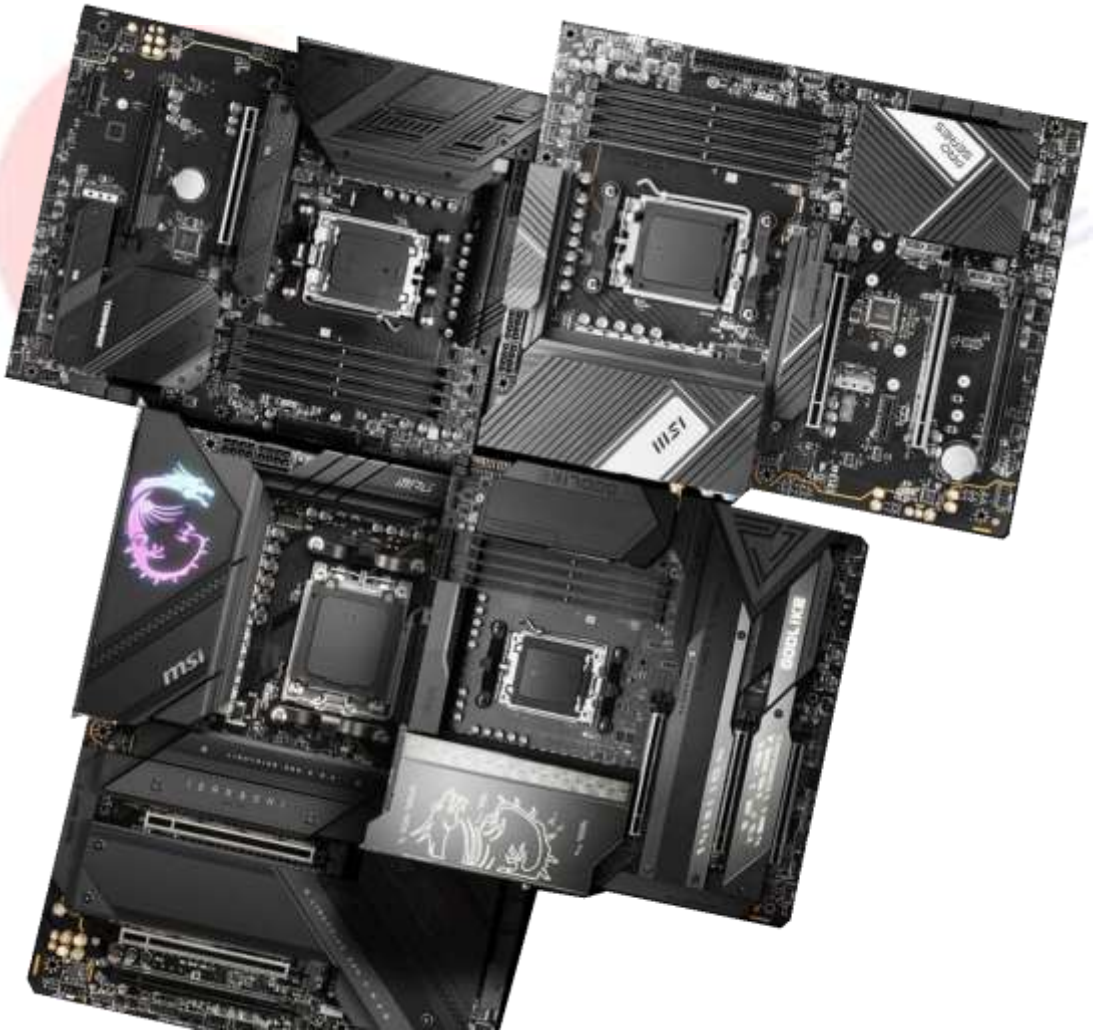


图 6-35

MSI's Initial AM5 Motherboards at a Glance

	MEG X670E Godlike	MEG X670E Ace	MPG X670E Carbon WiFi	Pro X670-P WiFi
VRM	24+2+1 power stages, 105A	22+2+1 power stages, 90A	18+2+1	14+2+1
PCB	10-layer	8-layer	8-layer	8-layer
PCIe 5.0 x16 slots	3x PCIe 5.0 x16 (x16, x8, x4 modes)	3x PCIe 5.0 x16 (x16, x8, x4 modes)	3x PCIe 5.0 x16 (x16, x8, x4 modes)	3x PCIe 4.0 x16 (x16, x4, x2 modes)
PCIe slots	-	-	-	PCIe 3.0 x1
M.2 slots	6 slots, PCIe 5.0 + PCIe 4.0	6 slots, PCIe 5.0 + PCIe 4.0	4 slots PCIe 5.0 + PCIe 4.0	4 slots PCIe 5.0/4.0/3.0
M.2 Expander	M.2 Xpander-Z Gen5 Dual	M.2 Xpander-Z Gen5 Dual	-	-
Ethernet	10GbE + 2.5GbE	10GbE	2.5GbE	2.5GbE
Wi-Fi	Wi-Fi 6E	Wi-Fi 6E	Wi-Fi 6E	Wi-Fi 6E
USB4/TB	-	-	-	-
USB	USB 3.2 Gen 2 + USB 3.2 Gen 2x2	USB 3.2 Gen 2 + USB 3.2 Gen 2x2	USB 3.2 Gen 2 + USB 3.2 Gen 2x2	USB 3.2 Gen 2 + USB 3.2 Gen 2x2
Audio Codec	ALC4082	ALC4082	Realtek ALC4080	Realtek ALC4080
Audio DAC	ESS ES9280AQ	ESS ES9280AQ	?	?
Add-ons	M-Vision Dashboard	-	-	-
Price	$1299.99	$699.99	$479.99	$289.99

图 6-36

All the motherboards are fairly advanced, but the flagship MEG-series come with M.2 Xpander-Z Gen5 Dual SSD adapter cards that add two M.2-2280 slots for SSDs and therefore enable a rather formidable storage system. In addition, the MEG X670E Godlike comes with an M-Vision Dashboard LCD that displays the status of the rig and allows system tuning with touch control.

MSI's AMD X670 and X670E motherboards will be available starting September 27, 2022.

6.2.2.2.2.4 AsRock X670E 主板

图 6-37

ASRock 主板价格相对于 ASUS 和 Gigabyte 较优惠，下面是一个国外的价格参考，注意：这些主板没有包含关税和增值税。

	French Price	French Price without VAT	U.K. Price	U.K. Price without VAT
X670E Taichi Carrara	€693	$578	£647	$618
X670E Taichi	€661	$550	£617	$589
X670E Steel Legend	€373	$343	£385	$368
X670E Pro RS	€412	$311	£290	$331
X670E PG Lightning	€334	$278	£312	$298

图 6-38

Model	X670E Taichi Carrara	X670E Taichi	X670E PG Lightning	X670E Pro RS	X670E Steel Legend

Form Factor	- EATX Form Factor: 12.0-in x 10.5-in, 30.5 cm x 26.7 cm - 8 Layer PCB	- EATX Form Factor: 12.0-in x 10.5-in, 30.5 cm x 26.7 cm - 8 Layer PCB	- ATX Form Factor: 12.0-in x 9.6-in, 30.5 cm x 24.4 cm - 8 Layer PCB	- ATX Form Factor: 12.0-in x 9.6-in, 30.5 cm x 24.4 cm - 8 Layer PCB	- ATX Form Factor: 12.0-in x 9.6-in, 30.5 cm x 24.4 cm - 8 Layer PCB
Unique Feature	**Superb Productivity** - PCIe Gen5 (Graphics, M.2) - 12CM Carrara Edition Cooling FAN - Blazing M.2 Gen5 Fan Heatsink - Dual Channel DDR5 - USB4 Type-C Ports (40 Gb / s) **Rock-Solid Durability** - Ultra Low-Loss PCB - 24+2+1 Power Phase, 105A SPS for VCore+GT with Enlarged Heatsink Armor - Flexible Integrated I / O Shield - Nichicon 12K Black Caps (100% Japan made high quality conductive polymer capacitors) **Ultrafast Gaming** - Killer 2.5G LAN, Killer 802.11ax Wi-Fi 6E Module - Killer DoubleShot™ Pro - ASRock Lightning Gaming Ports **EZ Update**	**Superb Productivity** - PCIe Gen5 (Graphics, M.2) - Blazing M.2 Gen5 Fan Heatsink - Dual Channel DDR5 - USB4 Type-C Ports (40 Gb / s) **Rock-Solid Durability** - Ultra Low-Loss PCB - 24+2+1 Power Phase, 105A SPS with Enlarged Heatsink Armor - Flexible Integrated I / O Shield - Nichicon 12K Black Caps (100% Japan made high quality conductive polymer capacitors) **Ultrafast Gaming** - Killer 2.5G LAN, Killer 802.11ax Wi-Fi 6E Module - Killer DoubleShotTM Pro - ASRock Lightning Gaming Ports **EZ Update** - BIOS Flashback Button	**Superb Productivity** - PCIe Gen5 (Graphics, M.2) - Dual Channel DDR5 **Rock-Solid Durability** - Server-Grade Low-Loss PCB - 14+2+1 Power Phase, 70A SPS for VCore+SOC - Pre-Installed I / O Shield **Ultrafast Connectivity** - 2.5G LAN - ASRock Lightning Gaming Ports **EZ Update** - BIOS Flashback Button - ASRock Auto Driver Installer **EZ Troubleshooter** - ASRock Post Status Checker< / a>	**Superb Productivity** - PCIe Gen5 (Graphics, M.2) - Dual Channel DDR5 **Rock-Solid Durability** - Server-Grade Low-Loss PCB - 14+2+1 Power Phase, 60A SPS with Enlarged Heatsink Armor - Flexible Integrated I / O Shield **Ultrafast Connectivity** - 2.5G LAN, 802.11ax Wi-Fi 6E Module **EZ Update** - BIOS Flashback Button - ASRock Auto Driver Installer **EZ Troubleshooter** - ASRock Post Status Checker< / a>	**Superb Productivity** - PCIe Gen5 (Graphics, M.2) - Dual Channel DDR5 **Rock-Solid Durability** - Server-Grade Low-Loss PCB - 16+2+1 Power Phase, 60A SPS with Enlarged Heatsink Armor - Flexible Integrated I / O Shield - Nichicon 12K Black Caps (100% Japan made high quality conductive polymer capacitors) **Ultrafast Connectivity** - 2.5G LAN, 802.11ax Wi-Fi 6E Module **EZ Update** - BIOS Flashback Button - ASRock Auto Driver Installer **EZ Troubleshooter** - ASRock Post Status Checker< / a>

	- BIOS Flashback Button - ASRock Auto Driver Installer	- ASRock Auto Driver Installer			
CPU	- Supports AMD Socket AM5 Ryzen™ 7000 Series Processors - Supports ASRock Hyper BCLK Engine	- Supports AMD Socket AM5 Ryzen™ 7000 Series Processors - Supports ASRock Hyper BCLK Engine	- Supports AMD Socket AM5 Ryzen™ 7000 Series Processors	- Supports AMD Socket AM5 Ryzen™ 7000 Series Processors	- Supports AMD Socket AM5 Ryzen™ 7000 Series Processors
Chipset	- AMD X670	- AMD X670	- AMD X670	- AMD X670	- AMD X670
Memory	- Dual Channel DDR5 Memory Technology - 4 x DDR5 DIMM Slots - Supports DDR5 ECC / non-ECC, un-buffered memory up to 6600+(OC)* - Max. capacity of system memory: 128GB - Supports Extreme Memory Profile (XMP) and EXTended Profiles for Overclocking (EXPO) memory modules	- Dual Channel DDR5 Memory Technology - 4 x DDR5 DIMM Slots - Supports DDR5 ECC / non-ECC, un-buffered memory up to 6600+(OC)* 1DPC 1R Up to 6600+ MHz (OC), 4800 MHz Natively. 1DPC 2R Up to 6000+ MHz (OC), 4800 MHz Natively. 2DPC 1R Up to 6000+ MHz (OC), 4000 MHz Natively. - Max. capacity of system memory: 128GB - Supports Extreme Memory Profile (XMP) and EXTended Profiles for Overclocking (EXPO) memory modules	- Dual Channel DDR5 Memory Technology - 4 x DDR5 DIMM Slots - Supports DDR5 ECC / non-ECC, un-buffered memory up to 6600+(OC)* - Max. capacity of system memory: 128GB - Supports Extreme Memory Profile (XMP) and EXTended Profiles for Overclocking (EXPO) memory modules	- Dual Channel DDR5 Memory Technology - 4 x DDR5 DIMM Slots - Supports DDR5 ECC / non-ECC, un-buffered memory up to 6600+(OC) - Max. capacity of system memory: 128GB - Supports Extreme Memory Profile (XMP) and EXTended Profiles for Overclocking (EXPO) memory modules	- Dual Channel DDR5 Memory Technology - 4 x DDR5 DIMM Slots - Supports DDR5 ECC / non-ECC, un-buffered memory up to 6600+(OC) - Max. capacity of system memory: 128GB - Supports Extreme Memory Profile (XMP) and EXTended Profiles for Overclocking (EXPO) memory modules
Slots	**CPU:** - 2 x PCIe 5.0 x16 Slots (PCIE1 and PCIE2), support x16 or x8 / x8 modes* **Chipset:** - 1 x Vertical M.2 Socket (Key E), supports type 2230	**CPU:** - 2 x PCIe 5.0 x16 Slots (PCIE1 and PCIE2), support x16 or x8 / x8 modes* **Chipset:** - 1 x Vertical M.2 Socket (Key E), supports type 2230	**CPU:** - 1 x PCIe 5.0 x16 Slot (PCIE1), supports x16 mode* - 1 x PCIe 4.0 x16 Slot (PCIE3), supports x4 mode* **Chipset:**	**CPU:** - 1 x PCIe 5.0 x16 Slot (PCIE1), supports x16 mode* **Chipset:** - 2 x PCIe 4.0 x1 Slots (PCIE2 and PCIE3)*	**CPU:** - 1 x PCIe 5.0 x16 Slot (PCIE1), supports x16 mode* - 1 x PCIe 3.0 x16 Slot (PCIE3), supports x4 mode* **Chipset:**

	WiFi / BT PCIe WiFi module - Supports AMD CrossFire™ - 15μ Gold Contact in VGA PCIe Slot (PCIE1) *Supports NVMe SSD as boot disks	WiFi / BT PCIe WiFi module - Supports AMD CrossFire™ - 15μ Gold Contact in VGA PCIe Slot (PCIE1) *Supports NVMe SSD as boot disks	- 1 x PCIe 4.0 x1 Slot (PCIE2)* - 1 x PCIe 4.0 x16 Slot (PCIE4), supports x1 mode* - 1 x M.2 Socket (Key E), supports type 2230 WiFi / BT PCIe WiFi module - Supports AMD CrossFire™ - 15μ Gold Contact in VGA PCIe Slot (PCIE1) *Supports NVMe SSD as boot disks	- 1 x M.2 Socket (Key E), supports type 2230 WiFi / BT PCIe WiFi module - 15μ Gold Contact in VGA PCIe Slot (PCIE1) *Supports NVMe SSD as boot disks	- 1 x PCIe 3.0 x1 Slot (PCIE2)* - 1 x Vertical M.2 Socket (Key E), supports type 2230 WiFi / BT PCIe WiFi module - Supports AMD CrossFire™ - 15μ Gold Contact in VGA PCIe Slot (PCIE1) *Supports NVMe SSD as boot disks
Graphics	Integrated AMD RDNA™ 2 graphics (Actual support may vary by CPU) - 1 x HDMI 2.1 TMDS / FRL 8G Compatible, supports HDR, HDCP 2.3 and max. resolution up to 4K 120Hz - 2 x USB4, support HDCP 2.3 and max. resolution up to 8K 60Hz* *Only the CPU's embedded graphics can be displayed through USB4 ports. If you want to display to a Type-C monitor, please use CPU models with	Integrated AMD RDNA™ 2 graphics (Actual support may vary by CPU) - 1 x HDMI 2.1 TMDS / FRL 8G Compatible, supports HDR, HDCP 2.3 and max. resolution up to 4K 120Hz - 2 x USB4, support HDCP 2.3 and max. resolution up to 8K 60Hz* *Only the CPU's embedded graphics can be displayed through USB4 ports. If you want to display to a Type-C monitor, please use CPU models with embedded graphics.	Integrated AMD RDNATM 2 graphics (Actual support may vary by CPU) - 1 x HDMI 2.1 TMDS / FRL 8G Compatible, supports HDR, HDCP 2.3 and max. resolution up to 4K 120Hz - 1 x DisplayPort 1.4 with DSC (compressed), supports HDCP 2.3 and max. resolution up to 4K 120Hz	Integrated AMD RDNATM 2 graphics (Actual support may vary by CPU) - 1 x HDMI 2.1 TMDS / FRL 8G Compatible, supports HDR, HDCP 2.3 and max. resolution up to 4K 120Hz - 1 x DisplayPort 1.4 with DSC (compressed), supports HDCP 2.3 and max. resolution up to 4K 120Hz	Integrated AMD RDNATM 2 graphics (Actual support may vary by CPU) - 1 x HDMI 2.1 TMDS / FRL 8G Compatible, supports HDR, HDCP 2.3 and max. resolution up to 4K 120Hz - 1 x DisplayPort 1.4 with DSC (compressed), supports HDCP 2.3 and max. resolution up to 4K 120Hz

	embedded graphics. USB4 graphics output may not be compatible with certain Type-C monitors. Please use graphics card outputs instead.	USB4 graphics output may not be compatible with certain Type-C monitors. Please use graphics card outputs instead.			
Audio	- 5.1 CH HD Audio with Content Protection (Realtek ALC4082 Audio Codec) - WIMA Audio Capacitors (For Front Outputs) - ESS SABRE9218 DAC for Front Panel Audio (130dB SNR) - Individual PCB Layers for R / L Audio Channel - Impedance Sensing on Rear Out port - Nahimic Audio	- 5.1 CH HD Audio with Content Protection (Realtek ALC4082 Audio Codec) - WIMA Audio Capacitors (For Front Outputs) - ESS SABRE9218 DAC for Front Panel Audio (130dB SNR) - Individual PCB Layers for R / L Audio Channel - Impedance Sensing on Rear Out port - Nahimic Audio	- 7.1 CH HD Audio (Realtek ALC897 Audio Codec) - Nahimic Audio	- 7.1 CH HD Audio (Realtek ALC897 Audio Codec) - Nahimic Audio	- 7.1 CH HD Audio with Content Protection (Realtek ALC1220 Audio Codec) - Impedance Sensing on Rear Out port - Individual PCB Layers for R / L Audio Channel - Nahimic Audio
Lan	- 2.5 Gigabit LAN 10 / 100 / 1000 / 2500 Mb / s - Killer E3100G - Supports Killer LAN Software - Supports Killer DoubleShot™ Pro	- 2.5 Gigabit LAN 10 / 100 / 1000 / 2500 Mb / s - Killer E3100G - Supports Killer LAN Software - Supports Killer DoubleShot™ Pro	- 2.5 Gigabit LAN 10 / 100 / 1000 / 2500 Mb / s - Dragon RTL8125BG - Supports Phantom Gaming LAN Software - Smart Auto Adjust Bandwidth Control - Visual User Friendly UI - Visual Network Usage Statistics - Optimized Default Setting for Game, Browser, and Streaming Modes	- 2.5 Gigabit LAN 10 / 100 / 1000 / 2500 Mb / s - Dragon RTL8125BG - Supports Dragon 2.5G LAN Software - Smart Auto Adjust Bandwidth Control - Visual User Friendly UI - Visual Network Usage Statistics - Optimized Default Setting for Game, Browser, and Streaming Modes	**1 x 2.5 Gigabit LAN 10 / 100 / 1000 / 2500 Mb / s (Dragon RTL8125BG)** - Supports Dragon 2.5G LAN Software - Smart Auto Adjust Bandwidth Control - Visual User Friendly UI - Visual Network Usage Statistics - Optimized Default Setting for Game, Browser, and Streaming Modes - User Customized Priority Control

			- User Customized Priority Control	- User Customized Priority Control	**1 x Gigabit LAN 10 / 100 / 1000 Mb / s (Realtek RTL8111)**
WiFi	- 802.11ax Wi-Fi 6E Module - Supports IEEE 802.11a / b / g / n / ac / ax - Supports Dual-Band 2x2 160MHz with extended 6GHz band support* - 2 antennas to support 2 (Transmit) x 2 (Receive) diversity technology - Supports Bluetooth + High speed class II - Supports MU-MIMO - Supports Killer LAN Software - Supports Killer DoubleShot™ Pro *Wi-Fi 6E (6GHz band) will be supported by Microsoft Windows 11. The availability will depend on the different regulation status of each country and region. It will be activated (for supported countries) through Windows Update	- 802.11ax Wi-Fi 6E Module - Supports IEEE 802.11a / b / g / n / ac / ax - Supports Dual-Band 2x2 160MHz with extended 6GHz band support* - 2 antennas to support 2 (Transmit) x 2 (Receive) diversity technology - Supports Bluetooth + High speed class II - Supports MU-MIMO - Supports Killer LAN Software - Supports Killer DoubleShot™ Pro *Wi-Fi 6E (6GHz band) will be supported by Microsoft Windows 11. The availability will depend on the different regulation status of each country and region. It will be activated (for supported countries) through Windows Update and software	- n/a	- 802.11ax Wi-Fi 6E Module - Supports IEEE 802.11a / b / g / n / ac / ax - Supports Dual-Band 2x2 with extended 6GHz band support* - 2 antennas to support 2 (Transmit) x 2 (Receive) diversity technology - Supports Bluetooth + High speed class II - Supports MU-MIMO *Wi-Fi 6E (6GHz band) will be supported by Microsoft Windows 11. The availability will depend on the different regulation status of each country and region. It will be activated (for supported countries) through Windows Update and software updates once available. A 6GHz compatible router is required for 6E functionality.	- 802.11ax Wi-Fi 6E Module - Supports IEEE 802.11a / b / g / n / ac / ax - Supports Dual-Band 2x2 with extended 6GHz band support* - 2 antennas to support 2 (Transmit) x 2 (Receive) diversity technology - Supports Bluetooth + High speed class II - Supports MU-MIMO *Wi-Fi 6E (6GHz band) will be supported by Microsoft Windows 11. The availability will depend on the different regulation status of each country and region. It will be activated (for supported countries) through Windows Update and software updates once available. A 6GHz compatible router is required for 6E functionality.

	and software updates once available. A 6GHz compatible router is required for 6E functionality.	updates once available. A 6GHz compatible router is required for 6E functionality.			
Rear Panel I/O	- 2 x Antenna Ports - 1 x HDMI Port - 1 x Optical SPDIF Out Port - 2 x USB4 Type-C Ports (40 Gb / s)* - 5 x USB 3.2 Gen2 Type-A Ports (10 Gb / s) (ReDriver) (USB32_12 are Lightning Gaming Ports. USB32_3 supports Ultra USB Power.) - 3 x USB 3.2 Gen1 Ports - 1 x RJ-45 LAN Port - 1 x Clear CMOS Button - 1 x BIOS Flashback Button - 1 x Line Out Jack (Gold Audio Jack) - 1 x Microphone Input Jack (Gold Audio Jack) *Supports USB PD 3.0 up to 9V@3A (27W) / 5V@3A (15W) charging	- 2 x Antenna Ports - 1 x HDMI Port - 1 x Optical SPDIF Out Port - 2 x USB4 Type-C Ports (40 Gb / s)* - 5 x USB 3.2 Gen2 Type-A Ports (10 Gb / s) (ReDriver) (USB32_12 are Lightning Gaming Ports. USB32_3 supports Ultra USB Power.) - 3 x USB 3.2 Gen1 Ports - 1 x RJ-45 LAN Port - 1 x Clear CMOS Button - 1 x BIOS Flashback Button - 1 x Line Out Jack (Gold Audio Jack) - 1 x Microphone Input Jack (Gold Audio Jack) *Supports USB PD 3.0 up to 9V@3A (27W) / 5V@3A (15W) charging	- 2 x Antenna Mounting Points - 1 x HDMI Port - 1 x DisplayPort 1.4 - 1 x USB 3.2 Gen2x2 Type-C Port (20 Gb / s) - 1 x USB 3.2 Gen2 Type-A Port (10 Gb / s) - 6 x USB 3.2 Gen1 Ports (USB32_34 are Lightning Gaming Ports) - 4 x USB 2.0 Ports - 1 x RJ-45 LAN Port - 1 x BIOS Flashback Button - HD Audio Jacks: Line in / Front Speaker / Microphone	- 2 x Antenna Ports - 1 x HDMI Port - 1 x DisplayPort 1.4 - 1 x Optical SPDIF Out Port - 1 x USB 3.2 Gen2 Type-A Port (10 Gb / s) (ReDriver) - 1 x USB 3.2 Gen2 Type-C Port (10 Gb / s) (ReDriver) - 4 x USB 3.2 Gen1 Ports - 4 x USB 2.0 Ports - 1 x RJ-45 LAN Port - 1 x BIOS Flashback Button - 1 x Line Out Jack (Gold Audio Jack) - 1 x Microphone Input Jack (Gold Audio Jack)	- 2 x Antenna Ports - 1 x HDMI Port - 1 x DisplayPort 1.4 - 1 x Optical SPDIF Out Port - 1 x USB 3.2 Gen2x2 Type-C Port (20 Gb / s) - 1 x USB 3.2 Gen2 Type-A Port (10 Gb / s) - 6 x USB 3.2 Gen1 Ports - 4 x USB 2.0 Ports - 2 x RJ-45 LAN Ports - 1 x BIOS Flashback Button - 1 x Line Out Jack (Gold Audio Jack) - 1 x Microphone Input Jack (Gold Audio Jack)

Storage	**CPU:** - 1 x Blazing M.2 Socket (M2_1, Key M), supports type 2280 PCIe Gen5x4 (128 Gb / s) mode* **Chipset:** - 1 x Hyper M.2 Socket (M2_2, Key M), supports type 2230 / 2242 / 2260 / 2280 / 22110 SATA3 6.0 Gb / s & PCIe Gen4x4 (64 Gb / s) modes* - 1 x Hyper M.2 Socket (M2_3, Key M), supports type 2280 PCIe Gen4x4 (64 Gb / s) mode* - 1 x Hyper M.2 Socket (M2_4, Key M), supports type 2280 PCIe Gen4x4 (64 Gb / s) mode* - 4 x SATA3 6.0 Gb / s Connectors **ASMedia ASM1061:** - 4 x SATA3 6.0 Gb / s Connectors** *Supports NVMe SSD as boot disks **If M2_2 is occupied by a SATA-type M.2 device, SATA3_A1 will be disabled.	**CPU:** - 1 x Blazing M.2 Socket (M2_1, Key M), supports type 2280 PCIe Gen5x4 (128 Gb / s) mode* **Chipset:** - 1 x Hyper M.2 Socket (M2_2, Key M), supports type 2230 / 2242 / 2260 / 2280 / 22110 SATA3 6.0 Gb / s & PCIe Gen4x4 (64 Gb / s) modes* - 1 x Hyper M.2 Socket (M2_3, Key M), supports type 2280 PCIe Gen4x4 (64 Gb / s) mode* - 1 x Hyper M.2 Socket (M2_4, Key M), supports type 2280 PCIe Gen4x4 (64 Gb / s) mode* - 4 x SATA3 6.0 Gb / s Connectors ASMedia ASM1061: - 4 x SATA3 6.0 Gb / s Connectors** *Supports NVMe SSD as boot disks Supports ASRock U.2 Kit ** If M2_2 is occupied by a SATA-type M.2 device,	CPU: - 1 x Blazing M.2 Socket (M2_1, Key M), supports type 2280 PCIe Gen5x4 (128 Gb / s) mode* Chipset: - 1 x Ultra M.2 Socket (M2_2, Key M), supports type 2280 SATA3 6.0 Gb / s & PCIe Gen3x4 (32 Gb / s) modes* - 1 x M.2 Socket (M2_3, Key M), supports type 2280 PCIe Gen4x2 (32 Gb / s) mode* - 1 x Hyper M.2 Socket (M2_4, Key M), supports type 2260 / 2280 PCIe Gen4x4 (64 Gb / s) mode* - 4 x SATA3 6.0 Gb / s Connectors *Supports NVMe SSD as boot disks Supports ASRock U.2 Kit	CPU: - 1 x Blazing M.2 Socket (M2_1, Key M), supports type 2280 PCIe Gen5x4 (128 Gb / s) mode* - 1 x Hyper M.2 Socket (M2_2, Key M), supports type 2260 / 2280 PCIe Gen4x4 (64 Gb / s) mode* Chipset: - 1 x Hyper M.2 Socket (M2_3, Key M), supports type 2260 / 2280 PCIe Gen4x4 (64 Gb / s) mode* - 1 x M.2 Socket (M2_4, Key M), supports type 2242 / 2260 / 2280 SATA3 6.0 Gb / s & PCIe Gen3x2 (16 Gb / s) modes* - 1 x Hyper M.2 Socket (M2_5, Key M), supports type 2260 / 2280 PCIe Gen4x4 (64 Gb / s) mode* - 6 x SATA3 6.0 Gb / s Connectors *Supports NVMe SSD as boot disks Supports ASRock U.2 Kit	**CPU:** - 1 x Blazing M.2 Socket (M2_1, Key M), supports type 2260 / 2280 PCIe Gen5x4 (128 Gb / s) mode* **Chipset:** - 1 x Hyper M.2 Socket (M2_2, Key M), supports type 2242 / 2260 / 2280 PCIe Gen4x4 (64 Gb / s) mode* - 1 x Hyper M.2 Socket (M2_3, Key M), supports type 2230 / 2242 / 2260 / 2280 PCIe Gen4x4 (64 Gb / s) mode* - 1 x Hyper M.2 Socket (M2_4, Key M), supports type 2260 / 2280 PCIe Gen4x4 (64 Gb / s) mode* - 4 x SATA3 6.0 Gb / s Connectors *Supports NVMe SSD as boot disks Supports ASRock U.2 Kit

		SATA3_A1 will be disabled.			
Connector	- 1 x Power LED and Speaker Header - 1 x RGB LED Header* - 3 x Addressable LED Headers** - 1 x CPU Fan Connector (4-pin)*** - 1 x CPU / Water Pump Fan Connector (4-pin) (Smart Fan Speed Control)**** - 6 x Chassis / Water Pump Fan Connectors (4-pin) (Smart Fan Speed Control)***** - 1 x 24 pin ATX Power Connector (Hi-Density Power Connector) - 2 x 8 pin 12V Power Connectors (Hi-Density Power Connector) - 1 x Front Panel Audio Connector (15μ Gold Audio Connector) - 2 x USB 2.0 Headers (Support 4 USB 2.0 ports) - 2 x USB 3.2 Gen1 Headers (Support 4 USB 3.2 Gen1 ports) - 1 x Front Panel Type C USB 3.2 Gen2x2 Header (20 Gb / s)	- 1 x Power LED and Speaker Header - 1 x RGB LED Header* - 3 x Addressable LED Headers** - 1 x CPU Fan Connector (4-pin)*** - 1 x CPU / Water Pump Fan Connector (4-pin) (Smart Fan Speed Control)**** - 6 x Chassis / Water Pump Fan Connectors (4-pin) (Smart Fan Speed Control)***** - 1 x 24 pin ATX Power Connector (Hi-Density Power Connector) - 2 x 8 pin 12V Power Connectors (Hi-Density Power Connector) - 1 x Front Panel Audio Connector (15μ Gold Audio Connector) - 2 x USB 2.0 Headers (Support 4 USB 2.0 ports) - 2 x USB 3.2 Gen1 Headers (Support 4 USB 3.2 Gen1 ports) - 1 x Front Panel Type C USB 3.2 Gen2x2 Header (20 Gb / s)	- 1 x SPI TPM Header - 1 x Power LED and Speaker Header - 1 x RGB LED Header* - 3 x Addressable LED Headers** - 1 x CPU Fan Connector (4-pin)*** - 1 x CPU / Water Pump Fan Connector (4-pin) (Smart Fan Speed Control)**** - 4 x Chassis / Water Pump Fan Connectors (4-pin) (Smart Fan Speed Control)***** - 1 x 24 pin ATX Power Connector - 1 x 8 pin 12V Power Connector (Hi-Density Power Connector) - 1 x 4 pin 12V Power Connector (Hi-Density Power Connector) - 1 x Front Panel Audio Connector - 1 x Thunderbolt™ AIC Connector (5-pin) (Supports ASRock Thunderbolt™ 4 AIC Card)	- 1 x SPI TPM Header - 1 x Power LED and Speaker Header - 1 x RGB LED Header* - 3 x Addressable LED Headers** - 1 x CPU Fan Connector (4-pin)*** - 1 x CPU / Water Pump Fan Connector (4-pin) (Smart Fan Speed Control)**** - 4 x Chassis / Water Pump Fan Connectors (4-pin) (Smart Fan Speed Control)***** - 1 x 24 pin ATX Power Connector (Hi-Density Power Connector) - 1 x 8 pin 12V Power Connector (Hi-Density Power Connector) - 1 x 4 pin 12V Power Connector (Hi-Density Power Connector) - 1 x Front Panel Audio Connector - 2 x USB 2.0 Headers (Support 4 USB 2.0 ports)	- 1 x SPI TPM Header - 1 x Power LED and Speaker Header - 1 x RGB LED Header* - 3 x Addressable LED Headers** - 1 x CPU Fan Connector (4-pin)*** - 1 x CPU / Water Pump Fan Connector (4-pin) (Smart Fan Speed Control)**** - 4 x Chassis / Water Pump Fan Connectors (4-pin) (Smart Fan Speed Control)***** - 1 x 24 pin ATX Power Connector (Hi-Density Power Connector) - 2 x 8 pin 12V Power Connectors (Hi-Density Power Connector) - 1 x Front Panel Audio Connector - 1 x Thunderbolt™ AIC Connector (5-pin) (Supports ASRock Thunderbolt™ 4 AIC Card) - 2 x USB 2.0 Headers (Support 4 USB 2.0 ports)

	- 1 x Dr. Debug with LED - 1 x Power Button with LED - 1 x Reset Button with LED *Supports in total up to 12V / 3A, 36W LED Strip **Support in total up to 5V / 3A, 15W LED Strip ***CPU_FAN1 supports the fan power up to 1A (12W). ****CPU_FAN2 / WP_3A supports the fan power up to 3A (36W). ***** CHA_FAN1~6 / WP support the fan power up to 2A (24W). CPU_FAN2 / WP_3A and CHA_FAN1~6 / WP can auto detect if 3-pin or 4-pin fan is in use.	- 1 x Dr. Debug with LED - 1 x Power Button with LED - 1 x Reset Button with LED *Supports in total up to 12V / 3A, 36W LED Strip **Support in total up to 5V / 3A, 15W LED Strip ***CPU_FAN1 supports the fan power up to 1A (12W). ****CPU_FAN2 / WP_3A supports the fan power up to 3A (36W). *****CHA_FAN1~6 / WP support the fan power up to 2A (24W). CPU_FAN2 / WP_3A and CHA_FAN1~6 / WP can auto detect if 3-pin or 4-pin fan is in use.	- 2 x USB 2.0 Headers (Support 4 USB 2.0 ports) - 2 x USB 3.2 Gen1 Headers (Support 4 USB 3.2 Gen1 ports) - 1 x Front Panel Type C USB 3.2 Gen2x2 Header (20 Gb / s) (ReDriver) *Supports in total up to 12V / 3A, 36W LED Strip **Support in total up to 5V / 3A, 15W LED Strip ***CPU_FAN1 supports the fan power up to 1A (12W). ****CPU_FAN2 / WP supports the fan power up to 2A (24W). *****CHA_FAN1~4 / WP support the fan power up to 2A (24W). CPU_FAN2 / WP and CHA_FAN1~4 / WP can auto detect if 3-pin or 4-pin fan is in use.	- 2 x USB 3.2 Gen1 Headers (Support 4 USB 3.2 Gen1 ports) - 1 x Front Panel Type C USB 3.2 Gen2x2 Header (20 Gb / s) *Supports in total up to 12V / 3A, 36W LED Strip **Support in total up to 5V / 3A, 15W LED Strip ***CPU_FAN1 supports the fan power up to 1A (12W). ****CPU_FAN2 / WP supports the fan power up to 2A (24W). *****CHA_FAN1~4 / WP support the fan power up to 2A (24W). CPU_FAN2 / WP and CHA_FAN1~4 / WP can auto detect if 3-pin or 4-pin fan is in use.	- 2 x USB 3.2 Gen1 Headers (Support 4 USB 3.2 Gen1 ports) - 1 x Front Panel Type C USB 3.2 Gen2x2 Header (20 Gb / s) (ReDriver) *Supports in total up to 12V / 3A, 36W LED Strip **Support in total up to 5V / 3A, 15W LED Strip ***CPU_FAN1 supports the fan power up to 1A (12W). ****CPU_FAN2 / WP supports the fan power up to 2A (24W). *****CHA_FAN1~4 / WP support the fan power up to 2A (24W). CPU_FAN2 / WP and CHA_FAN1~4 / WP can auto detect if 3-pin or 4-pin fan is in use.
BIOS	- 256Mb AMI UEFI Legal BIOS with GUI support	- 128Mb AMI UEFI Legal BIOS with multilingual GUI support	- 256Mb AMI UEFI Legal BIOS with GUI support	- 256Mb AMI UEFI Legal BIOS with GUI support	- 256Mb AMI UEFI Legal BIOS with GUI support

		- ACPI 6.0 Compliant wake up events - SMBIOS 2.7 Support - CPU, DRAM, PCH 1.0V, VCCIO, VCCSA, VCCST Voltage Multi-adjustment			
Support CD		- Drivers, Utilities, AntiVirus Software (Trial Version), Google Chrome Browser and Toolbar			
Software and UEFI	**Software** - ASRock Motherboard Utility (A-Tuning) - Killer Control Center - ASRock Polychrome SYNC* **UEFI** - ASRock Full HD UEFI - ASRock Auto Driver Installer - ASRock Instant Flash	**Software** - ASRock Motherboard Utility (A-Tuning) - Killer Control Center - ASRock Polychrome SYNC **UEFI** - ASRock Full HD UEFI - ASRock Auto Driver Installer - ASRock Instant Flash	**Software** - ASRock Motherboard Utility (Phantom Gaming Tuning) - ASRock Phantom Gaming LAN Software - ASRock Polychrome SYNC* **UEFI** - ASRock Full HD UEFI - ASRock Auto Driver Installer - ASRock Instant Flash *These utilities can be downloaded from ASRock Live Update & APP Shop.	**Software** - ASRock Motherboard Utility (A-Tuning) - ASRock Dragon 2.5G LAN Software - ASRock Polychrome SYNC* **UEFI** - ASRock Full HD UEFI - ASRock Auto Driver Installer - ASRock Instant Flash *These utilities can be downloaded from ASRock Live Update & APP Shop.	**Software** - ASRock Motherboard Utility (A-Tuning) - ASRock Dragon 2.5G LAN Software - ASRock Polychrome SYNC* **UEFI** - ASRock Full HD UEFI - ASRock Auto Driver Installer - ASRock Instant Flash *These utilities can be downloaded from ASRock Live Update & APP Shop.
Hardware Monitor					
OS	- Microsoft Windows 10 64-bit / 11 64-bit	- Microsoft Windows 10 64-bit / 11 64-bit	- Microsoft Windows 10 64-bit / 11 64-bit	- Microsoft Windows 10 64-bit / 11 64-bit	- Microsoft Windows 10 64-bit / 11 64-bit
Certifications	- FCC, CE - ErP / EuP ready (ErP / EuP ready power supply is required)	- FCC, CE - ErP / EuP ready (ErP / EuP ready power supply is required)	- FCC, CE - ErP / EuP ready (ErP / EuP ready power supply is required)	- FCC, CE - ErP / EuP ready (ErP / EuP ready power supply is required)	- FCC, CE - ErP / EuP ready (ErP / EuP ready power supply is required)
Accessories	- 1 x User Manual - 4 x SATA Data Cables - 1 x Wireless Dongle USB Bracket	- 1 x User Manual - 4 x SATA Data Cables - 1 x Wireless Dongle USB Bracket	- 1 x User Manual - 2 x SATA Data Cables - 4 x Screws for M.2 Sockets	- 1 x User Manual - 2 x SATA Data Cables - 1 x ASRock WiFi 2.4 / 5 / 6 GHz Antenna	- 1 x User Manual - 2 x SATA Data Cables - 1 x ASRock WiFi 2.4 / 5 / 6 GHz Antenna

	- 1 x Blazing M.2 Gen5 Fan Heatsink - 1 x 12CM Carrara Edition Cooling FAN - 1 x ASRock WiFi 2.4 / 5 / 6 GHz Antenna - 4 x Screws for M.2 Sockets - 1 x Standoff for M.2 Socket	- 1 x Blazing M.2 Gen5 Fan Heatsink - 1 x ASRock WiFi 2.4 / 5 / 6 GHz Antenna - 4 x Screws for M.2 Sockets - 1 x Standoff for M.2 Socket	- 1 x Standoff for M.2 Socket	- 4 x Screws for M.2 Sockets - 1 x Graphics Card Holder	- 4 x Screws for M.2 Sockets - 1 x Standoff for M.2 Socket - 1 x Graphics Card Holder

6.2.2.2.2.5 BioStar X670E 主板

图 6-39

图 6-40

6.2.3 PCIe Gen5 SSD

6.2.3.1 数据中心和企业级 Gen5 SSD

目前主流 PCIe Gen5 data center 和 enterprise NVMe SSD 主要是两家国际大厂有限对外销售，即 Kioxia 和 Samsung，但是仅限于大容量且有最低订单金额要求。另外，国产厂商大普和 MemBlaze 的 Gen5 U.2 SSD 也在逐步销售。

6.2.3.1.1 Kioxia Gen5 CD8/CM7 SSD

6.2.3.1.1.1 KIOXIA CD8 产品规格

Kioxia 规格参考如下：

Model Description

- **CD8-R Read Intensive 7.68TB**
- **CD8-R Read intensive 15.3TB**
- **CD8-V Mix Use 6.4TB**
- **CD8-V Mix Use 12.8TB**

- **CM7-R Read Intensive 7.68TB**
- **CM7-R Read Intensive 15.3TB**
- **CM7-V Mix Use 6.4TB**
- **CM7-V Mix Use 12.8TB**

说明：-R 的 DWPD 是 1，-V 的是 3

Drive writes per day (DWPD) is an endurance rating that manufacturers of NAND flash storage provide their customers. Unlike hard disk storage, solid state storage has a limited number of write/erase cycles before the oxide layer within the storage device's floating-gate transistors begins to break down, a process known as flash wear-out. The DWPD rating tells the customer how many times he can expect to overwrite the entire capacity of the solid state drive before it becomes unreliable.

上述 CD8 有 PCIe Gen5 x4 U.2 2.5' SSD；CM7 有 Gen5 x4 E3.S 和 U.2 两种规格。具体产品链接参见下面。

https://americas.Kioxia.com/en-us/business/ssd/data-center-ssd.html

图 6-41

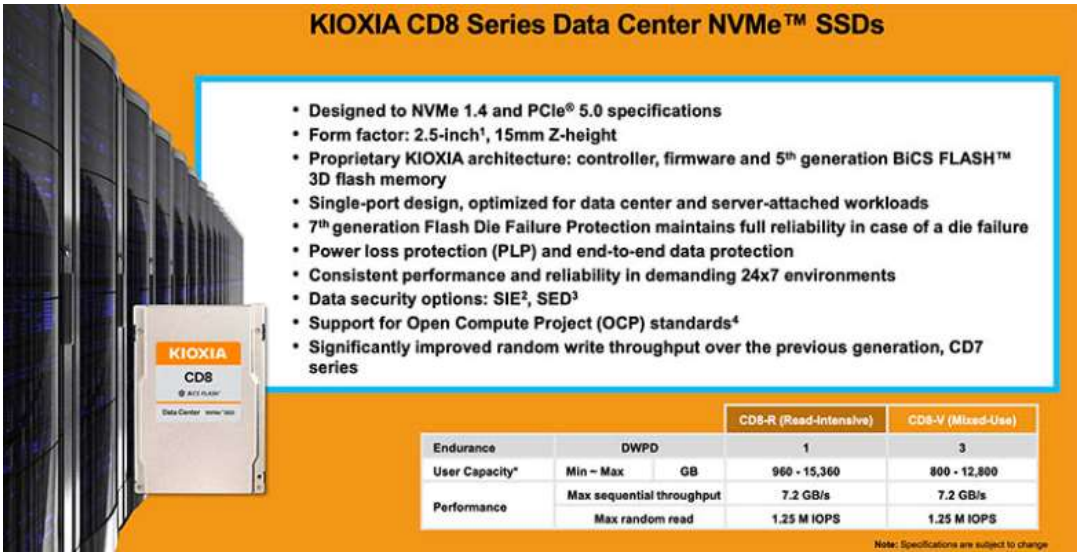


图 6-42

Product Category	Data Center Mixed Use SSD	Data Center Read Intensive SSD
Key Applications	Hyperscale IoT and big data analytics Online transaction processing (OLTP) (transactional and relational databases) Virtualized environments Streaming media and content delivery networks	Hyperscale IoT and big data analytics Online transaction processing (OLTP) (transactional and relational databases) Virtualized environments Streaming media and content delivery networks
DWPD	3	1
Interface	Designed to PCIe® 5.0 Specification, NVMe™ 1.4	Designed to PCIe® 5.0 Specification, NVMe™ 1.4
Maximum Interface Speed	128 GT/s (PCIe® Gen5 x4)	128 GT/s (PCIe® Gen5 x4)
Flash Memory Type	BiCS FLASH™ TLC	BiCS FLASH™ TLC
Storage Capacity (GB)	800 / 1,600 / 3,200 / 6,400 / 12,800	960 / 1,920 / 3,840 / 7,680 / 15,360
Security Option	SIE, SED	SIE, SED
Form Factor	2.5-inch	2.5-inch

6.2.3.1.1.2 KIOXIA CM7 产品规格

https://americas.Kioxia.com/en-us/business/ssd/enterprise-ssd.html

图 6-43

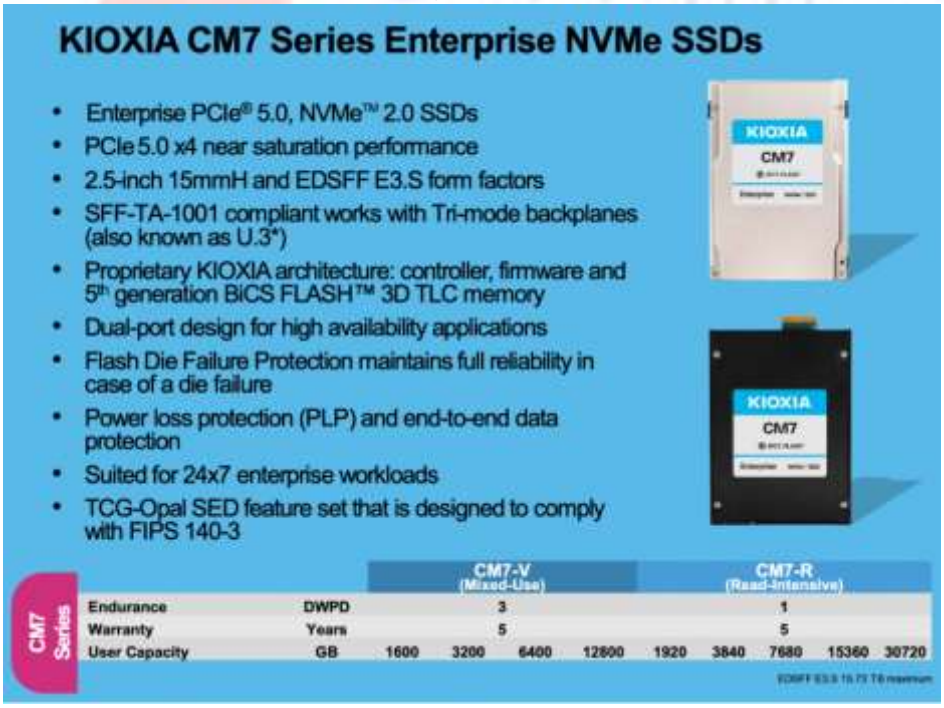


图 6-44

图 6-45

Product Category	Enterprise Mixed Use SSD	Enterprise Read Intensive SSD
Key Applications	Data warehousing Business intelligence Artificial intelligence and machine learning Online transaction processing (OLTP) (transactional and relational databases)	Data warehousing Business intelligence Artificial intelligence and machine learning Online transaction processing (OLTP) (transactional and relational databases)
DWPD	3	1
Interface	For U.2 host: PCIe® 5.0, Designed to NVMe™ 2.0 Specification For U.3 host: PCIe® 4.0, Designed to NVMe™ 2.0 Specification	For U.2 host: PCIe® 5.0, Designed to NVMe™ 2.0 Specification For U.3 host: PCIe® 4.0, Designed to NVMe™ 2.0 Specification
Maximum Interface Speed	For U.2 host: 128 GT/s (PCIe® Gen5 single x4, dual x2) For U.3 host: 64 GT/s (PCIe® Gen4 single x4, dual x2)	For U.2 host: 128 GT/s (PCIe® Gen5 single x4, dual x2) For U.3 host: 64 GT/s (PCIe® Gen4 single x4, dual x2)
Flash Memory Type	BiCS FLASH™ TLC	BiCS FLASH™ TLC

Storage Capacity (GB)	1,600 / 3,200 / 6,400 / 12,800	1,920 / 3,840 / 7,680 / 15,360 / 30,720
Security Option	SIE, SED, FIPS SED	SIE, SED, FIPS SED
Form Factor	2.5-inch	2.5-inch

6.2.3.1.1.3 Kioxia CM7 和 AMD Genoa CPU 测试数据

KIOXIA Demos CM7 PCIe Gen 5 NVMe SSD on AMD EPYC Genoa CPU Platform: 98% Sequential Read & 57% Higher Read Improvement

Hassan Mujtaba • Nov 11, 2022 10:59 AM EST

KIOXIA has showcased the first performance demo of its next-gen CM7 PCIe Gen 5 NVMe SSD on AMD's EPYC Genoa CPU platform.

KIOXIA Delivers Up To Double The Sequential Rate Improvement With CM7 PCIe Gen 5.0 NVMe SSDs, Demo Running On AMD's EPYC Genoa CPUs

Yesterday, AMD introduced its 4th Gen EPYC Genoa CPUs which mark a new beginning for the red team on their latest SP5 platform. The new platform is loaded with features such as PCIe Gen 5.0 support and KIOXIA is taking advantage of that to demonstrate their next-gen CM7 Series PCIe Gen 5.0 NVMe SSDs.

In the demo, KIOXIA used its recently launched CM7 Series SSD which comes in an EDSFF E3.S and 2.5 Inch form factor while featuring support for NVMe 2.0 and PCIe 5.0 specifications. The SSD can be found in two variants:

- **Read Intensive: 1 DWPD / Up To 30.72 TB Capacities**
- **Mixed Use: 3 DWPD / Up To 12.80 TB Capacities**

Server Information	
Model	4th Gen AMD EPYC reference system
CPUs	2x AMD EPYC 9354
No. of CPU Cores	32 per processor
CPU Frequency	3.80 GHz
Total Memory	64 GB DDR-5 DRAM
Memory Frequency	4,800 megatransfers/sec (MT/s)

Operating System Information	
Model	Ubuntu®
Version	22.04 LTS
Kernel	5.150-50-generic

SSD Information		
Model	KIOXIA CM7 Series	KIOXIA CM6 Series
Form Factor	2.5-inch	2.5-inch
Interface	PCIe 5.0 (single 32 GT/s x4, dual 32 GT/s x2)	PCIe 4.0 x4
Capacity	3.84 TB	3.84 TB
Flash Memory	BiCS FLASH™ 3D flash memory	BiCS FLASH™ 3D flash memory
Drive Write(s) per Day	1 (5 years)	1 (5 years)
Active Power	25 W (preliminary)	19 W

Test Software Information	
Model	Flexible I/O® (FIO)
Version	3.28

图 6-46

Both variants come with a dual-port design for High Availability (HA) applications, flash die failure protection, & a Self-Encrypting Drive (SED) supporting TCG Opal and TCG Ruby & a SED option of FIPS 130-3. The drive was tested on an AMD EPYC Genoa test platform comprising an EPYC 9354 32-core CPU in a dual-socket configuration. Other specs for the system included 64 GB of DDR5-4800 memory, Ubuntu 22.04 LTS OS, and a reference Trinite motherboard. For comparison, the older CX6 PCIe Gen 4.0 drive was used against the CX7 PCIe Gen 5.0. Both drives were 3.84 TB in capacity. The CX7 runs at a slightly higher 25W active power whereas the CX6 runs at a 19W active power mode.

CM7 Series highlights include:

- **EDSFF E3.S and 2.5-inch 15mm Z-height form factors (U.2 and U.3)**
- **Designed to the NVMe 2.0 and PCIe 5.0 specifications, and supports SFF-TA-1001/U.3 functionality**
- **SFF-TA-1001 (also known as U.3 capable of Universal Backplane Management enabled systems**
- **Read-intensive (1 DWPD) capacities up to 30.72TB**
- **Mixed-use (3 DWPD) capacities up to 12.80TB**
- **Dual-port design for high-availability applications**
- **Flash Die Failure Protection maintains full reliability in case of a die failure**

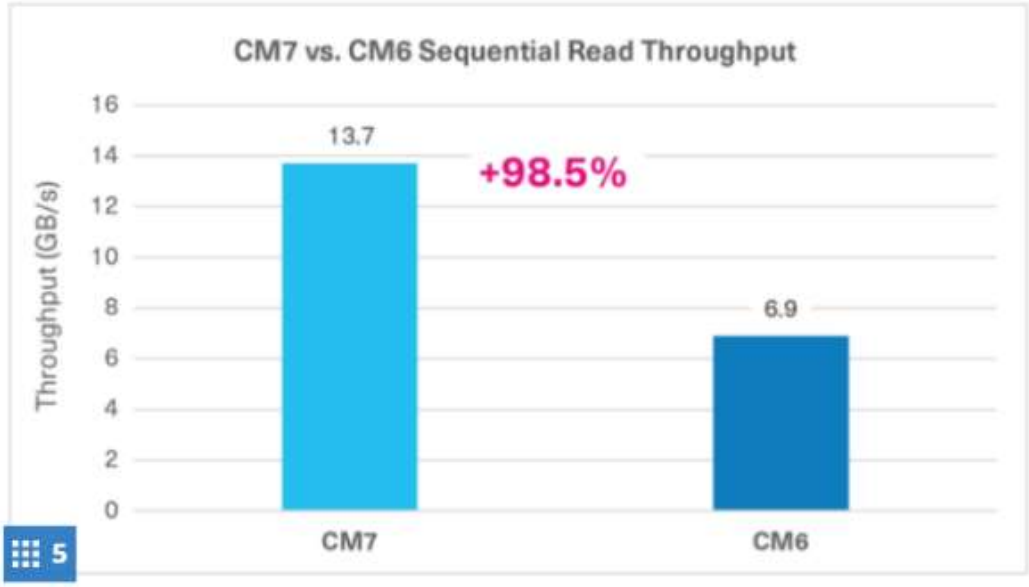
- **Cutting-edge feature support - SRIOV, CMB, multistream writes**

Procedures: The AMD EPYC reference system described above was configured with a 3.84 TB capacity CM7 Series SSD that performed a sequential read and a sequential write workload test that included a 128 kibibyte9 (KiB) block size, a queue depth of 32, and 1 CPU thread, and these results were recorded.

The same AMD EPYC reference system was then equipped with a 3.84 TB capacity CM6 Series SSD that performed a sequential read and a sequential write workload test that included a 128 KiB block size, a queue depth of 32, and 1 CPU thread. No other system changes were made other than changing the SSDs as described here.

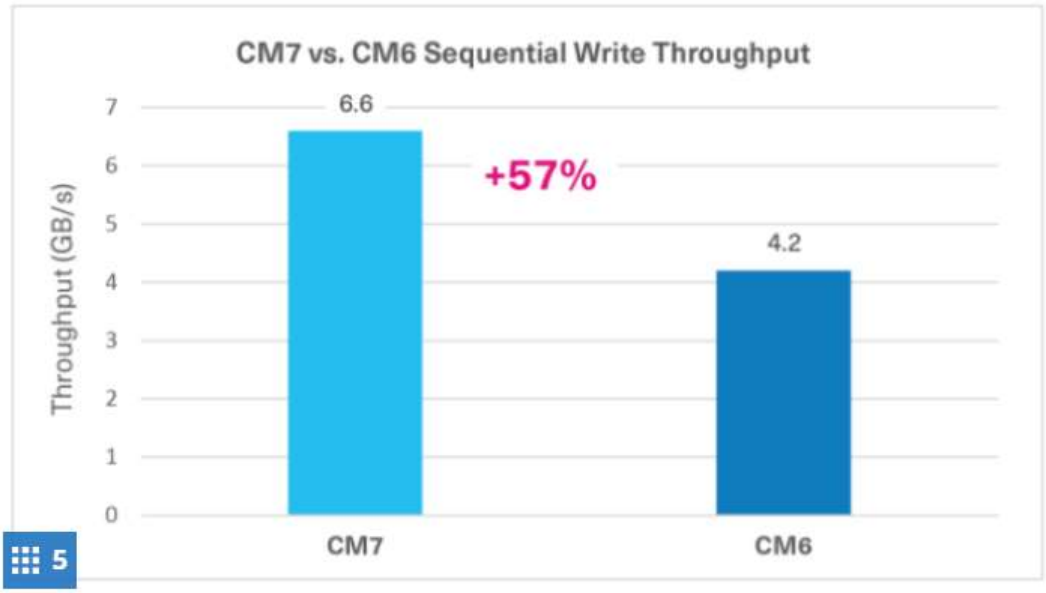
Both sets of test results were compared to determine the performance differences between the PCIe 5.0 CM7 Series and the PCIe 4.0 CM6 Series.

图 6-47



KIOXIA CM7 Series SSD vs CM6 Series SSD in sequential read throughput

图 6-48



KIOXIA CM7 Series SSD vs CM6 Series SSD in sequential write throughput

图 6-49

Starting with the performance comparisons, the KIOXIA CM7 PCIe Gen 5.0 SSD delivered a 98.5% higher Sequential Read Throughput and a 57% higher Sequential Write Throughput versus the older CM6 SSD. The drive posted up to a 13.7 GB/s transfer rate which is close to maxing out the NVMe Gen 5.0 standard that supports up to 14 GB/s. This is definitely one fast NVMe SSD & PCIe Gen 5.0 really proves that we are going to get much higher throughput speeds in the coming gen.

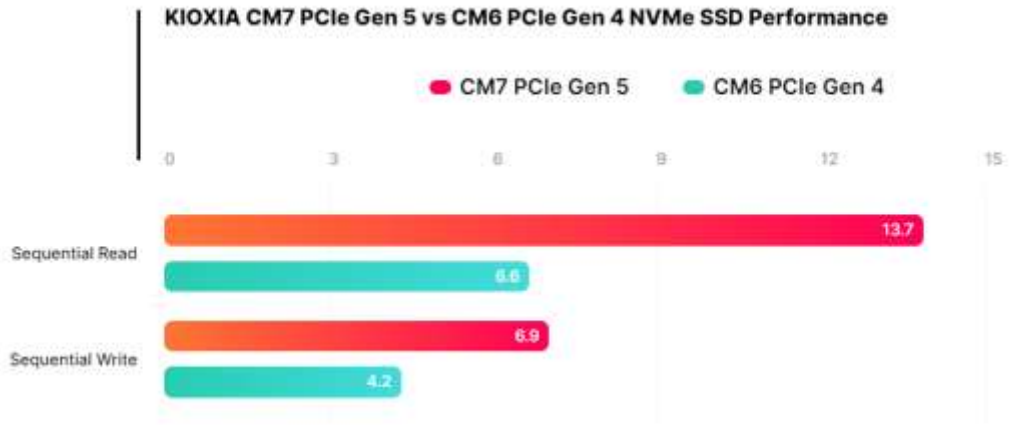


图 6-50

KIOXIA states that their CM7 PCIe Gen 5 NVMe SSDs are currently in pre-production and will be released soon. Server platforms such as AMD's EPYC Genoa and Intel's Xeon Max 'Sapphire Rapids' will take full advantage of these drives.

6.2.3.1.2 Samsung PM1743 PCIe Gen5 NVMe SSD

Samsung PM1743 PCIe Gen5 NVMe SSD 目前仅提供如下规格：

- **7.68TB U.2 2.5"**
- **15.36TB U.2 2.5"**
- **7.68TB E3.S**
- **15.36TB E3.S**

本文档相关章节已经列出 2021/12/31 日 Intel CPU 性能优化工程师使用 SerialCables 公司的 PCIe Gen5 U.2/AIC 转接卡 + PM1743 U.2 SSD + ASUS 使用 12 代酷睿 Gen5 CPU 测试 PM1743 达到 13.8G/S 的介绍，需要视频的可以添加微信联系我们。

图 6-51

下面其它图片分别为 Samsung PM1743 PCIe Gen5 U.2 和 E3.S SSD 以及配合测试使用的 SerialCables Gen5 U.2/AIC，E3.S/AIC adapter，Gen5 switch 卡，Quarch Gen5 U.2 和 E3.S 热插拔/故障注入测试卡，以及第三方 Retimer 卡。


SAMSUNG
PM1743
SAMSUNG
Controller
SAMSUNG
V-NAND
SAMSUNG
V-NAND

图 6-52


SAMSUNG
PM1743

图 6-53

图 6-54

6.2.3.1.2.1 Samsung PM1743 PCIe Gen5 SSD First Take Review

written by Lyle Smith January 13, 2022

The Samsung PM1743 SSD is their first enterprise SSD featuring the new PCIe 5.0 (Gen5) interface, combining the company's advanced sixth-generation V-NAND and a proprietary controller. Showcased at CES 2022, the new PCIe Gen5 interface offers a bandwidth of 32 gigatransfers per second (GT/s), doubling what we saw with PCIe Gen4 and making it a significant step up in performance.

图 6-55

Samsung quotes their new Gen5 SSD with sequential reads speeds of up to 13GB/s, while random read speeds are expected to reach upwards of a massive 2.5 million IOPS. Write

speeds are also projected to reach significantly improved numbers, with sequential and random speeds quoted at 6.6GB/s and 250,000 IOPS, respectively. This equates to roughly 1.9x and 1.7x faster speeds over the company's previous PCIe 4.0-based products (i.e., the previous generation PM1733).

Not only is the performance profile significantly improved, but the new Gen5 interface also allows Samsung to boost its power efficiency. For example, the new PM1743 SSD produces just 608MB/s per watt, which is approximately 30% less power than their previous Gen4 drives. This more effective and efficient power consumption has the potential to noticeably lower server and data center operating costs and reduces carbon footprint, especially in larger deployment use cases.

Samsung PM1743 Form Factors and Sever-Level Features

The PM1743 SSD will be available in capacities of 1.92TB to 15.36TB and in both the standard 2.5-inch size and growlingly popular 3-inch EDSFF (E3.S). Now that Gen5 is almost here, you will see more and more enterprises and technology supporting the ruler form factor, especially since it is capable of doubling storage density in systems compared to the traditional 2.5-inch configurations.

It is also expected that the PM1743 will be the first PCIe Gen5 SSD with dual-port support. This is hugely important, as it will help promote consistent operation and high availability of servers and storage arrays in the event of a failure between one of the port connections.

For security, the new Samsung Gen5 drive will feature an embedded security processor and Root of Trust (RoT). This will help will protect against malicious threats and data forgery, and will enable Secure Boot in server systems via attestation.

Testing the Samung PM1743 SSD

During its development, Samsung collaborated with Intel to optimize the PM1743, so we were excited when Intel dropped by the StorageReview lab with their test system to demonstrate what this new Gen5 drive can do.

The rig they brought was comprised of the ROG MAXIMUS Z690 APEX motherboard, which is equipped with five M.2 slots (including Gen4 and Gen5 slots) and six SATA 6Gb/s ports while supporting 12th Gen Intel Core CPUs and DDR5 RAM. They are outfitted with an Intel 12900K CPU and Corsair Vengeance DDR5 DRAM.

To connect the PM1743 SSDs to Intel's test system, they used two PCIe Gen5 X4 to U.2

vertical adapters from Serial Cables, a company that is known for making impressive products for users at the edge of technology. For example, the adaptor card actually has one of the newer PCIe power connectors available, very similar to what you would see on the new NVIDIA GPUs.

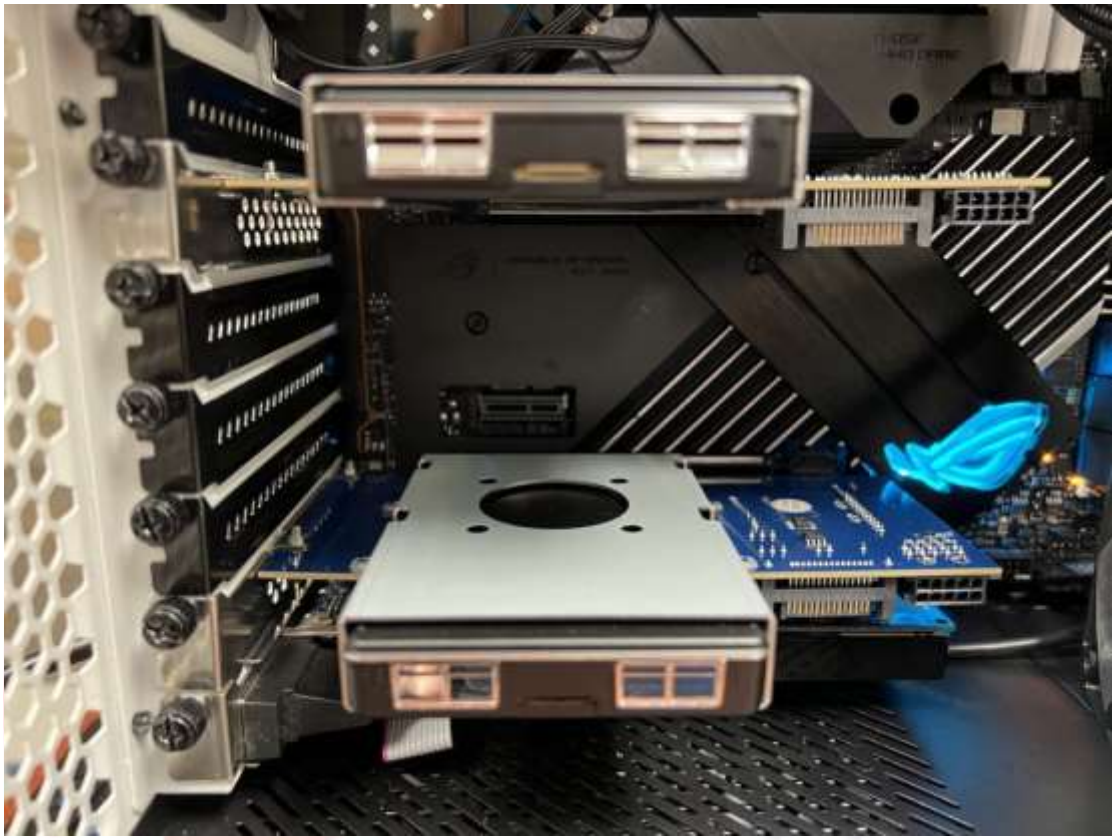


图 6-56

This card is ideal for testing the new Samsung Gen5 drive, as currently, shipping motherboards have a lot of limitations when it comes to Gen5 technology, specifically where Gen5 products can be inserted with the limited supply of Gen5 lanes. In our case, the ROG MAXIMUS Z690 APEX board has Gen5 lanes to two of the standards PCIe slots, which were used in this mini-review.

So, let's fire this rig up and show what a dual PM1743 SSD configuration can do for performance.

Samsung PM1743 SSD Performance

As mentioned, we installed two PM1743 drives via the Serial Cable adaptor cards inside the two designated PCIe Gen5 slots.

In IOmeter, we see it hover around 28GB/s, meaning that a single PM1743 drive is capable of producing a whopping 14GB/s, a little more than the Samsung spec. This is just large

block sequential read, but insane performance nonetheless.

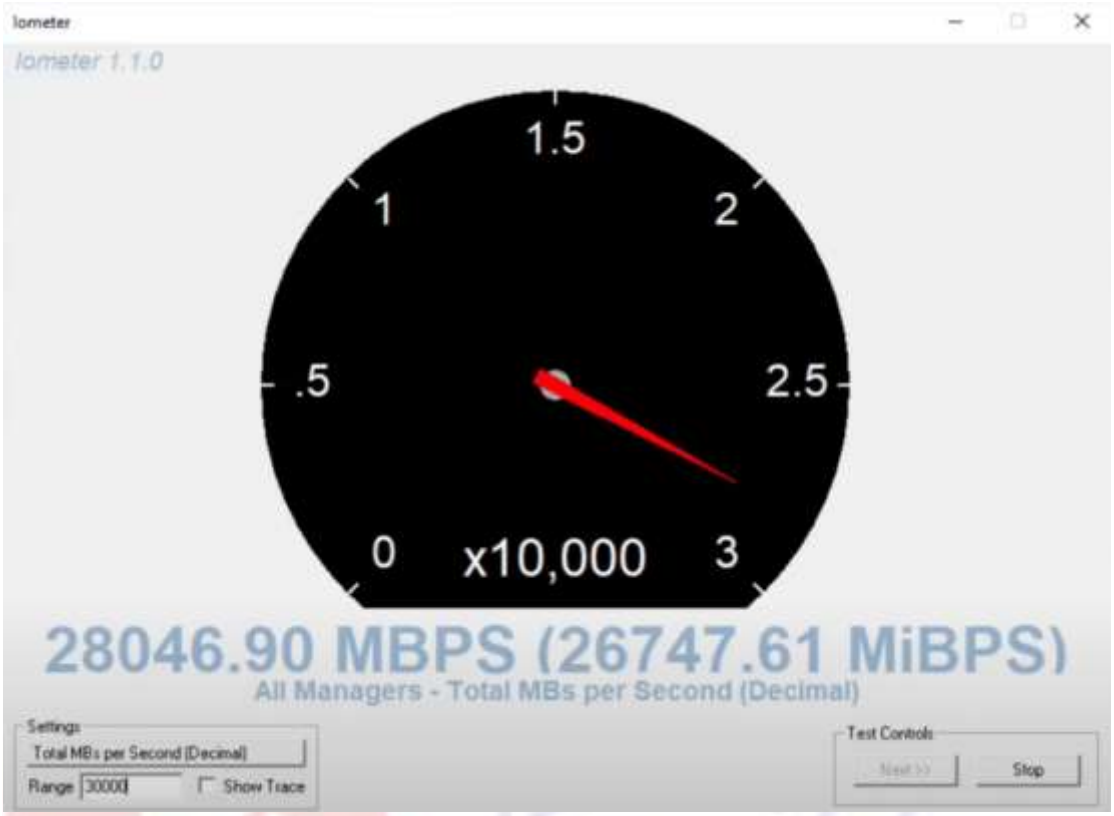


图 6-57

It should be noted that these PM1743 Gen5 SSDs are currently only in preproduction with a very early alpha firmware build. Intel also used a default configuration with no performance tuning whatsoever when putting this testing system together, so you might be able to squeeze even more performance out of it after it has been optimized.

Nonetheless, despite all of this, it's quite evident that the PM1743 Gen5 SSDs are still capable of saturating Gen5 x4 performance and reaching their quoted numbers indicated above. It really demonstrates a lot of confidence from Samsung to allow Intel to showcase these drives at such an early stage.

Conclusion

The new PCIe Gen5 interface is on the cusp of bringing storage performance that is more or less double Gen4, similar to the jump between Gen3 and Gen4. Enterprise SSDs led by the Samsung PM1743 are the closest to achieving this, with client drives expected to hit the market sometime in the mid-year. Though not every workload will require 14GB/s of straight read performance, many applications will still certainly benefit from it.

We can already see the glimmer in the eyes of data scientists everywhere as they contemplate being able to work with massive AI/ML (artificial intelligence/machine learning)

data sets on Gen5 SSDs. In addition, this will also eventually make a huge impact on the array business, as you can just imagine what 24 of these new drives loaded across the front will do.

New challenges will emerge though, as piping data in and out of the box will certainly require more sophisticated networking and possibly even Data Processing Units (DPUs) to maximize the performance benefits. There's also not broad industry support at the moment for any single form factor. SSD vendors like Samsung are being forced to more or less create one of everything to meet the needs of system designers and hyperscalers. That's unsustainable though and the industry needs to eventually coalesce on a couple of options. Either way, we're excited to see the PM1743 platform mature and can't wait to run our full suite of tests against this drive in the coming months.

6.2.3.2 消费类 Gen5 SSD

6.2.3.2.1 Q2~Q3/2023 计划发货的 Gen5 M.2 SSD

Where to buy PCIe Gen 5.0 SSDs: specs, potential release date

Last Updated on September 27, 2022

After much hype, the Ryzen 7000 series of processors has launched, which means support for PCIe Gen 5.0 SSDs is coming sooner rather than later. The chipset and graphics card manufacturer is the first confirmed company to offer support for the latest generation of storage. We're bringing you everything you need to know.

Now that the compatible AM5 motherboards have hit shelves both real and virtual, it shouldn't be long for PCIe Gen 5.0 SSDs to begin shipping. Fortunately, we've seen a little more action on the storage front from some of the biggest names in the industry, so there's a lot to look forward to.

We're bringing you all the information on where to buy PCIe Gen 5 SSDs where you are. What's most encouraging is that many memory makers have been confirmed to have support for AM5 socket's new standard.

This includes Asus, Crucial, MSI, Sabrent, Gigabyte, Seagate, and PNY, among others that we'll detail below. We've rounded up all the best retailers to get yourself one of the first Gen 5 SSDs when the next chipset generation is upon us.

We're going to be seeing unparalleled sequential performance compared to what's been

possible with Gen 4×4. PCIe 5.0 will offer exceptional rates of up to 13,000 MB/s read and 12,000 MB/s write, which is 60% faster than some of the best NVMe SSDs out right now.

Now, the first PCIe 5.0 Gen 5 SSDs will work with current-gen Intel processors, and the upcoming Raptor Lake, too. So things aren't limited to Zen 4. If you're at a crossroads of which processor gen to buy, our Ryzen 7000 vs Raptor Lake feature will bring you up to speed.

6.2.3.2.1.1 PCIe Gen 5.0 SSDs release date

The first PCIe Gen 5.0 SSDs should begin to roll out soon now that the Ryzen 7000 series has dropped. We do not yet know which day-one drives will launch alongside AMD's new processor line, however, we're expecting representation from the usual suspects.

A few Gen 5.0 SSDs have been unveiled ahead of the event and these are the models we're going to be focusing on. These include the Apacer AS2280F5, Corsair MP700, and Zadak TWSG5. All of these models are pushing the boundaries of the new tech straight out of the gate. We've also now got the confirmed existence of the Gigabyte AORUS Gen5 10000 SSD, too.

We'll be bringing you more confirmed models as they roll out ahead of the release. More established brand storage names are on the way, as we've already heard rumors about an upcoming Samsung 980 Pro successor. We'll keep you posted for more developments.

图 6-58 *The Apacer AS2280F5 – One of the first confirmed PCIe Gen 5.0 SSDs to date (Image Credit: Apacer)*

6.2.3.2.1.2 PCIe Gen 5.0 SSDs potential prices

No official pricing has been stated for the PCIe Gen 5.0 SSDs we're featuring, however, it's likely to remain largely in line with top-end Gen 4.0. For instance, the Corsair MP600 Pro LPX, the brand's current flagship, carried an MSRP of $370 at release. Based on this, we can estimate that many high-performing models will hover around the $350 – $400 mark depending on the manufacturer. We'll be bringing you more as we learn.

图 6-59 The technical specifications of the upcoming Phison E26 Gen 5.0 controller. (Image Credit: Phison)

6.2.3.2.1.3 How fast will PCIe Gen 5.0 SSDs be?

PCIe Gen 5.0 SSDs will have read and write speeds of up to 13,000 / 12,000 MB/s respectively. As a frame of reference, the current standard for NVMe drives, Gen 4.0, maxes out at around 8,000 MB/s. This is an increase of around 60% and consistent with the jump we saw between the previous generations. This is achieved through NVMe 2.0, significantly faster than NVMe 1.4.

6.2.3.2.1.4 Will PCIe Gen 5.0 SSDs require thicker heatsinks?

It's been confirmed by a representative from Phison that PCIe 5.0 SSDs could run significantly hotter than previous models. Because of this, thicker heatsinks could be needed than what we typically see accompanying motherboards or aftermarket solutions at the moment. Active cooling could be in question, as suggested by the ElecGear M.2 2280 SSD Cooler.

It's worth noting that Gen 4.0 NVMe SSDs ran hotter than their older Gen 3.0 equivalents. In a report published by WCCFTech, the company claimed that while it heavily recommended a heatsink for Gen 4.0 drives, a heatsink for the upcoming generation is

essential to avoid thermal throttling or damage.

6.2.3.2.1.5 What controllers will Gen 5.0 SSDs use?

Arguably the biggest name in upcoming Gen 5.0 SSDs is the Phison E26. This company is responsible for some of the fastest NVMe drives on the market with its existing E18 controller, so the anticipation for 5×4 compatibility shouldn't be overlooked. Also of note are the Silicon Motion MonTitan, Microchip Flashtec NVMe 4016 PM8667, and Marvell Bravera SC5 MV-SS1333.

These are just a few controllers that we know about, with more sure to come powering the newest line of storage drives. We're confident that manufacturers such as Western Digital and SK Hynix will have their own developed in-house controller, too.

6.2.3.2.1.6 Are NVMe SSDs better than SATA?

You can expect your average SATA SSD to deliver around 600 MB/s of sequential reads and writes. In contrast to this, the upcoming PCIe Gen 5 SSDs, which are built upon NVMe 2.0, can deliver up to 20x the performance. What's more, NVMe SSDs are significantly smaller and can be installed directly onto your motherboard.

The major advantage that SATA models have is price-per-gigabyte. Gen 5.0 SSDs will be expensive, especially in higher configurations such as 2TB and 4TB.

6.2.3.2.2 Phison E26 SSD 测试报告_2023

Phison E26 SSD Preview: PCIe 5 Storage Breaks Out For 2023

by Marco Chiappetta, Zak Killian — Monday, January 02, 2023, 08:50 AM EDT

6.2.3.2.2.1 Phison E26 SSD Preview: Next-Gen PCIe 5 Storage Performance Explored

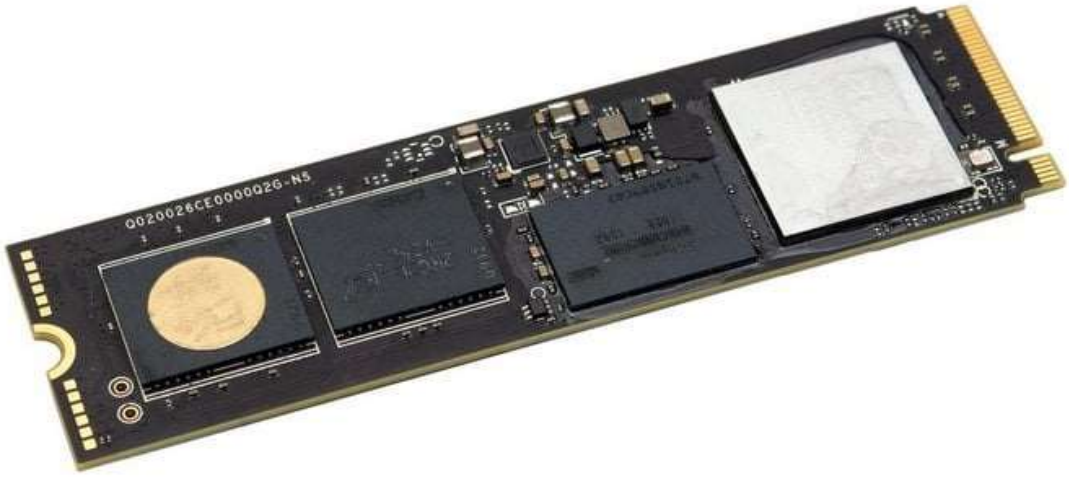


图 6-60

Phison E26 PCIe 5 NVMe SSD Preview: Coming Soon To An SSD Near You

A new controller, fresh firmware, and an updated design paradigm results in one of the fastest SSDs we've ever tested.

HOT

- Ultra High Sequential Transfers
- Low Latency
- Tuned For DirectStorage

NOT

- Didn't Always Lead With Random 4K Transfers
- Needs Faster NAND For Max Performance

Both AMD and Intel's desktop platforms now boast PCI Express 5.0 support—at least, as long as you get a premium motherboard. However, there simply aren't many PCIe 5.0 devices in the consumer arena to make use of said support. NVIDIA, AMD, and Intel's graphics cards are all based on PCIe 4.0, and so are the vast majority of solid-state storage devices out there. That's good news for folks that didn't shell out for a high-end mainboard, but what about those of us who did?

Naturally, PCIe 5.0 SSDs are on the way, and they'll be here sooner than you think. We have one of the very first of such drives here in the labs, thanks to the folks at Phison. The PS5026-E26 drive that we have is not a retail product, but a reference design for other

manufacturers to use when building their own SSDs. It boasts a brand-new E26 SSD controller with PCIe 5.0 x4 support, as well as NVMe 2.0 compliance and a fat DRAM cache. Let's take a closer look at the specifications as supplied by Phison.

Phison E26 PCIe 5 NVMe SSD Specifications And Features

图 6-61

To be honest, the specifications above are actually for the E26 controller, not this specific SSD. In fact, this particular drive is a bit behind the maximum possible performance of the E26 controller as it clocks its flash memory below its peak rate: at just 1600 MT/s, instead of the full 2400 MT/s. Phison says it isn't sure when drives with the full-speed flash will be available because it's up to Micron to actually make the hot-clocked NAND in the first place.

图 6-62

On the top side of the SSD, we can see the E26 controller itself, as well as two of the four Micron 4-terabit flash packages and the 4GB SK Hynix LPDDR4 DRAM. The Micron flash is 232-layer "B58R" TLC, while the LPDDR4 package is rated for 4266 MT/s per SK Hynix's catalog, but apparently running at 3200 MT/s when hooked up to the Phison controller.

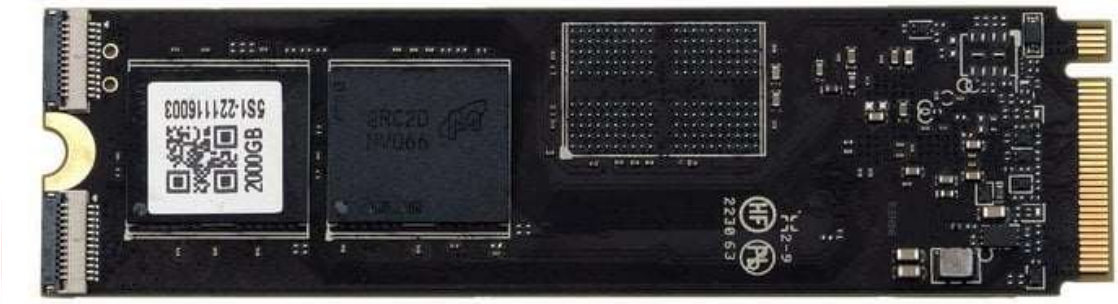


图 6-63

Meanwhile, on the bottom side of the SSD, there's two more of those Micron flash packages as well as some bare pads for another LPDDR4 IC. Presumably this device could accept denser flash packages, and you'd want more cache to maintain the same flash-to-DRAM ratio. There are also some peculiar interfaces on the end; we suspect those are diagnostic interfaces of some sort, as this is a pre-release product.

图 6-64

Lastly, this is the cooling device that Phison shipped along with the E26 reference SSD. Phison notes that the cooler is absolutely not required, and we found that the drive does indeed stay fairly cool most of the time, but such a beefy heatsink-and-fan will help this PCIe 5.0 SSD remain cool under sustained transfer loads like it's designed for.

On that topic, there's another part of this SSD that doesn't show up in photos, and that's the secret sauce in the firmware that Phison calls "IO+ Technology". This isn't exclusive to this SSD—drives sporting the company's E18 controller can also take advantage of it. IO+ Technology is a "cluster of firmware optimizations" that Phison says are targeted at increasing performance in sustained workloads.

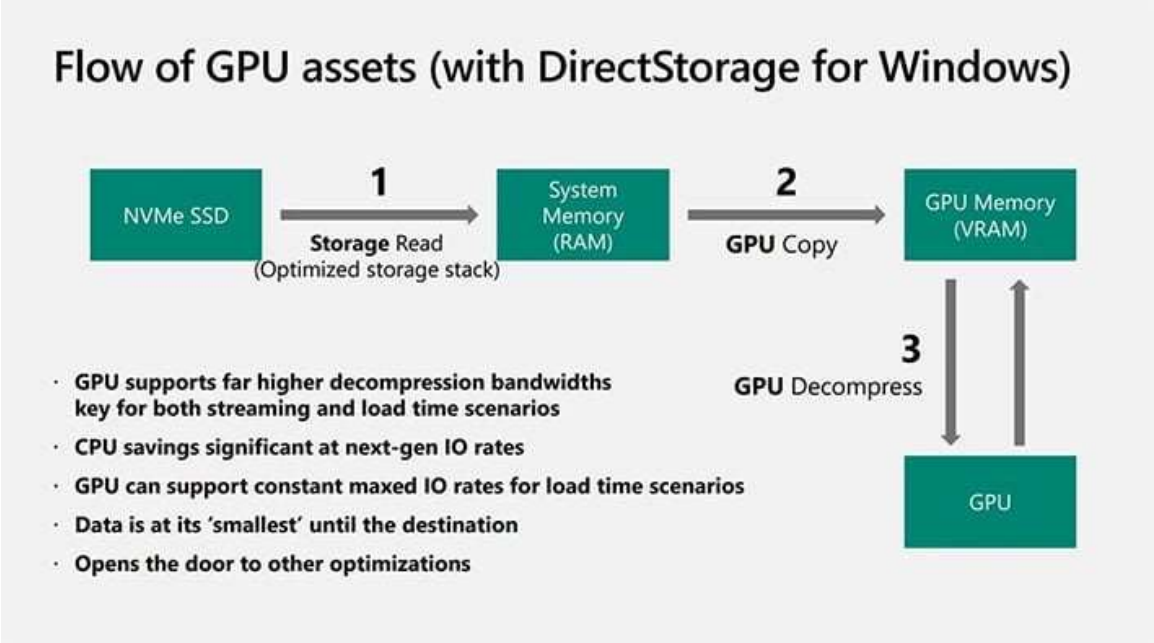


图 6-65 *Diagram from Microsoft showing the eventual goal of DirectStorage.*

Why focus on sustained workloads when most consumer workloads are more bursty? Because of the advent of DirectStorage. Microsoft's new I/O API, primarily targeted at games, is intended to allow game developers to use the system SSD in a similar way to how we used optical discs on 6th- and 7th-generation game consoles. Essentially, once DirectStorage 1.1 is enabled, games will be able to continually stream data from the SSD into memory.

Taking advantage of this feature will require an SSD that can maintain a high transfer rate of around 2.5 GB/second. To that end, every company's been shipping firmware optimized for such an unusual workload, and Phison says its own tweaks and fixes resulted in a 34% performance increase in a 4K sustained workload with a 70/30 split between reads and writes. In a still-synthetic but slightly-more-realistic DirectStorage test, it saw a 23% performance uplift on the E18.

图 6-66

We did some quick tests with a simulated DirectStorage workload on the E26 and the results are presented here. Effective bandwidth peaked at just shy of 15GB/s and 5.5GB of data loaded in about 1/3 of a second. Note, this tests was performed with a GeForce RTX 3080. With a higher-performing GPU, the GPU decompression data rate would have increased and effected the results.

Phison E26 PCIe 5 NVMe SSD Benchmarks

Under each test condition, the SSDs showcased here were installed as secondary volumes in our testbed, with a separate drive used for the OS and benchmark installations. Our testbed's motherboard was updated with the latest BIOS available at the time of publication and Windows 11 was fully updated. Windows firewall, automatic updates, and screen savers were all disabled before testing and Focus Assist was enabled to prevent any interruptions.

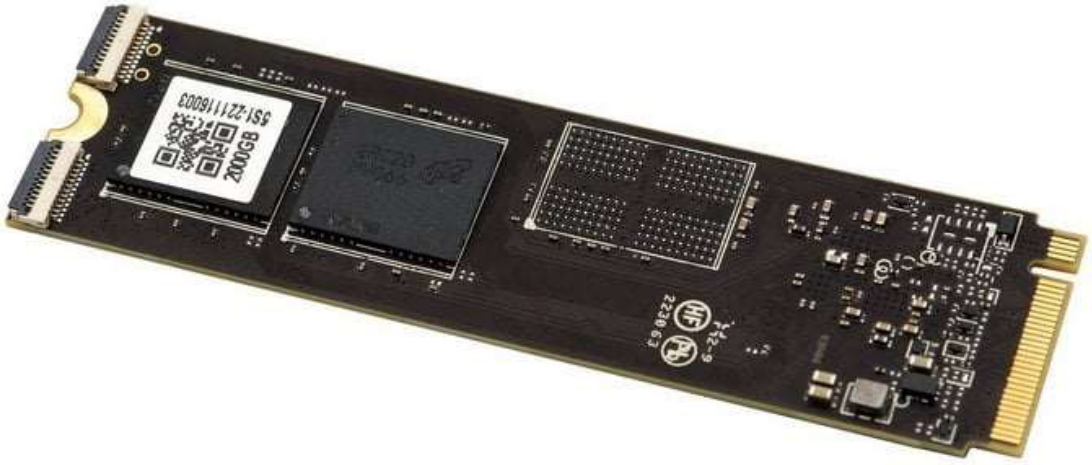


图 6-67

In all test runs, we rebooted the system, ensured all temp and prefetch data was purged, and waited several minutes for drive activity to settle and for the system to reach an idle state before invoking a test. All of the drives here have also been updated to their latest firmware as of press time. Where applicable, we would also typically use any proprietary NVMe drivers available from a given manufacturer, but all of the drives featured here used the Microsoft NVMe driver included with Windows 11.

HotHardware's Test System:

Processor:

AMD Ryzen 9 7950X

Motherboard:

Asus ROG CrossHair X670E Hero

Video Card:

GeForce RTX 3080

Memory:

32GB G.SKILL DDR5-5200

Storage:

ADATA XPG GAMMIX S70 Blade (OS Drive)

ADATA XPG GAMMIX S70 (2TB)

OS:

Windows 11 Pro x64

Chipset Drivers:

AMD v4.11.15.342

Benchmarks:

IOMeter 1.1

HD Tune v5.75

ATTO v4.01.01f

AS SSD

SiSoftware SANDRA

CrystalDiskMark v8.0.4 x64

Final Fantasy XIV: Endwalker

PCMark 10 Quick Storage Bench

Phison E18 B47R (2TB)
Samsung SSD 980 Pro (2TB)
Samsung SSD 990 Pro (2TB)
Solidigm P44 Pro (2TB)
Phison E26 Prototype (2TB)

3DMark Storage Tests

IOMeter Benchmarks

IOMeter is a well-respected industry standard benchmark. However, despite our results with IOMeter scaling as expected, it is debatable as to whether or not certain access patterns actually provide a valid example of real-world performance. The access patterns we tested may not reflect your particular workloads, for example. That said, we do think IOMeter is a reliable gauge for relative throughput, latency, and bandwidth with a given storage solution. In addition, there are certain highly-strenuous workloads you can place on a drive with IOMeter that you simply can't with most other storage benchmark tools.

In the following tables, we're showing two sets of access patterns; a custom Workstation pattern, with an 8K transfer size, consisting of 80% reads (20% writes) and 80% random (20% sequential) access and a 4K access pattern with a 4K transfer size, comprised of 67% reads (33% writes) and 100% random access. Queue depths from 1 to 16 were tested.

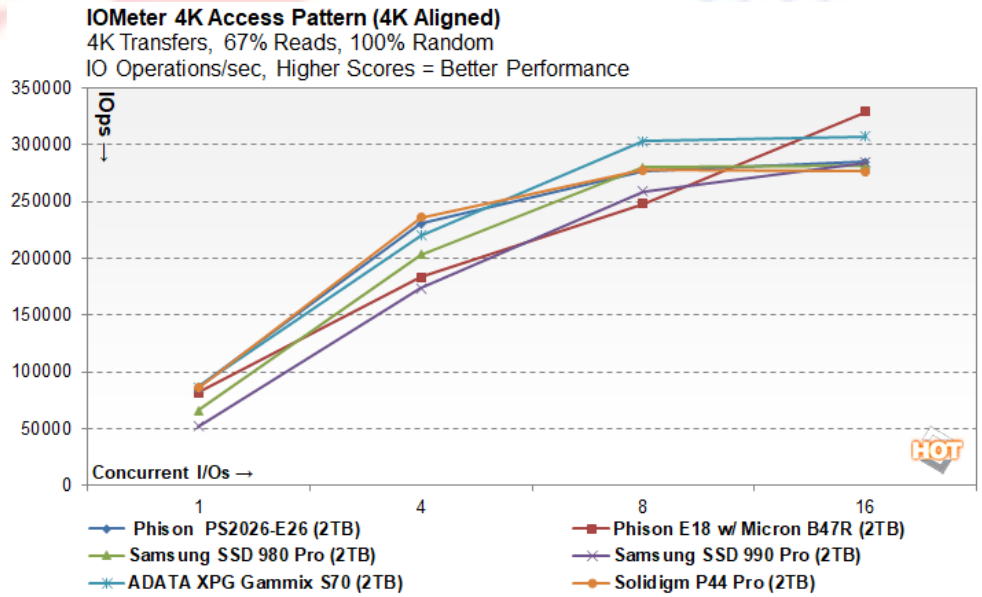


图 6-68

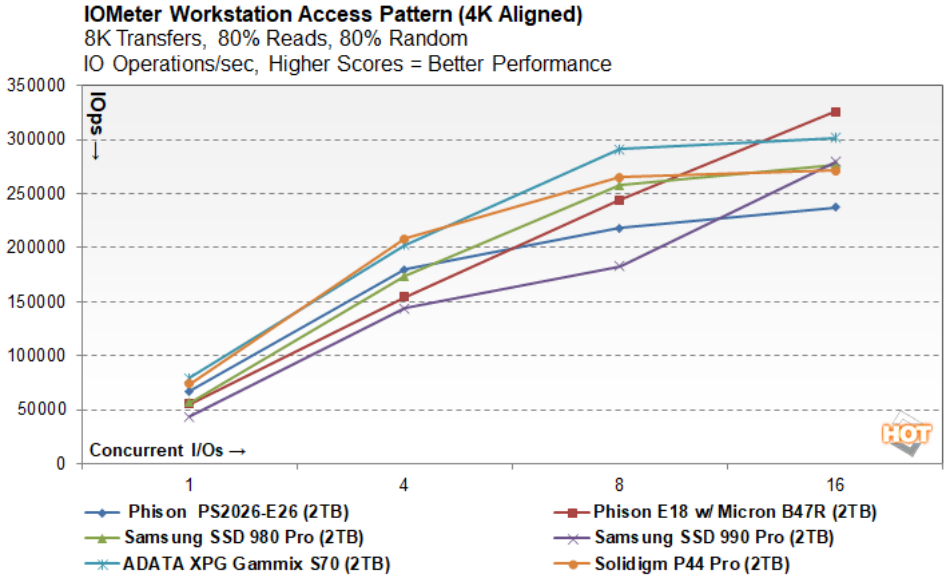


图 6-69

The main advantage this Phison reference platform has over the other SSDs in this benchmark is its PCIe 5.0 interface. That really doesn't help it in this 8K transfer workstation benchmark, where it ultimately ends up trailing at the higher queue depth. Still, it's important to keep in mind that these are some of the very fastest SSDs on the market, and none of the drives here actually perform "poorly" in an absolute sense. Notably, in the all-important QD1 and QD4 tests, it lands right in the middle of this esteemed pack.

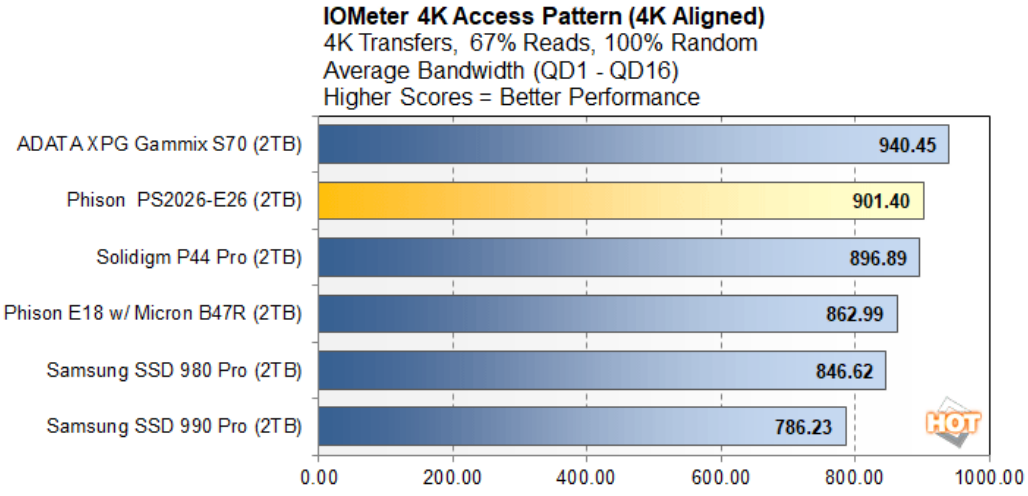


图 6-70

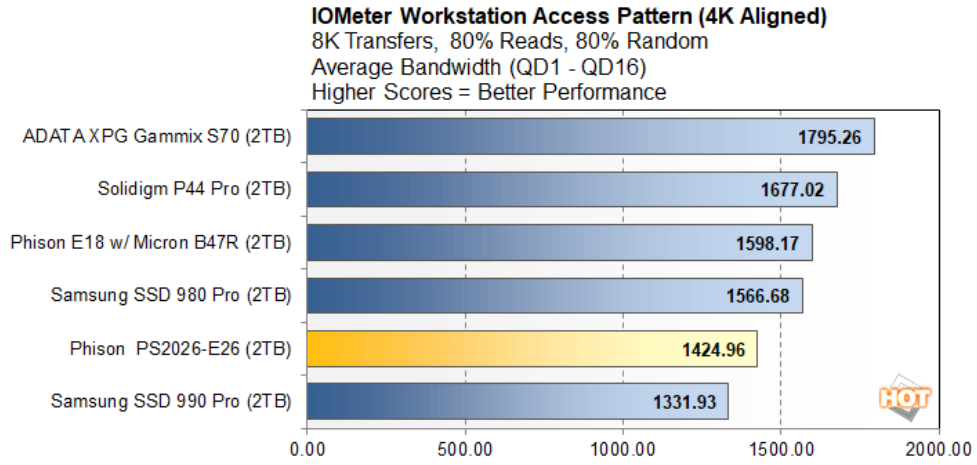


图 6-71

These numbers represent the average bandwidth for the drives we tested with both access patterns, across every queue depth. The ADATA XPG Gammix S70 drive tops the charts in these tests, but wait until our real-world application tests to see why these results aren't as indicative as you might think. We suspect the IO+ Technology firmware is tightly-optimized for 4K random accesses at lower queue depths, and that may be what's causing it some stress in the 8K Workstation pattern tests—though, again, we have to stress that this is not a "bad" score by any means.

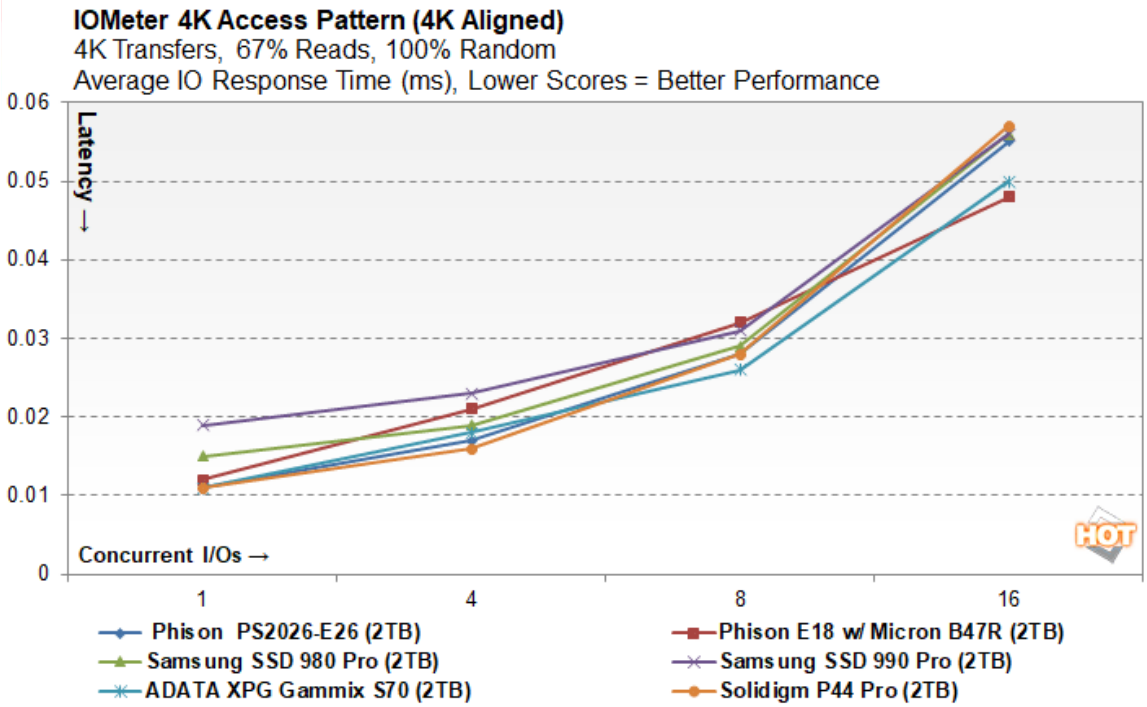


图 6-72

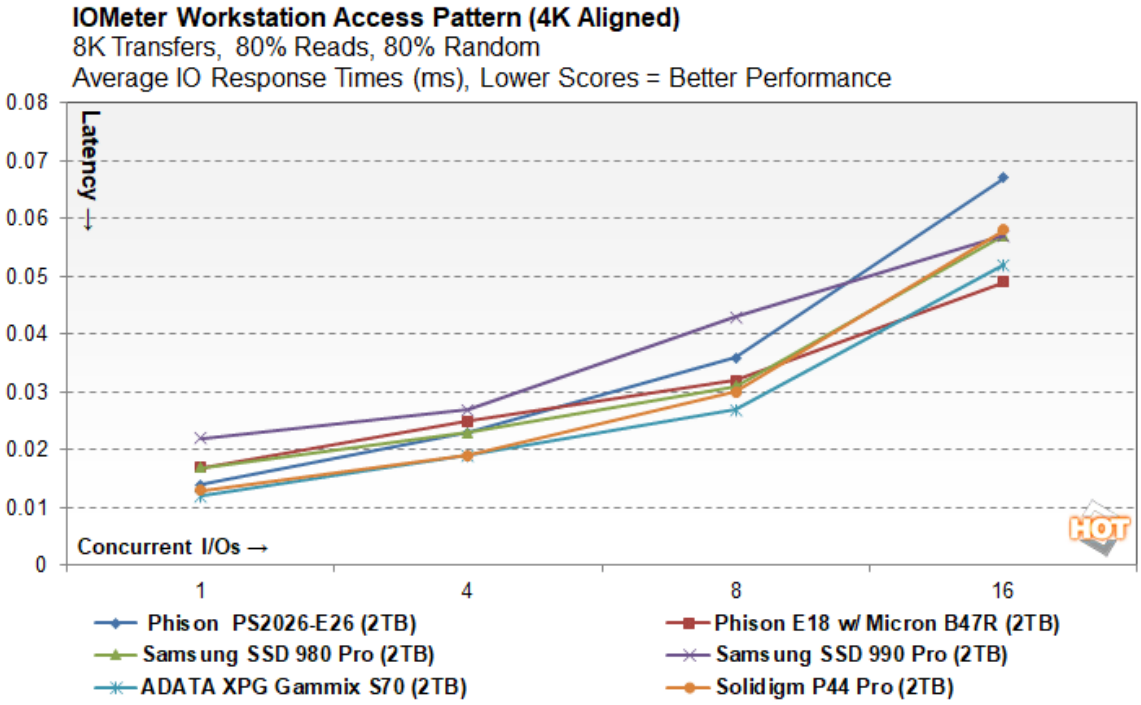


图 6-73

In these latency tests, the Phison reference platform and its E26 controller impress at 4K but once again falter a bit in the Workstation pattern. Clearly, this drive and its controller are tuned for client desktop usage and less for heavy workstation access patterns.

SiSoft SANDRA 2021

Next we used SiSoft SANDRA, the System ANalyzer, Diagnostic and Reporting Assistant for some quick tests. Here, we used the File System Test and provide the results from our comparison SSDs. Read and write performance metrics, along with the overall drive score, are detailed below.

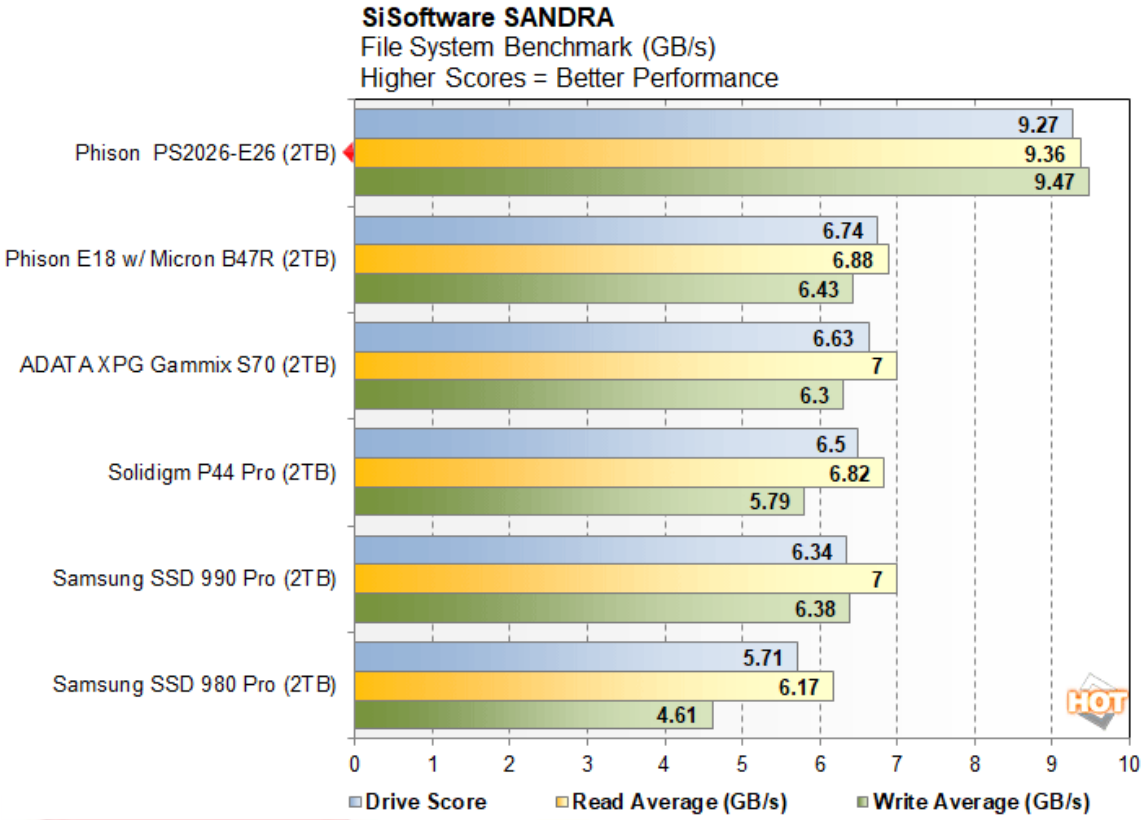


图 6-74

Sandra's File System Benchmark was originally created with hard drives in mind, and as a result it is primarily a test of sequential performance. Unsurprisingly, the new PCIe 5.0 SSD runs away from the competition in this test, putting up a top score that's almost half-again as fast as the next-fastest competitor. If you need to move big files around a bunch, you need a PCIe 5.0 SSD.

ATTO Disk Benchmark

ATTO is another "quick and dirty" type of disk benchmark that measures transfer speeds across a specific volume length. It measures raw transfer rates for both reads and writes and graphs them out in an easily interpreted chart. We chose .5KB through 64MB transfer sizes and a queue depth of 6 over a total max volume length of 256MB. ATTO's workloads are sequential in nature and measure raw bandwidth, rather than I/O response time, access latency, etc.

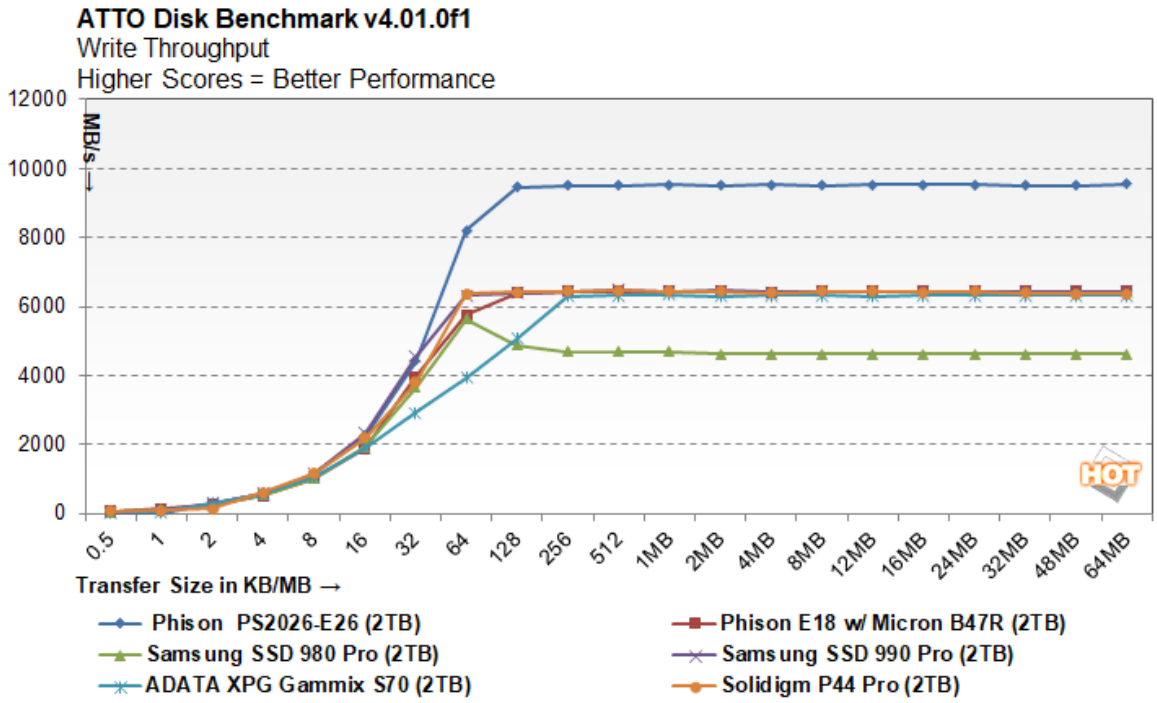


图 6-75

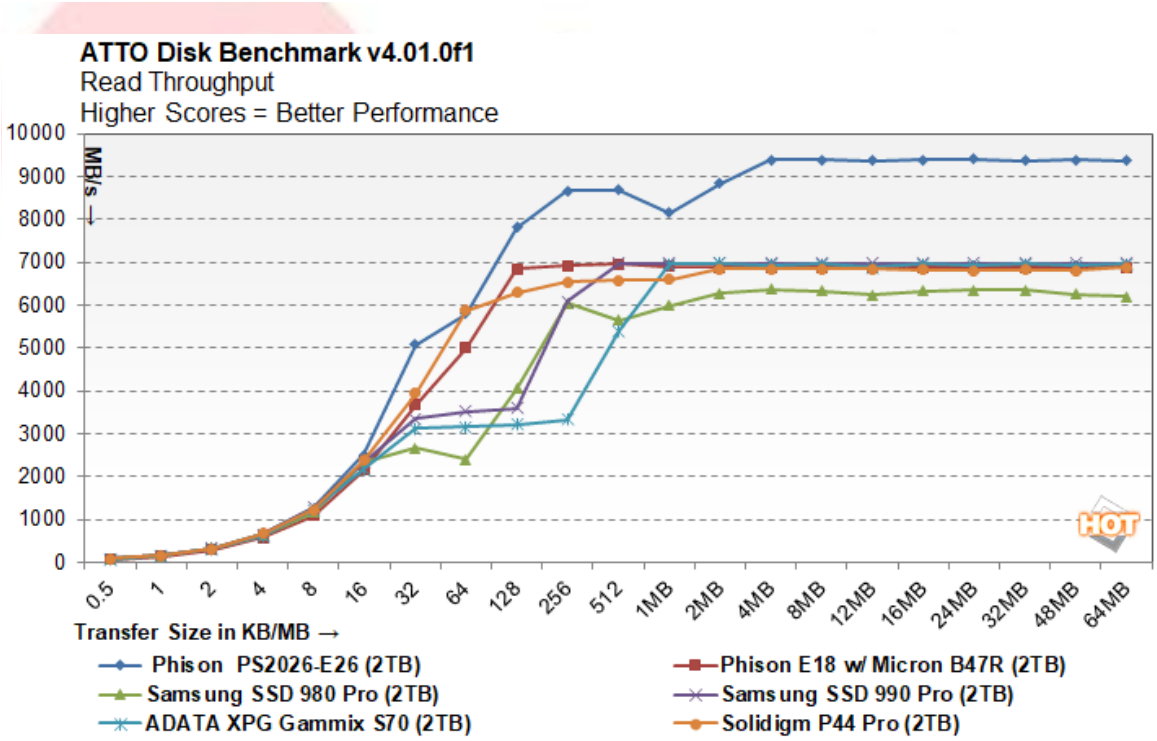


图 6-76

In these sequential tests, we once again see the PCIe 5.0 Phison drive completely outpace the competition. The rest of the drives, limited by their PCIe 4.0 interfaces as they are, cluster up around 7 GB/sec after the 1MB transfer size, although the E26 is creeping ahead in reads as early as 32KB. Meanwhile the Phison drive peaks at over 9 GB/second in both reads and writes.

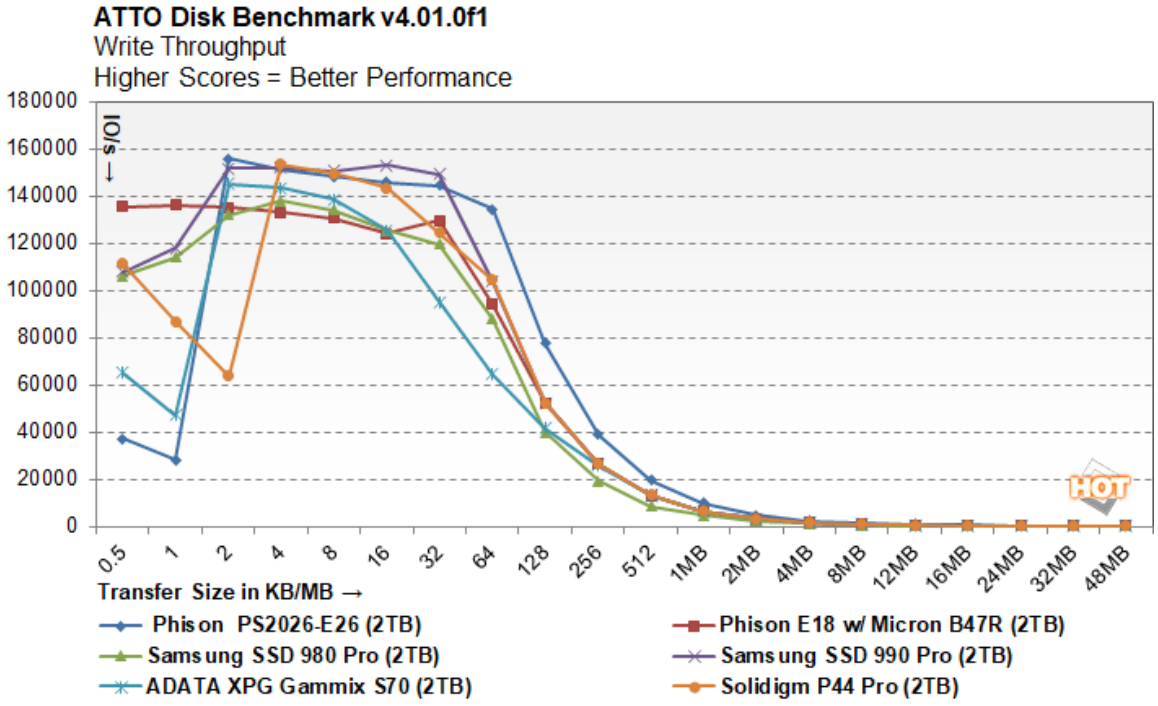


图 6-77

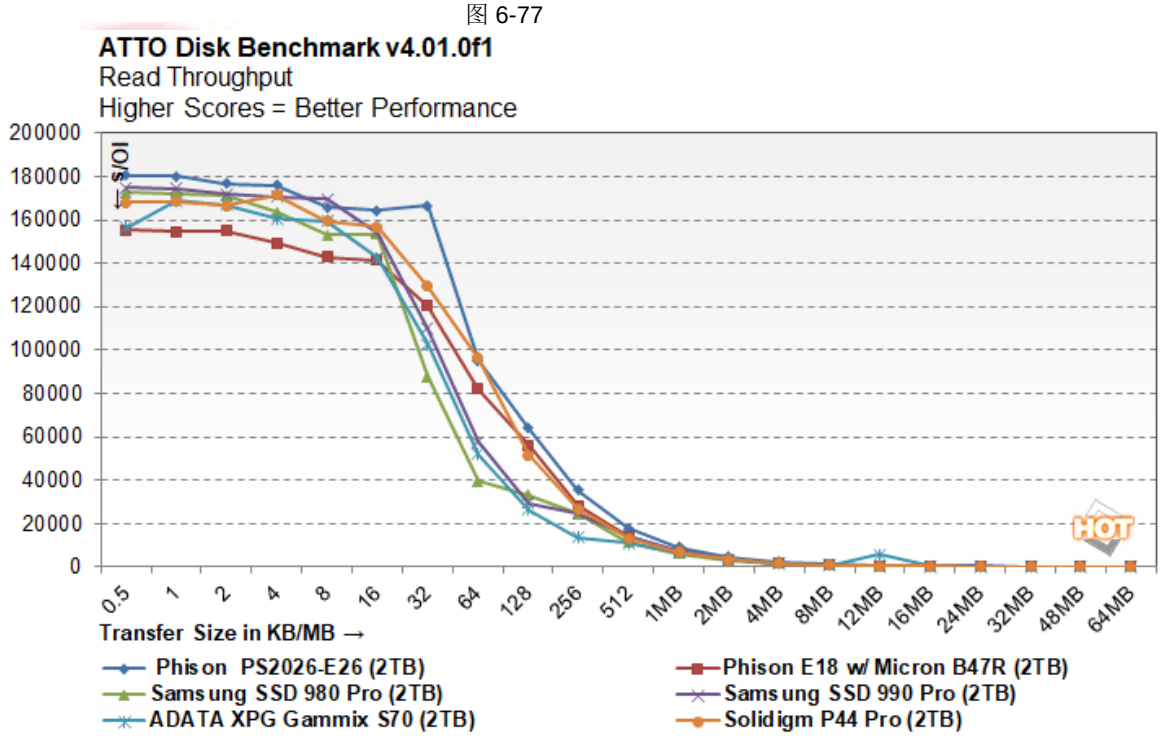


图 6-78

These benchmarks measure IOPS instead of transfer rate, so naturally the numerical performance of the drive falls off a cliff as transfer size rises. The Phison drive manages to lead the pack in read IOPS at every transfer size, but it clearly struggles a bit with write performance with very small transfer sizes. This might be a firmware optimization for Phison to work out, or it could simply be an oddity of ATTO's test.

AS SSD Compression Benchmark

Next up we ran the Compression Benchmark built-into AS SSD, an SSD specific benchmark being developed by Alex Intelligent Software. This test is interesting because it uses a mix of compressible and non-compressible data and outputs both Read and Write throughput of the drive. We only graphed a small fraction of the data (1% compressible, 50% compressible, and 100% compressible), but the trend is representative of the benchmark's complete results.

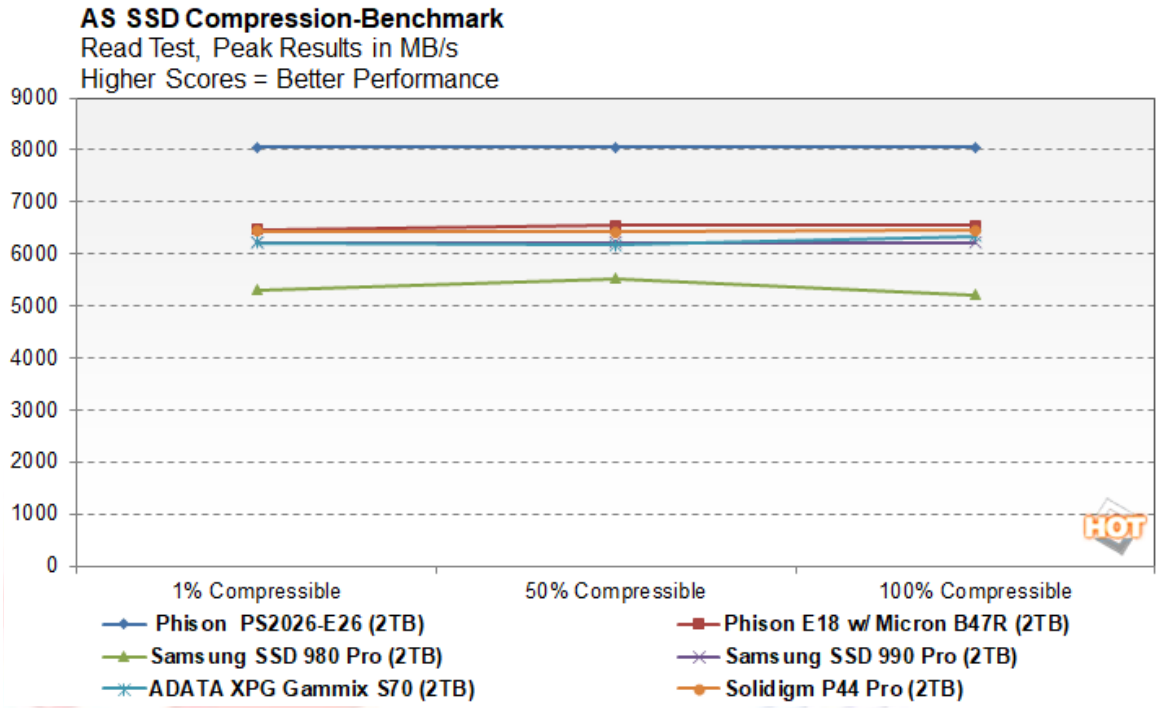


图 6-79

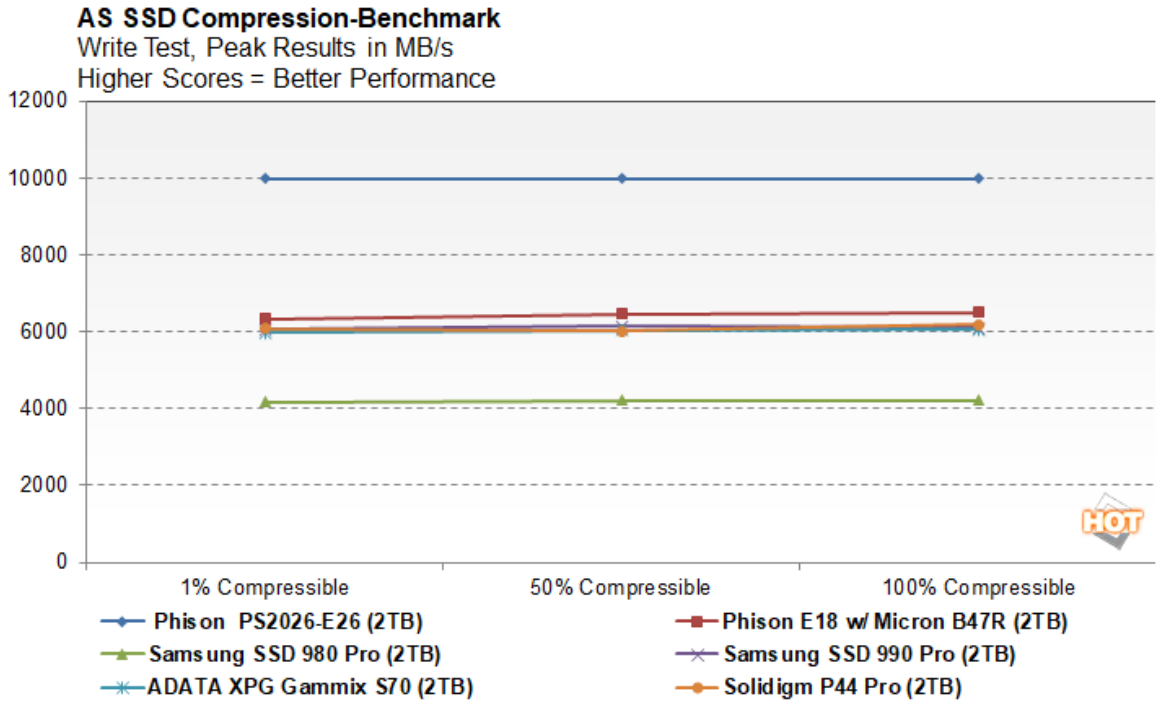


图 6-80

As none of these drives rely on compression to improve their storage performance, these tests are essentially completely flat aside from some margin-of-error differences. Notably, however, the Phison E26 reference drive crests 10 GB/second in the write test. We'll investigate sequential writes more and find the real peak performance on the next page.

6.2.3.2.2.2 Phison E26 SSD Preview: More Benchmarks, Gaming Tests And The Verdict

图 6-81

EFD Software's HD Tune is described on the company's website as such: "HD Tune is a hard disk utility with many functions. It can be used to measure the drive's performance, scan for errors, check the health status (S.M.A.R.T.), securely erase all data and much more." The latest version of the benchmark added temperature statistics and improved support for SSDs, among a few other updates and fixes.

HDTune v5.75 Benchmarks

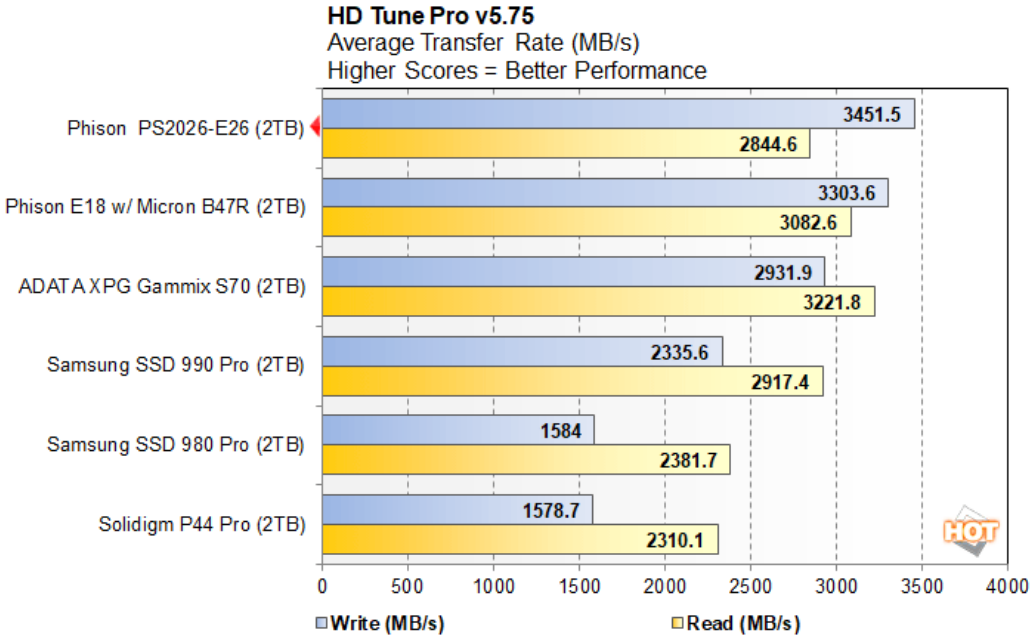

HD Tune Pro v5.75
Average Transfer Rate (MB/s)
Higher Scores = Better Performance
Phison PS2026-E26 (2TB)
3451.5
2844.6
Phison E18 w/ Micron B47R (2TB)
3303.6
3082.6
ADATA XPG Gammix S70 (2TB)
2931.9
3221.8
Samsung SSD 990 Pro (2TB)
2335.6
2917.4
Samsung SSD 980 Pro (2TB)
1584
2381.7
Solidigm P44 Pro (2TB)
1578.7
2310.1
HOT
0
500
1000
1500
2000
2500
3000
3500
4000
Write (MB/s)
Read (MB/s)

图 6-82

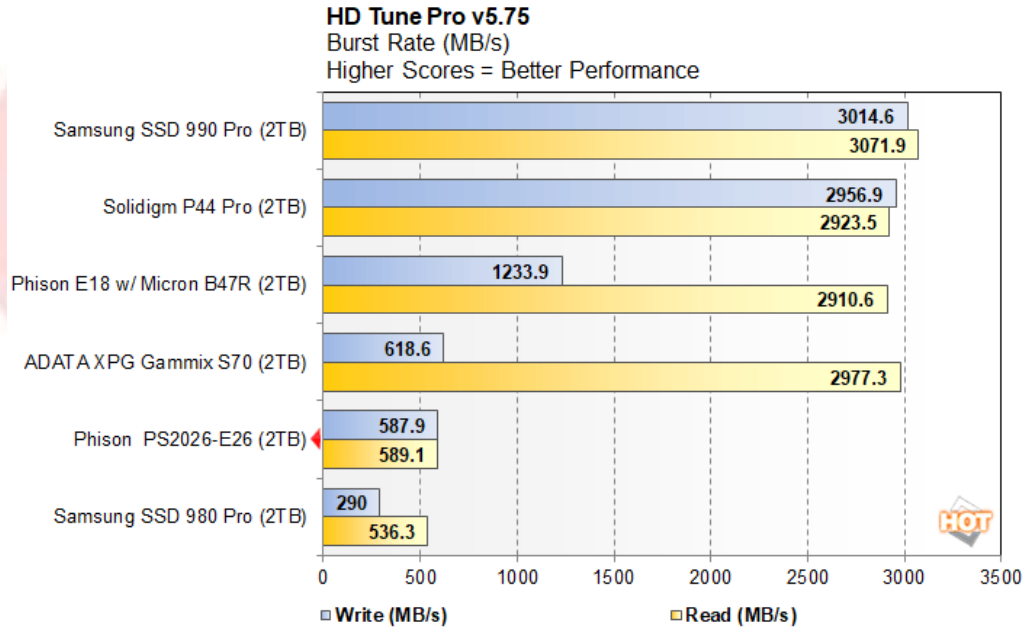

HD Tune Pro v5.75
Burst Rate (MB/s)
Higher Scores = Better Performance
Samsung SSD 990 Pro (2TB)
3014.6
3071.9
Solidigm P44 Pro (2TB)
2956.9
2923.5
Phison E18 w/ Micron B47R (2TB)
1233.9
2910.6
ADATA XPG Gammix S70 (2TB)
618.6
2977.3
Phison PS2026-E26 (2TB)
587.9
589.1
Samsung SSD 980 Pro (2TB)
290
536.3
HOT
0
500
1000
1500
2000
2500
3000
3500
Write (MB/s)
Read (MB/s)

图 6-83

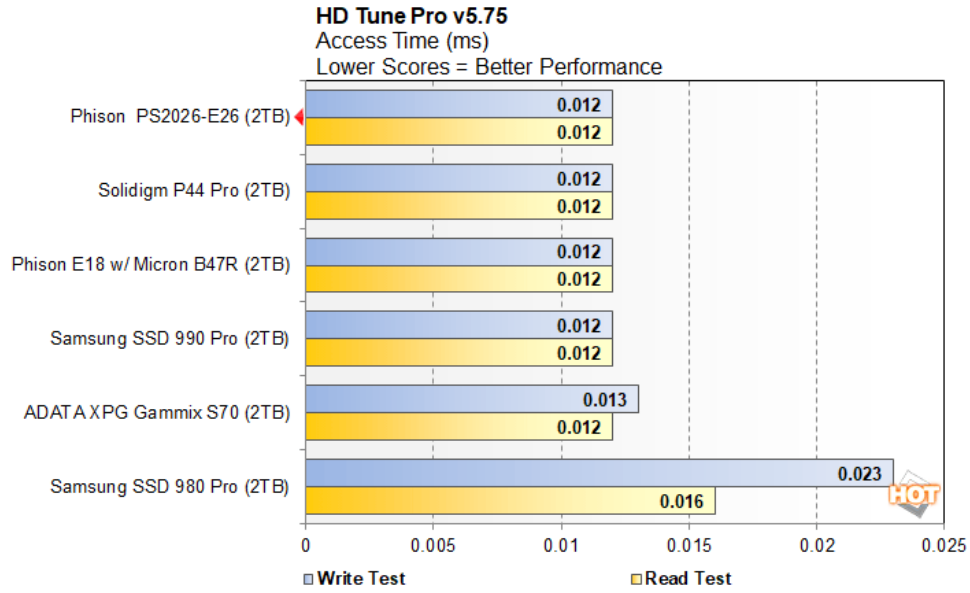


图 6-84

As expected, the new Phison drive and its fresh E26 controller offer very impressive average transfer rates in HD Tune, although the read performance actually lags behind some of the other drives in our tests. That's despite the PS2026-E26 having the lowest access time we've measured on this platform; we suspect all of these drives are hitting a sort of system bottleneck of sorts in that metric.

Surprisingly, the PS2026-E26 actually puts in a relatively poor result in HD Tune Pro's burst transfer rate test. We're loathe to speculate on the reasons for this, but if we were to hazard a guess, we'd assume that the IO+ Technology firmware is so heavily tuned for sustained transfers that the burst rate suffers slightly.

CrystalDiskMark x64 Benchmarks

CrystalDiskMark is a synthetic benchmark that tests both sequential and random small and mid-sized file transfers using incompressible data. It provides a quick look at best and worst case scenarios with regard to SSD performance, best case being larger sequential transfers and worse case being small, random transfers.

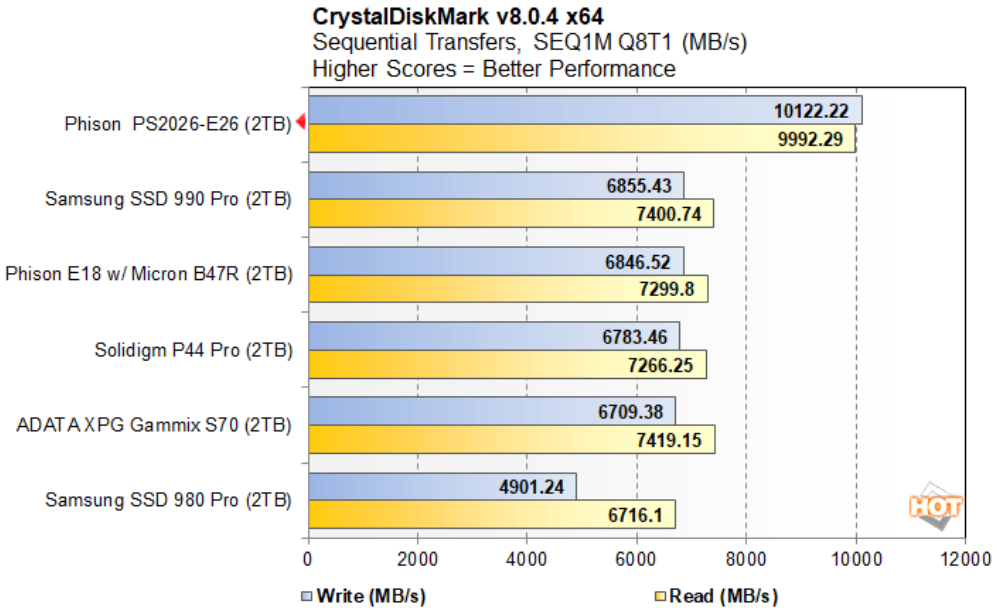


图 6-85

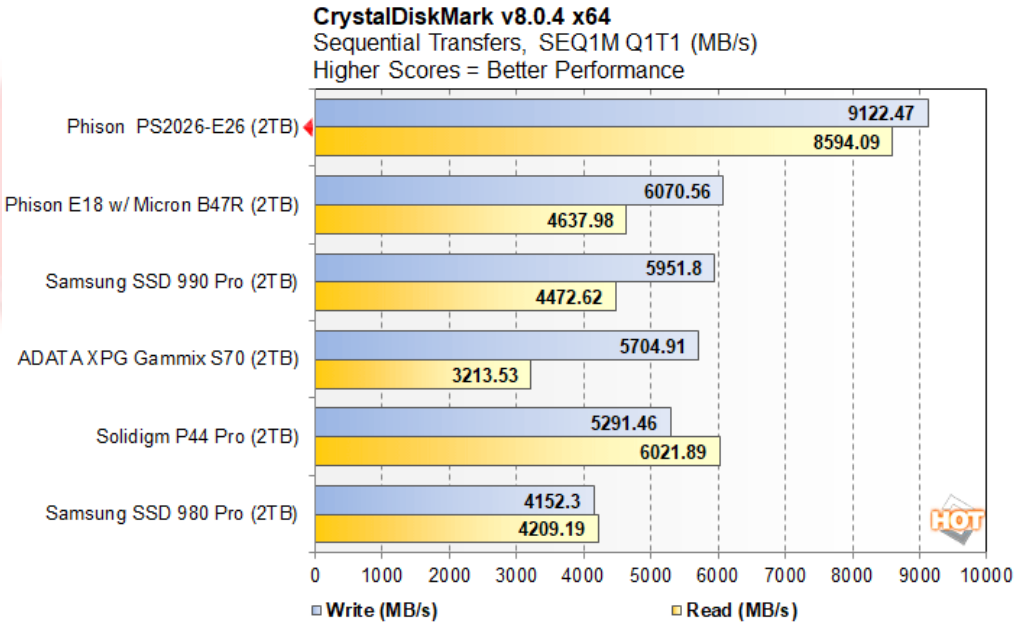


图 6-86

We are once again unsurprised that the Phison drive and its white-hot PCIe 5.0 interface absolutely destroy all comers in these sequential benchmarks. We feel like the Q8T1 results represent what is most likely the absolute maximum real-world performance of this drive, at least on the AMD platform we tested it with. Performance on Intel's 13th Gen platform is actually somewhat higher, but all of our reference data was captured on a socket AM5 system.

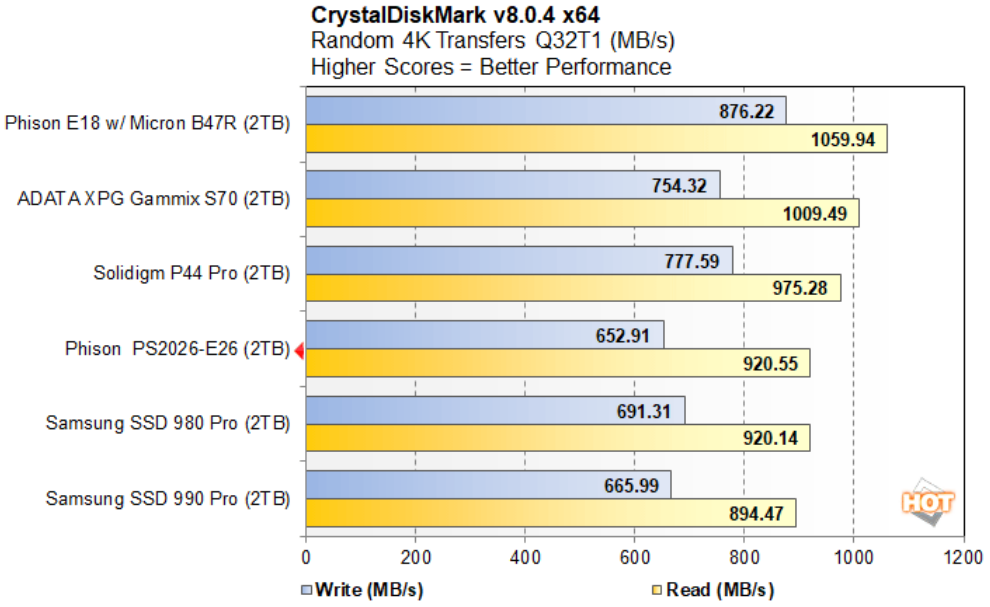


图 6-87

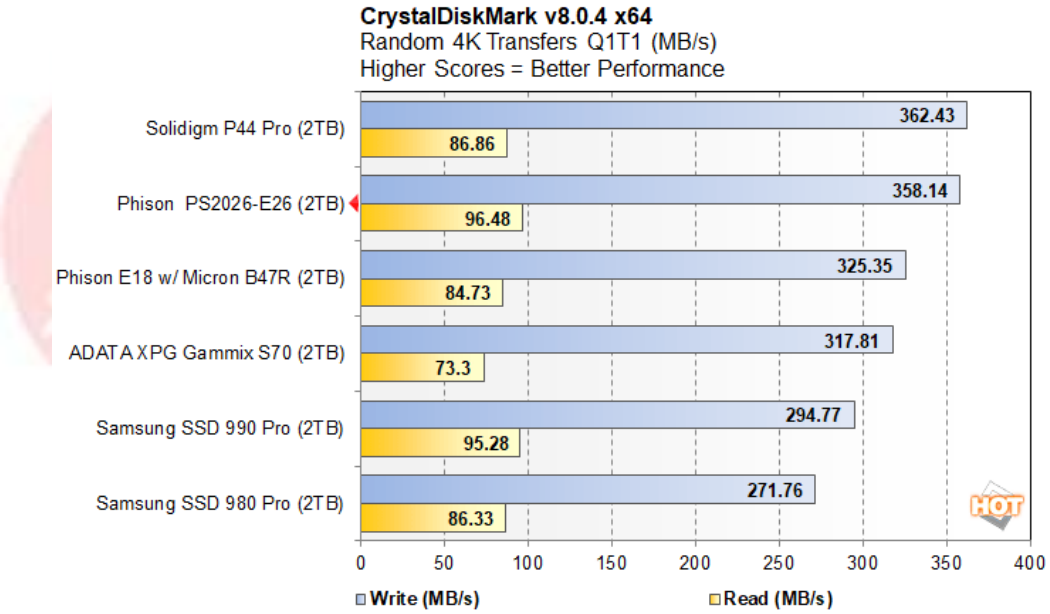


图 6-88

The random access results in CrystalDiskMark aren't as stunning at a glance as the sequential results, but note that this drive is the fastest of the bunch in single-threaded random 4K reads with a queue depth of 1, even beating out the Samsung SSD 990 Pro. The Q32T1 results are good but unremarkable among the other drives; as we've noted already, this SSD is very clearly optimized for a desktop workload that is very unlikely to see such extreme I/O queue depths.

Final Fantasy XIV: Endwalker Game Level Load Times

We also tested game level load times using the *Final Fantasy XIV: Endwalker* benchmark. This tool loads an array of different game levels during its graphics benchmark and outputs the average result when complete.

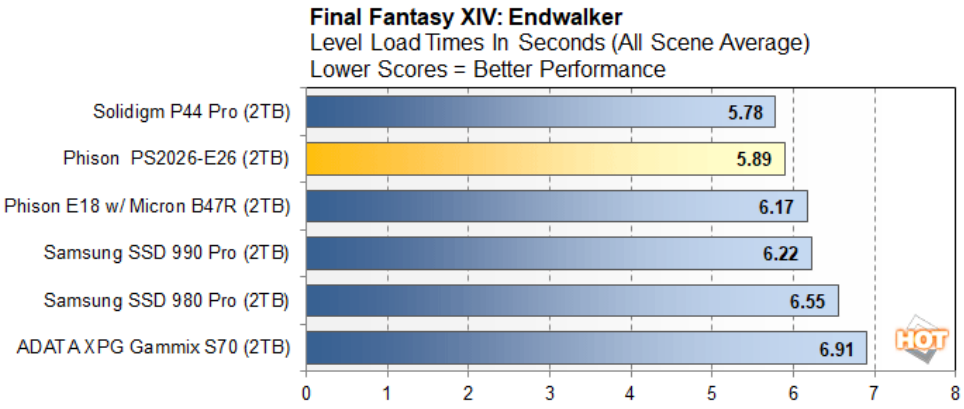


图 6-89

We honestly expected the new drive to best all comers in this test, which is largely a benchmark of how fast your SSD can chuck data into RAM. The Solidigm drive has exceptionally low access latency as well as a much better burst transfer rate than the Phison drive, so that's likely what allowed it to creep ahead. The difference in the two drives is a whopping 110 milliseconds, so it's essentially margin-of-error anyway; both are ludicrously fast.

UL's 3DMark Gaming Storage Benchmark

UL recently added a gaming-centric storage benchmark to 3DMark, that leverages trace-based tests of actual PC games and gaming-related activities (like streaming with OBS) to measure real-world gaming performance in a variety of scenarios. The tests include things like loading *Battlefield V*, *Call of Duty: Black Ops 4*, and *Overwatch* from the initial launch to the main menu, recording a 1080p gameplay video at 60 FPS with OBS while playing *Overwatch*, installing *The Outer Worlds* and saving game progress, then finally, copying the Steam folder for *Counter-Strike: Global Offensive* from one drive to another.

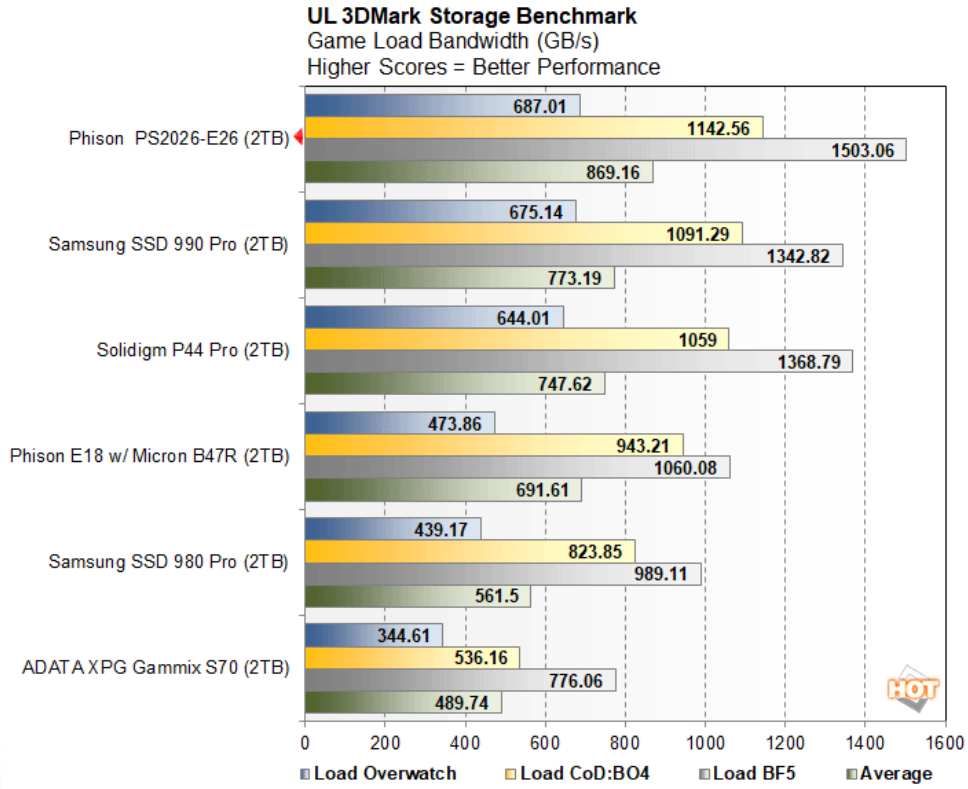


图 6-90

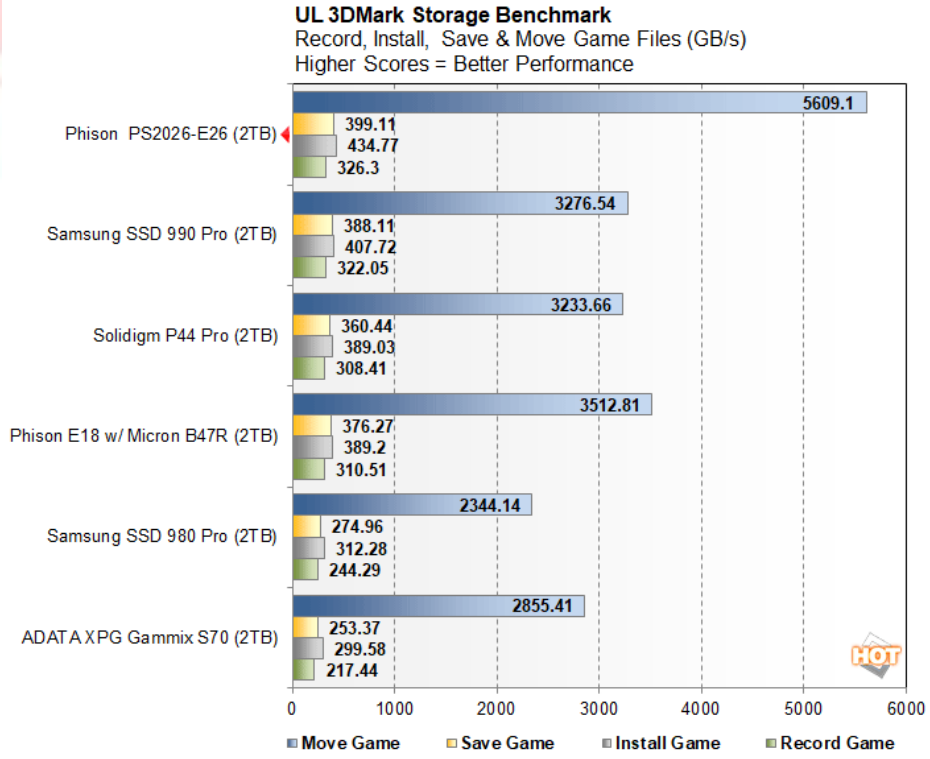


图 6-91

The *Battlefield V* test and the "Move Game" test both strongly favor the new PCIe 5.0 SSD, but it comes out at least slightly ahead of the other drives in just about every test in this benchmark.

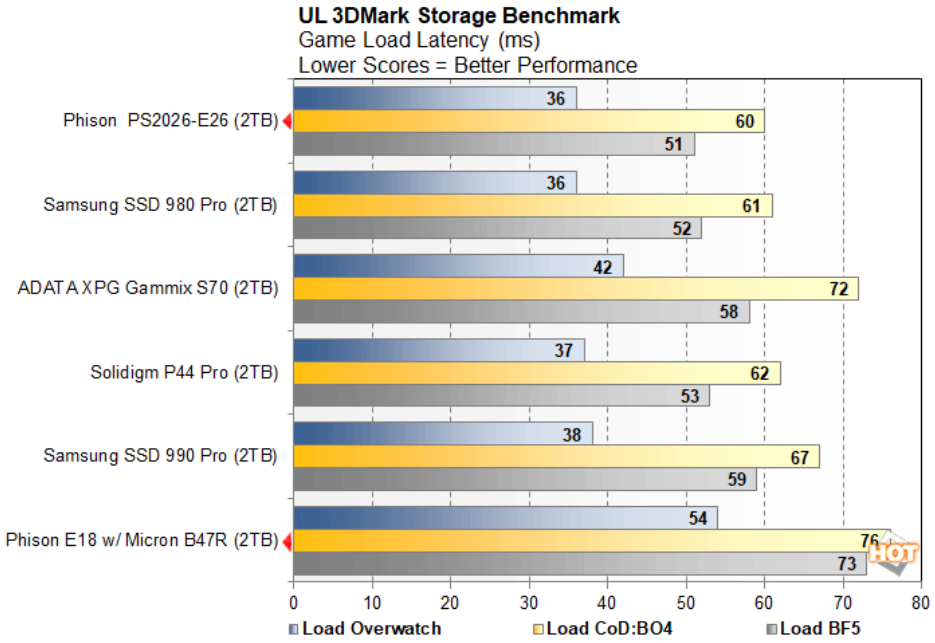


图 6-92

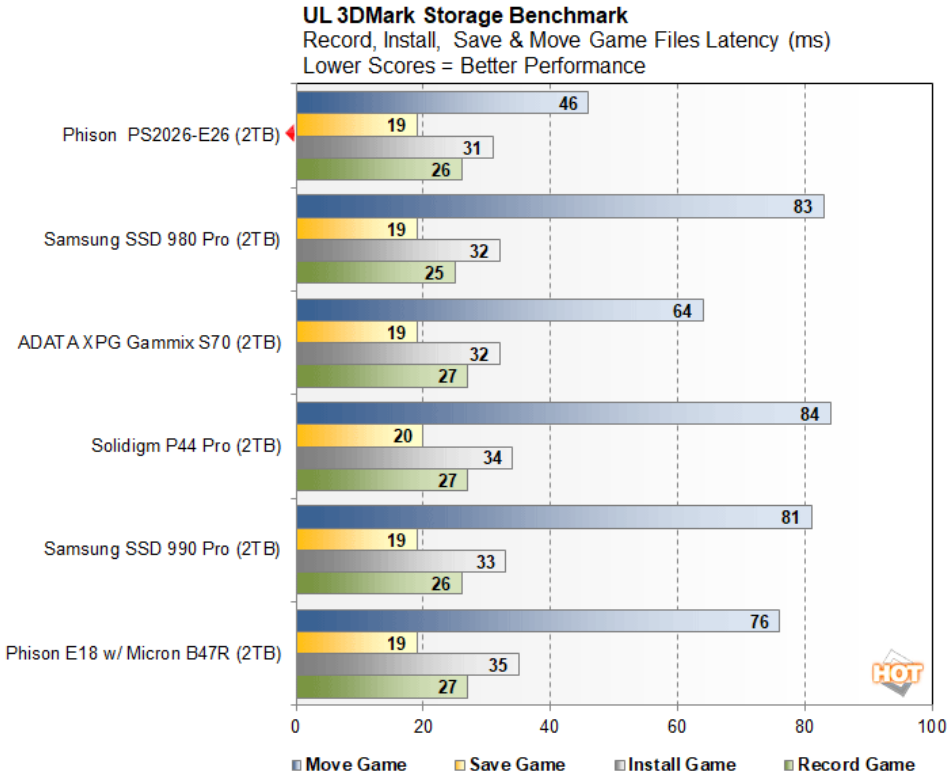


图 6-93

That pattern continues in these sets of tests that measure I/O latency instead of throughput. As we discussed before, higher throughput can also mean lower latency when it allows you to keep things flowing instead of filling up buffers, and that could well be why the Phison PS2060-E26 reference platform comes out on top or ties the fastest SSDs in virtually every test.

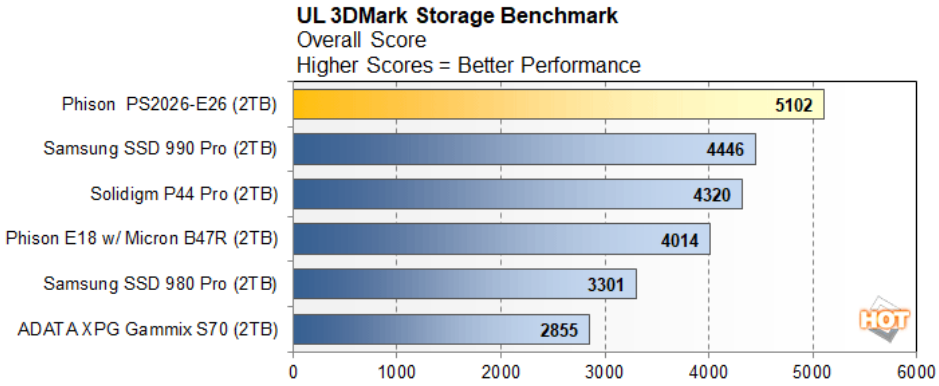


图 6-94

Naturally, given its performance in the individual benchmarks, the overall score for the new drive is a solid win over the other SSDs. Going all the way back to *3DMark '99*, the storied benchmark utility has always favored the latest hardware, and that holds true for the new storage benchmark, too.

UL PCMark 10 System Drive Storage Test

We like PCMark 10's new quick storage benchmark module for its real-world application measurement approach to testing. PCMark offers a trace-based measurement of system response times and bandwidth under various scripted workloads of traditional client / desktop system use cases.

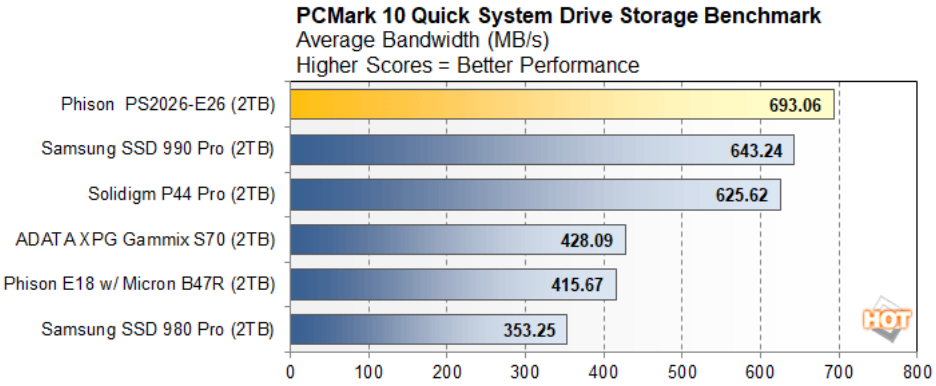


图 6-95

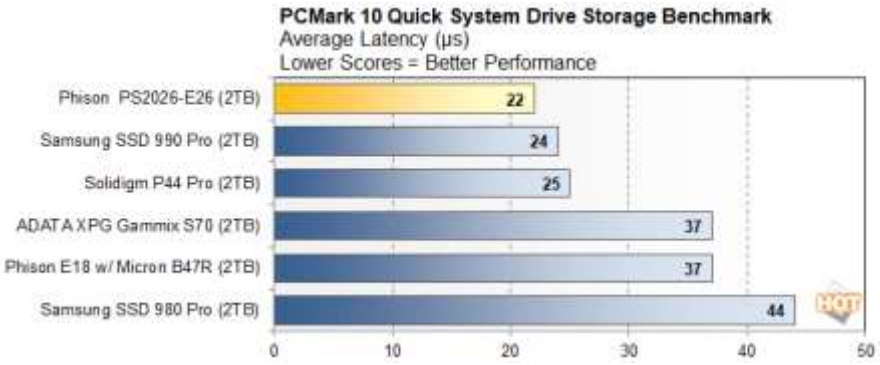


图 6-96

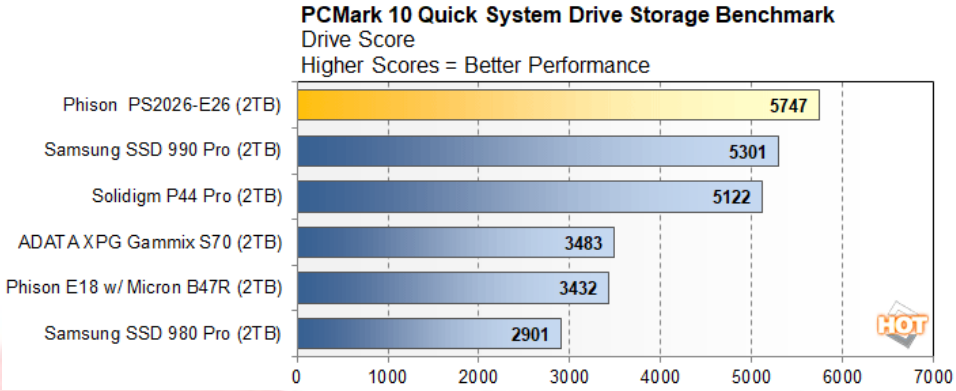


图 6-97

In the trace-based PCMark tests, which are comprised of a broad mix of consumer workloads, the E26-based drive takes the top spot in every metric, and by a conclusive margin, too. Aside from the *Final Fantasy XIV* test, these are probably the most "real-world" tests that we perform, and Phison's new SSD reference platform performs with panache, dispatching even the mighty Samsung SSD 990 Pro.

Phison E26 PCIe Gen 5 SSD Preview: The Preliminary Verdict

Overall, in the synthetic benchmarks the Phison drive blew away the competition in sequential tests but didn't always lead in random accesses. Don't get it twisted, though: we're testing it against some of the fastest PCIe 4.0 SSDs around, and even at this early stage it keeps up with the pack.

Where we were much more impressed was with the drive's practical and real-world performance. It blew away the competition in both the 3DMark storage test as well as the PCMark productivity tests. That's surely thanks in part to its PCIe 5.0 interface, but there's no way that's the whole story here. Phison's hard work with the firmware likely plays a role, and the large 4GB LPDDR4 cache surely helps as well.

图 6-98

Find Phison-Based NVMe SSDs @ Amazon

Despite DirectStorage being "released" on Windows 11, there's currently no commercial software at all beyond a few test applications that actually uses it. In the synthetic bulk workload DirectStorage test we experimented with, we're given a glimpse of that the drive is capable of.

This SSD is specifically tuned for DirectStorage-like workloads, so it would have been interesting to see how it holds up against other SSDs that haven't received such tuning when actual games using the technology arrive. We'll just have to wait until actual DirectStorage-enabled games hit the market, the first of which should be Square-Enix's Forspoken, releasing on January 24th.

This reference platform isn't a retail SSD. You won't be able to buy this drive in this exact form, because Phison, unlike, say, NVIDIA, doesn't sell its own branded SSDs. Instead, you'll be buying drives with these controllers from the likes of Kingston, Seagate, Sabrent, ADATA, and Aorus, among many others. Phison doesn't speak for its partners, but the company did tell us that this hardware has been released to manufacturers and that "production has started," so we expect these SSDs will be showing up sooner than later.

6.2.3.2.2.3 破 10000MB/s！但是……PCIe 5.0 SSD 技嘉 AG510K 笔电安装测试记录

2023-03-09 13:59:19 | 来源：公平评测 | 作者：song1118

最新上市的技嘉 PCIe 5.0 SSD，标准 M.2 2280 规格，号称顺序读写速率高达 10000 MB/s，如果在笔记本电脑上安装使用，又将如何呢？

文章前言

最新上市的技嘉 PCIe 5.0 SSD，标准 M.2 2280 规格，号称顺序读写速率高达 10000 MB/s，如果在笔记本电脑上安装使用，又将如何呢？ 笔者第一时间购买了技嘉 AG510X 2TB PCIe 5.0 SSD，进行了实测。 全文有 4000 余字 50 张图片，中文版在“公平评测“中文站首发，英文版将稍后在“FairReviews”英文站发布。 全文分为以下章节： **接友线报：突然到手 安装忧虑：散热太大 不装散热：当场消失 换用雷电：实测高温 装上散热：温度大降 解决方案：散热杂交 最终评价：高能高温**

接友线报：突然到手

3 月 4 日下午 2 点多，突有友告知笔者，PCIe 5.0 SSD 有卖了，并且是 M.2 2280 规格的。 闻讯当即直扑京东，一看有 2TB 的，容量合格，于是没细看就下单付款。 3 月 5 日，从京东武汉仓库发出的顺丰送达，这速度，可以！

图 6-99

收到之后，拆开看到外包装才知道是技嘉出品，京东上面标注的型号是大雕 510K。 实物彩盒正上方中央是技嘉雕头金色标志，下方标注 AORUS Gen5 10000 SSD。

图 6-100

彩盒底部标注的是型号是 AG510K2TB----AG5 自然是 AORUS Gen5 的缩写，10K 自然是 10000MB/s 的含义，2TB 代表容量。

图 6-101

彩盒其实是一个两头贯穿的套盒，两侧都有技嘉封签，取下这个套盒之后里面是黑色的内盒，上方中央还有技嘉雕头标志。

图 6-102

内盒第一层是 M.2 SSD，下面还有官方散热器：

图 6-103

内盒包装物取出一览图，除了 M.2 SSD 和散热器之外，还有两张纸质文件。

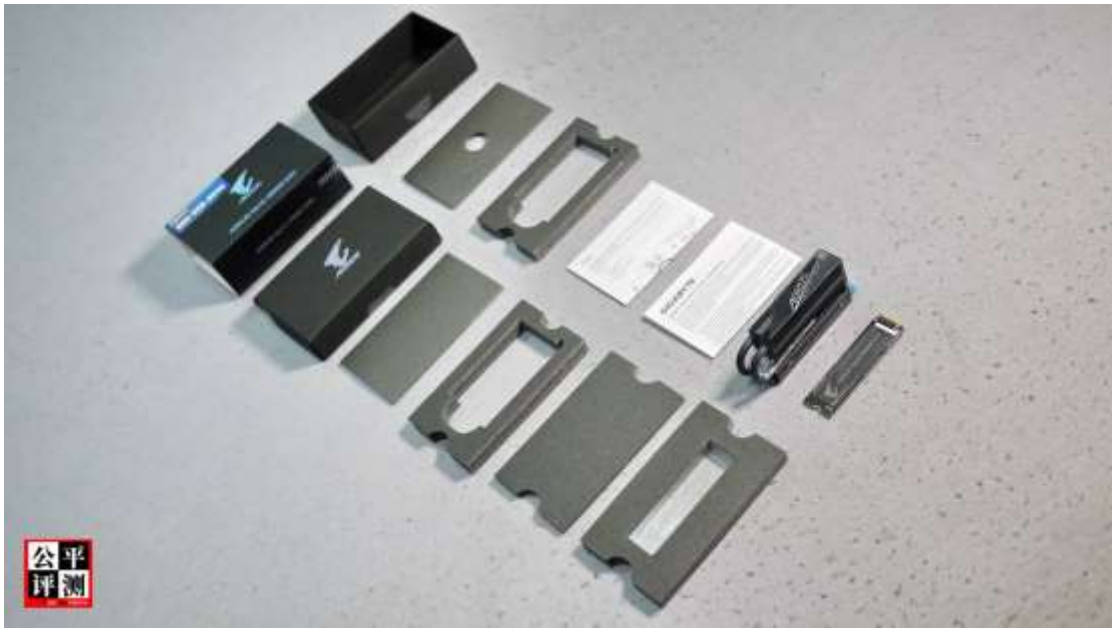


图 6-104

纸质文件分别为快速安装说明书和安装图解。

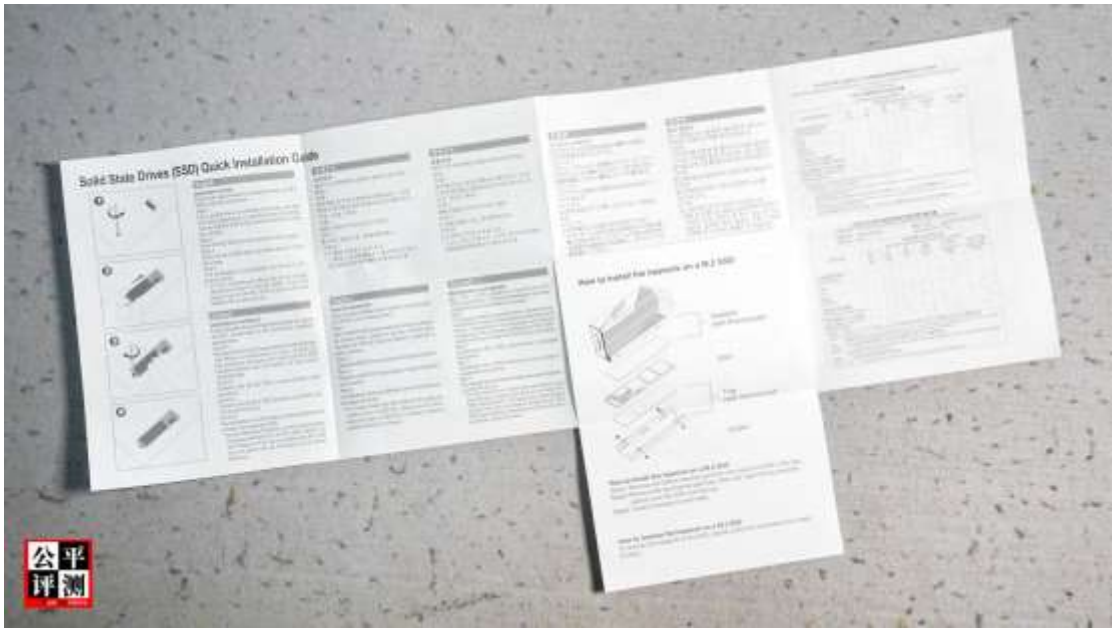


图 6-105

散热器中还有一个小塑料袋封装的螺丝，用来固定散热器。

图 6-106

相对 M.2 2280 SSD 来说，散热器的体积十分庞大惊人，使笔者感到深深的忧虑。

图 6-107

安装忧虑：散热太大

忧虑的原因，是因为笔者是想将这个 PCIe 5.0 M.2 SSD，安装在笔记本电脑上使用。

图 6-108

笔记本电脑的型号是微星 CreatorPro X17，是 2022 年上市的、当时唯二支持 PCIe 5.0 M.2 2280 SSD 的笔电，还有一个型号是其同门的同形异素体----微星 GT77。

图 6-109

将 AG510K2TB（下文简称为“AG510K”）的官方散热器，试放到微星 CreatorPro X17 唯一支持 PCIe 5.0 M.2 2280 SSD 的位置试试看，似乎可以安装：

图 6-110

但这散热器超高尺寸，要么将笔电底盖挖洞，要么将笔电机体反置或 90 度放置使用，才能使用。

图 6-111

再仔细看看 AG510K 的尺寸，确实是标准 M.2 2280：

图 6-112

看看背面，无颗粒似是单面的：

图 6-113

仔细观看侧面，发现其背面靠近 M.2 接口的地方，还有一些元器件，看来不能算是严格的单面 SSD 了，不支持双面 SSD 的机型怕是不方便

图 6-114

既然来都来了，就先试试不要散热器吧！

图 6-115

不装散热：当场消失

安装好 AG510K 的微星 CreatorPro X17，开机进入 BIOS，RST GIABYTE AG510K2TB 正常识别。

图 6-116

进入 intel RST 设置，和 PCIe 4.0 SSD 组建 RAID，也操作正常。 1

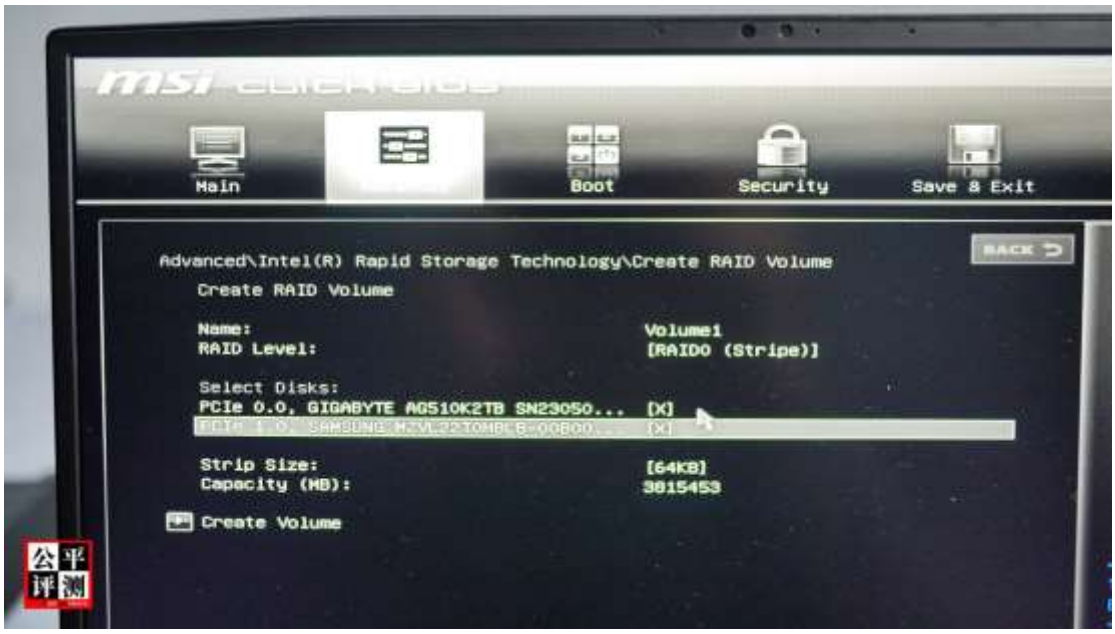


图 6-117

不过为了测试，还是先将微星 CreatorPro X17 的 VMD 关闭了，处于非 RST 模式，此时 BIOS 中硬盘信息会显示为 GIABYTE AG510K2TB。

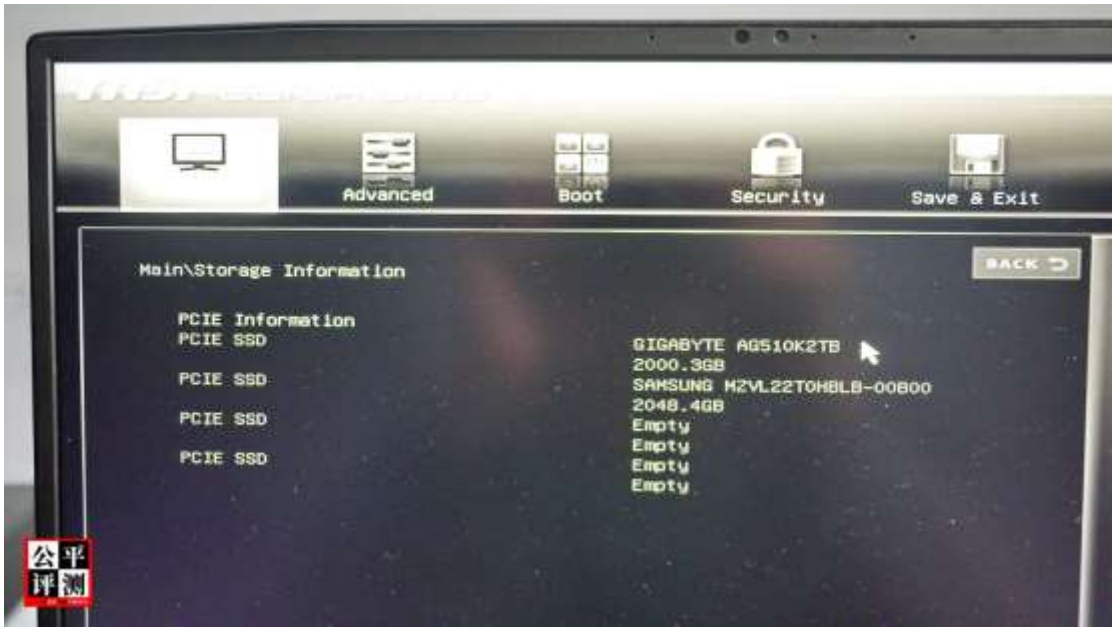


图 6-118

开进进入 Windows 11，设备管理器、AIDA64 和 HWiNFO64 都准确识别了 AG510K：PCIe 5.0 x 4，32.0 GT/s。 微星 CreatorPro X17 的 PCIe 5.0 x4 SSD 和独立显卡使用的 PCIe 5.0 x8，都是直连 CPU 的。

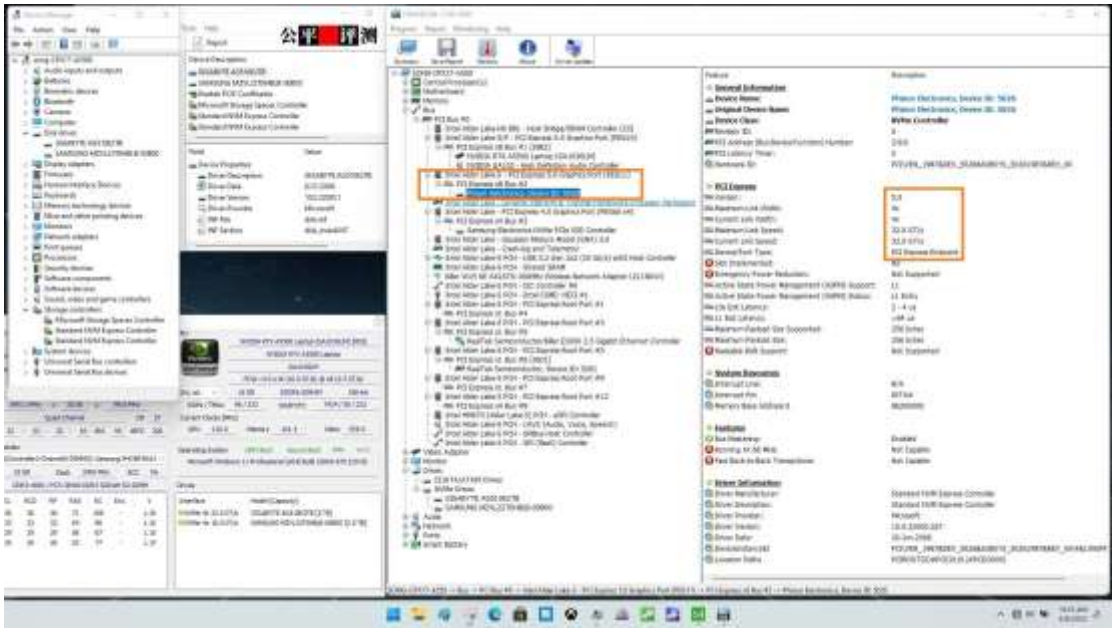


图 6-119

作为对比，先测试了同样是直连 CPU 的 PCIe 4.0 x4 SSD 的三星 MZVL22T0HBLB-00B00，需要注意的是，此盘是系统安装分区所在，对成绩会稍有不利。 CrystaDiskMark 标准测试，用时约 4 分 40 秒。

图 6-120

使用 Generic Log Viewer，对 HWiNFO64 记录的测试数据进行分析，得到下图，可以看到，三星的温度从开始的 37 摄氏度，到测试完成时上升到了 63 摄氏度。 测试时室温为 20 摄氏度左右。

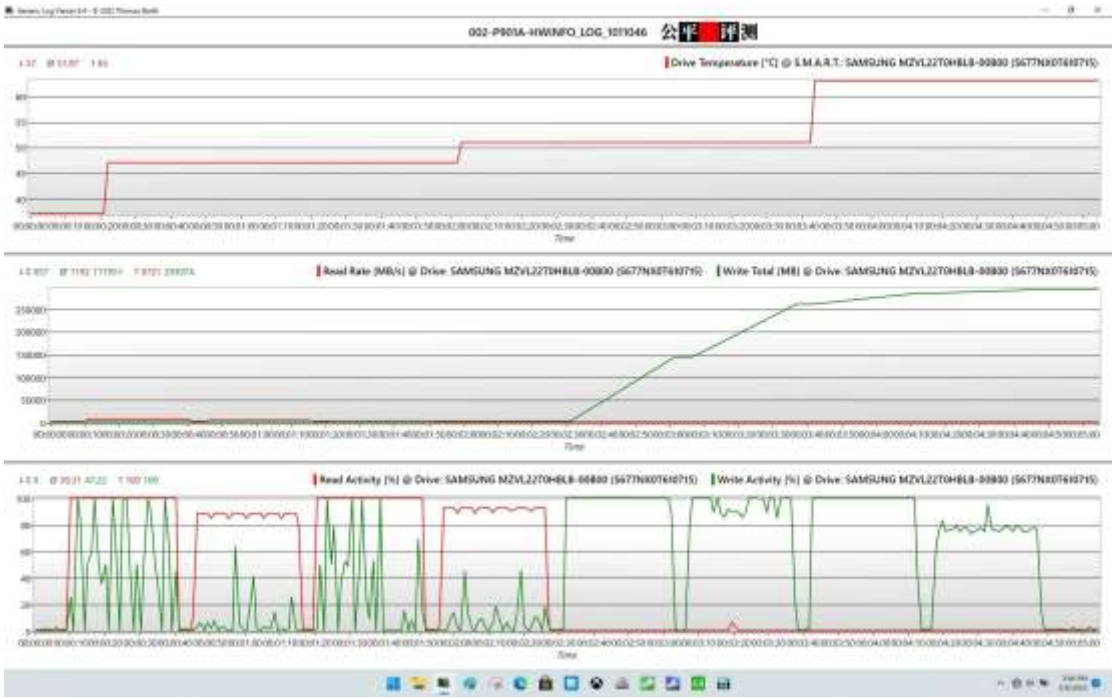


图 6-121

然后，对 AG510K 进行了同样的测试： 不错！顺序读取破万，有 10088.09MB/s！ 不对！怎么测试没全部完成，就自动停止了？ 请注意下图右侧两个橙色方框中的信息。

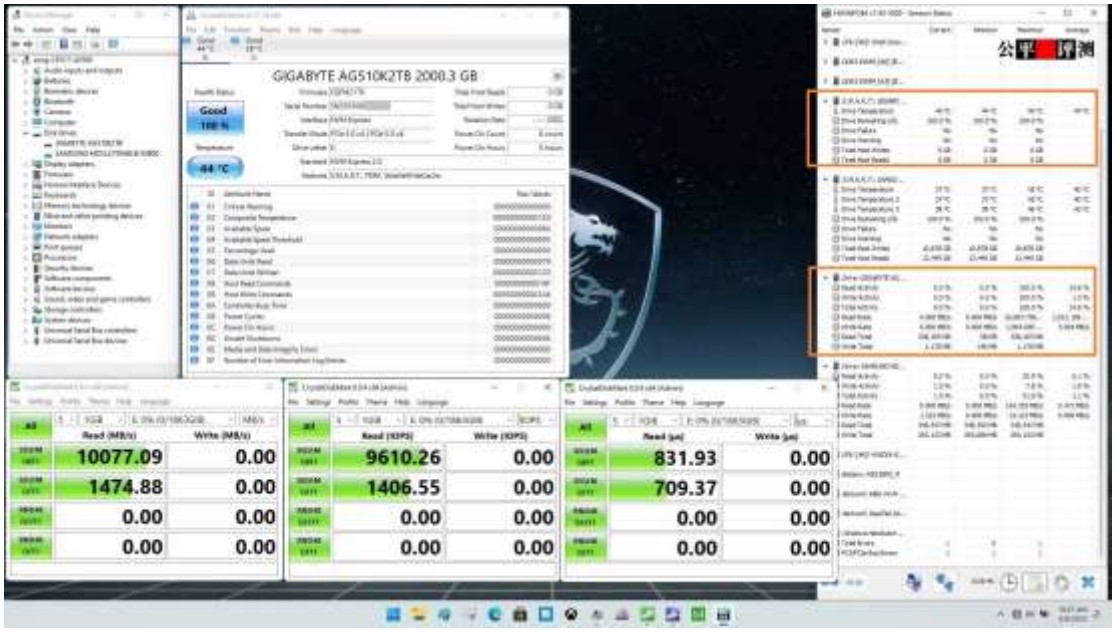


图 6-122

不要慌！一切都是技术调整！ 应该是在测试时笔者操作不小心鼠标点击了停止测试。 再次测试，结果是……显示的顺序读取速率是 43763.83 MB/s ？！

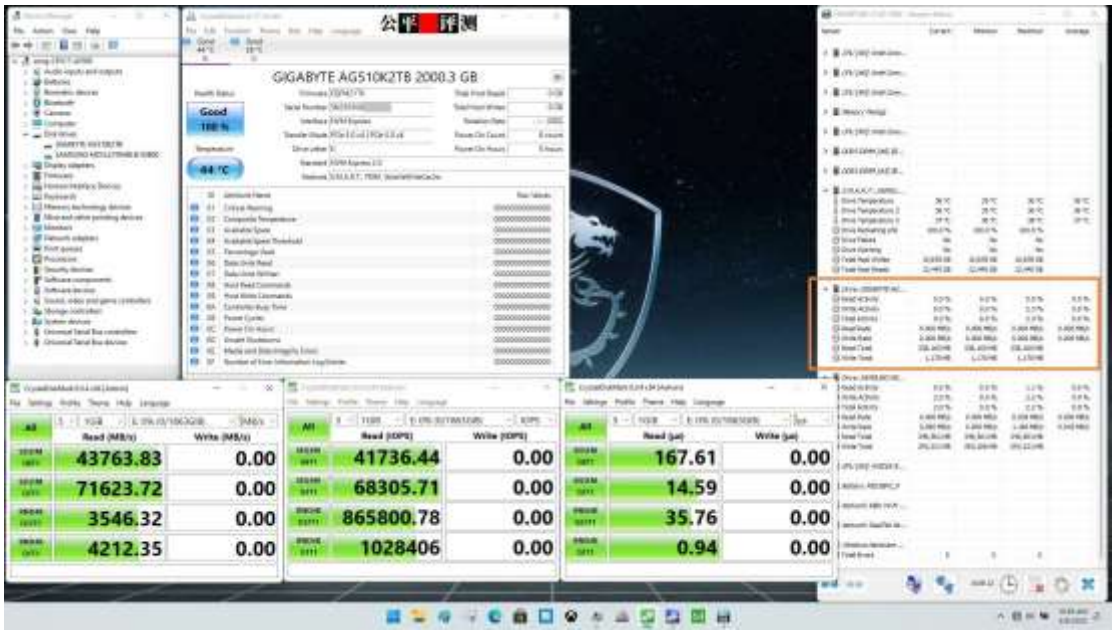


图 6-123

再一看右侧的 HWiNFO64，AG510K 的 S.M.A.R.T 信息栏已经不见了！ 只剩下一个信息栏？ 不妙！ 是不是这 AG510K 因为没有安装散热器，测试产生高温，然后挂了？ 在上面进行的第一次测试时，HWiNFO64 显示的 AG510K 的 S.M.A.R.T 信息中，其温度只有 44 摄氏度----即上文提示的右侧两个橙色方框中之一。 看来这个显示的温度信息有问题，加上现在这样的表现，一定是 AG510K 出现异常了，而且多半就是温度方面过高造成的。 尝试重启电脑，看看 AG510K 是否能恢复正常，结果重启之后，在 BIOS 和设备管理面找不到 AG510K 了！ 完了，难道就这样挂了？！ 赶紧关机。

换用雷电：实测高温

拆机取下依然带有余温的 AG510K，查看外观无异常，等待其彻底冷却下来之后，将其安装到雷电 3 硬盘盒中，使用雷电 3 连接到微星 CreatorPro X17 进行测试。 出于破罐子破摔的心态，依然没有给 AG510K 安装散热器，看看到底如何。

图 6-124

同时，既然又拆机了，干脆顺便给微星 CreatorPro X17 的四个 M.2 SSD 全部填满。AG510K 使用雷电 3 连接之后，开机正常识别，微星 CreatorPro X17 正确显示了 5 个 SSD，并且都是 PCIe 连接，HWiNFO64 显示分别有 NVMe 4x/8.0 GT/s、NVMe 4x/16.0 GT/s 和 NVMe 4x/32.0 GT/s----但实际上，显示为 NVMe 4x 32.0 GT/s 的 AG510K 采用雷电 3 连接，只能工作在 PCIe 3.0/8.0 GT/s 状态。

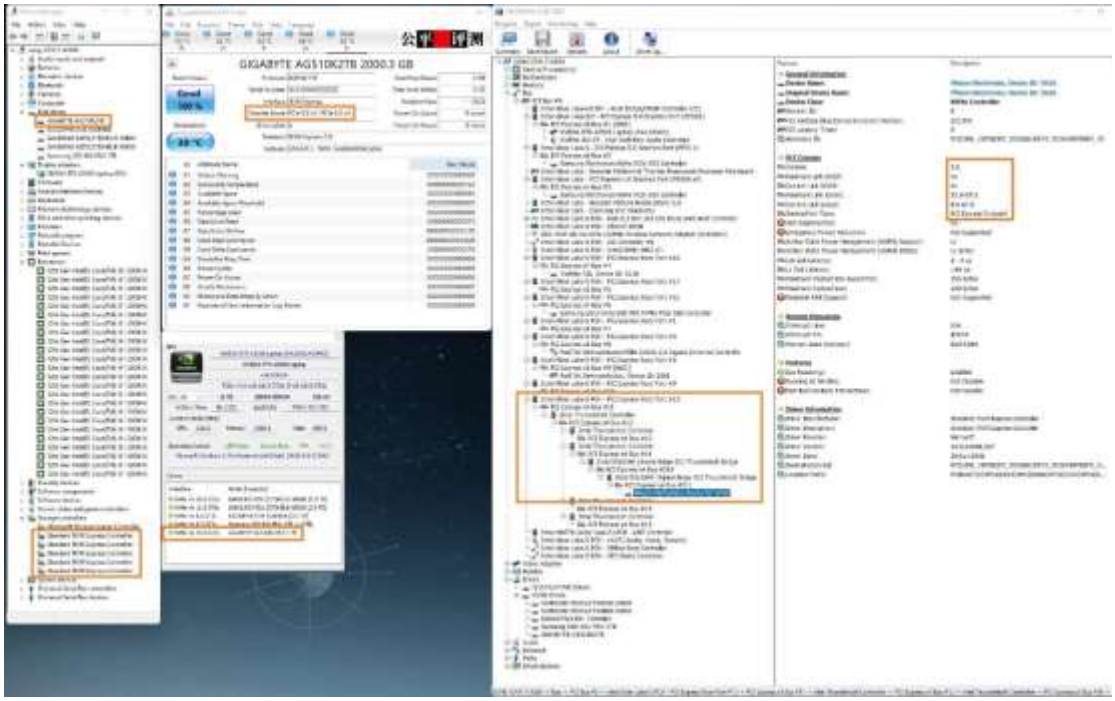


图 6-125

同样使用 CrystaDiskMark 进行测试，结果进度比第一次测试稍好，进行到顺序写入测试时，AG510K 才出现异常，CrystaDiskMark 自动中断了测试。 测试的成绩如下图所示：Q8T1 读取 2887.82 MB/s，Q8T1 顺序写入 1201.92 MB/s，和其他 SSD 在雷电 3 连接时的成绩差不多。但这当然不是 AG510K 的真实实力。 此时，HWiNFO64 显示的最高温度也只有 59 摄氏度，一般而言，SSD 温度不高于 60 摄氏度，不至于发生这样的异常现象。看来，雷电 3 连接之下，通过软件记录到的温度不一定准确。

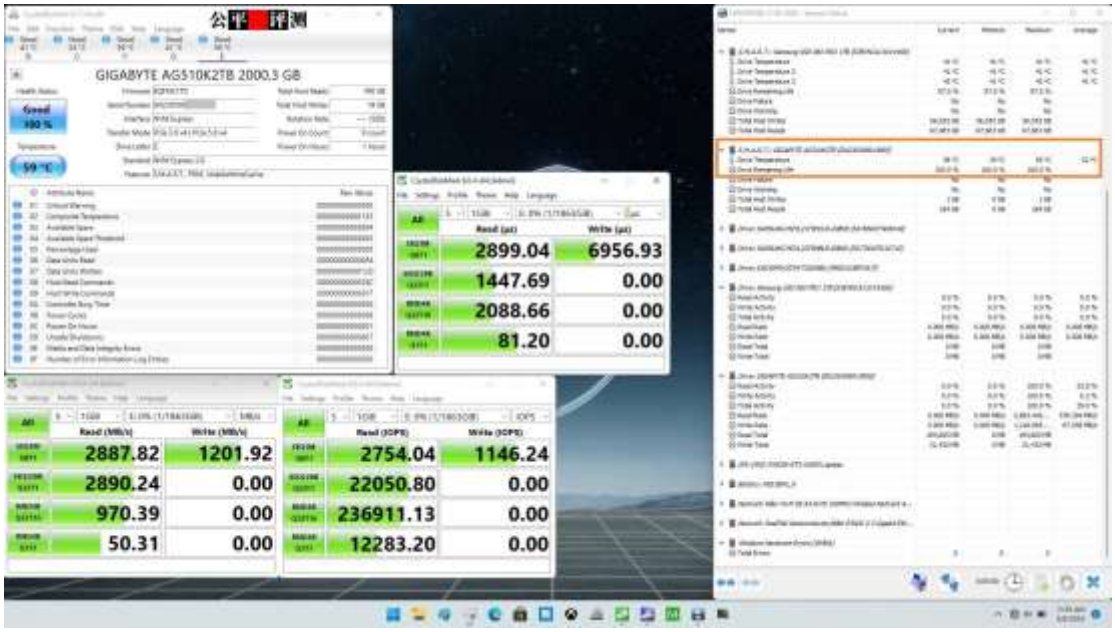


图 6-126

使用 Generic Log Viewer，对 HWiNFO64 记录的测试数据进行分析，得到下图. 可以看到在雷电 3 连接之下，没有安装散热器的 AG510K，CrystaDiskMark 前半部读取测试顺利完成，后半部写入测试只完成了 Q8T1 部分就停止了。 其起始温度为 37 摄氏度，最后温度一直维持在 59 摄氏度，这应该不是其真实温度。

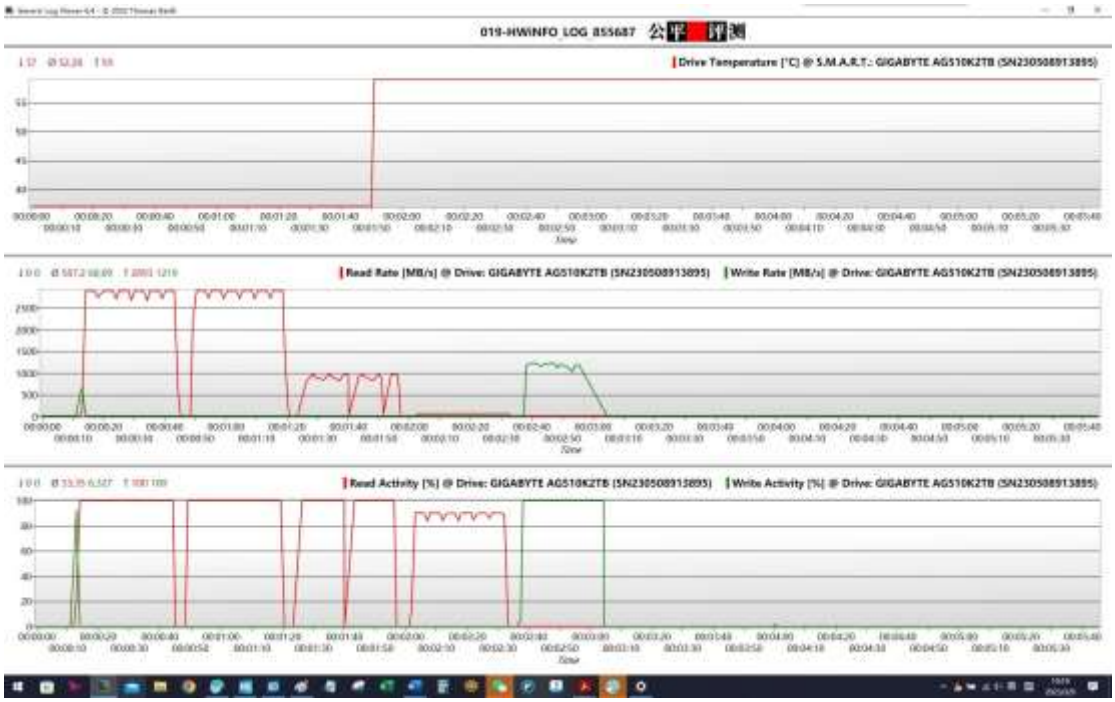


图 6-127

笔者在以上雷电 3 连接测试中，对安装在雷电 3 硬盘盒中的 AG510K 的外部温度进行了测量，记录到的其外表温度峰值达到 66.6 摄氏度，此时室温为 20 摄氏度左右，AG510K 直接裸露在空气中。

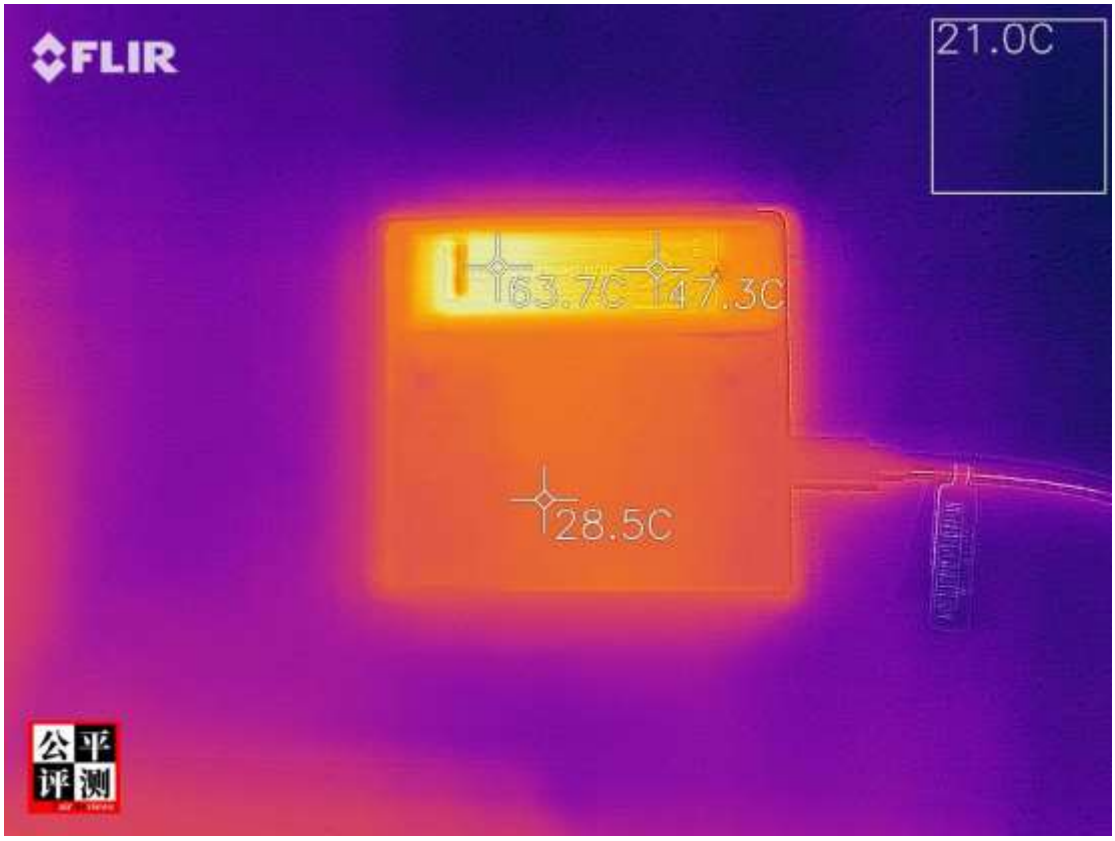


图 6-128

那么在雷电 3 硬盘盒中的 AG510K，安装上散热器又会如何？ 加上散热器之后其工作温度绝对要好，但是鉴于雷电 3 连接之下带宽不足，完全不能发挥 AG510K 的性能，这个测试意义不大----上面将其安装在雷电 3 中进行测试，笔者只是用来检测 AG510K 是否已经损坏而已。

图 6-129

通过以上测试，证实如下： 1.此致 AG510K 还没坏； 2.在全速运行之下，AG510K 的温度会很高，如没有有效散热，会出现异常甚至消失； 3.微星 CreatorPro X17 和雷电 3 硬盘盒，都不能采集到 AG510K 真实有效温度；

装上散热：温度大降

鉴于如果在微星 CreatorPro X17 上面安置 AG510K 的官方散热器，基本没法无损或稳定进行，笔者决定先将 AG510K 和官方散热器，安装到台式机上进行测试。 AG510K 的散热器的上部是金属铝散热板+两铜质热管和铝散热鳍片，下部是铝制保护底座，两者使用 4 颗十字螺丝锁定。 取下这 4 颗螺丝分离两者之后，可见两者和 SSD 接触面都有预制导热媒质，并使用了浅蓝色半透明塑料保护膜进行保护。

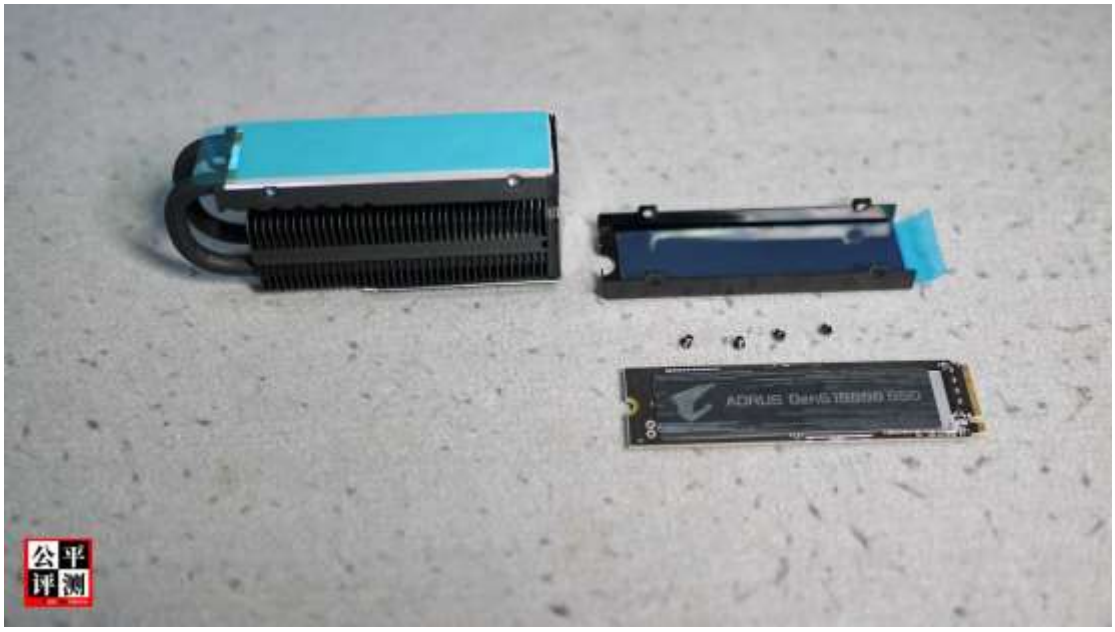


图 6-130

除去保护膜之后，可见上部的导热媒质为灰色，下部底座的导热媒质为灰黑色。

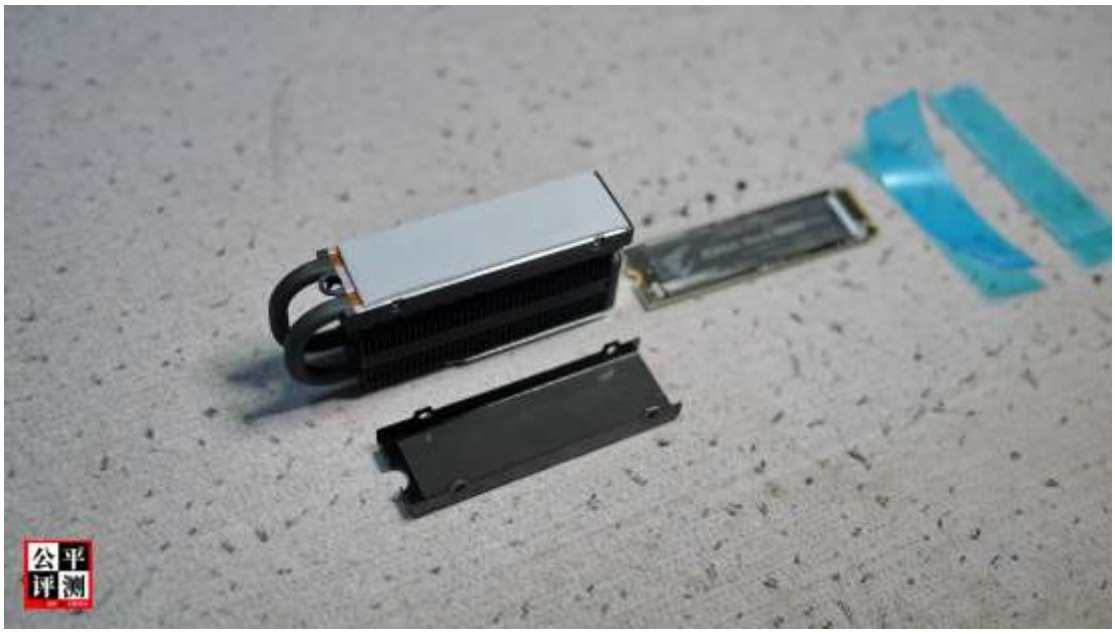


图 6-131

依据如下官方图片，官方宣称： 上部的灰色导热媒质带有“纳米碳涂层”，下部底座灰黑色的导热媒质为“双面高导热系统导热垫”。 使用之后，会“带来卓绝的无掉速性能”，“以保持全速顺序读写性能”。

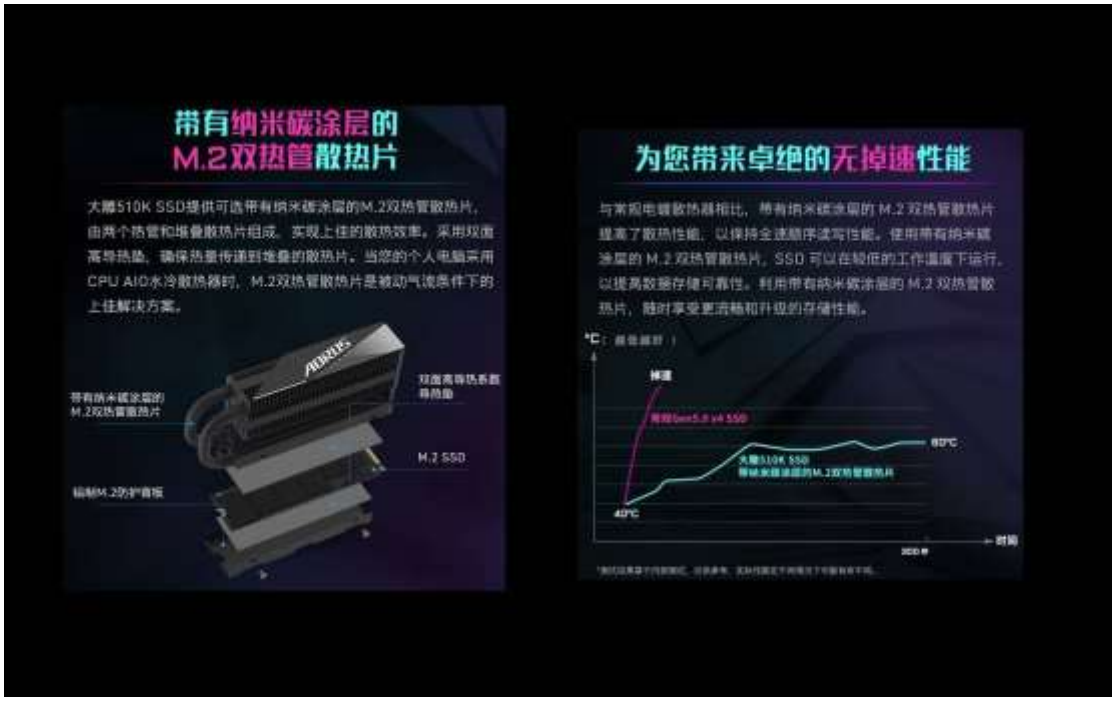


图 6-132

不过，由于条件有限，笔者测试使用的台式机是 Alienware Aurora R15----虽然它配备的是 i9-13900K+RTX 4090，但其 M.2 SSD 插槽，并不支持 PCIe 5.0 只支持 4.0。 但不管怎么，至少可以在 PCIe 4.0 之下、验证 AG510K 的性能和官方散热器的效果。

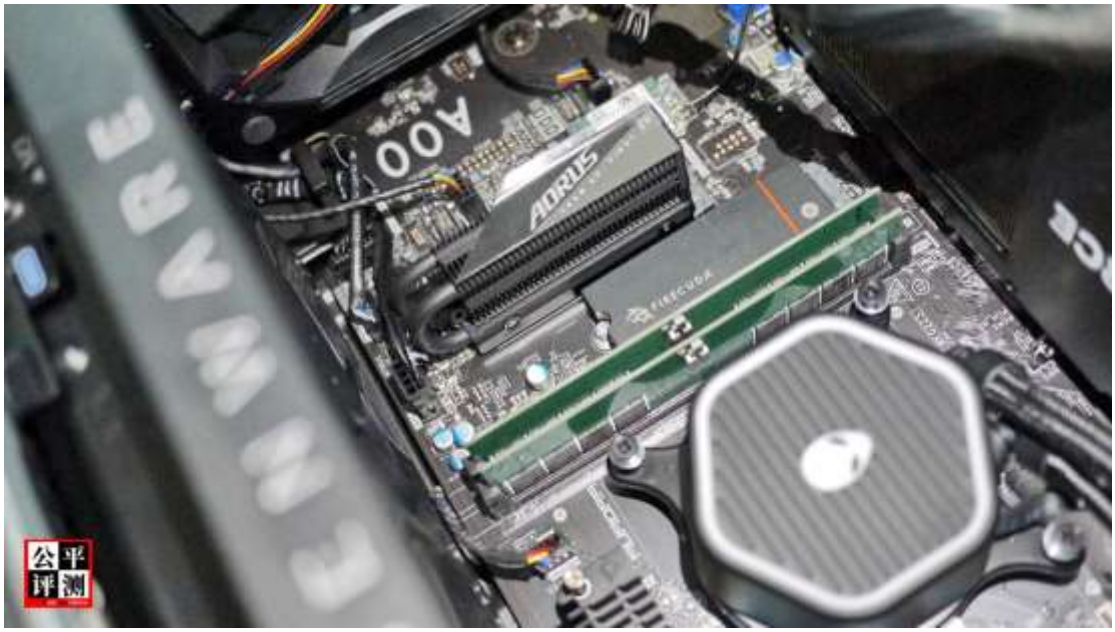


图 6-133

由于这次测试有了官方散热器的加持，CrystaDiskMark 和 AS SSD Benchmark 的均顺利完成测试，结果如下： Q8T1 顺序读写速率分别为 7094.78、6936.31 MB/s， Q32T16 4K 随机读写分别为 709.75、450.27 MB/s； 其他数据看图即可，不一一赘述。 AS SSD Benchmark 得分只有 3825 分。 没办法，这是在 PCIe 4.0 x4 连接 VMD 模式之下，

AG510K 的部分性能在理论上只发挥了一半。

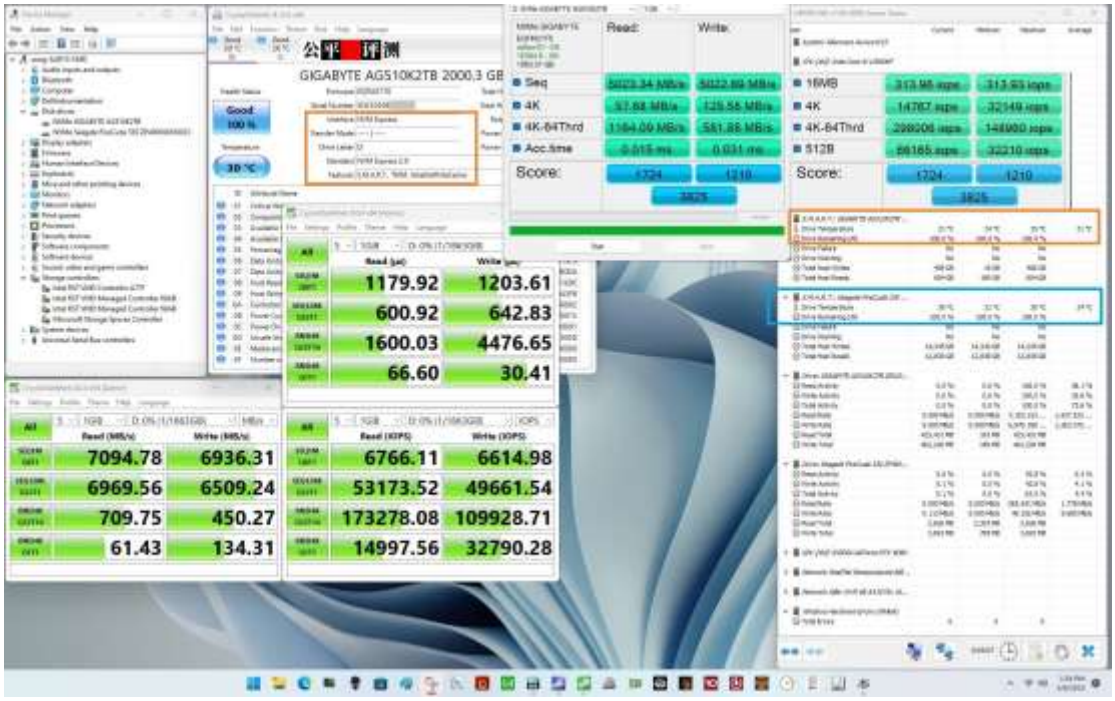


图 6-134

同时，上图也可以看到，AG510K 有了官方散热器的加持，橙色方框标注中的温度谷值仅为 24 摄氏度，峰值只有 35 摄氏度----官方散热器无愧其体积巨大，散热效果果然犀利！

同时，在 Alienware Aurora R15 上作为系统盘的 SSD，是 4TB 的希捷 FrieCuda 530 PCIe 4.0 SSD，它也有使用希捷官方散热器，其体积远远小于技嘉 AG510K 的散热器。其温度表现更好，蓝色方框标注中的温度谷值只有 21 摄氏度。 使用 Generic Log Viewer，对 HWiNFO64 记录的测试数据进行分析，得到下图.。 只能说明在这个 PCIe 4.0 连接测试中，AG510K 顺利完成测试，表现正常，PCIe 5.0 连接的行依然无法验证。

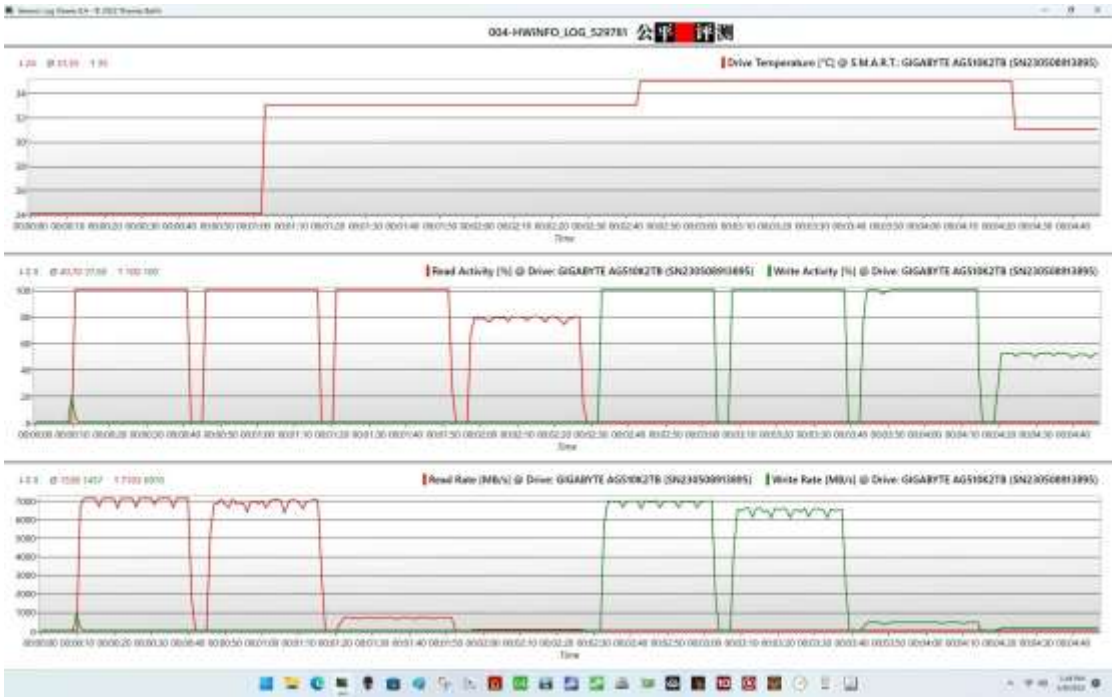


图 6-135

但是笔者手头没有技嘉官方的推荐主板，除了微星 CreatorPro X17 之外，再没有其他支持 PCIe 5.0 SSD 的电脑。

图 6-136

而支持 PCIe 5.0 SSD 的微星 CreatorPro X17，作为一台笔电，其空间确实无法在无损之下安装 AG510K 的官方散热器。 笔者测试了将微星 CreatorPro X17 的底盖取下，然后安装 AG510K 和散热器。

图 6-137

但经过仔细测量和试装，发现 AG510K 的原厂散热器的两根热管，会和微星 CreatorPro X17 显卡供电部分的散热模块发生冲突，造成散热器无法安装到位，强行安装的话有可能会造成 AG510K 电路板受压变形。 怎么办？ 绝不能放弃！ 经过笔者苦思冥想，翻箱倒柜，多次测量，最后诞生了一个的杂交解决方案：

解决方案：散热杂交

解决方案就是：搬出上文提到的 4TB 的希捷 FrieCuda 530 PCIe 4.0 SSD，将其希捷官方散热器的上半部取下来，用到 AG510K 之上。

图 6-138

希捷 FrieCuda 530 SSD 原厂散热器的上半部和 SSD 的接触面，同样有着淡蓝色的导热媒质，尽管其体积相对要小，更没有 AG510K 原厂散热器那样华丽的双热管和散热鳍片，但其金属材质绝对厚度也不低，应该有一定的热量缓冲和散热能力。

图 6-139

由于两者都是为标准 M.2 2280 SSD 的散热而生，所以尽管厂商、型号和设计都不同，但希捷 FrieCuda 530 SSD 原厂散热器的上半部，和技嘉 AG501K 原厂散热器的底座，在月老笔者的红线牵引之下，还是顺利地完美地结合在一起了----只是双方的 4 颗螺丝锁定空位对不上----而且结合的力度还不错，非常紧密无间。

图 6-140

接下来就简单了，使用希捷 FrieCuda 530 SSD 原厂散热器上半部的 AG501K 顺利完成安装，而微星 CreatorPro X17 的内部空间足够，对于这样厚度的散热器完全能够容纳，所以微星 CreatorPro X17 的底盖也能顺利安装复原。

图 6-141

使用 CrystaDiskMark 进行测试，终于完整完成测试，成绩也不俗： Q8T1 顺序读写速率，分别为 10086.77、10128.74 MB/s----这就是 PCIe 5.0 SSD 的实力展现！ 其他 ISO 和 us 不用赘述看图即可。 需要提示的是，安装这样的杂交散热器的 AG510K，在微星 CreatorPro X17 之上运行测试，不到 5 分钟其工作温度谷值为 47 摄氏度，峰值达到 74 摄氏度，均值为 62 摄氏度。

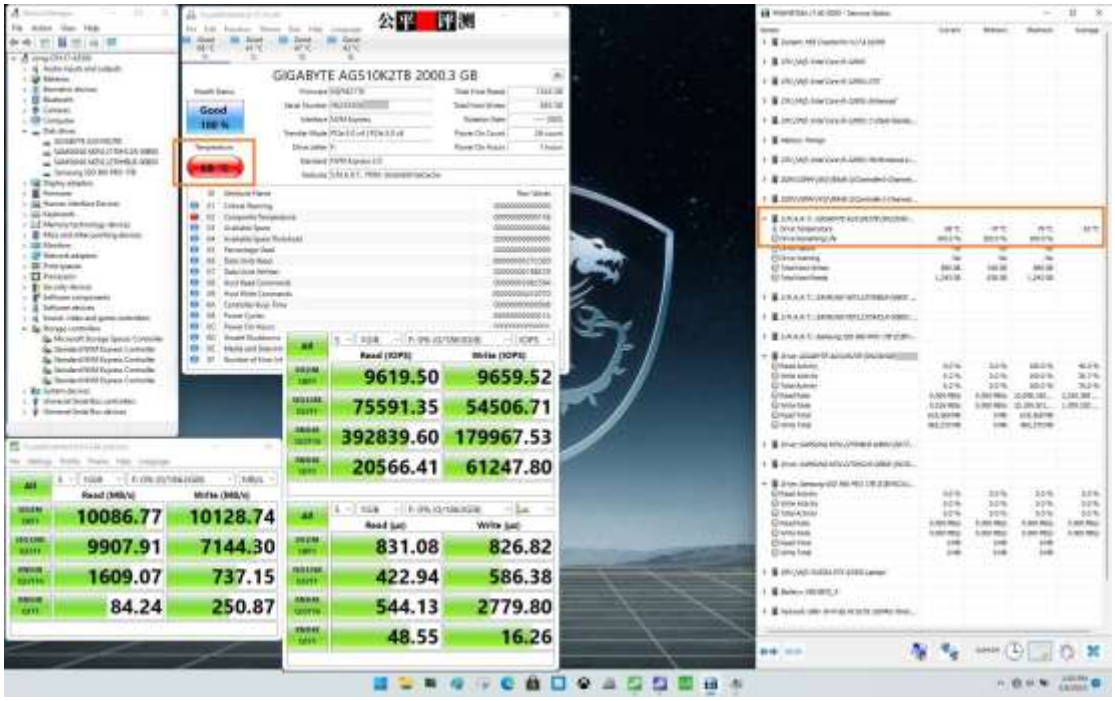


图 6-142

基于这样的工作温度，笔者特意进行了两次对比测试，测试均使用 AS SSD Benchmark 进行，第一次测试是微星 CreatorPro X17 开机闲置 10 分钟之后进行； 第二次测试是在第一次测试完成的 1 分钟之后马上进行。 得到的测试结果如下图所示，前后两次测试得分差距明显，分别是 9330 分和 8774 分。 仔细查看，会发现在读取方面差距不大，在写入方面发生了明显的性能下跌----这就是所谓的”SSD 过热掉速“现象。

图 6-143

最终评价：高能高温

AG510K 的官方技术规格参数如下图所示。 其使用的群联主控在以上多个截图中展现，2TB 容量标称的顺序读写速率也实测无误； 其待机功耗低于 85 毫瓦、读写功耗低于 11 瓦，笔者没法测量，但安装在微星 CreatorPro X17 能使用，说明其功耗还在笔电承受范围之内； 至于官方宣称的平均无故障时间 160 万小时……只能说：如果没有官方散热器，笔者购买的这个 AG510K，到目前为止启动不到 30 次，通电时间没超过 2 小时，就出现多次故障了。所以笔者觉得，官方在宣传上必须强调：**凡未正确安装使用官方原厂散热器者，所有性能参数和保修一律无效！**

图 6-144

附带提下： 当下上市的 M.2 2280 PCIe 5.0 SSD，使用的主控基本都是群联(PHISON)E26，镁光闪存颗粒。 比如，越南 SSTC 推出的 Tiger Shark”(虎鲨)，其型号简单粗暴----就是叫做”PHI-E262TB“，其散热器外观比技嘉还要高调，加上了 RGB-LED 光效，需要外部供电。

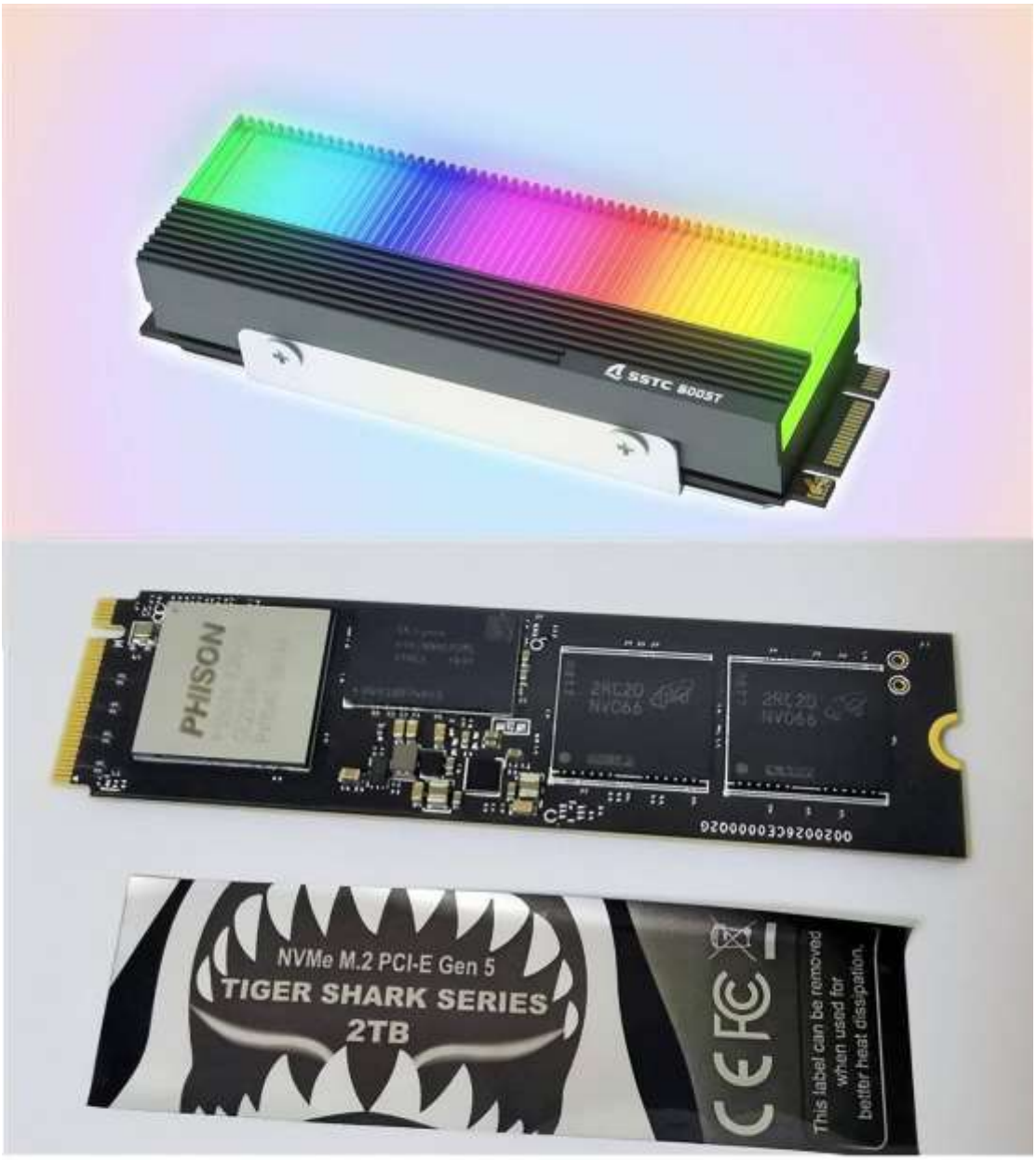


图 6-145

同一款 SSD，在不同平台之下性能有差异的问题，在最新的 PCIe 5.0 SSD 之上依然存在。 下面是 PHI-E262TB 分别在 intel 和 AMD 平台上的测试结果，在 RND4K 之时出现了明显的性能差异。 笔者当下没有支持 PCIe 5.0 SSD 的 AMD 平台的电脑，所以无法进行这样的对比测试。

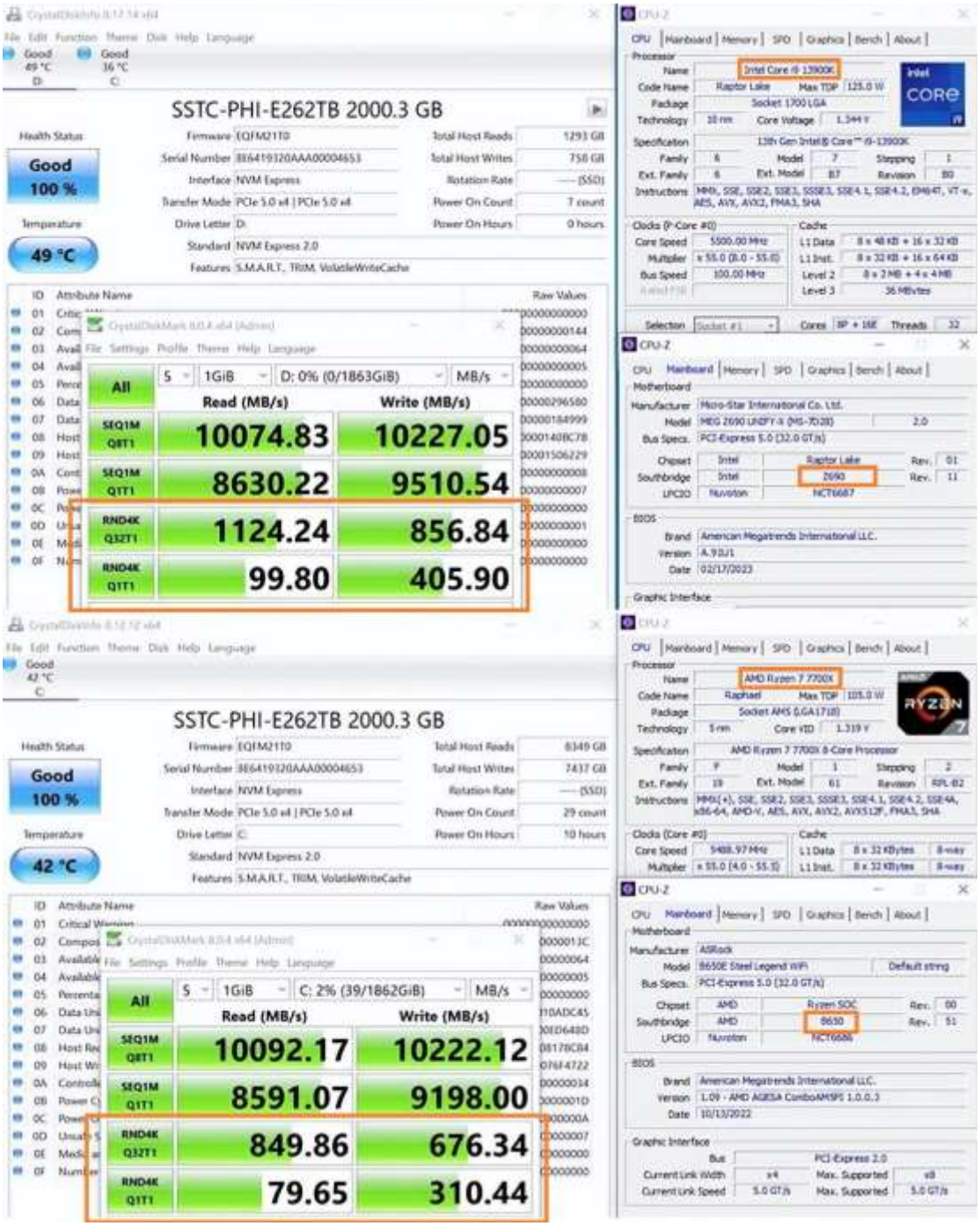


图 6-146

鉴于以上测试结果，笔者断定，就目前而言，此 AG510K 2TB M.2 2280 PCIe 5.0 SSD，无法安装在微星 CreatorPro X17 长期高负荷使用； 这不是微星 CreatorPro X17 的问题；也不是技嘉 AG510K 的问题；而是笔者的问题 —— 笔者企图在 M.2 2280 PCIe 5.0 SSD 刚刚面世之时，就将其使用到笔记本电脑之上。或者说，和技嘉 AG510K 一样，当下所有 M.2 2280 PCIe 5.0 SSD，离开散热器都无法正常运行！**笔电要早日普及 M.2 2280 PCIe 5.0 SSD，还有待诸君努力！**

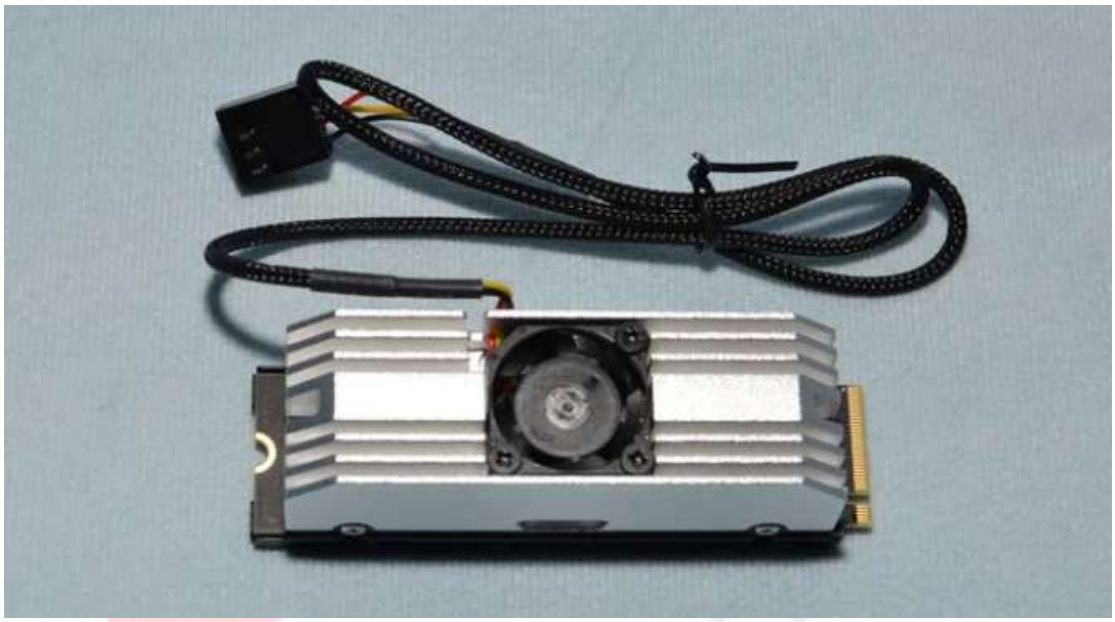
6.2.3.2.3 Phison E26 Max14um 2TB SSD 测试报告_2024 – Quarch PPM 测试能效比，平均功耗以及 ASPM 功耗！

This is the fastest SSD we've ever tested — Phison E26 Max14um 2TB performance preview

By Shane Downing

published January 13, 2024

Phison delivers its 14 GB/s reference design SSD.



PHISON MAX14UM INTRODUCTION

When we first looked at Phison's Max14um reference design, so-named because it was targeting the 14 GB/s performance point, we had some questions. Can it really achieve even higher levels of performance out of a PCIe 5.0 platform? Phison assured us that it honed its design to meet the original desired specifications while still operating within reasonable power limits. We expect multiple manufacturers to come out with their own versions of the design to compete with the best SSDs — particularly with custom cooling solutions, acknowledging that, yes, this level of performance is possible without any insane trade-offs, even if some of the heatsinks make it seem otherwise.

The ostensible lure of the Max14um design is that it can provide higher levels of bandwidth within the same nominal M.2 power limit. In fact, having to design around that power limit means that the reference model is potentially faster in other areas than the original drives as well, due to varied optimizations.

Make no mistake: Power consumption and thermal output are still notably high, but this

drive feels closer to what Phison first laid out a year ago. We are impressed with its across-the-board performance, including sustained write speeds. These upcoming SSDs are absolutely worth a look for the storage enthusiast.

MAX14UM SPECIFICATIONS

Product	1TB	2TB	4TB
Form Factor	M.2 2280-D2	M.2 2280-D2	M.2 2280-D2
Interface / Protocol	PCIe 5.0 x4 / NVMe 2.0	PCIe 5.0 x4 / NVMe 2.0	PCIe 5.0 x4 / NVMe 2.0
Controller	Phison E26	Phison E26	Phison E26
DRAM	LPDDR4	LPDDR4	LPDDR4
Flash Memory	232-Layer Micron TLC	232-Layer Micron TLC	232-Layer Micron TLC
Sequential Read	13,000 MB/s	14,000 MB/s	14,000 MB/s
Sequential Write	9,500 MB/s	12,000 MB/s	12,000 MB/s
Random Read	1,300K	1,500K	1,500K
Random Write	1,400K	1,600K	1,600K
Security	TCG OPAL 2.0	TCG OPAL 2.0	TCG OPAL 2.0
Endurance (TBW)	700TB	1,400TB	3,000TB
Active Power	10.7W	11W	TBD
Warranty	5-Year	5-Year	5-Year

Drives built on the Max14um platform will be available in capacities of 1TB, 2TB, and 4TB. While the exact specifications vary, the reference design is rated for up to 14,000 / 12,000 MB/s for sequential reads and writes and up to 1500K / 1600K random read/write IOPS. Drive warranty may also vary, but the standard warranty will be for five years and up to 700TB of writes per TB of capacity. The Phison E26 is built to support TCG OPAL

2.0 for hardware encryption, but this feature is optional.

The default mean time between failure (MTBF) rating is for 1.6 million hours, derived via reliability simulation — which can be largely discounted as being relevant for consumers. Although active power is not officially listed for the 4TB capacity, we know that Phison is bumping up against the M.2 11.555W average power limit on these drives. There are no other surprises here.

SOFTWARE AND ACCESSORIES

Software as provided by individual drive manufacturers will vary, but most of the prominent ones offer SSD toolboxes for basic functionality and optionally OEM software for imaging and cloning. There are free options for the majority of features, but the presence of a toolbox for drive sanitization and firmware updates is preferred. Phison provides the mechanism for firmware updates and it is up to the manufacturer to test and provide the relevant updates. For this reference drive, no software was provided.

A CLOSER LOOK AT THE PHISON MAX14UM REFERENCE DRIVE

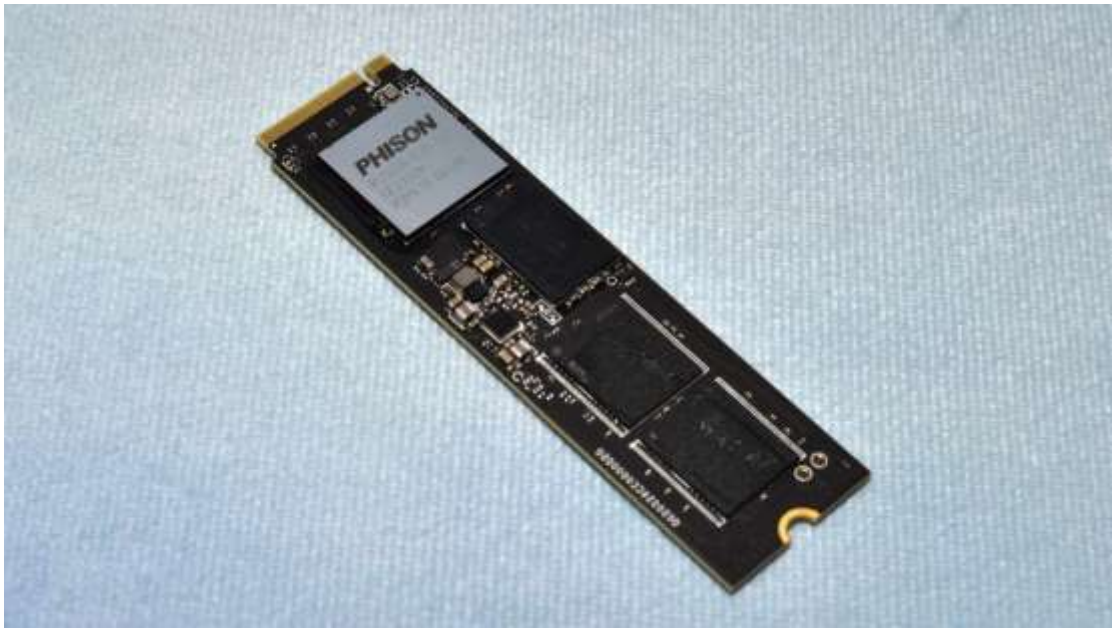
(Image credit: Tom's Hardware)

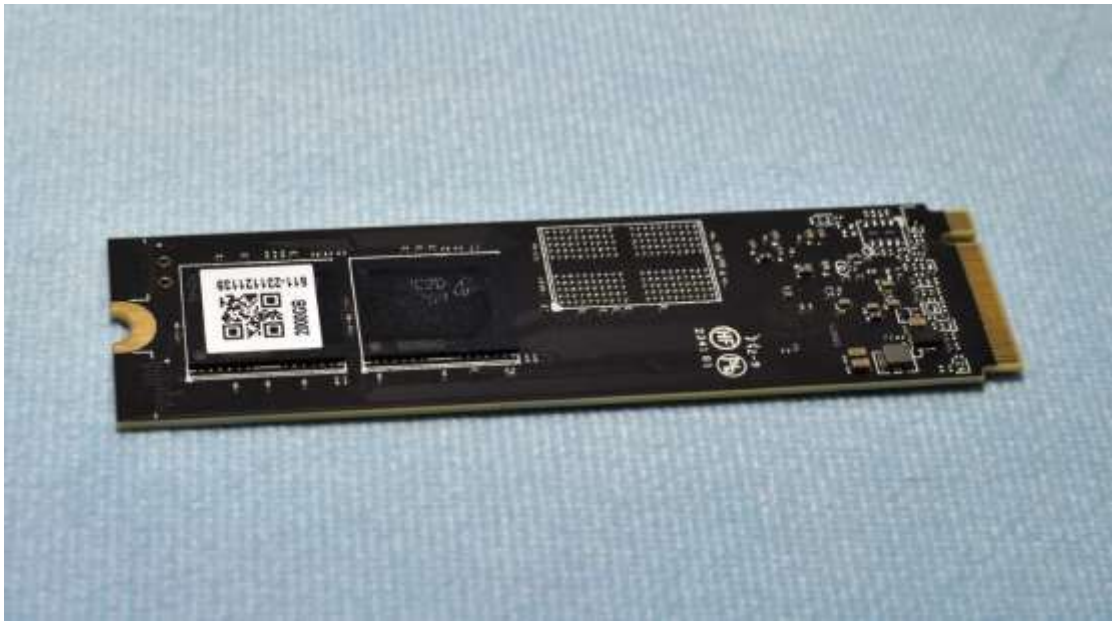
As is well-known by now, these Phison E26 PCIe 5.0 SSDs need a significant amount of cooling, and a heatsink is required for proper operation. The reference design also includes active cooling via a small fan, powered by a 4-pin fan connector with PWM support. In testing, we generally have found that you can get away without using the fan on these drives in a PC with decent airflow. In some extreme cases like sustained workloads, thermal throttling might occur without the fan, though it's audible and has an

annoyingly high pitch.

Non-reference designs from some manufacturers range from dual fans, as with the PNY XLR8 CS3150, to the MSI Spaptium M580's elaborate hybrid cooling solution. There's also the Patriot Viper VP573 with a low-profile blower design. Other drives, like the Sabrent Rocket 5, arrive bare or at least provide the option to purchase a bare drive, with the assumption that you will provide your own after-market solution. This could be cooling provided by the motherboard, or something like Thermalright's HR-09 Pro.

One active but quiet SSD cooling solution we saw at CES 2024 was the AirJet Mini, which can be configured to cool PCIe 5.0 drives. Micron was showing off four Phison E26 drives running in RAID, which provided sustained read speeds of over 40 GB/s. More development in this area is anticipated.





(Image credit: Tom's Hardware)

While we are now very familiar with the Phison E26 controller, having reviewed over ten drives with capacities up to 4TB, it's technically not the only PCIe 5.0 platform that will be made available. Phison revealed the E31T's specifications at CES, but that hardware is not intended to compete at the high-end. For that we have SMI's SM2508, demonstrated in the Adata Project NeonStorm SSD, and InnoGrit's IG5666, shown in the Teamgroup T-Force GE Pro. The former is expected to use TSMC's 6nm process node and should usher in some thermal relief versus the current Phison E26 12nm designs. We expect to see the Maxio MAP1802 at some point, too, also at 6nm.

There are also hybrid PCIe 5.0 drives possible, as the Samsung 990 EVO will run in both x4 PCIe 4.0 and x2 PCIe 5.0 modes. Its listed performance puts it in line with budget PCIe 4.0 SSDs such as the WD Black SN770 or WD Blue SN580, IG5220-based drives like the Patriot P400, E21T-based like the Silicon Power UD90, and SM2269XT-based like the Solidigm P41 Plus. Some MAP1602 drives also fall into this category. The 990 EVO's bandwidth is not sufficient to really put it into the PCIe 5.0 category, but it's an interesting development that's worth watching, particularly as a near-term solution for laptops.

Moving along, the Max14um reference design is double-sided in the M.2 2280-D2-M form factor with a maximum of four NAND flash packages. This form factor is designed for a maximum height of 3.5mm, excluding any heatsinks. The drive's intended capacity range is currently up to 4TB, which is achieved with 32 dies but only 16CE (chip enabled), the same as with the 2TB configuration. This certainly leaves room for 8TB, but that will be a difficult capacity to achieve at this level of performance given the M.2 power limitations.

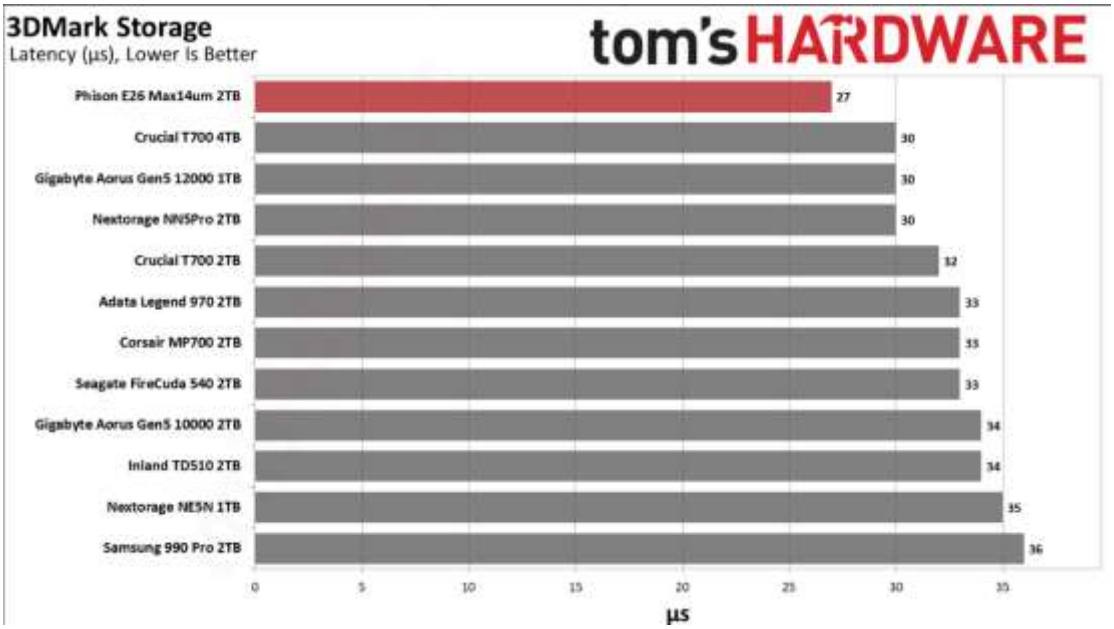
PHISON MAX14UM COMPARISON PRODUCTS

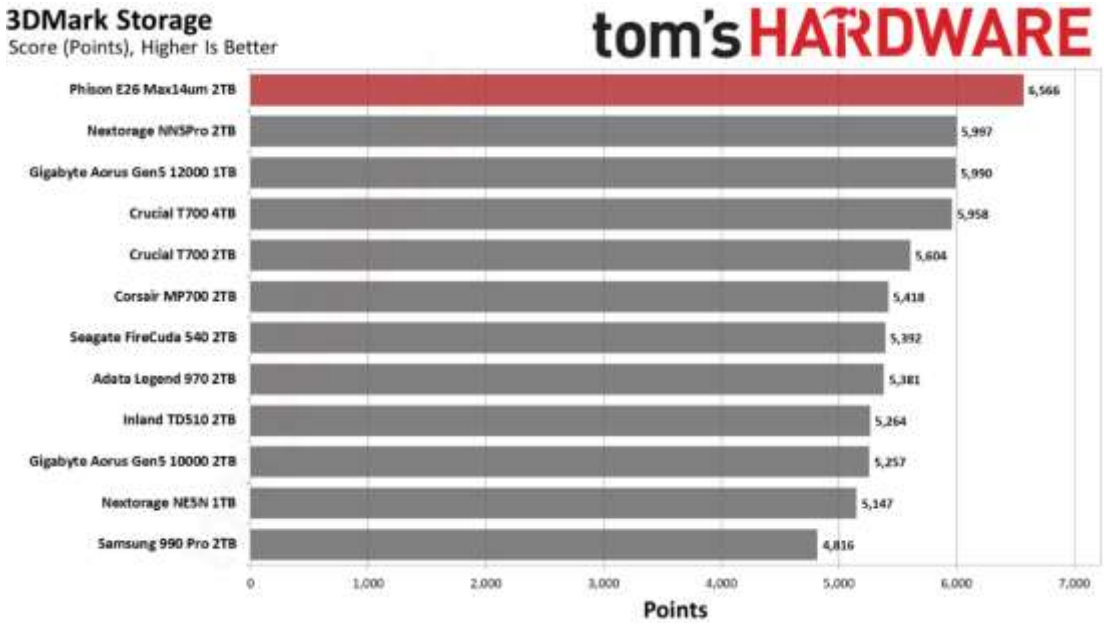
With the Max14um we are mostly interested in how it performs against existing PCIe 5.0 SSDs, but we've included the Samsung 990 Pro PCIe 4.0 as a comparison point — a drive that we've also reviewed at 4TB. Current PCIe 5.0 drives all use the Phison E26 controller, but variants start with 10 GB/s drives like the Seagate FireCuda 540, Nextorage NE5N, Inland TD510, Corsair MP700, and Gigabyte Aorus Gen 5 10000. Newer drives perform at up to 12 GB/s and include the Crucial T700, which we reviewed separately at 4TB, Nextorage NN5Pro, Gigabyte Aorus Gen5 12000, and Adata Legend 970.

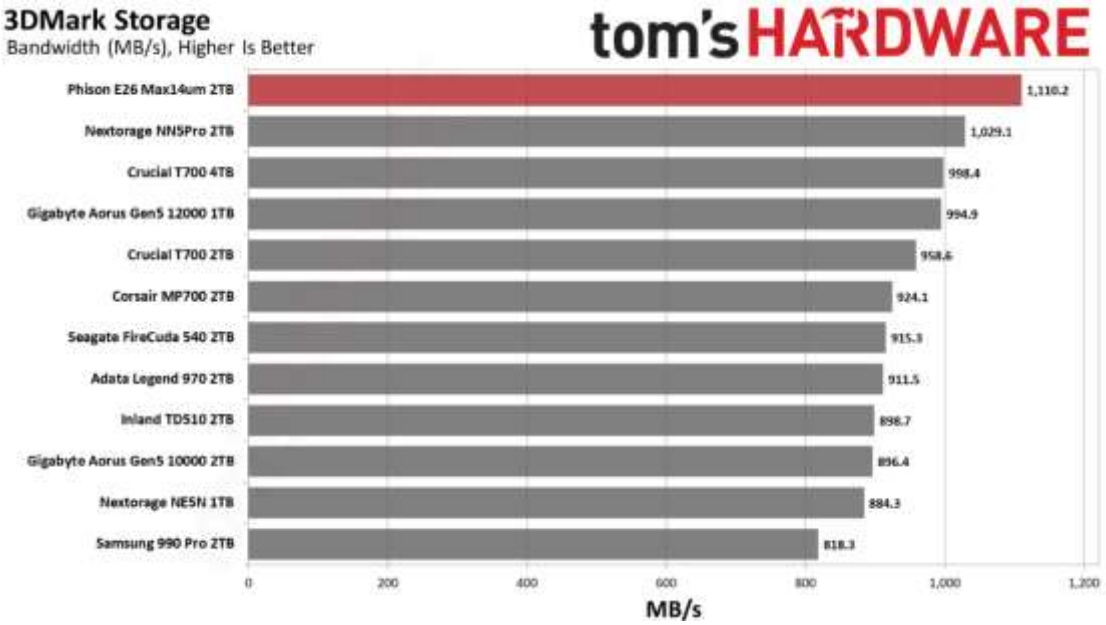
This is the first taste we've had of Phison E26 equipped with Micron 2400 MT/s NAND, and the only PCIe 5.0 drives currently shipping all have the same basic hardware. However, as noted above, we saw several PCIe 5.0 drives at CES 2024 using other controllers, so we could finally see some added competition in this segment in the near future.

TRACE TESTING — 3DMARK STORAGE BENCHMARK

Built for gamers, 3DMark's Storage Benchmark focuses on real-world gaming performance. Each round in this benchmark stresses storage based on gaming activities including loading games, saving progress, installing game files, and recording gameplay video streams. Future gaming benchmarks will be DirectStorage-inclusive and we include details of that where possible.





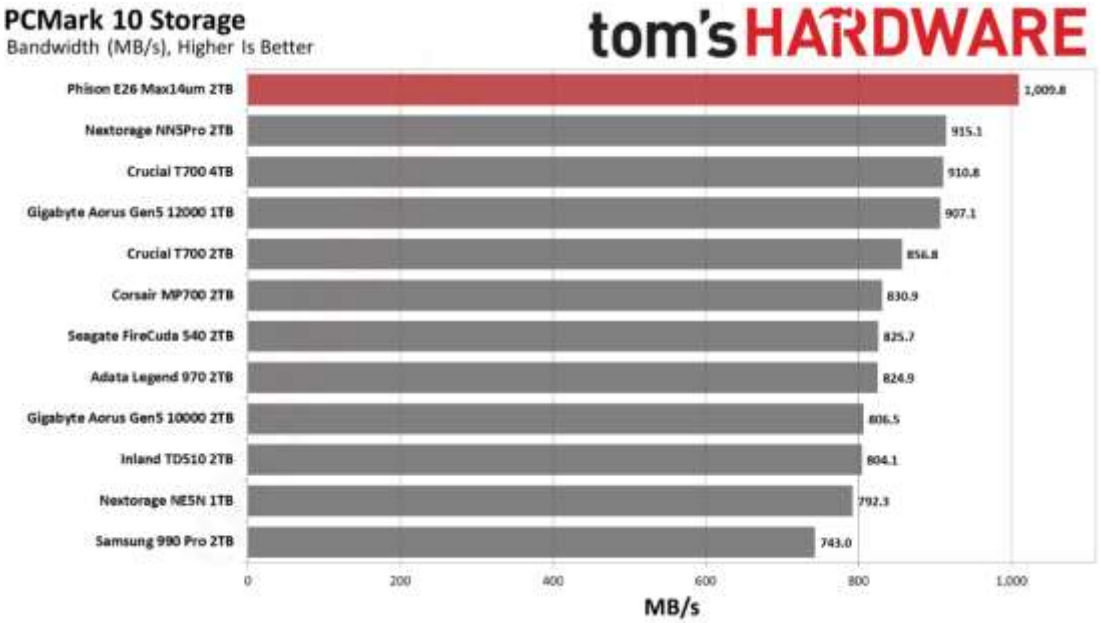


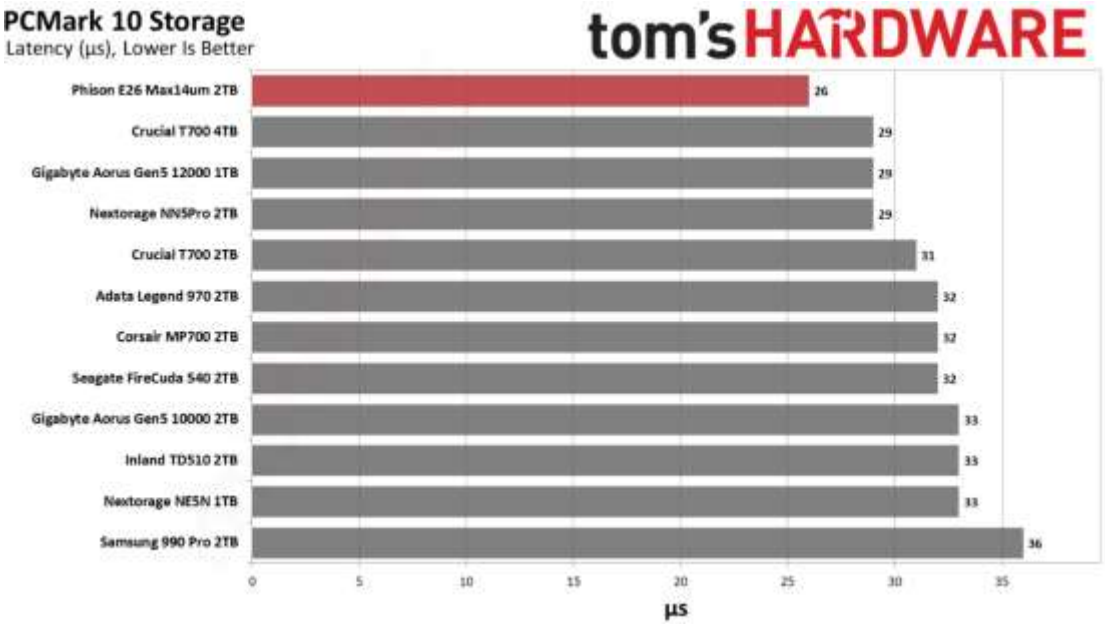
(Image credit: Tom's Hardware)

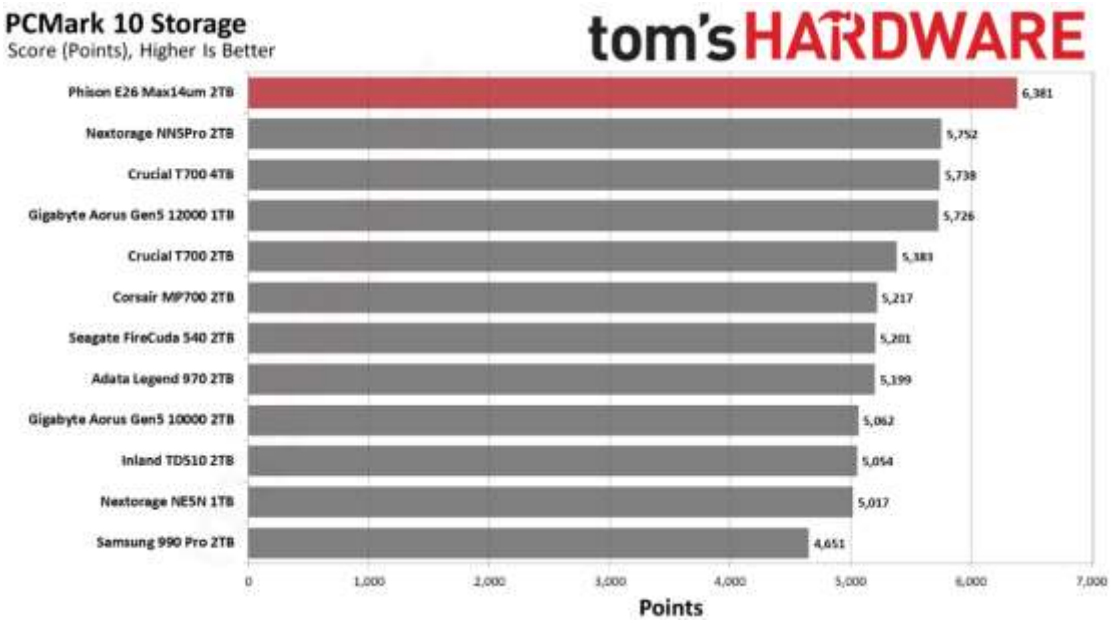
The Max14um sets a record in 3DMark with the best scores to date. Its firmware is DirectStorage-optimized so drives built on it will be a beast for both current and future game titles. Overall performance is nearly 10% higher than the next closest E26 12 GB/s drive.

TRACE TESTING — PCMARK 10 STORAGE BENCHMARK

PCMark 10 is a trace-based benchmark that uses a wide-ranging set of real-world traces from popular applications and everyday tasks to measure the performance of storage devices.


PCMark 10 Storage
Bandwidth (MB/s), Higher Is Better
tom's HARDWARE
Phison E26 Max14um 2TB
Nextorage NN5Pro 2TB
Crucial T700 4TB
Gigabyte Aorus Gen5 12000 1TB
Crucial T700 2TB
Corsair MP700 2TB
Seagate FireCuda 540 2TB
Adata Legend 970 2TB
Gigabyte Aorus Gen5 10000 2TB
Inland TD510 2TB
Nextorage NE5N 1TB
Samsung 990 Pro 2TB
1,009.8
915.1
910.8
907.1
856.8
830.9
825.7
824.9
806.5
804.1
792.3
743.0
0
200
400
600
800
1,000
MB/s


PCMark 10 Storage
Latency (µs), Lower Is Better
tom's HARDWARE
Phison E26 Max14um 2TB
Crucial T700 4TB
Gigabyte Aorus Gen5 12000 1TB
Nextorage NN5Pro 2TB
Crucial T700 2TB
Adata Legend 970 2TB
Corsair MP700 2TB
Seagate FireCuda 540 2TB
Gigabyte Aorus Gen5 10000 2TB
Inland TD510 2TB
Nextorage NE5N 1TB
Samsung 990 Pro 2TB
26
29
29
29
31
32
32
32
33
33
33
36
0
5
10
15
20
25
30
35
µs

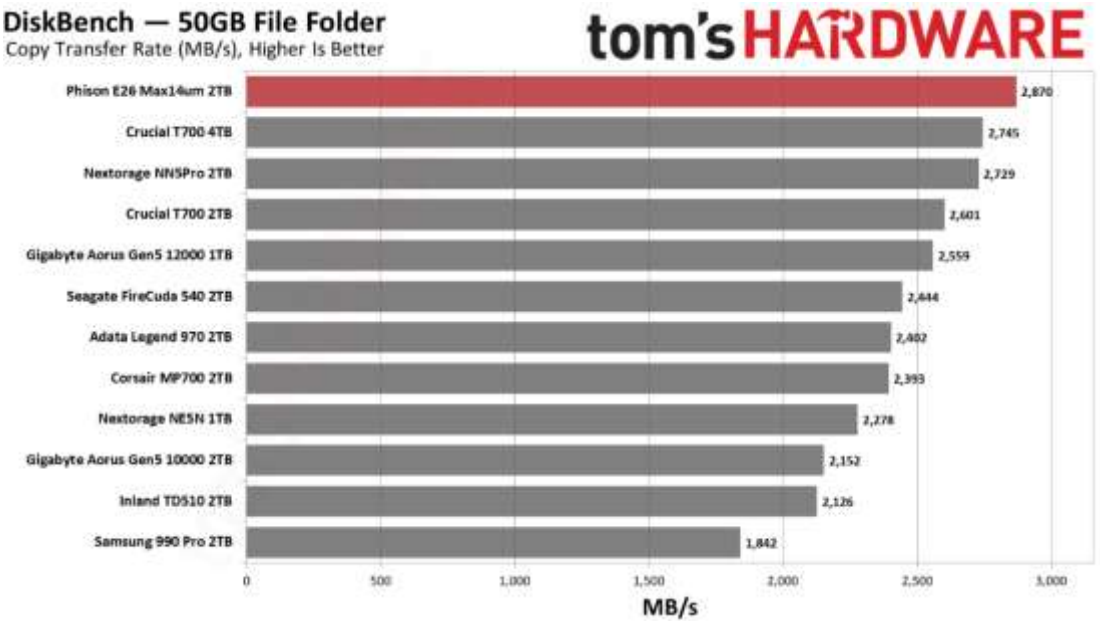


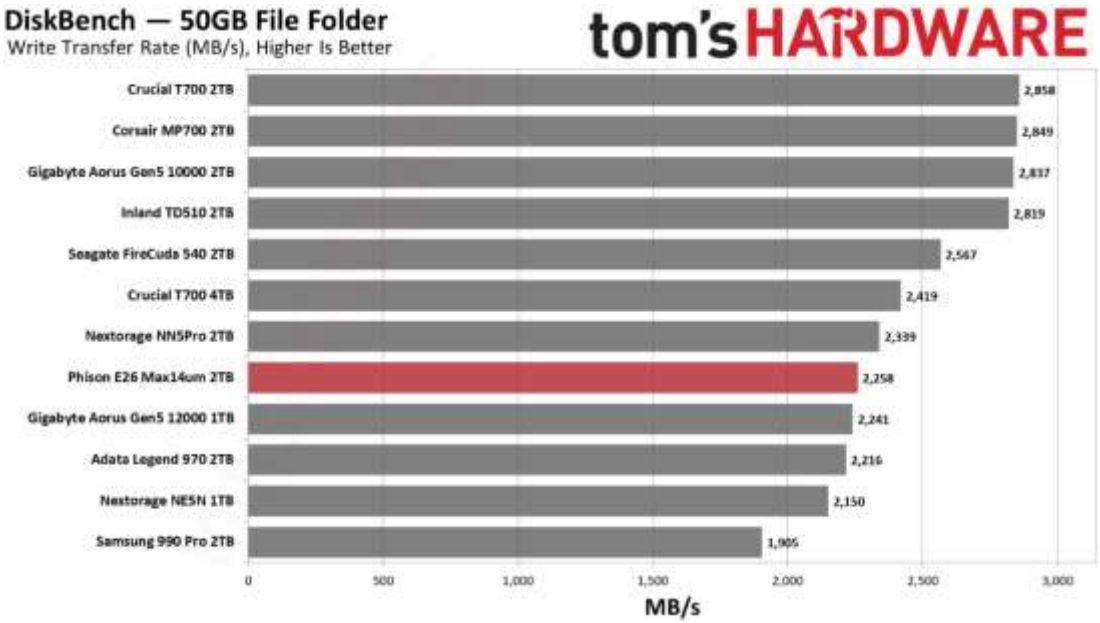
(Image credit: Tom's Hardware)

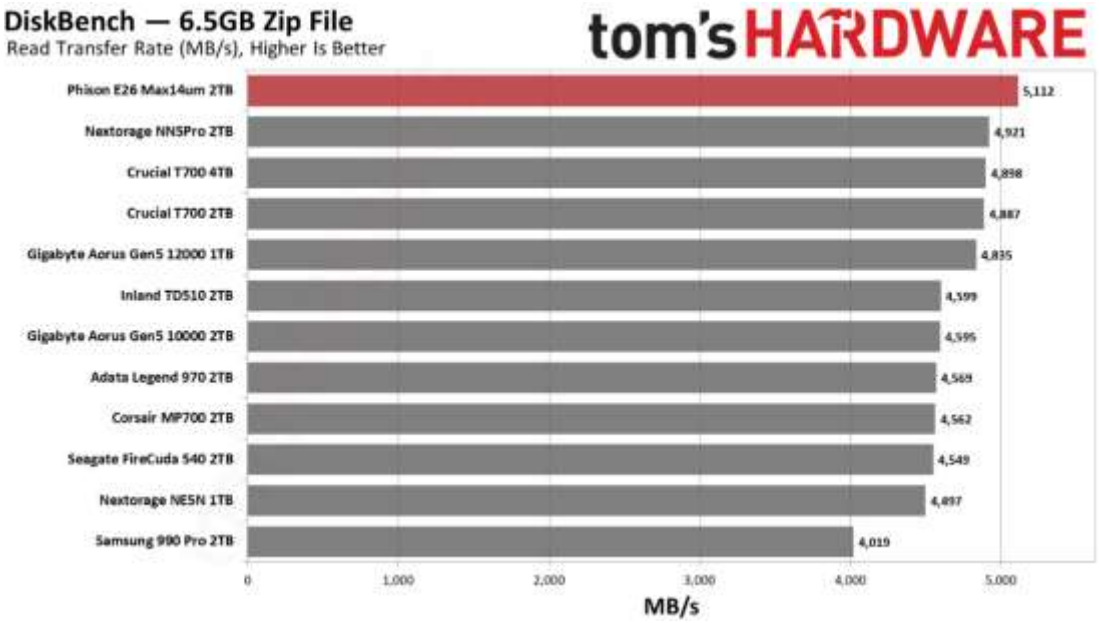
The Max14um sets new records in PCMark 10 as well, breaking the 1,000 MB/s bandwidth result for the first time with a single drive. Burst performance doesn't get better than this, although the translation to real world applications is questionable. You need a high-end system with specific workloads to take advantage.

DISKBENCH TRANSFER RATES

We use the DiskBench storage benchmarking tool to test file transfer performance with a custom, 50GB dataset. We write 31,227 files of various types, such as pictures, PDFs, and videos to the test drive. Then we copy all of that data to a new folder, and follow-up with a reading test of a newly-written 6.5GB zip file. This last test represents a real world type workload that fits into the cache of most drives.


DiskBench — 50GB File Folder
Copy Transfer Rate (MB/s), Higher Is Better
tom's HARDWARE
Phison E26 Max14um 2TB
Crucial T700 4TB
Nextorage NN5Pro 2TB
Crucial T700 2TB
Gigabyte Aorus Gen5 12000 1TB
Seagate FireCuda 540 2TB
Adata Legend 970 2TB
Corsair MP700 2TB
Nextorage NE5N 1TB
Gigabyte Aorus Gen5 10000 2TB
Inland TD510 2TB
Samsung 990 Pro 2TB
2,870
2,745
2,729
2,601
2,559
2,444
2,402
2,393
2,278
2,152
2,126
1,842
0
500
1,000
1,500
2,000
2,500
3,000
MB/s


DiskBench — 50GB File Folder
Write Transfer Rate (MB/s), Higher Is Better
tom's HARDWARE
Crucial T700 2TB
Corsair MP700 2TB
Gigabyte Aorus Gen5 10000 2TB
Inland TD510 2TB
Seagate FireCuda 540 2TB
Crucial T700 4TB
Nextorage NN5Pro 2TB
Phison E26 Max14um 2TB
Gigabyte Aorus Gen5 12000 1TB
Adata Legend 970 2TB
Nextorage NE5N 1TB
Samsung 990 Pro 2TB
2,858
2,849
2,837
2,819
2,567
2,419
2,339
2,258
2,241
2,216
2,150
1,905
0
500
1,000
1,500
2,000
2,500
3,000
MB/s



(Image credit: Tom's Hardware)

The Max14um's copy performance is unmatched. It gets an extra push from its higher bandwidth cap, but it's not a huge amount better than drives like the T700 and NN5Pro. You will see a larger benefit in comparison to Gen 3 and Gen 4 SSDs.

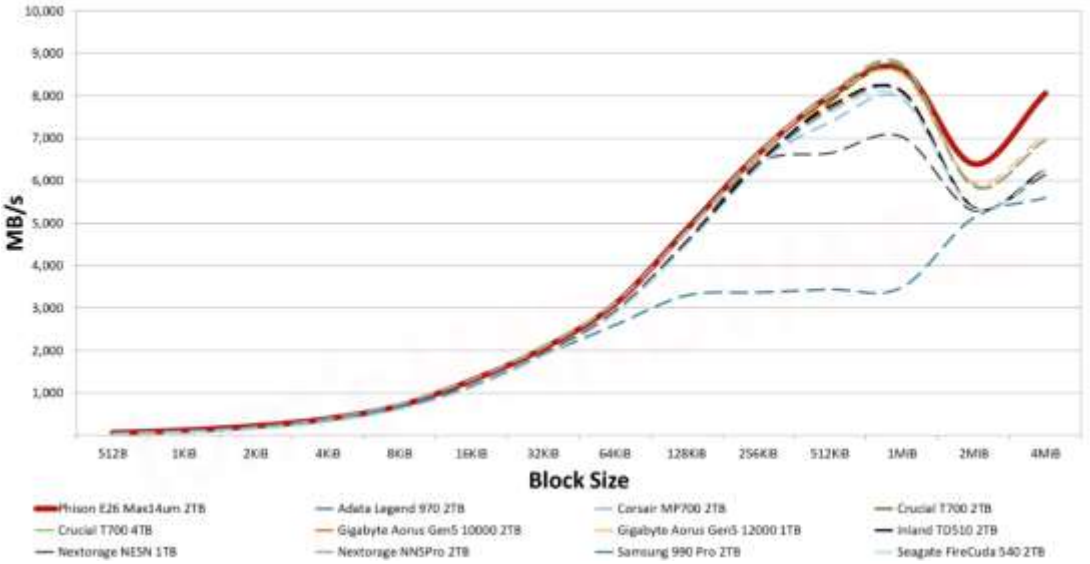
We generally ignore the write transfer rate for now, as the source PCIe 4.0 drive can be a bottleneck. While we had a few drives hit 2.8 GB/s in previous testing, most drives now max out at 2.2–2.3 GB/s.

ATTO AND CRYSTALDISKMARK

ATTO and CrystalDiskMark (CDM) are free and easy-to-use storage benchmarking tools that SSD vendors commonly use to assign performance specifications to their products. Both of these tools give us insight into how each device handles different file sizes and at different queue depths.

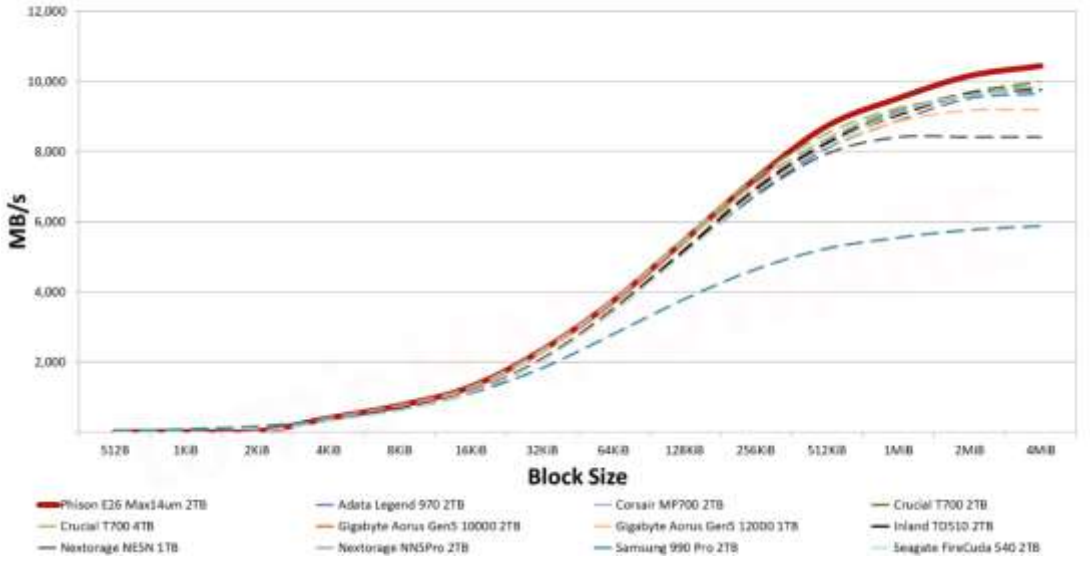
ATTO Disk Benchmark
SSD Read Throughput at QD1 and Various Block Sizes

tom's HARDWARE



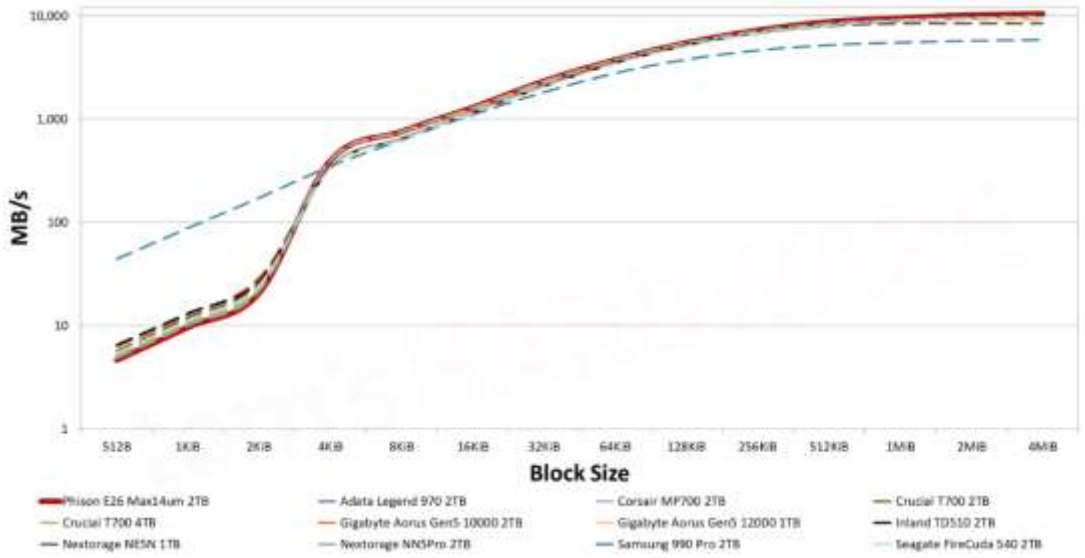
ATTO Disk Benchmark
SSD Write Throughput at QD1 and Various Block Sizes

tom's HARDWARE



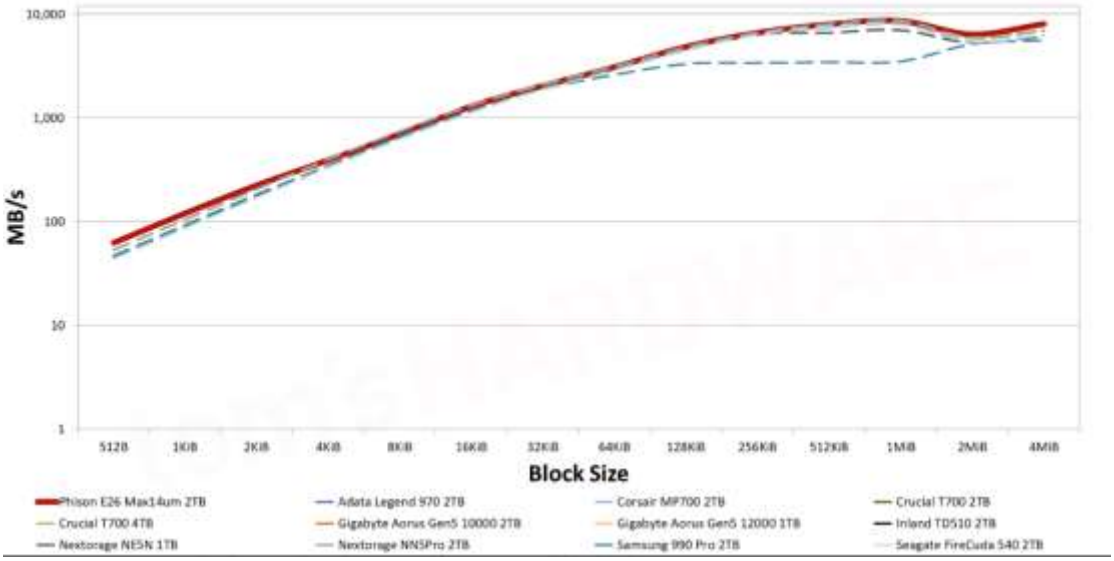
ATTO Disk Benchmark
SSD Write Throughput at QD1 and Various Block Sizes

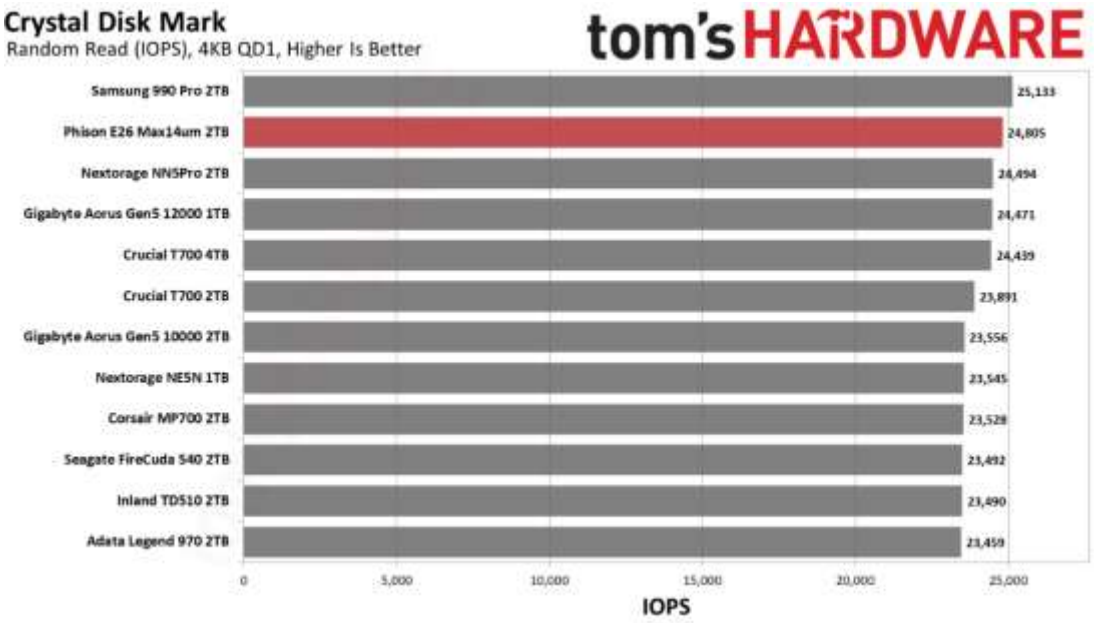
tom's HARDWARE

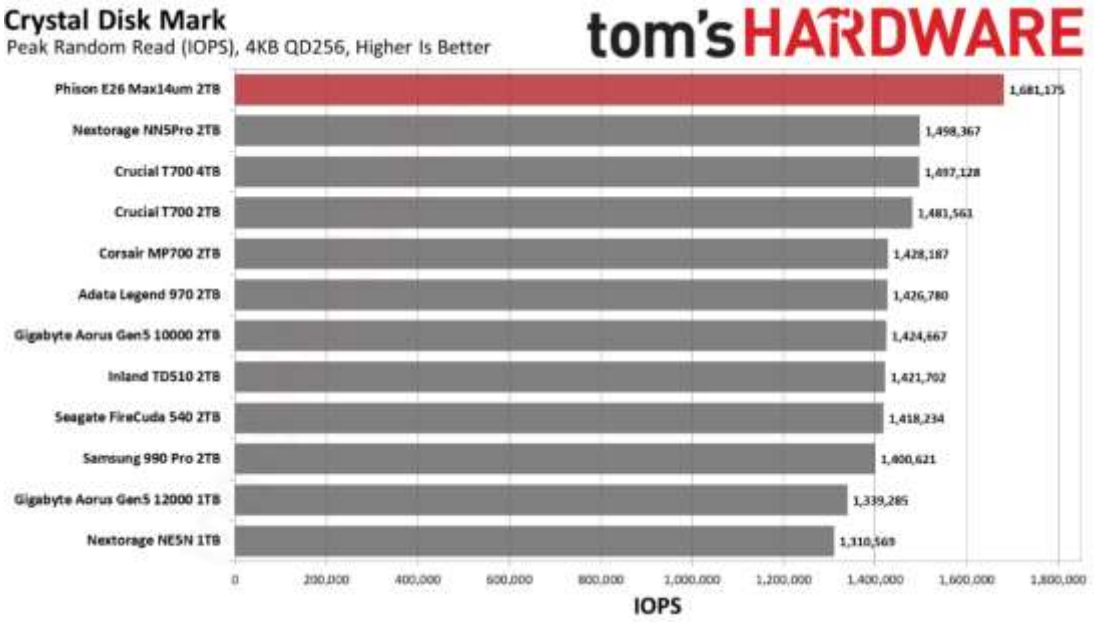


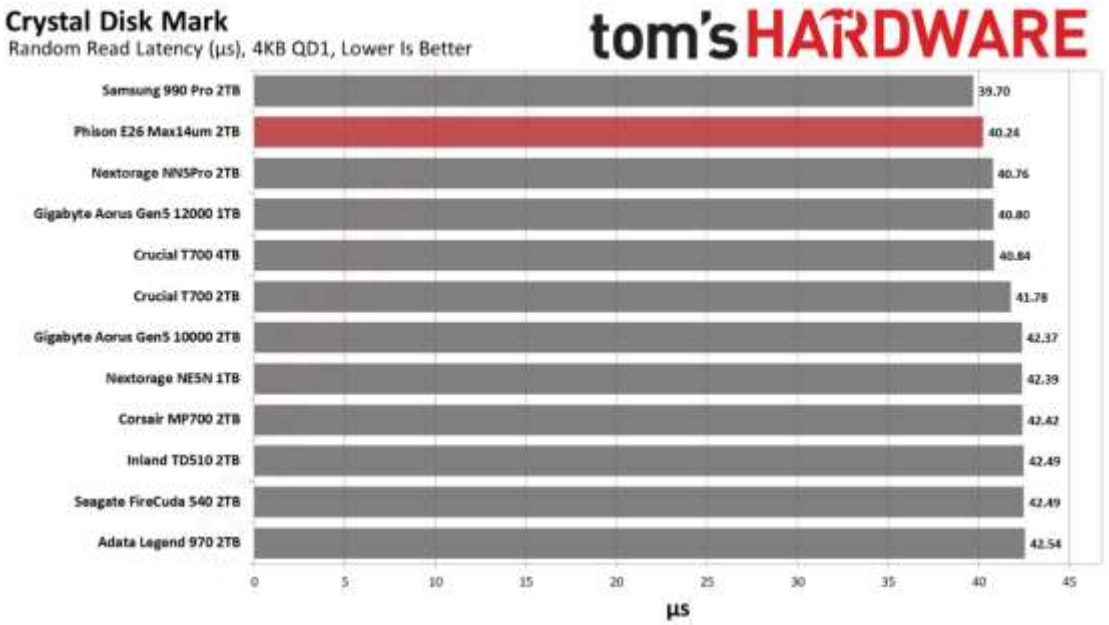
ATTO Disk Benchmark
SSD Read Throughput at QD1 and Various Block Sizes

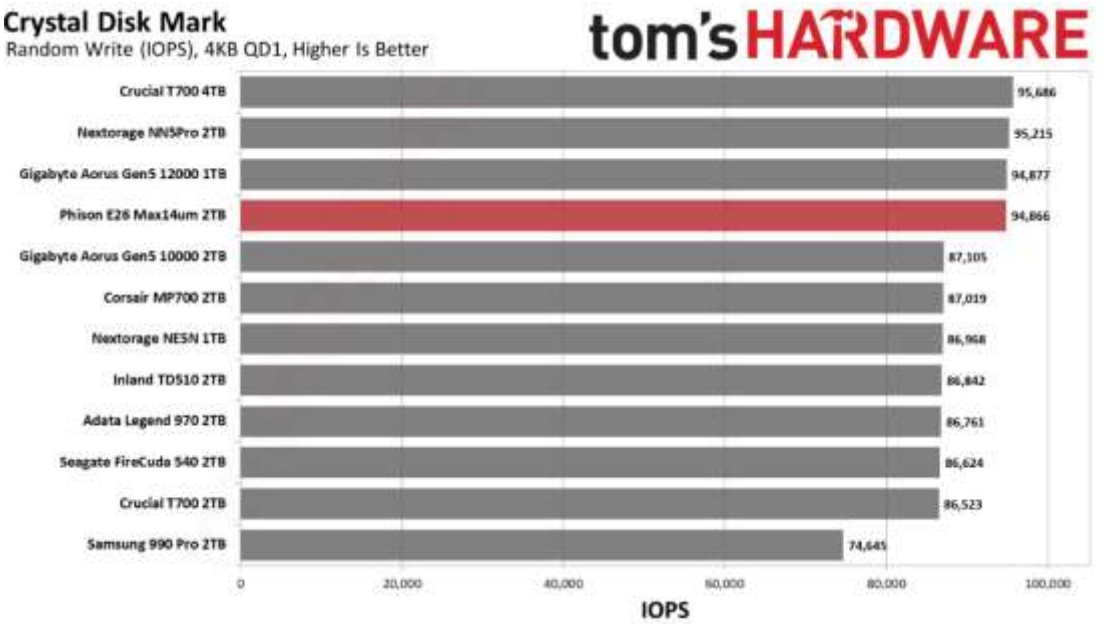
tom's HARDWARE

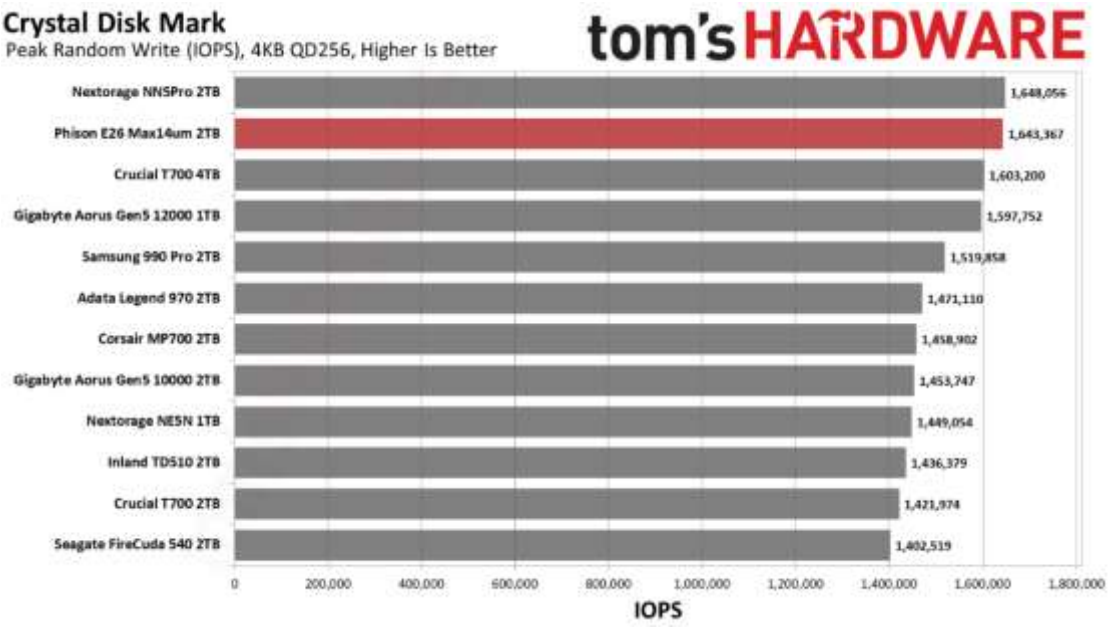


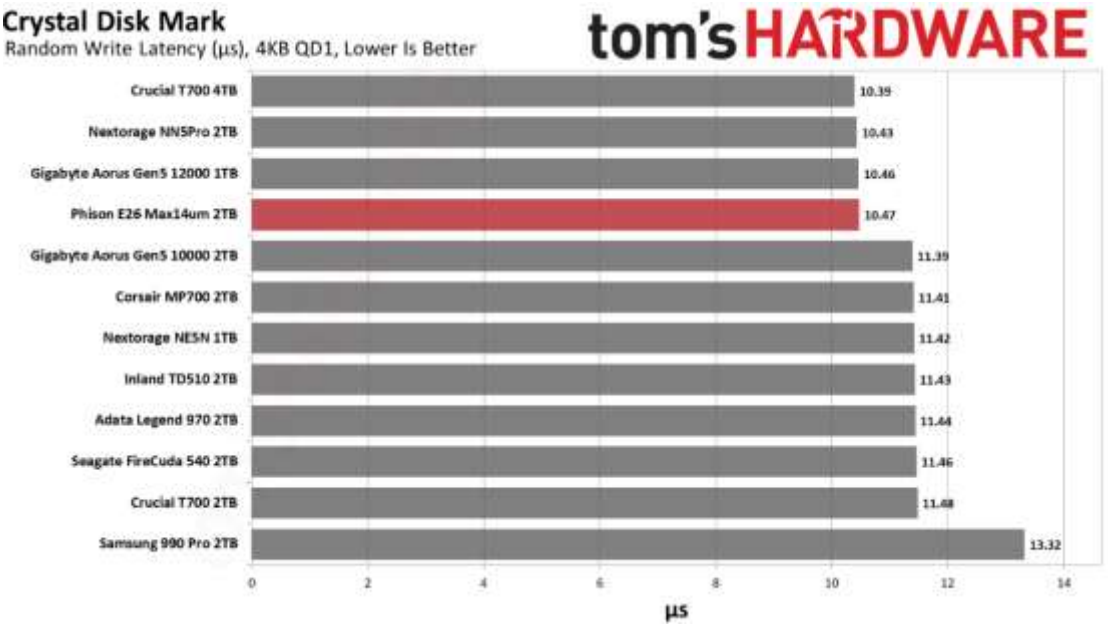

Crystal Disk Mark
Random Read (IOPS), 4KB QD1, Higher Is Better
tom's HARDWARE
Samsung 990 Pro 2TB
Phison E26 Max14um 2TB
Nextorage NN5Pro 2TB
Gigabyte Aorus Gen5 12000 1TB
Crucial T700 4TB
Crucial T700 2TB
Gigabyte Aorus Gen5 10000 2TB
Nextorage NE5N 1TB
Corsair MP700 2TB
Seagate FireCuda 540 2TB
Inland TD510 2TB
Adata Legend 970 2TB
25,133
24,805
24,494
24,471
24,439
23,891
23,556
23,545
23,528
23,492
23,490
23,459
0
5,000
10,000
15,000
20,000
25,000
IOPS

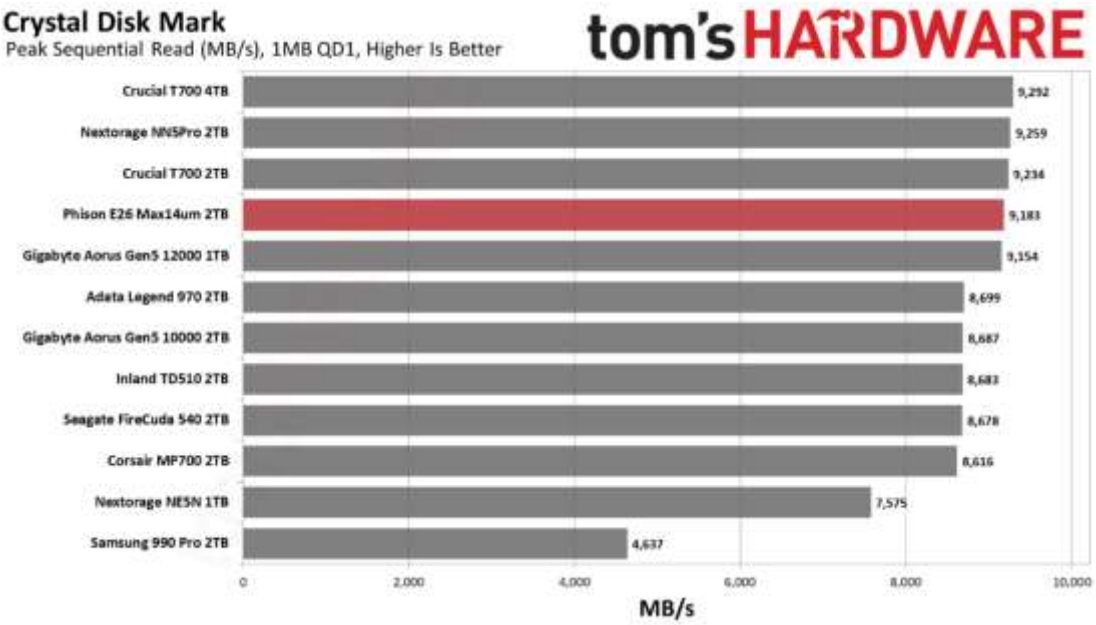

Crystal Disk Mark
Peak Random Read (IOPS), 4KB QD256, Higher Is Better
tom's HARDWARE
Phison E26 Max14um 2TB
Nextorage NN5Pro 2TB
Crucial T700 4TB
Crucial T700 2TB
Corsair MP700 2TB
Adata Legend 970 2TB
Gigabyte Aorus Gen5 10000 2TB
Inland TD510 2TB
Seagate FireCuda 540 2TB
Samsung 990 Pro 2TB
Gigabyte Aorus Gen5 12000 1TB
Nextorage NE5N 1TB
1,681,175
1,498,367
1,497,128
1,481,563
1,428,187
1,426,780
1,424,667
1,421,702
1,418,234
1,400,621
1,339,285
1,310,569
0
200,000
400,000
600,000
800,000
1,000,000
1,200,000
1,400,000
1,600,000
1,800,000
IOPS

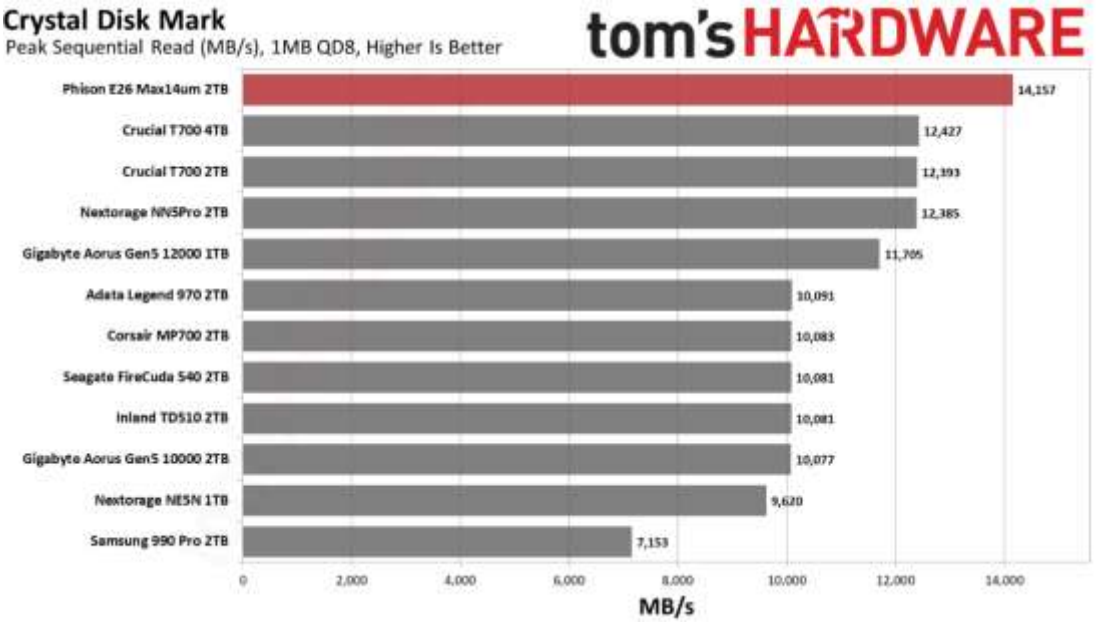

Crystal Disk Mark
Random Read Latency (µs), 4KB QD1, Lower Is Better
tom's HARDWARE
Samsung 990 Pro 2TB 39.70
Phison E26 Max14um 2TB 40.24
Nextorage NN5Pro 2TB 40.76
Gigabyte Aorus Gen5 12000 1TB 40.80
Crucial T700 4TB 40.84
Crucial T700 2TB 41.78
Gigabyte Aorus Gen5 10000 2TB 42.37
Nextorage NE5N 1TB 42.39
Corsair MP700 2TB 42.42
Inland TD510 2TB 42.49
Seagate FireCuda 540 2TB 42.49
Adata Legend 970 2TB 42.54
0 5 10 15 20 25 30 35 40 45
µs

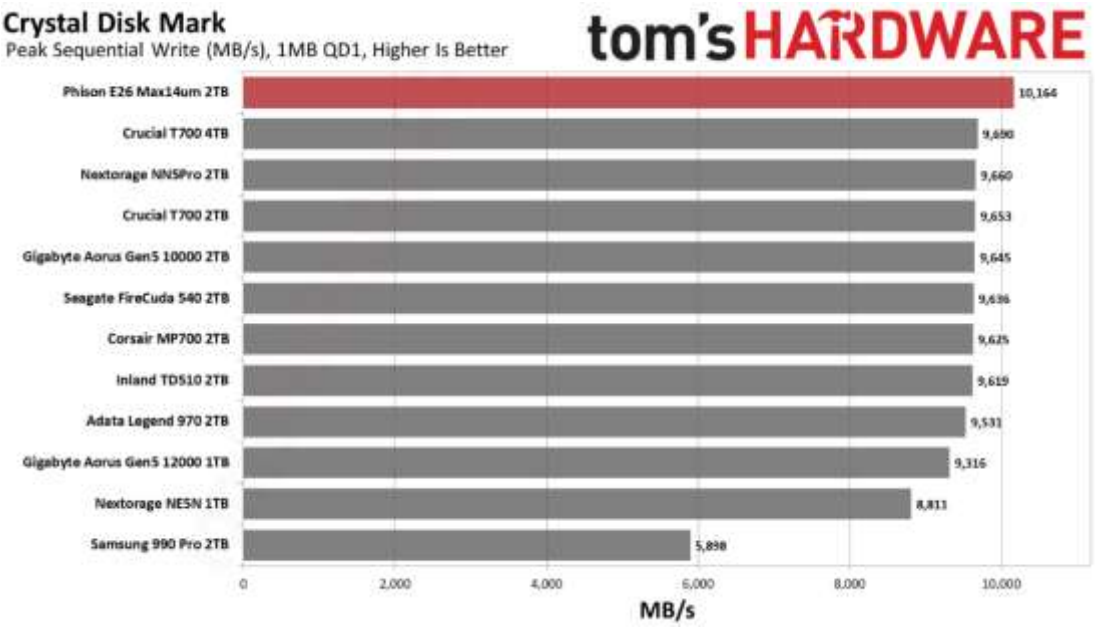

Crystal Disk Mark
Random Write (IOPS), 4KB QD1, Higher Is Better
tom's HARDWARE
Crucial T700 4TB 95,686
Nextorage NN5Pro 2TB 95,215
Gigabyte Aorus Gen5 12000 1TB 94,877
Phison E26 Max14um 2TB 94,866
Gigabyte Aorus Gen5 10000 2TB 87,105
Corsair MP700 2TB 87,019
Nextorage NE5N 1TB 86,968
Inland TD510 2TB 86,842
Adata Legend 970 2TB 86,761
Seagate FireCuda 540 2TB 86,624
Crucial T700 2TB 86,523
Samsung 990 Pro 2TB 74,645
0 20,000 40,000 60,000 80,000 100,000
IOPS

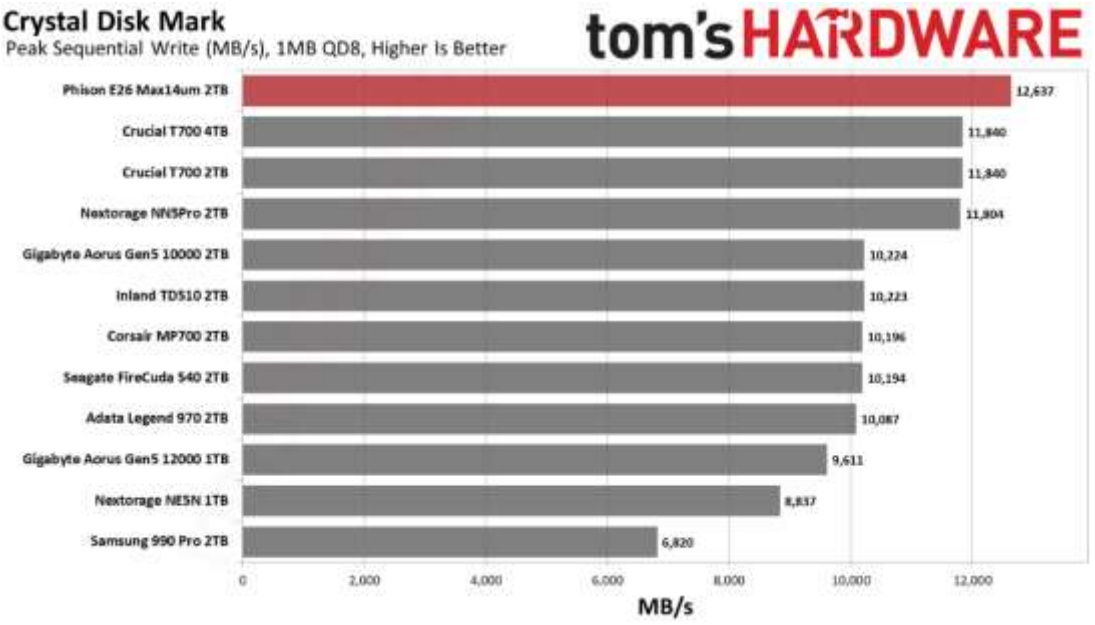

Crystal Disk Mark
Peak Random Write (IOPS), 4KB QD256, Higher Is Better
tom's HARDWARE
Nextorage NNSPro 2TB
Phison E26 Max14um 2TB
Crucial T700 4TB
Gigabyte Aorus Gen5 12000 1TB
Samsung 990 Pro 2TB
Adata Legend 970 2TB
Corsair MP700 2TB
Gigabyte Aorus Gen5 10000 2TB
Nextorage NE5N 1TB
Inland TD510 2TB
Crucial T700 2TB
Seagate FireCuda 540 2TB
1,648,056
1,643,367
1,603,200
1,597,752
1,519,858
1,471,110
1,458,902
1,453,747
1,449,054
1,436,379
1,421,974
1,402,519
0
200,000
400,000
600,000
800,000
1,000,000
1,200,000
1,400,000
1,600,000
1,800,000
IOPS


Crystal Disk Mark
Random Write Latency (µs), 4KB QD1, Lower Is Better
tom's HARDWARE
Crucial T700 4TB
Nextorage NNSPro 2TB
Gigabyte Aorus Gen5 12000 1TB
Phison E26 Max14um 2TB
Gigabyte Aorus Gen5 10000 2TB
Corsair MP700 2TB
Nextorage NE5N 1TB
Inland TD510 2TB
Adata Legend 970 2TB
Seagate FireCuda 540 2TB
Crucial T700 2TB
Samsung 990 Pro 2TB
10.39
10.43
10.46
10.47
11.39
11.41
11.42
11.43
11.44
11.46
11.48
13.32
0
2
4
6
8
10
12
14
µs


Crystal Disk Mark
Peak Sequential Read (MB/s), 1MB QD1, Higher Is Better
tom's HARDWARE
Crucial T700 4TB
9,292
Nextorage NV5Pro 2TB
9,259
Crucial T700 2TB
9,234
Phison E26 Max14um 2TB
9,183
Gigabyte Aorus Gen5 12000 1TB
9,154
Adata Legend 970 2TB
8,699
Gigabyte Aorus Gen5 10000 2TB
8,687
Inland TD510 2TB
8,683
Seagate FireCuda 540 2TB
8,678
Corsair MP700 2TB
8,616
Nextorage NE5N 1TB
7,575
Samsung 990 Pro 2TB
4,637
0
2,000
4,000
6,000
8,000
10,000
MB/s


Crystal Disk Mark
Peak Sequential Read (MB/s), 1MB QD8, Higher Is Better
tom's HARDWARE
Phison E26 Max14um 2TB
14,157
Crucial T700 4TB
12,427
Crucial T700 2TB
12,393
Nextorage NV5Pro 2TB
12,385
Gigabyte Aorus Gen5 12000 1TB
11,705
Adata Legend 970 2TB
10,091
Corsair MP700 2TB
10,083
Seagate FireCuda 540 2TB
10,081
Inland TD510 2TB
10,081
Gigabyte Aorus Gen5 10000 2TB
10,077
Nextorage NE5N 1TB
9,620
Samsung 990 Pro 2TB
7,153
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
MB/s





(Image credit: Tom's Hardware)

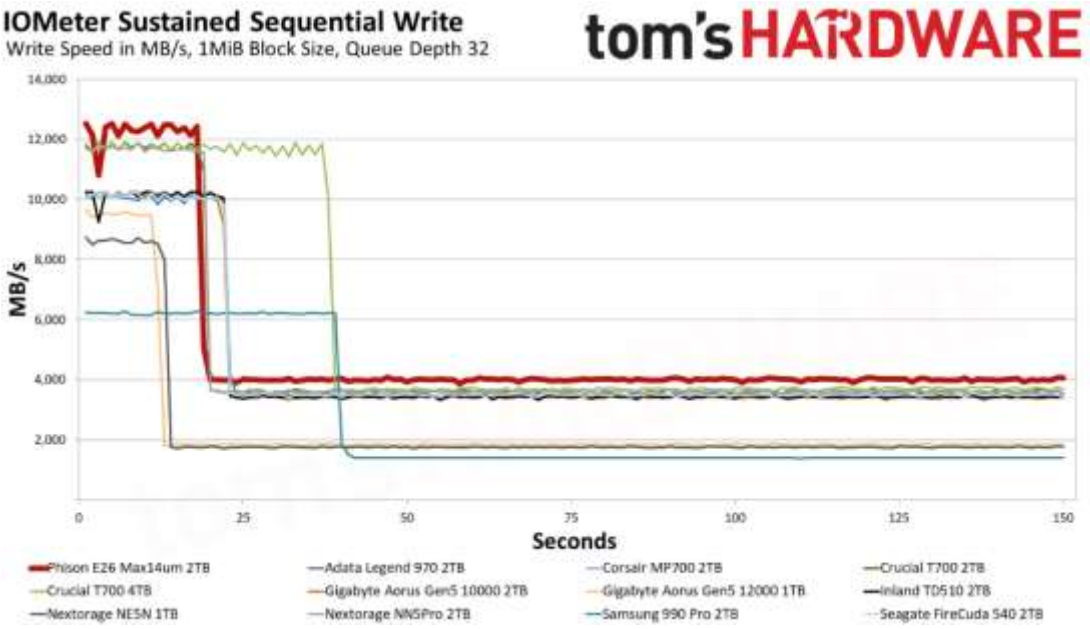
The Max14um performs as expected in ATTO, matching and the slightly exceeding other E26-based drives. It does push things a bit higher for sequential reads and writes at larger block sizes. QD1 sequential performance in CDM is not particularly impressive, but if you have multiple streams going on you will benefit from the extra bandwidth.

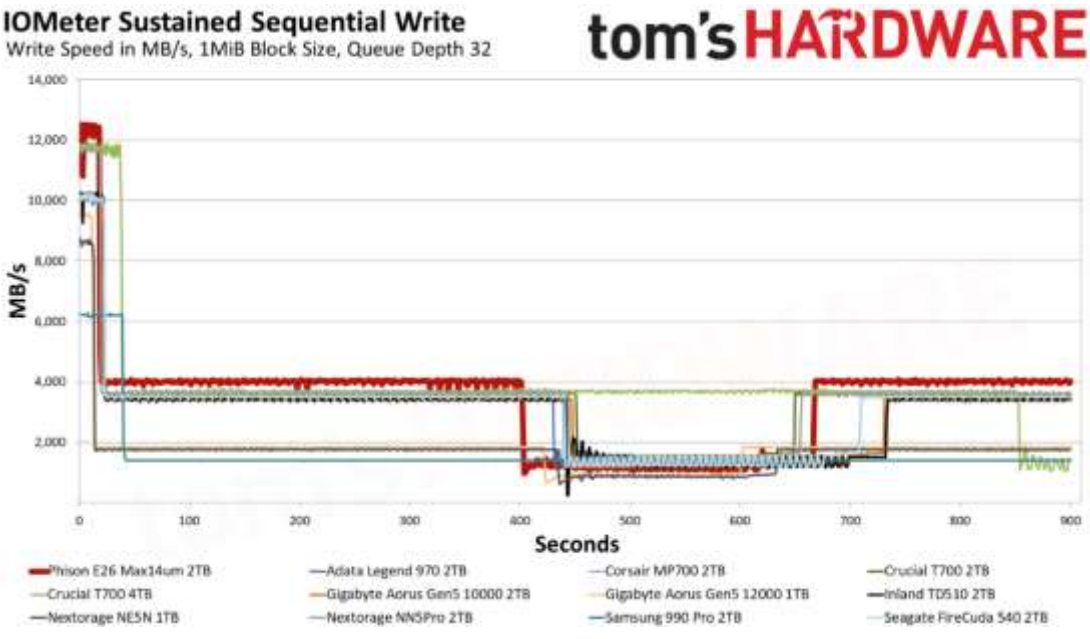
Random QD1 4KB latencies remain good, so there's no trade-off here except against the 990 Pro in reads. Current and future 990 Pros will be using a later generation of Samsung flash which might level this matchup, as latency may be slightly worse for it, but the difference here is small enough not to worry.

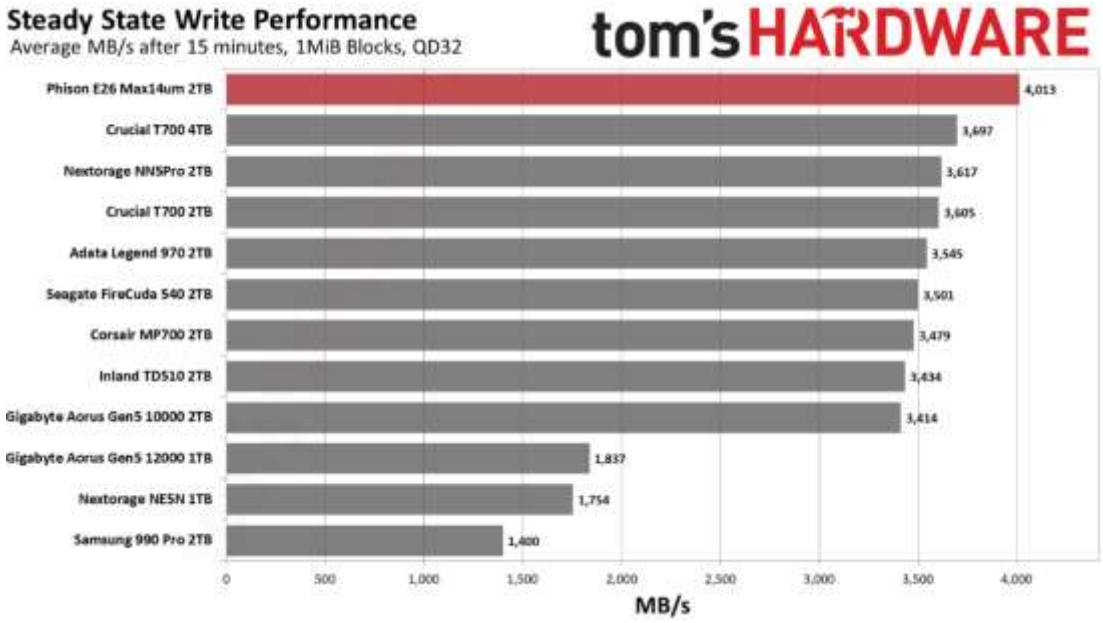
SUSTAINED WRITE AND TEMPERATURES

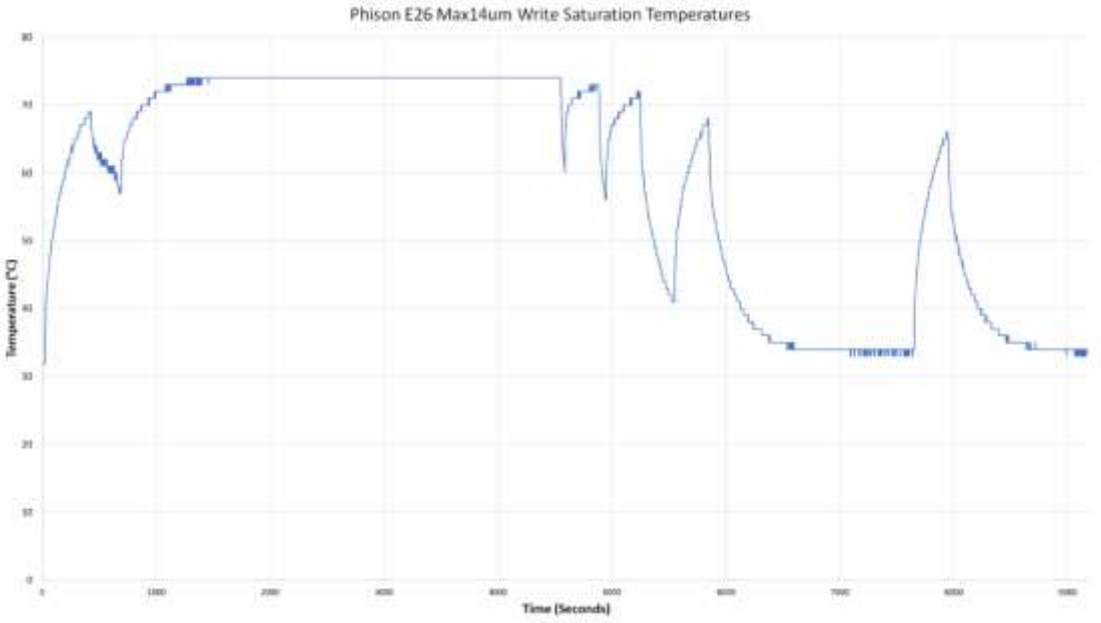
Official write specifications are only part of the performance picture. Most SSDs implement a write cache, which is a fast area of (usually) pseudo-SLC programmed flash that absorbs incoming data. Sustained write speeds can suffer tremendously once the workload spills outside of the cache and into the "native" TLC or QLC flash.

We use Iometer to hammer the SSD with sequential writes for at least 15 minutes (we tested for 120 minutes on the Max14um) to measure both the size of the write cache and performance after the cache is saturated. We also monitor cache recovery via multiple idle rounds. This process shows the performance of the drive in various states as well as the steady state write performance.









(Image credit: Tom's Hardware)

The 2TB Max14um sustains over 12 GB/s in pSLC mode with an SLC cache size matching the slower E26 rivals. Lower-end variants like the MP700 are just above 10 GB/s while the faster T700 is closer to 12 GB/s. While TLC performance has been pretty similar among the tested PCIe 5.0 drives, the Max14um manages to eke out 4 GB/s, which finally and definitively puts it above the fastest PCIe 4.0 drives. This is a very high-performance drive for larger sustained writes. It's also nice that the drive can recover SLC while still providing solid TLC-mode performance.

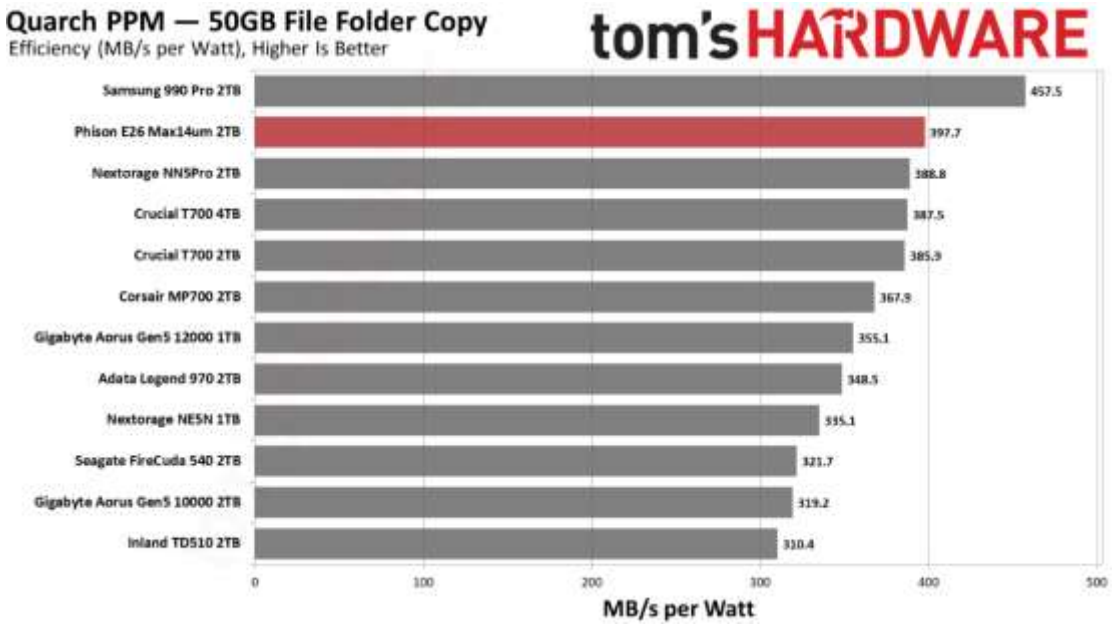
We've also included a plot of the drive temperature over the course of our write saturation testing. You can see the Max14um peak at 74C, about 1,300 seconds into our stress

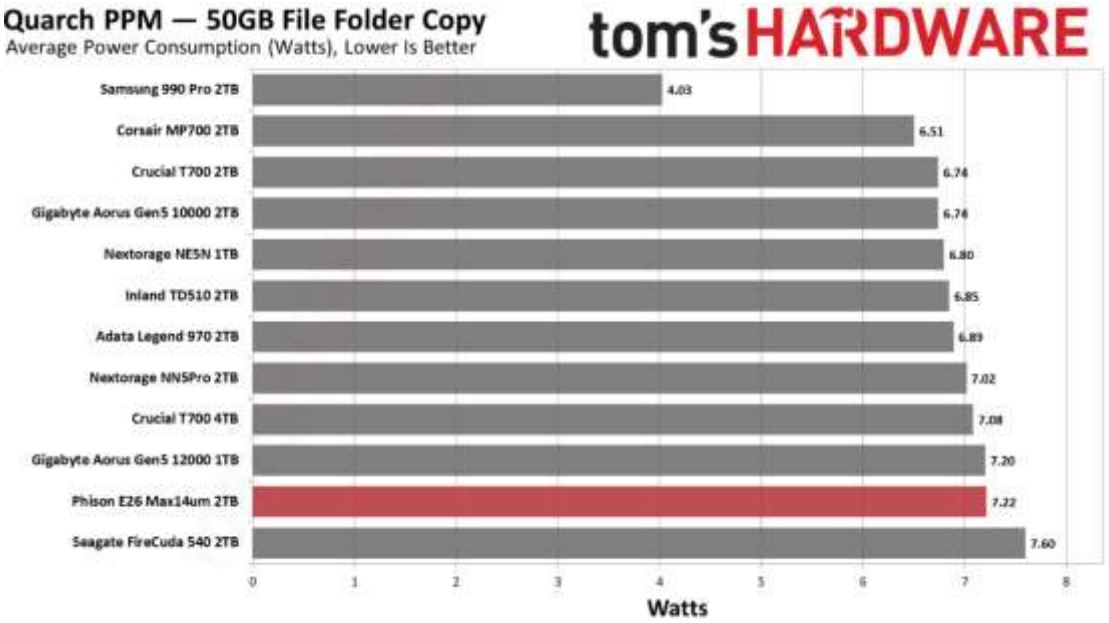
test. Note that at this point, we've done roughly two complete drive writes to the SSD, so this isn't even remotely a real-world consumer workload.

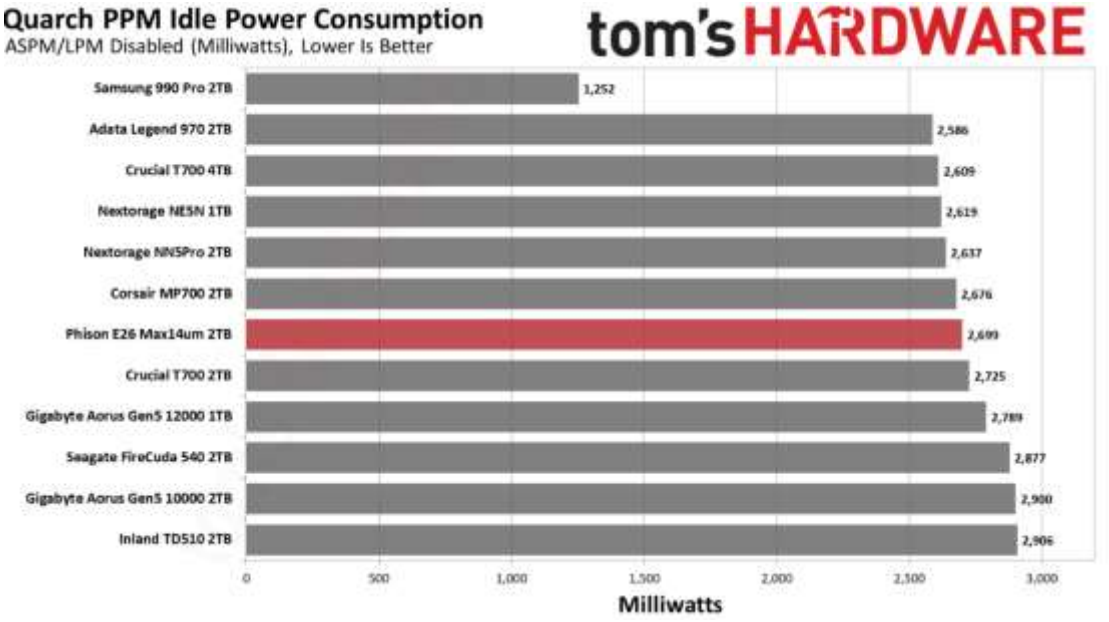
POWER CONSUMPTION

We use the Quarch HD Programmable Power Module to gain a deeper understanding of power characteristics. Idle power consumption is an important aspect to consider, especially if you're looking for a laptop upgrade as even the best ultrabooks can have mediocre stock storage. Desktops may be more performance-oriented with less support for power-saving features, so we show the worst-case.

Some SSDs can consume watts of power at idle while better-suited ones sip just milliwatts. Average workload power consumption and max consumption are two other aspects of power consumption but performance-per-watt, or efficiency, is more important. A drive might consume more power during any given workload, but accomplishing a task faster allows the drive to drop into an idle state more quickly, ultimately saving energy.







(Image credit: Tom's Hardware)

Power is one area where the E26 seems to have issues. The good news is that the higher performance achieved by the Max14um — which required a lot of fine-tuning to reach — does make the drive a bit more efficient. Idle power consumption is still disappointingly high, though. This drive is intended to be used in high-performance desktops so there's a reasonable chance you can live with high idle draw. Officially, though, the drive is rated to pull less than 144mW in PS3 and less than 85mW in PS4 power modes.

TEST BENCH AND NOTES

CPU	Intel Core i9-12900K

Motherboard	Asus ROG Maximus Z790 Hero
Memory	2x16GB G.Skill DDR5-5600 CL28
Graphics	Intel Iris Xe UHD Graphics 770
CPU Cooling	Enermax Aquafusion 240
Case	Cooler Master TD500 Mesh V2
Power Supply	Cooler Master V850 i Gold
OS Storage	Sabrent Rocket 4 Plus 2TB
Operating System	Windows 11 Pro

We use an Alder Lake platform with most background applications such as indexing, Windows updates, and anti-virus disabled in the OS to reduce run-to-run variability. Each SSD is prefilled to 50% capacity and tested as a secondary device. Unless noted, we use active cooling for all SSDs.

PHISON E26 MAX14UM CONCLUSION

The Max14um is simply the fastest SSD we've ever tested, and we gave Phison one of our Best of CES awards for moving the entire consumer storage industry forward. It's not perfect, though: Our original complaints about power consumption and thermal output remain valid. It's expected that a high-end PCIe 5.0 SSD will pull a lot of power, but idle desktop consumption remains high. Efficiency also leaves something to be desired, but we expect this will be cleared up in time with more efficient controllers and flash.

Part of that means needing a smaller process node as well as four-channel and DRAM-less designs, and part of it means iterative improvements to Micron's flash. We've seen great things out of the Maxio MAP1602 combined with YMTC's 232-Layer flash, so we know it's possible. Even though Phison's upcoming E31T won't reach 14 GB/s, it's likely to give excellent levels of performance with lower power use. With the E26 Max14um, we're looking more at cutting-edge, early adopter hardware. That's perfectly fine if you really want to push your storage.

There's also the question of capacity. We've only tested one E26 drive at 4TB and some manufacturers aren't even planning to go that high. 8TB meanwhile is nowhere to be

seen. Given that this controller needs at least 2TB to really stretch its legs, that means there's a lot of competition at 1TB and also at 4TB and up. Solid PCIe 4.0 drives like the Samsung 990 Pro, Solidigm P44 Pro, and WD Black SN850X are probably the best bet at 1TB. The 990 Pro and SN850X are also excellent at 4TB, but if you just want capacity then drives like the Lexar NM790 remain a good value. That does leave a 14 GB/s monster as a bit niche.

We are nevertheless interested to see what will come out from various brands. Having more drives available is a net benefit and we hope for better cooling solutions, too. We don't think that the thermal output of these drives is as bad as it first appears; it's really more a matter of requiring a heatsink. This adds cost and potential complexity but in the grand scheme of things, it just means we will need more efficient drives for future laptops. DirectStorage adoption remains far enough away that you can wait this one out on desktop, unless you really want top tier storage performance today.

6.3 测试端口扩展

目前有些公司专注于消费类 M.2 Gen 4 SSD 测试，有些专注于数据中心用 U.2 Gen 4 SSD 测试。购买了上述主板后面临的一个问题是，如何可以多测试一些 SSD？下面分 M.2 扩展 U.2 扩展，以及 PCIe Gen 4/5/6 插槽几个部分进行讲述。

6.3.1 M.2 Gen 4 NVMe SSD 扩展

一般的消费类 Gen 4 主板都配置一个或者 2 个 M.2 接口，如果要测试更多怎么办呢？常用的有下面几种方法。

6.3.1.1 M.2 – AIC 转接卡

参见下图，这样就可以将未使用的 Gen 4 x4 插槽转成 M.2 Slot 使用，这样就可以有效增加 M.2 的测试数量 – 该转接卡也有带掉电控制功能的版本，支持 Python 脚本进行上下电测试。

图 6-147 PCIe Gen 4/5/6 x4 M.2 to AIC 转接卡

6.3.1.1.1 Serial Cables Gen4 m.2 适配器评测 – *StorageReview***

https://www.storagereview.com/review/serial-cables-pcie-gen4-m-2-adapter-review

作者：莱尔 • 史密斯 2021 年 2 月 18 日

Serial Cables 公司的 PCIe Gen4 m.2 适配器 (PCI4-AD-x4M2-04-G4) 允许测试实验室和用户向台式机和服务器主板插槽添加额外的 PCIe Gen4 x4 M.2 slot。 传统上，m.2 适配器为企业提供了一种经济高效的服务器引导驱动器替代方案（不占用前置插槽）； 然而，这个高度可定制的适配器的用途远不止这些。

图 6-148

Serial Cables 公司 Gen4 适配器几乎与所有 m.2 SSD 尺寸兼容，因为它具有 2230（30 毫米）、2242、2260、2280 和全长 22110（110 毫米）外形尺寸镀孔。目前，大多数 Gen4 m.2 SSD 都提供一种外形尺寸 (2280)，但一系列支持允许用户添加其他不同尺寸

的非 Gen4 卡。尽管适配卡向后兼容 Gen3 驱动器，但您当然只能看到 Gen3 速度。这也适用于主机主板。如果在不支持 Gen4 的主板上添加此卡，则驱动器将受限于 PCIe 插槽的 Gen3 带宽。

就更换驱动器而言，这个家伙是独一无二的，而且用途广泛。大多数消费卡使用螺钉将 m.2 SSD 固定下来以获得更“永久”的意义，而该卡具有几种不同形式的快速释放卡舌。这些方法之一是一个小的弹簧加载功能，可以轻松地将 m.2 驱动器卡入到位。另一个是枢轴点上的塑料片（形状像吉他拨片）；只需将驱动器安装在 m.2 插槽中，轻轻按下它，然后将塑料卡舌滑到驱动器上以将其固定到位。取消对螺丝刀的需求肯定会受到那些使用案例涉及持续更换驱动器的人的欢迎，因为在传统适配器卡上安装 m.2 驱动器可能既麻烦又烦人。

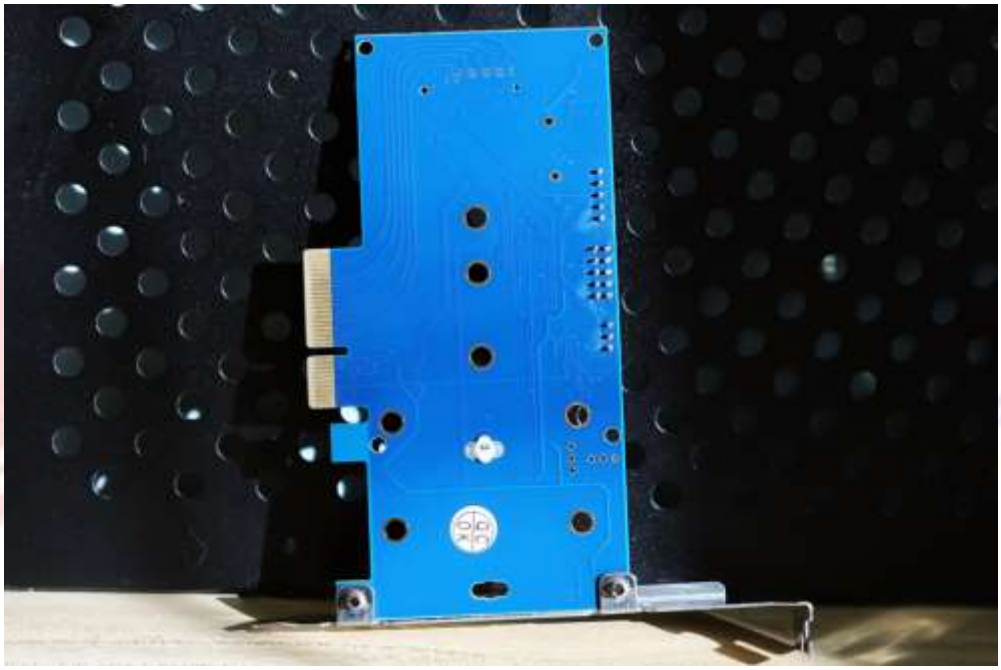


图 6-149

在适配器的顶部，您会看到一系列熟悉的引脚，包括 CLKREQ#（时钟请求信号）、WAKE#（唤醒功能）和 PERST#（用于指定电源电压参数）。您还会注意到位于 PCB 左下方的开关（即启动或停止电流沿电路流动的按钮）。这目前没有任何功能，因为 Serial Cables 公司为正在进行 SSD 开发的特定客户添加了此按钮。留意固件更新以在未来启用此功能。

性能表现

为了演示 Serial Cables 公司 PCIe Gen4 m.2 适配器的性能，我们使用了联想 P620（一个电源驱动的工作站，配备了支持 PCIe Gen4 的 AMD Threadripper PRO CPU）并使用以下配置运行 Blackmagic 基准测试：

为适配器配备了三星 980 Pro PCIe Gen4 SSD 并将其安装在工作站的 PCIe 插槽之一 P620 上

将三星 980 Pro 直接安装在工作站内部的原生插槽上

这样做的目的是显示适配器满足或超过第 1 层工作站的板载插槽。使用 P620 内部的 Serial Cables 公司适配器，980 Pro 达到 5.29Gb/s 读取和 4.36GB/s 写入。

图 6-150

这些结果与将驱动器直接安装在主板上时记录的速度几乎相同，因为三星 Pro 的读取速度为 5.28GB/s，写入速度为 4.34GB/s。

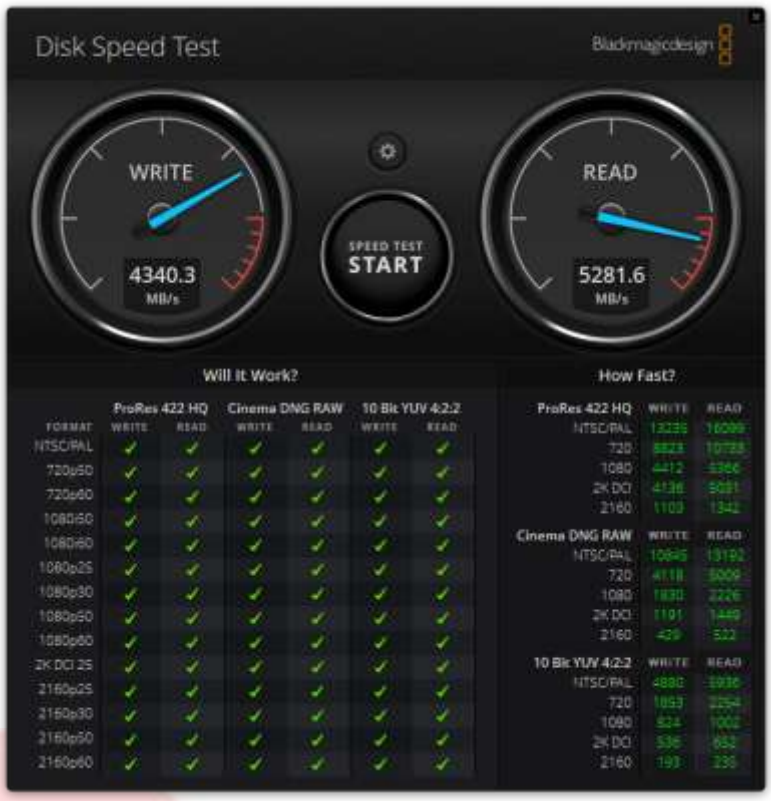


图 6-151

结论

虽然我们只是使用 Serial Cables 公司 PCIe Gen4 m.2 适配器作为 M.2 NVMe SSD 到支持 Gen4 的主机系统的传递，但您可以在测试实验室或工程实验室中使用它。 因此，Serial Cables 公司适配器当然不适合所有人，但比您可能从您最喜欢的在线零售商处找到的普通 15 美元的 m.2/AIC 适配器要好得多。同时，该适配器也提供带 micro-controller 掉电测试管理功能的版本。

这是一款非常小众的适配卡，可以配置为完成一系列高技术任务。 将所有这些与其超过板载速度的性能结合起来，像这样的卡并不多。

6.3.1.1.2 其它常见四盘位 Gen3/4/5 M.2 SSD 扩展卡

如果仅是转接，不需要通过 API 控制掉电/上电，也可以采用下面的四盘位 M.2 SSD 转接卡（下面三张图依次对应 Gen5, Gen4, Gen3 M.2 NVMe SSD 扩展卡，需要主板 x16 插槽支持 bifurcation）。

图 6-152 Gigabyte AORUS Gen5 AIC Adaptor

图 6-153 Gigabyte AORUS Gen4 AIC Adaptor

图 6-154 普通 PCIe Gen3 x16 4*M.2 NVMe SSD 扩展卡

6.3.1.2 PCIe Gen 4/5/6 Host 卡

由于目前市场上没有 Gen 4 x16 转接 4 个 M.2 Slot 的转接卡，如果觉得通过上述方法只能转接一个 M.2 端口还是太少怎么办呢？可以使用下面两种 Gen 4 x16 转接 4 个盘位的方法，但是两种方法都需要二次转接，不过不用担心，我们做过实际测试，通过使用高质量的转接卡可以实现对于 PCIe Gen 4 信号质量基本不造成影响。

6.3.1.2.1 Gen 4 x16 转接 2 个 Gen 4 x8 SlimSAS 接口

参见下图，该 Gen 4 x16 Host 卡顶部左上角转换为两个 Gen 4 x8 接口，通过两根独立的 Gen 4 x8 to 2*U.2 socket 线缆可以转接成 4 个 U.2 插槽，然后再次通过 M.2 to U.2 转接卡转接成 4 个 M.2 Gen 4 SSD。

图 6-155 图 6.2 PCIe Gen 4/5/6 x16 Host 卡（带 2 个 Gen 4 x8 internal port）

图 6-156 PCIe Gen 4/5/6 x16 Host 卡（带 2 个 x8 internal port）

上图为该 x16 插卡连接了两根 x8 to 2*U.2 Slot 线缆的图片，其中左上角的图片为每个转接线端部的细部照片，即每根线转接成了 2 个 U.2 Slot。然后通过下面 M.2 to U.2 转接卡将 M.2 接入每个 U.2 Slot 即可。

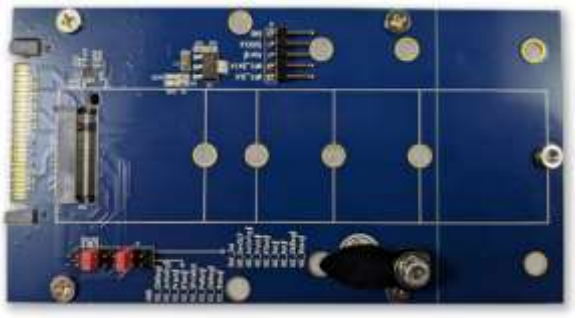


图 6-157 Gen 4 x4 M.2 to U.2 转接卡

下图是连接一根 x8 to 2*U.2 Slot 线缆的实拍图，从图中可以清晰地看到该转接线提供 2 个 U.2 插槽，以及一个 SATA 头端的电源接口，因为 U.2 SSD 需要单独供电。

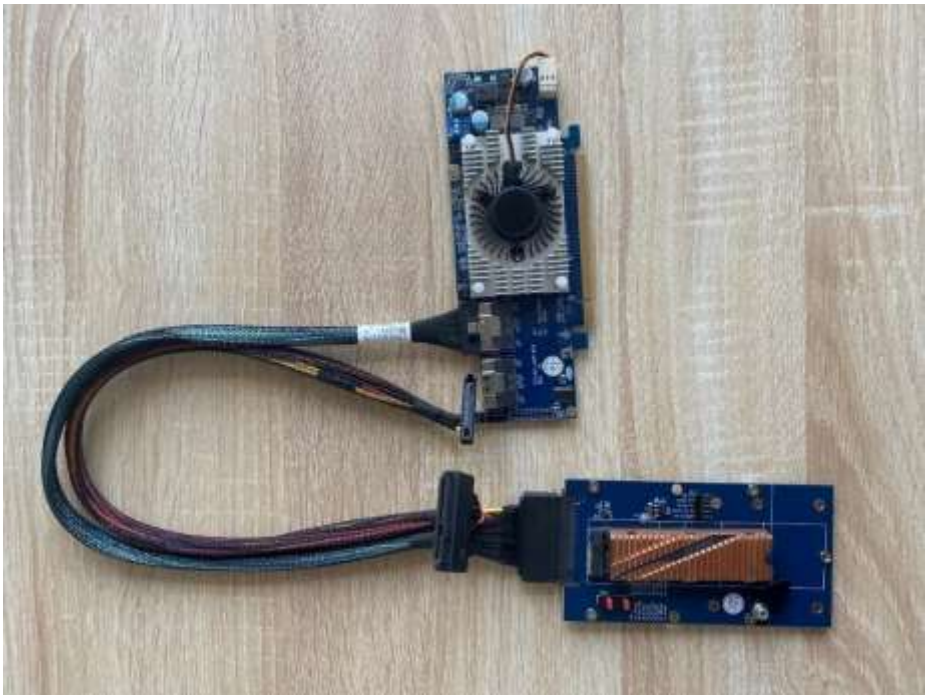


图 6-158 Gen 4 x4 M.2 to U.2 转接卡连接一个 M.2 NVMe SSD 示意图

6.3.1.2.2 Gen 4 x16 转接 4 个 Gen 4 x4 HD-MINI-SAS 接口

图 6-159 PCIe Gen 4/5/6 x16 Host 卡（带 4 个 Gen 4 x4 external cable port）

下图是连接一根 HD MINI SAS to U.2 Slot 线缆的实拍图，从图中可以清晰地看到该转接线提供转接成一个 U.2 插槽，以及一个 SATA 头端的电源接口，因为 U.2 SSD 需要单独供电。

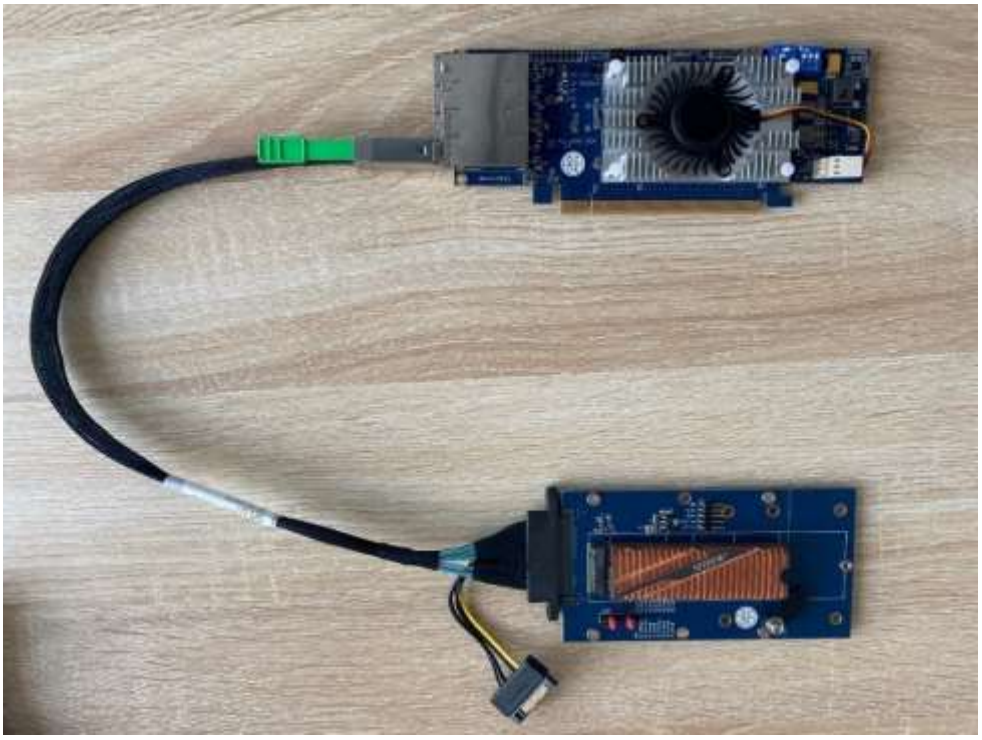


图 6-160 HD MINI SAS to U.2 Slot 线缆实拍图

6.3.1.3 PCIe Gen 4/5/6 Host 卡（带 M.2 Slot）

如果想增加 M.2 SSD 插槽，又不想减少一个 PCIe Gen 4/5/6 Slot 怎么办？可以使用下面的PCIe Gen 4/5/6 Host卡，该卡底部为Gen 4 x16金手指，上面提供一个Gen 4 x8 sot，右边提供一个 M.2 插槽。

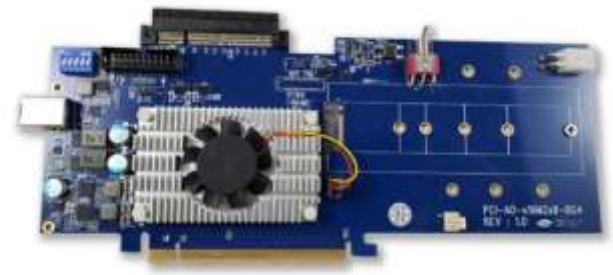


图 6-161 PCIe Gen 4/5/6 x16 Host 卡（带 M.2 和 x8 Slot）

6.3.1.4 PCIe Gen 4/5/6 M.2 NVMe SSD 功能测试扩展板（20 端口）

如果需要对于 Gen 4 NVMe SSD 进行批量功能性测试，那么我们提供如下的支持 20 个 Gen 4 x1 M.2 slot 的扩展板，可以结合用户的需求进行定制。主机端插入 Gen 4 Host Card 通过 slimsas to slimsas cable 连接该 M.2 扩展板，这样主机即可发现 20 个 NVMe SSD。

图 6-162

6.3.2 U.2 Gen 4 NVMe SSD 扩展

一般的消费类 Gen 4 主板不提供 U.2 接口，服务器主板目前也没有提供 Gen 4 U.2 插槽（基于 Intel 公司 PCIe Gen 3 CPU 的存储服务器一般前面板提供 Gen3 U.2 Slot），那么应该如何测试 Gen 4 U.2 NVMe SSD 呢？常用的有下面几种方法。

6.3.2.1 U.2 – AIC 转接卡

参见下面两图，可以将未使用的 Gen 4 x4 插槽转成 U.2 Slot 使用，这样可以有效增加 U.2 的测试数量。两种的功能完全一样，区别是一个是竖插，一个为横插，竖插的转接卡价格较贵。

图 6-163 Gen 4 U.2 to AIC 转接卡（竖插）

图 6-164 Gen 4 U.2 to AIC 转接卡（横插）

** *横插型号提供 1/2 Height Bracket 以及 Full Height Bracket 供选择*

6.3.2.2 PCIe Gen 4/5/6 Host 卡

参见 1.2.1 PCIe Gen 4/5/6 Host 卡，图 2，6，8。

图 6-165

6.3.2.3 PCIe Gen 4/5/6 NVMe SSD 盘柜

目前业内提供两种盘柜可供选择，一个为 Active 盘柜，一种为 Passive 盘柜。

两个产品的前面板看起来几乎一样，都提供 8 个 Gen 4 U.2 Slot，支持 Single Port Gen 4 x4，也支持 Dual Port Gen 4 x2，支持热插拔，支持对每一个盘位通过 API 命令进行异常掉电/上电，方便测试。两者的主要区别在于 Active 盘柜内部提供 PCIe Gen 4 交换芯片，Passive 盘柜内部没有交换芯片。如果测试 M.2 NVMe SSD 那么该盘柜需要配置 M.2 to U.2 转接卡（参加图 4）。

图 6-166 Passive 和 Active 盘柜前面板实拍图

下面是两种盘柜的连接图实拍，注意两者使用的线缆是不一样的，都是从 Host 卡连接到盘柜后面板的接口。

图 6-167 Passive 盘柜连接方式（使用特殊定制线缆）

图 6-168 Active 盘柜连接方式

我们提供 Gen 4 U.2 和 U.3 两种规格。

- **Gen 4 U.2 测试盘柜**

图 6-169

- **Gen 4 U.3 测试盘柜**

图 6-170

6.3.3 PCIe Gen 4/5/6 Slot 扩展

有些用户开发 PCIe Gen 4/5/6 NVMe SSD Controller，或者开发基于 PCIe Gen 4 的各类板卡，在前期验证阶段可以使用 FPGA 验证卡，或者后期 ASIC 流片以后往往一段时间内仍然使用 PCIe 插卡进行性能/功能验证，这个时候需要使用较多的 PCIe Gen 4/5/6 插槽。目前针对这方面的扩展需求有两种方案可供选择。

6.3.3.1 PCIe Gen 4 高扩展性服务器主板

可以选用下图所示的 7-Slot PCIe Gen 4/5/6 x16 主板，AMD CPU 提供 128 个 lane 可以直通每个 PCIe Slot。

如果采用 Gen 4 x16 Host Card，那么一张卡可以连接 4 块盘，那么该主板可以支持 4*7 = 28 块 NVMe SSD 盘进行测试。

图 6-171 PCIe Gen5 插槽主板，可以通过 Gen5 swich 进行 U.2 扩展

图 6-172 7-Slot Gen4 高扩展性主板

6.3.3.2 PCIe Gen 4/5/6 Slot 扩展板（5x16 或者 8x8）

6.3.3.2.1 Gen5 插槽扩展

Gen4 或者 Gen5 主机通过插入 Gen 5 host card，然后通过 Gen5 cable 连接到扩展板的 Gen5 target card，然后待测试或者使用的 Gen5 卡可以插入扩展板的插槽中。实现上述扩展需要下面的这些部件。

图 6-173

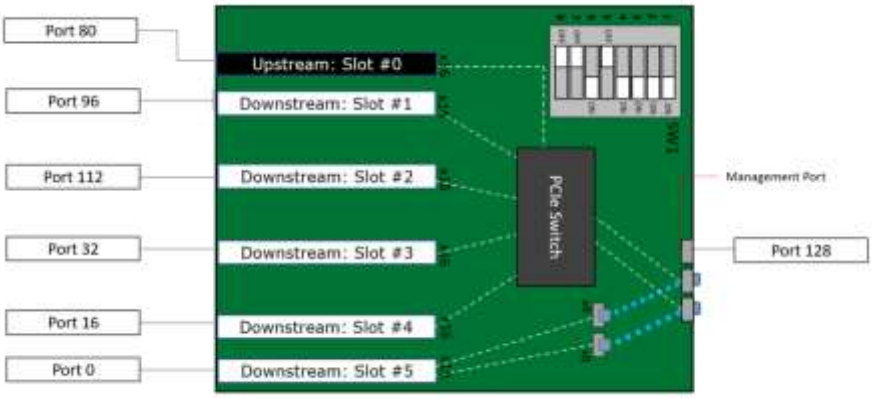


图 6-174 扩展板内部架构

扩展板典型应用场景

x16 用例

使用两条链路电缆连接在主机和 Gen5 背板或 扩展单元之间。

图 6-175

6.3.3.2.2 Gen4 插槽扩展

可以使用专用的 PCIe Gen 4/5/6 Slot 扩展板。参见下图，左图为 1 个 x16 上行扩展到 5 个 x16 下行，右图为一个 x16 上行扩展到 8 个 x8 下行。

图 6-176 PCIe Gen 4/5/6 Slot 扩展柜

6.3.3.3 PCIe Gen 4/5/6 Slot 扩展板（x4 slot）

参见下图，通过在主机上面插入 SerialCables 公司的带 slim-sas Gen 4 x8 的 host card，可以直接连接 PCIe Gen 4/5/6 Slot 扩展板（4x4 slot）。

该方案非常适合于 NVMe SSD 控制器厂商进行测试，将 NVMe SSD controller 做成 Gen4x4 验证板可以直接插入扩展板，该待测 controller 可以热插拔，不需要每次换卡的时候重启电脑，大大节省测试效率。

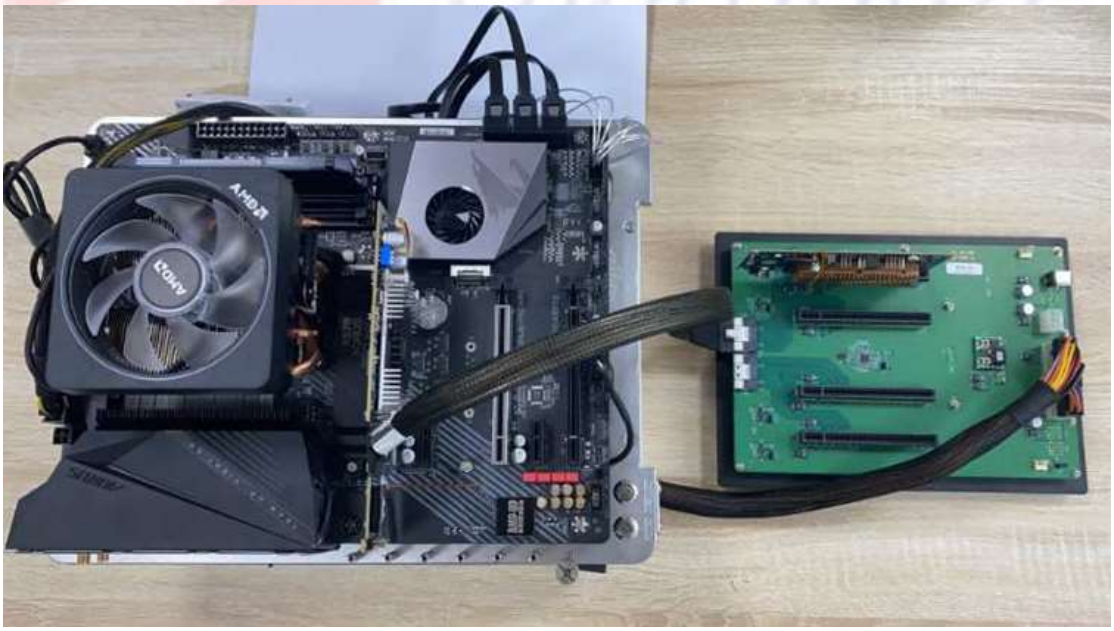


图 6-177

6.3.3.4 PCIe Gen 5 M.2 扩展桥接卡（Gen5 x4 M.2）

AORUS Gen5 AIC 适配器：高达 16TB 的 SSD，60GB/秒的读取速度

新的 AORUS Gen5 AIC 适配器支持多达 4 个 4TB PCIe 5.0 SSD，高达 16TB... 以 60GB/秒的超快速度实现终极速度。

安东尼加雷法

发表于 2022 年 8 月 23 日上午 3:33 CDT | |更新时间：2022 年 9 月 15 日星期四 12:09 PM CDT

图 6-178

AORUS 刚刚发布了支持 PCIe 5.0 的新 AORUS Gen5 AIC 适配器，该适配器将以超快的速度驱动大量 SSD 容量。

新的 AORUS Gen5 AIC 适配器具有 4 个 NVMe M.2 插槽，最多支持 4 个 PCIe 5.0 SSD，总容量为 16TB。这意味着您将购买 4 个 4TB PCIe 5.0 NVMe M.2 SSD，将它们安装到强大的 AORUS Gen5 AIC 适配器中，并享受 16TB 的 SSD 容量……速度不同于您使用过的任何产品：60GB/秒。

AORUS 宣布，当您在 AORUS Gen5 AIC 适配器内使用带有 PCIe 5.0 SSD 的 RAID 阵列时，您将提升至 60GB/秒（60,000MB/秒）。坦率地说，这很荒谬，但很高兴看到：SATA 6Gbps SSD 的速度约为 550MB/秒，而基于 PCIe 3.0 的 NVMe M.2 SSD 可以达到大约 3.5GB/秒，而基于 PCIe 4.0 的 NVMe M .2 SSD 最高可达 7.5GB/秒（7500MB/秒）。

因此，令人眼花缭乱的 60GB/秒真是太疯狂了。以下是使用 PCIe 5.0 x16 和 PCIe 5.0 x4 可以驱动的 SSD 带宽之间的一些快速比较。

PCIe 的带宽：

- **PCIe 5.0 x16：60GB/秒+**

- **PCIe 4.0 x16：30GB/秒+**
- **PCIe 3.0 x16：15GB/秒+**
- **PCIe 5.0 x4：15GB/秒+**
- **PCIe 4.0 x4：7.5GB/秒+**
- **PCIe 3.0 x4：3.9GB/秒+**

AORUS 刚刚以 AORUS Gen5 10000 SSD 的形式发布了其新的 PCIe 5.0 SSD，它本身的速度为 12.4GB/秒（12,400MB/秒）。如果您将其中 4 个坏男孩放在一起，您会看到 12.4 x 4 = 在 RAID 阵列中以 60GB/秒的速度运行，使用 AORUS Gen5 AIC 适配器。

图 6-179

技嘉渠道解决方案产品开发部总监 Jackson Hsu 解释道：“随着即将到来的 PCIe® 5.0 平台，高速存储可以达到 10GB/s 以上的访问速度。对于追求更高性能的用户，技嘉 AORUS Gen5 AIC 通过构建磁盘阵列引领极致性能 集成四个 PCIe® 5.0 插槽，用户可以选择具有自定义容量和性能的不同 NVMe M.2 SSD，以获得最大的灵活性。同时，先进的热设计防止高温下的热节流。极速运算，让技嘉 AORUS Gen5 AIC 成为提升储存效能的最佳选择”。

图 6-180

在内部，新的 AORUS Gen5 10000 SSD 包装了新的 Phison PS5026-E26 8 通道控制器，技嘉表示它将为用户提供“随机读取速度的终极控制”。还有超过 200 层的堆栈结构，最大 2400MT/s 带宽的 3D-TLC NAND 闪存和 LPDDR4 缓存设计。

技嘉表示，其全新 AORUS Gen5 10000 SSD 由群联 PS5026-E26 控制器的多核架构增强，因此它不仅将改善 AI 多任务操作，而且“将内容创作者、游戏玩家和渴望极致的用户性能更上一层楼”。

散热方面，技嘉指出，大部分次世代主机板都内建 M.2 散热片，因此 AORUS Gen5 10000 SSD 特别设计了易于拆卸的全覆铜散热片。该公司表示，用户将能够在主板上的内置散热器或带有 AORUS Gen5 10000 SSD 的封闭式散热器之间进行选择

阅读更多：https://www.tweaktown.com/news/88087/aorus-gen5-aic-adaptor-up-to-16tb-of-ssds-rocks-with-60gb-sec-reads/index.html

6.4 温箱专用 PCIe Gen 4/5/6 高低温测试背板

有的测试场景需要将 PCIe Gen 4 NVMe SSD 放入高低温箱进行测试，可以配置下图所示高低温测试背板或者夹具，支持-25 ~ 85 度，分别用于 U.2 和 M.2 接口 NVMe SSD。主机侧插入“图 3 PCIe Gen 4/5/6 x16 Host 卡（带 2 个 x8 internal port）”所示的 Gen 4 Host 卡，然后通过两个独立的 x8 to x8 cable 连接到下面背板的背面高速接口，线缆支持高低温。注意：一张 PCIe Gen 4/5/6 x16 Host 卡支持同时测试 4 个 PCIe Gen 4/5/6 x4 Single Port NVMe SSD。

图 6-181 置于高低温箱的 Gen 4 NVMe SSD 背板示意图

7. NAND 和 DDR5 测试工具和夹具

7.1 NAND 特性分析设备

对于 SSD controller 或者 SSD drive 研发中心来讲，设计 LDPC 纠错算法的前提需要事先了解所支持的 NAND 的特性。NAND 在不同的温度下的特性往往差异较大，尤其在实际工作环境中往往受到周围环境（例如控制器）影响会导致温度过高，所以这类特性测试往往需要在 NAND 支持的最大速率（1.6GT, 2.0/2.4GT）以及加温下，针对 NAND 进行非常细致的测试才能了解这些特性。

参加下图，我们可以很清晰地看出该颗 NAND 在 30 度，50 度，以及 70 度的时候的 BER 的变化曲线差异很大。

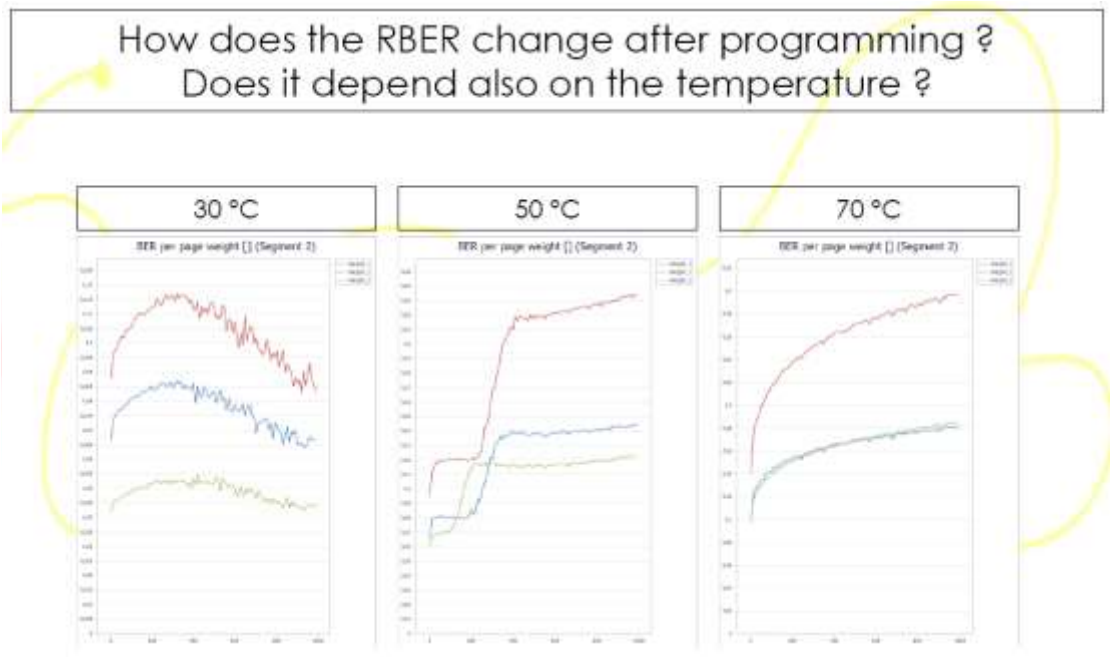


图 7-1

意大利公司 NplusT 从 2002 年起专注于 Memory 测试，其针对研发中心的 NAND 特性分析设备获得业内众多 SSD Controller 公司以及科研院所和高校的青睐，国内的控制器厂商如 Alibaba，平头哥，Tenafe，Yeestor，Dapu，浪潮，华芯，大学包括清华大学，复旦大学，南京大学，山东大学，另外还有中芯国际等芯片制造企业等等。

NplusT 提供的针对测试数据的后处理大数据分析软件也得到了众多应用。该 NAND 特性分析设备支持业内常用的 Kioxia, Micron, Sandisk, WDC, YMTC 等 NAND。

图 7-2

NAND 闪存技术 (ONFI5) 正在快速发展，以跟上大容量、高性能、高可靠性的存储需求：

- **同一封装中的 die 数量、LUN 和 plane 数量的增加；**
- **层数的增加；**
- **I/O 速度达到或超过 1.6 GT/sec；**
- **更短的 program 和 read 时间；**
- **具有极高峰值的更高的平均内核 core 和 IO 电流。**

100 多层 3D TLC 和 QLC NAND 设备在数据、信号和电源完整性方面对 SSD 设计人员提出了挑战。

有关 endurance, retention and disturb sensitivity 的知识对于优化介质的管理算法至关重要，这对于 3D 结构和拓扑的复杂性增加至关重要。

对于高效的系统级电源完整性解决方案，设计人员需要了解各种操作模式下的 NAND 电源功耗的概况分析。下面，作为一个例子，它显示了 NAND 电流分布以及 plane 和活跃的 die 数量的影响。

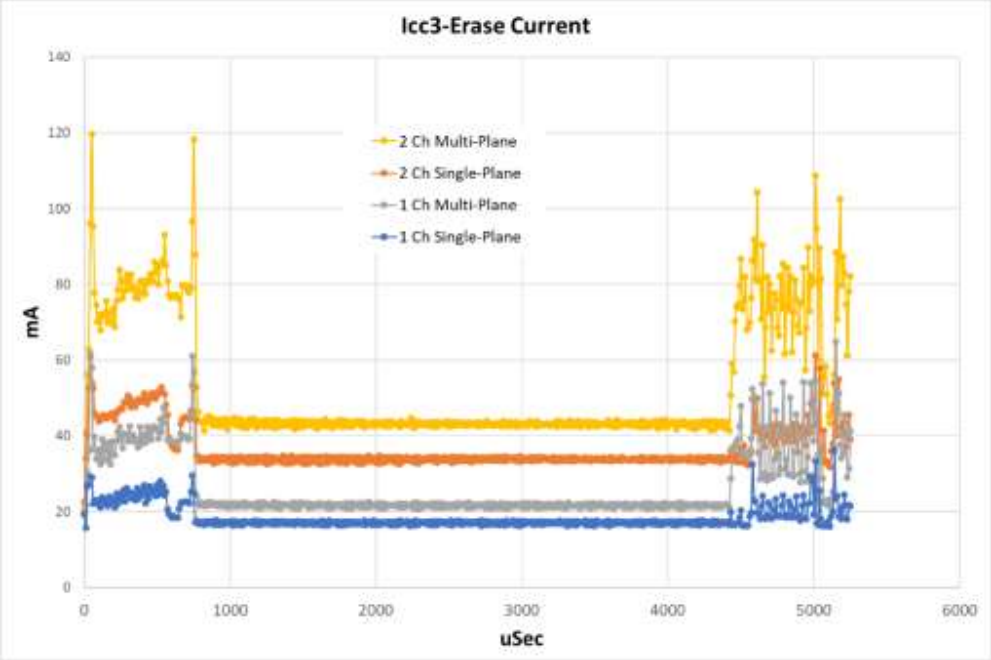


图 7-3

总的来说，并行多芯片测试对于以下方面很重要：

- **减少从大量 die 中收集数据的测试时间；**
- **模拟 SSD 控制器的行为，该控制器同时执行不同 die 的操作；**
- **提取和优化总功耗曲线并最大限度地减少尖峰和峰值；**
- **验证不同 die 之间操作的干扰/干扰；**
- **了解并减轻多芯片堆栈的热影响：堆栈中的温度变化、热分析和提高数据可靠性的方法；**
- **测量高速接口对不同 NAND 通道的影响。**

传统的 NAND 特性测试工具显示出它们的局限性：需要通过集成的实时功率波形捕获全速进行特性分析，以在各种操作条件（温度、老化、操作并行性）下监控 NAND 数据、功率和信号完整性 - 以复制 NAND 在 SSD 或类似目标应用程序中的工作方式，以满足产品规格、质量和可靠性。

此外，如果测试和特性分析工具是基本的并且不支持并行性和完整 I/O，那么开发测试程序/脚本并在 NAND 设备上执行它们所需的工作量可能会非常高，并且会消耗大量的时间和资源速度。

NplusT 新一代 NanoCycler NAND 特性测试工具的主要目标是能够在多个 NAND 设备上同时执行特性测试程序，以实时生成和提取大量读出和功率数据。

NanoCycler HS 是用于 NAND 测试和深入表征性能和可靠性的可扩展平台。与传统市场其它产品不同，NanoCycler 以全 I/O 速度执行 NAND 并行测试。该系统支持高达 ONFI 5.0， 2.4GT/s。

NanoCycler 系统可扩展性允许对每个系统最多 84 个 NAND 设备执行同时测试。

可以级联多个系统以将并行度扩展到 84 位以上。

NanoCycler 系统架构支持不同类型的并行测试操作：

- **可以在所有测试单元上执行相同的测试程序，以拥有大量 DUT 并收集具有统计代表性的数据；**
- **可以在具有不同测试参数（例如时序、数据模式等）或测试条件（温度、电压）的每个测试仪单元上执行相同的测试程序，以识别相关性、相关性、裕度；**
- **不同的测试单元可以执行不同的测试程序，以减少总表征时间。**

由于每个 Tester Unit 直接连接一个 NAND 设备，一个 NAND 设备最多包含 16 个 die（HDP 封装），NanoCyler 全系统配置最多可以同时测试 1344 个 NAND die。

这导致系统的 NAND 测试数据总带宽约为 240 GT/s，假设 NAND I/O 速度为 1.6GT/s，I/O 吞吐量利用率为 90%。多个 NanoCyler 系统可以级联，以进一步增加总系统容量和数据吞吐量。以全 I/O 速度运行 NAND 对于支持 NAND 设备内的芯片并行操作至关重要。

下表显示了 NanoCycler 可以并行测试执行 NAND 读取操作的芯片数量，例如提取 NAND 原始误码率。

NAND die test parallelism (2 Channels package)

Page size = 16kB tRead = 60uS		# Planes		
		1	2	4
I/O Speed	1,600 MB/s	10	6	2
	2,000 MB/s	14	6	4
	2,400 MB/s	16	8	4

图 7-4

NanoCycler 专有测试控制器基于与被测 NAND 器件紧密耦合的高性能 FPGA。它可以控制和监控测试条件（温度、电压、功率）、生成测试数据模式、对封装中的每个芯片应用背对背命令、传输数据并实时计算原始误码率数据，全速工作，并且没有开销。

总之，NplusT 的 NanoCycler 性能和能力完全符合 NAND 技术的发展。它可以符合当今实际 NAND 应用条件的全速和并行性测试测试并且分析待测 NAND 特性。同时，其经过验证的架构、硬件和软件可扩展性使不受限制的数量的 NAND 芯片能够同时独立地进行测试。

7.1.1 面向 SSD 开发的 NAND 特性分析

NAND 是 SSD 中的存储介质，就像磁盘于 HDD 一样。

与 HDD 磁盘一样，NAND 也是一种“不完美”的介质，会产生数据错误从而需要对于读取错误机型纠正。 NAND 介质的不同之处和挑战在于，错误会随着时间增加，包括 NAND 磨损、多次读取相同数据、超过数据保留时间，以及在不同温度条件下。

随着 TLC，QLC 以及未来 PLC 的每个单元的位数增加，“缺陷”的影响随着 3D-NAND 的代际更迭变得越来越具有挑战性。

同时，更高的性能、更高的速度、更高的功率和电流尖峰(Spike)进一步挑战了 NAND 信号和电源的完整性。

设计高性能、高容量、高可靠性的 SSD，了解和解决所有这些方面的问题对于交付在整个生命周期内满足可靠性、数据完整性和性能一致性的产品至关重要。

在和最终应用将要运行的相同条件下，适当的 NAND 特性分析对于以下方面的设计和验证至关重要：

- **NAND 管理算法通常称为“NAND 策略”；**
- **SSD 控制器纠错 (LDPC)；**
- **SSD 信号和电源完整性。**

需要特别注意的 NAND 特性分析的关键是：

- **闪存 Cell Vt 分布重叠，导致内存数据读取错误增加；**
- **编程/擦除周期、读操作干扰、数据保留时间、温度条件下的 Vt 分布变化；**
- **影响正常 NAND 功能和数据完整性的电流尖峰(Spike)；**
- **来自时序、端接、驱动强度、PCB 布局的 I/O 信号完整性；**
- **闪存 Cell 在生命周期内的退化导致的硬故障。**

Vt 分布的重叠尾部（TLC 为 7，QLC 为 15）是读取错误的根源。当太多单元位于读取阈值的错误一侧时，大量错误读取的位需要极大的努力

才能执行纠错，或者甚至超出 SSD 本身的 纠错能力。

在与最终应用类似的操作条件下对于 Vt 分布的特性分析，随着使用寿命、磨损、数据保留时间、读取干扰、温度等方面，对于了解错误行为、设计读取级别放置的策略和算法以及设计和调整 SSD 纠错 (LDPC)的至关重要。了解这些特性如何变化，以及 SSD 控制器的跟踪/预测策略和纠错系统，对于提供高数据可靠性、稳定的性能和良好的 QoS（SSD 存储应用程序的关键参数）非常关键。

SSD 容量和速度增加带来的第二个挑战是确保电源和信号完整性。 NAND 编程、擦除和读取内部操作会产生比平均电流值高几倍的电流尖峰。多个 NAND 裸片的并行和同步操作，通常在多（8 到 16 个及更多）SDD 闪存通道中会导致 SSD 级别的极端电流尖峰。

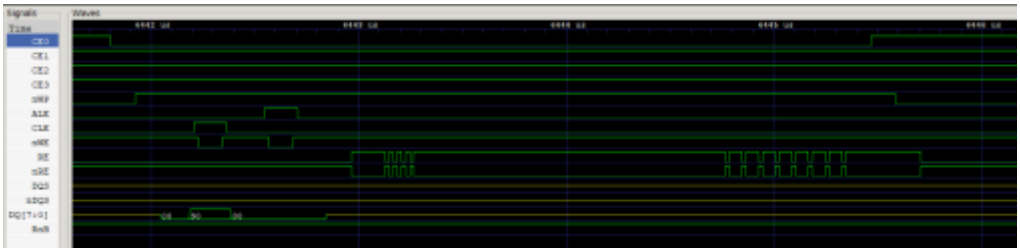


图 7-5

了解 NAND 时序和电流分布对于设计、仿真、调整大小、校准电源调节器以及制定策略来限制和避免这些电流尖峰的对齐至关重要。

除了电源完整性之外，SSD 设计还必须处理信号完整性：由于 I/O 数据走线在超过 1GT/s 数据速率时变成传输线，因此 PCB 设计必须越来越准确，并且需要特定的培训和校准程序来设置 NAND 和 SSD 控制器 I/O 参数。

提供可靠结果的高效特性分析需要强大的测试环境。 NanoCycler 是一个专门为 NAND 特性分析而创建的测试系统，支持 SSD 开发人员进行 NAND 探索。NanoCycler 支持如下功能：

- **能够针对每个 NAND Socket 运行独立的测试，优化多个测试条件下的统计数据生成；**
- **具有根据每个测试 Socket 进行准确、快速的温度控制；**
- **可从单 NAND Socket 开发测试平台完全扩展至多个 84 插槽机架；**
- **以与 NAND 在 SSD 中运行相同的速度执行数据传输，最高可达 2.0GT/sec（即将推出 2.4GT/sec）；**
- **以零开销生成写入和验证数据，包括“完美”随机模式；**
- **支持从/到被测设备的快速位图数据传输；**

- **使用智能滤波算法检测平均和峰值电流，捕获电流波形以增加对设备内部操作的可见性；**
- **提供内置的、可定制的界面训练算法，时间分辨率只有几皮秒 ps。**

NanoCycler 提供了一个开发环境来提高测试工程师的效率，通过一个测试库能够：

- **支持各大厂商的最新设备；**
- **在单个 LUN 或多个 LUN、die、CE、通道上执行操作；**
- **使用多种预定义和自定义模式运行由 ONFI 标准定义的命令；**
- **接受具有任意信号时序、命令代码等等的定制、供应商特定命令序列；**
- **生成易于后处理的数据日志。**

NanoCycler 的专业软件：

- **为测试执行和高级测试流程定义提供易于使用的图形用户界面；**
- **配备多个测试系统共享的高效数据采集系统；**
- **使用流行的 Python 语言实现测试流程和算法；**
- **包括调试工具，如可视化信号序列显示，使故障排除更加容易。**

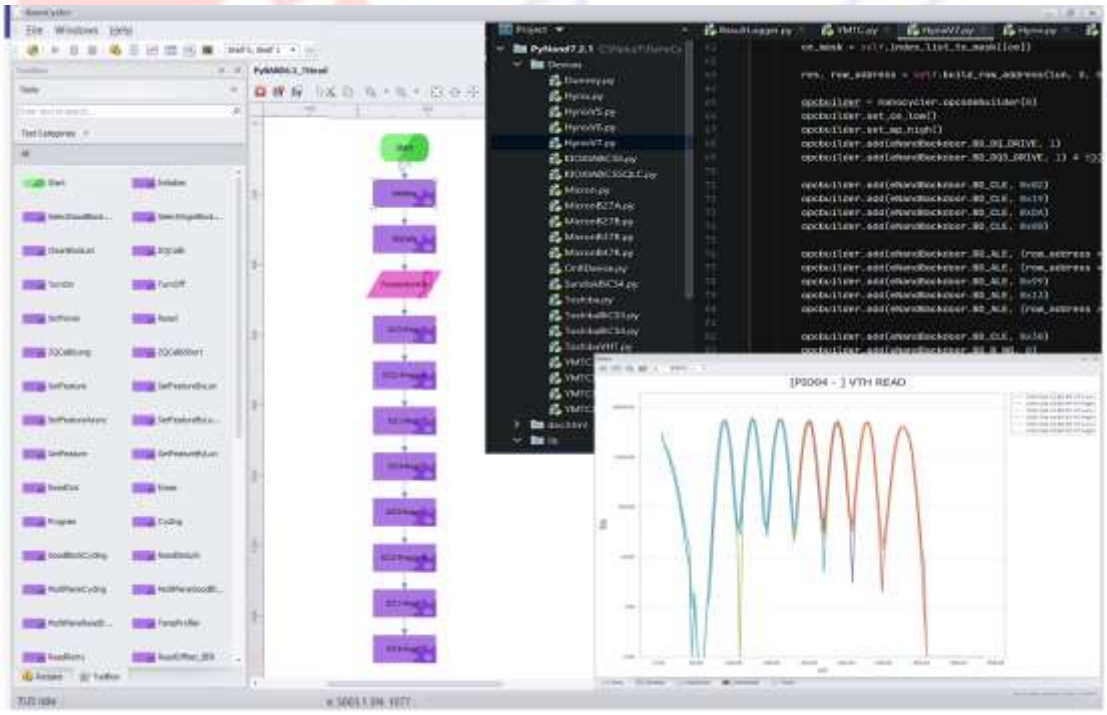


图 7-6

下面是一些 nanocycler GUI 的主界面展示：

- **创建 NAND 测试序列**

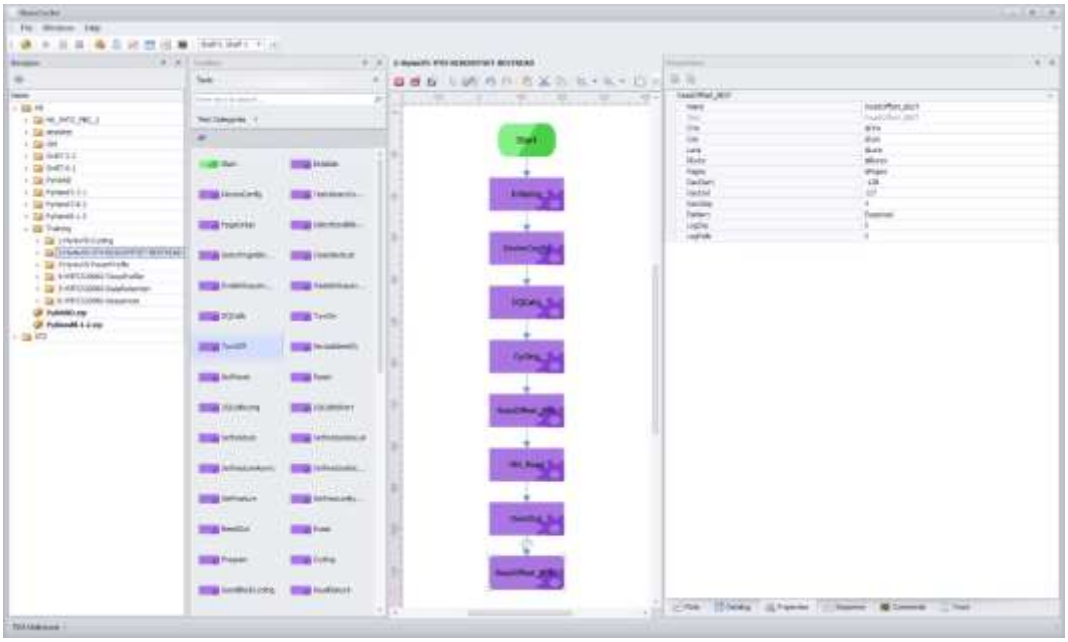


图 7-7

- **NAND 测试过程中各种主要特性的图形输出**

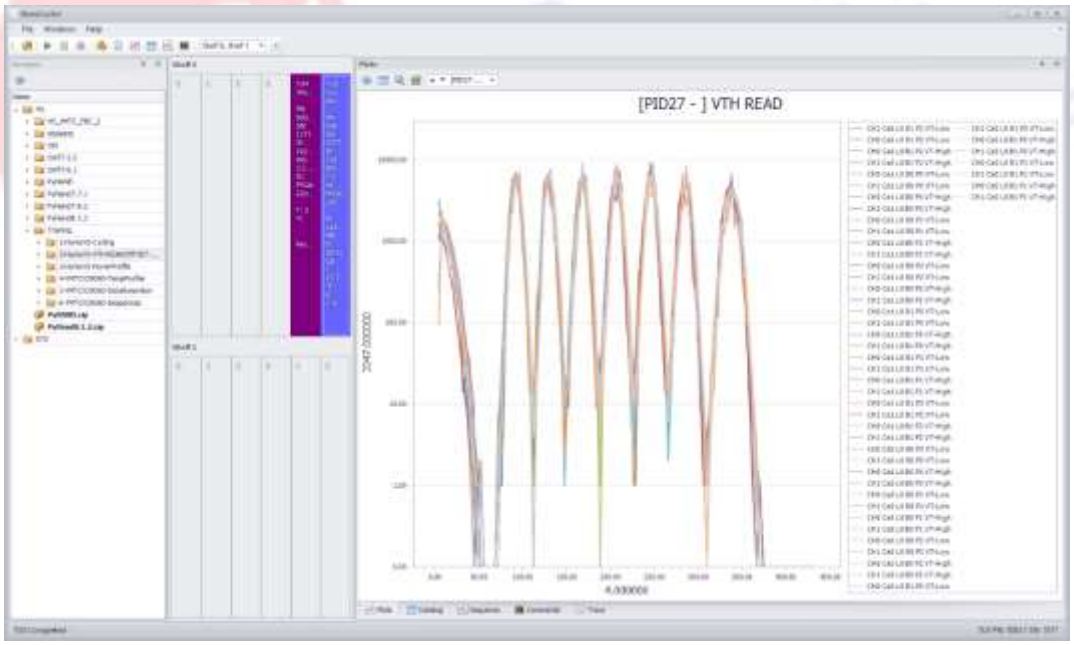


图 7-8

如果没有可靠的高性能 NAND 特性测试工具，就无法构建可靠的高性能 SSD。

对于需要了解其 NAND 特定特性的 SSD 开发人员来说，NanoCycler 是必不可少的帮助。 NanoCycler 与 ONFI 技术路线图保持同步，确保此类投资的长期有效性。

7.1.2 Nanocycler 产品图片

7.1.2.1 高密度 12 槽位设备(12-TU)

该新设计提供单台(shelf)提供 12 个 TU （Tester Unit），单个集成机架提供 7 台 shelf 总计可以达到 84 个 TU，可以同时测试 84 个 NAND Flash，非常适合于批量测试。

下图是单台 12-TU shelf 图片。该 shelf 可以单独使用，也可以安装在机架里面，下面呈现的是从机架中抽出的图片。

图 7-9

图 7-10 集成机架

7.1.2.2 传统 6 槽位测试设备(6-TU)

图 7-11

图 7-12

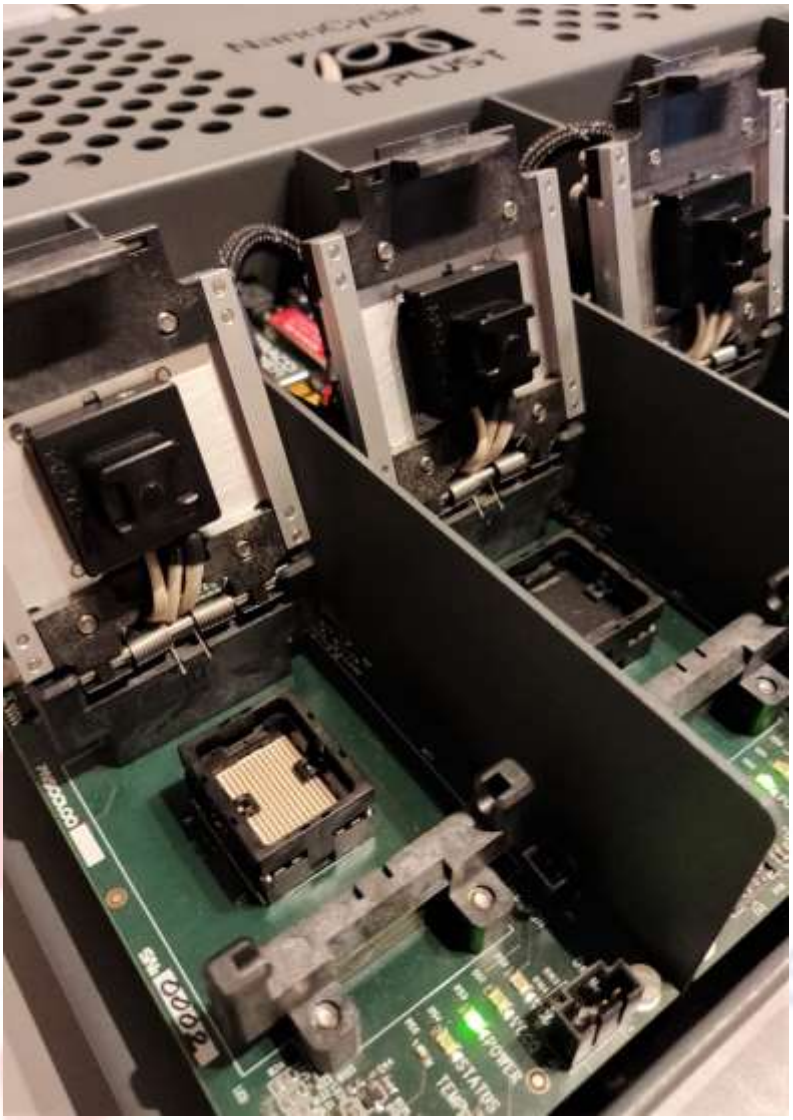


图 7-13

目前提供三种型号设备供选择：

- **Nanocycler Standard 标准版**
- **Nanocycler HS 高速版 (比标准版速度更高，1.6GT vs 800MT；提供电压拉偏和电流监控)**
- **Nanocycler FDE 版本 （主要用于 ECC/LDPC 算法优化的部门）**

7.1.3 Nanocycler Standard 和 HS 版本技术指标

产品功能	**STD800** *(Standard 800M)*	**HS16** *(High Speed 1.6GT)*	**HS24** *(High Speed 2.4GT)*
支持协议	- NV-SDR - NV-DDR - NV-DDR2 - NV-DDR3		
支持的 NAND 封装	BGA 152/132	BGA 152/132	BGA 154/152/132
支持的最大速度	800 MT/sec	1.6 GT/sec	2.4 GT/sec
Timing Modes 模式	Mode 0 - Mode 10	Mode 0 - Mode 15	Mode 0 - Mode 19
Pattern 发生器	- 高质量伪随机码 - 0 开销 - 全 0, 全 1, 替换轮流 - 用户从指定文件导入的 pattern		
数据采集	- 每 chunck 或者每 page 的 0 开销的 fail bit 计数 - 完整的 bitmap 上传到管理电脑硬盘		
时序分析	- 1us 信号捕获	- 1ns edge placement - 20ns 信号捕获	
电压拉偏功能	- 通过跳线设定 - 支持编程打开/关闭	- 支持 API 编程 - 支持编程打开/关闭 - Vccq: 0.95V~1.9V, 1000mA - Vcc: 1.0V~3.8V, 500mA - Vpp: 10V~15V, 100mA	
电流测量	分流电阻到外置探针进行测量	- 波形抓取 - 峰值和平均值捕获 - 2k 的采样 buffer 空间 - 硬件实现平均值计算 50ns 采样率 1mA 分辨率	
温度控制	- 支持室温，例如 25°C~125 °C (可以放置在 0°C 的温箱里面) - 1°C 的精准度 - 从室温加热到 125 度大概 5 分钟，然后再降温到室温约 3 分钟，总计大概 8 分钟		
产品架构	- 开发平台（单 socket） - 桌面系统（6 个 socket) - 机架系统（24 或者 48 个 socket，通过 4 或者 8 个桌面系统堆叠） - 高密度系统（84 个 socket，7 个 12 socket 的桌面系统堆叠） - 每个 socket 可以运行独立的测试：不同的测试脚本，nand 类型，温度访问，异步运行和停止 - NAND 的追踪信息再多个系统间共享		

7.1.4 Nanocycler FDE 版本技术指标

Protocols:	• NV-SRD • NV-DDR • NV-DDR2 • NV-DDR3
Max Transfer Rate:	• 1.6 GT/sec
Timing Modes:	• Mode 0 - Mode 12
Pattern Generator:	• High Quality Random • Solid, Alternate • User Pattern from File
Data Collection:	• RBER per Chunk • Readout to File
Voltages (programmable):	• Vcc: 1.0 .. 3.8V 500mA • Vccq: 0.8 .. 2.0V 1000mA • Vpp: 10..15V 100mA
Current Measurement:	• Icc, Iccq, Ipp • Dynamic, 1k samples • 10 kHz .. 1 MHz rate
Temperature Control:	• Room .. 125°C • 1°C accuracy • Full range in cc.10 min

图 7-14

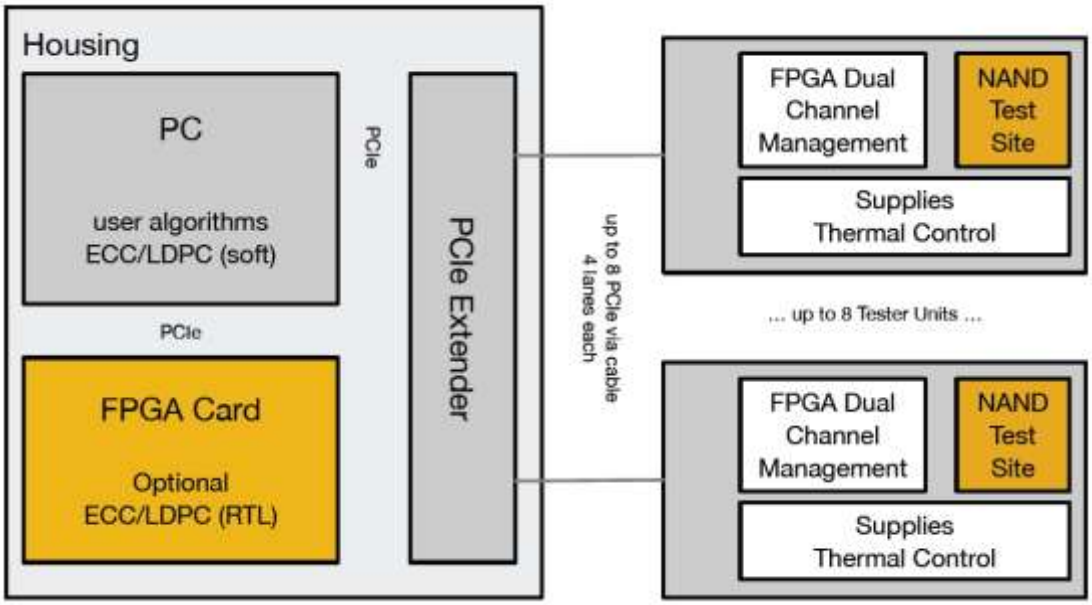


图 7-15

7.1.5 Nanocycler 标准版基本功能

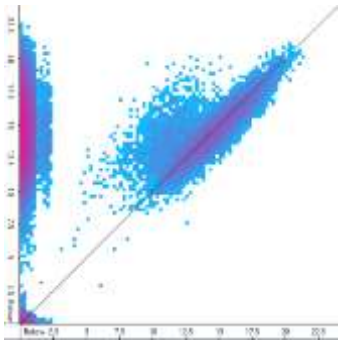
- **Straightforward management**

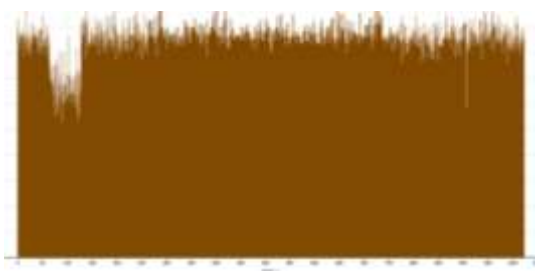
  - Easy test setup supported by a large set of built-in experiments
  - Integrated data collection
  - Desktop installation
- **Large number of independent experiments in parallel**
  - Per package thermal control
  - Independent test per package
  - Available in 1, 6, 12, 24 and 48 package configurations
- **ONFI Compatible**
  - BGA-152 and BGA-132 packages
  - Dual-channel multi-die testing
  - Up to 800 MT/sec

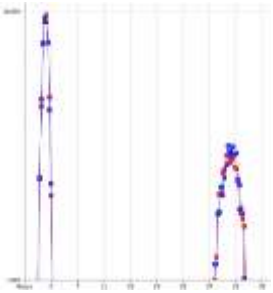
7.1.6 BarnieMAT 后处理分析软件

7.1.6.1 BarnieMAT 后处理分析基本功能

- **Worldwide N1 bitmap analysis software**
- **Very fast array data processing engine**
- **Easy adaptation for new device types / technologies**
- **Python programmability**
- **Reporting and charting tools**
- **Post-processing**
- **Injected in the BarnieMAT data analysis platform**
- **“Quite Big Data”**
- **AI experimental platform**







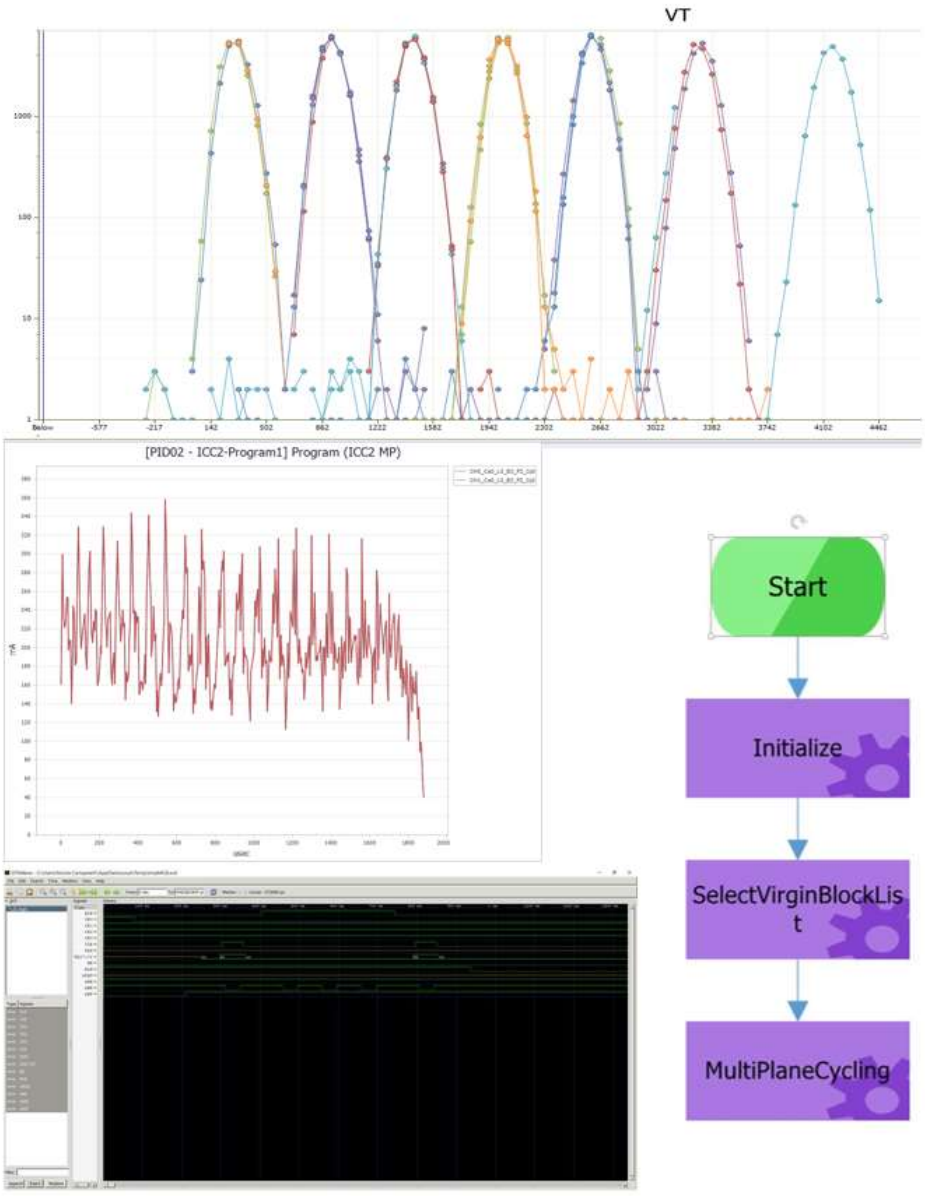

VT
[PID02 - ICC2-Program1] Program (ICC2 MP)
Start
Initialize
SelectVirginBlockLis
t
MultiPlaneCycling

图 7-16

7.1.6.2 BarnieMAT 后处理分析展示

7.1.6.2.1 BarnieMAT – Icc3-Erase Current

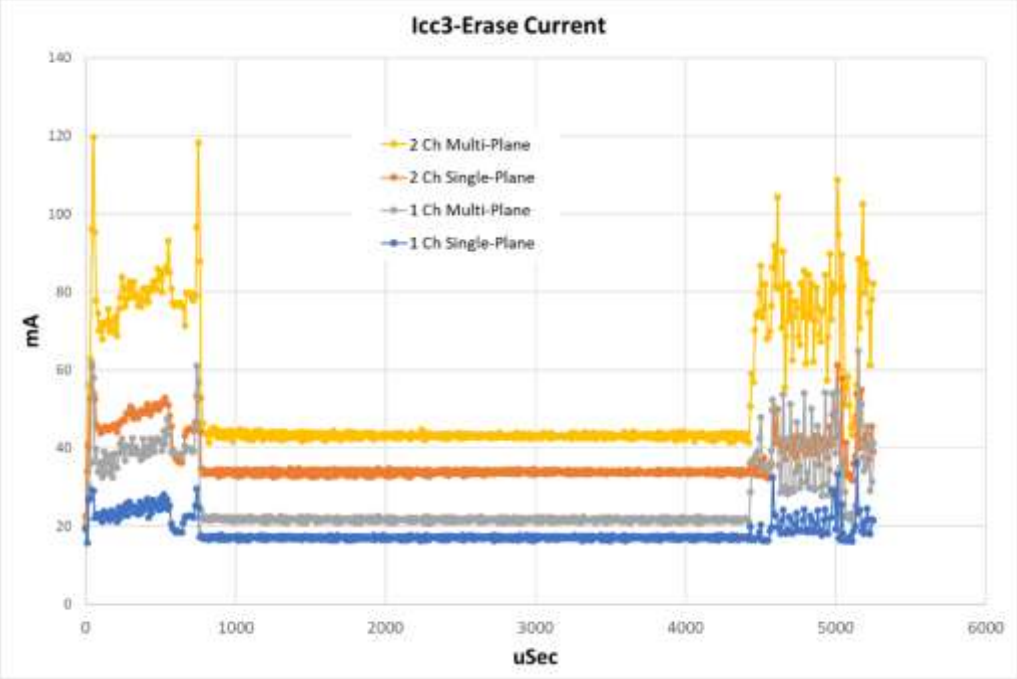


图 7-17

7.1.6.2.2 BarnieMAT - Improvement via Characterization

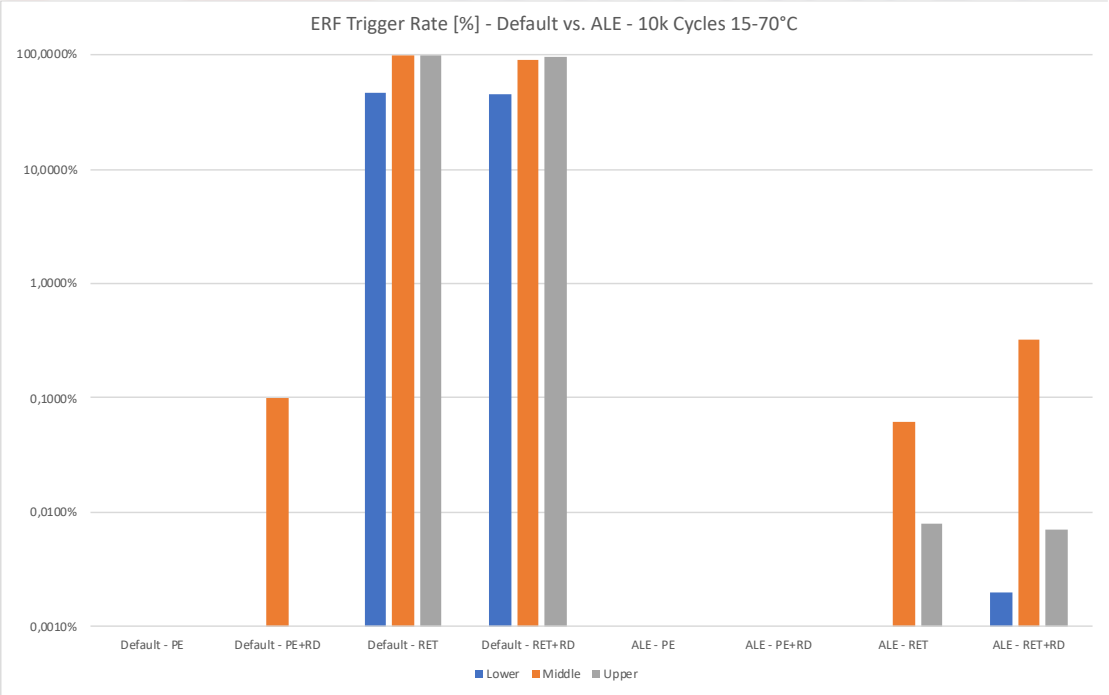


图 7-18

7.1.6.2.3 BarnieMAT - Temperature Profile with Dice

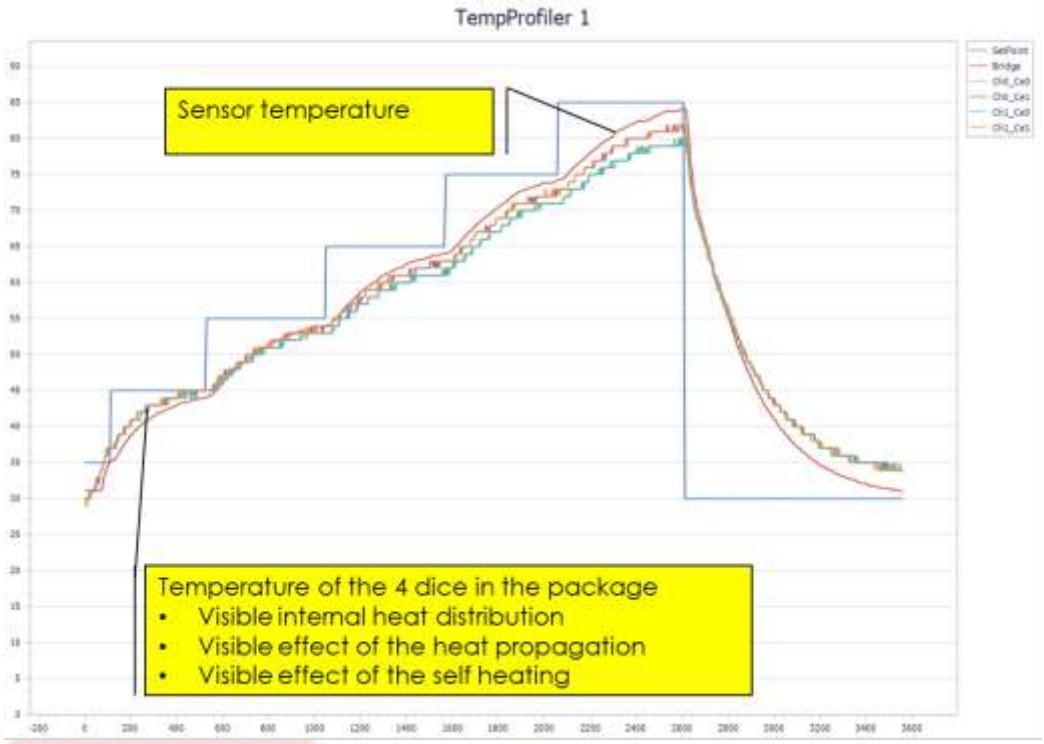


图 7-19

7.1.6.2.4 BarnieMAT - Program/Erase Times

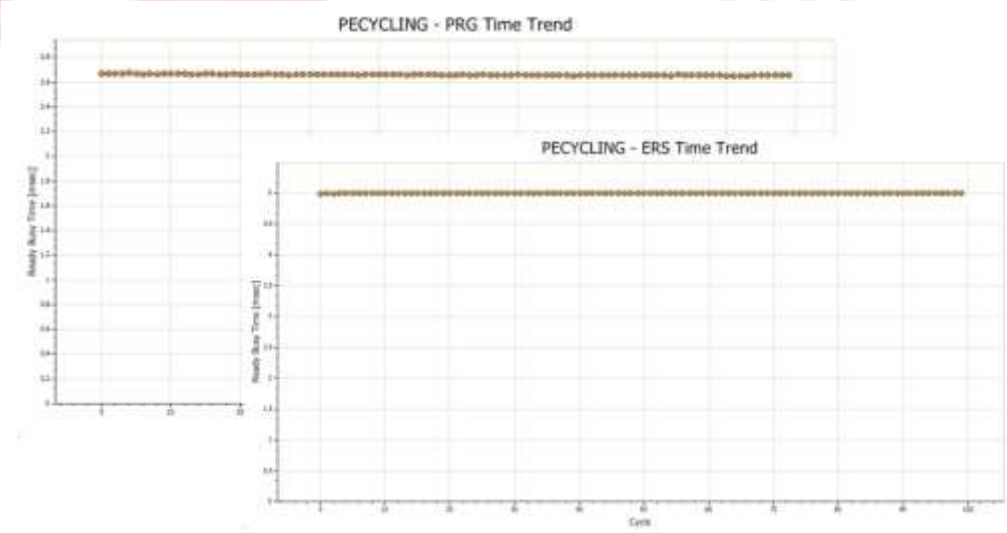


图 7-20

7.1.6.2.5 BarnieMAT - Read Times per Page Level

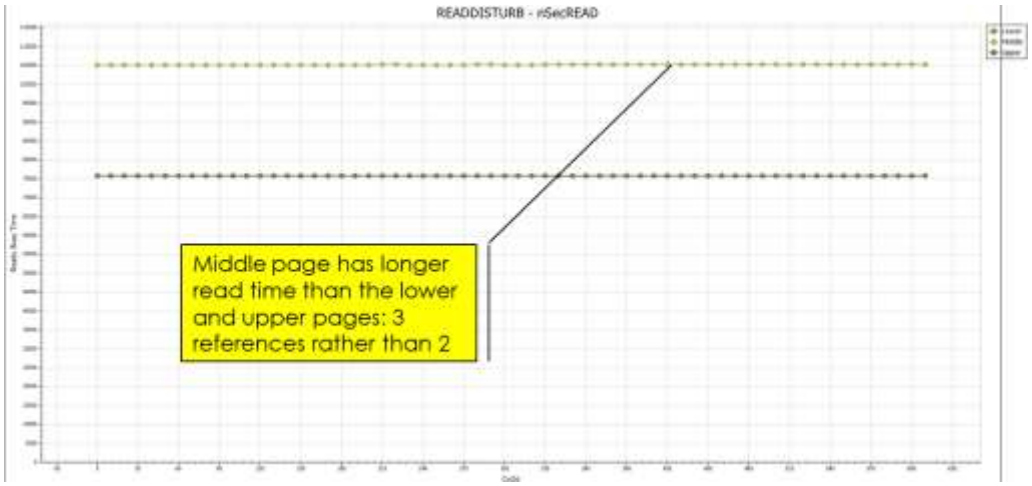


图 7-21

7.1.6.2.6 BarnieMAT - BER Distribution Trend

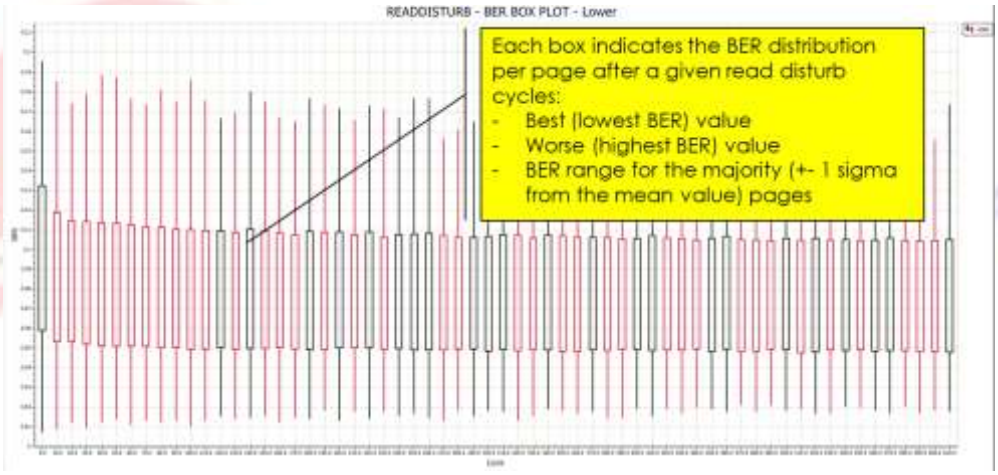


图 7-22

7.1.6.2.7 BarnieMAT - Page Fail Count Distribution

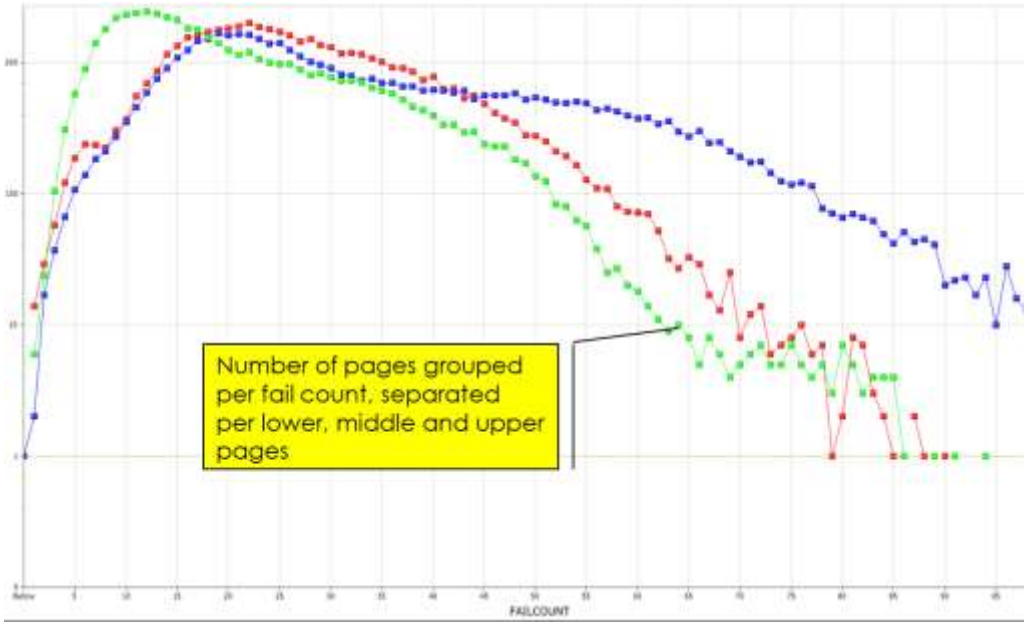


图 7-23

7.1.6.2.8 BarnieMAT - Read Retry Option Analysis

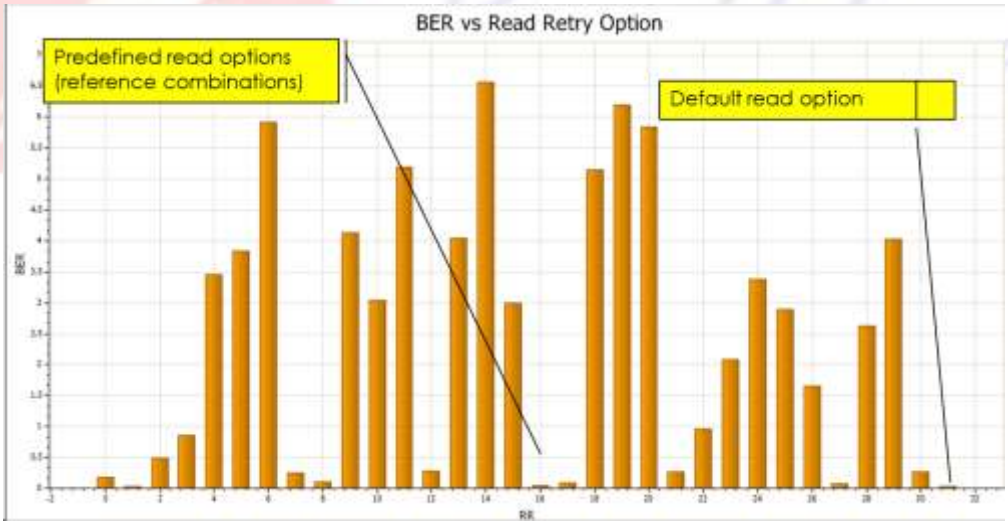


图 7-24

7.1.6.2.9 BarnieMAT - Number of Failing Bits per Level

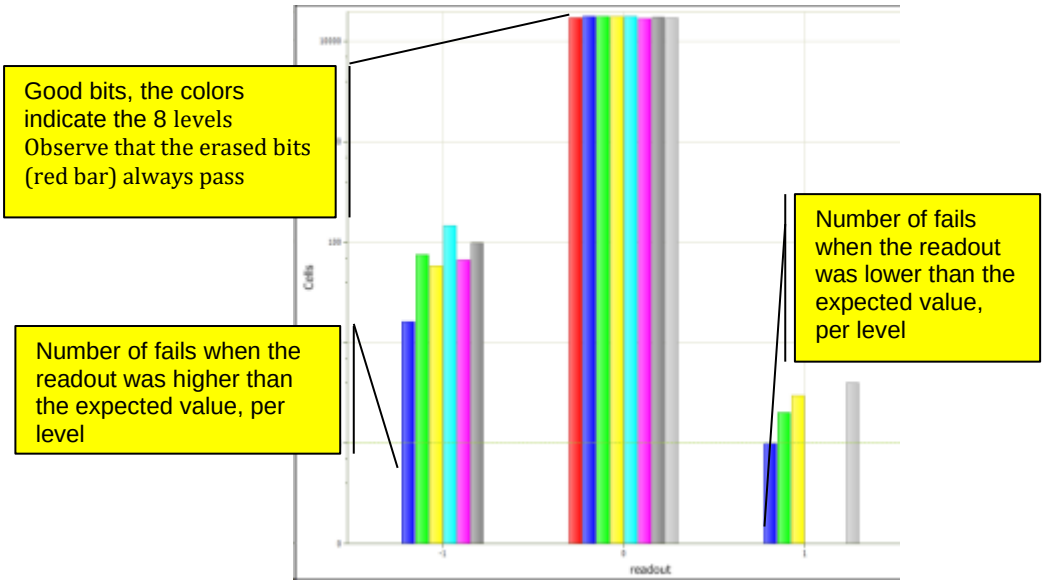


图 7-25

7.1.6.2.10 BarnieMAT - Vt Shift Moving References

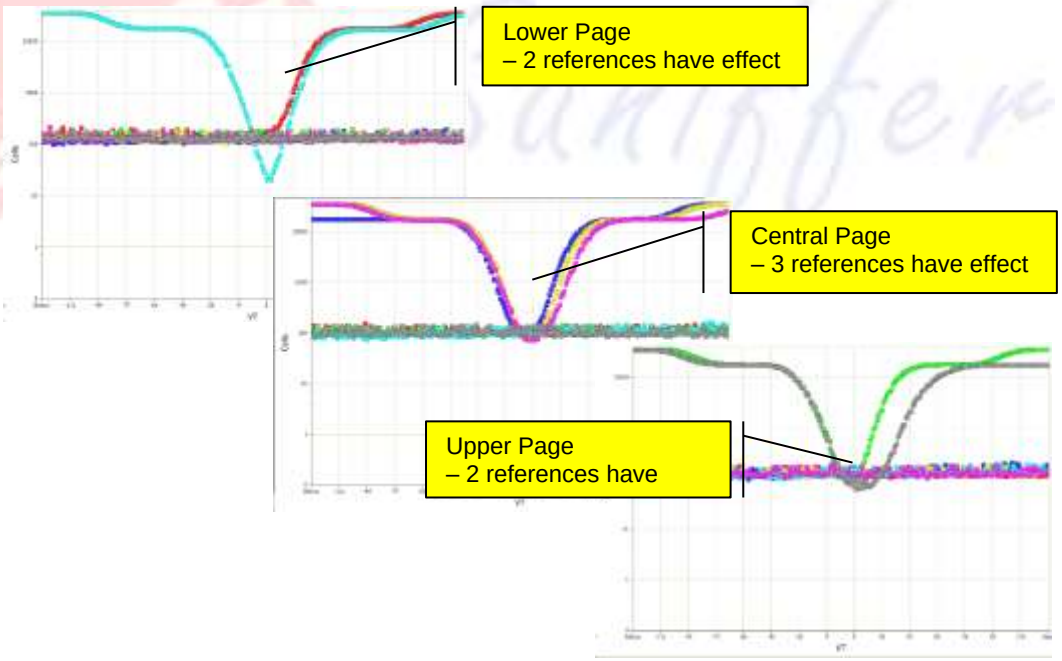


图 7-26

7.1.6.2.11 BarnieMAT - Vt Distribution

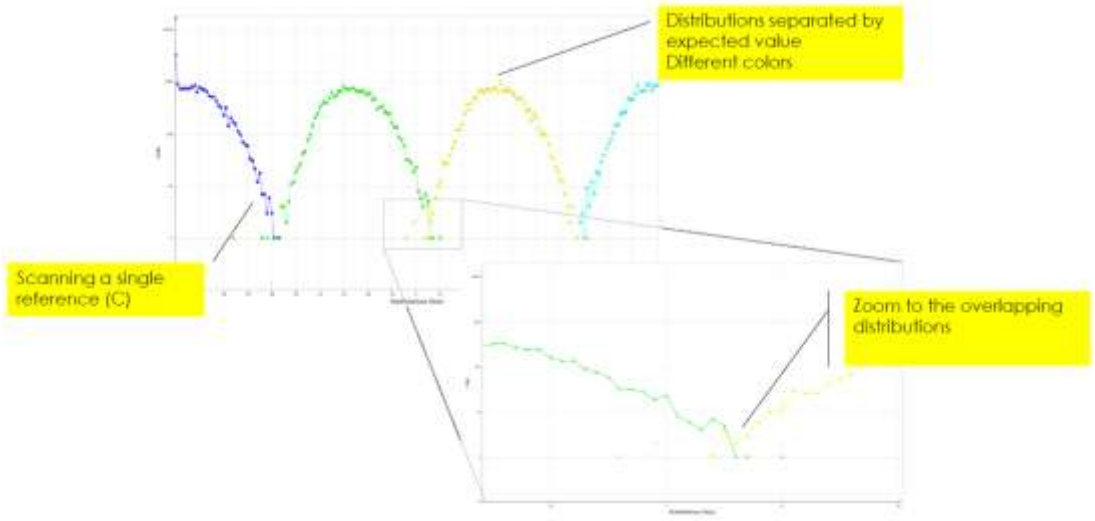


图 7-27

7.1.6.2.12 BarnieMAT - Cell Population Move

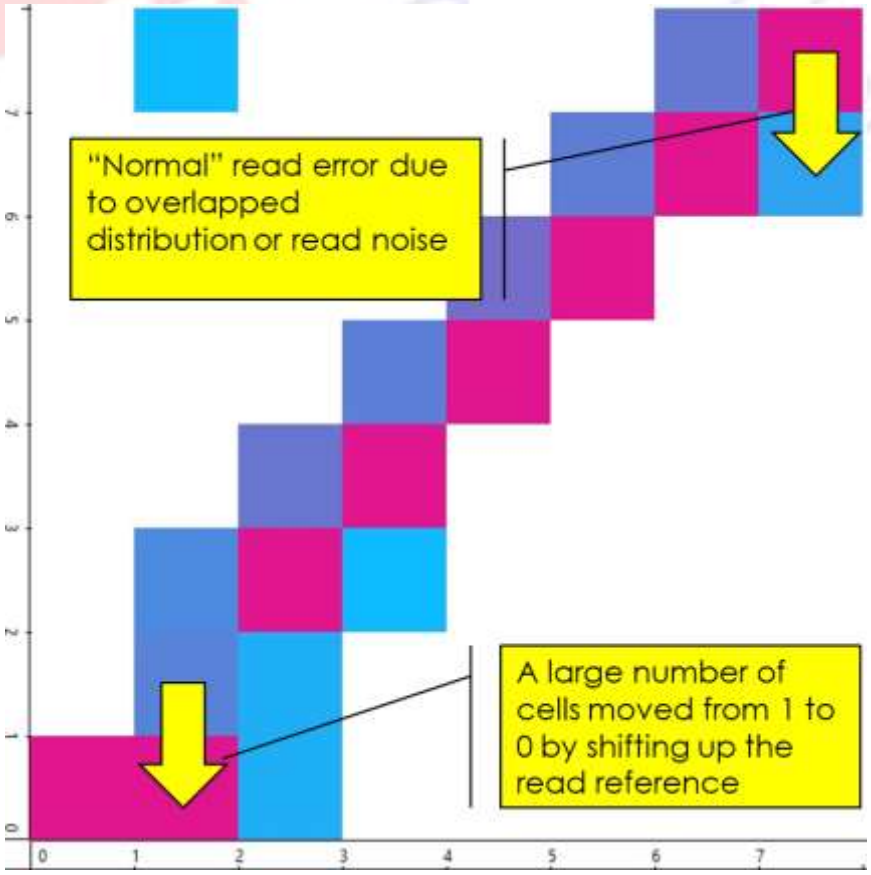


图 7-28

7.1.6.2.13 BarnieMAT - Topologic View of Fails

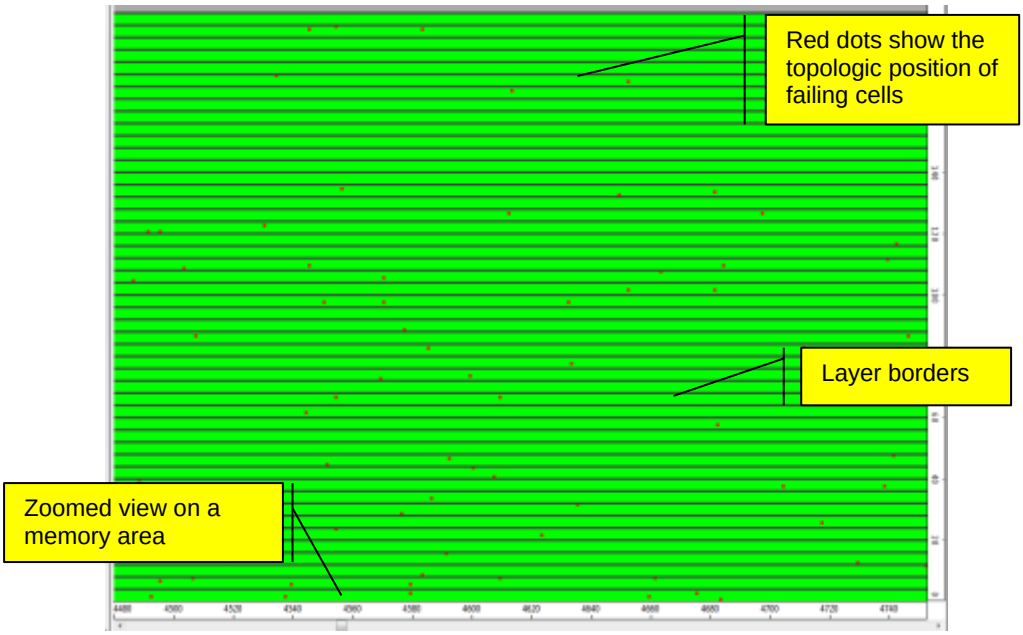


图 7-29

7.1.6.2.14 BarnieMAT - Fail Distribution per Layer

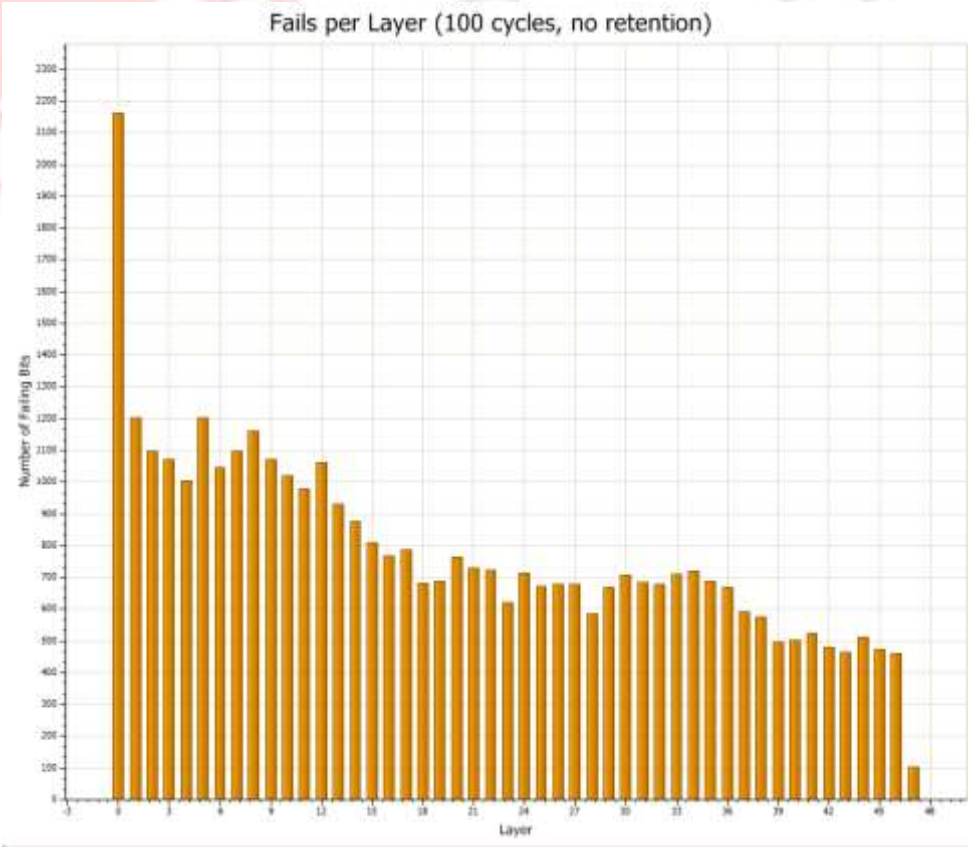
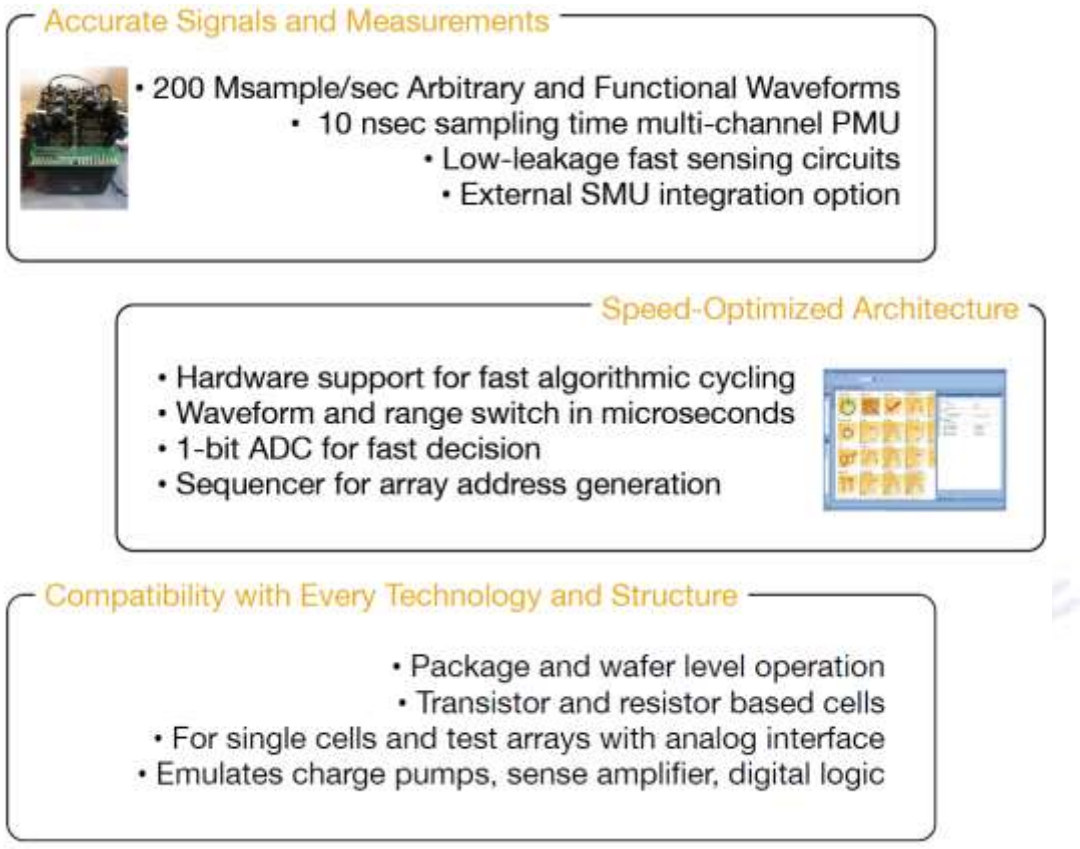


图 7-30

7.2 新型闪存开发电参数测量、特性测试和分析平台 TESTMESH TMA-100

Characterization Platform for NVM Technology Development

Millions of Cycles in a Flash



下面是该平台的一些基本功能 -

Feature	Performance
Waveforms	▪ up to 12 channels ▪ Arbitrary and algorithmic waveforms ▪ Multiple waveform storage with fast switching ▪ -12V..+12V, 5 nsec step time, 100mA, 0 or 50 Ohm impedance

PMU	▪ Operation modes - Voltage force current measurement - Current force voltage measurement - High impedance voltage measurement - Pass-through voltage measurement - Pass-through current measurement ▪ Measurement ranges -±1uA, ±10uA, ±100uA, ±1mA, ±10mA, ±100mA, 2A -±1.2V, ±12V ▪ Sampling - 10 nsec sampling time - On-the-fly averaging - 2k sample buffer - Features - Peak detection, averaging, 1-bit decision
Digital IO	• 32 channels (extendable) with per-pin direction control • Vih 1.8V .. 6.0V programmable in groups of 8 • 100 MHz engine (full speed in the low voltage range)
Power Supplies	• 4 channels • 0.5V..7V, 2A each • Programmable current limit • Built-in ramp generator
References	• Dual-channel voltage reference -±12V 20mA • Single-channel current reference -±1mA 5V
Architecture	• Dedicated PMU for each group of 4 waveform generators • Additional PMU for the digital IOs, supplies and references • Switch matrix for output configuration • Integration of external SMU

7.3 NAND 协议分析仪

NAND 闪存协议分析仪是一种帮助验证 NAND 闪存的功能的工具。NAND 闪存是一种非易失性存储技术，可用于固态硬盘（SSD）、移动电话存储、嵌入式存储卡、USB 设备等等，NAND 闪存协议分析仪提供了全面的协议、方法、验证和生产力功能，使用户能够实现加速验证闭环。

概述

NAND 协议分析仪由于当前 NAND 速度越来越快，同时分析仪需要同时抓取很多通道，所以采集内存 buffer 深度往往可以配置到高达 512GB，最高采样率高达 20GHz。这样高速、强大的内存跟踪解决方案使调试和验证高达 2.4 Gb/s 的基于 NAND 的子系统更便利。

下面是 NAND 协议分析仪的一些特性：

- **Maximum Sample Rate**
  - **Max. 20 GHz**
  - 1.25 / 2.5 / 5 / 10 /20 GHz options
- **Number of Channels: total 34 CH**
- **Single-ended channels (31 CH)**
  - **Per-pin termination level**
  - **Per-pin threshold level**
  - **Per-pin timing delay**
- **Differential channels (3 CH)**
  - **1 Differential clock input**
  - **2 Differential strobe input**
- **Analysis mode**
  - **Timing mode**
- **Acquisition Memory**
  - **Max. 512 GB (standard 128 GB)**
  - **Full memory depth up to 20 GHz**
  - **Hardware data compression**
- **Hysteresis settings for all channels**
- **Host Interface: PCIe Gen3 x8**
- **Target application**
- **High Speed NAND Device**

用于数据采集的大内存

高达 512GB 的大内存深度允许通过长时间捕获来调试非常复杂的问题。

超快的采样速度

高达 20 GHz 的超快采样速度使您能够以高达 2.4 Gb/s 的速率捕获所有通道的数据，而无需任何额外选项。

探针抓取信号方案

为了最大限度地减少探头系统中的信号衰减，模拟模块与数字模块分开，并尽可能靠近设备放置。

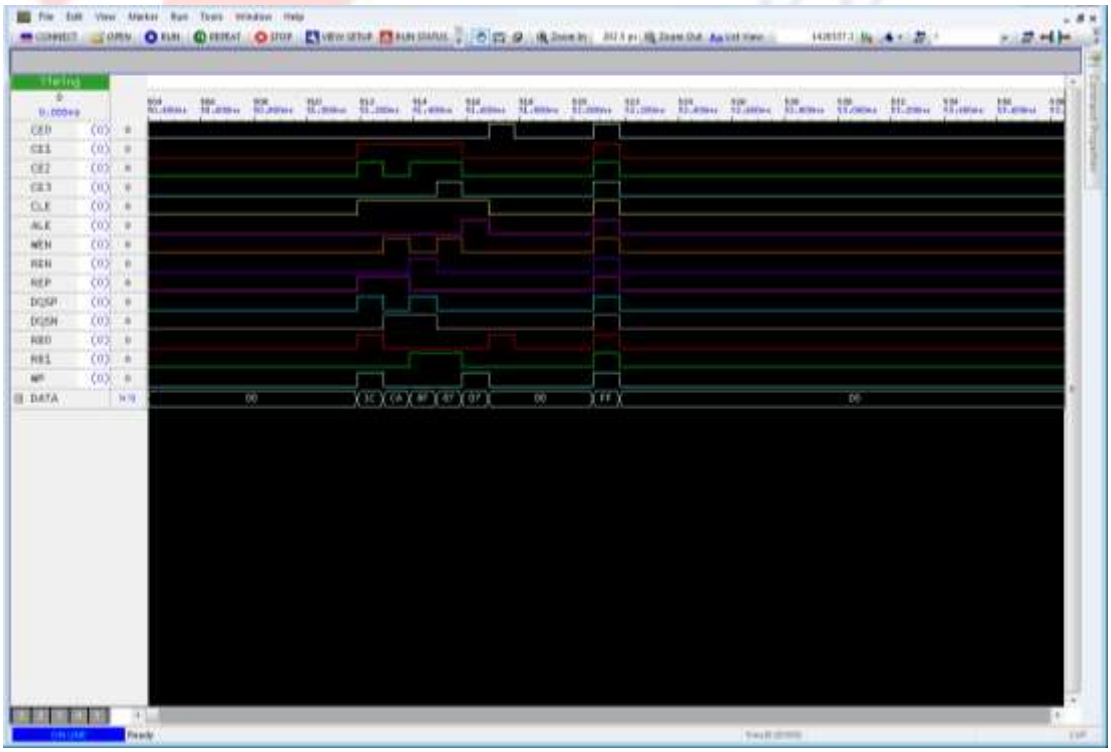
基于 interposer 的抓取信号方案 – 串接在 NAND 和基板之间将信号导出到 NAND 分析仪

我们提供各种中介层和连接解决方案

软件特点

- **通过 TeraView 2.0 GUI 图形用户界面可以快速地捕捉、 定位并分析 NAND 协议。**
- **轻松设置和配置 窗户。**
- **基于软件的定时模式眼 图表。**
- **可编程的 包 解码器 和 API 调用使得用户可以实现自动化测试。**

TeraView 2.0 Wave 和列表视图



Channel Specification	
Number of Channels	34 (31 single-ended + 3 differential)
Input Voltage Range	-0.2V to 1.2V (analyzer input)
Minimum Input Voltage Swing	40 mV (analyzer input)
Termination (Vtt) Voltage Range	0V to 1.2V (analyzer input)
Termination (Vtt) Voltage Resolution	1 mV
Threshold (Vref) Voltage Range	0V to 1.2V (analyzer input)
Threshold (Vref) Voltage Resolution	1 mV
Hysteresis (Vhys) Voltage Range	0 ~ 50 mV
Timing Delay Range	-2ns to 2ns
Timing Delay Resolution	50 ps @20 GHz
Input Impedance	50 Ohm (analyzer input) 250 Ohm (5:1 probe) 500 Ohm (10:1 probe)

Acquisition Specification	
Sample Rate	1.25 / 2.5 / 5 / 10 / 20 GHz
Analysis Mode	Timing mode, Filter mode, Eye mode
Acquisition Memory	Max. 512 GB (standard 128 GB)
Host Interface	PCIe Gen3 x8
Data Compression	Hardware compression
Data De-compression	Software de-compression

Trigger Specification	
Trigger Pattern	E(Either edge), R(Rising edge), F(Falling edge), H(High), L(Low), X(Don't care) for all channel
Trigger Sequence	5 burst trigger
Trigger Action	Start capture
Trigger Rate	1.25G
Pre-trigger Size	Max. 128Gbytes (depends on RDIMM Size)

Filtering Specification	
Filter Pattern	E(Either edge), R(Rising edge), F(Falling edge), for all channel
Filtering Rate	1.25G
Filtering Size per a event	8 bytes including raw data and time information

NAND 协议分析仪系统配置

Environmental and physical	
Power Consumption	Max. 150W
Operation Temperature (nom)	0 to +40 dec C
Humidity (nom)	0 to 80% relative humidity
Dimensions (W x H x D)	347 mm x 290 mm x 74 mm
Weight	4.7kg

模型	数量	描述
NAND-20G	1	NAND 协议分析仪（128GB – 标准）
- 选项 256G/512G	1	256GB/512GB 内存选项
P21034A	1	软接触无连接器探头，单端，34通道
PCIe-HIB38	1	PCIe 主机适配器 (Gen3 x8)
PCIe-0802	1	PCIe 电缆 2m (Gen3 x8)

7.4 NAND 颗粒筛选和 Burn-In 测试设备

对于生产环节用到的 NAND 颗粒进行筛选(sorting) 是国内大部分 SSD 厂商 SSD 贴片生产之前基本都要做的一个步骤，当然，个别 SSD 大厂如果能得到 NAND 原厂的批次质量保证可能除外。现在由于芯片短缺，国内有些厂家的 NAND 渠道来源不正，有些是 NAND 原厂的次品，甚至是拆机版 NAND，这些都需要在生产之前进行筛选。

当然，NAND 筛选将根据产品的特性分为常温筛选，RDT 高温筛选，或者用于工业级领域 SSD 的高、低温筛选。

围绕日韩的 NAND 高密度老化测试设备（可以同时测试 4608 颗 NAND）一般价格昂贵，如下：

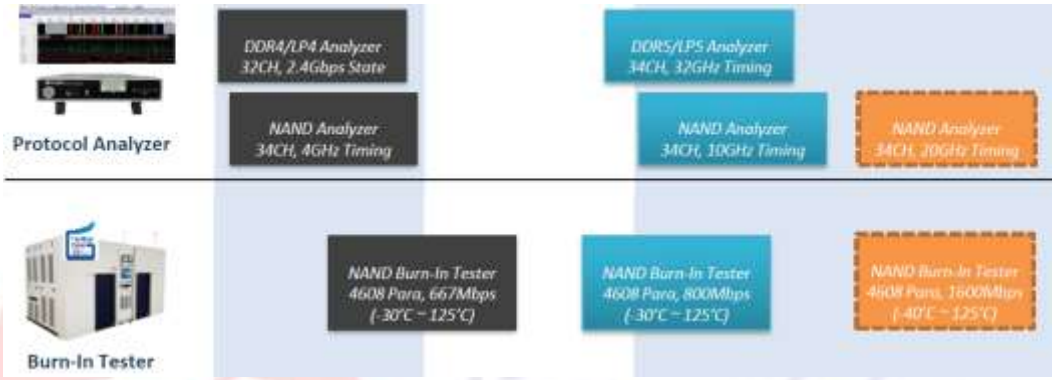


图 7-31

说明：由于 DDR5/LPDDR5 也分为日韩、美国两条阵线，所以上述产品也有针对 DDR5/LPDDR5 测试设备

7.4.1 便携式（标准版）- 8 槽位

- **容量 1-8 颗芯片**
- **尺寸 L22.5cm*W19.5cm*H2.5cm**
- **工作温度 0～+70℃**
- **产品概述 同时支持 8 颗粒同时测试，方便携带**
- **支持测试命令类型 基础测试（闪存信息检测、寿命预测、质量等级分类）等**
- **已支持测试封装规格 BGA152、BGA132（可定制扩展）**
- **已支持测试闪存品牌 Micron、Intel、YMTC、Hynix、Toshiba、Sandisk 等厂商的 SLC、MLC、TLC、QLC 类型 NAND Flash 芯片颗粒**

7.4.2 便携式（专业版）- 8 槽位

- **容量 1-8 颗芯片**
- **工作温度 0～+70℃**
- **产品概述 同时支持 8 颗粒同时测试，方便携带，功能全面**

- 支持测试命令类型 基础测试、实验测试、高阶测试（数据保持、抗干扰、自定义）等
- 支持测试范围可选 起始测试块数、块间间隔、循环数、测试时间等个性化设置
- 已支持闪存协议类型 ONFI/toggle 接口颗粒
- 已支持测试封装规格 BGA152、BGA132（可定制扩展）
- 已支持测试闪存品牌 Micron、Intel、YMTC、Hynix、Toshiba、Sandisk 等厂商的 SLC、MLC、TLC、QLC 类型 NAND Flash 芯片颗粒
- 支持 pattern 全 0，全 1，全 5，伪随机；棋盘格；字线随机等

7.4.3 生产版 – 240 槽位

- 容量 1-240 颗
- 尺寸 L180cm*W72.5cm*H187cm
- 工作温度 室温～+85℃
- 产品概述 最大支持 240 颗，可同时测 30 种不同颗粒，高温下的筛选测试
- 支持测试命令类型 基础测试（闪存信息检测、寿命预测、质量等级分类）等
- 已支持测试封装规格 BGA152、BGA132（可定制扩展）
- 已支持测试闪存品牌 Micron、Intel、YMTC、Hynix、Toshiba、Sandisk 等厂商的 SLC、MLC、TLC、QLC 类型 NAND Flash 芯片颗粒

7.4.4 科研版 – 200 槽位

- 容量 1-200 颗
- 尺寸 L185cm*W87.5cm*H183cm
- 工作温度 -40℃～+85℃
- 产品概述 可同时测试 1~200 颗芯片，宽温测试、可同时测 25 种不同颗粒
- 支持测试命令类型 基础测试
- 实验测试
- 高阶测试（数据保持、抗干扰、自定义）等
- 支持测试范围可选 起始测试块数、块间间隔、循环数、测试时间等个性化设置
- 已支持闪存协议类型 ONFI/toggle 接口颗粒
- 已支持测试封装规格 BGA152、BGA132（可定制扩展）
- 已支持测试闪存品牌 Micron、Intel、YMTC、Hynix、Toshiba、Sandisk 等厂商的 SLC、MLC、TLC、QLC 类型 NAND Flash 芯片颗粒
- 支持 pattern 全 0，全 1，全 5，伪随机；棋盘格；字线随机等

7.4.5 卓越版 – 512 槽位

- 容量 **1-512** 颗芯片
- 尺寸 **L140cm*W120cm*H230cm**
- 工作温度 **-40℃～+85℃**
- 产品概述 最高支持 **512** 颗颗粒同时测试、宽温测试
- 支持测试命令类型 基础测试、实验测试、高阶测试（数据保持、抗干扰、自定义）等
- 支持测试范围可选 起始测试块数、块间间隔、循环数、测试时间等个性化设置
- 已支持闪存协议类型 **ONFI/toggle** 接口颗粒
- 已支持测试封装规格 **BGA152**、**BGA132**（可定制扩展）
- 已支持测试闪存品牌 **Micron**、**Intel**、**YMTC**、**Hynix**、**Toshiba**、**Sandisk** 等厂商的 **SLC**、**MLC**、**TLC**、**QLC** 类型 **NAND Flash** 芯片颗粒
- 支持 **pattern** 全 **0**，全 **1**，全 **5**，伪随机；棋盘格；字线随机等

7.5 NAND 数据读取和恢复工具

7.5.1 VNR (Visual NAND Reconstructor) 软件

The built-in database of NAND chips and controller configurations provides a solution for most chips known to date. The trusted database is regularly updated by VNR developers and technological partners. The user database can be used to store your configs and solutions.

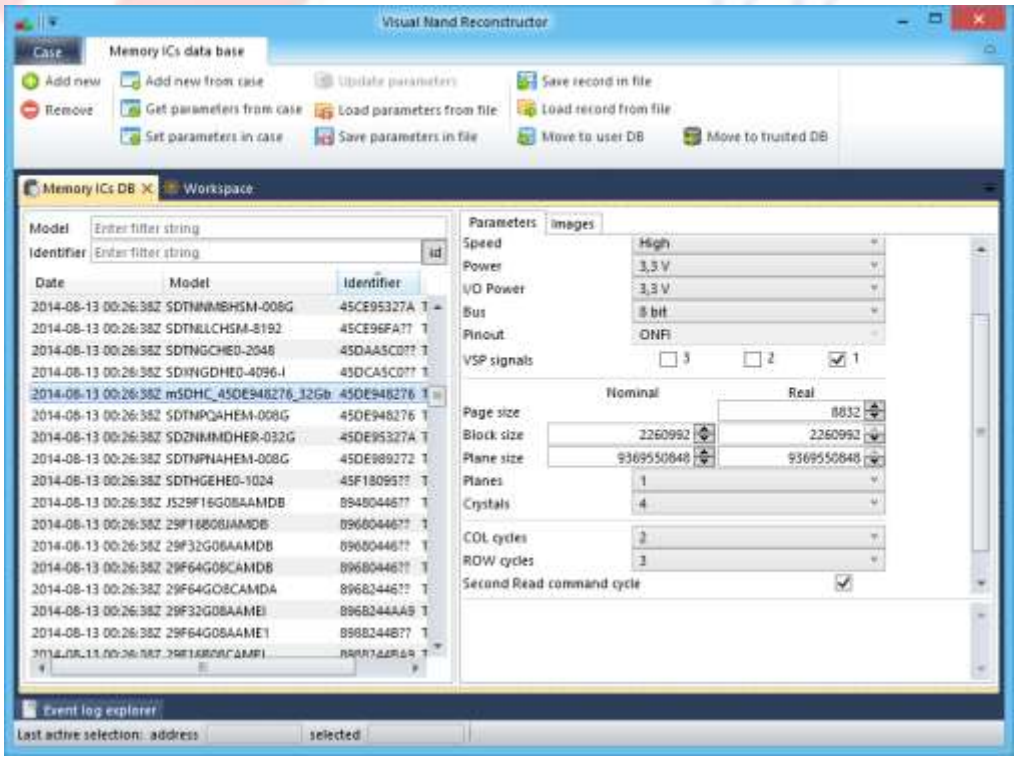


图 7-32

User friendly and intuitive software concept with unified functions helps to recover and analyze data faster and more efficient.

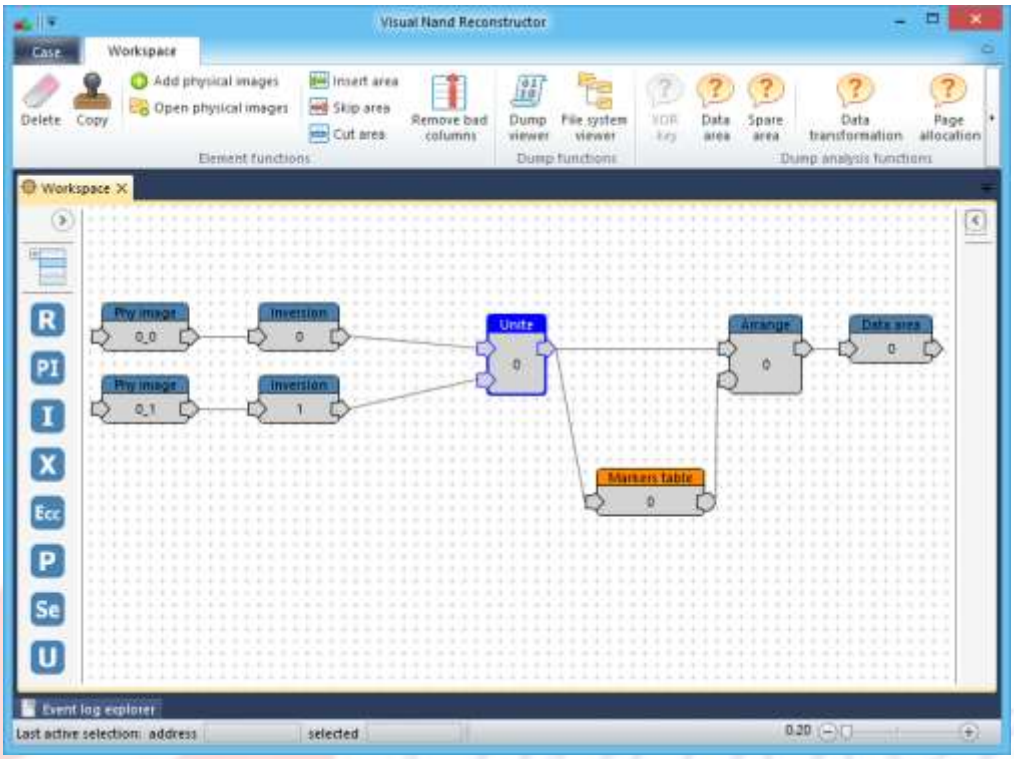


图 7-33

Automatic analysis modes make image reconstruction and data recovery process easier and faster.

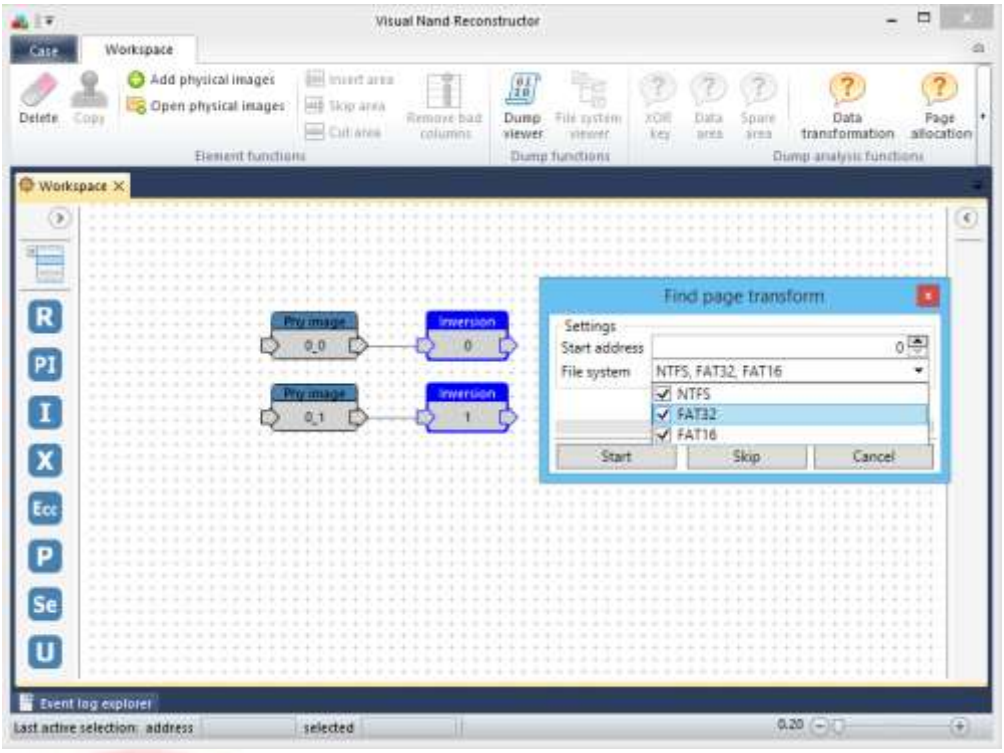


图 7-34

Unique modes of data visualization with multi-level image structure description makes reverse engineering easier.

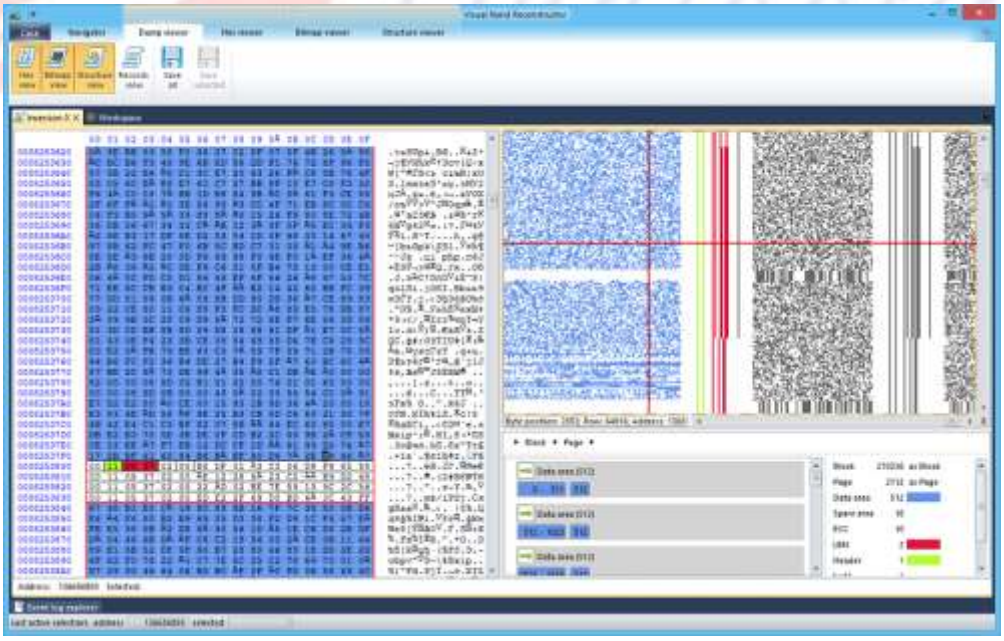


图 7-35

Special scrambler (XOR) extraction mode helps to “decrypt” data even in cases with new devices

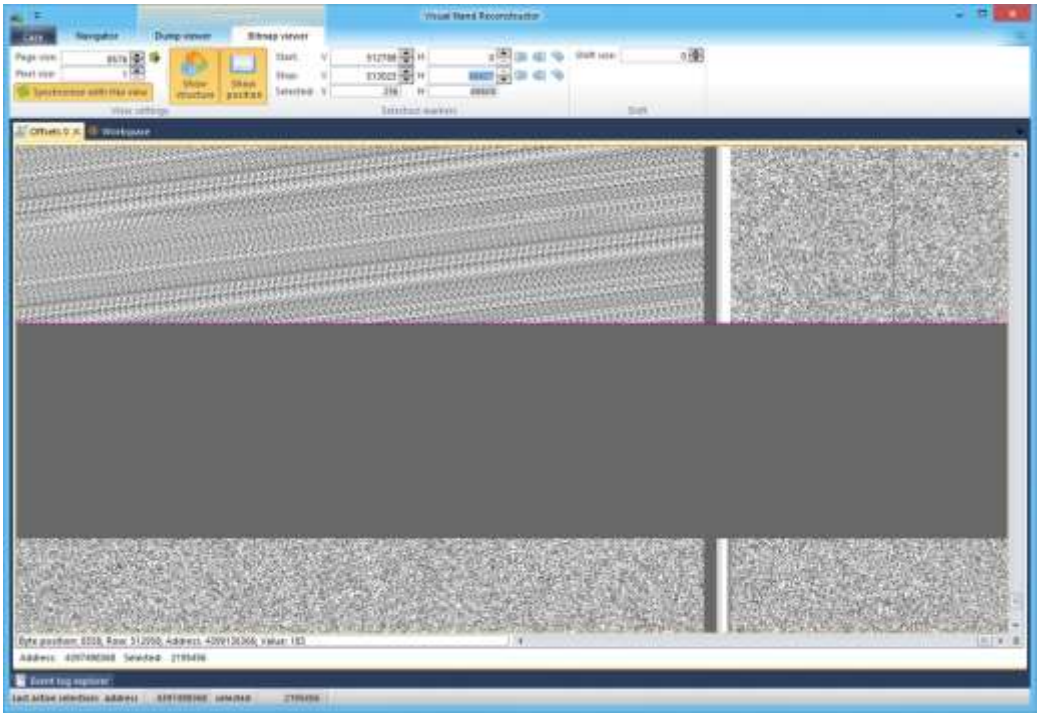


图 7-36

File system viewer works with most common file systems of flash devices – FAT and NTFS.

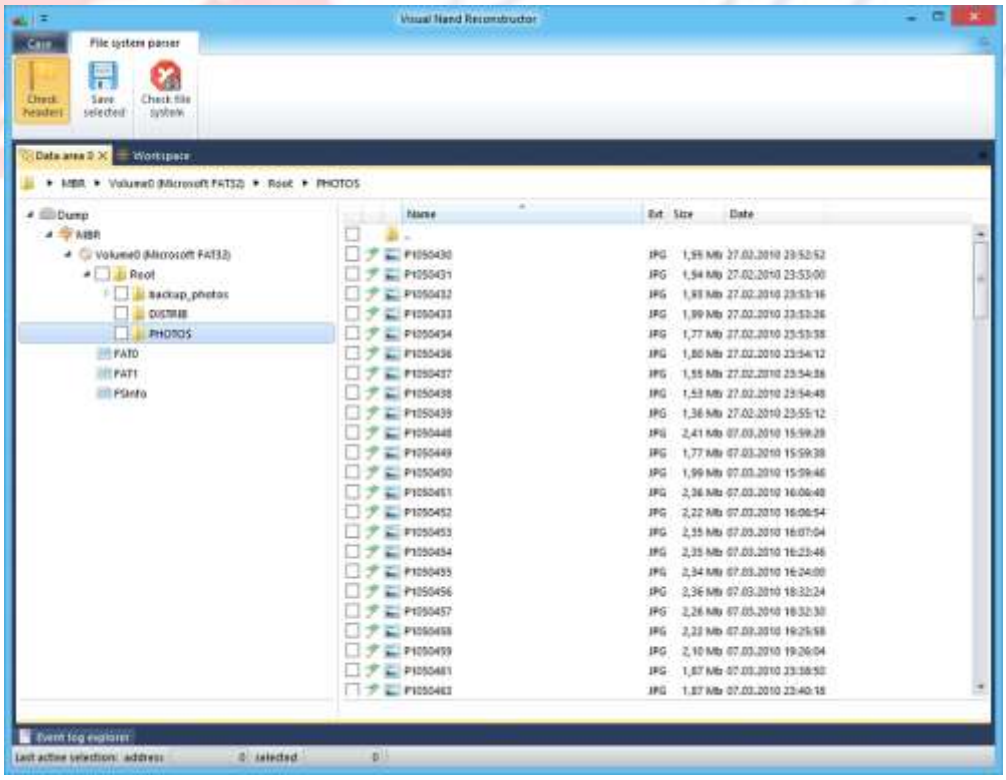


图 7-37

*** The software works with original reader and adapters only.*

Please note that the appearance of adapters is subject to change without prior notice.

7.5.2 VNR READER

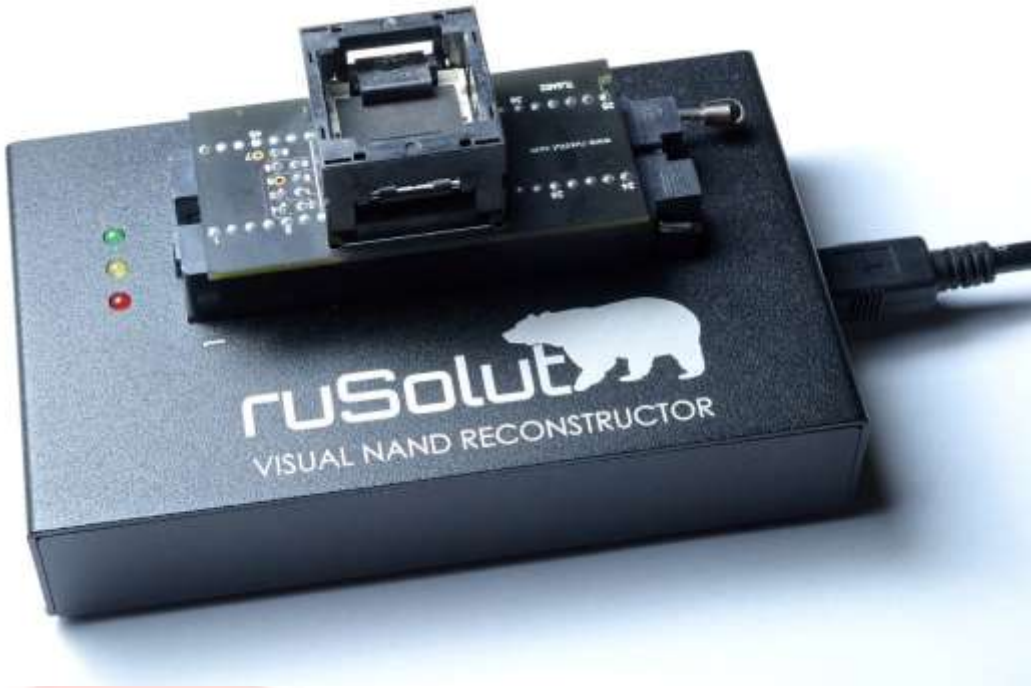


图 7-38

- **Functions**
  - Read NAND flash memory chips
- **Supported NAND packages**
  - TSOP48, LGA52, LGA60, TSOP56, BGA100, BGA152, BGA154, BGA224, Monolithic chips
  - Samsung, Sandisk, Hynix, Toshiba, Intel, Micron and others
- **Data transfer protocols**
  - Asynchronous ONFI, DDR, WL tripple address, WL tripple address with DDR
- **NAND architectures**
  - SLC, MLC, TLC
- **Power adjustment**
  - Power adjustment of Core and I/O ports of NAND chips from 1.6V to 4.0V. This feature is important for all 1.8V NAND chips. Power adjustment also helps to reduce bit errors that appear while memory chip reading under standard 3.3 Volts. Voltage level can be adjusted through software.
- **NAND access modes**
  - Read NAND physical image to file (data recovery and digital forensics)
  - Real-time access (Bit error estimation, NAND configuration analysis)
- **I/O data bus**
  - According to ONFI standard and Samsung recommendations the reader supports 8-bit and 16-bit data bus.

- **Speed**
  - Data transfer rate is 7-10 Mb/s depending on NAND chip
- **Interfaces**
  - Mini-USB 2.0 interface for connecting to PC
  - ZIF interface for adapter connection
- **LED indicators**
  - Green – USB power, Yellow – NAND power, Red – Error
- **Operating system requirements**
  - Windows driver for x64/x86 Windows7/8/10 platforms
  - Recommended operating system is windows 10 x64

STANDARD KIT

图 7-39

STARTER KIT

图 7-40

7.6 NAND 测试工装和夹具

7.6.1 NAND Flash Memory Interposers

Flash Memory Interposers

Interposers

Flash memory technologies are defined by the Open NAND Flash Interface (ONFI) industry workgroup and Joint Electronic Devices Engineering Council (JEDEC) for use in handheld devices or applications where low-power and small size is critical.

Nexus Technology offers high quality and high fidelity interposers, enabling the industry to confidently and accurately gain access to Flash memory buses for debug and compliance verification. ... more...

Flash Interposers

	Interposers				Options	
	Oscilloscope			Logic Analyzer		
Package	EdgeProbe™	Socketed	Direct Attach	Socketed	Riser	Component Socket
132 Ball NAND	✓	✓	*	✓	Yes	Yes
152 Ball NAND	✓	✓	*	✓	Yes	Yes
153 Ball eMMC	*	*	✓	*	Yes	No
172 Ball NAND	*	*	*	*	Yes	No
304 Ball NAND	✓(XH)	*	✓(XH)	*	Yes	No
316 Ball Onfi	*	*	*	✓	Yes	No
Custom	Custom designs are also available. Please contact us.					

*If you don't see what you need, please contact us for the most up to date information.

Attachment Service

Nexus Technology's expert attachment service provides a ready-to-go test solution customized to your application. We will attach the interposer and any additional accessories to your application's target. We can also power-on and test your application to confirm functionality.

Attachment Service

More Information on Interposer Types

Electrical Analysis

Electrical analysis is enabled by using either an EdgeProbe, High Density, or Socketed interposer to capture memory activity on an oscilloscope. The oscilloscope is then used to debug, analyze, and verify the analog characteristics of your design. Presenting an accurate representation of the signals under test to the oscilloscope is critical. Nexus interposers provide an unobtrusive interconnect and accurate signal to your oscilloscope. Learn more...

Logic / Compliance Analysis

Logic analysis is performed using a logic / compliance interposer to capture memory activity on a logic or memory analyzer. The logic or memory analyzer is then used to debug, analyze, and verify the logic (basic protocol) of your design. Compliance analysis uses the same interposers to capture activity on a memory analyzer. The memory analyzer is then used to debug, analyze, margin test, performance analyze, and verify the memory protocol. Learn more...

Product List

- NAND 304 Ball XH Series EdgeProbe Interposers
- NAND 304 Ball XH Series Direct Attach Interposer
- NAND 132 Ball Oscilloscope Socketed Interposer
- NAND 152 Ball Oscilloscope Socketed Interposer
- ONFI 316 Ball Logic Compliance Interposer
- eMMC 153 Ball Direct Attach Oscilloscope Interposers
- NAND 132 Ball Logic Compliance Interposer
- NAND 152 Ball Logic Compliance Interposer
- NAND 132 Ball EdgeProbe™ Interposer
- NAND 152 Ball EdgeProbe™ Interposer

图 7-41

7.6.2 NAND 152 Ball Logic Compliance Interposer

图 7-42

7.6.2.1 Logic/Compliance Interposer

Optimal Logic/Compliance validation requires analysis of the signals as seen by the memory components. This allows for the highest confidence that the signals captured are representative of the signals on the target.

7.6.2.2 Product Configuration Table

Nomenclature	Description	Component Socket
NEX-NAND152CA2	2 Channel ONFi 152 ball NAND Flash Logic Analyzer memory interposer with solderballs only (no memory socket).	No
NEX-NAND152CA2-SK	2 Channel ONFi 152 ball NAND Flash Logic Analyzer memory interposer with solderballs and memory socket installed.	Yes

**Single channel support is also available. Please contact us for more information.*

7.6.2.3 Available Accessories

Type	Desc.	Quantity	Nomenclature
Riser	Riser elevates interposer .050″	1	NEX-RSRNAND152
Socket	Memory socket	1	NEX-SOCKETNAND152
Solder Balls	Install solder balls on socket	1	NEX-OPT-SOLDERBALLS-SK

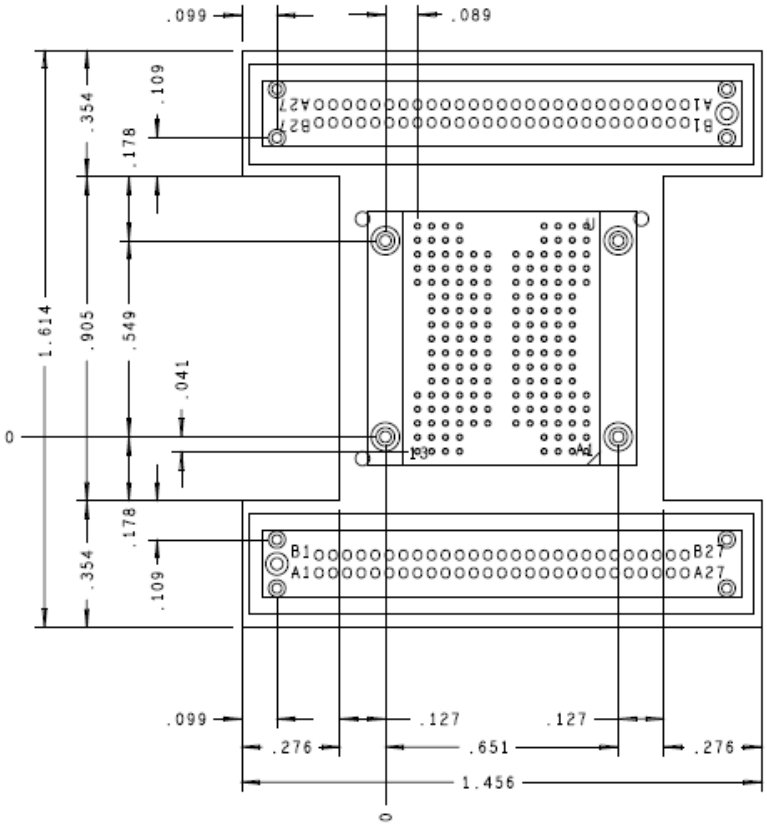


图 7-43 NAND 152 BALL LOGIC SOCKETED MECHANICAL

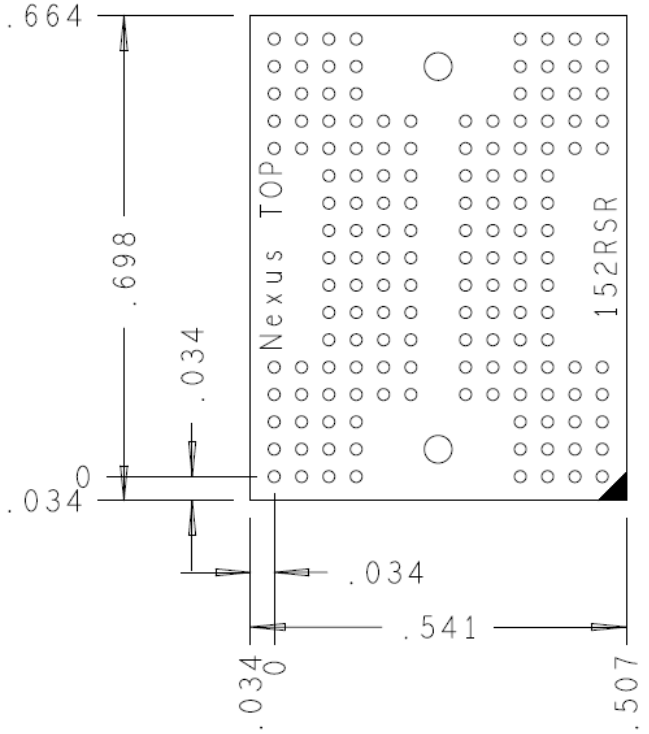


图 7-44 NFLASH 152 BALL RISER MECHANICAL

7.6.3 NAND 152 Ball Oscilloscope Socketed Interposer

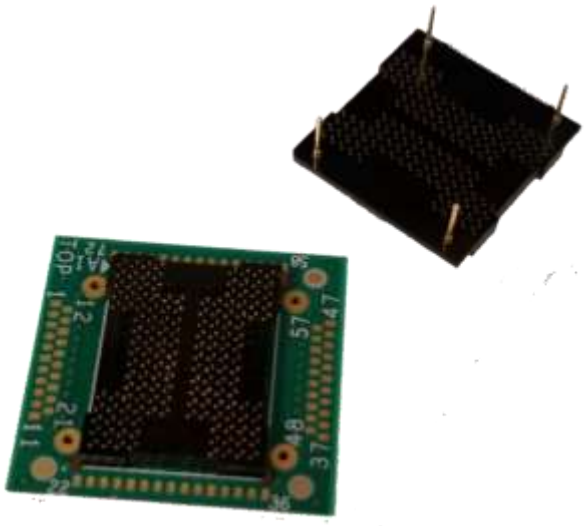


图 7-45

7.6.3.1 Premier Component Interposer Design

Optimal Flash validation requires analysis of the Flash signals as seen by the memory components. This allows for the highest confidence that the signals captured are representative, contain little interference, and present the maximum possible data eye size. Nexus Technology component interposers allow for any oscilloscope to be used for probing Flash signals extremely close to the memory components.

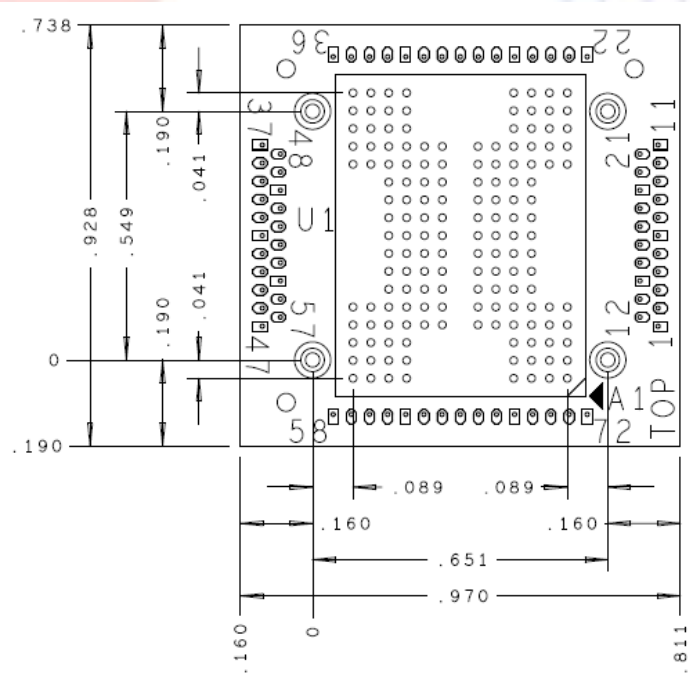


图 7-46 FLASH 152 Ball Oscilloscope Socketed Mechanical drawing

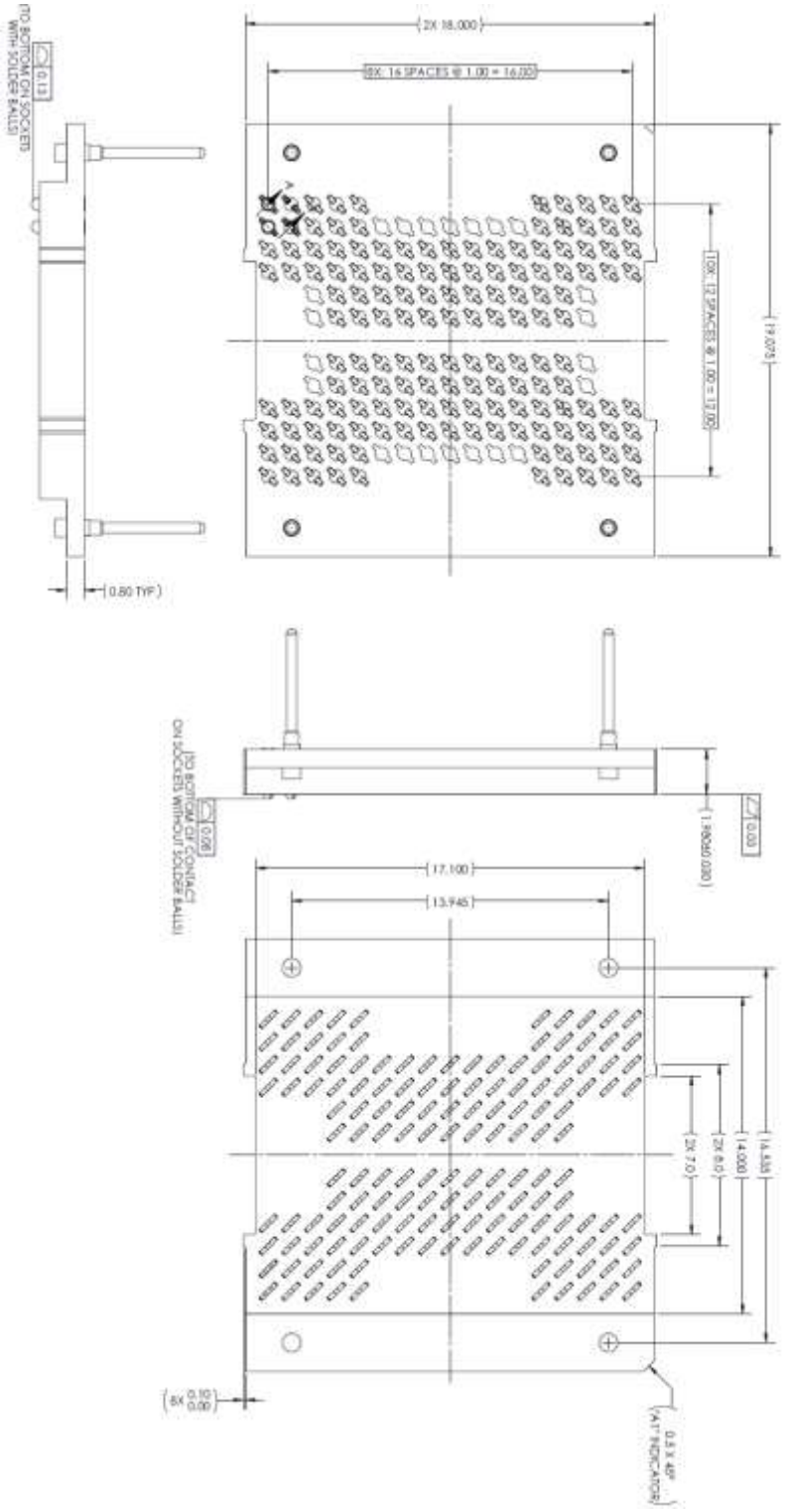


图 7-47 152 Ball Custom Target Socket

7.6.3.2 Product Configuration Table

Nomenclature	Package Size	Component Socket
NEX-NAND152BSC	152 Ball	No
NEX-NAND152BSCSK	152 Ball	yes

7.6.3.3 Available Accessories

Type	Desc.	Quantity	Nomenclature
Riser	Riser elevates interposer .050″	1	NEX-RSRNAND152
Custom Socket	Custom Target Socket	1	NEX-NAND152BGASKBA
Custom Socket	Custom Target Socket	3	NEX-NAND152BGASKBA-3
Solder Balls	Install solder balls on riser	1	NEX-OPT-SOLDERBALLS-RSR
Solder Balls	Install solder balls on standard memory socket	1	NEX-OPT-SOLDERBALLS-SK

7.6.4 NAND Target Socketed Interposer Technology

图 7-48 Target Socket Example

Target Socketed interposers consists of a target socket, an interposer, and – optionally – a component socket. The target socket has a standard ball-grid-array (BGA) interface which is used to install the socket on to the target. The target socket has a removable socket interface to which the interposer can be mechanically attached and re-attached by using a simple tool and your finger to press-fit the parts together.

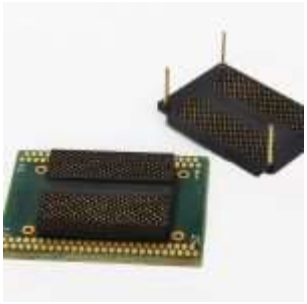


图 7-49 Socketed Interposer w/Target Socket

The target socket's design elevates the interposer up and over adjacent components, allowing this solution to fit in a wide variety of mechanically tight applications.

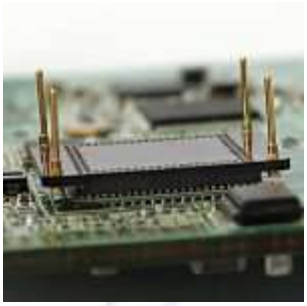


图 7-50 Target Socket Attached to Target

If an optional component socket is provided, this socket will reside on the interposer and provide a mechanically constrained and reusable interface for attaching standard BGA components. If the optional component socket is not provided, the standard BGA component is attached to the interposer using standard BGA component attachment techniques. Lastly, when testing is completed, the interposer can be removed and the BGA component may be press fit directly the custom socket on the target essentially removing the affect of the interposer in the target.

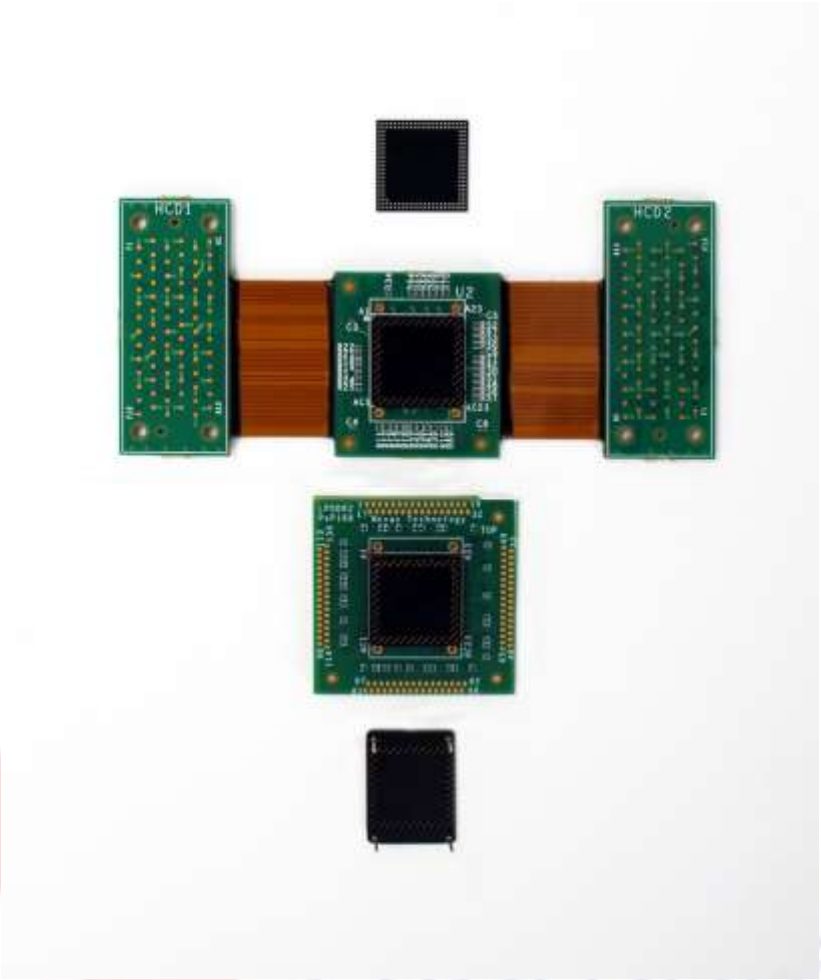


图 7-51 Sockets and Interposers Example

7.7 NAND/SSD HAST 测试母板

我们也提供温箱内使用的针对各种接口的 SSD 以及 NAND 的 HAST 测试母板的定制开发设计服务。

7.7.1 M.2 NVMe SSD HAST 测试母板

下面是针对 M.2 NVMe SSD slot 的 HAST 测试母板的一个用例，具体 form factor 和接口规格可以根据用户定制，包括 AIC, U.2, M.2 等，支持 110 摄氏度。

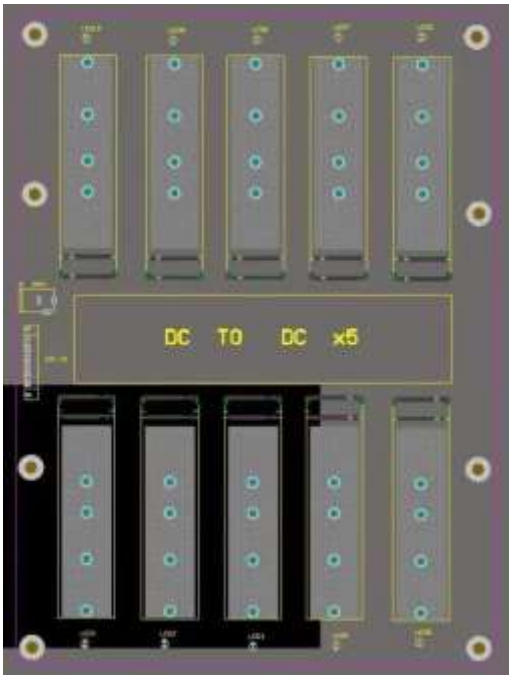


图 7-52

7.7.2 3D NAND 高密度 HAST 测试母板

下图是另外一个例子，对于 3D NAND 进行 HAST 高温测试的母板

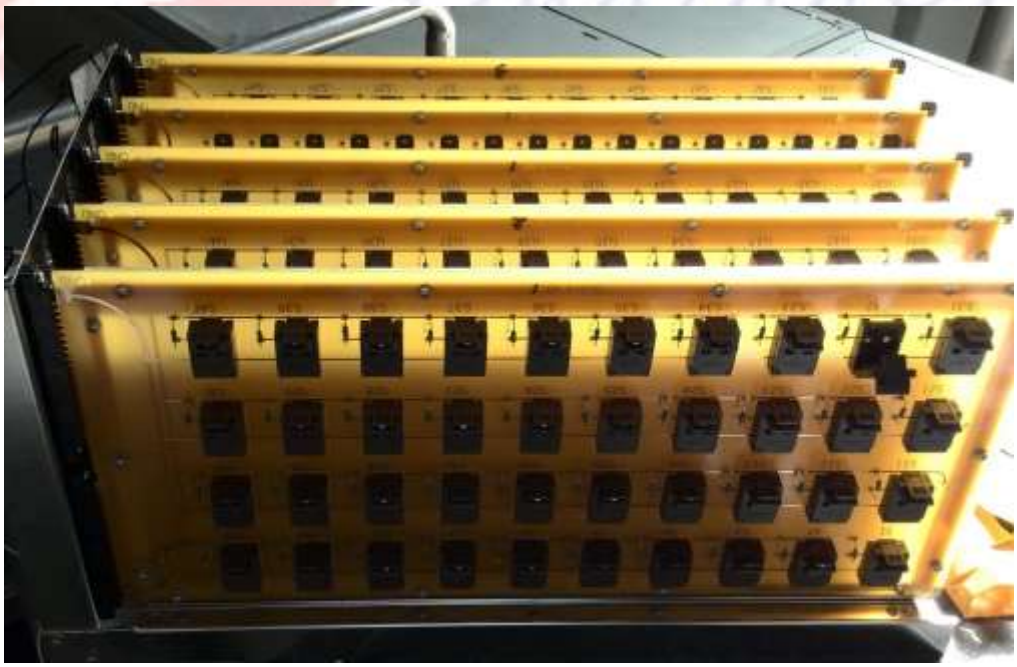


图 7-53

7.8 NAND BGA152/132 clamshell burn-in socket

BGA152/132 clamshell burn in test socket for SSD NAND flash test.

图 7-54

Technical Spec

Chip package	BGA
Pin count	88pin
Pin pitch	1.0mm
Insulation Resistance	1000mΩ Min.At DC500V
Dielectric Witdstanding Voltage For 1Min at AC700V	
Contact Resistance 30mΩ Max.at 10mA and 20mV Max.(initial)	
Current Rating	1A Max.
Operation temperature	-55℃~+175℃
Life Span	25,000 Times (Mechanical)

7.9 DDR5/4, LPDDR5/4 和 eMMC Interposer

7.9.1 DDR5 Interposer

DDR5 Main Memory Interposers

Double-data rate fifth-generation (DDR5) main memory technologies are developed by the Joint Electronic Devices Engineering Council (JEDEC) for use in servers, workstations, and high-performance portable applications that require deep memory.

Nexus Technology offers high quality and high fidelity interposers, enabling the industry to confidently and accurately gain access to DDR5 main memory buses for debug and compliance verification. … more...

Packaging/Modules

DDR5 main memory is available in standard Ball-Grid-Array (BGA) component packages as well as dual-inline-memory-modules of DIMM and SODIMM. Standard BGA packages are soldered directly to the printed circuit board (PCB) while modules comprise a series of packages in a standard PCB format with standard connections between the DIMM and the main board. Interposers are available for both component packages and DIMM and SODIMM modules.

BGA Interposers

		Interposers				Options	
		Oscilloscope			MA51x0 Analyzer		
Package	Balls	EdgeProbe™	Direct Attach	Target Socketed	Target Socketed	Riser	Component Socket
x4/x8	78/82	√(XH)	√(XH)	√(XH)	*	Yes	Yes
x16	102/106	√(XH)	√(XH)	√(XH)	*	Yes	Yes
RCD	240	*	√(XH)	*	*	Yes	No
Data Buffer	55	*	√(XH)	*	*	Yes	No
Custom		Custom designs are also available. Please contact us.					

* If you don't see what you need, please contact us for the most up to date information.

Slot Interposers

Type	Instrumentation	Product
UDIMM	Oscilloscope	DDR5-A-UDM-288
UDIMM	MA51x0	DDR5-D-UDM-288
UDIMM	MA51x0 and Oscilloscope	DDR5-DQ-UDM-288

Product List

- DDR5 240 Ball RCD XH Series Direct Attach Interposer
- DDR5 55 Ball Data Buffer XH Series Direct Attach Interposer
- DDR5 UDIMM Interposer for Oscilloscopes
- DDR5 x16 XH Series Direct Attach Interposer
- DDR5 x16 XH Series EdgeProbe Interposers
- DDR5 x16 XH Series Oscilloscope Target Socketed Interposer
- DDR5 x4/x8 XH Series Direct Attach Interposer
- DDR5 x4/x8 XH Series EdgeProbe Interposers
- DDR5 x4/x8 XH Series Oscilloscope Target Socketed Interposer

图 7-55

7.9.2 DDR4 Interposer

DDR4 Main Memory Interposers

Double-data rate fourth-generation (DDR4) main memory technologies are developed by the Joint Electronic Devices Engineering Council (JEDEC) for use in servers, workstations, and high-performance portable applications that require deep memory.

Nexus Technology offers high quality and high fidelity interposers, enabling the industry to confidently and accurately gain access to DDR4 main memory buses for debug and compliance verification. ... more...

Packaging/Modules

DDR4 main memory is available in standard Ball-Grid-Array (BGA) component packages as well as dual-inline-memory modules of DIMM and SODIMM. Standard BGA packages are soldered directly to the printed circuit board (PCB) while modules comprise a series of packages in a standard PCB format with standard connections between the DIMM and the main board. Interposers are available for both component packages and DIMM and SODIMM modules.

DIMM Interposers

Module Type	1N4168 Analyzer Interposer	MSO Interposer
DIMM	√	√
SODIMM	√	√

Component Interposers

		Interposers				Options	
		Oscilloscope			MA Instrument		
Package	Balls	EdgeProbe™	Direct Attach	Target Socketed	Target Socketed	Riser	Component Socket
x4/x8	78	√ (XH)	√ (XH)	√ (XH)	√	Yes	Yes
x16	96	√ (XH)	√ (XH)	√ (XH)	√	Yes	Yes
x32	144	√	-	-	-	No	Yes
Custom		Custom designs are also available. Please contact us.					

* If you don't see what you need, please contact us for the most up to date information.
(XH) XH Series

XH Series

Nexus Technology's XH Series interposers bring new enhancements to the EdgeProbe™ and Direct Attach types of memory component interposers that maintain signal integrity across the interposer path as well as provide for extremely high-fidelity oscilloscope probe points for both leading edge and emerging memory technologies.

Learn more...

Attachment Service

Nexus Technology's expert attachment service provides a ready-to-go test solution customized to your application. We will attach the interposer and any additional accessories to your application's target. We can also power-on and test your application to confirm functionality.

Attachment Service

More Information on Interposer Types

Electrical Analysis

Electrical analysis is enabled by using either an EdgeProbe, High Density, or Socketed interposer to capture memory activity on an oscilloscope. The oscilloscope is then used to debug, analyze, and verify the analog characteristics of your design. Presenting an accurate representation of the signals under test to the oscilloscope is critical. Nexus interposers provide an unobtrusive interconnect and accurate signal to your oscilloscope. Learn more...

Logic / Compliance Analysis

Logic analysis is performed using a logic / compliance interposer to capture memory activity on a logic or memory analyzer. The logic or memory analyzer is then used to debug, analyze, and verify the logic (basic protocol) of your design. Compliance analysis uses the same interposers to capture activity on a memory analyzer. The memory analyzer is then used to debug, analyze, margin test, performance analyze, and verify the memory protocol. Learn more...

Product List

- DDR4 144 Ball EdgeProbe™ Interposers
- DDR4 78 Ball Logic/Compliance Interposer
- DDR4 78 Ball XH Series EdgeProbe Interposers
- DDR4 78 Ball XH Series Oscilloscope Direct Attach Interposer
- DDR4 78 Ball XH Series Oscilloscope Target Socketed Interposer
- DDR4 96 Ball Logic/Compliance Interposer
- DDR4 96 Ball XH Series EdgeProbe Interposers
- DDR4 96 Ball XH Series Oscilloscope Direct Attach Interposer
- DDR4 96 Ball XH Series Oscilloscope Target Socketed Interposer
- DDR4 DIMM Compliance Interposer
- DDR4 DIMM Logic Compliance Interposer
- DDR4 DIMM Mixed Signal Oscilloscope Interposer
- DDR4 SODIMM Compliance Interposer
- DDR4 SODIMM Logic Compliance Interposer
- DDR4 SODIMM Mixed Signal Oscilloscope Interposer

图 7-56

7.9.3 LPDDR5 interposer

LPDDR5 Mobile Memory Interposers

Low power double-data rate fifth-generation (LPDDR5 and LPDDR5X) mobile memory technologies are developed by the Joint Electronic Devices Engineering Council (JEDEC) for use in handheld devices or applications where low power and small size is critical. Nexus Technology offers high quality and high fidelity interposers, enabling the industry to confidently and accurately gain access to LPDDR5(X) mobile memory buses for debug and compliance verification. ... more...

Packaging/Modules

LPDDR5 and LPDDR5X mobile memory is available in standard Ball Grid Array (BGA) component packages. Standard BGA packages are soldered directly to the printed circuit board (PCB). Interposers are available for component packages detailed below.

	Interposers			Options	
	Oscilloscope		MA Instrument		
Package (Ch×Dh)	EdgeProbe™	Direct Attach	Logic Compliance	Riser	Component Socket
297 Ball	*	√ 300	*	Yes	No
315 Ball (0×14)	√ 180	√ 344	√	Yes	No
416 Ball	*	*	*	Yes	No
441 Ball	*	√ 344	*	Yes	No
496 Ball (4×10)	*	√ 344	√	Yes	No
Custom	Custom designs are also available. Please contact us.				

* If you don't see what you need, please contact us for the most up to date information.

XH Series

Nexus Technology's XH Series interposers bring new enhancements to the EdgeProbe™ and Direct Attach types of memory component interposers that maintain signal integrity across the interposer path as well as provide for extremely high-fidelity oscilloscope probe points for both leading edge and emerging memory technologies. Learn more...

Attachment Service

Nexus Technology's expert attachment service provides a ready-to-go test solution customized to your application. We will attach the interposer and any additional accessories to your application's target. We can also power-on and test your application to confirm functionality.

Attachment Service

More Information on Interposer Types

Electrical Analysis

Electrical analysis is enabled by using either an EdgeProbe, High Density, or Socketed interposer to capture memory activity on an oscilloscope. The oscilloscope is then used to debug, analyze, and verify the analog characteristics of your design. Presenting an accurate representation of the signals under test to the oscilloscope is critical. Nexus interposers provide an unobtrusive interconnect and accurate signal to your oscilloscope. Learn more...

Logic / Compliance Analysis

Logic analysis is performed using a logic / compliance interposer to capture memory activity on a logic or memory analyzer. The logic or memory analyzer is then used to debug, analyze, and verify the logic (basic protocol) of your design. Compliance analysis uses the same interposers to capture activity on a memory analyzer. The memory analyzer is then used to debug, analyze, margin test, performance analyze, and verify the memory protocol. Learn more...

The following analyzers currently support LPDDR5:

Memory Analyzers Supporting LPDDR5

Product	Details
 MA5100 Series Memory Analyzer	Memory analyzer supporting DDR5, LPDDR5(X), LPDDR4(X) and LPDDR3 performance, margins and capture up to LPDDR5-6400, LPDDR4-4267 and LPDDR3-2133 with 1G-Sample acquisition depth, ClockSafe™ and Single Smart Frequency or Sixteen Smart Frequency Analysis.

Product List

- LP5-A CDA 315 Interposer
- LPDDR5 297 Ball XH Series Direct Attach Interposer
- LPDDR5 315 Ball Compliance Interposer
- LPDDR5 315 Ball XH Series EdgeProbe Interposers
- LPDDR5 441 Ball Direct Attach Interposer
- LPDDR5 496 Ball Compliance Interposer
- LPDDR5 496 Ball XH Series Direct Attach Interposer

图 7-57

7.9.4 LPDDR4 interposer

LPDDR4 Mobile Memory Interposers

Low power double data rate fourth generation (LPDDR4 and LPDDR4X) mobile memory technologies are developed by the Joint Electronic Devices Engineering Council (JEDEC) for use in handheld devices or applications where low power and small size is critical.

Nexus Technology offers high quality and high fidelity interposers, enabling the industry to confidently and accurately gain access to LPDDR4(X) mobile memory buses for debug and compliance verification. ... more...

Packaging/Modules

LPDDR4 and LPDDR4X mobile memory is available in standard Ball-Grid Array (BGA) component packages. Standard BGA packages are soldered directly to the printed circuit board (PCB). Interposers are available for component packages detailed below.

	Interposers			Options	
	Oscilloscope		MA Instrument		
Package (Ch x Dh)	EdgeProbe™	Direct Attach	Logic Compliance	Riser	Component Socket
14 Ball (2 x 5)	*	*	*	No	No
149 Ball (2+19)	*	✓ (XH)	✓	Yes	No
200 Ball (2+18)	✓ (XH)	✓ (XH)	✓	Yes	Yes
254 Ball (2+16)	*	✓ (XH)	✓	Yes	Yes
272 Ball (4+18)	✓	✓	✓	Yes	Yes
366 Ball (4+16)	*	✓	✓	Yes	Yes
376 Ball (4+16)	*	*	*	No	No
432 Ball (4+18)	*	✓ (XH)	✓	Yes	No
556 Ball	Contact Nexus	Contact Nexus	Contact Nexus	Yes	No
Custom	Custom designs are also available. Please contact us.				

* If you don't see what you need, please contact us for the most up to date information.
(XH) XH Series

XH Series

Nexus Technology's XH Series interposers bring new enhancements to the EdgeProbe™ and Direct Attach types of memory component interposers that maintain signal integrity across the interposer path as well as provide for extremely high-fidelity oscilloscope probe points for both leading edge and emerging memory technologies.

Learn more...

Attachment Service

Nexus Technology's expert attachment service provides a ready-to-go test solution customized to your application. We will attach the interposer and any additional accessories to your application's target. We can also power on and test your application to confirm functionality.

Attachment Service

More Information on Interposer Types

Electrical Analysis

Electrical analysis is enabled by using either an EdgeProbe, High Density, or Socketed interposer to capture memory activity on an oscilloscope. The oscilloscope is then used to debug, analyze, and verify the analog characteristics of your design. Presenting an accurate representation of the signals under test to the oscilloscope is critical. Nexus interposers provide an unobtrusive interconnect and accurate signal to your oscilloscope. Learn more...

Logic / Compliance Analysis

Logic analysis is performed using a logic / compliance interposer to capture memory activity on a logic or memory analyzer. The logic or memory analyzer is then used to debug, analyze, and verify the logic (basic protocol) of your design. Compliance analysis uses the same interposers to capture activity on a memory analyzer. The memory analyzer is then used to debug, analyze, margin test, performance analyze, and verify the memory protocol. Learn more...

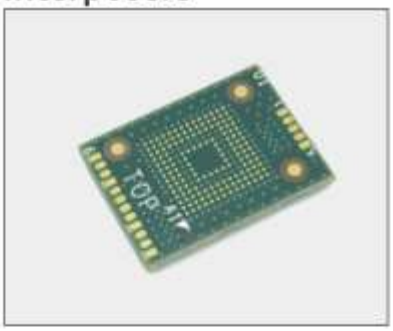
Product List

- LPDDR4 149 Ball Compliance Interposer
- LPDDR4 200 Ball Compliance Interposer
- LPDDR4 200 Ball XH Series Direct Attach Interposer
- LPDDR4 200 Ball XH Series EdgeProbe Interposers
- LPDDR4 254 Ball Compliance Interposer
- LPDDR4 254 Ball XH Series Direct Attach Interposer
- LPDDR4 272 Ball Compliance Interposer
- LPDDR4 272 Ball Direct Attach Interposer
- LPDDR4 272 Ball EdgeProbe™ Interposer
- LPDDR4 366 Ball Compliance Interposer
- LPDDR4 366 Ball Direct Attach Interposer
- LPDDR4 432 Ball Compliance Interposer
- LPDDR4 149 Ball XH Series Direct Attach Interposer
- LPDDR4 432 Ball XH Series Direct Attach Interposer

图 7-58

7.9.5 eMMC 153 Ball Direct Attach Oscilloscope Interposers

eMMC 153 Ball Direct Attach Oscilloscope Interposers



Direct Attach Interposers

Nexus Technology recommends direct attach flash component interposers as a low cost alternate to the target socketed component interposers for oscilloscope applications. Read more...

Simulation and De-embedding

Oscilloscope de-embedding software filters/removes interposer effects. Please contact Nexus for the appropriate de-embedding software for your interposer.

Product Configuration Table

Nomenclature	Interposer Type	Quantity Included
NEX-NAND153SCDS	Direct Attach	1

Available Accessories

Type	Desc.	Quantity	Nomenclature
Riser	Riser elevates interposer .050"	1	NEX-RSRNAND153

Attachment Service

Nexus Technology's expert attachment service provides a ready-to-go test solution customized to your application. We will attach the interposer and any additional accessories to your application's target. We can also power-on and test your application to confirm functionality. Please contact us for more information.

Attachment Service

Mechanical Outline

EMMC 153 Ball Oscilloscope Direct Attach
NEX-NAND153SCDS

EMMC 153 Ball Riser
NEX-RSRNAND153

图 7-59

7.10 高速接口批量测试高密度测试板

针对各种高速接口，我们有合作伙伴可以设计如下的各类测试板：

- **DRAM Application DDR2, DDR3, DDR4, DDR5, LPDDR2, LPDDR3, LPDDR4, LPDDR4X, LPDDR5, GDDR4, GDDR5, GDDR6**

- **Flash Application NAND Flash(Legacy) EX-PBA NAND, MCP HS NAND, Toggle NAND Micro SD, eMMC, eMCP SSD, S3E, UFS, S4E, S5E**

图 7-60

7.10.1 POGO SOCKET

图 7-61

7.10.1.1 What is Pogo Socket?

Pogo Pin sockets are a key element in the development and testing of semiconductor products. Different types of sockets provide specific electrical and mechanical parameters for specific steps in the product testing process. In high-performance sockets, Pogo Pin socket is still 90% of the application's contact. *Although often used as a generic name, pogo pin is a registered trademark of* ***Everett Charles Technologies*** *(ECT).*

A pogo pin or spring-loaded pin is a type of electrical connector mechanism that is used in many modern electronic applications and in the electronics testing industry.[1] They are used for their improved durability over other electrical contacts, and the resilience of their electrical connection to mechanical shock and vibration.[2]

The name pogo pin comes from the pin's resemblance to a pogo stick – the integrated helical spring in the pin applies a constant normal force against the back of the mating receptacle or contact plate, counteracting any unwanted movement which might otherwise cause an intermittent connection. This helical spring makes pogo pins unique, since most other types of pin mechanisms use a cantilever spring or expansion sleeve.

How does a Pogo Pin work? [Structure, Materials & Advantages] | C.C.P. Contact Probes (ccpcontactprobes.com)
https://www.ccpcontactprobes.com/introduction-pogo-pin

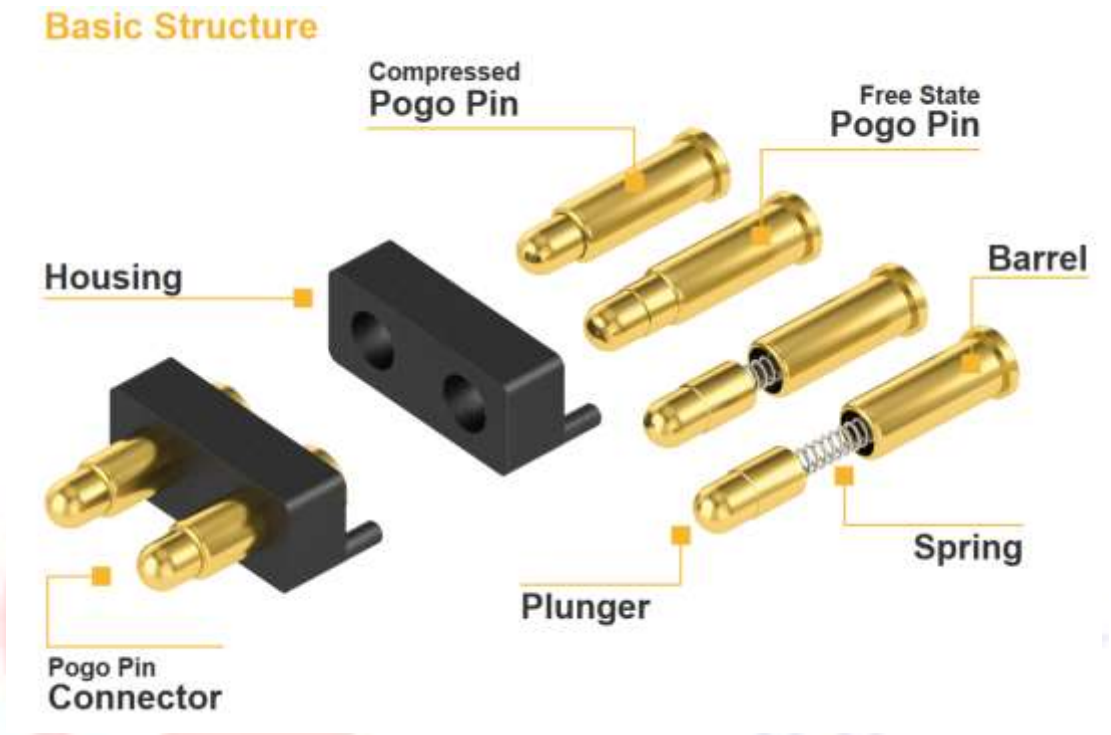


图 7-62

图 7-63

Let’s Discover More About Pogo Pin And Pogo Pin Socket(10)

TTS Group - Let’s Discover More About Pogo Pin And Pogo Pin Socket(10) (tts-grp.com)

In this article, we will be discussing the topic of pogo pins and pogo heads. A pogo pin or sometimes called spring-loaded pin is a variety of electric connector that can be found in most of the modern electronic devices and in the electronics testing industry. Due to the pin’s resemblance to a pogo stick, spring-loaded pin is also called pogo pin for this very reason. In most types of pin mechanics, they use a cantilever spring or expansion sleeve unlike pogo pins which utilise a helical spring. Material vise, copper alloys are usually the most common choice to make the springs. As for the barrel and plunger of pogo pins, brass and copper are usually selected as base material and then applied a thin layer of nickel on it. A gold plating is also applied to improve the contact resistivity.

Pogo pins are getting more and more popular in the electronics testing industry nowadays and it is not without reasons. Pogo pins have been created for roughly 5 decades and its functionality has been improving year after year. The first reason for why we are preferring pogo pins in this line of work is the low cost of pogo pins. In comparison, pogo pins are more efficient than the other connectors and the production costs are much cheaper. Almost all connectors face the same problem. They are either very stable but big in size or they are very fragile but small in size. Pogo pins on the other hand have both of two worlds. They are extremely small but their functionality doesn't decrease. Another advantage that the pogo pins have over the other pins is that pogo pins are the most durable among them. Pogo pins are fairly easy to assemble. Most engineers have to worry about the assembly when designing a device. Using pogo pins, they are able to quickly assemble the parts while having much more flexibility compared to using other connectors.

One of the companies that sells pogo pins is TTS Group. One of their products sold is pogo pins aka probe pins. The list of products sold can be categorised into three categories which are pogo pins, test sockets and hand socket lid. There is a wide selection of pogo pins to choose from. In the pogo pin category, there is a Kelvin Probe, Wiping Probe, WLCSP Probe and of course, the High Power Probe. Each of the probes has their own dimensions and their own functions.

The next item on the list is the pogo pin socket. It is confirmed that one of the key elements in developing and testing a semiconductor product is the pogo pin sockets. Everyone plays a different role in this world and so do sockets. Each type of socket provides different parameters in different product testing processes. Pogo pin sockets are most commonly used but if the parameters are too extreme, engineers will have to use another type of

connectors to solve this issue.

There are several advantages as to why we use pogo pin sockets. Just like their counterpart, the pogo pin, pogo pin sockets also have high durability. One of the most important things in starting a business is to know what the customer needs. Pogo pin sockets did a very good job in this case as it can be easily adapted to customer needs. Next, pogo pin sockets' performance are very excellent in their jobs and they are specially suitable for High-Frequency applications. The pogo pin sockets are also small in size which in turn are able to achieve space-saving compared to plugs in their line of work. TTS Group also sells pogo pin sockets and the dimensions of those sockets are listed in their website.

7.10.1.2 高速 Pogo Socket 举例

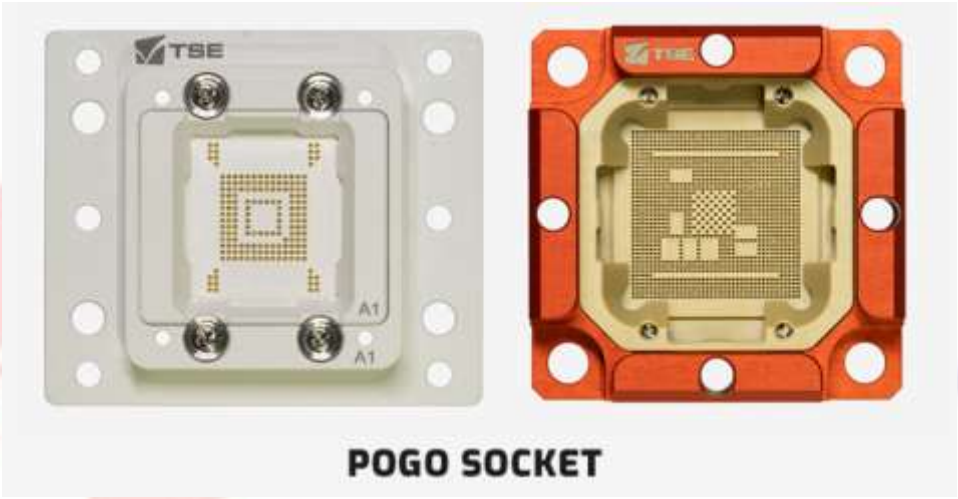


图 7-64

7.10.2 TEST SOCKET

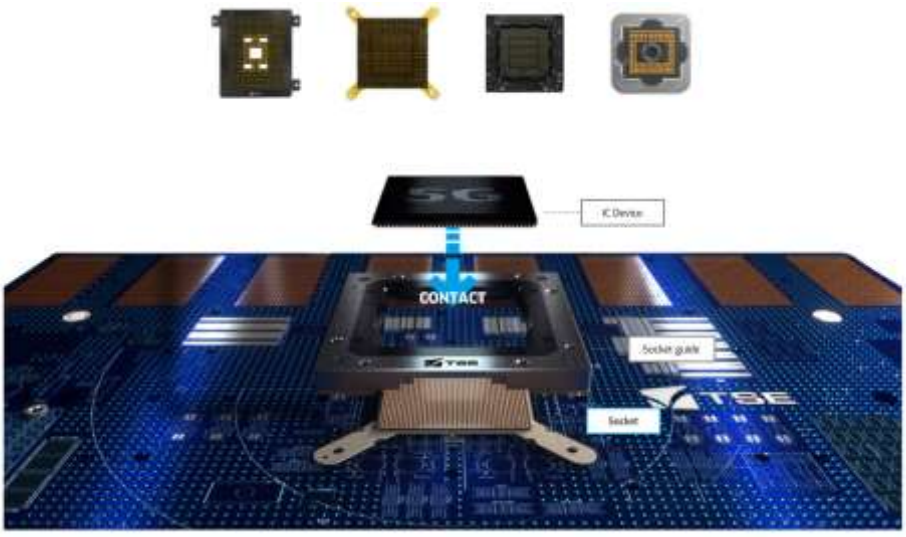


图 7-65

7.10.3 INTERFACE BOARD

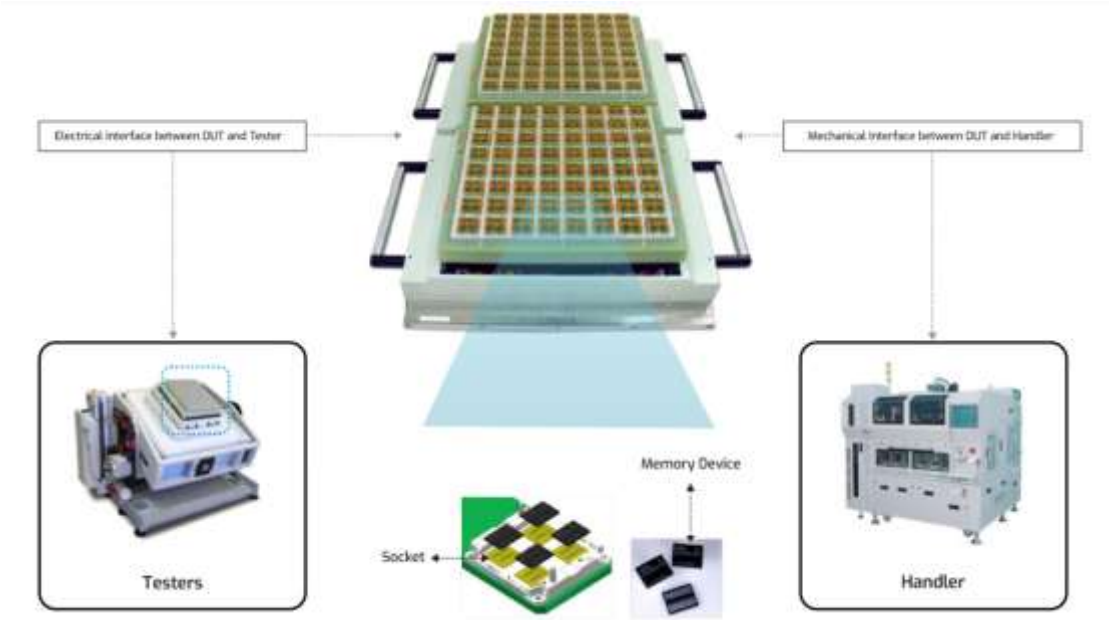


图 7-66

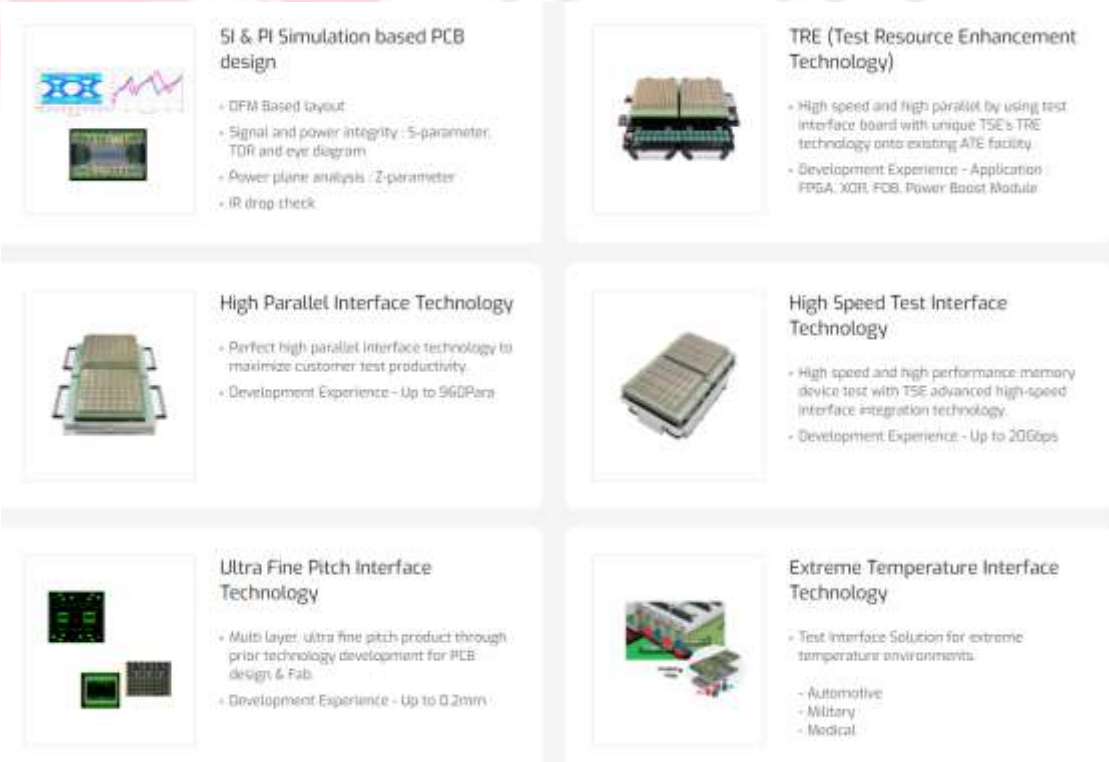


图 7-67

7.10.4 LOAD BOARD

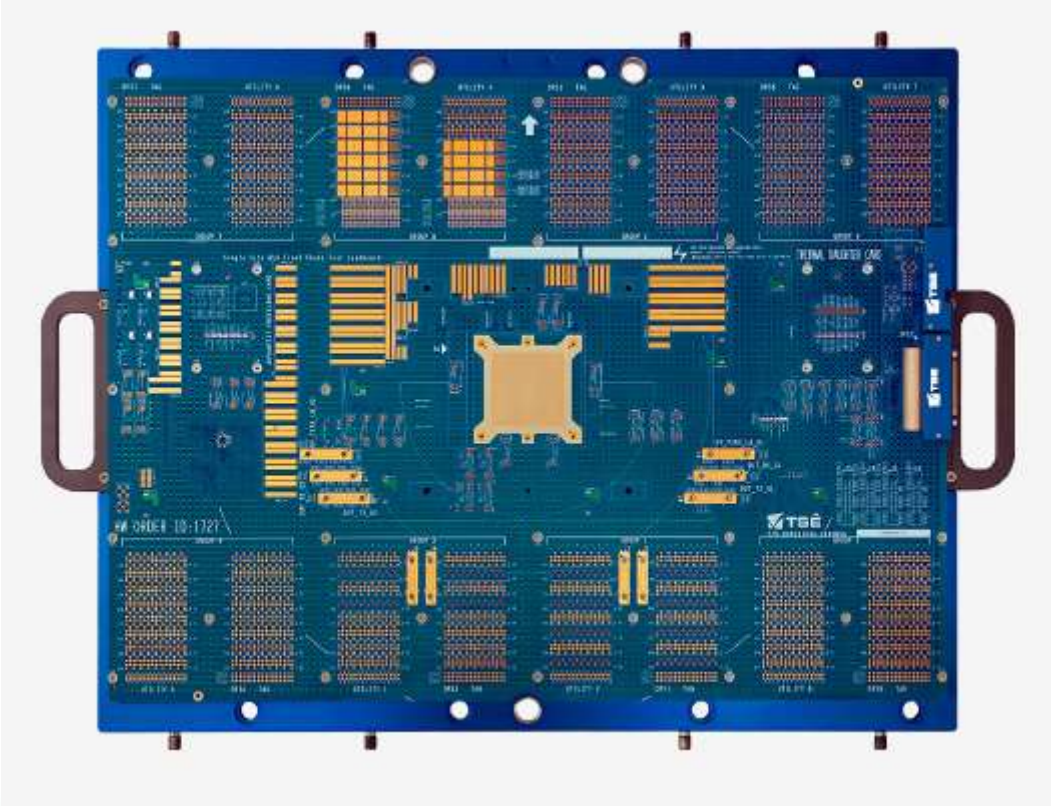


图 7-68

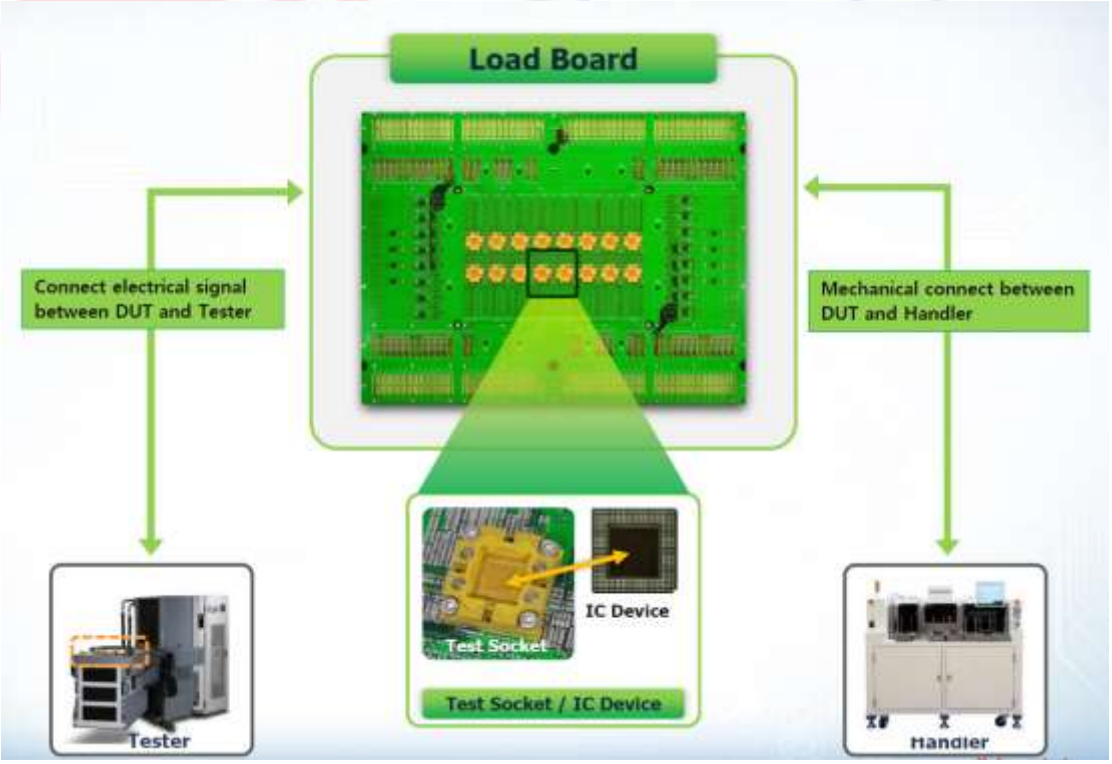


图 7-69

7.11 Zero Footprint Sockets

*** Ironwood Electronics Zero footprint (ZFP)*

Zero Footprint Socket - Memory Test Solutions

11/01/2019 | Ironwood Products | Created by Dagmar Oertel

***HSIO Technologies and Ironwood Electronics** announce that development, production, service and sale of the HSIO test socket has been aquired by Ironwood. Ironwood is extending its already wide range of sockets and adapaters with another important line. The well-known and proven Grypper socket is now available at EMC.*

Ironwood Electronics is already working on the next generation socket to meet the growing demand for signal integrity and interconnect density.

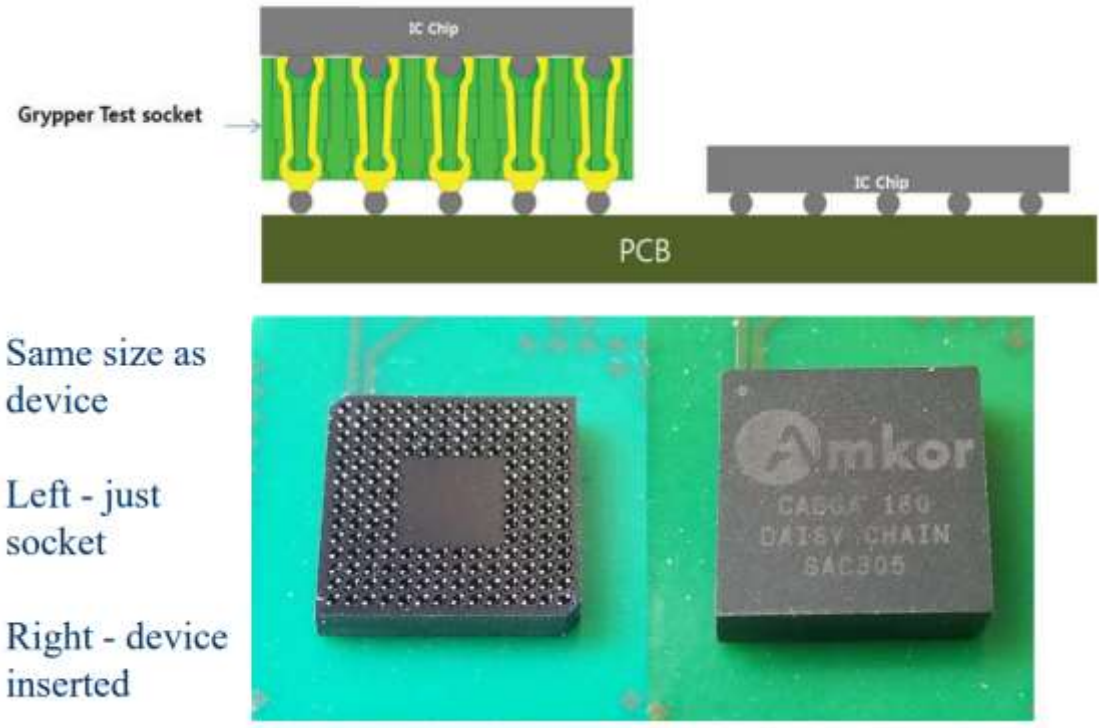


图 7-70

Standard catalog items available for:

- **eMMC**
- **LPDDR**
- **DDR**
- **ONFI(NAND)**

下面是一个 NVMe SSD 应用。

图 7-71

下面是一个手机应用。

Grypper socket mounted on iPhone

图 7-72

图 7-73

Grypper

- Zero Footprint Socket
- Same size as device
- No lid required
- Device is simply pressed in
- Solderable on original hardware

more...

Giga-snaP™ BGA Adapter

- Pitch 0.5mm to 1.27mm
- Female with balls
- Male with balls
- IC carrier
- Extensions
- RoHS and non RoHS
- For all Footprints

more...

Test & Burn In Socket for BGA & QFN

- Pitch 0.2mm to 1.27mm
- Optimal signal integrity
- Up to 20 GHz
- More than 300.000 test cycles
- Quicklock
- -55°C ... +180°C

more...

Package Converter

Fast Solution for

- fixing of layout errors
- obsolete parts
- when having soldering problems
- BGA to QFP
- QFP to PGA
- QFP to QFP

more...

Sockets for 94 GHz and above

- Pitch 0.2mm to 1.27mm
- Above 94 GHz @ -1dB
- Contact resistance 25 mΩ
- Temperature range -55°C to +180°C
- More than 200.000 test cycles

more...

GHz Elastomer Socket for BGA & QFN

- Pitch 0.25mm to 1.27mm
- Optimal signal integrity
- Up to 30 GHz
- More than 2000 test cycles

EMC

DE EN Search

Home Company Products News & Events Contact Downloads E-tec Group Distribution

- Heatsinks and Fans
- Quicklocks
- Sensor options
- SMT Adapter
- THT Adapter

more...

Standardized Socket

- Pitch 0.35mm to 1.27mm
- Spring pin contacts
- Standardized components
- Fast availability

more...

Socket Technology Overview

- Spring pins
- Elastomer variants

more...

Catalog

- Socket Catalog
- Grypper Catalog

We will be happy to send you a copy.

图 7-74

7.12 DDR5/LPDDR5 协议分析仪

图 7-75

7.12.1 概述

JKI JLA420A 协议分析仪系统是业界最高性能的 DDR5/LPDDR5 协议分析仪，具有高达 512GB 的深度采集存储深度和高达 32GHz 的最高采样率。其强大的高速内存跟踪解决方案使工程师能够调试和验证基于 DDR 的内存系统，速度高达 6.4 Gb/s。

该 DDR5 协议分析仪也可以配置 RDIMM DDR4 interposer 分析 3200M DDR4 内存条。

7.12.1.1 用于数据采集的大内存

高达 512GB 的大内存深度允许您调试非常复杂的需要捕获时间长的问题。

7.12.1.2 超快的采样速度

高达 32 GHz 的超快采样速度使您无需任何额外选项即可为所有通道捕获高达 6.4 Gb/s 的数据速率。

7.12.1.3 探头解决方案

为了最大限度地减少探头系统中的信号衰减，模拟模块与数字模块分开并放置在尽可能靠近设备的位置。

7.12.1.4 Interposer 解决方案

我们提供各种 Interposer 和连接解决方案：LPDDR4/5 interposer、DDR4/5 DIMM interposer

图 7-76

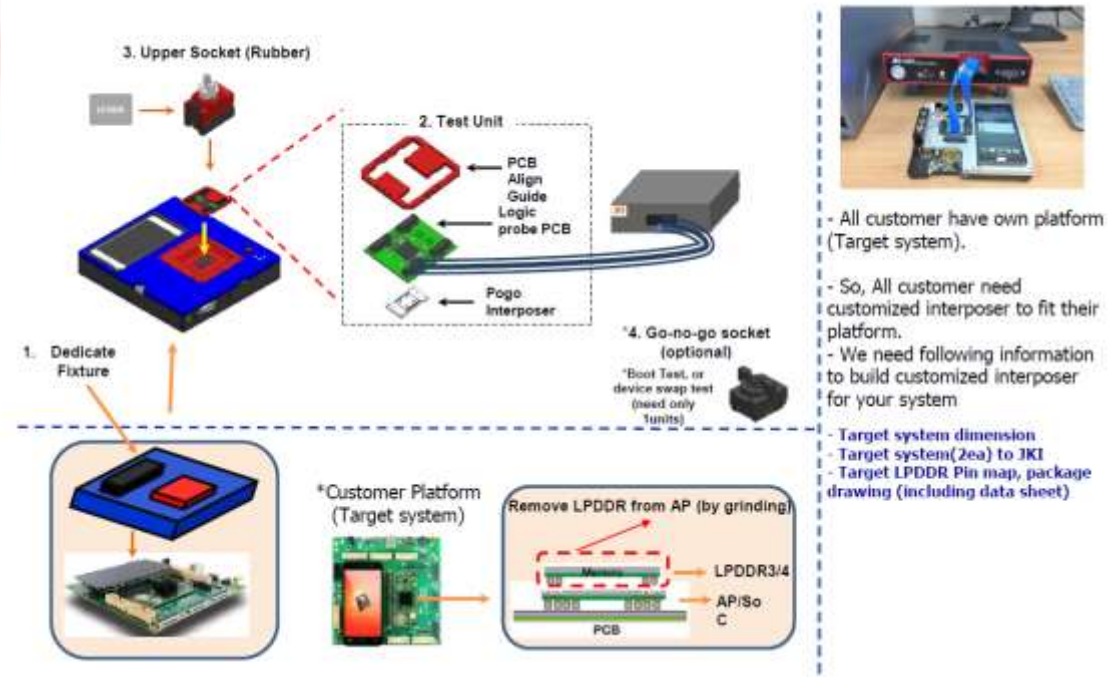


图 7-77

7.12.2 产品基本功能

- **最大样本 速度**
  - 最大限度。 32 千兆赫
  - 2/4/8/16/32 GHz 选项

- **通道数：总计 34 CH**
- **单端通道（ 31 中文)**
  - 每针端接 等级
  - 每引脚阈值 等级
  - 每引脚时序 延迟
- **差分通道（ 3 中文)**
  - 1 个差分时钟输入 (CK)
  - 2 差分频闪 输入(WCK/RDQS)
- **分析 模式**
  - timing 模式
  - eye diagram 模式
- **Capture buffer 大小**
  - 最大限度 512 GB（标准 128 国标）
  - 全内存深度高达 32 千兆赫
  - 硬件数据 压缩
- **支持所有通道 channel 的 Hysteresis 设置**
- **主机接口：PCIe Gen3 x8**
- **应用场景**
  - DDR4/5 协议 分析
  - LPDDR4/5 协议 分析

7.12.3 软件 GUI 功能

- **捕获， 导航, 和 分析 内存 记忆 协议 通过 快速地 TeraView 2.0 图形用户界面。**
- **简单的设置和配置 窗户。**
- **基于软件的定时模式眼图 图表。**
- **可编程的 包 解码器 和 API 函数方便 用户 事后自己二次开发。**

7.12.4 新的 TeraView 2.0 Wave & List View 视图

图 7-78

7.12.5 时序模式眼图

图 7-79

7.12.6 产品技术指标一览

Channel Specification	
Number of Channels	34 (31 single-ended + 3 differential)
Input Voltage Range	-0.2V to 1.2V (analyzer input)
Minimum Input Voltage Swing	40 mV (analyzer input)
Termination (Vtt) Voltage Range	0V to 1.2V (analyzer input)
Termination (Vtt) Voltage Resolution	1 mV
Threshold (Vref) Voltage Range	0V to 1.2V (analyzer input)
Threshold (Vref) Voltage Resolution	1 mV
Hysteresis (Vhys) Voltage Range	0 – 50 mV
Timing Delay Range	-2ns to 2ns
Timing Delay Resolution	31.25 ps @32 GHz
Input Impedance	50 Ohm (analyzer input) 250 Ohm (5:1 probe) 500 Ohm (10:1 probe)

Acquisition Specification	
Sample Rate	2.0 / 4.0 / 8 / 16 / 32 GHz
Analysis Mode	Timing mode, Filter mode, Eye mode
Acquisition Memory	Max. 512 GB (standard 128 GB)
Host Interface	PCIe Gen3 x8
Data Compression	Hardware compression
Data De-compression	Software de-compression

Trigger Specification	
Trigger Pattern	E(Either edge), R(Rising edge), F(Falling edge), H(High), L(Low), X(Don't care) for all channel
Trigger Sequence	2 ~ 4 burst trigger
Trigger Action	Start capture
Trigger Rate	TBD
Pre-trigger Size	TBD

Environmental and physical	
Power Consumption	Max. 150W
Operation Temperature (nom)	0 to +40 dec C
Humidity (nom)	0 to 80% relative humidity
Dimensions (W x H x D)	347 mm x 290 mm x 74 mm
Weight	4.7kg

图 7-80

7.12.7 DDR5 的推荐系统配置

模型	数量	描述
JLA420A	1个	JLA420A DDR5/LPDDR5 协议分析仪（128GB-标配）
P105Z0	1个	DDR5 DIMM Interposer
SW-DD5-BD	1个	软件 – DDR5 总线解码器
PCIe-HIB38	1个	PCIe 主机适配器 (Gen3 x8)
PCIe-0802	1个	PCIe 电缆 2m (Gen3 x8)

7.12.8 LPDDR5 的推荐系统配置

模型	数量	描述
JLA420A	1个	JLA420A DDR5/LPDDR5 协议分析仪（128GB-标配）
P1034B	1个	Samtec 探头电缆
SW-LP5-BD	1个	软件 – LPDDR5 总线解码器
PCIe-HIB38	1个	PCIe 主机适配器 (Gen3 x8)
PCIe-0802	1个	PCIe 电缆 2m (Gen3 x8)

7.12.9 产品信息

模型	描述
JLA420A	JLA420A DDR5/LPDDR5 协议分析仪（128GB-标配）
- 选项 256G	256 GB 内存选项
- 选项 512G	512 GB 内存选项
P1034B	Samtec 探头电缆
P2134A-08S	无连接器探头电缆
P104Z0	DDR4 DIMM Interposer
P105Z0	DDR5 DIMM Interposer
PCIe-HIB38	PCIe 主机适配器 (Gen3 x8)
PCIe-0802	PCIe 电缆 2m (Gen3 x8)
SW-LP5-BD	软件 – LPDDR5 总线解码器
SW-DD5-BD	软件 – DDR5 总线解码器
CAL-1032A	校准套件
W200R-01	1年软硬件技术服务计划
W200R-03	3 年保修和校准服务计划（返回 JKI）

7.13 DDR5 测试设备

7.13.1 DDR5 RDIMM 研发测试平台产品规格

DDR5 RDIMM Test System Technical Specifications	
Configuration	4 / 8 DUT per unit
Operation Modes	Manual loading / unloading. Parallel testing
Control Software	Control software installed on external Control PC with Windows 10 Professional
Number of Test Sock- ets	Up to 8x 288-pin sockets per chassis
Memory Type Sup- ported	DDR5 SDRAM 8/16/24/32 Gbit. x4/x8 density
Memory Standards Supported	DDR5 ECC RDIMM / LRDIMM / 3DS DDR3200 / 3600 / 4000 / 4400 / 4800 RDIMM / LRDIMM 8 rank max. 3DS DIMM 16 rank max.
Test Frequency	Clock 1600 / 1800 / 2000 / 2200 / 2400 MHz
I/O Interface	DDR5 CK/CA/DQ/DQS/DM POD 1.1V DDR5 RESETn CMOS 1.1V

	SPD CMOS Push-pull 1.0V PMC PWR_GOOD POD 1.1V	
Temperature Control Chamber	Optional heat chamber with control profile up to 85°C Optional cold air system with control profile down to 10°C	
Test Site[1]		
Address Generation	18X + 11Y + 2BA + 3BG + 3CID	
# of Rank Supported	RDIMM / LRDIMM 8-rank, 3DS 16-rank	
Data lines	2-set 40bits	
Control Lines	2-set CA/CSn/PAR set, 1-set ALERTn/RESETn	
DUT Clocks	1 pair differential	
Device Access Param- eters	tRCD, tCL, tWL, tRAS, tRC, tRP, tCSSR, tCWL, tWR, tWRPRE, tWRA, tRDA, tRTP tRW, tWW, tRR, tFAW, CmdRate Adjustable in nCLKs	
Refresh Control	Per row Refresh time adjustable from 1us up to 32768 memory clock periods	
DUT Power	VIN_BULK 12V nominal; adjustable +/-25%, 10mV step, 3A max VIN_MGMT 3.3V nominal; adjustable +/-25%, 10mV step, 1A max	
Current Measurement	VIN_BULK 0~3A, resolution 0.1mA, accuracy +/-(2% + 0.1mA) VIN_MGMT 0~1A, resolution 0.1mA, accuracy +/-(2% + 0.1mA) * DUT power rails have over-current protection	
Tests Available	Functional	Row Hammer, MarchA, MarchX, MarchC, MarchG, MATS, and many more industry standard tests.
	Application Others	ATSLWB, JumpLWB, Shift, MoveRot, MT64 Rank Margin test, SPD test, IDDx
SPD Operations	SPD Read, Write, Serialization, Import, Export, Address Check, Check Sum, Auto Setup , Program, Compare, Program Exclude, Compare Exclude, Hex View, Symbolic View	
I2C/I3C Operations	I2C/I3C access to DUT Hub and downstream devices; such as PMIC, RCD, TS	
Error Logging	Individual 80 DQ failures indicator Record last fail physical address (18X / 11Y / 2BA / 3BG) per rank	
Other Features	Shmoo and Burn-in tools Comprehensive test flow control and binning control Speed Grading and Sorting support Test and Results Data logging capability Address Register Read Control (RDIMM and LRDIMM)	
Optional Features	PPR (Post Package Repair) Support Device Core parameter Profiler DQ Eye Diagram	
Test System Environmental		

Tester Chassis	670mm x 360mm x 190mm (D x W x H) On-chassis front panel LCD for displaying tester status & DUT pass/fail. Optional Heater/Cooler user control if equipped with Temperature Chamber
Temperature Chamber Chassis Dimension	560 x 360 x 300mm (D x W x H) approx.
Connection to Control PC	100/1000Base-T Ethernet with CAT5 cable
Chassis cooling	Forced air cooling
Tester Station Weight	18Kg approx.
Temperature Chamber Weight	22Kg approx.
AC Power Source	1200W 100~127V@12A ,60/50Hz 200~240V@7.5A ,60/50Hz
Operating Temperature	10 ºC– 30ºC ambient
Operating Humidity	20% - 80 % (non-condensing)
Storage Temperature	5 - 40ºC
Storage Humidity	35% – 70% (non-condensing)

Notes

1.）Each 'Test Site' corresponds to one DIMM test socket

图 7-81 Chassis without chamber

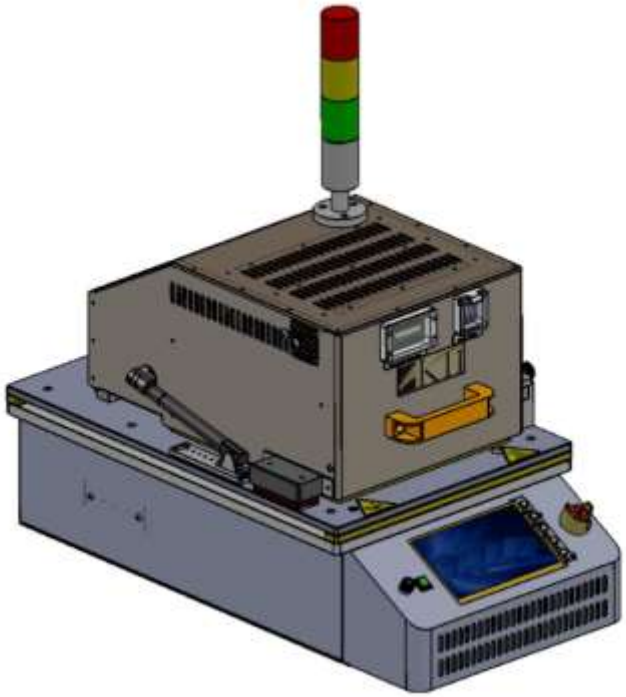


图 7-82 Chassis with chamber

7.13.2 DDR5 UDIMM 研发测试平台产品规格

DDR5 UDIMM Test System Technical Specifications		
Configuration		2 DUT per chassis
Operation Modes		Manual loading / unloading. Parallel testing
Control Software		Control software installed on external Control PC with Windows 10 Professional
Number of Test Sockets		UDIMM High Performance Model – 2x 288-pin ZIF test sockets per chassis UDIMM High Capacity Model – 4x 288-pin ZIF test sockets per chassis SODIMM Model - 4x 262-pin ZIF test sockets per chassis Note : Other chassis options are available in the future
Memory Type Supported		DDR5 SDRAM 8/16Gb density, x8/16 width; 32Gb future support
DIMM Standards Supported		DDR5 ECC and non-ECC UDIMM and non-ECC SODIMM DDR3200 / 3600 / 4000 / 4400 / 4800 / 6400; over-clock to 8400 Mbps 2 rank max. Support DDR5 to UDIMM adapter (Socket DIMM adapter)
Test Frequency		Clock 1600 / 1800 / 2000 / 2200 / 2400 up to 4200 MHz Fine adjustment per MHz using CPU internal multiplier

I/O Interface		DDR5 CK/CA/DQ/DQS/DM POD 1.1V DDR5 RESETn CMOS 1.1V SPD I2C CMOS 3.3V, open-drain PMC PWR_GOOD POD 1.1V
Temperature Control Chamber		Optional heat chamber with control profile up to 85°C, Uniformity: ±3℃
Test Site[1]		
Address Generation		17X + 10Y + 2BA + 3BG
# of Rank Supported		2-rank
Data lines		2-set 32/36bits
Control Lines		2-set CA/CSn/PAR set, 1-set ALERTn/RESETn
DUT Clocks		2 pair differential
Device Access Parameters		tRCD, tCL, tWL, tRAS, tRC, tRP, tCWL, tCSSR, tCWL, tWR, tWRPRE, tWRA, tRDA, tRTP tRW, tWW, tRR, tFAW, CmdRate Adjustable in nCLKs
Refresh Control		Per row Refresh time adjustable from 1us up to 32768 memory clock periods
DUT Power		VIN_BULK 5V nominal; adjustable +/-20%, 10mV step, 5A max Each DIMM is individually powered to enhance performance
Current Measurement		Normal mode - 0~5A, resolution 0.1mA, accuracy +/-(2% + 0.1mA) Precision mode - 0 -50mA, resolution 0.2uA, accuracy +/-(2% + 0.2uA) * DUT power rails have over-current protection
Tests Available		Functional Row Hammer, MarchA, MarchX, MarchC, MarchG, MATS, and many more industry standard tests. Application ATSLWB, JumpLWB, Shift, MoveRot, MT64 Others Rank Margin test, SPD test, IDDx
SPD Operations		SPD Read, Write, Serialization, Import, Export, Address Check, Check Sum, Auto Setup , Program, Compare, Program Exclude, Compare Exclude, Hex View, Symbolic View
I2C/I3C Operations		I2C and I3C access to DUT Hub and downstream devices - PMIC, TS
Error Logging		Individual 72 DQ failures indicator Record last fail physical address (17X / 10Y / 2BA / 3BG) per rank

Other Features		Shmoo and Burn-in tools Comprehensive test flow control and binning control Speed Grading and Sorting support Test and Results Data logging capability Front Panel LCD to display tester status information
Optional Features		PPR (Post Package Repair) Support Device Core parameter Profiler DQ Eye Diagram
Test System Environmental		
Tester Chassis Dimension		560 x 360 x 190mm (D x W x H) approx.
Temperature Chamber Chassis Dimension		560 x 360 x 300mm (D x W x H) approx.
Connection to Control PC		100/1000Mbase Ethernet with CAT5 cable
Chassis cooling		Liquid cooling
Tester Station Weight		18Kg approx.
Temperature Chamber Weight		22Kg approx.
AC Power Source		1200W， 100~127V@12A ,60/50Hz 200~240V@7.5A ,60/50Hz
Operating Temperature		10 °C– 25°C ambient
Operating Humidity		20% - 80 % (non-condensing)
Storage Temperature		5 - 40°C
Storage Humidity		35% – 70% (non-condensing)

图 7-83 Chassis

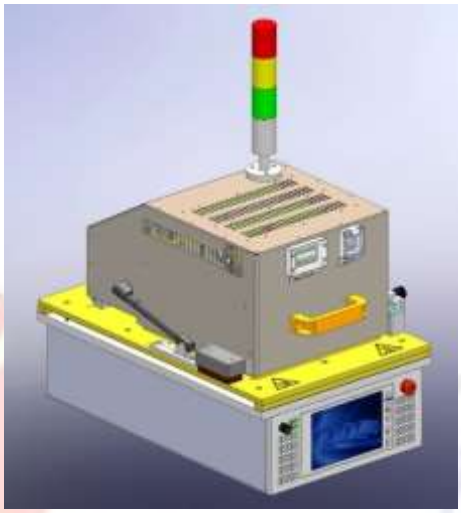


图 7-84 With Heat chamber

7.13.3 DDR5 Socket-DIMM 转接卡产品规格

7.13.3.1 DDR5 Socket-DIMM DDR5 x 16 1Rank Non-ECC UDIMM

DESCRIPTION

This Socket DIMM has been tested and verified to be reliable running at DDR5 2800MHz (PC5-5600) , low latency timing of 46-45-45 at 1.1V. It supports numerous DDR5 x16 FBGA package sizes, such as; (7.5mm x 13mm, 9mm x 14mm, 10mm x 14mm, 10.1mm x 14mm, etc …) . *It requires an appropriate guiding tool for a specific DDR5 package size.*

FEATURES

- **DDR5 functionality and operations supported as defined in the component data sheet**
- **288-pin, DDR5 unbuffered dual in-line memory module (DDR5 NON-ECC**

UDIMM)

- **Fast data transfer rate: PC5-4800, PC5-5200, PC5-5600**
- **Single-rank**
- **16 internal banks (x16): 4 groups of 4 banks each**
- **Sideband access with I3C-basic/I2C support**
- **Two independent I/O sub channels for increased bandwidth**
- **Halogen-free**
- **Fly-by topology**
- **Gold edge contacts**
- **Terminated clock, control and command/address bus**

Options Marking

- **Operating temperature**
  - Commercial (0°C ≤ TOPER ≤ 95°C) C
- **Frequency/CAS latency**
  - 0.416ns @ CL = 40 (DDR5-4800) 48B
  - 0.384ns @ CL = 42 (DDR5-5200) 52B
  - 0.357ns @ CL = 46 (DDR5-5600) 56B

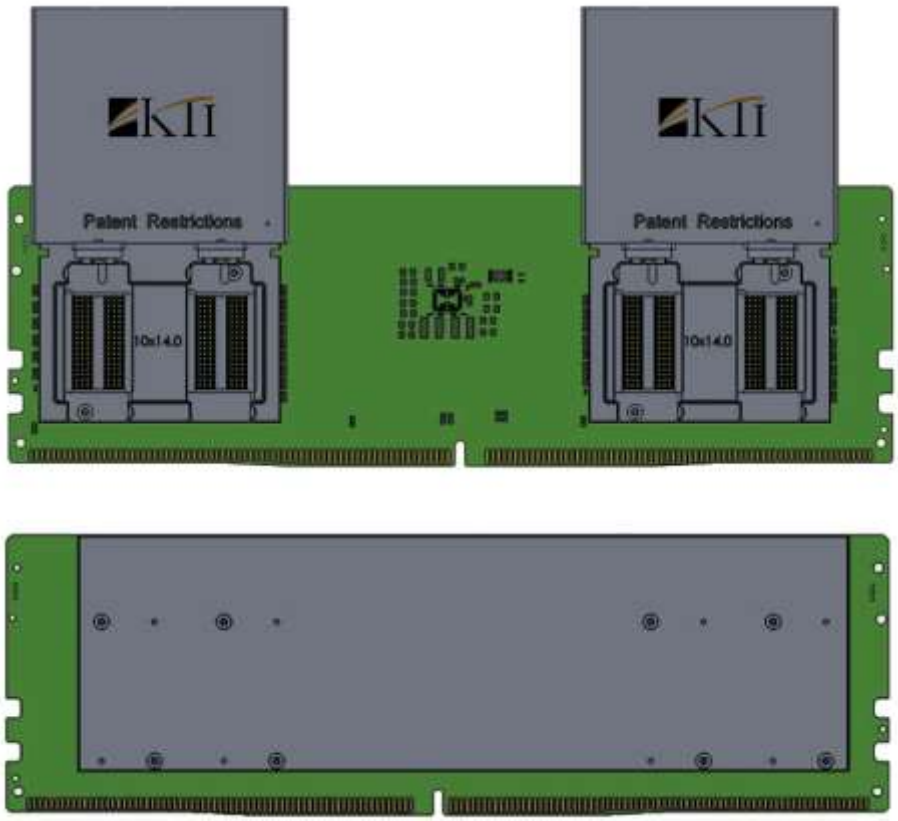


图 7-85

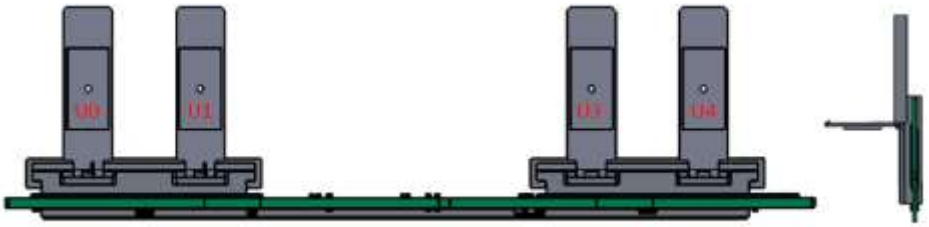


图 7-86

图 7-87 实物图

7.13.3.2 DDR5 Socket-DIMM DDR5 x 8 1Rank Non-ECC UDIMM

基本技术指标基本和 7.9.3.1 差不多。

DESCRIPTION

This Socket DIMM has been tested and verified to be reliable running at DDR5 2800MHz (PC5-5600) low latency timing of 46-45-45 at 1.1V. It supports numerous DDR5 x8 FBGA package sizes, such as; (9mm x 11mm, 10mm x 11mm, 10.1mm x 11mm, etc …) . *It requires an appropriate guiding tool for a specific DDR5 package size.*

FEATURES

- **DDR5 functionality and operations supported as defined in the component data sheet**

- **288-pin, DDR5 unbuffered dual in-line memory module (DDR5 NON-ECC UDIMM)**
- **Fast data transfer rate: PC5-4800, PC5-5200, PC5-5600**
- **Single-rank**
- **32 internal banks(x4, x8): 8 groups of 4 banks each**
- **Sideband access with I3C-basic/I2C support**
- **Two independent I/O sub channels for increased bandwidth**
- **Halogen-free**
- **Fly-by topology**
- **Gold edge contacts**
- **Terminated clock, control and command/address bus**

Options Marking

- **Operating temperature**
  - Commercial (0°C ≤ TOPER ≤ 95°C) C
- **Frequency/CAS latency**
  - 0.416ns @ CL = 40 (DDR5-4800) 48B
  - 0.384ns @ CL = 42 (DDR5-5200) 52B
  - 0.357ns @ CL = 46 (DDR5-5600) 56B

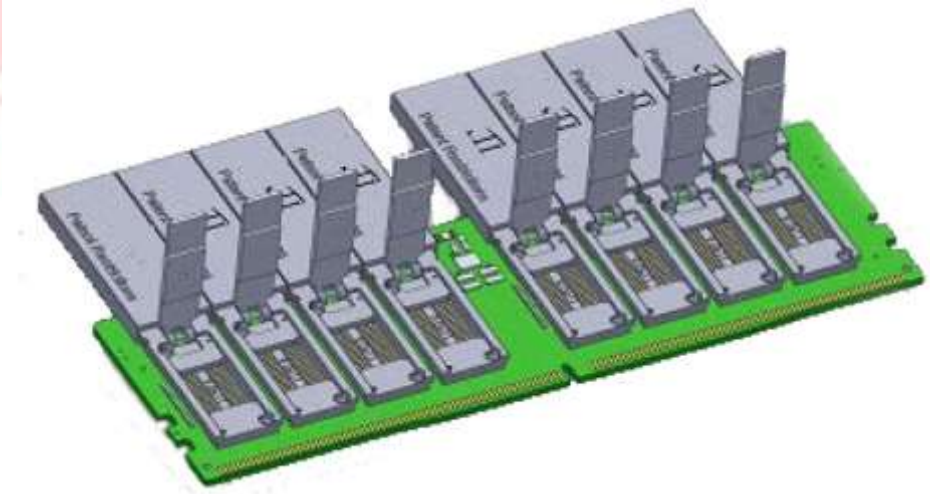


图 7-88

图 7-89 实物图

7.13.4 LPDD4X/LPDDR4/LPDDR3/eMCP/eMMC 测试平台规格

2019 年 2 月 20 日，JEDEC(固态存储协会)正式发布了 JESD209-5，即 Low Power Double Data Rate 5 (LPDDR5)全新低功耗内存标准。相较于 2014 年发布的第一代 LPDDR4 标准，LPDDR5 的 I/O 速度从 3200 MT/s 提升到 6400 MT/s(DRAM 速度 6400Mbps)，直接翻番。

如果匹配高端智能机常见的 64bit bus，每秒可以传送 51.2GB 数据;要是 PC 的 128bit BUS，每秒破 100GB 无压力。

固态协会认为，LPDDR5 有望对下一代便携电子设备(手机、平板)的性能产生巨大提升，为了实现这一改进，标准对 LPDDR5 体系结构进行了重新设计，转向最高 16 Bank 可编程和多时钟体系结构。

同时，还引入了数据复制(Data-Copy)和写 X(Write-X)两个减少数据传输操作的命令来降低整体系统功耗，前者可以将单个阵脚的数据直接复制到其它针脚，后者则减少了 SoC 和 RAM 传递数据时的耗电。

另外，LPDDR5 还引入了链路 ECC 纠错，信号电压 250mV，Vddq/Vdd2 电压还是 1.1V。

之前的 LPDDR4 由于输入/输出接口数据传输速度最高可达 3200Mbps，是通常使用的 DDR3 DRAM 的两倍，新推出的 8Gb LPDDR4 内存可以支持超高清影像的拍摄和播放,并能持续拍摄 2000 万像素的高清照片。

与 LPDDR3 内存芯片相比，LPDDR4 的运行电压降为 1.1 伏，堪称适用于大屏幕智能手机和平板电脑、高性能网络系统的最低功耗存储解决方案。以 2GB 内存封装为例，比起基于 4Gb LPDDR3 芯片的 2GB 内存封装，基于 8Gb LPDDR4 芯片的 2GB 内存封装因运行电压的降低和处理速度的提升，最大可节省 40%的耗电量。同时，新产品的输入/输出信号传输采用三星独有的低电压摆幅终端逻辑(LVSTL, Low Voltage Swing Terminated Logic)，不仅进一步降低了 LPDDR4 芯片的耗电量，并使芯片能在低电压下进行高频率运转，实现了电源使用效率的最优化。

现在市场上目前针对 LPDDR5 的应用主要有手机 SoC 和 SSD controller 等，但是针对 LPDDR5 的测试设备还基本都在开发完善过程中。所以，目前针对 LPDDR4X 的测试需求还有一些。下面是针对 LPDDR4X 内存颗粒的测试工具简单介绍。

图 7-90

图 7-91

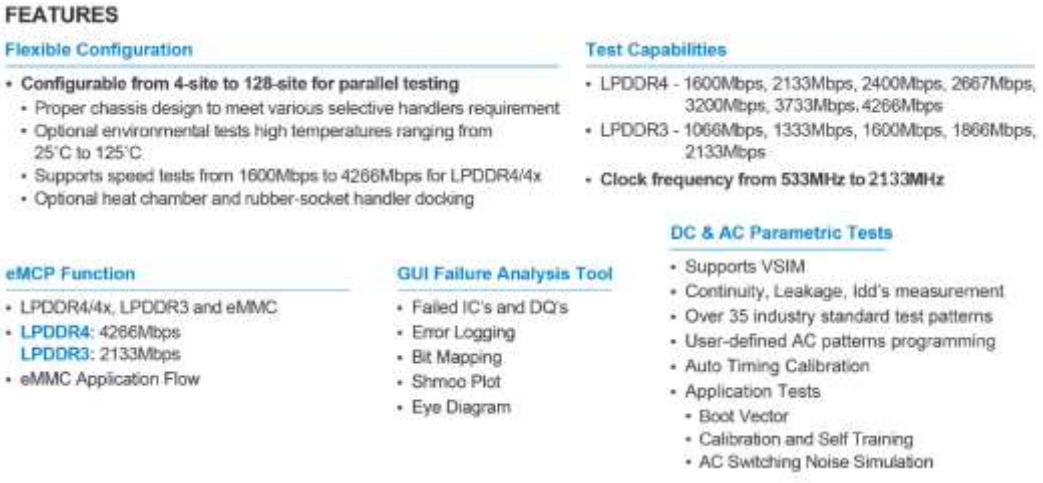


图 7-92

图 7-93

7.13.4.1 LPDDR4 Interposers

有的时候需要将 LPDDR4X 的信号拿出来给示波器，逻辑分析仪，或者协议分析仪进行问题调试，这个时候需要 LPDDR4 的 interposer。业内有标准的 LPDDR4 interposer，也有的场景需要结合用户的需求进行定制。

Key Features

- High fidelity interposers for LPDDR4
- Low and high-speed operation
- Data rates exceeding 4,267MT/s
- Enables oscilloscope or logic/memory analyzer probing
- Patented interposer/probe designs

Applications

- LPDDR4
  - Memory validation and debug
  - Monitoring bus traffic
  - Bus traffic measurement
  - Optimization of memory performance
  - Analog insight
- LPDDR4 rates above LPDDR4-4267

图 7-94

8. SSD 批量测试/RDT/高低温测试方案

8.1 SSD 批量测试设备

8.1.1 产品概览

- **PCI-Gen4 NVMe 256 para Tester**
- **16 AMD Server**
  - AMD EPYC Server CPU 7002 , PCI Gen4 x16 four slots, DRAM 1 TB, Master pc, KVM, 24 Inch Monitor, Ethernet hub
- **aardvark I2C emulator support SMBUS comm**
- **power on/off control, UART serial for NVMe debug**

Rack dimension : 19inch full rack

- **Height 2000mm, depth 900mm, width 450mm**

图 8-1

图 8-2

8.1.2 监控功能

- **Simple test-run with PC-GUI program**
  - Execution and monitoring for each test boards
  - Analysis and display for damaged blocks
  - Easy development platform for new test cases
  - Database for all test actions and logs

8.1.3 测试界面

图 8-3

8.2 SSD 专用 RDT 测试温箱

8.2.1 P41000 - PCIe/NVMe Burn-In tester

Key Features:

- P41000HT: 40 port (optional dual port available)
- PCIe-Gen3, 4 lanes per port (gen3, 2 lanes for dual port tester)
- Support Gen1, Gen2, Gen3 auto select or manual select
- Exchangeable loader adapter technology
  - [default] 2.5" SFF8639, U.2, U.3
  - [optional] M.2 card adapter

- PMU (Power Management Unit) – programmable via USB or Ethernet
  - Drive insertion detection circuitry
  - Power on/off control, LED control
  - Power Margin +/- 10%
- WEB based GUI control SW
- Customized UART connectivity
- Temperature control: -10 ~ 85C

图 8-4

8.2.2 老化测试平台 BI120A/BI-003

BI-120A

BI-003

图 8-5

8.2.3 桌面测试平台 BI-003/P8100/T400

BI-003

READ MORE

P8100

READ MORE

P12000

图 8-6

8.3 SSD 专用测试温箱

数据中心使用的 SSD 一般在相对恒温、恒湿的空调机房里面，例如温度常年 22±2℃左右，湿度 45~65%RH。但是，由于 SSD 一般都是在主机或者存储系统的机头和盘柜里面，受制于通风散热以及 I/O 读写等各方面的影响，一般在研发阶段会测试在 0-55 度温度范围 SSD 的各种性能和行为。

当然，对于一些用于汽车电子信息娱乐导航系统里面，或者嵌入式设备里面的 SSD，由于这些产品经常要在户外，所以温度往往会到 -55 度（中国漠河）或者 40 度以上（例如，阳光暴晒的车体），所以针对这些产品的测试往往需要更宽的温度范围，例如-45 ~ 85℃。

8.3.1 美日韩基于 FPGA 的测试温箱

针对 SSD 的测试温箱大体分为两大类，一类是专业的 SSD 测试温箱，参见下图为一个 4 个 chamber 箱体的专用温箱。这些专用测试设备大多来自美国，日本和韩国，价格非常昂贵，采用专门定制的 FPGA 测试板卡，支持 60/120/240 片 SSD，温度范围 -40~105℃，当然不同厂商、不同型号的产品支持的 SSD 类型、盘的数量、温度范围也有差异。这些设备提供专门的针对 SSD 测试图形化测试软件，通常也支持私有的脚本编程。

图 8-7

8.3.2 基于 X86 CPU 和 SWITCH 的测试温箱

第二类 SSD 测试温箱也是 Turnkey solution 测试温箱，这类系统一般是基于 X86 CPU 平台结合专门定制的测试卡以及测试 cage 等集成而成，在支持同样的温度范围以及测试同等数量的 SSD 的情况下，产品的整体拥有成本相对第一类专业设备要经济很多。

下面以国内 SSD 生产厂常用的 SSD 测试温箱为例简单介绍一下，整机测试系统主要包括高低温箱、测试主板、PM 板和测试软件等。

8.3.2.1 控制方式与特色

平衡调温调湿 BTC 控制系统,以 PID 方式控制 SSR,使系统之加热湿量等于热湿损耗量,故能长期稳定的使用。我们提供两类测试温箱，一种为普通升温/降温型，一种是快速升温/降温型号。

其中，快速升温/降温参数如下：

- **-20℃→+150℃ 10℃～15℃/min，约 12 分钟以内（机械制冷, 标准负载下)**
- **10℃～15℃/min 非线性可调（出风口处测量，机械制冷, 标准负载下）**
- **同时可以满足 5℃～8℃/min 线性可调（出风口处测量，机械制冷, 标准负载下）**

8.3.2.2 整机产品外型

图 8-8

整机主要有如下一些特性：

- **支持 SSD 测试片数定制化，例如 126 片、256 片、300 片、420、516 片等等（按需求定制）；**
- **支持研发微小型定制化便携式，例如 6 片等等；**
- **支持（-70~+180 度）的测试；**
- **支持异常断电测试和老化测试；**
- **支持自动化温控测试；**
- **支持全部采用软体进行智能化控制测试；**
- **支持测试测试软体的定制化；**
- **支持箱内风速与温度均衡；**
- **支持快速升降温控制；**
- **支持 PCIe/eMMC/UFS/DRAM/Flash 老化的定制化研发；**
- **支持网络化控制，可以异地控制测试并看测试结果；**
- **支持 APP 远程控制测试；**

8.3.2.3 整机主要组成部分

整机测试系统主要包括高低温箱、PC 主板、PM 板和测试软体等等，整机的各项组织如下图所示：

图 8-9

8.3.2.4 测试硬件示例

硬件部分主要由如下几部分组成：

- **测试箱外壳；**
- **PC 主板；**
- **PM 板；**
- **电源控制部件；**

8.3.2.5 测试软件示例

图 8-10

软体部分主要由如下部分组成：

- **Test PC：主要分为如下几种 PCT、BIT、MDT 与 FDS；**
- **Console：可以控制整个 Test PC 操作，是测试的控制接口，用于发送测试指令与配置脚本，是测试的指挥中心；**
- **DMS：用于保存所有测试结果；**
- **QMS：用于网络管控操作；**
- **Linux 操作系统；**

8.4 ThermoJet 快速高低温气流温度冲击系统

如果想在实验室常温测试环境下仅仅对运行中的板卡或者外设上某一个主控芯片做高低温测试，最合适的方式就是采用 ThermoJet 设备，参见下图，一般这种设备只能对一个 DUT 进行测试。

图 8-11

这种新型的快速高低温气流温度冲击（循环）系统—Thermojet ，对比于传统的温箱，有以下几个特征:

1）可以针对整块 PCB 板提供温度环境或高低温冲击。对比于传统的温箱，它的优势是升降温速率非常迅速, -55℃～125℃<10 秒；

2）如果一块 PCB 板上有很多元器件，但你只需要针对其中的某一个 IC 单独进行高低温冲击而不影响周边其它元器件，那么传统的温箱无法解决这类测试，只能用 Thermojet 进行隔离冲击。

3）对测试机平台 load board 上的 IC 进行温度循环或冲击；传统温箱无法针对此类测试。

上述测试场合传统的温箱无法解决！就必然要引入一种新型技术的高低温冲击/循环设备—ThermoJet

它完全颠覆了传统温度循环箱的工作模式。有着如下优点：

1) 做温度冲击时, 温变十分迅速：-55℃~+125℃ <10 秒;真正达到温度“冲击”目的；

2) 做恒温测试时，可以稳定的维持在某个温度点.精度可达±1℃;

3) 实时监控 IC 体表真实温度, 实现温度闭环反馈;

4) 升降温时间可控.可程序化操作。

非常适合于手动测试。用于 design house/实验室/研究所/QA 部门等；亦可用于生产线自动化大批量测试。

下图将 PCB 板上的某一个 IC 隔离出来单独进行温度冲击而不影响周边其它元器件，这个时候需要 case by case 设计将待测芯片罩起来的“罩子”，这个定制交付周期一般需要 3 个月时间。参见下图。

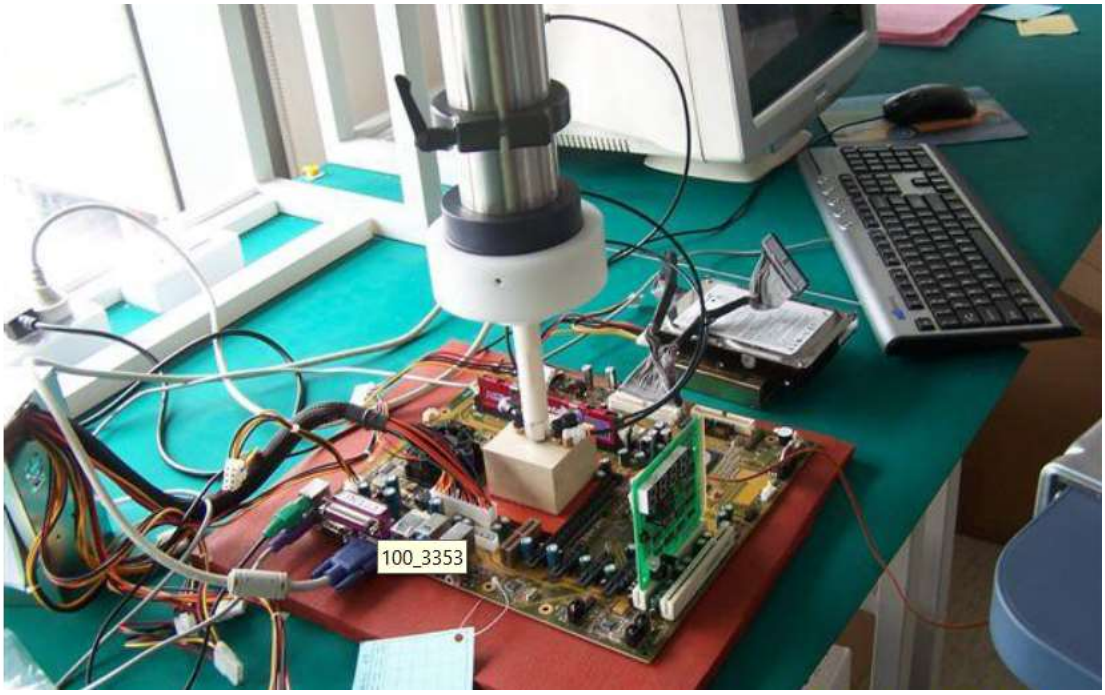


图 8-12

8.5 Peltier 高低温测试模组

有的时候需要测试的温度范围不大，例如，-20 ~ 85 度，这个时候也可以通过在是室温下使用 Peltier 技术和产品实现，参见下图。

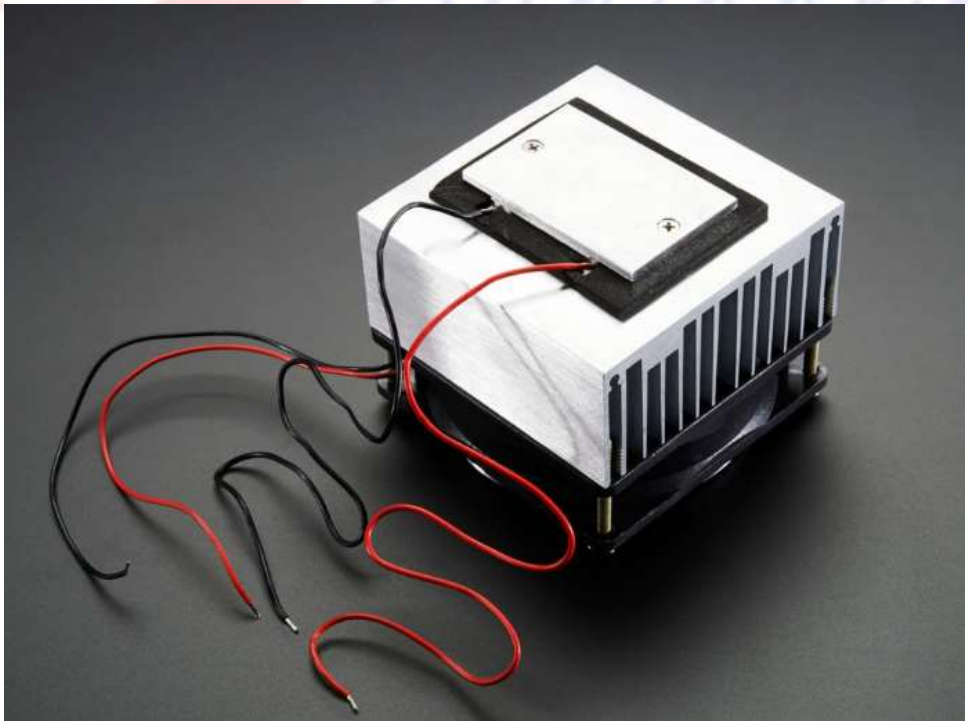


图 8-13

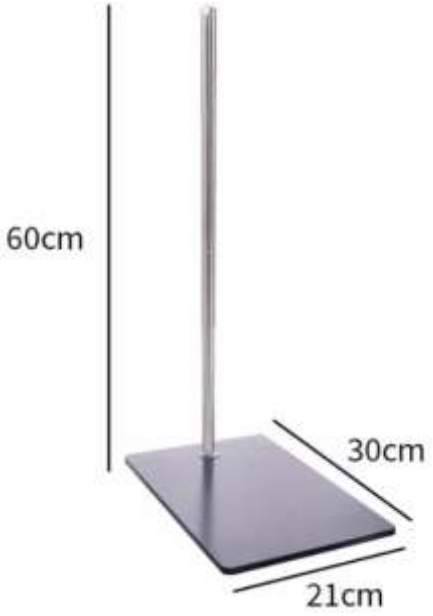
上面的图片是 Adafruit Industries 公司的 Peltier Thermo-Electric Cooler Module + Heatsink Assembly 示例。

8.6 PCIe Gen 4/5/6 SSD 测试托架和机架

8.6.1 PCIe 协议分析仪 slot interposer 托架+夹具

这里主要指的是针对插卡的 slot interposer，如果 interposer 顶部再次插入待测的插卡如果较大、较重的情况下，同时有没有 “信号过硬”的 PCIe Gen 5/6 延长线的情况下，我们简易采用下面的托架 + 多向夹，可以稳固住 interposer + 待测插卡（DUT）。

铁架台主要用来提供支撑，建议试用较粗的立柱，方便多向夹可以加紧（否则需要填充物），下面左侧的三种托架，推荐中间的一款，重量大概 5-6KG，右边的是它的具体尺寸。



下面是托架 + 2*大力多向夹固定好好以后空载的状态图片。

下面的图片是使用上述的托架 + 2*大力多向夹，分别用来夹紧 slot interposer 和 DUT 插卡。

8.6.2 测试台主板托架

8.6.2.1 主板托架顶视图

图 8-14 主板托架实拍图

8.6.2.2 主板托架侧视图

图 8-15 主板托架实拍图

8.6.3 SSD 测试实验室机架

SSD Rack 整体设计已获得国家发明专利，该机架主要用于机房大批量测试 SSD 使用，可以非常高效地利用机房的空间，也大大提高测试得便利性，测试人员可以在机房内部通过 KVM 或者远程直接查看每台主机的测试情况。下面是一些简单介绍。

8.6.3.1 SSD 机架技术规格

- SSD RACK 总体高度大概 2 米，查看最上层主机需要人字梯。
- SSD RACK 包含 5 层，每层放置 5 片主板，5 组 PC 电源，每个主板测试 4 个 SSD，单台 SSD RACK 总计可以测试 100 片 SSD。
- 使用 16A 8 口 PDU/智能 PDU(PDU 可配置智能 PDU，可远程监控开关电源。
- 配备千兆网络适配器，可远程监控 Rack 测试状态。
- PC 部分可根据客户需求配置。
- KVM 结合网络适配器，可实现远程查看，监视每个产品的测试状态。
- 是否需要远程功能根据实际需求配置。

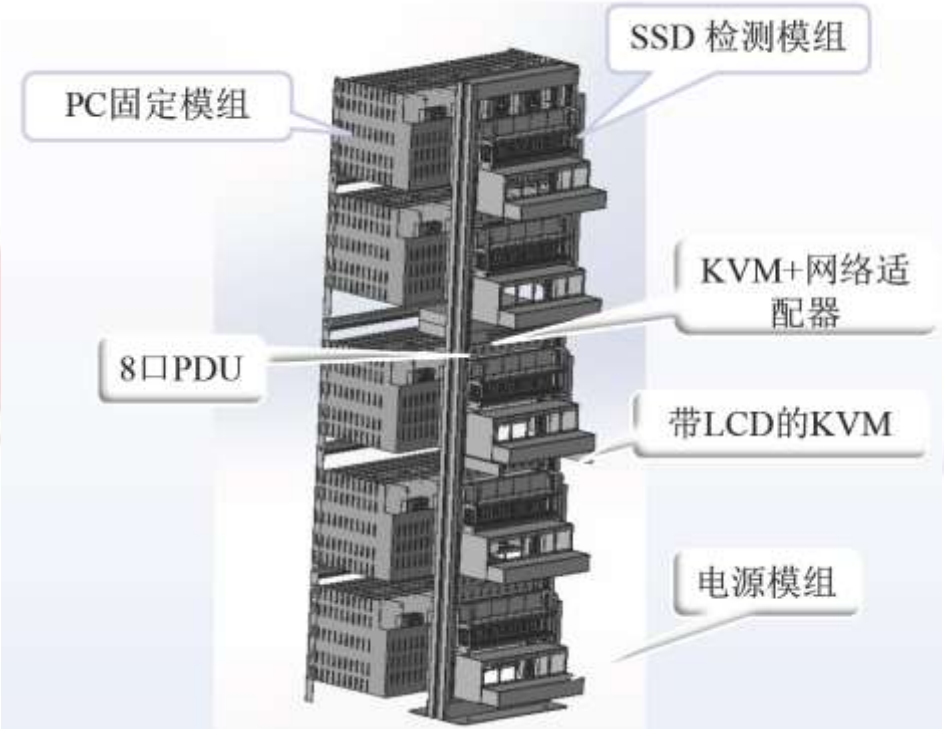


图 8-16 SSD 机架主要组件示意图

8.6.3.2 SSD 机架外形尺寸

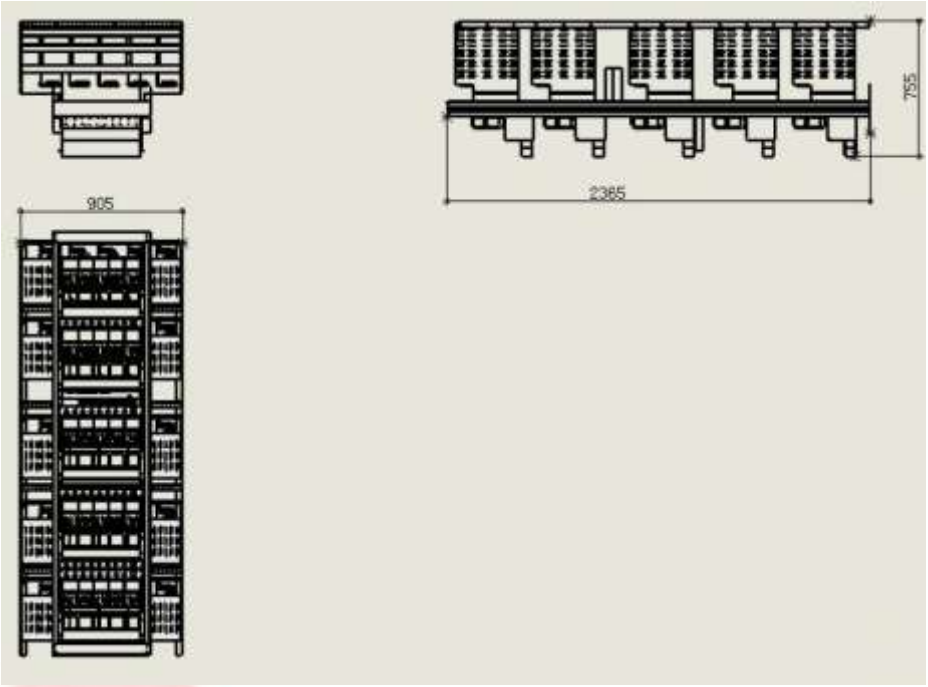


图 8-17 SSD 机架尺寸图

9. UFS 4.0&I3C 协议分析仪

9.1 UFS 4.0 协议分析仪

PROTOCOL INSIGHT®

UFS 4.0 协议分析仪是一种用于抓取并且解码、分析 UFS host 与 USF device 之间的通信的协议层分析仪工具。 它可以跨 MPHY、UniPro 和 UFS 协议层捕获和调试数据。目前业内最新的 UFS 4.0 协议分析仪支持速度高达 23Gbps 的 MIPI M-PHY HS-Gear 5。

业内知名的 UFS 协议分析仪公司是 Protocol Insight。

- **Protocol Insight 为开发移动和受移动影响的产品的客户提供测试和测量工具**
- **Protocol Insight 是 UFS 和 UniPro 协议分析和生成的市场领导者**

– 自 2014 年以来，Protocol Insight 一直在提供 UFS 和 UniPro 调试和分析工具

– Protocol Insight 的理念是无与伦比的产品领先地位

主要特性和优点

- 支持 UFS 4.0、UniPro 2.0 和 M-PHY 5.0 HS-G5

• Smart Tune™ 均衡允许每次突发时对 PHY 进行连续均衡。

• Trace Validation™是一项获得专利的人工智能(AI) 工具，它使用复杂的状态机逻辑以算法方式分析轨迹，无需用户推理或辨别。（仅限猎鹰系列）

- 流式捕获使用 Thunderbolt™ 3 的完整 40Gb/s 带宽将跟踪实时保存到磁盘。
- 事件视图将原始符号显示为独特的时间对齐显示中的数据包事件，并通过以下方式呈现所有事件的完整图片： 向下钻取到最低级别字节。
- 合规性/一致性验证执行一系列 CTS 测试用例，使用智能跟踪验证引擎分析结果并生成摘要报告。（仅限猎鹰系列）

- 验证 UFS 设备是否符合 UFSA UFS 合规性测试矩阵。（仅限猎鹰系列）
- 硬件中具有完整 UniPro 堆栈的刺激允许主机仿真和在链路上创建特定流量，并具有广泛的错误注入。（仅限猎鹰系列）
- Test Executive™ 压力测试控制 DUT 并自动执行测试。任意次数的循环或无结果测试后停止案例，然后使用跟踪验证分析结果。（仅限猎鹰系列）
- 设计自定义测试用例以引入极端情况、裕度或压力测试的错误。（仅限猎鹰系列）

产品系列描述

UniPro/UFS 协议分析仪/练习器的 Falcon 系列是既定的行业标准，具有强大的协议分析仪和练习器功能以及前所未有的灵活

性。

- Falcon G500C 是一款可以作为“嗅探器”捕获 x2 双向链路的分析仪。支持 UFS 4.0、UniPro 2.0 和 M-PHY 5.0 HS-G5。
- Falcon G500C – G4 是一款分析仪，可以作为“嗅探器”捕获 x2 双向链路。它支持 UFS 3.1、UniPro 1.8 和 M-PHY 4.1 HS-G4。 G500C – G4 可现场升级以支持 UFS 4.0 和 M-PHY HS-G5。

- Falcon G550C 是一款训练器/分析器；它与 G500C 分析仪相同，但也是一个协议练习器，可以在 x2 双向链路上生成链路流量，同时捕获来自 DUT 的响应流量。 Falcon G550C 练习器可以执行主机仿真并执行 UniPro 和 JEDEC 合规/一致性测试套件 (CTS) 。

- Falcon G550C – G4 练习器/分析器与 G500C – G4 分析器相同，但也是一个协议练习器，可以执行主机仿真并执行 UniPro 1.8 和 JEDEC JESD224A

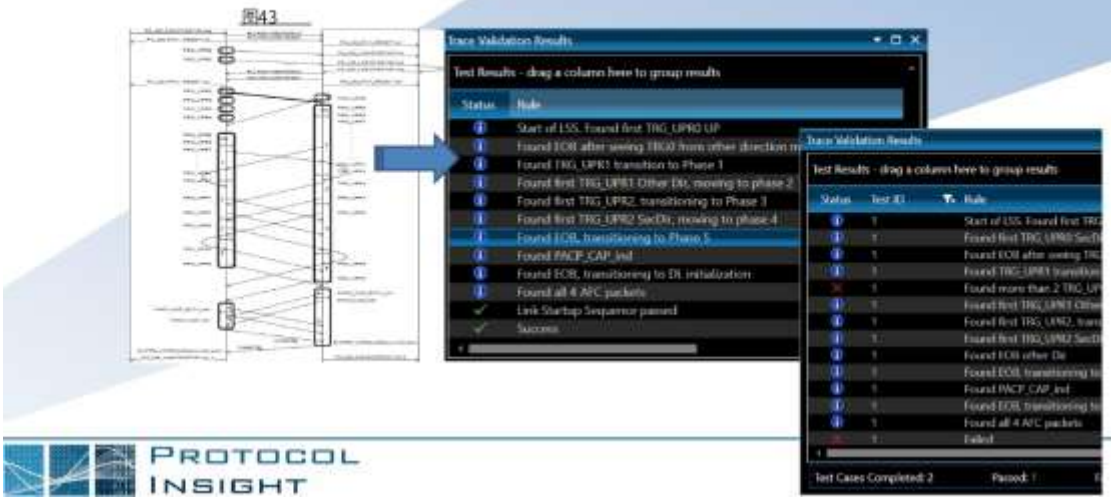
合规性/一致性测试套件 (CTS)。 G550C – G4 可现场升级以执行未来的任务 UniPro 2.0 和 UFS 4.0 合规/一致性测试套件 (CTS)。

- Falcon G400C 分析仪支持 UFS 3.1、UniPro 1.8 和 M-PHY 4.1 HS-G4，并可升级至 G450C 。该产品已被 Falcon G500C – G4 取代。

- Falcon G450C 练习器/分析仪与 G400C 分析仪相同，但也是协议练习器可以执行主机仿真并执行 UniPro 1.8 和 JEDEC JESD224A 合规性/一致性测试套件 (CTS)。该产品已被 Falcon G550C 取代 – G4。

TRACE分析和验证

链接启动顺序调试示例



PROTOCOL INSIGHT

CTS 测试用例执行

测试执行人员执行一致性测试用例,然后验证结果是否符合 CTS。

- UFS CTS 测试用例:符合 UFSA CTM v1.3
  和 JEDEC JESD224A,用于 UFS 2.0/2.1 和 UFS 外部卡扩展。
- 初步 UFS 3.1 合规性测试
- 计划进行初步 UFS 4.0 合规性测试

- UniPro CTS 测试用例:符合MIPI CTS v1.1。
- UniPro 1.8 一致性测试
- 计划进行初步 UniPro 2.0 一致性测试

广泛的报告和分析工具

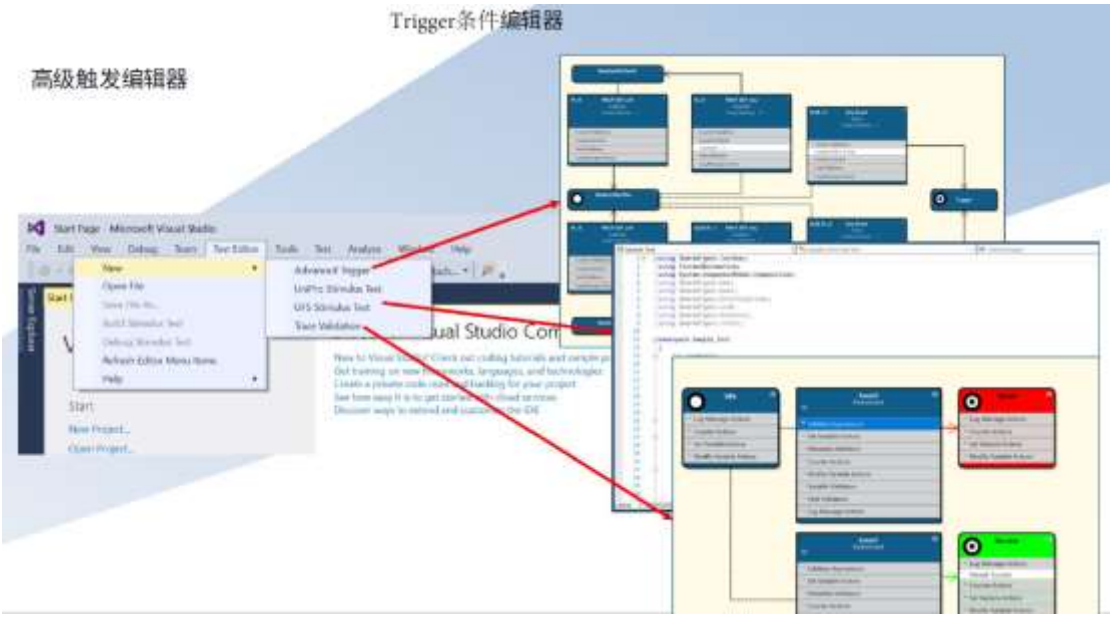
- 按测试参数 – 状态、单独测试或测试规则
- 按协议特征 – 数据包、字节、速度、链路等。 - 摘要和通过/失败报告

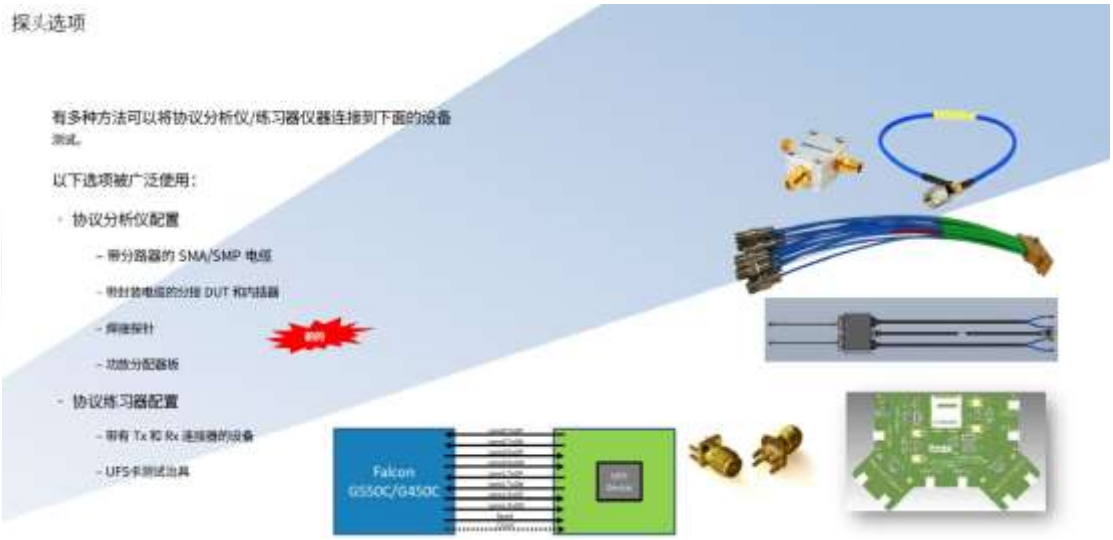
PROTOCOL INSIGHT

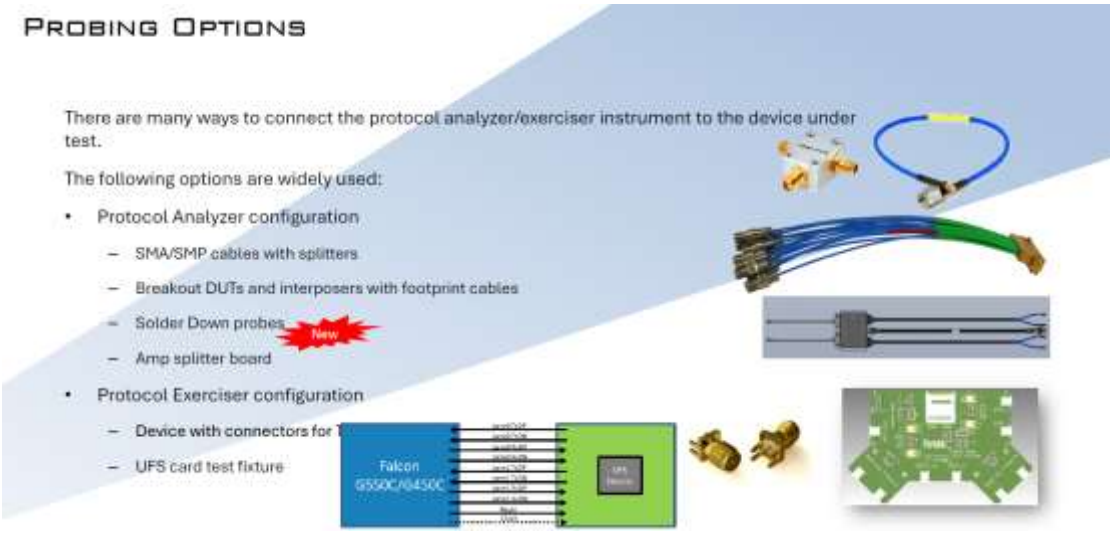
UFS 数据包延迟示例

活动视图

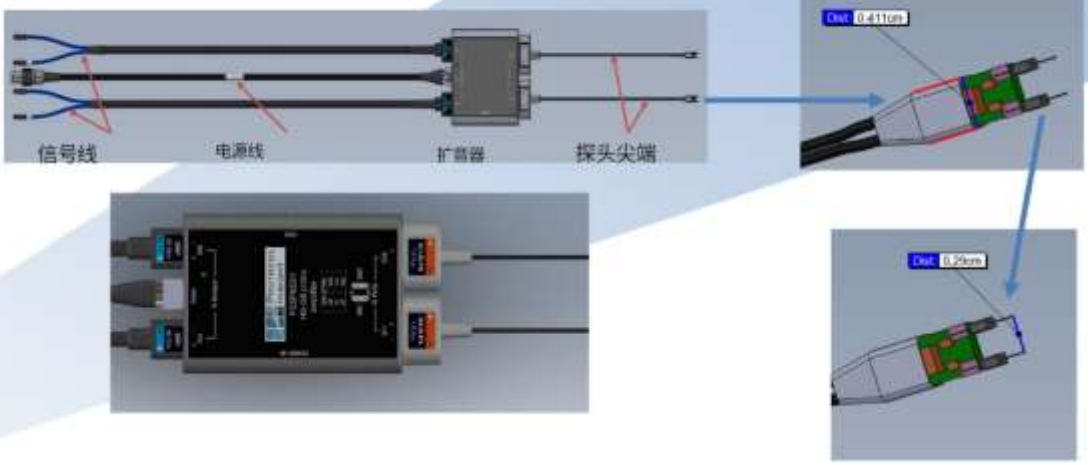
- 显示总线上的所有 UniPro 或 UFS 事件
- 独特的时间对齐显示允许缩小
- 显示低水平填充物、准备、同步、Hibern8、睡眠、停顿等M-PHY级别数据包


Trigger条件编辑器
高级触发编辑器

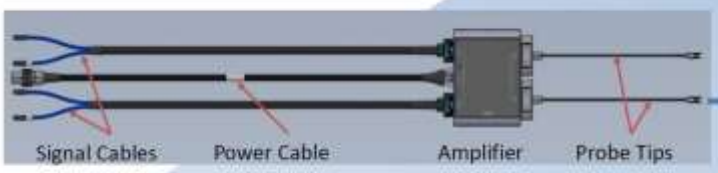

探头选项
有多种方法可以将协议分析仪/练习器仪器连接到下面的设备
测试。
以下选项被广泛使用：
- 协议分析仪配置
– 带分路器的 SMA/SMP 电缆
– 带封装电缆的分线 DUT 和内插器
– 焊接探针
– 功放分配器板
- 协议练习器配置
– 带有 Tx 和 Rx 连接器的设备
– UFS卡测试治具
Falcon
G550C/G450C

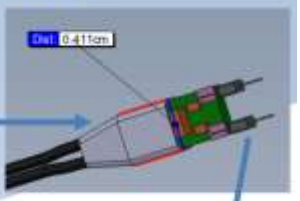

PROBING OPTIONS
There are many ways to connect the protocol analyzer/exerciser instrument to the device under test.
The following options are widely used:
• Protocol Analyzer configuration
– SMA/SMP cables with splitters
– Breakout DUTs and interposers with footprint cables
– Solder Down probes
New
– Amp splitter board
• Protocol Exerciser configuration
– Device with connectors for
– UFS card test fixture
Falcon
G550C/G450C

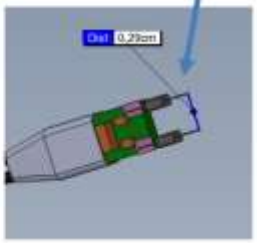
HS-G5 焊接探针



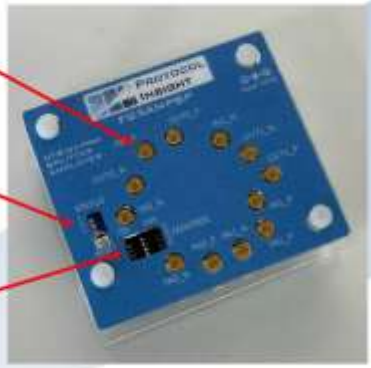
HS-G5 Solder-down Probe



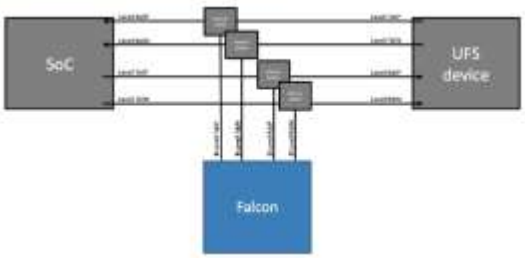




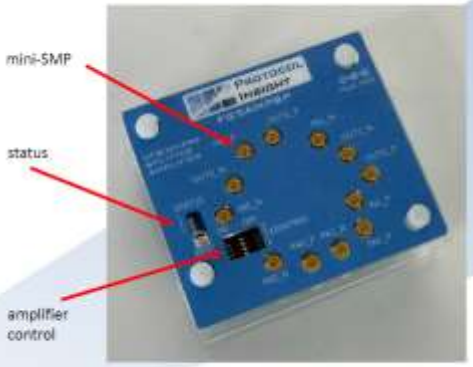
FG5AMPSP - 放大器分配器板



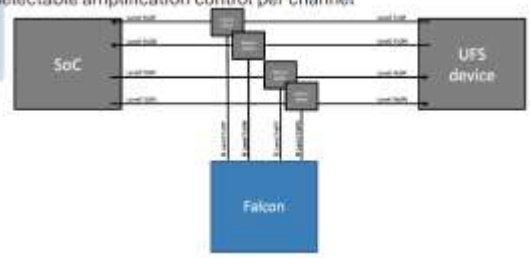
- 为 HS-G5 A/B 启用电阻分配器
- M-PHY v5.0 删除了TX 小幅度模式。 "标准"电阻分路器将面临 HS-G5 操作的挑战。
- 每个放大器分配器板提供一个子链路的连接 - 与标准 COTS 分路器相比，信号裕度提高约 5 dB - 12 个连接器 - 4 个用于仪器、主机、设备的 SMP-mini 连接 - 一块板取代了所有四个分路器，如下所示
- 每个通道可选择放大控制



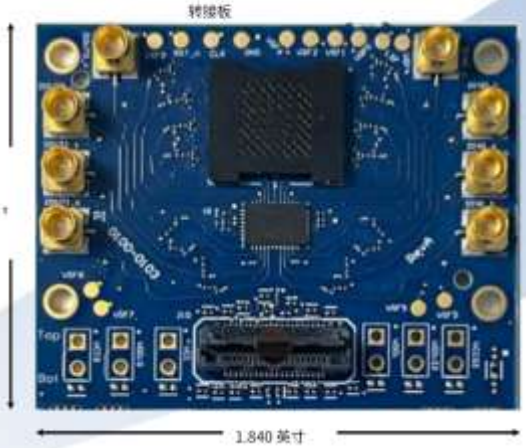
FG5AMPSP - Amp Splitter Board



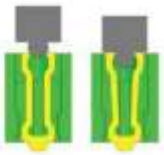
- Enables Resistive Splitters for HS-G5 A/B
  - M-PHY v5.0 removed the TX small amplitude mode. "Standard" resistive splitters will have challenges with HS-G5 operation.
- Each amp splitter board provides connection for one sublink
- Improves signal margin by ~5 dB versus a standard COTS splitter
- 12 connectors - 4 SMP-mini connections for instrument, host, device
- One board replaces all four of the splitters, shown below
- Selectable amplification control per channel



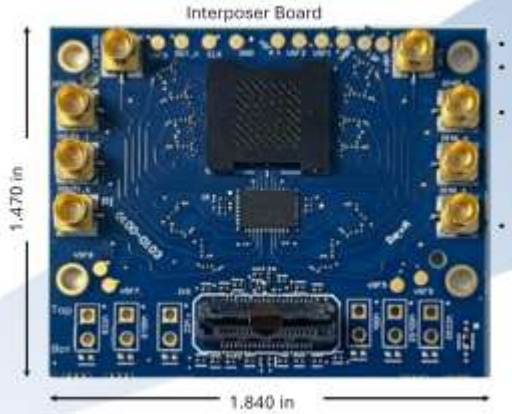
FG5AMPGR – "Grypper"插入



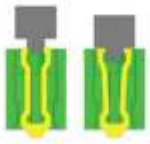
- 基于放大器分配器平台的功能构建
- 为来自设备和主机的仪器提供 4 个 SMP-mini 连接
  - 一块板支持主机和设备的 x2 连接
- 主板提供与客户系统和UFS 的连接
  设备
  - "Grypper"插座为 UFS 设备提供灵活的连接选项，也可用于连接到主机系统
  - 也可以直接焊接到电路板上
- "Grypper"插座直接卡入电路板或封装并直接夹住焊球。



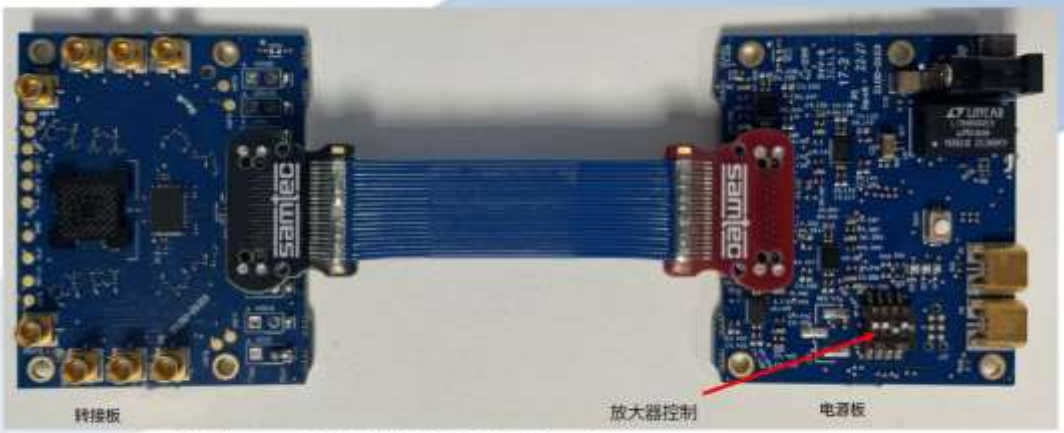
FG5AMPGR – "Grypper" Interposer



- Builds on the capabilities of amp splitter platform
- Provides 4 SMP-mini connections for instrument from device and host
  - One board supports a x2 connection from host and device
- Main board provides the connection to the customer system and UFS device
  - "Grypper" socket provides flexible connection option for UFS device and can also be used to attach to the host system
  - Can also be soldered directly to board
- "Grypper" socket snaps directly to the board or package and grips directly to the solder balls.

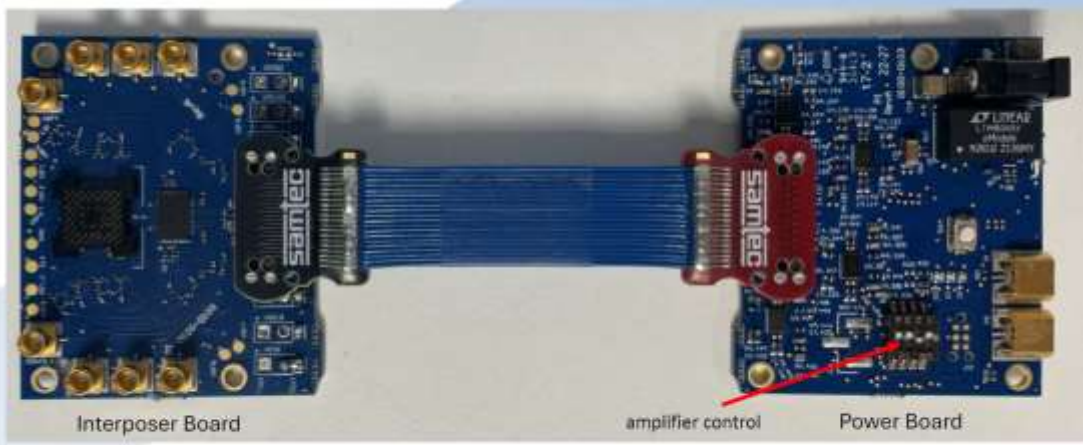


FG5AMPGR – 电源卡



- 与标准 COTS 分路器相比,信号裕度提高约 5 dB
- 为每个通道提供功率和可选择的放大控制
- 灵活的焊带可轻松将中介层与客户平台集成

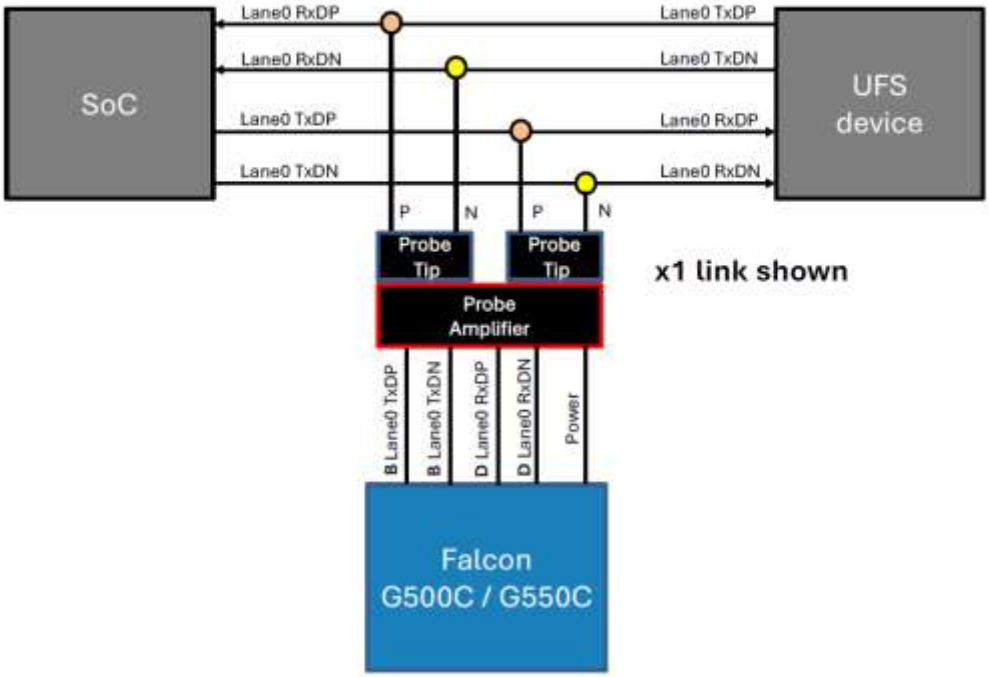
FG5AMPGR – Power Card



- Improves signal margin by ~5 dB versus a standard COTS splitter
- Provides power and selectable amplification control per channel
- Flexible ribbon allows easy integration of interposer with customer platform

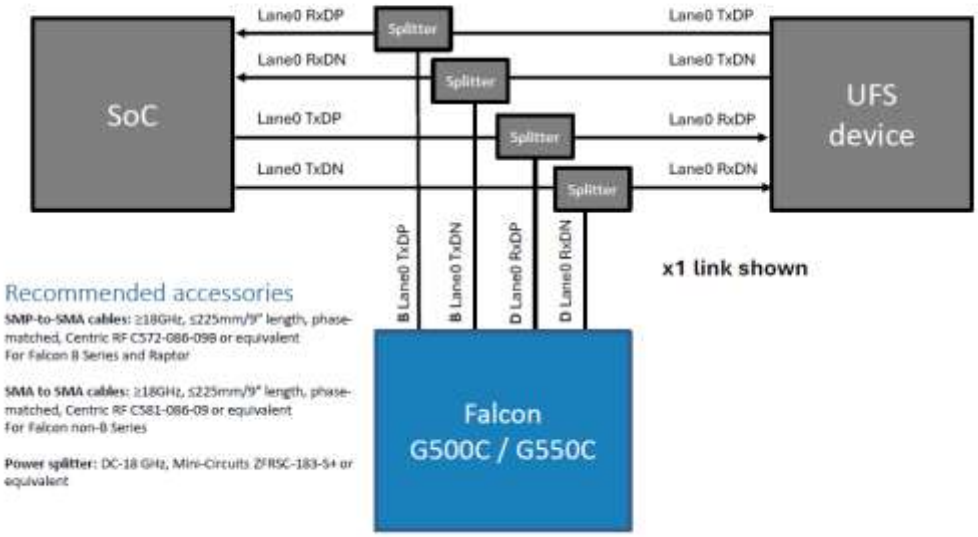
Analyzer

Solder Down HS-G5 Probe



Analyzer

SMA/SMP cables with splitters



Recommended accessories

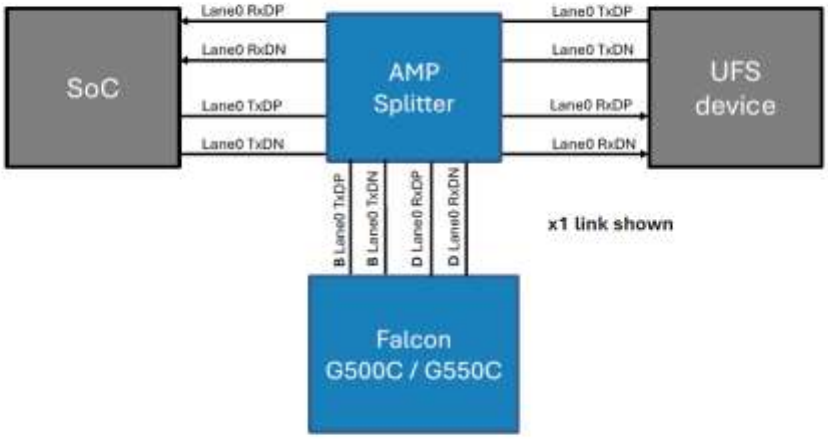
SMP-to-SMA cables: ≥18GHz, ≤225mm/9" length, phase-matched, Centric RF C572-086-098 or equivalent
For Falcon B Series and Raptor

SMA to SMA cables: ≥18GHz, ≤225mm/9" length, phase-matched, Centric RF C581-086-09 or equivalent
For Falcon non-B Series

Power splitter: DC-18 GHz, Mini-Circuits ZFRSC-183-S+ or equivalent

Analyzer

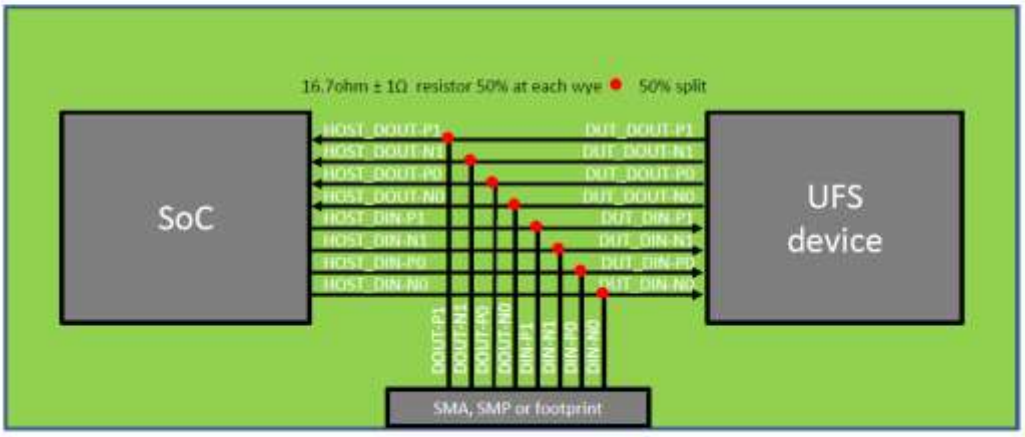
SMP/SMPm cables AMP splitter



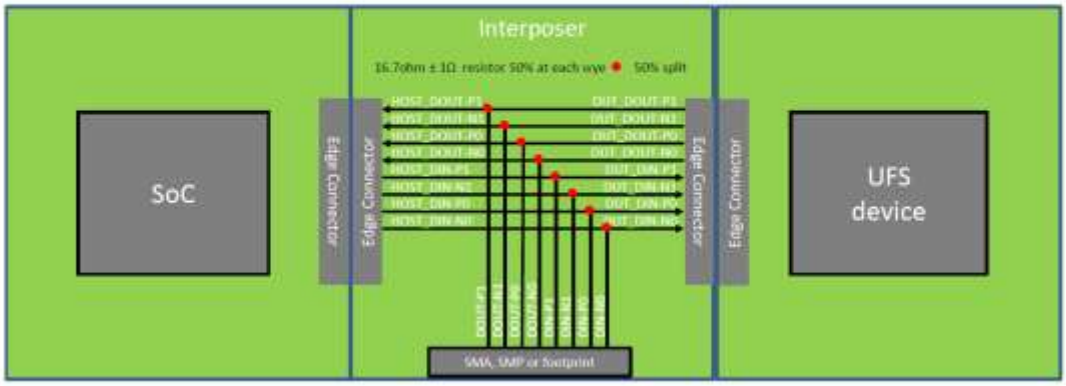
Includes

SMP-to-SMPm cable: FGSAMPSP-C one pair of 9", phase-matched cables, SMPm right-angle to SMP.

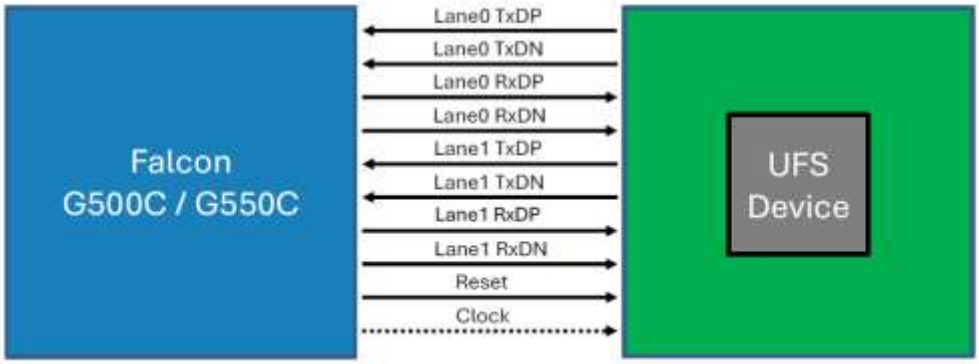
Analyzer - Breakout DUT



ANALYZER - INTERPOSER



EXERCISER



9.2 I3C 协议分析仪

Introspect I3C Exerciser & Protocol Analyzer & Scope

产品特色

- 集合了 Exerciser Master/Slaver、Analyzer 以及示波器功能。
- 支持 I3C Master Device 与 Slave Device CTS(Conformance Test Suite)。
- 单机可配置为 4 个 I3C 装置，作为 Master or Slave 来与系统上的其他 I3C 进行测试。
- 内建 500MHz 示波器能力，同步进行数位与逻辑的封包解析。

- 可生成 I3C 的数据流量并且支持 SDR、HDR 模式。
- 强大触发能力，轻松触发 CCC 上所有的协议以及 Broadcast 通信，包含 IBI 与 Hot-Join 等功能。
- 完整支援 MIPI I3C basic 以及 JEDEC 规范。

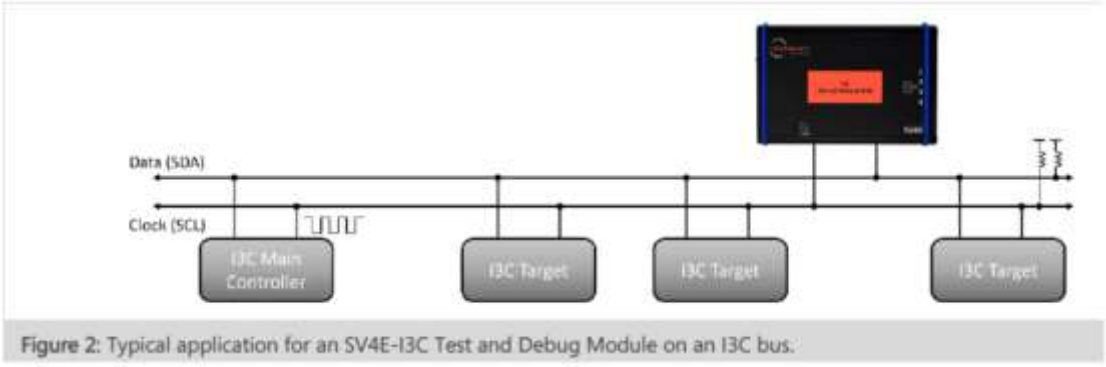
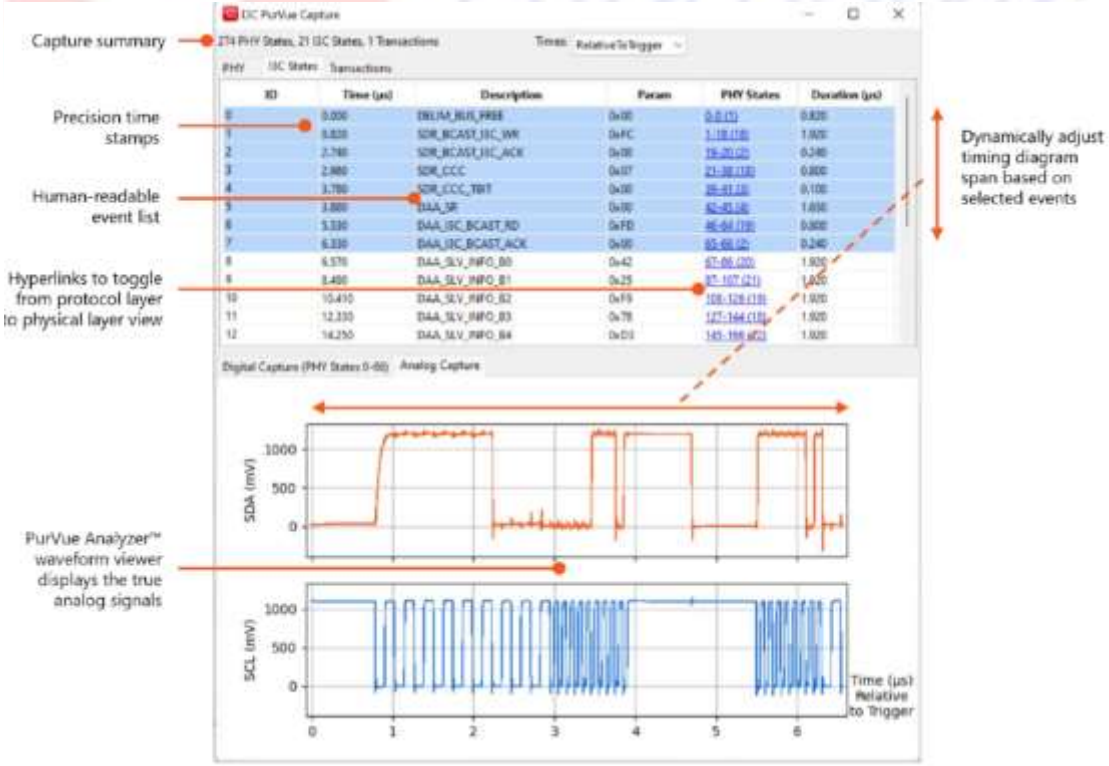


Figure 2: Typical application for an SV4E-I3C Test and Debug Module on an I3C bus.

Introspect technology 是 MIPI Alliance 与 JEDEC 的联盟成员。 Introspect 推出的 I3C 系列产品，使用 USB 连接电脑，透过 IESP 的软体介面进行读/写封包发送及协定撷取与分析的功能，支持 SDR、HDR 模式，向下相容 I2C，并可依照测试需求同时 扮演 Master/Slave 的角色，并且可以在调整封包发送时的电压/时序等参数，让工程师可以完整地对 I3C 产品进行物理层道协议层的功能测试。



9.3 UFS 3.0/4.0 开发板

常用的 UFS 3.0/4.0 开发板目前以美国、加拿大或者韩国为主。下面介绍加拿大的一款使用较多的开发板基于高通 888 平台，下面简介以及 System block diagram，仅供参考。

The HDK8350 development platform consists of two major components:

- ❑ Main board
- ❑ Display card (Display panel + Expansion card)

The following diagrams show the top and bottom on the development platform.

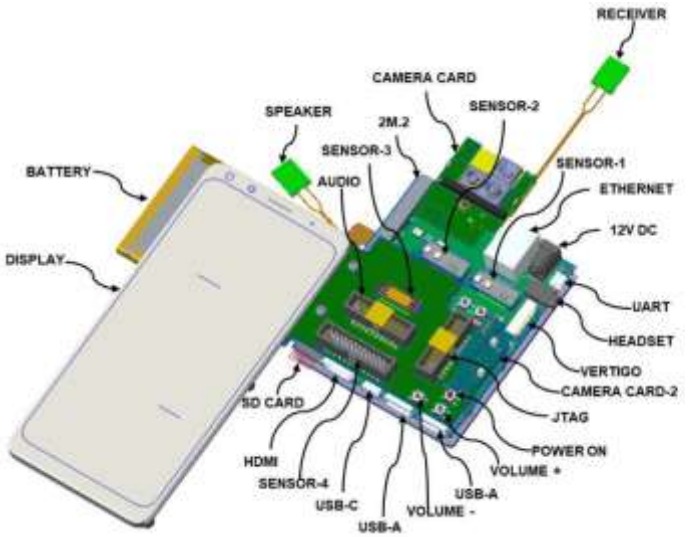


图 9-1

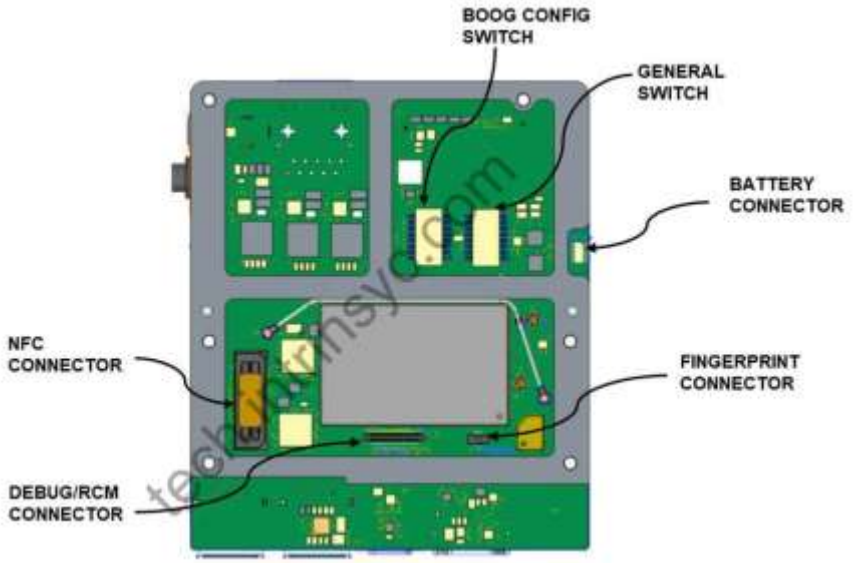


图 9-2

Qualcomm® Snapdragon™ 888 Mobile Hardware Development Kit

Qualcomm® Snapdragon™ 888 Mobile Hardware Development Kit

Comprehensive and expandable development and evaluation kit for the Snapdragon 888 Mobile Platform.

The Snapdragon 888 Mobile Hardware Development Kit provides an open-frame solution for technology companies to integrate and innovate devices based on the Snapdragon 888 Mobile Platform.

The Snapdragon 888 Mobile Hardware Development Kit is a feature-rich Android development platform that is designed to provide an ideal starting point for creating high-performance mobile devices and applications based on the Snapdragon 888 Mobile Platform. The kit includes the hardware, software tools and accessories needed to immediately begin your mobile development work.

With an advanced 5-nanometer design, Snapdragon 888 platform is engineered for innovative and intelligent on-device AI, gigapixel speed professional camera quality, desktop quality graphics, and Gigabit Class download speeds.

The Snapdragon 888 mobile development platform is designed to provide original equipment manufacturers (OEMs), hardware/software vendors, developers and engineers with next generation software technology and tools to accelerate development and testing of devices.

Solution Highlights

Kit Contents

- Single board computer (SBC) with Snapdragon 888
- 12V AC power adapter
- USB cable
- Setup guide

Display Expansion Card is an additional accessory

Development Platform

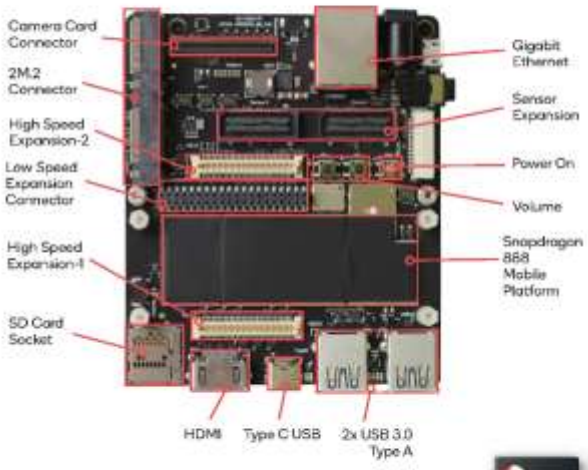


图 9-3

Snapdragon 888 Mobile Platform Applications

- Mobile PCs
- Apps Development
- IP Cameras
- Hexagon DSP
- Smart Phones/Tablets
- Artificial Intelligence

Snapdragon 888 Main Board

The Snapdragon 888 main board measuring 100mm x 85mm, is where all the processing occurs.

Display Expansion Board

The display expansion board includes a FHD+ AMOLED display with capacitive touch panel along with audio connectors, sensor GenX connector, Legacy sensor connector, 20p JTAG connector and front camera card connector.

To learn more visit:
developer.qualcomm.com

Qualcomm
snapdragon

Snapdragon 888 Specifications

Dimensions	100mm x 85mm (main board)
CPU	Qualcomm® Kryo™ 680 CPU
GPU	Qualcomm® Adreno™ 660 GPU delivers up to 35% faster graphics rendering*
DSP	Qualcomm® Hexagon™ 780 processor
Memory and Storage	12GB LPDDR5 PoP memory 256GB UFS 3.0
Connectivity	Wi-Fi: 802.11a/b/g/n/ac/ax 2.4/5GHz Bluetooth 5.1 + NFC card connector reserved
Camera Support	Qualcomm Spectra™ 580 image processing engine 6x MIPI CSI with support for 3-2 camera configuration
Display	2x MIPI dual 4-lane DSI + touch panel
Multimedia	HDMI 2.0 output - supports up to 4K UHD
I/O Interfaces	2M.2, HDMI, 1x USB 3.1 Type C, 2x USB 3.0 Type A, 1x USB 2.0 micro-B for UART, Gigabit Ethernet, 6x MIPI-CSI, 2x MIPI dual 4-lane DSI Expansion headers for additional features
Operating System	Android 11
Optional Accessories	Display: 6.65" AMOLED Display (2340 x 1080) with Touch Panel Camera Daughter Card: 16MP + 48MP + 13MP Rear Camera, TOF Sensor

*Compared to previous generation

Qualcomm Hexagon, Qualcomm Adreno, Qualcomm Spectra and Qualcomm Kryo are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

图 9-4

下面是一款高通的基于 UFS 4.0 的开发板介绍。

SNAPDRAGON® 8 GEN 2 MOBILE PLATFORM

The Snapdragon 8 Gen 2 Mobile Platform defines a new premium standard for connected computing. Intelligently engineered with groundbreaking AI across the board, this AI marvel enables truly extraordinary experiences.

Groundbreaking AI

Our fastest, most advanced Qualcomm® AI Engine ever puts a world of possibilities at your fingertips. Communicate with friends or colleagues across the globe with multi-language translation and transcription, or capture premium content with AI Cinematic video. And now, the Qualcomm® Sensing Hub features dual-AI processors for the first time, powering exciting new experiences like direct-to-app voice assistance for convenient control of your favorite apps.

- Up to 4.35x faster AI performance[1]
- First Snapdragon® Mobile Platform with INT4 precision support for 60% better performance/watt
- First Snapdragon Mobile Platform with Micro Tile Inferencing

Ingenious capture

Break photography barriers with the Qualcomm Spectra™ 18-bit triple Cognitive ISP—our first ever AI-powered camera processor. This new architecture powers real-time Semantic Segmentation, to recognize and optimize each aspect within a frame—like faces, hair, clothes, backgrounds, and more.

- First Snapdragon Mobile Platform with AI-powered Always-Sensing Camera
- Up to 200 MP photo capture
- 8K HDR video capture in 10-bit HDR

Next-level audio experiences

Be transported by Snapdragon Sound™ Technology, now featuring spatial audio with head-tracking[2] for complete surround-sound immersion. Now you're always at the center of sound—no matter how you move you'll always be captivated by the audio that's all around you.

- Hear every detail of your music with lossless music streaming
- Ultra-low latency Bluetooth® streaming <48ms for lag-free gaming

Champion-level gameplay

Accelerate your wins with the full arsenal of Snapdragon Elite Gaming™ Features. Best-in-class lighting, shadow, and illumination effects add astonishing authenticity to scenes, brought to you by the industry standard for real-time Hardware-Accelerated Ray Tracing. And, support for Unreal Engine 5 and Metahuman framework produces photorealistic human characters for unbeatable immersion. When combined with accelerated performance from our Qualcomm® Adreno™ GPU and Qualcomm® Kryo™ CPU, worlds of gaming will unfold and astound in real-time.

- First mobile platform to support Vulkan 1.3 APIs
- Adreno GPU delivers up to 25% faster performance, with up to 45% better power efficiency
- Kryo CPU improves performance up to 35%, while new micro-architecture allows up to 40% more power efficiency

Unparalleled connectivity

Featuring the Snapdragon X70 5G Modem-RF System, Snapdragon 8 Gen 2 is the world's first and only mobile platform with a dedicated 5G AI processor. Plus, gaming, streaming, and communication from home soar via Wi-Fi 7 (the industry's lowest latency offering), all brought to you by the Qualcomm® FastConnect™ 7800 Mobile Connectivity System.

- 5G Dual-SIM Dual-Active (DSDA) enables the simultaneous use of two 5G+5G or 5G+4G SIM cards for ultimate user flexibility
- Blazing Wi-Fi speeds of up to 5.8 Gbps—more than double Wi-Fi 6
- World's first commercial Wi-Fi 7 SoC, with advanced High Band Simultaneous Multi-Link

Vault-like security

Snapdragon Secure offers the latest in isolation, cryptography, key management, attestation, and more—all intricately designed to protect your data and privacy.

- Enhanced Face Unlock security system including liveness detection

1. On select models (Android 13+)
2. When used with certain compatible Snapdragon Sound products

图 9-5

SPECIFICATIONS & FEATURES

Artificial Intelligence

Qualcomm® Adreno™ GPU

Qualcomm® Kryo™ CPU

Qualcomm® Hexagon™ Processor
- Fused AI Accelerator Architecture
- Hexagon Tensor Accelerator
- Hexagon Vector eXtensions
- Hexagon Scalar Accelerator
- Hexagon Direct Link
- Support for mix precision (INT8+INT16)
- Support for all precisions (INT4, INT8, INT16, FP16)
- Micro Tile Inferencing

Qualcomm® Sensing Hub
- Dual AI Processors for audio and sensors
- Always-Sensing camera

5G Modem-RF System

Snapdragon® X70 5G Modem-RF System
- 5G mmWave and sub-6 GHz, standalone (SA) and non-standalone (NSA) modes, standalone mmWave and mmWave-sub6 dual connectivity, FDD, TDD
- mmWave: 8 carriers, 2x2 MIMO
- Sub-6 GHz: 4x4 MIMO
- Qualcomm® 5G AI Suite
- Qualcomm® AI-Enhanced Signal Boost
- Qualcomm® 5G PowerSave Gen 3
- Qualcomm® Smart Transmit™ 3.0 technology
- Qualcomm® Wideband Envelope Tracking
- Qualcomm® 5G Ultra-Low Latency Suite
- Global 5G multi-SIM, including 5G-5G/4G Dual-SIM Dual-Active (DSDA)

Downlink: Up to 10 Gbps

Uplink: Up to 3.5 Gbps

Multimode support: 5G NR, NR-DC, EN-DC, LTE, CBRS, WCDMA, HSPA, TD-SCDMA, CDMA 1x, EV-DO, GSM/EDGE

Wi-Fi & Bluetooth[1]

Qualcomm® FastConnect™ 7800 System
- Wi-Fi Generation: Wi-Fi 7
- Peak speed: 5.8 Gbps
- 802.11be, 802.11ax, 802.11ac, 802.11a/b/g/n
- Wi-Fi Spectral Bands: 5 GHz, 6 GHz
- Channel Bandwidth: 20/40/80/160 MHz
- 8-stream sounding (for 8x8 MU-MIMO)
- MIMO Configuration: 2x2 (2-stream)
- MU-MIMO (Uplink & Downlink)
- 4K QAM
- OFDMA (Uplink & Downlink)
- Dual-band simultaneous (2x2 + 2x2)
- Wi-Fi Security: WPA3-Enterprise, WPA3-Enhanced Open, WPA3 Easy Connect, WPA3-Personal

Integrated Bluetooth
- Bluetooth audio: Snapdragon Sound™ Technology with support for Qualcomm® aptX™ Voice, aptX Lossless, aptX Adaptive, and LE Audio
- Bluetooth features: Bluetooth 5.3, LE Audio, Dual Bluetooth antennas

Camera

Qualcomm Spectra™ Image Signal Processor
- Cognitive ISP: Triple 18-bit ISPs
- Up to 36 MP triple camera @ 30 FPS with Zero Shutter Lag
- Up to 64+36 MP dual camera @ 30 FPS with Zero Shutter Lag
- Up to 108 MP single camera @ 30 FPS with Zero Shutter Lag
- Up to 200 Megapixel Photo Capture

AI-based face detection, auto-focus, and auto-exposure

Rec. 2020 color gamut photo and video capture

Up to 10-bit color depth photo and video capture

8K HDR Video Capture + 64 MP Photo Capture

10-bit HEIF, HEIC photo capture, HEVC video capture

Video Capture HDR Formats: HDR10+, HDR10, HLG, Dolby Vision

8K HDR Video Capture @ 30 FPS

4K Video Capture @ 120 FPS

Slow-mo video capture at 720p @ 960 FPS

8K Video Playback @ 60 FPS

Engine for Visual Analytics 3.0

Real-time Semantic Segmentation photo and video processing

Bokeh Engine 2 for Video Capture

Pro Sight Video Capture

Computational HDR Video Capture: Up to 4 exposures (with QDOL image sensors)

Video super resolution

Multi-frame Noise Reduction (MFNR)

Locally Motion Compensated Temporal Filtering

Multi-Frame and triple exposure staggered/digital overlap HDR dual-sensor support

Audio

Qualcomm Aqstic™ audio codec

Qualcomm Aqstic smart speaker amplifier

Total Harmonic Distortion + Noise (THD+N), Playback: -108dB

Qualcomm Audio and Voice Communication Suite

Spatial Audio with Head Tracking[2]

Display

On-Device Display Support:
- 4K @ 60 Hz
- QHD+ @ 144 Hz

Maximum External Display Support:
- Up to 4K @ 60 Hz
- 10-bit color depth, Rec. 2020 color gamut
- HDR10, HDR10+, HDR vivid, and Dolby Vision

Demura and subpixel rendering for OLED Uniformity

OLED Aging Compensation

CPU

Kryo CPU
- 64-bit Architecture
- 1 Prime core, up to 3.2 GHz[3]
  - Arm Cortex-X3 technology
- 4 Performance cores, up to 2.8 GHz
- 3 Efficiency cores, up to 2.0 GHz

Visual Subsystem

Adreno CPU
- Real-time Hardware Accelerated Ray Tracing
- Snapdragon Game Post Processing Accelerator
- HDR gaming (10-bit color depth, Rec. 2020 color gamut)
- Snapdragon Shadow Denoiser
- API Support: OpenGL® ES 3.2, OpenCL™ 2.0 FP, Vulkan® 1.3
- Hardware-accelerated H.265 and VP9 decoder
- HDR Playback Codec support for HDR10+, HDR10, HLG, and Dolby Vision

Security

Platform Security Foundations, Trusted Execution Environment & Services, Secure Processing Unit (SPU)

Trust Management Engine

Qualcomm® wireless edge services (WES) and premium security features

Qualcomm® 3D Sonic Sensor and Qualcomm 3D Sonic Max (fingerprint sensor)

Qualcomm® Type-C® Hyperwise

Location

Concurrent: GPS, Glonass, BeiDou, Galileo, QZSS, NavIC

Triple frequency GNSS

Sensor-Assisted Positioning 6.0
- With global map aiding
- Urban pedestrian navigation with sidewalk accuracy
- Global freeway lane-level vehicle navigation

Charging

Qualcomm® Quick Charge™ 5 Technology

Memory

Support for LP-DDR5x memory up to 4200 MHz

Memory Density: up to 16 GB

General Specifications

Full suite of Snapdragon Elite Gaming™ Features

4 nm Process Technology

USB Version 3.1, USB Type-C Support

Part Number: SM8550-AB

Storage: UFS 4.0

Snapdragon

1. When used with certain compatible Snapdragon Sound smartphones.
2. Based on Snapdragon Sound Spatial Audio with Head Tracking.

图 9-6

10. 附录 A：PCIe 和 NVMe 协议基础知识

10.1 PCIe, NVMe，CXL，DDR，UFS 和 NAND 协议 Wiki

10.1.1 PCIe 协议 Wiki

PCI Express

PCI Express (**Peripheral Component Interconnect Express**), officially abbreviated as **PCIe** or **PCI-e**,[1] is a high-speed serial computer expansion bus standard, designed to replace the older PCI, PCI-X and AGP bus standards. It is the common motherboard interface for personal computers' graphics cards, hard disk drive host adapters, SSDs, Wi-Fi and Ethernet hardware connections.[2] PCIe has numerous improvements over the older standards, including higher maximum system bus throughput, lower I/O pin count and smaller physical footprint, better performance scaling for bus devices, a more detailed error detection and reporting mechanism (Advanced Error Reporting, AER),[3] and native hot-swap functionality. More recent revisions of the PCIe standard provide hardware support for I/O virtualization.

The PCI Express electrical interface is measured by the number of simultaneous lanes.[4] (A lane is a single send/receive line of data. The analogy is a highway with traffic in both directions.) The interface is also used in a variety of other standards — most notably the laptop expansion card interface called ExpressCard. It is also used in the storage interfaces of SATA Express, U.2 (SFF- 8639) and M.2.

Format specifications are maintained and developed by the PCI-SIG (PCI Special Interest Group) — a group of more than 900 companies that also maintains the conventional PCI specifications.

PCI Express

Peripheral Component Interconnect Express	
PCI Express logo	
Year created	2003
Created by	Intel · Dell · HP · IBM
Supersedes	PCI · PCI-X · AGP
Width in bits	1 per lane (up to 16 lanes)
No. of devices	1 on each endpoint of each connection.[a]
Speed	Dual simplex; examples in single- lane (x1) and 16-lane (x16): **Version 1.x:** 2.5 GT/s **x1:** 250 MB/s **x16:** 4 GB/s **Version 2.x:** 5 GT/s **x1:** 500 MB/s **x16:** 8 GB/s **Version:3.x:** 8 GT/s **x1:** 985 MB/s **x16:** 15.75 GB/s **Version 4.0:** 16 GT/s **x1:** 1.97 GB/s **x16:** 31.5 GB/s **Version 5.0:** 32 GT/s **x1:** 3.94 GB/s **x16:** 63 GB/s **Version 6.0:** 64 GT/s **x1:** 7.56 GB/s **x16:** 121 GB/s

Architecture

Conceptually, the PCI Express bus is a high-speed serial replacement of the older PCI/PCI-X bus.[7] One of the key differences between the PCI Express bus and the older PCI is the bus topology; PCI uses a shared parallel bus architecture, in which the PCI host and all devices share a common set of address, data, and control lines. In contrast, PCI Express is based on point-to-point topology, with separate serial links connecting every device to the root complex (host). Because of its shared bus topology, access to the older PCI bus is arbitrated (in the case of multiple masters), and limited to one master at a time, in a single direction. Furthermore, the older PCI clocking scheme limits the bus clock to the slowest peripheral on the bus (regardless of the devices involved in the bus transaction). In contrast, a PCI Express bus link supports full-duplex communication between any two endpoints, with no inherent limitation on concurrent access across multiple endpoints.

In terms of bus protocol, PCI Express communication is encapsulated in packets. The work of packetizing and de-packetizing data and status-message traffic is handled by the transaction layer of the PCI Express port (described later). Radical differences in electrical signaling and bus protocol require the use of a different mechanical form factor and expansion connectors (and thus, new motherboards and new adapter boards); PCI slots and PCI Express slots are not interchangeable. At the software level, PCI Express preserves backward compatibility with PCI; legacy PCI system software can detect and configure newer PCI Express devices without explicit support for the PCI Express standard, though new PCI Express features are inaccessible.

The PCI Express link between two devices can vary in size from one to 16 lanes. In a multi-lane link, the packet data is striped across lanes, and peak data throughput scales with the overall link width. The lane count is automatically negotiated during device initialization and can be restricted by either endpoint. For example, a single-lane PCI Express (x1) card can be inserted

into a multi-lane slot (x4, x8, etc.), and the initialization cycle auto-negotiates the highest mutually supported lane count. The link can dynamically down-configure itself to use fewer lanes, providing a failure tolerance in case bad or unreliable lanes are present. The PCI Express standard defines link widths of x1, x2, x4,x8, and x16. Up to and including PCIe 5.0, x12, and x32 links were defined as well but never used.[8] This allows the PCI Express bus to serve both cost-sensitive applications where high throughput is not needed, and performance-critical applications such as 3D graphics, networking (10 Gigabit Ethernet or multiport Gigabit Ethernet), and enterprise storage (SAS or Fibre Channel). Slots and connectors are only defined for a subset of these widths, with link widths in between using the next larger physical slot size.

As a point of reference, a PCI-X (133 MHz 64-bit) device and a PCI Express 1.0 device using four lanes (x4) have roughly the same peak single-direction transfer rate of 1064 MB/s. The PCI Express bus has the potential to perform better than the PCI-X bus in cases where multiple devices are transferring data simultaneously, or if communication with the PCI Express peripheral is bidirectional.

	Version 7.0: 128 GT/s **x1:** 15.13 GB/s **x16:** 242 GB/s
Style	Serial
Hotplugging interface	Yes (with ExpressCard, OCuLink, CFexpress or U.2)
External interface	Yes (with OCuLink or PCI Express External Cabling)
Website	pcisig.com(https://pcisig.com/)

Interconnect

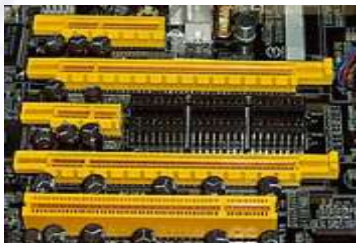
PCI Express devices communicate via a logical connection called an *interconnect*[9] or *link*. A link is a point-to-point communication channel between two PCI Express ports allowing both of them to send and receive ordinary PCI requests (configuration, I/O or memory read/write) and interrupts (INTx, MSI or MSI-X). At the physical level, a link is composed of one or more *lanes*.[9] Low-speed peripherals (such as an 802.11 Wi-Fi card) use a single-lane (x1) link, while a graphics adapter typically uses a much wider and therefore faster 16-lane (x16) link.

two types of PCIe slot on an Asus H81M-K motherboard

Lane

A lane is composed of two differential signaling pairs, with one pair for receiving data and the other for transmitting. Thus, each lane is composed of four wires or signal traces. Conceptually, each lane is used as a full-duplex byte stream, transporting data packets in eight-bit "byte" format simultaneously in both directions between endpoints of a link.[10] Physical PCI Express links may contain 1, 4, 8 or 16 lanes.[11][5]: 4,5 [9] Lane counts are written with an "x" prefix (for example, "x8" represents an eight-lane card or slot), with x16 being the largest size in common use.[12] Lane sizes are also referred to via the terms "width" or "by" e.g., an eight-lane slot could be referred to as a "by 8" or as "8 lanes wide."

For mechanical card sizes, see below.

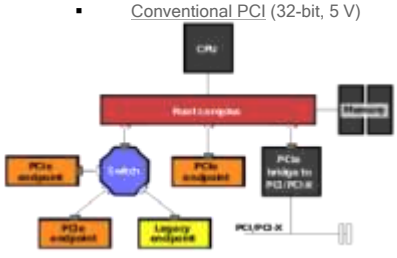


Various slots on a computer motherboard, from top to bottom:

- **PCI Express x4**
- **PCI Express x16**
- **PCI Express x1**
- **PCI Express X16**
- Conventional PCI (32-bit, 5 V)

Serial bus

The bonded serial bus architecture was chosen over the traditional parallel bus because of the inherent limitations of the latter, including half-duplex operation, excess signal count, and inherently lower bandwidth due to timing skew. Timing skew results from separate electrical signals within a parallel interface traveling through conductors of different lengths, on potentially different printed circuit board (PCB) layers, and at possibly different signal velocities. Despite being transmitted simultaneously as a single word, signals on a parallel interface have different travel duration and arrive at their destinations at different times. When the interface clock period is shorter than the largest time difference between signal arrivals, recovery of the transmitted word is no longer possible. Since timing skew over a parallel bus can amount to a few nanoseconds, the resulting bandwidth limitation is in the range of hundreds of megahertz.



Example of the PCI Express topology: white "junction boxes" represent PCI Express device downstream ports, while the gray ones represent upstream ports.[5]: 7

A serial interface does not exhibit timing skew because there is only one differential signal in each direction within each lane, and there is no external clock signal since clocking information is embedded within the serial signal itself. As such, typical bandwidth limitations on serial signals are in the multi-gigahertz range. PCI Express is one example of the general trend toward replacing parallel buses with serial interconnects; other examples include Serial ATA (SATA), USB, Serial Attached SCSI (SAS), FireWire (IEEE 1394), and RapidIO. In digital video, examples in common use are DVI, HDMI, and DisplayPort.

Multichannel serial design increases flexibility with its ability to allocate fewer lanes for slower devices.

PCI Express x1 card containing a PCI Express switch (covered by a small heat sink), which creates multiple endpoints out of one endpoint and lets multiple devices share it

Form factors

PCI Express (standard)

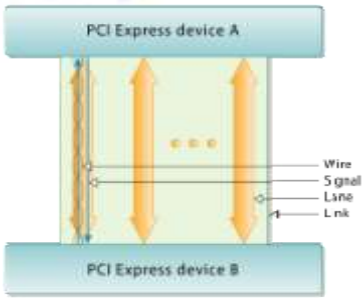
A PCI Express card fits into a slot of its physical size or larger (with x16 as the largest used), but may not fit into a smaller PCI Express slot; for example, a x16 card may not fit into a x4 or x8 slot. Some slots use open-ended sockets to permit physically longer cards and negotiate the best available electrical and logical connection.

The number of lanes actually connected to a slot may also be fewer than the number supported by the physical slot size. An example is a x16 slot that runs at x4, which accepts any x1, x2, x4, x8 or x16 card, but provides only four lanes. Its specification may read as "x16 (x4 mode)", while "mechanical @ electrical" notation (e.g. "x16 @ x4") is also common. The advantage is that such slots can accommodate a larger range of PCI Express cards without requiring motherboard hardware to support the full transfer rate. Standard mechanical sizes are x1, x4, x8, and x16. Cards with a differing number of lanes need to use the next larger mechanical size (i.e a x2 card uses the x4 size, or a x12 card uses the x16 size).

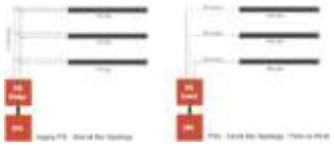
The cards themselves are designed and manufactured in various sizes. For example, solid-state drives (SSDs) that come in the form of PCI Express cards often use HHHL (half height, half length) and FHHL (full height, half length) to describe the physical dimensions of the card.[14][15]

PCI card type	Dimensions height × length × width, maximum	
	(mm)	(in)
Full-Length	111.15 × 312.00 × 20.32	4.376 × 12.283 × 0.8
Half-Length	111.15 × 167.65 × 20.32	4.376 × 6.600 × 0.8
Low-Profile/Slim	68.90 × 167.65 × 20.32	2.731 × 6.600 × 0.8

The PCIe slots on a motherboard are often labeled with the number of PCIe lanes they have. Sometimes what may seem like a large slot may only have a few lanes. For instance, an x16 slot with only 4 PCIe lanes is quite common.[6]



A PCI Express link between two devices consists of one or more lanes, which are dual simplex channels using two differential signaling pairs.[5]: 3



Highly simplified topologies of the Legacy PCI Shared (Parallel) Interface and the PCIe Serial Point- to-Point Interface[13]

Modern (since c.2012[16]) gaming video cards usually exceed the height as well as thickness specified in the PCI Express standard, due to the need for more capable and quieter cooling fans, as gaming video cards often emit hundreds of watts of heat.[17] Modern computer cases are often wider to accommodate these taller cards, but not always. Since full-length cards (312 mm) are uncommon, modern cases sometimes cannot fit those. The thickness of these cards also typically occupies the space of 2 PCIe slots. In fact, even the methodology of how to measure the cards varies between vendors, with some including the metal bracket size in dimensions and others not.


Intel P3608 NVMe flash SSD, PCI-E add-in card

For instance, a 2020 Sapphire card measures 135 mm in height (excluding the metal bracket), which exceeds the PCIe standard height by 28 mm.[18] Another card by XFX measures 55 mm thick (i.e. 2.7 PCI slots at 20.32 mm), taking up 3 PCIe slots.[19] The Asus GeForce RTX 3080 10 GB STRIX GAMING OC video card is a two-slot card that has dimensions of 318.5 mm × 140.1 mm × 57.8 mm, exceeding PCI Express' maximum length, height, and thickness respectively.[20]

Pinout

The following table identifies the conductors on each side of the edge connector on a PCI Express card. The solder side of the printed circuit board (PCB) is the A-side, and the component side is the B-side.[21] PRSNT1# and PRSNT2# pins must be slightly shorter than the rest, to ensure that a hot-plugged card is fully inserted. The WAKE# pin uses full voltage to wake the computer, but must be pulled high from the standby power to indicate that the card is wake capable.[22]

PCI Express connector pinout (x1, x4, x8 and x16 variants)

Pin	Side B	Side A	Description		Pin	Side B	Side A	Description
1	+12 V	PRSNT1#	Must connect to farthest PRSNT2# pin		50	HSOp(8)	Reserved	Lane 8 transmit data, + and −
2	+12 V	+12 V	Main power pins		51	HSOn(8)	Ground	
3	+12 V	+12 V			52	Ground	HSIp(8)	Lane 8 receive data, + and −
4	Ground	Ground			53	Ground	HSIn(8)	
5	SMCLK	TCK	SMBus and JTAG port pins		54	HSOp(9)	Ground	Lane 9 transmit data, + and −
6	SMDAT	TDI			55	HSOn(9)	Ground	
7	Ground	TDO			56	Ground	HSIp(9)	Lane 9 receive data, + and −
8	+3.3 V	TMS			57	Ground	HSIn(9)	
9	TRST#	+3.3 V			58	HSOp(10)	Ground	Lane 10 transmit data, + and −
10	+3.3 V aux	+3.3 V	Aux power & Standby power		59	HSOn(10)	Ground	
11	WAKE#	PERST#	Link reactivation; fundamental reset [23]		60	Ground	HSIp(10)	Lane 10 receive data, + and −
Key notch					61	Ground	HSIn(10)	
12	CLKREQ#[24]	Ground	Clock Request Signal		62	HSOp(11)	Ground	Lane 11 transmit data, + and −

13	Ground	REFCLK+	Reference clock differential pair	63	HSOn(11)	Ground	
14	HSOp(0)	REFCLK−	Lane 0 transmit data, + and −	64	Ground	HSIp(11)	Lane 11 receive data, + and −
15	HSOn(0)	Ground		65	Ground	HSIn(11)	
16	Ground	HSIp(0)	Lane 0 receive data, + and −	66	HSOp(12)	Ground	Lane 12 transmit data, + and −
17	PRSNT2#	HSIn(0)		67	HSOn(12)	Ground	
18	Ground	Ground		68	Ground	HSIp(12)	Lane 12 receive data, + and −
PCI Express x1 cards end at pin 18				69	Ground	HSIn(12)	
19	HSOp(1)	Reserved	Lane 1 transmit data, + and −	70	HSOp(13)	Ground	Lane 13 transmit data, + and −
20	HSOn(1)	Ground		71	HSOn(13)	Ground	
21	Ground	HSIp(1)	Lane 1 receive data, + and −	72	Ground	HSIp(13)	Lane 13 receive data, + and −
22	Ground	HSIn(1)		73	Ground	HSIn(13)	
23	HSOp(2)	Ground	Lane 2 transmit data, + and −	74	HSOp(14)	Ground	Lane 14 transmit data, + and −
24	HSOn(2)	Ground		75	HSOn(14)	Ground	
25	Ground	HSIp(2)	Lane 2 receive data, + and −	76	Ground	HSIp(14)	Lane 14 receive data, + and −
26	Ground	HSIn(2)		77	Ground	HSIn(14)	
27	HSOp(3)	Ground	Lane 3 transmit data, + and −	78	HSOp(15)	Ground	Lane 15 transmit data, + and −
28	HSOn(3)	Ground		79	HSOn(15)	Ground	
29	Ground	HSIp(3)	Lane 3 receive data, + and −	80	Ground	HSIp(15)	Lane 15 receive data, + and −
30	PWRBRK#[25]	HSIn(3)		81	PRSNT2#	HSIn(15)	
31	PRSNT2#	Ground		82	Reserved	Ground	
32	Ground	Reserved					
PCI Express x4 cards end at pin 32							
33	HSOp(4)	Reserved	Lane 4 transmit data, + and −				
34	HSOn(4)	Ground					
35	Ground	HSIp(4)	Lane 4 receive data, + and −				
36	Ground	HSIn(4)					
37	HSOp(5)	Ground	Lane 5 transmit data, + and −				
38	HSOn(5)	Ground					
39	Ground	HSIp(5)	Lane 5 receive data, + and −				
40	Ground	HSIn(5)					
41	HSOp(6)	Ground	Lane 6 transmit data, + and −				
42	HSOn(6)	Ground					
43	Ground	HSIp(6)	Lane 6 receive data, + and −				
44	Ground	HSIn(6)					
45	HSOp(7)	Ground	Lane 7 transmit data, + and −				
46	HSOn(7)	Ground					
47	Ground	HSIp(7)	Lane 7 receive data, + and −				

Legend	
Ground pin	Zero volt reference
Power pin	Supplies power to the PCIe card
Card-to-host pin	Signal from the card to the motherboard
Host-to-card pin	Signal from the motherboard to the card

48	PRSNT2#	HSIn(7)		Open drain	May be pulled low or sensed by multiple cards
49	Ground	Ground		**Sense pin**	Tied together on card
PCI Express x8 cards end at pin 49				**Reserved**	Not presently used, do not connect

All PCI express cards may consume up to 3 A at +3.3 V (9.9 W). The amount of +12 V and total power they may consume depends on the form factor and the role of the card:[26]: 35–36 [27][28]

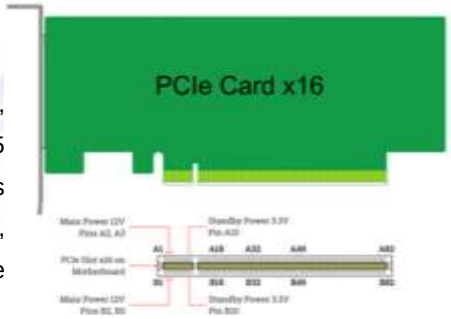
8-pin (left) and 6-pin (right) power connectors used on PCI Express card

- **x1 cards are limited to 0.5 A at +12 V (6 W) and 10 W combined.**
- **x4 and wider cards are limited to 2.1 A at +12 V (25 W) and 25 W combined.**
- **A full-sized x1 card may draw up to the 25 W limits after initialization and software configuration as a high-power device.**
- **A full-sized x16 graphics card may draw up to 5.5 A at +12 V (66 W) and 75 W combined after initialization and software configuration as a high-power device.[22]: 38–39**

Optional connectors add 75 W (6-pin) or 150 W (8-pin) of +12 V power for up to 300 W total (2 × 75 W + 1 × 150 W).

- **Sense0 pin is connected to ground by the cable or power supply, or float on board if cable is not connected.**
- **Sense1 pin is connected to ground by the cable or power supply, or float on board if cable is not connected.**

Some cards use two 8-pin connectors, but this has not been standardized yet as of 2018, therefore such cards must not carry the official PCI Express logo. This configuration allows 375 W total (1 × 75 W + 2 × 150 W) and will likely be standardized by PCI-SIG with the PCI Express 4.0 standard. The 8-pin PCI Express connector could be confused with the EPS12V connector, which is mainly used for powering SMP and multi-core systems. The power connectors are variants of the Molex Mini-Fit Jr. series connectors. [30]



The main 12 V power supply for the PCIe slot is pins B2, B3 (side B) and pins A2, A3 (side A). Power standby 3.3 V is pin B10 and A10. PCIe x1 cards can receive up to 25 W and x16 graphics cards can receive up to 75 W, combined.[29]

Molex Mini-Fit Jr. part numbers[30]

Pins	Female/receptacle on PS cable	Male/right-angle header on PCB
6-pin	45559-0002	45558-0003
8-pin	45587-0004	45586-0005, 45586-0006

6-pin power connector (75 W)[31]		8-pin power connector (150 W)[32][33][34]	
Pin	Description	Pin	Description
1	+12 V	1	+12 V
2	Not connected (usually +12 V as well)	2	+12 V
3	+12 V	3	+12 V
		4	Sense1 (8-pin connected[A])
4	Ground	5	Ground
5	Sense	6	Sense0 (6-pin or 8-pin connected)
6	Ground	7	Ground
		8	Ground

6 pin power connector pin map

8 pin power connector pin map

A. When a 6-pin connector is plugged into an 8-pin receptacle the card is notified by a missing *Sense1* that it may only use up to 75 W.

PCI Express Mini Card

PCI Express Mini Card (also known as **Mini PCI Express**, **Mini PCIe**, **Mini PCI-E**, **mPCIe**, and **PEM**), based on PCI Express, is a replacement for the Mini PCI form factor. It is developed by the PCI-SIG. The host device supports both PCI Express and USB 2.0 connectivity, and each card may use either standard. Most laptop computers built after 2005 use PCI Express for expansion cards; however, as of 2015, many vendors are moving toward using the newer M.2 form factor for this purpose.

A WLAN PCI Express Mini Card and its connector

Due to different dimensions, PCI Express Mini Cards are not physically compatible with standard full- size PCI Express slots; however, passive adapters exist that let them be used in full-size slots.[35]

Physical dimensions

Dimensions of PCI Express Mini Cards are 30 mm × 50.95 mm (width × length) for a Full Mini Card. There is a 52-pin edge connector, consisting of two staggered rows on a 0.8 mm pitch. Each row has eight contacts, a gap equivalent to four contacts, then a further 18 contacts. Boards have a thickness of

1.0 mm, excluding the components. A "Half Mini Card" (sometimes abbreviated as HMC) is also specified, having approximately half the physical length of 26.8 mm.

MiniPCI and MiniPCI Express cards in comparison

Electrical interface

PCI Express Mini Card edge connectors provide multiple connections and buses:

- **PCI Express x1 (with SMBus)**
- **USB 2.0**
- **Wires to diagnostics LEDs for wireless network (i.e., Wi-Fi) status on computer's chassis**
- **SIM card for GSM and WCDMA applications (UIM signals on spec.)**
- **Future extension for another PCIe lane**
- **1.5 V and 3.3 V power**

Mini-SATA (mSATA) variant

Despite sharing the Mini PCI Express form factor, an mSATA slot is not necessarily electrically compatible with Mini PCI Express. For this reason, only certain notebooks are compatible with mSATA drives. Most compatible systems are based on Intel's Sandy Bridge processor architecture, using the Huron River platform. Notebooks such as Lenovo's ThinkPad T, W and X series, released in March–April 2011, have support for an mSATA SSD card in their WWAN card slot. The ThinkPad Edge E220s/E420s, and the Lenovo IdeaPad Y460/Y560/Y570/Y580 also support mSATA.[36] On the contrary, the L-series among others can only support M.2 cards using the PCIe standard in the WWAN slot.

An Intel mSATA SSD

Some notebooks (notably the Asus Eee PC, the Apple MacBook Air, and the Dell mini9 and mini10) use a variant of the PCI Express Mini Card as an SSD. This variant uses the reserved and several non- reserved pins to implement SATA and IDE interface passthrough, keeping only USB, ground lines, and sometimes the core PCIe x1 bus intact.[37] This makes the "miniPCIe" flash and solid-state drives sold for netbooks largely incompatible with true PCI Express Mini implementations.

Also, the typical Asus miniPCIe SSD is 71 mm long, causing the Dell 51 mm model to often be (incorrectly) referred to as half length. A true 51 mm Mini PCIe SSD was announced in 2009, with two stacked PCB layers that allow for higher storage capacity. The announced design preserves the PCIe interface, making it compatible with the standard mini PCIe slot. No working product has yet been developed.

Intel has numerous desktop boards with the PCIe x1 Mini-Card slot that typically do not support mSATA SSD. A list of desktop

boards that natively support mSATA in the PCIe x1 Mini-Card slot (typically multiplexed with a SATA port) is provided on the Intel Support site.[38]

PCI Express M.2

M.2 replaces the mSATA standard and Mini PCIe.[39] Computer bus interfaces provided through the M.2 connector are PCI Express 3.0 (up to four lanes), Serial ATA 3.0, and USB 3.0 (a single logical port for each of the latter two). It is up to the manufacturer of the M.2 host or device to choose which interfaces to support, depending on the desired level of host support and device type.

PCI Express External Cabling

PCI Express External Cabling (also known as *External PCI Express*, *Cabled PCI Express*, or *ePCIe*) specifications were released by the PCI- SIG in February 2007.[40][41]

Standard cables and connectors have been defined for x1, x4, x8, and x16 link widths, with a transfer rate of 250 MB/s per lane. The PCI-SIG also expects the norm to evolve to reach 500 MB/s, as in PCI Express 2.0. An example of the uses of Cabled PCI Express is a metal enclosure, containing a number of PCIe slots and PCIe-to-ePCIe adapter circuitry. This device would not be possible had it not been for the ePCIe specification.

PCI Express OCuLink

OCuLink (standing for "optical-copper link", since *Cu* is the chemical symbol for Copper) is an extension for the "cable version of PCI Express", acting as a competitor to version 3 of the Thunderbolt interface. Version 1.0 of OCuLink, released in Oct 2015, supports up to PCIe

3.0 x4 lanes (8 GT/s (gigatransfers per second), 3.9 GB/s) over copper cabling; a fiber optic version may appear in the future.

In its latest version OCuLink-2, it supports up to 16 GB/s (PCIe 4.0 x8)[42] while the maximum bandwidth of a full speed Thunderbolt 4 cable is 5 GB/s. Some suppliers may design their connector product to be able to support next generation PCI Express 5.0 running at 32 GT/s per lane for future proofing and minimizing development costs over the next few years.[42] Initially, PCI-SIG expected to bring OCuLink into laptops for connection of powerful external GPU boxes. It turned out to be a rare use. Instead, OCuLink became popular for PCIe interconnections in servers.[43]

Derivative forms

Numerous other form factors use, or are able to use, PCIe. These include:

- Low-height card
- ExpressCard: Successor to the PC Card form factor (with x1 PCIe and USB 2.0; hot-pluggable)
- PCI Express ExpressModule: A hot-pluggable modular form factor defined for servers and workstations XQD card: A PCI Express-based flash card standard by the CompactFlash Association with x2 PCIe
- CFexpress card: A PCI Express-based flash card by the CompactFlash Association in three form factors supporting 1 to 4 PCIe lanes
- SD card: The SD Express bus, introduced in version 7.0 of the SD specification uses a x1 PCIe link XMC: Similar to the CMC/PMC form factor (VITA 42.3)
- AdvancedTCA: A complement to CompactPCI for larger applications; supports serial based backplane topologies

- AMC: A complement to the AdvancedTCA specification; supports processor and I/O modules on ATCA boards (x1, x2, x4 or x8 PCIe).

- FeaturePak: A tiny expansion card format (43 mm × 65 mm) for embedded and small-form-factor applications, which implements two x1 PCIe links on a high-density connector along with USB, I2C, and up to 100 points of I/O

- Universal IO: A variant from Super Micro Computer Inc designed for use in low-profile rack-mounted chassis.[44] It has the connector bracket reversed so it cannot fit in a normal PCI Express socket, but it is pin-compatible and may be inserted if the bracket is removed.

- M.2 (formerly known as NGFF)

- M-PCIe brings PCIe 3.0 to mobile devices (such as tablets and smartphones), over the M-PHY physical layer.[45][46]

- U.2 (formerly known as SFF-8639)

The PCIe slot connector can also carry protocols other than PCIe. Some 9xx series Intel chipsets support Serial Digital Video Out, a proprietary technology that uses a slot to transmit video signals from the host CPU's integrated graphics instead of PCIe, using a supported add-in.

The PCIe transaction-layer protocol can also be used over some other interconnects, which are not electrically PCIe:

- Thunderbolt: A royalty-free interconnect standard by Intel that combines DisplayPort and PCIe protocols in a form factor compatible with Mini DisplayPort. Thunderbolt 3.0 also combines USB 3.1 and uses the USB-C form factor as opposed to Mini DisplayPort.

- USB4

History and revisions

While in early development, PCIe was initially referred to as *HSI* (for *High Speed Interconnect*), and underwent a name change to *3GIO* (for *3rd Generation I/O*) before finally settling on its PCI-SIG name *PCI Express*. A technical working group named the *Arapaho Work Group* (AWG) drew up the standard. For initial drafts, the AWG consisted only of Intel engineers; subsequently, the AWG expanded to include industry partners.

Since, PCIe has undergone several large and smaller revisions, improving on performance and other features.

PCI Express link performance[47][48]

Version	Intro-duced	Line code		Transfer rate per lane[i][ii]	Throughput[i][iii]				
					x1	x2	x4	x8	x16
1.0	2003	RZ	8b/10b	2.5 GT/s	0.250 GB/s	0.500 GB/s	1.000 GB/s	2.000 GB/s	4.000 GB/s
2.0	2007			5.0 GT/s	0.500 GB/s	1.000 GB/s	2.000 GB/s	4.000 GB/s	8.000 GB/s
3.0	2010		128b/130b	8.0 GT/s	0.985 GB/s	1.969 GB/s	3.938 GB/s	7.877 GB/s	15.754 GB/s
4.0	2017			16.0 GT/s	1.969 GB/s	3.938 GB/s	7.877 GB/s	15.754 GB/s	31.508 GB/s
5.0	2019			32.0 GT/s	3.938 GB/s	7.877 GB/s	15.754 GB/s	31.508 GB/s	63.015 GB/s
6.0	2022	PAM-4 FEC	242B/256B FLIT	64.0 GT/s 32.0 GBd	7.563 GB/s	15.125 GB/s	30.250 GB/s	60.500 GB/s	121.000 GB/s
7.0	2025 (planned)			128.0 GT/s 64.0 GBd	15.125 GB/s	30.250 GB/s	60.500 GB/s	121.000 GB/s	242.000 GB/s

Notes

i. In each direction (each lane is a dual simplex channel).

ii. Transfer rate refers to the encoded serial bit rate; 2.5 GT/s means 2.5 Gbit/s serial data rate.

iii. Throughput indicates the unencoded bandwidth (without 8b/10b, 128b/130b, or 242B/256B encoding overhead). The PCIe 1.0 transfer rate of 2.5 GT/s per lane means a 2.5 Gbit/s serial bit rate corresponding to a throughput of 2.0 Gbit/s or 250 MB/s prior to 8b/10b encoding.

PCI Express 1.0a

In 2003, PCI-SIG introduced PCIe 1.0a, with a per-lane data rate of 250 MB/s and a transfer rate of 2.5 gigatransfers per second (GT/s).

Transfer rate is expressed in transfers per second instead of bits per second because the number of transfers includes the overhead bits, which do not provide additional throughput;[49] PCIe 1.x uses an 8b/10b encoding scheme, resulting in a 20% (= 2/10) overhead on the raw channel bandwidth.[50] So in the PCIe terminology, transfer rate refers to the encoded bit rate: 2.5 GT/s is 2.5 Gbps on the encoded serial link. This corresponds to 2.0 Gbps of pre-coded data or 250 MB/s, which is referred to as throughput in PCIe.

PCI Express 1.1

In 2005, PCI-SIG[51] introduced PCIe 1.1. This updated specification includes clarifications and several improvements, but is fully compatible with PCI Express 1.0a. No changes were made to the data rate.

PCI Express 2.0

A PCI Express 2.0 expansion card that provides USB 3.0 connectivity.[b]

PCI-SIG announced the availability of the PCI Express Base 2.0 specification on 15 January 2007.[52] The PCIe 2.0 standard doubles the transfer rate compared with PCIe 1.0 to 5 GT/s and the per-lane throughput rises from 250 MB/s to 500 MB/s. Consequently, a 16-lane PCIe connector (x16) can support an aggregate throughput of up to 8 GB/s.

PCIe 2.0 motherboard slots are fully backward compatible with PCIe v1.x cards. PCIe 2.0 cards are also generally backward compatible with PCIe 1.x motherboards, using the available bandwidth of PCI Express 1.1. Overall, graphic cards or motherboards designed for v2.0 work, with the other being v1.1 or v1.0a.

The PCI-SIG also said that PCIe 2.0 features improvements to the point-to-point data transfer protocol and its software architecture.[53]

Intel's first PCIe 2.0 capable chipset was the X38 and boards began to ship from various vendors (Abit, Asus, Gigabyte) as of 21 October 2007.[54] AMD started supporting PCIe 2.0 with its AMD 700 chipset series and nVidia started with the MCP72.[55] All of Intel's prior chipsets, including the Intel P35 chipset, supported PCIe 1.1 or 1.0a.[56]

Like 1.x, PCIe 2.0 uses an 8b/10b encoding scheme, therefore delivering, per-lane, an effective 4 Gbit/s max. transfer rate from its 5 GT/s raw data rate.

PCI Express 2.1

PCI Express 2.1 (with its specification dated 4 March 2009) supports a large proportion of the management, support, and troubleshooting systems planned for full implementation in PCI Express 3.0. However, the speed is the same as PCI Express 2.0. The increase in power from the slot breaks backward compatibility between PCI Express 2.1 cards and some older motherboards with 1.0/1.0a, but most motherboards with PCI Express 1.1 connectors are provided with a BIOS update by their manufacturers through utilities to support backward compatibility of cards with PCIe 2.1.

PCI Express 3.0

PCI Express 3.0 Base specification revision 3.0 was made available in November 2010, after multiple delays. In August 2007, PCI-SIG announced that PCI Express 3.0 would carry a bit rate of 8 gigatransfers per second (GT/s), and that it would be backward compatible with existing PCI Express implementations. At that time, it was also announced that the final specification for PCI Express 3.0 would be delayed until Q2 2010.[57] New features for the PCI Express 3.0 specification included a number of optimizations for enhanced signaling and data integrity, including transmitter and receiver equalization, PLL improvements, clock data recovery, and channel enhancements of currently supported topologies.[58]

Following a six-month technical analysis of the feasibility of scaling the PCI Express interconnect bandwidth, PCI-SIG's analysis found that 8 gigatransfers per second could be manufactured in mainstream silicon process technology, and deployed with existing low-cost materials and infrastructure, while maintaining full compatibility (with negligible impact) with the PCI Express protocol stack.

PCI Express 3.0 upgraded the encoding scheme to 128b/130b from the previous 8b/10b encoding, reducing the bandwidth overhead from 20% of PCI Express 2.0 to approximately 1.54% (= 2/130). PCI Express 3.0's 8 GT/s bit rate effectively delivers 985 MB/s per lane, nearly doubling the lane bandwidth relative to PCI Express 2.0.[48]

On 18 November 2010, the PCI Special Interest Group officially published the finalized PCI Express 3.0 specification to its members to build devices based on this new version of PCI Express.[59]

PCI Express 3.1

In September 2013, PCI Express 3.1 specification was announced for release in late 2013 or early 2014, consolidating various improvements to the published PCI Express 3.0 specification in three areas: power management, performance and functionality.[46][60] It was released in November 2014.[61]

On 29 November 2011, PCI-SIG preliminarily announced PCI Express 4.0,[62] providing a 16 GT/s bit rate that doubles the bandwidth provided by PCI Express 3.0 to 31.5 GB/s in each direction for a 16-lane configuration, while maintaining backward and forward compatibility in both software support and used mechanical interface.[63] PCI Express 4.0 specs also bring OCuLink-2, an alternative to Thunderbolt. OCuLink version 2 has up to 16 GT/s (16 GB/s total for x8 lanes),[42] while the maximum bandwidth of a Thunderbolt 3 link is 5 GB/s.

In June 2016 Cadence, PLDA and Synopsys demoed PCIe 4.0 physical-layer, controller, switch and other IP blocks at the PCI SIG's annual developer's conference.[64]

Mellanox Technologies announced the first 100 Gbit/s network adapter with PCIe 4.0 on 15 June 2016,[65] and the first 200 Gbit/s network adapter with PCIe 4.0 on 10 November 2016.[66]

In August 2016, Synopsys presented a test setup with FPGA clocking a lane to PCIe 4.0 speeds at the Intel Developer Forum. Their IP has been licensed to several firms planning to present their chips and products at the end of 2016.[67]

On the IEEE Hot Chips Symposium in August 2016 IBM announced the first CPU with PCIe 4.0 support, POWER9.[68][69]

PCI-SIG officially announced the release of the final PCI Express 4.0 specification on 8 June 2017.[70] The spec includes improvements in flexibility, scalability, and lower-power.

On 5 December 2017 IBM announced the first system with PCIe 4.0 slots, Power AC922.[71][72]

NETINT Technologies introduced the first NVMe SSD based on PCIe 4.0 on 17 July 2018, ahead of Flash Memory Summit 2018[73]

AMD announced on 9 January 2019 its upcoming Zen 2-based processors and X570 chipset would support PCIe 4.0.[74] AMD had hoped to enable partial support for older chipsets, but instability caused by motherboard traces not conforming to PCIe 4.0 specifications made that impossible.[75][76]

Intel released their first mobile CPUs with PCI express 4.0 support in mid-2020, as a part of the Tiger Lake microarchitecture.[77]

PCI Express 5.0

In June 2017, PCI-SIG announced the PCI Express 5.0 preliminary specification.[70] Bandwidth was expected to increase to 32 GT/s, yielding 63 GB/s in each direction in a 16-lane configuration. The draft spec was expected to be standardized in 2019. Initially, 25.0 GT/s was also considered for technical feasibility.

On 7 June 2017 at PCI-SIG DevCon, Synopsys recorded the first demonstration of PCI Express 5.0 at 32 GT/s.[78]

On 31 May 2018, PLDA announced the availability of their XpressRICH5 PCIe 5.0 Controller IP based on draft 0.7 of the PCIe 5.0 specification on the same day.[79][80]

On 10 December 2018, the PCI SIG released version 0.9 of the PCIe 5.0 specification to its members,[81] and on 17 January 2019, PCI SIG announced the version 0.9 had been ratified, with version 1.0 targeted for release in the first quarter of 2019.[82]

On 29 May 2019, PCI-SIG officially announced the release of the final PCI Express 5.0 specification.[83]

On 20 November 2019, Jiangsu Huacun presented the first PCIe 5.0 Controller HC9001 in a 12 nm manufacturing process.[84] Production started in 2020.

On 17 August 2020, IBM announced the Power10 processor with PCIe 5.0 and up to 32 lanes per single-chip module (SCM) and up to 64 lanes per double-chip module (DCM).[85]

On 9 September 2021, IBM announced the Power E1080 Enterprise server with planned availability date 17 September.[86] It can have up to 16 Power10 SCMs with maximum of 32 slots per system which can act as PCIe 5.0 x8 or PCIe 4.0 x16.[87] Alternatively they can be used as PCIe 5.0 x16 slots for optional optical CXP converter adapters connecting to external PCIe expansion drawers.

On 27 October 2021, Intel announced the 12th Gen Intel Core CPU family, the world's first consumer x86-64 processors with PCIe 5.0 (up to 16 lanes) connectivity.[88]

On 22 March 2022, Nvidia announced Nvidia Hopper GH100 GPU, the world's first PCIe 5.0 GPU.[89]

On 23 May 2022, AMD announced its Zen 4 architecture with support for up to 24 lanes of PCIe 5.0 connectivity on consumer platforms and 128 lanes on server platforms.[90][91]

PCI Express 6.0

On 18 June 2019, PCI-SIG announced the development of PCI Express 6.0 specification. Bandwidth is expected to increase to 64 GT/s, yielding 128 GB/s in each direction in a 16-lane configuration, with a target release date of 2021.[92] The new standard uses 4-level pulse- amplitude modulation (PAM-4) with a low-latency forward error correction (FEC) in place of non-return-to-zero (NRZ) modulation.[93] Unlike previous PCI Express versions, forward error correction is used to increase data integrity and PAM-4 is used as line code so that two bits are transferred per transfer. With 64 GT/s data transfer rate (raw bit rate), up to 121 GB/s in each direction is possible in x16 configuration.[92]

On 24 February 2020, the PCI Express 6.0 revision 0.5 specification (a "first draft" with all architectural aspects and requirements defined) was released.[94]

On 5 November 2020, the PCI Express 6.0 revision 0.7 specification (a "complete draft" with electrical specifications validated via test chips) was released.[95]

On 6 October 2021, the PCI Express 6.0 revision 0.9 specification (a "final draft") was released.[96]

On 11 January 2022, PCI-SIG officially announced the release of the final PCI Express 6.0 specification.[97]

PAM-4 coding results in a vastly higher bit error rate (BER) of 10^{-6} (vs. 10^{-12} previously), so in place of 128b/130b encoding, a 3-way interlaced forward error correction (FEC) is used in addition to cyclic redundancy check (CRC). A fixed 256 byte Flow Control Unit (FLIT) block carries 242 bytes of data, which includes variable-sized transaction level packets (TLP) and data link layer payload (DLLP); remaining 14 bytes are reserved for 8-byte CRC and 6-byte FEC. [98][99] 3-way Gray code is used in PAM-4/FLIT mode to reduce error rate; the interface does not switch to NRZ and 128/130b encoding even when retraining to lower data rates.[100][101]

PCI Express 7.0

On 21 June 2022, PCI-SIG announced the development of PCI Express 7.0 specification.[102] It will deliver 128 GT/s raw bit rate and up to 242 GB/s per direction in x16 configuration, using the same PAM4 signaling as version 6.0. Doubling of the data rate will be achieved by fine-tuning channel parameters to decrease signal losses and improve power efficiency. The specification is expected to be finalised in 2025.

Extensions and future directions

Some vendors offer PCIe over fiber products,[103][104][105] with active optical cables (AOC) for PCIe switching at increased distance in PCIe expansion drawers,[106][87] or in specific cases where transparent PCIe bridging is preferable to using a more mainstream standard (such as InfiniBand or Ethernet) that may require additional software to support it.

Thunderbolt was co-developed by Intel and Apple as a general-purpose high speed interface combining a logical PCIe link with DisplayPort and was originally intended as an all-fiber interface, but due to early difficulties in creating a consumer-friendly fiber interconnect, nearly all implementations are copper systems. A notable exception, the Sony VAIO Z VPC-Z2, uses a nonstandard USB port with an optical component to connect to an outboard PCIe display adapter. Apple has been the primary driver of Thunderbolt adoption through 2011, though several other vendors[107] have announced new products and systems featuring Thunderbolt. Thunderbolt 3 forms the basis of the USB4 standard.

Mobile PCIe specification (abbreviated to *M-PCIe*) allows PCI Express architecture to operate over the MIPI Alliance's M-PHY

physical layer technology. Building on top of already existing widespread adoption of M-PHY and its low-power design, Mobile PCIe lets mobile devices use PCI Express.[108]

Draft process

There are 5 primary releases/checkpoints in a PCI-SIG specification:[109]

- Draft 0.3 (Concept): this release may have few details, but outlines the general approach and goals.
- Draft 0.5 (First draft): this release has a complete set of architectural requirements and must fully address the goals set out in the 0.3 draft.
- Draft 0.7 (Complete draft): this release must have a complete set of functional requirements and methods defined, and no new functionality may be added to the specification after this release. Before the release of this draft, electrical specifications must have been validated via test silicon.
- Draft 0.9 (Final draft): this release allows PCI-SIG member companies to perform an internal review for intellectual property, and no functional changes are permitted after this draft.
- 1.0 (Final release): this is the final and definitive specification, and any changes or enhancements are through Errata documentation and Engineering Change Notices (ECNs) respectively.

Historically, the earliest adopters of a new PCIe specification generally begin designing with the Draft 0.5 as they can confidently build up their application logic around the new bandwidth definition and often even start developing for any new protocol features. At the Draft 0.5 stage, however, there is still a strong likelihood of changes in the actual PCIe protocol layer implementation, so designers responsible for developing these blocks internally may be more hesitant to begin work than those using interface IP from external sources.

Hardware protocol summary

The PCIe link is built around dedicated unidirectional couples of serial (1-bit), point-to-point connections known as *lanes*. This is in sharp contrast to the earlier PCI connection, which is a bus-based system where all the devices share the same bidirectional, 32-bit or 64-bit parallel bus.

PCI Express is a layered protocol, consisting of a *transaction layer*, a *data link layer*, and a *physical layer*. The Data Link Layer is subdivided to include a media access control (MAC) sublayer. The Physical Layer is subdivided into logical and electrical sublayers. The Physical logical- sublayer contains a physical coding sublayer (PCS). The terms are borrowed from the IEEE 802 networking protocol model.

The PCIe Physical Layer (*PHY*, *PCIEPHY*, *PCI Express PHY*, or *PCIe PHY*) specification is divided into two sub-layers, corresponding to electrical and logical specifications. The logical sublayer is sometimes further divided into a MAC sublayer and a PCS, although this division is not formally part of the PCIe specification. A specification published by Intel, the PHY Interface for PCI Express (PIPE),[111] defines the MAC/PCS functional partitioning and the interface between these two sub-layers. The PIPE specification also identifies the *physical media attachment* (PMA) layer, which includes the serializer/deserializer (SerDes) and other analog circuitry; however, since SerDes implementations vary greatly among ASIC vendors, PIPE does not specify an interface between the PCS and PMA.

Connector pins and lengths

Lanes	Pins		Length	
	Total	Variable	Total	Variable
x1	2×18 = 36[110]	2× 7 = 14	25 mm	7.65 mm
x4	2×32 = 64	2×21 = 42	39 mm	21.65 mm
x8	2×49 = 98	2×38 = 76	56 mm	38.65 mm
x16	2×82 = 164	2×71 = 142	89 mm	71.65 mm

At the electrical level, each lane consists of two unidirectional differential pairs operating at 2.5, 5, 8, 16 or 32 Gbit/s, depending on the negotiated capabilities. Transmit and receive are separate differential pairs, for a total of four data wires per lane.

A connection between any two PCIe devices is known as a *link*, and is built up from a collection of one or more *lanes*. All devices must minimally support single-lane (x1) link. Devices may optionally support wider links composed of up to 32 lanes.[112][113] This allows for very good compatibility in two ways:

- A PCIe card physically fits (and works correctly) in any slot that is at least as large as it is (e.g., an x1 sized card works in any sized slot);
- A slot of a large physical size (e.g., x16) can be wired electrically with fewer lanes (e.g., x1, x4, x8, or x12) as long as it provides the ground connections required by the larger physical slot size.

In both cases, PCIe negotiates the highest mutually supported number of lanes. Many graphics cards, motherboards and BIOS versions are verified to support x1, x4, x8 and x16 connectivity on the same connection.

The width of a PCIe connector is 8.8 mm, while the height is 11.25 mm, and the length is variable. The fixed section of the connector is 11.65 mm in length and contains two rows of 11 pins each (22 pins total), while the length of the other section is variable depending on the number of lanes. The pins are spaced at 1 mm intervals, and the thickness of the card going into the connector is 1.6 mm.[114][115]

An open-end PCI Express x1 connector lets longer cards that use more lanes be plugged while operating at x1 speeds

Data transmission

PCIe sends all control messages, including interrupts, over the same links used for data. The serial protocol can never be blocked, so latency is still comparable to conventional PCI, which has dedicated interrupt lines. When the problem of IRQ sharing of pin based interrupts is taken into account and the fact that message signaled interrupts (MSI) can bypass an I/O APIC and be delivered to the CPU directly, MSI performance ends up being substantially better. [116]

Data transmitted on multiple-lane links is interleaved, meaning that each successive byte is sent down successive lanes. The PCIe specification refers to this interleaving as *data striping*. While requiring significant hardware complexity to synchronize (or deskew) the incoming striped data, striping can significantly reduce the latency of the n^{th} byte on a link. While the lanes are not tightly synchronized, there is a limit to the *lane to lane skew* of 20/8/6 ns for 2.5/5/8 GT/s so the hardware buffers can re-align the striped data.[117] Due to padding requirements, striping may not necessarily reduce the latency of small data packets on a link.

As with other high data rate serial transmission protocols, the clock is embedded in the signal. At the physical level, PCI Express 2.0 utilizes the 8b/10b encoding scheme[48] (line code) to ensure that strings of consecutive identical digits (zeros or ones) are limited in length. This coding was used to prevent the receiver from losing track of where the bit edges are. In this coding scheme every eight (uncoded) payload bits of data are replaced with 10 (encoded) bits of transmit data, causing a 20% overhead in the electrical bandwidth. To improve the available bandwidth, PCI Express version 3.0 instead uses 128b/130b encoding (1.54% overhead). Line encoding limits the run length of identical-digit strings in data streams and ensures the receiver stays synchronised to the transmitter via clock recovery.

A desirable balance (and therefore spectral density) of 0 and 1 bits in the data stream is achieved by XORing a known binary polynomial as a "scrambler" to the data stream in a feedback topology. Because the scrambling polynomial is known, the data can be recovered by applying the XOR a second time. Both the scrambling and descrambling steps are carried out in hardware.

Data link layer

The data link layer performs three vital services for the PCIe link:

1. sequence the transaction layer packets (TLPs) thatare generated by the transaction layer,
2. ensure reliable delivery of TLPs between two endpoints via an acknowledgement protocol (ACK and NAK signaling) that explicitly requires replay of unacknowledged/bad TLPs,
3. initialize and manage flow control credits

On the transmit side, the data link layer generates an incrementing sequence number for each outgoing TLP. It serves as a unique identification tag for each transmitted TLP, and is inserted into the header of the outgoing TLP. A 32-bit cyclic redundancy check code (known in this context as Link CRC or LCRC) is also appended to the end of each outgoing TLP.

On the receive side, the received TLP's LCRC and sequence number are both validated in the link layer. If either the LCRC check fails (indicating a data error), or the sequence-number is out of range (non-consecutive from the last valid received TLP), then the bad TLP, as well as any TLPs received after the bad TLP, are considered invalid and discarded. The receiver sends a negative acknowledgement message (NAK) with the sequence-number of the invalid TLP, requesting re-transmission of all TLPs forward of that sequence-number. If the received TLP passes the LCRC check and has the correct sequence number, it is treated as valid. The link receiver increments the sequence-number (which tracks the last received good TLP), and forwards the valid TLP to the receiver's transaction layer. An ACK message is sent to remote transmitter, indicating the TLP was successfully received (and by extension, all TLPs with past sequence-numbers.)

If the transmitter receives a NAK message, or no acknowledgement (NAK or ACK) is received until a timeout period expires, the transmitter must retransmit all TLPs that lack a positive acknowledgement (ACK). Barring a persistent malfunction of the device or transmission medium, the link-layer presents a reliable connection to the transaction layer, since the transmission protocol ensures delivery of TLPs over an unreliable medium.

In addition to sending and receiving TLPs generated by the transaction layer, the data-link layer also generates and consumes data link layer packets (DLLPs). ACK and NAK signals are communicated via DLLPs, as are some power management messages and flow control credit information (on behalf of the transaction layer).

In practice, the number of in-flight, unacknowledged TLPs on the link is limited by two factors: the size of the transmitter's replay buffer (which must store a copy of all transmitted TLPs until the remote receiver ACKs them), and the flow control credits issued by the receiver to a transmitter. PCI Express requires all receivers to issue a minimum number of credits, to guarantee a link allows sending PCIConfig TLPs and message TLPs.

PCI Express implements split transactions (transactions with request and response separated by time), allowing the link to carry other traffic while the target device gathers data for the response.

PCI Express uses credit-based flow control. In this scheme, a device advertises an initial amount of credit for each received buffer in its transaction layer. The device at the opposite end of the link, when sending transactions to this device, counts the number of credits each TLP consumes from its account. The sending device may only transmit a TLP when doing so does not make its consumed credit count exceed its credit limit. When the receiving device finishes processing the TLP from its buffer, it signals a return of credits to the sending device, which increases the credit limit by the restored amount. The credit counters are modular counters, and the comparison of consumed credits to credit limit requires modular arithmetic. The advantage of this scheme (compared to other methods such as wait states or handshake-based transfer protocols) is that the latency of credit return does not affect performance, provided that the credit limit is not encountered. This assumption is generally met if each device is designed with adequate buffer sizes.

PCIe 1.x is often quoted to support a data rate of 250 MB/s in each direction, per lane. This figure is a calculation from the physical signaling rate (2.5 gigabaud) divided by the encoding overhead (10 bits per byte). This means a sixteen lane (x16) PCIe card would then be theoretically capable of 16x250 MB/s = 4 GB/s in each direction. While this is correct in terms of data bytes, more meaningful calculations are based on the usable data payload rate, which depends on the profile of the traffic, which is a function of the high-level (software) application and intermediate protocol levels.

Like other high data rate serial interconnect systems, PCIe has a protocol and processing overhead due to the additional transfer robustness (CRC and acknowledgements). Long continuous unidirectional transfers (such as those typical in high-performance storage controllers) can approach >95% of PCIe's raw (lane) data rate. These transfers also benefit the most from increased number of lanes (x2, x4, etc.) But in more typical applications (such as a USB or Ethernet controller), the traffic profile is characterized as short data packets with frequent enforced acknowledgements.[118] This type of traffic reduces the efficiency of the link, due to overhead from packet parsing and forced interrupts (either in the device's host interface or the PC's CPU). Being a protocol for devices connected to the same printed circuit board, it does not require the same tolerance for transmission errors as a protocol for communication over longer distances, and thus, this loss of efficiency is not particular to PCIe.

Efficiency of the link

As for any "network like" communication links, some of the "raw" bandwidth is consumed by protocol overhead:[119]

A PCIe 1.x lane for example offers a data rate on top of the physical layer of 250 MB/s (simplex). This isn't the payload bandwidth but the physical layer bandwidth – a PCIe lane has to carry additional information for full functionality.[119]

Gen 2 Transaction Layer Packet[119]: 3

Layer	**PHY**	**Data Link Layer**	**Transaction**			**Data Link Layer**	**PHY**
Data	Start	Sequence	Header	Payload	ECRC	LCRC	End
Size (Bytes)	1	2	12 or 16	0 to 4096	4 (optional)	4	1

The Gen2 overhead is then 20, 24, or 28 bytes per transaction.

Gen 3 Transaction Layer Packet[119]: 3

Layer	**G3 PHY**	**Data Link Layer**	**Transaction Layer**			**Data Link Layer**
Data	Start	Sequence	Header	Payload	ECRC	LCRC
Size (Bytes)	4	2	12 or 16	0 to 4096	4 (optional)	4

The Gen3 overhead is then 22, 26 or 30 bytes per transaction.

The $\text{Packet Efficiency} = \frac{\text{Payload}}{\text{Payload} + \text{Overhead}}$ for a 128 byte payload is 86%, and 98% for a 1024 byte payload. For small accesses like register settings (4 bytes), the efficiency drops as low as 16%.

The maximum payload size (MPS) is set on all devices based on smallest maximum on any device in the chain. If one device has an MPS of 128 bytes, *all* devices of the tree must set their MPS to 128 bytes. In this case the bus will have a peak efficiency of 86% for writes.[119]: 3

Applications

PCI Express operates in consumer, server, and industrial applications, as a motherboard-level interconnect (to link motherboard-mounted peripherals), a passive backplane interconnect and as an expansion card interface for add-in boards.

In virtually all modern (as of 2012) PCs, from consumer laptops and desktops to enterprise data servers, the PCIe bus serves as the primary motherboard-level interconnect, connecting the host system-processor with both integrated peripherals (surface-mounted ICs) and add-on peripherals (expansion cards). In most of these systems, the PCIe bus co-exists with one or more legacy PCI buses, for backward compatibility with the large body of legacy PCI peripherals.

As of 2013, PCI Express has replaced AGP as the default interface for graphics cards on new systems. Almost all models of graphics cards released since 2010 by AMD (ATI) and Nvidia use PCI Express. Nvidia uses the high-bandwidth data transfer of PCIe for its Scalable Link Interface (SLI) technology, which allows multiple graphics cards of the same chipset and model number to run in tandem, allowing increased performance. AMD has also developed a multi-GPU system based on PCIe called CrossFire. AMD, Nvidia, and Intel have released motherboard chipsets that support as many as four PCIe x16 slots, allowing tri-GPU and quad-GPU card configurations.

Asus Nvidia GeForce GTX 650 Ti, a PCI Express 3.0 x16 graphics card

The Nvidia GeForce GTX 1070, a PCI Express 3.0 x16 Graphics card

External GPUs

Theoretically, external PCIe could give a notebook the graphics power of a desktop, by connecting a notebook with any PCIe desktop video card (enclosed in its own external housing, with a power supply and cooling); this is possible with an ExpressCard or Thunderbolt interface. An ExpressCard interface provides bit rates of 5 Gbit/s (0.5 GB/s throughput), whereas a Thunderbolt interface provides bit rates of up to 40 Gbit/s (5 GB/s throughput).

In 2006, Nvidia developed the Quadro Plex external PCIe family of GPUs that can be used for advanced graphic applications for the professional market.[120] These video cards require a PCI Express x8 or x16 slot for the host-side card, which connects to the Plex via a VHDCI carrying eight PCIe lanes.[121]

Intel 82574L Gigabit Ethernet NIC, a PCI Express x1 card

In 2008, AMD announced the ATI XGP technology, based on a proprietary cabling system that is compatible with PCIe x8 signal transmissions.[122] This connector is available on the Fujitsu Amilo and the Acer Ferrari One notebooks. Fujitsu launched their AMILO GraphicBooster enclosure for XGP soon thereafter.[123] Around 2010 Acer launched the Dynavivid graphics dock for XGP.[124]

A Marvell-based SATA 3.0 controller, as a PCI Express x1 card

In 2010, external card hubs were introduced that can connect to a laptop or desktop through a PCI ExpressCard slot. These hubs can accept full-sized graphics cards. Examples include MSI GUS,[125] Village Instrument's ViDock,[126] the Asus XG Station, Bplus PE4H V3.2 adapter,[127] as well as more improvised DIY devices.[128] However such solutions are limited by the size (often only x1) and version of the available PCIe slot on a laptop.

The Intel Thunderbolt interface has provided a new option to connect with a PCIe card externally. Magma has released the ExpressBox 3T, which can hold up to three PCIe cards (two at x8 and one at x4).[129] MSI also released the Thunderbolt GUS II, a PCIe chassis dedicated

for video cards.[130] Other products such as the Sonnet's Echo Express[131] and mLogic's mLink are Thunderbolt PCIe chassis in a smaller form factor.[132]

In 2017, more fully featured external card hubs were introduced, such as the Razer Core, which has a full-length PCIe x16 interface.[133]

Storage devices

An OCZ RevoDrive SSD, a full- height x4 PCI Express card

The PCI Express protocol can be used as data interface to flash memory devices, such as memory cards and solid-state drives (SSDs).

The XQD card is a memory card format utilizing PCI Express, developed by the CompactFlash Association, with transfer rates of up to 1GB/s.[134]

Many high-performance, enterprise-class SSDs are designed as PCI Express RAID controller cards. Before NVMe was standardized, many of these cards utilized proprietary interfaces and custom drivers to communicate with the operating system; they had much higher transfer rates (over 1 GB/s) and IOPS (over one million I/O operations per second) when compared to Serial ATA or SAS drives.[135][136] For example, in

2011 OCZ and Marvell co-developed a native PCI Express solid-state drive controller for a PCI Express 3.0 x16 slot with maximum capacity of 12 TB and a performance of to 7.2 GB/s sequential transfers and up to 2.52 million IOPS in random transfers.[137]

SATA Express was an interface for connecting SSDs through SATA-compatible ports, optionally providing multiple PCI Express lanes as a pure PCI Express connection to the attached storage device.[138] M.2 is a specification for internally mounted computer expansion cards and associated connectors, which also uses multiple PCI Express lanes.[139]

PCI Express storage devices can implement both AHCI logical interface for backward compatibility, and NVM Express logical interface for much faster I/O operations provided by utilizing internal parallelism offered by such devices. Enterprise-class SSDs can also implement SCSI over PCI Express.[140]

Certain data-center applications (such as large computer clusters) require the use of fiber-optic interconnects due to the distance limitations inherent in copper cabling. Typically, a network-oriented standard such as Ethernet or Fibre Channel suffices for these applications, but in some cases the overhead introduced by routable protocols is undesirable and a lower-level interconnect, such as InfiniBand, RapidIO, or NUMAlink is needed. Local-bus standards such as PCIe and HyperTransport can in principle be used for this purpose,[141] but as of 2015, solutions are only available from niche vendors such as Dolphin ICS, and TTTech Auto.

Competing protocols

Other communications standards based on high bandwidth serial architectures include InfiniBand, RapidIO, HyperTransport, Intel QuickPath Interconnect, and the Mobile Industry Processor Interface (MIPI). The differences are based on the trade-offs between flexibility and extensibility vs latency and overhead. For example, making the system hot-pluggable, as with Infiniband but not PCI Express, requires that software track network topology changes.

Another example is making the packets shorter to decrease latency (as is required if a bus must operate as a memory interface). Smaller packets mean packet headers consume a higher percentage of the packet, thus decreasing the effective bandwidth. Examples of bus protocols designed for this purpose are RapidIO and HyperTransport.

PCI Express falls somewhere in the middle, targeted by design as a system interconnect (local bus) rather than a device interconnect or routed network protocol. Additionally, its design goal of software transparency constrains the protocol and raises its latency somewhat.

Delays in PCIe 4.0 implementations led to the Gen-Z consortium, the CCIX effort and an open Coherent Accelerator Processor Interface (CAPI) all being announced by the end of 2016.[142]

On 11 March 2019, Intel presented Compute Express Link (CXL), a new interconnect bus, based on the PCI Express 5.0 physical layer infrastructure. The initial promoters of the CXL specification included: Alibaba, Cisco, Dell EMC, Facebook, Google, HPE, Huawei, Intel and Microsoft.[143]

Integrators list

The PCI-SIG Integrators List lists products made by PCI-SIG member companies that have passed compliance testing. The list include switches, bridges, NICs, SSDs, etc.[144]

See also

- Active State Power Management (ASPM)
- Peripheral Component Interconnect PCI
- configuration space
- PCI-X
- PCI/104-Express
- PCIe/104
- Root complex Serial
- Digital Video Out (SDVO) List of
- device bit rates § Main buses
- UCIe

Notes

a. Switches can create multiple endpoints out of one to allow sharing it with multiple devices.

b. The card's Serial ATA power connector is present because the USB 3.0 ports require more power than the PCI Express bus can supply. More often, a 4-pin Molex power connector is used.

References

1. Mayhew, D.; Krishnan, V. (August 2003). "PCI express and advanced switching: Evolutionary path to building next generation interconnects". *11th Symposium on High Performance Interconnects, 2003. Proceedings*. pp. 21–29. doi:10.1109/CONECT.2003.1231473 (https://doi.org/10.1109%2FCONECT.2003.1231473). ISBN 0-7695-2012-X. S2CID 7456382 (https://api.semanticscholar.org/CorpusID:7456382).
2. "Definition of PCI Express" (https://www.pcmag.com/encyclopedia/term/48998/pci-express). *PCMag*.
3. Zhang, Yanmin; Nguyen, T Long (June 2007). "Enable PCI Express Advanced Error Reporting in the Kernel" (https://web.archive.org/web/20160310074031/https://ols.fedoraproject.org/OLS/Reprints-2007/zhang-Reprint.pdf) (PDF). *Proceedings of the Linux Symposium*. Fedora project. Archived from the original (https://ols.fedoraproject.org/OLS/Reprints-2007/zhang-Re print.pdf) (PDF) on 10 March 2016. Retrieved 8 May 2012.
4. https://www.hyperstone.com Flash Memory Form Factors—The Fundamentals of Reliable Flash Storage, Retrieved 19 April 2018
5. Ravi Budruk (21 August 2007). "PCI Express Basics" (https://web.archive.org/web/20140715120034/http://www.pcisig.com/developers/main/training_materials/get_document?doc_id=4e00a39acaa5c5a8ee44ebb07baba982e5972c67). PCI-SIG. Archived from the original (http://www.pcisig.com/developers/main/training_materials/get_document?doc_id=4e00a39acaa5c5a8ee44ebb07baba982e5972c67) (PDF) on 15 July 2014. Retrieved 15 July 2014.
6. "What are PCIe Slots and Their Uses" (https://pcguide101.com/motherboard/what-are-pcie-slots/). PC Guide 101. 18 May 2021. Retrieved 21 June 2021.
7. "How PCI Express Works" (http://computer.howstuffworks.com/pci-express.htm). *How Stuff Works*. 17 August 2005. Archived (https://web.archive.org/web/20091203053924/http://computer.howstuffworks.com/pci-express.htm) from the original on 3 December 2009. Retrieved 7 December 2009.
8. "4.2.4.9. Link Width and Lane Sequence Negotiation", *PCI Express Base Specification, Revision 2.1.*, 4 March 2009
9. "PCI Express Architecture Frequently Asked Questions" (https://web.archive.org/web/20081113163608/http://www.pcisig.com/news_room/faqs/faq_express/). PCI-SIG. Archived from the original (http://www.pcisig.com/news_room/faqs/faq_express/) on 13 November 2008. Retrieved 23 November 2008.
10. "PCI Express Bus" (https://web.archive.org/web/20071208162241/http://www.interfacebus.com/Design_Connector_PCI_Express.html). *Interface bus*. Archived from the original (http://www.interfacebus.com/Design_Connector_PCI_Express.html) on 8 December 2007. Retrieved 12 June 2010.
11. 32 lanes are defined by the *PCIe Base Specification* up to PCIe 5.0 but there's no card standard in the *PCIe Card Electromechanical Specification* and that lane number was never implemented.
12. "PCI Express — An Overview of the PCI Express Standard" (https://web.archive.org/web/20100105163040/http://zone.ni.com/devzone/cda/tut/p/id/3767). *Developer Zone*. National Instruments. 13 August 2009. Archived from the original (http://zone.ni.com/devzone/cda/tut/p/id/3767) on 5 January 2010. Retrieved 7 December 2009.
13. Qazi, Atif. "What are PCIe Slots?" (https://pcgearlab.com/motherboard/what-are-pcie-slots/). *PC Gear Lab*. Retrieved 8 April 2020.
14. "New PCIe Form Factor Enables Greater PCIe SSD Adoption" (http://www.nvmexpress.org/blog/new-pcie-form-factor-enables-greater-pcie-ssd-adoption/). *NVM Express*. 12 June 2012. Archived (https://web.archive.org/web/20150906180730/http://www.nvmexpress.org/blog/new-pcie-form-factor-enables-greater-pcie-ssd-adoption/) from the original on 6 September 2015.
15. "Memblaze PBlaze4 AIC NVMe SSD Review" (http://www.storagereview.com/memblaze_pblaze4_aic_nvme_ssd_review). *StorageReview*. 21 December 2015.
16. July 2015, Kane Fulton 20 (20 July 2015). "19 graphics cards that shaped the future of gaming" (https://www.techradar.com/news/gaming/19-graphics-cards-that-shaped-the-future-of-gaming-1289666). *TechRadar*.
17. Leadbetter, Richard (16 September 2020). "Nvidia GeForce RTX 3080 review: welcome to the next level" (https://www.eurogamer.net/articles/digitalfoundry-2020-nvidia-geforce-rtx-3080-review). *Eurogamer*.
18. "Sapphire Radeon RX 5700 XT Pulse Review | bit-tech.net" (https://bit-tech.net/reviews/tech/graphics/sapphire-radeon-rx-5700-xt-pulse-review/1/). *bit-tech.net*. Retrieved 26 August 2019.
19. "AMD Radeon™ RX 5700 XT 8GB GDDR6 THICC II—RX-57XT8DFD6" (http://xfxforce.com/en-gb/Products/product-category/amd-radeon-rx-5700xt-8gb-thicc-ii-rx-57xt8dfd6). *xfxforce.com*. Retrieved 25 August 2019.
20. "ROG Strix GeForce RTX 3080 OC Edition 10GB GDDR6X | Graphics Cards" (https://rog.asus.com/graphics-cards/graphics-cards/rog-strix/rog-strix-rtx3080-o10g-gaming-model/spec). *rog.asus.com*.
21. "What is the A side, B side configuration of PCI cards" (https://web.archive.org/web/20111102042843/http://www.adexelec.com/faq.htm#pcikeys). *Frequently Asked Questions*. Adex Electronics. 1998. Archived from the original (http://www.adexelec.

com/faq.htm#pcikeys) on 2 November 2011. Retrieved 24 October 2011.
22. *PCI Express Card Electromechanical Specification Revision 2.0*
23. "PCI Express Card Electromechanical Specification Revision 4.0, Version 1.0 (Clean)" (https://members.pcisig.com/wg/PCI-SIG/document/13446).
24. "L1 PM Substates with CLKREQ, Revision 1.0a" (https://pcisig.com/sites/default/files/specification_documents/ECN_L1_PM_Substates_with_CLKREQ_31_May_2013_Rev10a.pdf) (PDF). PCI-SIG. Retrieved 8 November 2018.
25. "Emergency Power Reduction Mechanism with PWRBRK Signal ECN" (https://web.archive.org/web/20181109193739/https://pcisig.com/sites/default/files/specification_documents/Emergency%20Power%20Reduction%20Mechanism%20with%20PWRBRK%20Signal%20ECN.pdf) (PDF). PCI-SIG. Archived from the original (https://pcisig.com/sites/default/files/specification_documents/Emergency%20Power%20Reduction%20Mechanism%20with%20PWRBRK%20Signal%20ECN.pdf) (PDF) on 9 November 2018. Retrieved 8 November 2018. *PCI Express Card Electromechanical Specification Revision 1.1*
26. Schoenborn, Zale (2004), *Board Design Guidelines for PCI Express Architecture* (http://e2e.ti.com/cfs-file/ key/communityserver-discussions-components-files/639/7851.PCIe_5F00_designGuides.pdf#page=19) (PDF), PCI-SIG, pp. 19–21, archived (https://web.archive.org/web/20160327185412/http://e2e.ti.com/cfs-file/ key/communityserver-discussions-compo nents-files/639/7851.PCIe_5F00_designGuides.pdf#page=19) (PDF) from the original on 27 March 2016
27. *PCI Express Base Specification,Revision 1.1* Page 332
28. "Where Does PCIe Cable Go?" (https://greatpcreview.com/guides/where-does-pcie-cables-go/). 16 January 2022. Retrieved 10 June 2022.
29. "Mini-Fit® PCI Express®* Wire to Board Connector System" (https://www.molex.com/pdm_docs/ps/PS-45558-001-001.pdf) (PDF). Retrieved 4 December 2020.
30. *PCI Express x16 Graphics 150W-ATX Specification Revision 1.0*
31. *PCI Express 225 W/300 W High Power Card Electromechanical Specification Revision 1.0*
32. *PCI Express Card Electromechanical Specification Revision 3.0*
33. Yun Ling (16 May 2008). "PCIe Electromechanical Updates" (https://web.archive.org/web/20151105083550/http://kavi.pcisig.com/developers/main/training_materials/get_document?doc_id=fa4ec3357012d69821baa0856011c665ac770768). Archived from the original (http://kavi.pcisig.com/developers/main/training_materials/get_document?doc_id=fa4ec3357012d69821baa0856011c665ac770768) on 5 November 2015. Retrieved 7 November 2015.
34. "MP1: Mini PCI Express / PCI Express Adapter" (http://www.hwtools.net/Adapter/MP1.html). *hwtools.net*. 18 July 2014. Archived (https://web.archive.org/web/20141003233055/http://www.hwtools.net/Adapter/MP1.html) from the original on 3 October 2014. Retrieved 28 September 2014.
35. "mSATA FAQ: A Basic Primer" (http://forum.notebookreview.com/lenovo-ibm/574993-msata-faq-basic-primer.html). Notebook review. Archived (https://web.archive.org/web/20120212164949/http://forum.notebookreview.com/lenovo-ibm/574 993-msata-faq-basic-primer.html) from the original on 12 February 2012.
36. "Eee PC Research" (http://beta.ivancover.com/wiki/index.php/Eee_PC_Research). *ivc* (wiki). Archived (https://web.archive.org/web/20100330035948/http://beta.ivancover.com/wiki/index.php/Eee_PC_Research) from the original on 30 March 2010. Retrieved 26 October 2009.
37. "Desktop Board Solid-state drive (SSD) compatibility" (http://www.intel.com/support/motherboards/desktop/sb/CS-032415.htm?wapkw=032415). Intel. Archived (https://web.archive.org/web/20160102233130/http://www.intel.com/support/motherboards/desktop/sb/CS-032415.htm?wapkw=032415) from the original on 2 January 2016.
38. "How to distinguish the differences between M.2 cards | Dell US" (https://www.dell.com/support/article/en-us/sln301626/how-to-distinguish-the-differences-between-m-2-cards?lang=en). *www.dell.com*. Retrieved 24 March 2020.
39. "PCI Express External Cabling 1.0 Specification" (http://www.pcisig.com/specifications/pciexpress/pcie_cabling1.0/). Archived (https://web.archive.org/web/20070210055546/http://www.pcisig.com/specifications/pciexpress/pcie_cabling1.0/) from the original on 10 February 2007. Retrieved 9 February 2007.
40. "PCI Express External Cabling Specification Completed by PCI-SIG" (https://web.archive.org/web/20131126064157/http://www.pcisig.com/news_room/news/press_release/02_07_07). PCI SIG. 7 February 2007. Archived from the original (http://www.pcisig.com/news_room/news/press_release/02_07_07) on 26 November 2013. Retrieved 7 December 2012.
41. "OCuLink connectors and cables support new PCIe standard" (https://web.archive.org/web/20170313215047/http://www.connectortips.com/oculink-connectors-cables-support-new-pcie-standard/). *www.connectortips.com*. Archived from the original (https://www.connectortips.com/oculink-connectors-cables-support-new-pcie-standard/) on 13 March 2017.
42. Mokosiy, Vitaliy (9 October 2020). "Untangling terms: M.2, NVMe, USB-C, SAS, PCIe, U.2, OCuLink" (https://mokosiy.mediu m.com/untangling-terms-m-2-nvme-usb-c-sas-pcie-6599c044f38e). *Medium*. Retrieved 26 March 2021.
43. "Supermicro Universal I/O (UIO) Solutions" (http://www.supermicro.com/products/nfo/uio.cfm). Supermicro.com. Archived (https://web.archive.org/web/20140324184437/http://www.supermicro.com/products/nfo/uio.cfm) from the original on 24 March 2014. Retrieved 24 March 2014.
44. "Get ready for M-PCIe testing" (http://www.edn.com/design/pc-board/4423319/Get-ready-for-M-PCIe-testing), *PC board design*, EDN
45. "PCI SIG discusses M-PCIe oculink & 4th gen PCIe" (https://www.theregister.co.uk/Print/2013/09/13/pci_sig_discusses_m_

pcie_oculink_and_fourth_gen_pcie/), *The Register*, UK, 13 September 2013, archived (https://web.archive.org/web/20170629201006/http://www.theregister.co.uk/Print/2013/09/13/pci_sig_discusses_m_pcie_oculink_and_fourth_gen_pcie/) from the original on 29 June 2017

46. "PCI Express 4.0 Frequently Asked Questions" (https://web.archive.org/web/20140518224913/http://www.pcisig.com/news_room/faqs/FAQ_PCI_Express_4.0/#EQ3). *pcisig.com*. PCI-SIG. Archived from the original (http://www.pcisig.com/news_room/faqs/FAQ_PCI_Express_4.0/#EQ3) on 18 May 2014. Retrieved 18 May 2014.

47. "PCI Express 3.0 Frequently Asked Questions" (https://web.archive.org/web/20140201172536/http://www.pcisig.com/news_room/faqs/pcie3.0_faq/#EQ2). *pcisig.com*. PCI-SIG. Archived from the original (http://www.pcisig.com/news_room/faqs/pcie3.0_faq/#EQ2) on 1 February 2014. Retrieved 1 May 2014.

48. "What does GT/s mean, anyway?" (http://www.tmworld.com/electronics-news/4380071/What-does-GT-s-mean-anyway-). *TM World*. Archived (https://web.archive.org/web/20120814002641/http://www.tmworld.com/electronics-news/4380071/What-does-GT-s-mean-anyway-) from the original on 14 August 2012. Retrieved 7 December 2012.

49. "Deliverable 12.2" (https://web.archive.org/web/20100817201815/http://www.eiscat.se/groups/EISCAT_3D_info/DeliverableWP12.2/preview_popup). SE: Eiscat. Archived from the original (http://www.eiscat.se/groups/EISCAT_3D_info/DeliverableWP12.2/preview_popup) on 17 August 2010. Retrieved 7 December 2012.

50. *PCI SIG* (http://www.pcisig.com/), archived (https://web.archive.org/web/20080706134414/http://pcisig.com/) from the original on 6 July 2008

51. "PCI Express Base 2.0 specification announced" (https://web.archive.org/web/20070304101327/http://www.pcisig.com/news_room/PCIe2_0_Spec_Release_FINAL2.pdf) (PDF) (Press release). PCI-SIG. 15 January 2007. Archived from the original (http://www.pcisig.com/news_room/PCIe2_0_Spec_Release_FINAL2.pdf) (PDF) on 4 March 2007. Retrieved 9 February 2007. — note that in this press release the term *aggregate bandwidth* refers to the sum of incoming and outgoing bandwidth; using this terminology the aggregate bandwidth of full duplex 100BASE-TX is 200 Mbit/s.

52. Smith, Tony (11 October 2006). "PCI Express 2.0 final draft spec published" (http://www.reghardware.co.uk/2006/10/11/pic-sig_posts_pcie_2_final_draft/). *The Register*. Archived (https://web.archive.org/web/20070129121731/http://www.reghardware.co.uk/2006/10/11/pic-sig_posts_pcie_2_final_draft/) from the original on 29 January 2007. Retrieved 9 February 2007.

53. Key, Gary; Fink, Wesley (21 May 2007). "Intel P35: Intel's Mainstream Chipset Grows Up" (http://www.anandtech.com/cpuch ipsets/showdoc.aspx?i=2993). AnandTech. Archived (https://web.archive.org/web/20070523055011/http://www.anandtech.c om/cpuchipsets/showdoc.aspx?i=2993) from the original on 23 May 2007. Retrieved 21 May 2007.

54. Huynh, Anh (8 February 2007). "NVIDIA "MCP72" Details Unveiled" (https://web.archive.org/web/20070210200616/http://w ww.dailytech.com/article.aspx?newsid=6021). AnandTech. Archived from the original (http://www.dailytech.com/article.asp x?newsid=6021) on 10 February 2007. Retrieved 9 February 2007.

55. "Intel P35 Express Chipset Product Brief" (http://download.intel.com/products/chipsets/P35/317304.pdf) (PDF). Intel. Archived (https://web.archive.org/web/20070926150158/http://download.intel.com/products/chipsets/P35/317304.pdf) (PDF) from the original on 26 September 2007. Retrieved 5 September 2007.

56. Hachman, Mark (5 August 2009). "PCI Express 3.0 Spec Pushed Out to 2010" (https://www.pcmag.com/article2/0,2817,2351266,00.asp). PC Mag. Archived (https://web.archive.org/web/20140107192535/http://www.pcmag.com/article2/0,2817,2351266,00.asp) from the original on 7 January 2014. Retrieved 7 December 2012.

57. "PCI Express 3.0 Bandwidth: 8.0 Gigatransfers/s" (http://www.extremetech.com/article2/0,1697,2169018,00.asp). ExtremeTech. 9 August 2007. Archived (https://web.archive.org/web/20071024140702/http://www.extremetech.com/article2/0,1697,2169018,00.asp) from the original on 24 October 2007. Retrieved 5 September 2007.

58. "PCI Special Interest Group Publishes PCI Express 3.0 Standard" (https://web.archive.org/web/20101121001048/http://www.xbitlabs.com/news/other/display/20101118151837_PCI_Special_Interest_Group_Publishes_PCI_Express_3_0_Standard.html). X bit labs. 18 November 2010. Archived from the original (http://www.xbitlabs.com/news/other/display/20101118151837_PCI_Special_Interest_Group_Publishes_PCI_Express_3_0_Standard.html) on 21 November 2010. Retrieved 18 November 2010.

59. "PCIe 3.1 and 4.0 Specifications Revealed" (http://www.eteknix.com/pcie-3-1-and-4-0-specifications-revealed/). *eteknix.com*. July 2013. Archived (https://web.archive.org/web/20160201102133/http://www.eteknix.com/pcie-3-1-and-4-0-specifications-revealed/) from the original on 1 February 2016.

60. "Trick or Treat… PCI Express 3.1 Released!" (http://blogs.synopsys.com/expressyourself/2014/11/12/trick-or-treat-pci-express-3-1-released/). *synopsys.com*. Archived (https://web.archive.org/web/20150323193106/https://blogs.synopsys.com/expressyourself/2014/11/12/trick-or-treat-pci-express-3-1-released/) from the original on 23 March 2015.

61. "PCI Express 4.0 evolution to 16 GT/s, twice the throughput of PCI Express 3.0 technology" (https://web.archive.org/web/20121223043627/http://www.pcisig.com/news_room/Press_Releases/November_29_2011_Press_Release_/) (press release). PCI-SIG. 29 November 2011. Archived from the original (http://www.pcisig.com/news_room/Press_Releases/November_29_2011_Press_Release_/) on 23 December 2012. Retrieved 7 December 2012.

62. "Frequently Asked Questions | PCI-SIG" (https://web.archive.org/web/20161020040106/https://pcisig.com/faq?field_category_value%5B%5D=pci_express_4.0). *pcisig.com*. Archived from the original (https://pcisig.com/faq?field_category_value%5B%5D=pci_express_4.0#4415) on 20 October 2016.

63. "PCIe 4.0 Heads to Fab, 5.0 to Lab" (http://www.eetimes.com/document.asp?doc_id=1330006). *EE Times*. 26 June 2016. Archived (https://web.archive.org/web/20160828221858/http://www.eetimes.com/document.asp?doc_id=1330006) from the original on 28 August 2016. Retrieved 27 August 2016.

64. "Mellanox Announces ConnectX-5, the Next Generation of 100G InfiniBand and Ethernet Smart Interconnect Adapter | NVIDIA" (https://www.mellanox.com/news/press_release/mellanox-announces-connectx-5-next-generation-100g-infiniband- and-ethernet-smart-interconnect). *www.mellanox.com*.

65. "Mellanox Announces 200Gb/s HDR InfiniBand Solutions Enabling Record Levels of Performance and Scalability | NVIDIA" (https://www.mellanox.com/news/press_release/mellanox-announces-200gbs-hdr-infiniband-solutions-enabling-re cord-levels-performance-and). *www.mellanox.com*.

66. "IDF: PCIe 4.0 läuft, PCIe 5.0 in Arbeit" (http://www.heise.de/newsticker/meldung/IDF-PCIe-4-0-laeuft-PCIe-5-0-in-Arbeit-32 97114.html). *Heise Online* (in German). 18 August 2016. Archived (https://web.archive.org/web/20160819153631/http://ww w.heise.de/newsticker/meldung/IDF-PCIe-4-0-laeuft-PCIe-5-0-in-Arbeit-3297114.html) from the original on 19 August 2016. Retrieved 18 August 2016.

67. Brian Thompto, POWER9 Processor for the Cognitive Era (https://old.hotchips.org/wp-content/uploads/hc_archives/hc28/H C28.23-Tuesday-Epub/HC28.23.90-High-Perform-Epub/HC28.23.921-.POWER9-Thompto-IBM-final.pdf)

68. 2016 IEEE Hot Chips 28 Symposium (HCS), 21–23 Aug. 2016 (https://ieeexplore.ieee.org/xpl/conhome/7932734/proceedin g)

69. Born, Eric (8 June 2017). "PCIe 4.0 specification finally out with 16 GT/s on tap" (https://techreport.com/news/32064/pcie-4-0 -specification-finally-out-with-16-gt-s-on-tap). Tech Report. Archived (https://web.archive.org/web/20170608155216/http://tec hreport.com/news/32064/pcie-4-0-specification-finally-out-with-16-gt-s-on-tap) from the original on 8 June 2017. Retrieved 8 June 2017.

70. "IBM Unveils Most Advanced Server for AI" (https://www-03.ibm.com/press/us/en/pressrelease/53452.wss). *www- 03.ibm.com*. 5 December 2017.

71. IBM Power System AC922 (8335-GTG) server helps you to harness breakthrough accelerated AI, HPDA, and HPC performance for faster time to insight, IBM Europe Hardware Announcement ZG17-0147 (https://www.ibm.com/common/ssi/ ShowDoc.wss?docURL=/common/ssi/rep_ca/7/877/ENUSZG17-0147/index.html)

72. "NETINT Introduces Codensity with Support for PCIe 4.0 – NETINT Technologies" (https://www.netint.ca/blog/netint-introdu ces-codensity-with-support-for-pcie-4-0/). *NETINT Technologies*. 17 July 2018. Retrieved 28 September 2018.

73. Mujtaba, Hassan (9 January 2019). "AMD Ryzen 3000 Series CPUs Based on Zen 2 Launching in Mid of 2019" (https://wcc ftech.com/amd-ryzen-3000-zen-2-desktop-am4-processors-launching-mid-2019/).

74. Alcorn, Paul (3 June 2019). "AMD Nixes PCIe 4.0 Support on Older Socket AM4 Motherboards, Here's Why" (https://www.to mshardware.com/news/amd-pcie-4.0-socket-am4-motherboard,39559.html). *Tom's Hardware*. Archived (https://archive.toda y/20190610025155/https://www.tomshardware.com/news/amd-pcie-4.0-socket-am4-motherboard,39559.html) from the original on 10 June 2019. Retrieved 10 June 2019.

75. Alcorn, Paul (10 January 2019). "PCIe 4.0 May Come to all AMD Socket AM4 Motherboards (Updated)" (https://www.tomsha rdware.com/news/amd-ryzen-pcie-4.0-motherboard,38401.html). *Tom's Hardware*. Archived (https://archive.today/20190610 025353/https://www.tomshardware.com/news/amd-ryzen-pcie-4.0-motherboard,38401.html) from the original on 10 June 2019. Retrieved 10 June 2019.

76. Cutress, Dr. Ian (13 August 2020). "Tiger Lake IO and Power" (https://www.anandtech.com/show/15971/intels-11th-gen-core -tiger-lake-soc-detailed-superfin-willow-cove-and-xelp/5). *Anandtech*.

77. "1,2,3,4,5... It's Official, PCIe 5.0 is Announced | synopsys.com" (https://blogs.synopsys.com/expressyourself/2017/08/15/1-2 -3-4-5-its-official-pcie-5-0-is-announced). *www.synopsys.com*. Retrieved 7 June 2017.

78. "PLDA Announces Availability of XpressRICH5™ PCIe 5.0 Controller IP | PLDA.com" (https://www.plda.com/plda-announce s-availability-xpressrich5tm-pcie-50-controller-ip). *www.plda.com*. Retrieved 28 June 2018.

79. "XpressRICH5 for ASIC | PLDA.com" (https://www.plda.com/products/pcie-solutions-asicsoc/pcie-controller-ip/pcie-soft-ip/p cie-40-soft-ip/xpressrich5-asic). *www.plda.com*. Retrieved 28 June 2018.

80. "Doubling Bandwidth in Under Two Years: PCI Express® Base Specification Revision 5.0, Version 0.9 is Now Available to Members" (http://pcisig.com/doubling-bandwidth-under-two-years-pci-express®-base-specification-revision-50-version-09-n ow). *pcisig.com*. Retrieved 12 December 2018.

81. "PCIe 5.0 Is Ready For Prime Time" (https://www.tomshardware.com/news/pcie-4.0-5.0-pci-sig-specification,38460.html). *tomshardware.com*. 17 January 2019. Retrieved 18 January 2019.

82. "PCI-SIG® Achieves 32GT/s with New PCI Express® 5.0 Specification" (https://www.businesswire.com/news/home/201905 29005766/en/PCI-SIG%C2%AE-Achieves-32GTs-with-New-PCI-Express%C2%AE-5.0-Specification). *www.businesswire.com*. 29 May 2019.

83. "PCI-Express 5.0: China stellt ersten Controller vor" (https://www.pcgameshardware.de/Mainboard-Hardware-154107/New s/PCI-Express50-China-stellt-ersten-Controller-vor-1337072/). *PC Games Hardware*. 18 November 2019.

84. IBM's POWER10 Processor, Hot Chips 32, August 16–18, 2020 (https://hc32.hotchips.org/assets/program/conference/day1/ HotChips2020_Server_Processors_IBM_Starke_POWER10_v33.pdf)

85. Power E1080 Enterprise server delivers a uniquely architected platform to help securely and efficiently scale core operational and AI applications in a hybrid cloud, IBM Europe Hardware Announcement ZG21-0059 (https://www.ibm.com/c ommon/ssi/rep_ca/9/877/ENUSZG21-0059/index.html)

86. IBM Power E1080 Technical Overview and Introduction (http://www.redbooks.ibm.com/redpapers/pdfs/redp5649.pdf)
87. "Intel Unveils 12th Gen Intel Core, Launches World's Best Gaming" (https://www.intel.com/content/www/us/en/newsroom/news/12th-gen-core-processors.html). Intel.com. Retrieved 16 February 2022.
88. "NVIDIA Announces Hopper Architecture, the Next Generation of Accelerated Computing" (https://nvidianews.nvidia.com/ne ws/nvidia-announces-hopper-architecture-the-next-generation-of-accelerated-computing).
89. "AMD Showcases Industry-Leading Gaming, Commercial, and Mainstream PC Technologies at COMPUTEX 2022" (https://www.amd.com/en/press-releases/2022-05-23-amd-showcases-growth-gaming-commercial-and-mainstream-mobile-and-ind ustry). AMD.com. Retrieved 23 May 2022.
90. "4th Gen AMD EPYC™ Processor Architecture" (https://www.amd.com/en/campaigns/epyc-9004-architecture). AMD.com. Retrieved 12 November 2022.
91. "PCI-SIG® Announces Upcoming PCI Express® 6.0 Specification to Reach 64 GT/s" (https://www.businesswire.com/news/home/20190618005945/en/PCI-SIG%C2%AE-Announces-Upcoming-PCI-Express%C2%AE-6.0-Specification-to-Reach-64-GTs). *www.businesswire.com*. 18 June 2019.
92. Smith, Ryan. "PCI Express Bandwidth to Be Doubled Again: PCIe 6.0 Announced, Spec to Land in 2021" (https://www.anandtech.com/show/14559/pci-express-bandwidth-to-be-doubled-again-pcie-60-announced-spec-to-land-in-2021). *www.anandtech.com*.
93. "PCI Express 6.0 Reaches Version 0.5 Ahead Of Finalization Next Year—Phoronix" (https://www.phoronix.com/scan.php?page=news_item&px=PCI-Express-6.0-v0.5). *www.phoronix.com*.
94. Shilov, Anton (4 November 2020). "PCIe 6.0 Specification Hits Milestone: Complete Draft Is Ready" (https://www.tomshardware.com/news/pcie-6-specification-hits-milestone-complete-draft-is-ready). *Tom's Hardware*.
95. Yanes, Al. "PCIe® 6.0 Specification, Version 0.9: One Step Closer to Final Release | PCI-SIG" (https://pcisig.com/blog/pcie®-60-specification-version-09-one-step-closer-final-release). *pcisig.com*. Retrieved 6 October 2021.
96. "PCI-SIG® Releases PCIe® 6.0 Specification Delivering Record Performance to Power Big Data Applications" (https://www.businesswire.com/news/home/20220111005011/en/PCI-SIG%C2%AE-Releases-PCIe%C2%AE-6.0-Specification-Deliv ering-Record-Performance-to-Power-Big-Data-Applications). Business Wire. 11 January 2022. Retrieved 16 February 2022.
97. "The Evolution of the PCI Express Specification: On its Sixth Generation, Third Decade and Still Going Strong" (https://pcisig.com/blog/evolution-pci-express-specification-its-sixth-generation-third-decade-and-still-going-strong). Pci-Sig. 11 January 2022. Retrieved 16 February 2022.
98. Debendra Das Sharma. "PCIe 6.0 Specification: The Interconnect for I/O Needs of the Future" (https://ghostarchive.org/varchive/youtube/20211030/jhehXwnu0Ss). PCI-SIG. p. 8. Archived from the original (https://www.youtube.com/watch?v=jhehX wnu0Ss) on 30 October 2021.
99. "Pushing the Envelope with PCIe 6.0: Bringing PAM4 to PCIe" (https://www.cadence.com/content/dam/cadence-www/global/en_US/documents/tools/ip/design-ip/pushing-the-envelope-with-pcie-6-wp.pdf) (PDF). Retrieved 16 February 2022.
100. "PowerPoint Presentation" (https://pcisig.com/sites/default/files/files/PCIe%206.0%20Webinar_Final_.pdf) (PDF). Retrieved 16 February 2022.
101. "PCI-SIG® Announces PCI Express® 7.0 Specification to Reach 128 GT/s" (https://www.businesswire.com/news/home/202 20621005137/en). Business Wire. 21 June 2022. Retrieved 25 June 2022.
102. "PLX demo shows PCIe over fiber as data center clustering interconnect" (http://www.cablinginstall.com/index/display/articl e-display/8876181966/articles/cabling-installation-maintenance/news/data-center/2011/6/plx-demo_shows_pcie.html). *Cabling install*. Penn Well. Retrieved 29 August 2012.
103. "Introduced second generation PCI Express Gen 2 over fiber optic systems" (http://www.adnaco.com/2011/09/03/new1). Adnaco. 22 April 2011. Archived (https://web.archive.org/web/20121004094357/http://www.adnaco.com/2011/09/03/new1/) from the original on 4 October 2012. Retrieved 29 August 2012.
104. "PCIe Active Optical Cable System" (http://www.samtec.com/cable-systems/active-optics/active-optical-cable/pcie.aspx). Archived (https://web.archive.org/web/20141230122105/http://www.samtec.com/cable-systems/active-optics/active-optical-cable/pcie.aspx) from the original on 30 December 2014. Retrieved 23 October 2015.
105. IBM Power Systems E870 and E880 Technical Overview and Introduction (http://www.redbooks.ibm.com/redpapers/pdfs/red p5137.pdf)
106. "Acer, Asus to Bring Intel's Thunderbolt Speed Technology to Windows PCs" (https://www.pcworld.com/article/240013/acer_asus_to_bring_intels_thunderbolt_speed_technology_to_windows_pcs.html). *PC World*. 14 September 2011. Archived (https://web.archive.org/web/20120118190546/http://www.pcworld.com/article/240013/acer_asus_to_bring_intels_thunderbolt_speed_technology_to_windows_pcs.html) from the original on 18 January 2012. Retrieved 7 December 2012.
107. Kevin Parrish (28 June 2013). "PCIe for Mobile Launched; PCIe 3.1, 4.0 Specs Revealed" (http://www.tomshardware.com/n ews/M-PCIe-M.2-PCIe-3.1-PCIe-4.0-OCuLink,23259.html). *Tom's Hardware*. Retrieved 10 July 2014.
108. "PCI Express 4.0 Draft 0.7 & PIPE 4.4 Specifications — What Do They Mean to Designers? — Synopsys Technical Article |

ChipEstimate.com" (https://www.chipestimate.com/PCI-Express-40-Draft-07-and-PIPE-44-Specifications-What-Do-They-Me an-to-Designers/Synopsys/Technical-Article/2017/02/21). *www.chipestimate.com*. Retrieved 28 June 2018.

109. "PCI Express 1x, 4x, 8x, 16x bus pinout and wiring @" (http://pinouts.ru/Slots/pci_express_pinout.shtml). RU: Pinouts. Archived (https://web.archive.org/web/20091125025800/http://pinouts.ru/Slots/pci_express_pinout.shtml) from the original on 25 November 2009. Retrieved 7 December 2009.

110. "PHY Interface for the PCI Express Architecture" (https://web.archive.org/web/20080317171752/http://download.intel.com/technology/pciexpress/devnet/docs/pipe2_00.pdf) (PDF) (version 2.00 ed.). Intel. Archived from the original (http://download.intel.com/technology/pciexpress/devnet/docs/pipe2_00.pdf) (PDF) on 17 March 2008. Retrieved 21 May 2008.

111. PCI Express System Architecture (https://www.mindshare.com/files/ebooks/PCI%20Express%20System%20Architecture.pd f)

112. PCI Express Architecture, intel.com (https://www.intel.ru/content/www/ru/ru/io/pci-express/pci-express-architecture-general.h tml)

113. "Mechanical Drawing for PCI Express Connector" (http://www.interfacebus.com/PCI-Express-Bus-PCIe-Description.html#d). Interface bus. Retrieved 7 December 2007.

114. "FCi schematic for PCIe connectors" (http://portal.fciconnect.com/Comergent/fci/drawing/10018783.pdf) (PDF). FCI connect. Retrieved 7 December 2007.

115. Reducing Interrupt Latency Through the Use of Message Signaled Interrupts (https://www.intel.com/content/dam/www/publi c/us/en/documents/white-papers/msg-signaled-interrupts-paper.pdf)

116. *PCI Express Base Specification, Revision 3.0* Table 4-24

117. "Computer Peripherals And Interfaces" (https://www.google.com/books?id=Xp7-NKsJ8_sC&pg=PA35&dq=frequent+enforced+acknowledgements/). Technical Publications Pune. Archived (https://web.archive.org/web/20140225203956/http://www.google.com/books?id=Xp7-NKsJ8_sC&pg=PA35&dq=frequent+enforced+acknowledgements%2F) from the original on 25 February 2014. Retrieved 23 July 2009.

118. Lawley, Jason (28 October 2014). "Understanding Performance of PCI Express Systems" (https://www.xilinx.com/support/do cumentation/white_papers/wp350.pdf) (PDF). 1.2. Xilinx.

119. "NVIDIA Introduces NVIDIA Quadro® Plex—A Quantum Leap in Visual Computing" (http://www.nvidia.com/object/IO_3452 7.html). *Nvidia*. 1 August 2006. Archived (https://web.archive.org/web/20060824225752/http://www.nvidia.com/object/IO_34 527.html) from the original on 24 August 2006. Retrieved 14 July 2018.

120. "Quadro Plex VCS—Advanced visualization and remote graphics" (http://www.nvidia.com/page/quadroplex.html). nVidia. Archived (https://web.archive.org/web/20110428042937/http://www.nvidia.com/page/quadroplex.html) from the original on 28 April 2011. Retrieved 11 September 2010.

121. "XGP" (https://web.archive.org/web/20100129170434/http://ati.amd.com/technology/xgp/). *ATI*. AMD. Archived from the original (http://ati.amd.com/technology/xgp/) on 29 January 2010. Retrieved 11 September 2010.

122. *Fujitsu-Siemens Amilo GraphicBooster External Laptop GPU Released* (http://www.ubergizmo.com/2008/12/fujitsu-siemens -amilo-graphicbooster-external-laptop-gpu-released/), 3 December 2008, archived (https://web.archive.org/web/20151016192734/http://www.ubergizmo.com/2008/12/fujitsu-siemens-amilo-graphicbooster-external-laptop-gpu-released/) from the original on 16 October 2015, retrieved 9 August 2015

123. *DynaVivid Graphics Dock from Acer arrives in France, what about the US?* (http://www.ubergizmo.com/2010/08/dynavivid-gr aphics-dock-from-acer-arrives-in-france-what-about-the-us/), 11 August 2010, archived (https://web.archive.org/web/201510 16192734/http://www.ubergizmo.com/2010/08/dynavivid-graphics-dock-from-acer-arrives-in-france-what-about-the-us/) from the original on 16 October 2015, retrieved 9 August 2015

124. Dougherty, Steve (22 May 2010), "MSI to showcase 'GUS' external graphics solution for laptops at Computex" (https://www.t weaktown.com/news/15382/msi_to_showcase_gus_external_graphics_solution_for_laptops_at_computex/), *TweakTown*

125. Hellstrom, Jerry (9 August 2011), "ExpressCard trying to pull a (not so) fast one?" (http://www.pcper.com/news/Editorial/Expr essCard-trying-pull-not-so-fast-one), *PC Perspective* (editorial), archived (https://web.archive.org/web/20160201160254/htt p://www.pcper.com/news/Editorial/ExpressCard-trying-pull-not-so-fast-one) from the original on 1 February 2016

126. "PE4H V3.2 (PCIe x16 Adapter)" (http://www.hwtools.net/Adapter/PE4H%20V3.2.html). Hwtools.net. Archived (https://web.a rchive.org/web/20140214012341/http://www.hwtools.net/Adapter/PE4H%20V3.2.html) from the original on 14 February 2014. Retrieved 5 February 2014.

127. O'Brien, Kevin (8 September 2010), "How to Upgrade Your Notebook Graphics Card Using DIY ViDOCK" (http://www.noteb ookreview.com/default.asp?newsID=5846&review=how+to+upgrade+laptop+graphics+notebook), *Notebook review*, archived (https://web.archive.org/web/20131213054647/http://www.notebookreview.com/default.asp?newsID=5846&review=how+to+upgrade+laptop+graphics+notebook) from the original on 13 December 2013

128. Lal Shimpi, Anand (7 September 2011), "The Thunderbolt Devices Trickle In: Magma's ExpressBox 3T" (http://www.anandtech.com/show/4743/the-thunderbolt-devices-trickle-in-magmas-expressbox-3t), *AnandTech*, archived (https://web.archive.org/web/20160304201352/http://www.anandtech.com/show/4743/the-thunderbolt-devices-trickle-in-magmas-expressbox-3t) from the original on 4 March 2016

129. "MSI GUS II external GPU enclosure with Thunderbolt" (https://www.theverge.com/2012/1/10/2698168/msi-GUS-II-external-thunderbolt-gpu-enclosure). *The Verge* (hands-on). 10 January 2012. Archived (https://web.archive.org/web/20120213123659/http://www.theverge.com/2012/1/10/2698168/msi-GUS-II-external-thunderbolt-gpu-enclosure) from the original on 13 February 2012. Retrieved 12 February 2012.

130. "PCI express graphics, Thunderbolt" (http://www.tomshardware.com/reviews/pci-express-graphics-thunderbolt,3263-2.html),

Tom's hardware, 17 September 2012

131. "M logics M link Thunderbold chassis no shipping" (https://www.engadget.com/2012/12/13/mlogics-mlink-thunderbolt-chass is-now-shipping-399/), *Engadget*, 13 December 2012, archived (https://web.archive.org/web/20170625182008/https://www.engadget.com/2012/12/13/mlogics-mlink-thunderbolt-chassis-now-shipping-399/) from the original on 25 June 2017

132. Burns, Chris (17 October 2017), "2017 Razer Blade Stealth and Core V2 detailed" (https://www.slashgear.com/2017-razer-b lade-stealth-and-core-v2-detailed-17504328/), *SlashGear*, archived (https://web.archive.org/web/20171017211631/https://www.slashgear.com/2017-razer-blade-stealth-and-core-v2-detailed-17504328/) from the original on 17 October 2017

133. "CompactFlash Association readies next-gen XQD format, promises write speeds of 125 MB/s and up" (https://www.engadget.com/2011/12/08/compactflash-association-readies-next-gen-xqd-format-promises-w/). Engadget. 8 December 2011. Archived (https://web.archive.org/web/20140519002107/http://www.engadget.com/2011/12/08/compactflash-association-rea dies-next-gen-xqd-format-promises-w/) from the original on 19 May 2014. Retrieved 18 May 2014.

134. Zsolt Kerekes (December 2011). "What's so very different about the design of Fusion-io's ioDrives / PCIe SSDs?" (http://ww w.storagesearch.com/ssd-29.html). storagesearch.com. Archived (https://web.archive.org/web/20130923065816/http://www.storagesearch.com/ssd-29.html) from the original on 23 September 2013. Retrieved 2 October 2013.

135. "Fusion-io ioDrive Duo Enterprise PCIe Review" (https://web.archive.org/web/20131004230049/http://www.storagereview.com/fusionio_iodrive_duo_enterprise_pcie_review). storagereview.com. 16 July 2012. Archived from the original (http://www.storagereview.com/fusionio_iodrive_duo_enterprise_pcie_review) on 4 October 2013. Retrieved 2 October 2013.

136. "OCZ Demos 4 TiB, 16 TiB Solid-State Drives for Enterprise" (https://web.archive.org/web/20130325121004/http://www.xbitlabs.com/news/storage/display/20120110180208_OCZ_Demos_4TB_16TB_Solid_State_Drives_for_Enterprise.html). X-bit labs. Archived from the original (http://www.xbitlabs.com/news/storage/display/20120110180208_OCZ_Demos_4TB_16TB_Solid_State_Drives_for_Enterprise.html) on 25 March 2013. Retrieved 7 December 2012.

137. "Enabling Higher Speed Storage Applications with SATA Express" (http://www.sata-io.org/technology/sataexpress.asp). SATA-IO. Archived (https://web.archive.org/web/20121127010238/http://www.sata-io.org/technology/sataexpress.asp) from the original on 27 November 2012. Retrieved 7 December 2012.

138. "SATA M.2 Card" (https://www.sata-io.org/sata-m2-card). SATA-IO. Archived (https://web.archive.org/web/20131003103042/ https://www.sata-io.org/sata-m2-card) from the original on 3 October 2013. Retrieved 14 September 2013.

139. "SCSI Express" (https://web.archive.org/web/20130127094133/http://www.scsita.org/library/scsi-express/). SCSI Trade Association. Archived from the original (http://www.scsita.org/library/scsi-express/) on 27 January 2013. Retrieved 27 December 2012.

140. Meduri, Vijay (24 January 2011). "A Case for PCI Express as a High-Performance Cluster Interconnect" (http://www.hpcwire. com/hpcwire/2011-01-24/a_case_for_pci_express_as_a_high-performance_cluster_interconnect.html). HPCwire. Archived (https://web.archive.org/web/20130114041356/http://www.hpcwire.com/hpcwire/2011-01-24/a_case_for_pci_express_as_a_high-performance_cluster_interconnect.html) from the original on 14 January 2013. Retrieved 7 December 2012.

141. Evan Koblentz (3 February 2017). "New PCI Express 4.0 delay may empower next-gen alternatives" (https://www.techrepublic.com/article/new-pci-express-4-0-delay-may-empower-next-gen-alternatives/). *Tech Republic*. Archived (https://web.archive.org/web/20170401143837/http://www.techrepublic.com/article/new-pci-express-4-0-delay-may-empower-next-gen-alternat ives/) from the original on 1 April 2017. Retrieved 31 March 2017.

142. Cutress, Ian. "CXL Specification 1.0 Released: New Industry High-Speed Interconnect From Intel" (https://www.anandtech.c om/show/14068/cxl-specification-1-released-new-industry-high-speed-interconnect-from-intel). *www.anandtech.com*. Retrieved 9 August 2019.

143. "Integrators List | PCI-SIG" (http://pcisig.com/developers/integrators-list). *pcisig.com*. Retrieved 27 March 2019.

Further reading

- Budruk, Ravi; Anderson, Don; Shanley, Tom (2003), Winkles, Joseph 'Joe' (ed.), *PCI Express System Architecture*, Mind share PC system architecture, Addison-Wesley, ISBN 978-0-321-15630-3, 1120 pp.
- Solari, Edward; Congdon, Brad (2003), *Complete PCI Express Reference: Design Implications for Hardware and Software Developers*, Intel, ISBN 978-0-9717861-9-6, 1056 pp.
- Wilen, Adam; Schade, Justin P; Thornburg, Ron (April 2003), *Introduction to PCI Express: A Hardware and Software Developer's Guide*, Intel, ISBN 978-0-9702846-9-3, 325 pp.

External links

- Media related to PCIe at Wikimedia Commons
- PCI-SIG Specifications (https://pcisig.com/specifications)

WIKIPEDIA

NVMe

NVM Express (**NVMe**) or **Non-Volatile Memory Host Controller Interface Specification** (**NVMHCIS**) is an open,logical-device interface specification for accessing a computer's non-volatile storage media usually attached via PCI Express (PCIe) bus. The initialism *NVM* stands for *non-volatile memory*, which is often NAND flash memory that comes in several physical form factors, including solid-state drives (SSDs), PCIe add-in cards, and M.2 cards, the successor to m SATA cards. NVM Express, as a logical-device interface, has been designed to capitalize on the low latency and internal parallelism of solid-state storage devices.[1]

Non-Volatile Memory Host Controller Interface Specification

Abbreviation	NVMe
Year started	2011
Organization	NVM Express Work Group (incorporated as NVM Express in 2014)
Website	nvmexpress.org (https://nvmexpress.o

Architecturally, the logic for NVMe is physically stored within and executed by the NVMe controller chip that is physically co-located with the storage media, usually an SSD. Version changes for NVMe, e.g., 1.3 to 1.4, are incorporated within the storage media, and do not affect PCIe-compatible components such as motherboards and CPUs.[2]

By its design, NVM Express allows host hardware and software to fully exploit the levels of parallelism possible in modern SSDs. As a result, NVM Express reduces I/O overhead and brings various performance improvements relative to previous logical-device interfaces, including multiple long command queues, and reduced latency. The previous interface protocols like AHCI were developed for use with far slower harddisk drives (HDD) where a very lengthy delay (relative to CPU operations) exists between a request and data transfer, where data speeds are much slower than RAM speeds, and where disk rotation and s eek time give rise to further optimization requirements.

NVM Express devices are chiefly available in the form of standard-sized PCI Express expansion cards[3] and as 2.5- inch form-factor devices that provide a four-lane PCI Express interface through the U.2 connector (formerly known as SFF-8639).[4][5] Storage devices using SATA Express and the M.2 specification which support NVM Express as the logical-device interface are a popular use-case for NVMe and have become the dominant form of solid-state storage for servers, desktops, and laptops alike.[6][7]

Specifications

Specifications for NVMe released to date include:[8]

- 1.0e (January 2013)
- 1.1b (July 2014)
- 1.2 (November 2014)
  - 1.2a (October 2015)
  - 1.2b (June 2016)
  - 1.2.1 (June 2016)
- 1.3 (May 2017)
  - 1.3a (October 2017)
  - 1.3b (May 2018)
  - 1.3c (May 2018)
  - 1.3d (March 2019)
- 1.4 (June 2019)
  - 1.4a (March 2020)
  - 1.4b (September 2020)
- 2.0 (May 2021)[9]
  - 2.0a (July 2021)
  - 2.0b (January 2022)
  - 2.0c (October 2022)

Background

Historically, most SSDs used buses such as SATA, SAS or Fibre Channel for interfacing with the rest of a computer system. Since SSDs became available in mass markets, SATA has become the most typical way for connecting SSDs in

personal computers; however, SATA was designed primarily for interfacing with mechanical hard disk drives (HDDs), and it became increasingly inadequate for SSDs, which improved in speed over time.[10] For example, within about five years of mass market mainstream adoption (2005–2010) many SSDs were already held back by the comparatively slow data rates available for hard drives—unlike hard disk drives, some SSDs are limited by the maximum throughput of SATA.

High-end SSDs had been made using the PCI Express bus before NVMe, but using non-standard specification interfaces. By standardizing the interface of SSDs, operating systems only need one common device driver to work with all SSDs adhering to the specification. It also means that each SSD manufacturer does not have to design specific interface drivers. This is similar to how USB mass storage devices are built to follow the USB mass-storage device class specification and work with all computers, with no per-device drivers needed.[11]

Intel SSD 750 series, an SSD that uses NVM Express, in form of a PCI Express 3.0 ×4 expansion card (front and rear views)

NVM Express devices are also used as the building block of the burst buffer storage in many leading supercomputers, such as Fugaku Supercomputer, Summit Supercomputer and Sierra Supercomputer, etc.[12][13]

History

The first details of a new standard for accessing non-volatile memory emerged at the Intel Developer Forum 2007, when NVMHCI was shown as the host-side protocol of a proposed architectural design that had Open NAND Flash Interface Working Group (ONFI) on the memory (flash) chips side.[14] A NVMHCI working group led by Intel was formed that year. The NVMHCI 1.0 specification was completed in April 2008 and released on Intel's web site.[15][16][17]

Technical work on NVMe began in the second half of 2009.[18] The NVMe specifications were developed by the NVM Express Workgroup, which consists of more than 90 companies; Amber Huffman of Intel was the working group's chair. Version 1.0 of the specification was released on 1 March 2011,[19] while version 1.1 of the specification was released on 11 October 2012.[20] Major features added in version 1.1 are multi-path I/O (with namespace sharing) and arbitrary-length scatter-gather I/O. It is expected that future revisions will significantly enhance namespace management.[18] Because of its feature focus, NVMe 1.1 was initially called "Enterprise NVMHCI".[21] An update for the base NVMe specification, called version 1.0e, was released in January 2013.[22] In June 2011, a Promoter Group led by seven companies was formed.

The first commercially available NVMe chipsets were released by Integrated Device Technology (89HF16P04AG3 and 89HF32P08AG3) in August 2012.[23][24] The first NVMe drive, Samsung's XS1715 enterprise drive, was announced in July 2013; according to Samsung, this drive supported 3 GB/s read speeds, six times faster than their previous enterprise offerings.[25] The LSI SandForce SF3700 controller family, released in November 2013, also supports NVMe.[26][27] A Kingston HyperX "prosumer" product using this controller was showcased at the Consumer Express products, the Intel SSD data center family that interfaces with the host through PCI Express bus, which includes the DC P3700 series, the DC P3600 series, and the DC P3500 series.[30] As of November 2014, NVMe drives are commercially available.

In March 2014, the group incorporated to become NVM Express, Inc., which as of November 2014 consists of more than 65 companies from across the industry. NVM Express specifications are owned and maintained by NVM Express, Inc., which also promotes industry awareness of NVM Express as an industry-wide standard. NVM Express, Inc. is directed by a thirteen-member board of directors selected from the Promoter Group, which includes Cisco, Dell, EMC, HGST, Intel, Micron, Microsoft, NetApp, Oracle, PMC, Samsung, SanDisk and Seagate.[31]

In September 2016, the CompactFlash Association announced that it would be releasing a new memory card specification, CFexpress, which uses NVMe.

NVMe Host Memory Buffer (HMB) added in version 1.2 of the NVMe specification.[32] HMB allows SSDs to utilize the host's DRAM, which can improve the I/O performance for DRAM-less SSDs.[33]

Form factors

There are many form factors of NVMe solid-state drive, such as AIC, U.2, U.3, M.2 etc.

AIC (add-in card)

Almost all early NVMe solid-state drives are HHHL (half height, half length) or FHHL (full height, half length) AIC, with a PCIe 2.0 or 3.0 interface. A HHHL NVMe solid-state drive card is easy to insert into a PCIe slot of a server.

U.2 (SFF-8639)

U.2, formerly known as **SFF-8639**, is a computer interface for connecting solid-state drives to a computer. It uses up to four PCI Express lanes. Available servers can combine up to 48 U.2 NVMe solid-state drives.[34]

U.3 (SFF-8639 or SFF-TA-1001)

U.3 is built on the U.2 spec and uses the same SFF-8639 connector. It is a 'tri-mode' standard, combining SAS, SATA and NVMe support into a single controller. U.3 can also support hot-swap between the different drives where firmware support is available. U.3 drives are still backward compatible with U.2, but U.2 drives are not compatible with U.3 hosts.

M.2

M.2, formerly known as the **Next Generation Form Factor** (**NGFF**), uses a M.2 NVMe solid-state drive computer bus. Interfaces provided through the M.2 connector are PCI Express 3.0 or PCI Express 4.0 (up to four lanes).

EDSFF

Main article: Enterprise and Data Center Standard Form Factor

NVMe-oF

NVM Express over Fabrics (NVMe-oF) is the concept of using a transport protocol over a network to connect remote NVMe devices, contrary to regular NVMe where physical NVMe devices are connected to a PCIe bus either directly or over a PCIe switch to a PCIe bus. In August 2017, a standard for using NVMe over Fibre Channel (FC) was submitted by the standards organization International Committee for Information Technology Standards (ICITS), and this combination is often referred to as FC-NVMe or sometimes NVMe/FC.[35]

As of May 2021, supported NVMe transport protocols are:

- FC, FC-NVMe[35][36]
- TCP, NVMe/TCP[37]
- Ethernet, RoCE v1/v2 (RDMA over converged Ethernet)[38]
- InfiniBand, NVMe over InfiniBand or NVMe/IB[39]

The standard for NVMe over Fabrics was published by NVM Express, Inc. in 2016.[40][41]

The following software implements the NVMe-oF protocol:

- Linux NVMe-oF initiator and target.[42] RoCE transport was supported initially, and with Linux kernel 5.x, native support for TCP was added.[43]
- Storage Performance Development Kit (SPDK) NVMe-oF initiator and target drivers.[44] Both RoCE and TCP transports are supported.[45][46]
- Starwind NVMe-oF initiator and target for Microsoft Windows, supporting both RoCE and TCP transports.[47]

Comparison with AHCI

The Advanced Host Controller Interface (AHCI) has the benefit of wide software compatibility, but has the downside of not delivering optimal performance when used with SSDs connected via the PCI Express bus. As a logical-device interface, AHCI was developed when the purpose of a host bus adapter (HBA) in a system was to connect the CPU/memory subsystem with a much slower storage subsystem based on rotating magnetic media. As a result, AHCI introduces certain inefficiencies when used with SSD devices, which behave much more like RAM than like spinning media.[6]

The NVMe device interface has been designed from the ground up, capitalizing on the lower latency and parallelism of PCI Express SSDs, and complementing the parallelism of contemporary CPUs, platforms and applications. At a high level, the basic advantages of NVMe over AHCI relate to its ability to exploit parallelism in host hardware and software, manifested by the differences in command queue depths, efficiency of interrupt processing, the number of uncacheable register accesses, etc., resulting in various performance improvements.[6][48]: 17–18

The table below summarizes high-level differences between the NVMe and AHCI logical-device interfaces.

High-level comparison of AHCI and NVMe[6]

	AHCI	NVMe
Maximum queue depth	One command queue; Up to 32 commands per queue	Up to 65535 queues;[49] Up to 65536 commands per queue
Uncacheable register accesses (2000 cycles each)	Up to six per non-queued command; Up to nine per queued command	Up to two per command
Interrupt	A single interrupt	Up to 2048 MSI-X interrupts
Parallelism and multiple threads	Requires synchronization lock to issue a command	No locking
Efficiency for 4 KB commands	Command parameters require two serialized host DRAM fetches	Gets command parameters in one 64-byte fetch
Data transmission	Usually half-duplex	Full-duplex

Operating system support

ChromeOS

On February 24, 2015, support for booting from NVM Express devices was added to ChromeOS.[51][52]

DragonFly BSD

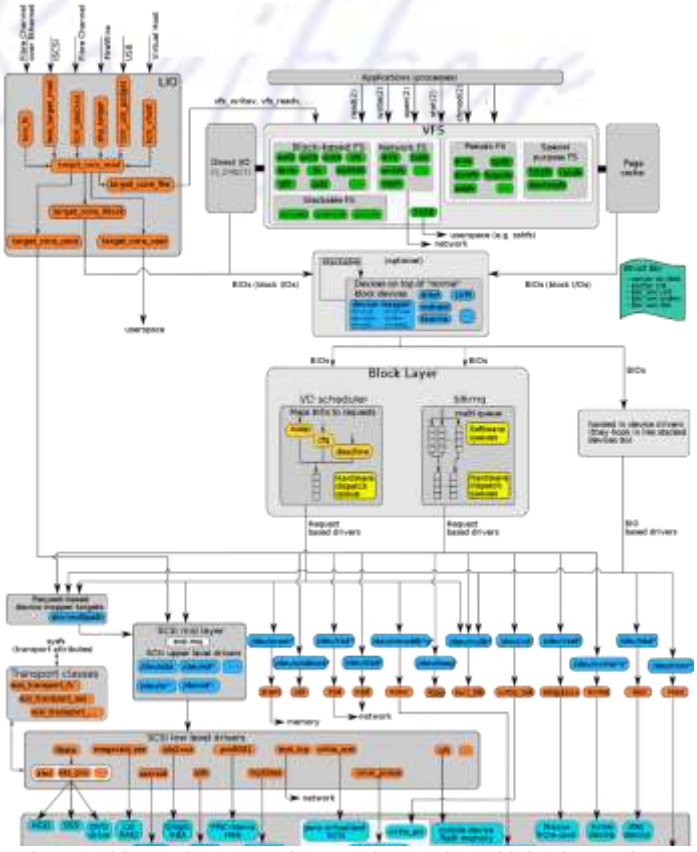
The first release of DragonFly BSD with NVMe support is version 4.6.[53]

FreeBSD

Intel sponsored a NVM Express driver for FreeBSD's head and stable/9 branches.[54][55] The nvd(4) and nvme(4) drivers are included in the GENERIC kernel configuration by default since FreeBSD version 10.2 in 2015.[56]

Genode

Support for consumer-grade NVMe was added to the Genode framework as part of the 18.05[57] release.



The position of NVMe data paths and multiple internal queues within various layers of the Linux kernel's storage stack.[50]

Haiku

Haiku gained support for NVMe on April 18, 2019.[58][59]

illumos

illumos received support for NVMe on October 15, 2014.[60]

iOS

With the release of the iPhone 6S and 6S Plus, Apple introduced the first mobile deployment of NVMe over PCIe in smartphones.[61] Apple followed these releases with the release of the first- generation iPad Pro and first-generation iPhone SE that also use NVMe over PCIe.[62]

Linux

Intel published an NVM Express driver for Linux on 3 March 2011,[63][64][65] which was merged into the Linux kernel mainline on 18 January 2012 and released as part of version 3.3 of the Linux kernel on 19 March 2012.[66] Linux supports NVMe Host Memory Buffer[67] from version 4.13.1[68] with default maximum size 128 MB.[69]

macOS

Apple introduced software support for NVM Express in Yosemite 10.10.3. The NVMe hardware interface was introduced in the 2016 MacBook and MacBook Pro.[70]

NetBSD

NetBSD added support for NVMe in NetBSD 8.0.[71] The implementation is derived from OpenBSD 6.0.

OpenBSD

Development work required to support NVMe in OpenBSD has been started in April 2014 by a senior developer formerly responsible for USB 2.0 and AHCI support.[72] Support for NVMe has been enabled in the OpenBSD 6.0 release.[73]

OS/2

Arca Noae provides an NVMe driver for ArcaOS, as of April, 2021. The driver requires advanced interrupts as provided by the ACPI PSD running in advanced interrupt mode (mode 2), thus requiring the SMP kernel, as well.[74]

Solaris

Solaris received support for NVMe in Oracle Solaris 11.2.[75]

VMware

Intel has provided an NVMe driver for VMware,[76] which is included in vSphere 6.0 and later builds, supporting various NVMe devices.[77] As of vSphere 6 update 1, VMware's VSAN software-defined storage subsystem also supports NVMe devices.[78]

Windows

Microsoft added native support for NVMe to Windows 8.1 and Windows Server 2012 R2.[48][79] Native drivers for Windows 7 and Windows Server 2008 R2 have been added in updates.[80] Many vendors have released their own

Windows drivers for their devices as well. There are also manually customized installer files available to install a specific vendor's driver to any NVMe card, such as using a Samsung NVMe driver with a non-Samsung NVMe device, which may be needed for additional features, performance, and stability.[81] Support for NVMe HMB was added in Windows 10 Anniversary Update (Version 1607) in 2016.[32]

The OpenFabrics Alliance maintains an open-source NVMe Windows Driver for Windows 7/8/8.1 and Windows Server 2008R2/2012/2012R2, developed from the baseline code submitted by several promoter companies in the NVMe workgroup, specifically IDT, Intel, and LSI.[82] The current release is 1.5 from December 2016.[83]

Software support

QEMU

NVMe is supported by QEMU since version 1.6 released on August 15, 2013.[84] NVMe devices presented to QEMU guests can be either real or emulated.

UEFI

An open source NVMe driver for UEFI is available on SourceForge.[85]

Management tools

nvmecontrol

The nvmecontrol tool is used to control an NVMe disk from the command line on FreeBSD. It was added in FreeBSD 9.2.[86]

nvme-cli

NVM-Express user space tooling for Linux.[87]

nvme-cli on Linux

See also

- M.2
- PCI Express
- SATA Express
- Solid-state drive
- Universal Flash Storage (UFS)

References

1. **"NVM Express" (https://nvmexpress.org/). NVM Express, Inc. Retrieved 2017-01-24. "NVMe is designed from the ground up to deliver high bandwidth and low latency storage access for current and future NVM technologies."**

2. "NVMe 1.4 Specification Published: Further Optimizing Performance and Reliability" (https://web.archive.org/web/20210127014339/https://www.anandtech.com/show/14543/nvme-14-specification-published). *www.anandtech.com*. Archived from the original (https://www.anandtech.com/show/14543/nvme-14-specification-published.) on 2021-01-27.
3. Drew Riley (2014-08-13). "Intel SSD DC P3700 800GB and 1.6TB Review: The Future of Storage" (http://www.tomshardware.com/reviews/intel-ssd-dc-p3700-nvme,3858-3.html). *tomshardware.com*. Retrieved 2014-11-21.
4. "Intel Solid-State Drive DC P3600 Series" (http://www.intel.com/content/dam/www/public/us/en/documents/product-specifications/ssd-dc-p3600-spec.pdf) (PDF). Intel. 2015-03-20. pp. 18, 20–22. Retrieved 2015-04-11.
5. Paul Alcorn (2015-06-05). "SFFWG Renames PCIe SSD SFF-8639 Connector To U.2" (http://www.tomshardware.com/news/sff-8639-u.2-pcie-ssd-nvme,29321.html). Tom's Hardware. Retrieved 2015-06-09.
6. Dave Landsman (2013-08-09). "AHCI and NVMe as Interfaces for SATA Express Devices – Overview" (https://www.sata-io.org/sites/default/files/documents/NVMe%20and%20AHCI%20as%20SATA%20Express%20Interface%20Options%20-%20Whitepaper_.pdf) (PDF). SATA-IO. Retrieved 2013-10-02.
7. Paul Wassenberg (2013-06-25). "SATA Express: PCIe Client Storage" (https://web.archive.org/web/20131004222635/https://www.sata-io.org/sites/default/files/documents/SATA%20Express%20-%20CS%202013.pdf) (PDF). SATA-IO. Archived from the original (https://www.sata-io.org/sites/default/files/documents/SATA%20Express%20-%20CS%202013.pdf) (PDF) on 2013-10-04. Retrieved 2014-11-21.
8. NVMe Specifications (https://nvmexpress.org/developers/nvme-specification/)
9. NVM Express Announces the Rearchitected NVMe 2.0 Library of Specifications (https://nvmexpress.org/nvm-express-announces-the-rearchitected-nvme-2-0-library-of-specifications/)
10. Walker, Don H. "A Comparison of NVMe and AHCI" (https://web.archive.org/web/20190212011912/http s://sata-io.org/sites/default/files/documents/NVMe%20and%20AHCI_%20_long_.pdf) (PDF). *31 July 2012*. SATA-IO. Archived from the original (https://www.sata-io.org/sites/default/files/documents/NVMe%20and%20AHCI_%20_long_.pdf) (PDF) on 12 February 2019. Retrieved 3 July 2013.
11. "NVM Express Explained" (https://nvmexpress.org/wp-content/uploads/2013/04/NVM_whitepaper.pdf) (PDF). *nvmexpress.org*. 9 April 2014. Retrieved 21 March 2015.
12. "Using LC's Sierra Systems" (https://hpc.llnl.gov/training/tutorials/using-lcs-sierra-system). *hpc.llnl.gov*. Retrieved 2020-06-25.
13. "SummitDev User Guide" (https://web.archive.org/web/20200806000635/https://docs.olcf.ornl.gov/systems/summitdev_user_guide.html). *olcf.ornl.gov*. Archived from the original (https://docs.olcf.ornl.gov/sy stems/summitdev_user_guide.html) on 2020-08-06. Retrieved 2020-06-25.
14. "Speeding up Flash... in a flash" (https://web.archive.org/web/20090918093831/http://www.theinquirer.net/inquirer/news/1018710/speeding-flash-flash). The Inquirer. 2007-10-13. Archived from the original (http://www.theinquirer.net/inquirer/news/1018710/speeding-flash-flash) on September 18, 2009. Retrieved 2014-01-11.
15. http://www.bswd.com/FMS09/FMS09-T2A-Huffman.pdf

16. "Flash new standard tips up" (https://web.archive.org/web/20140111131722/http://www.theinquirer.net/inquirer/news/1018442/nvram-standard-tips). The Inquirer. 2008-04-16. Archived from the original (http://www.theinquirer.net/inquirer/news/1018442/nvram-standard-tips) on January 11, 2014. Retrieved 2014-01-11.
17. http://www.flashmemorysummit.com/English/Collaterals/Proceedings/2008/20080813_T2A_Huffman.pdf
18. http://www.flashmemorysummit.com/English/Collaterals/Proceedings/2013/20130813_A12_Onufryk.pdf
19. "New Promoter Group Formed to Advance NVM Express" (https://nvmexpress.org/wp-content/uploads/ 2013/04/NVMe_Press_Release_New-Promoter-Group_20110601.pdf) (PDF). *Press release*. June 1, 2011. Retrieved September 18, 2013.
20. Amber Huffman, ed. (October 11, 2012). "NVM Express Revision 1.1" (https://nvmexpress.org/wp-conte nt/uploads/2013/05/NVM_Express_1_1.pdf) (PDF). *Specification*. Retrieved September 18, 2013.
21. David A. Deming (2013-06-08). "PCIe-based Storage" (https://web.archive.org/web/20130920064556/http://snia.org/sites/default/files2/SPDEcon2013/presentations/Storage%20Plumbing/DavidDeming_PCIe-based_Storage_r1.pdf) (PDF). *snia.org*. Archived from the original (http://snia.org/sites/default/files2/SPDEcon2013/presentations/Storage%20Plumbing/DavidDeming_PCIe-based_Storage_r1.pdf) (PDF) on 2013-09-20. Retrieved 2014-01-12.
22. Amber Huffman, ed. (January 23, 2013). "NVM Express Revision 1.0e" (https://nvmexpress.org/wp-cont ent/uploads/2013/04/NVM_10e_specification.pdf) (PDF). *Specification*. Retrieved September 18, 2013.
23. "IDT releases two NVMe PCI-Express SSD controllers" (https://web.archive.org/web/20120824032335/http://www.theinquirer.net/inquirer/news/2200157/idt-releases-two-nvme-pciexpress-ssd-controllers). The Inquirer. 2012-08-21. Archived from the original (http://www.theinquirer.net/inquirer/news/2200157/idt-releases-two-nvme-pciexpress-ssd-controllers) on August 24, 2012. Retrieved 2014-01-11.
24. "IDT Shows Off The First NVMe PCIe SSD Processor and Reference Design - FMS 2012 Update" (http://www.thessdreview.com/daily-news/latest-buzz/idt-shows-off-the-first-nvme-pcie-ssd-processor-and-reference-design-fms-2012-update/). The SSD Review. 2012-08-24. Retrieved 2014-01-11.
25. "Samsung Announces Industry's First 2.5-inch NVMe SSD | StorageReview.com - Storage Reviews" (https://web.archive.org/web/20140110200756/http://www.storagereview.com/samsung_announces_industry_s_first_25inch_nvme_ssd). *StorageReview.com*. 2013-07-18. Archived from the original (http://www.storagereview.com/samsung_announces_industry_s_first_25inch_nvme_ssd) on 2014-01-10. Retrieved 2014-01-11.
26. "LSI SF3700 SandForce Flash Controller Line Unveiled | StorageReview.com - Storage Reviews" (https://web.archive.org/web/20140111022326/http://www.storagereview.com/lsi_sf3700_sandforce_flash_controller_line_unveiled). *StorageReview.com*. 2013-11-18. Archived from the original (http://www.storagereview.com/lsi_sf3700_sandforce_flash_controller_line_unveiled) on 2014-01-11. Retrieved 2014-01-11.
27. "LSI Introduces Blazing Fast SF3700 Series SSD Controller, Supports Both PCIe and SATA 6Gbps" (https://web.archive.org/web/20160305231734/http://hothardware.com/news/lsi-introduces-blazing-fast-sf3700-series-ssd-controller-supports-both-pcie-and-sata-6gbps). *hothardware.com*. Archived from the original (https://hothardware.com/News/LSI-Introduces-Blazing-Fast-SF3700-Series-SSD-Controller-S

upports-Both-PCIe-and-SATA-6Gbps/) on 5 March 2016. Retrieved 21 March 2015.

28. Jane McEntegart (7 January 2014). "Kingston Unveils First PCIe SSD: 1800 MB/s Read Speeds" (htt p://www.tomshardware.com/news/kingston-pcie-ssd,25600.html). *Tom's Hardware*. Retrieved 21 March 2015.

29. "Kingston HyperX Predator PCI Express SSD Unveiled With LSI SandForce SF3700 PCIe Flash Controller" (https://web.archive.org/web/20160528221546/http://hothardware.com/news/kingston-hyper x-predator-pci-express-ssd-unveiled--with-lsi-sandforce-sf3700-flash-controller). *hothardware.com*. Archived from the original (https://hothardware.com/News/Kingston-HyperX-Predator-PCI-Express-SSD-Unveiled--With-LSI-Sandforce-SF3700-Flash-Controller/) on 28 May 2016. Retrieved 21 March 2015.

30. "Intel® Solid-State Drive Data Center Family for PCIe*" (http://www.intel.com/content/www/us/en/solid- state-drives/intel-ssd-dc-family-for-pcie.html). *Intel*. Retrieved 21 March 2015.

31. "NVM Express » NVM Express Organization History" (https://web.archive.org/web/20151123014820/ht tp://www.nvmexpress.org/about/company-history/). *nvmexpress.org*. Archived from the original (https://n vmexpress.org/about/company-history/) on 23 November 2015. Retrieved 23 December 2015.

32. The Toshiba RC100 SSD Review: Tiny Drive In A Big Market (https://www.anandtech.com/show/1281 9/the-toshiba-rc100-ssd-review)

33. Kim K, Kim T (2020) HMB in DRAM-less NVMe SSDs: Their usage and effects on performance. PLOS ONE 15(3): e0229645. (https://doi.org/10.1371/journal.pone.0229645)

34. "All-Flash NVME Servers for Advanced Computing Supermicro" (https://www.supermicro.com/en/produ cts/nvme). Supermicro. Retrieved 2022-07-22.

35. "NVMe over Fibre Channel (NVMe over FC) or FC-NVMe standard" (https://searchstorage.techtarget.c om/definition/NVMe-over-FC-Nonvolatile-Memory-Express-over-Fibre-Channel). *Tech Target*. January 1, 2018. Retrieved May 26, 2021.

36. "FC-NVMe rev 1.14 (T11/16-020vB)" (https://standards.incits.org/apps/group_public/download.php/873 64/T11-2017-00020-v003.pdf) (PDF). *INCITS*. April 19, 2017. Retrieved May 26, 2021.

37. "NVMe-oF Specification" (https://nvmexpress.org/developers/nvme-of-specification/). *NVMexpress*. Retrieved May 26, 2021.

38. "Supplement to InfiniBandTMArchitecture Specification Volume 1 Release 1.2.1" (https://cw.infinibandt a.org/document/dl/7148). *Infiniband*. September 2, 2014. Retrieved May 26, 2021.

39. "What is NVMe-oF?" (https://www.storagereview.com/review/nvme-nvme-of-background-overview). *Storage Review*. June 27, 2020. Retrieved May 26, 2021.

40. "NVM Express over Fabrics Revision 1.0" (https://nvmexpress.org/wp-content/uploads/NVMe_over_Fa brics_1_0_Gold_20160605.pdf) (PDF). NVM Express, Inc. 5 June 2016.

41. Woolf, David (February 9, 2018). "What NVMe over Fabrics Means for Data Storage" (https://www.netw orkcomputing.com/storage/what-nvme-over-fabrics-means-data-storage/1066956182).

42. Hellwig, Christoph (July 17, 2016). "NVMe Over Fabrics Support in Linux"

(https://events.static.linuxfou nd.org/sites/events/files/slides/nvme-over-fabrics.pdf) (PDF).
43. Petros Koutoupis (June 10, 2019). "Data in a Flash, Part III: NVMe over Fabrics Using TCP" (https://ww w.linuxjournal.com/content/data-flash-part-iii-nvme-over-fabrics-using-tcp). *Linux Journal*. Retrieved May 26, 2021.
44. Stern, Jonathan (7 June 2016). "Announcing the SPDK NVMf Target" (http://www.spdk.io/feature/2016/ 06/07/announce-nvmf/).
45. "SPDKNVMe-oFRDMA (Target & Initiator) Performance Report" (https://ci.spdk.io/download/performan ce-reports/SPDK_rdma_perf_report_2101.pdf) (PDF). *SPDK*. February 1, 2021. Retrieved May 26, 2021.
46. "SPDKNVMe-oFTCP (Target & Initiator) Performance Report" (https://ci.spdk.io/download/performance -reports/SPDK_tcp_perf_report_2101.pdf) (PDF). *SPDK*. February 1, 2020. Retrieved May 26, 2021.
47. "How is NVMe-oF doing? Part 3: StarWind NVMe-oF Initiator + Linux SPDK NVMe-oF Target" (https:// www.hyper-v.io/nvme-part-3-starwind-nvme-initiator-linux-spdk-nvme-target/). *Hyper-V Blog*. August 12, 2019. Retrieved May 26, 2021.
48. Andy Herron (2013). "Advancements in Storage and File Systems in Windows 8.1" (https://web.archiv e.org/web/20140110193117/http://snia.org/sites/default/files2/SDC2013/presentations/FileSystems/AndyHeron_Enhancements_To_Win81_Storage.pdf) (PDF). *snia.org*. Archived from the original (http://snia.org/sites/default/files2/SDC2013/presentations/FileSystems/AndyHeron_Enhancements_To_Win81_Storage.pdf) (PDF) on 2014-01-10. Retrieved 2014-01-11.
49. Amber Huffman (March 9, 2020). "NVM Express Base Specification Revision 1.4a" (https://nvmexpres s.org/wp-content/uploads/NVM-Express-1_4a-2020.03.09-Ratified.pdf) (PDF). *Specification*. section 1.2 Theory of Operation, p. 7. Retrieved May 16, 2020.
50. Werner Fischer; Georg Schönberger (2015-06-01). "Linux Storage Stack Diagram" (https://www.thoma s-krenn.com/en/wiki/Linux_Storage_Stack_Diagram). Thomas-Krenn.AG. Retrieved 2015-06-08.
51. "NVM Express » ChromeOS adds boot support for NVM Express" (https://nvmexpress.org/blog/chrome os-adds-boot-support-for-nvm-express/). *nvmexpress.org*. Retrieved 21 March 2015.
52. "4f503189f7339c667b045ab80a949964ecbaf93e - chromiumos/platform/depthcharge - Git at Google" (https://chromium.googlesource.com/chromiumos/platform/depthcharge/+/4f503189f7339c667b045ab8 0a949964ecbaf93e). *googlesource.com*. Retrieved 21 March 2015.
53. "DragonFly BSD 4.6" (https://www.dragonflybsd.org/release46/). *www.dragonflybsd.org*. Retrieved 2016-09-08.
54. "Log of /head/sys/dev/nvme" (http://svnweb.freebsd.org/base/head/sys/dev/nvme/?view=log). *FreeBSD source tree*. The FreeBSD Project. Retrieved 16 October 2012.
55. "Log of /stable/9/sys/dev/nvme" (http://svnweb.freebsd.org/base/stable/9/sys/dev/nvme/?view=log).

FreeBSD source tree. The FreeBSD Project. Retrieved 3 July 2013.

56. "FreeBSD 10.2-RELEASE Release Notes" (https://www.freebsd.org/releases/10.2R/relnotes.html#kern el-config). The FreeBSD Project. Retrieved 5 August 2015.

57. "Release notes for the Genode OS Framework 18.05" (https://genode.org/documentation/release-note s/18.05#NVMe_storage_devices). *genode.org*.

58. "#9910 NVMe devices support" (https://dev.haiku-os.org/ticket/9910). *dev.haiku-os.org*. Retrieved 2019-04-18.

59. "NVMe Driver Now Available - Haiku Project" (https://www.haiku-os.org/blog/kallisti5/2019-04-16_nvm e_driver_now_available/). *www.haiku-os.org*. Retrieved 2016-07-28.

60. "4053 Add NVME Driver Support to Illumos" (https://github.com/illumos/illumos-gate/commit/3c9168fa8 e9c30d55b3aa2fde74bd7da46df53f5). *github.com*. Retrieved 2016-05-23.

61. Ho, Joshua (September 28, 2015). "iPhone 6s and iPhone 6s Plus Preliminary Results" (http://www.an andtech.com/show/9662/iphone-6s-and-iphone-6s-plus-preliminary-results). *AnandTech*. Retrieved 2016-06-01.

62. Chester, Brandon (May 16, 2016). "The iPhone SE Review" (https://www.anandtech.com/show/10285/t he-iphone-se-review/2). *AnandTech*.

63. Matthew Wilcox (2011-03-03). "NVM Express driver" (https://archive.today/20120717195616/http://sb.l wn.net/Articles/431103/). LWN.net. Archived from the original (http://sb.lwn.net/Articles/431103/) on 2012-07-17. Retrieved 2013-11-05.

64. Keith Busch (2013-08-12). "Linux NVMe Driver" (http://www.flashmemorysummit.com/English/Collatera ls/Proceedings/2013/20130812_PreConfD_Busch.pdf) (PDF). *flashmemorysummit.com*. Retrieved 2013-11-05.

65. "IDF13 Hands-on Lab: Compiling the NVM Express Linux Open Source Driver and SSD Linux Benchmarks and Optimizations" (https://web.archive.org/web/20140111004350/https://intel.activeevent s.com/sf13/connect/fileDownload/session/FF44850B359CA1CD47D3E6A3437446FD/SF13_SSDL00 1_100.pdf) (PDF). *activeevents.com*. 2013. Archived from the original (https://intel.activeevents.com/sf1 3/connect/fileDownload/session/FF44850B359CA1CD47D3E6A3437446FD/SF13_SSDL001_100.pd f) (PDF) on 2014-01-11. Retrieved 2014-01-11.

66. "Merge git://git.infradead.org/users/willy/linux-nvme" (https://git.kernel.org/cgit/linux/kernel/git/torvalds/li nux.git/commit/?id=92b5abbb44e05cdbc4483219f30a435dd871a8ea). *kernel.org*. 2012-01-18. Retrieved 2013-11-05.

67. Kim, K.; Kim, T. (2020). "HMB in DRAM-less NVMe SSDs: Their usage and effects on performance" (htt ps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7051071). *PLOS ONE*. 15 (3): e0229645. Bibcode:2020PLoSO..1529645K (https://ui.adsabs.harvard.edu/abs/2020PLoSO..1529645K). doi:10.1371/journal.pone.0229645 (https://doi.org/10.1371%2Fjournal.pone.0229645). PMC 7051071 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7051071). PMID 32119705 (https://pubmed.ncbi.nlm.ni h.gov/32119705).

68. "Linux 4.13 has been released on Sun, 3 Sep 2017" (https://kernelnewbies.org/Linux_4.13#Storage).

69. "Pci.c « host « nvme « drivers - kernel/Git/Stable/Linux.git - Linux kernel stable tree"

(https://git.kernel.org/pub/scm/linux/kernel/git/stable/linux.git/tree/drivers/nvme/host/pci.c?h=v4.13.1).
70. "Faster 'NVM Express' SSD Interface Arrives on Retina MacBook and OS X 10.10.3" (http://www.macru mors.com/2015/04/11/nvme-mac-os-x/). *macrumors.com*. Retrieved 11 April 2015.
71. "nvme -- Non-Volatile Memory Host Controller Interface" (http://man.netbsd.org/nvme.4). *NetBSD manual pages*. 2021-05-16. Retrieved 2021-05-16.
72. David Gwynne (2014-04-16). "non volatile memory express controller (/sys/dev/ic/nvme.c)" (http://bxr.s u/OpenBSD/sys/dev/ic/nvme.c). *BSD Cross Reference*. Retrieved 2014-04-27.
73. David Gwynne (2016-04-14). "man 4 nvme" (http://man.openbsd.org/OpenBSD-current/man4/nvme.4). *OpenBSD man page*. Retrieved 2016-08-07.
74. "NVME" (https://www.arcanoae.com/wiki/nvme/). *Arca Noae wiki*. Arca Noae, LLC. 2021-04-03. Retrieved 2021-06-08.
75. "nvme(7D)" (https://docs.oracle.com/cd/E36784_01/html/E36884/esc-nxge-7d.html). Oracle. Retrieved 2014-12-02.
76. "Intel Solid-State for NVMe Drivers" (https://downloadcenter.intel.com/download/23929/Intel-Solid-Stat e-Drive-Data-Center-Family-for-NVMe-Drivers). *intel.com*. 2015-09-25. Retrieved 2016-03-17.
77. "VMware Compatibility Guide for NVMe devices" (http://www.vmware.com/resources/compatibility/vcl/result.php?search=NVMe&searchCategory=all). *vmware.com*. Retrieved 2016-03-17.
78. "VSAN Now Supporting NVMe Devices" (https://blogs.vmware.com/virtualblocks/2015/11/11/vsan-now-supporting-nvme-devices/). *vmware.com*. 2015-11-11. Retrieved 2016-03-17.
79. "Windows 8.1 to support hybrid disks and adds native NVMe driver" (http://www.myce.com/news/windo ws-8-1-to-support-hybrid-disks-and-native-nvme-driver-68663/). *Myce.com*. 2013-09-06. Retrieved 2014-01-11.
80. "Update to support NVM Express by using native drivers in Windows 7 or Windows Server 2008 R2" (h ttp://support.microsoft.com/kb/2990941/en-us). Microsoft. 2014-11-13. Retrieved 2014-11-17.
81. "Recommended AHCI/RAID and NVMe Drivers" (https://www.win-raid.com/t29f25-Recommended-AH CI-RAID-and-NVMe-Drivers.html). 10 May 2013.
82. "Windows NVM Express" (https://web.archive.org/web/20130612081416/https://www.openfabrics.org/r esources/developer-tools/nvme-windows-development.html). *Project web site*. Archived from the original (http://www.openfabrics.org/resources/developer-tools/nvme-windows-development.html) on June 12, 2013. Retrieved September 18, 2013.
83. "Nvmewin-Revision 157: /Releases" (https://web.archive.org/web/20170510114242/https://svn.openfabrics.org/svnrepo/nvmewin/releases/). Archived from the original (https://svn.openfabrics.org/svnrepo/n vmewin/releases/) on 2017-05-10. Retrieved 2016-08-13.
84. "ChangeLog/1.6" (http://wiki.qemu.org/ChangeLog/1.6). *qemu.org*. Retrieved 21 March 2015.
85. "Download EDK II from" (https://sourceforge.net/projects/edk2/files/EDK%20II%20Releases/other/Nvm ExpressDxe-alpha.zip/download). *SourceForge.net*. Retrieved 2014-01-11.

86. ***NVM Express control utility*** **(https://www.freebsd.org/cgi/man.cgi?query=nvmecontrol&sektion=8&manp ath=freebsd-release-ports)**, **The FreeBSD Project, 2018-03-12, retrieved 2019-07-12**

87. ***GitHub - linux-nvme/nvme-cli: NVMe management command line interface.*** **(https://github.com/linux-nv me/nvme-cli)**, **linux-nvme, 2019-03-26, retrieved 2019-03-27**

External links

- Official website
- NVMe info
- CompactFlash Association
- LFCS: Preparing Linux for nonvolatile memory devices, LWN.net, April 19, 2013, by Jonathan Corbet
- Multipathing PCI Express Storage, Linux Foundation, March 12, 2015, by Keith Busch
- NVMe, NVMe-oF and RDMA for network engineers, August 2020, by Jerome Tissieres

10.1.3 CXL 协议 Wiki

WIKIPEDIA

Compute Express Link

Compute Express Link (**CXL**) is an open standard for high-speed central processing unit (CPU)-to-device and CPU-to-memory connections, designed for high performance data center computers. CXL is built on the PCI Express (PCIe) physical and electrical interface and includes PCIe-based block input/output protocol (CXL.io) and new cache-coherent protocols for accessing system memory (CXL.cache) and device memory (CXL.mem).

Year created	2019
№ of devices	4096
Speed	Full duplex **1.x**, **2.0** (32 GT/s): 3.938 GB/s (×1) 63.015 GB/s (×16) **3.0** (64 GT/s): 7.563 GB/s (×1) 121.0 GB/s (×16)
Style	Serial
Website	www.computeexpresslink.org

History

The CXL technology was primarily developed by Intel. The CXL Consortium was formed in March 2019 by founding members Alibaba Group, Cisco Systems, Dell EMC, Meta, Google, Hewlett Packard Enterprise (HPE), Huawei, Intel Corporation and Microsoft,[5][6] and officially incorporated in September 2019.[7] As of January 2022, AMD, Nvidia, Samsung Electronics and Xilinx joined the founders on the board of directors, while ARM, Broadcom, Ericsson, IBM, Keysight, Kioxia, Marvell Technology, Mellanox, Microchip Technology, Micron, Oracle Corporation, Qualcomm, Rambus, Renesas, Seagate, SK Hynix, Synopsys, and Western Digital, among others, were contributing members.[8][9] Industry partners include the PCI-SIG,[10] Gen-Z,[11] SNIA,[12] and DMTF.[13]

On April 2, 2020, the Compute Express Link and Gen-Z Consortiums announced plans to implement interoperability between the two technologies,[14][15] with initial results presented in January 2021.[16] On November 10, 2021, Gen-Z specifications and assets were transferred to CXL, to focus on developing a single industry standard.[17] At the time of this announcement, 70% of Gen-Z members already joined the CXL Consortium.[18]

On August 1, 2022, OpenCAPI specifications and assets were transferred to the CXL Consortium,[19][20] which now includes companies behind memory coherent interconnect technologies such as OpenCAPI (IBM), Gen-Z (HPE), and CCIX (Xilinx) open standards, and proprietary InfiniBand / RoCE (Mellanox), Infinity Fabric (AMD), Omni-Path and QuickPath/Ultra Path (Intel), and NVLink/NVSwitch (Nvidia) protocols.[21]

Specifications

On March 11, 2019, the CXL Specification 1.0 based on PCIe 5.0 was released.[6] It allows host CPU to access shared memory on accelerator devices with a cache coherent protocol. The CXL Specification 1.1 was released in June, 2019.

On November 10, 2020, the CXL Specification 2.0 was released. The new version adds support for CXL switching, to allow connecting multiple CXL 1.x and 2.0 devices to a CXL 2.0 host processor, and/or pooling each device to multiple host processors, in distributed shared memory and disaggregated storage configurations; it also implements device integrity and data encryption.[22] There is no bandwidth increase from CXL 1.x, because CXL 2.0 still utilizes PCIe 5.0 PHY.

On August 2, 2022, the CXL Specification 3.0 was released, based on PCIe 6.0 physical interface and PAM-4 coding with double the bandwidth; new features include fabrics capabilities with multi-level switching and multiple device types per port, and enhanced coherency with peer-to-peer DMA and memory sharing.[23][24]

Implementations

On April 2, 2019, Intel announced their family of Agilex FPGAs featuring CXL.[25]

On May 11, 2021, Samsung announced a 128 GByte DDR5 based memory expansion module that allows for terabyte level memory expansion along with high performance for use in data centres and potentially next generation PCs.[26] An updated 512 GByte version based on a proprietary memory controller was released on May 10, 2022.[27]

In 2021, CXL 1.1 support was announced for Intel Sapphire Rapids processors[28] and AMD Zen 4 EPYC "Genoa" and "Bergamo" processors.[29]

CXL devices were shown at the ACM/IEEE Supercomputing Conference (SC21) by vendors including Intel,[30] Astera, Rambus, Synopsys, Samsung.

Protocols

The CXL standard defines three separate protocols:[34][22]

- **CXL.io - based on PCIe 5.0 with a few enhancements, it provides configuration, link initialization and management, device discovery and enumeration, interrupts, DMA, and register I/O access using non-coherent loads/stores.**
- **CXL.cache - allows peripheral devices to coherently access and cache host CPU memory with a low latency request/response interface.**
- **CXL.mem - allows host CPU to coherently access cached device memory with load/store commands for both volatile (RAM) and persistent non-volatile (flash memory) storage.**

CXL.cache and CXL.mem protocols operate with a common link/transaction layer, which is separate from the CXL.io protocol link and transaction layer. These protocols/layers are multiplexed together by an Arbitration and Multiplexing (ARB/MUX) block before being transported over standard PCIe 5.0 PHY using fixed-width 528 bit (66 byte) Flow Control Unit (FLIT) block consisting of four 16-byte data 'slots' and a two-byte cyclic redundancy check (CRC) value.[34] CXL FLITs encapsulate PCIe standard Transaction Layer Packet (TLP) and Data Link Layer Packet (DLLP) data with a variable frame size format.[35][36]

CXL 3.0 introduces 256-byte FLIT in PAM-4 transfer mode.

Device types

CXL is designed to support three primary device types:[22]

- Type 1 (CXL.io and CXL.cache) – specialised accelerators (such as smart NIC) with no local memory. Devices rely on coherent access to host CPU memory.
- Type 2 (CXL.io, CXL.cache and CXL.mem) – general-purpose accelerators (GPU, ASIC or FPGA) with high-performance GDDR or HBM local memory. Devices can coherently access host CPU's memory and/or provide coherent or non-coherent access to device local memory from the host CPU.
- Type 3 (CXL.io and CXL.mem) – memory expansion boards and persistent memory. Devices provide host CPU with low-latency access to local DRAM or byte-addressible non-volatile storage.

Type 2 devices implement two memory coherence modes, managed by device driver. In device bias mode, device directly accesses local memory and no caching is performed by the CPU; in host bias mode, the host CPU's cache controller handles all access to device memory. Coherence mode can be set individually for each 4 KB page, stored in a translation table in local memory of Type 2 devices. Unlike other CPU-to-CPU memory coherency protocols, this arrangement only requires the host CPU memory controller to implement the cache agent; such asymmetric approach reduces implementation complexity and reduces latency.[34]

CXL 2.0 added support for switching in tree-based device fabrics, allowing PCIe, CXL 1.1 and CXL 2.0 devices to form virtual hierarchies of single- and multi-logic devices that can be managed by multiple hosts.[37]

CXL 3.0 replaced bias modes with enhanced coherency semantics, allowing Type 2 and Type 3 devices to back invalidate the data in the host cache when the device has made a change to the local memory. Enhanced coherency also helps implement peer-to-peer transfers within a virtual hierarchy of devices in the same coherency domain. It also supports memory sharing of the same memory segment between multiple devices, as opposed to memory pooling where each device was assigned a separate segment.[38]

CXL 3.0 allows multiple Type 1 and Type 2 devices per each CXL root port; it also adds multi-level switching, helping implement device fabrics with non-tree topologies like mesh, ring, or spline/leaf. Each node can be a host or a device of any type. Type 3 devices can implement Global Fabric Attached Memory (GFAM) mode, which connects a memory device to a switch node without requiring direct host connection. Devices and hosts use Port Based Routing (PBR) addressing mechanism that supports up to 4,096 nodes.[38]

See also

- Coherent Accelerator Processor Interface (CAPI)
- UCIe
- Data processing unit (DPU)

References

1. **"ABOUT CXL" (https://www.computeexpresslink.org/about-cxl)**. *Compute Express Link*. **Retrieved 2019-08-09.**
2. **"Synopsys Delivers Industry's First Compute Express Link (CXL) IP Solution for Breakthrough Performance in Data-Intensive SoCs" (https://finance.yahoo.com/news/synopsys-delivers-industrys-first-compute-000000436.html)**. *finance.yahoo.com*. Yahoo! Finance. Retrieved 2019-11-09.
3. **"A Milestone in Moving Data" (https://newsroom.intel.com/editorials/milestone-moving-data/)**. *Intel Newsroom*. **Intel**. **Retrieved 2019-11-09.**
4. **"Compute Express Link Consortium (CXL) Officially Incorporates; Announces Expanded Board of Directors" (https://www.businesswire.com/news/home/20190917005948/en/Compute-Express-Link-Consortium-CXL-Officially-Incorporates)**. *www.businesswire.com*. Business Wire. 2019-09-17. Retrieved 2019-11-09.
5. Comment, Will Calvert. **"Intel, Google and others join forces for CXL interconnect" (https://www.dat**

acenterdynamics.com/en/news/intel-google-and-others-join-forces-cxl-interconnect/). *www.datacenterdynamics.com*.

6. Cutress, Ian. "CXL Specification 1.0 Released: New Industry High-Speed Interconnect From Intel" (https://www.anandtech.com/show/14068/cxl-specification-1-released-new-industry-high-speed-int erconnect-from-intel). *Anandtech*. Retrieved 2019-08-09.

7. "Compute Express Link Consortium (CXL) Officially Incorporates; Announces Expanded Board of Directors" (https://www.businesswire.com/news/home/20190917005948/en/Compute-Express-Link-Consortium-CXL-Officially-Incorporates-Announces-Expanded-Board-of-Directors). *www.businesswire.com*. September 17, 2019.

8. "Compute Express Link: Our Members" (https://www.computeexpresslink.org/members). *CXL Consortium*. 2020. Retrieved 2020-09-25.

9. Papermaster, Mark (July 18, 2019). "AMD Joins Consortia to Advance CXL, a New High-Speed Interconnect for Breakthrough Performance" (https://community.amd.com/community/amd-busines s/blog/2019/07/18/amd-joins-consortia-to-advance-cxl-a-new-high-speed-interconnect-for-breakthr ough-performance). *Community.AMD*. Retrieved 2020-09-25.

10. "CXL Consortium and PCI-SIG Announce Marketing MOU Agreement" (https://www.computeexpresslink.org/post/cxl-consortium-and-pci-sig-announce-marketing-mou-agreement). 23 September 2021.

11. "Industry Liaisons" (https://www.computeexpresslink.org/industry-liaisons).

12. "SNIA and CXL Consortium Form Strategic Alliance" (https://www.computeexpresslink.org/post/snia-and-cxl-consortium-form-strategic-alliance). 3 November 2020.

13. "DMTF and CXL Consortium Establish Work Register" (https://www.computeexpresslink.org/post/dmtf-and-cxl-consortium-establish-work-register). 14 April 2020.

14. "CXL Consortium and Gen-Z Consortium Announce MOU Agreement" (https://b373eaf2-67af-4a29-b28c-3aae9e644f30.filesusr.com/ugd/0c1418_efb1cff3f41d486ea85d50ec638ea715.pdf) (PDF). Beaverton, Oregon. April 2, 2020. Retrieved September 25, 2020.

15. "CXL Consortium and Gen-Z Consortium Announce MOU Agreement" (https://genzconsortium.org/cxl-consortium-and-gen-z-consortium-announce-mou-agreement/). April 2, 2020. Retrieved April 11, 2020.

16. "CXL™ Consortium and Gen-Z Consortium™ MoU Update: A Path to Protocol" (https://www.computeexpresslink.org/post/cxl-consortium-and-gen-z-consortium-mou-update-a-path-to-protocol). 24 June 2021.

17. Consortium, C. X. L. (November 10, 2021). "Exploring the Future" (https://www.computeexpresslink.org/post/exploring-the-future-cxl-consortium-gen-z-consortium). *Compute Express Link*.

18. "CXL Will Absorb Gen-Z" (https://www.eetimes.com/cxl-will-absorb-gen-z/). 9 December 2021.

19. OpenCAPI to Fold into CXL - CXL Set to Become Dominant CPU Interconnect Standard (https://www.anandtech.com/show/17519/opencapi-to-fold-into-cxl)

20. CXL Consortium and OpenCAPI Consortium Sign Letter of Intent to Transfer OpenCAPI Specifications to CXL (https://www.computeexpresslink.org/_files/ugd/0c1418_d3474155dc6e4929aa2a5658a894d1a6.pdf)

21. Morgan, Timothy Prickett (November 23, 2021). "Finally, A Coherent Interconnect Strategy: CXL Absorbs Gen-Z" (https://www.nextplatform.com/2021/11/23/finally-a-coherent-interconnect-strategy

-cxl-absorbs-gen-z/). *The Next Platform*.

22. **"Compute Express Link (CXL): All you need to know" (https://www.rambus.com/blogs/compute-express-link/)**. ***Rambus***.

23. **"Compute Express Link (CXL) 3.0 Announced: Doubled Speeds and Flexible Fabrics" (https://www.anandtech.com/show/17520/compute-express-link-cxl-30-announced-doubled-speeds-and-flexible-fabrics)**.

24. **"Compute Express Link (CXL) 3.0 Debuts, Wins CPU Interconnect Wars" (https://www.tomshardware.com/news/cxl-30-debuts-one-cpu-interconnect-to-rule-them-all)**. **2 August 2022**.

25. **"How do the new Intel Agilex FPGA family and the CXL coherent interconnect fabric intersect?" (https://blogs.intel.com/psg/how-do-the-new-intel-agilex-fpga-family-and-the-cxl-coherent-interconnect-fabric-intersect/)**. *PSG@Intel*. 2019-05-03. Retrieved 2019-08-09.

26. **"Samsung Unveils Industry-First Memory Module Incorporating New CXL Interconnect Standard" (https://news.samsung.com/global/samsung-unveils-industry-first-memory-module-incorporating-new-cxl-interconnect-standard)**. ***Samsung***. **2021-05-11. Retrieved 2021-05-11.**

27. **"Samsung Electronics Introduces Industry's First 512GB CXL Memory Module" (https://news.samsung.com/global/samsung-electronics-introduces-industrys-first-512gb-cxl-memory-module)**.

28. **"Intel Architecture Day 2021" (https://www.intel.com/content/www/us/en/newsroom/resources/press-kit-architecture-day-2021.html)**. *Intel*.

29. **Paul Alcorn (November 8, 2021). "AMD Unveils Zen 4 CPU Roadmap: 96-Core 5nm Genoa in 2022, 128-Core Bergamo in 2023" (https://www.tomshardware.com/news/amd-unveils-zen-4-cpu-roadmap-96-core-5nm-genoa-128-core-begamo)**. ***Tom's Hardware***.

30. **Patrick Kennedy (December 7, 2021). "Intel Sapphire Rapids CXL with Emmitsburg PCH Shown at SC21" (https://www.servethehome.com/intel-sapphire-rapids-cxl-emmitsburg-pch-sc21-astera-labs-synopsys/)**. *Serve the Home*. Retrieved November 18, 2022.

31. **"CXL Put Through Its Paces" (https://www.eetimes.com/cxl-put-through-its-paces/)**. **December 10, 2021.**

32. **"CXL Consortium Showcases First Public Demonstrations of Compute Express Link Technology at SC21" (https://www.hpcwire.com/off-the-wire/cxl-consortium-showcases-first-public-demonstrations-of-compute-express-link-technology-at-sc21/)**. ***HPCwire***.

33. **Consortium, C. X. L. (December 16, 2021). "CXL Consortium Makes a Splash at Supercomputing 2021 (SC21)" (https://www.computeexpresslink.org/post/cxl-consortium-makes-a-splash-at-supercomputing-2021-sc21)**. ***Compute Express Link***.

34. **"Compute Express Link Standard | DesignWare IP | Synopsys" (https://www.synopsys.com/designware-ip/technical-bulletin/compute-express-link-standard-2019q3.html)**. ***www.synopsys.com***.

35. **Consortium, C. X. L. (September 23, 2019). "Introduction to Compute Express Link (CXL): The CPU-To-Device Interconnect Breakthrough" (https://www.computeexpresslink.org/post/introduction-to-compute-express-link-cxl-the-cpu-to-device-interconnect-breakthrough)**. *Compute Express Link*.

36. **https://www.flashmemorysummit.com/Proceedings2019/08-07-Wednesday/20190807_CTRL-202A-1_Lender.pdf**

37. **Danny Volkind and Elad Shlisberg (June 15, 2022). "CXL 1.1 vs CXL 2.0 – What's the difference?" (https://www.computeexpresslink.org/_files/ugd/0c1418_74c3afe48bf340cdbe59af75a88f2370.pdf) (PDF). UnifabriX. Retrieved November 18, 2022.**

External links

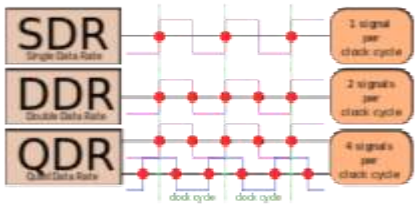
Official website (http://www.computeexpresslink.org)

10.1.4 DDR 协议 Wiki

WIKIPEDIA

Double data rate

In computing, a computer bus operating with **double data rate** (**DDR**) transfers data on both the rising and falling edges of the **c**lock signal.[1] This is also known as **double pumped**, **dual- pumped**, and **double transition**. The term **toggle mode** is used in the context of NAND flash memory.



A comparison between single data rate, double data rate, and quad data rate

Overview

The simplest way to design a clocked electronic circuit is to make it perform one transfer per full cycle (rise and fall) of a clock signal. This, however, requires that the clock signal changes twice per transfer, while the data lines change at most once per transfer. When operating at a high bandwidth, signal integrity limitations constrain the clock frequency. By using both edges of the clock, the data signals operate with the same limiting frequency, thereby doubling the data transmission rate.

This technique has been used for microprocessor front-side busses, Ultra-3 SCSI, expansion buses (AGP, PCI-X[2]), graphics memory (GDDR), main memory (both RDRAM and DDR1 through DDR5), and the HyperTransport bus on AMD's A thlon 64 processors. It is more recently being used for other systems with high data transfer speed requirements – as an example, for the output of analog-to-digital converters (ADCs).[3]

DDR should not be confused with dual channel, in which each memory channel accesses two RAM modules simultaneously. The two technologies are independent of each other and many motherboards use both, by using DDR memory in a dual channel configuration.

An alternative to double or quad pumping is to make the link self-clocking. This tactic was chosen by hfiniBand and PCI Express.

Relation of bandwidth and frequency

Describing the bandwidth of a double-pumped bus can be confusing. Each clock edge is referred to as a *beat*, with two beats (one upbeat and one downbeat) per cycle. Technically, the hertz is a unit of *cycles* per second, but many people refer to the number of *transfers* per second. Careful usage generally talks about "500 MHz,

double data rate" or "1000 MT/s", but many refer casually to a "1000 MHz bus," even though no signal cycles faster than 500 MHz.

DR SDRAM popularized the technique of referring to the bus bandwidth in megabytes per second, the product of the transfer rate and the bus width in bytes. DDR SDRAM operating with a 100 MHz clock is called DDR-200 (after its 200 MT/s data transfer rate), and a 64-bit (8-byte) wide D IMM operated at that data rate is called PC-1600, after its 1600 MB/s peak (theoretical) bandwidth. Likewise, 12.8 GB/s transfer rate DDR3-1600 is called PC3-12800.

Some examples of popular designations of DDR modules:

Names	Memory clock	I/O bus clock	Transfer rate	Theoretical bandwidth
DDR-200, PC-1600	100 MHz	100 MHz	200 MT/s	1.6 GB/s
DDR-400, PC-3200	200 MHz	200 MHz	400 MT/s	3.2 GB/s
DDR2-800, PC2-6400	200 MHz	400 MHz	800 MT/s	6.4 GB/s
DDR3-1600, PC3-12800	200 MHz	800 MHz	1600 MT/s	12.8 GB/s
DDR4-2400, PC4-19200	300 MHz	1200 MHz	2400 MT/s	19.2 GB/s
DDR4-3200, PC4-25600	400 MHz	1600 MHz	3200 MT/s	25.6 GB/s
DDR5-4800, PC5-38400	300 MHz	2400 MHz	4800 MT/s	38.4 GB/s
DDR5-6400, PC5-51200	400 MHz	3200 MHz	6400 MT/s	51.2 GB/s

DDR SDRAM uses double-data-rate signalling only on the data lines. Address and control signals are still sent to the DRAM once per clock *cycle* (to be precise, on the rising edge of the clock), and timing parameters such as CAS latency are specified in clock cycles. Some less common DRAM interfaces, notably LPDDR2, GDDR5 and XDR DRAM, send commands and addresses using double data rate.

DR5 uses two 7-bit double data rate command/address buses to each DIMM, where a registered clock

driver chip converts to a 14-bit SDR bus to each memory chip.

See also

- **DDR SDRAM**, **DDR2 SDRAM**, **DDR3 SDRAM**, **DDR4 SDRAM** and **DDR5 SDRAM**
- **GDDR SDRAM**, **GDDR3 SDRAM**, **GDDR4 SDRAM**, **GDDR5 SDRAM** and **GDDR6 SDRAM**
- **List of interface bit rates**
- **Pumping (computer systems)**
- **Quad data rate**

References

1. Hennessy, John L.; Patterson, David A. (2007). ***Computer architecture: a quantitative approach*** (https://books.google.com/books?id=pqYl3SWkA64C&pg=PA314). Amsterdam: Morgan Kaufmann. p. 314. ISBN 0-12-370490-1.

2. Schmid, Patrick. **"PCI Express Battles PCI-X" (https://www.tomshardware.com/reviews/pci-e xpress-**

battles-pci,1176-2.html). *Tom's Hardware Guide*.
3. "AD9467 ADC" (http://www.analog.com/static/imported-files/data_sheets/AD9467.pdf) (PDF) (data sheet). Analog Devices.

10.1.5 UFS 协议 Wiki

WIKIPEDIA

Universal Flash Storage

Universal Flash Storage (**UFS**) is a flash storage specification for digital cameras, mobile phones and onsumer electronic devices.[1][2] It was designed to bring higher data transfer speed and increased reliability to flash memory storage, while reducing market confusion and removing the need for different adapters for different types of cards.[3] The standard encompasses both packages permanently attached (embedded) within a device (**eUFS**), and removable UFS memory cards.

Overview

UFS uses NAND flash. It may use multiple stacked 3D TLC NAND flash dice (integrated circuits) with an integrated controller.[4]

The proposed flash memory specification is supported by consumer electronics companies such as Nokia, Sony Ericsson, Texas Instruments, STMicroelectronics, Samsung, Micron, and SK Hynix.[5] UFS is positioned as a replacement for eMMCs and SD cards. The electrical interface for UFS uses the M-PHY,[6] developed by the MIPI Alliance, a high-speed serial interface targeting 2.9 Gbit/s per lane with up- scalability to 5.8 Gbit/s per lane.[7][8] UFS implements a full-duplex serial LVDS interface that scales better to higher bandwidths than the 8-lane parallel and half-duplex interface of eMMCs. Unlike eMMC, Universal Flash Storage is based on the SCSI architectural model and supports SCSI Tagged Command Queuing[9] The standard is developed by, and available from, the JEDEC Solid State Technology Association.

The Linux kernel supports UFS.[10]

History

In 2010, the Universal Flash Storage Association (UFSA) was founded as an open trade association to promote the UFS standard.

In September 2013, JEDEC published JESD220B UFS 2.0 (update to UFS v1.1 standard published in June 2012). JESD220B Universal Flash Storage v2.0 offers increased link bandwidth for performance improvement, a security features extension and additional power saving features over the UFS v1.1.

On 30 January 2018 JEDEC published version 3.0 of the UFS standard, with a higher 11.6 Gbit/s data rate per lane (1450

MB/s) with the use of MIPI M-PHY v4.1 and UniProSM v1.8. At the MWC 2018, Samsung unveiled embedded UFS (eUFS) v3.0 and uMCP (UFS-based multi-chip package) solutions.[11][12][13]

On 30 January 2020 JEDEC published version 3.1 of the UFS standard.[14] UFS 3.1 introduces Write Booster, Deep Sleep, Performance Throttling Notification and Host Performance Booster for faster, more power efficient and cheaper UFS solutions. The Host Performance Booster feature is optional.[15]

In 2022 Samsung announced version 4.0 doubling from 11.6 Gbit/s to 23.2 Gbit/s with the use of MIPI M- PHY v5.0 and UniPro v2.0.

Notable devices

In February 2013, semiconductor company Toshiba Memory (now Kioxia) started shipping samples of a 64GB NAND flash chip, the first chip to support the then new UFS standard.[16]

In April 2015, Samsung's Galaxy S6 family was the first phone to ship with eUFS storage using the UFS 2.0 standard.[17]

On 7 July 2016, Samsung announced its first UFS cards, in 32, 64, 128, and 256 GB storage capacities.[18] The cards were based on the UFS 1.0 Card Extension Standard. The 256GB version was reported to offer sequential read performance up to 530 MB/s and sequential write performance up to 170 MB/s and random performance of 40,000 read IOPS and 35,000 write IOPS. However, they were apparently not actually released to the public.

On 17 November 2016, Qualcomm announced the Snapdragon 835 SoC with support for UFS 2.1.[19]

On 14 May 2019, OnePlus introduced the OnePlus 7 and OnePlus 7 Pro, the first phones to feature built-in eUFS 3.0 (The Galaxy Fold, originally planned to be the first smartphone to feature UFS 3.0 was ultimately delayed after the OnePlus 7's launch).[20]

The first UFS cards began to be publicly sold in early 2020. According to a Universal Flash Storage Association press release, Samsung planned to transition its products to UFS cards during 2020.[21] Several consumer devices with UFS card slots have been released in 2020.

Version comparison

UFS

UFS	1.0	1.1	2.0	2.1	2.2	3.0	3.1	4.0
Introduced	2011-02-24[22]	2012-06-25[23]	2013-09-18[24]	2016-04-04[25]	2020-08[26]	2018-01-30[27]	2020-01-30[14]	2022-08-17[28]
Bandwidth per lane	300 MB/s		600 MB/s			1450 MB/s		2900 MB/s
Max. number of lanes	1		2					
Max. total bandwidth	300 MB/s		1200 MB/s			2900 MB/s		5800 MB/s
M-PHY version	?	?	3.0		?	4.1		5.0
UniPro version	?	?	1.6		?	1.8		2.0

UFS Card

UFS Card	1.0	1.1	3.0
Introduced	2016-03-30[29]	2018-01-30[27]	2020-12-08[30]
Bandwidth per lane	600 MB/s		1200 MB/s
Max. number of lanes	1		
Max. total bandwidth	600 MB/s		1200 MB/s
M-PHY version	3.0		4.1
UniPro version	1.6		1.8

Implementation

- **UFS 2.0 has been implemented in Snapdragon 820 and 821. Kirin 950 and 955. Exynos 7420.**
- **UFS 2.1 has been implemented in Snapdragon 712 (710&720G), 730G, 732G, 835, 845 and 855. Kirin 960, 970 and 980. Exynos 9609,[31] 9610,[32] 9611,[33] 9810 and 980.[34]**
- **UFS 3.0 has been implemented in Snapdragon 855, 855+, 860, 865, Exynos 9820–9825,[35] and Kirin 990.[36]**
- **UFS 3.1 has been implemented in Snapdragon 855+/860, Snapdragon 865, Snapdragon 870, Snapdragon 888, Exynos 2100, and Exynos 2200.[37][38][39]**
- **UFS 4.0 has been implemented in MediaTek Dimensity 9200 and Snapdragon 8 Gen 2.[40]**

Complementary UFS standards

On 30 March 2016, JEDEC published version 1.0 of the UFS Card Extension Standard (JESD220-2), which offered many of the features and much of the same functionality as the existing UFS 2.0 embedded device standard, but with additions and modifications for removable cards.[41]

Also in March 2016, JEDEC published version 1.1 of the UFS Unified Memory Extension (JESD220- 1A),[42] version 2.1 of the UFS Host Controller Interface (UFSHCI) standard (JESD223C),[43] and version 1.1A of the UFSHCI Unified Memory Extension standard (JESD223-1A).[44]

On January 30, 2018, the UFS Card Extension standard was updated to version 1.1 (JESD220-2A),[45] and the UFSHCI standard was updated to version 3.0 (JESD223D), to align with UFS version 3.0.[46]

Rewrite cycle life

A UFS drive's rewrite life cycle affects its lifespan. There is a limit to how many write/erase cycles a flash block can accept before it produces errors or fails altogether. Each write/erase cycle causes a flash memory cell's oxide layer to deteriorate. The reliability of a drive is based on three factors: the age of the drive, total terabytes written over time and drive writes per day.[47] This is typical of flash memory in general.

See also

- Memory card
- Solid-state drive

References

1. "Nokia, Others Back Mobile Memory Standard" (https://web.archive.org/web/20080209210001/http://www.pcworld.com/article/id,137200-c,unresolvedtechstandards/article.html). *PC World*. Archived from the original (http://www.pcworld.com/article/id,137200-c,unresolvedtec hstandards/article.html) on 9 February 2008.
2. "JEDEC Announces Publication of Universal Flash Storage (UFS) Standard | JEDEC" (https://www.jedec.org/news/pressreleases/jedec-announces-publication-universal-flash-storage -ufs-standard). *www.jedec.org*.
3. Malykhina, Elena (14 September 2007). "Mobile Tech Companies Work On Flash Memory Standard" (https://web.archive.org/web/20120912190317/http://www.informationweek.com/ mobile-tech-companies-work-on-flash-memo/201806565). *Information Week*. Archived from the original (http://www.informationweek.com/mobile-tech-companies-work-on-flash-memo/2 01806565) on 12 September 2012. Retrieved 19 September 2012.
4. "Toshiba Begins to Sample UFS 3.0 Drives: 96L 3D TLC NAND, Up to 2.9 GB/s" (https://www.anandtech.com/show/13891/toshiba-samples-ufs-3-storage). *Anandtech*. 23 January 2019. Retrieved 18 August 2020.
5. Modine, Austin (14 September 2007). "Flash memory makers propose common card" (http://www.channelregister.co.uk/2007/09/14/flash_memory_makers_propose_ufs/). *The Channel*. Retrieved 19 September 2012.
6. "JEDEC Solid State Technology Association | MIPI Alliance" (https://web.archive.org/web/20110928215748/http://www.mipi.org/about-mipi/industry-associations/jedec-solid-state-techn ology-association). Archived from the original (http://www.mipi.org/about-mipi/industry-associ ations/jedec-solid-state-technology-association/) on 28 September 2011.Retrieved 15 August 2011.
7. "MIPI" (https://www.mipi.org/). *MIPI*.
8. "Universal Flash Storage (UFS) Eco-System | TOSHIBA Semiconductor & Storage Products Company | Europe(EMEA)" (https://web.archive.org/web/20151222124341/http://toshiba.se micon-storage.com/eu/application/ufs.html). Archived from the original (http://toshiba.semico n-storage.com/eu/application/ufs.html) on 22 December 2015. Retrieved 26 October 2015.
9. "Universal Flash Storage: Mobilize Your Data" (http://www.design-reuse.com/articles/30845/ universal-flash-storage-mobilize-your-data.html). *Design Reuse*. Retrieved 18 August 2020.
10. "Universal Flash Storage" (https://www.kernel.org/doc/Documentation/scsi/ufs.txt). *The Linux Kernel Archives*. Retrieved 13 November 2022.
11. "Evolving Mobile Solutions: Samsung at MWC 2018 | Samsung Semiconductor Global Website" (http://www.samsung.com/semiconductor/insights/news-events/evolving-mobile-so lutions-samsung-at-mwc-2018/). *www.samsung.com*.

12. "eUFS | Samsung Semiconductor Global Website" (http://www.samsung.com/semiconductor/estorage/eufs/). *www.samsung.com*.

13. "Samsung Starts Producing First 512-Gigabyte Universal Flash Storage for Next-Generation Mobile Devices | Samsung Semiconductor Global Website" (http://www.samsung.com/semic onductor/insights/news-events/samsung-starts-producing-first-512-gigabyte-universal-flash- storage-for-next-generation-mobile-devices/). *www.samsung.com*.

14. "JEDEC Publishes Update to Universal Flash Storage (UFS) Standard | JEDEC" (https://www.jedec.org/news/pressreleases/jedec-publishes-update-universal-flash-storage-ufs-standar d). *www.jedec.org*. Retrieved 31 January2020.

15. Shilov, Anton. "Faster, Cheaper, Power Efficient UFS Storage: UFS 3.1 Spec Published" (https://www.anandtech.com/show/15456/faster-cheaper-power-efficient-ufs-storage-ufs-31-spe c-published). *www.anandtech.com*. Retrieved 1 February2020.

16. "Toshiba ships first NAND flash chips with faster transfer standard" (https://www.pcworld.idg.com.au/article/453256/toshiba_ships_first_nand_flash_chips_faster_transfer_standard/). *PC World*. 8 February 2013. Retrieved 18 August 2020.

17. "The Samsung Galaxy S6 and S6 edge Review" (https://www.anandtech.com/show/9146/th e-samsung-galaxy-s6-and-s6-edge-review/7). *Anandtech*. 17 April 2015. Retrieved 18 August 2020.

18. Shilov, Anton. "Samsung Rolls Out Its First UFS Cards: SSD Performance in Card Form- Factor" (http://www.anandtech.com/show/10475/samsung-rolls-out-its-first-ufs-cards-ssd-perf ormance-in-card-formfactor). Retrieved 7 July 2016.

19. "Qualcomm Snapdragon 865 to sport LPDDR5X RAM, UFS 3.0, will come in 2 variants: Report" (https://www.firstpost.com/tech/news-analysis/qualcomm-snapdragon-865-to-sport-l pddr5x-ram-ufs-3-0-will-come-in-2-variants-report-6835011.html). *First Post*. 18 June 2019. Retrieved 18 August2020.

20. "OnePlus 7 Pro confirmed to feature UFS 3.0 flash storage" (https://www.androidcentral.co m/oneplus-7-pro-confirmed-feature-ufs-30-flash-storage). *Android Central*. 6 May 2019. Retrieved 18 August2020.

21. Universal Flash Storage Association (3 January 2020). "UFSA Widens UFS Ecosystem, Adding Vendors of Removable Mobile Cards and Related Technology" (https://www.busines swire.com/news/home/20200103005373/en/UFSA-Widens-UFS-Ecosystem-Adding-Vendors-of-Removable-Mobile-Cards-and-Related-Technology). *Business Wire*. Austin, Texas. Retrieved 22 November 2020. " "UFS Cards will play a critical role[...]" said Hangu Sohn, vice president of NAND Memory Planning at Samsung Electronics. "Moreover, with a royalty-free form factor and open standard design, we expect to see a rapid transition to these cards in 2020." "

22. "JEDEC Announces Publication of Universal Flash Storage (UFS) Standard | JEDEC" (https://www.jedec.org/news/pressreleases/jedec-announces-publication-universal-flash-storage -ufs-standard). *www.jedec.org*. Retrieved 8 May 2017.

23. "JEDEC Updates Universal Flash Storage (UFS) Standard | JEDEC" (https://www.jedec.org/news/pressreleases/jedec-updates-universal-flash-storage-ufs-standard). *www.jedec.org*. Retrieved 8 May 2017.

24. "JEDEC Publishes Universal Flash Storage (UFS) Standard v2.0 | JEDEC" (https://www.jedec.org/news/pressreleases/jedec-publishes-universal-flash-storage-ufs-standard-v20). *www.jedec.org*. Retrieved 8 May 2017.

25. "JEDEC Updates Universal Flash Storage (UFS) and Related Standards | JEDEC" (https://

www.jedec.org/news/pressreleases/jedec-updates-universal-flash-storage-ufs-and-related-s tandards). *www.jedec.org*. Retrieved 8 May 2017.

26. "UNIVERSAL FLASH STORAGE, UFS 2.2" (https://www.jedec.org/standards-documents/do cs/jesd220c-22). *www.jedec.org*. Retrieved 31 July 2021.

27. "JEDEC Publishes Universal Flash Storage (UFS & UFSHCI) Version 3.0 and UFS Card Extension Version 1.1 | JEDEC" (https://www.jedec.org/news/pressreleases/jedec-publishes -universal-flash-storage-ufs-ufshci-version-30-and-ufs-card). *www.jedec.org*. Retrieved 31 January 2018.

28. "JEDEC Updates Universal Flash Storage (UFS) and Supporting Memory Interface Standard" (https://www.jedec.org/news/pressreleases/jedec-updates-universal-flash-storage -ufs-and-supporting-memory-interface). *www.jedec.org*. Retrieved 18 September 2022.

29. "JEDEC Publishes Universal Flash Storage (UFS) Removable Card Standard | JEDEC" (htt ps://www.jedec.org/news/pressreleases/jedec-publishes-universal-flash-storage-ufs-remova ble-card-standard). *www.jedec.org*. Retrieved 30 October 2017.

30. "JEDEC Advances Universal Flash Storage (UFS) Removable Card Standard 3.0" (https://w ww.jedec.org/news/pressreleases/jedec-advances-universal-flash-storage-ufs-removable-ca rd-standard-30). *www.jedec.org*. Retrieved 31 July 2021.

31. "Exynos 9609 Mobile Processor: Specs, Features | Samsung Exynos" (https://www.samsun g.com/semiconductor/minisite/exynos/products/mobileprocessor/exynos-9609/). *Samsung Semiconductor*. Retrieved 26 January 2020.

32. "Exynos 9610 Processor: Specs, Features | Samsung Exynos" (https://www.samsung.com/s emiconductor/minisite/exynos/products/mobileprocessor/exynos-7-series-9610/). *Samsung Semiconductor*. Retrieved 26 January 2020.

33. "Exynos 9611 Mobile Processor: Specs, Features | Samsung Exynos" (https://www.samsun g.com/semiconductor/minisite/exynos/products/mobileprocessor/exynos-9611/). *Samsung Semiconductor*. Retrieved 26 January 2020.

34. "Exynos 980 5G Mobile Processor: Specs, Features | Samsung Exynos" (https://www.samsu ng.com/semiconductor/minisite/exynos/products/mobileprocessor/exynos-980/). *Samsung Semiconductor*. Retrieved 26 January 2020.

35. "Exynos 9 Series 9820 Processor: Specs, Features | Samsung Exynos"(https://www.samsu ng.com/semiconductor/minisite/exynos/products/mobileprocessor/exynos-9-series-9820/). *Samsung Semiconductor*. Retrieved 14 November 2018.

36. Cutress, Ian (6 September 2019). "Huawei Announces Kirin 990 and Kirin 990 5G: Dual SoC Approach, Integrated 5G Modem" (https://www.anandtech.com/show/14851/huawei-an nounces-kirin-990-and-kirin-990-5g-dual-soc-approach-integrated-5g-modem). *AnandTech*. Archived (https://web.archive.org/web/20190906140013/https://www.anandtech.com/show/1 4851/huawei-announces-kirin-990-and-kirin-990-5g-dual-soc-approach-integrated-5g-mode m) from the original on 6 September 2019.

37. "Qualcomm Snapdragon 888: specs and benchmarks" (https://nanoreview.net/en/soc/qualco mm-snapdragon-875). *NanoReview.net*. Retrieved 22 February 2021.

38. "Exynos 2100 5G Mobile Processor: Specs, Features" (https://www.samsung.com/semicond uctor/minisite/exynos/products/mobileprocessor/exynos-2100/). *Samsung.com*. Retrieved 27 June 2021.

39. "Exynos 2200 | Processor | Samsung Semiconductor" (https://www.samsung.com/semiconductor/minisite/exynos/products/mobileprocessor/exynos-2200/). *www.samsung.com*.
40. "Snapdragon-8-Gen-2-Product-Brief.pdf" (https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/Snapdragon-8-Gen-2-Product-Brief.pdf) (PDF). Qualcomm. Retrieved 19 November 2022.
41. "JEDEC Publishes Universal Flash Storage (UFS) Removable Card Standard | JEDEC" (https://www.jedec.org/news/pressreleases/jedec-publishes-universal-flash-storage-ufs-removable-card-standard). *www.jedec.org*. Retrieved 7 July 2016.
42. "Standards & Documents Search | JEDEC" (https://www.jedec.org/document_search?search_api_views_fulltext=jesd220-1a). *www.jedec.org*.
43. "Standards & Documents Search | JEDEC" (https://www.jedec.org/document_search?search_api_views_fulltext=jesd223c). *www.jedec.org*.
44. "Standards & Documents Search | JEDEC" (https://www.jedec.org/document_search?search_api_views_fulltext=jesd223-1a). *www.jedec.org*.
45. "UNIVERSAL FLASH STORAGE (UFS) CARD EXTENSION, Version 3.0 | JEDEC" (https://www.jedec.org/standards-documents/docs/jesd220-2). *www.jedec.org*.
46. "UFS (Universal Flash Storage) | JEDEC" (https://www.jedec.org/standards-documents/focus/flash/universal-flash-storage-ufs). *www.jedec.org*.
47. "SSD Lifespan: How Long Will Your SSD Work?" (https://www.enterprisestorageforum.com/storage-hardware/ssd-lifespan.html). *Enterprise Storage Forum*. 1 March 2019. Retrieved 18 August 2020.

External links

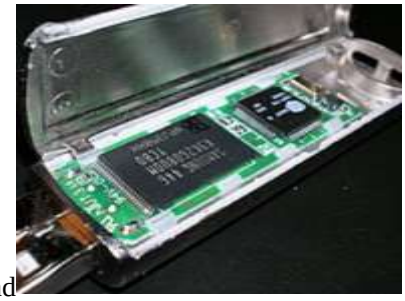
- JEDEC
- Universal Flash Storage Association
- Current standards of UFS and UFS Card
- Presentation by Scott Jacobson and Harish Verma at Flash Memory Summit 2013
- What is UFS 2.1 smartphone?

10.1.6 NAND 协议 Wiki

WIKIPEDIA

Flash memory

Flash memory is an electronic non-volatile computer memory storage medium that can be electrically erased and reprogrammed. The two main types of flash memory, **NOR flash** and **NAND flash**, are named for the NOR and NAND logic gates. Both use the same cell design, consisting of floating gate MOSFETs. They differ at the circuit level depending on whether the state of the bit line or word lines is pulled high or low: in NAND flash, the relationship between the bit line and the word lines resembles a NAND gate; in NOR flash, it resembles a NOR gate.



A disassembled USB flash drive. The chip on the left is flash memory. The controller is on the right.

Flash memory, a type of floating-gate memory, was invented at Toshiba in 1980 and is based on EEPROM technology. Toshiba began marketing flash memory in 1987.[1] EPROMs had to be erased completely before they could be rewritten. NAND flash memory, however, may be erased, written, and read in blocks (or pages), which generally are much smaller than the entire device. NOR flash memory allows a single machine word to be written – to an erased location – or read independently. A flash memory device typically consists of one or more flash memory chips (each holding many flash memory cells), along with a separate flash memory controller chip.

The NAND type is found mainly in memory cards, USB flash drives, solid-state drives (those produced since 2009), feature phones, smartphones, and similar products, for general storage and transfer of data. NAND or NOR flash memory is also often used to store configuration data in numerous digital products, a task previously made possible by EEPROM or battery-powered static RAM. A key disadvantage of flash memory is that it can endure only a relatively small number of write cycles in a specific block.[2]

Flash memory[3] is used in computers, PDAs, digital audio players, digital cameras, mobile phones, synthesizers, video games, scientific instrumentation, industrial robotics, and medical electronics. Flash memory has fast read access time, but it is not as fast as static RAM or ROM. In portable devices, it is preferred to use flash memory because of its mechanical shock resistance since mechanical drives are more prone to mechanical damage.[4]

Because erase cycles are slow, the large block sizes used in flash memory erasing give it a significant speed advantage over non-flash EEPROM when writing large amounts of data. As of 2019, flash memory costs much less than byte-programmable EEPROM and had become the dominant memory type wherever a system required a significant amount of non-volatile solid-state storage. EEPROMs, however, are still used in applications that require only small amounts of storage, as in serial presence detect.[5][6]

Flash memory packages can use die stacking with through-silicon vias and several dozen layers of 3D TLC NAND cells (per die) simultaneously to achieve capacities of up to 1 tebibyte per package using 16 stacked dies and an integrated flash controller as a separate die inside the package.[7][8][9][10]

Background

The origins of flash memory can be traced back to the development of the floating-gate MOSFET (FGMOS), also known as the floating-gate transistor.[11][12] The original MOSFET (metal–oxide–semiconductor field-effect transistor), also known as the MOS transistor, was invented by Egyptian engineer Mohamed M. Atalla and Korean engineer Dawon Kahng at Bell Labs in 1959.[13] Kahng went on to develop a variation, the floating-gate MOSFET, with Chinese engineer Simon Min Sze at Bell Labs in 1967.[14] They proposed that it could be used as floating-gate memory cells for storing a form of programmable read-only memory (PROM) that is both non-volatile and re-programmable.[14]

Early types of floating-gate memory included EPROM (erasable PROM) and EEPROM (electrically erasable PROM) in the 1970s.[14] However, early floating-gate memory required engineers to build a memory cell for each bit of data, which proved to be cumbersome,[15] slow,[16] and expensive, restricting floating-gate memory to niche applications in the 1970s, such as military equipment and the earliest experimental mobile phones.[11]

Invention and commercialization

Fujio Masuoka, while working for Toshiba, proposed a new type of floating-gate memory that allowed entire sections of memory to be erased quickly and easily, by applying a voltage to a single wire connected to a group of cells.[11] This led to Masuoka's invention of flash memory at Toshiba in 1980.[15][17][18] According to Toshiba, the name "flash" was suggested by Masuoka's colleague, Shōji Ariizumi, because the erasure process of the memory contents reminded him of the flash of a camera.[19] Masuoka and colleagues presented the invention of NOR flash in 1984,[20][21] and then NAND flash at the *IEEE 1987 International Electron Devices Meeting* (IEDM) held in San Francisco.[22]

Toshiba commercially launched NAND flash memory in 1987.[1][14] Intel Corporation introduced the first commercial NOR type flash chip in 1988.[23] NOR-based flash has long erase and write times, but provides full address and data buses, allowing random access to any memory location. This makes it a suitable replacement for older read-only memory (ROM) chips, which are used to store program code that rarely needs to be updated, such as a computer's BIOS or the firmware of set-top boxes. Its endurance may be from as little as 100 erase cycles for an on-chip flash memory,[24] to a more typical 10,000 or 100,000 erase cycles, up to 1,000,000 erase cycles.[25] NOR-based flash was the basis of early flash-based removable media; CompactFlash was originally based on it, though later cards moved to less expensive NAND flash.

NAND flash has reduced erase and write times, and requires less chip area per cell, thus allowing greater storage density and lower cost per bit than NOR flash. However, the I/O interface of NAND flash does not provide a random-access external address bus. Rather, data must be read on a block-wise basis, with typical block sizes of hundreds to thousands of bits. This makes NAND flash unsuitable as a drop-in replacement for program ROM, since most microprocessors and microcontrollers require byte-level random access. In this regard, NAND flash is similar to other secondary data storage devices, such as hard disks and optical media, and is thus highly suitable for use in mass-storage devices, such as memory cards and solid-state drives (SSD). Flash memory cards and SSDs store data using multiple NAND flash memory chips.

The first NAND-based removable memory card format was SmartMedia, released in 1995. Many others followed, including MultiMediaCard, Secure Digital, Memory Stick, and xD-Picture Card.

Later developments

A new generation of memory card formats, including RS-MMC, miniSD and microSD, feature extremely small form factors. For example, the microSD card has an area of just over 1.5 cm^2, with a thickness of less than 1 mm.

NAND flash has achieved significant levels of memory density as a result of several major technologies that were commercialized during the late 2000s to early 2010s.[26]

Multi-level cell (MLC) technology stores more than one bit in each memory cell. NEC demonstrated multi-level cell (MLC) technology in 1998, with an 80 Mb flash memory chip storing 2 bits per cell.[27] STMicroelectronics also demonstrated MLC in

2000, with a 64 MB NOR flash memory chip.[28] In 2009, Toshiba and SanDisk introduced NAND flash chips with QLC technology storing 4 bits per cell and holding a capacity of 64 Gbit.[29][30] Samsung Electronics introduced triple-level cell (TLC) technology storing 3-bits per cell, and began mass-producing NAND chips with TLC technology in 2010.[31]

Charge trap flash

Charge trap flash (CTF) technology replaces the polysilicon floating gate, which is sandwiched between a blocking gate oxide above and a tunneling oxide below it, with an electrically insulating silicon nitride layer; the silicon nitride layer traps electrons. In theory, CTF is less prone to electron leakage, providing improved data retention.[32][33][34][35][36][37]

Because CTF replaces the polysilicon with an electrically insulating nitride, it allows for smaller cells and higher endurance (lower degradation or wear). However, electrons can become trapped and accumulate in the nitride, leading to degradation. Leakage is exacerbated at high temperatures since electrons become more excited with increasing temperatures. CTF technology however still uses a tunneling oxide and blocking layer which are the weak points of the technology, since they can still be damaged in the usual ways (the tunnel oxide can be degraded due to extremely high electric fields and the blocking layer due to Anode Hot Hole Injection (AHHI).[38][39]

Degradation or wear of the oxides is the reason why flash memory has limited endurance, and data retention goes down (the potential for data loss increases) with increasing degradation, since the oxides lose their electrically insulating characteristics as they degrade. The oxides must insulate against electrons to prevent them from leaking which would cause data loss.

In 1991, NEC researchers including N. Kodama, K. Oyama and Hiroki Shirai described a type of flash memory with a charge trap method.[40] In 1998, Boaz Eitan of Saifun Semiconductors (later acquired by Spansion) patented a flash memory technology named NROM that took advantage of a charge trapping layer to replace the conventional floating gate used in conventional flash memory designs.[41] In 2000, an Advanced Micro Devices (AMD) research team led by Richard

M. Fastow, Egyptian engineer Khaled Z. Ahmed and Jordanian engineer Sameer Haddad (who later joined Spansion) demonstrated a charge-trapping mechanism for NOR flash memory cells.[42] CTF was later commercialized by AMD and Fujitsu in 2002.[43] 3D V-NAND (vertical NAND) technology stacks NAND flash memory cells vertically within a chip using 3D charge trap flash (CTP) technology. 3D V-NAND technology was first announced by Toshiba in 2007,[44] and the first device, with 24 layers, was first commercialized by Samsung Electronics in 2013.[45][46]

3D integrated circuit technology

3D integrated circuit (3D IC) technology stacks integrated circuit (IC) chips vertically into a single 3D IC chip package.[26]Toshiba introduced 3D IC technology to NAND flash memory in April 2007, when they debuted a 16 GB eMMC compliant (product number THGAM0G7D8DBAI6, often abbreviated THGAM on consumer websites) embedded NAND flash memory chip, which was manufactured with eight stacked 2 GB NAND flash chips.[47] In September 2007, Hynix Semiconductor (now SK Hynix) introduced 24-layer 3D IC technology, with a 16 GB flash memory chip that was manufactured with 24 stacked NAND flash chips using a wafer bonding process.[48] Toshiba also used an eight-layer 3D IC for their 32 GB THGBM flash chip in 2008.[49] In 2010, Toshiba used a 16-layer 3D IC for their 128 GB THGBM2 flash chip, which was manufactured with 16 stacked 8 GB chips.[50] In the 2010s, 3D ICs came into widespread commercial use for NAND flash memory in mobile devices.[26]

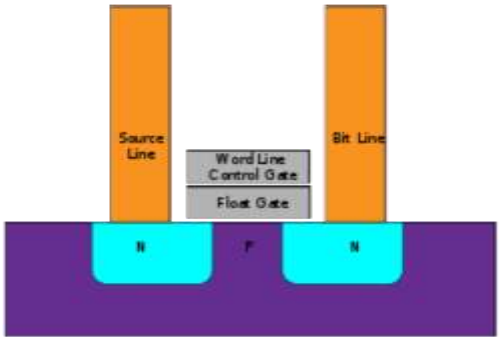
As of August 2017, microSD cards with a capacity up to 400 GB (400 billion bytes) are available.[51][52] The same year, Samsung combined 3D IC chip stacking with its 3D V-NAND and TLC technologies to manufacture its 512 GB KLUFG8R1EM flash memory chip with eight stacked 64-layer V-NAND chips.[53] In 2019, Samsung produced a 1024 GB flash chip, with eight stacked 96-layer V-NAND chips and with QLC technology.[54][55]

Principles of operation

Flash memory stores information in an array of memory cells made from floating-gate transistors. In single-level cell (SLC) devices, each cell stores only one bit of information. Multi-level cell (MLC) devices, including triple-level cell (TLC) devices, can store more than one bit per cell.

The floating gate may be conductive (typically polysilicon in most kinds of flash memory) or non-conductive (as in SONOS flash memory).[56]

Floating-gate MOSFET



A flash memory cell

In flash memory, each memory cell resembles a standard metal–oxide–semiconductor field-effect transistor (MOSFET) except that the transistor has two gates instead of one. The cells can be seen as an electrical switch in which current flows between two terminals (source and drain) and is controlled by a floating gate (FG) and a control gate (CG). The CG is similar to the gate in other MOS transistors, but below this, there is the FG insulated all around by an oxide layer. The FG is interposed between the CG and the MOSFET channel. Because the FG is electrically isolated by its insulating layer, electrons placed on it are trapped. When the FG is charged with electrons, this charge screens the electric field from the CG, thus, increasing the threshold voltage (V_T) of the cell. This means that the V_T of the cell can be changed between the *uncharged FG threshold voltage* (V_{T1}) and the higher *charged FG threshold voltage* (V_{T2}) by changing the FG charge. In order to read a value from the cell, an intermediate voltage (V_I) between V_{T1} and V_{T2} is applied to the CG. If the channel conducts at V_I, the FG must be uncharged (if it were charged, there would not be conduction because V_I is less than V_{T2}). If the channel does not conduct at the V_I, it indicates that the FG is charged. The binary value of the cell is sensed by determining whether there is current flowing through the transistor when V_I is asserted on the CG. In a multi-level cell device, which stores more than one bit per cell, the amount of current flow is sensed (rather than simply its presence or absence), in order to determine more precisely the level of charge on the FG.

Floating gate MOSFETs are so named because there is an electrically insulating tunnel oxide layer between the floating gate and the silicon, so the gate "floats" above the silicon. The oxide keeps the electrons confined to the floating gate. Degradation or wear (and the limited endurance of floating gate Flash memory) occurs due to the extremely high electric feld (10 million volts per centimeter) experienced by the oxide. Such high voltage densities can break atomic bonds over time in the relatively thin oxide, gradually degrading its electrically insulating properties and allowing electrons to be trapped in and pass through freely (leak) from the floating gate into the oxide, increasing the likelihood of data loss since the electrons (the quantity of which is used to represent different charge levels, each assigned to a different combination of bits in MLC Flash) are normally in the floating gate. This is why data retention goes down and the risk of data loss increases with increasing degradation.[57][58][36][59][60]The silicon oxide in a cell degrades with every erase operation. The degradation increases the amount of negative charge in the cell over time due to trapped electrons in the oxide and negates some of the control gate voltage, this over time also makes erasing the cell slower, so to maintain the performance and reliability of the NAND chip, the cell must be retired from use. Endurance also decreases with the number of bits in a cell. With more bits in a cell, the number of possible states (each represented by a different voltage level) in a cell increases and is more sensitive to the voltages used for programming. Voltages may be adjusted to compensate for degradation of the silicon oxide, and as the number of bits increases, the number of possible states also increases and thus the cell is less tolerant of adjustments to programming voltages, because there is less space between the voltage levels that define each state in a cell.[61]

Fowler–Nordheim tunneling

The process of moving electrons from the control gate and into the floating gate is called Fowler–Nordheim tunneling, and it fundamentally changes the characteristics of the cell by increasing the MOSFET's threshold voltage. This, in turn, changes the drain-source current that flows through the transistor for a given gate voltage, which is ultimately used to encode a binary value. The Fowler-Nordheim tunneling effect is reversible, so electrons can be added to or removed from the floating gate, processes traditionally known as writing and erasing.[62]

Internal charge pumps

Despite the need for relatively high programming and erasing voltages, virtually all flash chips today require only a single

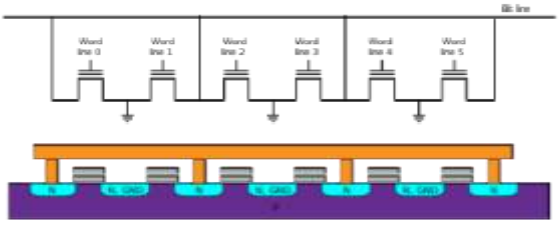
supply voltage and produce the high voltages that are required using on-chip charge pumps.

Over half the energy used by a 1.8 V NAND flash chip is lost in the charge pump itself. Since boost converters are inherently more efficient than charge pumps, researchers developing low-power SSDs have proposed returning to the dual Vcc/Vpp supply voltages used on all early flash chips, driving the high Vpp voltage for all flash chips in an SSD with a single shared external boost converter.[63][64][65][66][67][68][69][70]

In spacecraft and other high-radiation environments, the on-chip charge pump is the first part of the flash chip to fail, although flash memories will continue to work – in read-only mode – at much higher radiation levels.[71]

NOR flash

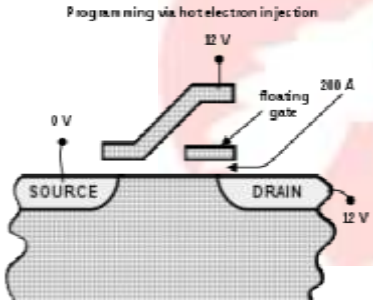
In NOR flash, each cell has one end connected directly to ground, and the other end connected directly to a bit line. This arrangement is called "NOR flash" because it acts like a NOR gate: when one of the word lines (connected to the cell's



NOR flash memory wiring and structure on silicon

CG) is brought high, the corresponding storage transistor acts to pull the output bit line low. NOR flash continues to be the technology of choice for embedded applications requiring a discrete non-volatile memory device. The low read latencies characteristic of NOR devices allow for both direct code execution and data storage in a single memory product.[72]

Programming



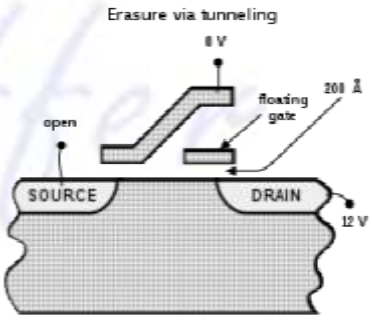
Programming a NOR memory cell (setting it to logical 0), via hot-electron injection

A single-level NOR flash cell in its default state is logically equivalent to a binary "1" value, because current will flow through the channel under application of an appropriate voltage to the control gate, so that the bitline voltage is pulled down. A NOR flash cell can be programmed, or set to a binary "0" value, by the following procedure:

- an elevated on-voltage (typically >5 V) is applied to the CG
- the channel is now turned on, so

electrons can flow from the source to the drain (assuming an NMOS transistor)

- the source-drain current is sufficiently high to cause some high energy electrons to jump through the insulating layer onto the FG, via a process called hot-electron injection.



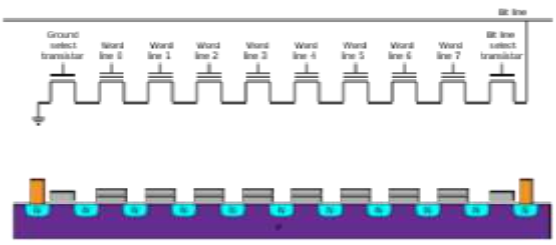
Erasing a NOR memory cell (setting it to logical 1), via quantum tunneling

Erasing

To erase a NOR flash cell (resetting it to the "1" state), a large voltage *of the opposite polarity* is applied between the CG and source terminal, pulling the electrons off the FG through quantum tunneling. Modern NOR flash memory chips are divided into erase segments (often called blocks or sectors). The erase operation can be performed only on a block-wise basis; all the cells in an erase segment must be erased together. Programming of NOR cells, however, generally can be performed one byte or word at a time.

NAND flash

NAND flash also uses floating-gate transistors, but they are connected in a way that resembles a NAND gate: several transistors are connected in series, and the bit line is pulled low only if all the word lines are pulled high (above the transistors' V_T). These groups are then connected via some additional transistors to a NOR-style bit line array in the same way that single transistors are linked in NOR flash.



NAND flash memory wiring and structure on silicon

Compared to NOR flash, replacing single transistors with serial-linked groups adds an extra level of addressing. Whereas NOR flash might address memory by page then word, NAND flash might address it by page, word and bit. Bit-level addressing suits bit-serial applications (such as hard disk emulation), which access only one bit at a time. Execute-in-place applications, on the other hand, require every bit in a word to be accessed simultaneously. This requires word-level addressing. In any case, both bit and word addressing modes are possible with either NOR or NAND flash.

To read data, first the desired group is selected (in the same way that a single transistor is selected from a NOR array). Next, most of the word lines are pulled up above V_{T2}, while one of them is pulled up to V_I. The series group will conduct (and pull the bit line low) if the selected bit has not been programmed.

Despite the additional transistors, the reduction in ground wires and bit lines allows a denser layout and greater storage capacity per chip. (The ground wires and bit lines are actually much wider than the lines in the diagrams.) In addition, NAND flash is typically permitted to contain a certain number of faults (NOR flash, as is used for a BIOS ROM, is expected to be fault-free). Manufacturers try to maximize the amount of usable storage by shrinking the size of the transistors.

NAND Flash cells are read by analysing their response to various voltages.[59]

Writing and erasing

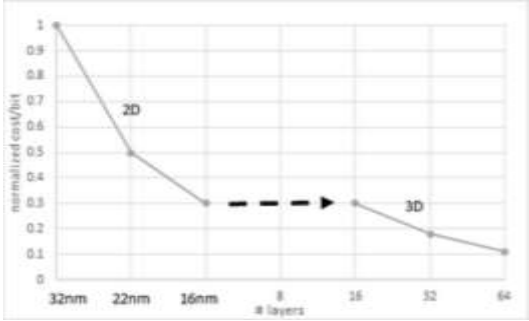
NAND flash uses tunnel injection for writing and tunnel release for erasing. NAND flash memory forms the core of the removable USB storage devices known as USB flash drives, as well as most memory card formats and solid-state drives available today.

The hierarchical structure of NAND flash starts at a cell level which establishes strings, then pages, blocks, planes and ultimately a die. A string is a series of connected NAND cells in which the source of one cell is connected to the drain of the next one. Depending on the NAND technology, a string typically consists of 32 to 128 NAND cells. Strings are organised into pages which are then organised into blocks in which each string is connected to a separate line called a bitline. All cells with the same position in the string are connected through the control gates by a wordline. A plane contains a certain number of blocks that are connected through the same bitline. A flash die consists of one or more planes, and the peripheral circuitry that is needed to perform all the read, write, and erase operations.

The architecture of NAND flash means that data can be read and programmed (written) in pages, typically between 4 KiB and 16 KiB in size, but can only be erased at the level of entire blocks consisting of multiple pages. When a block is erased, all the cells are logically set to 1. Data can only be programmed in one pass to a page in a block that was erased. Any cells that have been set to 0 by programming can only be reset to 1 by erasing the entire block. This means that before new data can be programmed into a page that already contains data, the current contents of the page plus the new data must be copied to a new, erased page. If a suitable erased page is available, the data can be written to it immediately. If no erased page is available, a block must be erased before copying the data to a page in that block. The old page is then marked as invalid and is available for erasing and reuse.[73]

Vertical NAND

Vertical NAND (V-NAND) or 3D NAND memory stacks memory cells vertically and uses a charge trap flash architecture. The vertical layers allow larger areal bit densities without requiring smaller individual cells.[74] It is also sold under the trademark *BiCS Flash*, which is a trademark of Kioxia Corporation (former Toshiba Memory Corporation). 3D NAND was first announced by Toshiba in 2007.[44] V-NAND was first commercially manufactured by Samsung Electronics in 2013.[45][46][75][76]



3D NAND continues scaling beyond 2D.

Structure

V-NAND uses a charge trap flash geometry (which was commercially introduced in 2002 by AMD and Fujitsu)[43] that stores charge on an embedded silicon nitride film. Such a film is more robust against point defects and can be made thicker to hold larger numbers of electrons. V-NAND wraps a planar charge trap cell into a cylindrical form.[74] As of 2020, 3D NAND Flash memories by Micron and Intel instead use floating gates, however, Micron 128 layer and above 3D NAND memories use a conventional charge trap structure, due to the dissolution of the partnership between Micron and Intel. Charge trap 3D NAND Flash is thinner than floating gate 3D NAND. In floating gate 3D NAND, the memory cells are completely separated from one another, whereas in charge trap 3D NAND, vertical groups of memory cells share the same silicon nitride material.[77]

An individual memory cell is made up of one planar polysilicon layer containing a hole filled by multiple concentric vertical cylinders. The hole's polysilicon surface acts as the gate electrode. The outermost silicon dioxide cylinder acts as the gate dielectric, enclosing a silicon nitride cylinder that stores charge, in turn enclosing a silicon dioxide cylinder as the tunnel dielectric that surrounds a central rod of conducting polysilicon which acts as the conducting channel.[74]

Memory cells in different vertical layers do not interfere with each other, as the charges cannot move vertically through the silicon nitride storage medium, and the electric fields associated with the gates are closely confined within each layer. The vertical collection is electrically identical to the serial-linked groups in which conventional NAND flash memory is configured.[74]

Construction

Growth of a group of V-NAND cells begins with an alternating stack of conducting (doped) polysilicon layers and insulating silicon dioxide layers.[74]

The next step is to form a cylindrical hole through these layers. In practice, a 128 Gibit V-NAND chip with 24 layers of memory cells requires about 2.9 billion such holes. Next, the hole's inner surface receives multiple coatings, first silicon dioxide, then silicon nitride, then a second layer of silicon dioxide. Finally, the hole is filled with conducting (doped) polysilicon.[74]

Performance

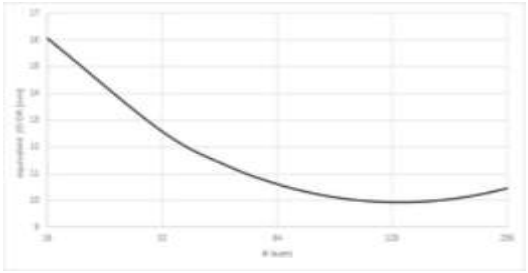
As of 2013, V-NAND flash architecture allows read and write operations twice as fast as conventional NAND and can last up to 10 times as long, while consuming 50 percent less power. They offer comparable physical bit density using 10-nm lithography but may be able to increase bit density by up to two orders of magnitude, given V-NAND's use of up to several hundred layers.[74] As of 2020, V-NAND chips with 160 layers are under development by Samsung.[78]

Cost

The wafer cost of a 3D NAND is comparable with scaled down (32 nm or less) planar NAND Flash.[79] However, with planar NAND scaling stopping at 16 nm, the cost per bit reduction can continue by 3D NAND starting with 16 layers. However, due to the non-vertical sidewall of the hole etched through the layers; even a slight deviation leads to a minimum bit cost, i.e., minimum equivalent design rule (or maximum density), for a given number of layers; this minimum bit cost layer number decreases for smaller hole diameter.[80]

Limitations

Block erasure



Minimum bit cost of 3D NAND from non-vertical sidewall. The top opening widens with more layers, counteracting the increase in bit density

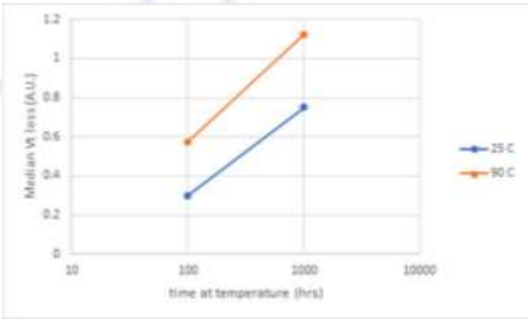
One limitation of flash memory is that it can be erased only a block at a time. This generally sets all bits in the block to 1. Starting with a freshly erased block, any location within that block can be programmed. However, once a bit has been set to 0, only by erasing the entire block can it be changed back to 1. In other words, flash memory (specifically NOR flash) offers random-access read and programming operations but does not offer arbitrary random-access rewrite or erase operations. A location can, however, be rewritten as long as the new value's 0 bits are a superset of the over-written values. For example, a nibble value may be erased to 1111, then written as 1110. Successive writes to that nibble can change it to 1010, then 0010, and finally 0000. Essentially, erasure sets all bits to 1, and programming can only clear bits to 0.[81] Some file systems designed for flash devices make use of this rewrite capability, for example Yaffs1, to represent sector metadata. Other flash file systems, such as YAFFS2, never make use of this "rewrite" capability—they do a lot of extra work to meet a "write once rule".

Although data structures in flash memory cannot be updated in completely general ways, this allows members to be "removed" by marking them as invalid. This technique may need to be modified for multi-level cell devices, where one memory cell holds more than one bit.

Common flash devices such as USB flash drives and memory cards provide only a block-level interface, or flash translation Layer (FTL), which writes to a different cell each time to wear-level the device. This prevents incremental writing within a block; however, it does help the device from being prematurely worn out by intensive write patterns.

Data Retention

Data stored on flash cells is steadily lost due to electron detrapping. The rate of loss increases exponentially as the absolute temperature increases. For example: For a 45 nm NOR Flash, at 1000 hours, the threshold voltage (Vt) loss at 25 deg Celsius is about half that at 90 deg Celsius.[82]



45nm NOR flash memory example of data retention varying with temperatures

Memory wear

Another limitation is that flash memory has a finite number of program – erase cycles (typically written as P/E cycles). Most commercially available flash products are guaranteed to withstand around 100,000 P/E cycles before the wear begins to deteriorate the integrity of the storage.[83] Micron Technology and Sun Microsystems announced an SLC NAND flash memory chip rated for 1,000,000 P/E cycles on 17 December 2008.[84] Longer P/E cycles of Industrial SSDs speak for their endurance level and make them more reliable for Industrial usage.

The guaranteed cycle count may apply only to block zero (as is the case with TSOP NAND devices), or to all blocks (as in NOR). This effect is mitigated in some chip firmware or file system drivers by counting the writes and dynamically remapping blocks in order to spread write operations between sectors; this technique is called wear leveling. Another approach is to perform write verification and remapping to spare sectors in case of write failure, a technique called bad block management (BBM). For portable consumer devices, these wear out management techniques typically extend the life of the flash memory beyond the life of the device itself, and some data loss may be acceptable in these applications. For high-reliability data storage, however, it is not advisable to use flash memory that would have to go through a large number of programming cycles. This limitation is meaningless for 'read-only' applications such as thin clients and routers, which are programmed only once or at most a few times during their lifetimes.

In December 2012, Taiwanese engineers from Macronix revealed their intention to announce at the 2012 IEEE International Electron Devices Meeting that they had figured out how to improve NAND flash storage read/write cycles from

10,000 to 100 million cycles using a "self-healing" process that used a flash chip with "onboard heaters that could anneal small groups of memory cells."[85] The built-in thermal annealing was to replace the usual erase cycle with a local high temperature process that not only erased the stored charge, but also repaired the electron-induced stress in the chip, giving write cycles of at least 100 million.[86] The result was to be a chip that could be erased and rewritten over and over, even when it should theoretically break down. As promising as Macronix's breakthrough might have been for the mobile industry, however, there were no plans for a commercial product featuring this capability to be released any time in the near future.[87]

Read disturb

The method used to read NAND flash memory can cause nearby cells in the same memory block to change over time (become programmed). This is known as read disturb. The threshold number of reads is generally in the hundreds of thousands of reads between intervening erase operations. If reading continually from one cell, that cell will not fail but rather one of the surrounding cells on a subsequent read. To avoid the read disturb problem the flash controller will typically count the total number of reads to a block since the last erase. When the count exceeds a target limit, the affected block is copied over to a new block, erased, then released to the block pool. The original block is as good as new after the erase. If the flash controller does not intervene in time, however, a **read disturb** error will occur with possible data loss if the errors are too numerous to correct with an error-correcting code.[88][89][90]

X-ray effects

Most flash ICs come in ball grid array (BGA) packages, and even the ones that do not are often mounted on a PCB next to other BGA packages. After PCB Assembly, boards with BGA packages are often X-rayed to see if the balls are making proper connections to the proper pad, or if the BGA needs rework. These X-rays can erase programmed bits in a flash chip (convert programmed "0" bits into erased "1" bits). Erased bits ("1" bits) are not affected by X-rays.[91][92]

Some manufacturers are now making X-ray proof SD[93] and USB[94] memory devices.

Low-level access

The low-level interface to flash memory chips differs from those of other memory types such as DRAM, ROM, and EPROM, which support bit-alterability (both zero to one and one to zero) and random access via externally accessible address buses.

NOR memory has an external address bus for reading and programming. For NOR memory, reading and programming are random-access, and unlocking and erasing are block-wise. For NAND memory, reading and programming are page-wise, and unlocking and erasing are block-wise.

NOR memories

Reading from NOR flash is similar to reading from random-access memory, provided the address and data bus are mapped correctly. Because of this, most microprocessors can use NOR flash memory as execute in place (XIP) memory, meaning that programs stored in NOR flash can be executed directly from the NOR flash without needing to be copied into RAM first. NOR flash may be programmed in a random-access manner similar to reading. Programming changes bits from a logical one to a zero. Bits that are already zero are left unchanged. Erasure must happen a block at a time, and resets all the bits in the erased block back to one. Typical block sizes are 64, 128, or 256 KB.

NOR flash by Intel

Bad block management is a relatively new feature in NOR chips. In older NOR devices not supporting bad block management, the software or device driver controlling the memory chip must correct for blocks that wear out, or the device will cease to work reliably.

The specific commands used to lock, unlock, program, or erase NOR memories differ for each manufacturer. To

avoid needing unique driver software for every device made, special Common Flash Memory Interface (CFI) commands allow the device to identify itself and its critical operating parameters.

Besides its use as random-access ROM, NOR flash can also be used as a storage device, by taking advantage of random-access programming. Some devices offer read-while-write functionality so that code continues to execute even while a program or erase operation is occurring in the background. For sequential data writes, NOR flash chips typically have slow write speeds, compared with NAND flash.

Typical NOR flash does not need an error correcting code.[95]

NAND memories

NAND flash architecture was introduced by Toshiba in 1989.[96] These memories are accessed much like block devices, such as hard disks. Each block consists of a number of pages. The pages are typically 512,[97] 2,048 or 4,096 bytes in size. Associated with each page are a few bytes (typically 1/32 of the data size) that can be used for storage of an error orrecting code (ECC) checksum.

Typical block sizes include:

- **32 pages of 512+16 bytes each for a block size (effective) of 16 KiB**
- **64 pages of 2,048+64 bytes each for a block size of 128 KiB**[99]
- **64 pages of 4,096+128 bytes each for a block size of 256 KiB**[100]
- **128 pages of 4,096+128 bytes each for a block size of 512 KiB.**

While reading and programming is performed on a page basis, erasure can only be performed on a block basis.[100]

NAND devices also require bad block management by the device driver software or by a separate controller chip. Some SD cards, for example, include controller circuitry to perform bad block management and wear leveling. When a logical block is accessed by high-level software, it is mapped to a physical block by the device driver or controller. A number of blocks on the flash chip may be set aside for storing mapping tables to deal with bad blocks, or the system may simply check each block at power-up to create a bad block map in RAM. The overall memory capacity gradually shrinks as more blocks are marked as bad.

NAND relies on ECC to compensate for bits that may spontaneously fail during normal device operation. A typical ECC will correct a one-bit error in each 2048 bits (256 bytes) using 22 bits of ECC, or a one-bit error in each 4096 bits (512 bytes) using 24 bits of ECC.[101] If the ECC cannot correct the error during read, it may still detect the error. When doing erase or program operations, the device can detect blocks that fail to program or erase and mark them bad. The data is then written to a different, good block, and the bad block map is updated.

Hamming codes are the most commonly used ECC for SLC NAND flash. Reed-Solomon codes and B CH codes (Bose-Chaudhuri-Hocquenghem codes) are commonly used ECC for MLC NAND flash. Some MLC NAND flash chips internally generate the appropriate BCH error correction codes.[95]

Most NAND devices are shipped from the factory with some bad blocks. These are typically marked according to a specified bad block marking strategy. By allowing some bad blocks, manufacturers achieve far higher yields than would be possible if all blocks had to be verified to be good. This significantly reduces NAND flash costs and only slightly decreases the storage capacity of the parts.

When executing software from NAND memories, virtual memory strategies are often used: memory contents must first be paged or copied into memory-mapped RAM and executed there (leading to the common combination of NAND + RAM). A memory management unit (MMU) in the system is helpful, but this can also be accomplished with overlays. For this reason, some systems will use a combination of NOR and NAND memories, where a smaller NOR memory is used as software ROM and a larger NAND memory is partitioned with a file system for use as a non-volatile data storage area.

NAND sacrifices the random-access and execute-in-place advantages of NOR. NAND is best suited to systems

requiring high capacity data storage. It offers higher densities, larger capacities, and lower cost. It has faster erases, sequential writes, and sequential reads.

Standardization

A group called the Open NAND Flash Interface Working Group (ONFI) has developed a standardized low-level interface for NAND flash chips. This allows interoperability between conforming NAND devices from different vendors. The ONFI specification version 1.0[102] was released on 28 December 2006. It specifies:

- **A standard physical interface (pinout) for NAND flash in TSOP-48, WSOP-48, LGA-52, and BGA-63 packages**
- **A standard command set for reading, writing, and erasing NAND flash chips**
- **A mechanism for self-identification (comparable to the serial presence detection feature of SDRAM memory modules)**

The ONFI group is supported by major NAND flash manufacturers, including Hynix, Intel, Micron Technology, and Numonyx, as well as by major manufacturers of devices incorporating NAND flash chips.[103]

Two major flash device manufacturers, Toshiba and Samsung, have chosen to use an interface of their own design known as Toggle Mode (and now Toggle). This interface isn't pin-to-pin compatible with the ONFI specification. The result is that a product designed for one vendor's devices may not be able to use another vendor's devices.[104]

A group of vendors, including Intel, Dell, and M icrosoft, formed a Non-Volatile Memory Host Controller Interface (NVMHCI) Working Group.[105] The goal of the group is to provide standard software and hardware programming interfaces for nonvolatile memory subsystems, including the "flash cache" device connected to the PCI Express bus.

Distinction between NOR and NAND flash

NOR and NAND flash differ in two important ways:

- The connections of the individual memory cells are different.
- The interface provided for reading and writing the memory is different; NOR allows random access, while NAND allows only page access.[106]

NOR and NAND flash get their names from the structure of the interconnections between memory cells. In NOR flash, cells are connected in parallel to the bit lines, allowing cells to be read and programmed individually. The parallel connection of cells resembles the parallel connection of transistors in a CMOS NOR gate. In NAND flash, cells are connected in series, resembling a CMOS NAND gate. The series connections consume less space than parallel ones, reducing the cost of NAND flash. It does not, by itself, prevent NAND cells from being read and programmed individually.

Each NOR flash cell is larger than a NAND flash cell – 10 F^2 vs 4 F^2 – even when using exactly the same semiconductor device fabrication and so each transistor, contact, etc. is exactly the same size – because NOR flash cells require a separate metal contact for each cell.[107]

Because of the series connection and removal of wordline contacts, a large grid of NAND flash memory cells will occupy perhaps only 60% of the area of equivalent NOR cells[108] (assuming the same C MOS process resolution, for example, 130 nm, 90 nm, or 65 nm). NAND flash's designers realized that the area of a NAND chip, and thus the cost, could be further reduced by removing the external address and data bus circuitry. Instead, external devices could communicate with NAND flash via sequential-accessed command and data registers, which would internally retrieve and output the necessary data. This design choice made random-access of NAND flash memory impossible, but the goal of NAND flash was to replace mechanical hard disks, not to replace ROMs

Attribute	NAND	NOR
Main application	File storage	Code execution
Storage capacity	High	Low
Cost per bit	Low	
Active power	Low	
Standby power		Low
Write speed	Fast	
Read speed		Fast
Execute in place (XIP)	No	Yes
Reliability		High

Write endurance

The write endurance of SLC floating-gate NOR flash is typically equal to or greater than that of NAND flash, while MLC NOR and NAND flash have similar endurance capabilities. Examples of endurance cycle ratings listed in datasheets for NAND and NOR flash, as well as in storage devices using flash memory, are provided.[109]

Type of flash memory	Endurance rating (erases per block)	Example(s) of flash memory or storage device
SLC NAND	100,000	Samsung OneNAND KFW4G16Q2M, Toshiba SLC NAND Flash chips,[110][111][112][113][114] Transcend SD500, Fujitsu S26361-F3298
MLC NAND	5,000 to 10,000 for medium-capacity applications; 1,000 to 3,000 for high-capacity applications[115]	Samsung K9G8G08U0M (Example for medium-capacity applications), Memblaze PBlaze4,[116] ADATA SU900, Mushkin Reactor
TLC NAND	1,000	Samsung SSD 840
QLC NAND	?	SanDisk X4 NAND flash SD cards[117][118][119][120]
3D SLC NAND	100,000	Samsung Z-NAND[121]
3D MLC NAND	6,000 to 40,000	Samsung SSD 850 PRO, Samsung SSD 845DC PRO,[122][123] Samsung 860 PRO
3D TLC NAND	1,000 to 3,000	Samsung SSD 850 EVO, Samsung SSD 845DC EVO, Crucial MX300[124][125][126], Memblaze PBlaze5 900, Memblaze PBlaze5 700, Memblaze PBlaze5 910/916,Memblaze PBlaze5 510/516,[127][128][129][130] ADATA SX 8200 PRO (also being sold under "XPG Gammix" branding, model S11 PRO)
3D QLC NAND	100 to 1,000	Samsung SSD 860 QVO SATA, Intel SSD 660p, Samsung SSD 980 QVO NVMe, Micron 5210 ION, Samsung SSD BM991 NVMe[131][132][133][134][135][136][137][138]
3D PLC NAND	Unknown	In development by SK Hynix (formerly Intel)[139] and Kioxia (formerly Toshiba Memory).[115]
SLC(floating-gate) NOR	100,000 to 1,000,000	Numonyx M58BW (Endurance rating of 100,000 erases per block); Spansion S29CD016J (Endurance rating of 1,000,000 erases per block)
MLC(floating-gate) NOR	100,000	Numonyx J3 flash

However, by applying certain algorithms and design paradigms such as wear leveling and memory over-provisioning, the endurance of a storage system can be tuned to serve specific requirements.[140]

In order to compute the longevity of the NAND flash, one must account for the size of the memory chip, the type of memory (e.g. SLC/MLC/TLC), and use pattern. Industrial NAND are in demand due to their capacity, longer endurance and reliability in sensitive environments.

3D NAND performance may degrade as layers are added.[121]

As the number of bits per cell increases, the performance of NAND flash may degrade, increasing random read times to 100µs for TLC NAND which is 4 times the time required in SLC NAND, and twice the time required in MLC NAND, for random reads.[141]

Flash file systems

Because of the particular characteristics of flash memory, it is best used with either a controller to perform wear leveling and error correction or specifically designed flash file systems, which spread writes over the media and deal with the long erase times of NOR flash blocks. The basic concept behind flash file systems is the following: when the flash store is to be updated, the file system will write a new copy of the changed data to a fresh block, remap the file pointers, then erase the old block later when it has time.

In practice, flash file systems are used only for memory technology devices (MTDs), which are embedded flash memories that do not have a controller. Removable flash memory cards, SSDs, eMMC/eUFS chips and USB flash drives have built-in controllers to perform wear leveling and error correction so use of a specific flash file system may not add benefit.

Capacity

Multiple chips are often arrayed or die stacked to achieve higher capacities[142] for use in consumer electronic devices such as multimedia players or GPSs. The capacity scaling (increase) of flash chips used to follow Moore's law because they are manufactured with many of the same integrated circuits techniques and equipment. Since the introduction of 3D NAND, scaling is no longer necessarily associated with Moore's law since ever smaller transistors (cells) are no longer used.

Consumer flash storage devices typically are advertised with usable sizes expressed as a small integer power of two (2, 4, 8, etc.) and a designation of megabytes (MB) or gigabytes (GB); e.g., 512 MB, 8 GB. This includes SSDs marketed as hard drive replacements, in accordance with traditional hard drives, which use decimal prefixes.[143] Thus, an SSD marked as "64 GB" is at least 64×1000^3 bytes (64 GB). Most users will have slightly less capacity than this available for their files, due to the space taken by file system metadata.

The flash memory chips inside them are sized in strict binary multiples, but the actual total capacity of the chips is not usable at the drive interface. It is considerably larger than the advertised capacity in order to allow for distribution of writes (wear leveling), for sparing, for error correction codes, and for other metadata needed by the device's internal firmware.

In 2005, Toshiba and SanDisk developed a NAND flash chip capable of storing 1 GB of data using multi-level cell (MLC) technology, capable of storing two bits of data per cell. In September 2005, Samsung Electronics announced that it had developed the world's first 2 GB chip.[144]

In March 2006, Samsung announced flash hard drives with a capacity of 4 GB, essentially the same order of magnitude as smaller laptop hard drives, and in September 2006, Samsung announced an 8 GB chip produced using a 40 nm manufacturing process.[145] In January 2008, SanDisk announced availability of their 16 GB MicroSDHC and 32 GB SDHC Plus cards.[146][147]

More recent flash drives (as of 2012) have much greater capacities, holding 64, 128, and 256 GB.[148]

A joint development at Intel and Micron will allow the production of 32-layer 3.5 terabyte (TB) NAND flash sticks and 10 TB standard-sized SSDs. The device includes 5 packages of 16 × 48 GB TLC dies, using a floating gate cell

design.[149]

Flash chips continue to be manufactured with capacities under or around 1 MB (e.g. for BIOS-ROMs and embedded applications).

In July 2016, Samsung announced the 4 TB Samsung 850 EVO which utilizes their 256 Gbit 48-layer TLC 3D V-NAND.[150] In August 2016, Samsung announced a 32 TB 2.5-inch SAS SSD based on their 512 Gbit 64-layer TLC 3D V-NAND. Further, Samsung expects to unveil SSDs with up to 100 TB of storage by 2020.[151]

Transfer rates

Flash memory devices are typically much faster at reading than writing.[152] Performance also depends on the quality of storage controllers, which become more critical when devices are partially full.[152] Even when the only change to manufacturing is die-shrink, the absence of an appropriate controller can result in degraded speeds.[153]

Applications

Serial flash

Serial Flash: Silicon Storage Tech SST25VF080B

Serial flash is a small, low-power flash memory that provides only serial access to the data - rather than addressing individual bytes, the user reads or writes large contiguous groups of bytes in the address space serially. Serial Peripheral Interface Bus (SPI) is a typical protocol for accessing the device. When incorporated into an embedded system, serial flash requires fewer wires on the PCB than parallel flash memories, since it transmits and receives data one bit at a time. This may permit a reduction in board space, power consumption, and total system cost.

There are several reasons why a serial device, with fewer external pins than a parallel device, can significantly reduce overall cost:

- **Many ASICs are pad-limited, meaning that the size of the die is constrained by the number of wire bond pads, rather than the complexity and number of gates used for the device logic. Eliminating bond pads thus permits a more compact integrated circuit, on a smaller die; this increases the number of dies that may be fabricated on a wafer, and thus reduces the cost per die.**
- **Reducing the number of external pins also reduces assembly and packaging costs. A serial device may be packaged in a smaller and simpler package than a parallel device.**
- **Smaller and lower pin-count packages occupy less PCB area.**
- **Lower pin-count devices simplify PCB routing.**

There are two major SPI flash types. The first type is characterized by small pages and one or more internal SRAM page buffers allowing a complete page to be read to the buffer, partially modified, and then written back (for example, the Atmel AT45 *DataFlash* or the Micron Technology Page Erase NOR Flash). The second type has larger sectors where the smallest sectors typically found in this type of SPI flash are 4 kB, but they can be as large as 64 kB. Since this type of SPI flash lacks an internal SRAM buffer, the complete page must be read out and modified before being written back, making it slow to manage. However, the second type is cheaper than the first and is therefore a good choice when the application is code shadowing.

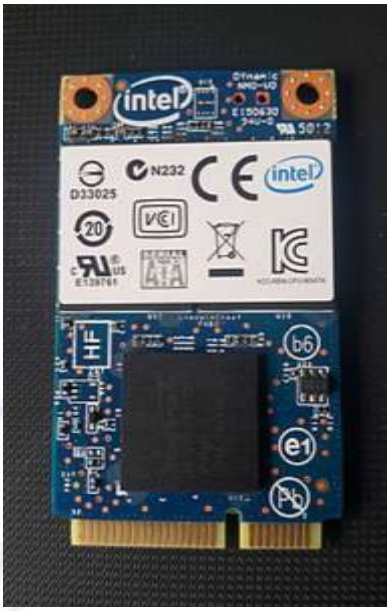
The two types are not easily exchangeable, since they do not have the same pinout, and the command sets are incompatible.

Most FPGAs are based on SRAM configuration cells and require an external configuration device, often a serial flash chip, to reload the configuration bitstream every power cycle.[154]

With the increasing speed of modern CPUs, parallel flash devices are often much slower than the memory bus of the computer they are connected to. Conversely, modern SRAM offers access times below 10 ns, while DDR2 SDRAM offers access times below 20 ns. Because of this, it is often desirable to shadow code stored in flash into RAM; that is, the code is copied from flash into RAM before execution, so that the CPU may access it at full speed. Device firmware may be stored in a serial flash chip, and then copied into SDRAM or SRAM when the device is powered-up.[155] Using an external serial flash device rather than on-chip flash removes the need for significant process compromise (a manufacturing process that is good for high-speed logic is generally not good for flash and vice versa). Once it is decided to read the firmware in as one big block it is common to add compression to allow a smaller flash chip to be used. Since 2005, many devices use serial NOR flash to deprecate parallel NOR flash for firmware storage. Typical applications for serial flash include storing firmware for hard drives, Ethernet network interface adapters, DSL modems, etc.

Flash memory as a replacement for hard drives

One more recent application for flash memory is as a replacement for hard disks. Flash memory does not have the mechanical limitations and latencies of hard drives, so a solid-state drive (SSD) is attractive when considering speed, noise, power consumption, and reliability. Flash drives are gaining traction as mobile device secondary storage devices; they are also used as substitutes for hard drives in high-performance desktop computers and some servers with RAID and SAN architectures.



An Intel mSATA SSD

There remain some aspects of flash-based SSDs that make them unattractive. The cost per gigabyte of flash memory remains significantly higher than that of hard disks.[156] Also flash memory has a finite number of P/E (*program/erase*) cycles, but this seems to be currently under control since warranties on flash-based SSDs are approaching those of current hard drives.[157] In addition, deleted files on SSDs can remain for an indefinite period of time before being overwritten by fresh data; erasure or shred techniques or software that work well on magnetic hard disk drives have no effect on SSDs, compromising security and forensic examination. However, due to the so-called *TRIM* command employed by most solid state drives, which marks the logical block addresses occupied by the deleted file as unused to enable garbage collection, data recovery software is not able to restore files deleted from such.

For relational databases or other systems that require ACID transactions, even a modest amount of flash storage can offer vast speedups over arrays of disk drives.[158][159]

In May 2006, Samsung Electronics announced two flash-memory based PCs, the Q1-SSD and Q30-SSD were expected to become available in June 2006, both of which used 32 GB SSDs, and were at least initially available only in South Korea.[160] The Q1-SSD and Q30-SSD launch was delayed and finally was shipped in late August 2006.[161]

The first flash-memory based PC to become available was the Sony Vaio UX90, announced for pre-order on 27 June 2006 and began to be shipped in Japan on 3 July 2006 with a 16Gb flash memory hard drive.[162] In late September 2006 Sony upgraded the flash-memory in the Vaio UX90 to 32Gb.[163]

A solid-state drive was offered as an option with the first MacBook Air introduced in 2008, and from 2010 onwards, all models were shipped with an SSD. Starting in late 2011, as part of Intel's Ultrabook initiative, an increasing number of ultra-thin laptops are being shipped with SSDs standard.

There are also hybrid techniques such as hybrid drive and ReadyBoost that attempt to combine the advantages of both technologies, using flash as a high-speed non-volatile cache for files on the disk that are often referenced, but rarely modified, such as application and operating system executable files.

Flash memory as RAM

As of 2012, there are attempts to use flash memory as the main computer memory, DRAM[164]

Archival or long-term storage

Floating-gate transistors in the flash storage device hold charge which represents data. This charge gradually leaks over time, leading to an accumulation of logical errors, also known as "bit rot" or "bit fading".[165]

Data retention

It is unclear how long data on flash memory will persist under archival conditions (i.e., benign temperature and humidity with infrequent access with or without prophylactic rewrite). Datasheets of Atmel's flash-based "ATmega" microcontrollers typically promise retention times of 20 years at 85 °C (185 °F) and 100 years at 25 °C (77 °F).[166]

The retention span varies among types and models of flash storage. When supplied with power and idle, the charge of the transistors holding the data is routinely refreshed by the firmware of the flash storage.[165] The ability to retain data varies among flash storage devices due to differences in firmware, data redundancy, and error correction algorithms.[167]

An article from CMU in 2015 states "Today's flash devices, which do not require flash refresh, have a typical retention age of 1 year at room temperature." And that retention time decreases exponentially with increasing temperature. The phenomenon can be modeled by the Arrhenius equation.[168][169]

FPGA configuration

Some FPGAs are based on flash configuration cells that are used directly as (programmable) switches to connect internal elements together, using the same kind of floating-gate transistor as the flash data storage cells in data storage devices.[154]

Industry

One source states that, in 2008, the flash memory industry includes about US$9.1 billion in production and sales. Other sources put the flash memory market at a size of more than US$20 billion in 2006, accounting for more than eight percent of the overall semiconductor market and more than 34 percent of the total semiconductor memory market.[170] In 2012, the market was estimated at $26.8 billion.[171] It can take up to 10 weeks to produce a flash memory chip.[172]

Manufacturers

The following were the largest NAND flash memory manufacturers, as of the first quarter of 2019.[173]

1. **Samsung Electronics – 34.9%**
2. **Kioxia – 18.1%**
3. **Western Digital Corporation – 14%**
4. **Micron Technology – 13.5%**
5. **SK Hynix – 10.3%**
6. **Intel – 8.7% Note: SK Hynix acquired Intel's NAND business at the end of 2021[174]**

Samsung remains the largest NAND flash memory manufacturer as of first quarter 2022.[175]

Shipments

Flash memory shipments (est. manufactured units)

Year(s)	Discrete flash memory chips	Flash memory data capacity (gigabytes)	Floating-gate MOSFET memory cells (billions)
1992	26,000,000[176]	3[176]	24[a]

1993	73,000,000[176]	17[176]	139[a]
1994	112,000,000[176]	25[176]	203[a]
1995	235,000,000[176]	38[176]	300[a]
1996	359,000,000[176]	140[176]	1,121[a]
1997	477,200,000+[177]	317+[177]	2,533+[a]
1998	762,195,122[178]	455+[177]	3,642+[a]
1999	12,800,000,000[179]	635+[177]	5,082+[a]
2000–2004		134,217,728,000 (NAND)[180]	1,073,741,824,000 (NAND)[180]
2005–2007	?		
2008	1,226,215,645 (mobile NAND)[181]		
2009	1,226,215,645+ (mobile NAND)		
2010	7,280,000,000+[b]		
2011	8,700,000,000[183]		
2012	5,151,515,152 (serial)[184]		
2013	?		
2014	?	59,000,000,000[185]	118,000,000,000+[a]
2015	7,692,307,692 (NAND)[186]	85,000,000,000[187]	170,000,000,000+[a]
2016	?	100,000,000,000[188]	200,000,000,000+[a]
2017	?	148,200,000,000[c]	296,400,000,000+[a]
2018	?	231,640,000,000[d]	463,280,000,000+[a]
2019	?	?	?
2020	?	?	?
1992–2020	**45,358,454,134+ memory chips**	**758,057,729,630+ gigabytes**	**2,321,421,837,044 billion+ cells**

In addition to individual flash memory chips, flash memory is also embedded in microcontroller (MCU) chips and system-on-chip (SoC) devices.[192] Flash memory is embedded in ARM chips,[192] which have sold 150 billion units worldwide as of 2019,[193] and in programmable system-on-chip (PSoC) devices, which have sold 1.1 billion units as of 2012.[194] This adds up to at least 151.1 billion MCU and SoC chips with embedded flash memory, in addition to the 45.4 billion known individual flash chip sales as of 2015, totalling at least 196.5 billion chips containing flash memory.

Flash scalability

Due to its relatively simple structure and high demand for higher capacity, NAND flash memory is the most aggressively scaled technology among electronic devices. The heavy competition among the top few manufacturers only adds to the aggressiveness in shrinking the floating-gate MOSFET design rule or process technology node.[89] While the expected shrink timeline is a factor of two every three years per original version of Moore's law, this has recently been accelerated in the case of NAND flash to a factor of two every two years.

ITRS or company	2010	2011	2012	2013	2014	2015	2016	2017	2018
ITRS Flash Roadmap 2011[195]	32 nm	22 nm	20 nm	18 nm	16 nm				
Updated ITRS Flash Roadmap[196]					17 nm	15 nm	14 nm		
Samsung[195][196][197] (Samsung 3D NAND)[196]	35–20 nm[31]	27 nm	21 nm (MLC, TLC)	19–16 nm 19–10 nm (MLC, TLC)[198]	19–10 nm V-NAND (24L)	16–10 nm V-NAND (32L)	16–10 nm	12–10 nm	12–10 nm
Micron, Intel[195][196][197]	34–25 nm	25 nm	20 nm (MLC + HKMG)	20 nm (TLC)	16 nm	16 nm 3D NAND	16 nm 3D NAND	12 nm 3D NAND	12 nm 3D NAND
Toshiba, WD (SanDisk)[195][196][197]	43–32 nm 24 nm (Toshiba)[199]	24 nm	19 nm (MLC, TLC)		15 nm	15 nm 3D NAND	15 nm 3D NAND	12 nm 3D NAND	12 nm 3D NAND
SK Hynix[195][196][197]	46–35 nm	26 nm	20 nm (MLC)		16 nm	16 nm	16 nm	12 nm	12 nm

As the MOSFET feature size of flash memory cells reaches the 15–16 nm minimum limit, further flash density increases will be driven by TLC (3 bits/cell) combined with vertical stacking of NAND memory planes. The decrease in endurance and increase in uncorrectable bit error rates that accompany feature size shrinking can be compensated by improved error correction mechanisms.[200] Even with these advances, it may be impossible to economically scale flash to smaller and smaller dimensions as the number of electron holding capacity reduces. Many promising new technologies (such as FeRAM, MRAM, PMC, PCM, ReRAM, and others) are under investigation and development as possible more scalable replacements for flash.[201]

Timeline

Date of introduction	Chip name	Memory Package Capacity Megabits (Mb), Gigabits (Gb), Terabits (Tb)	Flash type	Cell type	Layers or Stacks of Layers	Manufacturer(s)	Process	Area	Ref
1984	?	?	NOR	SLC	1	Toshiba	?	?	[20]
1985	?	256 kb	NOR	SLC	1	Toshiba	2,000nm	?	[28]
1987	?	?	NAND	SLC	1	Toshiba	?	?	[1]

1989	?	1 Mb	NOR	SLC	1	Seeq, Intel	?	?	[28]
		4 Mb	NAND	SLC	1	Toshiba	1,000nm		
1991	?	16 Mb	NOR	SLC	1	Mitsubishi	600 nm	?	[28]
1993	DD28F032SA	32 Mb	NOR	SLC	1	Intel	?	280mm²	[202][203]
1994	?	64 Mb	NOR	SLC	1	NEC	400 nm	?	[28]
1995	?	16 Mb	DINOR	SLC	1	Mitsubishi, Hitachi	?	?	[28][204]
			NAND	SLC	1	Toshiba	?	?	[205]
		32 Mb	NAND	SLC	1	Hitachi, Samsung, Toshiba	?	?	[28]
		34 Mb	Serial	SLC	1	SanDisk			
1996	?	64 Mb	NAND	SLC	1	Hitachi, Mitsubishi	400 nm	?	[28]
				QLC	1	NEC			
		128 Mb	NAND	SLC	1	Samsung, Hitachi	?		
1997	?	32 Mb	NOR	SLC	1	Intel, Sharp	400 nm	?	[206]
			NAND	SLC	1	AMD, Fujitsu	350 nm		
1999	?	256 Mb	NAND	SLC	1	Toshiba	250 nm	?	[28]
				MLC	1	Hitachi	1		
2000	?	32 Mb	NOR	SLC	1	Toshiba	250 nm	?	[28]
		64 Mb	NOR	QLC	1	STMicroelectronics	180 nm		
		512 Mb	NAND	SLC	1	Toshiba	?	?	[207]
2001	?	512 Mb	NAND	MLC	1	Hitachi	?	?	[28]
		1 Gibit	NAND	MLC	1	Samsung			
					1	Toshiba, SanDisk	160 nm	?	[208]
2002	?	512 Mb	NROM	MLC	1	Saifun	170 nm	?	[28]
		2 Gb	NAND	SLC	1	Samsung, Toshiba	?	?	[209][210]
2003	?	128 Mb	NOR	MLC	1	Intel	130 nm	?	[28]

		1 Gb	NAND	MLC	1	Hitachi			
2004	?	8 Gb	NAND	SLC	1	Samsung	60 nm	?	[209]
2005	?	16 Gb	NAND	SLC	1	Samsung	50 nm	?	[31]
2006	?	32 Gb	NAND	SLC	1	Samsung	40 nm		
Apr-07	THGAM	128 Gb	Stacked NAND	SLC		Toshiba	56 nm	252mm²	[47]
Sep-07	?	128 Gb	Stacked NAND	SLC		Hynix	?	?	[48]
2008	THGBM	256 Gb	Stacked NAND	SLC		Toshiba	43 nm	353mm²	[49]
2009	?	32 Gb	NAND	TLC		Toshiba	32 nm	113 mm²	[29]
		64 Gb	NAND	QLC		Toshiba, SanDisk	43 nm	?	[29][30]
2010	?	64 Gb	NAND	SLC		Hynix	20 nm	?	[211]
				TLC		Samsung	20 nm	?	[31]
	THGBM2	1 Tb	Stacked NAND	QLC		Toshiba	32 nm	374mm²	[50]
2011	KLMCG8GE4A	512 Gb	Stacked NAND	MLC		Samsung	?	192 mm²	[212]
2013	?	?	NAND	SLC		SK Hynix	16 nm	?	[211]
		128 Gb	V-NAND	TLC		Samsung	10 nm	?	
2015	?	256 Gb	V-NAND	TLC		Samsung	?	?	[198]
2017	eUFS 2.1	512 Gb	V-NAND	TLC	8 of 64	Samsung	?	?	[53]
		768 Gb	V-NAND	QLC		Toshiba	?	?	[213]
	KLUFG8R1EM	4 Tb	Stacked V-NAND	TLC		Samsung	?	150mm²	[53]
2018	?	1 Tb	V-NAND	QLC		Samsung	?	?	[214]
		1.33 Tb	V-NAND	QLC		Toshiba	?	158mm²	[215][216]
2019	?	512 Gb	V-NAND	QLC		Samsung	?	?	[54][55]
		1 Tb	V-NAND	TLC		SK Hynix	?	?	[217]
	eUFS 2.1	1 Tb	Stacked V-NAND[218]	QLC	16 of 64	Samsung	?	150mm²	[54][55][219]

See also

- eMMC
- Flash memory controller
- List of flash file systems
- List of flash memory controller manufacturers
- microSDXC (up to 2 TB), and the successor format Secure Digital Ultra Capacity (SDUC) supporting cards up to 128 TiB
- Open NAND Flash Interface Working Group
- Read-mostly memory (RMM)
- Universal Flash Storage
- USB flash drive security
- Write amplification

Notes

a. Single-level cell (1-bit per cell) up until 2009. Multi-level cell (up to 4-bit or half-byte per cell) commercialised in 2009.[29][30]

b. Flash memory chip shipments in 2010:
   - NOR – 3.64 billion[182]
   - NAND – 3.64 billion+ (est.)

c. Flash memory data capacity shipments in 2017:
   - NAND non-volatile memory (NVM) – 85 exabytes (est.)[189]
   - Solid-state drive (SSD) – 63.2 exabytes[190]

d. Flash memory data capacity shipments in 2018 (est.)
   - NAND NVM – 140 exabytes[189]
   - SSD – 91.64 exabytes[191]

References

1. "1987: Toshiba Launches NAND Flash" (https://www.eweek.com/storage/1987-toshiba-launches-nand-flash). *eWeek*. 11 April 2012. Retrieved 20 June 2019.
2. "A Flash Storage Technical and Economic Primer" (http://www.flashstorage.com/flash-storage-technical-ec onomic-primer/). *FlashStorage.com*. 30 March 2015. Archived (https://web.archive.org/web/201507202208 44/http://www.flashstorage.com/flash-storage-technical-economic-primer/) from the original on 20 July 2015.
3. "What is Flash Memory" (https://www.bitwarsoft.com/what-is-flash-memory.html).

Bitwarsoft.com. 22 July 2020.

4. "HDD vs SSD: What Does the Future for Storage Hold?" (https://www.backblaze.com/blog/ssd-vs-hdd-futu re-of-storage/). *backblaze.com*. 6 March 2018. Archived (https://web.archive.org/web/20221222025652/https://www.backblaze.com/blog/ssd-vs-hdd-future-of-storage/) from the original on 22 December 2022.

5. "TN-04-42: Memory Module Serial Presence-Detect Introduction" (https://www.micron.com/-/media/client/gl obal/documents/products/technical-note/dram-modules/tn_04_42.pdf?rev=e5a1537ce3214de5b695f17c3 40fd023) (PDF). Micron. Retrieved 1 June 2022.

6. "What is serial presence detect (SPD)? - Definition from Whatls.com" (https://whatis.techtarget.com/definiti on/serial-presence-detect-SPD). *Whatls.com*.

7. Shilov, Anton. "Samsung Starts Production of 1 TB eUFS 2.1 Storage for Smartphones" (https://www.anan dtech.com/show/13918/samsung-starts-production-of-1-tb-eufs-21-storage-for-smartphones). *AnandTech.com*.

8. Shilov, Anton. "Samsung Starts Production of 512 GB UFS NAND Flash Memory: 64-Layer V-NAND, 860 MB/s Reads" (https://www.anandtech.com/show/12120/samsung-starts-production-of-512-gb-ufs-chips). *AnandTech.com*.

9. Kim, Chulbum; Cho, Ji-Ho; Jeong, Woopyo; Park, Il-han; Park, Hyun-Wook; Kim, Doo-Hyun; Kang, Daewoon; Lee, Sunghoon; Lee, Ji-Sang; Kim, Wontae; Park, Jiyoon; Ahn, Yang-lo; Lee, Jiyoung; Lee, Jong-Hoon; Kim, Seungbum; Yoon, Hyun-Jun; Yu, Jaedoeg; Choi, Nayoung; Kwon, Yelim; Kim, Nahyun; Jang, Hwajun; Park, Jonghoon; Song, Seunghwan; Park, Yongha; Bang, Jinbae; Hong, Sangki; Jeong, Byunghoon; Kim, Hyun-Jin; Lee, Chunan; et al. (2017). "11.4 a 512Gb 3b/Cell 64-stacked WL 3D V-NAND flash memory". *2017 IEEE International Solid-State Circuits Conference (ISSCC)*. pp. 202–203. doi:10.1109/ISSCC.2017.7870331 (https://doi.org/10.1109%2FISSCC.2017.7870331). ISBN 978-1-5090- 3758-2. S2CID 206998691 (https://api.semanticscholar.org/CorpusID:206998691).

10. "Samsung enables 1TB eUFS 2.1 smartphones - Storage - News - HEXUS.net" (https://m.hexus.net/tech/n ews/storage/127010-samsung-enables-1tb-eufs-21-smartphones/). *m.hexus.net*.

11. "Not just a flash in the pan" (https://www.economist.com/technology-quarterly/2006/03/11/not-just-a-flash-i n-the-pan). *The Economist*. 11 March 2006. Retrieved 10 September 2019.

12. Bez, R.; Pirovano, A. (2019). *Advances in Non-Volatile Memory and Storage Technology*. Woodhead Publishing. ISBN 9780081025857.

13. "1960 - Metal Oxide Semiconductor (MOS) Transistor Demonstrated" (https://www.computerhistory.org/silic onengine/metal-oxide-semiconductor-mos-transistor-demonstrated/). *The Silicon Engine*. Computer History Museum.

14. "1971: Reusable semiconductor ROM introduced"

(https://www.computerhistory.org/storageengine/reusabl e-semiconductor-rom-introduced/). *Computer History Museum*. Retrieved 19 June 2019.

15. Fulford, Adel (24 June 2002). "Unsung hero" (https://www.forbes.com/global/2002/0624/030.html). *Forbes*. Archived (https://web.archive.org/web/20080303205125/http://www.forbes.com/global/2002/0624/030.htm

l) from the original on 3 March 2008. Retrieved 18 March 2008.

16. "How ROM Works" (https://computer.howstuffworks.com/rom5.htm). *HowStuffWorks*. 29 August 2000. Retrieved 10 September 2019.

17. US 4531203 (https://worldwide.espacenet.com/textdoc?DB=EPODOC&IDX=US4531203) Fujio Masuoka

18. Semiconductor memory device and method for manufacturing the same (https://patents.google.com/patent/ US4531203)

19. "NAND Flash Memory: 25 Years of Invention, Development - Data Storage - News & Reviews - eWeek.com" (http://www.eweek.com/c/a/Data-Storage/NAND-Flash-Memory-25-Years-of-Invention-Devel opment-684048/). *eweek.com*.

20. "Toshiba: Inventor of Flash Memory" (http://www.flash25.toshiba.com). *Toshiba*. Retrieved 20 June 2019.

21. Masuoka, F.; Asano, M.; Iwahashi, H.; Komuro, T.; Tanaka, S. (December 1984). "A new flash E2PROM cell using triple polysilicon technology". *1984 International Electron Devices Meeting*: 464–467. doi:10.1109/IEDM.1984.190752 (https://doi.org/10.1109%2FIEDM.1984.190752). S2CID 25967023 (https://api.semanticscholar.org/CorpusID:25967023).

22. Masuoka, F.; Momodomi, M.; Iwata, Y.; Shirota, R. (1987). "New ultra high density EPROM and flash EEPROM with NAND structure cell". *Electron Devices Meeting, 1987 International*. IEDM 1987. IEEE. pp. 552–555. doi:10.1109/IEDM.1987.191485 (https://doi.org/10.1109%2FIEDM.1987.191485).

23. Tal, Arie (February 2002). "NAND vs. NOR flash technology: The designer should weigh the options when using flash memory" (https://web.archive.org/web/20100728210327/http://www2.electronicproducts.com/N AND_vs_NOR_flash_technology-article-FEBMSY1-FEB2002.aspx). Archived from the original (http://www2.electronicproducts.com/NAND_vs_NOR_flash_technology-article-FEBMSY1-FEB2002.aspx) on 28 July 2010. Retrieved 31 July 2010.

24. "H8S/2357 Group, H8S/2357F-ZTATTM, H8S/2398F-ZTATTM Hardware Manual, Section 19.6.1" (http://documentation.renesas.com/doc/products/mpumcu/rej09b0138_h8s2357.pdf)

(PDF). Renesas. October 2004. Retrieved 23 January 2012. "The flash memory can be reprogrammed up to 100 times."

25. "AMD DL160 and DL320 Series Flash: New Densities, New Features" (http://www.spansion.com/Support/Application%20Notes/AMD%20DL160%20and%20DL320%20Series%20Flash-%20New%20Densities,%20New%20Features.pdf) (PDF). AMD. July 2003. Archived (https://web.archive.org/web/20150924104223/http://www.spansion.com/Support/Application%20Notes/AMD%20DL160%20and%20DL320%20Series% 20Flash-%20New%20Densities,%20New%20Features.pdf) (PDF) from the original on 24 September 2015. Retrieved 13 November 2014. "The devices offer single-power-supply operation (2.7 V to 3.6 V), sector architecture, Embedded Algorithms, high performance, and a 1,000,000 program/erase cycle endurance guarantee."

26. James, Dick (2014). "3D ICs in the real world" (https://www.researchgate.net/publication/271453642). *25th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC 2014)*: 113–119. doi:10.1109/ASMC.2014.6846988 (https://doi.org/10.1109%2FASMC.2014.6846988). ISBN 978-1-4799- 3944-2. S2CID 42565898 (https://api.semanticscholar.org/CorpusID:42565898).

27. "NEC: News Release 97/10/28-01" (http://www.nec.co.jp/press/en/9710/2801.html). *www.nec.co.jp*.

28. "Memory" (http://maltiel-consulting.com/Semiconductor_technology_memory.html). *STOL (Semiconductor Technology Online)*. Retrieved 25 June 2019.

29. "Toshiba Makes Major Advances in NAND Flash Memory with 3-bit-per-cell 32nm generation and with 4- bit-per-cell 43nm technology" (http://www.toshiba.co.jp/about/press/2009_02/pr1102.htm). *Toshiba*. 11 February 2009. Retrieved 21 June 2019.

30. "SanDisk ships world's first memory cards with 64 gigabit X4 NAND flash" (https://www.slashgear.com/san disk-ships-worlds-first-memory-cards-with-64-gigabit-x4-nand-flash-1360217/). *SlashGear*. 13 October 2009. Retrieved 20 June 2019.

31. "History" (https://www.samsung.com/us/aboutsamsung/company/history/). *Samsung Electronics*. Samsung. Retrieved 19 June 2019.

32. "StackPath" (https://www.electronicdesign.com/technologies/memory/article/21796009/interview-spansion s-cto-talks-about-embedded-charge-trap-nor-flash-technology). *www.electronicdesign.com*.

33. Ito, T., & Taito, Y. (2017). SONOS Split-Gate eFlash Memory. Embedded Flash Memory for Embedded Systems: Technology, Design for Sub-Systems, and Innovations, 209–244. doi:10.1007/978-3-319-55306- 1_7

34. Bez, R., Camerlenghi, E., Modelli, A., & Visconti, A. (2003). Introduction to flash memory. Proceedings of the IEEE, 91(4), 489–502. doi:10.1109/jproc.2003.811702

35. Lee, J.-S. (2011). Review paper: Nano-floating gate memory devices. Electronic Materials Letters, 7(3), 175–183. doi:10.1007/s13391-011-0901-5

36. Aravindan, Avinash (13 November 2018). "Flash 101: Types of NAND Flash" (https://www.embedded.com/ flash-101-types-of-nand-flash/).
37. Meena, J., Sze, S., Chand, U., & Tseng, T.-Y. (2014). Overview of emerging nonvolatile memory technologies. Nanoscale Research Letters, 9(1), 526. doi:10.1186/1556-276x-9-526
38. "Charge trap technology advantages for 3D NAND flash drives" (https://searchstorage.techtarget.com/tip/C harge-trap-technology-advantages-for-3D-NAND-flash-drives). *SearchStorage*.
39. Grossi, A., Zambelli, C., & Olivo, P. (2016). Reliability of 3D NAND Flash Memories. 3D Flash Memories, 29–62. doi:10.1007/978-94-017-7512-0_2
40. Kodama, N.; Oyama, K.; Shirai, H.; Saitoh, K.; Okazawa, T.; Hokari, Y. (December 1991). "A symmetrical side wall (SSW)-DSA cell for a 64 Mbit flash memory". *International Electron Devices Meeting 1991 [Technical Digest]*: 303–306. doi:10.1109/IEDM.1991.235443 (https://doi.org/10.1109%2FIEDM.1991.2354 43). ISBN 0-7803-0243-5. S2CID 111203629 (https://api.semanticscholar.org/CorpusID:111203629).
41. Eitan, Boaz. "US Patent 5,768,192: Non-volatile semiconductor memory cell utilizing asymmetrical charge trapping" (http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=5,768,192.PN.&OS=PN/5,768,192&RS=PN/5,768, 192). US Patent & Trademark Office. Retrieved 22 May 2012.
42. Fastow, Richard M.; Ahmed, Khaled Z.; Haddad, Sameer; et al. (April 2000). "Bake induced charge gain in NOR flash cells" (https://www.researchgate.net/publication/3253902). *IEEE Electron Device Letters*. 21 (4): 184–186. Bibcode:2000IEDL...21..184F (https://ui.adsabs.harvard.edu/abs/2000IEDL...21..184F). doi:10.1109/55.830976 (https://doi.org/10.1109%2F55.830976). S2CID 24724751 (https://api.semanticscholar.org/CorpusID:24724751).
43. "Samsung produces first 3D NAND, aims to boost densities, drive lower cost per GB" (https://www.extreme tech.com/computing/163221-samsung-produces-first-3d-nand-aims-to-boost-densities-drive-lower-cost-per-gb). *ExtremeTech*. 6 August 2013. Retrieved 4 July 2019.
44. "Toshiba announces new "3D" NAND flash technology" (https://www.engadget.com/2007/06/12/toshiba-an nounces-new-3d-nand-flash-technology/). ***Engadget***. 12 June 2007. Retrieved 10 July 2019.
45. "Samsung Introduces World's First 3D V-NAND Based SSD for Enterprise Applications | Samsung | Samsung Semiconductor Global Website" (https://www.samsung.com/semiconductor/insights/news-events/samsung-introduces-worlds-first-3d-v-nand-based-ssd-for-enterprise-applications/). *Samsung.com*.

46. Clarke, Peter. "Samsung Confirms 24 Layers in 3D NAND" (https://www.eetimes.com/author.asp?section_i d=36&doc_id=1319167). *EETimes*.
47. "TOSHIBA COMMERCIALIZES INDUSTRY'S HIGHEST CAPACITY EMBEDDED NAND FLASH MEMORY FOR MOBILE CONSUMER PRODUCTS" (https://web.archive.org/web/20101123023805/http://www.toshiba.com/taec/news/press_releases/2007/memy_07_470.jsp). *Toshiba*. 17 April 2007. Archived from the original (http://www.toshiba.com/taec/news/press_releases/2007/memy_07_470.jsp) on 23 November 2010. Retrieved 23 November2010.
48. "Hynix Surprises NAND Chip Industry" (http://www.koreatimes.co.kr/www/news/biz/2007/09/123_9628.htm l). *The Korea Times*. 5 September 2007. Retrieved 8 July 2019.
49. "Toshiba Launches the Largest Density Embedded NAND Flash Memory Devices" (https://www.toshiba.c o.jp/about/press/2008_08/pr0701.htm). Toshiba. 7 August 2008. Retrieved 21 June 2019.
50. "Toshiba Launches Industry's Largest Embedded NAND Flash Memory Modules" (https://www.toshiba.co.j p/about/press/2010_06/pr1701.htm). *Toshiba*. 17 June 2010. Retrieved 21 June 2019.
51. SanDisk. "Western Digital Breaks Boundaries with World's Highest-Capacity microSD Card" (https://www. sandisk.com/about/media-center/press-releases/2017/western-digital-breaks-boundaries-with-worlds-high est-capacity-microsd-card). *SanDisk.com*. Archived (https://web.archive.org/web/20170901035345/https:// www.sandisk.com/about/media-center/press-releases/2017/western-digital-breaks-boundaries-with-worlds -highest-capacity-microsd-card) from the original on 1 September 2017. Retrieved 2 September 2017.
52. Bradley, Tony. "Expand Your Mobile Storage With New 400GB microSD Card From SanDisk" (https://www. forbes.com/sites/tonybradley/2017/08/31/expand-your-mobile-storage-with-new-400gb-microsd-card-from- sandisk/#1f2c918d7cc7). *Forbes*. Archived (https://web.archive.org/web/20170901064146/https://www.forbes.com/sites/tonybradley/2017/08/31/expand-your-mobile-storage-with-new-400gb-microsd-card-from-sandisk/#1f2c918d7cc7) from the original on 1 September 2017. Retrieved 2 September 2017.
53. Shilov, Anton (5 December 2017). "Samsung Starts Production of 512 GB UFS NAND Flash Memory: 64- Layer V-NAND, 860 MB/s Reads" (https://www.anandtech.com/show/12120/samsung-starts-production-of- 512-gb-ufs-chips). *AnandTech*. Retrieved 23 June2019.
54. Manners, David (30 January 2019). "Samsung makes 1TB flash eUFS module" (https://www.electronicswe ekly.com/news/business/samsung-makes-1tb-flash-module-2019-01/). *Electronics Weekly*. Retrieved 23 June 2019.

55. Tallis, Billy (17 October 2018). "Samsung Shares SSD Roadmap for QLC NAND And 96-layer 3D NAND" (https://www.anandtech.com/show/13497/samsung-shares-ssd-roadmap-for-qlc-nand-and-96layer-3d-nan d). *AnandTech*. Retrieved 27 June 2019.

56. Basinger, Matt (18 January 2007), *PSoC Designer Device Selection Guide* (https://web.archive.org/web/20091031121330/http://www.psocdeveloper.com/uploads/tx_piapappnote/an2209_03.pdf) (PDF), AN2209, archived from the original (http://www.psocdeveloper.com/uploads/tx_piapappnote/an2209_03.pdf) (PDF) on 31 October 2009, "The PSoC ... utilizes a unique Flash process: SONOS"

57. "2.1.1 Flash Memory" (https://www.iue.tuwien.ac.at/phd/windbacher/node14.html). *www.iue.tuwien.ac.at*.

58. "Floating Gate MOS Memory" (http://www.princeton.edu/~chouweb/newproject/research/SEM/FloatMOSMem.html). *www.princeton.edu*.

59. Shimpi, Anand Lal. "The Intel SSD 710 (200GB) Review" (https://www.anandtech.com/show/4902/intel-ss d-710-200gb-review). *www.anandtech.com*.

60. "Flash Memory Reliability, Life & Wear » Electronics Notes" (https://www.electronics-notes.com/articles/ele ctronic_components/semiconductor-ic-memory/flash-wear-levelling-reliability-lifetime.php#:~:text=Flash%20memory%20wear%20out%20mechanism&text=The%20wear%2Dout%20mechanism%20for,the%20fl a sh%20memory%20wear%20issue.).

61. "Understanding TLC NAND" (https://www.anandtech.com/show/5067/understanding-tlc-nand/2).

62. "Solid State bit density, and the Flash Memory Controller" (https://www.hyperstone.com/en/Solid-State-bit- density-and-the-Flash-Memory-Controller-1235,12728.html). *hyperstone.com*. 17 April 2018. Retrieved 29 May 2018.

63. Yasufuku, Tadashi; Ishida, Koichi; Miyamoto, Shinji; Nakai, Hiroto; Takamiya, Makoto; Sakurai, Takayasu; Takeuchi, Ken (2009), *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design - ISLPED '09* (http://www.computer.org/csdl/proceedings/islped/2009/8684/00/86840087-abs.html), pp. 87–92, doi:10.1145/1594233.1594253 (https://doi.org/10.1145%2F1594233.1594253), ISBN 9781605586847, S2CID 6055676 (https://api.semanticscholar.org/CorpusID:6055676), archived (https://web.archive.org/web/20160305135918/https://www.computer.org/csdl/proceedings/islped/2009/8684/ 00/86840087-abs.html) from the original on 5 March 2016 (abstract) (http://adsabs.harvard.edu/abs/2010IE ITE..93..317Y).

64. Micheloni, Rino; Marelli, Alessia; Eshghi, Kam (2012), *Inside Solid State Drives (SSDs)* (https://books.goo gle.com/books?id=8LS3egzcBG4C&pg=PA188), Bibcode:2013issd.book.....M (https://ui.adsabs.harvard.e du/abs/2013issd.book.....M), ISBN 9789400751460, archived (https://web.archive.org/web/20170209234319/https://books.google.com/books?id=8LS3egzcBG4C&pg=PA188&lpg=PA188) from the original on 9 February 2017

65. Micheloni, Rino; Crippa, Luca (2010), *Inside NAND Flash Memories* (https://books.google.com/books?id=v aq11vKwo_kC&pg=PA530), ISBN 9789048194315, archived (https://web.archive.org/web/20170209164808/https://books.google.com/books?id=vaq11vKwo_kC&pg=PA530&lpg=PA530) from the original on 9 February 2017 In particular, pp 515-536: K. Takeuchi. "Low power 3D-integrated SSD" (https://link.springe r.com/content/pdf/10.1007%2F978-90-481-9431-5_18.pdf)

66. Mozel, Tracey (2009), *CMOSET Fall 2009 Circuits and Memories Track Presentation Slides* (https://books. google.com/books?id=XlbOf-m8fdYC&pg=RA5-PA3), ISBN 9781927500217, archived (https://web.archive.org/web/20170209213305/https://books.google.com/books?id=XlbOf-m8fdYC&pg=RA5-PA3&lpg=RA5- PA3) from the original on 9 February 2017

67. Tadashi Yasufuku et al., "Inductor and TSV Design of 20-V Boost Converter for Low Power 3D Solid State Drive with NAND Flash Memories" (https://www.researchgate.net/publication/220240029_Inductor_and_T SV_Design_of_20-V_Boost_Converter_for_Low_Power_3D_Solid_State_Drive_with_NAND_Flash_Me mories)Archived (https://web.archive.org/web/20160204025034/https://www.researchgate.net/publication/220240029_Inductor_and_TSV_Design_of_20-V_Boost_Converter_for_Low_Power_3D_Solid_State_Drive_with_NAND_Flash_Memories) 4 February 2016 at the Wayback Machine. 2010.

68. Hatanaka, T. and Takeuchi, K. "4-times faster rising VPASS (10V), 15% lower power VPGM (20V), wide output voltage range voltage generator system for 4-times faster 3D-integrated solid-state drives" (https://ieeexplore.ieee.org/document/5986104). 2011.

69. Takeuchi, K., "Low power 3D-integrated Solid-State Drive (SSD) with adaptive voltage generator" (https://i eeexplore.ieee.org/document/5488397). 2010.

70. Ishida, K. et al., "1.8 V Low-Transient-Energy Adaptive Program-Voltage Generator Based on Boost Converter for 3D-Integrated NAND Flash SSD" (https://ieeexplore.ieee.org/document/5759723). 2011.

71. A. H. Johnston, "Space Radiation Effects in Advanced Flash Memories" (http://trs-new.jpl.nasa.gov/dspac e/bitstream/2014/13431/1/01-2369.pdf)Archived (https://web.archive.org/web/20160304220536/http://trs-new.jpl.nasa.gov/dspace/bitstream/2014/13431/1/01-2369.pdf) 4 March 2016 at the Wayback Machine. NASA Electronic Parts and Packaging Program (NEPP). 2001. "... internal transistors used for the charge pump and erase/write control have much thicker oxides because of the requirement for high voltage. This causes flash devices to be considerably more sensitive to total dose damage compared to

other ULSI technologies. It also implies that write and erase functions will be the first parameters to fail from total dose.

... Flash memories will work at much higher radiation levels in the read mode. ... The charge pumps that are required to generate the high voltage for erasing and writing are usually the most sensitive circuit functions, usually failing below 10 krad(SI)."

72. Zitlaw, Cliff. "The Future of NOR Flash Memory" (http://www.eetimes.com/design/memory-design/421563 4). *Memory Designline*. UBM Media. Retrieved 3 May 2011.

73. "NAND Flash Controllers - The key to endurance and reliability" (https://www.hyperstone.com/en/NAND-Fl ash-controllers-The-key-to-endurance-and-reliability-1256,12728.html). *hyperstone.com*. 7 June 2018. Retrieved 1 June 2022.

74. "Samsung moves into mass production of 3D flash memory" (http://www.gizmag.com/samsung-v-nand-flas h-chip-ssd/28655). Gizmag.com. 27 August 2013. Archived (https://web.archive.org/web/20130827091835/http://www.gizmag.com/samsung-v-nand-flash-chip-ssd/28655/) from the original on 27 August 2013. Retrieved 27 August 2013.

75. "Samsung Electronics Starts Mass Production of Industry First 3-bit 3D V-NAND Flash Memory" (https://ne ws.samsung.com/global/samsung-electronics-starts-mass-production-of-industry-first-3-bit-3d-v-nand-flash-memory). *news.samsung.com*.

76. "Samsung V-NAND technology" (https://web.archive.org/web/20160327194431/http://www.samsung.com/us/business/oem-solutions/pdfs/V-NAND_technology_WP.pdf) (PDF). *Samsung Electronics*. September 2014. Archived from the original (http://www.samsung.com/us/business/oem-solutions/pdfs/V-NAND_technology_WP.pdf) (PDF) on 27 March 2016. Retrieved 27 March 2016.

77. Tallis, Billy. "Micron Announces 176-layer 3D NAND" (https://www.anandtech.com/show/16230/micron-an nounces-176layer-3d-nand). *www.anandtech.com*.

78. "Samsung said to be developing industry's first 160-layer NAND flash memory chip" (https://www.techspot. com/news/84905-samsung-developing-industry-first-160-layer-nand-flash.html). *TechSpot*.

79. "Toshiba's Cost Model for 3D NAND" (https://www.linkedin.com/pulse/toshibas-cost-model-3d-nand-frederi ck-chen). *www.linkedin.com*.

80. "Calculating the Maximum Density and Equivalent 2D Design Rule of 3D NAND Flash" (https://www.linke din.com/pulse/calculating-maximum-density-equivalent-2d-design-rule-frederick-chen). *linkedin.com*. Retrieved 1 June 2022.; "Calculating the Maximum Density and Equivalent 2D Design Rule of 3D NAND Flash"

(https://semiwiki.com/lithography/296121-calculating-the-maximum-density-and-equivalent-2d-desi gn-rule-of-3d-nand-flash/). *semwiki.com*. Retrieved 1 June 2022.

81. "AVR105: Power Efficient High Endurance Parameter Storage in Flash Memory" (http://ww1.microchip.co m/downloads/en/AppNotes/doc2546.pdf). p. 3
82. Calabrese, Marcello (May 2013). "Accelerated reliability testing of flash memory: Accuracy and issues on a 45nm NOR technology" (https://ieeexplore.ieee.org/document/6563298). *Proceedings of 2013 International Conference on IC Design & Technology (ICICDT)*: 37–40. doi:10.1109/ICICDT.2013.6563298 (https://doi.org/10.1109%2FICICDT.2013.6563298). ISBN 978-1-4673-4743-3. S2CID 37127243 (https://ap i.semanticscholar.org/CorpusID:37127243). Retrieved 22 June 2022.
83. Jonathan Thatcher, Fusion-io; Tom Coughlin, Coughlin Associates; Jim Handy, Objective-Analysis; Neal Ekker, Texas Memory Systems (April 2009). "NAND Flash Solid State Storage for the Enterprise, An In- depth Look at Reliability" (http://www.snia.org/sites/default/files/SSSI_NAND_Reliability_White_Paper_0.pdf) (PDF). Solid State Storage Initiative (SSSI) of the Storage Network Industry Association (SNIA). Archived (https://web.archive.org/web/20111014033413/http://snia.org/sites/default/files/SSSI_NAND_Reli ability_White_Paper_0.pdf) (PDF) from the original on 14 October 2011. Retrieved 6 December 2011.
84. "Micron Collaborates with Sun Microsystems to Extend Lifespan of Flash-Based Storage, Achieves One Million Write Cycles" (http://investors.micron.com/releasedetail.cfm?ReleaseID=440650) (Press release). Micron Technology, Inc. 17 December 2008. Archived (https://web.archive.org/web/20160304075718/http://investors.micron.com/releasedetail.cfm?ReleaseID=440650) from the original on 4 March 2016.
85. "Taiwan engineers defeat limits of flash memory" (http://phys.org/news/2012-12-taiwan-defeat-limits-memo ry.html). *phys.org*. Archived (https://web.archive.org/web/20160209010327/http://phys.org/news/2012-12-taiwan-defeat-limits-memory.html) from the original on 9 February 2016.
86. "Flash memory made immortal by fiery heat" (https://www.theregister.co.uk/2012/12/03/macronix_thermal_annealing_extends_life_of_flash_memory/). *theregister.co.uk*. Archived (https://web.archive.org/web/20170913183926/https://www.theregister.co.uk/2012/12/03/macronix_thermal_annealing_extends_life_of_flash_ memory/) from the original on 13 September 2017.
87. "Flash memory breakthrough could lead to even more reliable data storage"

(https://web.archive.org/web/2 0121221044513/http://news.yahoo.com/flash-memory-breakthrough-could-lead-even-more-reliable-12404 9340.html). *news.yahoo.com*. Archived from the original (https://news.yahoo.com/flash-memory-breakthrou gh-could-lead-even-more-reliable-124049340.html) on 21 December 2012.

88. "TN-29-17 NAND Flash Design and Use Considerations Introduction" (http://www.micron.com/~/media/do cuments/products/technical-note/nand-flash/tn2917.pdf) (PDF). Micron. April 2010. Archived (https://web.archive.org/web/20151212004340/https://www.micron.com/~/media/documents/products/technical-note/nan d-flash/tn2917.pdf) (PDF) from the original on 12 December 2015. Retrieved 29 July 2011.

89. Kawamatus, Tatsuya. "Technology For Managing NAND Flash" (https://web.archive.org/web/20180515164812/http://read.pudn.com/downloads151/ebook/654250/0808002.pdf) (PDF). Hagiwara sys-com co., LTD. Archived from the original (http://read.pudn.com/downloads151/ebook/654250/0808002.pdf) (PDF) on 15 May 2018. Retrieved 15 May 2018.

90. Cooke, Jim (August 2007). "The Inconvenient Truths of NAND Flash Memory" (http://www.dslreports.com/r0/download/1507743~59e7b9dda2c0e0a0f7ff119a7611c641/flash_mem_summit_jcooke_inconvenient_truths_nand.pdf) (PDF). Flash Memory Summit 2007. Archived (https://web.archive.org/web/20180215023326/http://www.dslreports.com/r0/download/1507743~59e7b9dda2c0e0a0f7ff119a7611c641/flash_mem_summit_jcooke_inconvenient_truths_nand.pdf) (PDF) from the original on 15 February 2018.

91. Richard Blish. "Dose Minimization During X-ray Inspection of Surface-Mounted Flash ICs" (http://www.spansion.com/Support/Application%20Notes/Dose_Minimization_Xray_Inspect_AN.pdf) Archived (https://web.archive.org/web/20160220204227/http://www.spansion.com/Support/Application%20Notes/Dose_Minim ization_Xray_Inspect_AN.pdf) 20 February 2016 at the Wayback Machine. p. 1.

92. Richard Blish. "Impact of X-Ray Inspection on Spansion Flash Memory" (http://www.spansion.com/Suppor t/Application%20Notes/X-ray_inspection_on_flash_AN.pdf) Archived (https://web.archive.org/web/20160304044211/http://www.spansion.com/Support/Application%20Notes/X-ray_inspection_on_flash_AN.pdf) 4 March 2016 at the Wayback Machine

93. "SanDisk Extreme PRO SDHC/SDXC UHS-I Memory Card" (https://www.sandisk.com/home/memory-card s/sd-cards/extremepro-sd-uhs-i). Archived (https://web.archive.org/web/20160127214859/https://www.sandisk.com/home/memory-cards/sd-cards/extremepro-sd-uhs-i) from the original on

27 January 2016. Retrieved 3 February 2016.

94. "Samsung 32GB USB 3.0 Flash Drive FIT MUF-32BB/AM" (http://www.samsung.com/us/computer/memor y-storage-accessories/MUF-32BB/AM). Archived (https://web.archive.org/web/20160203145010/http://www.samsung.com/us/computer/memory-storage-accessories/MUF-32BB/AM) from the original on 3 February 2016. Retrieved 3 February 2016.

95. Spansion. "What Types of ECC Should Be Used on Flash Memory?" (http://www.spansion.com/Support/Application%20Notes/Types_of_ECC_Used_on_Flash_AN.pdf) Archived (https://web.archive.org/web/20160304044226/http://www.spansion.com/Support/Application%20Notes/Types_of_ECC_Used_on_Flash_A N.pdf) 4 March 2016 at the Wayback Machine. 2011.

96. "DSstar: TOSHIBA ANNOUNCES 0.13 MICRON 1GB MONOLITHIC NAND" (https://web.archive.org/web/20121227020955/http://www.tgc.com/dsstar/02/0917/104762.html). Tgc.com. 23 April 2002. Archived from the original (http://www.tgc.com/dsstar/02/0917/104762.html) on 27 December 2012. Retrieved 27 August 2013.

97. Kim, Jesung; Kim, John Min; Noh, Sam H.; Min, Sang Lyul; Cho, Yookun (May 2002). "A Space-Efficient Flash Translation Layer for CompactFlash Systems". *Proceedings of the IEEE*. Vol. 48, no. 2. pp. 366–375. doi:10.1109/TCE.2002.1010143 (https://doi.org/10.1109%2FTCE.2002.1010143).

98. TN-29-07: Small-Block vs. Large-Block NAND flash Devices (http://download.micron.com/pdf/technotes/na nd/tn2907.pdf)Archived (https://web.archive.org/web/20130608040707/http://download.micron.com/pdf/technotes/nand/tn2907.pdf) 8 June 2013 at the Wayback Machine Explains 512+16 and 2048+64-byte blocks

99. AN10860 LPC313x NAND flash data and bad block management (http://www.nxp.com/documents/applica tion_note/AN10860.pdf)Archived (https://web.archive.org/web/20160303231413/http://www.nxp.com/documents/application_note/AN10860.pdf) 3 March 2016 at the Wayback Machine Explains 4096+128-byte blocks.

100. Thatcher, Jonathan (18 August 2009). "NAND Flash Solid State Storage Performance and Capability – an In-depth Look" (https://www.snia.org/sites/default/education/tutorials/2009/spring/solid/JonathanThatcher_NandFlash_SSS_PerformanceV10-nc.pdf) (PDF). SNIA. Archived (https://web.archive.org/web/20120907062956/http://www.snia.org/sites/default/education/tutorials/2009/spring/solid/JonathanThatcher_NandFlash_SSS_PerformanceV10-nc.pdf) (PDF) from the original on 7 September 2012. Retrieved 28 August 2012.

101. "Samsung ECC algorithm"

(http://www.elnec.com/sw/samsung_ecc_algorithm_for_256b.pdf) (PDF). Samsung. June 2008. Archived (https://web.archive.org/web/20081012043739/http://www.elnec.com/sw/samsung_ecc_algorithm_for_256b.pdf) (PDF) from the original on 12 October 2008. Retrieved 15 August 2008.

102. "Open NAND Flash Interface Specification" (https://web.archive.org/web/20110727145313/http://onfi.org/wp-content/uploads/2009/02/onfi_1_0_gold.pdf) (PDF). Open NAND Flash Interface. 28 December 2006. Archived from the original (http://onfi.org/wp-content/uploads/2009/02/onfi_1_0_gold.pdf) (PDF) on 27 July 2011. Retrieved 31 July 2010.

103. A list of ONFi members is available at "Membership - ONFi" (http://onfi.org/membership/). Archived (https://web.archive.org/web/20090829141114/http://onfi.org/membership/) from the original on 29 August 2009. Retrieved 21 September 2009.

104. "Toshiba Introduces Double Data Rate Toggle Mode NAND in MLC And SLC Configurations" (http://www.toshiba.com/taec/news/press_releases/2010/memy_10_599.jsp). *toshiba.com*. Archived (https://web.archive.org/web/20151225111800/http://www.toshiba.com/taec/news/press_releases/2010/memy_10_599.jsp) from the original on 25 December 2015.

105. "Dell, Intel And Microsoft Join Forces To Increase Adoption of NAND-Based Flash Memory in PC Platforms" (http://www.microsoft.com/en-us/news/press/2007/may07/05-30nvmhcipr.aspx). REDMOND, Wash: Microsoft. 30 May 2007. Archived (https://web.archive.org/web/20140812212921/http://www.microsoft.com/en-us/news/press/2007/may07/05-30nvmhcipr.aspx) from the original on 12 August 2014. Retrieved 12 August 2014.

106. Aravindan, Avinash (23 July 2018). "Flash 101: NAND Flash vs NOR Flash" (https://www.embedded.com/flash-101-nand-flash-vs-nor-flash/). *Embedded.com*. Retrieved 23 December 2020.

107. *NAND Flash 101: An Introduction to NAND Flash and How to Design It in to Your Next Product* (https://web.archive.org/web/20160604054353/https://www.micron.com/~/media/Documents/Products/Technical%20 Note/NAND%20Flash/tn2919_nand_101.pdf) (PDF), Micron, pp. 2–3, TN-29-19, archived from the original (http://www.micron.com/~/media/Documents/Products/Technical%20Note/NAND%20Flash/tn2919_nand_ 101.pdf) (PDF) on 4 June 2016

108. Pavan, Paolo; Bez, Roberto; Olivo, Piero; Zanoni, Enrico (1997). "Flash Memory

Cells – An Overview" (htt ps://ieeexplore.ieee.org/document/622505). *Proceedings of the IEEE*. Vol. 85, no. 8 (published August 1997). pp. 1248–1271. doi:10.1109/5.622505 (https://doi.org/10.1109%2F5.622505). Retrieved 15 August 2008.

109. "The Fundamentals of Flash Memory Storage" (http://electronicdesign.com/memory/fundamentals-flash-m emory-storage). 20 March 2012. Archived (https://web.archive.org/web/20170104163357/http://electronicdesign.com/memory/fundamentals-flash-memory-storage) from the original on 4 January 2017. Retrieved 3 January 2017.

110. "SLC NAND Flash Memory | TOSHIBA MEMORY | Europe(EMEA)" (https://web.archive.org/web/2019010 1193808/https://business.toshiba-memory.com/en-emea/product/memory/slc-nand/slc.html). *business.toshiba-memory.com*. Archived from the original (https://business.toshiba-memory.com/en-emea/ product/memory/slc-nand/slc.html) on 1 January 2019. Retrieved 1 January 2019.

111. "Loading site please wait..." (https://www.toshiba.com/tma/technologymoves/slc-nand.jsp) *Toshiba.com*.

112. "Serial Interface NAND | TOSHIBA MEMORY | Europe(EMEA)" (https://web.archive.org/web/2019010114 5411/https://business.toshiba-memory.com/en-emea/product/memory/slc-nand/serial.html). *business.toshiba-memory.com*. Archived from the original (https://business.toshiba-memory.com/en-emea/ product/memory/slc-nand/serial.html) on 1 January 2019. Retrieved 1 January 2019.

113. "BENAND | TOSHIBA MEMORY | Europe(EMEA)" (https://web.archive.org/web/20190101145413/https://b usiness.toshiba-memory.com/en-emea/product/memory/slc-nand/benand.html). *business.toshiba-memory.com*. Archived from the original (https://business.toshiba-memory.com/en-emea/product/memory/sl c-nand/benand.html) on 1 January 2019. Retrieved 1 January 2019.

114. "SLC NAND Flash Memory | TOSHIBA MEMORY | Europe(EMEA)" (https://web.archive.org/web/2019010 1145415/https://business.toshiba-memory.com/en-emea/product/memory/slc-nand.html). *business.toshiba-memory.com*. Archived from the original (https://business.toshiba-memory.com/en-emea/product/memory/sl c-nand.html) on 1 January 2019. Retrieved 1 January 2019.

115. Salter, Jim (28 September 2019). "SSDs are on track to get bigger and cheaper

thanks to PLC technology" (https://arstechnica.com/gadgets/2019/09/new-intel-toshiba-ssd-technologies-squeeze-more-bits-into-each-cell/). *Ars Technica*.

116. "PBlaze4_Memblaze" (http://memblaze.com/en/index.php?c=article&a=type&tid=54). *memblaze.com*. Retrieved 28 March 2019.

117. Crothers, Brooke. "SanDisk to begin making 'X4' flash chips" (https://www.cnet.com/news/sandisk-to-begin-making-x4-flash-chips/). *CNET*.

118. Crothers, Brooke. "SanDisk ships 'X4' flash chips" (https://www.cnet.com/news/sandisk-ships-x4-flash-chip s/). *CNET*.

119. "SanDisk Ships Flash Memory Cards With 64 Gigabit X4 NAND Technology" (https://phys.org/news/2009- 10-sandisk-ships-memory-cards-gigabit.html). *phys.org*.

120. "SanDisk Begins Mass Production of X4 Flash Memory Chips" (https://www.photoreview.com.au/news/san disk-begins-mass-production-of-x4-flash-memory-chips/). 17 February 2012.

121. Tallis, Billy. "The Samsung 983 ZET (Z-NAND) SSD Review: How Fast Can Flash Memory Get?" (https:// www.anandtech.com/show/13951/the-samsung-983-zet-znand-ssd-review). *AnandTech.com*.

122. Vättö, Kristian. "Testing Samsung 850 Pro Endurance & Measuring V-NAND Die Size" (http://www.anandt ech.com/show/8239/update-on-samsung-850-pro-endurance-vnand-die-size). *AnandTech*. Archived (https://web.archive.org/web/20170626155736/http://www.anandtech.com/show/8239/update-on-samsung-850-pro-endurance-vnand-die-size) from the original on 26 June 2017. Retrieved 11 June 2017.

123. Vättö, Kristian. "Samsung SSD 845DC EVO/PRO Performance Preview & Exploring IOPS Consistency" (http://www.anandtech.com/show/8319/samsung-ssd-845dc-evopro-preview-exploring-worstcase-iops/3). *AnandTech*. p. 3. Archived (https://web.archive.org/web/20161022231209/http://www.anandtech.com/show/8319/samsung-ssd-845dc-evopro-preview-exploring-worstcase-iops/3) from the original on 22 October 2016. Retrieved 11 June 2017.

124. Vättö, Kristian. "Samsung SSD 850 EVO (120GB, 250GB, 500GB & 1TB) Review" (http://www.anandtech. com/show/8747/samsung-ssd-850-evo-review/4). *AnandTech*. p. 4. Archived (https://web.archive.org/web/20170531043312/http://www.anandtech.com/show/8747/samsung-ssd-850-evo-review/4) from the original on 31 May 2017. Retrieved 11 June 2017.

125. Vättö, Kristian. "Samsung SSD 845DC EVO/PRO Performance Preview & Exploring IOPS Consistency" (http://www.anandtech.com/show/8319/samsung-ssd-845dc-evopro-preview-exploring-worstcase-iops/3). *AnandTech*. p. 2. Archived (https://web.archive.org/web/20161022231209/http://www.anandtech.com/show/8319/samsung-ssd-845dc-evopro-preview-exploring-worstcase-iops/3) from the original on 22 October 2016. Retrieved 11 June 2017.

126. Ramseyer, Chris (9 June 2017). "Flash Industry Trends Could Lead Users Back to Spinning Disks" (http:// www.tomshardware.com/news/consumer-optane-enterprise-ssd-market,34631.html). *AnandTech*. Retrieved 11 June 2017.

127. "PBlaze5 700" (http://memblaze.com/en/index.php?c=article&a=type&tid=100). *memblaze.com*. Retrieved 28 March 2019.

128. "PBlaze5 900" (http://memblaze.com/en/index.php?c=article&a=type&tid=101). *memblaze.com*. Retrieved 28 March 2019.

129. "PBlaze5 910/916 series NVMe SSD" (http://memblaze.com/en/index.php?c=article&a=type&tid=102). *memblaze.com*. Retrieved 26 March 2019.

130. "PBlaze5 510/516 series NVMe™ SSD" (http://memblaze.com/en/index.php?c=article&a=type&tid=103). *memblaze.com*. Retrieved 26 March 2019.

131. "QLC NAND - What can we expect from the technology?" (https://www.architecting.it/blog/qlc-nand/). 7 November 2018.

132. "Say Hello: Meet the World's First QLC SSD, the Micron 5210 ION" (https://www.micron.com/about/blog/2018/november/meet%20the%20worlds%20first%20qlc%20ssd%20the%20micron%205210%20ion). *Micron.com*.

133. "QLC NAND" (https://www.micron.com/products/advanced%20solutions/qlc%20nand). *Micron.com*.

134. Tallis, Billy. "The Intel SSD 660p SSD Review: QLC NAND Arrives For Consumer SSDs" (https://www.ana ndtech.com/show/13078/the-intel-ssd-660p-ssd-review-qlc-nand-arrives). *AnandTech.com*.

135. "SSD endurance myths and legends articles on StorageSearch.com" (http://www.storagesearch.com/ssdm yths-endurance.html). *StorageSearch.com*.

136. "Samsung Announces QLC SSDs And Second-Gen Z-NAND" (https://www.tomshardware.com/news/sam sung-qlc-z-nand-ssd-flash,37945.html). *Tom's Hardware*. 18 October 2018.

137. "Samsung 860 QVO review: the first QLC SATA SSD, but it can't topple TLC yet" (https://www.pcgamesn.c om/samsung-860-qvo-review-benchmarks-qlc-ssd). *PCGamesN*.

138. "Samsung Electronics Starts Mass Production of Industry's First 4-bit Consumer SSD" (https://news.samsu ng.com/global/samsung-electronics-starts-mass-production-of-industrys-first-4-bit-consumer-ssd). *news.samsung.com*.
139. Nellis, Hyunjoo Jin, Stephen (20 October 2020). "South Korea's SK Hynix to buy Intel's NAND business for $9 billion" (https://www.reuters.com/article/us-intel-divestiture-sk-hynix-idUSKBN2742IY). *Reuters* – via www.reuters.com.
140. "NAND Evolution and its Effects on Solid State Drive Useable Life" (https://web.archive.org/web/20111112000643/http://www.wdc.com/WDProducts/SSD/whitepapers/en/NAND_Evolution_0812.pdf) (PDF). Western Digital. 2009. Archived from the original (http://www.wdc.com/WDProducts/SSD/whitepapers/en/NAND_Evolution_0812.pdf) (PDF) on 12 November 2011. Retrieved 22 April 2012.
141. "Understanding TLC NAND" (https://www.anandtech.com/show/5067/understanding-tlc-nand/2).
142. "Flash vs DRAM follow-up: chip stacking" (https://web.archive.org/web/20121124042741/http://www.dailycircuitry.com/2012/04/as-follow-up-to-our-flash-vs-dram.html). The Daily Circuit. 22 April 2012. Archived from the original (http://www.dailycircuitry.com/2012/04/as-follow-up-to-our-flash-vs-dram.html) on 24 November 2012. Retrieved 22April 2012.
143. "Computer data storage unit conversion - non-SI quantity" (http://www.convertunits.com/type/computer+dat a+storage). Archived (https://web.archive.org/web/20150508070909/http://www.convertunits.com/type/com puter+data+storage) from the original on 8 May 2015. Retrieved 20 May 2015.
144. Shilov, Anton (12 September 2005). "Samsung Unveils 2GB Flash Memory Chip" (https://web.archive.org/web/20081224220204/http://www.xbitlabs.com/news/memory/display/20050912212649.html). X-bit labs. Archived from the original (http://www.xbitlabs.com/news/memory/display/20050912212649.html) on 24 December 2008. Retrieved 30 November2008.
145. Gruener, Wolfgang (11 September 2006). "Samsung announces 40 nm Flash, predicts 20 nm devices" (https://web.archive.org/web/20080323070752/http://www.tgdaily.com/content/view/28504/135/). TG Daily. Archived from the original (http://www.tgdaily.com/content/view/28504/135/) on 23 March 2008. Retrieved 30 November 2008.
146. "SanDisk Media Center" (http://www.sandisk.com/Corporate/PressRoom/PressReleases/PressRelease.aspx?ID=4079). *sandisk.com*. Archived

(https://web.archive.org/web/20081219084116/http://www.sandisk.com/Corporate/PressRoom/PressReleases/PressRelease.aspx?ID=4079) from the original on 19 December 2008.

147. "SanDisk Media Center" (http://www.sandisk.com/Corporate/PressRoom/PressReleases/PressRelease.aspx?ID=4091). *sandisk.com*. Archived (https://web.archive.org/web/20081219084247/http://www.sandisk.com/Corporate/PressRoom/PressReleases/PressRelease.aspx?ID=4091) from the original on 19 December 2008.

148. https://www.pcworld.com/article/225370/look_out_for_the_256gb_thumb_drive_and_the_128gb_tablet.html; "Kingston outs the first 256GB flash drive" (https://techcrunch.com/2009/07/20/kingston-outs-the-first-256gb-flash-drive/). Archived (https://web.archive.org/web/20170708012814/https://techcrunch.com/2009/07/20/ kingston-outs-the-first-256gb-flash-drive/) from the original on 8 July 2017. Retrieved 28 August 2017. 20 July 2009, Kingston DataTraveler 300 is 256 GB.

149. Borghino, Dario (31 March 2015). "3D flash technology moves forward with 10 TB SSDs and the first 48- layer memory cells" (http://www.gizmag.com/high-capacity-3d-flash-memory/36782). *Gizmag*. Archived (https://web.archive.org/web/20150518115212/http://www.gizmag.com/high-capacity-3d-flash-memory/3678 2/) from the original on 18 May 2015. Retrieved 31 March 2015.

150. "Samsung Launches Monster 4TB 850 EVO SSD Priced at $1,499 | Custom PC Review" (https://www.cust ompcreview.com/news/samsung-launches-4tb-850-evo-ssd-priced-1499/30838/). *Custom PC Review*. 13 July 2016. Archived (https://web.archive.org/web/20161009172049/https://www.custompcreview.com/new s/samsung-launches-4tb-850-evo-ssd-priced-1499/30838/) from the original on 9 October 2016. Retrieved 8 October 2016.

151. "Samsung Unveils 32TB SSD Leveraging 4th Gen 64-Layer 3D V-NAND | Custom PC Review" (https://w ww.custompcreview.com/news/samsung-unveils-32tb-ssd-leveraging-4th-gen-64-layer-3d-v-nand/31651/). *Custom PC Review*. 11 August 2016. Archived (https://web.archive.org/web/20161009170533/https://www.custompcreview.com/news/samsung-unveils-32tb-ssd-leveraging-4th-gen-64-layer-3d-v-nand/31651/) from the original on 9 October 2016. Retrieved 8 October 2016.

152. Master, Neal; Andrews, Mathew; Hick, Jason; Canon, Shane; Wright, Nicholas (2010). "Performance analysis of commodity and enterprise class flash devices" (http://www.pdsw.org/pdsw10/resources/papers/ master.pdf) (PDF). *IEEE Petascale*

Data Storage Workshop. Archived (https://web.archive.org/web/20160506160509/http://www.pdsw.org/pdsw10/resources/papers/master.pdf) (PDF) from the original on 6 May 2016.

153. "DailyTech - Samsung Confirms 32nm Flash Problems, Working on New SSD Controller" (https://web.archive.org/web/20160304003356/http://www.dailytech.com/article.aspx?newsid=16407). *dailytech.com*. Archived from the original (http://www.dailytech.com/article.aspx?newsid=16407) on 4 March 2016. Retrieved 3 October 2009.

154. Clive Maxfield. "Bebop to the Boolean Boogie: An Unconventional Guide to Electronics" (https://books.goo gle.com/books?id=u0xyEuXF3l4C). p. 232.

155. Many serial flash devices implement a *bulk read* mode and incorporate an internal address counter, so that it is trivial to configure them to transfer their entire contents to RAM on power-up. When clocked at 50 MHz, for example, a serial flash could transfer a 64 Mbit firmware image in less than two seconds.

156. Lyth0s (17 March 2011). "SSD vs. HDD" (https://web.archive.org/web/20110820095531/http://elitepcbuildi ng.com/ssd-vs-hdd). elitepcbuilding.com. Archived from the original (http://elitepcbuilding.com/ssd-vs-hdd) on 20 August 2011. Retrieved 11 July 2011.

157. "Flash Solid State Disks – Inferior Technology or Closet Superstar?" (http://www.storagesearch.com/bitmic ro-art1.html). STORAGEsearch. Archived (https://web.archive.org/web/20081224215032/http://www.storagesearch.com/bitmicro-art1.html) from the original on 24 December 2008. Retrieved 30 November 2008.

158. Vadim Tkachenko (12 September 2012). "Intel SSD 910 vs HDD RAID in tpcc-mysql benchmark" (http://w ww.mysqlperformanceblog.com/2012/09/11/intel-ssd-910-vs-hdd-raid-in-tpcc-mysql-benchmark/). *MySQL Performance Blog*.

159. Matsunobu, Yoshinori. "SSD Deployment Strategies for MySQL." (https://www.slideshare.net/matsunobu/s sd-deployment-strategies-for-mysql) Archived (https://web.archive.org/web/20160303224013/http://www.slideshare.net/matsunobu/ssd-deployment-strategies-for-mysql) 3 March 2016 at the Wayback Machine *Sun Microsystems*, 15 April 2010.

160. "Samsung Electronics Launches the World's First PCs with NAND Flash-based Solid State Disk" (http://w ww.samsung.com/he/presscenter/pressrelease/pressrelease_20060524_0000257996.asp). *Press Release*. Samsung. 24 May 2006. Archived (https://web.archive.org/web/20081220094813/http://www.samsung.com/he/presscenter/pressrelease/pressrelease_20060524_0000257996.asp) from the original on 20 December 2008. Retrieved 30 November 2008.

161. "Samsung's SSD Notebook" (https://web.archive.org/web/20181015192607/https://news.softpedia.com/news/Samsung-s-SSD-Notebook-33475.shtml). 22 August 2006. Archived from the original (https://news.soft pedia.com/news/Samsung-s-SSD-Notebook-33475.shtml) on 15 October 2018. Retrieved 15 October 2018.

162. "文庫本サイズの VAIO「type U」 フラッシュメモリー搭載モデル発売" (https://www.sony.jp/CorporateCrui se/Press/200606/06-0627/). *Sony.jp* (in Japanese).

163. "Sony Vaio UX UMPC – now with 32 GB Flash memory | NBnews.info. Laptop and notebook news, reviews, test, specs, price | Каталог ноутбуков, ультрабуков и планшетов, новости, обзоры" (http://nbne ws.info/en/news/397).

164. Douglas Perry (2012) (http://www.tomshardware.com/news/fusio-io-flash-ssdalloc-memory-ram,16352.htm

l) Princeton: Replacing RAM with Flash Can Save Massive Power.

165. "Understanding Life Expectancy of Flash Storage" (https://www.ni.com/en-us/support/documentation/suppl emental/12/understanding-life-expectancy-of-flash-storage.html). *www.ni.com*. 23 July 2020. Retrieved 19 December 2020.

166. "8-Bit AVR Microcontroller ATmega32A Datasheet Complete" (https://web.archive.org/web/20160409120244/http://www.atmel.com/Images/Atmel-8155-8-bit-Microcontroller-AVR-ATmega32A_Datasheet.pdf) (PDF). 19 February 2016. p. 18. Archived from the original (https://www.atmel.com/images/atmel-8155-8-bit-microcontroller-avr-atmega32a_datasheet.pdf) (PDF) on 9 April 2016. Retrieved 29 May 2016.
"Reliability Qualification results show that the projected data retention failure rate is much less than 1 PPM over 20 years at 85 °C or 100 years at 25 °C"

167. "On Hacking MicroSD Cards « bunnie's blog" (https://www.bunniestudios.com/blog/?p=3554).

168. "Data Retention in MLC NAND Flash Memory: Characterization, Optimization, and Recovery" (https://user s.ece.cmu.edu/~omutlu/pub/flash-memory-data-retention_hpca15.pdf) (PDF). 27 January 2015. p. 10. Archived (https://web.archive.org/web/20161007000927/https://users.ece.cmu.edu/~omutlu/pub/flash-mem ory-data-retention_hpca15.pdf) (PDF) from the original on 7 October 2016. Retrieved 27 April 2016.

169. "JEDEC SSD Specifications Explained" (https://www.jedec.org/sites/default/files/Alvin_Cox%20%5BCompatibility%20Mode%5D_0.pdf) (PDF). p. 27.

170. Yinug, Christopher Falan (July 2007). "The Rise of the Flash Memory Market: Its

Impact on Firm Behavior and Global Semiconductor Trade Patterns" (https://web.archive.org/web/20080529180622/http://www.usitc.gov/journal/Final_falan_article1.pdf) (PDF). *Journal of International Commerce and Economics*. Archived from the original (http://www.usitc.gov/journal/Final_falan_article1.pdf) (PDF) on 29 May 2008. Retrieved 19 April 2008.

171. NAND memory market rockets (http://www.tgdaily.com/hardware-brief/71015-nand-memory-market-rockets) Archived (https://web.archive.org/web/20160208114459/http://www.tgdaily.com/hardware-brief/71015-nand-memory-market-rockets) 8 February 2016 at the Wayback Machine, 17 April 2013, Nermin Hajdarbegovic, *TG Daily*, retrieved at 18 April 2013

172. "Power outage may have ruined 15 exabytes of WD and Toshiba flash storage" (https://appleinsider.com/articles/19/07/01/power-outage-may-have-ruined-15-exabytes-of-wd-and-toshiba-memory). *AppleInsider*.

173. "NAND Flash manufacturers' market share 2019" (https://www.statista.com/statistics/275886/market-share-held-by-leading-nand-flash-memory-manufacturers-worldwide/). *Statista*. Retrieved 3 July 2019.

174. "SK Hynix completes first phase of $9 bln Intel NAND business buy" (https://www.reuters.com/technology/sk-hynix-completes-first-phase-9-bln-intel-nand-business-buy-2021-12-29/). *Reuters*. 29 December 2021. Retrieved 27 June 2022.

175. "NAND Revenue by Manufacturers Worldwide (2014-2022)" (https://businessquant.com/nand-revenue-by-manufacturer-worldwide#:~:text=NAND%20manufacturers%20collectively%20generated%20%2417.91,third%20and%20fourth%20positions%2C%20respectively.). 26 May 2020. Retrieved 27 June 2022.

176. "The Flash Memory Market" (http://smithsonianchips.si.edu/ice/cd/MEMORY97/SEC05.PDF#page=4) (PDF). *Integrated Circuit Engineering Corporation*. Smithsonian Institution. 1997. p. 4. Retrieved 16 October 2019.

177. Cappelletti, Paulo; Golla, Carla; Olivo, Piero; Zanoni, Enrico (2013). *Flash Memories* (https://books.google.com/books?id=cHzrBwAAQBAJ&pg=PA32). Springer Science & Business Media. p. 32. ISBN 9781461550150.

178. "Not Flashing Quite As Fast" (https://books.google.com/books?id=e6mzAAAAIAAJ). *Electronic Business*. Cahners Publishing Company. 26 (7–13): 504. 2000. "Unit shipments increased 64% in 1999 from the prior year, and are forecast to increase 44% to 1.8 billion units in 2000."

179. Sze, Simon Min. "EVOLUTION OF NONVOLATILE SEMICONDUCTOR MEMORY: From Invention to Nanocrystal Memory" (https://indico.cern.ch/event/422861/attachments/891704/1255315/Sze_26APR05.pdf#page=41) (PDF). *CERN*. National Yang Ming Chiao Tung University. p. 41. Retrieved 22 October 2019.

180. Handy, Jim (26 May 2014). "How Many Transistors Have Ever Shipped?" (https://www.forbes.com/sites/jim handy/2014/05/26/how-many-transistors-have-ever-shipped/). *Forbes*. Retrieved 21 October 2019.

181. "【Market View】Major events in the 2008 DRAM industry; End application demand remains weak, 2009 NAND Flash demand bit growth being revised down to 81%" (https://www.dramexchange.com/WeeklyRes earch/Post/2/1911.html). *DRAMeXchange*. 30 December 2008. Retrieved 16 October 2019.

182. "NOR Flash Memory Finds Growth Opportunities in Tablets and E-Book Readers" (https://technology.ihs.c om/389310/nor-flash-memory-finds-growth-opportunities-in-tablets-and-e-book-readers). *IHS Technology*. IHS Markit. 9 June 2011. Retrieved 16 October 2019.

183. "Samsung to unveil new mass-storage memory cards" (http://www.koreatimes.co.kr/www/tech/2019/06/693_118515.html). *The Korea Times*. 29 August 2012. Retrieved 16 October 2019.

184. "Winbond Top Serial Flash Memory Supplier Worldwide, Ships 1.7 Billion Units in 2012, Ramps 58nm Production" (https://www.businesswire.com/news/home/20130410005060/en/Winbond-Top-Serial-Flash- Memory-Supplier-Worldwide). *Business Wire*. Winbond. 10 April 2013. Retrieved 16 October 2019.

185. Shilov, Anton (1 October 2015). "Samsung: NAND flash industry will triple output to 253EB by 2020" (http s://www.kitguru.net/components/hard-drives/anton-shilov/samsung-nand-flash-industry-will-triple-output-to- 253eb-by-2020/). *KitGuru*. Retrieved 16 October 2019.

186. "Flash memory prices rebound as makers introduce larger-capacity chips" (https://asia.nikkei.com/Busines s/Flash-memory-prices-rebound-as-makers-introduce-larger-capacity-chips). *Nikkei Asian Review*. Nikkei, Inc. 21 July 2016. Retrieved 16 October 2019.

187. Tidwell, William (30 August 2016). "Data 9, Storage 1 - NAND Production Falls Behind in the Age of Hyperscale" (https://seekingalpha.com/article/4002948-data-9-storage-1-nand-production-falls-behind-age-hyperscale). *Seeking Alpha*. Retrieved 17 October 2019.

188. Coughlin, Thomas M. (2017). *Digital Storage in Consumer Electronics: The Essential Guide* (https://books.google.com/books?id=K2dCDwAAQBAJ&pg=PA217). Springer. p. 217. ISBN 9783319699073.

189. Reinsel, David; Gantz, John; Rydning, John (November 2018). "IDC White Paper: The Digitization of the World" (https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper. pdf#page=14) (PDF). *Seagate Technology*. International Data Corporation. p. 14. Retrieved 17 October 2019.

190. Mellor, Chris (28 February 2018). "Who was the storage dollar daddy in 2017? S. S. D" (https://www.thereg ister.co.uk/2018/02/28/crossover_ssd_dollars_exceeded_disk_drive_dollars_in_20

17/). *The Register*. Retrieved 17 October 2019.

191. "Combined SSD, HDD Storage Shipped Jumps 21% to 912 Exabytes in 2018" (https://www.businesswire. com/news/home/20190307005812/en/TRENDFOCUS-Combined-SSD-HDD-Storage-Shipped-Jumps). *Business Wire*. TRENDFOCUS. 7 March 2019. Retrieved 17 October 2019.

192. Yiu, Joseph (February 2015). "Embedded Processors" (https://community.arm.com/cfs-file/key/telligent-e volution-components-attachments/01-2142-00-00-00-00-70-29/Embedded-SoC-Design-for-High-Reliabilit y-Systems-1.02.pdf) (PDF). *ARM*. Embedded World 2015. Retrieved 23 October 2019.

193. Smith, Ryan (8 October 2019). "Arm TechCon 2019 Keynote Live Blog (Starts at 10am PT/17:00 UTC)" (htt ps://www.anandtech.com/show/14959/arm-techcon-2019-keynote-live-blog). *AnandTech*. Retrieved 15 October 2019.

194. "2011 Annual Report" (https://web.archive.org/web/20191016115727/http://investors.cypress.com/static-fil es/62237288-5a22-4903-9ef8-3719d37ea699). *Cypress Semiconductor*. 2012. Archived from the original (http://investors.cypress.com/static-files/62237288-5a22-4903-9ef8-3719d37ea699) on 16 October 2019. Retrieved 16 October 2019.

195. "Technology Roadmap for NAND Flash Memory" (https://web.archive.org/web/20150109095122/http://www.techinsights.com/uploadedFiles/Public_Website/Content_-_Primary/Marketing/2013/Nand_Flash_Road map/NAND-Flash-Roadmap.ppt). techinsights. April 2013. Archived from the original (http://www.techinsights.com/uploadedFiles/Public_Website/Content_-_Primary/Marketing/2013/Nand_Flash_Roadmap/NAND-Flash-Roadmap.ppt) on 9 January 2015. Retrieved 9 January 2015.

196. "Technology Roadmap for NAND Flash Memory" (https://web.archive.org/web/20150109095119/http://www.techinsights.com/uploadedFiles/NAND-Flash-Roadmap-2014.ppt). techinsights. April 2014. Archived from the original (http://www.techinsights.com/uploadedFiles/NAND-Flash-Roadmap-2014.ppt) on 9 January 2015. Retrieved 9 January 2015.

197. "NAND Flash Memory Roadmap" (https://web.archive.org/web/20180625075602/http://www.techinsights.com/techservices/TechInsights-NAND-Flash-Roadmap-2016.pdf) (PDF). *TechInsights*. June 2016. Archived from the original (http://www.techinsights.com/techservices/TechInsights-NAND-Flash-Roadmap-2016.pdf) (PDF) on 25 June 2018. Retrieved 25 June 2018.

198. "Samsung Mass Producing 128Gb 3-bit MLC NAND Flash"

(https://web.archive.org/web/20190621175628/https://www.tomshardware.co.uk/NAND-128Gb-Mass-Production-3-bit-MLC,news-43458.html). *Tom's Hardware*. 11 April 2013. Archived from the original (https://www.tomshardware.co.uk/NAND-128Gb-Mass- Production-3-bit-MLC,news-43458.html) on 21 June 2019. Retrieved 21 June 2019.

199. "Toshiba : News Release (31 Aug, 2010): Toshiba launches 24nm process NAND flash memory" (http://www.toshiba.co.jp/about/press/2010_08/pr3101.htm?from=RSS_PRESS&uid=20100831-1112e). *Toshiba.co.jp*.

200. Lal Shimpi, Anand (2 December 2010). "Micron's ClearNAND: 25nm + ECC, Combats Increasing Error Rates" (http://www.anandtech.com/show/4043/micron-announces-clearnand-25nm-with-ecc). Anandtech. Archived (https://web.archive.org/web/20101203082325/http://www.anandtech.com/show/4043/micron-ann ounces-clearnand-25nm-with-ecc) from the original on 3 December 2010. Retrieved 2 December 2010.

201. Kim, Kinam; Koh, Gwan-Hyeob (16 May 2004). *2004 24th International Conference on Microelectronics (IEEE Cat. No.04TH8716)*. Vol. 1. Serbia and Montenegro: Proceedings of the 24th International Conference on Microelectronics. pp. 377–384. doi:10.1109/ICMEL.2004.1314646 (https://doi.org/10.110 9%2FICMEL.2004.1314646). ISBN 978-0-7803-8166-7. S2CID 40985239 (https://api.semanticscholar.org/ CorpusID:40985239).

202. "A chronological list of Intel products. The products are sorted by date" (https://web.archive.org/web/20070809053720/http://download.intel.com/museum/research/arc_collect/timeline/TimelineDateSort7_05.pdf) (PDF). *Intel museum*. Intel Corporation. July 2005. Archived from the original (http://download.intel.com/museum/research/arc_collect/timeline/TimelineDateSort7_05.pdf) (PDF) on 9 August 2007. Retrieved 31 July 2007.

203. "DD28F032SA Datasheet" (http://www.datasheetcatalog.com/datasheets_pdf/D/D/2/8/DD28F032SA.shtm l). Intel. Retrieved 27 June 2019.

204. "Japanese Company Profiles" (http://smithsonianchips.si.edu/ice/cd/PROF96/JAPAN.PDF) (PDF). Smithsonian Institution. 1996.Retrieved 27 June 2019.

205. "Toshiba to Introduce Flash Memory Cards" (http://www.toshiba.co.jp/about/press/1995_03/pr0201.htm). Toshiba. 2 March 1995. Retrieved 20 June 2019.

206. "WORLDWIDE IC MANUFACTURERS" (http://smithsonianchips.si.edu/ice/cd/STATUS98/SEC02.PDF) (PDF). Smithsonian Institution. 1997. Retrieved 10 July 2019.

207. "TOSHIBA ANNOUNCES 0.13 MICRON 1Gb MONOLITHIC NAND FEATURING

LARGE BLOCK SIZE FOR IMPROVED WRITE/ERASE SPEED PERFORMANCE" (https://web.archive.org/web/20060311224004/http://www.toshiba.com/taec/news/press_releases/2002/to-230.jsp). Toshiba. 9 September 2002. Archived from the original (http://www.toshiba.com/taec/news/press_releases/2002/to-230.jsp) on 11 March 2006. Retrieved 11 March 2006.

208. "TOSHIBA AND SANDISK INTRODUCE A ONE GIGABIT NAND FLASH MEMORY CHIP, DOUBLING CAPACITY OF FUTURE FLASH PRODUCTS" (http://www.toshiba.co.jp/about/press/2001_11/pr1202.ht m). Toshiba. 12 November 2001. Retrieved 20 June 2019.

209. "Our Proud Heritage from 2000 to 2009" (https://www.samsung.com/semiconductor/about-us/history-03/). *Samsung Semiconductor*. Samsung. Retrieved 25 June 2019.

210. "TOSHIBA ANNOUNCES 1 GIGABYTE COMPACTFLASH™CARD" (https://web.archive.org/web/20060311212118/http://www.toshiba.com/taec/news/press_releases/2002/to-231.jsp). Toshiba. 9 September 2002. Archived from the original (http://www.toshiba.com/taec/news/press_releases/2002/to-231.jsp) on 11 March 2006. Retrieved 11 March 2006.

211. "History: 2010s" (https://web.archive.org/web/20210517040328/https://www.skhynix.com/eng/about/history 2010.jsp). *SK Hynix*. Archived from the original (https://www.skhynix.com/eng/about/history2010.jsp) on 17 May 2021. Retrieved 8 July 2019.

212. "Samsung e·MMC Product family" (http://www.mt-system.ru/sites/default/files/klmxgxge4a-x001mmc4_41_2ynm_based_emmc1_1.pdf) (PDF). Samsung Electronics. December 2011. Retrieved 15 July 2019.

213. "Toshiba Develops World's First 4-bit Per Cell QLC NAND Flash Memory" (https://www.techpowerup.com/ 234729/toshiba-develops-worlds-first-4-bit-per-cell-qlc-nand-flash-memory). *TechPowerUp*. 28 June 2017. Retrieved 20 June 2019.

214. Shilov, Anton (6 August 2018). "Samsung Starts Mass Production of QLC V-NAND-Based SSDs" (https://w ww.anandtech.com/show/13170/samsung-starts-mass-production-of-qlc-vnandbased-ssds). *AnandTech*. Retrieved 23 June 2019.

215. "Toshiba's flash chips could boost SSD capacity by 500 percent" (https://www.engadget.com/2018/07/20/to shiba-flash-166-gb-per-chip/). *Engadget*. 20 July 2018. Retrieved 23 June 2019.

216. McGrath, Dylan (20 February 2019). "Toshiba Claims Highest-Capacity NAND" (https://www.eetimes.com/ document.asp?doc_id=1334344). *EE Times*. Retrieved

23 June 2019.

217. Shilov, Anton (26 June 2019). "SK Hynix Starts Production of 128-Layer 4D NAND, 176-Layer Being Developed" (https://www.anandtech.com/show/14589/sk-hynix-128-layer-4d-nand). *AnandTech*. Retrieved 8 July 2019.

218. Mu-Hyun, Cho. "Samsung produces 1TB eUFS memory for smartphones" (https://www.zdnet.com/article/s amsung-produces-1tb-eufs-memory-for-smartphones/). *ZDNet*.

219. "Samsung Breaks Terabyte Threshold for Smartphone Storage with Industry's First 1TB Embedded Universal Flash Storage" (https://news.samsung.com/global/samsung-breaks-terabyte-threshold-for-smartphone-storage-with-industrys-first-1tb-embedded-universal-flash-storage). Samsung. 30 January 2019. Retrieved 13 July 2019.

External links

- Semiconductor Characterization System has diverse functions
- Understanding and selecting higher performance NAND architectures Archived 31 October 2012 at the Wayback Machine
- How flash storage works, presentation by David Woodhouse from Intel
- Flash endurance testing
- NAND Flash Data Recovery Cookbook
- Type of Flash Memory by OpenWrt

Text is available under the Creative Commons Attribution-ShareAlike License 3.0; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

10.2 PCIe/NVMe 初始化过程分析

每年随着本科或者研究生毕业季的到来，国内各大 SSD 厂商以及芯片设计行业知名公司纷纷进入各个理工科高校进行抢人大战，新一批的工程师将很快步入工作岗位成为研发、测试方面的工程师。但是，这些来自不同专业的学生可能有不少人在学校期间并没有接触过 PCIe 和 NVMe 协议，或者对于这些协议仅有一些模糊的认识，急需进行相关协议的一些基础培训。

Saniffer 公司结合在 SSD 测试尤其是 PCIe/NVMe 协议分析领域的经验，通过剖析一个 PCIe Gen 4 NVMe SSD 在主机上加电启动过程的 trace 文件给大家就 PCIe 和 NVMe 初始化过程做一个简单的介绍，希望大家对于这两个初始化过程有个初步的认识。

说明：

本文讲解使用的 trace 文件是由 SerialTek PCIe 4 协议分析仪抓取的台湾 Phison 公司 PCIe Gen 4 M.2 NVMe SSD 在 Gigabyte AMD Gen 4 CPU 的主板上的加电初始化过程。由于篇幅关系，只能讲述 PCIe 和 NVMe 初始化过程中的关节环节，如果想深入了解初始化过程的每个细节，可以添加下面的微信获取 SerialTek PCIe 协议分析软件和完整的 trace 文件进行学习。

提示：对于 Saniffer 公司发布的 Gen 4/5/6 NVMe SSD 测试工具白皮书感兴趣的朋友，可以访问 saniffer 官方网站下面的连接下载最新更新的“PCIe Gen 4/5/6NVMe SSD 测试环境搭建和常用工具图解”文档。如果有其它问题可以点击本文左下角的“阅读原文”留下你的联系方式。

https://www.saniffer.com/cn/downloads/

NVMe SSD 底层使用 PCIe 总线，可以说 PCIe 总线是 NVMe SSD 的“高速公路”，NVMe SSD 硬件同时实现了 PCIe 协议和 NVMe 协议，因此上电初始化分为两步，首先是 PCIe 层的初始化，只有在 PCIe 初始化成功“路”通了以后，才进行第二步 NVMe 层的初始化。

注意：

1） 下面的 PCIe 初始化阶段的“盘”可以理解为任意的 PCIe end point device，例如网卡，GPU 卡等

2）PCIe 初始化发生在 BIOS 加电自检后扫描 PCIe 总线的设备的阶段。下面是 Legacy BIOS 和 UEFI BIOS 两种 BIOS 初始化的基本过程。

- **Legacy BIOS（Basic Input/Output System）执行过程**

1. 加电自检（POST）：计算机加电后，BIOS 开始执行自检程序，检查系统硬件是否正常，包括内存、硬盘、键盘等。

2. 初始化硬件：BIOS 初始化系统硬件，设置正确的工作模式、中断向量等。这包括设置时钟频率、内存参数、外围设备等。

3. 寻找可启动设备：BIOS 根据预先设定的启动顺序（通常是硬盘、光驱、USB 等），尝试从这些设备中找到可引导的操作系统。

4. 加载引导程序：BIOS 从可启动设备的引导扇区（通常是硬盘的主引导记录）读取引导程序，将控制权交给引导程序。

5. 启动操作系统：引导程序加载操作系统的内核镜像到内存，并开始执行操作系统的初始化过程。

在整个过程中，BIOS 起到了初始化硬件和启动系统的桥梁作用，是计算机启动的重要组成部分。

- **UEFI BIOS（Basic Input/Output System）执行过程**

UEFI（Unified Extensible Firmware Interface）BIOS 执行过程相比传统的 Legacy BIOS 有所不同，其主要步骤如下：

1. 启动固件初始化（EFI Firmware Initialization）：计算机加电后，UEFI 固件开始初始化硬件，执行自检和设备初始化。与传统的 POST 过程相比，UEFI 执行更快且更灵活，支持更多的硬件和功能。

2. UEFI 固件设置界面（UEFI Firmware Setup Interface）：与传统的 BIOS 设置界面类似，UEFI 提供了一个可视化的设置界面，允许用户配置系统设置，如启动顺序、硬件参数等。

3. 启动设备选择（Boot Device Selection）：UEFI 提供了一个更灵活的启动设备选择界面，用户可以直接从界面中选择启动设备，而不需要进入设置界面。

4. 启动管理器（Boot Manager）：UEFI 引入了一个启动管理器，负责管理系统中的可启动操作系统。用户可以在启动管理器中选择要启动的操作系统，而不需要手动设置启动顺序。

5. 加载 UEFI 应用程序（Load UEFI Applications）：UEFI 支持加载和运行 UEFI 应用程序，这些应用程序可以提供额外的功能和服务，如诊断工具、固件更新工具等。

6. 启动操作系统（Boot Operating System）：UEFI 通过加载操作系统的 Boot Loader 来启动操作系统。Boot Loader 通常位于 EFI 系统分区中，由 UEFI 固件直接加载并执行。

总体而言，UEFI BIOS 执行过程更加灵活、快速，并提供了更多的功能和特性，使得计算机的启动和管理更加方便和高效。

UEFI（Unified Extensible Firmware Interface）是由 Intel 公司发起的，最初于 2005 年提出，并由多家硬件和软件公司共同制定。UEFI 的发起者主要是为了解决 Legacy BIOS 存在的一些限制和缺陷，并提供更灵活、可扩展的固件接口。

- **UEFI 相对于 Legacy BIOS 的优势**

1. 更加灵活的固件接口：UEFI 提供了更强大、更灵活的固件接口，支持更多的硬件和功能，如硬盘容量更大、启动速度更快等。

2. 支持更大的硬盘和启动分区：Legacy BIOS 受限于 16 位系统，无法有效地支持超过 2TB 的硬盘或 2.2TB 以上的分区，而 UEFI 可以轻松支持更大容量的硬盘和分区。

3. 更快的启动速度：UEFI 的启动过程更快，因为它使用了现代的启动技术，如并行加载驱动程序和操作系统。

4. 更好的图形界面支持：UEFI 提供了更加直观、易用的图形界面，用户可以通过鼠标和触摸屏来进行设置和操作。

5. 更安全的启动过程：UEFI 支持安全启动（Secure Boot）功能，可以防止恶意软件在启动过程中进行植入或篡改，提高了系统的安全性。

总的来说，UEFI 相比 Legacy BIOS 具有更多的优势和功能，能够更好地适应现代计算机系统的需求，因此逐渐取代了传统的 Legacy BIOS。

- **以及 PCIe 设备初始化中 cold reset 和 hot reset 的区别**

在 PCIe 设备初始化中，cold reset 和 hot reset 是两种不同的重置方式，它们的主要区别在于重置时机和影响范围。

1. Cold Reset（冷重置）：

- 冷重置是在系统上电时执行的重置操作。

- 当系统上电或者重新启动时，所有 PCIe 设备都会进行冷重置。这意味着所有 PCIe 设备都会在启动时执行完整的初始化过程，包括重新进行 PCIe 链路协商和配置寄存器的初始化。

- 冷重置可以确保所有 PCIe 设备在系统启动时处于已知的初始状态，但是它可能会导致系统启动时间较长。

2. Hot Reset（热重置）：

- 热重置是在系统运行过程中执行的重置操作。

- 热重置仅影响执行重置操作的 PCIe 设备，而不会影响其他 PCIe 设备或系统其他部分。

- 热重置可以用于解决特定 PCIe 设备的问题，而不会影响整个系统的稳定性。

- 与冷重置不同，热重置不会重新进行 PCIe 链路协商，因此它可能会比冷重置速度更快，但它不会重新初始化 PCIe 设备的配置寄存器。

综上所述，冷重置和热重置在 PCIe 初始化中的差异主要体现在重置时机和影响范围上。冷重置是在系统上电或重新启动时执行的完整重置，而热重置是在系统运行过程中执行的局部重置。

- **PCIe LTSSM recovery 发生的原因以及过程**

在 PCIe（Peripheral Component Interconnect Express）架构中，LTSSM（Link Training and Status State Machine）负责管理和控制物理层链接的训练和状态转换。LTSSM 链路在进行训练和状态转换时可能会出现多次恢复（recovery）的情况，主要有以下几种可能的原因：

1. 信号干扰或噪声： 高速信号传输过程中，可能受到外部电磁干扰或信号噪声的影响，导致链路训练过程中发生错误，需要进行恢复。

2. 电缆质量问题： 使用低质量或损坏的电缆可能导致信号传输不稳定，从而触发 LTSSM 链路的恢复机制。

3. 设备兼容性问题： 如果连接的 PCIe 设备之间存在兼容性问题，可能导致链路训练失败或错误，需要进行恢复。

4. 时钟偏移： 时钟信号的偏移或不同步可能导致链路训练过程中的时序问题，触发恢复。

5. 硬件故障： PCIe 设备或接口的硬件故障可能导致链路训练过程中的错误，需要进行恢复。

6. 固件或驱动问题： 设备的固件或驱动程序可能存在 bug 或问题，导致链路训练失败或错误，需要进行恢复。

在进行 LTSSM 链路恢复时，系统会尝试重新训练链路或进行状态转换，以尽可能恢复正常的链接状态。

PCIe LTSSM 链路在发生恢复（recovery）时，通常会通过以下过程进行：

1. 错误检测： 链路状态机会不断检测链路上的错误情况，例如丢失的数据包、信号干扰等。当发现错误时，LTSSM 会记录错误并尝试识别其类型和原因。

2. 错误恢复触发： 当 LTSSM 检测到错误时，会根据错误的类型和严重程度触发相应的恢复动作。例如，可以触发 PHY 层重新训练、重新初始化链路、进行错误注销等操作。

3. PHY 层重新训练： 如果错误属于物理层（PHY）问题，例如时钟失步或信号干扰，LTSSM 可能会触发 PHY 层的重新训练，以尝试恢复正常的信号传输。

4. 状态转换： 在某些情况下，LTSSM 可能会尝试通过状态转换来恢复链路状态。例如，如果链路进入了错误状态，LTSSM 可能会尝试重新进入训练状态或重新初始化链路。

5. 错误报告和记录： 在恢复过程中，LTSSM 会记录错误信息，并可能向系统软件或固件报告错误情况，以便进一步分析和处理。

整个恢复过程通常是自动化的，并且由硬件和固件共同协作完成。LTSSM 会根据 PCIe 规范定义的流程和机制来进行恢复操作，以确保链路的稳定性和可靠性。

10.2.1 PCIe 初始化流程简介

PCIe（Peripheral Component Interconnect Express）协议中的 LTSSM（Link Training and Status State Machine）是用于管理和维护 PCIe 链路的状态机。以下是 LTSSM 的基本过程：

1. Detect（检测）：LTSSM 首先检测连接性，确定是否存在可用的链路。这涉及电缆和插槽之间的物理连接。

2. Polling（轮询）：LTSSM 在检测到物理连接后，开始轮询对端的状态，以便进行链路层的初始化。这包括寻找远端的 LTSSM 状态和确定通信参数。

3. Configuration（配置）：链路层初始化后，LTSSM 进入配置阶段，其中协商链路的参数，例如速度和宽度，以便进行高效的数据传输。

4. Training（训练）：在链路配置完成后，LTSSM 进入训练阶段，进行电气和时序训练，以确保链路的稳定性和可靠性。

5. Link-Up（链路建立）：当训练成功完成后，LTSSM 将链路状态转换为“Link-Up”，表示 PCIe 链路已建立，可以进行数据传输。

总体而言，LTSSM 通过一系列状态转换确保 PCIe 链路的稳定性和性能，并在链路建立后准备好进行数据交换。

PCIe 硬件初始化完成的标志是盘进入最大速率 L0 状态，进入 L0 状态后，主机和盘就能正常使用 TLP 报文进行数据传输。参见下图

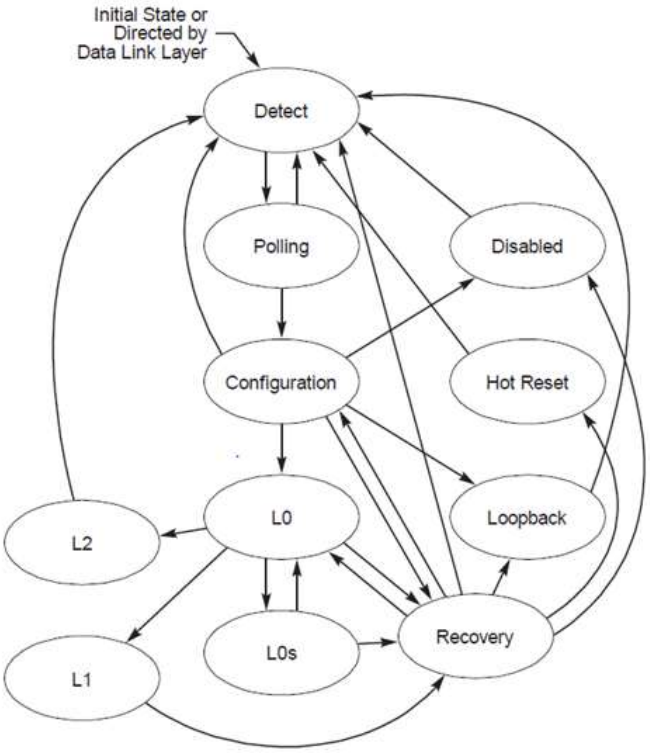


图 10-1

从状态机可以看到，盘进入 L0 只能是通过 Configuration 或者 Recovery 进入（L0s 只能通过 L0 状态进入，再退出到 L0）。

下图是抓取的一次盘的完整上电 LTSSM 跳转，左边是盘，右边是槽位。参见下图

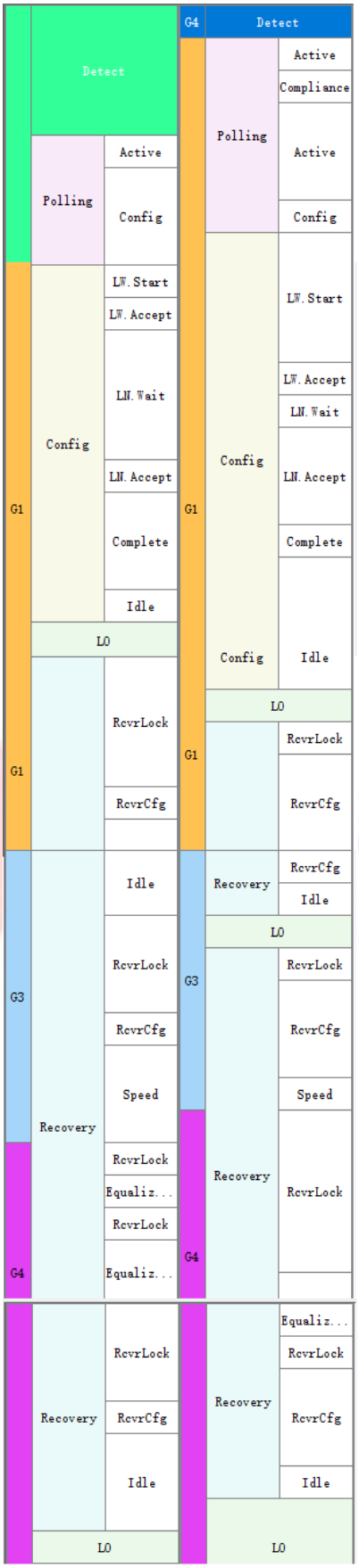

Detect
Polling
Active
Config
LW.Start
LW.Accept
LW.Wait
LW.Accept
Complete
Idle
L0
RcvrLock
RcvrCfg
Idle
RcvrLock
RcvrCfg
Speed
RcvrLock
Equaliz...
RcvrLock
Equaliz...
Recovery
RcvrLock
RcvrCfg
Idle
L0
G1
G3
G4
Detect
Polling
Active
Compliance
Active
Config
Config
LW.Start
LW.Accept
LW.Wait
LW.Accept
Complete
Idle
L0
RcvrLock
RcvrCfg
Recovery
RcvrCfg
Idle
L0
RcvrLock
RcvrCfg
Speed
RcvrLock
Recovery
Equaliz...
RcvrLock
RcvrCfg
Idle
L0

图 10-2

从整体的 LTSSM 可以看到，盘是从 Detect→polling→configuration→G1 L0→Recovery→G3→G4 L0；

接下来我们再来解释一下初始化过程中的每一个状态。

1）Detect

Detect 状态是设备上电复位或者热复位后的第一个状态，也就是 LTSSM 的入口状态，当前设备检测到对端设备在位后，就会往下进入 polling 状态。检测方法是发送端改变链路电压对链路充电，根据充电时间长短来判断对端是否在位。

2）Polling

进入此状态，说明两者已经相遇了，接下来就需要打个招呼，看下是否能够正常交流。

在这个状态下，两端设备通过互相发送 TS1 和 TS2 来确认 Bit Lock, Symbol Lock，Polarity Inversion 等，参见下图

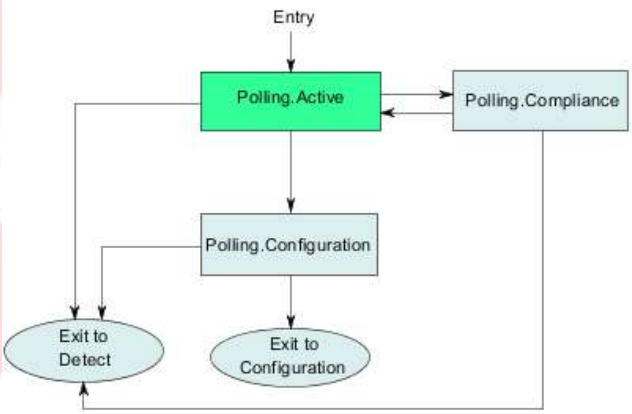


图 10-3

从 Trace 中可以看到这个过程中发送的内容都是 F7，也就是 PAD，无实际意义，只是通过发送一定数量的序列看是否满足协议要求。参见下图

Training Lanes	Training Links	Training Rates	Summary	Speed
				2.5
				2.5
F7	F7	2.5/5.0/8.0/16.0 GT/s		2.5
F7 F7 F7 F7	F7 F7 F7 F7	2.5/5.0/8.0/16.0 GT/s	Transmitt...	2.5
				2.5
F7 F7 F7 F7	F7 F7 F7 F7	2.5/5.0/8.0/16.0 GT/s	Transmitt...	2.5
				2.5
F7 F7 F7 F7	F7 F7 F7 F7	2.5/5.0/8.0/16.0 GT/s		2.5

图 10-4

3）Configuration

进入此状态，说明两者使用的是同一种语言，交流无障碍，为了能达成合作需要进行更深入的交流。

该阶段两边通过 TS1 和 TS2 来确认 Link number 和 Lane number。参见下图。

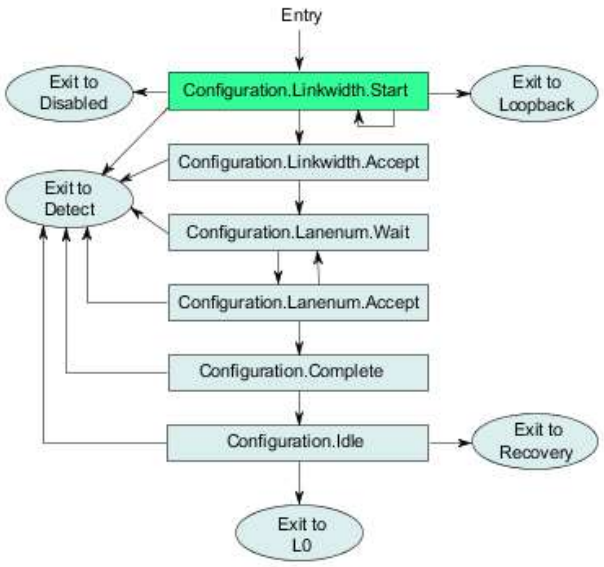


图 10-5

从 Trace 中可以看到 TS1,TS2 里面的 Training Lanes 和 Training Links 都是确切的值，通过这些值来进行信息交换。参见下图

Training Lanes	Training Links	Training Rates	Summary	Speed	Link Width
03 01 00 02	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		2.5	x4
				2.5	x1
				2.5	x1
03 01 00 02	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		2.5	x4
				2.5	x1
				2.5	x1
				2.5	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Transmitt...	2.5	x4
				2.5	x1
				2.5	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Transmitt...	2.5	x4
				2.5	x1
				2.5	x1
				2.5	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0 GT/s	Speed Change	2.5	x4
				2.5	x1
				2.5	x1
				2.5	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		2.5	x4
				2.5	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Speed Change	2.5	x4
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Speed Cha...	2.5	x4
				2.5	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0 GT/s	Speed Cha...	2.5	x4

图 10-6

4）L0

进入此状态后，意味着双方交流完成了，可以开始愉快合作了。也就是可以进行 DLLP 和 TLP 通信了。TLP packet 承载真正的命令（例如 MRd，CfgRd 等）和响应，每一个 TLP 发出后，接收端要回复一个 DLLP ACK 表示收到了 TLP，如果接收端收到的 TLP 有 bit error 或者 CRC error，那么就会回复 NAK，参见下图一个 CfgRd 的一个 transaction。

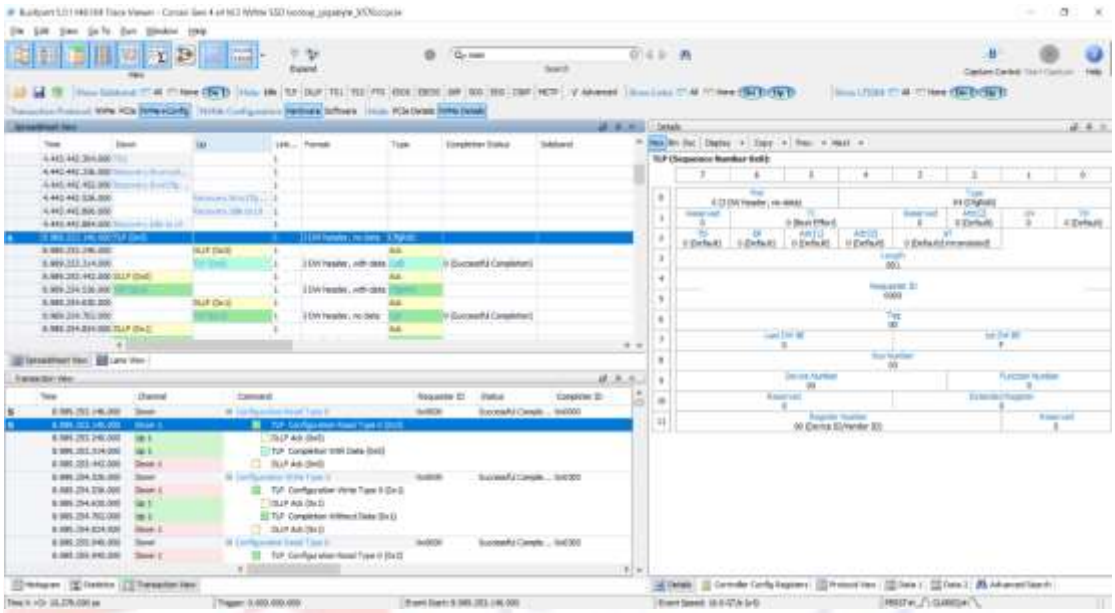


图 10-7

第一次进入的时候是 GEN1 的状态，但期望的是最大协商能力的 L0 状态，因此就需要再次重新交流一下，达到最佳合作状态，也就是跳入 Recovery 状态。

5）Recovery

进入此状态，说明双方认为还能继续交流一下，争取达到合作共赢的最佳状态。

这是一个重新训练状态，目的是达到最佳链路状态，进入该状态比如链路异常，未达到最大速率或者最大宽度。参见下图。

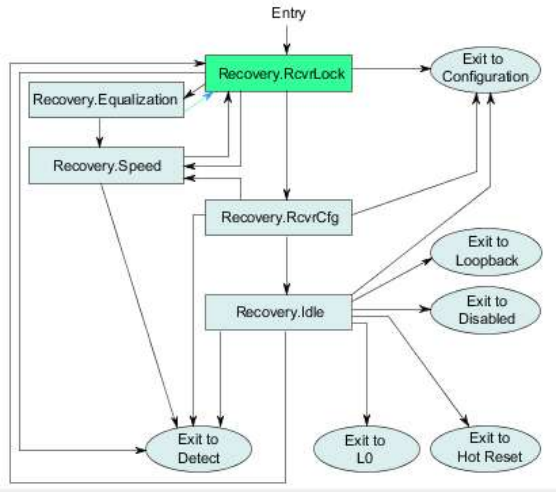


图 10-8

从 Trace 中可以看到，两端设备执行重协商动作都是通过 TS 序列中的一些特殊字段，比如 Speed Change Bit，EqCmd Bit 来指示实现的。参见下图。

0	00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Speed Change	2.5	x4
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Speed Cha...	2.5	x4
0					2.5	x1
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0 GT/s	Speed Cha...	2.5	x4
0					2.5	x1
0					2.5	x4
0					2.5	x1
0					2.5	x1
0					2.5	x1
0					2.5	x1
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		8.0	x4
0					8.0	x1
0					2.5	x1
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0 GT/s		8.0	x4
0					8.0	x1
0					8.0	x1
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		8.0	x4
0					8.0	x1
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0 GT/s		8.0	x4
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		8.0	x4
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0 GT/s		8.0	x4
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Use Prese...	8.0	x4
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0 GT/s		8.0	x4
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Use Prese...	8.0	x4
0	00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Use Prese...	8.0	x4

图 10-9

Training Lanes	Training Links	Training Rates	Summary	Speed	Link W
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Use Preset; Tr...	16.0	x4
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
				16.0	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
				16.0	x1
				16.0	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
				16.0	x1
				16.0	x1
				16.0	x1
				16.0	x1
				16.0	x1
			Cfg: Read Devi...	16.0	x4
				16.0	x4
			Cfg: Device ID...	16.0	x4
00 01 02 03	01 01 01 01	2.5/5.0/8.0 GT/s		8.0	x4
				8.0	x1
				8.0	x1
				8.0	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Speed Change	8.0	x4
				8.0	x1
				8.0	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		8.0	x4
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Speed Change	8.0	x4
				8.0	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Speed Change	8.0	x4
				8.0	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Speed Change	8.0	x4
				8.0	x1
				8.0	x4
				8.0	x1
				2.5	x1
				8.0	x1
				2.5	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
				16.0	x1
				2.5	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
				16.0	x1
				16.0	x1
				16.0	x1
				16.0	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4

图 10-10

经过 Recovery 后，盘和主机以 GEN4X4 的期望状态进入 L0，达到了最佳合作状态。

以上就是一次正常上电协商的状态机跳转，至于状态机中的其他状态可以参考协议。

至此 PCIe 硬件初始化已经全部完成，接下来就是主机软件对设备的处理，主要是设备的枚举以及资源分配，设备设置等。

从 Trace 中看到主机下发的第一个 TLP 报文，配置读取设备的 Device ID，这表明主机软件已经开始接管 PCIe 设备了。参见下图。

Training Lanes	Training Links	Training Rates	Summary	Speed	Link W
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s	Use Preset; Tr...	16.0	x4
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
				16.0	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
				16.0	x1
				16.0	x1
00 01 02 03	01 01 01 01	2.5/5.0/8.0/16.0 GT/s		16.0	x4
				16.0	x1
				16.0	x1
				16.0	x1
				16.0	x1
				16.0	x1
			Cfg: Read Devi...	16.0	x4
				16.0	x4
			Cfg: Device ID...	16.0	x4

图 10-11

关于主机软件对 PCIe 的初始化，我们暂且跳过，这是对所有 PCIe 设备的通用流程，接下来我们直接看一下 NVMe 层的初始化。

10.2.2 NVMe 初始化流程简介

主机软件初始化完 PCIe 后，开始加载 NVMe 驱动，也就是初始化 NVMe。（因为抓取的是上电 Trace，而这个主板 BIOS 支持 NVMe 设备，因此下面的 Trace 是 BIOS 下 NVMe 初始化流程，和 OS 下的 NVMe 驱动稍微有点差异，但整体原理和流程是一样的）。

首先看一下 NVMe Controller 的寄存器定义，有助于对照 Trace 解析。这些寄存器的基地址是设备的 BAR0 地址。参见下图。

Figure 68: Register Definition

Start	End	Symbol	Description
0h	7h	CAP	Controller Capabilities
8h	Bh	VS	Version
Ch	Fh	INTMS	Interrupt Mask Set
10h	13h	INTMC	Interrupt Mask Clear
14h	17h	CC	Controller Configuration
18h	1Bh	Reserved	Reserved
1Ch	1Fh	CSTS	Controller Status
20h	23h	NSSR	NVM Subsystem Reset (Optional)
24h	27h	AQA	Admin Queue Attributes
28h	2Fh	ASQ	Admin Submission Queue Base Address
30h	37h	ACQ	Admin Completion Queue Base Address
38h	3Bh	CMBLOC	Controller Memory Buffer Location (Optional)
3Ch	3Fh	CMBSZ	Controller Memory Buffer Size (Optional)
40h	43h	BPINFO	Boot Partition Information (Optional)
44h	47h	BPRSEL	Boot Partition Read Select (Optional)
48h	4Fh	BPMBL	Boot Partition Memory Buffer Location (Optional)
50h	57h	CMBMSC	Controller Memory Buffer Memory Space Control (Optional)
58h	5Bh	CMBSTS	Controller Memory Buffer Status (Optional)
5Ch	DFFh	Reserved	Reserved
E00h	E03h	PMRCAP	Persistent Memory Capabilities (Optional)
E04h	E07h	PMRCTL	Persistent Memory Region Control (Optional)
E08h	E0Bh	PMRSTS	Persistent Memory Region Status (Optional)
E0Ch	E0Fh	PMREBS	Persistent Memory Region Elasticity Buffer Size
E10h	E13h	PMRSWTP	Persistent Memory Region Sustained Write Throughput
E14h	E1Bh	PMRMSC	Persistent Memory Region Controller Memory Space Control (Optional)
E1Ch	FFFh	Reserved	Command Set Specific
1000h	1003h	SQ0TDBL	Submission Queue 0 Tail Doorbell (Admin)
1000h + (1 * (4 << CAP.DSTRD))	1003h + (1 * (4 << CAP.DSTRD))	CQ0HDBL	Completion Queue 0 Head Doorbell (Admin)

图 10-12

1）获取 NVMe 设备的基本信息

参见下图，可以看到 BAR0 基地址是 0Xfc80000,读取了偏移 0，8，14，1c 寄存器，从寄存器状态可以看到这是一个支持 NVMe 1.3 协议的控制器，并且 NVMe 层 Not Ready。

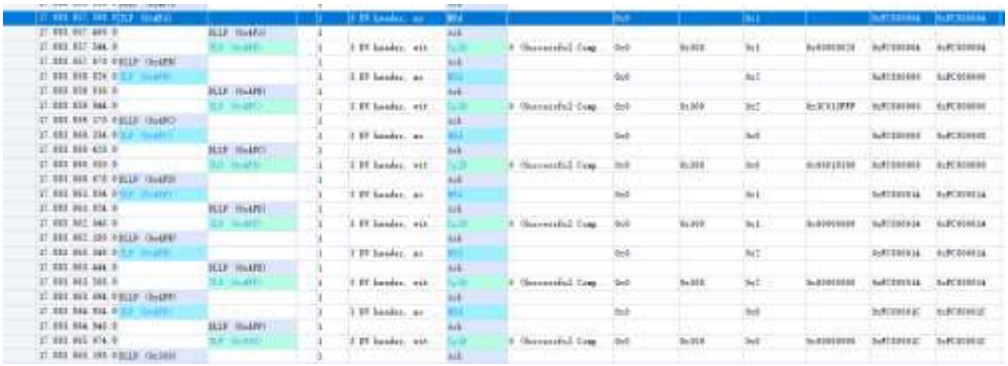


图 10-13

2）配置 NVMe 设备的 Admin Queue

参见下图 Trace 中可以看到主机写了偏移 14（Controller Configuration）,24(Admin Queue Attributes)，28(Admin SQ Base Addr Low)，2C(Admin CQ Base Addr High)，30(Admin CQ Base Addr Low)，34（Admin CQ Base Addr High）寄存器，其中 Admin SQ 的基地址为 0xbd551000,Admin CQ 的基地址为 0Xbd54c000。

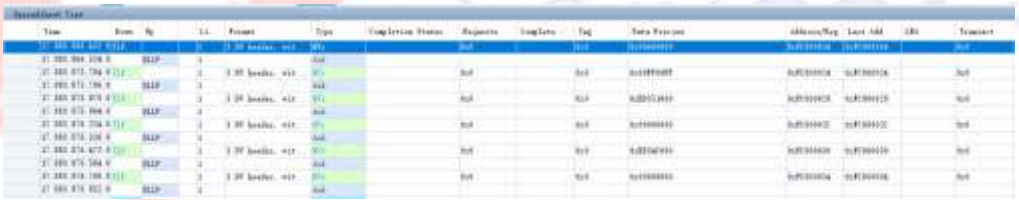


图 10-14

参见下图，这是一个典型的 NVMe storage 的架构图，从图中可以看出需要有 admin submission queue 以及 completion queue, 然后创建 IO submission queue 和 completion queue。

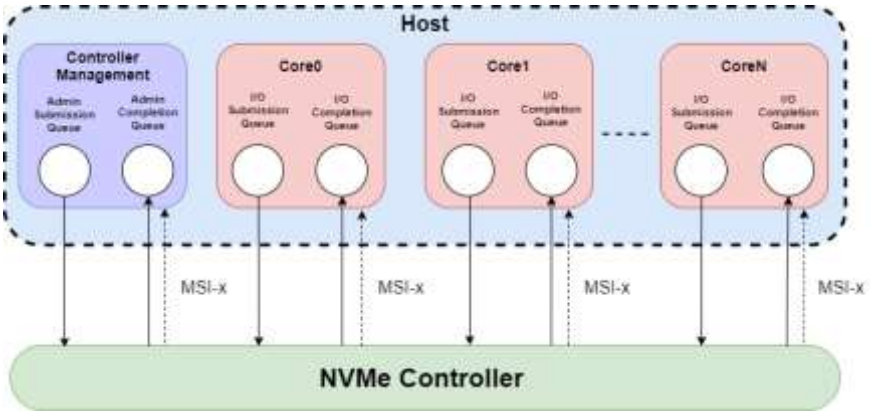


图 10-15

3）做 NVMe Controller Reset，等待 Reset 完成

参见下图，写偏移 14 寄存器的 Bit0,做 NVMe Controller Reset，然后轮询 1C 寄存器的 Bit0，等待 status 为 1，为 1 表明盘侧 NVMe reset 完成，NVMe Controller Ready。

这一步完成后，主机和盘之间可以通过 Admin Queue 进行管理消息通信。

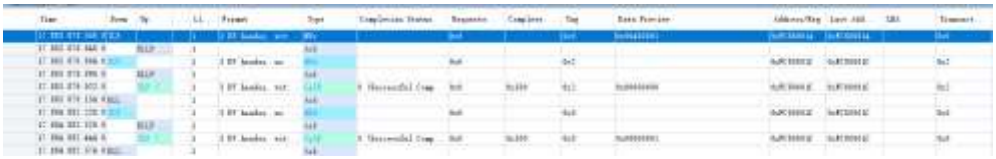


图 10-16

4）初始化 NVMe 字符设备

参见下图，盘硬件 NVMe 初始化完成后，盘能执行 Admin 命令，主机给盘发送一些管理命令从而获取到盘的信息，包括 set-feature 和 identify 这些命令。主机通过盘返回的信息，创建字符设备，完成 NVMe 字符设备初始化。下面是一个 Admin 命令（set-feature）的 Trace，从这个 Trace 中我们可以看到一个完整命令的执行过程，这些地址都能和前面 Trace 看到的设置地址一致。其他命令，包括 IO 命令也是类似的。

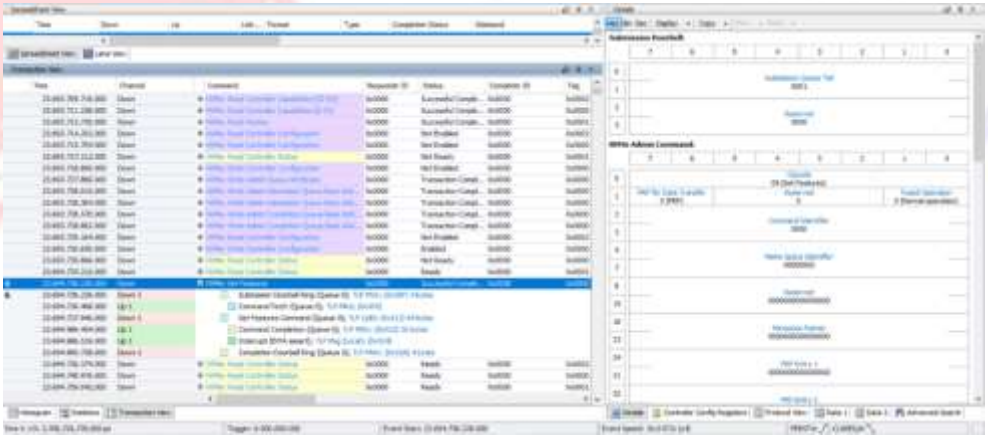


图 10-17

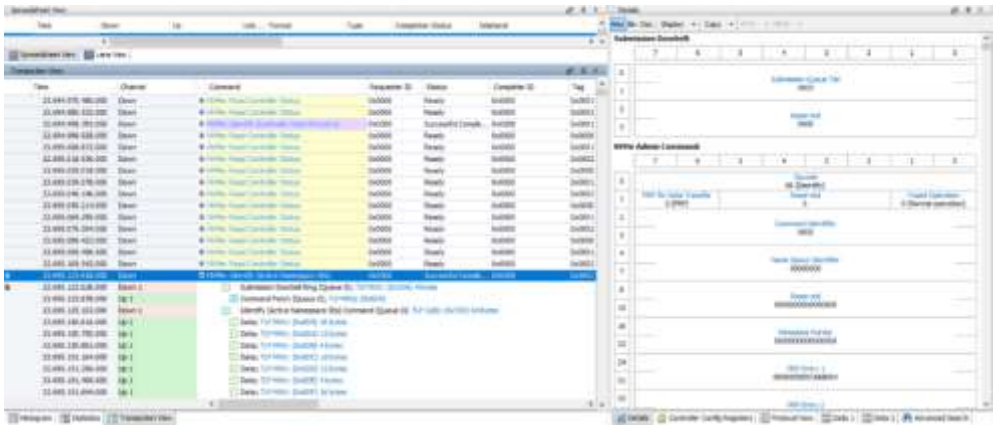


图 10-18

5）初始化 NVMe 块设备

参见下图，主机对盘数据的读写 IO 操作是通过块设备来完成的，盘的每个 NS 在主机上就是一个块设备，并且 IO 是通过 IO Queue 来通信，和 Admin Queue 分离。

首先主机会创建 IO CQ 和 IO SQ（queue 的个数以及 SQ/CQ 绑定关系由主机软件决定），然后发送 identify ns 枚举所有的 ns，并且为每个 ns 创建一个块设备，完成主机块设备初始化。

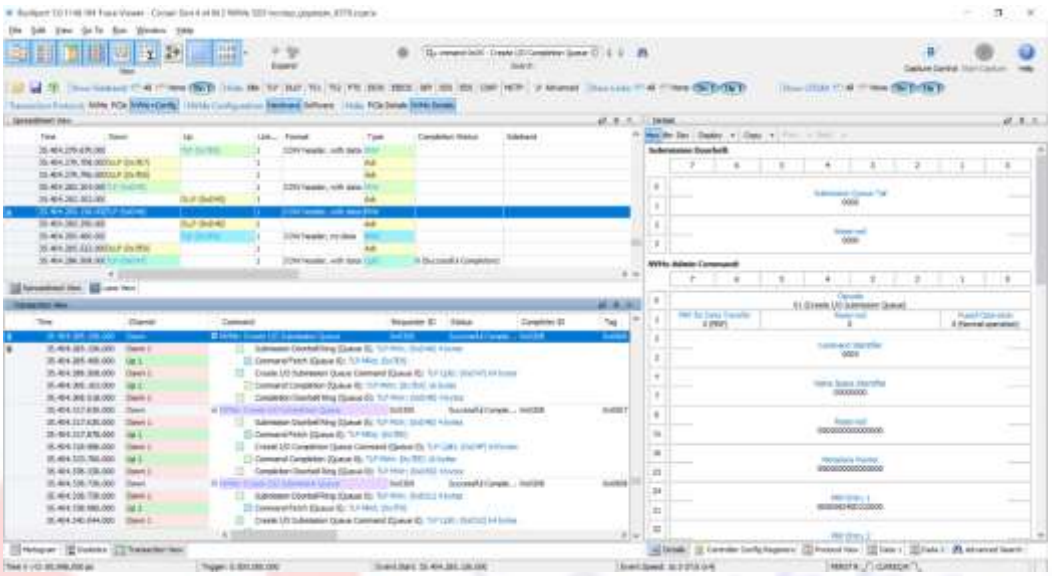


图 10-19 建 IO submission queue 的解码

下图为初始化过程中创建了多个 IO submission queue 和 completion queue, 然后才开始进行 read 读操作。

图 10-20

完成后主机就可以对盘上数据进行读写操作了，到此整个 NVMe SSD 初始化完成。

以下是一个 Read Cmd，其中下图为 transaction 图，

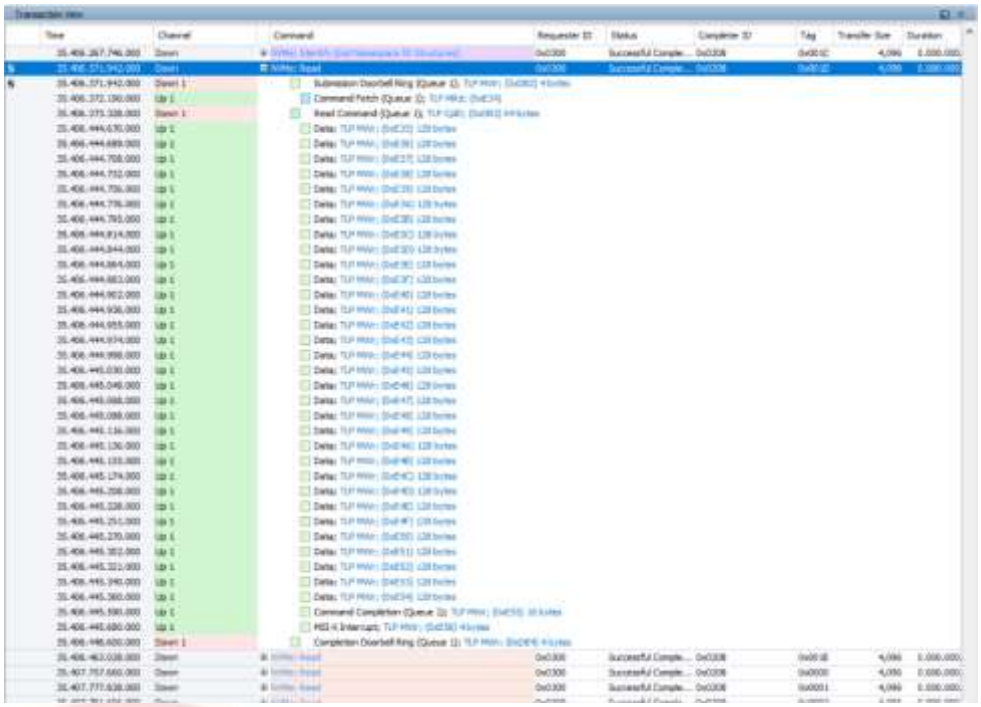


图 10-21 transaction 图

下图为 NVMe command 处理的过程，

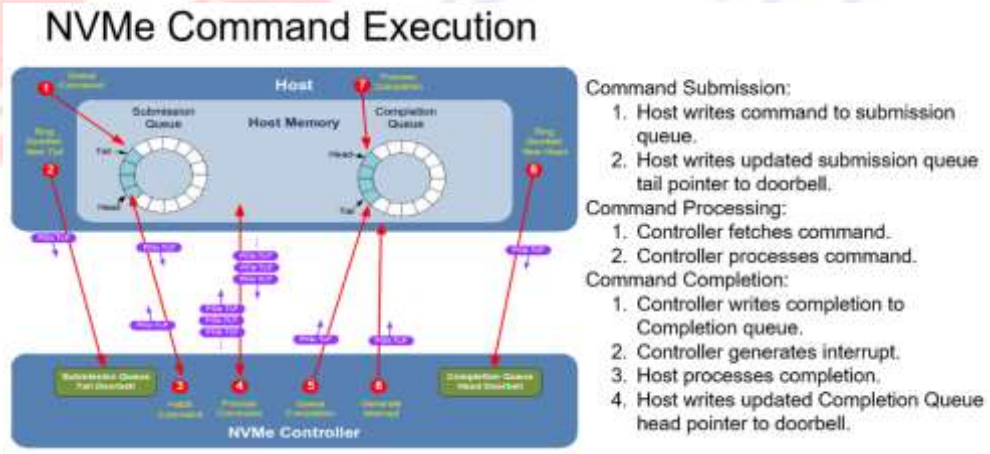


图 10-22 NVMe command 处理的过程

下图为该 Read Cmd 的 transaction 过程解码


Details
Submission Doorbell:
Submission Queue Tail
0002
Reserved
0000
NVMe IO Command:
Opcode
02 (Read)
PRP or SGL for Data Transfer
0 (PRP)
Reserved
0
Fused Operation
0 (Normal operation)
Command Identifier
0001
Name Space Identifier
00000001
Reserved
Metadata Pointer
PRP Entry 1
PRP Entry 2
Starting LBA
Number of Logical Blocks
0000
Reserved
000
Limited Retry
1
Protection Information Field
0
Incompressible
0 (no inform...
Access Latency
0 (None. No latency inform...
Access Frequency
0 (No frequency information provided)
Reserved
000000
Expected Initial Logical Block Reference Tag
00000000
Expected Logical Block Application Tag
0000
Expected Logical Block Application Tag Mask
0000
Completion Queue Entry:
Command Specific
00000000
Reserved
00000000
Submission Queue Head Pointer
0002
Submission Queue Identifier
0001
Command Identifier
0001
Status Code
00 (Successful Completion)
Phase Tag
1
More (M)
0
Command Retry Delay (CRD)
0
Status Code Type
0 (Generic Command Status)
MSI-X:
Message Data
00000007
Completion Doorbell:
Completion Queue Head
0002
Reserved
0000

图 10-23

10.3 蛋蛋读 NVMe 系列

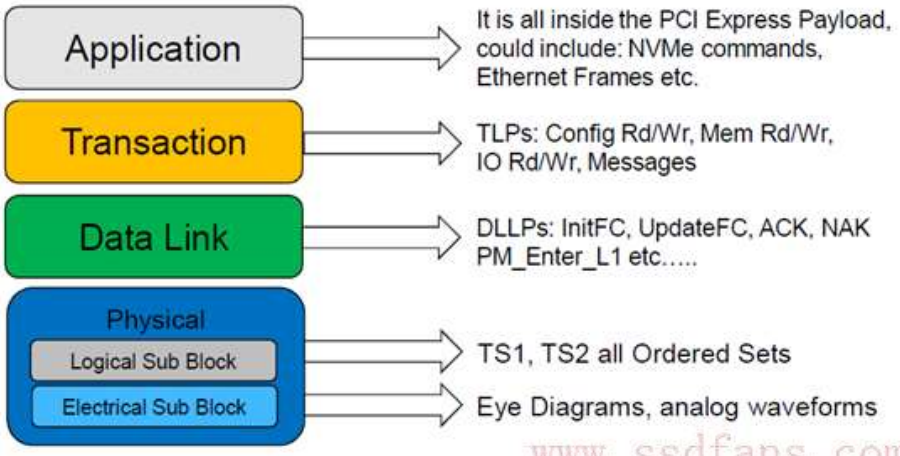
10.3.1 蛋蛋读 NVMe 之一

Posted on 2017 年 8 月 3 日 by SSD Fans

原创内容，转载请注明： [http://www.ssdfans.com] 谢谢！

没有前戏，直接进入。蛋蛋就是这么个人。

NVMe 是一种 Host 与 SSD 之间通讯的协议，它在协议栈中隶属高层。



NVMe 在协议栈中处于应用层或者命令层，它是指挥官，军师，在三国的话，就是诸葛亮的角色。”运筹帷幄之中，决胜千里之外”。军师设计好计谋，就交由手下五虎大将去执行。NVMe 的手下大将就是 PCIe，它所制定的任何命令，都交由虎将 PCIe 去完成。虽然 NVMe 的命令可能可以由别的接口协议完成，但 NVMe 与 PCIe 合作形成的战斗力无疑是最强的。

NVMe 是为 SSD 所生的。NVMe 出现之前，SSD 绝大多数走的是 AHCI 和 SATA 的协议，后者其实是为传统 HDD 服务的。与 HDD 相比，SSD 具有更低的延时和更高的性能，AHCI 已经不能跟上 SSD 性能发展的步伐了，已经成为制约 SSD 性能的瓶颈。所有 SATA 接口的 SSD，你去看性能参数，会发现都不会超过 600MB/s。如果碰到有人跟你说它的 SATA SSD 读取性能可以超过 600MB/s，直接拨打 110 报警。不是底层 Flash 带宽不够，是 SATA 接口速度限制了，因为 SATA 现在最高带宽就是 600MB/s。OK，既然 SATA 接口速度太慢，我用 PCIe 好了，不过上层协议还是 AHCI。五虎上将有了，由刘备指挥，让人不禁感叹暴殄天物呀。刘备什么水平，诸葛亮出现之前，居无定所，一会跟着曹操混，一会又跟着吕布混，谁肯收留就跟谁混。惨呀！AHCI 和刘备一个德行，只有一个命令队列，最多同时只能发 32 条命令，HDD 时代（群雄逐鹿）还能混混，SSD 时代（三足鼎立）就只有被灭的份。刘备需要三顾茅庐，需要诸葛亮的辅佐。同样，SSD 需要 PCIe，更需要 NVMe。

在这样的背景下，Intel 等巨头携天子以令诸侯，集大家智慧，制定出了 NVMe 规范，目的就是释放 SSD 性能潜力，解 SSD 倒悬之苦。

上面只列了几个巨头，参与的公司远不止这些。没有上榜的公司不要见怪。

NVMe 制定了 Host 与 SSD 之间通讯的命令，以及命令如何执行的。

NVMe 有两种命令，一种叫 **Admin Command**，用以 Host 管理和控制 SSD；另外一种就是 **I/O Command**，用以 Host 和 SSD 之间数据的传输。下面是 NVMe1.2 支持的命令列表：

NVMe 支持的 Admin Command：

Figure 40: Opcodes for Admin Commands

Opcode (07) Generic Command	Opcode (06:02) Function	Opcode (01:00) Data Transfer	Opcode[2]	O/M[1]	Namespace Identifier Used[3]	Command
0b	000 00b	00b	00h	M	No	Delete I/O Submission Queue
0b	000 00b	01b	01h	M	No	Create I/O Submission Queue
0b	000 00b	10b	02h	M	Yes	Get Log Page
0b	000 01b	00b	04h	M	No	Delete I/O Completion Queue
0b	000 01b	01b	05h	M	No	Create I/O Completion Queue
0b	000 01b	10b	06h	M	Yes	Identify
0b	000 10b	00b	08h	M	No	Abort
0b	000 10b	01b	09h	M	Yes	Set Features
0b	000 10b	10b	0Ah	M	Yes	Get Features
0b	000 11b	00b	0Ch	M	No	Asynchronous Event Request
0b	000 11b	01b	0Dh	O	Yes	Namespace Management
0b	001 00b	00b	10h	O	No	Firmware Commit
0b	001 00b	01b	11h	O	No	Firmware Image Download
0b	001 01b	01b	15h	O	Yes	Namespace Attachment
	I/O Command Set Specific					
1b	na	Na	80h – BFh	O		I/O Command Set specific
	Vendor Specific					
1b	na	Na	C0h – FFh	O		Vendor specific

NOTES:
1. O/M definition: O = Optional, M = Mandatory.
2. Opcodes not listed are reserved.
3. A subset of commands uses the Namespace Identifier field (CDW1.NSID). When not used, the field shall be cleared to 0h.

Figure 41: Opcodes for Admin Commands – NVM Command Set Specific

Opcode (07) Generic Command	Opcode (06:02) Function	Opcode (01:00) Data Transfer	Opcode[2]	O/M[1]	Namespace Identifier Used[3]	Command
1b	000 00b	00b	80h	O	Yes	Format NVM
1b	000 00b	01b	81h	O	Yes	Security Send
1b	000 00b	10b	82h	O	Yes	Security Receive

NOTES:
1. O/M definition: O = Optional, M = Mandatory.
2. Opcodes not listed are reserved.
3. A subset of commands uses the Namespace Identifier field (CDW1.NSID). When not used, the field shall be cleared to 0h.

NVMe 支持的 I/O Command：

Figure 149: Opcodes for NVM Commands

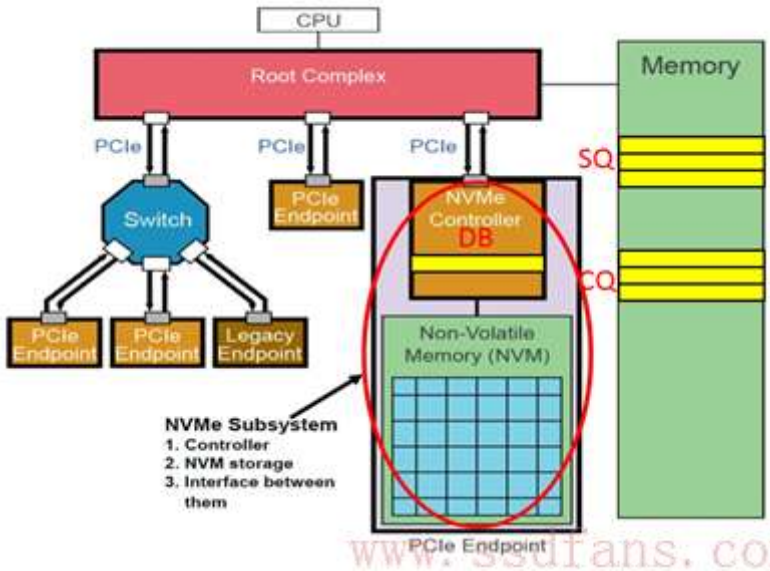
Opcode (07)	Opcode (06:02)	Opcode (01:00)	Opcode[2]	O/M[1]	Command[3]
Standard Command	Function	Data Transfer			
0b	000 00b	00b	00h	M	Flush
0b	000 00b	01b	01h	M	Write
0b	000 00b	10b	02h	M	Read
0b	000 01b	00b	04h	O	Write Uncorrectable
0b	000 01b	01b	05h	O	Compare
0b	000 10b	00b	08h	O	Write Zeroes
0b	000 10b	01b	09h	O	Dataset Management
0b	000 11b	01b	0Dh	O[4]	Reservation Register
0b	000 11b	10b	0Eh	O[4]	Reservation Report
0b	001 00b	01b	11h	O[4]	Reservation Acquire
0b	001 01b	01b	15h	O[4]	Reservation Release
	Vendor Specific				
1b	na	na	80h – FFh	O	Vendor specific

NOTES:
1. O/M definition: O = Optional, M = Mandatory.
2. Opcodes not listed are reserved.
3. All NVM commands use the Namespace Identifier field (CDW1.NSID).
4. Mandatory if reservations are supported as indicated in the Identify Controller data structure.

跟 ATA spec 中定义的命令相比，NVMe 的命令个数少了很多，完全是为 SSD 量身定制的。大家现在别纠结于具体的命令，了解一下就好。老板交代干活的时候，再找 spec 一个一个看吧。

命令有了，那么，Host 又是怎么把这些命令发送给 SSD 执行呢？

NVMe 有三宝：Submission Queue （SQ），Completion Queue（CQ）和 Doorbell Register （DB）。SQ 和 CQ 位于 Host 的内存中，DB 则位于 SSD 的控制器内部。上图：



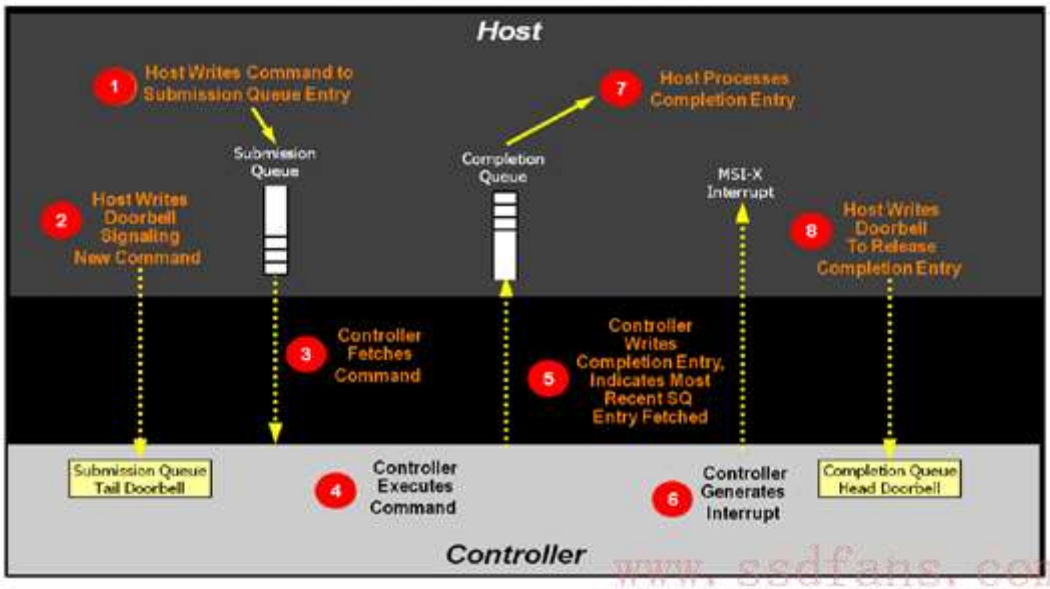
这张图信息量比较大，除了让我们知道 SQ 和 CQ 在 Host 的 memory 中以及 DB 在 SSD 端外，而且让我们对一个 PCIe 系统有一个具体的认识。上图中的 NVMe Subsystem 一般就是 SSD。请看这张图几秒钟，然后闭上眼，脑补 SSD 所处的位置：SSD 作为一个 PCIe Endpoint 通过 PCIe 连着 Root Complex （RC），然后 RC 连接着 CPU 和内存。RC 是什么？我们可以认为 RC 就是 CPU 的代言人，助理，或者小蜜。作为系统中最高层，CPU 说：我很忙的，你 SSD 有什么事情先跟我小蜜说！尽管如此，SSD 的地位还是较过去提升了一级，

过去 SSD 别说直接接触霸道总裁，就是连小蜜的面都见不到，SSD 和小蜜之间还隔着一座南桥呢。滚蛋吧，南桥君！

扯远了，刚才要说什么来着。对了，是三宝。SQ 位于 Host 内存中，Host 要发送命令时，先把准备好的命令放在 SQ 中，然后通知 SSD 来取；CQ 也是位于 Host 内存中，一个命令执行完成，成功或失败，SSD 总会往 CQ 中写入命令完成状态。DB（大宝？）又是干什么用的呢？Host 发送命令时，不是直接往 SSD 中发送命令的，而是把命令准备好放在自己的内存中，那怎么通知 SSD 来获取命令执行呢？Host 就是通过写 SSD 端的大宝寄存器来告知 SSD 的：饭已 OK 了，下来密西吧！

OK，具体的我们来看看 NVMe 是如何处理命令的，看图说话：

Figure 207: Command Processing



这是 NVMe1.2 规范中的第 207 张图。不知道是人家图画得好呢，还是 NVMe 就是这么简单，抑或是我比较聪明，反正上面的命令处理流程我一看就明白了。好吧，给没我聪明的人再解释一下。

说，把大象放冰箱一共要几步？答：三步。

第一步，打开冰箱门；

第二步，放进大象；

第三步，关上冰箱门。

说，NVMe 处理命令需要几步？答：八步：

第一步：Host 写命令到 SQ；

第二步：Host 写 DB，通知 SSD 取指；

第三步：SSD 收到通知，于是从 SQ 中取指；

第四步：SSD 执行指令；

第五步：指令执行完成，SSD 往 CQ 中写指令执行结果；

第六步：然后 SSD 发短信通知 Host 指令完成；

第七步：收到短信，Host 处理 CQ，查看指令完成状态；

第八步：Host 处理完 CQ 中的指令执行结果，通过 DB 回复 SSD：指令执行结果已处理，辛苦您了！

曹植七步作诗，NVMe 就比曹植差一点，需要八步。

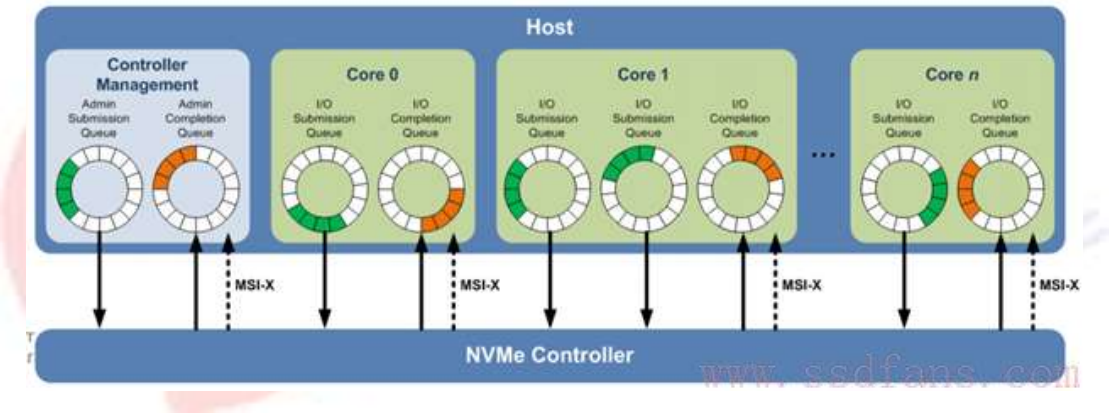
关于 NVMe，到现在相信大家有了一些基本认识。关于更多技术细节，今天我不打算讲了。我要吸取之前的教训，比如在一篇文章里就把 SSD 基本原理介绍了，而不是分别介绍。这样很不讨巧，一口气写完，对自己写文章是压力，对读者读文章也是压力，对网站的浏览量也不好。阿呆的做法值得学习，一个话题，采用连载的方式推出，有朋友也这么向我建议，于是我决定采取类似方式来谈 NVMe，毕竟 NVMe 是个大话题。于是，我把标题从”蛋蛋读 NVMe”改成”蛋蛋读 NVMe 之一”，后面还有之二，之三。。。接下来《蛋蛋读 NVMe 之二》我会详细解读 NVMe 的三宝 （SQ,CQ,DB），敬请期待。

10.3.2 蛋蛋读 NVMe 之二

Posted on 2017 年 8 月 3 日 by SSD Fans

上回书说道，NVMe 有三宝：SQ,CQ 和 DB。接下来我们就详细的看看这吉祥三宝。

Host 往 SQ 中写入命令， SSD 往 CQ 中写入命令完成结果。SQ 与 CQ 的关系，可以是一对一的关系，也可以是多对一的关系，但不管怎样，他们是成对的：有因就有果，有 SQ 就必然有 CQ。



有两种 SQ 和 CQ，一种是 Admin，另外一种是 I/O，前者放 Admin 命令，用以 Host 管理控制 SSD，后者放置 I/O 命令，用以 Host 与 SSD 之间传输数据。”你挑着担，我牵着马”（西游记的节奏呀），Admin SQ/CQ 和 I/O SQ/CQ 各司其职，你不能把 Admin 命令放到 I/O SQ 中，同样，你也不能把 I/O 命令放到 Admin SQ 里面。如果你不信这个邪，可以不遵守这个规矩试试，看看会发生什么，反正后果自负。

正如上图所示，系统中只有一对 Admin SQ/CQ，它们是一一对应的关系；I/O SQ/CQ 却可以很多，多达 65535（64K 减去一个 SQ/CQ）。行政人员少，干活的人多，很多公司都是这样的吧，所以 Admin SQ/CQ 少，I/O SQ/CQ 多就不难理解了。Host 端每个 Core 可以有一个或者多个 SQ，但只有一个 CQ。给每个 Core 分配一对 SQ/CQ 好理解，为什么一个 Core 中还要多个 SQ 呢？一是性能需求，一个 Core 中有多线程，可以做到一个线程独享一个 SQ；二是 QoS 需求，什么是 QoS？Quality of Service，服务质量。脑补一个场景，蛋蛋一边看小电影，同时迅雷在后台下载小电影，由于电脑配置差，看个小电影都卡。蛋蛋最讨厌看小电影的时候卡顿了，因为你刚刚燃起的激情会被那个缓冲浇灭。所以，蛋蛋不要卡顿！怎么办？NVMe 建议，你设置两个 SQ，一个赋予高优先级，一个低优先级，把看小电影所需的命令放到高优先级的 SQ，迅雷下载所需的命令放到低优先级的 SQ，这样，你那破电脑就能把有限的资源优先满足你看小电影了。至于迅雷卡不卡，下载慢不慢，这个时候已经不重要了。能让蛋蛋舒舒服服的看完一个小电影，就是好的 QoS。

实际系统中用多少个 SQ，取决于系统配置和性能需求，可灵活设置 I/O SQ 个数。关于系统中 I/O SQ 的个数，NVMe 白皮书给出如下建议：

Feature	Enterprise Recommended	Client Recommended
I/O Queues	16 to 128	2 to 8
Physically dis-contiguous queues	Design choice	No
Logical block size	4KB	4KB
Interrupt Support	MSI-X	MSI-X
Arbitration	WRR w/ Urgent or Round Robin	Round Robin
AER	Yes	Yes
Firmware Update	Required	Required
End-to-end data protection	Yes	No
SR-IOV support	Yes	No
Security Send and Receive	Yes	Yes

作为队列，每个 SQ 和 CQ 都有一定的深度：对 Admin SQ/CQ 来说，其深度可以是 2-4096（4K）；对 I/O SQ/CQ，深度可以是 2-65536(64K)。队列深度也是可以配置的。

SQ/CQ 的个数可以配置，每个 SQ/CQ 的深度又可以配置，因此 NVMe 的性能是可以通过配置队列个数和队列深度来灵活调节的。NVMe 太牛了吧，想胖就胖，想瘦就瘦；想高就高，想矮就矮，整一孙悟空呀！我们已经知道，AHCI 只有一个命令队列，且队列深度是固定的 32，就凡人一个，和 NVMe 相比，无论是在命令队列广度还是深度上，都是无法望其项背的；NVMe 命令队列的百般变化，更是 AHCI 无法做到的。说到百般变化，我突然又想到一件残忍的事情：PCIe 也是可以的。一个 PCIe 接口，可以有 1,2,4,8,12,16,32 条 lane！SATA 都要哭了，单挑都挑不过你，你还来群殴我。总之 AHCI/SATA 和 NVMe/PCIe 这么一比较，画面太美，蛋蛋不敢看。

蛋蛋在这里总是贬低 AHCI/SATA，有人要说蛋蛋忘恩负义，过河拆桥。怎么说？想当年，你 SSD 刚出来的时候，要不是 AHCI/SATA 收留了你，辛苦把你养大，都不知道你现在在哪里流浪。现在好了，你 SSD 翅膀硬了，不说一句感谢的话，倒反过来嫌弃我。各位看官，误会了，前面都是演戏，不说你 AHCI/SATA 不好，怎么能突出我 NVMe/PCIe 的好，毕竟后者才是男女一号，这么做完全是剧情需要。戏外，SSD 不会忘记你 AHCI/SATA 的好。忘恩负义？蛋蛋不是那种人。

虽然是在戏里，但总说 AHCI/SATA 的不好，这样真的好吗？蛋蛋是个怀旧的人，突然就有种蛋蛋的忧伤。好吧，以后就谈 NVME，不说 AHCI 了。孰好孰坏，留与读者评说。

戏还得继续演。

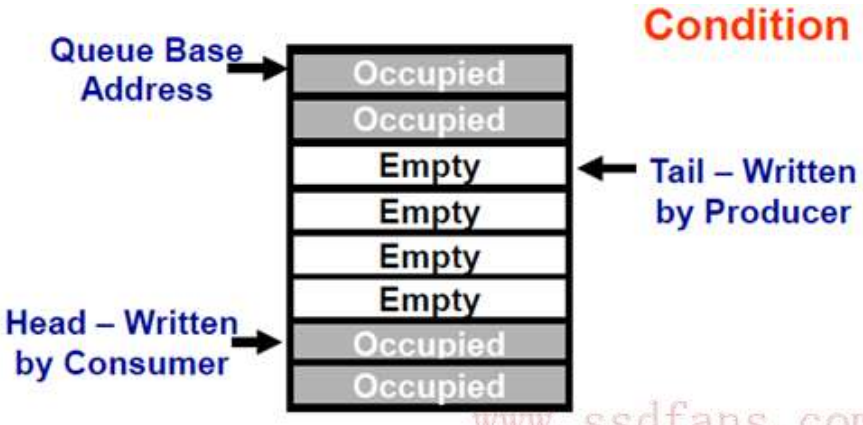
每个 SQ 放入的是命令条目，无论是 Admin 还是 I/O 命令，每个命令条目大小都是 64 字节；每个 CQ 放入的是命令完成状态信息条目，每个条目大小是 16 字节。

在继续谈大宝（DB）之前，先对 SQ 和 CQ 做个小结：

1. SQ 用以 Host 发命令，CQ 用以 SSD 回命令完成状态
2. SQ/CQ 可以在 Host 的内存中，也可以在 SSD 中，但一般在 Host 内存中（所有系列文章都是基于 SQ/CQ 在 Host 内存中讲的）；
3. 两种类型的 SQ/CQ：Admin 和 I/O，前者发送 Admin 命令，后者发送 I/O 命令;
4. 系统中只能有一对 Admin SQ/CQ，但可以有很多对 I/O SQ/CQ;
5. I/O SQ 与 CQ 可以是一对一的关系，也可以是一对多的关系；

6. I/O SQ 是可以赋予不同优先级的；
7. I/O SQ/CQ 深度可达 64K，Admin SQ/CQ 深达 4K；
8. I/O SQ/CQ 的广度和深度都可以灵活配置；
9. 每条命令大小是 64 字节，每条命令完成状态是 16 字节；
10. 不要过河拆桥。

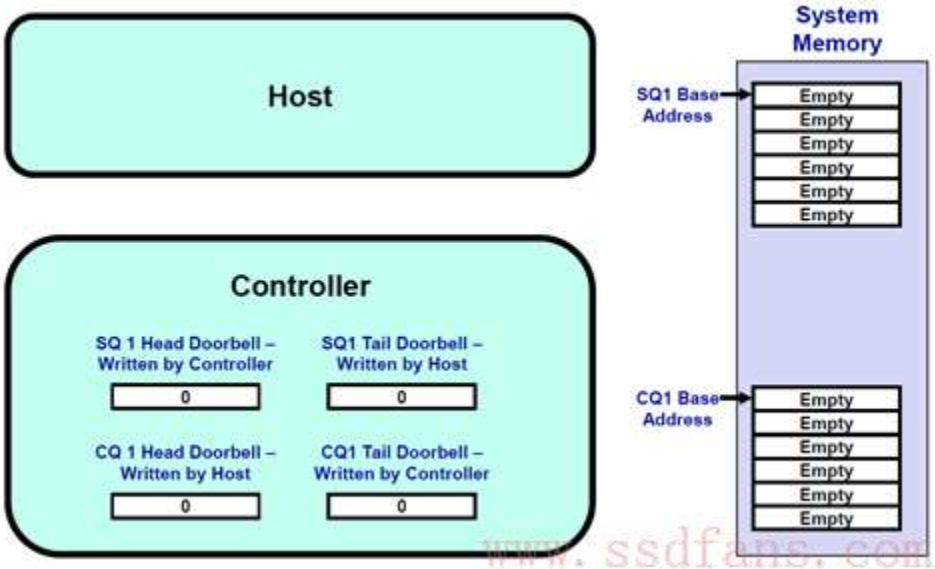
SQ/CQ 中的”Q”,是 Queue，队列的意思，无论 SQ 还是 CQ，都是队列，并且是环形队列。队列有几要素，除了队列深度，队列内容，还有两个重要的，就是队列的头（Head）和尾巴（Tail）。大家都排过队，你加入队伍的时候，都是站到队伍的最后，如果你插队，蛋蛋就会鄙视你。队伍最前头的那个，正在被服务或者等待被服务，一旦完成，就离开队伍。队列的头尾很重要，头决定谁会被马上服务，尾巴决定了新来的人站的位置。DB，就是用来记录了一个 SQ 或者 CQ 的 Head 和 Tail。每个 SQ 或者 CQ，都有两个对应的 DB: Head DB 和 Tail DB。DB 是在 SSD 端的寄存器，记录 SQ 和 CQ 的头和尾巴的位置。



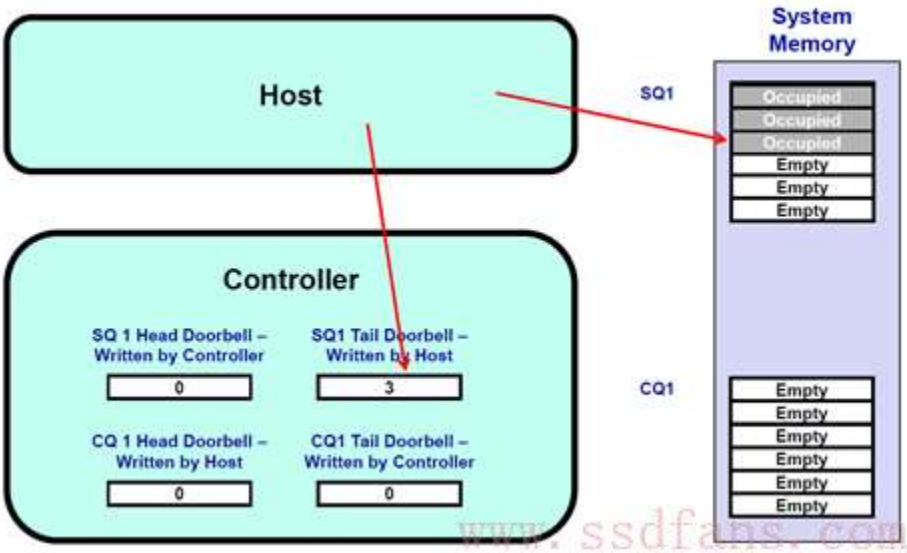
上面是一个队列的生产/消费模型。生产者往队列的 Tail 写入东西，消费者往队列的 Head 取出东西。对一个 SQ 来说，它的生产者是 Host，因为它往 SQ 的 Tail 位置写入命令，消费者是 SSD，因为它往 SQ 的 Head 取出指令执行；对一个 CQ 来说，刚好相反，生产者是 SSD，因为它往 CQ 的 Tail 写入命令完成信息，消费者则是 Host，它从 CQ 的 Head 取出命令完成信息。

举个例子，看图说话.

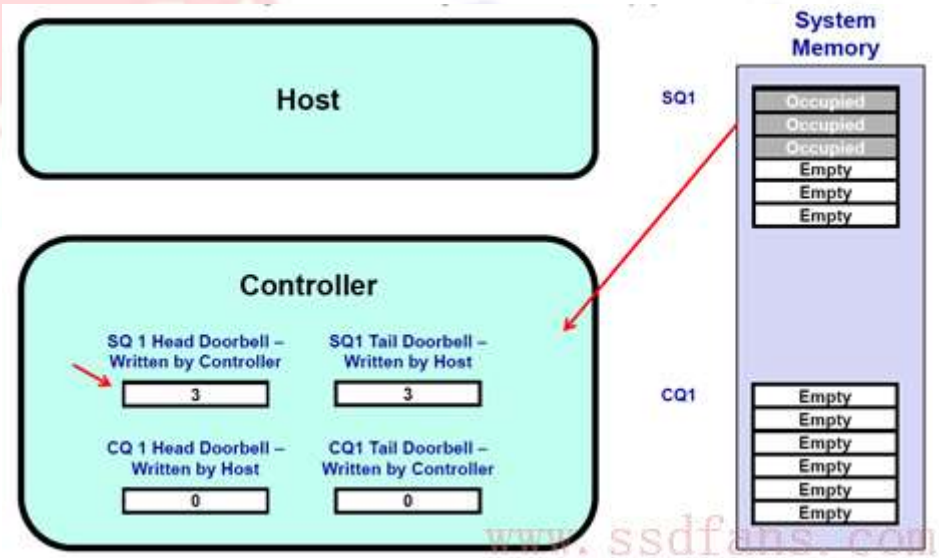
1. 开始假设 SQ1 和 CQ1 是空的，Head = Tail = 0.



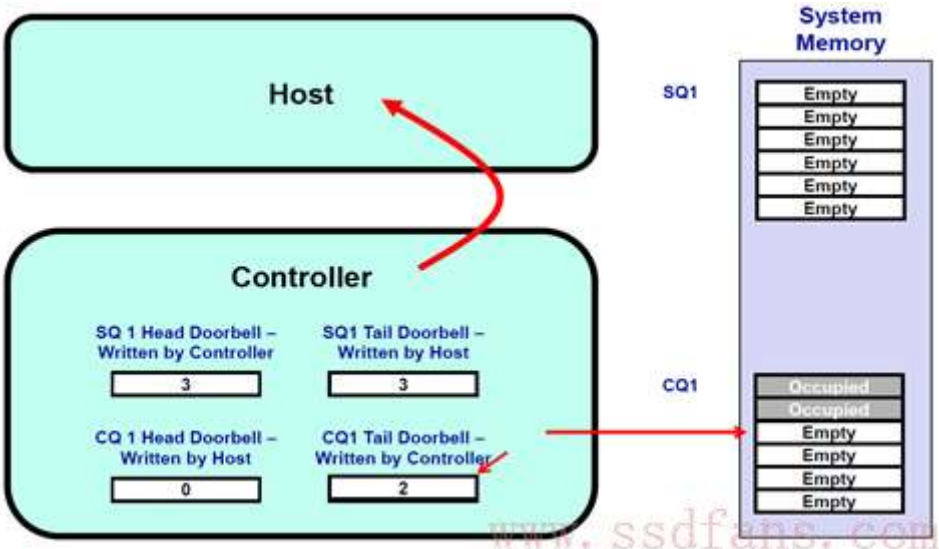
2. 这个时候，Host 往 SQ1 中写入了三个命令，SQ1 的 Tail 则变成 3。 Host 在往 SQ1 写入三个命令后，同时漂洋过海去更新 SSD Controller 端的 SQ1 Tail DB 寄存器，值为 3。Host 更新这个寄存器的同时，也是在告诉 SSD Controller：有新命令了，需要你去取。



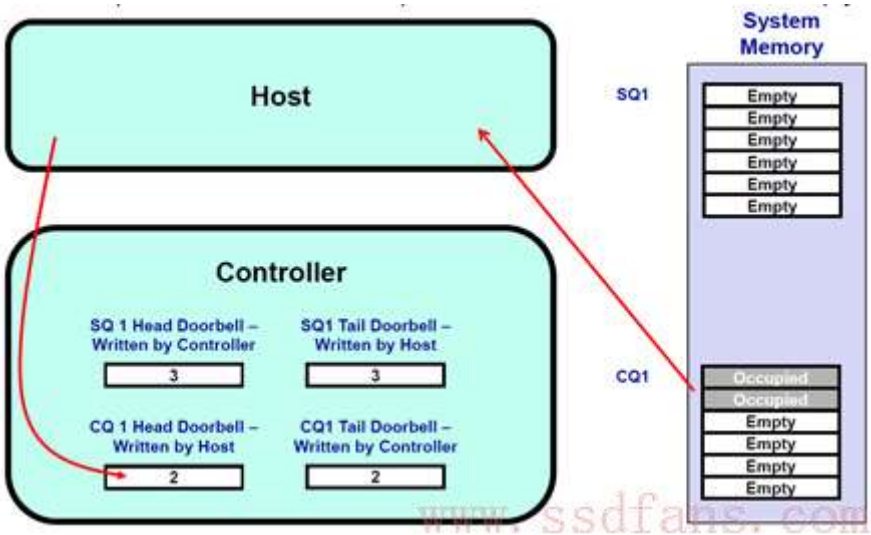
3. SSD Controller 收到通知后，于是派人去 SQ1 把 3 个命令都取回来执行。SSD 把 SQ1 的三个命令都消费了，SQ1 的 Head 从而也调整为 3，SSD Controller 会把这个 Head 值写入到本地的 SQ1 Head DB 寄存器。



4. SSD 执行完了两个命令，于是往 CQ1 中写入两个命令完成信息，同时更新 CQ1 对应的 Tail DB 寄存器，值为 2。SSD 并且发消息给 Host：有命令完成，请注意查看。



5. Host 收到 SSD 的短信通知，于是从 CQ1 中取出那两条完成信息处理。处理完毕，Host 又漂洋过海的往 CQ1 Head DB 寄存器中写入 CQ1 的 head，值为 2。



看完这个例子，又重温了一下命令处理流程。之前我们也许只记住了命令处理需要 8 步（距离曹植一步之遥），看完上面的例子，我们应该对命令处理流程有个更深入具体的认识。

那么，DB 在命令处理流程中起了什么作用呢？

首先，如前所示，它记住了 SQ 和 CQ 的头和尾。对 SQ 来说，SSD 是消费者，它直接和队列的头打交道，很清楚 SQ 的头在哪里，所以 SQ head DB 由 SSD 自己维护；但它不知道队伍有多长，尾巴在哪，后面还有多少命令等待执行，相反，Host 知道，所以 SQ Tail DB 由 Host 来更新。SSD 结合 SQ 的头和尾，就知道还有多少命令在 SQ 中等待执行了。对 CQ 来说，SSD 是生产者，它很清楚 CQ 的尾巴在哪里，所以 CQ Tail DB 由自己更新，但是 SSD 不知道 Host 处理了多少条命令完成信息，需要 Host 告知，因此 CQ Head DB 由 Host 更新。SSD 根据 CQ 的头和尾，就知道 CQ 能不能以及能接受多少命令完成信息。

DB 的另外一个作用，就是通知作用：Host 更新 SQ Tail DB 的同时，也是在告知 SSD 有新的命令需要处理；Host 更新 CQ Head DB 的同时，也是在告知 SSD，你返回的命令完成状态信息我已经处理，同时表示谢意。

这里有一个对 Host 不公平的地方，Host 对 DB 只能写，还仅限于写 SQ Tail DB 和 CQ Head DB，不能读取 DB。蛋蛋突然想唱首歌：

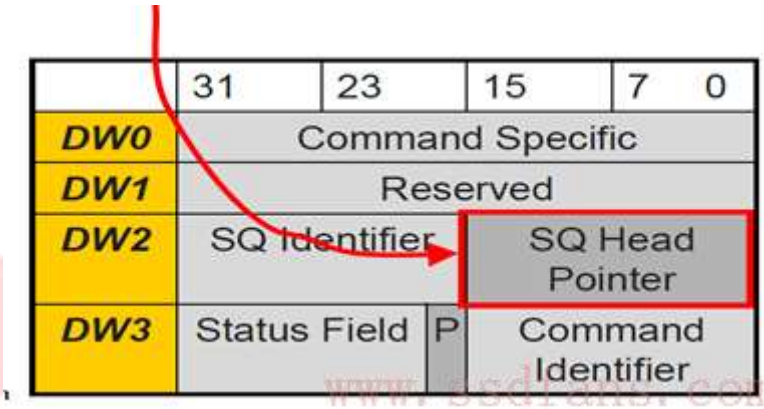
我俩太不公平

爱和恨全由你操纵

可今天我已离不开你

不管你爱不爱我

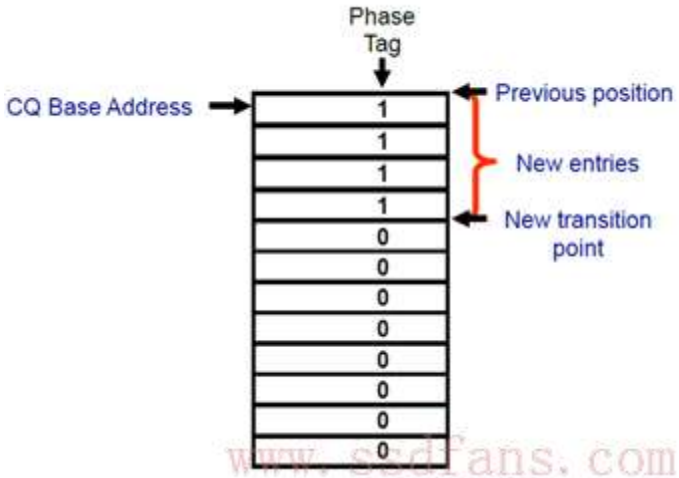
Host 就是这样痴情。在这个限制下，我们看看 Host 是怎样维护 SQ 和 CQ 的。SQ 的尾巴没有问题，Host 是生产者，对新命令来说，它清楚自己应该站在队伍哪里。但是 Head 呢？SSD 在取指的时候，是偷偷进行的，Host 对此毫不知情。Host 发了取指通知后，它并不清楚 SSD 什么时候去取命令，取了多少命令。怎么破？机智如你，如果是你，你会怎么做？山人自有妙计。给个提示：



这是什么鬼东西？这是 SSD 往 CQ 中写入的命令完成状态信息（16 字节）。

是的，SSD 往 CQ 中写入命令状态信息的同时，还把 SQ Head DB 的信息告知了 Host！！这样，Host 对 SQ 中 Head 和 Tail 的信息都有了，轻松玩转 SQ。

CQ 呢？Host 知道 Head，不知道 Tail。那怎么能知道 Tail 呢？思路很简单，既然你 SSD 知道，那你告诉我呗！SSD 怎么告诉 Host 呢？还是通过 SSD 返回命令状态信息中。哈哈，看到上图中的“P”吗？干什么用，做标记用。



具体是这样的：一开始 CQ 中每条命令完成条目中的”P” bit 初始化为 0，SSD 在往 CQ 中写入命令完成条目时，会把”P”写成 1。记住一点，CQ 是在 Host 端的内存中，Host 可以

检查 CQ 中的所有内容，当然包括”P”了。Host 记住上次的 Tail，然后往下一个一个检查”P”，就能得出新的 Tail 了。就是这样。

最后，给大宝做个小结：

1. DB 在 SSD Controller 端，是寄存器
2. DB 记录着 SQ 和 CQ 的 Head 和 Tail
3. 每个 SQ 或者 CQ 有两个 DB: Head DB 和 Tail DB
4. Host 只能写 DB，不能读 DB
5. Host 通过 SSD 往 CQ 中写入的命令完成状态获取 Head 或者 Tail

三宝介绍完了，今天就到这。《蛋蛋读 NVMe 之三》会是什么，我还没有想好，下次看了再说吧。

10.3.3 蛋蛋读 NVMe 之三

Posted on 2017 年 8 月 3 日 by SSD Fans

有个人一直在思考三个问题：我是谁？我从哪里来？我要去哪里？

你猜这个人最后怎么着？

成了哲学家?

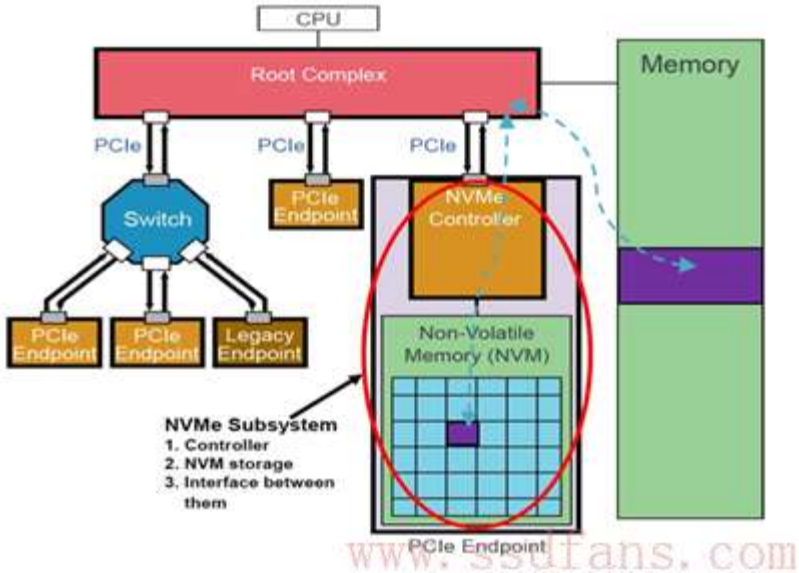
疯了?

疯了的哲学家?

我觉得无外乎这三种结果了。

相比人的世界，这三个问题在 NVMe 的世界就很容易得到答案了，至少不会把人逼疯。

我是数据，我从 Host 来，要到 SSD 去，或者，我从 SSD 来，要去到 Host。



Host 如果想往 SSD 上写入用户数据，需要告诉 SSD 写入什么数据，写入多少数据，以及数据源在内存中的什么位置，这些信息包含在 Host 向 SSD 发送的 Write 命令中。每笔用户数据对应着一个叫做 LBA（Logical Block Address）的东西，Write 命令通过指定 LBA 来告诉 SSD 写入的是什么数据。对 NVMe/PCIe 来说，SSD 收到 Write 命令后，通过 PCIe 去 Host 的内存数据所在位置读取数据，然后把这些数据写入到闪存中，同时得到 LBA 与闪存位置的映射关系。

Host 如果想读取 SSD 上的用户数据，同样需要告诉 SSD 需要什么数据，需要多少数据，以及数据最后需要放到 Host 内存的哪个位置上去，这些信息包含在 Host 向 SSD 发送的 Read 命令中。SSD 根据 LBA，查找映射表，找到对应闪存物理位置，然后读取闪存获得数据。数据从闪存读上来以后，对 NVMe/PCIe 来说，SSD 会通过 PCIe 把数据写入到 Host 指定的内存中。这样就完成了 Host 对 SSD 的读访问。

在上面的描述中，大家有没有注意到一个问题，那就是 Host 在与 SSD 的数据传输过程中，Host 是被动的一方，SSD 是主动的一方。你 Host 需要数据，是我 SSD 主动把数据写入到你的内存中；你 Host 写数据，同样是我 SSD 主动去你 Host 的内存中取数据，然后写入到闪存。SSD 跟快递小哥一样辛劳，不仅送货上门，还上门取件。之前蛋蛋还为 Host 不能读取 DB 打抱不平，现在看来，Host 不值得同情，太懒了。

无论送货上门，还是上门取件，你都需要告诉快递小哥你的地址，不然茫茫人海，快递小哥怎么就能找到你呢？同样的，Host 你不亲自传输数据，那总该告诉我 SSD 去你内存中什么地方取用户数据，或者要把数据写入到你内存中的什么位置。你在告诉快递小哥送货地址或者取件地址时，会说 XX 路 XX 号 XX 弄 XX 楼 XX 室，也可能会说 XX 小区 XX 楼 XX 室，anyway，快递小哥能找到就行。Host 也有两种方式来告诉 SSD 数据所在内存位置，一是 PRP (Physical Region Page, 不是 P2P!)，二是 SGL (Scatter/Gather List)。不过，后者感觉不怎么友善，因为怎么听起来都像”死过来”（SGL）。当然了，也可能是我误会了，人家只是在说”送过来”。

先说 PRP。

NVMe 把 Host 的内存划分为一个一个页（Page），页的大小可以是 4KB,8KB,16KB…128MB。

PRP 是什么，长什么样呢？

Figure 14: PRP Entry Layout

PRP Entry 本质就是一个 64 位内存物理地址，只不过把这个物理地址分成两部分：页起始地址和页内偏移。最后两 bit 是 0，说明 PRP 表示的物理地址只能四字节对齐访问。页内偏移可以是 0，也可以是个非零的值。

PRP Entry 描述的是一段连续的物理内存的起始地址。如果需要描述若干个不连续的物理内存呢？那就需要若干个 PRP Entry。把若干个 PRP Entry 链接起来，就成了 PRP List。

Figure 16: PRP List Layout

63 *n+1*	*n* 0
Page Base Address *k*	0h
Page Base Address *k+1*	0h
...	
Page Base Address *k+m*	0h
Page Base Address *k+m+1*	0h

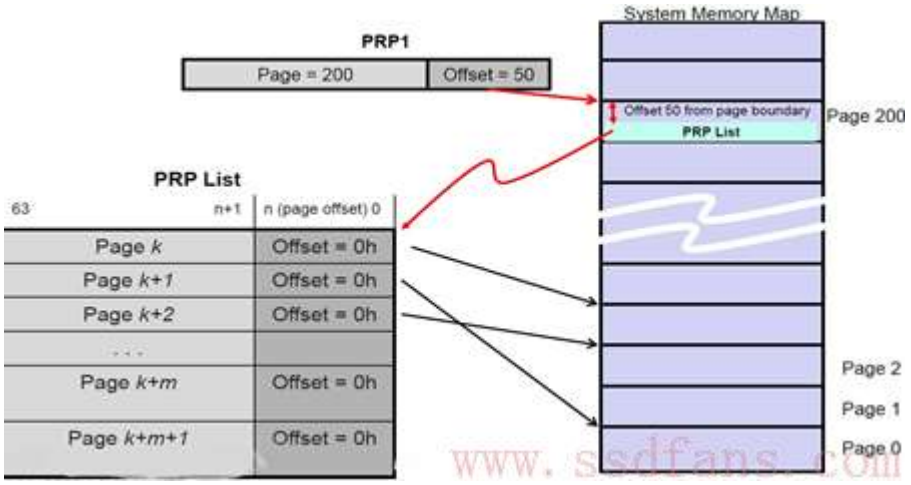
是的，正如你所见，PRP List 中的每个 PRP Entry 的偏移量都必须是 0，PRP List 中的每个 PRP Entry 都是描述一个物理页。它们不允许有相同的物理页，不然 SSD 往同一个物理页写入几次的数据，导致先写入的数据被覆盖。

每个 NVMe 命令中有两个域：PRP1 和 PRP2，Host 就是通过这两个域告诉 SSD 数据在内存中的位置或者数据需要写入的地址。

Bytes	Description
63:60	**Command Dword 15 (CDW15)**: command specific
59:56	**Command Dword 14 (CDW14)**: command specific
55:52	**Command Dword 13 (CDW13)**: command specific
51:48	**Command Dword 12 (CDW12)**: command specific
47:44	**Command Dword 11 (CDW11)**: command specific
43:40	**Command Dword 10 (CDW10)**: command specific
39:32	**PRP Entry 2 (PRP2)**: The 2nd address entry for commands that use it.
31:24	**PRP Entry 1 (PRP1)**: The first address entry for this command.
23:16	**Metadata Pointer (MPTR)**: Address of a contiguous metadata buffer.
15:8	Reserved
7:4	**Namespace Identifier (NSID)**: Namespace for this command. A value of 0h means NSID isn't used for this command; value of all F's means it applies to all namespaces on the device.
3:0	**Command Dword 0 (CDW0)**: Used by all commands

PRP1 和 PRP2 有可能指向数据所在位置，也可能指向 PRP List。类似 C 语言中的指针概念，PRP1 和 PRP2 可能是指针，也可能是指针的指针，还有可能是指针的指针的指针。别管你包的有多严实，根据不同的命令，SSD 总能一层一层的剥下包装，找到数据在内存的真正物理地址。SSD 善解人衣。

下面是一个 PRP1 指向 PRP List 的示例：



PRP1 指向一个 PRP List，PRP List 位于 Page 200，页内偏移 50 的位置。SSD 确定 PRP1 是个指向 PRP List 的指针后，就会去 Host 内存中（Page 200，Offset 50）把 PRP List 取过来。获得 PRP List 后，就获得数据的真正物理地址，SSD 然后就会往这些物理地址读入或者写入数据。

对 Admin 命令来说，它只用 PRP 告诉 SSD 内存物理地址；对 I/O 命令来说，除了用 PRP，Host 还可以用 SGL 的方式来告诉 SSD 数据在内存中写入或者读取的物理地址。

Figure 12: Command Format – NVM Command Set

Bytes	Description
63:60	**Command Dword 15 (CDW15):** This field is command specific Dword 15.
59:56	**Command Dword 14 (CDW14):** This field is command specific Dword 14.
55:52	**Command Dword 13 (CDW13):** This field is command specific Dword 13.
51:48	**Command Dword 12 (CDW12):** This field is command specific Dword 12.
47:44	**Command Dword 11 (CDW11):** This field is command specific Dword 11.
43:40	**Command Dword 10 (CDW10):** This field is command specific Dword 10.
39:24	**Data Pointer (DPTR):** This field specifies the data used in the command. If CDW0[15:14] is set to 00b, then the definition of this field is: 39:32 — **PRP Entry 2 (PRP2):** This field: a) is reserved if the data transfer does not cross a memory page boundary. b) specifies the Page Base Address of the second memory page if the data transfer crosses exactly one memory page boundary. E.g.: i. the command data transfer length is equal in size to one memory page and the offset portion of the PBAO field of PRP1 is non-zero or ii. the Offset portion of the PBAO field of PRP1 is equal to zero and the command data transfer length is greater than one memory page and less than or equal to two memory pages in size. c) is a PRP List pointer if the data transfer crosses more than one memory page boundary. E.g.: i. the command data transfer length is greater than or equal to two memory pages in size but the offset portion of the PBAO field of PRP1 is non-zero or ii. the command data transfer length is equal in size to more than two memory pages and the Offset portion of the PBAO field of PRP1 is equal to zero. 31:24 — **PRP Entry 1 (PRP1):** This field contains the first PRP entry for the command or a PRP List pointer depending on the command. If CDW0[15:14] is set to 01b or 10b, then the definition of this field is: 39:24 — **SGL Entry 1 (SGL1):** This field contains the first SGL segment for the command. If the SGL segment is a Data Block descriptor, then it describes the entire data transfer. If more than one SGL segment is needed to describe the data transfer, then the first SGL segment is a Segment, or Last Segment descriptor. Refer to section 4.4 for the definition of SGL segments and descriptor types.

Host 在命令中会告诉 SSD 采用何种方式。具体来说，如果命令当中 DW0[15：14]是 0，就是 PRP 的方式，否则就是 SGL 的方式。

SGL 是什么？SGL 是一个数据结构，用以描述一段数据空间，这个空间可以是数据源所在的空间，也可以是数据目标空间。SGL(Scatter Gather List)首先是个 List，是个链表，由一个或者多个 SGL Segment 组成，而每个 SGL Segment 又由一个或者多个 SGL Descriptor 组成。SGL Descriptor 是 SGL 最基本的单元，它描述了一段连续的物理内存空间：起始地址+空间大小。

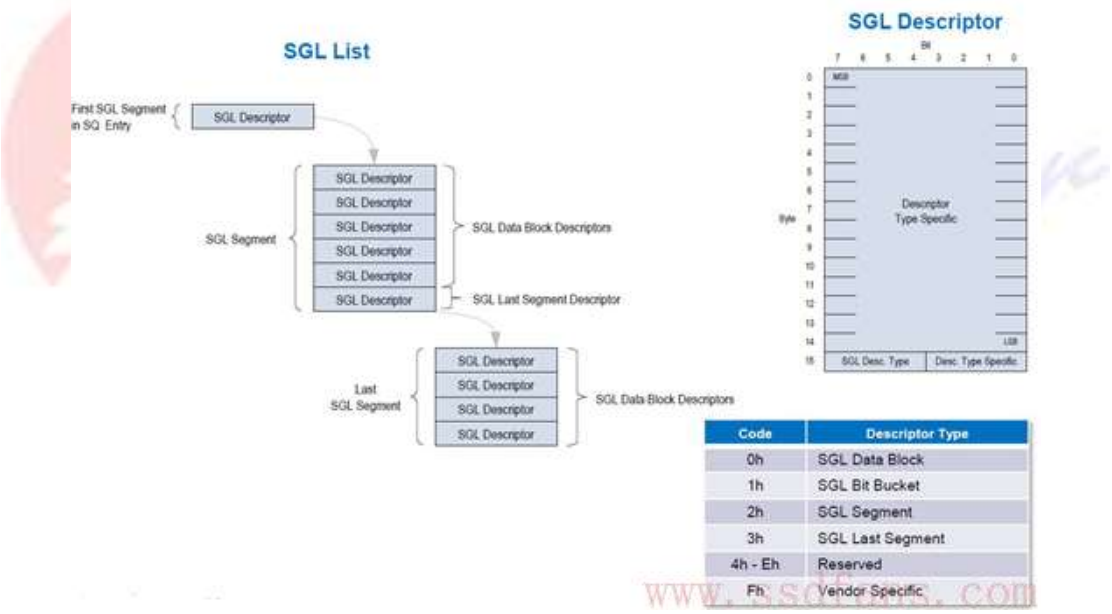
每个 SGL Descriptor 大小是 16 字节。一块内存空间，可以用来放用户数据，也可以用来放 SGL Segment，根据这段空间的不同用途，SGL Descriptor 也分几种类型。

Figure 19: SGL Descriptor Type

Code	Descriptor
0h	SGL Data Block descriptor
1h	SGL Bit Bucket descriptor
2h	SGL Segment descriptor
3h	SGL Last Segment descriptor
4h – Eh	Reserved
Fh	Vendor specific

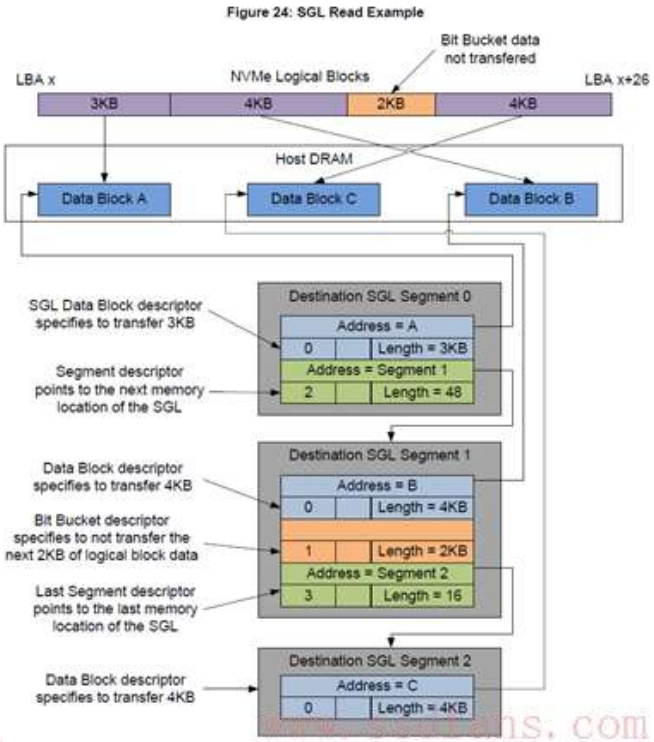
有 4 种 SGL Descriptor，一种是 Data Block，这个好理解，就是描述的这段空间是用户数据空间；一种是 Segment 描述符。SGL 不是由 SGL Segment 组成的链表吗？既然是链表，前面一个 Segment 就需要有个指针指向下一个 Segment，这个指针就是 SGL Segment 描述符，它描述的是它下个 Segment 所在的空间。特别地，对链表当中倒数第二个 Segment，它的 SGL Segment 描述符我们把它叫做 SGL Last Segment 描述符。它本质还是 SGL Segment 描述符，描述的还是 SGL Segment 所在的空间。为什么需要把倒数第二个 SGL Segment 描述符单独的定义成一种类型呢？我认为是让 SSD 在解析 SGL 的时候，碰到 SGL Last Segment 描述符，就知道链表快到头了，后面只有一个 Segement 了。那么，SGL Bit Bucket 是什么鬼？它只对 Host 读有用，用以告诉 SSD，你往这个内存写入的东西我是不要的。好吧，你既然不要，我也就不传了。

说了这么多，可能有点晕，结合下张图，可能会更明白点。



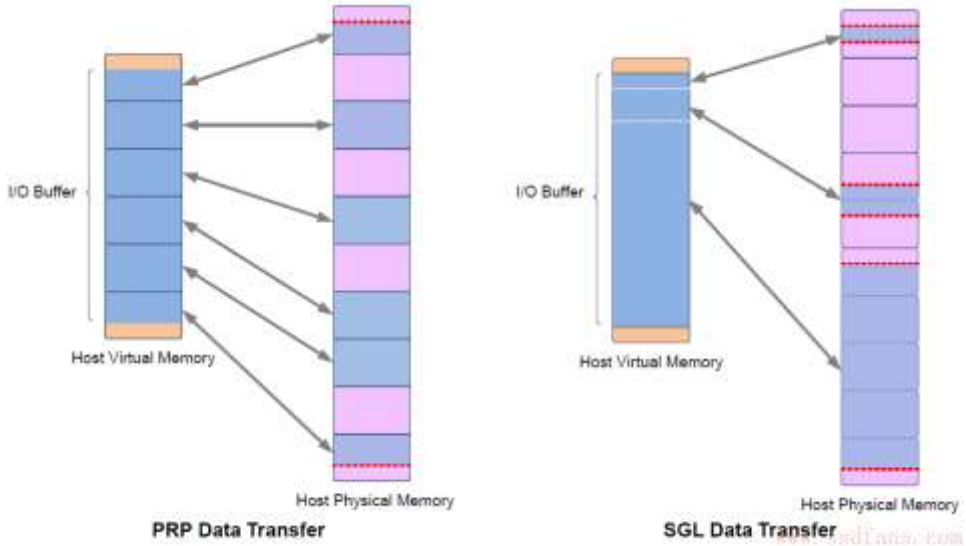
如果还是晕，看个例子吧。

这个例子中，假设 Host 需要往 SSD 中读取 13KB 的数据，其中真正只需要 11KB 数据，这 11KB 的数据需要放到 3 个大小不同的内存中，分别是：3KB,4KB 和 4KB。



无论是 PRP 还是 SGL，本质都是描述内存中的一段数据空间，这段数据空间在物理上可能连续的，也可能是不连续的。Host 在命令中设置好 PRP 或者 SGL，告诉 SSD 数据源在内存的什么位置，或者从闪存上读取的数据应该放到内存的什么位置。

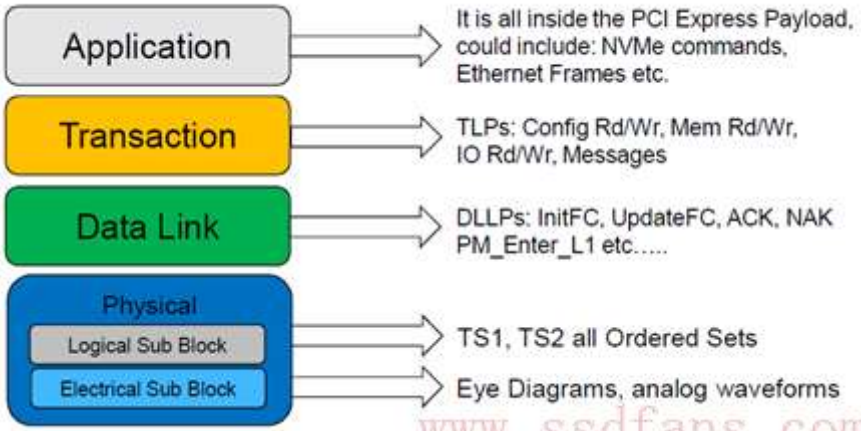
大家也许跟我有个同样的疑问(自作多情？)，那就是，既然有 PRP，为什么还需要 SGL?事实上，NVMe1.0 的时候的确只有 PRP，SGL 是 NVMe1.1 之后引入的。SGL 和 PRP 本质的区别在哪？下图道出了真相：一段数据空间，对 PRP 来说，它只能映射到一个个物理页，而对 SGL 来说，它可以映射到任意大小的连续物理空间。



这章就到这吧。下面《蛋蛋读 NVMe 之四》，蛋蛋会带大家走基层，看看一个 NVMe 读写命令在 PCIe 层是怎样实现的。精彩继续，不要错过。

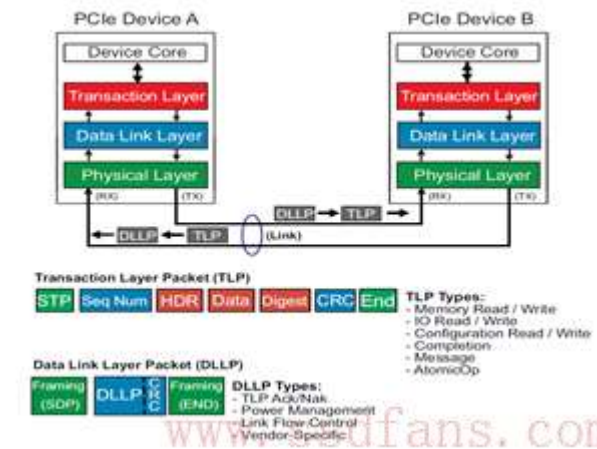
10.3.4 蛋蛋读 NVMe 之四

Posted on 2017 年 8 月 3 日 by SSD Fans



今天我又把这张图搬出来了。没错，它是《蛋蛋读 NVMe 之一》里面的第一张图。任何一种计算机协议，它都是采用这种分层结构的。下层总是为上层服务的。有些协议，上图所有的层次都有定义和实现，而有些协议，只定义了其中的几层。然而，要让一种协议能工作，它需要一个完整的协议栈，PCIe 定义了下三层，NVMe 定义了最上层，两者一拍即合，构成一个完整的 Host 与 SSD 通讯的协议。

PCIe 与 NVMe 最直接接触的是传输层。在 NVMe 层，我们能看到的是 64 字节的命令，16 字节的命令返回状态，以及跟命令相关的数据。而在 PCIe 的传输层，我们能看到的是 TLP （Transaction Layer Packet）。还是跟快递做类比，你要寄东西，可能是手机，可能是电脑，不管是什么，你交给快递小哥，他总是把你要寄的东西打包，快递员看到的就是包裹，他根本不关心你里面的内容。PCIe 传输层作为 NVMe 最直接的服务者，不管你 NVMe 发给我的是命令，还是命令状态，还是用户数据，我统统帮你放进包裹，打包后交给下一层，让数据链路层继续处理。



今天不打算深入讲解 PCIe，这又是一个大的话题。SSD FANS 可能后续会推出类似 NVMe 系列文章，来个 PCIe 系列的，大家可以期待一下。对 PCIe，我们今天只关注传输层，因为它跟 NVMe 接触是最直接最亲密的。PCIe 传输层传输的是 TLP，它就是个包裹，一般由包头和数据组成，当然也有可能只有包头没有数据。NVMe 传下来的数据都是放在 TLP 的数据部分的（Payload）。为实现不同的目的，TLP 可分为以下几种类型：

1. Configuration Read/Write
2. I/O Read/Write
3. Memory Read/write

4. Message
5. Completion

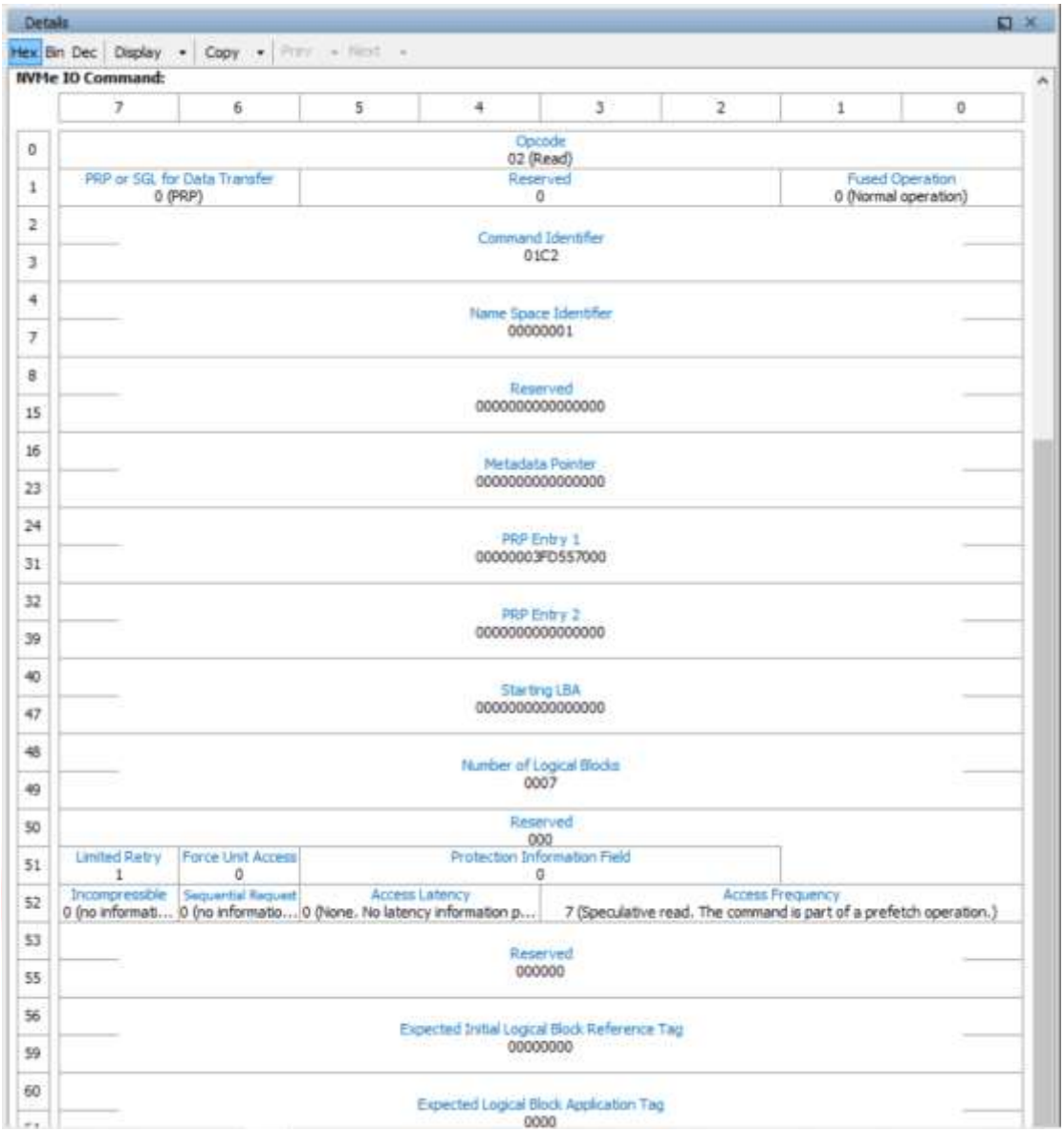
注意，这个 Completion 跟 NVMe 层的 Completion 不是同一个东西，他们处在不同层。在 NVMe 命令处理过程中，PCIe 传输层基本只用 Memory read/write TLP 来为 NVMe 服务，其他 TLP 我们不用管。

Host 发送一个 Read 命令，PCIe 是怎么服务的？今天主要目的，就是结合 NVMe 命令处理流程，蛋蛋带着大家把下面这张图看懂，看看 NVMe 和 PCIe 的传输层发生了什么。

我靠，密密麻麻的，什么鬼东西？别急，蛋蛋带你一步一步把它看懂。

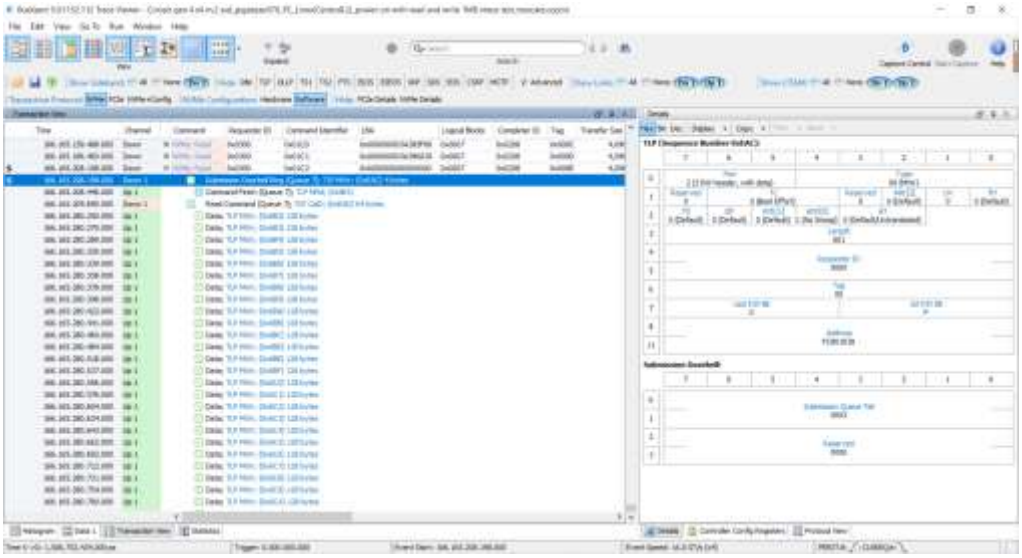
首先，Host 准备了一个 Read 命令给 SSD：

也许你对 NVMe Read 命令格式不是很清楚，说实话，我也不清楚，但从上图，我们还是能得到下面的信息： Host 需要从起始 LBA 0x00000000(LBA)上读取 4096 字节的数据，读到哪里去呢？PRP1 给出内存地址是 0x 3FD557000。这个命令放在编号为 3 的 SQ 里(SQID = 7)，CQ 编号也是 3 (CQID = 7)。我觉得知道这些就够了。相信看了蛋蛋读 NVMe 系列的，刚才说的这些都应该能懂。下面是针对 SSD 从 admin queue 取到的命令的解码：



当 Host 把一个命令准备好放到 SQ 后，接下来步骤是什么呢？回想一下 NVMe 命令处理的八个步骤。

第二步就是：**Host 通过写 SQ 的 Tail DB，通知 SSD 来取命令**。

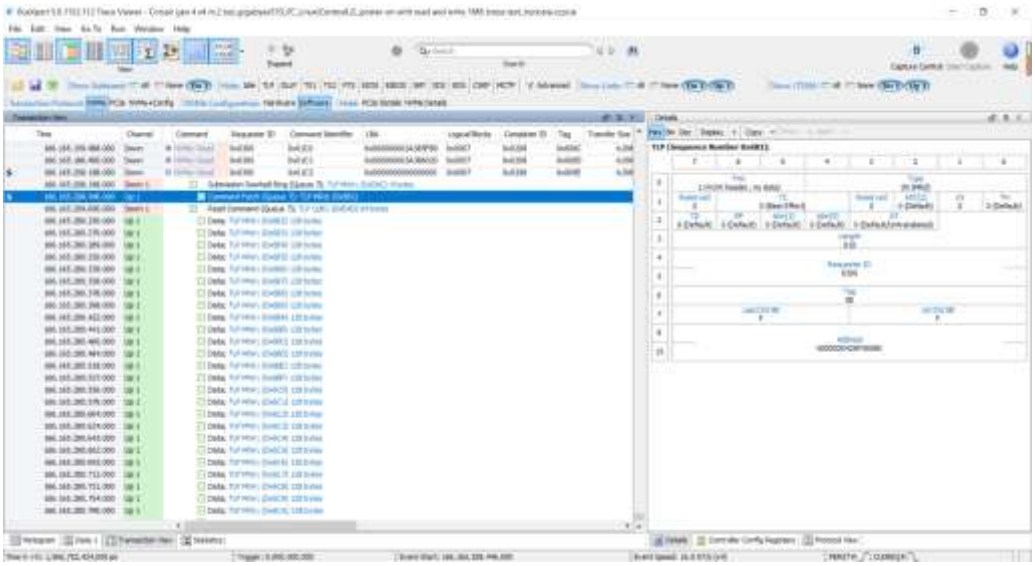


上图中，上层是 NVMe 层，下层是 PCIe 的传输层，这一层我们看到的是 TLP。Host 想往

SQ Tail DB 中写入的值是 3。PCIe 是通过一个 **Memory Write TLP** 来实现 Host 写 CQ 的 Tail DB 的。

一个 Host，下面可能连接着若干个 Endpoint，该 SSD 只是其中的一个 Endpoint 而已，那有个问题，Host 怎么能准确更新该 SSD Controller 中的 Tail DB 寄存器呢？怎么寻址？其实，在上电的过程中，每个 Endpoint 的内部空间都会通过内存映射(memory map)的方式映射到 Host 的内存中，SSD Controller 当中的寄存器会被映射到 Host 的内存，当然也包括 Tail DB 寄存器。Host 在用 Memory Write 写的时候，Address 只需设置该寄存器在 Host 内存中映射的地址，就能准确写入到该寄存器。以上图为例，该 Tail DB 寄存器应该映射在 Host 内存地址 FC801038，所以 Host 写 DB，只需指定这个物理地址，就能准确无误的写入到对应的寄存器中去。应该注意的是：Host 并不是往自己内存的那个物理地址写入东西，而是用那个物理地址作为寻址用，往 SSD 方向写。否则就太神奇了，往自己内存写东西就能改变 SSD 中的寄存器值，那不是量子效应吗？我们的东西还没有那么玄乎。

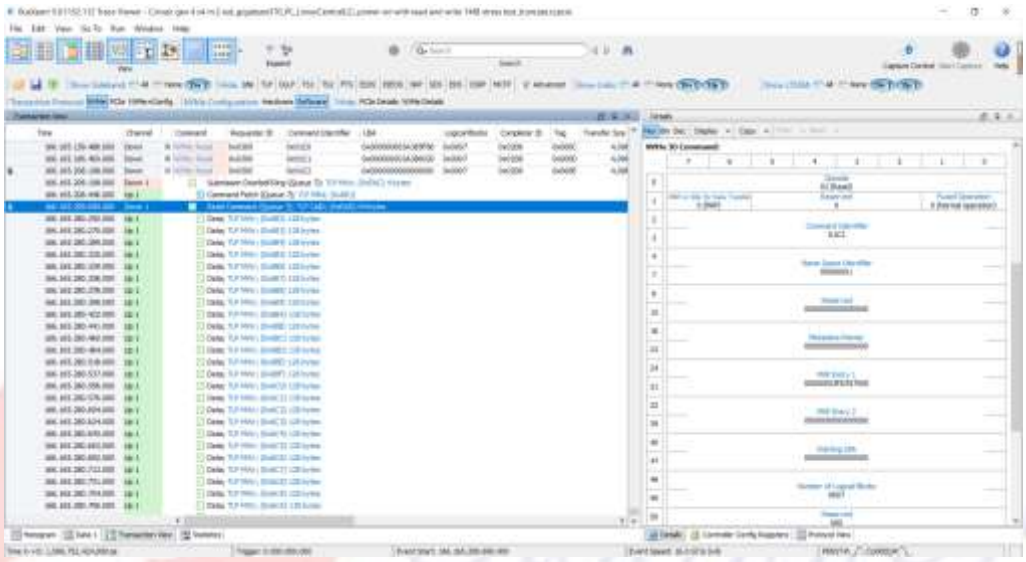
NVMe 处理命令的第三步：**SSD 收到通知，去 Host 端的 SQ 中取指**。



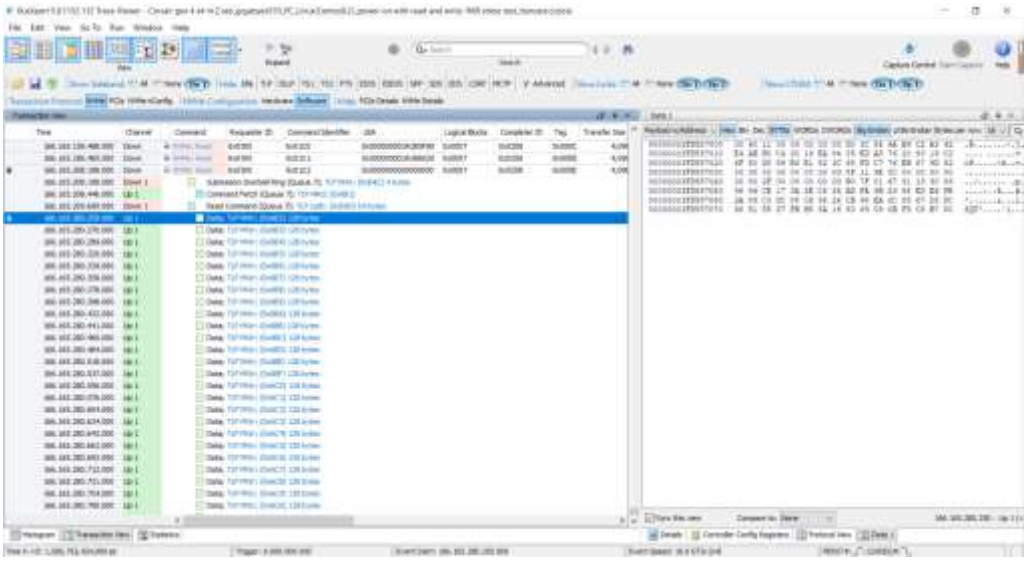
PCIe 是通过发一个 **Memory Read TLP** 到 Host 的 SQ 中取指的。可以看到，PCIe 需要往 Host 内存中读取 16 个 DWORD 的数据。为什么是 16 DWORD 数据，因为每个 NVMe 命

令的大小是 64 个字节。从上图中，我们可以推断 SQ 3 当前的 Head 指向的内存地址是 0x42AF50080？怎么推断来的？因为 SSD 总是从 Host 的 SQ 的 Head 取指的，而上图中，Address 就是 0x42AF50080，所以我们有此推断。

在上图中，SSD 往 Host 发送了一个 Memory Read 的请求，Host 通过 Completion 的方式把命令数据返回给 SSD。和前面的 Memory Write 不同，Memory Read 中是不含数据，只是个请求，数据的传输需要对方发个 Completion。像这种需要对方返回状态的 TLP 请求，我们叫它 Non-Posted 请求。怎么理解呢？Post，有"邮政"的意思，就像你寄信一样，你往邮箱中一扔，对方能不能收到，就看快递员的素养了，反正你是把信发出去了。像 Memory Write 这种，就是 Posted 请求，数据传给对方，至于对方有没有处理，我们不在乎；而像 Memory Read 这种请求，它就必须是 Non-Posted 了，因为如果对方不响应（不返回数据）给我，Memory Read 就是失败的。所以，每个 Memory read 请求都有相应的 Completion。



NVMe 处理命令的第四步：**SSD 执行读命令，把数据从闪存中读到缓存中，然后把数据传给 Host**。数据从闪存中读到缓存中，这个是 SSD 内部的操作，跟 PCIe 和 NVMe 没有任何关系，因此，我们捕捉不到 SSD 的这个行为。我们在 PCIe 接口上，我们只能捕捉到 SSD 把数据传给 Host 的过程。

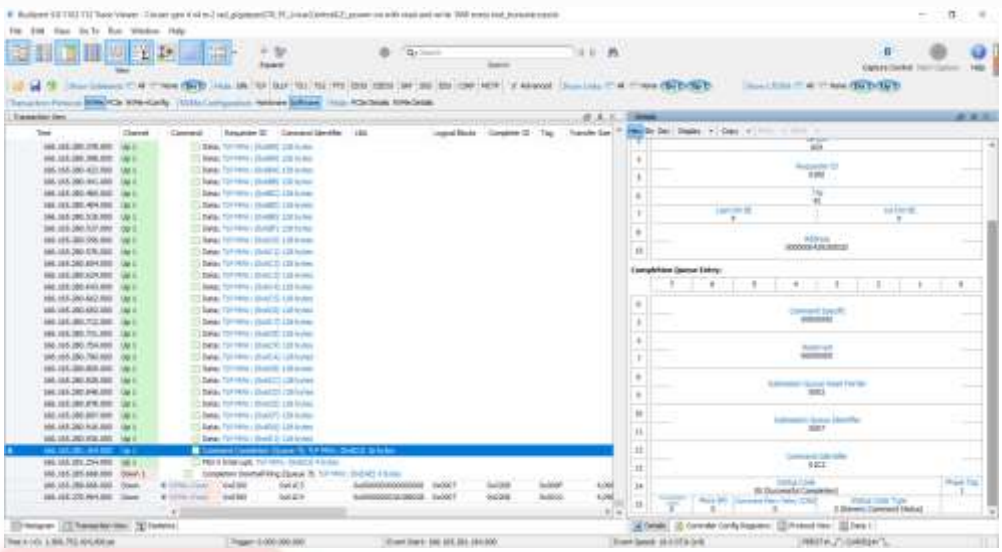


从上图中可以看出，SSD 是通过 **Memory write TLP** 把 Host 命令所需的 128 个 DWORD 数据写入到 Host 命令所要求的内存中去。SSD 每次写入 32 个 DWORD（128 Byte），一共

写了 32 次。同时，从上图右侧的解码的地址空间我们看到，第一个数据包确实是写在地址空间 0x3FD557000 起始的地方。

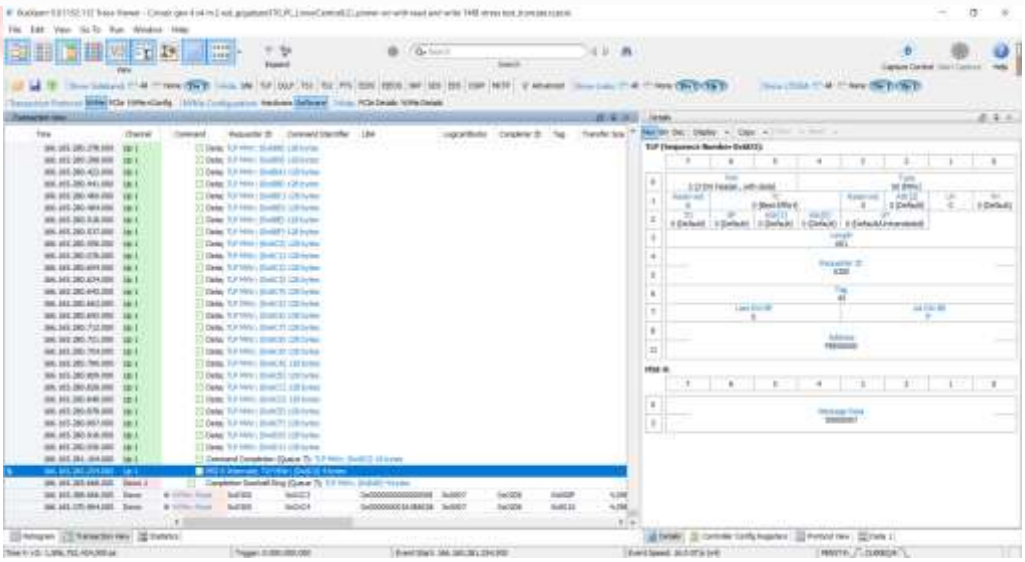
另外，正如之前所说，我们没有看到 Completion，合理。

SSD 一旦把数据返回给 Host，SSD 认为命令以及处理完毕，第五步就是：**SSD 往 Host 的 CQ 中返回状态**。



从上图中可以看出，SSD 是通过 **Memory write TLP** 把 16 个字节的命令完成状态信息写入到 Host 的 CQ 中。

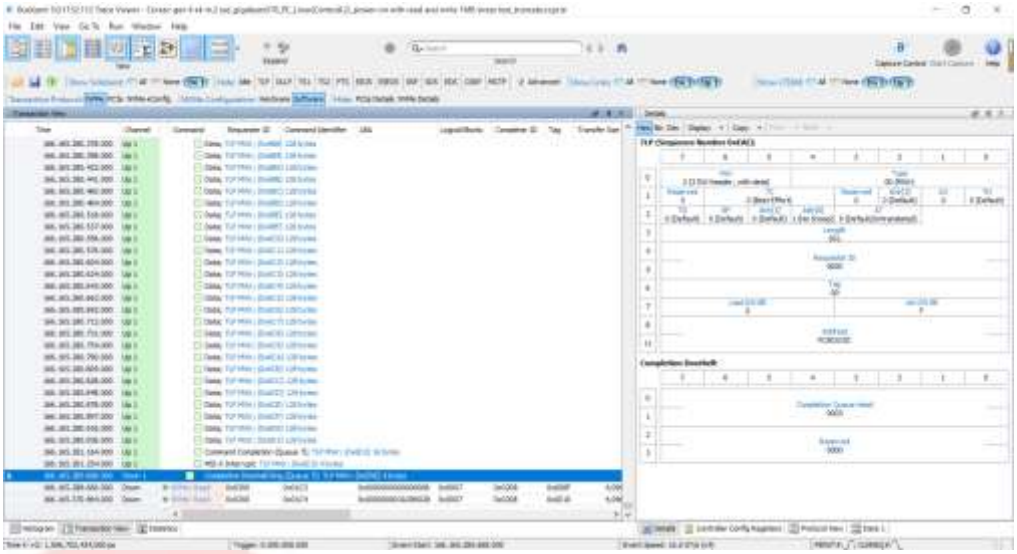
SSD 往 Host 的 CQ 中写入后，第六步就是：**SSD 采用中断的方式告诉 Host 去处理 CQ**。



SSD 中断 Host，NVMe/PCIe 有四种方式：Pin-based interrupt, single message MSI, multiple message MSI 和 MSI-X。关于中断，具体的可以参看 spec 第 171 页，有详细介绍，有兴趣的可以去看看。从上图中，这个例子中使用的是 MSI-X 中断方式。跟传统的中断不一样，它不是通过硬件引脚的方式，而是把中断信息和正常的数据信息一样，PCIe 打包把中断信息告知 Host。上图告诉我们，SSD 还是通过 **Memory Write TLP** 把中断信息告知 Host，这个中断信息长度是 1DWORD。

Host 收到中断后，第七步就是：**Host 处理相应的 CQ**。这步是在 Host 端内部发生的事情，在 PCIe 线上我们捕捉不到这个处理过程。

最后一步，Host 处理完相应的 CQ 后，**需要更新 SSD 端的 CQ Head DB,告知 SSD CQ 处理完毕**。



跟前面一样，Host 还是通过 **Memory Write TLP** 更新 SSD 端的 CQ Head DB。

从我们抓的 PCIe trace 上，我们从 PCIe 的传输层看到了一个 NVMe Read 命令是怎么处理的，看到传输层基本都是通过 Memory Write 和 Memory Read TLP 传输 NVMe 命令、数据和状态等信息；我们确实也看到了 NVMe 命令处理的八个步骤，蛋蛋没有欺骗大家。

上面举的是 NVMe 读命令处理，其他命令处理过程其实差不多，就不凑篇幅了。

最后，我再贴出完整 Trace，相信，也希望大家不会再有一团乱麻的感觉。

下面的图是针对上图的每个环节的解码的汇总。

注意：我们去掉了占用空间较多（32 个 TLP MWr 的 TLP 报文）的 Data 部分的解码，这样主要是为了看的清楚并且便于理解前面 NVMe Read 的执行的关键环节。

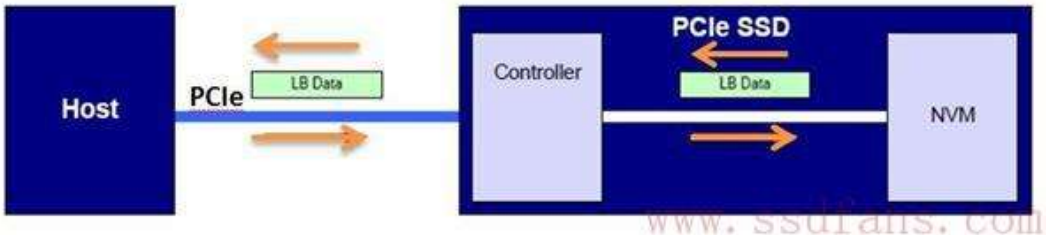
10.3.5 蛋蛋读 NVMe 之五

Posted on 2017 年 8 月 3 日 by SSD Fans

没错，这是李连杰在《中南海保镖》中的一张剧照。剧情已经很模糊，准备晚上回去怀旧一下。今天用这张图开始，是因为接下来，我们要说的话题就是 NVMe 中端到端数据保护功能，看看 NVMe 中的保镖是怎样为我们的数据保驾护航的。

我们需要保护的是数据。Host 与 SSD 之间，数据传输的最小单元是逻辑块（Logical Block,LB），每个逻辑块大小可以是 512/520/1024/2048/4096 字节等，Host 在格式化 SSD 的时候，逻辑块大小就确定了，以后两者就按这个逻辑块大小进行数据交互。

数据从 Host 到 NVM（Non-Volatile Memory，目前一般是闪存，后面我就用闪存来代表 NVM），首先要经过 PCIe 传输到 SSD 的 Controller，然后 Controller 把数据写入到闪存；反过来，Host 想从闪存上读取数据，首先 SSD Controller 从闪存上获得数据，然后经过 PCIe 把数据传送给 Host。



Host 与 SSD 之间，数据在 PCIe 上传输的时候，由于信道噪声的存在（说白了就是存在干扰），可能导致数据出错；另外，在 SSD 内部，Controller 与闪存之间，数据也可能发生错误。路途凶险。为确保 Host 与闪存之间数据的完整性，即 Host 写入到闪存的数据与最初 Host 写的数据一致，以及 Host 读到的数据与最初从闪存上读上来的数据一致，NVMe 提供了一个端到端数据保护功能。

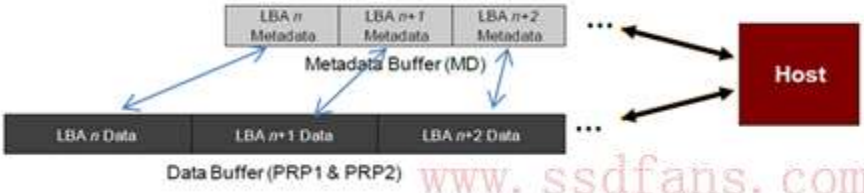
除了逻辑块数据本身，NVMe 还允许每个逻辑块带个助理，叫做元数据（Meta Data）。这个助理的职责，NVMe 虽然没有明确的要求，但如果数据需要保护，NVMe 要求这个助理必须能充当保镖的角色。

元数据有两种存在方式，一种是作为逻辑块数据的扩展，和逻辑块数据放一起存放，这是贴身保镖：

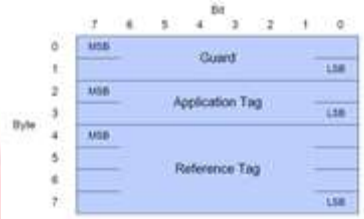
Figure 210: Metadata – Contiguous with LBA Data, Forming Extended LBA

另外一种方式就是逻辑块数据放在一起，元数据单独放在别处。虽不是贴身保护，但保镖在附近时刻注意着主人的安全，属非贴身保镖：

Figure 211: Metadata – Transferred as Separate Buffer

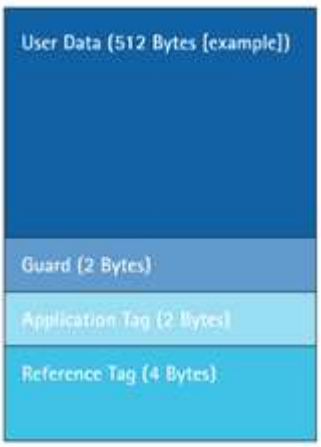


贴身保护与否，我们不关心形式，我们只关心元数据是如何保护逻辑块数据的。NVMe 要求每个逻辑块数据的保镖配备下面这把武器:



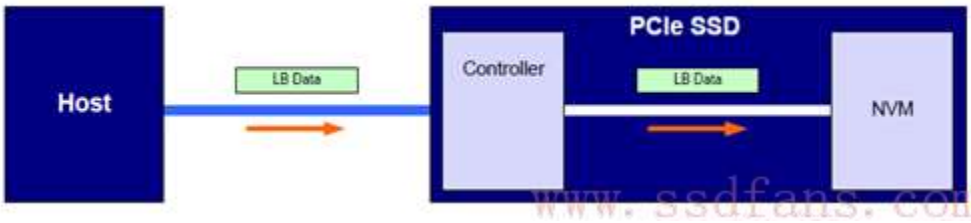
其中的”Guard”是 16 比特的 CRC （Cyclic Redundancy Check），它是逻辑块数据算出来的；”Application Tag”和”Reference Tag”包含该数据块的逻辑地址（LBA）等信息。CRC 校验能够检测出数据是否有错，后者则是保证数据不会出现张冠李戴的问题，比如我 LBA X 使用了 LBA Y 的数据，这种情况往往是 SSD 固件 Bug 导致的。Anyway，NVMe 能帮你发现这个问题。

佩了保镖的数据看起来就是下面这个样子（以 512 字节的数据块为例）：



在 Host 与 SSD 数据传输过程中，NVMe 可以让每个逻辑块数据都带上保镖，也可以让他们不带保镖，也可以在某个治安差的地方把保镖带上，然后在治安环境好的地方不用保镖。

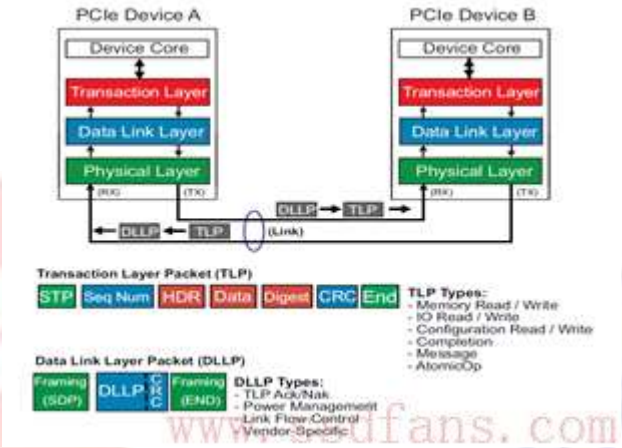
Host 往 SSD 写入数据，不带保镖：



什么情况下可以不带保镖？

如果你普通人一个，完全没有必要配保镖，原因有：1. 你请不起保镖；2. 谁有空来伤害你呢？3. 太平盛世。

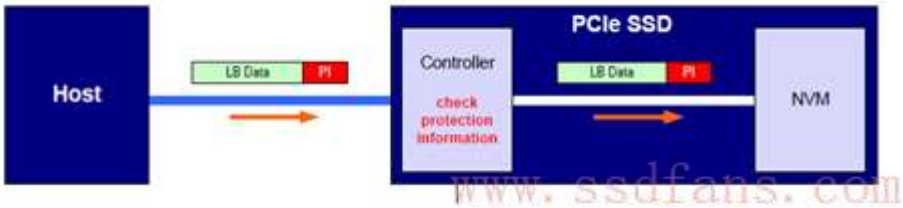
如果是无关紧要的数据（如小电影），完全没有必要进行端到端的保护，毕竟数据保护需要传输额外的数据 （每个逻辑数据块需要至少额外 8 字节的数据保护信息，有效带宽减少），还需要 SSD 做额外的数据完整性校验（耗时，性能变差），最关键的是 PCIe 通道上，其数据天然就能受到保护。怎么说？



对每个 TLP 来说，其中有个 Digest 域，就是对 HDR 和 Data 进行数据保护的，本质就是 CRC。这个 Digest 是可选的。如果使能了 Digest，数据在 PCIe 上传输是毫无风险的，因为有便衣警察保护，在 NVMe 层完全没有必要进行额外的数据保护。

当然，它不能发现数据张冠李戴的问题。

Host 往 SSD 写入数据，全程带上保镖的情况：



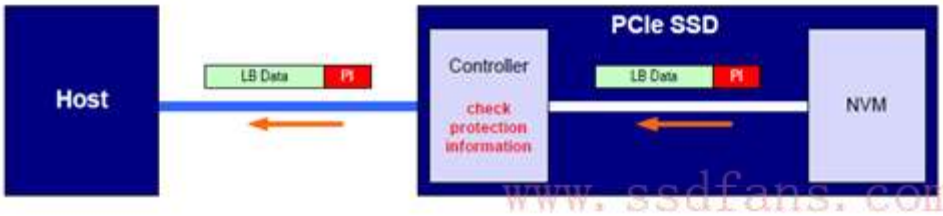
红色 **PI**，Protection Information，就是传说中的保镖。NVMe 居然给数据配这么一个大红显眼的保镖，我也算是服了。

Host 数据通过 PCIe 传输到 SSD Controller 之间，按理来说数据已经受到 PCIe 的保护，但 PCIe 保镖也有可能不在情况，那就是 TLP 中 Digest 域可能不存在，这是 PCIe 允许的。这个时候，如果要保证在 PCIe 上数据传输的可靠性，就需要 NVMe 自带保镖。数据到达 SSD Controller 时，SSD Controller 会重新计算逻辑块数据的 CRC，与保镖的 CRC 比较，如

果两者匹配，说明数据传输是没有问题的；否则，数据就是有问题的，这个时候，SSD Controller 就会给 Host 报错。

除了 CRC 校验，还要检测有没有张冠李戴的问题，通过检测 Reference Tag 和 Application Tag，看看这个没有 CRC 问题的数据是不是该笔 Host 写命令对应的数据，如果不匹配，同样需要向 Host 报错。

如果数据检测没有问题，SSD Controller 会把逻辑块数据和 PI 一同写入闪存中。这个 PI 一同写入到闪存中有什么意义呢？在读的时候有意义。

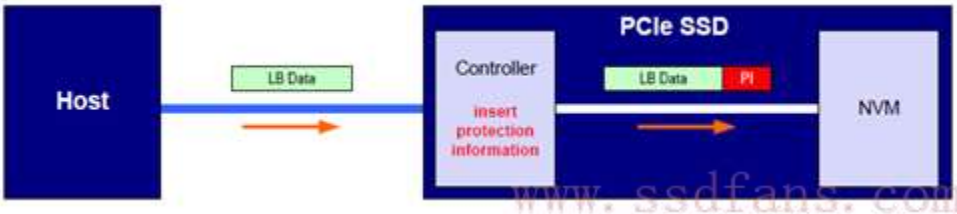


SSD Controller 读闪存的时候，会对读上来的数据进行 CRC 校验，如果写入的时候带有 PI，这个时候就能检测出读上来的数据是否正确，从而决定这个数据要不要传给 Host。有人要说，对闪存来说，数据不是受 ECC 保护吗？为什么还要额外进行数据校验？没错，写入到闪存中的数据是受 ECC 保护，这个没有问题，但在 SSD 内部，数据从 Controller 到闪存之间，一般都要经过 DRAM 或者 SRAM，在之前 SSD Controller 写入到闪存，或者这个时候从闪存读数据到 SSD Controller，可能就会发生比特翻转之类的小概率事件，从而导致数据不正确。如果在 NVMe 层再做个 CRC 保护，这类数据错误就能被发现了。

除了数据在 SSD 内发生反转，由于固件问题，或者别的原因，还是会出现数据张冠李戴的问题：数据虽然没有 CRC 错误，但是它不是我们想要的数据。因此，还需要做 Reference Tag 和 Application Tag 检测。

SSD Controller 通过 PCIe 把数据传给 Host，Host 端也会对数据进行校验，看 SSD 返回过来的数据是否有错。

Host 往 SSD 写入数据，半程带保镖的情况：



这种情况，Host 与 Controller 端之间是没有数据保护，因为 PCIe 已经能提供数据完整性保证了（TLP 中的 Digest 使能）。但在 SSD 内部，Controller 到闪存之间，由于乱七八糟的原因（数据反转，LBA 数据不匹配），存在数据错误的可能，NVMe 要求 SSD Controller 在把数据写入到闪存前，计算好数据的 PI，然后把数据和 PI 一同写入到闪存。

SSD Controller 读闪存的时候，会对读上来的数据进行 PI 校验，如果没有问题，剥除 PI，然后把逻辑块数据返回给 Host；如果校验失败，说明数据存在问题，SSD 需要向 Host 报错。如下图所示：

数据端到端保护是 NVMe 的一个特色，其本质就是在数据块当中加入 CRC 和数据块对应的 LBA 等冗余信息，SSD Controller 或者 Host 端利用这个这些信息进行数据校验，然后根据校验结果执行相应的操作。加入这些检错信息的好处是能让 Host 与 SSD Controller 及时发现数据错误，副作用就是：

1. 每个数据块需要额外的至少 8 字节的数据保护信息，有效带宽减少：数据块大小越小，带宽影响越大。
2. SSD Controller 需要做数据校验，影响性能。

但是，我觉得这二个副作用的影响是微乎其微的，跟数据安全性相比，这又算得了什么呢？

10.3.6 蛋蛋读 NVMe 之六

Posted on 2017 年 8 月 3 日 by SSD Fans

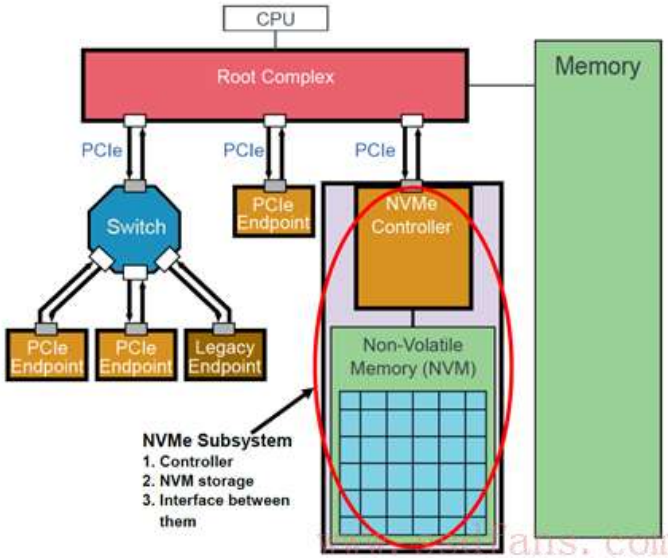
原创内容，转载请注明： [http://www.ssdfans.com] 谢谢！

前几天某个晚上，山哥给我发微信说：NVMe Namespace 不打算写篇文章吗？

我笑说，没有人关注，懒得写了。

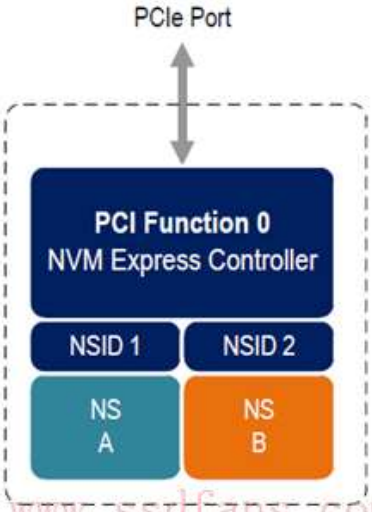
其实这只是一方面，主要的是（jiang）没有（lang）找到话题（cai）写下去（jin）。山哥这么一说，倒是提醒了我，可以写写 Namespace，毕竟这个东西我们在 SATA 上是看不到的。

那么什么是 Namespace （以下简称 NS，不打算翻译成中文）？



上图中红圈圈起来的是一个 NVMe 子系统，通常来说就是 SSD。一个 NVMe SSD 主要由 SSD Controller，闪存空间和 PCIe 接口组成。如果把闪存空间划分成若干个独立的逻辑空间，每个空间逻辑块地址（LBA）范围是 0 到 N-1 (N 是逻辑空间大小)，这样划分出来的每一个逻辑空间我们就叫做 NS。对 SATA SSD 来说，一个闪存空间只对应着一个逻辑空间，与之不同的是，NVMe SSD 可以是一个闪存空间对应多个逻辑空间。

每个 NS 都有一个名称与 ID，如同每个人都有名字和身份证号码，ID 是独一无二的，系统就是通过 NS 的 ID 来区分不同的 NS。



如上图例子，整个闪存空间划分成 2 个 NS，名字分别是 NS A 和 NS B，对应的 NS ID 分别是 1 和 2。如果 NS A 大小是 M （以逻辑块大小为单位），NS B 大小是 N，则他们的逻辑地址空间分别是 0 到 M-1 和 0 到 N-1。Host 读写 SSD，都是要在命令中指定读写的是哪个 NS 中的逻辑块。原因很简单，如果不指定 NS，对同一个 LBA 来说，假设就是 LBA 0，SSD 根本就不知道去读或者写哪里，因为有两个逻辑空间，每个逻辑空间都有 LBA 0。如同我只说德州，如果不告诉你是哪个国家的，你怎知道我说的是美国德州还是山东德州。

一个 NVMe 命令一共 64 字节，其中第 4 到第 7 个 Byte 指定了要访问的 NS。

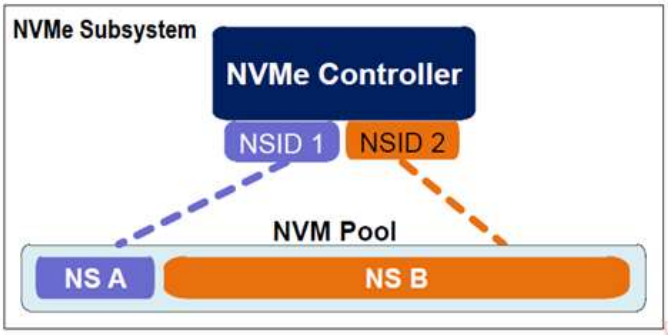
Bytes	Description
63:60	**Command Dword 15 (CDW15):** This field is command specific Dword 15.
59:56	**Command Dword 14 (CDW14):** This field is command specific Dword 14.
55:52	**Command Dword 13 (CDW13):** This field is command specific Dword 13.
51:48	**Command Dword 12 (CDW12):** This field is command specific Dword 12.
47:44	**Command Dword 11 (CDW11):** This field is command specific Dword 11.
43:40	**Command Dword 10 (CDW10):** This field is command specific Dword 10.
39:32	**PRP Entry 2 (PRP2):** This field: a) is reserved if the data transfer does not cross a memory page boundary. b) specifies the Page Base Address of the second memory page if the data transfer crosses exactly one memory page boundary. E.g.,: i. the command data transfer length is equal in size to one memory page and the offset portion of the PBAO field of PRP1 is non-zero or ii. the Offset portion of the PBAO field of PRP1 is equal to zero and the command data transfer length is greater than one memory page and less than or equal to two memory pages in size. c) is a PRP List pointer if the data transfer crosses more than one memory page boundary. E.g.,: i. the command data transfer length is greater than or equal to two memory pages in size but the offset portion of the PBAO field of PRP1 is non-zero or ii. the command data transfer length is equal in size to more than two memory pages and the Offset portion of the PBAO field of PRP1 is equal to zero.
31:24	**PRP Entry 1 (PRP1):** This field contains the first PRP entry for the command or a PRP List pointer depending on the command.
23:16	**Metadata Pointer (MPTR):** This field contains the address of a contiguous physical buffer of metadata. This field is only used if metadata is not interleaved with the logical block data, as specified in the Format NVM command. This field shall be Dword aligned.
15:08	Reserved
07:04	**Namespace Identifier (NSID):** This field specifies the namespace ID that this command applies to. If the namespace ID is not used for the command, then this field shall be cleared to 0h. If a command shall be applied to all namespaces accessible by this controller, then this field shall be set to FFFFFFFFh. Unless otherwise noted, specifying an inactive namespace ID in a command that uses the namespace ID shall cause the controller to abort the command with status Invalid Field in Command. Specifying an invalid namespace ID in a command that uses the namespace ID shall cause the controller to abort the command with status Invalid Namespace or Format.
03:00	**Command Dword 0 (CDW0):** This field is common to all commands and is defined in Figure 10.

对每个 NS 来说，都有一个 4KB 大小的数据结构来描述它。

Bytes	Man/Opt	Description
7:0	M	Namespace Size
15:8	M	Namespace Capacity
23:16	M	Namespace Utilization
24	M	Namespace Features
25	M	Number of LBA Formats
26	M	Formatted LBA Size
27	M	Metadata Capabilities
28	M	End-to-end Data Protection Capability
29	M	End-to-end Data Protection Type Settings
30	O	Namespace Multi-path I/O and Namespace Sharing Capabilities
31	O	Reservation Capabilities
119:32	-	Reserved
127:120	M	IEEE Extended Unique Identifier
131:128	M	LBA Format 0 Supported
191:132	O	LBA Formats 1 – 15 Supported
383:192	-	Reserved
4095:384	O	**Vendor Specific (VS):** This range of bytes is allocated for vendor specific usage.

该数据结构描述了该 NS 的大小，整个空间已经写了多少，每个 LBA 的大小，以及端到端数据保护相关设置，该 NS 是否属于某个 Controller 还是几个 Controller 可以共享，等等。

NS 由 Host 创建和管理，每个创建好的 NS，从 Host 操作系统角度看来，就是一个独立的磁盘，用户可在每个 NS 做分区等操作。



This example:
OS sees two drives
- NS A = Disk 0
- NS B = Disk 1
- Logical partitions on A and B

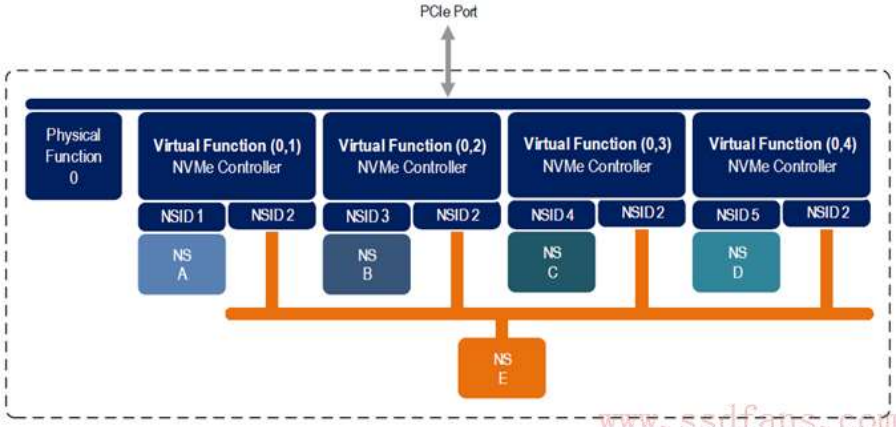
上例中，整个闪存空间划分成两个 NS，NS A 和 NS B，操作系统看到两个完全独立的磁盘。我的天呀，太神奇了，我买一个 SSD，居然得到两个磁盘，赚大发了。

每个 NS 是独立的，逻辑块大小可以不同，端到端数据保护配置也可以不同：你可以让一个 NS 使用保镖，另一个 NS 不使用保镖，再一个 NS 半程使用保镖（见《蛋蛋读 NVMe 之五》）。这样我就在想，是不是可以把我的 SSD 划分成两个 NS: 一个 NS 使用数据端到端保护，上面存放操作系统、软件和其他重要数据，另外一个 NS 不使用端到端数据保护，上面只存放小电影之类的数据。

其实，NS 更多的是应用在企业级，可以根据客户不同需求创建不同特征的 NS，也就是在一个 SSD 上创建出若干个不同功能特征的磁盘（NS）供不同客户使用。

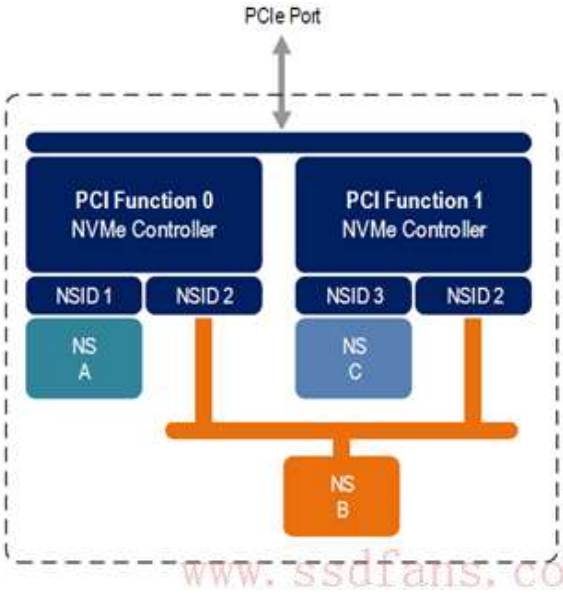
NS 的另外一个重要使用场合是：SR-IOV。

什么是 SR-IOV？ 英文全称为 Single Root- I/O Virtualization，SR-IOV 技术允许在虚拟机之间高效共享 PCIe 设备，并且它是在硬件中实现的，可以获得能够与本机性能媲美的 I/O 性能。单个 I/O 资源（单个 SSD）可由许多虚拟机共享。共享的设备将提供专用的资源，并且还使用共享的通用资源。这样，每个虚拟机都可访问唯一的资源。



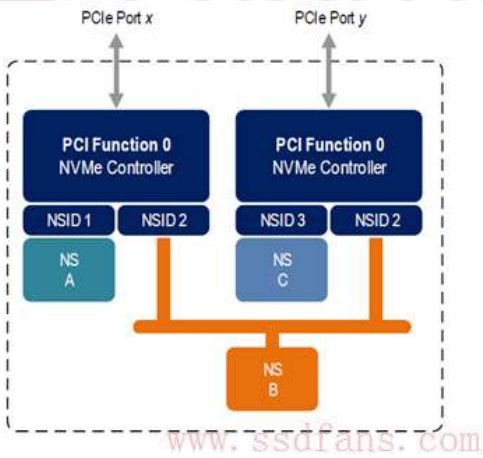
如上图所示，该 SSD 作为 PCIe 的一个 Endpoint，实现了一个物理功能 (Physical Function ,PF)，有 4 个虚拟功能（Virtual Function，VF）关联该 PF。每个 VF，都有自己独享的 NS，还有公共的 NS （NS E)。此功能使得虚拟功能可以共享物理设备，并在没有 CPU 和虚拟机管理程序软件开销的情况下执行 I/O。关于 SR-IOV 更多知识，请自行百度或者谷歌。这里我们只需知道 NVMe 中的 NS 有用武之地就可以。

对一个 NVMe 子系统来说，除了包含若干个 NS，还可以由若干个 SSD Controller。注意，这里不是说一个 SSD Controller 有多个 CPU，而是说一个 SSD 有几个实现了 NVMe 功能的 Controller。



如上图例子，一个 NVMe 子系统包含了两个 Controller，分别实现不同功能（也可以是相同功能）。整个闪存空间分成 3 个 NS，其中 NS A 由 Controller 0（左边）独享，NS C 由 Controller 1（右边）独享，而 NS B 是两者共享。独享的意思是说只有与之关联的 Controller 才能访问该 NS，别的 Controller 是不能对之访问的，上图中 Controller 0 是不能对 NS C 进行读写操作的，同样，Controller 1 也不能访问 NS A；共享的意思是说，该 NS（这里是 NS B）是可以被两个 Controller 共同 访问的。对共享 NS，由于几个 Controller 都可以对它进行访问，所以要求每个 Controller 对该 NS 的访问都是原子操作，从而避免同步问题。

事实上，一个 NVMe 子系统，除了可以有若干个 NS，除了可以有若干个 Controller，还可以有若干个 PCIe 接口。



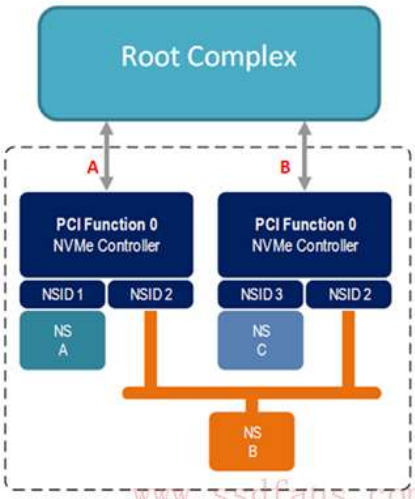
与前面的架构不一样，上图的架构是每一个 Controller 有自己的 PCIe 接口，而不是两者共享一个。Dual Port，哈哈，在 SATA SSD 上没有见过吧。这两个接口，往上有可能连着同一个主机，也可能连着不同的主机。现在能提供 Dual PCIe Port 的 SSD 接口只有 SFF-8639（关于这个接口，可参看站内文章《SFF-8639 接口来袭》），也叫 U.2，它支持标准的 NVMe 协议和 Dual-Port，号称 SSD 接口明日之星。

硬盘主要接口及特点

http://www.expreview.com (last update: 01/04/2016)

	SATA III	mSATA	SATA Express	M.2	U.2/SFF-8639	PCI-E（HHHL）
速度	6Gbps	6Gbps	10/16Gbps	10/32Gbps	32Gbps	20/32Gbps
规格/长度	2.5/3.5寸	51mm	2.5/3.5寸	30-110mm	2.5寸	167mm
界面	SATA	SATA	PCI-E x2	PCI-E x2、x4 SATA	PCI-E x2、x4 SATA	PCI-E x2、x4
工作电压	5V	3.3V	5V	3.3V	3.3V/12V	12V
体积	大	小	大	小	大	大
备注	绝对主流但已落伍	基本淘汰	尴尬之选	今日之星	不确定的明日之星	今日之星

下图是两个 PCIe 接口连着一个主机的情况：

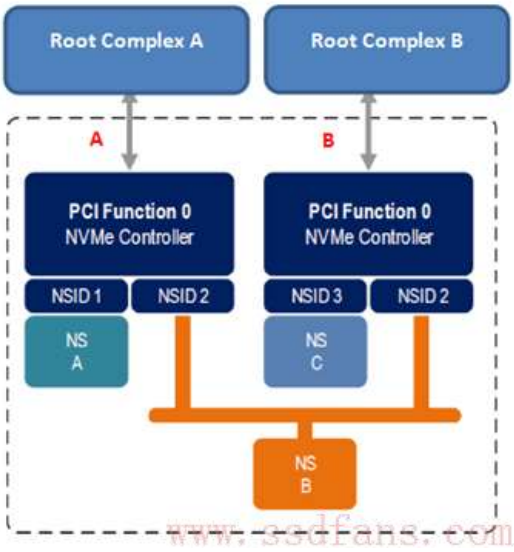


为什么要这么玩？

我认为，一方面，Host 访问 SSD，可以双管齐下，性能可能更好点。不过对访问 NS B 来说，同一时刻只能被一个 Controller 访问，双管齐下又如何。考虑到还可以同时操作 NS A 和 NS C，性能或多或少的有所提升。

我觉得，更重要的是，这种双接口冗余设计，可以提升系统可靠性。假设 PCIe A 接口出现问题，这个时候 Host 可以通过 PCIe B 无缝衔接，继续对 NS B 进行访问。当然了，NS A 是无法访问了。

如果 Host 突然死机怎么办？据小道消息，阿法狗输给李世石那盘，就是阿法狗死机了，然后重启再战，结果超时认输。哈哈，开个玩笑。在一些很苛刻的场景下，是不允许 Host 宕机的。但是，是电脑总有死机的时候，怎么办？最直接有效的办法还是采用冗余容错策略：SSD 有两个 Controller，有两个 PCIe 接口，那么我主机也弄个双主机：一个主机挂了，另一个主机接管任务，继续执行，你就慢慢重启吧。

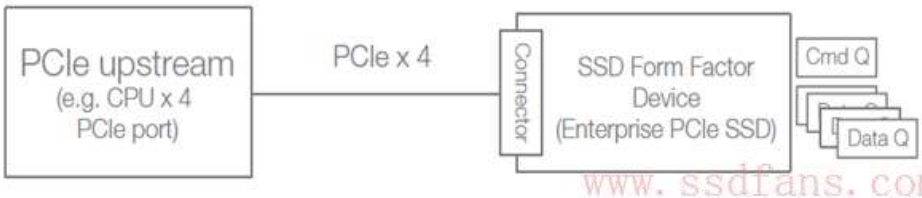


我们来看一个 Dual Port 的真实产品。

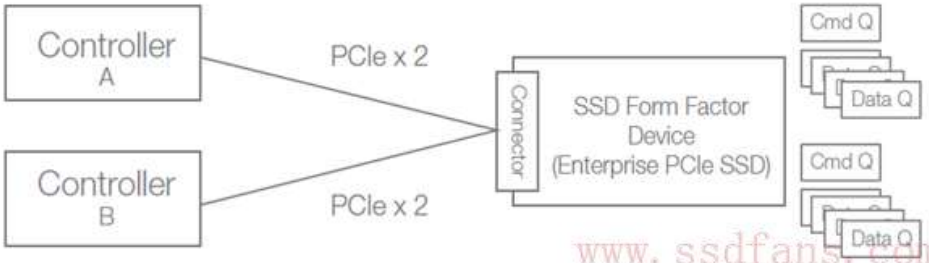
2015 年，OCZ 发布了业界第一个具有 Dual Port 的 PCIe NVMe 的 SSD： Z-Drive 6000 系列。

物理上，这些 SSD 都有两个 PCIe Port，但可以通过不同的固件，实现 Single Port 和 Dual Port 功能。

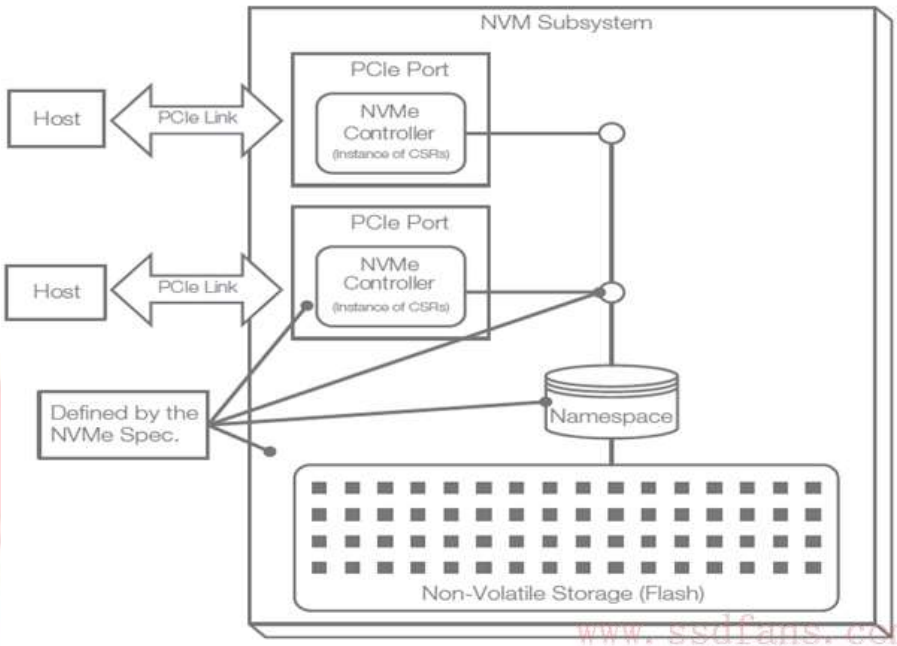
如果只用一个 Port，那么它就是一个 4 通道的 PCIe 接口，向上连接一个主机：



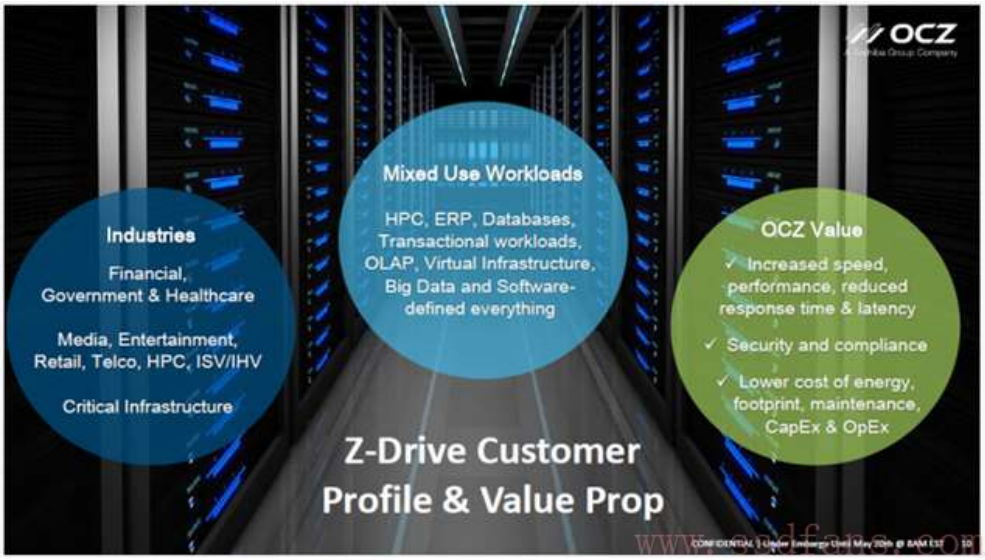
如果使能 Dual Port，那么可以配置成 2 个 2 通道的 PCIe 接口，即每个 Port 有两个通道。



具体来看，整个系统就是这个样子：



每个 Port 可以连接两个独立的 Host，Host 有两个独立的数据通道（Data Path）对闪存空间进行访问，如果其中一个数据通道发生故障，OCZ 的 Host 热交换（Hot-swap ）技术能让另外一个 Host 无缝低延时的接管任务。有些应用，比如银行金融系统、在线交易处理（OnLine Transaction Processing，OLTP)、在线分析处理（OnLine Analytical Processing，OLAP)、高性能计算（High Performance Computing ，HPC),大数据等，对系统可靠性和实时性要求非常高，这个时候，带有 Dual Port 的 SSD 就能派上用场了。



带有 Dual Port 的这种 SSD，主要是面向企业用户，特别是上面提到的那些应用行业。对我们普通用户来说，我感觉使用 Dual Port 就没有这个必要了。

多 NS，多 Controller，多 PCIe 接口，给 NVMe SSD 开发者，以及存储架构师很大的发挥空间。给不同的 NS 配置不同的数据保护机制，或者虚拟化，或者使用冗余容错提高系统可靠性，抑或别的设计，NVMe 提供了这些基础设施，怎么玩就看你的想象力了。

文章最后，我要特别感谢南山：首先，南山建议我写个 NS 相关的话题，才有了此文的开始；其次，本文定稿前，南山帮把关，提出了很多建议，和提供相关资料。如果没有南山的支持，本文就可能只是泛泛而谈，大家就看不到这么精彩的文章了😊。

10.4 阿呆实战 NVMe 系列

Posted on 2016 年 7 月 1 日 by SSD Fish

10.4.1 阿呆实战 NVMe 之一

看完了蛋蛋的 NVMe 系列，你是不是有点手痒痒了，毕竟这是一门技术，不是艺术，只能欣赏，却不能亵玩，实在是太不让人尽兴了。三俗的来说，估计你也是看了上面的女神进来看文章的，但是，是不是下面的苍老师才能让你真正动手实践？

八卦蛋蛋

作为 ssdfans 的招牌，蛋蛋现在已经有了很多粉丝。你估计也想知道蛋蛋到底是个什么样的人，阿呆有幸略知一二。首先，蛋蛋博士真的是一位博士，饱读诗书，而且智商也很高。肯定有人要问了，你怎么知道他智商高呢，把跑分亮出来看看。其实阿呆也没跑过分，只知道蛋蛋家族有高智商基因，出了不少高考状元。乡亲们因为有了这样一家邻居，教育孩子不得不格外卖力气: 天赋不如人已经输在了起跑线上，后天可得加倍努力啊！

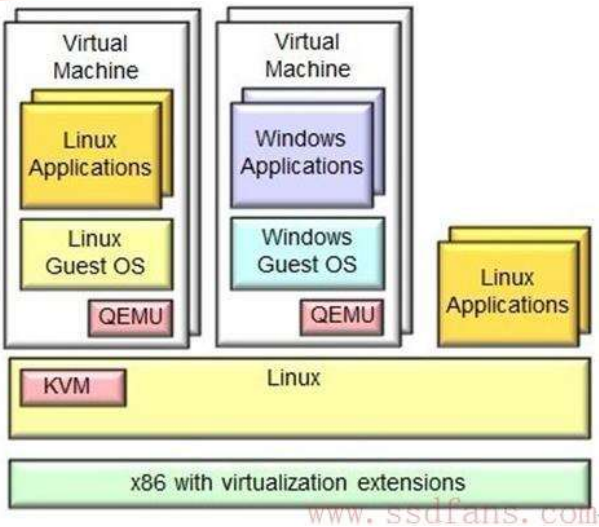
阿呆以前搞不懂 NVMe，都是英文文档，好不容易看完也不能消化，花个把星期很多也是云里雾里。但是，看了蛋蛋写的 NVMe 系列，通俗易懂，最多半天就能掌握精髓。想到这里，阿呆不得不像古人一样发出慨叹：

天不生蛋蛋，万古长如夜！

QEMU

好了，后面我们不扯蛋了，言归正传。尽管阿呆想动手实践 NVMe，可是舍不得花钱买昂贵的 PCIe SSD 练手，而且买了也是人家的控制器，自己怎么在上面搞开发啊，真是愁死人了。后来，阿呆想到一个好主意：真的玩不起，咱还不能搞个假的吗?果然找到一个免费的操作系统 Linux 上免费的虚拟机软件 QEMU。估计很多人玩过 VMWare，QEMU 差不多，只不过是开源的，所以手闲的人可以在上面搞开发。如下图，在 Linux 系统上用 QEMU 虚拟了一个 Windows 系统。

说到这里，不得不提一下大名鼎鼎的 KVM 了。总是听人家讲什么虚拟化，KVM 之类的，听起来挺时髦的，那 KVM 到底是个什么东东？现在所说的虚拟化，一般都是指在 CPU 硬件支持基础之上的虚拟化技术，就是有些 CPU 是支持虚拟化指令的。KVM 也一样依赖此项技术，全称 Kernel-based Virtual Machine，其实就是虚拟了一个 Linux 内核，而且通过 CPU 的专有虚拟化指令可以达到很高的效率。但是我们知道 Linux 操作系统分为内核层和用户层，光有 KVM 这个内核层，我们还跑不起来任何图形界面等用户应用。KVM 就相当于是开发商，只会盖房子，而且盖的非常坚固，但是盖好房子发现都是毛坯，不能住人，这时候他们发现了另一个开发商 QEMU，他们也盖房子，还搞装修，装修的本领很不错，盖得房子实在不咋地。于是，这两家开发商战略合作，KVM 负责盖房子，QEMU 负责搞装修，人们交了房就能住进精装修的房子了，又坚固又漂亮好用。所以现在往往是 KVM+QEMU 搞一个虚拟机，KVM 是内核态，QEMU 是用户态。如下图。



说来惭愧，阿呆的知识其实都是百度上搜的。如果你想更深入地了解 KVM，就得拜师了，上哪里找老师呢？哈哈，远在天边，近在眼前，我们藏龙卧虎的 ssdfans 微信群里就有位大牛，西山居的系统运维经理肖力，大名鼎鼎的《深度实践 KVM》就是他写的。

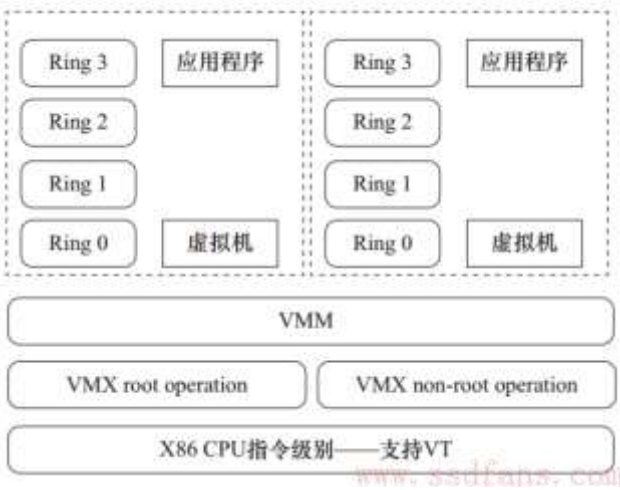
总是听人说云计算，虚拟化，为什么他俩关系这么铁？阿呆也不懂内情，但是知道云计算公司搞了很多虚拟机，因为我们 www.ssdfans.com 网站就是买了阿里云的 ECS 服务，所谓的 ECS 弹性云计算其实就是用虚拟机实现的。我们传统企业建设自己的机房，要配置很多电脑，交换机，存储阵列等，来搭建网站，实现邮件服务器，内部 OA 等服务。到了云计算这里，还是买很多机器，只不过这些机器都是虚拟的，所谓弹性就是系统的 CPU，内存，硬盘，带宽等资源都是可以配置的，谁叫人家是虚拟的呢，玩虚的就是任性啊。

不过最近总是有读者反映 ssdfans 网站从手机访问缓慢，电脑却还可以接受，阿呆还没好好看看是什么原因，希望有专家能够指点迷津。我们网站使用的是 wordpress，文件存放在阿里云 ECS，数据库采用了阿里云的 RDS，服务器只有杭州的，还没有买全国各地的 CDN 加速。

CPU 的虚拟化指令

后面阿呆要开始无耻的大段摘抄《深度实践 KVM》了， 嘿嘿，广告也不是白做的~~ 如下图，X86 的指令集一般分为 4 种特权模式：Ring 0，Ring 1，Ring 2，Ring 3，操作系统一般是 Ring 0，应用程序 Ring 3，驱动程序 Ring 1，Ring 2.可是到了虚拟机就遇到问题了：你敢让虚拟机的 Ring 0 指令真的跑在真实 CPU 的 Ring 0 模式下吗？毕竟都是虚拟的，搞不好就出问题了，但如果都跑在 Ring 3 模式下，那白白浪费了 CPU 的这个设计，性能又上不去了。

所以 Intel 等 CPU 厂商推出了虚拟化指令。Intel 的 VT-x 技术推出了两种操作模式：VMX root operation，VMX non-root operation。我们知道虚拟机包含两部分，虚拟层 VMM（virtual machine monitor）和虚拟出来的系统，虚拟层其实就是我们物理系统里的一个应用，所以它跑在 VMX root operation，和正常模式一样。而虚拟出来的系统跑在 VMX non-root operation，是处在 VMM 控制之下的一个环境。这样虚拟机跑在自己的 CPU 轨道里，物理机也跑在自己的原来的 CPU 轨道里，相互隔离，性能也得到了提高。



这就像是孩子长大娶了老婆，和父母住在一起，生活已经是两家人了，可是共用厨房和厕所，首先经常上厕所得排队，憋得受不了，其次，小两口的隐私也得不到保护。老爹没办法，可怜天下父母心啊，又花钱盖了一套厕所和厨房，大家同在屋檐下，再也不冲突了，相安无事。

I/O 虚拟化

一家人和睦了一段时间之后，新的问题又出现了。小两口消费观念新潮，总是需要买各种东西，但是又没时间去，所以只能让爹妈每天去超市代买，垃圾也让二老去倒。老人嫌累，小两口嫌他们少买了这个那个。老父亲好好想了想，找出症结了:尽管我们空间分开了，但是生活的原材料还没分开。后来，小两口不从超市购物了，他俩的东西全都选择网购，更自由，更方便，垃圾让快递小哥顺手帮忙带走，什么都搞定了。这就是 I/O 虚拟化啊：两家人的消费和废物处理隔离。

我们都知道，影响电脑性能的主要是 CPU 和内存，玩游戏还要显卡，下电影要网速和硬盘快。虚拟机也有这些问题，CPU 虚拟化提升了处理能力，但是数据操作也需要高性能啊，关键在于解决 I/O 设备与虚拟机数据交换的问题，而这部分主要相关的是 DMA 直接内存存取，以及 IRQ 中断请求，只要解决好这两个方面的隔离、保护以及性能问题，就是成功的 I/O 虚拟化。

慈祥的老父亲 Intel 开发了 VT-d 技术，在北桥（现在改叫 MCH，因为南桥退休了）提供 DMA 虚拟化和 IRQ 虚拟化。传统的 IOMMU（memory management unit，内存管理单元）集中管理所有 DMA，不容易实现 DMA 隔离，而 VT-d 实现了多个 DMA 保护区域的存在，实现 DMA 虚拟化。

传统设备中断请求有两种方式：一是通过 I/O 中断控制器路由，另一种是 DMA 写请求直接发出去的 MSI（message signaled interrupt，消息中断）。由于使用了 DMA，需要访问所有内存地址，没办法实现中断隔离，其实就是虚拟机的没办法区分。VT-d 的中断重映射架构重新定义了 MSI 的格式，尽管 MSI 依然是 DMA 写请求，但是不嵌入内存地址，而是消息 ID，通过消息 ID 区分不同的虚拟机区域。

啰啰嗦嗦说了这么多，怎么还没进入 NVMe 这个主题啊，没办法，阿呆就是这么个人，做足了前戏，慢慢的进入。

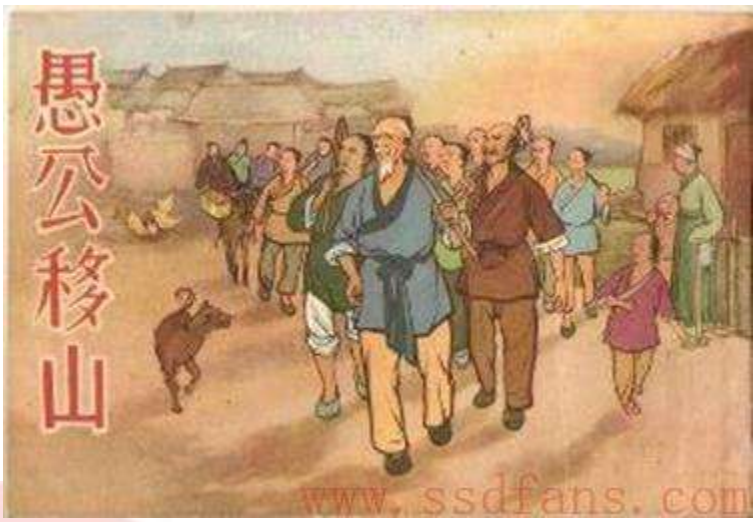
后面阿呆将会带你一起在 QEMU 虚拟机虚拟出一个 NVMe 设备，同时，在虚拟机 Linux 操作系统中开发一个 NVMe 驱动来使用它。

引用

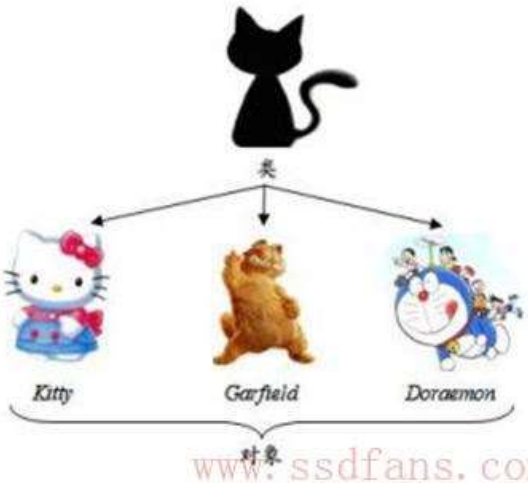
肖力等，《深度实践 KVM》

http://wenku.baidu.com/view/79bbd89851e79b89680226cb.html

10.4.2 阿呆实战 NVMe 之二



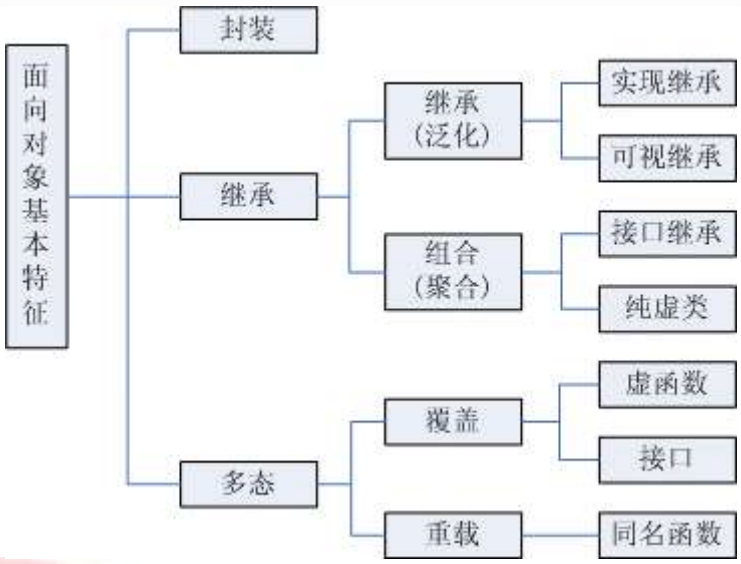
从前，计算机还是奢侈品，有一位屌丝程序员叫蛋蛋，尽管他每天开发电脑软件，但是码砖的住不起房子，码代码的买不起电脑。对蛋蛋来说，买个电脑跟愚公移山一样难，可是他太喜欢电脑了，怎么办？蛋蛋终于想出了一个主意：真的买不起，还不能搞一个假的吗？蛋蛋的想法是用软件模拟一个电脑，把所有的硬件都用软件来虚拟。隔壁 Cubicle 的程序员智叟听说了之后，讥笑他：”蛋蛋啊，你也太傻了，微软的软件高管说成熟的工程师一年只能写 4000 行代码，你什么时候才能模拟出那台电脑？”蛋蛋斩钉截铁地回答：”我死了，还有我的儿子，儿子死了还有孙子，子子孙孙无穷无尽地码代码，就能开发成功！”蛋蛋想出了一个子子孙孙继承事业的好办法，他用了面向对象的技术，自己只是写了顶层的几个类，确定好框架，后来的人想要加个新的硬件进来，就一层层继承各种类，完善接口就能使用了。没想到，子孙还没接班，其他世界各地的程序员也加入进来帮忙，很快就开发成了这个著名的虚拟机——QEMU。



记得以前有位同学说他去一家牛公司面试，一道面试题就是 C 语言怎么开发面向对象的代码？我听了这道题觉得还是有点难度，不过没好好细想。如今看了 QEMU 的代码，才发现这就是 C 语言实现面向对象的绝佳例子啊!他如果当初看过 QEMU 的源码，估计能和

面试官谈笑风生了。当然，本文的主要目的还是为了介绍 NVMe 在 QEMU 中的数据结构，教你面试题是副产品。代码不是阿呆自己写的，要那样我就牛了，QEMU 的 NVMe 虚拟机代码是 GIT 上的一个开源项目 https://github.com/nvmeqemu，谁都可以下载下来玩。

面向对象三大特点



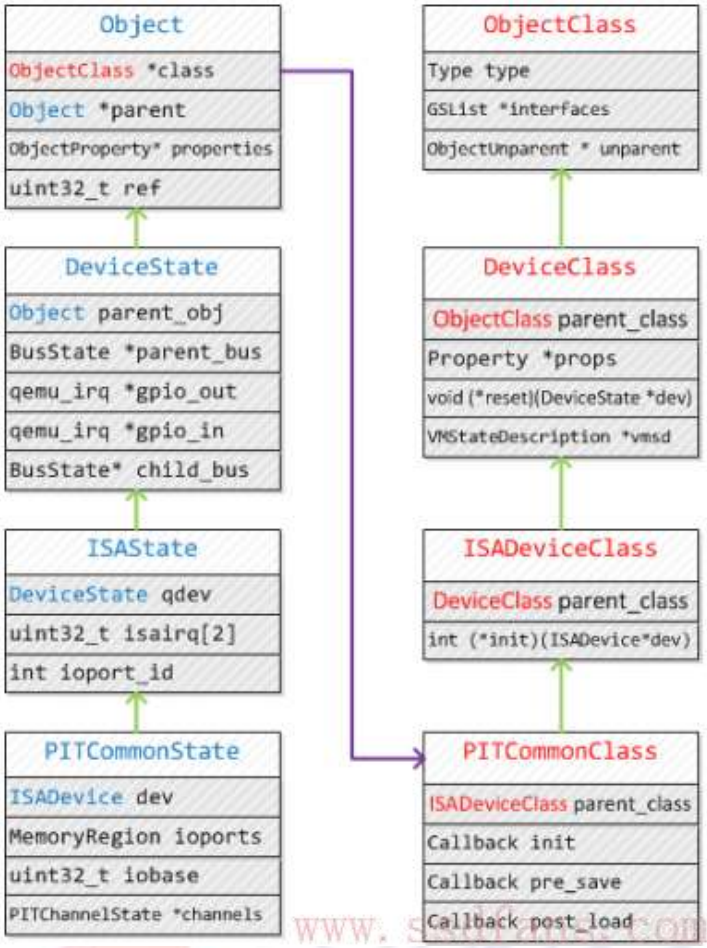
如上图，我们知道面向对象有三大特征：

1. 封装：这个好理解，就是在一个 struct/class 里面把各种变量，函数都打包塞进来。这个大部分人很快就能想到，很多宣称面向对象的 C 语言代码其实只是实现到了这一步。比如阿呆之前写的大话 EXT4 文件系统，Linux 内核的 VFS 说是面向对象设计，其实就是做了个封装而已。
2. 继承。继承的意思就程序员太懒了，要开发一个机器猫，想起以前搞过猫的对象，所以就让机器猫把猫这个对象包起来，这样猫对象所有的属性和接口都被继承了。再加一些机器猫独有的属性和接口，代码搞定。猫是父类，机器猫是子类，还可以继续子子孙孙，繁衍不息。这个咱 C 语言也有办法，就真的在 struct 里面包一个父类的 struct 不就完了，只不过纯手工，麻烦一点。
3. 多态。
    - 重载：就是函数名一样，参数不一样。这个比较好搞定，编译的时候编译成不一样的函数名就可以了。
    - 覆盖：父类和子类同样名称和参数的函数，子类重新定义覆盖父类的。执行的时候，如果子类对象赋给父类指针，还得执行子类的新的函数，这个叫做动态绑定。不太好弄，需要包含一个表格来实时查询。

我们下面来看看 NVMe 的数据结构，顺便看看 QEMU 实现了面向对象的哪些功能。

QEMU 面向谁的对象？

下图是一个 PIT 设备在 QEMU 中的结构，很明显，是个继承的关系，子类包含了一个父类的对象，继承了父亲的一切（其实应该是母亲的一切，谁叫码农是男的居多呢，parent 自然翻译成了父亲。但是生物上来说，那些没有两性生殖的生物，一切基因应该是来自于母亲。），不像我们两性生物，父母各取一半。不过新版的 QEMU 已经有点区别了，我看代码，NVMe 设备继承向上到了 DeviceState 就结束了。



QEMU 里面的设备分为三类：

```
typedef enum {
    MODULE_INIT_BLOCK,
    MODULE_INIT_DEVICE,
    MODULE_INIT_MACHINE,
    MODULE_INIT_MAX
} module_init_type;
```

Block 是 Linux 的块设备，比如磁盘 IO 之类，Device 就是各种外设，比如 PCIe 设备，Machine 是 CPU 等的虚拟。NVMe 是一个 PCI 设备，所以是 Device 类型。描述 NVMe 设备的对象叫 NVMEState，继承关系为：

```
struct DeviceState {
    const char *id;
    enum DevState state;
    QemuOpts *opts;
    int hotplugged;
    DeviceInfo *info;
    BusState *parent_bus;
    int num_gpio_out;
    qemu_irq *gpio_out;
    int num_gpio_in;
    qemu_irq *gpio_in;
    QLIST_HEAD(, BusState) child_bus;
    int num_child_bus;
    QLIST_ENTRY(DeviceState) sibling;
    int instance_id_alias;
    int alias_required_for_version;
};
```

```
struct PCIDevice {
    DeviceState qdev;
```

```
typedef struct NVMEState {
    PCIDevice dev;
```


那一个子类在 QEMU 中如果做到初始化？其实是要一级级初始化的，QEMU 用一种递归的方法，如果子类的父类还没有初始化，那就先初始化父类，调用父类的构造函数。

NVMe 设备怎么定义

尽管说 NVMe 设备的对象是 NVMEState，但是 nvme.c 里面只看到了一个 NVMe 相关的对象定义：

```
static PCIDeviceInfo nvme_info = {
    .qdev.name = "nvme",
    .qdev.desc = "Non-Volatile Memory Express",
    .qdev.size = sizeof(NVMEState),
    .qdev.vmsd = &vmstate_nvme,
    .qdev.reset = qdev_nvme_reset,
    .config_write = nvme_pci_write_config,
    .config_read = nvme_pci_read_config,
    .init = pci_nvme_init,
    .exit = pci_nvme_uninit,
    .qdev.props = (Property[]) {
        DEFINE_PROP_UINT32("namespaces", NVMEState, num_namespaces, 1),
        DEFINE_PROP_UINT32("size", NVMEState, ns_size, 512),
        DEFINE_PROP_END_OF_LIST(),
    }
};
```


那这个 nvme_info 是个什么东东?咱们且听下回分解。

引用

https://github.com/nvmegemu

QEMU 设备模拟，mnstory.net

10.4.3 阿呆实战 NVMe 之三

有人说阿呆你搞个虚拟的东西来玩 NVMe，看起来是很爽，代码随便写，然并卵有个什么鸟用呢？其实在实际产品研发中还是有点用的，QEMU 最大的好处是可以用 GDB Debug。一般实体机如果发生了 Panic，只能通过 Linux Kernel Panic Dump 的方式查看堆栈，

找原因，有些藏得比较深的问题比较难发现。而 QEMU 虚拟机在出错的时候会停下来，就能直接查看出错时各个变量的现场值，找到 Root Cause。你如果开发了 Linux NVMe 驱动，可以用 QEMU Debug，也能做一些简单的测试。

我们上回说到，蛋蛋定义了 QEMU 的一套顶层架构，这样后来的人加新的硬件进来，就可以直接套用。本文就通过 NVMe 设备的追根溯源来看看 QEMU 使用了何种巧妙的架构来驾驭那么多复杂的硬件。

怎样在 QEMU 注册一个 NVMe 设备？

上回说 nvme.c 里面定义了一个 PCIDeviceInfo 对象 nvme_info，如下图，包含了构造和析构函数，还有其他配置与接口赋值。那这个对象是怎样在 QEMU 之中被使用的呢？

```
static PCIDeviceInfo nvme_info = {
    .qdev.name = "nvme",
    .qdev.desc = "Non-Volatile Memory Express",
    .qdev.size = sizeof(NVMEState),
    .qdev.vmsd = &vmstate_nvme,
    .qdev.reset = qdev_nvme_reset,
    .config_write = nvme_pci_write_config,
    .config_read = nvme_pci_read_config,
    .init = pci_nvme_init,
    .exit = pci_nvme_uninit,
    .qdev.props = (Property[]) {
        DEFINE_PROP_UINT32("namespaces", NVMEState, num_namespaces, 1),
        DEFINE_PROP_UINT32("size", NVMEState, ns_size, 512),
        DEFINE_PROP_END_OF_LIST(),
    }
};
```


如下面代码，nvme_register_devices 注册了一个 PCI 设备 nvme_info。然后，用一个宏 device_init 来声明这个注册函数。看得出来，奥秘就在这个宏里面。

```
static void nvme_register_devices(void)
{
    pci_qdev_register(&nvme_info);
}

device_init(nvme_register_devices);
```


如下，可以看到 device_init 宏变成了 module_init，而这个 module_init 宏是个有 constructor attribute 的函数，起什么作用呢？在 GCC 中，这两个编译的 attribute 用于修饰某个函数，经过 constructor 属性修饰过的函数，可以在 main 函数运行前就先运行完毕，同理 destructor 在进程 exit 之前执行。相当于构造和析构函数。要知道 Linux 分为内核态和用户态，内核态的程序都是通过 modprobe .ko 文件形式加载，没有 main 函数一说。而用户态都是从 main 函数进来的，而虚拟机本身是跑在用户态的，所以也是从 main 函数进来。不过，module_init 宏声明的函数在 main 函数之前就执行了，就是为 main 函数做一些配置工作。

我们说 QEMU 把硬件分成了 Block，Device 和 Machine 三种，最终三种硬件的注册函数都汇总到了一个函数：register_module_init(void (*fn)(void), module_init_type type).

```
/* This should not be used directly.  Use block_init etc. instead.  */
#define module_init(function, type)                                         \
static void __attribute__((constructor)) do_qemu_init_ ## function(void) {  \
    register_module_init(function, type);                                   \
}

typedef enum {
    MODULE_INIT_BLOCK,
    MODULE_INIT_DEVICE,
    MODULE_INIT_MACHINE,
    MODULE_INIT_MAX
} module_init_type;

#define block_init(function) module_init(function, MODULE_INIT_BLOCK)
#define device_init(function) module_init(function, MODULE_INIT_DEVICE)
#define machine_init(function) module_init(function, MODULE_INIT_MACHINE)
```


于是，我们又来一探这个通用的注册函数干了些什么。如下面的代码，功能非常简单，就是每种硬件类型，在 QEMU 中都有一个设备链表，链表中每个节点的内容就是这个设备的注册函数。所以，在 QEMU 中注册一个设备就是把这个设备的 ModuleEntry 挂到链表的尾巴上。

```
void register_module_init(void (*fn)(void), module_init_type type)
{
    ModuleEntry *e;
    ModuleTypeList *l;

    e = qemu_mallocz(sizeof(*e));
    e->init = fn;

    l = find_type(type);

    QTAILQ_INSERT_TAIL(l, e, node);
}
```


QEMU 初始化

NVMe 设备的注册是完成了，非常的简单。那么 QEMU 又怎么来使用这个注册的设备呢？请你回过头看看，其实折腾了半天，只是在 main 函数调用之前把 nvme_register_devices 这个函数注册到了 device 设备的链表里面，真正的 NVMe 初始化还没弄呢。不过别着急，NVMe 设备初始化肯定是和 nvme_register_devices 函数内容分不开了。那我们就来看看这个函数做了什么。

```
void pci_qdev_register(PCIDeviceInfo *info)
{
    info->qdev.init = pci_qdev_init;
    info->qdev.unplug = pci_unplug_device;
    info->qdev.exit = pci_unregister_device;
    info->qdev.bus_info = &pci_bus_info;
    qdev_register(&info->qdev);
}
```


看起来是给继承的 DeviceInfo qdev 赋初值。下面是 PCIDeviceInfo 的定义，包含了构造和析构函数，PCI 设备的 config 读写接口，还有 vendor id 等参数。

```
typedef struct {
    DeviceInfo qdev;
    pci_qdev_initfn init;
    PCIUnregisterFunc *exit;
    PCIConfigReadFunc *config_read;
    PCIConfigWriteFunc *config_write;

    uint16_t vendor_id;
    uint16_t device_id;
    uint8_t revision;
    uint16_t class_id;
    uint16_t subsystem_vendor_id;       /* only for header type = 0 */
    uint16_t subsystem_id;              /* only for header type = 0 */

    /*
     * pci-to-pci bridge or normal device.
     * This doesn't mean pci host switch.
     * When card bus bridge is supported, this would be enhanced.
     */
    int is_bridge;

    /* pcie stuff */
    int is_express;   /* is this device pci express? */

    /* device isn't hot-pluggable */
    int no_hotplug;

    /* rom bar */
    const char *romfile;
} ? end [anonPCIDeviceInfo] ?  PCIDeviceInfo;
```


拿这些注册的 DeviceInfo 的 init 函数到底什么时候调用？要找到答案，我们就得去看看 main 函数做了些什么，毕竟注册 PCIDeviceInfo 对象是在 main 之前完成，那 main 肯定不会不用注册的结果。QEMU 包含了很多 main 函数，因为有很多工具，但是真正的 main 函数位于 Vl.c。

```
int main(int argc, char **argv, char **envp)
{
```


首先初始化最核心的 Machine 设备，CPU 之类的。

```
02085:      module_call_init(MODULE_INIT_MACHINE);
02086:      machine = find_default_machine();
02087:      cpu_model = NULL;
02088:      initrd_filename = NULL;
02089:      ram_size = 0;
02090:      snapshot = 0;
02091:      kernel_filename = NULL;
02092:      kernel_cmdline = "";
02093:      cyls = heads = secs = 0;
02094:      translation = BIOS_ATA_TRANSLATION_AUTO;
```


中间有 1000 多行代码不知道在干嘛，反正还没轮到 Device 设备。直到 1000 多行以后，Device 设备才被初始化。

```
03165:
03166:      module_call_init(MODULE_INIT_DEVICE);
03167:
```


但是当我们看到 module_call_init 函数的内容，还是很失望，它只是调用了 e->init()。

```
void module_call_init(module_init_type type)
{
    ModuleTypeList *l;
    ModuleEntry *e;

    l = find_type(type);

    QTAILQ_FOREACH(e, l, node) {
        e->init();
    }
}
```


e->init()是什么，请翻到本文开头，nvme_register_devices 函数就是 e->init()！！是不是觉得搞了半天，又回来了。悲哀啊，怎么这么麻烦啊，那本文开头的 nvme_info 的那些 init 函数到底什么时候调用？等下回阿呆搞明白了再告诉你。

> 引用

https://github.com/nvmegemu

10.4.4 阿呆实战 NVMe 之四

蝶恋花 【宋】欧阳修

庭院深深深几许？杨柳堆烟，帘幕无重数。玉勒雕鞍游冶处，楼高不见章台路。
雨横风狂三月暮，门掩黄昏，无计留春住。泪眼问花花不语，乱红飞过秋千去。

看完前面一篇，我相信你的心情就跟欧阳修这首《蝶恋花》中说的一样，感到：”庭院深深深几许？杨柳堆烟，帘幕无重数。”杨柳依依，一重又一重堵在前面，像一层层帘幕一样，我们翻了这么多层代码，根本看不到 NVMe 初始化在哪里(▼-▼)

再探 DeviceInfo

读者君，还记得 NVMe 设备的注册函数吗？我们再来过一遍，

首先有一个 static 类型的全局变量 nvme_info，是个 PCIDeviceInfo，内容为

```
static PCIDeviceInfo nvme_info = {
    .qdev.name = "nvme",
    .qdev.desc = "Non-Volatile Memory Express",
    .qdev.size = sizeof(NVMEState),
    .qdev.vmsd = &vmstate_nvme,
    .qdev.reset = qdev_nvme_reset,
    .config_write = nvme_pci_write_config,
    .config_read = nvme_pci_read_config,
    .init = pci_nvme_init,
    .exit = pci_nvme_uninit,
    .qdev.props = (Property[]) {
        DEFINE_PROP_UINT32("namespaces", NVMEState, num_namespaces, 1),
        DEFINE_PROP_UINT32("size", NVMEState, ns_size, 512),
        DEFINE_PROP_END_OF_LIST(),
    }
};
```


接着，通过一连串我们上文中发现的初始化流程，下面这个函数在 main 函数中通过调用 module_call_init(MODULE_INIT_DEVICE)而被执行。所以我们再来一级级深入剖析下面这个函数：

```
static void nvme_register_devices(void)
{
    pci_qdev_register(&nvme_info);
}
```


PCIDeviceInfo 对象 nvme_info 通过 pci_qdev_register 注册，而 PCIDeviceInfo 继承了 DeviceInfo 类，

```
typedef struct {
    DeviceInfo qdev;
    pci_qdev_initfn init;
    PCIUnregisterFunc *exit;
    PCIConfigReadFunc *config_read;
    PCIConfigWriteFunc *config_write;

    uint16_t vendor_id;
    uint16_t device_id;
    uint8_t revision;
    uint16_t class_id;
    uint16_t subsystem_vendor_id;       /* only for header type = 0 */
    uint16_t subsystem_id;              /* only for header type = 0 */

    /*
     * pci-to-pci bridge or normal device.
     * This doesn't mean pci host switch.
     * When card bus bridge is supported, this would be enhanced.
     */
    int is_bridge;

    /* pcie stuff */
    int is_express;   /* is this device pci express? */

    /* device isn't hot-pluggable */
    int no_hotplug;

    /* rom bar */
    const char *romfile;
} ? end {anonPCIDeviceInfo} ? PCIDeviceInfo;
```


所以类似，函数 pci_qdev_register 就是把这个 pci 设备的父对象 info->qdev 通过 qdev_register 函数注册。

```
void pci_qdev_register(PCIDeviceInfo *info)
{
    info->qdev.init = pci_qdev_init;
    info->qdev.unplug = pci_unplug_device;
    info->qdev.exit = pci_unregister_device;
    info->qdev.bus_info = &pci_bus_info;
    qdev_register(&info->qdev);
}
```


如下，nvme_info 的父对象 DeviceInfo 添加到了全局变量 device_info_list 当中，这个是所有 device 设备的链表。后面就好办了，我们只要查找这个全局变量调用的代码，就能找到初始化的地方了。

```
/* Register a new device type.  */
void qdev_register(DeviceInfo *info)
{
    assert(info->size >= sizeof(DeviceState));
    assert(!info->next);

    info->next = device_info_list;
    device_info_list = info;
}
```

```
struct DeviceInfo {
    const char *name;
    const char *fw_name;
    const char *alias;
    const char *desc;
    size_t size;
    Property *props;
    int no_user;

    /* callbacks */
    qdev_resetfn reset;

    /* device state */
    const VMStateDescription *vmsd;

    /* Private to qdev / bus.  */
    qdev_initfn init;
    qdev_event unplug;
    qdev_event exit;
    BusInfo *bus_info;
    struct DeviceInfo *next;
} ? end qdev_init ? ;
extern DeviceInfo *device_info_list;
```


离真相只有几步

功夫不负有心人，阿呆终于搞清楚了来龙去脉，下面就是揭晓奇迹的时刻。上文中我们看到 main 函数调用了

```
03166:      module_call_init(MODULE_INIT_DEVICE);
```

但是，没有继续留意后面的代码，其实再看几十行，就有一段有点不太直接的代码。

```
03197:      /* init generic devices */
03198:      if (qemu_opts_foreach(qemu_find_opts("device"), device_init_func, NULL, 1) != 0)
03199:          exit(1);
```


这段代码就是 device 设备初始化的地方，人世间很多事都是这样，往往我们在离真相很近的时候提前放弃了，等到花了大工夫搞明白之后，才追悔莫及。文科生就不像我们理科生这么纠结，他们能用文学化悲痛为美好：众里寻他千百度，蓦然回首，那人却在灯火阑珊处。

那么问题来了，qemu_opts_foreach 和 qemu_find_opts 都是来干嘛的？这里我们需要来看一条 QEMU 启动的命令:

qemu-system-x86_64 –enable-kvm -cpu host -smp cores=4,threads=2,sockets=4 -m 16384 -k en-us -hda /pps/guohongwei/vm_test/ubuntu.img -monitor stdio -device nvme

这里启动了 QEMU，并且配置了各种外设，使用了 KVM 内核，CPU 和内存配置，硬盘镜像，最后是 device 设备，加载了 nvme 设备。上面的两个函数 qemu_opts_foreach(qemu_find_opts(“device”), …)就是从参数列表中找到 device 设备，并且遍历。

这里遍历了注册的 Device 设备，通过 device_init_func 一个个初始化。我们来看初始化流程。

```
static int device_init_func(QemuOpts *opts, void *opaque)
{
    DeviceState *dev;

    dev = qdev_device_add(opts);
    if (!dev)
        return -1;
    return 0;
}
```


qdev_device_add 函数首先通过 qdev_find_info 查询，从我们前面看到的 DeviceInfo 注册的链表 device_info_list 中查到要初始化的 device。我们翻到开头 nvme_info 的初始值就知道，nvme_info 把里面 DeviceInfo 的 name 初始化为”nvme”，所以 qdev_find_info 就可以找到”-device nvme”对应的 DeviceInfo。

```
/* find driver */
info = qdev_find_info(NULL, driver);
if (!info || info->no_user) {
    qerror_report(QERR_INVALID_PARAMETER_VALUE, "driver", "a driver name");
    error_printf_unless_qmp("Try with argument '?' for a list.\n");
    return NULL;
}
```



```
static DeviceInfo *qdev_find_info(BusInfo *bus_info, const char *name)
{
    DeviceInfo *info;

    /* first check device names */
    for (info = device_info_list; info != NULL; info = info->next) {
        if (bus_info && info->bus_info != bus_info)
            continue;
        if (strcmp(info->name, name) != 0)
            continue;
        return info;
    }
```


得到 DeviceInfo 之后，创建 DeviceState 对象 qdev。

```
/* create device, set properties */
qdev = qdev_create_from_info(bus, info);
id = qemu_opts_id(opts);
if (id) {
    qdev->id = id;
}
if (qemu_opt_foreach(opts, set_property, qdev, 1) != 0) {
    qdev_free(qdev);
    return NULL;
}
if (qdev_init(qdev) < 0) {
    qerror_report(QERR_DEVICE_INIT_FAILED, driver);
    return NULL;
}
```


首先来看 qdev_create_from_info，创建了 DeviceState，真的吗？真的仅仅是创建了一个 DeviceState 对象吗？注意这里 mallocz 的大小是 info->size，而开头的时候，DeviceInfo 的 size 变量我们给的是 sizeof(NVMEState)！也就是说，其实这里创建的是一个 NVMEState 对象！并且给 props 赋初始值。那这些 props 是怎么来的？

```
static DeviceState *qdev_create_from_info(BusState *bus, DeviceInfo *info)
{
    DeviceState *dev;

    assert(bus->info == info->bus_info);
    dev = qemu_mallocz(info->size);
    dev->info = info;
    dev->parent_bus = bus;
    qdev_prop_set_defaults(dev, dev->info->props);
    qdev_prop_set_defaults(dev, dev->parent_bus->info->props);
    qdev_prop_set_globals(dev);
    QLIST_INSERT_HEAD(&bus->children, dev, sibling);
    if (qdev_hotplug) {
        assert(bus->allow_hotplug);
        dev->hotplugged = 1;
        qdev_hot_added = true;
    }
    dev->instance_id_alias = -1;
    dev->state = DEV_STATE_CREATED;
    return dev;
} ? end qdev_create_from_info ?
```



```
static PCIDeviceInfo nvme_info = {
    .qdev.name = "nvme",
    .qdev.desc = "Non-Volatile Memory Express",
    .qdev.size = sizeof(NVMEState),
```


请再回开头看看 nvme_info 的初始化，尾巴上有下面一段，通过 Property 结构体，给定了初始值。通过这种方式给 NVMEState 对象的变量赋初值。num_namespace 和 ns_size 都是 NVMEState 类的成员。

```
    .qdev.props = (Property[]) {
        DEFINE_PROP_UINT32("namespaces", NVMEState, num_namespaces, 1),
        DEFINE_PROP_UINT32("size", NVMEState, ns_size, 512),
        DEFINE_PROP_END_OF_LIST(),
    }
```


create 之后，再来看后面的 qdev_init 函数，其实就是调用了 DeviceInfo 里面的 init 函数。

```
/* Initialize a device.  Device properties should be set before calling
   this function.  IRQs and MMIO regions should be connected/mapped after
   calling this function.
   On failure, destroy the device and return negative value.
   Return 0 on success.  */
int qdev_init(DeviceState *dev)
{
    int rc;

    assert(dev->state == DEV_STATE_CREATED);
    rc = dev->info->init(dev, dev->info);
```


请往前翻 pci_qdev_register 函数内容，就知道 DeviceInfo 里的 init 函数是 pci_qdev_init。所以，我们又来看它里面做了什么。一进来，就通过 container_of 宏得到 DeviceInfo 的子类 PCIDeviceInfo。为什么，因为我们在 PCIDeviceInfo 里面有一个对象是 DeviceInfo qdev，知道了二者的偏移关系，就可以 qdev 的内存地址减去偏移量得到 PCIDeviceInfo 的内存地址，就是对象的指针。container_of 这个宏就是把这堆复杂的计算弄成了一个宏。在 linux 编程中还是很常见的，尤其是已知 struct 内部变量地址，要计算 struct 地址的时候。

```
static int pci_qdev_init(DeviceState *qdev, DeviceInfo *base)
{
    PCIDevice *pci_dev = (PCIDevice *)qdev;
    PCIDeviceInfo *info = container_of(base, PCIDeviceInfo, qdev);
```


还有，DeviceState 指针直接转成了 PCIDevice 指针，为什么可以这样？来看看 PCIDevice 的定义:第一个变量就是 DeviceState，很巧妙吧，这样就能做到子类到父类指针的轻松切换。为了面向对象，QEMU 真是煞费苦心啊。

```
struct PCIDevice {
    DeviceState qdev;
```

接着，又调用了 info->init 函数，不过请注意，这里的 info 已经变成了 PCIDeviceInfo，所以此 init 非彼 init 了。

```
if (info->init) {
    rc = info->init(pci_dev);
```

我们再回过头看 nvme_info 的初始化，就知道 init 函数是 pci_nvme_init。看到这里，憋了很久的阿呆忍不住大吼一声："终于轮到 NVME 的初始化了！"

```
static PCIDeviceInfo nvme_info = {
    .qdev.name = "nvme",
    .qdev.desc = "Non-Volatile Memory Express",
    .qdev.size = sizeof(NVMEState),
    .qdev.vmsd = &vmstate_nvme,
    .qdev.reset = qdev_nvme_reset,
    .config_write = nvme_pci_write_config,
    .config_read = nvme_pci_read_config,
    .init = pci_nvme_init,
    .exit = pci_nvme_uninit,
```

是啊，看了三篇代码，理清了 QEMU 买下的一个个，才轮到 NVMe 出场，不容易啊！下期我们就来看看 NVMe 的初始化，不过请先复习蛋蛋的《蛋蛋读 NVME 之 X》系列文章。

本文以一首北宋词开头，再以一首南宋词结尾，来表达阿呆此刻的心情。

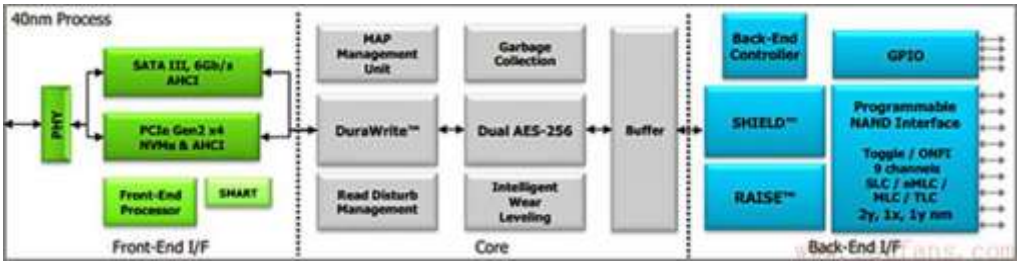
青玉案·元夕 【宋】辛弃疾

东风夜放花千树。更吹落、星如雨。宝马雕车香满路。凤箫声动，玉壶光转，一夜鱼龙舞。

> 蛾儿雪柳黄金缕。笑语盈盈暗香去。众里寻他千百度。蓦然回首，那人却在，灯火阑珊处。

引用
https://github.com/nvmegemu

10.4.5 阿呆实战 NVMe 之五



提要

我们知道，一般的 SSD 控制器里面分为前端，核心，后端，如上图是希捷收购的 SandForce SF3700 控制器的架构，三星不少 SSD 主控也是类似，三个部分分别用三个 CPU 管理。功能分别为：

- 前端：实现 SATA，PCIe 等物理电路和 ATA，NVMe 等协议。
- 核心：FTL，垃圾回收等等 SSD 核心算法。
- 后端：RAID，ECC，NAND Flash 接口实现。

所以，跟 NVMe 相关的是前端部分，从本文开始，我们正式进入 NVMe 内容，开发一个 NVMe SSD 控制器的前端协议逻辑。

前戏

记得去年阿呆家小呆呆还没出生，有人忽悠阿呆说胎教很管用。阿呆开始每天背唐诗宋词，晚上睡前隔着妹子肚皮背给娃听。不知道他有没有听进去，阿呆却发现原来诗词意境这么美，夜里躺床上听到雨滴落在别人家的雨篷上，滴滴答答，整个世界在春雨声中安静了，从耳朵到内心。我就想起老年的陆游进京面见宋孝宗，在客店里听了一夜的春雨，折腾了大半辈子，他已经不相信有人会派他去收复河山了，反而闲情逸致写了首诗，果然

宋孝宗找他也没什么事。不过那时候的人重视文化，陆游写的这首诗很快传遍了临安，传进了皇宫，宋孝宗反复诵读，赞叹不已。所以我建议各位读者晚上回家少刷朋友圈，看看诗词，这样睡得更香，第二天看 ssdfans 就有精神。

临安春雨初霁

年代: 宋 作者: 陆游

世味年来薄似纱，谁令骑马客京华。
小楼一夜听春雨，深巷明朝卖杏花。
矮纸斜行闲作草，晴窗细乳戏分茶。
素衣莫起风尘叹，犹及清明可到家。

最近老是贴代码，有点枯燥，蛋蛋没有工作后休息的这几周 ssdfans 公众号订阅人数增量看来不能达标了，希望大家多看看精华文章，分享一下蛋蛋的精品。不过别担心，往后不会贴那么多代码了（这是真的吗？）。

不知上次阿呆提醒之后，你有没有扫一下《蛋蛋读 NVMe 之 x》系列文章，尤其是最后一篇。从本篇起我们要进入 NVMe 相关的内容了，冒出个名词你得有心理准备，比如 namespace，这个是《蛋蛋读 NVMe 之六》里面提到的。说牛逼一点，我们要开始做一件激动人心的事：开发一个 NVMe Controller 了！尽管这个过程比较枯燥，但是只要你耐心看完，就一定有收获。毕竟理论联系实践才是王道。

NVMEState

今天的任务不重，只是看看 NVMEState 这个 NVMe 设备的数据结构。治大国如烹小鲜，NVMe 协议很庞杂，但只要我们读透 NVMe 的数据结构这把刀，后面看代码就如庖丁解牛般游刃有余。

typedef

```
struct NVMEState {

PCIDevice dev;


int mmio_index;


void

*bar0;


int bar0_size;

uint8_t nvectors;

/* Space for NVME Ctrl Space except doorbells */


uint8_t *cntrl_reg;

/* Masks for NVME Ctrl Registers */
```

```
uint8_t *rw_mask;
/* RW/RO mask */

uint8_t *rwc_mask;
/* RW1C mask */

uint8_t *rws_mask;
/* RW1S mask */

uint8_t *used_mask;
/* Used/Resv mask */

 struct nvme_features feature;


NVMEIOCQueue cq[NVME_MAX_QS_ALLOCATED];

NVMEIOSQueue sq[NVME_MAX_QS_ALLOCATED];

 DiskInfo *disk;

uint32_t ns_size;

uint32_t num_namespaces;

uint32_t instance;


time_t start_time;


/* Used to store the AQA,ASQ,ACQ between resets */


struct AQState aqstate;


/* TODO


* These pointers have been defined since

* present code uses the older defined strucutres

* which has been replaced by pointers.

* Once each and every reference is replaced by
```

```
* offset from cntrl_reg, remove these pointers

* becasue bit field structures are not portable

* especially when the memory locations of the bit fields

* have importance

*/

NVMECtrlCap *ctrlcap;

NVMEVersion *ctrlv;

NVMECtrlConf *cconf;
/* Ctrl configuration */

NVMECtrlStatus *cstatus;
/* Ctrl status */

NVMEAQA *admqattrs;
/* Admin queues attributes. */



QEMUTimer *sq_processing_timer;

int64_t sq_processing_timer_target;


/* Used for PIN based and MSI interrupts */

uint32_t intr_vect;


/* Page Size used by the hardware */

uint32_t page_size;


/* Pointer to Identify Controller Strucutre */

NVMEIdentifyController *idtfy_ctrl;


/* Pointer to Firmware slot info log page */

NVMEFwSlotInfoLog fw_slot_log;

uint8_t last_fw_slot;



uint8_t temp_warn_issued;
```

```
QEMUTimer *async_event_timer;



uint16_t async_cid[ASYNC_EVENT_REQ_LIMIT +
1];

uint16_t outstanding_asyncs;



QSIMPLEQ_HEAD(async_queue, AsyncEvent) async_queue;


/* Masks for async event requests */

uint8_t err_sts_mask;
/* error status event mask */

uint8_t smart_mask;
/* smart/health status event mask */

} NVMEState;
```

一个个来看。

- PCIDevice dev
  - **NVMe** 所继承的 PCI 设备对象。
- int mmio_index
  - NVMe 设备在 QEMU 内存中的索引。
- void
- *bar0
  - PCI 设备 BAR 空间的地址。NVMe 设备中的寄存器，doorbell（host 有事了就按门铃告诉设备:您有新短消息，请注意查收~）都通过这段内存地址实现。
- int bar0_size
  - bar0 空间大小。
- uint8_t nvectors

- 这个其实就是 NVMe 的队列个数。

- uint8_t *cntrl_reg
-

顾名思义，这就是 NVMe 的控制寄存器指针了。后面四个小兄弟是寄存器的 MASK，啥意思，就是每个 bit 加了个限制。

```
uint8_t *rw_mask;
/* RW/RO mask */ 可读写或只读

uint8_t *rwc_mask;
/* RW1C mask */ 写 1 清零

uint8_t *rws_mask;
/* RW1S mask */ 写 1 置 1

uint8_t *used_mask;
/* Used/Resv mask */ 是否在用
```

- struct nvme_features feature;
-

- NVMe 设备的各种特征值。

```
struct nvme_features {

uint32_t arbitration;

uint32_t power_management;

uint32_t LBA_range_type;
/* uses memory buffer */

uint32_t temperature_threshold;

uint32_t error_recovery;

uint32_t volatile_write_cache;

uint32_t number_of_queues;

uint32_t interrupt_coalescing;

uint32_t interrupt_vector_configuration;

uint32_t write_atomicity;

uint32_t asynchronous_event_configuration;

uint32_t software_progress_marker;

};
```

- NVMEIOCQueue cq[NVME_MAX_QS_ALLOCATED];

- NVMEIOSQueue sq[NVME_MAX_QS_ALLOCATED];
  - CQ 和 SQ，SQ 是 host 发命令的队列，CQ 是 NVMe 回结果的队列。
- DiskInfo *disk;

既然我们虚拟的是一个 NVMe SSD，那么里面肯定是有磁盘的，这个就是虚拟磁盘的数据结构。其实这个磁盘是放在文件里面，DIskInfo 里有文件指针，读写偏移，还有比如我们熟悉的 Identify 页内容等。

- uint32_t ns_size;
- uint32_t num_namespaces;

namespace 大小和个数。不知道 namespace 为何物的请查看精华文章，查看《蛋蛋读 NVMe 之六》。

- uint32_t instance;

注明这是第几个 NVMe 设备。

- time_t start_time

启动时间，用来算 SMART Log 里面的上电时间。

- struct AQState aqstate;

控制器 reset 之前保存 Admin 队列 AQA,ASQ,ACQ 的状态。

- NVMECtrlCap *ctrlcap;
- NVMEVersion *ctrlv;
- NVMECtrlConf *cconf;
  /* Ctrl configuration */
- NVMECtrlStatus *cstatus;
  /* Ctrl status */
- NVMEAQA *admqattrs;
  /* Admin queues attributes. */

这是 NVMe 设备 5 个寄存器的指针，设备初始化的时候把寄存器内存地址赋给它们。不过注释里也说了，这种指针表示法可移植性差。以后会直接用控制寄存器的偏移来访问。

- QEMUTimer *sq_processing_timer;
- int64_t sq_processing_timer_target;

sq_processing_timer 是处理 SQ 所用的 timer，里面注册了回调函数，sq_processing_timer_target 是触发 timer 的时间，当时间到了 target 之后，timer 里面注册的回调函数就会被调用。timer 的用途很广，比如让 NVMe 控制器定期处理 SQ 队列里的新命令，就可以每次检查完把 target 设置为一定时间以后再次触发，这样无穷无尽循环下去。

- uint32_t intr_vect;

NVMe 设备中断向量。

- uint32_t page_size;

NVMe 控制器和 Host 数据交互的页大小。

- NVMEIdentifyController *idtfy_ctrl;

NVMe 控制器 Identify Controller 的指针。里面就是设备相关的一些参数，比如 vendor id， subsystem vendor id 等，具体定义请参考 NVM Express 1.0b Chapter 5.11 Identify command。

- NVMEFwSlotInfoLog fw_slot_log;
- uint8_t last_fw_slot;

NVMe 控制器 Firmware slot info log 的指针。里面是固件相关的一些参数，比如固件版本。last_fw_slot 是上次固件的 slot 值。

- uint8_t temp_warn_issued;

NVMe 控制器温度预警了吗？

- QEMUTimer *async_event_timer;

```
uint16_t async_cid[ASYNC_EVENT_REQ_LIMIT +
1];

uint16_t outstanding_asyncs;

QSIMPLEQ_HEAD(async_queue, AsyncEvent) async_queue;
```

这又是一个 timer，搞虾米用的呢？顾名思义，NVMe Admin 命令有一个叫 Asynchronous Event Request，这个 timer 就是为这类异步请求设置的。Host 有时候给 NVMe 控制器打招呼:小弟，有个人帮大哥盯着，出现了就报告。NVMe 小弟赶快记下来，就是这个异步事件，有空了就留心一下。初始化注册回调函数，事件发生了就会调用回调函数。不相信？请看《蛋蛋读 NVMe 之一》，阿呆帮你把图贴过来。async_cid 是异步请求的 command id。outstanding_asyncs 是 NVMe 控制器还没完成的异步事件数。async_queue 就是异步事件队列。

Figure 40: Opcodes for Admin Commands

Opcode (07) Generic Command	Opcode (06:02) Function	Opcode (01:00) Data Transfer	Opcode[2]	O/M[1]	Namespace Identifier Used[3]	Command
0b	000 00b	00b	00h	M	No	Delete I/O Submission Queue
0b	000 00b	01b	01h	M	No	Create I/O Submission Queue
0b	000 00b	10b	02h	M	Yes	Get Log Page
0b	000 01b	00b	04h	M	No	Delete I/O Completion Queue
0b	000 01b	01b	05h	M	No	Create I/O Completion Queue
0b	000 01b	10b	06h	M	Yes	Identify
0b	000 10b	00b	08h	M	No	Abort
0b	000 10b	01b	09h	M	Yes	Set Features
0b	000 10b	10b	0Ah	M	Yes	Get Features
0b	000 11b	00b	0Ch	M	No	Asynchronous Event Request
0b	000 11b	01b	0Dh	O	Yes	Namespace Management
0b	001 00b	00b	10h	O	No	Firmware Commit
0b	001 00b	01b	11h	O	No	Firmware Image Download
0b	001 01b	01b	15h	O	Yes	Namespace Attachment
I/O Command Set Specific						
1b	na	Na	80h – BFh	O		I/O Command Set specific
Vendor Specific						
1b	na	Na	C0h – FFh	O		Vendor specific

NOTES:
1. O/M definition: O = Optional, M = Mandatory.
2. Opcodes not listed are reserved.
3. A subset of commands uses the Namespace Identifier field (CDW1.NSID). When not used, the field shall be cleared to 0h.


```
uint8_t err_sts_mask;
/* error status event mask */

uint8_t smart_mask;
/* smart/health status event mask */
```

当 NVMe 出现了 error 或者 smart 事件，对应的 mask 置 1.

你是不是还很疑惑 PCI 的 bar 和 mmio 到底是怎么用的？下期为你解惑。

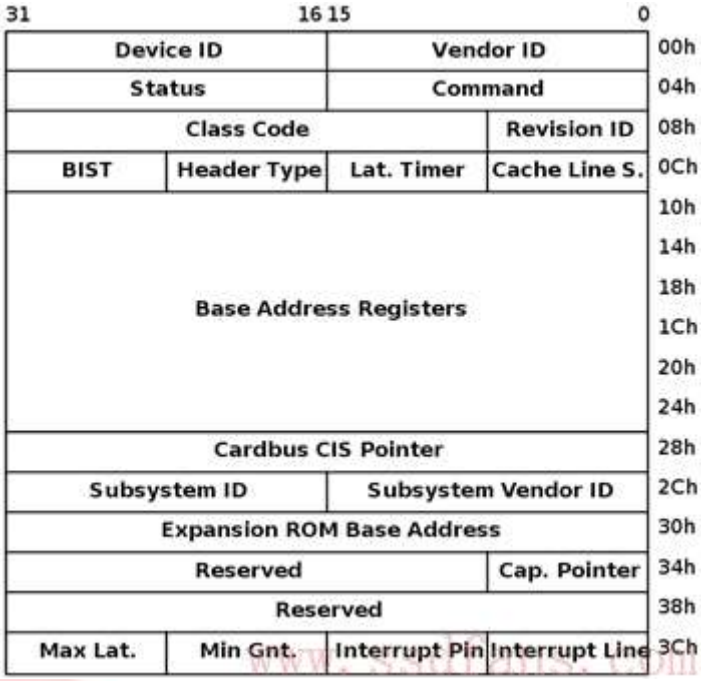
引用

https://github.com/nvmeqemu

10.4.6 阿呆实战 NVMe 之六

上文中我们提到 NVMe 设备其实是个 PCI 设备，里面用了 bar0 和 MMIO，但是不了解 PCI 的人肯定对这两个概念有所疑惑。阿呆本想写一些，不过发现网上有篇文章讲的很透彻，所以就不班门弄斧了，友情转载过来。

PCI 设备(PCI device)都有一个配置空间，大小为 256 字节，实际上是一组连续的寄存器，位于设备上。其中头部 64 字节是 PCI 标准规定的，格式如下：



剩余的部分是 PCI 设备自定义的。

PCI 配置空间头部有 6 个 BAR(Base Address Registers)，BAR 记录了设备所需要的地址空间的类型(memory space 或者 I/O space)，基址以及其他属性。BAR 的格式如下：

Memory Space BAR Layout

31 - 4	3	2 - 1	0
16-Byte Aligned Base Address	Prefetchable	Type	Always 0

I/O Space BAR Layout

31 - 2	1	0
4-Byte Aligned Base Address	Reserved	Always 1

可以看出，设备可以申请两类地址空间，memory space 和 I/O space，它们用 BAR 的最后一位区别开来。

说到地址空间，计算机系统中，除了我们常说的 memory address(包括逻辑地址、虚拟地址(线性地址)、CPU 地址(物理地址))，还有 I/O address，这是为了访问 I/O 设备(主要是设备中的寄存器)而设立的，大部分体系结构中，memory address 和 I/O address 都是分别编址的，且使用不同的寻址指令，构成了两套地址空间，也有少数体系结构将 memory address 和 I/O address 统一编址(如 ARM)。

有两套地址空间并不意味着计算机系统中需要两套地址总线，实际上，memory address 和 I/O address 是共用一套地址总线，但通过控制总线上的信号区别当前地址总线上的地址是 memory address 还是 I/O address。北桥芯片(Northbridge，Intel 称其 Memory Controller Hub，MCH)负责地址的路由工作，它内部有一张 address map，记录了 memory address，I/O address 的映射信息，一个典型的 address map 如图：

70h	DRAM Base 6		RW
74h	DRAM Limit 6		RW
78h	DRAM Base 7		RW
7Ch	DRAM Limit 7		RW
80h	Memory-Mapped I/O Base 0		RW
84h	Memory-Mapped I/O Limit 0		RW
88h	Memory-Mapped I/O Base 1		RW
8Ch	Memory-Mapped I/O Limit 1		RW
90h	Memory-Mapped I/O Base 2		RW
94h	Memory-Mapped I/O Limit 2		RW
98h	Memory-Mapped I/O Base 3		RW
9Ch	Memory-Mapped I/O Limit 3		RW
A0h	Memory-Mapped I/O Base 4		RW
A4h	Memory-Mapped I/O Limit 4		RW
A8h	Memory-Mapped I/O Base 5		RW
ACh	Memory-Mapped I/O Limit 5		RW
B0h	Memory-Mapped I/O Base 6		RW
B4h	Memory-Mapped I/O Limit 6		RW
B8h	Memory-Mapped I/O Base 7		RW
BCh	Memory-Mapped I/O Limit 7		RW
C0h	PCI I/O Base 0		RW
C4h	PCI I/O Limit 0		RW
C8h	PCI I/O Base 1		RW
CCh	PCI I/O Limit 1		RW

我们来看北桥是如何进行地址路由的。根据控制总线上的信号，北桥首先可以识别地址属于 memory space 还是 I/O space，然后分别做处理。

比如若是 memory space，则根据 address map 找出目标设备(DRAM 或 Memory Mapped I/O)，若是 DRAM 或 VGA，则转换地址然后发送给内存控制器或 VGA 控制器，若是其它 I/O 设备，则发送给南桥。

若是 I/O space，则发送给南桥(Southbridge，Intel 称其 I/O Controller Hub，ICH)，南桥负责解析出目标设备的 bus, device, function 号，并发送信息给它。

PCI 设备会向计算机系统申请很多资源，比如 memory space, I/O space, 中断请求号等，相当于在计算机系统中占位，使得计算机系统认识自己。

PCI 设备可以通过两种方式将自己的 I/O 存储器(Registers/RAM/ROM)暴露给 CPU：

在 memory space 申请地址空间，或者在 I/O space 申请地址空间。

这样，PCI 设备的 I/O 存储器就分别被映射到 CPU-relative memory space 和 CPU-relative I/O space，使得驱动以及操作系统得以正常访问 PCI 设备。对于没有独立 I/O space 的体系结构(如 ARM)，memory space 和 I/O space 是统一编址的，也就是说 memory space 与 I/O space 等价了，这时，即使 PCI 设备在 BAR 表明了要申请 I/O space，实际上也是分配在 memory space 的，所以驱动无法使用 I/O 端口指令访问 I/O，只能使用访存指令。在 Windows 驱动开发中，PCM_PARTIAL_RESOURCE_DESCRIPTOR 记录了为 PCI 设备分配的硬件资源，可能有 CmResourceTypePort, CmResourceTypeMemory 等，后者表示一段 memory 地址空间，顾名思义，是通过 memory space 访问的，前者表示一段 I/O 地址空间，

但其 flag 有 CM_RESOURCE_PORT_MEMORY 和 CM_RESOURCE_PORT_IO 两种，分别表示通过 memory space 访问以及通过 I/O space 访问，这就是 PCI 请求与实际分配的差异，在 x86 下，CmResourceTypePort 的 flag 都是 CM_RESOURCE_PORT_IO，即表明 PCI 设备请求的是 I/O 地址空间，分配的也是 I/O 地址空间，而在 ARM 或 Alpha 等下，flag 是 CM_RESOURCE_PORT_MEMORY，表明即使 PCI 请求的 I/O 地址空间，但分配在了 memory space，我们需要通过 memory space 访问 I/O 设备(通过 MmMapIoSpace 映射物理地址空间到虚拟地址空间，当然，是内核的虚拟地址空间，这样驱动就可以正常访问设备了)。

为了为 PCI 设备分配 CPU-relative space，计算机系统需要知道其所申请的地址空间的类型、基址等，这些信息记录在设备的 BAR 中，每个 PCI 配置空间拥有 6 个 BAR，因此每个 PCI 设备最多能映射 6 段地址空间(实际很多设备用不了这么多)。PCI 配置空间的初始值是由厂商预设在设备中的，于是设备需要哪些地址空间都是其自己定的，可能造成不同的 PCI 设备所映射的地址空间冲突，因此在 PCI 设备枚举(也叫总线枚举，由 BIOS 或者 OS 在启动时完成)的过程中，会重新为其分配地址空间，然后写入 PCI 配置空间中。

通过 memory space 访问设备 I/O 的方式称为 memory mapped I/O，即 MMIO，这种情况下，CPU 直接使用普通访存指令即可访问设备 I/O。

通过 I/O space 访问设备 I/O 的方式称为 port I/O，或者 port mapped I/O，这种情况下 CPU 需要使用专门的 I/O 指令如 IN/OUT 访问 I/O 端口。

常见的 MMIO 例子有，VGA card 将 framebuffer 映射到 memory space，NIC 将自己的片上缓冲映射到 memory space，实际上，最典型的 MMIO 应该是 DRAM，它将自己的存储空间映射到 memory space，是占用 CPU 地址空间最多的”设备”。

转载自

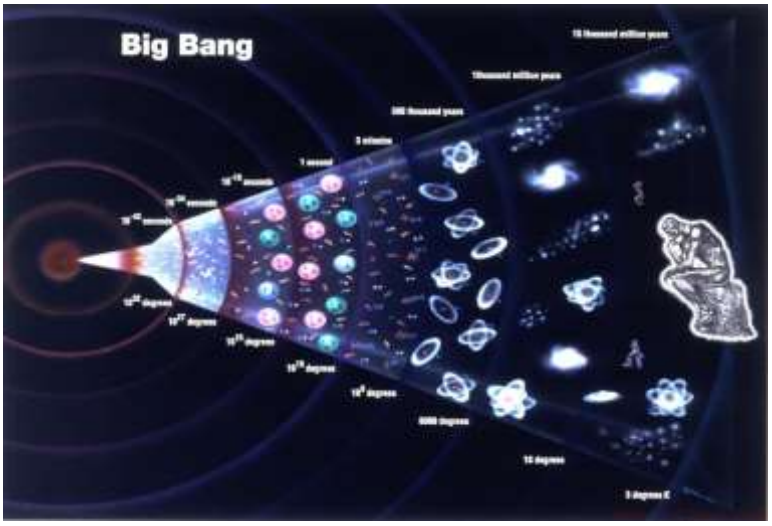
作者：使命召唤

@jilinxpd， *http://www.cnblogs.com/zszmhd/archive/2012/05/08/2490105.html*

10.4.7 阿呆实战 NVMe 之七

Posted on 2017 年 8 月 3 日 by SSD Fans

原创内容，转载请注明： [http://www.ssdfans.com] 谢谢！



提要

本系列文章，旨在带你开发一个 NVMe SSD 控制器的前端协议逻辑，只不过是在 QEMU 虚拟机环境中。前面我们介绍了 QEMU 中 PCI 设备的初始化代码，还有 NVMe 设备的主要描述类 NVMEState 的每个变量。本文来看看 NVMe 设备的初始化，不过从本文开始，阿呆将少贴一些代码，关于编程技巧方面的就不多写了，毕竟无关宏旨，我们重点还是关注 NVMe 协议的实现。

话说最近几十年来，很多聪明的科学家分析了许许多多的宇宙观测数据，最后得出一个惊人的结论：我们的宇宙是由一个致密炽热的奇点于 137 亿年前一次大爆炸后膨胀形成的。爆炸之初，物质只能以中子、质子、电子、光子和中微子等基本粒子形态存在。宇宙爆炸之后的不断膨胀，导致温度和密度很快下降。随着温度降低、冷却，逐步形成原子、原子核、分子，并复合成为通常的气体。气体逐渐凝聚成星云，星云进一步形成各种各样的恒星和星系，最终形成我们如今所看到的宇宙。

看得出来，宇宙一开始的时候定义了一些物理定律，同时赋予初始值，从此就有了丰富多彩的宇宙。我们的 NVMe 设备也是如此，在初始化的时候主要还是设定一些机制，同时为各种变量赋初值，这些基本的机制从此开始循环运转，满足用户实现各种各样的需求。

当然，宇宙并没有循环运转，《三体》的结尾告诉我们，大爆炸最终是要塌缩到奇点的，只不过被人为阻碍了，人类灭绝，只有程心和关一帆携手走到了新的宇宙。

NVMe 控制器初始化

初始化函数 pci_nvme_init 一进来，还是老套路，获得 NVMEState 指针。

```
1 static int pci_nvme_init(PCIDevice *pci_dev)
2 {
3     NVMEState *n = DO_UPCAST(NVMEState, dev, pci_dev);
```

接着，创建我们虚拟的 SSD 磁盘。看得出来，为每个 NVMe namespace 创建了一个 SSD，由此我们也可以理解 namespace 的用法，就是把 1 个 NVMe 出来的 SSD 划分成几个独立的 SSD，使用 namespace 来管理，在上层看来就有很多个盘，大家各干各的，互不干扰。真所谓鸡犬之声相闻，老死不相往来。

```
n->disk = (DiskInfo *)qemu_malloc(sizeof(DiskInfo)*n->num_namespaces);
```

接下来就是初始化 SQ，CQ 和 Admin Q。然后要从配置文件 NVME_device_PCI_config 把 PCI 相关寄存器的定义和参数加载进来。文件内容，举个例子：

```
1 <REG>
2     CFG_NAME = PCIHEADER
```

```
    NAME = "CMD"
    OFFSET = 0x04
    LENGTH = 0x02
    VALUE = 0x03
    RO_MASK = 0xFBF8
    RW_MASK = 0x0447
    RWC_MASK = 0x0
     RWS_MASK = 0x0
     DESC = "Command Register"
</REG>
```

看得出来，定义了偏移地址，长度，还有前文说过的各种 MASK。

MSI 中断初始化

接下来就是 MSI 中断初始化。哎，等等等等，MSI？什么东东？首先我们来看 PCI 设备的中断机制，是有一些物理的中断信号，与处理器的中断控制器相连，通过电平变化触发中断。

PCI 总线 V2.2 规范还定义了一种新的中断机制，即 MSI 中断机制。MSI 中断机制采用存储器写总线事务向处理器系统提交中断请求，其实现机制是向 HOST 处理器指定的一个存储器地址写指定的数据。这个存储器地址一般是中断控制器规定的某段存储器地址范围，而且数据也是事先安排好的数据，通常含有中断向量号。其实就是文明在进步，不玩硬的了，来软的。不用电平触发，而是往主机内存写段数据就能触发中断了。

HOST 主桥会将 MSI 这个特殊的存储器写总线事务进一步翻译为中断请求，提交给处理器。目前 PCIe 和 PCI-X 设备必须支持 MSI 中断机制，但是 PCI 设备并不一定都支持 MSI 中断机制。

MSI 中断机制虽然在 PCIe 总线上已经成为主流，但是在 PCI 设备中并不常用。即便是支持 MSI 中断机制的 PCI 设备，在设备驱动程序的实现中也很少使用这种机制。首先 PCI 设备具有 INTx#信号可以传递中断，而且这种中断传送方式在 PCI 总线中根深蒂固。其次 PCI 总线是一个共享总线，传递 MSI 中断需要占用 PCI 总线的带宽，需要进行总线仲裁等一系列过程，远没有使用 INTx#信号线直接。不过我们这里说的是 NVMe，就不考虑 PCI 了，而是关注 PCIe。

QEMU 里面用的是 MSI 的升级版 MSIX，MSI-X 中断机制与 MSI 的中断机制类似。PCIe 总线引出 MSI-X 机制的主要目的是为了扩展 PCIe 设备使用中断向量的个数，同时解决 MSI 中断机制要求使用中断向量号连续所带来的问题。

QEMU 中 NVMe 的 MSIX 中断向量初始化代码如下，其实就是申请了中断向量，页和前文介绍的 MMIO。

```
dev->msix_entry_used = qemu_mallocz(MSIX_MAX_ENTRIES *
                                    sizeof *dev->msix_entry_used);

dev->msix_table_page = qemu_mallocz(MSIX_PAGE_SIZE);
msix_mask_all(dev, nentries);

dev->msix_mmio_index = cpu_register_io_memory(msix_mmio_read,
                                      msix_mmio_write, dev,
                                      DEVICE_NATIVE_ENDIAN);
```

引用

https://github.com/nvmeqemu

sailing，《浅谈 PCIe 体系结构》，http://blog.sina.com.cn/s/blog_6472c4cc0100qfau.html

10.4.8 阿呆实战 NVMe 之八

Posted on 2017 年 8 月 3 日 by SSD Fans

原创内容，转载请注明： [http://www.ssdfans.com] 谢谢！

提要

本系列文章，旨在带你开发一个 NVMe SSD 控制器的前端协议逻辑，只不过是在 QEMU 虚拟机环境中。

澄清关于 PCIe BAR 空间的误解

前面我们说过了 NVMe 的 BAR 空间是个什么东东，结果在群里就被山哥指出来理解错误，BAR 寄存器里面的地址并不是 CPU 的存储器域空间，而是 PCI 域，只是因为 X86 架构这两个用一样的地址，所以给人一种误解。到了 PowerPC 处理器，就是不一样的地址，需要 Host PCI 桥或者 PCIe 的 RC 做地址转换。

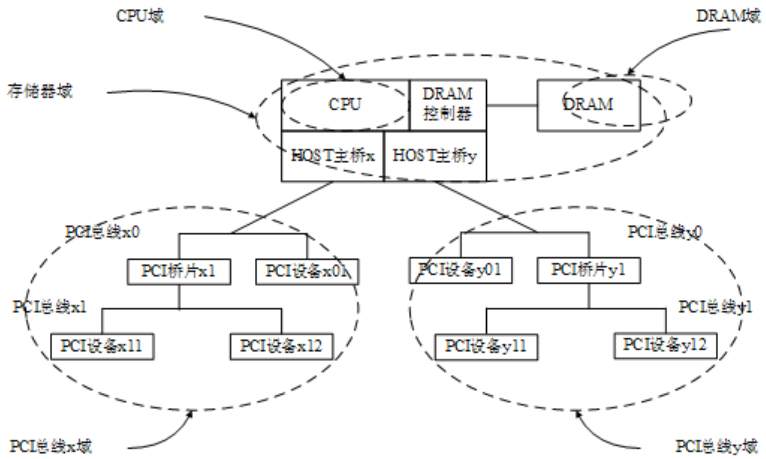


图 2-1 存储器域与 PCI 总线域的划分

上图所示的处理器系统由一个 CPU，一个 DRAM 控制器和两个 HOST 主桥组成。在这个处理器系统中，包含 CPU 域、DRAM 域、存储器域和 PCI 总线域地址空间。其中 HOST 主桥 x 和 HOST 主桥 y 分别管理 PCI 总线 x 域与 PCI 总线 y 域。PCI 设备访问存储器域时，也需要通过 HOST 主桥，并由 HOST 主桥进行 PCI 总线域到存储器域的地址转换；CPU 访问 PCI 设备时，同样需要通过 HOST 主桥进行存储器域到 PCI 总线域的地址转换。

Memory Address 和 IO Address

再来回顾一下之前的两个概念，PCIe 设备可以申请两类地址空间，memory space 和 I/O space，它们用 BAR 的最后一位区别开来。

说到地址空间，计算机系统中，除了我们常说的 memory address(包括逻辑地址、虚拟地址(线性地址)、CPU 地址(物理地址))，还有 I/O address，这是为了访问 I/O 设备(主要是设备中的寄存器)而设立的，大部分体系结构中，memory address 和 I/O address 都是分别编址的，且使用不同的寻址指令，构成了两套地址空间，也有少数体系结构将 memory address 和 I/O address 统一编址(如 ARM)。

有两套地址空间并不意味着计算机系统中需要两套地址总线，实际上，memory address 和 I/O address 是共用一套地址总线，但通过控制总线上的信号区别当前地址总线上的地址是 memory address 还是 I/O address。

QEMU 中 NVMe BAR 空间初始化

我们知道 QEMU 虚拟机其实本没有物理内存，他的内存是 QEMU 从宿主机的内存中虚拟出来的。我们虚拟的设备要获得内存地址或者 IO 地址，就需要向 QEMU 注册申请。我们这里 NVMe 是需要 IO 地址，所以首先通过 cpu_register_io_memory 注册，获得 io_mem_write/io_mem_read 的索引。该索引还有需要的 IO 空间大小等参数传给 cpu_register_physical_memory 函数，获得 QEMU 虚拟机的 IO 地址空间。

对于 PCI 设备来说，IO 地址注册就要多一步，因为要进行 PCI bar 地址与 IO 的映射，所以必须先调用下面函数来给 bar 注册 PCI 地址。关键参数说明：第一个是 PCI 设备指针，第三个是我们需要注册 IO 地址的空间长度，最后一个是我们要进行 IO 操作映射的初始化函数指针。

```
1    /* Register BAR 0 (and bar 1 as it is 64bit). */
2    pci_register_bar((struct PCIDevice *)&n->dev,
3        0, ((n->dev.cap_present & QEMU_PCI_CAP_MSIX) ?
4        n->dev.msix_bar_size : n->bar0_size),
5        (PCI_BASE_ADDRESS_SPACE_MEMORY ,
6        PCI_BASE_ADDRESS_MEM_TYPE_64),
7        nvme_mmio_map);
```

IO 初始化函数如下，关键参数说明：第一个依然是 PCI 设备指针，第三个是 PCI 地址映射的 PIO 起始地址，这个起始地址是在上面的函数注册 PCI 地址的时候，PCI 总线通过计算比较 PIO 地址空间得到的一个 PIO 地址起始空间。所以在我们注册设备 PIO 空间的时候必须将这个地址作为注册 IO 空间的起始地址。这个函数是在更新 bar 映射的时候被调用的，里面调用了前面说的 cpu_register_physical_memory，其实就是 QEMU 从宿主机申请一段内存用来映射到虚拟机。

```
1 static void nvme_mmio_map(PCIDevice *pci_dev, int reg_num, pcibus_t addr,
2                           pcibus_t size, int type)
```

在阿呆看的版本中，MSIX 中断向量数为 32 个，BAR 0 空间大小为 8KB，NVMe 队列为 64 个（包括 Admin 队列）。

NVMe 设备参数初始化

关键的工作搞定，后面的工作就简单了。首先是类似于 PCI 设备，从 NVME_device_NVME_config 文件读取 NVMe 相关寄存器的参数给对应变量赋值。例如：

```
1 <REG>
2    CFG_NAME = CNTRLREG
3    NAME = "AQA"
4    OFFSET = 0x24
5    LENGTH = 0x04
6    VALUE = 0x00000000
7    RO_MASK = 0xF000F000
8    RW_MASK = 0x0FFF0FFF
9    RWC_MASK = 0x00000000
10    RWS_MASK = 0x00000000
11    DESC = "Admin Queue Attributes"
12 </REG>
```

接下来的内容比较杂，主要是：

- 是各种配置参数初始化，比如温度的阈值，队列数。
- 给每个 namespace 的 Identify 页赋值。

- 一些 Log 页的赋值。
- SQ 处理和 Async 事件处理两个 Timer 的初始化。

Timer 初始化之后，其实 NVMe 设备就开始工作了，能响应主机的需求。

SSD 模拟器初始化

SSD 其实就是个存储空间，所以 QEMU 用了文件来模拟。读写 SSD 相当于读写宿主机上的一些文件。但是读写文件不太方便，所以这里使用了 Linux 的 mmap 函数，这个函数把文件映射到一个地址，后来写这个文件就跟写内存一样方便，可以按照偏移地址去写。

引用

https://github.com/nvmeqemu

sailing，《浅谈 PCIe 体系结构》，http://blog.sina.com.cn/s/blog_6472c4cc0100qfau.html

http://blog.csdn.net/yearn520/article/details/6560851

http://people.cs.nctu.edu.tw/~chenwj/dokuwiki/doku.php?id=qemu

10.4.9 阿呆实战 NVMe 之九

Posted on 2017 年 8 月 3 日 by SSD Fans

提要

本系列文章，旨在带你开发一个 NVMe SSD 控制器的前端协议逻辑，只不过是在 QEMU 虚拟机环境中。万事开头难，前面我们花了 8 篇文章来写 NVMe 在 QEMU 中的初始化，一方面说明初始化的重要和琐碎，另一方面也暴露了阿呆的水平不够，要是真完全掌握的人必然是高屋建瓴，言简意赅。阿呆学艺不精，只能一个函数一个函数来带你逛 NVMe 的集市了。

本篇我们来看看 NVMe 设备在 QEMU 的 Doorbell 如何实现，探探蛋蛋说过的大宝，阿呆改名成贴心暖男——大白。

NVMe 寄存器配置流程

前面我们讲到，NVMe 初始化的时候注册了 MMIO，如下，两个函数 nvme_mmio_read, nvme_mmio_write 就是来做 BAR 空间的读写，对 BAR 空间的读写最终都会调用这兄弟俩。

```
1 /* NVMe is Little Endian. */
2 n->mmio_index = cpu_register_io_memory(nvme_mmio_read, nvme_mmio_write,
3     n,  DEVICE_LITTLE_ENDIAN);
```

再来看看 NVMe 的 BAR 空间地址分配，如下，看得出来，前面 0x1000 之前都是寄存器空间。自然，每当 Host 来读写这些地址的时候，其实就是在读写对应的寄存器。0x1000 之后是每个 SQ，CQ 队列的大白（门铃，doorbell）的地盘。

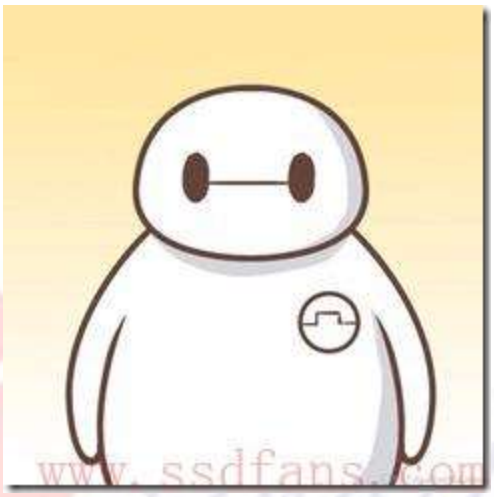
```
1 /* NVMe Controller Registers */
2 enum {
3     NVME_CAP      = 0x0000, /* Controller Capabilities, 64bit */
4 // 各种寄存器
5     NVME_CMD_SS    = 0x0F00, /* Command Set Specific*/
6     NVME_SQ0TDBL   = 0x1000, /* SQ 0 Tail Doorbell, 32bit (Admin) */
7     NVME_CQ0HDBL   = 0x1004, /* CQ 0 Head Doorbell, 32bit (Admin)*/
8     NVME_SQ1TDBL   = 0x1008, /* SQ 1 Tail Doorbell, 32bit */
```

```
 9     NVME_CQ1HDBL   = 0x100c, /* CQ 1 Head Doorbell, 32bit */
10
11     NVME_SQMAXTDBL = (NVME_SQ0TDBL + 8 * NVME_MAX_QID),
12     NVME_CQMAXHDBL = (NVME_CQ0HDBL + 8 * NVME_MAX_QID)
13 };
```

对于每个寄存器，QEMU 都有对应变量，所以调用前面的都写函数读写寄存器都会反映到这些内部变量，例如：

```
1 case NVME_ASQ:
2     nvme_cntrl_write_config(nvme_dev, NVME_ASQ, val, DWORD);
3     *((uint32_t *) &nvme_dev->sq[ASQ_ID].dma_addr) = val;
4     break;
```

大白是干啥的？



如果调用 nvme_mmio_write 写 SQ 的大白地址段，大白就开始工作了。大白工作的办公室名字叫 process_doorbell，参数如下。每次门铃一响，大白就通过地址 addr 推算出按的是哪个队列的门铃，其中偶数号的是 SQ，Host 给 device 的命令队列，奇数号的是 CQ，device 发给 host 的响应队列。

```
1 static void process_doorbell(NVMEState *nvme_dev, target_phys_addr_t addr,
2     uint32_t val)
```

先转一下蛋蛋总结的 SQ 和 CQ 的作用：

- SQ 用以 Host 发命令，CQ 用以 SSD 回命令完成状态
- SQ/CQ 在 Host 内存中;
- 两种类型的 SQ/CQ：Admin 和 I/O，前者发送 Admin 命令，后者发送 I/O 命令;
- 系统中只能有一对 Admin SQ/CQ，但可以有很多对 I/O SQ/CQ;
- I/O SQ 与 CQ 可以是一对一的关系，也可以是一对多的关系；
- I/O SQ 是可以赋予不同优先级的；
- I/O SQ/CQ 深度可达 64K，Admin SQ/CQ 深达 4K；
- I/O SQ/CQ 的广度和深度都可以灵活配置；
- 每条命令大小是 64 字节，每条命令完成状态是 16 字节；
- 不要过河拆桥。

再来看蛋蛋对大白的定义：“doorbell，DB，就是用来记录了一个 SQ 或者 CQ 的 Head 和 Tail。每个 SQ 或者 CQ，都有两个对应的 DB: Head DB 和 Tail DB。DB 是在 SSD 端的寄存器，记录 SQ 和 CQ 的头和尾巴的位置。”

一句话太简单，阿呆继续友情转载一大段：

“那么，DB 在命令处理流程中起了什么作用呢？

首先，如前所示，它记住了 SQ 和 CQ 的头和尾。对 SQ 来说，SSD 是消费者，它直接和队列的头打交道，很清楚 SQ 的头在哪里，所以 SQ head DB 由 SSD 自己维护；但它不知道队伍有多长，尾巴在哪，后面还有多少命令等待执行，相反，Host 知道，所以 SQ Tail DB 由 Host 来更新。SSD 结合 SQ 的头和尾，就知道还有多少命令在 SQ 中等待执行了。对 CQ 来说，SSD 是生产者，它很清楚 CQ 的尾巴在哪里，所以 CQ Tail DB 由自己更新，但是 SSD 不知道 Host 处理了多少条命令完成信息，需要 Host 告知，因此 CQ Head DB 由 Host 更新。SSD 根据 CQ 的头和尾，就知道 CQ 能不能以及能接受多少命令完成信息。

DB 的另外一个作用，就是通知作用：Host 更新 SQ Tail DB 的同时，也是在告知 SSD 有新的命令需要处理；Host 更新 CQ Head DB 的同时，也是在告知 SSD，你返回的命令完成状态信息我已经处理，同时表示谢意。”

现实中的大白

看得出来，SQ 的大白有消息，就意味着新的命令来了，赶快记下来，找人开工。CQ 的大白有消息，就意味着上次汇报之后，老板已经知道 SSD 最近干了什么活，所以需要把这些记录清理掉，以后不用汇报了。

我们来到基层实际看看上面的理论是怎么操作的。首先看 CQ 队列，只看核心的代码：

```
1 uint16_t new_head = val & 0xffff;
2 queue_id = (addr - NVME_CQ0HDBL) / QUEUE_BASE_ADDRESS_WIDTH;
3 nvme_dev->cq[queue_id].head = new_head;
4
5 if (nvme_dev->cq[queue_id].tail != nvme_dev->cq[queue_id].head) {
6     /* more completion entries, submit interrupt */
7     isr_notify(nvme_dev, &nvme_dev->cq[queue_id]);
8 }
```

可以看出来，通过大白的地址算出来队列编号 queue_id，接着更新 CQ 的 head 到上级领导写下来的位置，意味着下次从这个新的位置开始汇报工作。最后，再检查一下，是不是还有没处理完的 CQ，如果有，就触发中断，申请上级领导在百忙之中抽出时间听取最新工作汇报。

接下来，看看 SQ 队列的大白要干些啥子事情。计算队列编号和拿寄存器数据和 CQ 一样，只不过寄存器内容变成了队列的尾巴。因为写从 tail 往后填，读从 head 开始拿，所以领导往 SQ 尾巴填命令，SSD 就得知道尾巴填到哪里了。大白发现上级领导有新的工作任务发下来了，就设置定时器，5 微秒之后让负责的人执行。可见分工很严密啊，大白忠心耿耿，所以只负责传达命令。

```
 1 uint16_t new_tail = val & 0xffff;
 2 queue_id = (addr - NVME_SQ0TDBL) / QUEUE_BASE_ADDRESS_WIDTH;
 3 nvme_dev->sq[queue_id].tail = new_tail;
 4
 5 deadline = qemu_get_clock_ns(vm_clock) + 5000;
 6
 7 if (nvme_dev->sq_processing_timer_target == 0) {
 8     qemu_mod_timer(nvme_dev->sq_processing_timer, deadline);
 9     nvme_dev->sq_processing_timer_target = deadline;
10 }
```

大白的活是干完了，但是设置了定时器，到时候活得有人干啊，下文我们就来看看到底是谁在处理 SQ。

引用
https://github.com/nvmeqemu

10.4.10 阿呆实战 NVMe 之十

Posted on 2017 年 8 月 3 日 by SSD Fans

原创内容，转载请注明： [http://www.ssdfans.com] 谢谢！

提要

本系列文章，旨在带你开发一个 NVMe SSD 控制器的前端协议逻辑，只不过是在 QEMU 虚拟机环境中。上回说到了大白接受领导写到 SQ 队列的命令，设定计时器，让下面的小弟到时间后执行任务。本篇就来看 SQ 的命令怎么执行。你知道大部分 SSD 公司的工程师是工作在哪个 SQ 队列吗？阿呆将告诉你惊人的真相。

SQ 的定时回调函数

其实我们在 NVMe 初始化的时候，就注册了 SQ 计时器对应的回调函数，sq_processing_timer_cb，所以计时器时间一到，这个函数就被调用了。这个函数的任务很简单，就是把所有的 SQ 队列扫描一遍，只要那个队列 Head 和 Tail 不相等，就意味着还有活没干完，就调用干活的函数 process_sq 去料理一下。如果料理完，还有没干完的活，说明在干活过程中勤快的大白又分派领导下发的命令了，所以再设定一个定时器，稍后继续干。唉，大白啊大白，怪不得领导那么信任你，谁叫你那么勤快呢？就是苦了俺们这帮干活的函数，每次听大白门铃一响，就不敢偷懒了：赶快干完活回家带孩子。

看得出来，process_sq 是执行任务的核心，SQ 有 Admin 队列和普通 IO 队列，这些都要在这个函数去区分执行。核心流程很清晰，如下，就是如果 Admin 队列，就调用 nvme_admin_command 执行 Admin 命令，IO 队列，就调用 nvme_command_set 执行。执行完成后，填响应的内容到 CQ 条目中。post_cq_entry 函数会把 cqe 的内容复制到 CQ 队列的尾巴上，并触发中断，向上级领导汇报工作成果。

```
1 if (sq_id == ASQ_ID) {
2    nvme_admin_command(n, &sqe, &cqe);
3 } else {
4   nvme_command_set(n, &sqe, &cqe);
5 }
6
```

```
 7 /* Filling up the CQ entry */
 8 cqe.sq_id = sq_id;
 9 cqe.sq_head = n->sq[sq_id].head;
10 cqe.command_id = sqe.cid;
11 post_cq_entry(n, &n->cq[cq_id], &cqe);
```

Admin 队列的执行

上面说是所有的 SQ 都扫一遍，那第一个就是 Admin 队列了。nvme_admin_command 的任务很简单，就是看 Admin 命令是什么，就调用对应的处理函数，如下。adm_cmds_funcs 是个静态数组，里面是各种 Admin 命令的处理函数指针。阿呆哥还是决定把所有的函数都贴出来，满足你的好奇心。

```
1 f = adm_cmds_funcs[sqe->opcode];
2 ret = f(n, sqe, cqe);
```

```
 1 static adm_command_func * const adm_cmds_funcs[] = {
 2     [NVME_ADM_CMD_DELETE_SQ] = adm_cmd_del_sq,
 3     [NVME_ADM_CMD_CREATE_SQ] = adm_cmd_alloc_sq,
 4     [NVME_ADM_CMD_GET_LOG_PAGE] = adm_cmd_get_log_page,
 5     [NVME_ADM_CMD_DELETE_CQ] = adm_cmd_del_cq,
 6     [NVME_ADM_CMD_CREATE_CQ] = adm_cmd_alloc_cq,
 7     [NVME_ADM_CMD_IDENTIFY] = adm_cmd_identify,
 8     [NVME_ADM_CMD_ABORT] = adm_cmd_abort,
 9     [NVME_ADM_CMD_SET_FEATURES] = adm_cmd_set_features,
10     [NVME_ADM_CMD_GET_FEATURES] = adm_cmd_get_features,
11     [NVME_ADM_CMD_ASYNC_EV_REQ] = adm_cmd_async_ev_req,
12     [NVME_ADM_CMD_ACTIVATE_FW] = adm_cmd_act_fw,
13     [NVME_ADM_CMD_DOWNLOAD_FW] = adm_cmd_dl_fw,
14     [NVME_ADM_CMD_FORMAT_NVM] = adm_cmd_format_nvm,
15     [NVME_ADM_CMD_LAST] = NULL,
16 };
```

也许你觉得这些命令都很陌生，但是一旦你踏入 NVMe SSD Firmware 或者 QA 职业，那么很有可能，你每天的大部分时间就是在跟上面这些函数打交道，因为相比 SSD 控制器的 FTL 和后端 Flash 命令处理，前端的任务其实更琐碎，更多，而且一旦一个产品开发完成之后，FTL 和后端的活很少，大部分人都要到前端处理 NVMe 的命令了。所以，也许对你来说，Admin 队列才是养家糊口的饭碗，不要小看这些命令，每一个命令都有很多细节的协议，相当复杂，不夸张的说，如果公司大一点，这里面每个命令都可以养一个工程师。我们的社会也是如此，国家刚开始的时候，需要的是工程师领导人，建设祖国，等到经济发达了，大家都有钱了，需要的就是律师型领导人，这些人精通各种法律条文，擅长的不是建设，而是沟通，分配人群的利益，解决纠纷。本文题图就是喜剧明星金凯利演的律师，法律越复杂，律师的饭碗就越稳，协议越复杂，FW 和 QA 的饭碗就越稳，总是有干不完的活啊！为什么阿呆这么清楚，因为阿呆当年刚入行的时候也是每天靠 ATA 协议吃饭，搞各种 Smart Log，GPL Log 之类，忙得不亦乐乎。

对用户来说，这些函数都挺重要的，因为它们是用户观察 SSD 内部的窗口。比如，我们要了解 SSD 内部情况，经常要查看 Smart 信息，处理 Smart 的函数流程是 adm_cmd_get_log_page->adm_cmd_smart_info，里面是各种 SSD 的统计数据，比如 Host 读写命令的个数等。

PRP 的原理

先来补补课，回顾一下《蛋蛋读 NVMe 之三》里面讲的 PRP，这个是 Host 内存和 SSD 数据交互的内存页管理结构。

“NVMe 把 Host 的内存划分为一个一个页（Page），页的大小可以是 4KB,8KB,16KB…128MB。

PRP 是什么，长什么样呢？

Figure 14: PRP Entry Layout

PRP Entry 本质就是一个 64 位内存物理地址，只不过把这个物理地址分成两部分：页起始地址和页内偏移。最后两 bit 是 0，说明 PRP 表示的物理地址只能四字节对齐访问。页内偏移可以是 0，也可以是个非零的值。

PRP Entry 描述的是一段连续的物理内存的起始地址。如果需要描述若干个不连续的物理内存呢？那就需要若干个 PRP Entry。把若干个 PRP Entry 链接起来，就成了 PRP List。

Figure 16: PRP List Layout

63 … n+1	n … 0
Page Base Address *k*	0h
Page Base Address *k+1*	0h
…	
Page Base Address *k+m*	0h
Page Base Address *k+m+1*	0h

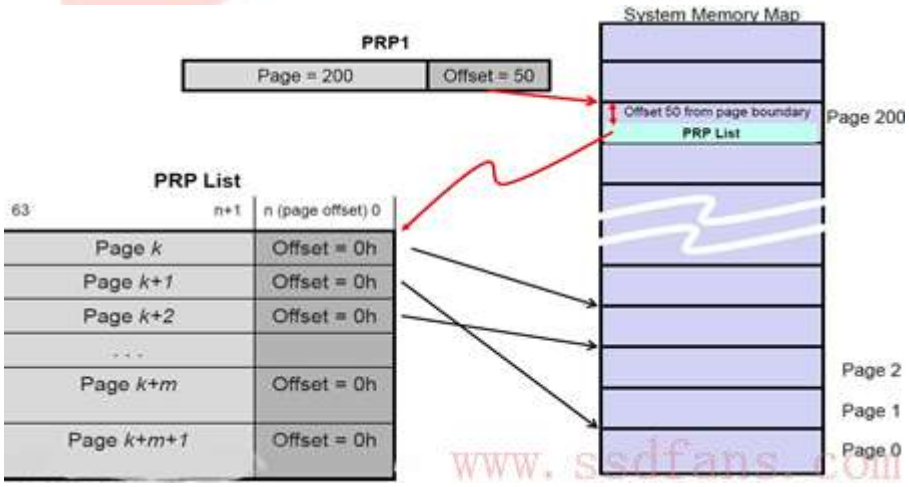
是的，正如你所见，PRP List 中的每个 PRP Entry 的偏移量都必须是 0，PRP List 中的每个 PRP Entry 都是描述一个物理页。它们不允许有相同的物理页，不然 SSD 往同一个物理页写入几次的数据，导致先写入的数据被覆盖。

每个 NVMe 命令中有两个域：PRP1 和 PRP2，Host 就是通过这两个域告诉 SSD 数据在内存中的位置或者数据需要写入的地址。

Bytes	Description
63:60	**Command Dword 15 (CDW15)**: command specific
59:56	**Command Dword 14 (CDW14)**: command specific
55:52	**Command Dword 13 (CDW13)**: command specific
51:48	**Command Dword 12 (CDW12)**: command specific
47:44	**Command Dword 11 (CDW11)**: command specific
43:40	**Command Dword 10 (CDW10)**: command specific
39:32	**PRP Entry 2 (PRP2)**: The 2nd address entry for commands that use it.
31:24	**PRP Entry 1 (PRP1)**: The first address entry for this command.
23:16	**Metadata Pointer (MPTR)**: Address of a contiguous metadata buffer.
15:8	Reserved
7:4	**Namespace Identifier (NSID)**: Namespace for this command. A value of 0h means NSID isn't used for this command; value of all F's means it applies to all namespaces on the device.
3:0	**Command Dword 0 (CDW0)**: Used by all commands

PRP1 和 PRP2 有可能指向数据所在位置，也可能指向 PRP List。类似 C 语言中的指针概念，PRP1 和 PRP2 可能是指针，也可能是指针的指针，还有可能是指针的指针的指针。别管你包的有多严实，根据不同的命令，SSD 总能一层一层的剥下包装，找到数据在内存的真正物理地址。SSD 善解人衣。

下面是一个 PRP1 指向 PRP List 的示例：



PRP1 指向一个 PRP List，PRP List 位于 Page 200，页内偏移 50 的位置。SSD 确定 PRP1 是个指向 PRP List 的指针后，就会去 Host 内存中（Page 200，Offset 50）把 PRP List 取过来。获得 PRP List 后，就获得数据的真正物理地址，SSD 然后就会往这些物理地址读入或者写入数据。”

读写 IO 命令执行

nvme_command_set 函数内容如下，看得出来，包括读写 IO 命令和 Trim 命令：

```
1 if (sqe->opcode == NVME_CMD_READ ,, (sqe->opcode == NVME_CMD_WRITE)){ // 读写 IO 命令
2   return nvme_io_command(n, sqe, cqe);
3 } else if (sqe->opcode == NVME_CMD_DSM) { // Data Set Management，其实这就是 Trim 命令
4   return nvme_dsm_command(n, sqe, cqe);
5 } else if (sqe->opcode == NVME_CMD_FLUSH) {
6   return NVME_SC_SUCCESS;
```

一说到读写，你或许会觉得很复杂，不过阿呆要善意的提醒一句：这里是 QEMU 虚拟机，读写就是个读写文件，操作起来相当的 easy。首先是通过 LBA 地址，计算出文件内将要读写的偏移地址和 LBA 个数。后面就是内存操作了，读就是往内存写数据，写就是从内存读数据。我们前文说过，QEMU 通过 mmap，把文件的操作等同于内存读写，读写文件的某个位置，就是操作指针和偏移地址。

比较绕的反而是 PRP 的操作。核心代码如下，首先对 prp1 进行读写，如果数据还没完，就看数据量是不是在一个 page 内，在的话，只需要读写 prp2 内存地址就可以了，数据量大于 1 个 page，就需要读出 prp list。

```
1 /* Writing/Reading PRP1 */
2 res = do_rw_prp(n, e->prp1, &data_size, &file_offset, mapping_addr,
3   e->opcode);
4
5 if (data_size > 0) {
6   if (data_size <= PAGE_SIZE) {
7     res = do_rw_prp(n, e->prp2, &data_size, &file_offset, mapping_addr,
8       e->opcode);
9   } else {
10    res = do_rw_prp_list(n, sqe, &data_size, &file_offset,
11      mapping_addr);
12  }
13 }
```

do_rw_prp_list 函数的内容和上面的理论一致，首先通过数据大小算出有多少个 prp 条目，然后从 prp2 内存地址读出 prp list，接着遍历每个 prp 条目，读写内存地址，与 SSD 对应的文件指针进行数据搬移。

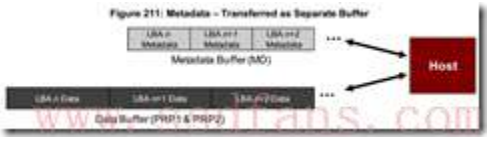
虚拟机的 Trim 命令

其实很简单，就是每个 namespace 维护了一个 bitmap，有 Trim 命令来，就在 bitmap 做个标记而已。

NVMe 的 metadata 机制

这里需要额外提一下，NVMe 支持一种 metadata 机制，就是每个 LBA 有一段 metadata，内容是什么完全看上级领导的心情，可以是校验位，也可以是其他的。metadata 有两种传输方法，一种如下图，紧跟在 LBA 之后连续传输。

另一种是单独传输。请你往上翻翻有张图是 NVMe 命令的解析，紧跟在 PRP1，PRP2 之后的就是 metadata pointer，指向的是 metadata 内存地址的 PRP 信息，所以虚拟机里面也有 metadata 文件来提供 metadata 的读写。代码很简单，就是 DMA 读写，所以阿呆就不复制粘贴了。



大结局

至此，阿呆实战 NVMe 算是写完了，真是虎头蛇尾啊，花了 8 篇写初始化，NVMe 的具体实现却只用了 2 篇。主要还是因为我们是虚拟机，不管 SSD 的具体实现，只负责 NVMe 协议实现。对协议来说，万事开头难，开头初始化流程理清楚了，基本就知道 NVMe 和 PCIe 的虚拟化技术，后面的真正实现反而很简单，水到渠成。

10.5 蛋蛋读 UFS 系列

10.5.1 蛋蛋读 UFS 之一：UFS 简介

Posted on 2018 年 5 月 30 日 by 蛋蛋

原创内容，转载请注明： [http://www.ssdfans.com] 谢谢！

我们知道，我们电脑由三大件组成：CPU，内存和硬盘。CPU 用以计算和控制，内存用以临时存储程序运行时所需的数据（掉电数据丢失），而硬盘用以长久保存数据（掉电数据不丢失）。

我们每天使用的手机，其本质是一个移动的小型计算机，同样由三大件组成：CPU，内存和存储设备。其中的存储设备相当于电脑的硬盘，用以长久保存手机上的数据，比如视频、照片、音乐、系统等数据。

电脑的硬盘有机械硬盘（HDD）和固态硬盘（SSD），前者是机械存储设备，存储介质是磁盘；而后者是电子存储设备，存储介质是闪存。我们不可能在小小的手机中塞入一个机械设备，所以手机上的存储设备只能是电子存储设备，存储介质也都是闪存。

现在是人手一个手机的时代，手机成了人们身体的一部分，一天不带手机，就感觉缺少了什么，吃嘛嘛不香。

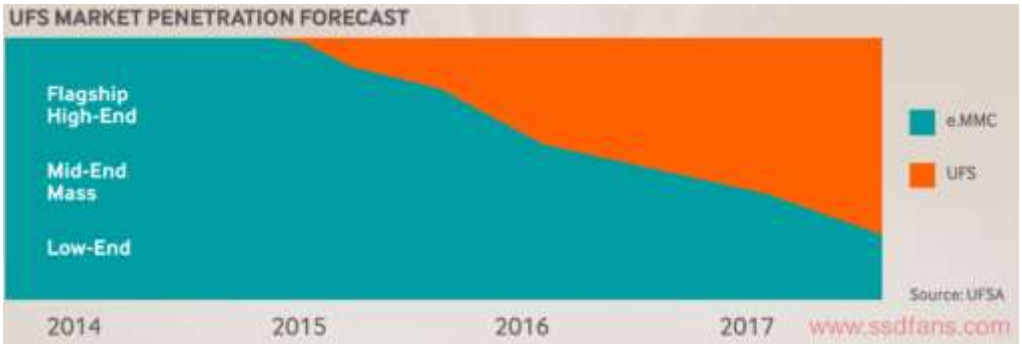
因此，人们对手机的要求也越来越高: 速度要快，容量要大，流畅不卡顿…

为了让手机更快，手机厂商使用更快、更多核的 CPU，加大系统内存（4GB 不够用 6GB，6GB 不够用 8GB），使用更快的存储设备。无论是电脑还是手机，三驾马车（CPU，内存和存储设备）中，跑得最慢的就是存储设备了。CPU 和内存的快步向前，促使最慢的存储设备也需要努力跟上，不然再快的 CPU 和再大容量的内存，你的手机用起来还是让你觉得不爽。

近年来，由于闪存技术的应用和发展，无论是电脑上的硬盘，还是手机中的存储设备，都在变得越来越快。

电脑上，从 HDD 到 SSD，从 SATA SSD 到 PCIe SSD，硬盘是越来越快；

手机上，从 SD 卡，到 eMMC 卡，再到 UFS 卡，存储卡的速度也是越来越快。现在一般手机配的是 eMMC，旗舰高端手机配的是 UFS。



我们这个系列的主角 UFS 已登场。为什么 UFS 是主角？为什么我要带大家去了解 UFS？因为，UFS 将是未来一段时间内手机存储的主流，我们有必要去了解 UFS 以及其相关的技术。

那么，什么是 UFS？Universal Flash Storage，通用闪存存储。它有两个意思，一是指手机存储接口协议，类似 SATA，PCIe/NVMe；二是使用该协议的存储设备。后面文章出现 UFS，读者请根据上下文理解。

为什么说 UFS 是手机存储的未来？无他，快也！

大家感受一下：

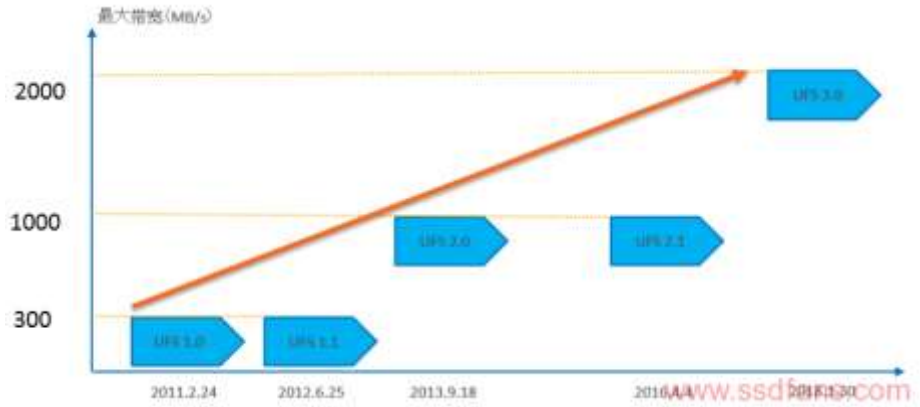
UFS 版本	2.0/2.1	3.0
单通道最大带宽（Mbps）	5830.4	11660.8
最大通道数	2	2
最大有效带宽（MB/s）	1081	2163

注：最大有效带宽去除了协议开销和 8/10 编码开销，二进制速率。

UFS 最新标准是 UFS3.0，于 2018 年 1 月 30 日发布。它最大带宽可以达到 2163MB/s！4 倍 SATA3.0 的速度 （600MB/s），超过 PCIe3.0x2 的速度（2GB/s 单向速度）。

不过，目前市面上的 UFS 产品还是 UFS2.0/2.1，其最大带宽 1081MB/s，也是秒杀一般的 SSD。

UFS 协议是 JEDEC（www.jedec.org）组织制定的，三星、海力士、东芝等公司力捧。下面是 UFS 协议的发展历程：



我们可以看到，UFS 协议一直在大踏步的朝着更高更快的目标前进。

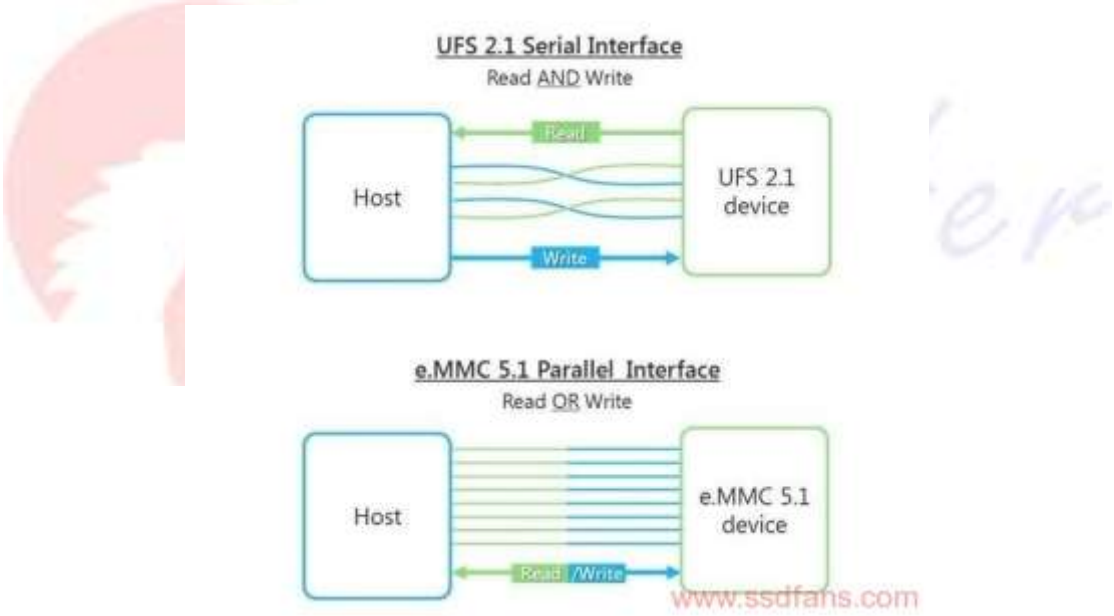
UFS 为什么能那么快？

首先，它在数据信号传输上，使用的是差分串行传输。这是 UFS 快的基础。所有的高速传输总线，如 SATA，PCIe，SAS，都是串行差分信号。串行，可以使用更快的时钟（时钟信息可以嵌在数据流中）；差分信号，即用两根信号线上的电平差表示 0 或者 1。与单端信号传输相比，差分信号抗干扰能力强，能提供更宽的带宽（跑得更快）。打个比方，假设用两个信号线上电平差表示 0 和 1，具体来讲，差值大于 0，表示 1，差值小于 0，表示 0。如果传输过程中存在干扰，两个线上加了近乎同样大小的干扰电平，两者相减，差值几乎不变，你大爷还是你大爷。但对单端信号传输来说，就很容易受干扰，比如 0-1V 表示 0,1-3V 表示 1，一个本来是 0.8V 的电压，加入干扰，变成 1.5V，相当于 0 变成 1，数据就出错了，你大妈已经不是你大妈了。抗干扰能力强，因而可以用更快的速度进行数据传输，从而能提供更宽的带宽了。

UFS 的前辈是 eMMC，使用的是并行数据传输。并行最大的问题是速度上不去，因为一旦时钟上去，干扰就变大，信号完整性无法保证。

其次，UFS 和 PCIe 一样，支持多通道数据传输，目前最多支持两个通道。多通道可以让 UFS 在成本、功耗和性能之间做取舍。

还有，它是全双工工作模式，就是读写可以并行。它的前辈 eMMC 是半双工，读写不能同时进行。



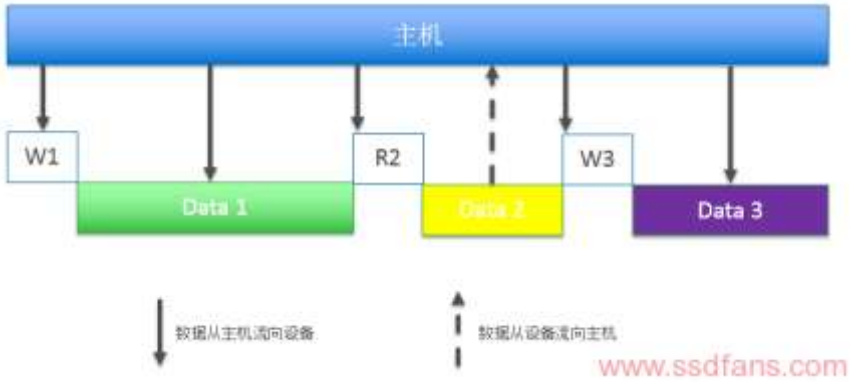
要让 UFS 速度快，这些基础设施是必须的。但要充分利用底层高速数据传输通道，还需要上层数据传输协议配合。就好比我们现在有一条又宽敞又平坦的高速公路，我们需要一辆高速的汽车行驶在上面。你如果让一辆拖拉机在上面跑，高速公路算是白修了。

UFS 协议上层，怎样来充分发挥底层速度快的优势呢？

UFS 支持命令队列，就是主机一下可以发很多个命令下去，然后 UFS 设备支持并行和乱序执行，谁先完成谁先返回状态。这种命令处理方式叫做异步命令处理。而它的前辈 eMMC，是不支持命令队列的，命令一个一个执行，或者一包一包（每个包里面含有若干个命令）执行，前面命令没有执行完成，后面的命令是不能发下去的。这种命令处理方式叫做同步命令处理。

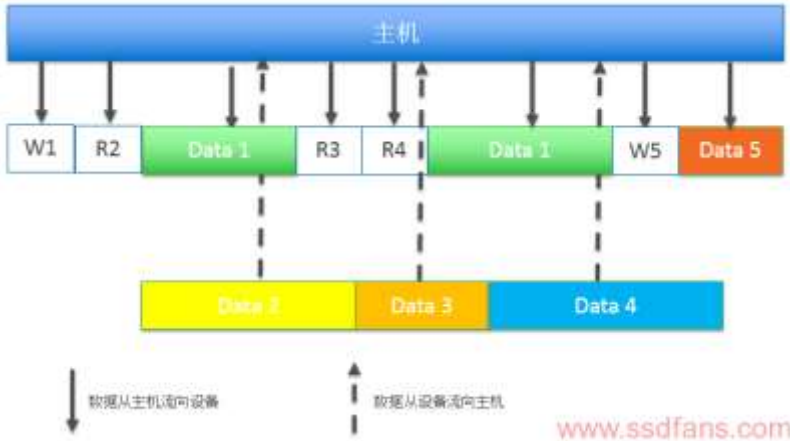
我们来比较一下“全双工+异步命令处理”和“半双工+同步命令处理”两者命令处理方式和命令执行效率。

• 半双工+同步



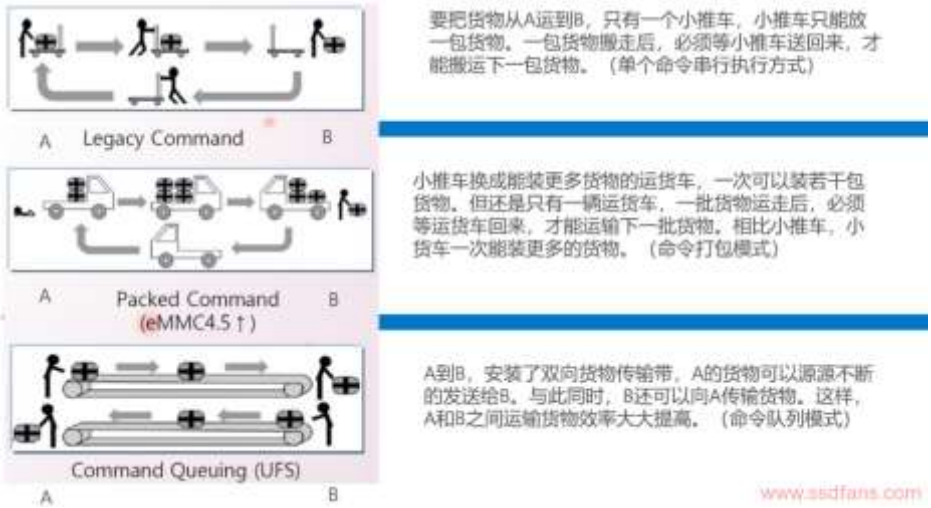
主机发了一个写命令 W1 给设备，然后主机把数据写到设备；由于是同步传输模式，命令处理是一个一个处理的，所以在发读命令 R2 之前，必须等前一个写命令 W1 完成；同样，在发送写命令 W3 之前，必须等 R2 命令完成。

• 全双工+异步



由于支持命令队列，主机一下可以发若干个命令给设备，如上图，主机一下发了一个写命令 W1 和读命令 R2 给设备。设备可以并行处理这两个命令，由于协议支持全双工操作，主机传输写命令 W1 的数据给设备的同时，设备也可以把读命令 R2 的数据返回给主机。后面命令 R3，R4，W5…的处理方式类似。

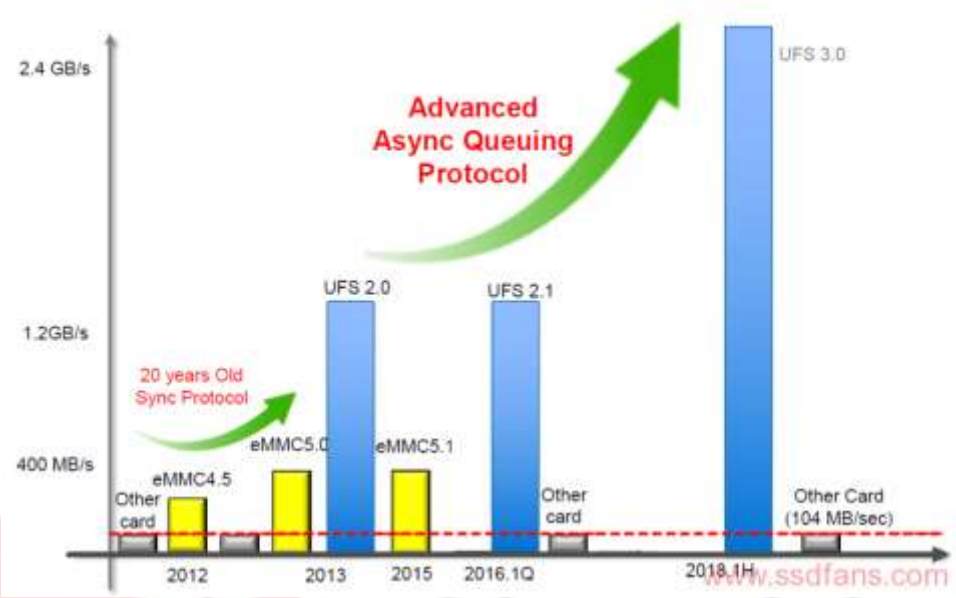
再形象一点，我们以搬运货物的例子来比较一下 eMMC 和 UFS 命令执行方式：



现在的手机，应用非常丰富，你要一边斗地主，一边听歌，还要聊微信，多线程操作。由于全双工和命令队列的存在，UFS 处理命令的效率大大提高，给用户极好的体验。

前面我们拿 UFS 和 eMMC 做了几个对比，但我好像忘了说什么是 eMMC。有人可能懵逼，什么是 eMMC？

eMMC，Embedded Multi Media Card，和 UFS 一样，也是 JEDEC 制定的移动存储协议，它是 UFS 前一代协议标准。



eMMC 最新标准是 2015 年发布的 eMMC5.1，最高速度是 400MB/s。JEDEC 已经有了 UFS，不确定会不会再发布新的 eMMC 标准。毕竟，并行传输的 eMMC 由于受限于物理信号，速度想要有个质的飞跃是不太现实。

行文至此，让我不由的想起去年那事件。

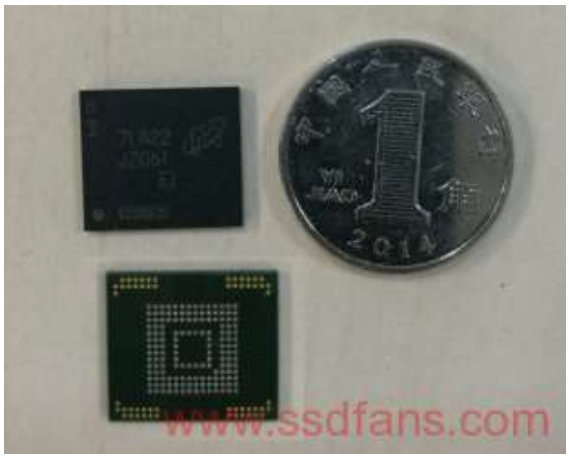
同一款手机，有人 UFS 和 eMMC 混着卖，手机还卖一个价钱，真是无德！UFS 和 eMMC，速度差异那么大（见下图，来自网络），价格能一样吗？你系统再优化，能把 eMMC 顺序读写速度优化到 401MB/s？别扯什么用户体验，用户都被耍猴了，体验还能好？水能载舟，亦能覆舟，不要太得意忘形。

华为P10（1）

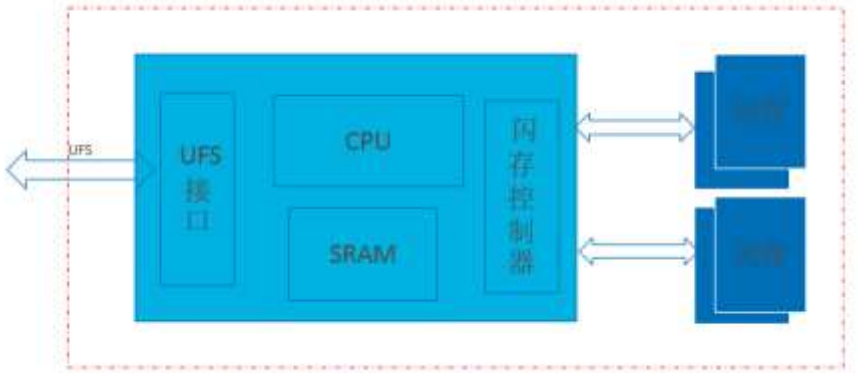
华为P10（2）

如果说 eMMC 是手机中的 HDD，那么 UFS 就是手机中的 SSD。UFS 取代 eMMC 成为主流手机存储协议，这是毫无疑问的。不过，UFS 一统天下的道路上还有一个拦路虎，那就是 NVMe。有人说，NVMe 不是 SSD 的协议标准吗？没错，不过，我要提醒大家的是，苹果现在手机中存储协议是 NVMe 而不是 UFS。在短期，UFS 和 NVMe 会分别在安卓和苹果手机中存在。长期来说，UFS 和 NVMe 是二分天下，还是合二为一，我们只能拭目以待了。

在本章结束前，给大家看看 UFS（BGA 形式）的实物图：



大小如大拇指手指盖大小。麻雀虽小，五脏俱全。UFS 存储芯片内部封装了 UFS 控制器和闪存阵列，和 SSD 结构很相似。不过和 SSD 相比，由于它的容量更小，因此闪存 die 比较少，闪存的通道数也少。另外，出于功耗和成本考虑，UFS 芯片一般是不带 DRAM 的架构。



UFS卡或芯片内部结构 www.ssdfans.com

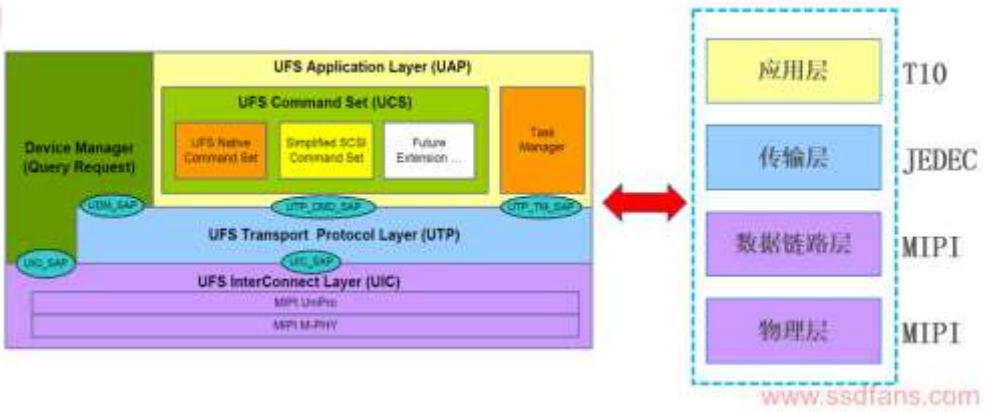
UFS 芯片内部设计与实现不是我们本系列的重点，本系列后续文章将专注于 UFS 协议。

10.5.2 蛋蛋读 UFS 之二：UFS 协议栈

Posted on 2018 年 5 月 30 日 by 蛋蛋

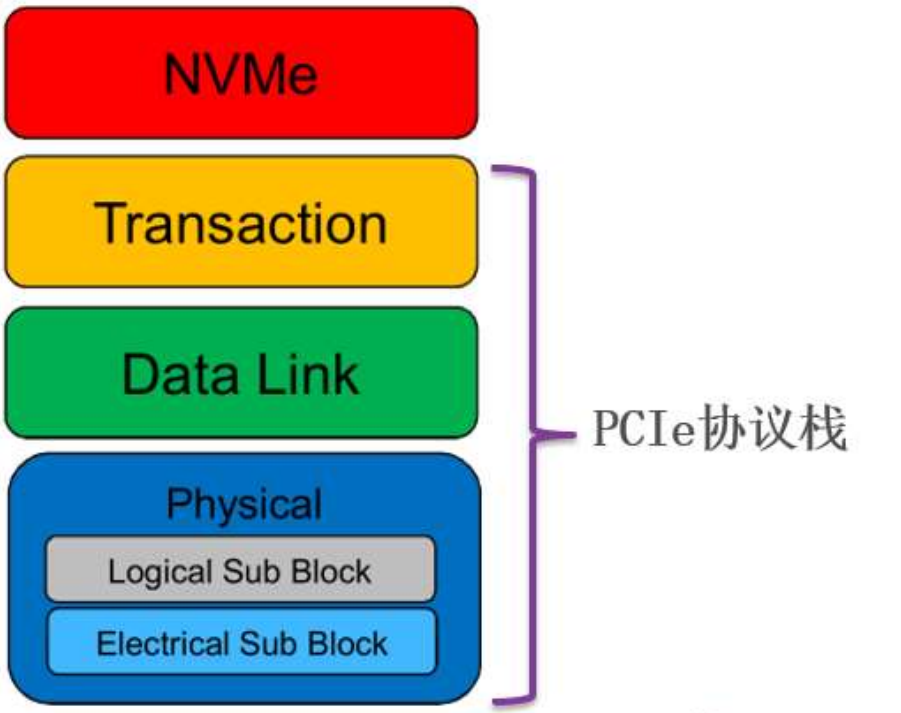
原创内容，转载请注明： [http://www.ssdfans.com] 谢谢！

任何一种接口或者协议，都是由一个完整的协议栈组成的。UFS 也不例外。



www.ssdfans.com

UFS 定义了一个完整的协议栈。从上到下，依次为应用层、传输层、数据链路层和物理层。UFS 使用 MIPI（Mobile Industry Processor Interface ，移动产业处理器接口）联盟的 UniPro 作为数据链路层和 MIPI 的 M-PHY 作为物理层，两者合起来称之为互连层（UFS InterConnect Layer）。与之相比，PCIe 接口只定义了下三层（如下图），没有应用层。只有加上上层 NVMe，才构成一个完整的 SSD 通讯协议。



目前 UFS 没有定义自己的命令（没有 UFS Native Command Set），使用的命令是简化的 SCSI 命令（基于 SBC 和 SPC），由 INCITS T10 组织定义的。关于 SCSI 相关协议，大家可以参看相应的 spec。

四层中，只有传输层是 JEDEC 自己定义的。所以，UFS 四层中有三层是别人的，命令层是 T10 的，数据链路层和物理层是 MIPI 的，传输层是 JEDEC 自己的。JEDEC 移花接木的水平真是高。不由的想到一个广告：“我们不生产水，我们只是大自然的搬运工！”

UFS 至今已经有五个版本，每层的版本也不尽相同。

协议层		规范	组织	UFS 1.1	UFS 2.0	UFS 2.1	UFS 3.0
应用层	SCSI	SPC-4 SBC-3 SAM-5	T10	Rev. 27 Rev. 24 Rev. 05	Rev. 27 Rev. 24 Rev. 05	Rev. 27 Rev. 24 Rev. 05	Rev. 27 Rev. 24 Rev. 05
传输层	UTP	UFS	JEDEC	UFS 1.1	UFS 2.0	UFS 2.1	UFS 3.0
互联层（数据链路层+物理层）	UIC	UniPro	MIPI	1.41	1.6	1.6	1.8
		M-PHY	MIPI	2.00	3.0	3.0	4.1

我们依次来看看这几层。

λUFS 应用层

应用层包括 UFS 命令集、设备管理器（Device Manager）和任务管理器（Task Manager）。应用层处于整个协议栈的最高层，所有的命令或者请求都来源于该层。它是最高统帅，所有的战术和策略都是它制定的，然后真正去冲锋陷阵的是将军和士兵（应用层下面的传输层和内联层）。

命令集

如前所述，目前 UFS 没有定义自己的命令，使用简化的 SCSI 命令。

其中包括一些 SPC（SCSI Primary Commands）命令：

命令名字
INQUIRY
MODE SELECT (10)
MODE SENSE (10)
REPORT LUNS
READ BUFFER
TEST UNIT READY
WRITE BUFFER
SECURITY PROTOCOL IN
SECURITY PROTOCOL OUT

和一些 SBC（SCSI Block Commands）命令：

命令名字
FORMAT UNIT
PRE-FETCH (10) / PRE-FETCH (16) *
READ (6) / READ (10) / READ (16) *
READ CAPACITY (10) / READ CAPACITY (16)
REQUEST SENSE
SEND DIAGNOSTIC
UNMAP
WRITE (6) / WRITE (10) / WRITE (16) *
START STOP UNIT
SYNCHRONIZE CACHE (10) / CACHE (16) *
VERIFY (10)

注：* 表示可选命令

UFS除了定义基本的读写命令，也有trim命令（UNMAP），还有其它一些命令。我们不打算深入其中。

设备管理器

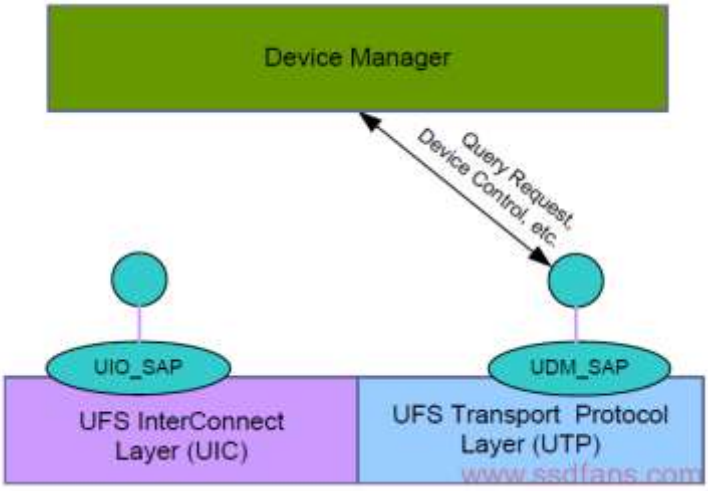
顾名思义，设备管理器用以管理UFS设备。

设备管理器有两个功能：一是处理设备级操作，二是管理设备级配置。

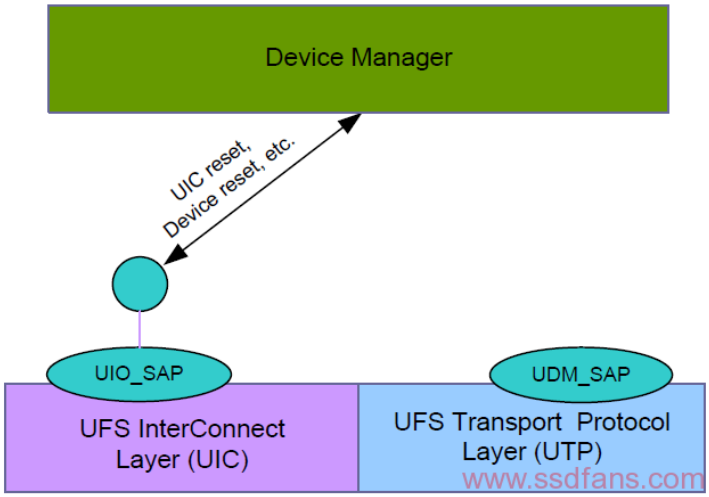
前者包括管理设备功耗、设置数据传输相关参数、使能/禁止设备后台操作（Background Operation）以及其它设备相关操作。

后者通过维护和存储一系列的描述符（Descriptor，后面有章节介绍），通过诸如Query请求修改或获取设备的配置信息。

从UFS层次架构图来看，设备管理器既可以通过下层的传输层为其服务（通过UDM_SAP）：



设备管理器也可以绕过传输层（通过 UIO_SAP），直接管理与控制互联层：



设备管理器可以通过互联层提供的接口（UIO_SAP），使用一系列的原语（Primitive）直接控制操作互联层（UIC）。这些原语包括重启设备、重启互联层、让物理层进入和退出休眠模式（Hibernate）等原语。

总之，设备管理器既可以走常规渠道（通过传输层，以数据包 UPIU 的形式），也可以走快速通道（发送 UIC 能理解的命令，原语的形式）管理和操作设备。

任务管理器

任务管理器用以管理命令队列中的命令。比如任务管理器可以发 Abort 命令，终止之前发下去的命令。它也可以清空命令队列中的所有命令。具体如下：

Function	Value	Description
Abort Task	01h	Abort specific task in queue in a specific LU. Identify by LUN and Task Tag
Abort Task Set	02h	Abort the task queue list in a specific LU. Identify by LUN.
Clear Task Set	04h	Clear the task queue list in specific LU. Identify by LUN. Equivalent to Abort Task Set.
Logical Unit Reset	08h	Reset the designated LU. Identify by LUN
Query Task	80h	Query a specific task in a queue list in a specific LU. Identify by LUN and Task Tag. If the specific task is present in the queue, Function Succeeded is returned in the response. If the specific task is not present in the queue, Function Complete is returned in the response.
Query Task Set	81h	Query a specific LU to see if there is any Task in queue. Identify by LUN. If there is one of more tasks present in the queue, Function Succeeded is returned in the response. If no task is present in the queue, Function Complete is returned in the response.

当某个命令超时时，系统可能发 Abort 命令把这个命令终止掉。

UFS 传输层

传输层为它上面的应用层服务。当传输层收到应用层命令或者请求后，它会产生 UPIU(UFS Protocol Information Unit)，把命令块或者请求封装成固定格式的数据结构，然后交由下层传到接收端的传输层。和命令相关的数据、状态，也有相应的 UPIU 数据包。UPIU 是主机和设备进行信息交换的基本数据单元。

UPIU，和 SATA 中的 FIS，PCIe 中的 TLP，是同一层次的东西，上层命令或者数据都是通过此类数据包封装起来，然后传输到接收端。

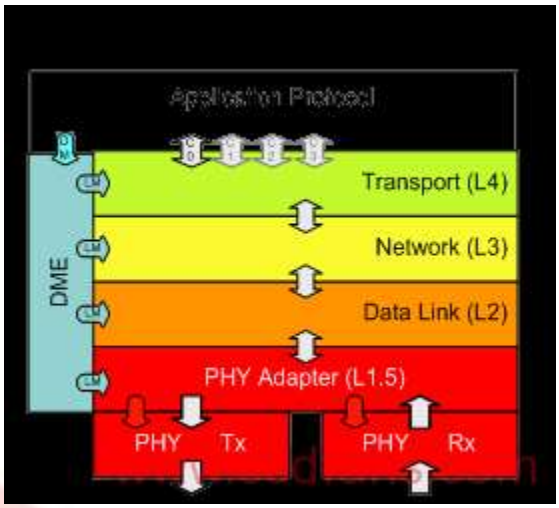
如果说应用层是统帅的话，传输层可以认为是将军了。

下一章节为专门介绍 UPIU，这里就不细讲。

UFS 互联层

UFS 互联层包括 MIPI UniPro 和 M-PHY，分别充当 UFS 数据链路层和物理层的角色。数据链路层负责主机和设备的链接，物理层传输实实在在的物理信号。

UniPro 其实不仅仅只定义了数据链路层，它本是也是一个比较完整的协议栈，如下图所示：



传输层（L4）支持多设备之间的双向连接，但 UFS 只支持 CPort0；

网络层（L3）支持通过设备 ID 寻址多达 128 个设备，但由于 UFS 是点到点传输，所以无需网络层；

数据链路层（L2）支持流控、CRC 生成和校验、重传机制等，UFS 利用了 UniPro 的数据链路层为主机和设备之间通讯提供可靠的连接。

物理层（M-PHY）使用 8/10 编码、差分信号串行数据传输。数据传输分高低速模式，每种模式下又有几种不同的速度档。

关于 MIPI UniPro 和 M-PHY，读者可以看相关的 spec，这里不细讲。

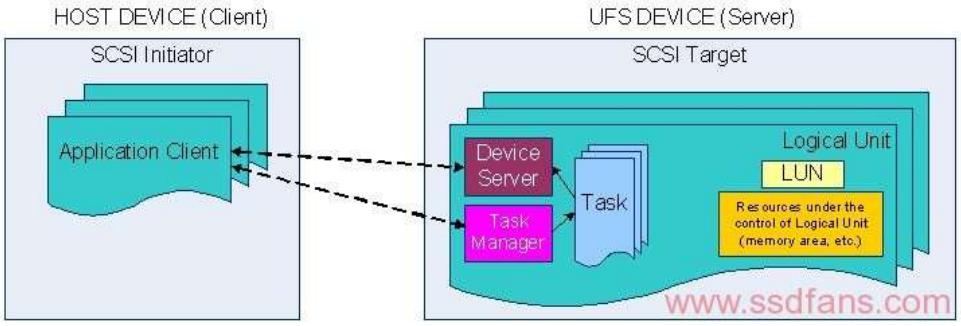
本章对 UFS 协议栈做了简单介绍，下一章将会对传输层发起的 UPIU 进行详细的介绍。

10.5.3 蛋蛋读 UFS 之三：UFS 数据包 UPIU

Posted on 2018 年 5 月 30 日 by 蛋蛋

UFS 中流淌的数据包叫做 UPIU（UFS Protocol Information Unit，UFS 协议信息单元），它是固定格式的数据结构，用以传输应用层发来的命令或者请求，以及跟它们相关的数据或者状态信息。它就是 SATA 中的 FIS，PCIe 中的 TLP。

我们看看 UFS 中命令或请求是怎么执行的。



UFS 采用“客户-服务器”或者说主从的命令架构，UFS 主机（Client，命令发起者，Initiator，他们都是一个意思）发送命令或者请求（Request）给 UFS 设备（服务器，Target），然后 UFS 设备执行命令并返回命令状态（Response）。

一个命令或者请求的执行包含下面几个阶段：

命令阶段：主机发起命令或请求给设备，这是“因”；

数据阶段：传输跟命令相关的数据，比如读写命令，都涉及到数据的传输；有些命令不涉及数据的传输，所以这个阶段并不是总是存在的，跟具体命令和请求相关。

状态阶段：设备执行完命令，必须给主机返回命令执行状态信息。这个是“果”，必不可少的。在 PCIe 中，有 Posted 和 Non-posted 的 TLP。对前者，命令执行者无需返回命令执行状态给命令发起者，对后者，命令执行者必须返回状态给命令发起者。对 UFS 来说，它的命令总是 non-posted，即设备必须返回命令状态给主机。

在命令执行过程中，无论是处在哪个阶段，UFS 主机和设备间都是通过 UPIU 进行信息的交互。

1. UFS 主机通过命令或者请求 UPIU 发命令请求给设备；

2. UFS 主机或者设备通过 UPIU 传输数据；

3. UFS 设备通过 UPIU 返回命令状态信息给主机。

下面我们看看 UFS 当中都有哪些 UPIU。

命令或者请求 UPIU

前一章看到，应用层包括 UFS 命令、设备管理器和任务管理器三个模块，传输层根据不同模块发来的命令或者请求，分别产生不同类型的 UPIU。

UFS 命令模块发送简化版本的 SCSI 命令，当传输层收到命令请求后，它会生成：**COMMAND UPIU**，把命令封装起来。

应用层通过任务管理器来管理任务队列，比如终止（Abort）和查询命令队列中的命令。当传输层收到来自任务管理器中的请求后，它会生成：**TASK MANAGEMENT REQUEST UPIU**，把请求封装起来。

UFS 通过设备管理器来管理 UFS 设备，比如设置和查询 UFS 设备的配置（Configuration）。当传输层收到来自设备管理器发来的请求后，它会生成：**QUERY REQUEST UPIU**，把请求封装起来。

应用层模块	对应的 UPIU	传输方向	作用
SCSI 命令	COMMAND UPIU	主机到设备	主机发送 SCSI 命令给设备
任务管理器	TASK MANAGEMENT REQUEST UPIU	主机到设备	主机管理命令队列中的命令，比如终止或者查询设备命令队列中的命令
设备管理器	QUERY REQUEST UPIU	主机到设备	主机通过设备管理器查询、配置、管理设备。

数据传输相关 UPIU

当主机发送了类似读命令给设备之后，设备需要返回数据给主机，设备通过 **DATA IN UPIU** 向主机传输数据。

当主机发送了类似写命令给设备之后，主机需要往设备写数据，主机通过 **DATA OUT UPIU** 向设备传输数据。

UFS 的主机是个暖男，它在向设备写数据的时候，会考虑到设备这个时候能不能接收数据（因为设备可能这个时候没有足够的空间接收主机数据），它在向设备发了写命令之后，不会立刻把数据传输给设备，而是在那里等设备的通知。当设备准备好接收数据，以及接收多少数据，设备通过 **READY TO TRANSFER UPIU （RTT）**告知主机。当主机接收到该 RTT 后，才开始按照 RTT 的信息传输数据。至于每次传输数据的多少，RTT 中包含这信息，主机根据 RTT 进行传输。

所以，主机只有在收到设备的 RTT，才能发 DATA OUT UPIU！

注意，读命令无需这种机制。因为设备从闪存中获得数据后，是设备控制数据的传输。对主机来说，它在发读命令之前，已经准备好足够的空间用以接收数据，所以不存在主机没有空间接收数据的情况。

数据传输相关的 UPIU	传输方向	作用
DATA IN UPIU	设备到主机	设备传输数据给主机
DATA OUT UPIU	主机到设备	主机写数据到设备。主机只有收到 RTT 后，才能往设备写数据。
READY TO TRANSFER UPIU	设备到主机	同步。处理写命令时，设备告诉主机可以传数据以及传多少数据。

状态 UPIU

前面看到，主机有三种请求：SCSI 命令，任务管理器发出的 Task Management Request，以及设备管理器发出的 Query request。针对不同的命令或者请求，设备在执行完相应的任务后，分别返回对应的状态 UPIU 给主机。

设备响应 UPIU	传输方向	作用
RESPONSE UPIU	设备到主机	设备返回命令执行状态
TASK MANAGEMENT RESPONSE UPIU	设备到主机	设备返回任务管理请求执行状态
QUERY RESPONSE UPIU	设备到主机	设备返回设备管理器的 Query 请求执行状态

其它 UPIU

除了以上常规的 UPIU，还有其它一些 UPIU 作为他用。

设备上电后，主机检测是否与之连接，会发 **NOP OUT UPIU** 给设备。我们平时想看看跟某个电脑或者网站能否连接上，会发一个 ping 命令。NOP OUT UPIU 跟 ping 命令作用类似。

当设备收到 NOP OUT UPIU 后，会返回 **NOP IN UPIU**。主机收到该 UPIU 后，确认与设备连接，然后可以进行后续操作。

最后一个 UPIU 就是 **REJECT UPIU**。当设备收到一个无效的 UPIU 时，它会发 REJECT UPIU 拒绝无效的 UPIU。

辅助 UPIU	传输方向	作用
NOP OUT UPIU	主机到设备	主机 ping 设备，查询主机是否与设备相连
NOP IN UPIU	设备到主机	设备响应 NOP OUT UPIU。主机收到该 UPIU 后，确认主机与设备相连。
REJECT UPIU	设备到主机	设备收到无效的 UPIU，发该 UPIU 给主机

UPIU 汇总

Table 10-1 — UPIU Transaction Codes

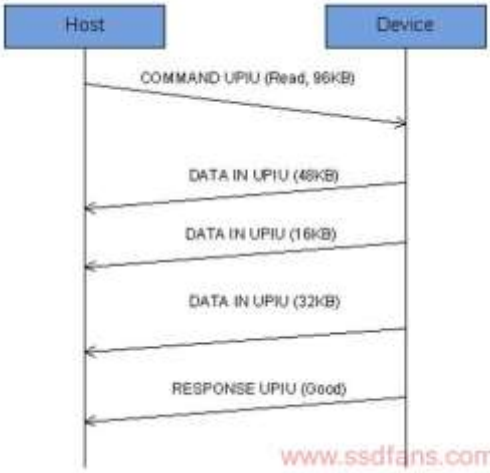
Initiator To Target	Transaction Code	Target to Initiator	Transaction Code
NOP OUT	00 0000b	NOP IN	10 0000b
COMMAND	00 0001b	RESPONSE	10 0001b
DATA OUT	00 0010b	DATA IN	10 0010b
TASK MANAGEMENT REQUEST	00 0100b	TASK MANAGEMENT RESPONSE	10 0100b
Reserved	01 0001b	READY TO TRANSFER	11 0001b
QUERY REQUEST	01 0110b	QUERY RESPONSE	11 0110b
Reserved	01 1111b	REJECT UPIU	11 1111b
Reserved	Others	Reserved	Others
NOTE 1 Bit 5 of the Transaction Code indicates the direction of flow and the originator of the UPIU: when equal '0' the originator is the Initiator device, when equal '1' the originator is the Target device.			

偷个懒，我就直接把 UFS spec 这张表贴这里。数了数，一共 12 个 UPIU。经过我之前的解释，读者现在应该清楚每个 UPIU 的作用了。

读写命令中 UPIU 交互例子

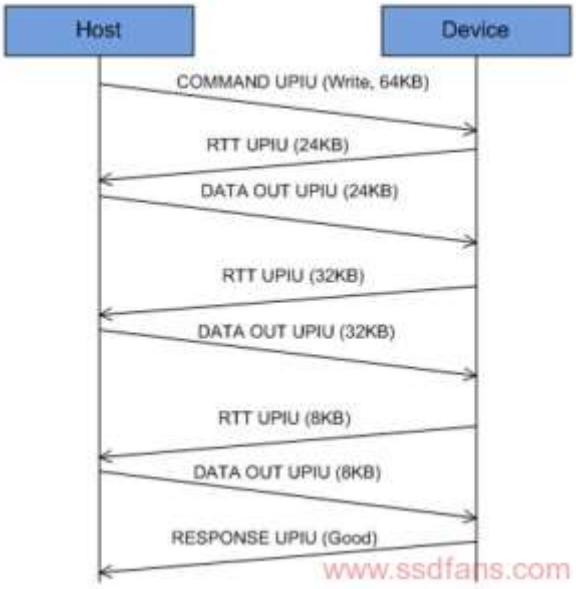
前面我们都是单个来看 UPIU，现在我们以读写命令为例，看看他们是如何组合完成命令处理的。

首先是一个**“主机往设备读取 96KB 数据”**的例子。



首先，主机发送读 96KB 数据的命令给设备，然后设备执行命令，分了三批把数据返回给主机，最后返回命令执行状态给主机。

然后是一个**“主机往设备写 64KB 数据”**的例子。



主机发送写 64KB 数据的命令给设备，然后在那里等设备响应。很快，设备说，你可以传 24KB 数据下来了，于是主机写 24KB 数据给设备；接着，设备又来通知说可以继续传 32KB 数据，主机照做。最后，设备通知说可以把最后 8KB 数据也传过来，主机于是写最后 8KB 数据。最后，主机收到设备命令执行完成的响应。

我们看到，主机必须等收到 RTT 后才能启动数据传输！

10.5.4 蛋蛋读 UFS 之四：UPIU 数据包格式

Posted on 2018 年 5 月 31 日 by 蛋蛋

UPIU 是命令、数据和状态信息传输的载体，是 UFS 协议栈的灵魂。UPIU 是有固定格式的数据包，我们分析数据包格式，有助于我们更深的理解 UPIU 以及整个 UFS 协议。这一章我们看看 UPIU 数据包的格式。

每个 UPIU 都有一个 12 字节的 Header，再加上跟每个 UPIU 相关的域。一个 UPIU（包括 Header）最小为 32 字节，最大为 65600 字节。

我们看通用的 Header，具体如下：

Table 10-4 — Basic Header Format

Basic UPIU Header Format				
Transaction Type		Flags	LUN	Task Tag
Initiator ID	Command Set Type	Query Function, Task Manag. Function	Response	Status
Total EHS Length		Device Information	Data Segment Length	

我们看看其中的一些域。

1. **Transaction Type**：就是指定该 UPIU 是前面 12 个 UPIU 中的哪一个,具体如下：

Table 10-1 — UPIU Transaction Codes

Initiator To Target	Transaction Code	Target to Initiator	Transaction Code
NOP OUT	00 0000b	NOP IN	10 0000b
COMMAND	00 0001b	RESPONSE	10 0001b
DATA OUT	00 0010b	DATA IN	10 0010b
TASK MANAGEMENT REQUEST	00 0100b	TASK MANAGEMENT RESPONSE	10 0100b
Reserved	01 0001b	READY TO TRANSFER	11 0001b
QUERY REQUEST	01 0110b	QUERY RESPONSE	11 0110b
Reserved	01 1111b	REJECT UPIU	11 1111b
Reserved	Others	Reserved	Others
NOTE 1 Bit 5 of the Transaction Code indicates the direction of flow and the originator of the UPIU: when equal '0' the originator is the Initiator device, when equal '1' the originator is the Target device.			

2. **Flags**：只对命令和其响应的 UPIU 有用，指定命令的属性。

Table 10-6 — UPIU Flags

UPIU Type	Operational Flags				Rsvd	CP[2]	Task Attribute	
	Bit 7	Bit 6	Bit 5	Bit 4	Bit 3	Bit 2	Bit 1	Bit 0
NOP Out	-	-	-	-	-	-	-	-
NOP In	-	-	-	-	-	-	-	-
Command	-	R	W	-	-	CP[2]	ATTR	
Response	-	O	U	D	-	-	-	-
Data Out	-	-	-	-	-	-	-	-
Data In	-	-	-	-	-	-	-	-
Ready to Transfer	-	-	-	-	-	-	-	-
Reject	-	-	-	-	-	-	-	-
Query Request	-	-	-	-	-	-	-	-
Query Response	-	-	-	-	-	-	-	-
Task Management Request	-	-	-	-	-	-	-	-
Task Management Response	-	-	-	-	-	-	-	-
NOTE 1 "-" denotes reserved values. NOTE 2 CP = Command Priority								

R: 如果该比特置起来，说明该命令是读命令；

W: 如果该比特置起来，说明该命令是写命令；

ATTR: 命令属性域。UFS 命令有 simple ， ordered 和 Head of Queue 命令。

ATTR Definition

Task Attribute	Bit 1	Bit 0
Simple	0	0
Ordered	0	1
Head of Queue	1	0
ACA (Not Used)	1	1

那么，这些命令有什么不一样呢。

Simple command：就是一般的命令，设备收到这样的命令无需特别处理，一般谁先到谁先执行。

Ordered command：设备收到这样的命令，应该把该命令之前的命令都处理完，才能处理该命令。（明星出场，先清个场。）

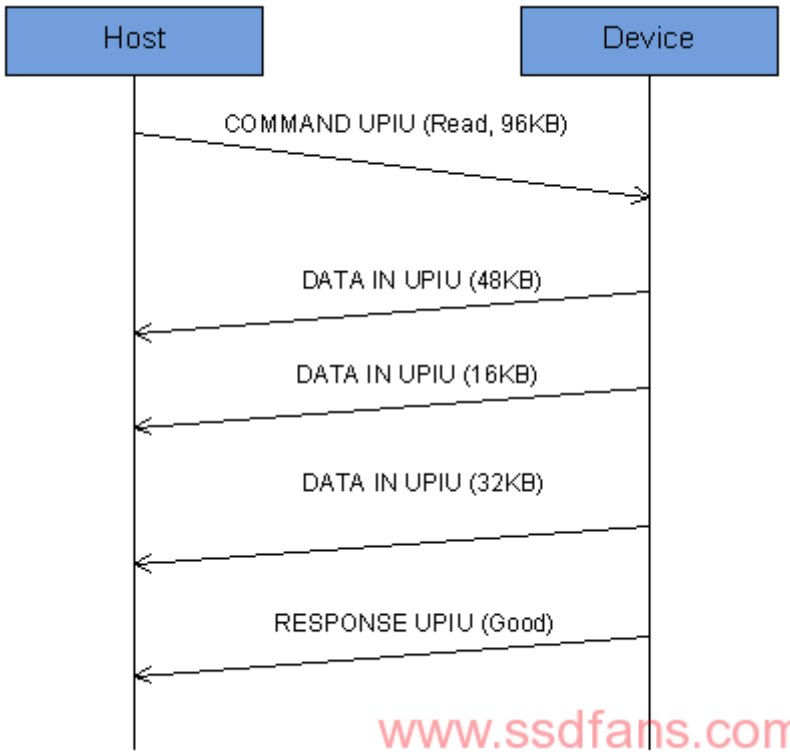
Head of Queue command：设备收到该命令后，放到命令队列的头部，立刻执行。（又见插队，这个没有上过幼儿园吧，连基本的排队意识都没有。）

CP: 表示命令的优先级。1 为高优先级，0 为低优先级。注意，该比特只适合简单命令（simple command）。

3. LUN: Logical Unit Number。UFS 上层协议来自 SCSI，它继承了 LU 的概念，即把存储物理空间划分成若干个逻辑空间，每个逻辑空间都是从 LBA 0 开始，用 LUN 标识。主机在发命令或者请求时，应该在命令中指定该命令是发给哪个 LU。LUN 用以寻址。UFS 的 LU 和 NVMe 中的 Namespace 一个概念。

4. Task Tag：UFS 支持命令队列，主机可以同时发送很多个命令给设备。为区分这些命令或者请求，主机需要为每个命令贴上标签 Tag。然后跟这个命令或者请求相关的数据 UPIU 和状态 UPIU，都具有跟这个命令 UPIU 一样的 Tag。

举例：



对这个读命令来说，COMMAND UPIU、所有的 DATA IN UPIU 和 RESPONSE UPIU 都具有同一个 task tag。

5. Command Type：命令类型。UFS 预期有三类命令：一是简化的 SCSI 命令，二是 UFS 自己原生的命令，三就是用户自定义命令。目前 UFS 的命令都是从别人家（SCSI）借来的，自己一个命令也没有制定。如用户无自定义命令，该域就是 0（SCSI 命令）。

6. Initiator ID： 主机的 ID，手机系统中一般一个主机连接一个 UFS 设备，所以主机 ID 一般为 0。

7. Response：设备告知主机命令或请求执行是否成功。

8. Status：设备返回命令执行状态。对 SCSI 命令的状态信息，UFS 有如下状态：

Table 10-15 — SCSI Status Values

Opcode	Response Description	Use
00h	GOOD	M
02h	CHECK CONDITION	M
04h	CONDITION MET	n/a
08h	BUSY	M
18h	RESERVATION CONFLICT	O
28h	TASK SET FULL	M
30h	ACA ACTIVE	n/a
40h	TASK ABORTED	n/a

GOOD - This status indicates that the device has completed the command without error.

CHECK CONDITION - This status indicates that the device has completed the command with error or other actions are required to process the result. Valid Sense Data for the last command processed will be returned within the response UPIU when this status occurs.

CONDITION MET - Not used for UFS.

BUSY - This status indicates that the logical unit is busy. When the logical unit is unable to accept a command this status will be returned. Issuing the command at a later time is the standard recovery action.

RESERVATION CONFLICT - This status is returned when execution of the command will result in a conflict of an existing reservation. UFS may support reserving areas of the device depending upon the device type and capabilities.

TASK SET FULL - This status is returned when the logical unit cannot process the command due to a lack of resources such as task queue being full or memory needed for command execution is temporarily unavailable.

ACA ACTIVE - This status is returned when an ACA condition exists. See [SAM] for further definition.

TASK ABORTED - This status shall be returned when a command is aborted by a command or task management function on another I_T nexus and the Control mode page TAS bit is set to one. Since in UFS TAS bit is zero TASK ABORTED status codes will never occur.

9. Query Function, Task Manag. Function：指定具体 Query 和 Task Management 功能。

任务管理器有如下功能（Function）：

Table 10-23 — Task Management Function values

Function	Value	Description
Abort Task	01h	Abort specific task in queue in a specific LU. Identify by LUN and Task Tag
Abort Task Set	02h	Abort the task queue list in a specific LU. Identify by LUN.
Clear Task Set	04h	Clear the task queue list in specific LU. Identify by LUN. Equivalent to Abort Task Set.
Logical Unit Reset	08h	Reset the designated LU. Identify by LUN
Query Task	80h	Query a specific task in a queue list in a specific LU. Identify by LUN and Task Tag. If the specific task is present in the queue, Function Succeeded is returned in the response. If the specific task is not present in the queue, Function Complete is returned in the response.
Query Task Set	81h	Query a specific LU to see if there is any Task in queue. Identify by LUN. If there is one of more tasks present in the queue, Function Succeeded is returned in the response. If no task is present in the queue, Function Complete is returned in the response.

设备管理器有如下功能：

Table 10-29 — Query Function field values

QUERY FUNCTION	
00h	Reserved
01h	STANDARD READ REQUEST
02h-3Fh	Reserved
40-7Fh	Vendor Specific Read Functions
80h	Reserved
81h	STANDARD WRITE REQUEST
82h-BFh	Reserved
C0h-FFh	Vendor Specific Write Functions

总的来说，就是读写设备属性（Attributes）、标识（flags）和描述符（descriptors）。

Table 10-31 — Query Function opcode values

OPCODE	Operation	QUERY FUNCTION
00h	NOP	Any value
01h	READ DESCRIPTOR	STANDARD READ REQUEST
02h	WRITE DESCRIPTOR	STANDARD WRITE REQUEST
03h	READ ATTRIBUTE	STANDARD READ REQUEST
04h	WRITE ATTRIBUTE	STANDARD WRITE REQUEST
05h	READ FLAG	STANDARD READ REQUEST
06h	SET FLAG	STANDARD WRITE REQUEST
07h	CLEAR FLAG	STANDARD WRITE REQUEST
08h	TOGGLE FLAG	STANDARD WRITE REQUEST
09h-EFh	Reserved	Reserved
F0h-FFh	Vendor Specific	Vendor Specific

关于设备属性、标识和描述符，后面有专门章节讲述。

10. Device Information：设备信息。该域往往跟该命令或者请求无关，属于设备夹带私货。因为 UFS 主机和设备是主从关系，如果 UFS 主机没有向设备发命令或者请求，UFS 设备是不能主动向主机报告设备状况的。如果 UFS 设备有特殊事件发生，它可以趁返回 RESPONSE UPIU 的时候把事件顺带告诉主机。所以该域只对 RESPONSE UPIU 有效。

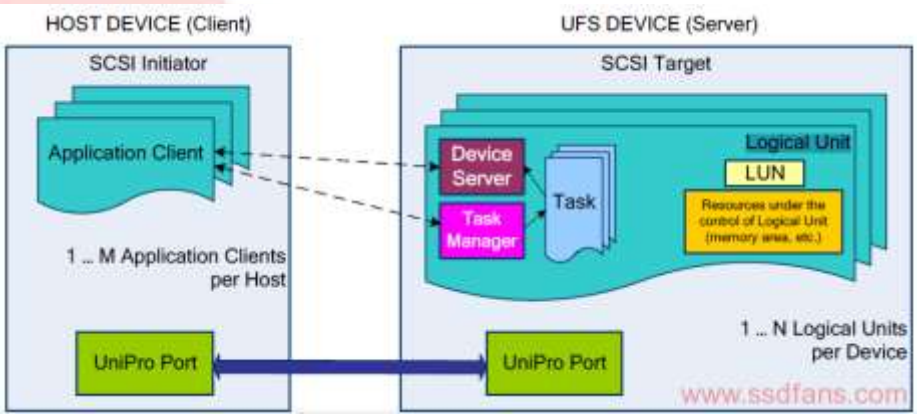
以上是 UPIU 头的基本信息，这个是所有 UPIU 都具有的。除此之外，每个 UPIU 有它独有的其它信息，UFS spec 上都有介绍，读者可以自行阅读。

10.5.5 蛋蛋读 UFS 之五：逻辑单元（LU）

Posted on 2018 年 5 月 31 日 by 蛋蛋

原创内容，转载请注明： [http://www.ssdfans.com] 谢谢！

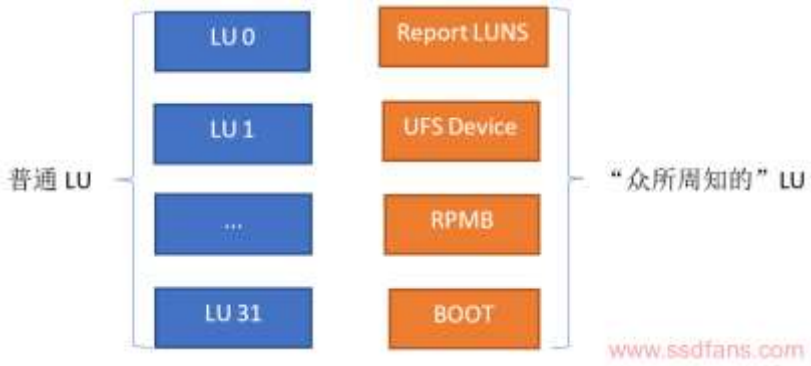
熟悉 NVMe 的朋友知道，NVMe 里面有 namespace 的概念，就是把 SSD 物理空间划分成若干个逻辑地址空间。在 UFS 的世界里，它也有这个特性。UFS 设备的物理存储空间可以有若干个独立的逻辑地址空间，我们把逻辑地址空间叫做 LU，即 Logical Unit，俗称“撸”。前面看到，在每个 UPIU 的 Header 中，有个 LUN（Logical Unit Number）的域，就是标识该 UPIU 关联的命令或者请求的目标逻辑单元。每个 LU 的地址空间是独立的，主机在发命令或者请求给设备的时候，须通过 LUN 指定目标逻辑单元。



如上图所示，UFS 设备有若干个 LU，每个 LU 接收主机发过来的命令或者请求，这些命令或者请求可来自应用层的 SCSI 模块、设备管理器或者任务管理器。每个 LU 都是独立的，“独立”表现在下面几个方面：

- 逻辑地址空间是独立的，都是从 LBA 0 开始；
- 逻辑块大小可以不同，可以为 4KB，..；
- 可以有不同的安全属性，比如可以设置不同的写保护属性；
- 每个 LU 可以有自己的命令队列；
- 不同的 LU 可以存储不同的数据，比如有的 LU 存储系统启动代码，有的 LU 存储普通的应用数据，有的 LU 存储用户特殊数据…
- 。。。

UFS2.1 中可以有最多 32 个普通 LU 和“四大名撸”（四个 Well known LU，众所周知的 LU）。



普通 LU 的逻辑块大小至少是 4KB，但 RPMB LU 逻辑块大小为 256B。至于什么是 RPMB LU，后面再讲。

普通 LU 我觉得没有什么好讲的，就是分别用来存储用户数据的。我们主要来讲讲“四大名撸”。

- **Report LUNS LU**

Report LUNS 主要用来代表设备向主机汇报设备 LU 清单。主机想知道设备 LU 的支持情况，就需要发命令或者请求给该 LU。UFS 其中有个命令“Report LUNS”（和该 LU 名字一样）用来访问 Report LUNS。

Command name	Opcode	Command Support
FORMAT UNIT	04h	M
INQUIRY	12h	M
MODE SELECT (10)	55h	M
MODE SENSE (10)	5Ah	M
PRE-FETCH (10)	34h	M
PRE-FETCH (16)	90h	O
READ (6)	08h	M
READ (10)	28h	M
READ (16)	88h	O
READ BUFFER	3Ch	O
READ CAPACITY (10)	25h	M
READ CAPACITY (16)	9Eh	M
REPORT LUNS	A0h	M
REQUEST SENSE	03h	M
SECURITY PROTOCOL IN	A2h	M
SECURITY PROTOCOL OUT	B5h	M
SEND DIAGNOSTIC	1Dh	M
START STOP UNIT	1Bh	M
SYNCHRONIZE CACHE (10)	35h	M
SYNCHRONIZE CACHE (16)	91h	O
TEST UNIT READY	00h	M
UNMAP	42H	M
VERIFY (10)	2Fh	M
WRITE (6)	0Ah	M
WRITE (10)	2Ah	M
WRITE (16)	8Ah	O
WRITE BUFFER	3Bh	M
M: mandatory, O: optional, R: RPMB		
NOTE 1 SECURITY PROTOCOL IN command and SECURITY PROTOCOL OUT command are supported by the RPMB well known logical unit.		

- **UFS Device LU**

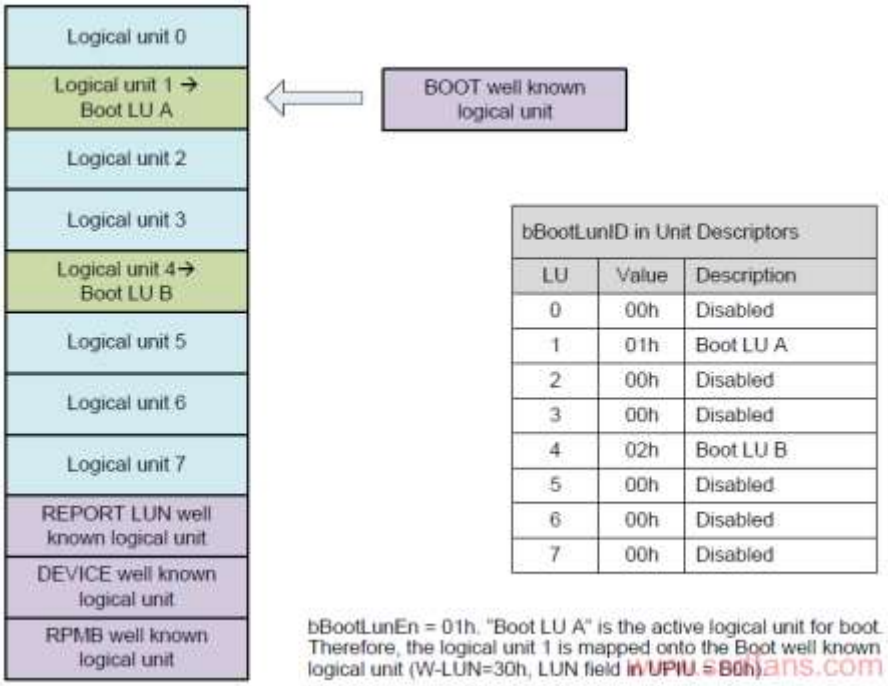
UFS 设备的法人。当 UFS 主机不针对某个具体 LU，而是对整个 UFS 设备发命令的时候，UFS Device LU 就成为该命令接收的对象，比如格式化 UFS 设备（FORMAT UNIT 命令）、切换 UFS 设备的功耗模式（START STOP UNIT 命令）等等。

- **BOOT LU**

顾名思义，就是用来存储启动代码的 LU。不过，BOOT LU 本身是不存储启动代码的，它只是个虚拟的 LU，启动代码物理上是存储在普通 LU 上的。

有两个 Boot LU，LU A 和 LU B，可以用来存储不同启动代码（比如一个新，一个旧），但在启动过程中，只有一个是活跃的（Active）的。32 个普通 LU 中的任意一个可以配成 Boot LU A 或者 Boot LU B。

举例说明：



bBootLunID in Unit Descriptors		
LU	Value	Description
0	00h	Disabled
1	01h	Boot LU A
2	00h	Disabled
3	00h	Disabled
4	02h	Boot LU B
5	00h	Disabled
6	00h	Disabled
7	00h	Disabled

在上例中，LU 1 充当 Boot LU A，LU 4 充当 Boot LU B。由于有两份启动代码，分别保存在 LU 1 和 LU 4，那启动的时候读取哪一份呢？

主机启动时，首先应该通过设备管理器，发送 Query 请求给设备，获取一个叫做“bBootLunEn ”的属性，该属性标识当前活跃（Active）的 Boot LU。

Table 13-1 — bBootLunEn Attribute

bBootLunEn	Description
00h	Boot LU A = disabled Boot LU B = disabled
01h	Boot LU A = enabled Boot LU B = disable
02h	Boot LU A = disable Boot LU B = enabled
Others	Reserved

在上例中，bBootLunEn = 01，说明 Boot LU A 是当前活跃的 Boot LU，因此主机会从 LU 1 上读取启动代码完成系统的启动。

值得一提的是，Boot LU 不是必须的。如果系统的启动代码不是存储在 UFS 设备上，那么 Boot LU 就不需要，因此 bBootLunEn = 0。

- **RPMB LU**

在 UFS 里，有这么一个 LU，主机往该 LU 写数据时，UFS 设备会校验数据的合法性，只有特定的主机才能写入；同时，主机在读取数据时，也提供了校验机制，保证了主机读取到的数据是从该 LU 上读的数据，而不是攻击者伪造的数据。这个 LU 就是 RPMB LU。

关于 RPMB，后面有专门章节介绍，这里不多说。

“四大名撸” 每个 LU 分工明确，分别执行不同的任务。下面把 “四大名撸” 能接收的命令列一下：

Well known logical unit	W-LUN	LUN Field in UPIU	Command name
REPORT LUNS	01h	81h	INQUIRY, REQUEST SENSE, TEST UNIT READY, REPORT LUNS
UFS Device	50h	D0h	INQUIRY, REQUEST SENSE, TEST UNIT READY, START STOP UNIT, FORMAT UNIT
Boot	30h	B0h	INQUIRY, REQUEST SENSE, TEST UNIT READY, READ (6), READ (10), READ (16)
RPMB	44h	C4h	INQUIRY,REQUEST SENSE, TEST UNIT READY, SECURITY IN, SECURITY OUT

他们能接收一些通用的命令（如上图绿色命令），还有只有该 LU 能执行的命令（如红色命令），具体命令可查看 Spec。

需要注意的是，写 Boot LU 和 RPMB LU 时，它是不支持 cache 操作的，就是说，数据必须写到闪存中以后，这笔写命令才算完成。而对一般 LU 的写，一般都是 cache 操作的，即主机数据到设备的内部 buffer，设备就会回命令完成状态给主机。

10.5.6 蛋蛋读 UFS 之六：UFS 设备初始化和启动

Posted on 2018 年 5 月 31 日 by 蛋蛋

原创内容，转载请注明： [http://www.ssdfans.com] 谢谢！

这一节讲讲 UFS 初始化。

初始化和启动包括三个阶段：部分初始化，加载启动代码（可选）和初始化完成。

- **部分初始化阶段**

这个阶段开始于上电或者设备重启，它涉及到整个 UFS 栈的初始化。

这个初始化阶段完成后，整个物理层（M-PHY）和数据链路层（UniPro）应该被初始化好，传输层可以和主机交互 Read 命令和“TEST UNIT READY”命令（主机发该命令给设备，查询设备是否准备好），主机也可以通过设备管理器访问设备描述符（Device Descriptor），获取设备配置信息。

- **加载启动代码**

如果启动代码不是存储在 UFS 设备上，则没有这一阶段。主机怎么知道启动代码是不是存储在 UFS 设备上呢？

经过前一阶段的初始化，主机可以访问设备描述符，获得“bBootLunEn ”属性，读取该属性可以知道启动代码是否在 UFS 设备上，以及具体在哪个 Boot LU 上面。

Table 13-1 — bBootLunEn Attribute

bBootLunEn	Description
00h	Boot LU A = disabled Boot LU B = disabled
01h	Boot LU A = enabled Boot LU B = disable
02h	Boot LU A = disable Boot LU B = enabled
Others	Reserved

如果 bBootLunEn = 01h 或者 02h，说明启动代码存储在 UFS 设备上。由于 Boot LU 是映射到普通的 LU 上的，要读取启动代码，还需要知道 Boot LU 和存储启动代码 LU 的映射。主机可以通过读取单元描述符（Unit Descriptor）知道，比如：

bBootLunID in Unit Descriptors		
LU	Value	Description
0	00h	Disabled
1	01h	Boot LU A
2	00h	Disabled
3	00h	Disabled
4	02h	Boot LU B
5	00h	Disabled
6	00h	Disabled
7	00h	Disabled

查找到具体存储代码的 LU，主机就可以读取该 LU 获得启动代码。

- **初始化完成**

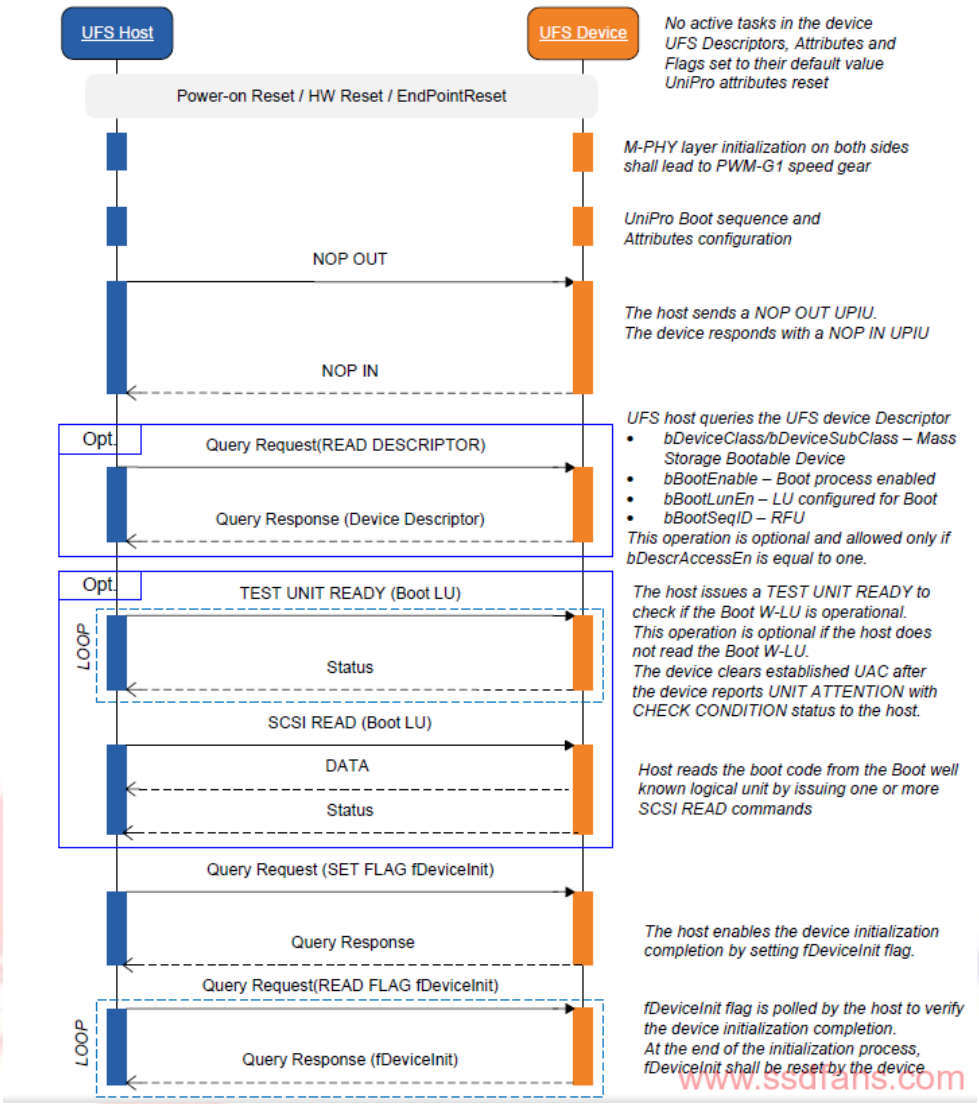
当主机完成前面两个阶段，主机会通过设备管理器，给设备设置 fDeviceInit = 1，这是一个标志（flag），用以初始化。 主机设置了该标志后，然后就一直在那查询该标志的值。

与此同时，UFS 设备的固件继续完成自己的初始化，当设备完成初始化，认为可以响应主机任何命令或者请求时，就清掉 fDeviceInit，即 fDeviceInit = 0。

当主机查询到 fDeviceInit =0，就可以发任何 UFS 协议中的任何命令或请求给 UFS 设备。

至此，整个 UFS 初始化和启动完成。

下图具体展示了 UFS 初始化和启动过程（可选的，Opt）：



再解释一下这个启动过程：

主机给设备上电或者重启设备，然后主机和设备端的物理层和数据链路层完成初始化，之后主机 ping 设备（通过 NOP OUT UPIU），确认设备双方连接正常。设备收到 NOP OUT UPIU，应该回 NOP IN UPIU，表明双方连接没有问题。

加载启动代码可选（上图蓝色方框中的步骤）。主机读取 UFS 设备描述符，如果 bDescrAccessEn = 0，设备描述符不可访问，那么，即使启动代码存储在 UFS 设备上，我们也无法在该阶段加载启动代码，因为诸如 bBootEnable 和 bBootLunEn 之类的信息无法获取，主机就无法知道存储代码存储在哪个 LU 上。因此，如果 bDescrAccessEn = 0，加载启动代码阶段不应该放在这里，而是在设备彻底初始化好后。

加载启动代码阶段，主机通过读取设备描述符，获得启动代码在哪个 LU 上，然后发个试探性命令“TEST UNIT READY”给该 LU，查看该 LU 是否准备好。如果 Boot LU 准备好，主机就通过发 READ 命令给设备，加载启动代码。

然后，主机设置 fDeviceInit = 1，然后一直轮询该标志，一旦 fDeviceInit 变成 0，标志 UFS 设备初始化完成。

最后，再把设备初始化过程中，双方交互的内容做个总结：

Table 13-2 — Valid UPIUs and SCSI Commands for Each Initialization Phase

Phase	Event	Valid UPIU	Valid SCSI command
Before Initialization	Power On Reset / HW Reset / EndPointReset	None	None
UIC Layer Initialization Phase	M-PHY layer initialization UniPro Boot sequence and UIC Layer Attributes configuration	None	None
UTP Layer Initialization Phase	Receive a single NOP OUT UPIU Send NOP IN UPIU for response to NOP OUT UPIU	NOP OUT UPIU NOP IN UPIU	None
Boot W-LU Ready Phase (Optional)	Read Device Descriptor (Optional) (1)	QUERY REQUEST UPIU (READ DESCRIPTOR Device Descriptor)	None
	Boot Transfer (Optional) (2)	COMMAND UPIU for Boot W-LU	INQUIRY, REQUEST SENSE, TEST UNIT READY, READ (6), READ (10), READ (16) (3)
Application Layer Initialization Phase	Receive QUERY REQUEST UPIU (SET FLAG fDeviceInit to '01h') Send QUERY RESPONSE UPIU with fDeviceInit = '00h'	QUERY REQUEST UPIU (SET FLAG fDeviceInit to '01h') QUERY REQUEST UPIU (READ FLAG fDeviceInit) QUERY RESPONSE UPIU	None
Device initialization completed Phase	Any normal operation	Any supported UPIU	Any supported SCSI commands

NOTE1 Device Descriptor may be read only if bDescrAccessEn is set to '01h'.

NOTE2 Boot well-known logical unit may be read if bBootEnable is set to '01h', at least one logical unit is configured for boot (bBootLunEn) and bBootLunID selects the desired boot logical unit.

10.5.7 蛋蛋读 UFS 之七：描述符、标识和属性

Posted on 2018 年 5 月 31 日 by 蛋蛋

原创内容，转载请注明： [http://www.ssdfans.com] 谢谢！

UFS 中也有吉祥三宝，那就是描述符（Descriptors）、标识（Flags）和属性（Attributes），主机通过这三宝，来控制与管理 UFS 设备。

- **描述符**

描述符是一块或者一页参数用以描述一个 UFS 设备，比如，UFS 有整个 UFS 设备的描述符 （Device Descriptor），UFS 设备的配置描述符（Configuration Descriptor），每一个 LU 还有其描述符（Unit Descriptor），等等。下面是 UFS 里面所有种类的描述符。

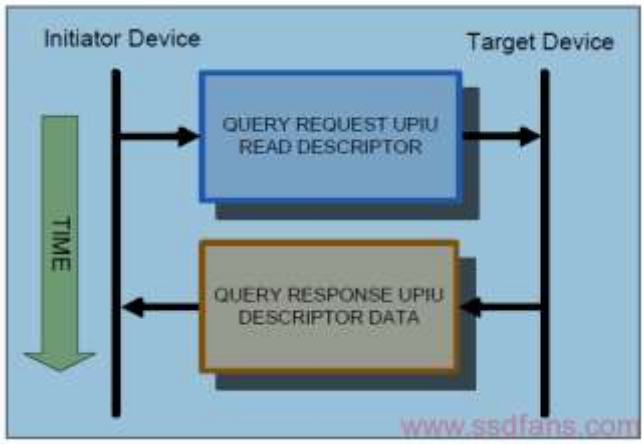
Table 14-1 — Descriptor identification values

Descriptor IDN	Descriptor Type
00h	DEVICE
01h	CONFIGURATION
02h	UNIT
03h	Reserved
04h	INTERCONNECT
05h	STRING
06h	Reserved
07h	GEOMETRY
08h	POWER
09h	DEVICE HEALTH
0Ah ... FFh	Reserved

除了配置描述符和 OEM_ID 字符串描述符，所有的描述符都是只读的，即 UFS 设备一旦出厂，主机是不能对它进行修改。

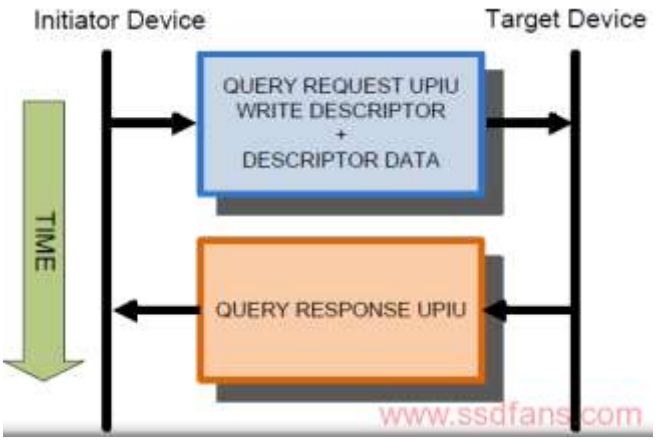
还记得 Query Request UPIU 吗？主机是通过设备管理器来访问这些描述符的。

主机读描述符：



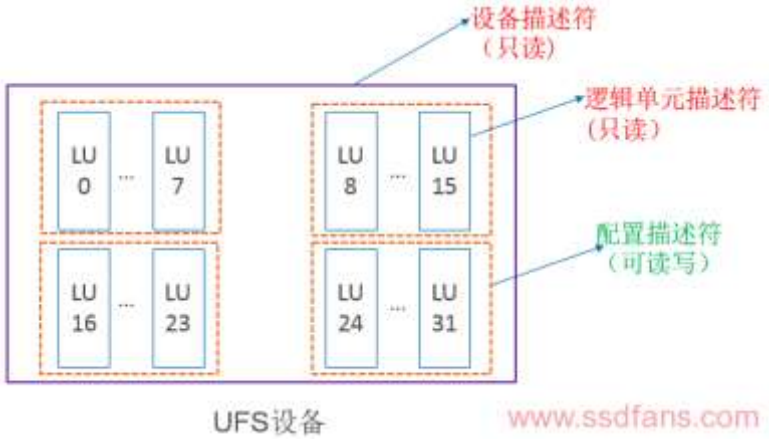
主机通过发 Query Request UPIU 给 UFS 设备，然后设备通过 Query Response UPIU 返回描述符数据。

主机写描述符：



主机如果想更改配置，可以写**配置描述符**。主机要写入的数据包含在 Query Request UPIU 中，一旦 UFS 设备更新完，返回 Query Response UPIU。

我们简单过一下上面的这些描述符，更加详细的描述大家可以自行看 UFS spec。



UFS 设备只有一个，所以只有一个设备描述符；

一共有 32 个普通 LU，每个 LU 有一个逻辑单元描述符，所以最多有 32 个逻辑单元描述符；

每 8 个 LU 有一个配置描述符，所以一共最多 4 个配置描述符。

1. 设备描述符

设备描述符就是描述整个 UFS 设备属性的描述符，这些参数在 UFS 设备出厂时就由厂家设置好，主机对它只可读不可写。

设备描述符					
偏移量	大小	名字	出厂默认设置	可配置?	说明
00h	1	bLength	40h	否	设备描述符大小。固定为 64 字节。
01h	1	bDescriptorIDN	00h	否	设备描述符类型编号
…	…	…	…	…	…
06h	1	bNumberLU	00h	是	普通 LU 的个数。
07h	1	bNumberWLU	04h	否	Well Known LU 的个数，UFS2.1 中有“四大名捕”。
08h	1	bBootEnable	00h	是	标识是否支持从 UFS 设备启动。如果该位为 0，表示不支持 UFS 设备启动，在初始化过程时，就没有加载启动代码的阶段。（见前一章）
09h	1	bDescrAccessEn	00h	是	标识设备在部分初始化（Partial）后，主机能否访问设备描述符。
…	…	…	…	…	…
12h	2	wManufactureDate	设备相关	否	设备出厂日期。比如 August 2010 = 0810h
14h	1	iManufactureName	设备相关	否	设备生产厂家名称。这里只是一个索引。索引到包含厂家名称的字符串中。
15h	1	iProductName	设备相关	否	设备产品名称。同样是个索引。索引到包含产品名称的字符串中。
16h	1	iSerialNumber	设备相关	否	设备序列号。同样是个索引。索引到包含产品序列号的字符串中。
…	…	…	…	…	…

上面只是截取了一部分设备描述符数据结构内容，有关完整的设备描述符内容，大家可以看 spec。

前面说了设备描述符是只读属性，为什么我们看到设备描述符里的有些项是可配置的呢？设备描述符的确只可能读不可写，但是主机通过写配置描述符（主机可写），然后这些项的变化就反映到设备描述符里来了。

UFS 设备只有一个，所以一共只有一个设备描述符。

2. 逻辑单元（LU）描述符

逻辑单元描述符用来描述某个具体 LU 的特性和能力，比如该 LU 逻辑块大小、该逻辑块是不是存有启动代码、该逻辑块内存类型等等。

逻辑单元描述符

偏移量	大小	名字	出厂默认设置	可配置?	说明
00h	1	bLength	23h	否	逻辑单元描述符大小，固定为 35 字节。
01h	1	bDescriptorIDN	02h	否	逻辑单元描述符类型编号
...	...	...	...	...	...
03h	1	bLUEnable	00h	是	逻辑单元使能。 1：该逻辑单元使能。 0：该逻辑单元禁止。
04h	1	bBootLunID	00h	是	Boot LUN ID: 00h：该逻辑单元没有映射到任何 Boot LU; 01h：该逻辑单元映射到 Boot LU A; 02h：该逻辑单元映射到 Boot LU B. 其它：保留。
05h	1	bLUWriteProtect	00h	是	标识该逻辑单元是否写保护。
06h	1	bLUQueueDepth	设备相关	否	每个逻辑单元有个命令队列，该域指明命令队列深度。 0：没有自己的命令队列，和其它 LU 共享命令队列; 1-255：该 LU 自己的队列深度
...	...	...	...	...	...
08h	1	bMemoryType	00h	是	指定 LU 内存类型: 00h：存储普通数据 01h：存储系统代码 02h：挥发性存储类型 03-06h：分别对应增强型存储类型 1-4，对应增强型数据，在闪存上，可能获存储在 SLC 上确保其存储可靠性。
...	...	...	...	...	...
0Ah	1	bLogicalBlockSize	0Ch	是	逻辑块大小，默认为 4KB，用户可配置逻辑块大小。

对于逻辑单元描述符中可配置的项，主机可以通过写配置描述符进行相应的更改。

3. 配置描述符

用户想对 UFS 做一些配置，或者使能/禁止一些 feature， 可以通过写配置描述符达到目的。这些项的更新会反映到设备描述符或者逻辑单元描述符上。注意只有在属性（Attribute）bConfigDescrLock = 0 时才可以写配置描述符，即配置描述符没有被锁住，配置描述符才能写，否则也是只读的。

UFS2.1 有 32 个普通的 LU，每 8 个 LU 有个配置描述符，所以一共有 4 个描述符。

拿出一个配置描述符来看看它的格式。

Table 14-6 — Configuration Descriptor Format (INDEX = 00h)

Offset	Description
00h ... (B-1)h	Configuration Descriptor header and Device Descriptor configurable parameters
(B)h ... (B+L-1)h	Unit Descriptor 0 configurable parameters
(B+L)h ... (B+2*L-1)h	Unit Descriptor 1 configurable parameters
...	...
(B+7*L)h ... (B+8*L-1)h	Unit Descriptor 7 configurable parameters

该配置描述符，包含了设备描述符可配置的参数和 LU 0-7 中可配置的参数。

Table 14-10 — Configuration Descr. Header and Device Descr. Conf. parameters (INDEX = 00h)

Configuration Descriptor Header and Device Descriptor configurable parameters				
Offset	**Size**	**Name**	**MDV (1,2)**	**Description**
00h	1	bLength	90h	Size of this descriptor
01h	1	bDescriptorIDN	01h	Configuration Descriptor Type Identifier
02h	1	bConfDescContinue	00h	00h : This value indicates that this is the last Configuration Descriptor in a sequence of write descriptor query requests. Device shall perform internal configuration based on received Configuration Descriptor(s). 01h : This value indicates that this is not the last Configuration Descriptor in a sequence of write descriptor query requests. Other Configuration Descriptors will be sent by host. Therefore the device should not perform the internal configuration yet.
03h	1	bBootEnable		Boot Enable Enables to boot feature.
04h	1	bDescrAccessEn		Descriptor Access Enable Enables access to the Device Descriptor after the partial initialization phase of the boot sequence.

.....

举例来说，bBootEnable 是设备描述符中一个可配置的项。它在出厂设置时 bBootEnable = 0，用户在使用 UFS 设备时，把启动代码存放在 UFS 设备上，因此，为使能 Boot feature, 用户须通过写配置描述符把该比特置起来：bBootEnable = 1。然后，主机在读取设备描述符的时候，会看到 bBootEnable 变成了 1。

除了配置 UFS 设备，配置描述符还可以对每一个 LU 进行配置。

Table 14-12 — Unit Descriptor configurable parameters

Unit Descriptor configurable parameters			
Offset	**Size**	**Name**	**Description**
00h	1	bLUEnable	Logical Unit Enable
01h	1	bBootLunID	Boot LUN ID
02h	1	bLUWriteProtect	Logical Unit Write Protect
03h	1	bMemoryType	Memory Type
04h	4	dNumAllocUnits	Number of Allocation Units Number of allocation units assigned to the logical unit. The value shall be calculated considering the capacity adjustment factor of the selected memory type
08h	1	bDataReliability	Data Reliability
09h	1	bLogicalBlockSize	Logical Block Size
0Ah	1	bProvisioningType	Provisioning Type
0Bh	2	wContextCapabilities	Context Capabilities
0Dh:0Fh	3	Reserved	

比如，主机可以通过写配置描述符，使能某个 LU，或者设置某个 LU 的逻辑块大小，以及其它和 LU 相关的配置。

UFS 中还有其它的一些描述符，这里就不一一细看。

我们接下里看看另一宝：标志（flags）。

- **标志（flags）**

UFS 中的标志其实就是一些开关，布尔型，非 0 即 1，打开或者关闭。这些标志可以用来使能或者禁止 UFS 设备的一些功能、模式或者状态。

在 UFS2.1 协议中，一共有以下一些标志：

名字	说明
fDeviceInit	用以设备初始化。主机在加载启动代码后设置该标志，设备初始化完毕后清除该标志。
fPermanentWPEn	该标志使能所有配置为永久保护的逻辑单元的写保护，一旦设置为永久写保护后，该标志不能被清除。 0：永久写保护禁止；1：永久写保护使能。
fPowerOnWPEn	该标志使能所有配置为上电写保护的逻辑单元的写保护，一旦设置为永久写保护后，该标志不能被清除。 0：上电写保护禁止；1：上电写保护使能。
fBackgroundOpsEn	0：UFS 设备被禁止后台操作（比如垃圾回收操作）； 1：UFS 设备允许运行后台操作。
fDeviceLifeSpanModeEn	0：设备寿命模式（牺牲性能获取寿命）禁止； 1：设备寿命模式使能；
fPurgeEnable	0：Purge 操作（把垃圾数据从闪存中擦除掉）被禁止； 1：Purge 操作使能；
fPhyResourceRemoval	☹ 看 UFS spec
fBusyRTC	Busy Real Time Clock: 0：设备没有执行跟 RTC 相关的内部操作； 1：设备正在执行跟 RTC 相关的内部操作。 当设备正在执行跟 RTC 相关的内部操作，建议这个时候主机不要打扰设备，不要发命令给设备。
fPermanentlyDisableFwUpdate	0：设备固件可以被更新； 1：禁止更新设备固件。

主机也是通过设备管理器的 Query Request UPIU 来读取或者写标志。

- **属性（Attributes）**

如果说 flags 是布尔类型，那么属性就是 C 语言中的枚举类型。属性的值不仅仅是 0 或者 1，它是一定数字范围的。属性可以表示设备的一些状态，比如当前设备后台任务的状态。有些属性，主机只可读，有些属性，主机可以写。

主机也是通过设备管理器的 Query Request UPIU 来读取或者写属性。

名字	说明
bBootLunEn	标识从哪个 Boot Lun 加载启动代码。 00h：禁止从 UFS 设备启动； 01h：从 Boot LU A 启动； 02h：从 Boot LU B 启动； 其它：保留。
bCurrentPowerMode	当前功耗模式。 00h：Idle； 10h：Pre-Active； 11h：Active； 20h：Pre-Sleep； 22h：Sleep； 30h：Pre-PowerDown； 33h：PowerDown； 其它：保留。
bActiveCCLevel	Active ICC 级别定义了在 UFS 设备在工作功耗（Active）模式下允许的最大电流消耗： 00h：最低 Active ICC 级别； … 0Fh：最高 Active ICC 级别； 其它：保留。
bOutOfOrderDataEn	乱序数据传输使能。 00h：禁止乱序数据传输； 01h：允许乱序数据传输，即一个命令当中的逻辑块数据，可以不按 LBA 的顺序传输。 其它：保留。
bBackgroundOpStatus	UFS 设备后台操作状态。 00h：不需要后台操作； 01h：需要，但不急（Critical）；

	02h：需要，对性能有影响； 03h：迫切需要做后台操作； 其它：保留。
bPurgeStatus	UFS 设备清除操作（把设备里无效的数据清除）状态。 00h：空闲，清除操作被禁止； 01h：UFS 设备正在执行清除操作； 02h：UFS 设备清除操作被 host 提前终止了； 03h：UFS 设备成功完成清除操作； 04h：清除操作失败，因为逻辑单元的命令队列非空； 05h：清除操作一般性失败； 其它：保留
bMaxDataInSize	指定 DATA IN UPIU 中最大传输数据大小（设备向主机传输数据）。该值以 512 字节为单位，不能超过 bMaxInBufferSize。
bMaxDataOutSize	指定 DATA OUT UPIU 中最大传输数据大小（主机向设备传输数据）。该值以 512 字节为单位，不能超过 bMaxOutBufferSize。设备向主机发 READY TO TRANSFER UPIU 时，请求的数据量不能超过 bMaxDataOutSize。
dDynCapNeeded	Dynamic Capacity Needed. UFS 支持动态调整某个 LU 的容量。该值用以指定需要某个 LU 的物理空间减小多少容量。
bRefClkFreq	参考时钟频率。 0h：19.2MHz； 1h：26MHz； 2h：38.4MHz； 3h：52MHz； 其它：保留。
bConfigDescrLock	配置描述符锁。 0h：配置描述符没有锁住，主机可以读写配置描述符；

	1h：配置描述符锁住了，主机不能写配置描述符（只读）； 其它：保留。
bMaxNumOfRTT	允许设备发的最多的 Outstanding RTTs 个数，不能超过 bDeviceRTTCap。
wExceptionEventControl	异常事件控制（Exception Event Control）。 Bit 0：用以使能/禁止动态容量调整事件； Bit 1：用以使能/禁止系统池（存储系统数据的存储空间）事件； Bit 2：用以使能/禁止紧急后台事件； Bit 3-15：保留。
wExceptionEventStatus	异常事件状态（Exception Event Status）。 Bit 0：标识是否需要动态调整容量，比如当闪存坏块出现很多时，动态的把用户容量缩小； Bit 1：用以标识系统池是否耗尽； Bit 2：用以标识是否有紧急后台操作； Bit 3-15：保留。
…	…

10.5.8 蛋蛋读 UFS 之八：RPMB

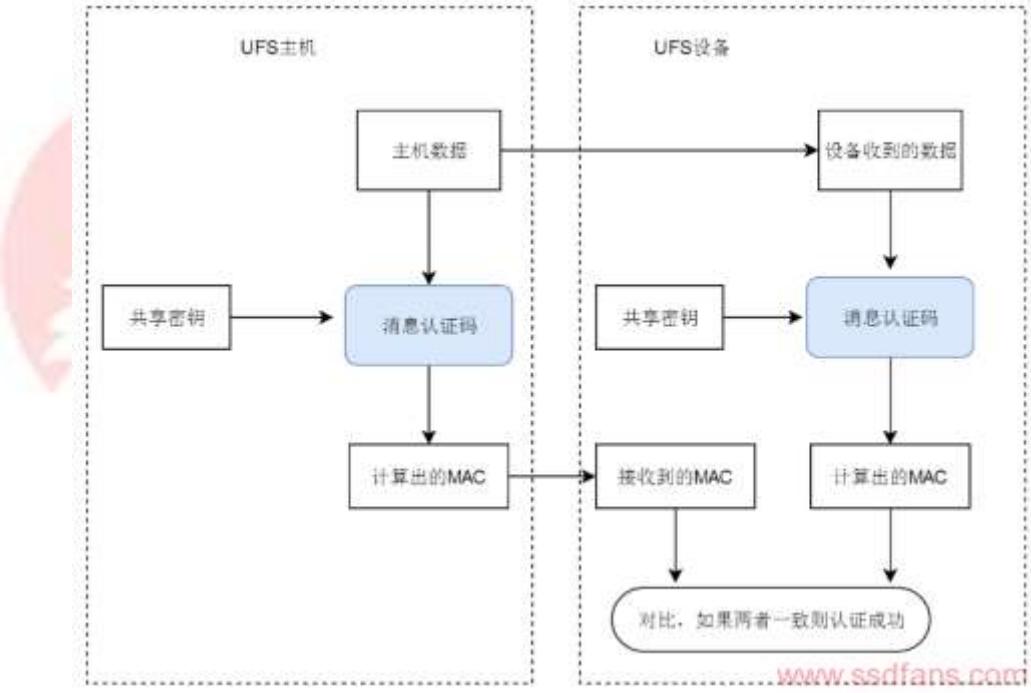
Posted on 2018 年 5 月 31 日 by 蛋蛋

在 UFS 里，有这么一个 LU，主机往该 LU 写数据时，UFS 设备会校验数据的合法性，只有特定的主机才能写入；同时，主机在读取数据时，也提供了校验机制，保证了主机读取到的数据是从该 LU 上读的数据，而不是攻击者伪造的数据。这个 LU 就是 RPMB（Replay Protected Memory Block）LU，四大“名撸”（四个 Well Known LU）之一。

有些人家里有保险箱，用以存放他们认为重要的东西，比如现金、存折、房产证、情书等。输入密码，打开密码箱，然后放东西进去；取的时候，首先需要密码打开保险箱，然后把东西取出。没有密码，老婆是万万看不到老公和他初恋之间的情书的。RPMB 就像是手机里的密码箱，用户可以把一些重要数据存储其中。

我们来看看 RPMB 这个保险箱。

UFS 主机通过认证（authenticated）的方式访问 RPMB LU。下图展示了数据写过程：



1. 首先，UFS 主机和 UFS 设备共享密钥，该密钥在 UFS 设备出厂时就保存在 UFS 设备；
2. UFS 主机在发送主机数据给 UFS 设备前，会用该密钥和哈希算法生成消息认证码 (Message Authentication Code, MAC)；
3. UFS 主机把主机数据连同 MAC 一起发给 UFS 设备；
4. UFS 设备把收到的主机数据和共享密钥在本地重新计算 MAC，然后把计算出的 MAC 和收到的 MAC 做对比，如果一致，则认证成功,写入到闪存；否则,拒绝该笔数据的写。

UFS 使用 HMAC（Hash-based Message Authentication Code）SHA-256 算法生成消息认证码。HMAC 运算利用哈希算法，以一个密钥和一个消息为输入，生成一个消息摘要作为

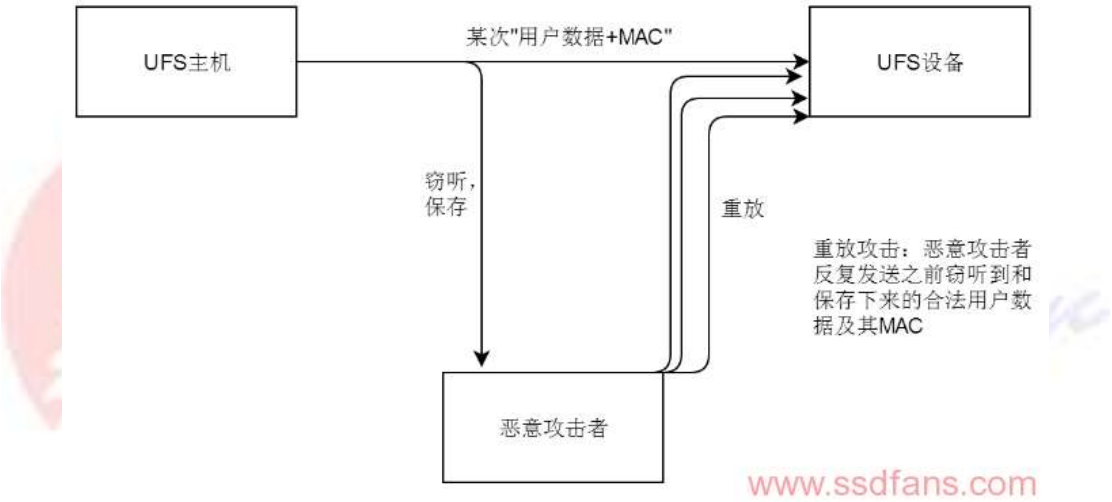
输出。关于 HMAC 具体算法，可参看 https://en.wikipedia.org/wiki/HMAC，我们这里不深入。

消息认证码本质是哈希值。哈希的一个特点是，即使只改变原数据一比特数据，两者的哈希值也是完全不同的。如果恶意攻击者在数据传输过程中篡改了用户数据，那么 UFS 设备根据收到的数据和共享密钥生成的 MAC 肯定与接收到的 MAC 不一样，认证通不过，数据就不会写入 UFS 设备。

这里的前提是共享密钥不能被恶意攻击者获取，否则，恶意攻击者完全可以模拟主机行为：把自己的恶意数据和共享密钥生成 MAC，然后把恶意数据和其对应的 MAC 发送给 UFS 设备。UFS 设备会认证成功,恶意数据被写入。所以，请保管好你的密码!

但是，恶意攻击者是狡猾的，即使他没有办法获得你的密钥，它还是有办法对你进行攻击的。

恶意攻击者监听到 UFS 主机和 UFS 设备之间某次数据传输，得到“主机数据 + MAC”，然后该恶意攻击者重复发送该“主机数据 + MAC”给 UFS 设备，由于“主机数据 + MAC”是合法的，认证通过，UFS 设备就会接收该数据并写到闪存。恶意攻击者如果一直重复发这些数据给 UFS 设备，UFS 设备 RPMB LU 将会被写爆！这就是重放攻击, Replay Attack。



RPMB 的全名是：Replay Protection Memory Block，它的名字暗示了 RPMB 是能抵御重放攻击的。那么 RPMB 是怎么对付重放攻击的呢？

UFS 维护了一个写计数（Write Counter），初始化为 0。UFS 设备每次成功处理完一个 RPMB 写命令，写计数加一。主机在往设备写入数据前，获得该计数。然后把用户数据和该计数一起做 MAC 计算。这样，即使恶意攻击者窃听到某次合法的“用户数据 + MAC”，往设备写入时，由于写计数发生变化，它无法生成写计数改变之后的 MAC 值，因此就无法一直重复往设备写入某次合法的“用户数据 + MAC”。魔高一尺，道高一丈，正义终战胜邪恶！

上面就是 RPMB 数据安全性背后的原理。下面再回到 UFS RPMB 协议上来。

UFS2.1 中，RPMB LU 最小逻辑空间为 128KB，最大为 16MB。它的逻辑块大小为 256B（普通 LU 逻辑块大小一般为 4KB）。应用层不是通过普通的 Read/Write 命令读/写 RPMB 上的数据，而是通过 SECURITY PROTOCOL OUT/IN 命令来访问 RPMB 的。

	逻辑空间大小	逻辑块大小	访问命令	安全性	存储数据类型
RPMB LU	128KB -16MB	256B	Security Protocol OUT/IN	数据认证	防篡改的重要数据
普通 LU	无限制	一般为 4KB	Read/Write	无	普通用户数据

UFS 主机在访问设备 RPMB 时，是通过下面消息交互完成的。

请求（Request）	响应（Response）	说明
Authentication Key programming request	Authentication Key Programming response	用在写认证密钥
Write Counter read request	Write Counter read response	用在读取写计数
Authenticated data write request	Authenticated data write response	用在写认证数据
Authenticated data read request	Authenticated data read response	用在读认证数据
Result read request	NA	读取操作结果
Secure Write Protect Configuration Block write request	Secure Write Protect Configuration Block write response	用于写安全写保护配置块
Secure Write Protect Configuration Block read request	Secure Write Protect Configuration Block read response	用于读安全写保护配置块

每条消息包含一条或者若干条消息数据帧。消息数据帧大小是 512 字节，具体如下：

Table 12-9 — RPMB Message Data Frame

Bit / Byte	7	6	5	4	3	2	1	0
0	(MSB)	Stuff Bytes						
195								(LSB)
196	(MSB)	Key / MAC						
227								(LSB)
228	Data [255]							
483	Data [0]							
484	(MSB)	Nonce						
499								(LSB)
500	(MSB)	Write Counter						
503								(LSB)
504	(MSB)	Address						
505								(LSB)
506	(MSB)	Block Count						
507								(LSB)
508	(MSB)	Result						
509								(LSB)
510	(MSB)	Request Message Type / Response Message Type						
511								(LSB)

从中，我们看到：

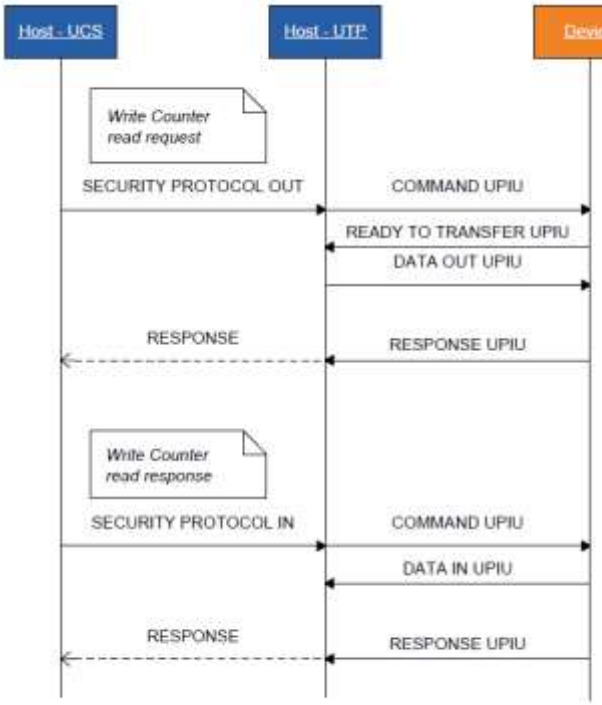
- 认证密钥（Key）是 32 字节；
- 1 使用 SHA-256 计算 MAC，就是任意长度的数据，产生的 MAC 值总是 256 比特，即 MAC 大小为 32 字节。
- 逻辑块数据大小为 256 字节。

- 写计数（Write Counter）为 4 字节，当该值涨到 0xFFFF FFFF，它就保持不动，不会继续增长了。
- Address，RPMB 的逻辑地址，同 LBA。两个字节，最多表示 65536 个逻辑块，每个逻辑块大小为 256 字节，因此 RPMB 逻辑空间最大为
- Block Count，逻辑块数，即指定读写多少个逻辑块。
- Result，RPMB 操作结果（状态）。

下面举几个 RPMB 操作例子来理解上面的消息：

- **主机读取写计数**

如前所述，写计数的目的是抵御重放攻击。写计数是 UFS 设备管理和维护的，UFS 设备递增该计数。主机在写数据时，需要知道该计数，然后加上用户数据，一起计算 MAC。



Write Counter Read Request

Offset	Field Name	Value
0:195	Stuff bytes	0...00h
196:227	MAC / Key	0...00h
228:483	Data	0...00h
484:499	Nonce	*Nonce from the host*
500:503	Write counter	0...00h
504:505	Address	0000h
506:507	Block Count	0000h
508:509	Result	0000h
510:511	Req./Resp.	0002h

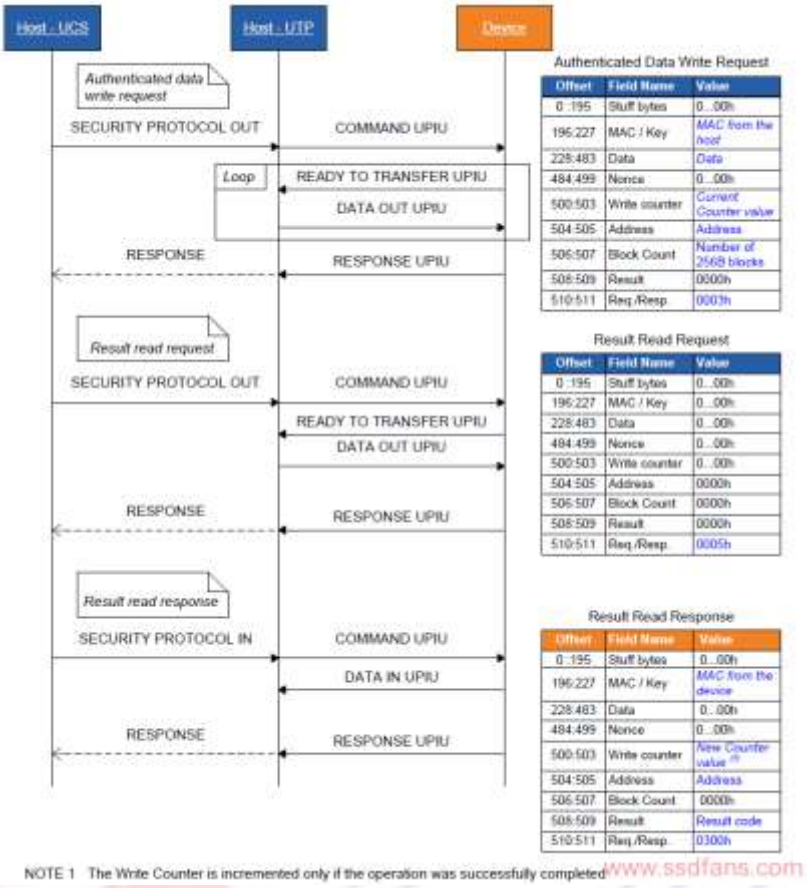
Write Counter Read Response

Offset	Field Name	Value
0:195	Stuff bytes	0...00h
196:227	MAC / Key	*MAC from the device*
228:483	Data	0...00h
484:499	Nonce	*Copy of the Nonce*
500:503	Write counter	*Counter Value*
504:505	Address	0000h
506:507	Block Count	0000h
508:509	Result	*Result code*
510:511	Req./Resp.	0200h

命令层发 SECURITY PROTOCOL OUT/IN 命令读取写计数，然后传输层生成相应的 UPIU 进行主机与设备之间的交互，具体见上图。

- **主机写认证数据**

主机命令层通过 SECURITY PROTOCOL OUT 命令把用户数据和对应的 MAC 发送给设备，然后通过 SECURITY PROTOCOL OUT 请求获取前面数据写结果，最后通过 SECURITY PROTOCOL IN 读取写结果。写结果中包含新的写计数，这样下次主机利用新的写计数计算 MAC。注意，只有本次写认证数据成功，设备才会递增该计数。

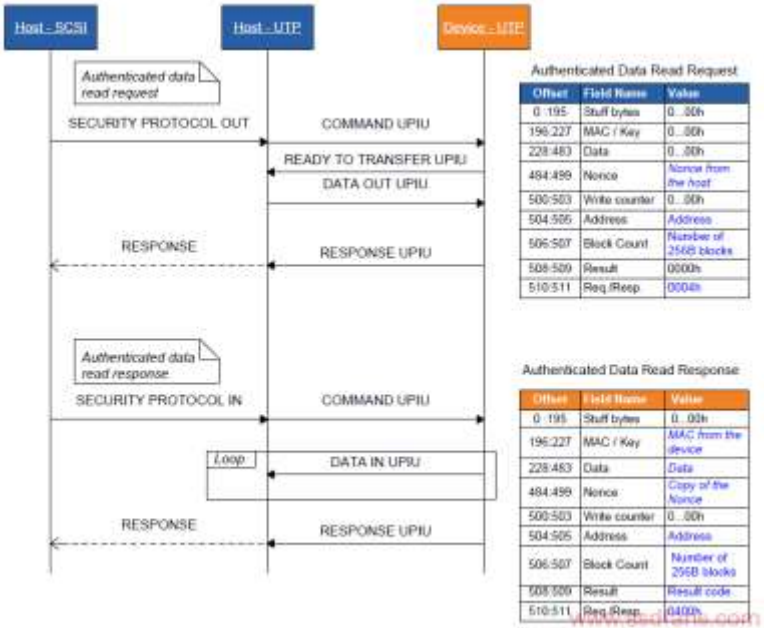


NOTE 1 The Write Counter is incremented only if the operation was successfully completed

- **主机读认证数据**

首先，主机通过 SECURITY PROTOCOL OUT 命令发送读取认证数据请求给设备，然后发送 SECURITY PROTOCOL IN 命令读取数据。

注意，主机读取数据也是需要认证的。在设备端，UFS 设备会计算 MAC，然后主机端根据 MAC 认证该数据。这样可以防止恶意攻击者在数据传输过程中（从设备到主机），用恶意数据更换原始数据。



RPMB 提供了认证访问方式和抵御重放攻击的机制，保证了存储在 RPMB LU 上数据的安全。因此，用户可以把一些敏感和重要的信息写在 RPMB 上。在实际应用中，它通常用于存储一些有防止非法篡改需求的数据，例如手机上指纹支付相关的公钥、序列号等敏感信息。

10.5.9 蛋蛋读 UFS 之九：UFS 数据安全

Posted on 2018 年 5 月 31 日 by 蛋蛋

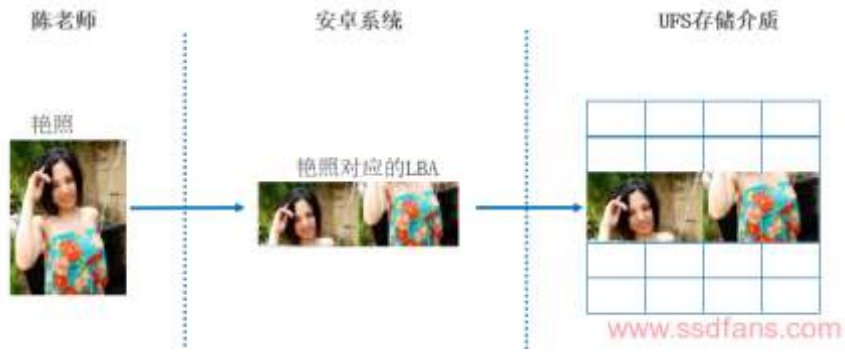
前面提到 RPMB 使用认证机制和抗重放攻击机制保障数据不被黑客攻击，除此之外，UFS 还有其它一些手段来保护用户数据安全，这一章节我们来关注 UFS 数据安全。

UFS 设备用来存储个人或者企业的数据信息，UFS 设备需要这样一种机制，就是必要时，数据能永久从设备（闪存）删除，这样就能防止别有用心的人通过反向工程获取你的数据。

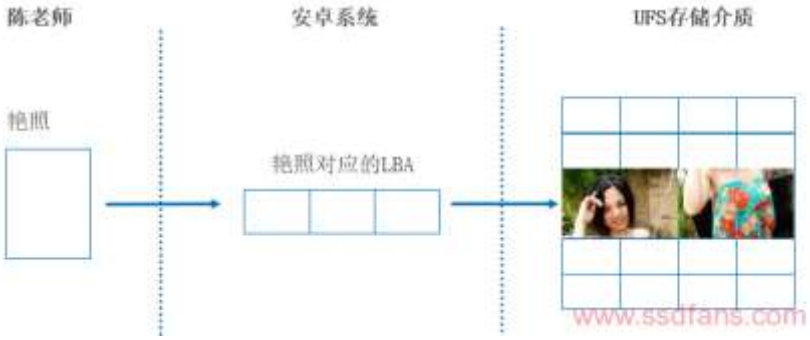
话说，陈老师吸取了上次教训，每次拍照后，事后“阅后即焚”。陈老师想：这样就没有人看到我们的照片了吧。陈老师很是得意。

没有想到，不久后网上又爆出陈老师新的“艳照门”事件。陈老师很是纳闷，我不是明明都删除了吗？？

我们帮陈老师分析一下为什么删了的照片还能被修手机的人弄出来。



手机文件系统把陈老师拍的照片数据用逻辑块管理，然后把这些逻辑块写到 UFS 设备的存储介质（如上图所示）。陈老师删除照片，删除的只是逻辑块数据，存在 UFS 存储介质上的数据还在原地，如下图所示：

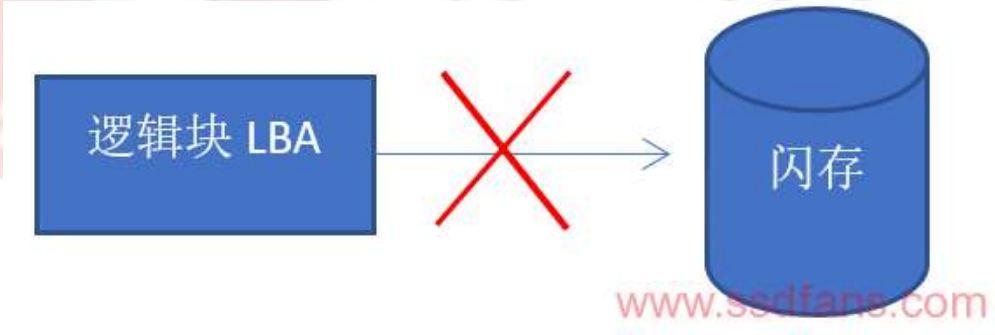


修手机的一看，这不陈老师吗？修手机的知道肯定能发现什么，嘴角不禁露出一丝不易觉察的笑。他从文件中没有找到照片，有点失望。小子吃一堑长一智呀！但猎奇的心不会让他轻易放弃的，有经验的他盯上了存储卡。功夫不负有心人，他从存储介质里把照片弄出来！

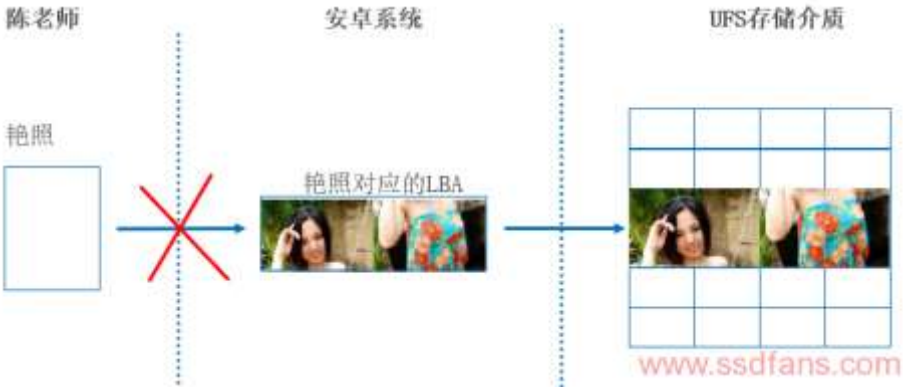
拍照不容易，且拍且珍惜。我们看看如何避免陈老师的悲剧。

- **擦除操作（Erase Operation）**

注意，这个“擦除”操作不是擦除存储介质，不是闪存层面的擦除操作，而是 UFS 层面的擦除操作。数据写在闪存上，UFS 设备内部有个逻辑地址到物理地址的映射，擦除操作通过切断这种映射，主机就不能获得擦除掉的数据。



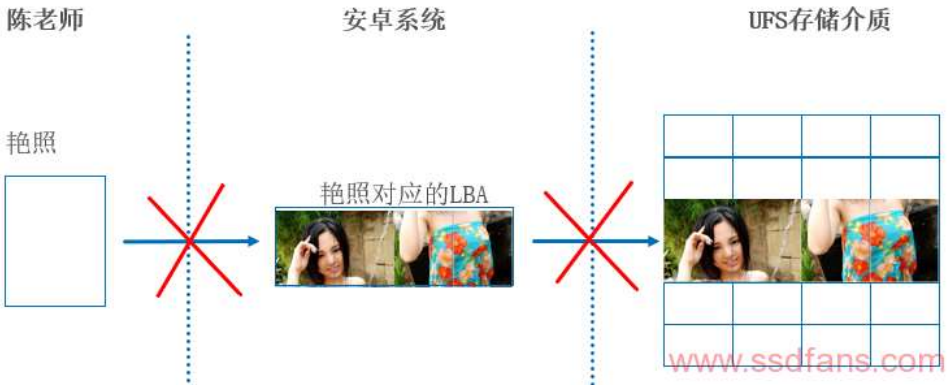
当陈老师删除照片时，它切断了用户直接访问照片的途径：



当陈老师删除照片后，手机系统会通过发送 UNMAP 命令（就是通常说的 TRIM）来告诉设备这些照片数据无效。设备收到该命令后，然后根据逻辑单元描述符中的 **bProvisioningType** 来确定执行具体操作。

bProvisioningType：
00h: Thin Provisioning is disabled (default)
02h: Thin Provisioning is enabled and TPRZ = 0 (Discard)
03h: Thin Provisioning is enabled and TPRZ = 1 (Erase)

即当 **bProvisioningType** = 03h 时，设备执行擦除操作，即切断逻辑地址到物理空间的映射。



一个逻辑块如果被擦除，那么主机访问这个逻辑块时，设备必须返回全 0 数据给主机。

注意，这个“擦除”操作不是擦除存储介质，只是主机让设备切断逻辑地址到物理地址的映射，因此不保证照片数据从闪存介质删除。但是，由于 UFS 设备知道该照片数据已经删除（没有逻辑块到物理空间的映射），在后续垃圾回收时，这些被删掉的数据很大概率会从介质上擦除掉。

- **舍弃操作（Discard Operation）**

和擦除操作类似，主机通过发送 UNMAP 命令来执行舍弃操作。当 bProvisioningType = 02h 时，设备执行舍弃操作。

舍弃操作和擦除操作的区别：主机访问一个被舍弃的逻辑块，可能获得任何数据，甚至包括舍弃前的数据，而擦除操作是主机获得全 0 数据。也就是说，对删除的照片，如果 UFS 设备执行的是舍弃操作，那么主机还可能获得原图片；如果 UFS 设备执行的是擦除操作，主机不可能再获得原照片。

但不管是舍弃操作还是擦除操作，都不能保证照片从存储介质上删除。像修手机这样的人，它不走寻常路（通过手机系统），直接操作闪存的话，还是有可能把删除的照片找回来。

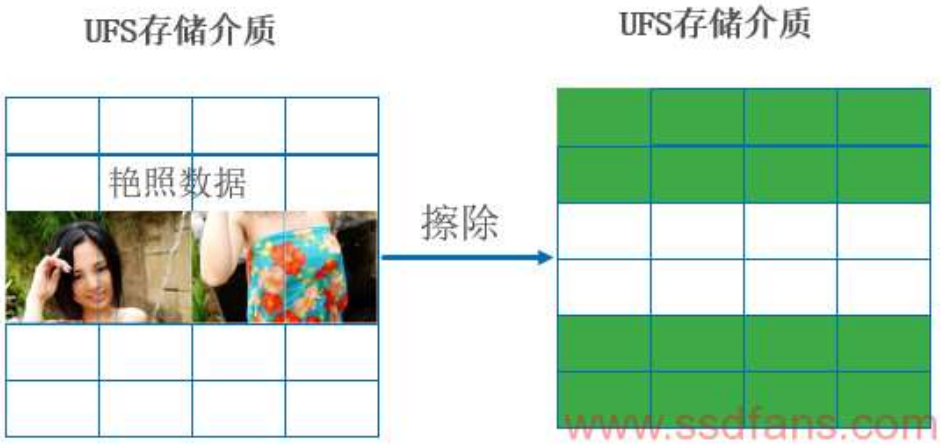
陈老师看到这里，急了，难道我以后再也不能拍照了吗？？

别急，小陈！你先坐下，听我慢慢讲。

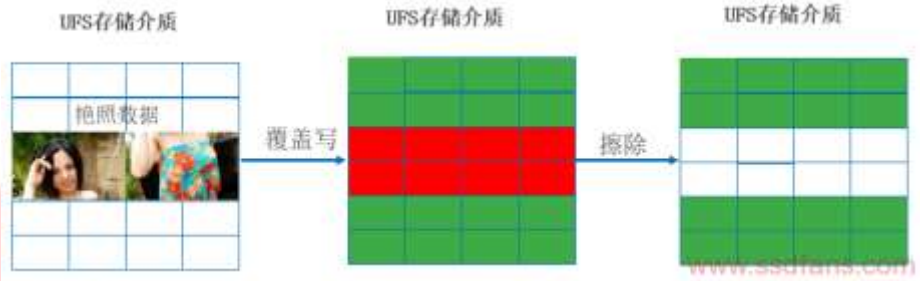
- **安全清除（Secure Removal）**

陈老师，有三种可选策略用以安全清除数据，你造吗？

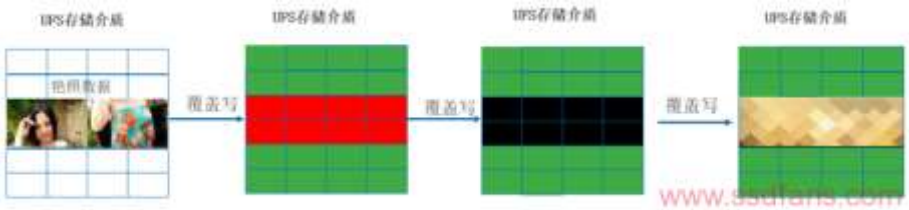
1. 设备控制器擦除（Erase）要被删除的逻辑块所对应的物理地址空间；



2. 设备控制器用单一字符覆盖写要被删除的逻辑块所对应的物理地址空间，然后擦除设备；



3. 设备控制器用单一字符、字符补码和随机字符，依次覆盖写要被删除的逻辑块所对应的物理地址空间。



又是覆盖写，又是擦除，照片是彻底从闪存中删除了。

陈老师听到这里，情绪缓和下来，终于是坐了下来。

- **清除操作（Purge Operation）**

清除操作是针对垃圾数据(比如陈老师删除的照片），让这些数据不仅不能通过正规渠道（操作系统）访问，还让这些数据无法从存储介质中获取，彻底把垃圾数据从 UFS 设备清除掉。

前面所说的擦除和舍弃操作，都是主机通过命令层的 UNMAP 命令来实施的。而清除操作则是主机通过设备管理器的 Query 功能来告诉设备的。

这里涉及到一个重要的标志（flag）和一个重要的属性（Attribute），分别是发 PurgeEnable 和 bPurgeStatus，前者用以使能/禁止清除操作，后者用以设备向主机提供清除操作的状态信息。

fPurgeEnable:

1. 上电或者重启，该标志位 0；
2. 主机通过设置或者清除该标志，使能或禁止清除操作；

3. 只有当所有逻辑单元的命令队列空的时候，主机才能设置该标志为 1 使能清除操作；
4. 当 UFS 设备执行完清除操作或者发生错误，该标志会被清零；

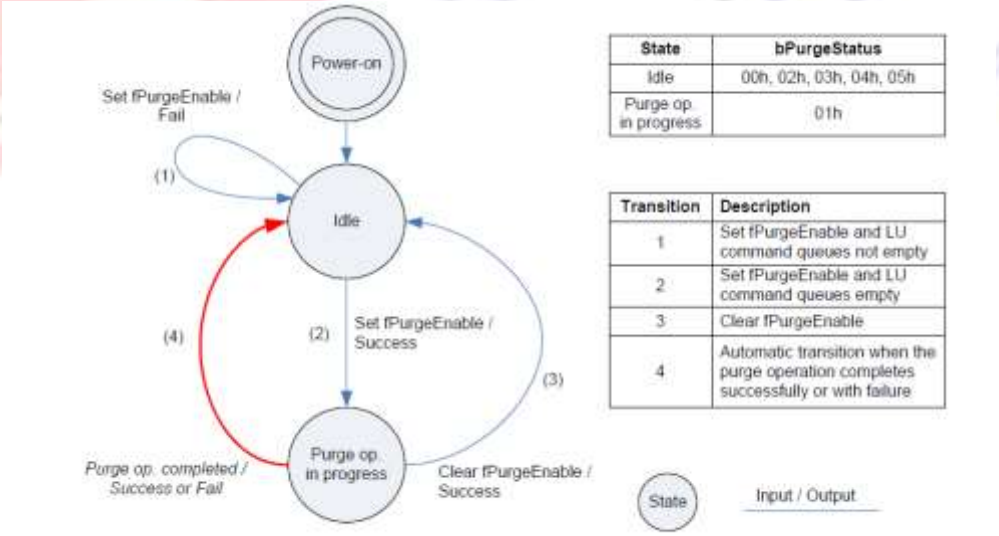
如果主机想终止设备执行清除操作，可以通过清除该标志达到目的。

bPurgeStatus：

bPurgeStatus 的值	状态
00h	清除操作被禁止
01h	设备正在执行清除操作
02h	清除操作被主机提前终止
03h	清除操作成功完成
04h	清除操作失败因为逻辑单元的命令队列非空
05h	清除操作一般性错误

主机为了让设备执行清除操作，主机通过 QUERY REQUEST UPIU 设置 fPurgeEnable = 1。如果当前逻辑单元的命令队列中没有任何命令，设备会执行清除操作。一旦设备开始执行清除操作，它不会响应主机发来的任何命令。如果这个时候主机需要让设备紧急响应命令，主机首先应该通过 QUERY REQUEST UPIU 设置 fPurgeEnable = 0 来提前终止设备的清除操作，然后再发送命令。

下图是清除操作的状态机图：



State	bPurgeStatus
Idle	00h, 02h, 03h, 04h, 05h
Purge op in progress	01h

Transition	Description
1	Set fPurgeEnable and LU command queues not empty
2	Set fPurgeEnable and LU command queues empty
3	Clear fPurgeEnable
4	Automatic transition when the purge operation completes successfully or with failure

NOTE 1 On each transition the input event (triggering the state transition) and the output of the transition itself are mentioned.

UFS 设备在执行清除操作时，对那些垃圾数据，有以下几种处理方式：

1. 默认是把这些垃圾数据从闪存空间擦除掉；
2. 或者先用单个字符（比如全 A）覆盖写，然后再擦除；
3. 抑或先用单个字符（比如 A）覆盖写，然后用它的补码（比如 5）覆盖写，最后用随机字符覆盖写；
4. 最后还可以使用用户自定义的方式处理。

这些手段前面已经介绍过。

- **格式化设备（Wipe Device）**

主机通过发送 FORMAT UNIT 命令格式化所有的逻辑单元（RPMB LU 除外）。不过，对那些写保护的逻辑单元，FORMAT UNIT 命令会失败。

FORMAT UNIT 的命令对象是 Device well know LU，它格式化除 RPMB 之外所有无写保护的逻辑单元。

FORMAT UNIT 会切断逻辑块到物理空间的映射。但如果要让数据彻底从设备上清除，UFS 设备还需要执行 Purge 操作，这样数据才能彻底删除。

- **写保护**

前面都是千方百计的清除数据，但有时候 UFS 设备需要保护写的数据。

每个逻辑逻辑单元（除了 RPMB）有写保护属性。写保护包括永久写保护和上电写保护，前者的意思是说，一旦该逻辑单元写保护使能，将终生是写保护（不能改回去了）；而后者写保护只对某次上电有效，如果设备重上电或者重启，写保护将失效。

最后总结一下 UFS 数据安全机制：

1. 安全擦除（本章重点讲述）；
2. 写保护（本章讲述）；

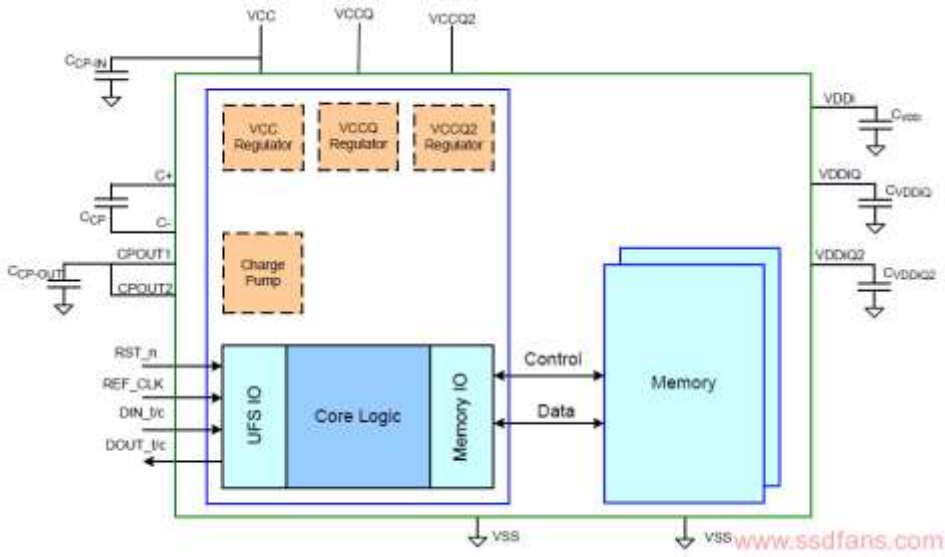
RPMB，提供数据认证和抗重放攻击机制保护一些重要数据，请参看 RPMB 章节。

10.5.10 蛋蛋读 UFS 之十：UFS 电源管理

Posted on 2018 年 5 月 31 日 by 蛋蛋

原创内容，转载请注明： [http://www.ssdfans.com] 谢谢！

UFS 是手机存储设备，因此对功耗要求很高。我们来看看 UFS 的电源管理。



三个供电电压，VCC,VCCQ 和 VCCQ2，分别给 UFS 设备模块供电。UFS 设备主要包括三部分：前端 UFS 接口（M-PHY），UFS 控制器和闪存介质（图中的 Memory 模块）。VCC 给闪存介质供电，VCCQ 一般给闪存输入输出接口和 UFS 控制器供电，VCCQ2 一般给 M-PHY 或其它一些低电压模块供电。

UFS2.1 中，三者电压值为：

	最小电压（V）	最大电压（V）
VCC	2.7	3.6
	1.70	1.95
VCCQ	1.1	1.3
VCCQ2	1.70	1.95

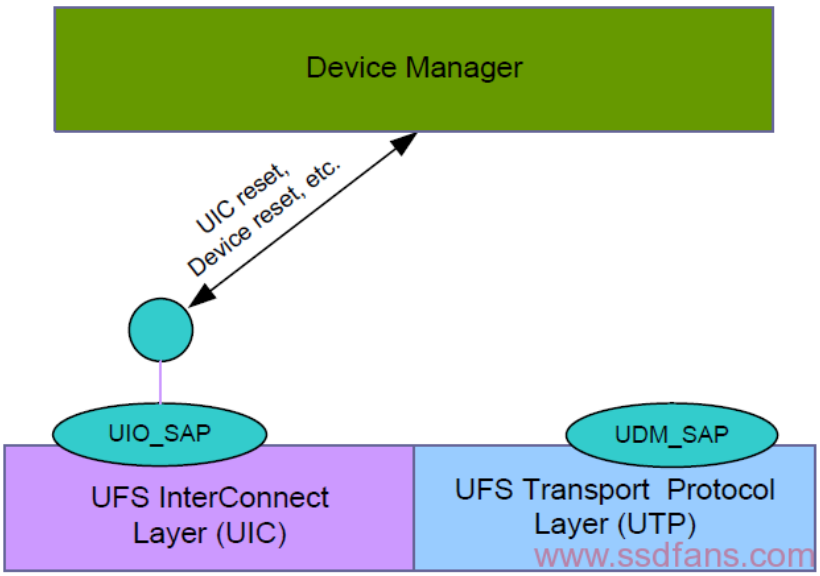
我们知道，UFS 协议采用 MIPI 的 M-PHY 作为物理层和 UniPro 作为其数据链路层。M-PHY 有高速模式（High Speed Mode, HS-MODE）和低速模式 （Low Speed Mode, LS-MODE）。其中，高速模式下，M-PHY 有两种状态：STALL 和 HS-BURST。

低速模式下，M-PHY 有三种状态：LINE-CFG，SLEEP 和 PWM-BURST。

当链路上没有数据传输时，M-PHY 会自动切换到 STALL 或者 SLEEP 状态下，这两种状态为省电状态。

除此之外，M-PHY 还有一种更加省电的状态，那就是 HIBERN8 （Hibernate，休眠状态），这种状态下，M-PHY 极为省电。UFS 主机和 UFS 设备不可能一直交互数据，总有闲下来的时候。当 UFS 主机没有读写 UFS 设备，它会让彼此链路进入休眠状态，即 HIBERN8。那 UFS 主机如何通知 M-PHY 切换到休眠状态呢？

前面提到，设备管理器可以略过传输层，直接管理与控制互联层：



主机设备管理器可以通过原语（Primitive）直接与 UFS 互联层（UIC，即 MIPI 的 UniPro 和 M-PHY）通信。除了上图中所示的 reset 原语，UFS 还包括让 UIC 进入和退出休眠的原语：DME_HIBERNATE_ENTER 和 DME_HIBERNATE_EXIT。

这是 UFS 主机和设备之间链路的省电模式，对 UFS 设备来说，链路只是整个 UFS 设备的一部分。一个 UFS 设备是否省电，除了看其链路，还需要考虑 UFS 控制器、存储介质等是否省电，即看整个 UFS 设备是否有好的电源管理。

UFS 定义了 4 种基本功耗模式：Active，Idle，Power Down 和 Sleep（简称 AIDS），外加 3 个过渡功耗模式：Pre-Active, Pre-Sleep 和 Pre-PowerDown，一共是 7 种功耗模式。非常 4+3！

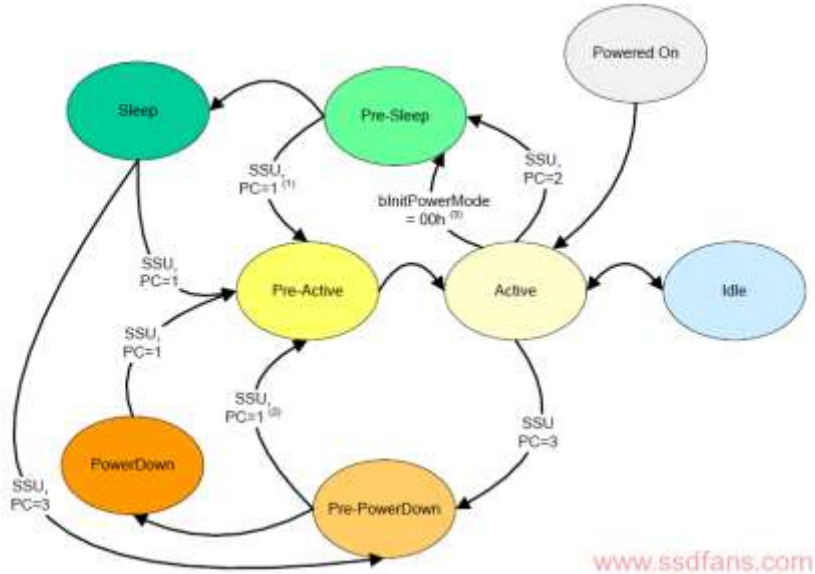
Active 模式：UFS 设备在执行命令或者做后台任务（Background Operation）时处于这种状态；

Idle 模式：UFS 设备空闲时，即既没有来自 UFS 主机的命令，自身也没有后台任务需要处理，设备就处于该状态；

Sleep 模式：闲得瞌睡了。睡眠模式下，VCC 电源可能被切断（取决 UFS 设备设计）。VCC 一般给闪存供电，即切断闪存供电。

Power Down 模式：掉电模式下，所有电源供电 VCC, VCCQ 和 VCCQ2 都可能被掐断（取决 UFS 设备设计），该模式是最省电的功耗模式了。

这些模式之间的转换如下图：



我们看到，触发模式之间转换的很多是 SSU，那么什么是 SSU? SSU 是 Start Stop Unit 的缩写，它是 UFS 协议中的一个基本命令，主机用它来切换 UFS 设备的功耗模式。

Table 11-1 — UFS SCSI Command Set

Command name	Opcode	Command Support
FORMAT UNIT	04h	M
INQUIRY	12h	M
MODE SELECT (10)	55h	M
MODE SENSE (10)	5Ah	M
PRE-FETCH (10)	34h	M
PRE-FETCH (16)	90h	O
READ (6)	08h	M
READ (10)	28h	M
READ (16)	88h	O
READ BUFFER	3Ch	O
READ CAPACITY (10)	25h	M
READ CAPACITY (16)	9Eh	M
REPORT LUNS	A0h	M
REQUEST SENSE	03h	M
SECURITY PROTOCOL IN	A2h	M
SECURITY PROTOCOL OUT	B5h	M
SEND DIAGNOSTIC	1Dh	M
START STOP UNIT	1Bh	M
SYNCHRONIZE CACHE (10)	35h	M
SYNCHRONIZE CACHE (16)	91h	O
TEST UNIT READY	00h	M
UNMAP	42H	M

具体命令可以参看 UFS spec。

注意，UFS 设备的这些功耗状态，和前面说的 M-PHY 接口的 STALL，SLEEP 或者 HIBERN8 状态是独立的，两者没有必然联系。比如，当前 M-PHY 处于 HIBERN8 状态，UFS 设备可以处于以上状态中的任何一种，比如 UFS 设备可以是处于 Active 状态，没有要求说你休眠了我也得跟着休眠。

一个优秀的员工，不是老板 push 一下，然后才往前走一步，而是能主动的去承担一些任务。一个好的 UFS 设备，不是等着主机发功耗切换命令来进入省电模式，而是自己能主动做一些事情来省电。

下面就是一个优秀 UFS 设备需要具备的素质。

比如，UFS 刚上电时，UFS 进入 Active 状态，一段时间如果没有来自主机的命令，自己内部也没有后台任务要处理，UFS 设备将进入 Idle 状态。Idle 意味着无事可做，这时候主机也没有发任何 SSU 命令要求 UFS 设备进入指定的状态（老板也没有叫你去做什么），好的 UFS 设备，这个时候就要想想怎么去省电。举例来说，如果当前 M-PHY 处于 HIBERN8 状态，说明主机目前不会访问 UFS 设备，因此，UFS 设备可以做一些节能设计：比如把当前 UFS 设备的软硬件上下文保存到闪存，然后切断所有电源以达到省电目的。待 M-PHY 接口退出 HIBERN8 状态，UFS 设备上电，然后把软硬件上下文加载运行。

老板没有叫你去干活，你主动的去把活干了，这样的员工哪个老板不喜欢？

省电和用户体验（命令响应快慢）其实是个矛盾。因为如果 UFS 设备休眠了，它就不能及时的响应主机的命令，因为它需要先退出休眠（比如需要把休眠之前保存的上下文重新加载，这往往比较花时间），然后再响应主机命令。睡觉是个技术活，在追求最大节能的同时，还要兼顾用户体验。

10.6 PCIe 协议底层杂谈

10.6.1 PCIe 基础概念

10.6.1.1 好大一棵树 – PCIE TREE

树?

树这个概念我们应该是很熟悉的了，生活周围到处都是。树的主要四部分是根、干、枝、叶。而树根一般在地下。我们常看见的也就是以上的树干、树枝、树叶。

PCIe 树?

学习 PCI/PCIe，最常听到的一个名词也是“树”，PCIe tree。所谓的 PCIe tree 就是 PCIe specification 定义的了 PCIe 系统的拓扑结构。定义了这个拓扑，我们所有的设备内部的 PCIe 设备都可以抽象对应到一颗“树”上来。

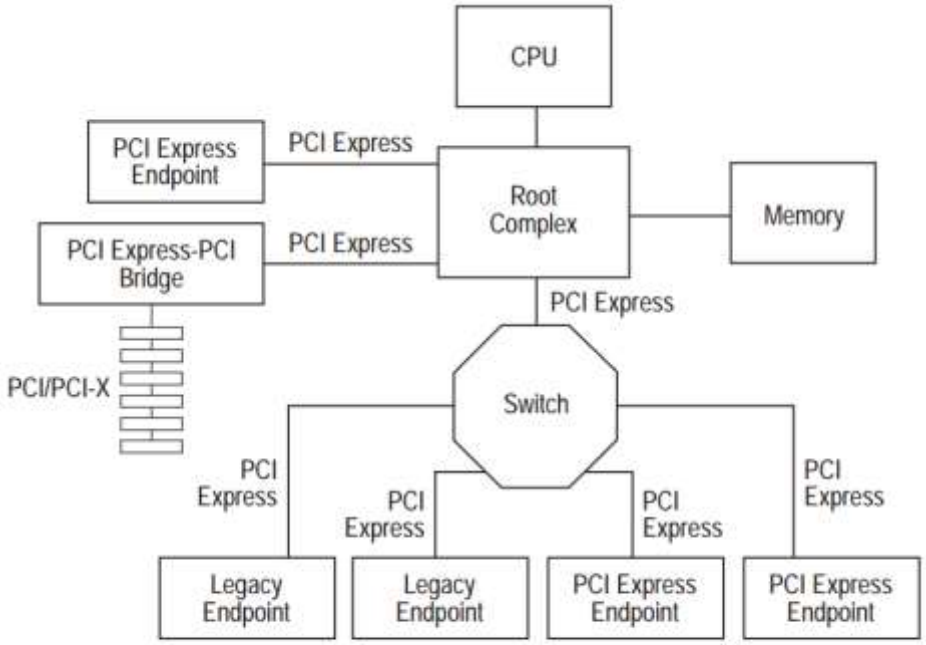


Figure 1-2: Example Topology

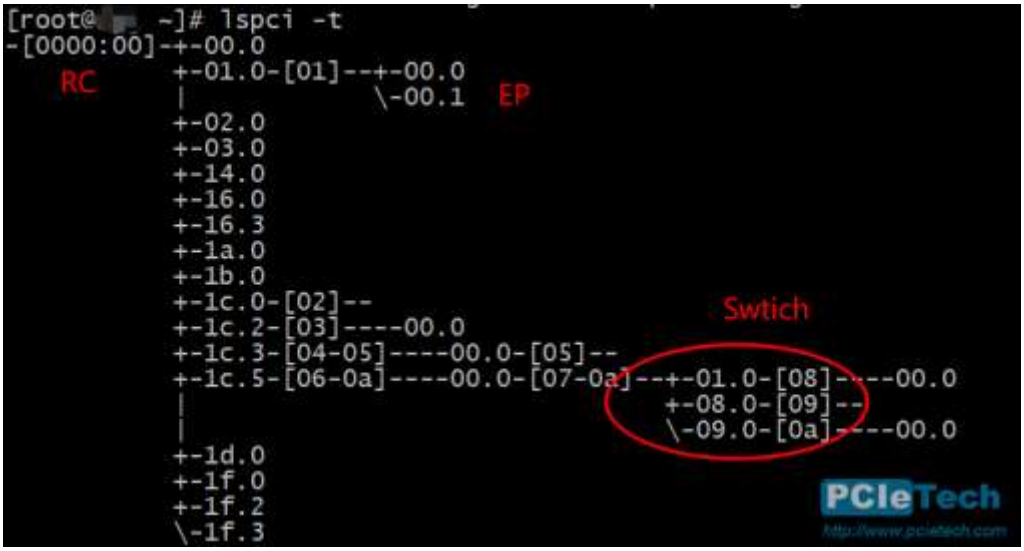
注意：spec 里面这个拓扑图的 RC 和 CPU 是两个独立的部分，事实上，我们常用的处理器目前大部分都是把 CPU 和 RC 集成到了一起。

这是一颗倒着生长的树！树的根即是 RC(Root Complex)，我们叫做根节点或者根联合体。RC 内部的结构和处理机制如同我们看不到的树在地面以下的错综复杂的根。管他呢，这部分我们并不十分 care。记住，这里是根，是一切的开始就好。

从根节点开始向下生长出树干和树枝（PCIe 链路），某些树枝还扩展生长出旁路的分支，这些可以扩展生长的地方，对应拓扑中的 Switch。最末端的叶/果实，对应拓扑中的 EP（End Point）。

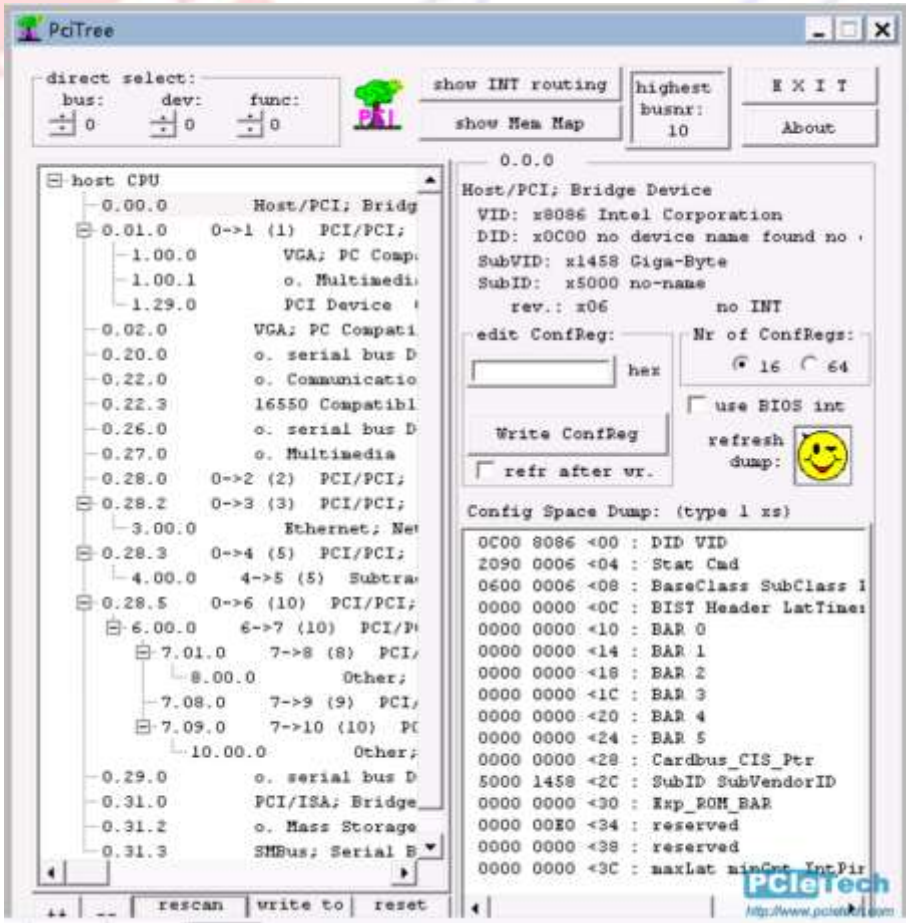
怎么看 PCIe 树?

Linux 提供了一个 lspci 的命令，用于显示当前主机的所有 PCI 总线信息，以及所有已连接的 PCI 设备信息。加上-t 参数（t：tree）即给我们显示了系统中的 PCIe tree 拓扑，只不过这颗显示的树是横向的，转了 90 度。



Lspci 其实是一个开源工具 pciutils（http://mj.ucw.cz/sw/pciutils/）的命令，对应命令还有 setpci（用于读写 PCIe 配置寄存器）。如果你的系统中找不到这个命令，可以执行 yum install pciutils 安装即可。工具官网或者内核官网可以下载到源码，阅读分析这些源码对于开发 PCIe 相关内容有极大的帮助。

而 Windows 下，有一个很久没有更新的工具，叫 pcitree （http://www.pcitree.de/）。有兴趣的可以下载来玩玩。注意:只能运行于 win7 以及之前的 windows 版本，在我的 win10 上无法运行。



树乱了？

做底层驱动开发特别是 PCIe 驱动相关的开发，经常看到的一个错误就是“树乱了”。树乱了从表面上看，即是 lspci -t 的内容发生了错乱，跟正常情况下的树不一样了。本质上，“树乱了”由于系统中的某些 PCIe 设备出错或者异常导致的。lspci -t 时，系统会重新访问这颗树上的所有设备，由于部分 PCIe 设备的异常，访问失败（返回异常值或者全 F），从而影响 pciutils 构建出一颗完整的树，出现部分树枝断裂(链路 link down)，树叶错位（EP Bar 空间异常）等现象。

10.6.1.2 PCIE 设备的资源

前面讲了树，理清了 PCIe 架构的大概框架；讲了链路，知道了框架中的路径。而对于树上的树叶/果实，即最终的 PCIe 设备，它们在系统中是是怎样存在的？今天开始逐步对 PCIe 设备的内部细节和协议的分层实现细节展开讲解。

在 PCIe 系统中，对于每一个 PCIe 设备，都具有三种类型的资源。具有了资源，PCIe 设备才具有了被访问、被使用的基本能力。这三种资源分别是：

- **ID Resource**

ID 资源：即设备在系统中的定位和身份证。用总线号（Bus）、设备号(Device)和功能号(Function)三个变量定义。

- **Memory Resource**

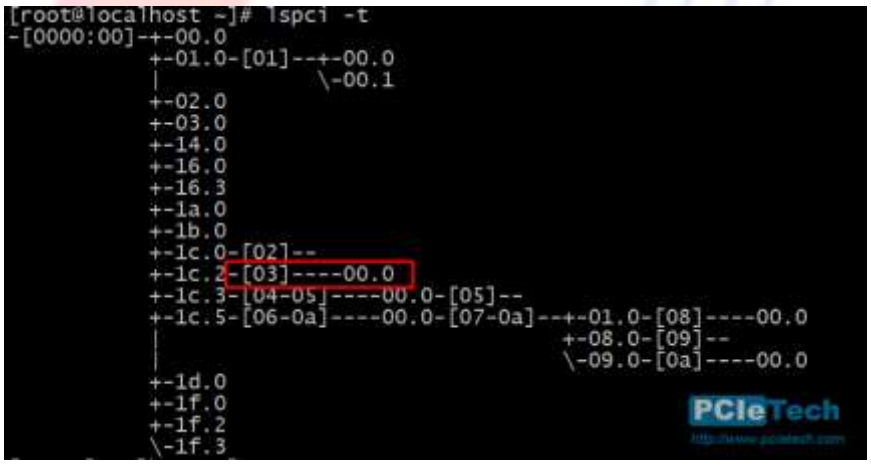
内存资源：即设备具备哪些可以提供给外部或内部使用的内存。

- **IO Resource**

IO 资源：即设备具有哪些可以使用的 IO 空间。

在一个 PCIe 系统中的每一个 PCIe 设备，如上的三种资源分配都是唯一的。不会有两个设备具备相同的资源。

我们看一个实例，如下这个 PCIe tree。红框标注的这个设备是张网卡。



这个设备的 ID 资源是：总线号 3，设备号 0，功能号 0。

我们可以使用 lspci 命令查看它的 memory 和 IO 资源：

```
[root@localhost ~]# lspci -s 3:0.0 -v
03:00.0 Ethernet controller: Realtek Semiconductor Co., Ltd. RTL8111/8168/8411
        Subsystem: Gigabyte Technology Co., Ltd Onboard Ethernet
        Flags: bus master, fast devsel, latency 0, IRQ 32
        I/O ports at d000 [size=256]
        Memory at f7a00000 (64-bit, non-prefetchable) [size=4K]
        Memory at f0000000 (64-bit, prefetchable) [size=16K]
        Capabilities: [40] Power Management version 3
        Capabilities: [50] MSI: Enable+ Count=1/1 Maskable- 64bit+
        Capabilities: [70] Express Endpoint, MSI 01
        Capabilities: [b0] MSI-X: Enable- Count=4 Masked-
        Capabilities: [d0] Vital Product Data
        Capabilities: [100] Advanced Error Reporting
        Capabilities: [140] Virtual Channel
        Capabilities: [160] Device Serial Number 01-00-00-00-68-4
        Kernel driver in use: r8169
        Kernel modules: r8169
```

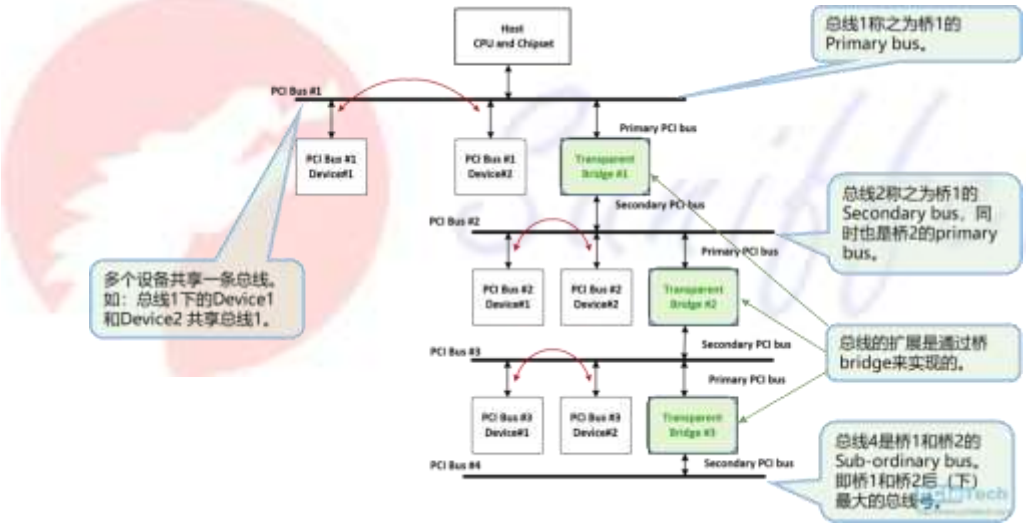
它的 IO 资源位于 d00 处，大小 256。

它的 Memroy 资源有两个，一个是位于 f7a00000 大小为 4K，一个是位于 f0000000 大小 16K。

后面会依次详细讲解每种资源的申请、分配和使用。

10.6.1.3 从 PCI 角度认识 PCIE

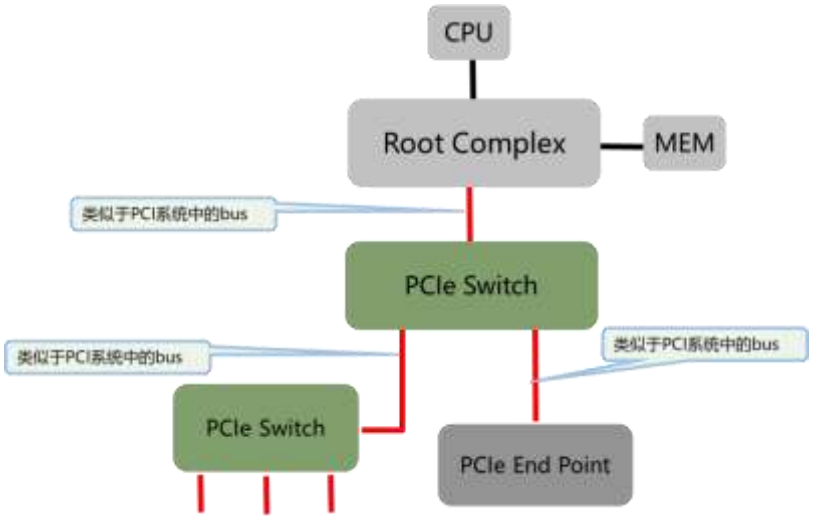
从软件角度或者说是从逻辑角度看，PCI 跟 PCIe 有着天然的继承性。让我们首先来看看 PCI 的逻辑关系。



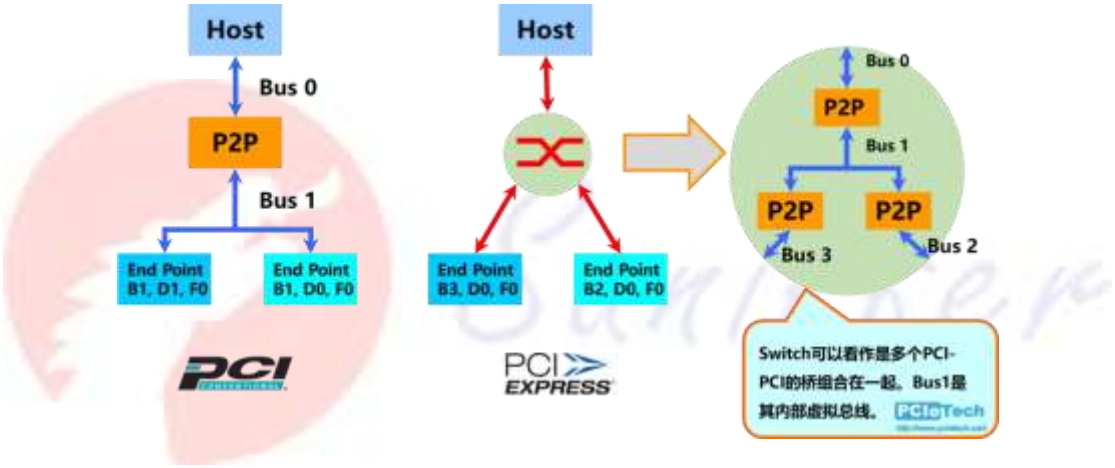
PCI 系统中，总线的扩展是依靠桥(Bridge)来扩展的。注意：这里是透明桥（Transparent Bridge），所谓透明是指这个桥对于经过它的报文或者数据，不做任何的处理和表更，直接往下游或者上游传递。既然有透明桥，那么一定就有相对应的非透明桥了么？是的，没错，确实如此，不过这里我们先卖个关子，后续慢慢再表。PCI 的系统中总线的命名关系图中写的比较清楚了，对于每一个桥，都有 Primary bus 和 Secondary bus 以及 Subordinate bus。

内核代码中 pci_bus 结构体有对应的定义，各位可以翻阅代码查看。

到了 PCIe 系统中，情况变成什么样了呢？让我们看一看，如下：



这里有个非常重要的部件取代了 PCI 系统中桥的功能，并且更为强大，它就是 PCIe Switch。可以说，PCIe Switch 是 PCIe 系统中的重中之重，掌握了 PCIe Switch，基本也就掌握了 PCIe 系统。为了便于大家理解 PCIe Switch，我们把它分解一下，就很容易看明白了：



从逻辑上看，Swith 可以看作是多个 PCI-PCI 桥的组合。内部有虚拟 PCI 总线。当然实际 PCIe Switch 内部构造远比这复杂，这里仅仅是从逻辑关系上说明 switch 和 PCI 的关系。

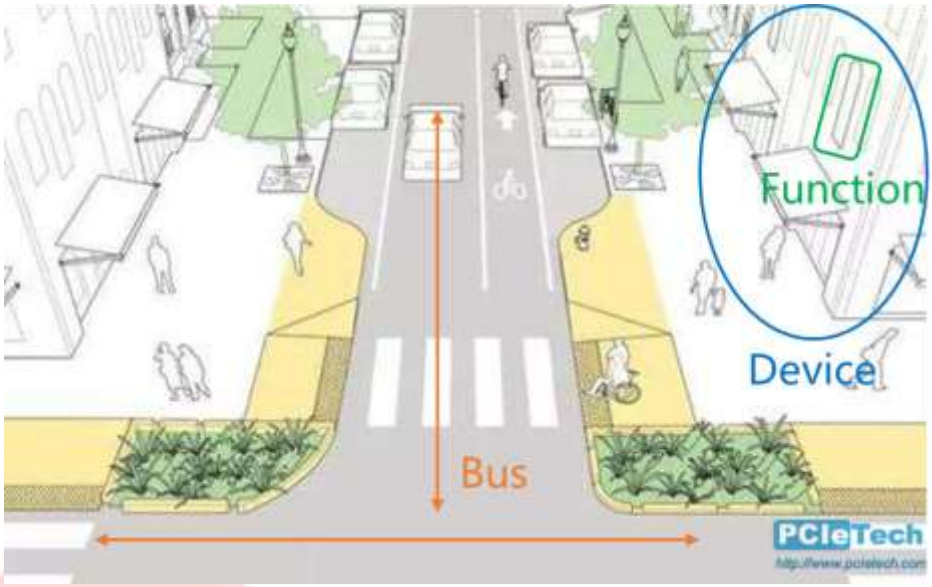
注意：图中的 P2P 指的是 PCI-PCI bridge，而不是 Peer-to-Peer。

10.6.1.4 PCIE 设备的身份证

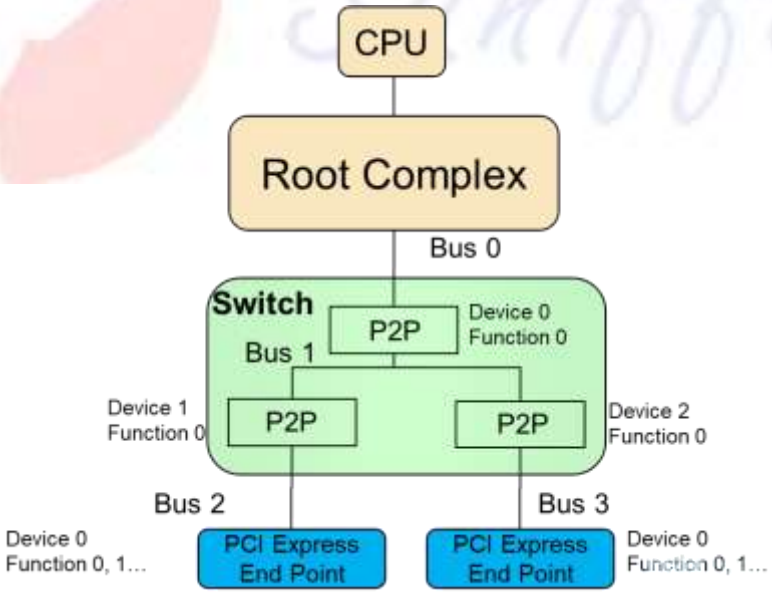
ID 资源是一个 PCIe 设备的最基本资源，每一个 ID 号都是独一无二的。分析 ID 资源前，我们需要先了解 Function（功能）的概念。对于一个 PCIe 设备，如果它只具有一个功能，我们称之为 Single Function Device；如果它有多个功能，则称之为 Multi Function Device。举个例子，在我的电脑上，有这么一个 AMD 的显卡，位于 01 号总线，00 号设备，它具有两个功能，功能 0 是显示用的 Radeon HD 6750，功能 1 是音频用的 Radeon HD 5700。

ID 是由总线号、设备号和功能号共同组成的。因为有多功能设备，所以 ID 实际是一个功能的名字和地址。在 Linux 系统中，表示 ID 的方式是[Bus:Device.Function]，如上的显示用的 6750 的 ID 号是 01:00.0。音频用的 5700 的 ID 号是 01:00.1。

说到地址，举个很形象的例子。总线号类似于街道的名字，设备号类似于这条街道上的某栋房子，而功能号类似于某一栋房子里的某个房间。



PCIeEP 的设备号（Device Number）都是 0，PCIe Switch 则不大一样。在 PCIe Switch 内部，实际是由多个 PCIe-To-PCIe 的桥组成，对于每一个桥的设备号可能是 0、1、2。。。如下图：



当然，这些桥的 Function 可能也有多个，上图仅示意了 Function0。

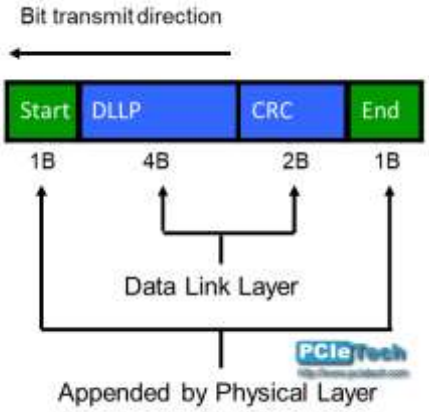
关于 ID 号，PCIe Spec 定义了如下规则：

- 总线号由 8bit 表示，因此，最大总线号为 256。
- 设备号由 5bit 表示，因此，最大设备号为 32。
- 功能号由 3bit 表示，因此，最大功能号为 8。

10.6.2 PCIe 数据链路层 DLLP 协议

10.6.2.1 DATA LINK LAYER PACKET (DLLP)简介

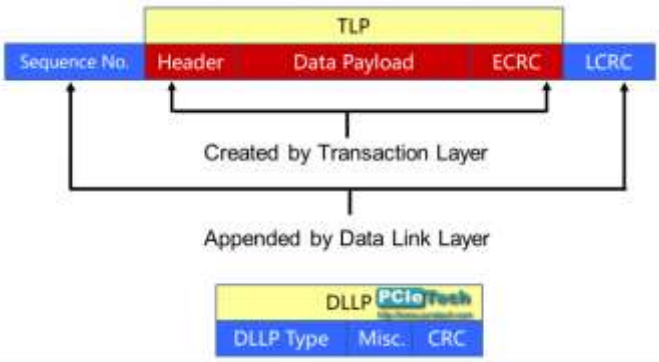
数据链路层的报文（DLLP）是由发送端的数据链路层生产，并在接收端被接收端的数据链路层接收，生活的范畴仅在一条链路上。DLLP 包括用于 Ack/Nak 机制、电源管理、流控（Flow Control）以及一些设备厂商自定义的功能。



如上图所示，DLLP 在物理层被添加了“Start”和“End”。数据链路层（DLL）生产 DLLP，并且在后面添加了 CRC 以校验这个 DLLP 是否正确。

前面一篇文章讲过，DLL 会给从 TL 层下来的每个 TLP 报文都添加两个东西：一是在 TLP 前面增加序列号（Sequence number）,二是在 TLP 尾部增加 LCRC，对端会检测这个 LCRC，以确保传输的 TLP 内容不会损坏。

每一个通过 DLL 层的 TLP 都会被加上序列号，序列号从 0 开始，最大 4095，超过 4095 后回绕到 0。例如：第一个 TLP 序列号为 0，第二个 TLP 序列号为 1，。。。第 4096 个 TLP 序列号为 4095，第 4097 个 TLP 序列号回绕为 0。



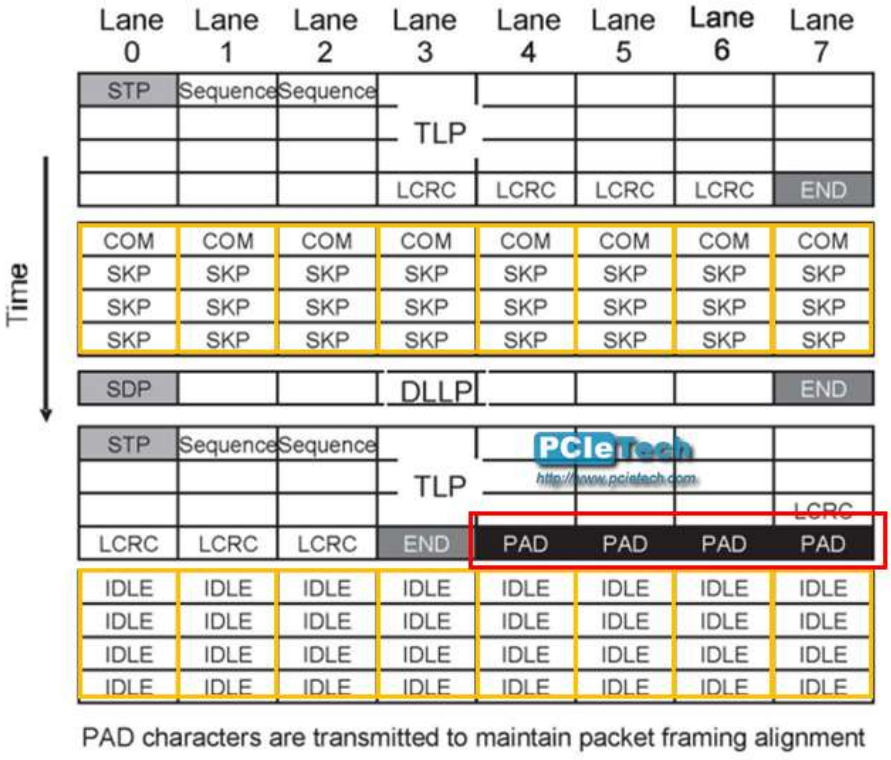
注意，这里有个地方很容易混淆： TLP 在 DLL 层被添加上 sequence number 和 LCRC 后，并不是 DLLP！上图的两种报文格式，可以简单理解为添加了 sequence number 和 LCRC 的 TLP 是数据 transfer_data，而 DLLP 报文是控制命令 transfer_cmd。transfer_cmd 确保了 transfer_data 传输的正确与否。

transfer_data 的 LCRC 错了，有相关 transfer_cmd 命令（Nak DLLP）回给发送端要求重传。而 transfer_cmd 的 CRC 错了，会怎样呢？答案是接收端会丢弃这条命令。依赖超时等机制继续处理事务。

10.6.2.2 ORDERED SETS 简介

物理层的控制字符除了昨天讲的用于 TLP/DLLP 报文的 STP/SDP/END/EDB 之外，我们来看看其他几个控制字符的用途：

PAD 字符：前面我们讲过字节流经过字节拆分后分布到不同的 lane 上发送。Spec 要求数据流对齐，PAD 字符就是在不对齐的情况下填充用。如下图 X8 的链路，红色框线中填充了 4 个 PAD 字符。



PAD characters are transmitted to maintain packet framing alignment

COM 字符：COM 字符用作有序集的首字符。有序集下面叙述。

SKP、IDL、FTS、EIE 字符都是是某个特殊功能有序集的一部分。

由 COM 开头组成的一系列字符，组成了有序集（Ordered Sets），用于链路管理等特殊功能。有序集又叫做物理层报文（PLP：Physical Layer Packet）。注意：不同于数据流的字节拆分到各个 lane 上，有序集是需要在每条 lane 上同时发送的。

Spec 定义了如下有序集：

- TS1&TS2（Training Sequence ）训练序列 1 和 2：用于链路初始化、链路训练，协商链路的速率、宽度等。
- SKP 有序集：用于发送时钟和接收时钟的补偿。
- EIOS 有序集（Electrical Idle Ordered Set）：用于通知链路进入低功耗模式。
- FTS 有序集：（Fast Training Sequence）用于通知链路从低功耗模式退回了正常模式。

10.6.2.3 SYMBOL 简介

物理层接收部分的处理即是发送反向，因此，不再展开讨论。今天我们讨论下 Symbol。

我们知道，从数据链路层下来的数据流（TLP或者DLLP）需要经过编码，一个字节8bit编码后变成10bit，这个10bit我们称之为Symbol，中文称之为符号，也有很多文章叫字符，为了统一，本文也称字符。因为是数据编码，我们称为数据字符。表示为 Dxx.y。

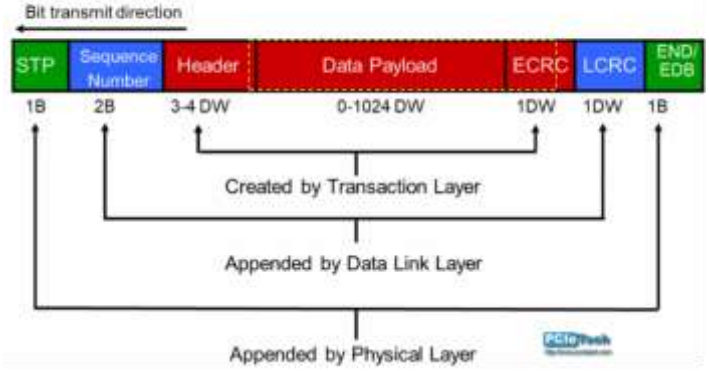
物理层本身还有一些控制字符，表示为 Kxx.y。其中的 xx 表示低五位，y 表示高三位。如数据 0x23(001 00011)的数据字符表示为 D3.1。

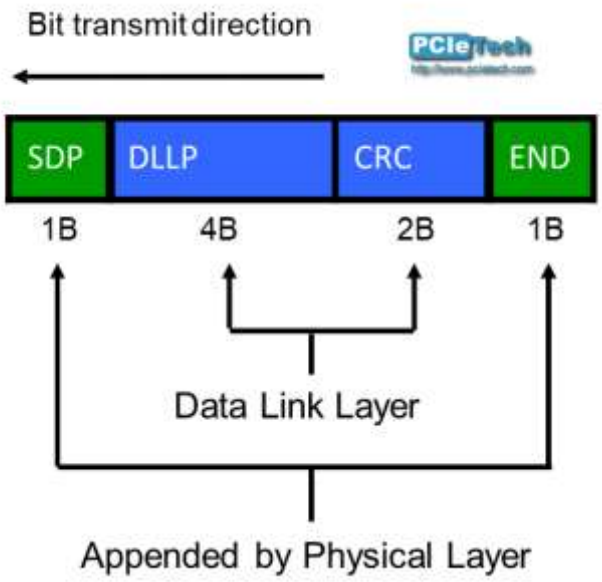
物理层常用的控制字符如下表：

Table 4-1: Special Symbols

Encoding	Symbol	Name	Description
K28.5	COM	Comma	Used for Lane and Link initialization and management
K27.7	STP	Start TLP	Marks the start of a Transaction Layer Packet
K28.2	SDP	Start DLLP	Marks the start of a Data Link Layer Packet
K29.7	END	End	Marks the end of a Transaction Layer Packet or a Data Link Layer Packet
K30.7	EDB	EnD Bad	Marks the end of a nullified TLP
K23.7	PAD	Pad	Used in Framing and Link Width and Lane ordering negotiations
K28.0	SKP	Skip	Used for compensating for different bit rates for two communicating Ports
K28.1	FTS	Fast Training Sequence	Used within an Ordered Set to exit from L0s to L0
K28.3	IDL	Idle	Used in the Electrical Idle Ordered Set (EIOS)
K28.4			Reserved
K28.6			Reserved
K28.7	EIE	Electrical Idle Exit	Reserved in 2.5 GT/s Used in the Electrical Idle Exit Ordered Set (EIEOS) and sent prior to sending FTS at speeds other than 2.5 GT/s

STP（Start of TLP）控制字符表示后续跟着的是一个 TLP 报文，SDP（Start of DLLP）表示后续跟着的是一个 DLLP 报文。结束符 END/EDB 前面我们已经讲过了。其他的 COM、PAD、SKP、FTS、IDL 控制字符都是物理层本身用于链路训练用。

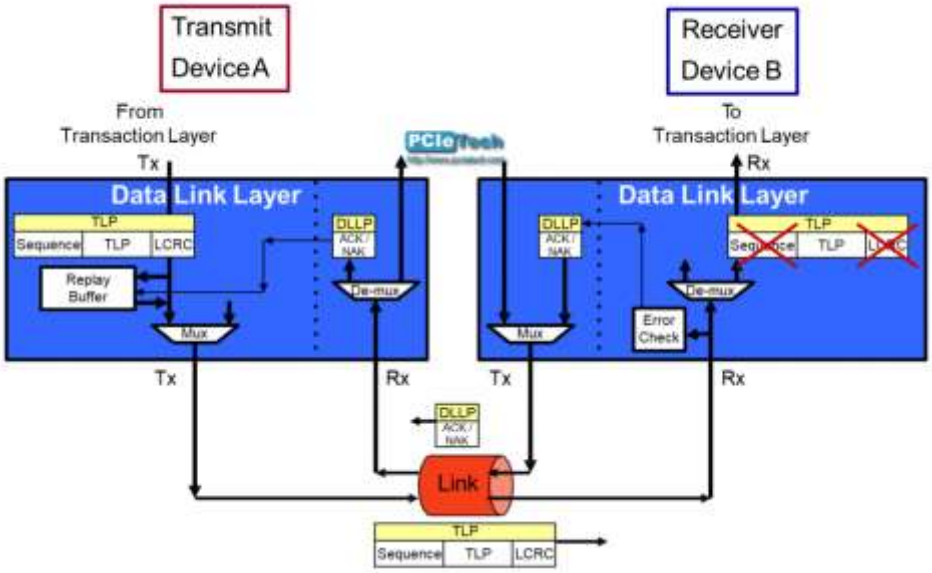




注意：*上述以及后面物理层相关的，我们讨论的都是以 Gen1、Gen2 为准，Gen3 及其以后更改了编码方式和部分控制字符，略有不同，但并不影响我们对协议的整体理解。因为协议规定，链路训练首先要求双方都从 Gen1 开始。*

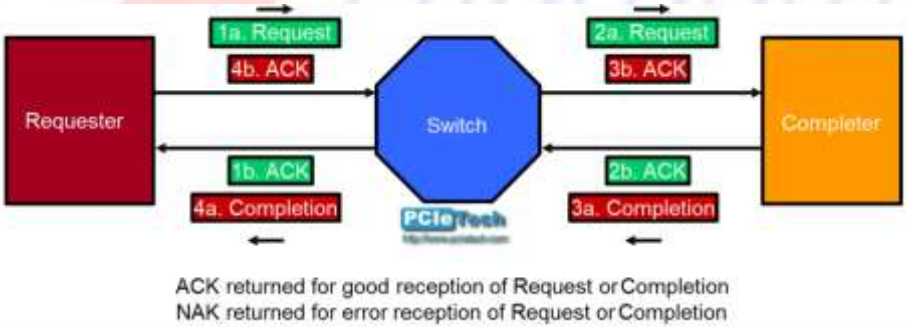
10.6.2.4 ACK & NAK 简介

TLP 报文在链路上传输的可靠性，是由 ACK/NAK 机制来保证的。事实上很多其他协议也有类似的机制。ACK 就是 acknowledge，ACK DLLP 表示 TLP 接收完成，NAK 就是 Negative acknowledge，意思就是拒绝接收这个 TLP。



如上图所示，每一个从 A 发到 B 的 TLP 报文，发出后，都会先备份缓存在 A 的重发缓冲区（Replay Buffer）里面。B 的接收器会检测 TLP 是否正确。如果是正确的，B 会返回 ACK 确认接收成功，此时 A 回删除重发缓冲区的对应备份。如果 TLP 由错误，B 会返回 NAK。收到 NAK 的 A 设备会从重发缓冲区重新发送对应的 TLP。（注意：并不是每一个 TLP 都必须返回 ACK/NAK，协议规定可以几个 TLP 之后用一个 ACK/NAK 来返回。）

ACK/NAK 是在两个设备间发送的，当然，通过 Switch 的情况下，无论是 Requester 还是 Completer，跟 Switch 交互的报文，同样也是分别有对应的 ACK/NAK 的。如下图。

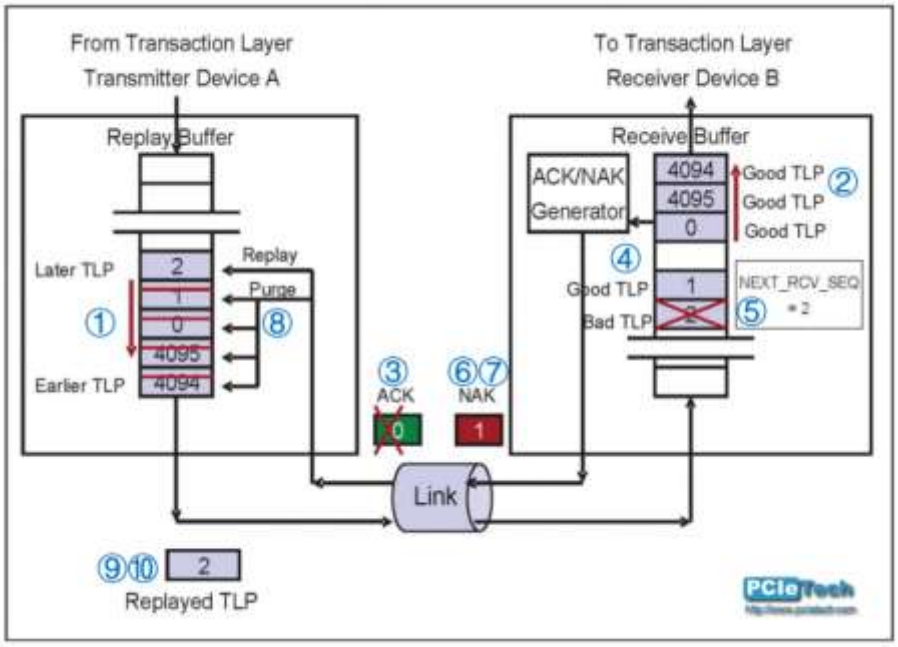


假设链路一切 OK，Requester 首先发送 request 报文，Switch 接收并回复给 Requester ACK。Switch 继续转发这个 Request 到 Completer，Completer 接收并回复给 Switch ACK。Completion 完成报文返回的路径上，同样如此，不再累述。

10.6.2.5 ACK & NAK SAMPLE 举例

举一个实际的 ACK NAK 的例子，这个例子来源于《PCI EXPRESS 系统体系结构标准教材》一书，英文书名叫《PCI Express System Architecture》。话说这本书当年是我学习 PCIe 的葵花宝典，目前市面已经绝迹了。顺便说一下，王齐的那本《PCI Express 体系结构导读》写的很好，讨论了很多 PCIe 的内部实现细节。

Figure 5-16: Lost ACK DLLP Handling



1、 发送端 A 发送序列号为 4094、4095、0、1、2 的 TLP 报文。（计数到 4095 翻转到 0 开始计数）

2、 接收端 B 接收序列号为 4094、4095、0 的 TLP 报文，NEXT_RCV_SEQ 计数增加，下一个待收取的 TLP 报文序列号是 1。当 ACK TIMER 超时时，发送 ACK 报文给 1、 发送端 A，对应序列号 0，我们称之为 ACK 0。代表之前的 4094、4095、0 这三个报文我已经接收成功了。

3、 不幸的是，这个回复给发送端的 ACK 0 报文，在链路上由于某种原因丢失了。发送端并没有收到这个 ACK，序列号为 4094、4095、0 的 TLP 报文仍然保存在发送端 A 的 Replay buffer 里面。

4、 随后，发送端 A 发送的序列号为 1、2 的报文 TLP 1、TLP 2 到达了接收端 B。TLP 1 校验完全正确，NEXT_RCV_SEQ 计数增加，下一个待收取的 TLP 报文序列号是 2。

5、 雪上加霜的是 TLP 2 由于某种原因损坏，校验不通过，接收端 B 不接收，NEXT_RCV_SEQ 计数不变，仍为 2。

6、 由于 TLP 2 不成功，接收端 B 需要回复一个 NAK 给发送端 A，其中的序列号为 1。

7、 NAK 1 报文顺利的到达了发送端 A。

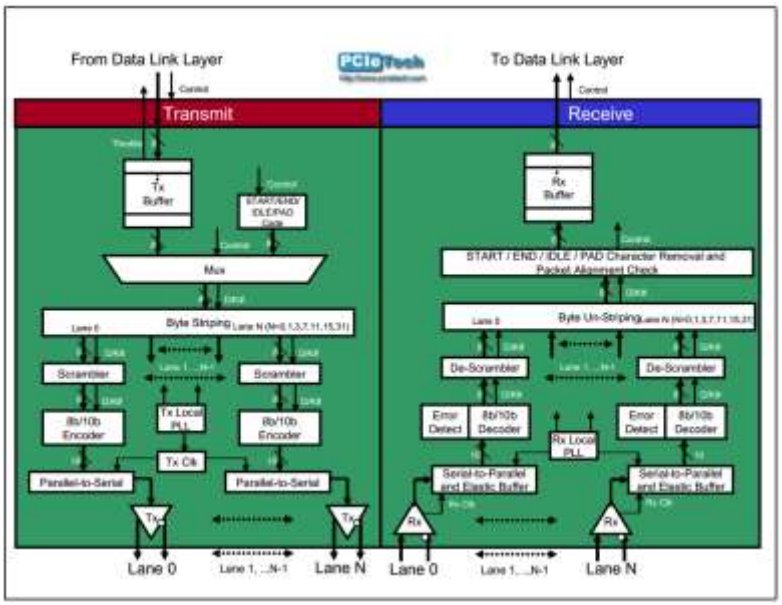
8、 发送端 A 收到 NAK 1，表示前面发送的 TLP 2 有错，但是 TLP 4094、TLP 4095、TLP 0、TLP 1 都已经被接收端正确接收了。因此，清除 Replay buffer 里面的序列号为 4094、4095、0、1 的报文。

9、 发送端尝试重新发送 TLP 2。

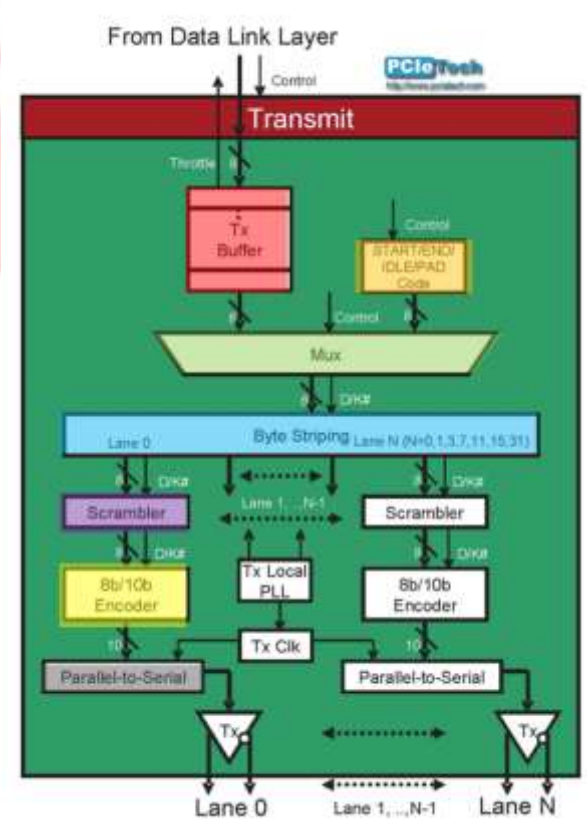
10、 TLP 2 这次成功的被接收端接收，NEXT_RCV_SEQ 计数增加，下一个待收取的 TLP 报文序列号是 3。

10.6.2.6 ACK/NAK 协议的发送漫谈

如上节所提到，物理层分为逻辑处理部分和电气部分。整个逻辑部分的处理框图如下：



简化一下，我们先来看看逻辑发送的部分：从数据链路层下来的数据报文 TLP 和链路层报文 DLLP 被放在发送缓冲区，被添加上控制字符（如 start 和 end 等），经过多路复用器分发到各条 lane 上。然后被扰频器加扰通过 8b/10b 编码或者 128/130 编码后，并转换发送到物理链路上。



发送缓冲区（Tx Buffer）：没啥好说的，缓存数据链路层下来的数据。并在缓冲区溢出时，反馈给数据链路层流控。

控制字符（Control Code）：区分发送的数据还是控制，生成不同的控制字符。

多路复用器（Mux）：根据发送的内容不同选通不同的控制字符添加。

字节拆分（Byte Stripping）：按照一定的规则，把数据流拆分成单个字节，并分发到不同的通道（lane）上，如 2、4、8、16、32 lane。单 lane 的情况下不需要拆分。

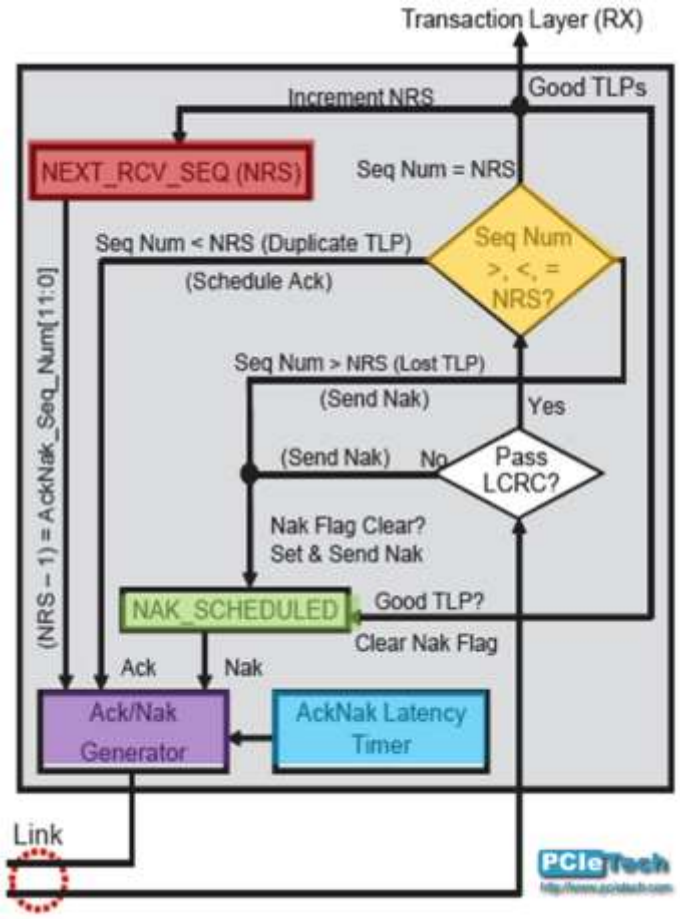
扰频（Scrambler）：也叫扰码，主要目的时防止高速链路上的 EMI 噪音。各位有兴趣可以参考相关的文档。注意：控制字符是不经过扰频的。

编码（8b/10b 或者 128b/130）：按照一定的规则按字节编码。（前面文章已经讲过一些）

并转串（Parallel to Serial）：没啥好讲的。：）

注:Gen3 的发送逻辑增加了一些内容，但是整体上不影响对于发送逻辑的理解。

10.6.2.7 ACK/NAK 协议的接收漫谈



备注：图中没有画出接收端的 receive buffer。

A. NEXT_RCV_SEQ Count

12 位的 NEXT_RCV_SEQ 计数器保存期望接收的下一个 TLP 的序列号。每收到一个正确的 TLP 报文，此计数器+1。到达 4095 后翻转为 0。如果 TLP 报文错误（CRC 错误、无效的 TLP 或者是序列号错误），计数器不变。

B. Sequence Number Checker

如果 LCRC 检测通过，接下来会检测接收到的 TLP 报文的序列号是否和上述 NEXT_RCV_SEQ count 值相等。如果匹配，则接收这个报文，NEXT_RCV_SEQ 计数器+1。

注意：这个时候，并不会直接返回 ACK 报文给发送端，而是等待 ACKNAK_LATENCY_TIMER 超时后再发送，此时发送 ACK 带有最后一个收到的完好的 TLP 的序列号，表示此序列号之前的 TLP 报文都是正确接收完成的。还记得前面讲的不是每一个 TLP 报文都有一个 ACK 对应么?这就是原因。

C. NAK_SCHEDULED Flag

当准备计划返回一个 NAK 给发送端时，NAK_SCHEDULED flag 标志被置位。当接收到第一个被要求重发的 TLP 时，此标志位清零。在 NAK_SCHEDULED flag 置位的时候，如果继续收到错误的 TLP，是否需要继续再发送一个 NAK？Spec 并没有讲。只有 if，没有 else。如下：

- If the NAK_SCHEDULED flag is clear,
  - schedule a Nak DLLP for transmission immediately
  - set the NAK_SCHEDULED flag

实际上，当 NAK_SCHEDULED flag 置位的时候，NEXT_RCV_SEQ Count 并没有增加，所以后续到来的 TLP 一定是序列号不匹配的。在当前 NAK 掉的 TLP 报文没有被发送端重新发送且在接收端正确接收之前，没有必要做其他任何事情，等着就好。

D. ACKNAK_LATENCY_TIMER

ACKNAK_LATENCY_TIMER 定时器溢出时，接收端会向发送端返回一个 ACK 或者 NAK。这个时间大约是发送器 REPLAY_TIMER 值得 1/3,以确保在发送端重发 REPLAY_TIMER 超时之前，发送端能够及时接收到接收端得 ACK/NAK 响应。

ACKNAK_LATENCY_TIMER 的值和 REPLAY_TIMER 类似，也跟链路速率、宽度即 MPS 有关，Spec 定义了一套计算公式，Gen3 情况下的值如下：

Table 3-9: Ack Transmission Latency Limit and AckFactor for 8.0 GT/s Mode Operation by Link Width and Max Payload (Symbol Times)

Max_Payload_Size (bytes)	Link Operating Width x1	x2	x4	x8	x12	x16	x32
128	333 AF = 1.4	224 AF = 1.4	169 AF = 1.4	163 AF = 2.5	154 AF = 3.0	144 AF = 3.0	129 AF = 3.0
256	512 AF = 1.4	313 AF = 1.4	214 AF = 1.4	203 AF = 2.5	186 AF = 3.0	168 AF = 3.0	141 AF = 3.0
512	655 AF = 1.0	385 AF = 1.0	250 AF = 1.0	182 AF = 1.0	205 AF = 2.0	182 AF = 2.0	148 AF = 2.0
1024	1167 AF = 1.0	641 AF = 1.0	378 AF = 1.0	246 AF = 1.0	290 AF = 2.0	246 AF = 2.0	180 AF = 2.0
2048	2191 AF = 1.0	1153 AF = 1.0	634 AF = 1.0	374 AF = 1.0	461 AF = 2.0	374 AF = 2.0	244 AF = 2.0
4096	4239 AF = 1.0	2177 AF = 1.0	1146 AF = 1.0	630 AF = 1.0	802 AF = 2.0	630 AF = 2.0	372 AF = 2.0

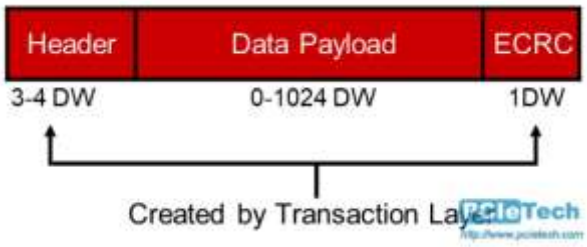
E. ACK/NAK DLLP Generator

即生成返回给发送端的 ACK/NAK DLLP 报文。报文包含 NEXT_RCV_SEQ count 计数器-1 的值，即最后一个校验正确的 TLP 报文的序列号。

10.6.3 PCIE 事务层 TLP 协议

10.6.3.1 TLP FORMAT 简介

今天我们来看一看 Transaction Layer 的 TLP 报文的具体组成。回顾一下前面我们讲到的 Transaction Layer 的报文。分为三个部分：Header、Data 和 ECRC。



Header 包含了这个 TLP 的类型、格式、路由地址、数据长度等重要信息，是 TLP 报文的核心，通常长度为 3DW 或 4DW。Data 则是具体的数据区域，4K 长度，如果数据的长度超过 4K，则会有多个包含数据的 TLP 报文。ECRC 是对 Header 和 Data 区域的校验和，在接收端这个校验和会重新计算一次并和发送的 ECRC 做比对，以确认这个报文的内容是否合法。

Header 的第一个字节定义了这个报文的格式：

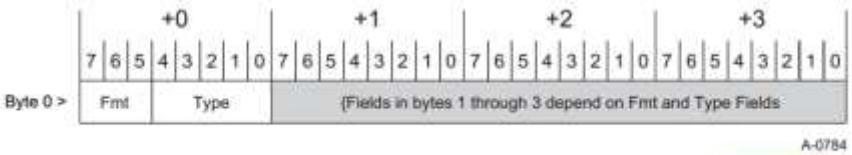


Figure 2-4: Fields Present in All TLPs

- Fmt 定义了这个 TLP Header 是 3DW 长度的，还是 4DW 长度的。这个 TLP 是带 Data 的，还是不带 Data 的，比如读请求 TLP 就是不带数据的。
- Type 定义了这个报文是 Memory W/R 还是 Config W/R，或者是 Message、Completion。

当 Fmt 和 Type 确认之后，不同类型的 TLP 后面的字段各不相同，具体可以参考 Spec。

以下是一个 64 位地址的 memory 访问的 Header，它是 4DW 长度，可以看到 Header 里面 Byte4 和 Byte5 明确的表明这个报文的 Requester 是谁，byte8 开始，表示访问的地址是多少：

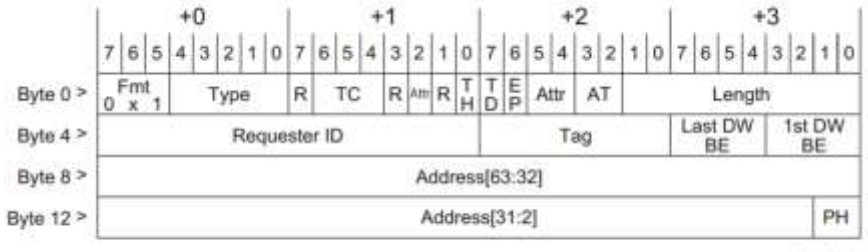


Figure 2-15: Request Header Format for 64-bit Addressing of Memory

10.6.3.2 TRANSACTION 类型

Transaction（事务）这个词在 wikipedia（https://en.wikipedia.org/wiki/Transaction_processing）里是这样描述的：

Transaction processing is information processing in computer science that is divided into individual, indivisible operations called transactions. Each transaction must succeed or fail as a complete unit; it can never be only partially complete.

有几个关键词：individual、indivisible、complete。换个好理解的说法，事务就是一件最小的事，成功或是失败无所谓，但必须完成。

PCIe 的数据/信息的传递都是一个事务。好比一辆汽车运送一批货物到一个地方去。这就是一个事务。PCIe 总线好比汽车要走的路，货物就是数据，两个 PCIe 设备分别是起点和终点。

还有两个概念：事务是需要发起和完成的。事务的发起者称为 Requester，响应这个事务的称为 Completer。

根据事务访问的地址空间（也就是前面我们讲的几种资源）不同，可以分成如下四类：Configuration, IO, Memory, and Message，分别用于访问系统内的不同的资源。

Table 2-1: Transaction Types for Different Address Spaces

Address Space	Transaction Types	Basic Usage
Memory	Read Write	Transfer data to/from a memory-mapped location
I/O	Read Write	Transfer data to/from an I/O-mapped location
Configuration	Read Write	Device Function configuration/setup
Message	Baseline (including Vendor–defined)	From event signaling mechanism to general purpose messaging

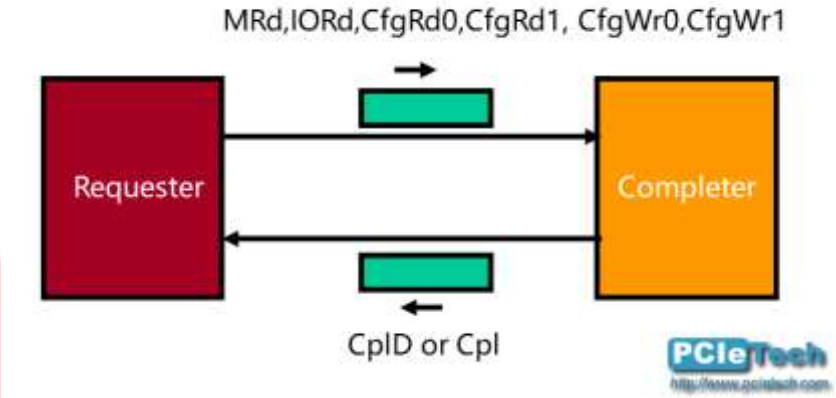
这四种类型的事务还可以根据是否需要完成报文分成三组：Posted、Non-Posted 和 Completion。

- Posted 的事务是指不需要完成报文的，比如存储器写事务，即写某个存储空间。
- Non-Posted 的事务是需要完成的报文返回的，比如读存储空间的事务，是需要读到的具体的值组成的完成报文返回的。
- Completion 则是具体返回的报文的事务。

10.6.3.3 NON-POSTED TRANSACTION 简介

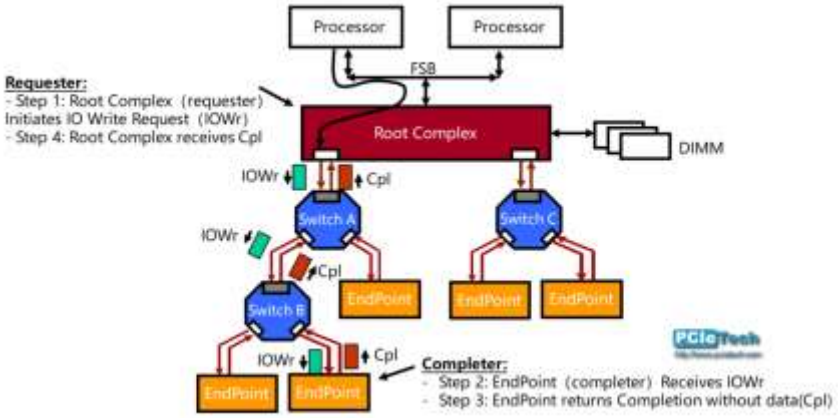
与 Posted 事务的关键差异是，Non-Posted 事务的发起者是需要收到从目标设备返回的完成报文后，事务才算结束。如果目标设备由于某种原因并未返回完成，那将会导致发起者接收完成超时 CO（Completion Timeout）。Configuration read and write, IO read and write, and Memory read 都属于 Non-Posted 事务。读很好理解，需要返回数据。而 Config write 和 IO write 这两种对设备的访问，都是必须要求得到设备的响应的，所以它们也是 Non-Posted 事务。

Completer 返回的完成报文（Cpl）可能是成功（SC: Success Comletion）,也可能是 CA（Completer Abort）或者 UR（Unsupported Request）。当然，Non-Posted 的发起者是知道这个报文是不是到达了目的地。

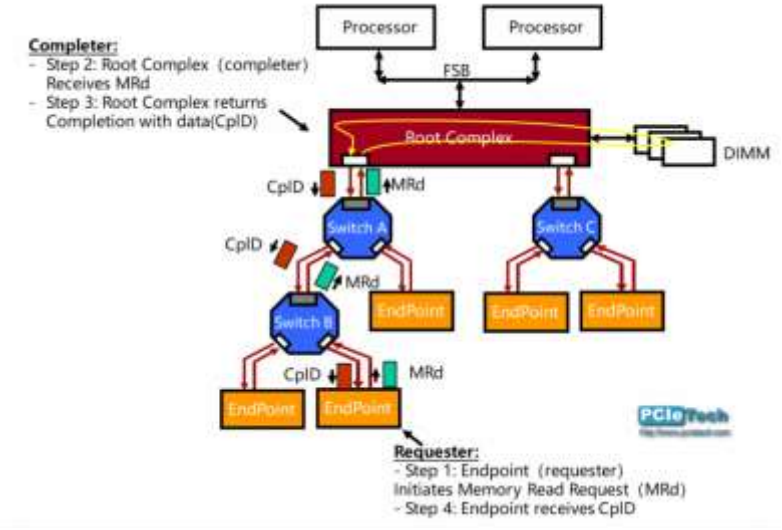


我们看几个例子：

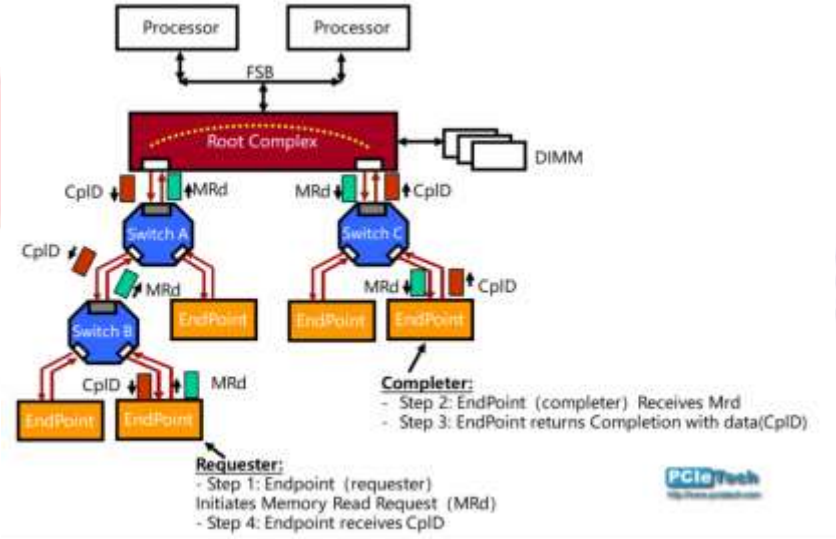
1、首先是 IO Write 的例子，RC 向 EP 发起 IO 写请求，EP 收到后，返回给 RC 的完成报文是不带数据的 Cpl。



2、EP 发起内存读请求，通常是 EP 上的 DMA 发起，RC 把数据返回给 EP，完成报文是带数据的。



3、EP 向另一个 EP 发起读请求，注意 RC 中的黄色虚线，前面我们提过，这种称之为 Peer-to-peer 传输。另一个 EP 返回带数据的完成报文。



附：文中几个缩写的含义。

缩写	含义
MRd	Memory Read
IORd	IO REad
CfgRd0, CfgRd1	Configuration Read（Type 0 and Type 1）
CfgWr0,CfgWr1	Configuration Write（Type 0 and Type 1）
CplD	Completion with Data

Cpl	Completion without Data

Config 有 type0 和 type1 两种，分别对应 EP 和 Switch。

10.6.3.4 POSTED TRANSACTION 简析

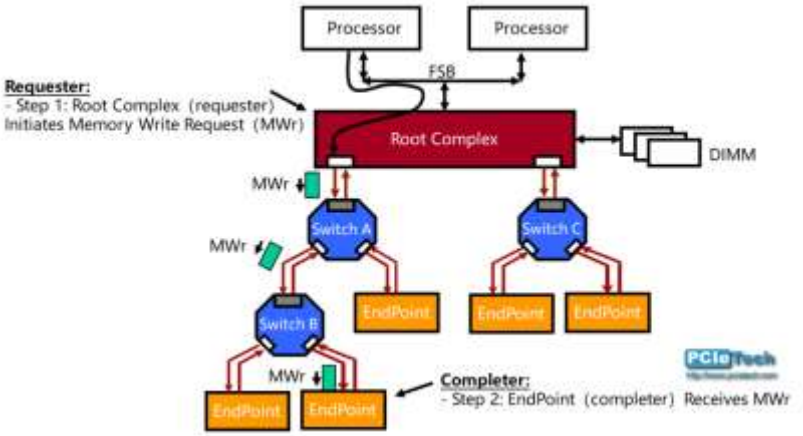
所谓 Posted Transaction，意思是指当数据发送到接收端的设备后，当前事务就完成了，发起事务的 Requester 就可以继续做下一件事了。因此，Posted Transaction 不会导致 Requester hang 住。发起事务的 Requester 并不知道什么时候数据到达，甚至不知道是否真的到达了目的地。事务在传输路径上丢失，系统也不会收到任何的通知。

Posted Transaction 主要包括 Memory write 和消息，并通过地址路由或者隐式路由。

图中的 MWr 表示 Memory Write Request。

注意：没有从 Completer 返回给 Requester 的完成报文。

我们来看看 Posted Transaction 通过 Switch 的情况，如下图：



- 正常情况下，报文从 SwitchA 的 ingress port（图中是 upstream port）进入，并被 Switch A 路由到 egress port，再进入 Switch B 的 upstream port，继续上述的过程，直到最终到达本次事务的接收者（即最下面的 EP）。
- 假如报文进入 Switch A 以后，Switch A 的下行口（egress port）的 link 已经 down 掉了，即 SwitchA 和 Switch B 之间的链路断掉。那么这个报文是不能从 SwitchA 的下行口发送出去的。Switch A 只能默默的把这个报文给“吃掉”，并且很遗憾，Requester 并不知道。

最后提醒一下，有个概念很容易迷惑初学者。所谓的 Posted Transaction，仅是指没有完成报文的返回，并不是说 Completer 没有应答，PCIe 的点对点传输，采用了 Ack/Nak 的机制，后续我们会讲到。

举个例子：我白给你发红包，并且你不需要给我打收据，这就是 Posted 事务。

你说收到，谢谢，这个是应答 Ack。你说不要，使不得，这也是个应答 Nak。

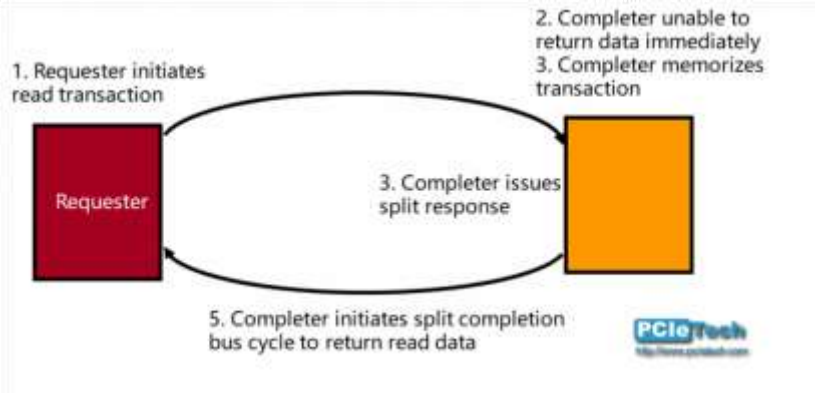
我把红包给了另一个人，告诉他转交给你，类似于报文经过了 Switch。他答应了我转交，我认为这个 Posted 事务就完成了。至于另一个人是不是把红包真的给了你，我不得而知。🤦

10.6.3.5 COMPLETION 介绍

在学习 Non-Posted Transaction 之前，我们首先来了解一下什么是 Completion（完成）。

PCIe 系统里，所有的读请求（Read request）都是分离事务（split transactions）。所谓分离，是指读命令和读到的数据并不是同步返回的。Requester（发起读请求的设备）发起一个读请求，Completer（返回读取的数据给 requester 的设备）稍后才会把完成数据返回给 Requester。数据读事务被分离成两个步骤：读请求（Read request）和完成（Completion）。

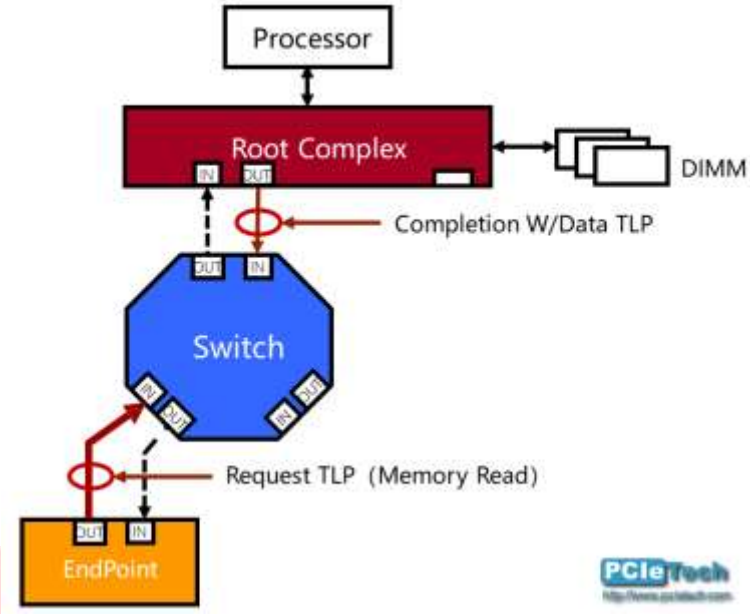
IO 和配置写也是分离事务，同样有完成报文。不过 IO 和配置写的完成报文纯粹是为了确认当前事务真正完成而已，完成报文并无真正意义的数据。



完成报文是由 Completer 在 Non-Poster 事务中返回给 Requester 的。跟 Posted 事务类似，Completer 发出完成包后，同样不知道这个包是否则真正到达了目标设备。

由于 PCIe 是点对点的传输，对于 Completer 来说，只要这个完成报文到达与 Competer 连接的设备，就认为这此事务完成了。

不同的是：如果完成报文并没有最终到达目标设备，Requester 端会触发完成超时（Completion timeout）。



注意：完成报文是基于 ID 路由的。如上图：EP 发起了一个读请求到 RC，同时，开启一个定时器，RC 收到读请求之后，稍后会把完成报文按 ID 路由发送给 EP。如果这个完成报文在 Switch 中路由到下行口时，Switch 到 EP 的链路 down 掉，那么这个完成报文同样会被 Switch“吃掉”。同时，当 Requester 一直收不到完成报文，定时器超时，会触发一个完成超时（Completion timeout）的异常。

10.6.4 PCIe 错误处理

10.6.4.1 PCIE 错误类型简析

我们开始讨论一下有关于 PCIe 错误相关的内容。每种协议都有各自定义的错误。每种硬件也会有各种错误。从某种意义上来说，开发人员存在的价值不仅在于实现各种协议的实现，更多的是去解决这些千奇百怪的错误。错误的相关内容包括：错误分类、错误检测、错误记录、错误通知、错误报告、错误处理等等。

Spec 早期（1.0）沿用了很多 PCI/PCI-X 协议的内容，要求每种 PCIe 设备都必须实现基础的错误模式。在 PCIe 1.1，新增了高级错误报告 AER（Advanced Error Reporting），但协议并没有要求每个设备必须实现，属于可选项。

首先，我们来看看错误的分类：

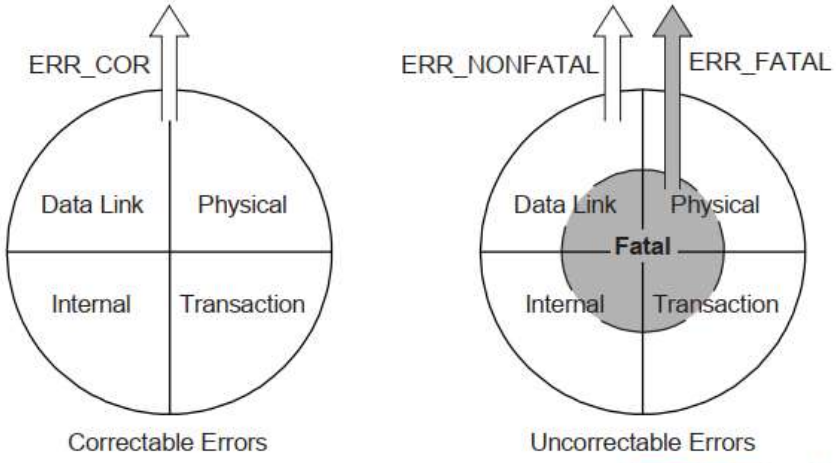


Figure 6-1: Error Classification

PCIe 的错误可以分成两类：不可修复错误（Uncorrectable errors）和可修复错误（Correctable errors）。这些错误在 PCIe 协议的各个层次（物理层、数据链路层、事务层）都会存在。另外 Internal error 是 PCIe 2.0 之后新增的，是指 PCIe 设备硬件的内部错误。

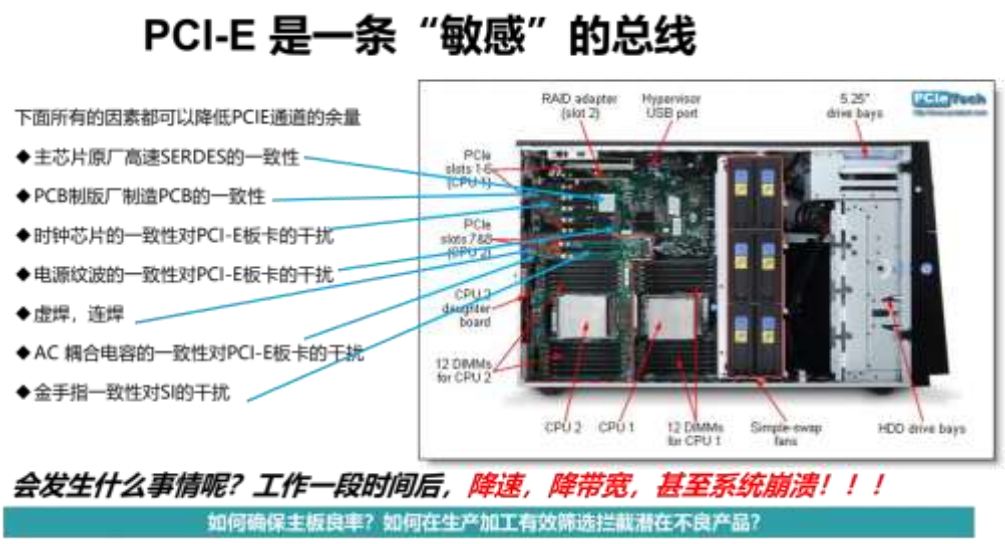
很显然，所谓可修复错误，是指这类错误是有可能被修复的。注意，是有可能！当修复不了，就会变成不可修复错误。可修复错误往往由硬件自我修复，如重发报文、重协商等等。这类错误只对系统的性能有所影响。可修复错误由硬件修复不需要软件参与，并且修复行为不会导致任何信息的丢失。在软件处理上，可以记录错误发生的首次时间点、发生频率等，以便于后期定位追溯问题。

不可修复错误是指无法自我修复的错误，可能导致某次事务的失败。不可修复错误又可以细分为致命（Fatal）和非致命（Non-Fatal）两种。非致命错误往往只会影响单次 PCIe 事务的失败。而致命错误可能会导致系统的崩溃！

不可修复致命错误是链路或者硬件不可靠导致的，对于不可修复致命错误需要复位链路上的组件。不可修复致命错误，没有统一的修复方法，每家都有自己的处理方法。平台设计者需要根据硬件设计不同，PCIe 器件承担的作用不同，业务流程不同进行不同的处理，原则上都需要复位链路上的组件。

我们再总结一下错误类型：

	可修正错误	不可修正错误
发生范围	物理层、数据链路层、事务层、设备内部	
自我修复	是	否
系统影响	影响性能	影响功能
软件处理	无需处理。可以考虑记录日志。	建议处理。



最后，需要提醒大家注意的是：对于 PCIe 错误，各个厂商和公司的要求不尽一致。很多厂商的设备，在生产基本都没有做任何的检测，就直接发货。这样会存在一定的隐患。对于个人 PC 领域，影响还属可控。如果是服务器、存储，就会存在业务性能降低、业务中断等重大影响。

10.6.4.2 ECRC VS LCRC 是做什么用途的呢？

TL 层的 TLP 报文有个 ECRC，而 DLL 层对 TLP 会加上序列号和 LCRC。两个 CRC 有何不同呢？

LCRC：

- LCRC 是 Link CRC，是用来检测在两个设备之间的链路（Link）上发生的错误，这些错误大多是由于物理层的信号质量问题引起的。
- 因为仅是针对 Link 和互联的两个设备，LCRC 换个说法也可以叫 Local CRC。
- 当 TLP 报文通过 Switch 时，在入口（ingress port）被 DLL 层校验 LCRC。而在 Switch 的出口（egress port）会重新生成 LCRC。
- 当接收端检测到 TLP 报文的 LCRC 不正确时，将会发送一个 NAK 报文给发送端，要求重传此 TLP。因此，LCRC 错误是可修正错误，因为可以通过重传来修正。
- 重传的过程并不需要软件参与。

ECRC：

- ECRC 是 End-to-End CRC，是检测 TLP 报文的内容是否在 Requester 和 Completer 之间的路径上是否损坏。例如 switch 在转发 TLP 报文时由于某些 bug 发生了数据不一致。
- ECRC 是 requester/initiaor 的 TL 层生成的。如果 Completer 检测到 ECRC 错误，会上报通知系统处理。并没有重传机制，因此是不可修正错误。
- 当 TLP 报文进入 Switch 时，Switch 并不会修改 ECRC，并完整把 TLP 报文发出去。换言之，TLP 经过路径上的 Switch 内容不会发生任何改变，甚至 ECRC 是错误的，Switch 也视而不见继续发出去。
- TLP 的错误，需要软件参与，由上层软件决定是否重发。

下表列出了两者的差异:

类别	LCRC	ECRC
含义	Link CRC	End-to-End CRC
长度	32-bit	
源/目的	两个连接的设备之间	Requeseter 和 Completer 之间
通过 Switch 时	在出口重新计算生产	Switch 不会修改
Switch 收到错误报文时	要求对端重发	记录错误，继续转发
完成者收到错误报文	要求对端重发	发出一个不可修正错误消息通知系统处理
是否可修正	可修正	不可修正

10.6.4.3 Linux Kernel 的 AER 是怎么工作的？

在 Linux 下，AER（Advanced Error Report）驱动都是注册给通用的 Port Bus Driver，称之为 service driver。除了 AER，还有热插拔等驱动也是采用类似的注册方式。PCIe Port Bus 驱动的内容及具体实现我们后面再详细讲。

```
/**
 * aer_probe - initialize resources
 * @dev: pointer to the pcie_dev data structure
 *
 * Invoked when PCI Express bus loads AER service driver.
 */
static int aer_probe(struct pcie_device *dev)
{
    int status;
    struct aer_rpc *rpc;
    struct device *device = &dev->device;

    rpc = devm_kzalloc(device, sizeof(struct aer_rpc), GFP_KERNEL);
    if (!rpc) {
        dev_printk(KERN_DEBUG, device, "alloc AER rpc failed\n");
        return -ENOMEM;
    }
    rpc->rpd = dev->port;
    set_service_data(dev, rpc);

    status = devm_request_threaded_irq(device, dev->irq, aer_irq, aer_isr,
                    IRQF_SHARED, "aerdrv", dev);
    if (status) {
        dev_printk(KERN_DEBUG, device, "request AER IRQ %d failed\n",
                dev->irq);
        return status;
    }

    aer_enable_rootport(rpc);
    dev_info(device, "AER enabled with IRQ %d\n", dev->irq);
    return 0;
} ? end aer_probe ?
```

```
static struct pcie_port_service_driver aerdriver = {
    .name       = "aer",
    .port_type  = PCI_EXP_TYPE_ROOT_PORT,
    .service    = PCIE_PORT_SERVICE_AER,

    .probe      = aer_probe,
    .remove     = aer_remove,
    .reset_link = aer_root_reset,
};

/**
 * aer_service_init - register AER root service driver
 *
 * Invoked when AER root service driver is loaded.
 */
int __init pcie_aer_init(void)
{
    if (!pci_aer_available() || aer_acpi_firmware_first())
        return -ENXIO;
    return pcie_port_service_register(&aerdriver);
}
```


AER service 和其他的 PCIe service 一样，都是注册给 pci port bus driver。Port bus driver 会调用 aer_probe，aer_probe 函数注册了中断服务函数和中断线程，然后调用 aer_enable_rootport 来使能中断。aer_enable_rootport 其实就是清除 error status reg，并且使能前面“Error message 控制”提到的相关控制寄存器。简而言之，就是把所有该使能的地方都使能起来，确保错误能够正常的上报上去。

```
/**
 * aer_irq - Root Port's ISR
 * @irq: IRQ assigned to Root Port
 * @context: pointer to Root Port data structure
 *
 * Invoked when Root Port detects AER messages.
 */
static irqreturn_t aer_irq(int irq, void *context)
{
    struct pcie_device *pdev = (struct pcie_device *)context;
    struct aer_rpc *rpc = get_service_data(pdev);
    struct pci_dev *rp = rpc->rpd;
    struct aer_err_source e_src = {};
    int pos = rp->aer_cap;

    pci_read_config_dword(rp, pos + PCI_ERR_ROOT_STATUS, &e_src.status);
    if (!(e_src.status & (PCI_ERR_ROOT_UNCOR_RCV|PCI_ERR_ROOT_COR_RCV)))
        return IRQ_NONE;

    pci_read_config_dword(rp, pos + PCI_ERR_ROOT_ERR_SRC, &e_src.id);
    pci_write_config_dword(rp, pos + PCI_ERR_ROOT_STATUS, e_src.status);

    if (!kfifo_put(&rpc->aer_fifo, e_src))
        return IRQ_HANDLED;

    return IRQ_WAKE_THREAD;
} ? end aer_irq ?
```


中断服务函数 aer_irq 就是读取 AER CAP 中的 status 和 source id 寄存器来确定是否产生了 AER，收到的第一个错误 message 的 id 是多少，把 status 和 id 推到 fifo 中，然后就启动线程。

```c
/**
 * aer_isr - consume errors detected by root port
 * @work: definition of this work item
 *
 * Invoked, as DPC, when root port records new detected error
 */
static irqreturn_t aer_isr(int irq, void *context)
{
    struct pcie_device *dev = (struct pcie_device *)context;
    struct aer_rpc *rpc = get_service_data(dev);
    struct aer_err_source uninitialized_var(e_src);

    if (kfifo_is_empty(&rpc->aer_fifo))
        return IRQ_NONE;

    while (kfifo_get(&rpc->aer_fifo, &e_src))
        aer_isr_one_error(rpc, &e_src);
    return IRQ_HANDLED;
}
```



```c
/**
 * find_source_device - search through device hierarchy for source device
 * @parent: pointer to Root Port pci_dev data structure
 * @e_info: including detailed error information such like id
 *
 * Return true if found.
 *
 * Invoked by DPC when error is detected at the Root Port.
 * Caller of this function must set id, severity, and multi_error_valid of
 * struct aer_err_info pointed by @e_info properly.  This function must fill
 * e_info->error_dev_num and e_info->dev[], based on the given information.
 */
static bool find_source_device(struct pci_dev *parent,
        struct aer_err_info *e_info)
{
    struct pci_dev *dev = parent;
    int result;

    /* Must reset in this function */
    e_info->error_dev_num = 0;

    /* Is Root Port an agent that sends error message? */
    result = find_device_iter(dev, e_info);
    if (result)
        return true;

    pci_walk_bus(parent->subordinate, find_device_iter, e_info);

    if (!e_info->error_dev_num) {
        pci_printk(KERN_DEBUG, parent, "can't find device of ID%04x\n",
                e_info->id);
        return false;
    }
    return true;
} ? end find_source_device ?
```


线程 aer_isr 从 Root Port 开始 walk_bus 遍历该 root port 下面的所有 PCIe 设备，读取设备 aer status 寄存器和 aer mask 寄存器，如果 status 对应 bit 为 1 且 mask 对应 bit 为 0，则加入 struct aer_err_info *e_info 的数据结构中等待处理。

由于 kernel 需要为大多数使用者负责，追求通用性，导致 aer 的处理乏善可陈，就不在这里介绍了。事实上，这个 AER 在实际的工程应用中，基本不可用。需要根据客户自己的实际环境，做大量的修改和优化。

AER 处理方面，需要考虑产品可靠性要求不同、单板设计不同、芯片在系统中承担的角色等方面，通用性和可靠性不可兼得。Kernel 要为所有硬件服务，因此选择了通用性。系统要想提高可靠性，就要适当牺牲通用性。

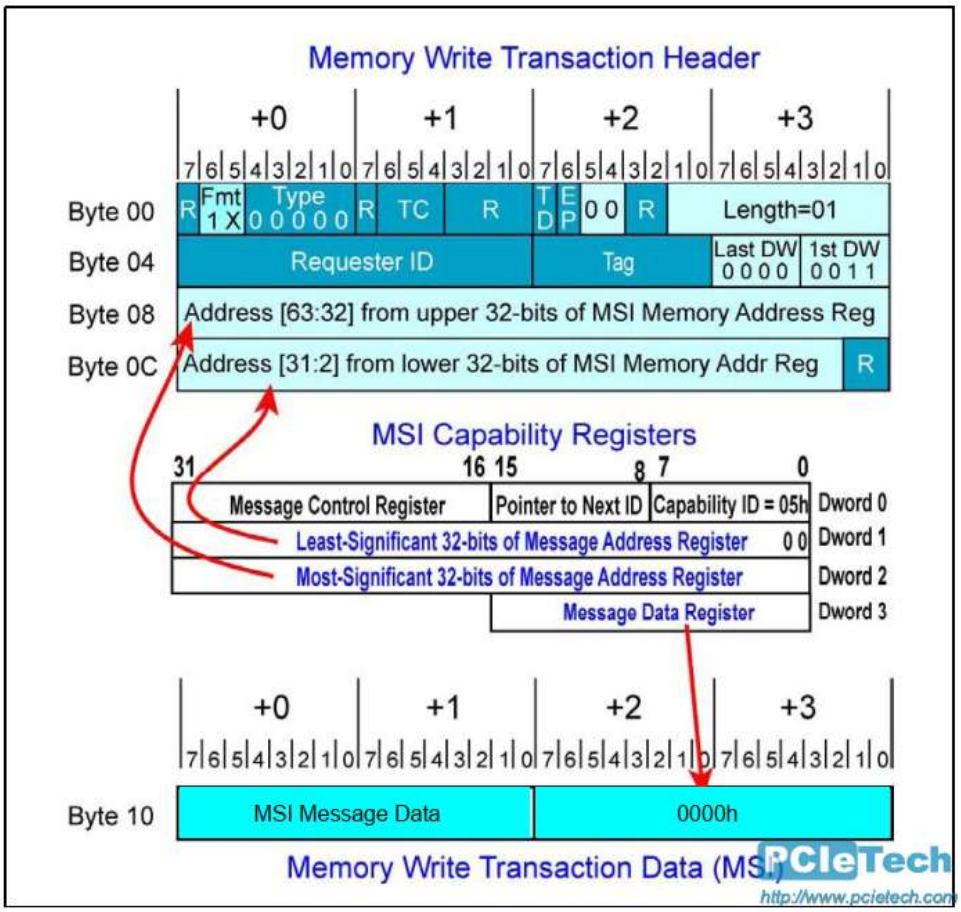
10.6.5 MSI-X 中断解析

10.6.5.1 MSI-X（一）

1. MSI 中断

MSI 中断本质上是一个 memory write，memory write 的地址就是设备配置空间的 MSI address 寄存器的值，memory write 的数据就是设备配置空间的 MSI data 寄存器的值。Message address 寄存器和 message data 寄存器是调用 pci_enable_msi 时，系统软件填入的。

也就是说，一个设备想产生一个 MSI 中断话，只需要使用配置空间的 message address 寄存器和 message data 寄存器发起一个 memory write 的请求，即往 message address 寄存器写入 memory data。在 X86 系统下，message address 对应的 LAPIC 的地址。




```
root@ubuntu:~# cat /proc/iomem | grep apic -i
  ec8f0000-ec8f03ff : IOAPIC 4
  ed1f0000-ed1f03ff : IOAPIC 3
  edaf0000-edaf03ff : IOAPIC 2
  edff0000-edff03ff : IOAPIC 1
  fee00000-fee00fff : Local APIC
Capabilities: [a0] MSI: Enable+ Count=1/1 Maskable- 64bit+
        Address: 00000000fee06000  Data: 4021
```


- 为何要引入 MSI-X

前面讲了 MSI 中断的机制，其实 MSI-X Capability 中断机制与 MSI Capability 的中断机制类似。既然机制类似，为啥还要需要引入 MSI-X 呢？

回答这个问题前，我们先看看 MSI 有哪些限制？

- MSI 相关的寄存器都是在配置空间中，从 Message Control 寄存器 multiple message Capble 字段可以看出 MSI 最多支持 32 个中断向量，且必须是 2^n，也就是说如果一个 function 需要 3 个中断向量，必须申请 4 个才可以满足。
- MSI 要求中断控制器分配给该 function 的中断向量号必须连续。

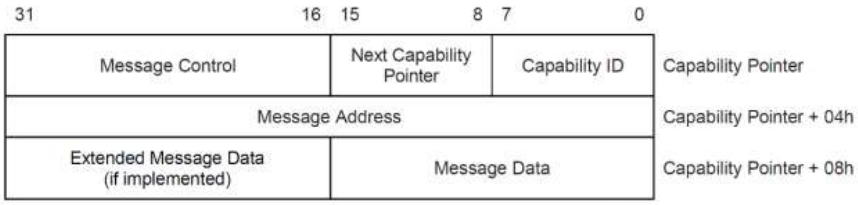


Figure 7-42: MSI Capability Structure for 32-bit Message Address

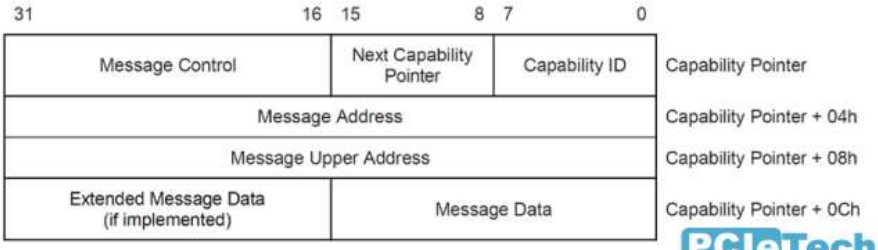


Figure 7-43: MSI Capability Structure for 64-bit Message Address

Bit Location	Register Description	Attributes
0	**MSI Enable** – If Set and the MSI-X Enable bit in the MSI-X Message Control register (see Section 7.9.2) is Clear, the Function is permitted to use MSI to request service and is prohibited from using INTx interrupts. System configuration software Sets this bit to enable MSI. A device driver is prohibited from writing this bit to mask a Function's service request. Refer to Section 7.5.1.1 for control of INTx interrupts. If Clear, the Function is prohibited from using MSI to request service. Default value of this bit is 0b.	RW
3:1	**Multiple Message Capable** – System software reads this field to determine the number of requested vectors. The number of requested vectors must be aligned to a power of two (if a Function requires three vectors, it requests four by initializing this field to 010b). The encoding is defined as: **Encoding** / **# of vectors requested** 000b 1 001b 2 010b 4 011b 8 100b 16 101b 32 110b Reserved 111b Reserved MSI最多支持32个中断向量	RO

MSI-X 的出现就是为了解决上面两个问题，主要是第二个问题。

10.6.5.2 MSI-X（二）

1、 MSI-XCAP 结构

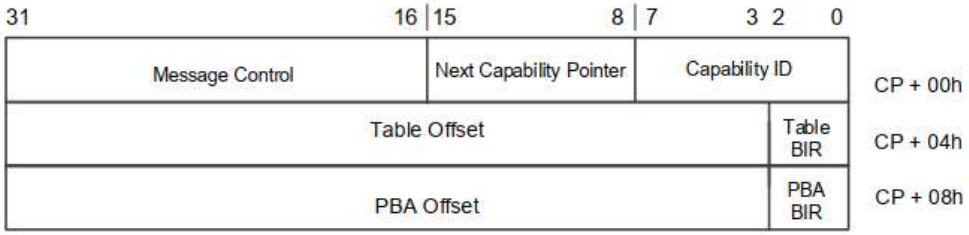


Figure 7-54: MSI-X Capability Structure

MSI-X 和 MSI 最大的不同是 messagedata、message address 和 status 字段没有存放在设备的配置空间中，而是使用 MSI-X Table structure 和 MSI-X PBA structure 来存放这些字段。

MSI-X Table structure 和 PBA structure 存放在设备的 BAR 空间里，这两个 structure 可以 map 到相同 BAR，也可以 map 到不同 BAR，但是这个 BAR 必须是 memory BAR 而不能是 IO BAR，也就是说这两个 structure 要 map 到 memory 空间。

注意：一个 function 只能支持一个 MSI-X CAP。

DWORD 3	DWORD 2	DWORD 1	DWORD 0		
Vector Control	Msg Data	Msg Upper Addr	Msg Addr	entry 0	Base
Vector Control	Msg Data	Msg Upper Addr	Msg Addr	entry 1	Base + 1*16
Vector Control	Msg Data	Msg Upper Addr	Msg Addr	entry 2	Base + 2*16
...	...	...	...	...	...
Vector Control	Msg Data	Msg Upper Addr	Msg Addr	entry (N-1)	Base + (N-1)*16

Figure 7-55: MSI-X Table Structure

存放在设备BAR空间中

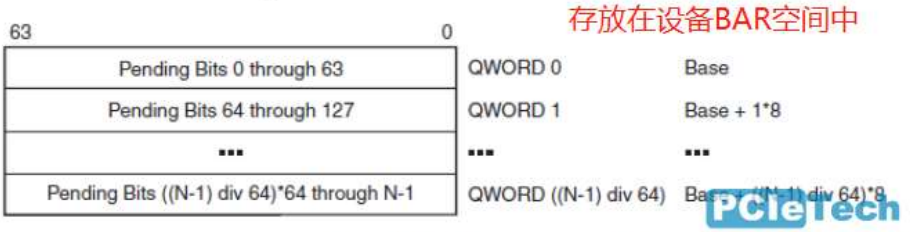


Figure 7-56: MSI-X PBA Structure

1.1 配置空间的 Message control 寄存器

配置空间 Message Control 寄存器中的 table size 字段表示 MSI-X 的 table 的大小。软件读取该字段获取 table size，table size+1 就是 MSI-X table entry 的个数，也就是 Figure 7-36 中的 entry(N-1)中的 N，每个 entry 对应一个中断向量。从 table size 可以看出 1 个 function 最多可以支持 2^{11}=2048 个 MSI-X 中断。

Figure 7-58: Message Control Register for MSI-X

Table 7-45: Message Control Register for MSI-X

Bit Location	Register Description	Attributes
10:0	**Table Size** – System software reads this field to determine the MSI-X Table Size N, which is encoded as N-1. For example, a returned value of 000 0000 0011b indicates a table size of 4.	RO

pci_msix_vec_count 读取 Message Control 寄存器的 table size 字段获取 MSI-X table entry 的个数。

调用关系如下：

pci_enable_msix_range->__pci_enable_msix_range->__pci_enable_msix->pci_msix_vec_count

```
/**
 * pci_msix_vec_count - return the number of device's MSI-X table entries
 * @dev: pointer to the pci_dev data structure of MSI-X device function
 * This function returns the number of device's MSI-X table entries and
 * therefore the number of MSI-X vectors device is capable of sending.
 * It returns a negative errno if the device is not capable of sending MSI-X
 * interrupts.
 **/
int pci_msix_vec_count(struct pci_dev *dev)
{
    u16 control;

    if (!dev->msix_cap)
        return -EINVAL;

    pci_read_config_word(dev, dev->msix_cap + PCI_MSIX_FLAGS, &control);
    return msix_table_size(control);
}
EXPORT_SYMBOL(pci_msix_vec_count);
```



```
#define msix_table_size(flags)  ((flags & PCI_MSIX_FLAGS_QSIZE) + 1)
```

1.2 配置空间的 Table offset/Table BIR 寄存器

配置空间中 Table offset/Table BIR 寄存器的 Table BIR 字段指示使用哪个 BAR 来映射的 MSI-X Table structure。该字段的 0-5 也对应 function 的 BAR0-5。

Table offset 字段代表 MSI-X Table structure entry 0 存放在 BAR 空间的偏移。

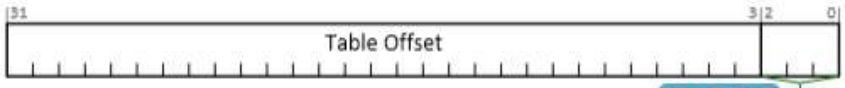


Figure 7-59: Table Offset/Table BIR Register for MSI-X

Bit Location	Register Description	Attributes
2:0	**Table BIR** – Indicates which one of a Function's Base Address registers, located beginning at 10h in Configuration Space, or entry in the Enhanced Allocation capability with a matching BEI, is used to map the Function's MSI-X Table into Memory Space. **BIR Value** / **Base Address register** 0 – 10h 1 – 14h 2 – 18h 3 – 1Ch 4 – 20h 5 – 24h 6 – Reserved 7 – Reserved For a 64-bit Base Address register, the Table BIR indicates the lower DWORD. For Functions with Type 1 Configuration Space headers, BIR values 2 through 5 are also Reserved.	RO
31:3	**Table Offset** – Used as an offset from the address contained by one of the Function's Base Address registers to point to the base of the MSI-X Table. The lower 3 Table BIR bits are masked off (set to zero) by software to form a 32-bit QWORD-aligned offset.	RO

kernel 代码先读取配置空间的 Table offset/Table BIR 寄存器，通过 Table BIR 字段获取使用哪个 BAR 来映射 MSI-X Table structure。然后计算出 MSI-X table entry0 相对 BAR 空间偏移的物理地址（phys_addr），最后 ioremap 得到虚拟地址。

调用关系如下：

pci_enable_msix_range->__pci_enable_msix_range->__pci_enable_msix->msix_capability_init

```
static void __iomem *msix_map_region(struct pci_dev *dev, unsigned nr_entries)
{
    resource_size_t phys_addr;
    u32 table_offset;
    unsigned long flags;
    u8 bir;

    pci_read_config_dword(dev, dev->msix_cap + PCI_MSIX_TABLE,
                          &table_offset);
    bir = (u8)(table_offset & PCI_MSIX_TABLE_BIR);
    flags = pci_resource_flags(dev, bir);
    if (!flags || (flags & IORESOURCE_UNSET))
        return NULL;

    table_offset &= PCI_MSIX_TABLE_OFFSET;
    phys_addr = pci_resource_start(dev, bir) + table_offset;

    return ioremap_nocache(phys_addr, nr_entries * PCI_MSIX_ENTRY_SIZE);
}
```


1.3 配置空间的 PBA offset/PBA BIR 寄存器

配置空间中 PBA offset/PBA BIR 寄存器的 PBA BIR 字段指示使用哪个 BAR 来映射的 PBA structure。该字段的 0-5 也对应 function 的 BAR0-5。

PBA offset 字段代表 PBA structure 存放在 BAR 空间的偏移。

7.7.2.4 PBA Offset/PBA BIR Register for MSI-X (Offset 08h)

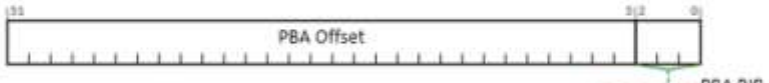


Figure 7-60: PBA Offset/PBA BIR Register for MSI-X

Table 7-47: PBA Offset/PBA BIR Register for MSI-X

Bit Location	Register Description	Attributes
2:0	**PBA BIR** – Indicates which one of a Function's Base Address registers, located beginning at 10h in Configuration Space, or entry in the Enhanced Allocation capability with a matching BEI, is used to map the Function's MSI-X PBA into Memory Space. The PBA BIR value definitions are identical to those for the MSI-X Table BIR.	RO
31:3	**PBA Offset** – Used as an offset from the address contained by one of the Function's Base Address registers to point to the base of the MSI-X PBA. The lower 3 PBA BIR bits are masked off (set to zero) by software to form a 32-bit QWORD-aligned offset.	RO

2、 Memory 空间的 MSI-X table structure

Message address 和 Message Upper address 字段存放的是 MSI-X memory write 请求需要使用的地址。

Message Data 字段存放的是 MSI-X memory write 请求需要使用的 data。该地址和 CPU 的架构相关，是使能 MSI-X 时，系统软件写入的。

DWORD 3	DWORD 2	DWORD 1	DWORD 0		
Vector Control	Msg Data	Msg Upper Addr	Msg Addr	entry 0	Base
Vector Control	Msg Data	Msg Upper Addr	Msg Addr	entry 1	Base + 1*16
Vector Control	Msg Data	Msg Upper Addr	Msg Addr	entry 2	Base + 2*16
...	...	...	...	...	...
Vector Control	Msg Data	Msg Upper Addr	Msg Addr	entry (N-1)	Base + (N-1)*16

Figure 7-55: MSI-X Table Structure

Figure 7-61: Message Address Register for MSI-X Table Entries

Table 7-48: Message Address Register for MSI-X Table Entries

Bit Location	Register Description	Attributes
1:0	**Message Address** – For proper DWORD alignment, software must always write zeroes to these two bits; otherwise the result is undefined. Default value of this field is 00b. These bits are permitted to be read-only or read-write.	RO or RW
31:2	**Message Address** – System-specified message lower address. For MSI-X messages, the contents of this field from an MSI-X Table entry specifies the lower portion of the DWORD-aligned address for the Memory Write transaction.	RW

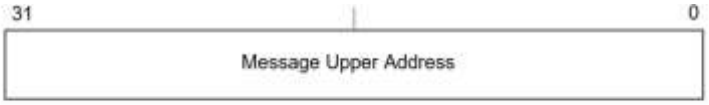


Figure 7-62: Message Upper Address Register for MSI-X Table Entries

Table 7-49: Message Upper Address Register for MSI-X Table Entries

Bit Location	Register Description	Attributes
31:0	**Message Upper Address** – System-specified message upper address bits. If this field is zero, 32-bit address messages are used. If this field is non-zero, 64-bit address messages are used.	RW

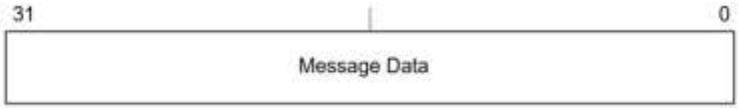


Figure 7-63: Message Data Register for MSI-X Table Entries

Table 7-50: Message Data Register for MSI-X Table Entries

Bit Location	Register Description	Attributes
31:0	**Message Data** – System-specified message data. For MSI-X messages, the contents of this field from an MSI-X Table entry specifies the 32-bit data payload of the DWORD Memory Write transaction. All 4 Byte Enables are Set. In contrast to message data used for MSI messages, the low-order message data bits in MSI-X messages are not modified by the Function. This field is read-write.	RW

Kernel 函数__pci_write_msi_msg 把 message address、message data 写入对应的 entry。

如果是 X86 的 CPU，存放 message address 和 message data 的结构 struct msi_msg *msg 是在 irq_msi_compose_msg 中初始化的，这个值和 CPU 架构相关。

desc->mask_base 是 MSI-X table entry 0 对应的虚拟地址，desc->msi_attrib.entry_nr 是 MSI-X table entry 编号。desc->mask_base 和 desc->msi_attrib.entry_nr 都是在 msix_setup_entries 赋值的（pci_enable_msix_range->__pci_enable_msix_range->

__pci_enable_msix->msix_capability_init->msix_setup_entries）。

X86 下调用关系：

pci_enable_msix_range->__pci_enable_msix_range->__pci_enable_msix->msix_capability_init-> pci_msi_setup_msi_irqs-> arch_setup_msi_irqs->native_setup_msi_irqs

msi_domain_alloc_irqs->irq_domain_activate_irq->__irq_domain_activate_irq->msi_domain_activate->irq_chip_write_msi_msg->pci_msi_domain_write_msg->__pci_write_msi_msg

```
=> __pci_write_msi_msg
=> msi_domain_activate
=> __irq_domain_activate_irq
=> irq_domain_activate_irq
=> msi_domain_alloc_irqs
=> native_setup_msi_irqs
=> __pci_enable_msix_range
=> igb_set_interrupt_capability
=> igb_init_interrupt_scheme
=> igb_probe
=> local_pci_probe
=> work_for_cpu_fn
=> process_one_work
=> worker_thread
=> kthread
=> ret_from_fork
```



```
struct x86_msi_ops x86_msi __ro_after_init = {
    .setup_msi_irqs     = native_setup_msi_irqs,
    .teardown_msi_irq   = native_teardown_msi_irq,
    .teardown_msi_irqs  = default_teardown_msi_irqs,
    .restore_msi_irqs   = default_restore_msi_irqs,
};

/* MSI arch specific hooks */
int arch_setup_msi_irqs(struct pci_dev *dev, int nvec, int type)
{
    return x86_msi.setup_msi_irqs(dev, nvec, type);
}
```



```
static void __iomem *pci_msix_desc_addr(struct msi_desc *desc)
{
    return desc->mask_base +
        desc->msi_attrib.entry_nr * PCI_MSIX_ENTRY_SIZE;
}
```



```
void __pci_write_msi_msg(struct msi_desc *entry, struct msi_msg *msg)
{
    struct pci_dev *dev = msi_desc_to_pci_dev(entry);

    if (dev->current_state != PCI_D0 || pci_dev_is_disconnected(dev)) {
        /* Don't touch the hardware now */
    } else if (entry->msi_attrib.is_msix) {
        void __iomem *base = pci_msix_desc_addr(entry);

        writel(msg->address_lo, base + PCI_MSIX_ENTRY_LOWER_ADDR);
        writel(msg->address_hi, base + PCI_MSIX_ENTRY_UPPER_ADDR);
        writel(msg->data, base + PCI_MSIX_ENTRY_DATA);
    } else {
        int pos = dev->msi_cap;
        u16 msgctl;

        pci_read_config_word(dev, pos + PCI_MSI_FLAGS, &msgctl);
        msgctl &= ~PCI_MSI_FLAGS_QSIZE;
        msgctl |= entry->msi_attrib.multiple << 4;
        pci_write_config_word(dev, pos + PCI_MSI_FLAGS, msgctl);

        pci_write_config_dword(dev, pos + PCI_MSI_ADDRESS_LO,
                               msg->address_lo);
        if (entry->msi_attrib.is_64) {
            pci_write_config_dword(dev, pos + PCI_MSI_ADDRESS_HI,
                                   msg->address_hi);
            pci_write_config_word(dev, pos + PCI_MSI_DATA_64,
                                  msg->data);
        } else {
            pci_write_config_word(dev, pos + PCI_MSI_DATA_32,
                                  msg->data);
        }
    } ? end else ?
    entry->msg = *msg;
} ? end __pci_write_msi_msg ?
```



```
static void irq_msi_compose_msg(struct irq_data *data, struct msi_msg *msg)
{
    struct irq_cfg *cfg = irqd_cfg(data);

    msg->address_hi = MSI_ADDR_BASE_HI;

    if (x2apic_enabled())
        msg->address_hi |= MSI_ADDR_EXT_DEST_ID(cfg->dest_apicid);

    msg->address_lo =
        MSI_ADDR_BASE_LO |
        ((apic->irq_dest_mode == 0) ?
            MSI_ADDR_DEST_MODE_PHYSICAL :
            MSI_ADDR_DEST_MODE_LOGICAL) |
        MSI_ADDR_REDIRECTION_CPU |
        MSI_ADDR_DEST_ID(cfg->dest_apicid);

    msg->data =
        MSI_DATA_TRIGGER_EDGE |
        MSI_DATA_LEVEL_ASSERT |
        MSI_DATA_DELIVERY_FIXED |
        MSI_DATA_VECTOR(cfg->vector);
} ? end irq_msi_compose_msg ?
```

Vector Control 字段存放的是控制字段，当 Mask Bit 为 1 时，PCIe 设备不能使用该 MSI-X tableentry 来发送中断消息。

如果其他的 MSI-X table entry 也是使用的相同的 vector，只要对应 entry 的 vector control 寄存器的 mask bit 字段不为 1，仍然可以使用该 vector 发送 MSI-X 中断消息。这个意思是说 Mask Bit 的作用范围是该 entry 的，如果两个 entry 使用相同的 vector（对 X86 来说就是 Message Data 字段低 8 bit 相同），Mask Bit 不为 1 的 entry 是可以使用该 vector 发出 message 中断。

Figure 7-64: Vector Control Register for MSI-X Table Entries

Table 7-51: Vector Control Register for MSI-X Table Entries

Bit Location	Register Description	Attributes
0	**Mask Bit** – When this bit is Set, the Function is prohibited from sending a message using this MSI-X Table entry. However, any other MSI-X Table entries programmed with the same vector will still be capable of sending an equivalent message unless they are also masked. Default value of this bit is 1b (entry is masked)	RW

此时 MSI-X 中断还没有完全初始化完毕，Kernel 代码是把 MSI-X vector control 寄存器的 Mask bit 置 1 来 mask 所有 vector 的中断。

调用关系：

pci_enable_msix_range->__pci_enable_msix_range->__pci_enable_msix->msix_capability_init->msix_program_entries

```
/*
 * This internal function does not flush PCI writes to the device.
 * All users must ensure that they read from the device before either
 * assuming that the device state is up to date, or returning out of this
 * file.  This saves a few milliseconds when initialising devices with lots
 * of MSI-X interrupts.
 */
u32 __pci_msix_desc_mask_irq(struct msi_desc *desc, u32 flag)
{
    u32 mask_bits = desc->masked;

    if (pci_msi_ignore_mask)
        return 0;

    mask_bits &= ~PCI_MSIX_ENTRY_CTRL_MASKBIT;
    if (flag)
        mask_bits |= PCI_MSIX_ENTRY_CTRL_MASKBIT;
    writel(mask_bits, pci_msix_desc_addr(desc) + PCI_MSIX_ENTRY_VECTOR_CTRL);

    return mask_bits;
}

static void msix_mask_irq(struct msi_desc *desc, u32 flag)
{
    desc->masked = __pci_msix_desc_mask_irq(desc, flag);
}
```

```
static void msix_program_entries(struct pci_dev *dev,
                 struct msix_entry *entries)
{
    struct msi_desc *entry;
    int i = 0;

    for_each_pci_msi_entry(entry, dev) {
        if (entries)
            entries[i++].vector = entry->irq;
        entry->masked = readl(pci_msix_desc_addr(entry) +
                      PCI_MSIX_ENTRY_VECTOR_CTRL);
        msix_mask_irq(entry, 1);
    }
}
```


3、 什么时候 umask 的 vector 中断呢？

以网卡为例，在 request_irq 的时候才把 MSI-X 的使用的 vector 给 unmask 的。

__igb_open->request_threaded_irq->__setup_irq->irq_startup->__irq_startup->unmask_irq->pci_msi_unmask_irq->msi_set_mask_bit->msix_mask_irq->__pci_msix_desc_mask_irq

```
=> do_vfs_ioctl
=> ksys_ioctl
=> __x64_sys_ioctl
=> do_syscall_64
=> entry_SYSCALL_64_after_hwframe
      ifconfig-1664  [024] d...  1130.023872: myprobe: (msix_mask_irq+0x0/0x40) flag=0
      ifconfig-1664  [024] d...  1130.023875: <stack trace>
=> msix_mask_irq
=> msi_set_mask_bit
=> unmask_irq.part.40
=> __irq_startup
=> irq_startup
=> __setup_irq
=> request_threaded_irq
=> __igb_open
=> __dev_open
=> __dev_change_flags
=> dev_change_flags
=> devinet_ioctl
=> inet_ioctl
=> sock_do_ioctl
=> sock_ioctl
=> do_vfs_ioctl
=> ksys_ioctl
=> __x64_sys_ioctl
=> do_syscall_64
=> entry_SYSCALL_64_after_hwframe
```



```
u32 __pci_msix_desc_mask_irq(struct msi_desc *desc, u32 flag)
{
    u32 mask_bits = desc->masked;
    void __iomem *desc_addr;

    if (pci_msi_ignore_mask)
        return 0;
    desc_addr = pci_msix_desc_addr(desc);
    if (!desc_addr)
        return 0;

    mask_bits &= ~PCI_MSIX_ENTRY_CTRL_MASKBIT;
    if (flag)
        mask_bits |= PCI_MSIX_ENTRY_CTRL_MASKBIT;

    writel(mask_bits, desc_addr + PCI_MSIX_ENTRY_VECTOR_CTRL);

    return mask_bits;
}

static void msix_mask_irq(struct msi_desc *desc, u32 flag)
{
    desc->masked = __pci_msix_desc_mask_irq(desc, flag);
}

static void msi_set_mask_bit(struct irq_data *data, u32 flag)
{
    struct msi_desc *desc = irq_data_get_msi_desc(data);

    if (desc->msi_attrib.is_msix) {
        msix_mask_irq(desc, flag);
        readl(desc->mask_base);      /* Flush write to device */
    } else {
        unsigned offset = data->irq - desc->irq;
        msi_mask_irq(desc, 1 << offset, flag << offset);
    }
}

/**
 * pci_msi_mask_irq - Generic IRQ chip callback to mask PCI/MSI interrupts
 * @data:   pointer to irqdata associated to that interrupt
 */
void pci_msi_mask_irq(struct irq_data *data)
{
    msi_set_mask_bit(data, 1);
}
EXPORT_SYMBOL_GPL(pci_msi_mask_irq);

/**
 * pci_msi_unmask_irq - Generic IRQ chip callback to unmask PCI/MSI interrupts
 * @data:   pointer to irqdata associated to that interrupt
 */
void pci_msi_unmask_irq(struct irq_data *data)
{
    msi_set_mask_bit(data, 0);
}
EXPORT_SYMBOL_GPL(pci_msi_unmask_irq);
```

10.6.5.3 MSI-X（三）

1. MSI 和 MSI-X 对比

从 INTx 过渡到 MSI，可以说是完全两套天壤之别的中断上报架构。一个是带外，一个是带内。而 MSI-X 则是以 MSI 为基础发展起来的，很多特性很类似。MSI-X 某种程度上可以看做的 MSI 的一个超集。它和 MSI 一样，可以生成 Memory Write 事务，来向中断控制器报告中断。

但是，MSI-X 定义了一套新的 capability structure，并且在 Memory 空间存储 MSI-X 中断地址表，可以处理更多的中断，也更灵活，在软件处理上也无法兼容 MSI。

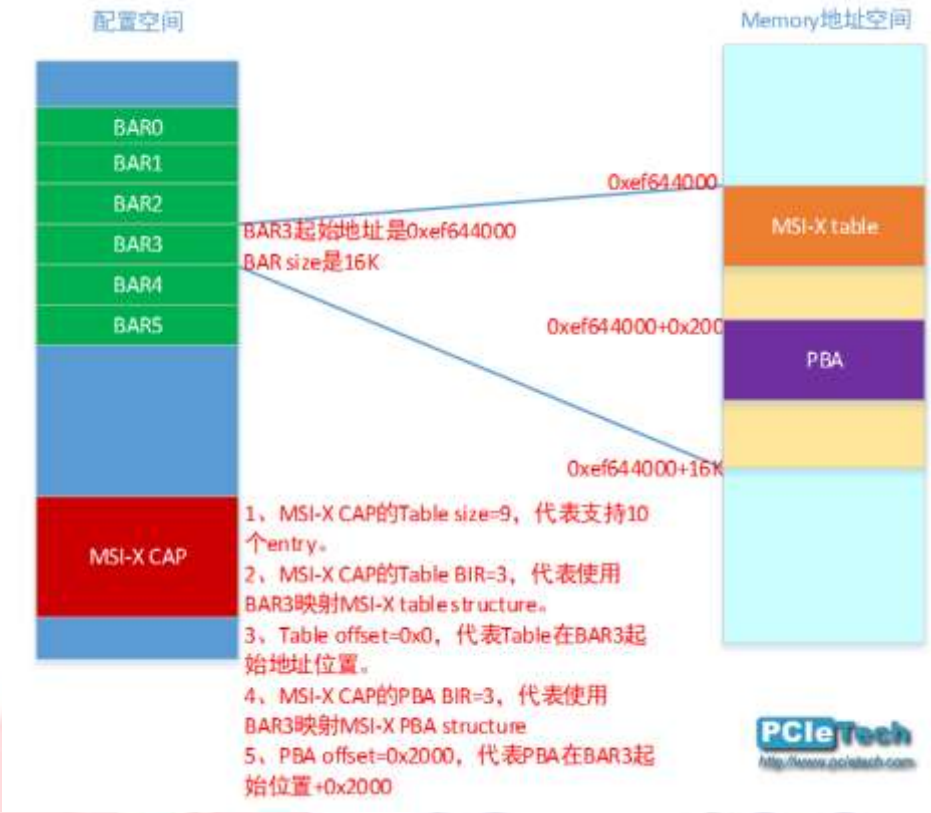
下面简单列表对比一下两者的差异：

对比项	MSI	MSI-X
Message Address	存放在 MSI 相关配置空间	存放在 BAR 空间 MSI-X table structure
Message Data	存放在 MSI 相关配置空间	存放在 BAR 空间 MSI-X table structure
Sataus 相关	存放在 MSI 相关配置空间	存放在 BAR 空间 PBA structure
每个设备支持的 Vector 数量	32	2048
中断号连续？	是	否

- 举个例子

说了这么多拿网卡举个例子吧。我们的 I350 网卡位于 bus 3，device0，function 0。从配置空间可以看出网卡申请了一个 BAR3，这正是 MSI-X 所使用的 BAR3，MSI-X table structure

存放在 BAR3 起始地址+0 的位置，PBA structure 存在 BAR3 起始地址+0x2000 的位置。



我们来读一下该地址，发现使用的 entry 的 message 地址为 LAPIC 的地址。

10.6.6 SR-IOV 浅谈

10.6.6.1 SR-IOV（一）

SR-IOV 是一项由 PCI-SIG 组织定义的规范，这个规范的完整名字叫《Single Root I/O Virtualization and Sharing Specification》，可以在 SIG 网站下载到。SR 的意思是 Single Root（单个根联合体，或者换句话讲，就是单个 CPU，一个领导容易理解）。IOV 即 I/O virtualization，也就是 I/O 虚拟化。

当前，虚拟化技术是非常火热的一个技术方向，包括的内容也非常的多。

维基上对于虚拟化是这样定义的：

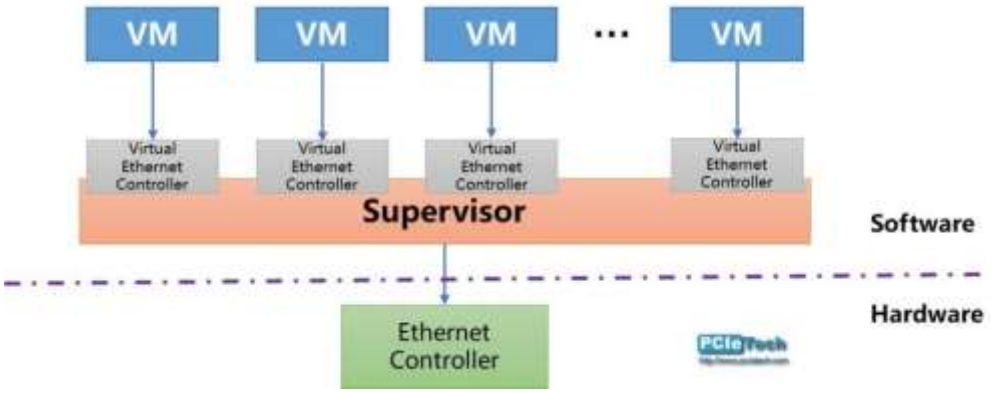
In computing, virtualization refers to the act of creating a virtual (rather than actual) version of something, including virtual computer hardware platforms, storage devices, and computer network resources。

虚拟化技术又可分为基于硬件的虚拟化和基于软件的虚拟化。我们这里主要讨论的是基于硬件的虚拟化技术，特别是针对 PCIe 设备的硬件虚拟化技术。SR-IOV 是这些虚拟化技术中的一种。

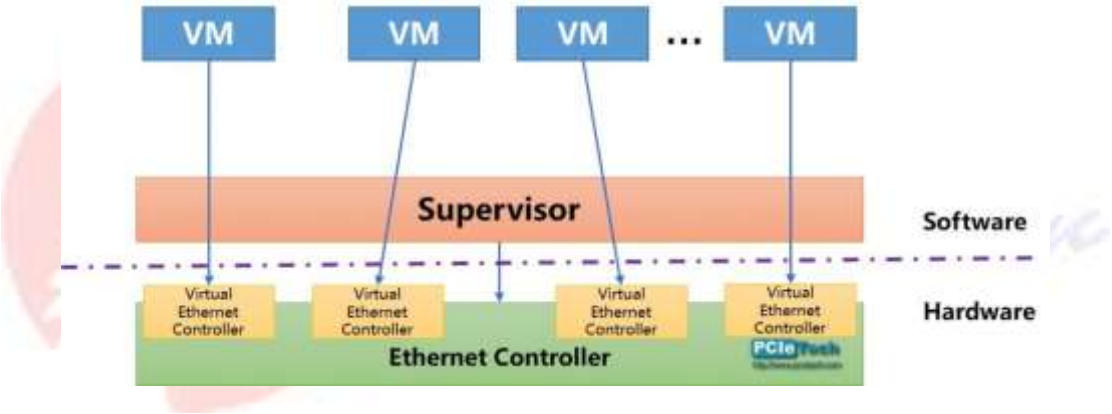
虚拟化技术中心需要解决的几个关键问题，一是安全，如果能够做到物理隔离是最好的。另一个是性能，软件模拟的终究会有天生的性能弱点，受制于 CPU 处理性能。

传统的虚拟化系统中，软件实现了一个管理硬件和虚拟机之间的中间层。通常称之为 Supervisor。它提供了硬件和虚拟机之间的接口，负责安全性，并确保虚拟机之间的隔

离和安全。Supervisor 必须通过为每个虚拟机模拟（实例化）一个虚拟的以太网控制器设备并支持 VM 来访问。显然，这样的系统很容易造成性能瓶颈。



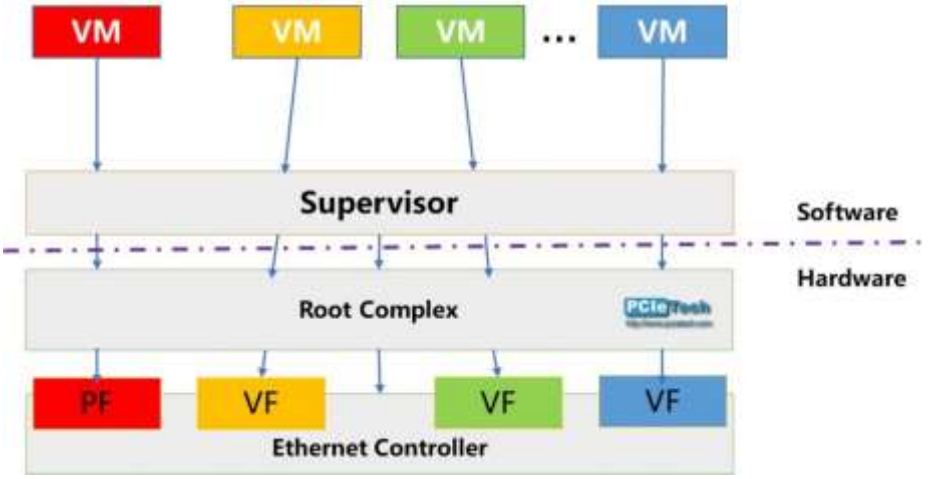
而 PCI SIG 定义的 SR-IOV，则是跟传统虚拟化完全不同的玩法。是一种不需要软件模拟就可以共享 PCIe I/O 设备的物理功能（PF: Physic Function）的方法。SR-IOV 规范定义了可以创建 PCIe 设备物理端口的虚拟功能（VF: Virtual Function）。每个虚拟功能都可以被直接分配给一个虚拟机使用,每个虚拟机都可访问唯一的硬件资源。因此，SR-IOV 实现了设备的单独分配和使用，且实现了虚拟机直接访问硬件，性能也得到极大的提升。



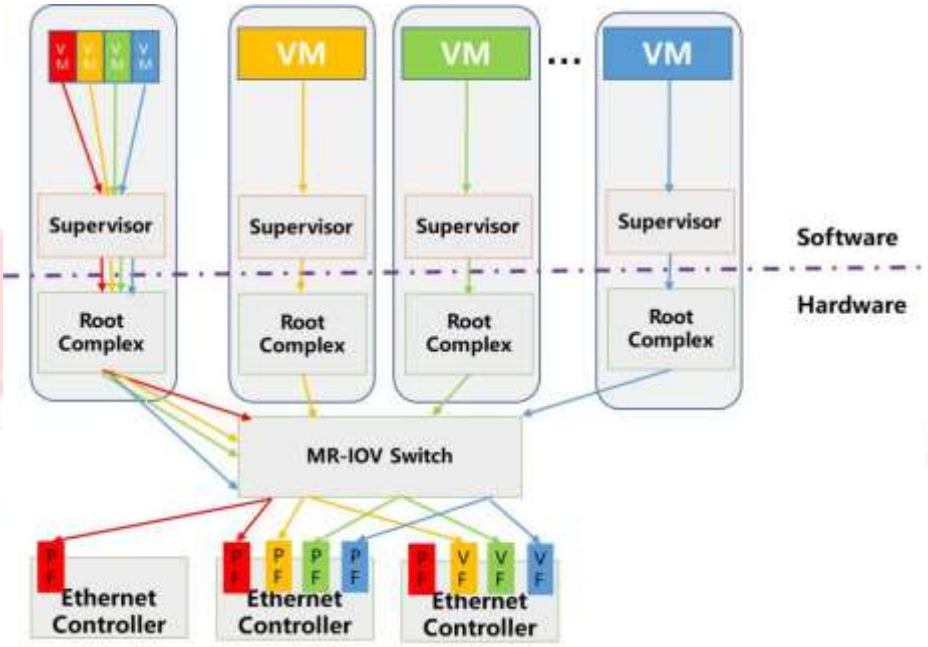
使用带有 SR-IOV 功能的设备，可以在硬件中实现虚拟设备。使用 SR-IOV 而不是更传统的网络虚拟化的好处是，在 SR-IOV 虚拟化中，VM 通过直接内存访问（DMA）直接与硬件网络适配器通信。可以实现最佳的系统性能，提供接近“裸机”的性能。

大概了解 IOV 之后，我们再回头来看看 SR 的定义，SR 就是 Single Root，单根节点。PCI-SIG 在定义 SR-IOV 之后，同样定义了另一套规范，叫 MR-IOV（Multi-Root I/OVirtualization and SharingSpecification）。之所以叫 MR，是为了和 SR 区分，因为 MR-IOV 支持多个根节点。

尝试简单对比一下 SR-IOV 和 MR-IOV 的逻辑框图。如下：



SR-IOV



MR-IOV

很明显，SR-IOV 只支持一个 Root Complex，而 MR-IOV 支持多个 Root Complex，并且，需要特殊的支持 MR-IOV 的 Switch 才能够组网。MR-IOV 的各种组合更加的复杂和多变。但归根结底，都是有 SR-IOV 演变的。因此，首先需要了解清楚 SR-IOV。

10.6.6.2 SR-IOV（二）

SR-IOV 的 spec（《Single Root I/O Virtualization and Sharing Specification》）不仅对于 SR-IOV 的细节定义的很清楚，而且从架构上讲解了 SR-IOV 的演进过程。了解历史，更有助于我们理解技术细节。

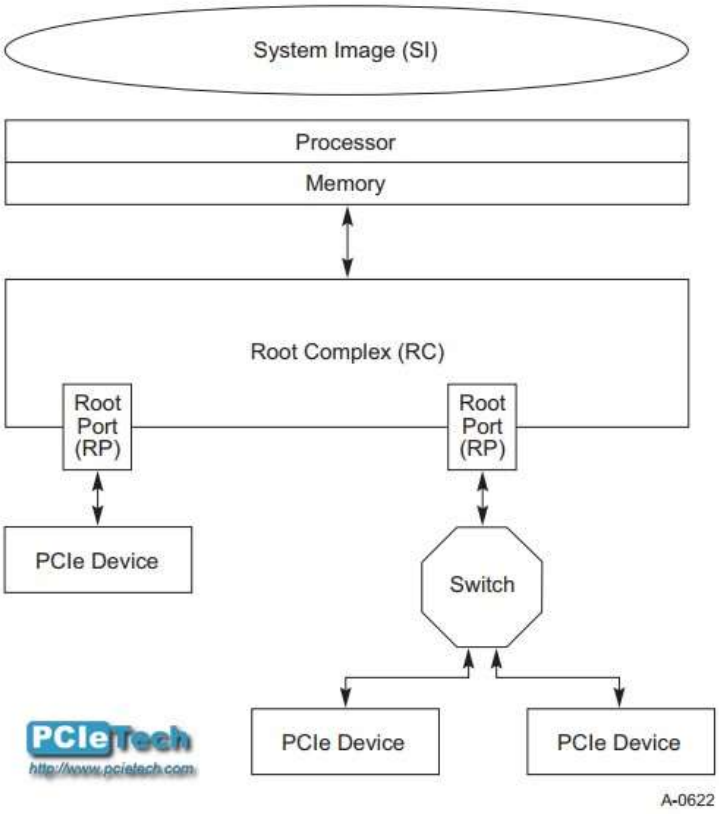


Figure 1-1: Generic Platform Configuration

上图是一个典型的早期系统架构图。系统中只有一个根节点（Root Complex）用于扩展连接外部的设备。这些连接走的是 PCIe 协议，当然，RC 可以直接连接外设（EP），也可以通过 PCIe Switch 扩展。（这部分内容是 PCIe 的基础协议，不了解的可以参考前期的文章《好大一棵树 – PCIe Tree》以及《从 PCI 角度认识 PCIe》，这里的 SI（System Image）相当于操作系统。

后面为了提高有效硬件资源利用率，引入了虚拟化技术，在不修改、增加硬件的前提下，软件模拟了硬件的行为。出现了一个软件的中间层 Virtualization Intermediary (VI)，也就是虚拟化中间层。VI 提供了对于底层硬件的抽象和访问控制，提供了某个硬件设备实例的虚拟化实例。因此，可以在 VI 之上安装更多的操作系统镜像或者虚拟机。

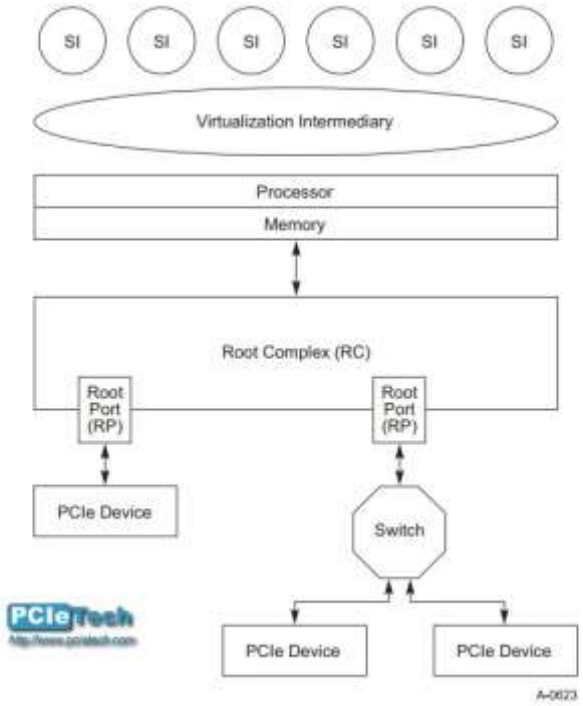


Figure 1-2: Generic Platform Configuration with a VI and Multiple SI

如前一篇所述，这种情况下，所有的 IO 访问都要通过软件实现的 VI 统一调度和统一转发。因此，性能损失极大。在这种背景下，SR-IOV 应运而生：

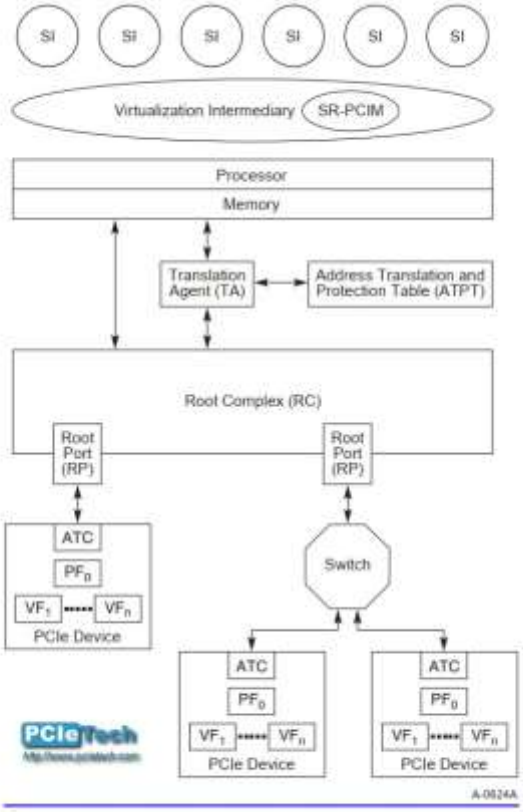


Figure 1-3: Generic Platform Configuration with SR-IOV and IOV Enablers

首先我们注意到：传统的 PCIe Device（EP）拥有了两中类别的功能（Function）：PF（Physic Function）和 VF（Virtual Function）。。其中 PF（图中的 PF0）跟以前的 Spec

定义的 PF 一样，不同的是，支持 SR-IOV 的 PF 可以被系统配置为支持多个 VF。当使能 SR-IOV 后，这个 PF 会生成多个 VF（图中的 VF1…VFn）。这些 VF 是可以被不同的操作系统（SI）来访问到的。对于 SI 或者应用程序来讲，访问这些 VF 和 PF 的方法一致，没有任何差异。注意，这个图中只显示了一个 PF0，实际上，支持 SR-IOV 的 PCIe Device 是可以有多个 PF 存在的。

其次，为了支持能够配置 PF、管理 PF、VF 等，软件方面需要对应增加，这个部分叫：SR-PCIM（Single Root PCI Manager）。可以理解为系统的 SR-IOV 管理驱动。

其他几个如：Address Translation and Protection Table（ATPT）、Translation Agent（TA）、Address Translation Cache（ATC）。这些都是有关 PCIe 设备地址翻译的问题，所谓地址翻译，是指存储地址空间到 PCIe 外设设备地址空间的转换。这部分内容不是我们当前内容考虑的重点，有兴趣的可以看看 ATS 的 Spec（Address Translation Services Specification）。

10.6.6.3 SR-IOV（三）

Physical Function (PF)

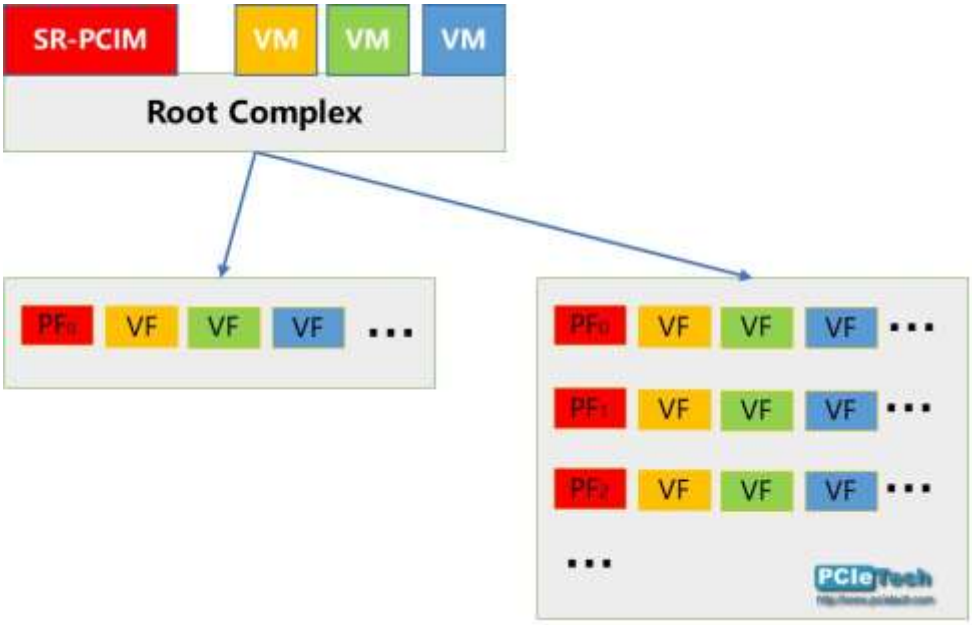
- 简而言之，这是基本规范定义的 PCI 功能（Function）的名称。
- 具有 SR-IOV 能力的 PF 负责用于配置 IOV 功能以及管理 PF 功能。
- 这些 PF 由上一篇文章提到的 SR PCIM 用于管理虚拟功能 VF。

Virtual Function (VF)

- 简而言之，就是虚拟化后的功能的名称。
- VF 有自己的 Bus、device 和 Function 号和自己的 BAR 地址。
- SI 可以直接访问这些功能上的资源。
- VF 由 SR-PCIM 创建/管理。
- 一个 VF 只与一个 PF 相关联。
- VF 创建后，可以通过 RC（Root Complex）按照传统 PCIe 方法进行配置和访问。

Physical Function 被 SR-PCIM 来管理配置虚拟化 VF 功能。

Physical Function 0 (PF0) 通常还负责设备的物理层的行为和属性，如物理层错误以及 link 事件响应等。



SR-IOV 规范对于支持 SR-IOV 的 Phyesical Function 增加了一个新的能力寄存器（SR-IOV Capability）用于管理、配置、使能 SR-IOV 能力。例如，下图截取了 Intel 82599 的相关 SR-IOV 能力寄存器的定义：

9.4.4 IOV Capability Structure

This is the new structure used to support the IOV capabilities reporting and control. The following tables shows the possible implementations of this structure in the 82599.

Byte Offset	Byte 3	Byte 2	Byte 1	Byte 0
0x150	Next Capability Ptr. (0x160)	Version (0x1)	Capability ID (0x000E)	
0x154	Control Register		Capabilities	
0x160	Next Capability Offset (0x0)	Version (0x1)	IOV Capability ID (0x0010)	
0x164	SR IOV Capabilities			
0x168	SR IOV Status		SR IOV Control	
0x16C	Total VFs (RO)		Initial VF (RO)	
0x170	Reserved	Function Dependency Link (RO)	Num VF (RW)	
0x174	VF Stride (RO)		First VF Offset (RO)	
0x178	VF Device ID		Reserved	

下图是 Spec 定义的 SR-IOV 能力寄存器的完整定义：

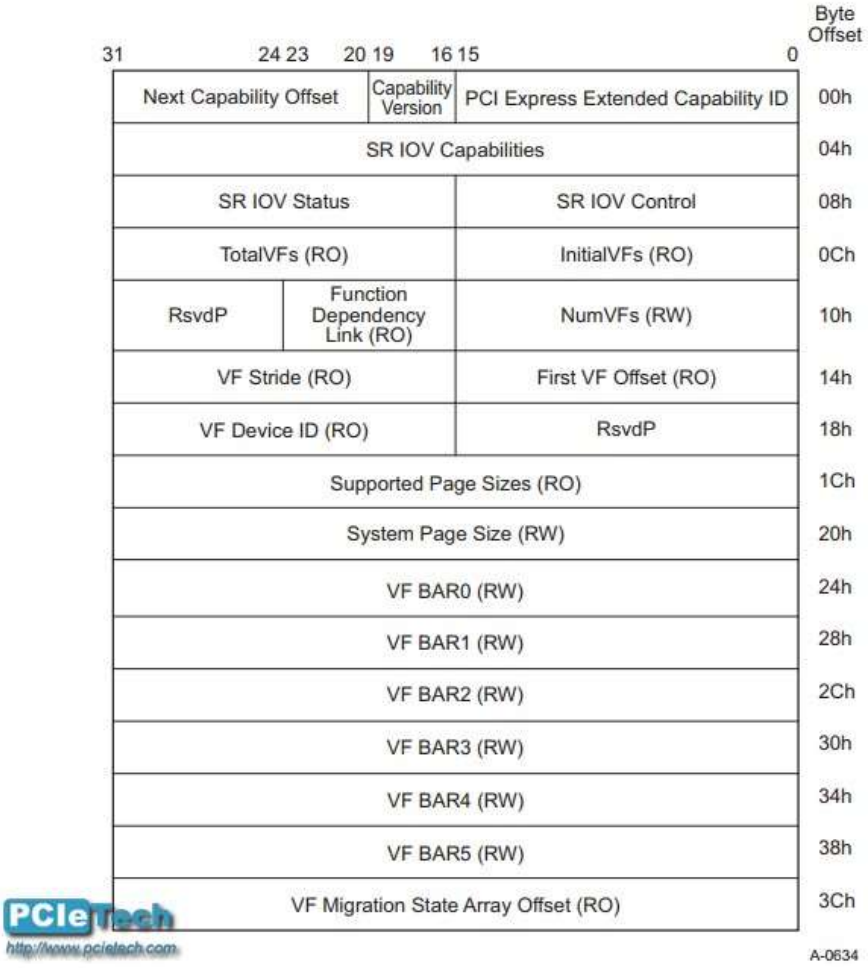


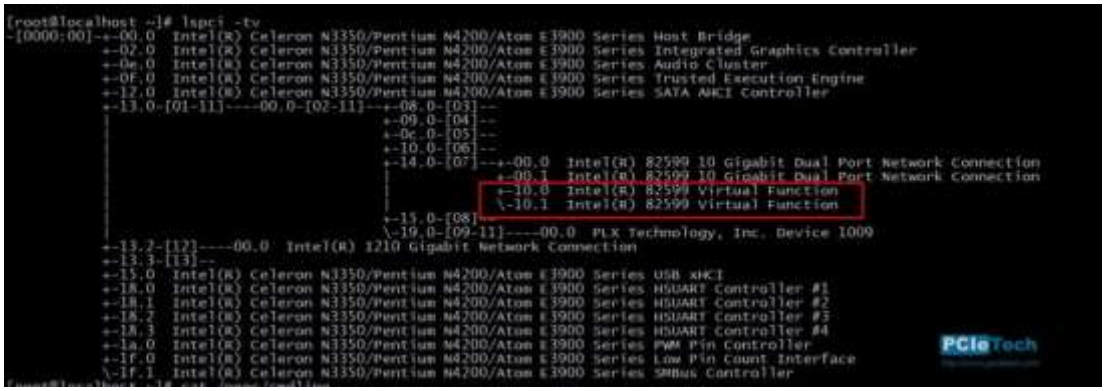
Figure 3-1: Single Root I/O Virtualization Extended Capabilities

在 Linux 下，内核是支持 SR-IOV 特性的。每个支持 SR-IOV 的设备厂商的驱动往往也是两个，一个是基本 PF 驱动，一个是虚拟 VF 驱动。比如 Intel 的网卡驱动包括：ixgbe-5.3.7.tar.gz, ixgbevf-4.3.5.tar.gz。如果需要使用 VF，需要加载 ixgbevf 驱动。

举个例子，Intel 82599 有两个原生的 PF，我们如果需要在一个 PF 下创建一个 VF，则加载驱动方式如下：

modprobe ixgbe max_vfs=1,1

如果想在两个 PF 上分别创建 3 个 VF，4 个 VF，参数修改为 max_vfs=3,4



打开日志。我们可以看到：

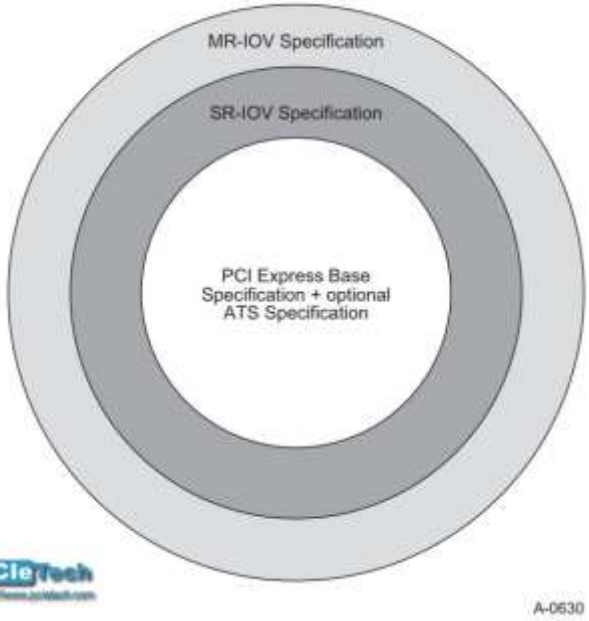
```
[root@localhost ~]# modprobe ixgbe max_vfs=1,1
[root@localhost ~]# dmesg
[ 128.279468] Intel(R) 10GbE PCI Express Linux Network Driver - version 5.3.7
[ 128.279475] Copyright(c) 1999 - 2018 Intel Corporation.
[ 128.301008] ixgbe: I/O Virtualization (IOV) set to 1
[ 128.301018] ixgbe: 0000:07:00.0: ixgbe_check_options: FCoE Offload feature enabled
[ 129.368851] ixgbe 0000:07:00.0: Enabling SR-IOV VFs using the max_vfs module parameter is deprecated.
[ 129.368861] ixgbe 0000:07:00.0: Please use the pci sysfs interface instead. Ex:
[ 129.368871] ixgbe 0000:07:00.0: echo '1' > /sys/bus/pci/devices/0000:07:00.0/sriov_numvfs
[ 129.469727] pci 0000:07:10.0: [8086:10ed] type 00 class 0x020000
[ 129.469845] pci 0000:07:10.0: can't set Max Payload Size to 256; if necessary, use "pci=pcie_bus_safe" and report a bug
[ 129.470141] iommu: Adding device 0000:07:10.0 to group 5
[ 129.470331] ixgbe 0000:07:00.0: SR-IOV enabled with 1 VFs
[ 129.470339] ixgbe 0000:07:00.0: configure port vlans to keep your VFs secure
[ 129.470406] ixgbe 0000:07:00.0: FCoE offload feature is not available. Disabling FCoE offload feature
```

上面驱动成功对每一个 PF 都创建、使能了一个 VF。

```
[ 129.508940] ixgbevf 0000:07:10.0: enabling device (0000 -> 0002)
[ 129.564099] ixgbevf 0000:07:10.0: PF still in reset state. Is the PF interface up?
[ 129.564111] ixgbevf 0000:07:10.0: Assigning random MAC address
[ 129.564220] ixgbevf 0000:07:10.0: irq 148 for MSI/MSI-X
[ 129.564241] ixgbevf 0000:07:10.0: irq 149 for MSI/MSI-X
[ 129.564286] ixgbevf 0000:07:10.0: Multiqueue Disabled: Rx Queue count = 1, Tx Queue count = 1
[ 129.565015] ixgbevf 0000:07:10.0: 92:31:0c:eb:c5:99
[ 129.565022] ixgbevf: eth1: ixgbevf_probe: Intel(R) 82599 Virtual Function
[ 129.565026] ixgbevf: eth1: ixgbevf_probe: GRO is enabled
[ 129.565030] ixgbevf: eth1: ixgbevf_probe: Intel(R) 10GbE PCI Express Virtual Function Driver
```

而 VF 驱动访问这些 VF 的方式，包括资源访问、MSI 中断分配等等都跟 PF 无任何差异。

最后，对于 SR-IOV 协议来讲。SR-IOV 规范的核心是 PCI Express 基本规范。 所有 IOV 实现必须符合 PCI Express 基本规范。SR-IOV 规范在基于 PCIe base spec 的基础上做了增加和补充。而 MR-IOV 规范是在 SR-IOV 规范基础上增加了新的支持 Multi RootComplex 的内容。换句话说，支持 MR-IOV 的设备一定是支持 SR-IOV 的。



另外，对于 CPU 支持虚拟化的规范，不是 PCI-SIG 的责任范畴，不同架构体系的 CPU 的实现方式和规范都不一致。

10.6.7 PCIe 热插拔

10.6.7.1 HOT-PLUG（一）

从今天起，我们讨论一下有关于“热插拔”的话题。

为什么要有“热插拔”？

从历史上到目前，IT 系统设备（PC、服务器、存储等等）中，对于 RAS（Reliability, Availability, Serviceability，也就是：可靠性、可用性和可维护性）的要求变得越来越高。这也就是我们所谓的 DFX 设计相关内容。

系统怎么样才更可靠，可用性才更高？这是有专门的可靠性设计的，涉及太多的领域，我们不展开。常规的看，对于系统可能出现的问题和故障，需要做：故障检测、故障隔离、故障告警、故障恢复等等。

故障的检测可能会使用检测链路状况、器件状态等等，在 PCIe 领域，常见的是检测链路协商宽度、速率，检测 AER 寄存器等等。

故障的隔离主要的目的是让故障或即将故障失效的器件、部件从系统中隔离出去。比如把失效的 PCIe 卡从系统中移除，不再接管业务，避免故障扩散。

故障恢复的手段就更多样化了，有复位修复、上下电修复、备份冗余、故障时切换备机等等。比如把失效的 PCIe 设备重新上下电做修复，或者把业务切换到备用的卡上。或者更换故障 PCIe 卡，也就是我们常说的更换 FRU（Field Replace Unit 现场可更换单元）

抱歉，做了比较长一段时间的可靠性工程师工作，貌似扯得有点远了。：)

网上很多的文章和书籍讲热插拔，都是讲的标准的热插拔，主要目的是为了现场快速更换 PCIe 设备。热插拔的基本目的是要让 PCIe 设备按照规定的顺序、原则，从系统中移除或插入到系统中来，并能正常的工作，且不影响系统的正常运行。**事实上，PCIe“热插拔”的关键目的就是为前面面所提到的系统 RAS 服务的，是提升系统 RAS 能力的非常重要的手段！**

热插拔有三个重要的功能：

1. 在线替换发生故障的 PCIe 设备。不需要关闭、重启系统。
2. 热插拔器件，系统及其他功能服务继续运行，不受影响。
3. 热插拔 PCIe 设备的相关驱动/软件自动加载/卸载。

Spec 中对于热插拔是这样定义的：

Hot-Plug
Insertion and/or removal of a card into an active backplane or system board as defined in PCI Standard Hot-Plug Controller and Subsystem Specification, Revision. 1.0. No special card support is required.
Hot swap
Insertion and/or removal of a card into a passive backplane. The card must satisfy specific requirements to support Hot swap.

注意，这里有两个相关概念，前面我们讲的热插拔，其实是广义的“热插拔”。这里的 Hot-Plug，为了方便，我们叫做 PCIe 热插拔，也就是我们现在及后面将要讨论的内容。

Hot Swap，即热交换，也叫热切换，主要指的是 CPCI（Compact PCI，紧凑型 PCI）领域所使用的。关于 Hot Swap，CPCI 有专门定义的一套规范，叫《Compact PCI® Hot Swap Specification》，也有中文版本，有兴趣的可以自行研究。

如无特别的指出，我们后续所说的热插拔都是指的是 PCIe Hot-Plug。

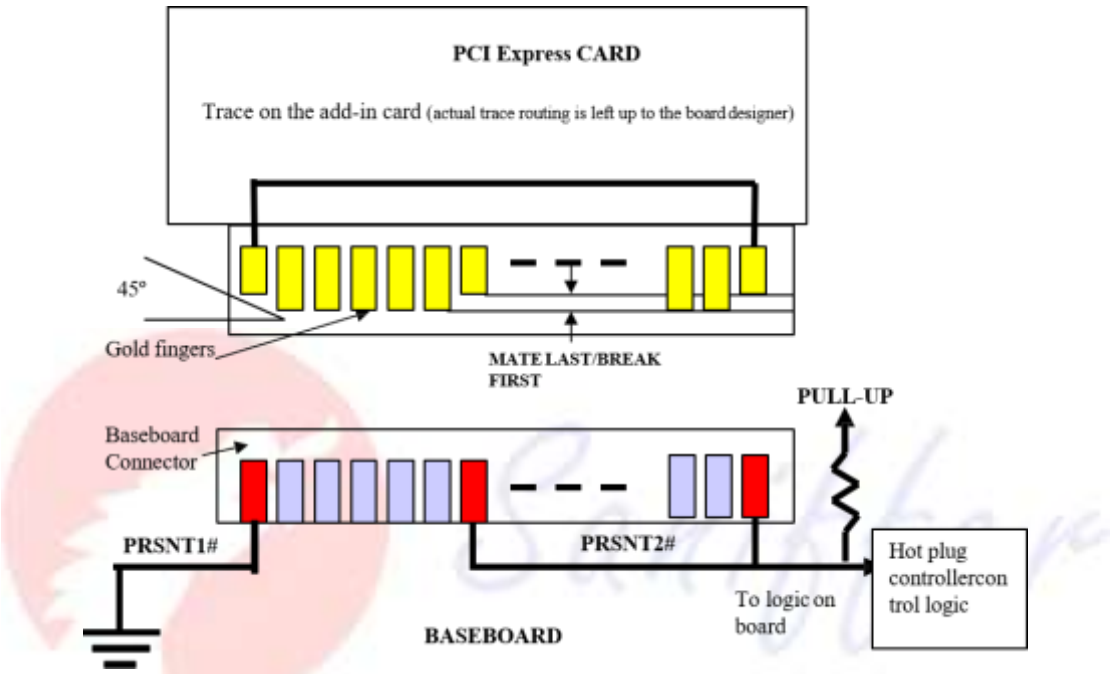
顺便说一下：在 PCI 的年代，就已经提出了 PCI 的相关热插拔规范。2001 年，PCI 定义了 PCI 标准热插拔控制器 (SHPC：PCI Standard Hot-plug Controller and Subsystem Specification) 规范，PCIe 沿用了这套规范。

10.6.7.2 HOT-PLUG（二）

上一节讲到，Spec 定义的热插拔是把一个 PCIe 卡（设备）从一个正在运行的背板或者系统中插入/或者移除。这个过程需要不影响系统的其他功能。插入的新的设备可以正确工作。

显然，这里面需要考虑的问题有硬件和软件两方面的事情。

硬件上看，很显然，一个新的设备插入系统，肯定是需要硬件上支持识别到这个插入动作的。因此，Spec 定义了一个在位（Present）的 pin 脚，硬件上用作判断卡是否插入。卡插入时，这个 pin 被拉低。当然，因为 PCIe 金手指的长度较长，插入卡时有可能前后高低差异。因此，需要有前后多个 present pin 来确保金手指完全插入。



需要注意的是，检测卡是否在位，除了使用 Present Pin 之外，也可以通过链路的负载检测等来完成。也就是所谓带内检测在位。或者一些特殊的场景比如 NT-NT 背靠背，就没有 Present 信号。NT-NT 的插入时，链路自动重新 Link。

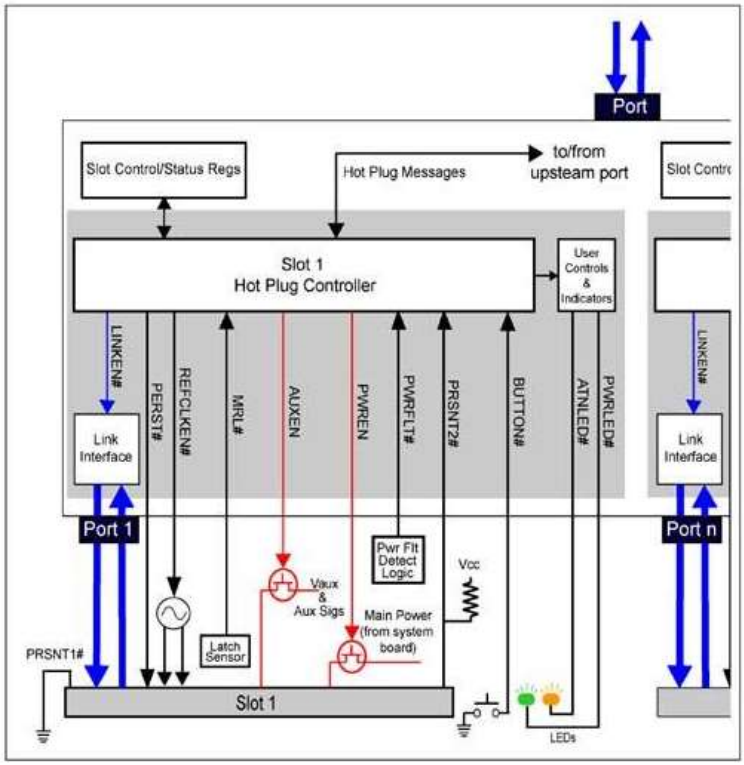
除了在位，硬件上还需要实现的是：电源控制（不用多讲，这个是基础），复位逻辑（可控的给出 PERST 信号复位新插入的设备）。另外，Spec 定义的热插拔是 graceful 的，也就是可控的、优雅的，不是突然地、暴力的。因此，热插拔的控制逻辑，需要一个按钮，来告诉系统我需要插入/拔出了。另外，人机界面上，一个 LED 指示灯来提示用户热插拔的工作状态也是蛮友好的事情。

综上，热插拔所需要的的硬件基本元素如下：

Element	Purpose
Power Indicator	绿色的 LED 灯，表示电源 On/Off
Attention Indicator	黄色 LED 等。热插拔过程的一些状态表示。亮或是闪烁。
MRL	Manually-operated Retention Latch 锁止开关，用户固定卡。

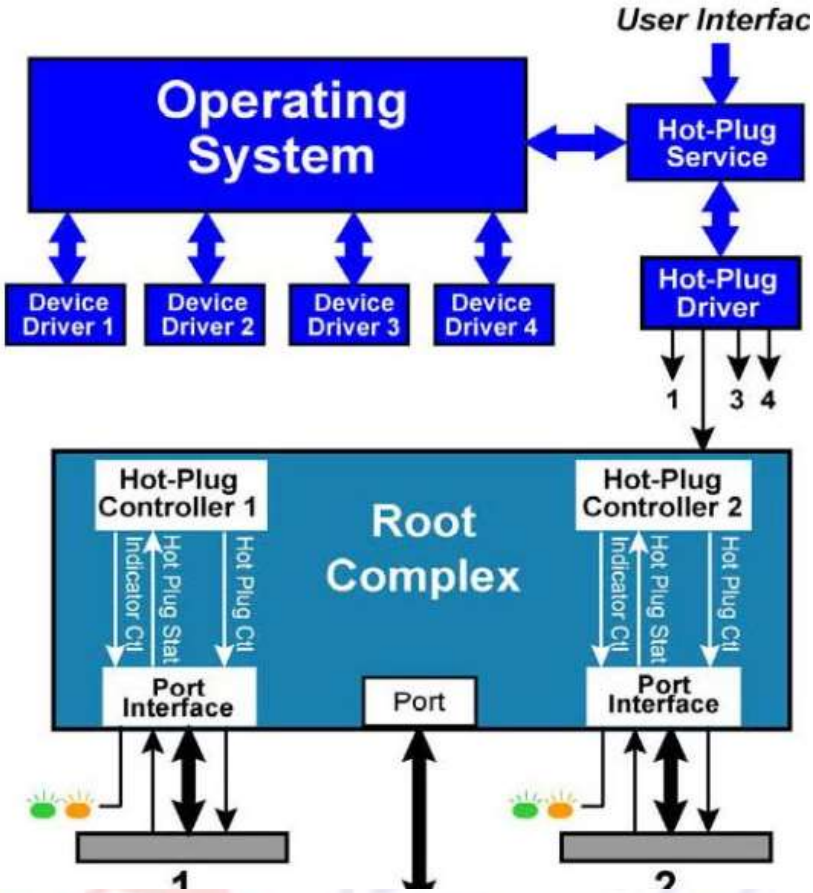
MRL Sensor	检测锁止开关是否打开的光传感器。协议定义的 MRL 和 MRL 传感器。但实际应用基本没什么用。反正我没见过。属于可选的。
Electromechanical Interlock	机电联锁。一种确保不能随意移除卡的机制。同上 MRL，我没见过。属于可选的。
Attention Button	按钮。用于告诉系统开始/取消 热插拔过程。
Slot numbering	热插拔槽位标号，这个取决于每个公司的自定义。
Power Controller	上下电控制部分。包括电源故障检测。

上面是 Spec 定义的一些元素，实际应用中，可以根据自己的实际项目需求来取舍。下图是一个支持热插拔的 PCIe Switch 的典型信号示意图：



软件方面的元素，其实简单来说就两个：

1. 热插拔驱动：系统的驱动程序，支持 PCIe 的热插拔过程、移除拔出卡的资源、重新分配插入卡的资源、控制上下电等等。
2. 支持热插拔的设备驱动：主要是需要支持插入时的初始化和移除时的资源释放。典型的是 linux 设备驱动的 probe/remove。



10.6.7.3 HOT-PLUG（三）

对于热插拔，Spec 对于所谓插槽（插卡）的 ON、OFF 状态都做了比较清晰的定义，如下：

ON（上电）状态的属性：

• Power is applied to the slot.

• REFCLK is on.

• The link is active or in the standby (LOs or Ll) low power state due to Active State Power Management.

• The PERST# signal is deasserted.

简而言之：电源打开、时钟开启、链路 link、复位信号拉高。

OFF（上电）状态的属性：

• Power to the slot is turned off.

• REFCLK is off.

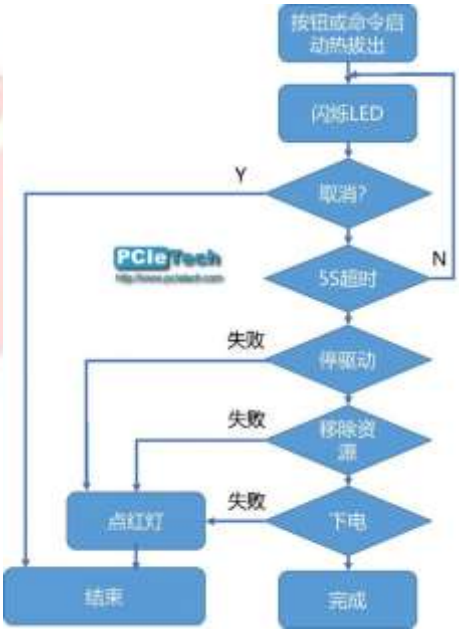
• The link is inactive. (Driver at the root of switch port is in Hi Z state)

• The PERST# signal is asserted.

简而言之：电源关闭、时钟关闭、链路断开、复位信号拉低生效。

热插拔的插入、拔出过程大致如下流程。先看卡的拔出过程：（注意，我们这里的插拔不是暴力的拔出，所谓暴力，是指不通知系统和软件，直接暴力的、突然的把 PCIe 卡从系统槽位中拔出！Spec 定义的热插拔是有严格的顺序要求和处理过程的）

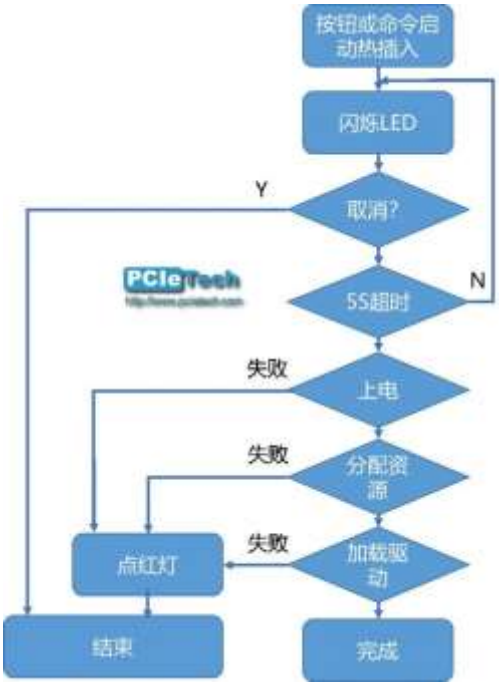
卡拔出流程如下：



1. 按 Attention Button 按钮或者通过软件的命令。通知热插拔控制器开始进行热拔出操作。
2. 闪烁 LED 灯，表示这张卡准备从系统中移除。五秒内没有其他异常，流程继续。五秒内可以取消此次操作。
3. 系统把这个 PCIe 设备在系统中拥有的资源删除，如总线资源、Memory 资源等。确保再没有流程访问这种卡。
4. 关闭这个卡所在槽位的电源。关闭 LED 灯，表示该卡可以从系统中拔出了。
5. （可选操作）打开 MRL。
6. 用户移除这张卡。

如果过程中有错误，则跳出移除过程，可以点 LED 红灯表示错误。需要人工干预处理。

卡插入流程基本与上述流程相反：



1. PCIe 插插入槽位，触发硬件检测到 Present 在位。
2. （可选操作）关闭 MRL。
3. 按 Attention Button 按钮或者通过软件命令，通知热插拔控制器开始进行热插入操作。
4. 闪烁 LED 灯，表示这张卡准备在这个槽位插入。五秒内可以取消此次操作。否则继续。
5. 打开这个槽位的电源。
6. 时钟打开、复位信号生效。插入的 PCIe 设备硬件可以正常工作，链路 link up。
7. 系统重新枚举插入槽位的总线，分配相应的资源。系统可以正常识别新插入的 PCIe 设备。
8. 插入 PCIe 设备对应的驱动进行相关的配置和初始化。PCIe 设备对应功能初始化完成，可以正常工作。

如果过程中有错误，则跳出插入过程，可以点 LED 灯表示错误。

10.6.7.4 HOT-PLUG（四）

Spec 为标准热插拔控制器定义了一套标准的寄存器组，即 PCI Express Capbilities 结构里面包含的内容。PCI Express Capbilities 结构的 Cap ID 为 0x10，可以容易的根据 ID 找到这个结构。这个结构里面包含 Device Capability、Device Control/Status、Link Control/Status、Slot Capbility/Control/Status 等多个寄存器。

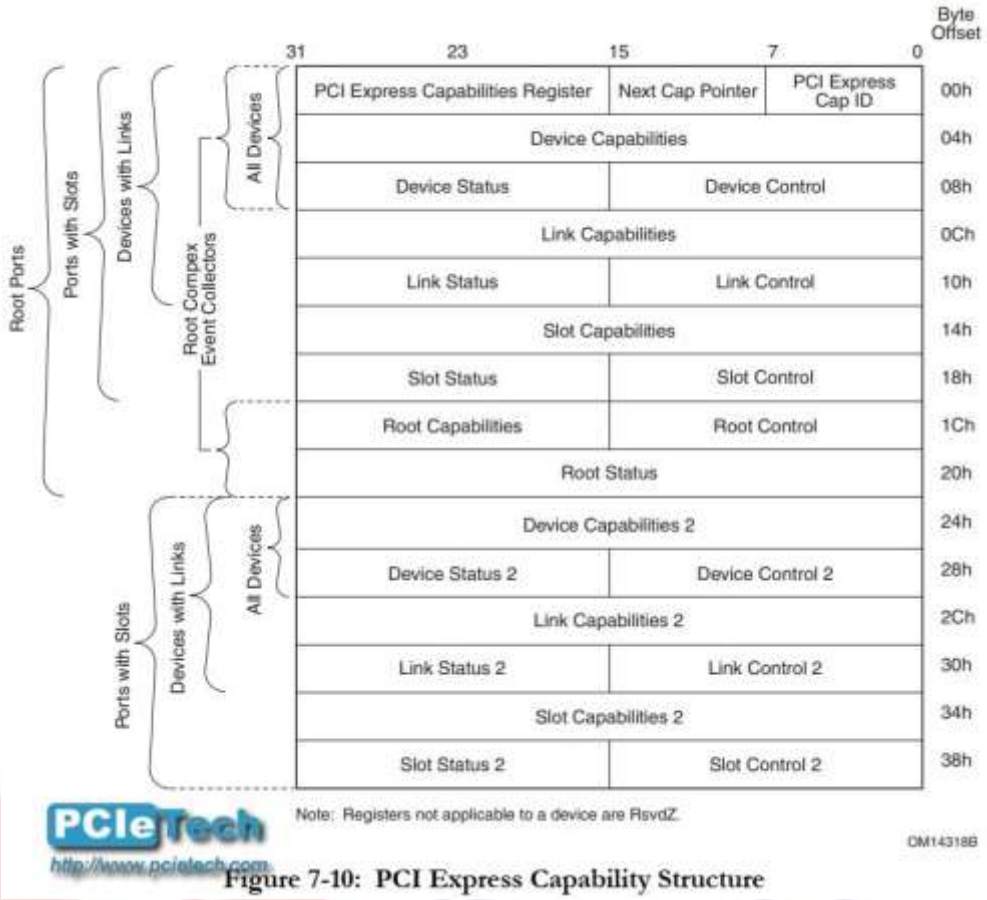
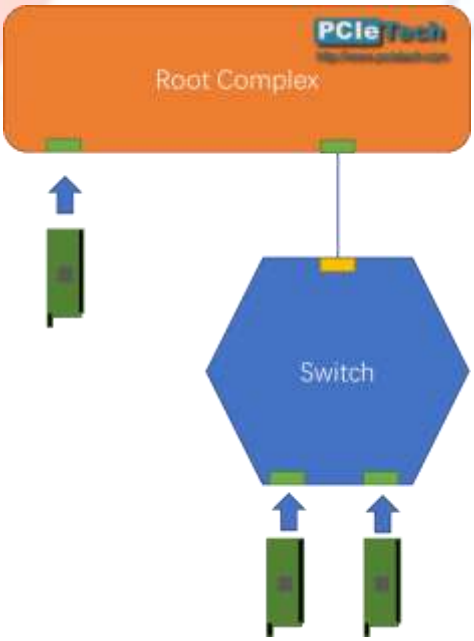


Figure 7-10: PCI Express Capability Structure

不是每个 PCIe 设备都完全具备这个图里的所有寄存器的。我们关注的热插拔，其实主要是 Root Complex 和 PCIe Switch 的下行端口是有热插拔能力的。



在这些寄存器中，对于热插拔而言，我们重点关注 Slot 相关的几个寄存器：Slot Capabilities、Slot Status、Slot Control。在讲这几个寄存器之前，我们首先要了解一下 Spec 定义的关于热插拔可能产生的事件（中断，可以通过设置 Slot Control 寄存器的 bit5Hot-Plug Interrupt Enable 来使能）：

插槽事件：即插槽检测到的一些动作或状态变化

• 按钮按下事件

• 检测到电源失效（Power Fault）

• MRL 传感器变化事件

• 在位检测变化事件

命令完成事件：设置 Slot Control 的一些命令，Spec 要求都必须要芯片执行完成后，返回一个命令完成。并且要求这些命令要在 1 秒内完成，否则超时。

数据链路层状态变化事件（Data Link Layer State Changed）这个事件主要告诉驱动软件 Link 是否建立成功。比如，新热插入的 PCIe 设备上电成功，并 Link OK，上报此事件。根据协议，软件在获取到 link 成功后，需要等待 100ms 执行后续的配置等操作。

下面，我们简单看看几个相关的寄存器：

Slot Capabilities：表征了这个槽位的能力，是否支持热插拔。

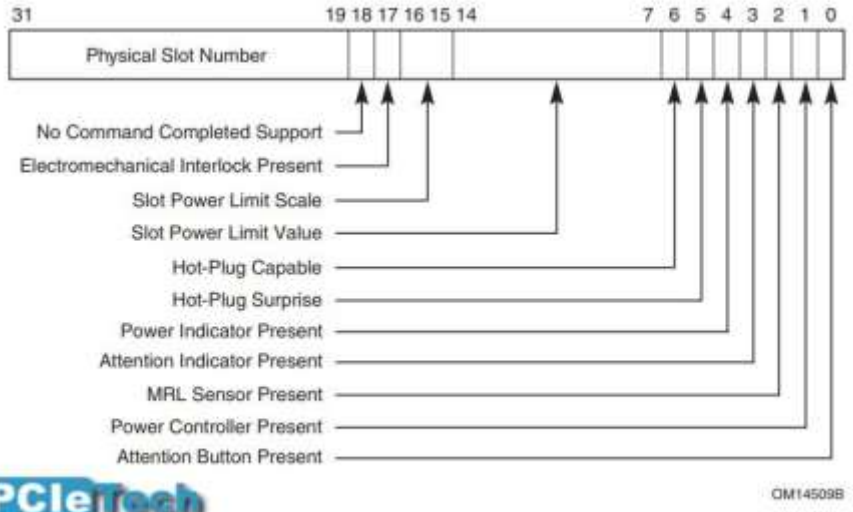


Figure 7-19: Slot Capabilities Register

比如，Bit0 Attention Button Present 表示了这个槽位的按钮是否存在。其他 bit 位根据命名基本上都能清楚的看出来是做什么用的，不再赘述。注意一下 Physical Slot Number 这个字段，这个字段表征当前槽位的槽位号，往往是 PCIe Switch/RC 的内部定义端口号。

Slot control Register：主要控制热插拔功能的寄存器。

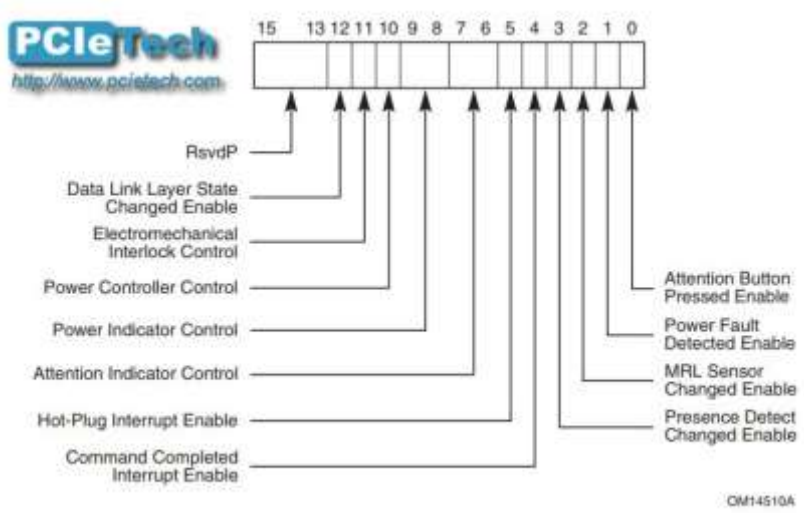


Figure 7-20: Slot Control Register

Slot Status Register：主要表示热插拔状态的寄存器。

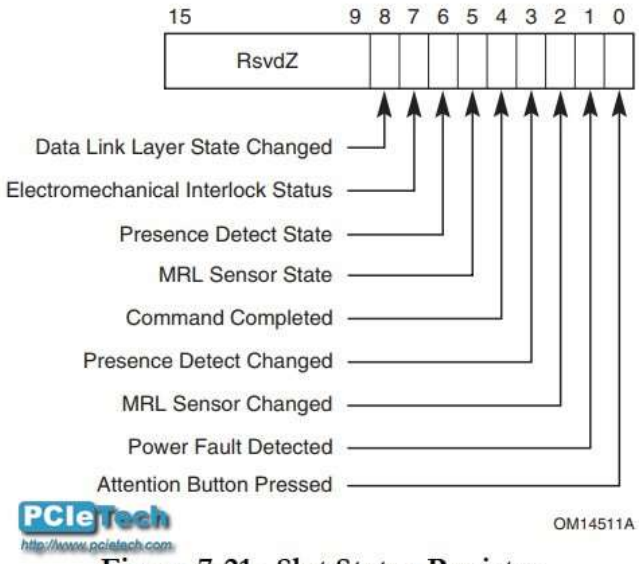


Figure 7-21: Slot Status Register

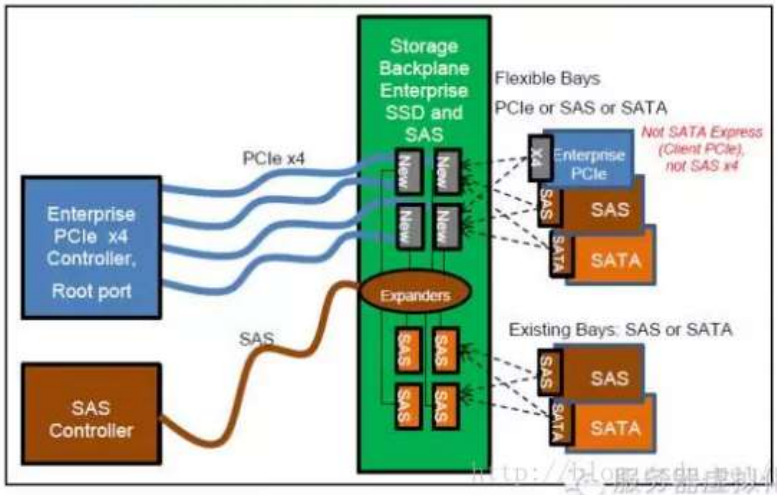
10.6.7.5 从用户的角度理解 NVMe SSD 热插拔时需要注意什么

转载于：https://blog.csdn.net/Memblaze_2011/article/details/52870727

热插拔是大家每天都有可能做的事情。比如，将一块 U 盘从 PC 中拔出，将一个鼠标从一台电脑换到另一台电脑。这些都是再平常不过的事情了。对于数据中心的运维人员来说，更换硬盘也是一件很频繁的事情。

NVMe SSD 已经从实验阶段进入到大量业务部署时期，热插拔这个 feature 变得非常关键。最开始 NVMe SSD 只是以 PCIe 接口的形式出现，跟网卡一样放在背板的卡槽上固定，这种形态的 NVMe 还不适合热插拔。随着 U.2 接口（如下图）的推出，NVMe SSD 可以直接如 SATA/SAS 硬盘一样放置在前面板，此时的 NVMe SSD 对热插拔的支持变得理所当然而且必须。

Figure 40: Backplane for Enterprise SSD (PCIe x4), SAS & SATA



服务器厂商和 SSD 厂商也都非常支持 U.2 接口。如 Dell 的 PowerEdge R730xd 带有 4 个 U.2 盘位，超威甚至发布了 24 盘位的服务器。SSD 方面，Memblaze 等 PCIe SSD 厂商都已经开始推出对应的产品。

Memblaze 的 PBlaze4 系列 PCIe SSD，发布于 2015 年年中

当客户采用 U.2 的 SSD 前，会对其热插拔功能做评估。在实施过程中，由于测试人员对 NVMe SSD 不了解导致系统崩溃的事情时有发生；或者对如何测试 NVMe SSD 热插拔无从下手。那么这篇文章，就跟大家分享在 NVMe SSD 热插拔过程中需要注意的问题（Linux 环境下）。

我们先来了解下 SAS/SATA 和 NVMe 在硬件上的差别。对 SAS 和 SATA 比较熟悉的人知道，SAS 和 SATA 设备通过控制器接入系统（如下图的 SATA）。SAS 和 SATA 设备的热插拔是由其控制器管理的。对于 SAS 来说，以常用的 MegaRAID 为例，其定义了一个热插拔 event，当设备插入或者设备拔出时，MegaRAID 会产生一个 event 并交由 MegaRAID 驱动处理。对于 SATA 而言，AHCI 协议规定了控制器对热插拔的处理流程，并确定控制器必须在热插拔产生时触发一个中断，这样内核的 AHCI 驱动就可以在中断中处理热插拔事件。

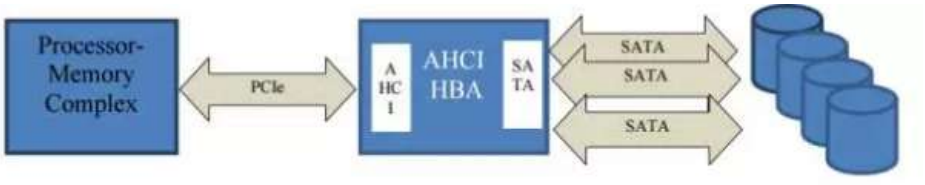


Figure 1 Host Bus Adaptor

NVMe SSD 是不需要控制器的，NVMe 直接连接到通用的 PCIe Bus 上（如下图），跟 SAS/SATA 控制器一个级别。NVMe SSD 热插拔完全依赖于 Host 的 PCIe 处理机制。

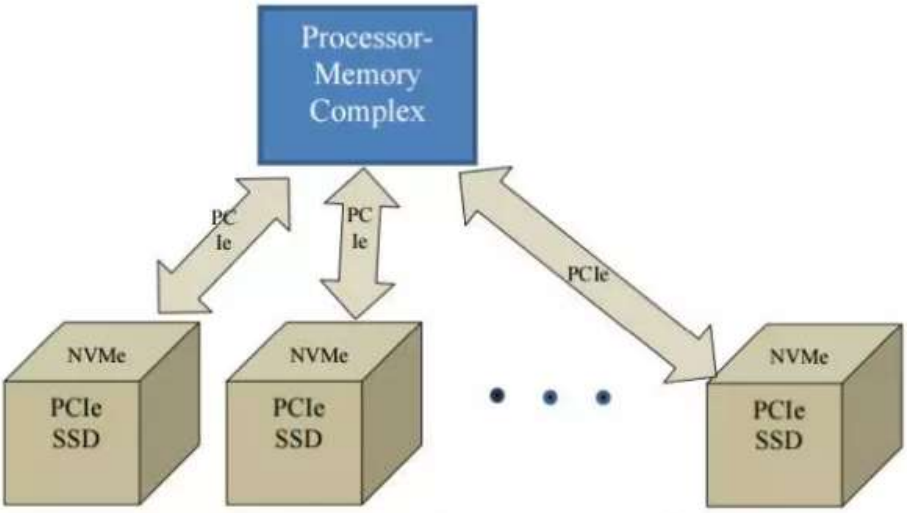


Figure 2 PCIe Attached Solid State Storage

因而，对于使用控制器的 SAS 和 SATA 硬盘来说，可以由控制器厂家进行测试并保证热插拔功能。但是 NVMe SSD 却依赖于 Host 端的 PCIe 的配合，PCIe 由于其通用性，在支持 NVMe SSD 上没有 SAS/SATA 控制器那么完美。对于用户来说，需要认识到 SAS/SATA 和 NVMe 在这方面的不同之处。

省去了控制器的 NVMe 比 SAS/SATA 的热插拔要复杂的多。在进行热插拔测试之前，第一步就是要确认当前的系统是否支持热插拔。

1，确认 SSD 的支持

对于 SSD，热插拔需要保证在插盘的过程中不会产生电流波峰而损坏器件；拔盘的时候，不会因为突然掉电而丢失数据。这个可以向 SSD 供应商确定或者查看产品规格书。

2，确认 PCIe 卡槽的支持

上面提到，NVMe 是直接连接到 PCIe Bus 上的，U.2 接口也是直接跟 PCIe 相连（当判断插入的设备为 NVMe SSD 时）。某些 U.2 接口内部连接的 PCIe 卡槽并不支持热插拔。PCIe Spec 规定了热插拔寄存器。下图（通过 lspci -vvv 获取）显示了一个 PCIe 卡槽的 Capabilities 寄存器信息。其中 LnkSta，SltCap，SltCtl 和 SltSta 4 个部分在热插拔过程中比较有用（具体意义请参考 PCIe Spec）。HotPlug 和 Surprise 是最基础的判断热插拔的标志位。SltSta 中有一个 PresDet 位指示当前是否有 PCIe 设备插入卡槽。

```
Capabilities: [68] Express (v2) Downstream Port (Slot+), MSI 00
        DevCap: MaxPayload 512 bytes, PhantFunc 0, Latency L0s <64ns, L1 <1us
                ExtTag- RBE+ FLReset-
        DevCtl: Report errors: Correctable- Non-Fatal- Fatal- Unsupported-
                RlxdOrd+ ExtTag- PhantFunc- AuxPwr- NoSnoop+
                MaxPayload 128 bytes, MaxReadReq 128 bytes
        DevSta: CorrErr+ UncorrErr+ FatalErr- UnsuppReq+ AuxPwr- TransPend-
        LnkCap: Port #6, Speed 8GT/s, Width x4, ASPM L1, Latency L0 <4us, L1 <4us
                ClockPM- Surprise+ LLActRep+ BwNot+
        LnkCtl: ASPM Disabled; Disabled- Retrain- CommClk-
                ExtSynch- ClockPM- AutWidDis- BWInt- AutBWInt-
        LnkSta: Speed 2.5GT/s, Width x0, TrErr- Train- SlotClk- DLActive- BWMgmt- ABWMgmt-
        SltCap: AttnBtn+ PwrCtrl+ MRL- AttnInd+ PwrInd+ HotPlug+ Surprise+
                Slot #70, PowerLimit 25.000W; Interlock- NoCompl-
        SltCtl: Enable: AttnBtn+ PwrFlt- MRL- PresDet+ CmdCplt+ HPIrq+ LinkChg-
                Control: AttnInd Off, PwrInd On, Power+ Interlock-
        SltSta: Status: AttnBtn- PowerFlt- MRL- CmdCplt- PresDet- Interlock-
                Changed: MRL- PresDet- LinkState+
        DevCap2: Completion Timeout: Not Supported, TimeoutDis-, LTR+, OBFF Via message ARIFwd+
        DevCtl2: Completion Timeout: 50us to 50ms, TimeoutDis-, LTR-, OBFF Disabled ARIFwd-
        LnkCtl2: Target Link Speed: 8GT/s, EnterCompliance- SpeedDis-, Selectable De-emphasis: -6dB
                 Transmit Margin: Normal Operating Range, EnterModifiedCompliance- ComplianceSOS-
                 Compliance De-emphasis: -6dB
        LnkSta2: Current De-emphasis Level: -3.5dB, EqualizationComplete+, EqualizationPhase1+
                 EqualizationPhase2+, EqualizationPhase3+, LinkEqualizationRequest-
```


我们可以通过下面的方法判断 NVMe 设备连接的 PCIe 卡槽是否支持热插拔。

找到 NVMe SSD（如 nvme0n1）对应卡槽的地址（如 0000:04:06.0）；通过 lspci 获得卡槽的热插拔寄存器信息（如果显示为 hotplug+,Surprise+则支持热插拔）

```
[root@        rver-60AFBF ~]# find /sys -name nvme0n1
/sys/devices/pci0000:00/0000:00:03.0/0000:03:00.0/0000:04:06.0/0000:07:00.0/block/nvme0n1
/sys/class/block/nvme0n1
/sys/block/nvme0n1
[root@        rver-60AFBF ~]# lspci -s 0000:04:06.0 -vvv | grep -i hotplug
                SltCap: AttnBtn+ PwrCtrl+ MRL- AttnInd+ PwrInd+ HotPlug+ Surprise+
```


3，确认操作系统的支持

PCIe 热插拔并不是完全由操作系统处理的，也有可能由 BIOS 处理，这完全取决于服务器 BIOS 的设计。当操作系统启动时，会根据 ACPI 提供的信息来了解到底由谁处理 PCIe 热插拔。如果由操作系统处理，则会根据 PCIe 卡槽发送的中断获知热插拔事件。对于 Linux 系统来说，一般使用 pciehp 驱动来干这件事情。所以，最简单的判断方法就是看系统中是否注册了热插拔中断服务程序。

```
[root@LetvWebServer-60AFBF ~]# cat /proc/interrupts | grep pciehp
 85:          2          0          0          0          0          0          0          0          0
          0          0          0  IR-PCI-MSI-edge      pciehp
 86:          2          0          0          0          0          0          0          0          0
          0          0          0  IR-PCI-MSI-edge      pciehp
 87:          9          0          0          0          0          0          0          0          0
          0          0          0  IR-PCI-MSI-edge      pciehp
```


对于 Linux 的 NVMe 热插拔支持将会单独用一篇文章讲解，此处不再多说。

4，确认 NVMe 驱动的支持

与其说驱动的支持，不如说驱动中是否有 Bug。Linux 内核提供了 NVMe 驱动，但是在实际的测试中，驱动的处理不当容易导致系统 Crash 和 Hang 住。产生这些问题的原因基本上可以归纳为 NVMe 驱动 release 设备和 pciehp release 设备产生竞争，出现空指针；NVMe 驱动 release 设备时，上层调用 sync 函数导致进程 block 住。这个最好跟 SSD 厂商沟通好自己的测试环境，以便提前了解可能出现的问题。

如果这些环节都通过，基本上可以确认当前的系统可以进行热插拔了。但是目前，Linux 系统和 PCIe 热插拔驱动存在不少问题，我们在操作中还需要避免出现下面的情况：

避免在一个服务器上短时间内频繁地（或者同时对多个设备）进行热插拔操作

原因：这是 pciehp 驱动中热插拔处理的 bug，centos7 都没有解决。

潜在的问题：可能导致 pciehp 进程 block 住，之后插入的盘无法识别。

解决办法：当对多个盘操作时，顺序进行热插拔，并打开 pciehp 的 debug 功能，通过 dmesg 获得 pciehp 热插拔处理进度。

避免对带有 I/O 的设备进行热插拔（尤其是启用了 Cache 的 I/O）

原因：这是由于 Linux Block 层与 PCIe 热插拔的配合问题导致的。

潜在的问题：可能导致系统某些进程 block 住，或者系统 crash。

解决办法：通过设置卡槽的 power 值，在拔盘之前通知操作系统先移除设备。

避免对已经 mount 文件系统的设备进行热插拔

原因：mount 无法感知热插拔事件。

潜在的问题：文件系统无法使用，数据丢失。

解决办法：提前 umount 文件系统。

按照上面的方法，能够避免绝大多数问题。但是还是可能出现错误，尤其在一些新的服务器厂商的产品中，由于兼容性问题导致 NVMe 设备无法识别。那么我们可以通过卡槽的 Capabilities 寄存器信息判断。如果设备没有被 PCIe 系统正确识别，那么就需要咨询厂商了。

总结

这篇文章主要介绍了在进行 NVMe SSD 热插拔时需要注意的事项。首先，我们检查系统是否支持 NVMe 热插拔，然后避免出现上面提到的 3 种情况。PCIe 目前还无法做到如 SATA/SAS 一样的支持力度，这个需要服务器厂商和 SSD 厂商共同推进，相信在未来会越来越好。用户在这个阶段，只有尽量和厂商多沟通，才能避免操作中造成系统崩溃，数据丢失等风险。

10.6.8 Linux 查看 PCIe 版本/速率以及 ASPM 的方法

https://www.jb51.net/article/172616.htm

查看主板上的 PCI 插槽

```
1    # dmidecode | grep --color "PCI"
```

```
root@      # dmidecode | grep --color "PCI"
                PCI is supported
        Internal Reference Designator: J9C1 - PCIE DOCKING CONN
        Type: x16 PCI Express
        Type: x1 PCI Express
        Type: x1 PCI Express
        Type: x1 PCI Express
        Type: x1 PCI Express
root@      #
```

在 Linux 下要如何得知 PCI-E Bus 使用的是 Gen(Generation) 1 還是 Gen2 還是新一代的 Gen 3 雖然使用

#lspci 只要可以看到目前系统所有的装置.但是好像看不到 PCI-E Bus 所採用的是哪一代的 PCI-E.

```
root@XXX# lspci
00:00.0 Host bridge: Intel Corporation Haswell DRAM Controller (rev 06)
00:01.0 PCI bridge: Intel Corporation Haswell PCI Express x16 Controller (rev 06)
00:01.1 PCI bridge: Intel Corporation Haswell PCI Express x8 Controller (rev 06)
00:02.0 VGA compatible controller: Intel Corporation Haswell Integrated Graphics Controller (rev 06)
00:03.0 Audio device: Intel Corporation Haswell HD Audio Controller (rev 06)
00:14.0 USB controller: Intel Corporation Lynx Point USB xHCI Host Controller (rev 05)
00:16.0 Communication controller: Intel Corporation Lynx Point MEI Controller #1 (rev 04)
00:1a.0 USB controller: Intel Corporation Lynx Point USB Enhanced Host Controller #2 (rev 05)
00:1c.0 PCI bridge: Intel Corporation Lynx Point PCI Express Root Port #1 (rev d5)
00:1c.4 PCI bridge: Intel Corporation Lynx Point PCI Express Root Port #5 (rev d5)
00:1c.5 PCI bridge: Intel Corporation Lynx Point PCI Express Root Port #6 (rev d5)
00:1d.0 USB controller: Intel Corporation Lynx Point USB Enhanced Host Controller #1 (rev 05)
00:1f.0 ISA bridge: Intel Corporation Lynx Point LPC Controller (rev 05)
00:1f.2 IDE interface: Intel Corporation Lynx Point 4-port SATA Controller 1 [IDE mode] (rev 05)
00:1f.3 SMBus: Intel Corporation Lynx Point SMBus Controller (rev 05)
00:1f.6 Signal processing controller: Intel Corporation Lynx Point Thermal Management Controller (rev 05)
01:00.0 PCI bridge: PLX Technology, Inc. Unknown device 8724 (rev ca)
02:01.0 PCI bridge: PLX Technology, Inc. Unknown device 8724 (rev ca)
02:02.0 PCI bridge: PLX Technology, Inc. Unknown device 8724 (rev ca)
02:08.0 PCI bridge: PLX Technology, Inc. Unknown device 8724 (rev ca)
02:09.0 PCI bridge: PLX Technology, Inc. Unknown device 8724 (rev ca)
03:00.0 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)
03:00.1 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)
03:00.2 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)
03:00.3 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)
04:00.0 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)
04:00.1 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)
04:00.2 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)
04:00.3 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)
06:00.0 Ethernet controller: Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection (rev 01)
06:00.1 Ethernet controller: Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection (rev 01)
07:00.0 PCI bridge: PLX Technology, Inc. PEX 8732 32-lane, 8-Port PCI Express Gen 3 (8.0 GT/s) Switch (rev ca)
08:01.0 PCI bridge: PLX Technology, Inc. PEX 8732 32-lane, 8-Port PCI Express Gen 3 (8.0 GT/s) Switch (rev ca)
08:08.0 PCI bridge: PLX Technology, Inc. PEX 8732 32-lane, 8-Port PCI Express Gen 3 (8.0 GT/s) Switch (rev ca)
08:09.0 PCI bridge: PLX Technology, Inc. PEX 8732 32-lane, 8-Port PCI Express Gen 3 (8.0 GT/s) Switch (rev ca)
08:0a.0 PCI bridge: PLX Technology, Inc. PEX 8732 32-lane, 8-Port PCI Express Gen 3 (8.0 GT/s) Switch (rev ca)
09:00.0 Ethernet controller: Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection (rev 01)
09:00.1 Ethernet controller: Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection (rev 01)
0e:00.0 Ethernet controller: Intel Corporation I210 Gigabit Network Connection (rev 03)
0f:00.0 Ethernet controller: Intel Corporation I210 Gigabit Network Connection (rev 03)
root@XXX#
```

```
root@XXX# lspci -tv
-[0000:00]-+-00.0 Intel Corporation Haswell DRAM Controller
  +-01.0-[0000:01-06]----00.0-[0000:02-06]--+-01.0-[0000:03]--+-00.0 Intel Corporation I350 Gigabit Network Connection
  |           |    +-00.1 Intel Corporation I350 Gigabit Network Connection
  |           |    +-00.2 Intel Corporation I350 Gigabit Network Connection
  |           |    \-00.3 Intel Corporation I350 Gigabit Network Connection
  |           +-02.0-[0000:04]--+-00.0 Intel Corporation I350 Gigabit Network Connection
```

```
8       |       |   +-00.1 Intel Corporation I350 Gigabit Network Connection
9       |       |   +-00.2 Intel Corporation I350 Gigabit Network Connection
10      |       |   \-00.3 Intel Corporation I350 Gigabit Network Connection
11      |       +-08.0-[0000:05]--
12      |       \-09.0-[0000:06]--+-00.0 Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection
13      |           \-00.1 Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection
14      +-01.1-[0000:07-0c]----00.0-[0000:08-0c]--+-01.0-[0000:09]--+-00.0 Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network
15      |       |   \-00.1 Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection
16      |       +-08.0-[0000:0a]--
17      |       +-09.0-[0000:0b]--
18      |       \-0a.0-[0000:0c]--
19      +-02.0 Intel Corporation Haswell Integrated Graphics Controller
20      +-03.0 Intel Corporation Haswell HD Audio Controller
2       +-14.0 Intel Corporation Lynx Point USB xHCI Host Controller
        +-16.0 Intel Corporation Lynx Point MEI Controller #1
        +-1a.0 Intel Corporation Lynx Point USB Enhanced Host Controller #2
        +-1c.0-[0000:0d]--
        +-1c.4-[0000:0e]----00.0 Intel Corporation I210 Gigabit Network Connection
        +-1c.5-[0000:0f]----00.0 Intel Corporation I210 Gigabit Network Connection
        +-1d.0 Intel Corporation Lynx Point USB Enhanced Host Controller #1
        +-1f.0 Intel Corporation Lynx Point LPC Controller
        +-1f.2 Intel Corporation Lynx Point 4-port SATA Controller 1 [IDE mode]
        +-1f.3 Intel Corporation Lynx Point SMBus Controller
        \-1f.6 Intel Corporation Lynx Point Thermal Management Controller
      root@XXX#
```

如果有装置是 unknown 的,需要更新 /usr/local/share/pci.ids.gz 請參考更新方式 http://benjr.tw/node/88

先查询 Inetl 82599EB 网卡的识别号(bus:device.function)

```
1   root@XXX# lspci | grep --color 82599
2   06:00.0 Ethernet controller: Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection (rev 01)
3   06:00.1 Ethernet controller: Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection (rev 01)
4   09:00.0 Ethernet controller: Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection (rev 01)
5   09:00.1 Ethernet controller: Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection (rev 01)
6   root@XXX#
```

在 PCI 的装置使用三个编号用来当作识别值,个别为 1. "汇流排(bus number)", 2. "装置(device number) 以及 3. "功能(function number)".

所以刚刚的 06:00.0 就是 bus number = 06 ,device number = 00 function = 0 .

这 3 个编号会组合成一个 16-bits 的识别码,

汇流排(bus number) 8bits 2^8 至多可连接 256 个汇流排(0 to ff),

装置(device number) 5bits 2^5 至多可接 32 种装置(0 to 1f) 以及

功能(function number) 3bits 2^3 至多每种装置可有 8 项功能(0 to 7).

关于更多 #lspci 的资讯请参考 http://benjr.tw/node/543

然后查看 vendor id 和 device id

```
root@XXX# lspci -n | grep -i 06:00.0
06:00.0 0200: 8086:10fb (rev 01)
root@XXX#
```

Linux 使用 Class ID + Vendor ID + Device ID 來代表裝置,如剛剛的 0200: 8086:10fb 所代表裝置名稱為 (Class ID = 0200 , Vendor ID = 8086, Device ID = 10fb)

最后查看指定 PCI 设备的带宽

```
root@XXX# lspci -n -d 8086:10fb -vvv | grep --color Width
  LnkCap: Port #9, Speed 5GT/s, Width x8, ASPM L0s, Latency L0 <1us, L1 <8us
  LnkSta: Speed 5GT/s, Width x8, TrErr- Train- SlotClk+ DLActive- BWMgmt- ABWMgmt-
  LnkCap: Port #9, Speed 5GT/s, Width x8, ASPM L0s, Latency L0 <1us, L1 <8us
  LnkSta: Speed 5GT/s, Width x8, TrErr- Train- SlotClk+ DLActive- BWMgmt- ABWMgmt-
  LnkCap: Port #1, Speed 5GT/s, Width x8, ASPM L0s, Latency L0 <1us, L1 <8us
  LnkSta: Speed 5GT/s, Width x8, TrErr- Train- SlotClk+ DLActive- BWMgmt- ABWMgmt-
  LnkCap: Port #1, Speed 5GT/s, Width x8, ASPM L0s, Latency L0 <1us, L1 <8us
  LnkSta: Speed 5GT/s, Width x8, TrErr- Train- SlotClk+ DLActive- BWMgmt- ABWMgmt-
root@XXX#
```

LnkSta : 目前系統所提供的速度 PCI-Express 2.0 (5GT/s)

LnkCap : 裝置目前所採用的速度.

LnkSta 和 LnkCap 這兩個速度有可能不一樣 , 典型情況下: 系統所提供的是 PCI Express 是 3.0 但裝置還是使用 2.0 的.

10.6.9 芯片中的数学 — 均衡器 EQ 和它在高速外部总线中的应用

转自：https://zhuanlan.zhihu.com/p/48343011

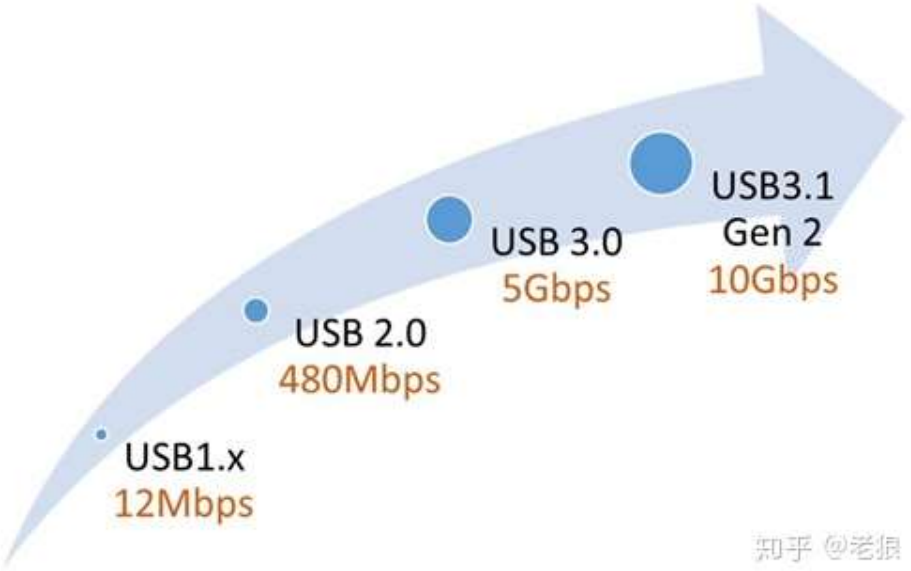
高速的串行总线逐渐淘汰了系统中的并行总线，作为并行总线最后堡垒的内存总线也越来越多的吸收了其中关键技术，尤其是均衡器（Equalization，EQ）技术。为什么串行总线会替代并行总线？EQ 被广泛使用，程序员甚至固件工程师为什么接触不多？本文主要介绍为什么要引入 EQ，EQ 对信号完整性的好处，眼图的作用，最后讨论各种均衡器在 PCIe、USB 等等高速串行总线中的应用及固件 BIOS 和驱动为什么没有涉及众多 EQ。

并行总线到串行总线的转变

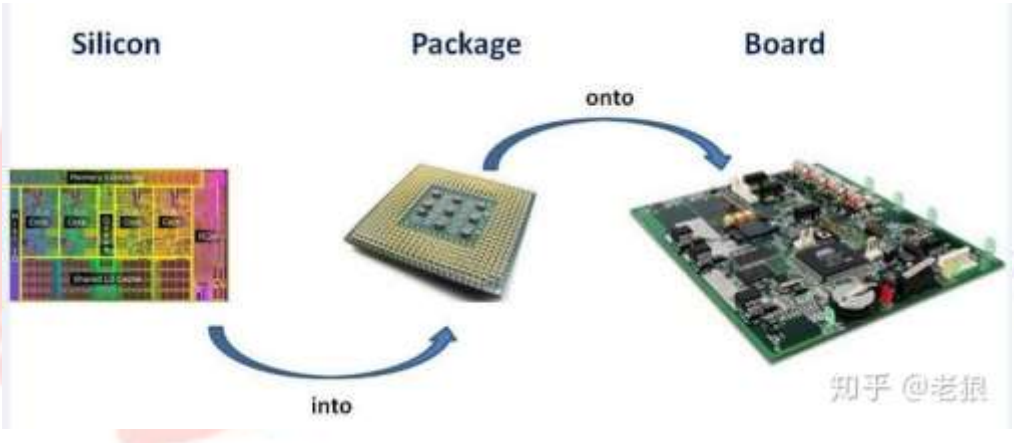
过去都认为串行比并行慢，串口比并口慢，就像四车道比单车道通行速度高一样很好理解。然而近十几年来，并行总线发展遇到了瓶颈。并行总线因为抗干扰能力差，时钟与数据同时传输的并行传输方式和线路串扰等等问题导致很难达到 1Gb/s 以上带宽，内存总线为了对齐/校准时钟与数据付出了极大的代价。而串行总线自从引入了差分信号后，对共模干扰抵抗能力很强，信道中没有时钟信号，时钟是在数据接收端进行恢复。

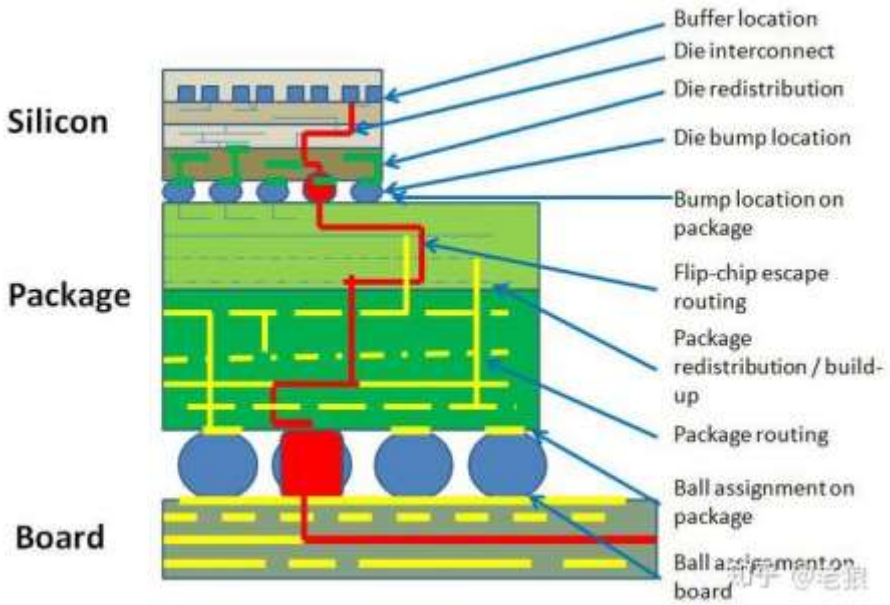
这些优点让串行总线频率可以越来越高，应用串行总线的 USB、PCIe、SATA、QPI、HDMI 等等外部总线将并行总线挤压到只剩下内存总线这个最后的堡垒。甚至连接 Flash 的总线也变成了 SPI 串行总线。

在传输速度不断提高后，即使如差分信号，它的信号完整性的问题也慢慢变得越来越严重。最新的接口版本如 PCI-Express 3.0 和 USB 3.1 等速率 都已经提升到 8Gbps 和 10Gpbs 以上：



而我们主板却还是沿用 FR4 低成本板材，信号在经过多次连接和传输后衰竭严重：





这是两根 PCB 走线的插入损耗图：

红色 10 英寸，蓝色 5 英寸。可以看出频率越高，插入损耗越大，信号的衰减在 10 英寸后可达 26dB，而 USB 3.0 最大只能允许 23dB 的衰减。所以不经过处理，信号从发送端 Tx 到达接收端 Rx 时，已经有了较大的损耗，可能导致 Rx 无法正确还原和解码信号，从而出现误码，甚至眼图完全闭合。

为降低误码率，USB 3.1 要求发送端 Tx 必须实现 **FFE**（Feed-forward Equalizer，前馈均衡器），接收端 Rx 必须实现 **CTLE** (Continuous Time Linear Equalizer，连续时间轴性均衡器)和 **DFE** (Decision Feedback Equalizer，判决反馈均衡器)。PCIe 在 1 代（gen1）和 2 代（gen2）中使用了去加重（**De-emphasis**）技 术和 **Preshoot** 技术，在 3 代（gen3）8GHz 时钟的要求下

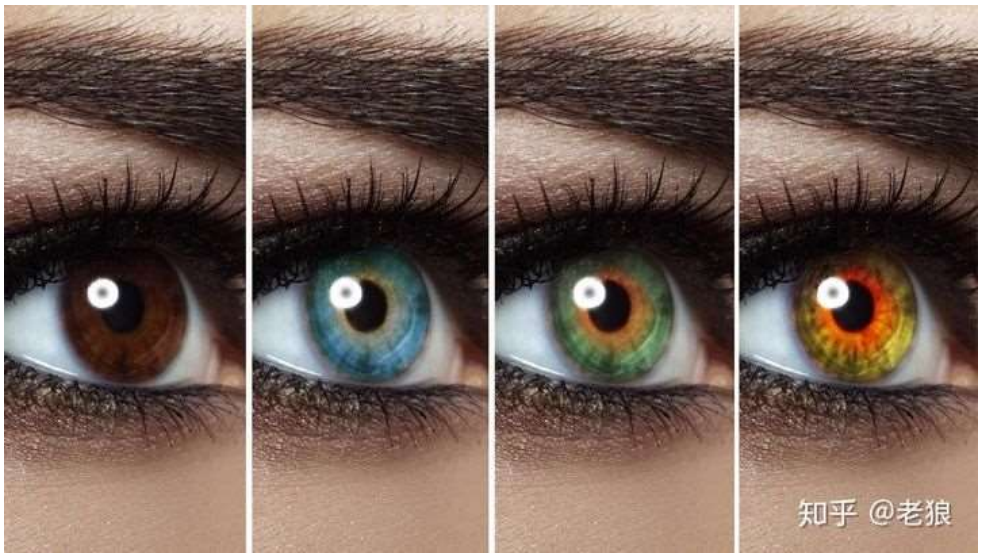
标准	时钟	传输位宽	每时钟数据	带宽
ISA	4.77 MHz	8	1	4.77 MB/s
ISA	8 MHz	16	0.5	8 MB/s
MCA	5 MHz	16	1	10 MB/s
MCA	5 MHz	32	1	20 MB/s
EISA	8.33 MHz	32	1	33.3 MB/s (16.7 MB/s typically)
VLB	33 MHz	32	1	133 MB/s
PCI	33 MHz	32	1	133 MB/s
PCI-X 66	66 MHz	64	1	533 MB/s
PCI-X 133	133 MHz	64	1	1,066 MB/s
PCI-X 266	133 MHz	64	2	2,132 MB/s
PCI-X 533	133 MHz	64	4	4,266 MB/s
AGP x1	66 MHz	32	1	266 MB/s
AGP x2	66 MHz	32	2	533 MB/s
AGP x4	66 MHz	32	4	1,066 MB/s
AGP x8	66 MHz	32	8	2,133 MB/s
PCIe 1.0 x1	2.5 GHz	1	1	250 MB/s
PCIe 1.0 x4	2.5 GHz	4	1	1,000 MB/s
PCIe 1.0 x8	2.5 GHz	8	1	2,000 MB/s
PCIe 1.0 x16	2.5 GHz	16	1	4,000 MB/s
PCIe 2.0 x1	5 GHz	1	1	500 MB/s
PCIe 2.0 x4	5 GHz	4	1	2,000 MB/s
PCIe 2.0 x8	5 GHz	8	1	4,000 MB/s
PCIe 2.0 x16	5 GHz	16	1	8,000 MB/s
PCIe 3.0 x1	8 GHz	1	1	1,000 MB/s
PCIe 3.0 x4	8 GHz	4	1	4,000 MB/s
PCIe 3.0 x8	8 GHz	8	1	8,000 MB/s
PCIe 3.0 x16	8 GHz	16	1	16,000 MB/s

也引入了 FFE、CTLE、DFE 和 CDR 均衡器（EQ）。

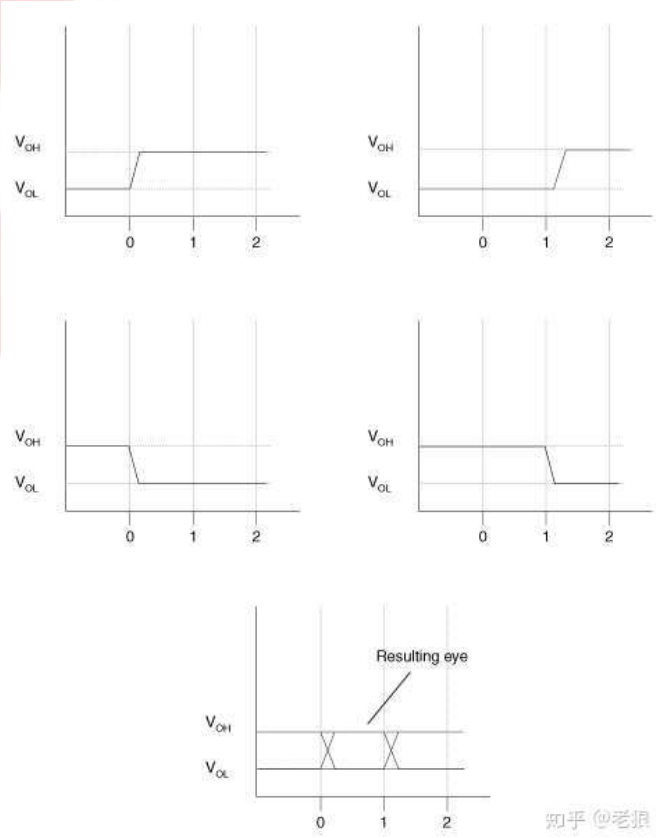
在我们介绍各种 EQ 之前，我们先来了解一下什么是眼图和眼图对于信号完整性的重要作用。

什么是眼图？

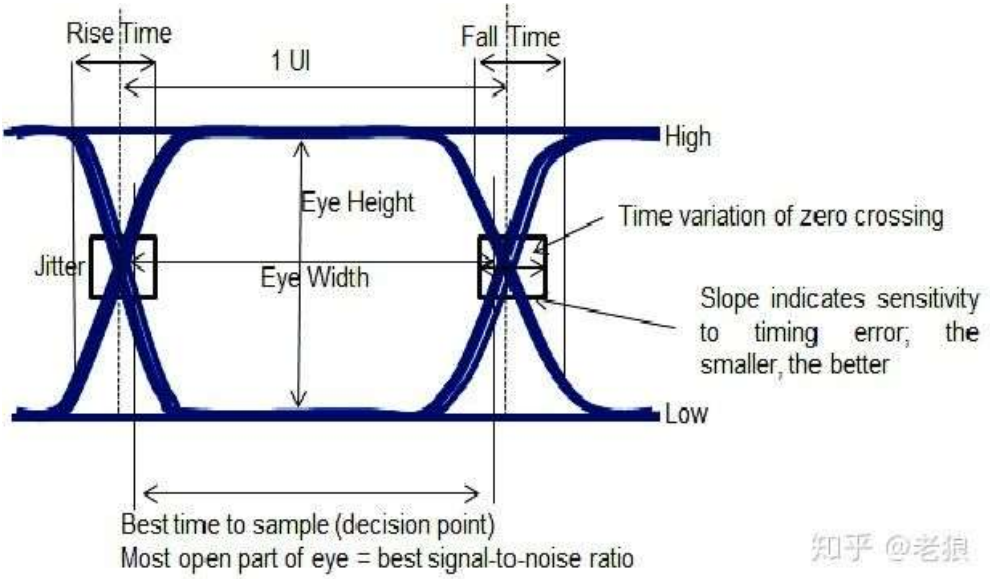
眼图并不是眼睛的图：



所谓眼图就是把一连串信号(000,001,010,011, 100, 101,110,111)叠加在一起，形成一个类似眼睛的图像，通常是在示波器上。

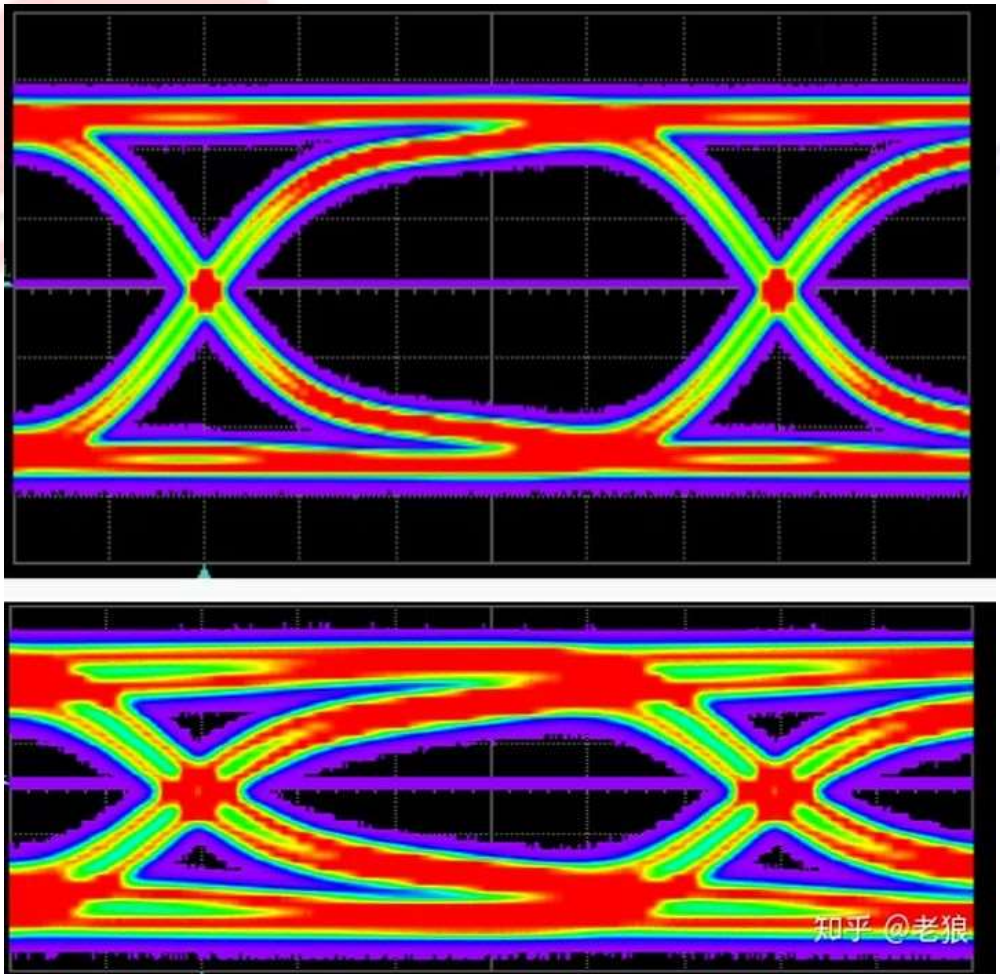


如这个示意图，把 011、001、100 和 110 叠加在一起形成一个眼图。它有不少术语：



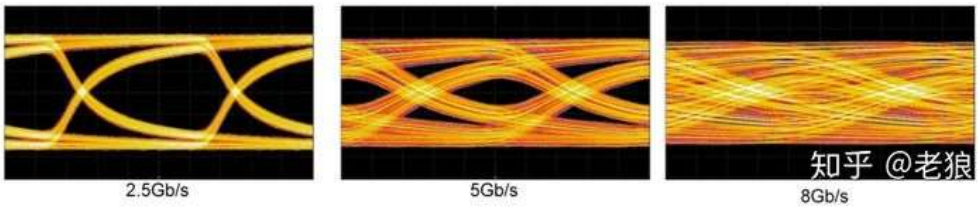
其中包括：高电平，低电平，周期(UI,Unit Interval)，眼高，眼宽，上升时，下降时和抖动 Jitter。

眼宽大，眼高高，Jitter 窄，眼图就好，我们叫做眼图睁开；眼宽扁，眼高低，Jitter 窄，信号就差，甚至难以采样和辨识，这时我们就叫眼图闭合。



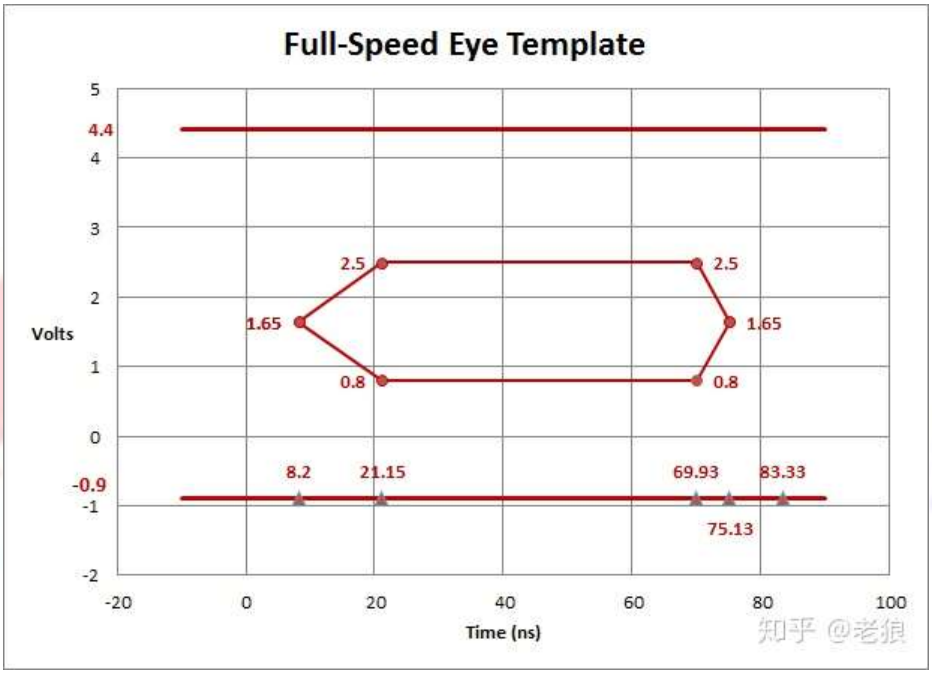
好眼图 VS 差眼图

举个例子：

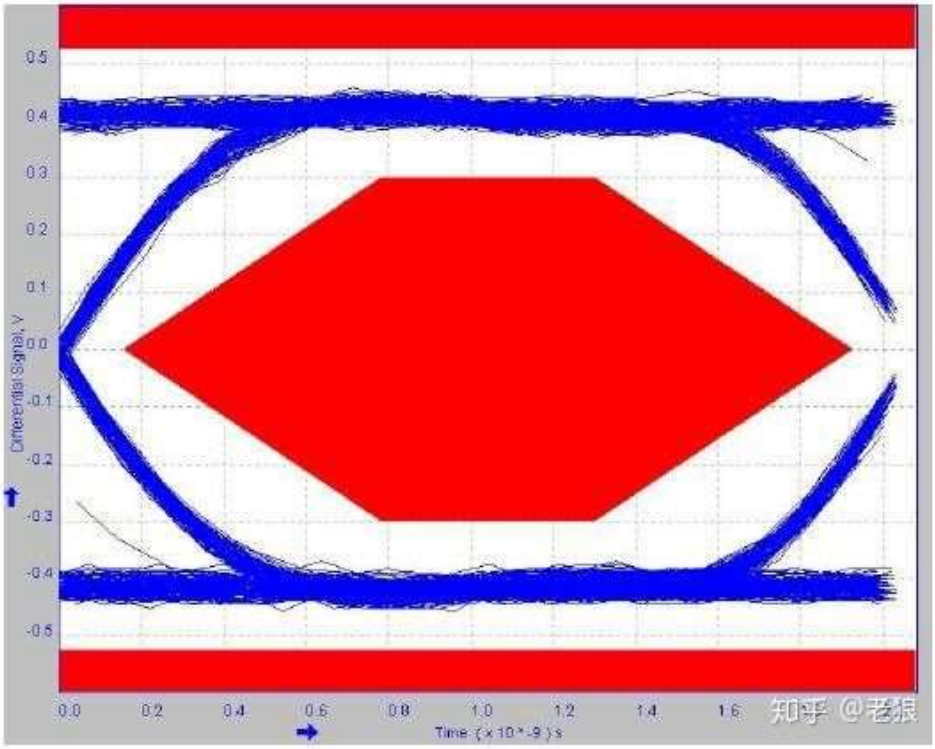


这是 PCIe 在 10 英寸的 PCB 版上 Rx 端接收到的眼图。2.5G 是眼图睁开，5G 则是半闭，而 8G 就完全闭合了，这时是不能够辨识数据的。

为了确保信号传输后的完整性，各个高速协议组织都公布了测试标准，例如 USB 协会发布了眼图模板：



这些红线部分是眼图不能碰的，碰到就属于不符合标准。一个符合标准的眼图如下：



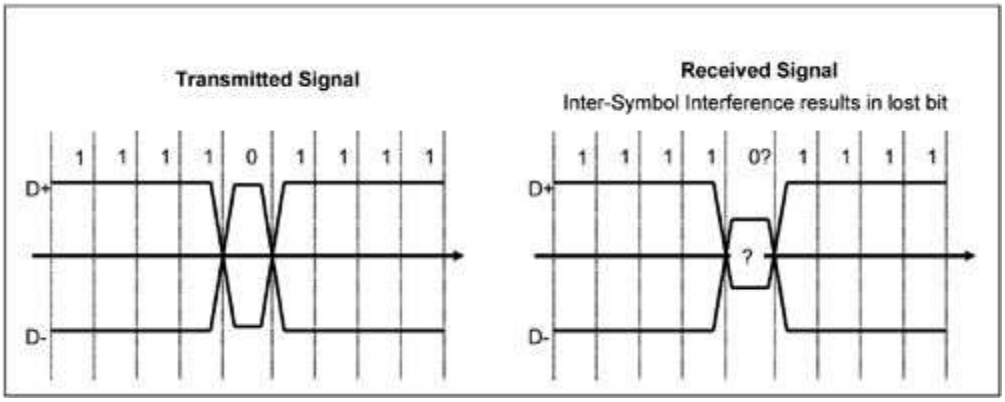
而不符合标准的质量比较差的电缆则眼图十分糟糕。

如前面所说，为了对抗高频信号的衰减和干扰，各种方法如去加重（**De-emphasis**）和 **Preshoot** 技术，以及各种 EQ 被引入传输协议。下面就以 PCIe 为例，介绍一下我们 BIOS 工程师和电脑爱好者可能感兴趣的其中几个关键技术。

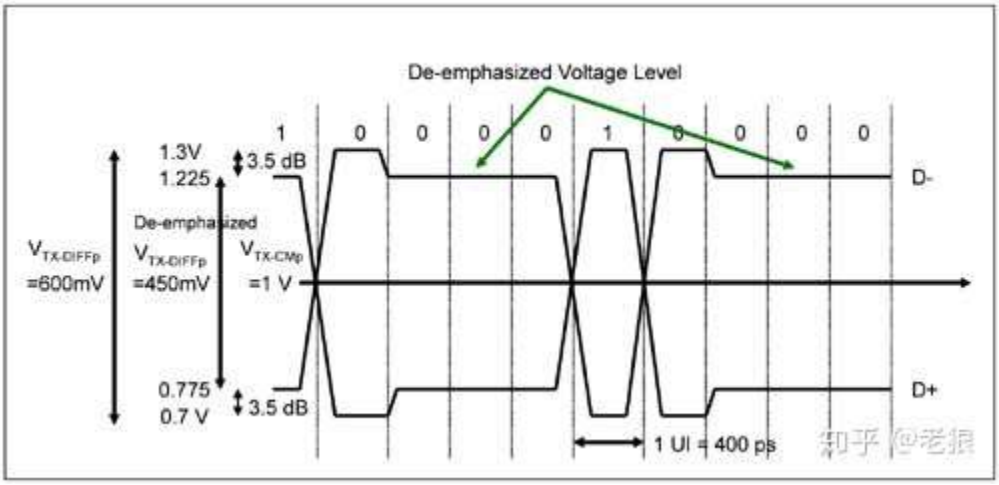
什么是去加重和 preshoot？

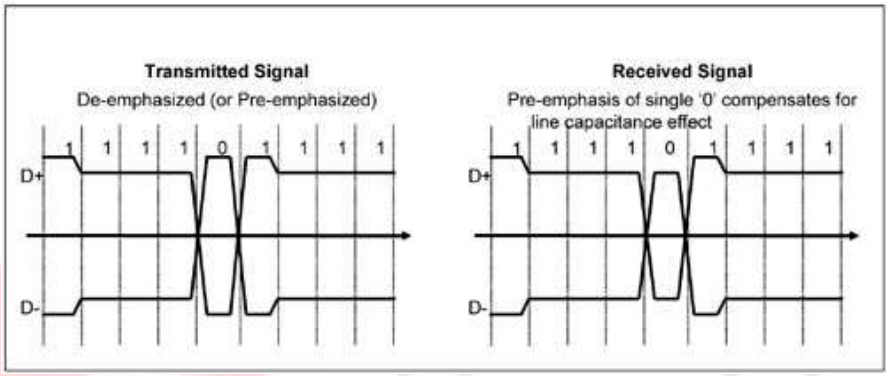
去加重（**De-emphasis**）和 preshoot 是为了对抗码间干扰的（**inter-symbol interference， ISI**）。

什么是码间干扰呢？我们可以这么理解，当我们发送 111101111 这样的数据是，忽然变化的 0，让电路里的电容很难迅速放电达到 0，后面又被迅速拉到 1，造成 0 的信号眼图很小：

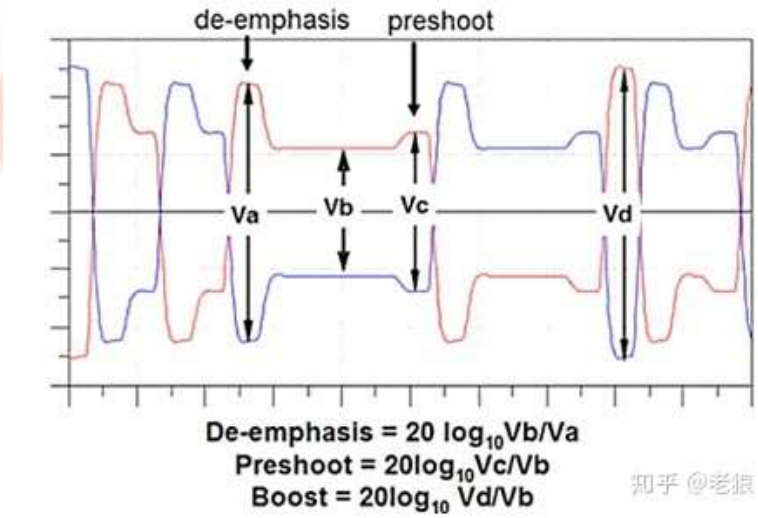


而这种情况随着频率的提高越来越严重。从信号的角度来看，也就是信道对高频衰减大，而对低频衰减小。那怎么办呢？通过压低 1 的幅度来张开 0 的眼图：





而 Preshoot 是将跳变前一个增大幅度：



PCIE 3 代中规定了共 11 种不同的 Preshoot 和 De-emphasis 的组合（Preset）

Preset Number	Preshoot (dB)	De-emphasis (dB)
P4	0.0	0.0
P1	0.0	-3.5 ± 1 dB
P0	0.0	-6.0 ± 1.5 dB
P9	3.5 ± 1 dB	0.0
P8	3.5 ± 1 dB	-3.5 ± 1 dB
P7	3.5 ± 1 dB	-6.0 ± 1.5 dB
P5	1.9 ± 1 dB	0.0
P6	2.5 ± 1 dB	0.0
P3	0.0	-2.5 ± 1 dB
P2	0.0	-4.4 ± 1.5 dB
P10	0.0	See Note 2.

在 PCIe root port 链路初始化 Training 中，Rx 发送 TxEQ preset 设置 要求给 Tx，此过程叫做动态均衡。是的，他们本质上是一种 **FFE**（Feed-forward Equalizer，前馈均衡器），发送端 Tx 通过它提高信号完整性。那么接受端 Rx 呢？

什么是 CTLE 和 DFE？

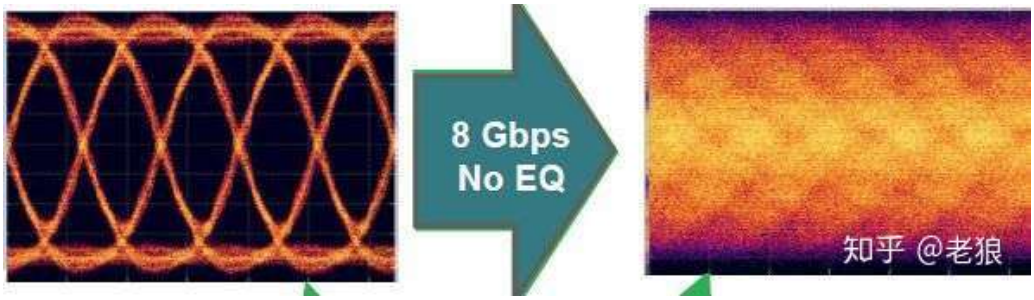
Rx 端采用 **CTLE** (Continuous Time Linear Equalizer，连续时间轴性均衡器) 和 **DFE** (Decision Feedback Equalizer，判决反馈均衡器)。限于篇幅，本文只简要介绍一下。

CTLE 是利用连续的信号曲线，减缓低频部分，用来补偿高频部分，因为高频部分损耗较大，所谓削峰填谷。它有个缺点是会放大高频噪声。

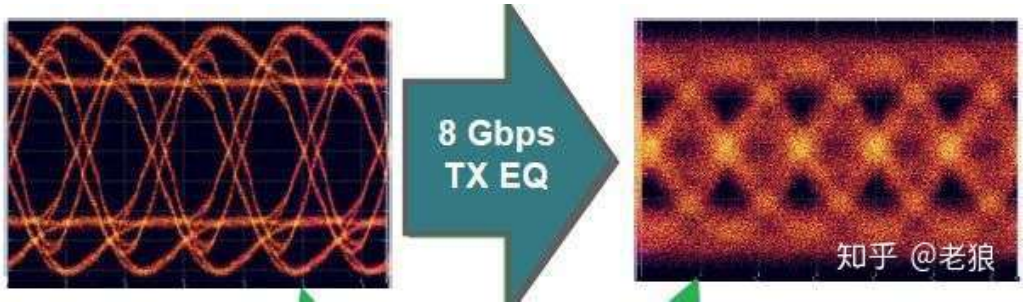
DFE 也是一种回馈均衡器，是用上次信道的输出经过判断后加权反馈到输入上。它不会放大高频噪声，但是只能处理码后干扰，不能消除码前干扰，且设计复杂和耗电。

效果如何？

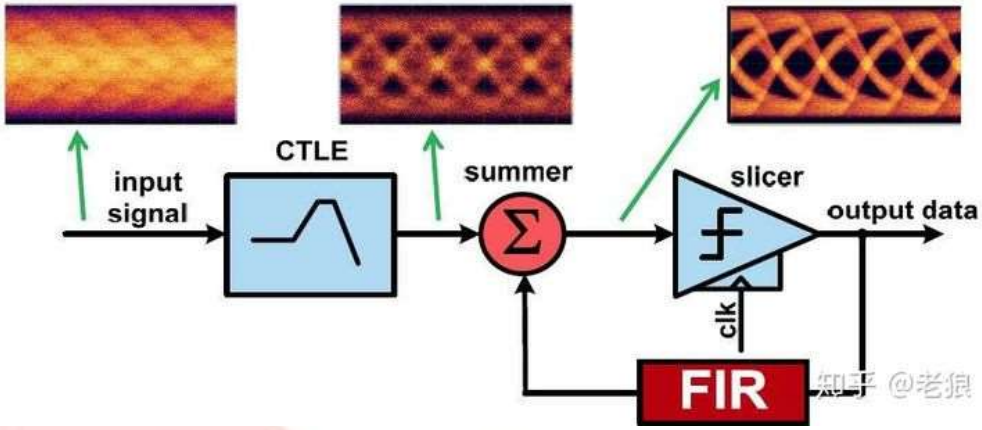
PCIe 3.0 信号不经过 EQ 处理是这样，眼图关闭：



Tx 经过 EQ 变成：



再在 Rx 经过 CTLE 和 DFE 后：



眼图才全部张开。

结论

我们把 FFE 和 DFE 这种具有回馈和自动调整的 EQ 叫做自适应均衡器（Adaptive Equalization），将 CTLE 这种叫做固定均衡器（Fixed Equalization）。普通程序员和一般 BIOS 工程师尽管可能接触了不少 PCIe、USB、HDMI 甚至是 QPI 的内容，但几乎都不会接触 EQ。这是因为 EQ 大部分是自适应的，是在链路 train 的时候，由硬件自动完成的，芯片组完成了 Tx 的部分，板卡或者设备中 TI 等的芯片完成了 Rx 的部分，极少需要固件和驱动参与。只有在链路出现问题后的调试甚至 workaround 错误时才需要手动设置 EQ 参数，来解决不能 train 通或者 train 到更高速率的问题。

不仅仅高速串行总线，现在内存的并行总线中也引入了 DFE 和 CTLE 算法，但它是 MRC 程序实现而不是硬件实现。

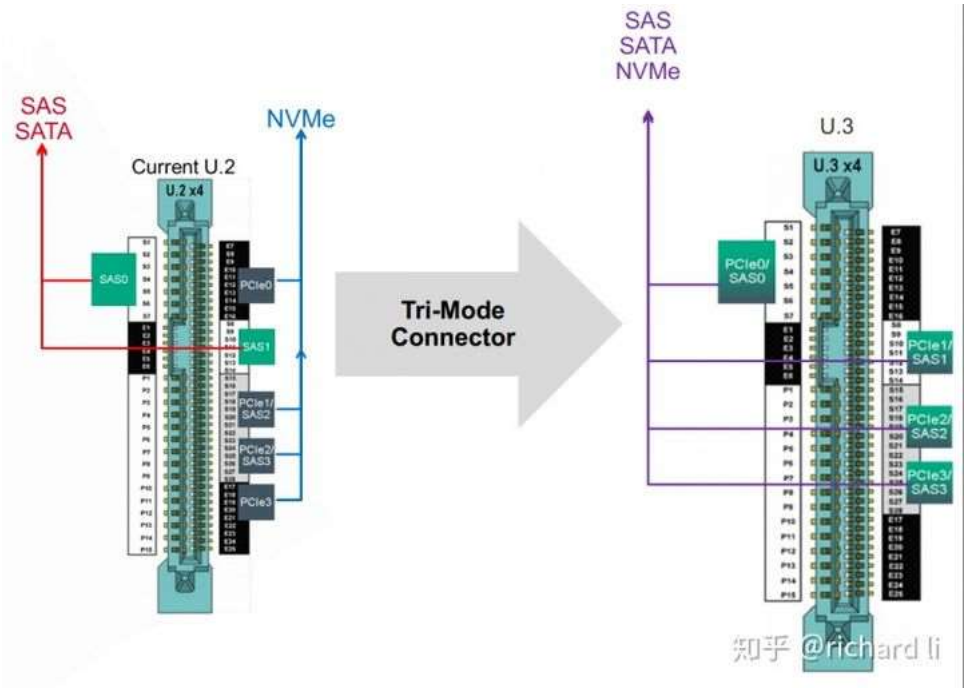
10.7 PCIe NVMe SSD 各种接口简介

10.7.1 PCIe U.2/U.3 接口的区别

https://zhuanlan.zhihu.com/p/133923652

同样使用 SFF-8639 的连接器，但是引脚定义不同。如图

- U.2 的 SAS/SATA 与 NVMe 使用不同的高速引脚。
- U.3 的 SAS/SATA 与 NVMe 共用高速引脚



U.3 主要是为了上行连接到支持 Tri-mode 的控制器，比如 Broadcom 的 ROC 与 IOC。

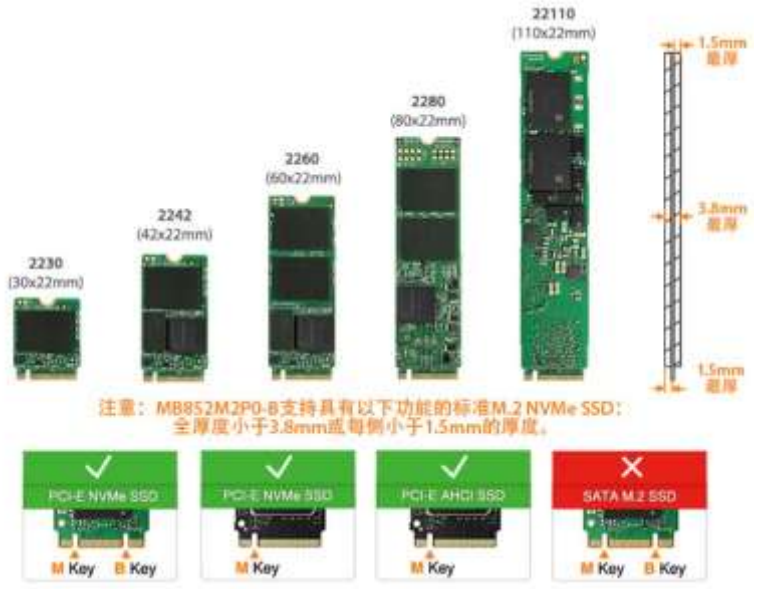
10.7.2 SATA 和 NVMe M.2 接口介绍

首先说一下目前固态硬盘常用的两个接口（与主板相连的接口形状）SATA3 和 M.2。

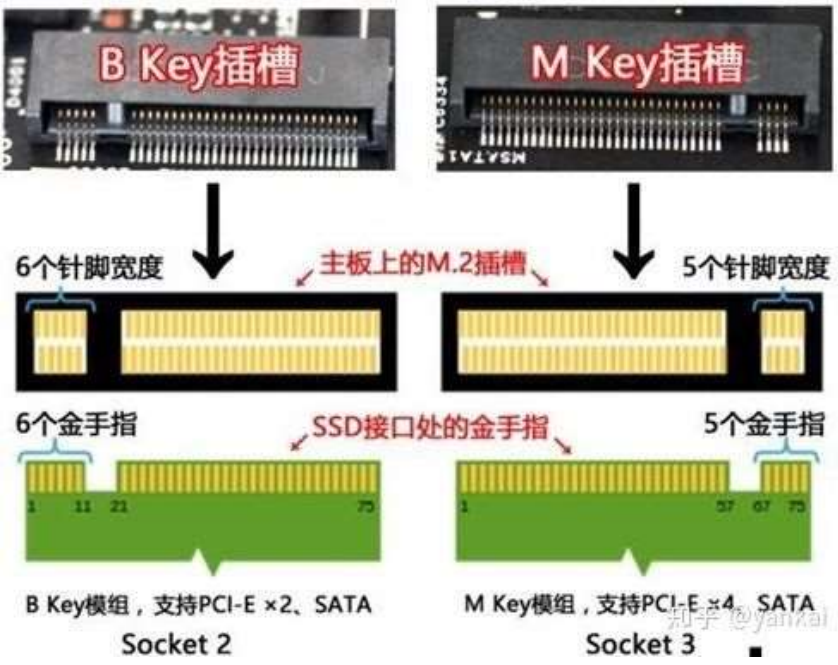
（1）采用 SATA3 接口（目前机械硬盘采用的接口方式）的固态硬盘，在传输方式上与 SATA3 的机械硬盘一样，速度的提升完全依靠固态硬盘的存储读取的速度比机械硬盘速度快。大约为传统机械硬盘的 5 倍。

（2）采用 M.2 接口的固态硬盘介绍如下：

尺寸：分为 2242/2260/2280/22110，就是长短不一样，越长可放的 flash 块儿越多，容量越大。



模组：金手指分为 B key 又称 Socket2 和 M key 又称 Socket3，不同的接口支持的总线不同。如下图所示：



在匹配主板的时候，记得查看主板支持的类型，目前基本上都是 M key 模组。

总线方式：分为 SATA3 和 PCI-E。

A.采用 SATA3 方式的和采用 SATA3 接口的固态硬盘类似，速度相当，属于固态硬盘中速度最慢的。

B.采用 PCI-E 总线固态硬盘， 不是台式机上的 PCI-E 的接口，目前 PCI-E 接口表示为 PCIe 3×4，全称为 PCI Express Gen 3×4，总线宽度大 32Gbps，实际传输速率超过 1000 MB/s.

协议：采用 PCI-E 线的 SSD 通常带有 NVMe，NVMe 其实与 AHCI 一样都是逻辑设备接口标准。NVMe 中文名称非易失性存储器标准，是使用 PCI-E 通道的 SSD 一种规范。NVMe 的优点在于更低的延时，更高的传输速率，更低的功耗控制。采用 M.2 接口支持 PCI-E3×4 的 SSD 速率可达 1000 MB/s，如果在支持 NVMe 协议，速率将突破 2000 M/s.

发布于 2018-07-14 14:59

10.7.3 数据中心 NVMe SSD 和 EDSFF 前瞻

来自 Intel、HPE、Dell & SNIA 等

Q：EDSFF 中 E1.S、E1.L、E3 都是怎么划分定位的？

Q：EDSFF 的发展与 PCIe 规范演进之间的关系是？

Q：为什么服务器厂商还在用 U.2 SSD，这背后的始作俑者是？

Q：如果 EDSFF 设备不只有 SSD，还能支持啥？

这两天看到一份很不错的技术资料，上面那几条都有答案，赶紧写点东西分享给大家。

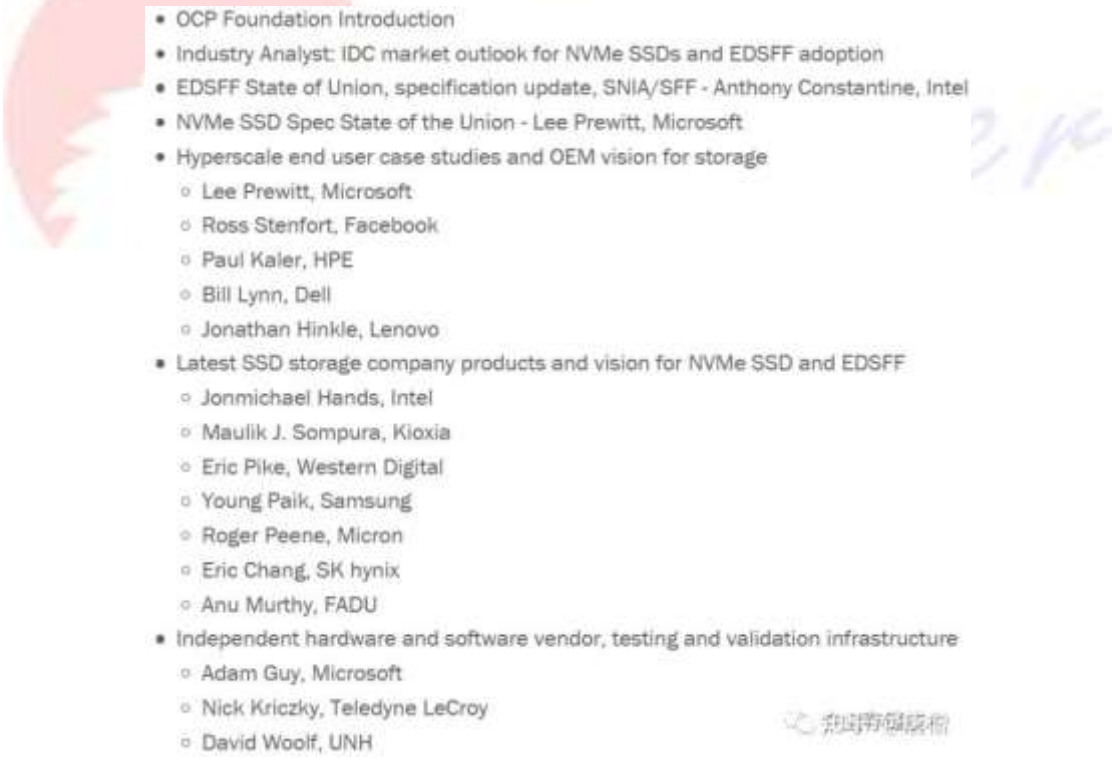
关于下一代数据中心 SSD 规格形态，无论 E1.S、E1.L 还是 E3（EDSFF 3 英寸），我以前都曾讨论过：

《EDSFF 3 英寸企业级 SSD 会成为下一代标准吗？》

《下一代数据中心 SSD 形态之争：调查报告篇》

《下一代数据中心 SSD 形态之争：来自 Azure 架构师的观点》

今天要给大家分享的内容，基本上可以把 EDSFF 比较全面、系统地呈现了。我的参考资料来自 OCP（Open Compute Project）的一场 Webinar，下面可以看到其演讲阵容的强大：

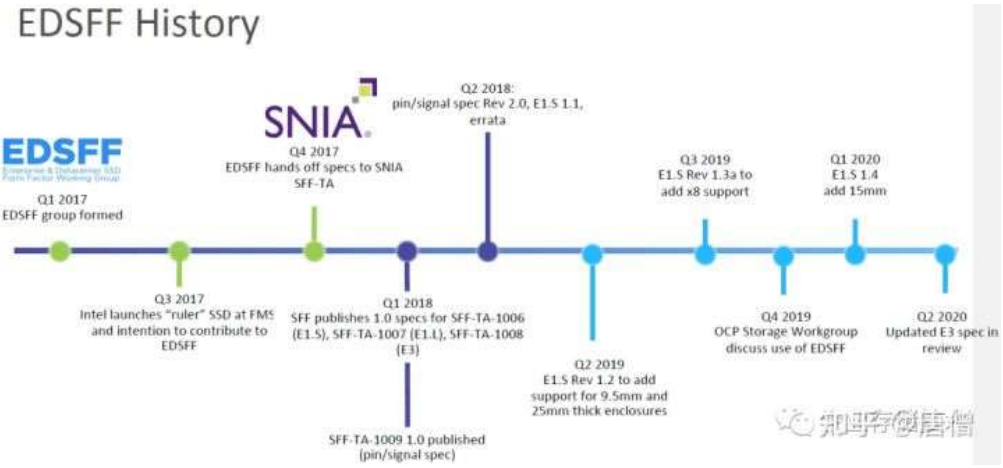


- OCP Foundation Introduction
- Industry Analyst: IDC market outlook for NVMe SSDs and EDSFF adoption
- EDSFF State of Union, specification update, SNIA/SFF - Anthony Constantine, Intel
- NVMe SSD Spec State of the Union - Lee Prewitt, Microsoft
- Hyperscale end user case studies and OEM vision for storage
  - Lee Prewitt, Microsoft
  - Ross Stenfort, Facebook
  - Paul Kaler, HPE
  - Bill Lynn, Dell
  - Jonathan Hinkle, Lenovo
- Latest SSD storage company products and vision for NVMe SSD and EDSFF
  - Jonmichael Hands, Intel
  - Maulik J. Sompura, Kioxia
  - Eric Pike, Western Digital
  - Young Paik, Samsung
  - Roger Peene, Micron
  - Eric Chang, SK hynix
  - Anu Murthy, FADU
- Independent hardware and software vendor, testing and validation infrastructure
  - Adam Guy, Microsoft
  - Nick Kriczky, Teledyne LeCroy
  - David Woolf, UNH

其中 Hyperscale 超大规模数据中心用户和（服务器）OEM 厂商，包括：微软、Facebook、HPE、Dell 和联想；SSD 存储厂商包括：Intel、Kioxia（前东芝）、WD、三星、Micron、SK hynix 等。

与其说 SNIA（存储网络工业协会）是 EDSFF 的规范制定者，不如说是 Intel 一直在背后推动。看看下图的配色风格吧：）

Intel 分享：EDSFF Overview



根据这个时间轴，E1.S（SFF-TA-1006）、E1.L（SFF-TA-1007）和 E3（SFF-TA-1008）早在 2018 年 Q1 就发布了 1.0 规范，3 年后的今天，也到了该成熟和大规模应用的时候了。

对于上述 3 类新的数据中心 NVMe SSD 尺寸和标准，E1.S 定位是可扩展和灵活的性能型存储，E1.L 针对高容量存储（如：QLC），E3 则是用于 2U 服务器/存储中的高性能 SSD。

Intel Recommended Platform Design Guidance

	2U Server	1U Server	Storage/JBOF	Enterprise Storage Array	Boot
(PCIe 4.0)	U.2 15mm	OEM: U.2 Hyperscale: U2 or E1.S	E1.L or U.2	U.2 Dual Port	M.2
(PCIe 4.0 → 5.0)	U.2 & E3.S	OEM: U.2/E1.S/E3 Hyperscale: E1.S	E1.L	U.2 Dual Port	M.2
(PCIe 5.0)	E3.S	E1.S & E3.S	E1.L	E3.S or E3.L	M.2 or E1.S

可能有的朋友会问，OEM 大厂的服务器新品中有些还清一色沿用 U.2，比如 Dell PowerEdge 15G。从上面这个“Intel 推荐平台设计指导”不难找到答案。

在 PCIe 4.0 阶段之所以 OEM 仍青睐 U.2，主要就是机箱的驱动器槽位、连接器部分能够与 2.5 英寸 SATA、SAS 乃至 HDD 机械盘通用。Hyperscale 大型互联网和云服务提供商相对没有这个包袱，针对应用划分出的全闪存服务器机型可以更早试水 E1.S。

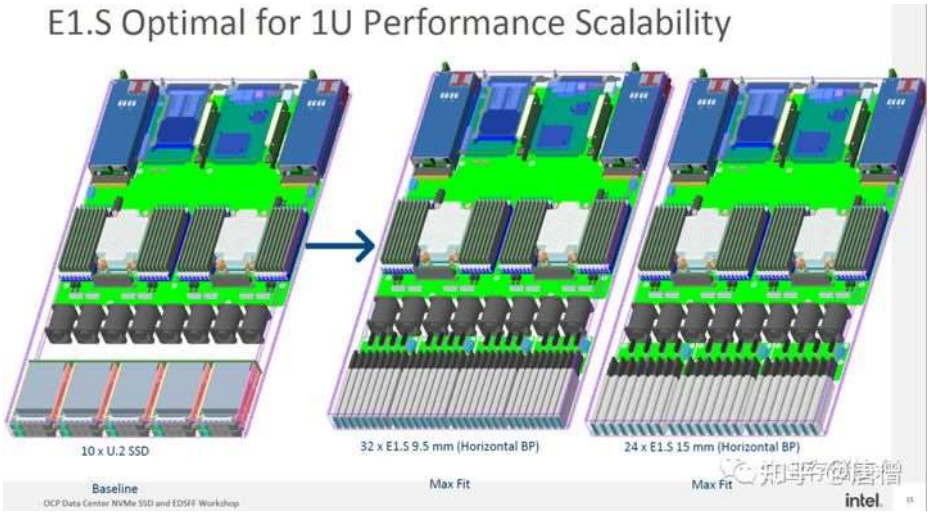
而最终过渡到 PCIe5.0 的时候，U.2 和 M.2 应该是电气上达不到要求了。到时 2U 服务器可能会普遍采用 E3.S；用于服务器扩展的存储/JBOF 使用 E1.L；企业级存储阵列用 E3.S 和 E3.L 都能实现双端口。

PCIe Gen4 和 Gen5 作为 SSD 接口的普及速度可能不算快，毕竟 Gen3 x4 对于许多应用来说也不慢了。上图右侧应该是数据中心 SSD 的出货 Unit 数量，M.2 的占比似乎偏大？不过可以看出从 2022 年往后 U.2/U.3 的占比开始下降，E1.S 和 E1.L 逐渐成为主流，而 E.3 还会晚一些。

上图是 E1.L 的应用场景，这个我想已经不用太多介绍，9.5mm 厚度的 E1.L SSD 在 1U 标准机箱中可以放 32 个。

Enclosure Parameter	5.9mm Device	Device with Heat Spreader (8.01mm)	Device with Symmetric Enclosure (9.5mm)	Device with Asymmetric Enclosure (15mm)	Device with Asymmetric Enclosure (25mm)
Recommended sustained power (W)	12	16	20	20	25
Enclosure Max Inlet air temperature, 950 m to 3050 m (° C)	35 - (1° C for 175 m of elevation gain)	35 - (1° C for 175 m of elevation gain)	35 - (1° C for 175 m of elevation gain)	35 - (1° C for 175 m of elevation gain)	35 - (1° C for 175 m of elevation gain)
Add in card to add in card pitch (mm)	9	11	13	17	26
Recommended Fan Pressure loss across device (Pascal)	83	52	64	40	21
Airflow, average min per device (CFM). 1 CFM = 1.7 m3/h)	1.41 – (0.01 CFM for every 1° C below 35° C inlet temp)	1.71 – (0.06 CFM for every 1° C below 35° C inlet temp)	2.02 - (0.02 CFM for every 1° C below 35° C inlet temp)	1.5 - (0.02 CFM for every 1° C below 35° C inlet temp)	1.5 - (0.02 CFM for every 1° C below 35° C inlet

E1.S SSD 的功耗和散热选项。上面这个图表，其实去掉那几张 SSD 图片[我在去年就给大家列出过](#)，这里也不想再细讲了。



E1.S 针对 1U（存储）性能扩展型服务器优化。对比当前机箱前部 10 个 U.2 热插拔 NVMe SSD 的 Baseline 设计，32 个 E1.S 9.5mm 或者 24 个 E1.S 15mm 的密度提升显而易见。

如上图右侧红圈：EDSFF E3 的连接器具备从 PCIex4 扩展到 x8 或者 x16 的空间

EDSFF E3 SSD 如果按照上图中的放置方向，其高度都是 76mm，E3.S 和 E3.L 的深度分别为 112.75mm、142.2mm，而宽度（厚度）都有标准（7.5mm）和 2x（16.8mm）两种。4 款规格的功耗覆盖了 20-25W 到 70W 的范围。目标用途如下：

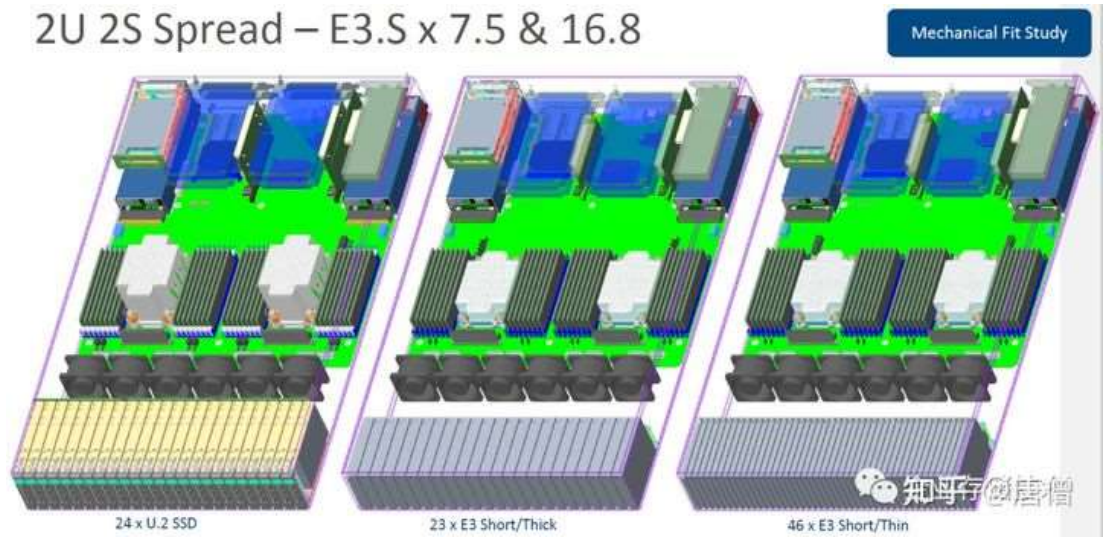
- E3.S：为 NANDSSD 主存储服务器优化；

- E3.S, 2x：支持高功耗设备，如基于 SCM 的 CXL（未来的 Optane SSD，其实叫持久化内存盘更合适）；

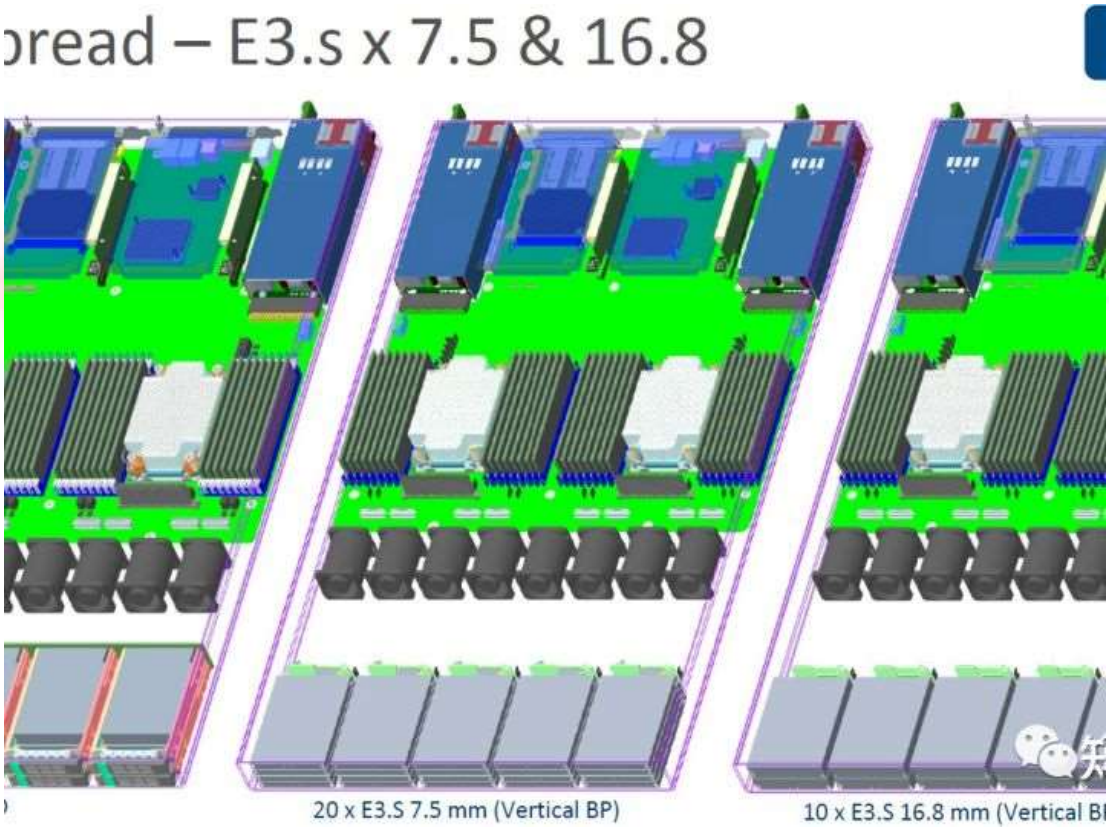
- E3.L：支持高容量 NAND 存储；

- E3.L, 2x：支持像 FPGA 和加速器那样的高功耗设备（PCIe 卡前置）。

注：关于在 PCIe 底层上面跑 CXL 等不同协议，我打算在下一篇中再跟大家讨论。



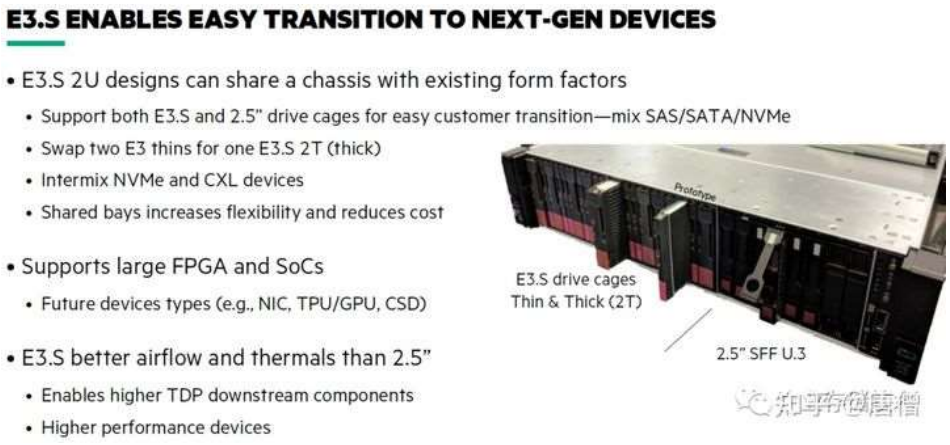
传统 2U 服务器，前置 U.2 SSD 数量最多 24 个，换成 E3.S, 2x/Thick（厚盘）之后可以放 23 个；而 E3.S/Thin（薄盘）则最多可以做到 46 个热插拔位。



E3 SSD 在 1U 机箱中需要横过来放，16.8mm 的厚盘可以放 10 个（与 U.2 数量相同），7.5mm 薄盘则可以放 20 个。

注：Intel 分享的内容就先讲这么多，下面将按照厂商在 Webinar 资料中的出现顺序来写，当然我会挑重点的东西。

HPE 分享：E3.S ENABLES NEXT-GEN DEVICES AND OPEN NVME SSD SPECS



根据 HP 在这里的原型机和介绍，E3.S 2U 设计可以和已有的 2.5 英寸驱动器共享机箱——即混用 SAS/SATA/NVMe（注意背板的连接器不通用），同时还能够：

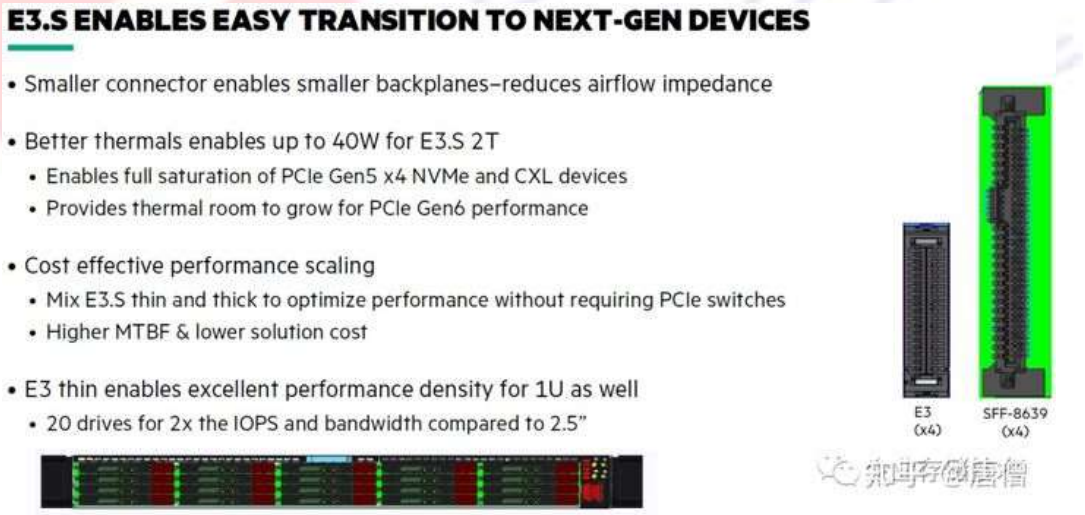
- *切换 2 个 E3 thins 薄盘为 1 个 E3.S 2T（thick）厚盘；*

- *混用 NVMe 和 CXL 设备；*

- *共享仓位提升灵活性和降低成本。*

E3.S 可以支持较大的 FPGA 和 SoC（不过前文中只有 E3.LThick 才有 70W 功耗），未来的设备类型如：NIC 网卡、TPU/GPU、ComputerStorage 计算型存储设备。

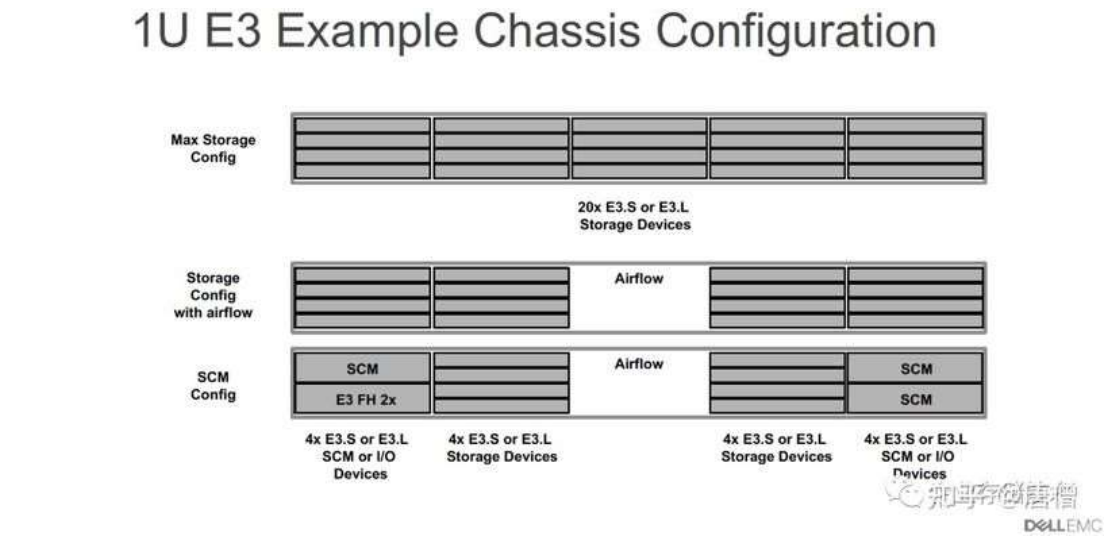
还有一点，就是 E3.S 的空气流动和散热比 2.5 寸驱动器更好，这样可以支持较高性能/TDP 的设备。



同样是 PCIe x4，E3 设备的连接器尺寸比 SFF-8639（同时还要提供向下兼容的 2 个 SAS/SATA 端口）明显小了许多，这样可以把背板做的更小——降低散热的流阻。

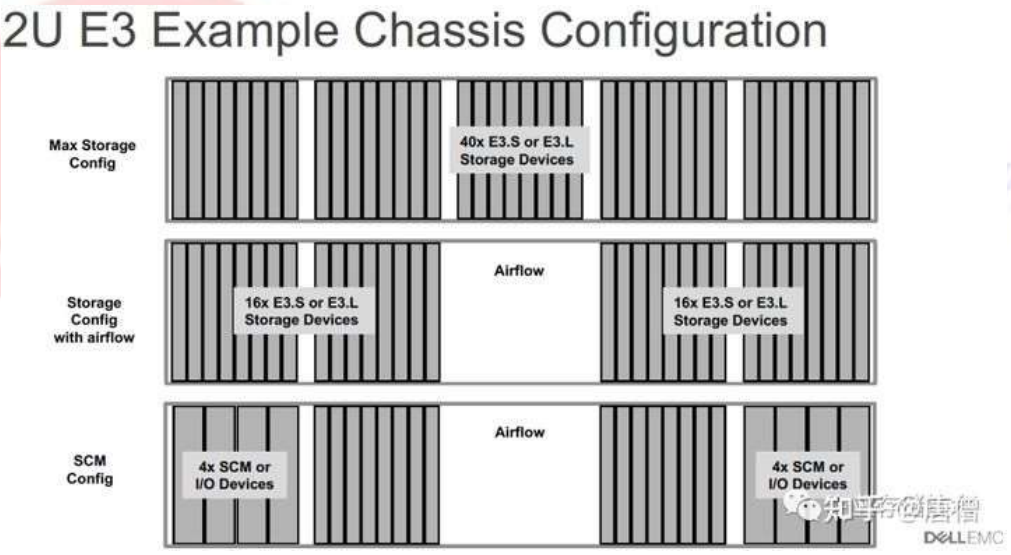
更好的散热，使 E3.S 双倍厚度支持到 40W，从而适配 PCIe Gen5 和未来的 Gen6 性能。

Dell EMC 分享：EDSFF E3 Form Facror，不只是 SSD



记得我在[《1U 双路风冷 350W？点评方升服务器散热设计》]中讨论过流阻对服务器中后部 CPU 等组件的影响。上图就是一种设计参考，比如在 1U 机箱配置 E3.S/E3.L SSD 时空出中间一列用于空气流动，还能保持 16 个 SSD 的密度。

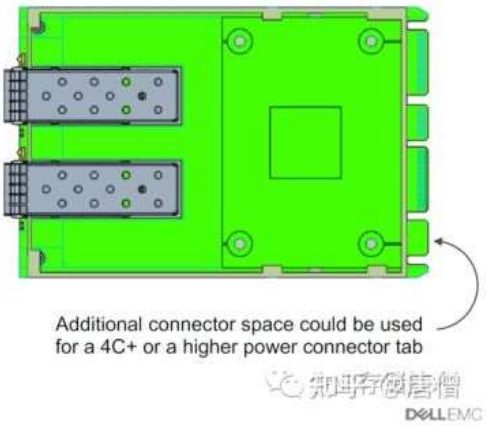
在 SCM（存储级内存）配置中，2 个 E3 薄盘位置可以替换成 1 个 E3 FH 2x 厚盘，除了支持 SCM 之外还可以是 I/O 设备（我看了第一反应就是 OCP 网卡，稍后会讲到）。



2U E3 机箱的情况类似，如果觉得放 40 个 E3.S 或者 E3.L SSD 对服务器进风流阻影响大，可以考虑空出中间，两边各保留 16 个仍有 32 盘的密度。SCM 配置与前面的 1U 机箱同理。

Future Device Types

- Moving the E3 connector to 19.54mm allows for the use of a 4C+ connector used by OCP 3 NIC
- Allows standard networking connectors in an E3 2T form factor
- Allows for potential future higher power devices



如上图，我理解这里应该是把 E3（2T 厚度）的连接器部分增加 19.54mm（4C+），就能允许使用 OCP3 网卡了，也可以支持未来的高功耗设备。

我在《*下一代 Xeon 服务器：Ice Lake-SP 通用云平台设计预览*》介绍过“2 种中板、（前置）OCP 网卡/U.2/EDSSF SSD 灵活设计”。期待未来将网卡等统一到 EDSFF E3，那时服务器设计将更通用模块化。

Samsung、Micron、SK hynix 和 FADU 的 PCIe Gen5 SSD

Samsung EDSFF is ready

Model	PM9A3	PM1743
Cell Technology	V6 TLC	V6 TLC
Interface	PCIe Gen4 1x4	PCIe Gen5 1x4, 2x2
Form Factors	E1.S 9.5mmT, 15mmT, 25mmT E1.L 9.5mmT	E3.S 1T
Endurance	1 DWPD	1 DWPD
Features	 • Optimized performance and latency for hyperscale environments • Improved health monitoring and debugging features • Enhanced security features	 • High performance PCIe Gen5 SSD for enterprise applications • Provides enhanced data encryption and attestation
Recommendation	For hyperscale datacenter, M.2 → E1.S/E1.S	For enterprise Server/Storage, U.2 → E3.S
Schedule	Available now	Available 1H'22

Notes: All product plans and roadmaps are subject to change without notice.

根据三星的介绍，其 PM1743 PCIe Gen5 SSD 将采用 E3.S 1T 尺寸，计划 2022 年 Q2 上市，实际情况为 2023 年 7 月只有非常小批量生产，我们在 2023 年一月份向 Samsung 原厂全款订购的 Samsung E3.S SSD 在 2024 年 3 月份都未能发货。

Micron and OCP: What's Next

Broad form factor range consolidates around demand-driven standards

EDSFF: Accelerating Adoption
Industry sees value in form factor optimization for flash

E1.S dominant EDSFF variant

Consolidation is coming:
Industry can't sustain offering 11 form factors (in addition to long tail of legacy form factors)

Near term: Micron sees focus around E1.S

Longer term: E1.S still dominant, E1.L support for large capacities and E3 growth aligned with PCIe Gen5

在美光的介绍中，M.2 在数据中心的占比没有前文里那么高（这个统计应该是 PB 容量）。正如 E1/E3 在 2022 年之后逐渐替代 U.2/U.3/2.5 的趋势那样，“行业不能维持 11 种 Form Factor，近期的焦点在 E1.S，而长期还会有 E1.L 支持大容量、E3 伴随 PCIe Gen5 增长。”

扩展阅读：《OCP 曝光 NVMe/SAS RAID 卡、U.3 混合背板》

SKhynix Products Supporting DC NVMe SSD Spec

- **SKhynix has been offering products that support the Datacenter NVMe SSD spec**
  - PE8111 E1.L and PE8110 E1.S are developed based on the Datacenter NVMe SSD spec v1.0a and those are being shipped to customers
  - New products coming up next aim to meet DC NVMe SSD Spec V2.0
- **PCIe Gen5 SSDs are expected to be developed in EDSFF form factor and Datacenter NVMe SSD spec**
  - Not only E1.S/E1.L but E3 SSD is being planned for PCIe Gen5 and it will be based on version 2.0

	E1.S 15mm	E1.L 18mm / 9.5mm	E3.x (TBD)
Product	PE8110	PE8111	Next generation
Interface	PCIe Gen4x4	PCIe Gen3x4	PCIe Gen5
Capacity	1920GB ~ 7680GB	15360 ~ 30720GB	TBD
Read / Write Bandwidth	6500 / 4400 MB/s	3550 / 3300 MB/s	TBD
Read / Write IOPS	1100 / 160 KIOPS	750 / 105 KIOPS	TBD

SK hynix 也提到了下一代 E3.xPCIe Gen5 SSD，不过我觉得随着对 IntelSSD 的收购，海力士自己原有的数据中心产品线存在不确定性。

FADU NVMe Datacenter SSDs

FADU OCP SSD Offering	Bravo Gen3x4 7%OP , 2 TB, E1.S	Delta Gen4x4 7%OP 4 TB, E1.S/E3?	Echo Gen5x4 7% OP 8 TB E1.S/E3?
SR in MB/s	3500	7300	14,000
SW MB/s	2700	4600	12,000
RR in KIOPS	800	1490	3500
RW in	100	180	410

Come talk to us

In Real workloads
We get excellent
Random Read in
Mix workloads/
Recovery after Burst/
QOS and Max latency

FADU 这个品牌我还不太熟，这里他们给出了 Gen5 x4 NVMe 数据中心 SSD（7% OP）的性能指标：顺序读 14,x00MB/s、顺序写 12,000 MB/s、随机读 3,500 KIOPS（350 万）、随机写 410 KIOPS（41 万）。

更多资料

未来 EDSFF 规格的设备将不只是 SSD

EDSFF 规范的文档大家可以从 SNIA 网站下载。最后还是老规矩，我把这份 OCP 的参考资料也共享出来：

链接：https://pan.baidu.com/s/1LLbuBRIyLgm3NIr8nyxx3w

提取码：k2x4

扩展阅读：《企业存储技术》文章分类索引（微信公众号专辑）》

注：本文只代表作者个人观点，与任何组织机构无关，如有错误和不足之处欢迎在留言中批评指正。如果您想在这个公众号上分享自己的技术干货，也欢迎联系我：）

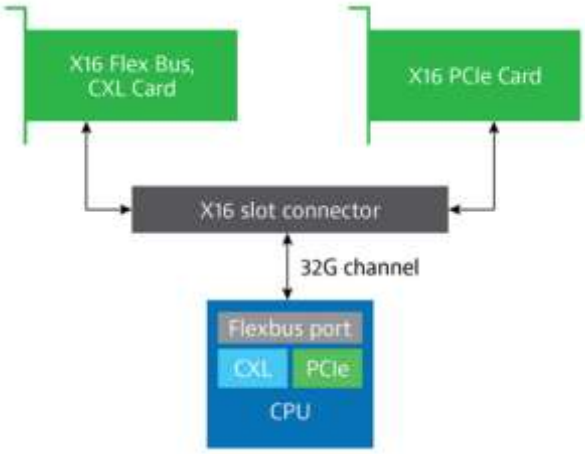
10.8 CXL 协议基础

10.8.1 Compute Express Link 基础

1. What is Compute Express Link?

CXL is a cache-coherent open interconnect standard for high-speed CPU connection to memory and other devices. Compute Express Link leverages the standard PCIe 5.0 physical layer and runs as a supported alternate protocol. By creating a common memory space for connected devices, the CXL standard brings performance advantages for hyperscalers and other advanced applications.

- **Compute express link** utilizes a flexible processor port that can operate in either PCIe 5.0 or CXL modes. Both device classes can achieve data rates of 32 GT/s or up to 64 GB/s in each direction over 16 lanes.
- **The CXL Consortium** was founded in 2019 by nine industry-leading organizations to develop technical specifications, support emerging use case models, and advance CXL technology development and adoption.
- **Artificial intelligence (AI)**, machine learning, and cloud infrastructure are among the applications that benefit most from the from extremely low latency and coherent memory access provided by the CXL interface.



2. How Does CXL Work?

In summary, the Compute Express Link framework establishes coherency between the

memory of the CPU and each connected device. This allows storage resources to be pooled and shared efficiently even as the software stack complexity is reduced. To enable memory pooling, both the host and peripheral device(s) must be CXL-enabled. Data transfer is completed using 528-bit flow control units or “flits”.

- **Single level switching** allows the host to fan out to multiple devices while maintaining high throughput in each direction. Resources including accelerators and any available CXL storage can be dynamically re-assigned as the server workload changes.

- **The CXL 2.0** specification also includes a standardized fabric manager. This ensures a seamless user experience with consistent configurations and error reporting regardless of the pooling type, host, or usage model.

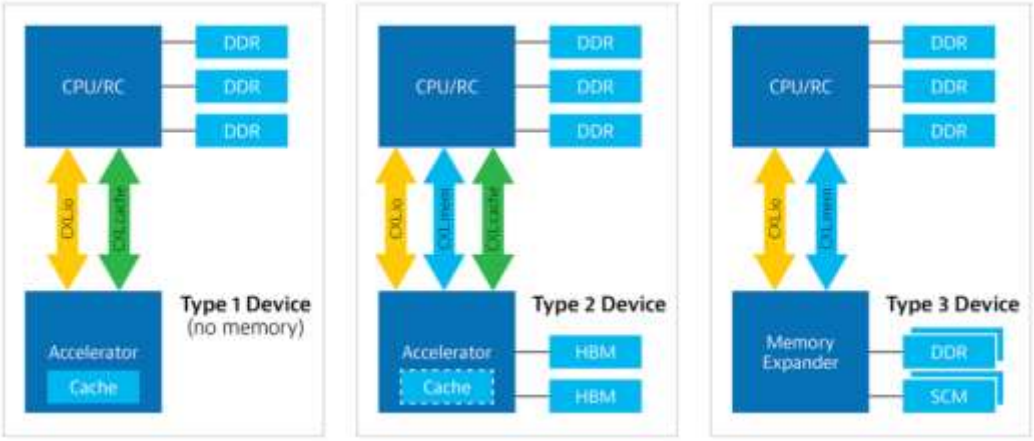
3. Common Compute Express Link Use Cases

As the interface has evolved, unique use cases and applications have led the CXL Consortium to define three discrete device types.

- **Type 1 Devices:** Accelerators and other devices that lack local memory and therefore must rely on the CPU are classified as Type 1. The CXL.io and CXL.cache protocols enable these devices to communicate and transfer memory capacity from the host processor more efficiently.

- **Type 2 Devices:** Products that include their own data storage capabilities but also leverage CPU memory are known as Type 2. All three CXL protocols combine to promote coherent memory sharing between these devices and the CPU.

- **Type 3 Devices:** Memory expanders or devices designed to augment existing CPU memory are classified as Type 3. The CXL.io and CXL.memory protocols enable the CPU to access these external sources with improved bandwidth and latency performance.



4. Compute Express Link Benefits

By streamlining connectivity and resource sharing, CXL technologies provide numerous enhancements that improve high-capacity workload performance while reducing system complexity and cost. These attributes become increasingly valuable as next-generation data centers and emerging technologies drive demand for faster data processing and lower total cost of ownership (TOC).

- **Coherency** enables CXL memory pools to remain consistent with respect to data validity. This allows for faster and more efficient resource sharing between devices and processors.

- **Heterogeneous** architecture, combining processors of varying types and generations, is fully accommodated by the CXL standard. This is especially useful for complex AI neural networks and machine learning systems as elements of the infrastructure evolve.

- **Lower latency** is the result of strategically pooled persistent memory, improved CXL switching efficiency, and standardized memory management. Reduced latency is considered a key enabler of next generation use cases and future PCIe 6.0 adoption.

5. CXL Protocols and Standards

The release of the CXL 1.0 standard in 2019 was a significant milestone marked by CPU access to shared accelerator device memory. Compute express link protocols and standards have continued to improve and expand since this successful debut.

- **CXL 1.1** improved compliance and interoperability aspects of the original standard while maintaining backwards compatibility with release 1.0.

- **CXL 2.0** added switching capabilities for fan-out configurations, resource pooling, and persistent memory support while minimizing the need to overprovision resources. Link-level Integrity and Data Encryption (CXL IDE) were also incorporated to improve security.

CXL Version Features

Function	CXL 1.0	CXL 1.1	CXL 2.0
16-lane, 32GT/s operation	✓	✓	✓
Flex Bus.cxl	✓	✓	✓
Bias based coherency	✓	✓	✓
Three sub-protocols (io, cache, mem)	✓	✓	✓
FLIT based transfers	✓	✓	✓
ARB/MUX management	✓	✓	✓
Enumeration	✓	✓	✓
PM/ASPM	✓	✓	✓
RAS	✓	✓	✓
Retimers	✓	✓	✓
Compliance		✓	✓
Switching			✓
Memory Pooling			✓
Fabric management			✓
Security/IDE			✓

- **Sub-protocols** developed for CXL specification 1.0 have remained consistent throughout the compute express link lifecycle:

  - **CXL.io** is based on the PCIe 5.0 protocol and is used for discovery, configuration, and register access functions. CXL.io must be supported by all compute express link devices in order to function.
  - **CXL.cache** manages interactions between the CPU (host device) and other compute express link enabled devices. This sub-protocol supports the efficient, low latency caching of host memory and direct device access to CPU memory using a request and response process.
  - **CXL.memory** provides modes of access for the host to provision attached device memory using load and store commands. In this configuration, the CPU acts as a master with the compute express link device(s) acting as subordinates.

6. Impact of CXL on Storage

The heterogeneous, open computing model defined by the CXL protocol specification creates a more flexible storage landscape with efficient data movement contributing to lower latency and cost. Beyond the implicit value of pooled coherent memory, the reduction of proprietary memory interconnects compliments the diversity of emerging technology devices.

This shift in storage dynamics will also increase the size of cached memory pools, helping

to meet the temporary storage needs of hyperscale data centers and other large computing enterprises. The additional capacity created through CXL memory pooling can be called upon as needed for high-volume workloads. This storage paradigm shift is consistent with the overall trend towards data center disaggregation and open architecture.

7. CXL and PCIe

PCI Express (PCIe) has become the de-facto high speed serial bus architecture over the past two decades, with point-to-point topology providing high speed links to connected devices. Despite the proficiency of PCIe for bulk data transfer, shortcomings become obvious in larger data center applications. Memory pools remain isolated from one another, which makes significant resource sharing nearly impossible and adds to the latency deficit for new connected devices.

- **PCIe 5.0** is the latest backwards compatible generation of the Peripheral Component Interconnect Express standard. Released in 2019, the 5th generation included a requisite doubling of throughput along with support for alternate protocol deployment that is now being utilized by the CXL interface.

- **Operating** over the PCIe 5.0 physical layer, CXL protocols build upon the versatility of standard PCIe architecture by integrating new memory sharing functionality within the transaction layer.

- **Standard PCIe devices** and CXL software can be supported over the same link. A flexible processor port can quickly negotiate either standard PCIe or alternate protocol CXL interconnect transactions.

- **PCIe 6.0 speed** will continue with the doubling convention of previous generations. This makes the melding of device and system memory an essential consideration for latency reduction and accelerator performance.

8. CXL vs CCIX

Rather than acting as a direct rival technology to CXL, Cache Coherent Interconnect for Accelerators (CCIX®) takes an alternate approach to memory pooling.

- **The CCIX Consortium:** The stated mission of the CCIX Consortium is to develop and promote adoption of an industry standard specification to enable coherent interconnect technologies between general-purpose processors and acceleration devices for efficient heterogeneous computing.

- **CCIX Specification:** The open-source specification defines a peer-to-peer, symmetrical mode of memory coherence and has been developed in parallel (chronologically) with the CXL specification. Storage capacity from multiple devices is pooled together using non-uniform memory access architecture (NUMA). Both the host device and accelerator(s) must support CCIX operation.

- **Alternate Protocols:** CCIX and compute express link each leverage the PCIe 5.0 physical layer and support of alternate protocols. Using FlexBus connectivity, CXL technology defaults to the CPU for cache consistency, while CCIX does not establish a device hierarchy. This distinction has led to measurable speed and latency advantages for compute express link architecture.

10.8.2 CXL 使用场景及概念解读

1. CXL 介绍（一）

一波 教授

本系列主要介绍 2019 年由 intel 推出的开放性互联协议 Compute Express Link (CXL)，包括使用场景以及重要概念。

1. CXL 简介

Compute Express Link (CXL)是业界支持的用于处理器、内存扩展和加速器的高速缓存一致性互连协议。CXL 技术在 CPU 内存空间和附加设备上的内存之间保持一致性，这允许资源共享以获得更高的性能，减少软件堆栈的复杂性，并降低整体系统成本。这使用户能够简单地关注目标工作负载，而不是加速器中多余的内存管理硬件。

CXL 被设计为高速通信的行业开放标准接口，因为加速器越来越多地被用来补充 CPU，以支持人工智能和机器学习等新兴应用。

2020 年 10 月推出了 CXL 2.0 规范，相比 1.1 版本增加了对扇出切换的支持，以连接到更多的设备；为提高内存利用效率和按需提供内存容量的内存池；以及对持久性内存的支持。

2. CXL & PCIe

CXL 是基于 PCIe5.0 的。PCIe 是一种高速串行计算机扩展总线标准，已经使用了很多年。在最近完成的 PCIe 5.0 版本中，CPU 和外围设备能够以每秒 32 千兆次(32GT/s)的速度进行传输。但是，在具有大型共享内存池和许多需要高带宽的设备的环境中，PCIe 具有一些局限性。PCIe 中没有指定支持一致性的机制，不能有效地管理隔离的内存池，也无法有效管理系统中多个设备之间的共享内存。

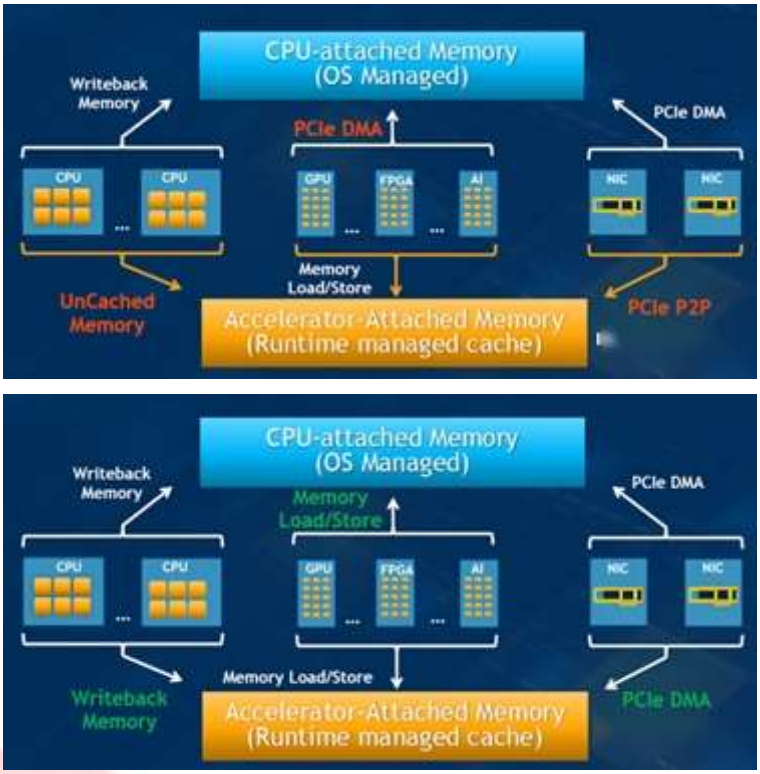


图 1 PCIe（上）和 CXL（下）内存访问方式的对比图

在图 1 中可以看到，PCIe 设备要访问主机内存，一般是使用直接存储器访问技术 DMA，且主机无法缓存 PCIe 设备的数据。在 CXL 中，利用三个子协议：http://CXL.io，CXL.cache 以及 CXL.mem（该部分内容将在第二部分中介绍），为主机和需要共享内存资源的设备（例如加速器和内存扩展器）之间的内存访问提供了低延迟的访问路径以及缓存一致性保证。

3. CXL 应用场景

传统的非一致性 IO 设备主要依赖于标准的生产者-消费者模型，与主机的互动很少。这样的加速器也倾向于在数据流或大型连续数据对象上工作。因此，这些设备通常不需要 CXL 提供的高级功能，而传统的 PCIe 足以作为加速器和主机的连接桥梁。

如图 2 所示，在 CXL 规范中，定义了三种适用于 CXL 协议的设备：

(1)想要在本地缓存 CPU 主存中的数据的设备。在这种情况下，设备仅需使用 http://CXL.io 和 CXL.cache。

(2)加速器上有内存，并且希望 CPU 和加速器之间有相互作用的设备。因此使用 http://CXL.io 协议允许 CPU 发现并配置设备，然后使用 CXL.cache 以允许设备访问 CPU 的内存和使用 CXL.mem 以允许 CPU 访问设备的内存。

(3)内存缓冲区，在这种情况下，需要 http://CXL.io 协议来发现和配置设备，以及 CXL.mem 协议以使 CPU 可以访问连接到内存缓冲区的内存。

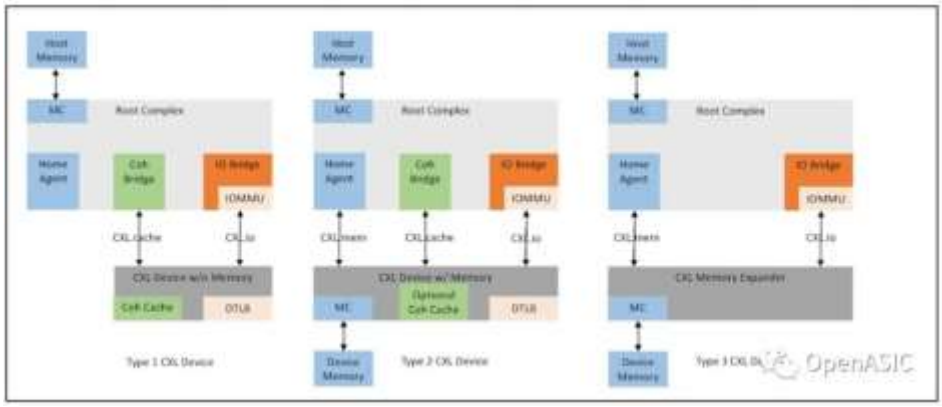


图 2 CXL 适用的三种设备概念图

针对第二种 CXL 设备，基于偏置的一致性模型为设备连接的内存定义了两种偏置模式：主机偏置模式和设备偏置模式。当设备附加的内存处于主机偏置模式时，它在设备上的显示方式与常规主机附加的内存相同。也就是说，如果设备需要访问它，则需要向主机发送一个请求，该请求将解决所请求行的一致性问题。另一方面，当设备连接的内存处于设备偏置模式时，可以确保设备没有被主机缓存的行。这样，设备可以访问它而无需向主机发送任何事务（请求，监听等）。

主机偏置模式和设备偏置模式中，主机和设备的访存方式可如图 3 所示。

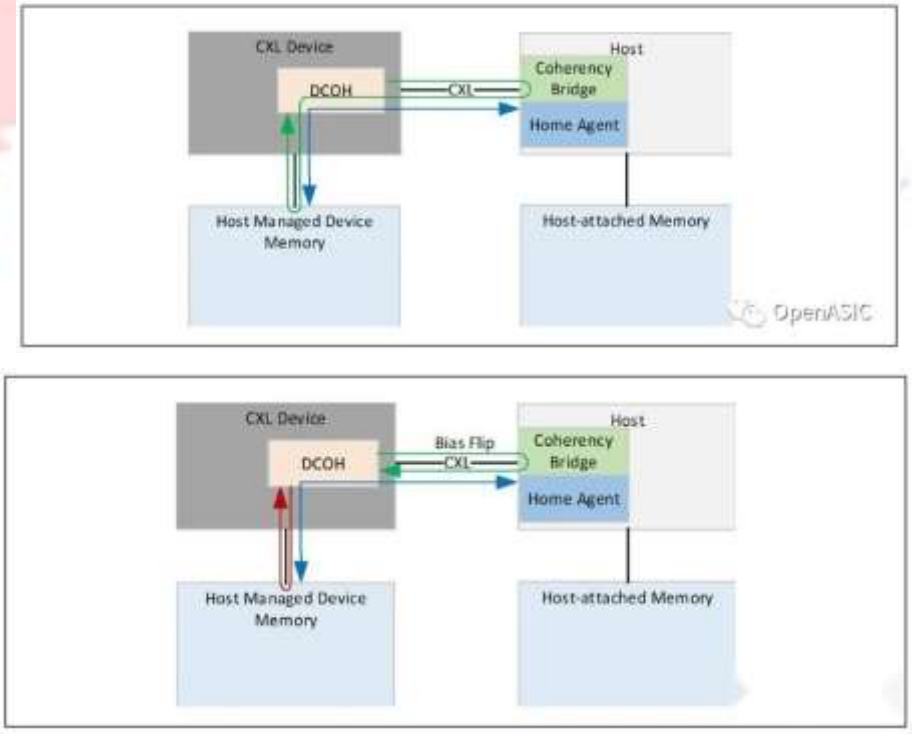


图 3 主机偏置模式（上）和设备偏置模式（下）的访存方式对比

4. CXL 分层概述

类似 PCIe，CXL 分为三层：事务层(Transaction Layer)、链路层(Link Layer)和物理层(Physical Layer)。CXL 的分层概念图如图 3 所示。

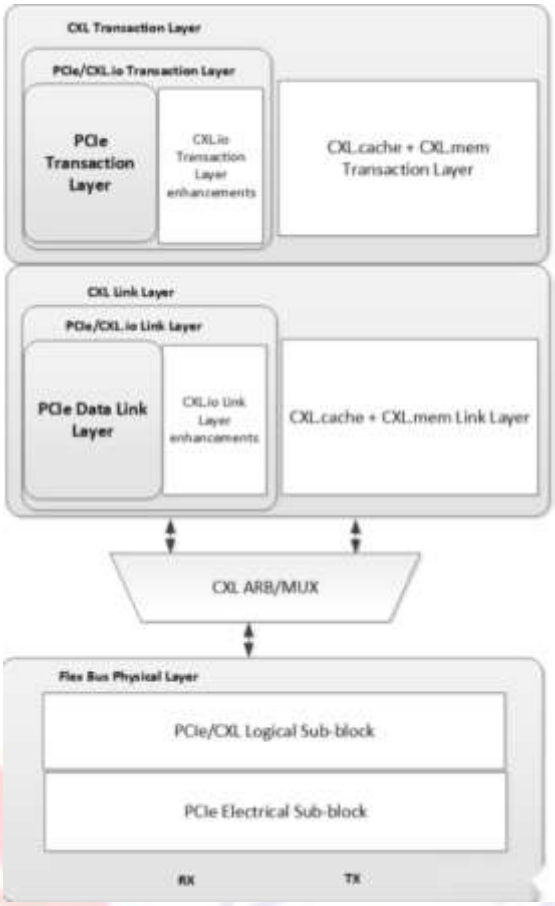


图 4 CXL 分层概念图

为了同时兼容 PCIe 和 CXL，CXL 中提供了 Flex Bus 端口使设备可以在 PCIe 和 CXL 中进行选择。CXL.cache 和 CXL.mem 协议被组合在一起共享一个公共的事务层和链路层，而 http://CXL.io 拥有自己的事务层和链路层。CXL 链路层与 CXL ARB / MUX 交互，从而交织来自两个逻辑流的流量。接下来是物理层，物理层中还包含两个子层，分别是逻辑子层和电气子层。其中逻辑子层可以在 PCIe 模式和 CXL 模式之间切换，而电气子层则遵循 PCIe 规范。在事务层中，http://CXL.io 为 I / O 设备提供了一个非一致性的加载/存储接口。

2. CXL 介绍（二）

本系列主要介绍 2019 年由 intel 推出的开放性互联协议 Compute Express Link (CXL)，包括使用场景、重要概念以及 CXL 对缓存一致性的支持。

本系列包含两部分内容，本文是系列中的第二部分。

1. CXL 子协议概述

CXL 中定义了三个协议，为主机访问服务器和需要共享内存资源的设备（例如加速器和内存扩展器）之间的内存访问和一致性缓存提供了极低的延迟路径。

CXL 的三个子协议简介如下：

- http://CXL.io 协议是 PCIe 5.0 协议的增强版本，可用于初始化、链接、设备发现和枚举以及寄存器访问。它为 I / O 设备提供了非一致性的加载/存储接口。
- CXL.cache 协议定义了主机与设备之间的交互，从而允许 CXL 设备使用请求和响应的方法以极低的延迟高效地缓存主机内存。

- CXL.mem 协议使主机处理器可以使用加载和存储命令访问 CXL 设备的内存。

由于 http://CXL.io 协议与 PCIe 协议非常类似，因此这里只简单介绍。CPU 和设备之间进行交互的相关问题主要在 CXL.cache 协议和 CXL.mem 协议中，这两个协议会在后续进行详细介绍。

2. CXL.cache 协议

CXL.cache 协议允许 CXL 设备使用请求和响应的方法以极低的延迟高效缓存主机内存。此协议将设备与主机之间的交互定义为多个请求，每个请求至少具有一个关联的响应消息，有时还包含数据传输。如图 1 所示，在设备到主机（D2H）和主机到设备（H2D）两个方向上都分别包含三个通道：请求，响应和数据。（Request,Response, and Data）。

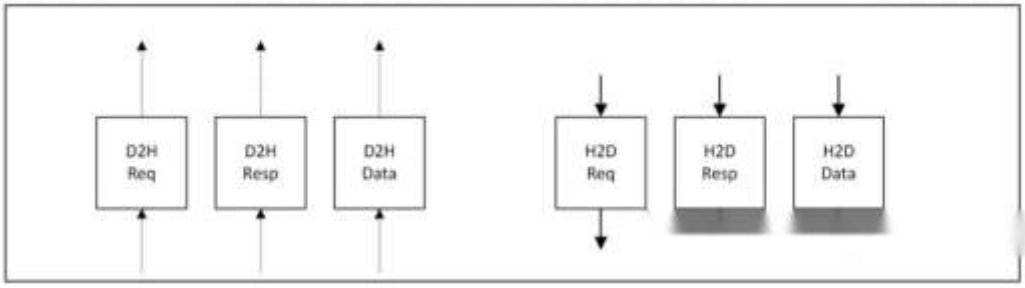


图 1 CXL.cache 通道示意图

D2H

对于设备到主机（D2H）的请求，有四种不同的语义：CXL.cache read，CXL.cache read0，CXL.cache read0 /write 和 CXL.cache write。所有设备到主机 CXL.cache 的事务都属于这四种语义之一，尽管给定语义内每种请求类型的允许响应和限制是不同的。详细分类可见图 2。

CXL.cache. – Device to Host Requests (Sheet 1 of 2)

CXL.cache Opcode	Semantic	Opcode
RdCurr	Read	00001
RdOwn	Read	00010
RdShared	Read	00011
RdAny	Read	00100
RdOwnNoData	Read0	00101
ItoMWr	Read0-Write	00110
MemWr	Read0-Write	00111
CLFlush	Read0	01000
CleanEvict	Write	01001

CXL.cache. – Device to Host Requests (Sheet 2 of 2)

CXL.cache Opcode	Semantic	Opcode
DirtyEvict	Write	01010
CleanEvictNoData	Write	01011
WOWrInv	Write	01100
WOWrInvF	Write	01101
WrInv	Write	01110
CacheFlushed	Read0	10000

图 2 设备到主机的请求分类

以 Rdshared 和 DirtyEvict 为例介绍上述 D2H 请求。

Rdshared 是自设备的完整的缓存行读取请求，用于以共享状态（S）来缓存行。一般来讲，Rdshared 会收到主机向设备（H2D）发送的（GO-S）响应来转换到 S 状态。

DirtyEvict 是向主机发出的从设备驱逐完整的 64 字节修改高速缓存行的请求。通常，DirtyEvict 从主机接收 GO-WritePull，此时设备必须放弃对线路的监听所有权，并按正常方式发送数据。DirtyEvict 请求还向主机保证设备不再包含该行的任何缓存副本。

D2H Response Encodings

Device CXL.cache Rsp	Opcode
RspIHitI	00100
RspVHitV	00110
RspIHitSE	00101
RspSHitSE	00001
RspSFwdM	00111
RspIFwdM	01111
RspVFwdV	10[illegible]

图 3 设备到主机的的响应分类

图 3 介绍了 H2D 响应的分类。这里以 RspIHitSE 为例介绍上述 D2H 响应。

RspIHitSE 表示以共享（S）或者独占（E）状态命中缓存行，并且该行现在无效时提供给监听（H2D 请求）的响应。如果设备为监听返回 RspIHitSE，则主机可以假定该行已从该设备清除。

H2D

主机到设备（H2D）的请求是一种为了保持一致性而进行的监听（snoop）。H2D 请求一共有三种，分别是 SnpData，SnpInv 和 SnpCurr。

- SnpData 的目的是在请求者处以共享或独占状态来缓存行（仅当所有设备均以 RspI 响应时，独占状态才能在请求者处缓存）。这种监听通常是由数据读取请求触发的。接收到此监听的设备必须使所有缓存行无效或降级为共享状态。如果设备保留脏数据，则必须将其返回给主机。

- SnpInv 的目的是在请求者处以独占状态来缓存行。这种监听通常是由写请求触发的。接收到此监听的设备必须使所有缓存行降级为无效状态。如果设备保留脏数据，则必须将其返回给主机。
- SnpCurr 获取该行的当前状态，但不需要更改任何缓存状态。它仅代表 RdCurr 请求发送。如果设备将数据保存为“已修改”状态，则必须将其返回给主机。设备和主机中的缓存状态都可以保持不变，并且主机不应更新其缓存。

主机到设备（H2D）一共有如图 4 所示的 8 种响应。这些响应的格式如图 5 所示，其中 Opcode 编码即图 4 中的 8 种。当编码为 GO（Global Observation）全局观察时，指示 RspData 中的 4 位为 MESI 编码，如图 6 所示。

H2D Response Opcode Encodings

H2D Response Class	Encoding	RspData
WritePull	0001	UQID
GO	0100	MESI
GO_WritePull	0101	UQID
ExtCmp	0110	Don't Care
GO_WritePull_Drop	1000	UQID
Fast_GO	1100	Don't Care
Fast_GO_WritePull	1101	UQID
GO_ERR_WritePull	1111	UQ

图 4 H2D 响应的 8 种类型

CXL.cache - H2D Response Fields

H2D Response	Width	Description
Valid	1	The Valid field indicates that this is a valid response to the device.
Opcode	4	The Opcode field indicates the type of the response being sent. Details in Table 20
RspData	12	The response Opcode determines how the RspData field is interpreted as shown in Table 20. Thus, depending on Opcode, it can either contain the UQID or the MESI information in bits [3:0] as shown in Table 13.
RSP_PRE	2	RSP_PRE carries performance monitoring information for requests that do not receive data. Details in Table 12
CQID	12	Command Queue ID: This is a reflection of the CQID sent with the D2H Request and indicates which device entry is the target of the response.
RSVD	1	
Total	**32**	

图 5 H2D 响应的格式

Cache State Encoding for H2D Response

Cache State	Encoding
Invalid (I)	4'b0011
Shared (S)	4'b0001
Exclusive (E)	4'b0010
Modified (M)	4'b0110
Error (Err)	4'b0100

图 6 H2D 响应的缓存行状态编码

3. CXL.mem 协议

CXL 内存协议称为 CXL.mem，它是 CPU 和内存之间的事务接口，使主机可以使用加载和存储命令访问 CXL 设备的内存。

CPU 中的一致性引擎使用 CXL.mem 请求和响应与内存相连。在这种配置中，CPU 一致性引擎被视为 CXL.mem Master，而 CXL 设备被视为 CXL.memSubordinate。CXL.memMaster 负责 CXL.mem 的请求（读取，写入等），而 CXL.memSubordinate 负责响应 CXL.mem 的请求（数据，完成等）。

当 Subordinate 是加速器时，CXL.mem 假定存在设备一致性引擎（DCOH），该引擎负责解决设备缓存方面的一致性并管理偏置状态。

从 Master 到 Subordinate 的 CXL.mem 事务称为“ M2S”，从 Subordinate 到 Master 的事务称为“ S2M”。在 M2S 事务中，有两种消息类别：

- 没有数据的请求（Req）
- 数据请求（RwD）
- 同样，在 S2M 事务中，也有两个消息类别：
- 没有数据的响应（NDR）
- 数据响应（DRS）

M2S

M2SReq 中的 MetaValue 中定义了三种请求：无效（I）、任意（A）和共享（S），介绍如下：

Invalid（I）：表示主机没有该行的可缓存副本。设备一致性引擎 DCOH 可以使用此信息将行的独占权授予设备。

Any（A）：表示主机可能具有该行的共享，独占或修改后的副本。DCOH 可以使用此信息来表明主机可能要更新行，并且在未先向主机发送请求之前，不应向设备提供该行的副本。

Shared（S）：表示主机最多可以具有该行的共享副本。DCOH 可以使用此信息来解释主机没有该行的独占或修改后的副本。如果设备需要该行的共享或当前副本，则 DCOH 可以提供此副本而无需向主机发送请求。如果设备需要该行的独占副本，则 DCOH 将必须先向主机发送请求。

带数据的请求（RwD）消息类通常包含从 Master 到 Subordinate 的写入。RwD 的操作码指定需要对数据和相关信息执行哪些操作，详见图 7。以 MemWr 为例，这是一种内存写入命令，用于全行写入。

Opcode	Description	Encoding
MemWr	Memory write command. Used for full line writes. If MetaField contains valid commands, perform Meta Data updates. If SnpType field contains valid commands, perform required snoops. If the snoop hits a Modified cacheline in the device, the DCOH will invalidate the cache and write the data from the Host to device-attached memory.	'0001
MemWrPtl	Memory Write Partial. Contains 64 byte enables, one for each byte of data. If MetaField contains valid commands, perform Meta Data updates. If SnpType field contains valid commands, perform required snoops. If the snoop hits a Modified cacheline in the device, the DCOH will need to perform a merge, invalidate the cache and write the contents back to device-attached memory.	'0010
Reserved	Reserved	

图 7RwD 的分类

S2M

NDR 消息类包含从 Subordinate 到 Master 的完成和指示。NDR 根据操作码分为 3 类：Cmp，Cmp-S 和 Cmp-E。以 Cmp 为例，这是一种写回、读取和无效的完成指示。

DRS 消息类包含从 Subordinate 到 Master 的内存读取数据。DRS 只有一种操作码：MemData，表明这个响应是内存读取数据。

部分有效的请求和响应的对应关系如图 8 所示。

M2S Req	Meta Field	Meta Value	SnpType	S2M NDR	S2M DRS	Description
MemRd	MetaD-State	A	SnpInv	Cmp-E	MemData	The Host wants an exclusive copy of the line
MemRd	MetaD-State	S	SnpData	Cmp-S or Cmp-E	MemData	The Host wants a shared copy of the line
MemRd	No-Op	NA	SnpCur	Cmp	MemData	The Host wants a non-cacheable but current value of the line
MemRd	No-Op	NA	SnpInv	Cmp	MemData	The Host wants a non-cacheable value of the line and the device should invalidate the line from its caches
MemInv	MetaD-State	A	SnpInv	Cmp-E	NA	The Host wa... line without data

图 8 部分有效的请求和响应的对应关系

4. 支持缓存一致性的写流程示例

写流程分为三个步骤：申请占有权、默写和逐出缓存行。

如图 9-1 所示，第一步是申请缓存行的占有权，也就是要将缓存行状态从无效（I）状态转换到独占（E）状态。具体操作如下：CXL 设备向 HA 发出 RdOwn 请求，HA 向 Memory Controller 发出 MemRd 请求进行内存数据的读取的同时会向 peercache 发出监听信号 SnpInv 使 peercache 从共享（S）状态转换为无效（I）状态。当 peercache 完成转换后会向 HA 发出 RspIHitSE 响应，表示已经完成从 S 状态到 I 状态的转换，此时可以结束监听。当 Memory Controller 也将读取得到的数据返回给 HA 后，HA 会向 CXL 设备发出 GO-E 响应（这是一种 H2D 响应），然后再单独发送 Data 给 CXL 设备。

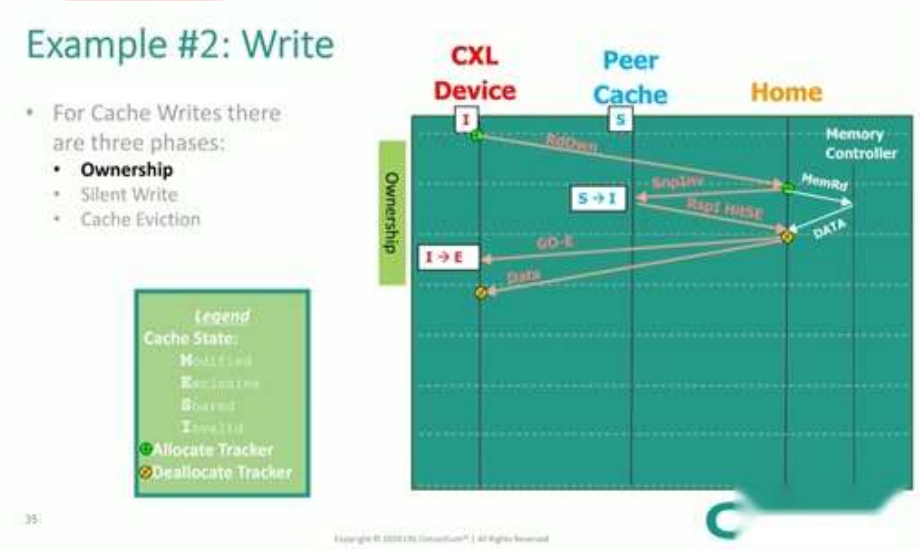


图 9-1 申请占有权流程示例

如图 9-2 所示，申请占有权之后的下一步是默写，这意味着 CXL 设备可以完成修改操作而不需要告知主机。这一步之后，该缓存行在 CXL 设备中的状态将由独占（E）状态转换为修改（M）状态。

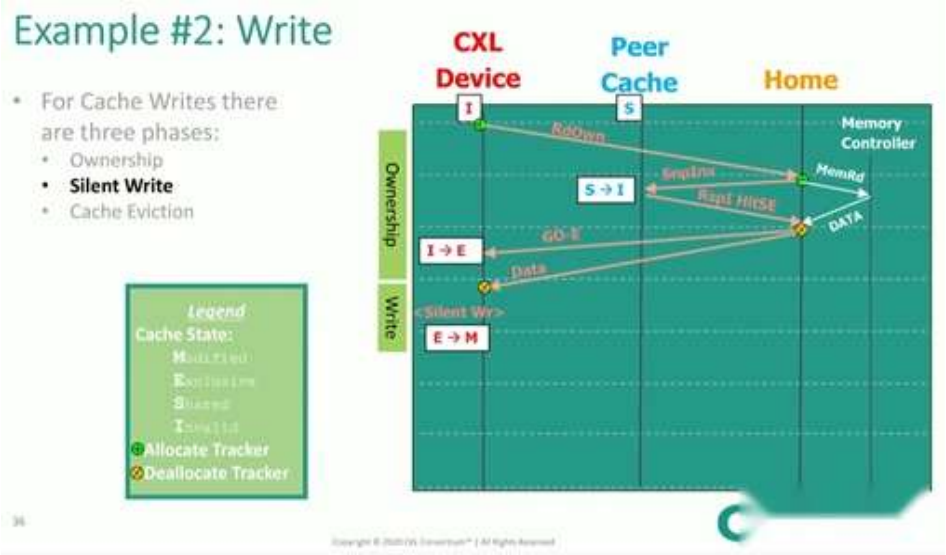


图 9-2 默写流程示例

如图 9-3 所示，当设备完成对缓存行的修改后，需要将该缓存行重新写入内存中。此时，CXL 设备需要向 HA 发送一条 DirtyEvict 请求，该请求收到 GO_WritePull 响应后可以将数据发往 HA，再由 HA 向 Memory Controller 发出 MemWr 请求，这是一种 M2S RwD，用于全行内存数据写入。在完成写入请求后，Memory Controller 向 HA 发送 cmp 响应，这是一种 S2M NDR，表示请求已完成。至此，数据的写入和缓存行逐出就完成了。

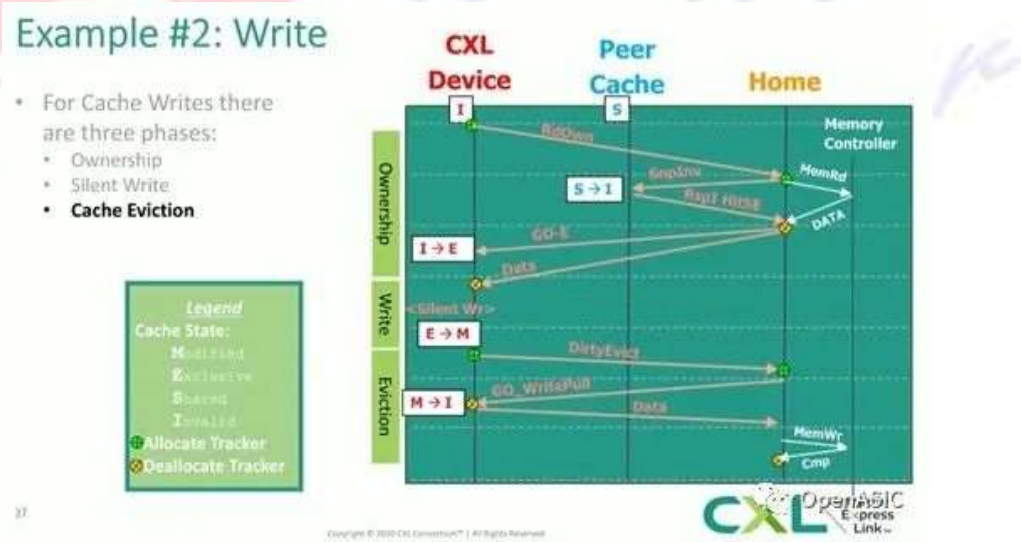


图 9-3 逐出缓存行流程示例

参考资料：

[1]CXL Specification 1.1

[2]PCI Express® Base Specification Revision 5.0

[3]"Compute Express Link™ (CXL™): Exploring Coherent Memory and Innovative Use Cases" (https://www.youtube.com/watch?v=It1_mHsor9g)

关注我们

实验室网站：http://viplab.fudan.edu.cn/

OpenASIC 官方网站：http://www.openasic.org

10.9 PCIe Retimer 基础

10.9.1 What is PCIe Retimer?

A PCIe Retimer is usually implemented as an integrated circuit (IC) chip that can be used, when placed on a PCB, to extend the length of a PCIe bus. It is particularly used it has to pass through a connector to a cable or to another PCB and then to another PCB (i.e. mid-plane or back-plane layouts). The discontinuities caused by the interconnect, PCB/cable changes, etc. produce reflections and increase inter-symbol-interference. These signal challenges will cause the PCIe signal to be too poor at the end point to be received without errors (or a high risk of errors), without some active circuitry to work past those discontinuities. That active ciruictry is the Retimer.

It takes as inputs a PCIe signal and outputs a re-generated signal as if it were a fresh PCIe device, in both directions.

This means it re-establishes a new PCIe link going forth, including re-training and establishing proper equalisation. To be more specific, the physical layer protocol and LTSSM training, including equalisation training, is terminated at the Retimer, such that the communication link is broken into two completely independent physical segments, on each side of the Retimer (upstream and downstream segments). The retimer can perform an optimal equalisation for each segment between the retimer and its link partner.

Retimers can be compared with repeaters and redrivers.

A redriver is an analog only device. It boosts the signal to counteract the attenuation caused by the interconnect. However, this means it could also boost the noise, so using more and more redrivers will eventually lead to the noise overtaking the signal.

A retimer, to summarise from above, is a digital and analog device. It receives the signal and extracts the digital part of it, then it regenerates it as a separately trained link, in both directions. Therefore, the noise (and other imperfections such as jitter), that was originally present, will be eliminated, it is like a fresh start from the re-generated signal.

In some applications, a redriver will be appropriate, in others, a retimer will be required. Generally, a retimer is the superior solution and can cover a super-set of where a redriver

may be used.

The term repeater is best not used in the context of PCIe, because it is less clear whether its function is more like a retimer or a redriver. It could be both, it could be in between, you could say there are two kinds of repeaters, a redriver and retimer, but it is probably clearer to just say either redriver or retimer. Having said that, the term repeater is still used in the context for other types of protocols, to refer to any device that helps extend the signal length.

10.9.2 PCI Express® Retimers vs. Redrivers: An Eye-Popping Difference

Retimers and redrivers have enabled longer physical channels in servers and storage systems since Peripheral Component Interface (PCI) Express (PCIe®) 3.0 was first introduced almost 10 years ago. Now that PCIe 4.0 is ramping up and PCIe 5.0 is just around the corner, how do these reach extension tools stack up in the face of new challenges in high-speed connectivity?

Redrivers vs Retimers

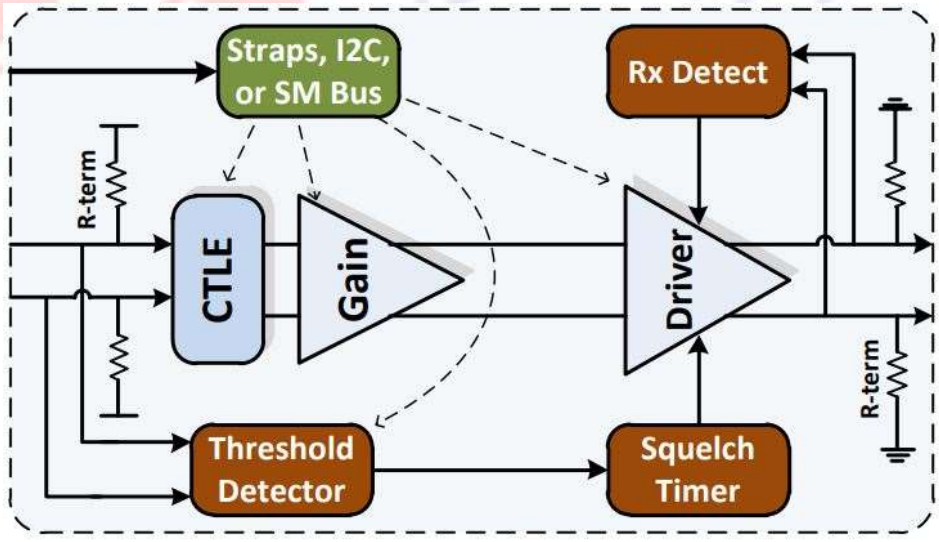


Figure 1: Redriver block diagram [1]

A **redriver** is a mostly analog reach extension device designed to boost the high-frequency portions of a signal to counteract the frequency-dependent attenuation caused by the interconnect: the central processing unit (CPU) package, system board, connectors and so on. A redriver's data path typically includes a continuous time linear equalizer (CTLE), a wideband gain stage and a linear driver. In addition, redrivers often have input loss-of-signal threshold and output receiver (Rx) detection capability. Figure 1 illustrates a typical redriver block diagram.

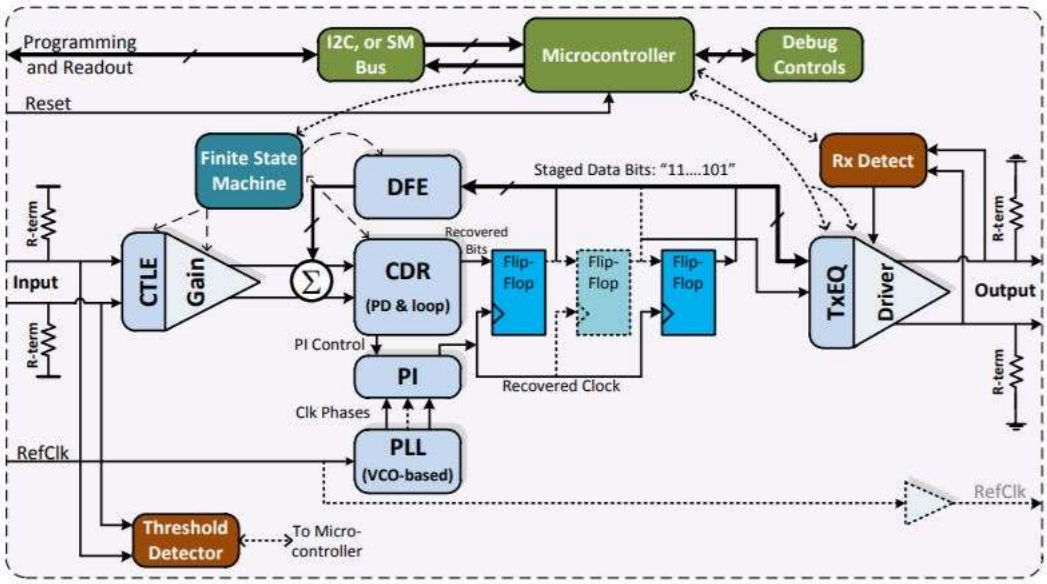


Figure 2: Retimer block diagram [1]

A **retimer** is a mixed signal analog/digital device that is protocol-aware and has the ability to fully recover the data, extract the embedded clock and retransmit a fresh copy of the data using a clean clock. In addition to the CTLE and wideband gain stages also found in a redriver, retimers contain a clock and data recovery (CDR) circuit, a decision feedback equalizer (DFE) and a transmit (Tx) finite impulse response (FIR) driver. Finite state machines (FSMs) and/or a microcontroller typically manage the automatic adaptation of the CTLE, wideband gain, DFE and FIR driver, and implement the PCIe® link training and status state machine (LTSSM). Figure 2 illustrates a typical retimer block diagram.

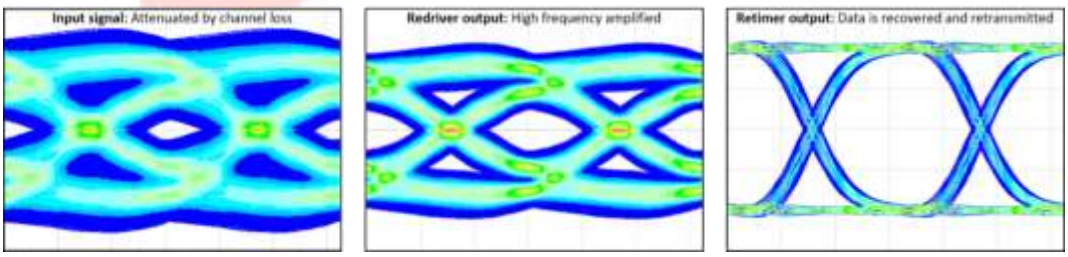


Figure 3: Example of an eye attenuated by a channel (left), the eye after a redriver (middle) and the eye after a retimer (right)

In simple terms, a redriver amplifies a signal, whereas a retimer retransmits a fresh copy of the signal. Figure 3 illustrates this and shows how an attenuated eye opening is boosted by a redriver and completely regenerated by a retimer.

The PCIe 4.0 specification took the unprecedented step of formally defining the terms “retimer,” “redriver” and the superset term “repeater,” all of which are types of extension devices or components whose purpose is to extend the physical length of a link. The definitions are:

- **Repeater**: An imprecise term for an extension device [2]. (This term causes confusion … please don't use it!)

- **Redriver**: A non-protocol-aware software-transparent extension device [2].

- **Retimer**: A physical layer protocol-aware, software-transparent extension device that forms two separate electrical link segments [2].

Use Cases for Retimers and Redrivers

Reach extension devices are necessary whenever the channel – the electrical path between the root complex (RC) and endpoint (EP) – is longer than the PCIe specification allows. The specification defines the maximum channel length in terms of insertion loss at the Nyquist frequency (an informative specification, but easy to validate) and in terms of a reference receiver's ability to sufficiently equalize and recover the data assuming a worst-case link partner transmitter (a normative specification, but time-consuming to validate). Suffice it to say, at PCIe 4.0 speeds, reach extension devices are necessary for:

- Multiconnector topologies.

- Cabled topologies.

- Single-connector add-in card (AIC) topologies with baseboard channels longer than 9.5 inches.

Figure 4 shows an example of a two-connector "riser card" topology, which ordinarily would exceed the PCIe 4.0 loss budget of 28 dB. A redriver or retimer will enable reliable, error-free communication between the RC and EP. But how do you choose which one is the right tool for the job? Well, it helps to know more about the fundamental differences in their capabilities.

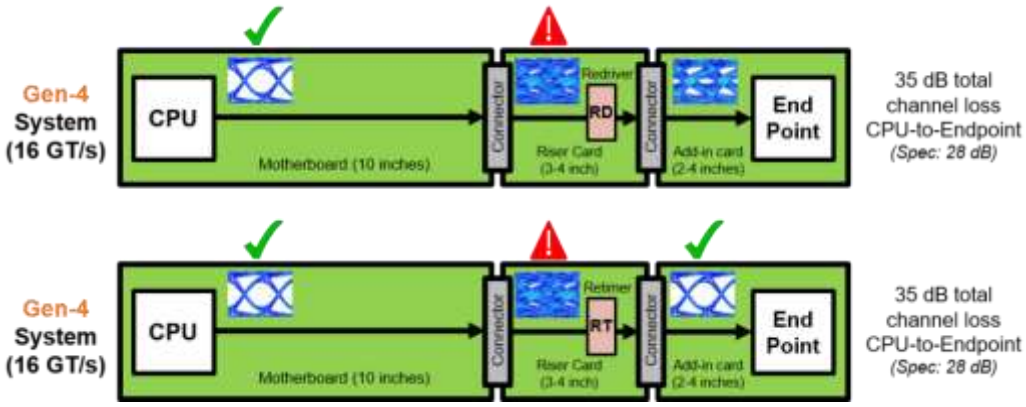


Figure 4: Example of redriver (top) and retimer (bottom) used in a two-connector topology

Comparing Retimer and Redriver Capabilities

Not all redrivers and retimers are the same. There are many distinctions between the two, which are universally true for all PCIe reach extension devices. For example:

Retimers actively participate in the PCIe protocol; redrivers do not.

The PCIe base specification spells out how and to what extent retimers participate in the protocol during Detect, Recovery, L0 and other LTSSM states. Equalization to the L0 and L1 link states requires value-added functionality from the retimer (handshakes, timeouts, bit manipulation, etc.). Redrivers are unaware of and unparticipating in the protocol. If the link works reliably the first time, that's great! But if the link experiences marginality of any sort, it becomes exceedingly difficult to pinpoint whether the problem is physically before the redriver or after it, since the redriver's role in link formation is undefined and unknown to its link partners.

Retimers reset the jitter and insertion loss budgets; redrivers do not.

A retimer's CDR fully recovers the data stream and retransmits it on a clean clock. Starting with a fresh copy of the data enables the extension of the channel to twice the original specification. Without a CDR, the best a redriver can do is attenuate (not reset) the data-dependent jitter (DDJ) caused by intersymbol interference (ISI). A redriver cannot attenuate uncorrelated or random jitter (RJ). In fact, a redriver will always add to RJ due to its own device thermal noise in a root-mean-square (RMS) manner [1].

Retimers have a DFE; redrivers do not.

A DFE compensates for reflections in the channel response caused by impedance discontinuities in board vias, connectors and package socket-board interfaces. The nice thing about a DFE is that it is unaffected by crosstalk. The DFE equalizes just as well in the presence of crosstalk, and once the data is sampled by the retimer's CDR, crosstalk is eliminated for good. Redrivers use a CTLE that boosts both the signal and the noise [1]. Crosstalk is not eliminated or even attenuated through a redriver; in fact, it gets amplified.

Retimers automatically adapt their receive and transmit equalizers to match the characteristics of the channel and the link partner's needs; redrivers do not.

A retimer will examine the signal it receives and adjust the CTLE and DFE to minimize its own bit error rate (BER). Likewise, the retimer's transmitter will adjust its de-emphasis and pre-shoot equalization to minimize the link partner's BER according to PCIe equalization

protocol. A redriver, conversely, operates with a static equalizer setting. The optimal setting (which can be different for every channel in the system) is often painstakingly selected following an exhaustive search in Input/Output Buffer Information Specification (IBIS) algorithmic modeling interface (AMI) simulations and again in lab testing – a process fondly referred to as “tuning.”

Retimers have built-in features to help diagnose link issues (both electrical and protocol); redrivers do not.

Retimers have tools for assessing the electrical performance (internal eye monitors, pattern generators, pattern checkers) and protocol performance (link state history monitors, timeout adjustments). Redrivers cannot offer such diagnostic features because they are neither protocol-aware nor aware of the actual data passing through. Redrivers do not know what state the link is in.

Retimers correct for lane-to-lane skew; redrivers do not.

PCIe has a tight requirement on the physical skew between lanes on a board (1.6 ns for PCIe 4.0), typically caused by mismatches in channel routing length [3]. Retimers are required to compensate and reset any lane-to-lane skew, effectively doubling the specification budget. Redrivers cannot compensate for lane-to-lane skew, and what’s worse is that they may degrade the skew depending on how symmetric the redriver package is across all lanes.

Retimers can be placed anywhere between two PCIe-compliant channels; redrivers cannot.

By definition, retimers extend the total PCIe channel reach by two times the specification. A redriver’s reach extension, however, depends on where it is placed in the channel – how much loss is before the redriver versus how much is after [1]. The specific placement of a redriver must be carefully determined by IBIS-AMI simulation and experimentation. Too close to the root complex transmitter, and the redriver’s CTLE will enter nonlinear operation and will have limited benefit. Placed too far from the transmitter, the redriver’s device noise may significantly degrade the signal-to-noise ratio (SNR) of the data signal. It’s not all bad news for redrivers. They do have lower power consumption and lower input-to-output latency compared to retimers. But if the link does not form in the first place or if the BER is too high, none of that matters!

Property	Retimer	Redriver
PCIe Protocol Participation	Protocol-aware	Protocol-unaware

Jitter Reduction	Resets entire jitter budget (DDJ, RJ, etc.)	Attenuates DDJ; amplifies RJ
Equalization Capabilities	CTLE, DFE, Tx FIR	CTLE
Adaptation	CTLE, DFE and Tx FIR automatically adapt to the channel	CTLE setting must be hand-selected based on simulation/experimentation
Diagnostics Capabilities	Receiver margining, eye diagram, eye width/height measurement, link state debugging information	None to speak of
Lane-to-Lane Skew Compensation Capabilities	Resets entire skew budget	Does not reset skew budget; may increase total skew
Placement	Anywhere with PCIe-compliant channels on the input and output sides	Not too close to the source transmitter, but not far away either
Usage in *Closed* Systems (i.e., systems where all endpoints are known and validated before release of the system)	Recommended; sanctioned use case in PCIe base specification	Highly discouraged; use at your own risk after extensive simulation and testing [1]
Usage in *Open* Systems (i.e., systems designed to be interoperable with any PCIe-compliant AIC)	Recommended; sanctioned use case in PCIe base specification	Not recommended / discouraged [1]

Table 1: Comparison of retimer and redriver capabilities and usage

Outlook for PCIe 4.0 Systems

Looking ahead to the upcoming PCIe 4.0 systems, all signs are pointing to an increased need for reach extension devices – and retimers in particular – due to several trends and challenges:

- CPUs have more PCIe lanes per socket (>100 in some cases [4]) compared to PCIe 3.0. This leads to a greater number of PCIe slots and riser cards, denser routing, and an increased use of multiconnector topologies.
- PCIe is shifting from an I/O bus to a multipurpose system interconnect. This means that more servers will be designed to be modular, allowing an array of compute, storage and networking resources to plug in to an increasing number of PCIe slots. This type of open, “plug anything in and it will work” server architecture requires a reach extension solution that is PCIe compliant with plug-and-play interoperability.
- The disaggregation of resources such as modular servers, storage trays and accelerator trays is pushing endpoints physically away from CPUs, requiring cables or carrier cards to connect everything together. These longer physical topologies will increasingly need reach extension devices.

- Systems are adopting a variety of interconnect styles – M.2, optical/copper link (OCuLink) cables and so on – all of which have unique lane-to-lane skew and crosstalk challenges that must be handled by the reach extension solution.
- PCIe 5.0 is coming right on the heels of PCIe 4.0, and pent-up demand across the industry for higher bandwidth will cause PCIe 4.0 to be short-lived. System designers are looking for a reach extension solution that can easily and quickly scale from PCIe 4.0 to PCIe 5.0.

In the end, system designers benefit from having multiple options for reach extension solutions. The exact performance requirements, physical constraints and cost targets for a server or storage system will guide the decision-making process, and the industry will benefit from knowing the trade-offs between retimer and redriver devices.

Citations

[1] Samaan, S., Froelich, D., and Johnson, S. (2015). High-Speed Serial Bus Repeater Primer: Re-driver and Re-timer Micro-Architecture, Properties, and Usage [White paper]. ***https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/serial-bus-white-paper.pdf***

[2] PCI Express Base Specification Revision 5.0 Version 0.9, 2019. ***https://pcisig.com/***

[3] PCI Express Card Electromechanical Specification Revision 4.0 Version 0.9, 2019. ***https://pcisig.com/***

[4] Why AMD EPYC Rome 2P Will Have 128-160 PCIe Gen4 Lanes and a Bonus. ***https://www.servethehome.com/why-amd-epyc-rome-2p-will-have-128-160-pcie-gen4-lanes-and-a-bonus/***

11. SSD/服务器/存储测试转接卡以及延长线等夹具速查手册

实验室常用 PCIe Gen 5 主机卡，转接卡和延长线图解速查

随着国内越来越多的公司开始进行 PCIe Gen 5 相关产品的开发、验证和测试，例如各类芯片，模块，插卡，主机、网络、存储系统等，针对 PCIe Gen4/5/6 总线的问题诊断分析和测试工具也需要提前提到议事日程。

本文针对实验室测试中各种高品质的支持 PCIe Gen5 信号质量的常用主机卡（switch card），Retimer 卡，各类接口转接卡，转接线，延长线进行了图文介绍和产品描述，同时提供了产品价格供参考。

本文推荐的产品都是在全球业内各大知名芯片设计、系统集成设计公司获得普遍使用的针对 PCIe Gen 5 总线测试搭建环境常用的各种工具，这些产品广泛适用于当前国内从事计算、网络、存储、SSD、AI、大数据、GPU/DPU/MaPU 等数据加速产品，SmartNIC，移动通讯，嵌入式系统设计，汽车电子等诸多领域的 PCIe Gen 5 高速总线的开发和测试过程中。

11.1 PCIe GEN5 转接卡/适配卡

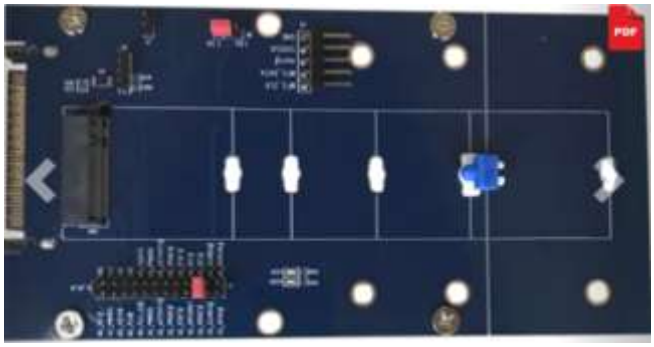
11.1.1 PCIe GEN5 U.2 ADAPTERS

PCIE GEN5 X4 TO U.2 VERTICAL ADAPTER

Part #: PCI5-AD-x439-01V-U2

Description: PCIe Gen5 Vertical x4 slot – U.2 adapter

11.1.2 PCIe GEN5 M2/U2 ADAPTERS



PCIE GEN5 X4 U.2 TO M.2 ADAPTER

Part #: PCI5-AD-U2M2-04

Description: PCIe Gen5 capable x4 U.2 to M.2 Adapter

11.1.3 PCIe GEN5 M2/AIC ADAPTERS

PCIE GEN5 X4 AIC TO M.2 ADAPTER

Part #: PCI5-AD-x4M2-04

Description: PCIe Gen5 capable x4 AIC to M.2 Adapter

11.1.4 PCIe GEN5 U.3 ADAPTERS

PCIE GEN5 X4 TO U.3 VERTICAL ADAPTER

Part #: PCI5-AD-x439-01V-U3

Description: PCIe Gen5 Vertical x4 slot – U.3 adapter

11.1.5 PCIe GEN5 EDSFF ADAPTERS

PCIE GEN5 X8 SLOT TO E3 EDSFF VERTICAL ADAPTER

Part #: PCI5-AD-x8EDSFF-E3V-01/E3

Description: PCIe Gen5 capable x8 slot – x8 EDSFF E3.S/L and E3.S.2T/E3.L.2T. Supports x4 as well

E1A

PCIE GEN5 X8 SLOT TO E3 EDSFF VERTICAL ADAPTER

Part #: PCI5-AD-x8EDSFF-E3V-01/E1A

Description: PCIe Gen5 capable x8 slot – x8 EDSFF E1.S 5.9mm. Supports x4 as well

E1B

PCIE GEN5 X8 SLOT TO E3 EDSFF VERTICAL ADAPTER

Part #: PCI5-AD-x8EDSFF-E3V-01/E1B

Description: PCIe Gen5 capable x8 slot – x8 EDSFF E1.S 8.01mm and 9.5mm. Supports x4 as well

Part #: PCI5-AD-U2E1.S

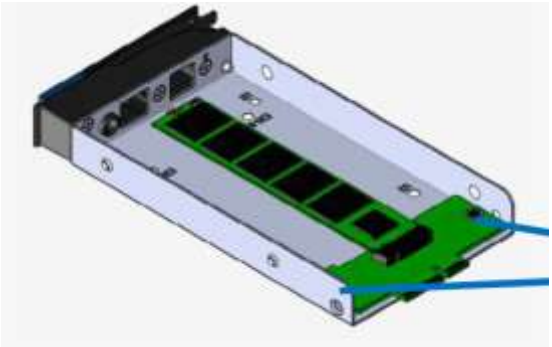
Description: Gen5 PCIe Gen5 E1.S to U.2 Adapter

Part #: PCI5-E3-U2-INT

Description: Gen5 PCIe U.2 Paddle Card for use with the Gen5 PCIe 8 bay JBOF

Part #: PCI5-E3-U3-INT

Description: Gen5 PCIe U.3 Paddle Card for use with the Gen5 PCIe 8 bay JBOF



Part #: PCI5-AD-E3M2-INT

Description: Gen5 PCIe M.2 Paddle Card for use with the Gen5 PCIe 8 bay JBOF

11.1.6 PCIe GEN5 OTHER ADADPTERS

PCIE GEN5 M.2 PADDLE CARD

Part #: PCI5-AD-E3M2-INT

Description: Gen5 PCIe M.2 Paddle Card for use with the Gen5 PCIe 8 bay JBOF

PCIE GEN5 U.3 PADDLE CARD

Part #: PCI5-E3-U3-INT

Description: Gen5 PCIe U.3 Paddle Card for use with the Gen5 PCIe 8 bay JBOF

PCIe Gen5 E3 to M.2 2×2 Adapter PCIE GEN5 E3 TO M.2 2×2 ADAPTER

Part #: PCI5-AD-E3M2-2×2

Description: PCIe Gen5 E3 EDSFF to M.2 2×2 adapter

PCIe Gen5 x16 QSFP-DD to x16 AIC Adapter

Part #: PCI5-AD-QDDX16

Description: PCIe Gen5 x16 AIC w/2, x8 QSFP-DD receptacles adapter. Includes metal base for stability.

PCIE GEN5 X16 LANE REASSIGNMENT ADAPTER KIT

Part #: PCI5-AD-x16LS-KIT

Description: PCIe Gen5 x16 Lane Reassignment Adapter Kit

PCIE GEN5 X16 LANE REVERSAL ADAPTER

Part #: PCI5-AD-x16LR

Description: PCIe Gen5 x16 Lane Reversal Adapter (Lane 15 – 0)

PCIE GEN5 X4 TO X16 LANE REDUCER

Part #: PCI5-AD-x4-x16

Description: PCIe Gen5 x4 to x16 Lane Reducer

PCIE GEN5 X8 LANE REASSIGNMENT ADAPTER

Part #: PCI5-AD-x8LS

Description: PCIe Gen5 x8 Lane Reassignment Adapter

PCIE GEN5 X8 TO X16 LANE REDUCER

Part #: PCI5-AD-x8-x16

Description: PCIe Gen5 x8 to x16 Lane Reducer

11.2 PCIe GEN5 转接线/延长线

11.2.1 GEN5 MCIO CABLES

PCIE Gen 5 MCIO x4(SFF-TA-1016) 38P to MCIO x4 (SFF-TA-1016) 38P, ultra low loss 29AWG wire

Part #: MCIO5-4XL-4XL-0.5M

Description: Gen 5 MCIO x4(SFF-TA-1016) 38P to MCIO x4 (SFF-TA-1016) 38P, ultra low loss 29AWG wire, 0.5M

Part #: MCIO5-4XL-4XL-1M

Description: Gen 5 MCIO x4(SFF-TA-1016) 38P to MCIO x4 (SFF-TA-1016) 38P, ultra low loss 29AWG wire, 1M

MCIO x4 to EDSFF 1X4 cable assembly with Gen6 wires and tighter impedance control

Part #: MCIO5-4XSC-EDSFFLL-1×4-0.5M

Description: MCIO x4 to EDSFF 1X4 cable assembly with Gen6 wires and tighter impedance control, 0.5M

Part #: MCIO5-4XSC-EDSFFLL-1×4-0.5M

Description: MCIO x4 to EDSFF 1X4 cable assembly with Gen6 wires and tighter impedance control, 1M


PCIE GEN5 MCIO 4X (SFF-TA-1016) 38P TO GENZ EDSFF 1C (SFF-TA-1009) 56P RECEPTACLE WITH 15P POWER

Part #: MCIO5-4XSC-EDSFF-1X4-0.5M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P to GenZ EDSFF 1C (SFF-TA-1009) 56P Receptacle with 15P power, for 1X4 Testing with Serial Cables Gen5 Switch card, 0.5m

Part #: MCIO5-4XSC-EDSFF-1X4-1M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P to GenZ EDSFF 1C (SFF-TA-1009) 56P Receptacle with 15P power, for 1X4 Testing with Serial Cables Gen5 Switch card , 1m

PCIe Gen 5 MCIO x4(SFF-TA-1016) 38P to Gen 5 Multilink Drive Receptacle(SFF-8639) 68P, for U.2 1X4, for use with Serial Cables Gen 5 switch cards (WITH ULTRA LOW LOSS WIRE)

Part #: MCIO5-4X-39U2LL-1X4-0.5M

Description: PCIe Gen 5 MCIO x4(SFF-TA-1016) 38P to Gen 5 Multilink Drive Receptacle(SFF-8639) 68P, for U.2 1X4, for use with Serial Cables Gen 5 switch cards 0.5M(WITH ULTRA LOW LOSS WIRE)

Part #: MCIO5-4X-39U2LL-1X4-1M

Description: PCIe Gen 5 MCIO x4(SFF-TA-1016) 38P to Gen 5 Multilink Drive Receptacle(SFF-8639) 68P, for U.2 1X4, for use with Serial Cables Gen 5 switch cards 1M(WITH ULTRA LOW LOSS WIRE)

PCIE GEN5 MCIO 4X (SFF-TA-1016) 38P TO GENZ EDSFF 1C (SFF-TA-1009) 56P RECEPTACLE WITH 15P POWER

Part #: MCIO5-4XSC-EDSFF-2X2-0.5M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P to GenZ EDSFF 1C (SFF-TA-1009) 56P Receptacle with 15P power, for 2X2 Testing with Serial Cables Gen5 Switch card, 0.5M

Part #: MCIO5-4XSC-EDSFF-2X2-1M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P to GenZ EDSFF 1C (SFF-TA-1009) 56P Receptacle with 15P power, for 2X2 Testing with Serial Cables Gen5 Switch card, 1M

PCIE GEN5 MCIO 4X (SFF-TA-1016) 38P TO MULTILINK GEN5 PCIE U.2 DRIVE RECEPTACLE (SFF-8639) 68P 1×4 CABLE

Part #: MCIO5-4X-39U2-1X4-0.5M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P TO Multilink Gen5 PCIe Drive Receptacle (SFF-8639) 68P, 30AWG, 0.5 meter For connecting Gen5 Serial Cables Host/Switch to U.2 device 1×4

Part #: MCIO5-4X-39U2-1X4-1M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P TO Multilink Gen5 PCIe Drive Receptacle (SFF-8639) 68P, 30AWG, 1 meter For connecting Gen5 Serial Cables Host/Switch to U.2 device 1×4

PCIE GEN5 MCIO 4X (SFF-TA-1016) 38P TO MULTILINK GEN5 PCIE U.2 DRIVE RECEPTACLE (SFF-8639) 68P 2×2 CABLE

Part #: MCIO5-4X-39U2-2X2-0.5M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P TO Multilink Gen5 PCIe Drive Receptacle (SFF-8639) 68P, 30AWG, 0.5 meter For connecting Gen5 Serial Cables Host/Switch to U.2 device 2×2

Part #: MCIO5-4X-39U2-2X2-1M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P TO Multilink Gen5 PCIe Drive Receptacle (SFF-8639) 68P, 30AWG, 1 meter For connecting Gen5 Serial Cables Host/Switch to U.2 device 2×2

PCIE GEN5 MCIO 4X (SFF-TA-1016) 38P TO MULTILINK GEN5 PCIE U.3 DRIVE RECEPTACLE (SFF-8639) 68P 1×4 CABLE

Part #: MCIO5-4X-39U3-1X4-0.5M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P TO Multilink Gen5 PCIe U.3 Drive Receptacle (SFF-8639) 68P, 30AWG, 0.5 meter For connecting Gen5 Serial Cables Host/Switch to U.2 device 1×4

Part #: MCIO5-4X-39U3-1X4-1M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P TO Multilink Gen5 PCIe U.3 Drive Receptacle (SFF-8639) 68P, 30AWG, 1 meter For connecting Gen5 Serial Cables Host/Switch to U.3 device 1×4

PCIE GEN5 MCIO 4X (SFF-TA-1016) 38P TO MULTILINK GEN5 PCIE U.3 DRIVE RECEPTACLE (SFF-8639) 68P 2×2 CABLE

Part #: MCIO5-4X-39U3-2X2-0.5M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P TO Multilink Gen5 PCIe U.3 Drive Receptacle (SFF-8639) 68P, 30AWG, 0.5 meter For connecting Gen5 Serial Cables Host/Switch to U.3 device 2×2

Part #: MCIO5-4X-39U3-2X2-1M

Description: Gen5 MCIO 4x (SFF-TA-1016) 38P TO Multilink Gen5 PCIe U.3 Drive Receptacle (SFF-8639) 68P, 30AWG, 1 meter For connecting Gen5 Serial Cables Host/Switch to U.3 device 2×2

PCIE GEN5 MCIO X8 (SFF-TA-1016) 74P TO *2 MCIO X4 (SFF-TA-1016) 38P, Y-CABLE

Part #: MCIO5-8X-4X2-0.5M

Description: Gen5 MCIO x8 (SFF-TA-1016) 74P to *2 MCIO x4 (SFF-TA-1016) 38P, Y-Cable, 0.5M

PCIE GEN5 MCIO X8 (SFF-TA-1016) 74P TO *2 MULTILINK PCIE DRIVE RECEPTACLE (SFF-8639) 68P W/15P POWER

Part #: MCIO5-8X-39X2U2-1X4-0.5M

Description: Gen5 MCIO x8 (SFF-TA-1016) 74P to *2 Multilink PCIe drive receptacle (SFF-8639) 68P w/15P power, for 1×4 testing, 0.5M

PCIE GEN5 MCIO X8 (SFF-TA-1016) 74P TO MCIO X8 (SFF-TA-1016) 74P

Part #: MCIO5-8X-8X-0.5M

Description: Gen5 MCIO x8 (SFF-TA-1016) 74P to MCIO x8 (SFF-TA-1016) 74P, 0.5M

Part #: MCIO5-8X-8X-1M

Description: Gen5 MCIO x8 (SFF-TA-1016) 74P to MCIO x8 (SFF-TA-1016) 74P, 1M

PCIE GEN5 MCIO x8 to QSFP-DD with Sidebands

Part #: MCIO5-8X-QDDS-1M

Description: MCIO x8 to QSFP-DD with Sidebands, 1M

11.2.2 GEN5 EDSFF CABLES

GEN5 *2 CEM X2 64P TO EDSFF (SFF-TA-1009) 56P 1C RECEPTACLE FOR 2×2 TESTING

Part #: PCI5-CEMX2-EDSFF-0.5M

Description: Gen5 *2 CEM x2 64P to EDSFF (SFF-TA-1009) 56P 1C receptacle for 2X2 testing, 0.5 meter

GEN5 CEM X4 64P TO EDSFF (SFF-TA-1009) 56P 1C RECEPTACLE FOR 1X4 TESTING

Part #: PCI5-CEM-EDSFF-1X4-1M

Description: Gen5 CEM x4 64P to EDSFF (SFF-TA-1009) 56P 1C receptacle for 1X4 testing, 1 meter

GEN5 EDSFF 1C (SFF-TA-1009) 56P TO MULTILINK DRIVE PLUG (SFF-8639) U.2 68P CABLE

Part #: PCI5-39MU2x4-EDSFF-0.5M

Description: Gen5 EDSFF 1C (SFF-TA-1009) 56P to Multilink Drive Plug (SFF-8639) U.2 68P, 30AWG, Extension cable, 0.5M

PCIE GEN5 EDSFF (SFF-TA-1009) MALE/PLUG 56P TO EDSFF (SFF-TA-1009) FEMALE RECEPTACLE 56P, 0.5M

Part #: PCI5-EDSFFMF-0.5M

Description: Gen5 EDSFF (SFF-TA-1009) Male/Plug 56P to EDSFF (SFF-TA-1009) Female Receptacle 56P, 0.5M

Part #: PCI5-EDSFFMF-1M

Description: Gen5 EDSFF (SFF-TA-1009) Male/Plug 56P to EDSFF (SFF-TA-1009) Female Receptacle 56P, 1M

GEN5 QSFP-DD x8 76P to *2 EDSFF (SFF-TA-1009) 1C 56P w/15P power

Part #: QDD5-8X-EDSFF1C-1X4-0.5M

Description: Gen5 QSFP-DD x8 76P to *2 EDSFF (SFF-TA-1009) 1C 56P w/15P power, 0.5M

Part #: QDD5-8X-EDSFF1C-1X4-1M

Description: Gen5 QSFP-DD x8 76P to *2 EDSFF (SFF-TA-1009) 1C 56P w/15P power, 1M

11.2.3 GEN5 U.2 CABLES

GEN5 *2 CEM X2 64P TO U.2 (SFF-8639) RECEPTACLE FOR 2×2 TESTING

Part #: PCI5-CEMX2-39U2-0.5M

Description: Gen5 *2 CEM x2 64P to U.2 (SFF-8639) receptacle for 2×2 testing, 0.5 meter

GEN5 AIC/CEM X4 PLUG TO GEN5 MULTILINK PCIE (SFF-8639) 68P RECEPTACLE, W/15P POWER, FOR 1×4 TESTING, 0.5M LENGTH

Part #: PCI5-CEM-39U2-1X4-0.5M

Description: Gen5 AIC/CEM x4 Plug to Gen5 Multilink PCIe (SFF-8639) 68P Receptacle, w/15P power, for 1×4 testing, 0.5M length

PCIe Gen 5 MCIO x4(SFF-TA-1016) 38P to Gen 5 Multilink Drive Receptacle(SFF-8639) 68P, for U.2 1X4

Part #: MCIO5-4X-39U2LL-1X4-0.5M
Description: PCIe Gen 5 MCIO x4(SFF-TA-1016) 38P to Gen 5 Multilink Drive Receptacle(SFF-8639) 68P, for U.2 1X4, for use with Serial Cables Gen 5 switch cards 0.5M(WITH ULTRA LOW LOSS WIRE)

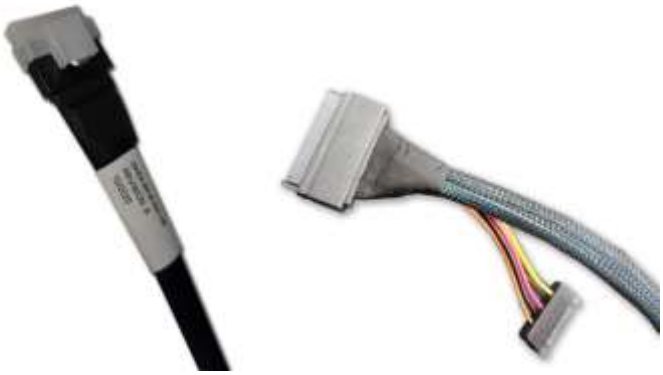
Part #: MCIO5-4X-39U2LL-1X4-1M
Description: PCIe Gen 5 MCIO x4(SFF-TA-1016) 38P to Gen 5 Multilink Drive Receptacle(SFF-8639) 68P, for U.2 1X4, for use with Serial Cables Gen 5 switch cards 1M(WITH ULTRA LOW LOSS WIRE)

Gen5 QSFP-DD x8 76P to *2 U.2 (SFF-8639) 68P w/15P power

Part #: QDD5-8X-39X2U2-1X4-0.5M
Description: Gen5 QSFP-DD x8 76P to *2 U.2 (SFF-8639) 68P w/15P power, 0.5M

Part #: QDD5-8X-39X2U2-1X4-1M
Description: Gen5 QSFP-DD x8 76P to *2 U.2 (SFF-8639) 68P w/15P power, 1M

11.2.4 GEN5 SlimSAS CABLES



PCIE GEN5 *2 SLIMSAS 8I (SFF-8654) STR. TO *1 MULTILINK GEN5 PCIE DRIVE RECEPTACLE (SFF-8639) 68P, 2×2

Part #: SLS5-8X-39X2U2-2X2-0.5M

Description: PCIe Gen5 *2 SlimSAS 8i (SFF-8654) Str. to *1 Multilink Gen5 PCIe drive Receptacle (SFF-8639) 68P, 2×2 connectivity, 0.5M. For use with BRCM Gen5 Host boards with SlimSAS

Part #: SLS5-8X-39X2U2-2X2-1M

Description: PCIe Gen5 *2 SlimSAS 8i (SFF-8654) Str. to *1 Multilink Gen5 PCIe drive Receptacle (SFF-8639) 68P, 2×2 connectivity, 1M. For use with BRCM Gen5 Host boards with SlimSAS

PCIE GEN5 SLIMSAS 8I (SFF-8654) STR. 74P TO *2 EDSFF 1C DRIVE RECEPTACLE (SFF-TA-1002) 56P FOR 1X4

Part #: SLS5-8X-1002X2-1X4-0.5M

SI Parameters: RAR

Description: PCIe Gen5 SlimSAS 8i (SFF-8654) str. 74P to *2 EDSFF 1C drive receptacle (SFF-TA-1002) 56P for 1X4 connections, 0.5M length

PCIE GEN5 *2 SLIMSAS 8I (SFF-8654) STR. 74P TO *1 DUAL PORT EDSFF 1C DRIVE RECEPTACLE (SFF-TA-1002) 56P FOR 2X2

Part #: SLS5-8X-1002-2X2-0.5M

Description: PCIe Gen5 *2 SlimSAS 8i (SFF-8654) str. 74P to *1 Dual Port EDSFF 1C drive receptacle (SFF-TA-1002) 56P for 2X2 connections, 0.5M length. Lanes 4-7 are NC

PCIE GEN5 SLIMSAS 8I (SFF-8654) STR. TO *2 MULTILINK GEN5 PCIE DRIVE RECEPTACLES (SFF-8639) 68P, 1×4

Part #: SLS5-8X-39X2U2-1X4-0.5M

Description: PCIe Gen5 SlimSAS 8i (SFF-8654) Str. to *2 Multilink Gen5 PCIe drive Receptacles (SFF-8639) 68P, 1×4 connectivity, 0.5M. For use with BRCM Gen5 Host boards with SlimSAS

Part #: SLS5-8X-39X2U2-1X4-1M

Description: PCIe Gen5 SlimSAS 8i (SFF-8654) Str. to *2 Multilink Gen5 PCIe drive Receptacles (SFF-8639) 68P, 1×4 connectivity, 1M. For use with BRCM Gen5 Host boards with SlimSAS

11.2.5 GEN5 OTHER CABLES

Gen5 CEM x16 164P Male Plug/Gold Finger to CEM x16 164P Straddle Mount Cable, 0.3M

Part #: PCI5-CEMX16MF-0.3M

Description: Gen5 CEM x16 164P Male Plug/Gold Finger to CEM x16 164P Straddle Mount Cable, 0.3M

Gen5 CEM x16 164P Male Plug/Gold Finger to CEM x16 164P Straddle Mount Cable, 0.45M

Part #: PCI5-CEMX16MF-0.45M

Description: Gen5 CEM x16 164P Male Plug/Gold Finger to CEM x16 164P Straddle Mount Cable, 0.45M

Gen5 EDSFF (SFF-TA-1002) 2C Male Plug/Gold Finger to EDSFF (SFF-TA-1002) 2C Female Receptacle Cable, 0.5M

Part #: PCI5-EDSFF2CMF-0.5M

Description: Gen5 EDSFF (SFF-TA-1002) 2C Male Plug/Gold Finger to EDSFF (SFF-TA-1002) 2C Female Receptacle Cable, 0.5M

Gen5 EDSFF (SFF-TA-1002) CEM/AIC Male Plug/Gold Finger to EDSFF (SFF-TA-1002) 2C Female Receptacle Cable, 0.5M

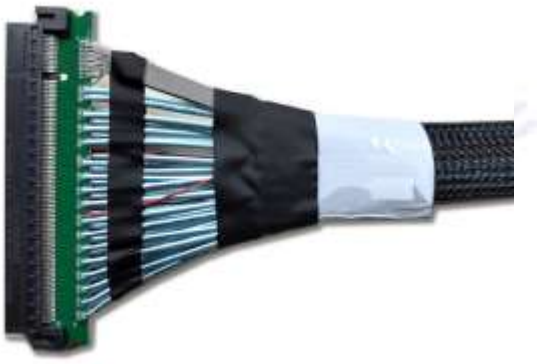
Part #: PCI5-CEMX8M-EDSFF2CF-0.5M

Description: Gen5 EDSFF (SFF-TA-1002) CEM/AIC Male Plug/Gold Finger to EDSFF (SFF-TA-1002) 2C Female Receptacle Cable, 0.5M

Gen5 CEM x16 164P Male Plug/Gold Finger to EDSFF (SFF-TA-1002) 4C+ Cable, 0.5M

Part #: PCI5-CEMX16M-4C+F-0.5M

Description: Gen5 CEM x16 164P Male Plug/Gold Finger to EDSFF (SFF-TA-1002) 4C+ Cable, 0.5M

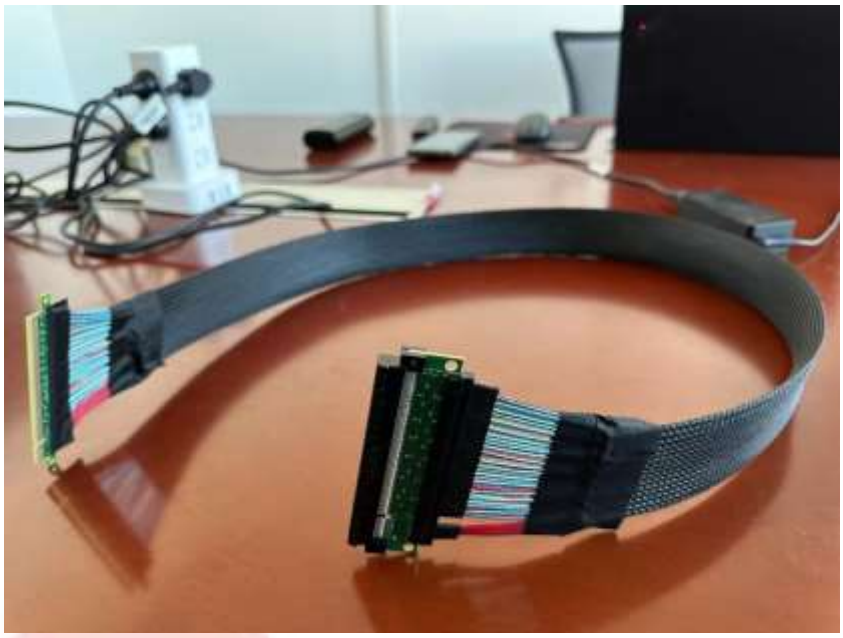


Gen5 CEM x16 164P Straddle Mount to EDSFF (SFF-TA-1002) 4C+ Male Plug/Gold Finger Cable, 0.5M

Part #: PCI5-CEMX16F-4C+-0.5M

Description: Gen5 CEM x16 164P Straddle Mount to EDSFF (SFF-TA-1002) 4C+ Male Plug/Gold Finger Cable, 0.5M

11.2.6 GEN5 PCIE CEM CABLES



Product Status	Active
Cable Connectors	PCIe x16
Number of Positions	164
Connector Type	Receptacle to Plug
Gender	Female to Male
Fastening Type	Push-Pull
Features	32Gbps PCIe Gen5 Riser Cable
Cable Type	Flat, Twin Axial
Usage	Internal
Length	1.31' (400.0mm)
Color	Black
Shielding	Shielded
Wire Gauge	-
Frequency - Max	25GHz
Operating Temperature	-
Base Product Number	PC-PCIE-16X

规格

400mm

500mm

700mm

1000mm

11.3 PCIe GEN5 主机卡/switch card

PCIE GEN5 X16 MCIO HOST CARD WITH BROADCOM ATLAS2 A0 PCIE SWITCH

Part #: PCI5-AD-x16HI-A0-BG5

Description: PCIe Gen5 capable x16 host card. Based on Broadcom's new **PEX89000 PCIe switch chip this x16 host card has 4, x4 MCIO connectors (special PCI bracket to expose internal MCIO connectors outside the enclosure) and 1, x16 Gen5 receptacle. This host card supports bifurcation down to 8×2's per/station and is user configurable through onboard CLI commands. Auxiliary power connector available for powering the DUT(s) vs slot supplied power.

****Supports Broadcom's PEA (PCIe Embedded Analyzer) technology through SerialTek's iTAP software.*

*** A0 version of PEX89000 PCIe switch chip is not a production chip and could have some errata.*

****iTAP software not included and functions on station 0 only*

*****Station 0 at MCIO*

PCIE GEN5 X16 MCIO HOST CARD WITH BROADCOM ATLAS2 PRODUCTION LEVEL LLC BROADCOM PCIE SWITCH

Part #: PCI5-AD-x16HI-BG5

Description: PCIe Gen5 x16 host card w/4, x4 MCIO receptacles based on the production level LLC (low lane count) Broadcom Atlas II switches. ***Supports Broadcom's PEA (PCIe Embedded Analyzer) technology through SerialTek's iTAP Panda Hardware.

***iTAP Panda Hardware not included

PCIE GEN5 X16 QSFP-DD HOST CARD WITH BROADCOM ATLAS2 B0 PCIE SWITCH

Part #: PCI5-AD-x16HE-BG5-QDD

Description: PCIe Gen5 capable x16 host card. Based on Broadcom's new **PEX89000 PCIe switch chip this x16 host card has 2, x8 WDD connectors.

***Supports Broadcom's PEA (PCIe Embedded Analyzer) technology through SerialTek's iTAP Panda Hardware.

***iTAP Panda Hardware not included

11.4 PCIe GEN5 Retimer 和 Redriver 卡

SerialCables 公司推出了基于 Astera Labs 的 Gen5 retimer 卡，参见下面的描述。

PCIE GEN5 X16 QSFP-DD RETIMER

Part #: PCI5-AD-x16HE-RT-A

Description: PCIe Gen5 x16 retimer w/2, x8 QSFP-DD receptacles based on the production level Astera Retimer.

Part #: PCI5-AD-x16HE-RT-P

Description: PCIe Gen5 x16 retimer w/2, x8 QSFP-DD receptacles based on the production level Parade Retimer.

下面是 astera labs 公司的 gen5 x16 retimer 卡。

The PCI Express® 5.0 add-in-card are intended for in-system evaluation of the Aries PCIe 5.0 x16 Smart Retimer. The low-profile active add-in card has a x16 PCIe CEM-compliant edge finger to be plugged into a Gen-5 system, and features a x16 CEM

connector on top to install an endpoint add-in card. It is configured for plug-and-play operation, meaning no retimer configuration is required and the Root Complex (e.g. CPU) and Endpoint (e.g. NIC) will automatically form a Link through the Aries Smart Retimer on power-up and de-assertion of PERST#.

下面是 Phison 公司的 Gen5 x16 redriver 卡，实际测试在各种平台上面对于信号增强的效果不错。

1.5 Pin Mode Setting

1.5.1 Pin Mode Default Setting

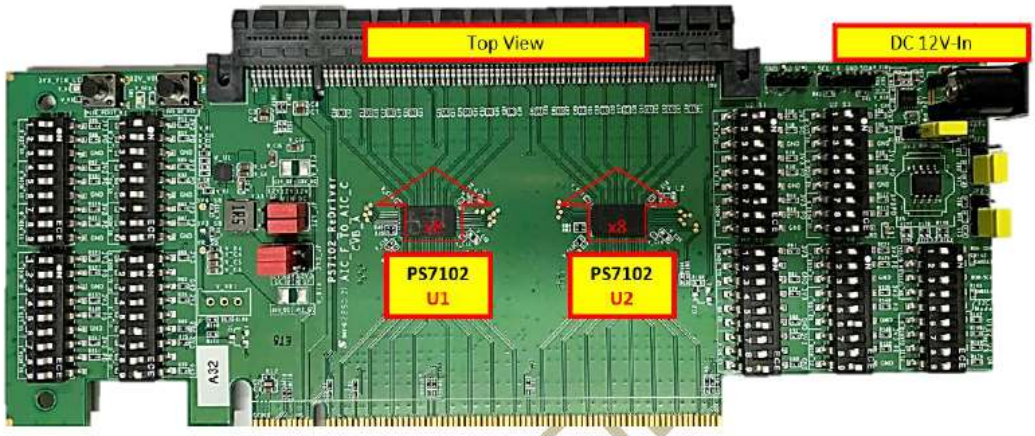


Figure 3: PS7102 Pin Mode Default Setting (TOP View)

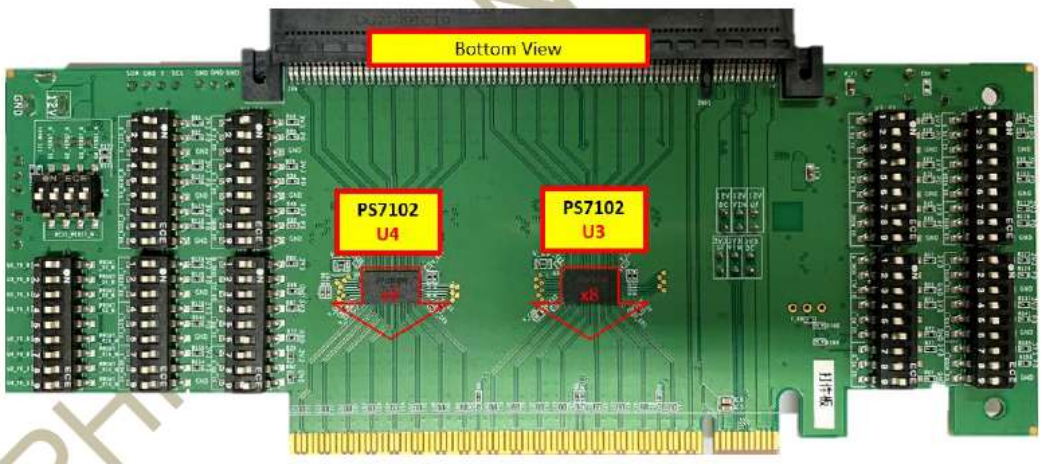
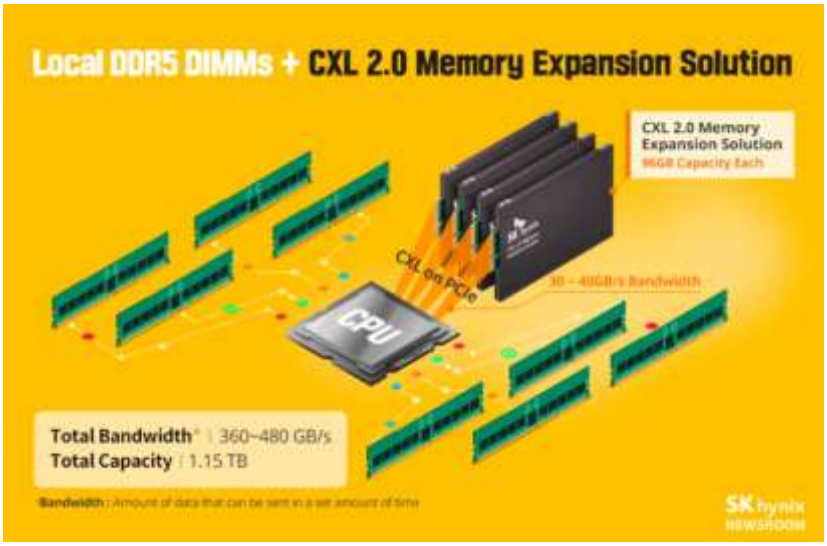


Figure 4: PS7102 Pin Mode Default Setting (Bottom View)

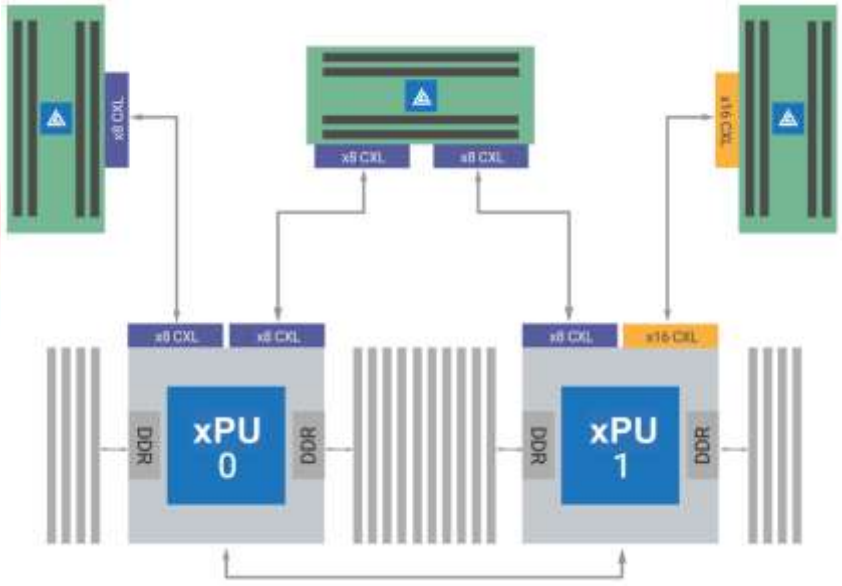
11.5 PCIe Gen5 CXL type3 Smart Memory Card

基于 CXL 技术的内存扩展获得了内存条厂商和高性能计算厂商的青睐，业内主流公司都在开发该类产品的过程中，如下图的 SK Hynix 的基于 EDSFF 的内存扩展。



下面以市场上可以测试的 CXL Type3 的卡为例简单介绍一下。

11.5.1 SYSTEM BLOCK DIAGRAM



11.5.2 Leo-SVB-RevA Add-in card

Leo System Validation Board (SVB) is an add-in card (AIC) for customer evaluation of the Leo CXL Smart Memory Controllers. The Leo SVB interfaces with the host enabling memory expansion, pooling and sharing over Compute Express Link (CXL)

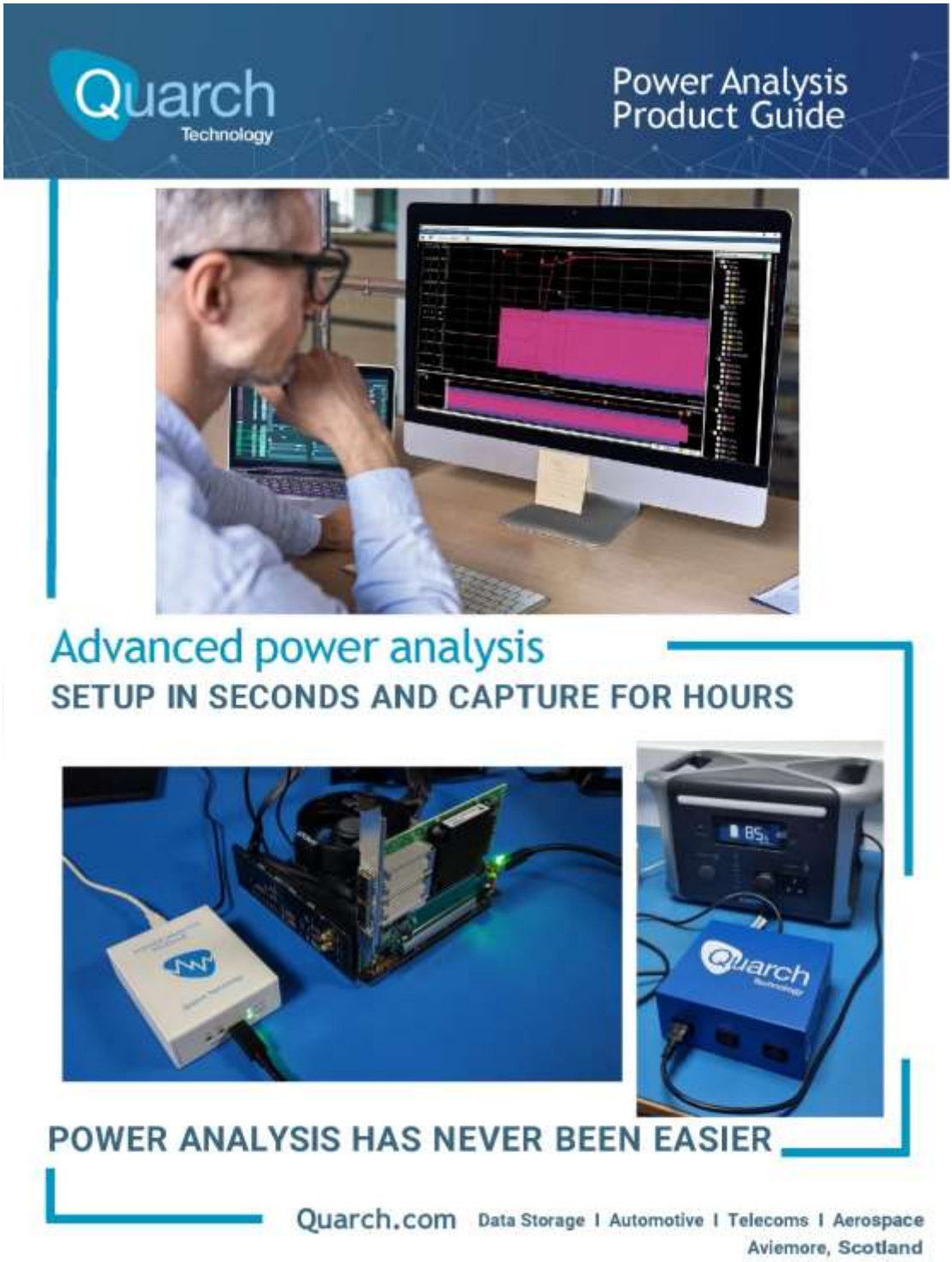
Leo CXL Smart Memory Controller - System Validation Board

11.5.3 A1000-1P4AA Smart Memory Card

Aurora A-Series CXL Smart Memory Solution in an add-in-card form factor is Astera Lab's turnkey offering for quick high volume deployment of use cases such as memory expansion and pooling featuring Leo P-Series CXL Smart Memory Controllers.

Aurora A-Series CXL Smart Memory Add-in Card

12. 附录 C： Quarch 功耗测试/分析速查



12.1 DC power analysis

DC power analysis

DC applications require a power analysis module and a power analysis fixture

Capture analog and digital signals over long time periods

12.2 Storage power analysis

Quarch Technology

Power Analysis Product Guide

GEN5 SFF PAM FIXTURE (SAS/SATA/U.2/U.3)

SFF PAM: ANALOG CHANNELS: 12V, [illegible] SAMPLE RATE 250KS/S, VOLTAGE: 0 – 15V ±(2mV+1%), 12V CURRENT: 100uA-1mA ± (25uA+1%) 1mA-13A ± (2mA+1%), 5V CURRENT: 100uA-1mA ± (25uA+1%) 1mA-13A ± (2mA+1%), 3.3VAUX CURRENT: 0-85mA ± (25uA+1%) DIGITAL CHANNELS: PERST#, CLKREQ#_PERSTB#, WAKE#, SMBCLK, SMBDAT, ACTIVITY#, PWRDIS, PRSNT#, IFDET, IFDET2, HPT0, HPT1, DUAPLORTEN#, P2, SAMPLE RATE 1MS/S

Storage power analysis

GEN5 M.2 M-KEY PAM FIXTURE

M.2 PAM: ANALOG CHANNELS: 3.3V, VIO_1V8, SAMPLE RATE 250KS/S, VOLTAGE: 0 – 15V ±(2mV+1%), CURRENT: 0-1mA ±(15uA+1%) 1mA-13A ±(2mA+1%) DIGITAL CHANNELS: CLKREQ#, PERST#, PEWAKE#, SUSCLK, PEDET, ALERT#, SMB_DATA, SMB_CLK, LED_1#, DEVSLP, MFG_DATA, MFG_CLK, VIO_CFG, PWRDIS, PLA_S3#, PLN# SAMPLE RATE 1MS/S

12.3 Storage and beyond

Quarch Technology

Power Analysis Product Guide

GEN5 EDSFF PAM FIXTURE (E1.S/E1.L/E3/E3

EDSFF PAM: ANALOG CHANNELS: 12V, 3.3VAUX, SAMPLE RATE 250KS/S, VOLTAGE: 0 – 15V ±(2mV+1%), 12V CURRENT: 100uA-1mA ± (25uA+1%) 1mA-13A ± (2mA+1%), 3.3VAUX CURRENT: 100uA-85mA ± (25uA+1%) DIGITAL CHANNELS:PRSNT0#, PERST1#_CLKREQ#, LED, SMBRST#, SMBDAT, SMBCLK, PWRDIS, PERST0#, DUALPORTEN#, RFU, MFG, SAMPLE RATE 1MS/S

Storage and beyond

GEN5 AIC/SLOT X16 PAM FIXTURE

Supports all slot powered AIC devices: SSDs, NICs, HBAs and more

AIC PAM: ANALOG CHANNELS: 12V, 3.3V, 3.3VAUX, SAMPLE RATE 250KS/S, VOLTAGE: 0 – 15V ±(2mV+1%), 12V CURRENT: 100uA-1mA ± (10uA+1%) 1mA-13A ± (2mA+1%), 3V3 CURRENT: 100uA-1mA ± (10uA+1%) 1mA-13A ± (2mA+1%), 3.3VAUX CURRENT: 0-400mA ± (10uA+1%) DIGITAL CHANNELS: PERST#, CLKREQ#, WAKE#, SMBCLK, SMBDAT, SAMPLE RATE 1MS/S

12.4 GPU and AI Analysis

Quarch Technology

Power Analysis Product Guide

GEN5 AIC/SLOT X16 PAM FIXTURE +AUX

AIC +AUX PAM: ANALOG CHANNELS: 12V, 3.3V, 3.3VAUX, 12VAUX, SAMPLE RATE 250KS/S, VOLTAGE: 0 – 15V ±(2mV+1%), 12V CURRENT: 0-32.5A ± (5mA+1%), 3.3V CURRENT: 0-13A ± (2mA+1%), 3.3VAUX CURRENT: 0-3.25A ± (0.5mA+1%), 12VAUX CURRENT: 0-162.5A ± (25mA+1%) DIGITAL CHANNELS: PERST#, CLKREQ#, WAKE#, SMBCLK, SMBDAT, REFCLK_LOS#, SAMPLE RATE 1MS/S

GPU and AI Analysis

AUX FIXTURES (DUAL PCIE, TRIPLE PCIE, 12VHP)

Calibrated fixtures support all major GPU / Accelerator power supplies

Includes power sequencing

Dual PCIe, Triple PCIe and 12VHP versions

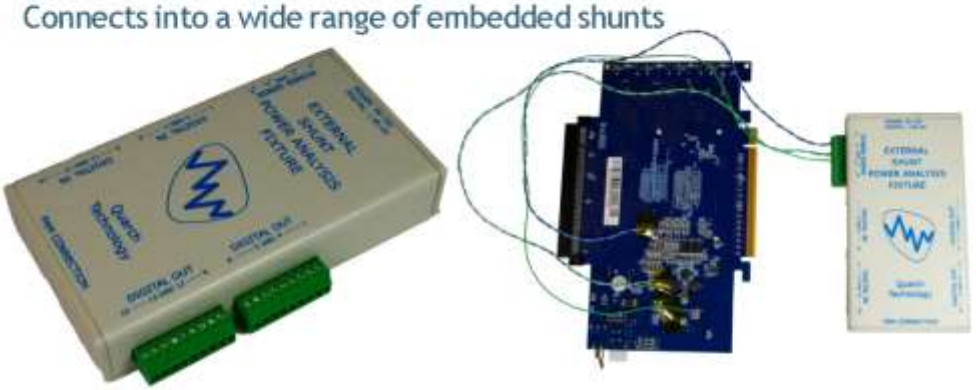
12.5 Multi-channel fixtures

Quarch Technology

Power Analysis Product Guide

EXTERNAL SHUNT PAM

Connects into a wide range of embedded shunts



EXTERNAL SHUNT: 4 CHANNELS, SAMPLE RATE 250KS/S, VOLTAGE: 0 - 15V ±(2mV+1%), CURRENT SENSE: 10uV-65mV ±(10uV+1%) DIGITAL CHANNELS: 16 CHANNELS SAMPLE RATE 1MS/S, VOLTAGE RANGE 1.8-5V

Multi-channel fixtures

EXTERNAL SHUNT PAM WITH DIMM FIXTURE

Connected to DIMM adaptor board

Allowing analyis of RAM power consumption

Supports 50mV current sense resistors

Quarch Technology

Power Analysis Product Guide

2 CHANNEL

2 CHANNEL: ANALOG CHANNELS: 2 CHANNELS, SAMPLE RATE 250KS/S, VOLTAGE: 0 – 15V ±(2mV+1%), CURRENT: 0-1mA ±(15uA+1%) 1mA-13A ±(2mA+1%) DIGITAL CHANNELS: 16 CHANNELS SAMPLE RATE 1MS/S, VOLTAGE RANGE 1.8-5V

Multi-channel fixtures

4 CHANNEL PAM

4 CHANNEL: ANALOG CHANNELS: 4 CHANNELS, SAMPLE RATE 250KS/S, VOLTAGE: 0 – 15V ±(2mV+1%), CURRENT: 10mA-13A ±(10mA+1%) DIGITAL CHANNELS: 16 CHANNELS SAMPLE RATE 1MS/S, VOLTAGE RANGE 1.8-5V

12.6 AC Power Analysis

12.7 Quarch Power Studio (QPS)

Quarch Technology

Power Analysis Product Guide

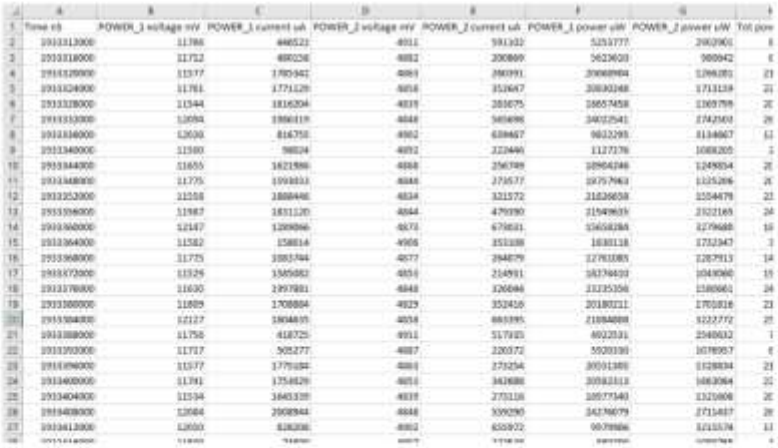
SIMPLE AUTOMATION API

Automate capture, annotations, custom channels, statistic calculations and more. Code examples: www.quarch.com/support/application-note

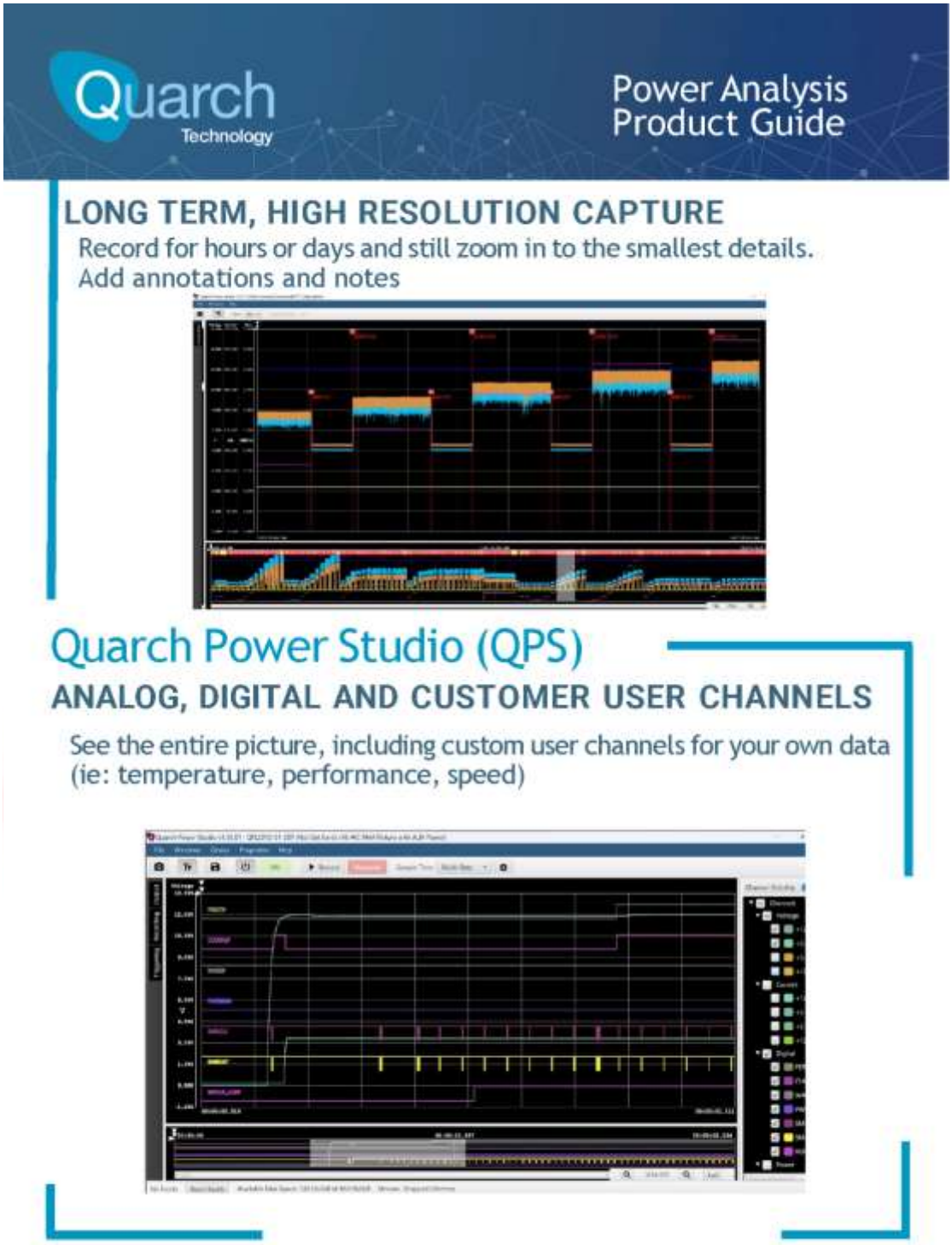
Quarch Power Studio (QPS)

EASY SHARING AND POST PROCESSING

Share full traces or smaller sections. Export to CSV format for custom post-processing



	A	B	C	D	E	F	G	H
1	Time nS	POWER_1 voltage mV	POWER_1 current uA	POWER_2 voltage mV	POWER_2 current uA	POWER_1 power uW	POWER_2 power uW	Tot pow
2	1911312000	11786	446521	4911	591102	5251777	2902901	8
3	1911316000	11752	480258	4882	200869	5623610	980942	6
4	1911320000	11577	1785942	4863	286991	20668994	1396181	21
5	1911324000	11781	1771129	4858	352647	20830268	1713139	21
6	1911328000	11544	1816204	4839	281675	20657458	1369799	20
7	1911332000	12094	1986319	4848	565698	24022541	2742503	26
8	1911336000	12030	816750	4902	698467	9822295	3134667	12
9	1911340000	11580	98024	4892	222446	1127176	1088205	1
10	1911344000	11655	1621586	4868	256749	18904246	1249854	20
11	1911348000	11775	1916811	4846	273577	22575963	1325206	23
12	1911352000	11558	1888446	4834	321572	21826658	1554479	23
13	1911356000	11987	1831120	4844	479390	21949635	2322165	24
14	1911360000	12147	1269866	4870	676581	15658284	3279668	19
15	1911364000	11582	158814	4905	351038	1838118	1722347	3
16	1911368000	11775	1881744	4877	284879	22161085	1287913	24
17	1911372000	11529	1585882	4851	214931	18278410	1043060	19
18	1911376000	11630	1957881	4848	326046	23235356	1580661	24
19	1911380000	11809	1708804	4829	352416	20180211	1701816	23
20	1911384000	12127	1804635	4858	663395	21884089	3222772	25
21	1911388000	11750	418725	4911	517101	4922031	2540632	7
22	1911392000	11717	505277	4887	220372	5920330	1078957	6
23	1911396000	11577	1775184	4863	273254	20551389	1328834	21
24	1911400000	11781	1754829	4851	342688	20582313	1663064	22
25	1911404000	11534	1665339	4839	275116	19577340	1325808	20
26	1911408000	12084	2008944	4848	339290	24276079	1711467	26
27	1911412000	12010	826206	4892	650972	9978866	3213574	13



12.8 Automation options

Quarch Technology

Power Analysis Product Guide

60W DUAL RAIL PROGRAMMABLE POWER SUPPLIES

12V and 3.3V/5V dual rail supply for SSDs, HDDs and beyond

Fully compatible with Power Studio and automated power capture

HD PPM: 2 PROGRAMMABLE OUTPUTS, 0-14.4V AND 0-6V, 1024 PATTERN POINTS PER CHANNEL, 1V/US NO-LOAD SLEW, SAMPLE RATE 250KS/S, VOLTAGE: 0 - 14.4V ±(1%), CURRENT: 0-1mA ±(2uA+2%) 1mA-4A ±(2mA+1%)

Programmable Power Modules

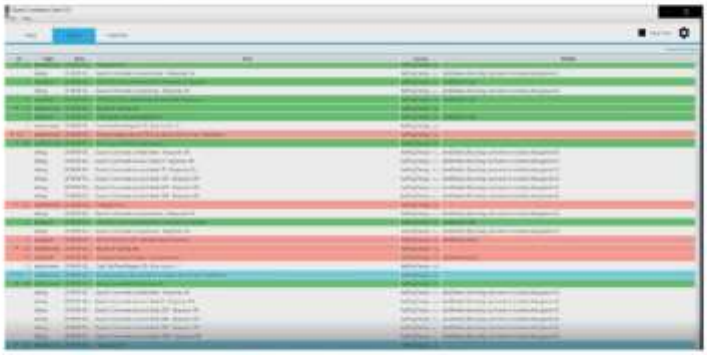
PLUG AND PLAY FIXTURING FOR MANY INTERFACES

Quarch Technology

Power Analysis Product Guide

QUARCH COMPLIANCE SUITE

Run standard automated workload, voltage margining tests and more



Automation options

QIS & QUARCHPY

Java Instrumentation Server allowing simple TCP based control of any Quarch Power Device. Full Python API available for fast integration

```
from quarchpy.device import *

# Specify the module to control
myDeviceID = "USB:QTL1999-05-005"

# Connect to the module
myQuarchDevice = getQuarchDevice(myDeviceID, ConType = "QIS")

# Convert the base device class to a power device, which provides additional controls, such as data streaming
myPowerDevice = quarchPPM(myQuarchDevice)

module.startStream("Stream1.csv")
```

Download QIS from: **quarch.com/downloads**
Download quarchpy from: **pypi.org/project/quarchpy/**

13. 附录 D：PCIe Gen 4/5/6 测试工具定制开发

结合我们在存储业界多年针对用户需求的关注，我们也提供结合每个用户的测试工具定制的发，尤其是当前热度较高的 PCIe Gen 4/5/6 相关的测试制具，夹具，工装，主机卡，交换板，自动化批量测试装置等，我们有业内最完善的 PCIe/NVMe 方面的技术专家，可以提供最高性价比的产品定制方案，可以灵活沟通 NRE。我们已经开发了大概 20 多款产品，获得了业内众多用户的好评。

下图是我们的业务定制开发能力一览。

14. 附录 E：PCIe 互操作性和兼容性测试夹具

PCI-SIG 合规性研讨会主持互操作性和合规性测试。互操作性测试使成员能够针对其他成员的产品测试他们的产品。一致性测试允许针对 PCI-SIG 测试模块进行产品测试，例如和 SerialTek Gen5 协议分析的互操作性测试，或者和 SanBlaze PCIe Gen5 SSD 测试设备的互操作性测试等。

两种测试类型都会针对每个检查的测试区域发布“通过”或“失败”结果。要正式将产品标记为合规，它们必须在互操作性测试中获得至少 80% 的分数并通过所有必需的合规性测试。

经过测试的规范互操作性和合规性测试侧重于最新的 PCI 规范，包括：

- **PCI Express 5.0**
- **PCI Express 4.0**
- **PCI Express U.2® / SFF-8639/ (8 GT/s)**
- **PCI Express M.2™ (8 GT/s)（仅供初步参考 ）**

PCI Express 物理层一致性测试主要是电气测试- 检查平台和附加卡 transmitter 发射器和 receiver 接收器特性。***大多数情况下，如果购买这些物理层兼容性测试设备将是一笔昂贵的实验室固定资产投资，目前国内在上海和北京等有很多第三方实验室都提供 PCIe Gen5 物理层和协议层兼容性测试，只要支付很少费用即可提前了解开发的芯片、板卡或者系统是否可以通过这些测试。***

对于一定要自己购买这些物理层兼容性测试夹具和治具的公司，下面你的列表可供参考。

物理层兼容性测试需要的夹具和治具

Product Code	Description
NR66	PCI Express M.2 Specification Revision 4.0, Version 1.1
NR65	PCI Express Base Specification Revision 6.0.1, Version 1.0
NR64	PCI Express Base Specification Revision 6.0 (hard copy)
NR63	PCI Express Card Electromechanical Specification Revision 5.0 (hard copy)
NR62	PCI Express SFF-8639 Module Specification Revision 4.0, Version 1.0 (hard copy)
NR61	PCI Express M.2 Specification Revision 4.0, Version 1.0 (hard copy)
NR60	PCI Express OCuLink Specification Revision 1.1 (hardcopy)
NR59	PCI Express External Cabling Specification Revision 3.0a (hard copy)

NR58	PCI Express Mini Card Electromechanical Specification Revision 2.1 (hard copy)
NR56	PCI Express Card Electromechanical Specification Revision 4.0 (hard copy)
NR55	PCI Express Base Specification Revision 5.0 (hard copy)
NR54	PCI Express SFF-8639 Module Specification Revision 3.0 (hard copy)
NR53	PCI-SIG ECN/Errata (hard copy). A full list can be found here. Please specify in the notes of the order form below which ECN/Errata you would like to purchase.
NR52	PCI Express Base Specification Revision 4.0 (hard copy)
NR50	PCI Express Base Specification Revision 3.1a (hard copy - includes both the Base and Card Electromechanical 3.0 specification documents)
NR48	PCI Express M.2 Specification Revision 3.0, Version 1.2 (hard copy)
NR47	PCI Express External Cabling 2.0 (hard copy)
NR46	PCI Firmware Specification 3.1 (hard copy)
NR45	PCI Express Base 3.0 Specification (hard copy - includes both the Base and Card Electromechanical 3.0 specification documents)
NR44	PCI Code and ID Assignment Specification 1.0 (hard copy)
NR43	Single Root I/O Virtualization Specification 1.1 (hard copy)
NR42	PCI Express Label and Usage Guidelines Revision 1.1 (hard copy)
NR41	PCI Express Base Specification Revision 2.1 (hard copy)
NR40	Address Translation Services Revision 1.1 (hard copy)
NR39	Multi-Root I/O Virtualization 1.0 (hard copy)
NR38	PCI Express 225W/300W High Power Card Electromechanical Specification 1.0 (hard copy)
NR37	Single Root I/O Virtualization Specification 1.0 (hard copy)
NR36	PCI Express Mini Card Electromechanical Specification Revision 1.2 (hard copy)
NR35	Address Translation Services Revision 1.0 (hard copy)
NR34	PCI Express External Cabling 1.0 Specification (hard copy)
NR33	PCI Express Base 2.0 Specification (hard copy - includes both the Base and Card Electromechanical 2.0 specification documents)
NR31	PCI Firmware Specification 3.0 (hard copy)
NR30	PCI Express ExpressModule Electromechanical Specification 1.0 (hard copy)
NR29	PCI Express x16 Graphics 150W-ATX Specification 1.0 (hard copy)
NR28	PCI Express Mini Card Electromechanical Specification 1.1 (hard copy)
NR27	PCI Express to PCI/PCI-X Bridge Specification Revision 1.0 (hard copy)
NR26	PCI-X 2.0a Protocol and Electrical Specification (hard copy)
NR25	PCI Express Specification 1.1 (hard copy - includes both the Base and Card Electromechanical 1.1 specification documents)
NR23	Standard Hot Plug Controller Specification 1.0 (hard copy)
NR16	PCI Local Bus Specification, Rev 3.0 (hard copy)
NR15	PCI-X Specification 1.0b (hard copy)
NR14	PCI-to-PCI Bridge Specification 1.2 (hard copy)
NR13	Mobile Design Guide Specification 1.1 (hard copy)
NR12	PCI Hot Plug Specification 1.1 (hard copy)
NR11	Power Management Interface Specification 1.2 (hard copy)
NR10	Mini PCI Specification 1.0 (hard copy)
NR9	PCI BIOS Specification 2.1 (hard copy)
CD1	CD ONLY, with PDF versions of all current Specifications.

PCIe M.2 3.0 CLB/CBB	PCIe M.2 3.0 (8.0 GT/s) CLB/CBB Test Fixture Kit Kit Includes: PCIe M.2 3.0 (8.0 GT/s) Fixture Kit (CBB & CLB) PCIe 4.0 Variable ISI board Mini Bend Cable Kit **AVAILABLE TO PCI-SIG MEMBERS ONLY**
4.0 CEM Kit	PCIe 4.0 (16.0 GT/s) CEM Electrical Test Fixture Kit includes: PCIe-CLB-x1x16, PCIe-CLB-x4x8, PCIe-CBB-MAIN, PCIe-VAR-ISI, **SMP to 3.5mm** short cable assembly adaptor & **SMP** 1 foot cable assembly ****Available to PCI-SIG Members Only****
CBB3	Rev. 3.0 of the PCI Express Compliance Base Board (CBB) for testing PCI Express Add-in Cards. (8.0 GT/s) **CURRENTLY OUT OF STOCK**
x1/x16 CLB3	PCI Express Compliance Load Board (CLB) for testing PCI Express Platforms (8.0 GT/s).
x4/x8 CLB3	PCI Express Compliance Load Board (CLB) for testing PCI Express Platforms (8.0 GT/s).
U.2 Electrical Pair	PCI Express U.2® (SFF-8639) Pair: Compliance Base Board and Compliance Load Board (8.0 GT/s) ****Available to PCI-SIG Members Only****
U.2 Adapter DP	PCI Express U.2® (SFF-8639) Dual Port Adapter (8.0 GT/s) ****Available to PCI-SIG Members Only****
U.2 Adapter SP	PCI Express U.2® (SFF-8639) Single Port Adapter (8.0 GT/s) ****Available to PCI-SIG Members Only****

另外，在 PCIe 物理层测试过程中，有的时候也会用到第三方的一些治具配合 Tek 或者 Keysight 的示波器一起使用，参见下页的常用的一些测试治具示例。

Resources

PCIE GEN4

PCIe test adapters and test fixture kits facilitate Source and Sink compliance testing.

PCIe Gen4

M.2 GEN4 SOCKET2

The PCIe Gen4 M.2 Socket 2 test adapter or test fixture kit facilitates Device and Host compliance testing for PCIe Gen4 M.2 Socket 2 Source and Sink devices.

Shop M.2 Gen4 Socket2

PCIe Gen4

M.2 GEN4 SOCKET3

The PCIe Gen4 M.2 Socket 3 test adapter or test fixture kit facilitates Device and Host compliance testing for PCIe Gen4 M.2 Socket 3 Source and Sink devices.

Shop M.2 Gen4 Socket3

PCIe Gen4

SFF 8639 GEN4 (U.2)

The 8639G4-TPA-PR test adapter or test fixture kit facilitates Source and Sink compliance testing for SFF-nsons and Devices. These SFF-8639 Gen4 fixtures utilize the multifunctional connector system supporting single/dual port SATA, dual and MultiLink SAS, and/or up to four (4) port PCIe device configurations. The SFF-8639 Gen4 receptacle fixture will accept devices with plug connectors that are in accordance with SFF-8482, SFF-8630, and SFF-8680.

Shop SFF-8639 Gen4 (U.2)

RESOURCES

User Manual & Specs	M.2 Socket 2	SFF-8639
	M.2 Socket 3	
Mechanical Drawings	SFF-8639 Universal Receptacle	PCIe M.2 Socket 2 Receptacle
	SFF-8639 Universal Plug	PCIe M.2 Socket 3 Plug
	PCIe M.2 Socket 2 Plug	PCIe M.2 Socket 3 Receptacle

15. 附录 F：PCIe 5.0 协议诊断、分析、测试常用工具和经验分享及 CXL 技术研讨

15.1 PCIe 5.0 协议诊断、分析、测试常用工具和经验分享

Saniffer

Agenda

- PCIe Gen5产品的市场现状
- PCIe Gen5分析和测试常用到哪些工具

Agenda

- PCIe Gen5产品的市场现状
  - 主机 – server和workstation, PC
  - 板卡和SSD
- PCIe Gen5分析和测试常用到哪些工具

PCIe Gen5主机 – server和workstation, PC

- Intel
  - Sapphire Rapids is a codename for Intel's server (fourth generation Xeon Scalable) and workstation processors based on Intel 7. There are 3 future Gen 5 platforms underway now.
  - Intel Core 12th, 13th CPU (Z690, Z790 chipset) – Asus, Gigabyte, Asrock, MSI
- AMD
  - AMD Genoa (4th Gen AMD EPYC) has launched with up to 96 cores and world-beating performance
  - X670E and B650 motherboards unleash the full performance of AMD Ryzen 7000 Series processors on AMD's Zen 4 AM5 platform and feature PCIe® 5.0 – Asus, Gigabyte, Asrock, MSI

PCIe Gen5主机 – server和workstation, PC

PCIe Gen5板卡和SSD

- Gen5 switch卡
  - SerialCables Gen5 x16 switch card
- Gen5 retimer卡
  - Astera Labs
- Gen5 GPU
  - MTS S80, S3000
- Gen5 网卡
  - Mellanox CX7
- Gen5 E3.S SSD
  - Kioxia CD7; CD8**, CM7**
- Gen5 M.2 SSD
  - Gigabyte, Inland Gen5 M.2 SS (Phison E26 controller)
- Gen5 U.2 SSD
  - Samsung PM1743 **
- Gen5 CXL Type 3 内存扩展卡
  - Samsung, SK Hynix, Montage, Astera Labs

Agenda

- PCIe Gen5产品的市场现状
- PCIe Gen5分析和测试常用到哪些工具
  - PCIe Gen5信号和速度给协议分析仪设计带来的挑战
  - 客户对于PCIe Gen5进行问题分析的痛点和难点

PCIe Gen5分析和测试常用到哪些工具

- 示波器，误码仪，VNA**
- 协议分析仪
- 掉电测试、热插拔、异常故障注入
- 电压拉偏和功耗测试设备
- 电压、电流、功耗、 Side-band边带信号监测设备
- SSD测试设备**
- 搭建PCIe Gen5测试环境用到各种产品

PCIe Gen5分析和测试常用到哪些工具

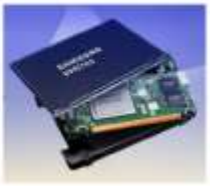
- 示波器，误码仪，VNA**
- 协议分析仪
- 掉电测试、热插拔、异常故障注入
- 电压拉偏和功耗测试设备
- 电压、电流、功耗、 Side-band边带信号监测设备
- SSD测试设备**
- 搭建PCIe Gen5测试环境用到各种产品

PCIe Gen5分析和测试常用到哪些工具

- 协议分析仪

PCIe Gen5分析和测试常用到哪些工具

- 协议分析仪的组成
  - 协议分析仪主机 – 专注于抓包、处理、分析
  - 各种PCIe接口的插板(Interposer) – 将信号引到旁路的协议分析仪主机分析
    - AIC插卡，支持x16, x8, x4
    - U.2
    - U.3
    - E1.S
    - E1.L
    - E3.S
    - E3.L
    - M.2
    - OCP
    - MCIO cable
    - HD Mini SAS cable
    - Slim SAS cable
    - Oculink cable
    - 光模块接口？


PCIe Gen5分析和测试常用到哪些工具

- EDSFF 接口

EDSFF Family

- Family of form factors and standards for data center NVMe SSDs
- E1.S for scalable & flexible performance storage
- E1.L for high capacity storage (e.g. QLC)
- E3 high performance SSD for 2U server / storage

PCIe Gen5分析和测试常用到哪些工具

- MCIO接口
  - MCIO cables are designed for data center, networking and telecommunications markets that use SAS, PCIe, Ethernet and other signal protocols. The solution can support cable to board and card to board applications in system, which include chip to chip, chip to module, chip to board and card edge option.
  - Mini Cool Edge IO(MCIO) is a flexible, robust and high performance connector and cable assembly solution that helps server and networking equipment design flexibility, reduces overall space, and extends the reach for high data rate signals. MCIO cable assemblies are provided with both discrete and ribbon raw cable 34AWG to 30AWG.

PCIe Gen5分析和测试常用到哪些工具

- Gen5 MCIO – NVMe SSD转接线缆

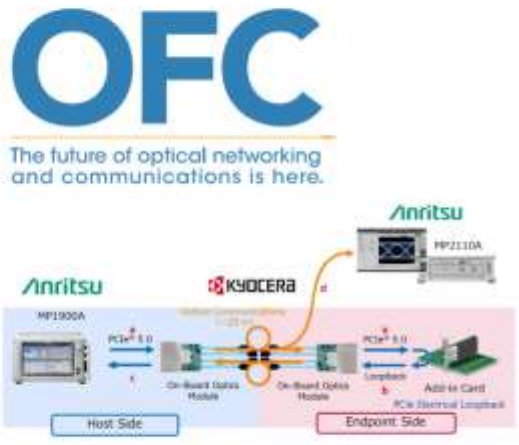
- MCIO to U.2 (1x4)
- MCIO to U.2 (2X2)
- MCIO to U.3 (1x4)
- MCIO to U.3 (2X2)
- MCIO to EDSFF(1x4)
- MCIO to EDSFF (2X2)

PCIe Gen5分析和测试常用到哪些工具

- Gen5 MCIO to MCIO cable

PCIe Gen5分析和测试常用到哪些工具

- PCIe Gen5光纤通讯接口
  - 安立公司和KYOCERA公司使用安立MP1900A信号质量分析仪和KYOCERA的板载光模块成功完成了PCI Express®5.0（PCIe®5.0）光传输测试。这是世界上首次*A使用端点附加卡（AIC）将PCIe®5.0（32 Gbps传输速度）电信号转换为光信号的测试。安立公司和KYOCERA已于**2023年3月7日至9日**在美国加州圣地亚哥的OFC2023现场公开演示测试



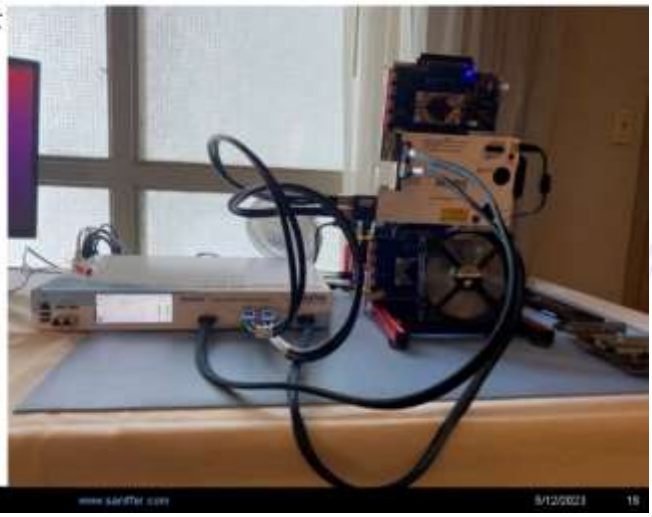
PCIe Gen5分析和测试常用到哪些工具

- PCIe Gen5分析仪的interposer展示

PCIe Gen5分析和测试常用到哪些工具

- 协议分析仪典型连接场景
  - Gen5 x16插卡

信号问题
解码速度
存储速度



PCIe Gen5分析和测试常用到哪些工具

- 协议分析仪典型连接场景
  - Gen5 x4 E3.S SSD
  - Gen5 x4 U.2 SSD

远程协作
远程分析
随时断网
Web管理

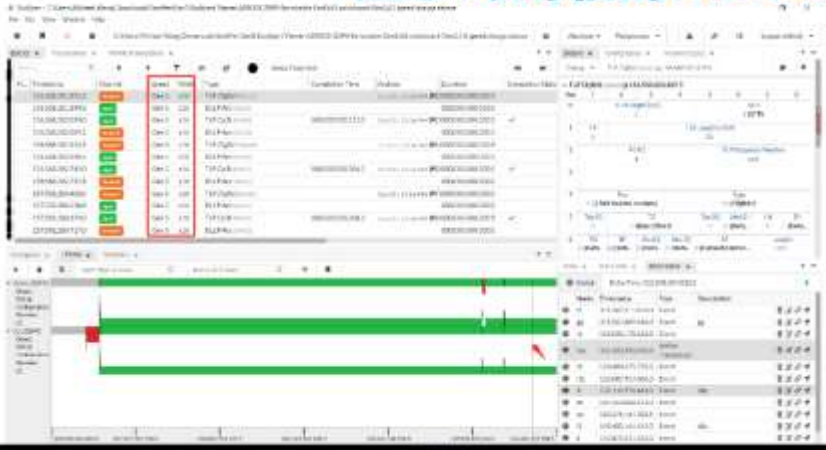
PCIe Gen5分析和测试常用到哪些工具

- 协议分析仪典型连接场景
  - Gen5 x4 M.2 SSD

低功耗分析
无需抓取上
电过程
配件经济

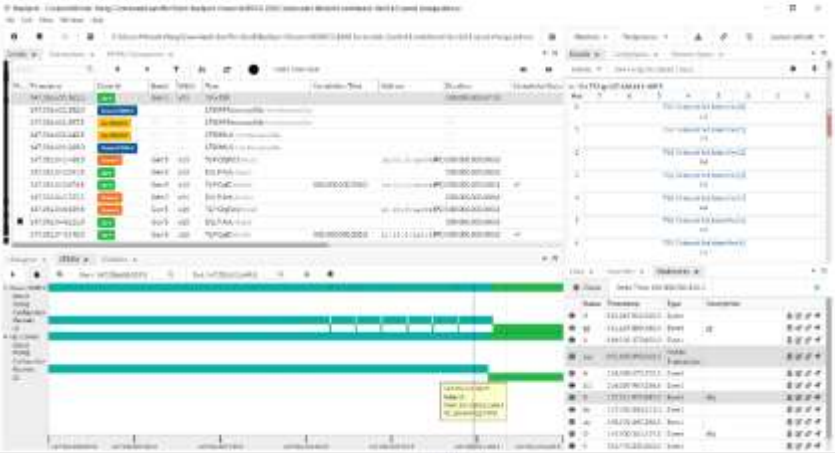
PCIe Gen5分析和测试常用到哪些工具

- PCIe协议分析仪解码界面 基于时间轴的LTSSM分析



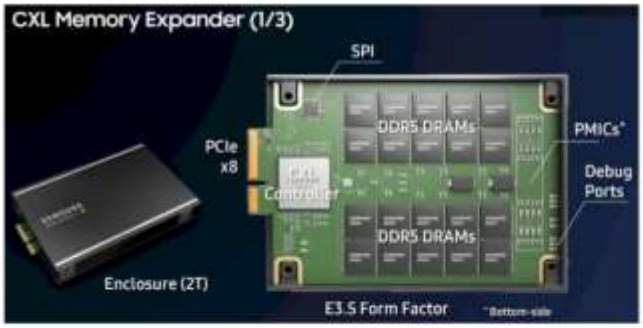
PCIe Gen5分析和测试常用到哪些工具

- LTSSM状态机变化



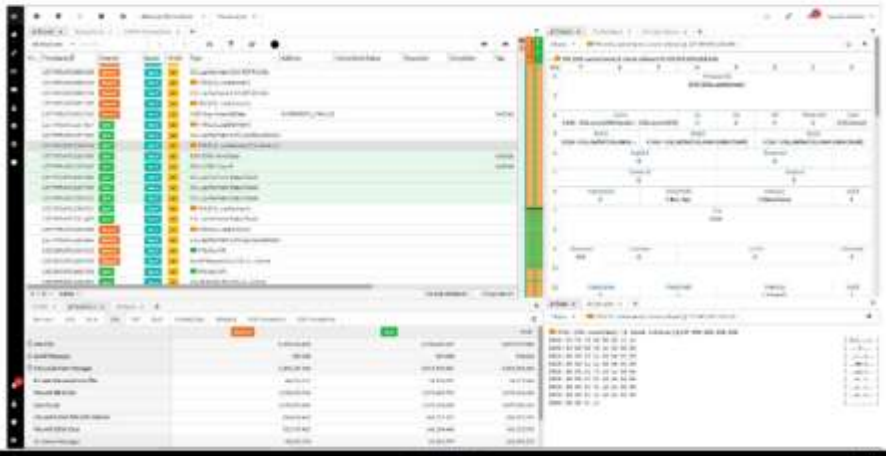
PCIe Gen5分析和测试常用到哪些工具

- Samsung CXL Server Memory Expander - Gen5 x8 E3.S
  - 使用SerialTek PCIe/CXL协议分析仪 + SerialCables Gen5 x8 E3.S/AIC adapter



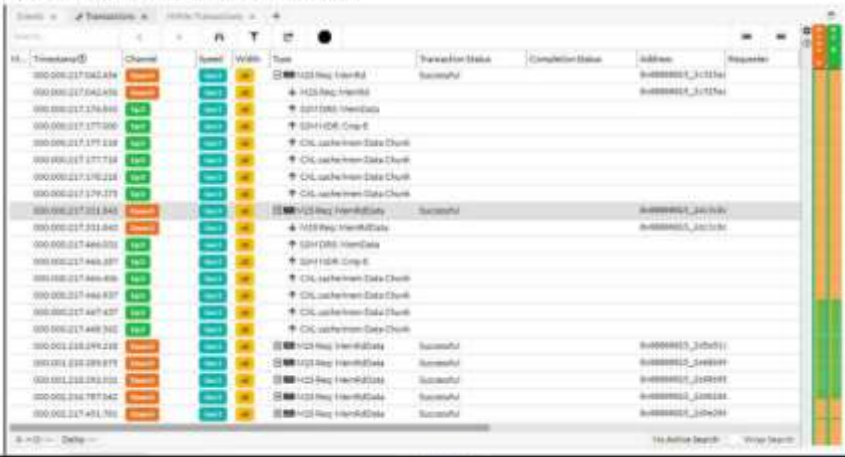
PCIe Gen5分析和测试常用到哪些工具

- CXL解码界面 - events



PCIe Gen5分析和测试常用到哪些工具

- CXL解码界面 - transactions

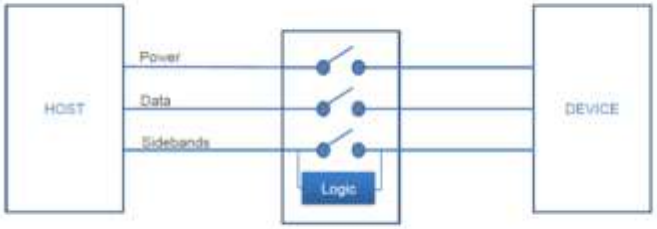


PCIe Gen5分析和测试常用到哪些工具

- 示波器，误码仪，VNA**
- 协议分析仪
- 掉电测试、热插拔、异常故障注入
- 电压拉偏和功耗测试设备
- 电压、电流、功耗、 Side-band边带信号监测设备
- SSD测试设备**
- 搭建PCIe Gen5测试环境用到各种产品

PCIe Gen5分析和测试常用到哪些工具

- 掉电测试、热插拔、异常故障注入
  - “热插拔”仅对企业级SSD试用，一般是U.2/U.3/E1.S/E1.L/E3.S/E3.L或者SAS以及企业级SATA SSD。
  - 针对M.2, AIC插卡我们一般称该模块为Card Control Module （插卡控制模块），导入物理层和链路层问题。



PCIe Gen5分析和测试常用到哪些工具

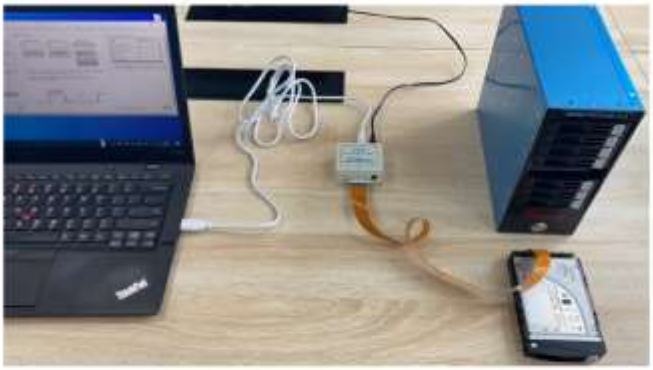
- 掉电测试、热插拔、异常故障注入 （以SSD盘为例，插卡类似）
  - 模拟盘的热插拔
  - 模拟盘热插拔过程中导致的pin bounce接触不好的情况
  - 模拟某些针脚断掉
  - 模拟某些针脚长通
  - 模拟某些针脚上面有信号毛刺
    - 物理毛刺的多少？注入一次毛刺，还是一直有毛刺？间隔时间多长？
    - 毛刺的高低，疏密，持续的时间长短
  - 模拟某个Lane中的某些差分信号有毛刺，或者某个Lane不通
  - 模拟非常快速的通/断测试
  - 模拟各个边带信号拉高或者拉低，例如PERST #, CLKREQ#等

PCIe Gen5分析和测试常用到哪些工具

- 掉电测试、热插拔、异常故障注入

PCIe Gen5分析和测试常用到哪些工具

- 掉电测试、热插拔、异常故障注入
  - U.2 SSD – NVMe SSD



PCIe Gen5分析和测试常用到哪些工具

- 掉电测试、热插拔、异常故障注入
  - Modules available for U.2, U.3, E1, E3 and more
  - Also relevant to fixed interfaces (AIC and M.2)
    - Lane width reduction
    - Power cycle / reset (Driving PERST)
  - Fault injection tests
    - Simulate bad cables, damaged connectors and data corruption

Real world capture of pin-bounce

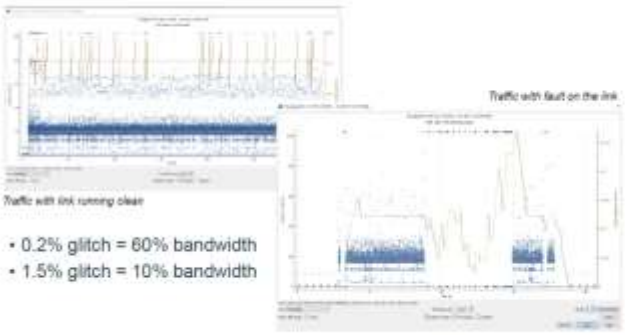
PCIe Gen5分析和测试常用到哪些工具

- 掉电测试、热插拔、异常故障注入
  - 插卡

PCIe Gen5分析和测试常用到哪些工具

- 掉电测试、热插拔、异常故障注入

• Random 'glitch' (physical layer interruption) of data
• Glitch long enough to hit a packet, but not enough to take down the link



• 0.2% glitch = 60% bandwidth
• 1.5% glitch = 10% bandwidth

PCIe Gen5分析和测试常用到哪些工具

- 示波器，误码仪，VNA**
- 协议分析仪
- 掉电测试、热插拔、异常故障注入
- 电压拉偏和功耗测试设备
- 电压、电流、功耗、 Side-band边带信号监测设备
- SSD测试设备**
- 搭建PCIe Gen5测试环境用到各种产品

PCIe Gen5分析和测试常用到哪些工具

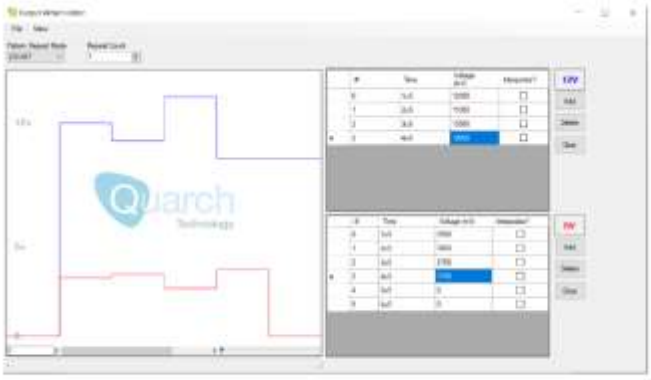
- 电压拉偏和功耗测试设备
  - PPM - Programmable Power Module

PCIe Gen5分析和测试常用到哪些工具

- 电压拉偏和功耗测试设备
  - PPM - Programmable Power Module

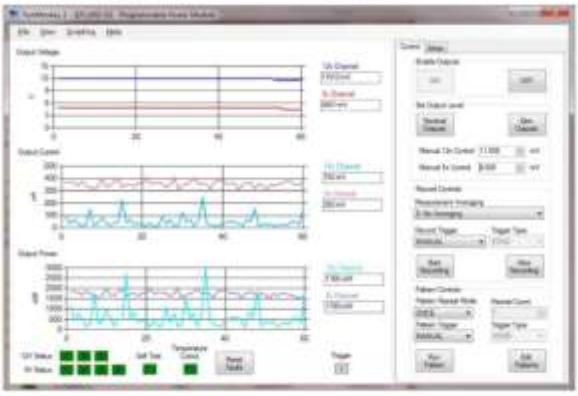
PCIe Gen5分析和测试常用到哪些工具

- 电压拉偏和功耗测试设备
  - PPM - Programmable Power Module



PCIe Gen5分析和测试常用到哪些工具

- 电压拉偏和功耗测试设备
  - PPM - Programmable Power Module



PCIe Gen5分析和测试常用到哪些工具

- 示波器，误码仪，VNA**
- 协议分析仪
- 掉电测试、热插拔、异常故障注入
- 电压拉偏和功耗测试设备
- 电压、电流、功耗、 Side-band边带信号监测设备
- SSD测试设备**
- 搭建PCIe Gen5测试环境用到各种产品

PCIe Gen5分析和测试常用到哪些工具

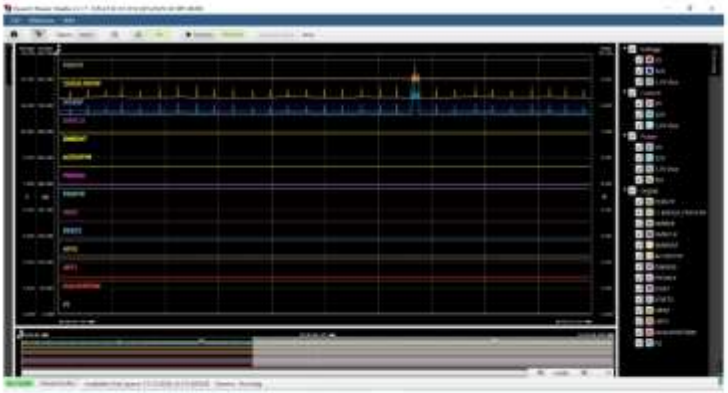
- 电压、电流、功耗、 Side-band边带信号监测设备

PCIe Gen5分析和测试常用到哪些工具

- 电压、电流、功耗、 Side-band边带信号监测设备
  - 插卡

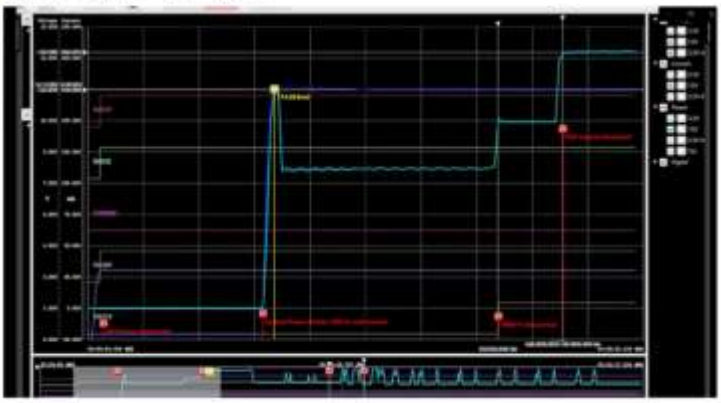
PCIe Gen5分析和测试常用到哪些工具

- 电压、电流、功耗、 Side-band边带信号监测设备
  - GUI



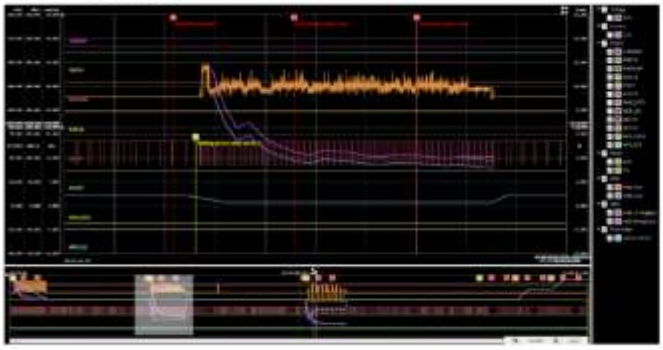
PCIe Gen5分析和测试常用到哪些工具

- 电压、电流、功耗、 Side-band边带信号监测设备
  - 放大、缩小、看区间值



PCIe Gen5分析和测试常用到哪些工具

- 电压、电流、功耗、 Side-band边带信号监测设备
  - 10 watts average
  - 45k IOPS average
  - 195 MB/s average



PCIe Gen5分析和测试常用到哪些工具

- 示波器，误码仪，VNA**
- 协议分析仪
- 掉电测试、热插拔、异常故障注入
- 电压拉偏和功耗测试设备
- 电压、电流、功耗、 Side-band边带信号监测设备
- SSD测试设备**
- 搭建PCIe Gen5测试环境用到各种产品

PCIe Gen5分析和测试常用到哪些工具

- SSD测试设备
  - NVMe SSD作为PCIe的一个重要应用，从2012年以来获得广泛应用

PCIe Gen5分析和测试常用到哪些工具

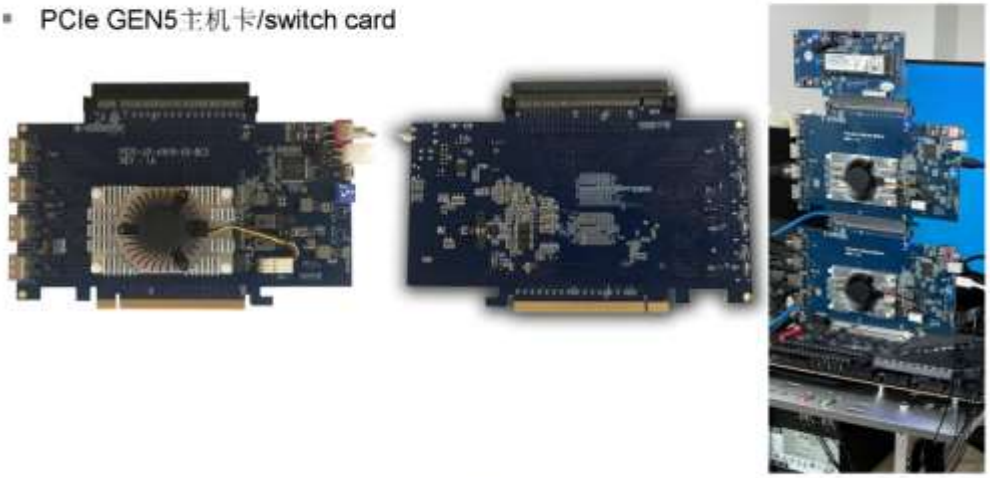
- NVMe 预封装测试脚本（涵盖18大类测试, ~1000个测试用例）
  - NVMe Commands test
  - NVMe I/O Tests
  - NVMe Resets (all supported reset methods)
  - NVMe Namespace Management
  - NVMe Basic Management Commands
  - NVMe-MI Full Command Set
  - NVMe Dual Port Drive Tests
  - NVMe SBExpress Hotplug and Link Testing
  - NVMe Quarch Testing Pull/Plug Glitch
  - NVMe Miscellaneous Commands (e.g. SR-IOV)
  - NVMe ZNS test
  - NVMe VDM test
  - NVMe Clocking Mode Test (SRIS)
  - NVMe TCG Opal/Ruby test
  - NVMe DSSD conformance test
  - UNH IOL NVMe Certification
  - UNH IOL NVMe-MI Certification
  - SSD Endurance JEDEC Spec. (long runtime)

PCIe Gen5分析和测试常用到哪些工具

- 示波器，误码仪，VNA**
- 协议分析仪
- 掉电测试、热插拔、异常故障注入
- 电压拉偏和功耗测试设备
- 电压、电流、功耗、 Side-band边带信号监测设备
- SSD测试设备**
- 搭建PCIe Gen5测试环境用到各种产品

搭建PCIe Gen5测试环境用到各种产品

- PCIe GEN5主机卡/switch card



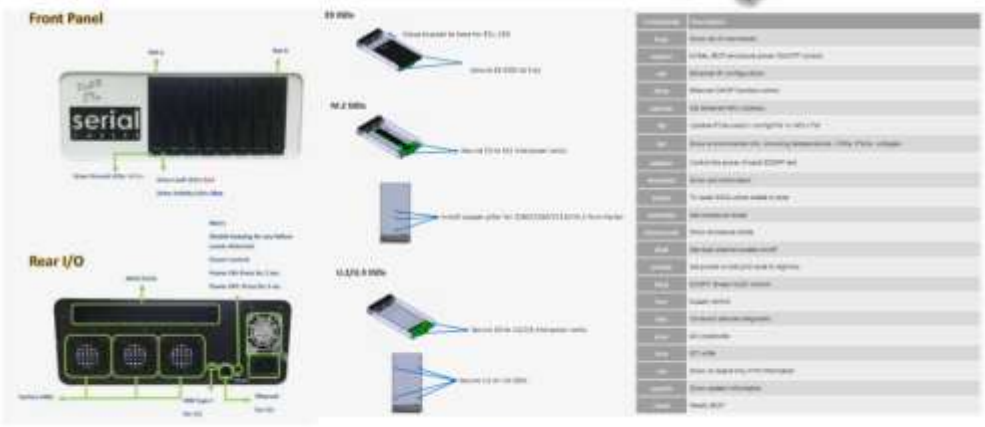
搭建PCIe Gen5测试环境用到各种产品

- 主板插槽信号和Gen5 switch插槽信号

Z690主板Gen5 x16插槽出来的信号　　Gen5 x16 switch卡顶部出来的信号

搭建PCIe Gen5测试环境用到各种产品

- PCIe GEN5主机卡/switch card + Gen5 JBOF



搭建PCIe Gen5测试环境用到各种产品

- PCIe GEN5转接卡/适配卡
  - U.2/U.3 转接卡
  - PCIe GEN5 M2/U2 ADAPTERS
  - PCIe GEN5 M2/AIC ADAPTERS
  - PCIe GEN5 EDSFF ADAPTERS
  - PCIe GEN5 OTHER ADAPTERS
    - Lane re-assignment, reversal, reducer

搭建PCIe Gen5测试环境用到各种产品

- PCIe GEN5 转接线/延长线
  - GEN5 MCIO CABLES
  - GEN5 EDSFF CABLES
  - GEN5 U.2 CABLES
  - GEN5 SlimSAS CABLES
  - GEN5 PCIE CEM CABLES

搭建PCIe Gen5测试环境用到各种产品

- PCIe GEN5 转接线/延长线的物理测量

搭建PCIe Gen5测试环境用到各种产品

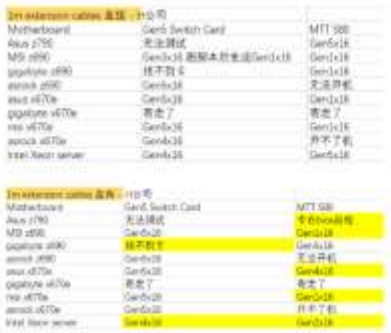
- PCIe GEN5 x16延长线的实际场景测试

0.5m extension cables		
Motherboard	Gen5 Switch Card	MIT G80
Asus z790	无法测试（不能延长线也开不了机）	找不到device
MSI z690	Gen3x16，跑测速只能到Gen4	找不到device
gigabyte z690	找不到device，不接延长线能找到device	找不到device
asrock z690	gen3 x16 运行脚本后lspci -vvv -s 去定查看，	找不到device
asus x670e	Gen5x8	找不到device
gigabyte x670e	Gen3 x16 运行脚本后只能到gen4	Gen4 x16
msi x670e	Gen4x16，运行脚本转换成Gen3，但是识别gen4	Gen4 x16
asrock x670e	找不到device	找不到device
Intel Xeon server	Gen4 x16 脚本也到不到gen5	找不到device

A公司 – 非常差（尽管其PCB增加了Gen5 re-driver)芯片

搭建PCIe Gen5测试环境用到各种产品

- PCIe GEN5 x16延长线的实际场景测试



H公司 – 改良后质量提升有限，

** 增加了中间两列，Gen5 x4 E3/M.2 SSD

** 上述E3.S和M.2 SSD通过SerialCables Gen5转接卡转接成插卡后再连接延长线

H公司 – 相对于A公司质量好一点 – Passive设计

搭建PCIe Gen5测试环境用到各种产品

- PCIe GEN5 x16延长线的实际场景测试

A线1米	switch	E3	M.2	moore
supermicro 13	Gen5x16	Gen5x4	Gen5x4	Gen5x16
intel	Gen5x16	Gen5x4	Gen4x4	Gen5x16
Asrock X670E	Gen5x16	Gen5x4	Gen5x4	Gen5x16
Asrock Z690	Gen5x16	Gen5x4	Gen5x4	找不到设备

A线0.5米	switch	E3	M.2	moore
supermicro 13	Gen5x16	Gen5x4	Gen5x4	Gen5x16
intel	Gen5x16	Gen5x4	Gen5x4	Gen5x16
Asrock X670E	Gen5x16	Gen5x4	Gen5x4	Gen5x16
Asrock Z690	Gen5x16	Gen5x4	Gen5x4	Gen5x16

L公司，A线缆改良后质量较好

B线1米	switch	E3	M.2	moore
supermicro 13	Gen5x16	Gen4x4	Gen5x4	Gen1x16
intel	Gen4x16	Gen4x4	Gen5x4	Gen2x16
Asrock X670E				Gen5x16
Asrock Z690				显示器没信号

B线0.5米	switch	E3	M.2	moore
supermicro 13	Gen5x16	Gen5x4	Gen5x4	Gen5x16
intel	Gen5x16	Gen2x4	Gen5x4	Gen5x16
Asrock X670E				Gen5x16
Asrock Z690				显示器没信号

L公司，改良B线缆，质量还是不大好

** 上述E3.S和M.2 SSD通过SerialCables Gen5转接卡转接成插卡连接延长线

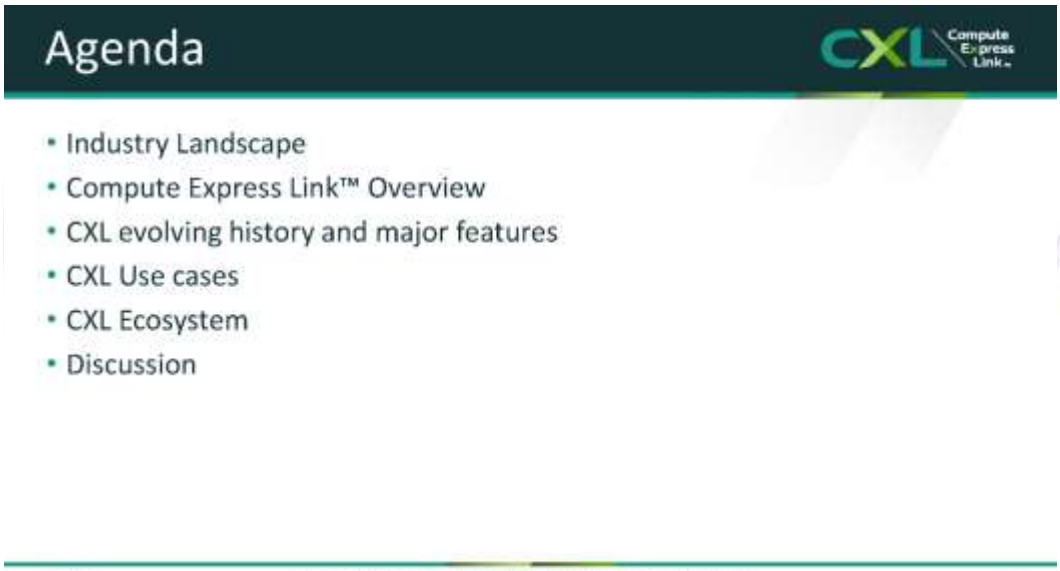
搭建PCIe Gen5测试环境用到各种产品

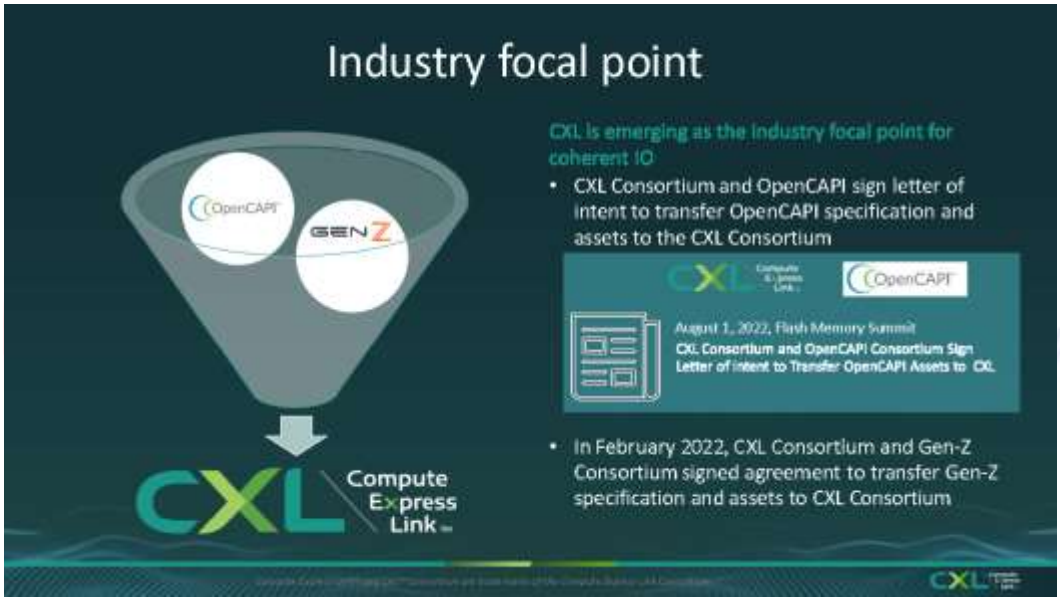
- Retimer卡和CXL卡

搭建PCIe Gen5测试环境用到各种产品

- PCIe5.0, CXL, NVMe, NAND, DDR5, UFS4测试技术和工具白皮书Ver 8.2链接:
  https://pan.baidu.com/s/1YPmVg0eqWiuKdShm2CYq3Q?pwd=egqt
- 有其它问题请添加微信，或者Sniffer公司公众号

15.2 CXL 1.1/2.0/3.0 技术研讨





Caching Overview

- Caching temporarily brings data closer to the consumer
- Improves latency and bandwidth using prefetching and/or locality
- Definitions for cache use:
  - Prefetching: Loading Data into cache before it is required
  - Spatial Locality (locality is space): Access address X then X+n
  - Temporal Locality (locality in Time): Multiple access to the same Data

CXL™ Approach

Coherent Interface

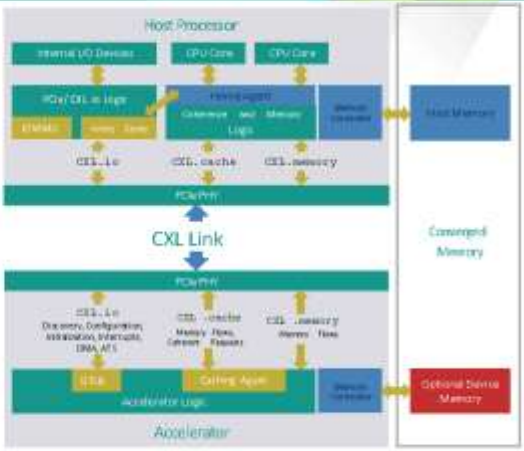
- Leverages PCIe with three multiplexed protocols
- Built on top of PCIe® infrastructure

Low Latency

- CXL.Cache/CXL.Memory targets near CPU cache coherent latency (<200ns load to use)

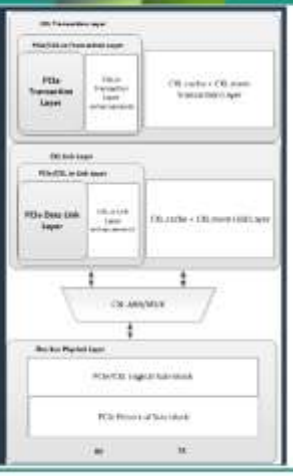
Asymmetric Complexity

- Eases burdens of cache coherence interface designs for devices

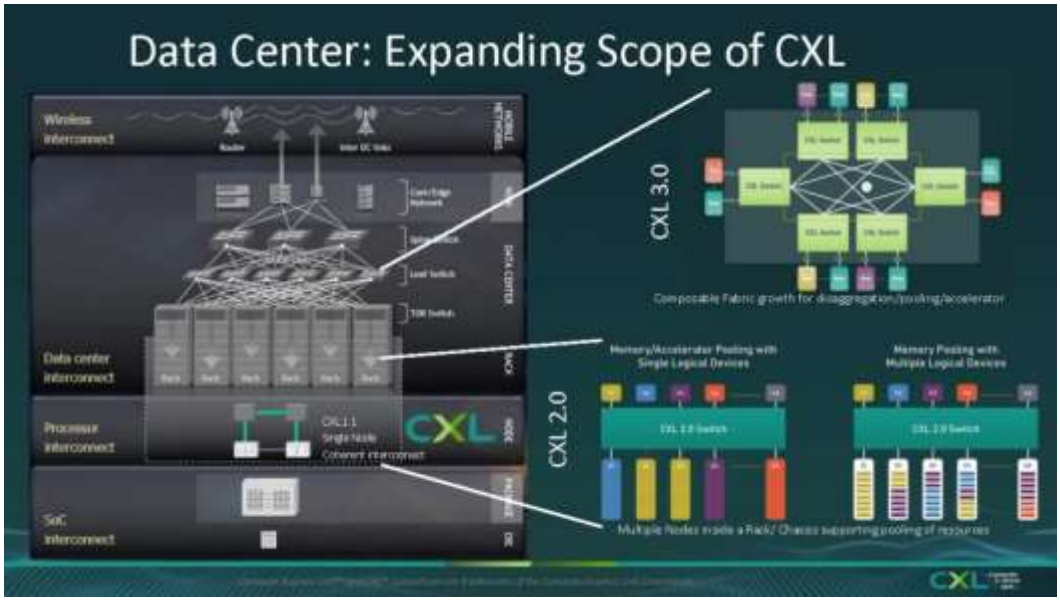


CXL Layering of Protocols

- 3 protocols sharing same PHY but different Link/ Transaction Layers
- Flits (528 b) is basic unit of transfer – protocol muxing at flit level
  - Protocol ID (16b) included; CRC (16b) additional for each flit (544 b)
- CXL.io: TLPs/ DLLPs as PCIe does – overlaid on payload part of the flit
  - Can be interrupted by CXL.Cache/ CXL.Memory at flit boundary - do not want a large DMA to block latency-sensitive protocols such as cache and memory
  - Optimized for large payload DMA transfers
  - Expect to leverage the PCIe stack
- CXL.cache and CXL.memory – natively flit based



CXL Specification Release Timeline

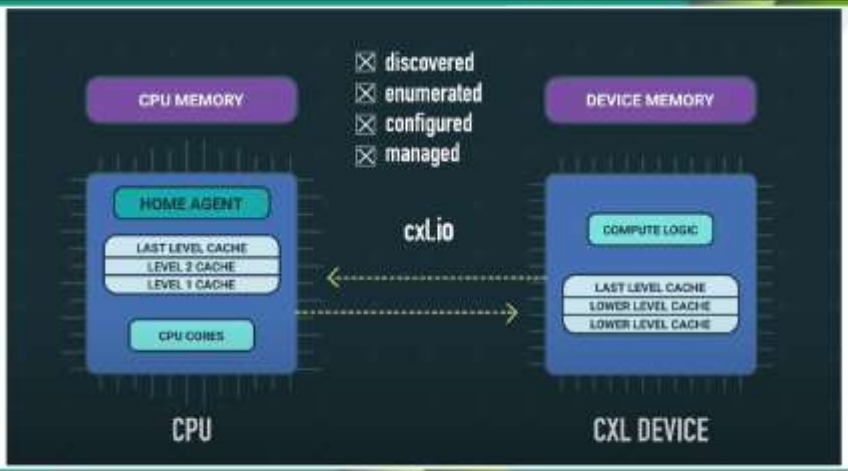


CXL3.0 Spec Feature Summary

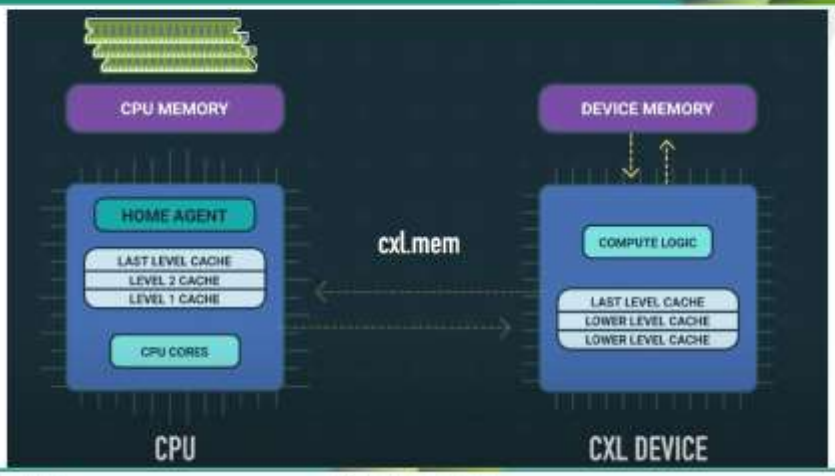
Features	CXL 1.0 / 1.1	CXL 2.0	CXL 3.0
Release date	2019	2020	1H 2022
Max link rate	32GTs	32GTs	64GTs
Flit 68 byte (up to 32 GTs)	✓	✓	✓
Flit 256 byte (up to 64 GTs)			✓
Type 1, Type 2 and Type 3 Devices	✓	✓	✓
Memory Pooling w/ MLDs		✓	✓
Global Persistent Flush		✓	✓
CXL IDE		✓	✓
Switching (Single-level)		✓	✓
Switching (Multi-level)			✓
Direct memory access for peer-to-peer			✓
Symmetric coherency (256 byte flit)			✓
Memory sharing (256 byte flit)			✓
Multiple Type 1/Type 2 devices per root port			✓
Fabric capabilities (256 byte flit)			✓

Not supported
✓ Supported

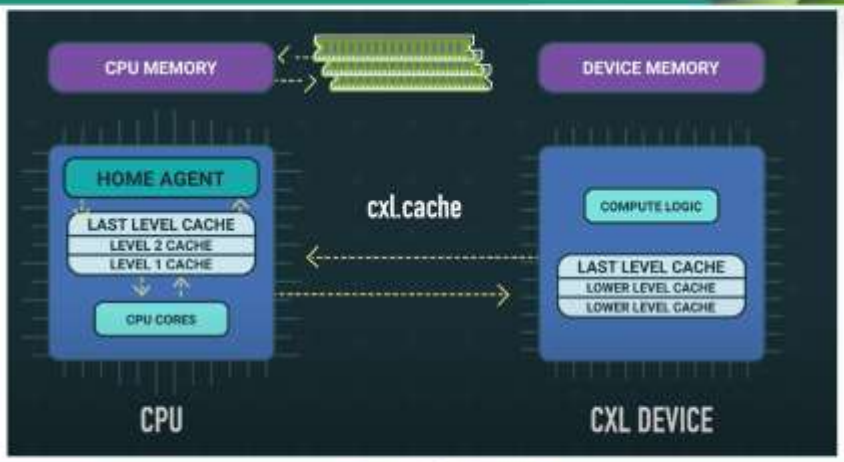
CXL Protocol Overview: CXL.io



CXL Protocol Overview: CXL.mem



CXL Protocol Overview: CXL.cache

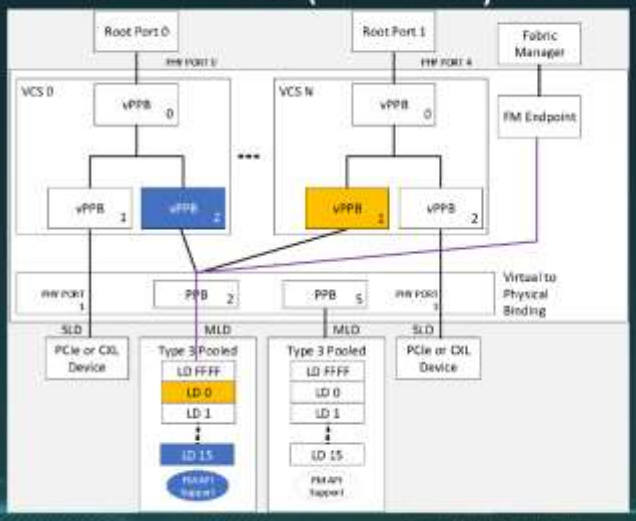


CXL 2.0 Switching Benefits – Pooling

Memory/Accelerator Pooling with Single Logical Devices

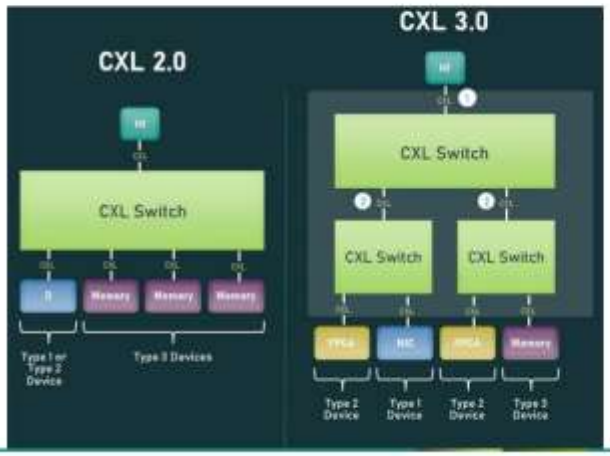
Memory Pooling with Multiple Logical Devices

CXL 2.0 HBR(Hierarchy Based Routing) Switch



FM Binds a Logical Device to VCS VPPB
Architecture allows for > 16 RPs on a switch with flexible binding to Logical Devices on a link

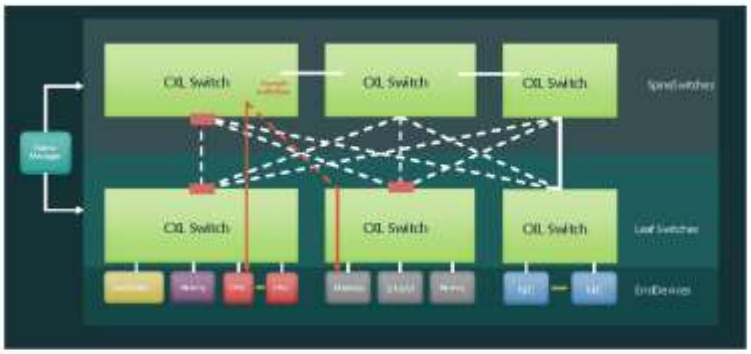
CXL 3.0: Multiple Level Switching, MultipleType-1/2



1. Each host's root port can connect to more than one device type (up to 16 CXL.cache devices)
2. Multiple switch levels (aka cascade)
   - Supports fanout of all device types

CXL 3.0 PBR(Port Based Routing) Switch

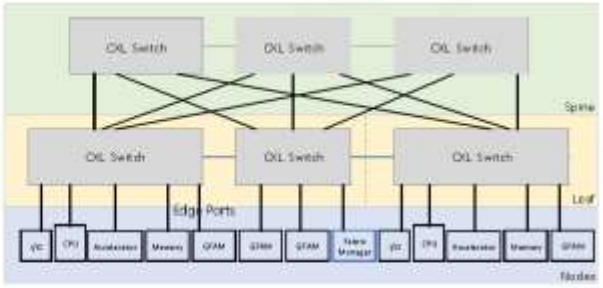
Composable Systems with Spine/Leaf Architectureat Rack/Pod Level



CXL 3.0 Fabric Architecture
- Interconnected Spine Switch System
- Leaf Switch NIC Enclosure
- Leaf Switch CPU Enclosure
- Leaf Switch Accelerator Enclosure
- Leaf Switch Memory Enclosure

CXL Fabric - Basics and Scope

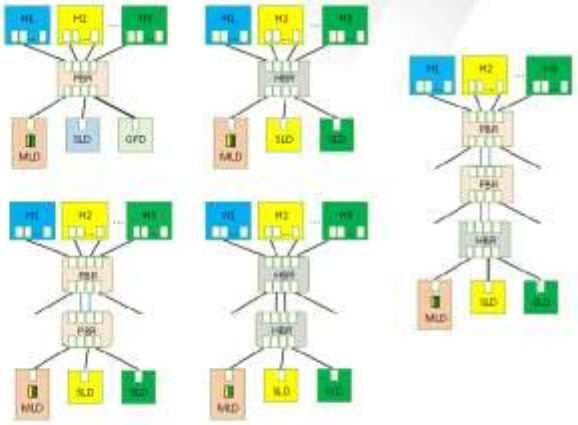
- PBR identifiers (PIDs)
  - 4096 PIDs (12-bits)
  - Source PID (SPID)
  - Destination PID (DPID)
- Most edge ports are assigned a unique PID
- G-FAM devices (GFDs)
  - Scalable memory resource
  - Accessible by all host and devices in the cluster
- Load/store memory semantics



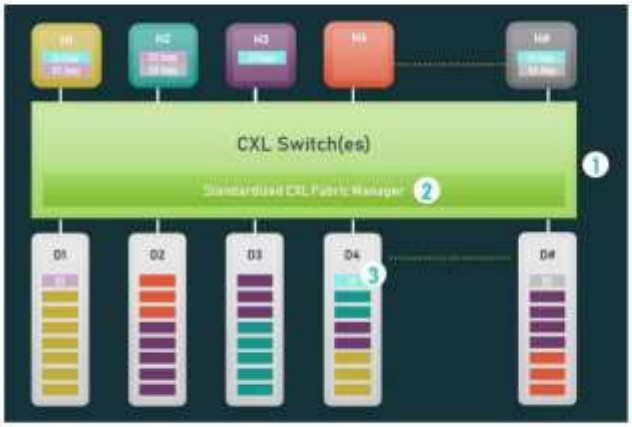
CXL Fabric - Port-Based Routing

- CXL 3.0 augments the previously defined Hierarchy-Based Routing with Port-Based Routing
- Interoperability is defined for configurations with a combination of Hierarchy- and Port-Based Routing CXL switches

Acronyms
- FAM = Fabric-Attached Memory
- HBR = Hierarchy-Based Routing
- PBR = Port-Based Routing
- SLD = Single Logical Device
- MLD = Multi-Logical Device
- GFD = Global FAM Device

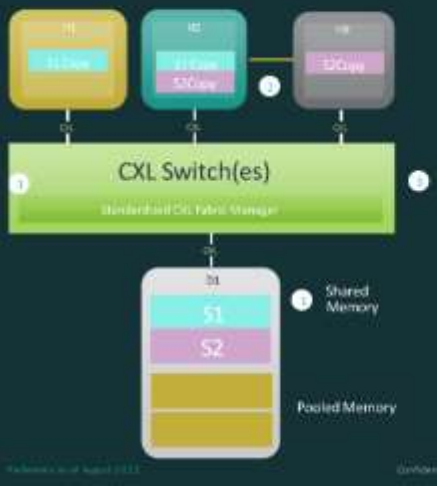


CXL 3.0: Pooling & Sharing



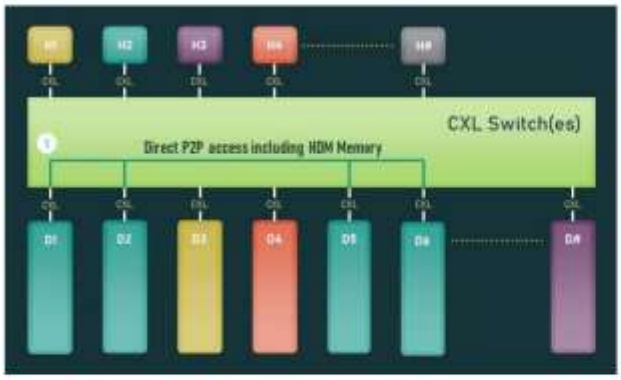
1. Expanded use case showing memory sharing and pooling
2. CXL Fabric Manager is available to setup, deploy, and modify the environment
3. Shared Coherent Memory across hosts using hardware coherency (directory + Back-Invalidate Flows). Allows one to build large clusters to solve large problems through shared memory constructs. Defines a Global Fabric Attached Memory (GFAM) which can provide access to up to 4095 entities

CXL3.0: COHERENT MEMORY SHARING



1. Device memory can be shared by all hosts to increase data flow efficiency and improve memory utilization
2. Host can have a coherent copy of the shared region or portions of shared region in host cache
3. CXL 3.0 defined mechanisms to enforce hardware cache coherency between copies

CXL3.0 Protocol Enhancements (UIO and BI) for Device to Device Connectivity



CXL 3.0 enables **non-tree topologies and peer-to-peer communication (P2P)** within a virtual hierarchy of devices

- Virtual hierarchies are associations of devices that maintains a coherency domain
- P2P to HDM-DB memory is I/O Coherent: a new Unordered I/O (UIO) Flow in CXL.io – the Type-2/3 device that hosts the memory will generate a new Back-Invalidation flow (CXL.Mem) to the host to ensure coherency if there is a coherency conflict

CXL 3.0 Conclusions

CXL 3.0 features

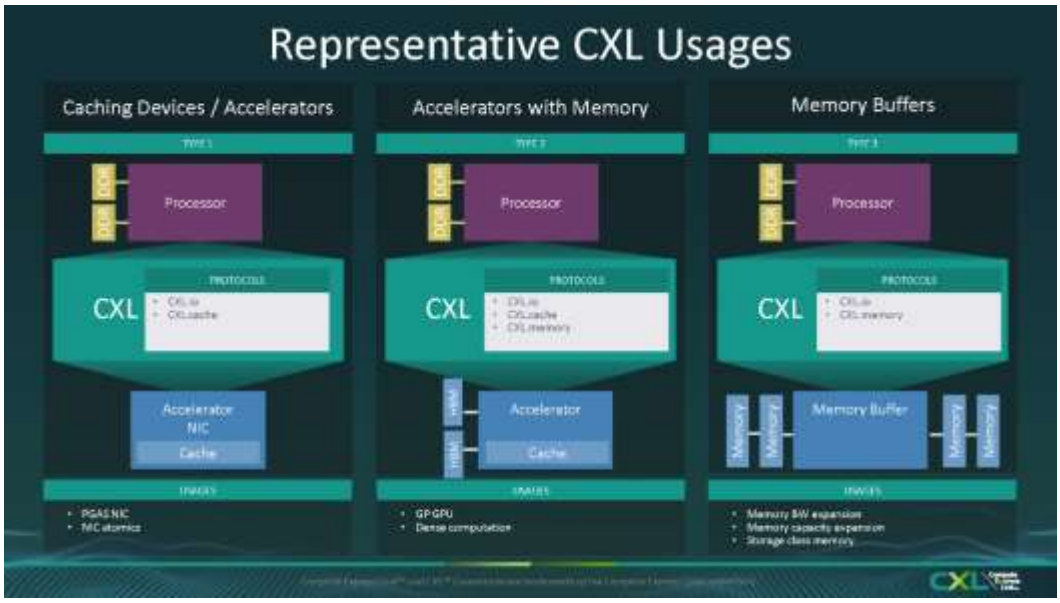
- Full fabric capabilities and Fabric management
- Expanded switching topologies
- Symmetric coherency capabilities
- Peer-to-peer resource sharing
- Double the bandwidth and zero added latency compared to CXL 2.0
- Full backward compatibility with CXL 2.0, CXL 1.1, and CXL 1.0

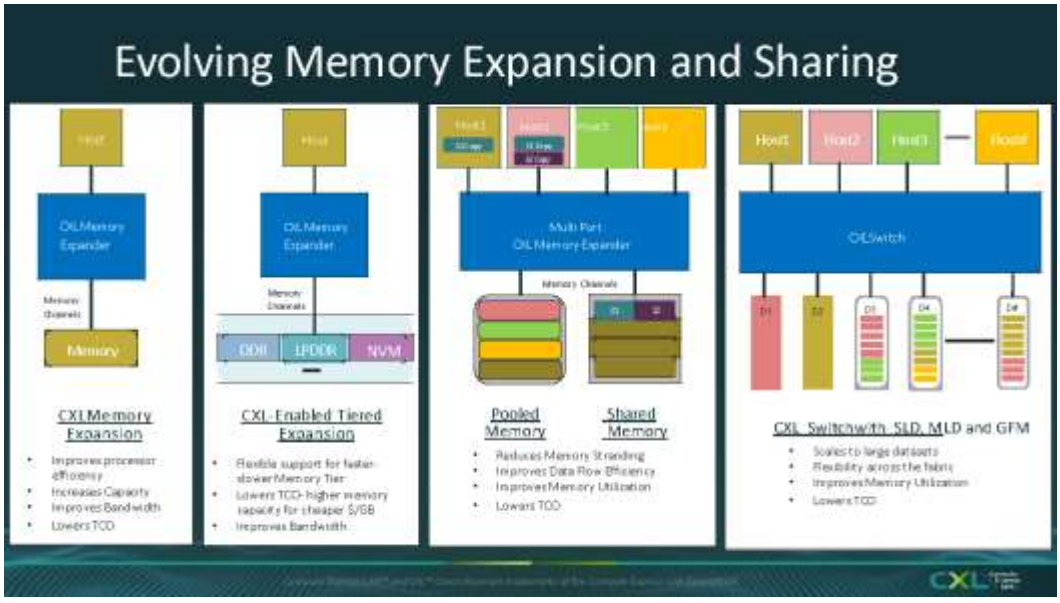
Enabling new usage models

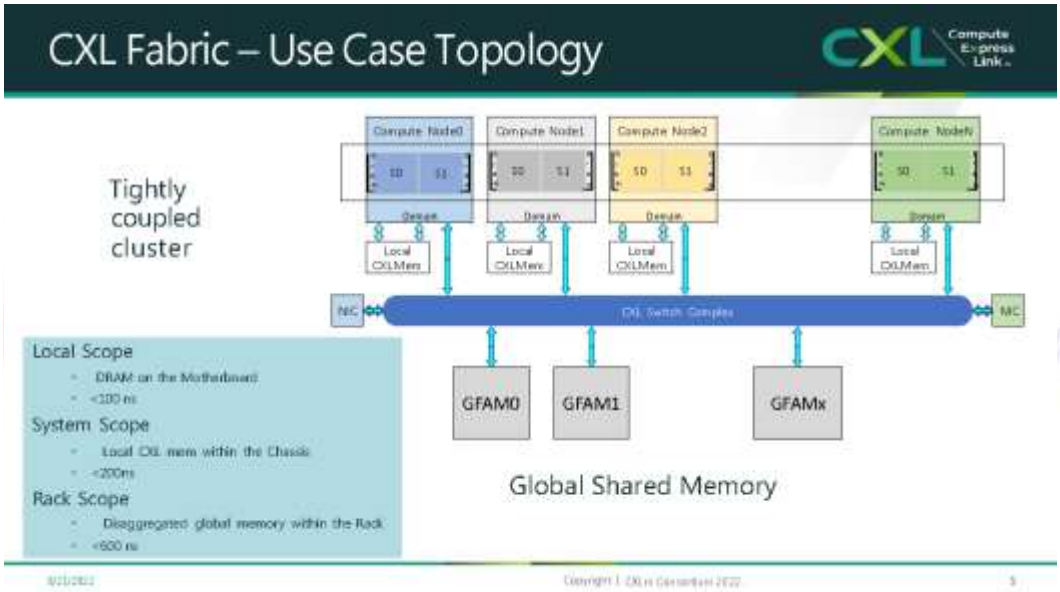
- Memory sharing between hosts and peer devices
- Support for multi-headed devices
- [Symmetric coherency capabilities use case]
- Expanded support for Type-1 and Type-2 devices
- GFAM provides expansion capabilities for current and future memory

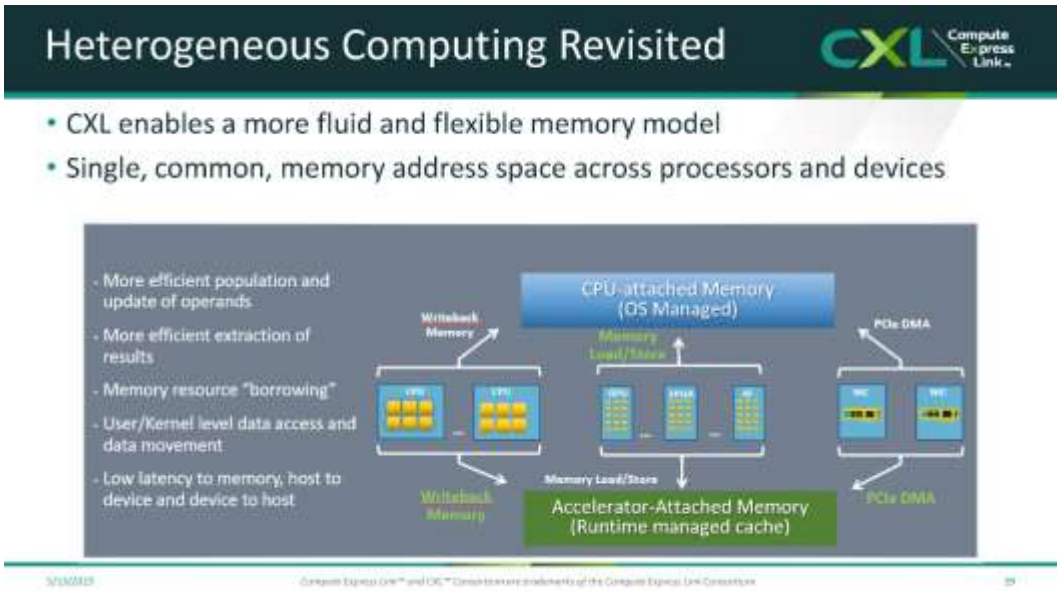
CXL Enables New Use-Cases

- Memory Expansion beyond CPU attached Memory
- Accelerators & NICs sharing memory with the Host CPU
- Modular Systems
- Disaggregrated computing



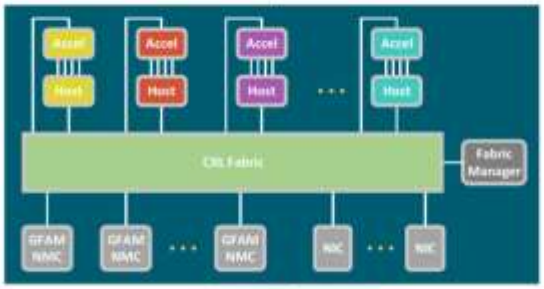

Evolving Memory Expansion and Sharing
CXL Memory Expansion
CXL Enabled Tiered Expansion
Pooled Memory
Shared Memory
CXL Switch with SLD, MLD and GFM


CXL Fabric – Use Case Topology
Tightly coupled cluster
Local Scope
System Scope
Rack Scope
GFAM0
GFAM1
GFAMx
Global Shared Memory


Heterogeneous Computing Revisited
• CXL enables a more fluid and flexible memory model
• Single, common, memory address space across processors and devices
- More efficient population and update of operands
- More efficient extraction of results
- Memory resource "borrowing"
- User/Kernel level data access and data movement
- Low latency to memory, host to device and device to host
CPU-attached Memory (OS Managed)
Accelerator-Attached Memory (Runtime managed cache)

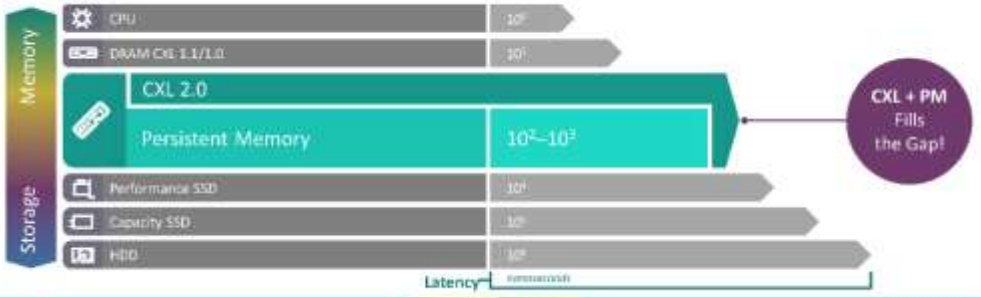
CXL Fabric – Near or In-Memory Compute

- HPC/AI/HPDA workloads consume significant amount of power
- Optimize compute to happen near the data
- Enable near and in-memory compute

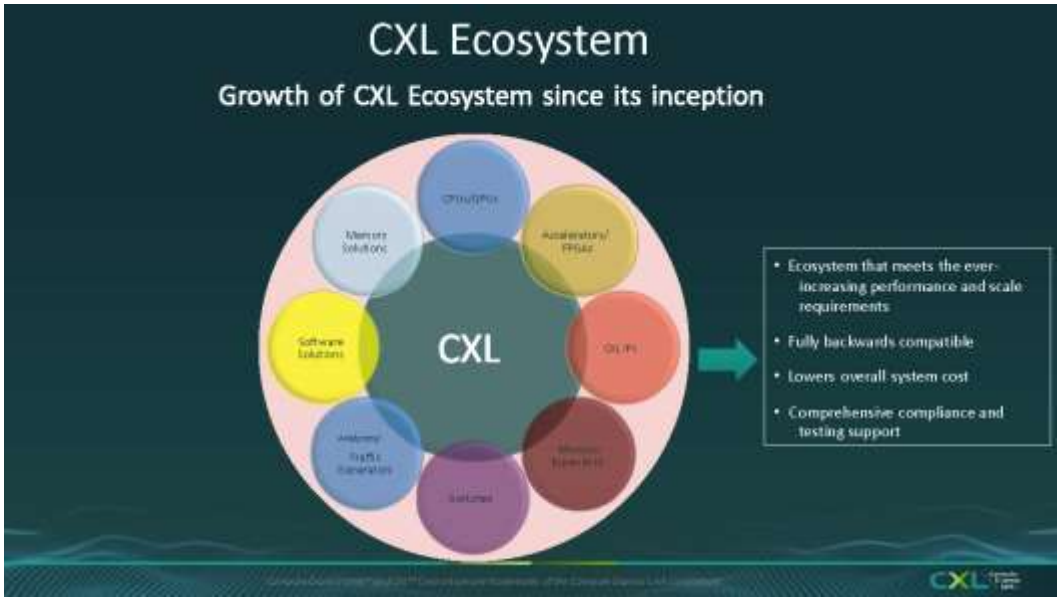


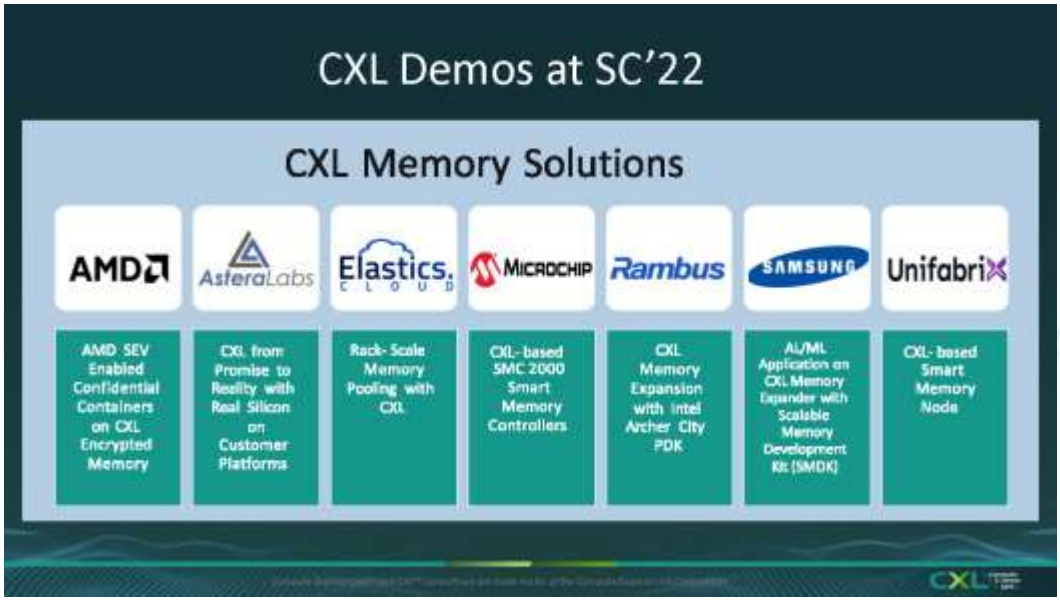
CXL 2.0 Benefits and Persistent Memory

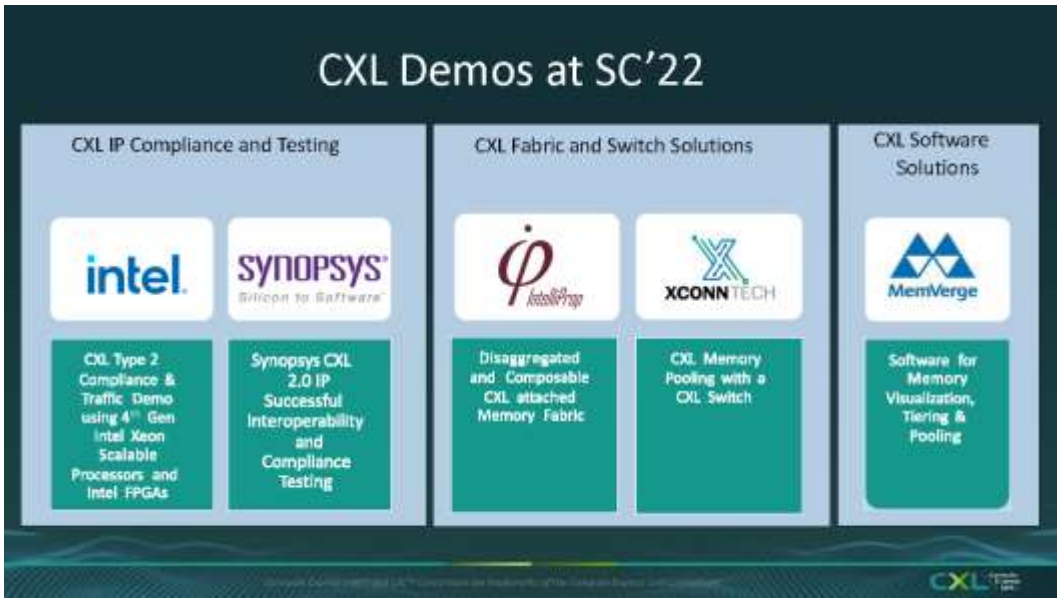
Moves Persistent Memory from Controller to CXL	Enables Standardized Management of the Memory and Interface	Supports a Wide Variety of Industry Form Factors



CXL Ecosystem

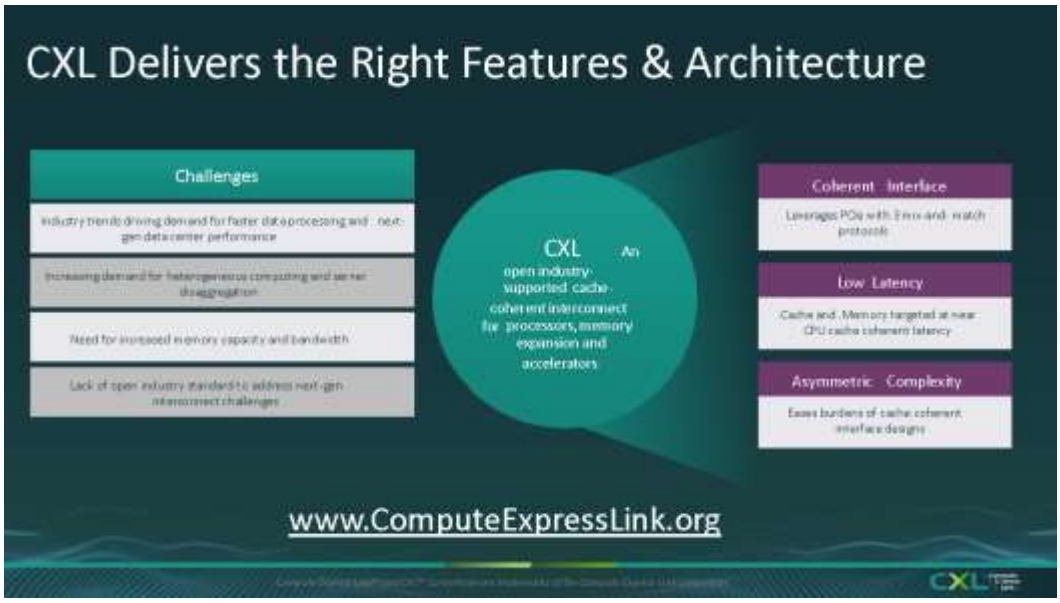

CXL Ecosystem
Growth of CXL Ecosystem since its inception
CXL
• Ecosystem that meets the ever-increasing performance and scale requirements
• Fully backwards compatible
• Lowers overall system cost
• Comprehensive compliance and testing support


CXL Demos at SC'22
CXL Memory Solutions
AMD
AsteraLabs
Elastics.
MICROCHIP
Rambus
SAMSUNG
UnifabriX
AMD SEV Enabled Confidential Containers on CXL Encrypted Memory
CXL from Promise to Reality with Real Silicon on Customer Platforms
Rack-Scale Memory Pooling with CXL
CXL-based SMC 2000 Smart Memory Controllers
CXL Memory Expansion with Intel Archer City PDK
AI/ML Application on CXL Memory Expander with Scalable Memory Development Kit (SMDK)
CXL-based Smart Memory Node

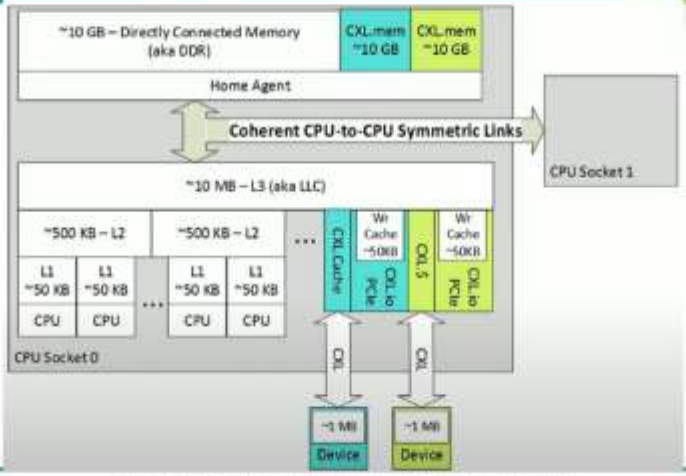

CXL Demos at SC'22
CXL IP Compliance and Testing
CXL Fabric and Switch Solutions
CXL Software Solutions
intel
synopsys
Silicon to Software
XCONN TECH
MemVerge
CXL Type 2 Compliance & Traffic Demo using 4th Gen Intel Xeon Scalable Processors and Intel FPGAs
Synopsys CXL 2.0 IP Successful Interoperability and Compliance Testing
Disaggregated and Composable CXL attached Memory Fabric
CXL Memory Pooling with a CXL Switch
Software for Memory Visualization, Tiering & Pooling

New opportunities

- AI/HPC
- Computational Storage for databases, etc
- Network accelerator / DPU
- DRAM memory and Stroage Class memory
- CXL Interface SSD
- Composable Architecture
- CXL VS. Proprietary Interfaces



Intel cache example



Cache Coherence

How do we make sure updates in cache are visible to other agents?

- Invalidate all peer caches prior to update
- Can managed with software or hardware → CXL uses hardware coherence

Define a point of "Global Observation" (aka GO) when new data is visible from writes

Tracking granularity is a "cacheline" of data → 64-bytes for CXL

Cache Coherence Protocol

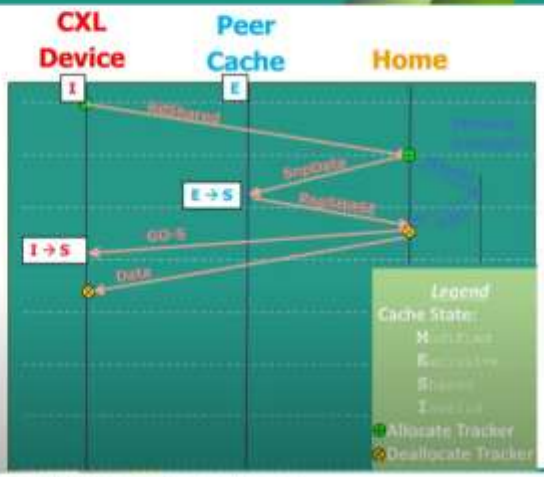
- Modern CPU caches are built on M,E,S,I protocol/states
  - **M**odified – Only in one cache, Can be read or written, Data **NOT** up-to-date in memory
  - **E**xclusive – Only in one cache, Can be read or written, Data **IS** up-to-date in memory
  - **S**hared – Can be in many caches, Can only be read, Data IS up-to-date in memory
  - **I**nvalid – Not in cache
- M,E,S,I is tracked for each cacheline address in each cache
  - Cacheline address in CXL is Addr[51:6]
- Notes:
  - Each level of the CPU cache hierarchy follows MESI and layers above must be consistent
  - Other extended states and flows are possible but not covered in context of CXL

How are Peer Caches Managed

- All peer caches managed by the "Home Agent" within the cache level
- A "Snoop" is the term for the Home to check cache state and causing cache state changes
- Example CXL Snoops:
  - Snoop Invalidate (SnpInv): Cache to degrade to I-state, and must return any Modified data
  - Snoop Data (SnpData): Cache to degrade to S-state, and must return any Modified data.
  - Snoop Current (SnpCurr): Cache state does not change, but must return any Modified data

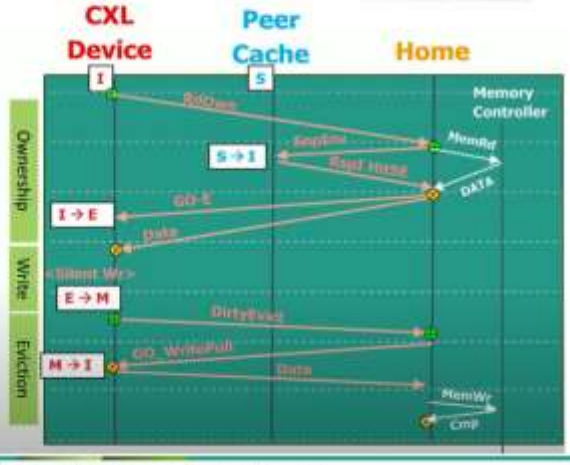
CXL.cache Read Flow example

- Diagram to show message flows in time
  - X-axis: Agents
  - Y-axis: Time

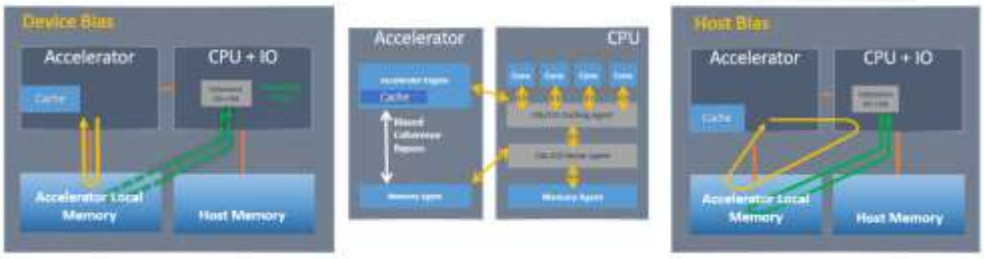


CXL.cache Write Flow example

- For Cache Writes there are three phases:
  - Ownership
  - Silent Write
  - **Cache Eviction**



CXL 2.0 Coherence Bias



Critical access class for accelerators is "device engine to device memory"

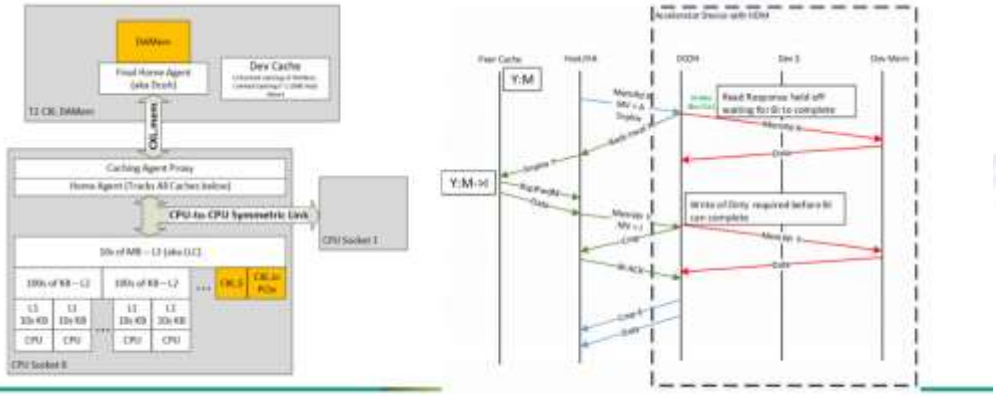
"Coherence Bias" allows a device engine to access its memory coherently without visiting the processor

Two driver managed modes or "Biases"

Both biases guaranteed correct/coherent

CXL 3.0 Protocol Enhancements: Mapping Large memory in Type-2 Devices to HDM with Back Invalidate

Existing Bias: Flip mechanism needed HDM to be tracked fully since device could not back snoop the host.
Back Invalidate with CXL 3.0: enables snoop filter implementation resulting in large memory that can be mapped to HDM



How does Conherence help

- Break out of IO ordering model:
  - Prefetching data before reading flag → saves latency
  - Streaming WRITEs to memory get a completion → PGAS NIC use model
  - WRITE data committed within local cache
- Data Reuse (cache locality) reduces latency and bandwidth demand
  - Descriptor or Semaphore Polling
- Complex Atomics → Needed by accelerators and NICs
- CXL.mem can expose Device memory as coherent → No DMA required
  - Host or peer can pull current data from Device when it is ready to consume → reduce data copies
  - Enables host/peer to prefetch data
- Not Risk Free!
  - Badly behaving Device or use model can impact system performance
  - Conflicts to single cache-line may cause serializing events → Spinlock across many devices/threads
  - Device should adopt best known practices for CPU coherence flow: TicketLock, LRU cache, etc.

15.3 R&S 罗德与施瓦茨公司 VNA 测试 PCIe Gen5 延长线缆信号质量

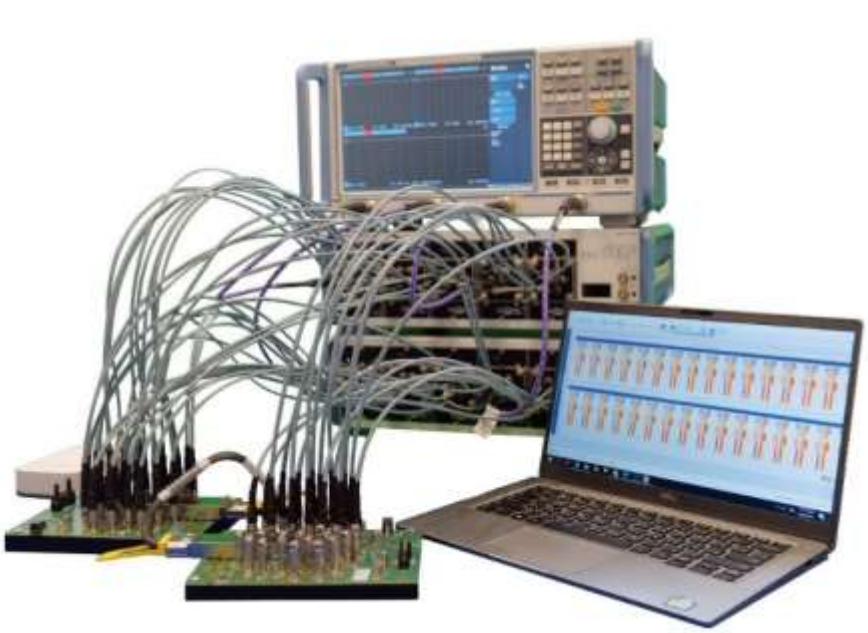
前面 14.1 章节 PPT 最后几页提到 Gen5 x16 延长线的问题，目前市场上的 PCIe Gen5 延长线缆或者延长转接线缆质量问题堪忧，大部分的线缆厂商确认有效的测试工具量测延长后的信号质量，包括插损，回损，以及串扰等参数。R&S 罗德与施瓦茨公司的 VNA 产品配合其自动化信号切换设备非常好地解决了测试 PCIe Gen5 延长线缆信号质量的问题。下面是相关的简介，需要的朋友可以联系我们首页的二维码获得样机或者预约进一步的技术交流。

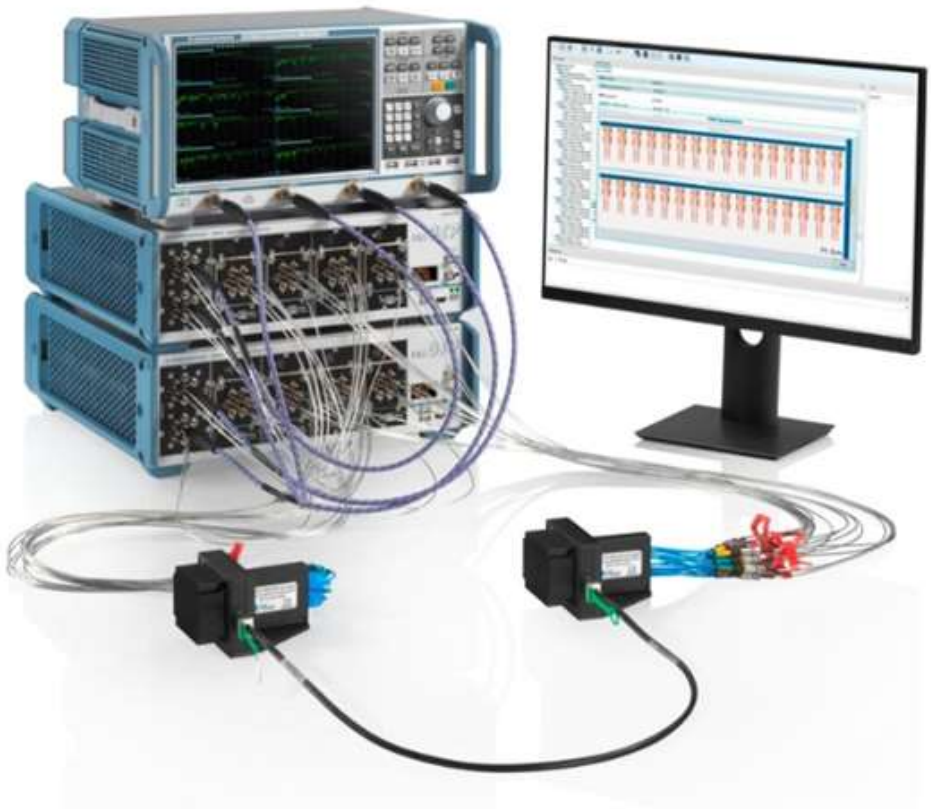
下面是一个典型的测试工具（VNA + OSP 自动化信号切换设备）搭配。

	Lane	VNA	OSP320	SWITCH	count
PCIE 5.0	x4	ZNB26	2	B122E 26.5GHz SP6T (terminal)	x8
	x8	ZNBT26	3		x12
	x16		6		x24 +B121Ex2
PCIE 6.0	x4	ZNB43	2	B122H 40GHz SP6T (terminal)	x8
	x8	ZNBT43	3		x12
	x16		6		x24 +B121Hx2

下面是可以测试各种常见的 PCIe Gen5 线缆类型。

下面两张图是典型的连接了待测 PCIe 线缆的示意图，电脑上安装了测试软件。





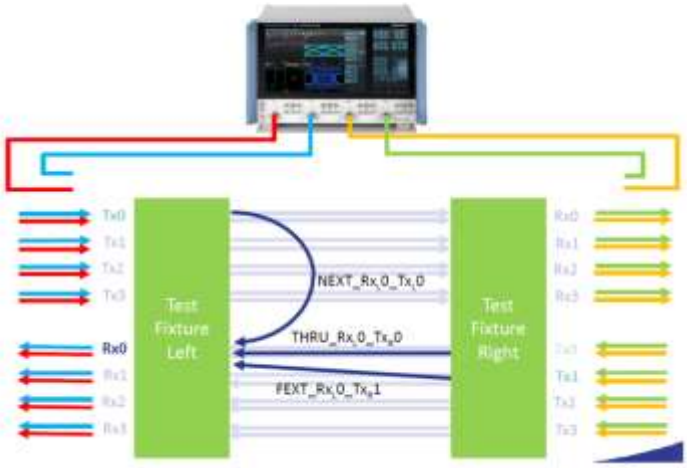
VNA 测试 PCIe Gen5 线缆的关键指标

VERIFICATION OF PCIE 5.0 / 6.0 CABLES AND CONNECTORS: GENERAL CONSIDERATIONS

Measurements:
- insertion loss: Sdd21
- return loss: Sdd11 and Sdd22
- near-end crosstalk NEXT
- far-end crosstalk FEXT

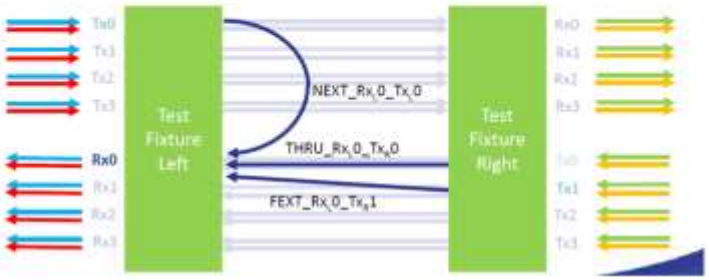
with 4 port VNA;
multiple 4-port measurements

Postprocessing:
- integrated return loss iRL
- power sum MDNEXT and ccICNNEXT
- power sum MDFEXT and ccICNFEXT
- intra-pair skew: EIPS
- inter-pair skew (lane-to-lane)



Measurements:
automation with switch matrix:
example for PCIe x8

	PCIe x4	PCIe x8	PCIe x16
number of lanes (Tx + Rx)	8	16	32
number of ports for full testing (all lanes and all crosstalk combinations)	32	64	128
number of 4-port measurements for full testing (all lanes and all crosstalk combinations)	8 x THRU 4 x 4 = 16 x NEXT_L 4 x 4 = 16 x NEXT_R 3 x 4 = 12 x FEXT_L 3 x 4 = 12 x FEXT_R total: 64 4-port meas.	16 x THRU 8 x 8 = 64 x NEXT_L 8 x 8 = 64 x NEXT_R 7 x 8 = 56 x FEXT_L 7 x 8 = 56 x FEXT_R total: 256 4-port meas.	32 x THRU 16 x 16 = 256 x NEXT_L 16 x 16 = 256 x NEXT_R 15 x 16 = 240 x FEXT_L 15 x 16 = 240 x FEXT_R total: 1024 4-port meas.



下面是采用该 VNA 的典型的测试报告输出。

ROHDE&SCHWARZ
Make ideas real

ZNrun Cable Test

Measurement Results (S-Parameters)	Evaluation Results	Overall Result
FAIL		

Supported Communication Standard

Specification	Link Speed [Gb/s]	Symbol Rate [Gbaud/s]	Medium Type	TX Lanes
PCI-SIG PCIe Gen 5 Cable	16		Cable	4

ccICNNEXT [uV]

Lane pair	ccICNNEXT_L [uV]	ccICNNEXT_R [uV]
1.00	192.92	348.72
2.00	330.52	259.34
3.00	284.97	235.40
4.00	154.06	261.27
5.00	347.62	374.89
6.00	266.07	365.32
7.00	243.08	206.31
8.00	348.14	117.62
9.00	314.86	351.95
10.00	180.11	320.65
11.00	286.66	278.25
12.00	389.57	303.92
13.00	98.21	99.96
14.00	130.65	387.66
15.00	223.97	206.43
16.00	378.76	351.18

ccICNFEXT [uV]

Lane pair	ccICNFEXT_L [uV]	ccICNFEXT_R [uV]
1.00	259.06	476.75
2.00	165.96	412.12
3.00	122.71	85.34
4.00	376.46	322.79
5.00	291.25	164.30
6.00	287.81	234.34
7.00	194.48	200.74
8.00	214.02	510.94
9.00	272.03	295.23
10.00	331.29	311.17
11.00	238.38	174.21
12.00	297.59	295.97

iRL [dB]

Lane	iRL_L [dB]	iRL_R [dB]
1.00	34.67	33.06
2.00	35.93	35.40
3.00	33.02	33.54
4.00	35.49	34.28

EIPS [ps]

Lane	EIPS_L [ps]	EIPS_R [ps]
1.00	0.08	-0.01
2.00	0.06	-0.57
3.00	0.15	-1.84
4.00	-0.52	-0.41

Chart 1: Insertion Loss

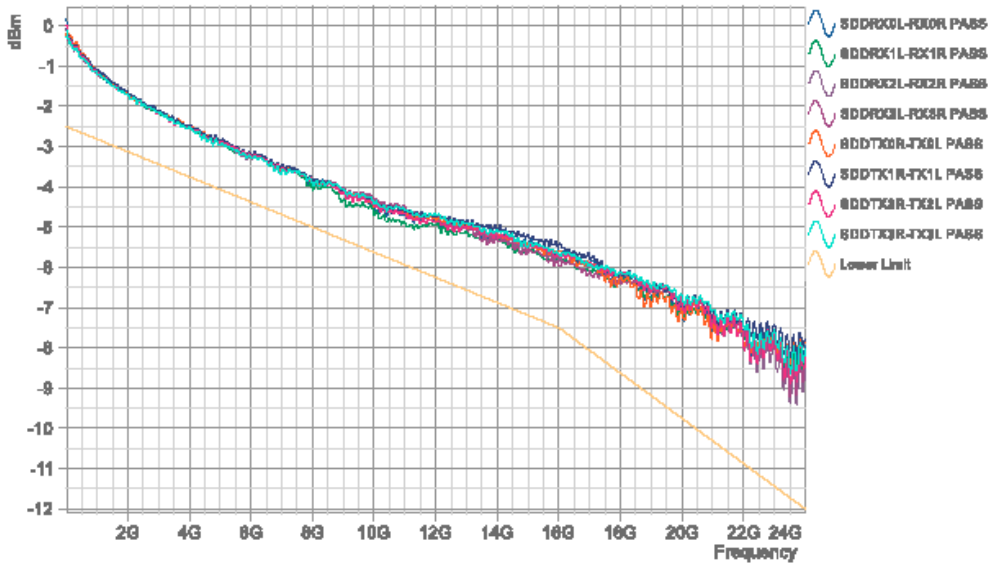


Chart 2: Return Loss (TX)

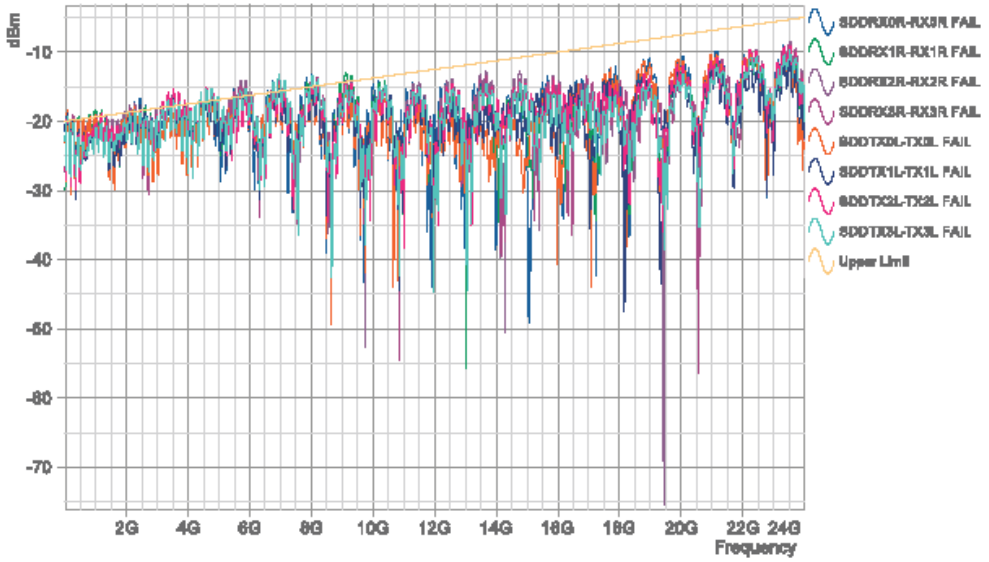


Chart 3: Return Loss (RX)

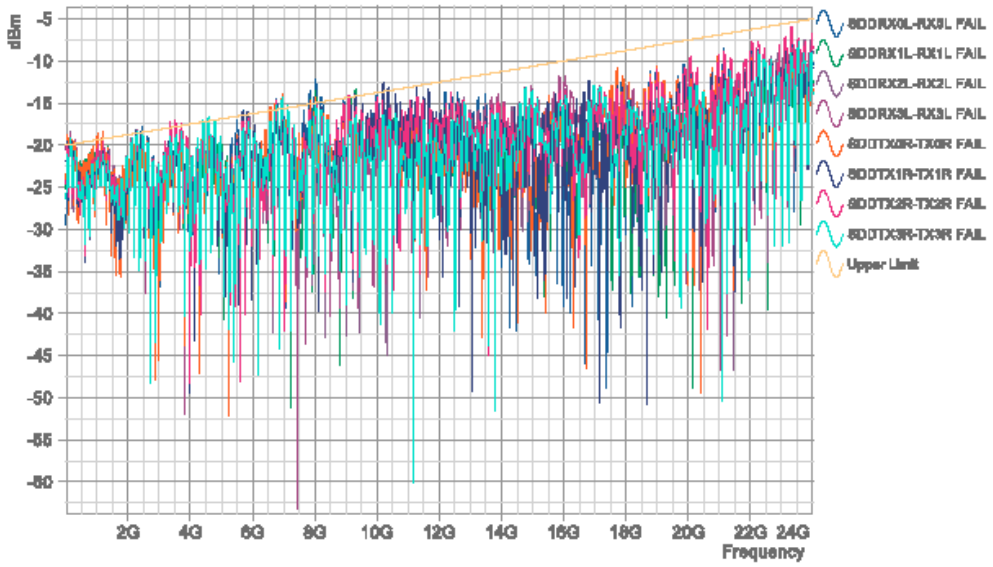


Chart 4: Powersum (NEXT)

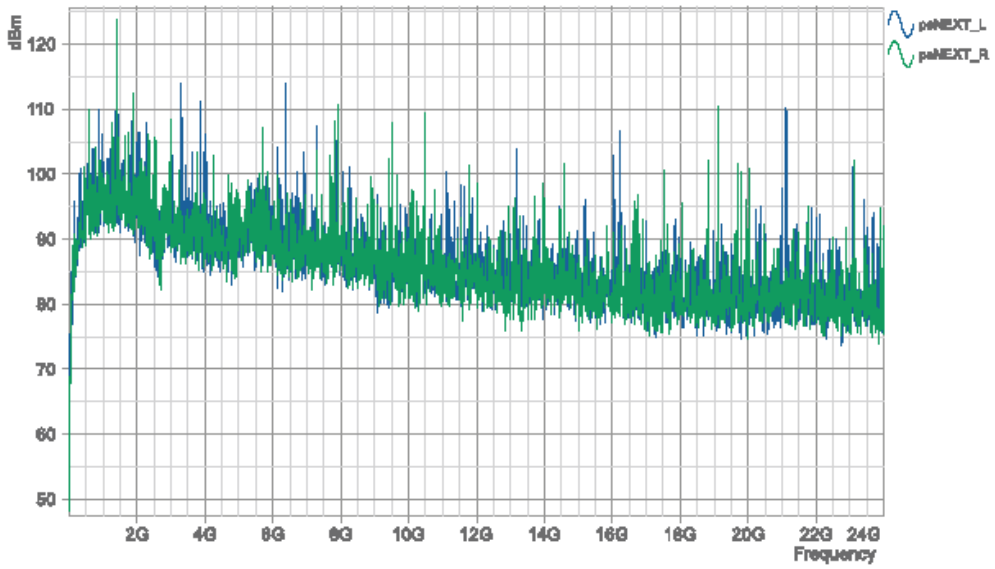
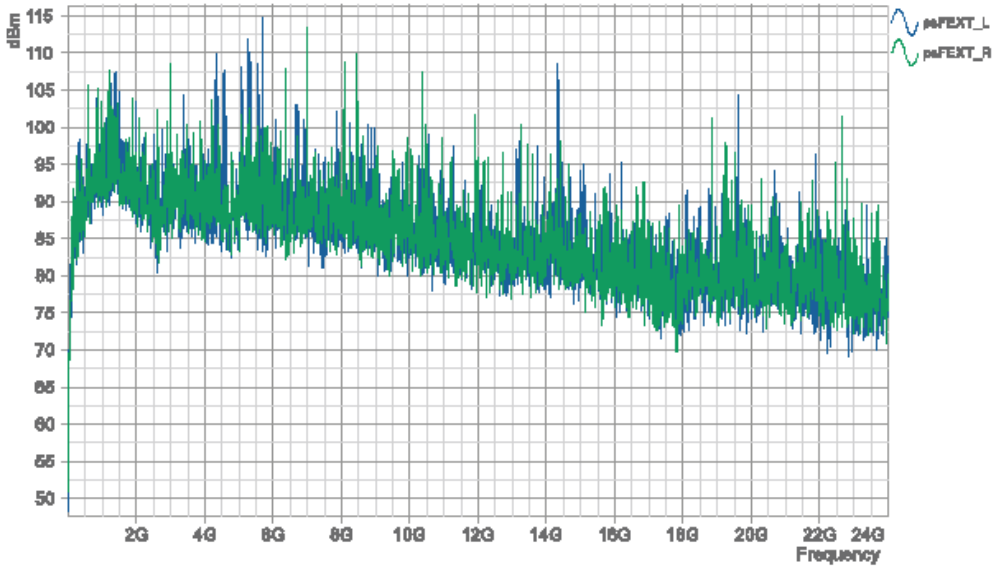


Chart 5: Powersum (FEXT)



16. 附录 G：针对 Gen5 M.2 SSD 和超薄笔记本散热的新方案

16.1 How It Works - How do you cool ultra-thin devices?

Published: SEPTEMBER 30, 2023

Devices are getting increasingly compact. Every fraction of a millimeter of internal device space is carefully designed by manufacturers to create the smallest and most powerful devices possible. Designs aimed at satisfying consumers' demand for increased performance in ever smaller devices.

Great examples of this are smart phones and ultra-thin notebooks as thin as 9.5mm, with only 3.5mm of internal space available for placement of critical system components. Performance expectations of these devices however are continually increasing with more powerful CPU's and GPU's and growing demands from 5G and even faster future networks. But this increased performance generates heat… a lot of heat.

What is throttling?

Throttling is a protective mechanism aimed at preventing overheating and processor damage. It lowers the performance of your devices well below its real capability -

protecting your device by leaving the user frustrated with low performance and slow operation.

To ensure peak sustained performance designers must match up a thermal solution that removes all the heat. If not, processors will slow down (throttle), thus delivering only a fraction of the performance.

In ultra-thin devices, designing a fully capable thermal solution has been a challenge because they are entirely reliant on passive thermal solutions like heat spreaders and vapor chambers. Active cooling using fans to enable higher performance has not been an option. Fans cannot provide active cooling in ultra-thin devices because as fans get thinner, they get less capable, with low air flow and back pressure (or suction power), which is unable to overcome the system resistance inside ultra-thin devices. This effectively reduces the air flow even lower and the amount of additional heat dissipation they remove to near zero.

What is System Resistance?

System resistance is the amount of friction that air encounters as it flows through a confined space such as inside ultra-thin electronic devices. The higher the system resistance, the higher the backpressure (also known as static pressure) is needed to create enough suction force to overcome the resistance and pull air through compact enclosed spaces.

This is why you'll only find fans in thicker Notebooks with air inlet vents in the back cover; the vents need to be located as near as possible to the fan to reduce system resistance. This inlet vent location, necessitated by the lack of adequate backpressure, is a poor design choice because as soon as the Notebook, intended to be portable and versatile, is placed on your lap (or a cushion as the devices gets uncomfortably warm), the vents are blocked and the fan, with its low back pressure, cannot pull air into the device. This blockage of airflow renders the fans almost useless and increases throttling.

Given the system resistance challenges for ultra-thin devices with only 3.5 mm of internal air gap, actively cooling them to increase performance has been considered impossible… until now.

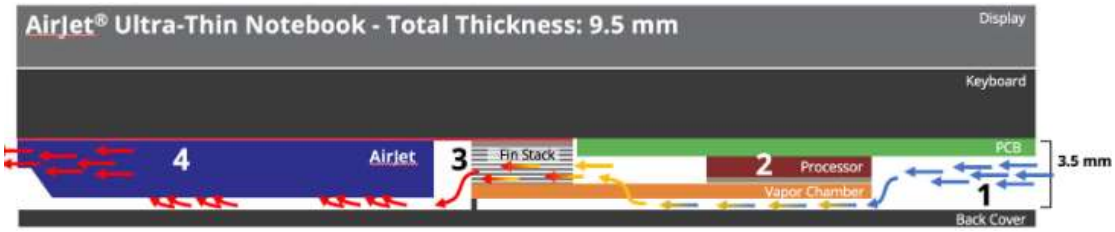
Airjet's compact size, only 27.5mm x 41.5mm and, importantly, only 2.8mm in vertical height, is slim enough to fit inside these ultra-thin devices. AirJet's massive back pressure of 1750 pascals (more than 10 times the suction force of a high-end notebook fan) enables air to be pulled into the device, overcoming the system resistance inherent to ultra-thin devices. Moreover, the inlet vents can be located on the sides and not in the back cover of the Notebook.

AirJet's massive back pressure also opens new possibilities for active cooling design in these ultra-thin devices. Until now, the low back pressure or suction force available from fans has limited the ability of manufacturers to maximize surface area on fin stacks for heat removal, as increasing the number of 'fins' in a more tightly configured form increases system resistance to unacceptable levels. Consider Fin Stacks A and B below.

Fin Stack A has double the surface area of Fin Stack B, doubling the potential heat removal. So why can't manufacturers use the more tightly configured fin stack approach in Fin Stack A for traditional active cooling? System Resistance! While Fin Stack A doubles the surface area available for heat transfer, it also significantly increases the system resistance. In fact, decreasing the space between the fins by half, increases the system resistance a massive 700%. So the more tightly fins are packed in the available space, the more backpressure, or suction it takes to pull air through the fin stack. Fin Stack A requires 700% more back pressure to move air though the device and over the surfaces to enable heat to be transferred. Fans simply don't have the back pressure to overcome this level of system resistance.

This is where AirJet enables a completely new approach to heat removal in these ultra-thin devices. AirJet's 1750 pascals of backpressure is more than enough to pull air into ultra-thin devices and through unique tightly configured fin stacks enabling active cooling and meaningful heat removal in ultra-thin compact devices for the first time.

The illustration below explains how it all works.



1. AirJet's massive backpressure enables ambient air to be pulled into the most compact device: through a discrete air inlet on the side of the device, and even through dustproof filters.

2. Heat is efficiently moved from the heat source (CPU, GPU, modem, or other source) to a fin stack using a 0.7mm thick vapor chamber.

3. The Fin Stack, specifically designed for use with AirJet, features a much higher number of tightly configured fins. This innovative fin stack design has never been possible in ultra-thin devices. Traditional fans, with their low suction power, are incapable of overcoming the increased system resistance created by the tightly packed fins. AirJet's 1750 pascals of back pressure easily pulls air though the design, significantly increasing the exposed surface area for heat transfer.

4. The heat saturated air is pulled through the fin stack, by the AirJet and is then expelled completely out of the device through discrete vents on the side of the device.

This innovative approach to active cooling using AirJet, the world's first solid-state active cooling chip, has opened new opportunities for active cooling in ultra-thin devices delivering up to 100% higher performance in these form factors. To find out more about how AirJet can enable your device to do more, contact Frore Systems.

16.2 OWC 使用 mini 冷却器开发 32TB 和 64TB SSD 设备

经过安东·希洛夫 (Anton Shilov) 于 2023 年 8 月 15 日上午 10:00 美国东部时间

OWC 和 Frore Systems 在 2023 年闪存峰会上展示了使用 Frore AirJet Mini 冷却器的静音 32 TB 和 64 TB 固态存储设备。 这两款设备都承诺始终如一的高性能，同时保持完全安静。此外，使用 Frore 的 AirJet 冷却系统为不使用风扇的高容量 SSD 存储解决方案打开了大门。

OWC 在展会上展示了其 内置 8 TB M.2-2280 SSD 的 64 TB Mercury Pro U.2 Dual 以及内置 4 个 8 TB M.2 驱动器的 32 TB U2 Shuttle 。该外设制造商表示，凭借 AirJet Mini 技术，其 64 TB OWC Mercury Pro U.2 Dual 实现了 2200 MB/s 至 2600 MB/s 的一致顺序写入速度。然而，它没有公开当仅使用内部风扇时这种配置有多快。OWC 在其网站上表示，当配备多个驱动器时，该机箱可以使 Thunderbolt 3 互连达到 2800 MB/s 的读/写速度。

OWC 的 Mercury Pro U.2 Dual 和 U2 Shuttle 本质上是 PCIe 3.0 SSD 载体，此类驱动器不会变得很热，因此将 Frore 的优质 AirJet 冷却技术应用于它们听起来有点大材小用。但这里有两件事需要注意。首先，OWC 的 Mercury Pro U.2 Dual 配备了 3,000 rpm 的风扇，并且该设备开箱后并不是完全安静，因此使用 AirJets 可以移除风扇并使其完全安静。其次，通过将 Frore 的 AirJet 与现有运营商结合使用，OWC 为后续几代 Mercury Pro U.2 Dual 和 U2 Shuttle 奠定了基础，这些产品将容纳更快、更热的 SSD，需要适当的冷却以确保一致的性能。

Frore 的薄膜 AirJet Mini 尺寸为 41.5 毫米 × 27.5 毫米 × 2.8 毫米，重 11 克；它的设计散热量为 5W，集成多个芯片可成比例地放大散热能力。现代 PCIe Gen5 驱动器在高负载下的功耗可能远远超过 5W，因此它们需要多个 AirJet Mini。

OWC 首席执行官拉里·奥康纳 (Larry O'Conner) 表示：“*看到 Frore Airjet 系统所提供的解决方案的应用和优势，我们感到非常兴奋。*”“*我期待与 Frore 合作，让我们的解决方案走得更远。这项技术使我们能够通过多种方式提高产能、长期升级设计、改善客户体验和应用适用性，拥有无限的机会。*”

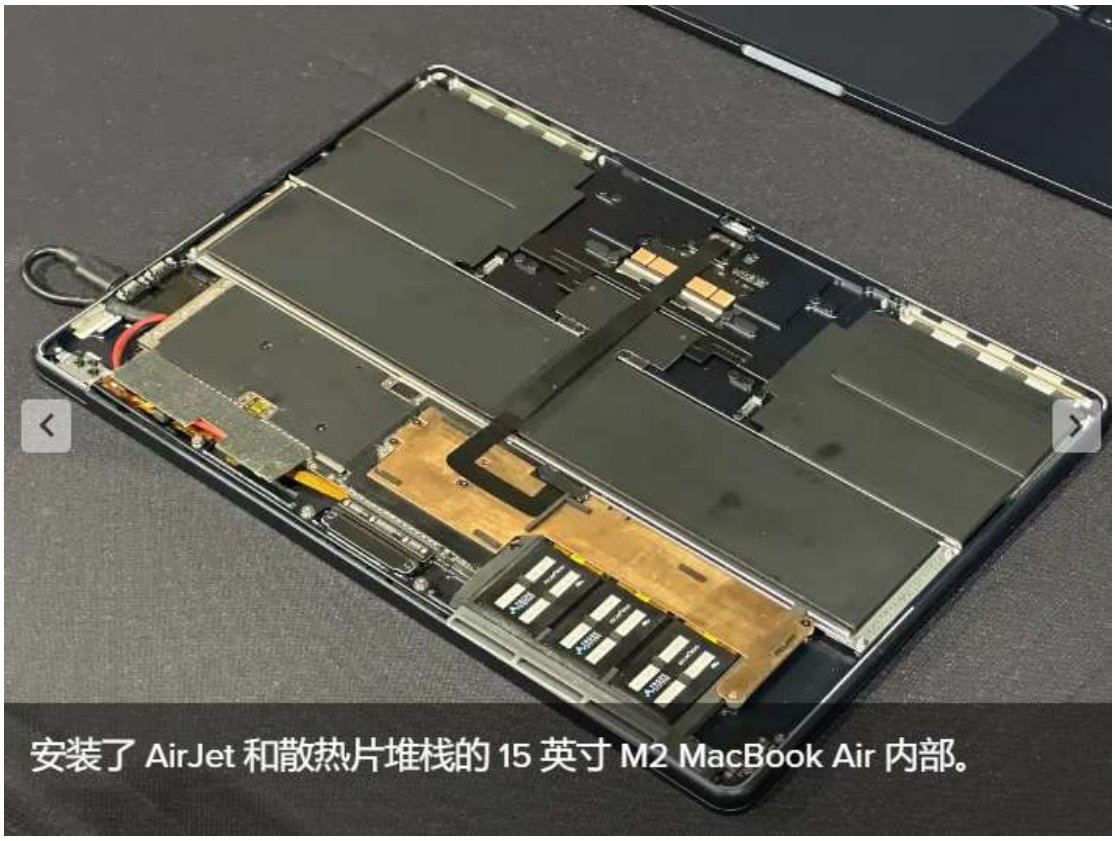
资料来源： 弗洛尔系统公司

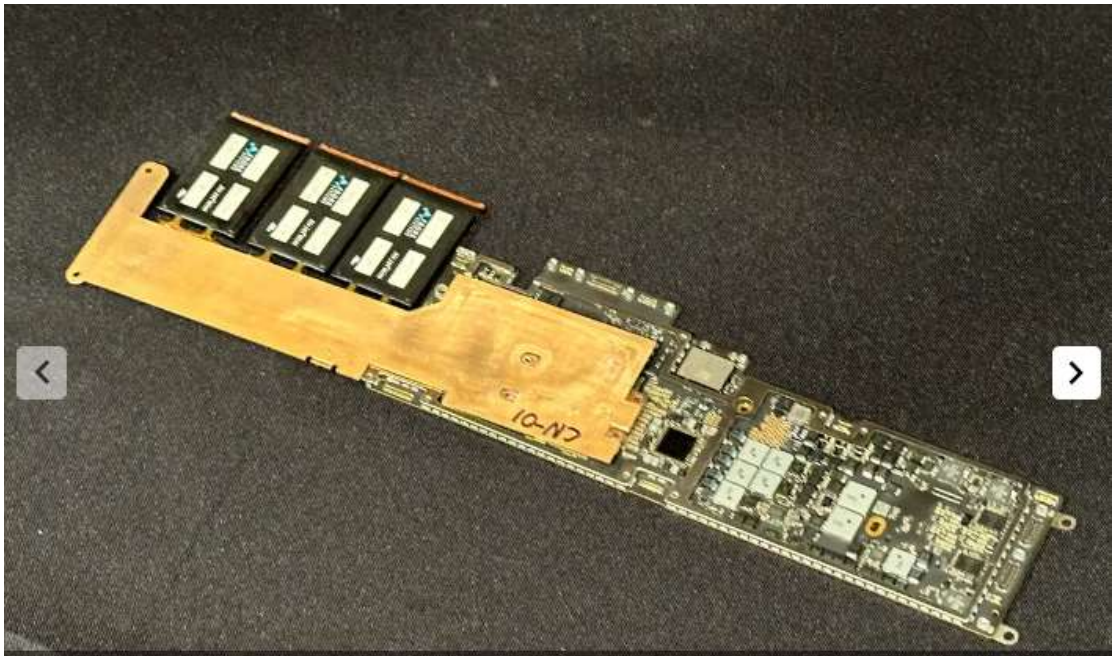
16.3 超薄 mini 冷却器风扇将 MacBook Air 变成 MacBook Pro

超薄冷却芯片适合苹果最薄的笔记本电脑。

罗曼·洛约拉

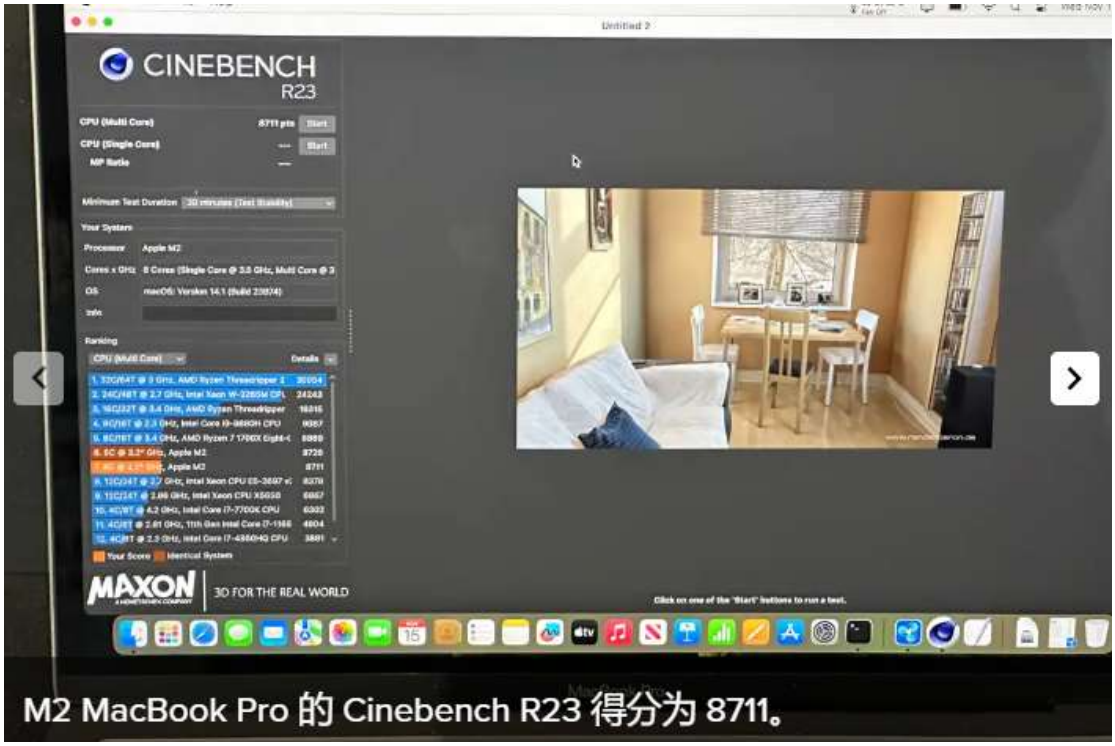
高级编辑， 麦克世界 2023 年 11 月 28 日上午 10:57（太平洋标准时间）

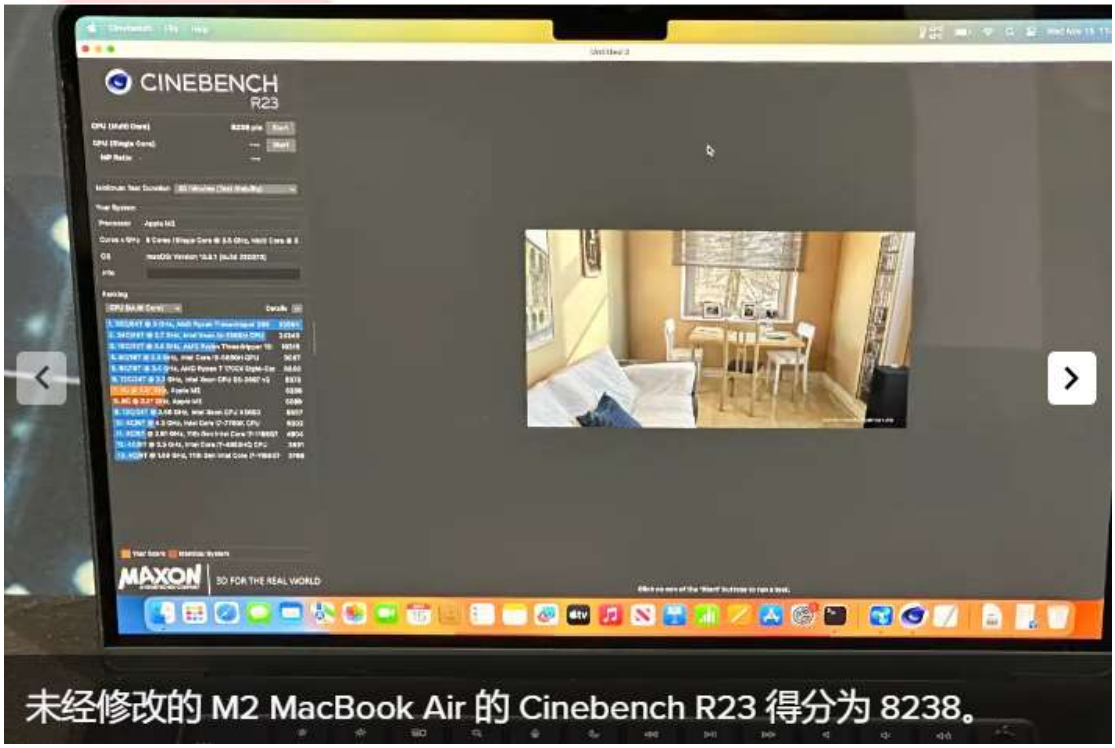




该修改使用翅片组将热量从 M2 吸走并转移到 AirJet 冷却系统。

为了安装 AirJet，Frore 改装了一台带有进气口的 MacBook Air。


CINEBENCH
R23
MAXON
3D FOR THE REAL WORLD
M2 MacBook Pro 的 Cinebench R23 得分为 8711。


CINEBENCH
R23
MAXON
3D FOR THE REAL WORLD
未经修改的 M2 MacBook Air 的 Cinebench R23 得分为 8238。

大多数电脑都使用风扇来帮助其芯片保持最佳工作温度。这些风扇笨重且噪音大，这就是超薄 MacBook Air 没有风扇的原因。但 Frore Systems 有一个解决方案 - 其名为 AirJet 的新型冷却系统超薄 - 足够薄，可以安装在当前的 MacBook Air 中，并在重负载下提高其性能。

由于 MacBook Air 没有适用于 SoC 的风扇，因此在处理器密集型工作期间其性能会降低，以保持适当的工作温度。另一方面，M2 13 英寸 MacBook Pro 有一个风扇，可以排出多余的热量，这样芯片就可以继续运转。

AirJet 被 Frore 称为“固态主动冷却芯片”，尺寸为 27.5 x 41.5 x 2.8 毫米，比典型的电脑风扇小得多、薄得多。它非常薄，Frore 能够在 M2 15 英寸 MacBook Air 上安装一组 AirJet 芯片。AirJet 可以将 MacBook Air 的温度保持在适当的水平，这样芯片就不必降速。使用 Cinebench R23 基准测试，现成的 M2 MacBook Air 比 M2 MacBook Pro 慢 7%。但配备 AirJet 设置的改良版 M2 MacBook Air 与 MacBook Pro 的 Cinebench 得分相当。下面我们的姊妹网站 PCWorld 的视频解释了它的工作原理。

AirJet 为何能够如此薄？该公司有一篇博客文章更详细地介绍了 AirJet 的工作原理，但本质上，其内部是用于振动膜的几种不同材料。这会产生背压，通过设备（例如笔记本电脑）外壳上的进气口吸入空气。当热量从芯片转移到散热片堆叠时，AirJet 使用空气将

热量从通风口排出。在 MacBook Air 演示中，Frore 在铰链附近添加了进气口，并使用扬声器孔作为排气口。

AirJet 不是消费品——消费者无法购买它并改装自己的 MacBook Air，Frore 也不打算创建用于消费者销售的套件。Frore 的 MacBook Air 演示证明了它在 MacBook 中的优势。其纤薄的外形可能为苹果产品带来可能性；例如，如果苹果公司使用 AirJet 代替 14 英寸和 16 英寸 MacBook Pro 中使用的风扇，则可以释放公司可用于安装更大电池的空间。另一个例子是 Mac Studio，正如你在 iFixit 的 Mac Studio 拆解中看到的那样，它有一个巨大的散热器和两个大风扇——AirJet 可以让苹果创建一个更小的冷却系统，从而缩小 Mac Studio 的尺寸。

有关 AirJet 的更多信息，请查看下面的幻灯片以及 PCWorld 更多报道的链接：

- AirJet 拆解：了解固态冷却革命的内部
- 实验室参观！深入了解 AirJet 的未来固态笔记本电脑冷却技术

作者：Roman Loyola，高级编辑

Roman 自 20 世纪 90 年代初就开始报道技术。他的职业生涯始于 MacUser，并曾在 MacAddict、Mac|Life 和 TechTV 工作。

16.4 Mini cooling device teardown: See inside the solid-state cooling revolution

We can finally see (at least some of) the guts of the AirJet sold-state cooling device, and watch it in action cooling an SSD

By Michael Crider

Staff Writer, PCWorld MAY 31, 2023 8:24 AM PDT

Image: Adam Patrick Murray/Foundry

One of the most exciting new technologies we've seen this year is Frore Systems' AirJet, a device that can cool PC components and other electronics using a super-thin, solid-state build with no moving parts. The initial device was shown off at CES at the beginning of the year, but at Computex in Taiwan Gordon and Adam got a chance to peek under the hood and see what makes the AirJet tick. Or not tick, I guess, but vibrate membranously on a tiny bit of power. You know what I mean.

If you haven't been following our coverage, the Airject uses several layers of exotic materials and precision geometry to vibrate tiny membranes, intaking air, blowing it over a copper heat spreader, and exhausting it at up to 200 kilometers per hour. A tiny AirJet Mini unit, about the size of a few postage stamps, can replace an active cooler ten times its size and many times its 2.8mm thickness.

Gordon's teardown demonstration lets you see all the pieces that go into an AirJet unit, showing just how tiny it is — and most of it is the copper heat spreader surrounding the microjets sealed under the casing. Millimeter-thin portions of the material design allow the expelled air to mix with cooler air, preventing an uncomfortable blast of heat from annoying the user.

What does the AirJet look like when used in a real product? AirJet maker Frore Systems has a demo to show how it can work on an M.2 SSD, which is where a lot of the extreme cooling needs are showing up these days. An external drive equipped with two AirJet Minis, compared to the same Sabrent retail design with passive cooling, shows a huge improvement in cooling while being used. The AirJet-equipped drive was 55

degrees Celsius (131 Fahrenheit) under load, with slightly better read and write performance, versus a toasty 68 degrees C (154 F) on the passively cooled drive.

The first retail product scheduled to ship with AirJet cooling is the Zotac Zbox PI430AJ (the AJ stands for AirJet!), which PCWorld also checked out at a Computex. We've visited Frore Systems for an AirJet lab tour, too. For the latest news on the world of PC components, be sure to subscribe to PCWorld on YouTube.

Author: Michael Crider, Staff Writer

Michael is a former graphic designer who's been building and tweaking desktop computers for longer than he cares to admit. His interests include folk music, football, science fiction, and salsa verde, in no particular order.

16.5 Lab tour! Go inside this mini cooling devi ce futuristic solid-state laptop cooling tech

Frore Systems' AirJet tech might just change the way we cool all kinds of gadgets.

By Michael Crider
Staff Writer, PCWorld FEB 24, 2023 11:52 AM PST

Image: Willis Lai/IDG

One of the most impressive products to come out of CES 2023 was AirJet, new solid-

state cooling system that can replace fans with cooling components using a fraction of the power and space. Seshu Madhavapeddy, founder and CEO of Frore Systems, was kind enough to give us a tour of the company's offices and development labs in San Jose.

A quick refresher on AirJet: It's a new cooling solution that uses a tiny, solid-state block with no separate moving parts to remove heat from components like CPUs and GPUs. The AirJet uses tiny vibrating membranes above a copper head spreader to draw in and expel air at up to 120 miles per hour, creating incredibly efficient air cooling in a space just 2.8mm thick. On a watt-per-watt basis, Frore claims that AirJet could more than double the cooling power in conventional fan-based systems, while working in an even smaller space.

The vibrating material is waterproof and dustproof, and AirJet already has functioning dust filters for the individual cooling units, down to one micron. Testing in windy, dusty chambers has shown that even in the most extreme conditions the AirJet's performance isn't adversely affected. Similar testing for heat, cold, humidity, and longevity shows the product is ready for the long haul.

Madhavapeddy freely admits that the AirJet is more costly than comparable fan-based solutions for laptops. "Is it priced [in a way] that would be reasonable and acceptable by manufacturers? The answer is an emphatic 'yes.'"

The AirJet isn't currently installed in any consumer products, but Frore has modified several current models to replace their internal cooling systems with the AirJet, just to demonstrate its efficacy. In a base model with no special engineering for AirJet, three of the solid-state coolers were able to replace the primary CPU heatsink and fan setup, boosting its core wattage from 12.5w to 15w with no ill effects, and an overall reduction in noise. Madhavapeddy claims that under more ideal design circumstances, it's possible to use four AirJet Mini units and boost it up to 21 watts.

Frore is working with "several device manufacturers," and "several commercial projects" are slated to hit shelves before the end of the year. At the moment the company is focusing on conventional notebook designs. Handheld gaming devices, mini PCs, M.2 storage drives, and digital cameras are also potential vectors for expansion.

Author: Michael Crider, Staff Writer

Michael is a former graphic designer who's been building and tweaking desktop computers for longer than he cares to admit. His interests include folk music, football, science fiction, and salsa verde, in no particular order.