# CXL、GenZ、CCIX架构以及未来的PM、内存和SSD形态

原创 唐僧 huangliang 企业存储技术 2021-05-17 01:15

收录于话题
#服务器 65 #Optane（傲腾）、非易失内存 29 #会议/技术资料分享、解读 32

目录

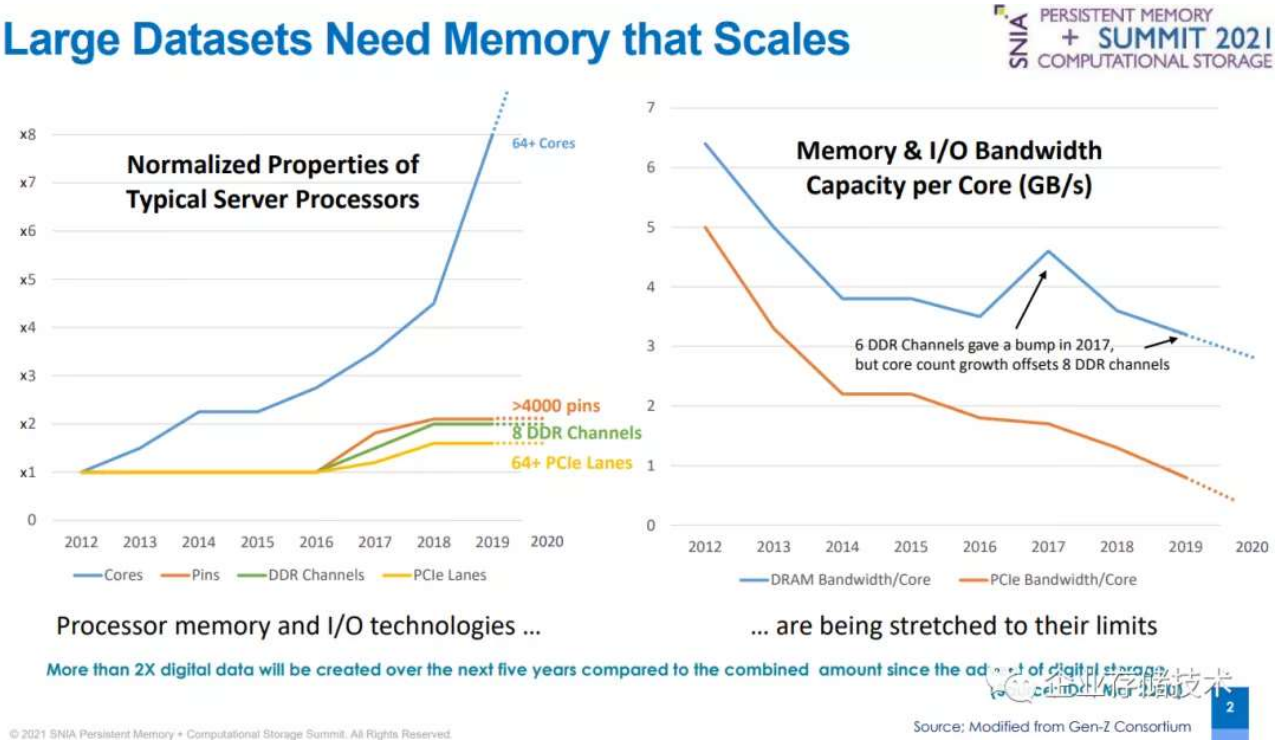


上周的一个技术会议，我听到有位演讲人的相当一部分ppt内容，与我在《*数据中心NVMe SSD和EDSFF前瞻：来自Intel、HPE、Dell & SNIA等*》中引用的图片恰好来自同一份资料。而在周末聚会时，另一位前同事也认为该文章挺有帮助。这里不敢说英雄所见略同，毕竟那两位同仁都是在专业的SSD厂商，只能说明我对技术资料的价值还有点判断力吧。偶尔熬夜写东西分享出来，对大家有帮助就没白忙活：）

今天要讨论的主题，参考资料主要来自SNIA技术会议2021 Persistent Memory +Computational Storage Summit（文末会列出下载链接）的一场分享《*Futureof Persistent Memory, DRAM and SSD Form Factors Aligned with New SystemArchitectures*》，演讲者是SMARTModular Technologies的产品市场总监ArthurSainio——就是那家主要做“非常规”内存和SSD模组的厂商。

还是继续展望持久内存（PM）、内存和SSD形态的方向，不过前面一篇主要围绕物理尺寸，而这次则是重点针对架构和接口协议。

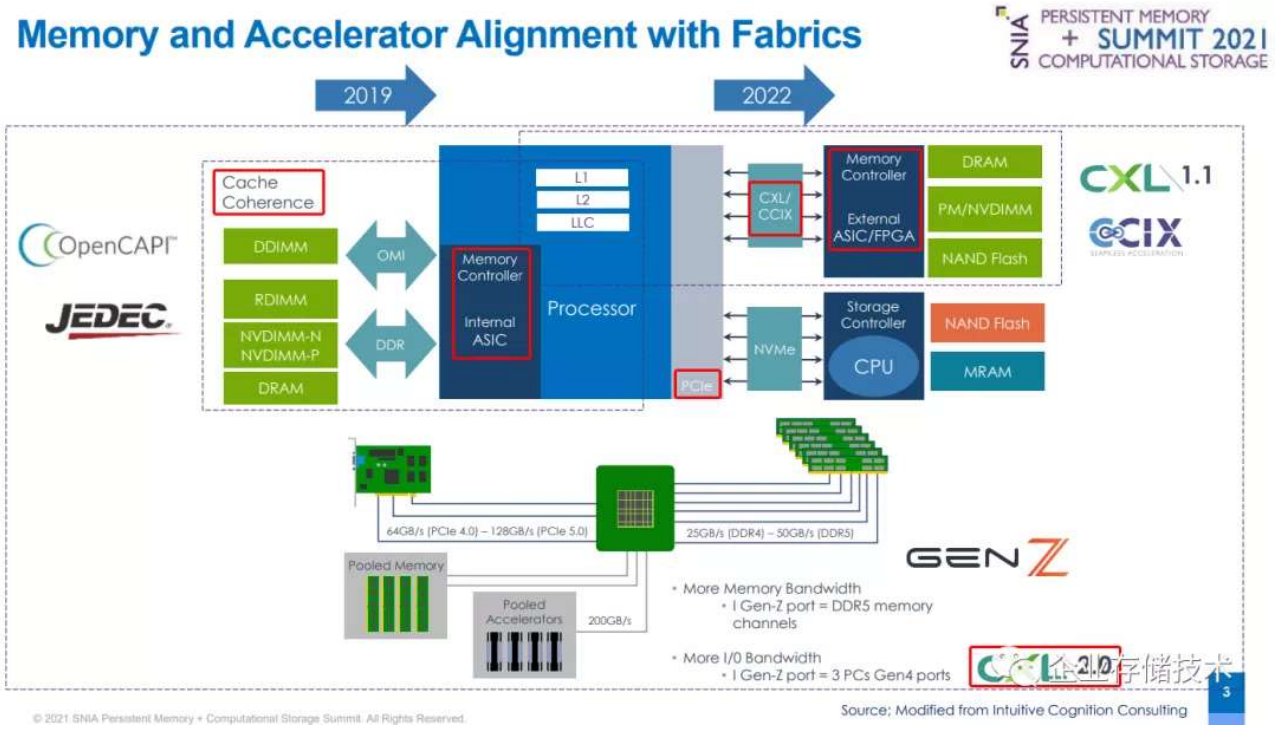
## 为什么要扩展内存/IO带宽：跟不上CPU计算核心发展




*点开图片可放大显示，以下同*

如上图，左半边显示从2012-2019年，服务器CPU核心（从8核到64核）增长了8倍，而Pin针脚数量（从LGA-2011到4094/4189）和内存通道数（从4到8）仅增长了2倍，PCIe lane（从40到64，如果是AMD则按照双路中每CPU支持来算）甚至不到2倍。可以看出I/O跟不上计算密度的增长。

右边是通过计算得出的平均每核心DRAM内存带宽，以及每核心PCIe带宽，都是呈不断下降趋势。



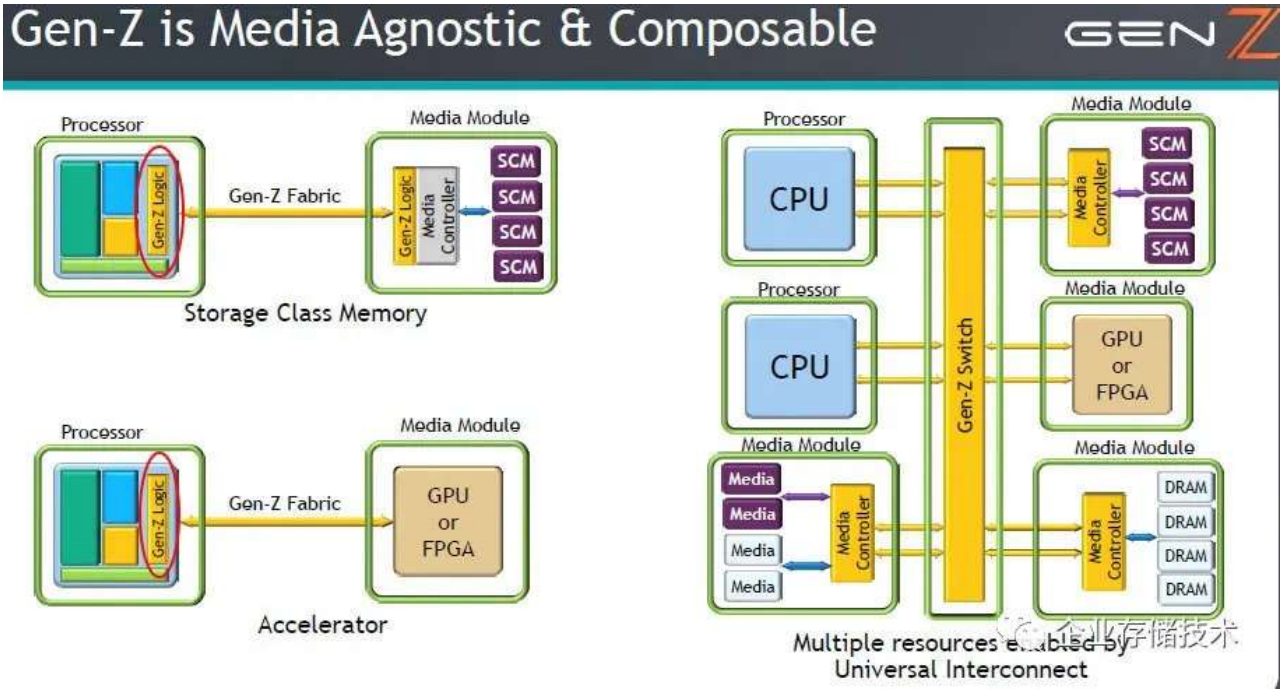
上面这张图信息量不小。在CPU左侧的“内存控制器 / 内部ASIC”，除了可以提供JEDEC标准化DDR内存接口之外；还有IBM 2019年之前就在Power平台上使用的OpenCAPI——具备缓存一致性（Cache Coherence）支持的OMI（开放内存接口）。

CPU右侧的物理连接是PCIe。具体到2022年PCIe 5.0服务器发布的时候，除了用于NAND/SCM SSD的传统NVMe协议之外，CXL和CCIX预计也会出现在x86平台。通过这2种协议连接的“内存控制器 / 外部ASIC”，同样支持缓存一致性。

下半部分，则是通过外部Bridge/Switch芯片连接的池化内存 / 加速器，这一块则是Gen-Z和CXL 2.0的天下。

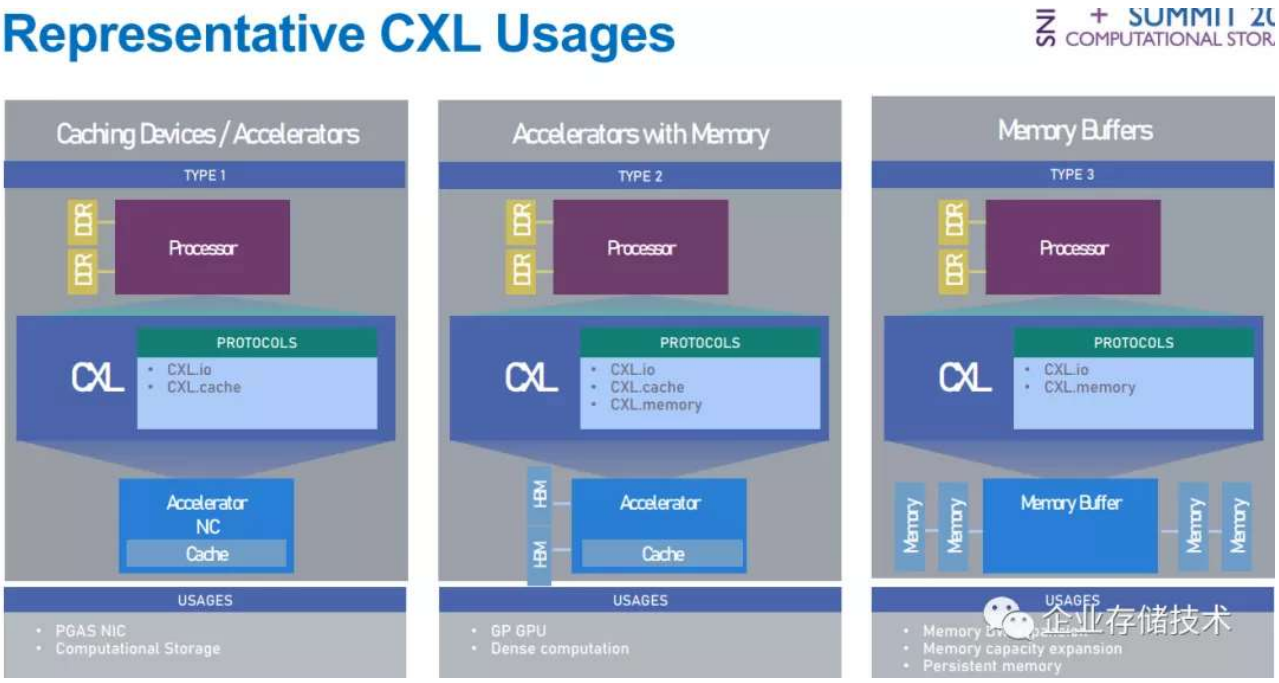
**先行者Gen-Z：不得不向Intel CXL低头？**

在讨论CXL之前有必要简单回顾一下Gen-Z，包括我之前写过2篇中的《*Gen-Z互连(下)：第一步25-100GB/s、PCI-SIG的反应*》，比如“复用PCIe pin”等，可以把Gen-Z看成CXL的先行者。




如果用Gen-Z Fabric直连SCM（存储级内存）或者GPU / FPGA加速器模块，需要在CPU一端提供Gen-Z Logic的支持。Intel仍占据80%以上的服务器市场，由于他们在该联盟中缺席，八卦一点地说：Gen-Z就有点“玩不下去”了。

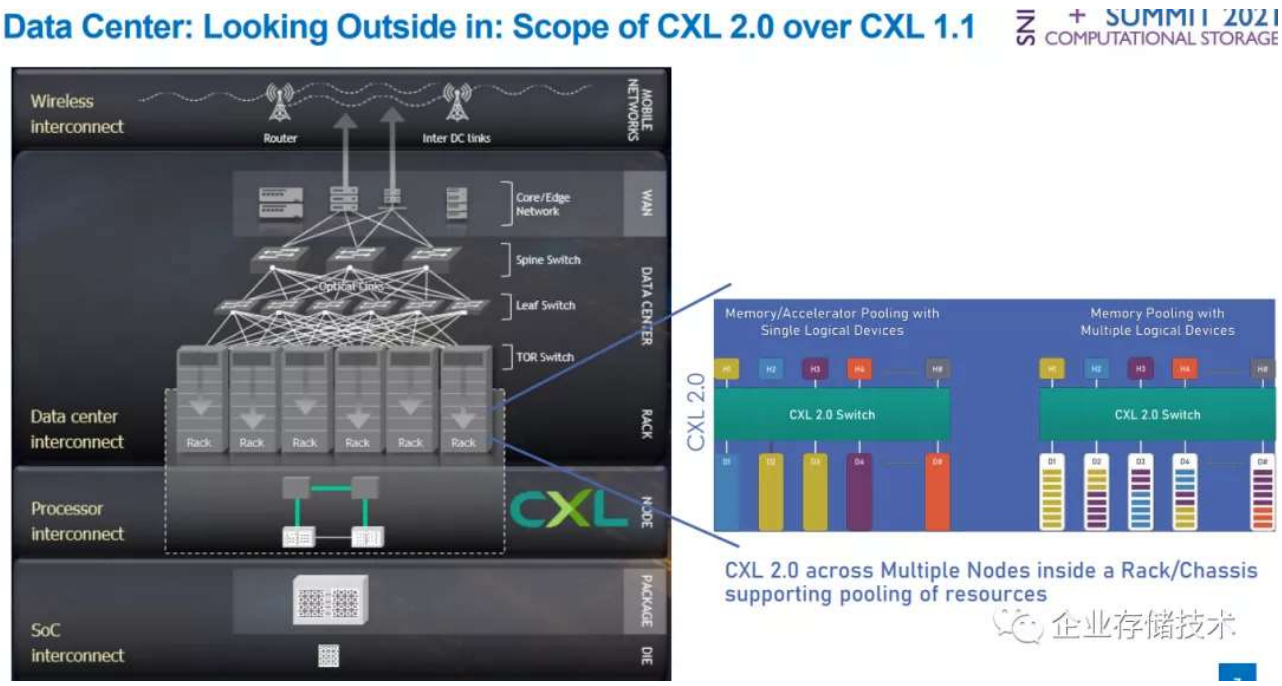
等到Intel开始推广CXL的时候，Gen-Z只好说“*大哥我错了，CPU直连那块我不碰了*”。于是大家就看到，现在只剩下右边通过Gen-Z Switch互连的部分。



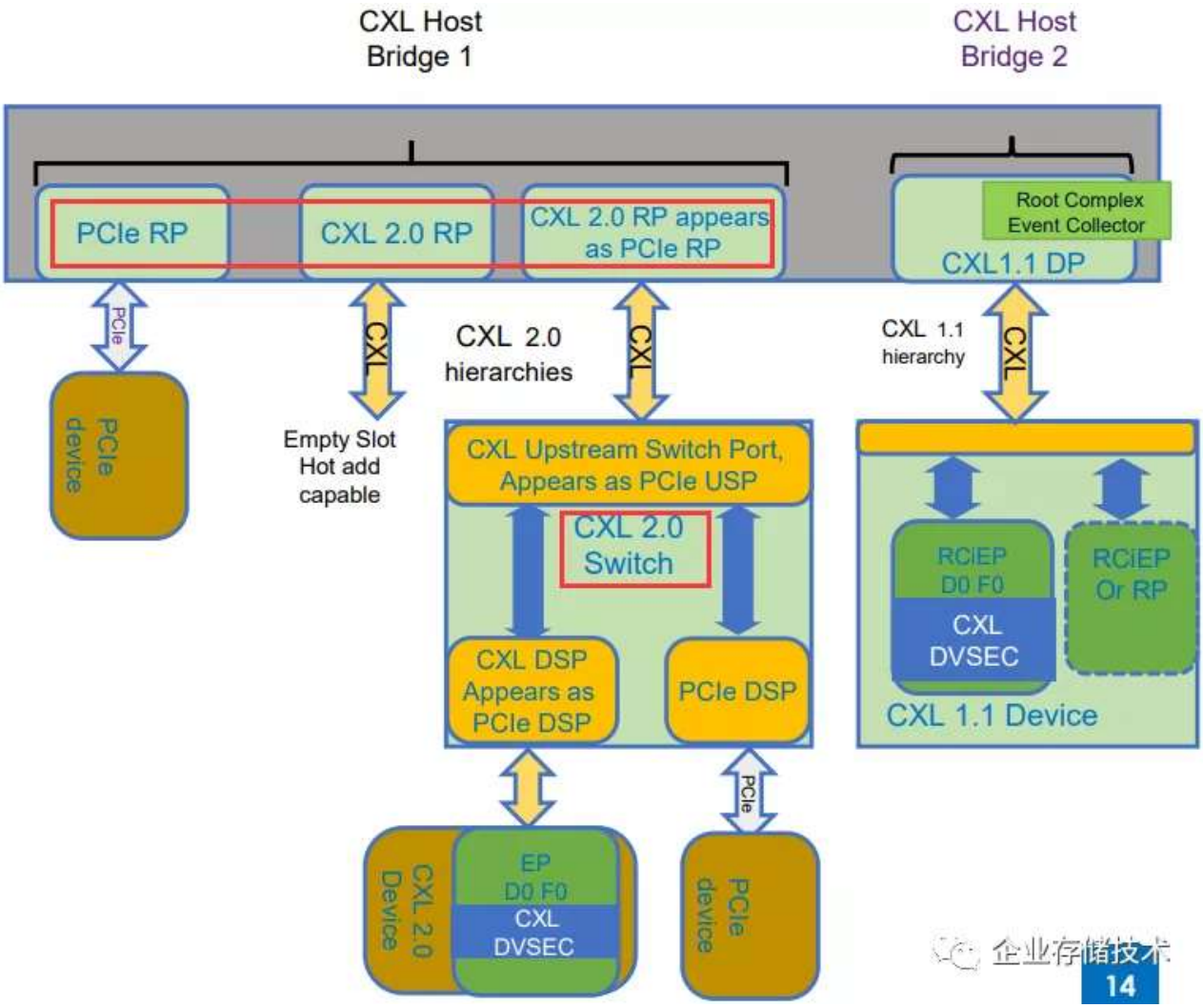

上图是CLX的3种典型用例：

- *Caching Device / Accelerators：加速器上的缓存，比如智能网卡、计算型存储器；*
- *Accelerators with Memory：带有内存的加速器，比如GPGPU、深度学习计算卡；*
- *Memory Bufffer：用于内存带宽、容量扩展，以及连接持久内存。*

具体的用到的协议包括CXL.io、CXL.cache和CXL.memory三种，在这里就不展开了。

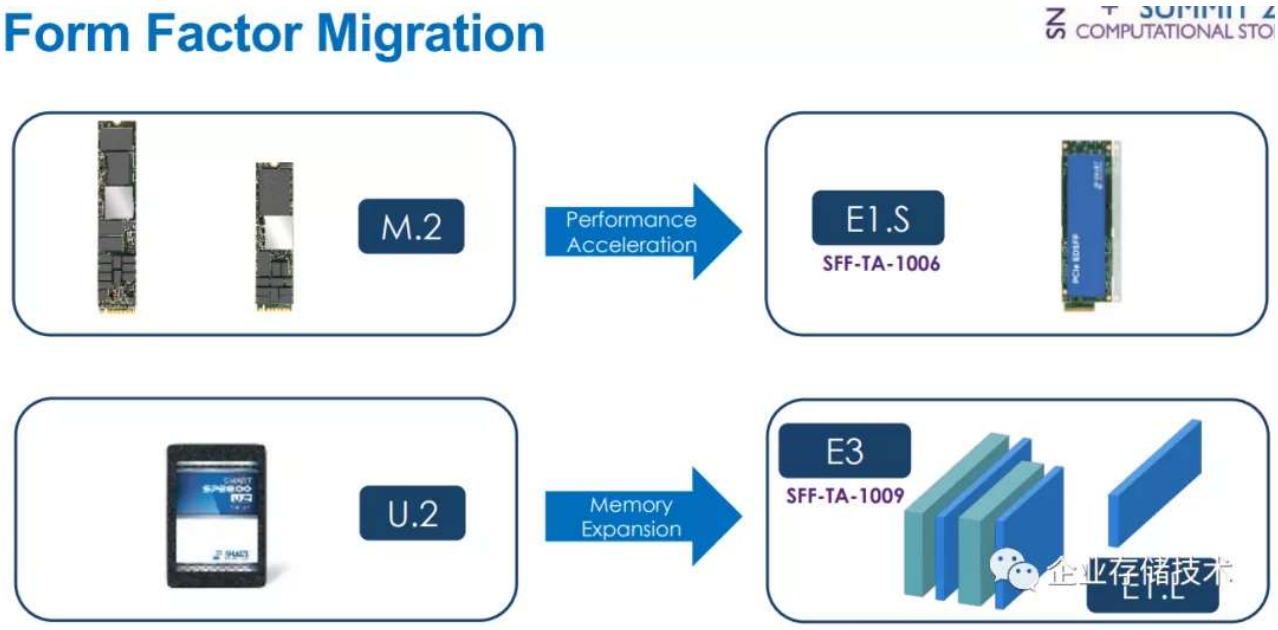


CXL 2.0规范引入了Switch，这样就能实现在机架/机箱内部的跨多节点互连，支持资源池化。

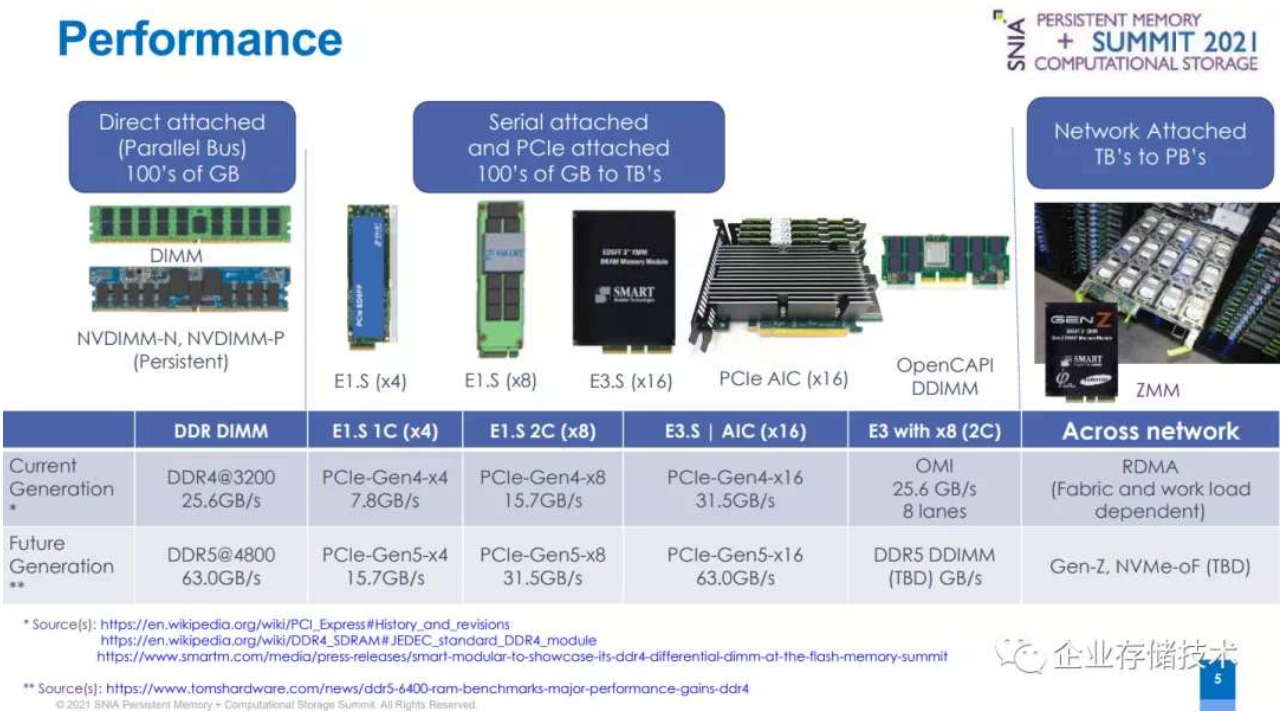



以我的技术水平看上图有些难度。从协议的角度，CXL RP（Root Port）与PCIe RP应该是处于同一层级，目前来看CXL底下也是PCIe 5.0物理层。

**350ns缓存一致性协议：在PCIe上超越NVMe的性能**




这里显示的（PCIe存储设备）**Form Factor迁移**：从M.2到E1.S（SFF-TA-1006）提高了性能和热插拔；从U.2到E3（SFF-TA-1009）和E1.L增加了内存扩展能力（由于CXL等的缓存一致性支持），至于还有存储/模块密度的提高，不是这里讨论的重点。




| | DDR DIMM | E1.S 1C (x4) | E1.S 2C (x8) | E3.S \| AIC (x16) | E3 with x8 (2C) | Across network |
|---|---|---|---|---|---|---|
| Current Generation * | DDR4@3200 25.6GB/s | PCIe-Gen4-x4 7.8GB/s | PCIe-Gen4-x8 15.7GB/s | PCIe-Gen4-x16 31.5GB/s | OMI 25.6 GB/s 8 lanes | RDMA (Fabric and work load dependent) |
| Future Generation ** | DDR5@4800 63.0GB/s | PCIe-Gen5-x4 15.7GB/s | PCIe-Gen5-x8 31.5GB/s | PCIe-Gen5-x16 63.0GB/s | DDR5 DDIMM (TBD) GB/s | Gen-Z, NVMe-oF (TBD) |

* Source(s): https://en.wikipedia.org/wiki/PCI_Express#History_and_revisions
https://en.wikipedia.org/wiki/DDR4_SDRAM#JEDEC_standard_DDR4_module
https://www.smartm.com/media/press-releases/smart-modular-to-showcase-its-ddr4-differential-dimm-at-the-flash-memory-summit

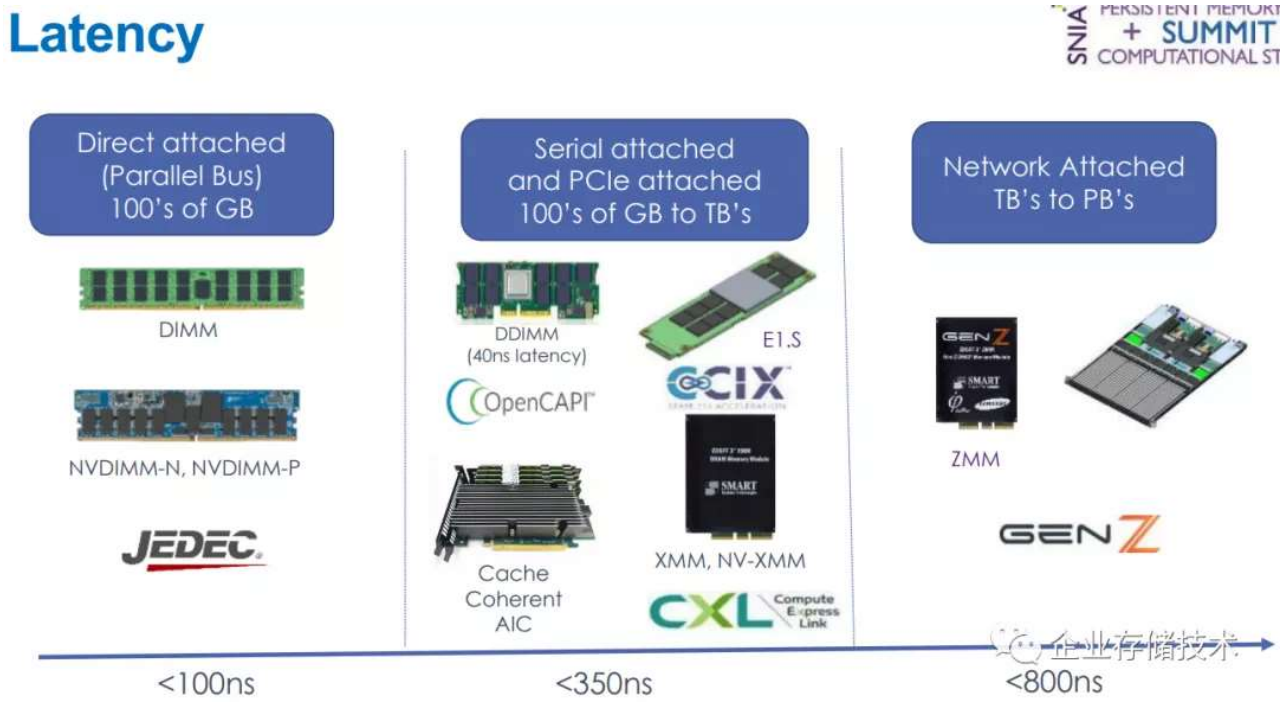
** Source(s): https://www.tomshardware.com/news/ddr5-6400-ram-benchmarks-major-performance-gains-ddr4



上图列出的是**带宽性能**。最左边的DDR DIMM内存是用并行总线直连，单个模组容量可达数百GB，当前和未来一代分别是DDR4@3200和DDR5@4800。

中间是串行连接（包括PCIe连接，从Gen4到Gen5都是多lane）模组，存储密度在数百GB到TB级别。其中包括E1.S单通道（x4）、E1.S双通道（x8）、E3.S / AIC（最多x16）以及OpenCAPI DDIMM（Differential内存，由于串行总线所以是差分信号）。

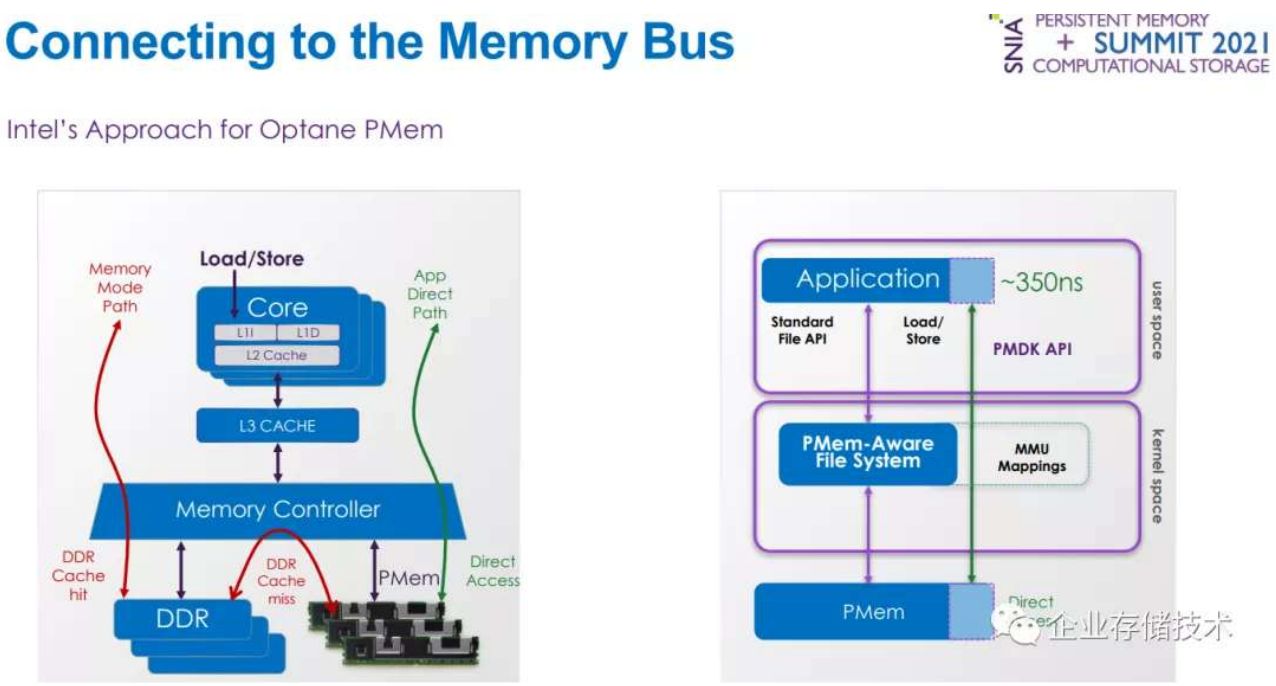
最右边是网络连接的存储介质，当前是用RDMA，未来除了Gen-Z，NVMe-oF还有不确定性。




接着对比**延时**。使用DRAM介质直连CPU的内存和NVDIMM不到100ns（后文中更进一步列出20ns）；通过PCIe串行连接的缓存一致性协议CXL（XMM、NV-XMM模组和AIC）、CCIX可以达到350ns延时；OpenCAPI的DDIMM也只有40ns；而Gen-Z这样经过外部Switch/网络连接的在800ns水平。


这里插一张图——某服务器厂商研发的一台原型机。已经被AMD收购的Xilinx，其VersalACAP（号称FPGA之后的新一代计算平台）卡插在AMD服务器的PCIe插槽上，实现了通过CCIX连接到DDR4控制器。这里的带宽，x8 CCIX相当于PCIe Gen4的16Gbps——或者x4 @ 32Gbps（PCIe Gen5）也就是16GB/s全双工。

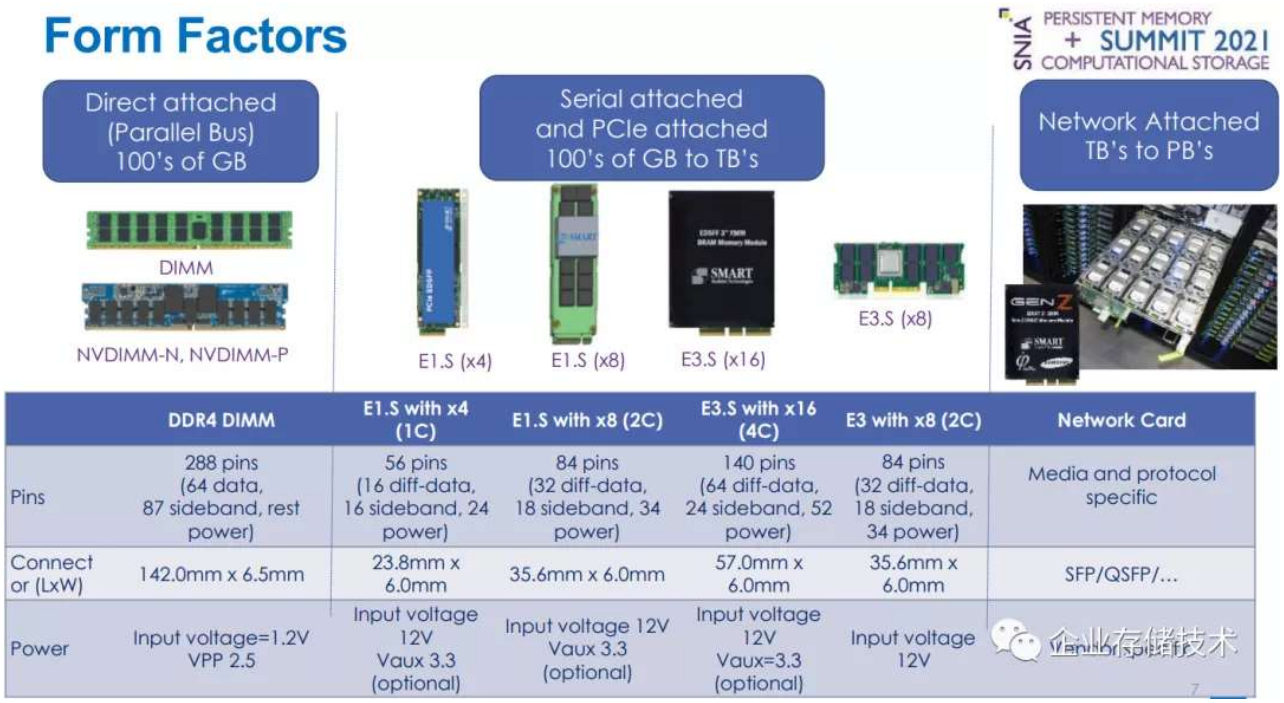
AMD也在CXL组织中，但是借助赛灵思的力量在CCIX上留一手我觉得也不错。此外，CCIX早就是ARM平台上的常客了，印象中多路CPU间互连就是用的这个吧。




再插一张图：Intel Optane Pmem（傲腾持久内存）的延时，即通过PMDK API访问也是350ns。从此处也可以侧面看出，在PCIe上那些新的缓存一致性协议的价值。

*注：NVMe的延时达到10μs以内就很不错了，即使是用户态的SPDK访问。毕竟还是块设备。*

**Form Factor：EDSFF用于内存扩展的价值**




| | DDR4 DIMM | E1.S with x4 (1C) | E1.S with x8 (2C) | E3.S with x16 (4C) | E3 with x8 (2C) | Network Card |
|---|---|---|---|---|---|---|
| Pins | 288 pins (64 data, 87 sideband, rest power) | 56 pins (16 diff-data, 16 sideband, 24 power) | 84 pins (32 diff-data, 18 sideband, 34 power) | 140 pins (64 diff-data, 24 sideband, 52 power) | 84 pins (32 diff-data, 18 sideband, 34 power) | Media and protocol specific |
| Connector (LxW) | 142.0mm x 6.5mm | 23.8mm x 6.0mm | 35.6mm x 6.0mm | 57.0mm x 6.0mm | 35.6mm x 6.0mm | SFP/QSFP/... |
| Power | Input voltage=1.2V VPP 2.5 | Input voltage 12V Vaux 3.3 (optional) | Input voltage 12V Vaux 3.3 (optional) | Input voltage 12V Vaux=3.3 (optional) | Input voltage 12V | Vendor specific |

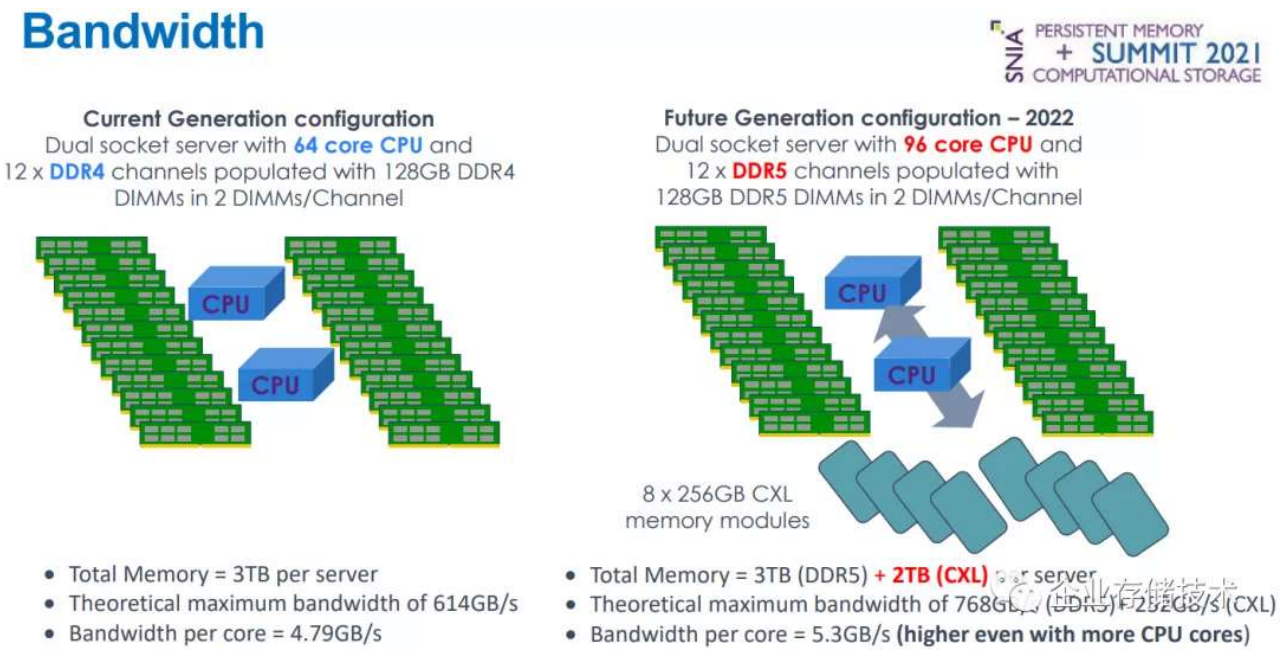


Form Factor包含的不只是物理尺寸，还有pin引脚数量和定义。

在DDR4 DIMM的288pin中，64条用于64bit并行总线（加上ECC校验实际上是72条）。

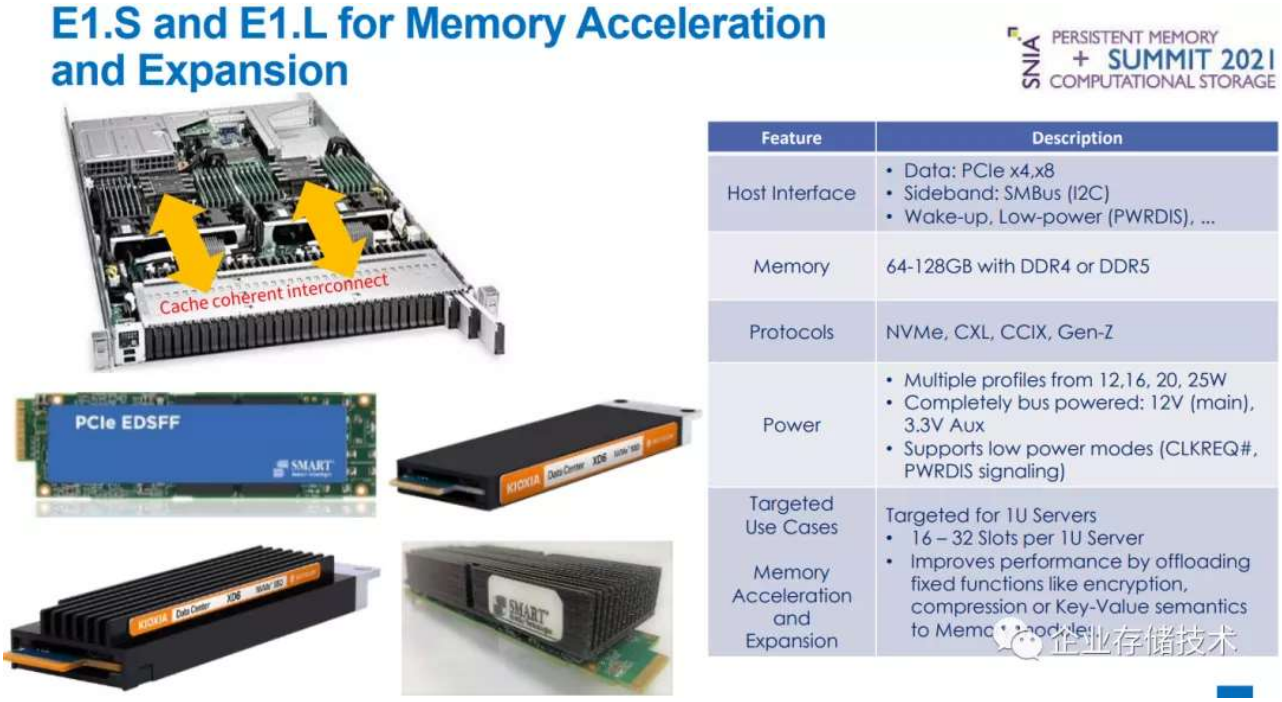
基于串行连接的PCIe部分，E1.S x4是在56pin中的16pin差分信号传输数据（PCIe每个lane双向共使用4pin）；以此类推，E1.S x8就是84pin中的32pin；E3.S x16是140pin中的64pin。

OpenCAPI看来也会使用E3.S x8，不过它只需要输入12V电压，而不像PCIe那样还有3.3Vaux。



具体到整个服务器系统的内存带宽，这张ppt左边的取值来源可能有点问题？双插槽64核CPU可以对应现在的AMD EPYC，而12个DDR4通道内存控制器目前x86还没有这么多的（会不会是ARM呢，毕竟也在CXL董事会里面）。

右边举例的意思是，当未来2022年CPU达到96核，除了DDR5内存速率更高（2x12通道总带宽768GB/s）之外，还会有CXL用于补充连接内存（252GB/s）。这样平均每核心带宽甚至可以比当前这一代更高点。




当**E1.S**和**E1.L**用于内存加速和扩展时，单个模组可以做到64-128GBDDR4或DDR5，除了当前的NVMe协议之外，还有CXL、CCIX和Gen-Z（我对后两者用在Intel平台上直连CPU不太看好）。相对于传统DIMM上插槽内存的价值，可以把一些固定功能如：加密、压缩或者Key-Value语义卸载到内存模组，来改进性能。

## E3.S and E3.L for Memory Expansion

SNIA + SUMMIT 202
COMPUTATIONAL STORAG


| Feature | Description |
|---|---|
| Host Interface | • Data: PCIe x16<br>• Sideband: SMBus (I2C)<br>• Wake-up, Low-power (PWRDIS), ... |
| Memory | Up to 256GB with DDR4 or DDR5<br>* Non-volatile persistent memory feature could be support on this form-factor using back-up and restore functionality like in NVDIMM-N. |
| Protocols | NVMe, CXL, CCIX, Gen-Z |
| Power | • 2 profiles 25W (thin) and 40W (thick)<br>• Bus powered: 12V (main), 3.3V Aux<br>• Supports low power modes (CLKREQ#, PWRDIS signaling) |
| Targeted Use Cases<br><br>Memory Expansion | Targeted for 2U Server<br>Enables 4TB – 8 TB of Memory expansion with 16 E3.S modules in single server chassis, better throughput than direct attached DDR4 DIMM. |



当**E3.S和E3.L**用于内存扩展时，单个模组可以做到256GB DDR4或DDR5，其实我想在上图中脑补加上一个三星最近宣布的CXL标准DDR5内存（如下图）。在这个尺寸上可以像NVDIMM-N那样使用备份（到闪存）和恢复功能，来支持非易失持久内存。

在**2U服务器**上，可以使用16个E3.S模组来扩展4TB-8TB内存，达到比直连DDR4 DIMM更好的吞吐。

## OpenCAPI High Bandwidth Memory - DDIMM

SNIA PERSISTENT MEMORY + SUMMIT 2021 COMPUTATIONAL STORAGE

OpenCAPI™

| Feature | Description |
|---|---|
| Host Interface | • OpenCAPI |
| Memory | • Up to 256GB |
| Protocols | • OMI – Open Memory Interface<br>• The memory bus is defined with one read port and one write port per channel, each having eight unidirectional differential lanes |
| Performance | • DDR4-3200<br>• Latency 40ns<br>• Data throughput rate of 25.6GB/s with 8 lanes<br>• The DDIMM/OMI approach delivers up to 4TB of memory on a server at about 320GB/second or 512GB at up to 650GB/s sustained rates. |
| Targeted Use Cases | • Targeted for servers<br>• High bandwidth, low latency serial connection for memory, accelerators, network, storage, and other devices like ASICs |



**OpenCAPI**是用于哪家服务器一看上图就清楚吧。OMI是一种高带宽低延时、串行连接的内存总线，读和写接口各有8对单向的差分信号，DDR4-3200数据带宽25.6GB/s。

除了支持内存之外，理论上还可以用于网卡、存储（SSD?）、ASIC等加速器。我记得Power9可以将OpenCAPI重定义为NVLINK来连接GPU。

**非易失内存：从NVDIMM到CXL定义的NV-XMM**

## NVDIMM for Persistent Memory

SNIA PERSISTENT MEMORY + SUMMIT 2021 COMPUTATIONAL STORAGE

**Key Features of DDR4 and DDR5 NVDIMM-N**

- Operation like DRAM
- Fast recovery from system power loss
- Software overhead can be eliminated

JEDEC

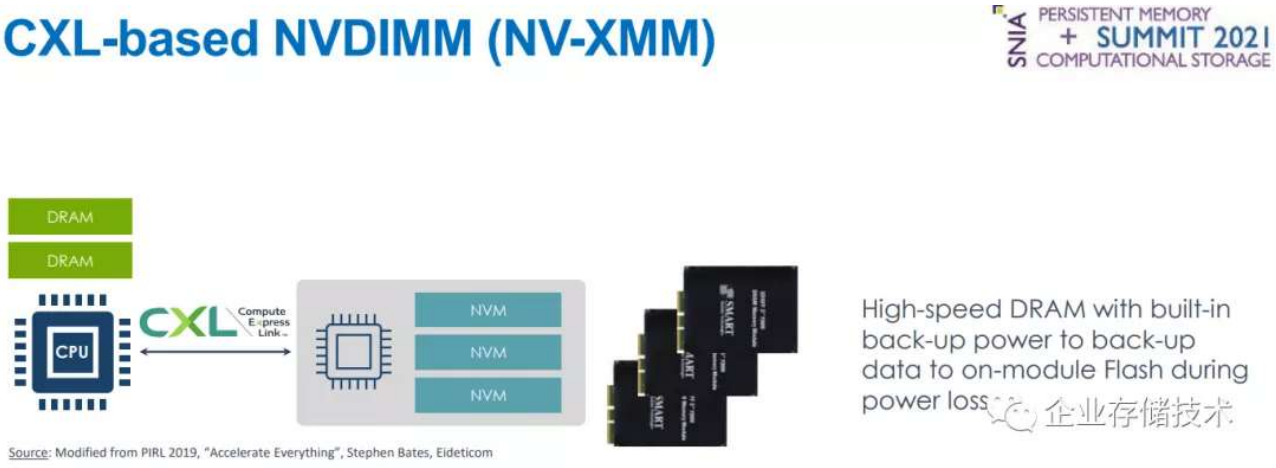
SNIA

+

Backup Power

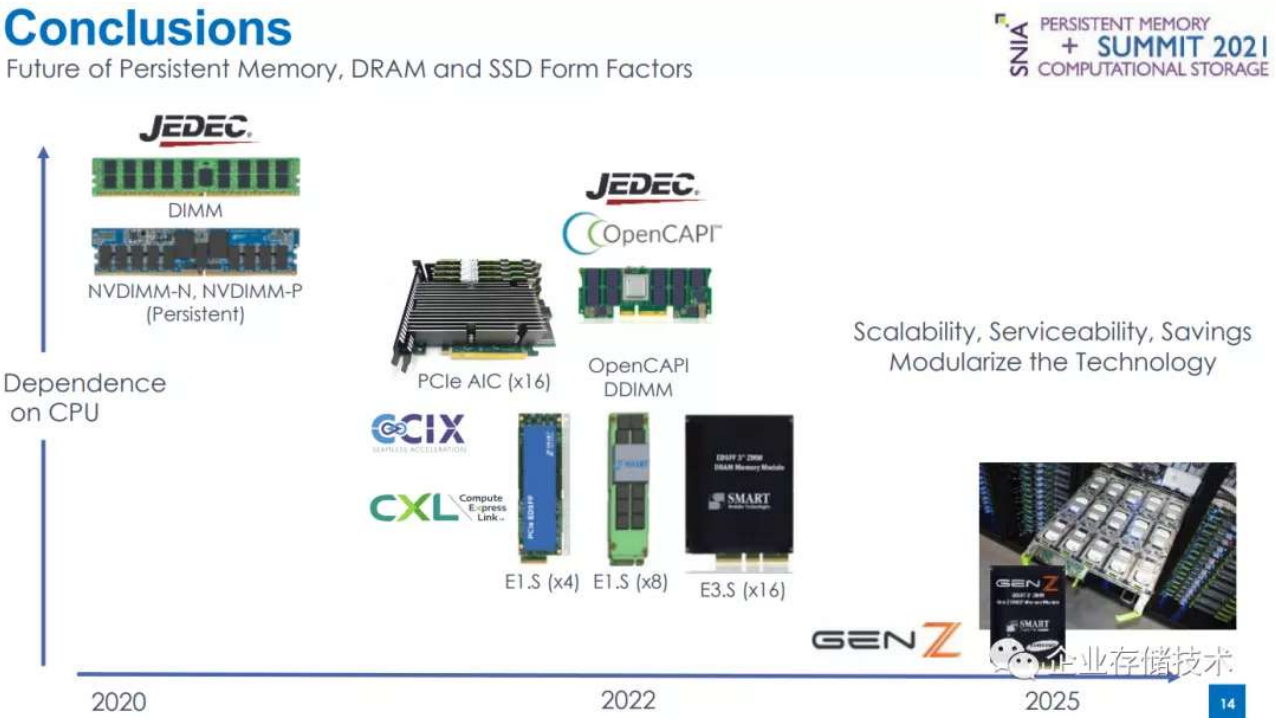
| Feature | Description |
|---|---|
| Host Interface | • DDR |
| Memory | • DDR4 16GB, 32GB<br>• DDR5 32GB, 64GB |
| Protocol | • JEDEC Compliant DDR4 / DDR5 |
| Features | • Throughput of 25.6GB/s (DDR4)<br>• Latency ~20ns<br>• AES 256 bit Encryption |
| Targeted Use Cases | • All Flash Arrays, Storage Servers, HPC, AI Training Servers<br>• Needed for very low latency tiering, caching, write buffering, storage, checkpointing<br>• Needed for AI/ML algorithm processing |

传统的**NVDIMM-N**最大的优势就是像内存一样访问，消除了软件上的开销，它的延时可以达到约20ns。比如在全闪存阵列和存储服务器中用于Cache、写缓冲、元数据存储，以及数据库等的checkpoint创建。




基于CXL的NVDIMM被称为**NV-XMM**，它像NVDIMM一样在模组上配备了闪存，支持掉电备份数据。

大家还记得Dell EMC PowerStore全闪存阵列上，NVMe（U.2）接口的NVRAM写缓存盘吗？NV-XMM有个同样的好处就是支持双端口，能让双控制器同时访问，这是DIMM形态NVRAM不具备的。另外就是未来的PCIe 5.0 x16接口，理论带宽可达当前PCIe 3.0 U.2 NVMe SSD的大约16倍。




*图中横坐标为年份，纵坐标是对CPU的依赖性*

简单总结一下。从时间点上来看，本文讨论的未来持久内存、DRAM和SSD的形态，新协议CXL、CCIX预计会伴随DDR5、PCIe 5.0，在新一代Intel Sapphire Rapids以及AMD服务器平台上出现；OpenCAPI也可能过渡到DDR5；而Gen-Z恐怕要等到2025年了...

*2021 Persistent Memory + Computational Storage Summit 会议 & 参考资料打包分享*
链接：https://pan.baidu.com/s/14RVFMMBjSfCdpfTEmxNJdA
提取码：exy8

SNIA官网链接 https://www.snia.org/pm-summit（含视频，需翻墙）

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。进一步交流技术，可以加我的微信/QQ：490834312。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：）*



长按二维码可直接识别关注

历史文章汇总：http://www.toutiao.com/c/user/5821930387/
http://www.zhihu.com/column/huangliang

点击下方“阅读原文”，查看更多历史文章

↓↓↓

收录于话题 #服务器 65

上一篇
AMD EPYC Genoa ‘7004’：若DDR5换OMI，CXL.mem服务器内存池化

下一篇
冷板式液冷标准化：PowerEdge 15G服务器散热杂谈

阅读原文　阅读 4198　文章已于2021/05/17修改

分享　收藏　赞 27　在看 26

写下你的留言

精选留言

Chinαχing
cpu内存通道的每个通道pin睡独立的72个?

企业存储技术(作者)
是啊，这还只是数据引脚

游侠
intel第四代 xeon scalable 4477针