# CXL技术调研报告

一、CXL简要介绍与背景

二、CXL应用场景

三、CXL协议的演进

四、CXL子协议

五、CXL的发展

六、CXL未来展望

CXL
Compute
Express
Link

2024/9/12

# CXL简要介绍

**Compute Express Link (CXL)** 是一种较新的高速互联标准，旨在实现处理器和设备（加速器、内存缓冲区、敏捷互联网卡和固态存储等）之间的高速通信。 CXL提供一致性和内存语义，带宽基于PCIe，带宽可以与PCIe的带宽一致，同时实现比PCIe低的多的延迟。

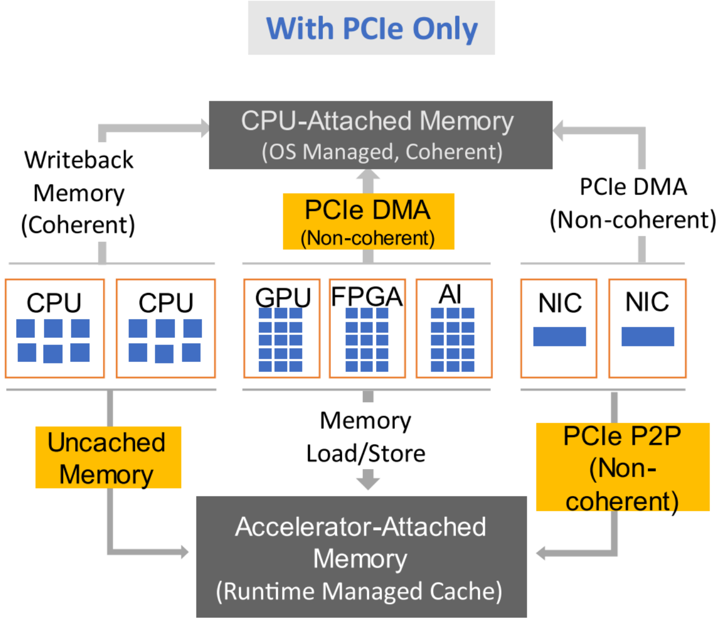
CXL于2019年由Inter提出，并于2019年3月与阿里巴巴、思科、戴尔、谷歌、华为、Meta、Microsoft和HPE成立了CXL联盟。

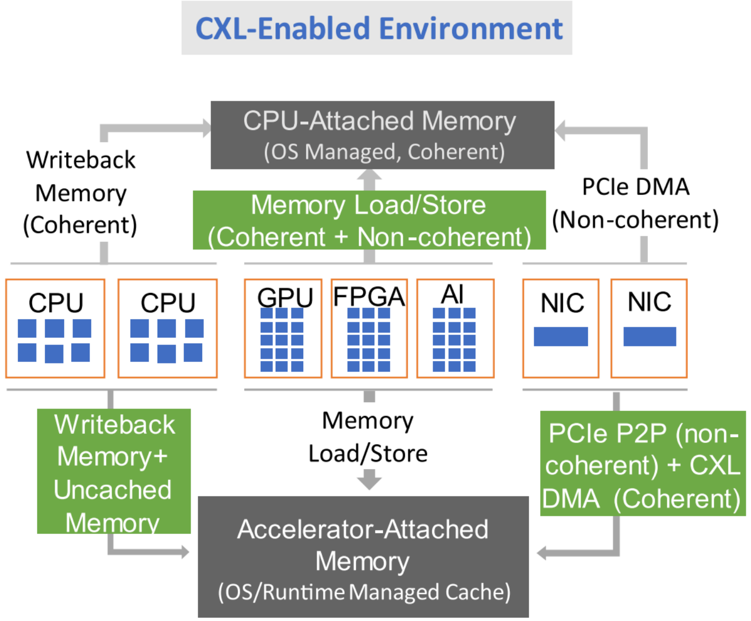
# CXL要解决的挑战

1. 对系统和设备内存的一致访问
2. 内存的可扩展性
3. 浪费资源和计算效率低下
4. 分布式系统中细粒度数据共享的延迟问题

# 对系统和设备内存的一致性访问

挑战：在传统架构中，系统内存通常通过DDR接口与CPU相连，并由CPU缓存层次结构进行缓存，而PCIe设备访问系统内存时通常是非一致性（non-coherent）的。这导致PCIe设备无法缓存系统内存，无法利用缓存的时间局部性或空间局部性，无法进行原子操作。而设备本地内存也无法映射到系统的可缓存地址空间。


With PCIe Only
CPU-Attached Memory
(OS Managed, Coherent)
Writeback
Memory
(Coherent)
PCIe DMA
(Non-coherent)
PCIe DMA
(Non-coherent)
CPU
CPU
GPU
FPGA
AI
NIC
NIC
Uncached
Memory
Memory
Load/Store
PCIe P2P
(Non-
coherent)
Accelerator-Attached
Memory
(Runtime Managed Cache)


CXL-Enabled Environment
CPU-Attached Memory
(OS Managed, Coherent)
Writeback
Memory
(Coherent)
Memory Load/Store
(Coherent + Non-coherent)
PCIe DMA
(Non-coherent)
CPU
CPU
GPU
FPGA
AI
NIC
NIC
Writeback
Memory+
Uncached
Memory
Memory
Load/Store
PCIe P2P (non-
coherent) + CXL
DMA (Coherent)
Accelerator-Attached
Memory
(OS/Runtime Managed Cache)

# 内存的可扩展性

挑战：随着计算需求的指数级增长，内存带宽和容量的需求也在不断增加，但传统的DDR内存由于引脚效率低下，扩展成本高且带宽受限，无法满足这些需求。尤其是对于多服务器环境，增加DDR通道不仅提高了平台成本，还带来了信号完整性问题。

**CXL**的解决方案：CXL利用PCIe物理层的高引脚带宽扩展内存带宽和容量，通过CXL.memory协议支持内存扩展，允许使用不同类型的内存，并且可以通过CXL接口将这些内存连接到处理器，实现更高效的内存扩展。

# 浪费资源和计算效率低下

挑战：在现代数据中心中，计算和内存资源往往是紧耦合的（tight coupling），每个服务器的资源只能供自己使用，导致资源利用率低。例如，当一个服务器的计算资源被完全利用时，可能还剩余大量未使用的内存，反之亦然。为了应对峰值需求，通常会为每个服务器过度配置资源，这导致了资源的浪费和效率低下。

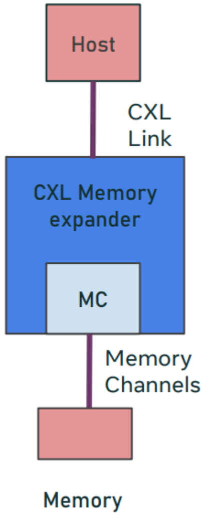
**CXL**的解决方案：CXL通过引入资源池化机制（Resource Pooling），允许在多个服务器之间动态分配内存和其他资源。这样，未充分利用的资源可以在多个服务器之间共享，提高了数据中心的整体资源利用率，避免资源浪费。

# 分布式系统中细粒度数据共享的延迟问题

挑战：在分布式系统中，工作负载往往需要进行细粒度的同步和数据共享，这些操作对延迟非常敏感，尤其是在更新较小数据块时。传统的网络延迟较高，无法满足这些场景的实时需求。

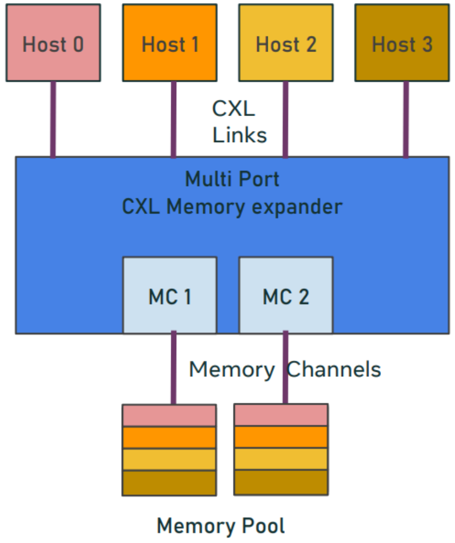
**CXL**的解决方案：CXL提供了一致性内存共享模型，支持多台主机之间的细粒度数据共享。通过CXL的低延迟互连，多个系统可以共享数据结构，执行同步操作或使用内存传递消息，这极大地降低了分布式系统中的通信延迟，提高了系统的响应速度。

# CXL应用场景


Host
CXL Link
CXL Memory expander
MC
Memory Channels
Memory

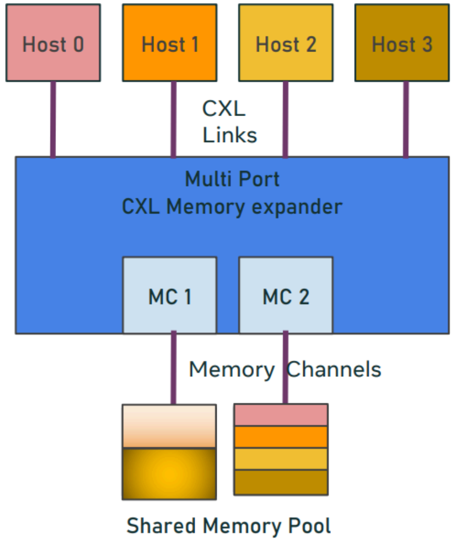
## Direct attached

Add Capacity
Add Bandwidth
Slower-cheaper tier


Host 0
Host 1
Host 2
Host 3
CXL Links
Multi Port CXL Memory expander
MC 1
MC 2
Memory Channels
Memory Pool

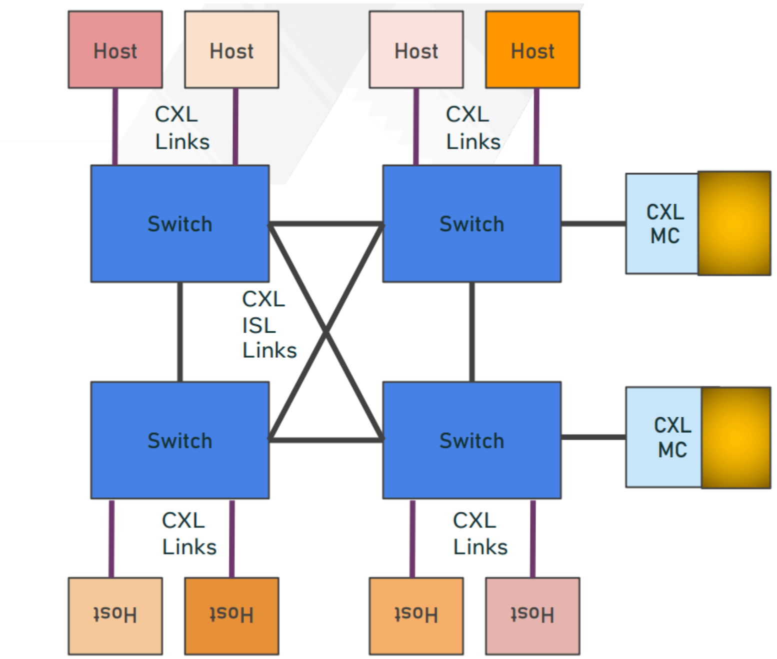
## Pooled Memory

Amortize CXL infra cost
Flexible allocation


Host 0
Host 1
Host 2
Host 3
CXL Links
Multi Port CXL Memory expander
MC 1
MC 2
Memory Channels
Shared Memory Pool

## Shared Memory

Deduplication
Host2host communication
large datasets

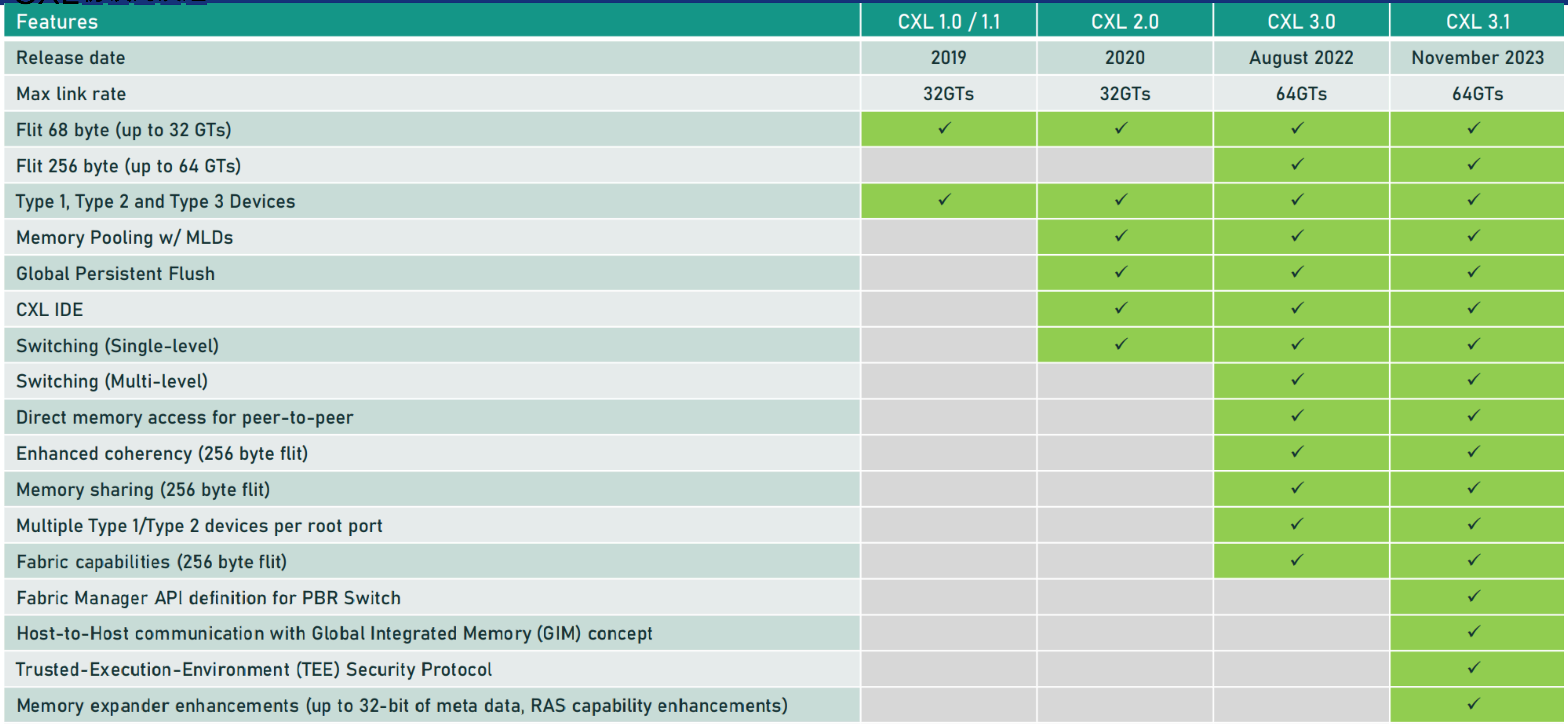

Host
Host
Host
Host
CXL Links
CXL Links
Switch
Switch
CXL MC
CXL ISL Links
Switch
Switch
CXL MC
CXL Links
CXL Links
Host
Host
Host
Host

## Fabric Memory

Scaling to huge datasets

2024/9/12

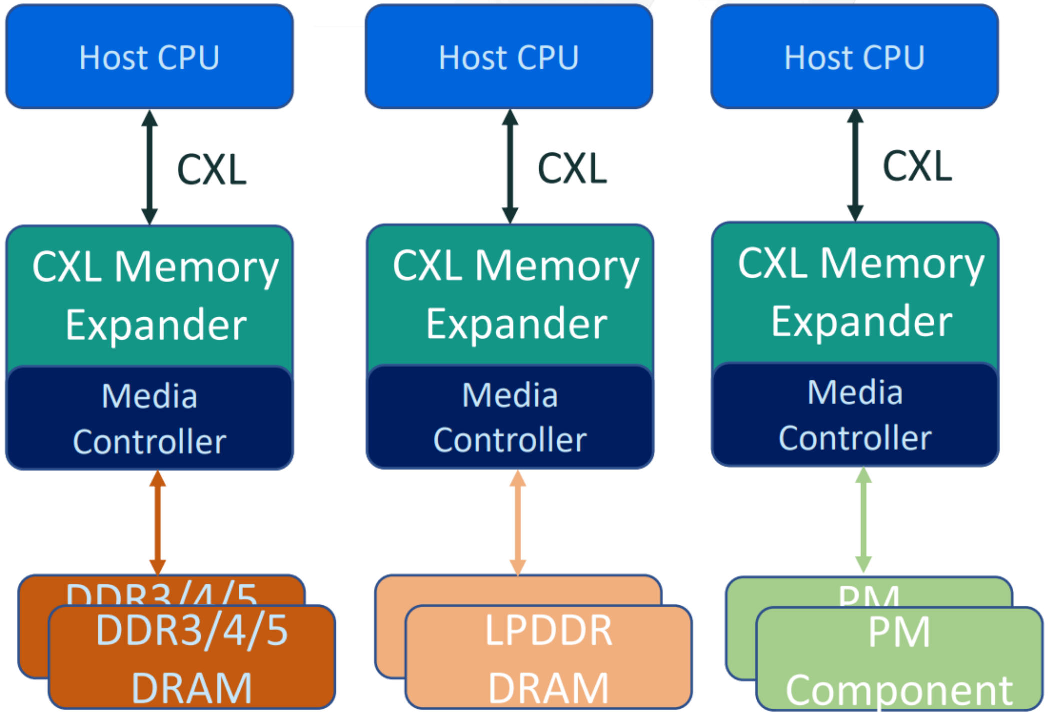
# CXL 协议的演进



| Features | CXL 1.0 / 1.1 | CXL 2.0 | CXL 3.0 | CXL 3.1 |
|---|---|---|---|---|
| Release date | 2019 | 2020 | August 2022 | November 2023 |
| Max link rate | 32GTs | 32GTs | 64GTs | 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 |  |  | ✓ | ✓ |
| Enhanced coherency (256 byte flit) |  |  | ✓ | ✓ |
| Memory sharing (256 byte flit) |  |  | ✓ | ✓ |
| Multiple Type 1/Type 2 devices per root port |  |  | ✓ | ✓ |
| Fabric capabilities (256 byte flit) |  |  | ✓ | ✓ |
| Fabric Manager API definition for PBR Switch |  |  |  | ✓ |
| Host-to-Host communication with Global Integrated Memory (GIM) concept |  |  |  | ✓ |
| Trusted-Execution-Environment (TEE) Security Protocol |  |  |  | ✓ |
| Memory expander enhancements (up to 32-bit of meta data, RAS capability enhancements) |  |  |  | ✓ |

# CXL1.1内存拓展

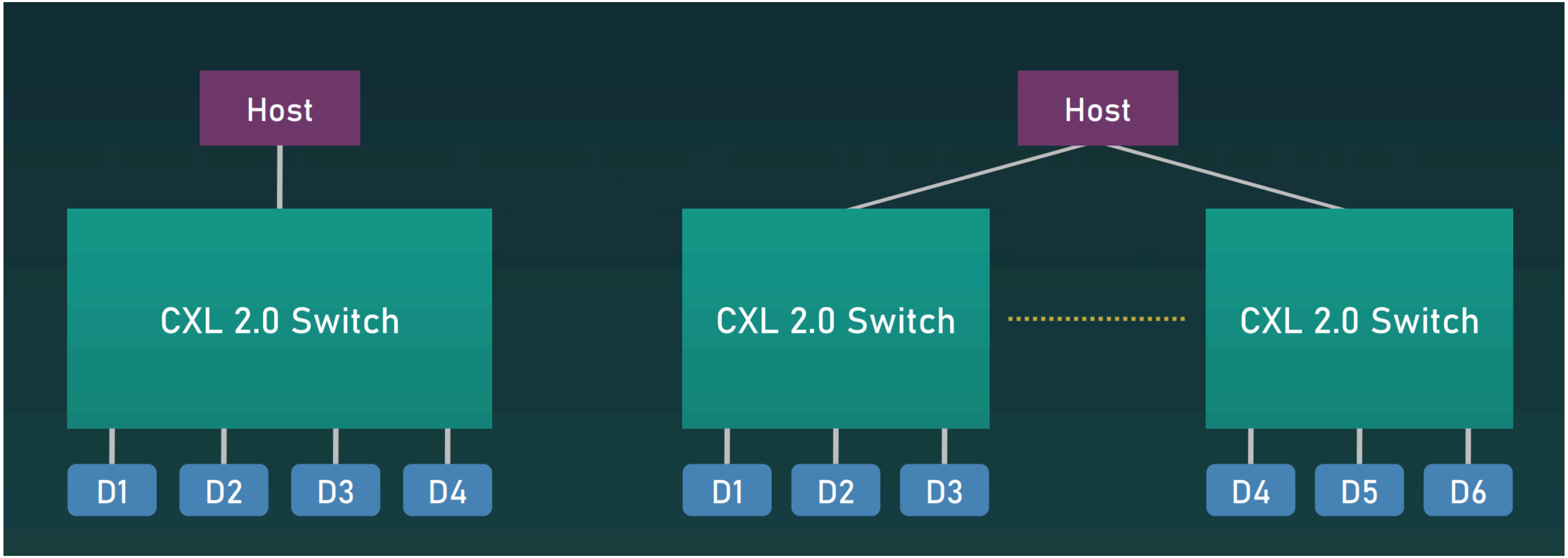
- CXL 给各种内存设备提供标准串行接口
- 远端内存与本地内存异步迭代：远端内存可以是 DDR3/4/5, LPDDR3/4/5, 非易失内存等。
- 融合不同内存特点(persistence, latency, BW, endurance, etc)


Host CPU
CXL
CXL Memory Expander
Media Controller
DDR3/4/5
DDR3/4/5 DRAM
Host CPU
CXL
CXL Memory Expander
Media Controller
LPDDR DRAM
Host CPU
CXL
CXL Memory Expander
Media Controller
PM
PM Component

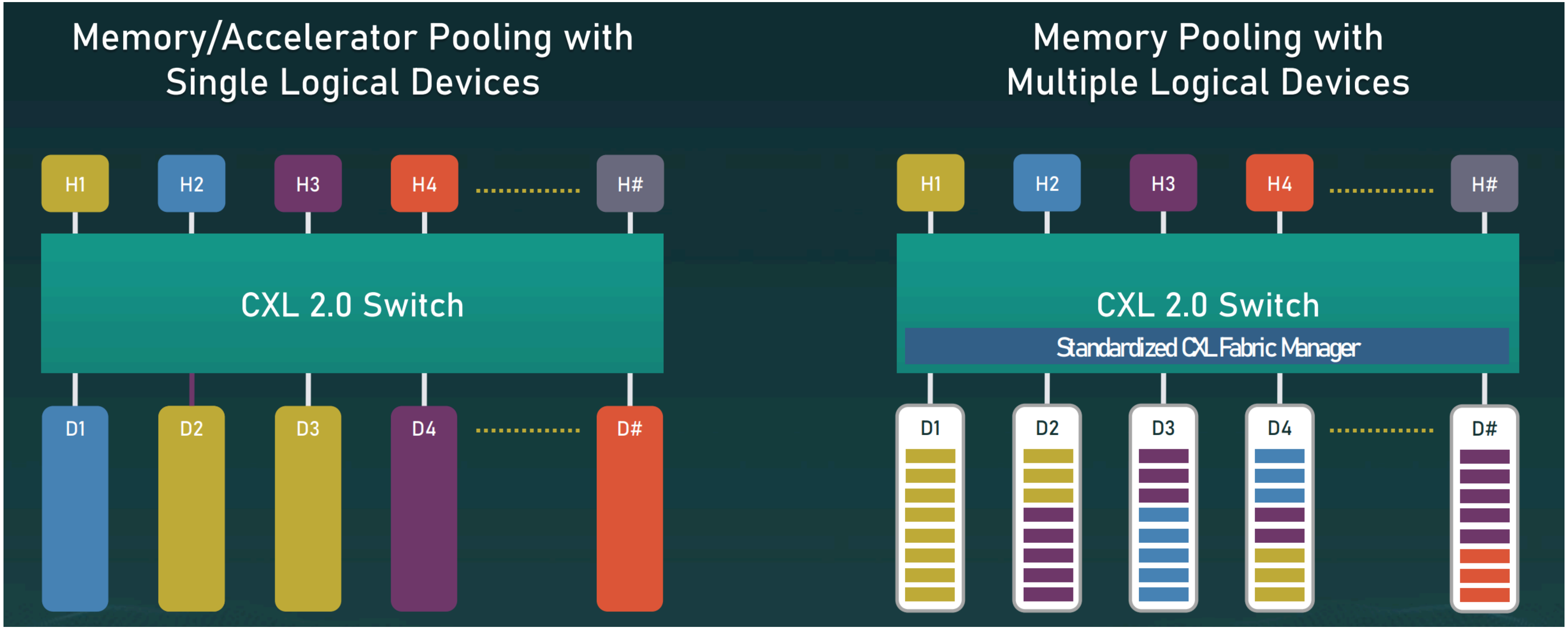
2024/9/12

# CXL2.0 switch

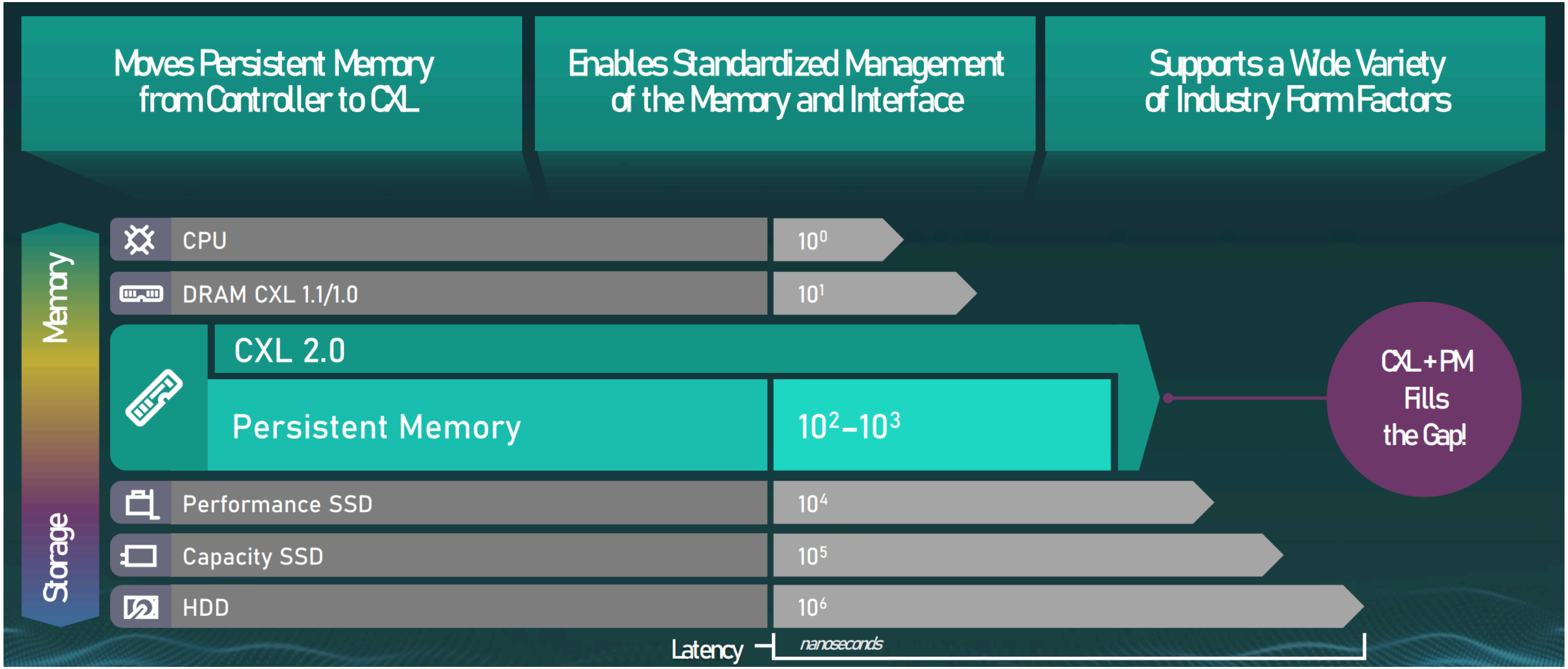
CXL2.0 □□single switch


Host
CXL 2.0 Switch
D1
D2
D3
D4
Host
CXL 2.0 Switch
CXL 2.0 Switch
D1
D2
D3
D4
D5
D6

# CXL2.0内存池化


Memory/Accelerator Pooling with Single Logical Devices
H1
H2
H3
H4
H#
CXL 2.0 Switch
D1
D2
D3
D4
D#
Memory Pooling with Multiple Logical Devices
H1
H2
H3
H4
H#
CXL 2.0 Switch
Standardized CXL Fabric Manager
D1
D2
D3
D4
D#

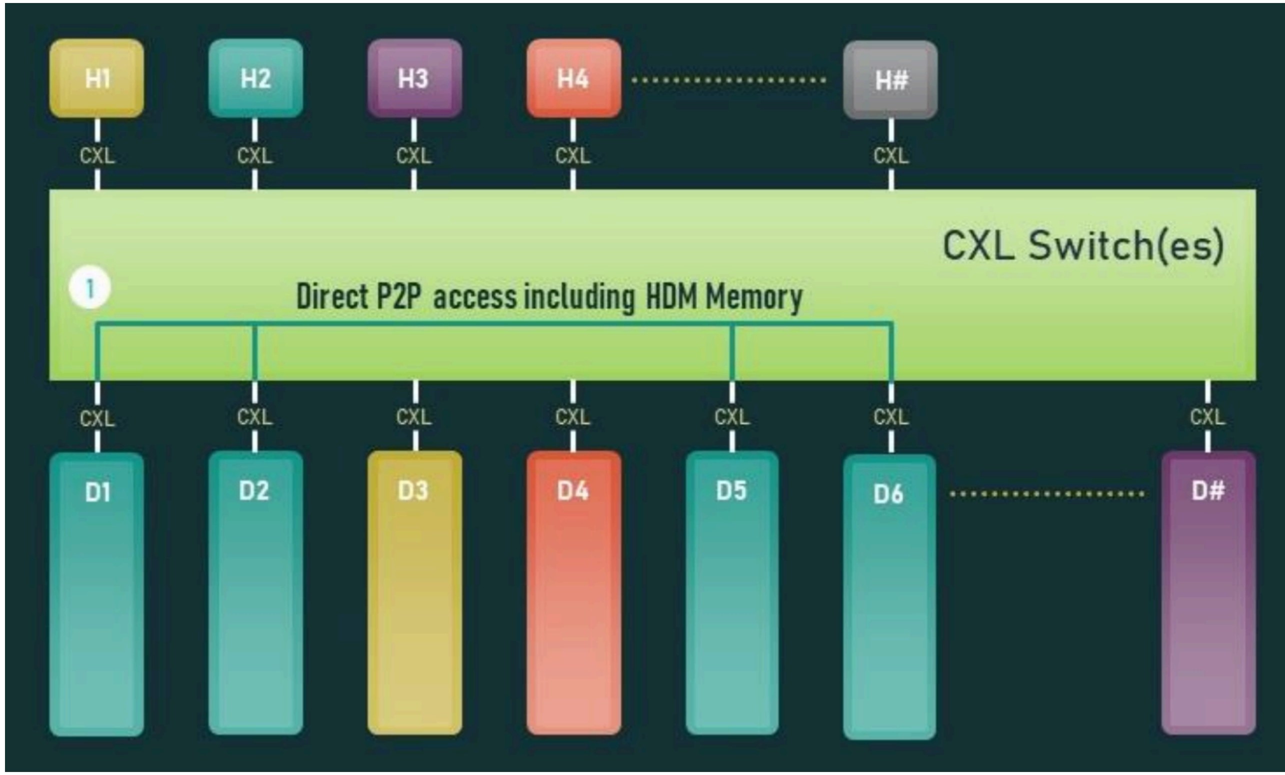
# CXL2.0 持久化内存


Moves Persistent Memory from Controller to CXL
Enables Standardized Management of the Memory and Interface
Supports a Wide Variety of Industry Form Factors
Memory
Storage
CPU
$10^0$
DRAM CXL 1.1/1.0
$10^1$
CXL 2.0
Persistent Memory
$10^2$–$10^3$
CXL + PM Fills the Gap!
Performance SSD
$10^4$
Capacity SSD
$10^5$
HDD
$10^6$
Latency
nanoseconds

2024/9/12

# CXL支持p2p交互的对等通信

- 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 conflic

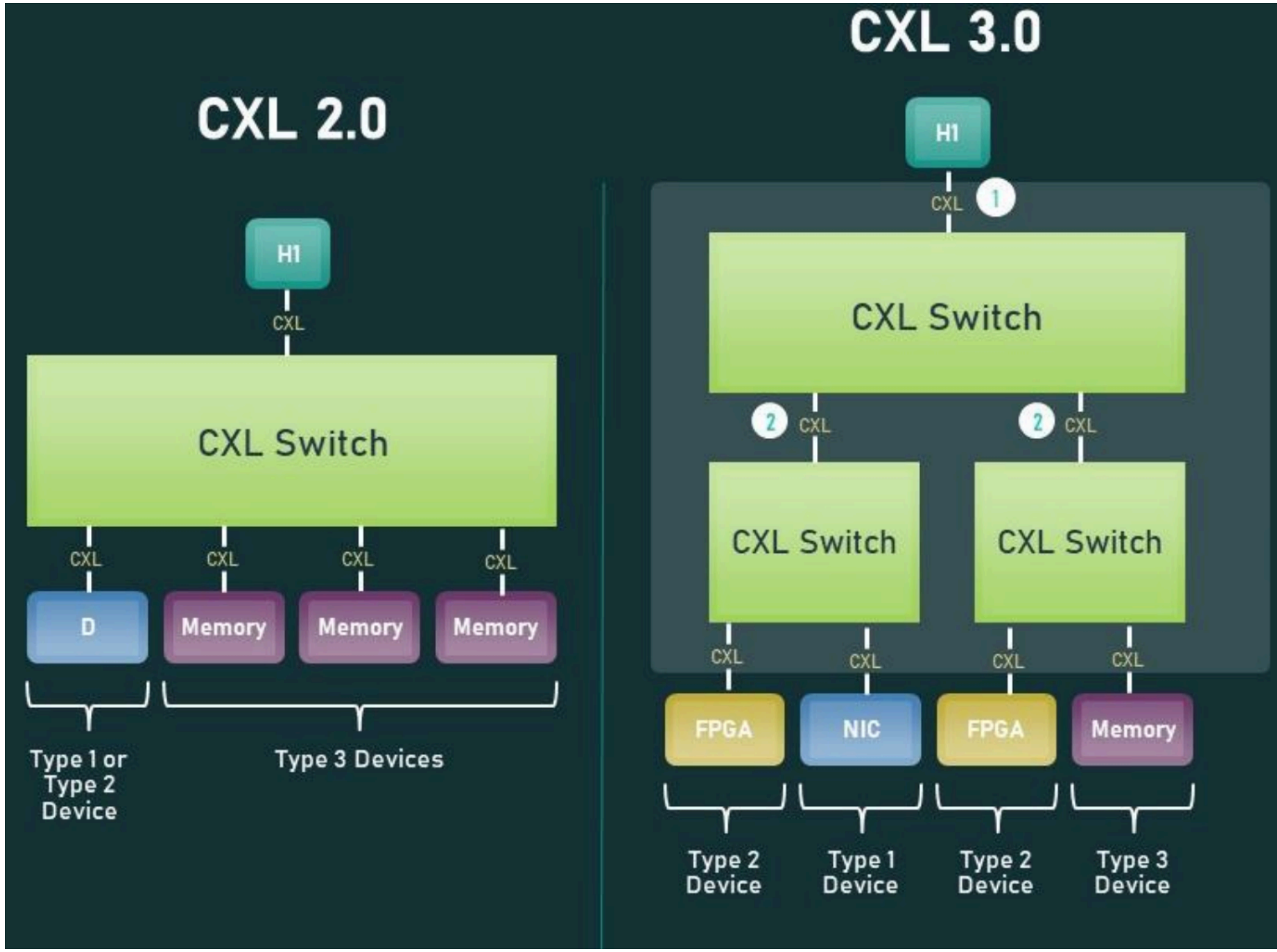

H1
H2
H3
H4
H#
CXL
CXL
CXL
CXL
CXL
CXL Switch(es)
1
Direct P2P access including HDM Memory
CXL
CXL
CXL
CXL
CXL
CXL
CXL
D1
D2
D3
D4
D5
D6
D#

2024/9/12

# CXL3.0特性

每个主机都可以连接多种设备（最多**16**个**CXL.cache**设备）

支持多级交换，允许多个**CXL**设备通过交换机连接，增强系统的可扩展性和灵活性。


CXL 2.0
H1
CXL
CXL Switch
CXL
CXL
CXL
CXL
D
Memory
Memory
Memory
Type 1 or Type 2 Device
Type 3 Devices
CXL 3.0
H1
CXL
1
CXL Switch
2
CXL
2
CXL
CXL Switch
CXL Switch
CXL
CXL
CXL
CXL
FPGA
NIC
FPGA
Memory
Type 2 Device
Type 1 Device
Type 2 Device
Type 3 Device

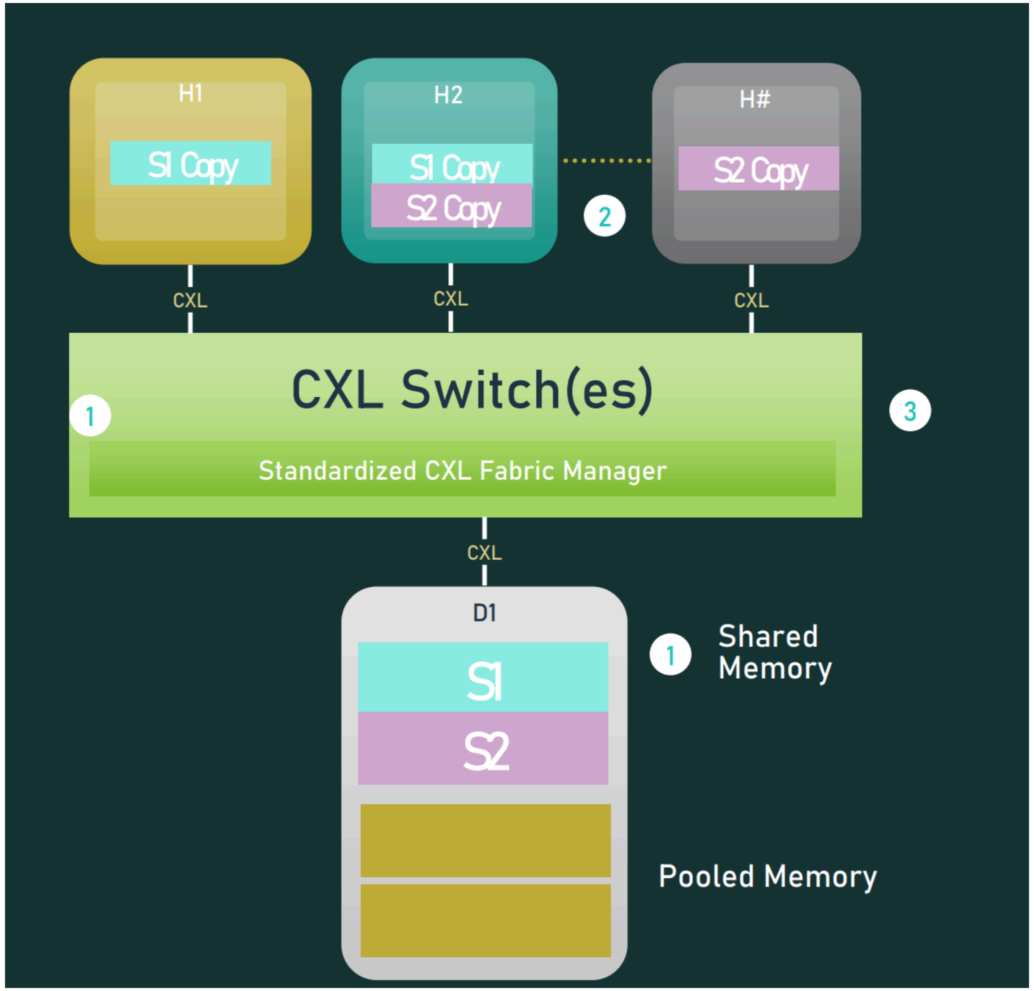
2024/9/12

# CXL3.0特性

设备内存可以被主机共享

主机可以在主机缓存中拥有共享区域或共享区域部分的一致性副本

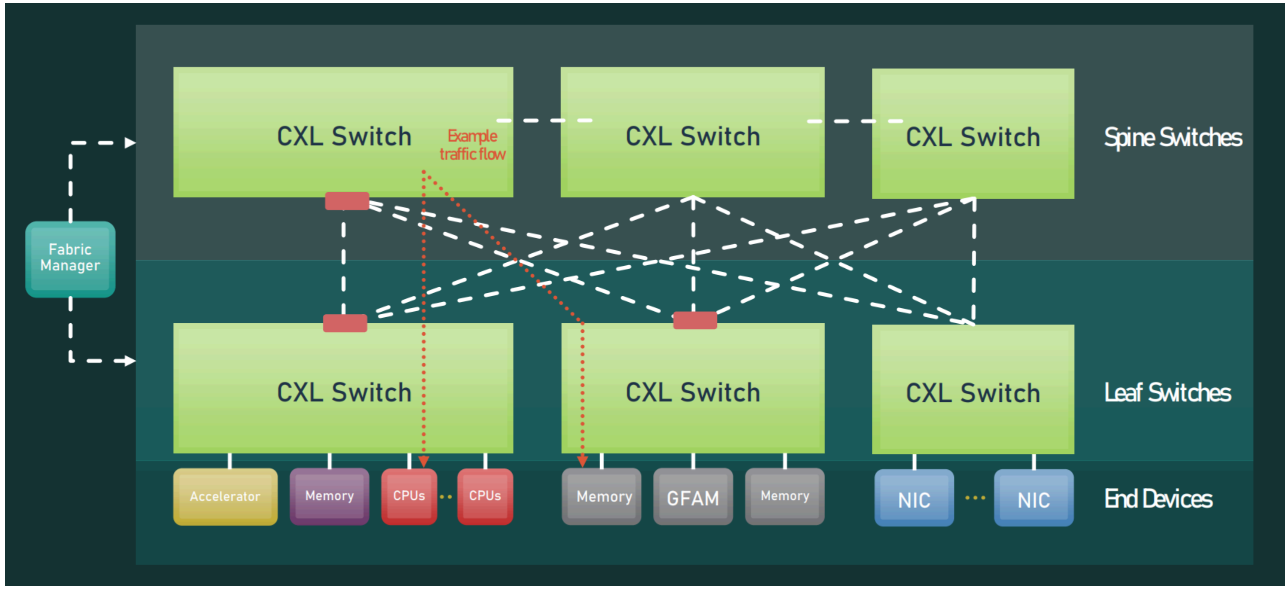
**CXL 3.0** 定义了在副本之间强制执行硬件缓存一致性的机制，确保在多主机或多设备访问共享资源时，缓存数据保持一致。


H1
S1 Copy
H2
S1 Copy
S2 Copy
H#
S2 Copy
2
CXL
CXL
CXL
1
CXL Switch(es)
3
Standardized CXL Fabric Manager
CXL
D1
S1
S2
1
Shared Memory
Pooled Memory

2024/9/12

# CXL3.0特性

引入了**CXL Fabric**架构，支持构建更大规模的系统互连，允许更多设备和主机通过**CXL Fabric**进行互联。这有助于实现更复杂的拓扑结构，并显著提高系统的可扩展性。


Fabric Manager
CXL Switch
Example traffic flow
CXL Switch
CXL Switch
Spine Switches
CXL Switch
CXL Switch
CXL Switch
Leaf Switches
Accelerator
Memory
CPUs
CPUs
Memory
GFAM
Memory
NIC
NIC
End Devices

2024/9/12

CXL3.1□□

## CXL Fabic Improvements/Extensions

- Scale out of CXL fabrics using PBR(Port Based Routing)

## Trusted-Execution-Enviroment Security Protocol(TSP)

- Allows for Virtualization-based, Trusted-Execution-Environments (TEEs) to host Confidential Computing workloads

## Memory Expander Improvements

- Up to 32-bit of meta-data and RAS capabiility enhancements

# CXL子协议

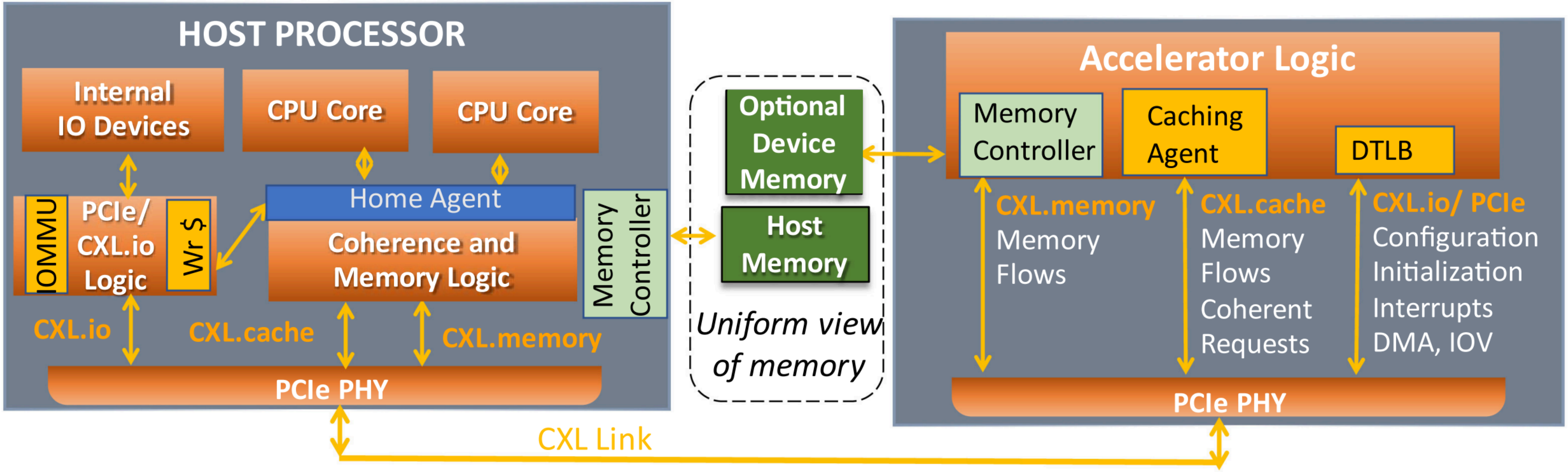
CXL由三个协议组成，分别是**CXL.io**、**CXL.cache**和**CXL.memory**，它们在PCIe物理层上多路复用。

**CXL.io**协议基于PCIe。它用于设备发现、状态报告、虚拟到物理地址的转换以及直接内存访问（DMA）。

**CXL.cache**协议允许设备（如加速器、FPGA等）缓存主机内存，并与主机共享一致性缓存的语义。

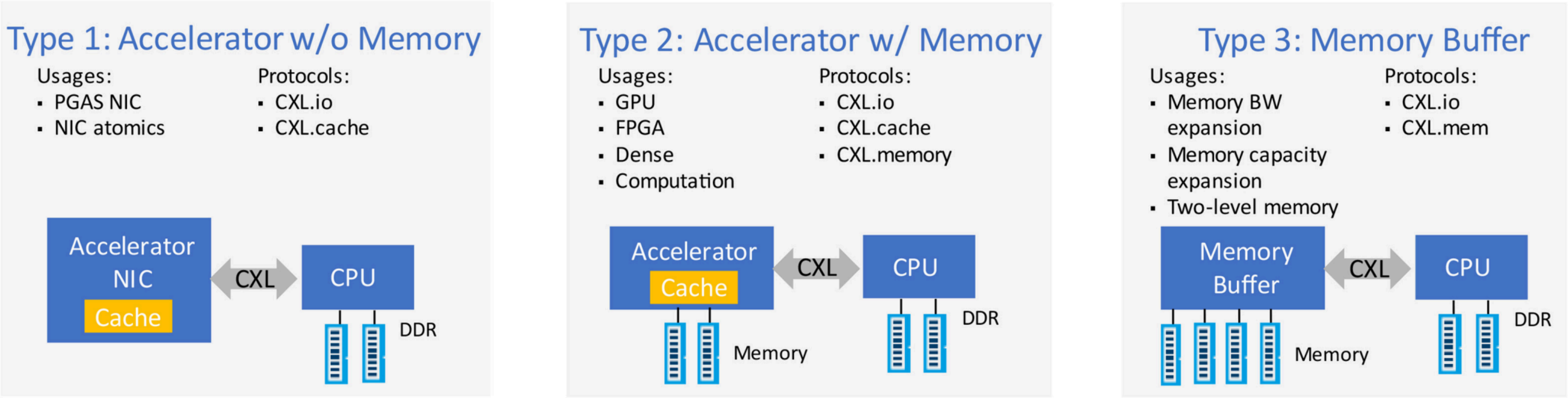
**CXL.memory**协议使设备能够直接访问主机管理的设备内存（Host-Managed Device Memory, HDM），并允许处理器将设备内存映射为系统内存的一部分。

# CXL子协议


HOST PROCESSOR
Internal IO Devices
CPU Core
CPU Core
IOMMU
PCIe/ CXL.io Logic
Wr $
Home Agent
Coherence and Memory Logic
Memory Controller
CXL.io
CXL.cache
CXL.memory
PCIe PHY
Optional Device Memory
Host Memory
Uniform view of memory
Accelerator Logic
Memory Controller
Caching Agent
DTLB
CXL.memory
Memory Flows
CXL.cache
Memory Flows
Coherent Requests
CXL.io/ PCIe
Configuration
Initialization
Interrupts
DMA, IOV
PCIe PHY
CXL Link

2024/9/12

# CXL□□



### Type 1: Accelerator w/o Memory

Usages:
- PGAS NIC
- NIC atomics

Protocols:
- CXL.io
- CXL.cache

### Type 2: Accelerator w/ Memory

Usages:
- GPU
- FPGA
- Dense
- Computation

Protocols:
- CXL.io
- CXL.cache
- CXL.memory

### Type 3: Memory Buffer

Usages:
- Memory BW expansion
- Memory capacity expansion
- Two-level memory

Protocols:
- CXL.io
- CXL.mem

Accelerator NIC
Cache
CXL
CPU
DDR
Accelerator
Cache
CXL
CPU
Memory
DDR
Memory Buffer
CXL
CPU
Memory
DDR

2024/9/12

# CXL应用实例

CXL 生态系统需要 CPU 和设备支持。CXL 1.1 和 CXL 2.0 已部署在一些商业产品中，CXL 3.0 产品预计将在几年内问世。

英特尔支持在 Sapphire Rapids（SPR）CPU和Agilex7 FPGA 上开始的 CXL，支持所有三种协议。AMD在Genoa和Bergamo CPU 中支持 CXL，并已宣布支持SmartNIC设备，ARM已宣布在V2、N2和E2系列CPU中支持CXL 2.0。

三星构建了CXL1.1内存扩展设备，并公开分享了其基准测试结果。ontage、SK Hynix、Microchip、Micron和Astera Labs宣布推出 CXL Type3存储设备。美光已经设计了一种CXL1.1近内存计算设备的原型。

2024/9/12

# CXL的未来方向

在计算机体系结构中，CXL允许独立开发内存控制器功能，如自适应DRAM刷新和低成本可靠性改进，推动更快的定制迭代。研究可能涉及内存延迟、缓存管理和预取策略的优化。

在计算机系统中，CXL内存池化将改变本地和分布式内存管理，需应对内存压力、网络拥塞和QoS的挑战。

在计算机工程中，研究重点是降低CXL内存和加速器的延迟，同时改进错误管理，并通过UCIe支持跨机架的可组合系统，提升性能和降低成本。

# 汇报完毕，恳请指正

**宋敬炎**

2024/9/12