Compute Express Link (CXL)

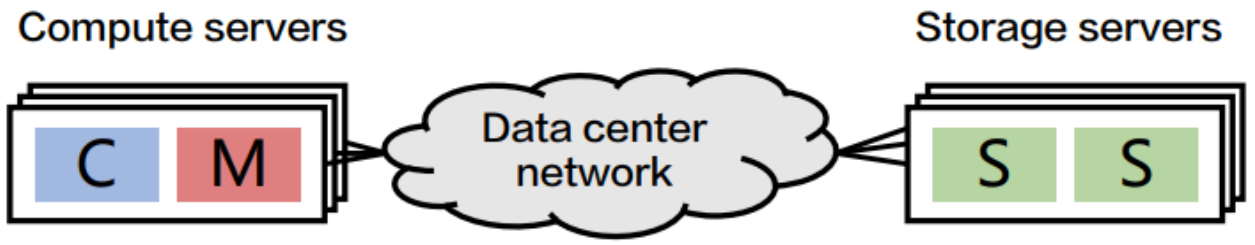
20231202

Keynote

- Memory disaggregation
- Compute Express Link
  - CXL protocols
  - Types of CXL device
- DirectCXL [ATC’22]

Storage Disaggregation

- Separates secondary storage devices (e.g., SSDs) from compute servers

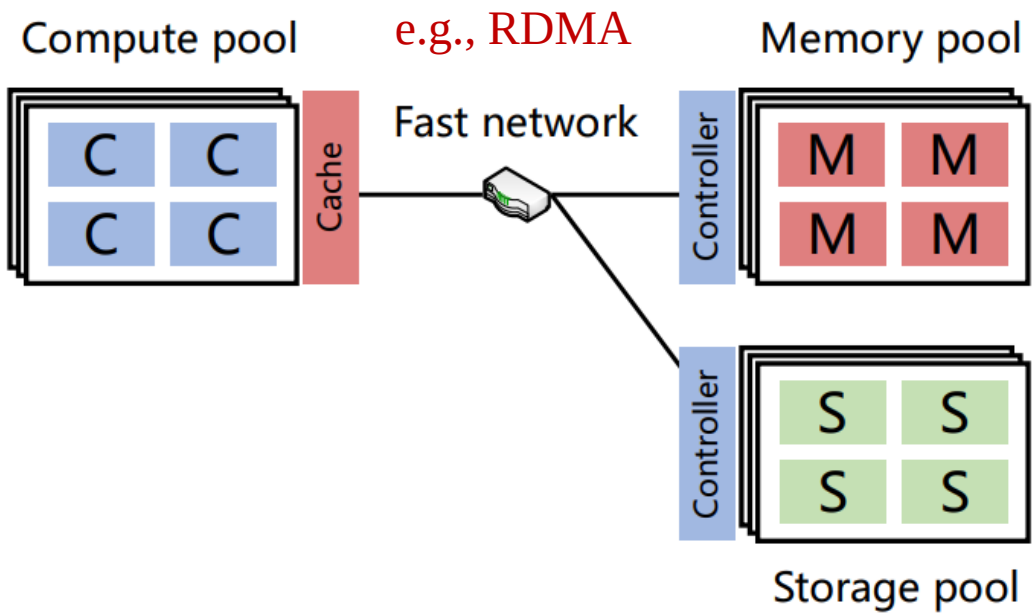


C, M tight coupling

S decoupling

Memory Disaggregation

- On the basis of storage disaggregation
- Memory Disaggregation
  - Split compute and memory with Remote Direct Memory Access (RDMA)

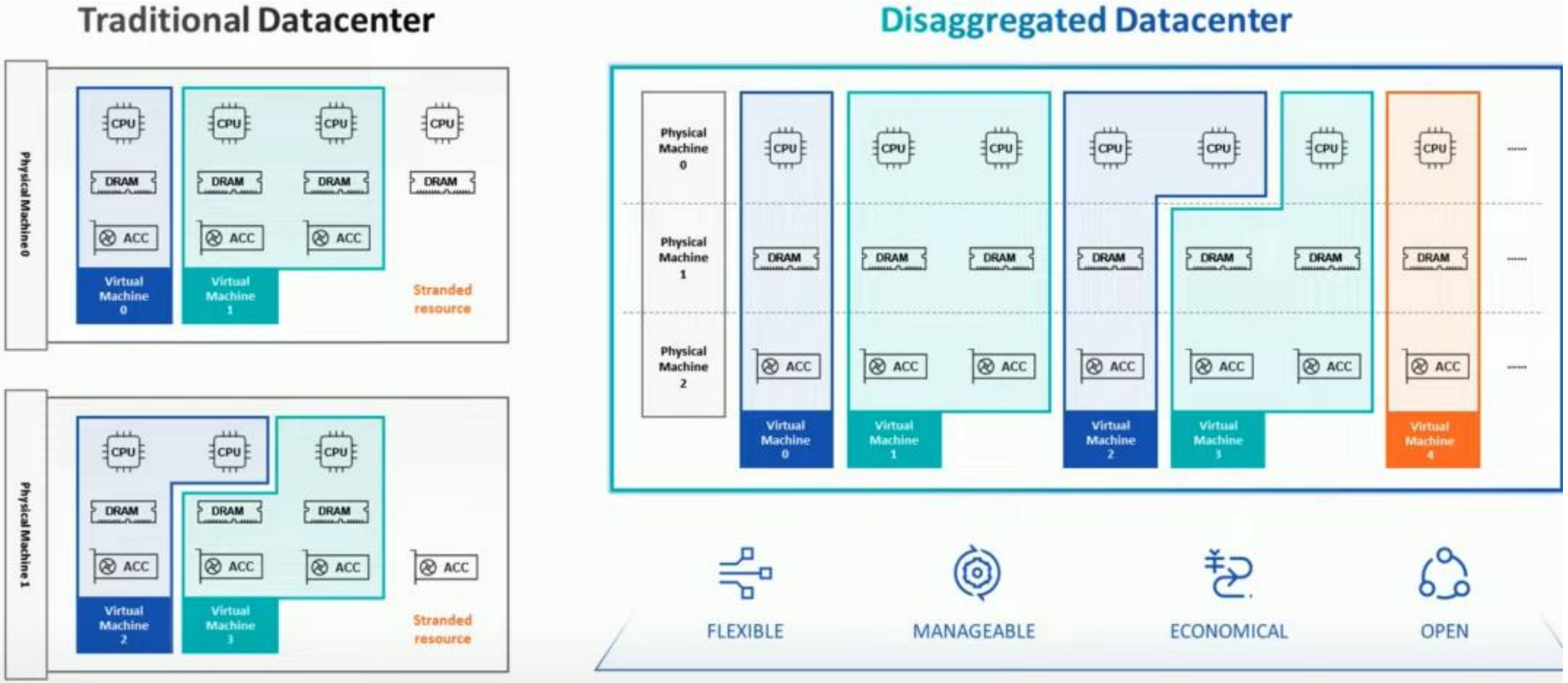


Disaggregation

- In disaggregated datacenter,
  - Compute (e.g., CPUs and GPUs),
  - Memory (e.g., DRAM),
  - Storage (e.g., SSDs and NVMs)
- are **physically separated** from each other, managed in independent resource pools

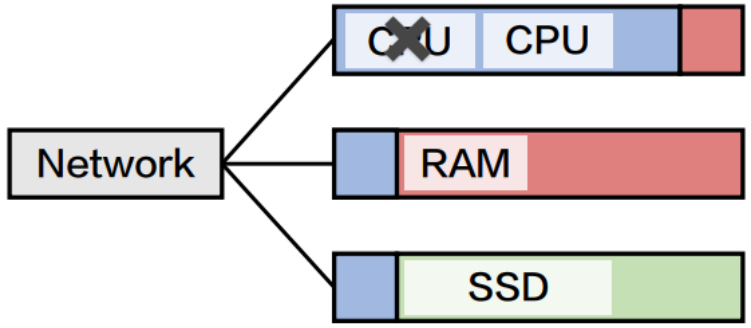
- Compute nodes (Hosts), memory nodes (CXL devices)

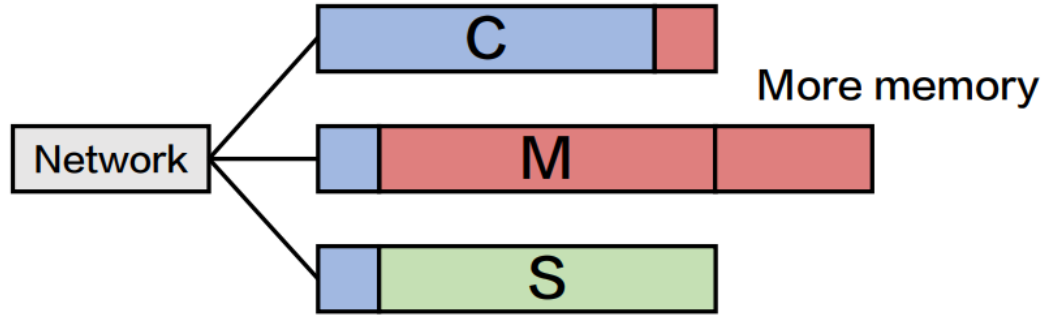
Disaggregation is the future of datacenter

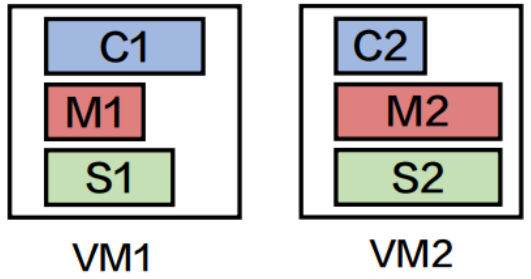


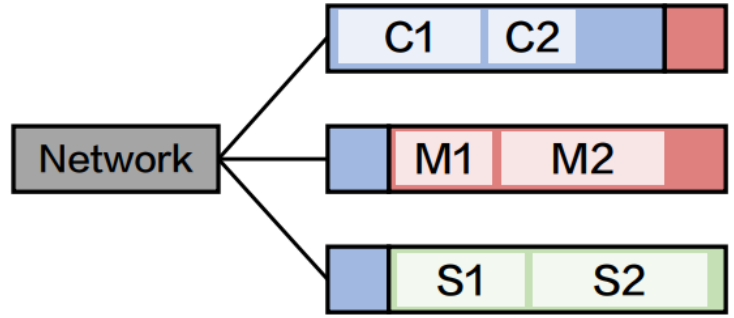
Operational Benefits of Disaggregation

- Independent failures
- Independent expansion
- Independent allocation





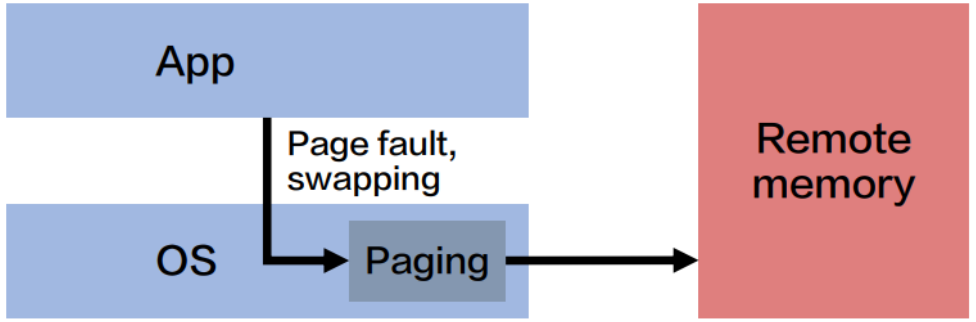




Types of Memory Disaggregation

- Kernel-space approaches

Page-based



Pros
- Unmodified applications
- Transparent infra evolution

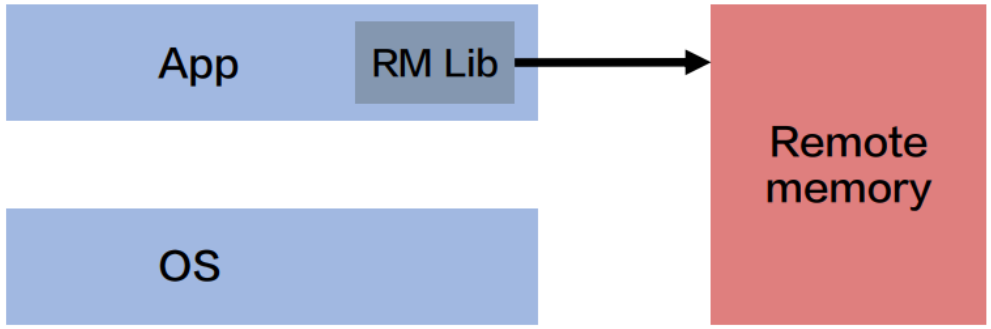
Cons
- High performance cost
- High development cost

Page Fault

Software runtimes have been proposed to enable applications to transparently, without code changes, use remote memory.

- User-space approaches

Object-based



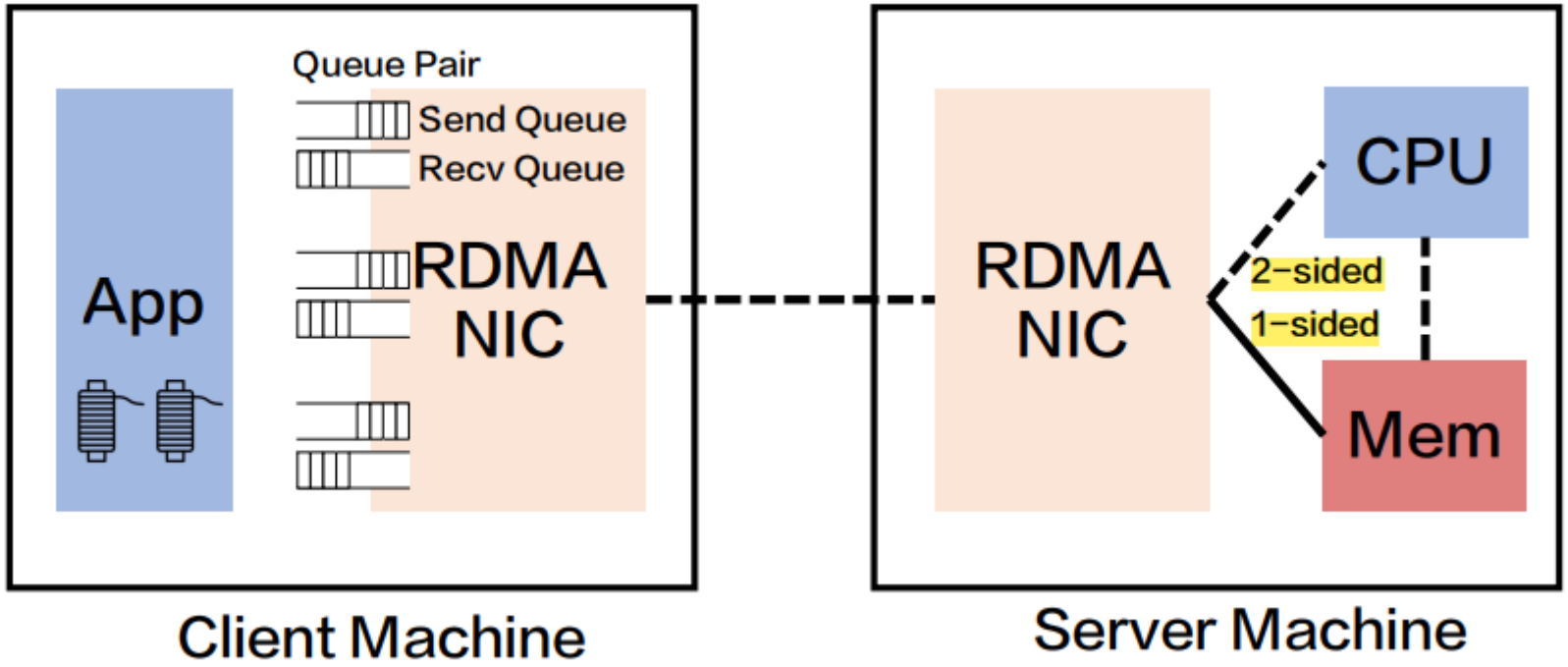
Pros
- No kernel overhead
- Fine-grained control
- Customized optimizations

Cons
- Application modifications

RDMA

- Remote Direct Memory Access
  - Low CPU utilization
  - High network speed

1-sided RDMA: RDMA operations are one-sided, since an RDMA operation can complete without any knowledge of the remote process.



RDMA

- Existing memory disaggregation approaches reply on RDMA
  - Host and multiple memory node
  - RDMA NIC
  - Memory regions (MRs)
  - Memory translation table (MTT)

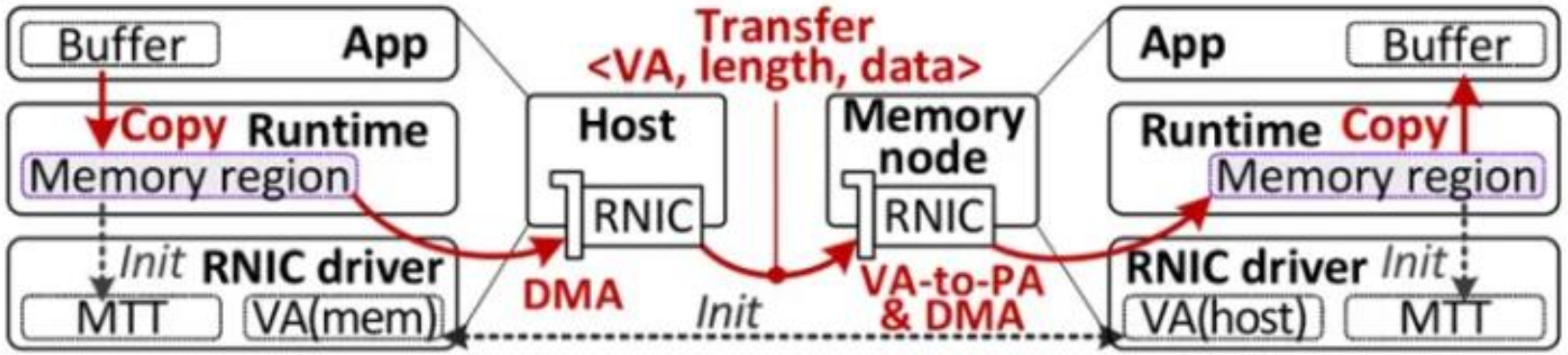


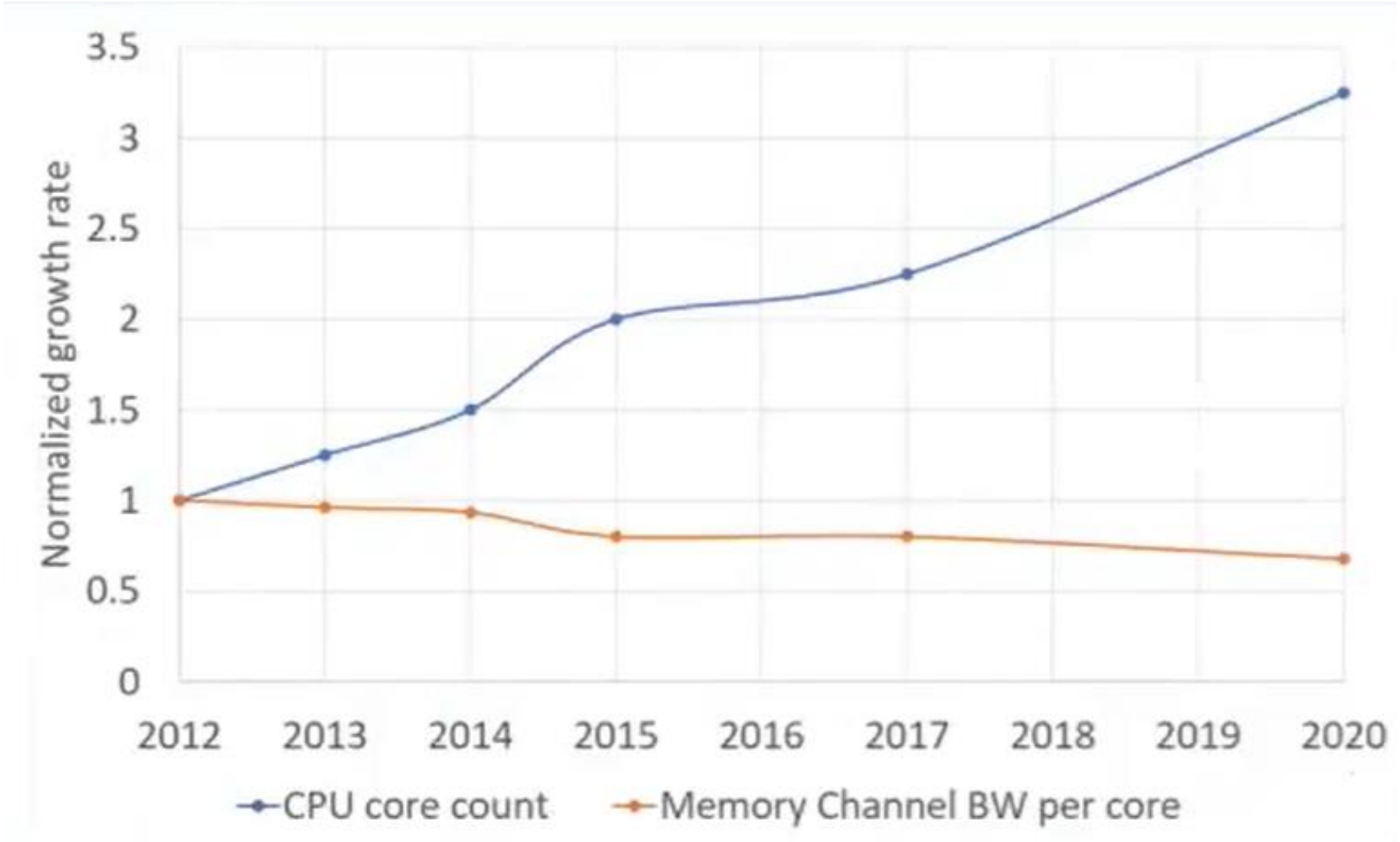
Figure 1: Data movement over RDMA.

Compute Express Link®: An open industry-standard interconnect enabling heterogeneous data-centric computing

Debendra Das Sharma

Intel Senior Fellow, Data Platforms and AI Group, Intel Corporation, Santa Clara, CA 95052, USA
debendra.das.sharma@intel.com

Background of CXL



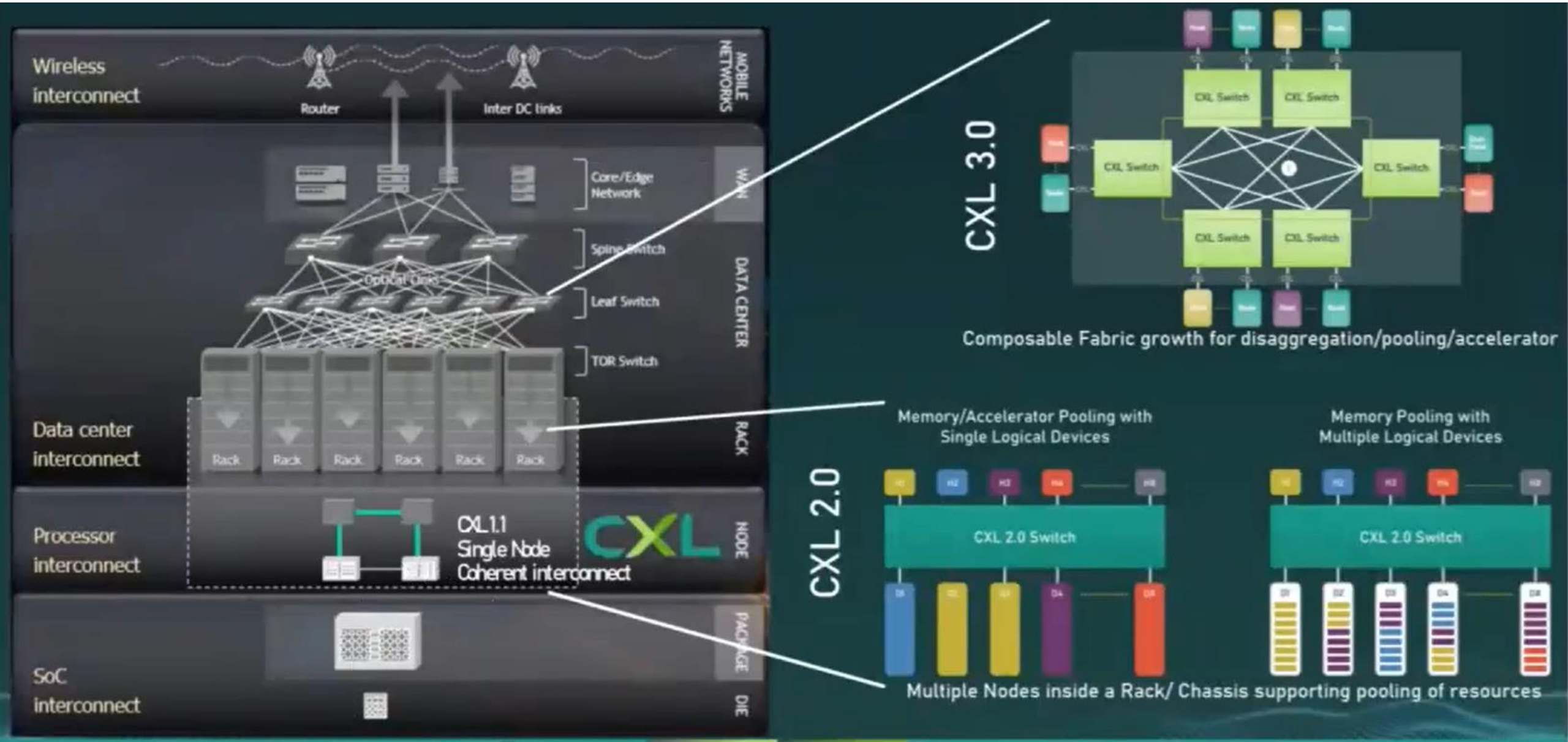
Access to memory matters

Compute Express Link

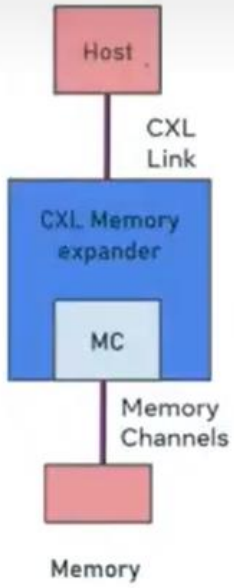
- Intel 2019
- 解决异构（heterogeneous）设备的缓存和内存访问一致性（coherency）的问题
- Based on PCIe
- Device: GPU, FPGA, Smart NIC, etc.
- Host: CPU 所在的主机

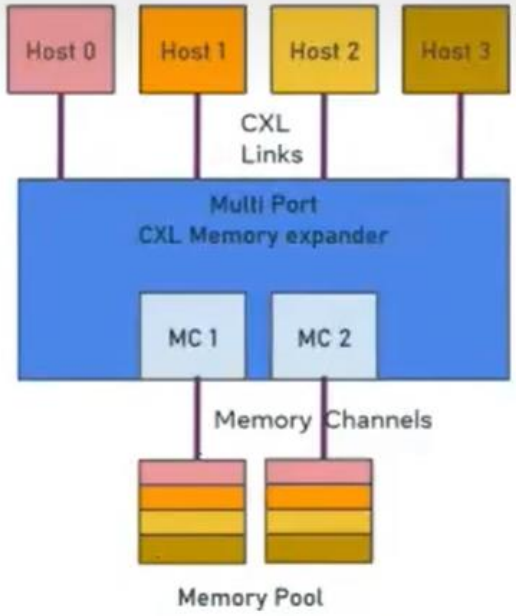
- Enable the communication between
  - memory of devices
  - memory of the server
  - cache of the server’s CPU

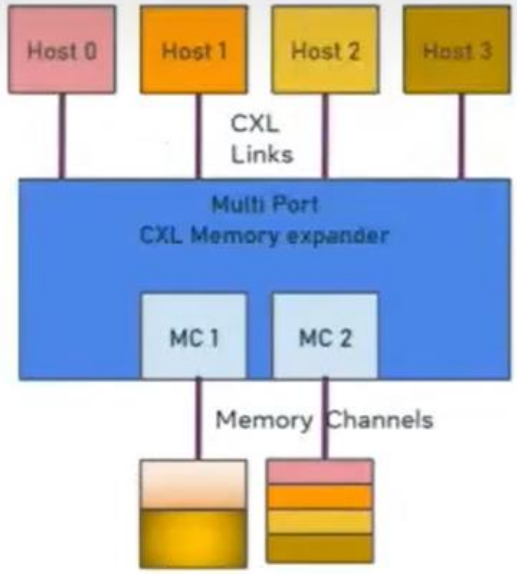
Datacenter: Expanding Scope of CXL

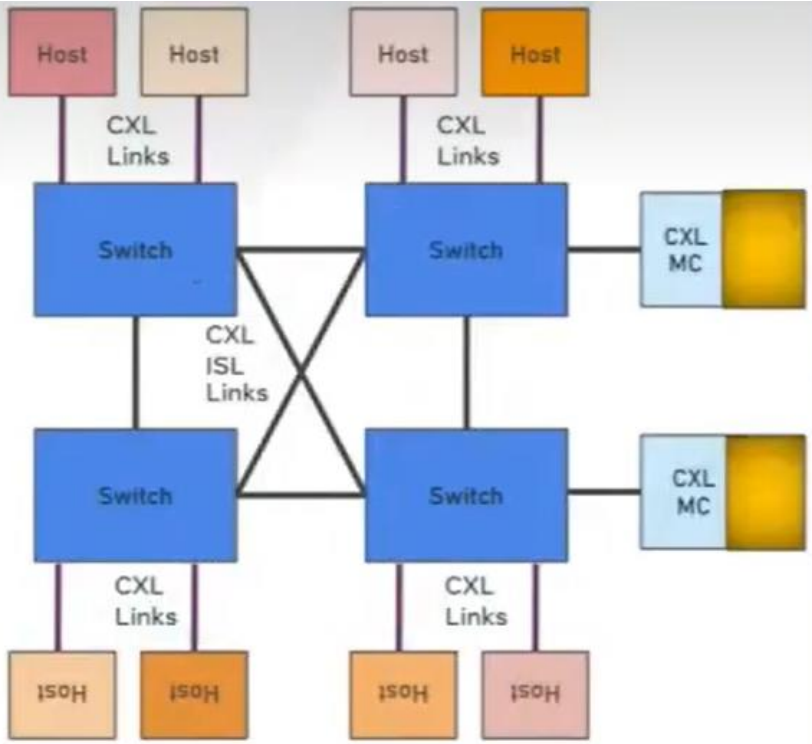


Scenarios of CXL 1.0, 2.0, 3.0









Direct attached

Add Capacity
Add Bandwidth
Slower-cheaper tier

Pooled Memory

Amortize CXL infra cost
Flexible allocation

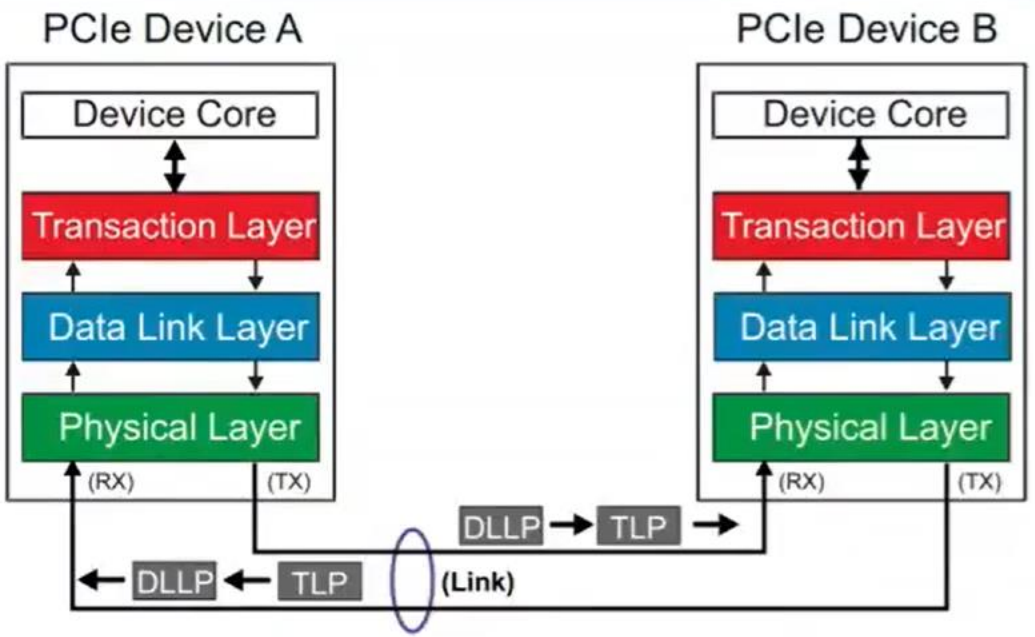
Shared Memory

Dedup lication
Host2host communication
large datasets

Fabric Memory

Scaling to huge datasets

PCIe 协议



Transaction Layer Packet (TLP)

TLP Types:
- Memory Read / Write
- IO Read / Write
- Configuration Read / Write
- Completion
- Message
- AtomicOp

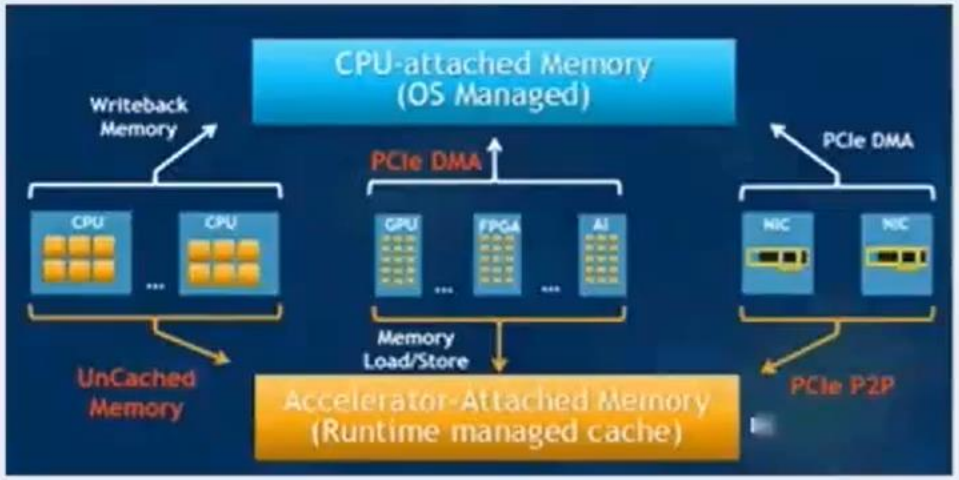
Data Link Layer Packet (DLLP)

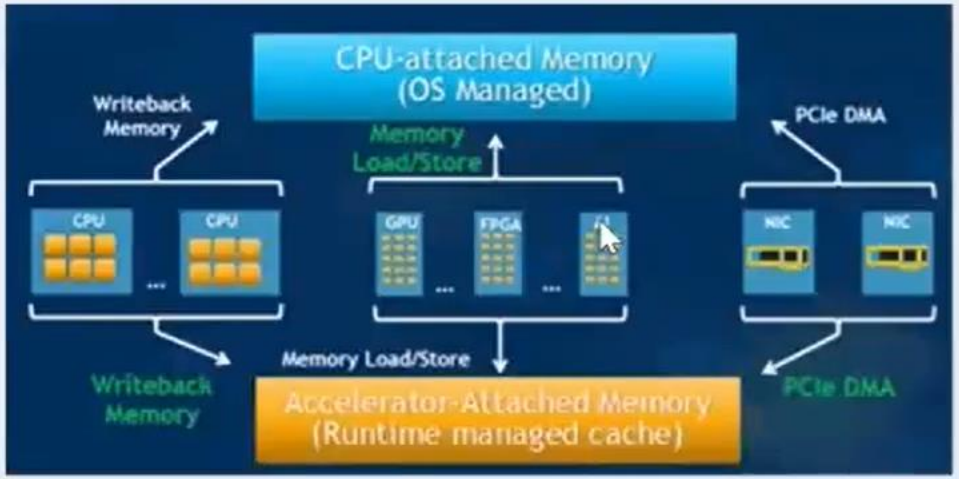
DLLP Types:
- TLP Ack/Nak
- Power Management
- Link Flow Control
- Vendor-Specific

信息以包的形式在PCIe设备之间传递：TLP DLLP PLP

TLP 用于承载PCIe事务，由发送设备根据Device Core提供的信息在事务层生成，穿过RC或Switch，到达最终的目标设备的事务层后才得以处理。

CXL协议缓存一致性





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。

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：CXL.io，CXL. cache以及CXL.mem，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。

CXL.cache可以提升效率的根本原因就是，原来需要DMA+软件才能完成的事儿，使用硬件加速了，它的本质也是硬件卸载加速。

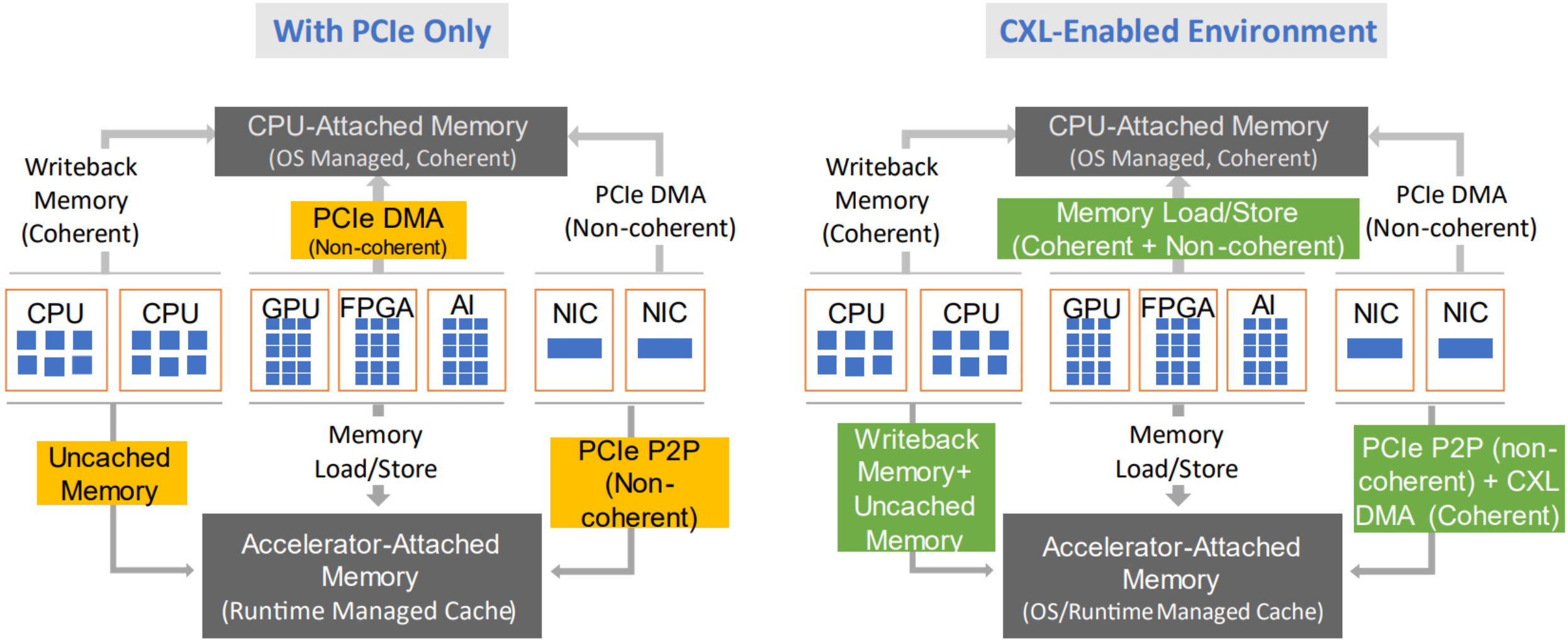
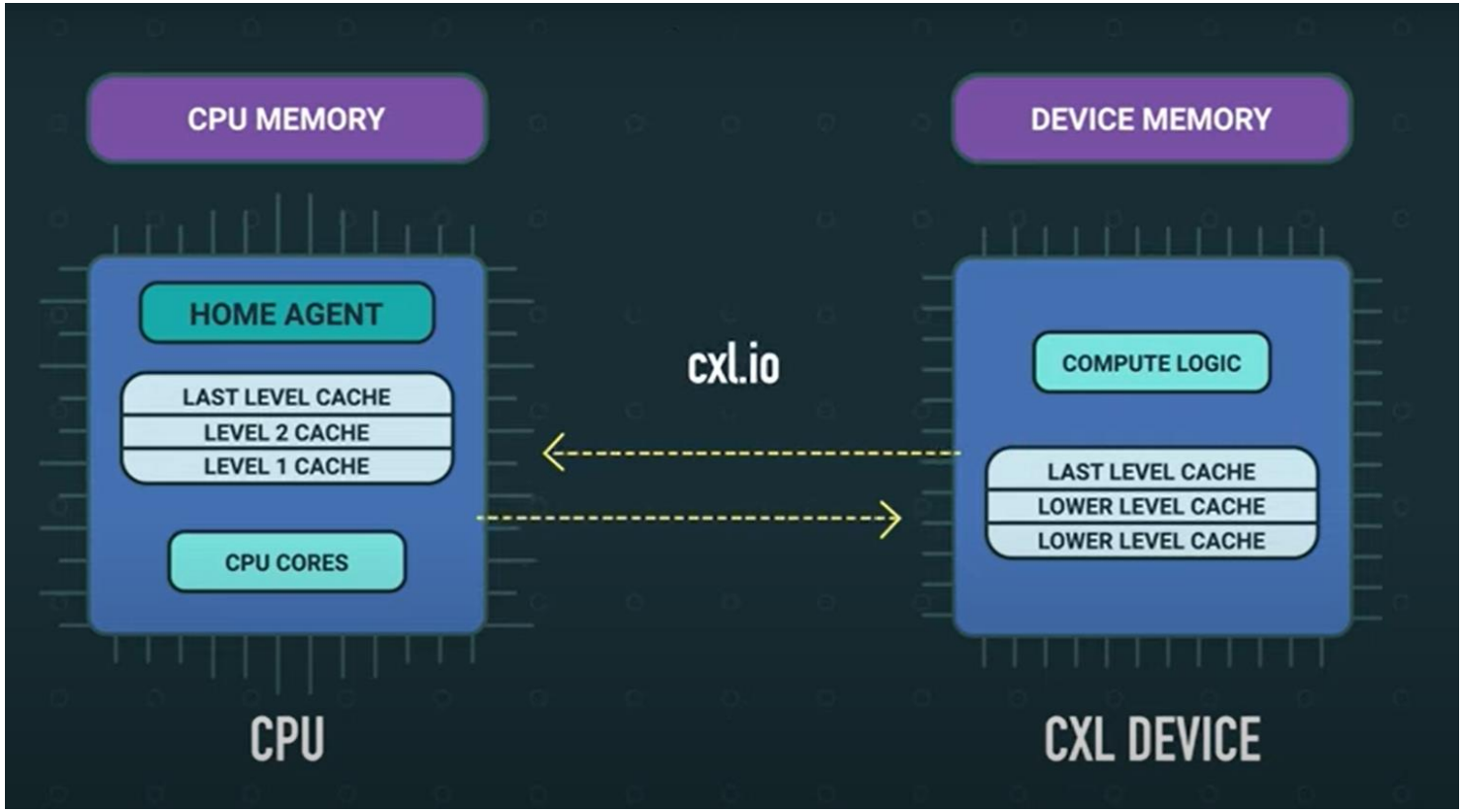


Figure 1: CXL enables coherency and memory semantics and builds on top of PCIe' s physical subsystem.

CXL Protocols & Standards

- The CXL standard supports a variety of use cases via 3 protocols
- **CXL.io**
- **CXL.cache**
- **CXL.memory**

cxl.io protocol

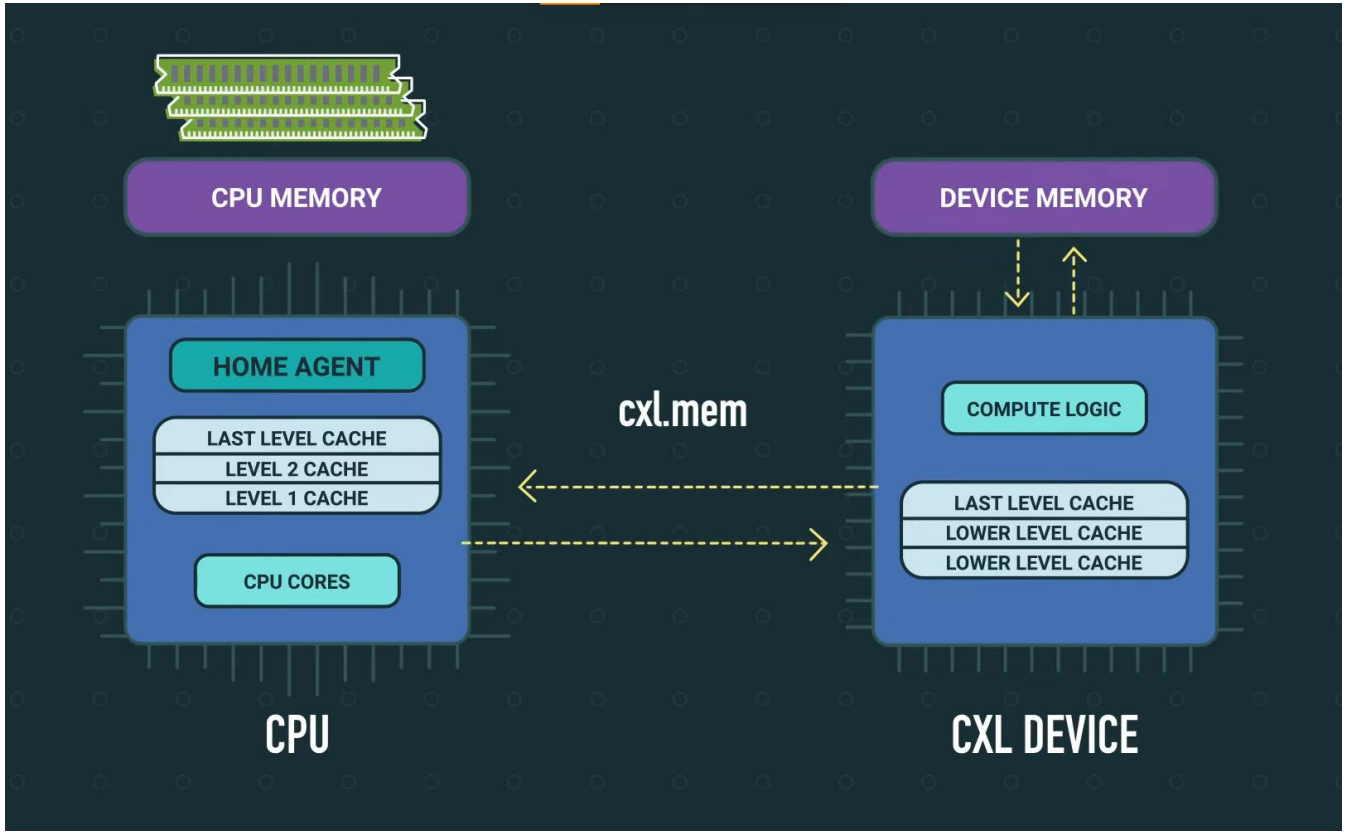


Discovery, configuration, interrupts, register access, enumeration, etc.

Similar to PCIe

This allows the home agent to view the accelerator memory in the same way it views its own memory

cxl.mem protocol



The home agent can access the device (accelerator) memory w/o software interventions.
The accelerator memory looks like DDR attached to the application, resulting in lower latency.

cxl.cache protocol

- **cxl.mem** enables processor to access device attached memory
- **cxl.cache** enables device to access processor memory

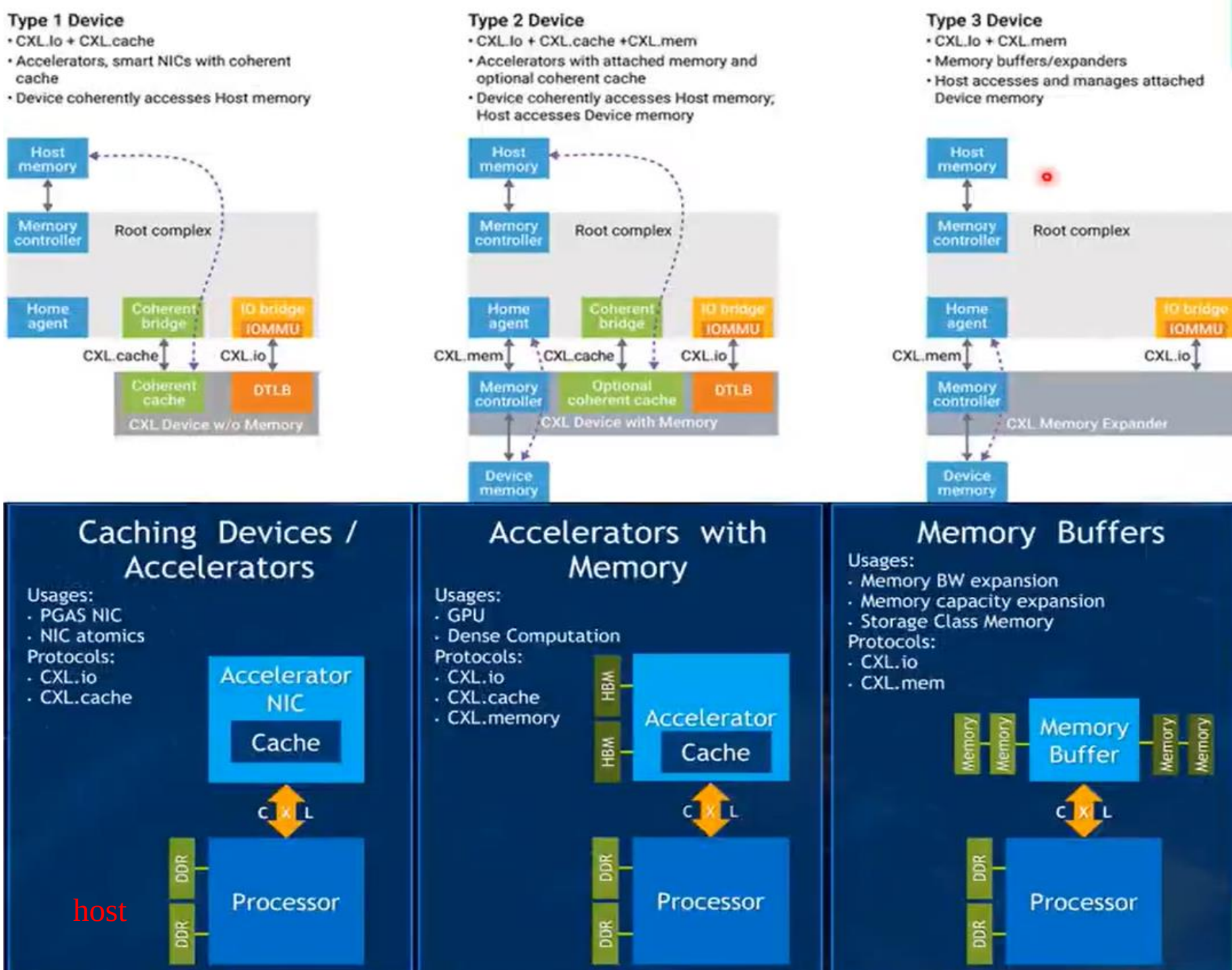
- **cxl.cache** establishes coherency and fast communication between host processor and CXL devices.

Three Types of CXL Devices

- Cache-coherent interconnects for connectivity between CPUs, accelerators, and I/O devices

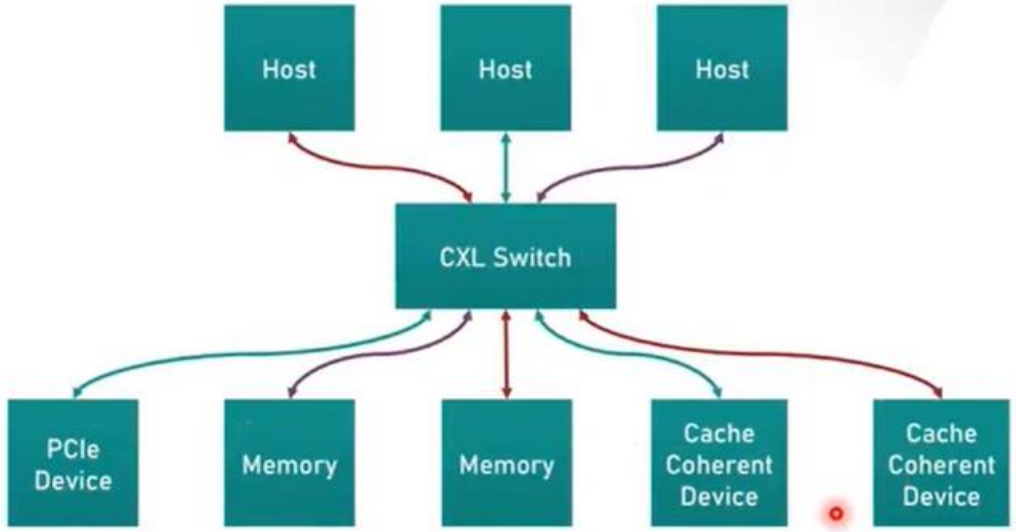
- Supports all devices, from accelerators to memory
  - Type 1: device accessing host memory
  - Type 2: device and host accessing each other's memory
  - Type 3: host accessing device memory

1. Smart NIC, …
2. GPU, FPGA, …
3. Memory, …

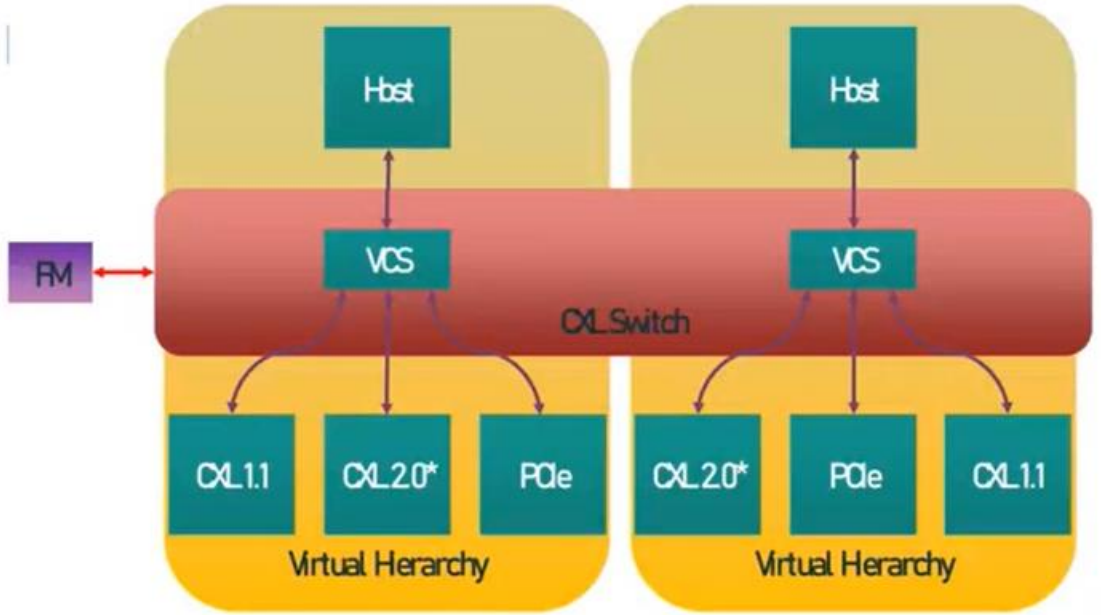


CXL 2.0

CXL交换器：在一层内实现多个Host和多个异构Device的互联



Heterogeneous devices



Switch可以实现不同Type CXL互联，同时支持PCIe协议的设备

Switch内可以划分VCS (Virtual CXL Switch)实现细粒度资源分配和隔离

Switch可以进行结构管理(FM)，实现每个交换机的中心化管理

Direct Access, High-Performance Memory Disaggregation with DIRECTCXL

Donghyun Gouk, Sangwon Lee, Miryeong Kwon, Myoungsoo Jung
Computer Architecture and Memory Systems Laboratory,
Korea Advanced Institute of Science and Technology (KAIST)
http://camelab.org

DirectCXL

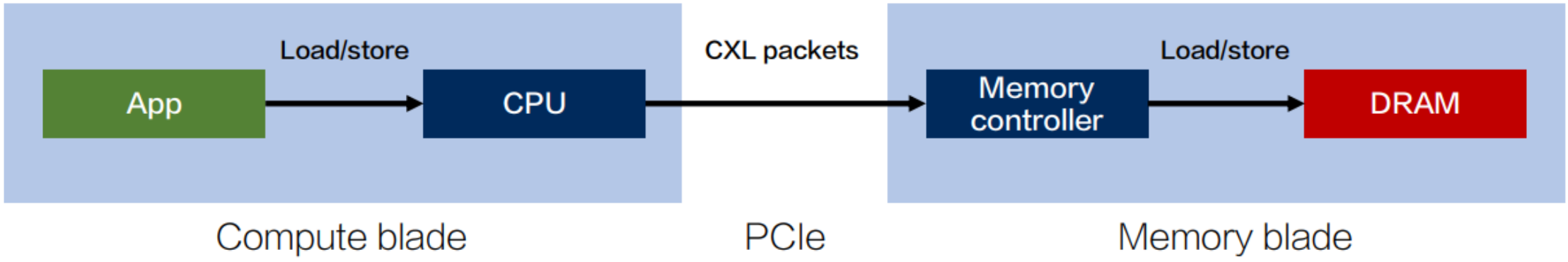
- An alternative approach to disaggregating memory using CXL
- Motivation: RDMA Cost
  - Data is copied over the network
    - Network latency
    - DMA operations on both sides
  - Data is copied between applications and NIC-registered memory region

Compared to RDMA

- Direct PCIe access through load/store instructions
  - No network latency
  - No extra data copies

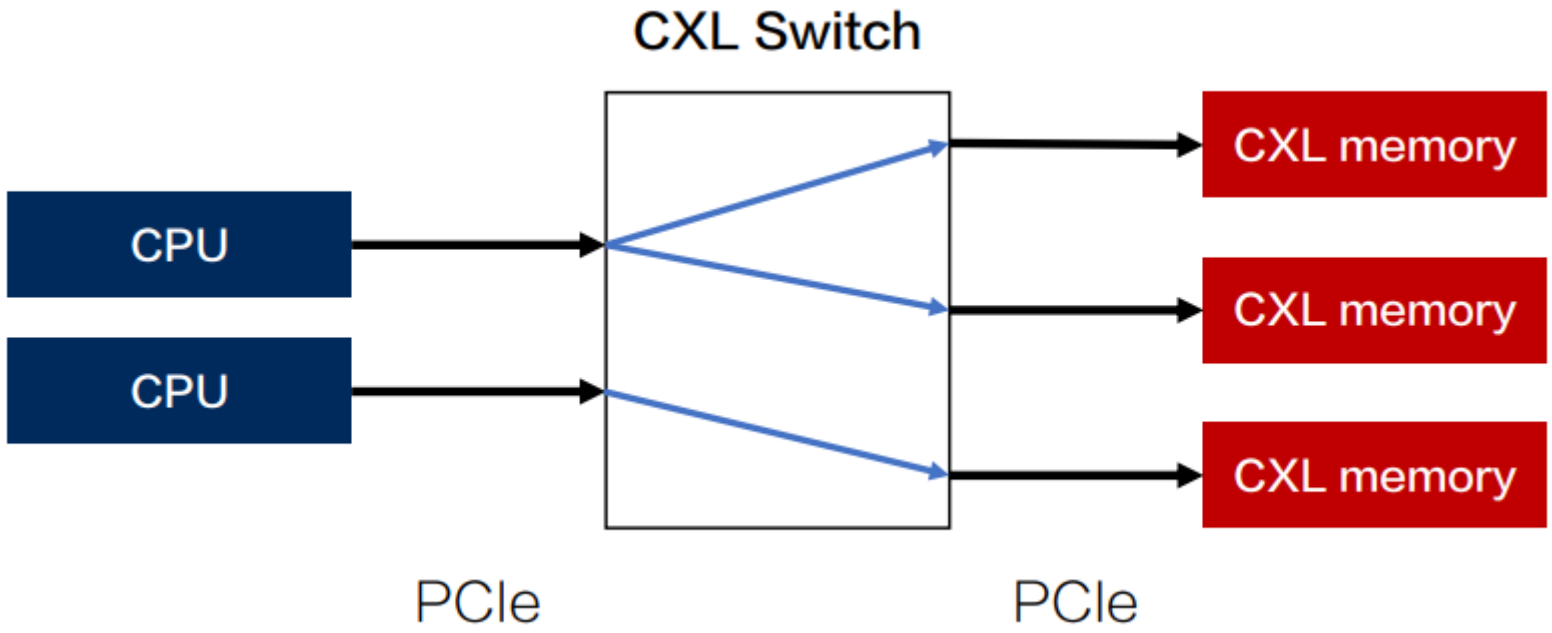
DirectCXL Design

- How to enable direct access to CXL memory?
  - Convert load/store instructions to CXL packets
  - An FPGA-based controller converts them back



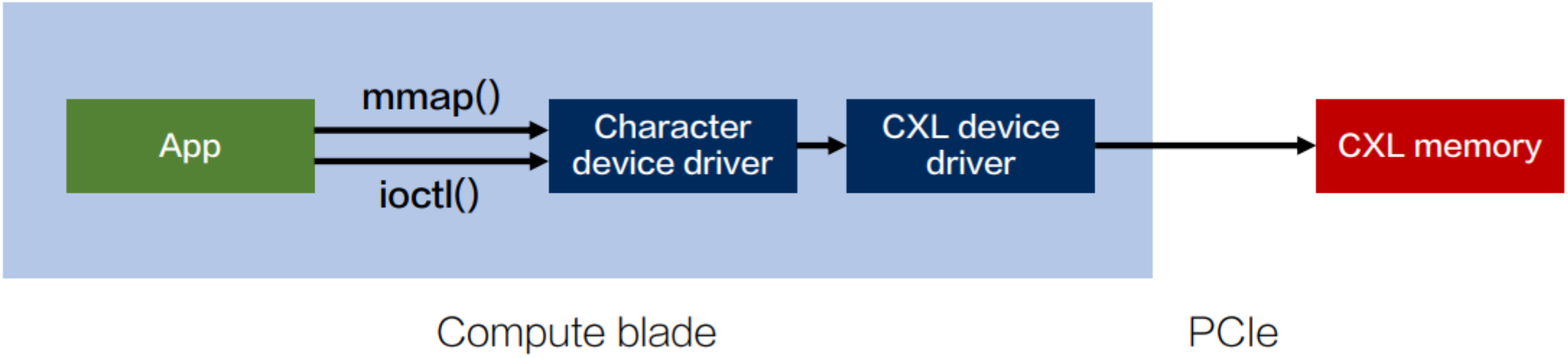
DirectCXL Design

- How to enable flexible memory configuration?
  - A CXL switch with a reconfigurable crossbar



DirectCXL Design

- How to present CXL memory to applications?
  - Leveraging Linux virtual memory system



DirectCXL Design

- RDMA-based memory disaggregation incurs **networking overhead** and **extra memory copies**
- DirectCXL provides a CXL solution via direct PCIe access, a CXL switch, and a software runtime
- Application performance is significantly improved **without modifications**

- 结束