

My CXL Pool Obviates Your PCIe Switch

Yuhong Zhong 

Daniel S. Berger  

Pantea Zardoshti 

Enrique Saurez 

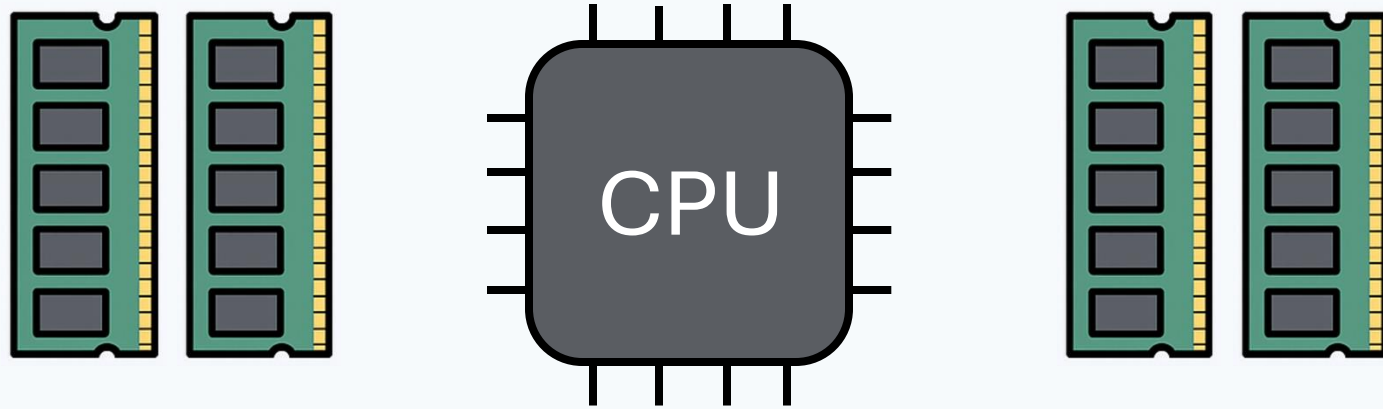
Jacob Nelson 

Antonis Psistakis 

Joshua Fried 

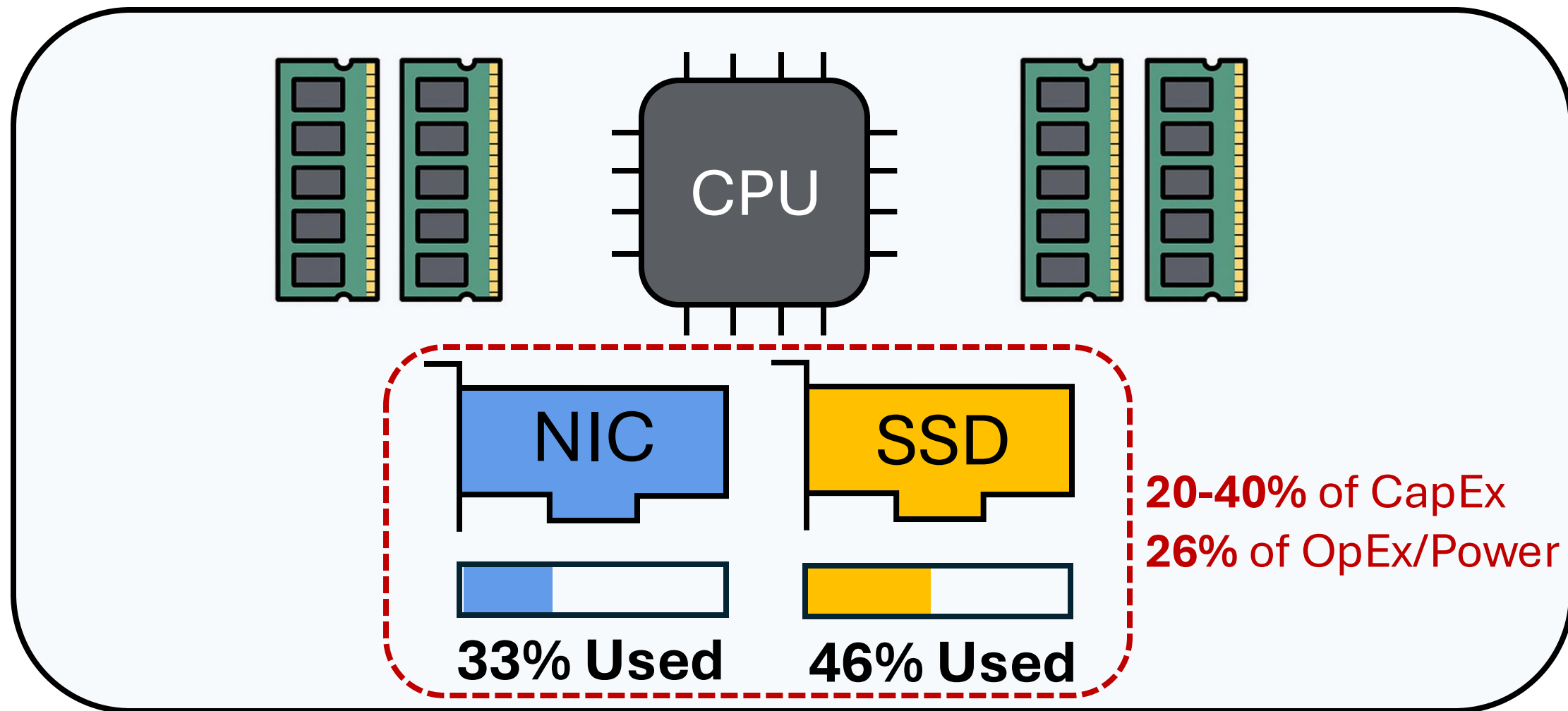
Asaf Cidon 

Lots of Work on Improving CPU and Memory Utilization



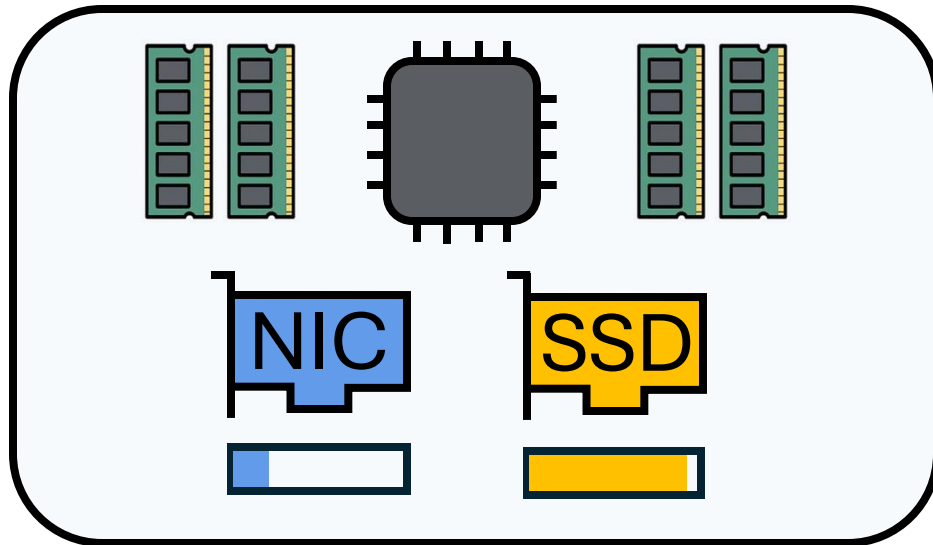
Pond (ASPLOS '23), HarvestVM (OSDI '20), AIFM (OSDI '20),
Shenango (NSDI '19), LegoOS (OSDI '18), InfiniSwap (NSDI '17), ...

Datacenters Are Full of Idle PCIe Devices

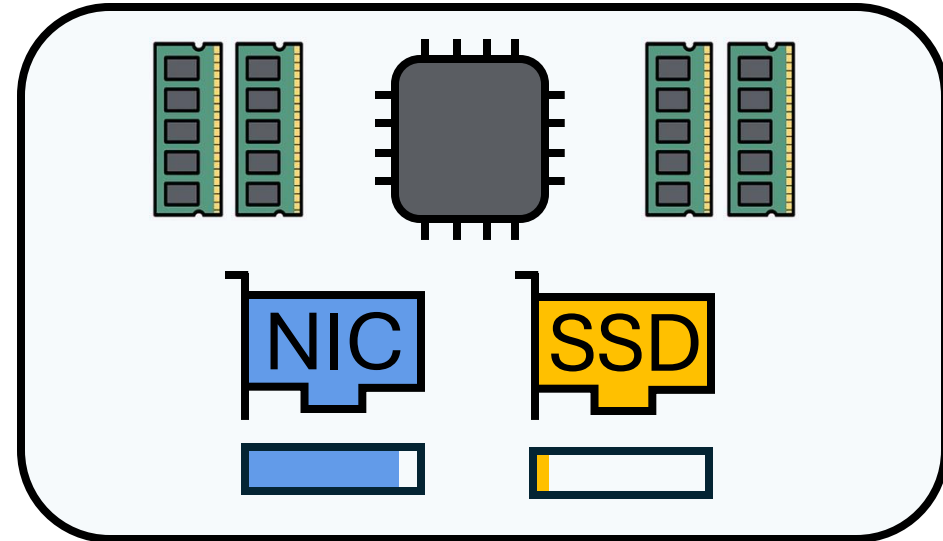


Overprovisioning Causes Low Utilization

Host A



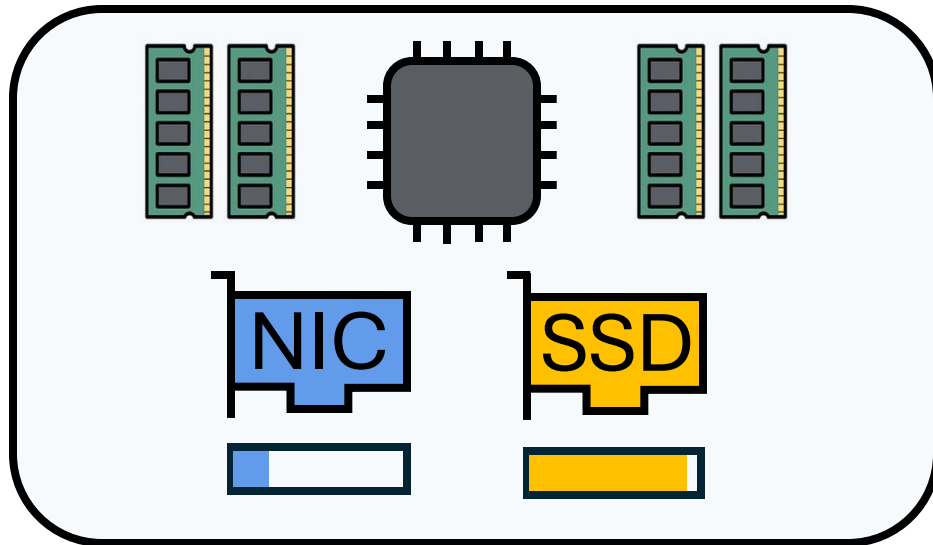
Host B



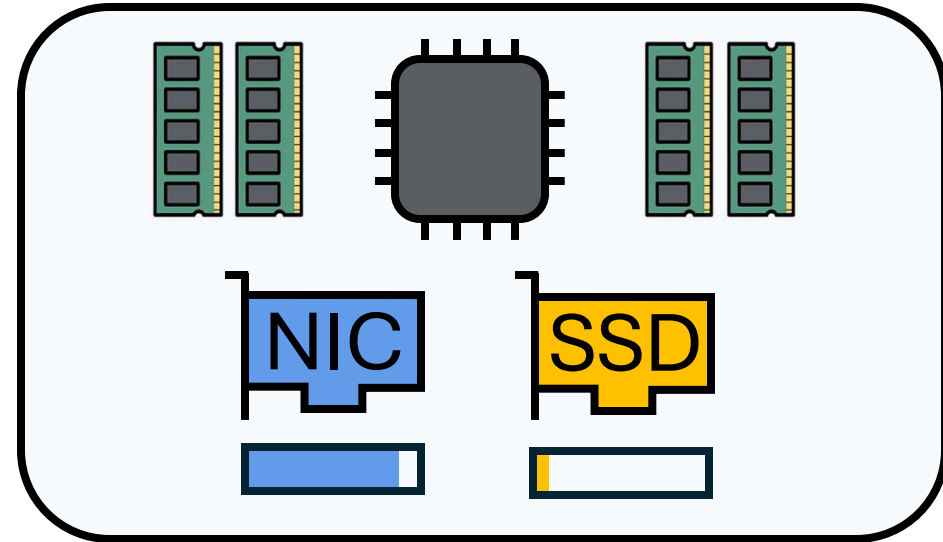
- PCIe resources are overprovisioned for **peak demand per-host**
- Idle resources cannot be used by other hosts
- Redundant devices are provisioned per-host

Pooling PCIe to Boost Utilization

Host A

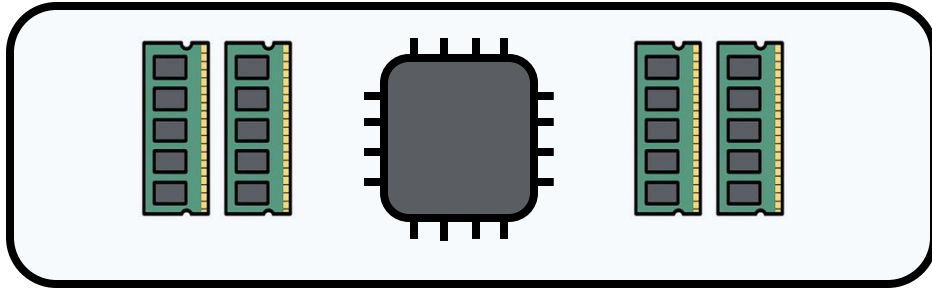


Host B

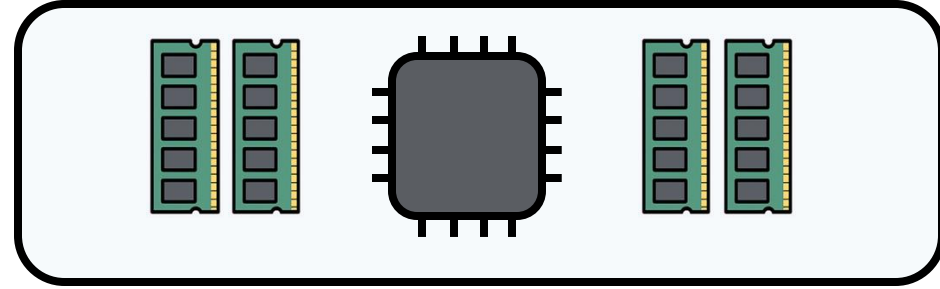


Pooling PCIe to Boost Utilization

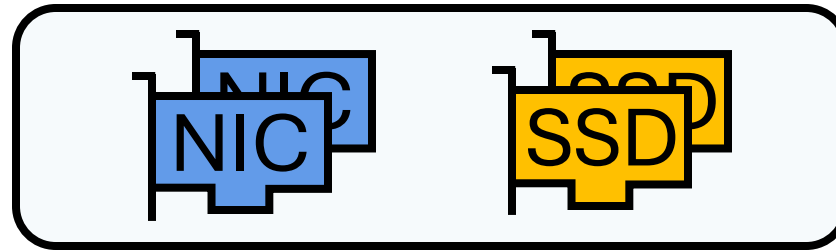
Host A



Host B



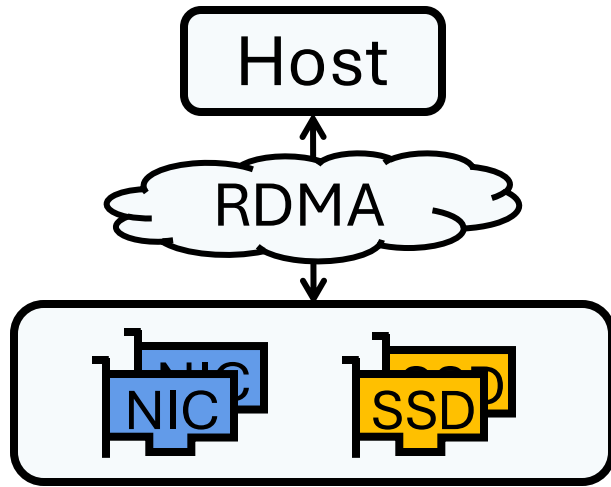
PCIe Device Pool



- Provision for the peak demand **across hosts**
- Idle resources can be used by any hosts
- A single backup device can be shared by many hosts

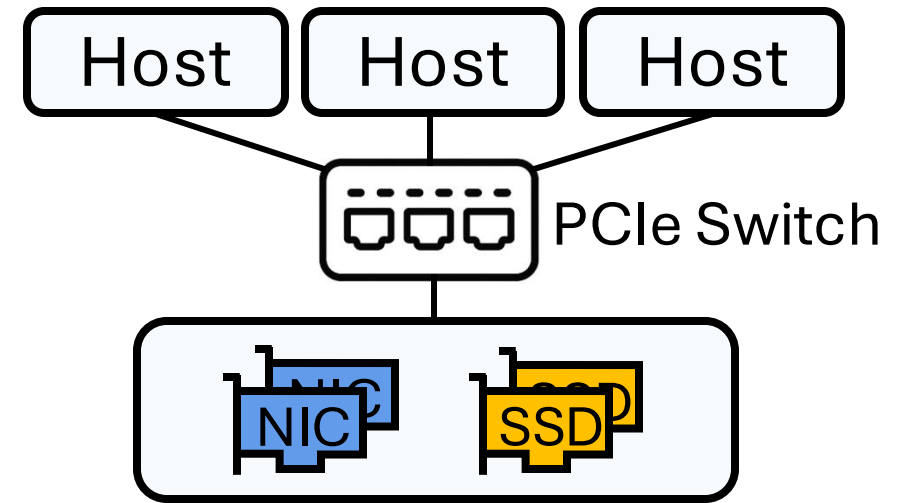
RDMA Is Limited, PCIe Switches Are Expensive

Option 1: **RDMA**



- ✗ Cannot pool NICs
- ✗ High latency overhead and limited IOPS

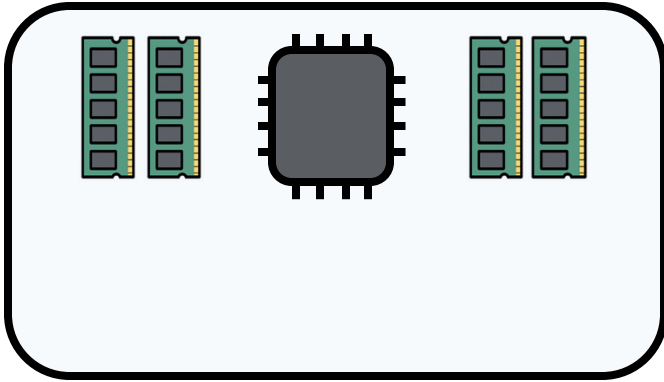
Option 2: **PCIe Switches**



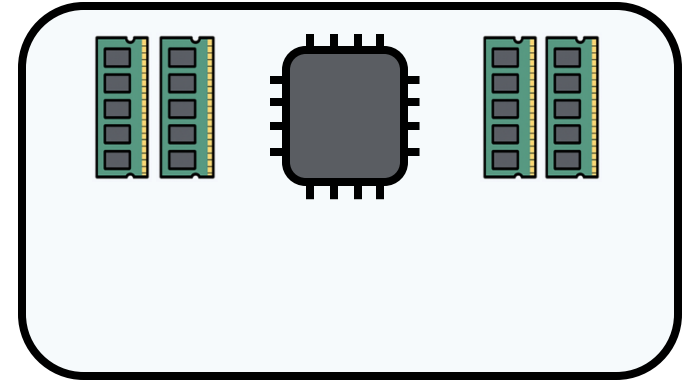
- ✗ Expensive to deploy
(switches + switch software
+ host adapter cards + cabling)
= **\$80,000 per rack**

CXL to Pool Memory

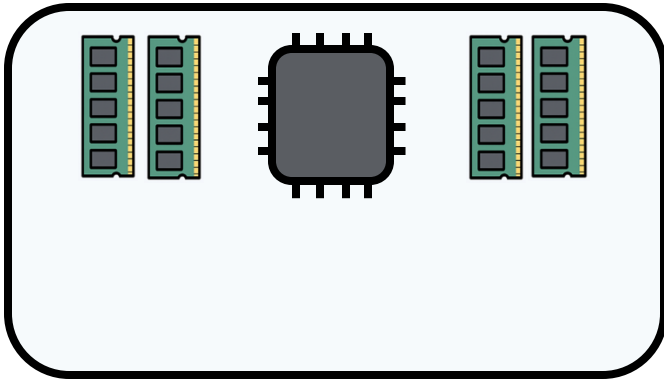
Host A



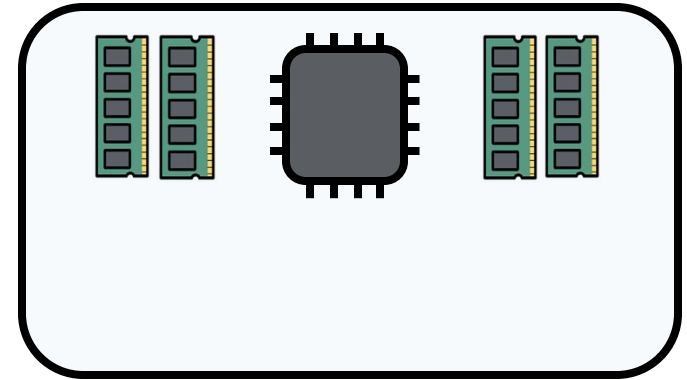
Host B



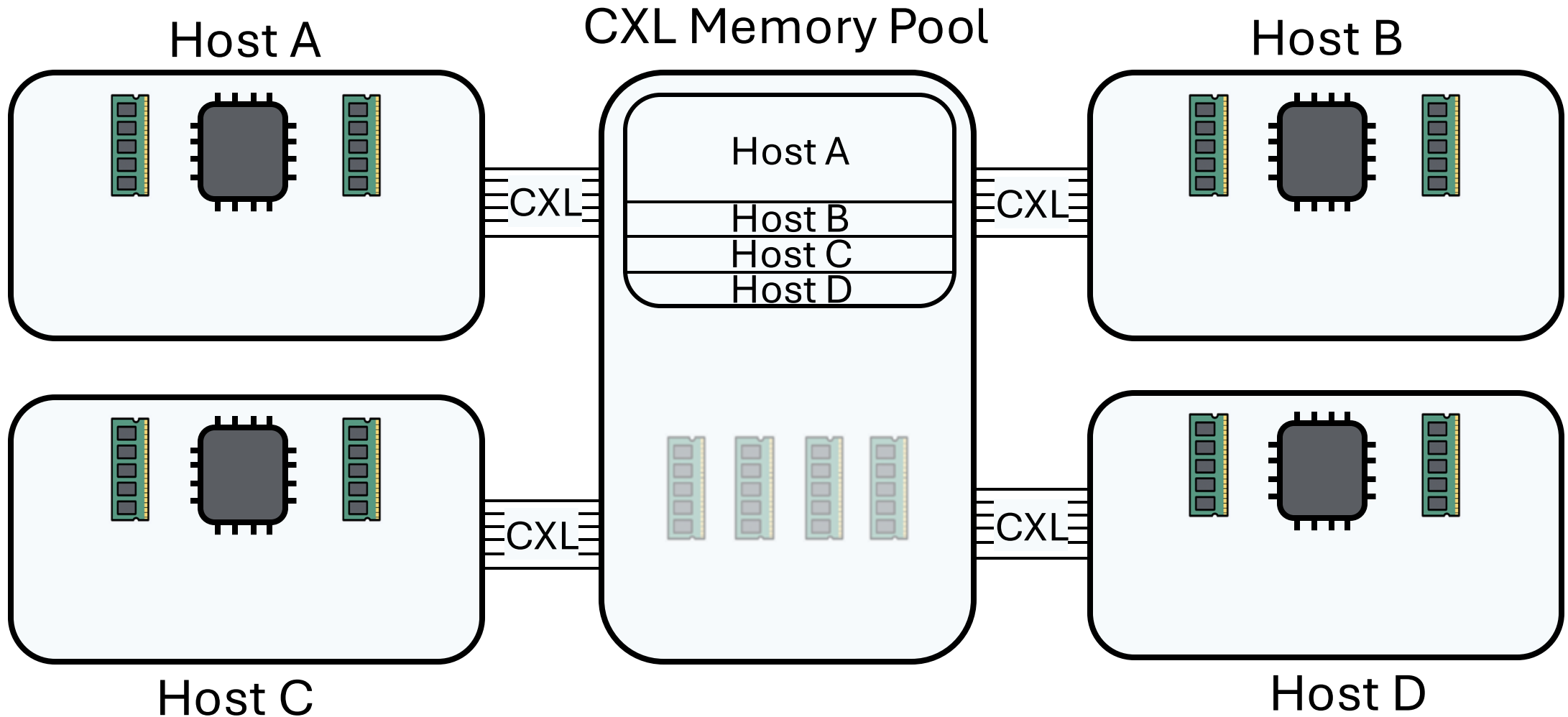
Host C



Host D

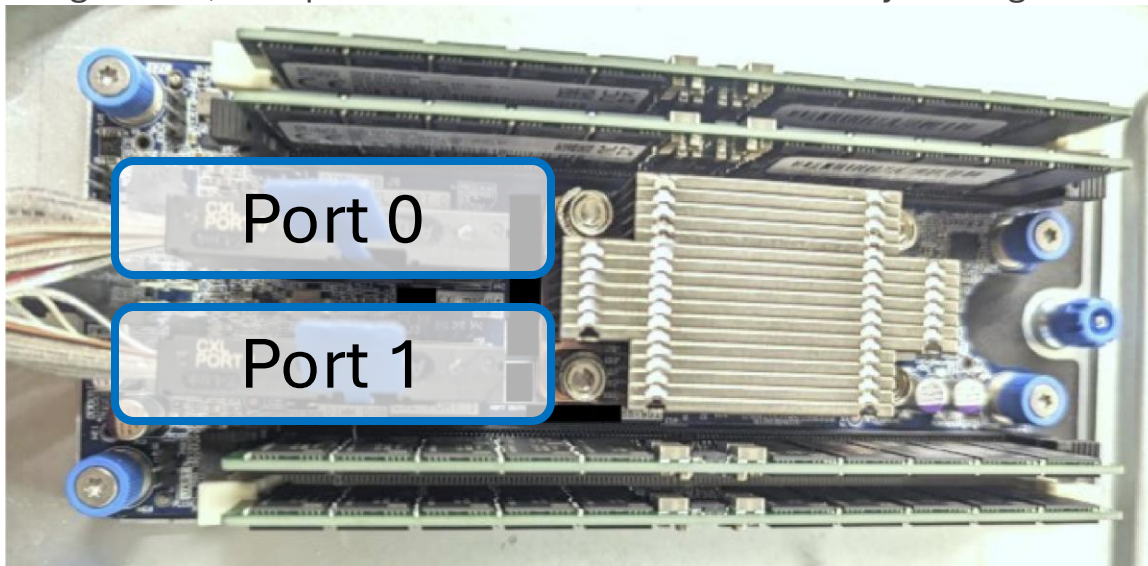


CXL to Pool Memory



Industry Is Proposing to Deploy CXL Pools

Berger et al., Octopus: Scalable Low-Cost CXL Memory Pooling



2 CXL ports

- 2-port CXL devices: Marvell Structera, Astera Labs Leo
- 8-port CXL devices: Seagate FPGA, SKH Niagra
- Expensive CXL switches are **not required!**

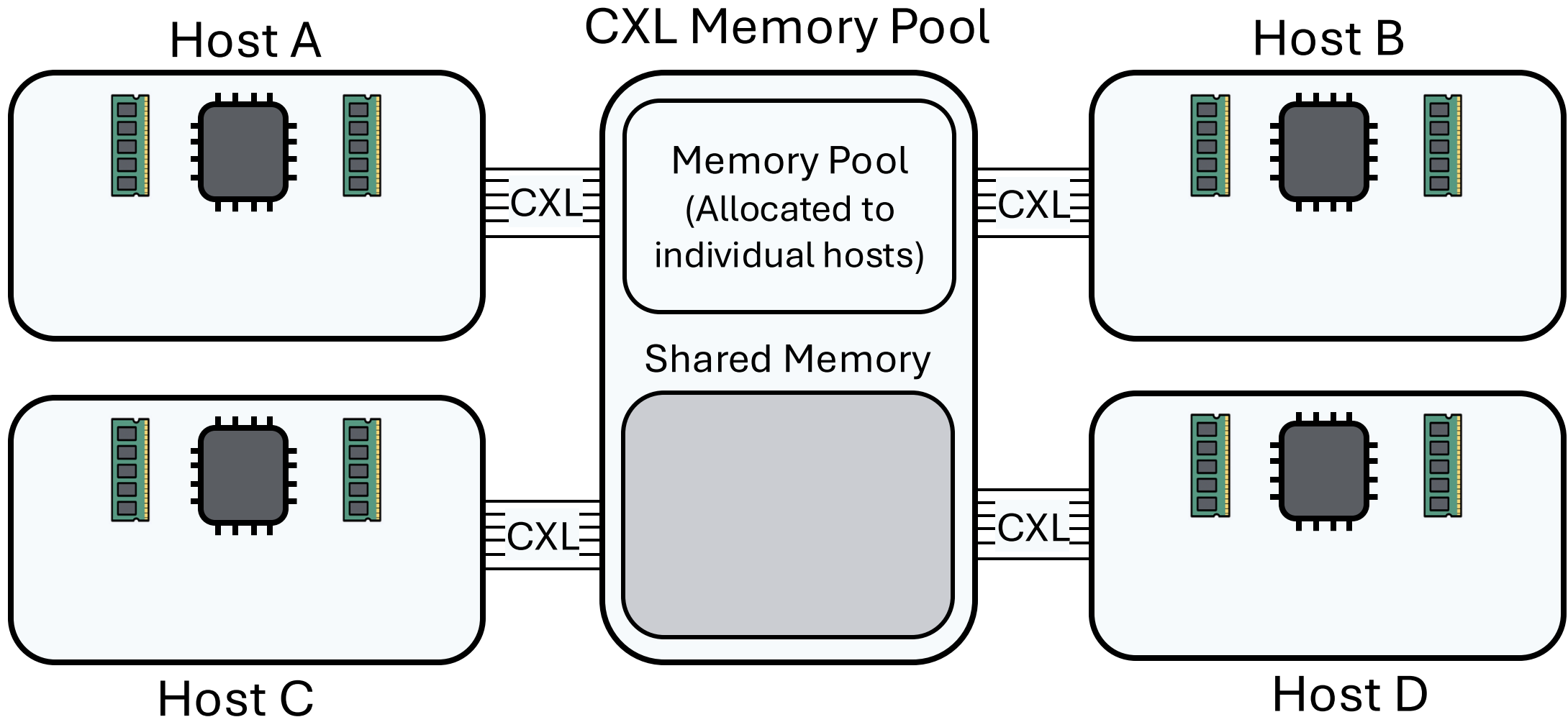
CXL device →



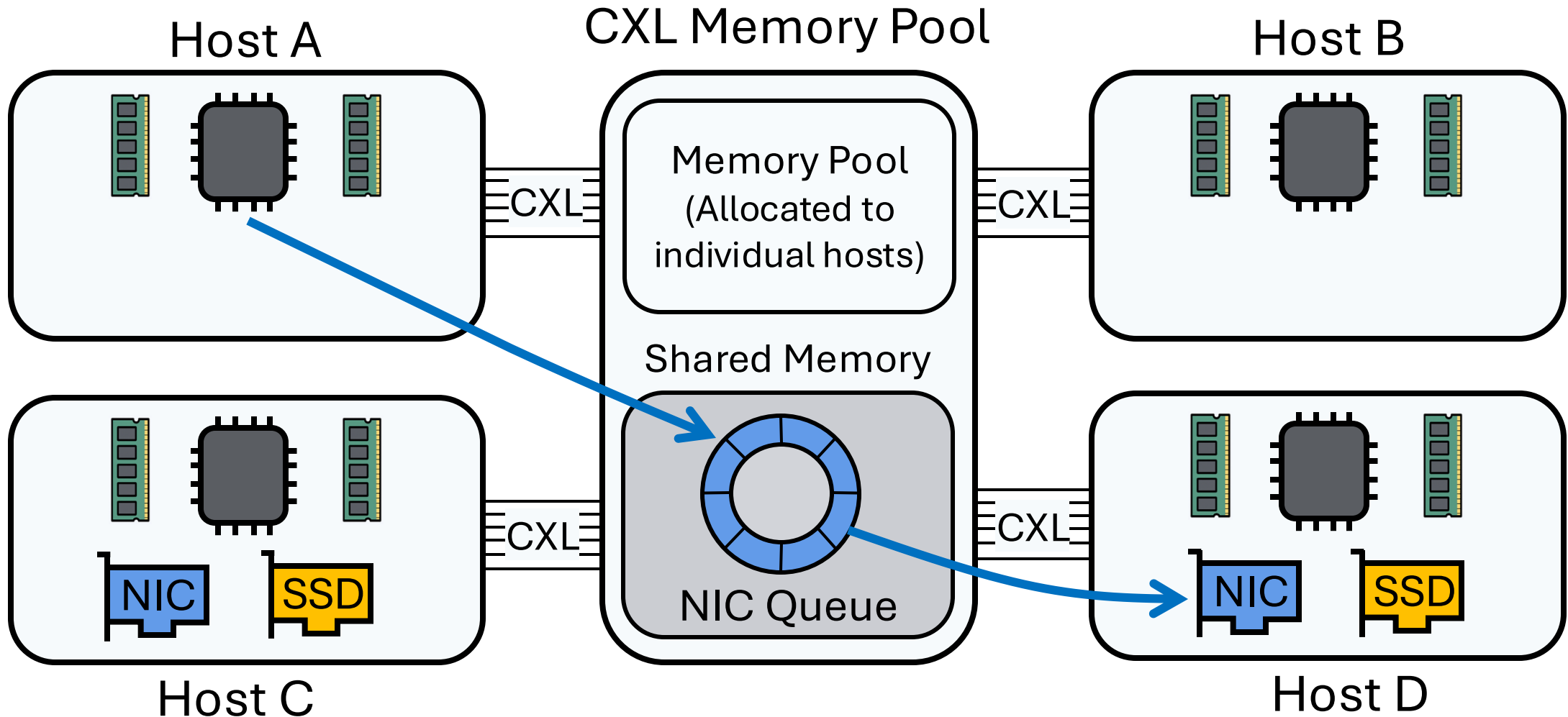
Samsung CMM-B, photo from Xconn Tech

3 CXL ports

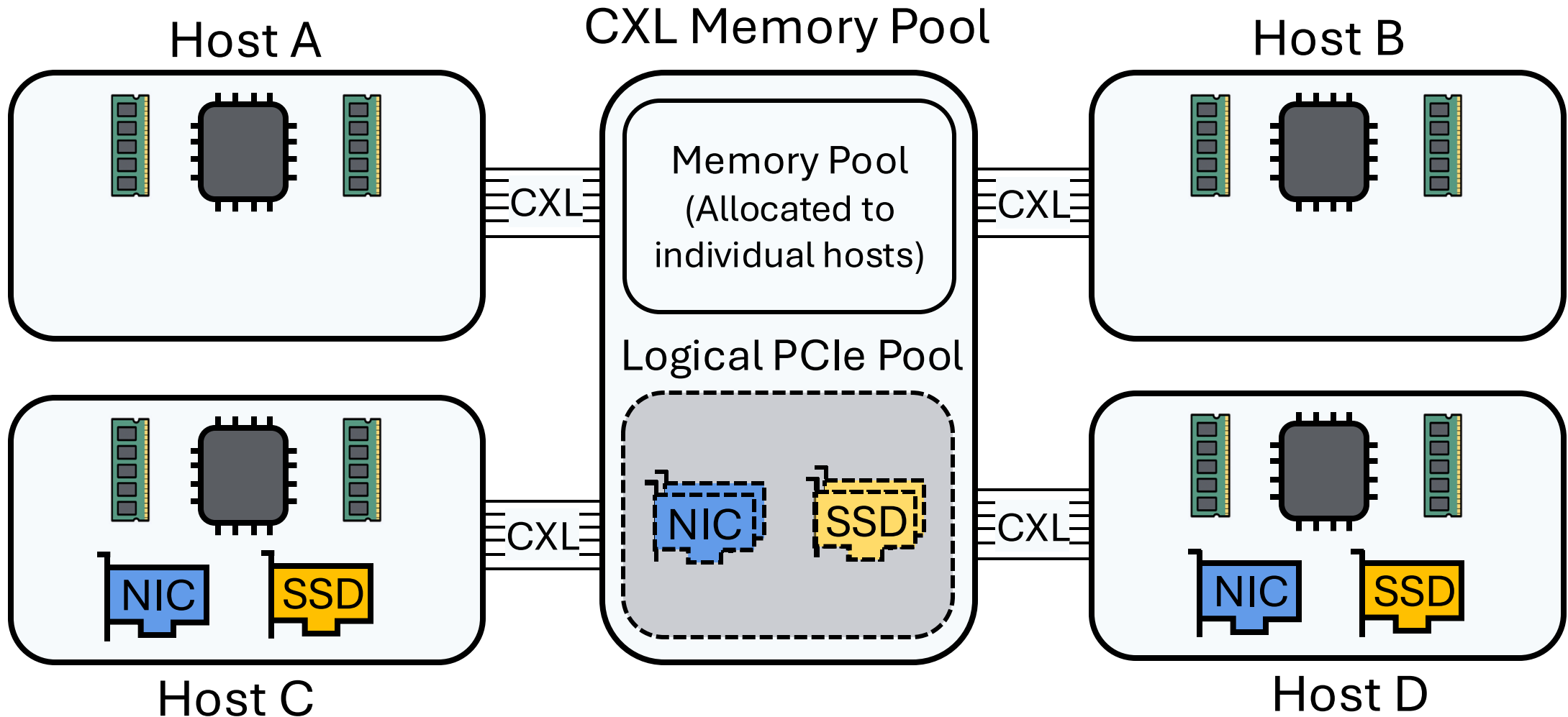
CXL to Pool Memory... and Now PCIe Devices!



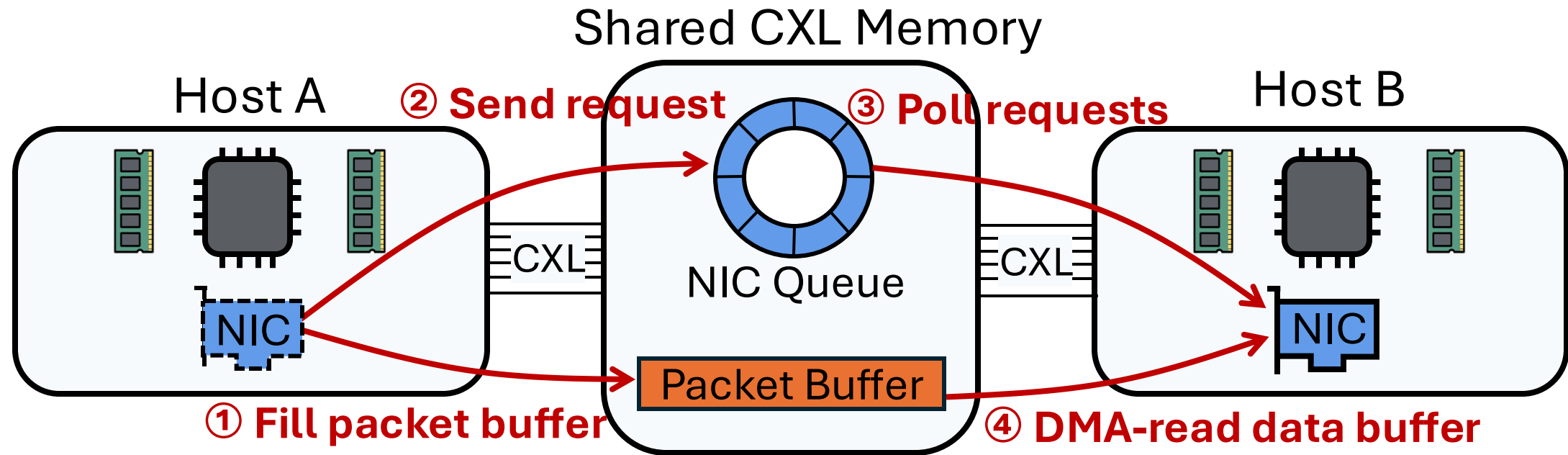
CXL to Pool Memory... and Now PCIe Devices!



CXL to Pool Memory... and Now PCIe Devices!



Example: NIC Datapath



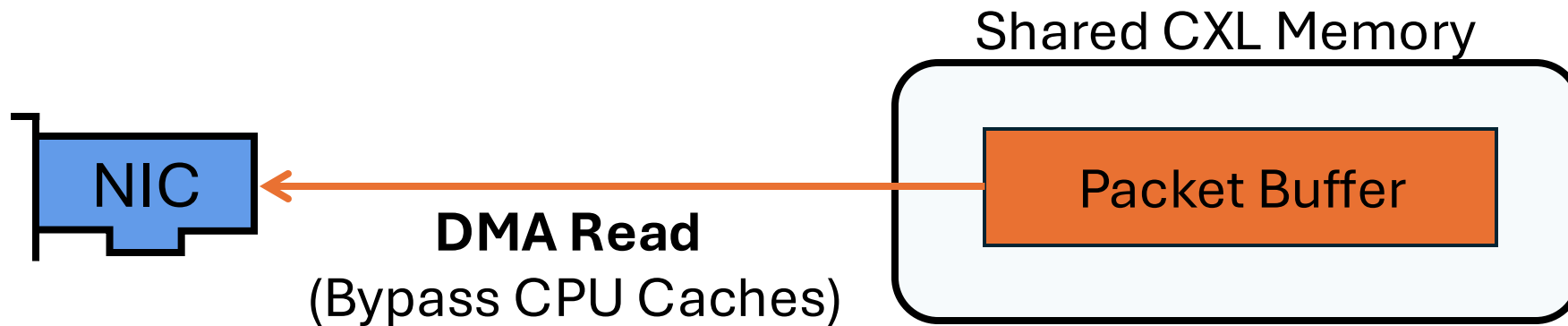
Challenges:

- Cross-host cache coherence
- CXL latency overhead
- CXL link bandwidth

Cache Coherence Is Not Required for PCIe Pooling

CXL memory devices available today do *not* support cross-host cache coherence

Key observation: PCIe devices often **bypass CPU caches** when accessing memory

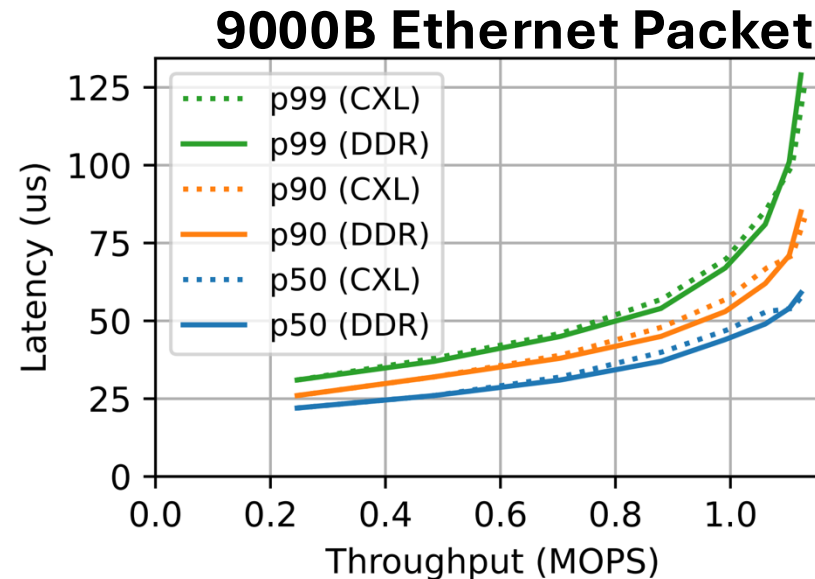
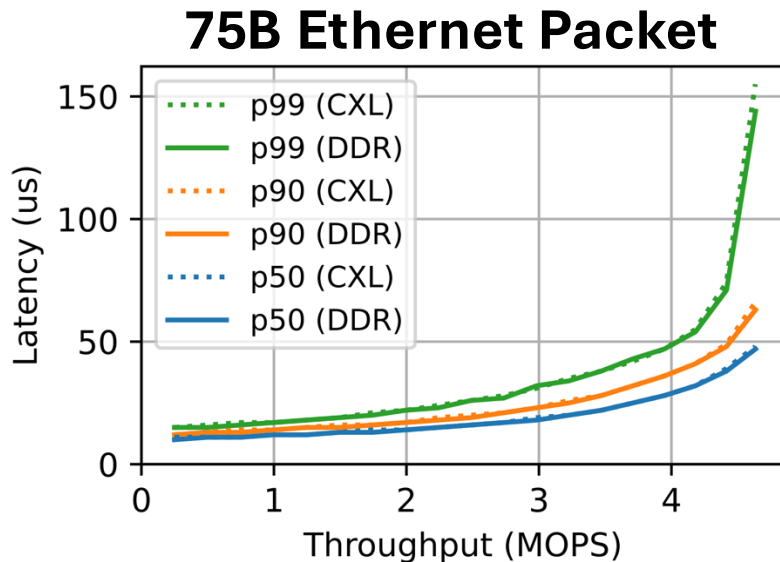


Use minimal software cache line flushes to ensure coherence

CXL Latency Incurs Small Overhead

CXL latency (220 ns) $\approx$ 2x local memory latency (100 ns)

- However, I/O latencies (e.g., network, storage) are at **μ s-scale**



Remaining challenge: signaling requests and completions over non-coherent CXL memory

CXL Links Provide Sufficient Bandwidth

Recent platforms (e.g., Intel Xeon 6) provide **64** CXL 2.0 / PCIe5 lanes per CPU socket

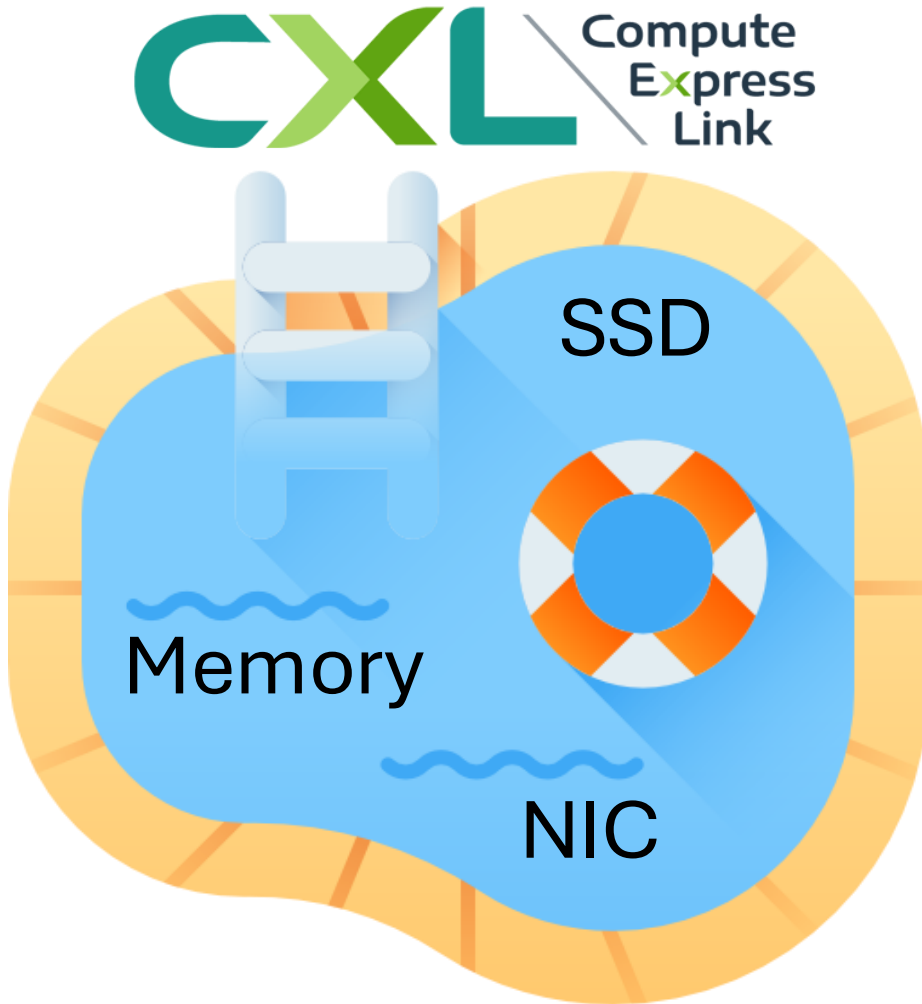
- Each PCIe lane provides 4 GB/s/direction bandwidth
- 64 lanes provide 512 GB/s/direction bandwidth in total

Use Case	Peak Bandwidth	# CXL Lanes
Multiple hosts sharing a single NIC	400 Gbps (50 GB/s)	16
Multiple hosts sharing a set of 6 NVMe SSDs	6 × 10 GB/s	16

Potential Research Directions 🤩

- **Datacenter network without ToRs**
 - Traditional datacenters have one ToR per rack, which could be the single point of failure
 - Can we eliminate ToRs and connect NICs to aggregation switches?
- **Load balancing**
 - Each host can send and receive network packets through multiple NICs
 - Dynamic flow migration between NICs, avoid high fan-in/fan-out by spreading traffic
- **Handling PCIe device failure**
 - How to detect device failure and fail over to other devices with minimal interruption
- **Pooling accelerators**
 - Accelerators with narrow use cases (e.g., FPGA, smartSSD, video decoding, *not* including GPU) have low utilization and low adoption rate
- **Scope of PCIe pooling (# hosts)**
 - How does the cost saving of pooling scale with the number of hosts

Thank you, and let's keep pooling!



- Pooling improves utilization and saves costs
- What are the challenges to implement pooling?
- What other devices can we pool?



yz@cs.columbia.edu