SNIA DEVELOPER CONFERENCE

SDC 25

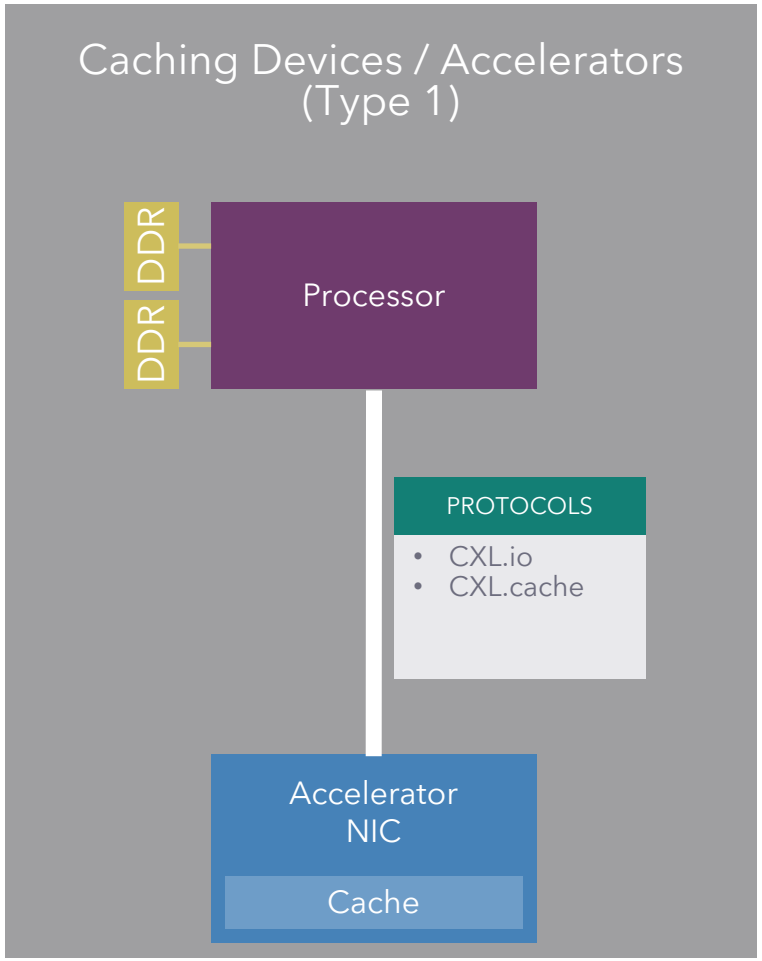
By Developers FOR Developers

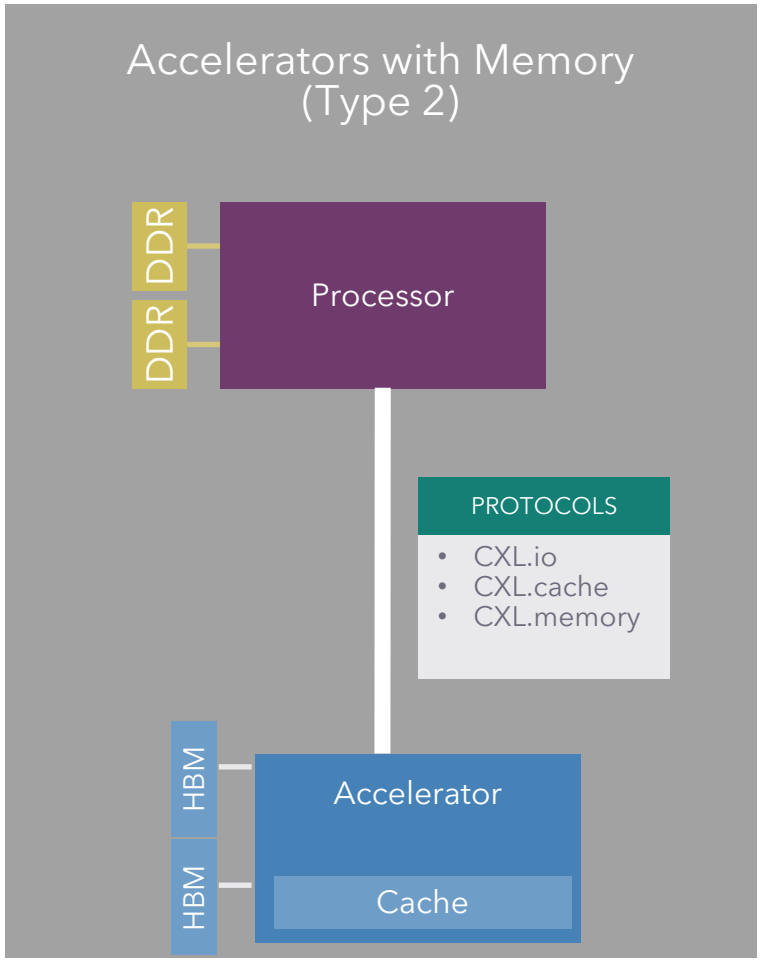
Hyatt Regency Santa Clara, CA
September 15-17, 2025

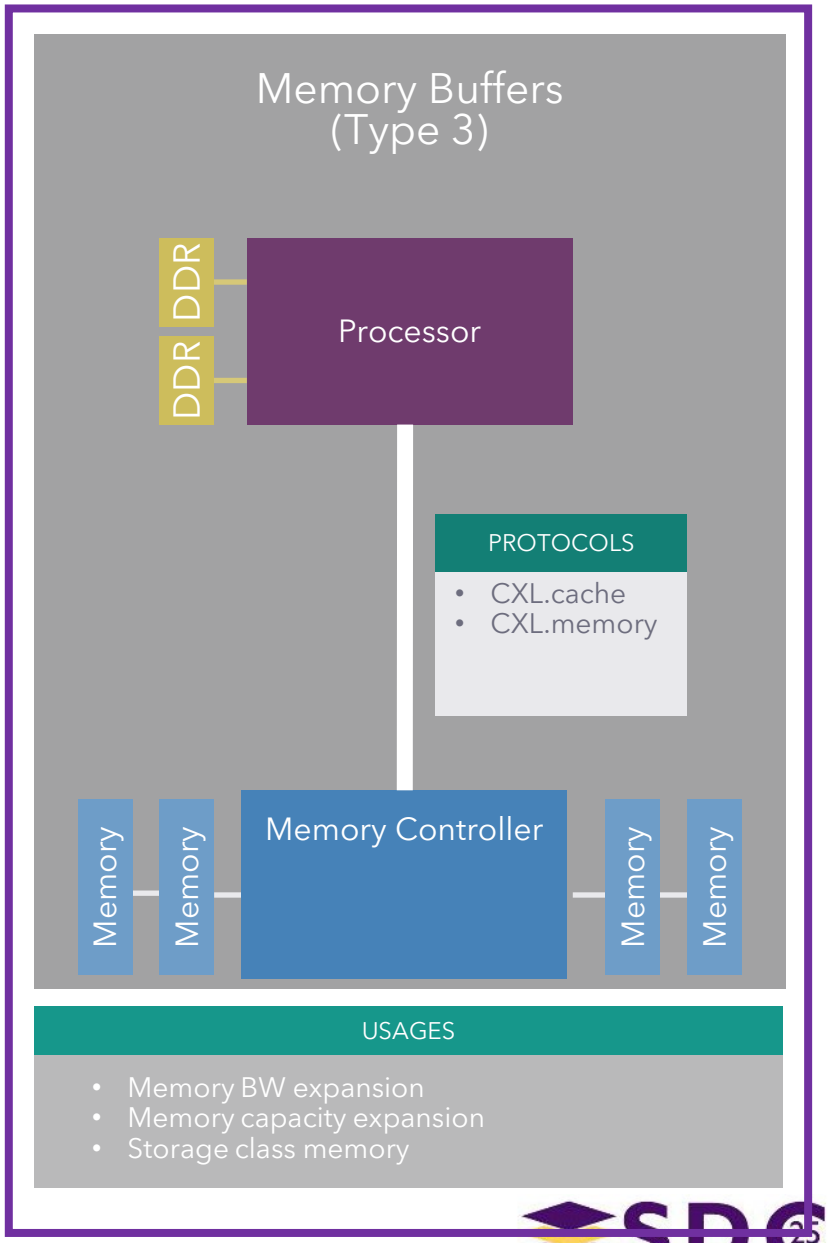
Advantages of CXL for Storage Applications

Anil Godbole, CXL Consortium MWG Chair

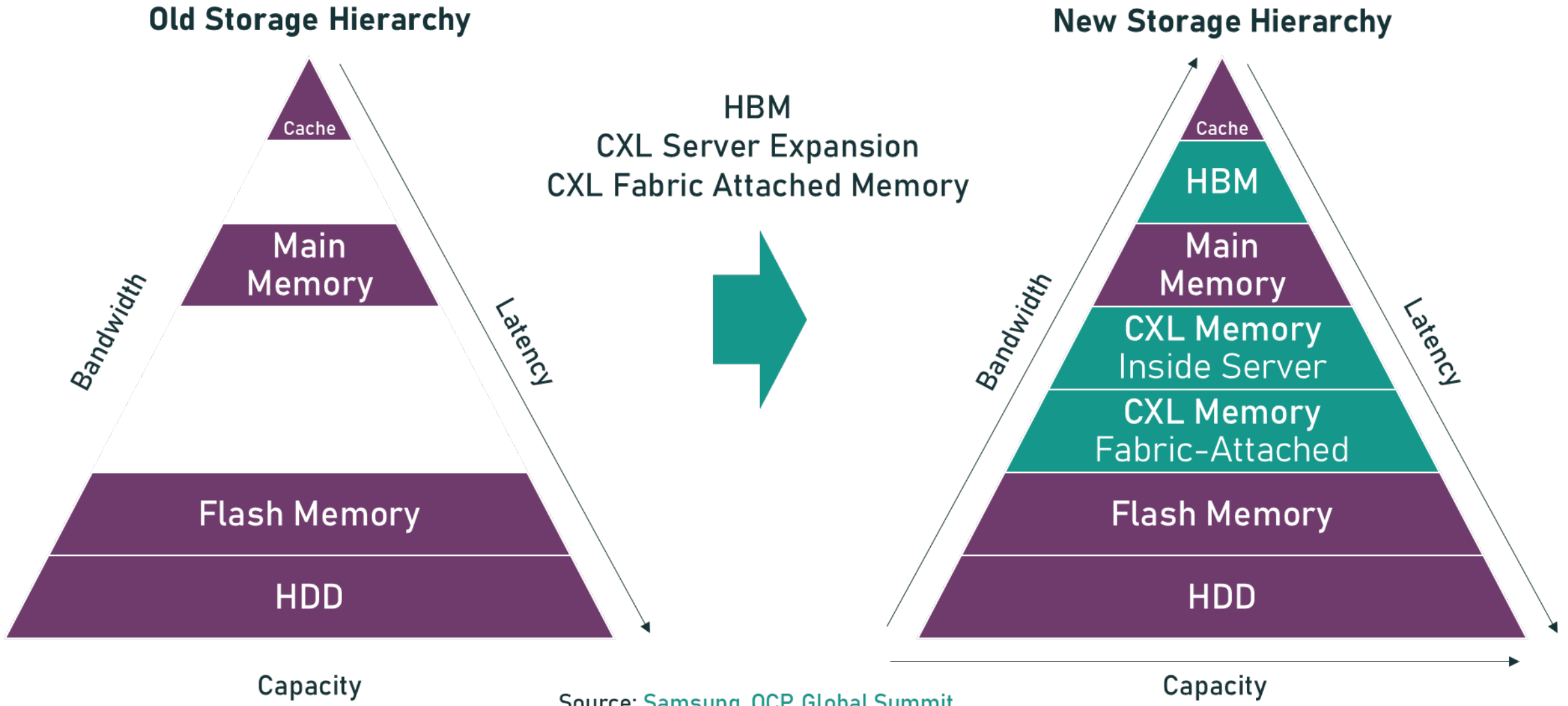
What is CXL?





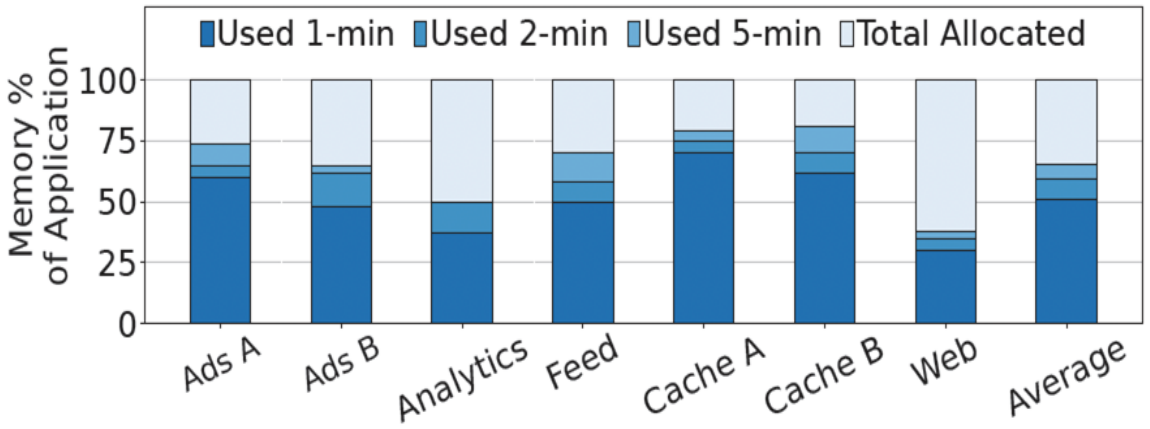


CXL Memory in New Storage Hierarchy

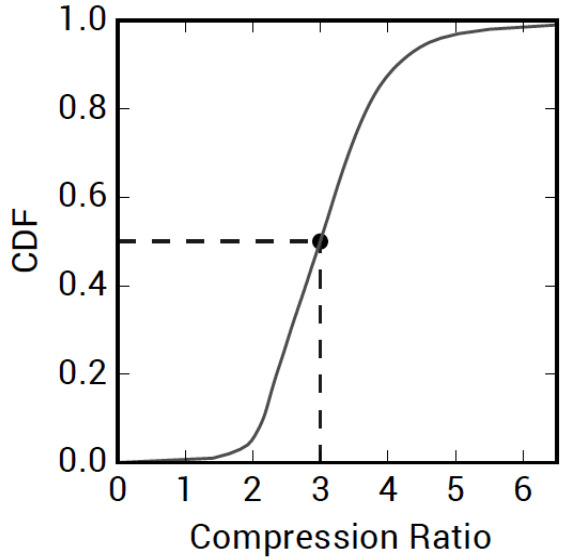


Source: Samsung, OCP Global Summit

Hyperscaler CXL-Memory based Swap Space Utility



Half of memory hasn't been used in the past minute



Cold data can compress at a 3:1 ratio; More like 2:1 once we account for incompressible pages

O/S-based Zswap like utility:

- Compress 'cold' pages & move them to CXL memory tier
- Make more room in main memory for active pages

Source: Making Memories at HyperScale with CXL; Webinar presented by Google/Meta

Optimizing for cost

DIMMs

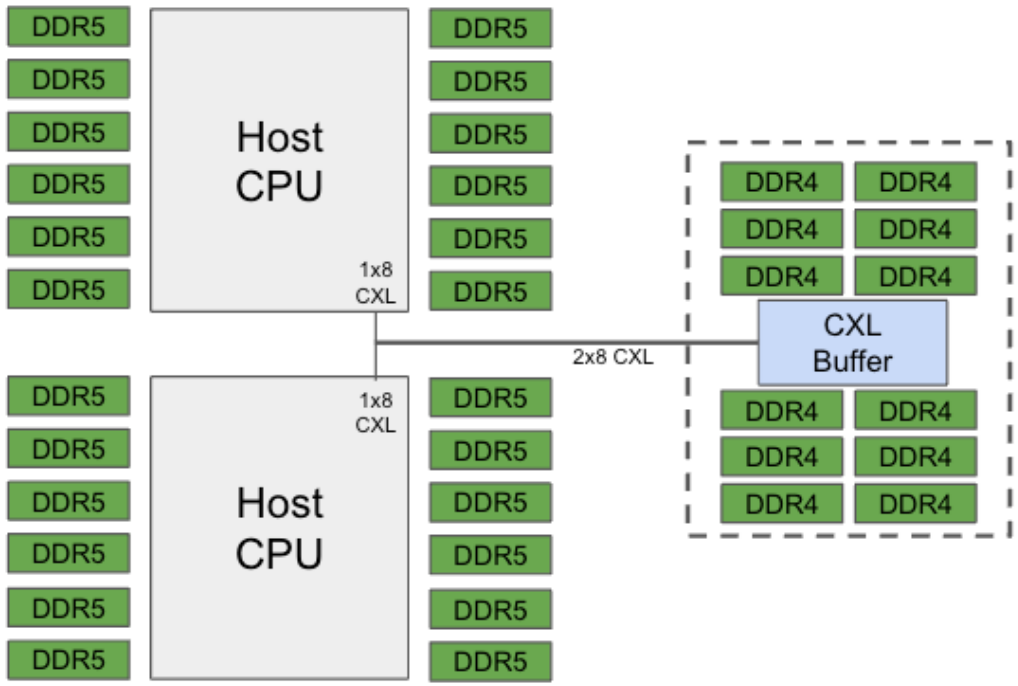
- Modular
- Serviceable
- Existing supply chain

DDR4

- Reuse & overstock
- Saves $
- Minimize embodied carbon footprint

Amortized costs over lots of DRAM!

- 4 channels, 3 DP ⇒ 12 DIMMs
- One CXL buffer per 432 DRAMs

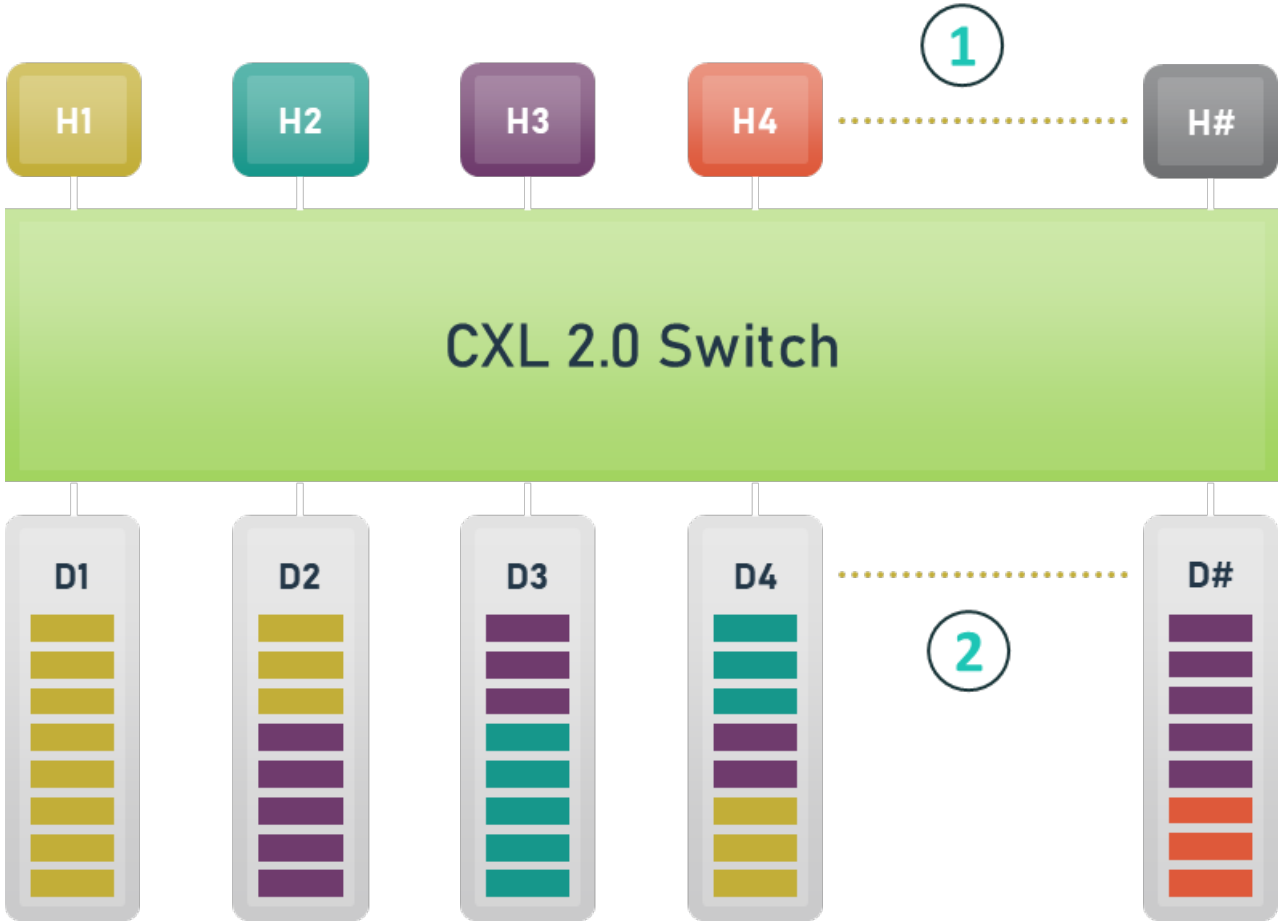


Example System Configuration

Claim: Much more performant than SSD-based Swap space utility

Source: Making Memories at HyperScale with CXL Webinar

CXL 2.0: Memory Pooling



Device memory allocated across multiple Hosts

Can also perform Memory Sharing using S/w techniques today

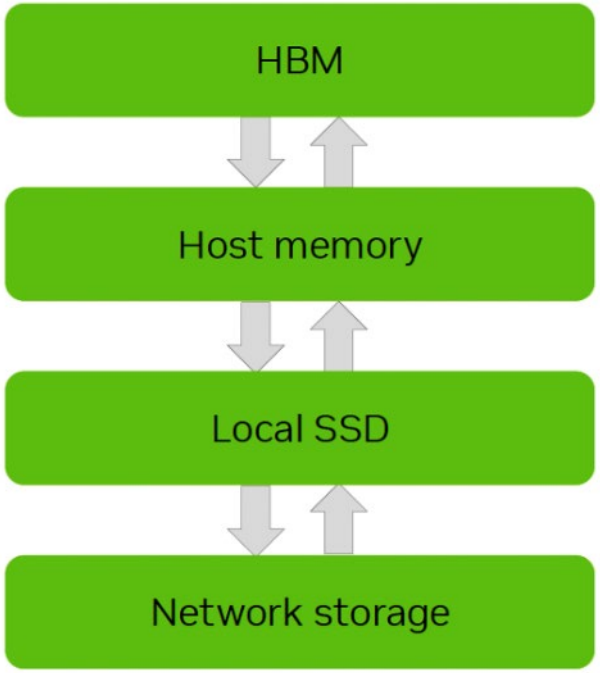
Backward compatibility of CXL Switch with PCIe devices can help with SSD-sharing or Backups/Checkpointing

Distributed Disaggregated Inferencing

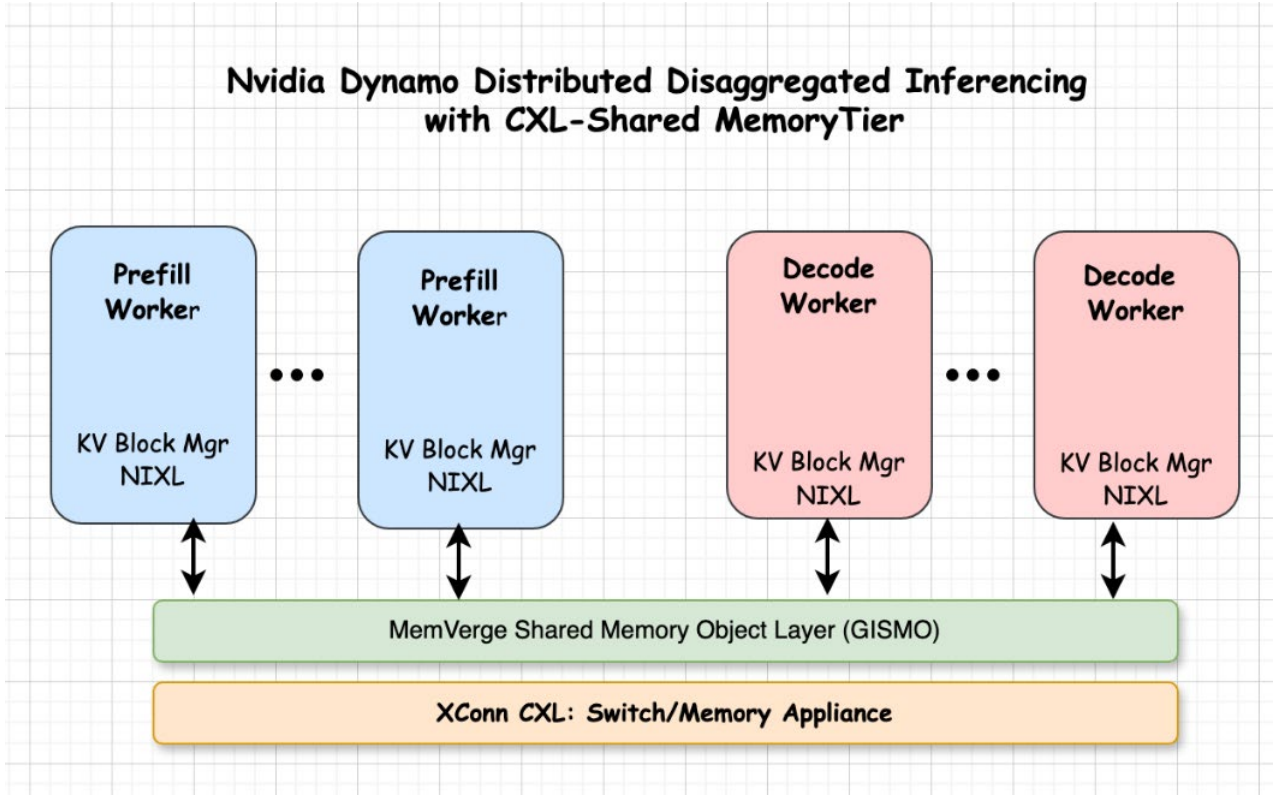
Distributed disaggregated inferencing platforms such as Nvidia's Dynamo help solve scaling challenges in three key areas: performance, correctness, and efficiency.

- ***GPU Thruput***: Disaggregate prefill and decode stages allows each stage to be optimized independently.
- ***Avoid KV cache re-computation***: Efficient, intelligent routing of workloads to pre-computed KV caches
- ***Memory bottlenecks***: Large-scale/Long context inference workloads require high amounts of performant KV cache storage that extend beyond the GPU's HBM memory capacity.
- ***New data transfer protocols such as NIXL***: For asynchronous, efficient movement of KV cache blocks in a highly dynamic inferencing cluster with tiered memory-storage.

Dynamo Memory-Storage Hierarchy



Augmenting Nvidia Dynamo KV Cache Hierarchy



NVIDIA

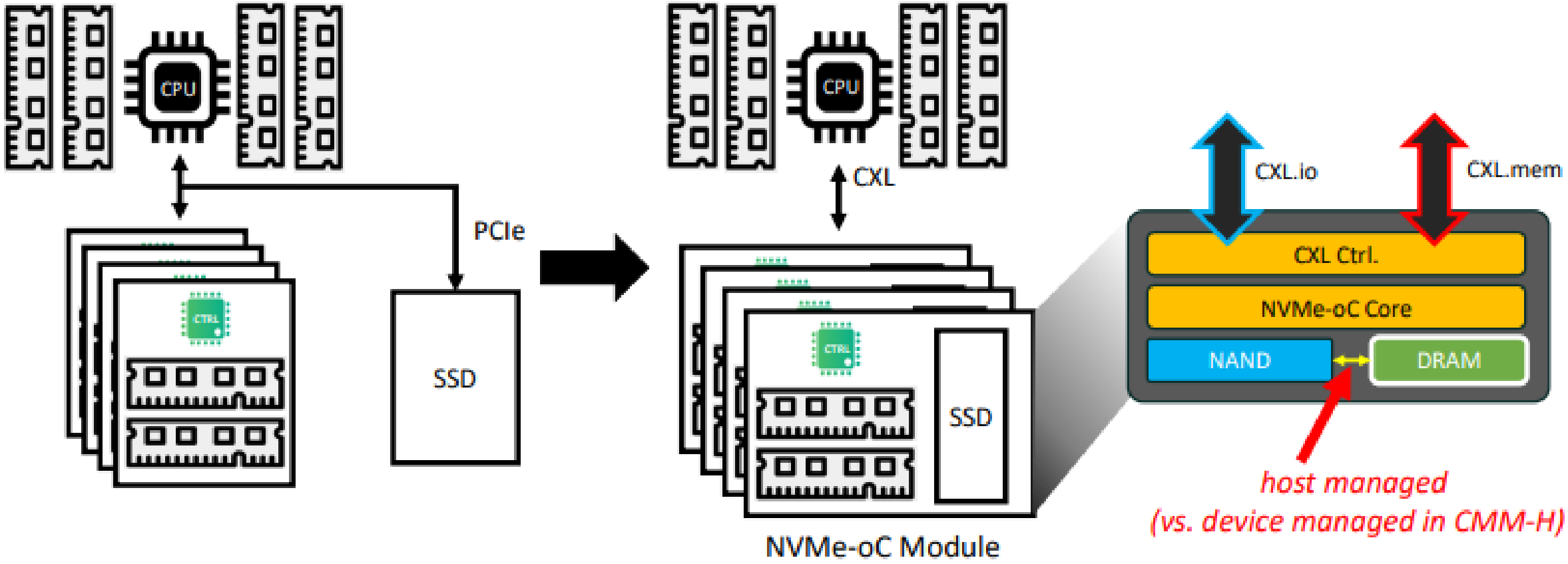
MemVerge

XCONN TECH

- CXL Shared Memory can be plugged into the NIXL framework & KV Block Manager to augment the KV cache's performance and add a new tier of memory situated between host memory and local SSD that can scale to10TB+.

NVMe-over-CXL: Combo Memory Expansion & Storage

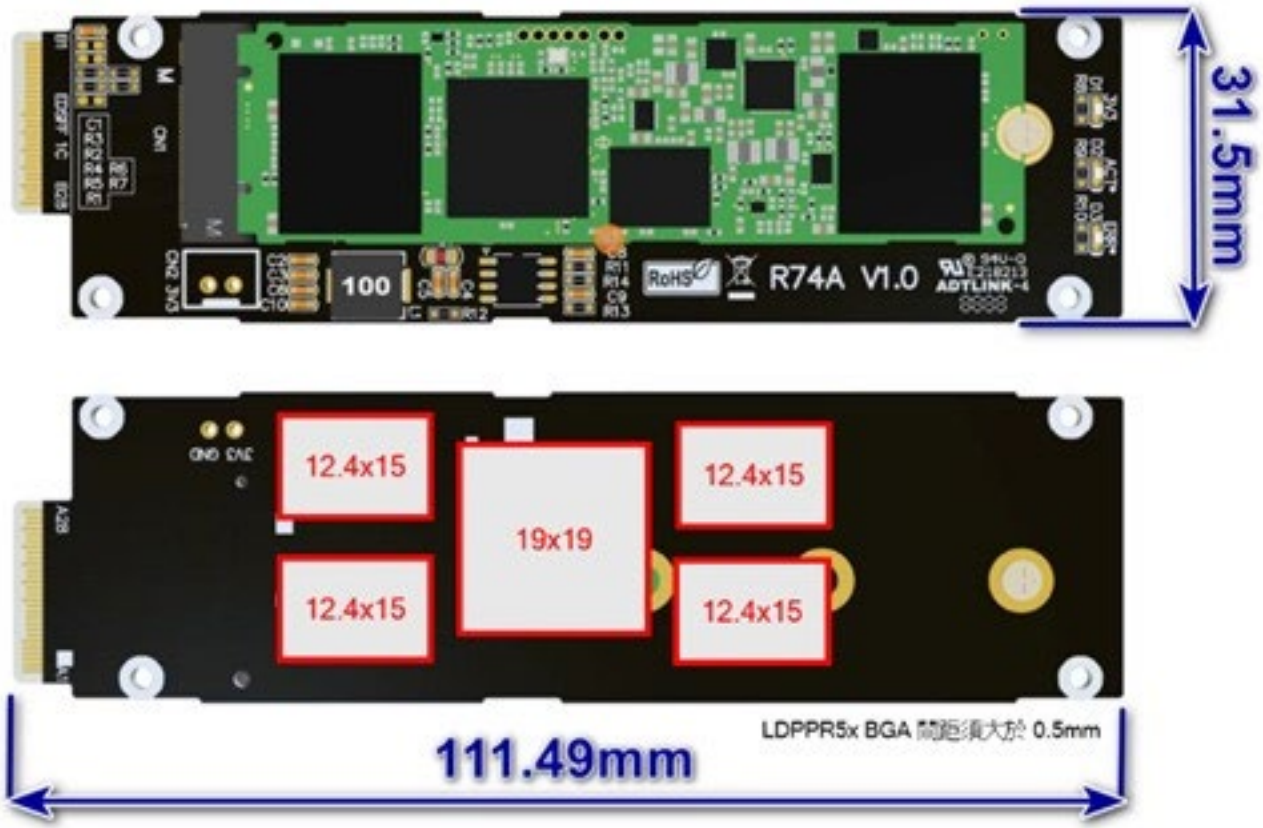
"NVMe-oC, with DAX-tiering, unifies DRAM, CXL, and SSD into a seamless and scalable memory architecture"



Suited for PCIe link constrained CPU SKUs (HEDT, Future Laptops)

Source: Wolley's presentation at FMS 2025; HEDT: High-end Desktops, Laptops

Wolley's E1.S CXL Memory Module



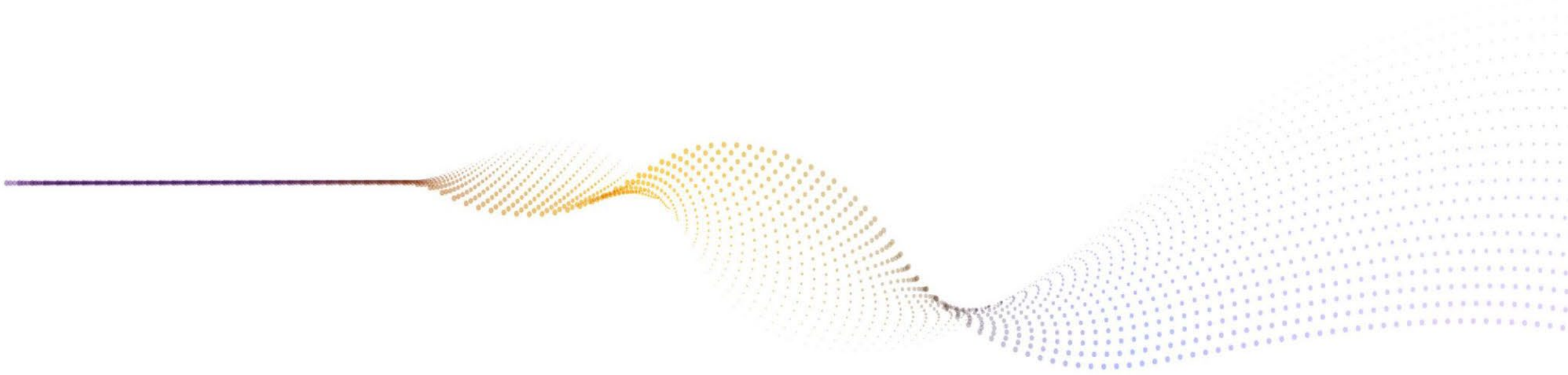
- **Form Factor:** E1.S
- **Device Interface:** PCIe Gen5 x8 (32 GB/s)
- **CXL Support:** CXL 2.0
- **Memory:** LPDDR5X-6400, 128-bit, 32/64 GB
- **Storage**: >1TB PCIe Gen5 x4 NVMe SSD
- **Layout**: storage (top picture), memory (bottom picture)
- **Use Case:** memory expansion, AI acceleration

Key Advantages

- Expand memory bandwidth without sacrificing storage connectivity by sharing a common PCIe/CXL interface
- Increase effective memory capacity through DRAM+SSD virtualization, lowering system TCO

Summary

- CXL Memory based Applications for Storage are beginning to emerge
- Usage of CXL Memory as an in-between tier between DRAM & SSD
  - Helps with Memory Capacity & Bandwidth expansion for CPUs
  - Helps to alleviate pressure on SSD layer in many emerging applications
- Emergence of CXL Switches & Memory Pooling can be leveraged for many modern applications including AI



Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference

