OCP FUTURE TECHNOLOGIES SYMPOSIUM

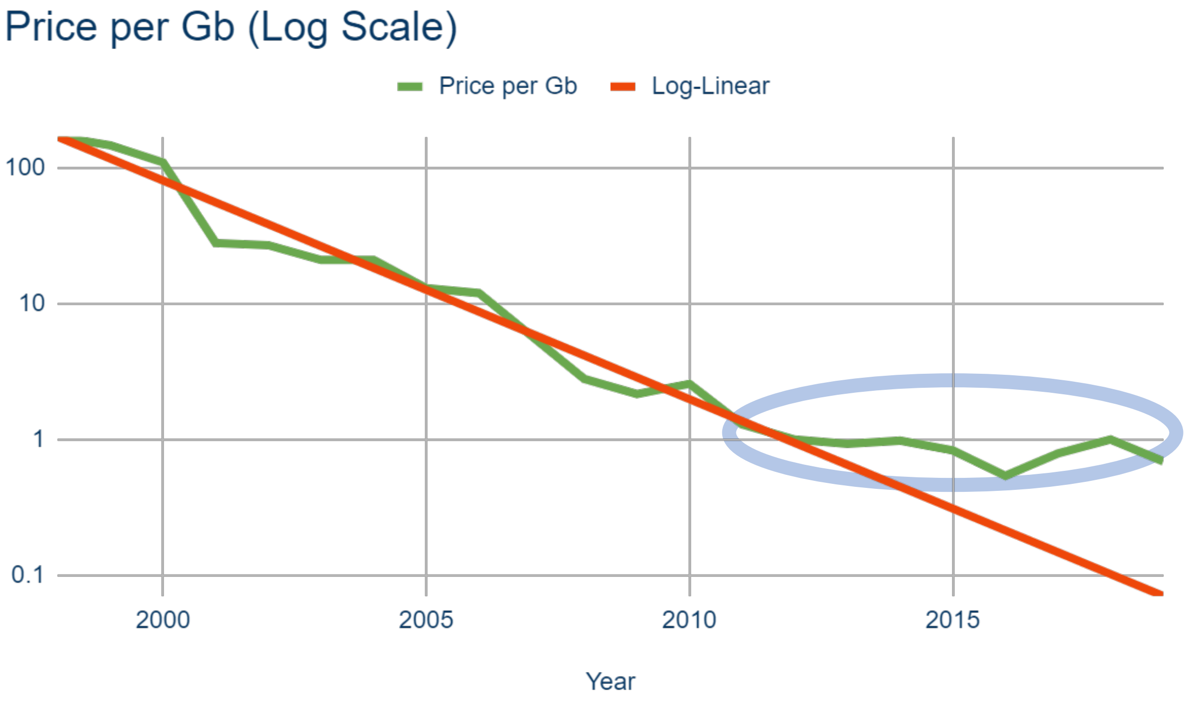
OCP Global Summit

November 8, 2021 | San Jose, CA

Software Defined Memory: A Meta perspective

Chris Petersen, Hardware Systems Technologist, Meta

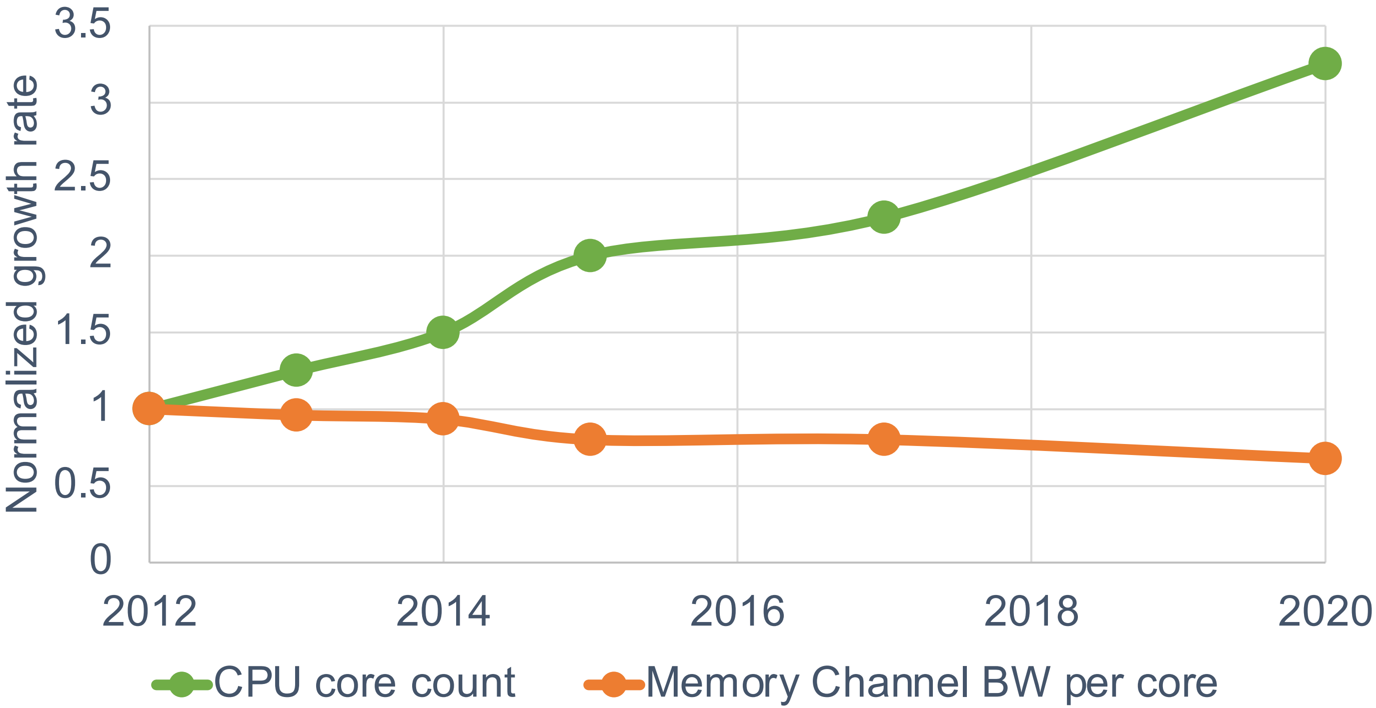
Increasing Memory Cost and Power



Memory an increasing % of system power and cost

- Memory price (cost/bit) flat due to scaling challenges
- Memory power scaling with speed

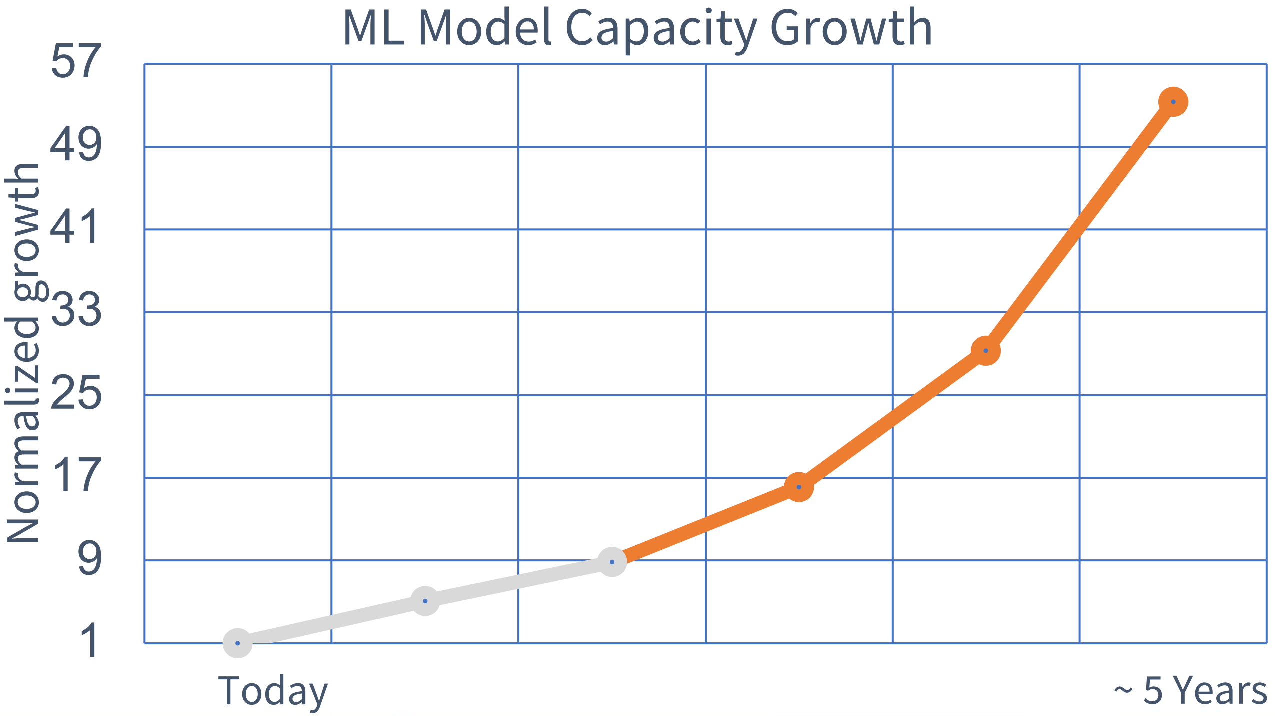
Increasing Core Counts Drives Growth



Increasing core counts driving memory demand

- Increased Bandwidth
- Increased Capacity

Machine Learning Growth

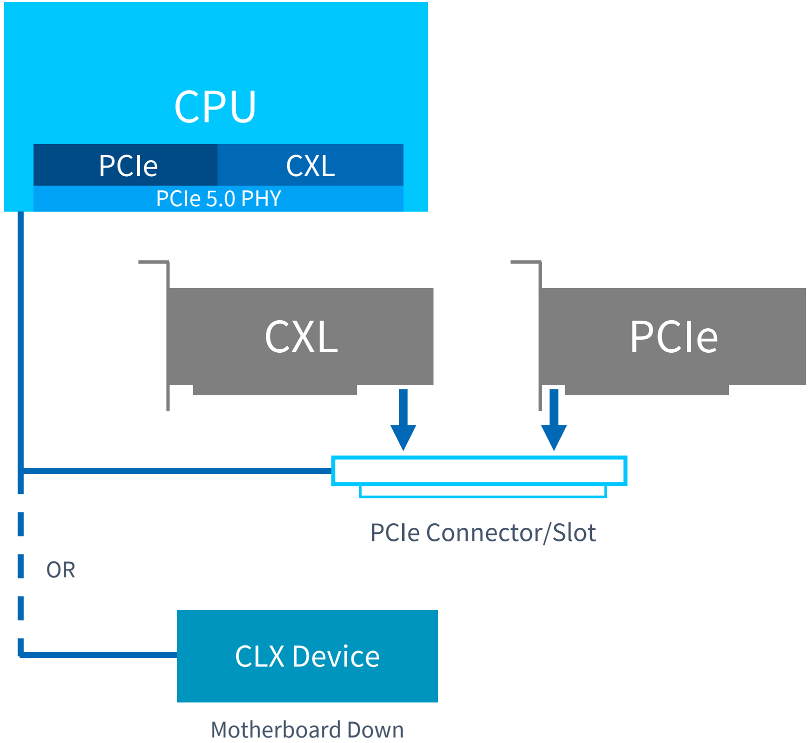


ML Models are growing rapidly

- ~50X growth in ~5 years
- Existing memory hierarchy can't keep pace

Compute Express Link (CXL) Introduction

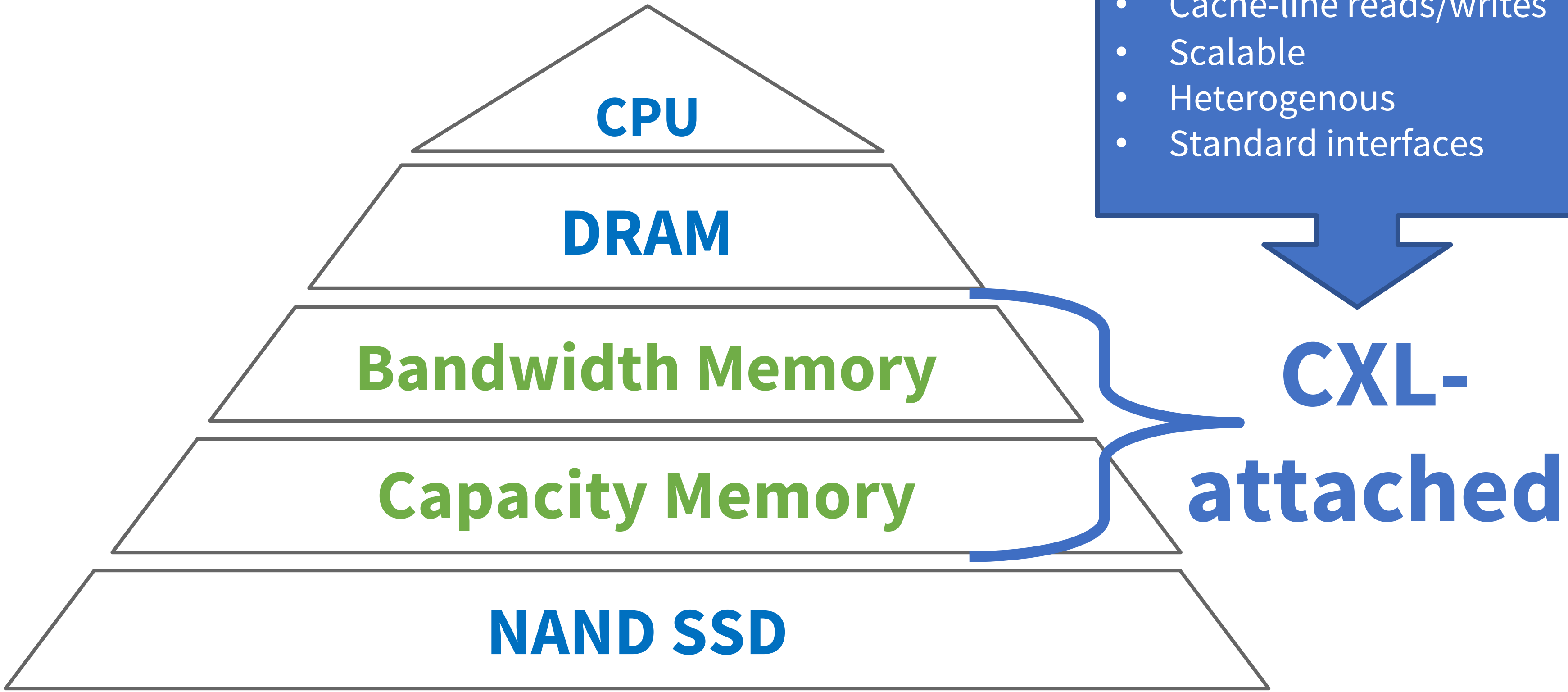
- Processor Interconnect:
  - Open industry standard
  - High-bandwidth, low-latency
  - Coherent interface
  - Leverages PCI Express®
  - Widths: x4, x8, x16



CXL Memory Tiers

Requirements:

- Memory, not storage
- Cache-line reads/writes
- Scalable
- Heterogenous
- Standard interfaces



CPU

DRAM

Bandwidth Memory

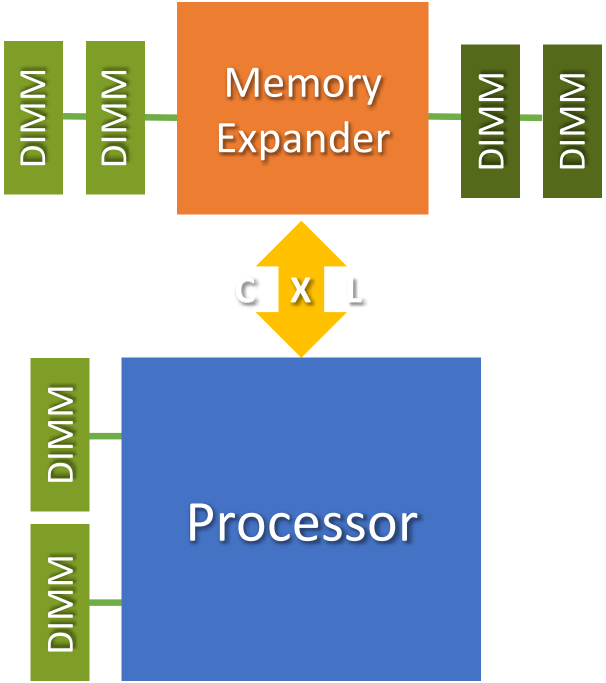
Capacity Memory

NAND SSD

CXL-attached

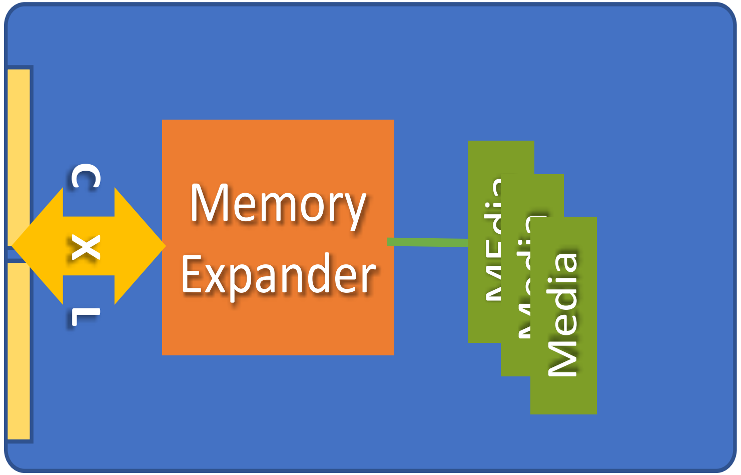
Bandwidth Memory Tier

- **Use Cases:** Warm Pages, Page Migration
- **BW:** BW per GB close to that of DDR4 memory
- **Latency:** NUMA-like
- **Power:** ~90% of DDR5 at ISO capacity
- **Capacity:** Scales with standard RDIMMs
- **Form factor:** Initial solutions focused on “chip down + DIMMs”



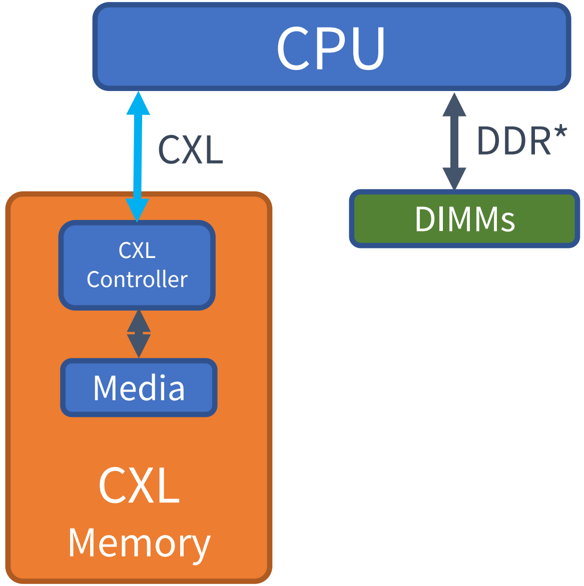
Capacity Memory Tier

- Use Cases: Caching and ML Models
- BW: BW per GB 5-10% of DDR5 memory
- Latency: Hundreds of ns
- Power: ~50% of DDR5 at ISO capacity
- Capacity: 256GB - 1TB
- Form factors: Use hot-pluggable form factors (like E1 or E3)

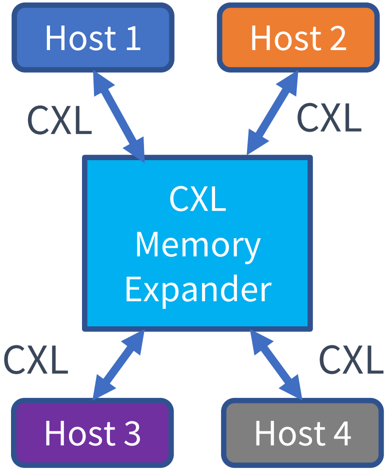


CXL Memory Evolution

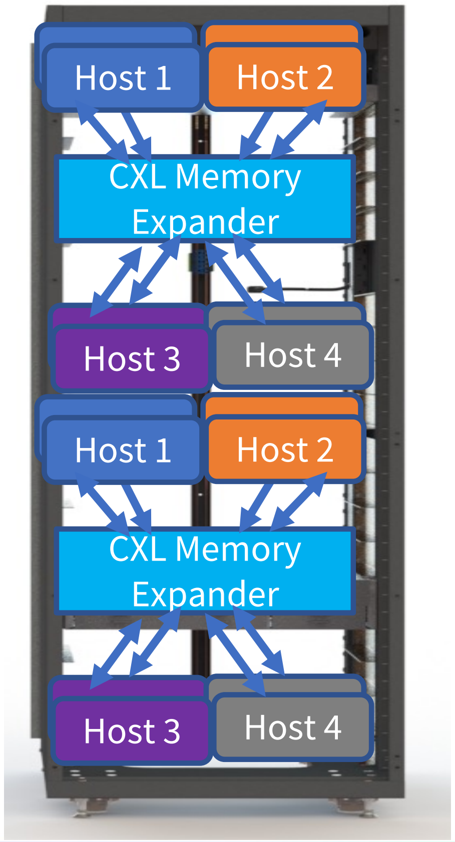
Direct-attach



Small Pools



Rack-scale Pools



Parting thoughts

- Lots of work ahead of us! Industry collaboration is critical.
- Think at the system level including SW integration, and also in phases
- Multiple CXL memory tiers are needed for multiple use cases. One size does **not** fit all!

OCP
GLOBAL SUMMIT
HIGHLY SCALABLE HPC & AI
TOTAL SOLUTIONS
MARVELL

OCP FUTURE TECHNOLOGIES SYMPOSIUM

2021 OCP Global Summit | November 8, 2021, San Jose, CA