Software Defined Memory (SDM) with Faster Interconnects and Tiered Memory

Manoj Wadekar, Hardware Systems Technologist, Meta

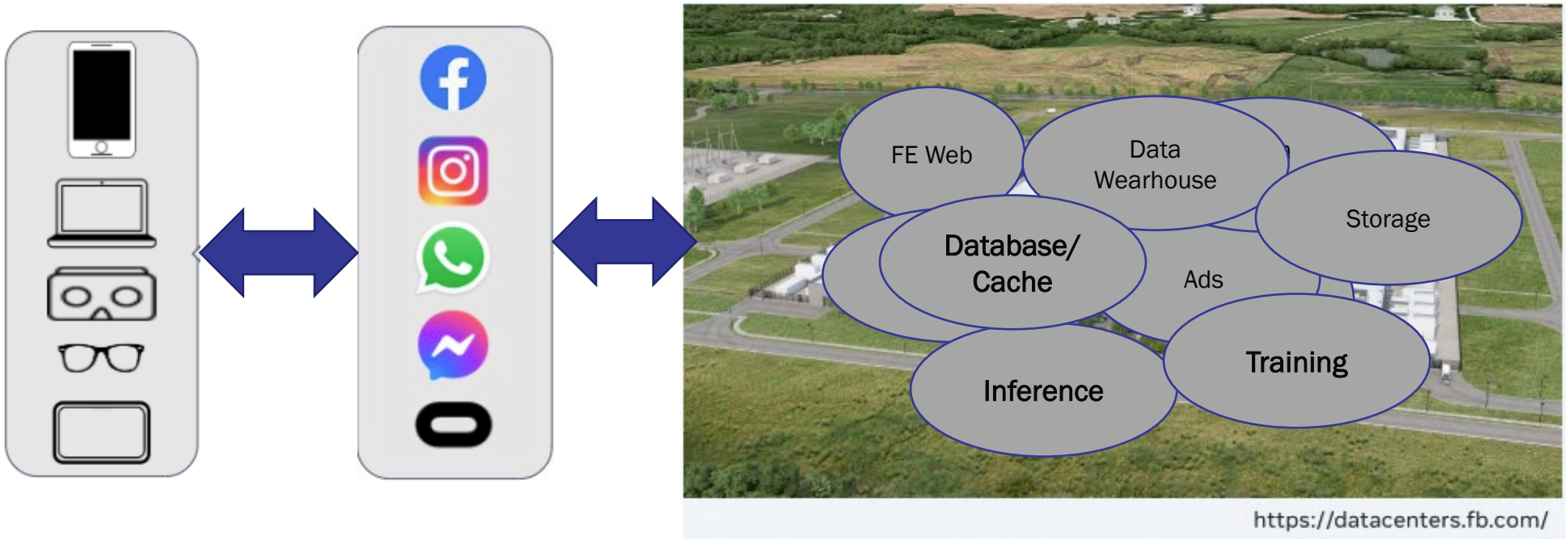
Anjaneya "Reddy" Chagam, Sr. Principal Engineer, Intel

Meta intel

Agenda

- Memory Challenges in Hyperscale Infrastructure
- Need for Software Defined Memory (SDM)
- SDM with Compute Express Link (CXL)
- SDM use-cases
- Summary

Hyperscale Infrastructure



- Application performance and growth depends on
  - DC, System, Component performance and growth
  - Compute, Memory, Storage, Network..
- Focusing on Memory discussion

Meta

Memory Challenges

Bandwidth and Capacity

- The Gap between bandwidth and capacity is widening
- Applications ready to trade between bandwidth and capacity

Power

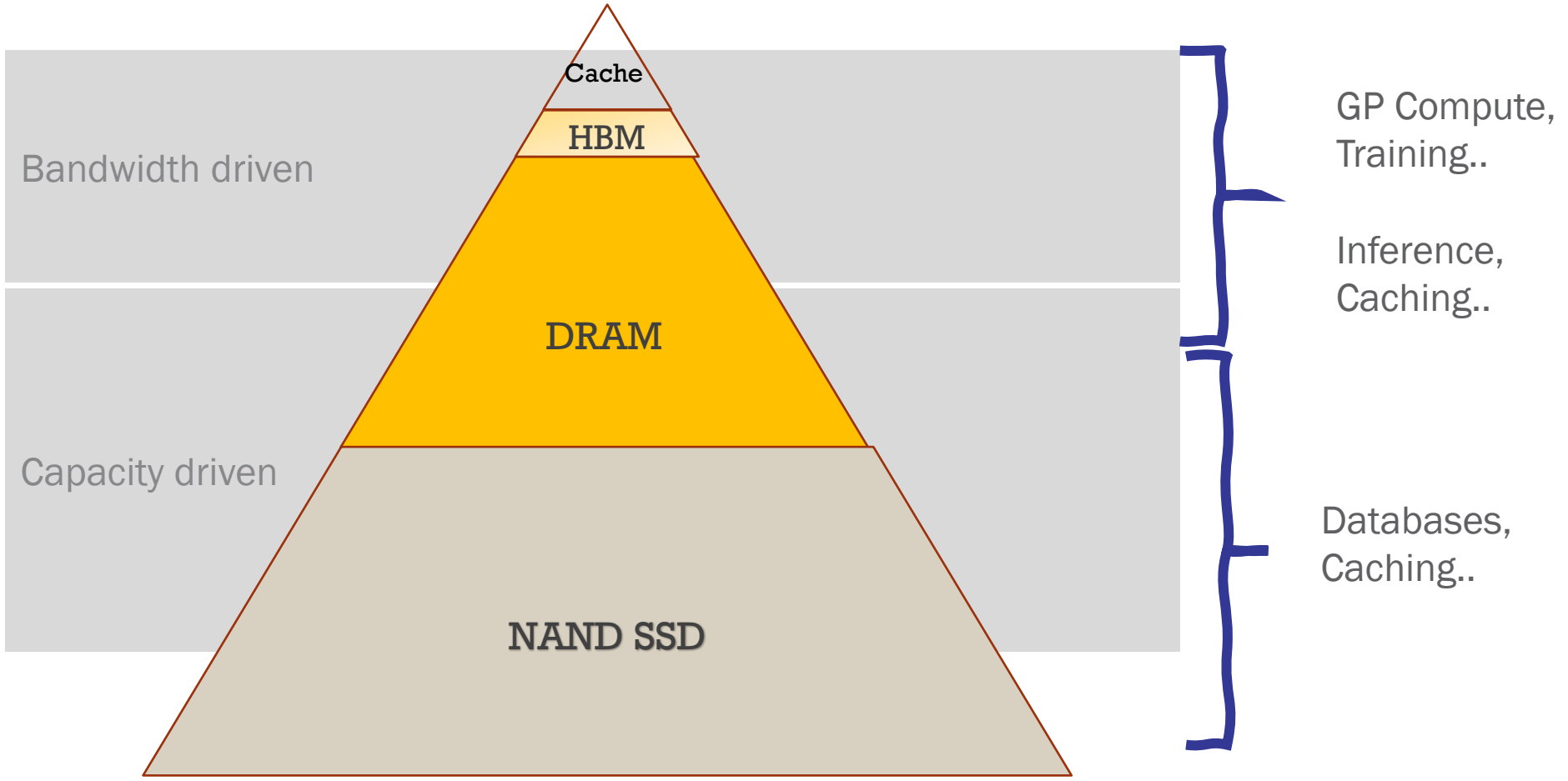
- DIMMs consume significant share of rack power
  - DDR5 exacerbates this
- Applications co-design to achieve higher capacity at optimized power

TCO

- Cost impact of min capacity increase and Die/ECC overheads
- Applications can trade performance/capacity to achieve optimal TCO

Meta

“Memory” Pyramid today



Meta

Hierarchical Memory: Need for SDM

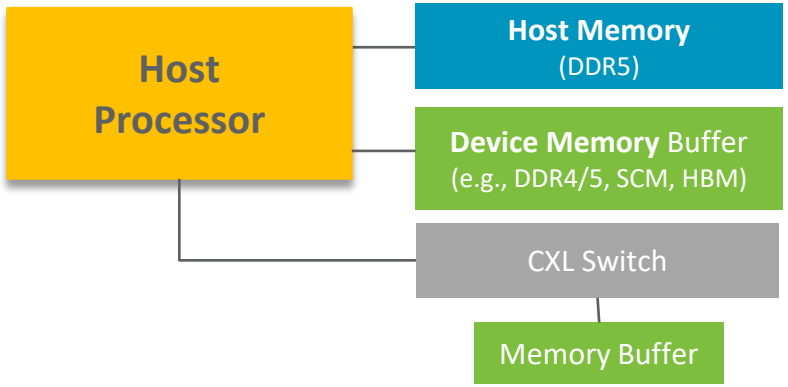
Emerging workloads driving need for **faster data processing**

Need for increased **memory capacity and bandwidth**

Lack of **open interconnect standard**

CXL

An open industry-supported cache-coherent interconnect for processors, memory expansion and accelerators

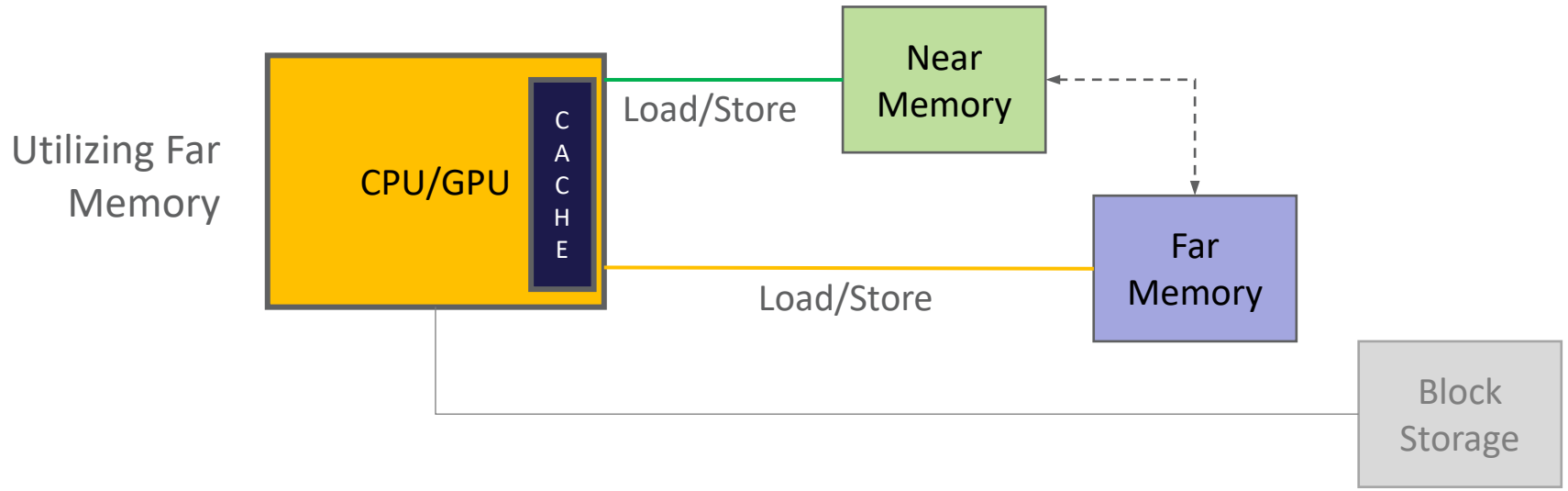


- Hierarchical memory with varying bandwidth/latency characteristics
- Late binding/on-demand provisioning
- Bare-metal v/s virtualized
- Memory Pooling (rack level)
- Kernel managed v/s application managed

Software-Defined Memory (SDM) is needed for optimizing resource utilization to deliver lower TCO and power-efficiency

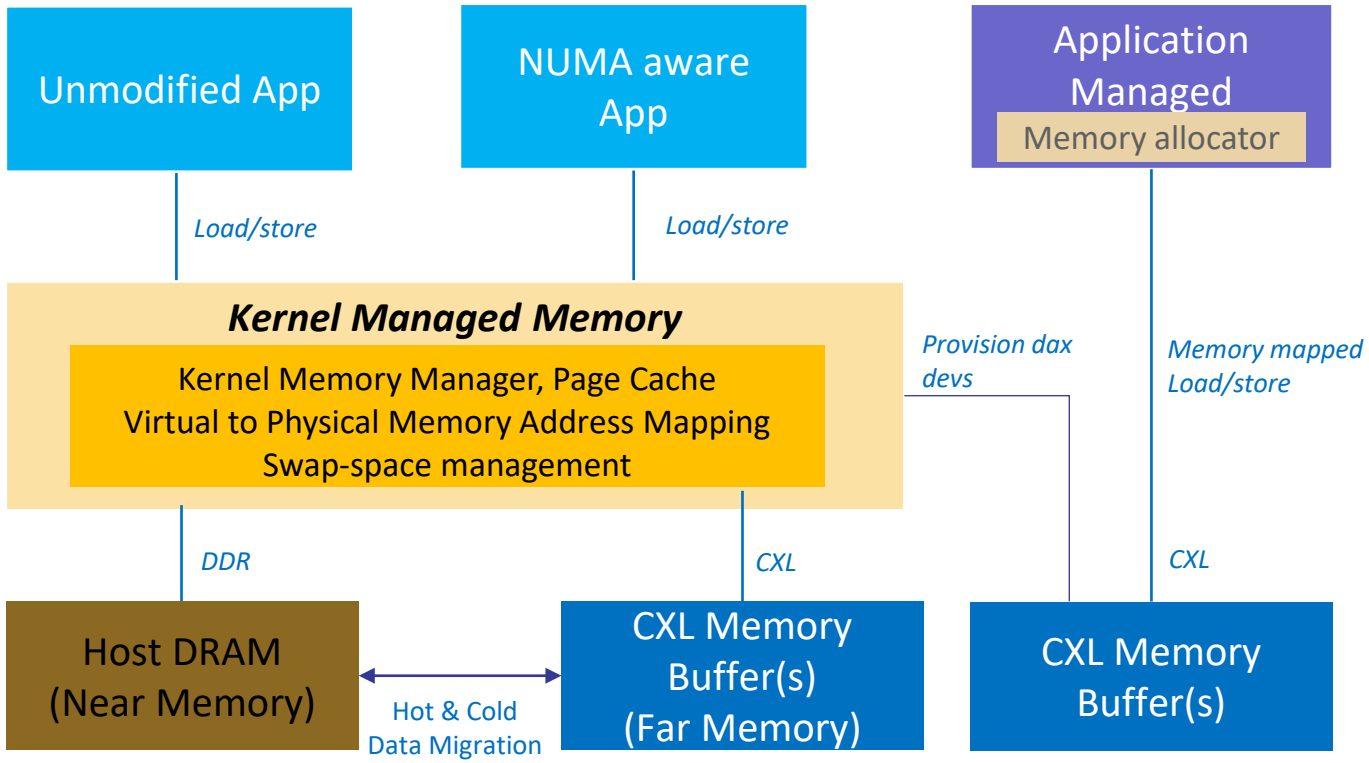
SDM - Introducing Far Memory

- Memory needs are growing faster than underlying memory technology
- Tiered memory can provide additional capacity at appropriate performance (with load/store) to sustain application needs
- Initial SDM opportunity is around CXL 1.1 Memory expansion for workload acceleration and lower costs with CXL connected DRAM



CXL Memory tier can enable performance/capacity trade off to achieve TCO gains

SDM - Kernel & Application Managed Memory



Kernel Managed

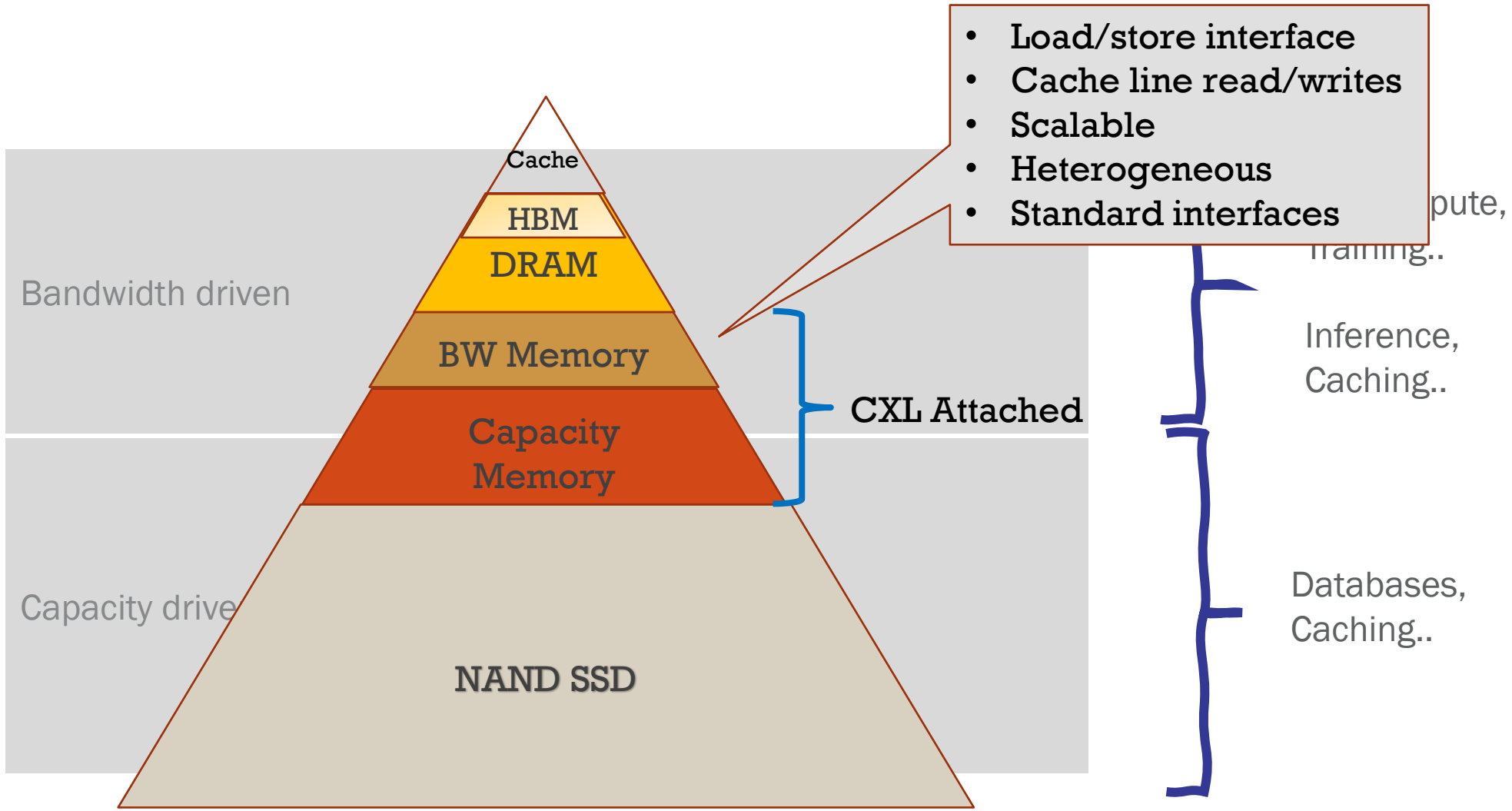
- OS can **map far memory** into application's **virtual address space**
- Applications **can execute from pages in far memory** (albeit more slowly)
- Kernel memory manager can implement **varying policies for migrating hot & cold pages** between tiers

Application Managed

- **Allows apps** to access CXL **memory as memory-mapped files**
- **Built on top of DAX** (Direct Access) file system

Linux Kernel tiering in early development stages

“Tiered Memory” Pyramid with CXL



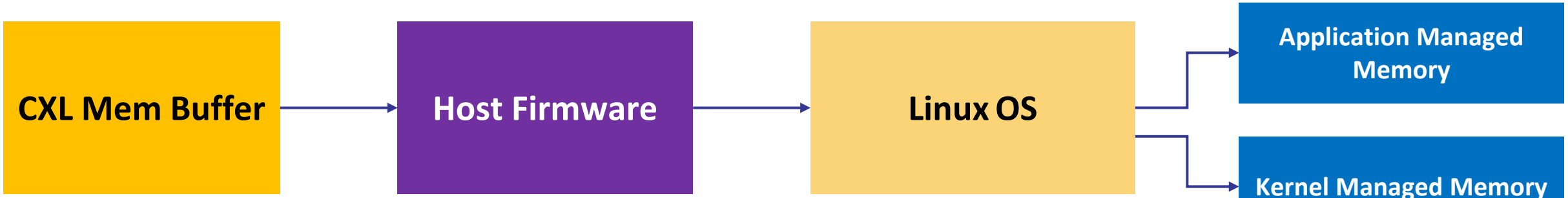
Memory Technologies

	Compute	Storage	Training	Inference
DDR4	☑	☑	☑	☑
DDR5	☑	☑	☑	☑
HBM			☑	☑
CXL+DDR	☑	☑		☑
SCM (PCIe/CXL) [Exploration Phase]		☑	☑	☑

Use Case Examples

- **Caching (e.g. Memcache/Memtier (Cachelib), Redis etc.)**
  - Need to achieve higher QPS while satisfying “retention time”
  - Higher memory capacity needed
  - Current solutions include ”tiered memory” with DRAM+NAND, but need load/store
- **Databases (E.g. RocksDB/MongoDB etc.)**
  - Need to achieve efficient storage capacity per node and deliver QPS SLA
  - Higher amount of memory enables more storage per node
- **Inference (E.g. DLRM)**
  - Petaflops and Number of parameters are increasing rapidly
  - AI Models are scaling faster than the underlying memory technology
  - Current solutions include ”tiered memory” with DRAM+NAND, but need load/store

Intel 4th Gen Xeon Processor (Sapphire Rapids): CXL Provisioning (*Caching Use Case*)



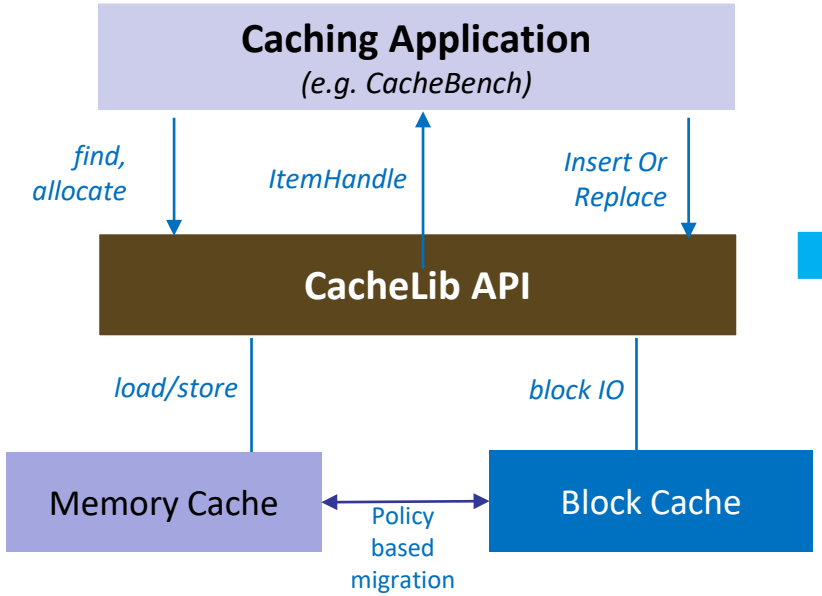
Host-managed Device Memory (HDM)

- **Setup Knobs**
  - Port bi-furcation
  - CXL security level etc.
- **CXL Training**
- **SAD Programming**
- **ACPI Table Setup**
  - CEDT, SRAT, HMAT etc.
- **CXL Memory Map**
  - Special Purpose (EFI_MEMORY_SP)
  - Conventional (EFI_MEMORY_WB)

- **CXL Memory Configs**
  - Conventional (Memory only NUMA domain)
  - HMEM/DAX device (/dev/dax<x>.<y>)
- **DAX Memory Config Knobs**
  - efi=nosoftreserve kernel command line
  - daxctl to promote/demote to system ram
- **Kernel Tiering for Conventional Memory**
- **CXL Mailbox Driver**
- **CXL Command Line Tools**

Intel 4th Gen Xeon Processor (Sapphire Rapids): Caching Use Case (CacheBench)

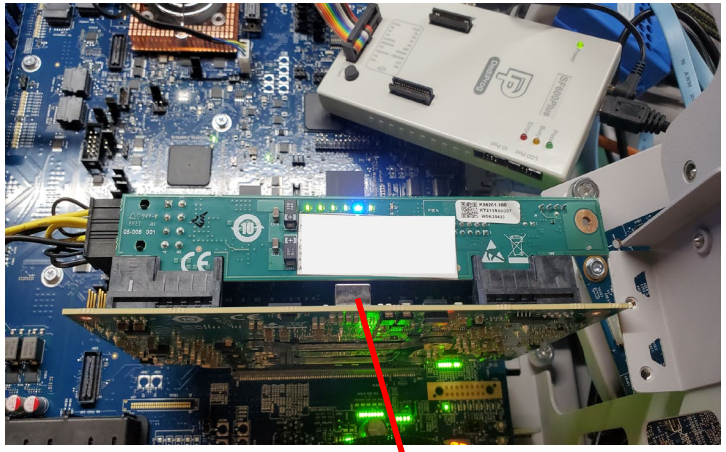
CacheBench



Github: https://github.com/facebook/CacheLib

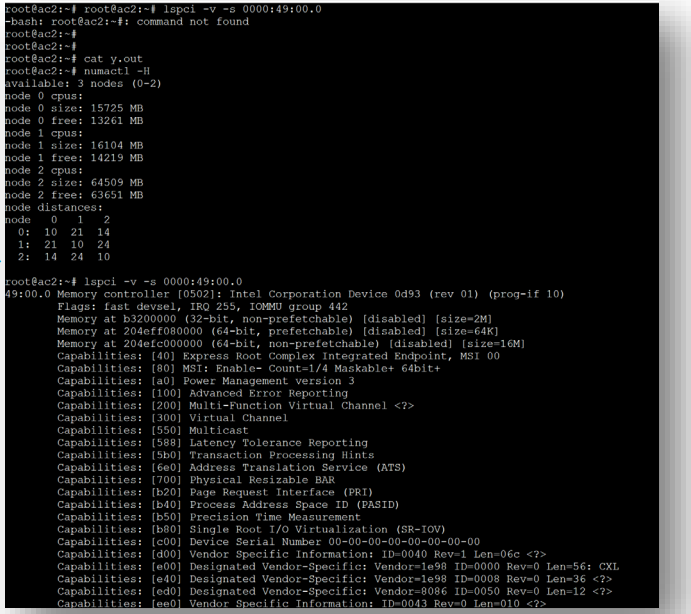
Open-source benchmarking tool for evaluating caching performance

Intel Sapphire Rapids Pre-production Platform



Intel Pre-production CXL FPGA Memory Buffer

CXL Memory in Linux OS



OCP SDM activity and progress

- SDM's focus: Apply emerging memory technologies in the development of use cases
- The OCP SDM group has three real-world memory focus areas:
  - Databases/Caching
  - AI/ML & HPC
  - Virtualized Servers
- SDM Team Members: AMD, ARM, Intel, Meta, Micron, Microsoft, Omdia, Samsung, VMWare
- Vendors are demonstrating CXL Capable CPUs and devices
- Meta and others are investigating solutions to real world memory problems

Call to Action: Join OCP SDM workstream!

Notices & Disclaimers

Intel technologies may require enabled hardware, software or service activation. Your costs and results may vary.

No product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/benchmarks .

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

.

intel.