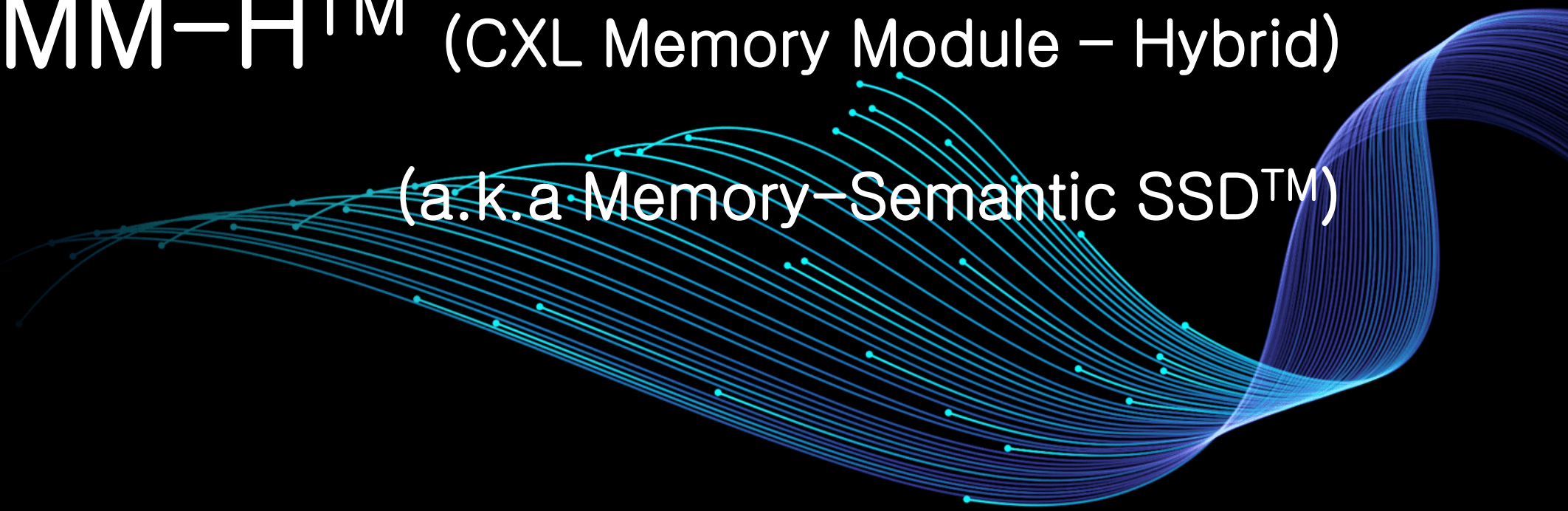


DLRM Acceleration with CMM-H™ (CXL Memory Module – Hybrid) (a.k.a Memory–Semantic SSD™)

A decorative graphic consisting of numerous thin, light blue lines that curve and flow across the lower half of the slide. Small, glowing blue dots are scattered along these lines, creating a sense of motion and data flow.

Jeong–Uk Kang

Master, Memory Business

Samsung Electronics

October 19, 2023

References

SAMSUNG

Innovation with CXL and Storage

Wonseb Jeong
Staff Engineer, Memory Business
Samsung Electronics

July 04, 2023

SAMSUNG

STORAGE DEVELOPER CONFERENCE
SDC²³
BY Developers FOR Developers

Is SSD with CXL Interfaces Brilliantly Stupid or Stupidly Brilliant?

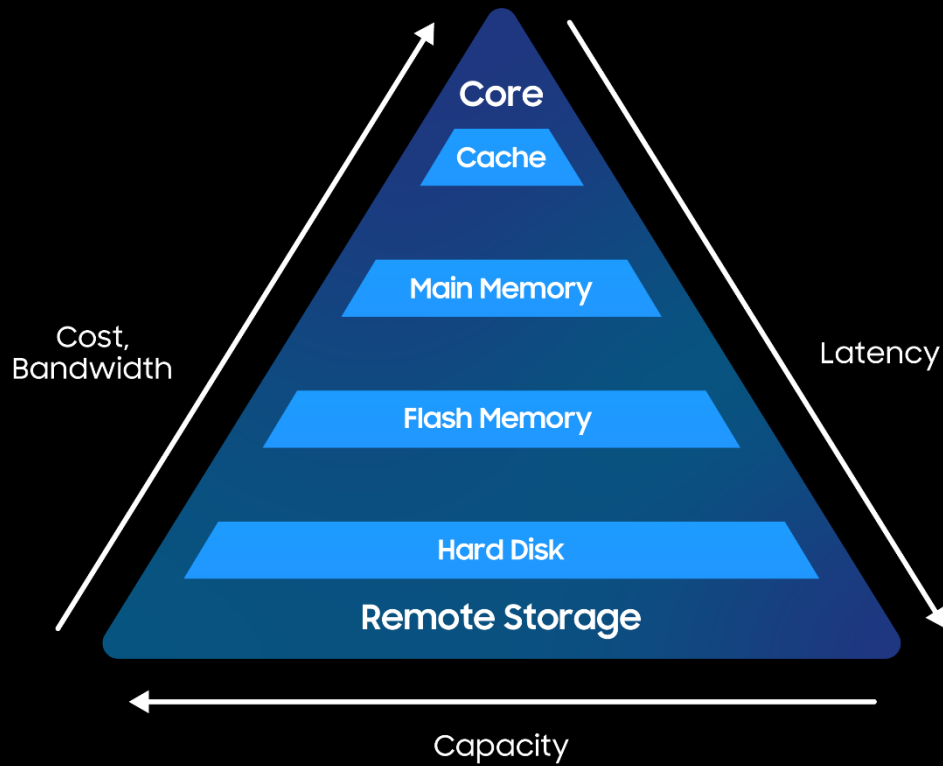
YANG SEOK KI, Ph.D.
Vice President, Device Solutions Research America - Memory, Samsung Electronics

September 19, 2023

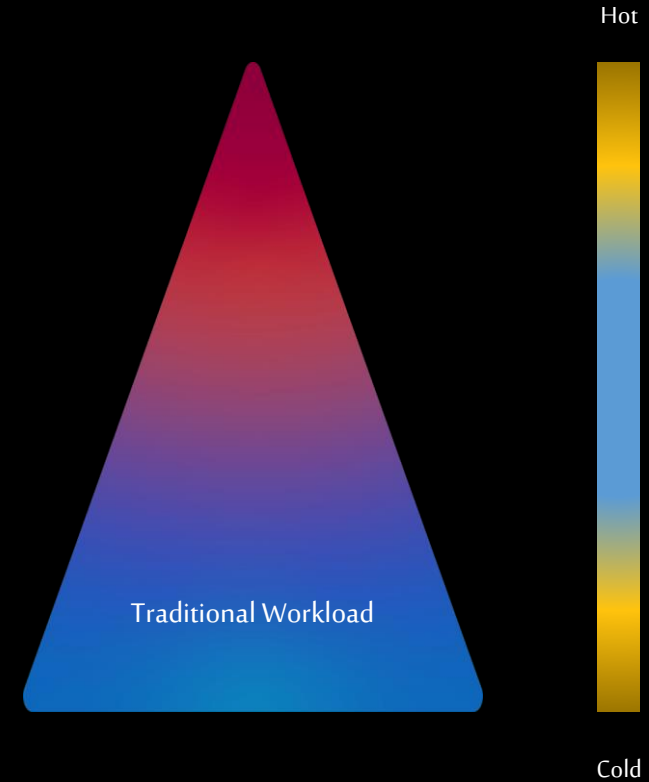
<https://storagedeveloper.org/conference/agenda/sessions/ssd-cxl-interfaces-brilliantly-stupid-or-stupidly-brilliant>

Memory Hierarchy

Keep hot data close to CPU using data locality



Memory Hierarchy



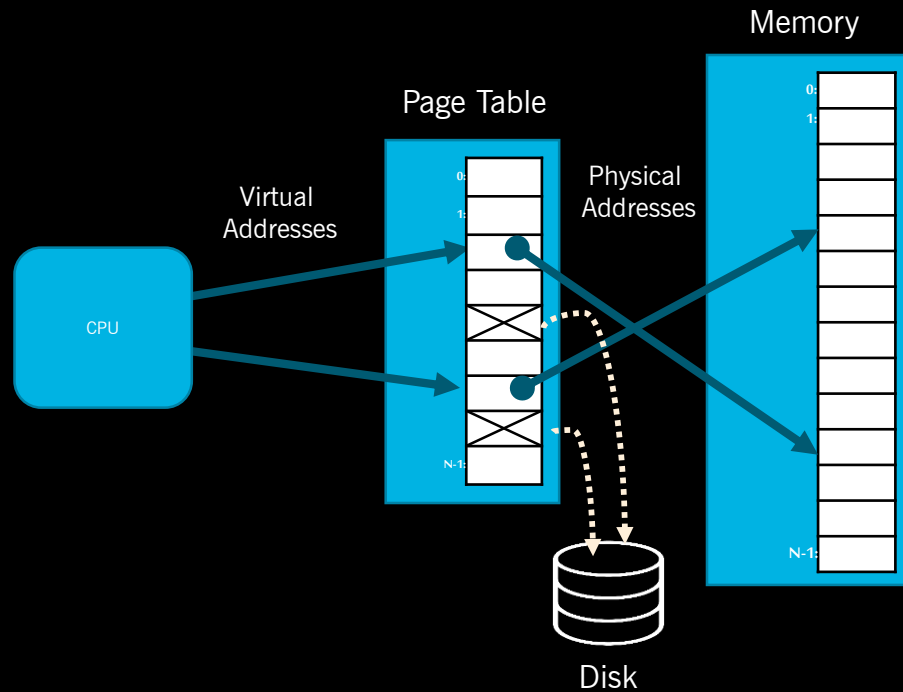
Traditional Workload

Needs (1): Secondary Memory

High overhead of virtual memory implementation

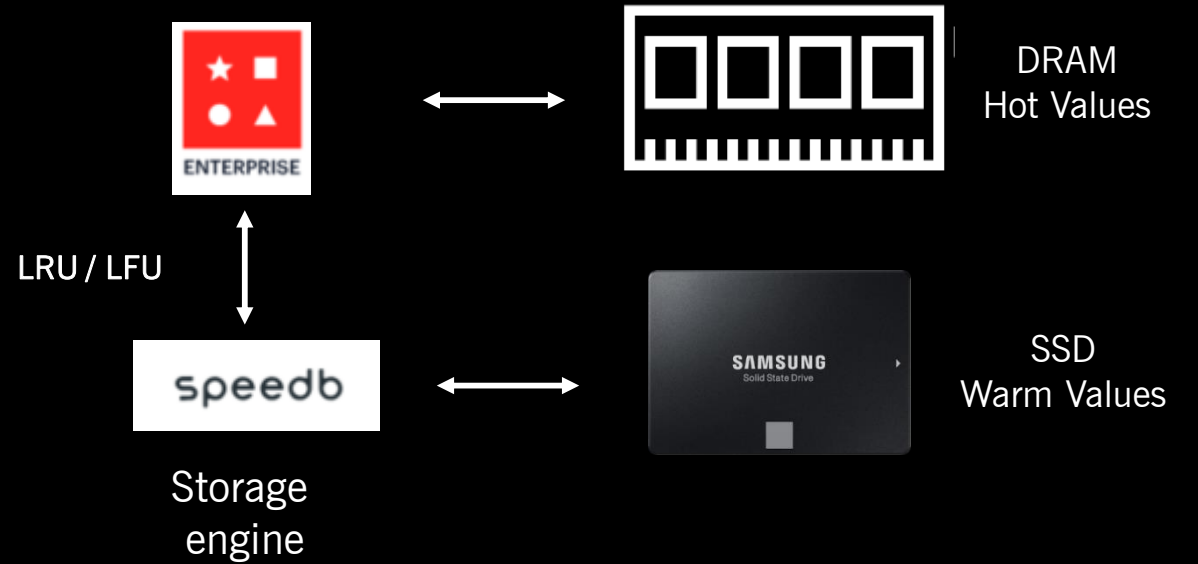
<OS-level>

Swap for memory extension on disk



<User-level>

Redis Auto Tiering for memory extension on SSD



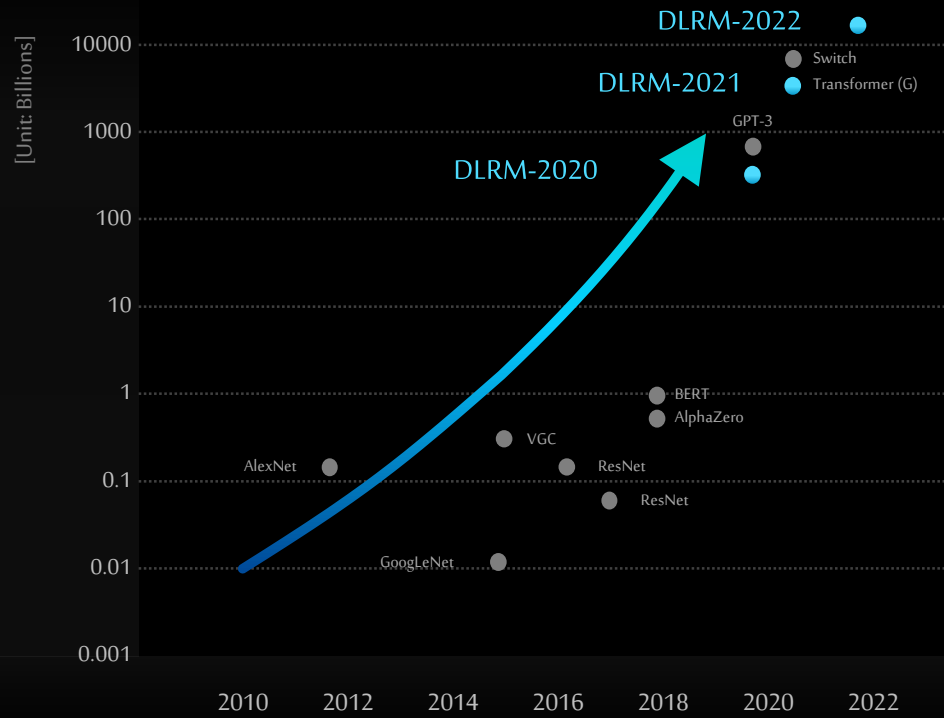
https://media.kingston.com/pdfs/MemoryandStorageBestPracticesforDesktopVirtualization_lr.pdf

<https://redis.com/wp-content/uploads/2023/08/redis-enterprise-auto-tiering-datasheet.pdf>

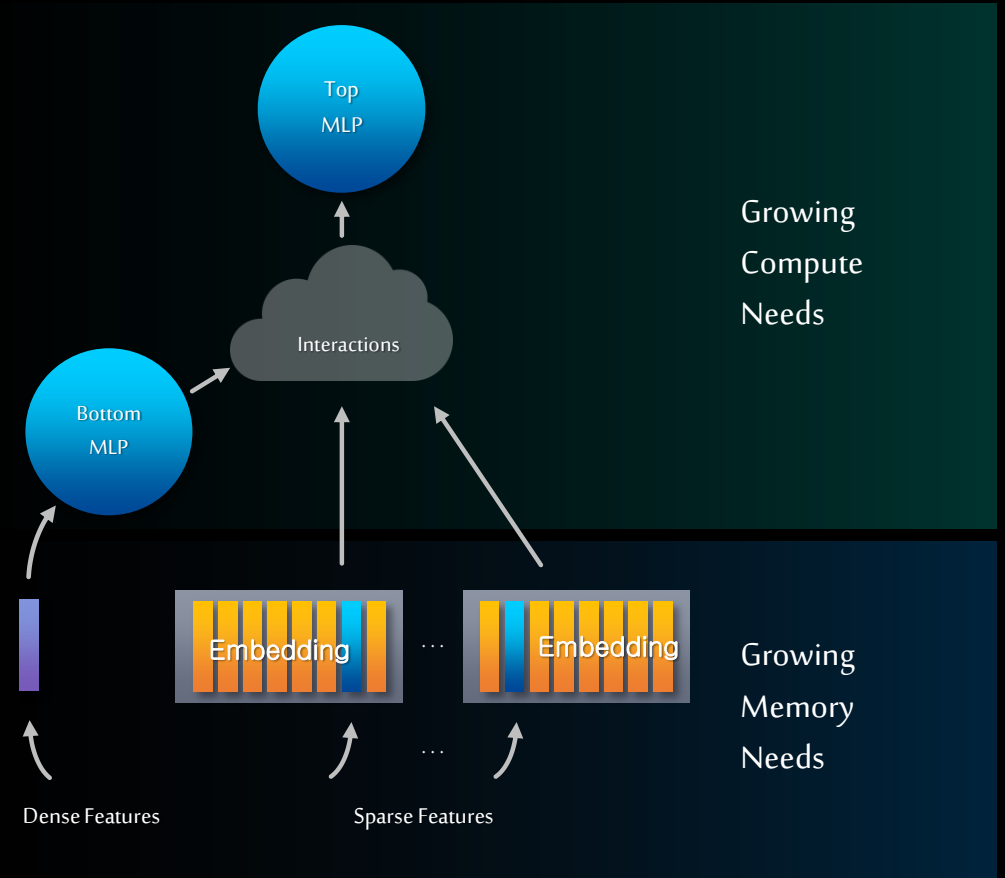
Needs (2): Fast Small IO

High overhead of IOs smaller than 4KB

DLRM size is rapidly growing

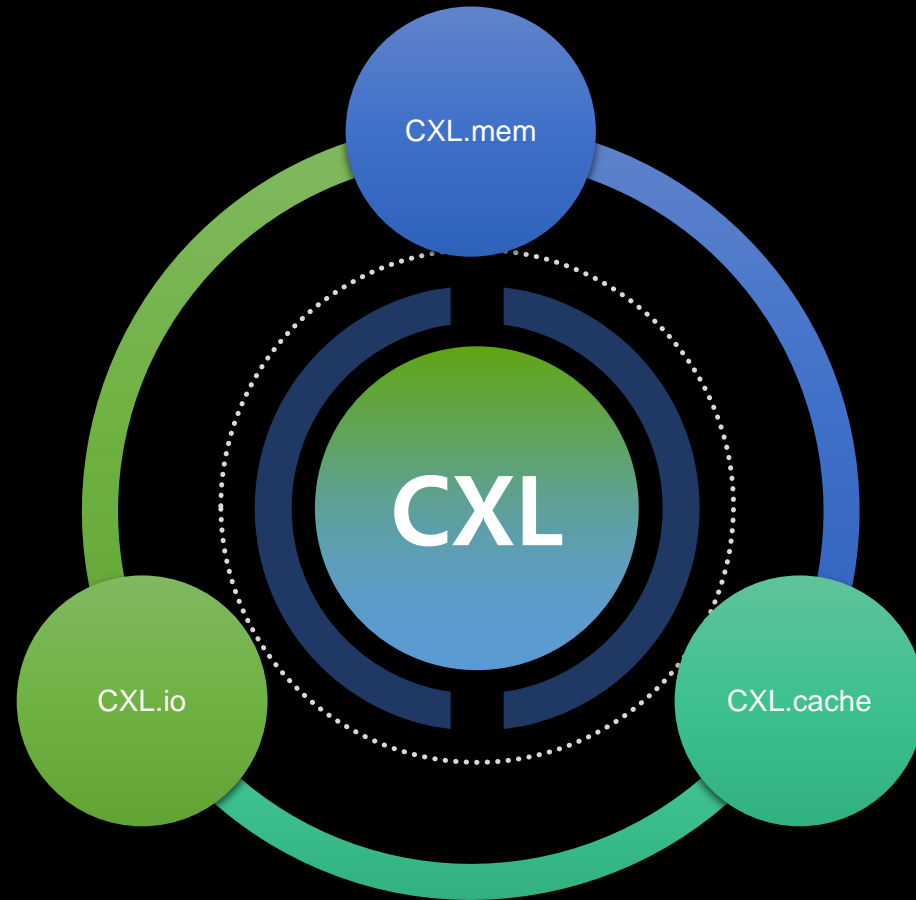


Source: Meta



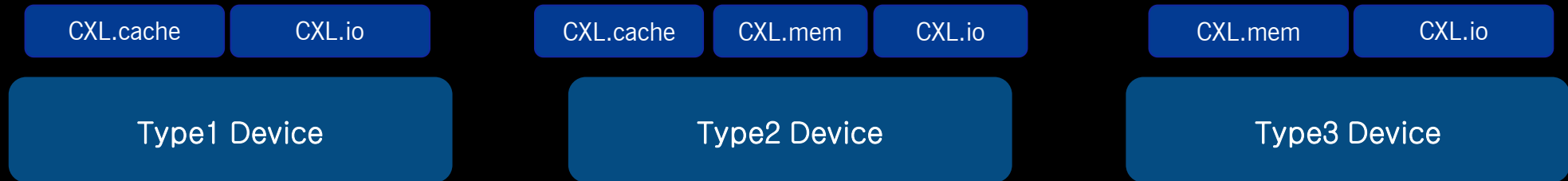
CXL (Compute Express Link)

Asynchronous blocking memory interface with optional coherency



CXL Device Types

Device types based on protocols, not functions



CMM-H™ (CXL Memory Module – Hybrid)

a.k.a Memory–Semantic SSD™

A Hybrid device of DRAM and NAND with CXL interfaces

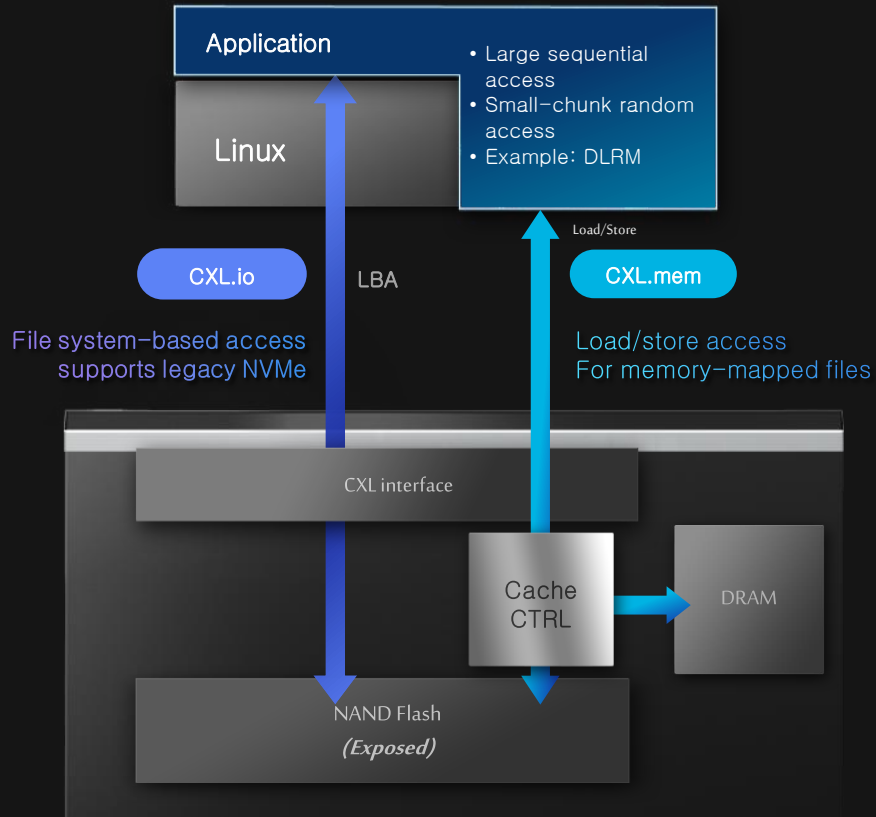
A new SSD that provides a *memory-like interface* based on CXL as a way to access NAND flash at the same time with the legacy block interface

- Byte-addressable
- Storage area mapped to physical address

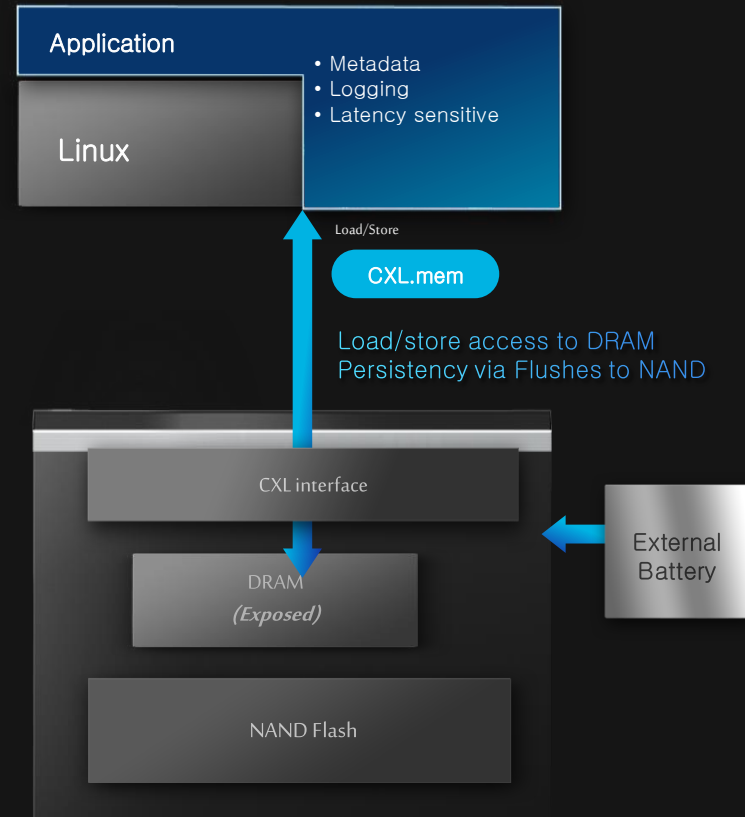


Concepts of CMM-H™

High-Performance Mode



Persistent Memory Mode



Memory Solutions with CXL

Memory Expander

CXL Type 3 device

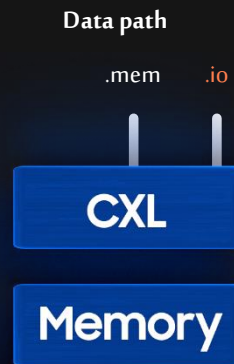
CXL device with high bandwidth and low latency without a long tail



Tiered Memory Solution

CXL Type 3 device

CXL device with .mem and .io as active data path



Accelerator Attached Solution

CXL Type 2/3 device

Accelerator with CXL interface



CMM-H™ (CXL Memory Module – Hybrid)

SAMSUNG

Expanding Capacity and Utilization of Memory for AI



Better System TCO

Larger capacity memory device at lower TCO
Best suited for Tiered Memory Solutions



Small Granularity Access

64-byte cache-granular fine grained access
to meet modern AI/ML workload needs



Persistent Memory Option

Speed comparable to DRAM with NAND storage backed and external battery power supply



CMM-H™ Architecture

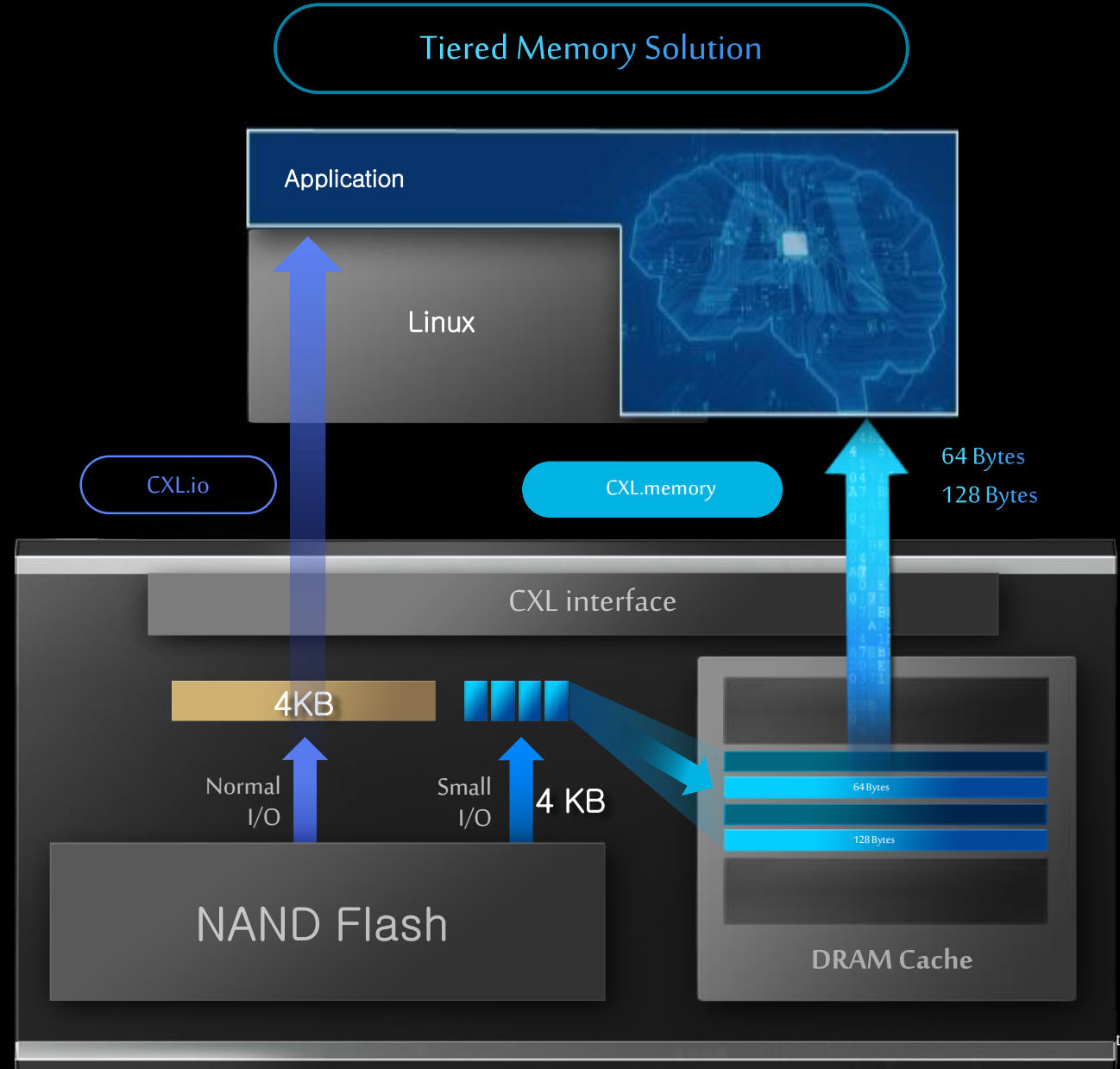
Optimized for AI Workload

CXL Protocol for AI

- Low latency enabled by CXL. mem protocol

Built-in DRAM Cache for AI

- DRAM cache to move **small-sized** data chunks suitable for AI/ML Applications
- Improve data store efficiency by writing data at the DRAM speed

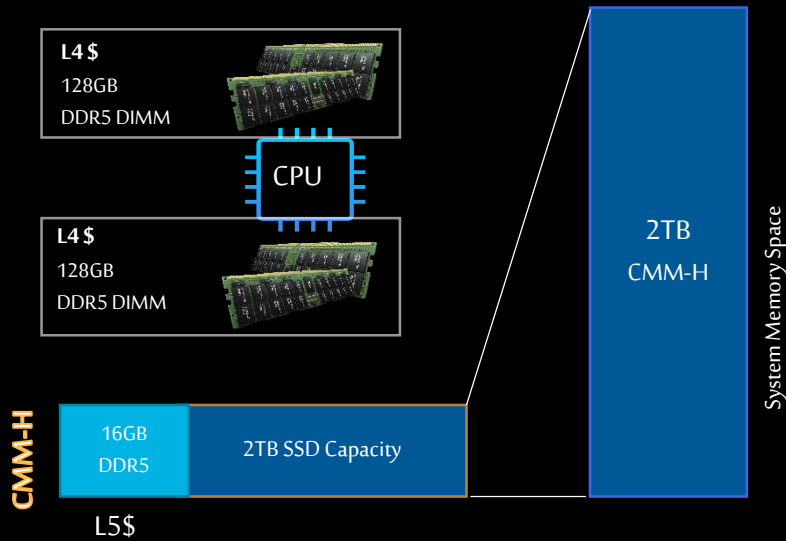


Secondary Memory Options

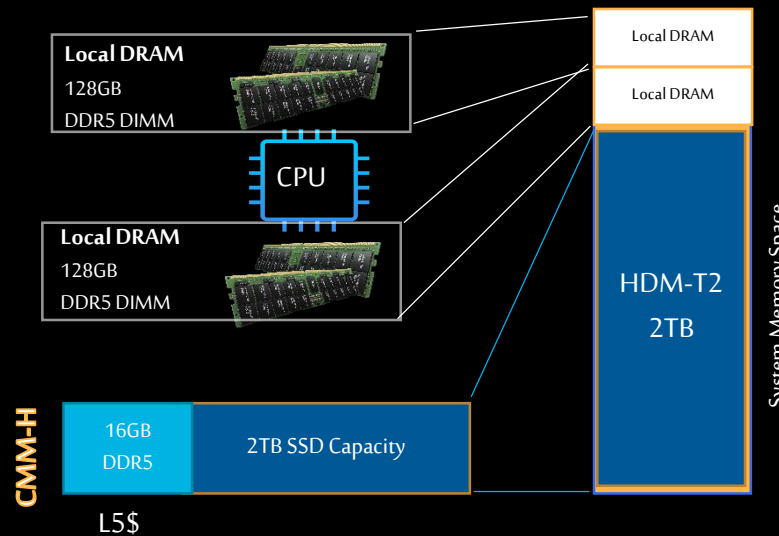
Example of Memory Configuration with TM Mode

- No page migration between storage and DRAM
- No extra I/O traffic for small size access
- Cacheable, H/W cache coherence protocol supported
- Supporting user-level prefetching

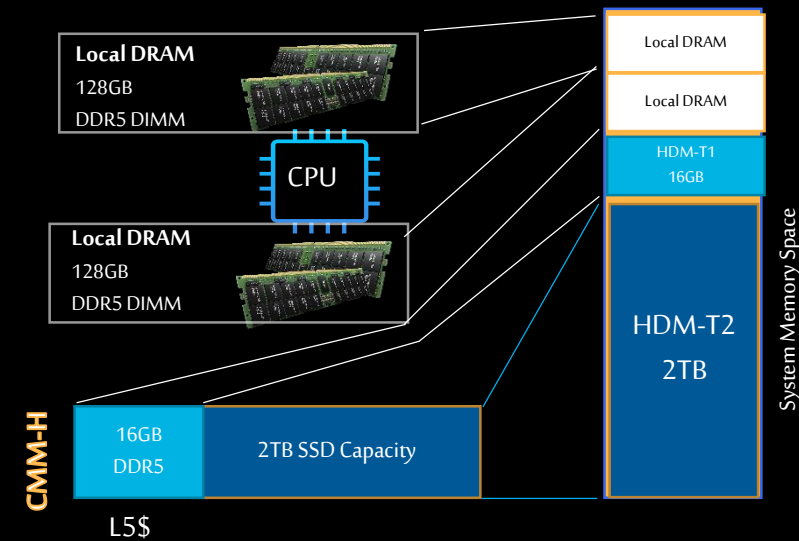
<2TB Main Memory (Case-1)>



<1-Tier Host-managed Device Memory (Case-2)>



<2-Tier Host-managed Device Memory (Case-3)>



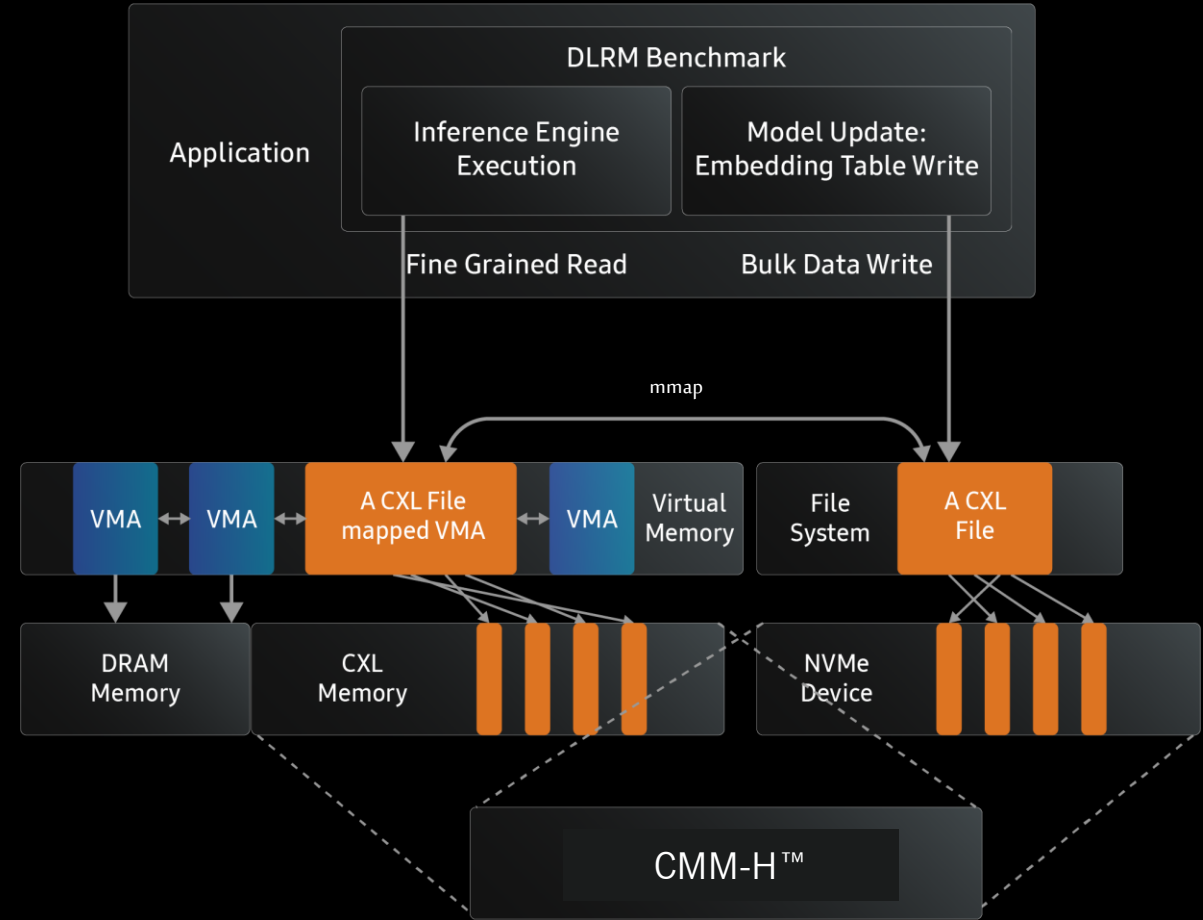
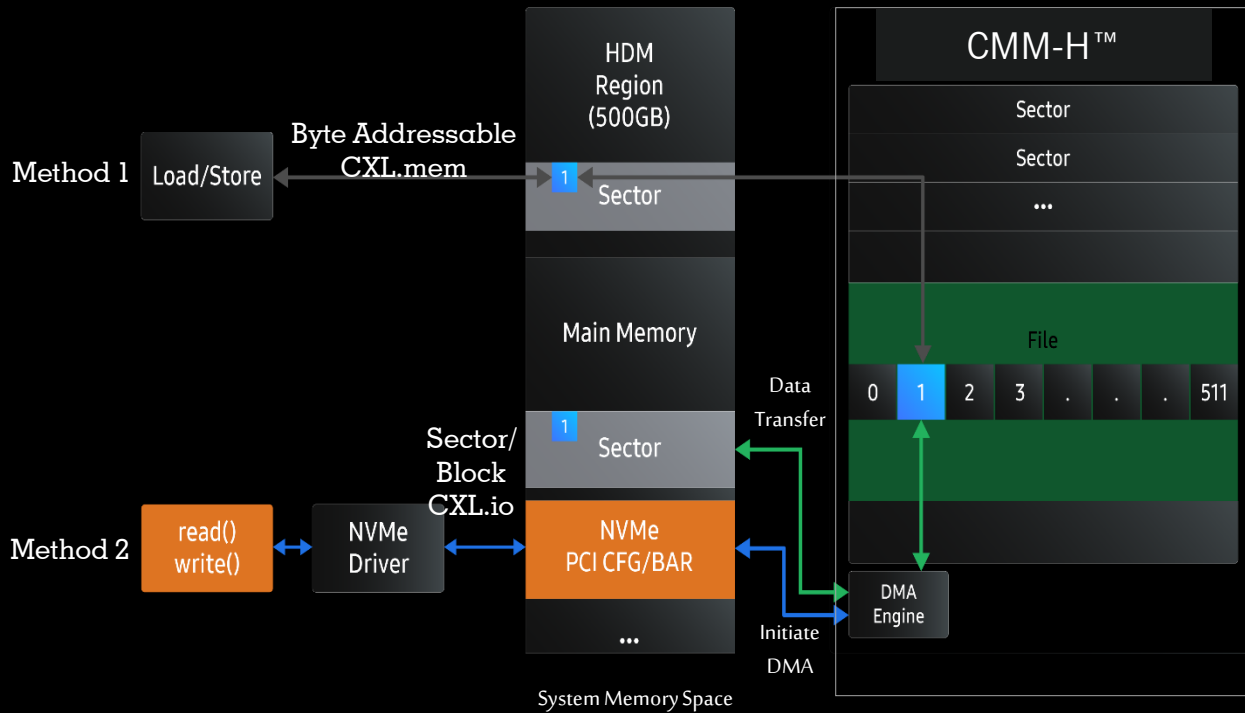
HDM: Host managed Device Memory

Fine-grain Access to Storage Data

Writes a file via NVMe interface (CXL.io), and performs *mmap()*

Reads the data via memory interface (CXL.mem)

*DLRM: Deep Learning Recommendation Model

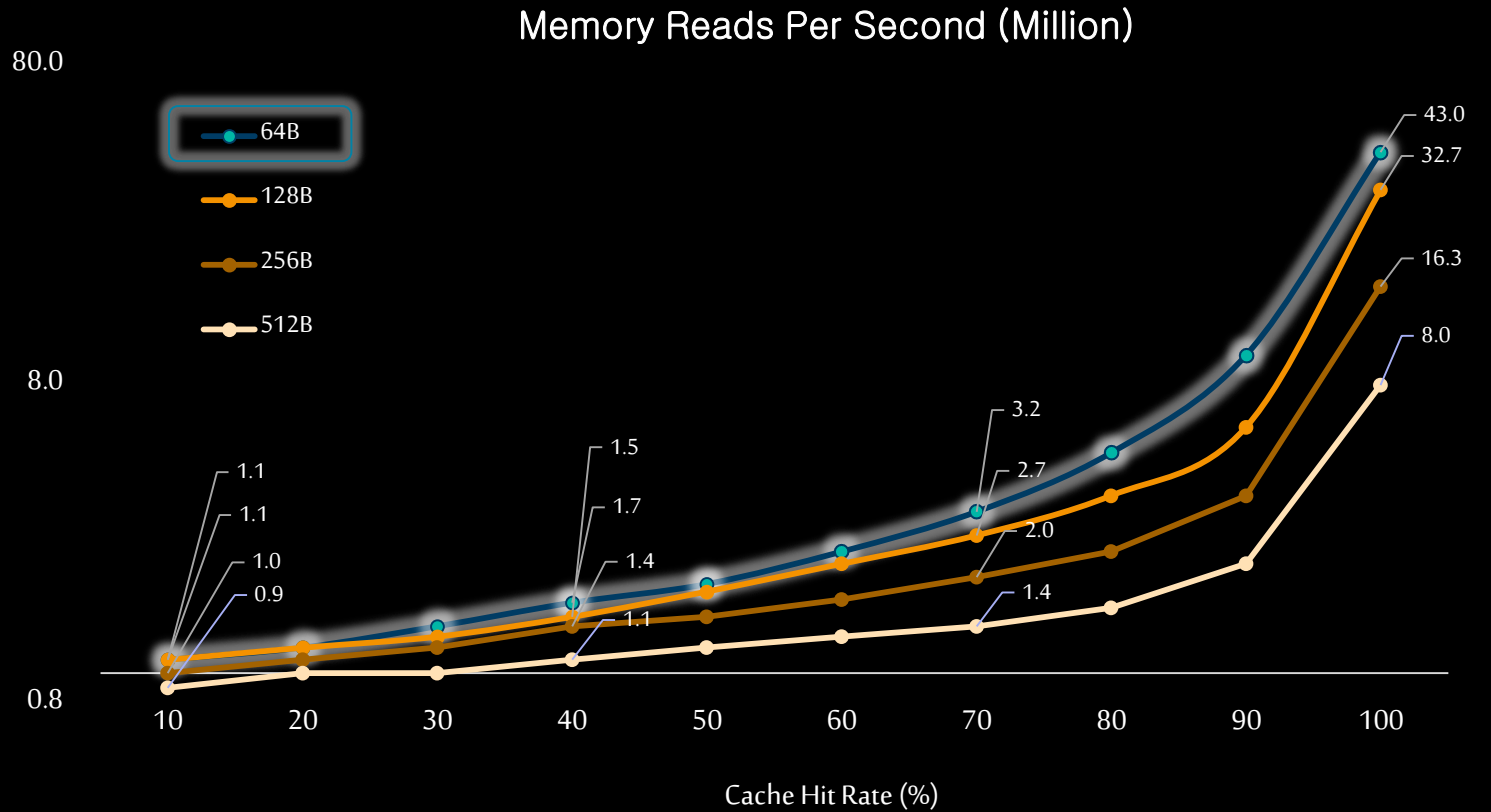


Tiered Memory

Secondary Memory Option 1 Performance

Key Features & Benefits

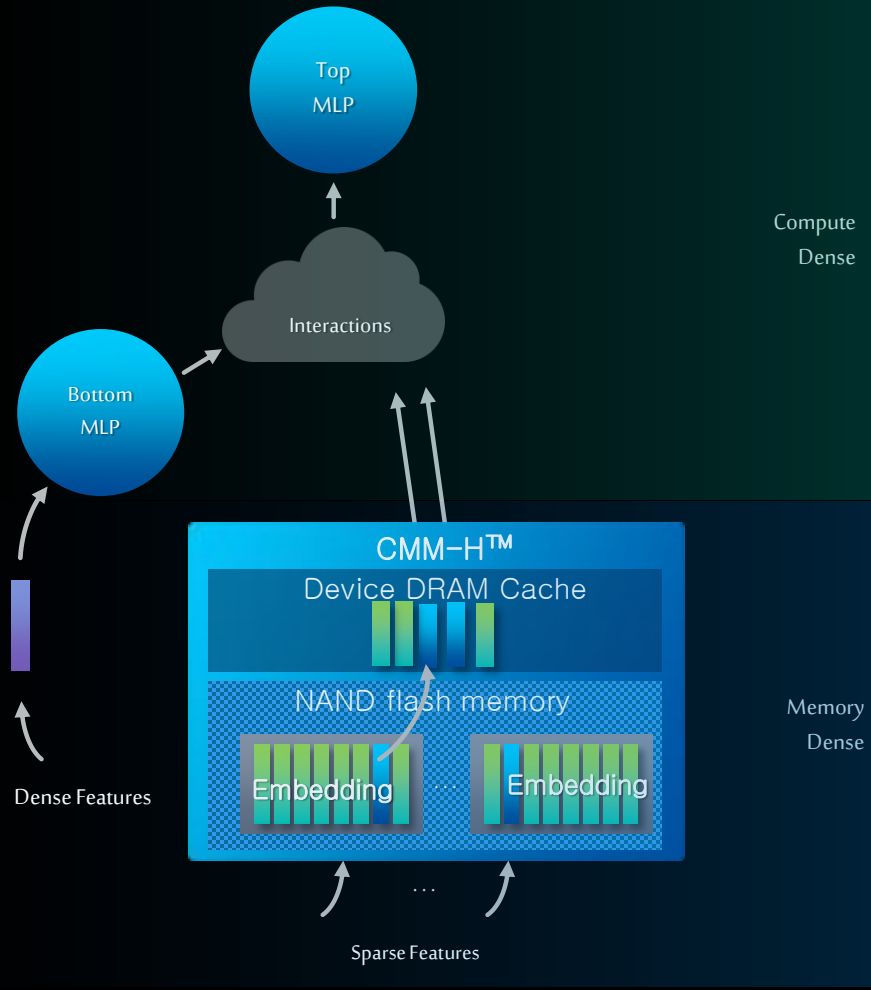
- Small granularity data access enable performance scales with cache hits
- Direct memory access advantage; no software cache overhead
- Large memory capacity at lower TCO



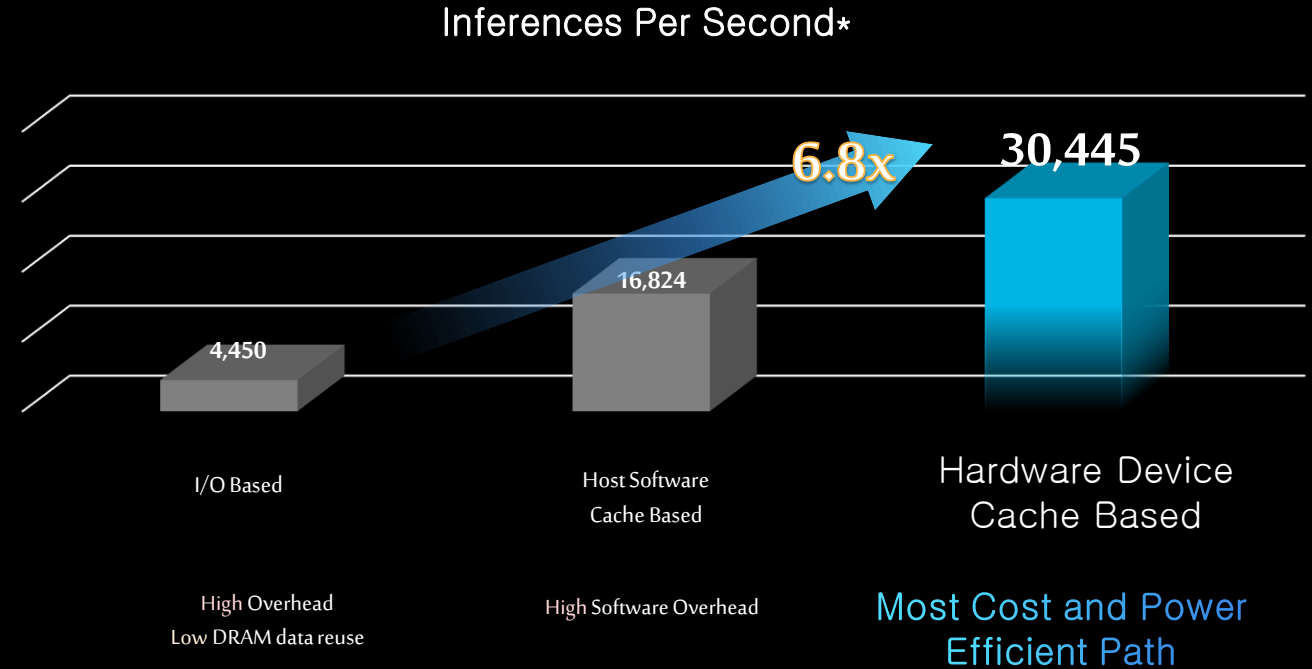
**Compared to PCIe Gen4 NVMe SSD

Efficient AI Recommendation system

DLRM Performance with Fast Small IOs



DLRM** performance (Meta)

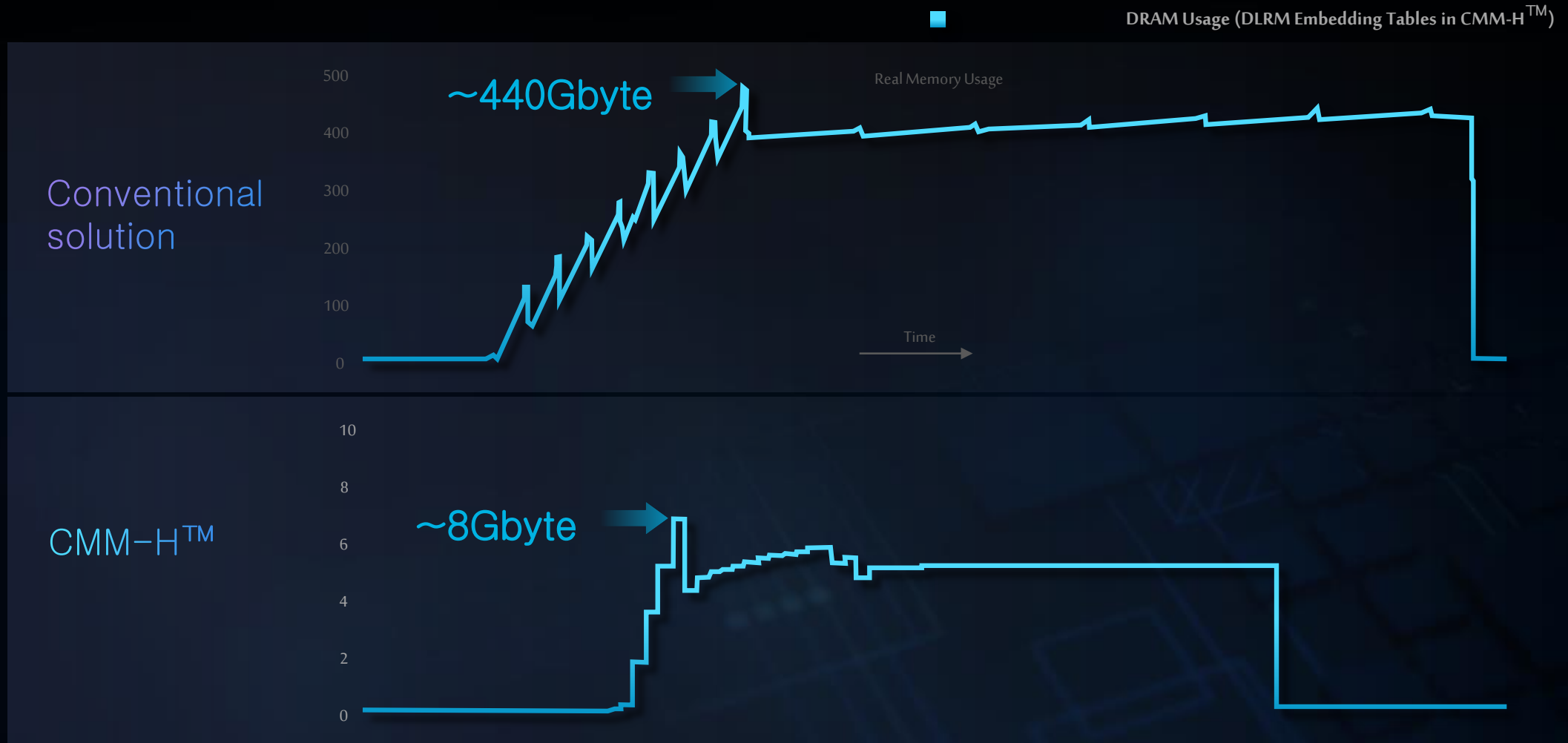


* Results based on publicly available [DLRM workload traces from Meta](#) and FPGA based PoC CMM-H™

** DLRM : Deep Learning Recommendation Model

DRAM Utilization

Comparison for Large AI Workload



Better DRAM Utilization with CMM-H™

Movie Recommendation System Demo

Original Movies Rated by User # 1000



DRAM-only DLRM Recommendations



MS-SSD Based Recommendations



Much Better Recommendation Accuracy both Visually and Numerically

DRAM-only

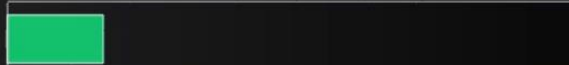
MS-SSD Based

Performance Comparable to DRAM (100% Hit Ratio)

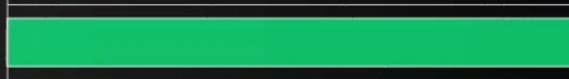
Relative Inference Time (lower is better)



1X

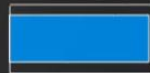


Cache Hit 1X

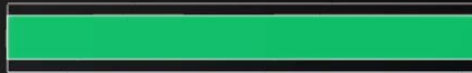


Cache Miss 6X

Recommendation Accuracy (higher is better)



1X



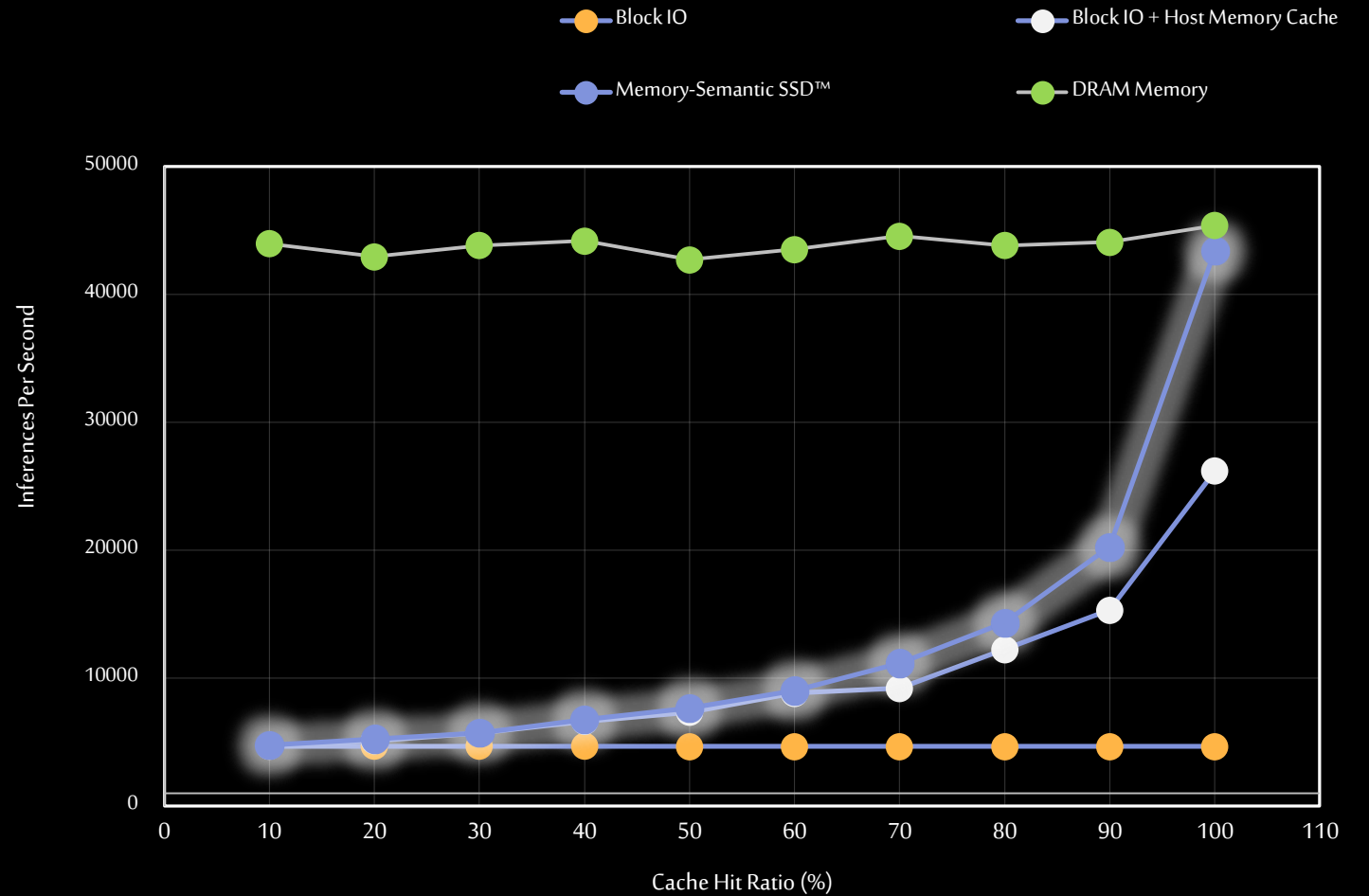
2.6X

End-to-End Performance

Managing in-device DRAM is the key!

Key Features & Benefits

- Close to DRAM end-to-end performance at a lower TCO*
- Up to ~10x better end-to-end performance with FPGA-based PoC**



*When 100% hit ratio

**Compared to PCIe Gen4 NVMe SSD

Thank You