SAMSUNG

CXL for Data Centric Computing

CXL Hybrid Memory Modules for Memory-Driven AI Applications

David McIntyre

Director, Product Solutions Planning
Samsung Semiconductor Inc.

Large Language Models Continue to Grow in Size

- **Model Complexity**
  - Larger and more complex models
  - GPT- MOE: 1.8T parameters
- **Data and Resources:**
  - Exponential growth in training and data
  - Compute requirements increasing WITH Memory
- **Efficiency Approaches**
  - Improved training techniques
  - Selecting the right models
  - Scalable models to address application requirements

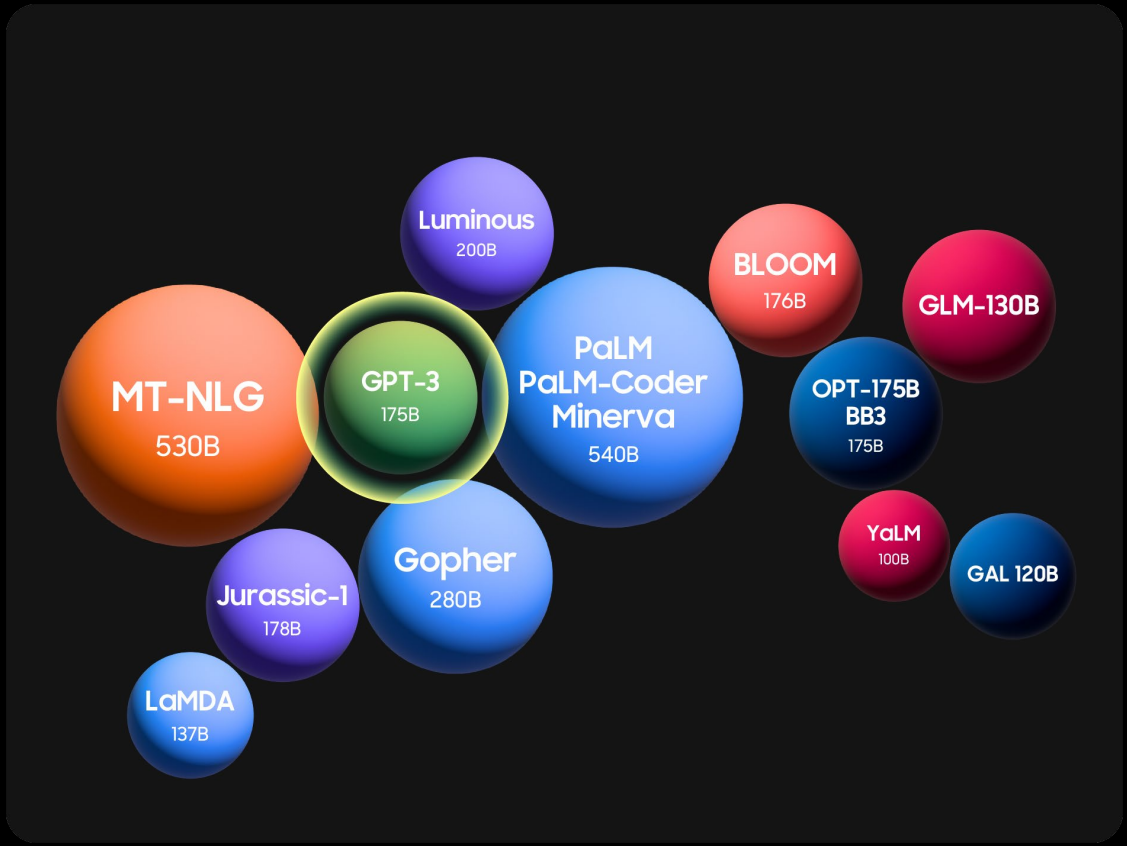
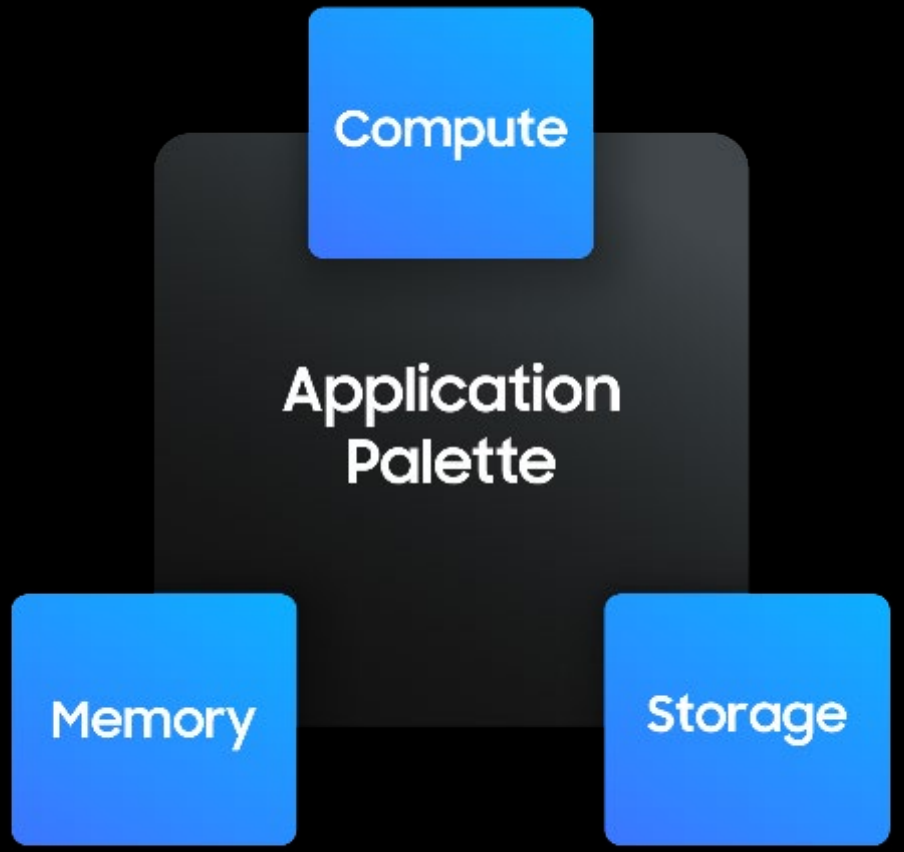


Figure 1: Exponential Growth of number of parameters in DL Models

SAMSUNG

Balancing Application-Driven Resources

- Local Orientation
- Application Specific
- Scalable Performance

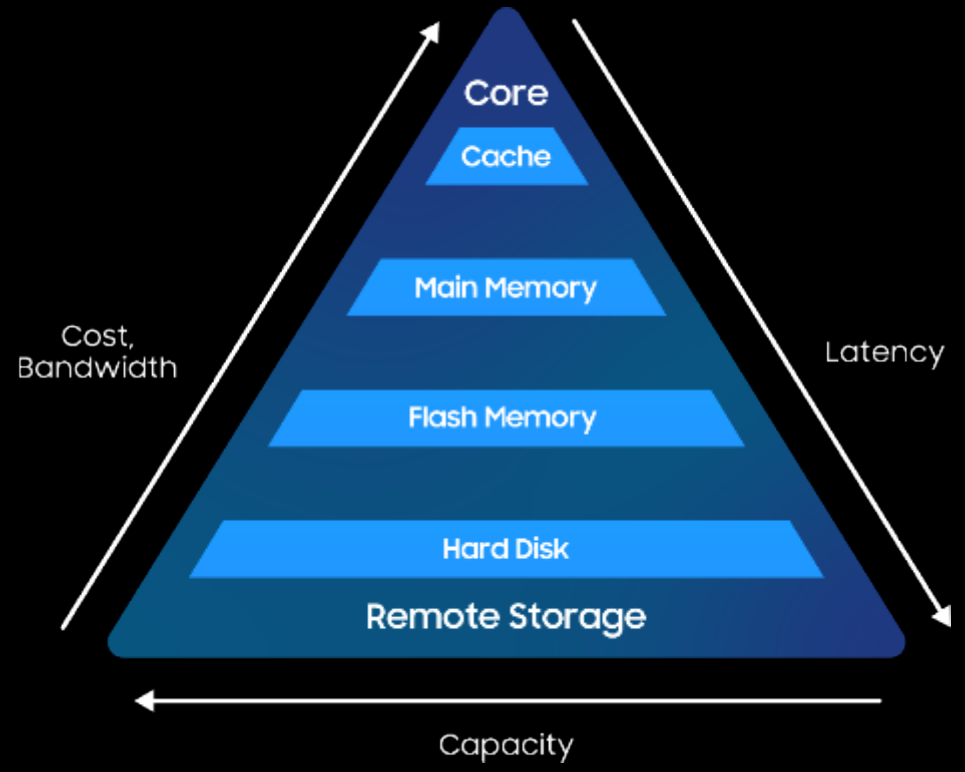


- **HBM Bandwidth**
- **Scalable capacity**
- **Tiering and Persistence**

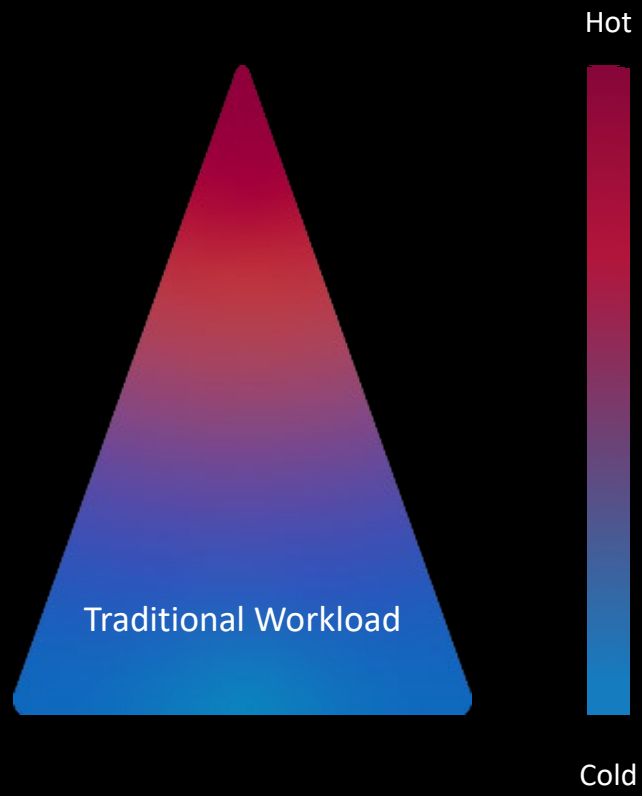
- Scalable capacity
- Tiered Performance
- Block/File/Object

Memory Hierarchy

Keep hot data close to CPU using data locality



Memory Hierarchy

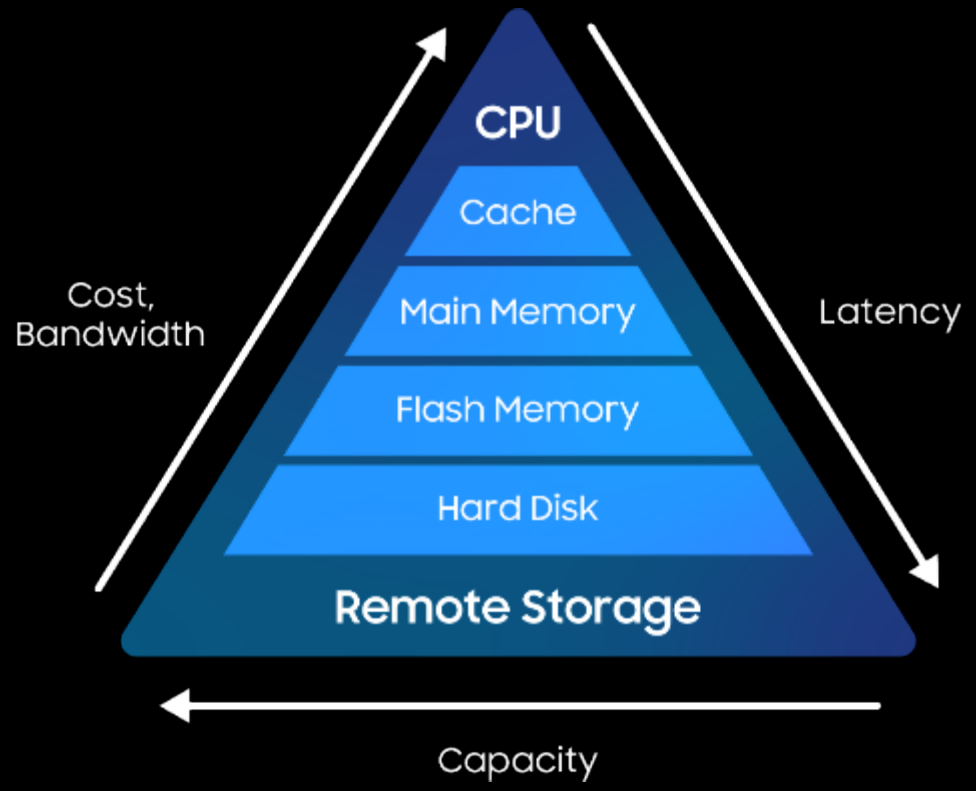


Traditional Workload

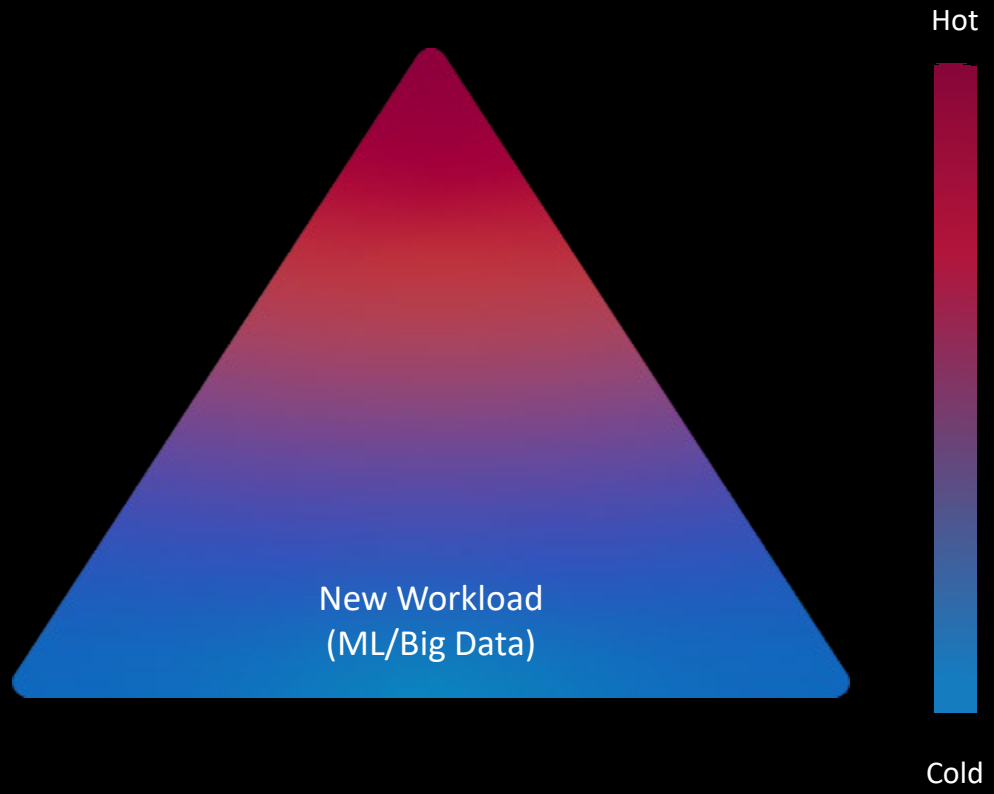
SAMSUNG

Memory Hierarchy Disparity for Modern Workloads

Not all workloads exhibit the conventional pattern of data locality



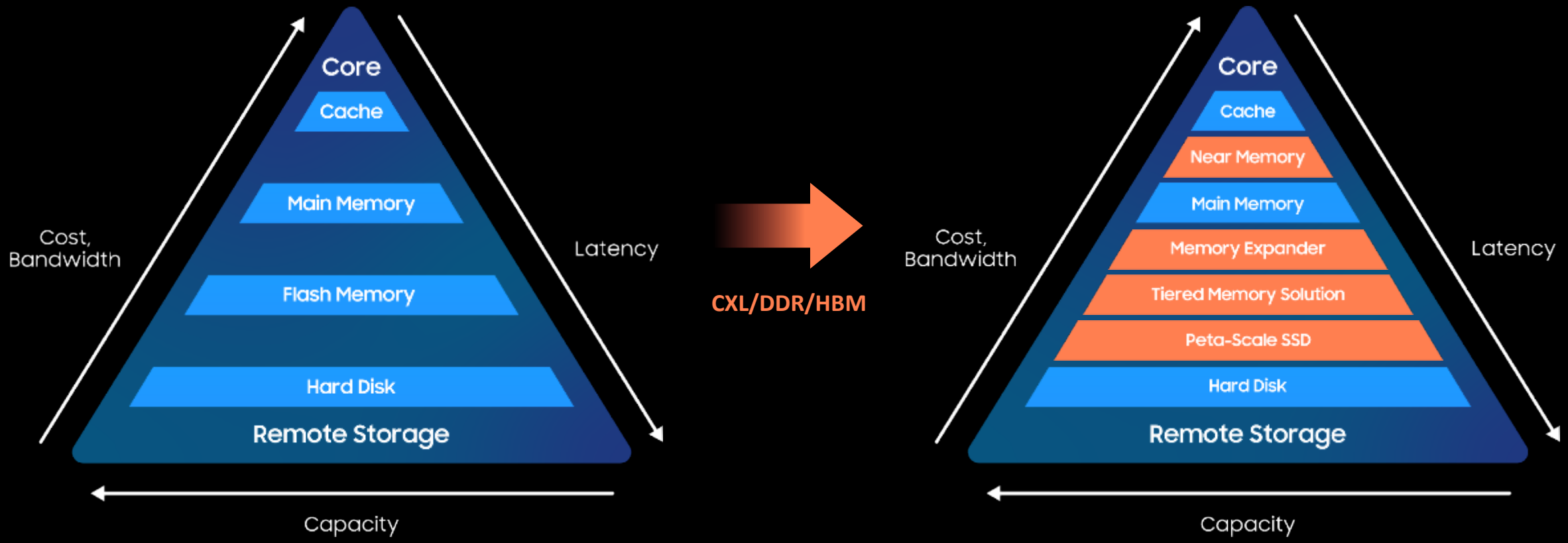
Memory Hierarchy



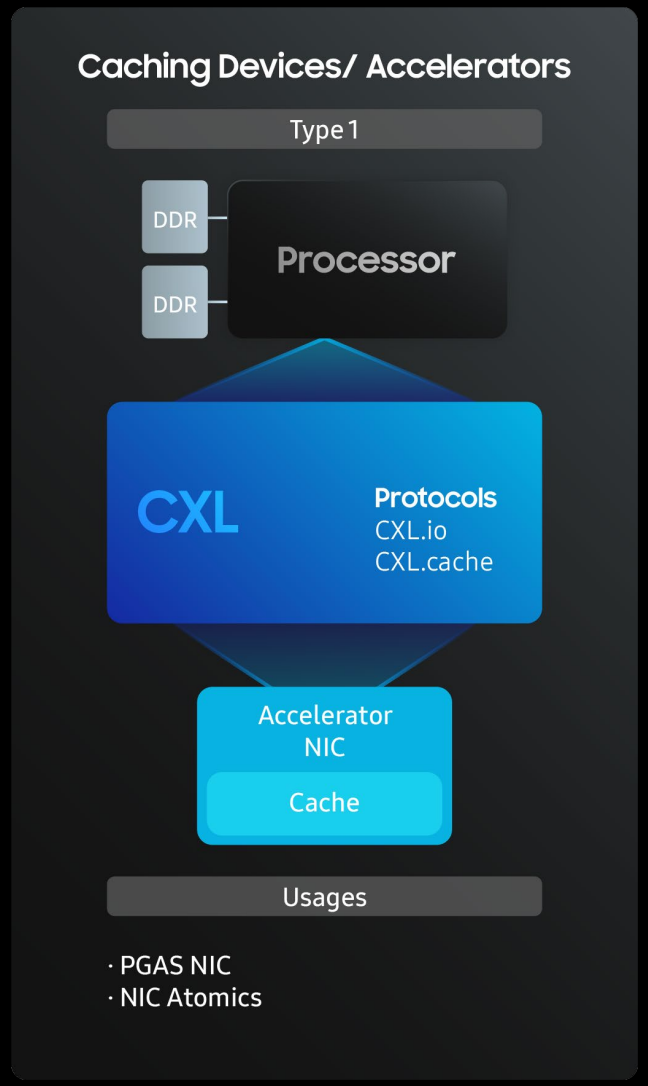
Modern Workload

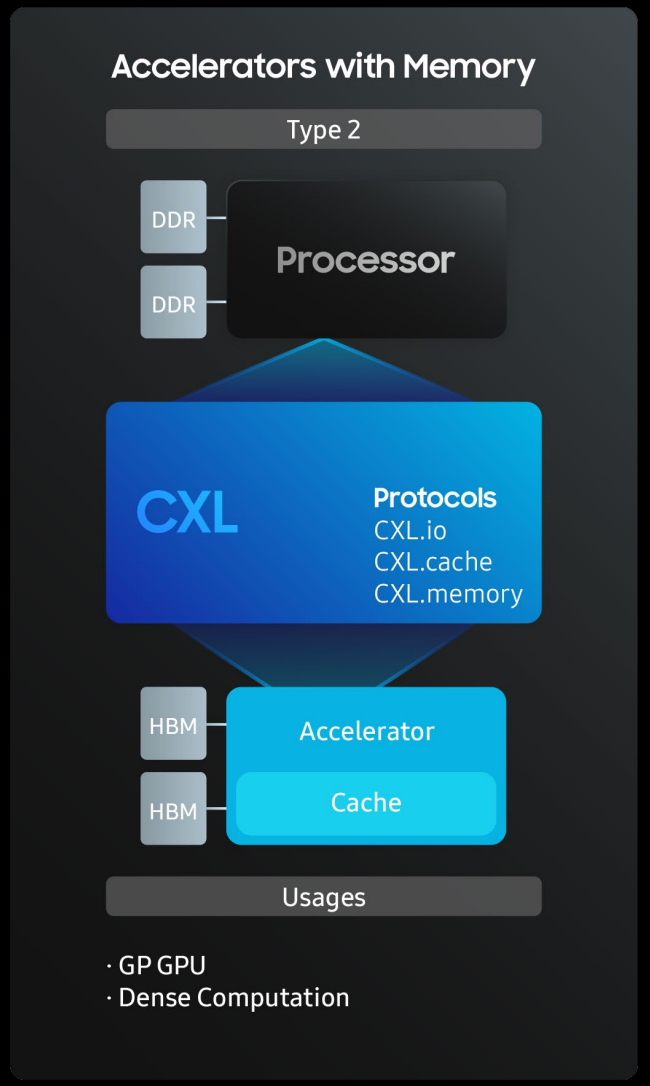
New Memory Hierarchy

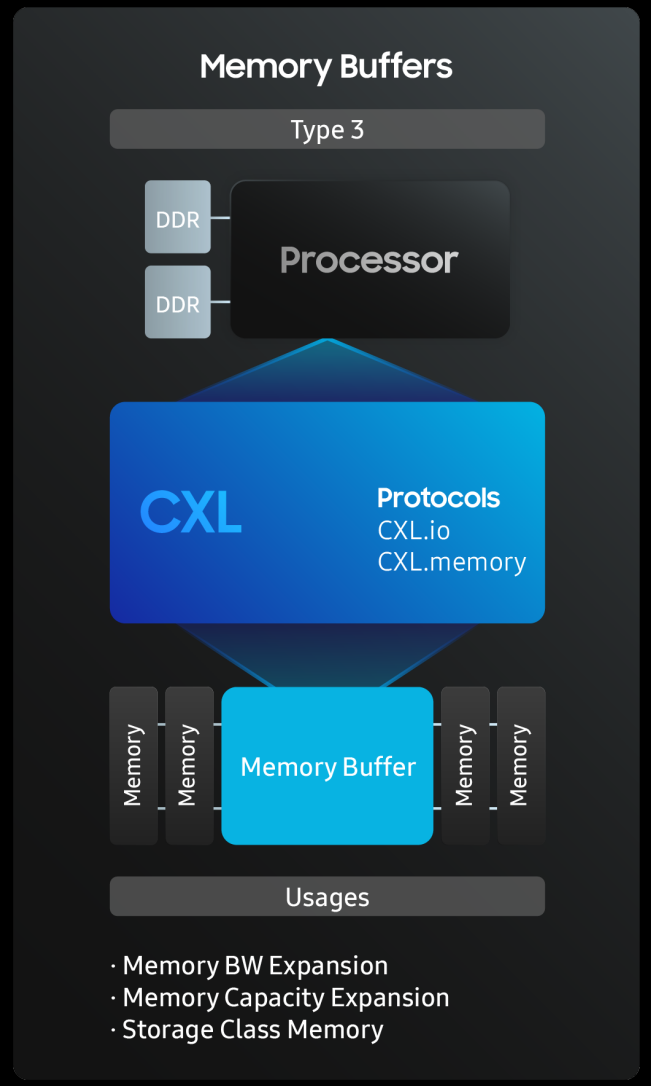
Deeper and more efficient memory hierarchy to fill the performance gap



CXL Device Types







Re: computeexpresslink.org

SAMSUNG

Samsung CXL Memory Module Device Portfolio

CMM-D
Memory Expander

CXL Type 3 device

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

CMM-H
Tiered Memory Solution

CXL Type 3 device

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

CMM-HC
Accelerator Attached

CXL Type 2/3 device

Accelerator with CXL interface

SAMSUNG

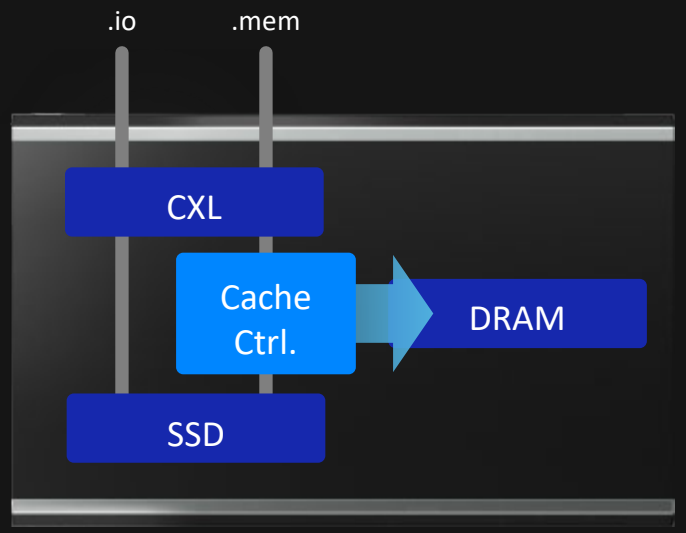
CXL Compared with PCIe Adoption Pathway

Attribute	PCIe	CXL	Comments
Spec 1.0 to 3.1	7 Years	4.5 Years	CXL benefits from established PCIe link layer
OEM Adoption	PCI to PCIe	CPUs released	• CXL spec and use cases closely coupled • Application targeting is the priority
Market Timing	Performance-based Evolution	TCO Optimization	• Memory-bound Workloads driving the opportunities for memory Expansion, Pooling and Tiering. • Supports Composable infrastructure and Heterogeneous Compute
Ecosystem	CPU and OEMs	Solutions Orientation (overall infrastructure)	Industry Alignment to three protocols

CMM-H Modes

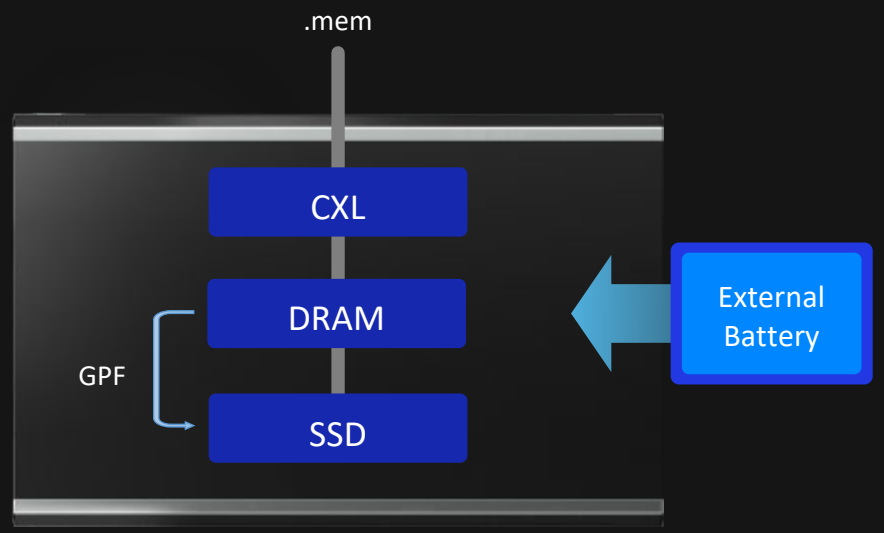
Tiered Memory

With .mem and
DRAM Cache



Persistent Memory

With CXL GPF
(Global Persistent Flush)
and external battery



CMM-H™ Architecture

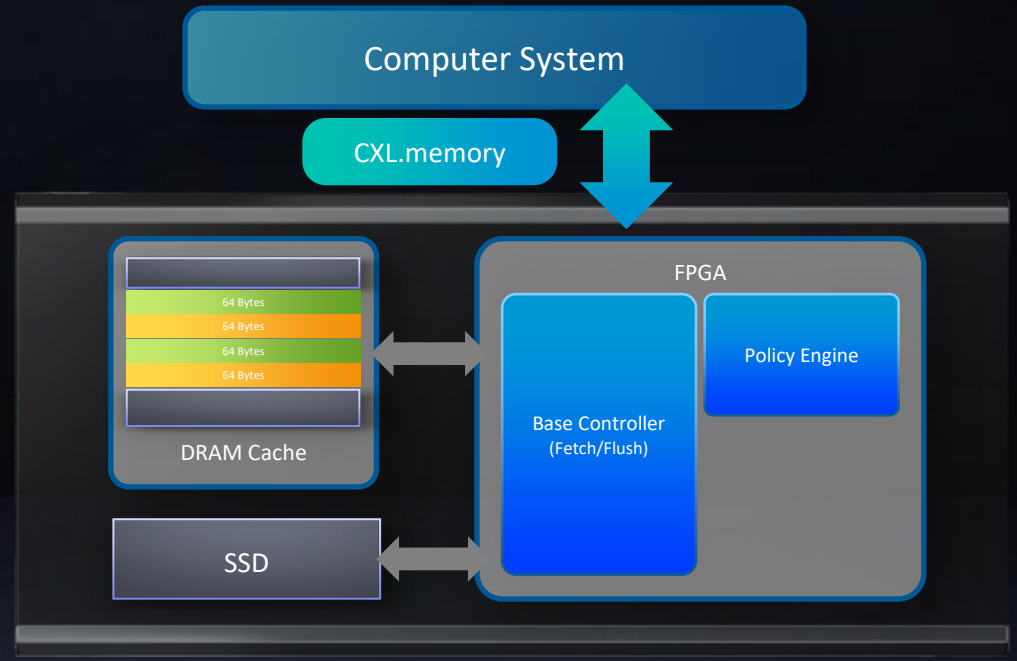
CXL load/store protocol

- Designed to be low latency

Built-in local DRAM cache

- DRAM cache to load/store small-sized data chunks
- Improve data store efficiency by accessing data at the DRAM speed

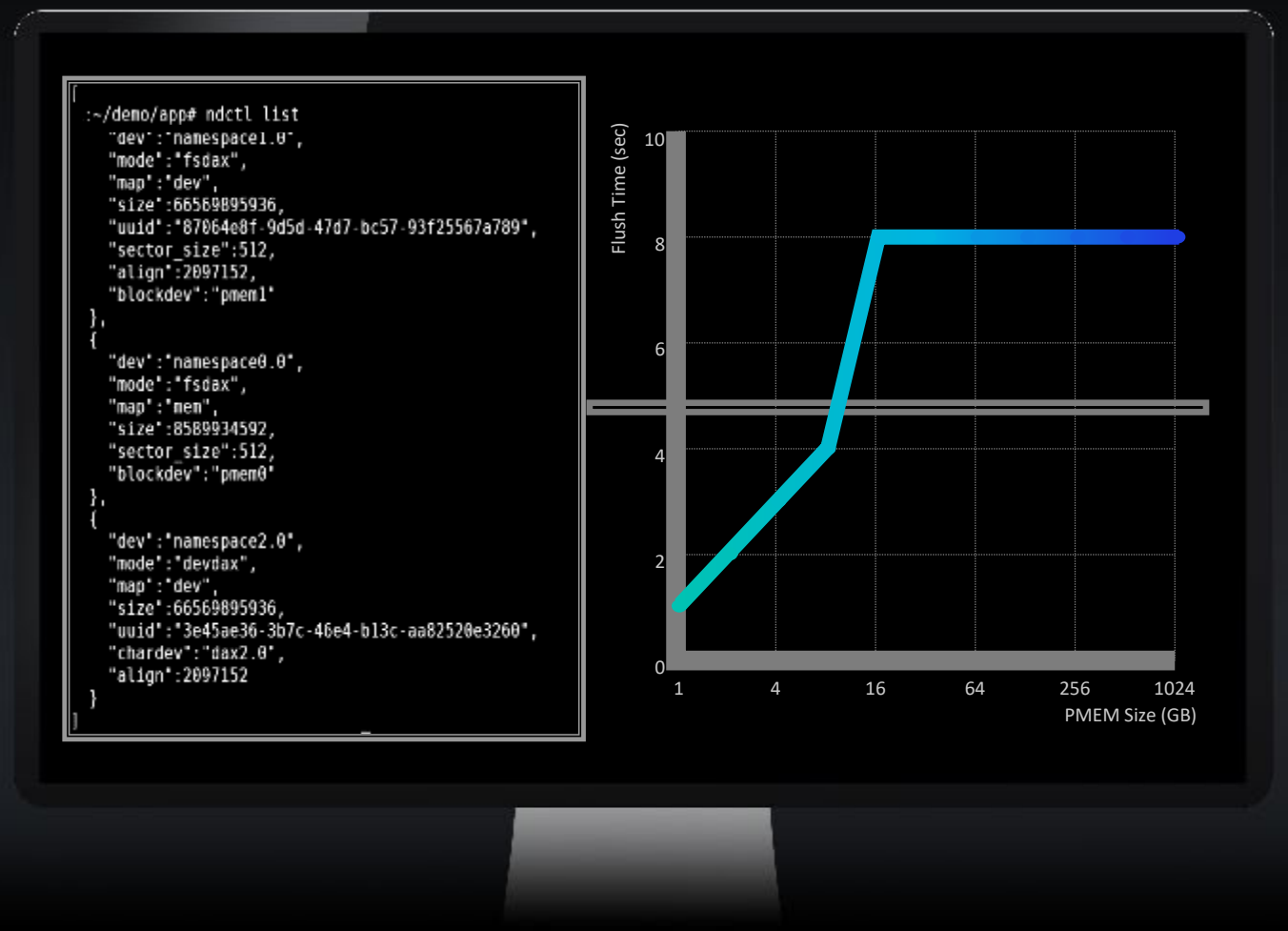
CMM-H Architecture



CMM-H : Persistent Memory (PMEM)

Key features & benefits

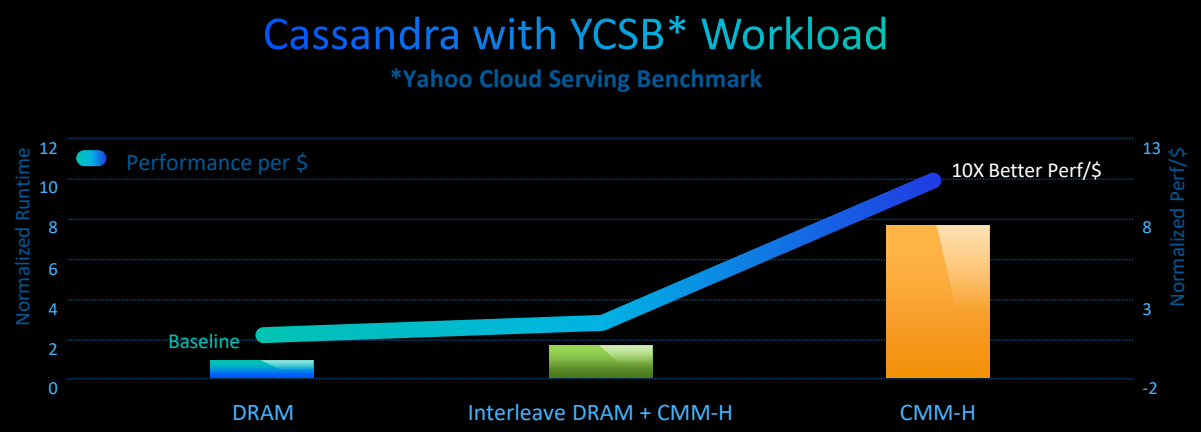
- Large Capacity Persistent Memory
- Orchestrated by NVDIMM framework
  - Uses same PMEM NDCTL utility
  - Compatible with NVDIMM code
- Supports CXL Global Persistent Flush (GPF)
- Prevents data loss in the case of sudden power outage

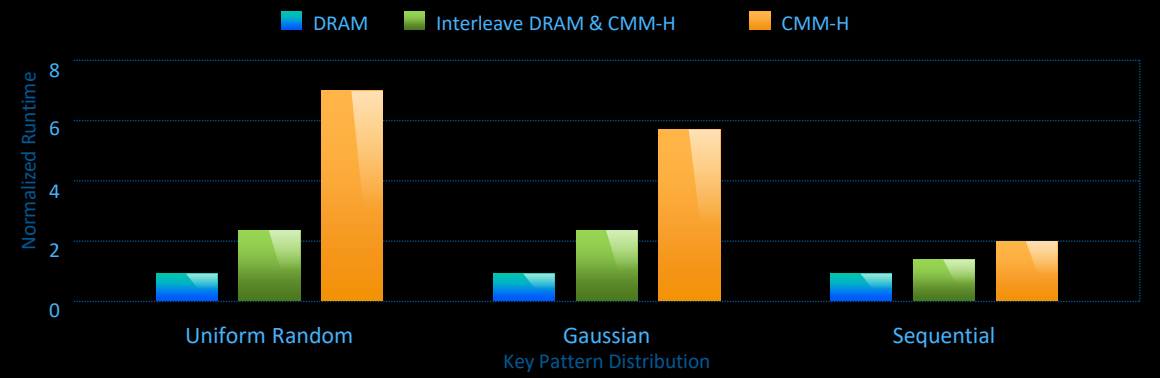


Application Examples

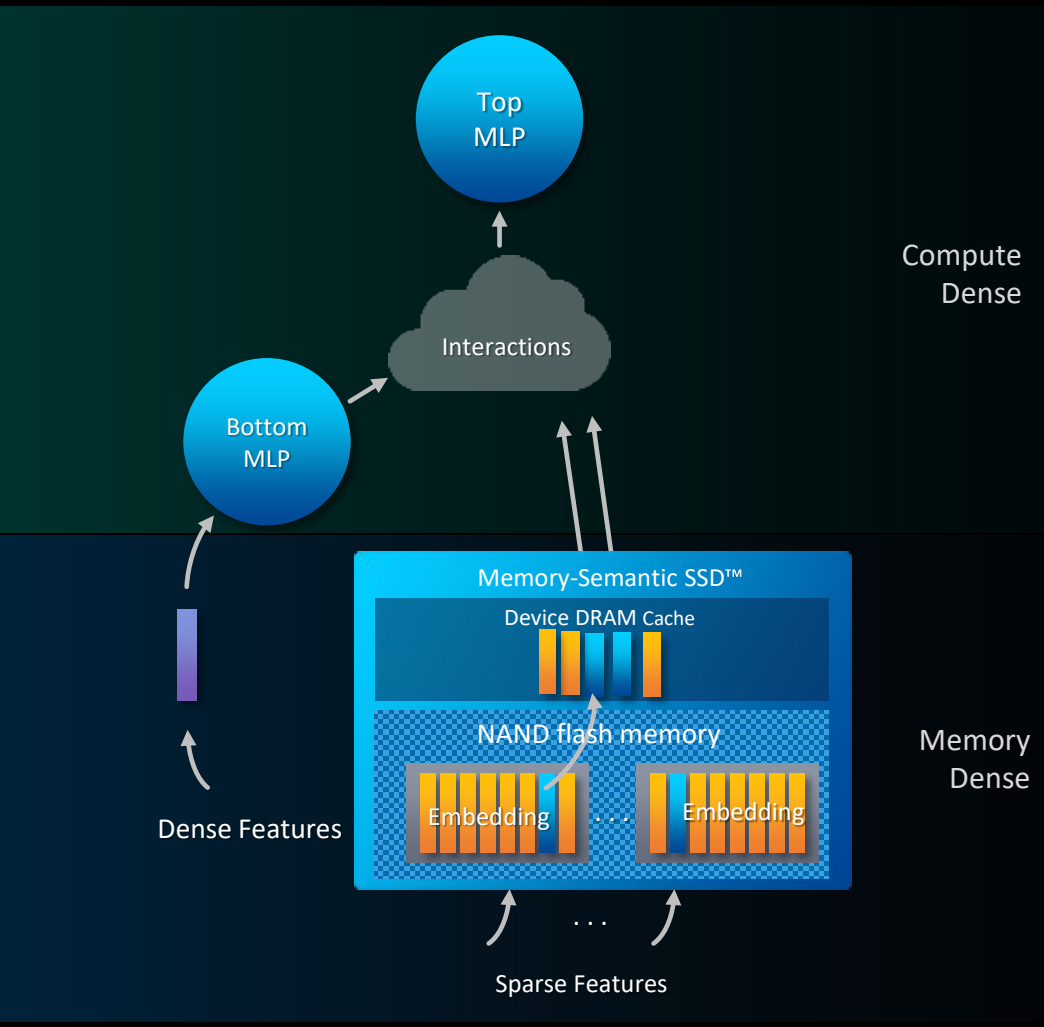
No changes in application required

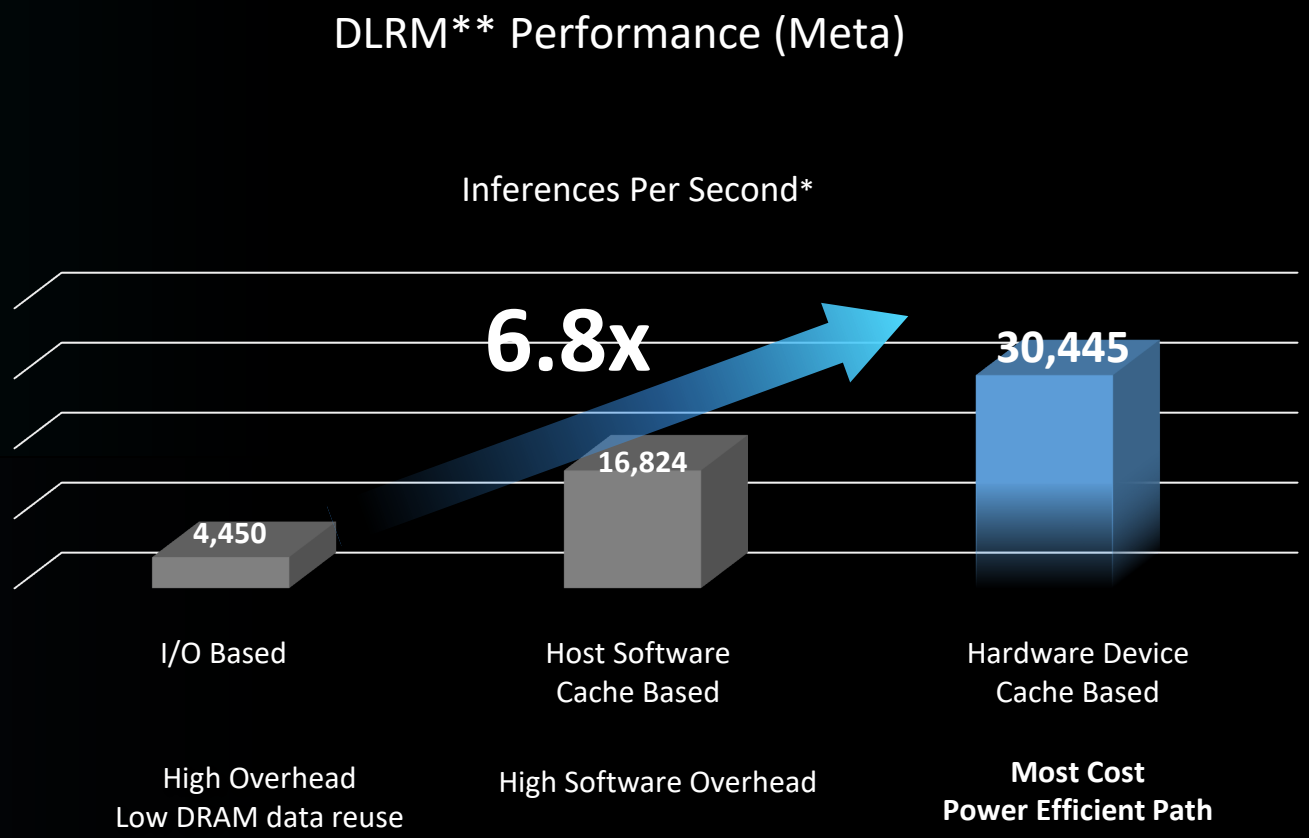
- Cassandra
- Redis
- GreenPlum
- DLRM
- KVM
- Hammer DB
- Graph500





Efficient AI Recommendation system: CMM-H



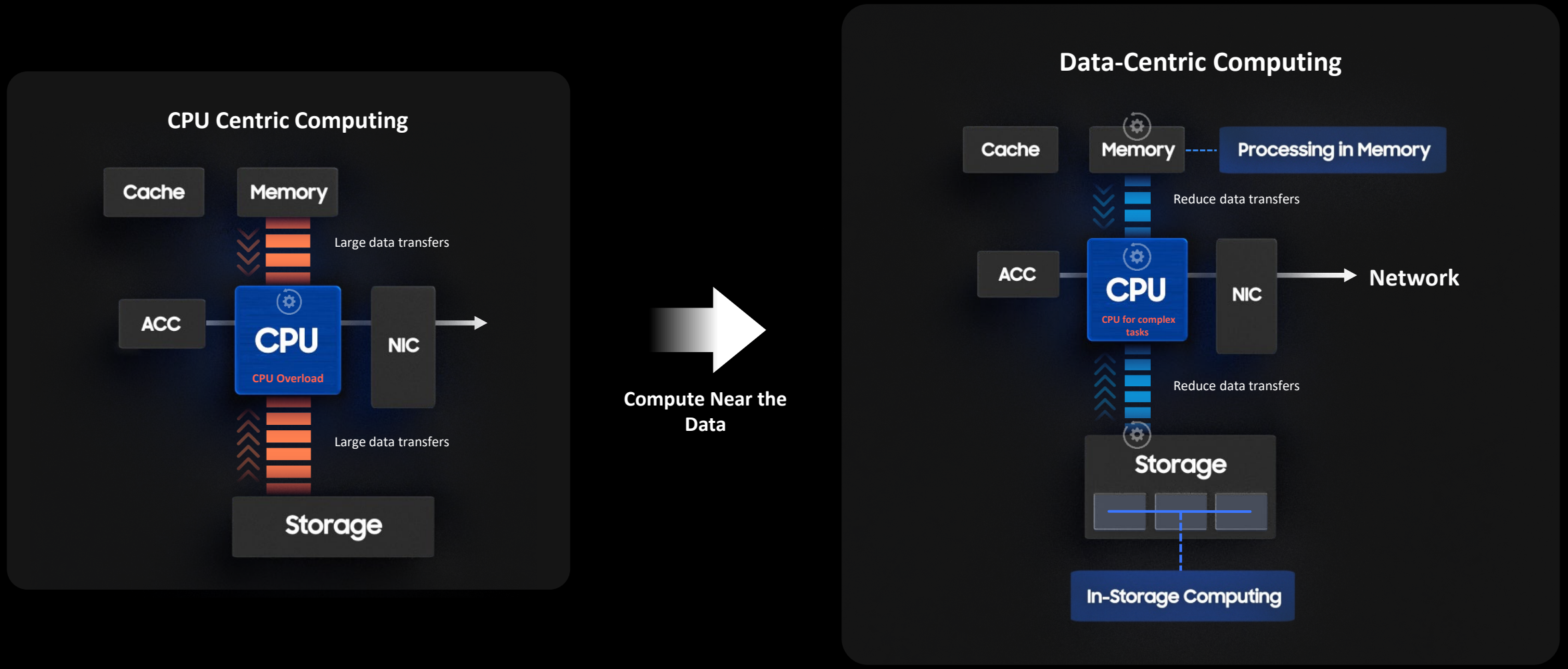


* Results based on publicly available DLRM workload traces from Meta and FPGA based PoC Memory-CMM-H™

** DLRM : Deep Learning Recommendation Model

Data-Centric Computing Concept

Move the computation to the data for large datasets



SAMSUNG

CXL Memory Module Device Types

CMM-D

Memory Expander

CXL Type 3 device

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

CMM-H

Tiered Memory Solution
Persistence

CXL Type 3 device

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

CMM-HC

Accelerator Attached
Solution

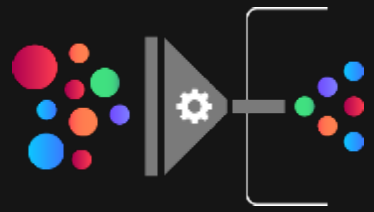
CXL Type 2/3 device

Accelerator with CXL interface

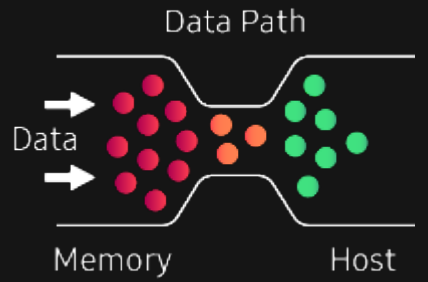
Data-centric Computing Benefits

Power-optimized scalable processing for large data

Low Power Computing



Data Reduction



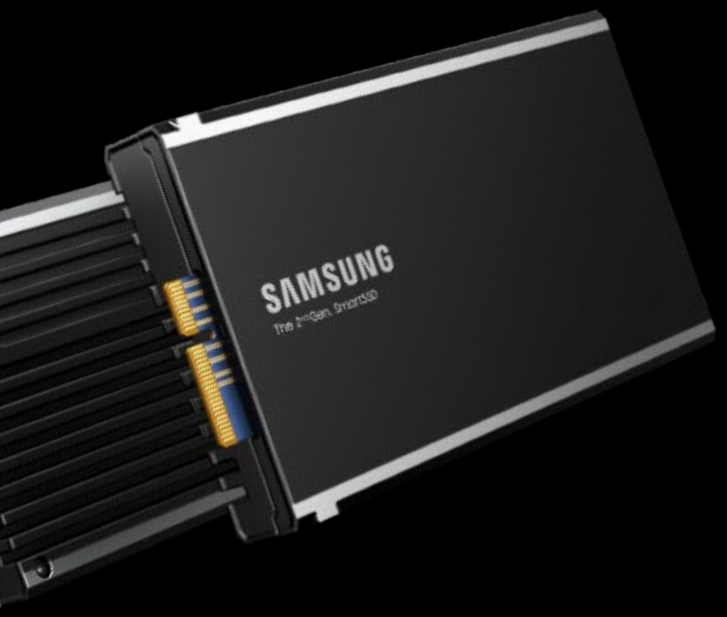
High Effective Bandwidth

Scalable Computing

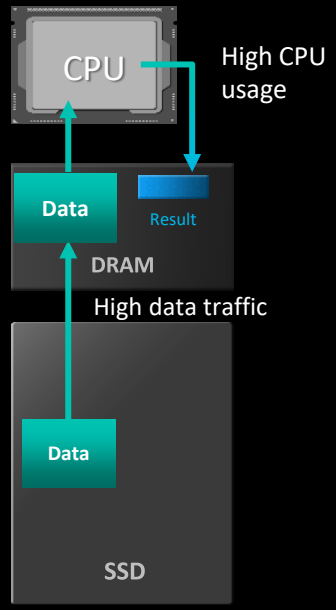
Accelerator Attached Tiered Memory Solution

CMM-HC* as a tiered memory solution with accelerator in package

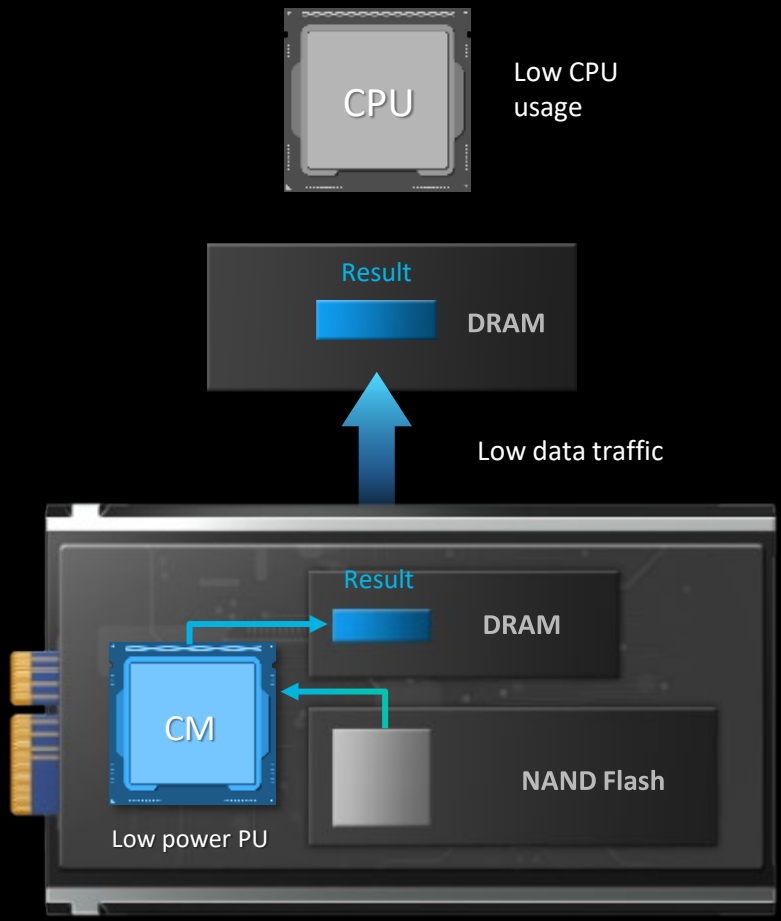
CMM-HC*: CXL Memory Module-Hybrid Computing



Server with Normal SSD

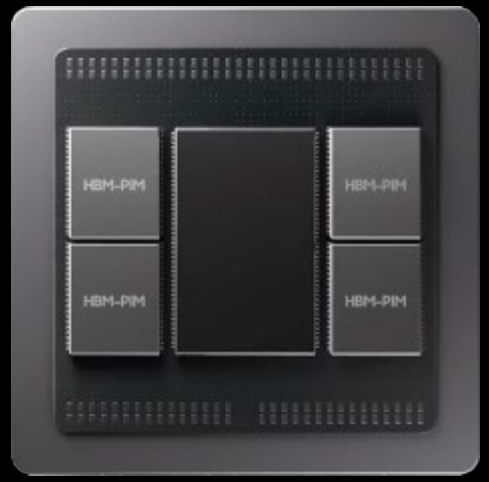


Server with CMM-HC



Samsung’s Solutions to Data-Centric Computing

From chip to device across DDR, CXL, and NVMe



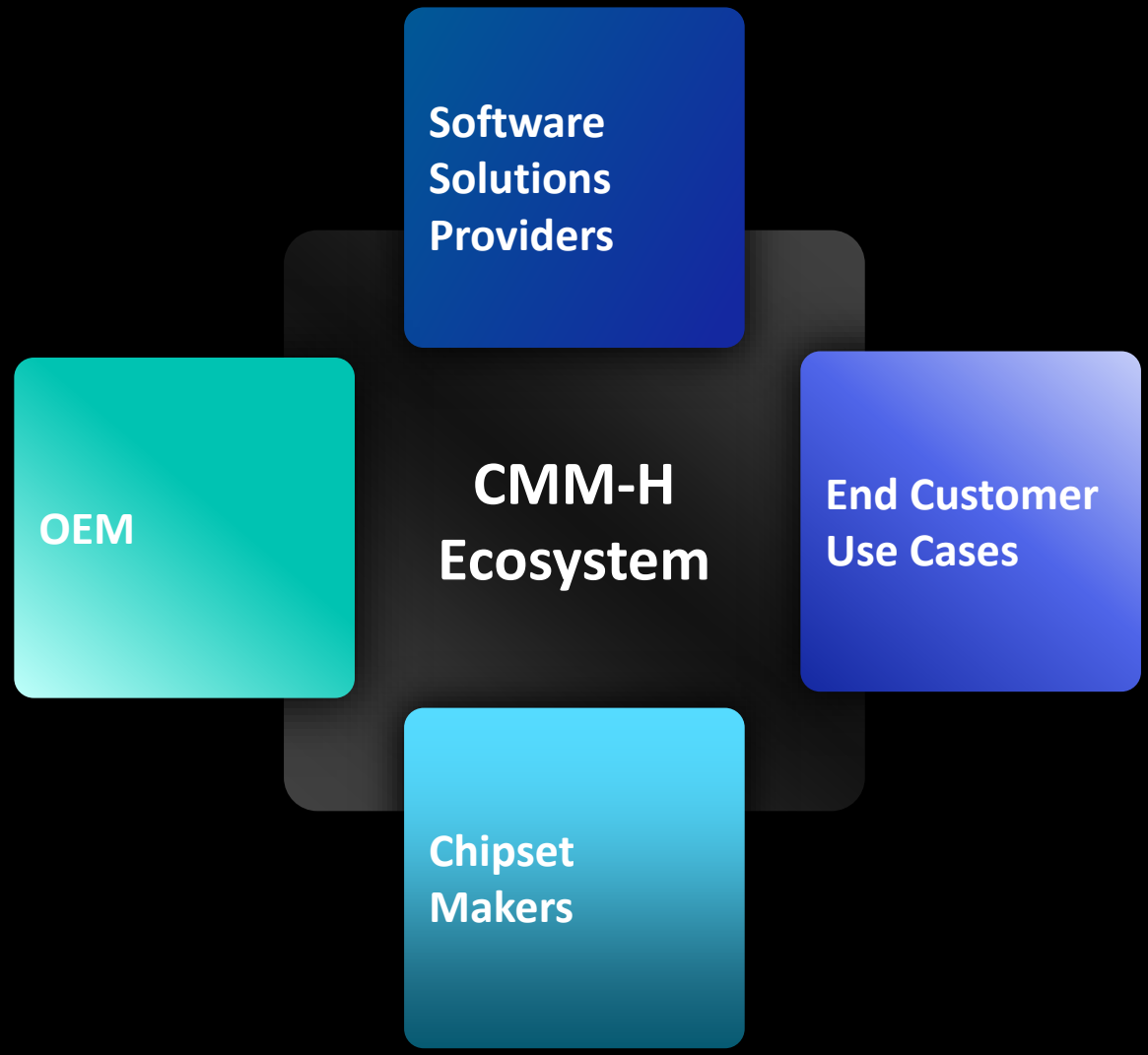
Processing In Memory (HBM-PIM)

Accelerator Attached Memory Expander (CXL-PNM)

Accelerator Attached Tiered Memory Solution (CMM-H)

CMM-H Ecosystem

- End Customers
  - Persistent memory
  - Tiered Memory TCO advantage
- Software Solutions Providers
  - IMDB companies
  - SW infrastructure providers
- OEMs
  - Go-to-market solutions
  - Sales channel enablement
- Chipset Makers
  - Roadmap alignment
  - CMM-H validation and certification



CXL Board of Directors

CXL Compute Express Link®

Industry Open Standard for High Speed Communications

255+ Member Companies

Your Invitation

Connect with Samsung regarding your memory-tiering and persistence applications

CXL Solutions Enablement is Here

Samsung CMM-H (CXL Memory Module-H)

CMM-H TCO Advantage

Balance of compute, memory and storage resources

Persistent Memory Support

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

Thank You