Flash Memory Summit

Controller Design Considerations for Samsung's Memory-Semantic SSD

Brian Myungjune JUNG

Samsung Electronics

35 YEARS OF NAND FLASH MEMORY INNOVATION

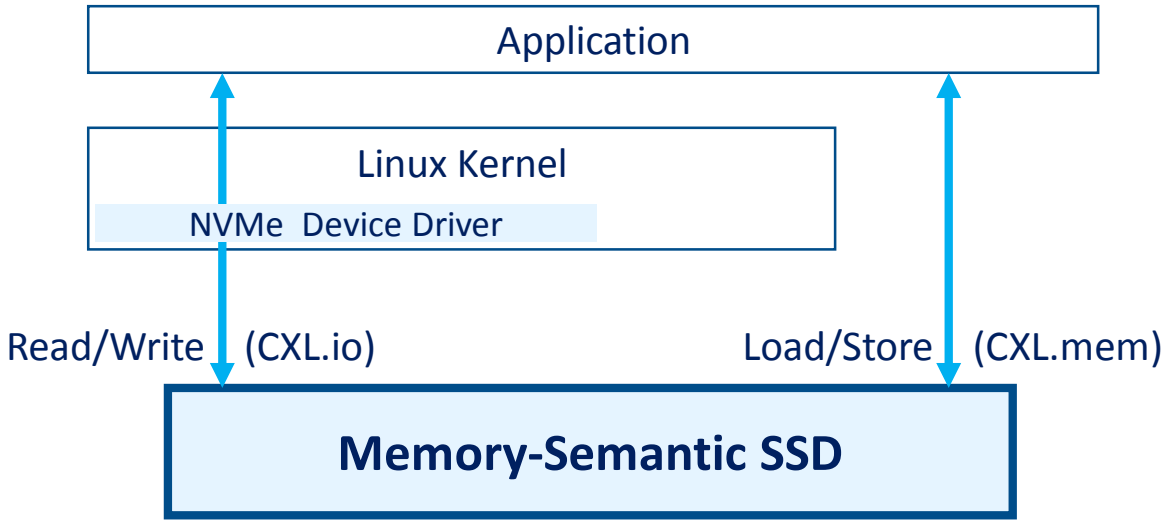
Table of Contents

Introduction to Memory-Semantic SSD

Memory-Semantic SSD

- CXL-based, dual-interface (.io & .mem)
- DRAM-like latency with NAND capacity

※ The entire NAND capacity is mapped to the CXL.mem memory address space

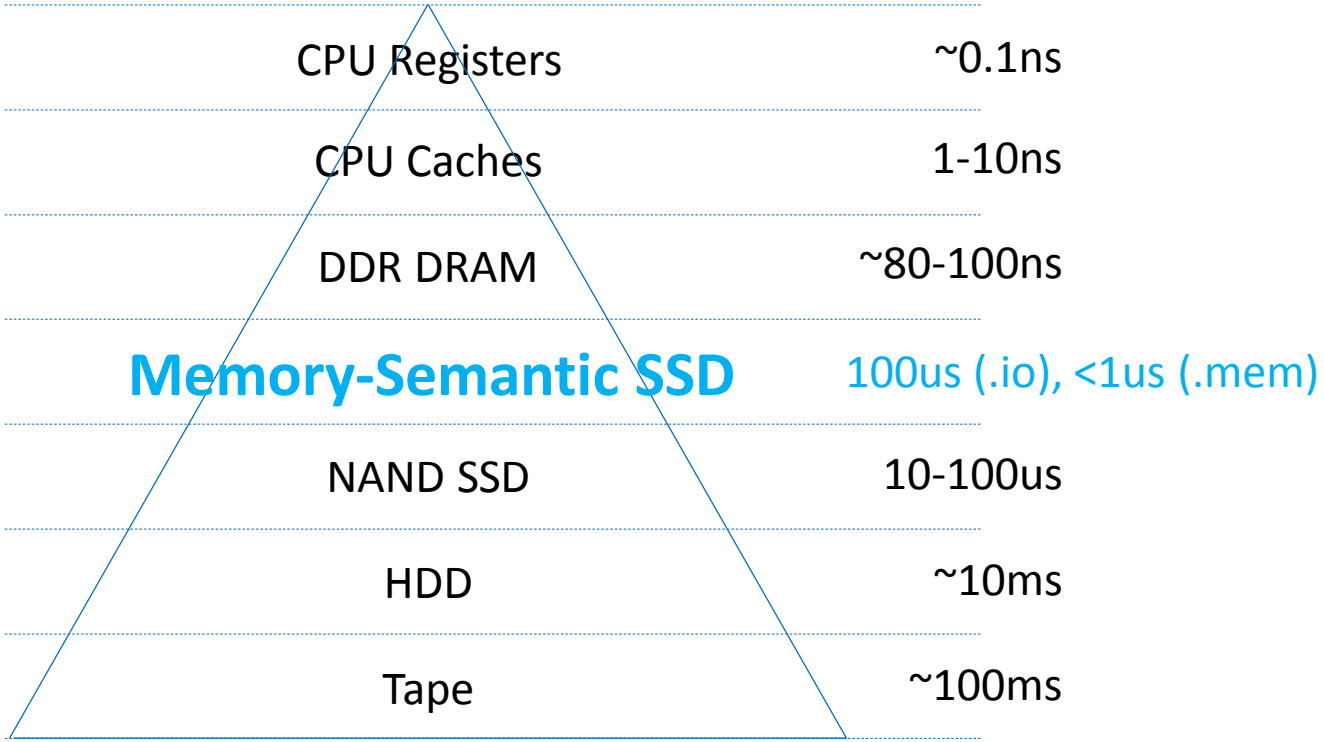


Position of Memory-Semantic SSD

- In-between DDR DRAM and NAND SSD

[Memory Characteristics]
Volatile Memory
Load/Store Instructions
Cache Line Granularity

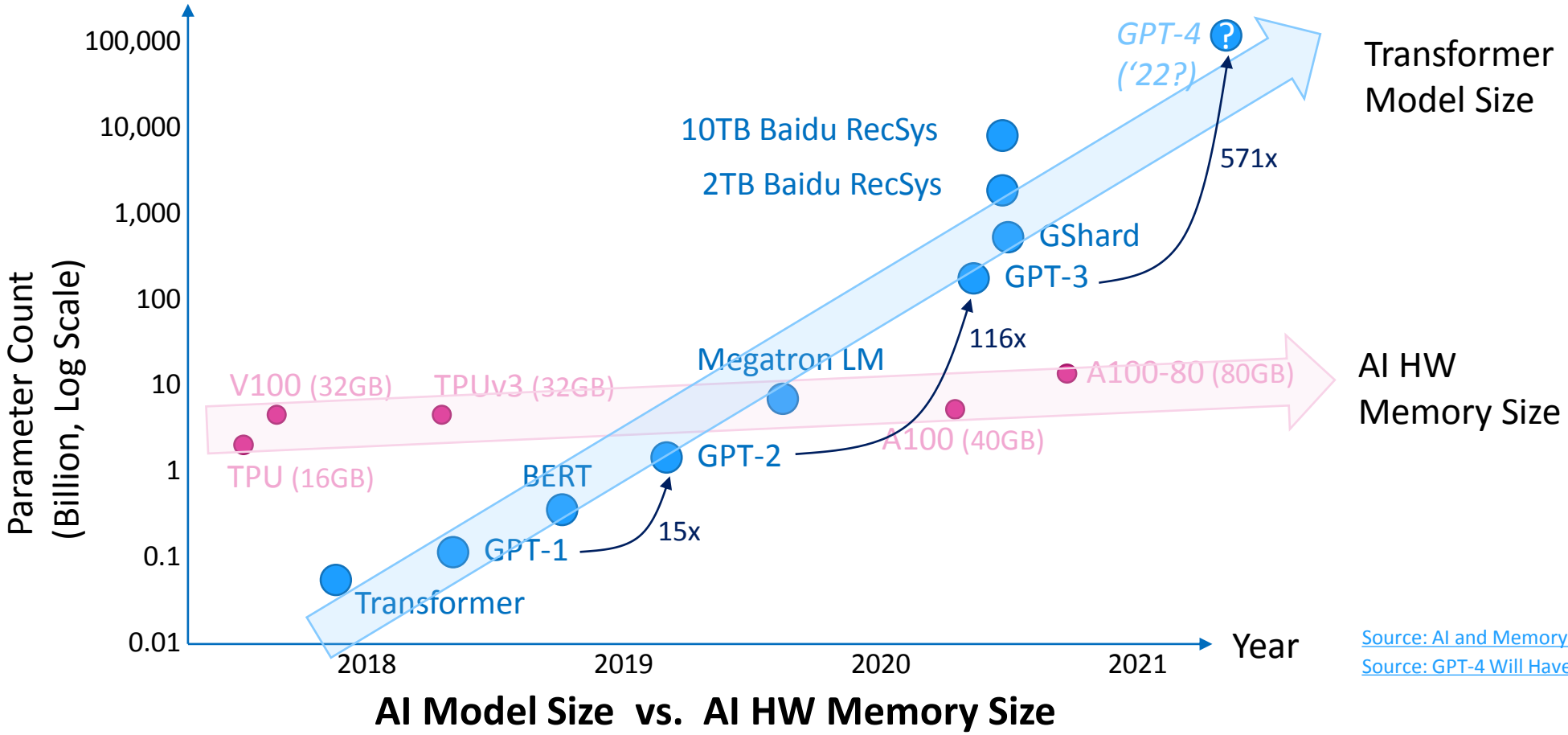
[Storage Characteristics]
Non-Volatile Storage
I/O Commands
Block Granularity



Memory Hierarchy

Target Use Case: Very Large AI Model

- Even larger-scale memory with lower cost



Source: AI and Memory Wall, by Amir Gholami

Source: GPT-4 Will Have 100 Trillion Parameters

Target Use Case: Non-Volatile Memory Systems

- Non-volatile memory with lower cost

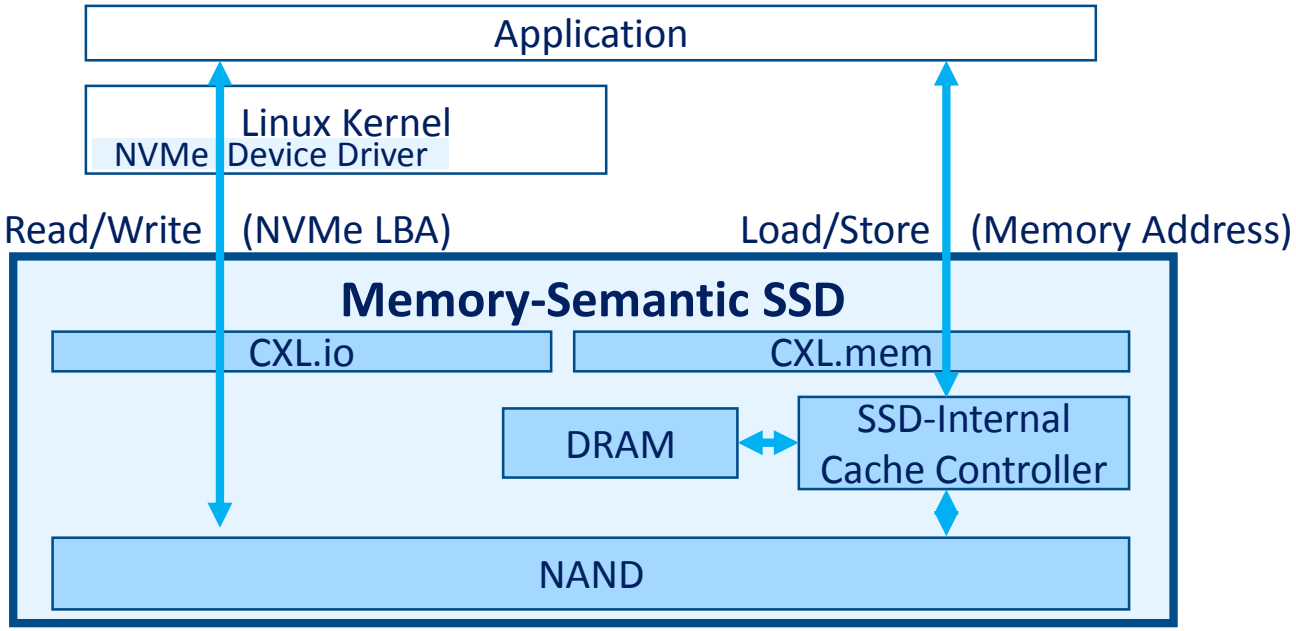
1. Reduces data copy (from storage to DRAM)
2. Data persists in memory after power interruption
3. Higher IOPS for smaller data (< 512-byte)
4. Easier to expand than DIMM based solution
5. Lower latency than flash storage (when internal cache hits)

Thanks to CXL

Memory-Semantic SSD Prototype

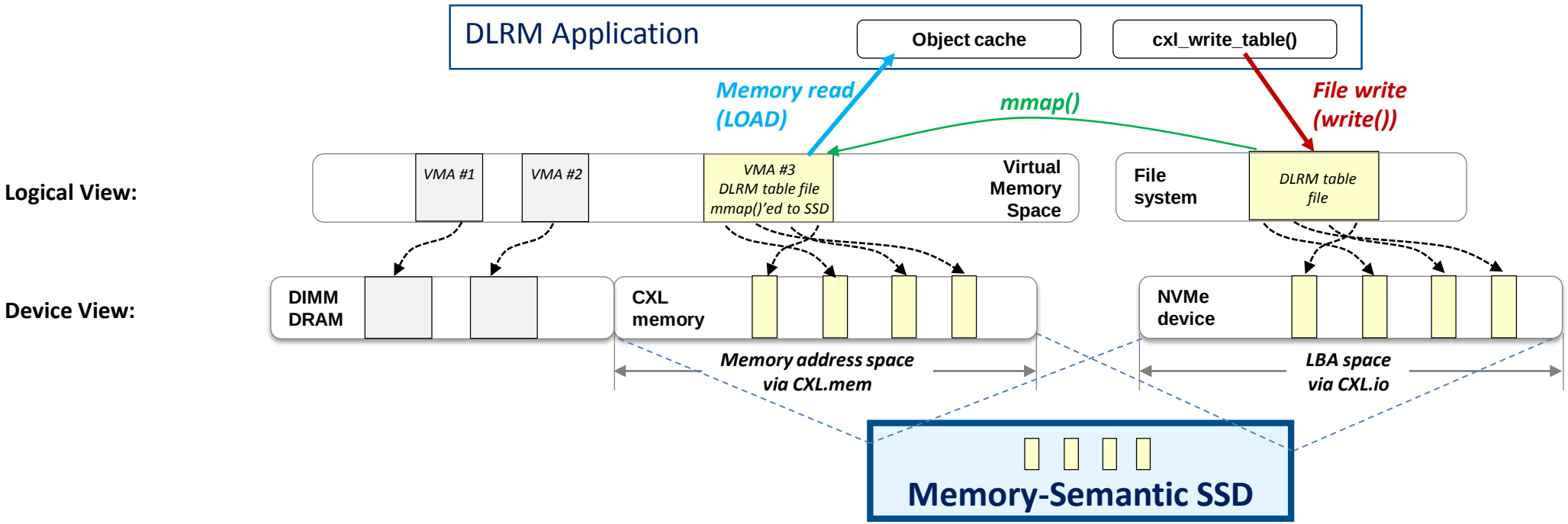
Memory-Semantic SSD, Based on CXL

- Memory-Semantic SSD provides dual-interface: CXL.io & CXL.mem
- Application can use any interface to access the same data
  - Application reads/writes data via NVMe interface (CXL.io)
  - Application uses memory interface to get high IOPS with low latency (CXL.mem)



Memory-Semantic SSD, How It Works

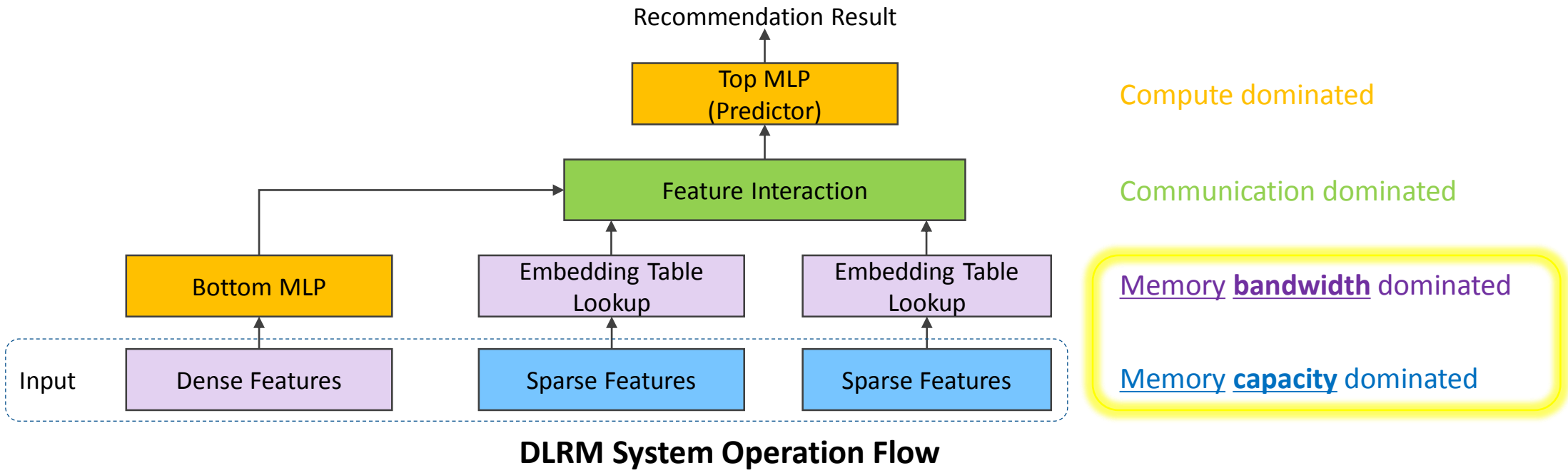
- An application (e.g., DLRM)
  - Writes a file via NVMe interface (CXL.io), and performs mmap()
  - Reads the data via memory interface (CXL.mem)



Motivation: Very Large AI Model

Memory Hungry

- An example: DLRM
  - DLRM system is memory capacity and bandwidth hungry
  - Memory-Semantic SSD is a great fit for DLRM

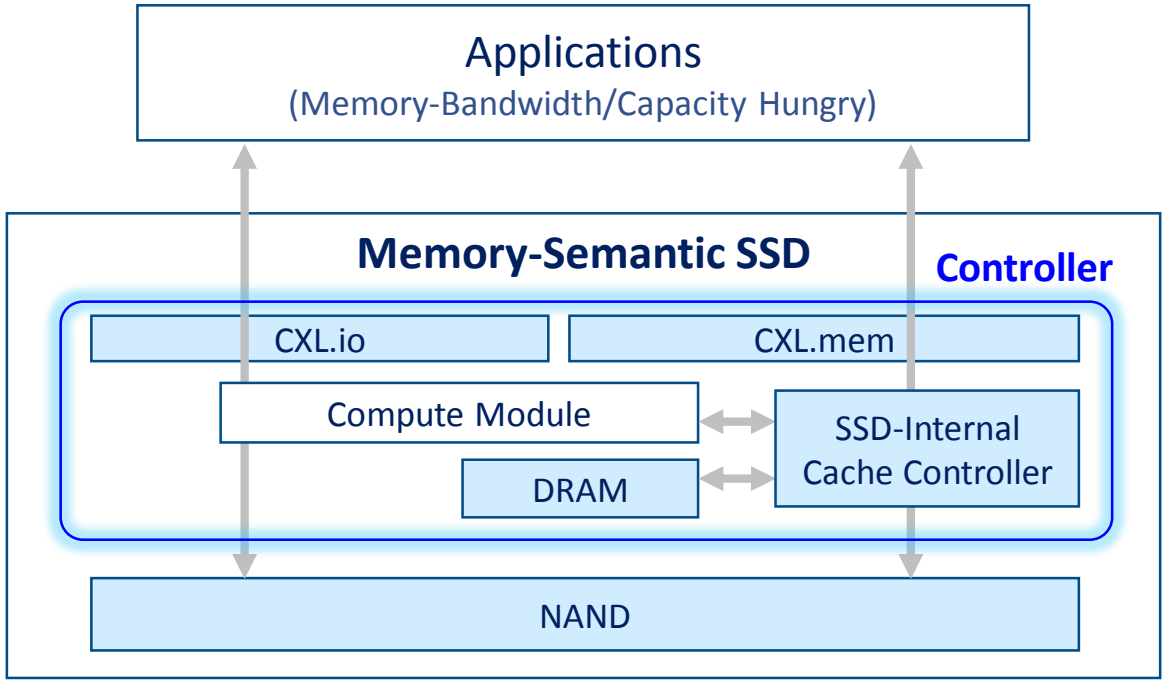


DLRM System Operation Flow

Source: DLRM: An advanced, open source deep learning recommendation model (facebook.com)

Controller Design Considerations

- Prototype v1
  - Dual-Interface Support
    - Address Translation (LBA-to-Mem)
  - High Performance & Deterministic Latency
    - Read Cache (LRU)
    - Prefetching
    - Request Coalescing
    - HW Logic-Based Data Path
- Prototype v2
  - Non-Volatile Memory Features
    - DRAM-Based Write Buffer w/ POR
- Prototype v3
  - Near Data Processing
    - Compute Module



Memory-Semantic SSD Prototype v1

- Features
  - E3 form-factor
  - FPGA-based prototype
  - Maximum 8 lanes, up-to Gen4
  - 16GB DRAM (for .mem read)
  - 2TB NAND
- See also Samsung booth for
  - E3 prototype sample
  - Memory-Semantic SSD + DLRM operation demo

Performance of E3 Prototype v1

- Random read over .mem (100% SSD-internal cache hit case)
  - Throughput (measured by sysbench)
    - ~ 20MIOPS (128-byte random read)
  - Latency (measured by Memory Latency Checker)
    - < 1us (64-byte, 128-byte)

What's the Next Step?

Roadmap of Memory-Semantic SSD

- FPGA-based prototype (E3.L 2T)
  - Prototype v1 further E3 samples will be available soon
  - Prototype v2 with ‘Write buffer with POR’ features (‘22.Q4)
  - Prototype v3 with ‘Near Data Processing’ features (’23.H2)
- ASIC-based E3 sample (E3.S 1T/2T)
  - Based on prototype v2 (’24.H1)
  - Based on prototype v3 (’25.H2)

Call for Collaboration (Memory-Semantic SSD)

Flash Memory Summit

- Hardware development
  - Server & switch which supports E3 form-factor Memory-Semantic SSD
  - Rack-scale disaggregated system with Memory-Semantic SSDs
- Standardization
  - Management scheme to monitor/control Memory-Semantic SSD
- Reference systems & ecosystems
  - Large-scale AI applications
  - Non-volatile memory applications and system software
  - HPC

35 YEARS OF NAND FLASH MEMORY INNOVATION

Thank You

