Flash Memory Summit

SAMSUNG

Solving AI Workload's Storage Problem with Memory-Semantic SSDs and Tiered Memory

Rekha Pitchumani

PhD, Senior Manager,
Memory Solutions Lab,
Samsung Semiconductor Inc.

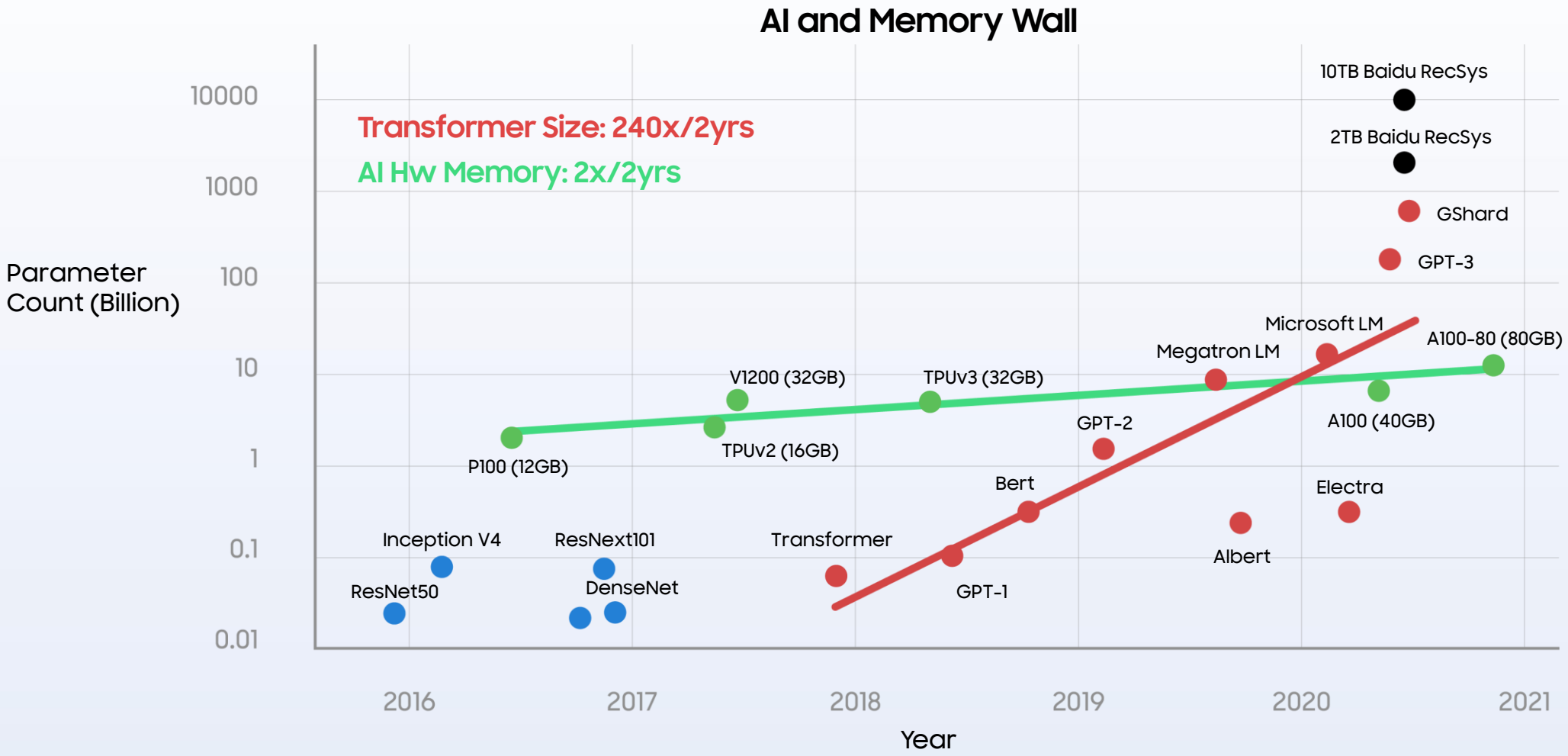
August 4th 2022

Agenda

Flash Memory Summit

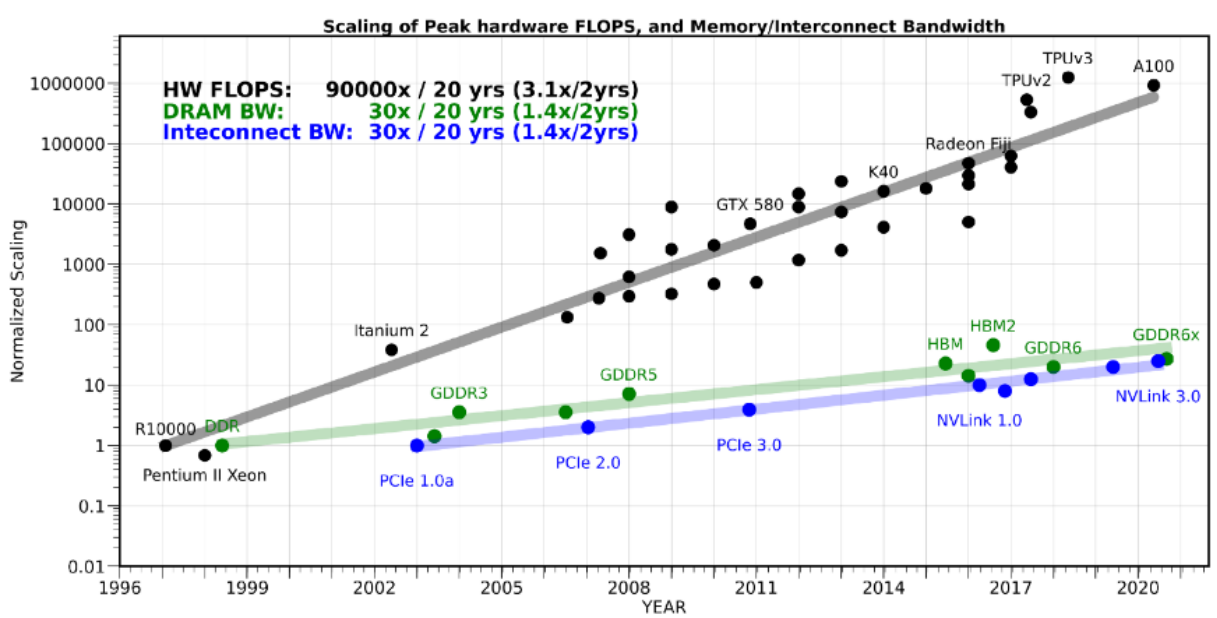
1. Growing AI Memory Needs
2. Samsung Memory Semantic SSD (MS SSD)
3. MS SSD for AI Memory/Storage
4. Memory Tiering for AI
5. Summary

Growing AI Models & Memory Needs



* Source: https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8

Accelerator Growth & AI Memory Wall



* Source: https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8

Memory capacity and bandwidth lagging behind compute growth. To keep up with workload demands,

- Rethink cache hierarchy
- Scale memory separately with caching, tiering and new interconnect technologies
- Compute near memory where applicable

Samsung's Memory-Semantic SSDs

Flash Memory Summit

With CXL, Memory, and Storage occupy the same physical slot

- Interchangeability means room for Memory-Storage convergence

Memory-Semantic SSD (MS SSD) supports dual (Memory/Storage) mode via the CXL.mem/CXL.io protocols

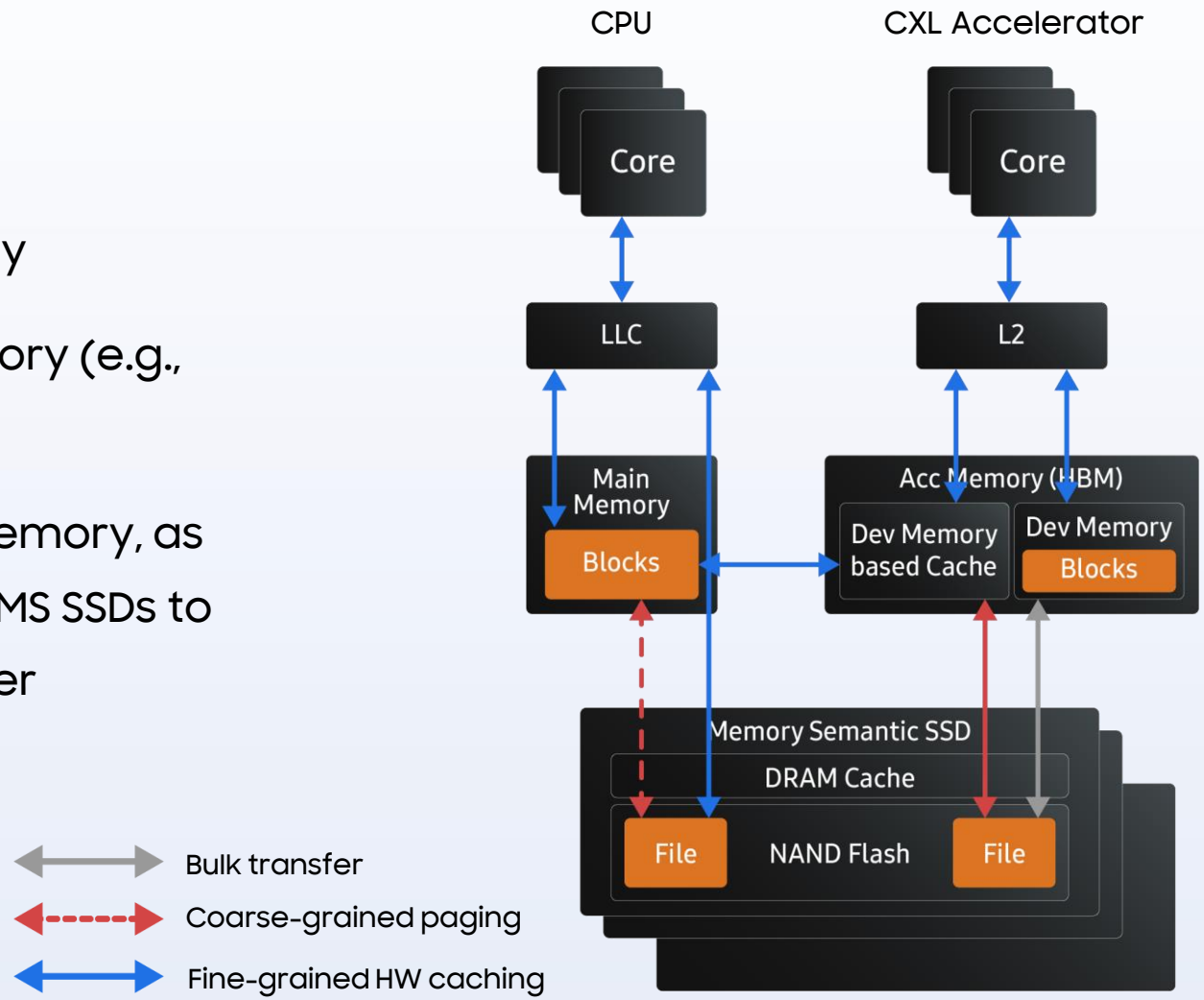
- Access the same data at a smaller granularity (64B) in memory mode than in IO mode (4KB)

To learn more, check out 'Controller Design Considerations for Memory-Semantic SSD' talk at FMS '22.

CXL Based Solution with MS SSD

CXL Based Accelerator with

- Redesigned on-chip cache hierarchy
- Ability to partition accelerator memory (e.g., HBM) into LLC or device memory
- CXL based MS SSDs, accessible as memory, as the slow memory tier, with multiple MS SSDs to meet the bandwidth needs of the tier

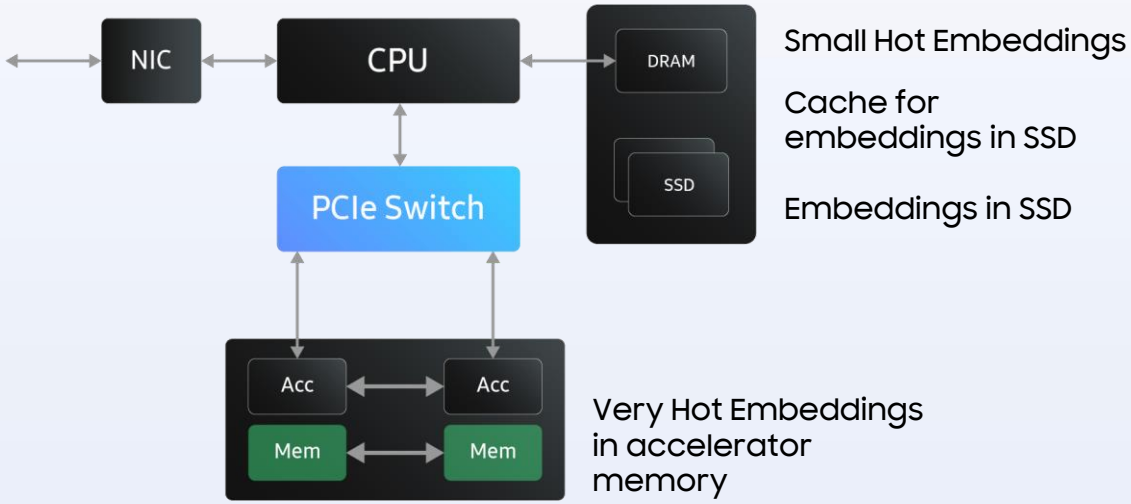


Recommendation Systems (RecSys)

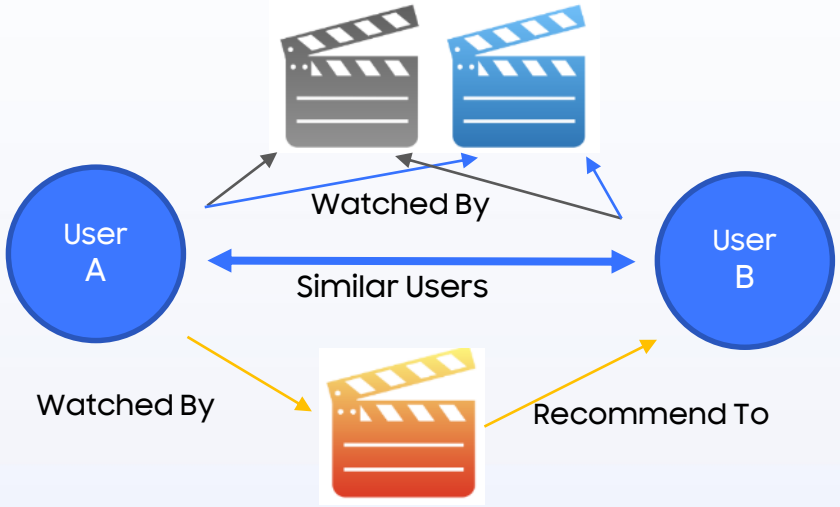
Flash Memory Summit

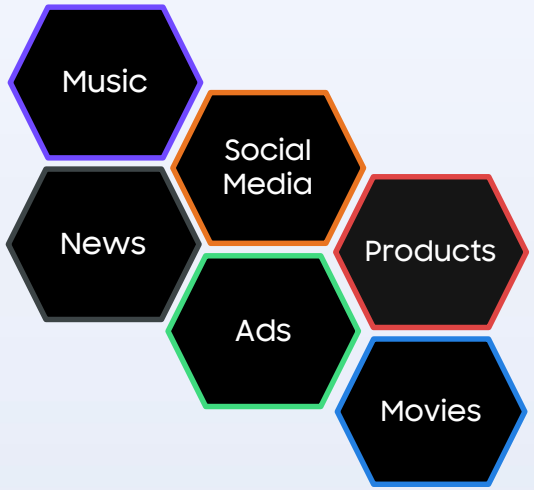
Meta's Deep Learning Recommendation Model (DLRM) to represent huge RecSys models

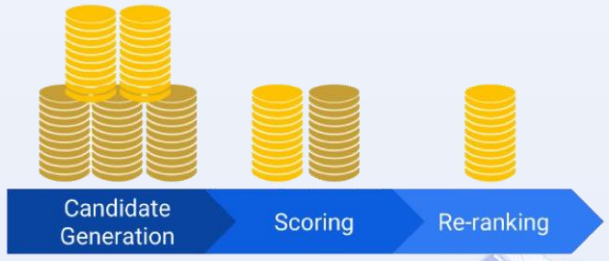
NVMe SSD based Software Defined Memory exploration for DLRM models at Meta



Source: "Memory Requirements of Meta AI Workloads"





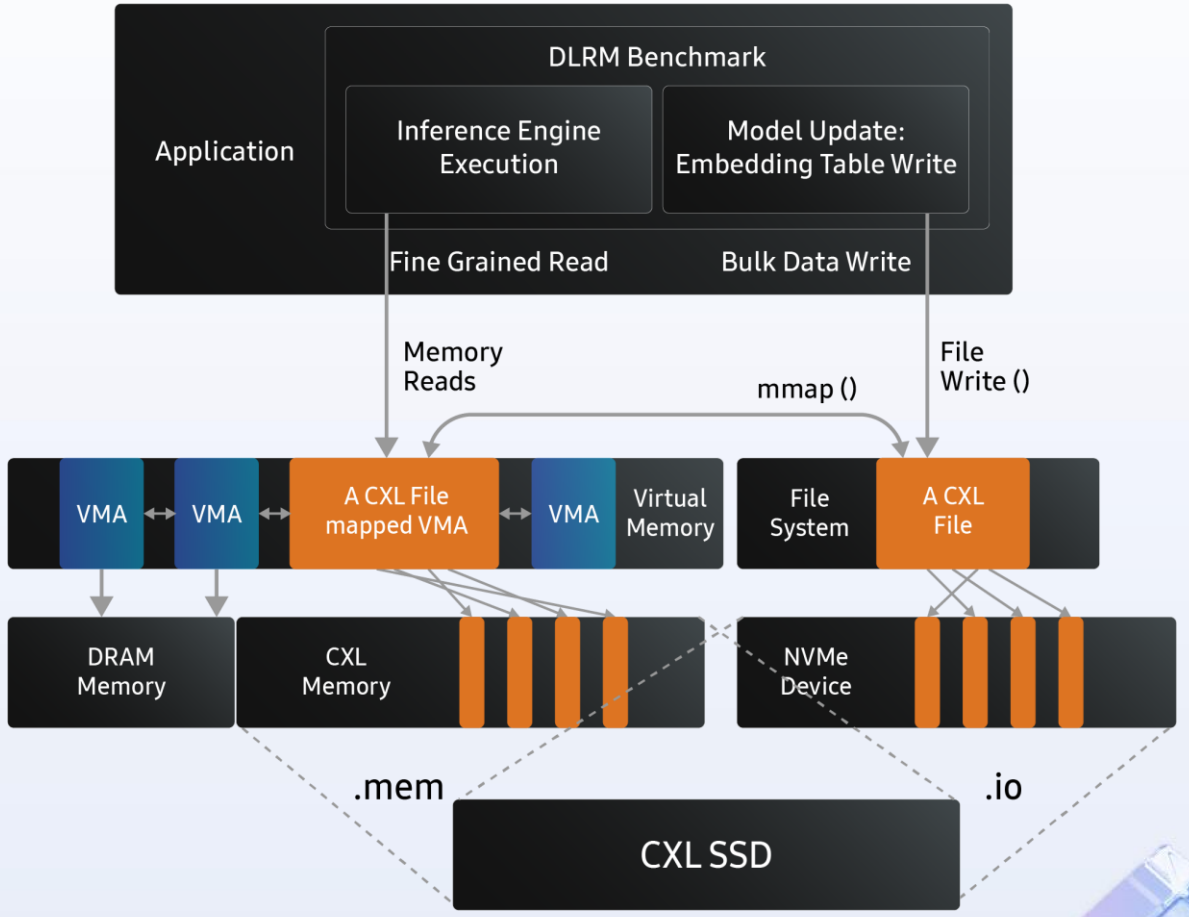


MS SSD for DLRM Workload

Flash Memory Summit

Dual-Mode Access

- Use as block device via CXL.io (file-system based access)
- Use as byte addressable memory via CXL.mem with load/store for memory-mapped files



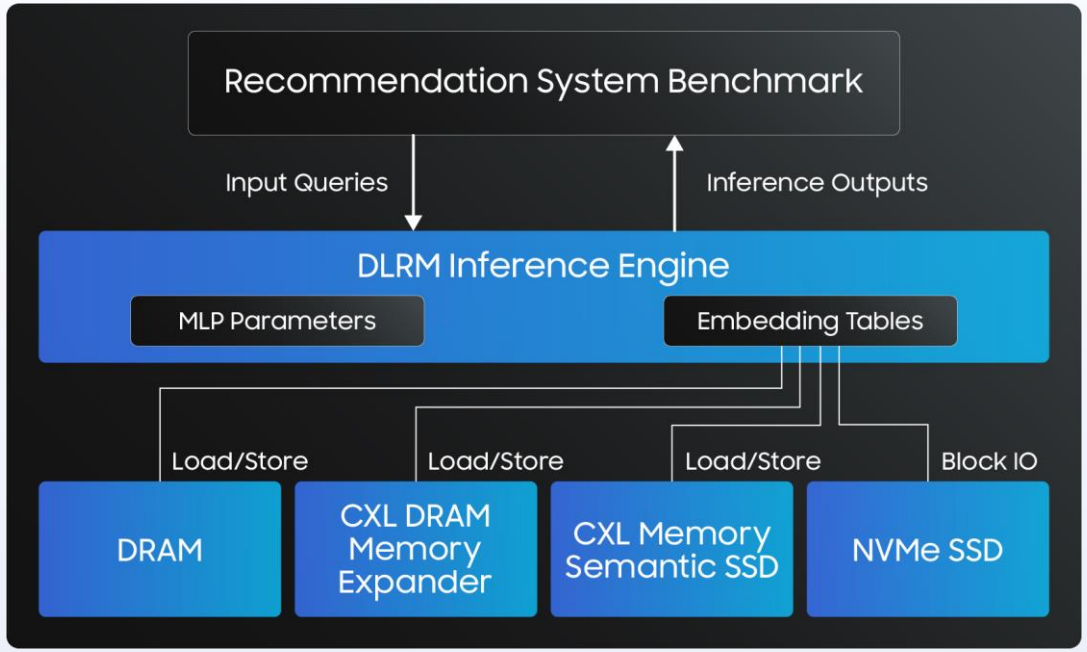
Visit Samsung demo booth (#407) for functional demo of Movie RecSys using MS SSD

RecSys Benchmark

Flash Memory Summit

Started with open recommendation system benchmark for a variety of RecSys including DLRM

- Added options to store and serve embedding tables from CXL Memory Expander, CXL based MS SSD, and NVMe SSD
- Added DRAM cache option and static partitioning of tables
- Adding memory tiering options



https://github.com/harvard-acc/DeepRecSys
https://github.com/facebookresearch/dlrm

Memory Tiering for AI

Flash Memory Summit

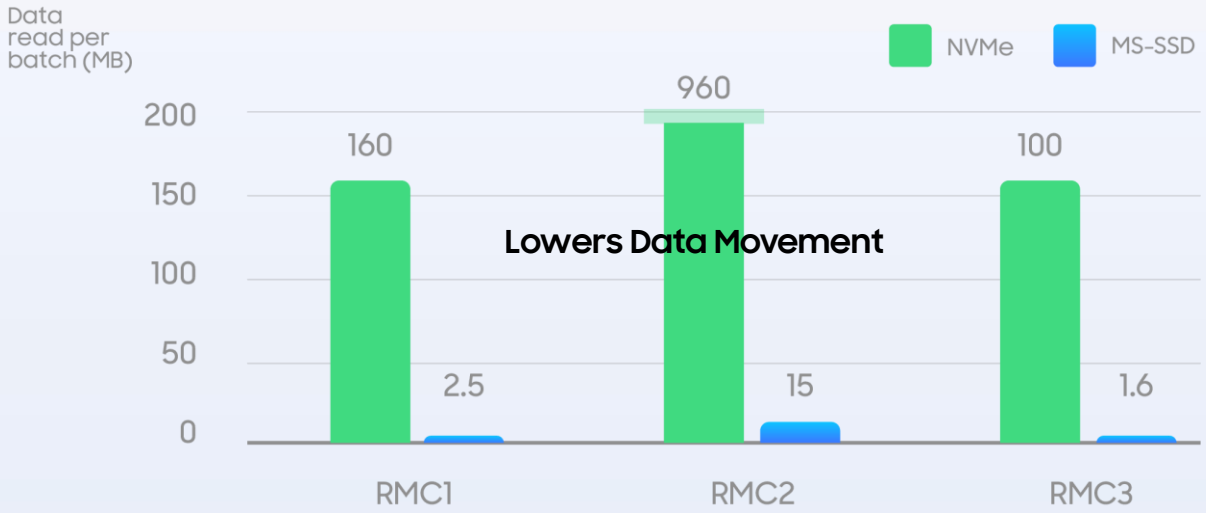
Has to operate at tensor (or embedding vector in case of DLRM) level

- Both static and dynamic options
- Needs and what works change for different AI domain such as Computer Vision, NLP, RecSys, etc.,
  - E.g., Hot/Warm/Cold embedding vector in different tiers in case of RecSys
- Call for more research and more AI framework level tiering support

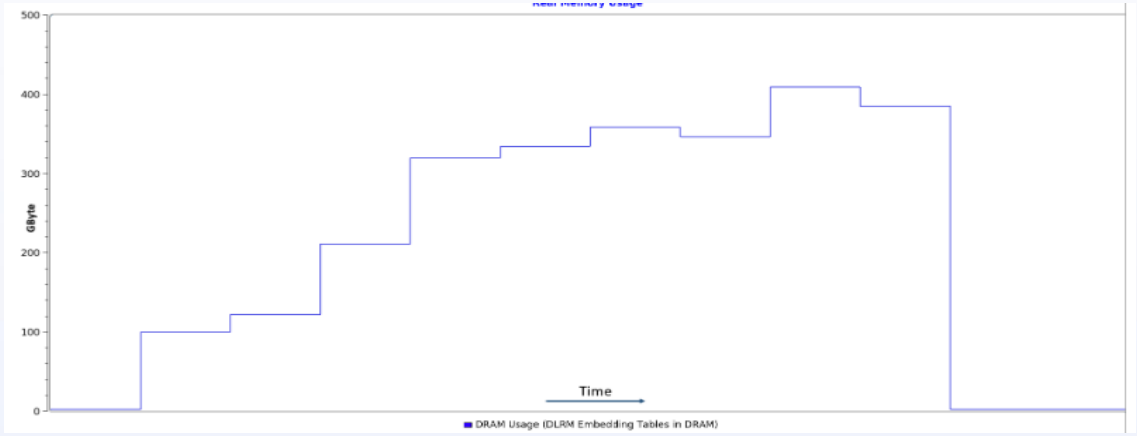
MS SSD Benefit

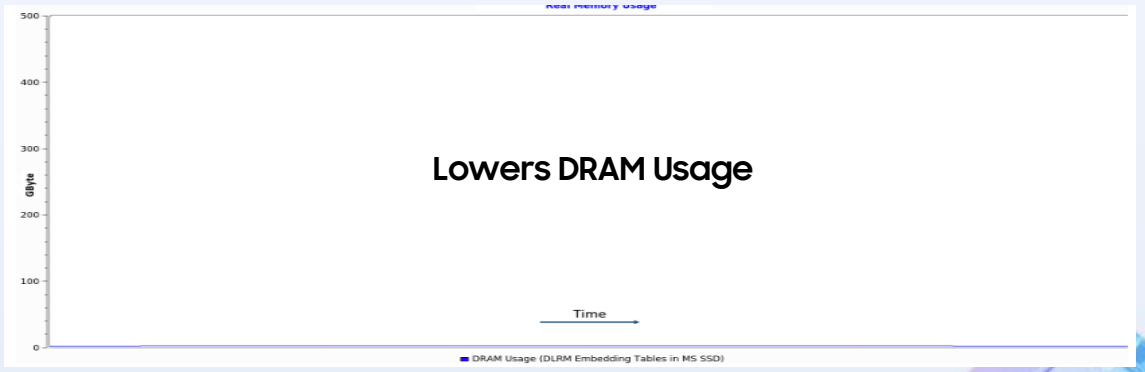
Finer-Granular access leads to less data movement from the SSD

Amount of data read for 3 DLRM models if 4K block is read for every embedding vector read



Amount of DRAM used to run same model with embedding tables on DRAM vs MS SSD





Summary

AI model sizes are fast growing

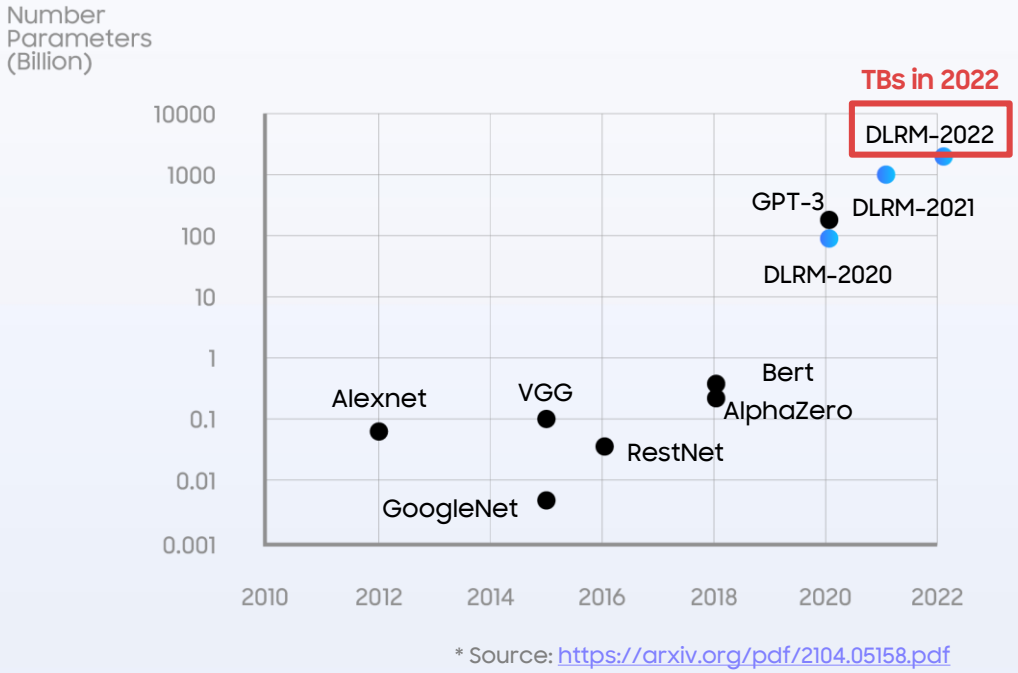
- RecSys, NLP, Video, Medical Imaging, etc.,

NAND flash SSD have better TCO compared to DRAM to store large models

- But NVMe SSDs have higher I/O stack overhead and unnecessary block data movement and copies to DRAM

Tiered Memory in AI framework with MS SSD can help solve AI's Storage/Memory Problem

- Enables fine-grained access and hardware caching, reduces data movement costs, simpler software stack, no copies to DRAM necessary to access non-temporal data



* Source: https://arxiv.org/pdf/2104.05158.pdf

Please visit Samsung demo booth (#407) to learn more!

SAMSUNG