Flash Memory Summit

SAMSUNG

Next-Gen System Architectures with Memory-Semantic SSDs

Rekha Pitchumani

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

August 3rd 2022

Agenda

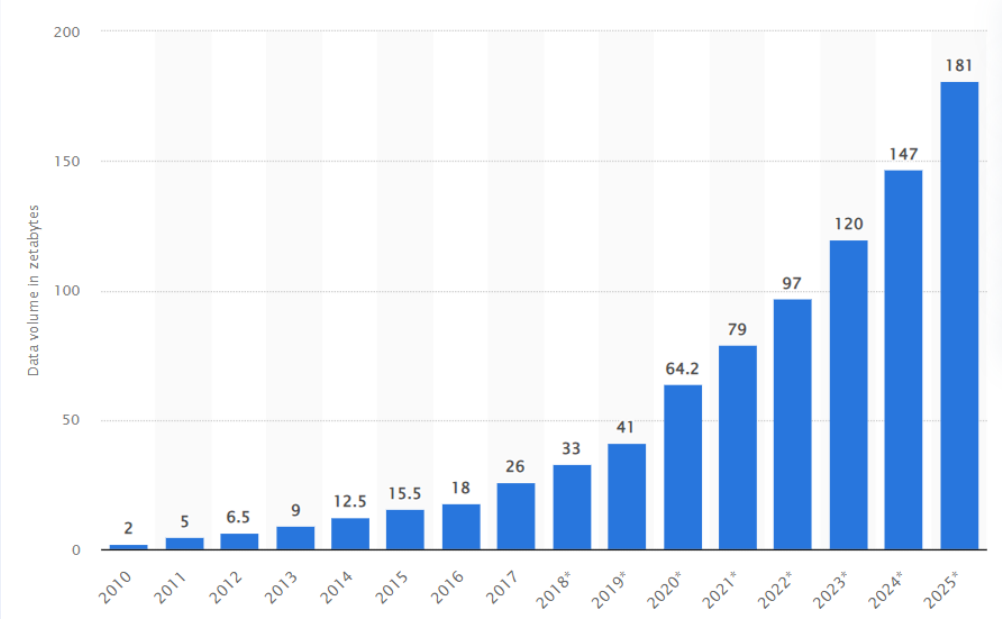
Flash Memory Summit

1. Industry Trend – Caches, Memory and Storage
2. Samsung Memory Semantic SSD (MS SSD)
3. Data Movement in Traditional Architectures vs Next-Gen Architectures with MS SSD
4. Benefits of MS SSD
5. Summary

Data Explosion & Technology Growth

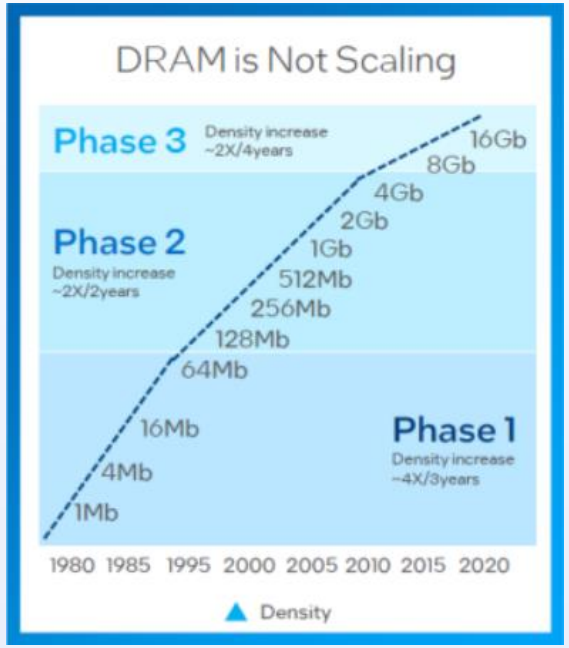
Flash Memory Summit

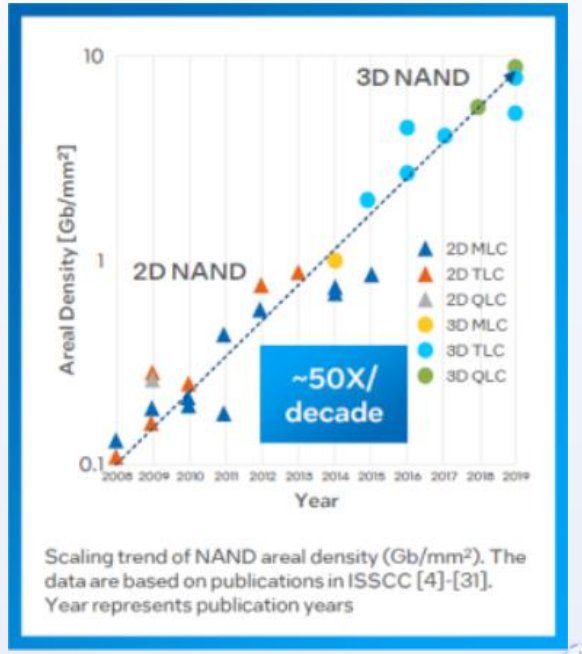
Amount of data created, consumed and copied worldwide



* Source: https://www.statista.com/statistics/871513/worldwide-data-created/

DRAM is not scaling at same pace, and is typically supplemented by NAND to keep up





* Source: https://infohub.delltechnologies.com/l/new-vmware-vsphere-memory-techniques-in-dell-emc-hyperconverged-infrastructure-solutions/technology-background

Accelerator Growth & AI Memory Wall

Flash Memory Summit

Prior work from Microsoft shows modern workloads need fat last level caches. To keep up with workload demands,

- We need larger L4 caches and memory side caches
- Large capacity memory tier
- Compute near memory

On-chip caches are getting bigger, but per core sizes not so much

	Intel Pentium Pro (1996)	Third Gen Intel Xeon Scalable (2021)
Cores (C) and Threads (T)	1C, 1T on 0.35 μm	Up to 40C, 2T per core on 10 nm
Core Frequency	150–200 MHz	Up to 3.9 GHz base frequency
L1 Cache (code, data)	8 KB Code 8 KB Data	32 KB Code 48 KB Data
L2 Cache (unified)	256 KB to 1 MB	1.25 MB per core on-die
L3 Cache	None	Up to 60 MB shared by cores; on-die distributed layout w/1.5 MB per core

Intel, the Intel logo and Xeon are trademarks of Intel Corporation or its subsidiaries.

* Source: "Advances in Microprocessor Cache Architectures Over the Last 25 Years"

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)

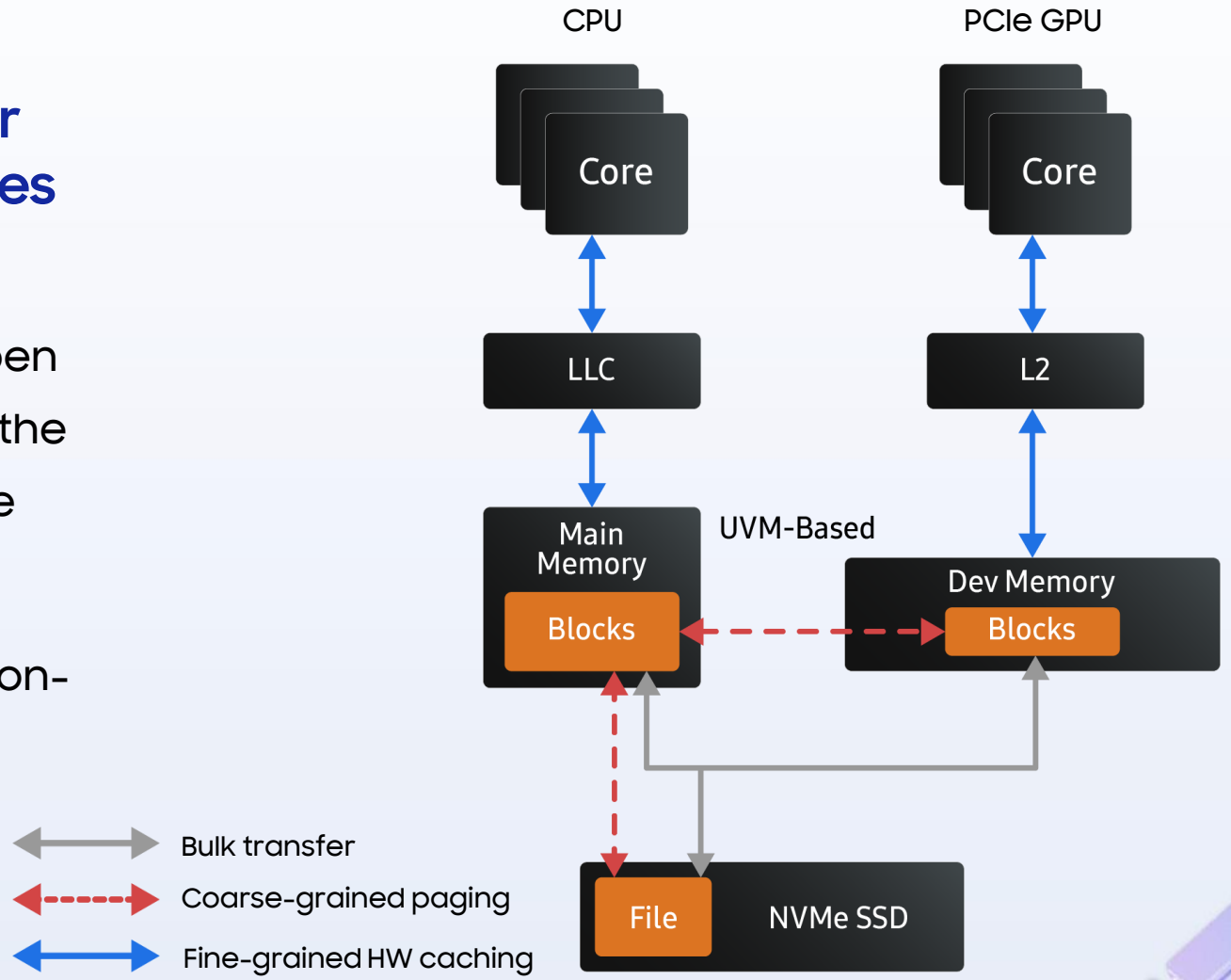
To learn more, check out 'Controller Design Considerations for Memory-Semantic SSD' talk at FMS '22.

Traditional Architecture

Flash Memory Summit

Both main memory and accelerator device memory are memory devices and the NVMe SSD is an IO device

- Data transfer between these can happen only via explicit bulk copies initiated by the application or coarse-grained software based paging and prefetching
- Fine-grained hardware caching to the on-chip caches is available only for the memory devices

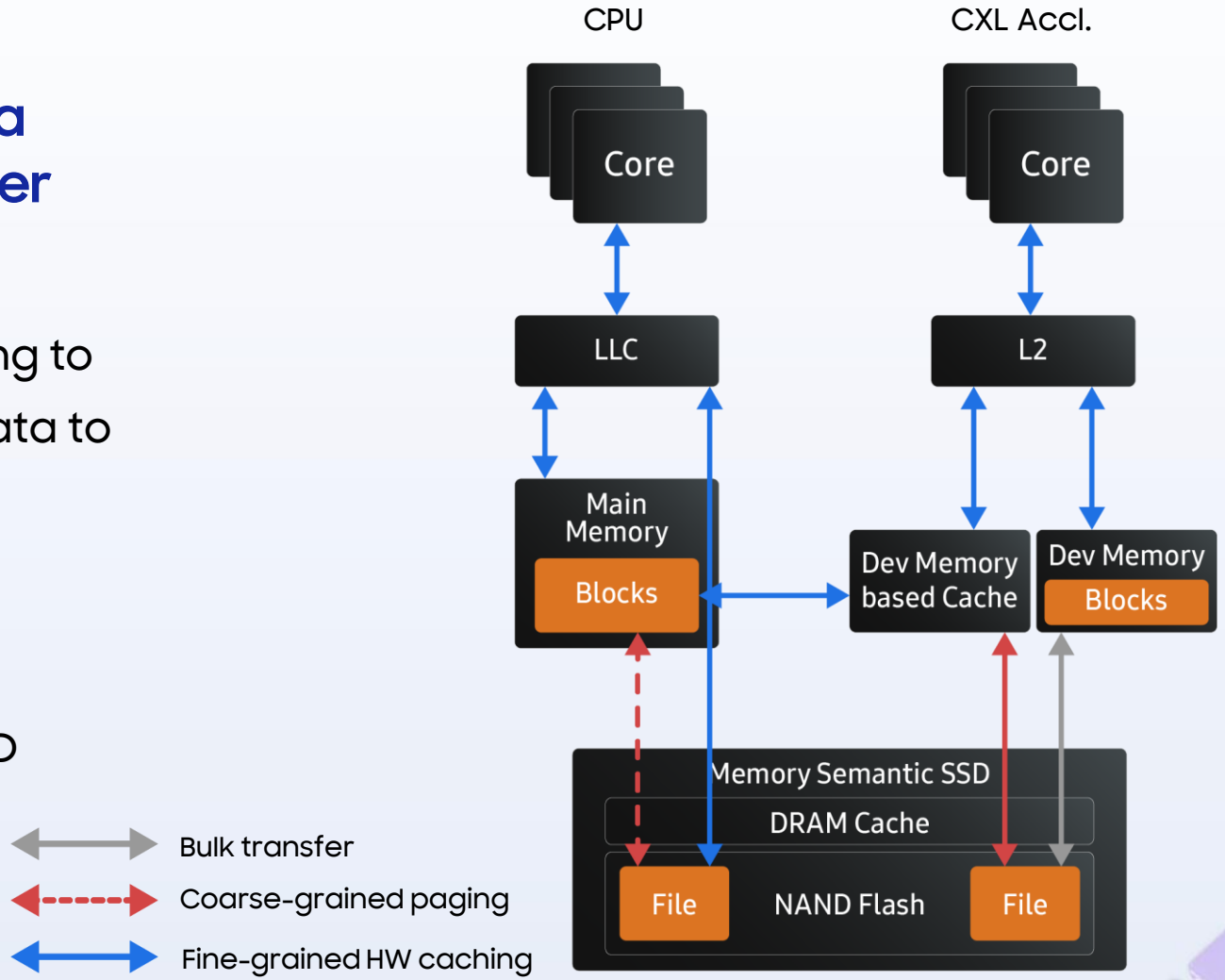


Next-Gen Architecture with MS SSD

Flash Memory Summit

Dual mode MS SSD can be used as a memory devices or an IO device per application need

- Enables fine-grained hardware caching to the on-chip caches without moving data to other memory devices
- If server or accelerator supports, can configure part of main or accelerator memory as a last level cache to MS SSD

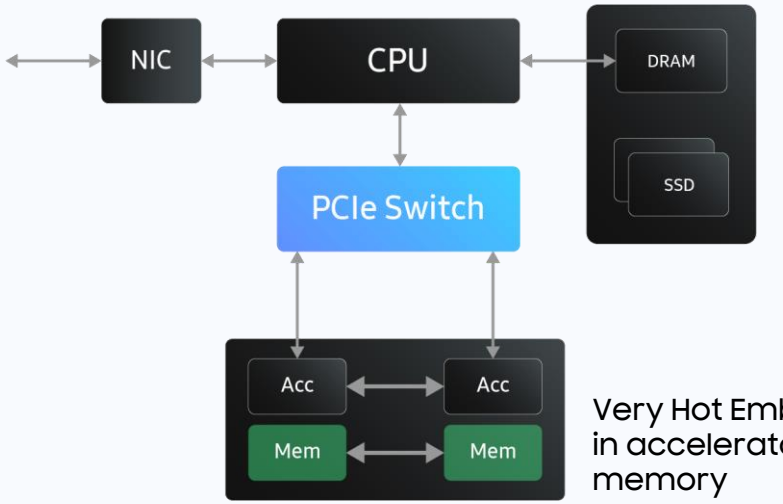


MS SSD-based Memory Expansion

Flash Memory Summit

Not a general-purpose memory expander, but for the benefit of applications that are looking for SSD-based memory expansion and can tolerate the higher latency

- E.g., Huge AI models such Meta's DLRM and Microsoft's DeepSpeed, that use NVMe SSDs for memory expansion and paying the IO tax
- Datacenter caching, also paying NVMe IO tax, such as Cachelib



Source: "Memory Requirements of Meta AI Workloads"

Source: "ZeRO-Infinity and DeepSpeed: Unlocking unprecedented model scale for deep learning training"

Source: "From DRAM to SSDs, Challenges with Caching at FB Scale"

MS SSD Benefit

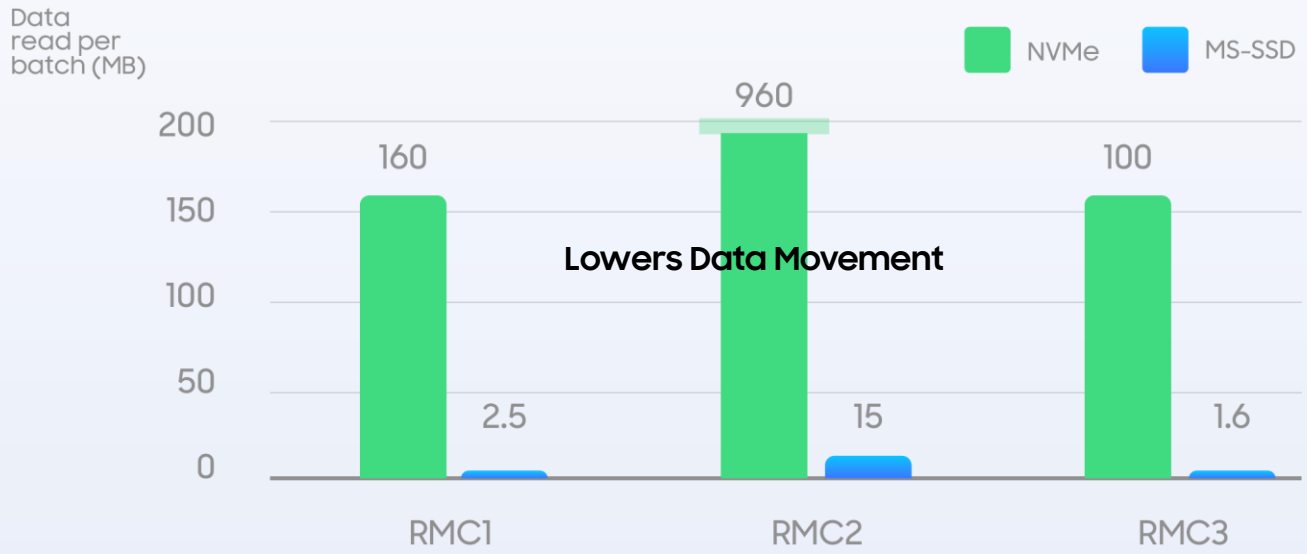
Flash Memory Summit

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

Reduced software stack complexity due to Memory Semantics

- Load/store access to SSD instead of block IO stack

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



MS SSD Challenges & Work-In-Progress

Flash Memory Summit

AI model sizes are fast growing

- Core latency tolerance, especially tail latency tolerance
- Need features in-place to tolerate higher latency

Work-In-Progress

- Quantify latency impact on E2E performance and workloads running on other cores with PoC device and reference servers
- Work with server vendors and in the future accelerator vendors to put necessary features in-place

Summary

Flash Memory Summit

Memory Semantic SSD with CXL interface provides load/store access to NAND flash media

- Enable finer-grained access required by modern applications

Lower overall system TCO by systematically adding appropriate storage at the bottom of the memory hierarchy

- Enable finer-grained hardware caching without software involvement and reduced data movement increases efficiency

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

SAMSUNG