FUTUREWEI Technologies

Whitepaper

Emerging Storage Class Memory

version	date
1.0	10/27/2023

Trademarks and Permissions

FUTUREWEI Technologies and other Futurewei® trademarks are trademarks of Futurewei® Technologies. Huawei trademarks are trademarks of Huawei Technologies Co., Ltd. All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services, and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services, and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees, or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

FUTUREWEI® TECHNOLOGIES, INC.

Boston Research Center

Address: 111 Speen Street, Suite 114
Framingham, MA 01701
United States of America

Website: http://www.futurewei.com/

Contents

1 EXECUTIVE SUMMARY

In this whitepaper, we will examine emerging storage class memory options after the demise of Intel® Optane. Our key takeaways:

- ***Intel Optane's downfall underscores the significance of economies of scale in the memory business.***

- ***None of the emerging memory technologies, MRAM, FERAM, and ReRAM, are particularly new, yet none have crossed the chasm into mass production.***

- ***CXL (Compute Express Link) has received widespread industry endorsement as the most promising memory interconnect. Vendors have developed CXL use cases in memory expansion, memory tiering, near-memory compute, and memory-semantic SSD.***

- ***Memory-semantic SSD has the potential to emerge as the new storage class memory and improve an enterprise storage system's performance by providing low-latency, shared access to its metadata.***

- ***Two technologies have the potential to emerge as the storage class memory options for enterprise storage. SLC SSD makes sense as a write cache in front of QLC SSDs, but its relatively higher latency makes it less ideal as a metadata store. For metadata store, CXL-based memory-semantic SSDs might be a better choice.***

- ***With CXL's support for persistent memory and cache-coherent memory sharing, memory-semantic SSD can improve an enterprise storage system's performance by providing low-latency, shared access to its metadata store.***

2 The Demise of Intel Optane

Optane was a memory technology that Intel introduced in 2015 to bridge the cost/performance gap between DRAM and SSD. Aimed to cost half as much as DRAM per GB while performing at near-DRAM speeds, Optane's adoption was expected to reduce the overall server and storage system costs and improve their performance. Intel invested approximately $10B trying to make Optane a success before deciding to discontinue it in 2022 [1].

The underlying technology for Optane was *3D XPoint memory,* a type of phase-change memory (PCM)[1]. Crosspoint arrays are the densest-possible layout of bits, and the more bits you get onto a chip the cheaper the cost per bit. At the 2015 introduction, Intel and its then-partner Micron® assumed that since a 3D XPoint memory die was half the size of a DRAM die, it could be produced at half the cost. However, this potential cost advantage can only be realized if Intel sold enough Optane products to drive large volumes, large enough to drive down the costs.

Initially, Intel launched Optane SSDs to drive volumes before introducing Optane DIMMs, which required modifications to processor architecture and operating systems. Unfortunately, Optane SSDs never became very popular as their performance was slowed down by their NVMe interface to the point that they were not all that much faster than NAND SSDs, while they were sold at a hefty price premium over NAND SSDs.

By the time Optane DIMMs were released, 3D XPoint memory's production volume was still low, keeping costs high. Unable to price Optane DIMMs significantly below DRAM level, crucial for driving Optane DIMM sales volumes, Intel struggled for another four years before discontinuing Optane in 2022.

Intel Optane's downfall underscores the significance of economies of scale in the memory business. 3D XPoint memory's volume never reached even 1/10th of DRAM volume, preventing it from achieving the anticipated cost benefits. Intel could have potentially matched Optane SSD prices with those of NAND SSDs to boost volume and reduce costs, but such a move would have exacerbated Intel's financial losses far beyond the actual $10B. As we evaluate emerging memory technologies in the next sections, the emphasis on economies of scale becomes paramount.

[1] PCM is a non-volatile memory technology that uses phase change materials to stores data.

3 Emerging Memory Technologies

3.1 MRAM, ReRAM, FERAM

Several memory technologies have emerged apart from Optane's underlying PCM technology, such as MRAM, FERAM (Ferro-Electric RAM), and ReRAM (Resistive RAM) (see Figure 1, [2]). Typically, these are nonvolatile memories that offer DRAM-like access speeds and have roadmaps for increased densities. They aim to surpass DRAM's scaling limitations and consume less power compared to the continuously refreshed DRAM and SRAM. Several companies have seen moderate level of financial success in their emerging memory efforts. Everspin® and Renesas® sell MRAMs, Adesto® sells ReRAMs, and Infineon® has a successful FERAM business.

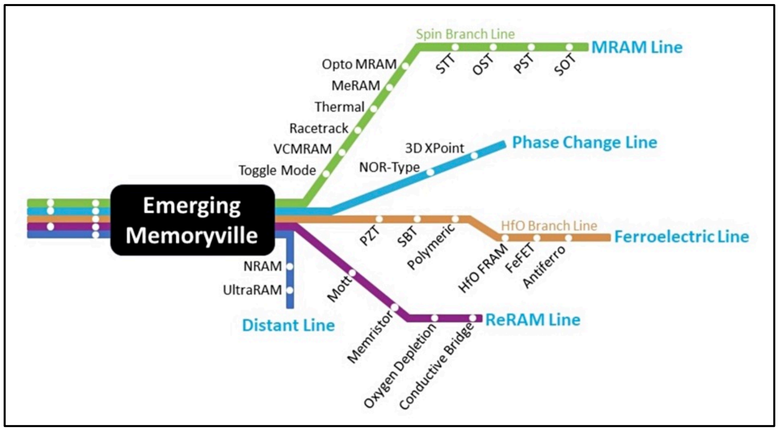
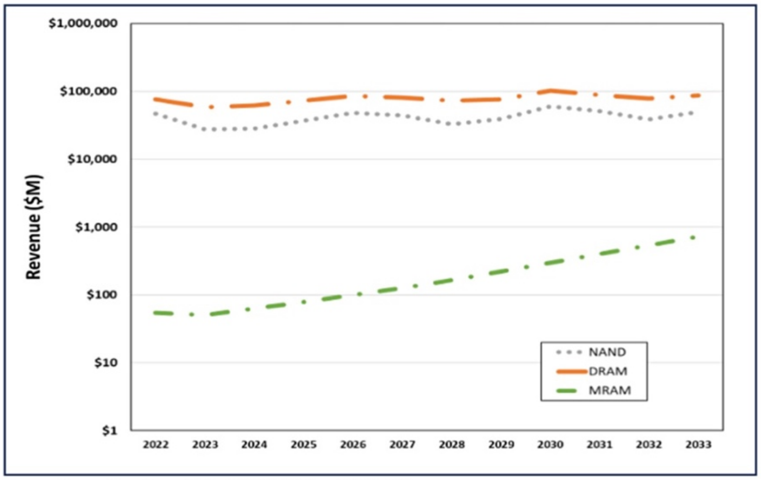


Figure 1: Emerging Memory Technologies

These technologies are usually chosen for applications that require nonvolatile memory that either consumes very low energy or can be exposed to high levels of radiation. Radiation is primarily a concern in space, but it’s also an issue for surgical instruments. While these applications have limited volume, low energy consumption is essential for wearable devices, ranging from wearable fitness devices to pacemakers and IoT endpoints. These are fast-growing markets. Nonetheless, none of these markets currently consumes a significant amount of memory and this limited demand is projected to persist in the near future, preventing these technologies from reaching volumes crucial for cost reduction. The challenge is figuring out how to expand the use cases of these emerging memory technologies beyond their current niche markets.

None of the emerging memory technologies are particularly new, yet none have crossed the chasm into mass production. MRAM has the largest chance of becoming widely used. In 2022, 133TB of MRAM was produced, yielding $118M in revenues. This could rise to 4.56EB and revenues of $980M by 2033. In contrast, DRAM and NAND revenues dwarf these figures (see Figure 2), both nearing $100B.



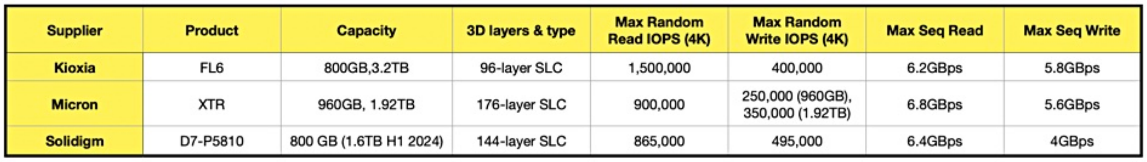
Comparison of projected DRAM, NAND and MRAM revenues

Figure 2: Projected DRAM, NAND, and MRAM revenues

3.2 SLC SSD

Solidigm® has unveiled its SLC SSD, D7-P5810, following similar product releases from Micron (XTR in 2023) and Kioxia® (FL6 in 2021) (see Figure 3, [3]). These high-performance SSDs are designed as the SCM for write caching and metadata store, as well as standalone storage for HPC and AI/ML applications.

With the rising adoption of QLC SSDs for capacity storage, SLC NAND can serve effectively as a fast write cache in front of the slower QLC SSDs. For example, SLC SSDs can be paired with QLC SSDs in a CSAL (Cloud Storage Acceleration Layer) deployment. CSAL, an open-source software, utilizes SCM as a write cache to reshape various write workloads into large sequential writes, enhancing the performance and endurance of QLC SSDs. Furthermore, SLC SSDs can also serve as the SCM for write caching and metadata store in enterprise storage systems. For instance, VAST DATA has been employing Kioxia's FL6 as the write cache and metadata store its Disaggregated Shared Everything (DASE) systems, replacing its initial use of Optane SSDs.



Supplier	Product	Capacity	3D layers & type	Max Random Read IOPS (4K)	Max Random Write IOPS (4K)	Max Seq Read	Max Seq Write
Kioxia	FL6	800GB,3.2TB	96-layer SLC	1,500,000	400,000	6.2GBps	5.8GBps
Micron	XTR	960GB, 1.92TB	176-layer SLC	900,000	250,000 (960GB), 350,000 (1.92TB)	6.8GBps	5.6GBps
Solidigm	D7-P5810	800 GB (1.6TB H1 2024)	144-layer SLC	865,000	495,000	6.4GBps	4GBps

Figure 3: Specifications for Kioxia FL6, Micron XTR, and Solidigm D7-P5810

Solidigm's D7-P5810 boasts an endurance of 50 drive writes per day (DWPD) for random writes, higher than Micron's 960GB XTR's 35 DWPD. In terms of latency, Solidigm's D7-P5810 offers 53µs for reads and 15µs for writes. Micron's XTR comes in at 60µs for reads while matching the 15µs write latency. Kioxia's FL6 stands out with 29µs read latency and 8µs write latency, making it the fastest among the trio in latency.

SLC SSD makes sense as a write cache in front of QLC SSDs, but its read latency of 29µs or higher makes it less ideal as a metadata store, especially when compared to Optane's 6µs read latency. This higher read latency partially explains why VAST Data excels in read bandwidth yet struggles with IOPS performance. For metadata store, memory-semantic SSDs might be a better choice, which we will discuss next.

4 Compute Express Link

4.1 Current Architecture and Limitations

The current CPU centric server architecture, along with its tightly coupled memory subsystem, has several limitations, particularly when it comes to virtualized environments, in-memory databases, emerging AI/ML applications, and next-gen storage architectures.

In the current architecture, a server platform houses one or more CPUs, each consisting of multiple CPU cores. Each CPU is connected to DDR/DIMM memory via multiple memory channels. Together, the CPU, its memory channels, and its local memory constitute a NUMA (Non-Uniform Memory Access) node. Memory access within a NUMA node benefits from low latency, while cross-NUMA memory access introduces additional latency[2]. Multiple memory channels within a NUMA node are typically interleaved to maximize memory bandwidth. However, as the average CPU core count has increased by 3x over the past eight years, the memory bandwidth has only increased by 2x, resulting in lower memory bandwidth per CPU core. CPU cache, a small, high-speed memory located on the CPU chip, is used to bridge the gap between the CPU's processing speed and the available memory bandwidth.

PCIe, a high-speed serial bus interconnect, is used to connect CPUs to network cards, storage controllers, as well as specialized computational devices such as GPU, FPGA, and ASICs, each maybe equipped with its own local memory. Data transfer between CPU memory and the local memory on a PCIe device via an DMA (Direct Memory Access) engine entail relatively higher overheads and latencies. Furthermore, interconnecting multiple servers typically involves using Ethernet or InfiniBand, which exhibit even higher latencies and lower bandwidths.

The current server architecture faces several challenges. First, despite the growing demand for more memory capacity from emerging applications, such as in-memory databases and AL/ML applications, it has become exceedingly difficult to scale up DDR/DIMM capacity. According to Micron Datacenter Segment Lead Ryan Baxter, while 512GB capacity is possible in DDR5, it relies on the expensive TSV (Through Silicon Vias) stacking. The cost per bit goes nonlinear extremely quickly and will increase DRAM cost as a percentage of server costs from 35-45% to 60-70% [4]. Consequently, memory cell scaling has reached its physical limits (see Figure 4), and the

[2] NUMA-local latency is in the tens of nanoseconds while cross-NUMA access adds 60-80 nanoseconds.

maximum number of DIMMs per memory channel is decreasing[3]. Moreover, increasing the number of memory channels and DIMMs faces space, power, and thermal constraints[4]. As a result, memory capacity expansion requires external memory solutions that are disaggregated from the CPU complex and freed of space, power, thermal limitations.

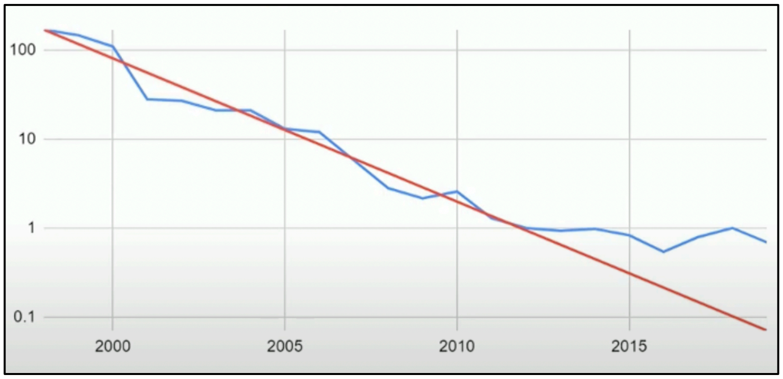


Figure 4: Memory Price per Bit has stayed flat since 2013.

DRAM has also become a major portion of hardware costs due to its poor scaling properties, often representing 50% of server costs. *Memory stranding,* caused by the current tightly coupled memory architecture, contributes predominantly to memory waste and increased server costs in virtualized environments. Memory stranding occurs when all CPU cores of a server are allocated to VMs, but unallocated memory capacity remains and cannot be utilized by other servers. Studies have shown that up to 25% of DRAM can become stranded [5]. *Memory pooling*, which allows VMs to access available memory beyond server boundaries, promises to improve memory utilization and bring substantial cost savings.

In addition to memory pooling, *memory sharing* is a valuable capability with use cases in in-memory databases and enterprise storage. Unlike memory pooling, which assigns non-overlapping external memory segments to multiple CPUs, memory sharing enables multiple CPUs to access the same external memory segment. In the case of enterprise storage, multiple storage controllers can leverage memory sharing to achieve cache-coherent, low-latency, shared access to in-memory metadata, thereby improving the overall system performance.

Furthermore, there is a growing demand for new memory media that can bridge the cost/performance and persistence gap between DDR/DIMM and NAND SSD (see Figure 6). While the interleaving of memory channels is necessary to maximize memory bandwidth, it is optimized for the same generation of DDR memory, making it impractical to rely on memory channels to support new media memory. External memory solutions are required to match the performance, capacity, and persistence capabilities of emerging memory technologies.

[3] The max number of DIMMs per channel has decreased from 3 for DDR3, 2 for DDR4, to 1 for DDR5.
[4] Each memory channel requires 200+ pins; DIMM slots have a power budget of 15-18W.

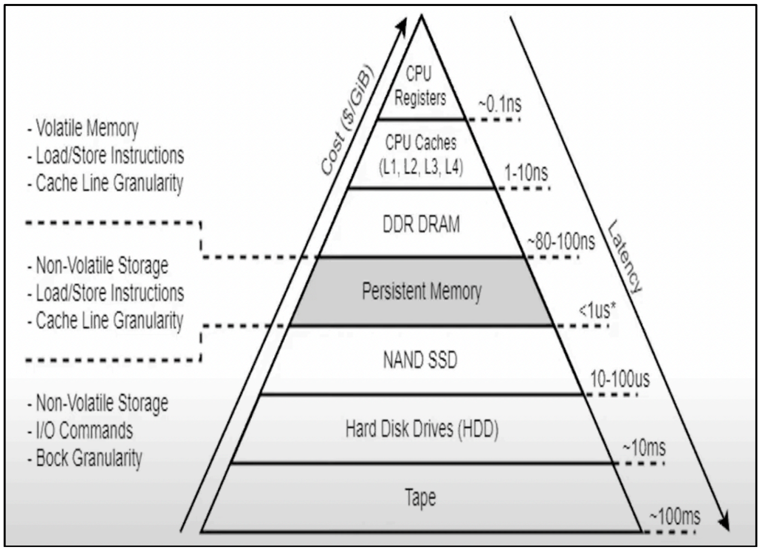


Figure 5: Cost/Performance Gap between DDR Memory and NAND SSD

Lastly, AI/ML and analytics applications require specialized computational devices, such as GPU and FPGA, often referred to as *accelerators*, to scale computing power. However, CPUs and accelerators cannot directly access each other's memory and have to rely on DMA to transfer data between the CPU memory and the local memory of an accelerator. DMA data transfer is less efficient for small data access, such as retrieving the computational result from a GPU. Furthermore, interconnecting multiple servers and accelerators through Ethernet or InfiniBand introduces higher latency and lower bandwidth compared to PCIe. Addressing these challenges calls for cache-coherent, byte-addressable memory access that allows CPUs and PCIe devices to access each other's memory with minimal latency.

4.2 Compute Express Link

The existing tightly coupled memory architecture faces challenges when it comes to supporting memory expansion, memory pooling, memory sharing, new memory media, and accelerators. Addressing these challenges calls for a new memory interconnect that disaggregates external memory from the CPU complex and supports cache-coherent, low-latency memory access.

There have been several attempts in the industry to introduce a new memory interconnect. In 2014, Nvidia® introduced NVLink to enable fast data transfer and access for GPUs. IBM® led the OpenCAPI standards effort in 2016 while HPE® was behind the Gen-Z initiative. AMD® developed a proprietary solution called Infinity Fabric as well as supported the open CCIX solution with Arm®. No industry ecosystem could develop around these solutions due to the lack of buy-in from Intel, which held over 90% CPU market share. In 2019, Intel took a significant step by donating their proprietary specification known as Compute Express Link (CXL) 1.0 to the newly formed CXL Consortium. Since then, CXL has garnered extensive support from various industry players, including AMD, Arm, Nvidia, Micron, Samsung®, Sky Hynix®, Microsoft®, Meta®, Google®, and other vendors. OpenCAPI and Gen-Z have subsequently been absorbed into CXL. This widespread industry endorsement has positioned CXL as the most promising memory interconnect solution.

CXL piggybacks off the existing ecosystem of PCIe 5.0 by using its physical and electrical layer, but with improved transaction and data link layers. Through the release of version 1.0/1.1 (March 2019), 2.0 (November 2020). and 3.0 (August 2022), CXL has introduced the following capabilities [6]:

High Bandwidth, ***Low Latency*** – CXL is aligned with 64GBps PCIe 5.0 and is expected to drive the adoption of 128GBps PCIe 6.0. CXL has been designed from the ground up to provide low-latency access. Each CXL data link frame, known as a FLIT (Flow Control Unit), consists of a 64-byte payload aligned with the CPU cacheline size, along with a small overhead (2-byte CRC and 2-byte header). The 64-byte granularity enables low-latency and cache-coherent access. The use of fixed frame size and CRC contributes to faster processing compared to the variable-sized frames used in PCIe data link layer.

Cache Coherency – With CXL 1.0, a host can access the local memory on an accelerator and an accelerator can access the host memory in a cache coherent manner. Both the host and the accelerator see the most up to date data. A home agent ensures that data cannot be changed simultaneously, and when a change occurs, it's propagated to all the cached copies of the data. CXL 3.0 enhances cache coherency by allowing CXL devices to *back-invalidate* data that's being cached by multiple hosts.

Heterogeneous Memory – The serial nature of PCIe bus allows CXL-attached memory device to be placed at more optimal locations in EDSFF (Enterprise & Datacenter Storage Form Factor) format, free from the space, power, and thermal constraints of the CPU complex. Compared to DIMMs, EDSFF has a larger power budget (25W+), scales up better, and makes hot-plug more feasible. CXL is a transactional protocol that allows CXL-attached memory devices to have different latency and bandwidth profiles. CXL-attached memory devices can report their QoS synchronously to hosts to prevent head of line blocking.

Persistent Memory –CXL 2.0 introduced *global persistent flush* in support of persistent memory. In the event of a power loss, a global persistent flush event is sent across the cache coherency domain to flush cached data from CPU cache, CXL device cache, and memory device write buffers.

Memory Pooling – CXL 2.0 added support for pooling accelerator and memory resources across multiple hosts. Non-overlapping segments on a CXL-attached memory device can be assigned dynamically to different hosts and moved from host to host in a hot-plug flow (see Figure 7). Memory pooling reduces memory stranding, improves memory utilization, and ultimately lowers server costs.

Memory Sharing – CXL 3.0 introduced memory sharing, where CXL-attached memory can be coherently shared across multiple hosts using hardware coherency. Unlike memory pooling, memory sharing allows the a given segment of memory to be simultaneously accessed by more than one host and still guarantee that every host sees the most up to date data at that location, without the need for software-managed coordination. When a host wants exclusive write access to a shared memory segment, a *back-invalidate* flow is launched to all the other hosts to ensure cache coherency.

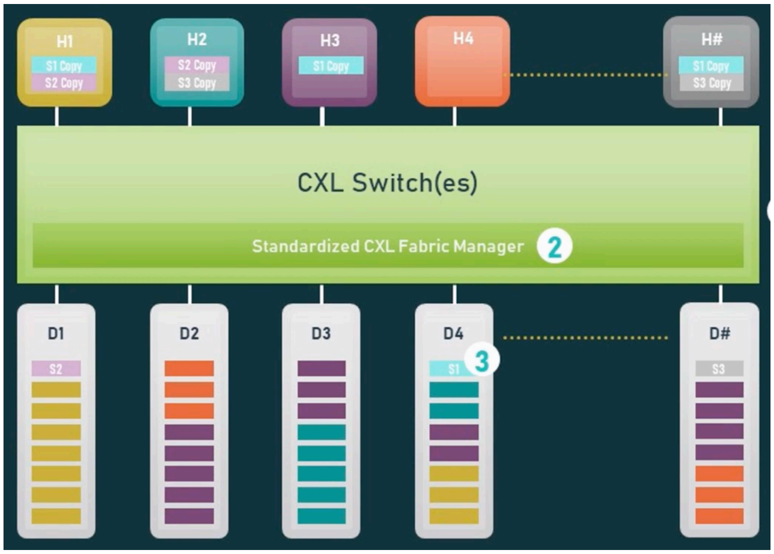


Figure 6: CXL Memory Pooling and Memory Sharing

Fabric Connectivity – CXL 1.0 was focused on single host connectivity. Subsequent CXL 2.0 and CXL 3.0 introduced single level switching and multi-level switching, respectively. Up to 4095 fabric-attached devices can be supported. The serial nature of PCIe, as compared to the parallel memory channel, allows CXL to extend its reach to an entire rack. Going forward, multi-rack connectivity may become feasible with the use of optical connections between racks.

Editor's Note: CXL is still in its early stages, with most real-world implementations based on CXL 1.1 in the form of memory expansion to a single host. Uncertainties remain regarding CXL 2.0 and CXL 3.0, such as switch latency and memory sharing overhead, which only real-world implementations will clarify. In the next section, we will explore CXL's existing and potential future use cases.

4.3 Use Cases

CXL enables various use cases including SmartNIC, accelerators (such as GPUs and FPGAs), and memory devices. In this section, we will focus on memory use cases. Among these use cases, some already have vendor products available, while others are still awaiting implementations.

4.3.1 Memory Expansion

Sky Hynix has made its CXL memory card available for sampling. The CXL memory card features 96GB DDR5 capacity packaged in the ES.3 EDSFF format and supports PCIe 5.0 x 8 Lane connectivity. In a memory expansion use case, an X86 CPU can have eight DDR5 channels with 768GB of DRAM capacity (8x 96GB DIMMs) and a total bandwidth of 320GBps. By adding 4x 96GB CXL memory cards and implementing memory interleaving across local and CXL memory, the server can reach a total of 1.15TB memory capacity and 480GBps bandwidth.

4.3.2 Memory Tiering

The Microsoft Azure team has identified memory stranding as a primary cause of memory waste and a potential source of cost savings in a virtualized environment. Memory stranding occurs when all cores of a server have been allocated to VMs, but unallocated memory capacity remains and cannot be utilized by other servers. The team discovered that up to 25% of memory becomes stranded across 100 production cloud clusters at Azure.

CXL-based memory pooling promises to improve memory utilization and thereby reduce costs. However, memory pooling faces challenges in meeting customer performance requirements, as CXL accesses experience higher latencies than same-NUMA-node accesses. Specifically, CXL adds 70-90ns to access latencies over same-NUMA-node DRAM within an 8-16 CPU pool and more than 180ns for rack-scale pooling.

The Microsoft team has devised a memory tiering system that dramatically reduces the impact of this higher latency. The memory tiering system relies on a machine learning model that identifies a subset of latency-insensitive workloads and cold memory pages within a VM to be allocated on CXL memory. This allows the system to both meet customer performance goals and significantly reduce memory cost. As a result, the memory pooling reduces memory cost by 7% while maintaining performance within 1-5% of the optimal performance without memory pooling. This translates into an overall reduction of 3.5% in cloud server cost.

It's important to note that the memory tiering system is implemented on top of an emulation layer on production servers given that CXL memory pooling devices are still some time away from deployment. The implementation allows the Microsoft team to evaluate key concepts and demonstrate their potential benefits.

4.3.3 Near Memory Compute

Panmnesia®, a Korean startup, claims to have developed a CXL-based near memory compute solution that speeds up vector search methods by 110x and recommendation models by 5x. Panmnesia was founded to commercialize technologies initially developed at the Korea Advanced Institute of Science & Technology. [7]

Recommendation models rely on embedded vectors, which are multi-dimensional numeric values used to describe complex data items such as words, phrases, paragraphs, images, or videos. A recommendation model takes an input vector and searches for similar items in a vector database. Microsoft, for example, manages over 100 billion vectors in its search engine, which consumes more than 40 terabytes of memory space.

Panmnesia has developed a CXL memory controller that provides much larger external memory compared to the local memories found on CPUs and GPUs. The CXL-attached memory allows the data for a large recommendation model to be stored in memory rather than much slower SSD storage. Panmnesia's memory controller also features near memory processing capabilities that preprocesses vector data in its local memory, reduces its size and complexity, and makes the results available for direct memory access by CPUs and GPUs. This near memory compute solution reduces latency, minimizes data movement, parallelizes vector search operations, and ultimately accelerates recommendation models.

4.3.4 Memory Semantic SSD

Originally Intel envisioned CXL as the ideal interconnect to Optane. Now without Optane, only two types of memory make sense in a CXL system: DRAM and something cheaper than DRAM. Today, all the emerging memory technologies are more costly than DRAM, rendering them unsuitable as CXL memory.

Both Samsung and Sky Hynix have been actively promoting their memory-semantic SSDs. Samsung showcased a prototype at last year's Flash Memory Summit. The prototype drive combines both DRAM and NAND with CXL connectivity (see Figure 8, [8]). The drive's DRAM is accessed with load-store memory-semantic commands and functions as a cache for the NAND. Notably, load-store memory access features cacheline granularity at 64 bytes while the minimum I/O size for an NVMe SSD is 4KB. According to Samsung, the memory-semantic SSD provides 20x improvement in small random read performance compared to an NVMe SSD. To support data persistence, a battery backup unit is required to flush the DRAM cache.

Additionally, Samsung's memory-semantic SSD supports dual-mode interface, allowing its DRAM to be accessed with CXL memory load/store commands, while its NAND is accessed with NVMe read/write commands. Applications, such as DLRM (Deep Learning Recommendation Engine) applications, can leverage this dual-mode interface to direct fine-grained reads from its inference engine to the CXL memory and bulk data writes from its model update engine to the NAND. Samsung is targeting this drive for AI/ML applications that require fast processing of smaller data sets.

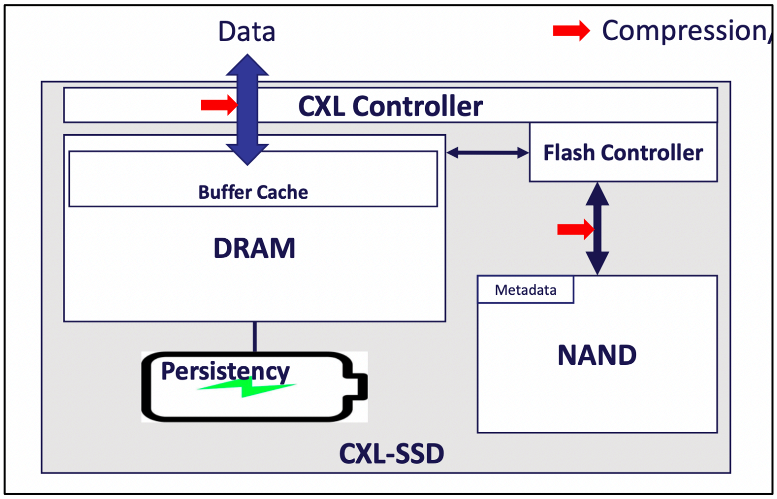


Figure 7: Memory Semantic SSD in Cache Mode

With CXL’s support for persistent memory and cache-coherent memory sharing, memory-semantic SSD can improve an enterprise storage system’s performance by providing low-latency, shared access to its metadata store.

5 Conclusion

After the demise of Intel Optane, two technologies have the potential to emerge as the storage class memory options for enterprise storage. SLC SSD makes sense as a write cache in front of QLC SSDs, but its relatively higher read latency makes it less ideal as a metadata store. For metadata store, CXL-based memory-semantic SSDs might be a better choice. With CXL’s support for persistent memory and cache-coherent memory sharing, memory-semantic SSD can improve an enterprise storage system’s performance by providing low-latency, shared access to its metadata store.

.

BIBLIOGRAPHY

[1] Research Whitepaper from Jim Handy, Objective Analysis
[2] Emerging Memory Technologies, Blocks and Files
[3] SLC SSDs, Blocks and Files
[4] Micron Datacenter Lead Ryan Baxter Interview with Blocks & Files
[5] Microsoft CXL Memory Pooling Whitepaper
[6] CXL Consortium Presentations and Whitepapers
[7] Panmnesia Memory Pooling Whitepaper
[8] Samsung Memory Semantic SSD Presentation, Mass Storage 2023