

FAQ

Low-Latency NAND Flash as Storage Class Memory (SCM): Filling the Gap Between DRAM and Storage

Low-latency NAND flash as storage class memory (SCM) sits between DRAM and flash-based storage in the memory hierarchy, where it offers higher memory capacity than DRAM and lower latency than conventional flash. It provides high performance and scalability for hyperscale data centers and enterprise applications, and it also shows promise for AI computing at the edge. In this FAQ, learn about the classes of memory used in computing environments, the alternative types of NAND flash memory, and the benefits that low-latency NAND memory offers.

What is the memory hierarchy?

Various types of memory are used in most computing environments, offering trade-offs between access speed (latency), data storage capacity, and cost per gigabyte.

Essentially the memory hierarchy is structured organization of data storage in computer systems, designed to balance speed, capacity, and cost. It arranges memory from fastest (which is also the lowest capacity and most expensive) at the top to the slowest (highest capacity and least expensive) at the bottom. SRAM (static random-access memory) and DRAM (dynamic random access memory) are volatile memories, meaning they lose their data when power is off, while secondary storage, such as triple-level cell (TLC)/quad-level cell (QLC) flash memory and hard disk drives (HDDs), are non-volatile, meaning they retain

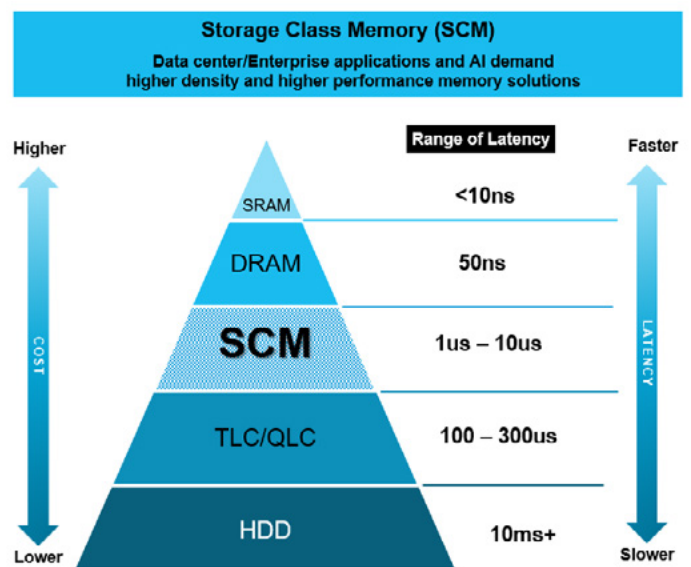
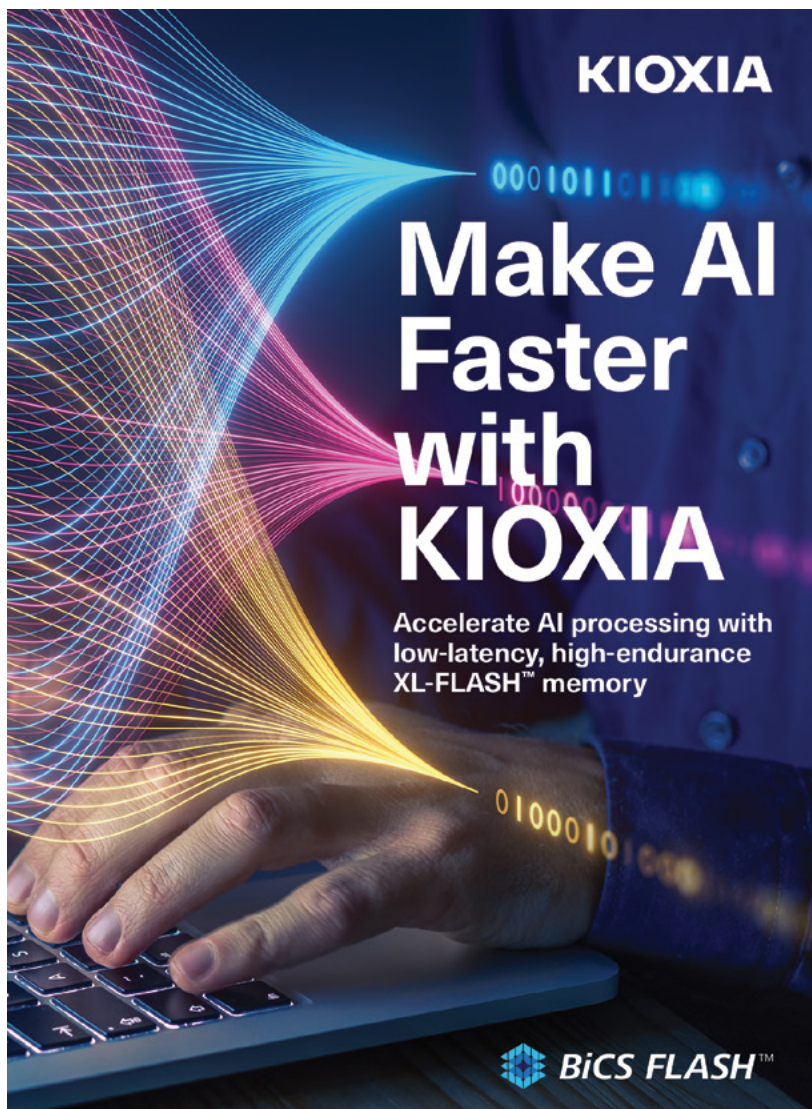


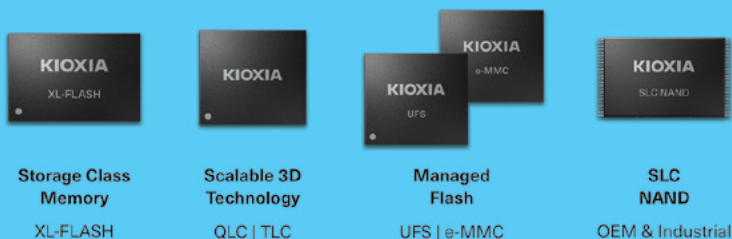
Figure 1. SCM fills the gap between DRAM and TLC/QLC flash in the memory hierarchy.



Optimized Storage Class Memory

KIOXIA XL-FLASH storage class memory delivers low-latency, high-endurance storage that is ideal for performance and latency-sensitive applications. As part of the broad KIOXIA portfolio, XL-FLASH memory enables scalable, reliable solutions for AI, data centers, high-IOPS SSDs and more.

KIOXIA Flash Memory Solutions



kioxia.com

Images are for illustration purposes only.
Copyright © 2026 KIOXIA America, Inc. All rights reserved.

data even when powered down. SRAM and DRAM are also limited in capacity and cannot scale like non-volatile memories.

Another important factor is power efficiency. In the memory hierarchy, power consumption decreases on a per-gigabyte basis as it descends from top to bottom. The hierarchy also has a clear gap between DRAM and most flash-based storage. SCM is intended to bridge this gap between fast volatile memories and slow non-volatile memories.

What is low-latency NAND memory, and how is it addressing this issue?

Low-latency NAND memory is a type of SCM that meets the need for memory with higher capacity than is economically achievable with DRAM but with much lower latencies than other flash-based storage. Low-latency NAND memory offers the endurance and non-volatility of NAND flash memory but with access speeds closer to those of DRAM.

What are the applications and use cases for this technology?

Key applications for using low-latency NAND memory as SCM would include the massive server farms fielded by hyperscalers to host artificial intelligence/machine learning (AI/ML) applications. Low-latency NAND memory can also be applied to run large databases or provide a variety of other cloud services. Additional potential applications include industrial automation, digital healthcare, transportation, surveillance, 5G networking, and write-intensive operations such as data logging and caching.

What is the future for low-latency NAND memory?

The future presents significant opportunities for low-latency NAND memory to meet market needs not addressed by other

Sponsored by

KIOXIA

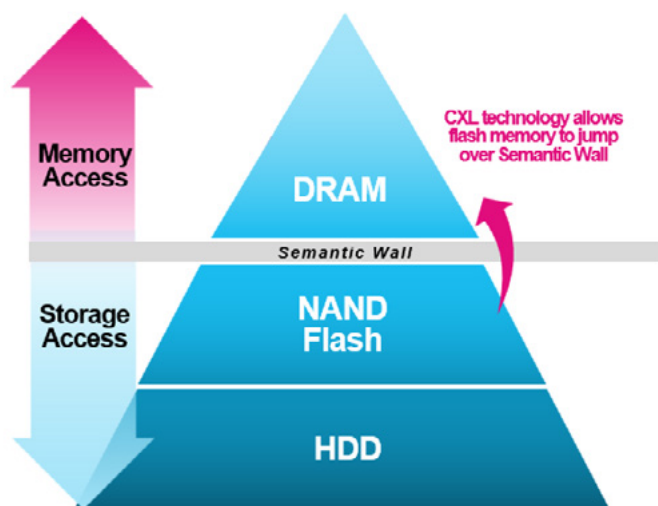


Figure 2. CXL technology enables low-latency NAND memory to overcome the semantic wall between DRAM and flash.

options. For example, low-latency NAND memory will continue to meet the escalating performance and capacity demands of AI and hyperscale data centers. Applications could extend down to general-purpose computers, for which low-latency NAND memory could represent a cost-effective, high-density solution for memory expansion.

Some low-latency NAND memory solutions also include a Compute Express Link (CXL) interface that will become increasingly useful in memory expansion. Memory has traditionally been expanded using the dual in-line memory modules (DIMMs), but using DIMM slots imposes expansion constraints that CXL overcomes, leading to its adoption. CXL technology enables flash memory to overcome the “semantic wall” that results from the different ways a host processor communicates with DRAM versus flash, even low-latency flash capable of small block reads, and high input/output operations per second (IOPS).

The low-latency, high-performance, and high-endurance attributes of this flash technology could also be used to create solid-state drives (SSDs) that deliver up to 100 million or more IOPS for near-GPU AI application.

Finally, the robotic industry has received attention lately as robots evolve from robotic vacuum cleaners, lawn mowers, and pool cleaners to autonomous humanoid configurations. Such robots will have to run heavy AI workloads, often without an internet connection and therefore without cloud support. SCM could cost-effectively provide them with the high-capacity memory they need to efficiently execute AI workloads at the edge. For all these applications, low-latency NAND provides the optimum balance of speed, capacity, and cost.

Need more information?

Learn More >