the Future of Memory and Storage

Memory Expansion with CXL® Interface with Low-latency Flash

Mahinder Saluja
Director of Technology and Storage Pathfinding
KIOXIA America, Inc.

August 2025

KIOXIA

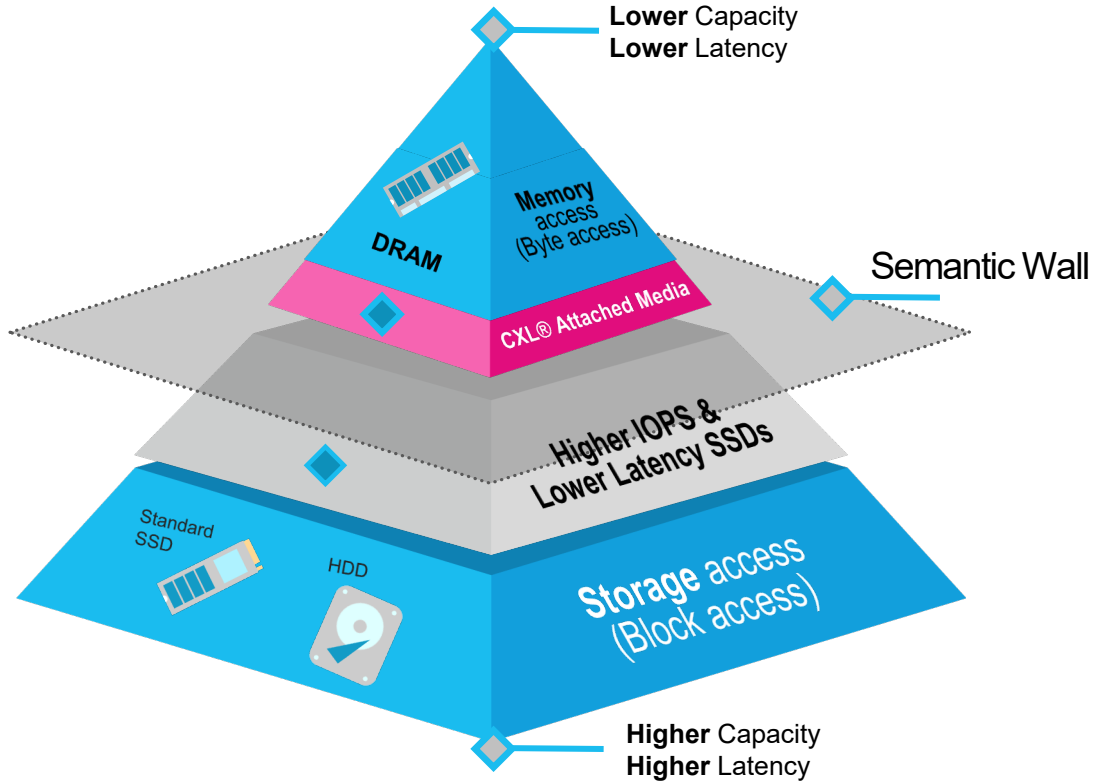
Agenda

- Infrastructure for Big Data Era
- Low Latency Flash Media
- System Stability and Applications
- Challenges and Opportunities

Dynamic Infrastructure for Big Data Era

The conventional infrastructure requirements are continuously evolving, so is the boundary between memory and storage.

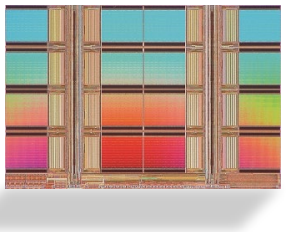
- High Bandwidth Memory and DRAM tier
- CXL® enables high-bandwidth and capacity media
- Higher IOPS SSDs optimized for GPU
- Fast SSDs for efficient checkpointing
- Ultra High-capacity SSDs : 128 TB[1] - 256 TB path to 1 PB

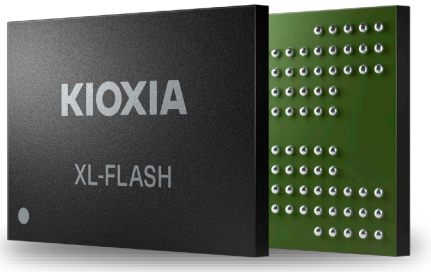


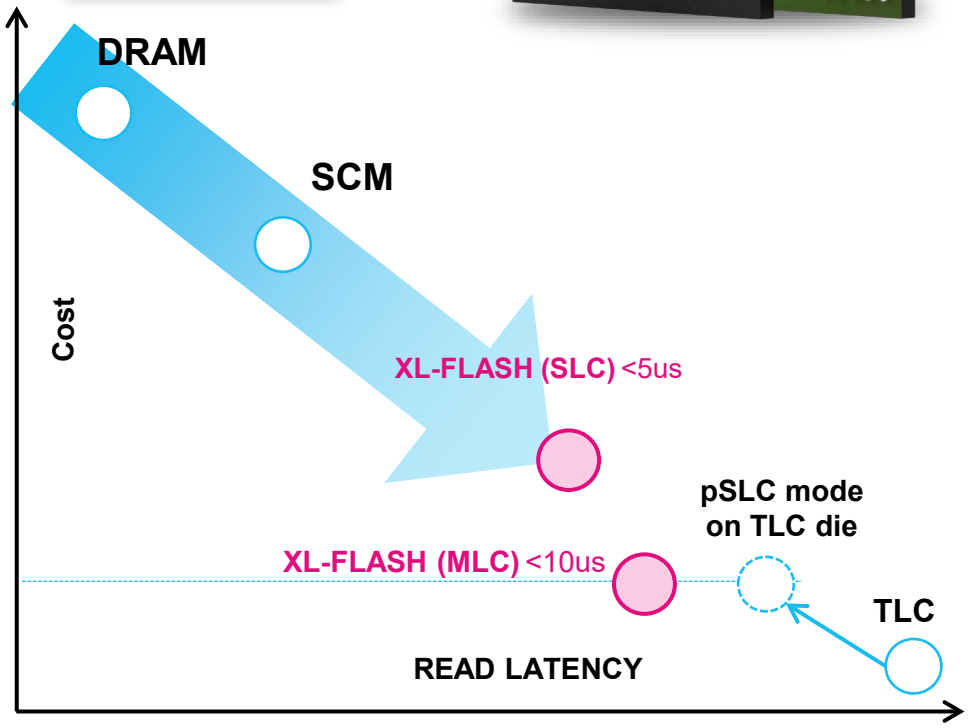
Can systems leverage CXL® -attached flash media for memory expansion?

 1. Definition of capacity: KIOXIA Corporation defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes, a terabyte (TB) as 1,000,000,000,000 bytes and a petabyte (PB) as 1,000,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1Gbit = 2^30 bits = 1,073,741,824 bits, 1GB = 2^30 bytes = 1,073,741,824 bytes, 1TB = 2^40 bytes = 1,099,511,627,776 bytes and 1PB = 2^40 bytes = 1,125,899,906,842,624 bytes and therefore shows less storage capacity. Available storage capacity (including examples of various media files) will vary based on file size, formatting, settings, software and operating system, and/or pre-installed software applications, or media content. Actual formatted capacity may vary. CXL and Compute Express Link are registered trademarks of the Compute Express Link Consortium, Inc..

Low Latency FLASH Introduction





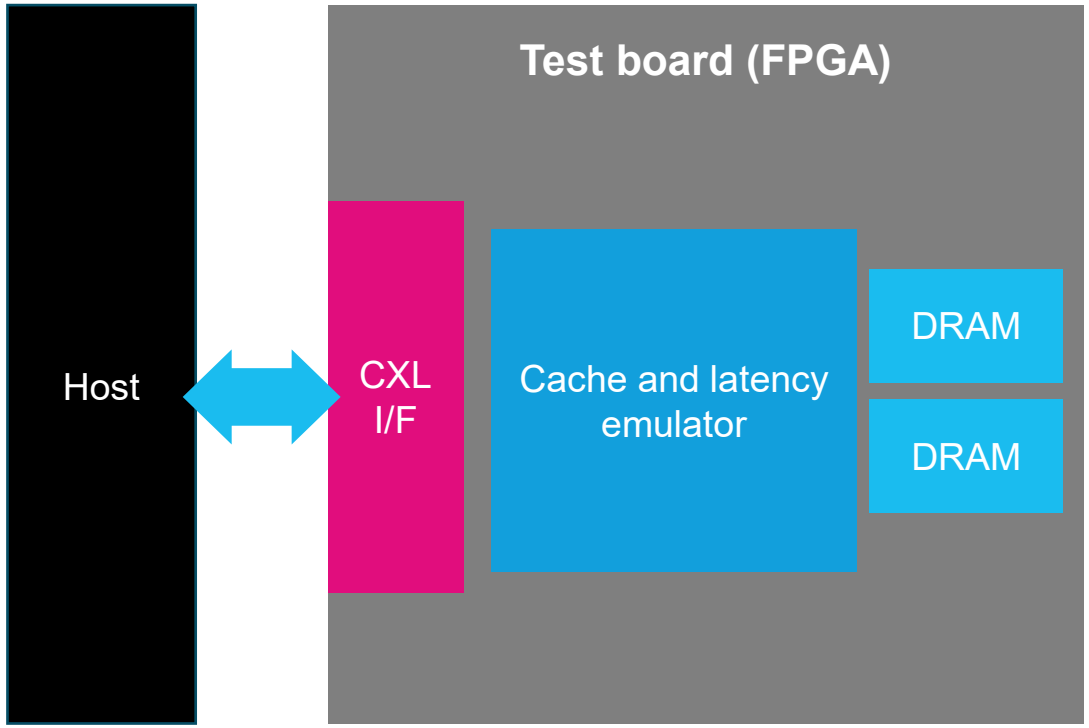


	2nd Gen. XL-FLASH	
	MLC	SLC
Capacity	256Gb[1]/die	128Gb/die
Page Size	4096B/16 Planes	4096B / 16 Planes
Read Latency	**<10 us**	**<5 us**

- Based on BiCS FLASH™ 3D flash memory technology
- 128Gb die (SLC) / 256Gb die (MLC) -- 2/4/8-die packages
- High cell reliability

Stress Testing in Progress

System Stability Test Environment with FPGA



Host* : SYS-741GE-TNRT: https://www.supermicro.org.cn/en/products/system/gpu/tower/sys-741ge-tnrt

Stress Testing Software:

- ✓ Open-source software kernel stress test suites
  - stress-ng, Linux® Test Project tests, xfstests, blktests
- ✓ Storage workload
  - FIO with numa_mem_policy, FIO hipri
- ✓ Network workload
  - iPerf3, Netperf, NetStress
- ✓ Memory workload
  - MASIM, FIO mmap
- ✓ Real world application benchmarks
  - Redis®, Memcached, SPEC CPU® 2017

All tests were run Independently and simultaneously

Results

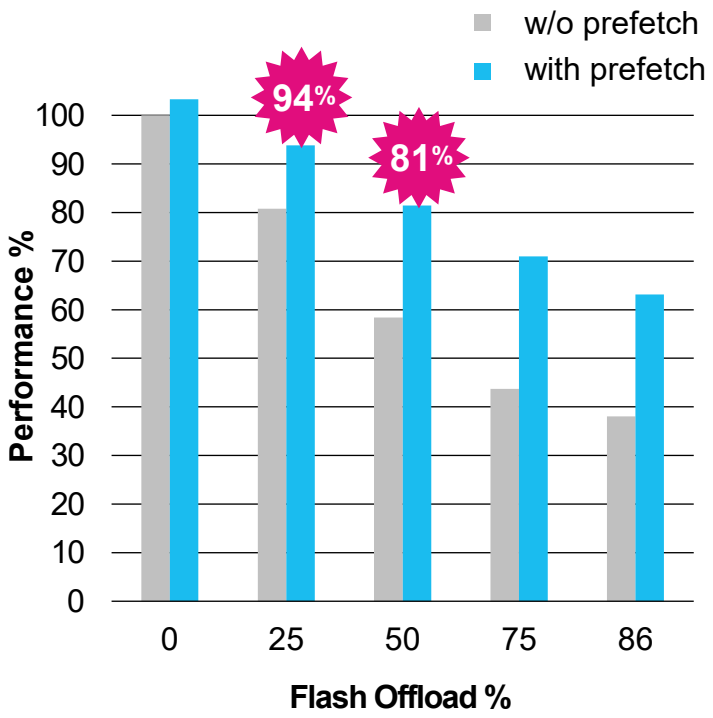
- No severe (i.e. non-recoverable) errors unique to CXL® memory in latest Linux kernels* up to 30us
- There were few warning and info level alerts from kernel like CXL® DRAM modules

Application Benchmark Redis™ In-Memory Database with Low Latency Flash

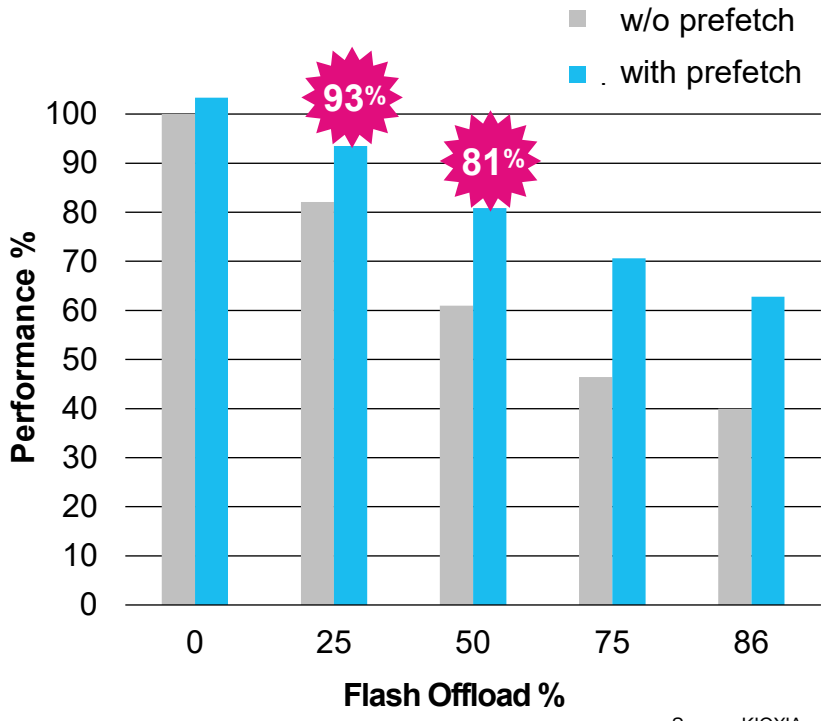
Tested with Yahoo!™ Cloud Serving Benchmark (**YCSB**) **tool**
Setup: 10M records(14 GB), 32 client threads

Data Type: 100B*10 fields/record
Offload with Linux® TPP (Transparent Page Placement)

Test C: Get 100%

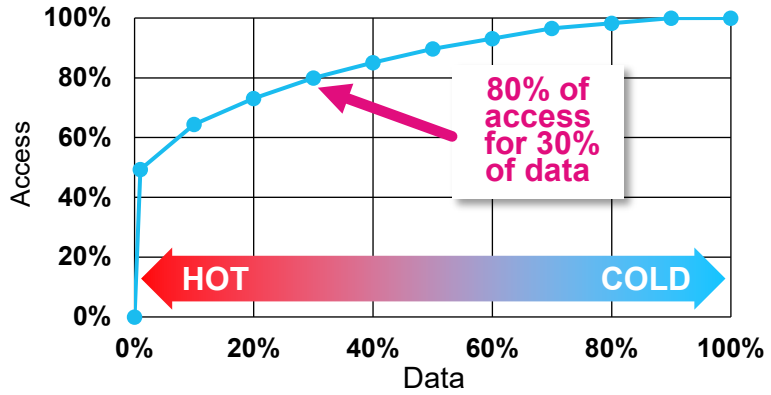


Test A: Get 50%, Put 50%

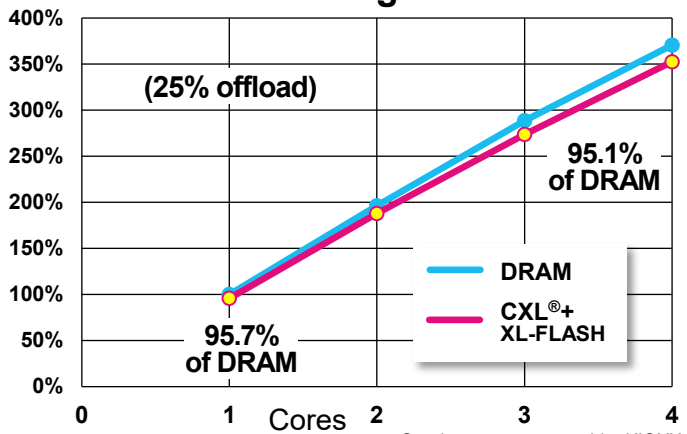


Source: KIOXIA

Zipf Distribution Workload A,C



Core Scaling for Redis

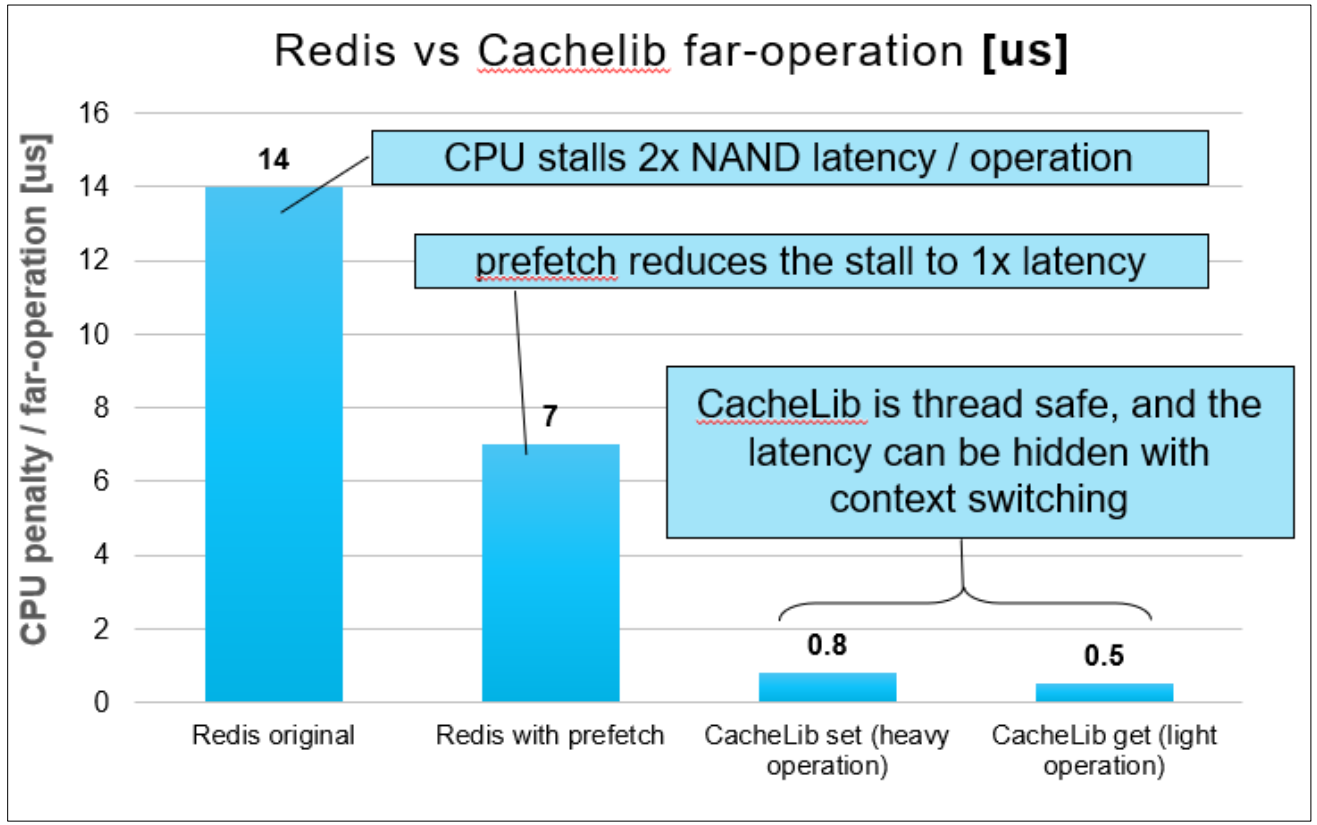


Graphs source: created by KIOXIA

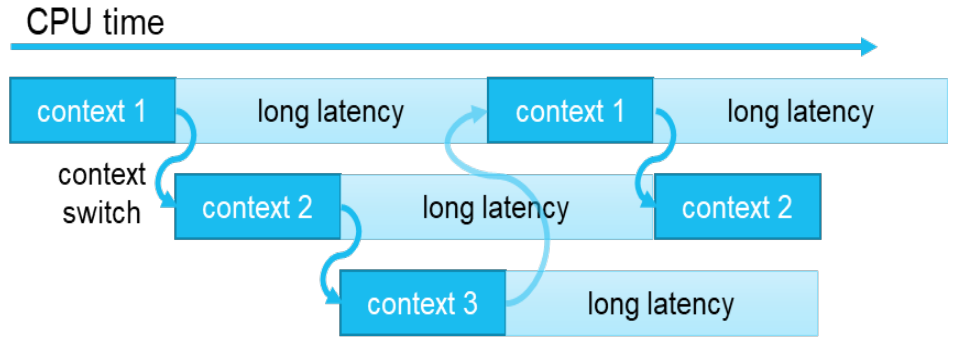
YCSB demonstrates CXL® and XL-FLASH technologies can offload 25% of memory with ~5% of performance degradation.

Application Benchmark CacheLib with High Latency Memory

CPU-time Penalty per Operation Targeting Far Memory



Hiding the Latency with Multiple Context



✓ **How to hide long latency:**

- Run multiple contexts in a core
- Request data in far memory with prefetch instructions and switch to another context

Results

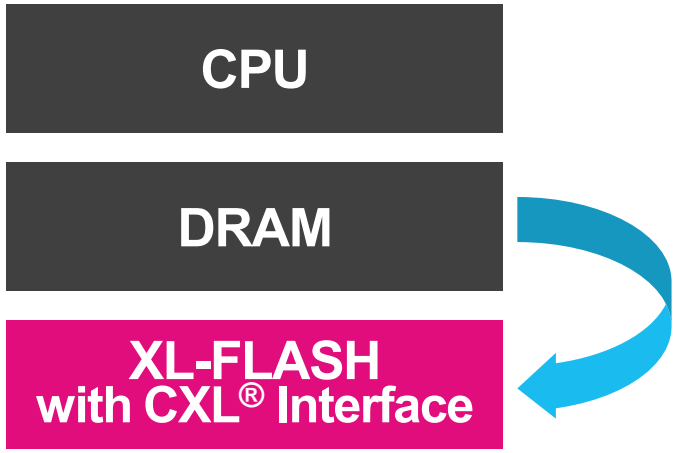
- CPU penalty is 0.5 to 0.8us per far operation in DRAM-CXL® tiering
- Even with 10M/s far operations, system performance will be 95% in a 128-core system

Estimation Result of Mix Applications for General Purpose Computing

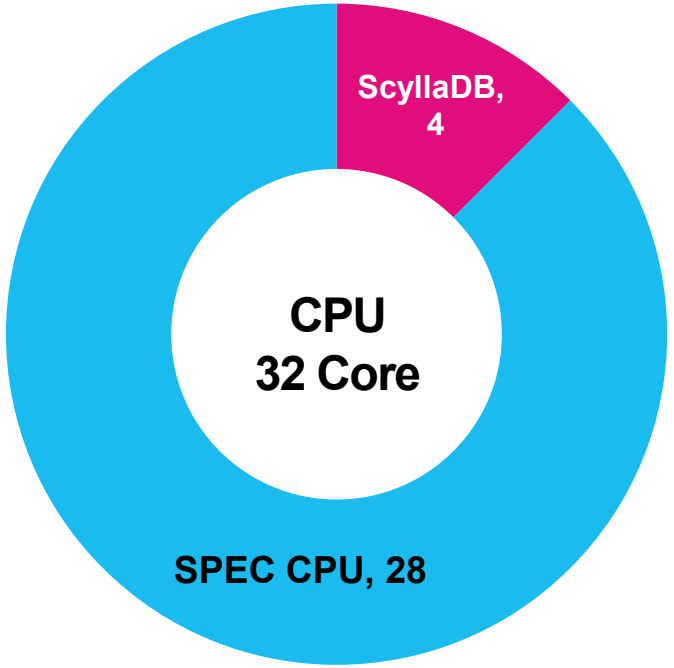
Mix applications for general purpose computing

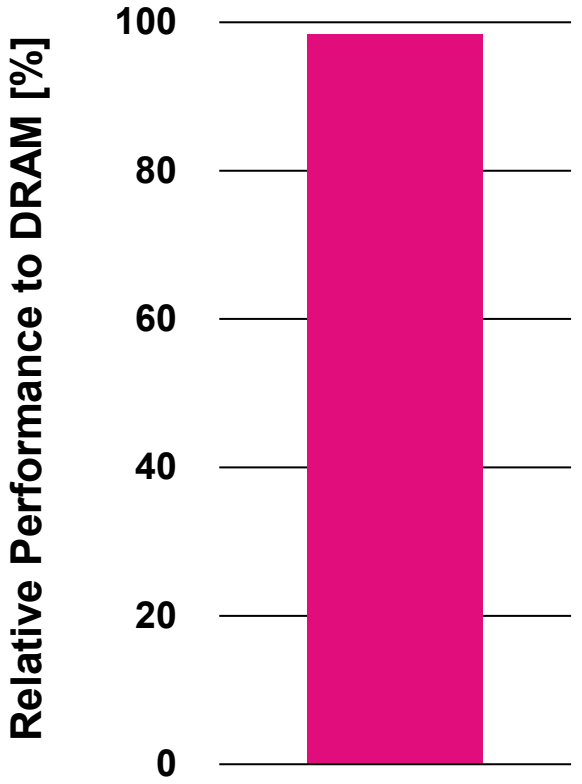
- Memory intensive application : ScyllaDB®
- Compute intensive application : SPEC CPU®

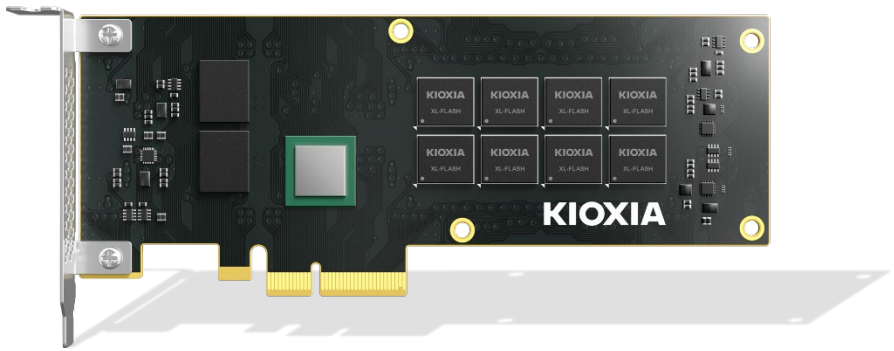
approx. **98.4% Performance with 33% Offload**



33%
Offload







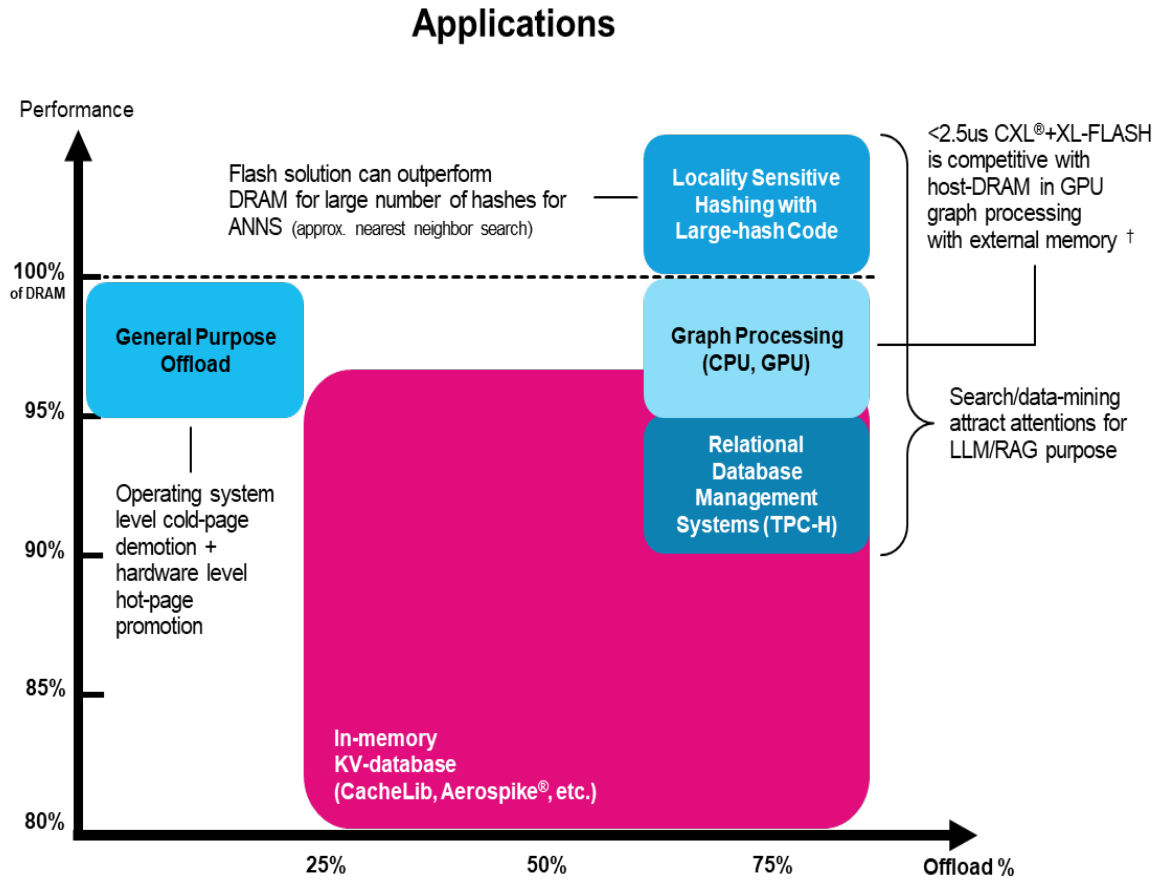
Challenges and Opportunities with CXL® Attached Flash Memory

All Applications Are Not The Same

- It is not suitable for latency/bandwidth sensitive applications
- Applications may not be tuned for leveraging memory hierarchy optimally

Leverage Industry Efforts

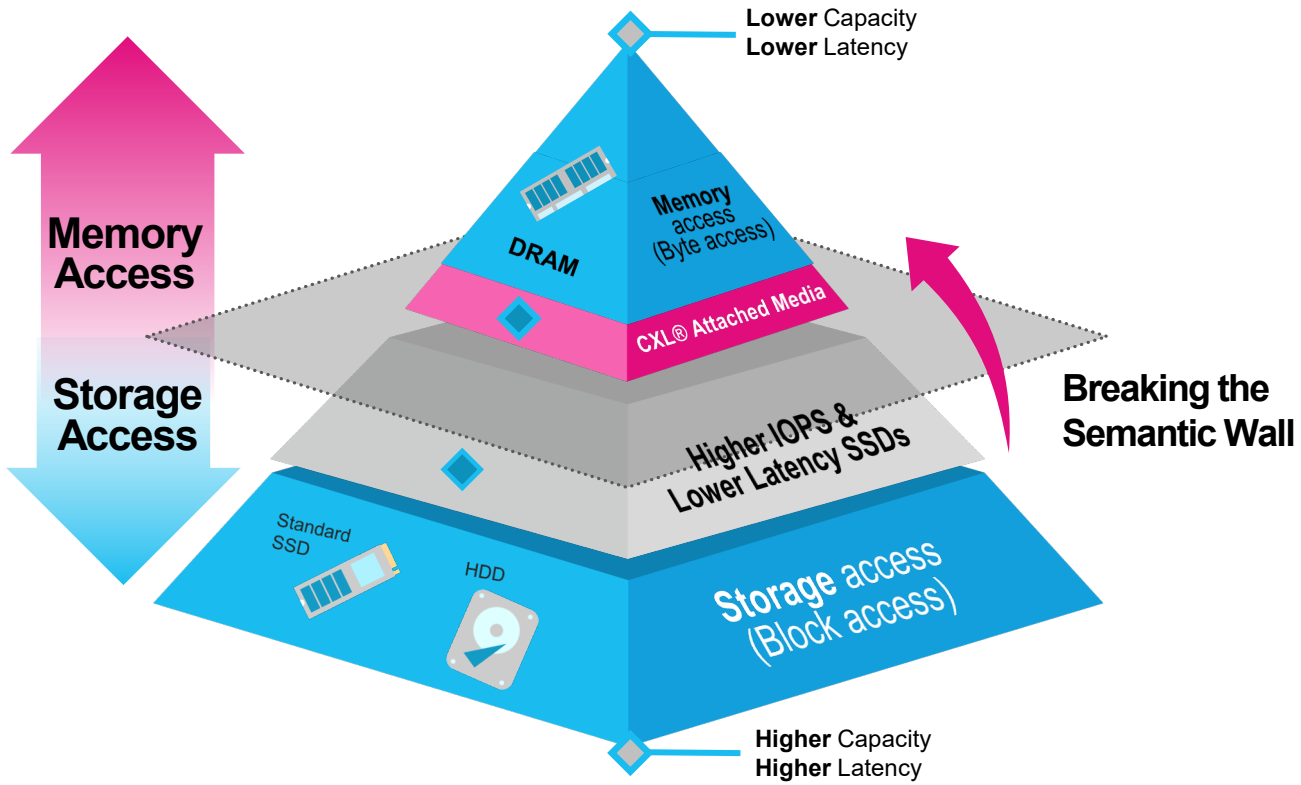
- Transparent Page Placement technique automatically manages large memory pages
- Transparent memory tiering techniques optimize data placement across different memory types
- Application specific libraries can further increase the efficiency and reduce cost



† GPU Graph Processing on CXL®-Based Microsecond-Latency External Memory
Source: KIOXIA testing engineers

- KIOXIA will be integrating CXL Hotness Monitoring Unit method hints-based patch to augment Linux kernel memory tiring

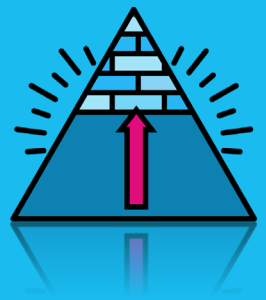
CXL® Bridges the Memory and Storage Semantic Wall



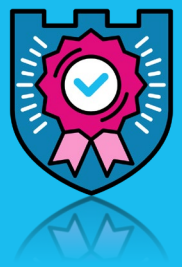
	CXL® and XL-FLASH Technologies
Media	BiCS FLASH™ 3D flash memory (XL-FLASH)
Value Pillar	Low latency <5us (single-level cell), <10 us (multi-level cell); DRAM cache tier
CXL® Access	CXL.mem, CXL.io (config)
Capacity	>= 512 gigabytes (GB[1])
Suitable Applications	In-memory data bases (DB), graph processing, cache, tiering, general purpose computing
Sample Availability	CQ2'26

- CXL® abstracts the media interface for systems
- Low Latency Flash media can break the semantic wall to expands the system memory cost effectively

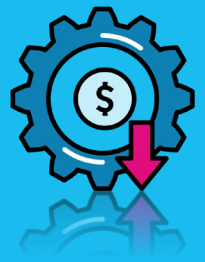
 1. Definition of capacity: Kioxia Corporation defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. Some computer operating systems, however, reports storage capacity using powers of 2 for the definition of 1GB = 2^30 bytes = 1,073,741,824 bytes and 1TB = 2^40 bytes = 1,099,511,627,776 bytes and therefore shows less storage capacity. Available storage capacity (including examples of various media files) will vary based on file size, formatting, settings, software and operating system, and/or pre-installed software applications, or media content. Actual formatted capacity may vary.



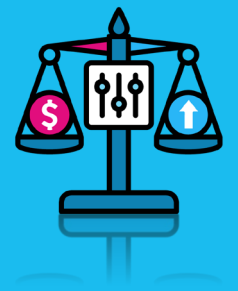
Flash memory can jump the semantic wall.



Flash memory is proven and reliable media.



Flash memory lowers the system TCO.



Flash memory can further perform and reduce cost with software.

If you are working on large memory intensive applications like **Data Mining, Analytics, High Performance Computing (HPC), Graph Processing Applications,** Please visit **KIOXIA Booth #307** for collaboration opportunities.

FMS
the Future of Memory and Storage

KIOXIA