KIOXIA

CXL® Attached Flash as Host Memory Expansion

- Mahinder Saluja, KIOXIA America, Inc.

AI LEADERSHIP FORUM



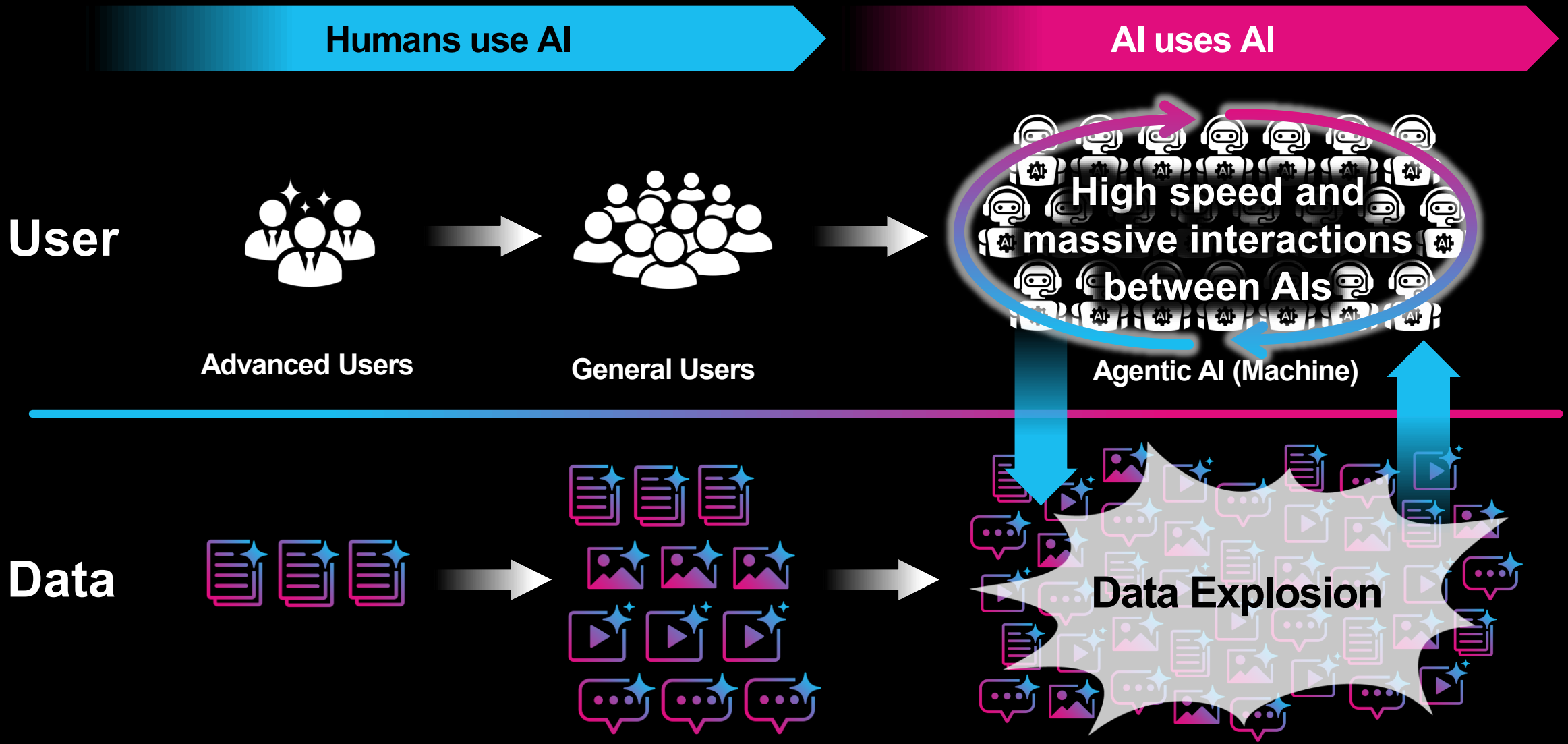
© 2026 KIOXIA Corporation. All Rights Reserved.

Agenda

- Evolution of AI Data Systems and Data Explosion
- DRAM and Flash Memory Outlook
- CXL® Attached Low Latency Flash Memory
- Applications and Use Cases

© 2026 KIOXIA Corporation. All Rights Reserved.

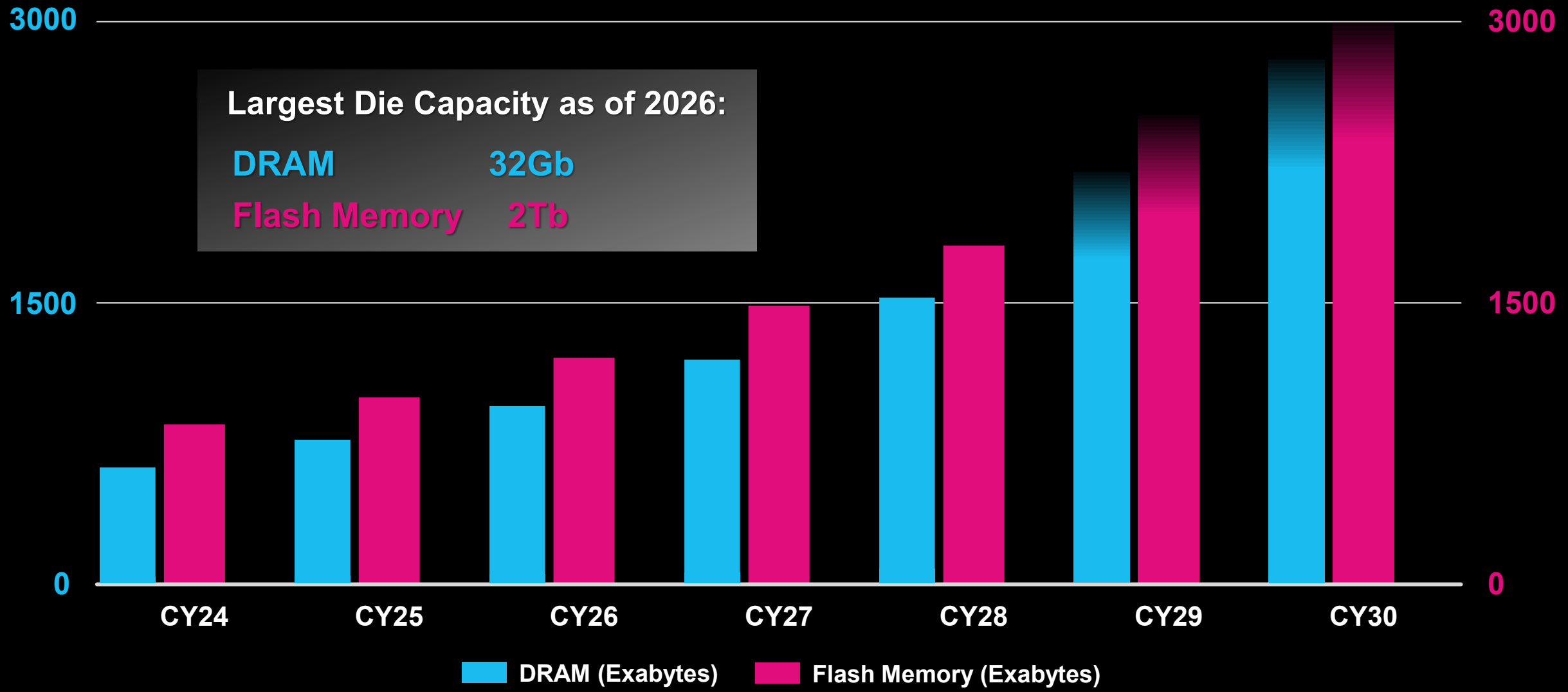
Evolution of AI Inference System and Data Explosion



KIOXIA

© 2026 KIOXIA Corporation. All Rights Reserved.

DRAM and Flash Memory Market Outlook

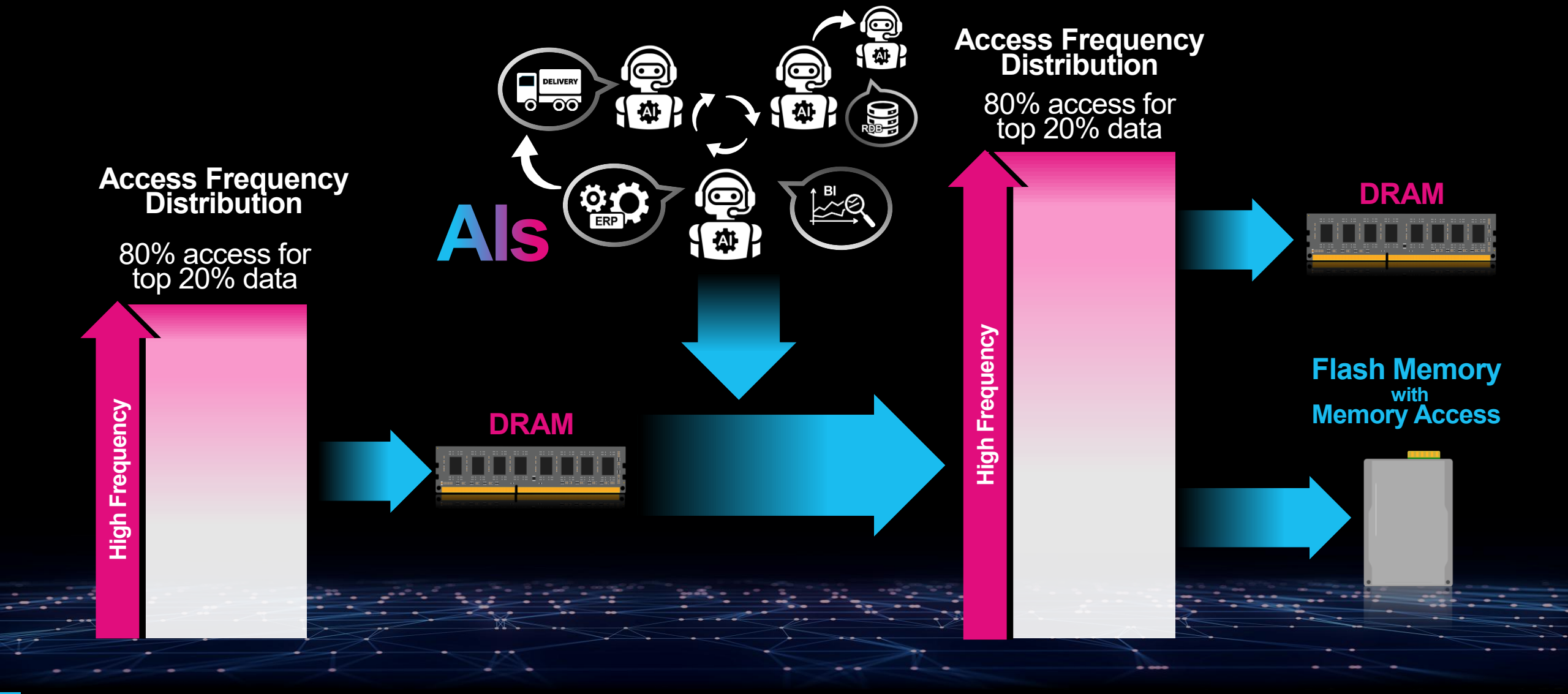


Sources: For 2024-2028 – TechInsights Inc. "NAND Market Report Q2 2026"; "DRAM Market Report Q2 2026"
For 2029-2030 – Illustrative

KIOXIA

© 2026 KIOXIA Corporation. All Rights Reserved.

How can flash memory contribute to memory expansion?



KIOXIA

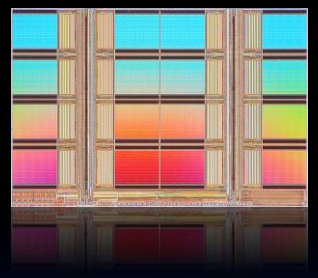
© 2026 KIOXIA Corporation. All Rights Reserved.

Solution for Memory Expansion with Flash Memory – XL-FLASH™

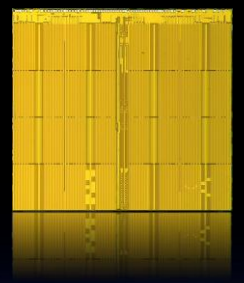
XL-FLASH™

- Low Read Latency: <5us
- High Endurance: 150-250K

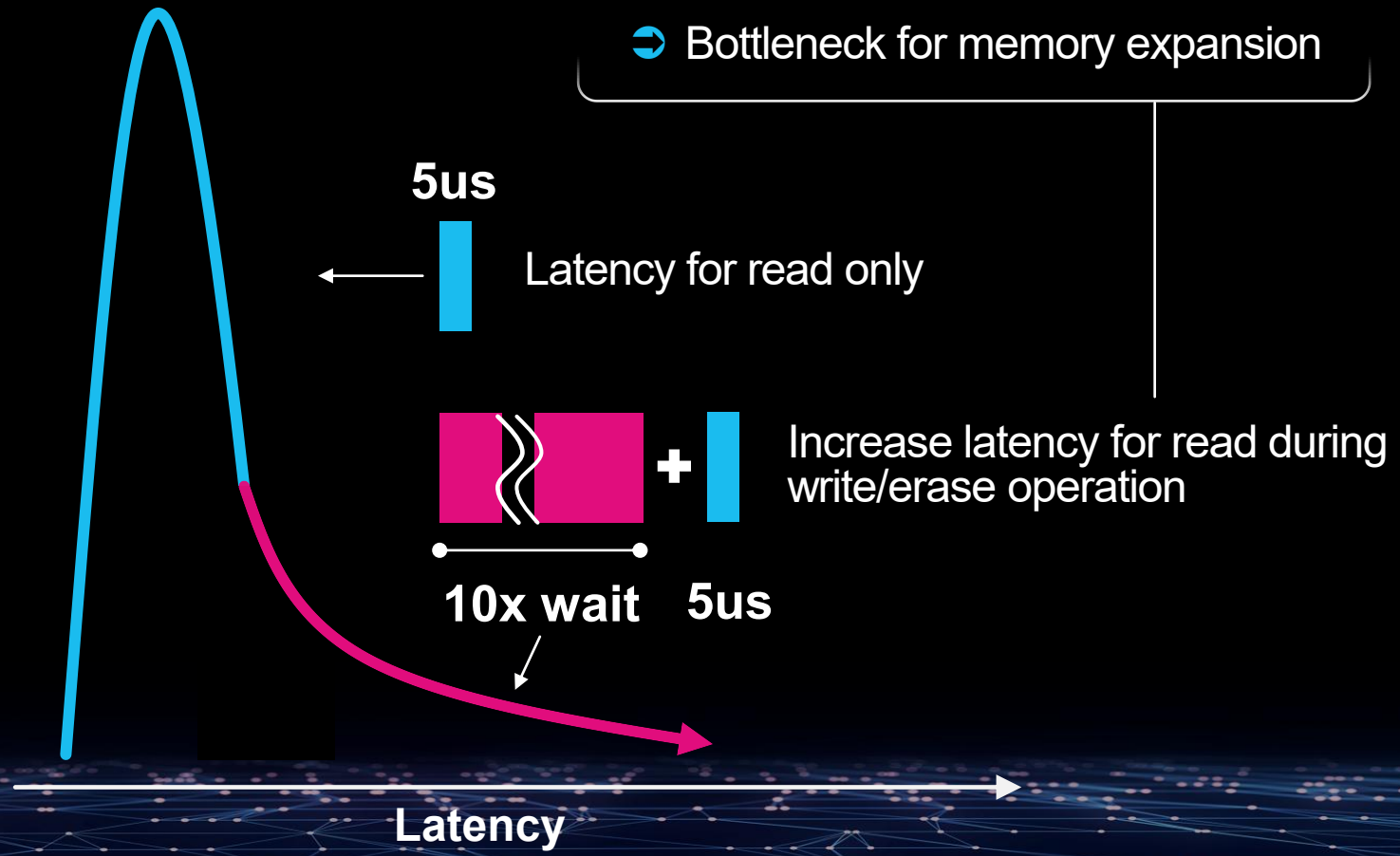
XL-FLASH™ gen. 2



XL-FLASH™ gen. 3



- Tail latency exceeds 100us
- Bottleneck for memory expansion

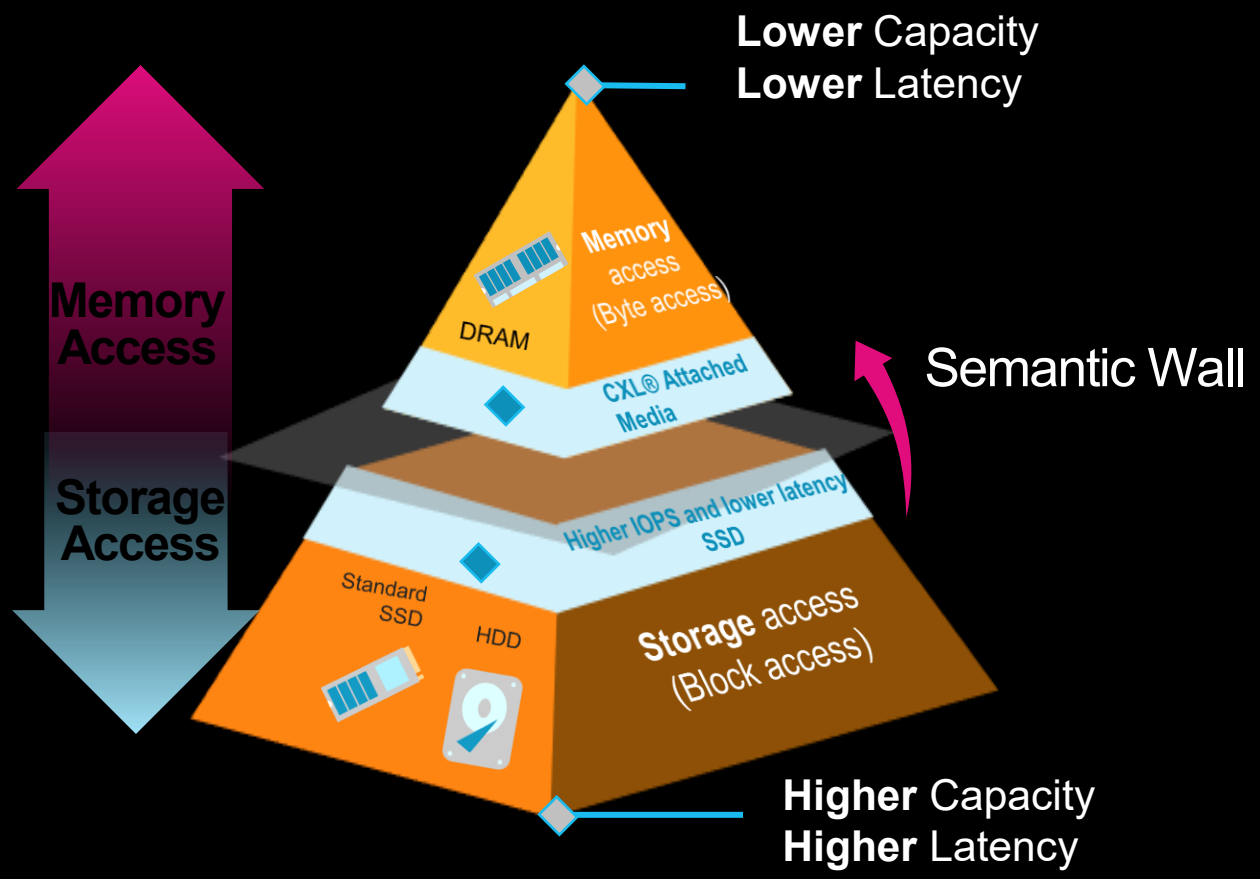


KIOXIA

gen. = generation

© 2026 KIOXIA Corporation. All Rights Reserved.

Semantic Wall with CXL® Attached XL-FLASH ™



	CXL® attached XL-FLASH™
Media	XL-FLASH™
Tiering within Device	DRAM cache tier, Low latency <5us (single-level cell), <10us (multi-level cell)
CXL Access	CXL.mem, CXL.io (config)
Applications	Agentic AI, General Purpose Computing, Memory Expansion, In-memory Databases, Cache Tiering, KV Store

KIOXIA

© 2026 KIOXIA Corporation. All Rights Reserved.

Solution for Memory Expansion with Flash Memory – CXL® Memory Module

XL-FLASH™

- Low Read Latency: <5us
- High Endurance: 150-250K

Dedicated Controller

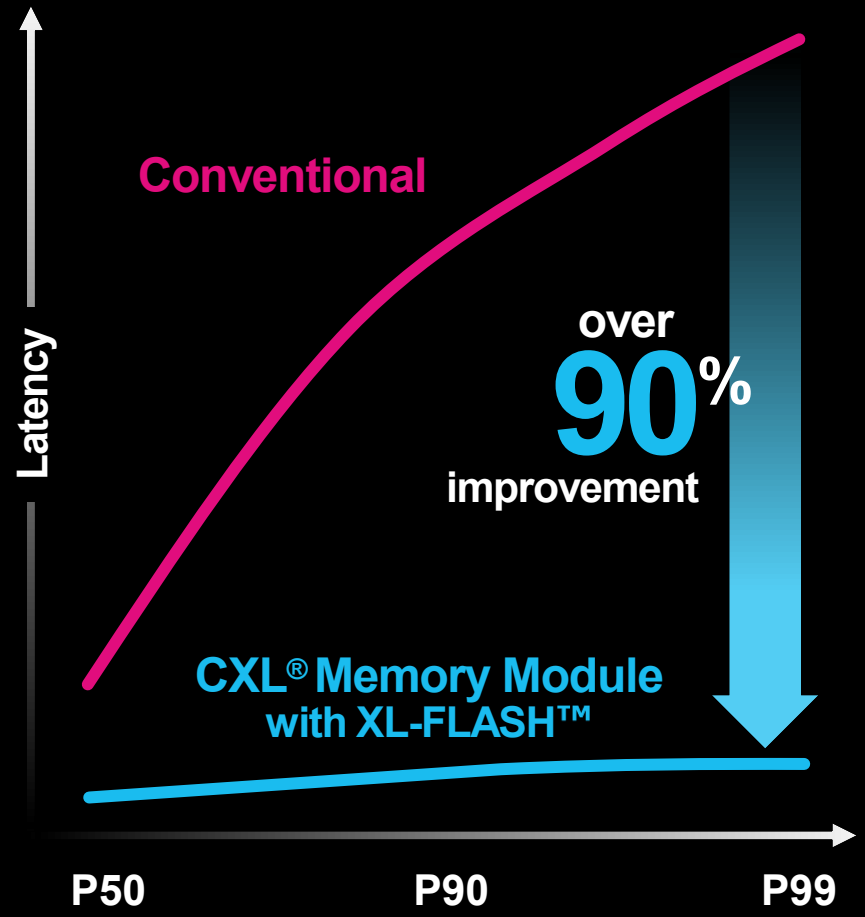
- CXL® Interface (= Memory Semantic)
- Latency Reduction Feature

=

CXL® Memory Module



- Full Memory Semantics
- Read Latency: <10us



Based on KIOXIA simulation results conducted in June 2026. Results may vary.

© 2026 KIOXIA Corporation. All Rights Reserved.

Applications to Leverage CXL® Attached Flash Memory

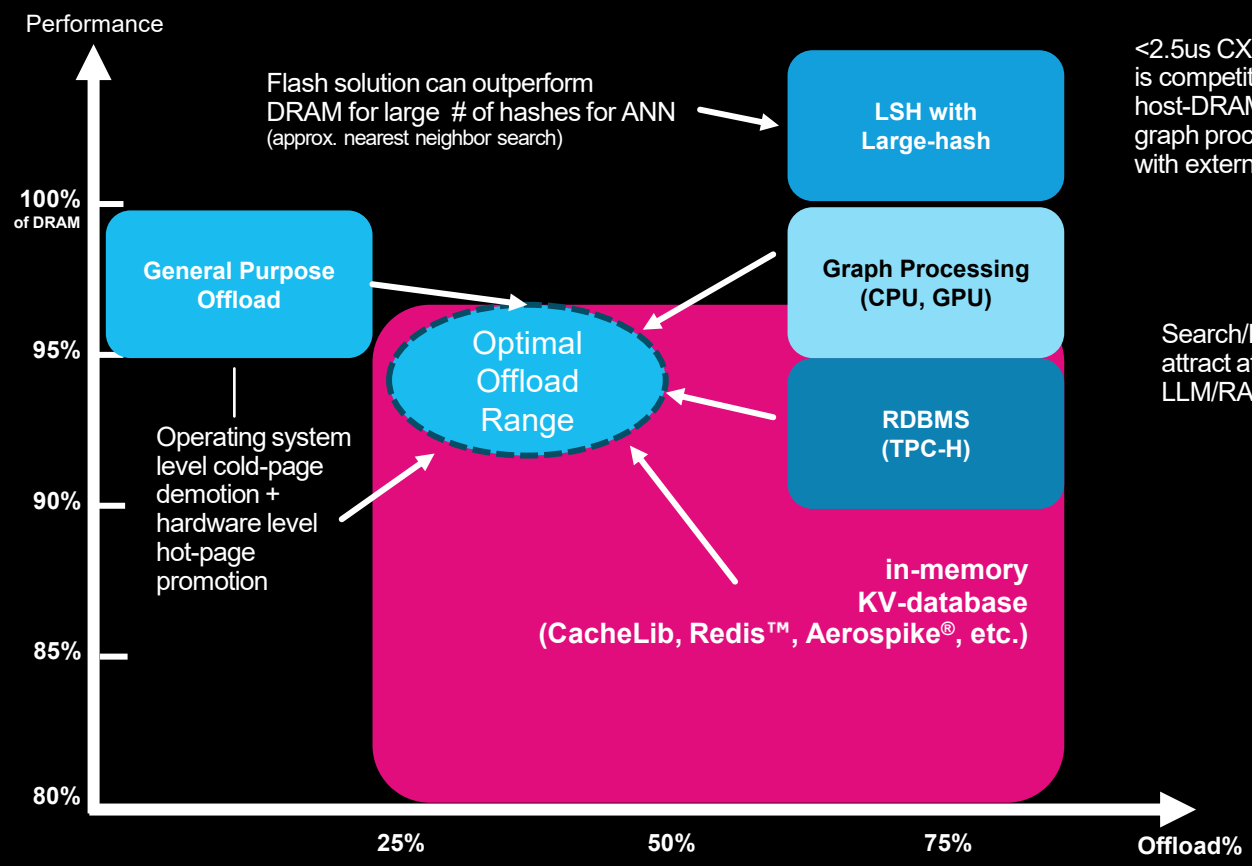
All Applications Are Not The Same

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

Leverage Industry Efforts

- Transparent Page Placement - Automatically manages large memory pages
- Transparent memory tiering solutions optimizes data placement across different memory types
- CXL® Hotness Monitoring Unit (CHMU)
- Application specific libraries can further increase the efficiency and reduce cost

Applications



<2.5us CXL®+XL-FLASH is competitive with host-DRAM in GPU graph processing with external memory

Search/Data-mining attract attentions for LLM/RAG purpose

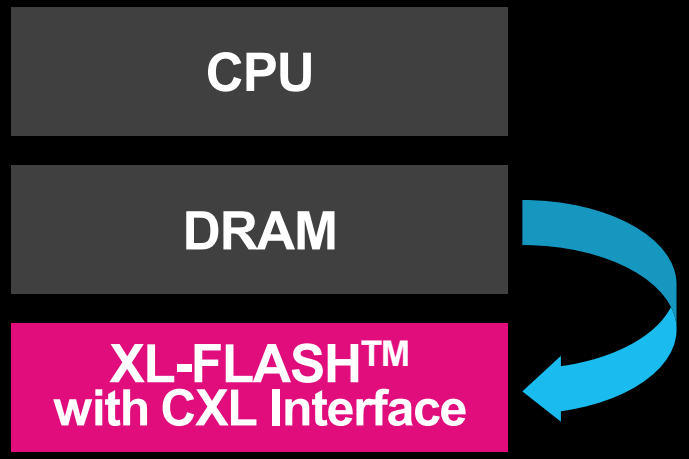
Source: KIOXIA

© 2026 KIOXIA Corporation. All Rights Reserved

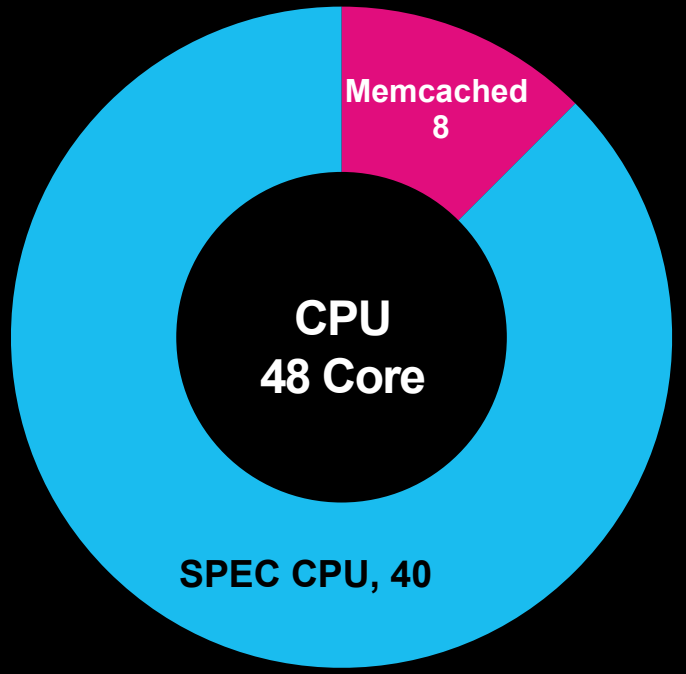
Linux® Kernel Patches for CXL® Hotness Monitoring Unit (CHMU)

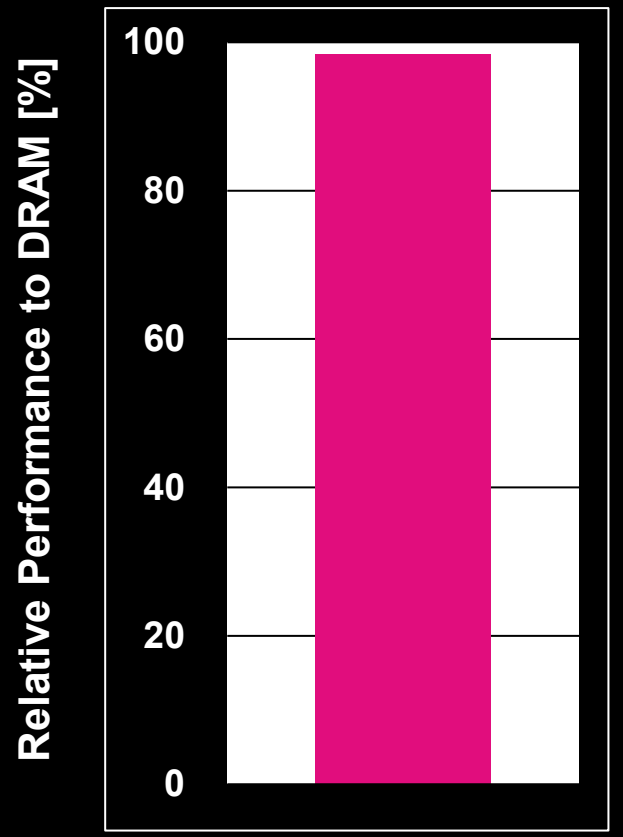
Mix Application Workload :
- **Memory intensive application : Memcached**
- **Compute intensive application : SPEC CPU®**

approx. **98% Performance with 36% Offload**
(DAMON overhead included)



36% Offload





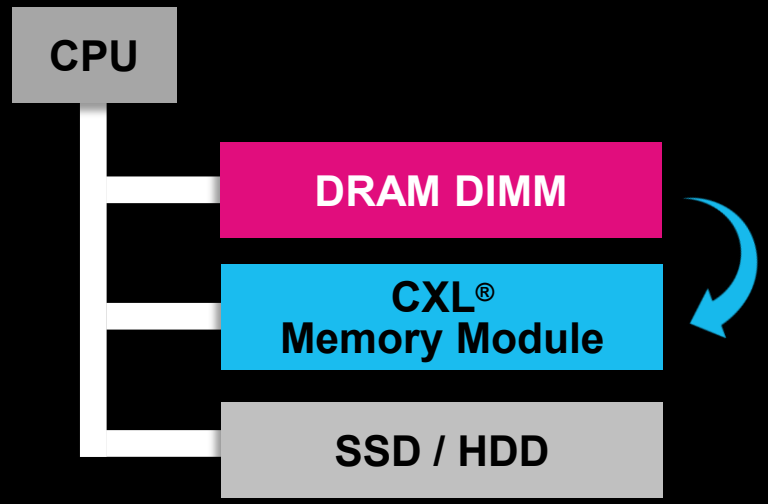
KIOXIA Contribution to Linux DAMON
- https://github.com/k-hiwada/CXL_Migration
- https://github.com/mita/linux/tree/damon-perf-for-v7.1-rc1
- https://github.com/mita/damo/tree/damo-perf-for-v3.2.2

Images/graphics on this slide are the property of KIOXIA America, Inc. (KAI) and are reproduced with the permission of KAI
SPEC CPU is a registered trademark of the Standard Performance Evaluation Corporation (SPEC).

KIOXIA

© 2026 KIOXIA Corporation. All Rights Reserved.

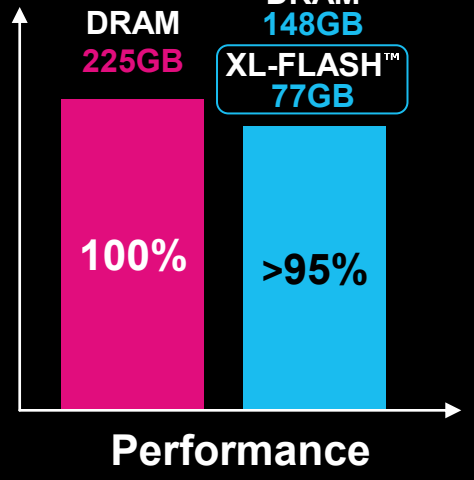
CXL® Memory Module with XL-FLASH™ Effectiveness



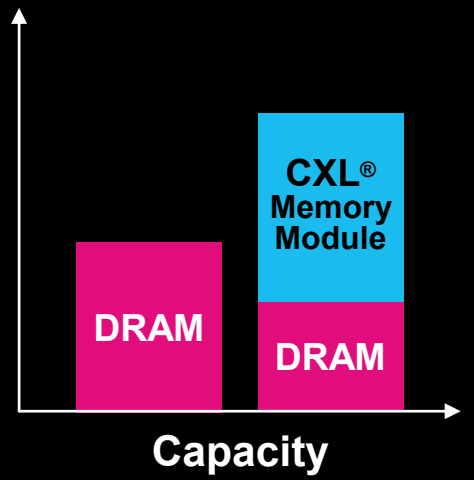
over
33%
data offload

- High performance
- Low latency < 10us
- Larger capacity

Same Capacity
>0.95x Performance



2.0x Capacity
1.3x Performance





The CXL® Memory Module with XL-FLASH™ offers more memory capacity along with enhanced cost performance.

Announced at FMS, XL1

For inquiries on XL Flash Contact:mahinder.saluja@us.kioxia.com

KIOXIA

© 2026 KIOXIA Corporation. All Rights Reserved.

Disclaimer

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. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1 GB = 2^30 = 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.

Images are for illustration purposes only.

CXL and Compute Express Link are trademarks of the Compute Express Link Consortium, Inc
Aerospike is a registered trademark of Aerospike, Inc.
Linux is a registered trademark of Linus Torvalds in the U.S. and other countries
Redis is a registered trademark of Redis Ltd. Any rights therein are reserved to Redis Ltd.

Other company names, product names, and service names may be trademarks of third-party companies.

© 2026 KIOXIA Corporation. All Rights Reserved. Information, including product pricing and specifications, content of services, and contact information is current and believed to be accurate on the date of the announcement, but is subject to change without prior notice. Technical and application information contained here is subject to the most recent applicable KIOXIA product specifications.

© 2026 KIOXIA Corporation. All Rights Reserved.

KIOXIA