High IOPS SSDs for AI Use Cases

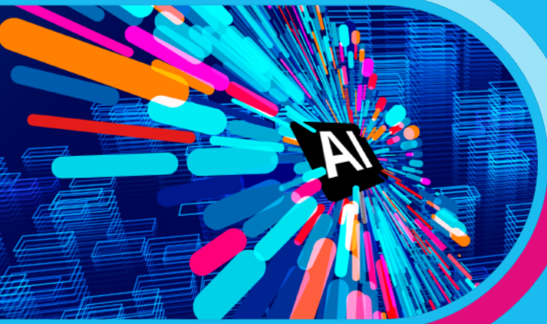
Rory Bolt
KIOXIA America, Inc.

SSDT-201-1
August 6th, 2025

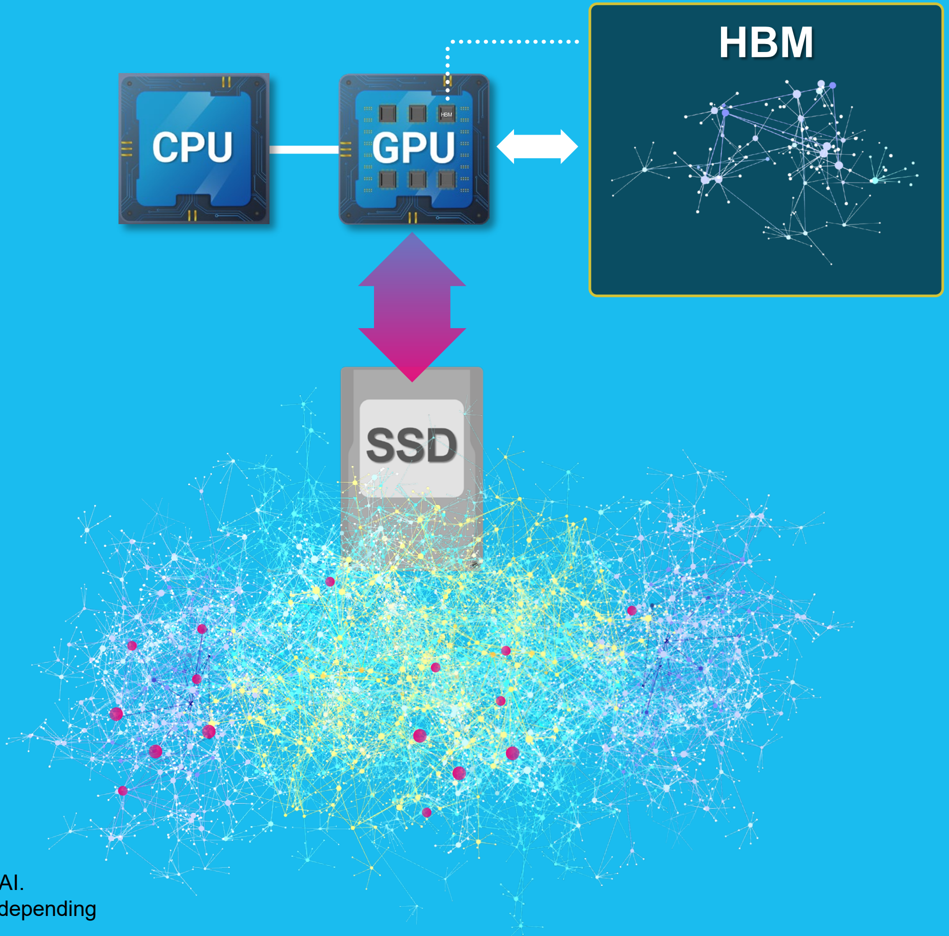
the Future of Memory and Storage

© 2025 KIOXIA America, Inc. All Rights Reserved.

Emerging AI Use-Case: GPU Memory Extension



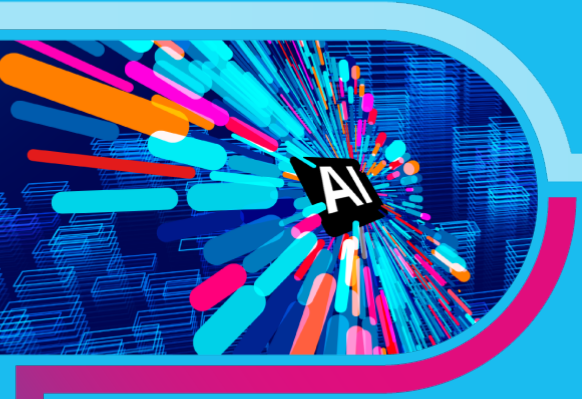
- **Addresses HBM expansion limitations and high costs**
- **Allows 10x – 100x larger datasets[1]**
- **GPU-initiated I/O**
  - up to 200M IOPS/GPU



All images and/or graphics within this slide are the property of KIOXIA America, Inc. (KAI) and are reproduced with the permission of KAI. Product images shown are a representation of a design model and not an accurate product depiction. 1. Comparing the size of HBM (depending on GPU) to the capacity of multiple SSDs, with the goal to handle datasets that are 10x- 100x larger than the largest possible today.

© 2025 KIOXIA Corporation. All Rights Reserved.

Emerging AI Use-Case: Near-GPU Caching

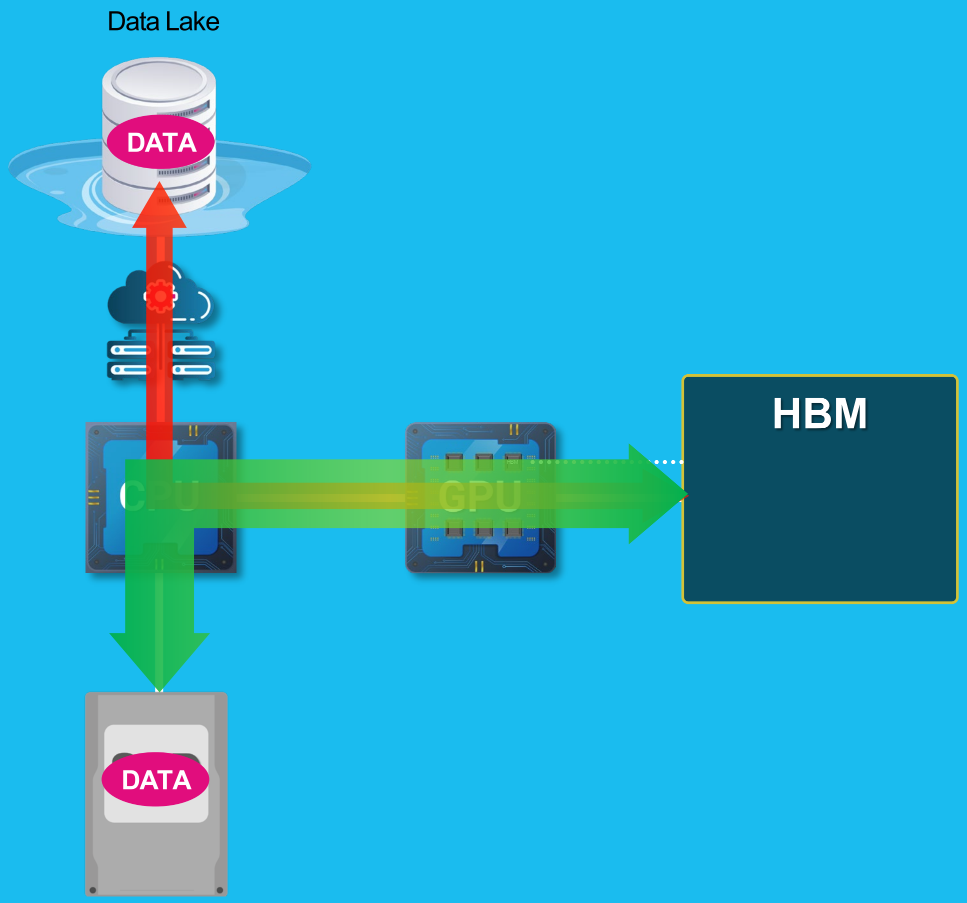


- Addresses inefficiency of small data accesses over very high-speed networks

- Large, efficient transfers from data lake to load cache

- Small reads serviced from local SSD

- CPU-initiated I/O

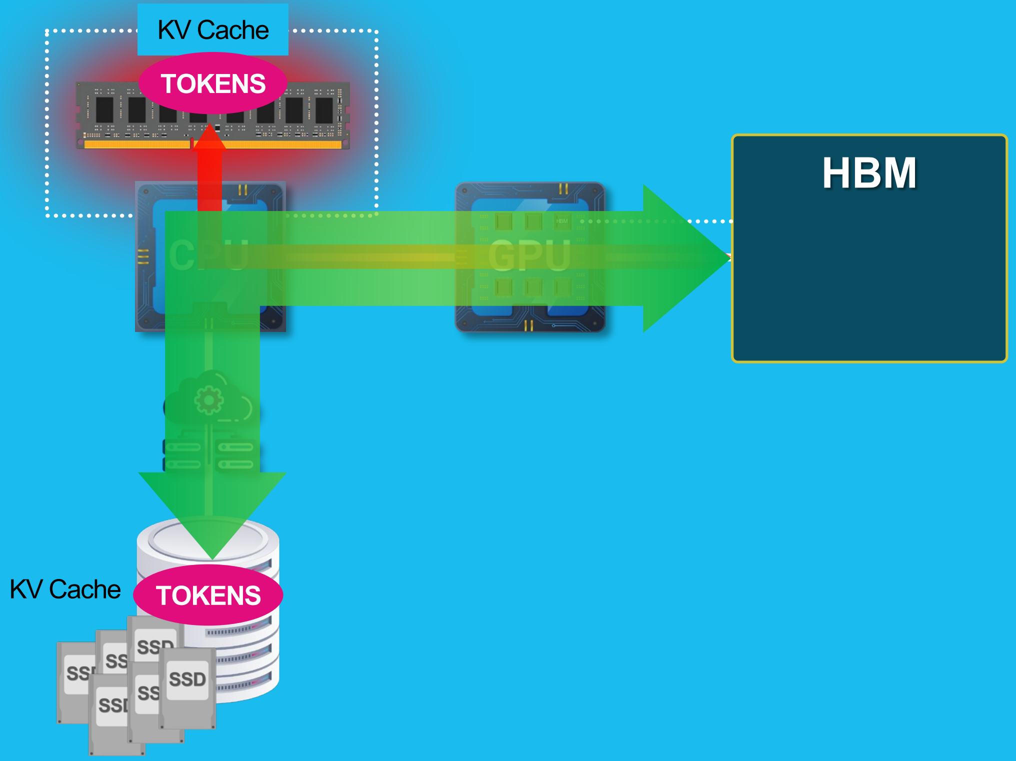


All images and/or graphics within this slide are the property of KIOXIA America, Inc. (KAI) and are reproduced with the permission of KAI. Product images shown are a representation of a design model and not an accurate product depiction.

© 2025 KIOXIA Corporation. All Rights Reserved.

Emerging AI Use-Case: Key Value Caching

- **Prevents recomputation of previously generated tokens**
- **Extends local memory-based caches**
  - Error recovery & routing benefits
- **CPU initiated I/O**



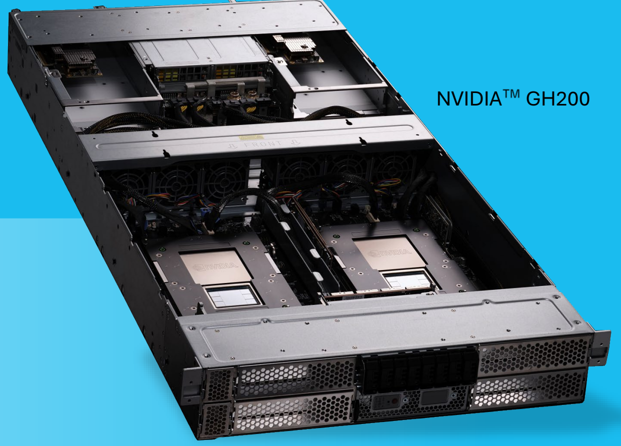
All images and/or graphics within this slide are the property of KIOXIA America, Inc. (KAI) and are reproduced with the permission of KAI. Product images shown are a representation of a design model and not an accurate product depiction.

© 2025 KIOXIA Corporation. All Rights Reserved.

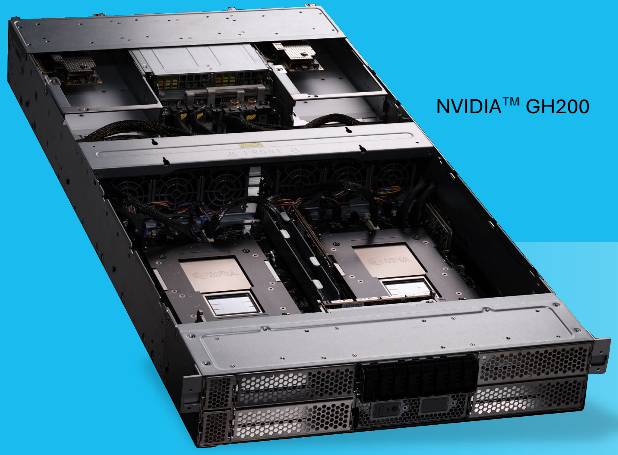
The Case for High IOPS

Alternate Paths to 200M IOPS

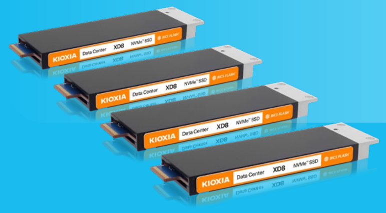
- Requires PCIe® switches
- Consumes lots of physical space
- Wasted capacity
- 32 TLC SSDs (800W for each GPU)



NVIDIA™ GH200



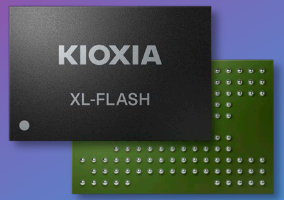
NVIDIA™ GH200



- 2~4 SSDs (100M or 50M IOPS) per GPU
- No PCIe switch needed
- ~120W

© 2025 KIOXIA Corporation. All Rights Reserved.

KIOXIA's Path to 100 Million I/Os Per Second



Enabled by XL-FLASH

2025

3M

4K Random Read IOPS
TLC Flash
PCIe® 5.0
14GB[2]/sec

2026

10M

512B[1] Random Read IOPS
XL-FLASH Gen. 2
PCIe® 6.0
23GB/sec

2027

100M

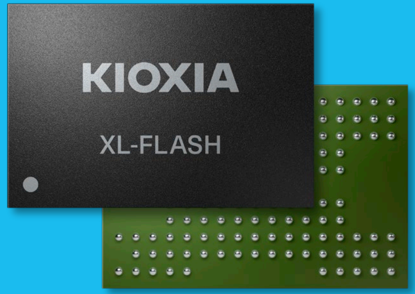
512B Random Read IOPS
XL-FLASH Gen. 3
PCIe® 7.0
50GB/sec

1. KIOXIA is collaborating with systems designers on liquid cooling solutions, in an effort to reach this targeting. 2. 512 byte block size. 2. Gigabytes (GB). All images and/or graphics within this slide are the property of KIOXIA America, Inc. (KIOXIA) and are reproduced with the permission of KIOXIA. Product images shown are a representation of a design model and not an accurate product depiction. PCIe is a registered trademark of PCI-SIG. NVMe is a registered or unregistered trademark of NVM Express, Inc. Other company and product names may be trademarks of the respective companies with which they are associated.

© 2025 KIOXIA Corporation. All Rights Reserved.

Low Latency Media is Key: Do The Math

- 100 Million IOPS requires a read to complete every 10 nsec[1]
- Typical TLC tRead ~ 60 usec[2]
- 60 usec / 10 nsec = 6,000 pipelined reads
- XL-FLASH tRead ~ 5 usec
- 5 usec / 10 nsec = 500 pipelined reads



Enabled by XL-FLASH

[1] Nanosecond (nsec), [2] Microsecond (usec). Product images shown are a representation of a design model and not an accurate product depiction. All images and/or graphics within this slide are the property of Kioxia Corporation (KIOXIA) and are reproduced with the permission of KIOXIA .

© 2025 KIOXIA Corporation. All Rights Reserved.

GPUs Use SSDs Differently

- Massive parallelism is the key to GPU performance
- Typical x86 system can issue ~50M IOPS consuming 100% CPU
- An NVIDIA Hopper™ GPU can generate ~200M IOPS with a projected <10% utilization
- It is not unusual for GPUs to drive device queue depths into the 10s of thousands!

NVIDIA Hopper is a trademark of NVIDIA Corporation.

© 2025 KIOXIA Corporation. All Rights Reserved.

Liquid Cooling Is In Your Future

- Faster flash media can be more power efficient!
- IOPS/Watt TLC: 480K vs XL-FLASH: 1.6M
- XL-FLASH @ 50M IOPS: ~ 35 Watts
- XL-FLASH @ 100M IOPS: ~ 60 Watts
- E3 may be required for surface area!

© 2025 KIOXIA Corporation. All Rights Reserved.

Performance / Power Preliminary Comparison with TLC SSDs

1st Gen 10M IOPS SSD

	Best in Class TLC	Best in Class TLC	High IOPS Gen1 XL-FLASH
PCIe® Gen.	**Gen5 x4**	**Gen6 x4**	**Gen6 x4**
512B Random Read [MIOPS]	N/A	N/A	10.0
4KB Random Read [MIOPS]	3.0	6.0	4.2
Power [W]	25	25	25
IOPS/Power Ratio	**0.5**	**1.0**	**1.7**

2nd Gen 50M/100M IOPS SSD

	Best in Class TLC	High IOPS Gen2 XL-FLASH	High IOPS Gen2 XL-FLASH
PCIe Gen.	**Gen7 x4**	**Gen7 x4**	**Gen7 x4**
512B Random Read [MIOPS]	N/A	50	100
4KB Random Read [MIOPS]	12	TBD	TBD
Power [W]	25	<=35	<=60
IOPS/Power Ratio	**1.0**	**>=3.0**	**>=3.5**

All images and/or graphics within this slide are the property of KIOXIA America, Inc. (KIOXIA) and are reproduced with the permission of KIOXIA. PCIe is a registered trademark of PCI-SIG.

© 2025 KIOXIA Corporation. All Rights Reserved.

KIOXIA