

The logo for SDC | StorageAI, featuring a stylized icon of three stacked horizontal bars to the left of the text "SDC | StorageAI™".

SDC | StorageAI™

A SNIA  Event

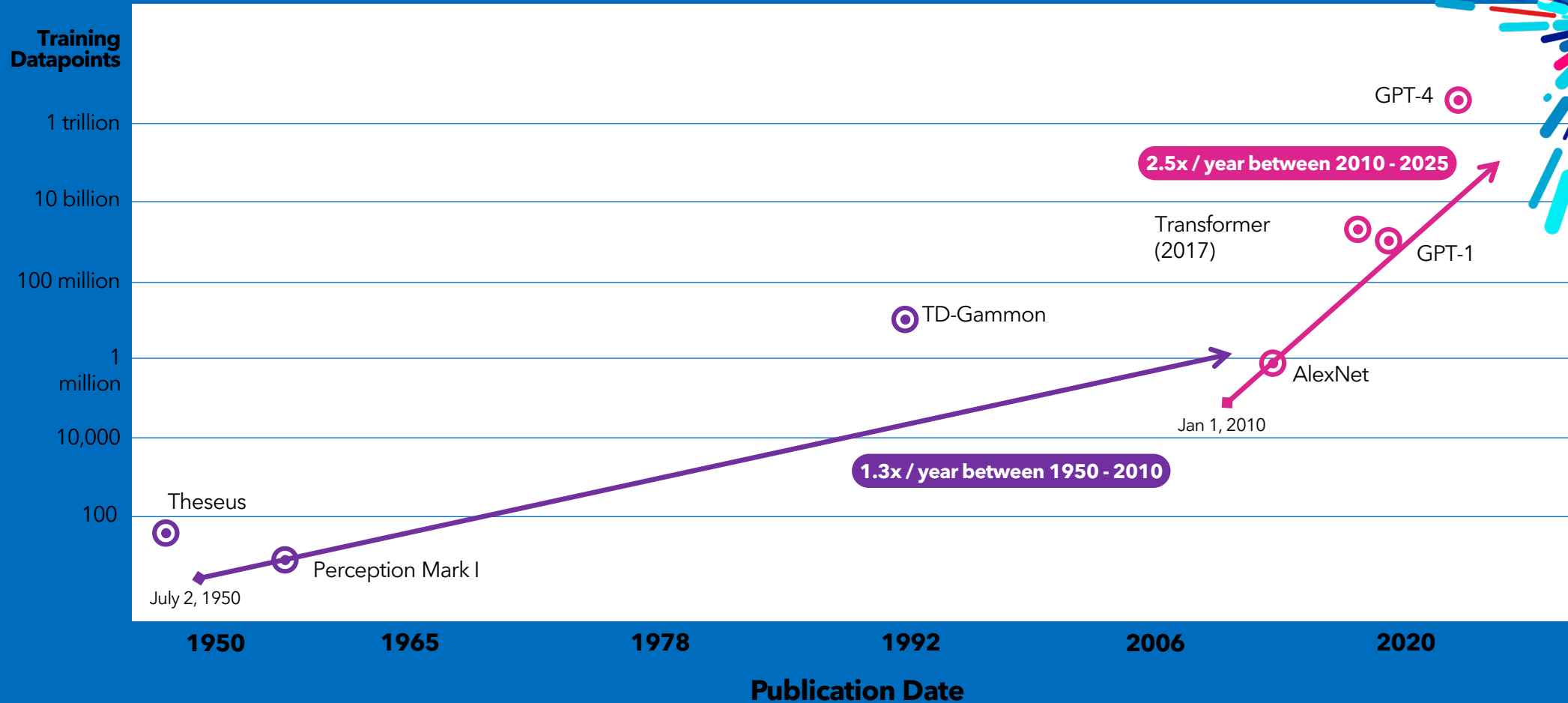
April 29, 2026 • Denver, Colorado

AI Impact On Storage

Rory Bolt, Senior Fellow
KIOXIA America, Inc.

AI Trend and Challenges

Exponential Growth of Datapoints Used to Train Notable AI Systems



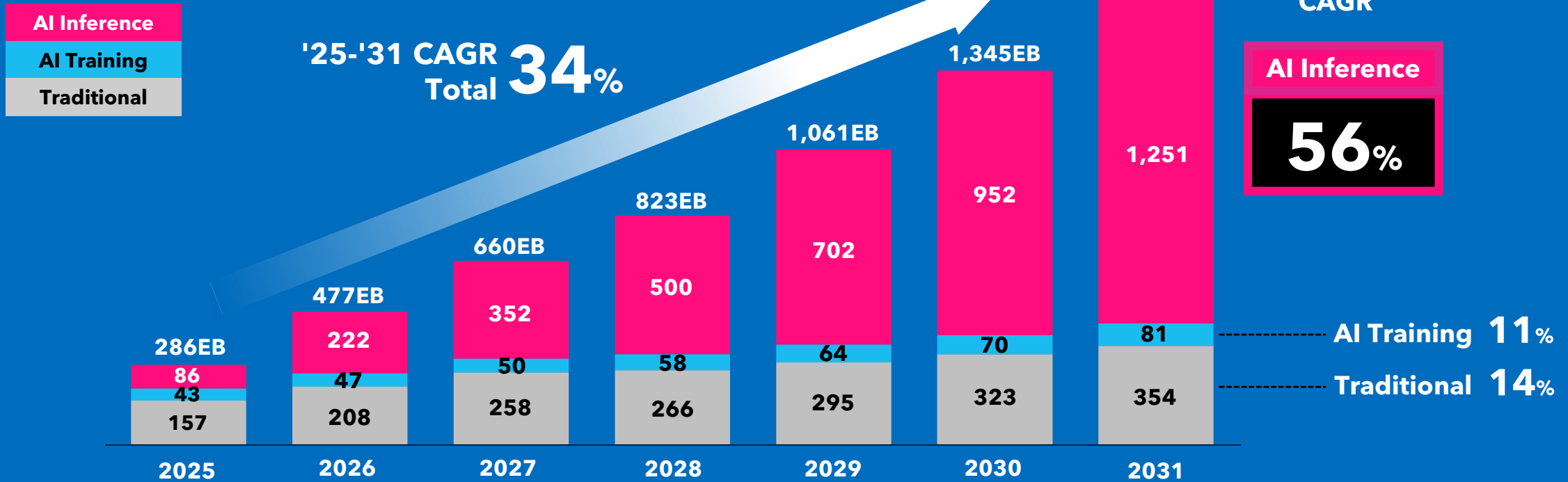
Source: OurWorldinData.org/artificial-intelligence

Data Source: Epoch (2025) - with major processing by Our World in Data

Data Center NAND Bit Demand by Detailed Workload

- Focus : Training → Inference (Token/sec and Token/\$)
 - ▶ Agentic AI, Edge AI, and Physical AI

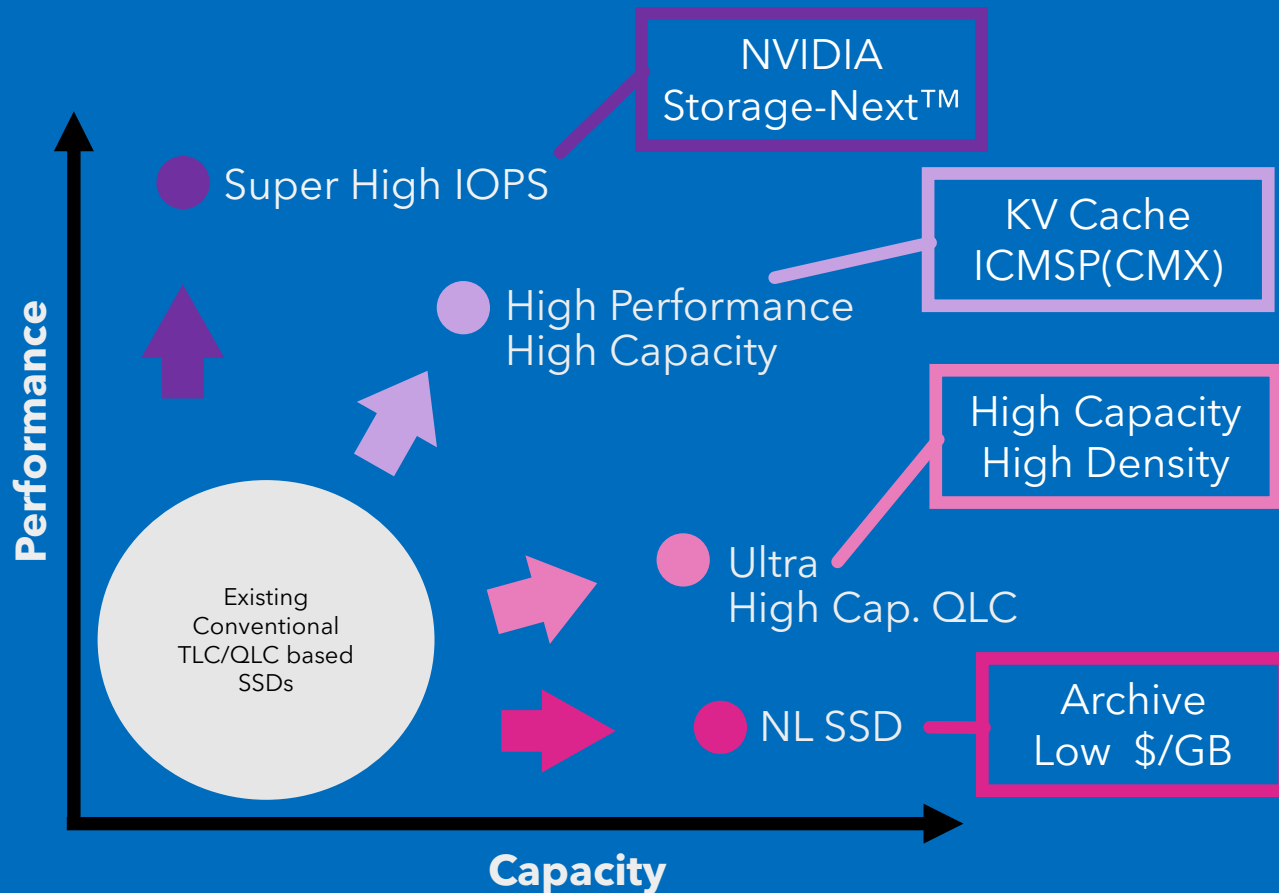
Bit Demand By Workload (EB)



NAND Market Report Q1 2026 TechInsights

4 Key Directions for SSD Products

Generative AI evolution has created and accelerated 4 direction for SSDs



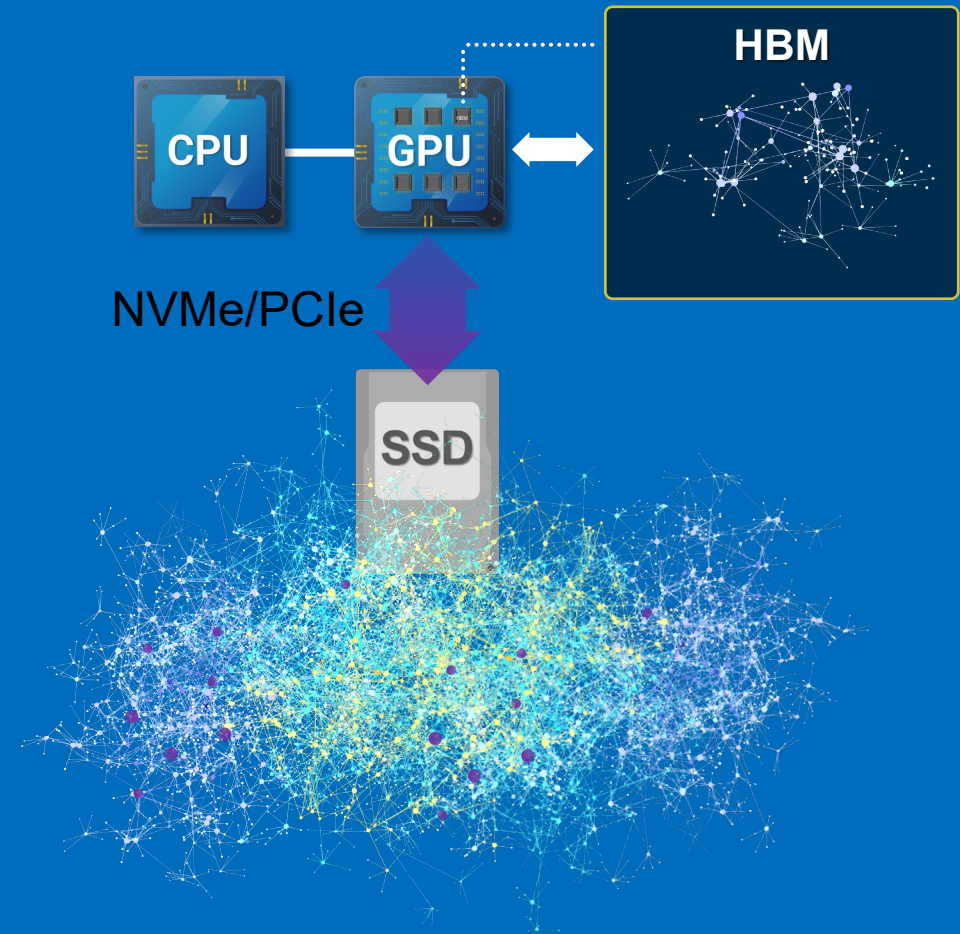
Segment	Usage	Key Technology
<ul style="list-style-type: none"> ● Super High IOPS KIOXIA GP Series 	NVIDIA Storage-Next™	PCIe® Gen6~ XL-FLASH™ 512 B access
<ul style="list-style-type: none"> ● High Performance High Capacity KIOXIA CM Series 	KV Cache-Reuse	PCIe® Gen5~ High OP TLC
<ul style="list-style-type: none"> ● High Capacity KIOXIA LC Series 	Ingestion RAG	QLC (122 TB/245 TB+)
<ul style="list-style-type: none"> ● HDD replacement Under Planning 	Archive	Low Cost QLC 256 TB+

KV : Key Value
RAG: Retrieval-Augmented Generation

NVIDIA Storage-Next™: GPU Memory Extension by NVMe SSDs



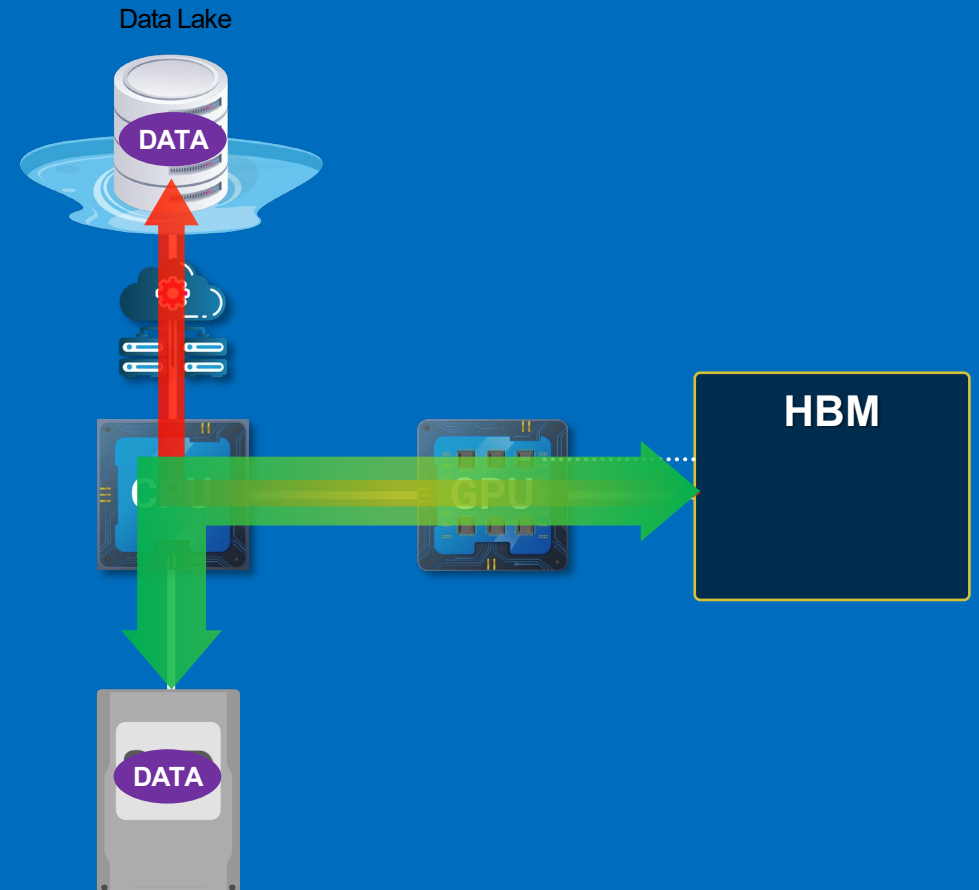
- **Addresses HBM expansion limitations & high costs**
- **Allows 10x - 100x larger datasets**
- **NVIDIA Storage-Next™**
 - GPU-initiated I/O (Software Stack: NVIDIA SCADA™)
 - NVMe™/PCIe®
 - Fine Grain I/O size (512 byte)
 - Super High IOPS (~200M IOPS/GPU)



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**



Another Topic: NVIDIA ICMS (Inference Context Memory Storage Platform)

NVIDIA

Newsroom | [NVIDIA in Brief](#) | [Exec Bios](#) | [NVIDIA Blog](#) | [Podcast](#) | [Media Assets](#) | [In the](#)

Press Release

NVIDIA BlueField-4 Powers New Class of AI-Native Storage Infrastructure for the Next Frontier of AI

January 5, 2026



News Summary:

- › NVIDIA BlueField-4 powers NVIDIA Inference Context Memory Storage Platform, a new kind of AI-native storage infrastructure designed for gigascale inference, to accelerate and scale agentic AI.
- › The new storage processor platform is built for long-context-processing agentic AI systems with lightning-fast long- and short-term memory.
- › Inference Context Memory Storage Platform extends AI agents' long-term memory and enables high-bandwidth sharing of context across clusters of rack-scale AI systems — boosting tokens per seconds and power efficiency by up to 5x.
- › Enabled by NVIDIA Spectrum-X Ethernet, extended context memory for multi-turn AI agents improves responsiveness, increases throughput per GPU and supports efficient scaling of agentic inference.

Key capabilities of the NVIDIA BlueField-4-powered platform include:

- › NVIDIA Rubin cluster-level KV cache capacity, delivering the scale and efficiency required for long-context, multi-turn agentic inference.
- › Up to 5x greater power efficiency than traditional storage.
- › Smart, accelerated sharing of KV cache across AI nodes, enabled by the NVIDIA DOCA™ framework and tightly integrated with the NVIDIA NIXL library and NVIDIA Dynamo software to maximize tokens per second, reduce time to first token and improve multi-turn responsiveness.
- › Hardware-accelerated KV cache placement managed by NVIDIA BlueField-4 eliminates metadata overhead, reduces data movement and ensures secure, isolated access from the GPU nodes.
- › Efficient data sharing and retrieval enabled by NVIDIA Spectrum-X™ Ethernet serves as the high-performance network fabric for RDMA-based access to AI-native KV cache.

Storage innovators including AIC, Cloudbian, DDN, Dell Technologies, HPE, Hitachi Vantara, IBM, Nutanix, Pure Storage, Supermicro, VAST Data and WEKA are among the first building next-generation AI storage platforms with BlueField-4, which will be available in the second half of 2026.

- ICMS: Storage Platform for **KV cache**
- PCIe® Gen5/Gen6 TLC NVMe™ SSDs will be utilized in the storage box
- Kioxia is working with NVIDIA® to clarify the requirements

High Level Requirements For SSDs

512B random read optimization

Increased endurance

High queue depths

Liquid cooling

Multi-initiator access

Larger capacities

512B Random Read Optimization

New ECC layouts

Not necessarily tied to IU size

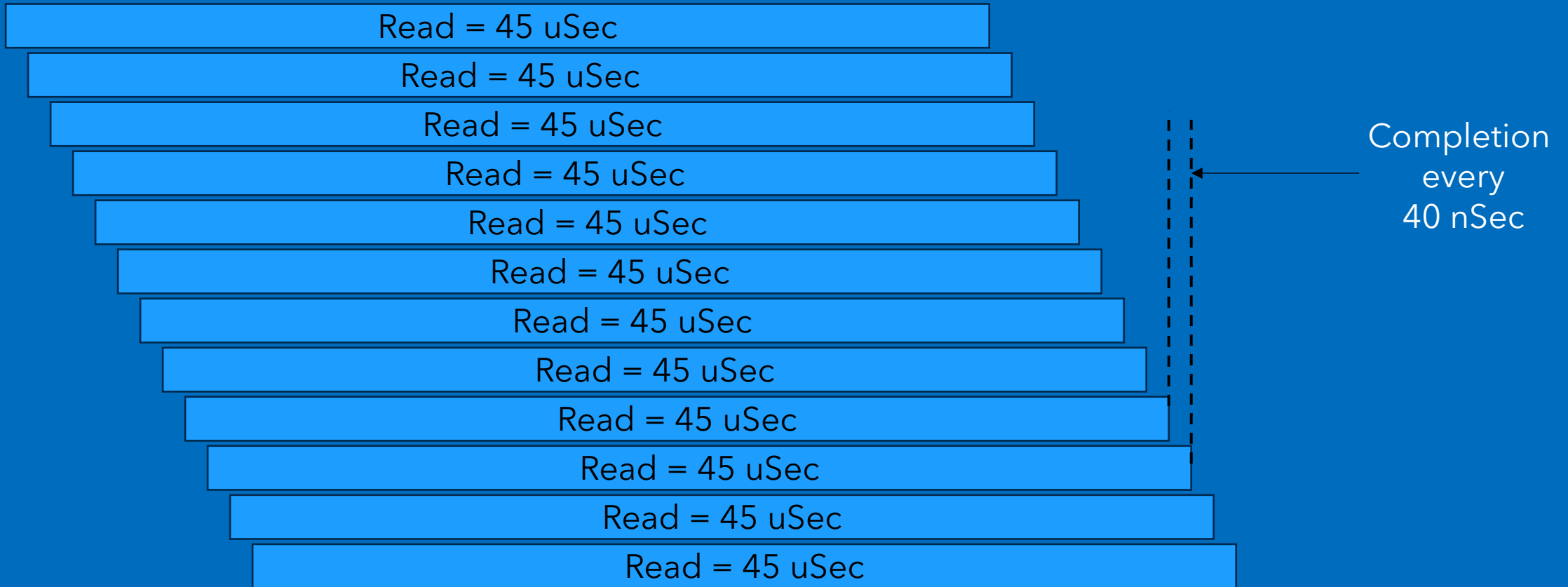
Concurrency requirements

Hypothetical case: 25M IOPS = 40nS per I/O

if tRead = 45uSec must overlap 1125 parallel IOs

if tRead = 25uSec must overlap 625 parallel IOs

512B Random Read Optimization



Increased Endurance

Workloads appear fundamentally different

Motivation for pSLC and pMLC

Overprovisioning as part of the solution

Caches vs long term storage

3 DWPD to 100 DWPD

High Queue Depths

Scheduling impact

Head of line blocking

Overall latency

Liquid Cooling

Power

Data center efficiency

Standardization of design?

Multi-initiator Access

Direct access to filesystem data

Leases on LBA ranges

Mapping leases to initiators

Fast path enforcement

Fragmentation?

Data protection?

Larger Capacities

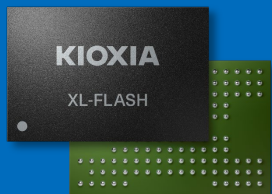
Data growth trends

pSLC and pMLC

SKU minimization and inventory management

KIOXIA develops SSDs that comply with NVIDIA Storage-Next™

Super High IOPS SSD - KIOXIA GP Series



Enabled by
XL-FLASH™

Today

Super High IOPS Emulator

Up to 100+ MIOPS
GPU Initiated I/O (w/ NVIDIA SCADA™)
KIOXIA GP Series Performance Emulation
(Latency histogram setting & IOPS clipping)

2026

10M < 25W

512B Random Read IOPS
XL-FLASH™ generation 2
PCIe® 6.0



**Evaluation samples
by the end of 2026**

2027

~100M

512B Random Read IOPS
XL-FLASH™ generation 3
PCIe® 7.0



Super High IOPS Emulator Testing (Up to 100+ MIOPS)
Software Stack Enablement, System Study with KIOXIA GP Series Performance

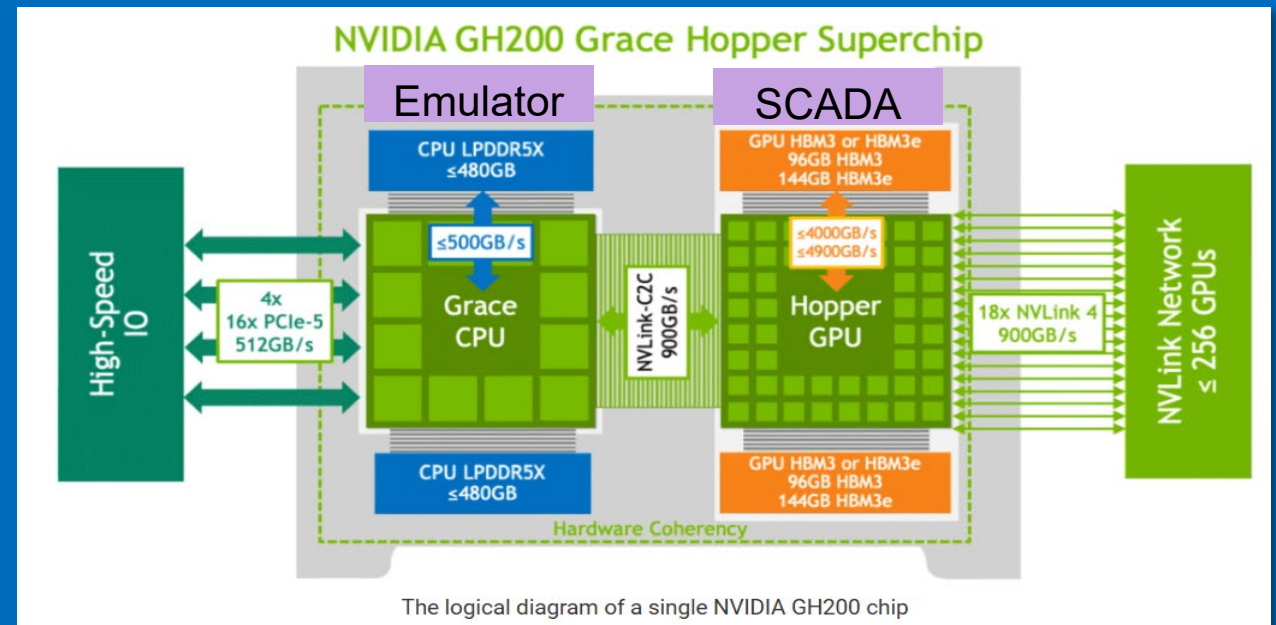
Subject to change without notice. Product image may be different than actual product.
PCIe is a registered trademark of PCI SIG

High IOPS Emulator

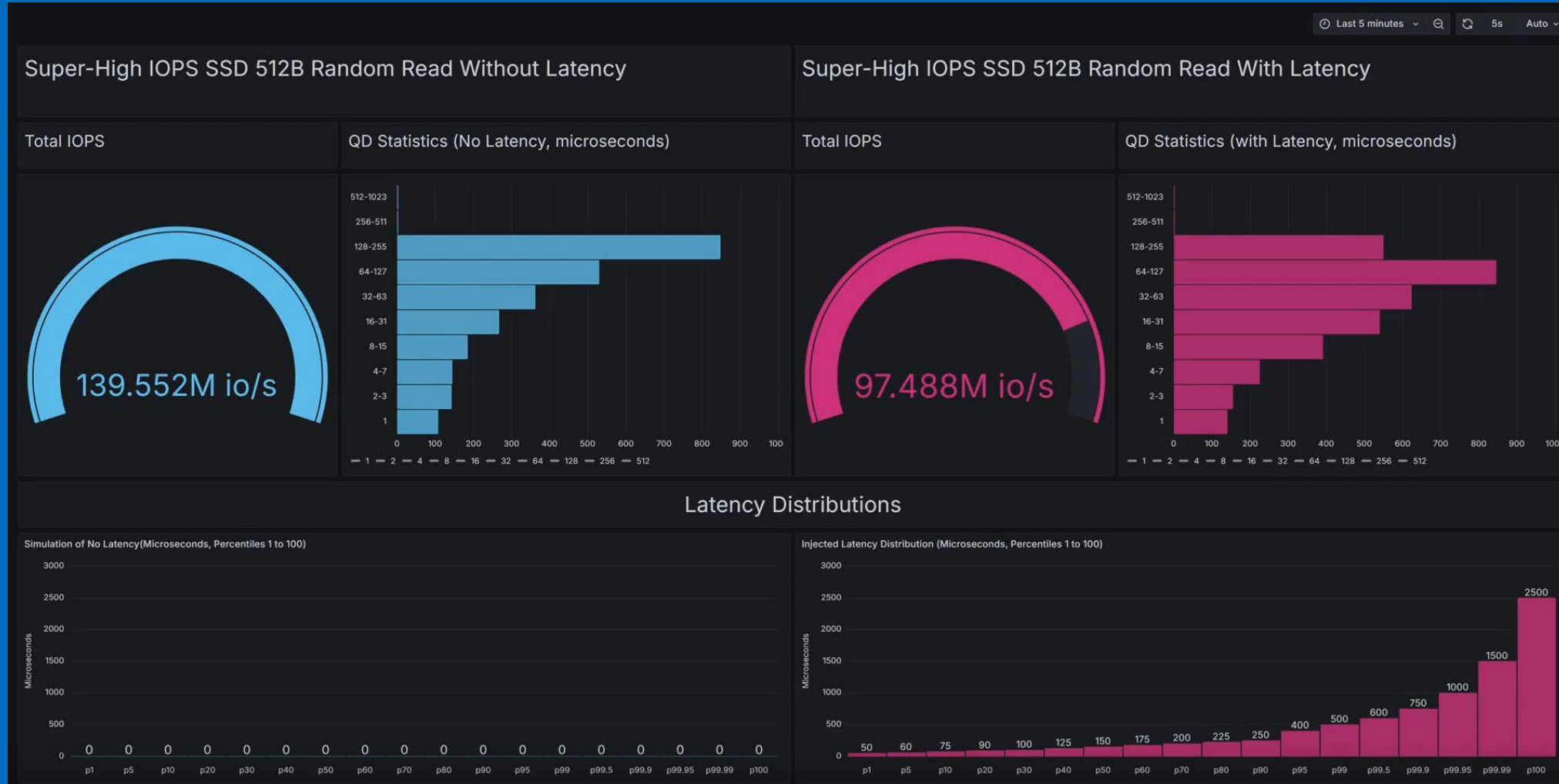
- Phase 1: Maximum performance done in Aug 2025
 - ▶ 140M IOPS in single instance generated on GH200.
- Phase 2: Model device performance done in Sept 2025
 - ▶ Add synthetic latency on IO execution
 - Dynamic latency adjustment
 - ▶ Telemetry
 - IO counters to track IO statistics
- Phase 3: Running on SCADA™
 - ▶ Application experiments
 - How apps behave under SCADA
 - ▶ Device experiments
 - Analyze SCADA IO characteristics with over 100M IOPS

Prerequisites:

- GH/GB system
- X86 based GPU system won't generate 100M IOPS due to narrow bandwidth between GPU and CPU



Emulator Demo (effect of higher latency profile)

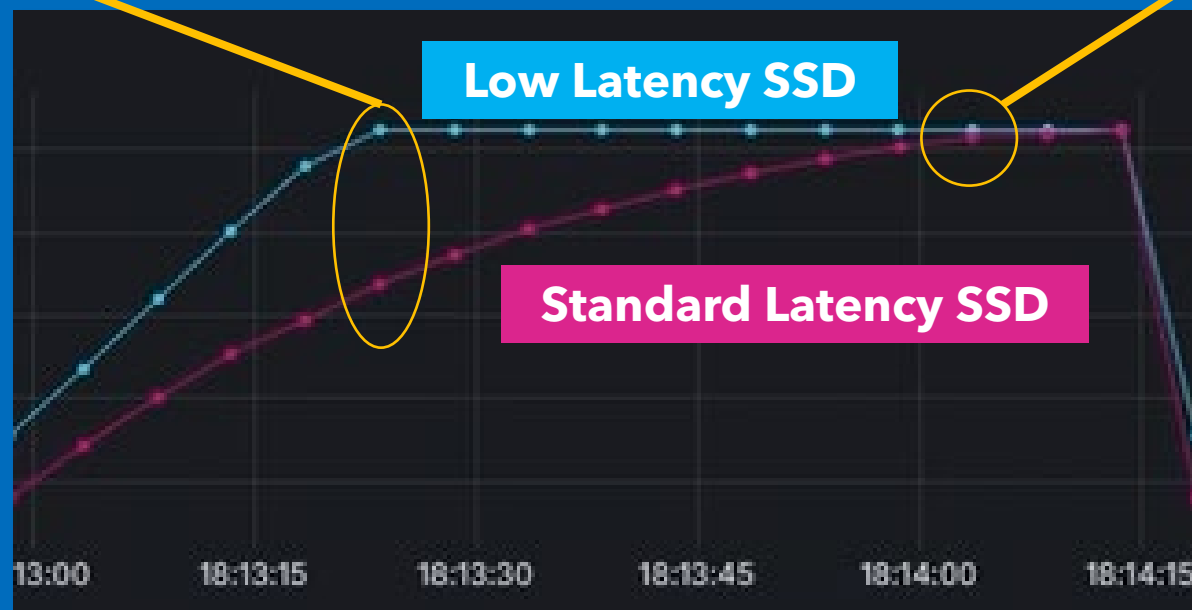


KIOIXA Super High IOPS Emulator - NVIDIA GTC Booth Demo



3072 Threads

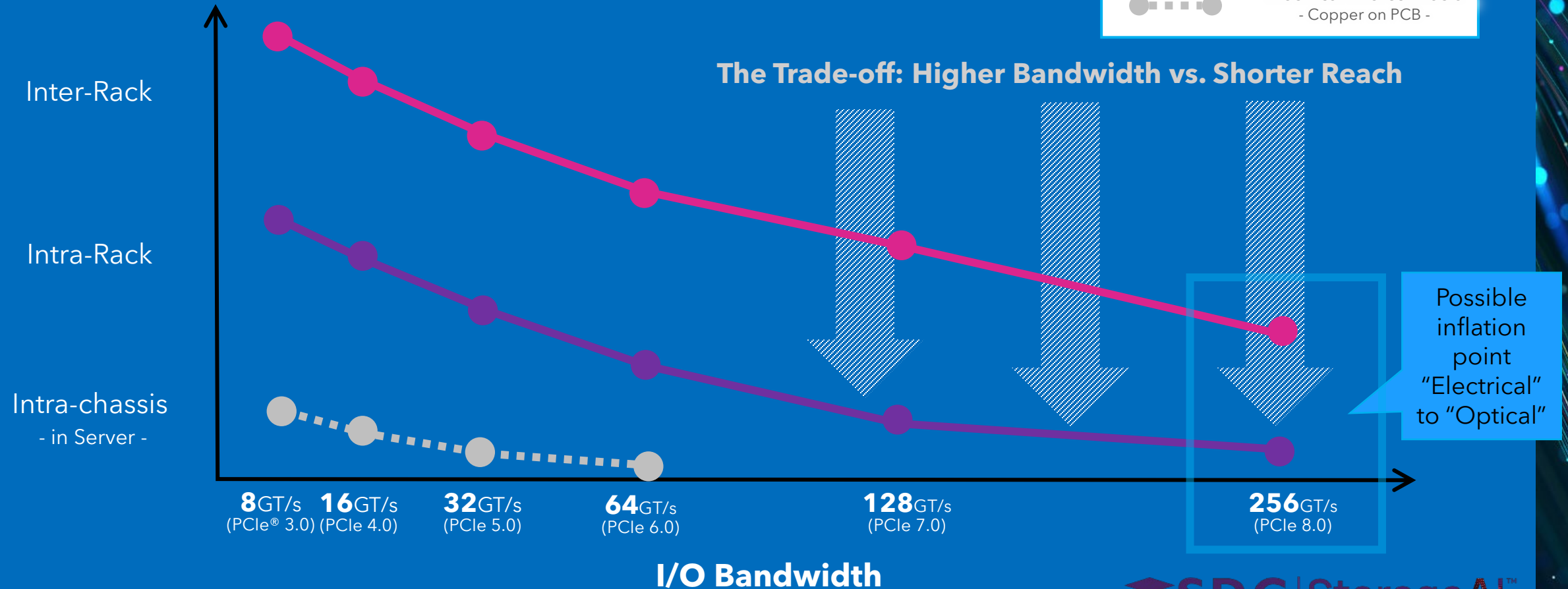
8192 Threads



Challenges of Increasing PCIe Bandwidth: High Bandwidth vs. Shorter Interconnect Lengths

Scaling Trends of Transmission Distance: Optical vs. Electrical Link*

Coverage Area



The logo for SDC | StorageAI. It features a stylized icon of three stacked, slightly offset rectangular blocks to the left of the text "SDC | StorageAI™". The text is in a clean, white, sans-serif font. The background of the entire slide is a dark blue gradient with abstract, glowing light trails and particles in shades of blue, cyan, and green, creating a sense of motion and data flow.

SDC | StorageAI™

A SNIA  Event

Thank You