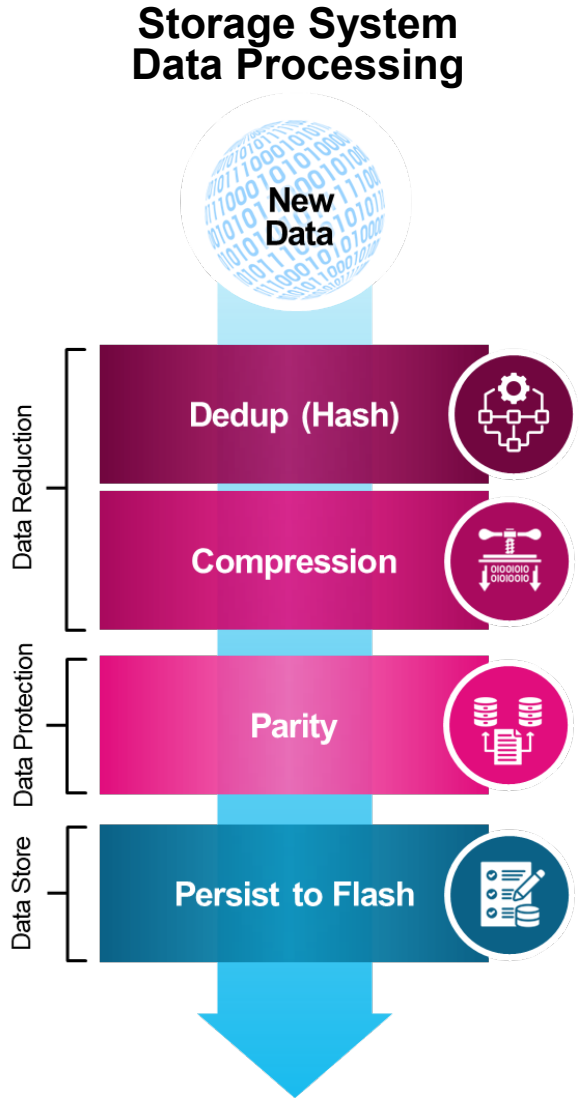
Offload Fixed Function Storage Services to Storage Subsystem

Mahinder Saluja,
Director of SSD Technology and Pathfinding
KIOXIA America, Inc.
August 2025

FMS
the Future of Memory and Storage

KIOXIA

Storage Systems Data Management Scalability Challenges

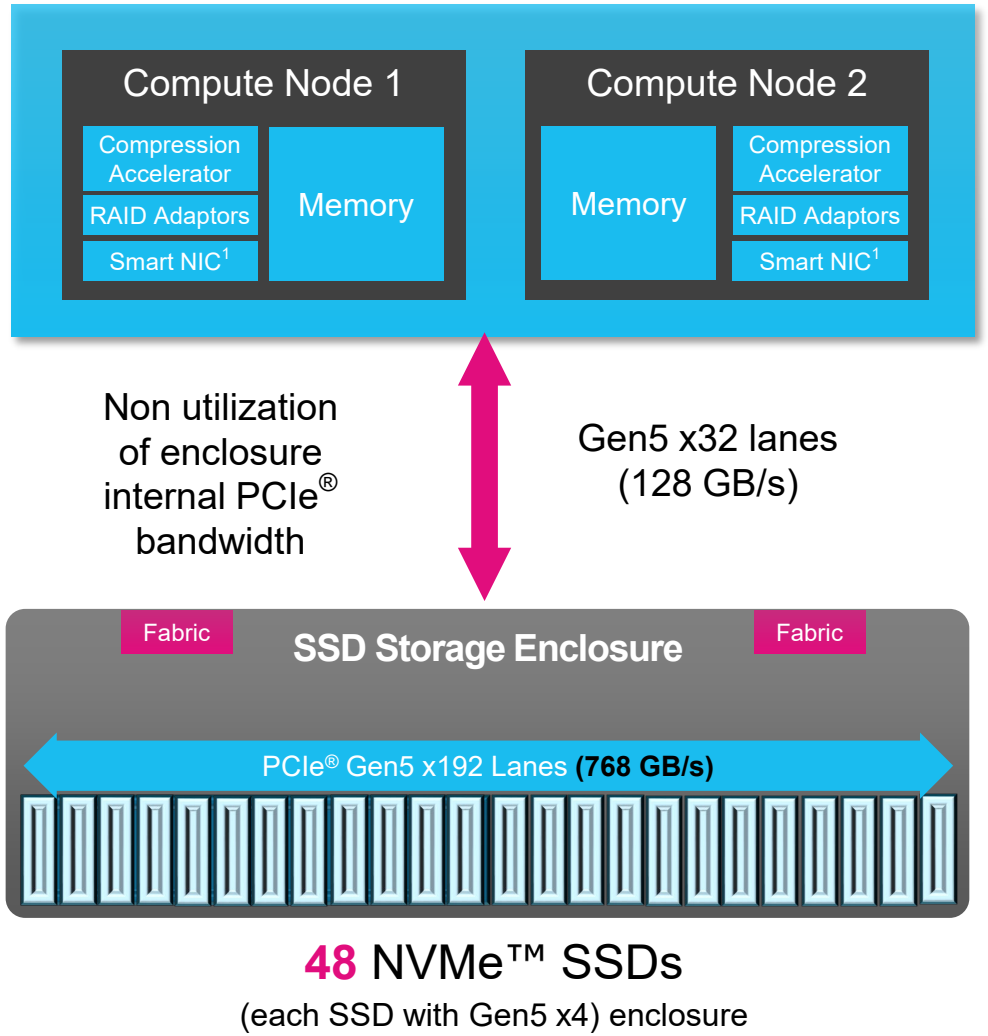


- **Data ingestion in a storage system involves a series of compute intensive operations**
- **With increased SSD performance, data management tasks require increased processing power usually solved by**
  - *General purpose CPU cores*
    - Current compute architecture and memory hierarchy increase total cost of acquisition and total cost of ownership
    - Overprovisioning cores and memory
  - *Accelerators*
    - Data processing units
    - Performance limited by number of PCIe® lanes assigned or system memory
    - Cannot scale with added SSDs
    - Consumes additional power from system slots

1. Deduplication (dedupe) hashes are unique identifiers generated from data blocks using hashing algorithms to identify duplicate data within a system. All images and/or graphics within this slide are the property of KIOXIA America, Inc. (KAI) and are reproduced with the permission of KAI. PCIe is a registered trademark of PCI-SIG

Storage Systems Infrastructure

Compute for Storage Functions



48 NVMe™ SSDs
(each SSD with Gen5 x4) enclosure

Compute Node Requirements

- Significant memory footprint for data management
- Accelerators to augment compute node

Storage Enclosure **48** NVMe SSDs

- **Available Resources**
  - Useable interface bandwidth (assuming 32 PCIe lanes): 32x4 = **128 GB/s**
  - PCIe bandwidth (Gen5) (48 Gen5 x4 Lanes) : **48x**4x4 = **768 GB/s**
    (SSDs*Lanes*per lane bandwidth)
  - **Spare PCIe bandwidth: 768-128 = 640 GB/s**

Host Orchestrated Compute Offload Building Blocks

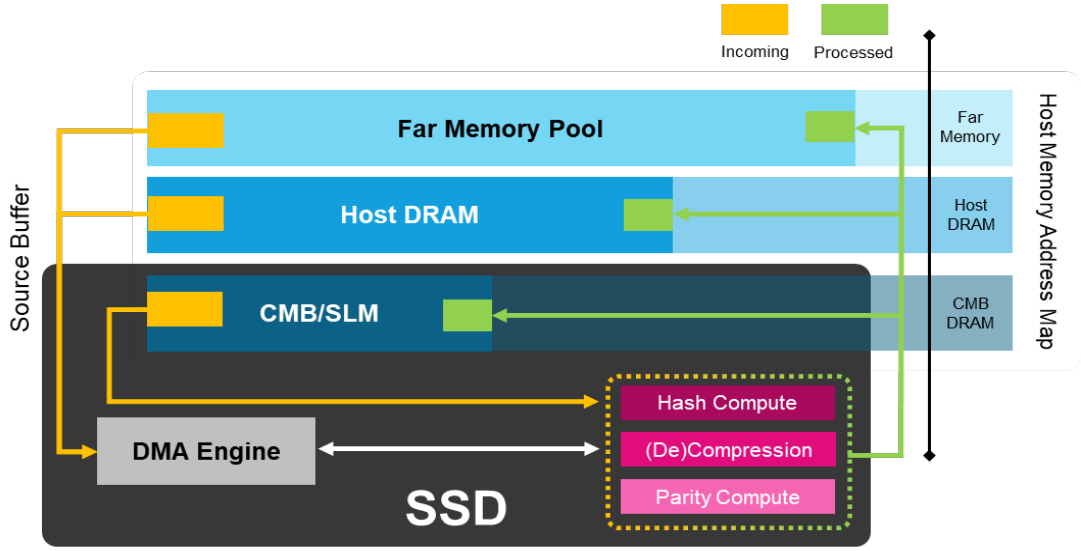
Power-efficient compute engines

DRAM bandwidth saving

Host orchestrated standard based

Applications of Offload Engines

- **Hash/CRC[3]:** Dedupe, Object/File signature/scrubbing, buffer integrity
- **(De)Compression:** Compression with levels, decompress and filter
- **Parity Compute**: Erasure code (EC), compare, **Data scrubbing, RAID Rebuild**



Host places data in appropriate buffer (host memory/ far memory/ CMB[1]) → Host issues command to the SSD → Compute engine on the SSD executes the command → Processed data is placed into the destination buffer → Command completion returns key attributes

1. Controller memory buffer (CMB). 2. Subsystem local memory (SLM). 3. Cyclic redundancy check (CRC). All images and/or graphics within this slide are the property of KIOXIA America, Inc. (KAI) and are reproduced with the permission of KAI.

Data Ingestion Pipeline with Offload

Storage Controller

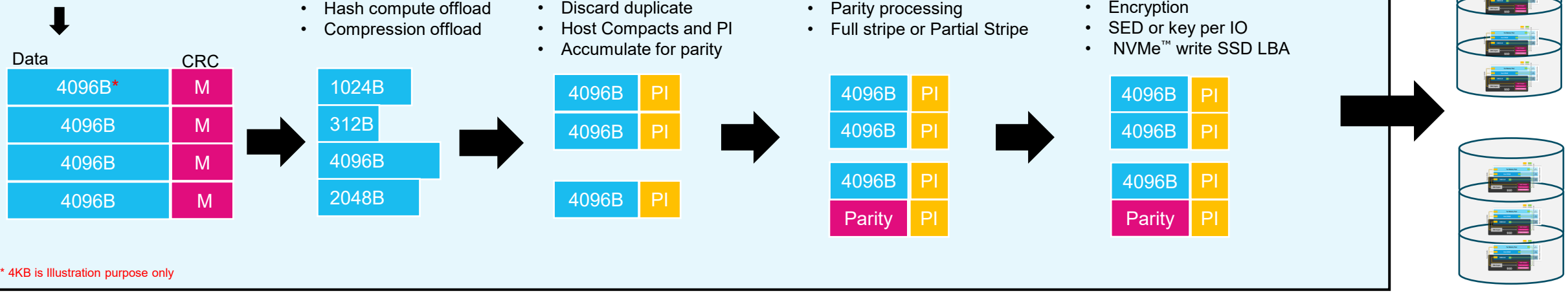
Step1: Host places data in appropriate buffer

Step2: Data Reduction and Compaction

Step3: Parity Compute Offload

Step4: Store Data to SSD

Host Memory or CMB[1] or SLM[2] Namespace

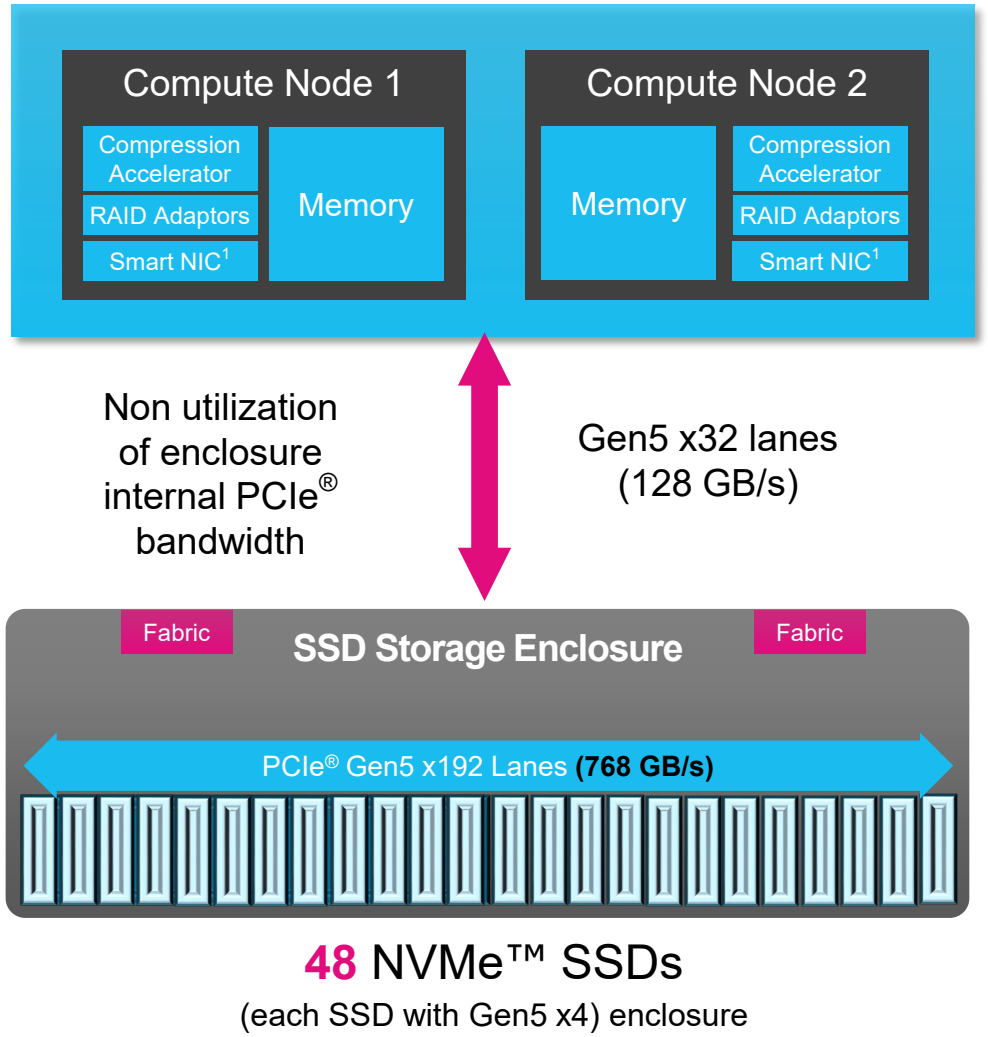


* 4KB is Illustration purpose only

 1. Controller Memory Buffer (CMB). 2. Subsystem Local Memory (SLM). 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 = 230 bits = 1,073,741,824 bits, 1GB = 230 bytes = 1,073,741,824 bytes, 1TB = 240 bytes = 1,099,511,627,776 bytes and 1PB = 240 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.

Storage Systems Discard PCIe® Bandwidth

Compute for Storage Functions



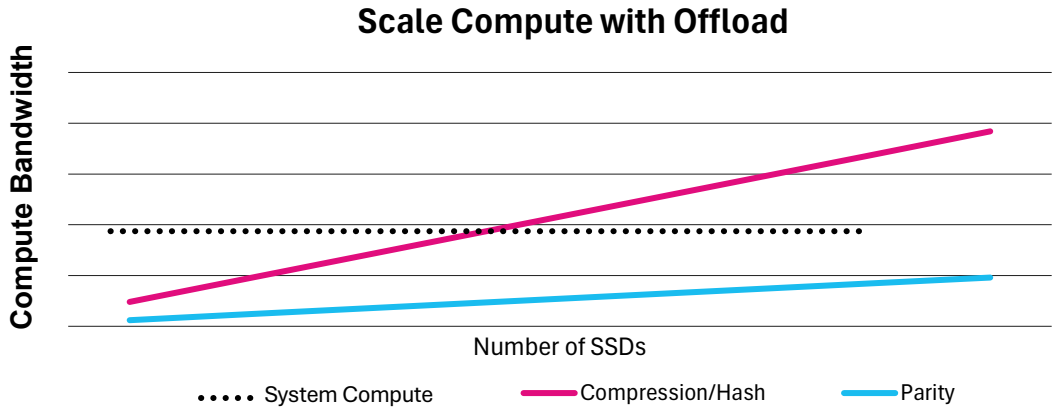
48 NVMe™ SSDs

(each SSD with Gen5 x4) enclosure

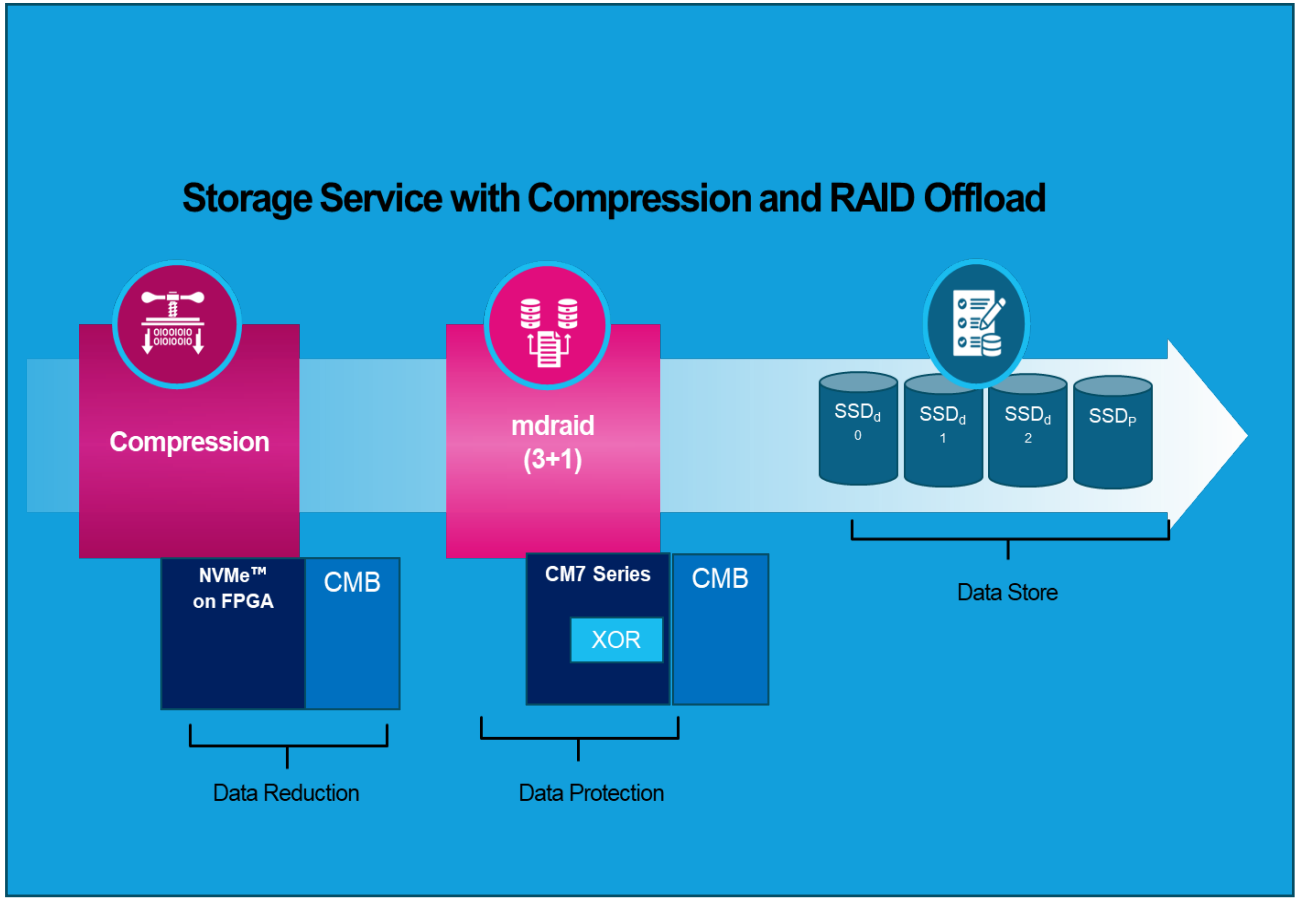
Compute Node

- Use SSD's accelerators and reduce memory footprint
- Scale up/out with every added SSD
- Leverage storage enclosure unused PCIe Bandwidth
- Cost effective and power efficient

- **With Offload Capable SSDs** (for illustration, not official specs.)
  - CMB/SLM memory bandwidth @ 10 GB/s/ SSD :480 GB/s
  - Parity compute bandwidth @ 2 GB/s/ SSD : 96 GB/s
  - (De)compression @ 10 GB/s :480 GB/s



Results from Data Pipeline Offload Proof of Concept (on Host with CPU)



No Offload (CPU) vs Offload

	No Offload	Offload	% Benefit
Write Bandwidth	~140 MB/s	~140 MB/s	-
Compression Ratio	2.5x	2.5x	-
Compression (gzip) CPU Core	200%	~1%	~199%
RAID CPU Resources	4%	4%	-
DRAM Bandwidth	~600 MB/s	~160 MB/s	3.8x

Results from RAID Offload Demonstration

RAID Offload: Proof of Concept (PoC) Results
(with KIOXIA CM7 Series SSDs and mdraid 5)

System	KIOXIA CM7 PCIe ®Gen4 x4 – mdraid 5	RAID Offload	% Benefit
CPU Utilization	**42%**	**37%**	**12% Reduction**
DRAM Bandwidth (in MiB^2/s)	**3450**	**340**	**91% Reduction**

System: DELL® PowerEdge™ R650xs Intel Xeon® Gold 6338N 2.2GHz (2 Socket, 32 Cores) PCIe® 4.0 , SSDs: 5xCM7 Gen4 (1.92 TB[3])

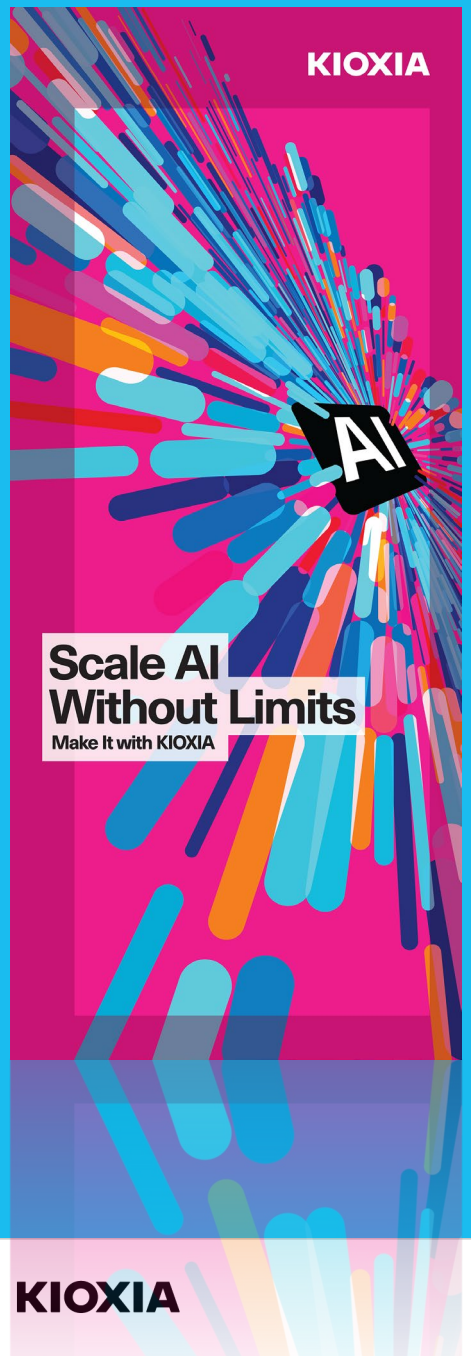
I/O workload: FIO[1] 512K Random Write @ 950 MB/s

Data Scrubbing PoC Results

	Offload Disabled	Offload Enabled
Scrubbing Time	**129s**	**91s**
DRAM Bandwidth	**10.24 GB/s**	**1.43 GB/s**
Total CPU Utilization	**99.5%**	**~70%**
L3 Cache Misses	**14.7M**	**4M**
Total PCIe® Write (MB/s)	**3694**	**159**

Demonstration from the KIOXIA booth at FMS 2024. All images and/or graphics within this slide are the property of KIOXIA America, Inc. (KAI) and are reproduced with the permission of KAI. PCIe is a registered trademark of PCI-SIG. Dell and PowerEdge are trademarks of Dell Inc. or its subsidiaries. Intel and Xeon are trademarks of Intel Corporation or its subsidiaries. All other company names, product names, and service names mentioned herein may be trademarks of their respective companies. 1. Flexible I/O Test (FIO). 2. Mebibyte 3. 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 = 230 bits = 1,073,741,824 bits, 1GB = 230 bytes = 1,073,741,824 bytes, 1TB = 240 bytes = 1,099,511,627,776 bytes and 1PB = 240 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.

Summary



- Storage systems data services is series of compute functions and ready to offload
- The unused PCIe® bandwidth can be leveraged efficiently
- Host managed standard fixed compute functions can be integrated to existing applications with nominal effort
- Scale out/up storage systems sustainably with every added SSDs

See the Offload PoC Demonstration at the KIOXIA Booth!

KIOXIA Booth #307

KIOXIA