# Data Scrubbing and Rebuild Using Offload

Aug 7, 2025

Devesh Rai
KIOXIA America, Inc.

FMS
the Future of Memory and Storage



KIOXIA

# Host Orchestrated Compute Offload Building Blocks



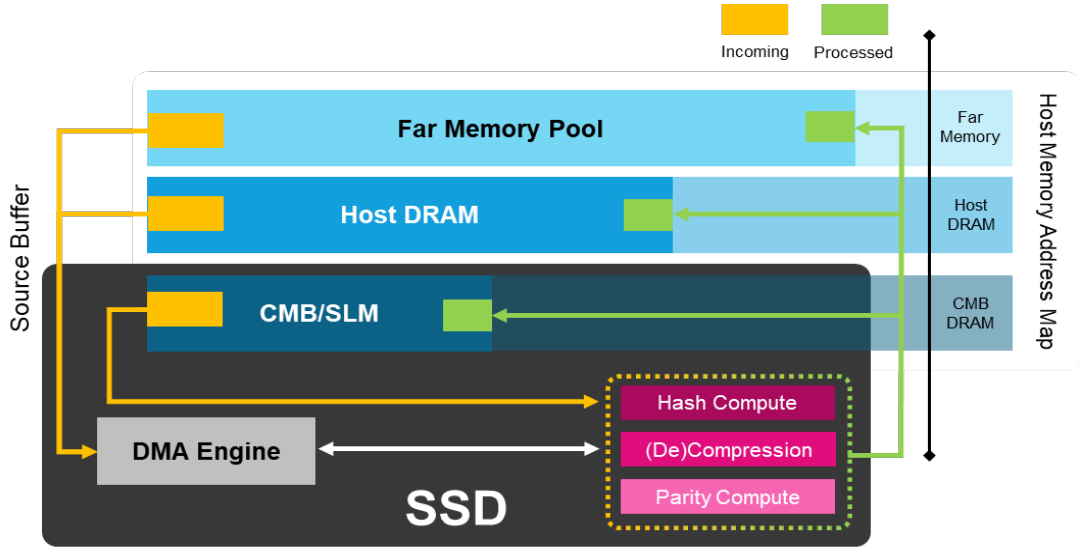
**Power-efficient compute engines**

**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**

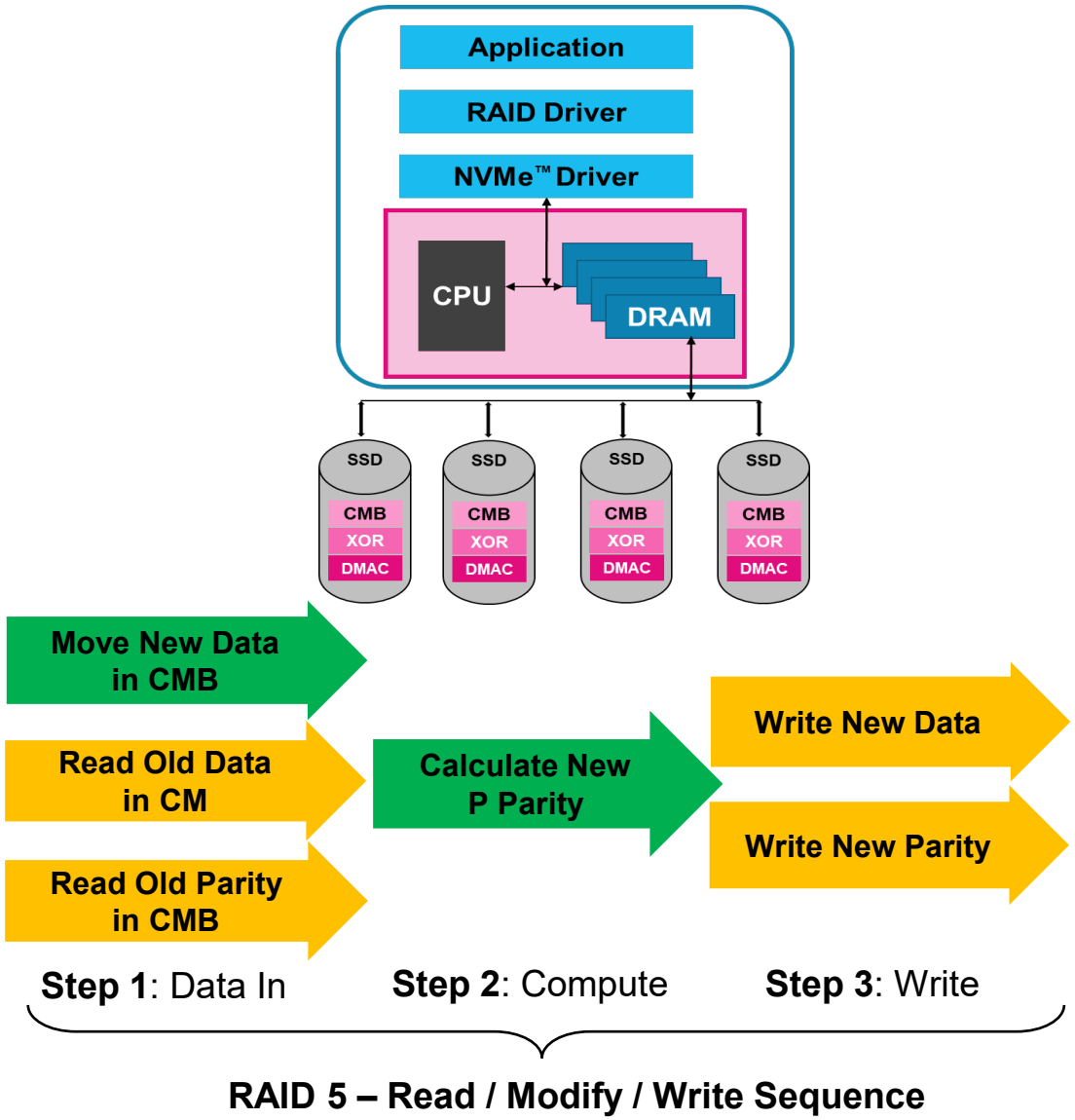


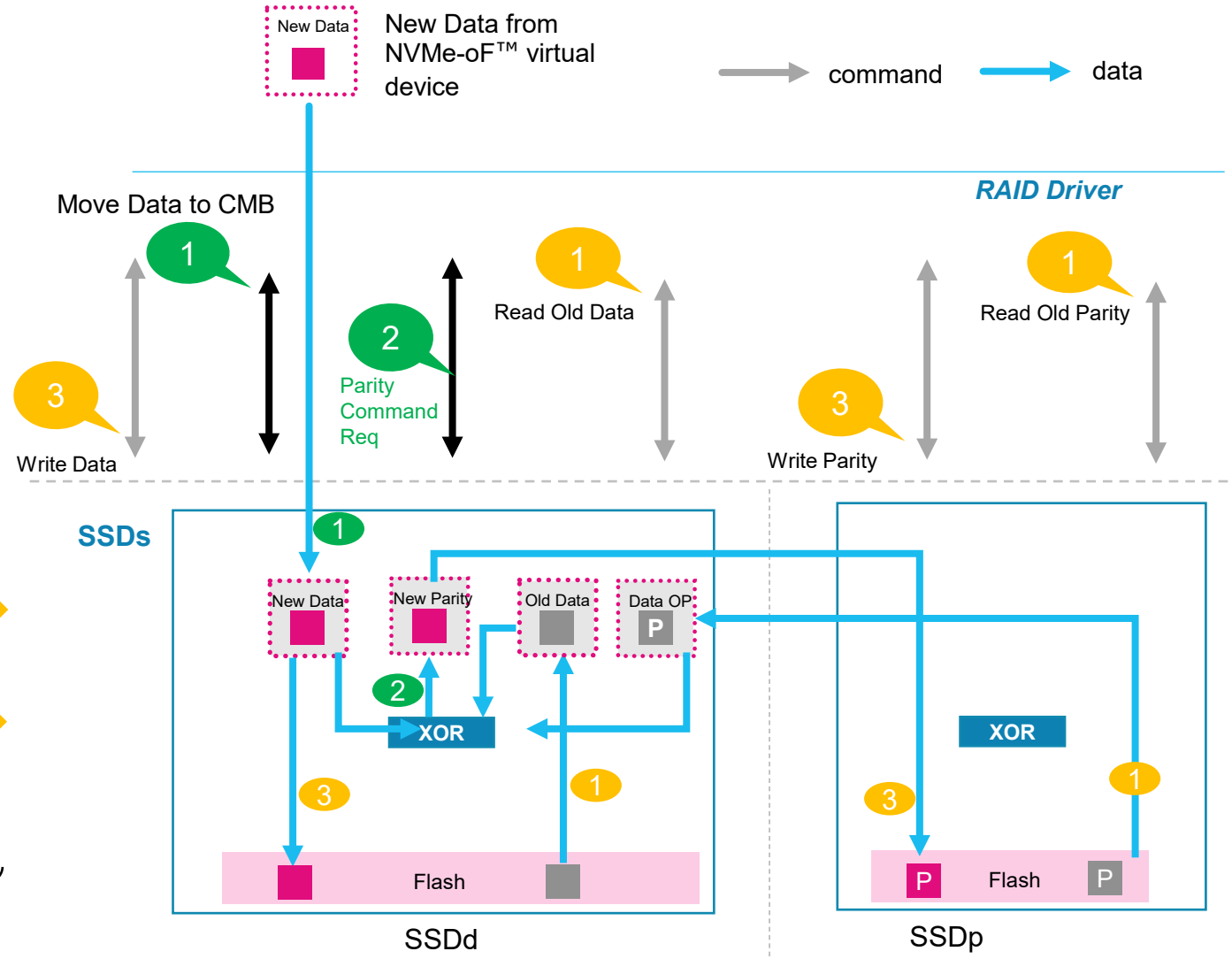

Host places data in appropriate buffer (host memory/ far memory/ CMB[1]) → Host issues command to the SSD → Compute engine on the SSD executes the command → Processed data is placed into the destination buffer → Command completion returns key attributes

KIOXIA

1. Controller memory buffer (CMB). 2. Subsystem local memory (SLM). 3. Cyclic redundancy check (CRC). 

# Command and Data Flow Example for RAID 5 Write



# Fail in Place Mechanism at the SSD Level

# KIOXIA Built-in SSD Level Failure Mitigation



| Failure Mitigation Features | Description |
|---|---|
| **Wear Leveling** | Internally developed “Always on” algorithm to allow even media wearing |
| **Die Fail Tolerance and Recovery** | Enterprise and data center SSDs historically supports die fail tolerance and recovery |
| **Customizable End-of-Life Behavior** | When and what actions to take upon reaching certain thresholds |
| **Overprovisioning (OP) and Available Spare** | Available Spare, critical warning in SMART[1] / health information<br>**Read Only mode if spare capacity falls below threshold** |
| **Dual-Port Functionality in Enterprise SSDs** | Ensures high availability in case of path failure |

**Data verification mechanism still required to ensure data integrity at File/Object/Stripe level.**



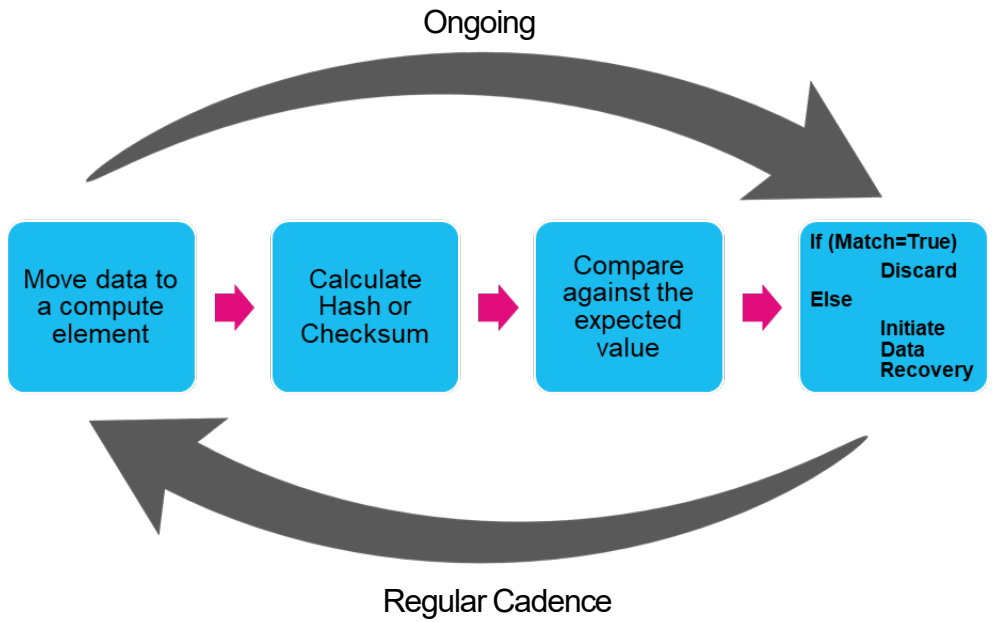
1. Self-Monitoring, Analysis and Reporting Technology (SMART) attributes

# Data Scrubbing Using Offload

# Data verification using Data Scrubbing in Conventional Setup



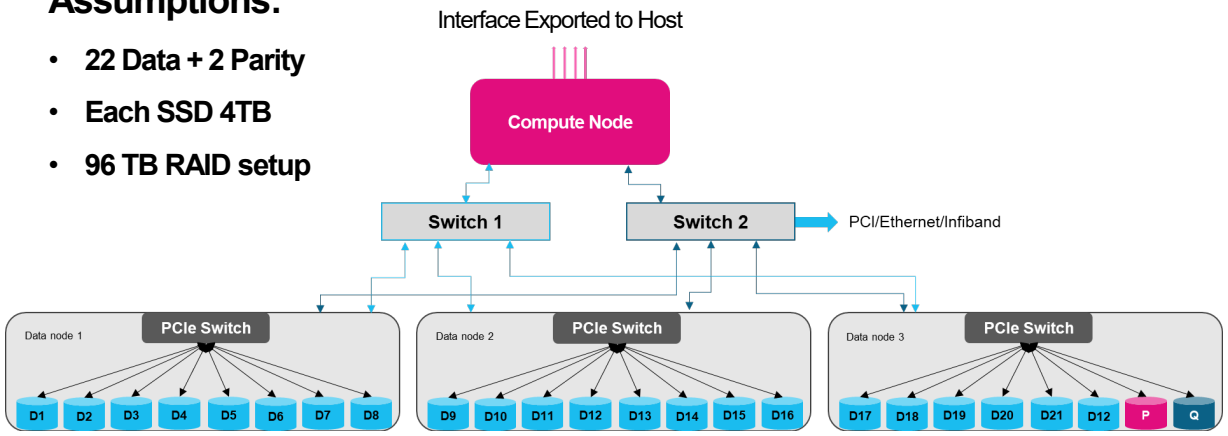
- **Data Scrubbing: Early detection and correction of errors**
- **Method employed : Hash, checksum or RAID technology**
- **Supported Level : File, Object and RAID Stripe**




**All data movement during scrubbing operation is an overhead penalty paid to ensure data integrity**

## Assumptions:

- **22 Data + 2 Parity**
- **Each SSD 4TB**
- **96 TB RAID setup**




**Compute node performing data and parities verification for one stripe using RAID**

1. P + D1 + D2 + D3 + D4 …… + D22 = 0
2. Q + g1.D1 + g2.D2 + g3.D3 + ….. + g22.D22 = 0

**For each Scrubbing Cycle:**

- **96TB data moves over PCIe®, network and CPU**
- **192TB through memory subsystem**

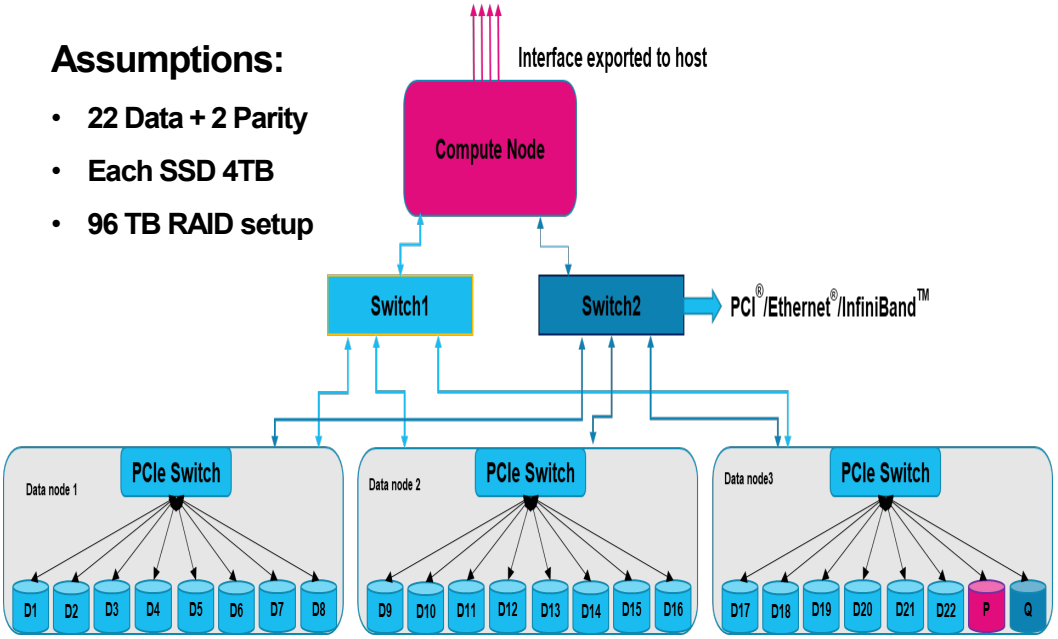
Assumptions created by KIOXIA in-house engineering team.

# Error Detection using Data Scrubbing with RAID/EC Offload

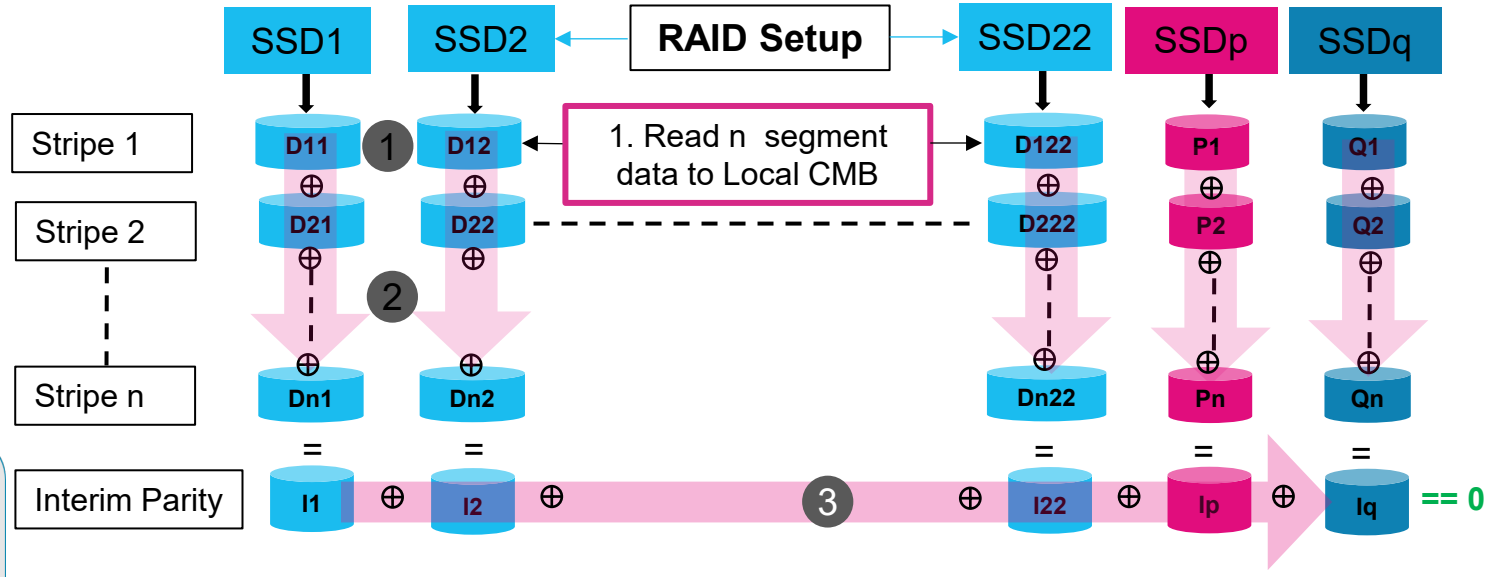


**Assumptions:**

- **22 Data + 2 Parity**
- **Each SSD 4TB**
- **96 TB RAID setup**




Assumptions created by KIOXIA in-house engineering team




2. On each SSD, interim parity by I1 = D11 + g1.D11 + D21 + g2.D21 + ……+ Dn1 + gn.Dn1
3. On any SSD, final parity by I1 + I2 + I3 + ……. + Ip + Iq == 0 ➔ **Success**

- **Using 3 step process, system resource usages reduced by ~99%**
- No data passes through CPU and DRAM on compute node
- For n stripes, only one stripe moves over network and PCIe®
- **A sustainable and scale out solution**
- Dnm=> Data segment D of nth stripe on SSDm

| System Resources to Scrub 96 TB (subject to change) | Conventional Method | KIOXIA Data Scrubbing Offload | % of Savings |
|---|---|---|---|
| Data passes through CPU for parity computation | 96 TB | 0 TB | 100% |
| Data passes through DRAM controller | 192 TB | 0 TB | 100% |
| Data passes through network | 96 TB | 0.4 TB | 99% |
| Data passes through PCIe® interface | 96 TB | 0.4 TB | 99% |

**A very flexible, geometry agnostic and adoptable solution for early error deduction**

KIOXIA

# Rebuild Using RAID Offload

# Efficient Rebuild of 2 SSDs using RAID offload

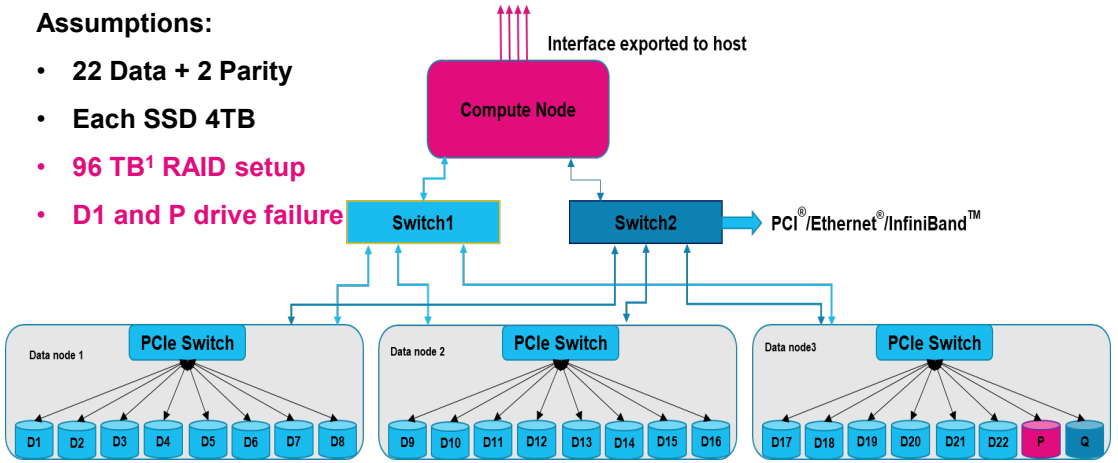


1. *Read good data segments of one RAID strip on compute node*
2. *Compute these two equations in order*
   a. *D1 = (g2.D2 + g3.D3 + ….. + g22.D22 + Q)/g1*
   b. *P = D1 + D2 + D3 + D4 …… + D22*
3. *Write D1 and P to new SSDs on respective stripe*
4. *Repeat steps 1-3 for each RAID stripe generation*

**Without offload**

**Assumptions:**

- **22 Data + 2 Parity**
- **Each SSD 4TB**
- **96 TB[1] RAID setup**
- **D1 and P drive failure**




1. *Read Dm to Dn segments in the CMB of one SSD in each data node*
2. *Calculate node specific intermediate data $D_{in}$ and parity $P_{in}$. For example, on node 1*
   1. *$P_{i1}$ = D2 + …….+ D8*
   2. *$D_{i1}$ = (g2.D2 + …….+g8.D8) / g1*
3. *Read intermediate data $D_{in}$ and parity $P_{in}$ from each data node on compute node and compute D1 followed by P parity*
   1. *$D1 = D_{i1} + D_{i2} + D_{i3}$*
   2. *$P = D1 + P_{i1} + P_{i2} + P_{i3}$*
4. *Write P and D1 to new SSDs on respective stripe*
5. *Repeat steps 1-4 for rebuilding each RAID stripe*

**With offload**

| System Resources to build two drives (subject to change[2]) | Conventional Method | KIOXIA Data Rebuild with Offload | % of Savings |
|---|---|---|---|
| Data passes through CPU for parity computation | 88 TB | 24 TB | 73% |
| Data passes through DRAM controller | 192 TB | 64 TB | 66% |
| Data passes through network | 96 TB | 32TB | 66% |
| Data passes through PCIe® interface | 96 TB | 108TB | -12% |

**Rebuild with offload can saturate write line speed**



System resource usage testing performed in-house by KIOXIA engineers. 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. 2. Numbers subject to change based on system used and SDD features. PCIe is a registered trademark of PCI-SIG. Ethernet is a registered trademark of Fuji Xerox Co., Ltd. InfiniBand is a trademark of the InfiniBand Trade Association. I Other company and product names may be trademarks of the respective companies with which they are associated. Tables and charts created by KIOXIA engineers. Images and/or graphics within this slide are the property of KIOXIA America, Inc. (KIOXIA) and are reproduced with the permission of KIOXIA.

# Summary

- SSD level data loss mitigation to
  - Avoid rebuild process specially with high-capacity SSDs
  - Recovery of data even in case of die failure
- Efficient data scrubbing method to detect failure early while saving significant system resource cost
- Efficient rebuild process saving significant system resource cost and time
- Data scrubbing and rebuild is based on KIOXIA RAID Offload Technology, and 'Best of Show' award winner at the Future of Memory and Storage (FMS) 2024 conference

KIOXIA

Definition of capacity: KIOXIA 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 1GB = 2^30 = 1,073,741,824 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, such as Microsoft Operating System and/or pre-installed software applications, or media content. Actual formatted capacity may vary.

All company names, product names and service names may be trademarks of their respective companies.

Images are for illustration purposes only.