



SNIA DEVELOPER CONFERENCE



BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

Optimizing HDD Interface in the Generative AI Era

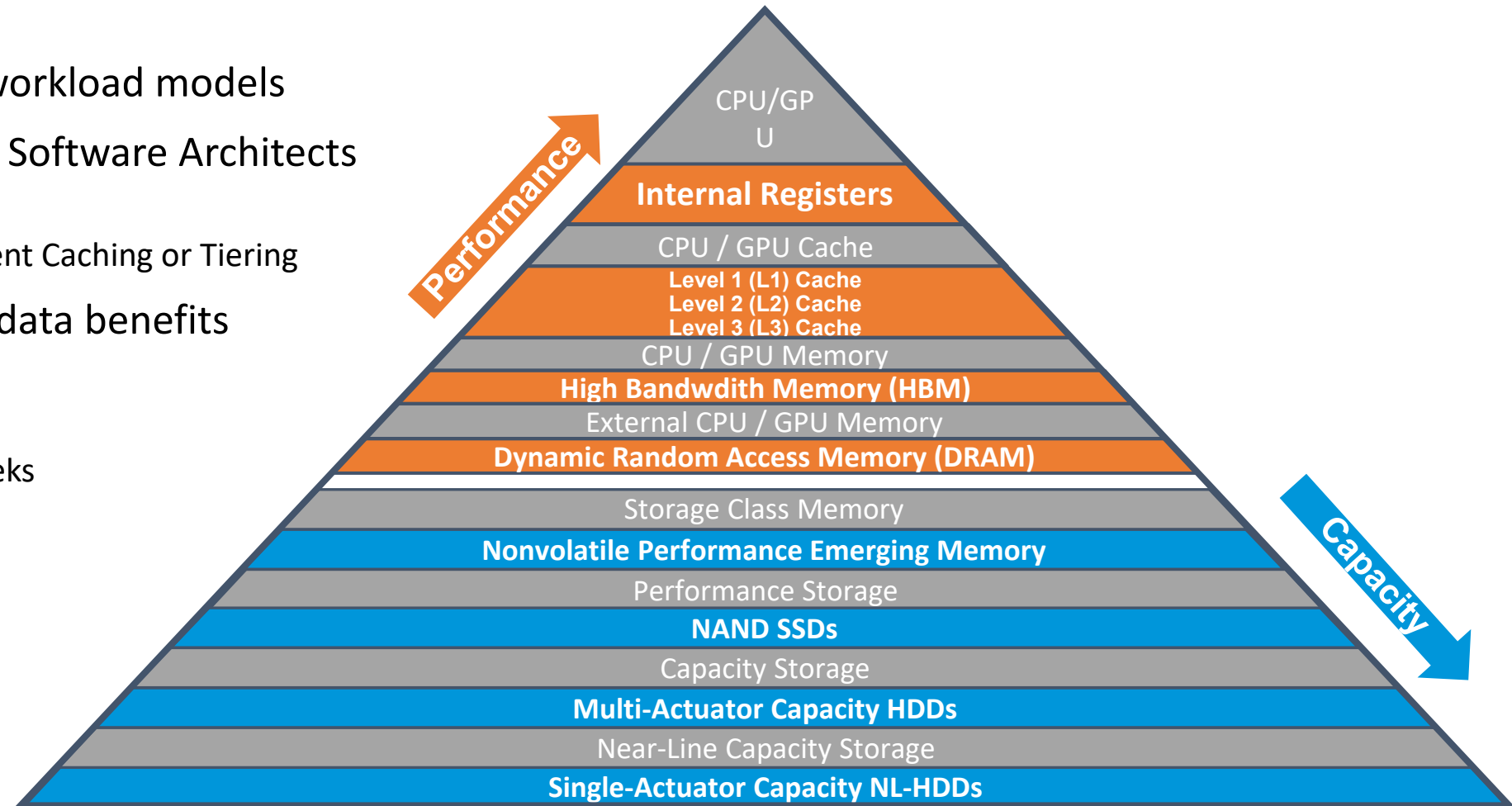
Seagate Technology

Mohamad El-Batal

Seagate OCTO Technologist & CTO Storage Systems

Generative AI-Memory & Storage Stack Hierarchy

- Different AI usage and workload models
- Workload innovation by Software Architects
 - Intelligent burst-buffer
 - Optimized-data-placement Caching or Tiering
- Streamlined sequential data benefits Hard Drives and SSDs
 - Write Amplification
 - Mechanical Actuator Seeks



Tier Class Volatile Non-Volatile



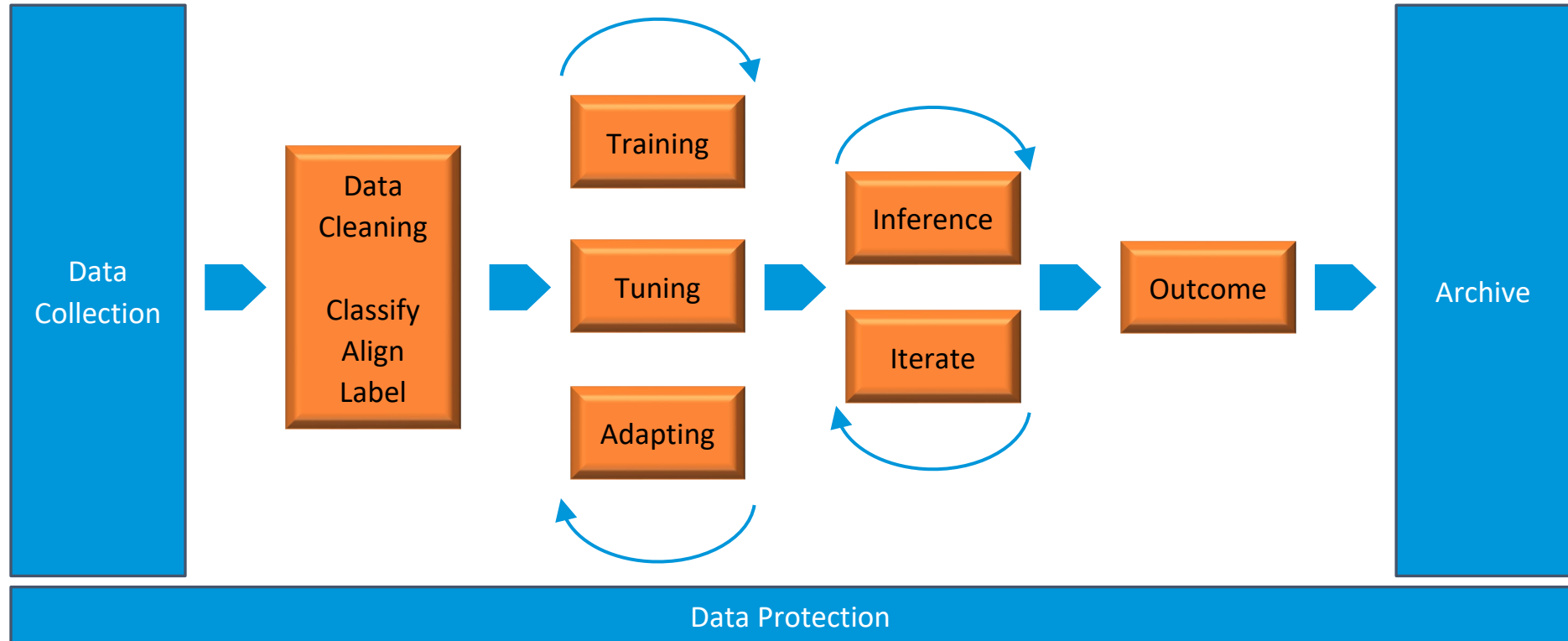
AI Requires GB/s/TB Layered TCO Efficiency

- Tiers can be higher latency and lower bandwidth
- Tiers must have same total aggregate bandwidth for 100% non-blocking bidirectional flow:
 - Little's law help absorb the Write sequentialization disparity with calculate buffer ratio
 - Read sequentialization is more complex and requires software-data-placement controls
- HDDs must maintain GB/s/TB Bandwidth/Capacity ratio as unit capacity grows
 - Typical SSD to HDD ratios in the tiered storage stack is:
 - 1 to 20 in bandwidth matching applications
 - 1 to 10 in cloud Storage applications
 - Dual- Actuator HDDs help keep this ratios stable with greater than 30TB capacities

Device Capacity	SA-HDD BW	DA-HDD BW	SA-HDD Ratio	DA-HDD Ratio
16TB	0.3 GB/s	0.6 GB/s	0.02 GB/s/TB	0.04 GB/s/TB
32TB	0.3 GB/s	0.6 GB/s	0.01 GB/s/TB	0.02 GB/s/TB

AI Data Management Workflow

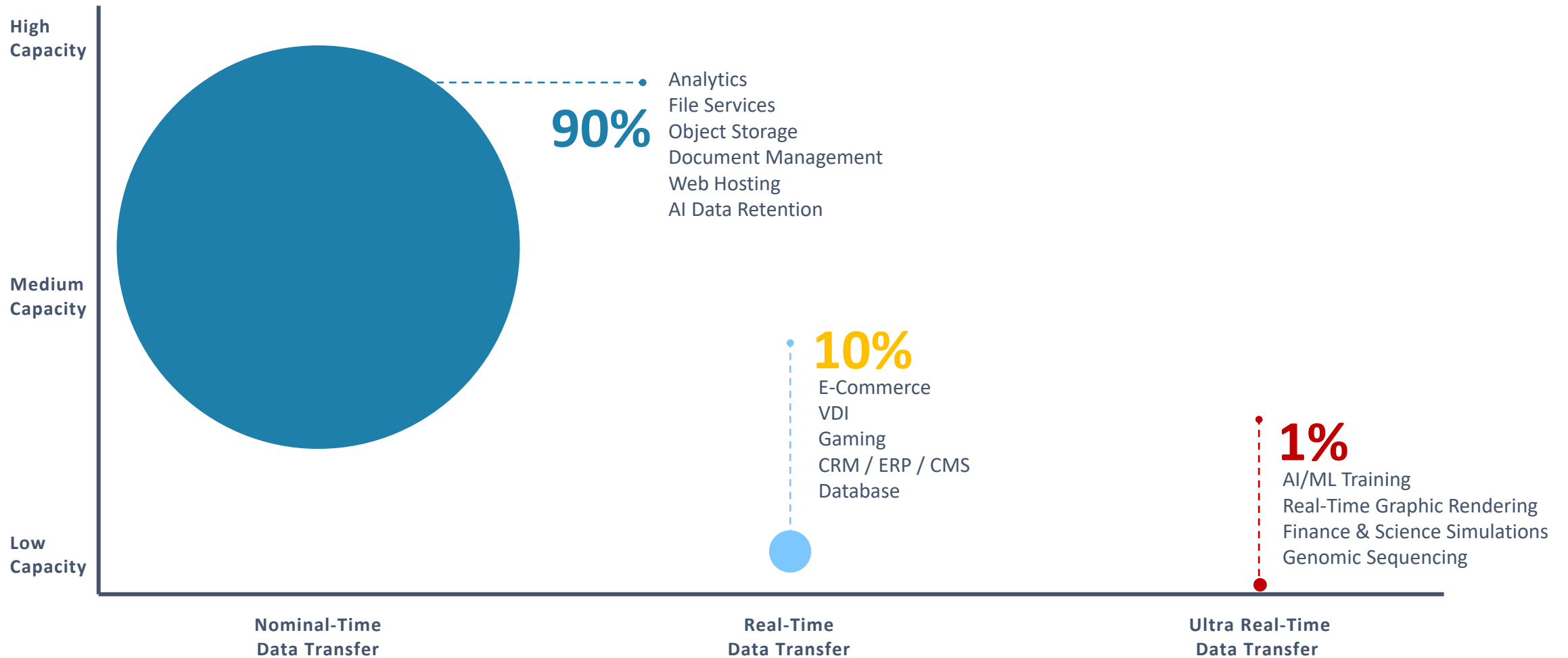
90% of total data is stored on **hard drives** & **10%** data is stored on **NAND/SSD**
in close proximity of Compute Resources



Most Enterprise Data Associated with General-Purpose Workloads

Image: Mapping enterprise workloads to EBs stored.

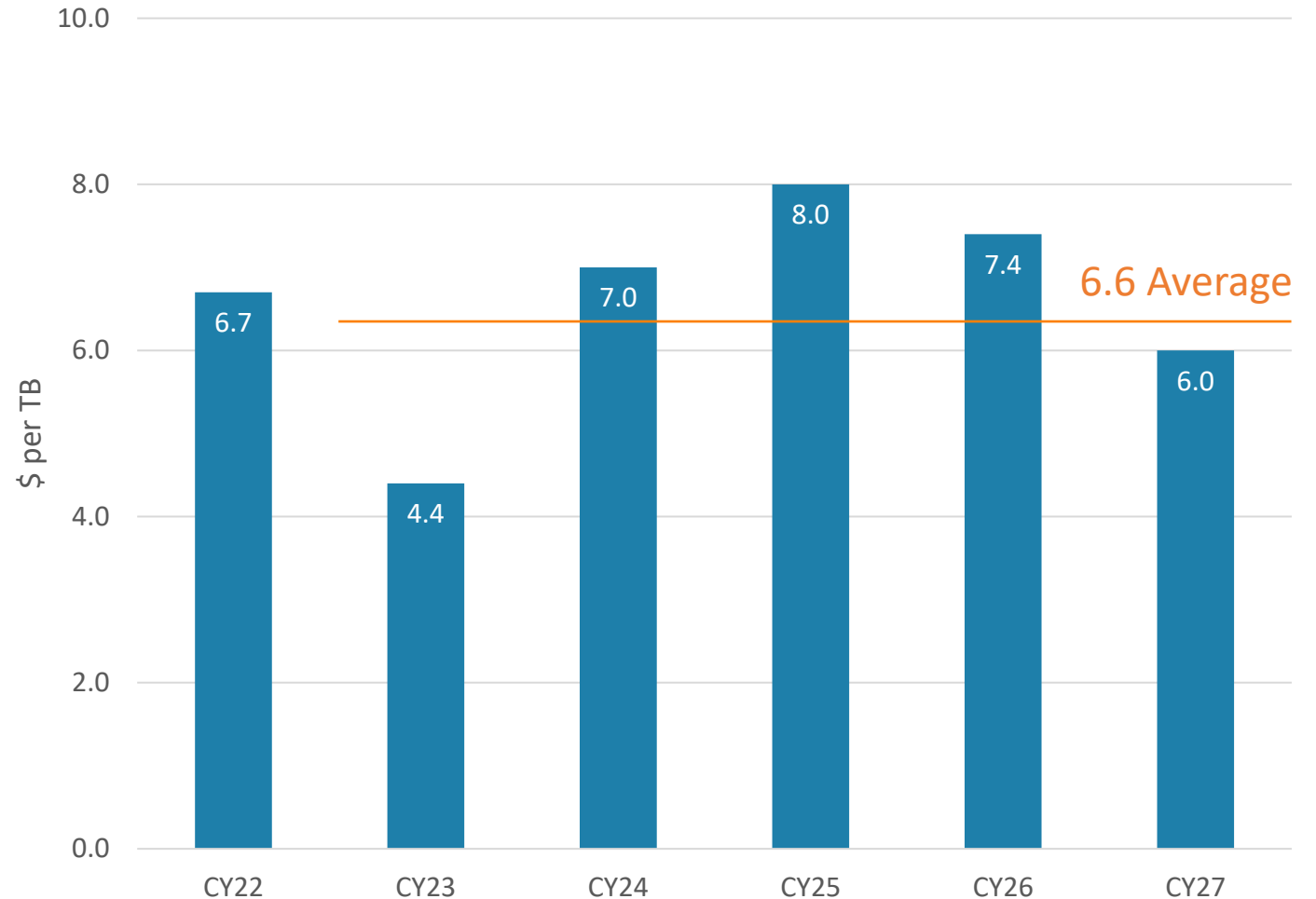
Source: Seagate analysis of IDC, (May 2023). *Streaming and Real-Time Data Redefined and Superimposed on IDC's Global DataSphere, 2022 (May 2023).*



HDDs will Maintain > 6:1 \$/TB Advantage over SSDs

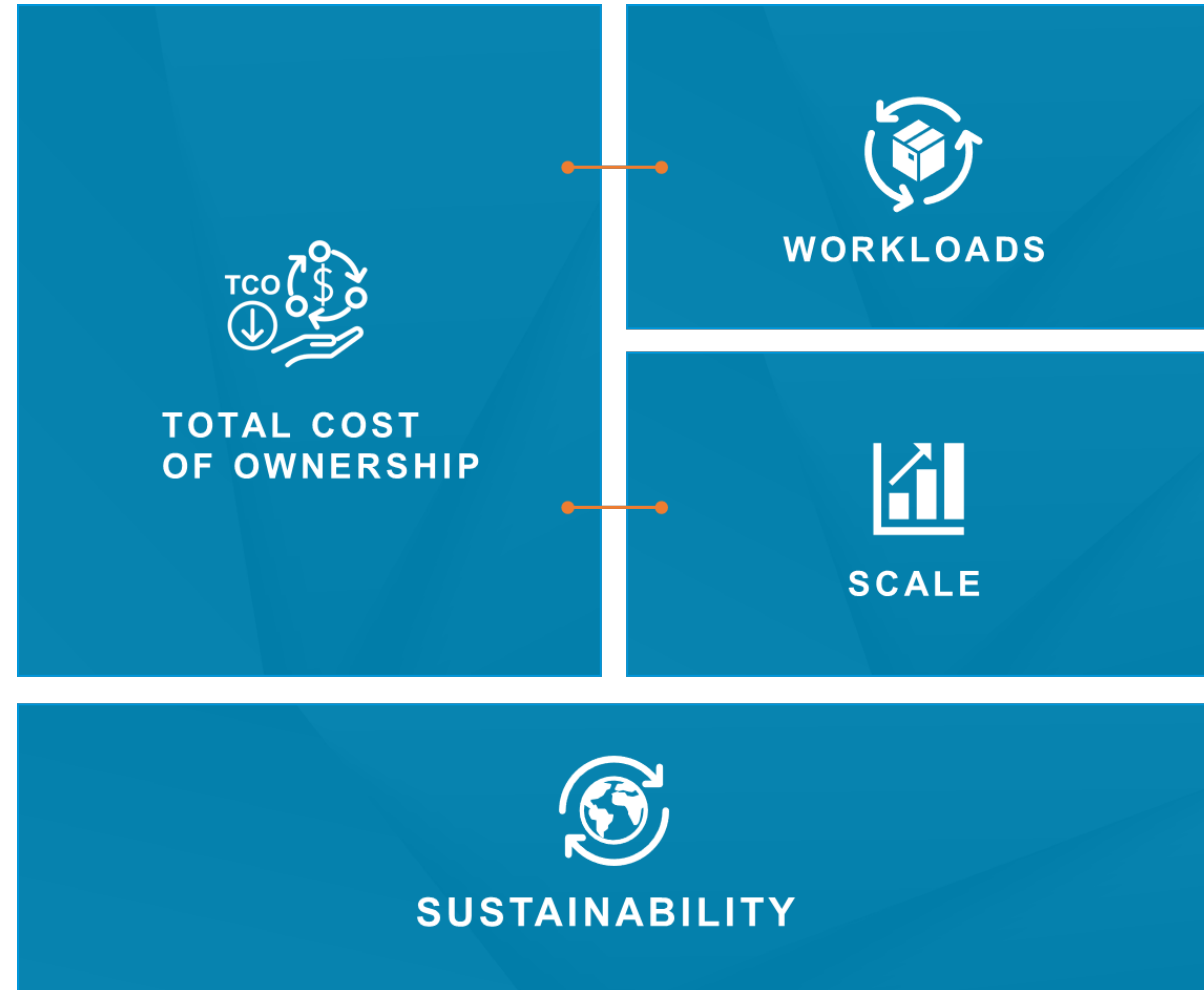
- SSD price-per-TB multiple compared with nearline hard drive price-per-TB
- Analysts project the price ratio to remain greater than 6 to 1 through 2027
- The average for this period is 6.6 to 1
- While dips happen, the pricing tends to rebound and equalize

Sources: Seagate's analysis based on *Forward Insights Q323 SSD Insights, Aug. 2023*; *IDC Worldwide Hard Disk Drive Forecast 2022-2027, Apr. 2023, Doc. #US50568323*; *TRENDFOCUS SDAS Long-Term Forecast, Aug. 2023*.



The World's Mega DataCenters Choose Hard Drives

Why?



Replacing all HDD Capacity w/ NAND is Cost-Prohibitive



\$10 investment required for a \$1 return




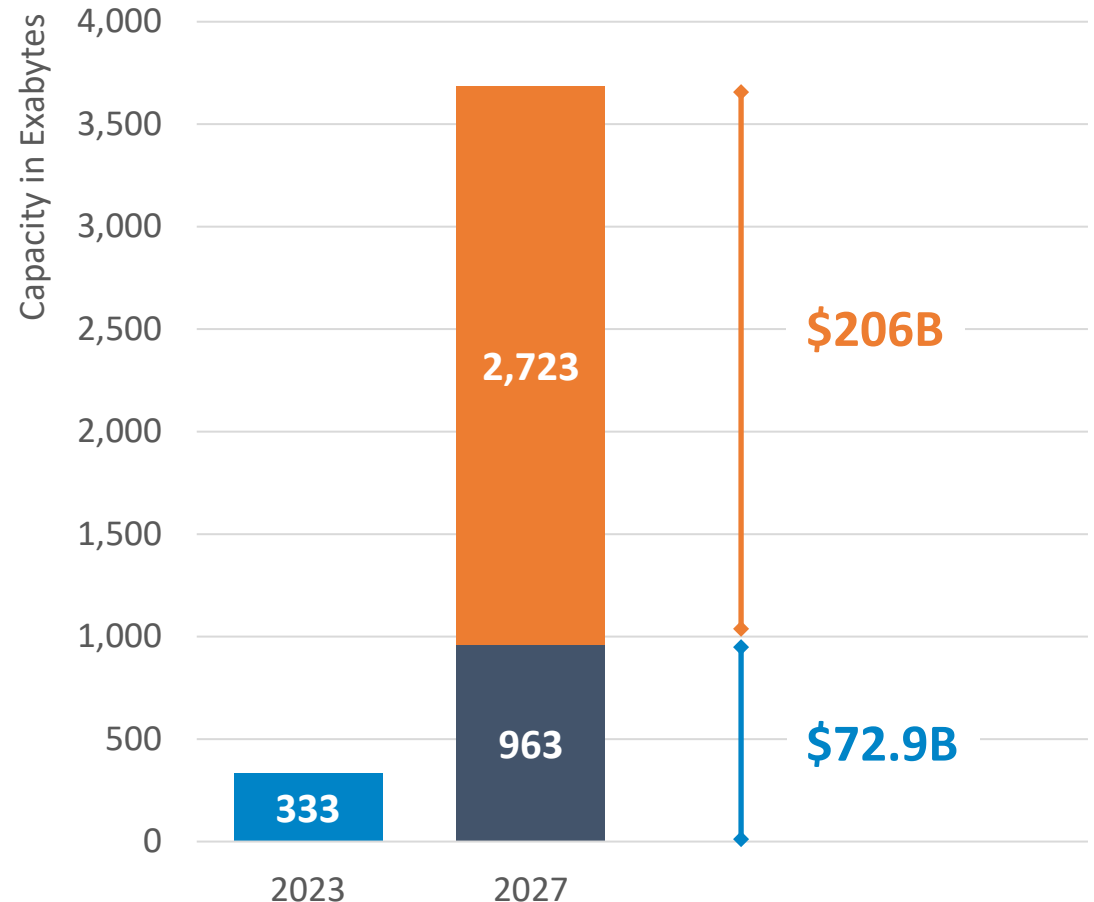
-  Exabytes the NAND industry produced
-  Exabytes the NAND industry is projected
-  Exabytes the NAND industry would need to produce, if they were to replace all HDDs

Image: Producing enough NAND Ebs to fulfill hard drive demand would be cost-prohibitive. Sources: Seagate analysis based on NAND Flash Platinum Datasheet by TrendForce and IDC Global Storage Forecast, 2023-2027 Doc. #US50851423, June 2023.

NAND Exabytes Production



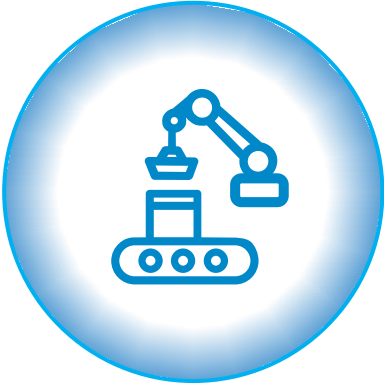
Hard Drives Deliver the Scale Needed for AI



Economies of Scale

> 6x lower

Acquisition cost than SSDs, enabling optimized TCO for AI.¹



Production at Scale

9x better

CapEx efficiency than the NAND industry.²



Sustainability at Scale

4x less operating power

10x lower EB/TB

Embodied Carbon (EB) per TB.³

1. Seagate's analysis based on *Forward Insights Q323 SSD Insights, Aug. 2023*; *IDC Worldwide Hard Disk Drive Forecast 2022-2027, Apr. 2023*, Doc. #US50568323; *TRENDFOCUS SDAS Long-Term Forecast, Aug. 2023*.
2. Seagate analysis of Yole Intelligence NAND Market Monitor, Q4 2023, CY2015-CY2023 compared to Seagate financials. Capex Efficiency: NAND 43%; Seagate 5%
3. Hotcarbon.org MS Azure white paper, SSD Storage Rack vs. Hard Drive Storage Rack: [A call for research on Storage Emissions.](#)

HDD in Generative AI Data Center Key Takeaways

- Generative AI experts predict exponential growth in data creation & consumption feedback loops that have insatiable appetites for all data center resources
- Such growth is impossible to accomplish without a sustainability focused Supply chain strategy that requires us to re-architect every growing critical component
- Today's Cloud & Hyperscale Storage stacks are very well balanced, but the Generative AI ecosystem will teach us all a few new lessons as it grows in scale

NVMe Storage Interface Consolidation



- OCP NVMe HDD Specification V1.0 Contributed in 2022 with 129 Active Members
- SATA cost replacement with Single-Port/Single-Lane customer demand
- Value propositions of NVMe-HDDs very clear to key CSPs & OEMs
- This is an opportunity to modernize and consolidate all mass storage with NVMe protocol
- Most Consumers, Providers and Partners are ready to enable this transition

Consumers

Drive Vendors

Systems Providers

Ecosystem

MSFT (Co-Chair)

Seagate (Co-Chair)

Seagate

Broadcom

Other CSPs

WDC (Co-Chair)

Wiwynn

Microchip

OEMs

Toshiba (Participated)

Super Micro

Amphenol TE

... others

Quanta

Molex

...etc.

Marvell

ARM

...etc.

NVMe Storage Interface Consolidation

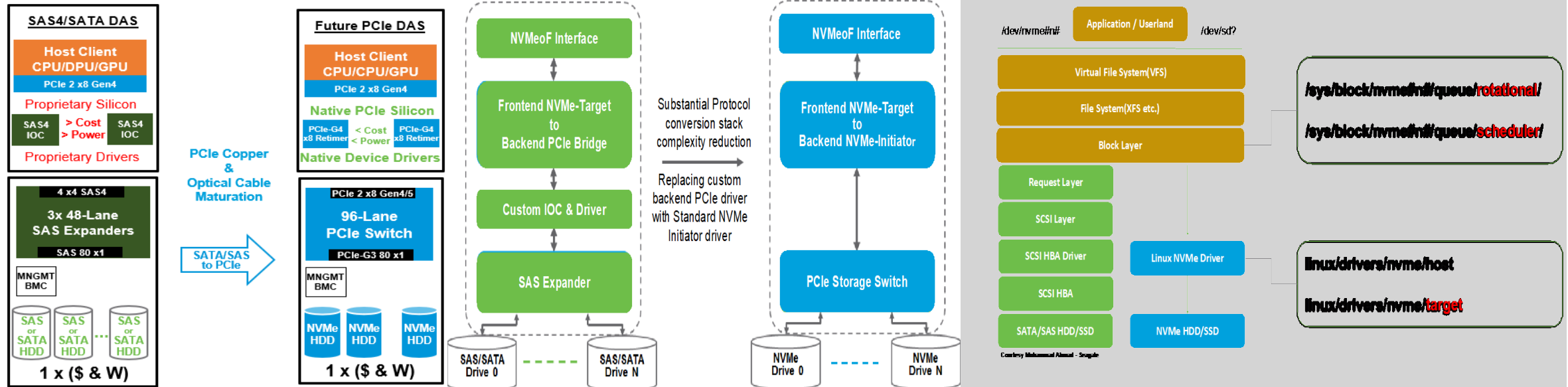
NVMe Committee TP4088 Rotational Media ratified in V2.0 – Q2 2021
Additional NVMe features planned for future NVMe HDDs:

- NVMe Committee TPAR 4091 Computational Programs (eBPF) – In Process
- NVMe Committee TPAR 4144 Command Duration Limit (CDL) – In Process
- NVMe Committee TPAR 4149 Elastic-Capacity (for Reman) – In Process
- Native SOC Tri-Mode (SAS, SATA & NVMe) ports are necessary for the transition
 - EDU1 – OCP Demo 2022 – Held back for SATA connector spec compliance
 - EDU2 – Sampled 2023 – Passed 80% Drivers Linux & MSFT vendor SIE testing
 - EDU3 – In development – NVMe2.0 with added firmware features possible by 2Q2025
 - Productization Likely to happen by 2027 to 2028, with major market pull from key CSPs

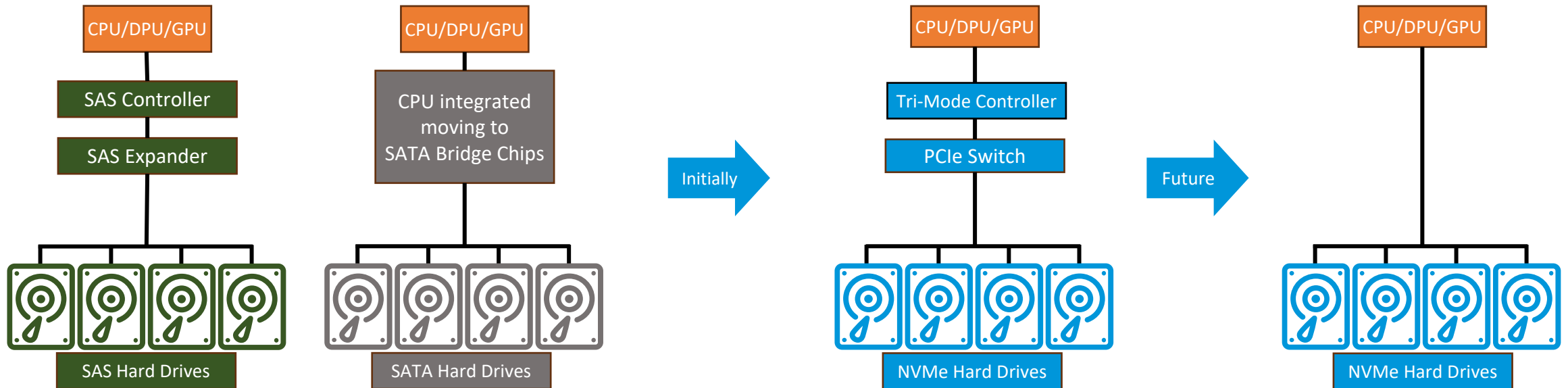


NVMe-HDD Unified Storage Stack & TCO Efficiency

- Provide > 50% reduction of Server-DAS & JBOD storage specific Silicon TCO & Power
- Standards based Silicon, OS-Stack, Driver and Software
- Unified Data-at-Rest, Data-in-Flight and Attestation security architecture consolidation for SSDs & HDDs
- CPU/GPU Direct-Attached optimizes Power, Bandwidth, Data-Reduction, Computational Storage ...etc.
- Efficient NVMeoF & GPU-Direct with NVMe-HDD Controller-Memory-Buffer(CMB) DPU/RNIC HW acceleration
- HDD Performance & Power reduction with NVMe Host-Memory-Buffer(HMB) NVC & Metadata optimization
- Dual-Actuator HDD enablement for IOPs, Bandwidth, Queuing & Rebuild performance efficiency



NVMe Simplifies The HDD Topology

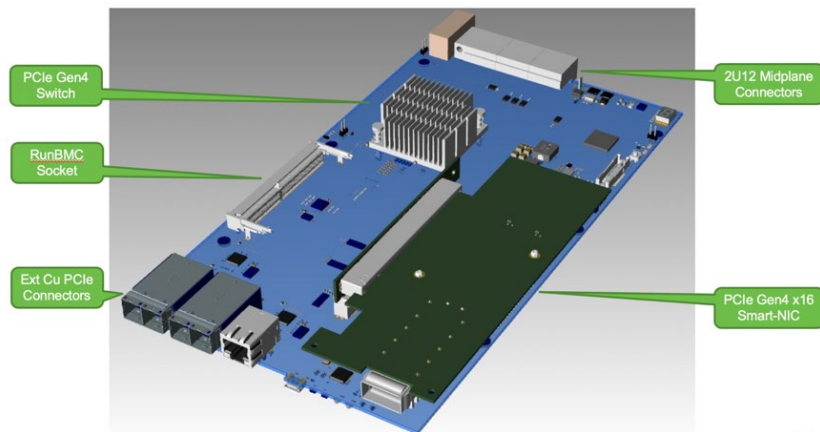
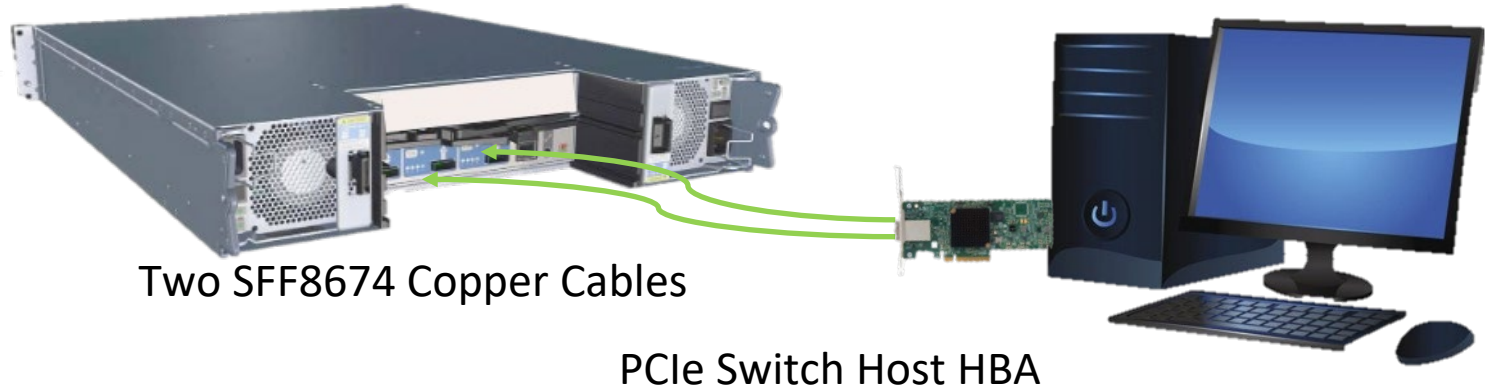


- Eliminates proprietary SAS & SATA proprietary Silicon and Drivers
- Common NVMe™ Driver/OS stack for HDD and SSD
- Native OS Namespace support for Multi-actuator
- Optimizes streamlined NVMe-oF™ Composability

EDU1 Demo at OCP Summit 2022



- PCIe-Gen4 IOMs Demoed 2022
- Supports NVMe-oF x16 RNICs

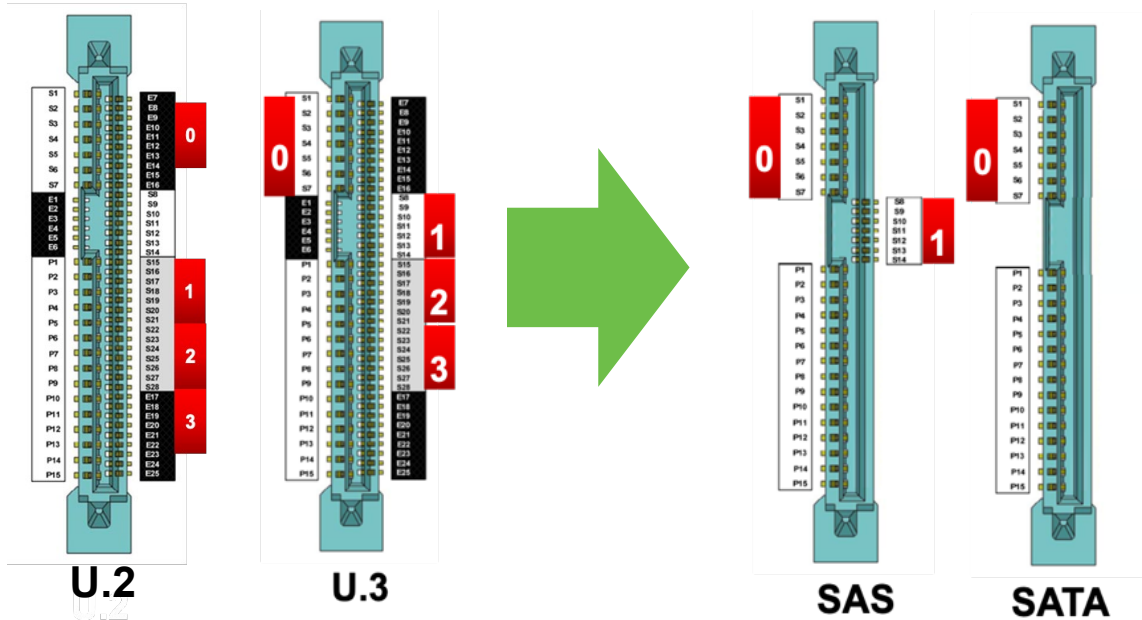


2U12-3.5" NVMe-HDD POC JBOF



Started With U.3 Then Moved to SAS/SATA Connectors

- Initially EDU used SFF8639 conn to validate all possible connections
 - Validated SRIS/SRNS Clockless PCIe-Gen3 capable Designs
 - No PCIe-Clocks, No PERST, No I2C/I3C/SMBus and No Dual-Port-Enable Signals ...etc.
- Current EDUs to use the SATA connector:
 - SATA Connector for Single-Port (Same HW SKU as SATA HDD)
 - SAS Connector for Dual-Port (Same HW SKU as SAS HDD) - TBD pending volume justification



https://docs.google.com/document/d/1pbgiMRCLb9im3j1MvPMZhtFZByP6M-C_/edit

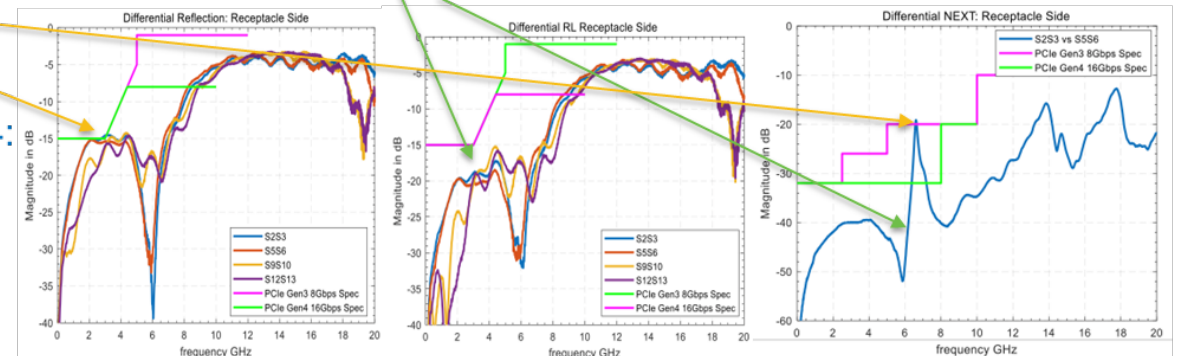
Market Timing	Current SKUs	Near-Term Transition	Long-Term Convergence
Hyperscalers	SATA(Single-Port)	SATA(Single-Port) <i>NVMe(Single-Port)</i>	<i>NVMe(Single-Port)</i>
Both	SAS(Dual-Port)	SAS(Single-Port)	<i>NVMe(Dual-Port)</i>
Enterprise		SAS(Dual-Port) <i>NVMe(Dual-Port)</i>	
	System Connector Receptacle	Device Connector Plug	Max PCIe link speed (8/16GT)
	6Gb SATA	6Gb SATA	
	SFF8482 - 12Gb SAS	6Gb SATA	8Gb/s
	SFF8482 - 24Gb SAS	6Gb SATA	
	SFF8482 - 12Gb SAS	12Gb SAS	8Gb/s
	SFF8482 - 24Gb SAS	24Gb SAS	8Gb/s & 16Gb/s
	SFF8482 - 24Gb SAS	12Gb SAS	
	SFF8639 - Gen4	6Gb SATA	
	SFF8639 - Gen4	12Gb SAS	8Gb/s
	SFF8639 - Gen4	24Gb SAS	8Gb/s & 16Gb/s

Connector SI Test Results Analysis

Conn-Pair-Comb*	PCIe Gen 3				PCIe Gen 4			
SI Test Type	IL	RL	NEXT	FEXT	IL	RL	NEXT	FEXT
SATA P/R-Pair	P	F	P	N/A	-	-	-	-
SAS3 P/R-Pair	P	P/M	P	P	P	F	P	P
SATA-P + SAS3-R Pair	P	P	P/M	P	P	F	F	N/A
SATA-P + SAS4-R Pair	P	P	P	P	F	F	F	N/A
SAS4 P/R-Pair	PP	PP	PP	PP	P	P	P	P
SAS3-P + SAS4-R Pair	P	P	P	P	F	F	P	P

We now have the targeted connector pairs SI test results from Amphenol:

- **Cost Optimized Drive w/ PCIe-Gen 3:**
 - SAS3-Plug/Receptacle connector pair **Marginal** with 85Ω backplane PCB return-loss; however, it **Passes** with backplane PCB at 92Ω impedance
 - SATA-Plug + SAS3-Receptacle connector pair **Marginal** regarding near-end-crosstalk; however, it **Passes** given the frequency is above 1.5x
 - SATA-Plug + SAS4-Receptacle connector pair passed
 - SAS4-Plug/Receptacle connector pair passed
 - SAS3-Plug + SAS4-Receptacle connector pair passed
- **Performance/HA Optimized Drive w/ PCIe Gen 3 or Gen 4:**
 - Only the SAS4-Plug/Receptacle connector pair passed

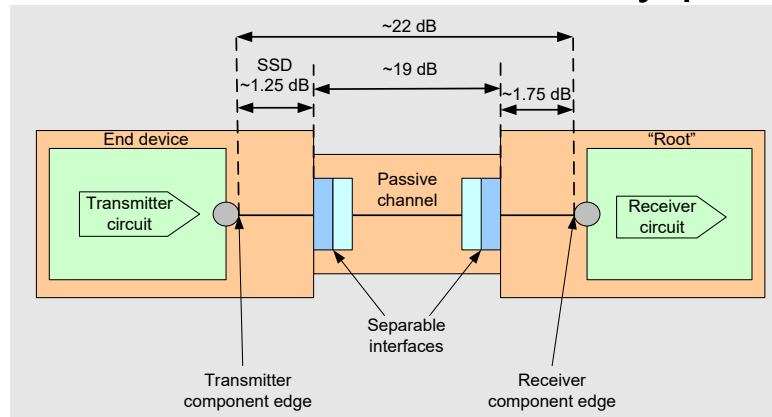


PCIe-Gen3 & SAS-3 Signal Integrity Channel Models

Low-Density and High-Density JBODs can be built with the typical <19" of PCIe-Gen3 to stay at $\sim 1 \times 10^{-15}$

PCIe-Gen 3

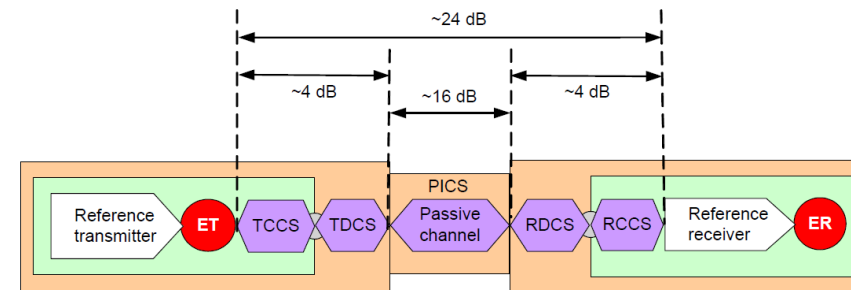
Insertion loss values at Nyquist



Backplane IL of 0.7/inch @ 4GHz
~25.7 inches (1 dB for connectors)
Shorten to increase BER from 1×10^{-12}
Data rate 128b/130b @ 8 GT/s

SAS-3

Insertion loss values at Nyquist



Backplane IL of 0.98/inch @ 6GHz
~15.3 inches (1 dB for connectors)
BER 1×10^{-15}
Data rate 8b/10b @ 12 Gb/s

PCIe Physical Layer Initiator Silicon Requirements

1. PCIe Root-Port SRIS or SRNS Initiator support
 - No RefCLK+/- requirements
 - No PERST# requirements
 - No SMBus I2C Sideband management requirements

2. PCIe Root-Port Hot-Plug Support
 - Down Stream Port Containment (DSPC) Support
 - Advanced queued commands cleanup and retry
 - Test and Validate on Insertion

3. PCIe Out-Of-Band transport layer attestation
 - Vendor Defined Messaging (VDM) for attestation Layer

https://docs.google.com/document/d/1pbgjMRCLb9im3j1MvPMZhtFZByP6M-C_/edit

Ac Coupling Termination

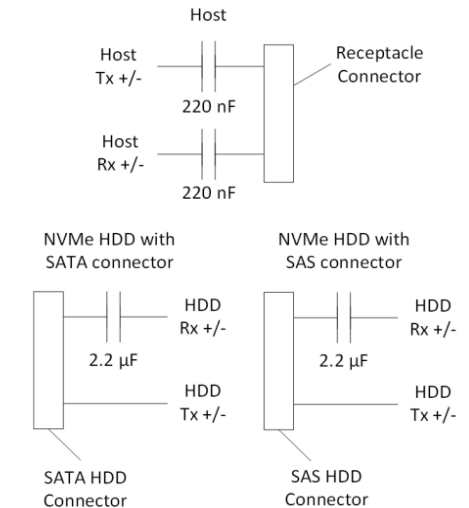


Figure 1: PCIe Required AC-Coupling for NVMe Host & Device

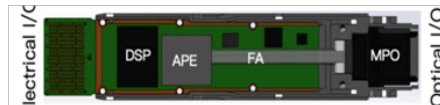
Location	Pin	Value	Tolerance
Host	S2 (Tx+)	220 nF	10%
	S3 (Tx-)	220 nF	10%
	S5 (Rx-)	220 nF	10%
	S6 (Rx+)	220 nF	10%
	S9 (Tx+)	220 nF	10%
	S10 (Tx-)	220 nF	10%
	S12 (Rx-)	220 nF	10%
Single port NVMe HDD	S2 (Rx+)	2.2 μF	20%
	S3 (Rx-)	2.2 μF	20%
Dual port NVMe HDD	S2 (Rx+)	2.2 μF	20%
	S3 (Rx-)	2.2 μF	20%
	S9 (Rx-)	2.2 μF	20%
	S10 (Rx+)	2.2 μF	20%

OCP PCIe Extended Connectivity Requirements

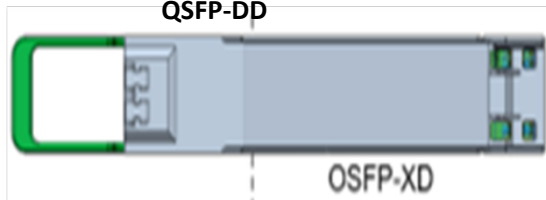
- OCP PCIe Extended Connectivity Requirements Rev-1.0 Summary:
 - Intra-Rack – Within a 21” or 19” rack connectivity DAC & AEC up to 3.0m in length
 - Inter-Rack – Adjacent rack connectivity AOC & DOC up to 7.0m for CXL & 10.0m for NVMe
 - SRIS/SRNS required and PCIe sideband signals are optional including SMBus
 - Common DAC, AEC & AOC Connector & Socket solutions on Root & Endpoint
- SIG EWG & OWG teams currently defining a CDFP equivalent functionality with:
 - Focus on PCIe x8 QSFP-DD pinout with a common socket for DAC, AEC & AOC
 - Working on x16 OSFP-XD next with common pinout for DAC, AEC & AOC
 - Retimer enabled AEC & AOC solutions must meet the OCP Power & Latency requirements
 - CXL prefers latency optimized Direct-Drive/Linear & Co-packaged Optics multi-wavelength solutions
 - Using CDR/Repeater/Redriver solutions that add <4ns Total-RTL are acceptable



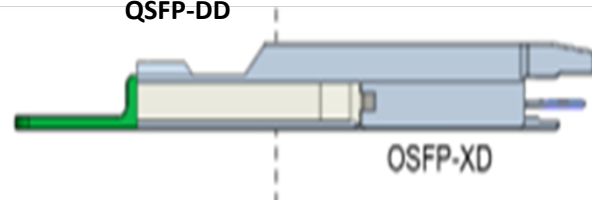
QSFP-DD



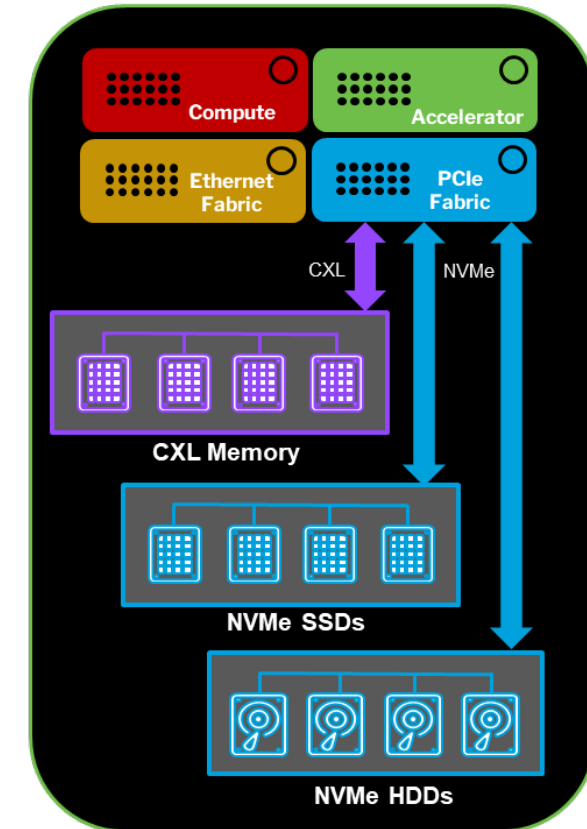
QSFP-DD



OSFP-XD



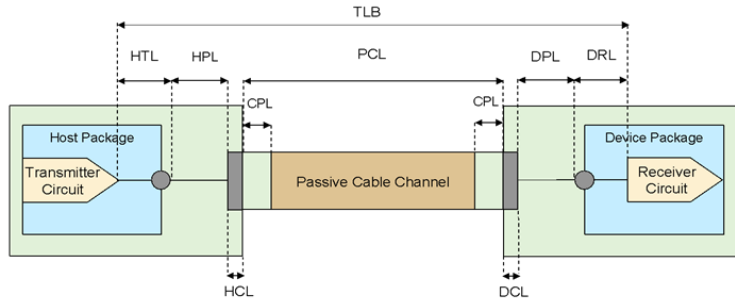
OSFP-XD



DAC = Direct Attached Connection
AEC = Active Electrical Connection
DOC = Direct Attached Optical
AOC = Active Optical Connection

PCIe 5.0 and 6.0 NVMe DAC, AEC & AOC Specific Requirements

PCIe-Gen5 Total Loss Budget (TLB) = 36dB => 10⁻¹² BER
 PCIe-Gen6 Total Loss Budget (TLB) = 32dB => 10⁻⁶ BER



TLB	Total Loss Budget	Cable AWG	IL* (db/m) @16GHz	IL* (db/m) @16GHz
PCL	Passive Cable Loss		100-Ohm	85-Ohm
HTL	Host Transmitter Loss	26	3.6	4.1
HPL	Host PCB Loss			
CPL	Connector PCB Loss	28	4.3	5.2
DPL	Device PCB Loss			
DRL	Device Receiver Loss	30	5.3	6.4
HCL	Host Connector Loss			
DCL	Device Connector Loss	32	7.2	8.3

Example viability test for 1-meter DAC cable in PCIe-5.0:

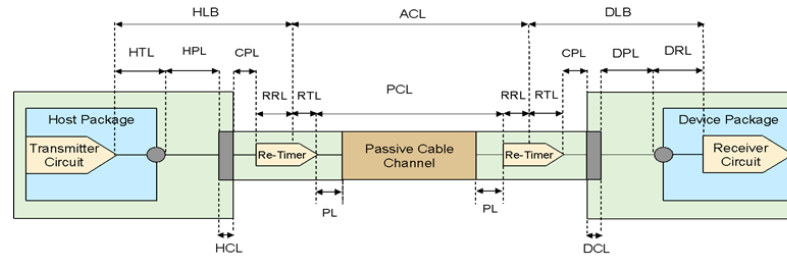
Cable AWG	IL* (db/m) @16GHz (PCIe-5.0 NRZ 36 db)	HTL(db)	HCL(db)	CPL(db)	DRL(db)	HPL(db)	DPL(db)
30	5.3	9	1	1	4	7.6	7.6
32	7.2	9	1	1	4	6.6	6.6

Example viability test for 1-meter DAC cable in PCIe-6.0:

Cable AWG	IL* (db/m) @16GHz (PCIe-6.0 PAM4 32db)	HTL(db)	HCL(db)	CPL(db)	DRL(db)	HPL(db)	DPL(db)
30	5.3	8	1	1	4	6.6	6.6
32	7.2	8	1	1	4	5.6	5.6

Up to 1-meter DAC

PCIe-Gen5 Total Loss Budget (TLB) = 36dB => 10⁻¹² BER
 PCIe-Gen6 Total Loss Budget (TLB) = 32dB => 10⁻⁶ BER



TLB	Total Loss Budget	Cable AWG	IL* (db/m) @16GHz	IL* (db/m) @16GHz
HLB	Host Loss Budget			
DLB	Device Loss Budget			
ACL	Active Cable Loss		100-Ohm	85-Ohm
PCL	Passive Cable Loss			
HTL	Host Transmitter Loss	26	3.6	4.1
HPL	Host PCB Loss			
CPL	Connector PCB Loss	28	4.3	5.2
PL	PCB Loss			
RRL	Retimer Receiver Loss	30	5.3	6.4
RTL	Retimer Transmitter Loss			
DPL	Device PCB Loss			
DRL	Device Receiver Loss	32	7.2	8.3
HCL	Host Connector Loss			
DCL	Device Connector Loss			

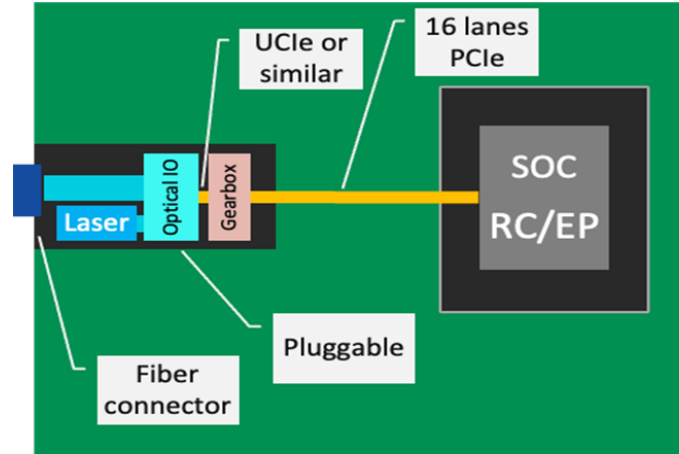
Example viability test for 3-meter AEC cable in PCIe-5.0:

Cable AWG	IL* (db/m) @16GHz (PCIe-5.0 NRZ 36db)	RTL(db)	2x PL(db)	RRL(db)	Margin (db)
30	16	4	2	4	10
32	22	4	2	4	4

Example viability test for 3-meter AEC cable in PCIe-6.0:

Cable AWG	IL* (db/m) @16GHz (PCIe-6.0 PAM4 32db)	RTL(db)	2x PL(db)	RRL	Margin (db)
30	16	4	2	4	6
32	22	4	2	4	0

Up to 3-meter EAC



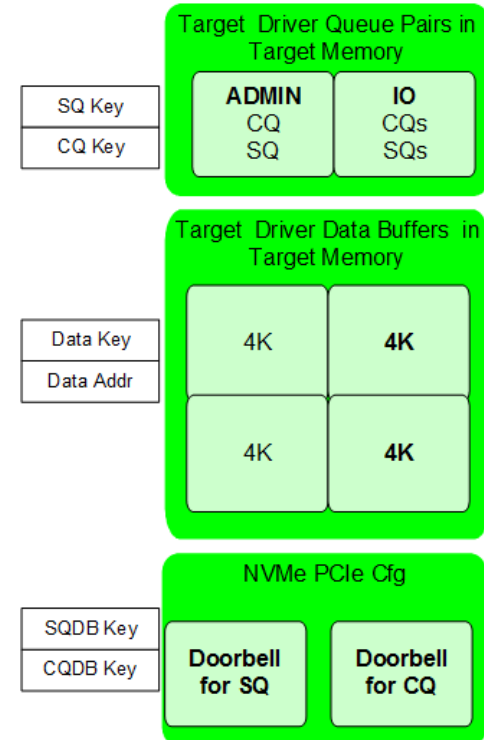
Rate	# of Lanes	Max Retimer plus Optical Circuit/Lane in (W)	Package Size (mm)
PCIe-5.0	1 x 4	1.0	tdb
PCIe-5.0	1 x 8	1.0	tdb
PCIe-5.0	1 x 16	1.0	tdb
PCIe-6.0	1 x 4	1.0	tdb
PCIe-6.0	1 x 8	1.0	tdb
PCIe-6.0	1 x 16	1.0	tdb

Up to 10-meter AOC

OCP PCIe Extended Connectivity Requirements v1.0b.docx - Google Docs

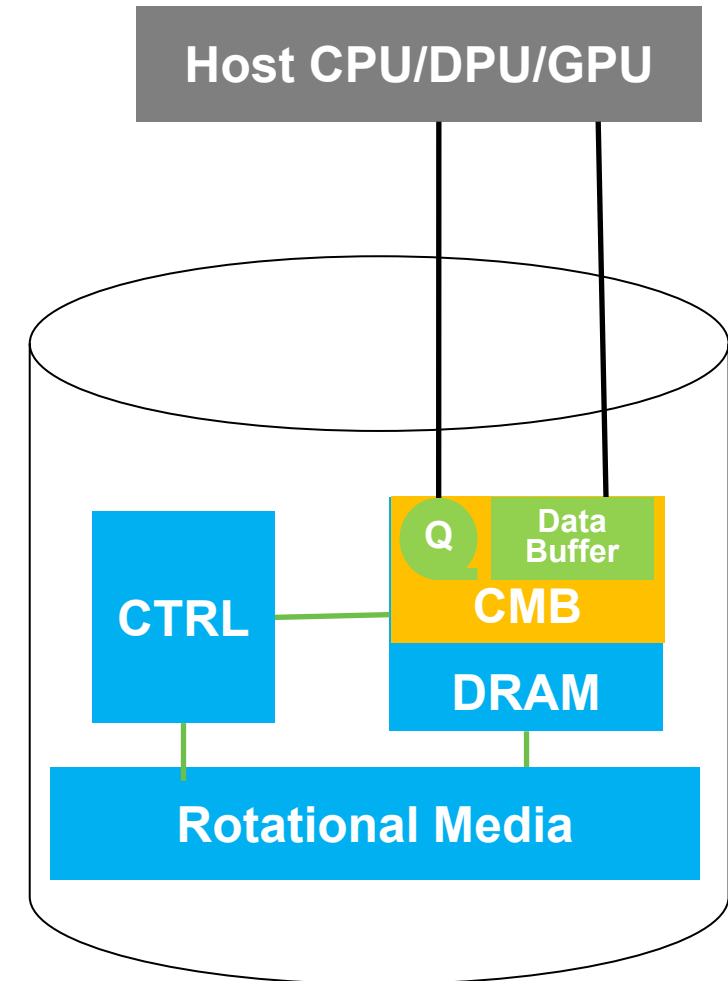
NVMe-oF Driver Target Initialization

1. Target Driver creates Queue Pairs in Target memory
2. Target Driver creates Data Buffers for IO transfer
3. Target Driver creates RDMA Keys for
 - Admin Queue and IO Queues
 - Data buffers
 - SQ Doorbell and CQ Doorbell
4. Client connect to Target via discover process
5. Client gets all the keys that are created by Target, also gets physical address of Data buffers



Controller Memory Buffer(CMB)

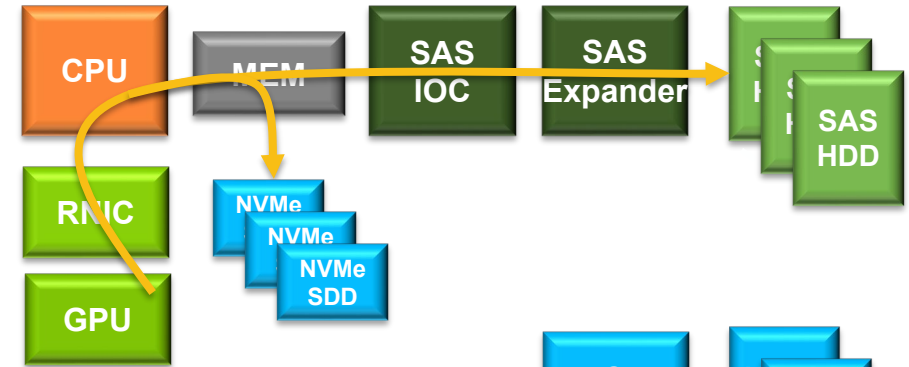
- Key Enhancements
 - Currently Supported in most RNICs
 - CMB allows commands and data to be move directly into storage controller via RNIC HW engine
 - Reduces TCO for Fabric attached Bunch of Drives(FBOD) appliances
- Implementation overview
 - HDDs require a very small CMB → 8-16MB given its lower IOPs
 - HW support to expose internal CMB DRAM to the Host PCIe bus
 - NVMe Silicon support to fetch commands/data from internal DRAM



NVMe-oF GPU-Direct HDD Topology Evolution

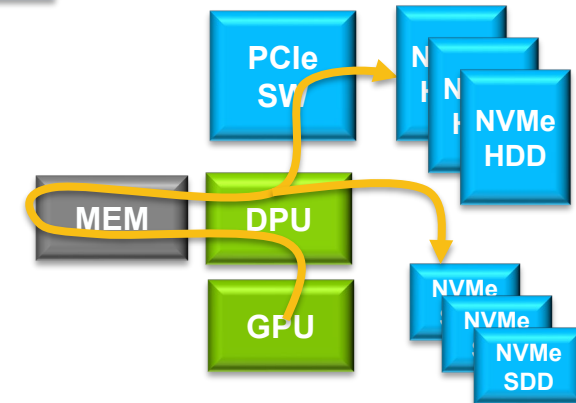
1. RNIC + CPU + SAS Enabled NVMe-oF GPU-Direct Driver:

- RNIC + CPU + Mem + SAS IOC + SAS Expander Overhead
- ~10us Fabric Latency Overhead
- Uses CPU DRAM for NVMe Queues & Data-Buffers



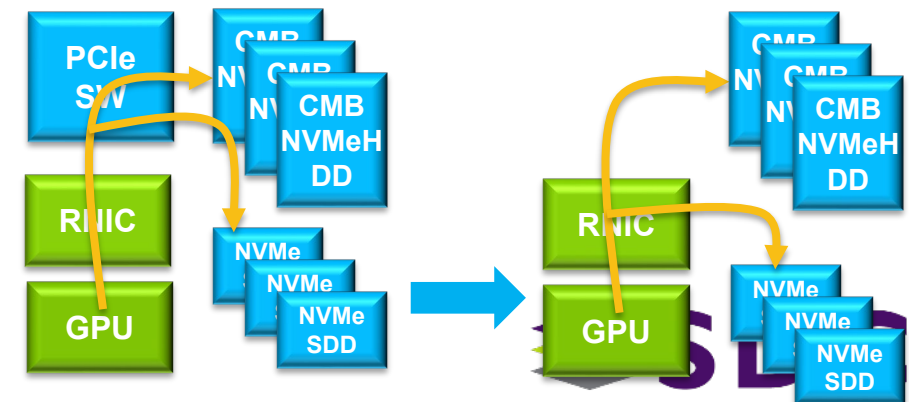
2. DPU + NVMe HDD Enabled NVMe-oF Driver:

- DPU + PCIe Switch Overhead
- ~5us Fabric Latency Overhead
- 100% CPU, DRAM & SAS Silicon offload
- Uses DPU DRAM for NVMe Queues & Data-Buffers



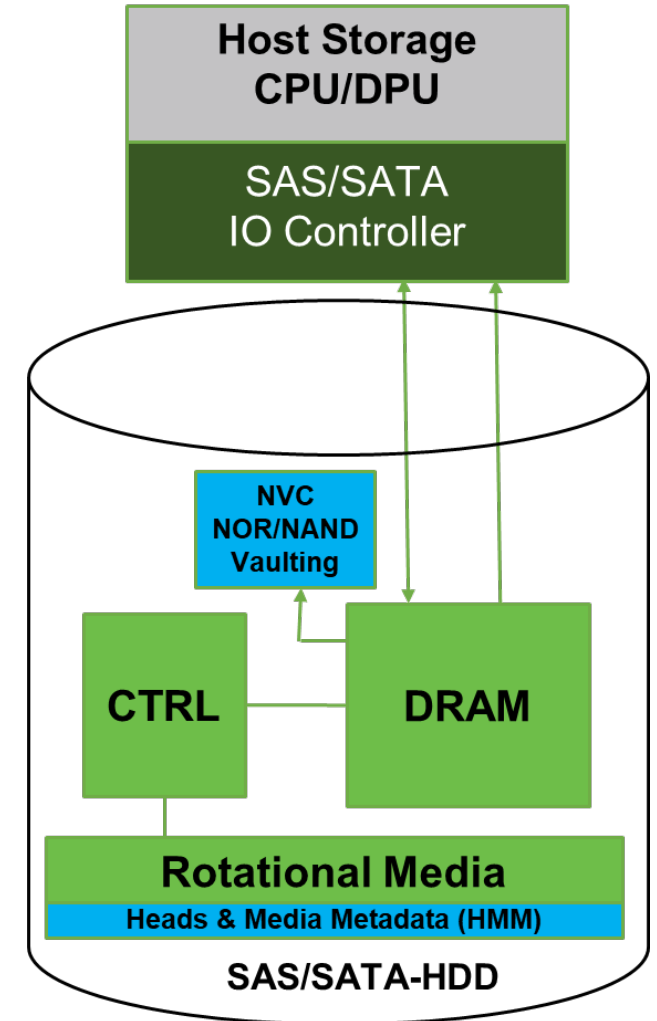
3. RNIC + NVMe HDD CMB Enabled NVMe-oF Driver:

- 2-3us Fabric Latency Overhead
- 100% CPU/DPU DRAM & SAS/Switch Silicon offload
- Use NVMe HDD CMB for NVMe Queues & Data-Buffers
- Connect HDDs & SSDs directly to RNIC and remove PCIe Switch Overhead



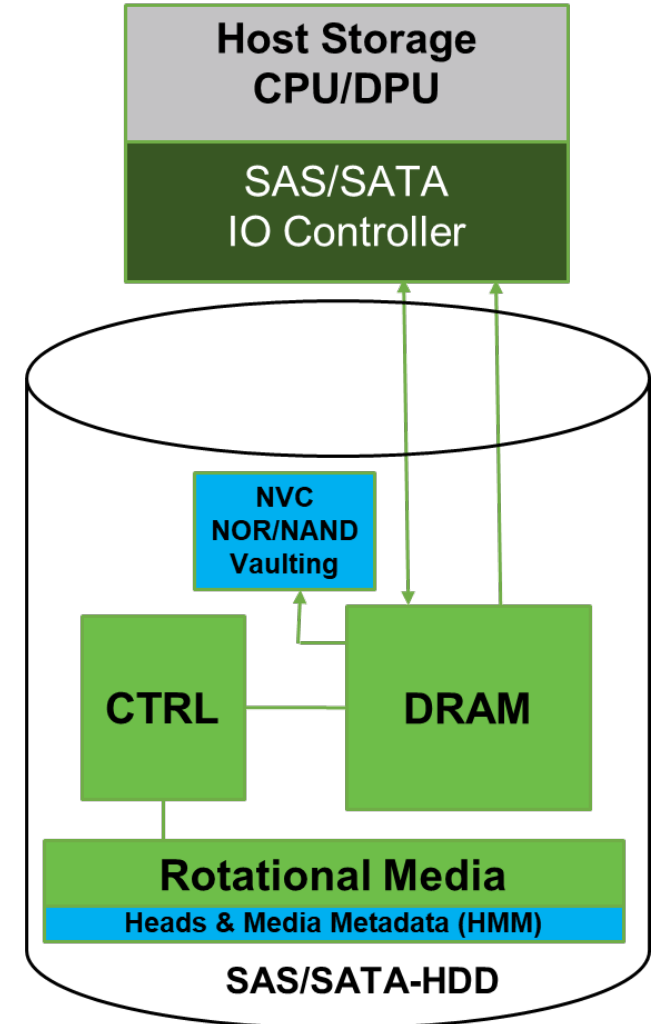
HDDs Metadata on Rotational Media

- All HDD's critical track indexing information tables and head calibration Metadata is currently stored on rotational media wasting storage space
- Before the HDD can spin-up and be ready for data transfers the drive must read all this data from rotational media
- This causes slower than desirable spin-up time at system startup and consumes customer visible data media tracks



HDDs Offer Limited NVC

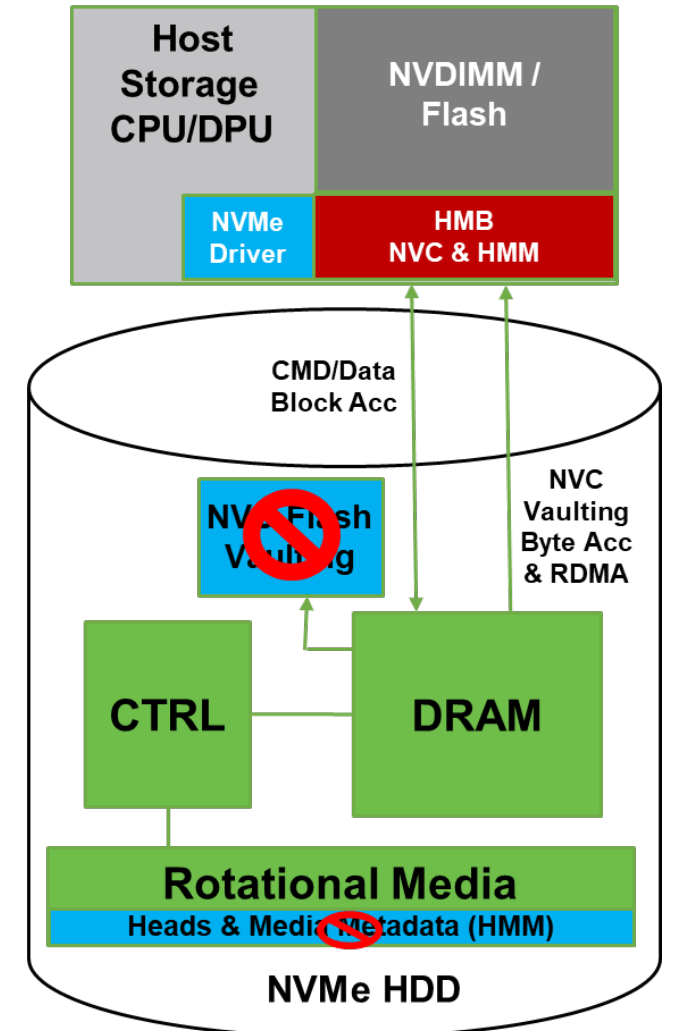
- HDDs must Data-Vault all DRAM Write Cache in NVM (Non-Volatile-Memory) upon a power loss condition, hence the term NVC (Non-Volatile-Cache)
- NVC Data-Vaulting uses Stored Rotational Energy from spinning disk platters converted into electrical power that is mostly consumed by slow NOR and NAND writes
- This limited amount of rotational Energy and the small/slow HDD NOR/NAND limit the HDD's NVC capacity thus increase latency and reduce write intensive applications performance



HMB Enables Optimal HDD Performance & TCO

Enterprise Datacenters Storage Servers in most cases BBU backed and have large DRAM capacity, Fast NVMe SSDs and/or SuperCap backed NVDIMMs

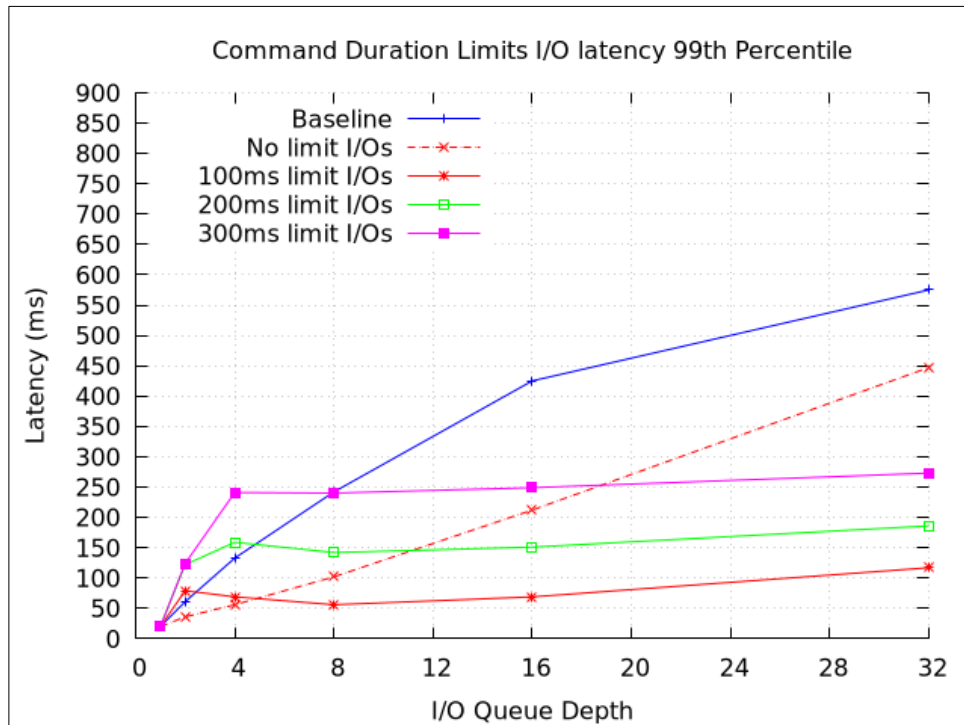
1. The NVMe Driver creates Host Memory Buffer (HMB) windows for each HDD in the storage domain within its NVM space. The NVMe Driver will then populate each unique HDD serial# Metadata onto each HDD associated HMB which enables HDDs to Spin-up
 - A. More storage capacity can be exposed to customers by freeing up Metadata sectors
 - B. The HMB resident HDD Metadata is Byte/RDMA accessible and much faster than from Rotational Media, which greatly speeds up spin-up time, so HDDs with 2 to 3 sec spin-up become possible, which could re-enable power friendly Massive Arrays of Inactive Disks (MAID) warm-storage applications in the industry
 - C. If an HDD is removed/stolen from its server, it is unreadable without its Metadata
2. The HDD NVC can be either Journaled or Single-shot Vaulted into the faster HMB partitions of these CPU/DPU NVMs instead on HDD's slower small NOR or NAND, which enables much larger HDD NVC capacity thus Lowering latency, improving performance and reducing power



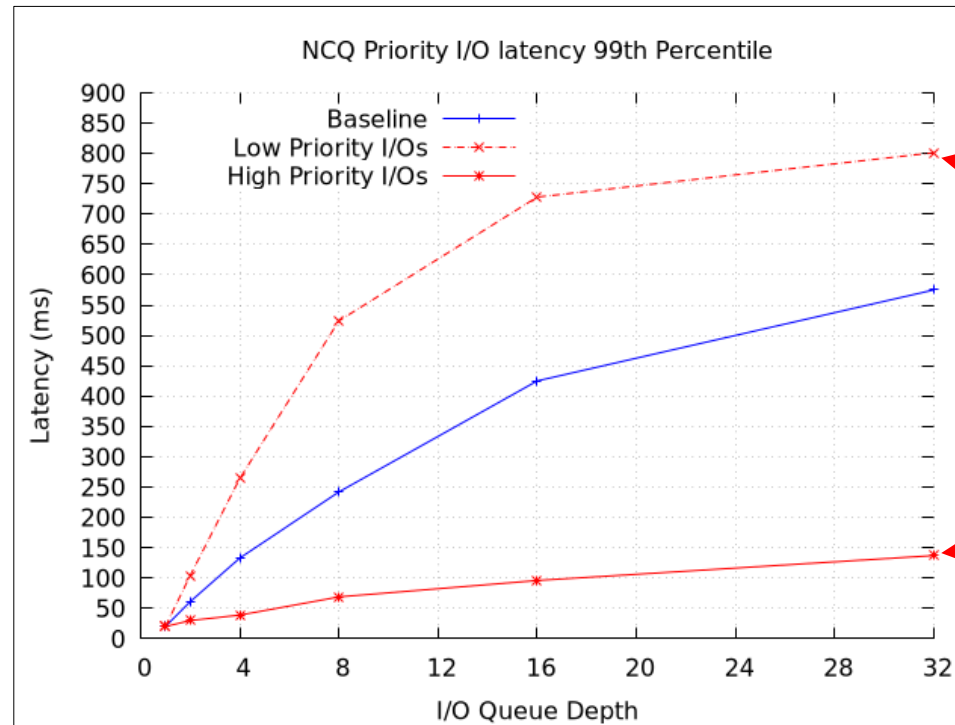
NVMe Command Duration Limits vs NCQ Priority

- CDL maintains good performance and tail latencies even for complex workloads
 - ✓ Four I/O CDL Service levels: 10% 100ms, 10% 200ms and 20% 300ms and 60% no-limit
 - ✓ With NCQ 40% of high-priority I/Os, priority overall performance degrades further (158 vs 165)

Command Duration Limits: maximum IOPS 165



NCQ Priority: maximum IOPS 158



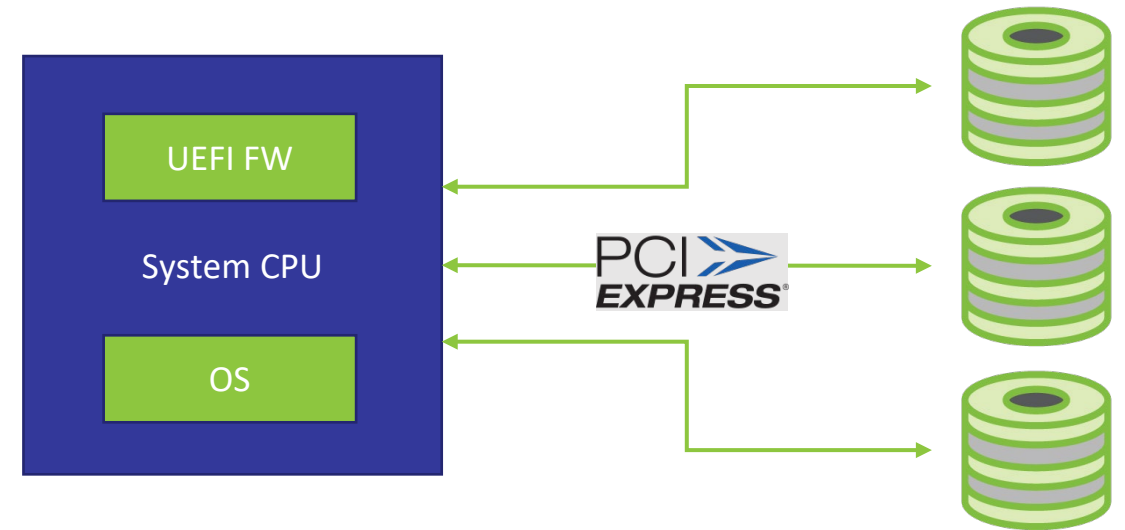
Higher tail latency for low priority commands

128KB Random Read, 40% high-priority I/Os

Higher latency at high queue depth

Direct-Attached NVMe-HDD Attestation

- Uses same model as in-data-path attestation today
 - NVMe in-band SECURITY IN/OUT
 - MCTP¹ over VDM²
 - SPDM³ over DOE⁴ (in proposal)
- Before OS boot
 - UEFI FW verifies devices
- After OS boot
 - OS Kernel verifies device
 - Hot-plugged HDDs



¹Management Component Transport Protocol

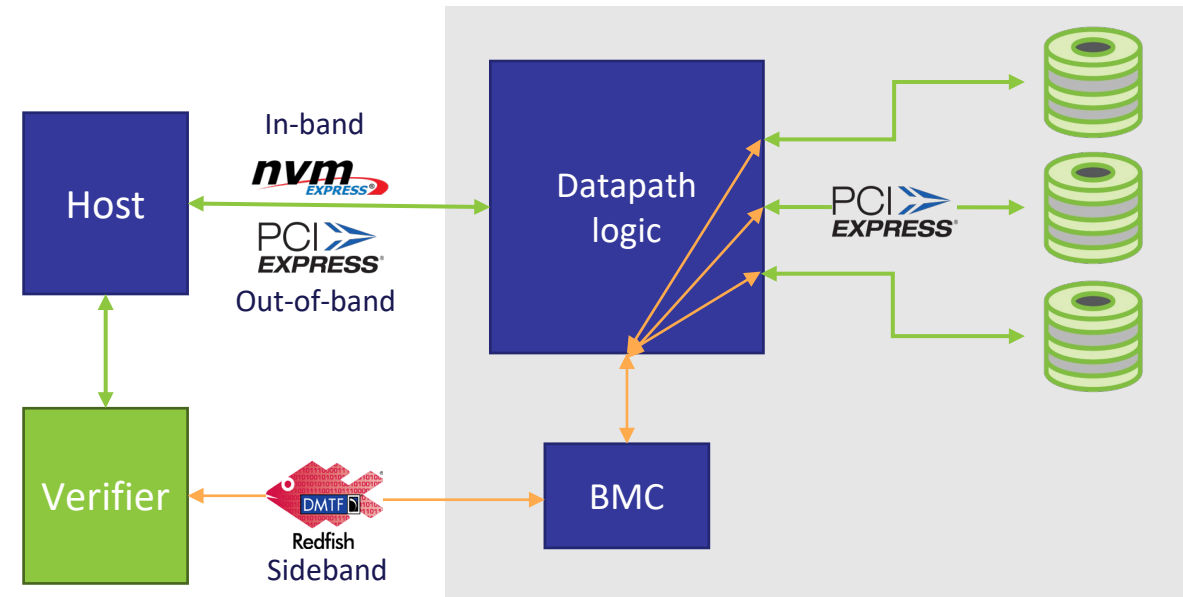
³Security Protocol and Data Model

²Vendor Defined Messages

⁴Document Object Exchange

JBOD NVMe-HDD Topologies Attestation

- **Data-path:** Storage system presents HDD to host via data-path.
 - Direct attestation performed by host using:
 - “In-band” (NVMe command set)
 - “Out-of-band” (VDM or DOE)
- **“Sideband”:** Storage system bridges attestation path to data-path
 - External verification performed at the appliance level (hierarchical) or the device level (passthrough)
 - Must be achievable using the simplest JBOF hardware components
 - BMCs for out-of-band management
 - Intelligent/non-intelligent PCIe switches
 - Storage system manages device accessibility during attestation to achieve isolation



NVMe-HDD Data-path Isolation

Manage security considerations traditionally addressed with sideband attestation over SMBus

- Ensures host is only exposed to trusted devices

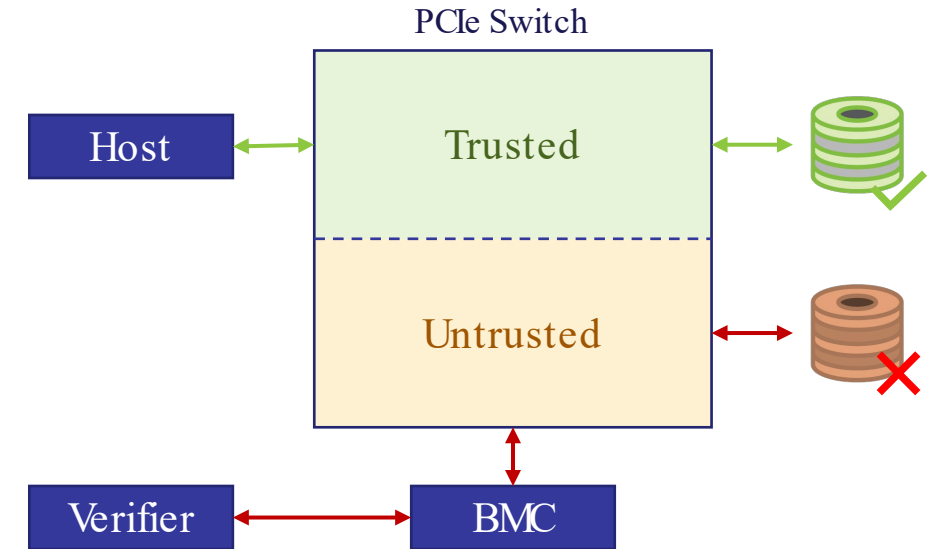
PCIe switch implements two partitions

- Trusted partition - Enumerated/managed by Host
- Untrusted partition - Enumerated/managed by BMC

Device placed into untrusted partition on insertion/power-on-reset

Verifier attests devices in untrusted partition via OOB path

- Transfers attested devices to Trusted partition
- Quarantines unattested/rogue devices in Untrusted partition



Partition Name	Permissions	
	Host Ports	BMC Port
Trusted	✓	✗
Untrusted	✗	✓

NVMe-HDD Systems Optimizations Key Takeaways

- Significant contributions from both Customers & Vendors needed before reaching our efficiency goals of: Optimized Capacity Utilization, Reduced Power, Enhanced Performance and Lower TCO
- A unified NVMe SSD & HDD Architecture enables full advantage of features like:
 - Leverage existing 3.5” storage chassis backplane/connectors and reduces IOM components
 - Data-Reduction works best with both NVMe SSDs and HDDs
 - Enable enhanced Multi-Actuator Support with Namespace (NS) existing OS support
 - Optimize Offline/Online HDD Reman/ADR with Elastic Capacity or NS Management
 - Secure Device identification with SPDM & MCTP over VDM/DOE Attestation
 - Controller Memory Buffer (CMB) enable NVMe-oF/TCP & GPU-Direct AI simplification
 - Host Memory Buffer (HMB) enhances performance & reduces power
 - Replaces the need for sNAND or iNAND with HMB space allocation
 - Better performance with larger HMB based NVC
 - Free up more capacity for user data with HMB Based Heads & Media Metadata tables
 - Faster Drive Spin-up time with HMB RDMA based Fast Metadata table read
 - HDDs would lose the ability to find their tracks when the Metadata is lost so super-secure erase
 - Host managed Zoned Namespaces (ZNS) ideal for SMR
 - Command Duration Limit (CDL) Quality-of-Service optimizations
 - Flexible Data Placement (FDP) Application friendly data placement
 - Computation storage application ready with eBPF
 - ...etc.



Please take a moment to rate this session.

Your feedback is important to us.