Powering the Data Center With NVM Express

NVM Express Webcast

June 11, 2019

Prepared by Mark Carlson & John Kim

Presenters

Mark Carlson, Principal Engineer, Toshiba Memory

John Kim, Director of Storage Marketing, Mellanox

Introduction

NVM Express™:

NVM Express is a non-profit industry organization developing an open collection of standards and information designed to fully expose the benefits of non-volatile memory in all types of computing environments from mobile to data center. is designed from the ground up to deliver high bandwidth and low latency storage access for current and future NVM technologies.

The NVMe™ set of specifications remove the bottlenecks in legacy storage infrastructure designed for hard drives, with a streamlined protocol, scalable performance, and industry standard software and drivers. NVM Express is a scalable host controller interface designed to address the needs of Enterprise, Data Center and Client systems that utilize PCI Express® (PCIe®) based solid state drives. The interface provides an optimized command issue and completion path. It includes support for parallel operation by supporting up to 64K commands within a single I/O queue to the device. Additionally, support has been added for many Enterprise capabilities like end-to-end data protection (compatible with T10 DIF and DIX standards), enhanced error reporting, and virtualization.

This talk

Powering the Data Center with NVM Express™
The NVMe™ standard is evolving to support data center deployment of SSDs. Hyperscalers have brought their requirements into NVM Express regarding isolation, predictable latency and write amplification. As a result, new features such as I/O determinism and predictable latency has been added to the upcoming NVMe 1.4 specification to address some of these concerns, and NVMe-oF has added support for TCP as a transport. In this NVM Express, Inc. hosted webinar, Mark Carlson of Toshiba and John Kim of Mellanox will cover the issues and show the new features from multiple data center perspectives.

Data Center Customers

When we say “data center” we mean any organization that is building data centers and populating them with modern, scale out components and systems.

- Hyperscalers
- Tier two data center customers (applying Hyperscaler techniques)

They specify components and have ODMs assemble the solutions for the next Datacenter

- OCP often used to develop then open source the hardware designs
- They all have unique requirements (for SSDs, etc.) but there are common requirements for which standards can play a role
- For drive vendors, the more that is standardized, the less custom code

Hyperscale – Most Common Use Cases

Boot and Log

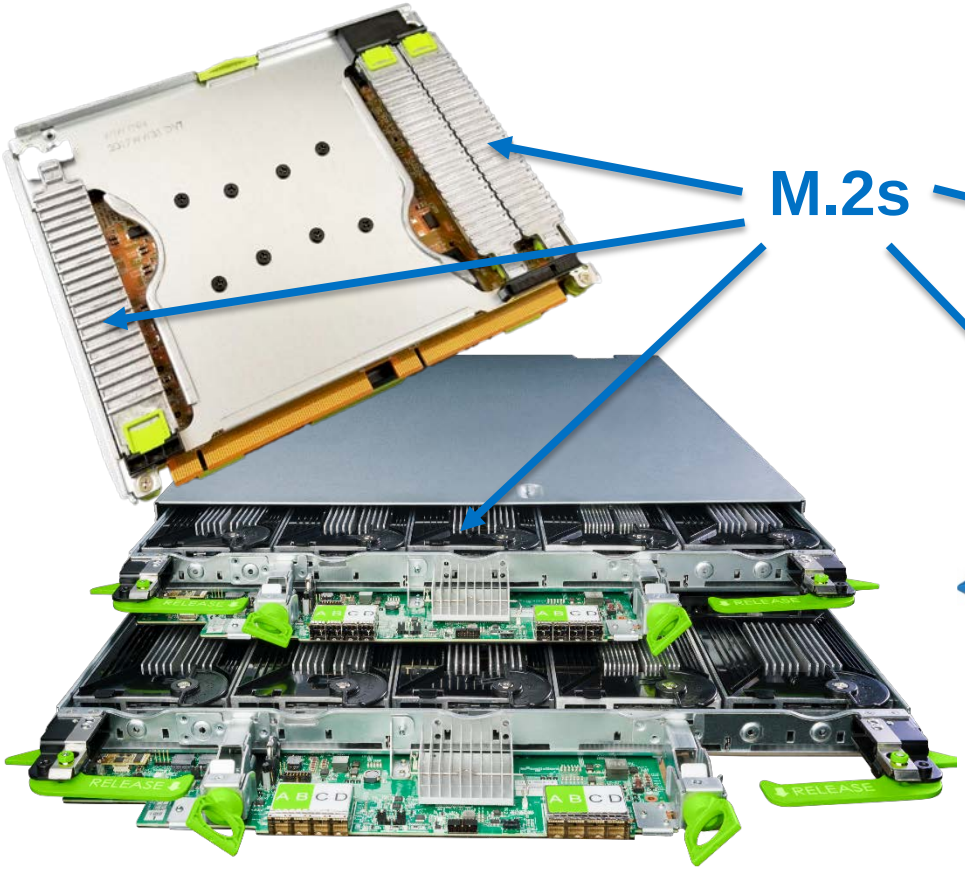
- OS boot drive
- OS and application logs

Databases

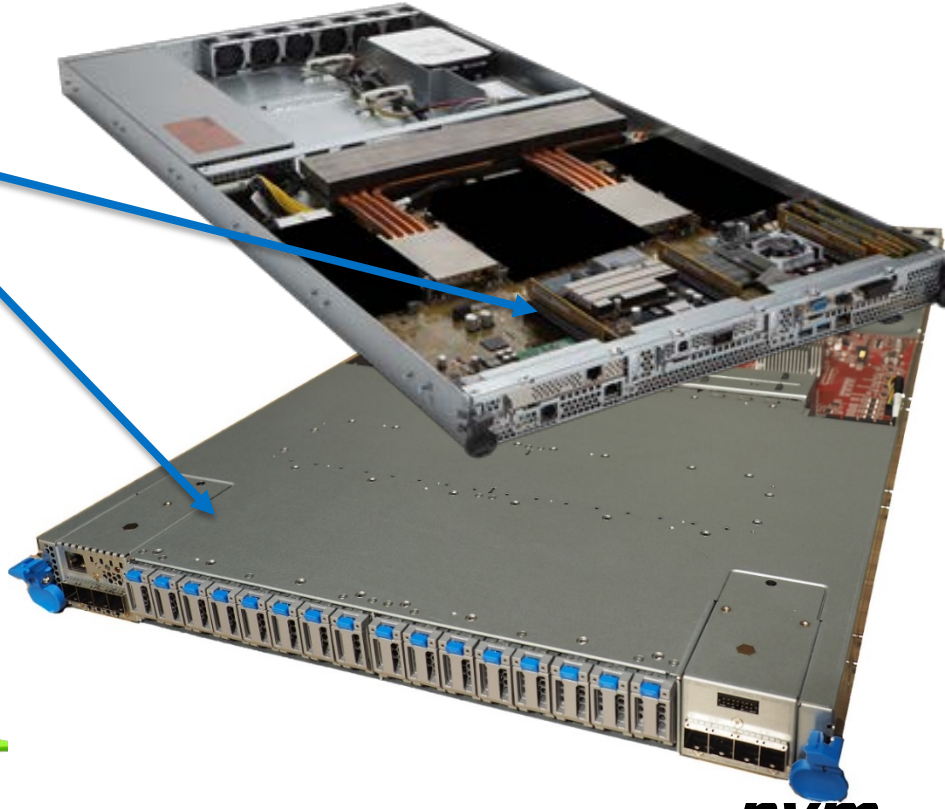
Cache

- Content caching
- Object caching
- Indexing

Where do Hyperscalers Use Flash Today?



M.2s



M.2 Carriers

Hyperscale NVM Characteristics and Challenges

Important:

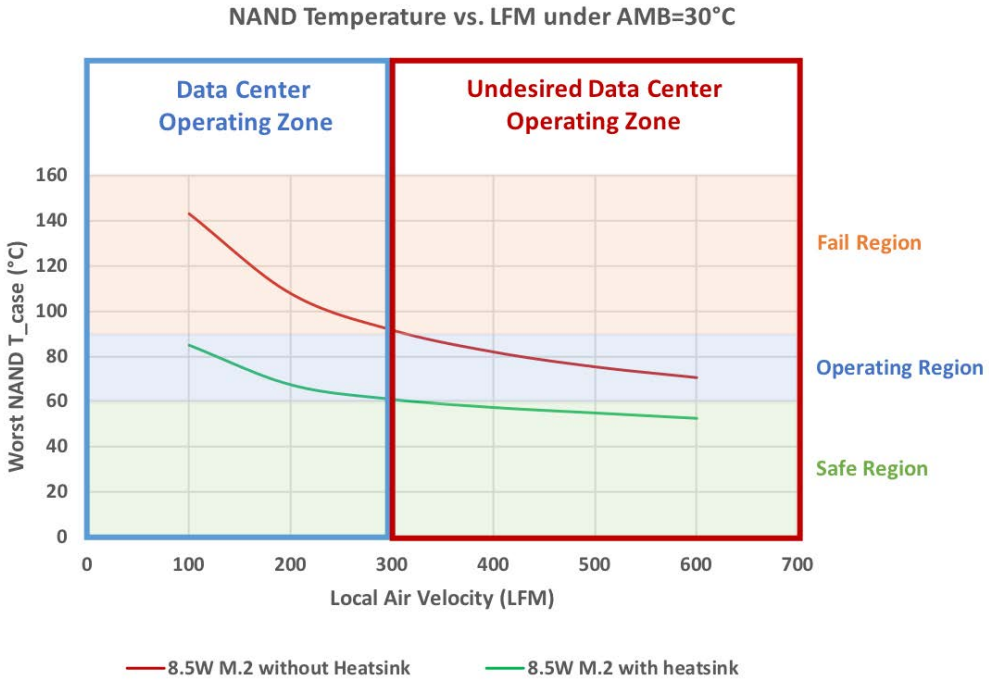
- Scalable & Flexible
- High volume & Low cost
- Power & Thermal Efficiency
- Hot-swappable & Serviceable
- Performance per TB & Quality of Service (low long-tail latency)

Less important:

- Backwards compatible
- Support for non-NVM media
- Maximum density
- Peak peformance (peak IOPs/BW)

Hyperscale Efficiency

Power and thermal efficiency are critical

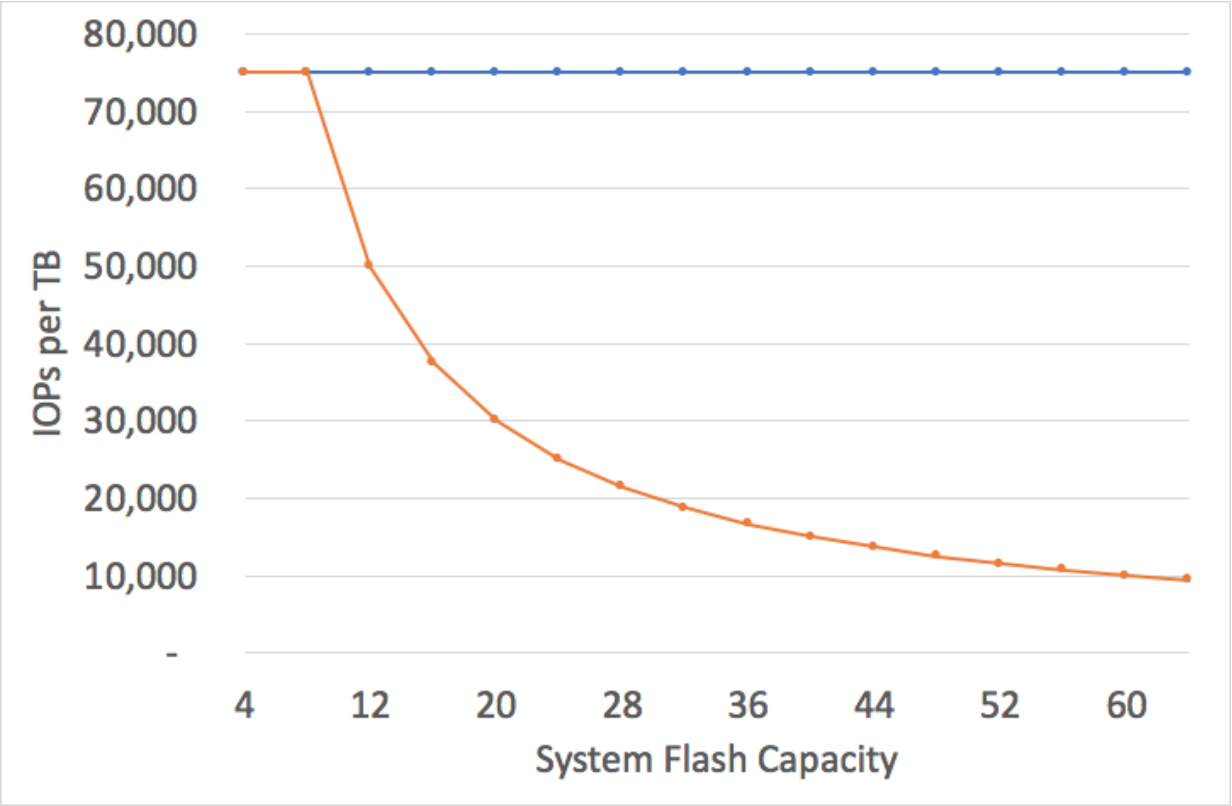


- Limited airflow and power is available in data centers
- Temperature increase across servers is large (delta T)
- OPEX matters (cooling is expensive)

NVMe™-MI enables effective thermal management!

Scalable Performance

IOPs scales with capacity



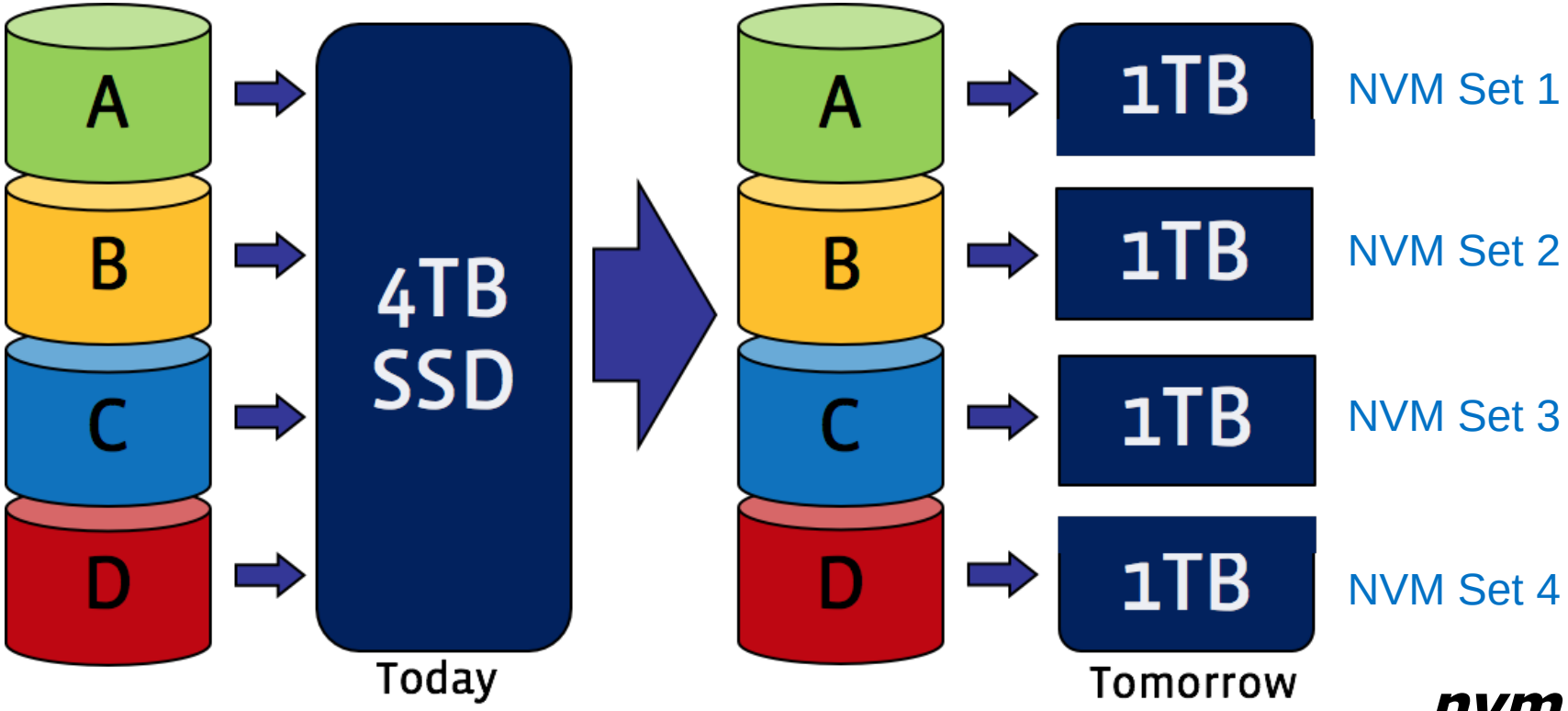
NVMe™ SSDs

SATA SSDs

*Basic assumptions: 4TB SSDs @ 300k 4k IOPs and 600k IOPs SATA limitation

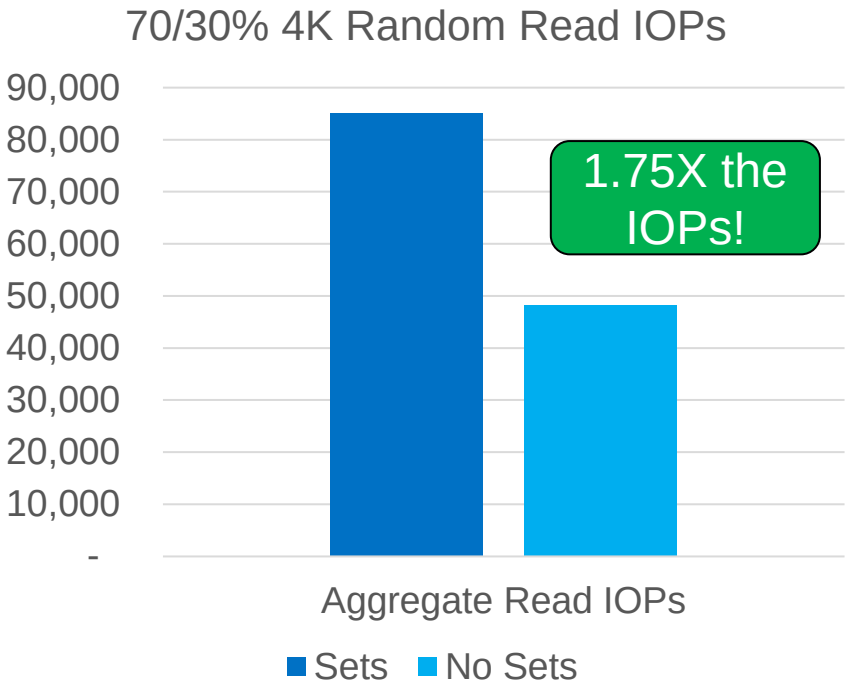
Scalable Performance

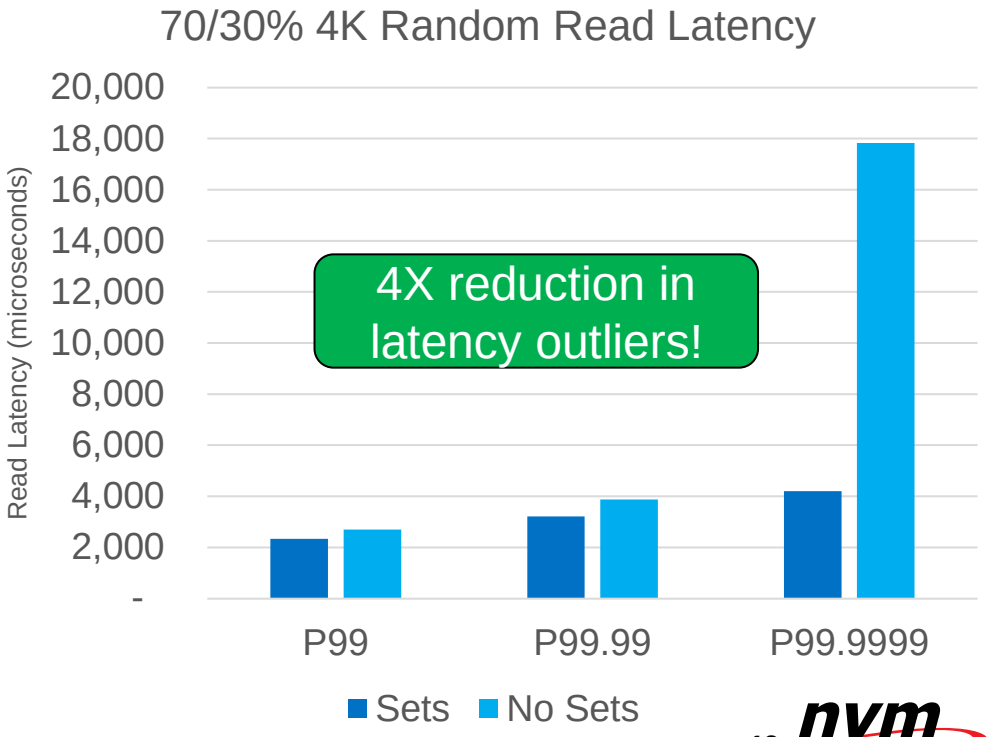
NVMe™ I/O Determinism



Scalable Performance

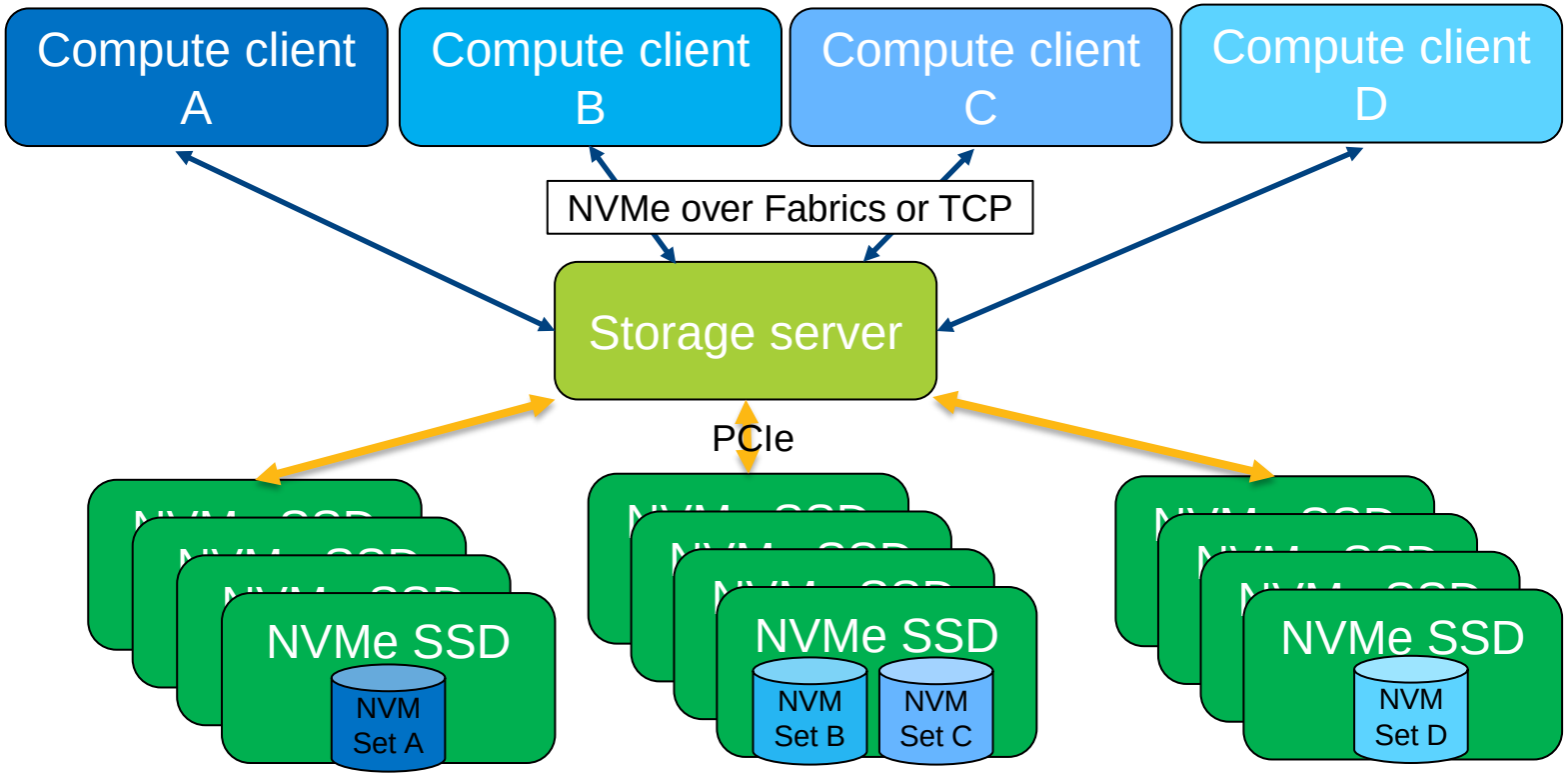
NVMe™ I/O Determinism





Scalable Performance

NVMe™ provides fabric connectivity



Common Requirements

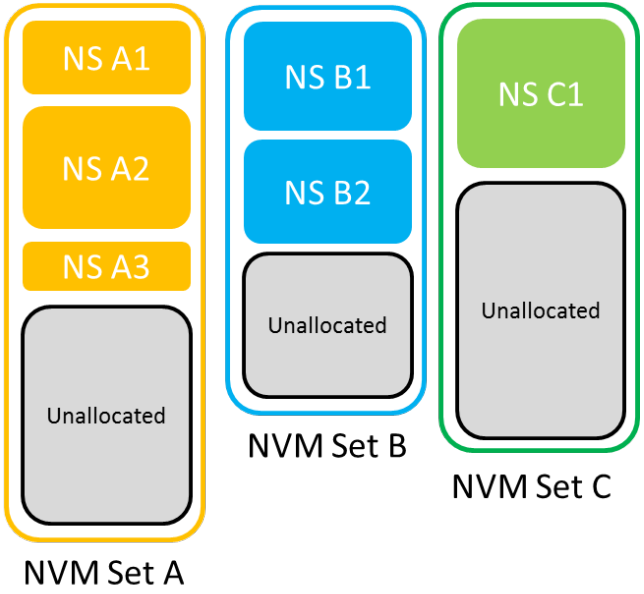
- **Tail Latency** – becomes a huge problem at scale
  - Need predictable performance
  - Need to fail fast rather than heroic recovery
- **Isolation** – drives are becoming too large and the needed multiple applications (for capacity utilization) interfere with each other
- **Wear** – better coordinate between drive and host to reduce write amplification
  - More control over the FTL (e.g. wear leveling)
  - Which are done on the host vs. which are done on the drive
- **TCP** for storage networking (no special networking hardware, scales farther)

I/O Determinism

- Multiple Hyperscalers came to NVMe™ Board of Directors with an I/O Determinism proposal
- After two years of work, the resulting Ratified Technical Proposals (against NVMe 1.3) include:
  - NVMe - TP 4003b IO Determinism 2018.09.17 -Ratified.pdf
  - NVMe - TP 4018a NVM Sets and Read Recovery Level 2018.07.23 - Ratified.pdf
- These TPs are integrated into the upcoming 1.4 version

NVM Sets

- An NVM Set is a collection of NVM with particular attributes from which one or more namespaces may be allocated
- An NVM subsystem supports one or more NVM Sets, discoverable in Identify
- NVM Sets are isolated from each other in regards to quality of service for IO commands
  - E.g., A write command to NS A1 should not measurably impact the quality of service for a read command to NS B1
- NVM Set creation/deletion is outside the scope of the current work



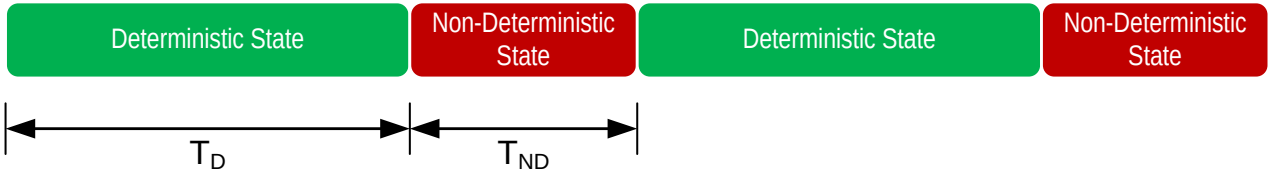
Read Recovery Level

- Read Recovery Level trades off the maximum time for a read before the SSD gives up on getting the data versus the unrecoverable bit error rate (UBER) that is warrantied
  - The host may make this tradeoff when the data is available in multiple locations
- The maximum time for a 4KB read is measured at queue depth = 1 with all other NVM Sets in an idle condition
- Level 0 is the “Fail Fast” –SSD fails at first sign of issues reading the data
- Only the Read Recovery Level that as been selected is reported as an attribute (UBER is not).

Read Recovery Level	QD1 P99 Read Max (examples)	QD1 P9999 Read Max (examples)	UBER (examples)
0 – “Fail Fast”	50 µs	500 µs	1e-14
1	100 µs	1 ms	1e-16
2	200 µs	2 ms	1e-17
3	500 µs	5 ms	1e-18

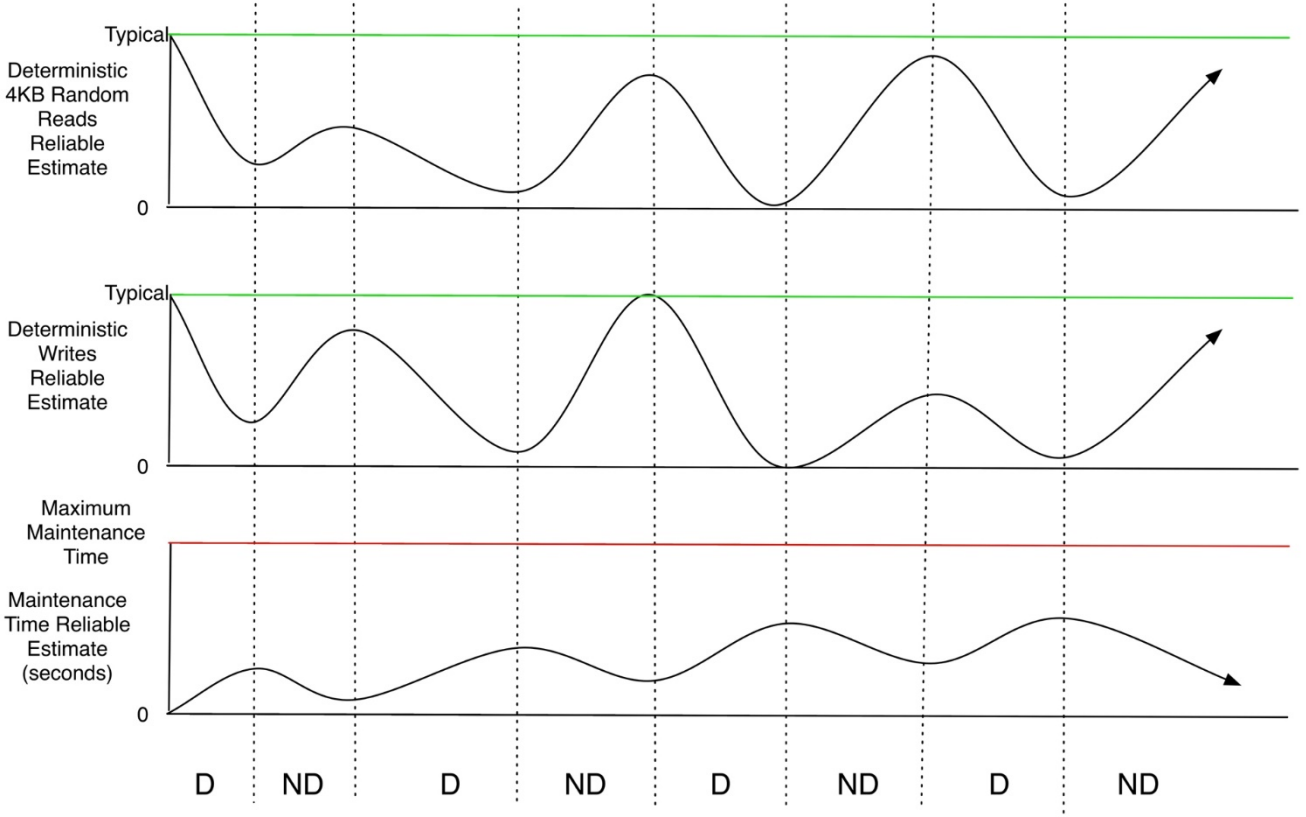
Note: Above table is for reference only and contains an example of values.

I/O Determinism States



- T_D – Time in deterministic state
- T_{ND} – Time in non-deterministic state
- The controller provides reliable estimates to guide transition into and out of deterministic state

Correlated


Typical
Deterministic 4KB Random Reads Reliable Estimate
0
Typical
Deterministic Writes Reliable Estimate
0
Maximum Maintenance Time
Maintenance Time Reliable Estimate (seconds)
0
D
ND
D
ND
D
ND
D
ND

NVMe™ Meeting Common Requirements

- ✔ **Tail Latency – becomes a huge problem at scale**
  - **Need predictable performance**
  - **Need to fail fast rather than heroic recovery**
- ✔ **Isolation – drives are becoming too large and the needed multiple applications (for capacity utilization) interfere with each other**
- Wear – better coordinate between drive and host to reduce write amplification
- More control over which parts of the FTL (e.g. wear leveling) are done on the host and which are done on the drive
- TCP for storage networking (no special networking hardware, scales farther)

Endurance Group Management

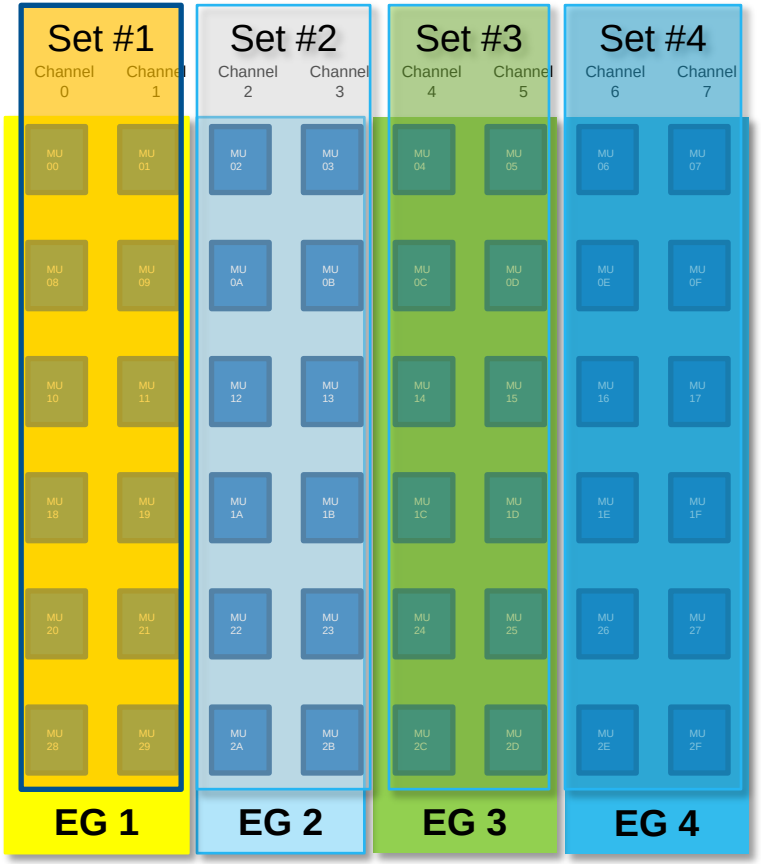
- Two methods:
  - Media Unit Endurance Group Management
  - Capacity Endurance Group Management
- Media Unit Endurance Group Management would be used for drives.
  - The operation will select from a fixed set of complete configurations; the selected configuration typically will be for the lifetime of the NVM subsystem.
  - This should satisfy the requirements of Hyperscalers.
    - Incrementally configuring endurance groups / NVM sets will not be supported for this method (not needed). Changing the configuration after the media has been used will not be supported in this method. If a use case for this is found, it could be the basis for a future TP.

Capacity Endurance Group Management

- Capacity Endurance Group Management is for systems to dynamically create Endurance Groups and NVM Sets. The operation specifies a capacity for endurance groups and NVM sets without understanding of the underlying media units.

IOD users

- Need to create NVM Sets according to their requirements (e.g., 1TB sets)
  - Once at the beginning of drive life
  - Supported Media Unit configurations are available indicating NVM Sets formed from Media Units along channels for isolation
  - Drive may only support two configurations (e.g. ½ TB and 1 TB sets) for this market



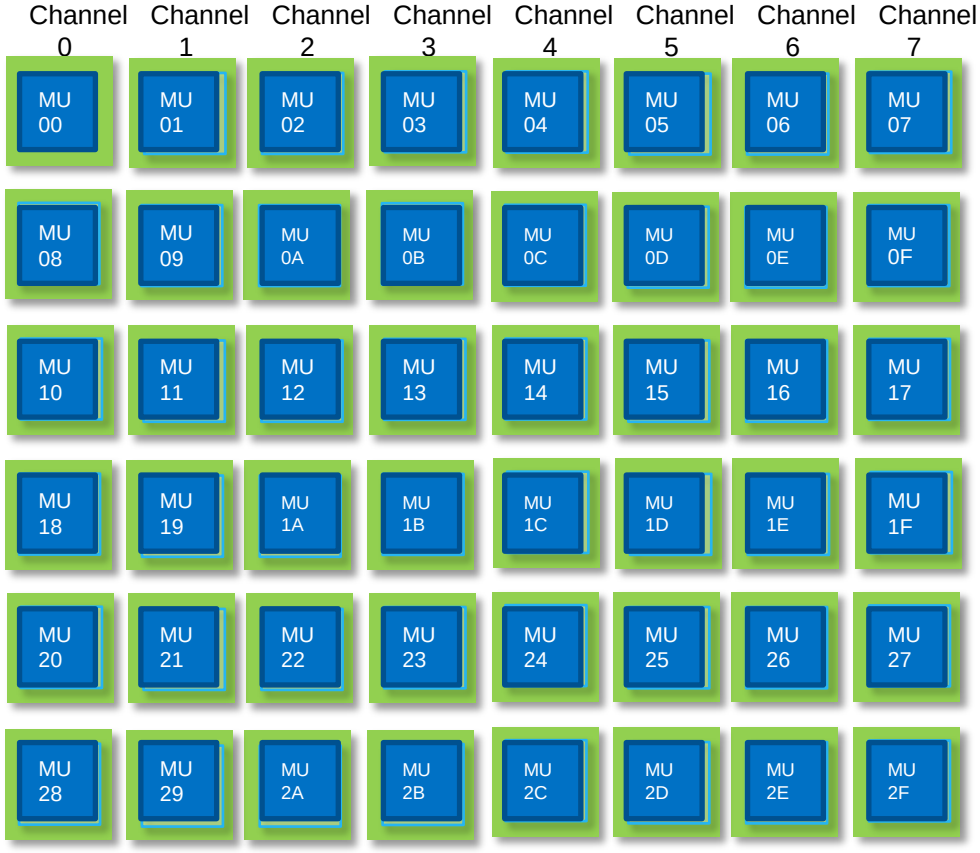
Get Log Page – Media Unit Status

Host Managed Media Users

- Need to closely manage placement of data and accommodate append behavior
  - Big concerns about Write Amplification and managing wear

1 EG / 1 NVM Set / 1 MU

- No predictable latency
- RAW UBER
- 1 Namespace / MU



Get Log Page – Media Unit Status

Media Unit Status Descriptor

Bytes	Description
01:00	**Media Unit Identifier**: This field indicates the identifier of the Media Unit in the domain that contains the controller processing the command.
03:02	Reserved
05:04	**Endurance Group Identifier**: This field indicates the Endurance Group for this Media Unit. Refer to section TP4018_4.TBD. A value of zero indicates that this Media Unit is not part of an Endurance Group.
07:06	**NVM Set Identifier**: This field indicates the NVM Set for this Media Unit. Refer to section TP4018_4.TBD. A value of zero indicates that this Media Unit is not part of an NVM Set.
09:08	**Channel Identifier**: This field indicates the Channel that the Media Unit is accessed through. A value of 0h indicates that the Channel Identifier is not specified.
11:10	**Capacity Adjustment Factor**: This field indicates the ratio between the capacity obtained with this Media Type and the capacity obtained with TBD . If this field is cleared to 0h, then the capacity adjustment factor is not specified. Refer to Figure Fig_MUAsgtD.
12	**Available Spare**: Contains a normalized percentage (0 to 100%) of the remaining spare capacity available for the Media Unit.
13	**Percentage Used**: Contains a vendor specific estimate of the percentage of life used for the Media Unit based on the actual usage and the manufacturer's prediction of NVM life. A value of 100 indicates that the estimated endurance of the NVM in the Media Unit has been consumed, but may not indicate an NVM failure. The value is allowed to exceed 100. Percentages greater than 254 shall be represented as 255. This value shall be updated once per power-on hour when the controller is not in a sleep state. Refer to the JEDEC JESD218A standard for SSD device life and endurance measurement techniques.
127:14	Reserved

Media Unit Descriptor

Bytes	Description
1:0	**Media Unit Identifier**: This field indicates the identifier of the Media Unit in the domain that contains the controller processing the command.
3:2	Reserved
5:4	**Endurance Group Identifier**: This field indicates the Endurance Group to which this Media Unit is assigned. Refer to section TP4018_4.TBD. A value of zero indicates that this Media Unit is not assigned to an Endurance Group.
7:6	**NVM Set Identifier**: This field indicates the NVM Set to which this Media Unit is assigned. Refer to section TP4018_4.TBD. A value of zero indicates that this Media Unit is not assigned to an NVM Set.
9:8	**Capacity Adjustment Factor**: This field indicates the ratio between the capacity obtained with this Media Type and the capacity obtained with TBD . A value of 0h indicates that the Capacity Adjustment Factor is not specified.
15:10	Reserved

Storage Systems Users

- Need to create, resize and delete Endurance Groups within a Domain
  - No need to implement Media Unit Management
  - Primarily tied to TP 4009 (Domains) – but are proceeding in parallel
- Capacity Endurance Group Management
  - Capacity is drawn from the Domain
  - Separate operations to create Endurance Groups and NVM Sets
  - Deletion of Endurance Group also Deletes NVM Set(s), Namespace(s)

Endurance Group Management – Command Dword 10

Bits	Description		
31:16	Object Identifier: This field contains a value specific to the value of the Operation field.		
15:04	Reserved		
03:00	Operation : Specifies the operation to be performed by the controller:		
	Value	**Description**	**Object Identifier**
	0h	**Select Media Unit Configuration:** Endurance Groups are configured as indicated by the Media Unit Configuration Descriptor specified by this command.	Media Unit Configuration Identifier (refer to Figure Fig_MUConfD).
	1h	**Release Media Unit Configuration:** All namespaces, NVM Sets, and Endurance groups are deleted. [initial definition]	Reserved
	2h	**Create Endurance Group:** An Endurance Group is created with the capacity specified by Command Dword 11 (refer to Figure Fig_EGMC11) and Command Dword 12 (refer to Figure Fig_EGMC12).	Reserved
	3h	**Delete Endurance Group:** The Endurance Group specified by this command is deleted. All namespaces and NVM Sets contained by the Endurance Group shall be deleted.	Endurance Group identifier of the Endurance Group to be deleted
	4h	**Create NVM Set:** An NVM Set is created with the capacity specified by Command Dword 11 and Command Dword 12.	Endurance Group identifier of the Endurance Group in which the NVM Set is to be created.
	5h	**Delete NVM Set:** The NVM Set specified by this command is deleted. All namespaces in the NVM Set are deleted.	NVM Set identifier of the NVM Set to be deleted
	6h to Fh	Reserved	

NVMe™ Meeting Common Requirements

- ✔ Tail Latency – becomes a huge problem at scale
  - Need predictable performance
  - Need to fail fast rather than heroic recovery
- ✔ Isolation – drives are becoming too large and the needed multiple applications (for capacity utilization) interfere with each other
- ✔ **Wear – better coordinate between drive and host to reduce write amplification**
- ✔ **More control over which parts of the FTL (e.g. wear leveling) are done on the host and which are done on the drive**
- TCP for storage networking (no special networking hardware, scales farther)

NVMe™-oF Using TCP

- NVMe™ over Fabrics increases flexibility, efficiency of NVMe deployments
- Some customers don’t want a special fabric
  - Don’t want to build separate Fibre Channel SAN or InfiniBand network
  - Don’t want to deploy rNICs or configure switches for RDMA
- Want NVMe-oF solution that runs on existing networks
  - Doesn’t require any changes to switch configurations
  - Doesn’t require building a new network or changing adapters
- TCP scales farther

NVMe™-oF Fabric Options

Fabric and Transport	Features	Special Network Requirements	Notes
Fibre Channel	Fast, Lossless	Separate network	Mostly in enterprise; rare in hyperscalers
InfiniBand	Fast, Lossless	Separate network	Mostly in specialized networks for HPC, AI/ML
RoCE – Ethernet	Fast, Flexible	rNICs, switch settings	Requires rNICs; may require switch reconfig
TCP – Ethernet	Ubiquitous	None (low latency nice but not required)	Uses standard Ethernet adapters and switches

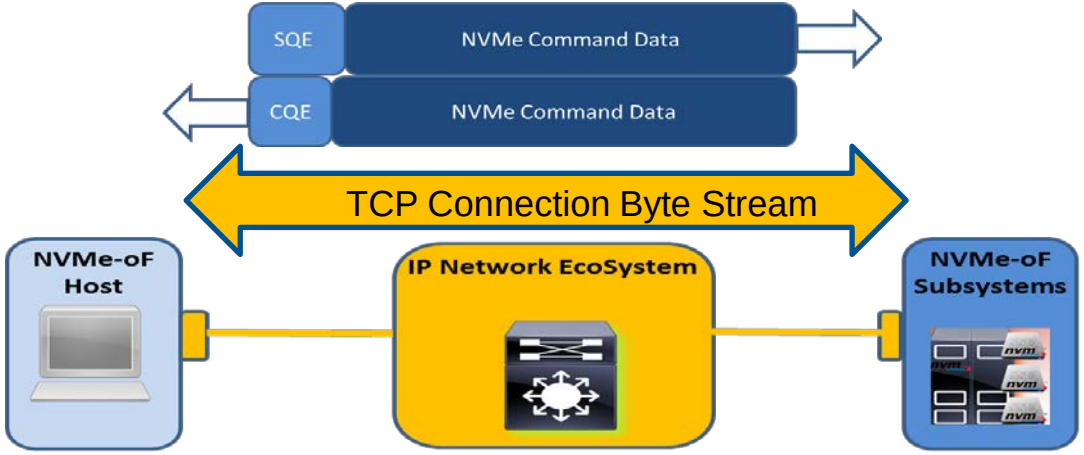
NVMe™-TCP Transport Basics

NVMe-TCP is constructed over the IETF TCP Transport

- TCP provides a reliable ordered byte stream between two IP endpoints
- Same IP transport used by iSCSI

NVMe-TCP defines methods used to exchange NVMe-oF Capsules and Command Data over TCP byte stream connections

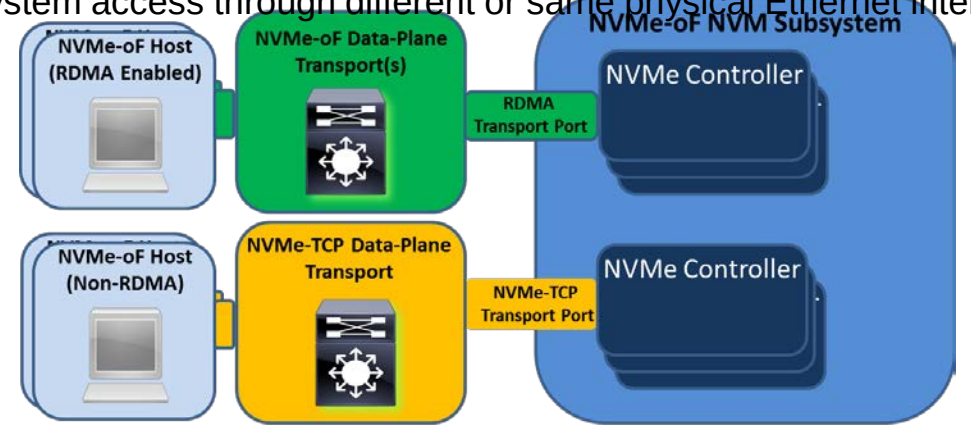
- Provides both Admin and I/O Queue operations



NVMe™-TCP Data-Plane Usage

- NVMe™-TCP: End-to-end NVMe I/O Queue operations on IP Datacenter networks
  - NVMe-oF Command and Response Capsule and Command Data Exchanges
  - NVMe-based alternative to iSCSI for IP block storage

- NVMe-oF Subsystem end-points may provide NVMe over both RDMA and TCP
  - Enables RDMA and non-RDMA hosts to establish Controller Associations
  - Host transport selection may be based on physical connectivity or policy
  - Subsystem access through different or same physical Ethernet interface



TCP (NVMe TP 8000) modifies NVMe™-oF 1.0

- Allows NVMe™-oF to run on TCP
- Uses existing switch configurations
- Customers accept potentially slower performance
- Network could be optimized for storage performance, but not required
- Customers can support multiple fabric types if desired

NVMe™ Meeting Common Requirements

- ✔ Tail Latency – becomes a huge problem at scale
  - Need predictable performance
  - Need to fail fast rather than heroic recovery
- ✔ Isolation – drives are becoming too large and the needed multiple applications (for capacity utilization) interfere with each other
- ✔ Wear – better coordinate between drive and host to reduce write amplification
- ✔ More control over which parts of the FTL (e.g. wear leveling) are done on the host and which are done on the drive
- **✔ TCP for storage networking (no special networking hardware, scales farther)**

Summary

As we have shown, NVM Express™ continues to be responsive to Hyperscaler requirements

New NVMe™ Features are making their way into products

NVMe testing/certification for these features will be shortly added

Attend Flash Memory Summit and SDC for more details on individual features

Contact: nvme@nereus-worldwide.com