NVMe™: Hardware Implementations and Key Benefits in environments

Sponsored by NVM Express® organization, the owner of NVMe™, NVMe-oF™ and NVMe-MI™ standards

NVMe™ A-11b Track Speakers

Chris Petersen

Chander Chadha

Jonmichael Hands

NVMe™ Agenda

Hyperscale Challenges and NVMe Solutions

NVMe for Data Center Enterprise Needs

NVMe Client Implementations

Q&A

NVM Express® Sponsored Track for Flash Memory Summit 2018

	Track	Title	Speakers	
NVMe-101-1	8/7/18 8:30-9:35	NVM Express: NVM Express roadmaps and market data for NVMe, NVMe-oF, and NVMe-MI - what you need to know for the next year.	Janene Ellefson, Micron J Metz, Cisco	Amber Huffman, Intel David Allen, Seagate
	8/7/18 9:45-10:50	NVMe architectures for in Hyperscale Data Centers, Enterprise Data Centers, and in the Client and Laptop space.	Janene Ellefson, Micron Chris Peterson, Facebook	Chander Chadha, Toshiba Jonmichael Hands, Intel
NVMe-102-1	3:40-4:45 8/7/18	NVMe Drivers and Software: This session will cover the software and drivers required for NVMe-MI, NVMe, NVMe-oF and support from the top operating systems.	Uma Parepalli, Cavium Austin Bolen, Dell EMC Myron Loewen, Intel Lee Prewitt, Microsoft	Suds Jain, VMware David Minturn, Intel James Harris, Intel
	4:55-6:00 8/7/18	NVMe-oF Transports: We will cover for NVMe over Fibre Channel, NVMe over RDMA, and NVMe over TCP.	Brandon Hoff, Emulex Fazil Osman, Broadcom J Metz, Cisco	Curt Beckmann, Brocade Praveen Midha, Marvell
NVMe-201-1	8/8/18 8:30-9:35	NVMe-oF Enterprise Arrays: NVMe-oF and NVMe is improving the performance of classic storage arrays, a multi-billion dollar market.	Brandon Hoff, Emulex Michael Peppers, NetApp Clod Barrera, IBM	Fred Night, NetApp Brent Yardley, IBM
	8/8/18 9:45-10:50	NVMe-oF Appliances: We will discuss solutions that deliver high-performance and low-latency NVMe storage to automated orchestration-managed clouds.	Jeremy Werner, Toshiba Manoj Wadekar, eBay Kamal Hyder, Toshiba	Nishant Lodha, Marvell Yaniv Romem, CTO, Excelero
NVMe-202-1	8/8/18 3:20-4:25	NVMe-oF JBOFs: Replacing DAS storage with Composable Infrastructure (disaggregated storage), based on JBOFs as the storage target.	Bryan Cowger, Kazan Networks	Praveen Midha, Marvell Fazil Osman, Broadcom
	8/8/18 4:40-6:45	Testing and Interoperability: This session will cover testing for Conformance, Interoperability, Resilience/error injection testing to ensure interoperable solutions base on NVM Express solutions.	Brandon Hoff, Emulex Tim Sheehan, IOL Mark Jones, FCIA	Jason Rusch, Viavi Nick Kriczky, Teledyne

Follow NVMe™

nvmexpress.org

@NVMExpress

Hyperscale Challenges and NVMe® Solutions

Chris Petersen, Facebook

Hyperscale use cases

Boot and Log

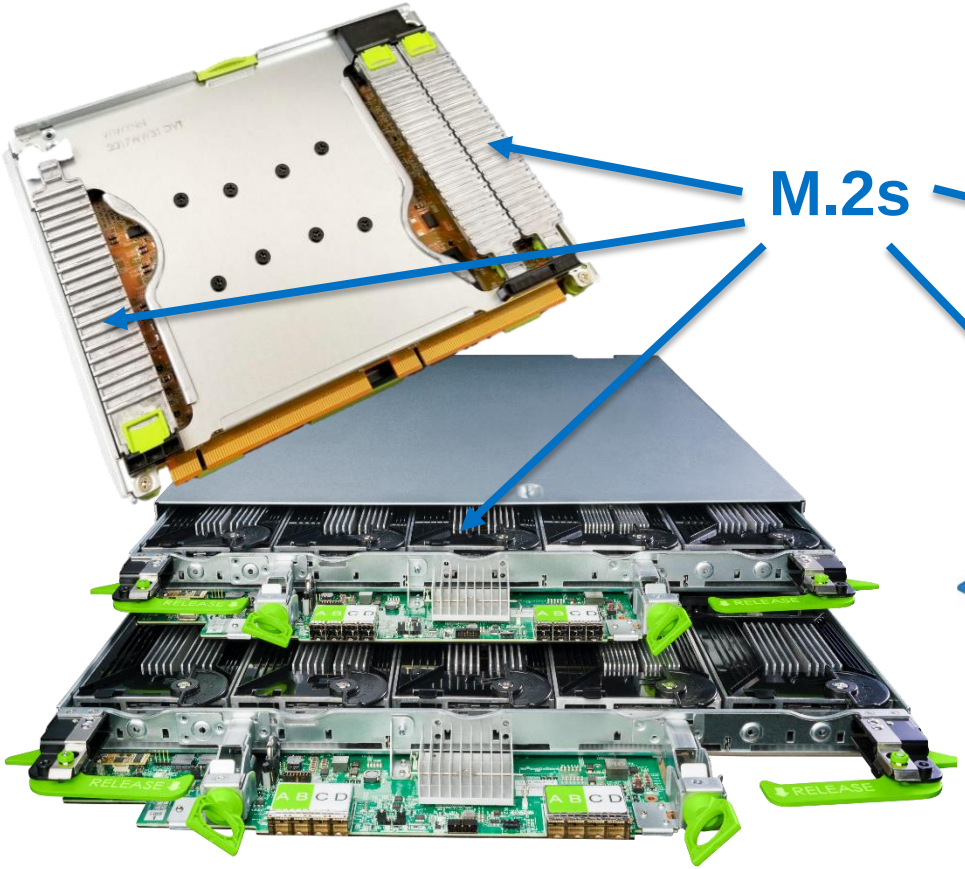
- OS boot drive
- OS and application logs

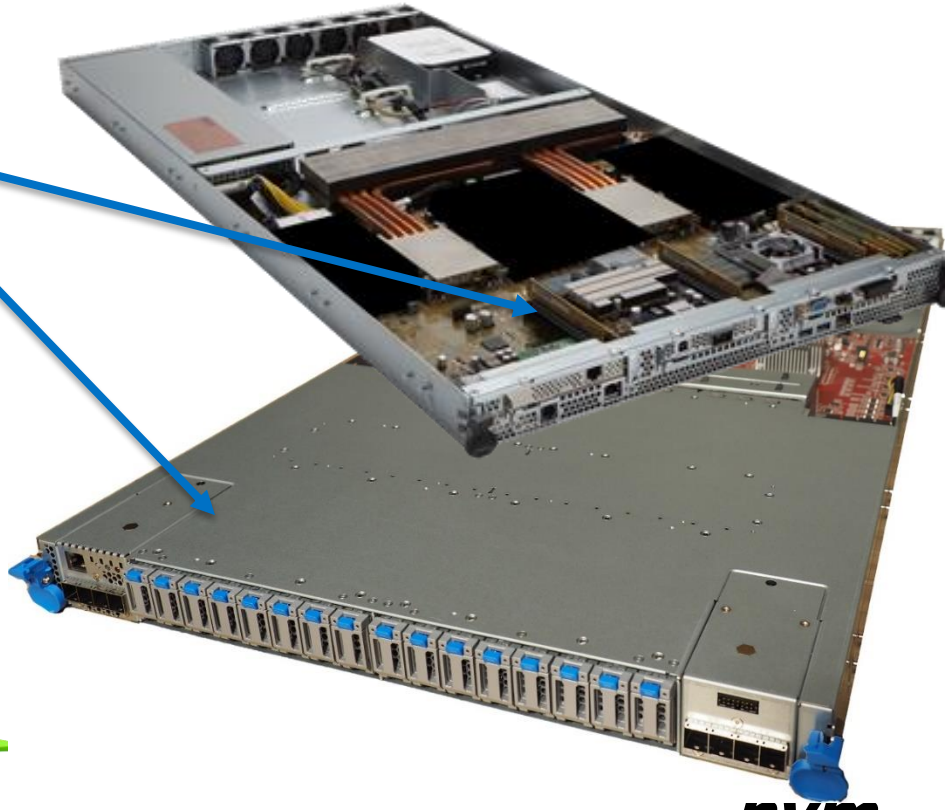
Databases

Cache

- Content caching
- Object caching
- Indexing

Where do Hyperscalers use flash today?





M.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

Less important:

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

Hyperscale NVM Characteristics and Challenges

Important:

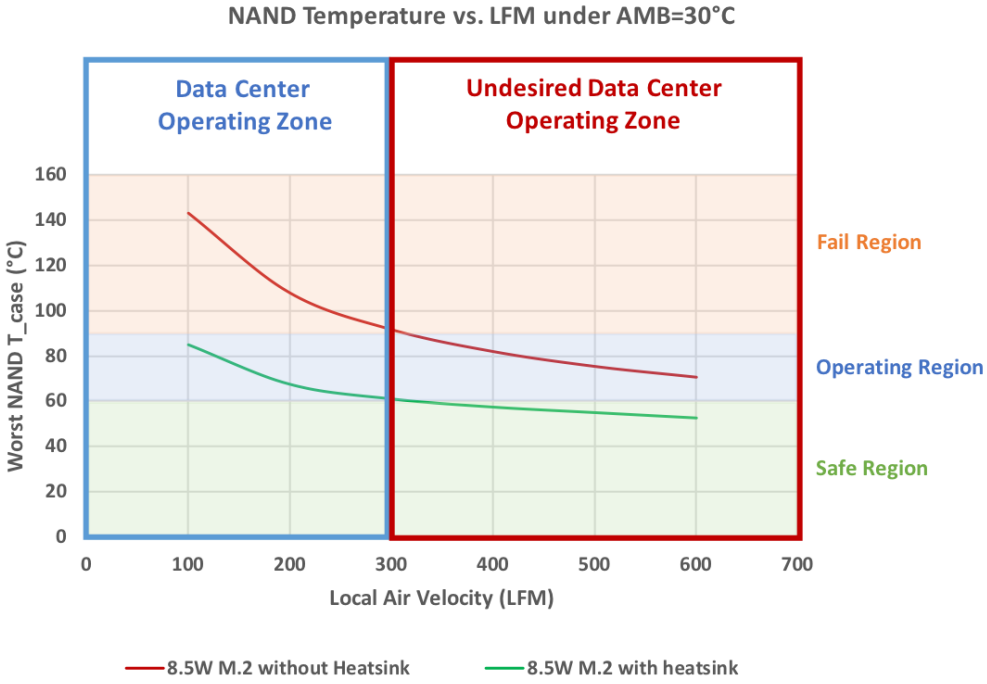
- Scalable & Flexible
- High volume & Low cost
- Power & Thermal Efficiency
- Hot-swappable & Serviceable
- Performance per TB & Quality of Service

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 datacenters
- Temperature increase across servers is large (delta T)
- OPEX matters

NVMe-MI™ enables effective thermal management!

Hyperscale NVM Characteristics and Challenges

Important:

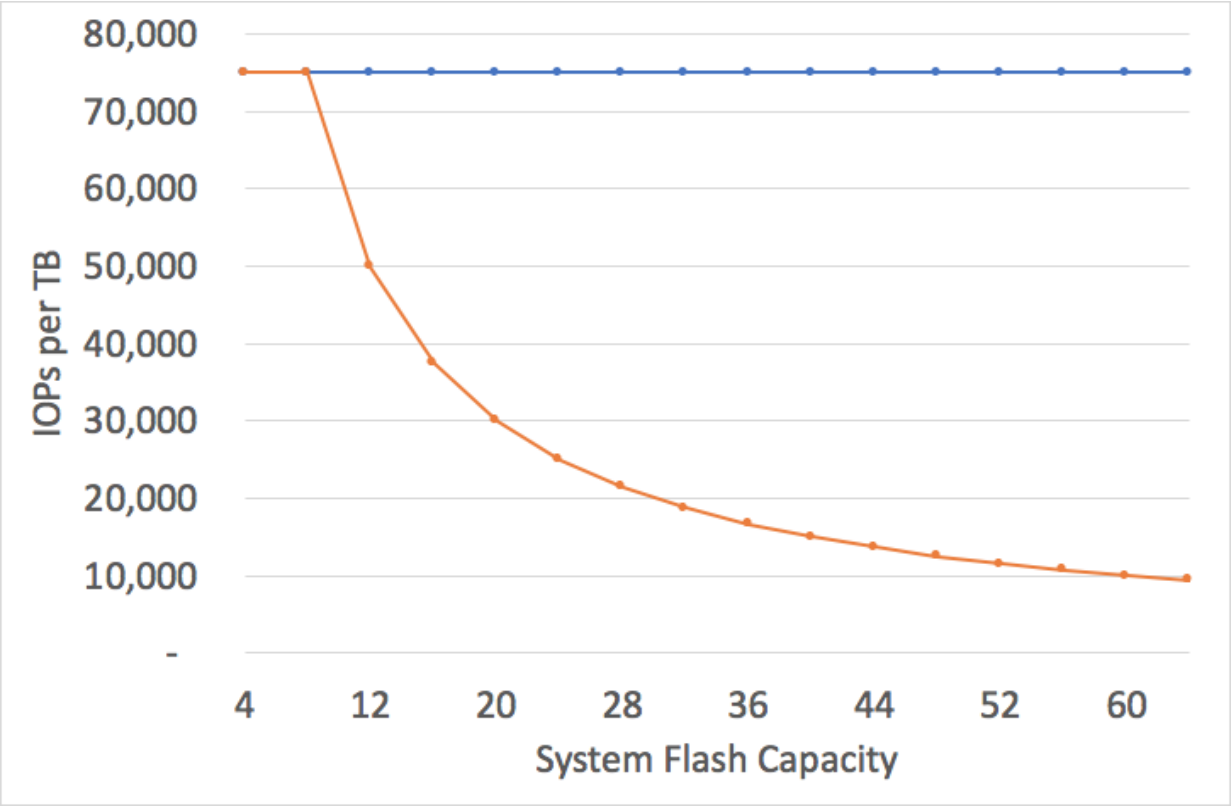
- Scalable & Flexible
- High volume & Low cost
- Power & Thermal Efficiency
- Hot-swappable & Serviceable
- Performance per TB & Quality of Service

Less important:

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

Scalable Performance

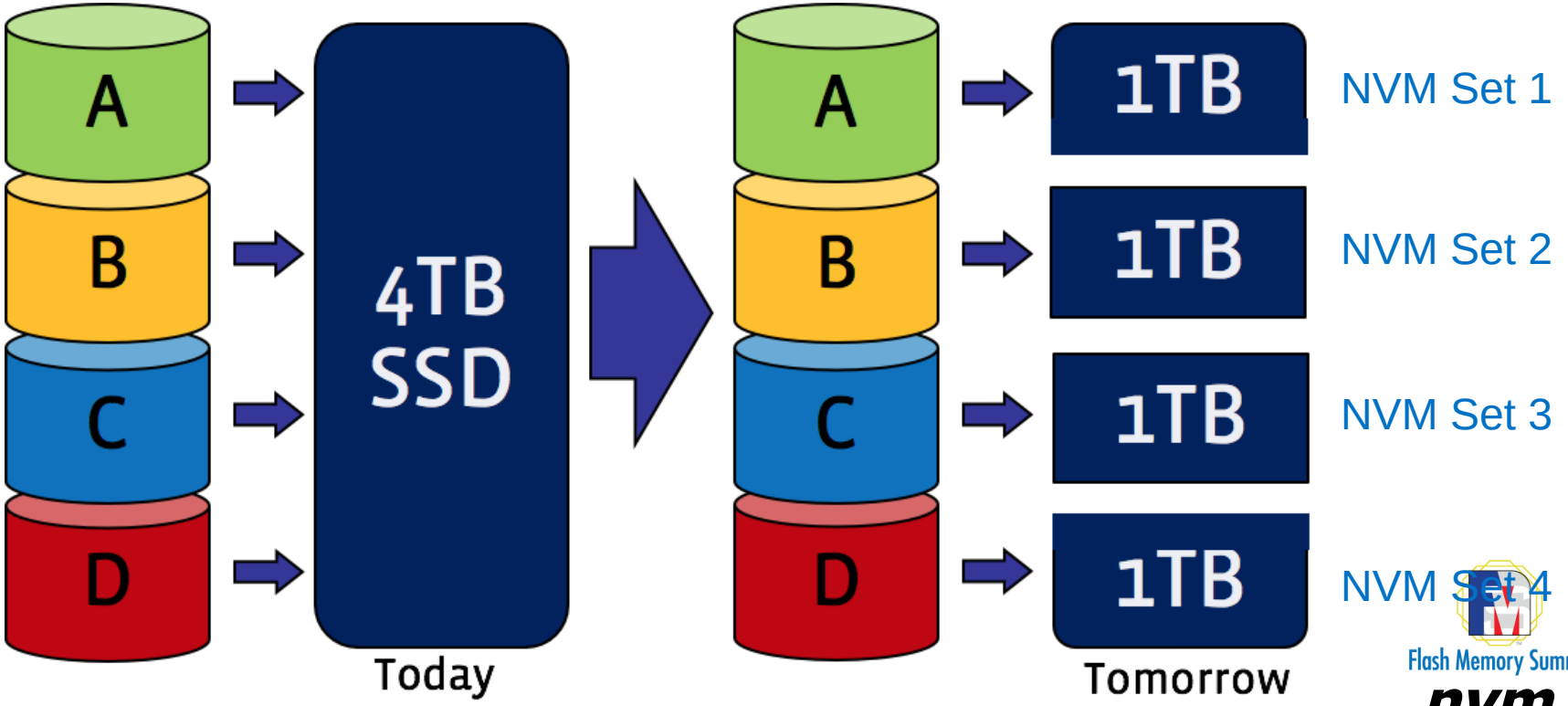
IOPs scales with capacity



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

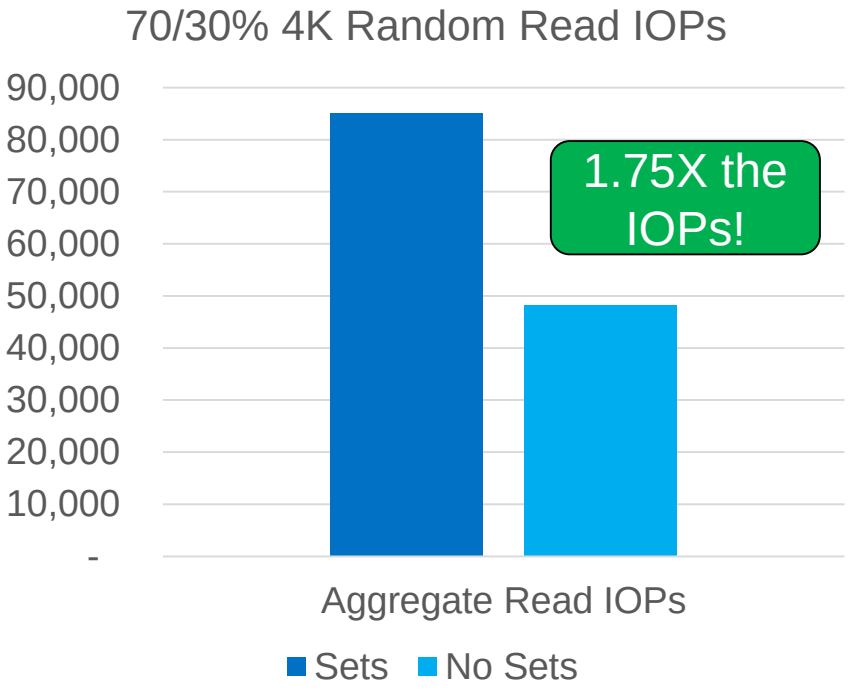
Scalable Performance

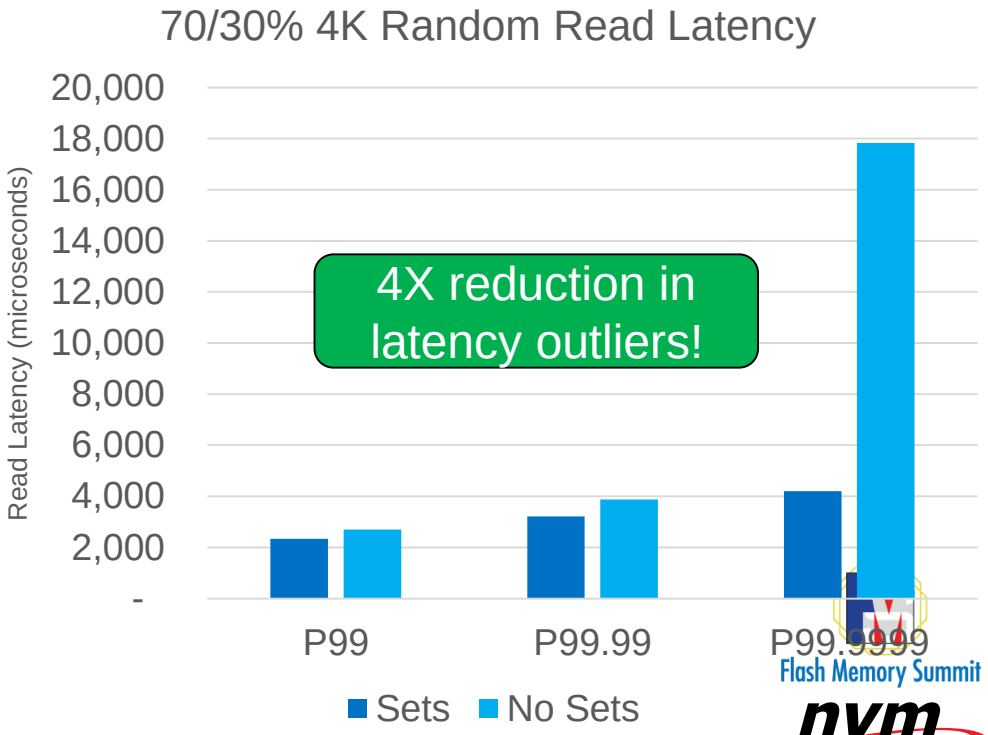
NVMe™ I/O Determinism



Scalable Performance

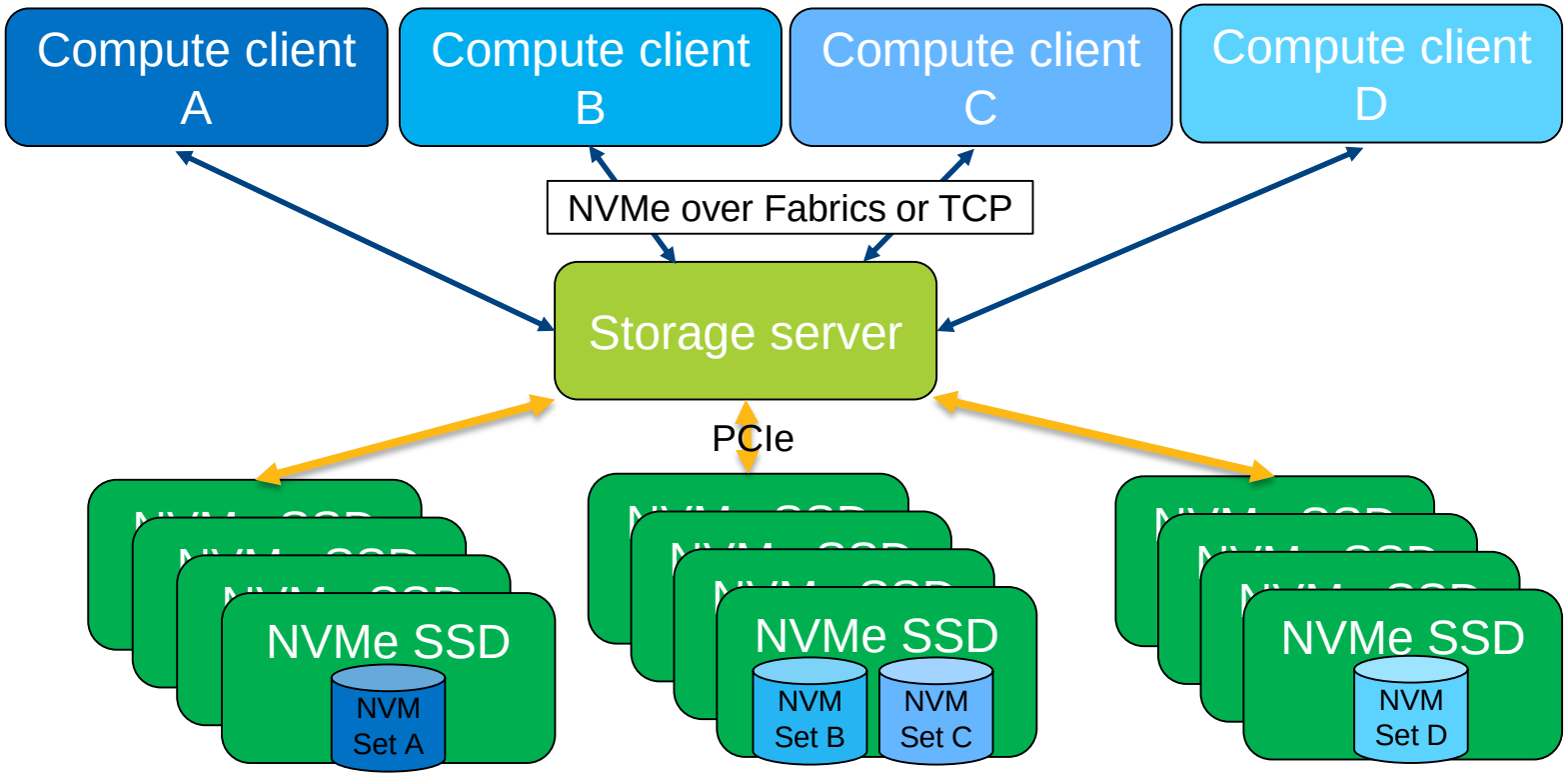
NVMe™ I/O Determinism





Scalable Performance

NVMe™ provides fabric connectivity



Hyperscale NVM Characteristics and Challenges

Important:

- Scalable & Flexible
- High volume & Low cost
- Power & Thermal Efficiency
- Hot-swappable & Serviceable
- Performance per TB & Quality of Service

Less important:

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

There may be many challenges, but innovative, standardized solutions are the key to scaling for the future!

NVMe™ for Enterprise Datacenters Needs

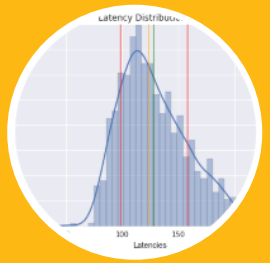
Chander Chadha, Toshiba

Enterprise Datacenter needs from storage

Scale

Performance

Pooling
Disaggregated

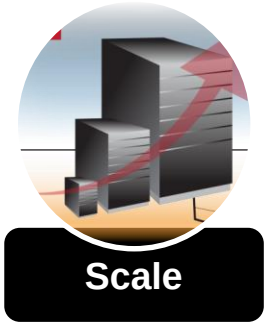


QoS

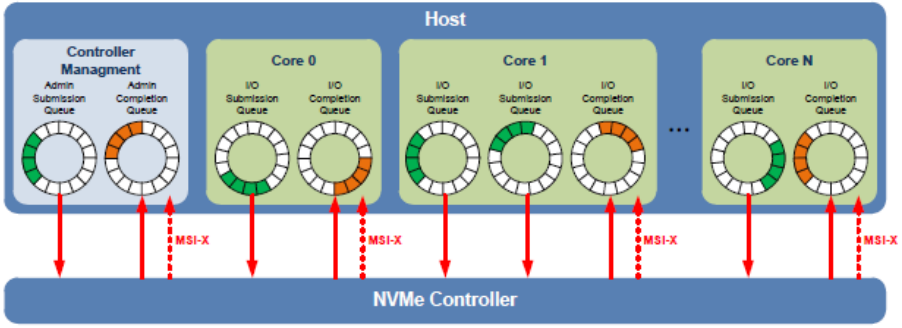
Data Integrity

Fault Tolerance

How NVMe™ benefits Enterprise Datacenter needs …



- ✓ Multiple core architecture with deeper queue depth
  - Non-locking cores for faster and parallel threads execution
  - Saleable queuing as system needs more storage resources while keeping the performance high
  - Host front end interface can maximize advantage of Flash parallelism



How NVMe™ benefits Enterprise Datacenter needs …

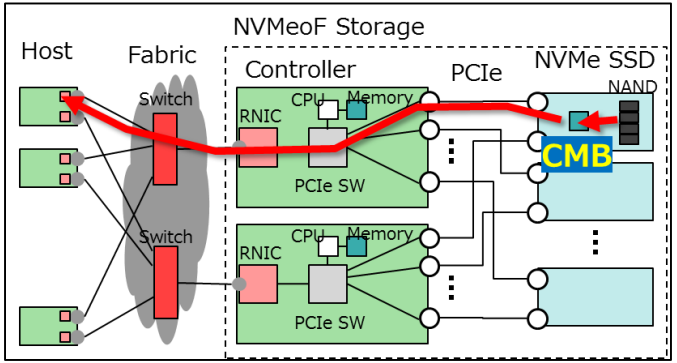
Performance

- ✓ PCIe® interface (1GB/lane Gen3)
  - Scalable interface with options to add lanes
  - Higher bandwidth and Random Performance over legacy SATA & SAS
  - Faster response time due to HBA elimination as PCIe direct attached
  - 4KB sector size (NVMe) better aligned (compared to 512B) for application performance acceleration

How NVMe™ benefits Enterprise Datacenter needs …

Pooling Disaggregated

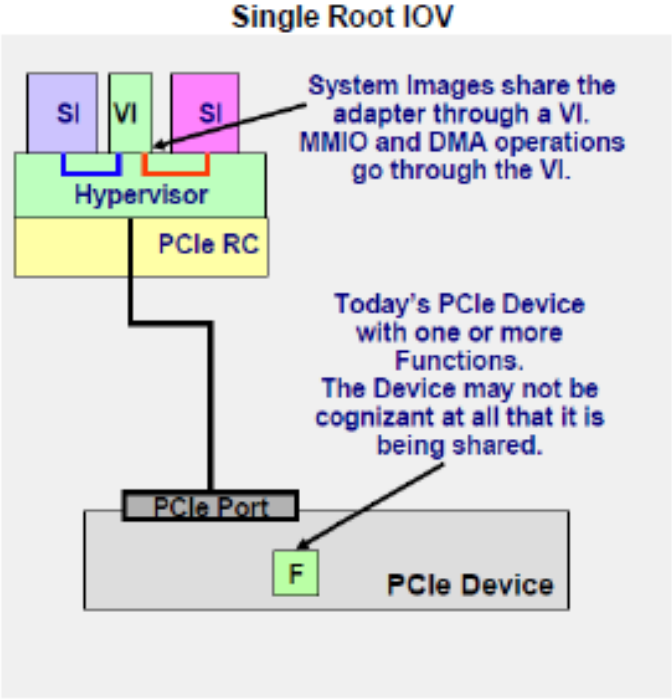
- ✓ SGL to connect fragmented Host Memory data to NVMe SSD to reduce IO and improve efficiency
- ✓ CMB, PMR(for persistency)as DRAM buffer for RDMA NICs for directly placing NVMe-oF™ queues and data into the NVMe SSD
  - Improves latency and eliminates Host CPU intervention, improving system performance



How NVMe™ benefits Enterprise Datacenter needs …

Pooling Disaggregated

- ✓ NVMe SRIOV enables single storage device to be exposed as multiple PCIe functions.
  - Improves latency as storage gets directly virtualized (native storage virtualization)
  - Multiple namespace sharing or Global namespace sharing options across multiple VF's
  - With multiple VF's HostBandwidth and IO's utilization across multiple Host with SRIOV

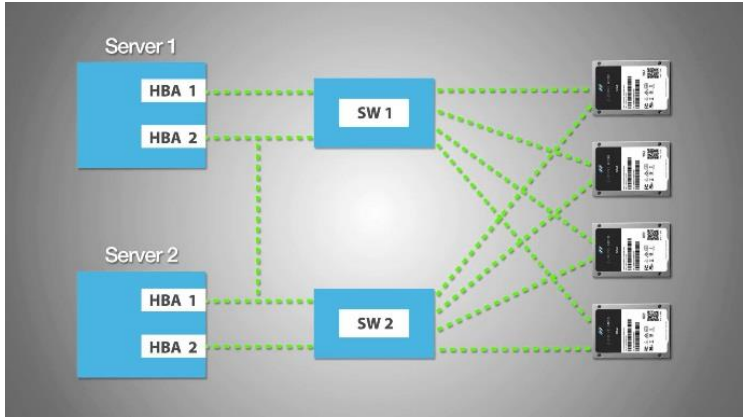


Source: "IO Virtualization and Sharing: PCI-SIG Technical Seminar 2007" – Michael Krause (HP), Renato Recio (IBM)

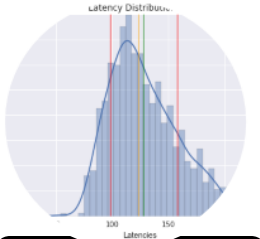
How NVMe™ benefits Enterprise Datacenter needs …

Fault Tolerance

- ✓ NVMe Dual Port
  - Redundant host physical access for failover
  - Reservation capabilities allow recovery from failing host

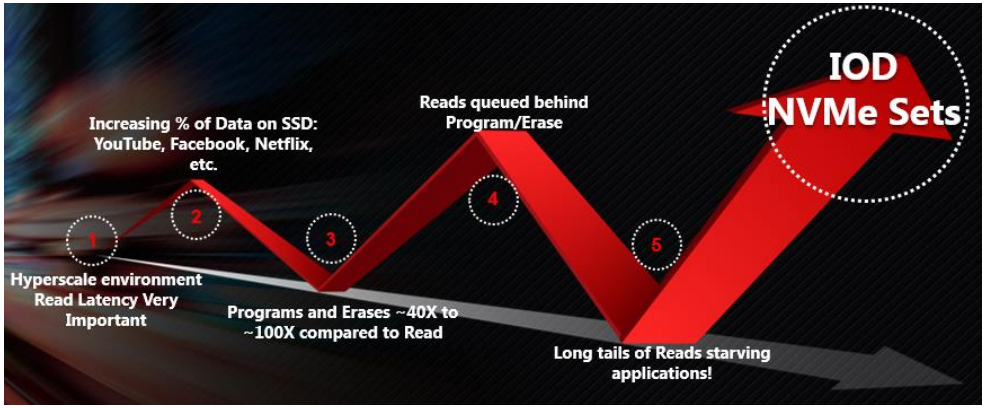


How NVMe™ benefits Enterprise Datacenter needs …



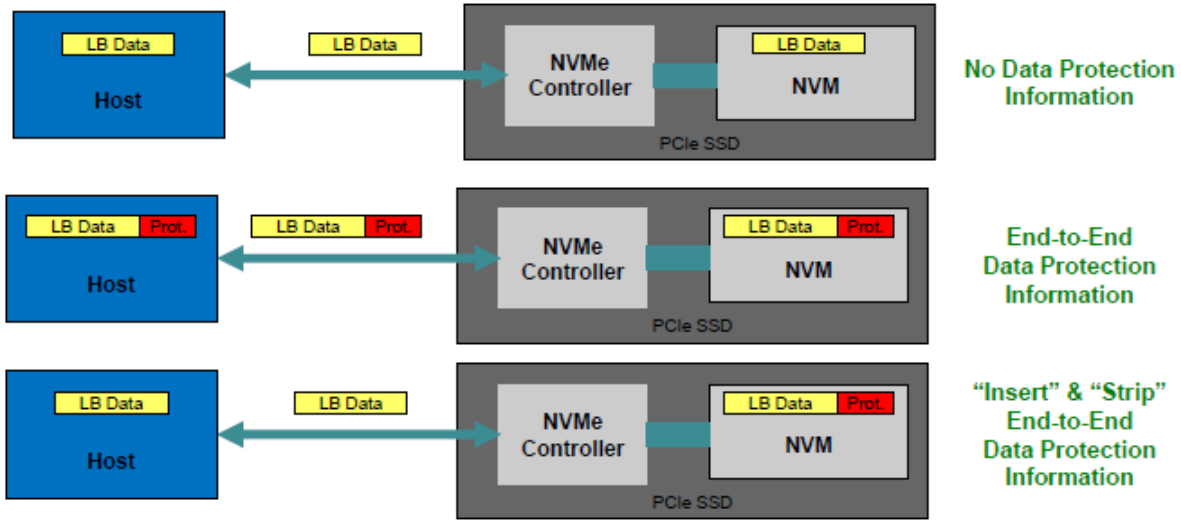
QoS

- ✓ NVMe Sets to address specific QoS needs for applications
  - NVMe SSD configured as multiple sets for QoS targets
    - Example: Sets targeted for Read QoS to prioritize read operations
  - Host scheduling of IO's based on deterministic time windows



How NVMe™ benefits Enterprise Datacenter needs …

✓ Fully compatible with T10 DIF & DIX, including DIF Type 1, 2, and 3



Flash Memory Summit

nvm EXPRESS®

NVMe™ Client Implementations

Jonmichael Hands, Intel

Client use cases for NVMe™

Gaming

Opens up the opportunity for unparalleled realism, with high quality textures and decreased load times

Content Creation

NVMe creates opportunity for new workflows for content creation when working with large data sets. Creators frequently move, backup, and duplicate storage

Workstation

Opportunity to accelerate any WS workload with large data requirements, reduce CPU idle time

Speed up design, CAD, simulations

Client / Mobile

High performance is driving NVMe into client. Efficiency and features of NVMe lead to better battery life. Lower latency and better QoS delivers better application responsiveness

Media Creation

Rendering, high resolution (4k, 8k editing), audio production

Consumer product storage priorities

What are consumer storage needs

- Low cost
- Small form factor
- Optimal thermal and power management
- High performance
- Low active power usage
- Compatibility

Why is NVMe™ great for all consumer storage?

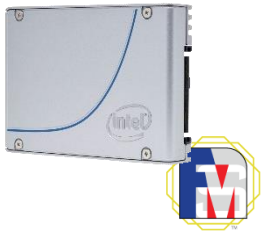
- Scalable streamlined storage stack
- Low latency
- Industry standard drivers in all OS
- Robust features to address power/thermals
- Scalability /w PCIe® and next gen NVM
- Built in security and manageability features

Client Desktop PCIe® Storage Form Factors

Add-in-card

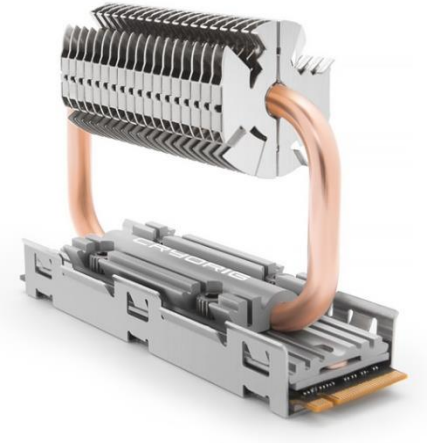
M.2

U.2



Source: https://www.msi.com/Motherboard/X299-XPOWER-GAMING-AC.html

M.2 mania!



http://www.cryorig.com/news.php?id=80
https://www.asus.com/us/Motherboard-Accessory/HYPER-M-2-X16-CARD/gallery/
https://www.ekwb.com/shop/ek-m-2-nvme-heatsink-black

Choose the right laptop (hint…it needs NVMe™ SSD)

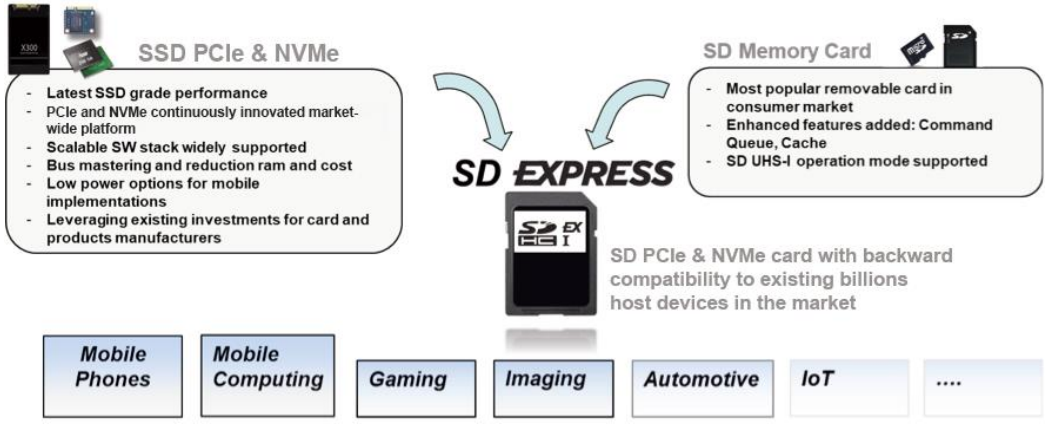
Choose the ultimate in form, funCtion & style!

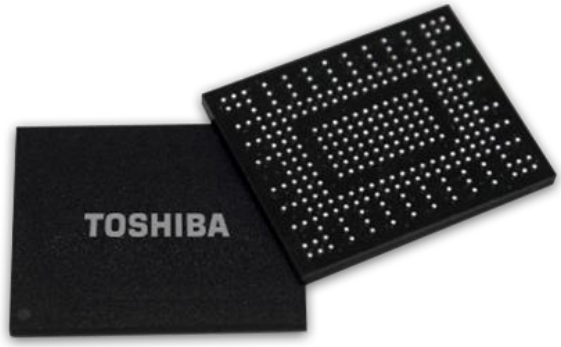
Choosing a balance of performance, mobility & battery life in the right form factor is essential.

	2 in 1 personal laptops equipped with Intel® Core™ Processor (Y-Series)	Versatile laptops equipped with Intel® Core™ Processor (U-Series)	Intel® Core™ Processor- based clam shell form factor laptops (H-Series)	Intel® Core™ Processor- based clam shell laptops supporting overclocking (HK-Series)*
				
	• BGA or M.2 NVMe	• M.2 NVMe	• M.2 NVMe and 2.5in SATA	• Dual M.2 NVMe slots and 2.5in SATA
PortaBility				
PerformanCe				
Battery life				

*Altering clock frequency or voltage may damage or reduce the useful life of the processor and other system components, and may reduce system stability and performance. Product warranties may not apply if the processor is operated beyond its specifications. Check with the manufacturers of system and components for additional details.

NVMe™ Scales to Mobile and Removable Storage





BGA 11.5x13mm

Learn more tomorrow at
CMOB-201B-1: New PCIe/NVMe Memory Cards Open up New High-Speed Applications

Source:
https://www.sdcard.org/downloads/pls/latest_whitepapers/SD_Express_Cards_with_PCIe_and_NVMe_Interfaces_White_Paper.pdf
https://business.toshiba-memory.com/en-us/product/storage-products/client-ssd.html

Power Consumption

BAPCo MOBILEMARK 2014

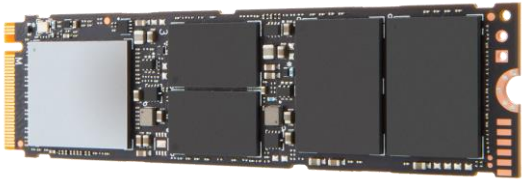
55.32 mW

Device Idle Power

19.32 mW

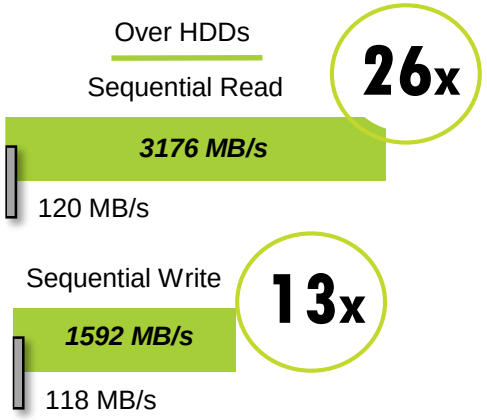
4K Video Playback

112.19 mW

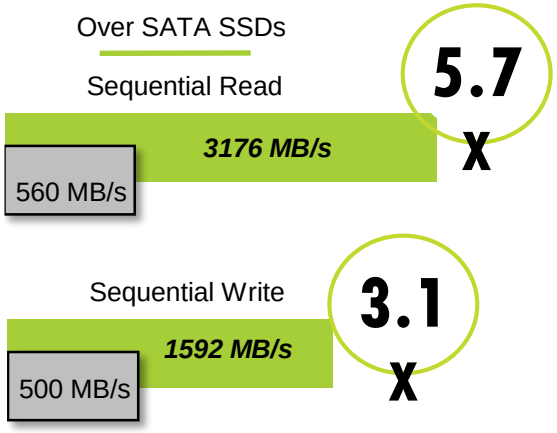


Intel® 760P SSD
Data is collected by Intel on Key sight 6705B* data logger by running Mobilemark* 2014 Office Productivity test for 2 hrs on Lenovo* Ideapad 720s. Windows* apps and other services are turned off for measurement consistency.
Data is collected by Intel on Key sight 6705B data logger by leaving the Lenovo Ideapad 720s for 10 mins and measuring the L1.2+PS3 power. Windows apps, radios, and other services are turned off for measurement consistency.
Data is collected by Intel on Key sight 6705B data logger by running 4K Video on the Lenovo Ideapad 720s for 1 hour and taking average of the measured power. Windows apps, radios, and other services are turned off for measurement consistency.
*Other names and brands may be claimed as the property of others.

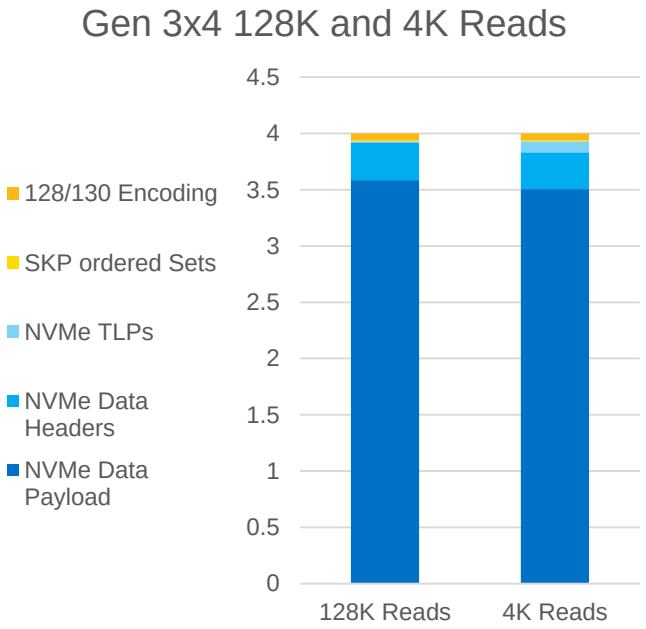
NVMe™ removes the SATA performance bottleneck



Intel® SSD 7 Series
versus
WD Blue* 5400RPM 500 GB HDD

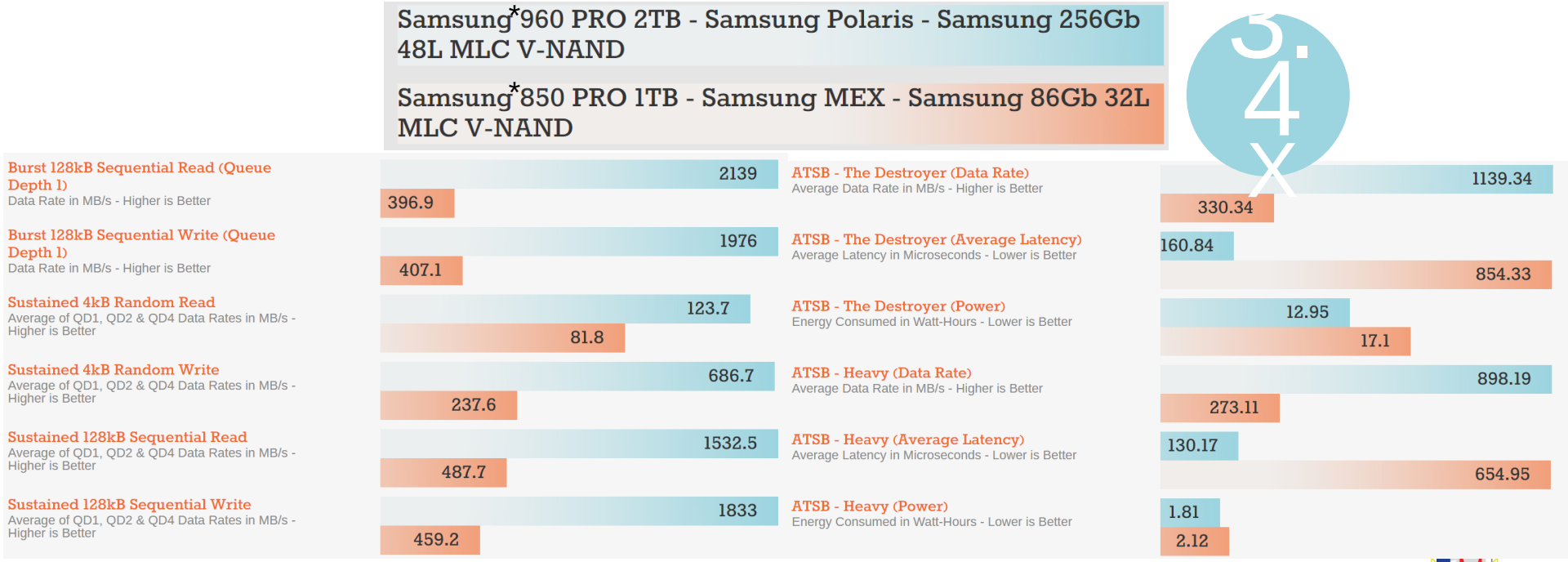


Intel® SSD 7 Series
versus
Intel® 545 SATA-based SSD



Storage performance comparison workload by Intel: CrystalDiskMark V5.2*. Drives being compared: Intel® 7600p vs Intel® 760p. System: processor: Intel® Core™ i7-7700K processor @ 4.5GHz Turbo Frequency, 8T/4C, 8MB cache, 91 W TDP, on motherboard: Asus Z270-A* Prime, memory: 2 X 4GB Corsair Vengeance DDR4* 3000MHz 8GB, operating system: Windows 10 Pro* (x64) OS RS2, storage: Intel® 600P and Intel® 760p 512GB. Drive under test is configured as a primary drive plugged into M.2 slot directly. System power profile set to performance mode. Data is collected at 500GB span size at Queue Depth 32 and thread 1 for sequential read and write. *Other names and brands may be claimed as the property of others.

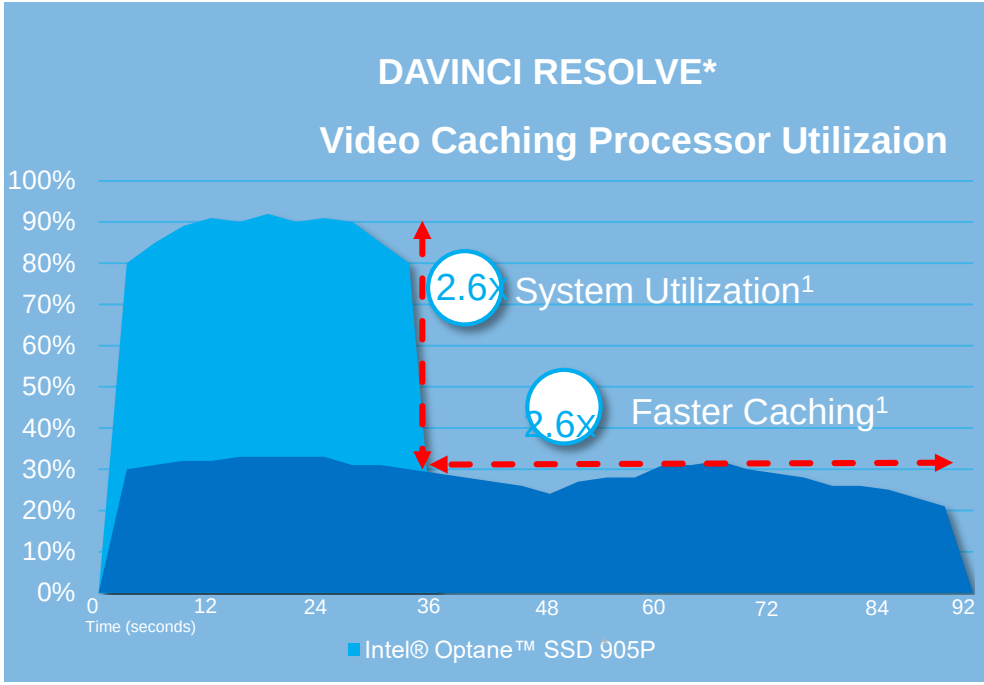
NVMe™ vs SATA Application Performance



*Other names and brands may be claimed as the property of others.

https://www.anandtech.com/bench/product/1922?vs=1982

NVMe™ required for next gen NVM
Intel® Optane™ Technology Proof Point



Up to 4x Faster renders[2]

Rendering 4K video
5.2mins to **1.26mins**[2]

Up to 2.6x System Utilization[1]

System utilization
30% to **90%**[1]

Up to 2.6x Faster Video Cache[1]

Transcoding 4K video
1.5mins to **36secs**[1]

Performance results are based on testing as of July 2018 and may not reflect the publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.
1.Test: Blackmagic DaVinci resolve 14* Video Caching of a 3.5mins @4K by using the command "media optimization." Test done by Intel. System Configurations: Intel® Core™ i9-7980X, Gigabyte X299 motherboard, NVIDIA* Geforce GTX1080, Memory 64GB (4X16GB) DDR4-2133, OS Win 10, Storage 1TB Intel® SSD 760p vs. 960GB Intel® Optane™ SSD 905P.
2.Test: Blackmagic DaVinci resolve 14* Video Rendering of a 3.5mins @4K by rendering it to DPX file format at 4K/24FPs/10b. Test done by Intel. System Configurations: Intel® Core™ i9-7980X, Gigabyte X299 motherboard, NVIDIA* Geforce GTX1080, Memory 64GB (4X16GB) DDR4-2133, OS Win 10, Storage 1TB Intel® SSD 760p vs. 960GB Intel® Optane™ SSD 905P.
*Other names and brands may be claimed as the property of others.

Flash Memory Summit
nvm EXPRESS®

NVMe™ 1.2 Improvements for Client

RTD3

Allows safe shutdown to the storage to save platform power

Platform Value

- Enables safe shutdown of device
- Power savings

Specification Details:

- Spec provides registers for providing device details for entry/exit latencies.

Additional Power State Info

Provides host additional info to the power levels supported by the device

Platform Value

- Additional details of power states to assist in transitions.
- Power/thermal benefit

Specification Details

- Spec allocates details in SMART

NVMe innovations enable additional features for client to help manage power/thermals.

NVMe™ 1.2 Improvements for small form factors

Host Memory Buffer

Allows the host driver to allocate system memory for the SSD's exclusive use

Platform Value

- Enables DRAM savings & smaller BGA packages
- E.g., Allocate translation tables in host DRAM

Specification Details:

- Device indicates preferred HMB size
- Host enables/disables via Set Features

Composite Temperature

Allows host to monitor temperature of the SSD

Platform Value

- Platform has feedback to the device temperature.
- If the host believes the temperature is out of its limits, it can set a lower power state on the NVMe device

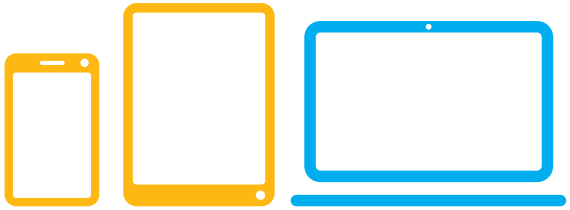
Specification Details

- Device indicates temperature in SMART
- Power State can be changed in power management

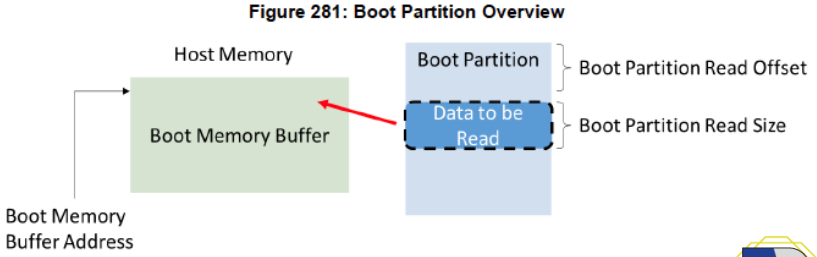
NVMe innovations enable scaling into smaller form factors delivering new differentiated platforms.

NVMe™ 1.3 - Boot Partitions

- Optional storage area that can be read with “fast” initialization method (not standard NVMe queues). Example: UEFI bootloader
- Saves cost and space by removing the need for another storage medium (like SPI flash, EPROM)
- Write using standard NVMe Firmware Download and Firmware Commit
- Can be protected with **Replay Protected Memory Block**



Makes NVMe more accessible for mobile and client form factors



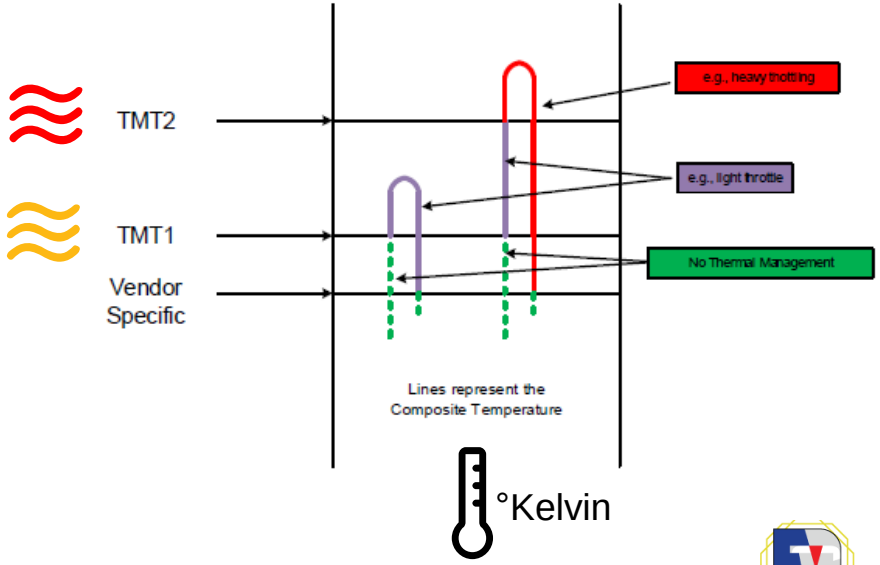
NVMe™ 1.3 - Host Controlled Thermal Management

Better thermal management in client systems like laptops and desktops.

Host can set **Thermal Management Temperature** at which a device should start going into a lower power state / throttling

- TMT1 – host tells SSD what temp in degrees K it should start throttling at
- TMT2 – threshold where the SSD should start heavy throttling regardless of impact to performance

Figure 264: HCTM Example



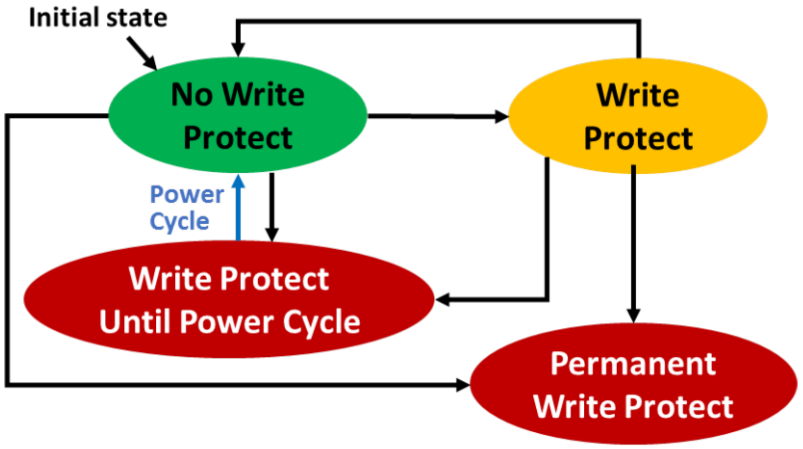
NVMe™ 1.4 –Namespace Write Protection

Namespace Write Protection is an optional configurable controller capability that enables the host to control the write protection state of a namespace.

(exactly what you think it does)

Could be used for secure space on drive, bootloader, backup image, important system files

Figure TBD1 – Namespace Write Protection State Machine Model




nvm
EXPRESS®
Architected for Performance

Back-Up Slides

NVMe™ 1.3 improvement for Enterprise NVMe

- Sanitize improvements
- Device Self Test
- Boot Partitions
- Error Log Updates
- Globally Unique NGUID/EUI64
- SGL Dword Simplify
- Streams Directive
- Telemetry
- Host Controlled Thermal Management
- NVMe-MI™ Tunneling