NVMe Annual Update, NVMe 1.3 and NVMe Market Data
Session A11 Part A
8:30 to 9:35

NVM Express Annual Update	Dave Landsman	Director Industry Standards, Western Digital
Major additions in the NVMe 1.3 standard and looking to the Future	Peter Onufryk	Fellow/NVM Solutions, Microsemi, and Member, NVMe Workgroup Board
NVMe in the Market with market data from IDC	Eric Burgener	Research Director, Storage, IDC

NVM Express Workgroup Update

Flash Memory Summit 2017

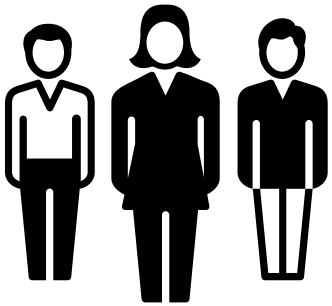
Dave Landsman
Director Industry Standards – Western Digital
NVMe BoD Member

NVM Express, Inc.
120+ Companies defining NVMe together

Board of Directors

13 elected companies, stewards of the technology & driving processes

Chair: Amber Huffman



Marketing Workgroup

NVMexpress.org, webcasts, tradeshows, social media, and press

Co-Chairs: Janene Ellefson and Jonmichael Hands

Technical Workgroup

NVMe Base and NVMe Over Fabrics

Chair: Amber Huffman

Management I/F Workgroup

Out-of-band management over SMBus and PCIe® VDM

Chair: Peter Onufryk

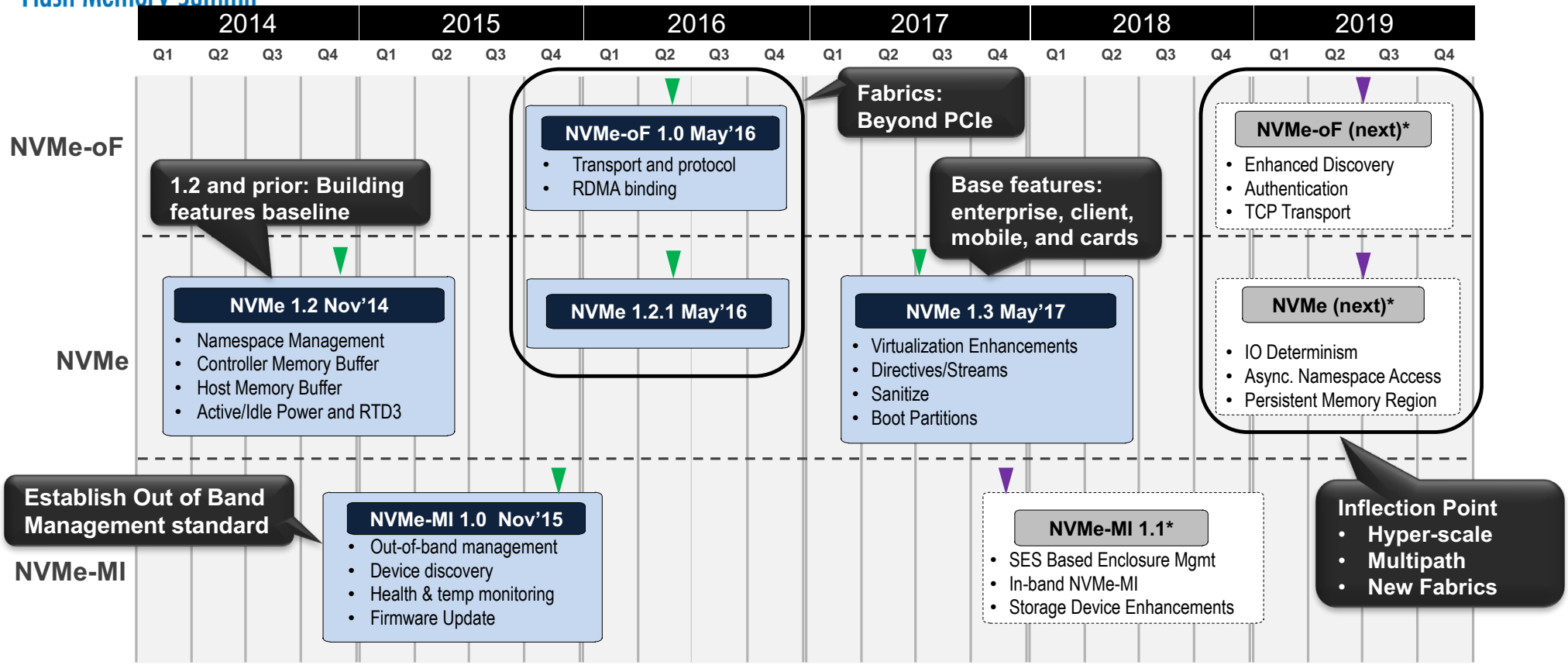
Vice Chair: Austin Bolen

Interop (ICC) Workgroup

Interop & Conformance Testing in collaboration with UNH-IOL

Chair: Ryan Holmqvist

NVMe Roadmap



Specification Release Planned Specification Release

* Subject to change

Follow NVMe!

- www.nvmexpress.org
  - Product Information
  - Events
  - Spec Info
  - Webcasts
  - News & Blogs
  - How to be a member

Flash Memory Summit

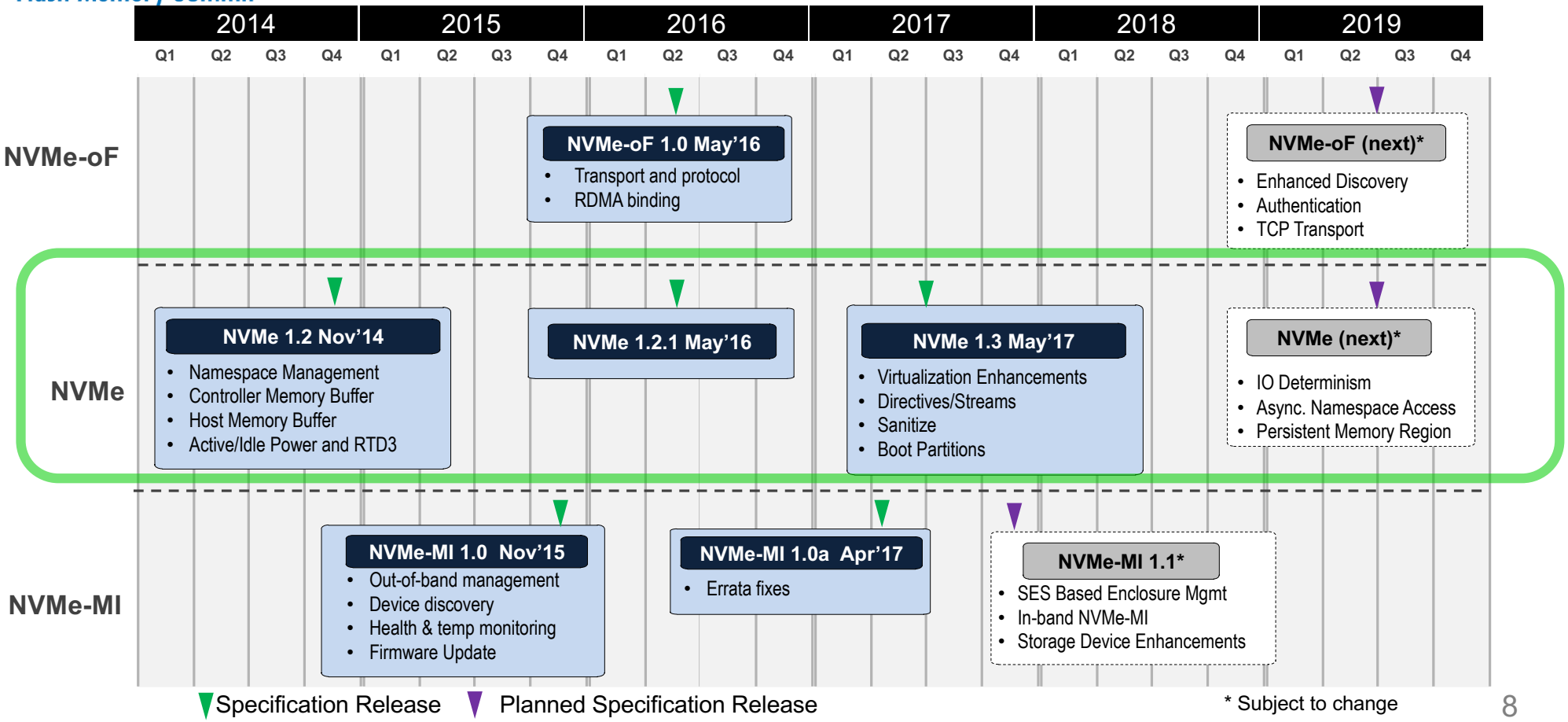
nvm EXPRESS®

THANK YOU

Major New Features in NVMe 1.3 and Looking to the Future

Peter Onufryk
Microsemi Corporation

NVMe Roadmap

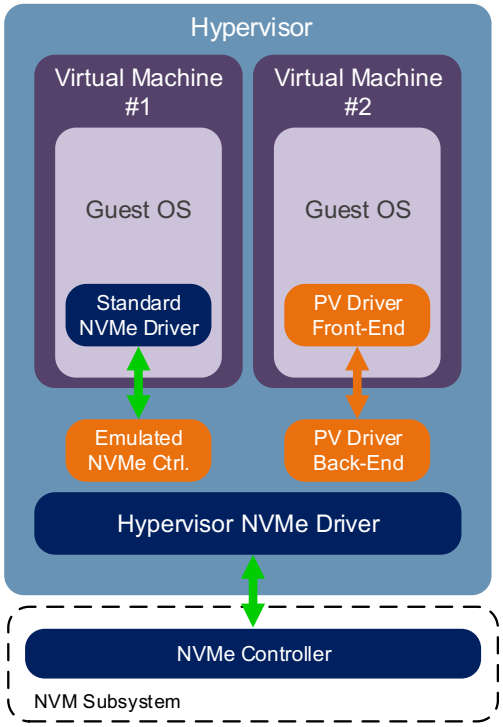


NVMe 1.3 New Feature Summary

Category	Description	Benefit
Client/Mobile	Boot Partitions	Enables bootstrapping of an SSD in a low resource environment
	Host Controlled Thermal Management	Host control to better regulate system thermals and device throttling
Data Center & Enterprise	Directives	Enables exchange of meta data between device and host. First use is Streams to increase SSD endurance and performance
	Virtualization	Provides more flexibility with shared storage use cases and resource assignment, enabling developers to flexibly assign SSD resources to specific virtual machines
	Emulated Controller Optimization	Better performance for software defined NVMe controllers
Debug	Timestamp	Start a timer and record time from host to controller via set and get features
	Error Log Updates	Error logging and debug, root cause problems faster
	Telemetry	Standard command to drop telemetry data, logs
Management	Device Self-Test	Internal check of SSD health, ensure devices are operating as expected
	Sanitize	Simple, fast, native way to completely erase data in an SSD, allowing more options for secure SSD reuse or decommissioning
	Management Enhancements	Allows same management commands in or out-of-band

Virtualization Enhancements Motivation

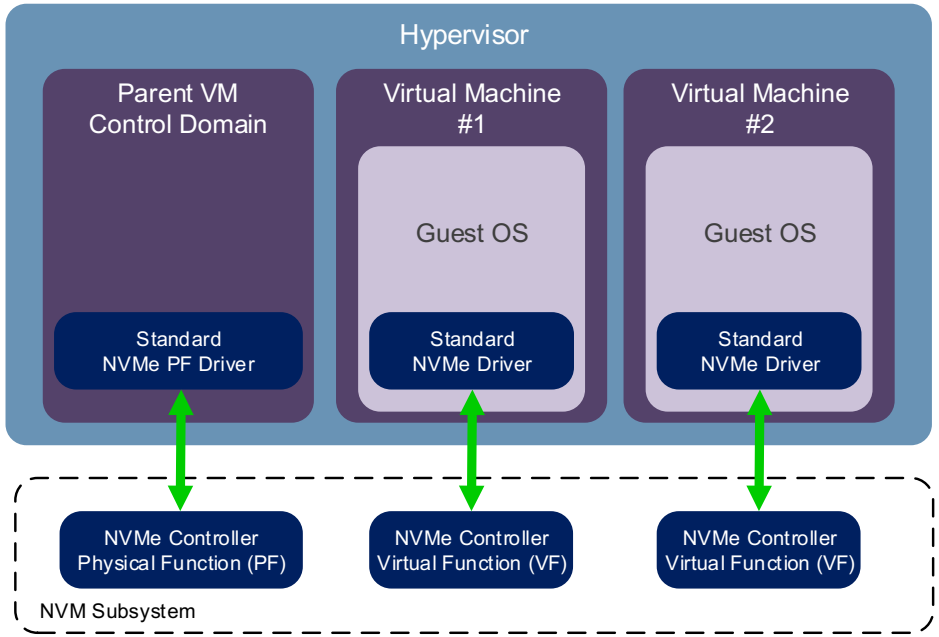
- Cloud hosters win with high density, oversubscription, and differentiation
  - Multi-tenancy is the norm
  - Premium differentiator offering high speed storage
- Virtual machines are inherently mobile and hosts are inherently dynamic



Software sharing limits NVMe throughput and increases latency

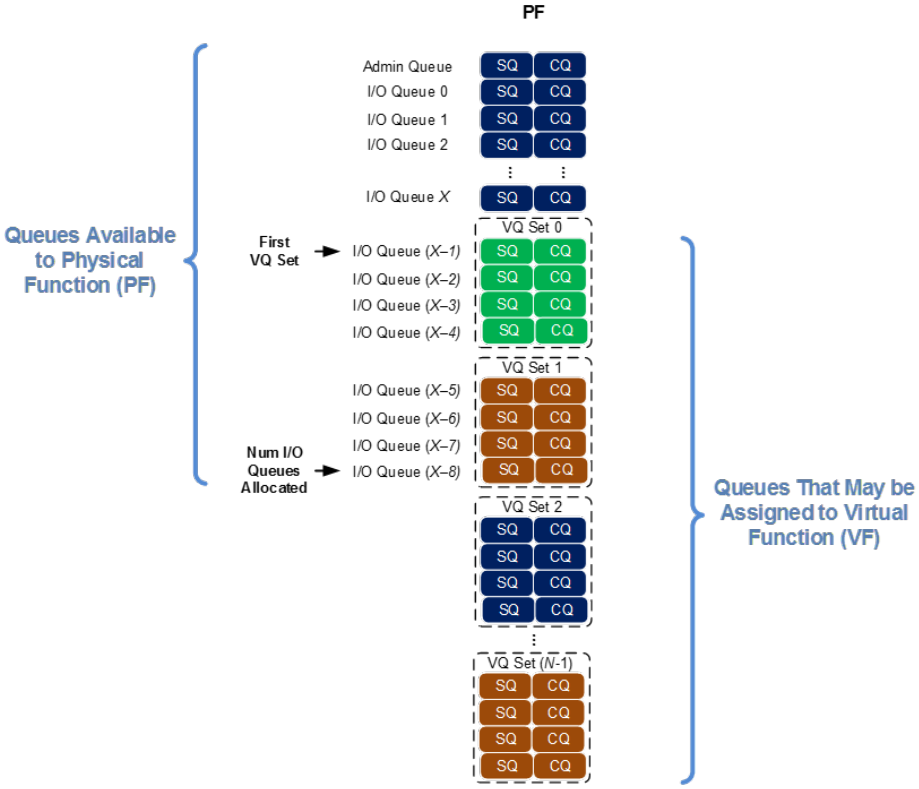
Virtualization Enhancements Direct Assignment

- Direct assignment enables each VM to directly access NVMe Controller
  - Eliminates software overhead
  - Allows guest OSes to use standard NVMe driver
- NVMe already supports PCI-SIG SR-IOV
- Virtualization enhancements standardize PF functionality and allow flexible dynamic mapping of resources
- Abstraction allows future mechanisms beyond SR-IOV



Virtualization Enhancements Allocating Resources

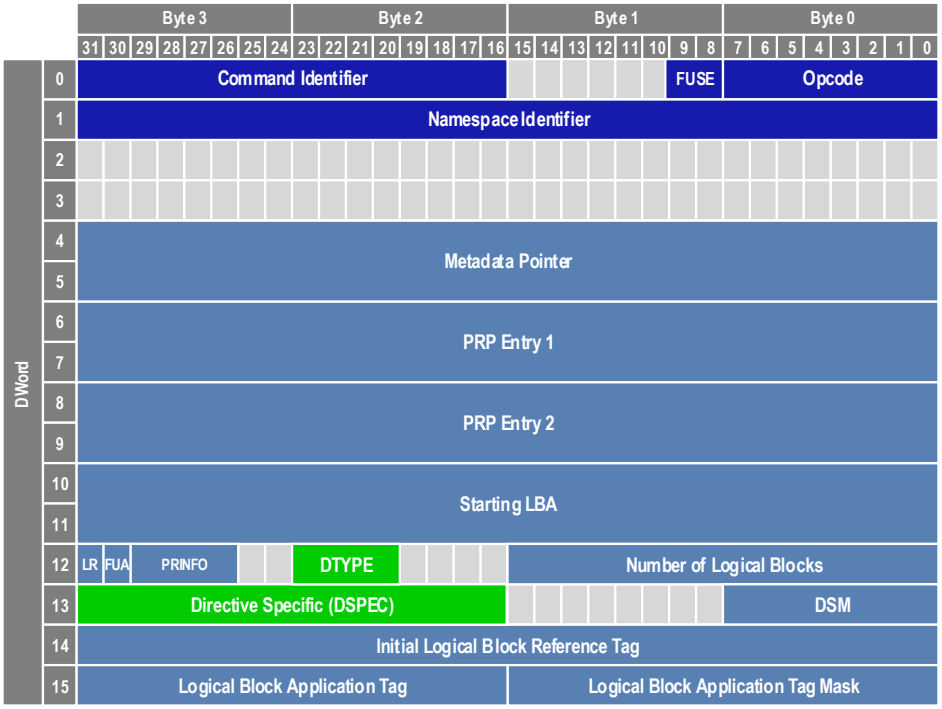
- Resources may be dynamically moved between the PF and VF(s)
- **VQ Set** – A set of (four) Submission Queue (SQ) and Completion Queue (CQ) pairs that may be assigned to a VF
- **VI Set** – A set of (four) MSI-X interrupt resources that may be assigned to a VF



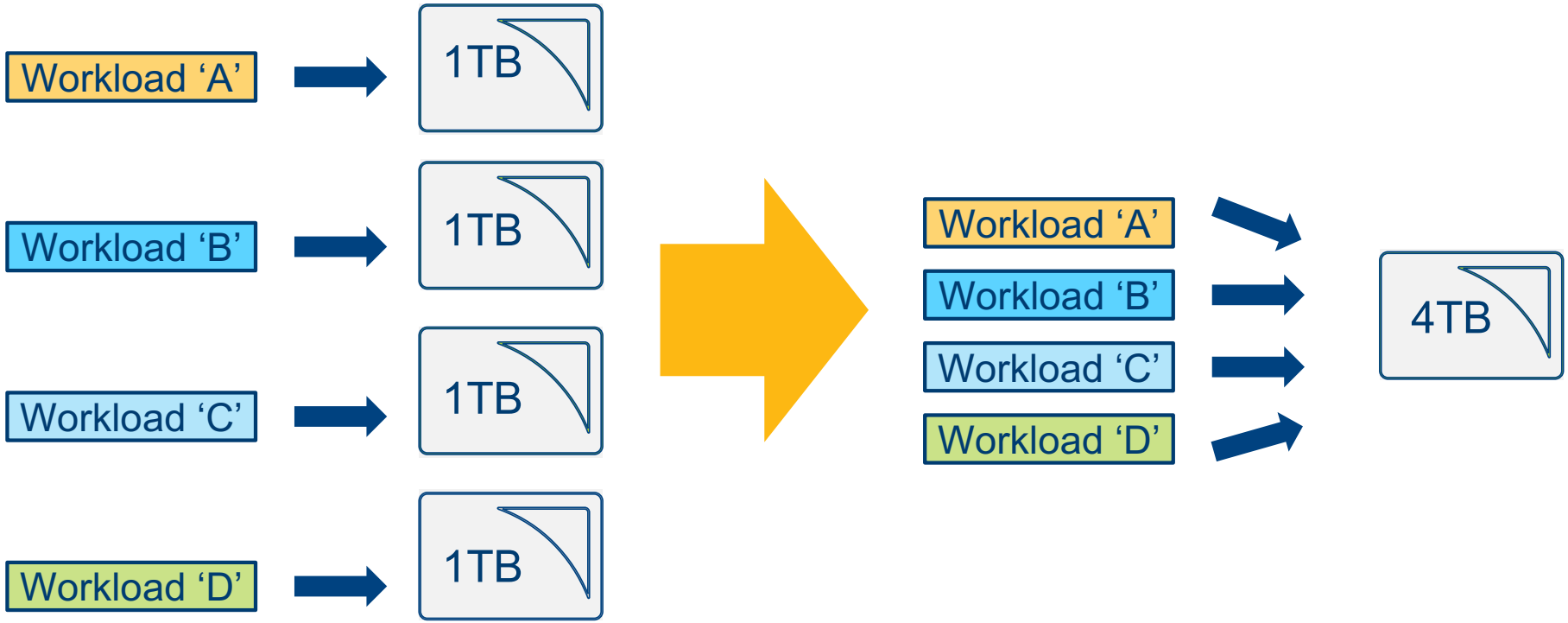
Directives

- NVMe commands are always 64-bytes in size
- Remaining unused space in commands needs to be allocated with care
- Reusable Directive Specific (DSPEC) field whose contents/format is defined by the Directive Type (DTYPE) field
- DSPEC is used for Streams today and future ideas tomorrow

Write Command



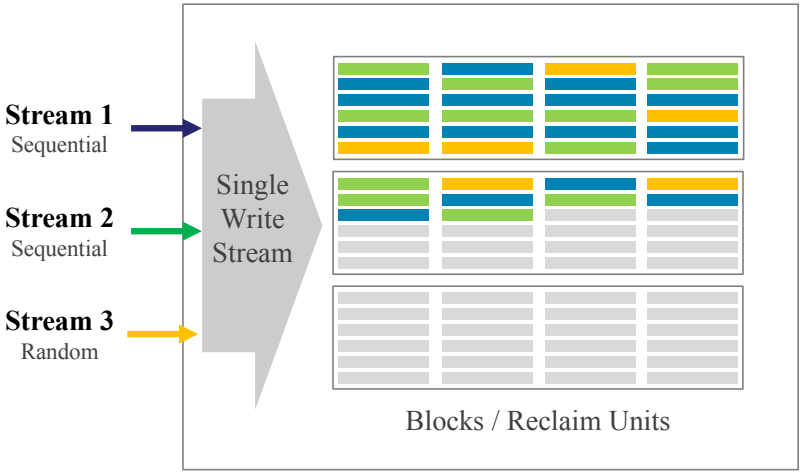
Streams Motivation



The Benefit of Streams

Standard SSD

No Stream Separation

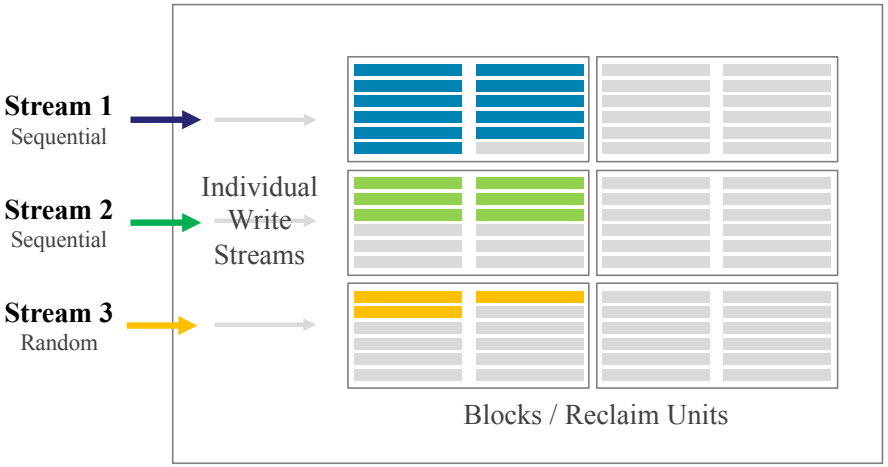


Mixed data needs garbage collection to reclaim blocks

Higher Write Amplification

SSD With Streams

Separation of streams into different reclaim units

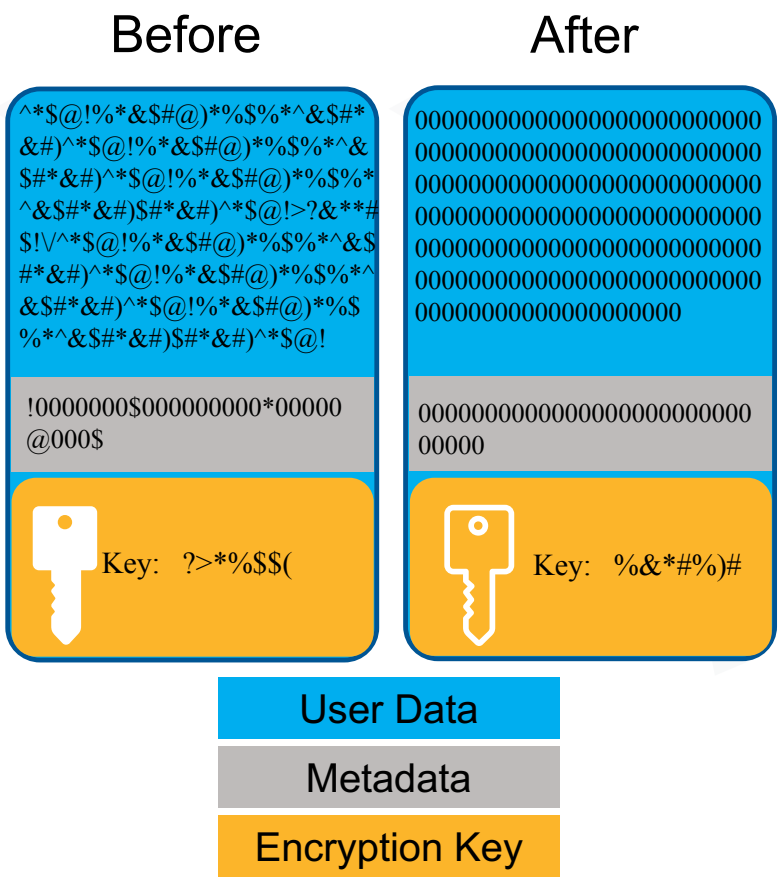


Separated data can be trimmed or self-invalidated to reclaim blocks

Lower Write Amplification

Sanitize

- Sanitize is used when retiring SSD from use, reusing for new use case, or end of life
- Alters user data so that is unrecoverable by erasing media, metadata, and cache
  - **Block Erase** – low level block erase on media (physically erase NAND blocks)
  - **Crypto Erase** - change media encryption key
  - **Overwrite** – overwrite with data patterns (not good or recommended for NAND based SSDs due to endurance)
- Benefits of Sanitize vs existing Format command
  - Keeps running after reset or pow
  - Erases all metadata, sensitive log pages and status during operation
  - Clears all user data in caches



Before

^*$@!%*&$#@)*%$%*^&$#*
&#)^*$@!%*&$#@)*%$%*^&
$#*&#)^*$@!%*&$#@)*%$%*
^&$#*&#)$#*&#)^*$@!>?&**#
$!\/^*$@!%*&$#@)*%$%*^&$
#*&#)^*$@!%*&$#@)*%$%*^
&$#*&#)^*$@!%*&$#@)*%$
%*^&$#*&#)$#*&#)^*$@!

!0000000$000000000*00000
@000$

Key: ?>*%$$(

After

00000000000000000000000000
00000000000000000000000000
00000000000000000000000000
00000000000000000000000000
00000000000000000000000000
00000000000000000000000000
00000000000000000000

0000000000000000000000000
00000

Key: %&*#%)#

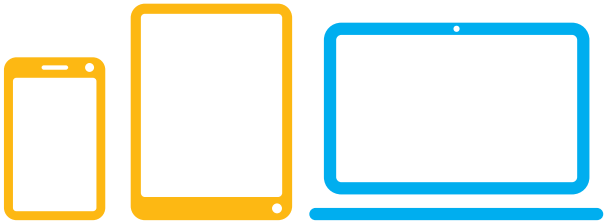
User Data

Metadata

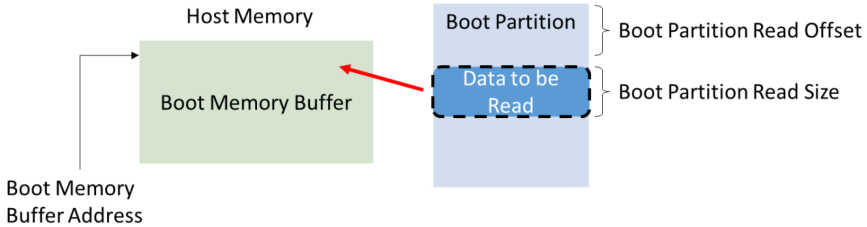
Encryption Key

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 (e.g., SPI Flash or EPROM)
- Boot partitions may be updated using standard NVMe Firmware Download and Firmware Commit
- Boot partitions may be write protected

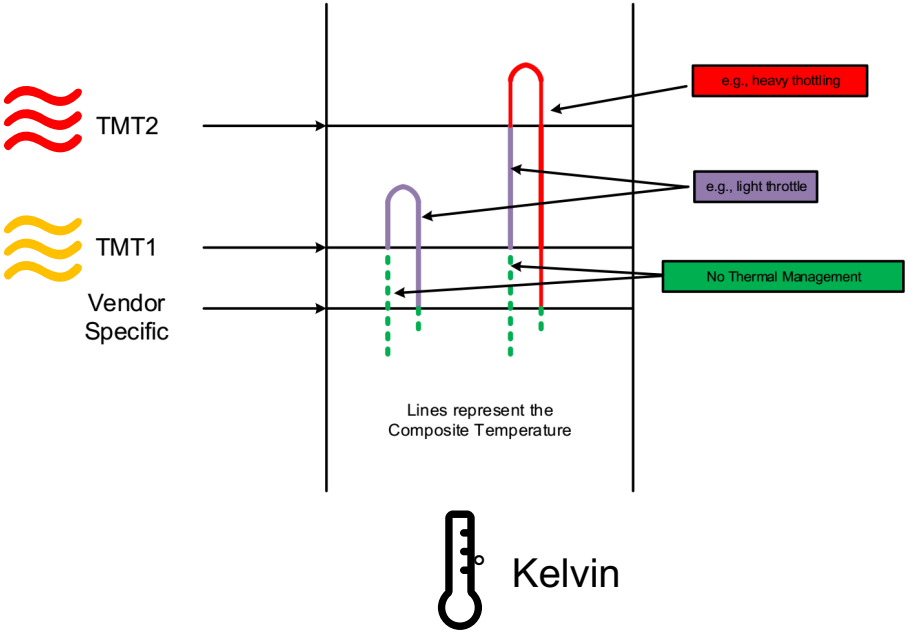


Makes NVMe more accessible for mobile and client form factors



Host Controlled Thermal Management

- Better thermal management in client systems like laptops and desktops
- Host can set **Thermal Management Temperature** at which an SSD should start going into a lower power state or throttling
- TMT1 – Threshold where SSD should start attempting to reduce temperature while minimizing impact on performance
- TMT2 – Threshold where the SSD should start reducing temperature regardless of impact to performance



New Debug Features

Timestamp

- Enables host to set a timestamp in controller via set features NVMe command, and read with get features

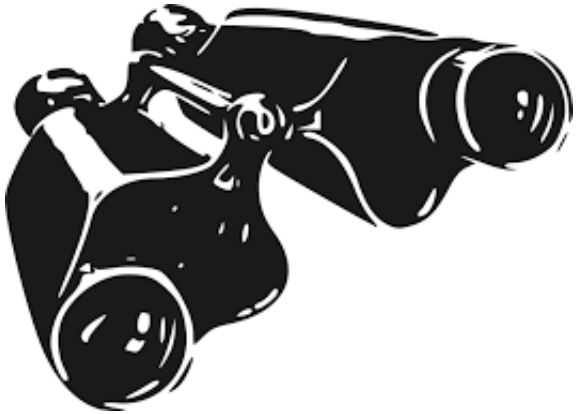
Error Log Updates

- Get Log NVMe command now returns more info on where the error occurred (queue, command, LBA, namespace, etc.) and error count

Telemetry

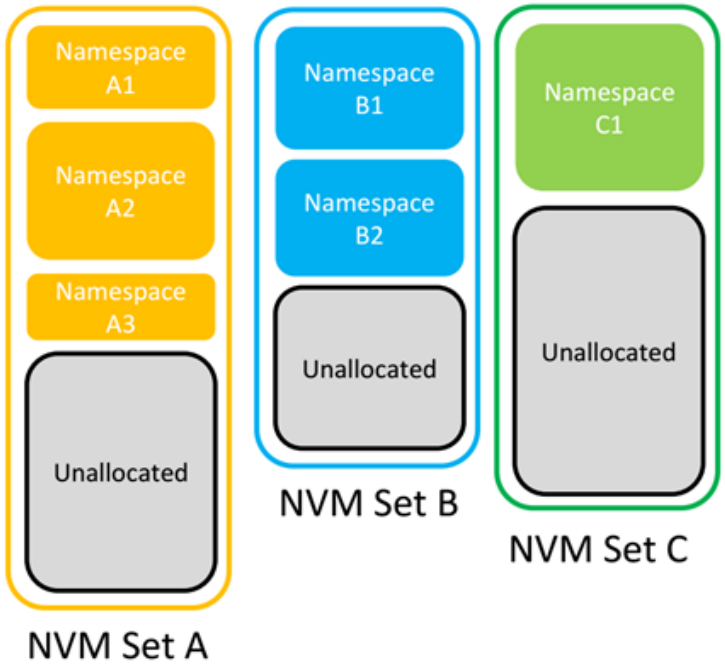
- vendor unique logs that can be dumped with industry standard commands and tools

Looking to the Future

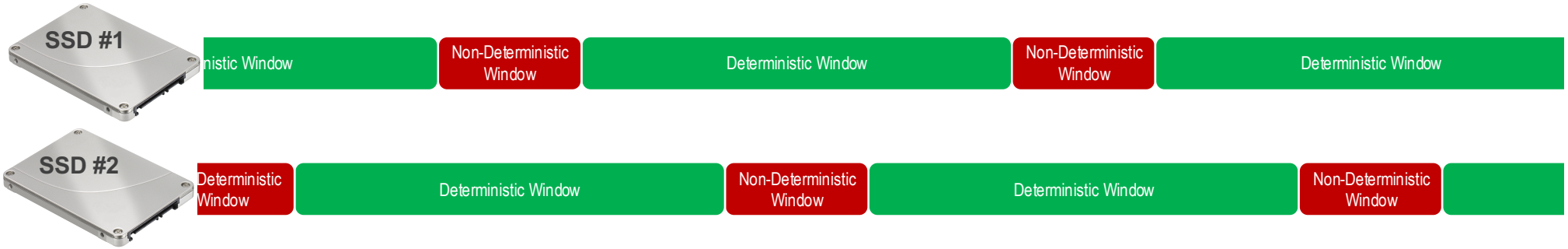


I/O Determinism - NVM Sets

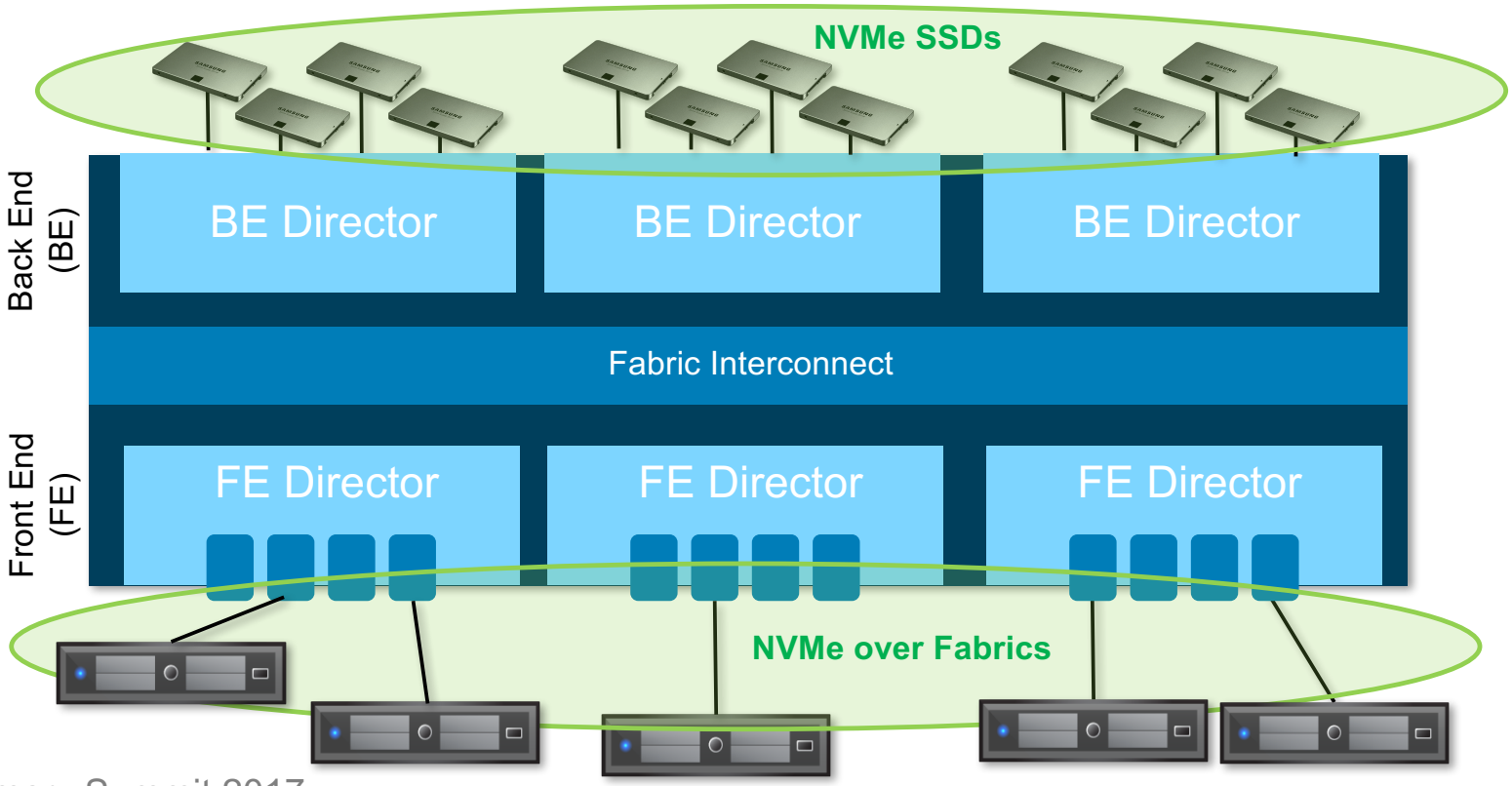
- NVM Sets are QoS Isolated
  - Write to namespace A1 does not impact QoS associated with namespace B2
- NVM Sets have attributes
  - Endurance
- NVM Subsystem may support one or more NVM Sets
- One or more Namespaces may be allocated to an NVM Set



I/O Determinism - Predictable Latency Mode

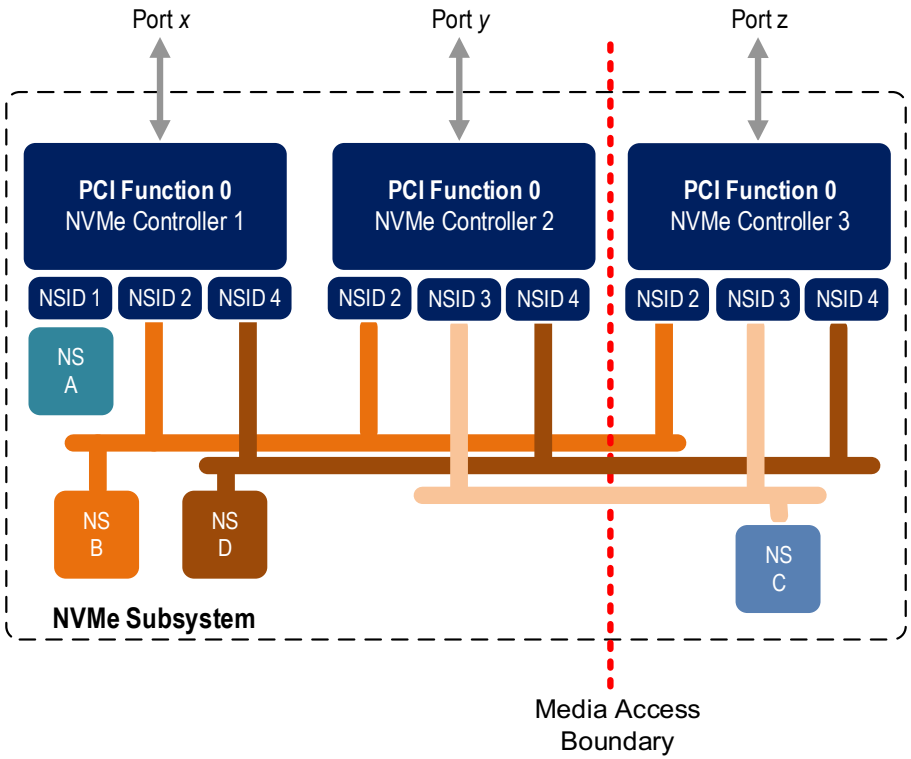


NVMe in High End Storage Systems



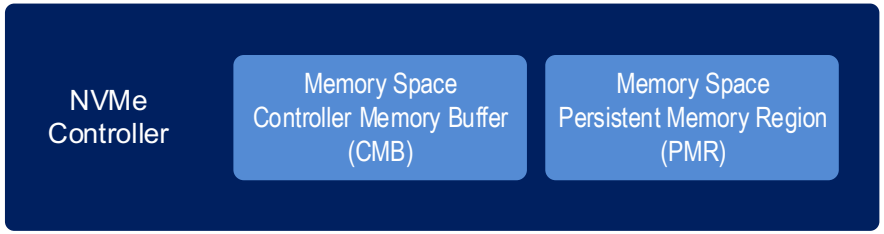
Slide credit: Dell EMC

Asynchronous Namespace Access



Persistent Memory Region (PMR)

- Controller Memory Buffer (CMB)
  - Introduced in NVMe 1.2
  - PCI memory space exposed to host
  - May be used to store commands & command data
  - Contents **do not** persist across power cycles and resets



- Persistent Memory Region (PMR)
  - PCI memory space exposed to host
  - May be used to store command data
  - Content persist across power cycles and resets

Summary

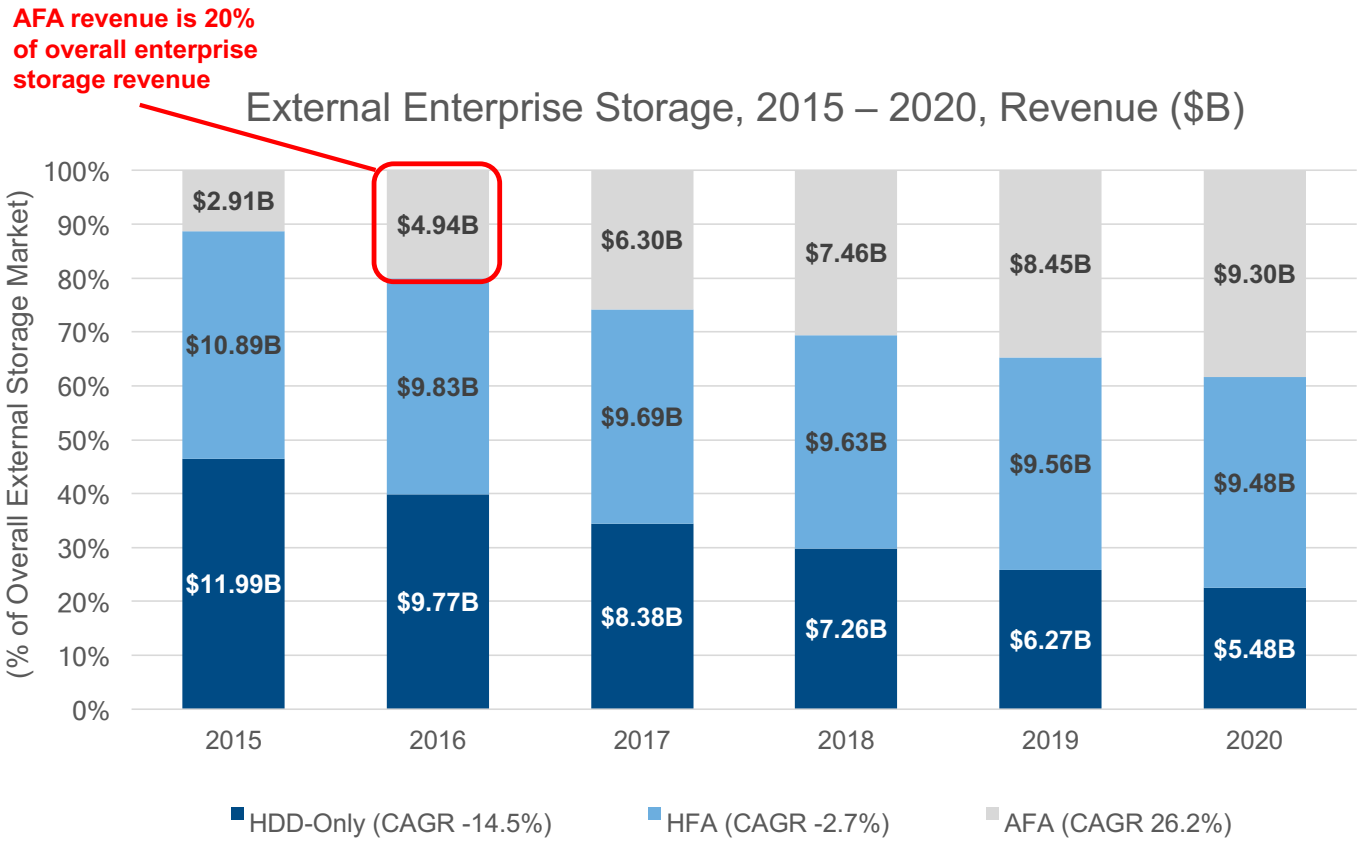
- NVMe continues to evolve with new innovative data center, enterprise, and client features
- NVMe 1.3 was ratified on May 1st and is available at www.nvmexpress.org
- Work is in progress on new features

IDC
Analyze the Future

NVMe in Enterprise Storage Systems

Eric Burgener
Research Director, Storage
IDC

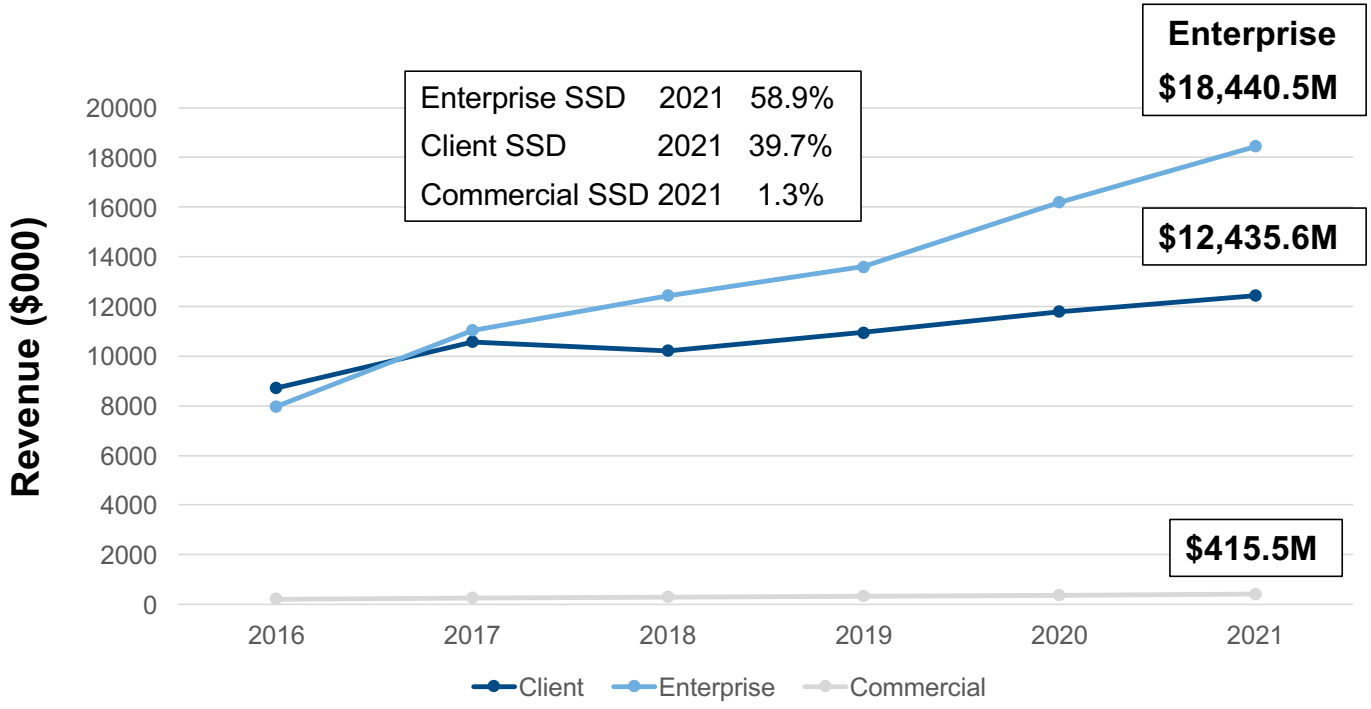
Flash Driving Enterprise Storage



Source: IDC, March 2017

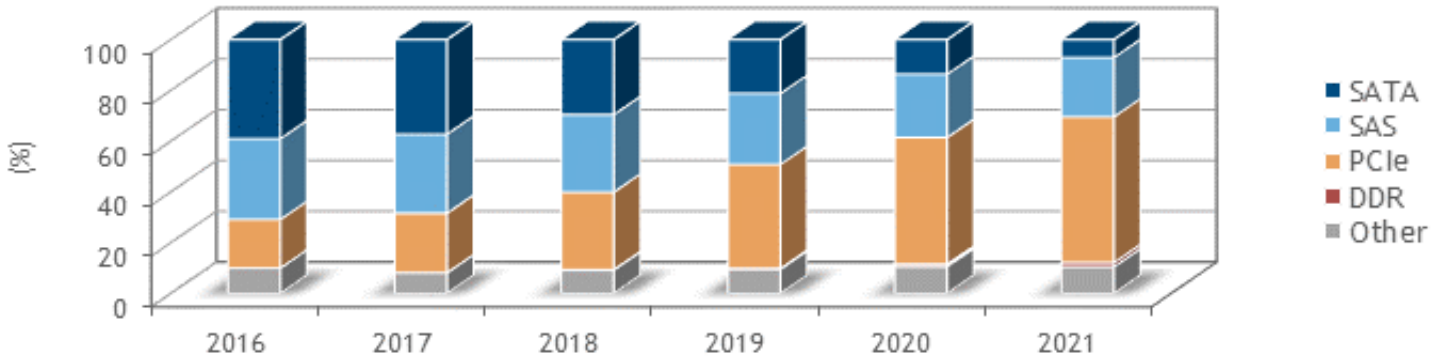
IDC Analyze the Future

WW SSD Market Forecast, 2017-2021



IDC Analyze the Future

WW SSD Revenue Share by Interface



- **PCIe will grow from a 12.4% revenue share in 2016 to a 61.4% share by 2020**
- **In 2020 NVMe SSD revenue alone will be $9.94B**
- **In 2020 rack scale flash systems revenues are still expected to be well under $1B**
- **In 2020 total SSD revenues will be $16.19B (all interface types)**

IDC Analyze the Future

Rack Scale Flash Systems Emerging

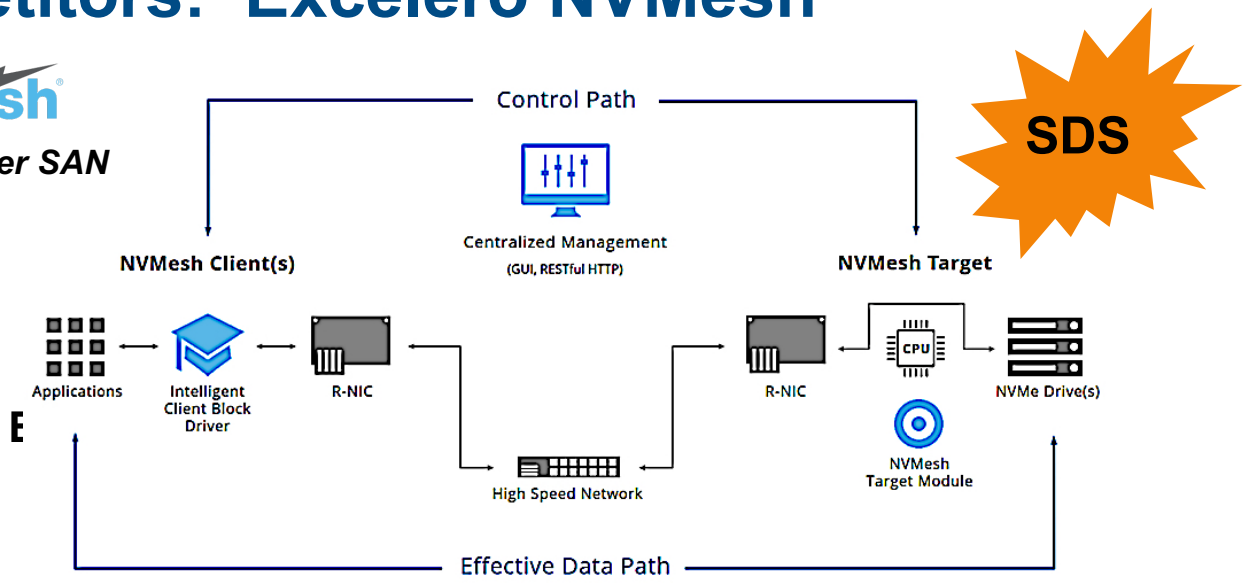
…and don't forget Pure Storage FlashArray//X

- Webscale infrastruct[illegible]o SCSI)
- Internal NVMe storag[illegible]e over Fabric
- Primary positioning is as an easily scalable "SSD" that offers the efficiencies of shared storage
- All require custom drivers on the host and some include hardware customizations (for "enterprise" features like RAID, snapshots, etc.)
- Primary workload targets include real-time big data analytics and super high performance databases
- First shipments in 2016 and industry revenues under $50M in 2017

RSF Competitors: Excelero NVMesh

NVMesh

Server SAN



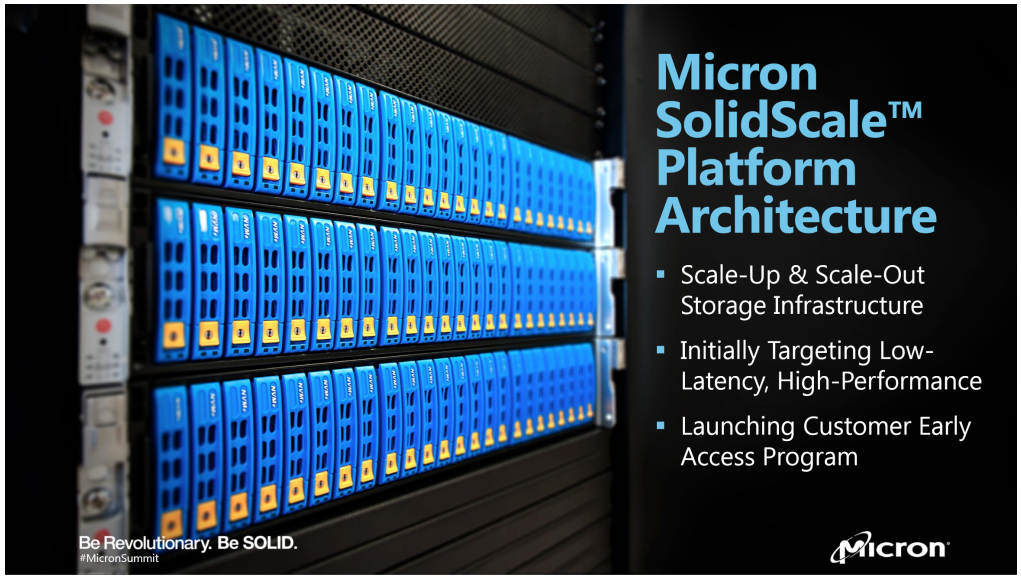
SDS

- Enables NVMe deployment at scale
- RDDA protocol for remote access
- Data services run on clients
- Logical volumes, data protection, multi pathing, tiering

- Converged/disaggregated
- SAN efficiencies/local latencies
- Runs on webscale infrastructure
- RoCE v2, Infiniband-based mesh
- Open19 Foundation member

IDC Analyze the Future

RSF Competitors: Micron SolidScale



- 24 NVMe SSDs in 2U
- Packaged with Excelero SDS
- RAID, dedupe, snapshots, replication
- Mellanox RoCE

- Announced in May 2017
- Targeted for use with real-time big data analytics, databases and VDI
- RESTful API supports easy automation

IDC Analyze the Future

The Importance of NVMe in Enterprise Storage

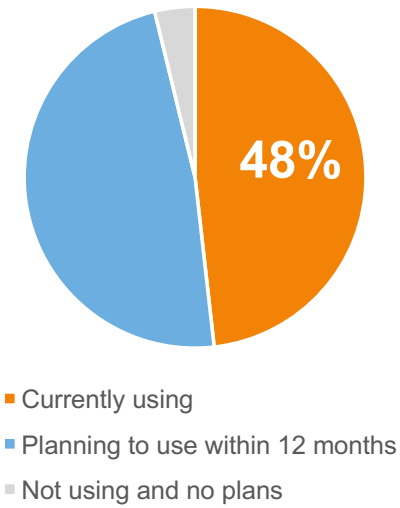
- NVMe vs SCSI advantages
- Lighter weight I/O stack optimized for memory
- Lower latencies and much higher throughput
- Supports much higher degrees of parallelism

- New workloads and data access patterns require much higher storage performance
- Real-time big data analytics need an ability to support high degrees of concurrency
- Big data exacerbates the data mobility problem

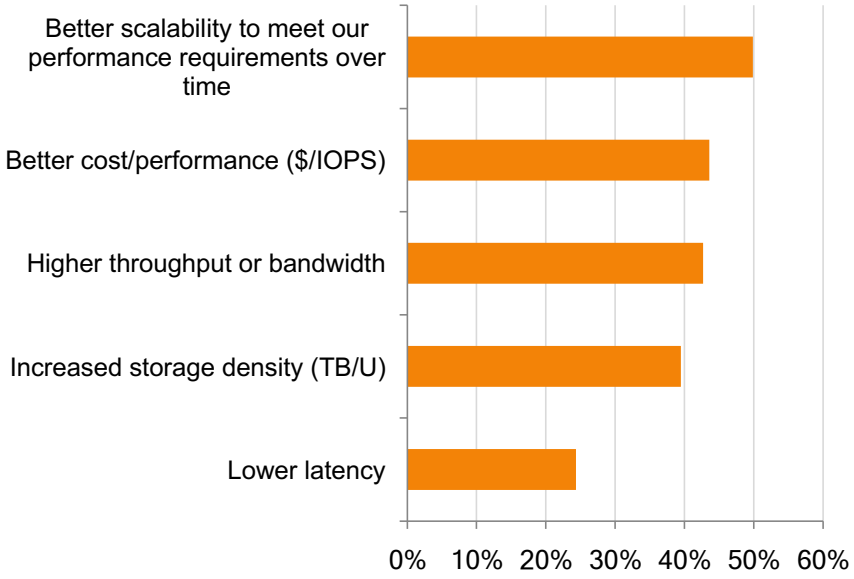
- Improved efficiencies for “at scale” computing
- Higher infrastructure densities
- NVMe interface bandwidth needed as drive sizes increase

NVMe Adoption and Drivers

PCIe or NVMe Flash



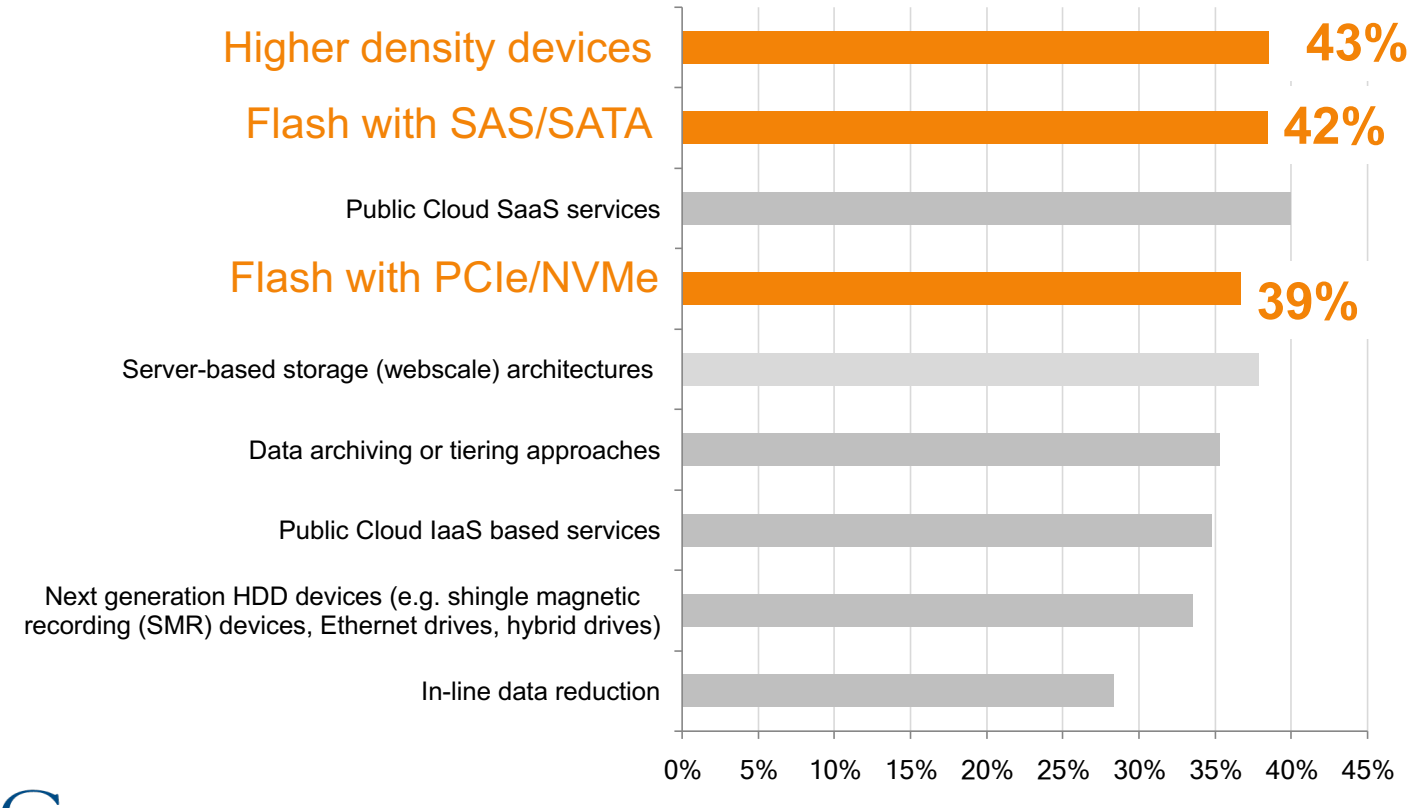
Drivers to NVMe Adoption



IDC Analyze the Future

N=804-Source: Micron Survey: Managing High Data Growth Survey, IDC, April, 2017

Flash Strategies To Manage Data Growth

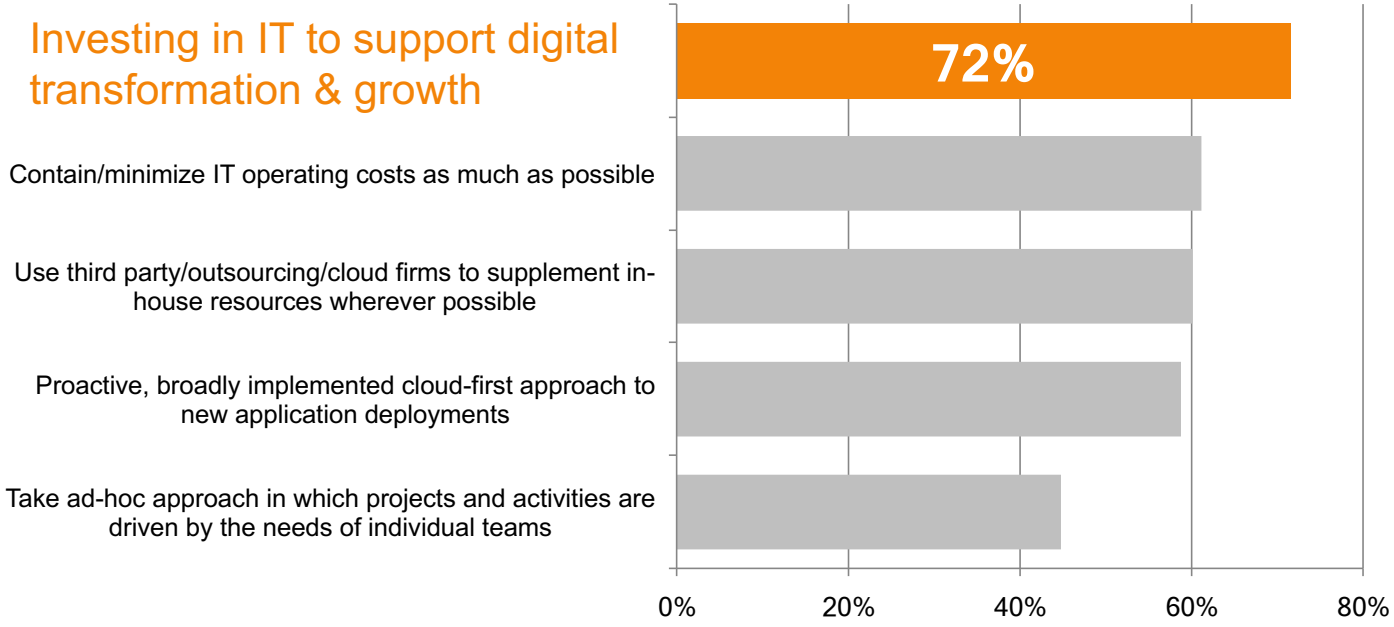


IDC Analyze the Future

N=804-Source: Micron Survey: Managing High Data Growth Survey, IDC, April, 2017

DX* Is Driving Infrastructure Strategies

Rational for IT Infrastructure Decisions

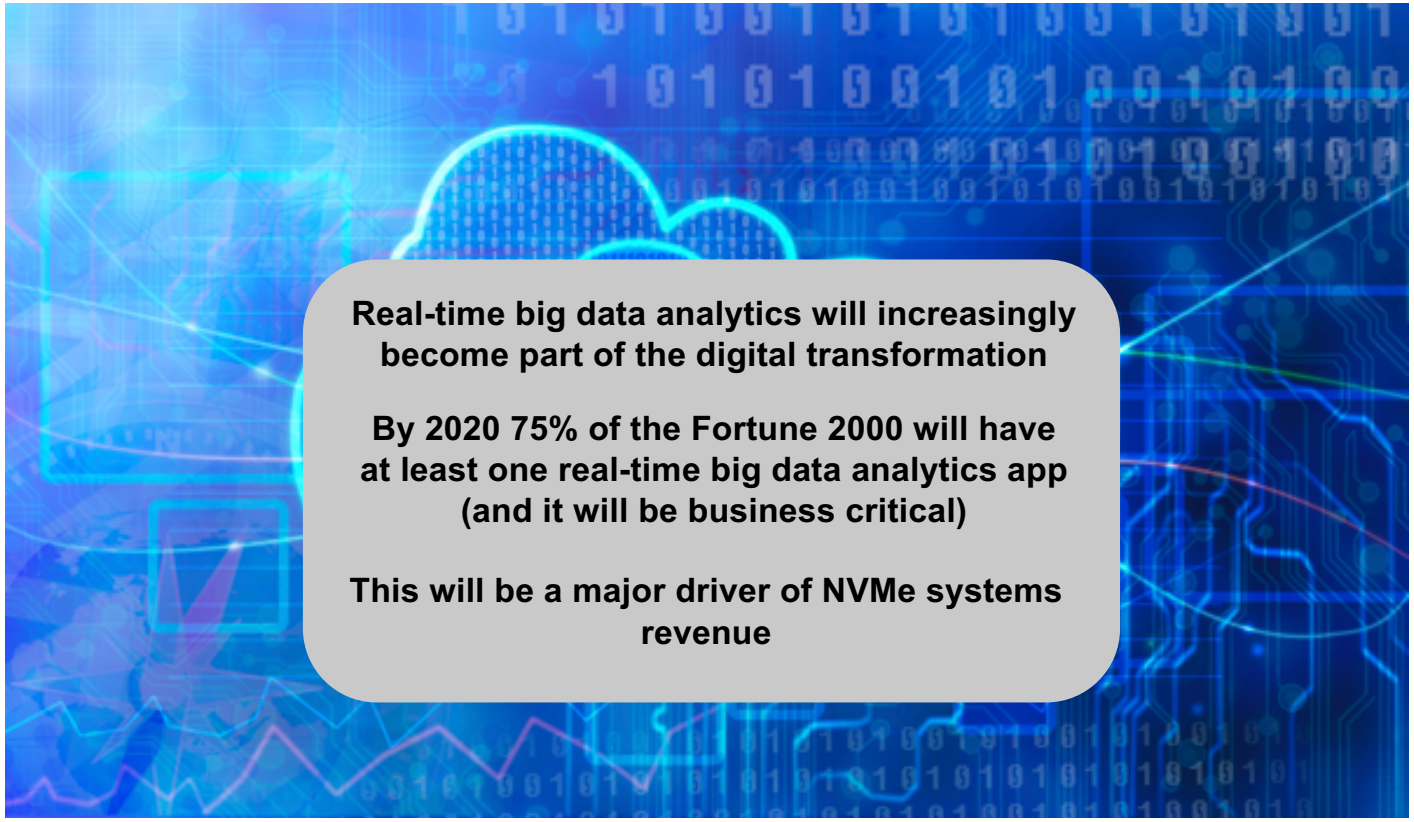


* Digital Transformation

Source: IDC Cloud Study 2017, N = 1007

IDC Analyze the Future

Real-Time Big Data Analytics



Real-time big data analytics will increasingly become part of the digital transformation

By 2020 75% of the Fortune 2000 will have at least one real-time big data analytics app (and it will be business critical)

This will be a major driver of NVMe systems revenue

Essential Guidance

- **NVMe will become the mainstream foundation technology for enterprise storage by 2020**
- **An increasing number of select workloads will require NVMe performance (starting now)**
- **There are other reasons to consider NVMe now besides just low latency**
  - High throughput, storage density/rebuild times
- **Established vendors are taking an incremental approach to NVMe integration…**
- **…but the rack scale flash architectures of the future are based on webscale designs**

IDC Analyze the Future

Thank You

Eric Burgener
Research Director
Storage
eburgener@idc.com

© IDC Visit us at IDC.com and follow us on Twitter: @IDC