SDC 20

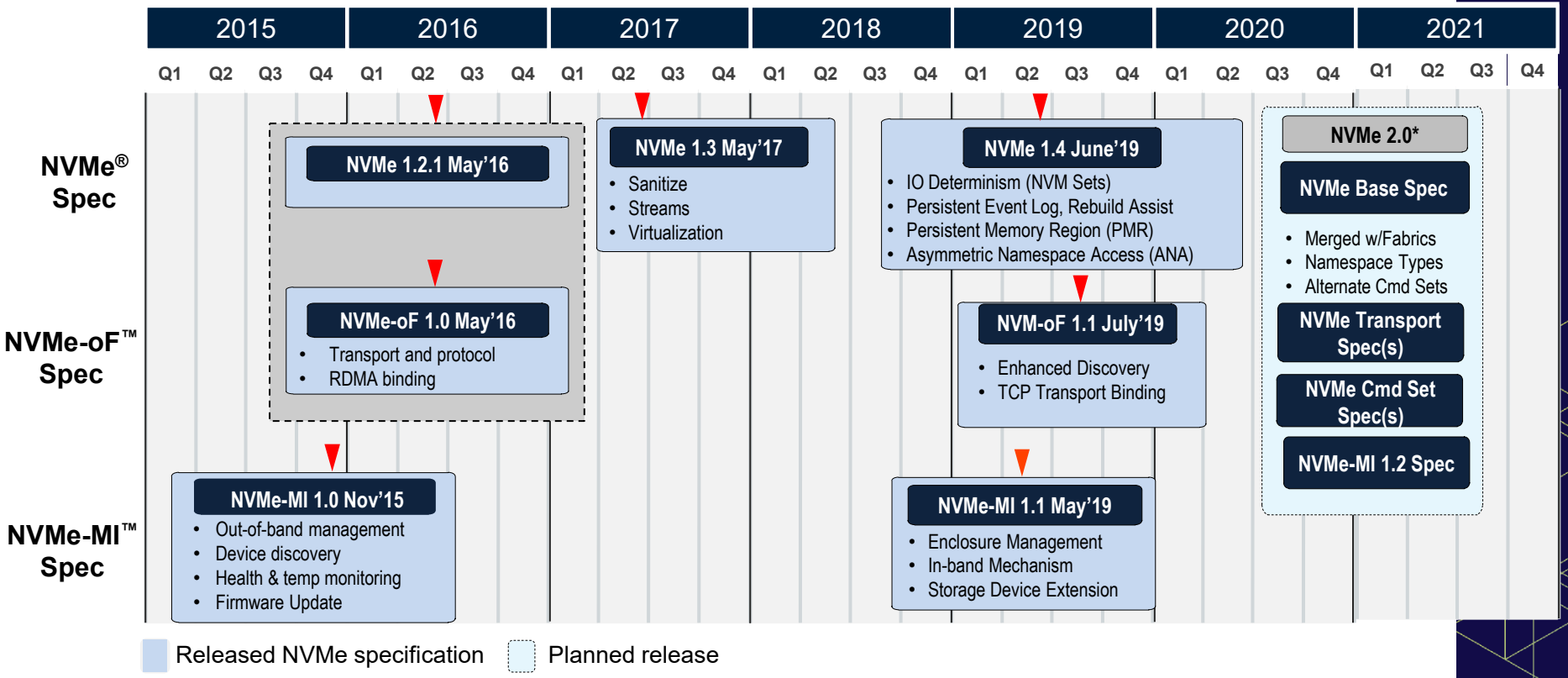
BY Developers FOR Developers

Storage Developer Conference
September 22-23, 2020

NVMe® 2.0 Specification Preview

Jonmichael Hands, Intel
Bill Martin, Samsung

NVM Express® Technology Specification Roadmap



NVMe® Specification- Cleaning Up and Bug Fixes

- NVMe® 1.4 Specification
  - TP 4042a Further Events for the Persistent Event Log
  - TP 4004b ANA Based protocol
  - TP 4005c Namespace Write Protect
- NVMe Next
  - TP 4052a Endurance Group Management
  - TP 4059a CMB Write Elasticity Status
  - TP 4065a Simply Copy Command

Ratified TP

TP that completed member review

NVMe® Specification - Enhancements

- NVMe® 1.4 Specification
  - TP 4054 CMB/PMR DMA Enhancements
- NVMe Next
  - TP 4059 CMB Write Elasticity Status
  - TP 4063 Telemetry Enhancements
  - TP 4078 Namespace Attachment Limit
  - TP 4040 Non-Data-Transfer (non-MDTS) Command Size Limits
  - TP 4047 Security Commands During Format NVM Commands
  - TP 4064 SGL Optimization
  - TP 4071 Commands and Effect Log Enhancements
  - TP 4079 Telemetry Log Size Change

Ratified TP

TP that completed member review

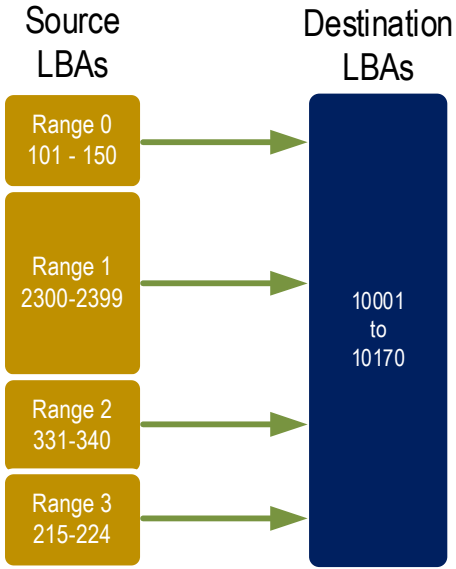
NVMe® Specification - Innovations

- NVMe® Next
  - TP 4009 ANA Domains and Partitioning
  - TP 4052 Endurance Group Management
  - TP 4065 Simple Copy Command
  - TP 4046 Command Group Control
  - TP 4055 Key per I/O
  - TP 4056 Namespace Types
  - TP 4053 Zoned Namespaces
  - TP 4015 NVMe Key Value

Ratified TP

TP that completed member review

Simple Copy Command (TP 4065)

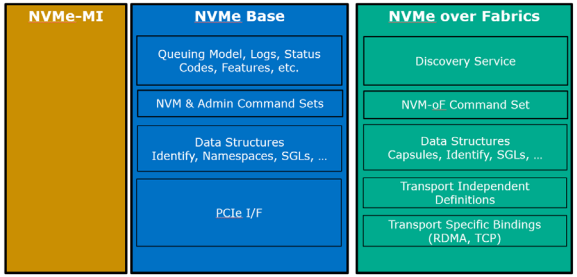


- New NVM I/O command that copies logical blocks from one or more logical block ranges to a single contiguous destination logical block range
  - Source logical block ranges described by Source Range Entries transferred from host
  - Supports protection information

Command Group Control (TP 4046)

- Defines new Lockdown admin command
  - May be used to prohibit execution of a command or modification of a feature in an NVM subsystem
    - Admin command
    - Set Feature for a specified Feature Identifier
    - Management Interface Command Set command
    - PCIe Command Set command
  - Provides interface level granularity
    - Ability to lockdown in-band, out-of-band, or both
- Once a command or feature is locked down, then it remains locked down until re-enabled by the Lockdown command or NVM subsystem power cycle

Refactoring NVMe® Specification

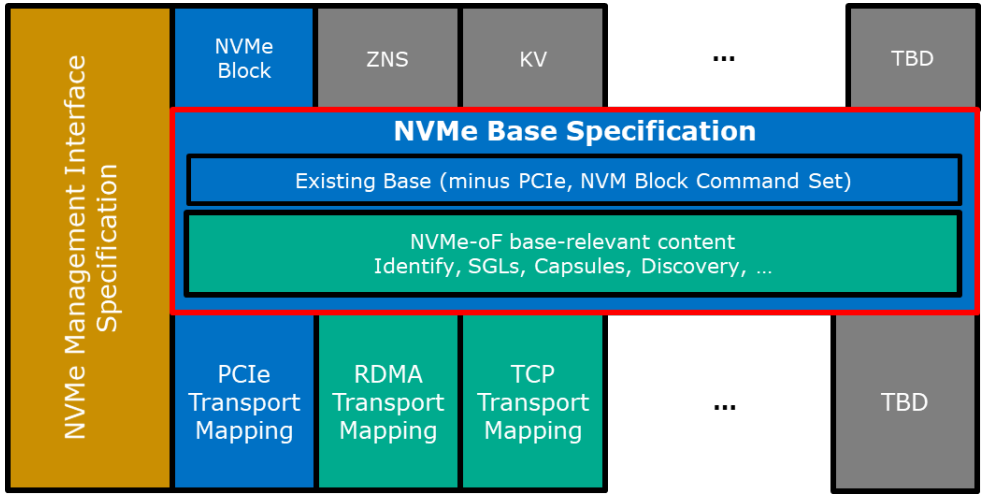


Key Aspects Driving the Refactor

- Back to the core values… Fast, Simple, Scalable
- Foster areas of innovation while minimizing impact to broadly deployed solutions
- Creating an extensible spec infrastructure that will take the industry through the next phase of growth for NVMe® technology!

Current

New Structure



New Specifications
Command Sets & Transports

NVMe® I/O Command Set Specs	Transport Command Set Specs

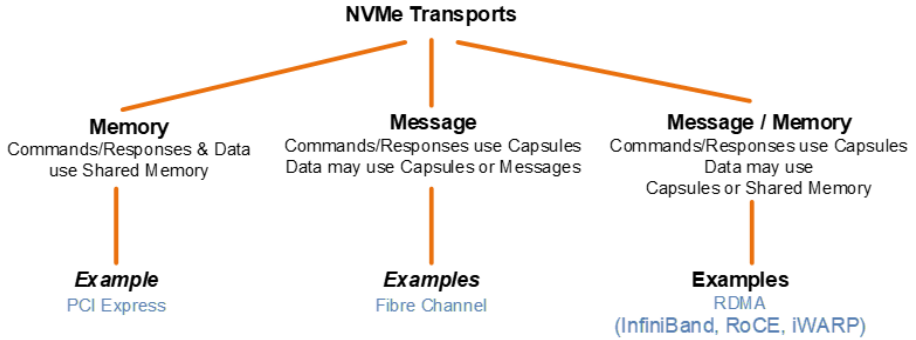
• NVM Command Set
  • Source: NVMe 1.4 Base Specification
	• PCIe Transport
	  • Source: NVMe 1.4 Base Specification

• Zoned Namespace Command Set
  • Source: TP4053
	• RDMA Transport
	  • Source: NVMe-oF™ 1.1 Specification

• Key Value Command Set
  • Source: TP4015
	• TCP Transport
	  • Source: NVMe-oF 1.1 Specification

Key Changes within the Base Specification

- Fabrics Specification integrated into the Base Spec
- Theory of Operation section enhanced with two main concepts
  - Include content for Domains, Endurance Groups, NVM Sets & Namespaces
  - Memory Based Theory (PCIe) & Message Based Theory (Fabrics)
- Created an NVM Express® Architecture section
- PCIe Registers and concepts moved to Transports Spec
- Improved organization of Controller
- Architecture, Data Structures & Features
- Data Structures section re-organized
- to only cover data and not concepts



NVM Express® Architecture Section

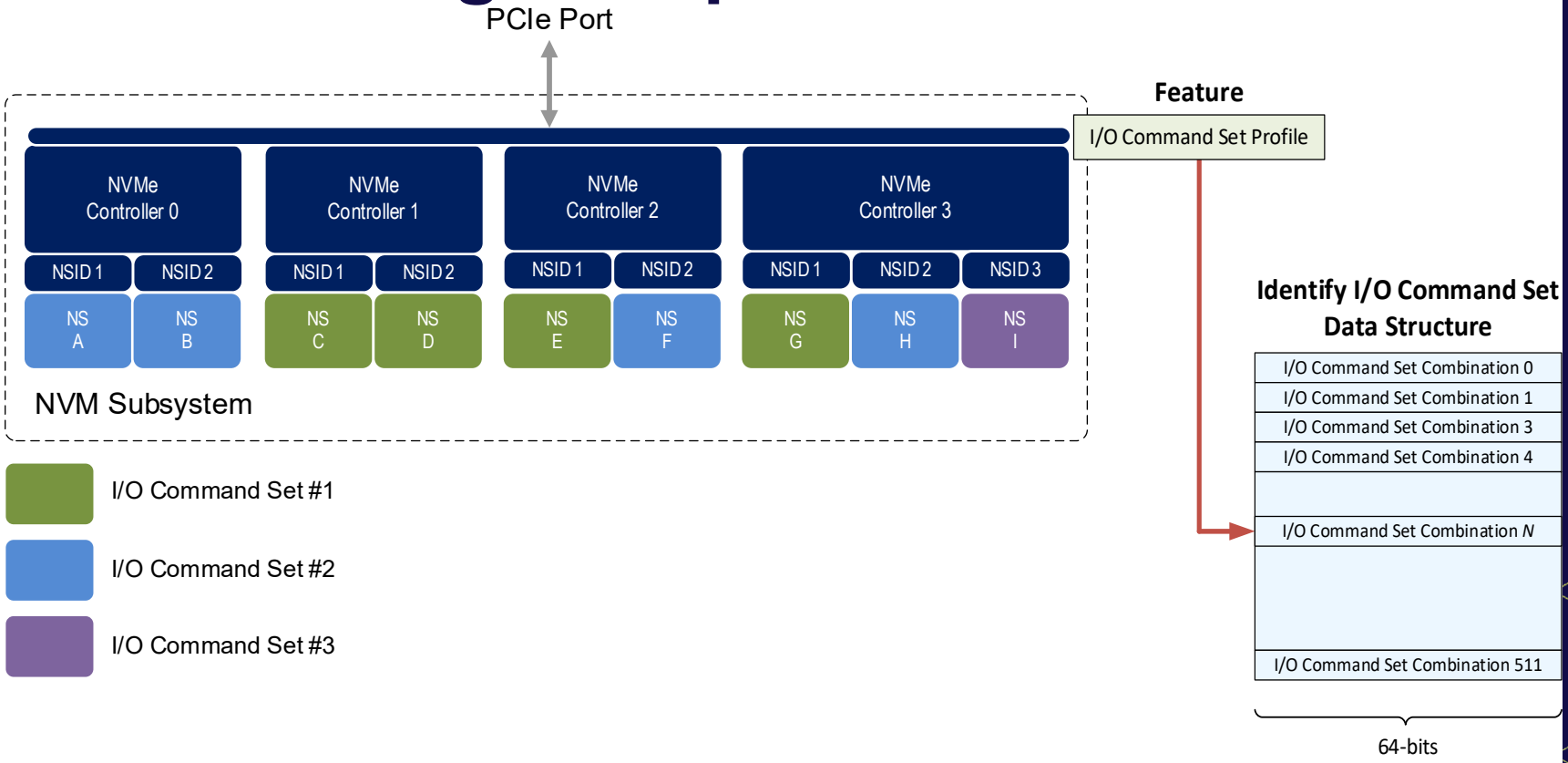
Early in the document to set context for the reader…

- NVM Controller Architecture
  - Includes Controller Model, Controller Types & Controller Properties sections
- NVM Subsystem Entities
  - Includes Namespaces, NVM Sets & Endurance Groups
- NVM Queue Models Status of Overall Execution Plan
  - Includes sections on Memory- & Message-based Queue Models & Queueing Data Structures
- Command Architecture
  - Includes Command Ordering Req's, Fused Operations, Atomic Operations & Command Arbitration
- Controller Initialization & Shutdown
  - Includes Memory- & Message-based Shutdown & Initialization
- Sections for Reset Types, Keep Alive, Privileged Actions & Firmware Updates

Status of Overall Execution Plan

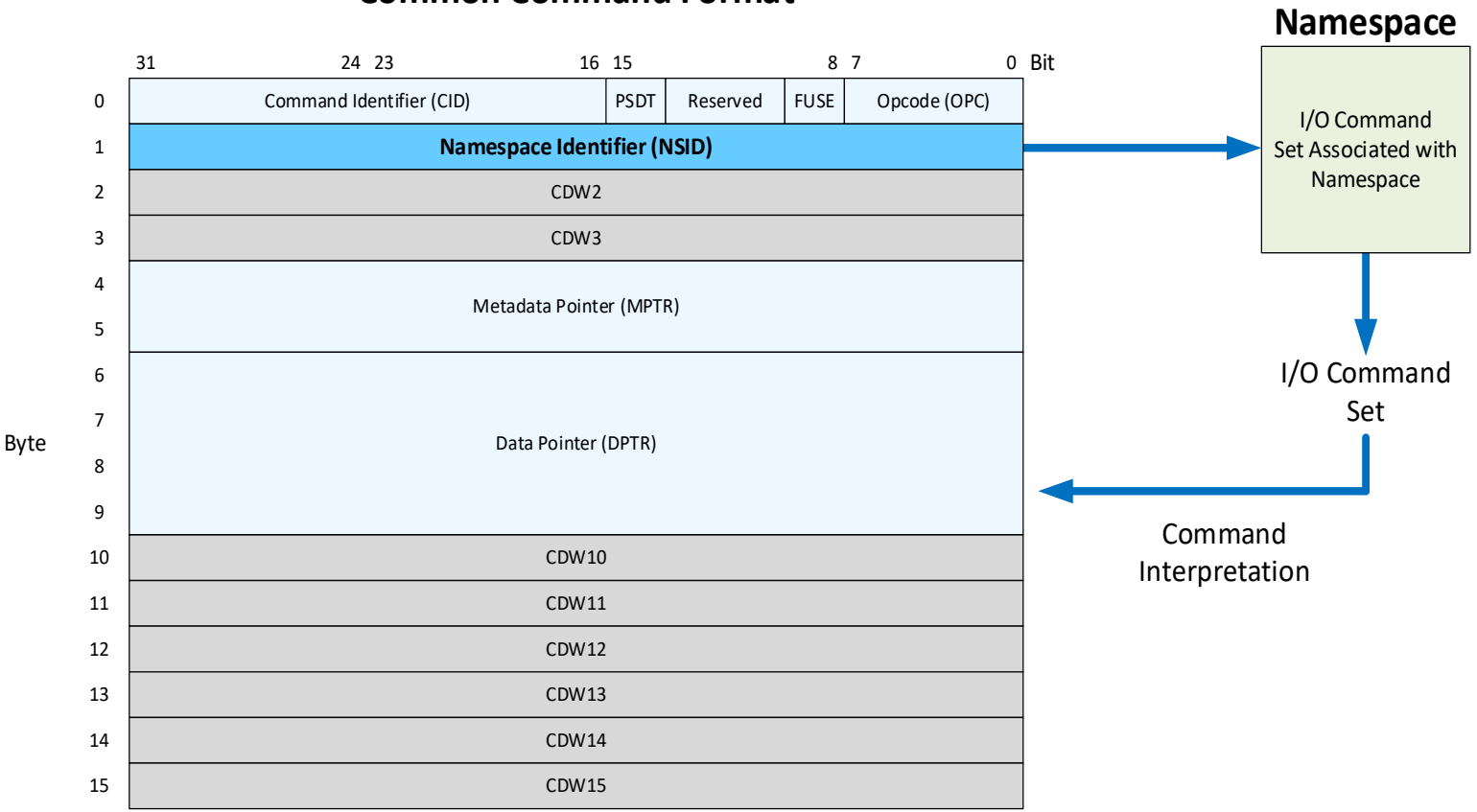
- ✓ Preparation – Create a new Outline
- ✓ Integrate the Fabrics Spec into the Base Spec
- ✓ Reorganize the Merged Spec for better readability & flow
- ✓ Add coverage of missing topics in the Theory of Operations section
- ✓ Create separate Fabrics Transport Template & Specs (PCIe, RDMA, & TCP)
- ❑ Generate the Command Set Specifications
  - ✓ Create a separated Command Set Specification Template
  - ✓ Create the NVM Command Set Specification
  - ❑ Create the initial ZNS & KV Command Set Specs based on TPs and Command Set Template
- ✓ Integrate TP4056 - Namespace Types
- ❑ Generate a “Final” set of NVMe® Specifications including:
  - ❑ Base Spec: NVMe 1.4 Base Spec, NVMe-oF™ 1.1, TP4056 & updated Theory of Operations
  - ❑ Final Transport Specs (PCIe, RDMA, TCP)
  - ❑ Final Command Set Specs (NVM, ZNS & KV)
  - ❑ Integration of all Ratified TPs into appropriate specifications
  - ❑ Aligned release of the NVMe-MI™ 1.2 Specification

Enabling Multiple Command Sets

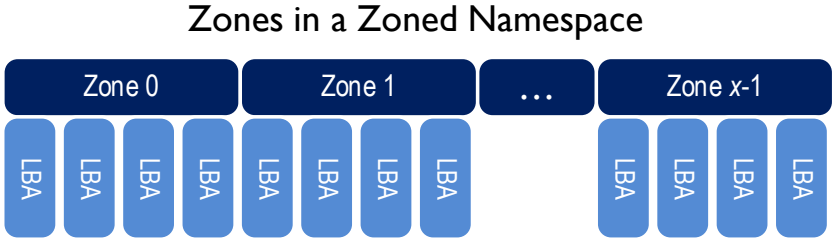


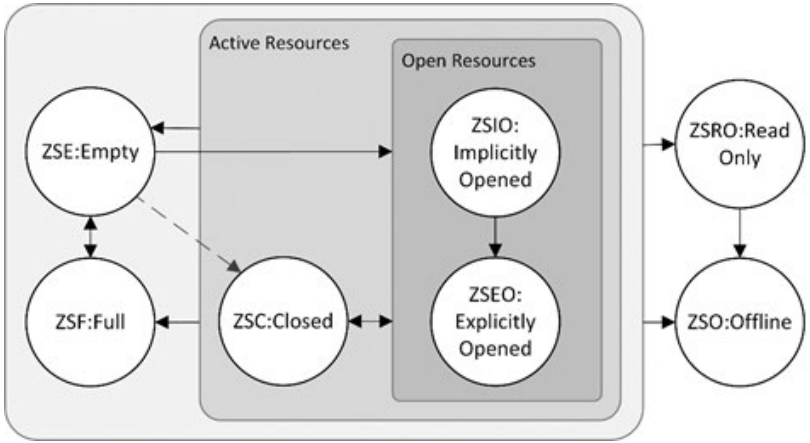
I/O Command Interpretation

Common Command Format



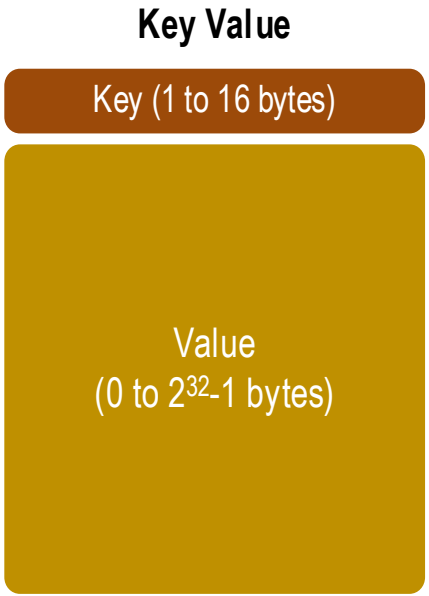
Zoned Namespaces Command Set





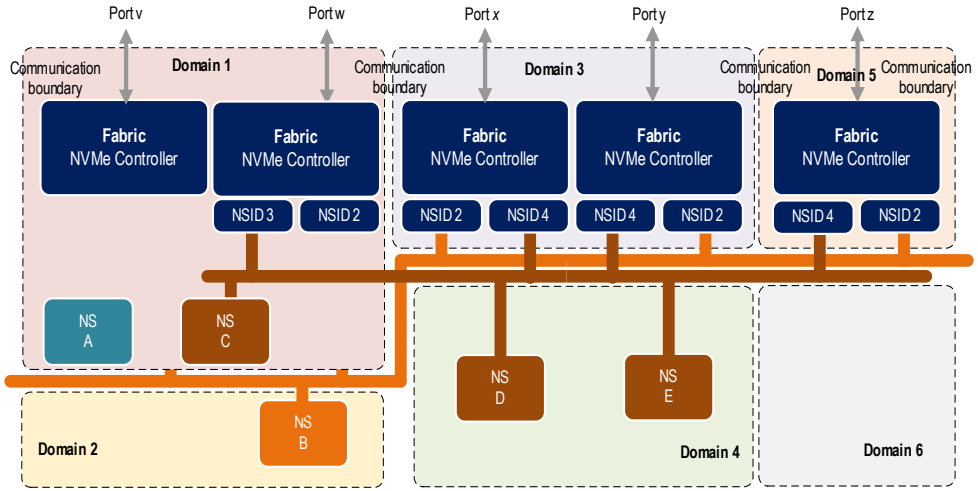
- Logical blocks are grouped into zones
  - Logical blocks are written sequentially within a zone
- State machine associated with each zone
  - Controls operational characteristics of each zone
  - State transitions may be explicitly controlled by the host or implicitly by host actions
- Benefits
  - Reduced write amplification
  - Reduced overprovisioning
  - Reduced memory on Storage Device (DRAM)

Key Value Command Set



Command	Description
Delete	Delete Key and Value associated with a specified Key
List	Lists Keys that exist in a Key Value Namespace starting at a specified Key
Retrieve	Retrieve Value associated with a specified Key
Exist	Returns status indicating whether a Key Value exists for a specified Key
Store	Stores a Key Value to a Key Value Namespace

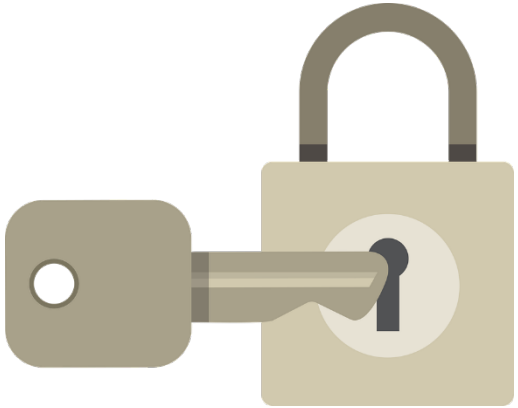
Domains and Partitions (TP 4009)



- An NVM subsystem may represent a warehouse-scale storage system
- A warehouse-scale storage system may be constructed from multiple Domains
  - Capacity, controllers, and ports, may be partitioned among Domains
  - Domains may be added, removed, reconfigured, partitioned, or fail
- NVMe® technology now defines Domains as an architectural element

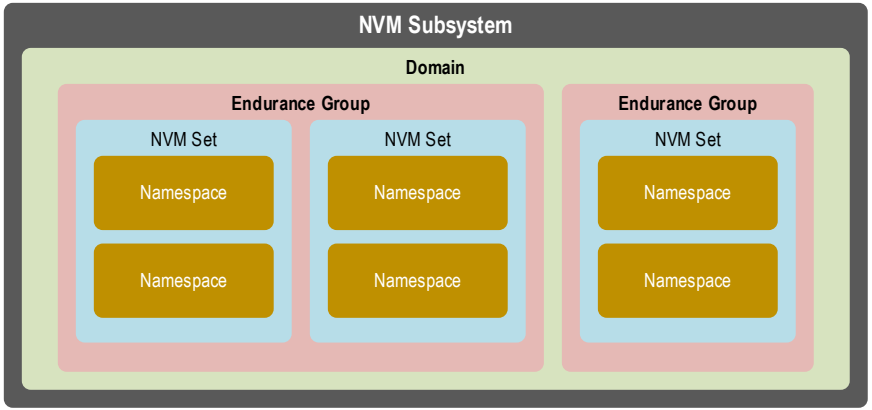
Key Per I/O (TP 4055)

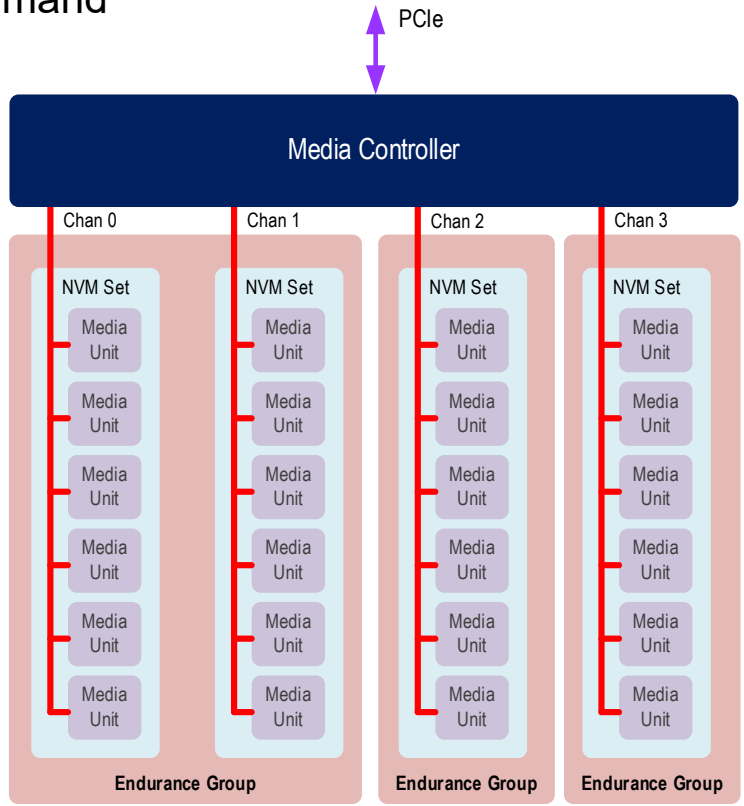
- Allows a unique key to be used on a per I/O basis to encryption/decrypt logical blocks stored in a Namespace
  - Key Tag in command specifies encryption key to use
    - NVM Subsystem supports up to 64K keys
  - Configuring keys and management of Key Tags will be defined in the TCG

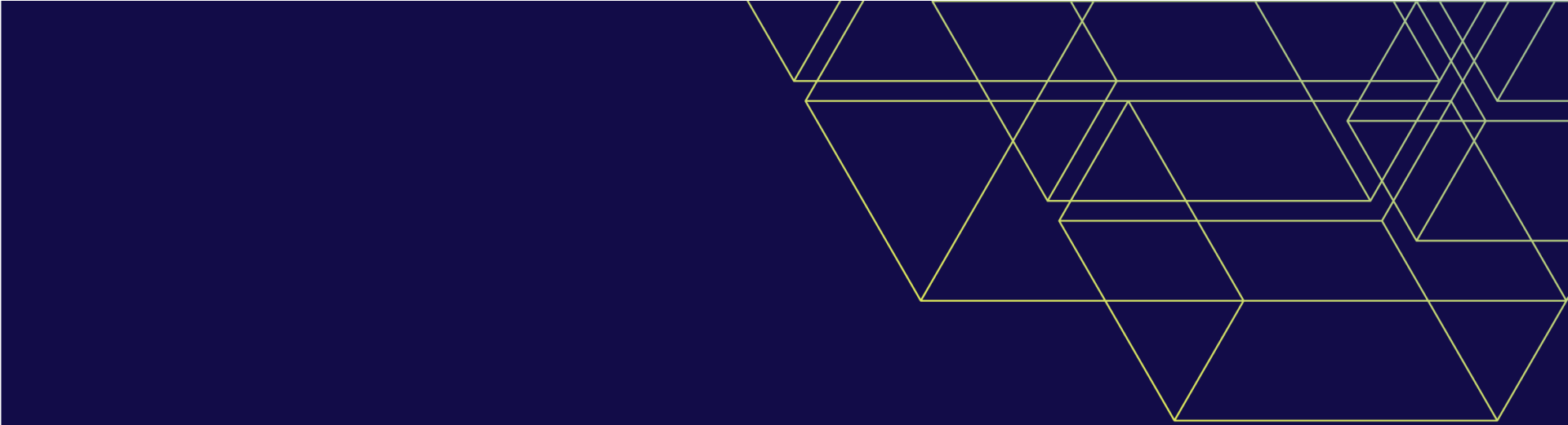


Endurance Groups (TP 4052)

- Defines new Capacity Management admin command
  - Creation/deletion of NVM Sets
  - Creation/deletion of Endurance Groups
  - Allocation of Media Units to Endurance Groups
  - Allocation of Media Units to NVM Sets







Please take a moment
to rate this session.

Your feedback matters to us.