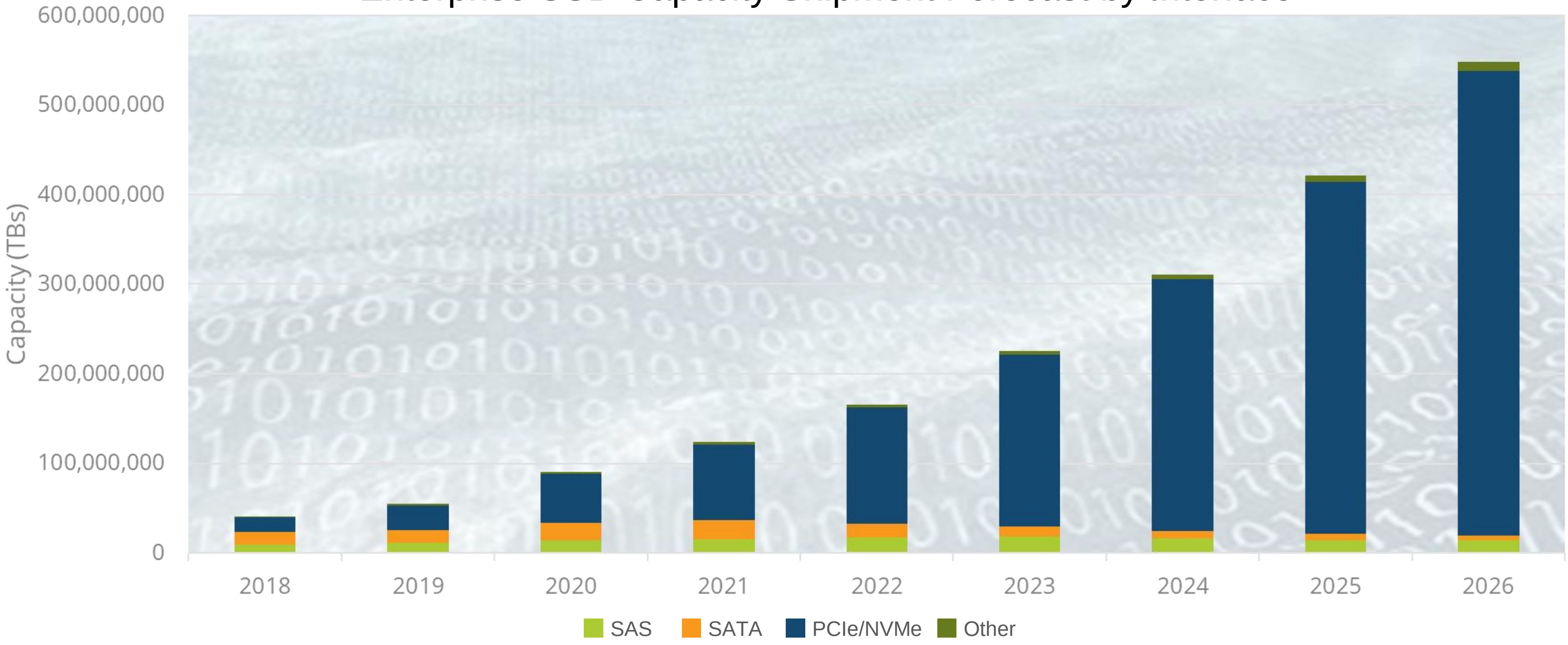
NVM Express State of the Union

Peter Onufryk, Intel Fellow

NVMe® Technical Workgroup Chair

NVMe® Specifications – The Language of Storage

Enterprise SSD Capacity Shipment Forecast by Interface



Source: Worldwide Solid-State Storage Forecast, 2022–2026 (May 2022) IDC #US47831722

NVMe® Technology Powers the Connected Universe

Units (Ku)*	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Enterprise	364	749	1,069	2,045	5,183	7,007	8,705	12,108	14,652	17,589
Cloud	2,051	3,861	10,369	12,276	19,105	22,981	27,916	32,469	40,080	49,147
Client	33,128	48,951	82,587	143,236	226,221	307,518	368,978	446,958	482,792	522,273

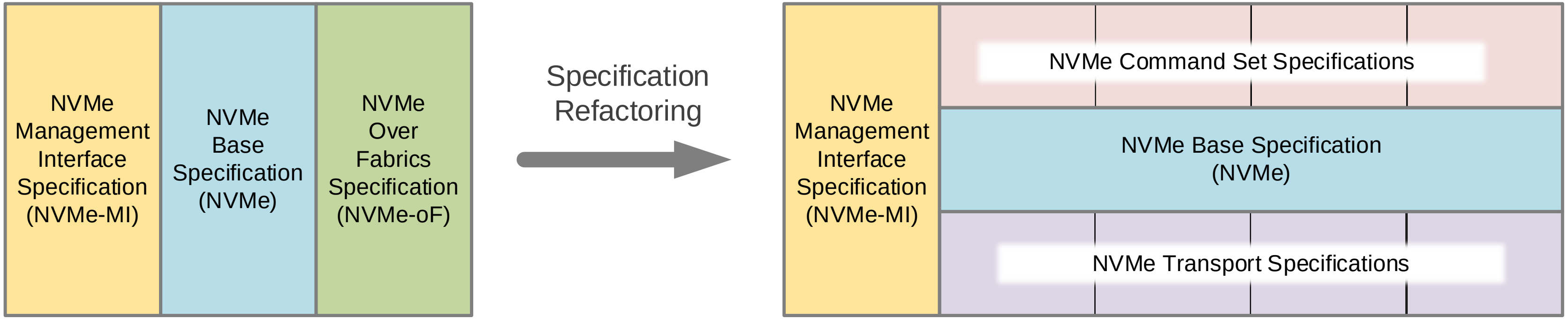
* Data and projections provided by Forward Insights Q2’21 & Q1’22

Consumer ● Client ● Embedded ● Enterprise ● Cloud

Cell Phones Tablets Laptops Desktops Storage Arrays Data Centers

NVMe® Specification Refactoring

- Why Refactor?
  - Ease development of NVMe-based technology
  - Enable rapid innovation while minimizing impact to broadly deployed solutions
  - Create extensible spec infrastructure that enables the next phase of growth for NVMe technology



NVMe Management Interface Specification (NVMe-MI)

NVMe Base Specification (NVMe)

NVMe Over Fabrics Specification (NVMe-oF)

Specification Refactoring

NVMe Management Interface Specification (NVMe-MI)

NVMe Command Set Specifications

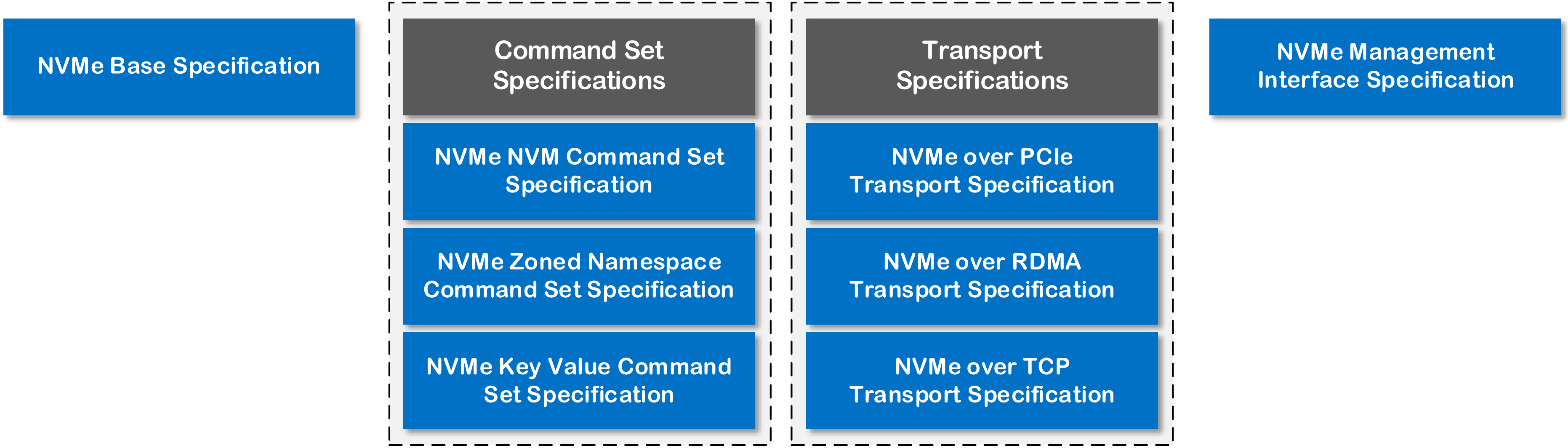
NVMe Base Specification (NVMe)

NVMe Transport Specifications

NVMe 1.x specifications

NVMe 2.0 specifications

NVMe® 2.0 Family of Specifications



NVMe Base Specification

Command Set Specifications

- NVMe NVM Command Set Specification
- NVMe Zoned Namespace Command Set Specification
- NVMe Key Value Command Set Specification

Transport Specifications

- NVMe over PCIe Transport Specification
- NVMe over RDMA Transport Specification
- NVMe over TCP Transport Specification

NVMe Management Interface Specification

NVMe 2.0 specifications were released on June 3, 2021
Refer to nvmexpress.org/developers

Activity Since Release of NVMe® 2.0 Family of Specifications*

New Authorized Technical Proposals

27

Ratified Technical Proposals

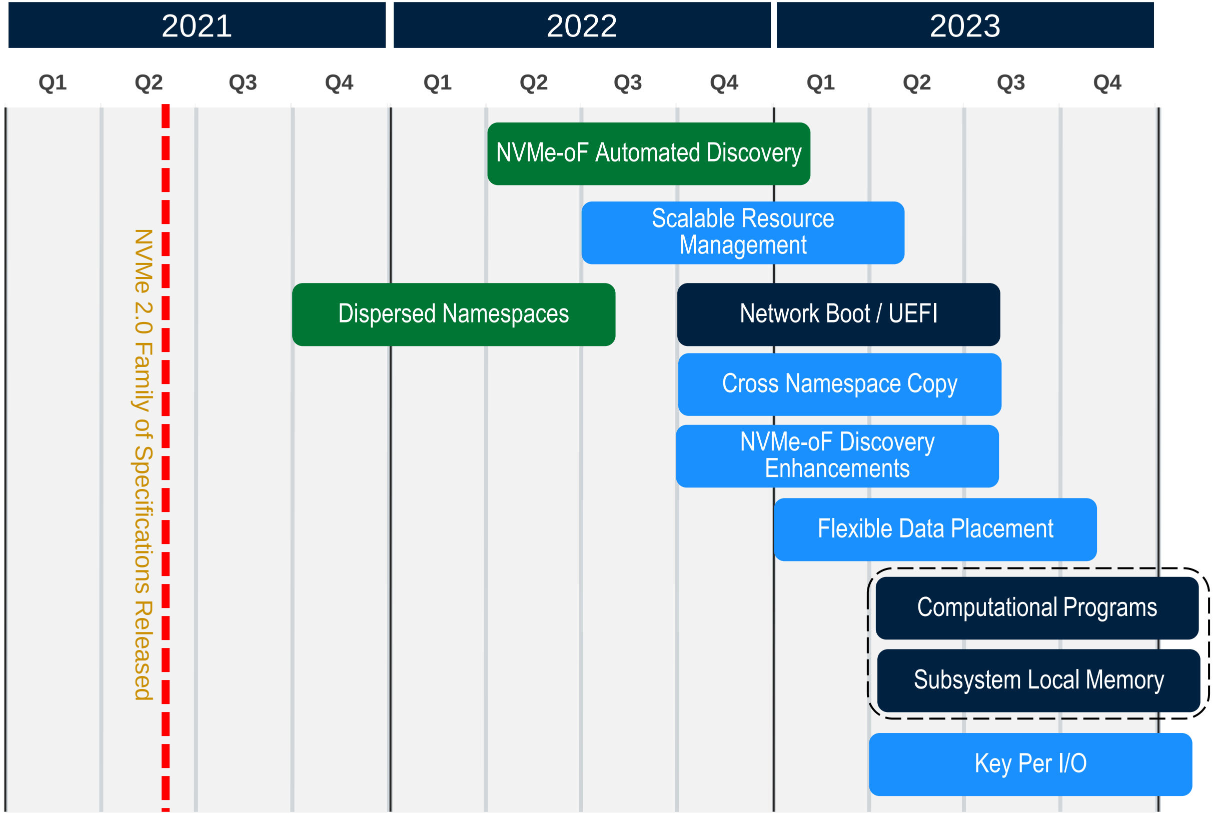
30

Ratified ECNs

5

* Activity as of 5/21/2022

NVMe® Specifications Feature Roadmap



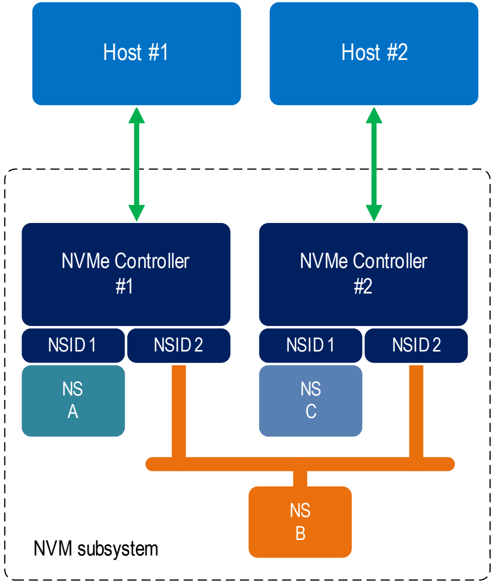
🟩 **Ratified Feature** (left edge indicates ratification quarter)

🟦 **Planned Feature** (left edge indicates planned ratification quarter)

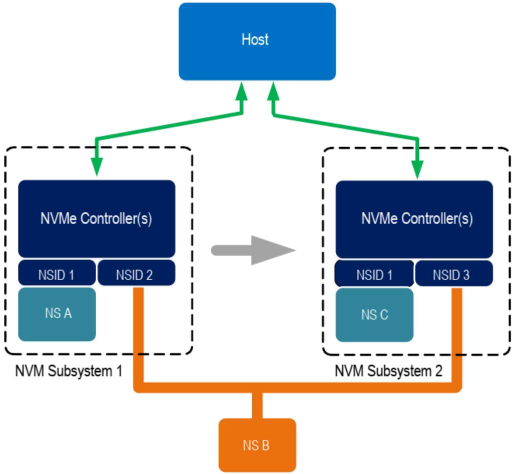
■ **Planned New Specification** (left edge indicates planned ratification quarter)

Dispersed Namespaces

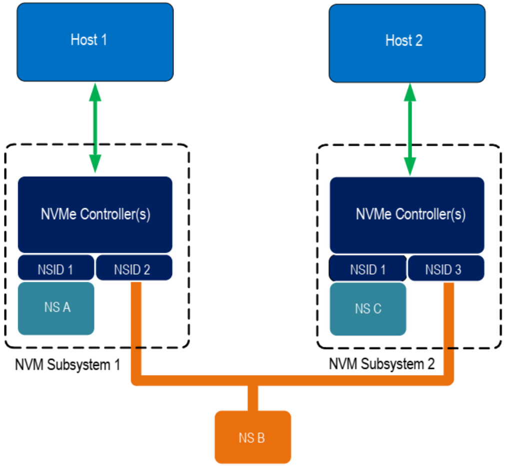
- Background
  - An NVM subsystem includes one or more controllers, zero or more namespaces, and one or more ports
  - Controller is the interface between host and NVM subsystem
  - Namespace is a formatted quantity of non-volatile memory
- A dispersed namespace is a shared namespaces that may be concurrently access by controllers in two or more NVM subsystems
  - Log page that provides a list of NQNs for all NVM subsystems that contain controllers able to accesses a dispersed namespace
  - An NVM subsystem may support reservations on dispersed namespaces



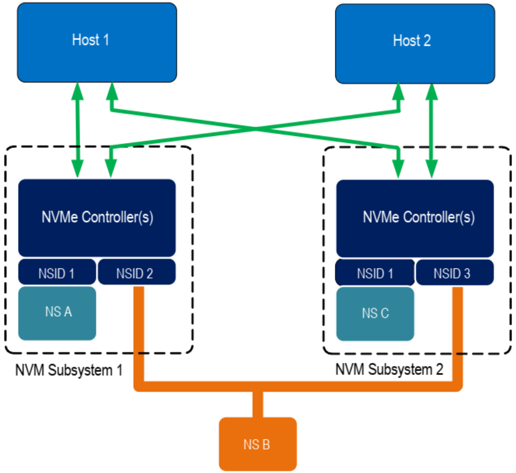
Dispersed Namespaces Applications



Online Data Migration



Data Replication



High Availability Data Replication

NVMe-oF™ Discovery Enhancements

- NVMe-oF Automated Discovery

  1. Automated Discovery of NVMe-oF Discovery Controllers for IP Networks (TP 8009 - ratified)
  2. NVMe-oF Centralized Discovery Controller (TP 8010 - ratified)

- NVMe-oF Discovery Enhancements

  3. Subsystem Driven Zoning with Pull Registrations (TP 8016 – in development)

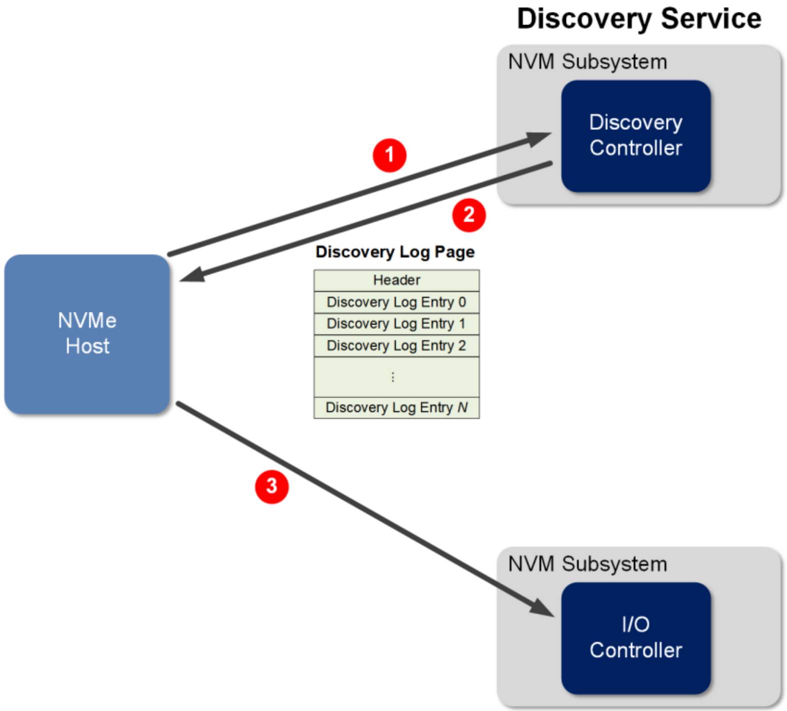
- All three discovery enhancements are only applicable for IP-based fabric transports

TP 8009

Automated Discovery of NVMe-oF™ Discovery Controllers for IP Networks

- Simplifies provisioning of Hosts by allowing them to locate NVMe®/TCP Discovery controllers

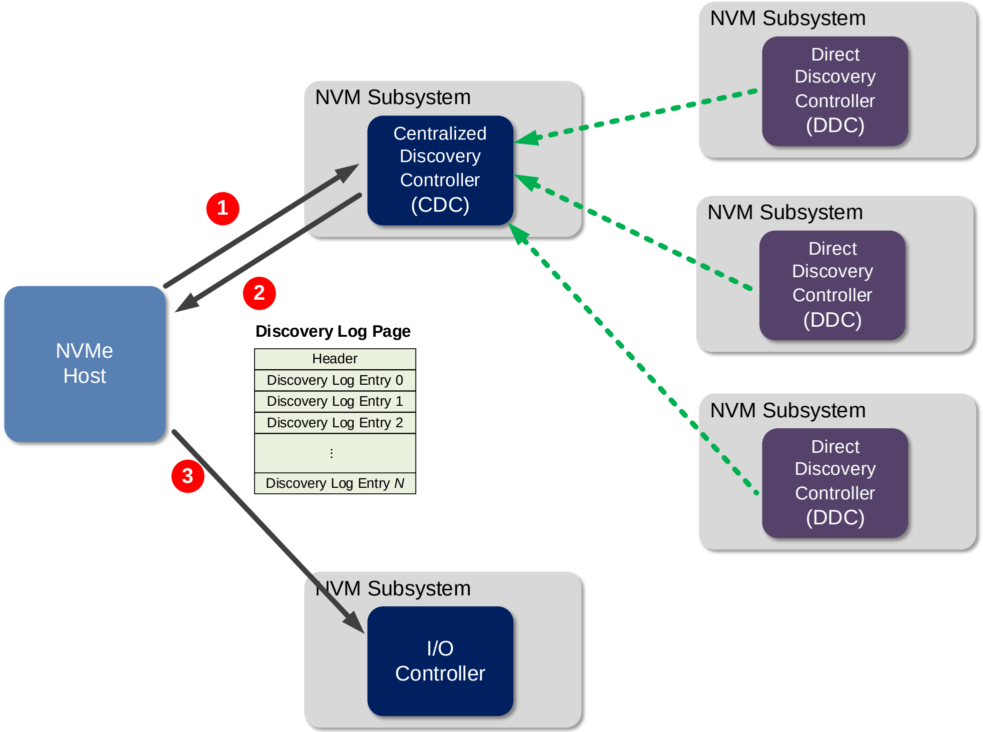
- IP Address of a Discovery controller may be determined by:
  - Administrative configuration
  - Means outside the specification
  - New capability using Domain Name System Service Discovery (DNS-SD) record



TP 8010

NVMe-oF™ Centralized Discovery Controller

- Enable discovery information to be consolidated and retrievable from a single Discovery Service
  - **Centralized Discovery Controller (CDC)**: a Discovery controller that reports discovery information registered by Direct Discovery Controllers and hosts
  - **Direct Discovery Controller (DDC)**: a Discovery controller capable or registering discovery information with a CDC
- A DDC registers with a CDC by one of the following methods
  - A push registration using a Discovery Information Management command
  - Notifying the CDC that a pull registration is required
  - Administration configuration


NVM Subsystem
Direct Discovery Controller (DDC)
NVM Subsystem
Centralized Discovery Controller (CDC)
NVM Subsystem
Direct Discovery Controller (DDC)
NVM Subsystem
Direct Discovery Controller (DDC)
1
2
3
NVMe Host
NVM Subsystem
I/O Controller

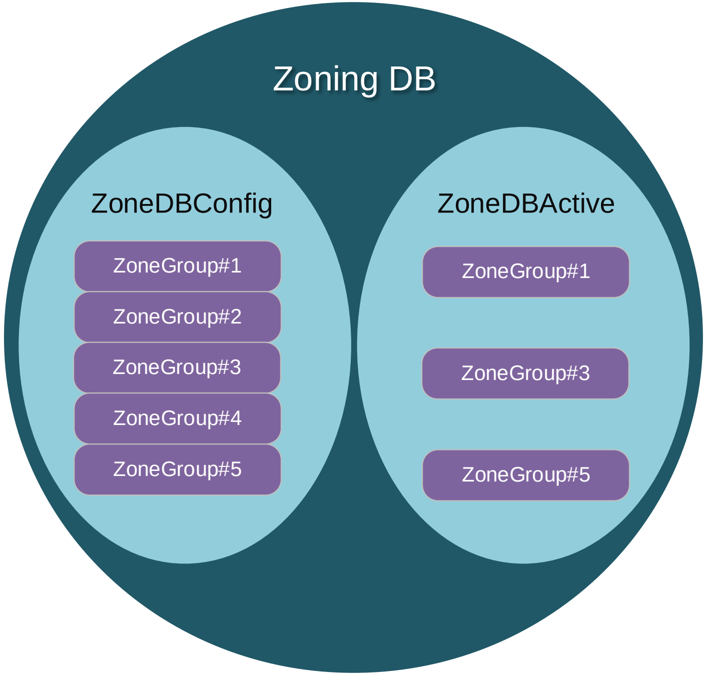
Discovery Log Page

Header
Discovery Log Entry 0
Discovery Log Entry 1
Discovery Log Entry 2
⋮
Discovery Log Entry *N*

TP8010
TP8016

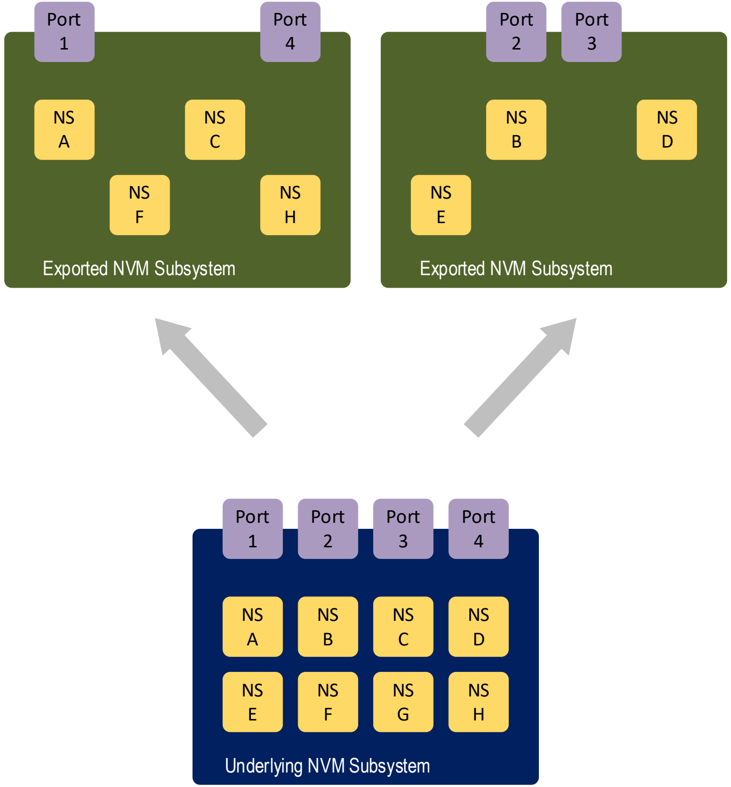
Fabric Zoning and Pull Registrations

- NVMe® architecture adds support for Fabric Zoning
  - Using Fabric Zoning a Centralized Discovery Controller (CDC) may filter Discovery Log Page information so that a host only has access to namespaces allocated to the host
- A ZoneGroup is a set of access control rules enforced by the CDC
  - A ZoneGroup contains Zones
  - A Zone is the unit of access control and members of the same Zone are allowed to communicate between each other
- Zoning database (ZoneDB) is maintained by CDC
  - ZoneDBConfig – List of configured ZoneGroups
  - ZoneDBActive – List of enforced ZoneGroups
- A DDC may provide Fabric Zoning formation to a CDC using push or pull registrations



Scalable Resource Management

- Defines a standard framework to dynamically construct, configure, and provision “Exported” NVM subsystems from underlying physical resources in an “Underlying” NVM subsystem
- New Admin Commands that enable
  - Creation and management of an Exported NVM subsystem
  - Manage Exported namespaces
  - Manage Exported ports
- Ability to manage host access to an Exported NVM subsystem using an “Allowed Host List”



Network Boot / UEFI

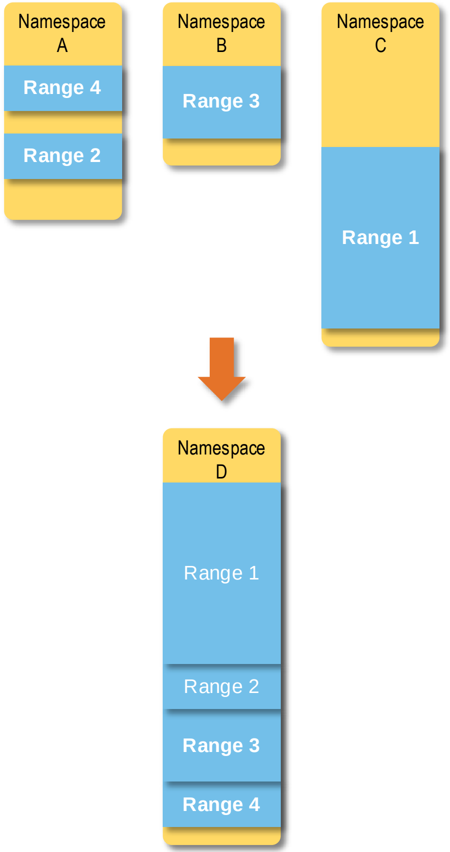
TP8012
TP4126

- NVMe-oF™ hosts require a HostNQN and HostID
  - Currently HostNQN and HostID needs to be configured by an administrator
  - This feature specifies how to construct a default HostNQN and HostID from a platform identifier (SMBIOS system UUID)
- New NVM Express® Boot Specification
  - The specification defines construct and guidelines for booting from NVMe® technology
  - While the specification covers all transports, the current specification only describes mechanisms for NVMe/TCP technology

TP4130

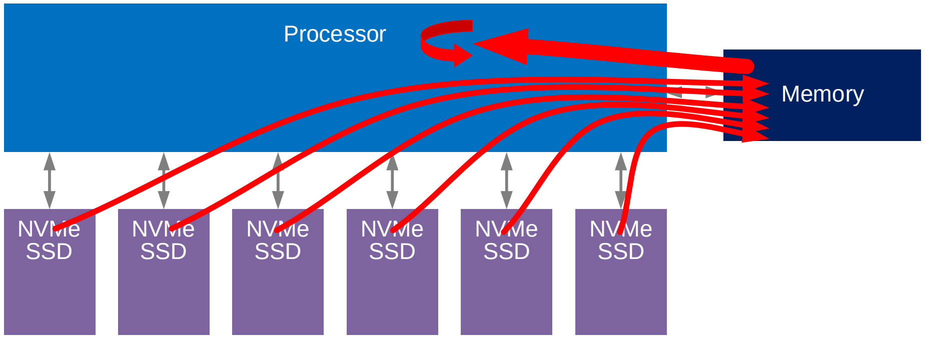
Cross Namespace Copy

- Copy command enhancement to copy data across namespaces
  - "Original" Copy Command
    - One or more source logical blocks ranges in a namespace to a single contiguous destination logical block range in the same destination namespace
  - "Enhanced" Copy Command
    - One or more source logical blocks in one **or more namespaces** to a single consecutive destination logical block range in a **destination** namespace
- Copy command does not reformat data
  - Logical block data and metadata format must be the same
  - End-to-End Data Protection type and size must be the same
  - Logical Block Storage Tag Mask and Storage Tag Size must be the same

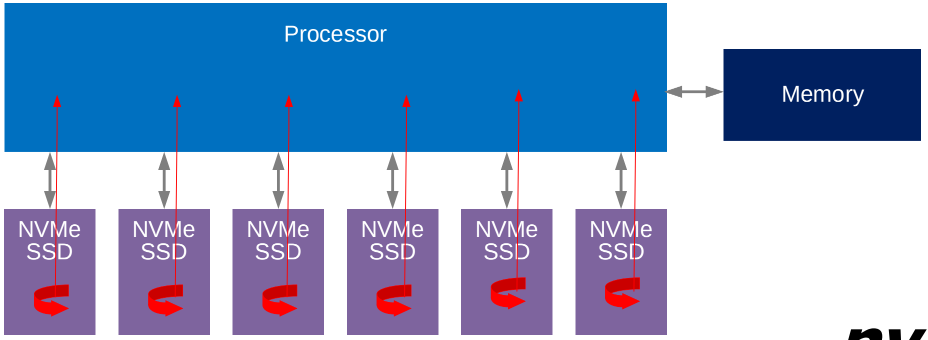


The Promise of Computational Storage

- **Higher performance and reduced latency** due to multiple SSDs operating in parallel
- **Reduced power** due to less data movement
- **Higher performance and reduced latency** due to elimination of processor I/O and memory bottlenecks



Data Processing in Main Processor

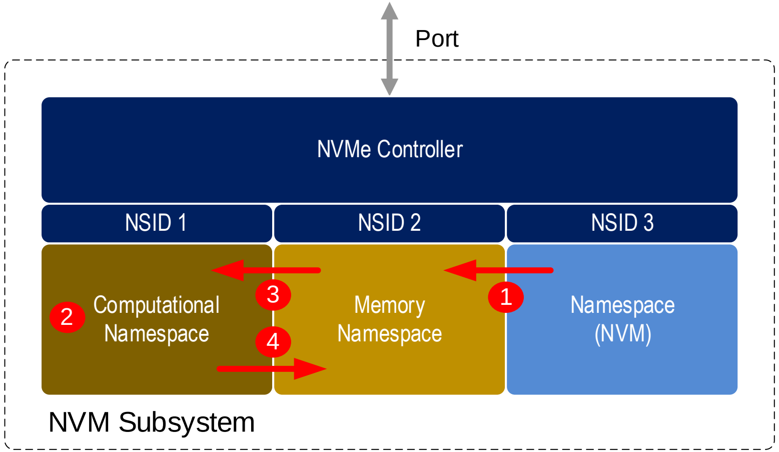


Data Processing in NVMe SSDs (Computational Storage)

Computational Programs

TP4091
TP4134

- Standardized framework for computational storage
- New command set for operating on Computational Namespaces
  - Fixed function programs
  - Downloadable eBPF programs
    - Used by Linux to run sandboxed programs
    - Vendor Agnostic
    - Widely supported (e.g., LLVM)



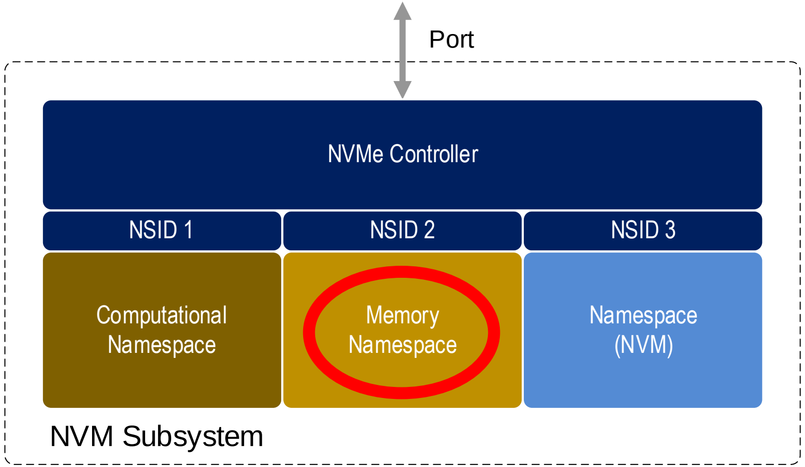
Example Operation:

1. Read data from NVM namespace into memory namespace
2. Execute program associated with computational namespace
3. Program reads data from memory namespace
4. Program stores result into memory namespace

TP4091
TP4134

Subsystem Local Memory

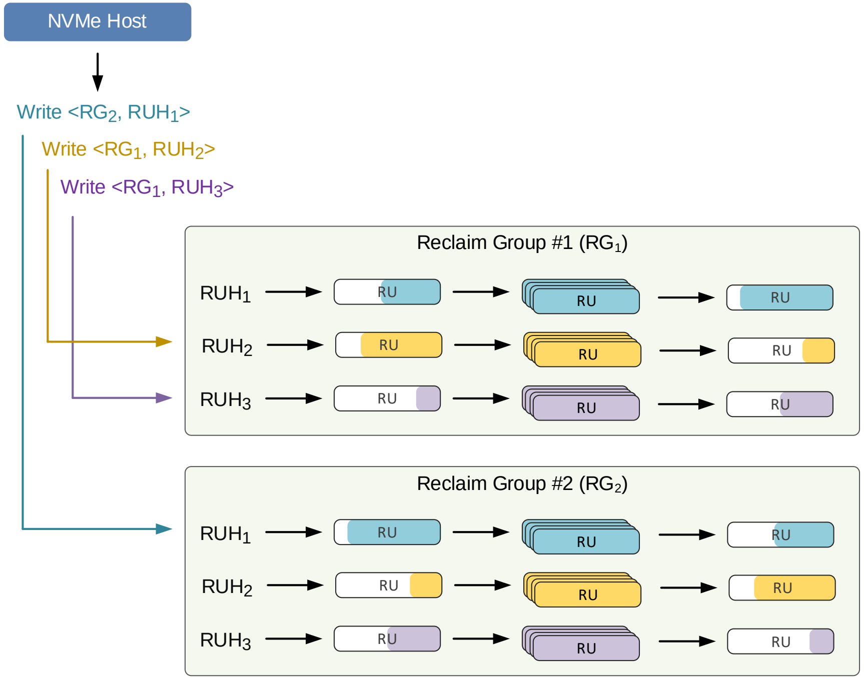
- eBPF operates on byte addressable memory
- Memory Namespaces and Memory command set
  - Required for computational programs but is new general NVMe® architectural element
  - Mechanism to copy to and from any other type of NVM namespace to memory namespace



TP4146

Flexible Data Placement

- Enhancement to the NVM Command Set to enable host guided data placement
- Reclaim Unit (RU) is a unit of NVM storage that may be independently read, written, and erased
- A Reclaim Groups (RG) is an independent collection one or more RUs
  - Limited interference between RGs
  - Each RG has one or more Reclaim Unit Handles (RUH) that each point to an RU
- Data Placement Directive allows host to specify RG and RU of where to place written data



Key Per I/O

- Self encrypting drives perform encryption on LBA ranges within namespaces
- Key per I/O provides dynamic fine grain encryption control by indicating which encryption key to use per I/O
  - Assigning an encryption key to a sensitive file or host object
  - Easier support of General Data Protection Regulation (GDPR)
  - Easier support of erasure when data is spread and mixed with other data that should be preserved (e.g., RAID and erasure coding)
- Mechanisms to download and manage keys are outside the scope of the specification
  - Keys are stored in volatile memory and are lost when powered off
- Liaison agreement between NVM Express® and TCG Storage Work Group
  - Ratification of TP will occur when work in both organizations has been completed

Summary

- NVMe® technology has succeeded in unifying client, cloud, and enterprise storage around a common architecture and adoption continues to grow

- Following the refactoring that created the NVMe 2.0 family of specifications, NVMe architecture is focusing on communicating new features and capabilities and not on specification releases
  - Technical Proposals are publicly released when ratified and may be immediately implemented

- The NVMe technical community continues to maintain and enhance existing specifications while developing new innovations
  - 27 new Technical Proposals authorized
  - 30 Technical Proposals ratified
  - 5 ratified ECNs

nvm EXPRESS®

Architected for Performance