The Evolution and Future of NVMe™

David Allen
NVMe™ Board Member and Seagate's Senior Dir. Product Marketing

Dr. J Metz
NVMe™ Board Member and R&D Engineer Cisco

What this presentation is… and isn't

This presentation is:

- A cursory overview of the status of the development of NVMe™, NVMe Management Interface, and NVMe over Fabrics™

This presentation is not:

- Comprehensive
- Carved in stone

Some of these topics are subject to change!

About NVM Express

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

NVM Express Base Specification

The register interface and command set for PCI Express attached storage with industry standard software available for numerous operating systems. NVMe™ is widely considered the defacto industry standard for PCIe SSDs.

NVM Express Management Interface (NVMe-MI™) Specification

The command set and architecture for out of band management of NVM Express storage (i.e., discovering, monitoring, and updating NVMe™ devices using a BMC).

NVM Express Over Fabrics (NVMe-oF™) Specification

The extension to NVM Express that enables tunneling the NVM Express command set over additional transports beyond PCIe. NVMe over Fabrics™ extends the benefits of efficient storage architecture at scale in the world's largest data centers by allowing the same protocol to extend over various networked interfaces.

Evolution of NVMe™

2011
- NVM Express Specification 1.0 published by industry leaders on March 1

2012
- NVM Express Specification 1.1 released on October 11

2014
- NVM Express Specification 1.2 released on November 3
- NVM Express Work Group was incorporated at NVM Express, Inc., the consortium responsible for the development of the NVM Express specification
- Work on the NVM Express over Fabrics (NVMe-oF™) Specification kicked-off

2015
- NVM Express Management Interface (NVMe-MI™) Specification officially released. Provides out-of-band management for NVMe™ components and systems and a common baseline management feature set across all NVMe™ devices and systems.

2016
- NVM Express over Fabrics (NVMe-oF™) Specification published; extending NVMe™ onto fabrics such as Ethernet, Fibre Channel and InfiniBand®, providing access to individual NVMe™ devices and storage systems.

2017
- NVM Express Specification 1.3 published. Addresses the needs of mobile devices, with their need for low power consumption and other technical features, making it the only storage interface available for all platforms from mobile devices through data center storage systems.

nvm EXPRESS®

NVMe™ Adoption – Industry

NVMe™ displacing SAS and SATA SSDs in server/PC markets

- PCIe NAND Flash SSDs primarily inside servers
- Lower latency Storage Class Memory (i.e., 3D Xpoint™) SSDs - NVMe™-only
- Extensive Client (i.e., laptop, tablet) use of smaller form factor SSDs – M.2 and BGA

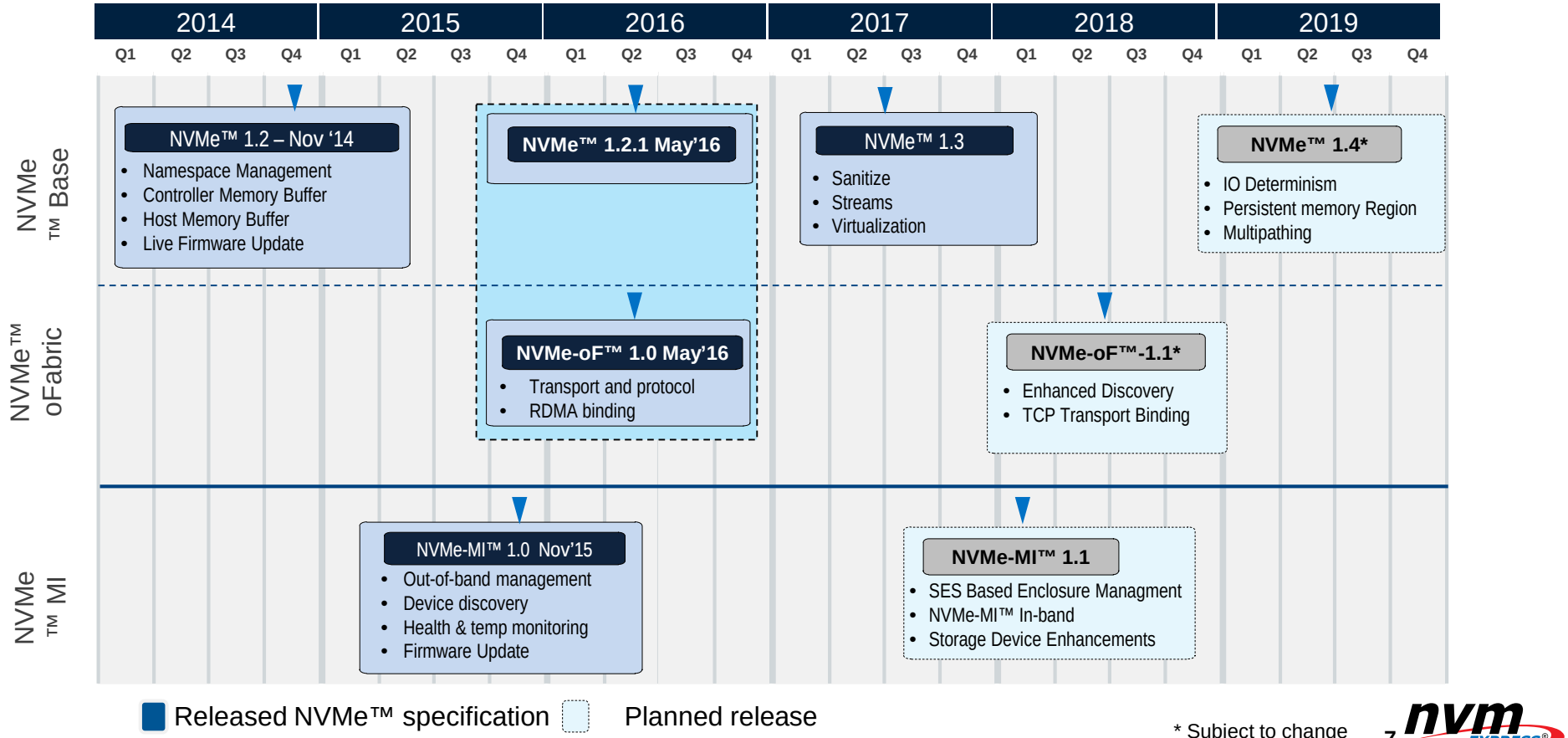
NVMe™ ecosystem and recognition growing quickly

- Many servers offer NVMe™ slots - different server configurations and form factors
- Startups already shipping NVMe™ and NVMe over Fabrics™ (NVMe-oF™) solutions
- Storage class NVMe™ SSDs emerging – enable high availability (HA) in storage arrays
  - Dual port support, enterprise level features (i.e., Data Integrity Field, TCG-stored data encryption)
- NVMe-oF™ emerging as a solution to limited scale of PCIe as a fabric
- Expanding ecosystem (i.e., Analyzers, NVMe-oF™ adapters)

Special Thanks to David Black, Dell/EMC

So...what's next?

NVMe™ Feature Roadmap



* Subject to change

The Future of NVMe™

NVMe™ 1.4
- IO Determinism
- Persistent Controller Mem Buffer and Event Log
- Multipathing

NVMe-MI™ 1.1
- SCSI Enclosure Services (SES)
- NVMe-MI™ In-band
- Native Enclosure Management

NVMe-oF™ 1.1
- Enhanced Discovery
- TCP Transport Binding

Quick Definition of Terms

Time to Level-Set our Words!

Key Terminology - NVM Express (NVMe™)

The command protocol to address non-volatile memory ***used with PCIe***

From the specification:

NVMe™ is a scalable host controller interface designed to address the needs of Enterprise and Client systems that utilize PCIe-based SSDs. The interface provides optimized command submission and completion paths. It includes support for parallel operation by supporting up to 65,535 I/O Queues with up to 64K outstanding commands per I/O Queue. Support has been added for many Enterprise capabilities such as end-to-end data protection (compatible with SCSI Protection Information, commonly known as T10 DIF, and SNIA DIX standards), enhanced error reporting and virtualization.

NVMe over Fabrics™ (NVMe-oF™): The command protocol to address non-volatile memory ***using a non-PCIe network***

Understanding Workgroup Terms

TPAR - Technical Proposal Authorization Request

- Early-stage document to propose new working projects
- First phase of any technical idea to be worked on
- If approved for work, ‘graduates’ to a TP (below)

TP - Technical Proposal

- Authorized working proposal (also known as “in Phase 2”)

ECN - Engineering Change Notice

- After a specification is finalized, sometimes there are errors that need to be fixed
- Errata are fixed and sent out as an ECN

Useful Terminology

NVM Subsystem

NVMe™ Domain*

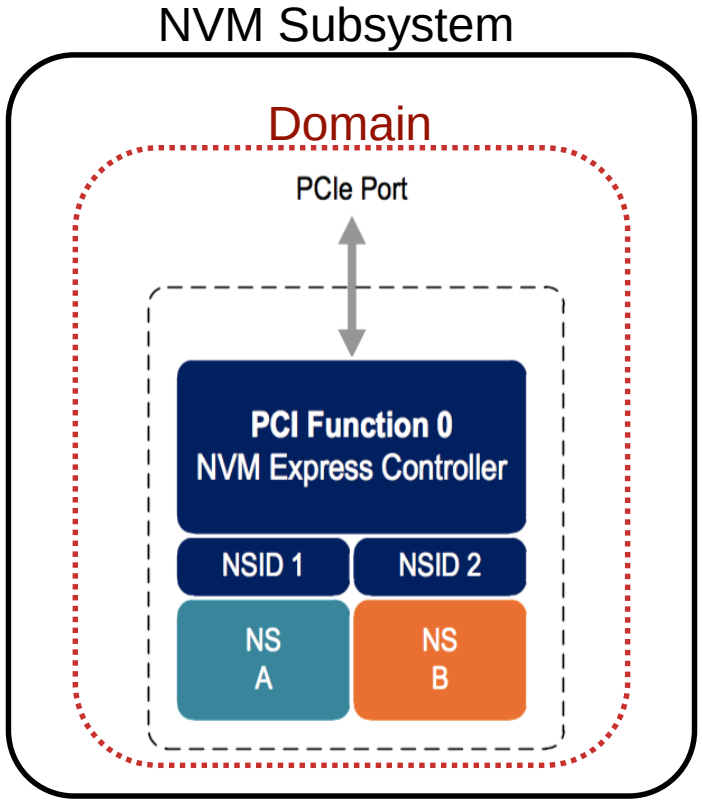
NVMe™ Controller

NVMe™ Namespace & Namespace ID

NVM Sets*

Consists of the components that make up a NVMe™ target:

- Domain(s)*
- NVMe™ Port(s)
- NVMe™ Controller(s)
- Namespace(s)
- Media



**New* concept

Useful Terminology

NVM Subsystem

NVMe™ Domain*

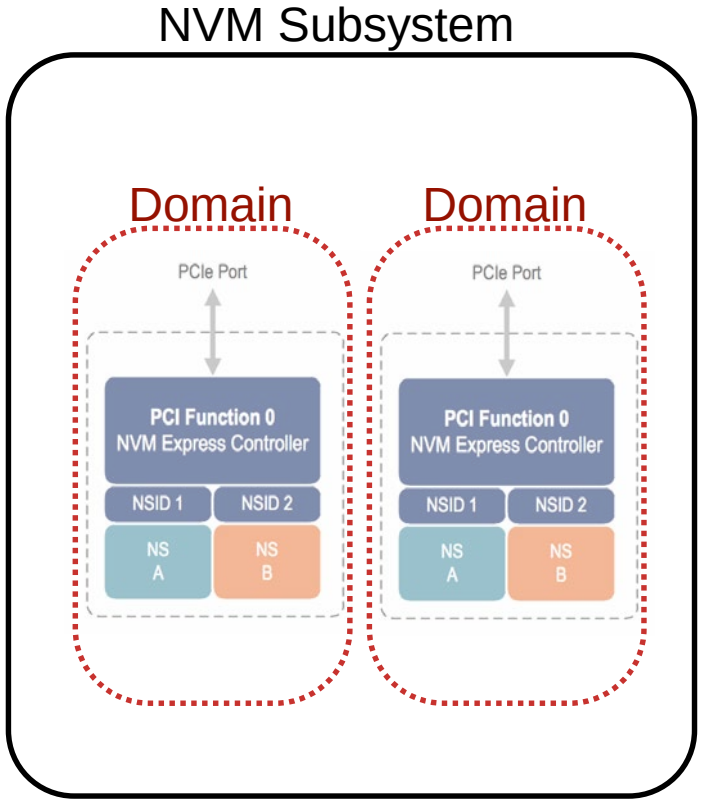
NVMe™ Controller

NVMe™ Namespace & Namespace ID

NVM Sets*

An entity within a NVMe™ subsystem

- Two or more domains can exist
- They are independent
- Can be cooperative
- If they are not cooperative, a NVM Subsystem is considered to be *partitioned*



**New* concept

Useful Terminology

NVM Subsystem

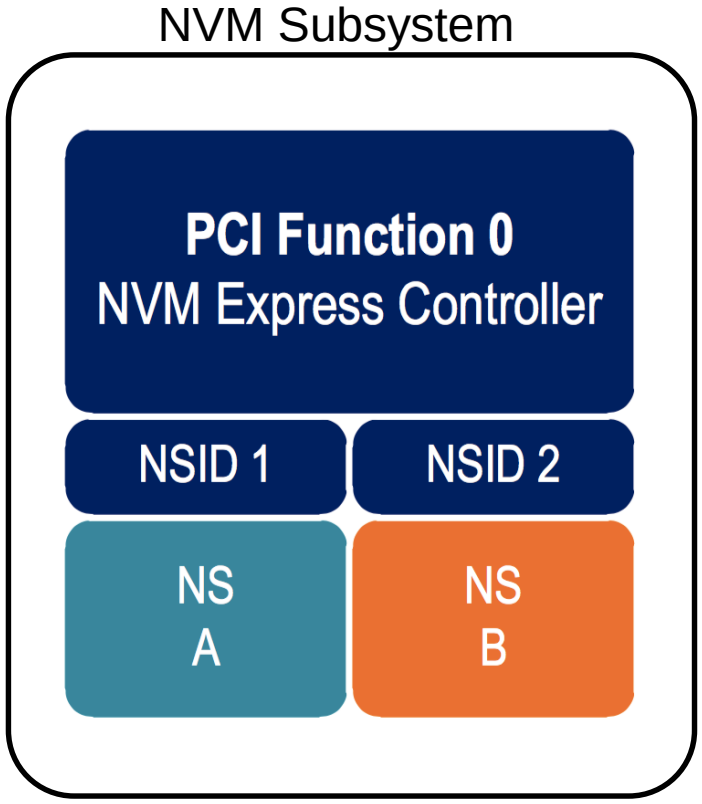
NVMe™ Domain*

NVMe™ Controller

NVMe™ Namespace & Namespace ID

NVM Sets*

- An entity within a NVMe™ subsystem
- Controller can support multiple namespaces



**New* concept

Useful Terminology

NVM Subsystem

NVMe™ Domain*

NVMe™ Controller

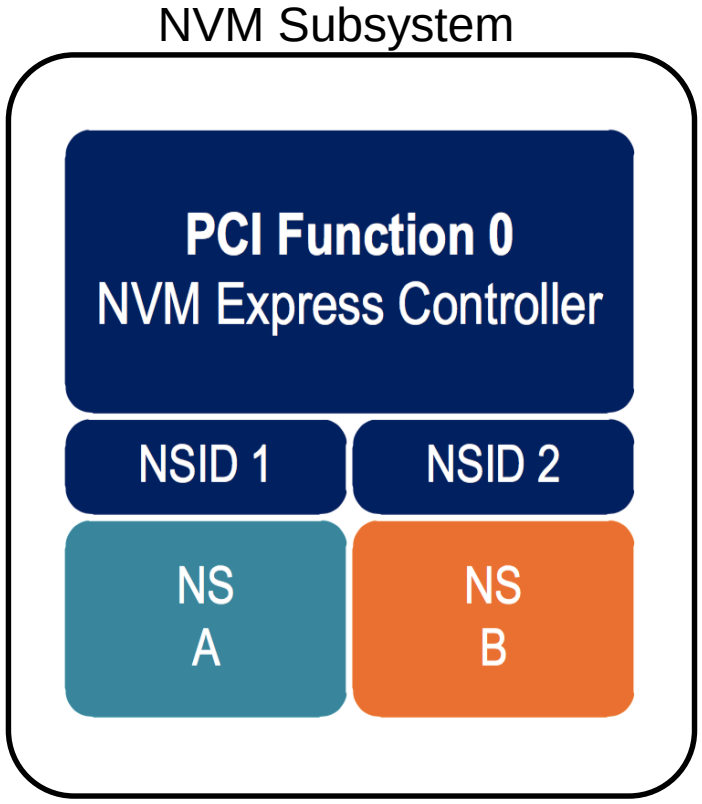
NVMe™ Namespace & Namespace ID

NVM Sets*

A quantity of non-volatile memory that may be formatted into logical blocks

- Can be shared
- Can be private

A NamespaceID (NSID) is an identifier used by a controller to provide access to a namespace



**New* concept

Useful Terminology

NVM Subsystem

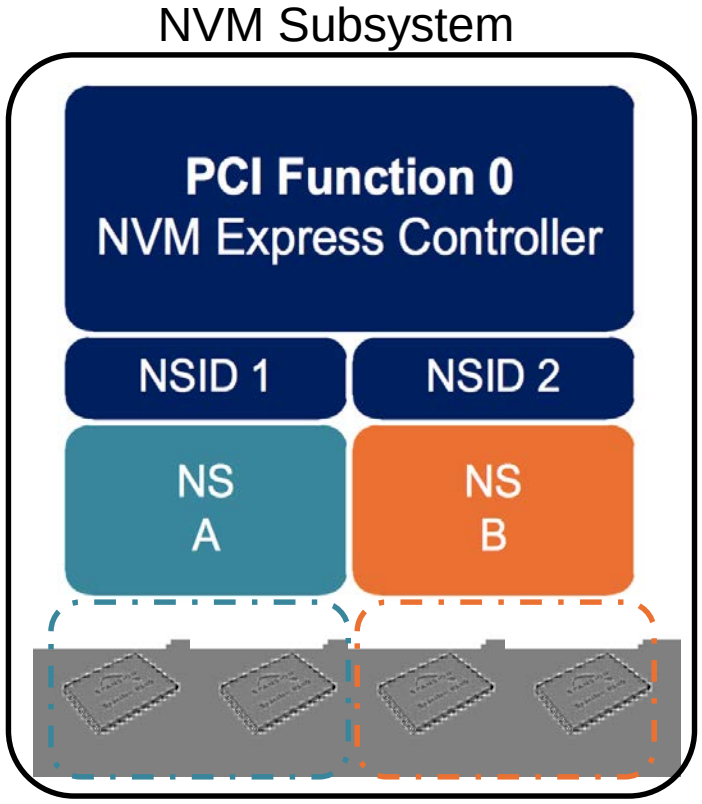
NVMe™ Domain*

NVMe™ Controller

NVMe™ Namespace & Namespace ID

NVM Sets*

A collection of non-volatile media with particular attributes from which one or more namespaces may be allocated



**New* concept

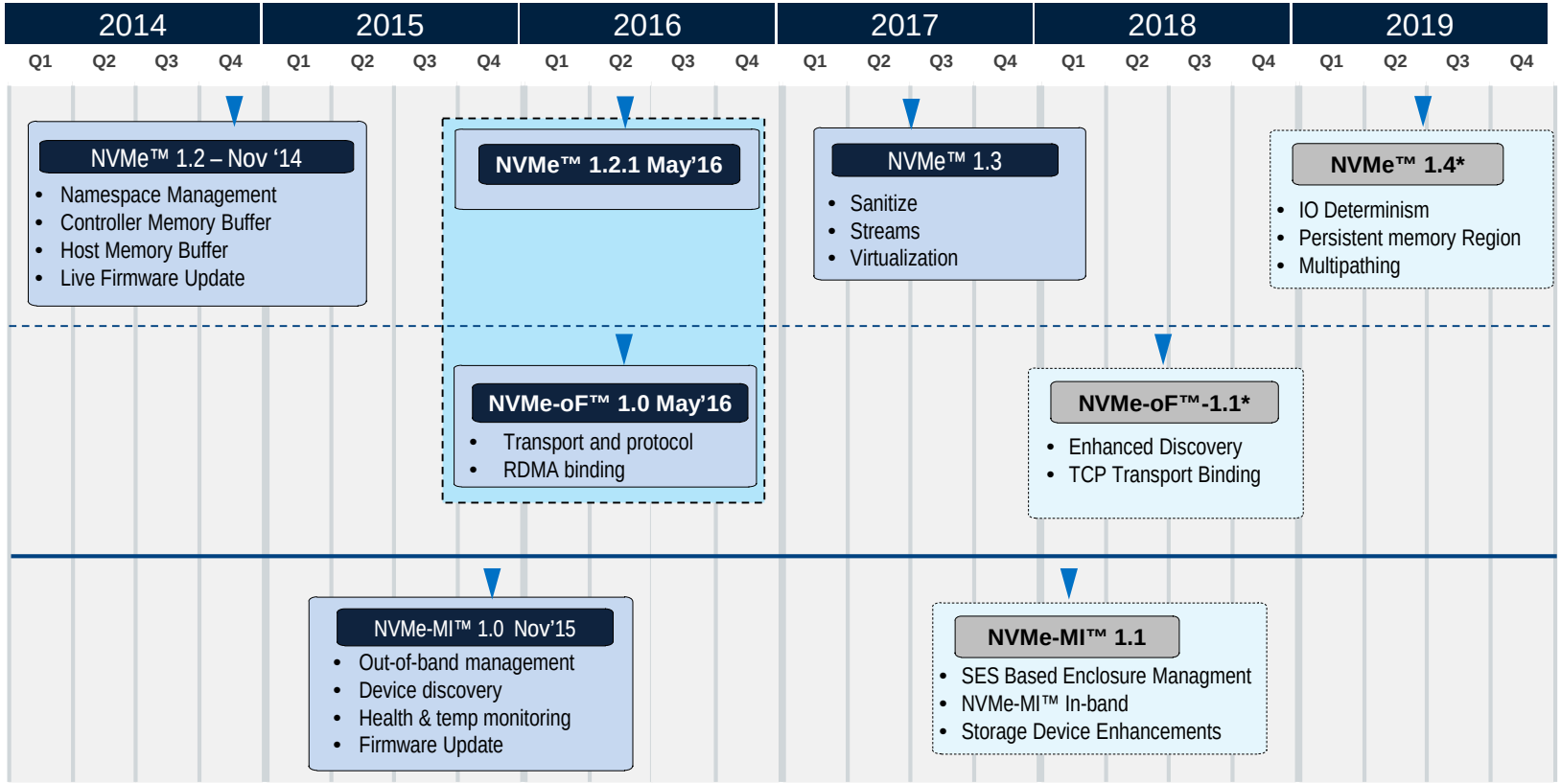
What does an NVMe Subsystem Look Like?

Examples:

NVMe Roadmap
Continuous Improvements

Projected completion: Early 2018

NVMe™ Feature Roadmap



Released NVMe™ specification Planned release

* Subject to change

Ever-Advancing Performance and Features

NVMe™ 1.4*

- I/O Determinism
- Persistent memory Region
- Multipathing

Data latency

- Improvement: I/O Determinism (IOD)

High Performance Non-Volatile data needs

- Improvement: Persistent Memory Region

Ease of Data sharing

- Improvements: Multi-Pathing access

Management Needs

NVMe-MI™ 1.1

- SES Based Enclosure Management
- NVMe-MI™ In-band
- Storage Device Enhancements

Standardized Management for ease of adoption

- Industry standard tools and compliance

Improvements and updates to managing the subsystems and end devices

- Event logging
- Incorporating robust industry adopted enclosure management
- Diverse connections to end devices (SSDs)
    - Additional In-band mechanisms

Enterprise Networking Needs

NVMe-oF™-1.1*

- Enhanced Discovery
- TCP Transport Binding

- Robustness in networking topologies
  - Congestion Management
- New and interesting transport capabilities
  - TCP bindings for NVMe-oF™
- Improvements in automation
  - Discovery
- Security Enhancements
  - In-band authentication

NVMe™ 1.4

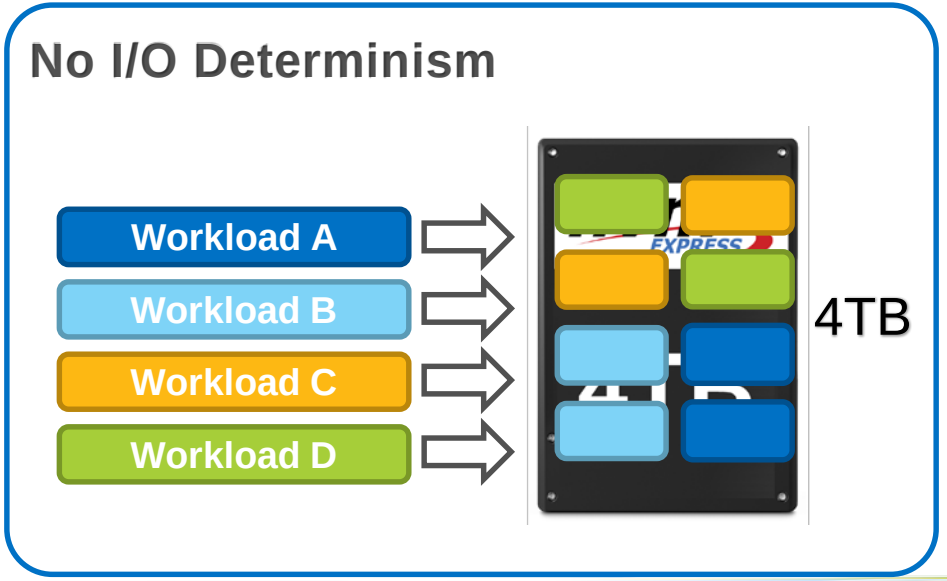
Projected completion: 2019

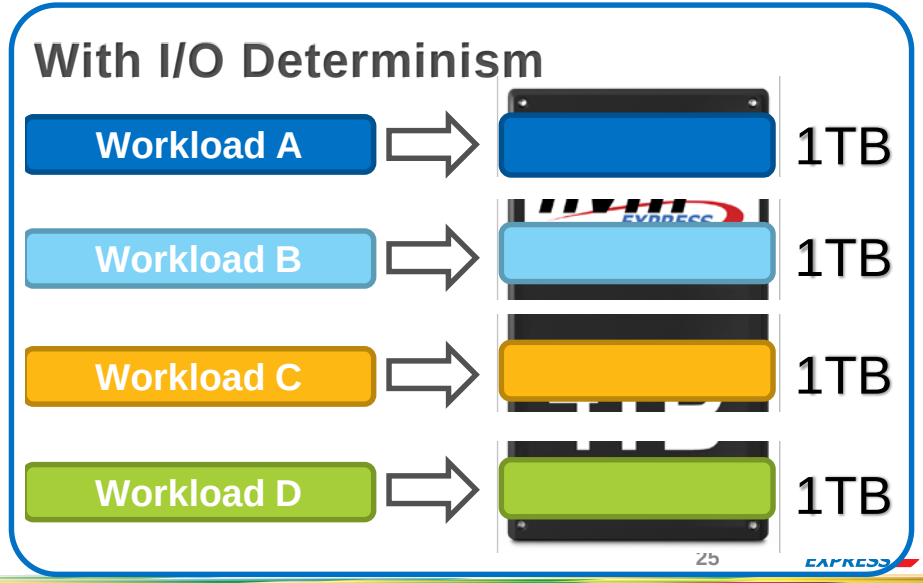
What is I/O Determinism?

- Enables hosts to treat an SSD as many small sub-SSDs and process I/O in parallel in each small sub-SSD
  - Enables host threads to process I/O independently in small sub-SSDs without blocking from other thread I/Os
- Can reduce average read latency significantly for higher performance and for better Quality of Service (QoS)
  - These improvements are due to the parallel execution of I/Os without any conflict to the media
- I/O Determinism is gaining considerable amount of industry attention

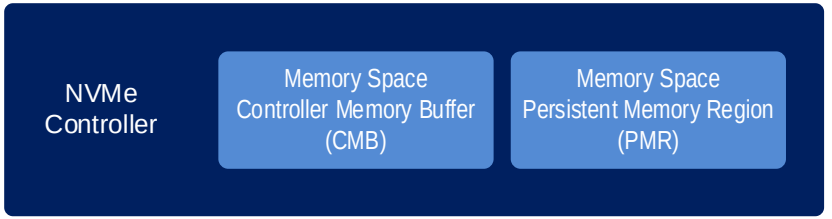
What is NVMe I/O Determinism?

- Service isolation region
- Increase Read I/OPs and reduce max latency
- Provides strict QoS profile
- Significantly improves P99 and P9999 for a well-behaved host





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 and command data
- Contents do not persist across power cycles and resets

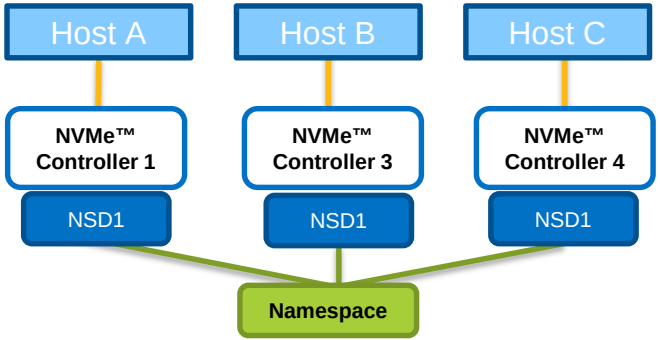
Persistent Memory Region (PMR)

- PCI memory space exposed to host
- May be used to store command data
- Content persist across power cycles and resets

NVMe™ Multipathing and Namespace Sharing

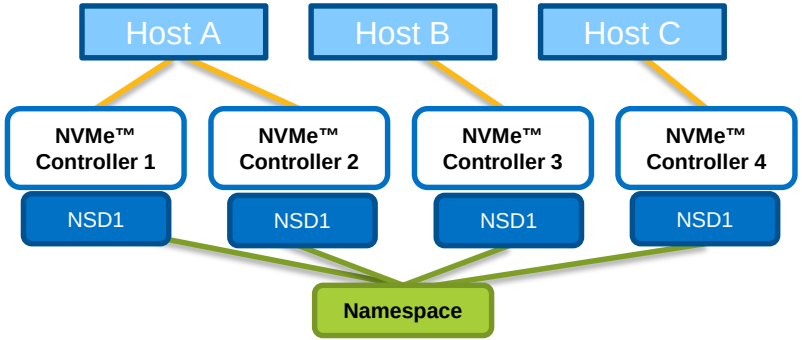
Technical Term: Asymmetric Namespace Access (ANA)

NVMe™ Multipathing I/O refers to two or more completely independent PCI Express paths between a single host and a namespace



NVMe™ Multipathing

Namespace sharing enables two or more hosts to access a common shared namespace using different NVM Express controllers



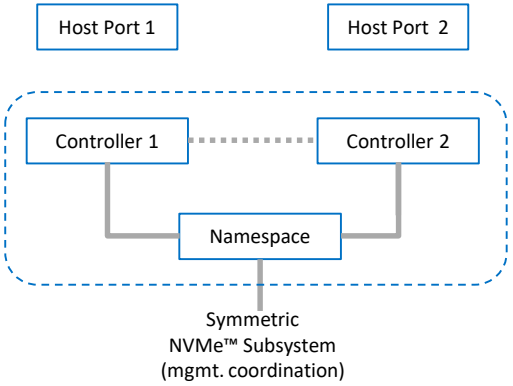
Namespace Sharing

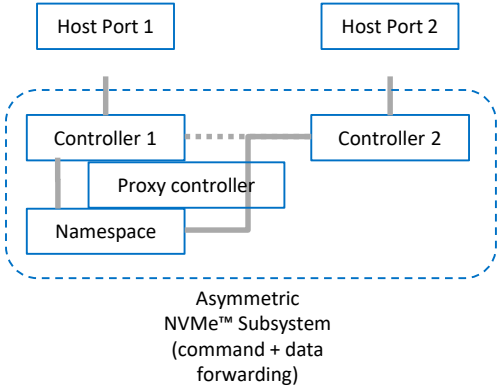
Both multi-path I/O and namespace sharing require that the NVM subsystem contain two or more controllers

Multi-Pathing Related Capabilities

- Symmetric Multi-Pathing already exists in the NVMe™ specification
  - Symmetric access: The same access characteristics on all paths
    - The host doesn't care which path is used – because they are all the same
  - Identify controller data (CMIC field)
  - Identify namespace data (NMIC field)

- Fabrics provide additional opportunities for multiple asymmetric paths
  - Asymmetric access: Different access characteristics on different paths
    - The host cares which path is used – because they are NOT all the same
  - New capabilities to be added to the NVMe™ specification





Special Thanks to Fred Knight, NetApp

NVMe™ Management Interface (NVMe-MI™) 1.1

Projected completion: Early 2018

NVMe-MI™ 1.1 Key Work Items

NVMe-MI™ 1.1

- SES Based Enclosure Management
- NVMe-MI™ In-band
- Storage Device Enhancements

- SCSI Enclosure Services (SES) Based Enclosure Management
  - Draft completed, working through final technical items
  - SCSI translation (completed)
- Support for In-Band NVMe-MI™
  - Draft complete and in workgroup review
- NVMe™ Storage Device Enhancement – In work

Enclosure Management

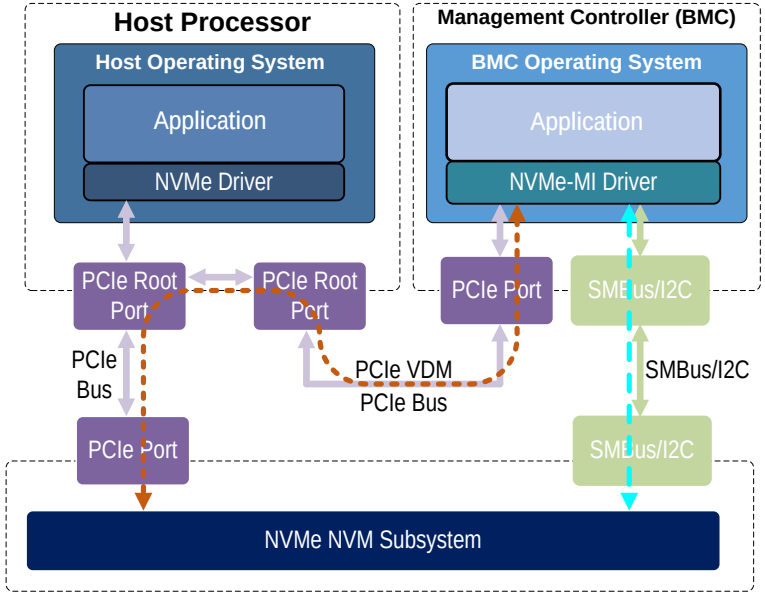
- Native PCIe Enclosure Management (NPEM)
  - Transport specific basic enclosure management
  - Submitted to the PCI-SIG Protocol Workgroup (PWG) on behalf of the NVMe™ Management Interface Workgroup
  - Approved by PCI-SIG on August 10, 2017
- SES Based Enclosure Management
  - Technical proposal being developed in NVMe-MI™ workgroup
  - Comprehensive enclosure management

Management – In-Band? Out-of-Band? Rock Band?

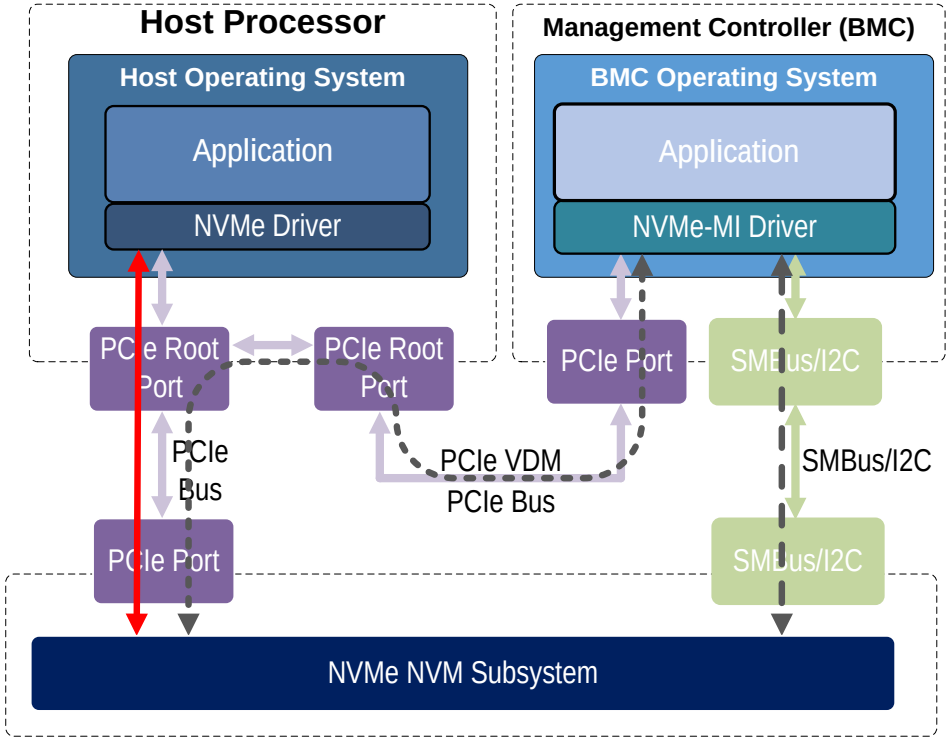
- Out-of-Band Management – Independent connection separate from the main IO path and operation system (Usually SMBus/I2C physical interface)
- In-Band Management – Utilizes the NVMe driver and the main data path interface (Usually PCIe Bus)
- Provides “Rockin” NVMe-MI management solutions and flexibility of implementations

NVMe-MI™ Out-of-Band Management

- **Out-of-Band Management** – Management that operates with hardware and components that are ***independent of the operation system control***
- **NVMe™ Out-of-Band Management Interfaces**
  - SMBus/I2C
  - PCIe Vendor Defined Messages (VDM)
  - IPMI FRU Data (VPD) accessed over SMBus/I2C



In-Band Management and NVMe-MI™



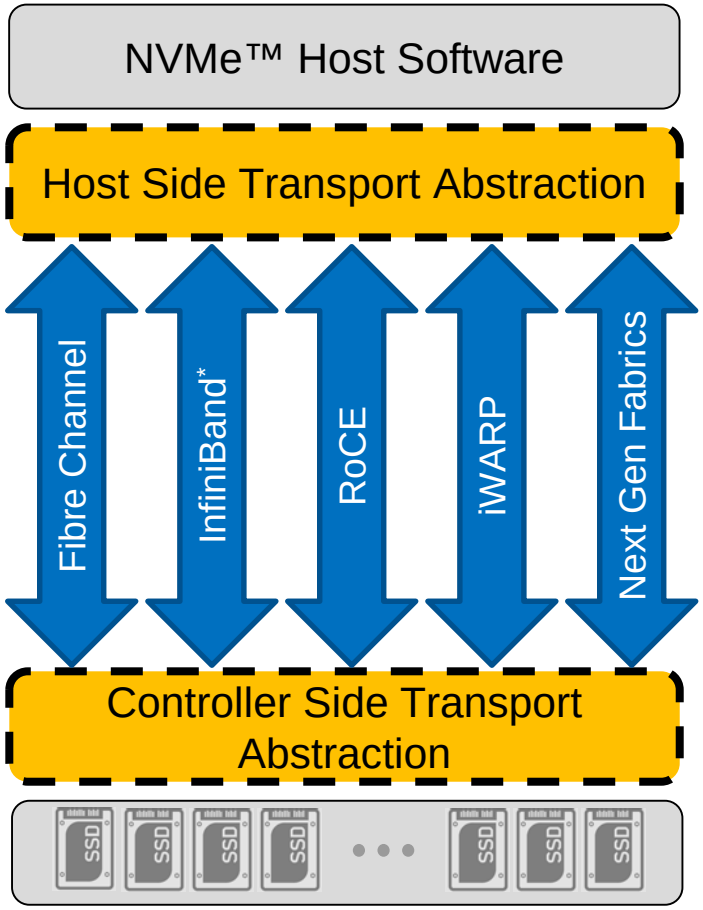
- In-band mechanism allows application to tunnel NVMe-MI™ commands through NVMe™ driver
  - Two new NVMe™ Admin commands
    - NVMe-MI™ Send
    - NVMe-MI™ Receive
- Benefits
  - Provides management capabilities not available in-band via NVMe™ commands
    - Efficient NVM subsystem health status reporting
    - Ability to manage NVMe™ at a FRU level
    - Vital Product Data (VPD) access
    - Enclosure management

NVMe over Fabrics™ 1.1

Projected completion: 2018

NVMe over Fabrics™

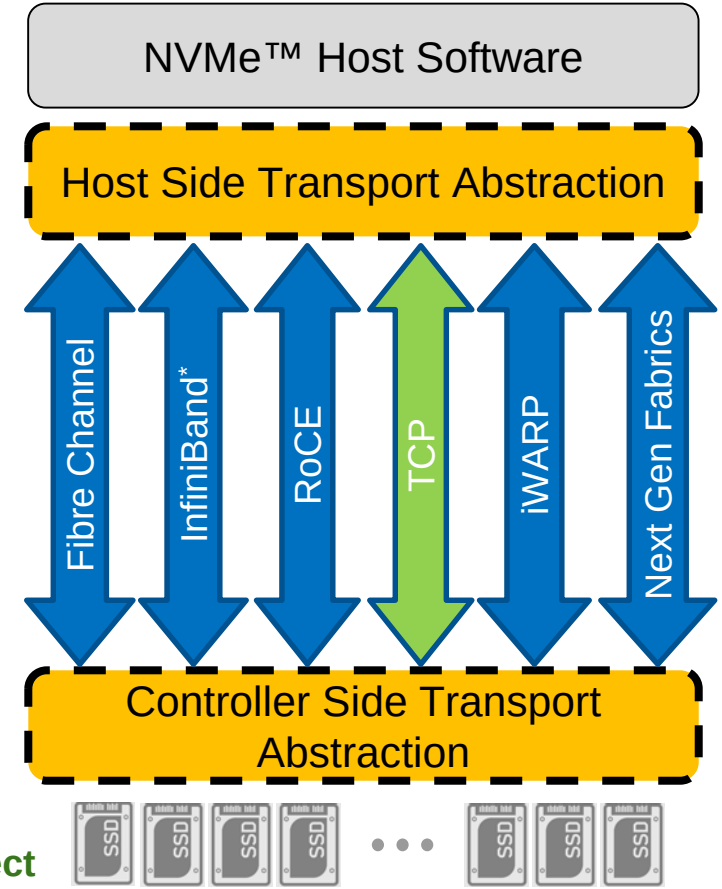
- Use NVMe™ end-to-end to get the simplicity, efficiency and low latency benefits
- NVMe over Fabrics™ is a thin encapsulation of the base NVMe™ protocol across a fabric
  - No translation to another protocol (i.e. SCSI)
- NVMe-oF™ Fabrics v1.0 include RDMA-based transports and Fibre Channel
  - RDMA-based i.e. InfiniBand™, RoCE, iWARP



Slide credit: NVM Express and Intel

NVMe-TCP

- NVMe™ block storage protocol over standard TCP/IP transport
- Enables disaggregation of NVMe™ SSDs ***without compromising latency*** and ***without requiring changes to networking infrastructure***
- Independently scale storage & compute to maximize resource utilization and optimize for specific workload requirements
- Maintains NVMe™ model: sub-systems, controllers namespaces, admin queues, data queues

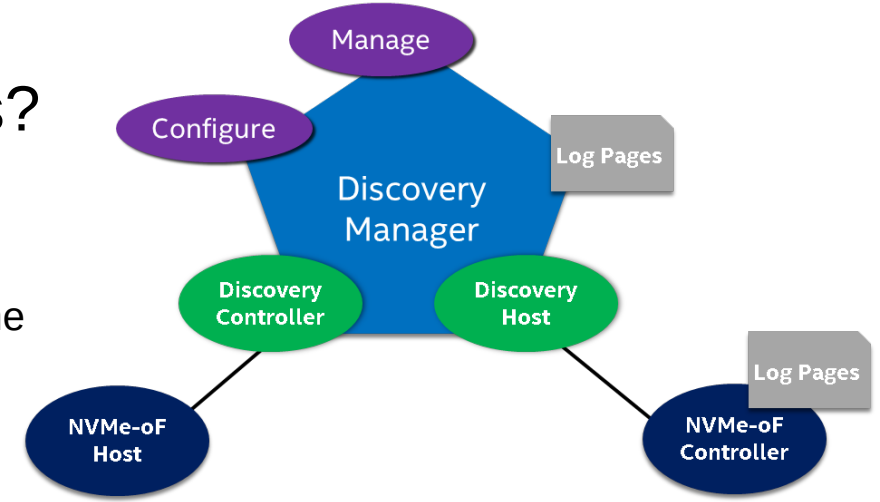


Designed to be a simple, efficient, and scalable way to connect compute nodes to a pool of remote NVMe™ SSDs

Special Thanks to Cam Eshgi, LightBits

Enhanced Discovery

- How do I connect storage consumers to storage suppliers?
  - Specification enhancement for efficient, dynamic resource management
  - Fabric-transport specific mechanisms to determine where to get provisioning information from
  - Linux kernel driver stack changes as the specification evolves
  - Management tools to enable NVMe-oF™ management and scale-out



Special Thanks to Phil Cayton, Intel

Summary

Summary

NVMe has over 50 ongoing projects in the technical working groups

There are over 130 participating companies working on these projects

Key improvements to the NVMe base spec, Fabrics, and Management will help facilitate more robust, resilient, and powerful tools for data centers.

Special Thanks:

- David Black (Dell/EMC)
- Fred Knight (NetApp)
- Kam Eshghi (Lightbits Labs)
- Phil Cayton (Intel)
- Dave Minturn (Intel)
- Peter Onufryk (Microsemi)

Questions?

Visit www.nvmexpress.org or more information on NVM Express technology

Follow us:

Twitter: twitter.com/nvmexpress
LinkedIn: linkedin.com/company/11106843/


nvm
EXPRESS®
Architected for Performance