Speakers

Ross Stenfort

Hardware System Engineer

facebook


Facebook @ Scale
1 Billion
1.3 Billion
3.1 Billion

What Does Hyperscale And NVMe® Technology Look Like?

Next Generation NVMe® Form Factor: E1.S

- E1.S is a **next generation NVMe form factor**
  - *With same PCB and firmware supports a diverse family of thermal options for the market:*
    - High density (5.9mm)
    - High performance at low airflow (25mm)
- Supports performance scaling
  - Support for Gen 5 PCIe® and beyond
- Hot plug support
- Excellent for both storage and compute in 1OU
- Broad market support

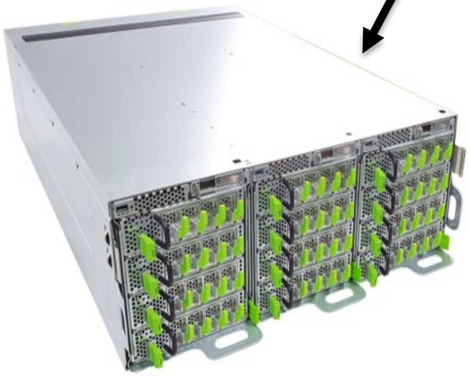
Extensive profolio of thermal options from high performance to high density

Scalable and Efficient NVMe® Platforms

25mm E1.S Latch

1 OU Blade

Chassis with 1 OU Blades

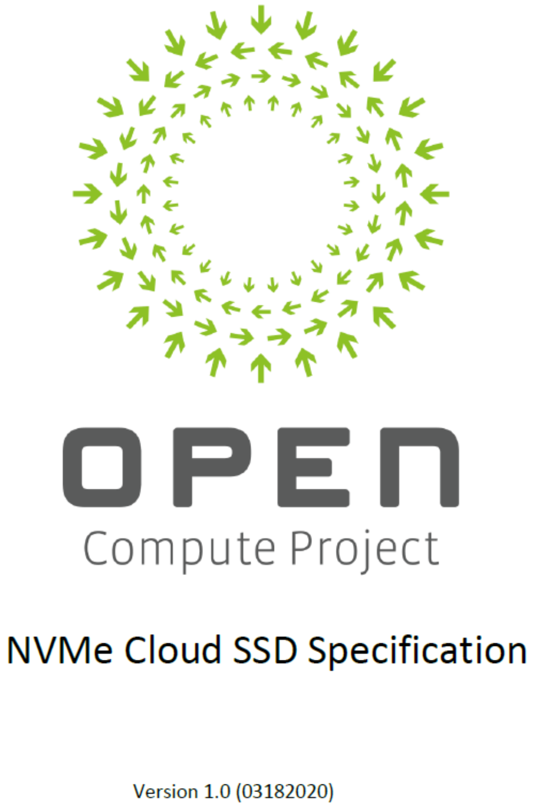


	1 OU Blade Platform	2 OU Blade Platform
Chassis Height	4OU	
SSD	25mm E1.S @25W	
Number of SSDs per Blade	4	6
Number of SSDs per Chassis	48	36
Max Blade TLC Capacity	64 TB	96 TB
Max Chassis TLC Capacity	768 TB	576 TB
Efficiency	Excellent: Less Than 0.145 CFM/W	

Chassis with 2 OU Blades

Flash Memory Summit

What are Hyperscale NVMe® Device Requirements?


OPEN
Compute Project

NVMe Cloud SSD Specification

Version 1.0 (03182020)

Link to specification can be found under OCP Contributions:
https://www.opencompute.org/documents/nvme-cloud-ssd-specification-v1-0-3-pdf

What does the NVMe® Cloud SSD Cover?

Table of Contents

- There are ~70 pages of requirements of what is needed to build a NVMe® Cloud SSD

- This includes requirements around:
  - NVM Express®
  - PCI Express®
  - SMART Logs
  - Reliability
  - Thermal
  - Power
  - Security
  - Form Factor
  - SMBUS
  - Tooling

Everything Needed To Build a NVMe Cloud SSD

Flash Memory Summit

nvm EXPRESS®

Building innovative and highly efficient data centers using NVMe® technology

Speakers

Rupin Mohan

Director R&D, CTO (SAN

Hewlett Packard Enterprise

Agenda for 2020

Data Center Trends

New I/O Stack Refresher

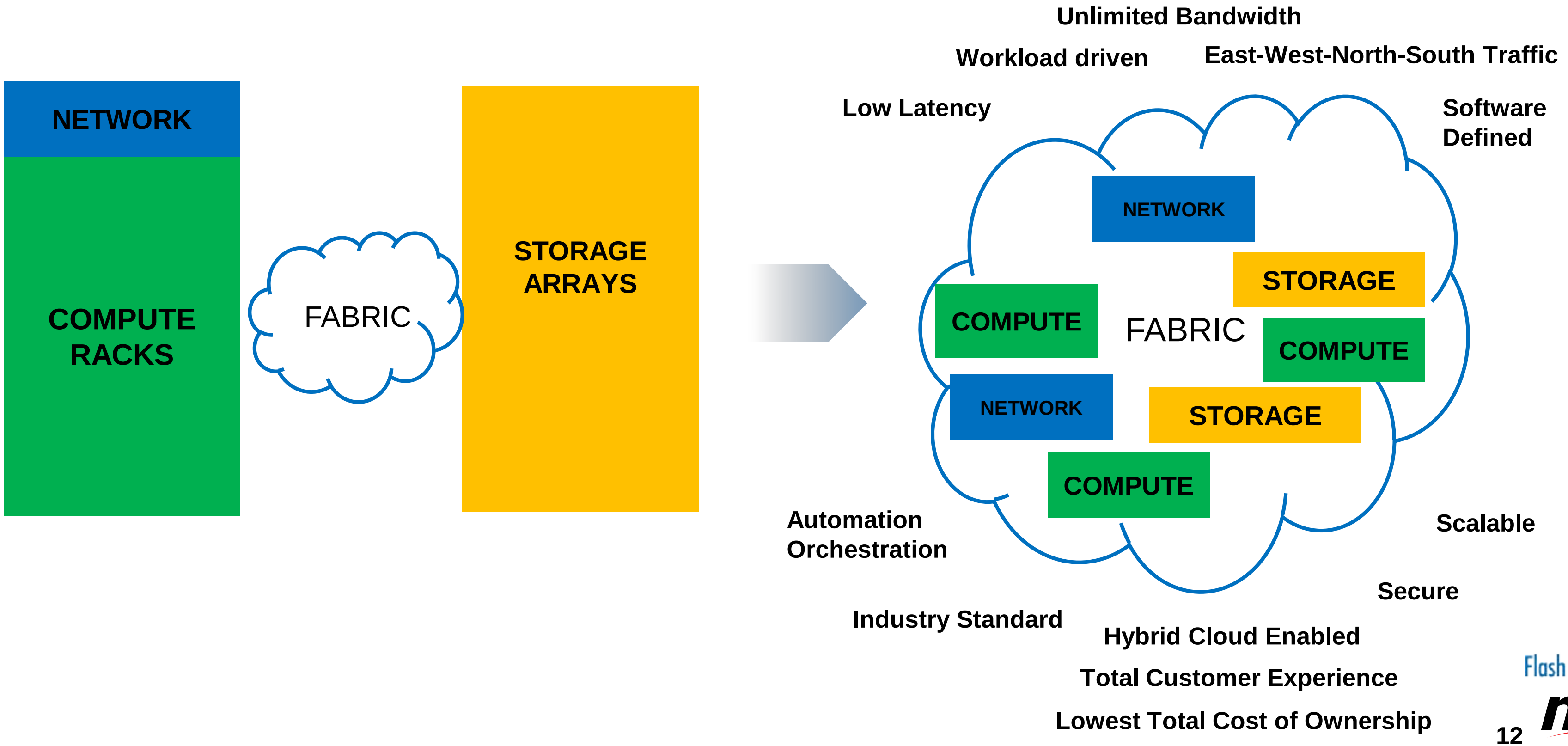
Hybrid Cloud – Storage Networking Protocol Comparison

NVMe® Centralized Discovery Controller

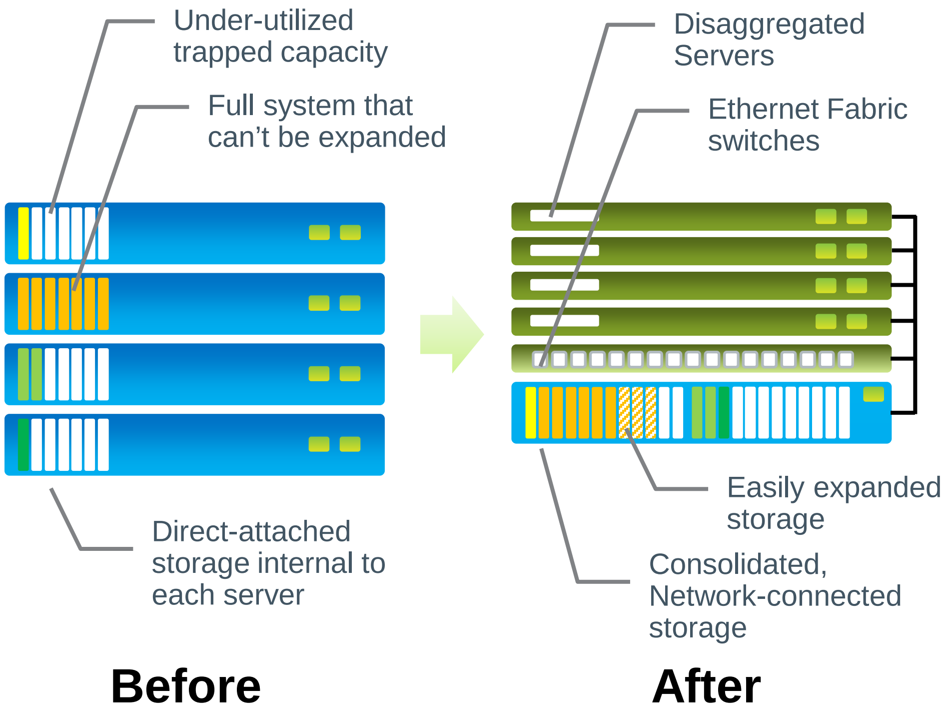
Next Steps

Data Center Trends

Disaggregation – What does it mean?



NVMe-oF™ Technology Use Case - Redefining Internal DAS



- Advantages:
  - Delivers the performance of DAS
  - Improves utilization of flash and facilitates data sharing
  - Increases availability of storage with HA and network connectivity
  - Reduces rack space and power requirements
  - Delivers better Total Cost of Ownership
  - Improves customer experience deploying NVMe-oF™ technology

The New I/O Stack

with NVMe® over Fabrics specification

A new language for accessing solid state media

Traditional Storage Arrays

1. Storage Controller runs SCSI
2. Front end FC/iSCSI
3. Backend SAS/SATA
4. Software Feature Rich based on SCSI

Hybrid Storage Arrays

1. Storage Controller runs SCSI. Upgraded back end (partial/full)– Controller does SCSI-NVMe translation with NVMe® drives in the backend
2. Memory-Driven Flash
3. Software Feature Rich based on SCSI

Next Gen. Storage Arrays

1. Controller runs NVMe
2. Backend NVMe Drives (PCIe®, NVMe over Fabrics)
3. Frontend NVMe (FC-NVMe, NVMe over Ethernet)
4. Software Features running NVMe, expect parity in 3 years

Flash Memory Summit

nvm EXPRESS®

I/O Stack evolution

Management of NVMe Namespaces
- Redfish/Swordfish API's

Layer	Evolution
Applications	Enterprise Apps taking advantage of SPDK (RDMA)
OS – Storage Stack	Volume Manager optimized to NVMe® technology, new protocol
Host Adapter – Driver	FC-NVMe, NVMe over Ethernet (RoCEv2, TCP) – Lim. OS Support
SAN – Switch	FC, Ethernet switches
Host Port on Array – Front End Fabric	FC-NVMe, NVMe over Ethernet (RoCEv2, TCP)
SCM - Cache	3D X-point as read cache (Memory Driven Flash)
Storage Controller Core	Transition to NVMe technology, including all features (RC, etc)
Drives in Head Shell	Partial # of NVMe drives to full cage
JBOF – for scale	Scale to multiple shelves over PCIe® or Switching Fabric

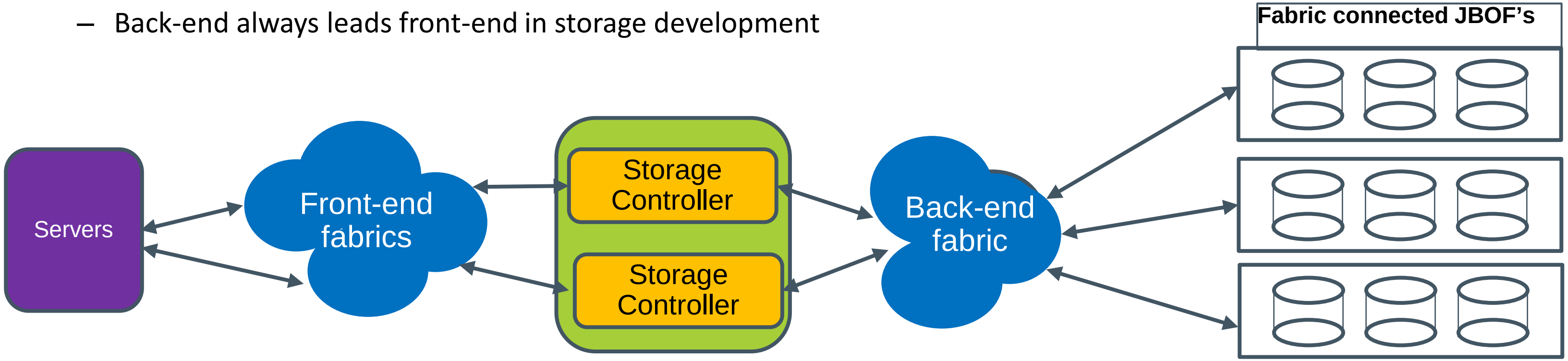
Use Cases for the Enterprise

with NVMe® over Fabrics technology

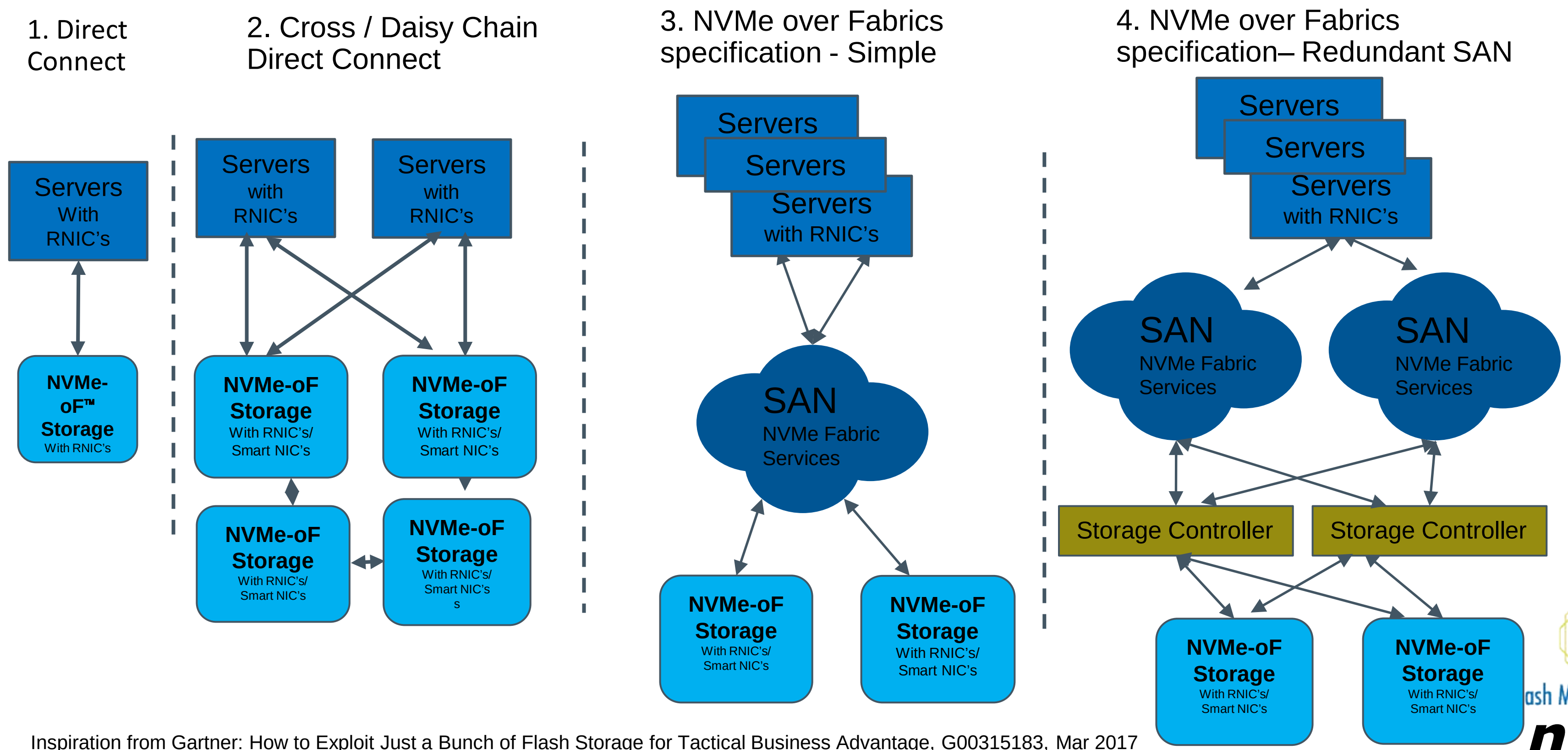
NVMe® over Fabrics Specification– Enterprise Storage

Shared storage will require NVMe® primary arrays to have FABRIC connectivity

- Initially on the ***back-end*** of the array and on the ***front-end*** as well
- Back-end always leads front-end in storage development



NVMe® over Fabrics Technology Deployment Scenarios



Inspiration from Gartner: How to Exploit Just a Bunch of Flash Storage for Tactical Business Advantage, G00315183, Mar 2017

Hybrid Cloud - Protocol Comparison

NVMe® over Fabrics technology

The landscape today….

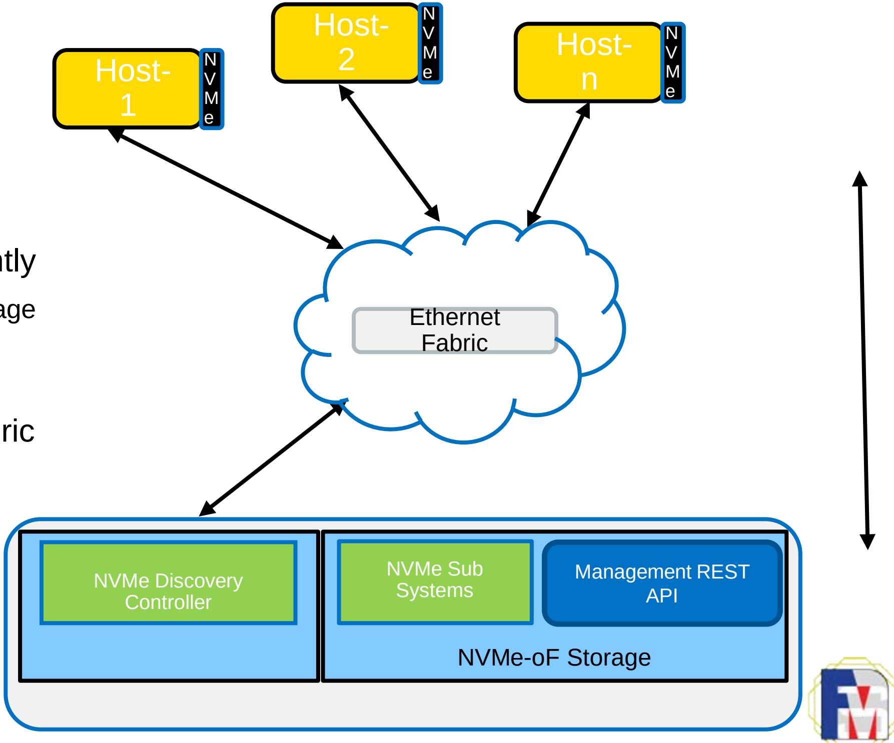
Protocol	Latency	Scalable	Performance	Hybrid Enterprise
Fibre Channel	Lower	Yes	High	Teir 0, On-Prem
RoCEv2	Lowest	Yes	High	Tier 0, Hybrid
TCP	Low-Medium w/Offload	Yes	Medium-High	Tier 1, Hybrid
InfiniBand	Lowest	Limited	High	None
iWARP	Medium	Yes	Medium	None

Centralized Discovery Controller for Ethernet Storage Fabrics

Centralized Discovery Controller

Problem we are trying to solve

- Lack of Centralized Discovery Service for Hosts and Storage Devices in NVMe® Ethernet Fabrics
  - No single location to get consolidated resource information (hosts, discovery controllers) without referrals
- NVMe-oF™ infrastructure scalability hurdles due to discovery/configuration of every resource independently
  - Complex and manual configuration for initial discovery of storage sub-systems
- No standardized mechanism to share information between discovery controllers for NVMe IP-based fabric transports
- No resource visibility management mechanism like iSNS discovery domains or FC soft zoning
- Handling fabric generated events and subsequent notifications
  - E.g. topology changes, grouping changes etc.



NVMe® specifications activity related to centralized discovery

Two technical proposals under development in FMDS (Fabric and Multi Domain Subsystem) NVM Express® task group

- TP 8009, Automated Discovery of Ethernet Discovery Controllers
- TP 8010, NVMe-oF Centralized Discovery (CD)

TP 8009

- An automated discovery mechanism of Discovery Controllers using existing mDNS and DNS-SD protocols:
  - A host can use mDNS query and a discovery controller can respond with its IP address, transport supported and hosts can thereby connect
  - A host or subsystem can send query and discover a centralized discovery controller (see TP 8010 below) in a fabric
  - Maintain compatibility with existing implementations and standard

TP 8010

- Uses TP8009 mDNS mechanism to discover Centralized Discovery Controller (CDC)
- CDC aggregates discovery information for NVMe-oF™ hosts and subsystems
- Groups host and subsystem information, e.g. for access control (zoning) enabling resource visibility management
- Generates fabric events to report changes

Active members: HPE, Dell-EMC, NetAPP, Intel, Lightbits, Mellanox, Marvell, Samsung, VMware

NVMe® specifications activity related to centralized discovery – TP 8009

An automated discovery mechanism using

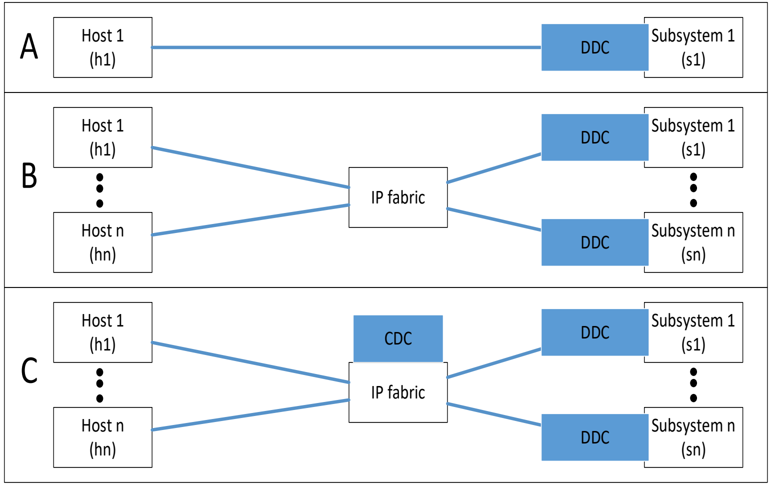
- mDNS (Multicast DNS): Multicast protocol for accessing stored DNS information (see RFC 6762)
- DNS-SD (DNS Service Discovery): Format of service discovery information to store in DNS (see RFC 6763)

Solution supports

- Direct Connect (A)
- Single broadcast domain (B)
- Multiple broadcast domains (C) – with TP 8010

Benefits of the Technical Proposal

- Automated discovery of Discovery Controllers – no manual explicit Host, Subsystem, or Discovery Controller configuration required.
- Dynamic solution that enables NVMe-oF™ entities to detect Discovery Controllers coming and going (mDNS announce).
- Does not preclude High Availability
- Provides a scalable solution to support Point-to-point, single broadcast domain, Multiple broadcast domain, etc. configurations
- Implementation details in the standard are still work in progress



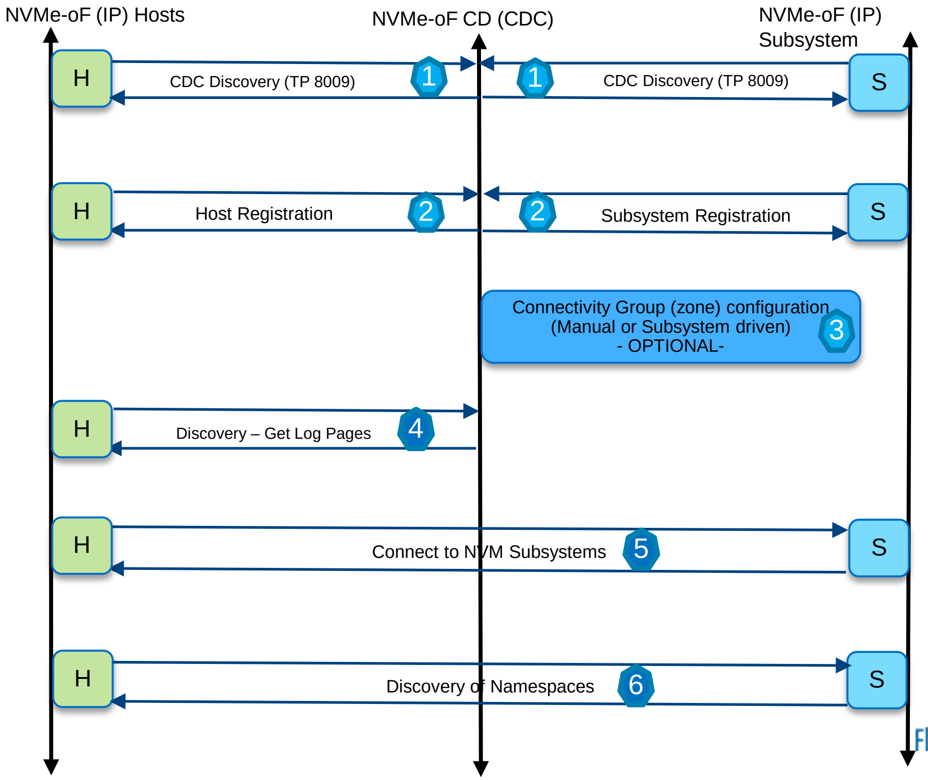
NVMe® specifications activity related to centralized discovery controller – TP 8010

Introduction

- An automated discovery mechanism to get consolidated resource information (hosts, discovery controllers) from a single location
- Co-existence of prior subsystems (CDC unaware) should be supported

Proposed Mechanism

- Uses proposed NVMe® automated discovery of IP discovery controllers TP 8009
- Aggregates discovery information for NVMe-oF™ hosts and subsystems using different discovery controllers
- Groups host and subsystem information, e.g. for access control (Connectivity Groups)
  - Generates fabric events to report changes
  - e.g., topology changes, grouping changes etc.
- Supports high availability
- Implementation details in the standard are still work in progress



Next Steps

Key Design Takeaways

- NVMe-oF™ SAN offers significant opportunities to service low latency, high performance disaggregated storage architectures
- Hybrid Cloud Enterprise is real and is the future
- Low latency and Higher IOPs is the name of the game in the new NVMe® technology world
- Ethernet Storage Fabric is where the Enterprise and Cloud intersects (Hybrid)
- New storage architectures are in development, across the industry
- New NVMe standards (TP 8009, TP 8010) will really simplify deployment and management of NVMe over Ethernet Fabrics

Thank You

Rupin.mohan@hpe.com

Credits: Babu Puttagunta, Curtis Ballard, HPE for driving this work in the NVMe® Group

nvm EXPRESS®

Architected for Performance