STORAGE DEVELOPER CONFERENCE
SDC 21
BY Developers FOR Developers

Virtual Conference
September 28-29, 2021

NVMe/TCP in the Enterprise

Next-Gen End-to-End Paradigm for Storage Connectivity

Mukesh Gupta, Dell-EMC

Murali Rajagopal, VMWare

Agenda

- Market Drivers & Challenges
- Evolution: SCSI, NVMe, NVMeoF, NVMe/TCP
- Next-Gen NVMe IP SAN
- Dell Technologies’ NVMe IP SAN Ecosystem
- Dell-EMC NVMe IP SAN Software
- VMware NVMe Support
- Dell-EMC PowerStore Storage System
- Summary

Agenda

- Market Drivers & Challenges
- Evolution: SCSI, NVMe, NVMeoF, NVMe/TCP
- Next-Gen NVMe IP SAN
- Dell Technologies’ NVMe IP SAN Ecosystem
- Dell-EMC NVMe IP SAN Software
- VMware NVMe Support
- Dell-EMC PowerStore Storage System
- Summary

Market Drivers

- The rapid adoption of **NVMe all flash arrays** across storage platforms
- The market shift towards **next generation disaggregation and composability** of compute, networking and storage infrastructure
- The continued growth of **software defined storage** implementations, especially at the Edge
- The rapid integration of **25/100GbE connectivity** in data centers

Customer Challenges

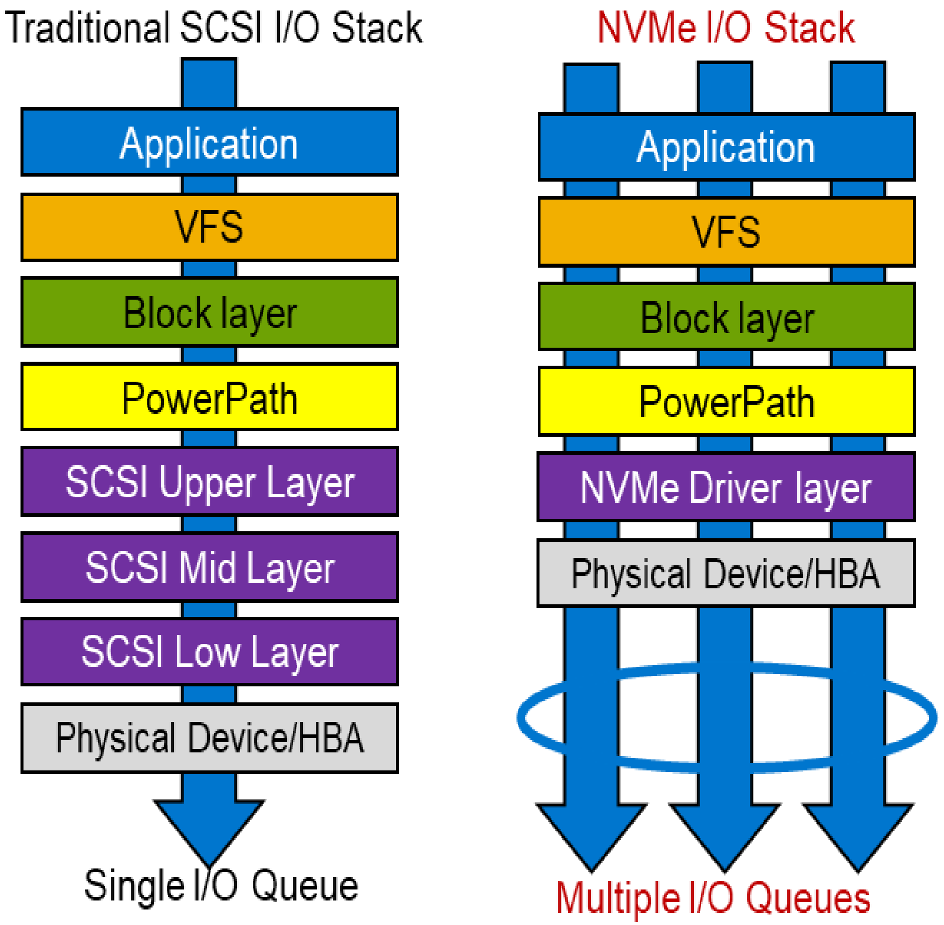
- How to meet the exploding data storage and traffic demands of today's workloads and applications while also preparing for the future
- Desire to leverage modern, cost-effective, open standards-based interconnect solutions without excessive complexity
- How to reduce on-going, multiple platform, operational and support costs while reducing complexity across silos of expertise

Agenda

- Market Drivers & Challenges
- Evolution: SCSI, NVMe, NVMeoF, NVMe/TCP
- Next-Gen NVMe IP SAN
- Dell Technologies' NVMe IP SAN Ecosystem
- Dell-EMC NVMe IP SAN Software
- VMware NVMe Support
- Dell-EMC PowerStore Storage System
- Summary

NVMe vs SCSI

Multi-queue architecture = higher performance



Traditional SCSI I/O Stack

- Application
- VFS
- Block layer
- PowerPath
- SCSI Upper Layer
- SCSI Mid Layer
- SCSI Low Layer
- Physical Device/HBA

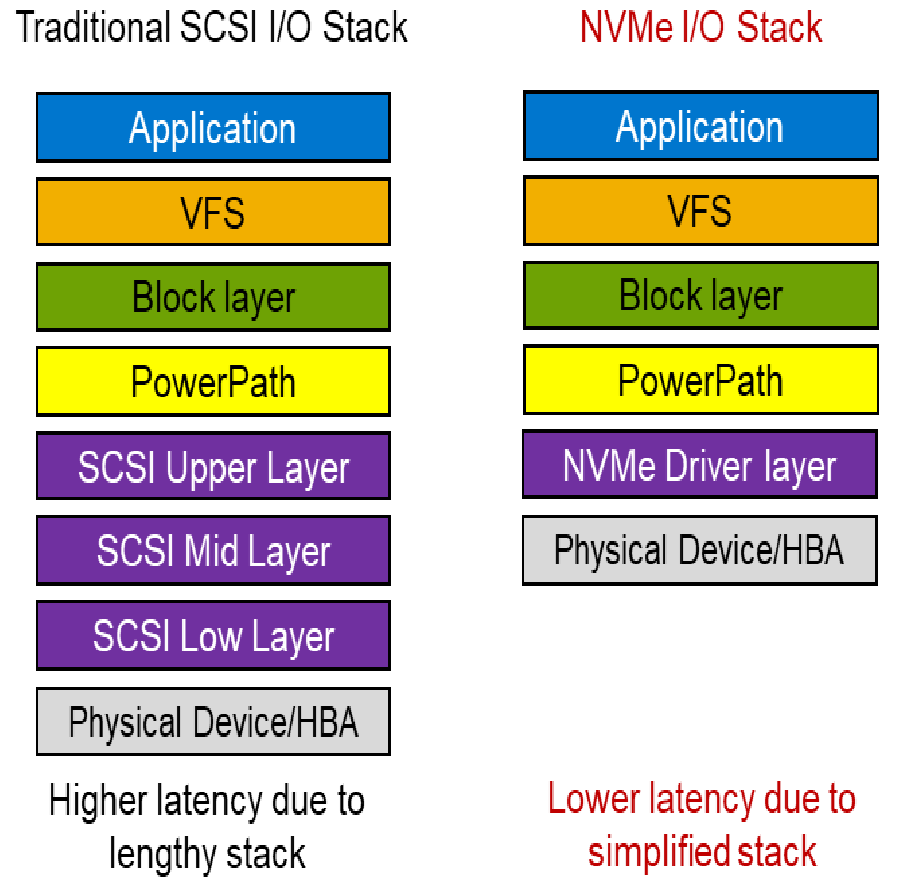
Single I/O Queue

NVMe I/O Stack

- Application
- VFS
- Block layer
- PowerPath
- NVMe Driver layer
- Physical Device/HBA

Multiple I/O Queues

Simplified stack = lower latency



Traditional SCSI I/O Stack

- Application
- VFS
- Block layer
- PowerPath
- SCSI Upper Layer
- SCSI Mid Layer
- SCSI Low Layer
- Physical Device/HBA

Higher latency due to lengthy stack

NVMe I/O Stack

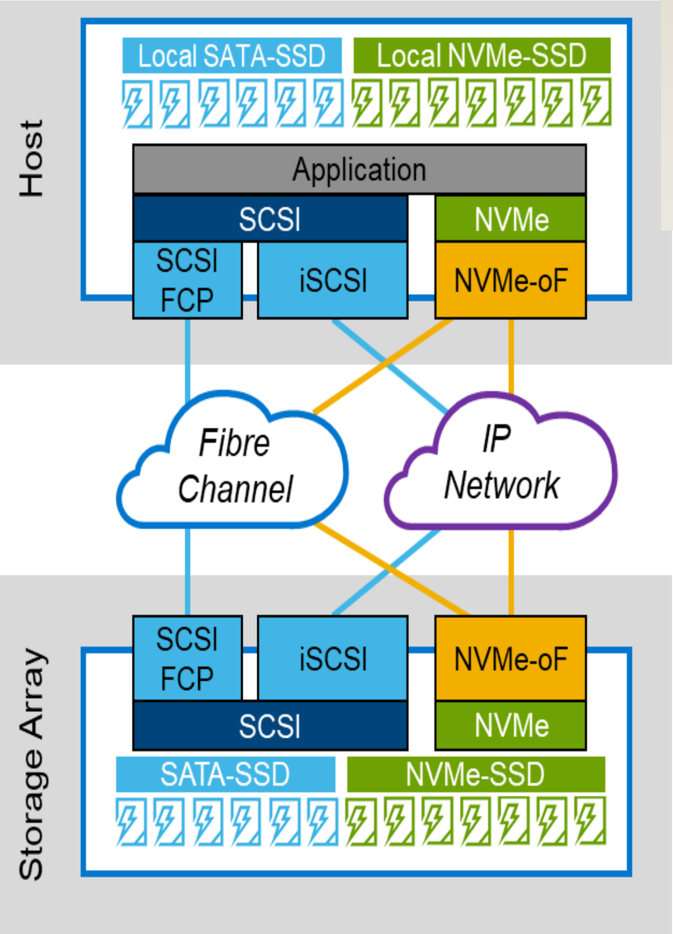
- Application
- VFS
- Block layer
- PowerPath
- NVMe Driver layer
- Physical Device/HBA

Lower latency due to simplified stack

NVMe-oF™ Evolution

From SCSI to NVMe

- Application running on a host that is accessing external array-based storage via either FC or iSCSI.
- NVMe Drives were first introduced on the host in 2015 and were used mainly for caching and boot drives
- NVMe-SSDs improve storage array performance but using the SCSI protocol can add significant latency.
- NVMe-oF™ can run over either Ethernet or Fibre channel with low latency.
    - NVMe™ over Fabrics (NVMe-oF™) protocols: NVMe/TCP, NVMe/FC, NVMe/RoCEv2, NVMe/IB, and NVMe/iWARP



Why NVMe/TCP?

- NVMe-oF/Ethernet gives customers standards-based, interoperable, high-speed, light-weight, low-latency, cost effective block storage access
  - 25 GbE is effectively the same speed as 32G FC at fraction of the cost
- NVMe-oF/TCP delivers better performance and reduced overhead compared to typical iSCSI
  - Realistic performance is similar to RoCEv2 (and it’s getting better as NVMe/TCP offloads emerge)
- TCP/IP for NVMe-oF transport just works by default
  - Specialized configuration of TCP ***is not required***
- TCP/IP is a better fit to Edge, IoT, Client deployments due to price & hardware
- TCP/IP allows a wide variety of network topologies (fully routable and fully flow controlled as needed)

NVMe/TCP - NVMe-oF Discovery Problem

- End-point Centric (vs Network-Centric) Model
  - Configuration for storage access on each Host
  - Addition / removal of NVM subsystems
- Scalability concerns with more than a few Hosts and NVM subsystems
  - Lack of automation; therefore, complexity in NVMe/TCP environments

Dell Contributions to NVM Express Standards

Tech Proposal (TP)	Status	Description
TP-8006	Published	Authentication
TP-8011	Published	Encryption (TLS 1.3)
TP-8009	Phase 3	Automatic discovery of NVMe-oF Discovery Controllers
TP-8010	Phase 3	Centralized Discovery Controller (CDC)
TP-8012 (boot)	In progress	Boot from NVMe-oF (Standard nBFT)
TP-4126 (boot)	In progress	Incorporate (FC-NVMe) requirements into NVM Express specification.

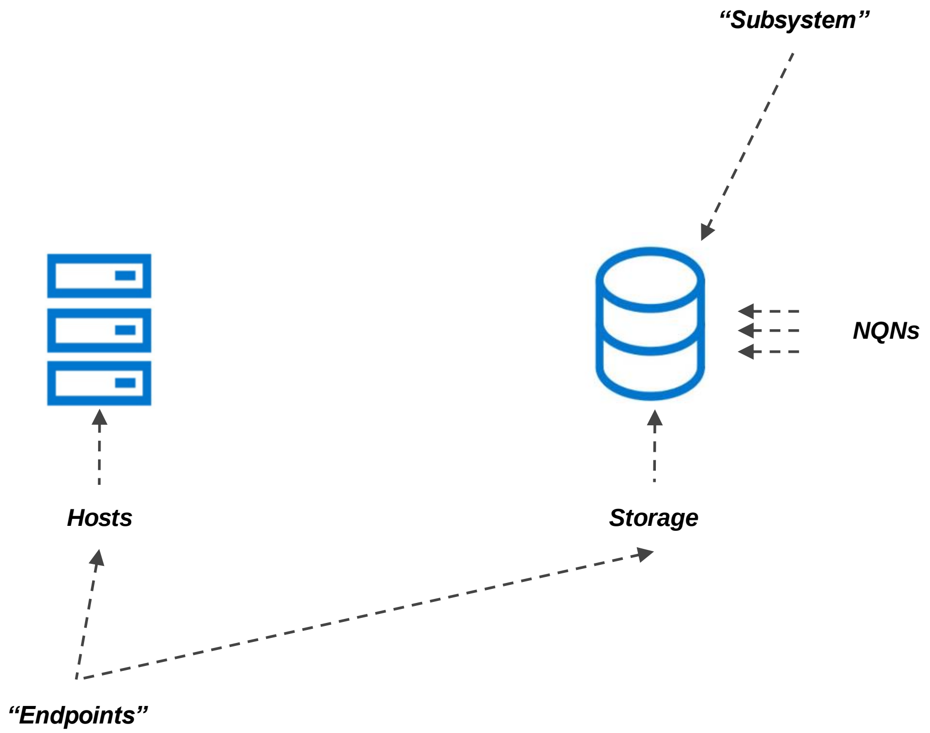
- Dell Technologies driving industry ecosystem through NVM express standards committees
- Significant Dell Technologies investment in standards development along with many other companies.
- TP-8009 and TP-8010 ratification expected later this year.

Agenda

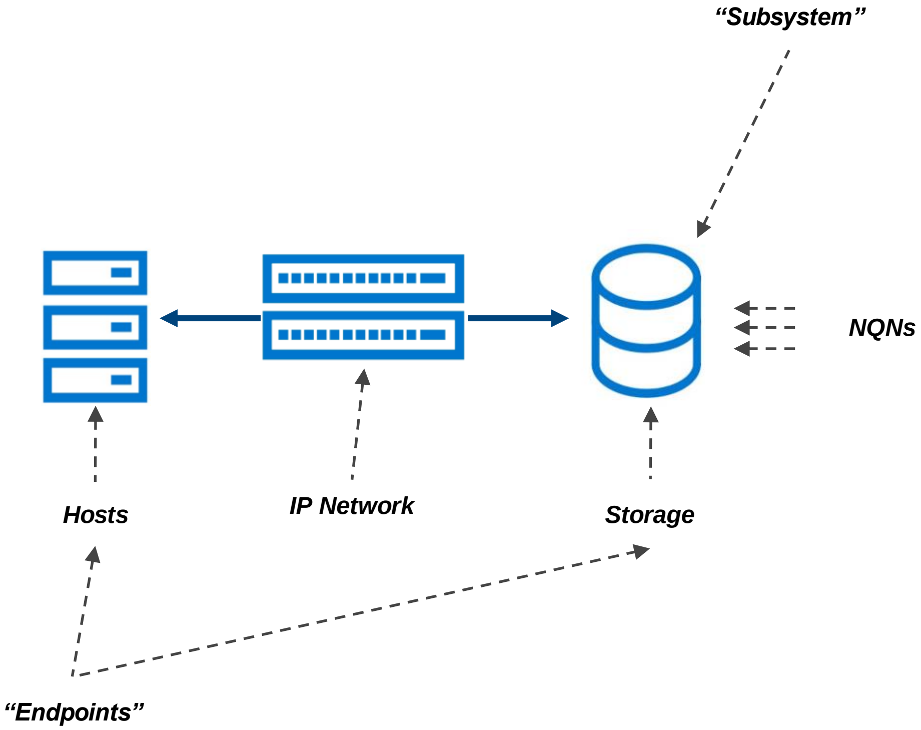
- Market Drivers & Challenges
- Evolution: SCSI, NVMe, NVMeoF, NVMe/TCP
- Next-Gen NVMe IP SAN
- Dell Technologies’ NVMe IP SAN Ecosystem
- Dell-EMC NVMe IP SAN Software
- VMware NVMe Support
- Dell-EMC PowerStore Storage System
- Summary

Next-Gen NVMe IP SAN

- **Endpoints**: Standard-compliant hosts and storage systems
  - e.g., ESXi & Dell-EMC PowerStore
  - NVMe subsystems are present on storage systems and defined by NVMe Qualified Names
    - (NQN, e.g. nqn.1988-11.com.dell:powerstore:00:6c1ee24aa6adACBC6314)

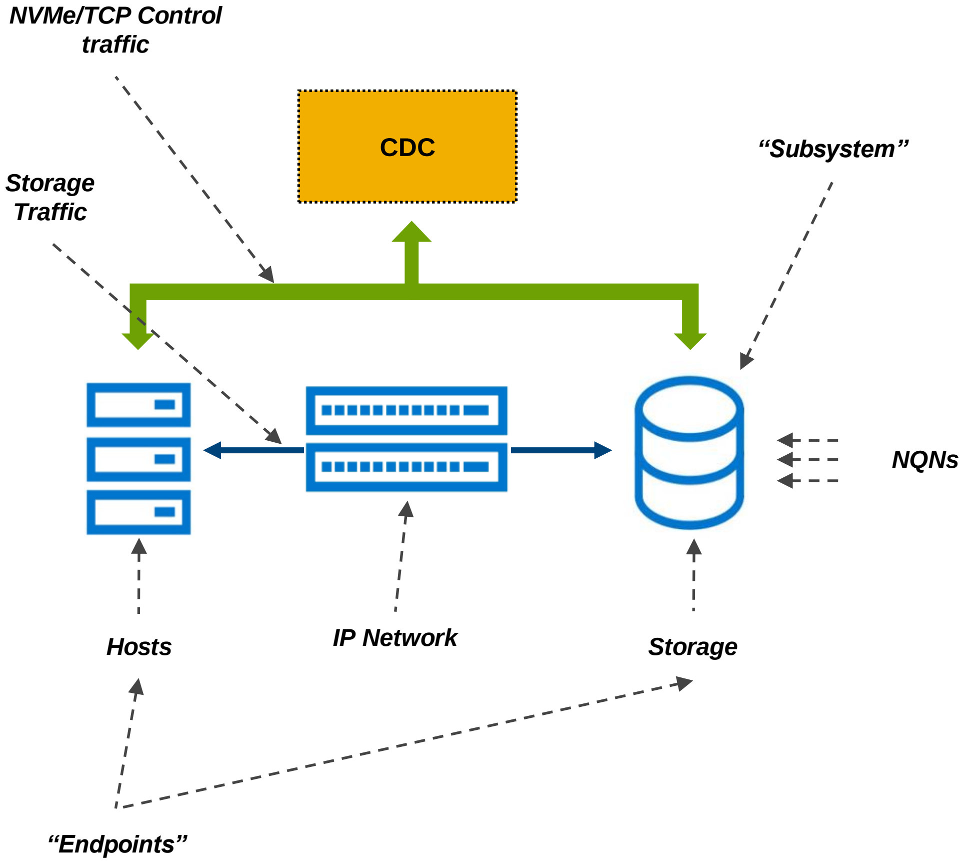


Next-Gen NVMe IP SAN



- **Endpoints**: Standard-compliant hosts and storage systems
  - ESXi & Dell-EMC PowerStore
  - NVMe subsystems are present on storage systems and defined by NVMe Qualified Names
    - (NQN, e.g. nqn.1988-11.com.dell:powerstore:00:6c1ee24aa6adACBC6314)
- **IP Network**:
  - Dell PowerSwitch for end-end solution but can be any vendor

Next-Gen NVMe IP SAN



- **Endpoints**: Standard-compliant hosts and storage systems
  - ESXi & Dell-EMC PowerStore
  - NVMe subsystems are present on storage systems and defined by NVMe Qualified Names
    - (NQN, e.g. nqn.1988-11.com.dell:powerstore:00:6c1ee24aa6adACBC6314)
- **IP Network**:
  - Dell PowerSwitch for end-end solution but can be any vendor
- **CDC:** Centralized Discovery Controller
  - A CDC is a controller for NVMe/TCP Endpoints that are taking part in NVMe™ over TCP protocol standards.
  - **Dell-EMC NVMe IP SAN Software** is Dell's implementation of a CDC.

CDC – Main Functions

- Discovery Service
  - NVMe/TCP endpoints dynamically discover the CDC instance
  - Listen and respond to mDNS queries from endpoints in the fabric
- End Point Registration Service
  - NVMe/TCP endpoints – host or subsystem registers their information with CDC
- Endpoint Query Service
  - NVMe/TCP hosts and subsystems can query CDC to discover each other
- Zone Service
  - Soft Zoning: GetLogPage responses only include NVMe subsystems zoned for the query host
  - Hard Zoning: Enforcement in the network with ACLs (integration with network switches)
- Asynchronous Notifications
  - Subscribe to state change notifications from endpoints and send these notifications to other endpoints for state changes

Equivalent of Fibre Channel

Name Server Database

Zone Server Database

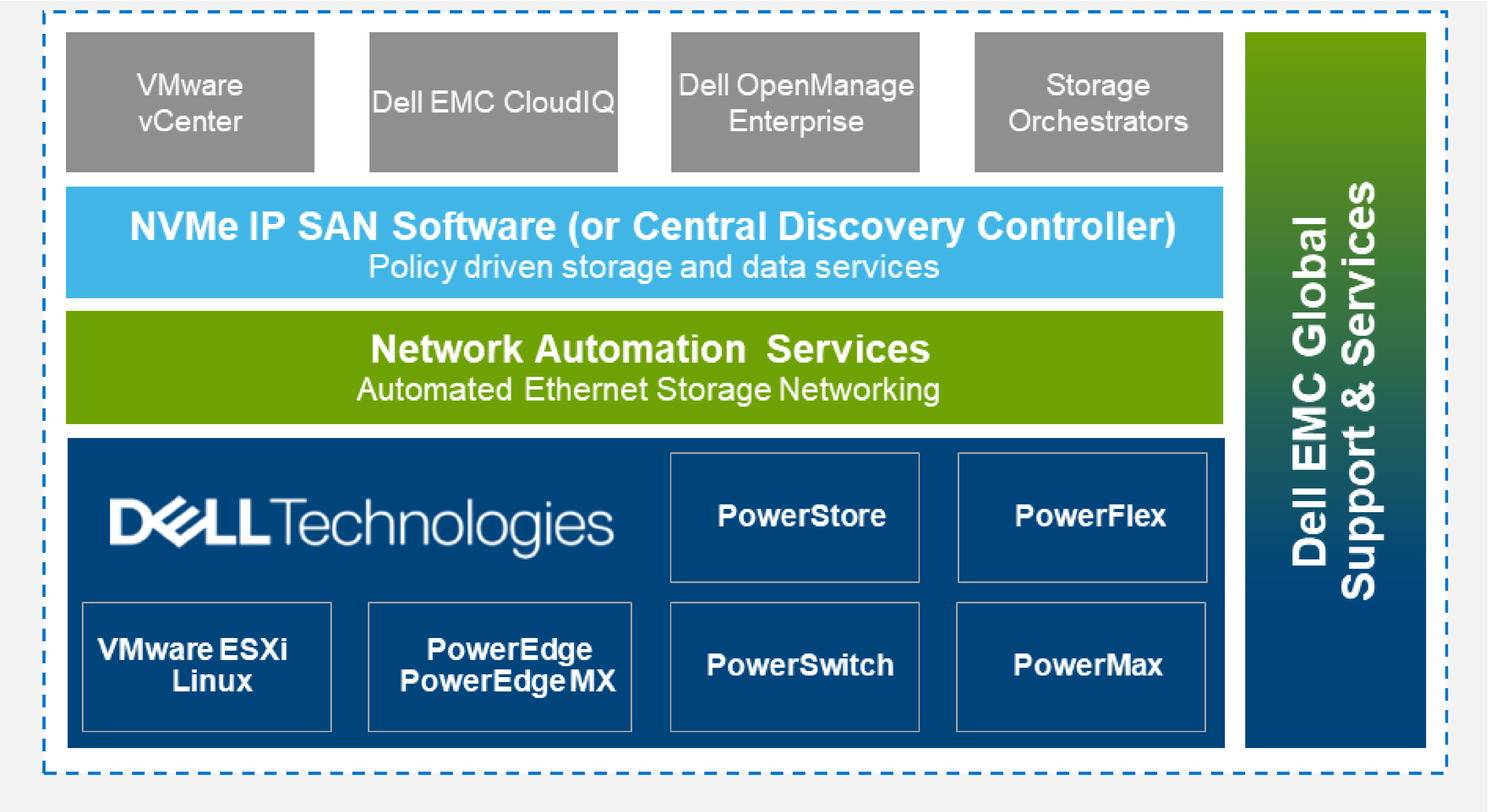
Registered State Change Notification (RSCN)

Agenda

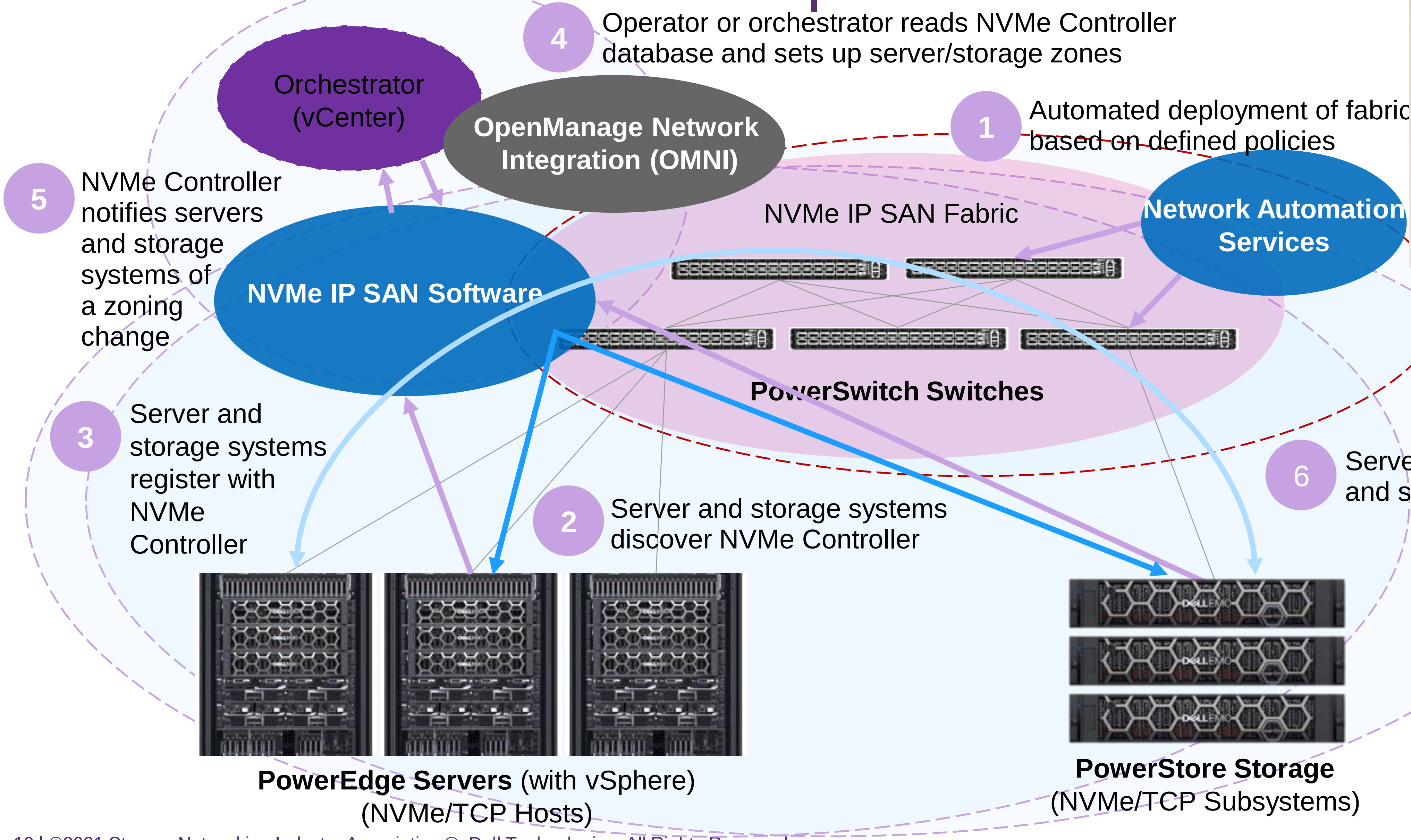
- Market Drivers & Challenges
- Evolution: SCSI, NVMe, NVMeoF, NVMe/TCP
- Next-Gen NVMe IP SAN
- Dell Technologies' NVMe IP SAN Ecosystem
- Dell-EMC NVMe IP SAN Software
- VMware NVMe Support
- Dell-EMC PowerStore Storage System
- Summary

Dell Tech NVMe IP SAN Ecosystem

Standard-based, end-end NVMe/TCP storage connectivity


VMware vCenter
Dell EMC CloudIQ
Dell OpenManage Enterprise
Storage Orchestrators
NVMe IP SAN Software (or Central Discovery Controller)
Policy driven storage and data services
Network Automation Services
Automated Ethernet Storage Networking
DELL Technologies
PowerStore
PowerFlex
VMware ESXi Linux
PowerEdge PowerEdge MX
PowerSwitch
PowerMax
Dell EMC Global Support & Services

Dell Tech NVMe IP SAN operations



1 Automated deployment of fabric based on defined policies

2 Server and storage systems discover NVMe Controller

3 Server and storage systems register with NVMe Controller

4 Operator or orchestrator reads NVMe Controller database and sets up server/storage zones

5 NVMe Controller notifies servers and storage systems of a zoning change

6 Servers connect to storage and start transferring data

PowerEdge Servers (with vSphere)
(NVMe/TCP Hosts)

PowerStore Storage
(NVMe/TCP Subsystems)

Dell Tech Next-Gen Storage Connectivity

Fast, simple, cost-effective, Ethernet-based alternative to Fibre Channel

Performance

- Comparable performance to FC at a fraction of the cost
- Superior performance & less complexity compared to iSCSI
- Consistent IO and latency performance under varying load conditions

Better Economics than FC

- Ethernet infrastructure lower cost than FC
- Leverage existing ethernet infrastructure
- Less operational overhead than iSCSI
- Standards-based infrastructure

Simple & Automated

- NVMe IP SAN Software automates storage configuration
- Network Automation Services automates network configuration
- Standards-based and interoperable
- Supported across Dell
- Accelerates transition to modular composable infrastructure

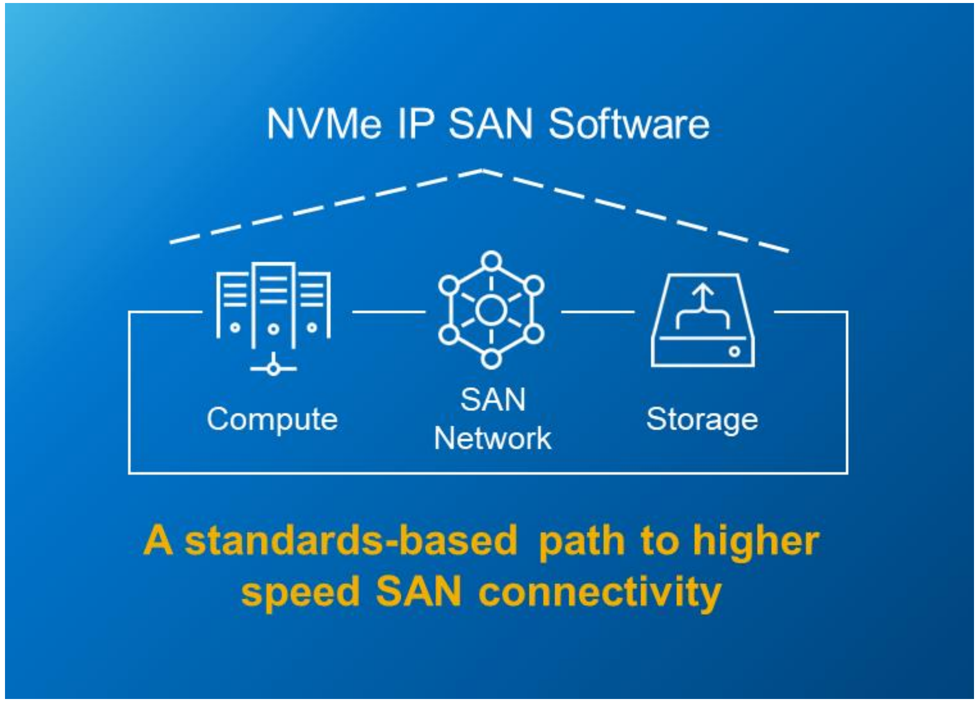
NVMe-oF / TCP: Standard, interoperable, light-weight, low latency, cost effective block storage access

Agenda

- Market Drivers & Challenges
- Evolution: SCSI, NVMe, NVMeoF, NVMe/TCP
- Next-Gen NVMe IP SAN
- Dell Technologies' NVMe IP SAN Ecosystem
- Dell-EMC NVMe IP SAN Software
- VMware NVMe Support
- Dell-EMC PowerStore Storage System
- Summary

Dell-EMC NVMe IP SAN Software

Standards based Centralized Discovery Controller (CDC) for NVMe/TCP hosts and storage subsystems



- **A containerized application,** enabling an end to end automated and integrated NVMe IP SAN fabric.
- **A policy-driven, Centralized Discovery Controller** to provide automated, NVMe-oF storage service discovery, end-point registration, connectivity and zoning services
- **Implements enhancements to the NVMe Standards**
  - TP-8009 – Automated Discovery of NVMe-oF Discovery Controllers
  - TP-8010 - Centralized Discovery Services

Agenda

- Market Drivers & Challenges
- Evolution: SCSI, NVMe, NVMeoF, NVMe/TCP
- Next-Gen NVMe IP SAN
- Dell Technologies’ NVMe IP SAN Ecosystem
- Dell-EMC NVMe IP SAN Software
- VMware NVMe Support
- Dell-EMC PowerStore Storage System
- Summary

Disclaimer

VMware 2019

- This presentation may contain product features or functionality that are currently under development.

- This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.

- Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.

- Technical feasibility and market demand will affect final delivery.

- Pricing and packaging for any new features/functionality/technology discussed or presented, have not been determined.

- This information is confidential.

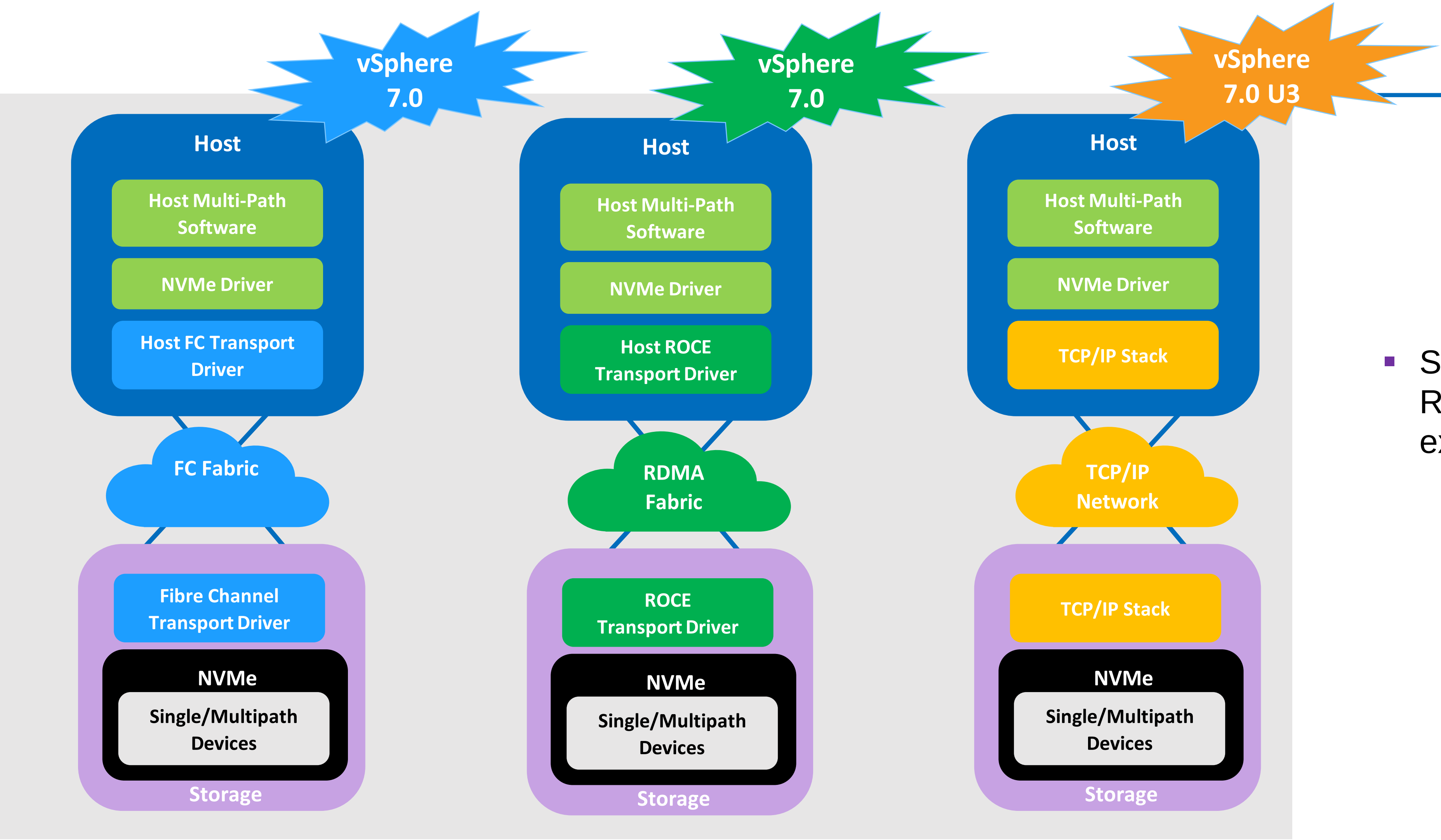
The information in this presentation is for informational purposes only and may not be incorporated into any contract. There is no commitment or obligation to deliver any items presented herein.

VMware NVMe Support

Murali Rajagopal

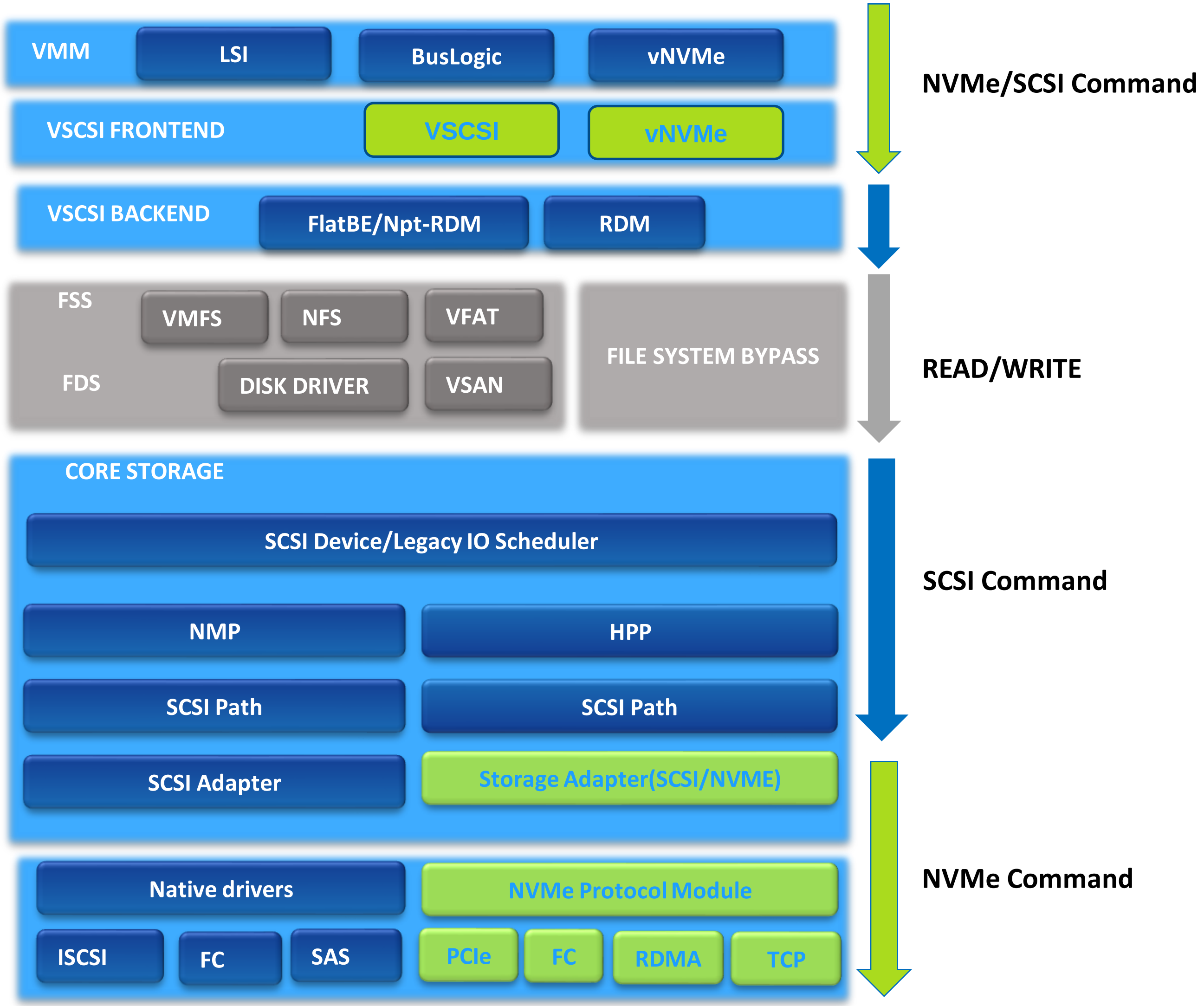
NVM Express Over Fabric (NVMe-oF)

vSphere external storage



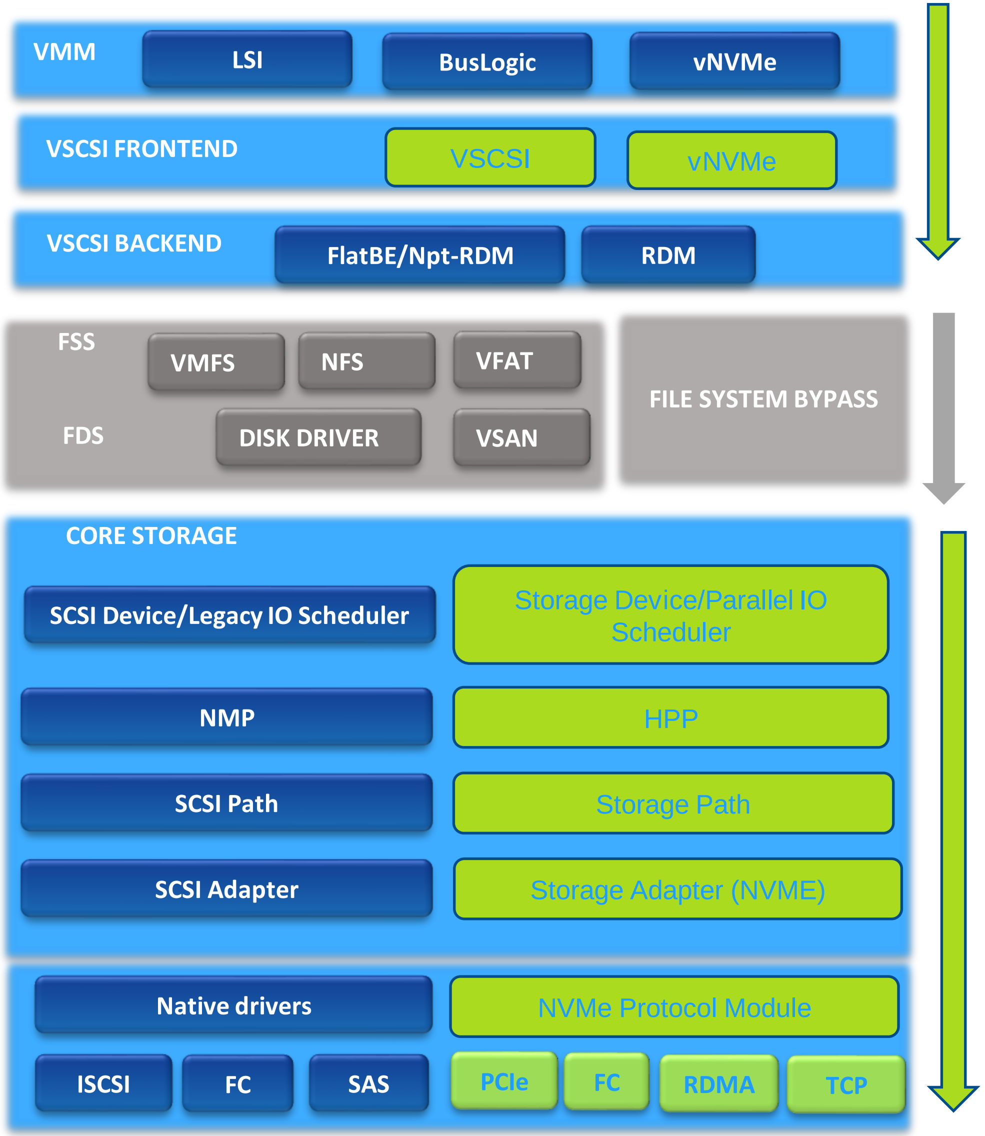
- Support NVMe-oF for FC, RoCEv2 and TCP based external storage arrays

ESXi Storage Stack – vSphere 7.0 U3



- Supports both NVMe and NVMe-oF (A/A and ANA)
- SCSI to NVMe translation at Storage (PSA) Adapter layer
- Supports Native NVMe driver
- Improved performance at vNVMe

Next Generation Storage Stack – vSphere Future



NVMe/SCSI Command

READ/WRITE

NVMe Command

- E2E NVMe
- High Performance (lower stack latency)
- A number of feature enhancements (e.g., Discovery, Metro Cluster, Boot)

NVMe support in 7.0 U2

vSphere 7.0 - 7.0 U2

- NGUID, UUID, EUI64
- Multipath support (ANA)
- Compare & Write Fused Operation
- NVMe over Fabric (FC, RoCEv2)
- NVMe Write Zeroes
- Persisted Controller Connection
- NVMe-oF Discovery (Partial support for TP 8002)

NVMe support in 7.0 U3

vSphere 7.0 - 7.0 U2

- NGUID, UUID, EUI64
- Multipath support (ANA)
- Compare & Write Fused Operation
- NVMe over Fabric (FC, RoCEv2)
- NVMe Write Zeroes
- Persisted Controller Connection
- NVMe-oF Discovery (Partial support for TP 8002)

vSphere 7.0 U3

- NVMe TCP Initiator
- Abort Enhancements (TP 4097)
- NVMe-oF Centralized Discovery Controller (TP 8010)

NVMe support Future Outlook

vSphere 7.0 - 7.0 U2

- NGUID, UUID, EUI64
- Multipath support (ANA)
- Compare & Write Fused Operation
- NVMe over Fabric (FC, RoCEv2)
- NVMe Write Zeroes
- Persisted Controller Connection
- NVMe-oF Discovery (Partial support for TP 8002)

vSphere 7.0 U3

- NVMe TCP Initiator
- Abort Enhancements (TP 4097)
- NVMe-oF Centralized Discovery Controller (TP 8010)

vSphere (Future)

- Virtual Volumes w/NVMe
- E2E NVMe
- NVMe Telemetry (and OCD 1.0)
- Enhanced Command Retry (TP 4033)
- NVMe Reservations for clustered VMDK
- Non-Data-Transfer (non-MDTS) Command Size Limits (TP 4040)
- Metro Cluster (Dispersed Namespaces TP 4034)
- NVMe-oF Discovery (TP 8002 Full support)
- Automated Discovery of NVMe-oF Discovery Controllers for IP Networks (TP 8009)
- NVMe-oF Inband Authentication (TP 8006)
- TLS 1.3 Profile (TP 8011 – TCP Transport only))
- Copy Across Namespaces (TP 4130)
- NVMe Boot for Ethernet Network(TP 8012)

NVMe-oF TCP Support

NVMe-oF TCP Support

vSphere 7.0 - 7.0 U2

- NGUID, UUID, EUI64
- Multipath support (ANA)
- Compare & Write Fused Operation
- NVMe over Fabric (FC, RoCEv2)
- NVMe Write Zeroes
- Persisted Controller Connection
- NVMe-oF Discovery (Partial support for TP 8002)

vSphere 7.0 U3

- NVMe TCP Initiator
- Abort Enhancements (TP 4097)
- NVMe-oF Centralized Discovery Controller (TP 8010)

vSphere (Future)

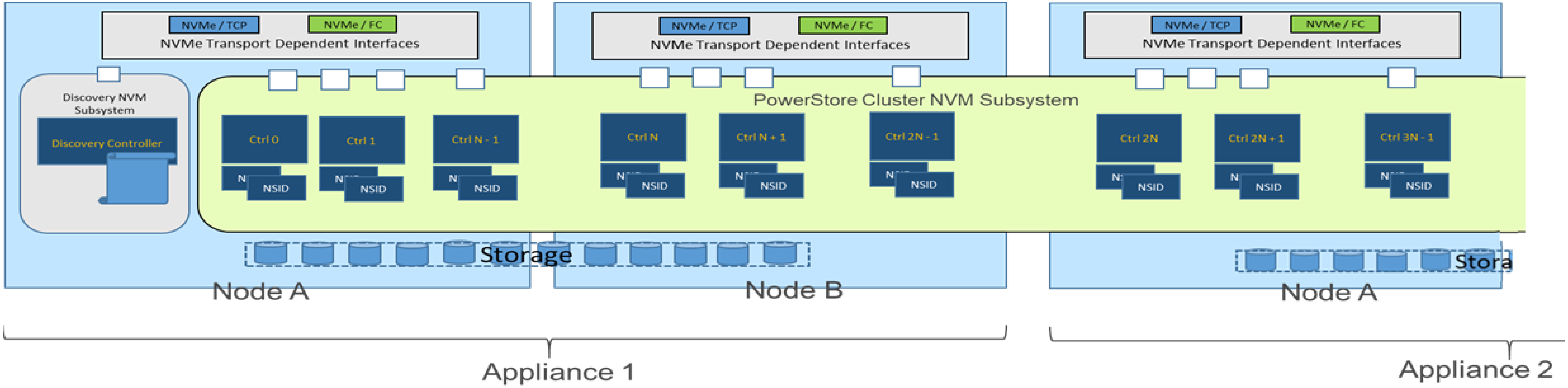
- Virtual Volumes w/NVMe
- E2E NVMe
- NVMe Telemetry (and OCD 1.0)
- Enhanced Command Retry (TP 4033)
- NVMe Reservations for clustered VMDK
- Non-Data-Transfer (non-MDTS) Command Size Limits (TP 4040)
- Metro Cluster (Dispersed Namespaces TP 4034)
- NVMe-oF Discovery (TP 8002 Full support)
- Automated Discovery of NVMe-oF Discovery Controllers for IP Networks (TP 8009)
- NVMe-oF Inband Authentication (TP 8006)
- TLS 1.3 Profile (TP 8011 – TCP Transport only))
- Copy Across Namespaces (TP 4130)
- NVMe Boot for Ethernet Network(TP 8012)

Agenda

- Market Drivers & Challenges
- Evolution: SCSI, NVMe, NVMeoF, NVMe/TCP
- Next-Gen NVMe IP SAN
- Dell Technologies’ NVMe IP SAN Ecosystem
- Dell-EMC NVMe IP SAN Software
- VMware NVMe Support
- Dell-EMC PowerStore Storage System
- Summary

Dell-EMC PowerStore

- PowerStore is an active-active NVMe, scale up and scale out, next-gen container-based storage platform.
- Each appliance has its own captive storage from which volumes (or Namespaces) would be exposed to hosts
- PowerStore supports NVMe-FC, and will support NVMe-TCP soon



Dell-EMC PowerStore NVMe/TCP Features

- **NVMeoF Specs v1.4 Compliant**
- Compatibility with standard TCP/IP and Ethernet networks
- Discovery subsystem with support for enhanced discovery, specifically TP-8002
  - Persistent Connections to the Direct Discovery Controller and Async Notifications
- Abort Command Processing
- Data & Header Digest Support
- ANA / Multi-pathing Support
  - ANA Group Change Notifications
- Volume Creation / Deletion & Namespace Attribute Change Notifications
- NVMe Reservations
- In-capsule Data Support for NVMe WRITE commands
- NGUID Support
- Support for Fused Operation (e.g., Compare & Write)
- Support Offload commands [such as Dataset Management (deallocation) and Write Zeros]
- **CDC Integration**

Dell-EMC PowerStore NVMe/TCP Features

- **NVMeoF Specs v1.4 Compliant**
- Compatibility with standard TCP/IP and Ethernet networks
- Discovery subsystem with support for enhanced discovery, sp cally TP-8002
  - Persistent Connections to the Direct Discovery Controller and Async Notifica
- Abort Command Processing
- Data & Header Digest Support
- ANA / Multi-pathing Support
  - ANA Group Change Notifications
- Volume Creation / Deletion & Namespace Attribute Change Notifications
- NVMe Reservations
- In-capsule Data Support for NVMe WRITE commands
- NGUID Support
- Support for Fused Operation (e.g., Compare & Write)
- Support Offload commands [such as Dataset Management (deallocation) and Write Zeros]
- **CDC Integration**

Futures

- NVMeoF Specs v2.0 Compliant
- Abort Enhancements (TP 4097)
- Automated Discovery of NVMe-oF Discovery Controllers for IP Networks (TP 8009)
- NVMe-oF Centralized Discovery Controller (TP 8010)
- Metro Clusters (Dispersed Namespaces TP 4034)
- NVMe-oF Inband Authentication (TP 8006)
- NVMe Copy Commands

Agenda

- Market Drivers & Challenges
- Evolution: SCSI, NVMe, NVMeoF, NVMe/TCP
- Next-Gen NVMe IP SAN
- Dell Technologies' NVMe IP SAN Ecosystem
- Dell-EMC NVMe IP SAN Software
- VMware NVMe Support
- Dell-EMC PowerStore Storage System
- Summary

Summary

- NVMe/TCP is uniquely positioned to become de facto choice for a SAN
- Next-Gen NVMe IP SAN is a standards-based, fast, simple, cost-effective, Ethernet-based alternative to Fibre Channel SAN
- Dell Technologies and VMWare working together to bring NVMe/TCP innovation and solutions

Please take a moment to rate this session.

Your feedback is important to us.