SNIA NSF | NETWORKING STORAGE

NVMe-oF™: Discovery Automation for NVMe® IP-based SANs

Live Webcast

November 4, 2021

12:00 pm PT / 3:00 pm ET

Today's Presenters

Tom Friend
Principal
Illuminosi

Erik Smith
Distinguished Member of Technical Staff
Dell Technologies

Curtis Ballard
Distinguished Technologist
Hewlett Packard Enterprise

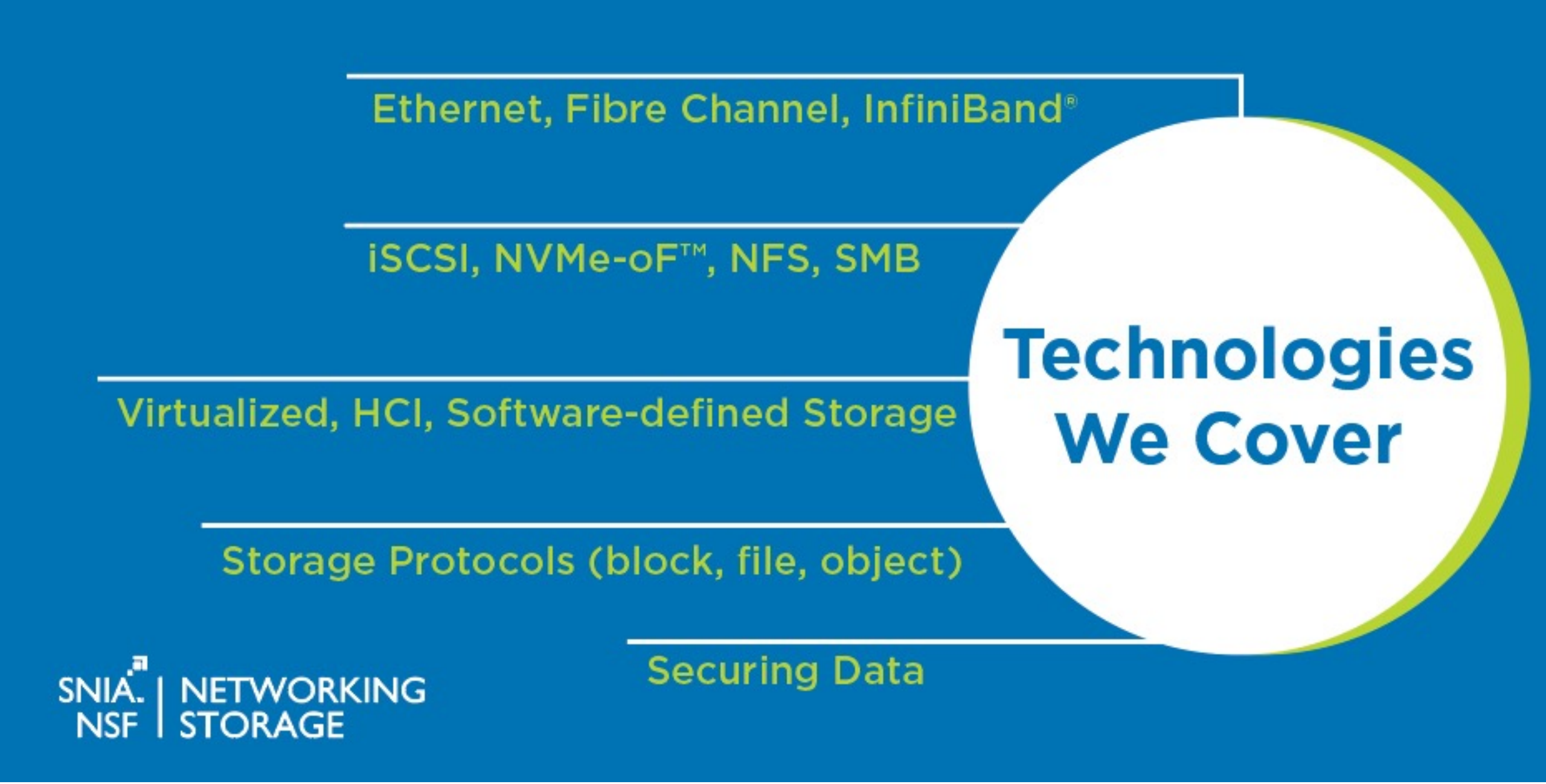
SNIA-at-a-Glance

185
industry leading organizations

2,000
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: **snia.org/technical** **@SNIA**



SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Agenda

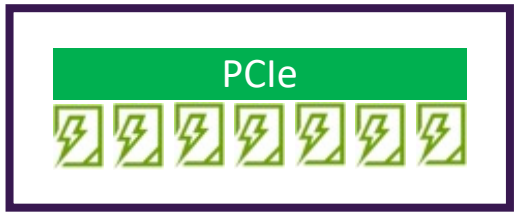
- NVMe-oF Overview and Discovery
- Direct Discovery
- Centralized Discovery
- Discovery Example

NVMe-oF: Overview and Discovery

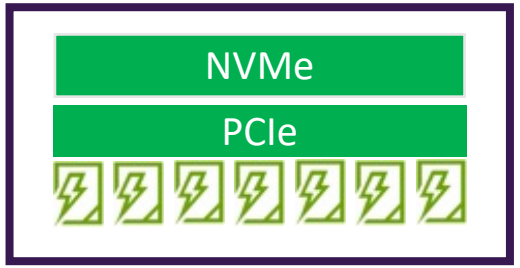
Erik Smith

NVMe, and NVMe over Fabrics

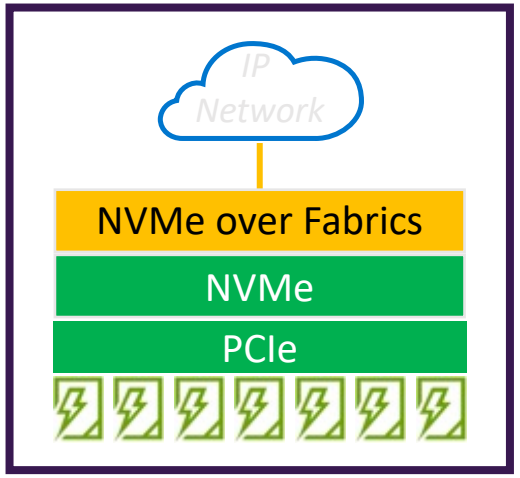
- **Non-Volatile Memory** Devices (NVM/SSD/Flash drives) can use a PCIe interface rather than a serial I/O controller.



- **Non-Volatile Memory Express (NVMe)** defines commands that allow interaction with an SSD/Flash (NVM) drive over PCIe.
  - 64K queues with up to 64K commands per queue



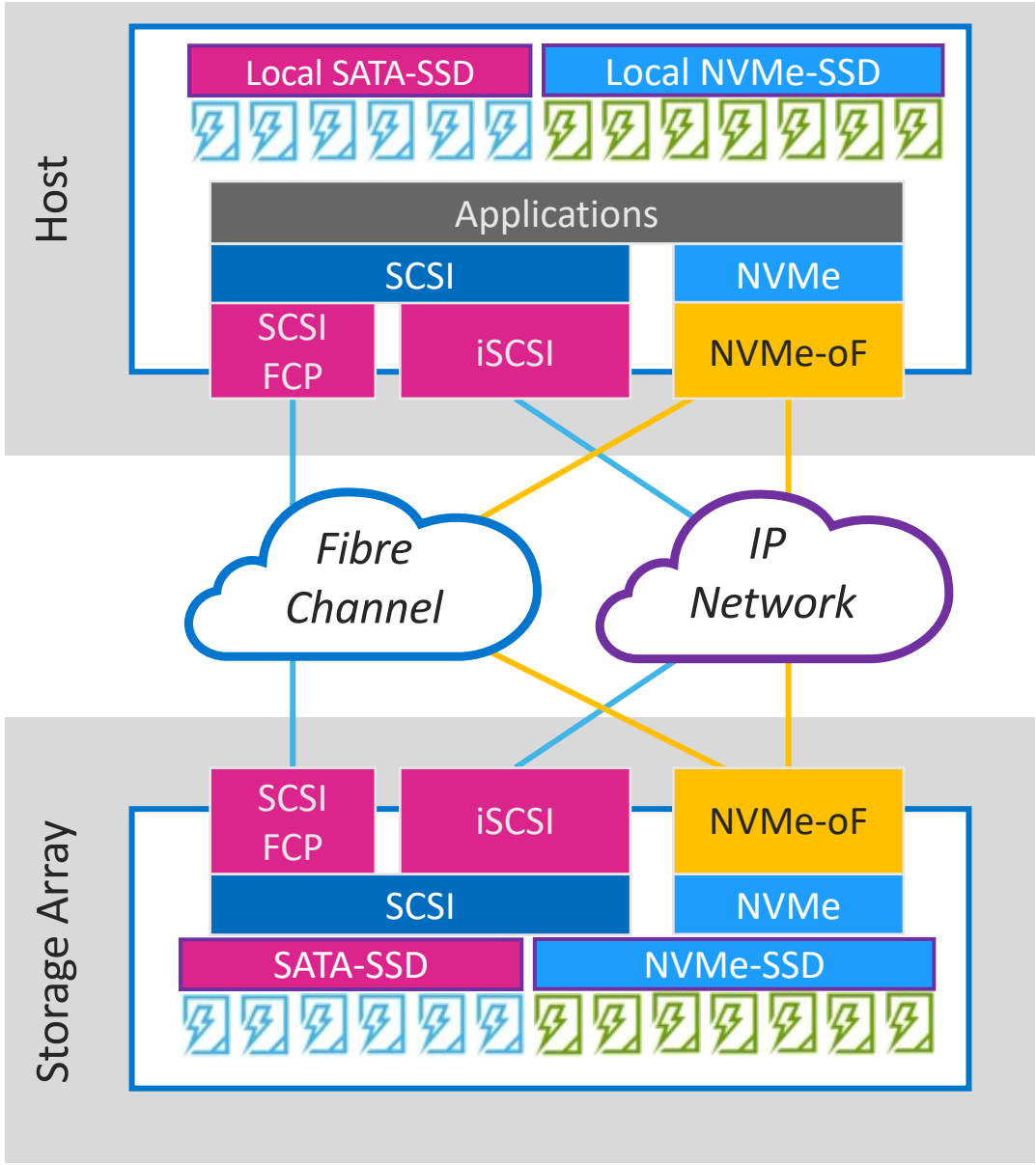
- **NVMe™ over Fabrics** describes how NVMe commands are transported over a network.
  - Fabric types include NVMe/FC, NVMe/TCP, NVMe/RoCEv2, NVMe/IB, and NVMe/iWARP.
  - *New* functionality described in TP-8009 and TP-8010 describe a standardized and scalable automated discovery process for IP networks.



SAN Evolution

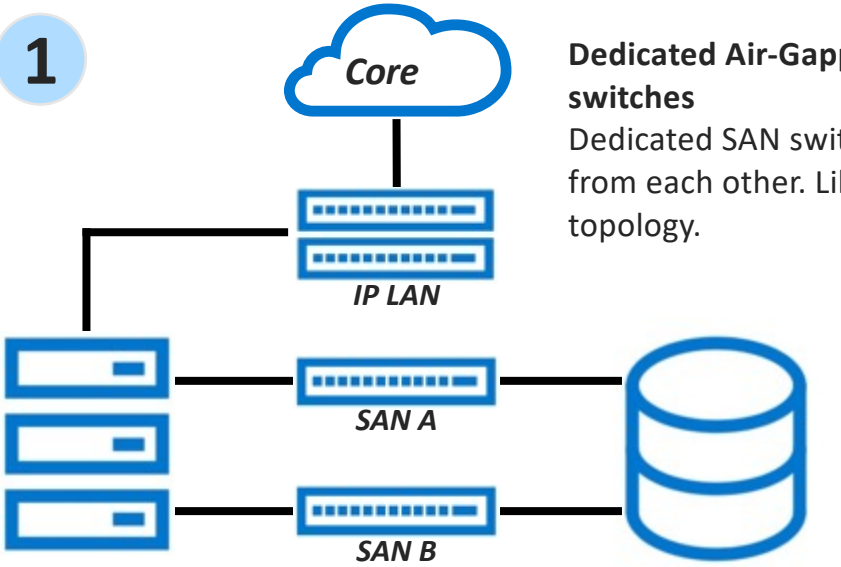
From SCSI to NVMe

- Applications running on a host that are 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 locally but using the SCSI protocol can add significant latency.
- NVMe-oF™ can run over either Fibre Channel or Ethernet.



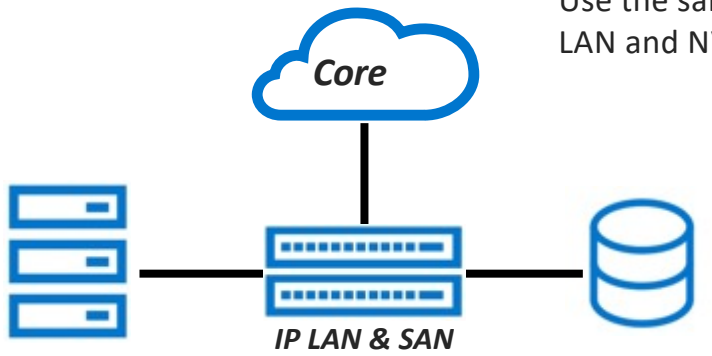
Network Topologies - Layer 1 Connectivity Requirements

1



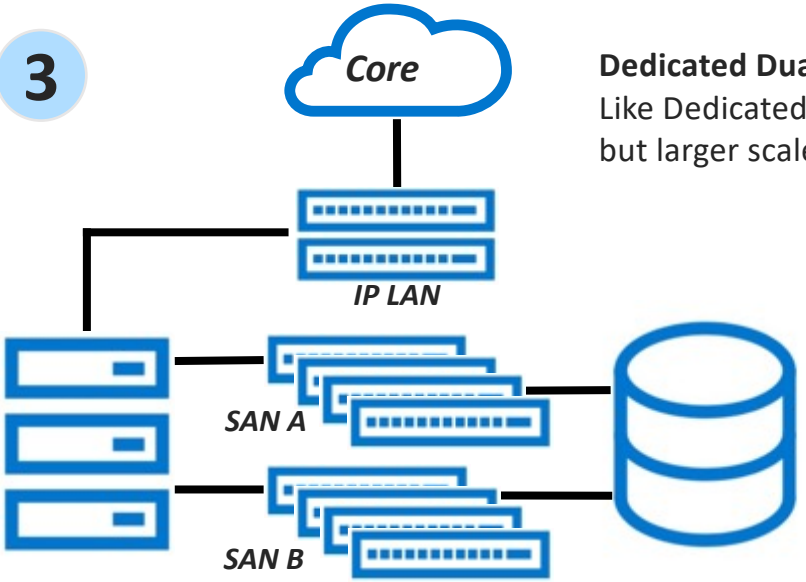
Dedicated Air-Gapped Dual SAN switches
Dedicated SAN switches, air-gapped from each other. Like **Dual FC SAN** topology.

2



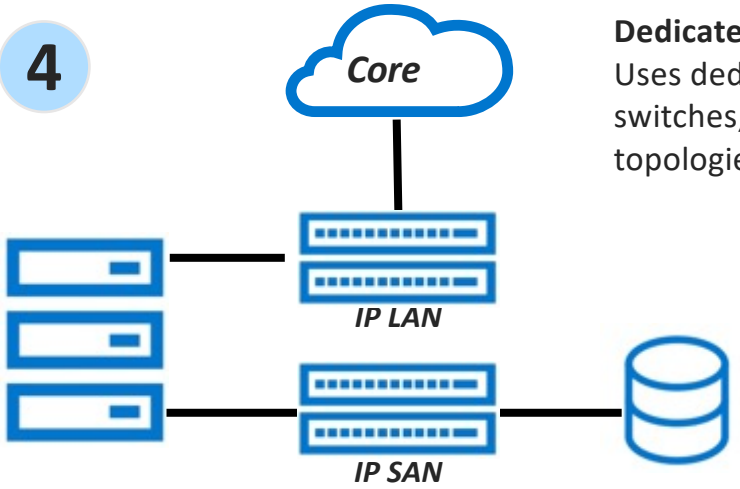
Converged LAN & SAN switches
Use the same IP Network for both LAN and NVMe/TCP storage traffic.

3



Dedicated Dual SAN Fabrics
Like Dedicated Dual SAN Switches, but larger scale

4



Dedicated SAN Switches
Uses dedicated “Back End” switches, like iSCSI and FC topologies

So… What's the Problem?

NVMe-oF's IP based Discovery Problem is well-documented and acknowledged in the standard.

> *"The method that a host uses to obtain the information necessary to connect to the initial Discovery Service is implementation specific. This information may be determined using a host configuration file, a hypervisor or OS property or some other mechanism."* – **NVMe-oF 1.1**

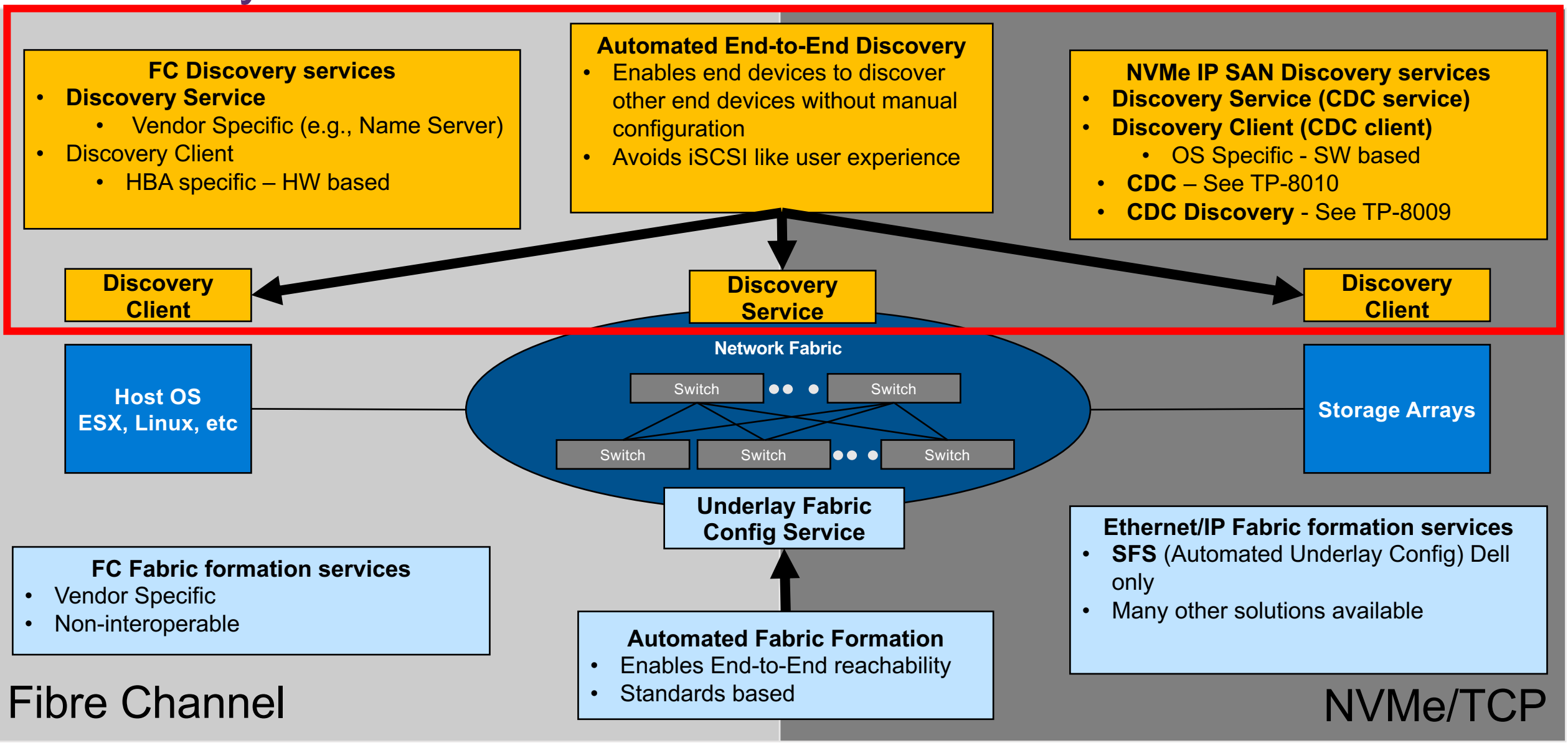
The problem? The methods described above all limit the scale and interop of any IP based NVMe-oF solution.

To address this limitation, in late 2019 a group of companies got together to see if we could agree on a standardized automated discovery process.

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.

We decided to base our approach on Fibre Channel's Fabric services. Why? FC already provides a very robust automated discovery protocol and almost everyone involved in the project had some amount of FC expertise. It turned out to be a bit more complicated than we hoped and required two separate Technical Proposals to get it done TP-8009 and TP-8010.

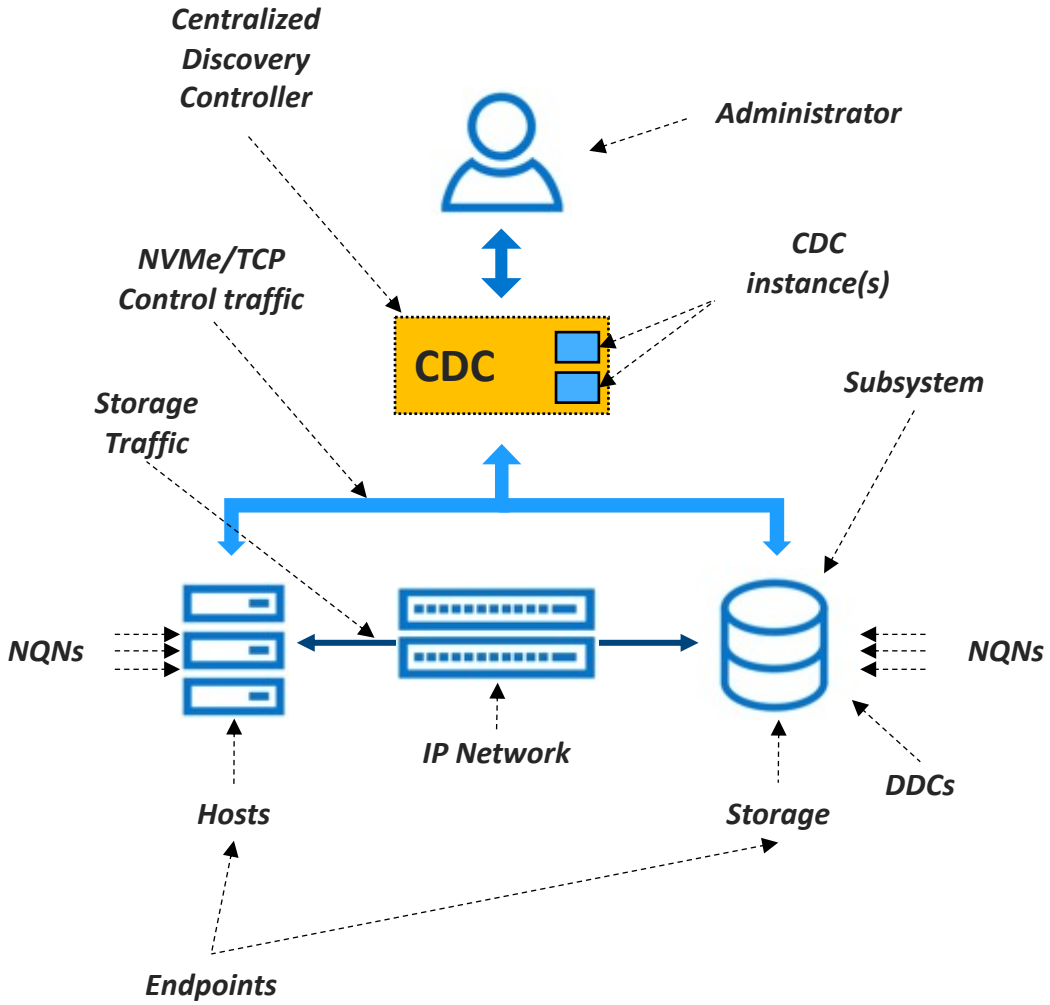
Discovery: FC vs NVMe IP-based SANs



NVMe-oF Discovery Built Around Discovery Controllers

- A Discovery controller is a single location that reports all known NVM subsystem interfaces
- Simplifies administration – A single Discovery controller IP **can** provide information about subsystem interfaces for multiple subsystems (arrays)
- The concept of a “referral” allows a Discovery controller to point to other Discovery controllers
- Common implementation today: every storage subsystem contains a discovery controller that only describes interfaces on that subsystem
- Until recently, there was no standardized method for Hosts, Subsystems or Discovery controllers to register information with a single Discovery controller (The Centralized Discovery section will cover this)

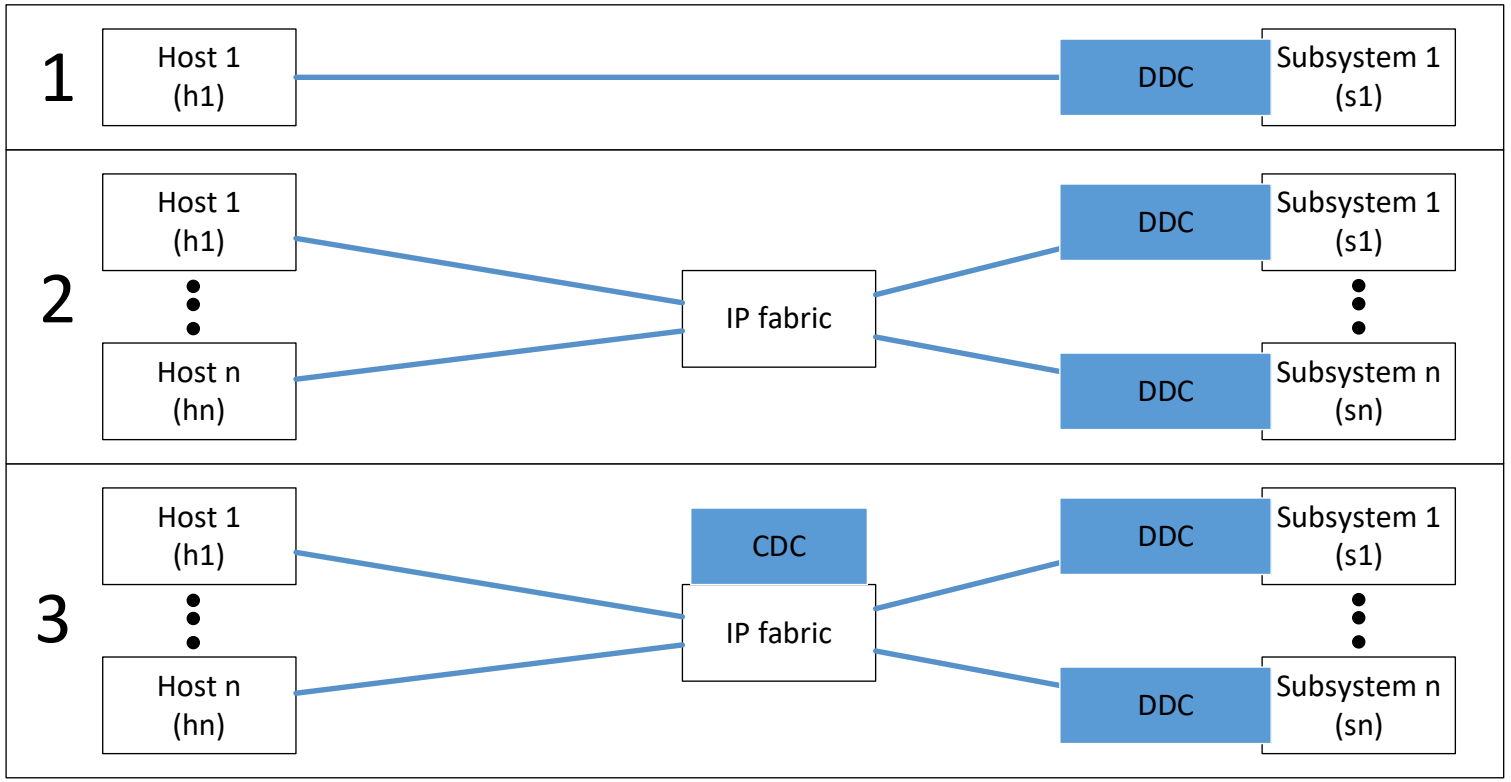
NVMe IP-based SAN Terminology



- **Endpoints**: On hosts and storage systems
  - Identified by NVMe Qualified Name (NQN) and IP Address
- **IP Network:**
  - Most modern switches (e.g., 25GbE capable and above) will work.
- **Subsystem**: Storage array, analogous to SCSI target
  - Identified by NVMe Qualified Names (NQN). NQN has a similar function to FQN in FC, and IQN in iSCSI.
- **CDC**: Centralized Discovery Controller Instances
  - Each CDC instance provides a Discovery controller for Endpoints that are taking part in a particular NVMe IP-based SAN instance.
- **DDC**: Direct Discovery Controllers
  - An NVMe Discovery controller that resides on Subsystems
  - Hosts could connect directly to storage via the DDC, but would lose the advantages of Centralization

Deployment Types that Support Automated Discovery

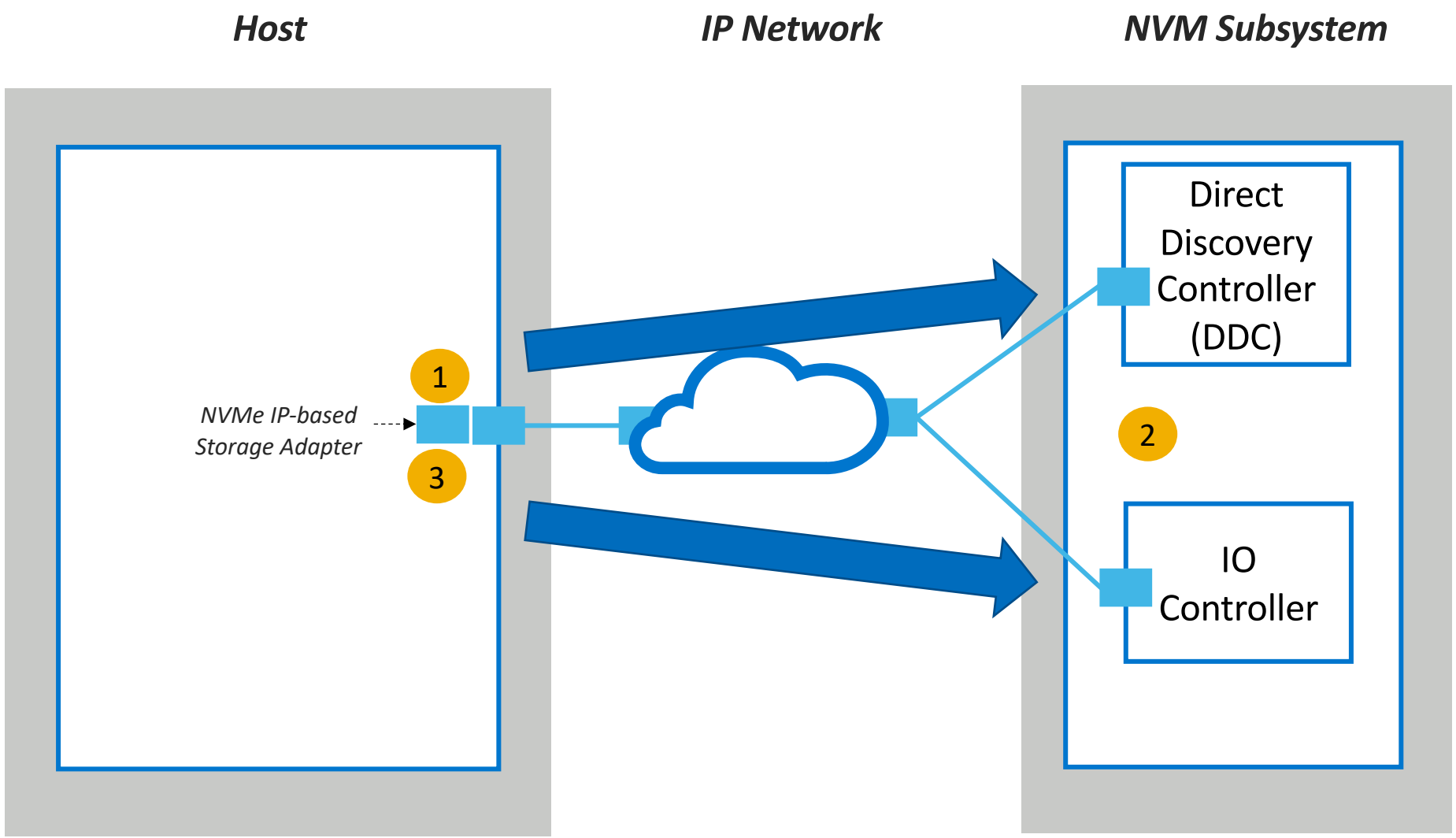
1. **Physically connected**: Host and Storage subsystem are connected by a cable
2. **Direct Discovery**: Multiple Hosts and subsystems without a CDC in the network
3. **Centralized Discovery**: Multiple Hosts and subsystems with a CDC in the network



CDC (Centralized Discovery Controller) – A Discovery controller that supports registration and zoning. Typically runs stand-alone (as a VM) or embedded on a switch in the fabric.

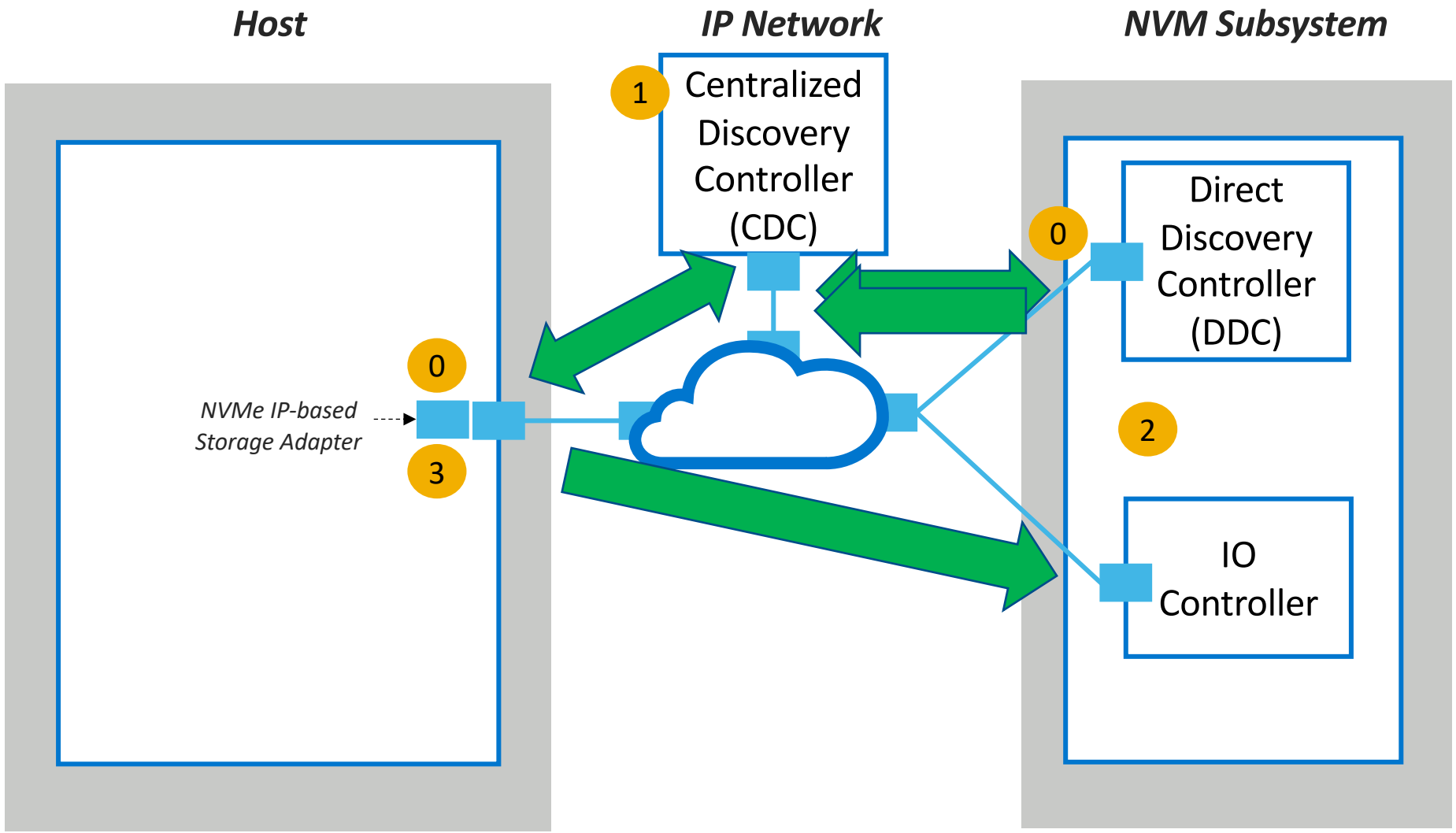
DDC (Direct Discovery Controller) – A Discovery controller that is not a CDC. Typically associated with a storage subsystem

Configuration Steps with Direct Discovery (Existing)



1. Host sends connect to Discovery Controller at IP Address supplied by admin
2. Storage admin provisions Namespaces (Storage) to the Host NQN
3. Host Admin uses nvme connect-all to Discover and connect to IO Controllers on that subsystem.
4. Repeat 1-3 on all Hosts for each subsystem

Configuration Steps with Centralized Discovery (New)



0. Host and subsystems automatically discover the CDC, connect to it and Register Discovery info
1. Zoning performed on CDC (optional)
2. Storage admin provisions namespaces to the Host NQN. Storage may send zoning info to CDC
3. After zoning, Host receives AEN, uses get log page, and connects to each IO Controller
4. Repeat 1-2 for each Hosts on each subsystem

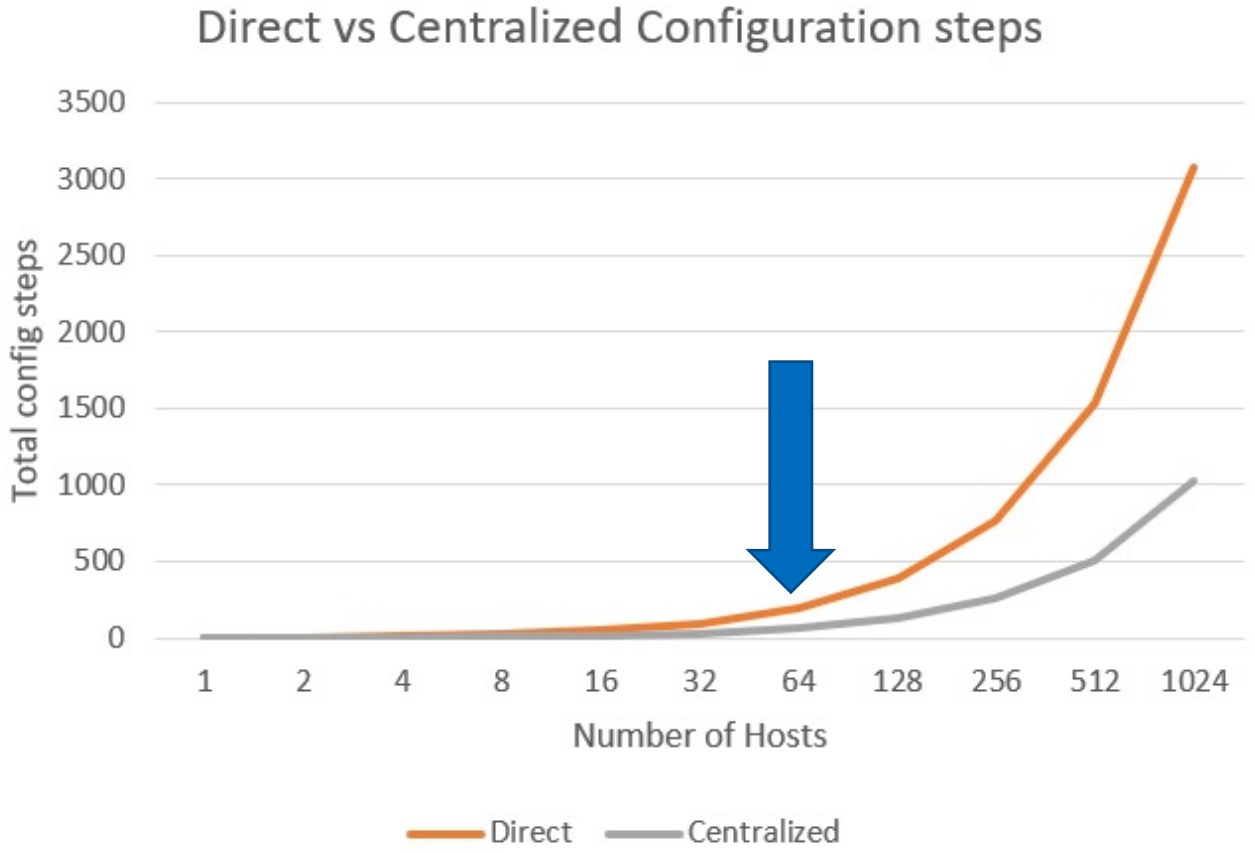
Direct vs Centralized Discovery at Scale

Direct Discovery config steps

1. **Host**: Determine subsystem Discovery controller IP -> connect
2. **Storage**: Provision storage
3. **Host**: Discover / connect all

Centralized Discovery config steps

1. **Host**: N/A
2. **CDC**: Configure Zoning (optional)
3. **Storage**: Provision storage



What the chart doesn't show

1. **Direct becomes impractical @ >64 hosts**
2. **Direct requires interaction with each host every time a storage subsystem is added or removed.**
3. **Direct may lead to extended discovery time if many subsystem interfaces are present.**

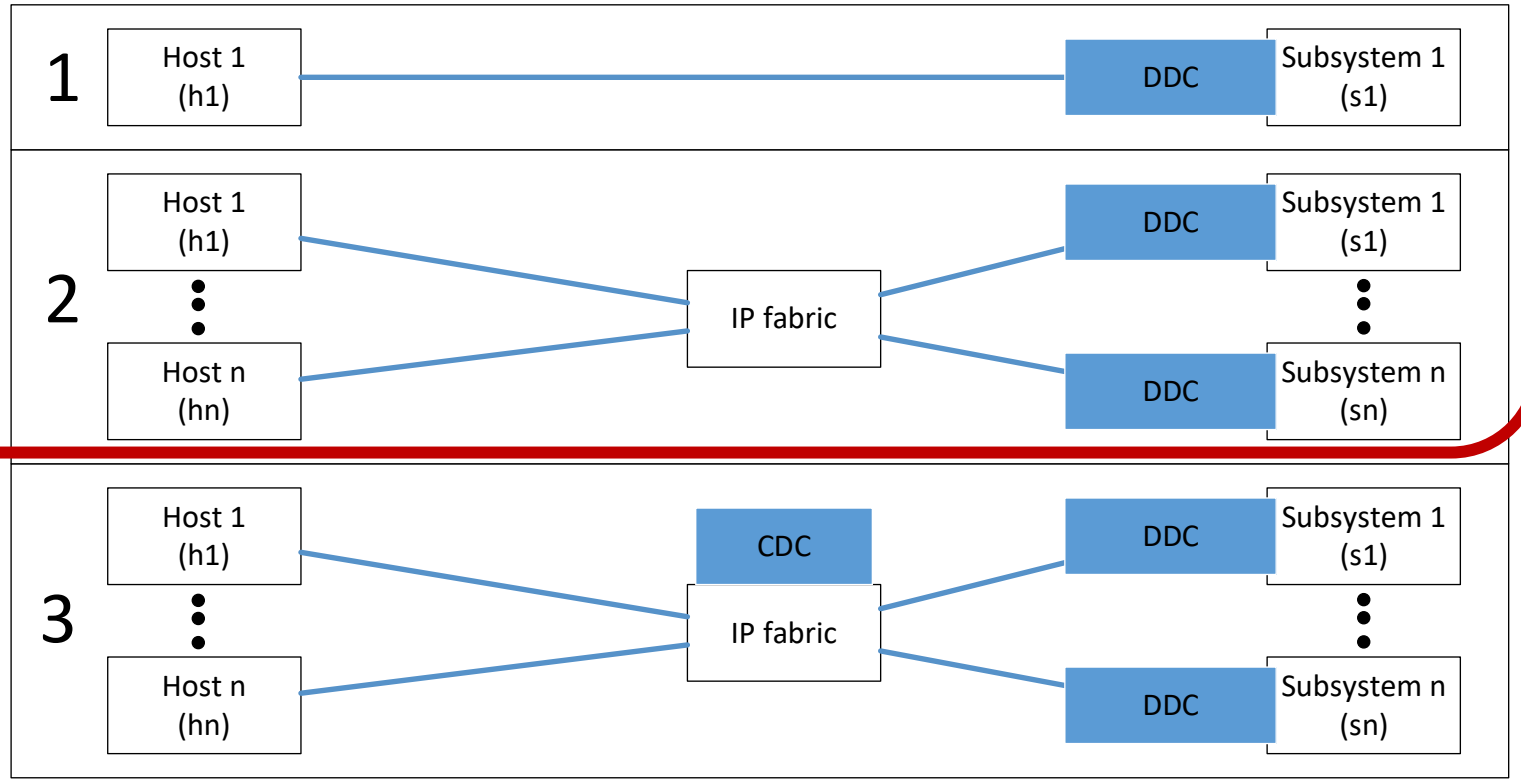
Additional Points about Discovery Automation

- Discovery Automation does not depend entirely upon a Centralized Discovery Controller (CDC).
- Smaller scale environments can make use of mDNS (as described in TP-8009) to automatically discover NVMe Discovery Controllers.
- This approach does not allow for Centralized Control, and this means:
  - Access control at the network is much more complicated/impractical
  - Hosts will not be notified when a new storage subsystem is added to the environment
- mDNS can become excessively chatty in larger configurations
  - Especially when there are more than 1000 ports in a single broadcast domain

Direct Discovery

Deployment Types that Support Automated Discovery

1. **Physically connected**: Host and Storage subsystem are connected by a cable
2. **Direct Discovery**: Multiple Hosts and subsystems without a CDC in the network
3. **Centralized Discovery**: Multiple Hosts and subsystems with a CDC in the network

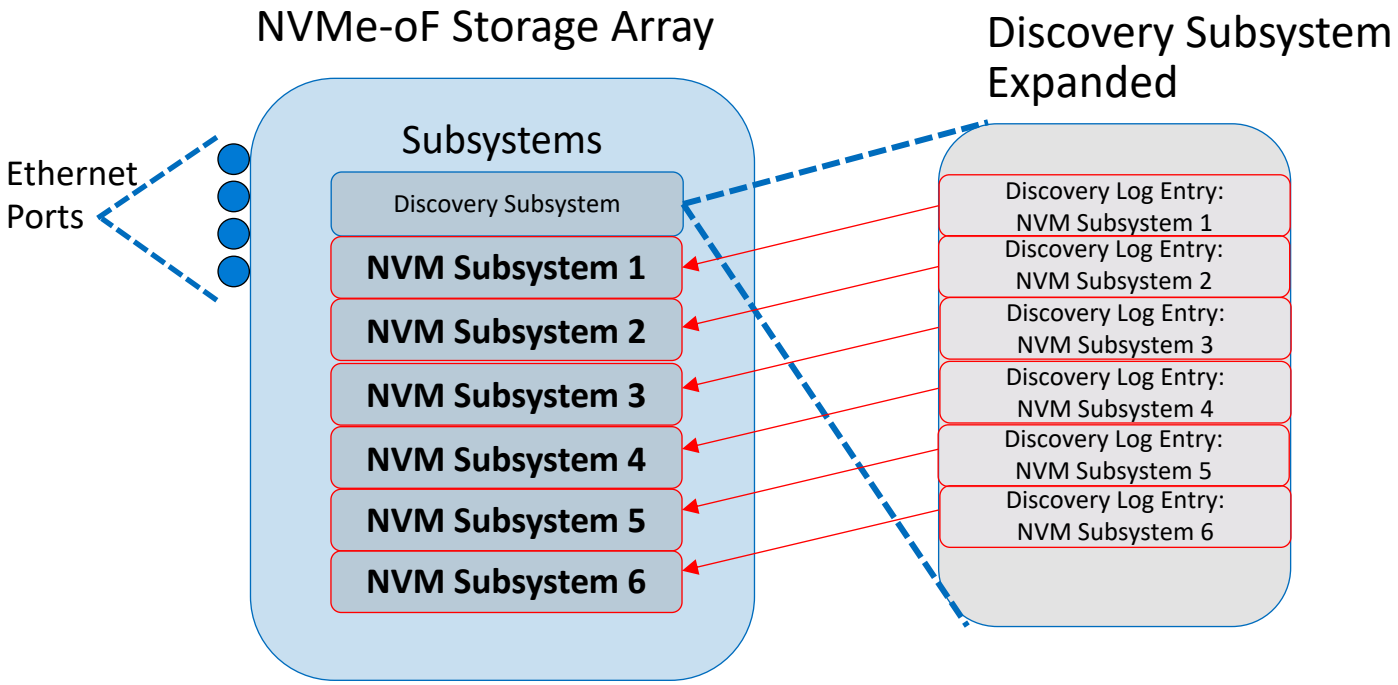


CDC (Centralized Discovery Controller) – A Discovery controller that supports registration and zoning. Typically runs stand-alone (as a VM) or embedded on a switch in the fabric.

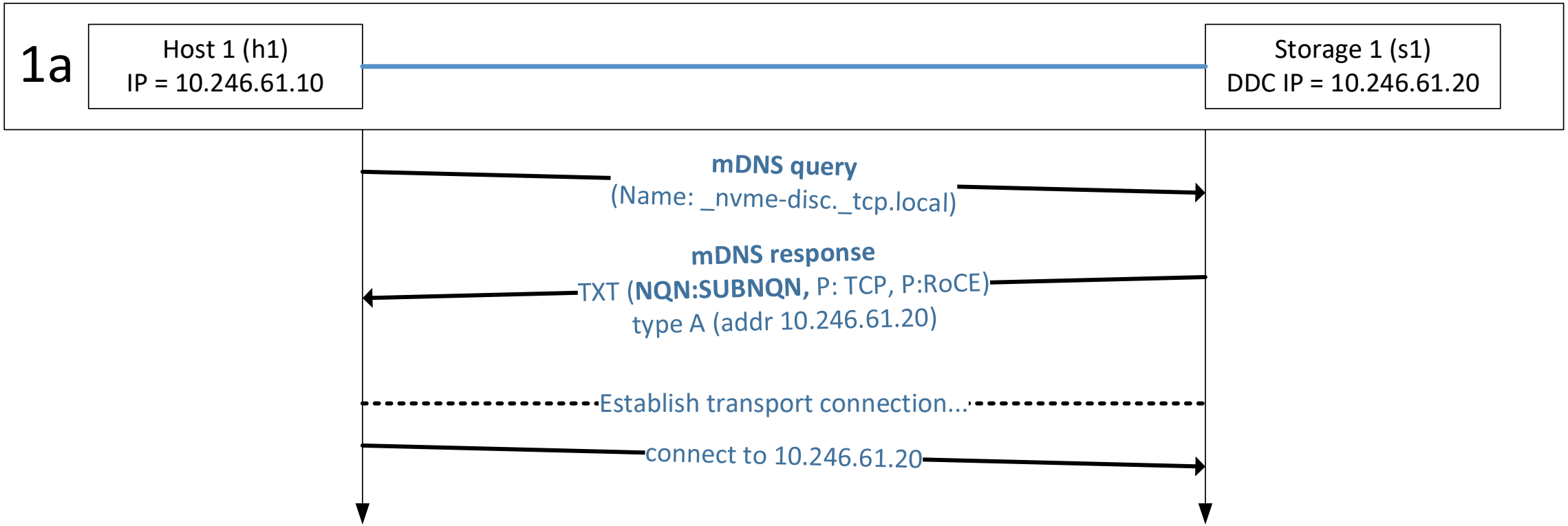
DDC (Direct Discovery Controller) – A Discovery controller that is not a CDC. Typically associated with a storage subsystem

Discovery of NVM Subsystems

- Administrator either:
  - provides IP address and transport type for a Storage System Port that contains a Discovery Controller, or
  - Discovers this IP Address (e.g., mDNS)
- Host connects to Discovery Controller in an entity called a Discovery Subsystem
- Host reads Discovery Log entries describing NVM Subsystems that the host is allowed to access
- Host connects to each accessible subsystem one path/connection at a time

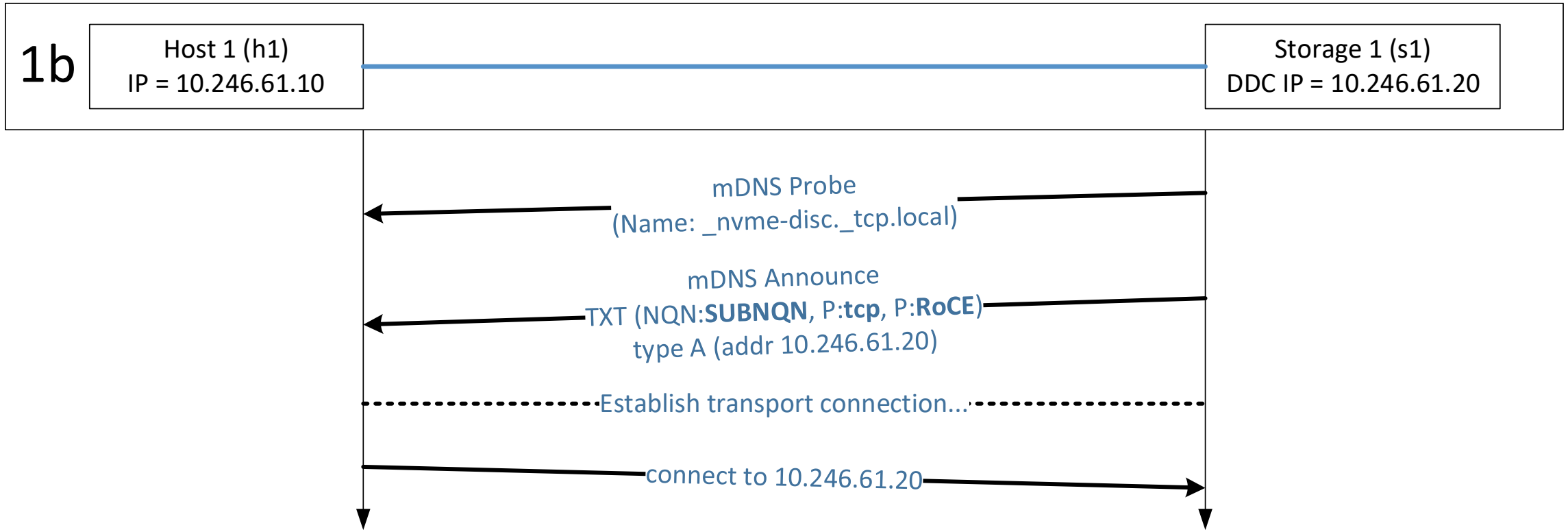


Direct Connect: A New Host Comes Online



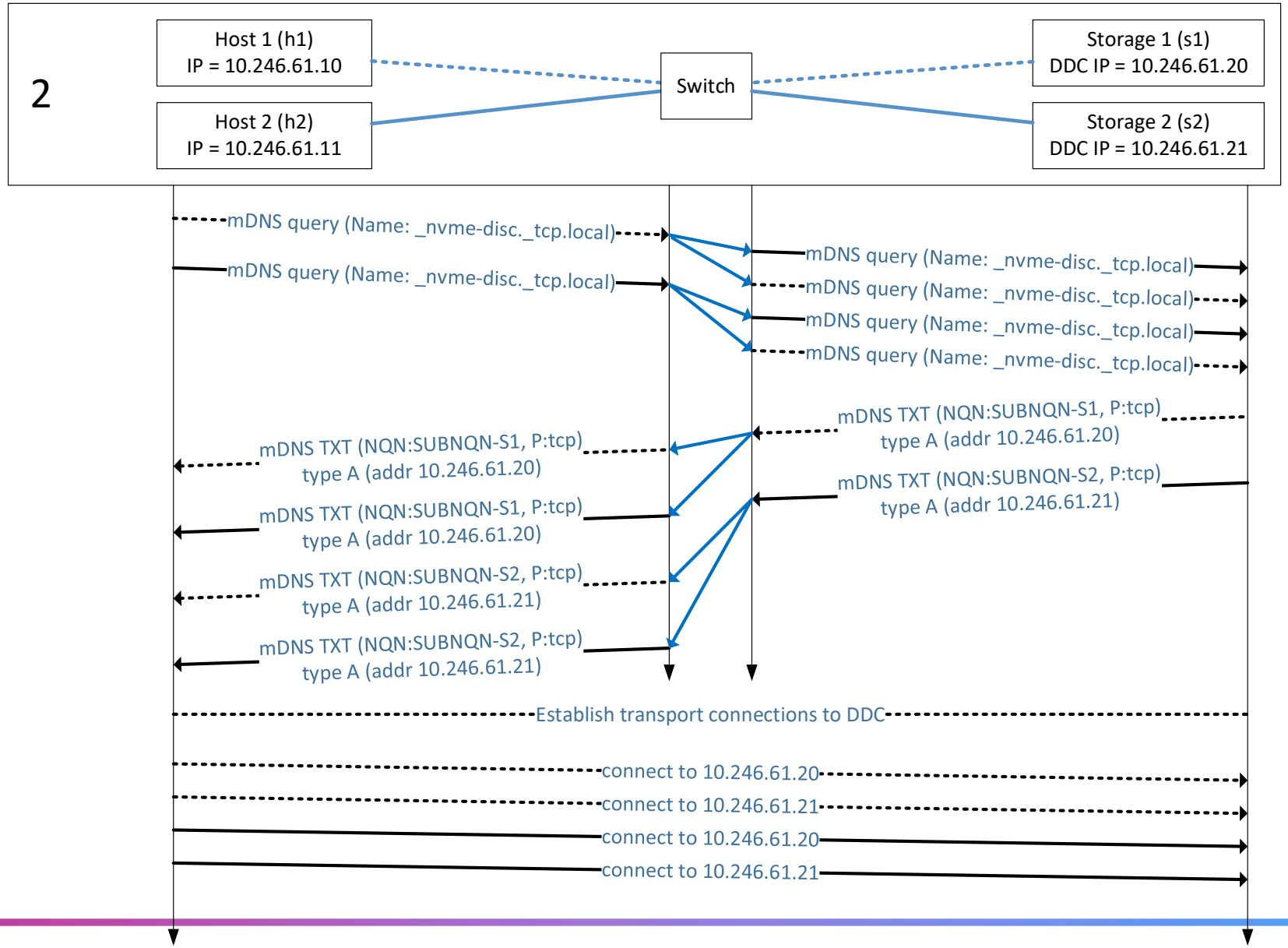
- Host (h1) uses mDNS to query for the "_nvme-disc" service
- Storage (s1) mDNS response includes DNS-SD records:
  - TXT contains the SUBNQN, as well as the protocols supported (e.g., tcp, roce)
  - "A" provides the IPv4 address of the DC on Storage (s1)

Direct Connect: Subsystem Comes Online after Host



- Storage (s1) comes online and transmits mDNS query to probe for the "_nvme-disc" service
- Storage (s1) mDNS announce includes DNS-SD records:
  - TXT contains the SUBNQN, as well as the protocols supported (e.g., tcp, roce)
  - "A" provides the IPv4 address of the DC on Storage (s1)

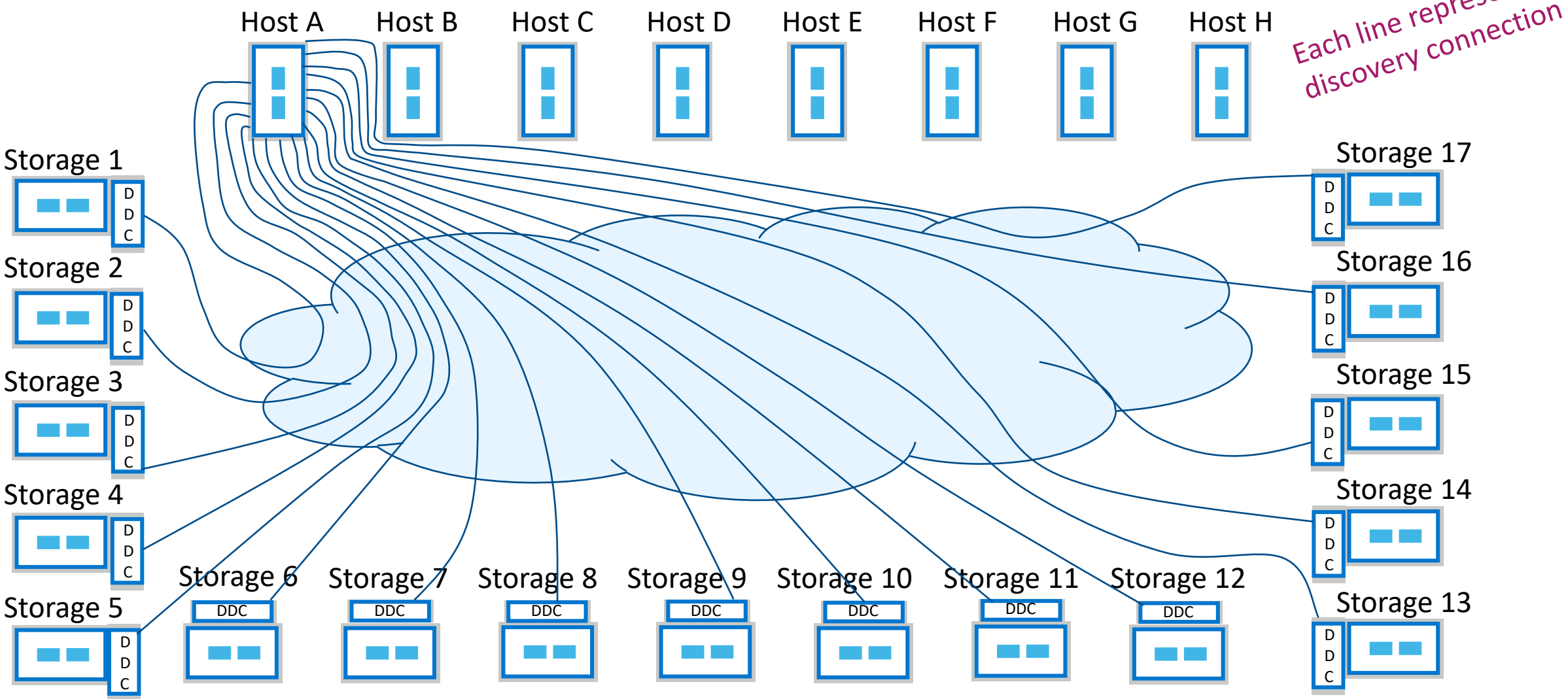
Multiple Hosts No CDC



Centralized Discovery

Curtis Ballard

The Scaling Problem



Each "Storage" represents a storage system presenting a Discovery Controller for one or more NVM Subsystems

What about Referrals

- Discovery Controllers today support a model called “Referrals”
- One Discovery Controller can point to “refer” to another Discovery Controller
- Simplifies Host manual configuration
  - only manually configure the “first” Discovery Controller

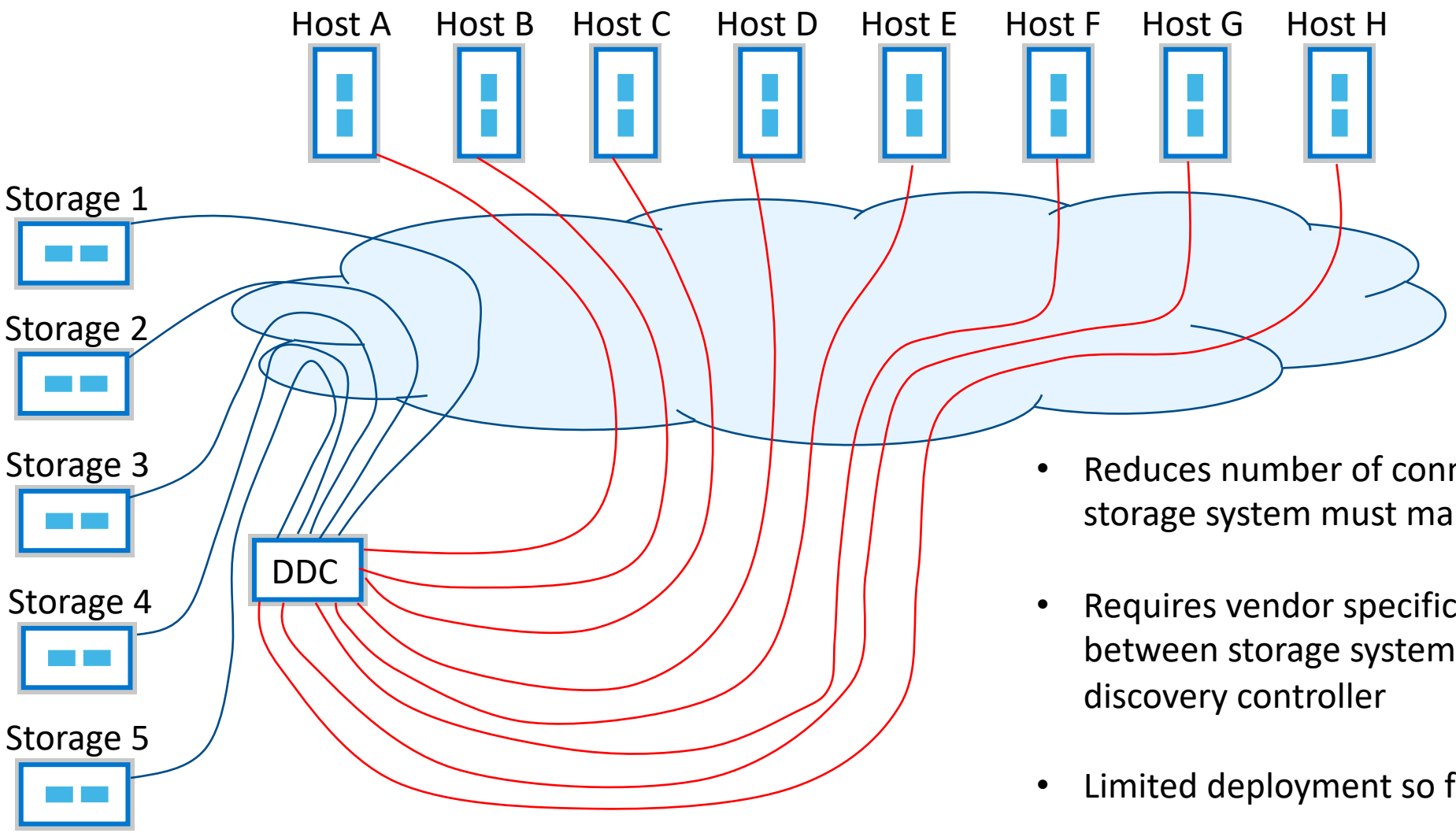
    – But –

- Referrals don’t scale well
  - Still require all of the Host to Discovery controller connections discussed in last slide
  - Can increase discovery time by reducing parallelism

Cooperating Storage Systems

- Storage systems could cooperate to exchange information with a common Discovery Controller
  - But –
- NVM Express 2.0, the latest specification released in June 2021, does not provide a method to enable storage system cooperation

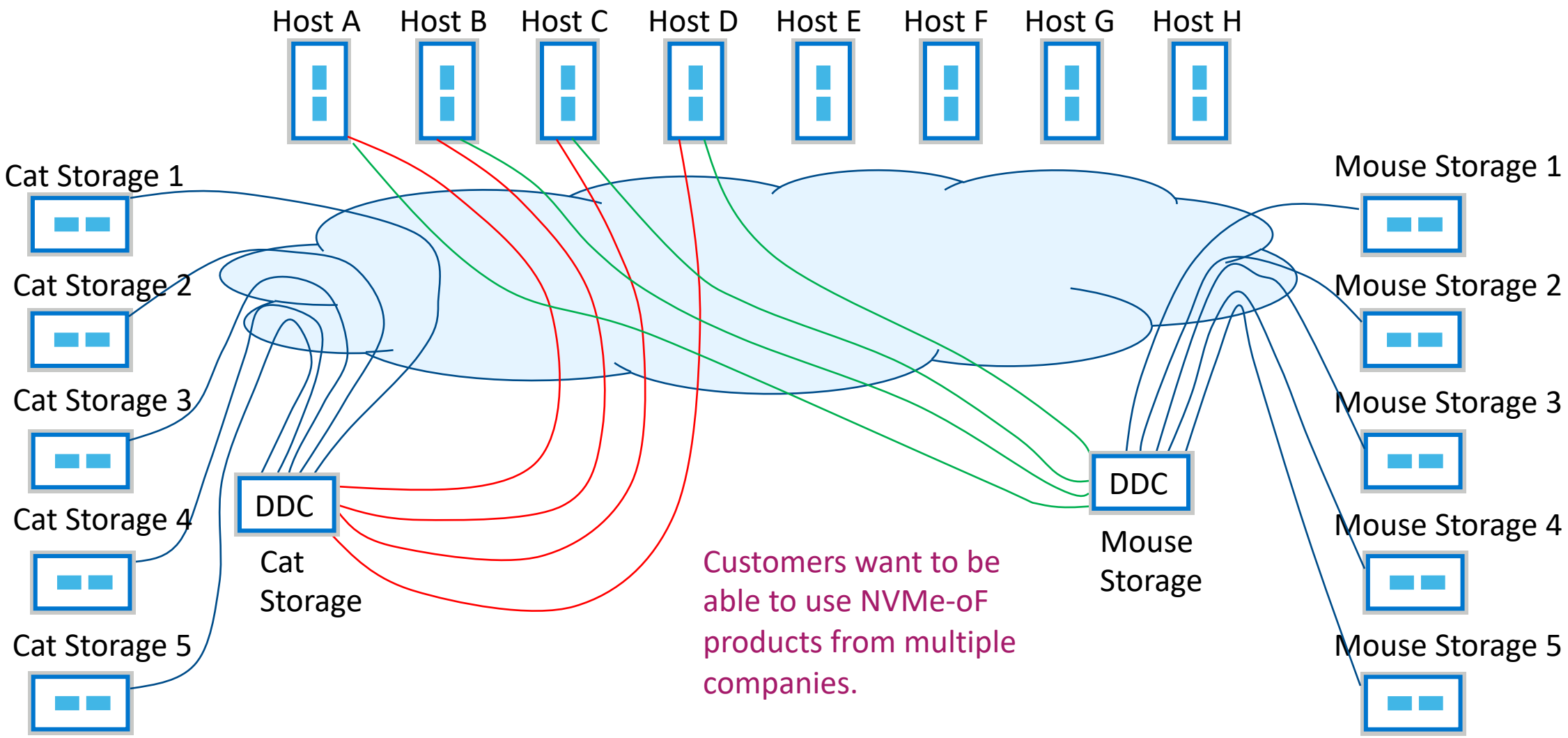
Cooperating Storage Systems



- Reduces number of connections each host and storage system must maintain
- Requires vendor specific information exchange between storage systems and shared direct discovery controller
- Limited deployment so far

Each “Storage” represents a cooperating storage system sharing a direct discovery controller

The Interoperability Problem



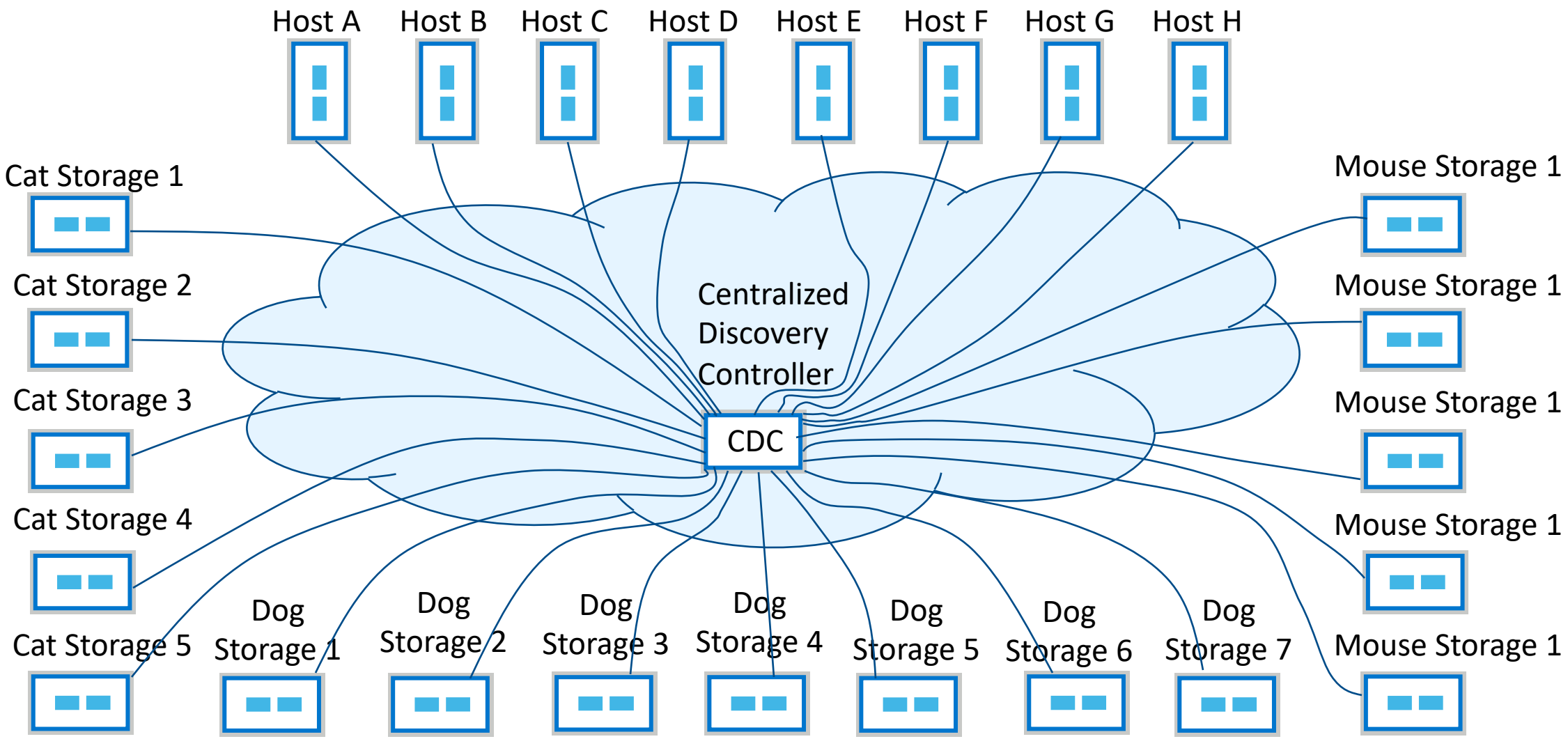
Each “Storage” represents a cooperating storage system sharing a PEER direct discovery controller

The Solution: Centralized Discovery Controllers

- NVM Express standard model for cooperating Discovery Controllers
  - Single Fabric entity that aggregates NVMe subsystem information
  - Standard API for sharing discovery information between a Centralized Discovery Controller (CDC), Hosts, and NVM Express storage system Discovery Controllers (Direct Discovery Controllers, DDCs)
  - Single location for storage systems to register discovery information
  - Single location for Hosts to query discovery information

- Additional new functionality for both CDCs and DDCs
  - Mechanism for Hosts to register Host information into discovery controllers
  - Mechanism for sharing connectivity rules, “Fabric Zoning”, information

Playing Nicely Together



Each “Storage” represents a cooperating storage system sharing a Centralized discovery controller

Host Discovery of Accessible Subsystems Model

- Clean evolution of existing host discovery
- CDC reports available NVM Subsystems
  - Same format discovery log pages as today
  - Same host specific accessible NVM Subsystems filtering as today is allowed
- Only completely new functionality is Host registering with the storage fabric

CDC Discovery of Accessible Subsystems Model

Two models defined

- Storage system sends registration information to the CDC: “Push Registration”

- or -

- CDC Pulls registration information from the storage system “Pull Registration”

- Why are there two models?
  - Storage system registering with CDC is the most direct but requires the storage system to implement some Host functionality in the Discovery Controller and send NVMe commands. Existing storage system Discovery Controllers do not have Host functionality and the CDC reading the discovery information is the best fit for their architecture.

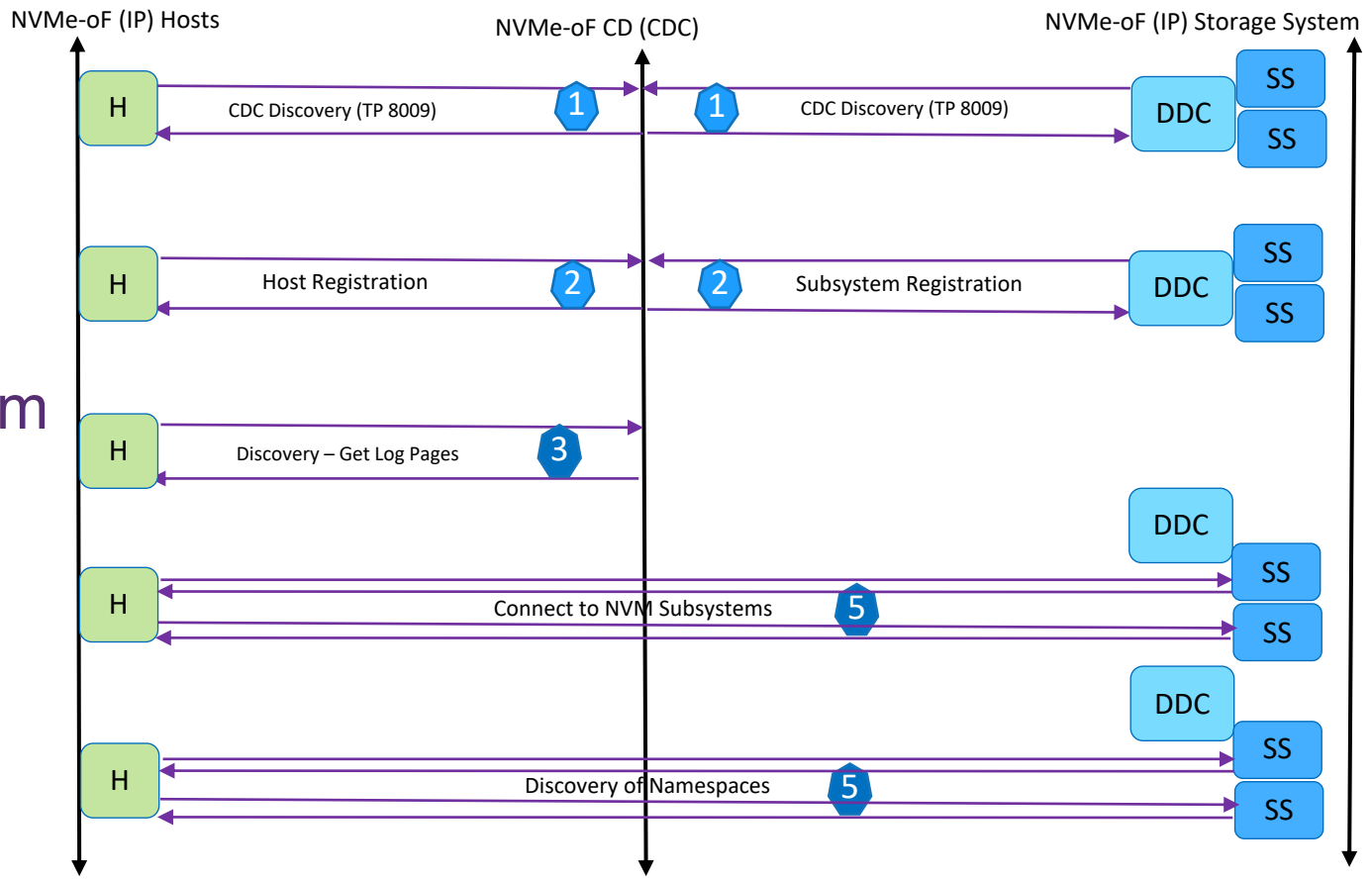
New: Registrations

- Push Registrations
  - Proactive registration from a Host or a Storage System into a CDC
  - Only registration model defined for Hosts
  - For Direct Discovery Controllers this requires host functionality and ability to send commands to CDC
  - NVMe-oF connections established with CDC
  - Hosts and CDCs use same registration commands with slightly different data formats

- Pull Registrations
  - Only for Direct Discovery Controllers in storage systems
  - Storage systems discover CDC and request pull registration
  - CDC uses existing Get Log Page commands to read existing Discovery Log Pages
  - DDC that supports CDC discovery reports information for all hosts and NVM subsystems to CDC

Example Simple Centralized Discovery Sequence

1. Hosts and storage systems discover the CDC
2. Hosts and storage systems register with the CDC
3. Hosts read discovery information from CDC (accessible NVM subsystems)
4. Hosts connect to NVM Subsystems
5. Hosts discover namespaces
6. Go!



How does the CDC Filter Responses by Host?

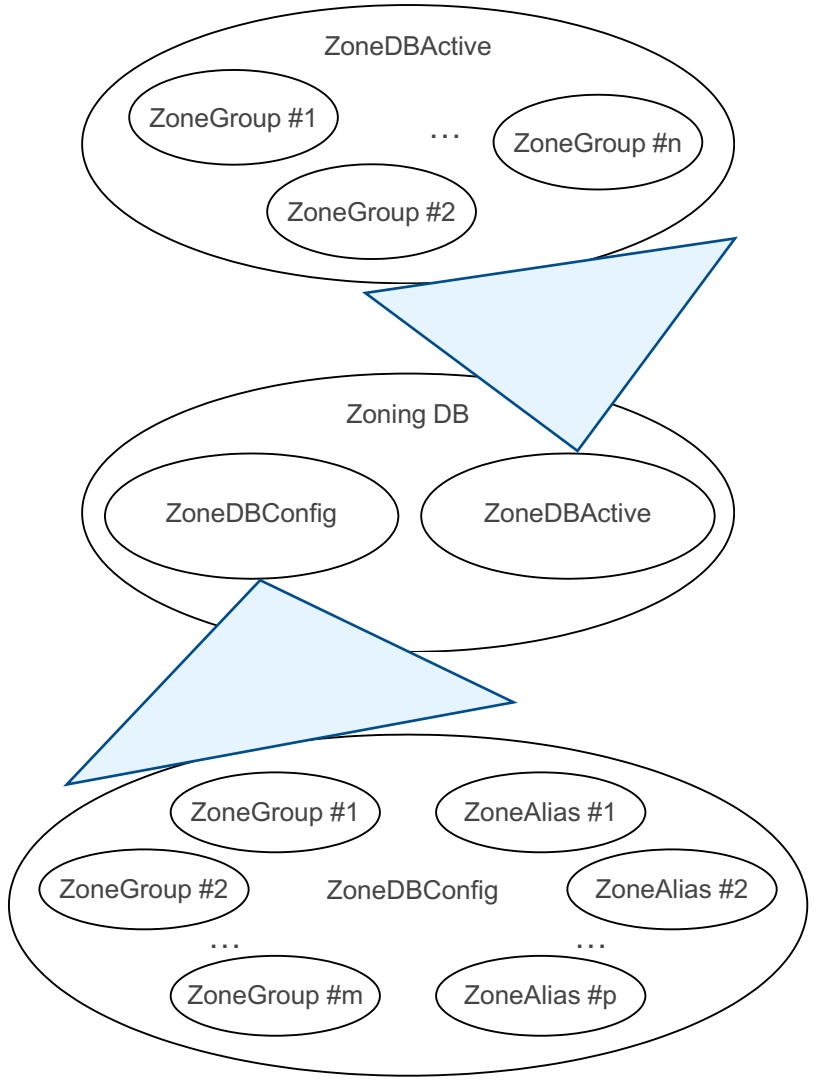
- Today's storage systems often implement access controls and the Discovery Controllers only report information about "accessible NVM Subsystems"
- The CDC has to get the full list of all "available" NVM Subsystems
- How does the CDC know which NVM Subsystems are "accessible" by which Hosts?

- Answer: Fabric Zoning

Fabric Zoning Quick Intro

- Zoning database in CDC stores configured and active zones
  - Configured zones is list of
    - All Fabric ZonesGroups; and
    - All Fabric ZoneAliases (a related set of Zone members)
  - Active Zones is list of ZoneGroups that are being enforced
- ZoneGroups contain Zone members
  - Hosts, NVM Subsystems, ZoneAliases
  - Member identification NQN, NQN/IP tuple, NQN/PortID tuple, etc.
- Admin commands defined for CDC and NVM subsystem to share Fabric Zone information

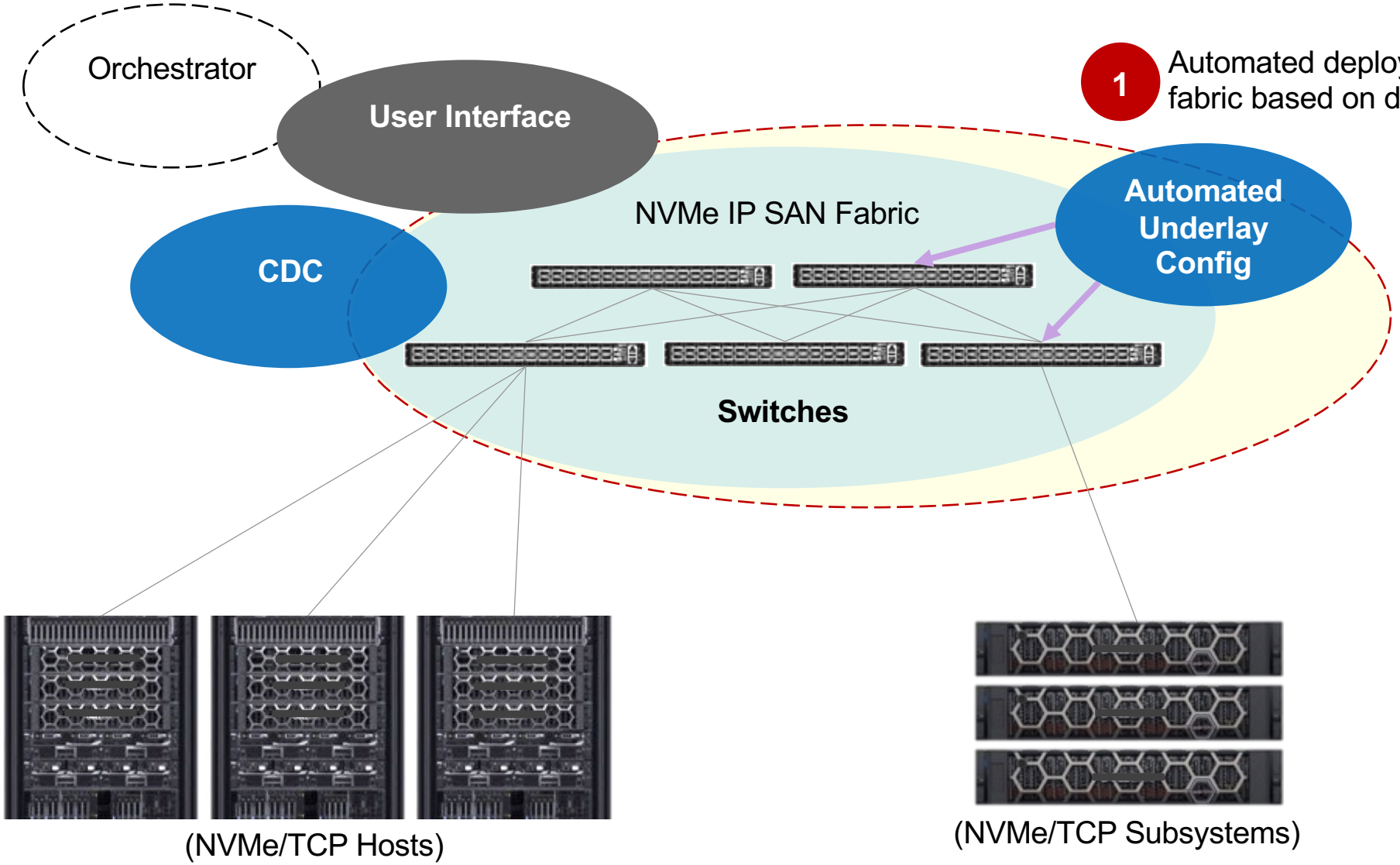
- Multiple active ZoneGroups allowed



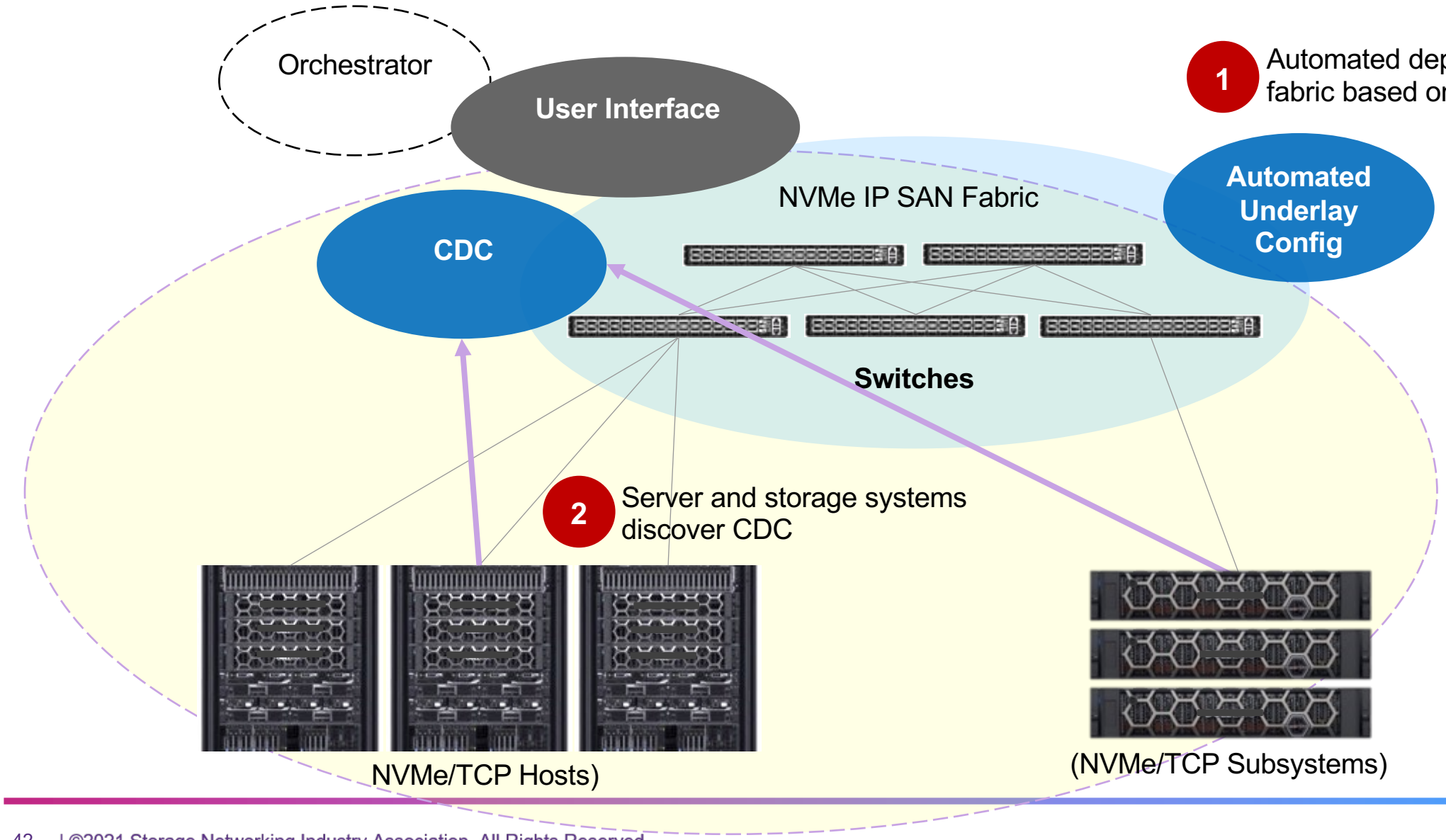
End-to-End Automated Discovery Example

Erik Smith

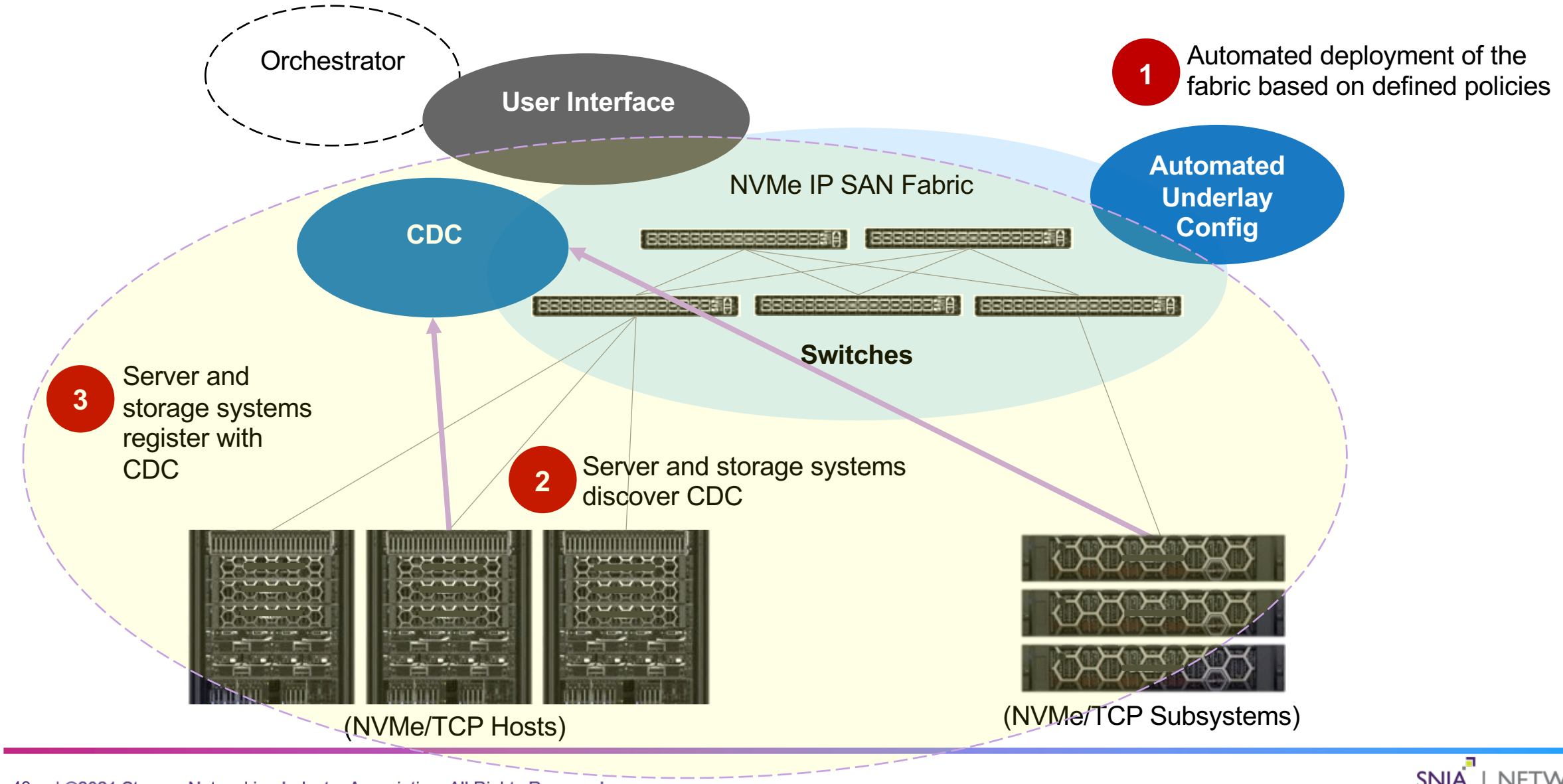
NVMe IP-Based SAN Operations



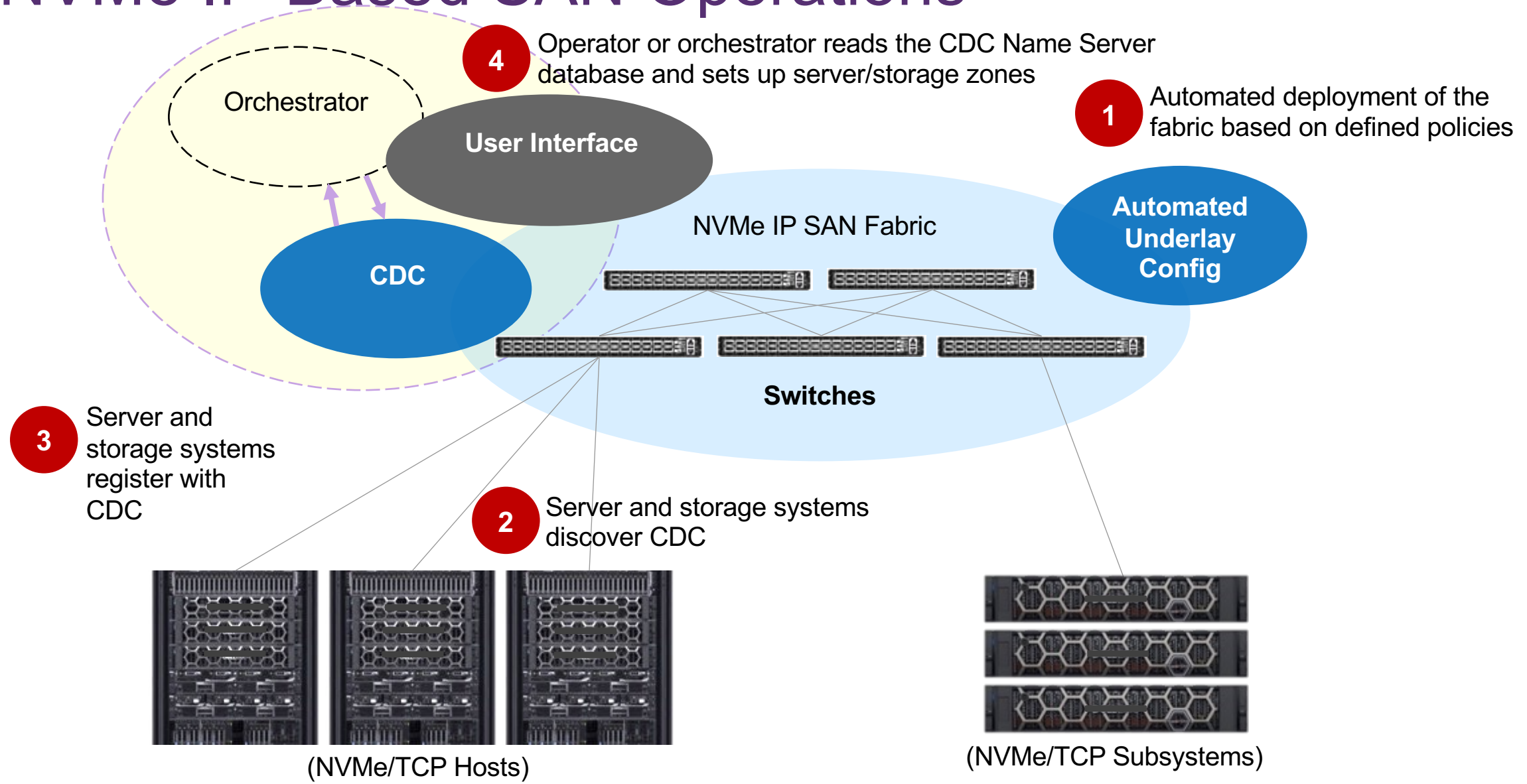
NVMe IP-Based SAN Operations



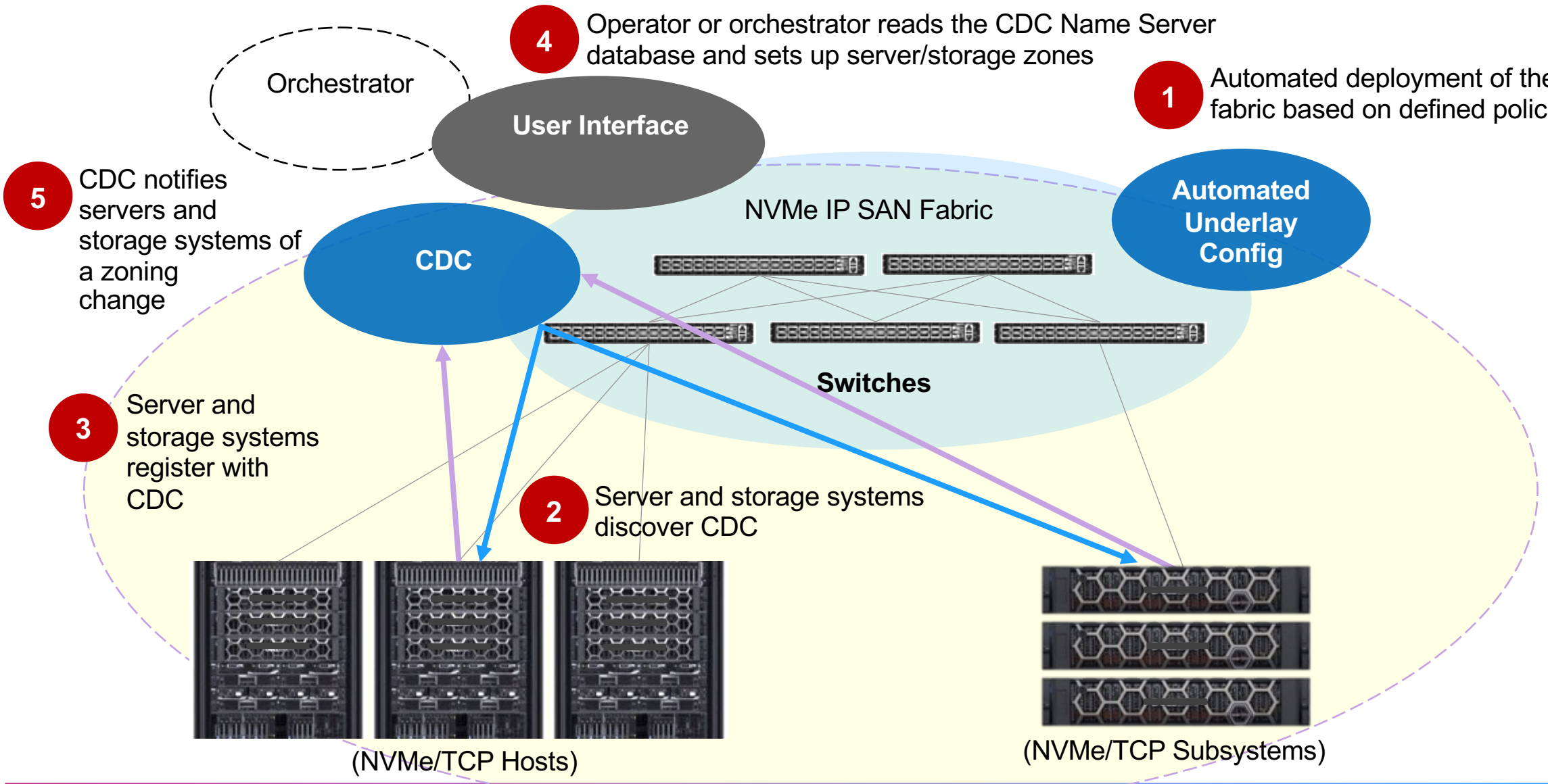
NVMe IP-Based SAN Operations



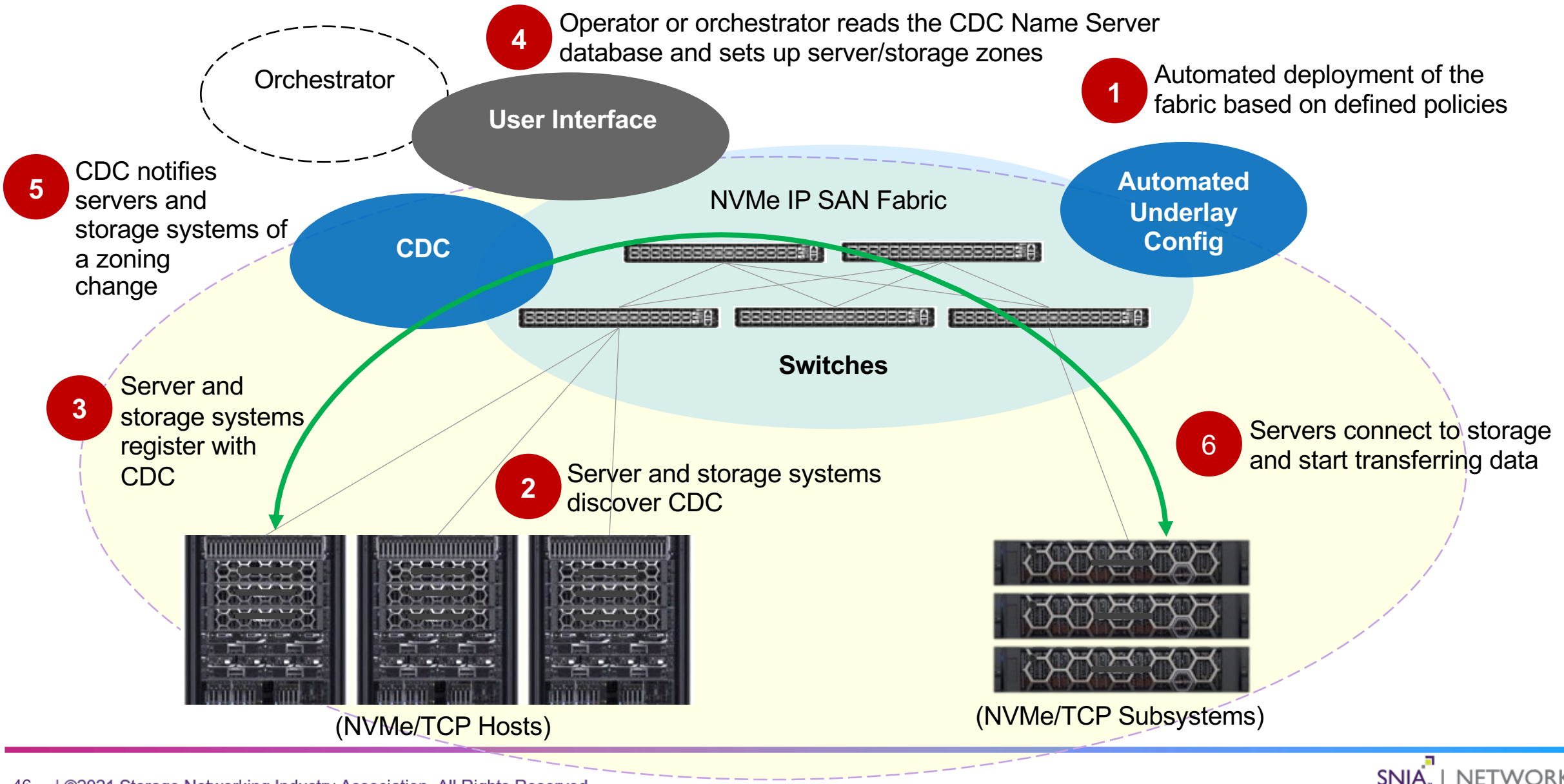
NVMe IP-Based SAN Operations



NVMe IP-Based SAN Operations



NVMe IP-Based SAN Operations



Key Takeaways

- Discovery Automation does not entirely depend upon the presence of a Centralized Discovery Controller (CDC).
  - Smaller scale environments can make use of mDNS (as described in TP-8009) to automatically discover NVMe Discovery Controllers.
- CDCs and subsystems that will support interacting with them should
  - Use Port-Local Log pages – Provides a much better UX and prevents leaking information between tenants.
  - Make use of Subsystem Driven Zoning (SDZ) – Storage admins only need to interact with one UI for storage provisioning.
  - Make use of extended attributes and register symbolic names that are meaningful to end-users.
  - Contribute to the open-source NVMe-oF Discovery client “nvme-stas” being led by Dell. Available for review after 8009 and 8010 are ratified (~end of the year).

After this Webcast

- Please rate this webcast and provide us with your feedback
- This webcast and a copy of the slides will be available at the SNIA Educational Library https://www.snia.org/educational-library
- A Q&A from this webcast, including answers to questions we couldn't get to today, will be posted on our blog at https://sniansfblog.org/
- Follow us on Twitter @SNIANSF

Thank You