STORAGE DEVELOPER CONFERENCE
SDC 21
BY Developers FOR Developers
Virtual Conference
September 28-29, 2021
A SNIA Event

Automating the discovery of NVMe-oF subsystems over an IP Network

Erik Smith
Distinguished Member of Technical Staff
Dell Technologies CTIO group

Agenda

- NVMe-oF's discovery problem
- Network topologies that support automated Discovery
- The differences between a FC SAN and an NVMe IP SAN used to transport NVMe/TCP
- An in-depth explanation of the discovery protocol

How did we get here?

NVMe-oF's IP based Discovery Problem is well-documented and was even 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**

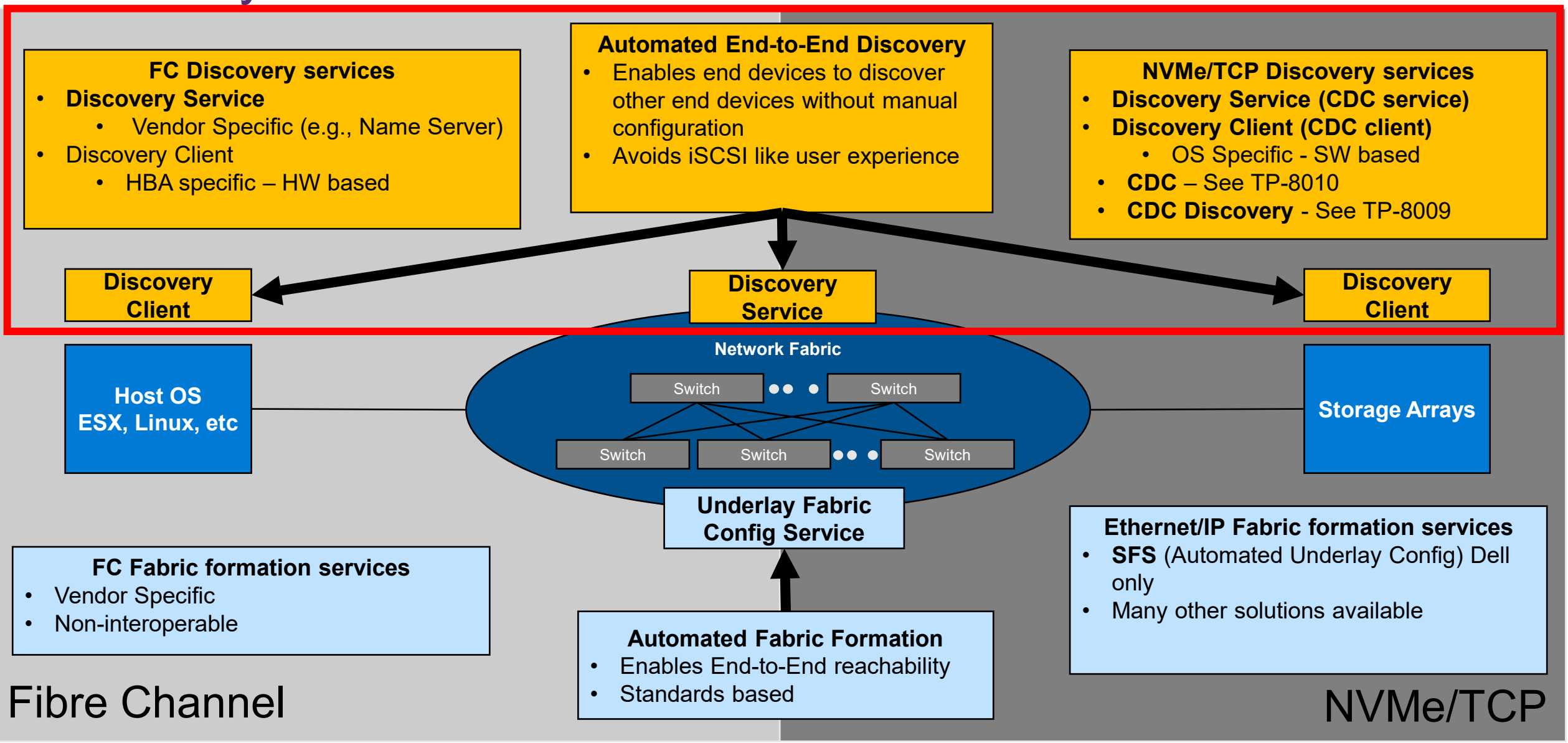
The problem? The methods described above all limit the scale and interop of any IP based NVMe-oF solution.

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/TCP

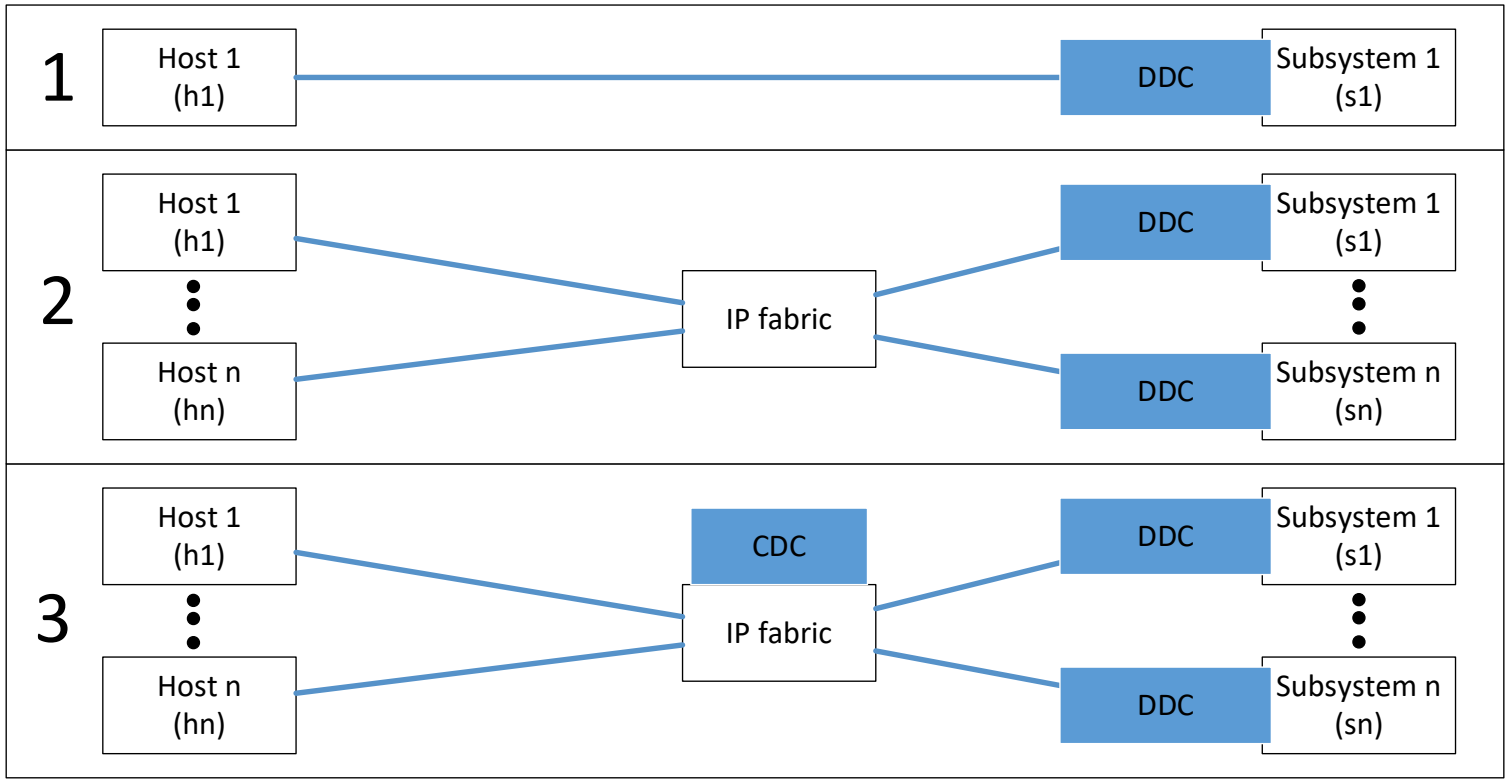


Deployment types that support Automated Discovery

1. Direct Connect

2. Multiple Hosts and subsystems without a CDC in the network

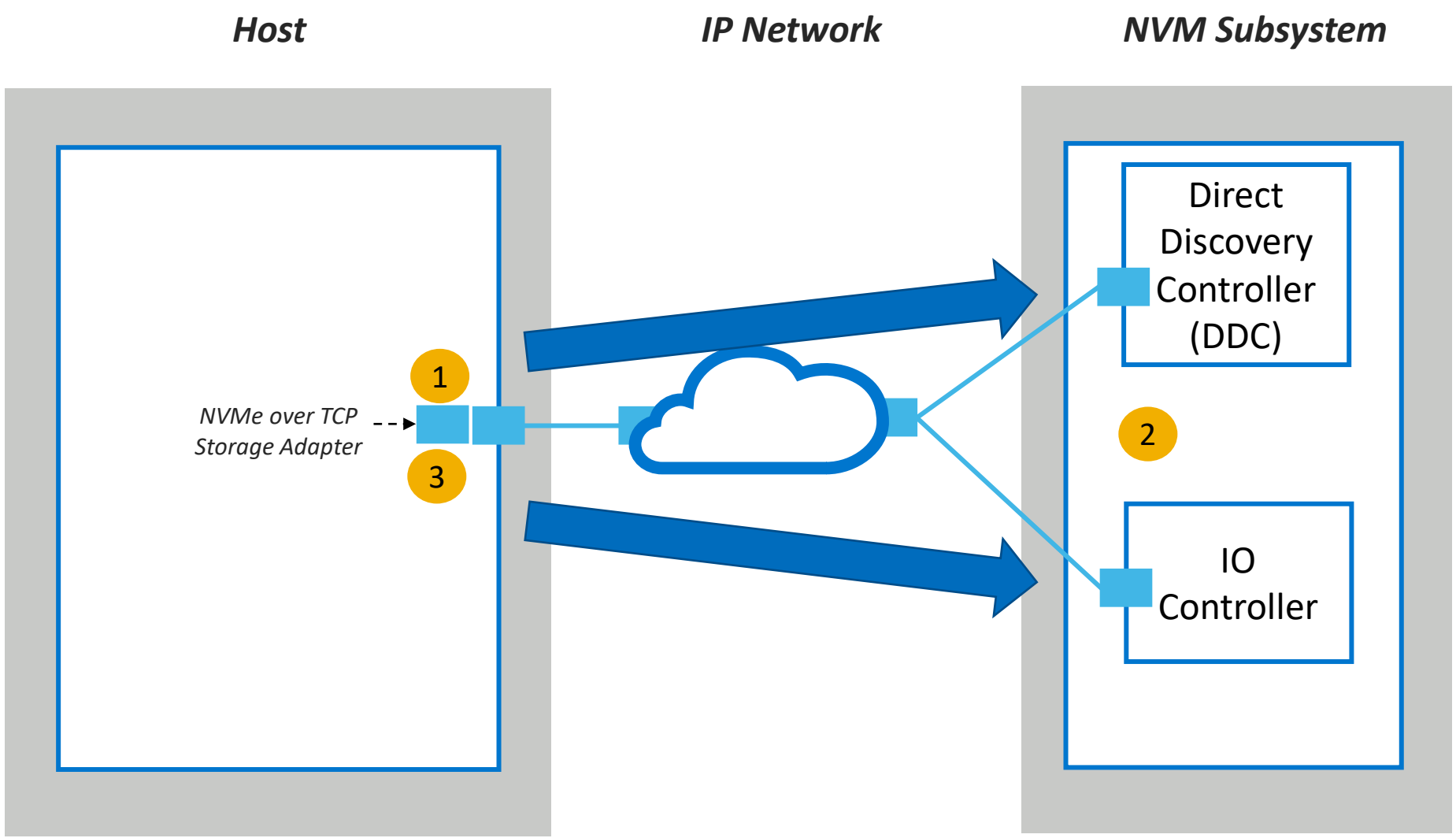
3. Multiple Hosts and subsystems with a CDC in the network



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.

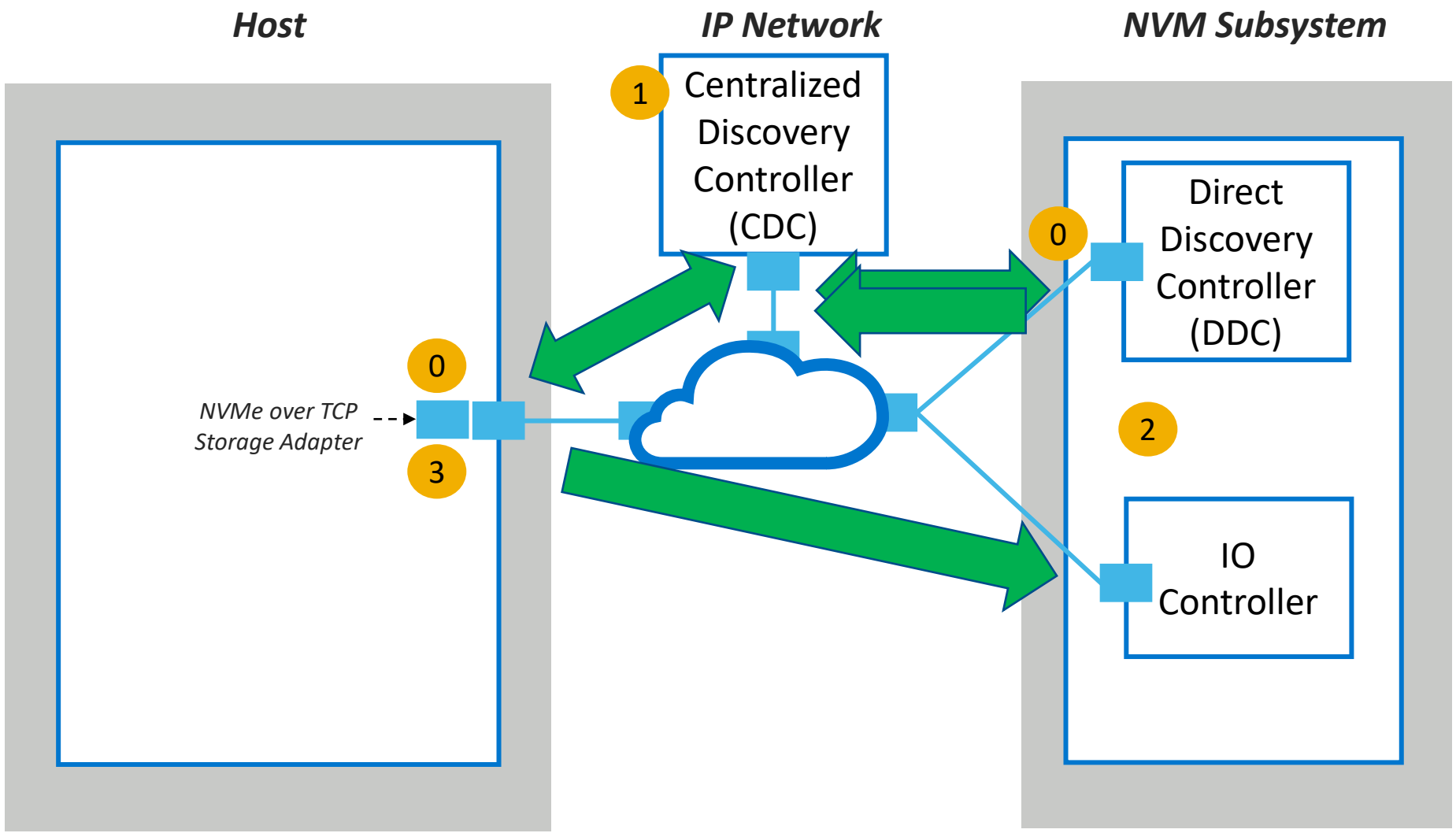
DDC (Direct Discovery Controller) – A Discovery controller that is not a CDC. Typically associated with a storage subsystem

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

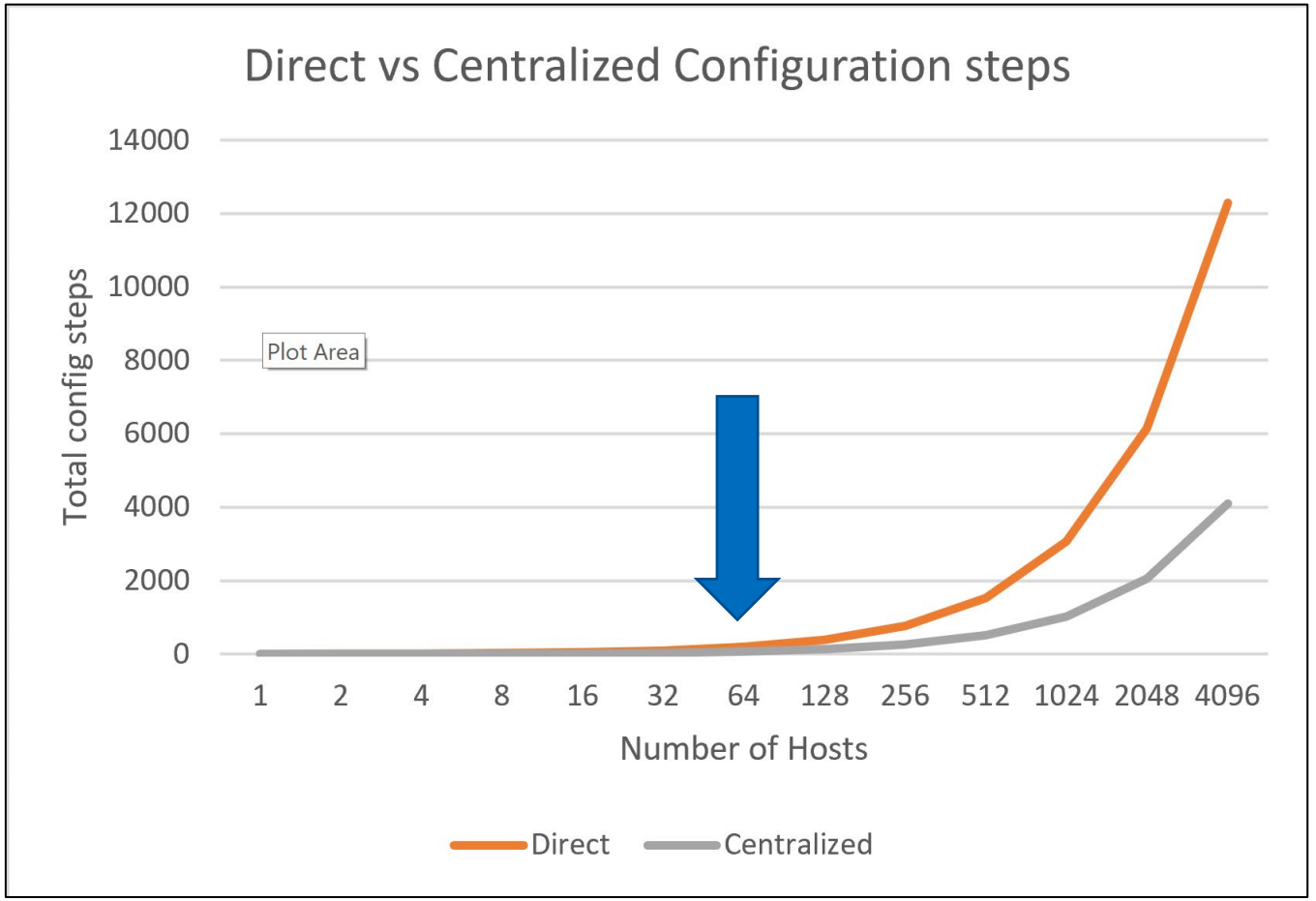
Direct vs Centralized Discovery at scale

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

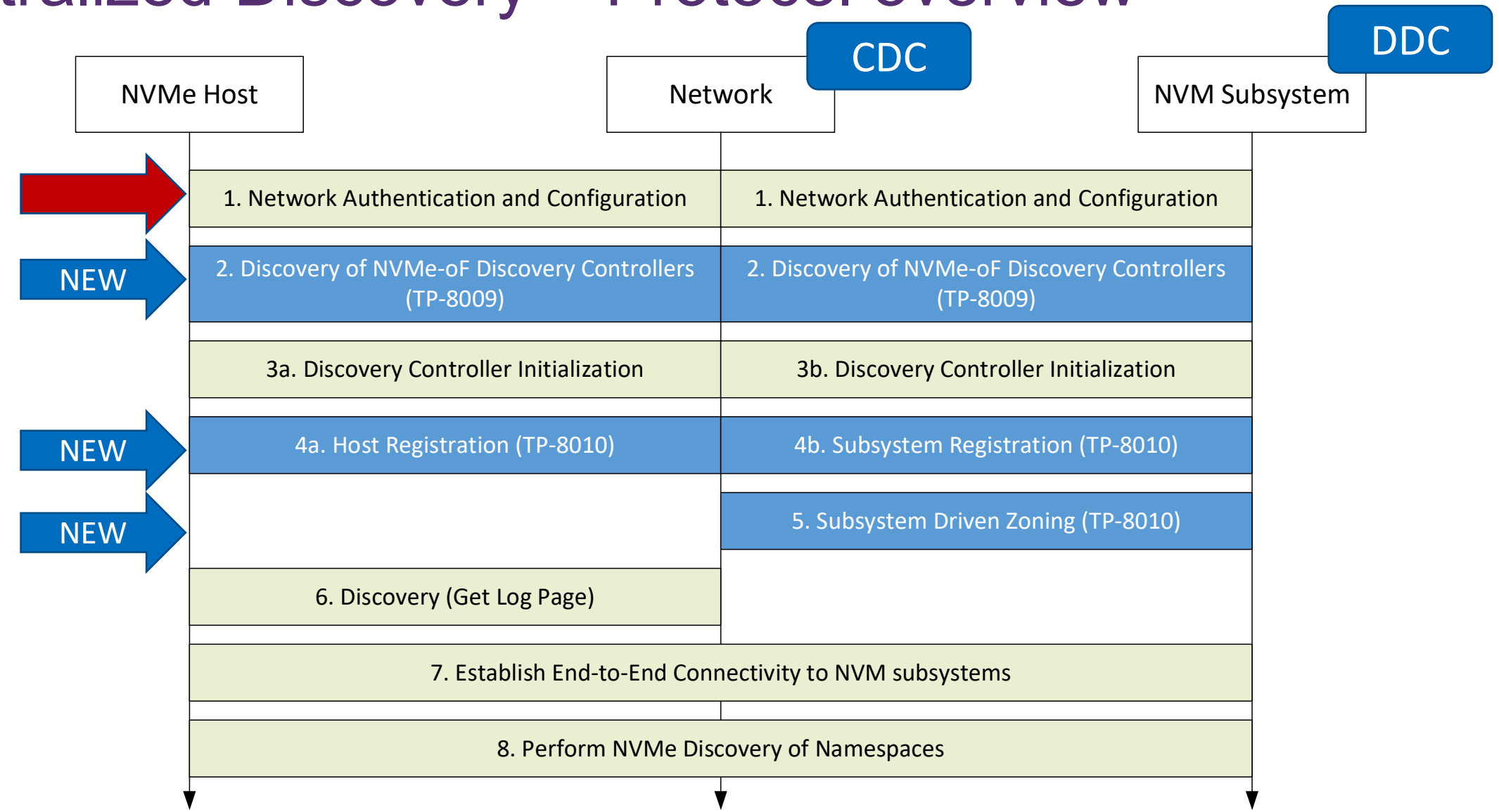
1. **Direct becomes impractical @ >64 hosts**
2. **Direct requires interaction with each host every time a storage subsystem is added or removed.**
3. **Direct may lead to extended discovery time if many subsystem interfaces are present.**

Additional points about Discovery Automation

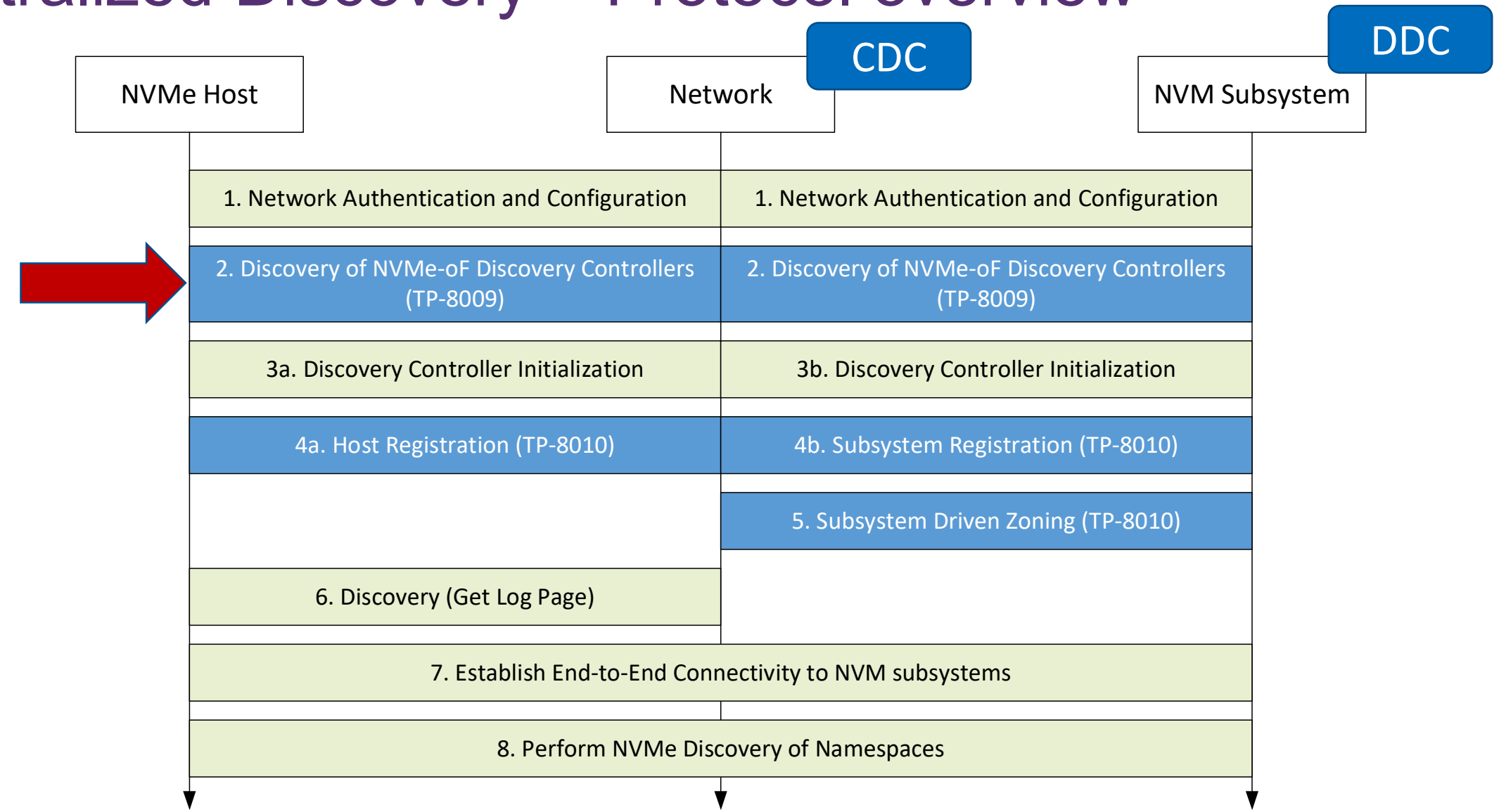
- Discovery Automation does not depend entirely upon a Centralized Discovery Controller (CDC).
- Smaller scale environments can make use of mDNS (as described in TP-8009) to automatically discover NVMe Discovery Controllers.
- This approach does not allow for Centralized Control, and this means:
  - 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

Discovery Protocol Overview

Centralized Discovery – Protocol overview



Centralized Discovery – Protocol overview

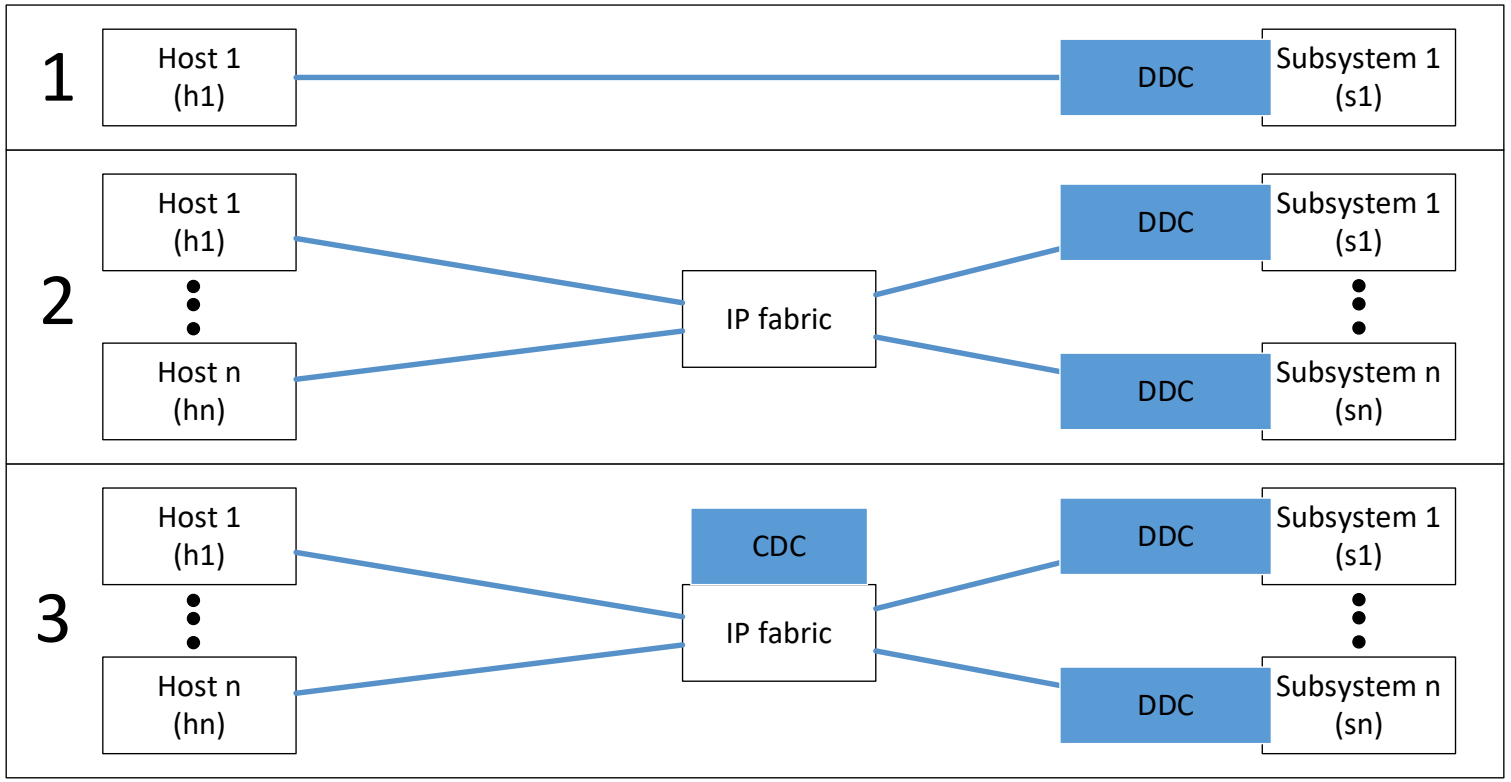


Deployment types that support Automated Discovery

1. Direct Connect

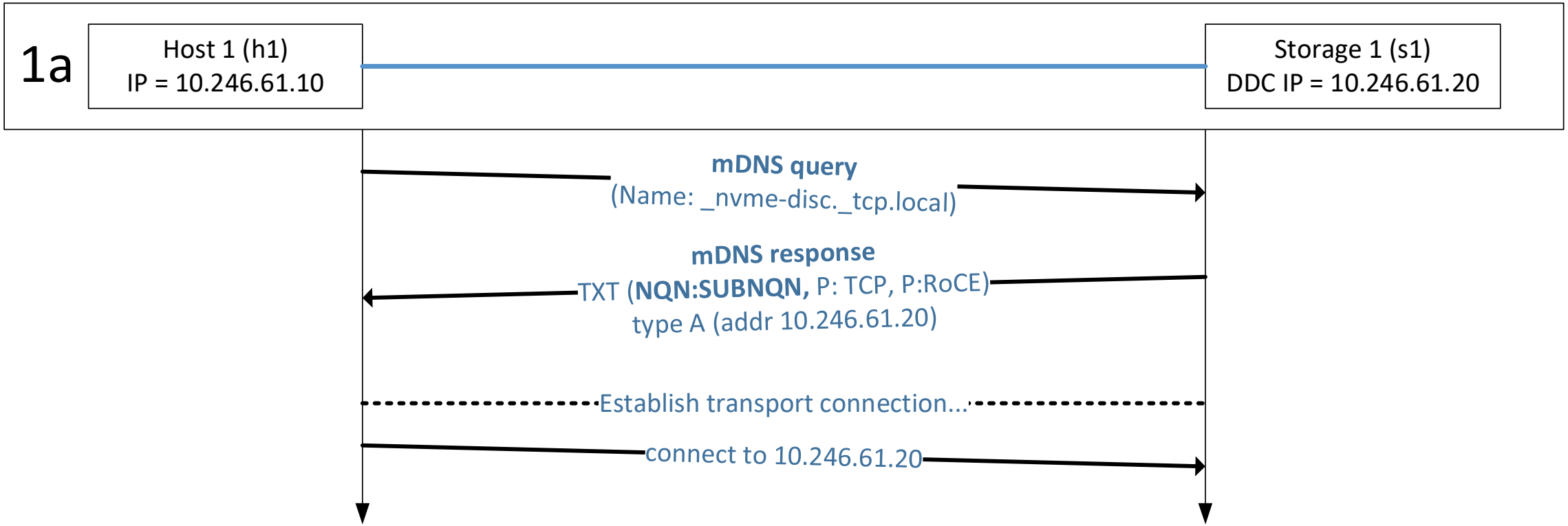
2. Multiple Hosts and subsystems without a CDC in the network

3. Multiple Hosts and subsystems with a CDC in the network



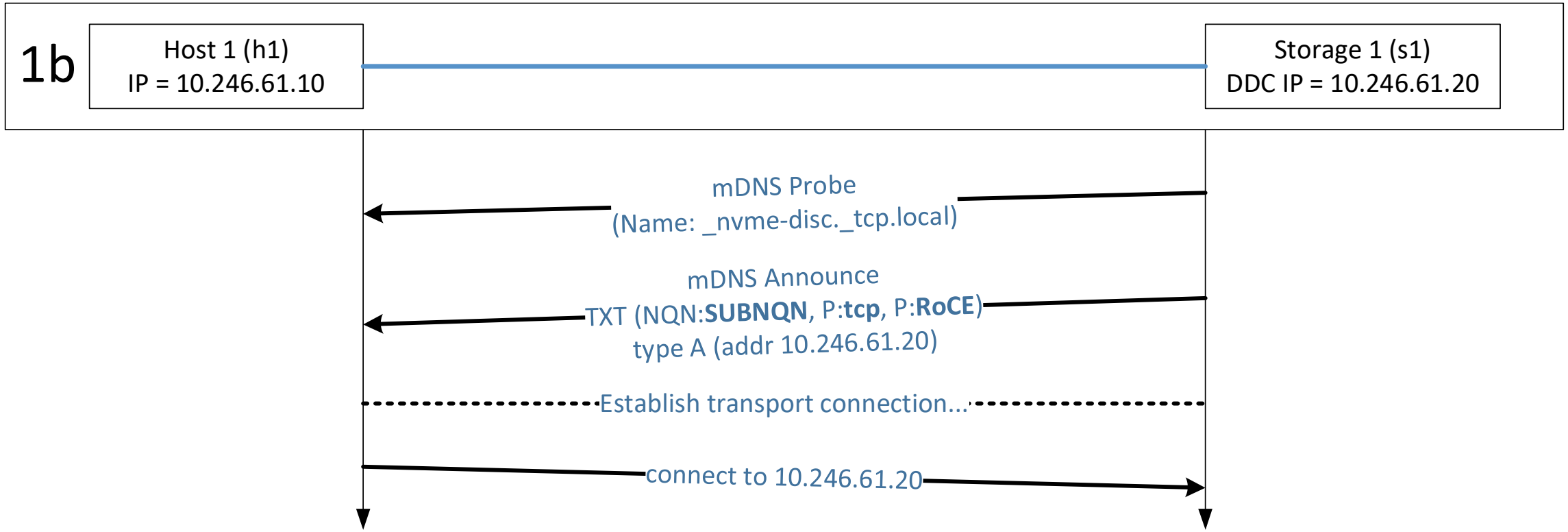
CDC (Centralized Discovery Controller) – A Discovery controller that supports registration and zoning. Typically runs stand-alone (as a VM) or embedded on a switch in the fabric.
DDC (Direct Discovery Controller) – A Discovery controller that is not a CDC. Typically associated with a storage subsystem

Direct Connect: A New Host comes online



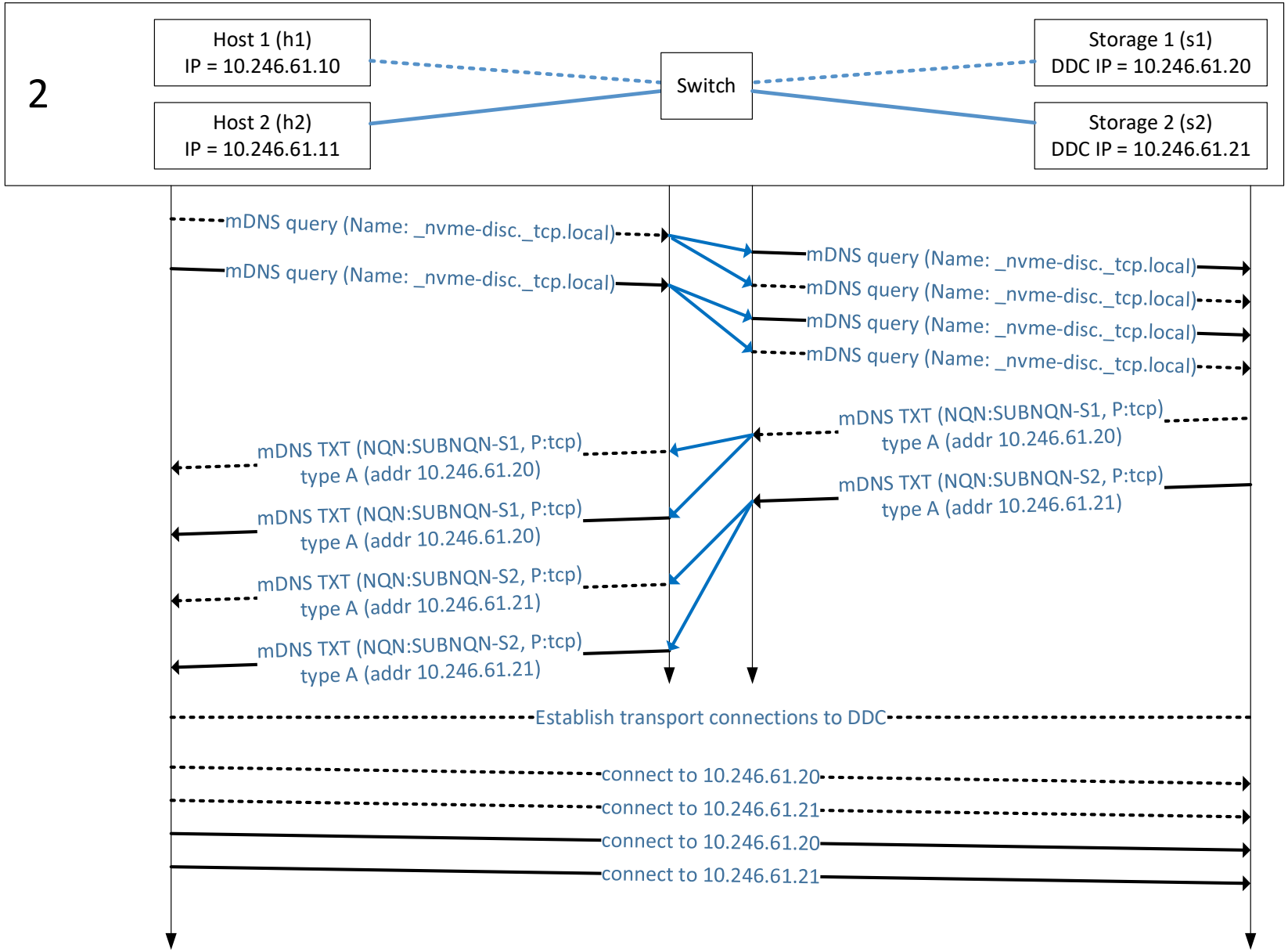
- Host (h1) uses mDNS to query for the "_nvme-disc" service
- Storage (s1) mDNS response includes DNS-SD records:
  - TXT contains the SUBNQN, as well as the protocols supported (e.g., tcp, roce)
  - "A" provides the IPv4 address of the DC on Storage (s1)

Direct Connect: Subsystem comes online after host

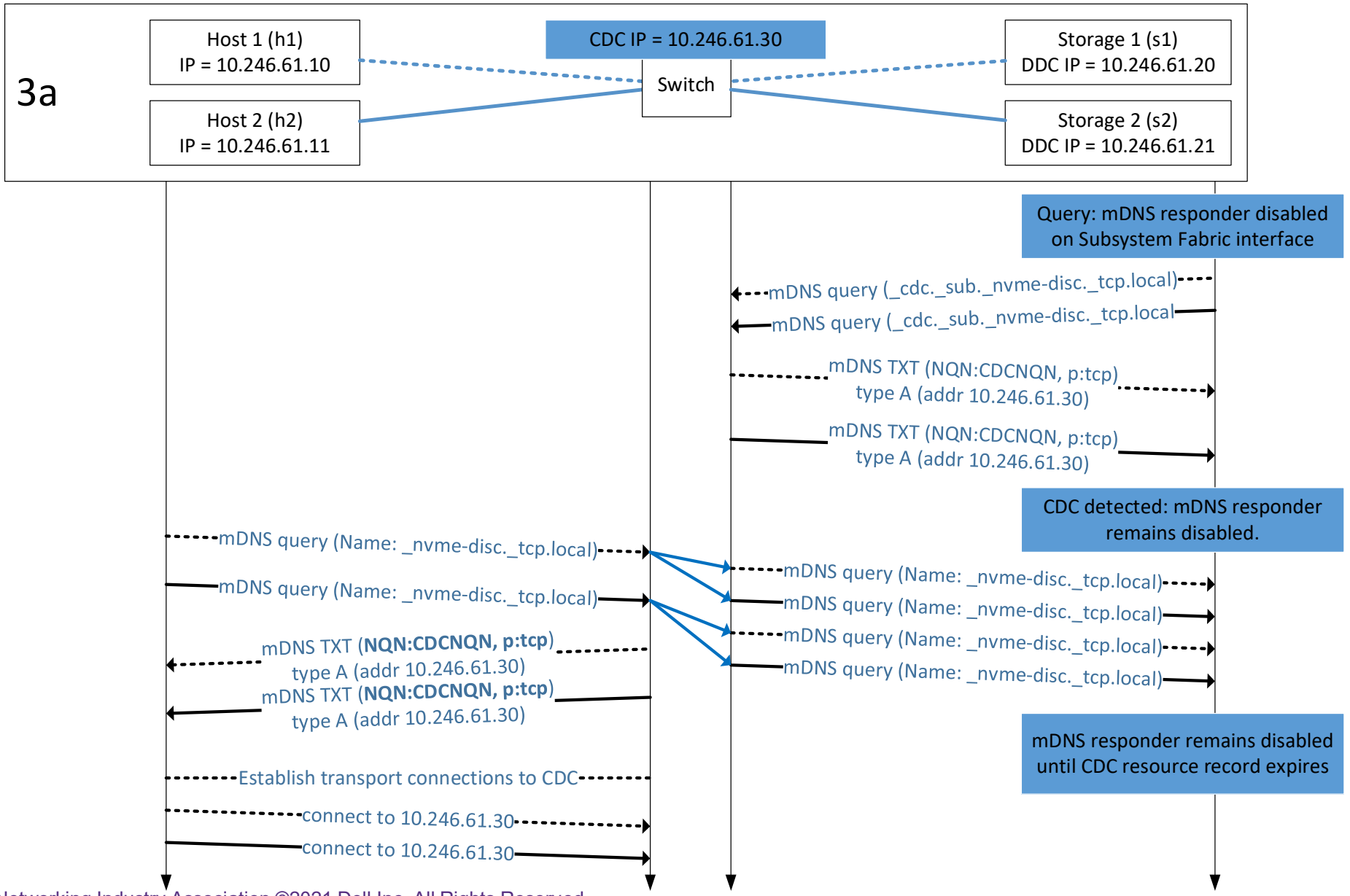


- Storage (s1) comes online and transmits mDNS query to probe for the "_nvme-disc" service
- Storage (s1) mDNS announce includes DNS-SD records:
  - TXT contains the SUBNQN, as well as the protocols supported (e.g., tcp, roce)
  - "A" provides the IPv4 address of the DC on Storage (s1)

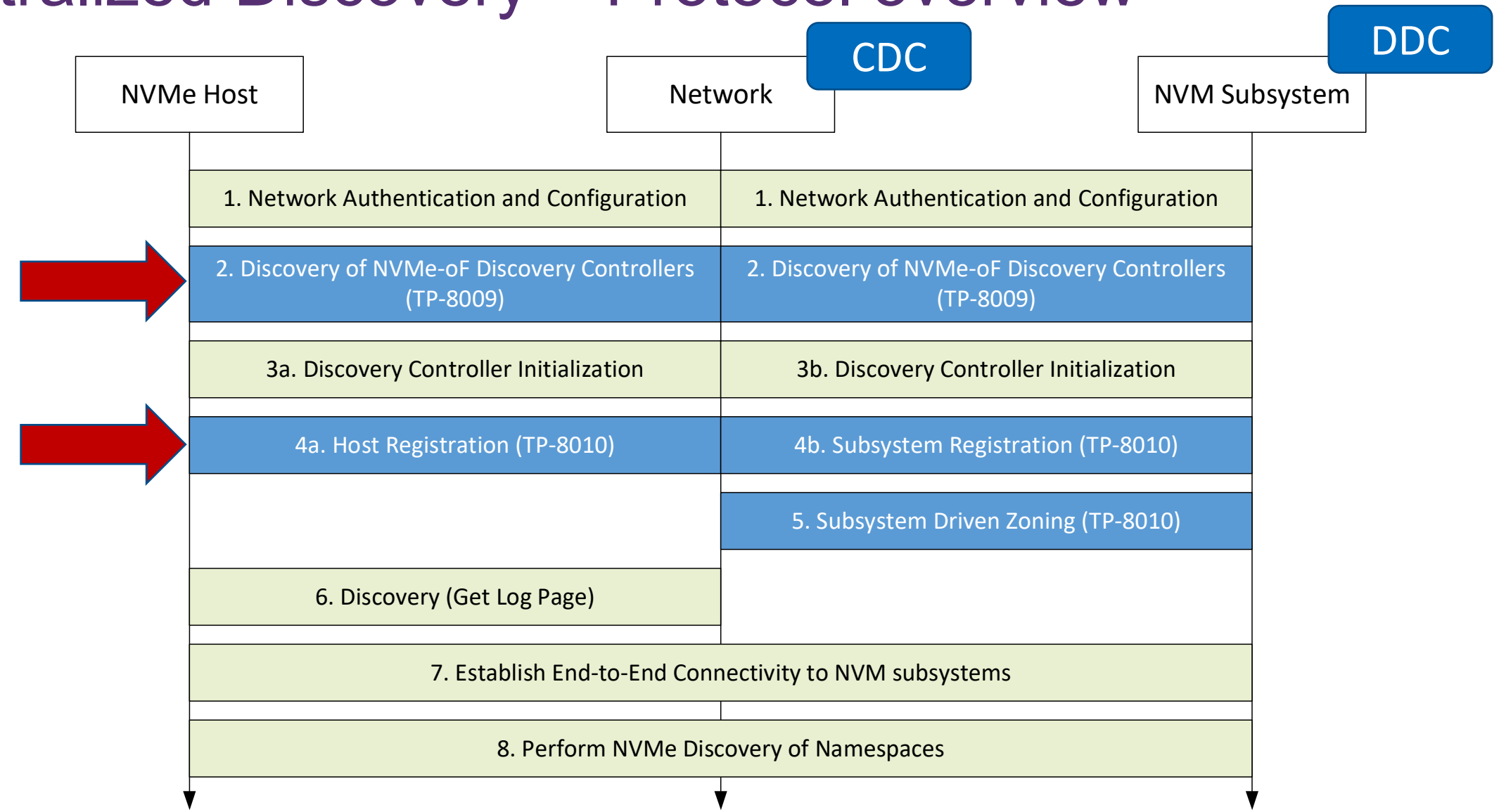
Multiple Hosts no CDC



Multiple Hosts with CDC



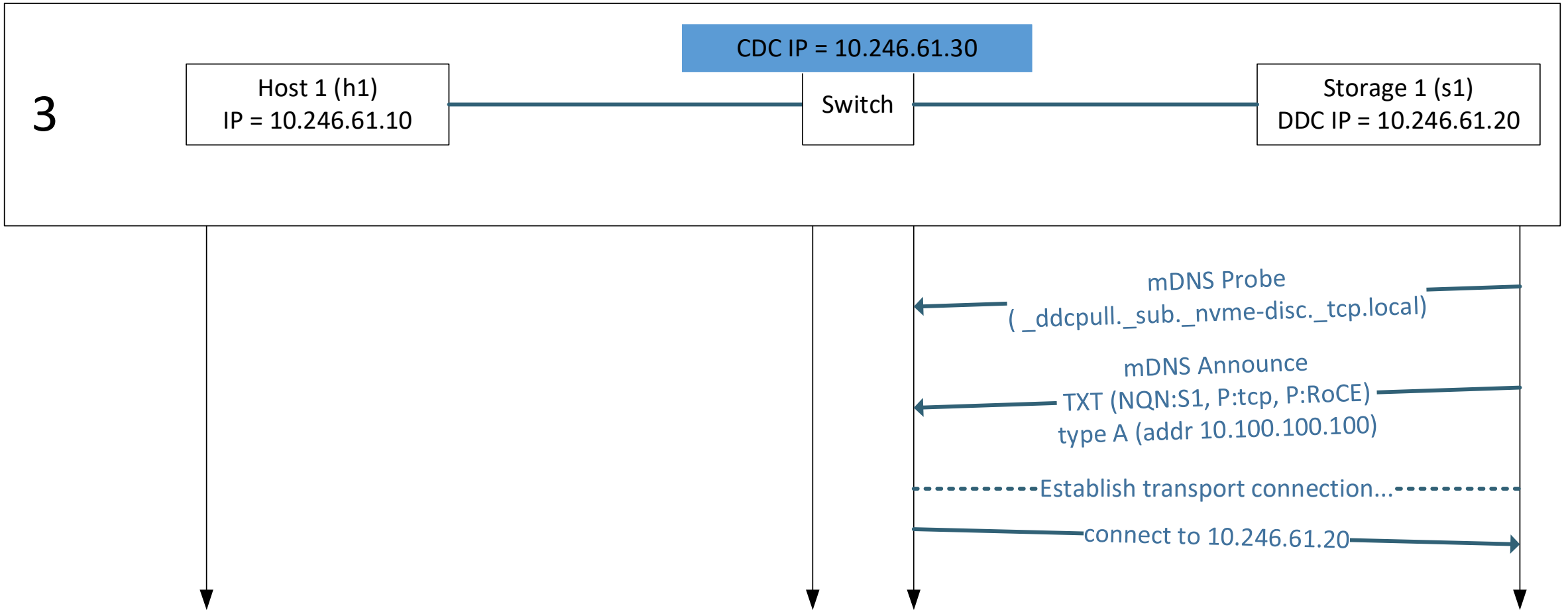
Centralized Discovery – Protocol overview



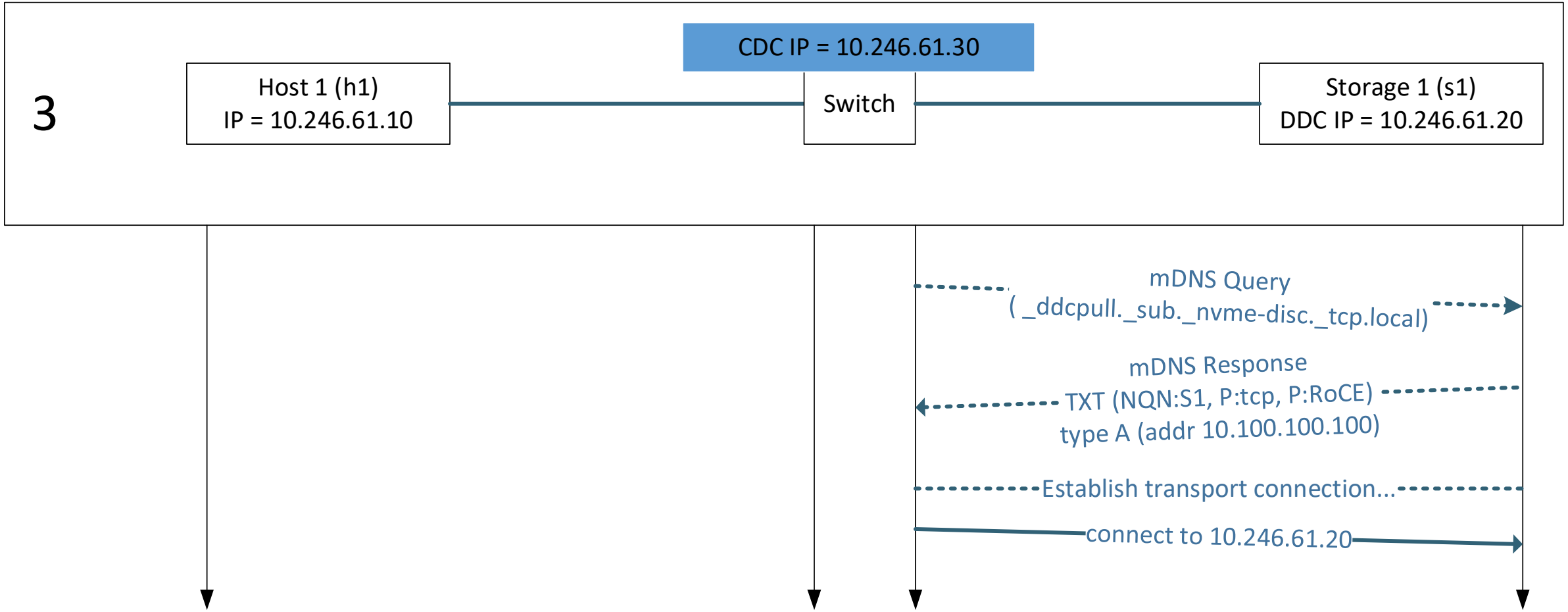
Discovery Information registration techniques

- We've defined two registration techniques; Push and Pull.
- Push registration uses the same basic approach as FC
  - Each Host or storage interface sends a registration command to the fabric services.
  - Fabric services store the registration information in a database (e.g., name server)
- Pull Registration is a new concept
  - Only allowed to be used by subsystem interfaces
  - Each subsystem interface informs the CDC that it has information that it would like to register
  - The subsystem does this by either using mDNS or by sending a "Kickstart Discovery Request" to the CDC.
  - The CDC will then Connect to the subsystem interface and retrieve the Discovery Log Page.
  - The Discovery Log Page entries are then added to the CDCs name server database.
  - Originally intended to allow legacy implementations to take advantage of Centralized Discovery
  - Some new subsystem implementations have chosen to use it because it is simpler for them to implement.

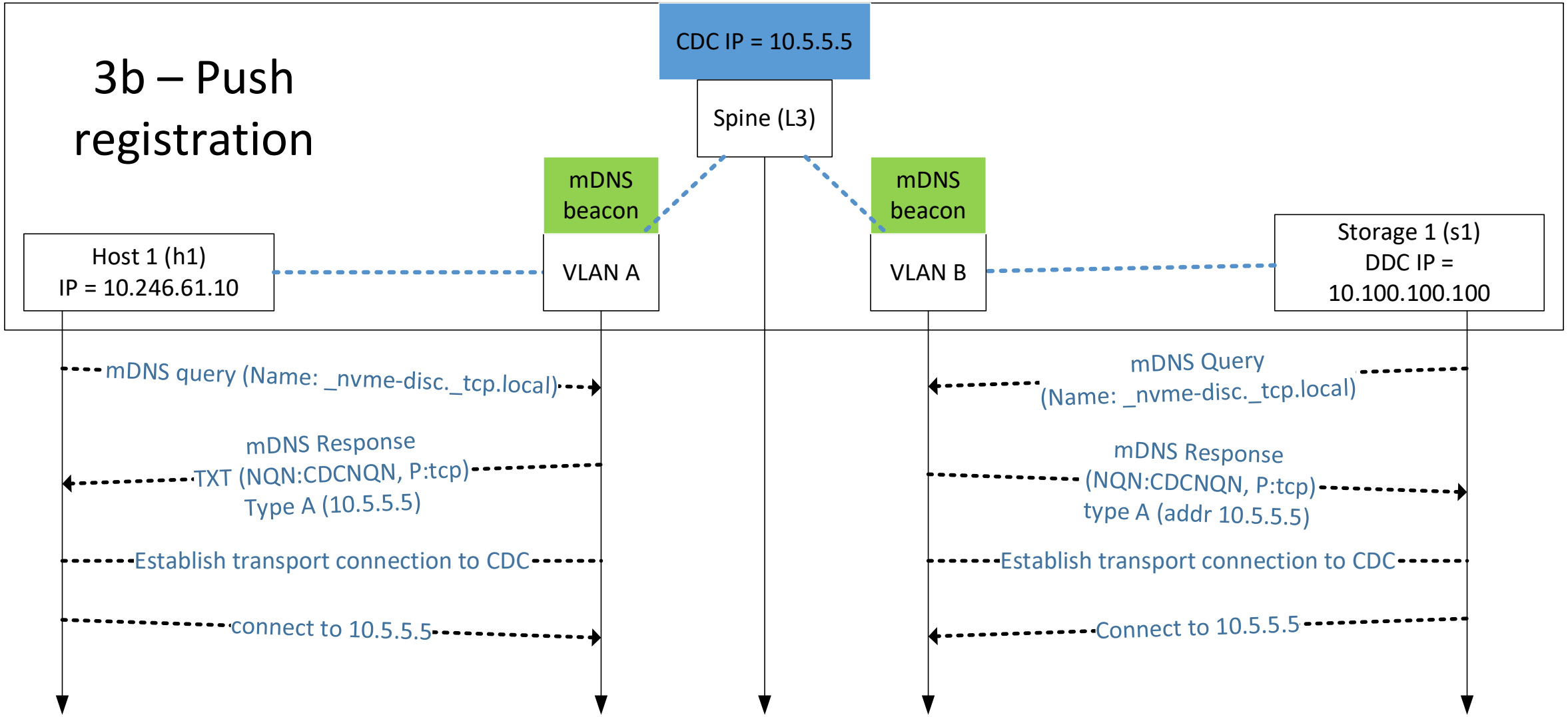
Single broadcast domain with CDC
DDC interface uses mDNS to announce that it requires Pull registration



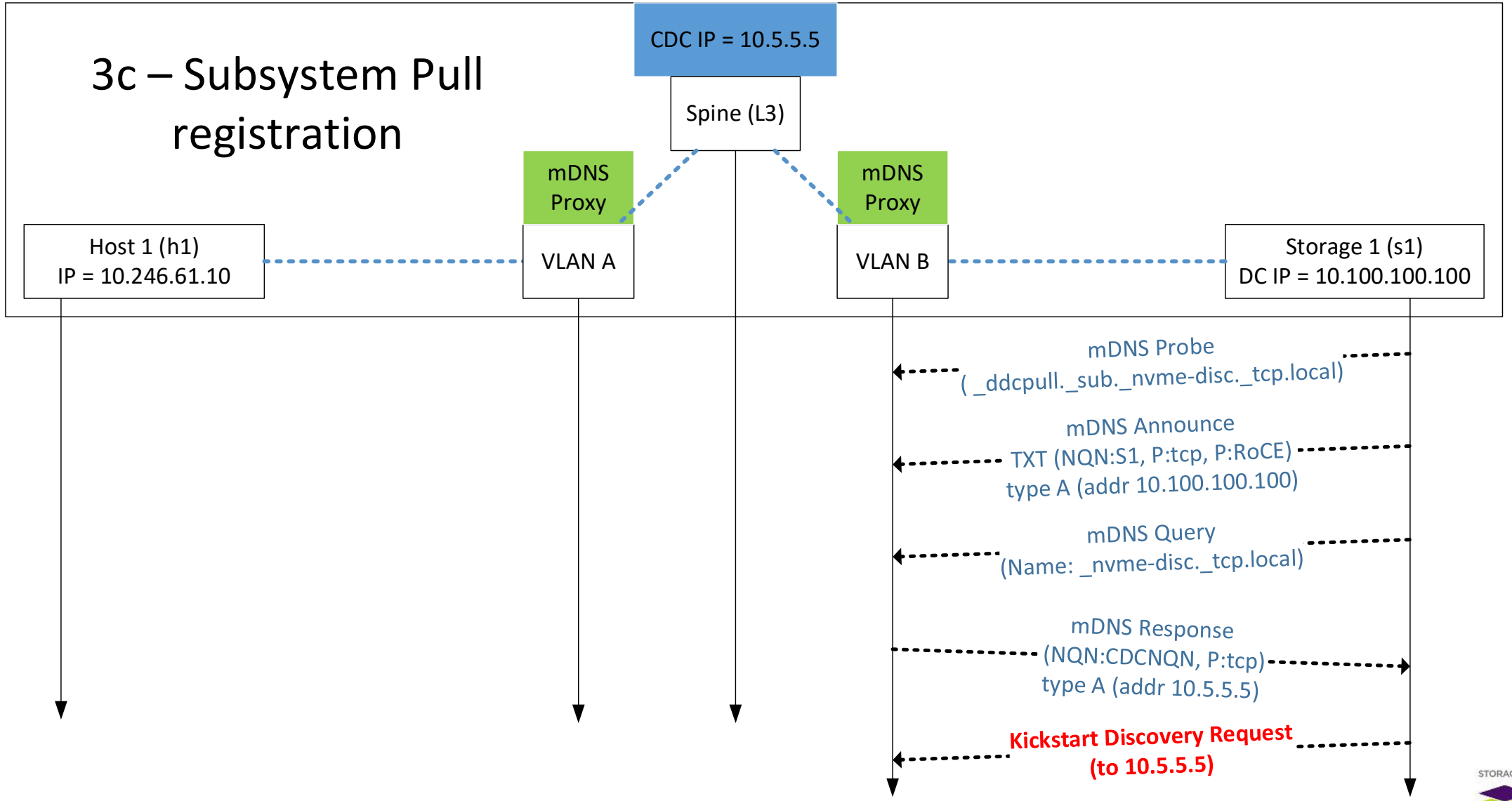
Single broadcast domain with CDC
CDC uses mDNS to discover DDC interfaces that require Pull registration



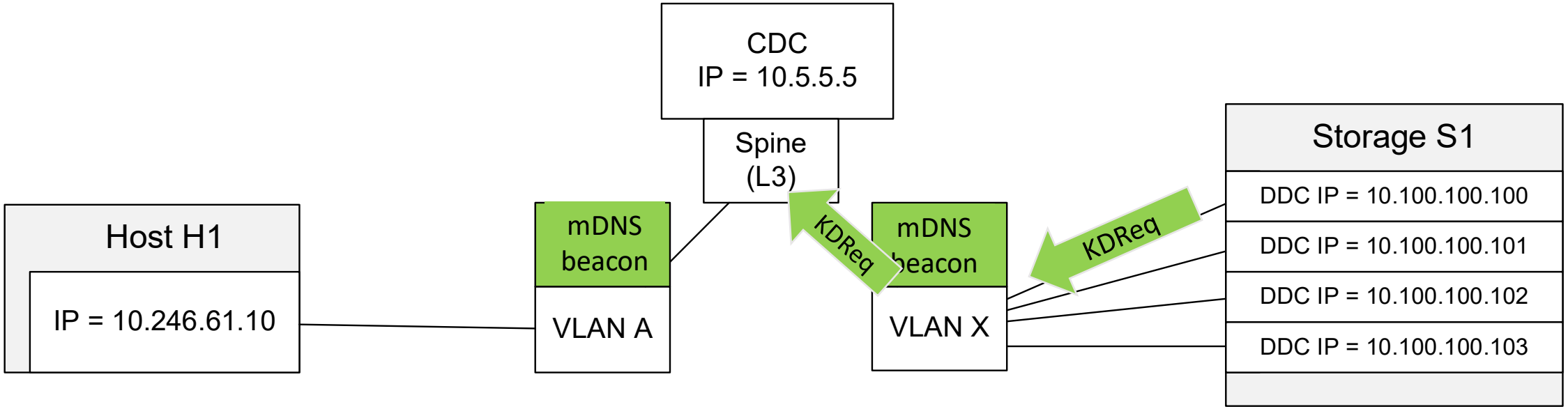
Multiple broadcast domains with CDC
Host and DDC will perform Push registration



Multiple broadcast domains with CDC
DDC will request Pull registration

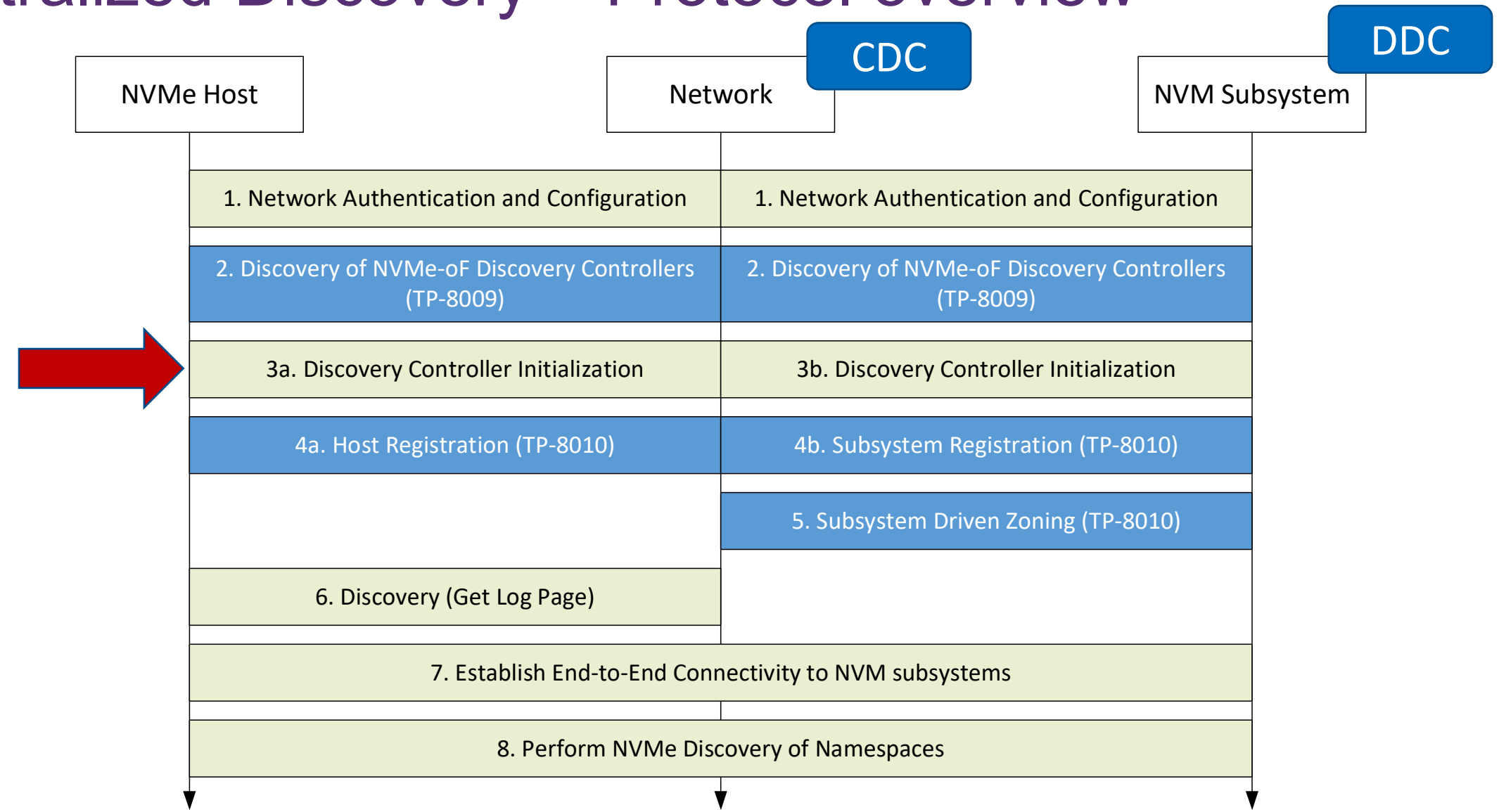


Why is a kickstart necessary?

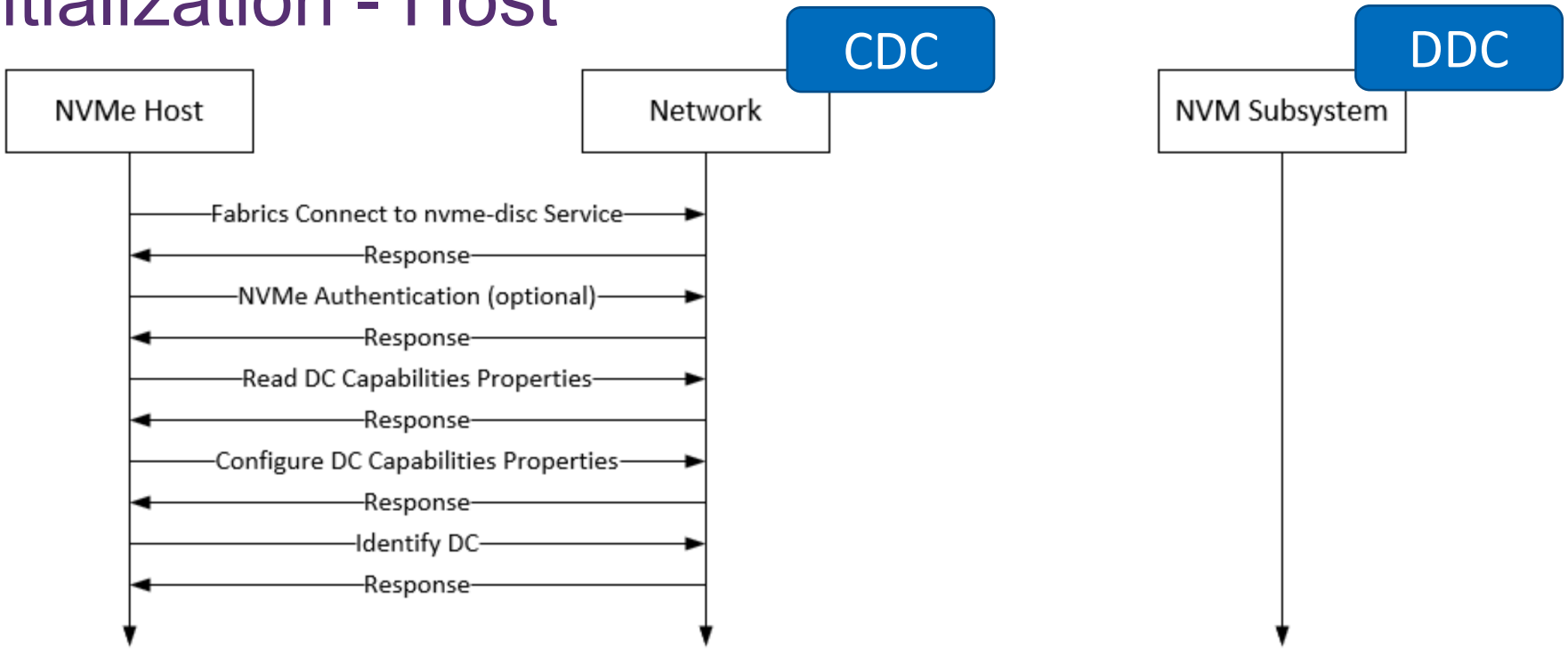


- After the CDC IP Address has been discovered (via mDNS beacon)
  - How does the CDC know that the subsystem needs to have information pulled from it?
  - What causes the CDC to send connect to the subsystem in the first place?
  - We need to be able to differentiate between a DDC that hasn't registered yet and a DDC that wants pull registration.
- If we use Kickstart Discovery Request
  - The mDNS beacon functionality can be a simple mDNS responder
- Without KDReq, the mDNS beacon would need to be an mDNS proxy (much more complicated)

Centralized Discovery – Protocol overview

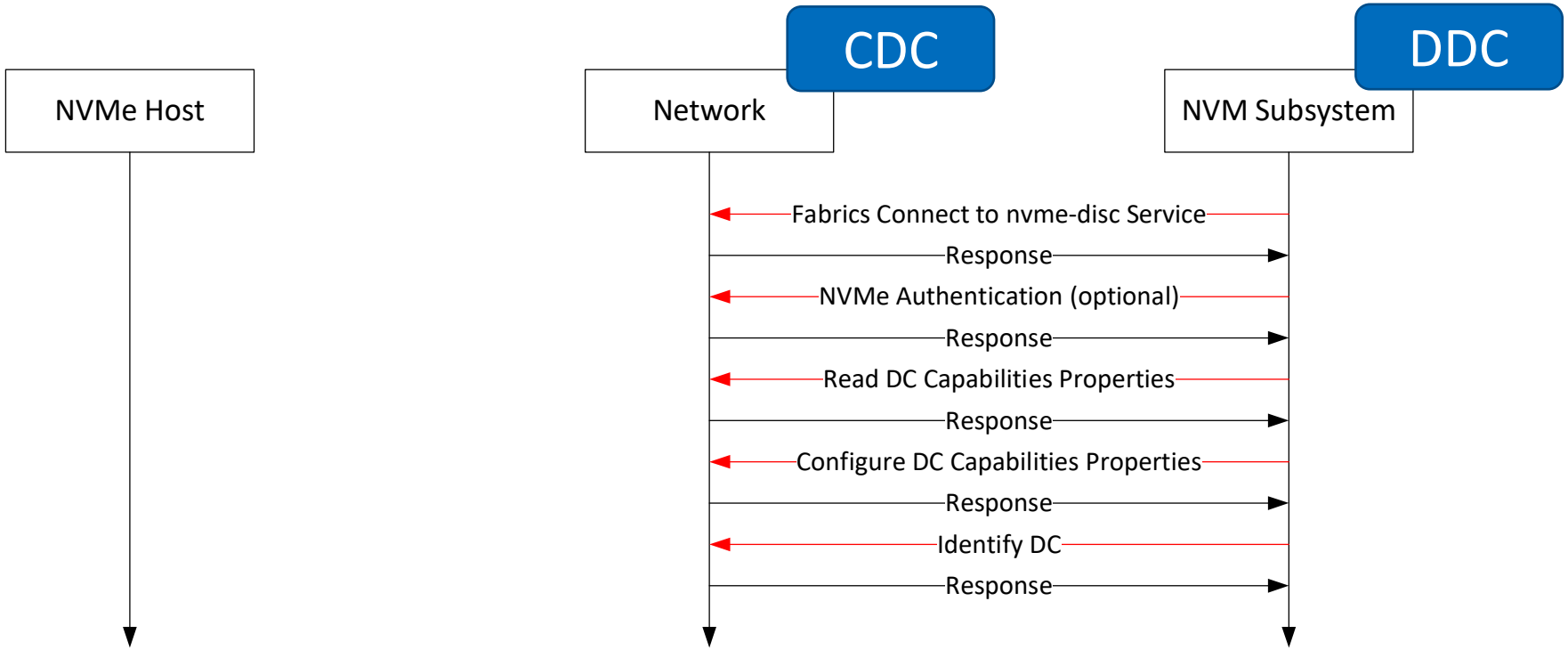


3a. DC Initialization - Host



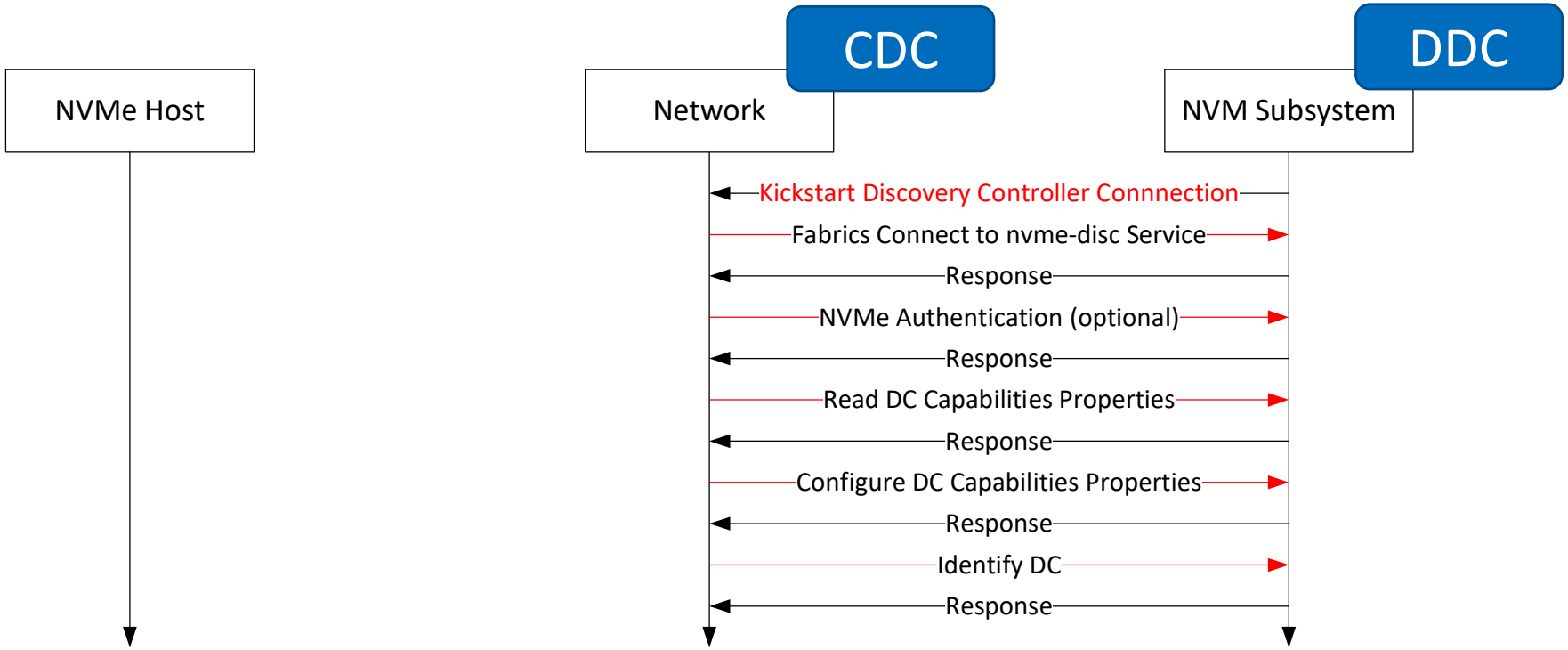
- After the Discovery Controller IP Addresses have been either:
  - Discovered (mDNS)
  - Configured (CLI, etc)
- The NVMe Host will establish a connection to the Discovery Controller(s) and initialize them.
  - This process is defined in the NVMe 2.0a Base Specification
  - With a Centralized Discovery Controller (shown), there will typically only be one CDC per VLAN
  - Without a Centralized Discovery Controller, there will be multiple DC per VLAN, at least one per NVM subsystem Interface
- End of this step: NVMe Host has established a connection to the CDC

3b. DC Initialization – Subsystem Push registration



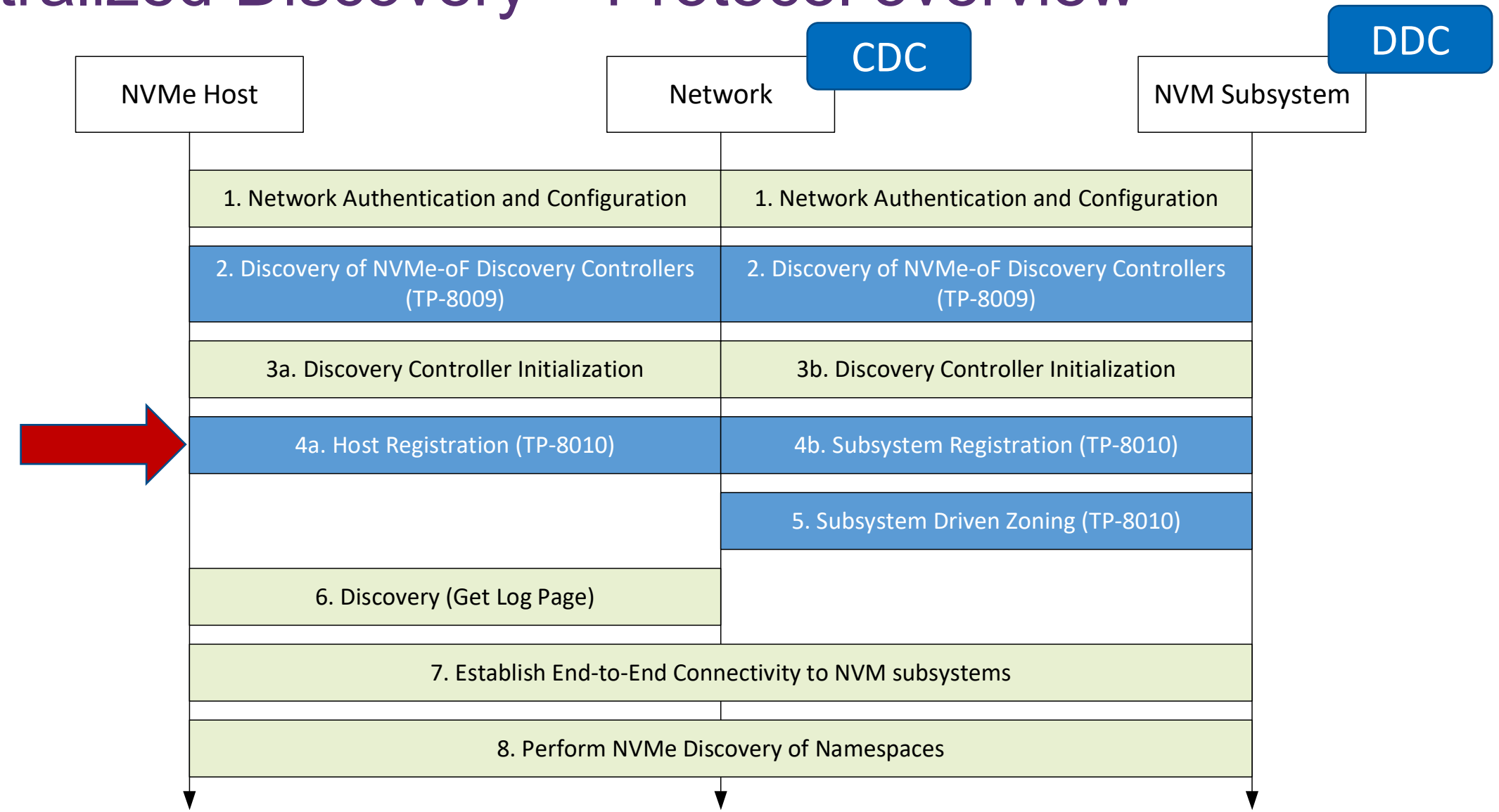
- After the CDC IP Address has been either:
  - Discovered (mDNS)
  - Configured (CLI, etc)
- The NVM subsystem will establish a connection to the CDC and initialize it.
  - This process is defined in NVMe 2.0a Base Specification
  - With an CDC (shown), there will usually only be one CDC per VLAN
- End of this step: NVM subsystem has established a connection to the CDC.

3b. DC Initialization – Subsystem Pull registration

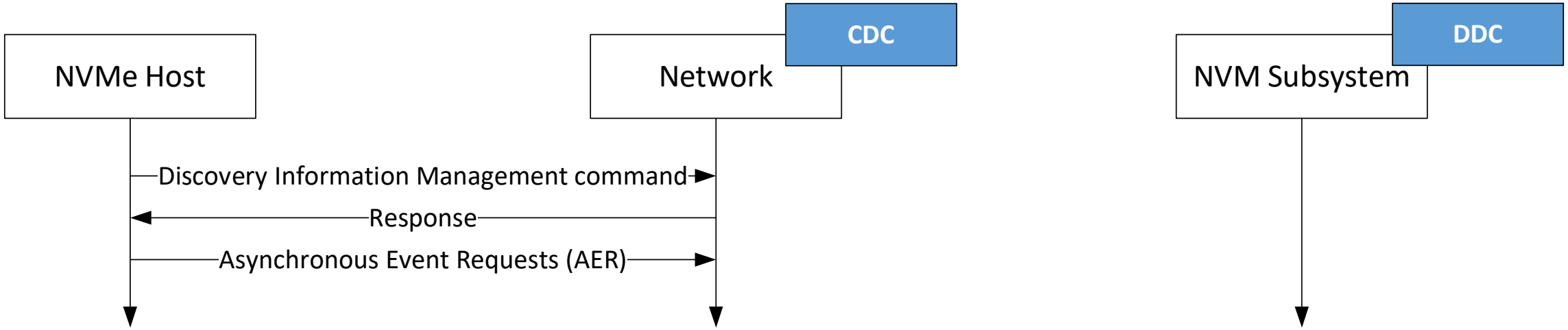


- After the CDC IP Address has been either:
  - Discovered (mDNS)
  - Configured (CLI, etc)
- The NVM subsystem will send the CDC a Kickstart Discovery Request command
- CDC will respond by sending connect to the subsystem and initializing the DDC on it.
  - This process is defined in NVMe 2.0a Base Specification
- End of this step: CDC has established a connection to the DDC on the NVM subsystem.

Centralized Discovery – Protocol overview

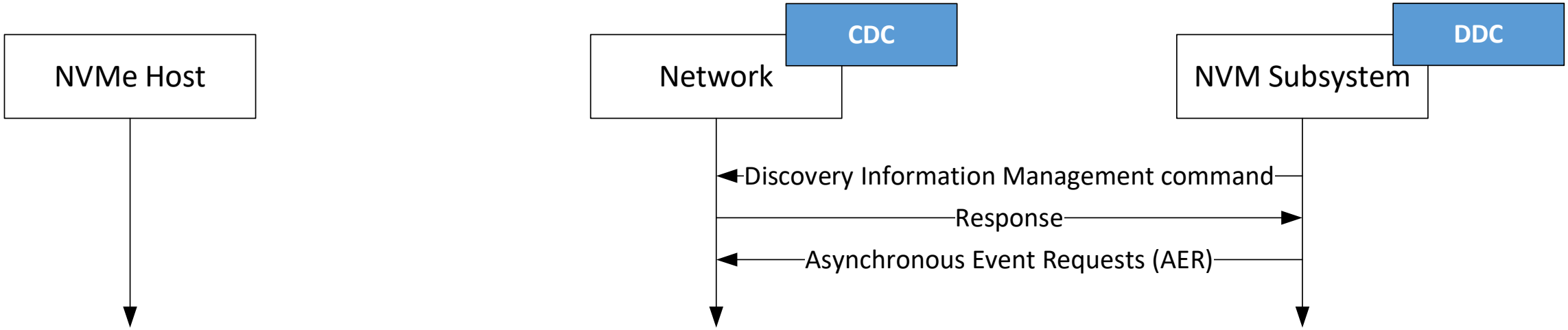


Step 4a - Host registers with CDC using Push registration



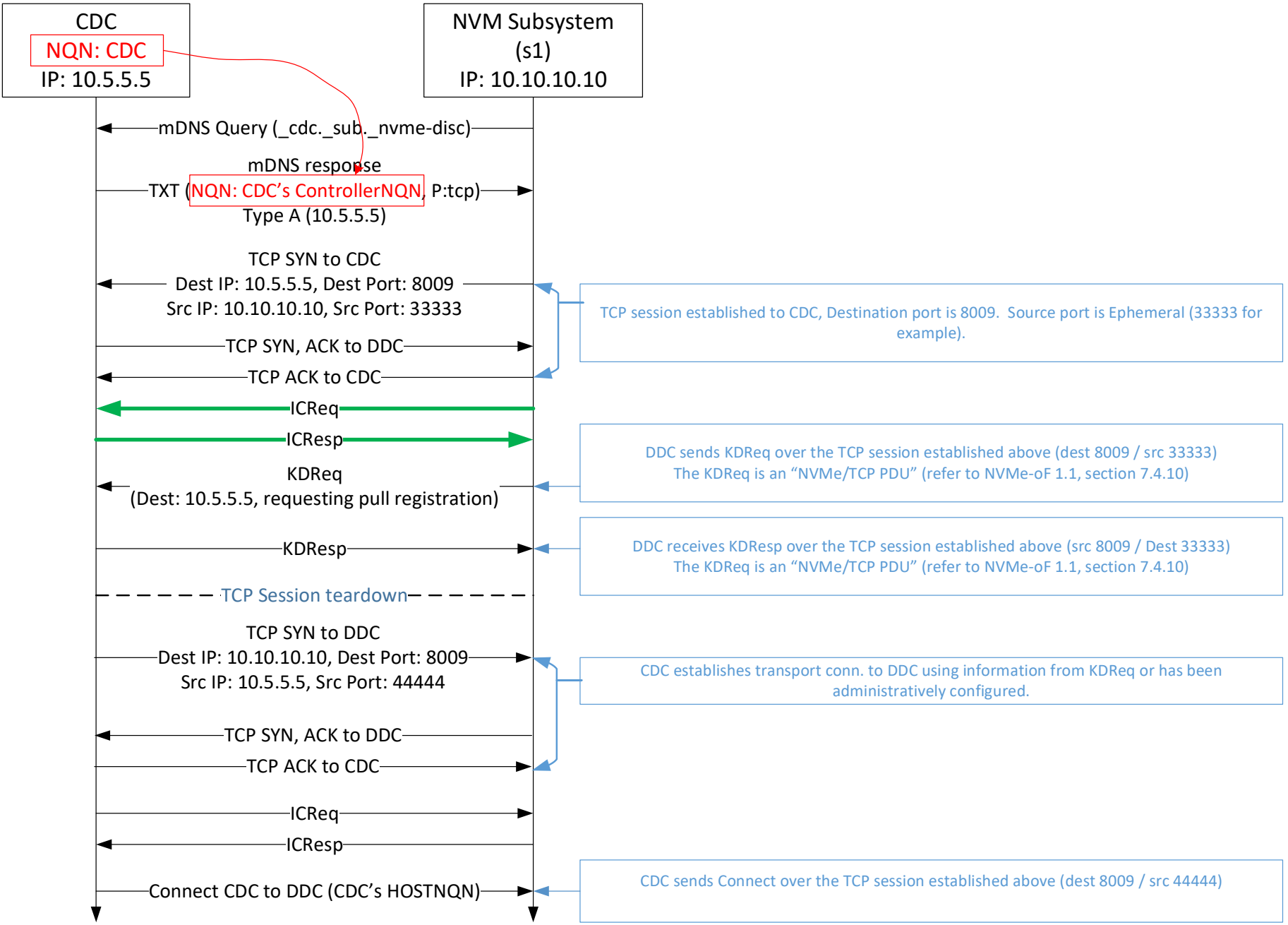
- During “3. Discovery Controller Initialization”, the type of Discovery Controller will have been discovered
  - i.e., Centralized Discovery Controller (CDC) or Direct Discovery Controller (DDC)
- If one or more CDCs were discovered, the NVMe Host will Register with them.
  - The information to be registered is effectively an enhanced NVMe Discovery Log Page (as defined in NVMe 2.0a Base Specification)
  - The extra information registered is a Symbolic Name
    - Could be a user-friendly name, a Group name, or Both
- Whether or not the DC is a CDC or DDC, the Host will register for Asynchronous Event Notifications (AEN) by using AER
- End of this step: Each Interface that discovered a CDC will have registered a log page and transmitted AER. **In the CDC case, the Host is now Discoverable by the subsystem**

Step 4b - Subsystem registers with CDC using Push registration



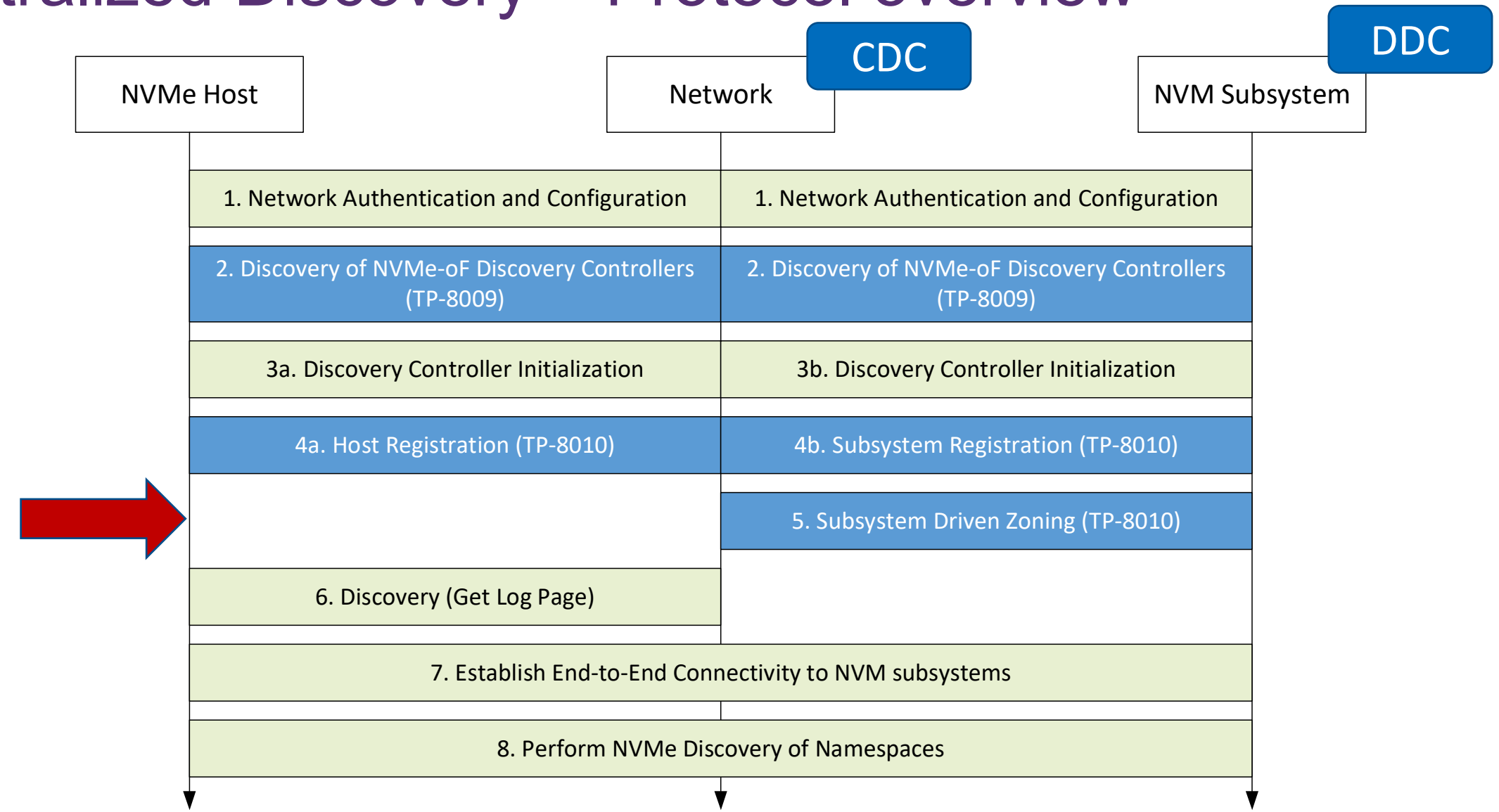
- During “3. Discovery Controller Initialization”, the type of Discovery Controller will have been discovered
  - i.e., CDC or DDC
- If one or more CDCs were discovered, the subsystem will Register with them.
  - The information to be registered is effectively an enhanced NVMe Discovery Log Page (as defined in NVMe 2.0a Base Specification)
  - The extra information registered is a Symbolic Name
    - Could be a user-friendly name, a Group name, or Both
- The subsystem will register for asynchronous notifications (AER)
- End of this step: Each Interface that discovered a CDC will have registered a log page and transmitted AER.
- The Subsystem is now Discoverable by the Host

Step 4b – Pull reg using Kickstart

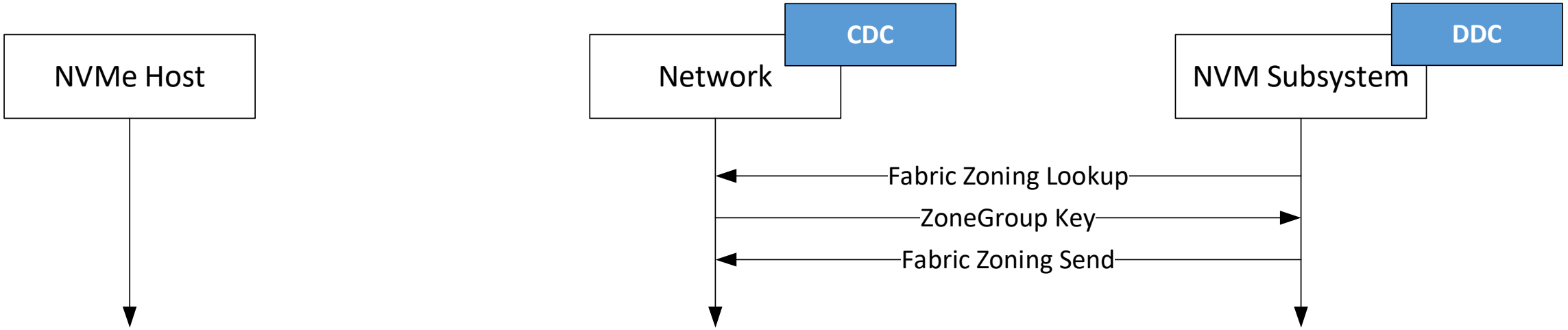


- ICReq and ICResp are used first.
- Update to ICReq allows for an indication that the connection will only be used for Kickstart (KDReq/KDResp)
- After connect from CDC to DDC, the CDC will transmit get log page and specify a new log page identifier that requests "**Port-Local DLPEs**" only.

Centralized Discovery – Protocol overview



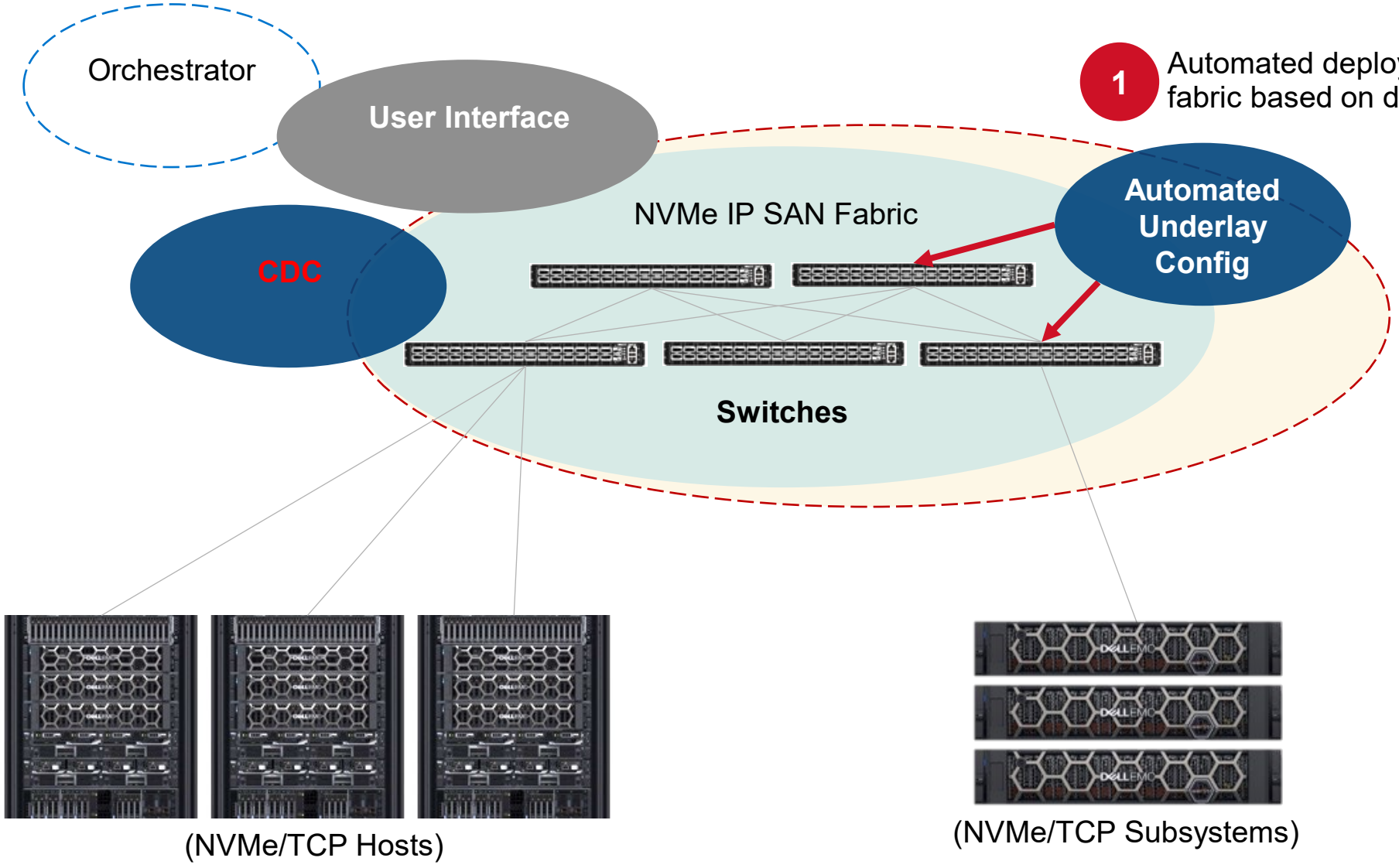
Step 5 - Subsystem Driven Zoning (SDZ) – A.K.A. Target Driven Zoning



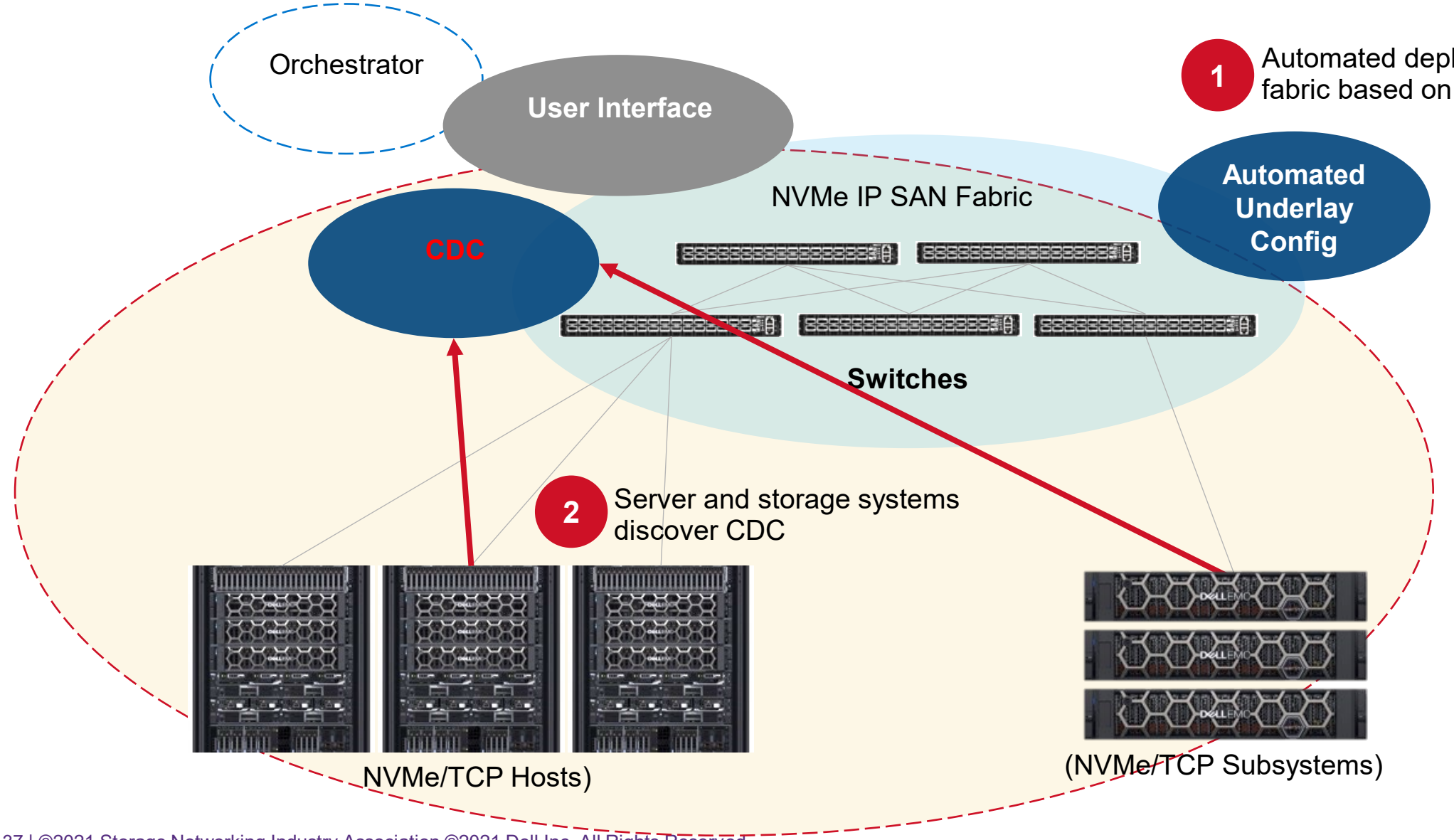
- As a part of the storage provisioning process, the subsystem may send a ZoneGroup to the CDC.
- The ZoneGroup describes which Hosts are allowed to access each subsystem interface.
- In the context of a CDC, the ZoneGroup is the unit of activation (like a FC zone set).
- A CDC instance may have multiple ZoneGroups active at the same time to avoid potential configuration clashes between multiple administrators.
- The process starts by the subsystem retrieving a Zoning Data Key
- The subsystem can then send the ZoneGroup definition using Fabric Zoning Send.
- The ZoneGroup definition SHOULD match the namespace masking definition. This allows for single-pane of glass management.

End-to-End Automated Discovery example

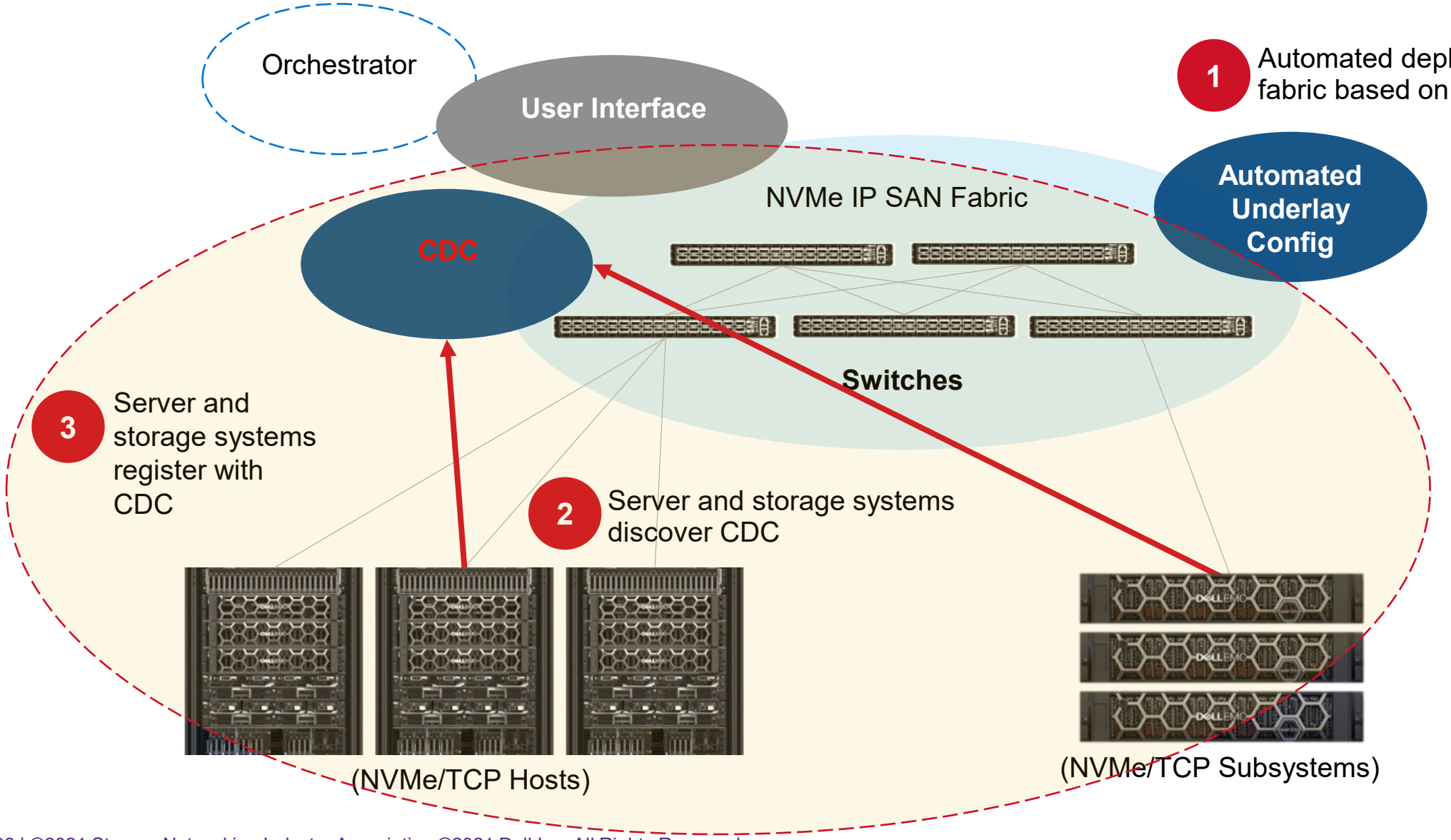
NVMe IP SAN operations



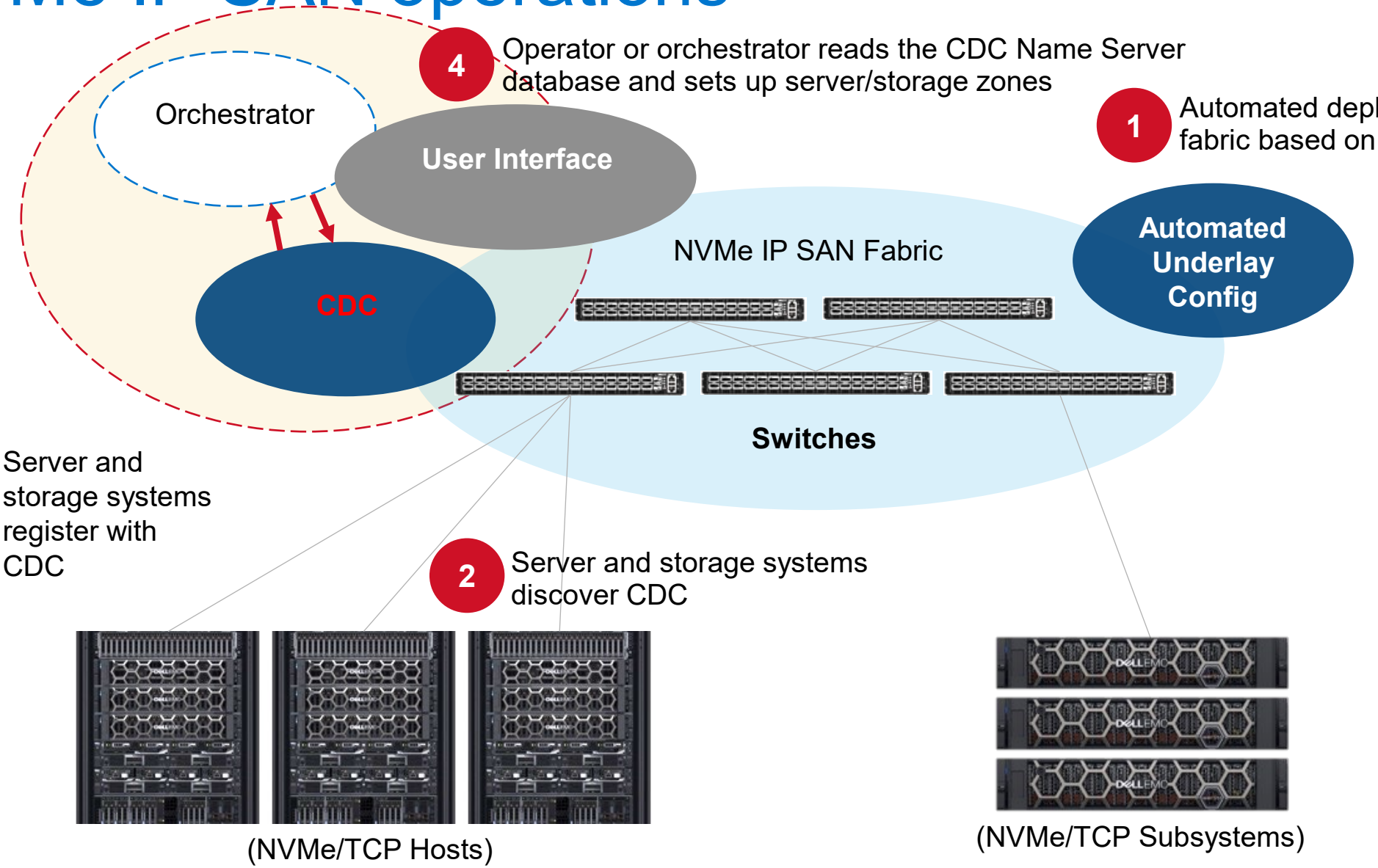
NVMe IP SAN operations



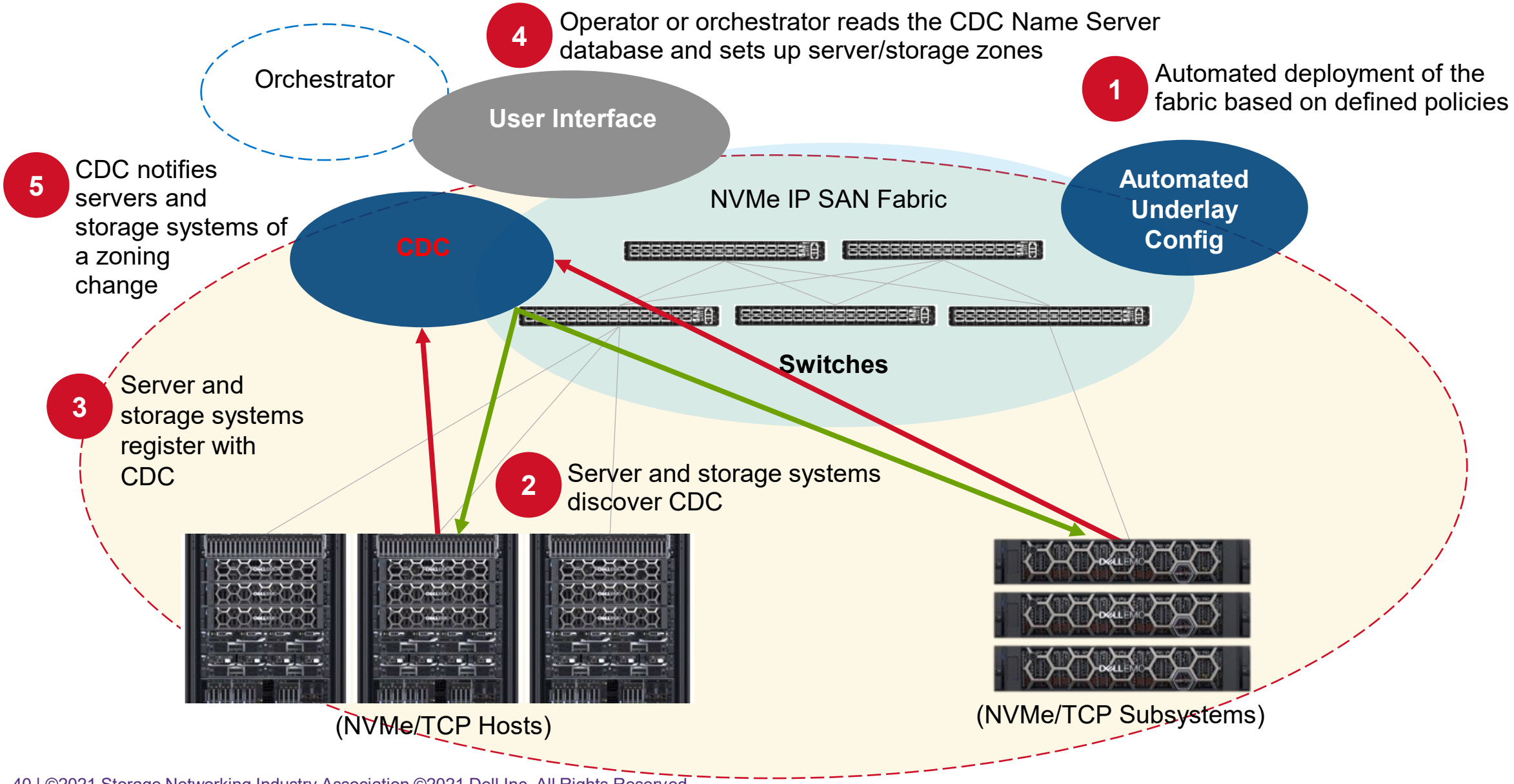
NVMe IP SAN operations



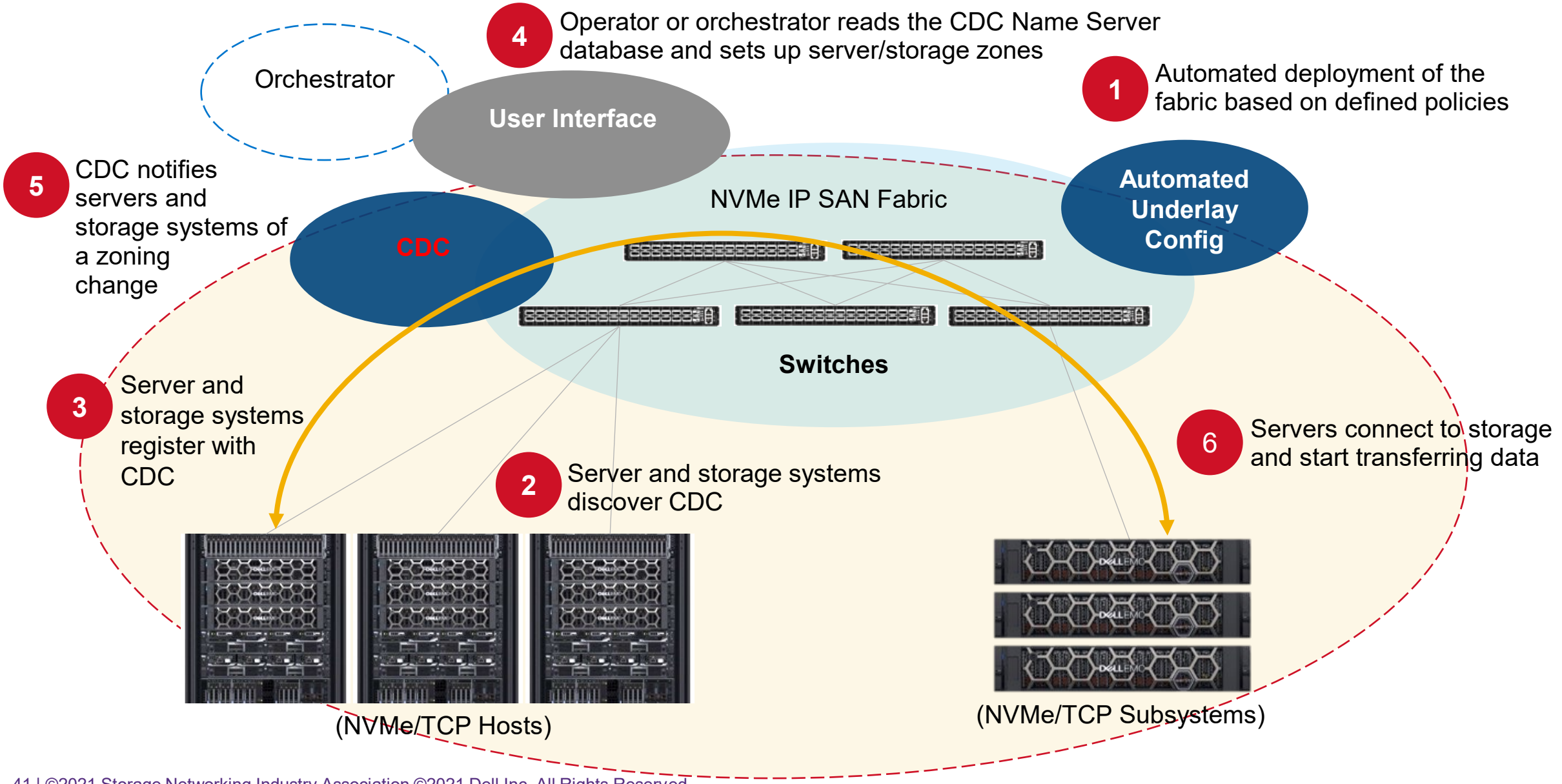
NVMe IP SAN operations



NVMe IP SAN operations



NVMe IP 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).

Please take a moment to rate this session.

Your feedback is important to us.

