SNIA DEVELOPER CONFERENCE
SDC 24
BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

Evaluating Discovery Automation on a Large scale NVMe-oF deployment

Raj Kumar Dani (Associate Director)
Samsung Semiconductor Inc

Swati Chawdhary (Senior Manager)
Nehal Kumar Ram (Staff Engg)
Sathish Kumar M (Associate Tech Director)

Agenda

- NVMeOF Overview and Discovery Controller
- Automated Discovery Methods (DDC and CDC) †
- Large-scale automated discovery network
- Impact of scaling in DDC vs CDC
- Demo
- Observations and recommendation
- Summary

†DDC - Direct Discovery Controller
†CDC - Centralized Discovery Controller

NVMeOF Overview and Discovery Controller

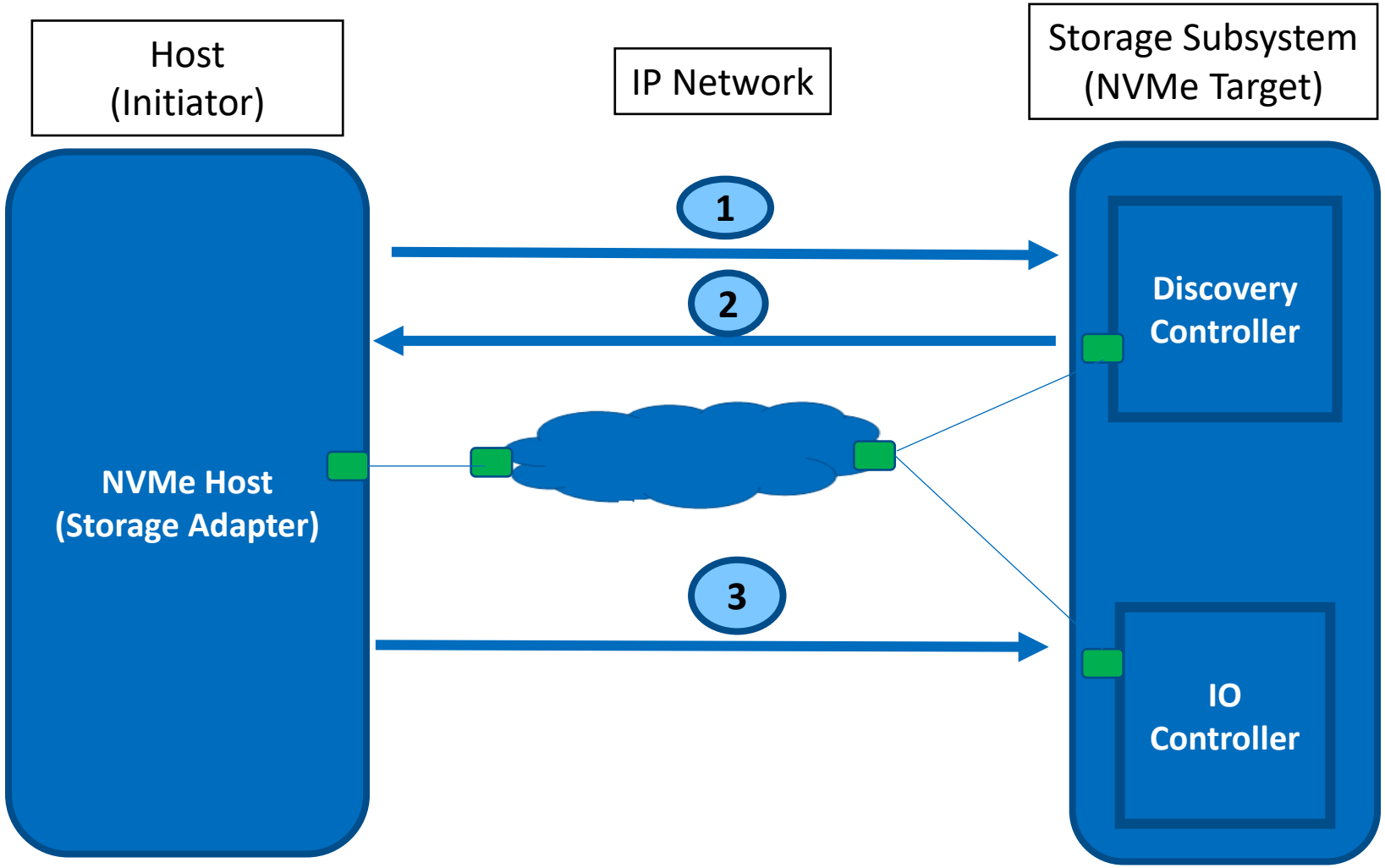
- **NVMe-oF** : Non-Volatile Memory Express over Fabrics
  - Enable NVMe commands to transfer data between host and storage subsystem, over a network fabric
  - Several automated discovery mechanisms to simplify process

- IP-based fabric transports, require each Host configure to connect to each Discovery Controller.

- Discovery controller: Single location of all known NVM subsystem interfaces for discovery

- Administrators mostly configure individually Discovery Controller on each Host
  - Example: “nvme discover -t tcp -a 192.168.1.2 -s 4420”
  - Challenge - Limits the scale and interop of any IP based NVMe-oF solution

NVMeOF Discovery Connection (without DDC and CDC)



1. Using configured IP, hosts query to a Discovery controller
2. Host finds the Discovery log page
   - Contains Network addresses and unique NVMe Qualified Names (NQNs)
3. Using Discovery Log page info, host connect to IO controller NVM subsystem

Repeat all the 1-3 steps for each IP subsystems

Automated Discovery Methods

- New functionality described in TP-8009 and TP-8010 (NVM Express) describe a standardized and scalable automated discovery process for IP network, as follows :
  - Direct Discovery Controllers (DDC) TP8009:
    - Provide information about subsystem interfaces using mDNS
    - Allow hosts to directly discover and connect to storage resources without needing a centralized discovery mechanism.
  - Centralized Discovery Controller (CDC) TP8010:
    - Controller aggregates discovery information and connectivity constraints for all hosts / subsystems in NVMe IP-based SAN
    - Each host / subsystem automatically discovers CDC, simplifies administration and reduces overhead

Direct Discovery Controller (TP8009)

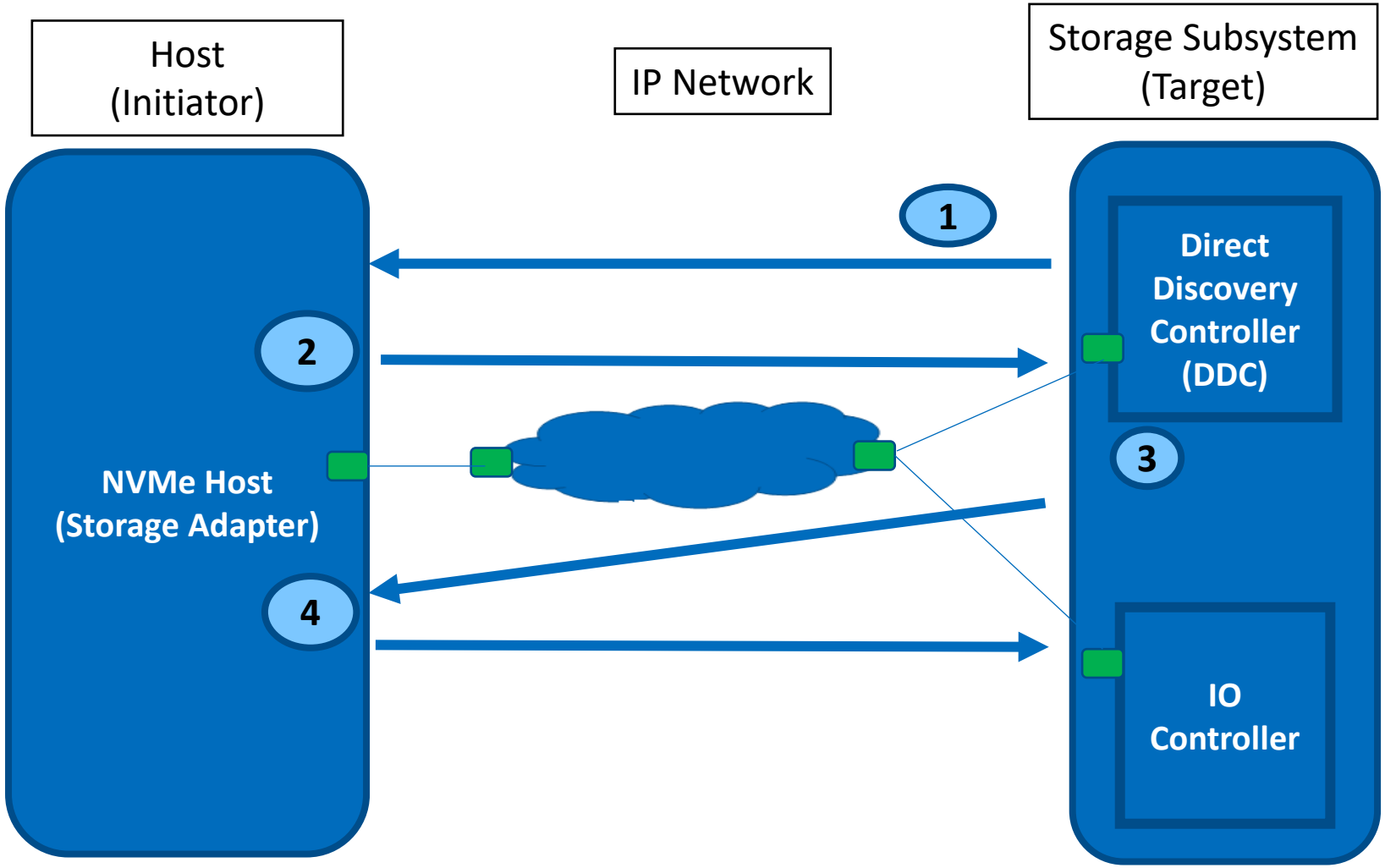
Direct Discovery Controllers (DDCs)

- NVMe-oF utilize mDNS (multicast DNS) and DNS-SD (DNS Service Discovery) to simplify and automate discovery

- mDNS (multicast DNS):
  - Allows devices on local network to discover each other without needing a central DNS

- DNS-SD (DNS Service Discovery):
  - Works with mDNS to advertise and discover services on a local n/w
  - Allows devices to announce services and discover services offered by other devices.

Configuration Steps with Automated Discovery of DDCs (TP8009)



1. Host and Subsystems discovers automatically via mDNS

2. Host automatically sends connect to DDC at given IP address via mDNS

3. Storage admin provides Namespace to the Host NQN

4. Host receives Aysnc Event Notification from DDCs, retrieve the list of IO controller and connect them

Repeat 2-4 on all hosts for each subsystem

Direct Discovery Controller – Advantages / Limitations

Direct Discovery Controller

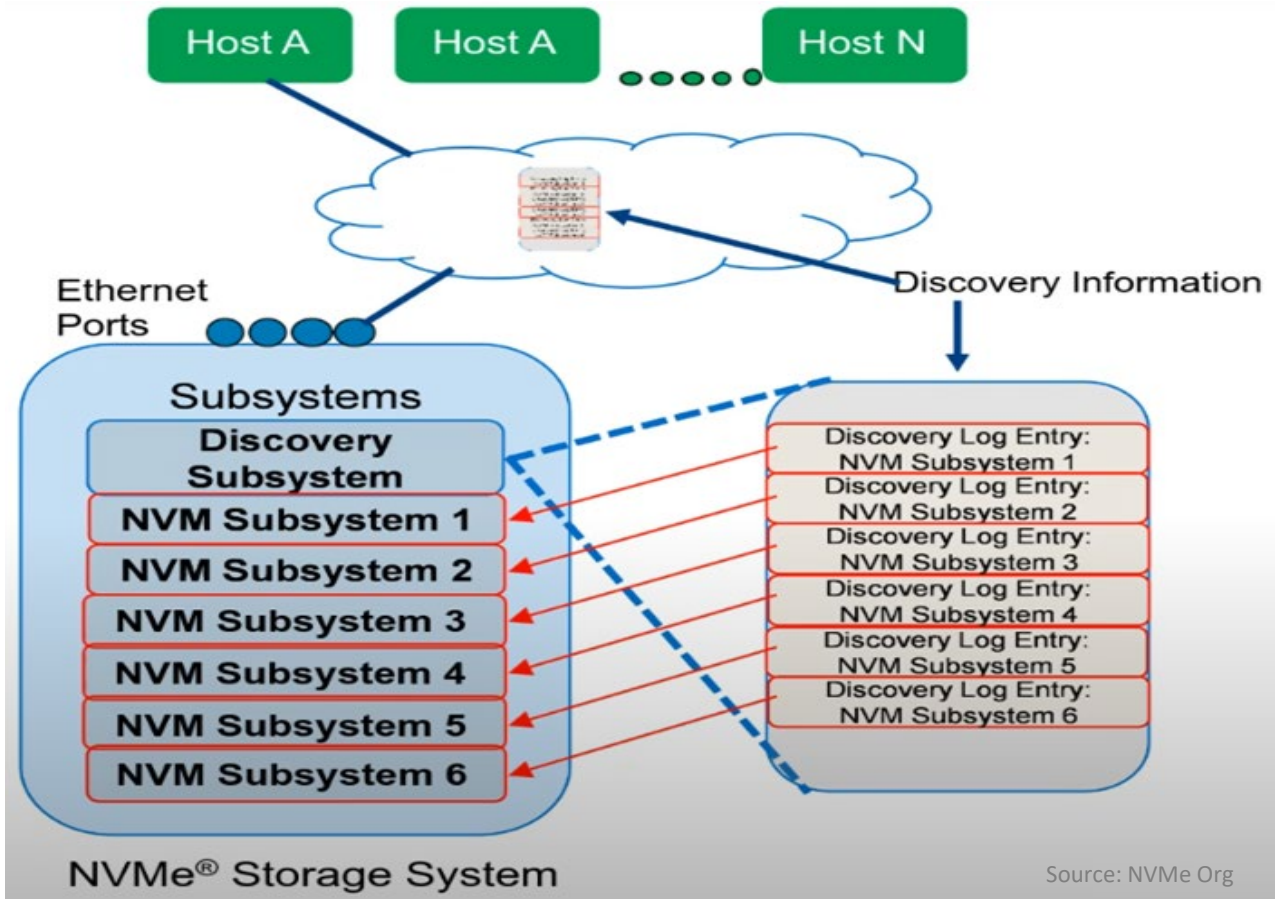
- Advantages (DDC):
  - Simpler and more straightforward
  - Smaller networks where the number of hosts and subsystems is manageable.
  - Reduces the need for additional infrastructure (centralized discovery is not necessary)

- Limitations:
  - More complex environments
  - In larger, direct discovery will become complex (manually manage and configure)
  - Here CDCs offer significant advantages by automating and simplifying discovery process

Direct Discovery

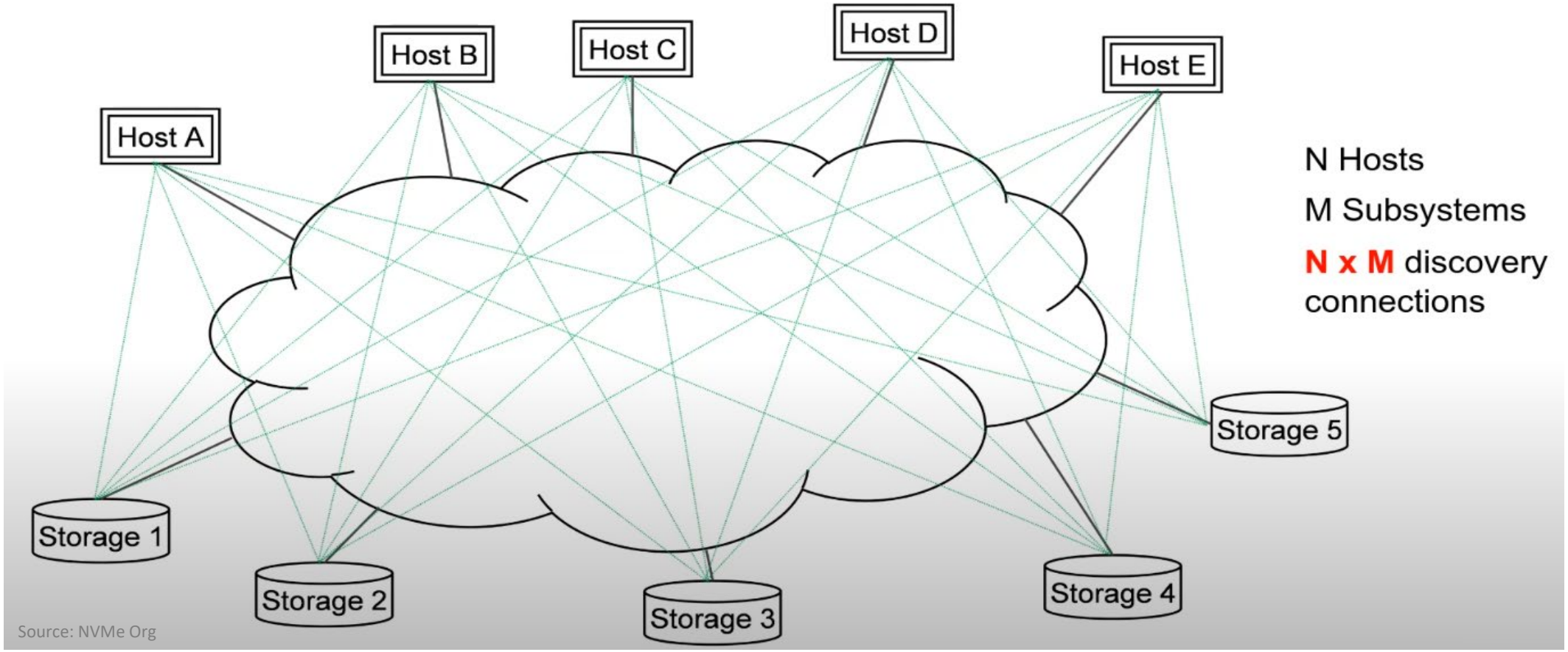
- Administrator configures path to a Discovery Subsystem
- Host connects to Discovery Controller in the Discovery Subsystem
- Discovery information is reported in discovery log page entries

Note: Significant Discovery log entry's carry inside network



Direct Discovery Controller Scaling Problem

- DDC : Discovery controller becomes a challenge on scaled full mesh network



Centralized Discovery Controller (TP8010)

Centralized Discovery Controllers(CDCs):

- Simplify and enhance discovery process in larger, more complex storage networks.

Following are the key aspects involved during discovery :

Centralized Management

- Aggregate discovery information for all hosts / subsystems - reduces administrative overhead
- Eliminates the need for manual configuration and reduces the risk of errors

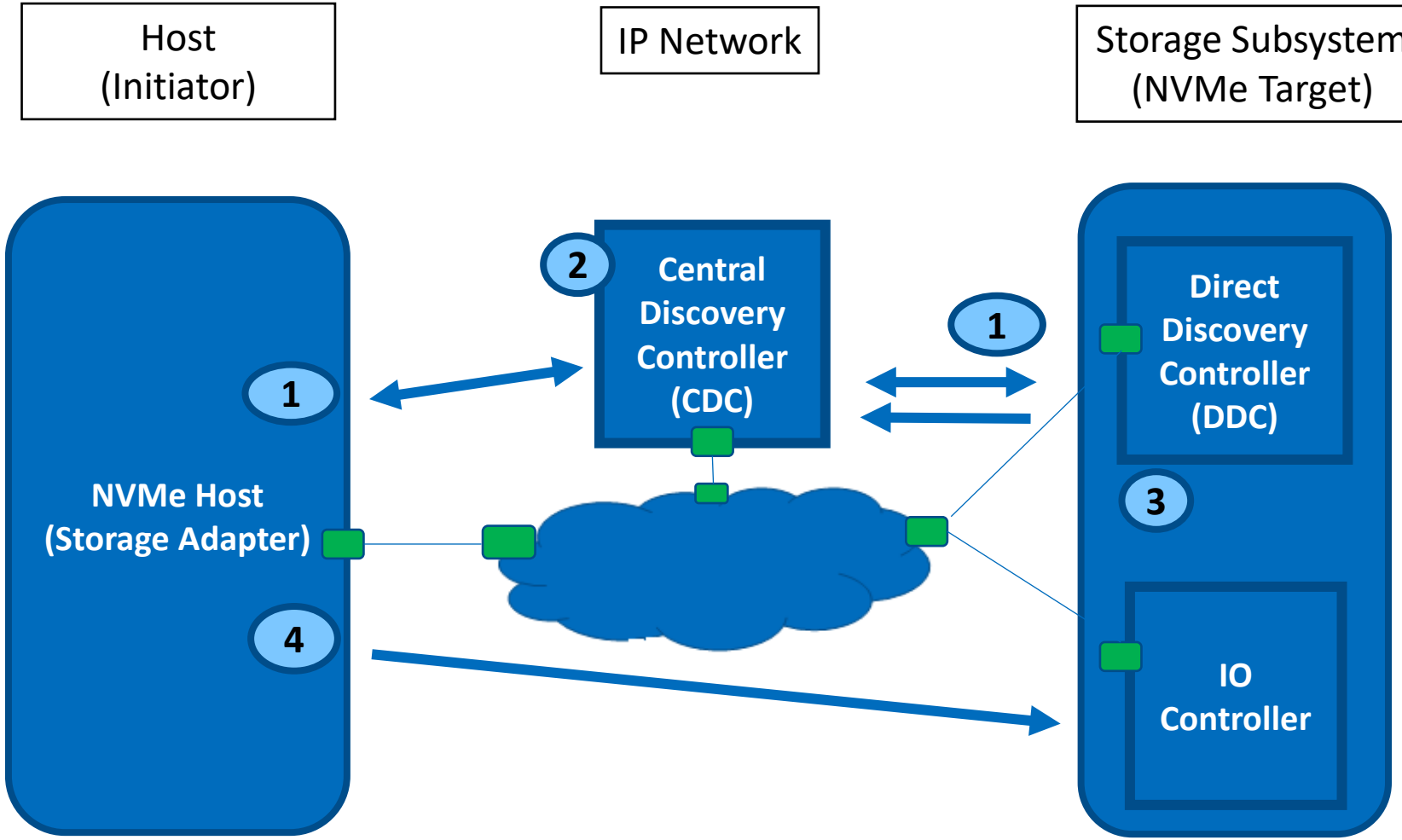
Zoning and Security

- CDCs support zoning: allows administrators to define and enforce connectivity constraints.
- Enhances security and ensures that only authorized hosts can access specific storage resources

Scalability:

- Centralizing the discovery process, CDCs make it easier to scale the network.
- Dynamic hosts / subsystems addition without significant reconfiguration

Configuration Steps with Automated Discovery of CDC (TP8010)



1. Host / Subsystems discovers automatically CDC & register discovery info

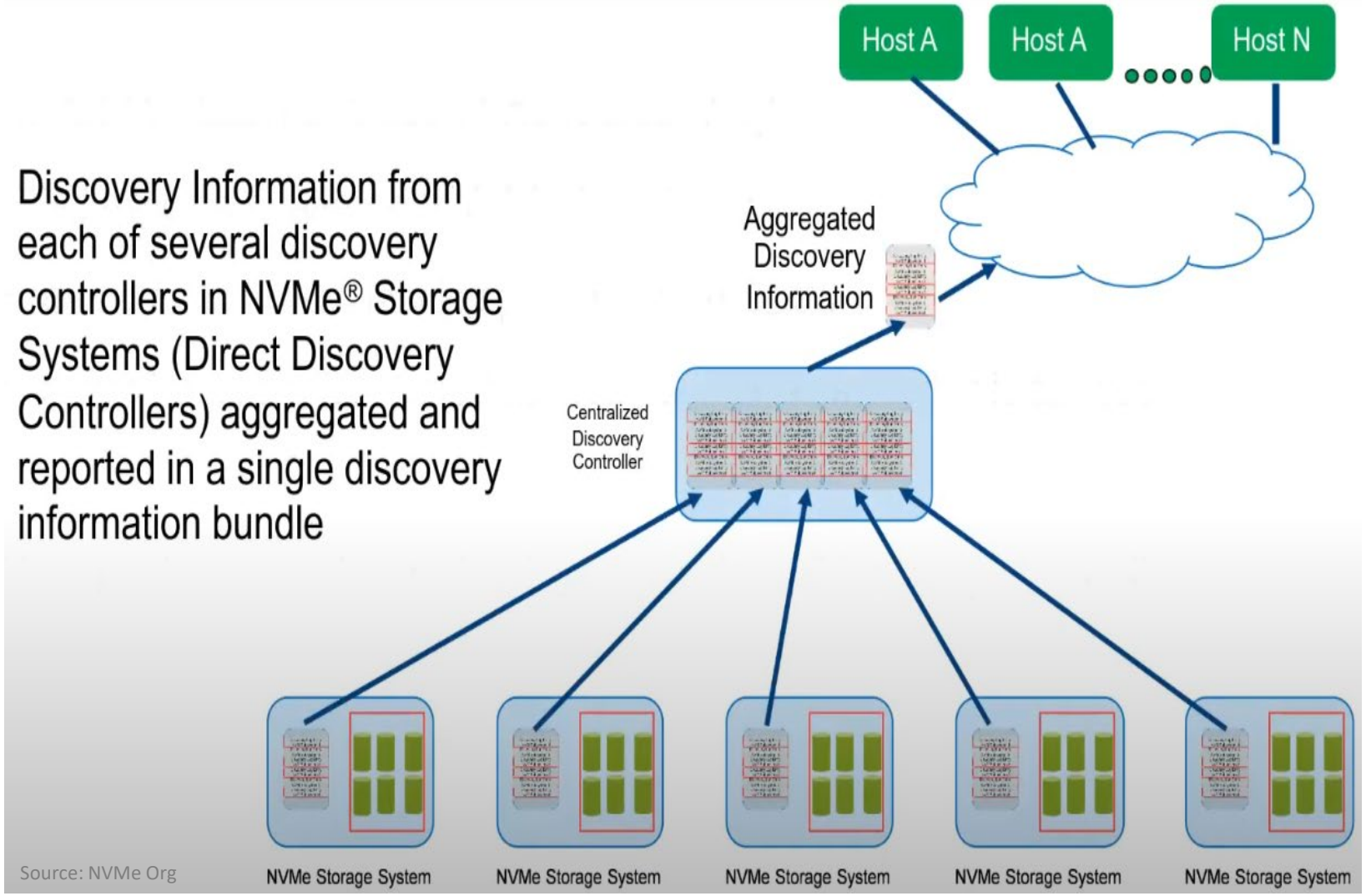
2. Zoning performed on CDC (optional)

3. Storage admin provides Namespace to the Host NQN, may need zoning to send to CDC

4. After zoning, Host receives Aysnc Event Notification, use get log page and connect to each IO controller

Repeat 2-3 on all hosts for each subsystem

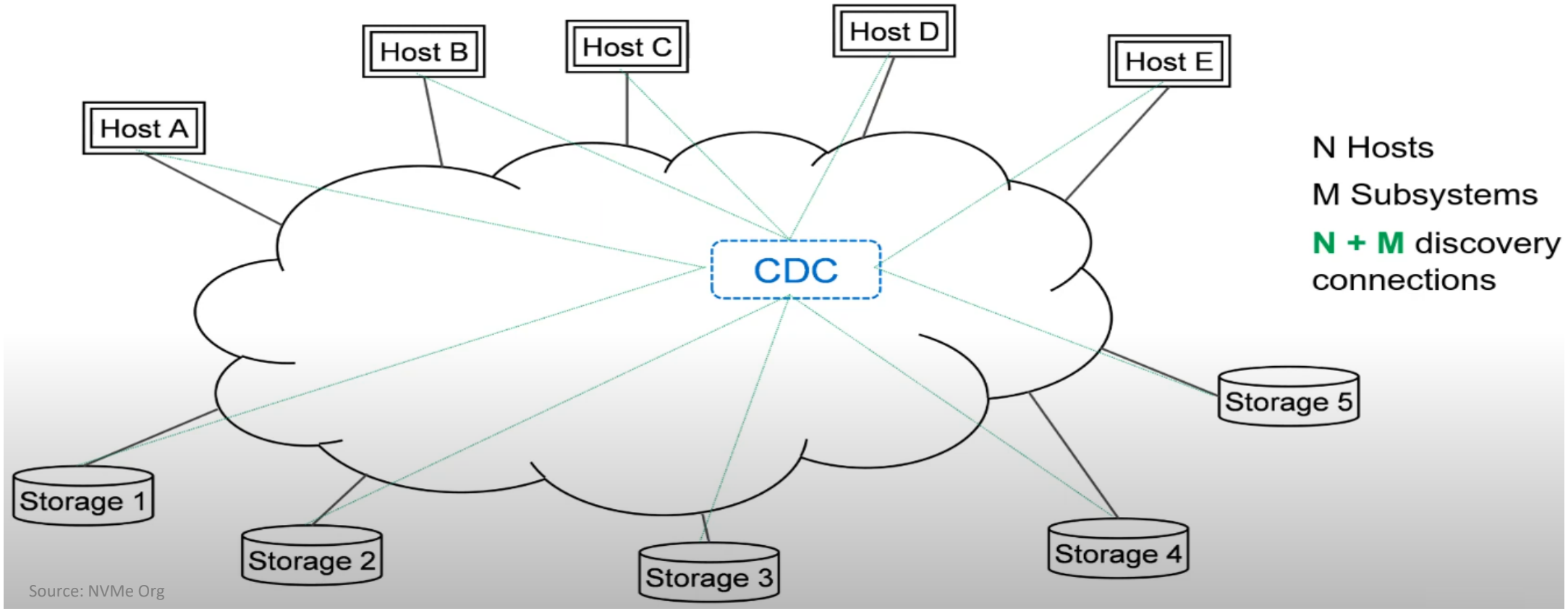
Centralized Discovery



- CDC reports available NVM Subsystems
- Same format Discovery log pages as with DDCs
- Same Host specific accessible NVM Subsystems filtering is allowed
- Host is able to register with the Centralized Discovery Controller

Scaling with Centralized Discovery Controller

- CDC : Centralized discovery controller - aggregated information on scaled network
  - Respected host connect to required storage subsystems only



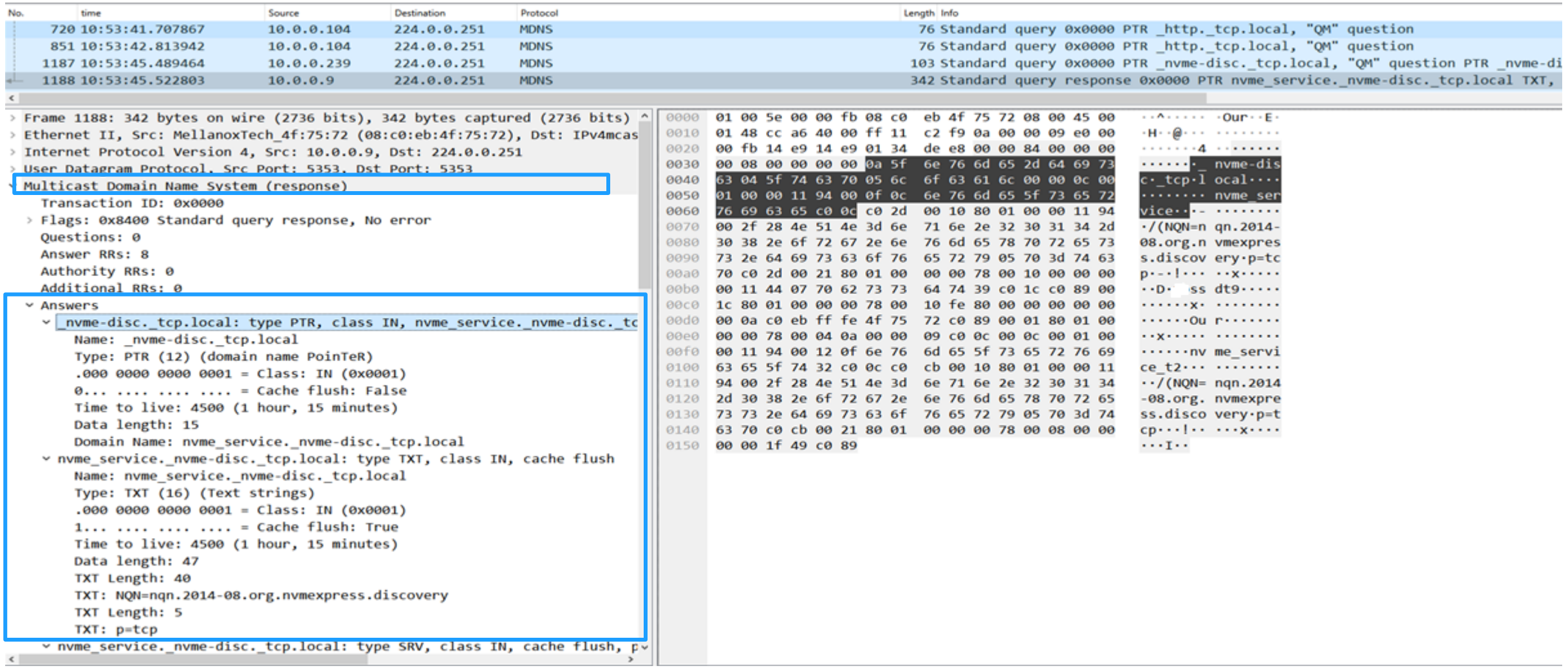
Direct Discovery (mDNS Query) - Trace

Destination	Protocol	Length	Info
224.0.0.251	MDNS	14…	Standard query 0x0000 SRV nvme_service._nvme-disc._tcp.local, "QM" question TXT spdk0_10.0.0.233_1153._nvme-disc._tcp.lo
224.0.0.251	MDNS	614	Standard query 0x0000 TXT A 10.0.0.236 AAAA fe80::966d:aeff:fe97:3b7c SRV 0 0 1153 9F-SMC-Type1-236.local TXT SRV 0 0 11
224.0.0.251	MDNS	239	Standard query response 0x0000 PTR, cache flush 9F-SMC-TYPE3-239.local A, cache flush 10.0.0.239 PTR, cache flush 9F-SMC
10.0.0.9	NVMe	142	NVMe Get Log Page Unspecified
10.0.0.239	NVMe	94	NVMe CQE for Get Log Page Unspecified
10.0.0.9	NVMe	142	NVMe Get Log Page NVMeOF Discovery
10.0.0.239	NVMe	118	NVMeOF Data for Get Log Page NVMeOF Discovery, offset 0
10.0.0.9	NVMe	142	NVMe Get Log Page NVMeOF Discovery
10.0.0.239	NVMe	21…	NVMeOF Data for Get Log Page NVMeOF Discovery, offset 0
10.0.0.9	NVMe	142	NVMe Get Log Page NVMeOF Discovery
10.0.0.239	NVMe	118	NVMeOF Data for Get Log Page NVMeOF Discovery, offset 0
224.0.0.251	MDNS	14…	Standard query 0x0000 PTR _nvme-disc._tcp.local, "QM" question TXT spdk0_10.0.0.232_1153._nvme-disc._tcp.local, "QM" que
224.0.0.251	MDNS	836	Standard query 0x0000 SRV 10-0-150-102:09/02/24:05:11:55._nvme-disc._tcp.local, "QM" question TXT spdk0_10.0.0.233_1153.
224.0.0.251	MDNS	11…	Standard query 0x0000 PTR _nvme-disc._tcp.local, "QM" question TXT nvme_service._nvme-disc._tcp.local, "QM" question SRV
224.0.0.251	MDNS	239	Standard query response 0x0000 PTR, cache flush 9F-SMC-TYPE3-239.local A, cache flush 10.0.0.239 PTR, cache flush 9F-SMC
10.0.0.9	NVMe/TCP	194	Initialize Connection Request
10.0.0.239	NVMe/TCP	194	Initialize Connection Response

```
> Ethernet II, Src: MellanoxTech_98:c2:bd (94:6d:ae:98:c2:bd), Dst: IPv4mcas
> Internet Protocol Version 4, Src: 10.0.0.239, Dst: 224.0.0.251
> User Datagram Protocol, Src Port: 5353, Dst Port: 5353
v Multicast Domain Name System (query)
  > Transaction ID: 0x0000
  > Flags: 0x0000 Standard query
    Questions: 18
    Answer RRs: 27
    Authority RRs: 0
    Additional RRs: 0
  v Queries
    v _nvme-disc._tcp.local: type PTR, class IN, "QM" question
        Name: _nvme-disc._tcp.local
        [Name Length: 21]
        [Label Count: 3]
        Type: PTR (12) (domain name PoinTeR)
        .000 0000 0000 0001 = Class: IN (0x0001)
        0... .... .... .... = "QU" question: false
    > nvme_service._nvme-disc._tcp.local: type TXT, class IN, "QM" question
```

```
01b0  00 96 67 c0 a6 00 21 00  01 00 00 00 78 00 08 00   ··g···!· ····x···
01c0  00 00 00 04 81 c0 92 c0  a6 00 10 00 01 00 00 11   ········ ········
01d0  94 00 2f 05 70 3d 74 63  70 28 6e 71 6e 3d 6e 71   ··/·p=tc p(nqn=nq
01e0  6e 2e 32 30 31 34 2d 30  38 2e 6f 72 67 2e 6e 76   n.2014-0 8.org.nv
01f0  6d 65 78 70 72 65 73 73  2e 64 69 73 63 6f 76 65   mexpress .discove
0200  72 79 c0 92 00 01 00 01  00 00 00 78 00 04 0a 00   ry······ ···x····
0210  00 e7 c0 92 00 1c 00 01  00 00 00 78 00 10 fe 80   ········ ···x····
0220  00 00 00 00 00 00 96 6d  ae ff fe 97 3b 9d c0 70   ·······m ····;··p
0230  00 21 00 01 00 00 00 78  00 08 00 00 00 00 04 80   ·!·····x ········
0240  c0 92 c0 70 00 10 00 01  00 00 11 94 00 2f 05 70   ···p···· ·····/·p
0250  3d 74 63 70 28 6e 71 6e  3d 6e 71 6e 2e 32 30 31   =tcp(nqn =nqn.201
0260  34 2d 30 38 2e 6f 72 67  2e 6e 76 6d 65 78 70 72   4-08.org .nvmexpr
0270  65 73 73 2e 64 69 73 63  6f 76 65 72 79 c0 54 00   ess.disc overy·T·
0280  21 00 01 00 00 00 78 00  08 00 00 00 00 1f 49 c0   !·····x· ·····I·
0290  40 c0 54 00 10 00 01 00  00 11 94 00 2f 28 4e 51   @·T····· ····/(NQ
02a0  4e 3d 6e 71 6e 2e 32 30  31 34 2d 30 38 2e 6f 72   N=nqn.20 14-08.or
02b0  67 2e 6e 76 6d 65 78 70  72 65 73 73 2e 64 69 73   g.nvmexp ress.dis
02c0  63 6f 76 65 72 79 05 70  3d 74 63 70 c0 40 00 01   covery·p =tcp·@··
02d0  00 01 00 00 00 78 00 04  0a 00 00 09 c0 40 00 1c   ·····x·· ·····@··
02e0  00 01 00 00 00 78 00 10  fe 80 00 00 00 00 00 00   ·····x·· ········
02f0  0a c0 eb ff fe 4f 75 72  c0 27 00 21 00 01 00 00   ·····Our ·'·!····
```


Direct Discovery (mDNS Response) - Trace



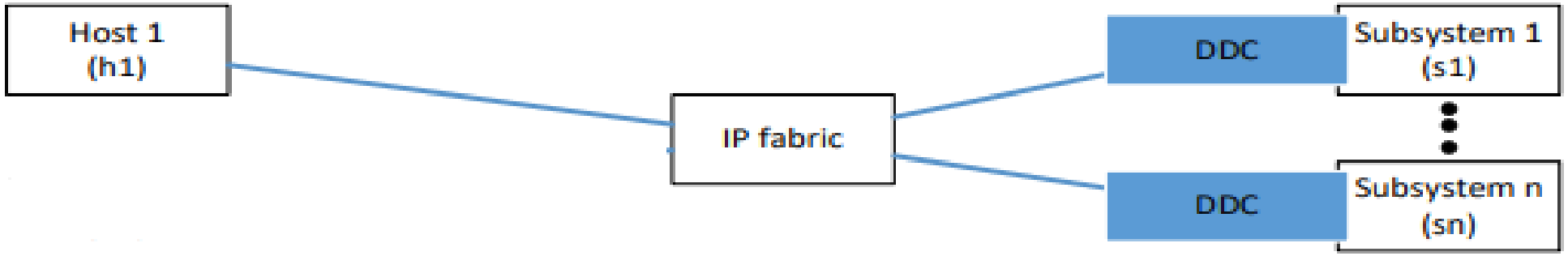
Manual vs Automated / Centralized Discovery at Scale

Discovery Methods	Manual	Automated - Direct/ Centralized
Steps Details	Host: IP and connect Storage : Provision Storage Host: Discover and connect	Host: Not required CDC : Configure Zoning (Opt.) Storage : Provision Storage
Number of steps	3 Ex: 1k systems = 3k steps	1 Ex: 1k systems = 1k steps

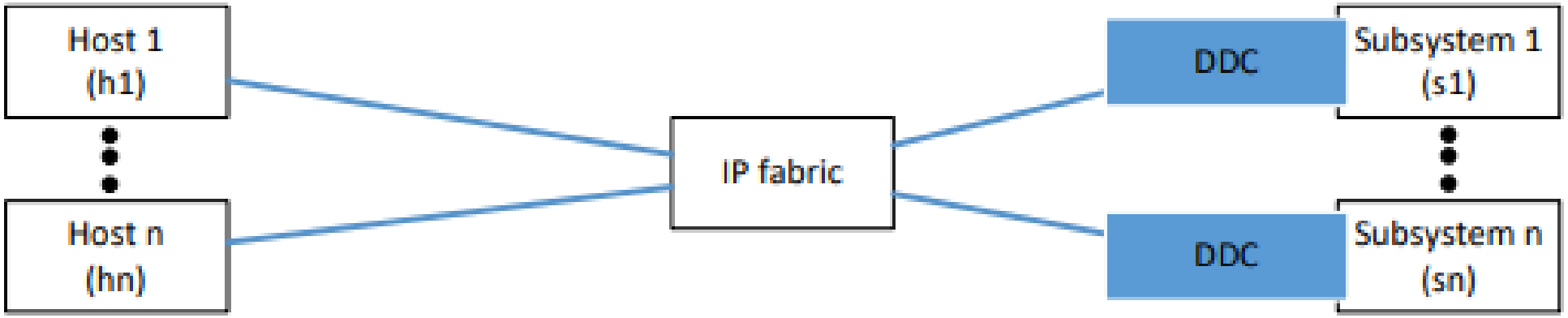
- Manual discovery non-viable for increase host / Storage subsystem
- Required 3x more steps in manual vs automated

Evaluation Demo

Scenario 1 – Single Host doing automated discovery (DDC) with multiple storage subsystems



Scenario 2 – Multiple Host doing automated discovery (DDC) with multiple storage subsystems



Continued.. On Scaled Environment

Test 2 | **Target 12 (64 Subsystems) X 18 Host**

```
----------------------------------------
Start of test 2
----------------------------------------

Number of Targets : 12 (64 Subsystems)
Number of Hosts : 18


[host_1] List of subsystems discovered and connected with automated discovery (nvme list-subsys)
        nqn.2023-01.nvmeof_target_7:subsystem0
        nqn.2023-01.nvmeof_target_1:subsystem6
                                :
                                :
                                :
host_18] List of subsystems discovered and connected with automated discovery (nvme list-subsys)
       nqn.2023-01.nvmeof_target_3:subsystem1
       nqn.2023-01.nvmeof_target_2:subsystem3
       nqn.2023-01.nvmeof_target_9:subsystem3
                                :
                                :
                                :


Average Time taken for automated discovery by each Host :: 31.544483343760174 seconds
```

Test 3 | **Target 12 (64 Subsystems) X 36 Host**

```
----------
Start of test 3
----------


Number of Targets : 12 (64 Subsystems)
Number of Hosts : 36

[host_1] List of subsystems discovered and connected with automated discovery (nvme list-subsys)
       nqn.2023-01.nvmeof_target_7:subsystem0
       nqn.2023-01.nvmeof_target_9:subsystem0
       nqn.2023-01.nvmeof_target_1:subsystem6
                        :
                        :
[host_36] List of subsystems discovered and connected with automated discovery (nvme list-subsys)
       nqn.2023-01.nvmeof_target_3:subsystem1
       nqn.2023-01.nvmeof_target_10:subsystem0
       nqn.2023-01.nvmeof_target_2:subsystem3
                        :
                        :


iverage Time taken for automated discovery by each Host :: 36.443747202555336 seconds
```


Observations / Recommendations (Scaled Environment)

Discovery Methods	DDC	CDC
Discovery Time	• Time increase's significantly • Lead to delays and inefficiencies	• Significantly reduce the time required for discovery • Minimizes the overall discovery time.
Network Congestion	• Congestion increase's due to high volume discovery traffic • Impact overall network performance	• Reduce discovery traffic due Centralized discovery • Better network performance (avoiding congestion)
Optimize to Use	• Pre-provision and careful network configuration would help	• Ensure properly configured to maintain better network perf & avoiding congestion
Tools : nvme-stas (Open Source) - help automate the discovery of instances, reducing manual config efforts		

Note: Single CDC instance will be challenging. However multiple CDC instance could be configured for HA

Summary

- In summary, while DDCs can be used in large-scale environments
  - Require careful planning
  - Potential network congestion with increased discovery time

- Scaling with CDCs with more number of nodes offers significant advantages
  - Reduced discovery time
  - Improved network performance
  - Simplified management.

These benefits make CDCs a preferred choice for large-scale deployments.

Source of Reference: NVM Express Org and TP8009 / TP8010

Thank You!

Please take a moment to rate this session.

Your feedback is important to us.


```
--------------------------------------------------------------------------------
Start of test 1
--------------------------------------------------------------------------------

Number of Targets : 12 (64 Subsystems)
Number of Hosts : 1


[host_1] List of subsystems discovered and connected with automated discovery (nv
        nqn.2023-01.nvmeof_target_1:subsystem3
        nqn.2023-01.nvmeof_target_7:subsystem0
        nqn.2023-01.nvmeof_target_1:subsystem6
        nqn.2023-01.nvmeof_target_3:subsystem3
        nqn.2023-01.nvmeof_target_2:subsystem1
        nqn.2023-01.nvmeof_target_1:subsystem0
        nqn.2023-01.nvmeof_target_3:subsystem0
        nqn.2023-01.nvmeof_target_4:subsystem4
        nqn.2023-01.nvmeof_target_9:subsystem0
        nqn.2023-01.nvmeof_target_11:subsystem3
        nqn.2023-01.nvmeof_target_1:subsystem2
        nqn.2023-01.nvmeof_target_3:subsystem2
        nqn.2023-01.nvmeof_target_5:subsystem3
        nqn.2023-01.nvmeof_target_2:subsystem0
        nqn.2023-01.nvmeof_target_1:subsystem4
                        :
        nqn.2023-01.nvmeof_target_3:subsystem5
        nqn.2023-01.nvmeof_target_6:subsystem1
        nqn.2023-01.nvmeof_target_1:subsystem5
        nqn.2023-01.nvmeof_target_2:subsystem4
        nqn.2023-01.nvmeof_target_6:subsystem2
        nqn.2023-01.nvmeof_target_2:subsystem5
        nqn.2023-01.nvmeof_target_4:subsystem3
        nqn.2023-01.nvmeof_target_12:subsystem2
        nqn.2023-01.nvmeof_target_2:subsystem3
        nqn.2023-01.nvmeof_target_3:subsystem7
        nqn.2023-01.nvmeof_target_10:subsystem2
        nqn.2023-01.nvmeof_target_12:subsystem0
        nqn.2023-01.nvmeof_target_9:subsystem1
        nqn.2023-01.nvmeof_target_4:subsystem2

Total time taken for automated discovery with 1 Host for 12 Targets (64 subsystems) : 25.09260058403015
```

```
-------------------------------------------
Start of test 2
-------------------------------------------

Number of Targets : 12 (64 Subsystems)
Number of Hosts : 18

[host_1] List of subsystems discovered and connected with automated discovery (nvme list-subsys)
        nqn.2023-01.nvmeof_target_7:subsystem0
        nqn.2023-01.nvmeof_target_1:subsystem6
                                :
host_18] List of subsystems discovered and connected with automated discovery (nvme list-subsys)
        nqn.2023-01.nvmeof_target_3:subsystem1
        nqn.2023-01.nvmeof_target_2:subsystem3
        nqn.2023-01.nvmeof_target_9:subsystem3
                                :
Total time taken for automated discovery with 18 Host for 12 Targets (64 subsystems) : 567.8007001876831
Average Time taken for automated discovery by each Host :: 31.544483343760174 seconds
```

```
----------  ----------------------------
Start of test 3
--------------------------------------

Number of Targets : 12 (64 Subsystems)
Number of Hosts : 36

[host_1] List of subsystems discovered and connected with automated discovery (nvme list-subsys)
        nqn.2023-01.nvmeof_target_7:subsystem0
        nqn.2023-01.nvmeof_target_9:subsystem0
        nqn.2023-01.nvmeof_target_1:subsystem6
                        :
[host_36] List of subsystems discovered and connected with automated discovery (nvme list-subsys)
        nqn.2023-01.nvmeof_target_3:subsystem1
        nqn.2023-01.nvmeof_target_10:subsystem0
        nqn.2023-01.nvmeof_target_2:subsystem3
                        :

Total time taken for automated discovery with 36 Host for 12 Targets (64 subsystems) : 1311.9748992919922
Average Time taken for automated discovery by each Host :: 36.443747202555336 seconds
```

