STORAGE DEVELOPER CONFERENCE

SDC22 | Fremont, CA
September 12-15, 2022

BY Developers FOR Developers

A SNIA Event

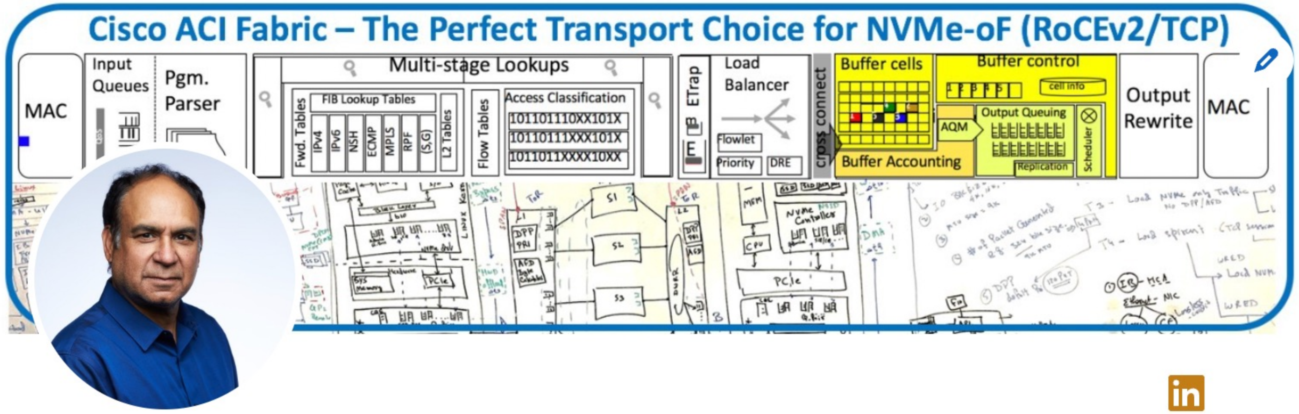
NVMe/FC or NVMe/TCP

In-depth Packet & Flow Level Comparison

Presented by: Kamal Bakshi

Director Technical Marketing, Cisco Systems

About me.. (I help Customers in adopting “New Technologies”)



Kamal Bakshi
Director Cisco -Technology Evangelist, Distinguished Speaker(Cisco)
San Jose, California, United States

Cisco
1996: (CCIE #2316) Cisco Certified Internetwork Expert

Technology Adoption Mantra (5A)

1-Awareness	(heard of it/marketing)
2-Advantages	(value proposition/match)
3-Attitude	(liking/personal experience)
4-Adoption	(migration process/non disruptive)
5-Acceptance	(easy to maintain/ROI)

Director Storage Transport Solutions
Cisco
Nov 2020 - Present · 1 yr 11 mos

-MDS/FC Storage Technical Marketing group
-NVMe-RoCEv2/TCP IP storage solutions

NVMe Promoters Group Board Member (Cisco)
NVM Express
Jan 2021 - Jan 2022 · 1 yr 1 mo

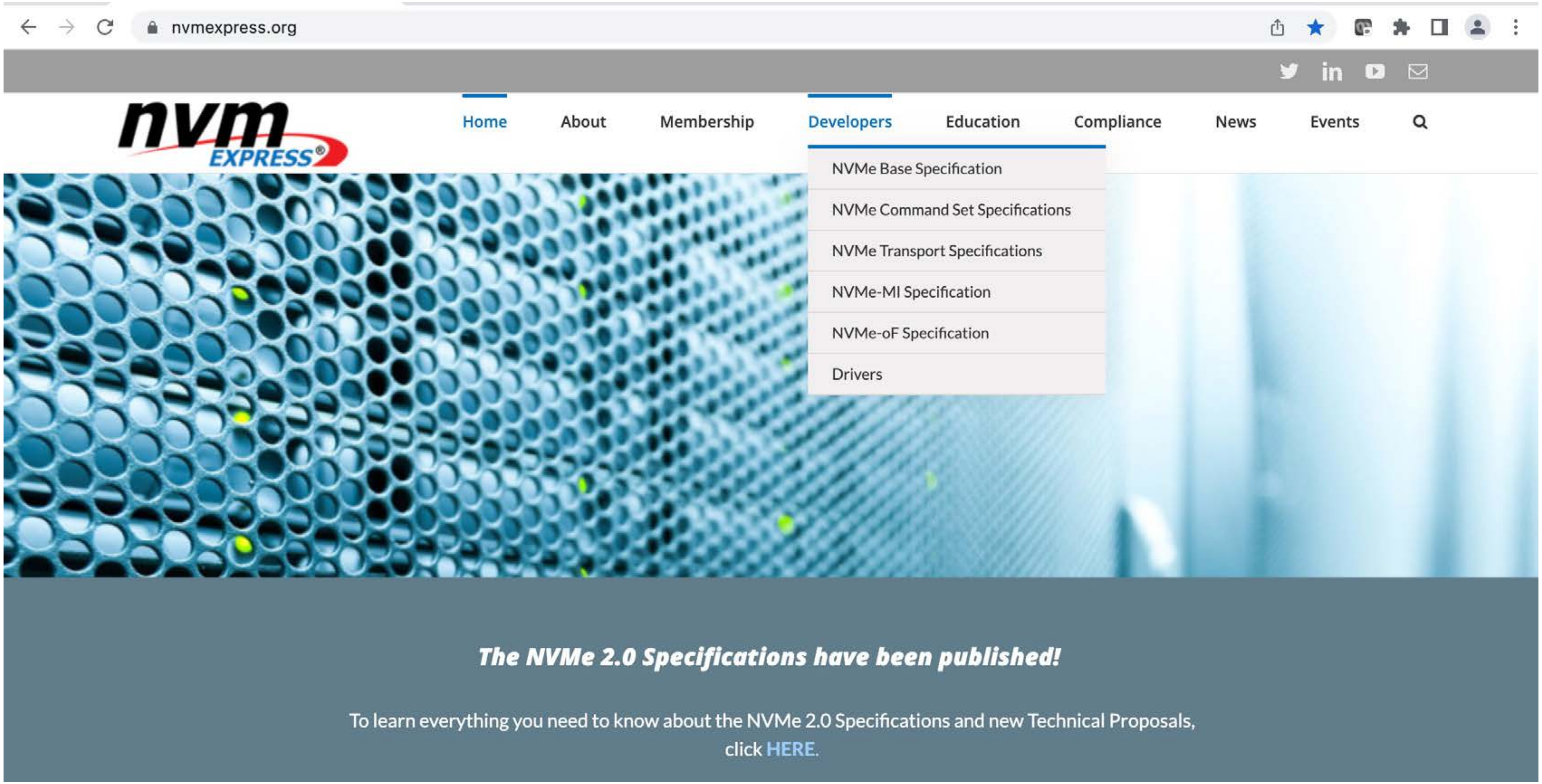
Principal Engineer
Cisco
Feb 2016 - Jan 2020 · 4 yrs
San Jose, CA, USA

Datacenter Technology Evangelist
-Intent Based DC Networking
-Next Generation DC Solutions Lab
-NVMe-oF Transport
-Sales Enablement / Technology Adoption

Director Technical Marketing
Cisco Systems
Oct 1994 - Jul 2011 · 16 yrs 10 mos
San Jose, CA

(Data Center Business Unit Group)
Managed multiple technical functions. Storage TME team, Competitive team, DC architecture showcase center, and tools development. Brought multiple internal BUs & partners together (EMC, Netapp, VMware) to stage Cisco DC3.0 architecture.

Download the latest NVMe specifications 2.0



Kamal Bakshi, Cisco

NVMe Adoption

- Today (2022) total NVMe market size is over $80 Billion
- By 2030 NVMe market will exceed $175 Billion (CAGR 28%)
- Nearly ALL servers shipping today support NVMe drives
- All enterprise networking adapters sold today are NVMe-oF capable
- Over 80% of the All Flash Storage Arrays are based on NVMe
- By 2026 SSD/flash will be cheaper than enterprise HDD/disks

Sources: G2M Research, Wikibon, & others

Future-Proof your IT Infrastructure by upgrading to NVMe today

Kamal Bakshi, Cisco

What is NVMe/oF?

What problem are
we trying to solve?

Why should I care?

What is the value
proposition & advantages
of this technology?

What to watch out for?

What are the Do's & Don'ts
for best experience?

Reap Benefits!

Better performance,
Easy to maintain,
High ROI

KNOWLEDGE IS THE KEY TO SUCCESS

- **PCIe4.0** 1053 pages
- **NVMe-PCIe** 37 pages
- **NVMe-Base** 455 pages
- **NVMe-oF** 84 pages
- **NVMe-CMD** 105 pages
- **NVMe-TCP** 35 pages
- **NVMe-RDMA** 16 pages
- **FC-FCP** 117 pages
- **NVMe-FC2** 106 pages
- **FC-LS** 253 pages
- **FC-FS** 504 pages
- **FC-GS** 430 pages
- **NVMe-CDC** TP8010 82pages
- **NVMe-ZNS** 46 pages
- **NVMe-KV** 23 pages
- **NVMe-MI** 161 pages

(…but over 3000 pages!)

Kamal Bakshi, Cisco

Agenda

1. Data Center Storage Architecture
2. NVMe/FC Architecture
3. NVMe/TCP Architecture

Appendix

- NVMe Evolution
- NVMe/PCIe Architecture
- NVMe/FC Packets
- NVMe/RoCEv2 Architecture
- NVMe Advanced Features

https://www.ciscolive.com/on-demand/on-demand-library.html?search=kamal%20bakshi#/

(Cisco Live video session that covers the above Appendix topics)

Kamal Bakshi, Cisco

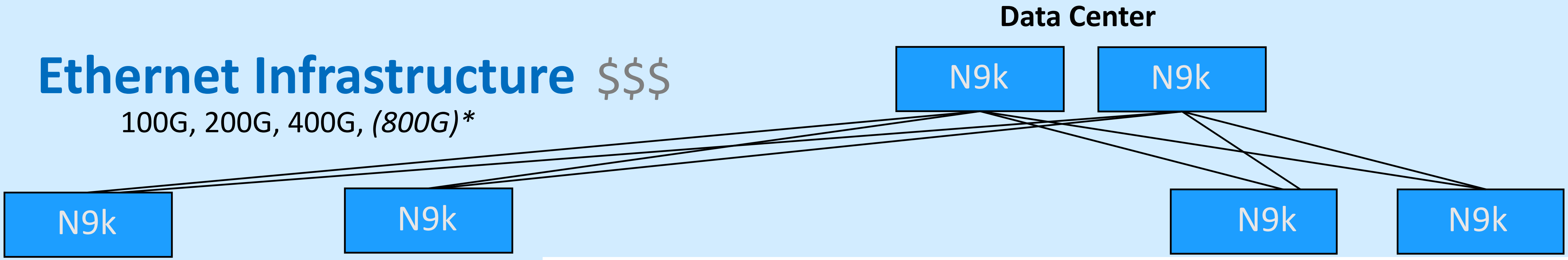
DC Storage Infrastructure

Kamal Bakshi, Cisco

Data centers are fast adopting 100G/400G Ethernet

Ethernet Infrastructure $$$

100G, 200G, 400G, *(800G)**



BUSINESS > MARKET RESEARCH

800G data center switch ports to top 400G by 2025: Dell'Oro

The availability of 800G optics and 25.6T chips has already created a demand for 800G switch ports, according to Dell'Oro.

Stephen Hardy

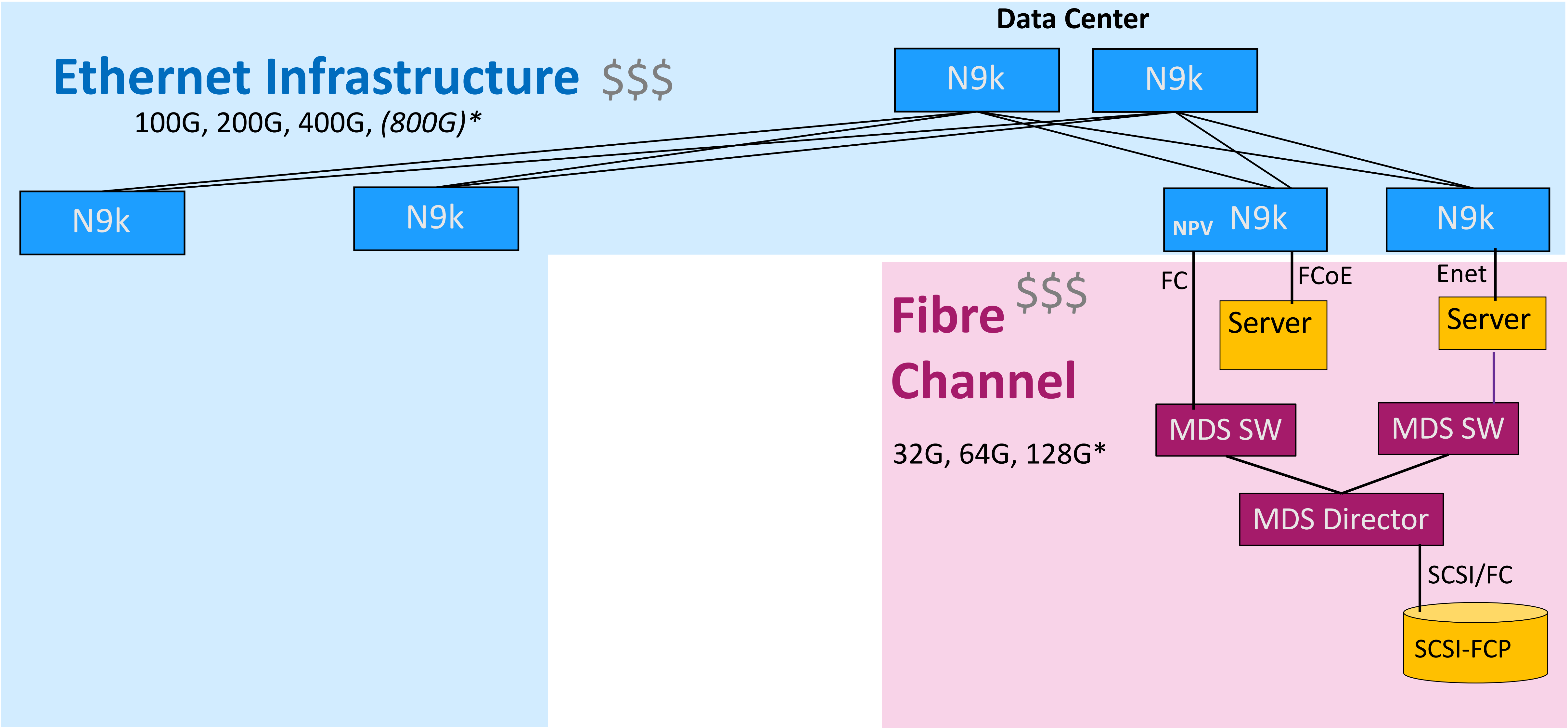
July 19, 2022

Quick Google Search

N9K -Cisco Nexus Ethernet Switch

Kamal Bakshi, Cisco

1994: Traditional Storage Infrastructure is Fibre Channel

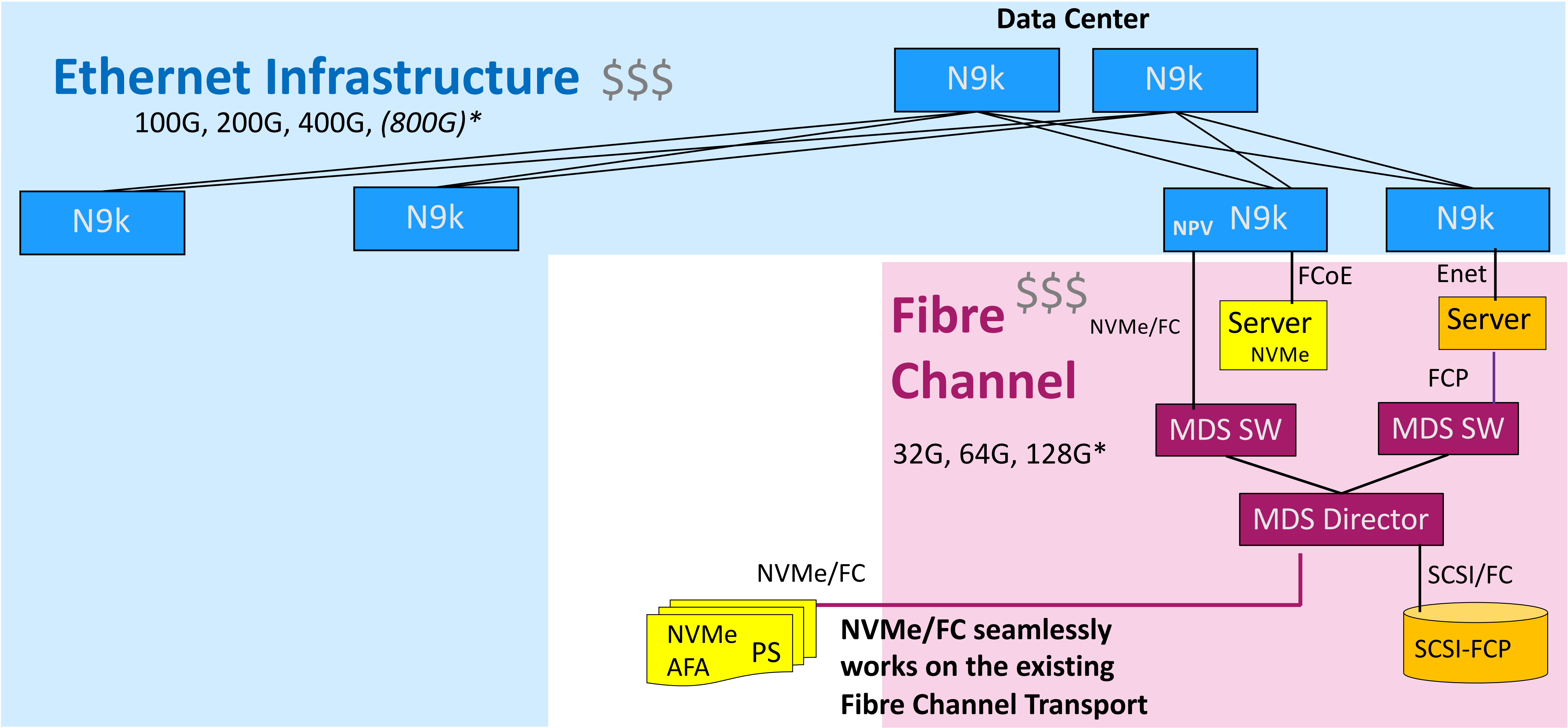


N9K -Cisco Nexus Ethernet Switch

MDS -Cisco Fibre Channel Switch

Kamal Bakshi, Cisco

2017: New NVMe/FC technology works on existing FC



N9K -Cisco Nexus Ethernet Switch

MDS -Cisco Fibre Channel Switch

PS -Dell PowerStore All-Flash Storage

Kamal Bakshi, Cisco

Is Fibre Channel Meeting Your Storage Criteria?

Check List

1-Security
- Dedicated Fabric, FC-SP, Hardware Zoning

2-High Availability
- Dual Fabric A/B, Multipathing, Lossless/B2B

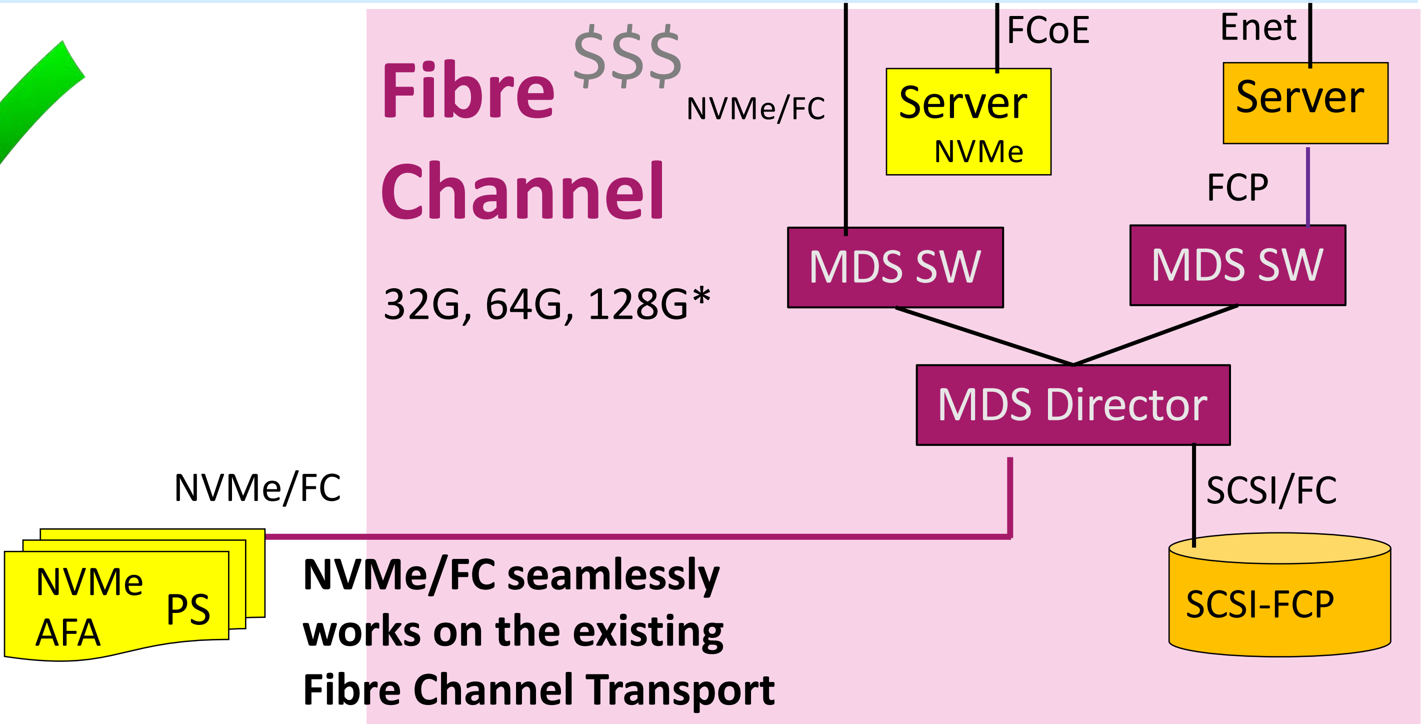
3-Built in SAN Fabric Services/Automation
- Addressing, Directory Services, Name Server, Change Notifications

4-High Performance
- Congestion Control/Slow Drain (B2B, DIRL, FPIN),
- Bandwidth, Throughput, IOPS, Latency, Zero Copy

5-Management
- Storage Troubleshooting, Topology, Storage Analytics

6-Scalability
- Fabric Architecture, Max. Switches/domain, Virtual SAN

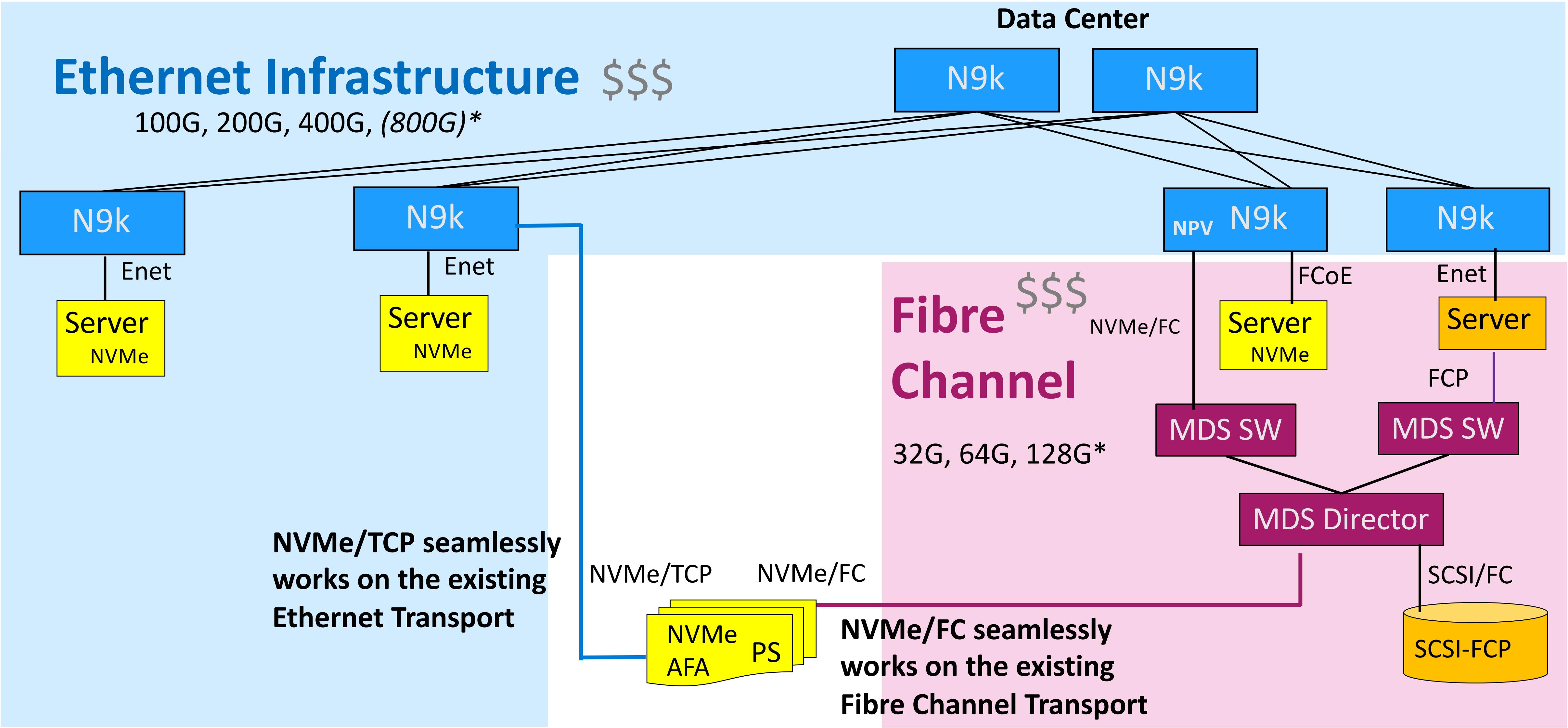
7-Ecosystem/Interop/Certification
- Driver support, Adapters, Optics, Storage arrays, Switches



MDS -Cisco Fibre Channel Switch
PS -Dell PowerStore All-Flash Storage

Kamal Bakshi, Cisco

2020: NVMe/TCP transport binding specification released



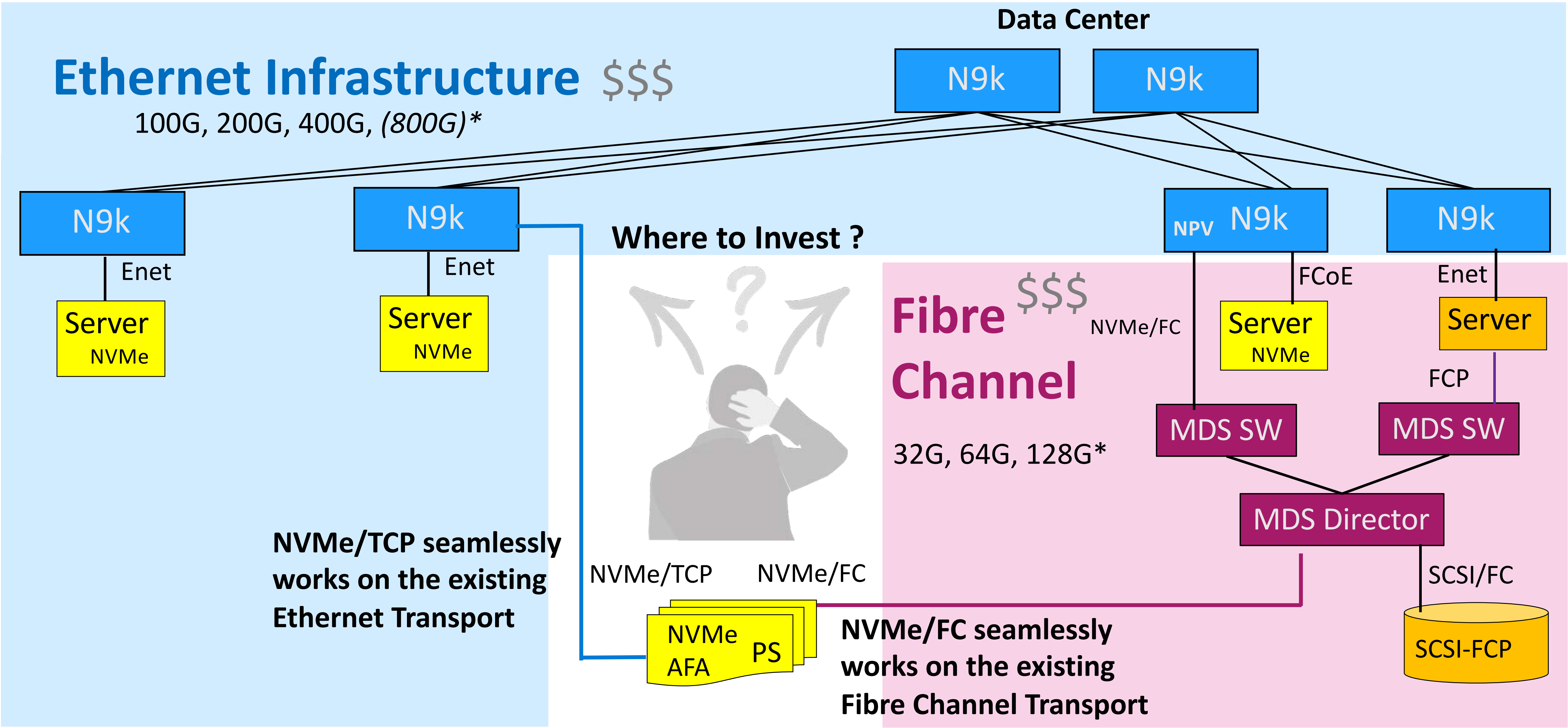
N9K -Cisco Nexus Ethernet Switch

MDS -Cisco Fibre Channel Switch

PS -Dell PowerStore All-Flash Storage

Kamal Bakshi, Cisco

Q: Does NVMe/TCP offers better Price/Performance than NVMe/FC?



N9K -Cisco Nexus Ethernet Switch

MDS -Cisco Fibre Channel Switch

PS -Dell PowerStore All-Flash Storage

Kamal Bakshi, Cisco

Can NVMe/TCP provide FC SAN like services?

1-Security

2-High Availability

3-Built in SAN Fabric Services/Automation

4-High Performance

5-Management

6-Scalability

7-Ecosystem/Interop/Certification

Kamal Bakshi, Cisco

2022: NVMe/TCP is not plug & play as compared to NVMe/FC

1-Security

-Dedicated Fabric (Y), FC-SP (TLS1.3), Hardware Zoning (SDN)

2-High Availability

-Dual Fabric A/B (Y), Multipathing (Y), Lossless/B2B (ECN/PFC)

3-Built in SAN Fabric Services/Automation

-Addressing (SDN), Directory Services, Name Server, Change Notifications (CDC -TP8009/10)

4-High Performance

-Congestion Control/Slow Drain (DIRL, FPIN) (TCP Congestion Control Methods ?)

-Bandwidth/Throughput/IOPS (100G/400G), Latency (Smart Buffering), Zero Copy (DPU?)

5-Management

-Storage Troubleshooting/Topology (Y), Storage Analytics (SOC Analytics ?, HBA Analytics)

6-Scalability

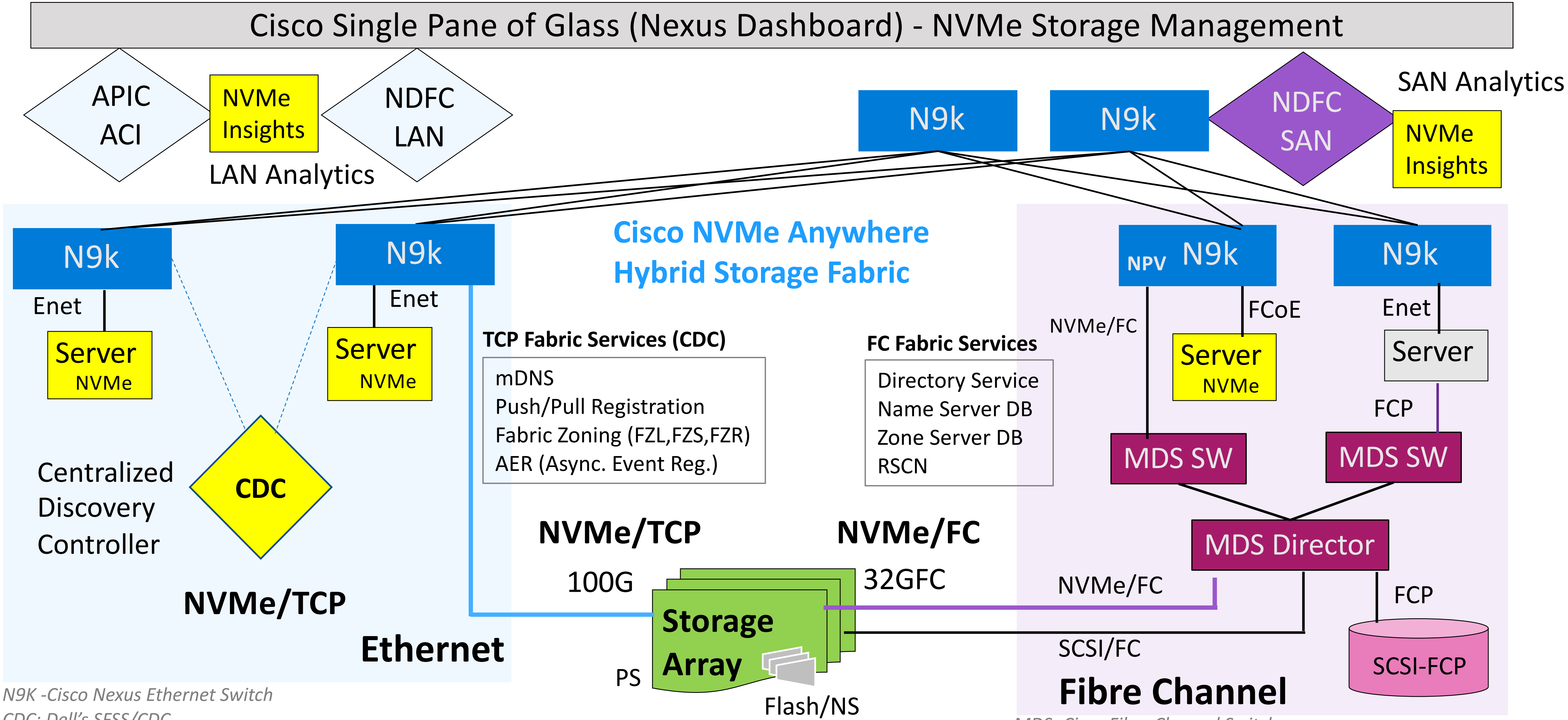
-Fabric Architecture, Max. Switches/domain, Virtual SAN (TCP is highly scalable)

7-Ecosystem/Interop/Certification (TBD: In pilot testing)

-Driver support (Linux, ESXi, Windows), CDC, Adapters, Optics, Storage arrays, Switches

Kamal Bakshi, Cisco

With NVMe you can take Advantage of both (FC & Enet) Infrastructure Investments!

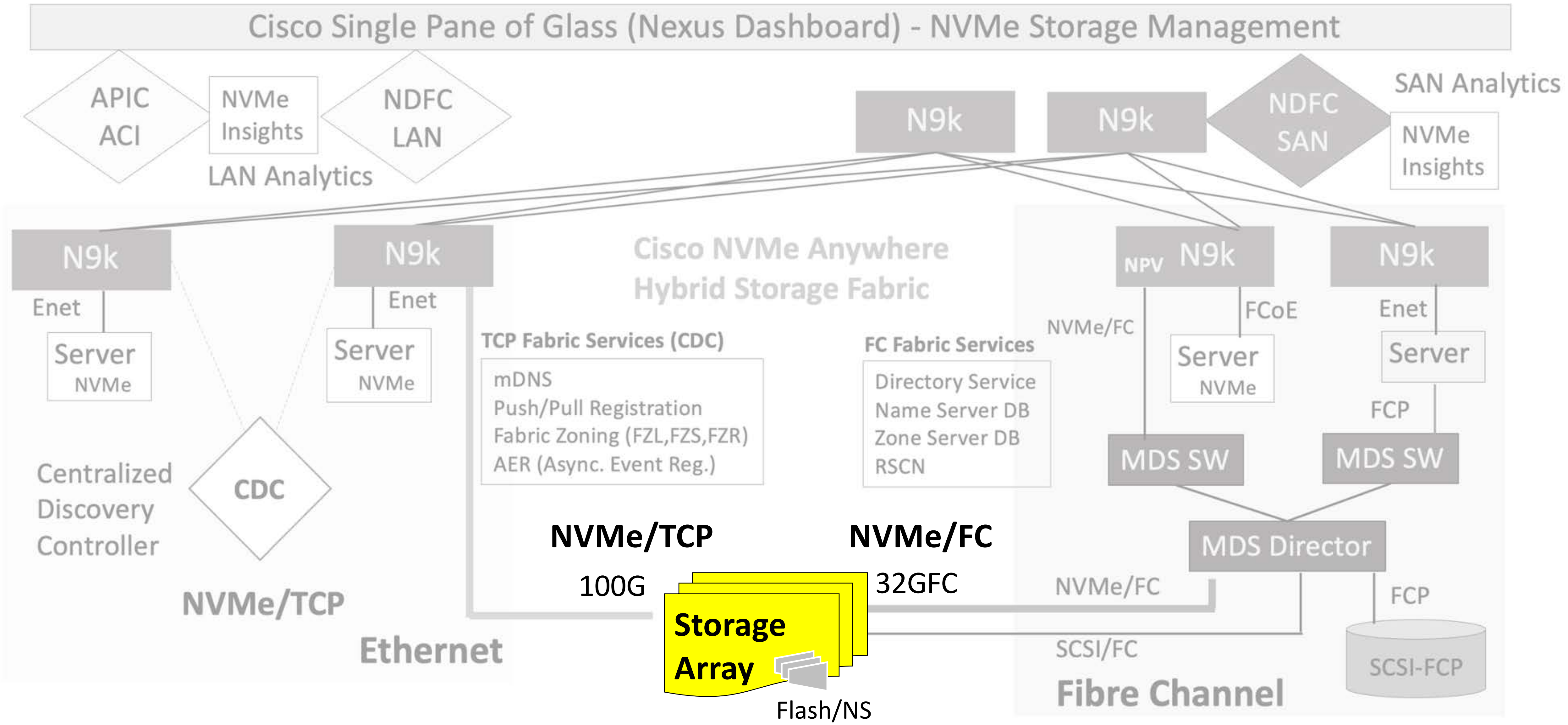


N9K -Cisco Nexus Ethernet Switch
CDC: Dell's SFSS/CDC

MDS -Cisco Fibre Channel Switch
PS -Dell PowerStore All-Flash Storage

Kamal Bakshi, Cisco

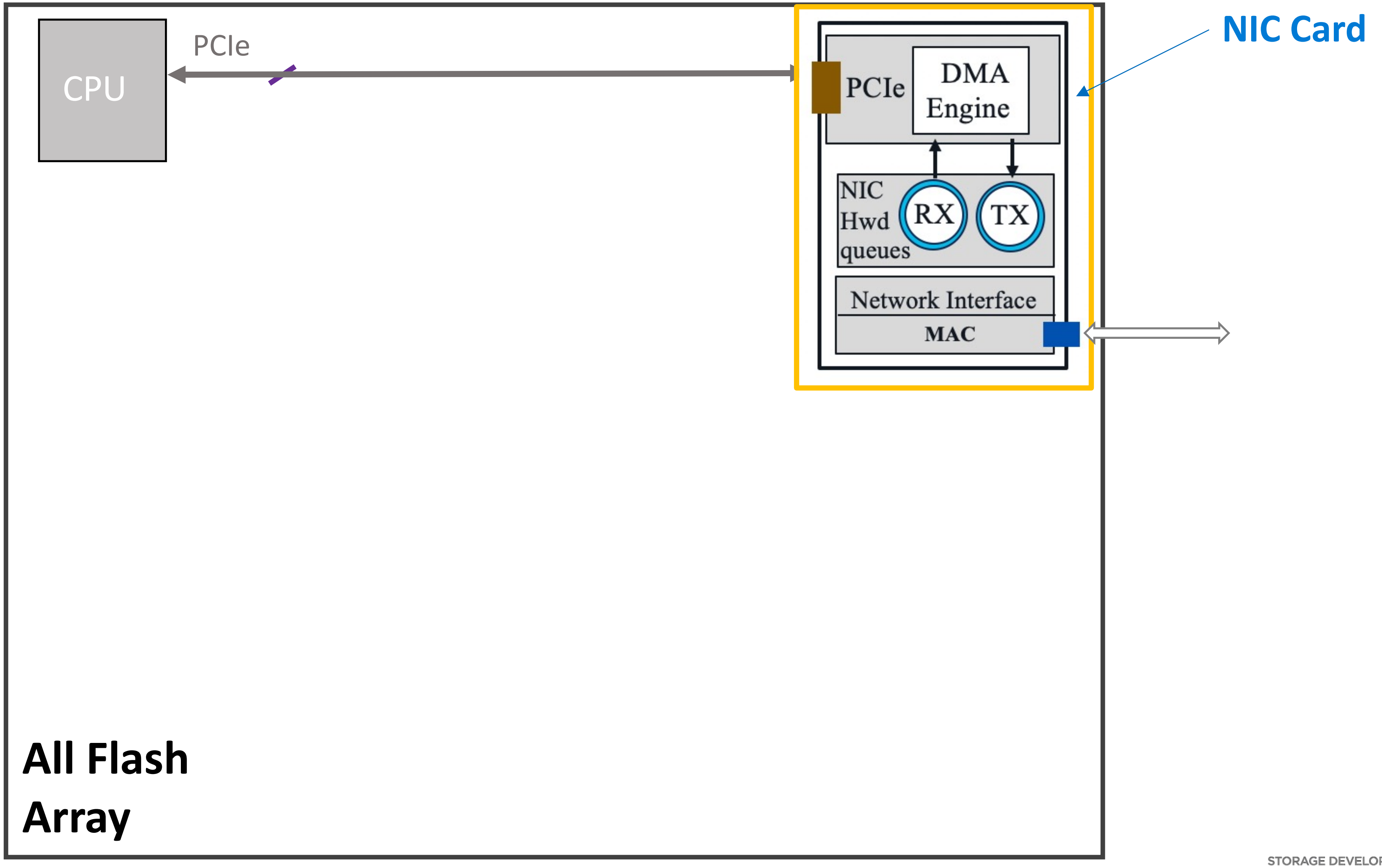
NVMe Storage Architecture



Kamal Bakshi, Cisco

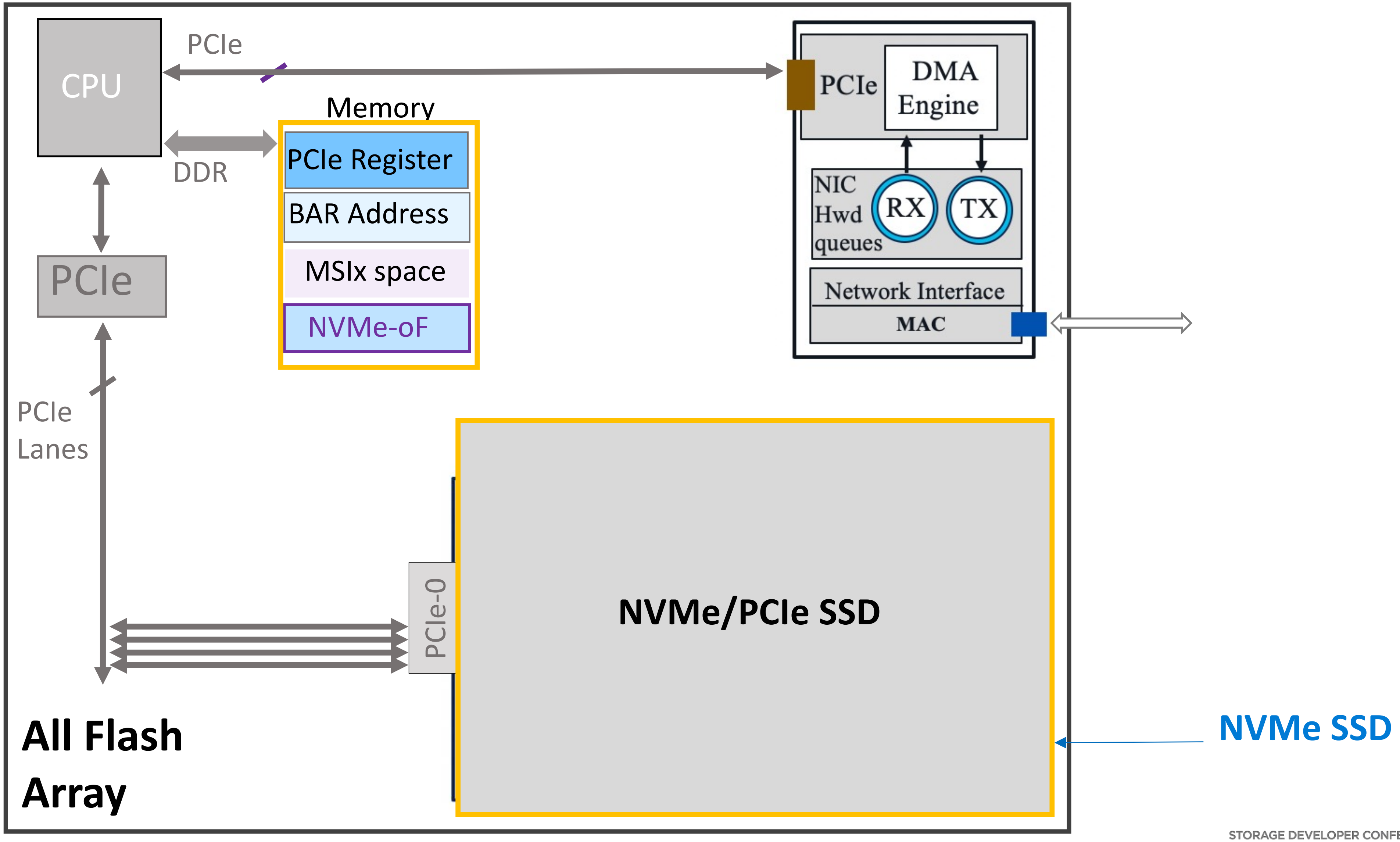
STORAGE DEVELOPER CONFERENCE SDC 22

NVMe Storage Architecture (Enet NIC)



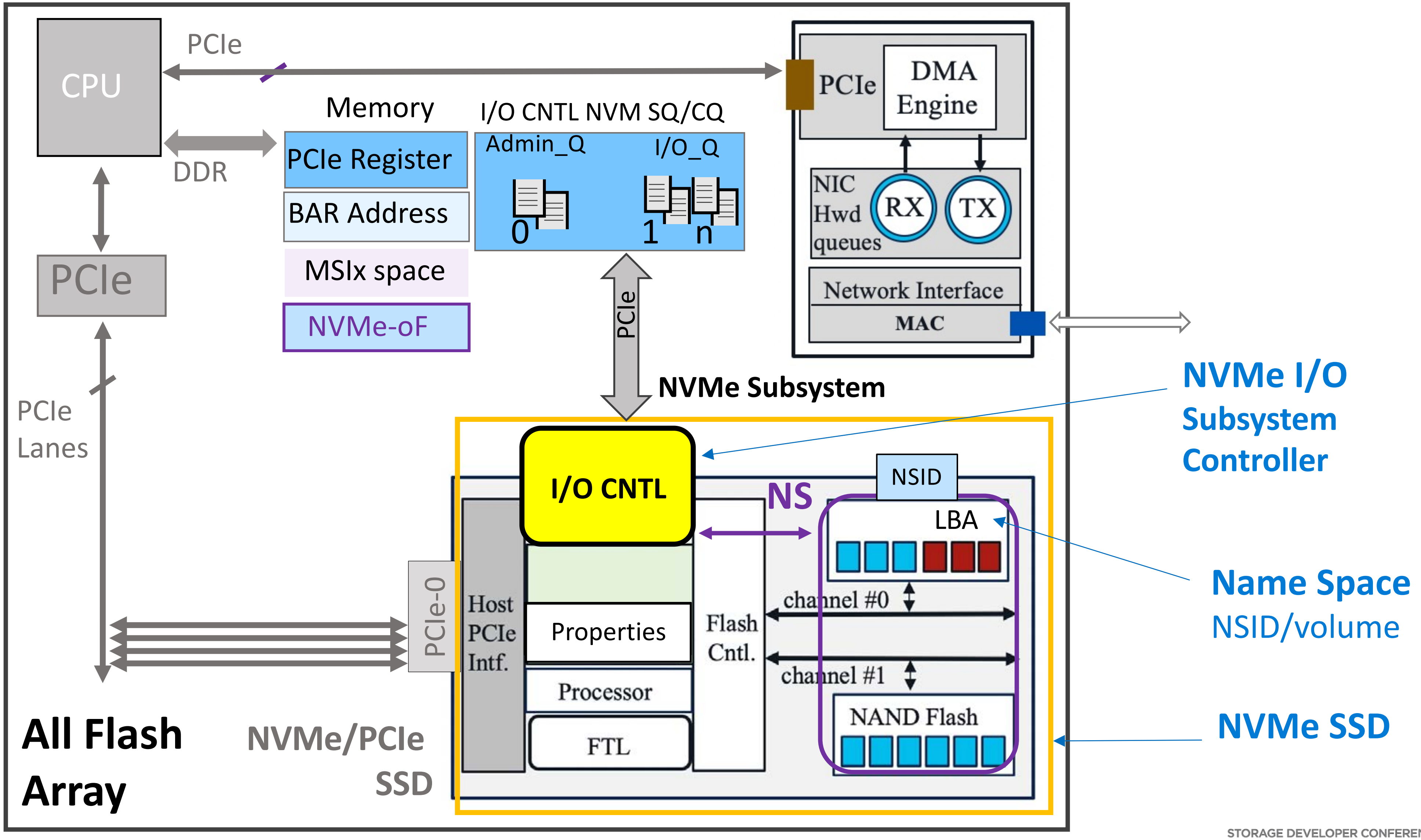
Kamal Bakshi, Cisco

NVMe Storage Architecture (PCIe SSD)



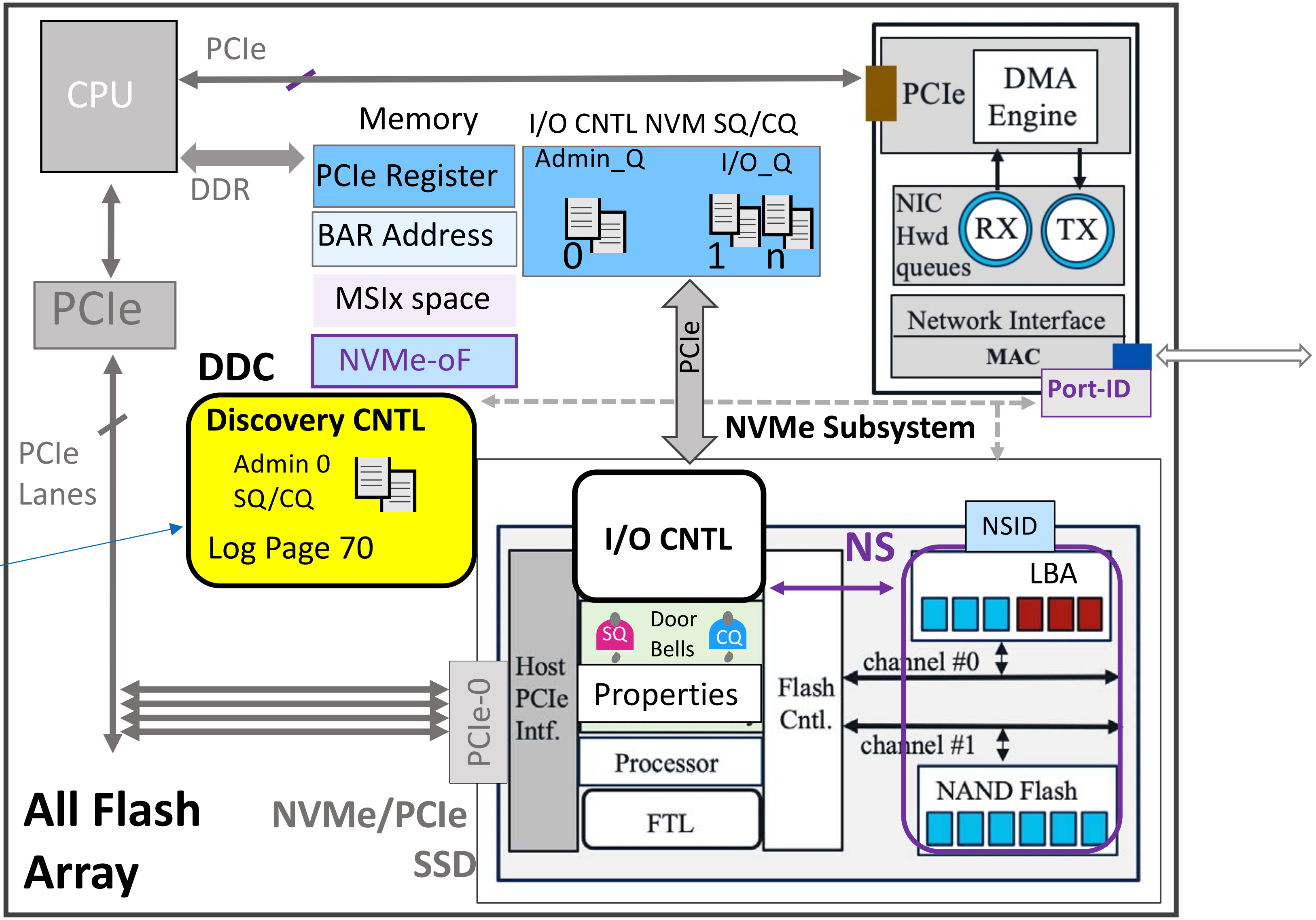
Kamal Bakshi, Cisco

NVMe Storage Architecture (I/O Controller)



Kamal Bakshi, Cisco

NVMe Storage Architecture (Discovery Controller/DDC)

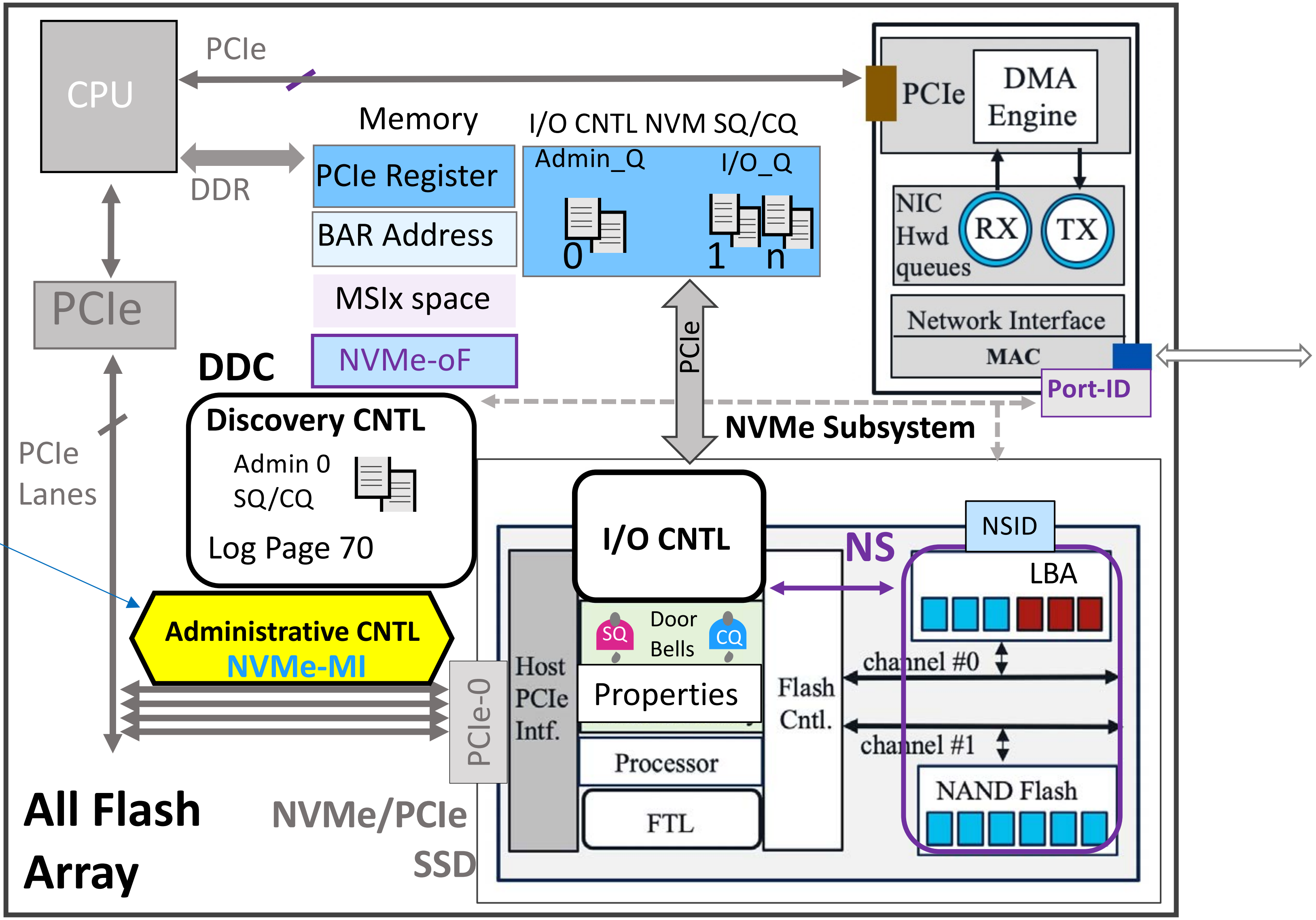


DDC
Direct Discovery
Controller (no I/O_Q)

Kamal Bakshi, Cisco

NVMe Storage Architecture (Administrative Controller)

Administrative Controller

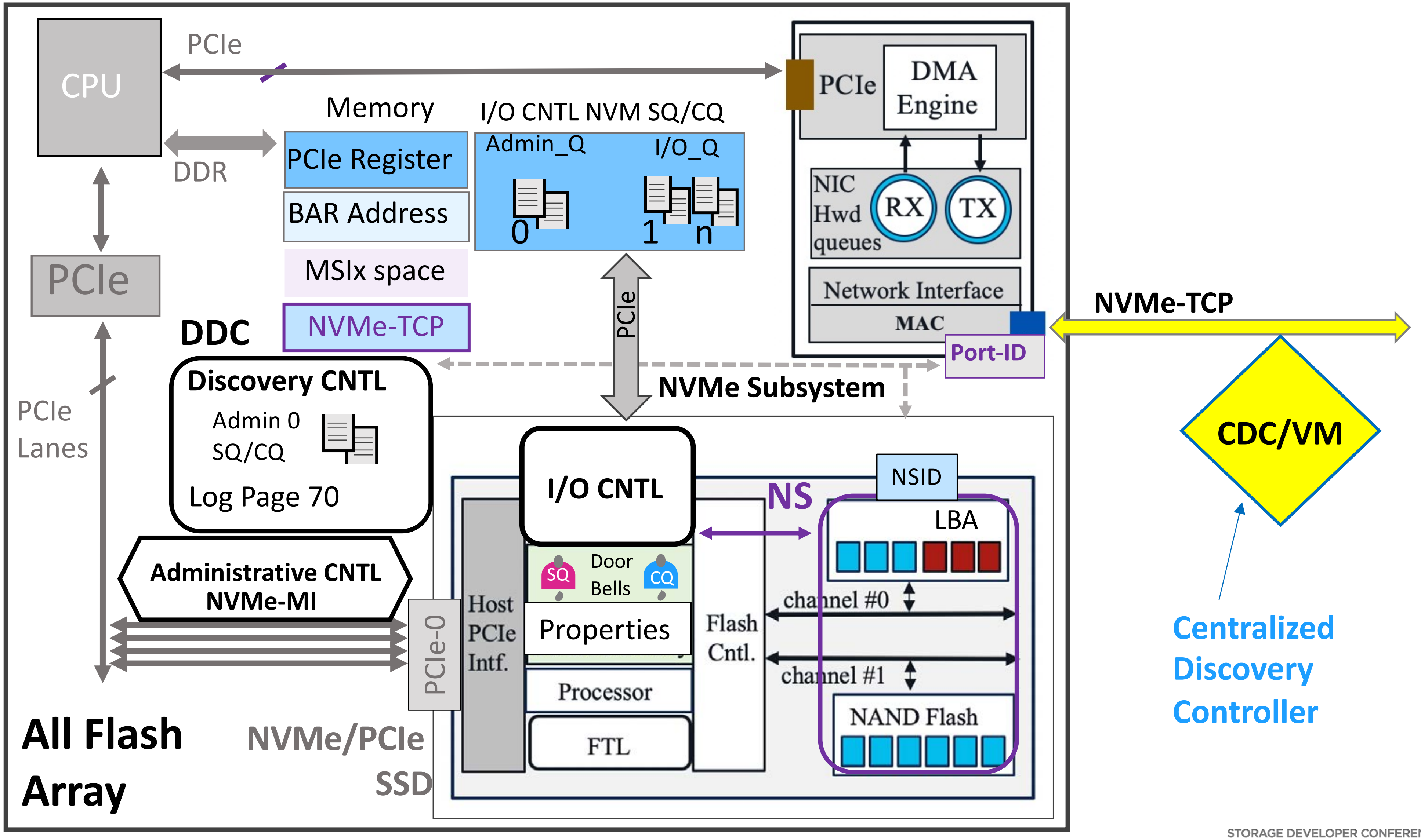


 |

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

NVMe/TCP Storage Architecture (CDC Controller)



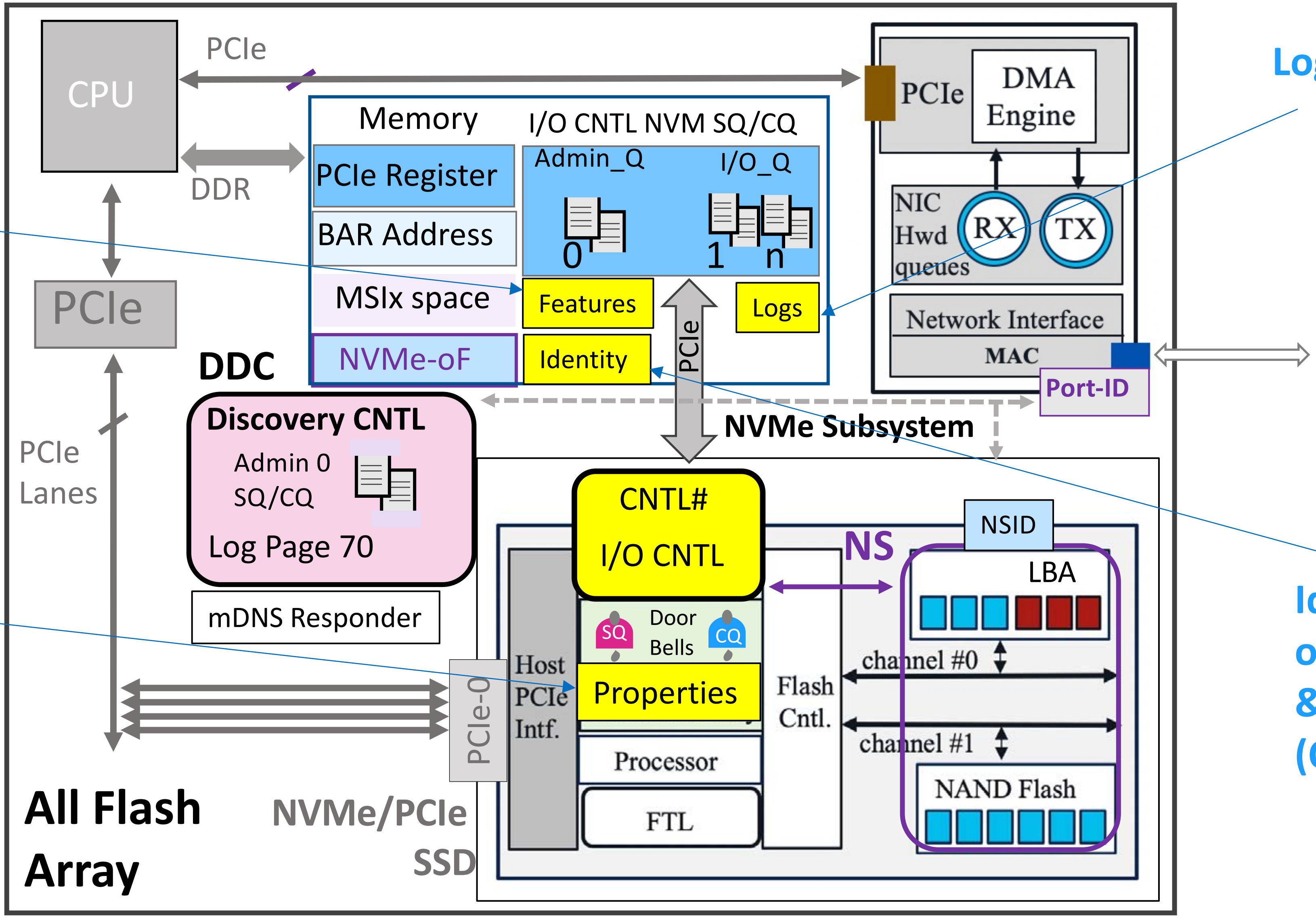
 |

Kamal Bakshi, Cisco

NVMe Storage Architecture (Information)

Tunable Features

Properties are on the controller's memory space



Log Pages, ID#

Identification of Controller & NameSpace (CNS)

Kamal Bakshi, Cisco

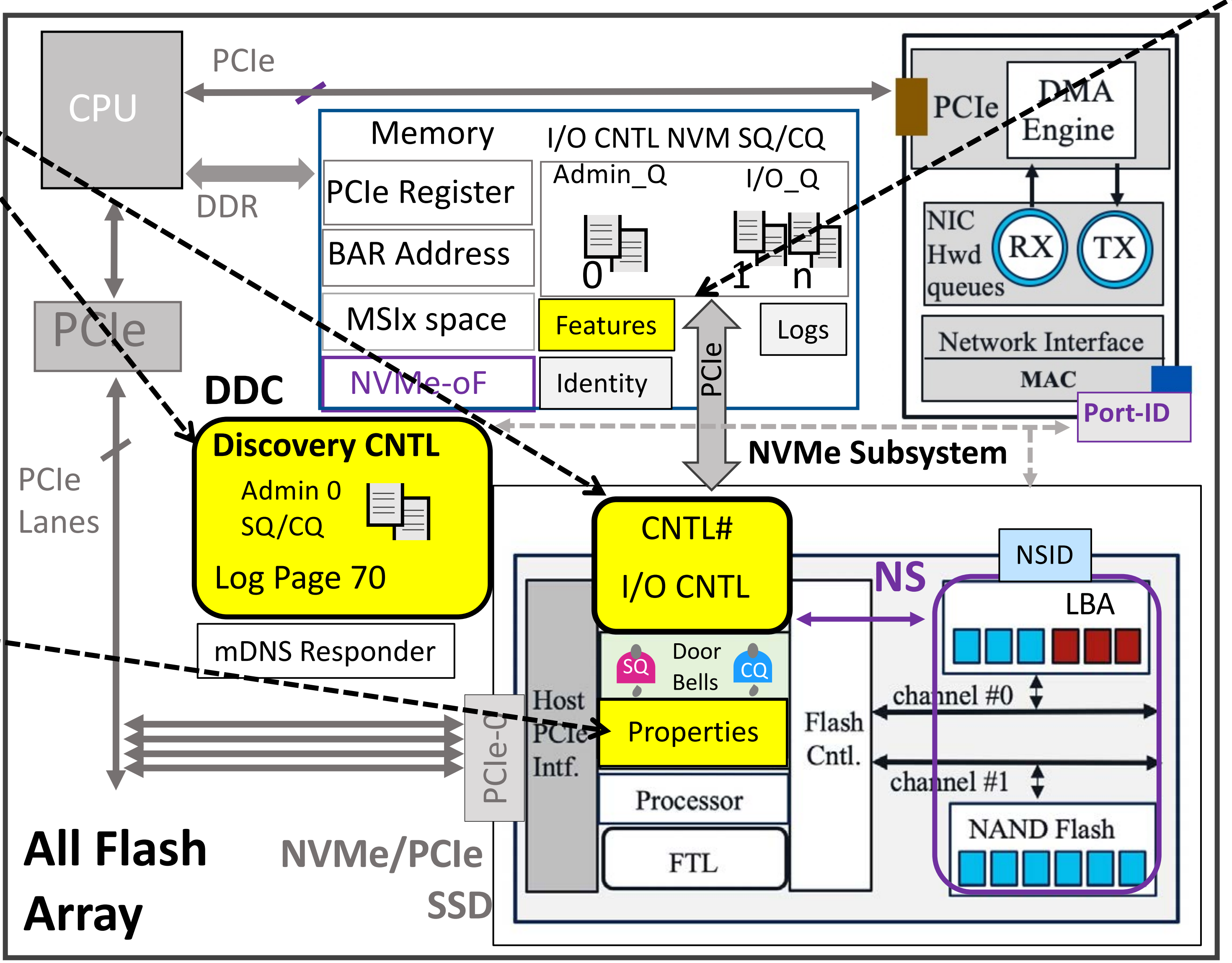
NVMe Storage Architecture (Properties, Features)

Controller Type

- Discovery Controller
- I/O Controller
- Administrative Controller

Properties (Get/Set)

00-07 Controller Capability
08-0B Version
14-17 Controller Configuration
1C-1F Controller Status
20-23 NVM Sub. Reset
F00-FFF Command Set Specific



Feature ID (Get/Set)

01 Arbitration
02 Power Mgmt.
04 Temperature Threshold
06 Volatile Write Cache
07 Number of Queues
08 Interrupt Coalescing
0B Async. Event Config.
0C Autonomous Power Trans.
0D Host Memory Buffer
0E Timestamp
0F Keep Alive Timer
10 Host Controlled Thermal Mgmt.
11 Non-Operational Power Transition
12 Read Recovery Level Config
13 Predictable Latency Mode Cfg.
14 Predictable Latency Mode window
16 Host Behavior Support
17 Sanitize Config
18 Endurance Group Event Cfg.
19 I/O Command Set Profile
20 Key Value Command set
7D Enhanced Controller Metadata
7E Controller Metadata
7F Namespace Metadata
80 Software Progress Marker
81 Host Identifier
82 Reservation Notification mask
83 Reservation Persistence
84 Namespace Write Protection Cfg.

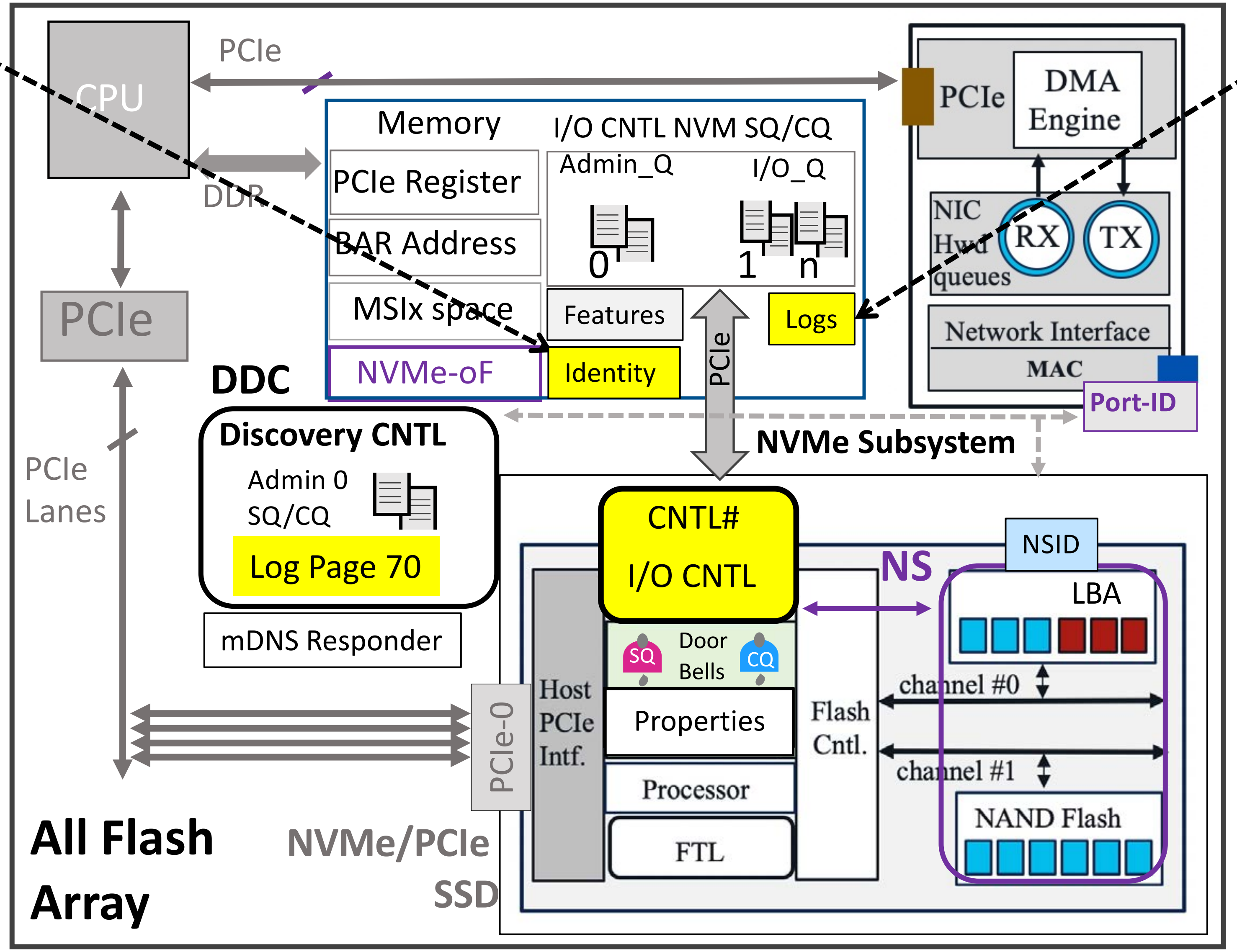
Kamal Bakshi, Cisco

NVMe Storage Architecture (Log Pages buffer, Identify/CNS buffer)

Identify (Controller/NS)
CNS value

Active Namespace Mgmt.
00 Identify Namespace for NSID
01 Identify Controller
02 Active Namespace ID List
03 NS ID descriptor for NSID
04 NVM Sets List
05 I/O command set / NSID
06 I/O command set /Controller
07 Active NS ID List (I/O cmd. set)
08 I/O cmd set Independent

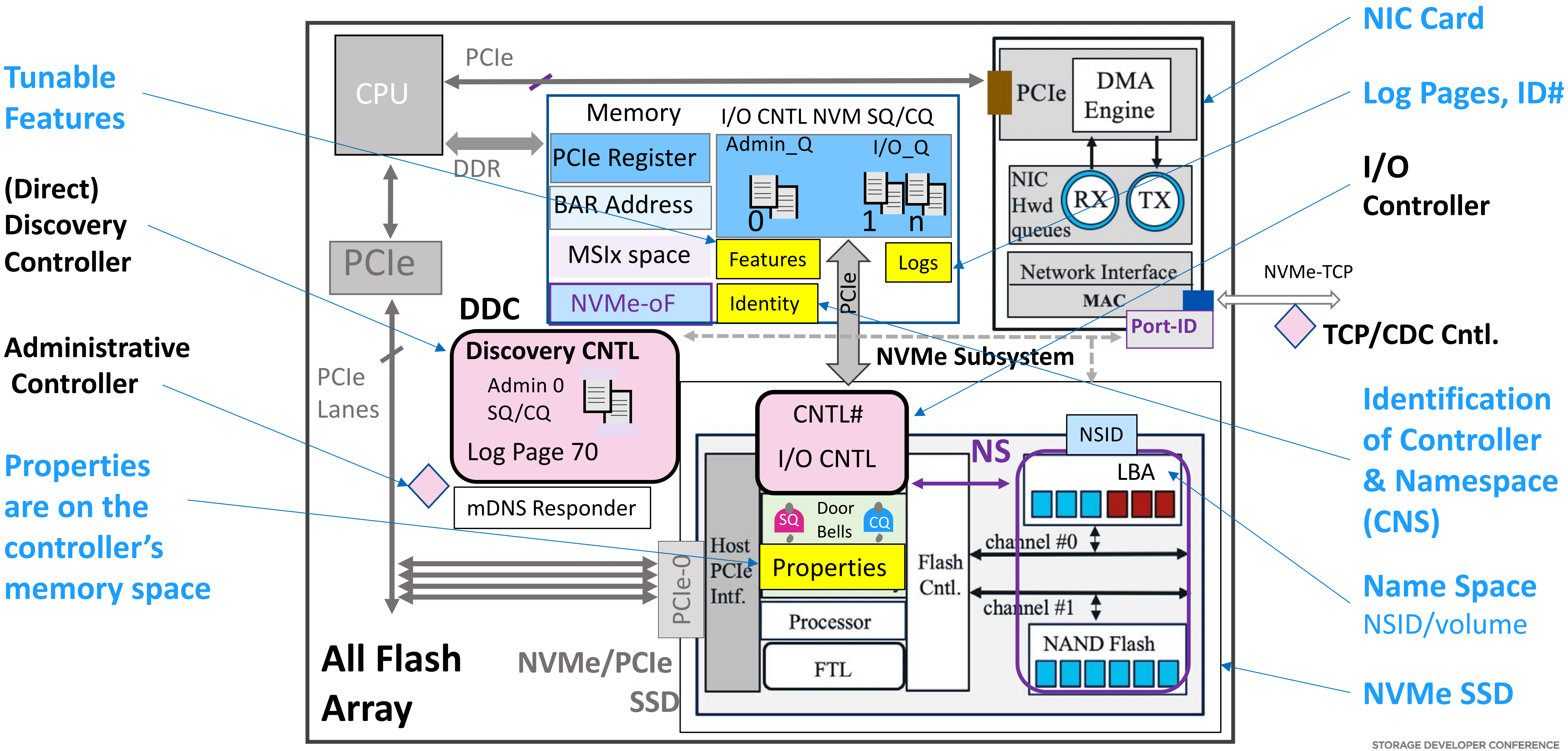
Controller NS Mgmt.
10 Allocated Namespace ID List
11 Identify NS / NSID
12 Controller List attached to NSID
13 Controller Lists in NVM Subsys.
14 Primary Controller Capabilities
15 Secondary Controller List
16 Namespace Granularity List
17 UUID List return to Host
18 Domain List
19 Endurance Group List
1A I/O command set / NSID
1B I/O command set /Identify NS
1C I/O command set data structure



Log Page ID
00 Supported Log Pages
01 Error Information
02 Smart Health Information
03 Firmware Slot Information
04 Changed Namespace List
05 Commands Supported
06 Device Self-test
07 Telemetry Host-Initiated
08 Telemetry Controller-Initiated
09 Endurance Group Information
0A Predictable Latency /NVMe set
0B Predicable Latency Event
0C Asymmetric Namespace Access
0D Persistent Event Log
0F Endurance Group Event
10 Media Unit Status
11 Supported Capacity Cfg. List
12 Feature Identifiers Supported
13 NMVe-MI Commands Supported
14 Command & Feature Lockdown
15 Boot Partition
16 Rotational Media Information
70 Discovery
71 Host Discovery
80 Reservation Notification
81 Sanitize Status

Kamal Bakshi, Cisco

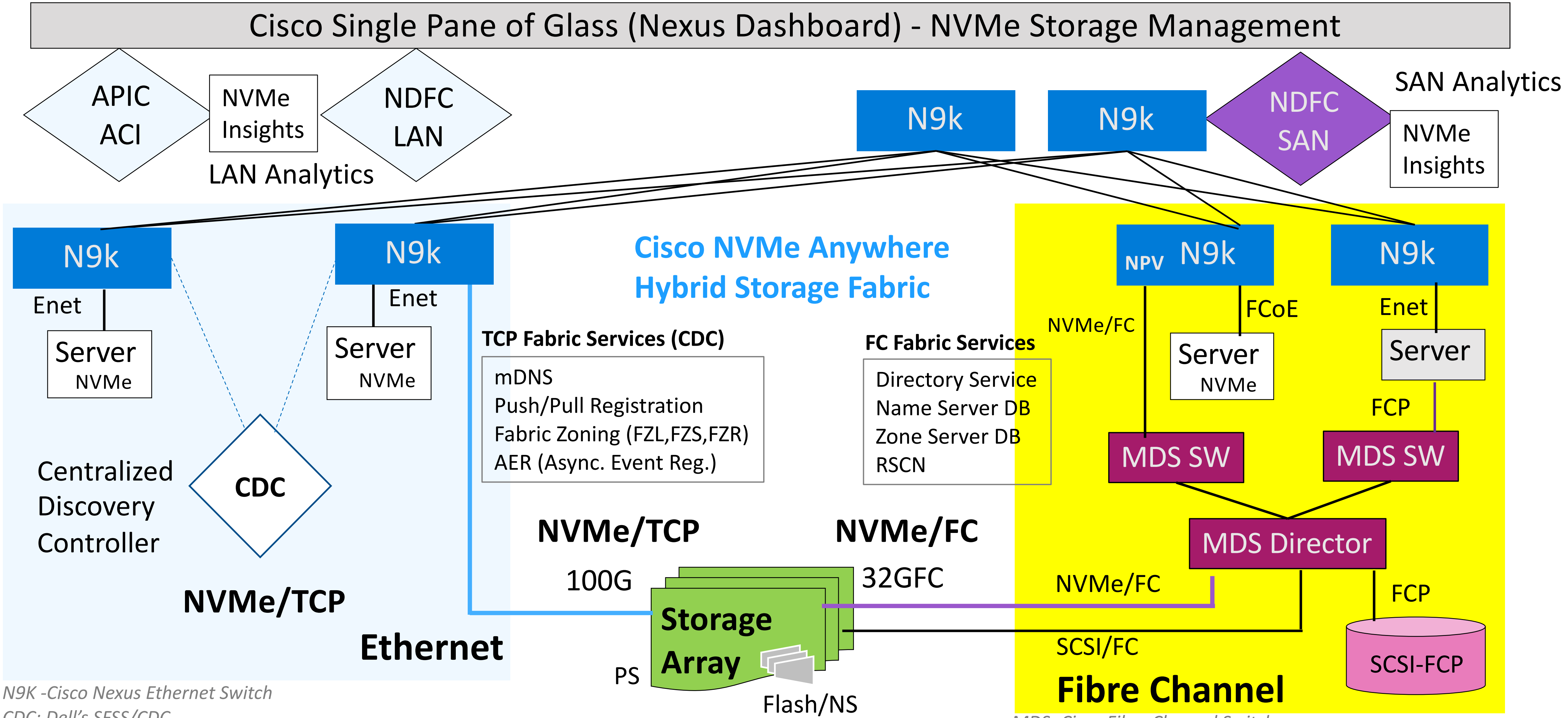
NVMe Storage Architecture (Architecture)



Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

NVMe/FC Architecture



N9K -Cisco Nexus Ethernet Switch
CDC: Dell's SFSS/CDC

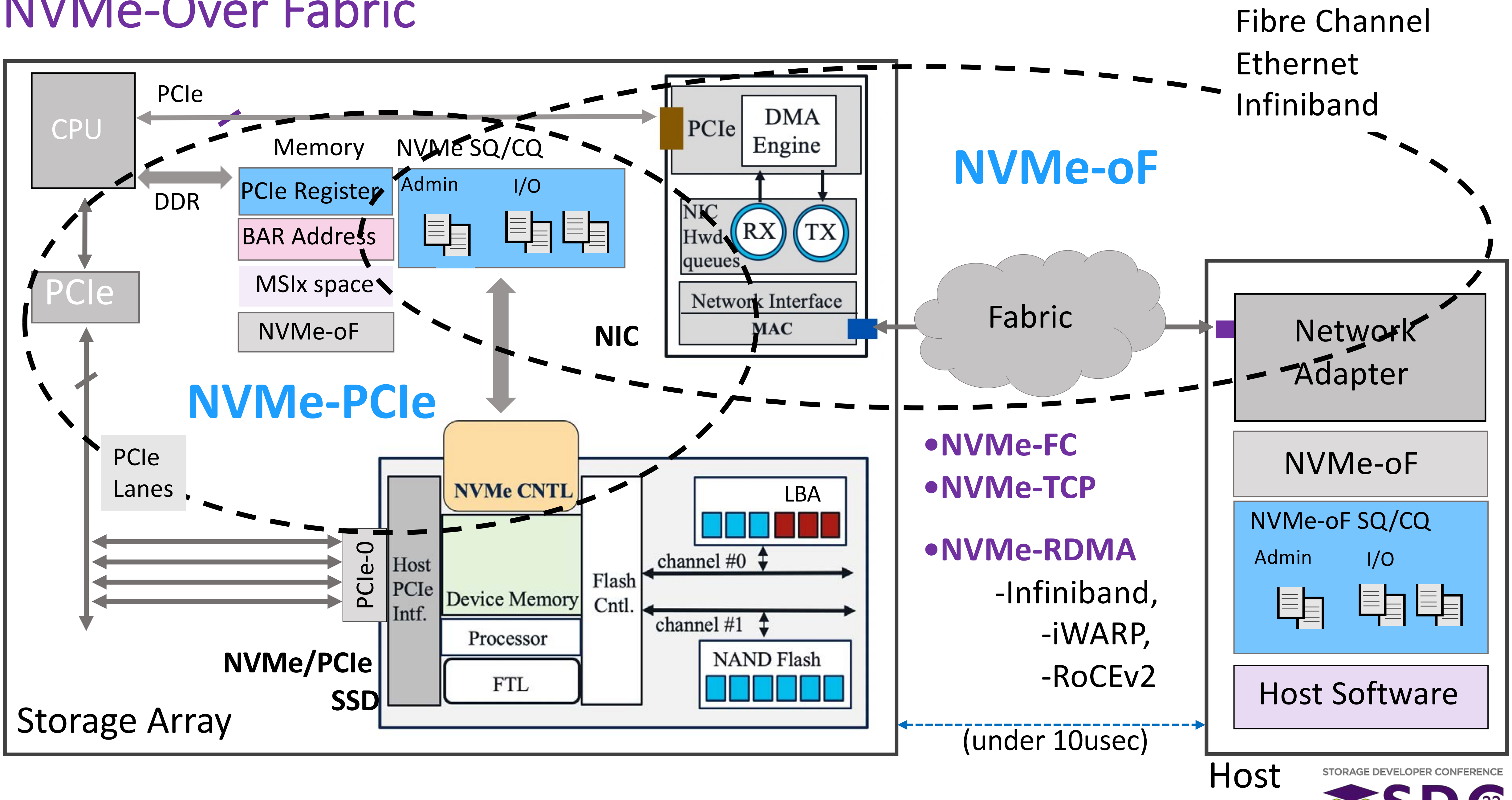
MDS -Cisco Fibre Channel Switch
PS -Dell PowerStore All-Flash Storage

Kamal Bakshi, Cisco

NVMe/FC Architecture

Kamal Bakshi, Cisco

NVMe-Over Fabric



 |

Kamal Bakshi, Cisco

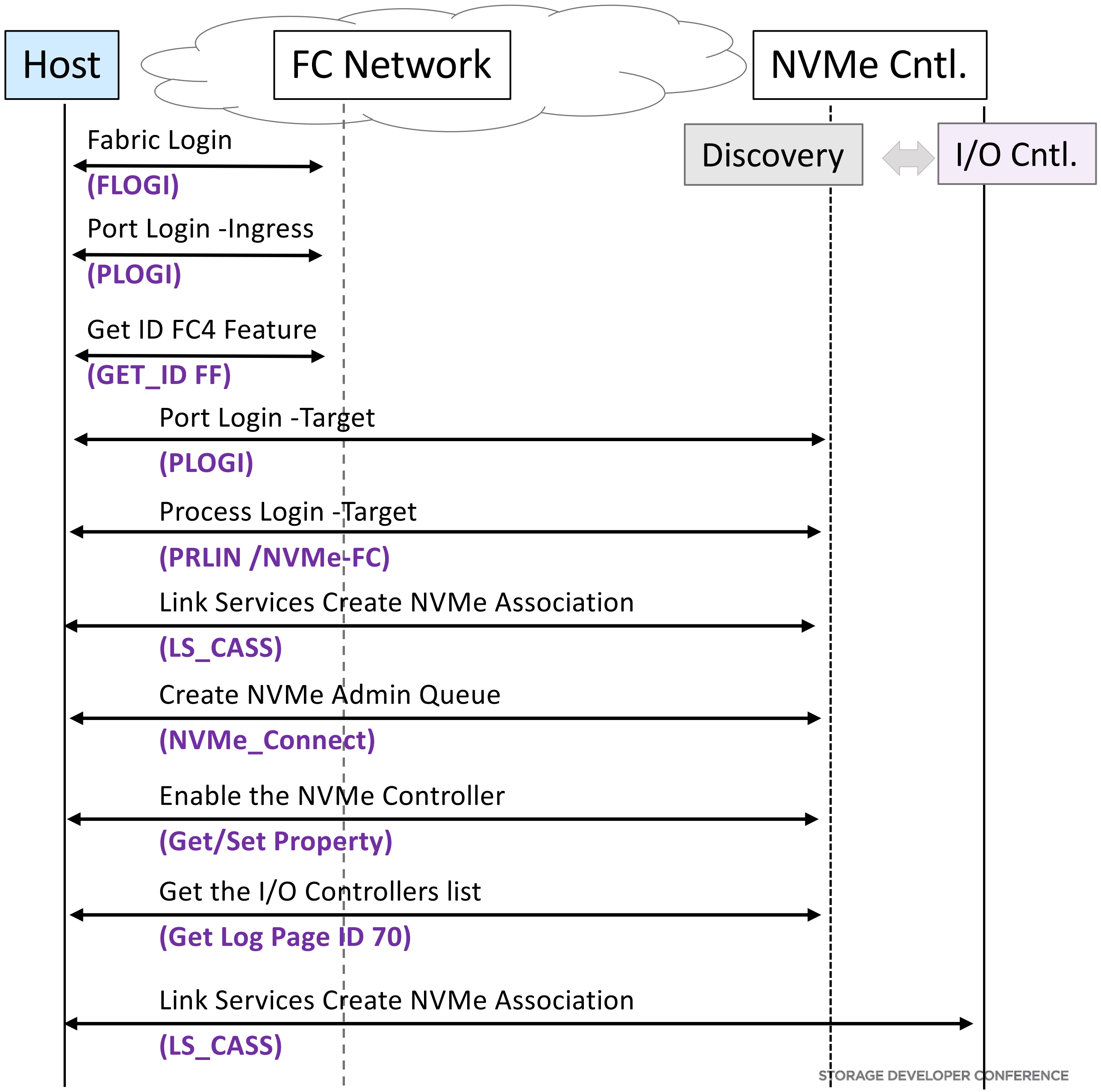
NVMe-FC Transport

I — Connect to NVMe Cntl.

II — Create NVMe Queues

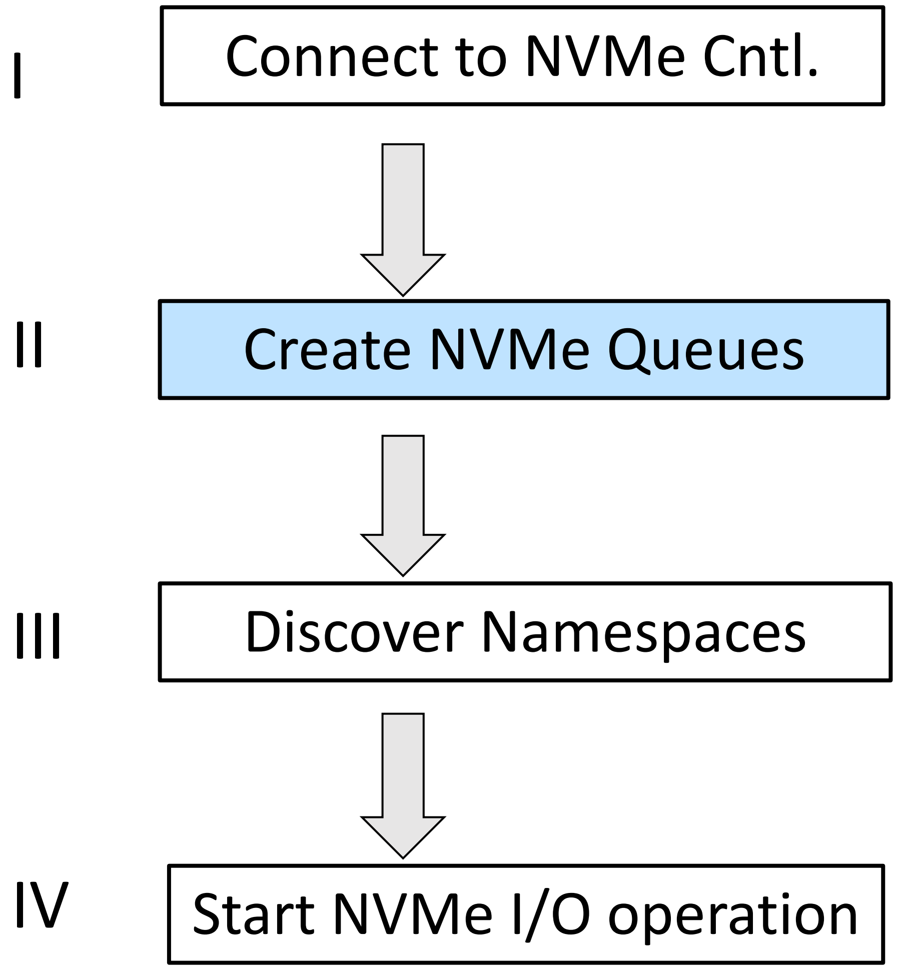
III — Discover Namespaces

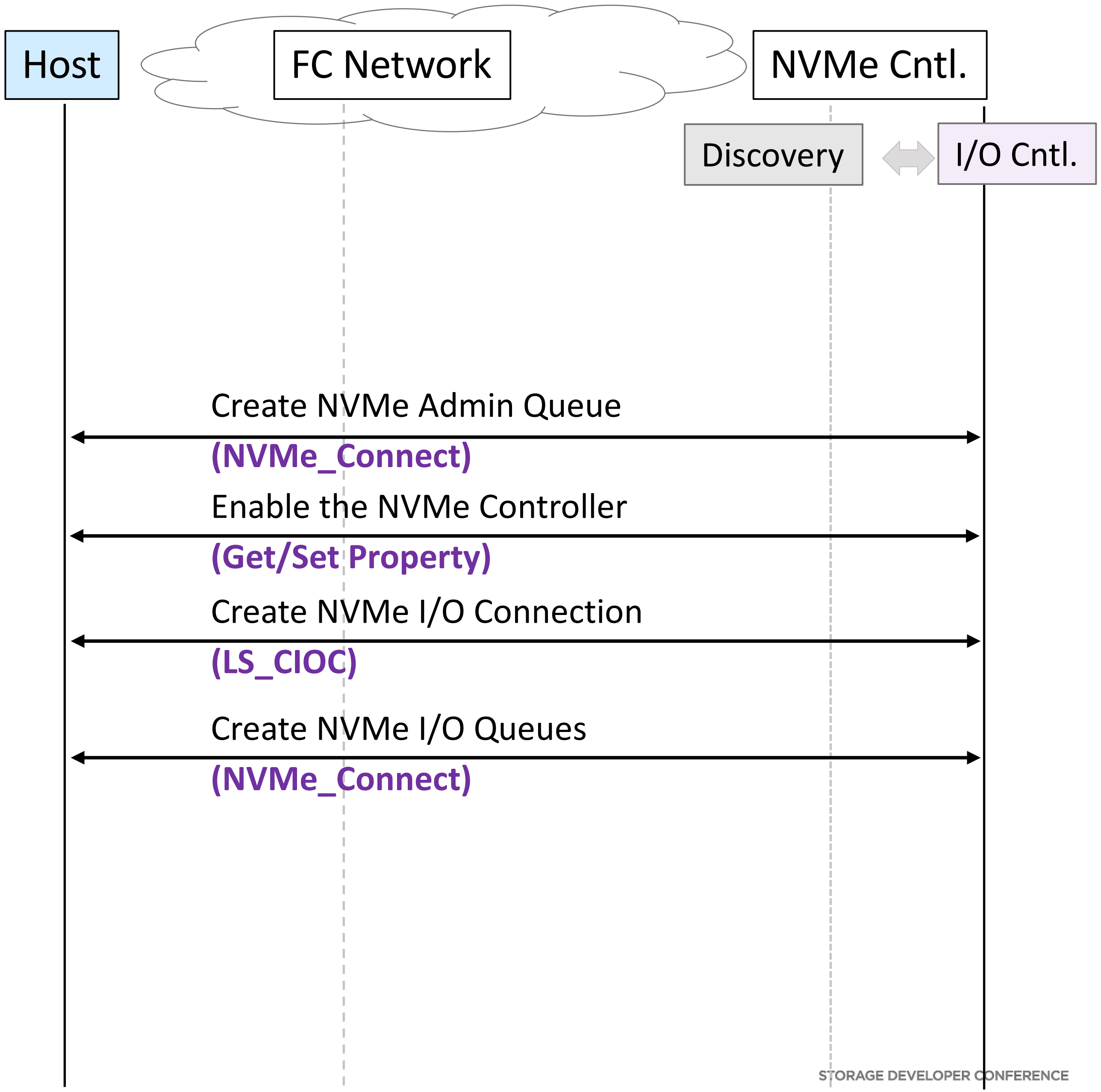
IV — Start NVMe I/O operation



Kamal Bakshi, Cisco

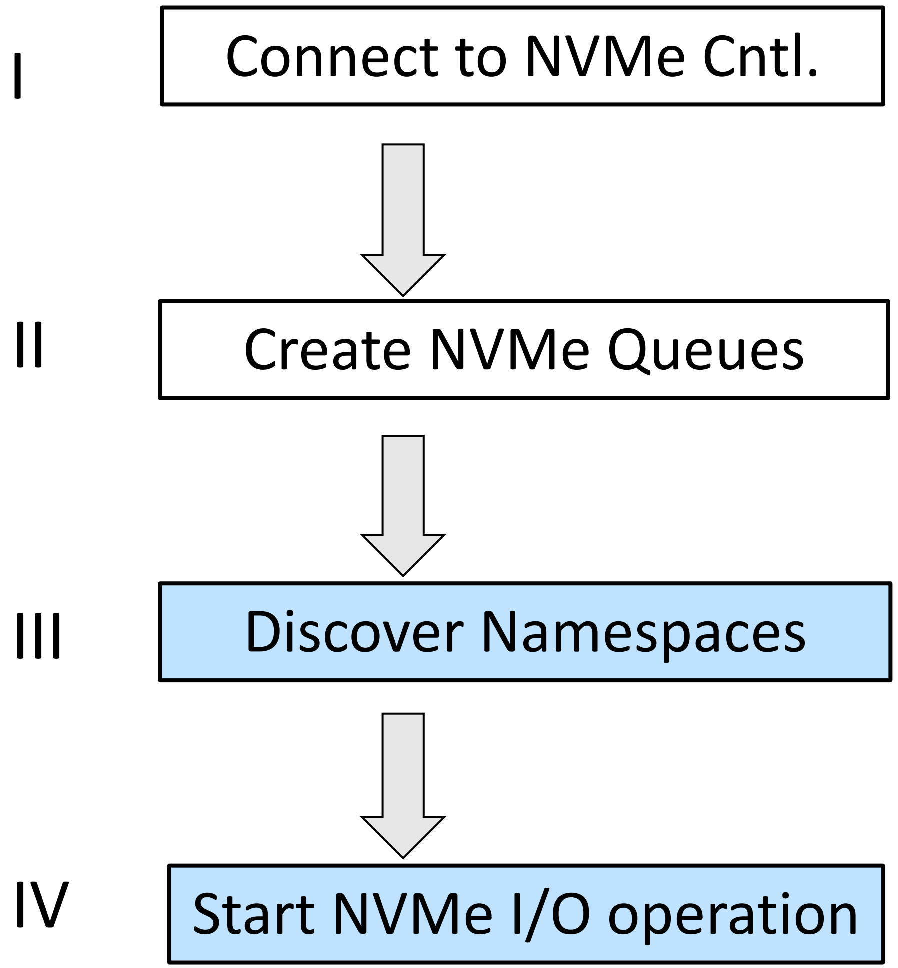
NVMe-FC Transport

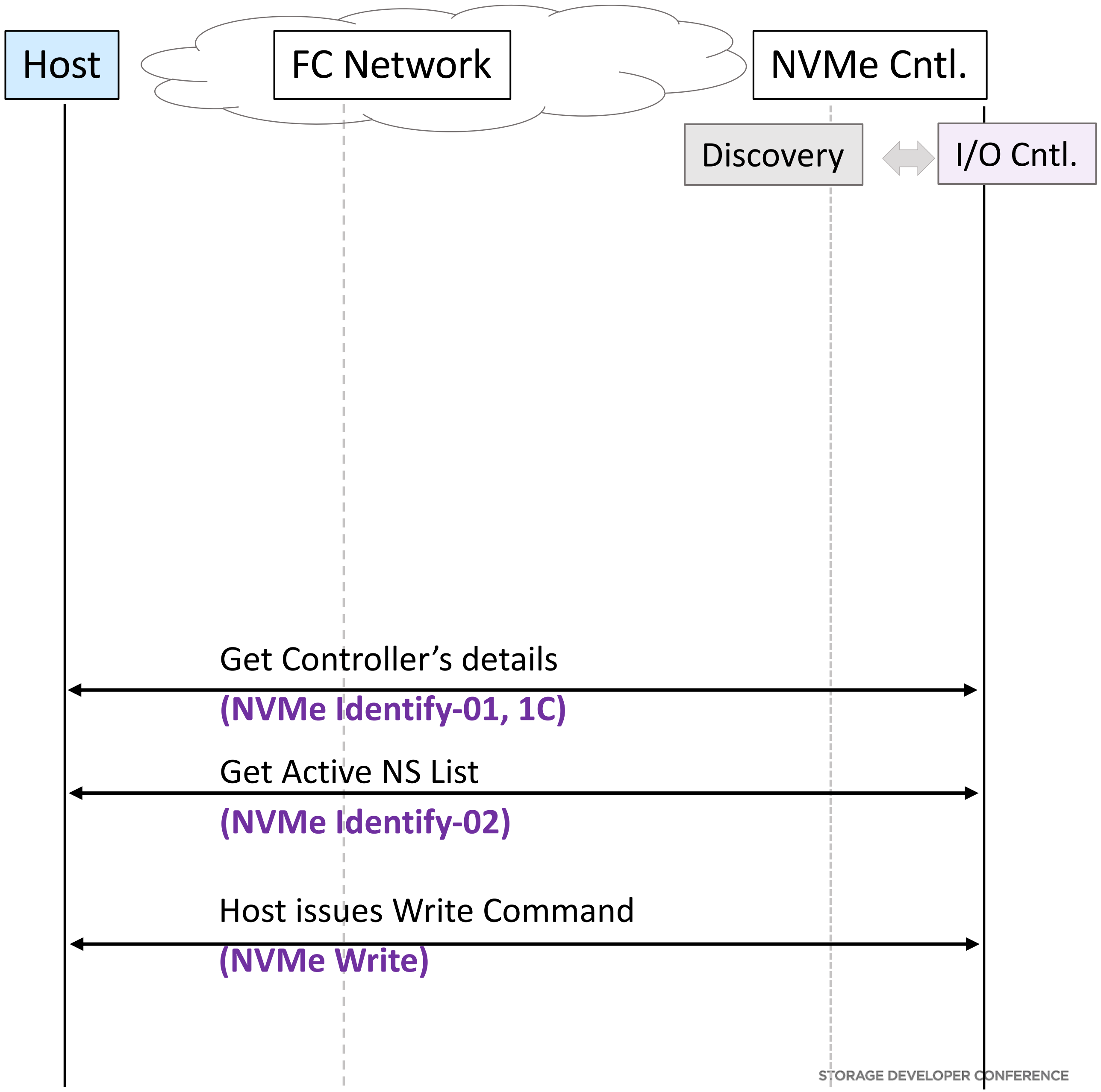




Kamal Bakshi, Cisco

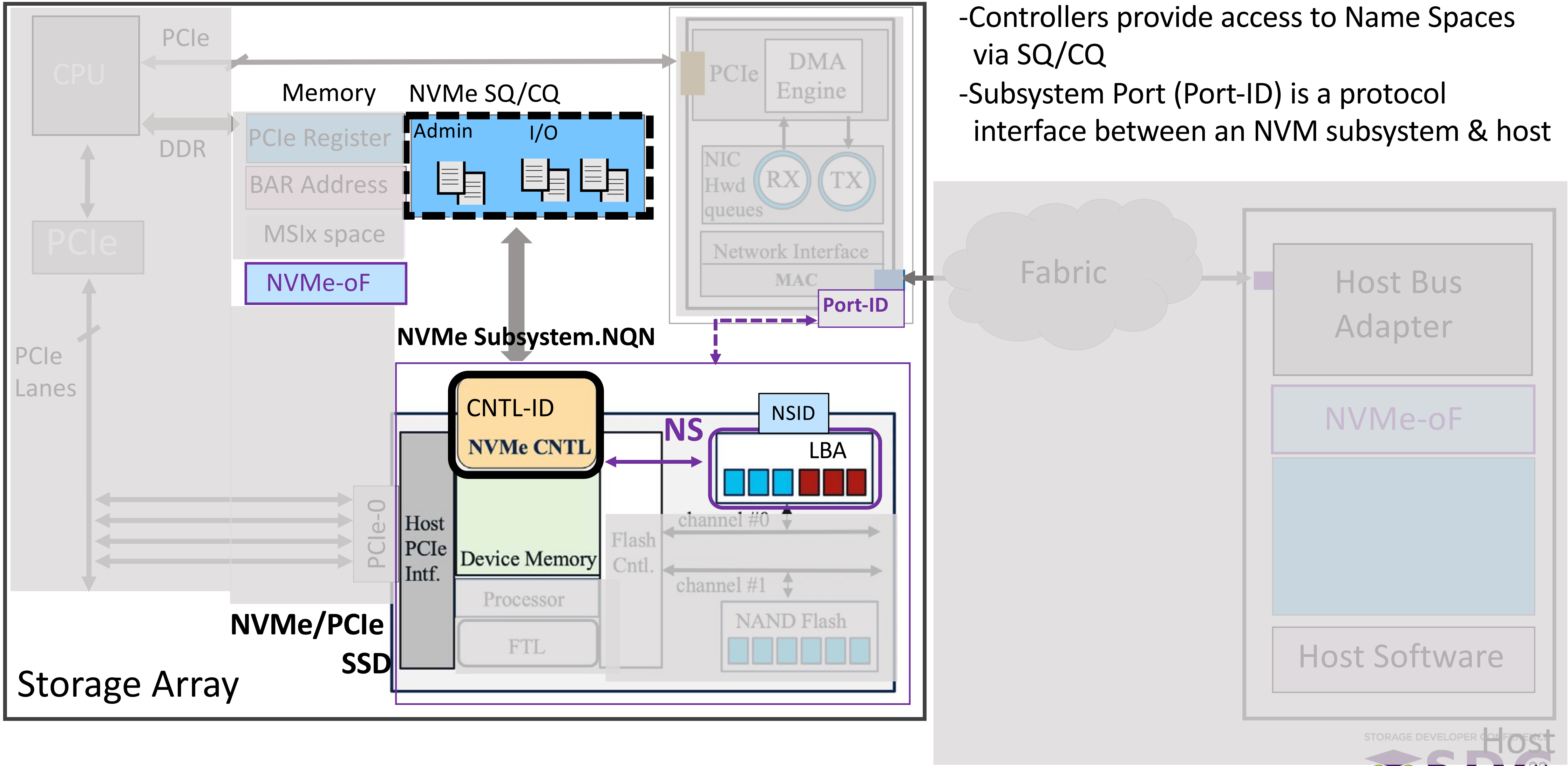
NVMe-FC Transport





Kamal Bakshi, Cisco

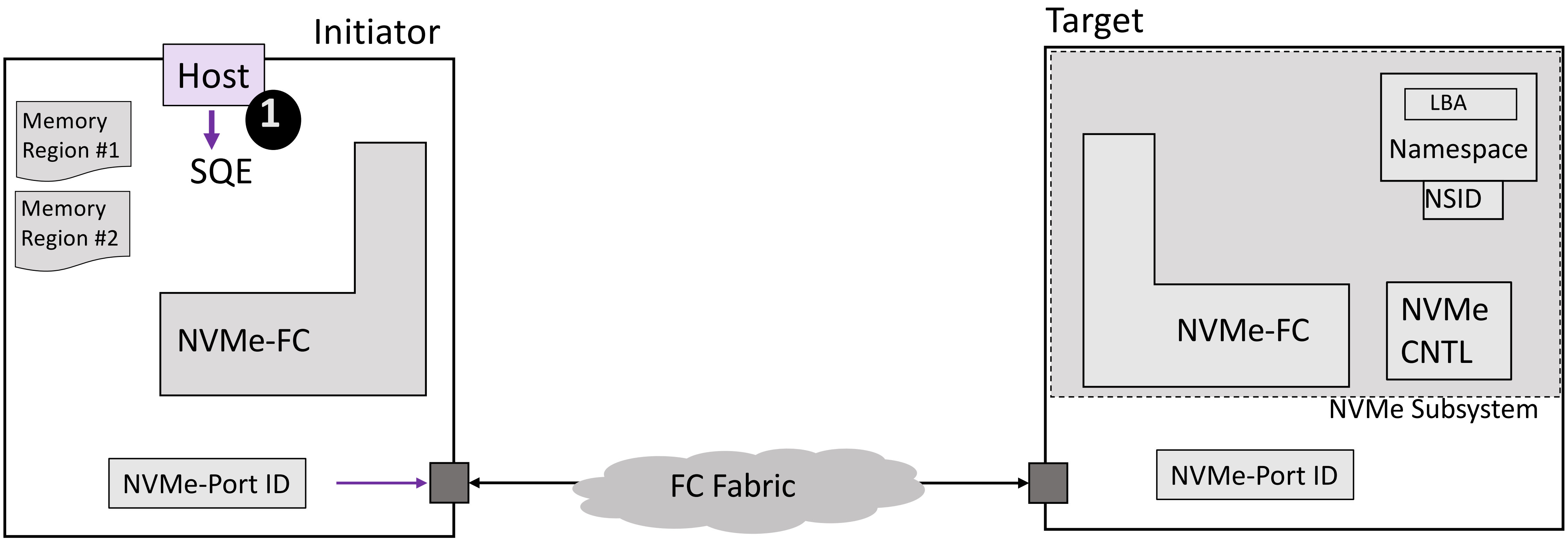
NVMe-oF (NVMe Subsystem)



- NVMe Subsystem consists of multiple CNTLs
- Controllers provide access to Name Spaces via SQ/CQ
- Subsystem Port (Port-ID) is a protocol interface between an NVM subsystem & host

Kamal Bakshi, Cisco

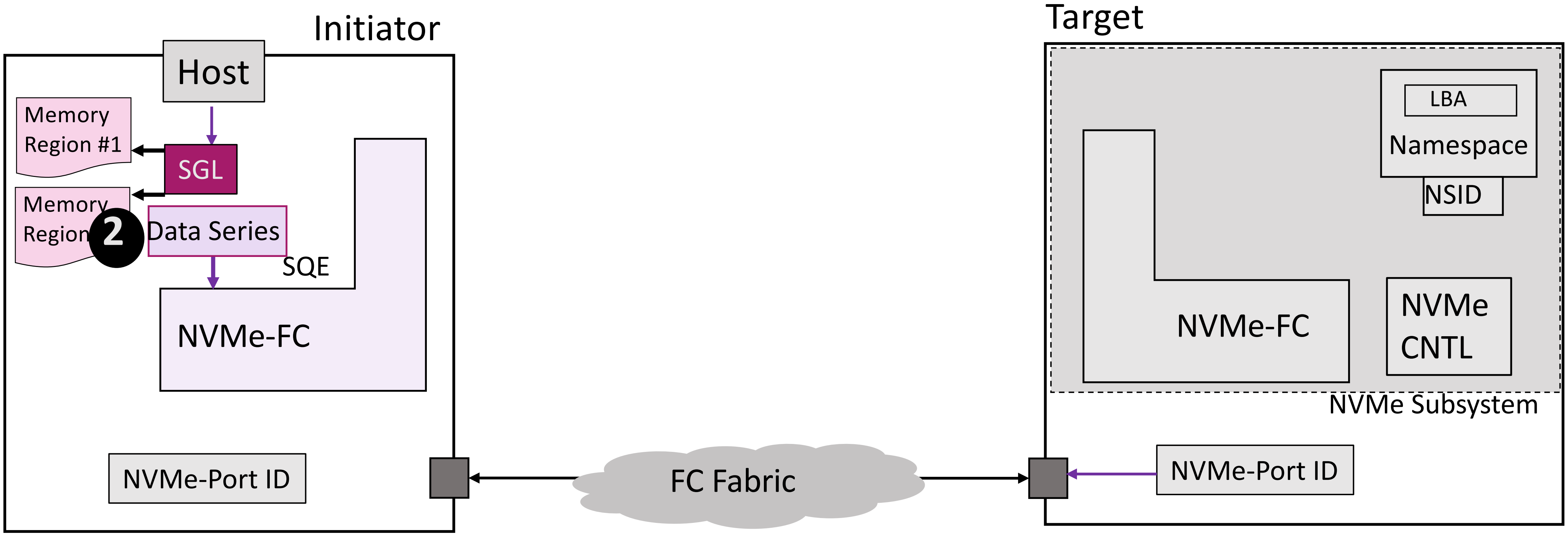
NVMe-oF (FC Mapping Abstractions)



1 NVMe Host Submits a NVMe_Write command as SQE (Submission Queue Entry)

Kamal Bakshi, Cisco

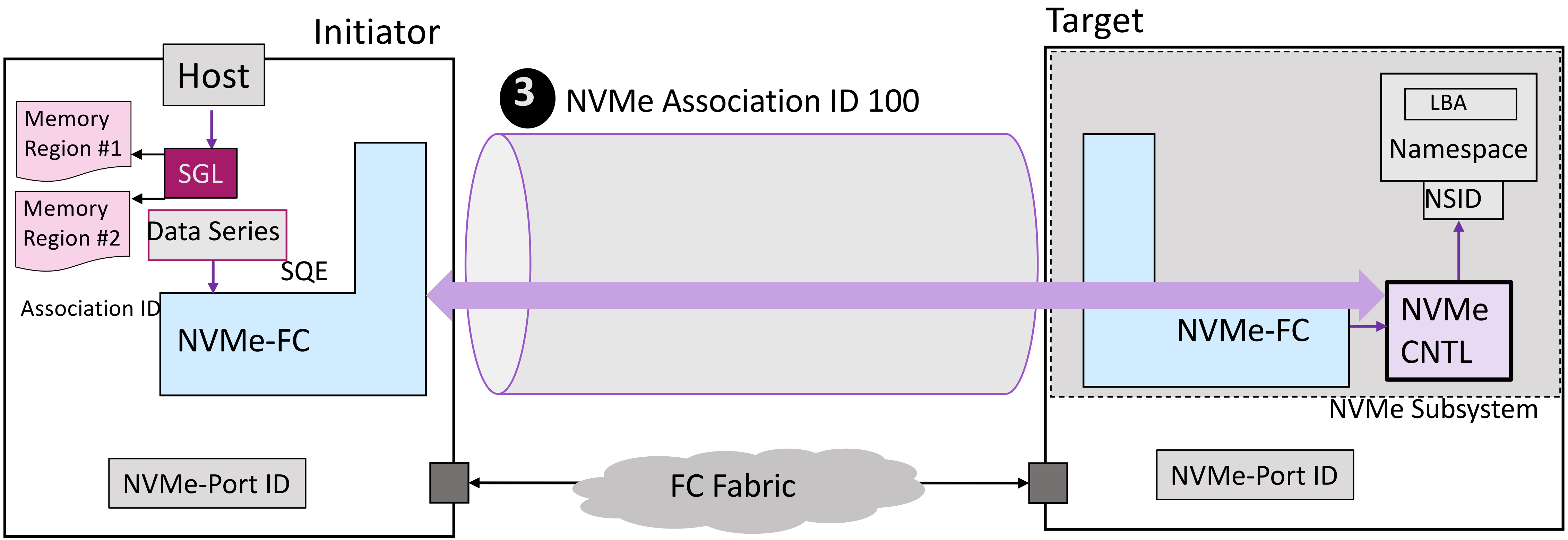
NVMe-oF (FC Mapping Abstractions)



2 Data pointed by the Host SGL is placed in a Data Series and command is passed to NVMe-FC layer

Kamal Bakshi, Cisco

NVMe-oF (FC Mapping Abstractions) -Association ID

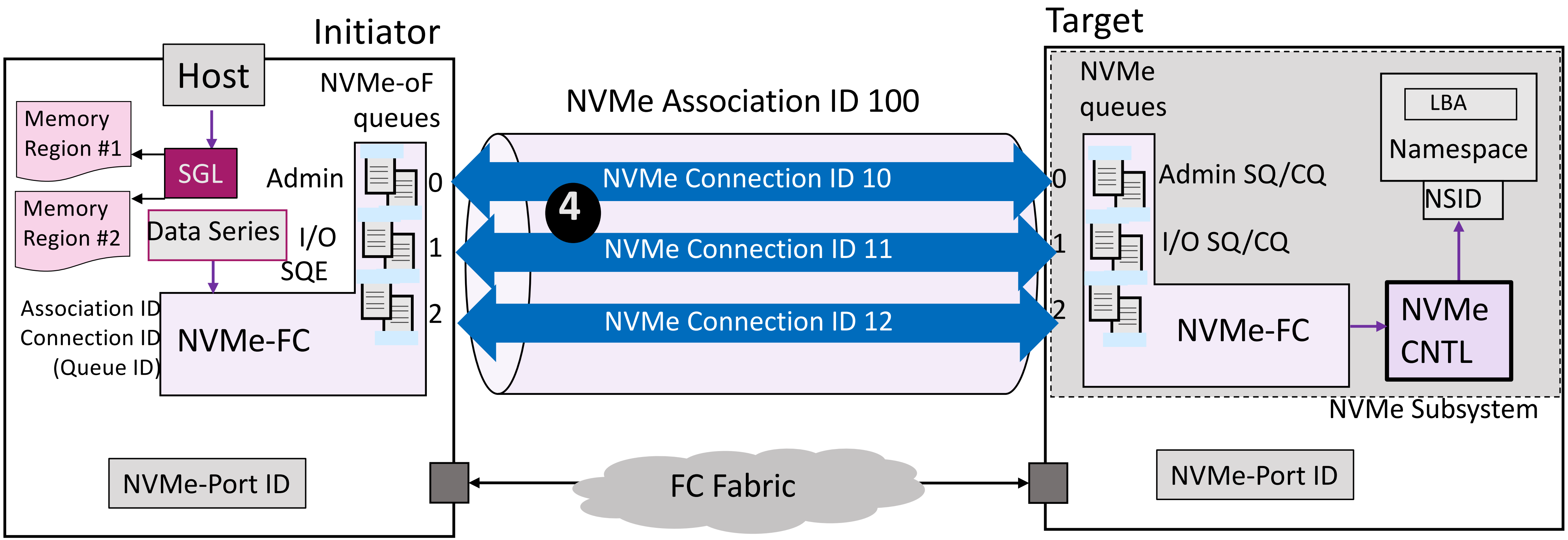


3 The Host NVMe-FC layer specifies the NVMe-FC association with the NVMe controller

 |

Kamal Bakshi, Cisco

NVMe-oF (FC Mapping Abstractions) -Connection IDs

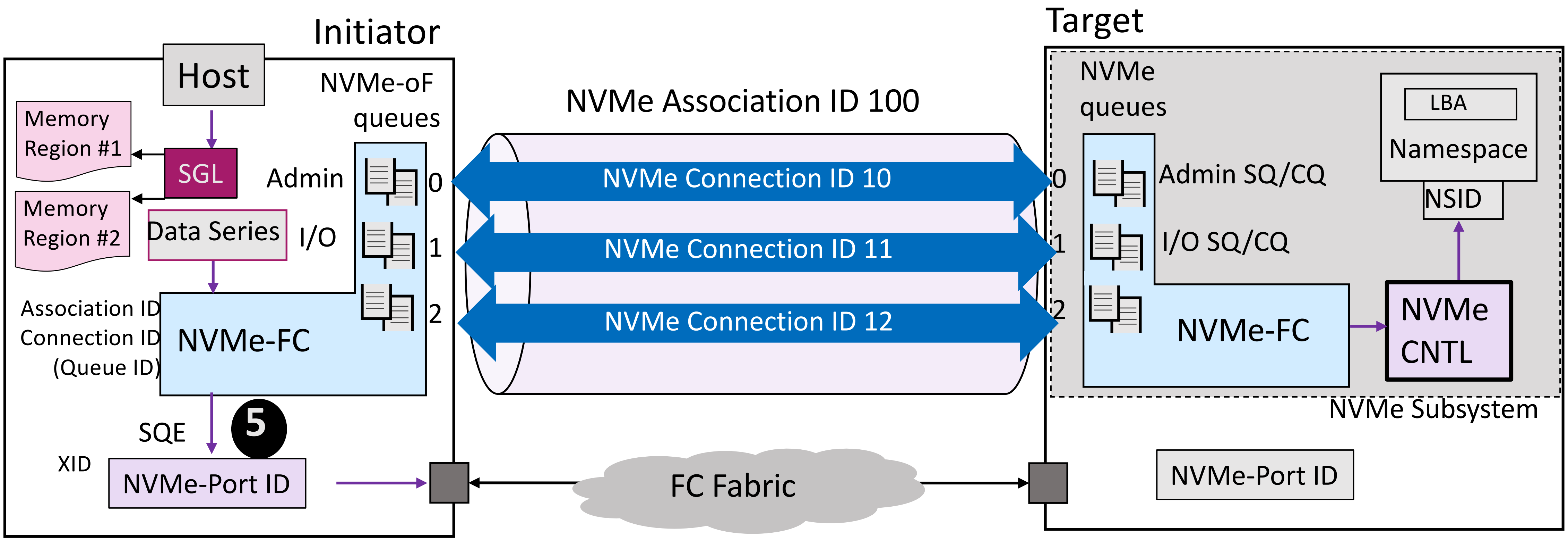


4 The Host NVMe-FC layer maintains a mapping of Host queues (NVMe-oF) to the NVMe controller's NVMe queues (SQ/CQ) via connection IDs.

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

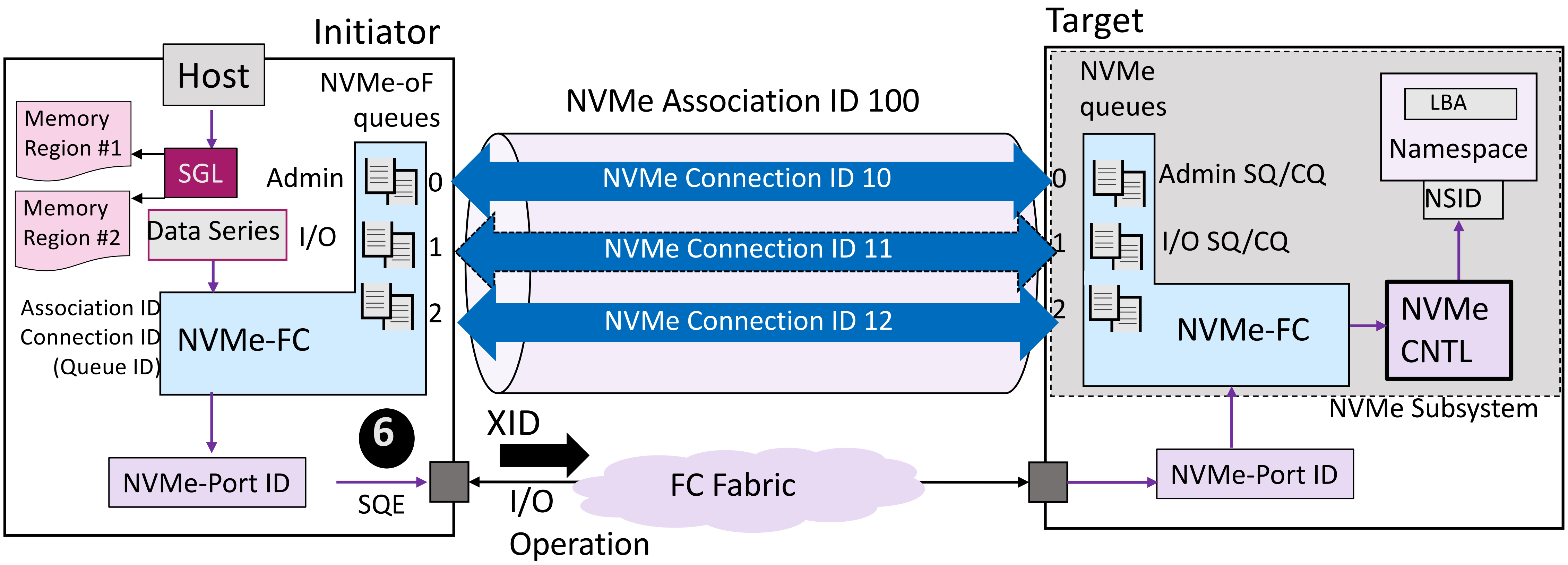
NVMe-oF (FC Mapping Abstractions) -Exchange IDs



5 Upon receiving the SQE command NVMe_Port allocates XID for the NVMe-FC I/O operation and associates the NVMe command in the SQE to the Exchange. All NVMe IUs for the NVMe-FC I/O operation are transmitted as part of this Exchange.

Kamal Bakshi, Cisco

NVMe-FC (Association ID, Connection ID, Exchange ID, Queue ID)

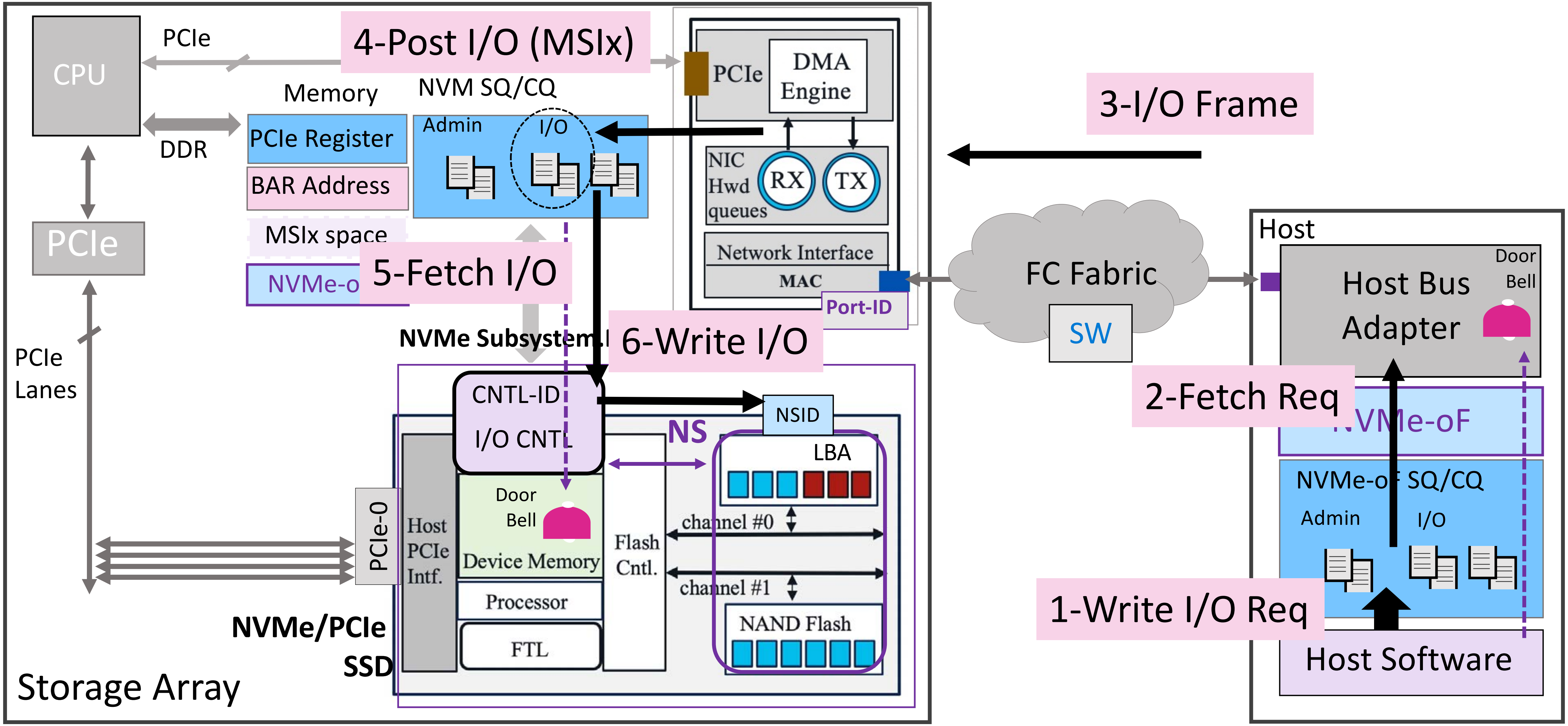


6 The initiator NVMe_Port transmits the NVMe_CMND IU payload to start the NVMe-FC I/O operation.

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE

NVMe-oF (HBA/MSIx Interrupts)

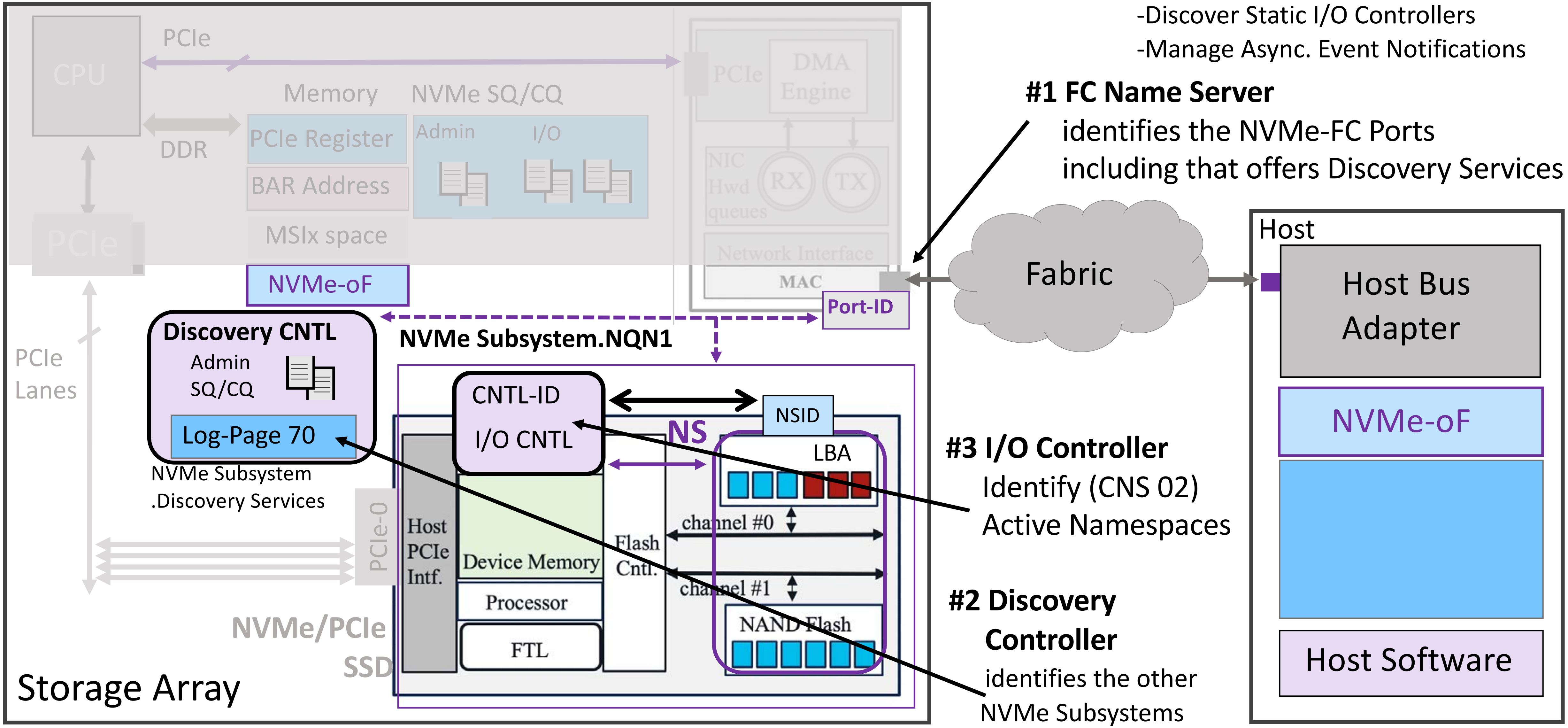


 |

Kamal Bakshi, Cisco

NVMe-oF (Discovery Services Subsystem)

-Discover Subsystems with Namespaces
-Discover Multiple Paths to Subsystems
-Discover Static I/O Controllers
-Manage Async. Event Notifications



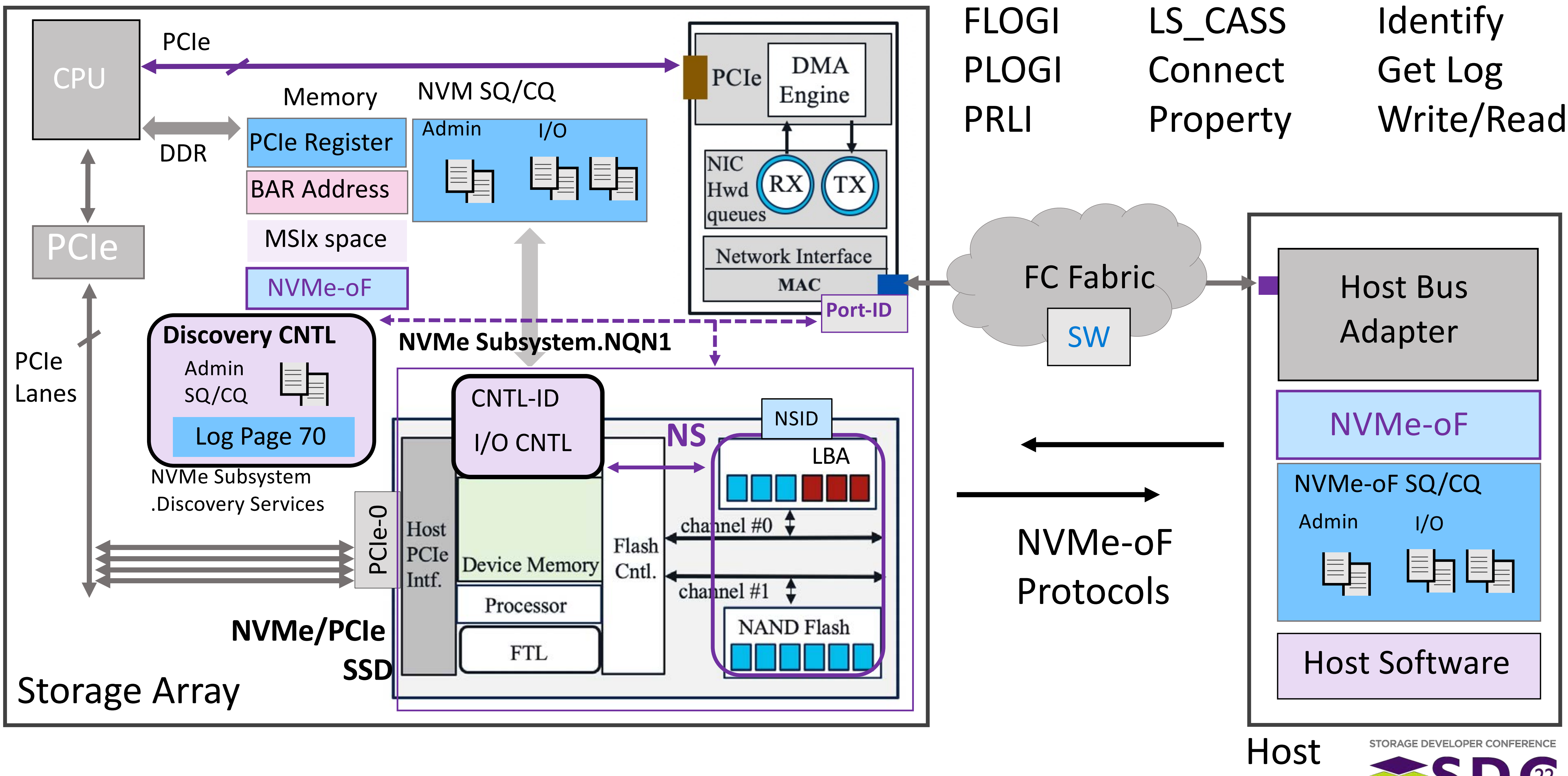
#1 FC Name Server
identifies the NVMe-FC Ports including that offers Discovery Services

#3 I/O Controller
Identify (CNS 02) Active Namespaces

#2 Discovery Controller
identifies the other NVMe Subsystems I/O CNTL, Discovery CNTL

Kamal Bakshi, Cisco

NVMe-FC Protocol Flows



Kamal Bakshi, Cisco

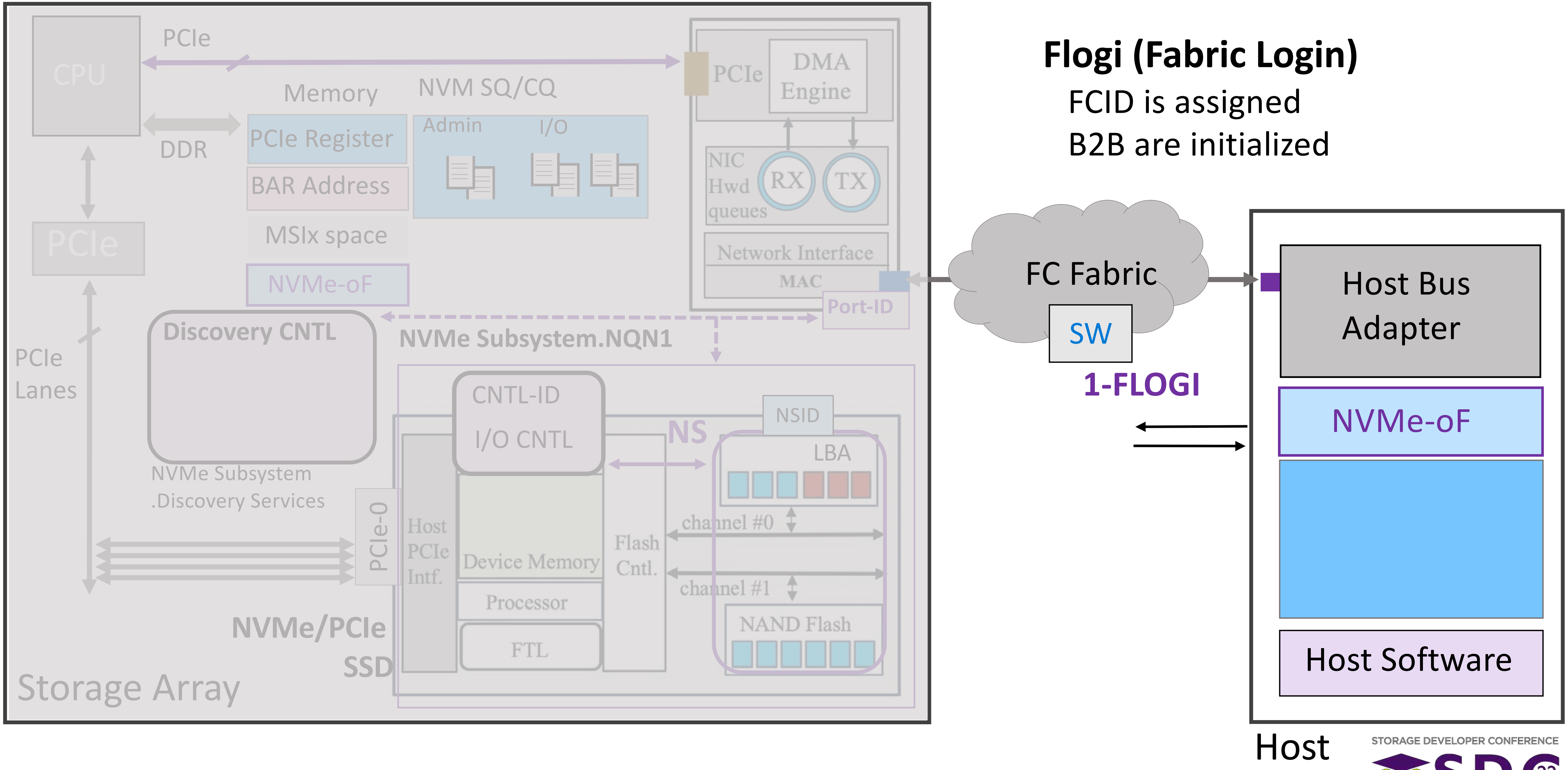
STORAGE DEVELOPER CONFERENCE SDC 22

NVMe-FC Protocol Flows (FLOGI)

Flogi (Fabric Login)

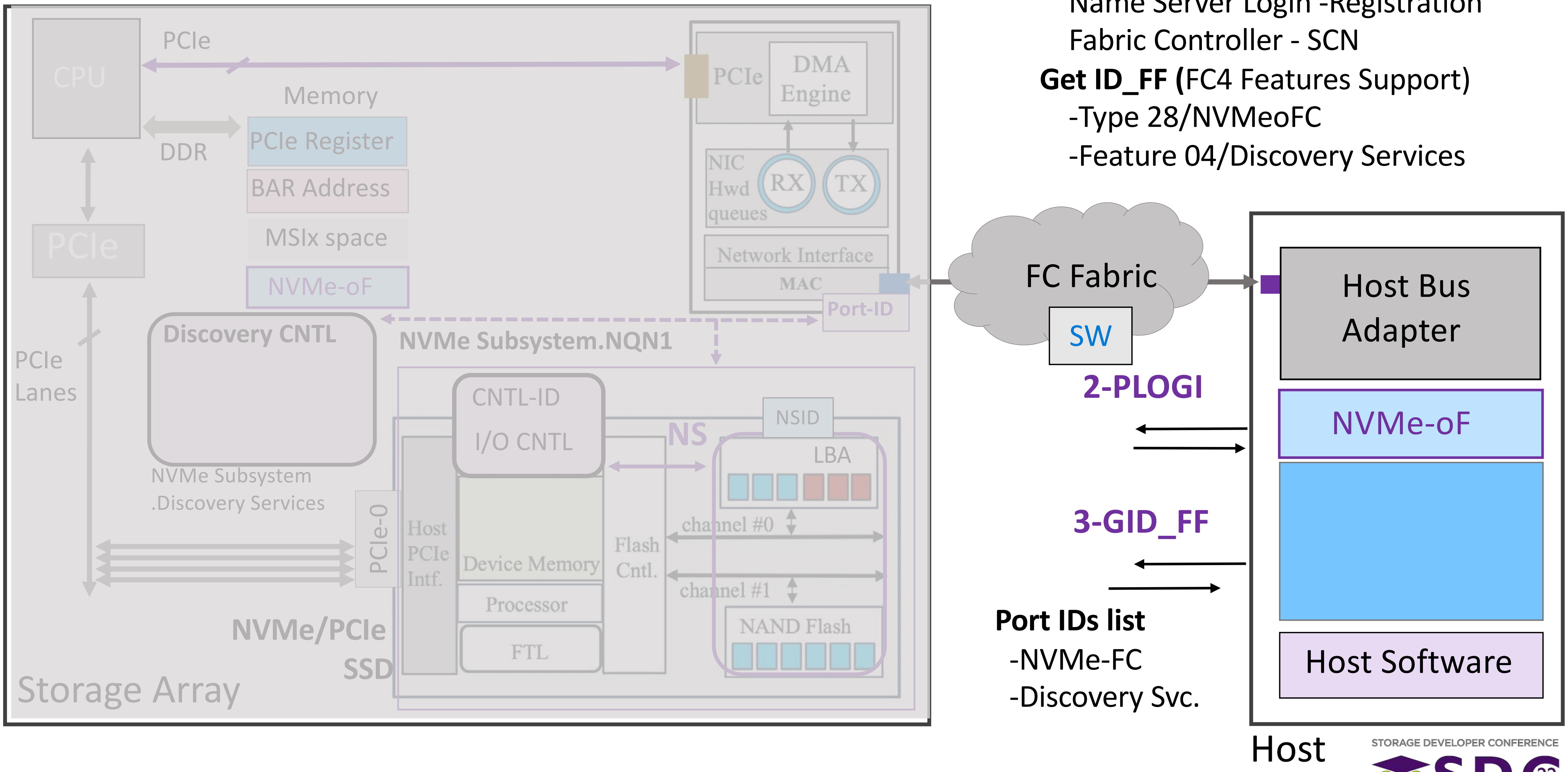
FCID is assigned

B2B are initialized



Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (PLOGI)



Port Logins

Name Server Login -Registration

Fabric Controller - SCN

Get ID_FF (FC4 Features Support)

-Type 28/NVMeoFC

-Feature 04/Discovery Services

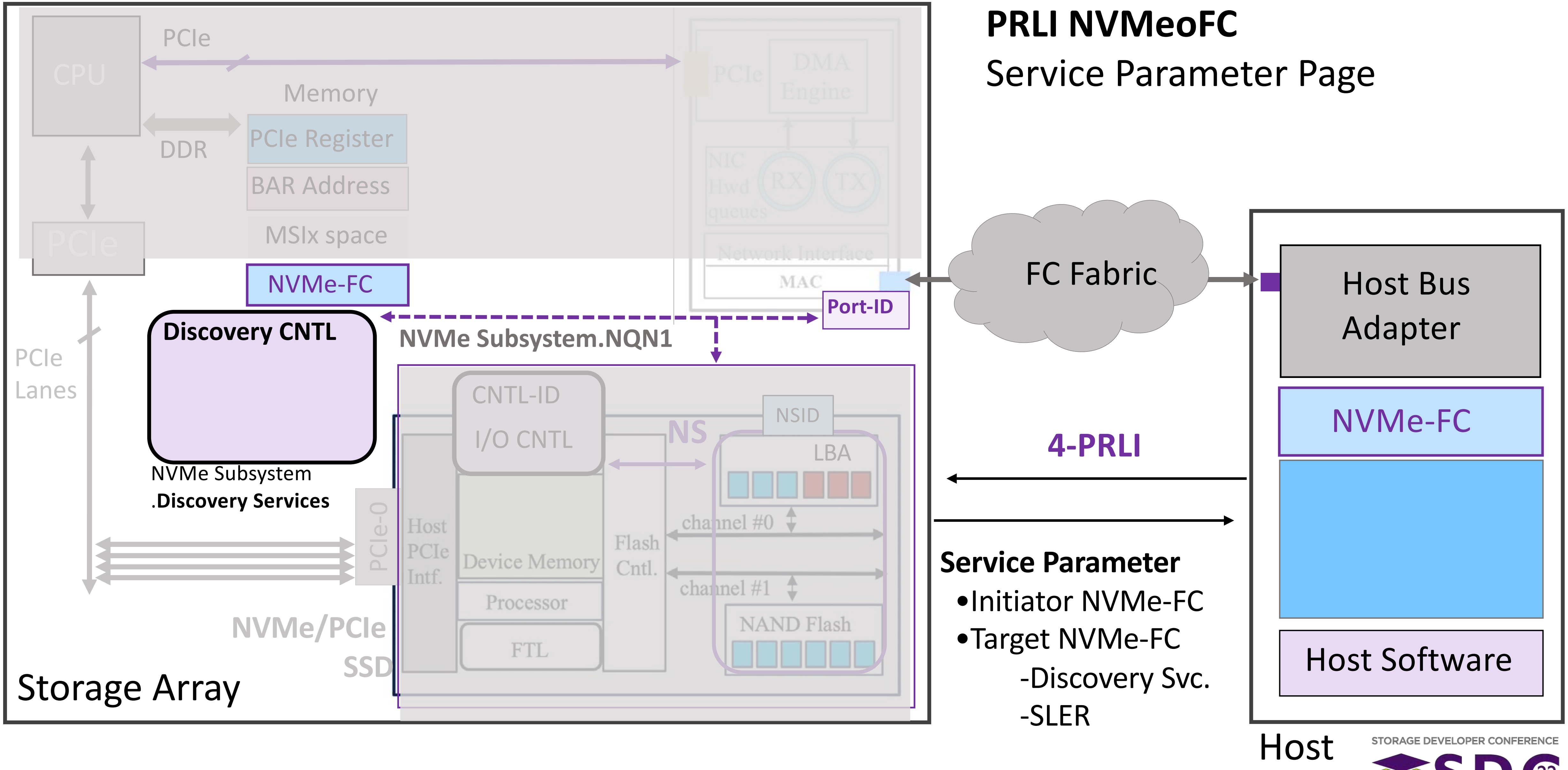
Port IDs list

-NVMe-FC

-Discovery Svc.

Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (PRLI)



PRLI NVMeoFC

Service Parameter Page

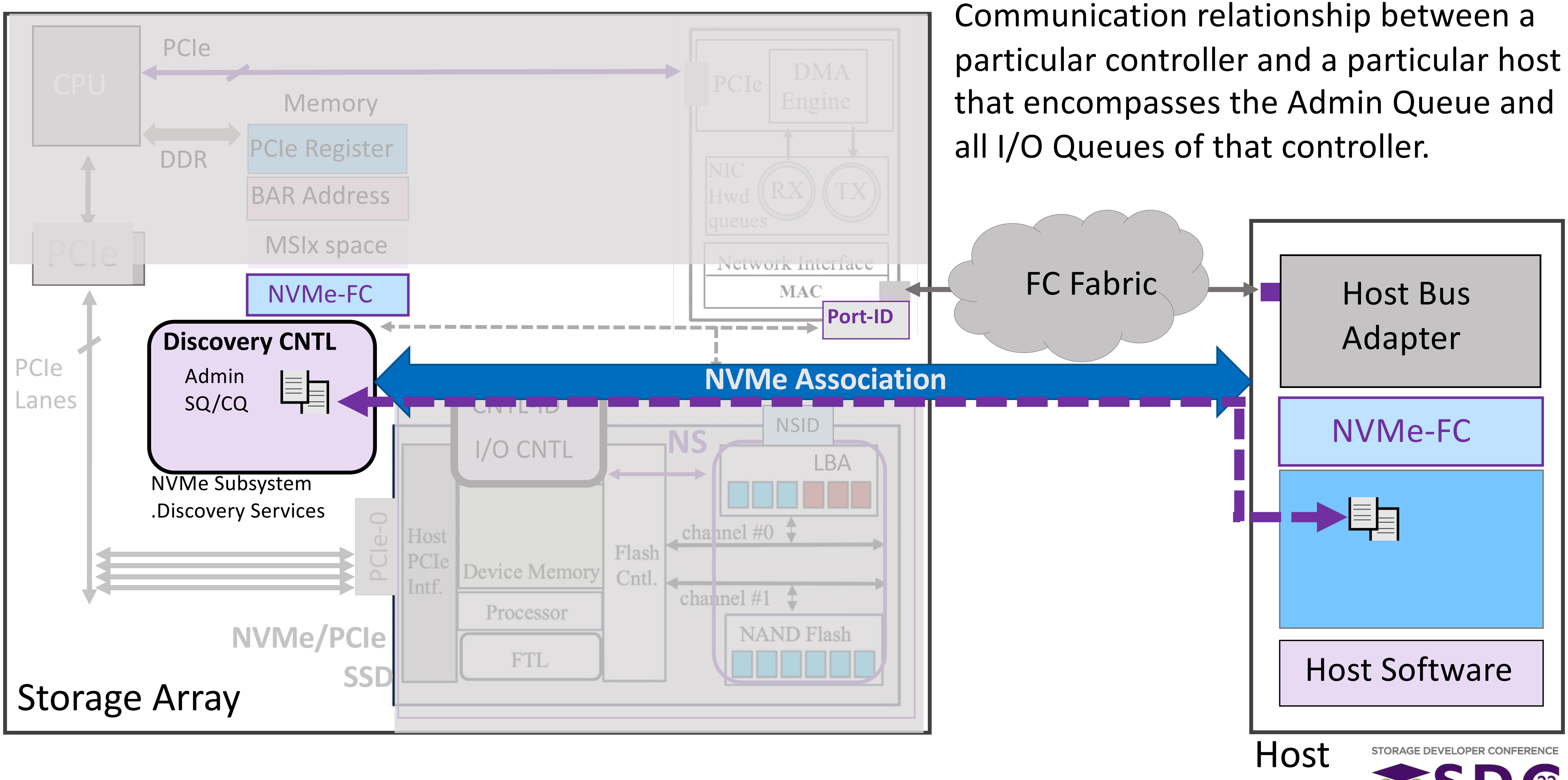
Service Parameter

- Initiator NVMe-FC
- Target NVMe-FC
  - -Discovery Svc.
  - -SLER

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

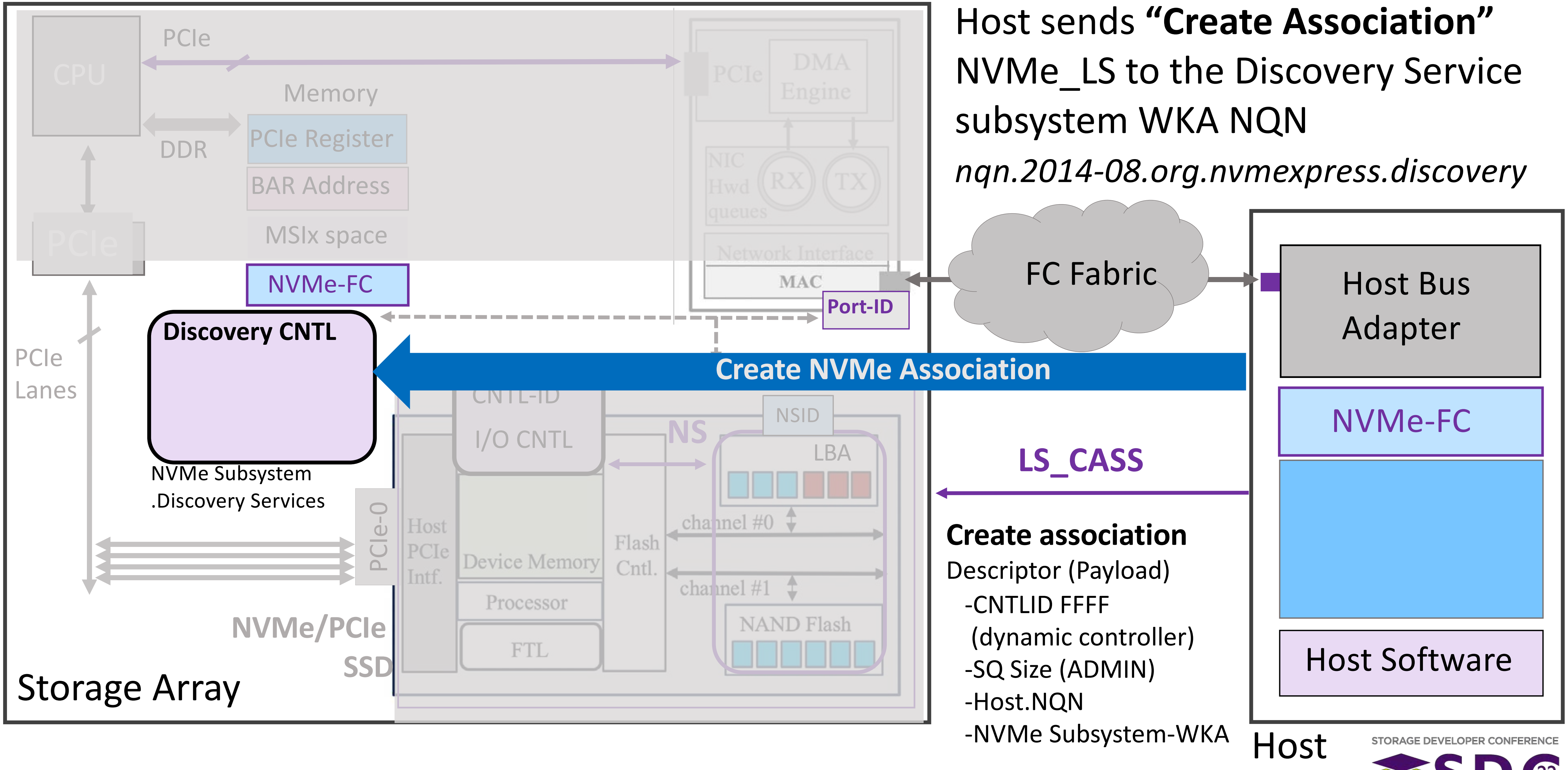
NVMe-FC Protocol Flows (NVMe Association)



Communication relationship between a particular controller and a particular host that encompasses the Admin Queue and all I/O Queues of that controller.

Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (LS _CASS)



Host sends **"Create Association"** NVMe_LS to the Discovery Service subsystem WKA NQN *nqn.2014-08.org.nvmexpress.discovery*

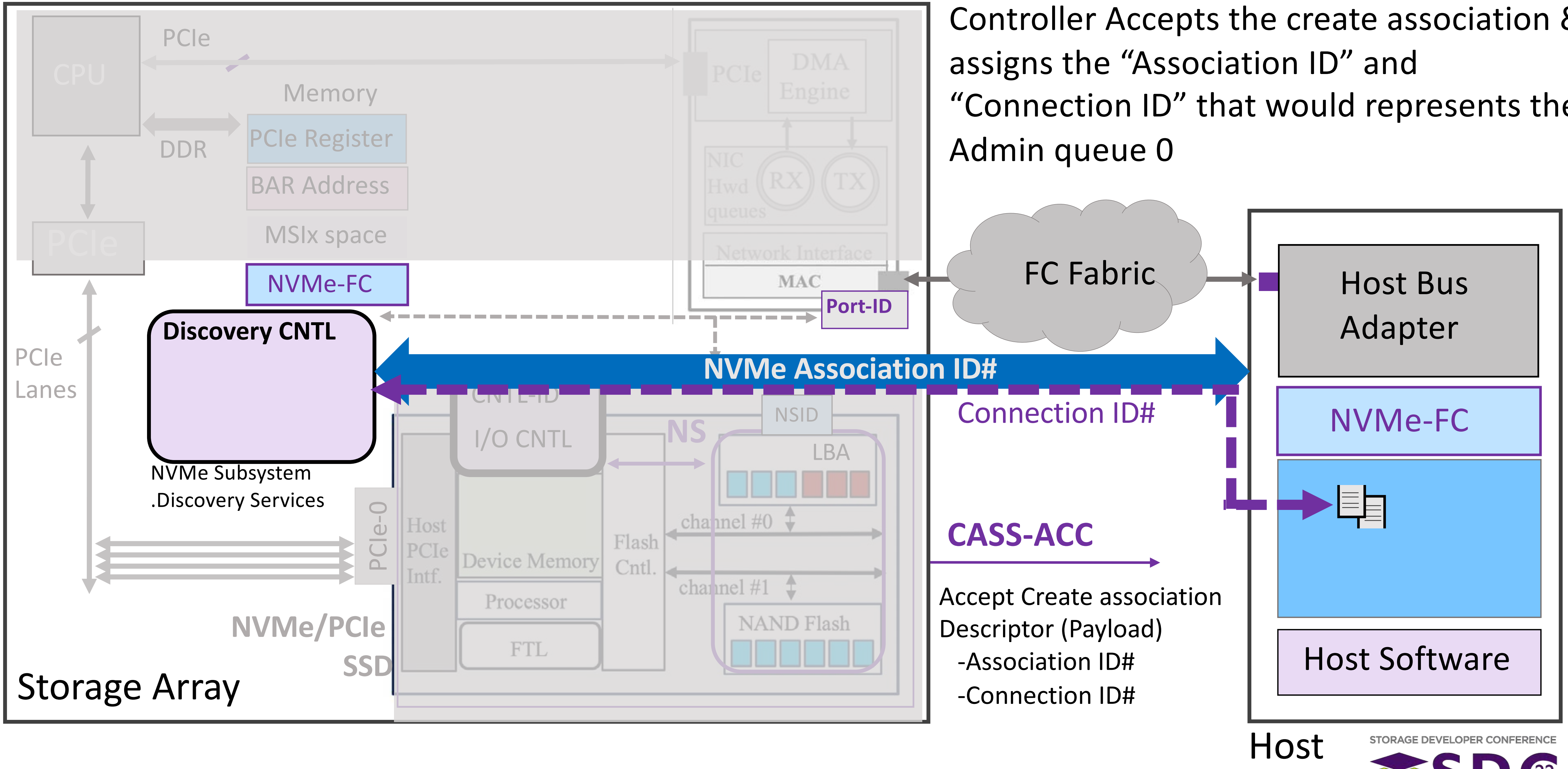
Create association

Descriptor (Payload)

- -CNTLID FFFF (dynamic controller)
- -SQ Size (ADMIN)
- -Host.NQN
- -NVMe Subsystem-WKA

Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (LS CASS_ACC)

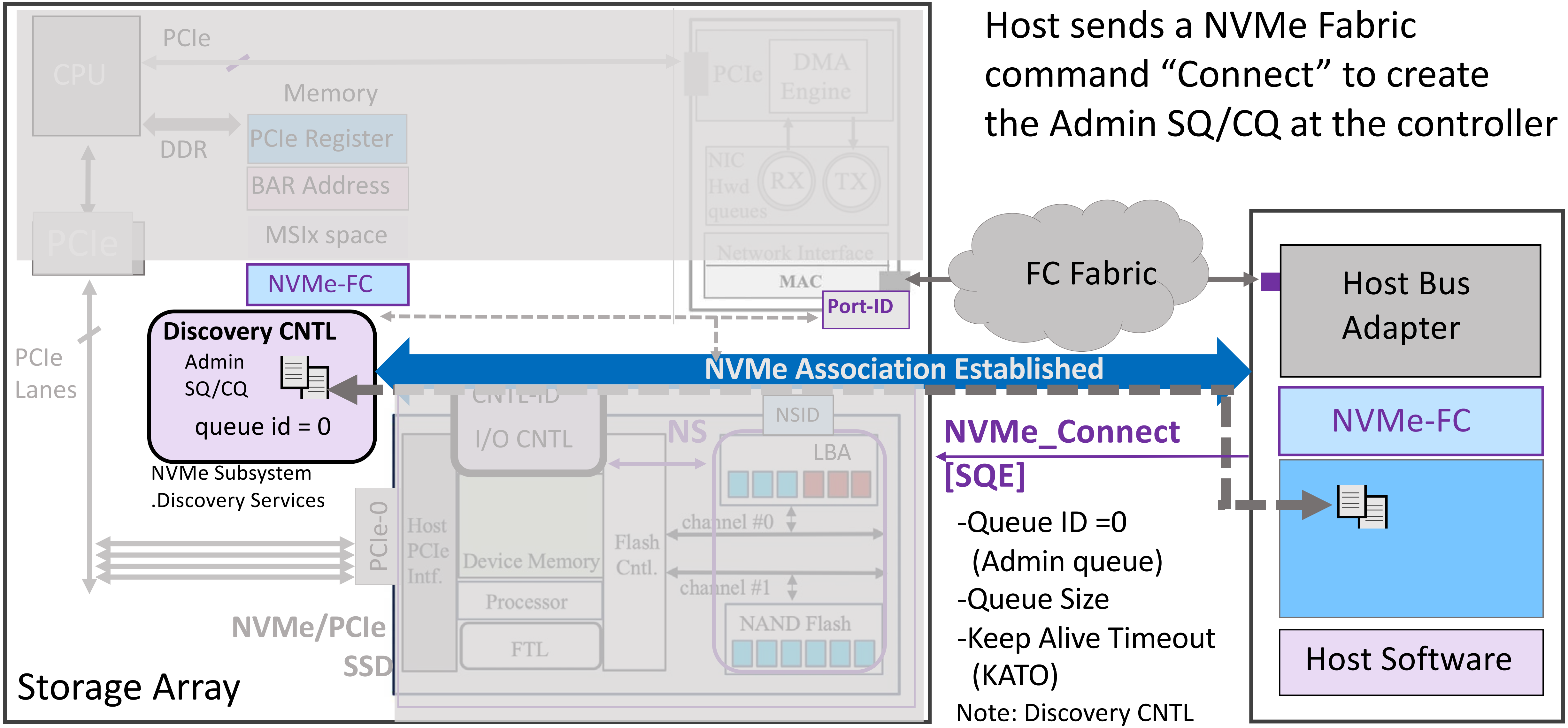


Controller Accepts the create association & assigns the "Association ID" and "Connection ID" that would represents the Admin queue 0

Accept Create association Descriptor (Payload)
- -Association ID#
- -Connection ID#

Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (Connect Command SQE)



Host sends a NVMe Fabric command "Connect" to create the Admin SQ/CQ at the controller

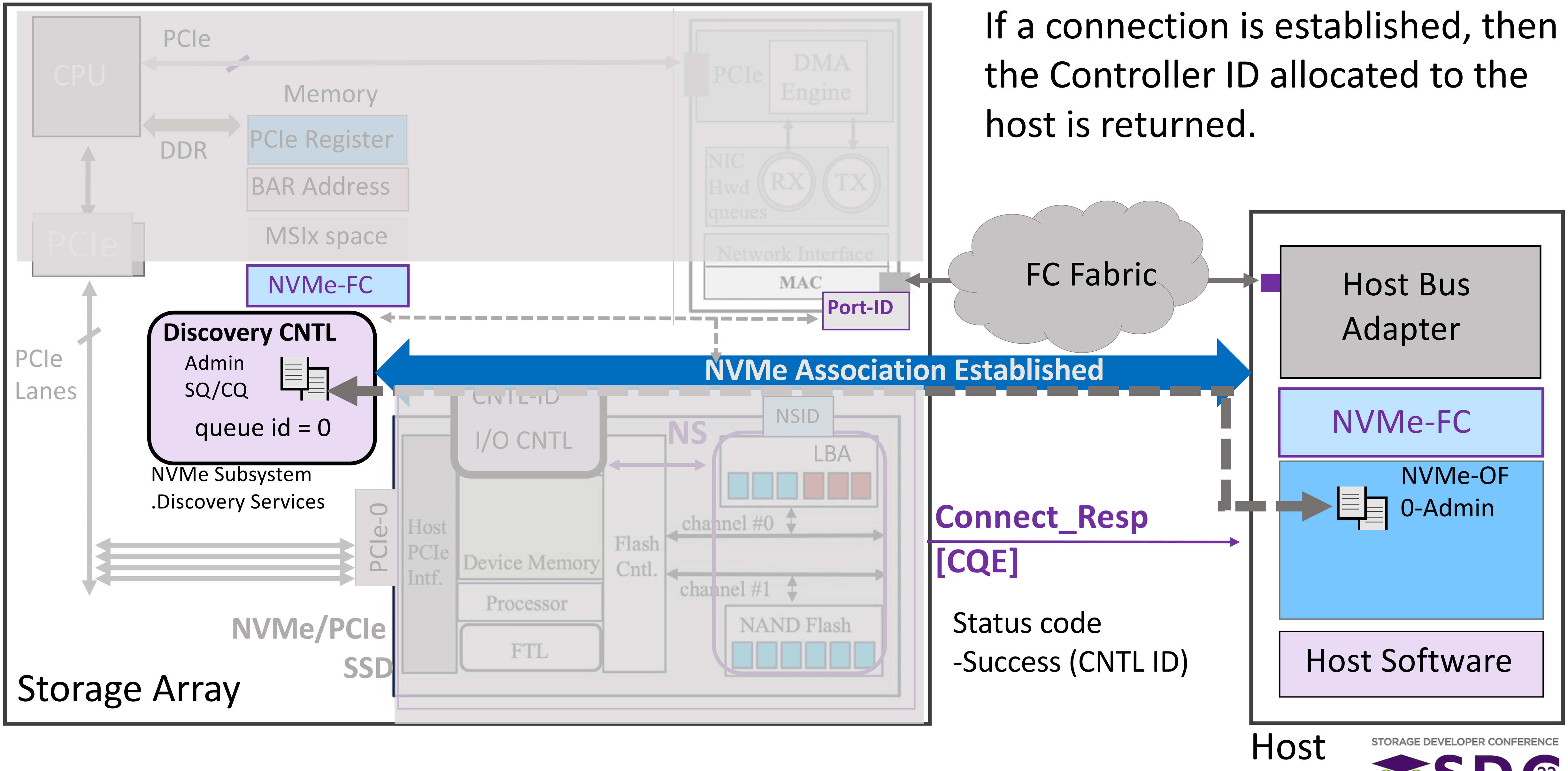
- -Queue ID =0 (Admin queue)
- -Queue Size
- -Keep Alive Timeout (KATO)

Note: Discovery CNTL doesn't have I/O queues

Kamal Bakshi, Cisco

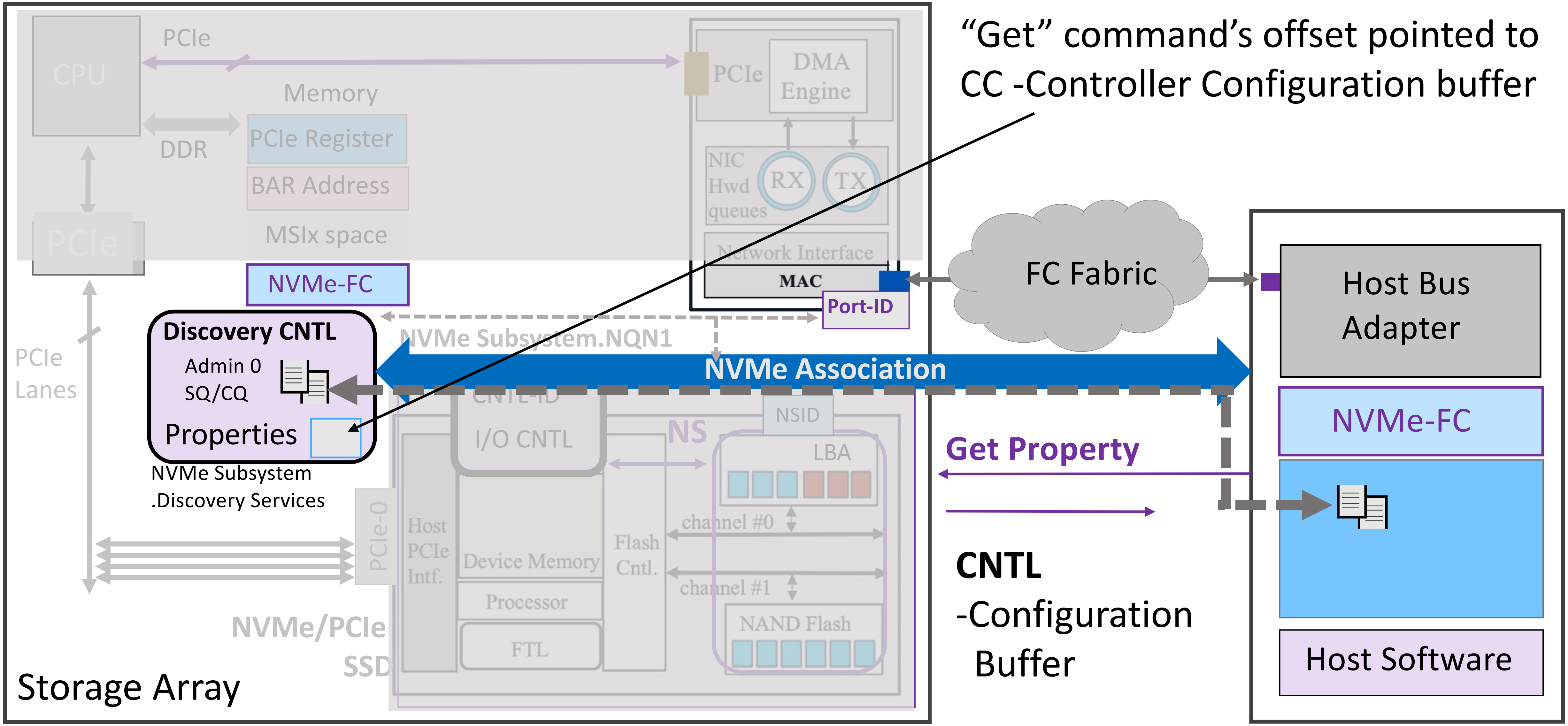
NVMe-FC Protocol Flows (Connect Response CQE)

If a connection is established, then the Controller ID allocated to the host is returned.



Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (Get Property CC)

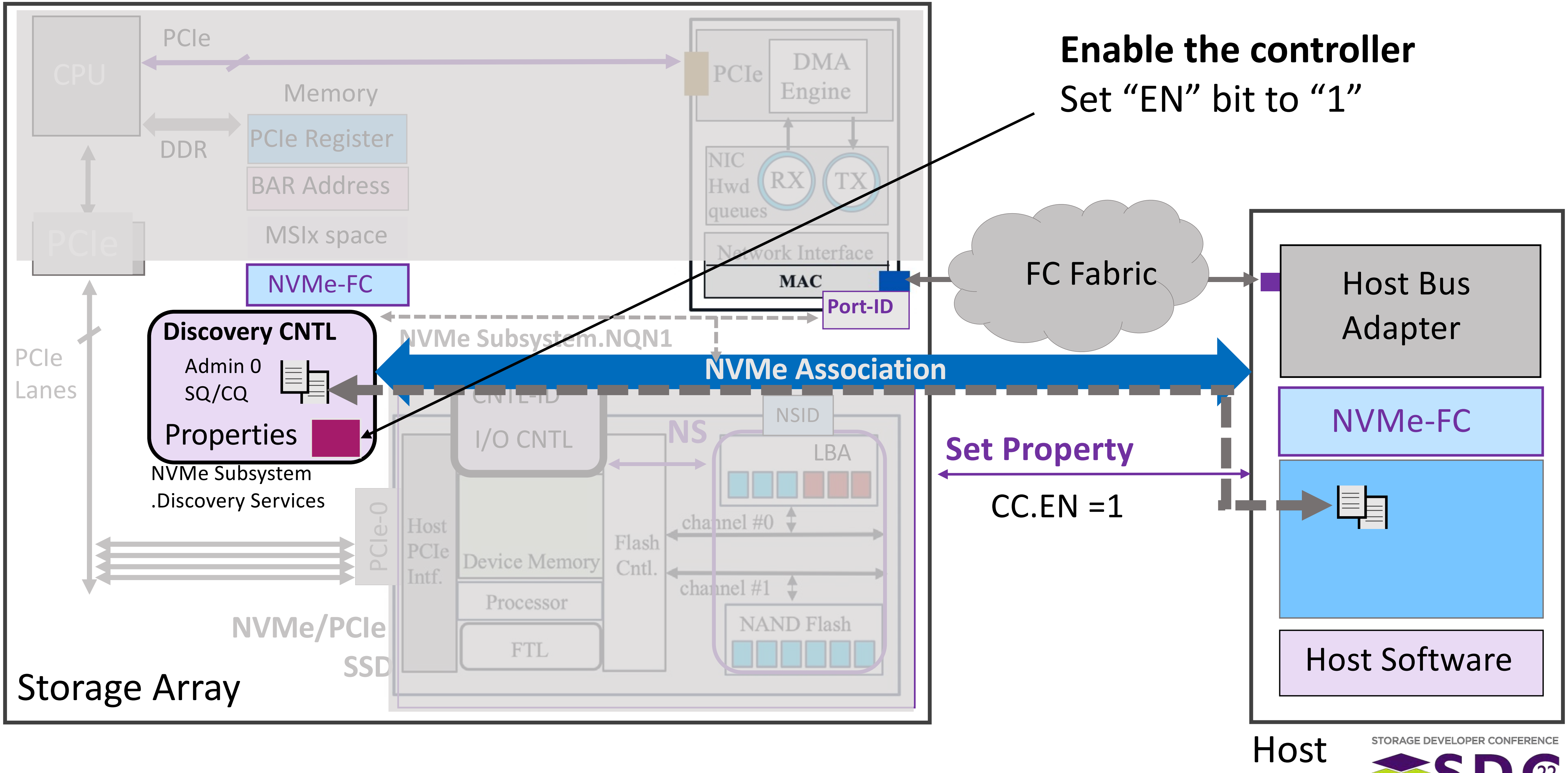


"Get" command's offset pointed to CC -Controller Configuration buffer

CNTL
-Configuration
Buffer

Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (Set Property CC.EN)



Enable the controller

Set "EN" bit to "1"

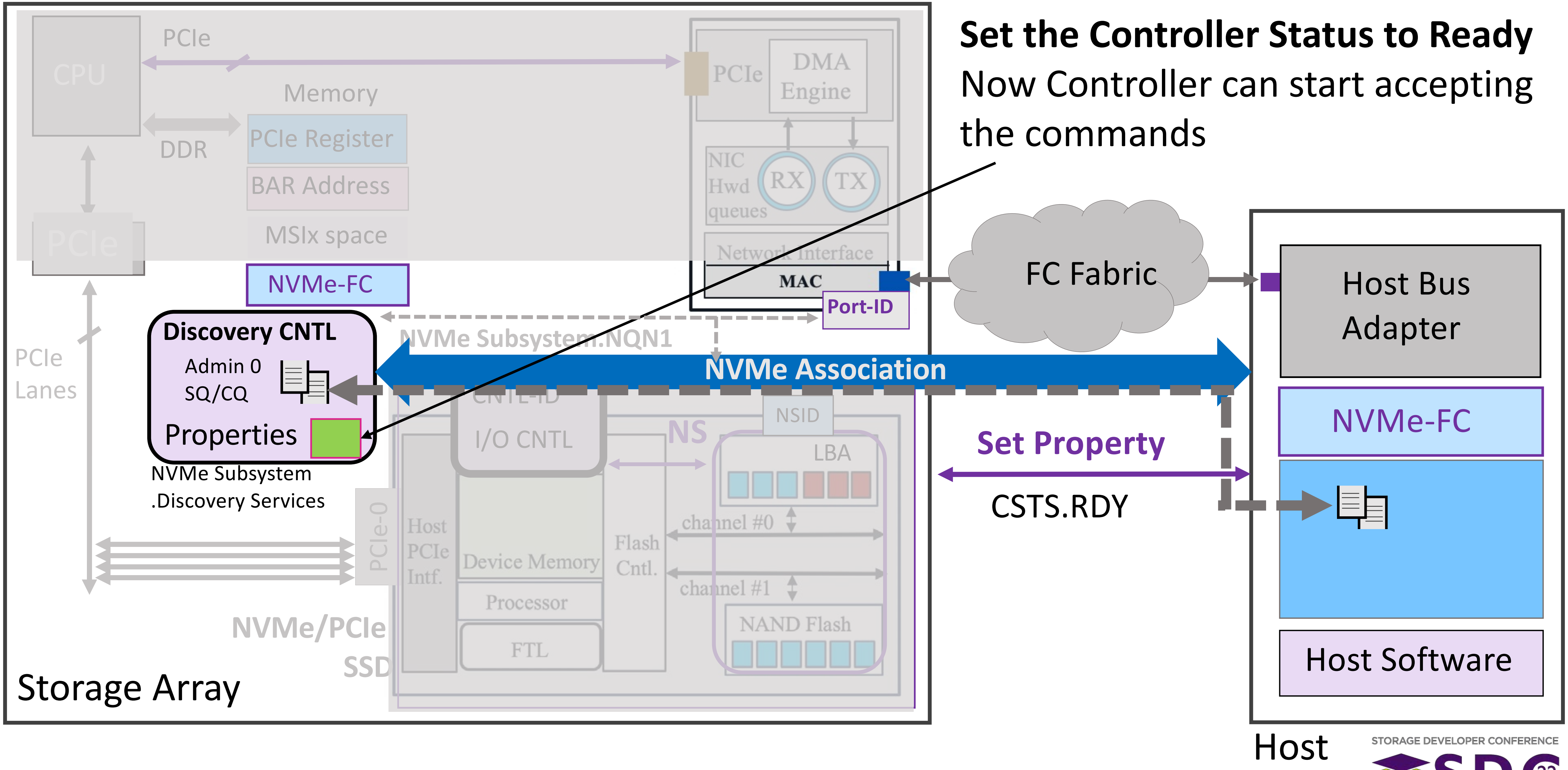
 |

Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (Set Property CSTS.RDY)

Set the Controller Status to Ready

Now Controller can start accepting the commands

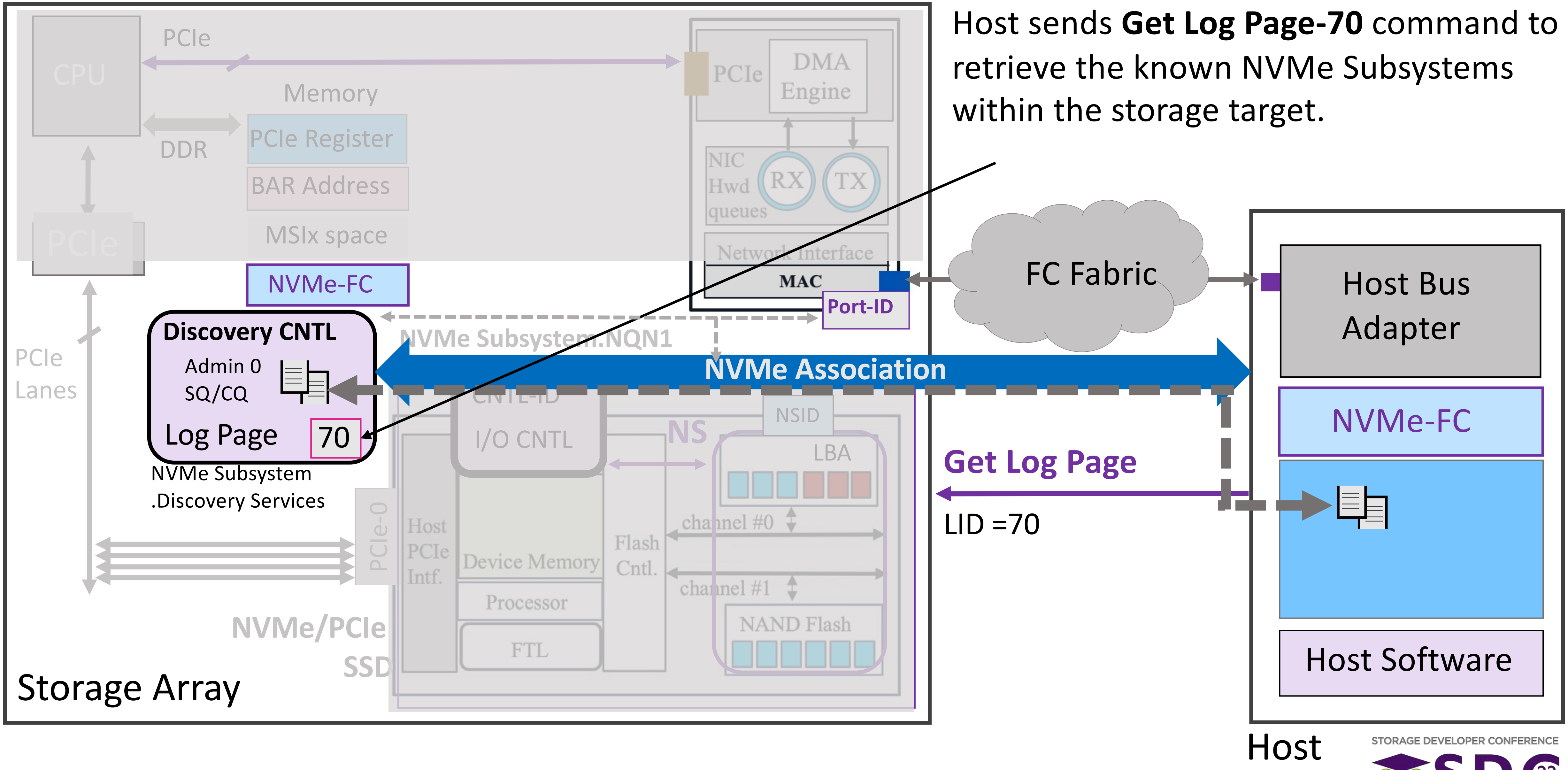


 |

Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (Get Log Page)

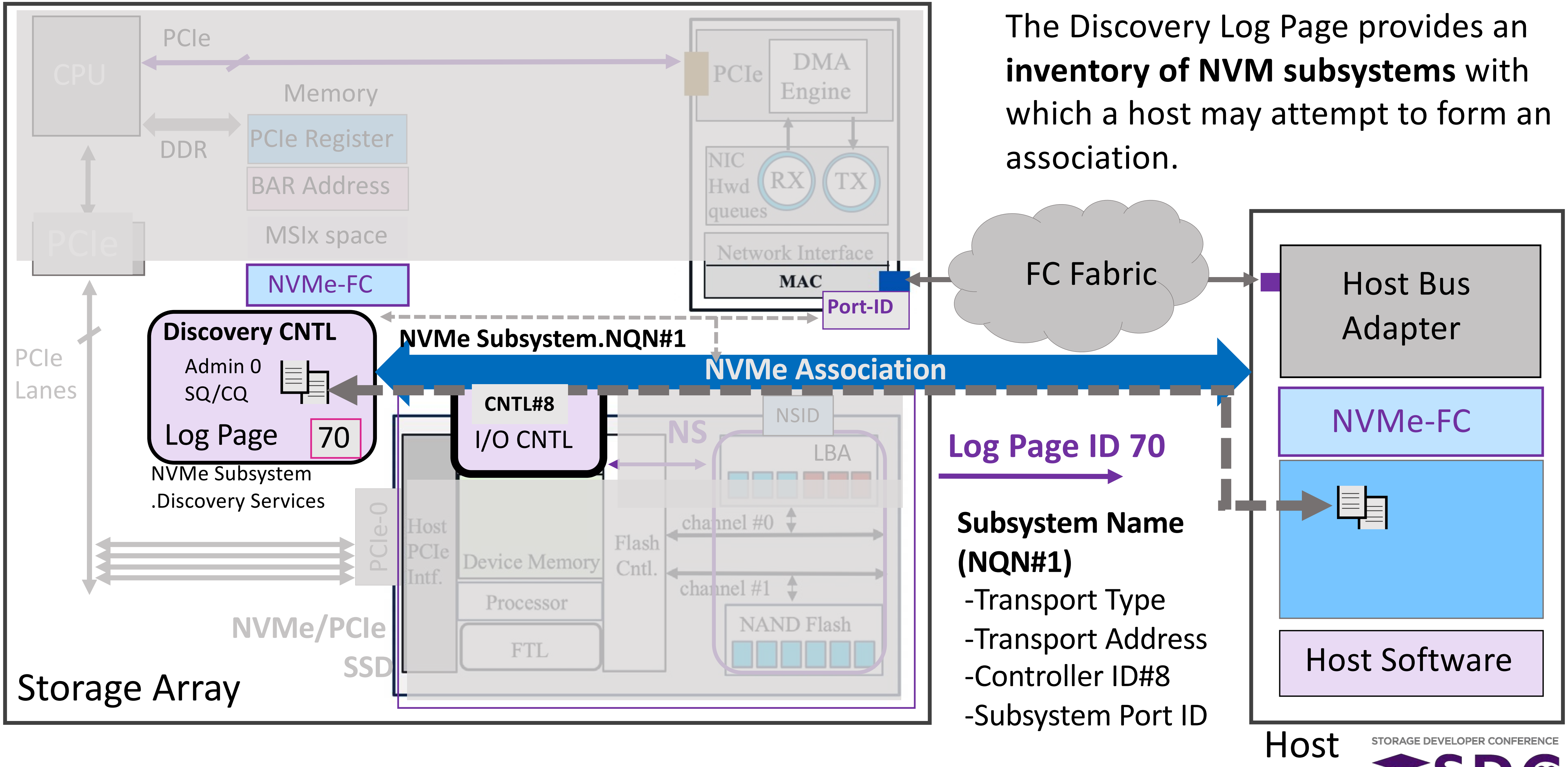
Host sends **Get Log Page-70** command to retrieve the known NVMe Subsystems within the storage target.



Kamal Bakshi, Cisco

 |

NVMe-FC Protocol Flows (Get Log Page)

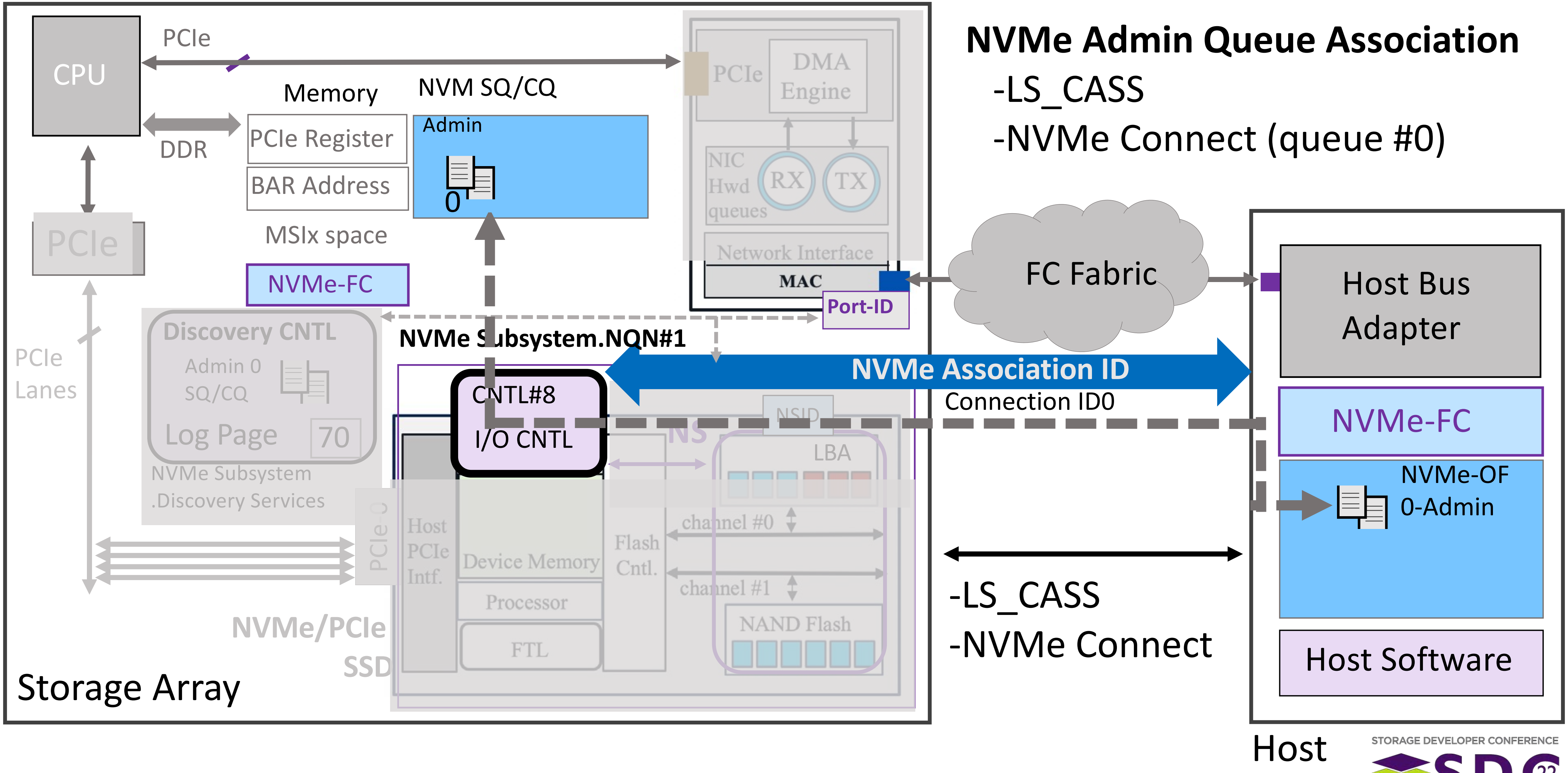


The Discovery Log Page provides an **inventory of NVM subsystems** with which a host may attempt to form an association.

Subsystem Name (NQN#1)
- -Transport Type
- -Transport Address
- -Controller ID#8
- -Subsystem Port ID

Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (Create Association with I/O CNTL)



NVMe Admin Queue Association

- -LS_CASS
- -NVMe Connect (queue #0)

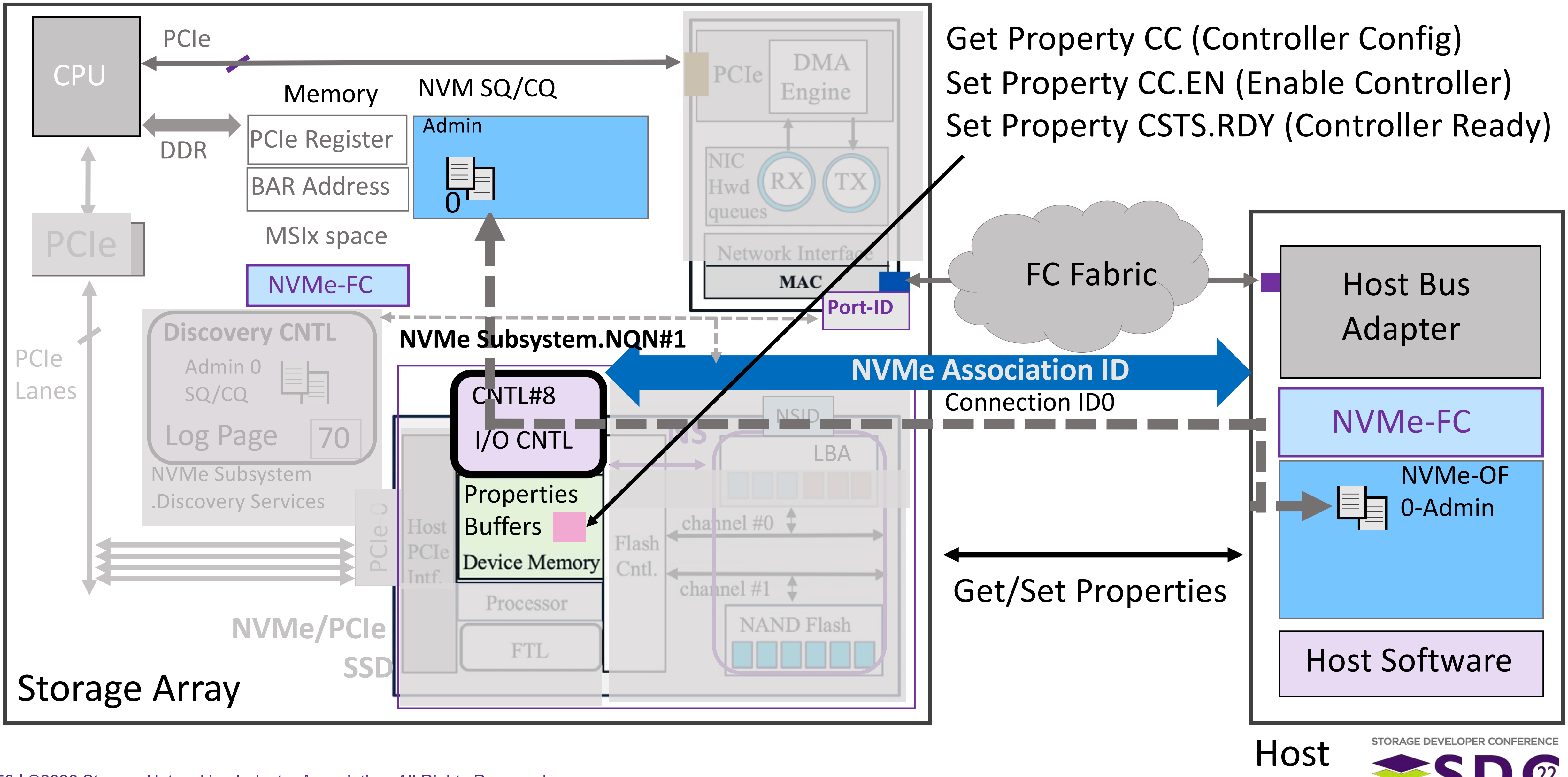
-LS_CASS

-NVMe Connect

Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (I/O CNTL Ready to accept commands)

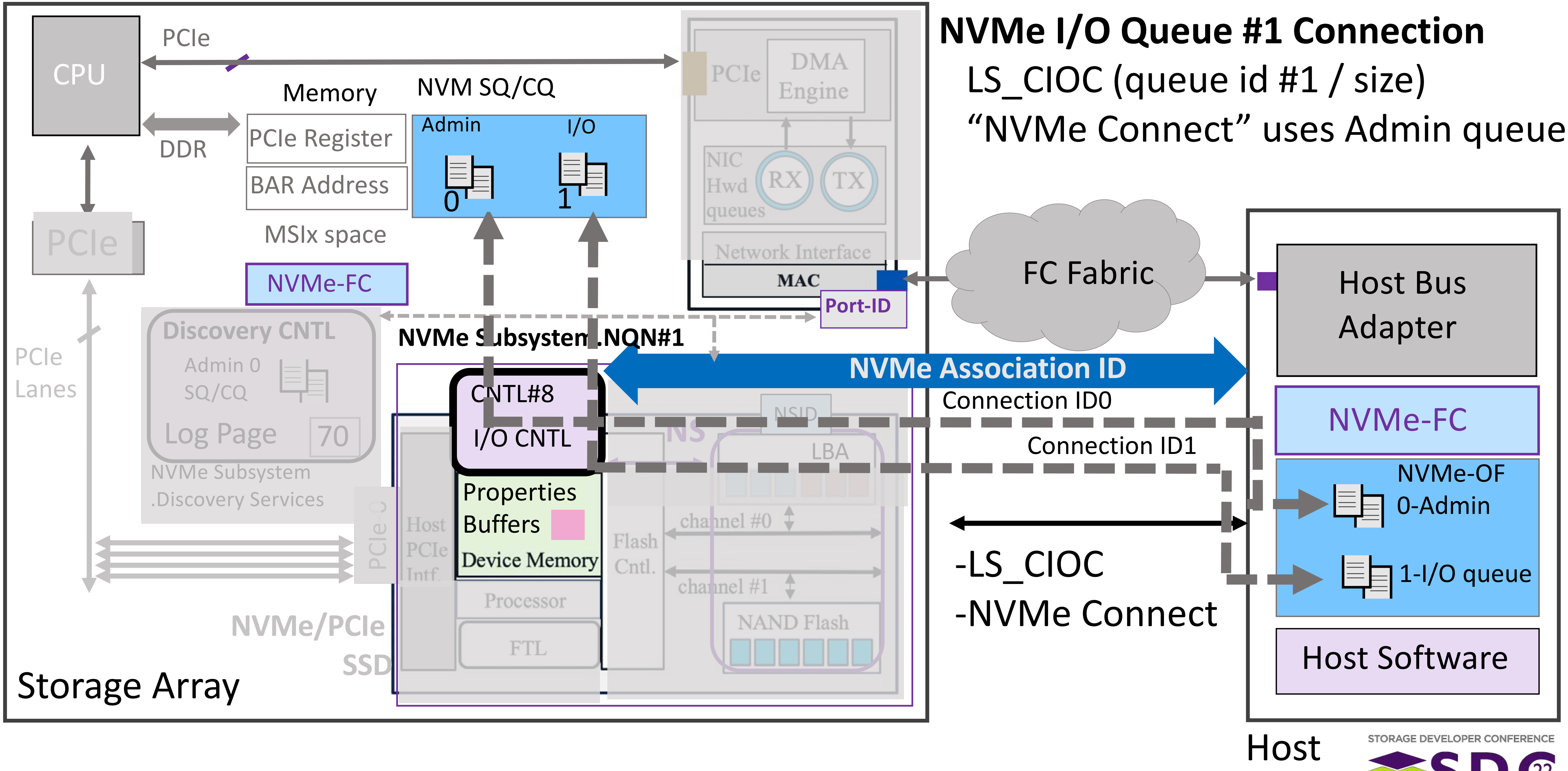
Get Property CC (Controller Config)
Set Property CC.EN (Enable Controller)
Set Property CSTS.RDY (Controller Ready)



 |

Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (Create I/O Queues)



NVMe I/O Queue #1 Connection

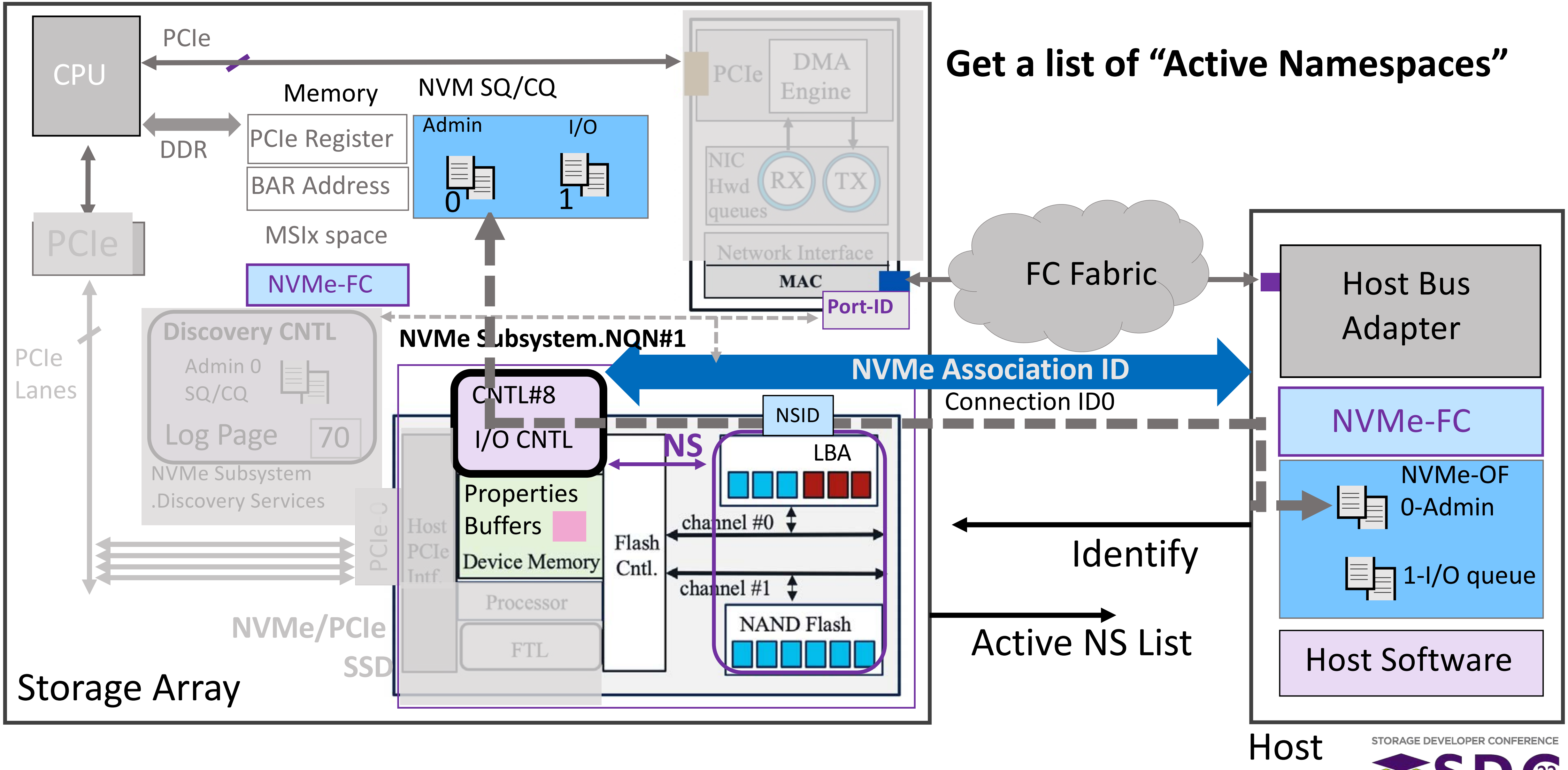
LS_CIOC (queue id #1 / size)

"NVMe Connect" uses Admin queue

Kamal Bakshi, Cisco

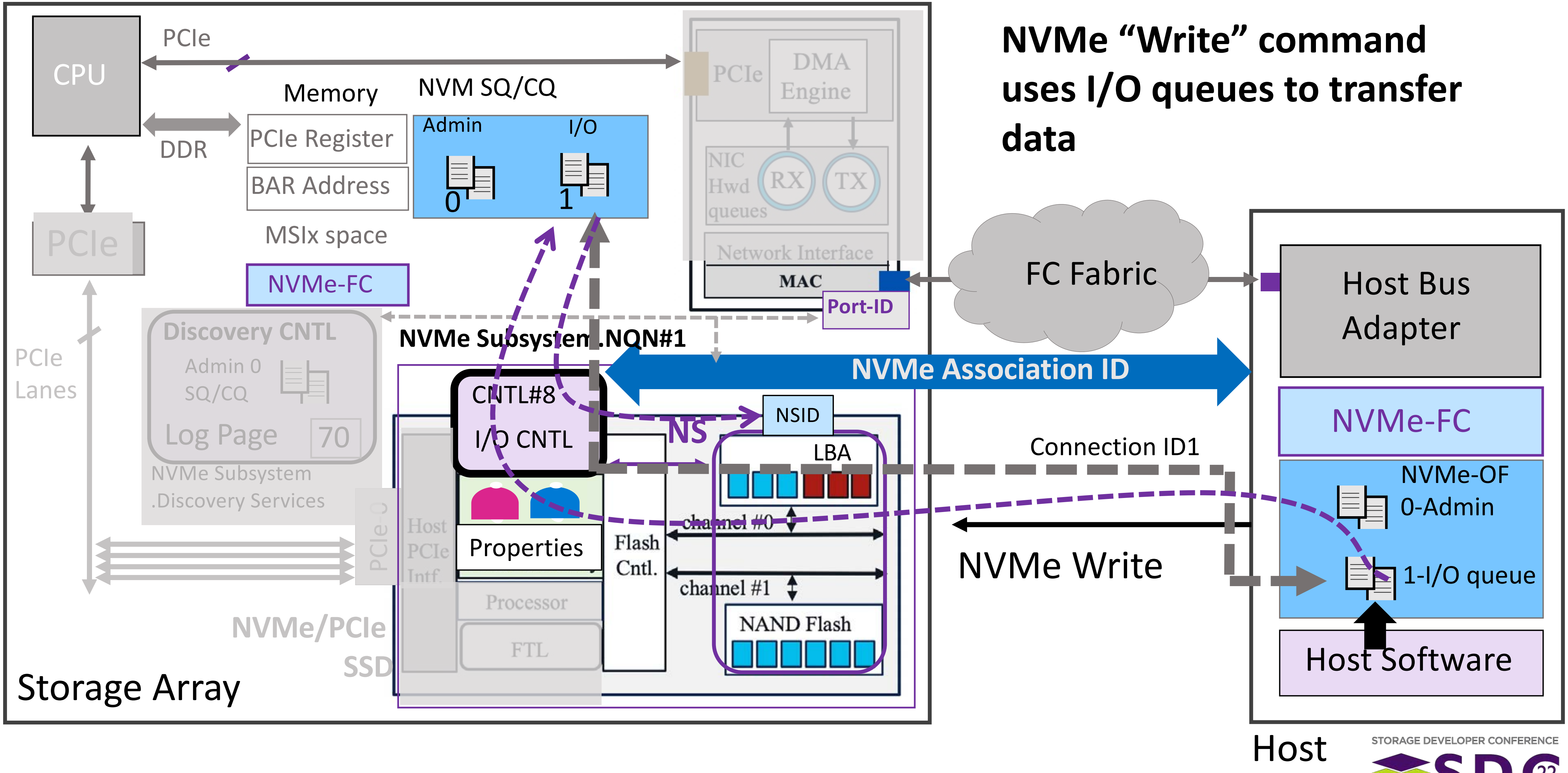
 |

NVMe-FC Protocol Flows (NVMe Identify CNS 02)



Kamal Bakshi, Cisco

NVMe-FC Protocol Flows (NVMe Write)

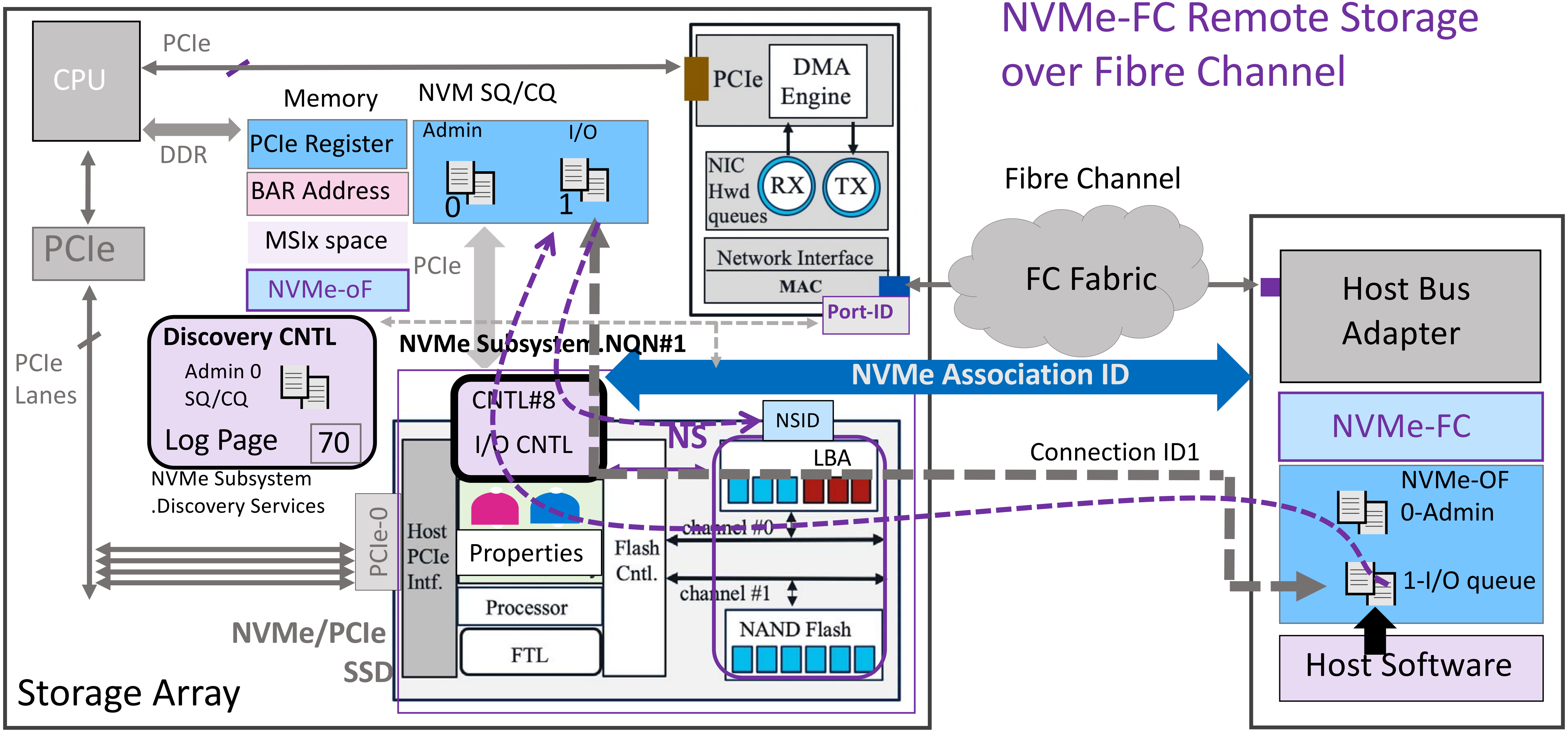


NVMe "Write" command uses I/O queues to transfer data

Kamal Bakshi, Cisco

NVMe-FC

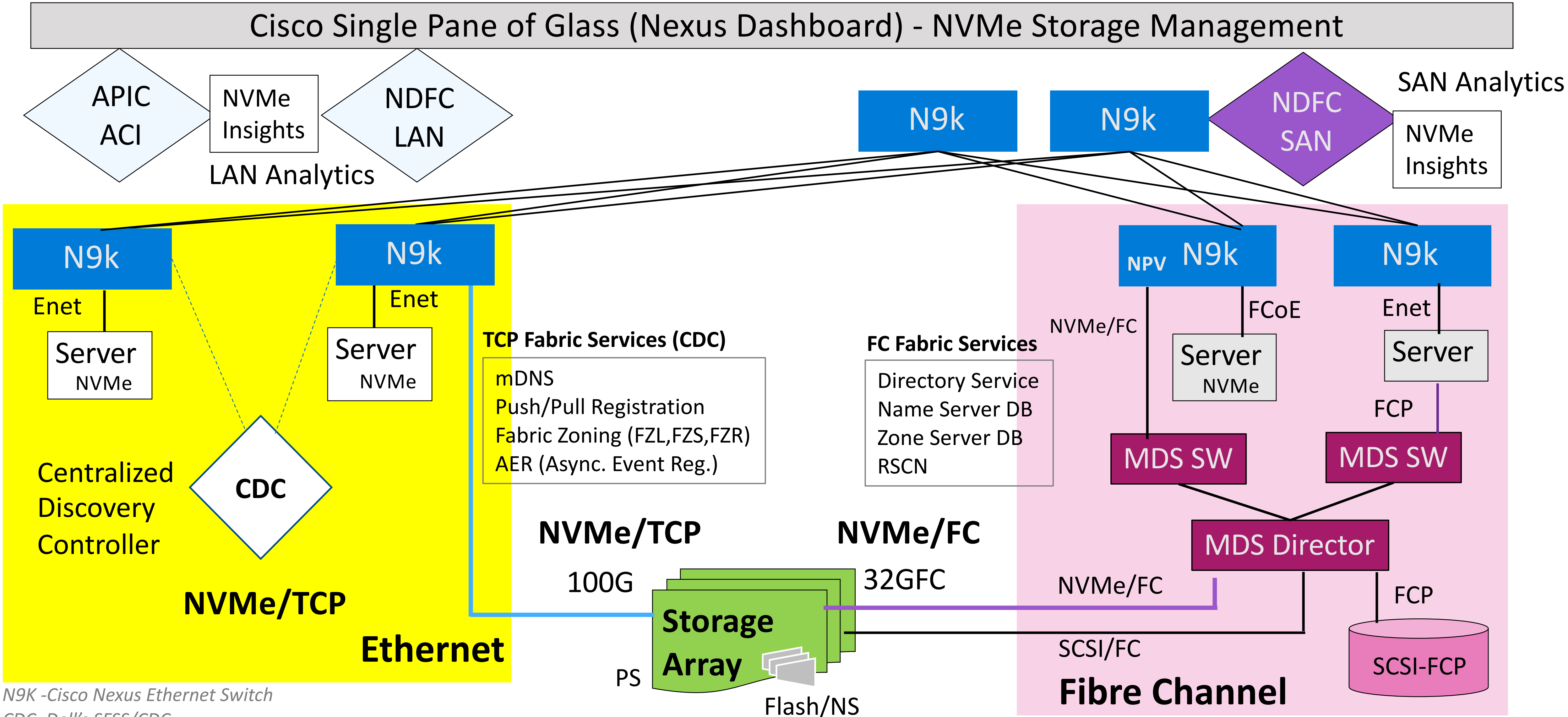
NVMe-FC Remote Storage over Fibre Channel



Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

NVMe/TCP Architecture



N9K -Cisco Nexus Ethernet Switch
CDC: Dell's SFSS/CDC

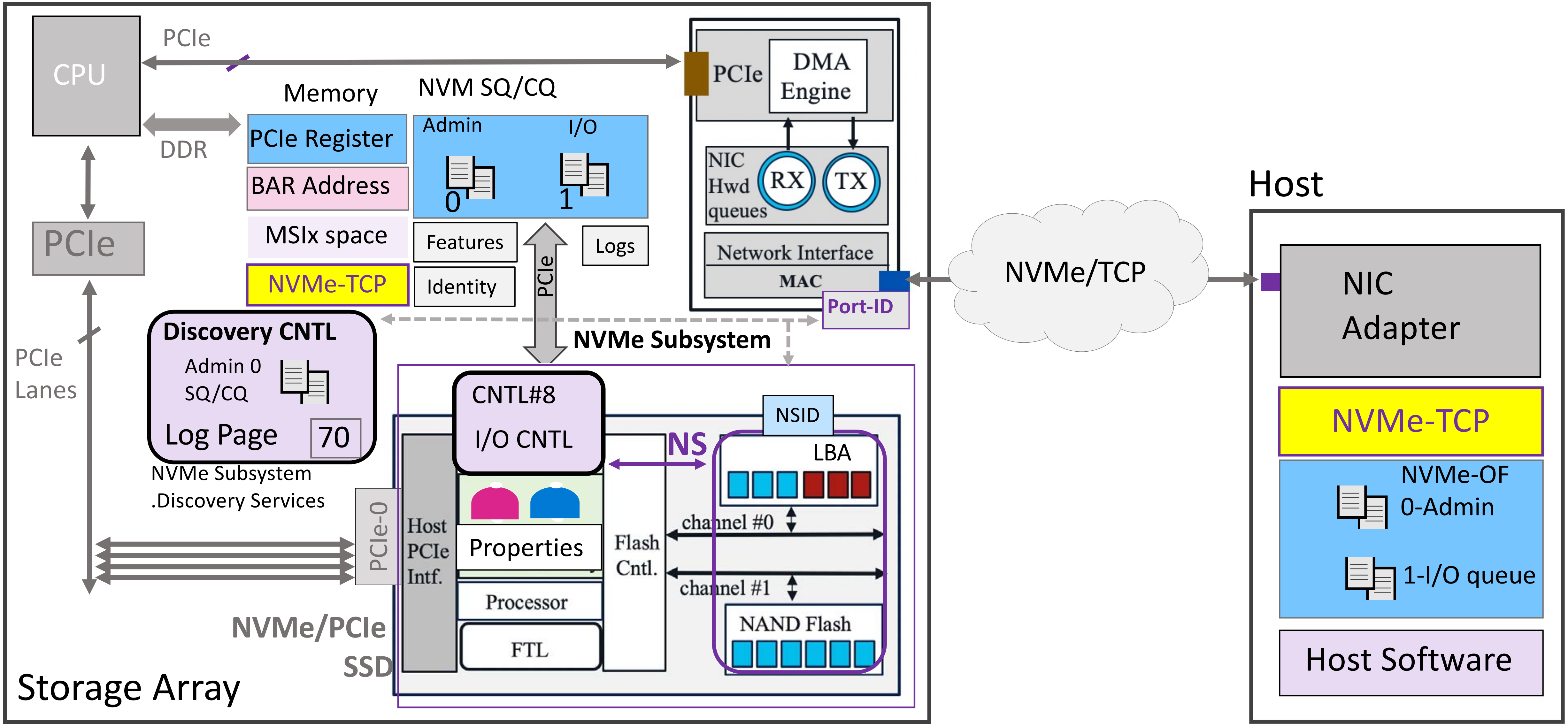
MDS -Cisco Fibre Channel Switch
PS -Dell PowerStore All-Flash Storage

Kamal Bakshi, Cisco

NVMe/TCP Architecture

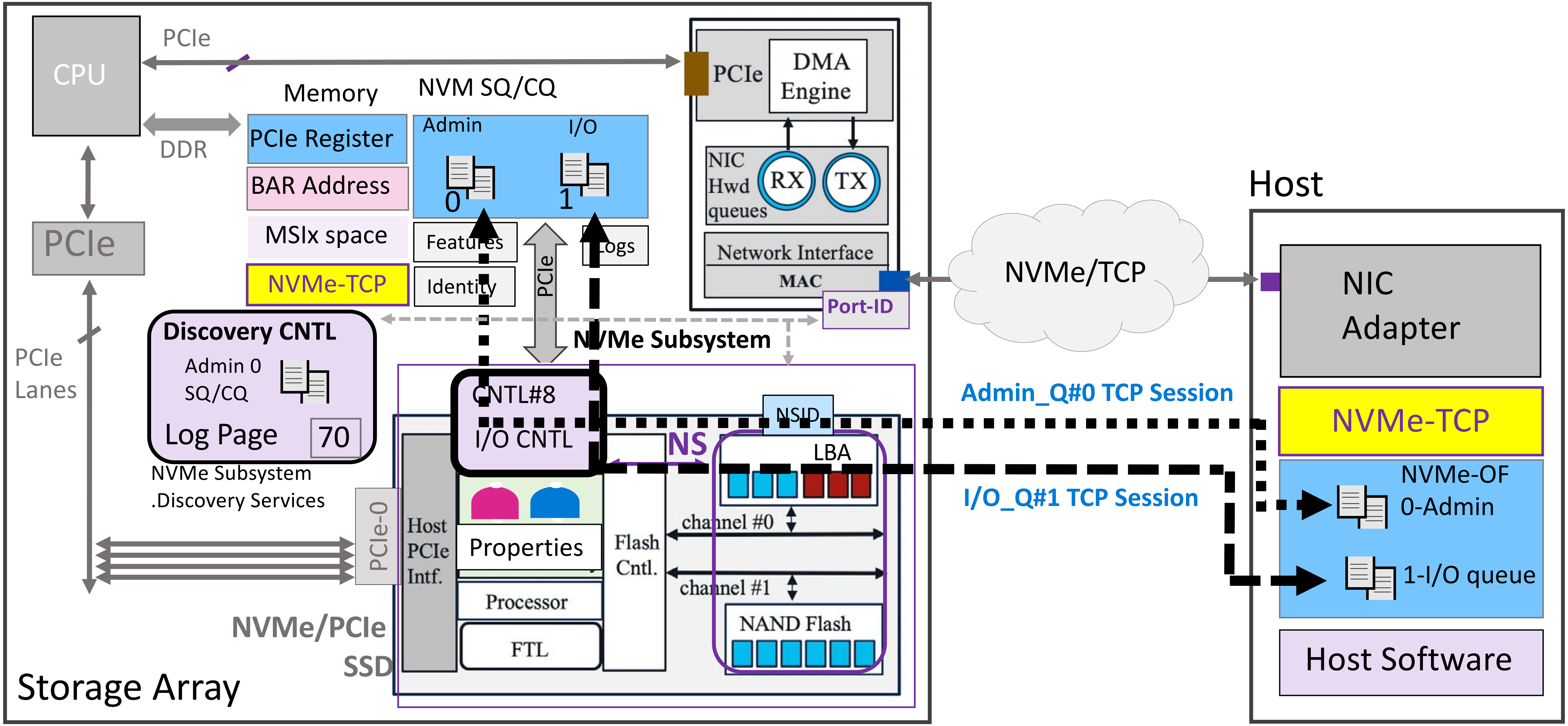
Kamal Bakshi, Cisco

NVMe/TCP Fabric (Storage⟷Host)



Kamal Bakshi, Cisco

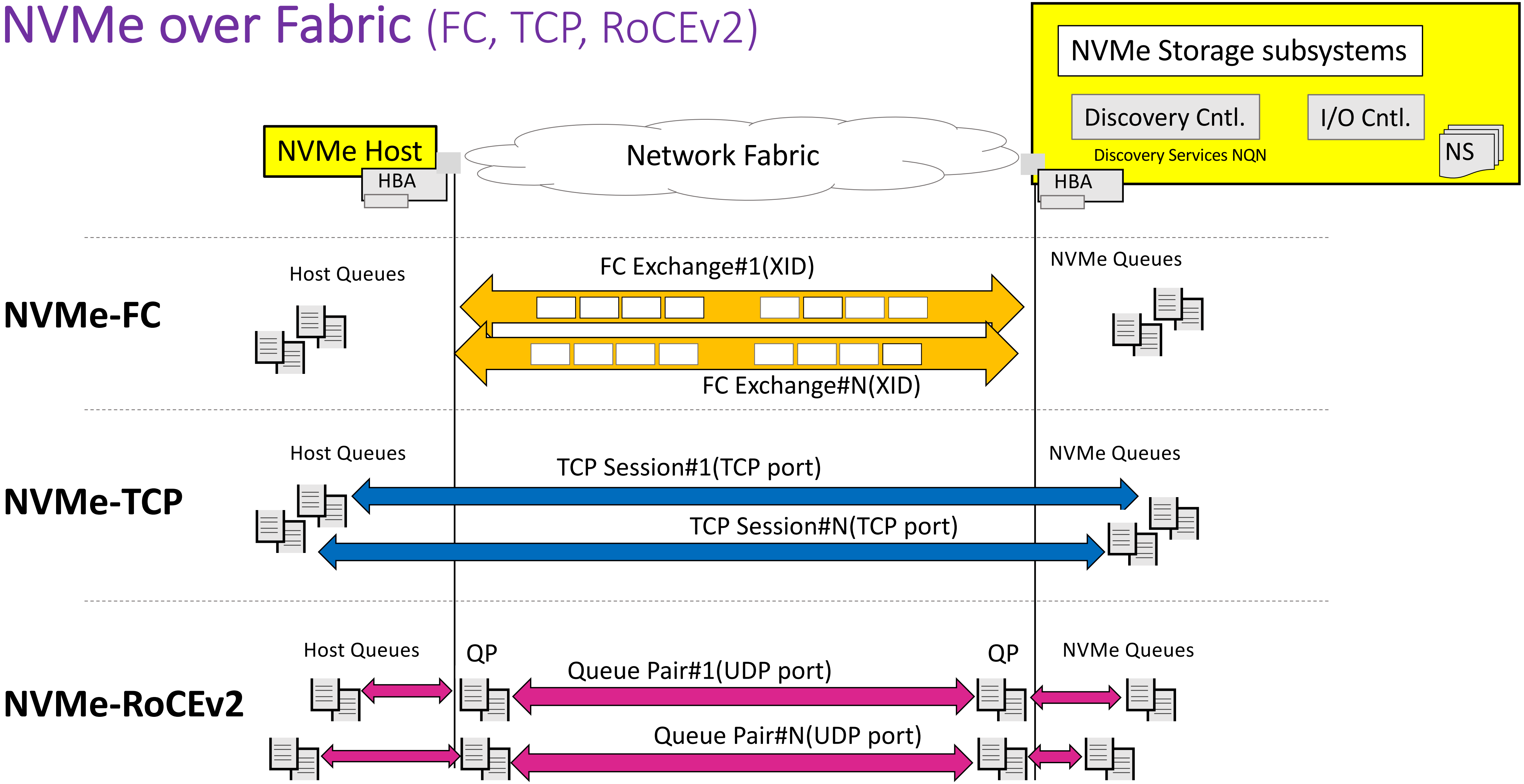
NVMe/TCP Fabric (Storage ⟷ Host)



Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

NVMe over Fabric (FC, TCP, RoCEv2)



Kamal Bakshi, Cisco

NVMe-TCP Port Numbers

TCP port 4420 has been assigned for use by NVMe over Fabrics

TCP port 8009 has been assigned by IANA for use by NVMe over Fabrics discovery. TCP port 8009 is the default TCP port for NVMe/TCP discovery controllers.

There is no default TCP port for NVMe/TCP I/O controllers, the Transport Service Identifier (TRSVCID) field in the Discovery Log Entry indicates the TCP port to use.
The TCP ports that may be used for NVMe/TCP I/O controllers include TCP port 4420, and the Dynamic and/or Private TCP ports (i.e., ports in the TCP port number range from 49152 to 65535). NVMe/TCP I/O controllers should not use TCP port 8009. TCP port 4420 shall not be used for both NVMe/iWARP and NVMe/TCP at the same IP address on the same network.

The TRSVCID field in a Discovery Log Entry for the NVMe/TCP transport shall contain a TCP port number in decimal representation as an ASCII string. If such a TRSVCID field does not contain a TCP port number in decimal representation as an ASCII string, then the host shall not use the information in that Discovery Log Entry to connect to a controller.

source: NVMe Specifications

Kamal Bakshi, Cisco

NVMe/TCP -(11) Types of PDUs

The PDU header (HDR) consists of a PDU common header **(CH)** which has a fixed length of 8 bytes and a PDU specific header **(PSH)** which has a variable length

CH Common Header (8 bytes) **PSH**

Enet hdr	IP hdr	TCP hdr	PDU Opcode (xx)	Flags	hlen	PDU offset	PDU len	Packet Specific Header	*HDGST*	PAD	Data	*DDGST*	Enet FCS
			00	01	02	03	04 07						

TCP Port#

8009 NVMe/TCP Discovery
4420 NVMe over Fabric
No default NVMe/TCP I/O cntl.

A host and a controller in an NVM subsystem communicate over TCP by exchanging NVMe/TCP Protocol Data Units (NVMe/TCP PDUs).

An NVMe/TCP PDU may be used to transfer a capsule, data, or control/status information.

(xx) PDU Opcode

00-ICReq -H2C
01-ICResp -C2H

02-H2CTermReq
03-C2HTermReq

04-CapsuleCmd -H2C
05-CapsuleResp -C2H

06-H2CData
07-C2HData

09-R2T-C2H

0A-KDReq
0B-KDResp

NVMe/TCP PDU Types

PDU Name	Opcode by field: Function (07:01)	Opcode by field: PDU Direction (00) [1]	Combined Opcode [2]	Section	PDU Description
ICReq	0000000b	0b	00h	3.6.2.2	**Initialize Connection Request:** A PDU sent from a host to a controller to communicate NVMe/TCP connection parameters and establish an NVMe/TCP connection
ICResp	0000000b	1b	01h	3.6.2.3	**Initialize Connection Response:** A PDU sent from a controller to a host to accept a connection request and communicate NVMe/TCP connection parameters
H2CTermReq	0000001b	0b	02h	3.6.2.4	**Host to Controller Terminate Connection Request:** A PDU sent from a host to a controller in response to a fatal transport error
C2HTermReq	0000001b	1b	03h	3.6.2.5	**Controller to Host Terminate Connection Request:** A PDU sent from a controller to a host in response to a fatal transport error
CapsuleCmd	0000010b	0b	04h	3.6.2.6	**Command Capsule:** A PDU sent from a host to a controller to transfer an NVMe over Fabrics Command Capsule
CapsuleResp	0000010b	1b	05h	3.6.2.7	**Response Capsule:** A PDU sent from a controller to a host to transfer an NVMe over Fabrics Response Capsule
H2CData	0000011b	0b	06h	3.6.2.8	**Host to Controller Data:** A PDU sent from a host to a controller to transfer data to the controller
C2HData	0000011b	1b	07h	3.6.2.9	**Controller to Host Data:** A PDU sent from a controller to a host to transfer data to the host
R2T	0000100b	1b	09h	3.6.2.10	**Ready to Transfer:** A PDU sent from a controller to a host to indicate that the controller is ready to accept data

(0A) Kickstart Discovery Request: A PDU sent from a DDC to a CDC to request a pull registration and communicate connection parameters to be used during the subsequent pull registration.

(0B) Kickstart Discovery Response: A PDU sent from a CDC to a DDC to accept a pull registration request and connection parameters.

NVMe/TCP facilitates an optional PDU Header Digest (HDGST) and Data Digest (DDGST). The presence of each digest is negotiated at the connection establishment.

source: NVM Express TCP Transport Specification 1.0b

Kamal Bakshi, Cisco

PDU Type (00) ICReq -Initiate Connection Request

			CH Common Header					PSH				
Enet hdr	IP hdr	TCP hdr	PDU Opcode (**00**)	Flags	hlen	PDU reser	PDU len	Packet Specific Header	*HDGST*	MAXR2T	reserved	Enet FCS
			00	01	02	03	04	08	11	12	16	127

An NVMe Transport connection is established between a host and an NVM subsystem prior to the transfer of any capsules or data.
The mechanism used to establish an NVMe Transport connection is NVMe Transport specific and defined by the corresponding NVMe Transport binding specification.

The first step is to establish a TCP connection between a host and a controller. A controller acts as the passive side of the TCP connection and is set to “listen” for host-initiated TCP connection establishment requests.

Once a TCP connection has been established, the host sends an Initialize Connection Request (ICReq) PDU to the controller.

Key Info carried in ICReq

-DDGST/HDGST enable or disable

-Max outstanding R2T (Ready to Transmit)

Opcode: 00 ICReq -Initialize Connection Request

Bytes	PDU Section	Description
00	CH	**PDU-Type:** 00h
01	CH	**FLAGS:** Reserved
02	CH	**HLEN:** Fixed length of 128 bytes (80h).
03	CH	**PDO:** Reserved
07:04	CH	**PLEN:** Fixed length of 128 bytes (80h).
09:08	PSH	**PDU Format Version (PFV):** Specifies the format version of NVMe/TCP PDUs. The format of the record specified in this definition shall be cleared to 0h.
10	PSH	**Host PDU Data Alignment (HPDA):** Specifies the data alignment for all PDUs transferred from the controller to the host that contain data. This value is 0’s based value in units of dwords in the range 0 to 31 (e.g., values 0, 1, and 2 correspond to 4 byte, 8 byte, and 12 byte alignment).
11	PSH	**DGST:** Host PDU header and data digest enable options. **Bits** – **Definition** 7:2 – Reserved 1 – **DDGST_ENABLE:** If set to ‘1’, the use of data digest is requested by the host for the connection. If cleared to ‘0’, data digest shall not be used for the connection. 0 – **HDGST_ENABLE:** If set to ‘1’, the use of header digest is requested by the host for the connection. If cleared to ‘0’, header digest shall not be used for the connection.
15:12	PSH	**Maximum Number of Outstanding R2T (MAXR2T):** Specifies the maximum number of outstanding R2T PDUs for a command at any point in time on the connection. This is a 0’s based value.
127:16	PSH	Reserved

source: NVM Express TCP Transport Specification 1.0b

Kamal Bakshi, Cisco

PDU Type (00) ICReq -example

CH Common Header

PSH

Enet hdr	IP hdr	TCP hdr	PDU Opcode (**00**)	Flags	hlen	PDU reser	PDU len	Packet Specific Header	*HDGST*	MAXR2T	reserved	Enet FCS
			00	01	02	03	04	08	11	12	16	127

ICReq

Initiate Connection Request

```
0000  00 50 56 bf 37 26 50 6b  4b 4b df 3a 08 00 45 00
0010  00 b4 21 75 40 00 40 06  04 6b 09 01 01 2c 09 01
0020  01 37 db cc 1f 49 01 78  d0 ce 02 84 1f e6 80 18
0030  01 f6 8c e0 00 00 01 01  08 0a 31 46 3d de a1 ec
0040  d2 21 00 00 80 00 80 00  00 00 00 00 00 00 00 00
0050  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0060  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0070  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0080  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0090  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00a0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00b0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00c0  00 00 f2 89 d9 92
```

```
▷ Ethernet II, Src: Mellanox_4b:df:3a (50:6b:4b:4b:df:3a), Dst: VMware_bf:37:26 (00:50:56:bf:37:26)
▷ Internet Protocol Version 4, Src: 9.1.1.44, Dst: 9.1.1.55
▷ Transmission Control Protocol, Src Port: 56268, Dst Port: 8009, Seq: 1, Ack: 1, Len: 128
◢ NVM Express Fabrics TCP Discovery Controller
    [Cmd Qid: 0 (AQ)]
    Pdu Type: ICReq (0)
    Pdu Specific Flags: 0x00 Non-Kickstart discovery NVMe/TCP connection
  ◢ Pdu Specific Flags: 0x00
      .... ...0 = PDU Header Digest: Not set
      .... ..0. = PDU Data Digest: Not set
      .... .0.. = PDU Data Last: Not set
      .... 0... = PDU Data Success: Not set
    Pdu Header Length: 128
    Pdu Data Offset: 0
    Packet Length: 128
  ◢ ICReq
      Pdu Version Format: 0
      Host Pdu data alignment: 0
      Digest Types Enabled: 0
      Maximum r2ts per request: 0
```

PDU Opcode

Kamal Bakshi, Cisco

PDU Type (01) ICResp -Initiate Connection Response

			CH Common Header					PSH				
Enet hdr	IP hdr	TCP hdr	PDU Opcode (**01**)	Flags	hlen	PDU offset	PDU len	Packet Specific Header	*HDGST*	MAXH2CDATA	*reserved*	Enet FCS
			00	01	02	03	04	08	11	12	16	127

When a controller receives an ICReq PDU, that controller responds with an Initialize Connection Response **(ICResp)** PDU. The exchange is used to both establish a connection and exchange connection configuration parameters.

When a connection is established, the host and controller are ready to exchange capsules and command data.

Opcode: 01 ICResp -Initialize Connection Response

Bytes	PDU Section	Description
00	CH	**PDU-Type:** 01h
01		**FLAGS:** Reserved
02		**HLEN:** Fixed length of 128 bytes (80h).
03		**PDO:** Reserved
07:04		**PLEN:** Fixed length of 128 bytes (80h).
09:08	PSH	**PDU Format Version (PFV):** Specifies the format version of NVMe/TCP PDUs. The format of the record specified in this definition shall be cleared to 0h.
10		**Controller PDU Data Alignment (CPDA):** Specifies the data alignment for all PDUs that transfer data in addiion to the PDU Header (refer to section 2). This is a 0's based value in units of dwords in the range 0 to 31 (e.g., values 0, 1, and 2 correspond to 4 byte, 8 byte, and 12 byte alignment).
11		**DGST:** Controller PDU header and data digest enable options. **Bits** / **Definition** 7:2 — Reserved 1 — **DDGST_ENABLE:** If set to '1', data digest is used for the connection. If cleared to '0', data digest is not used for the connection. 0 — **HDGST_ENABLE:** If set to '1', header digest is used for the connection. If cleared to '0', header digest is not used for the connection.
15:12		**Maximum Host to Controller Data length (MAXH2CDATA):** Specifies the maximum number of PDU-Data bytes per H2CData PDU in bytes. This value is a multiple of dwords and should be no less than 4,096.
127:16		Reserved

source: NVM Express TCP Transport Specification 1.0b

Kamal Bakshi, Cisco

PDU Type (01) ICResp -example

			CH Common Header					**PSH**				
Enet hdr	IP hdr	TCP hdr	PDU Opcode (**01**)	Flags	hlen	PDU offset	PDU len	Packet Specific Header	*HDGST*	MAXH2CDATA	*reserved*	Enet FCS
			00	01	02	03	04	08	11	12	16	127

Initiate Connection Response

0010 0000 (1048576)

```
0000  50 6b 4b 4b df 3a 00 50  56 bf 37 26 08 00 45 00
0010  00 b4 33 53 40 00 3e 06  f4 8c 09 01 01 37 09 01
0020  01 2c 1f 49 db cc 02 84  1f e6 01 78 d1 4e 80 18
0030  01 f6 7b 5e 00 00 01 01  08 0a a1 ec d2 23 31 46
0040  3d de 01 00 80 00 80 00  00 00 00 00 00 00 00 00
0050  10 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0060  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0070  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0080  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0090  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00a0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00b0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00c0  00 00 13 e9 19 f5
```

```
▷ Ethernet II, Src: VMware_bf:37:26 (00:50:56:bf:37:26), Dst: Mellanox_4b:df:3a (50:6b:4b:4b:df:3a)
▷ Internet Protocol Version 4, Src: 9.1.1.55, Dst: 9.1.1.44
▷ Transmission Control Protocol, Src Port: 8009, Dst Port: 56268, Seq: 1, Ack: 129, Len: 128
◢ NVM Express Fabrics TCP Discovery Controller
    [Cmd Qid: 0 (AQ)]
    Pdu Type: ICResp (1)
    Pdu Specific Flags: 0x00
  ◢ Pdu Specific Flags: 0x00
      .... ...0 = PDU Header Digest: Not set
      .... ..0. = PDU Data Digest: Not set
      .... .0.. = PDU Data Last: Not set
      .... 0... = PDU Data Success: Not set
    Pdu Header Length: 128
    Pdu Data Offset: 0
    Packet Length: 128
  ◢ ICResp
      Pdu Version Format: 0
      Controller Pdu data alignment: 0
      Digest types enabled: 0
      Maximum data capsules per r2t supported: 1048576
```


Kamal Bakshi, Cisco

PDU Type (0A) KDReq -Kickstart Discovery Request

Enet hdr	IP hdr	TCP hdr	PDU Opcode (**0A**)	Flags	hlen	PDU offset	PDU len	# of Kickstart Records	NUMKR	HDGST	KRN/Data	DDGST	Enet FCS
			00	01	02	03	04	08	10	12.	16		

CH Common Header · **PSH** · **PDU Data**

# of Discovery Info Entries

Kickstart Record #1, #2, #3 ....

KDReq (Kickstart Discovery Request)

Bytes		PDU Section	Description
00		CH	**PDU-Type:** 0Ah
01		CH	**FLAGS:** Bits 7:2 — Reserved Bit 1 — **DDGSTF:** If set to '1', then the DDGST field follows the PDU Data and contains a valid value. If cleared to '0', then the DDGST field is not present. Bit 0 — **HDGSTF:** If set to '1', then the HDGST field follows the PDU Header and contains a valid value. If cleared to '0', then the HDGST field is not present.
02		CH	**HLEN:** Fixed length of 12 bytes (i.e., Ch).
03		CH	**PDO:** Data offset within PDU. This value shall be a multiple of the data alignment specified by the CPDA field in the ICResp PDU (refer to section 3.6.2.3) that was previously sent by the CDC on this TCP connection.
07:04		CH	**PLEN:** Variable length in bytes. If HDGST and DDGST are both not present, then the length will be (NUMKR * 290 bytes) + 12 bytes. If only HDGST or only DDGST is present, then the length will be (NUMKR * 290 bytes) + 16 bytes. If HDGST and DDGST are both present, then the length will be (NUMKR * 290 bytes) + 20 bytes.
09:08		PSH	**Number of Kickstart Records (NUMKR):** This field specifies the number kickstart records included in the PDU DATA field.
11:10		PSH	**Number of Discovery Information Entries (NUMDIE):** This field specifies the maximum number of discovery information entires that the DDC is expected to return if a pull registration is requested. This field shall be cleared to 0h if a pull de-registration is being requested. Refer to the Pull Registrations and Pull De-Registrations section in the NVMe Base Specification.
HDGSTF=1	HDGSTF=0		
15:12	Not present	HDGST	**HDGST:** If HDGSTF is set to '1' in the FLAGS field, this field is valid and contains the header digest (refer to section 3.3.1.1). If the HDGSTF bit is cleared to '0', then this field is not present.
305:16	301:12	DATA	**Kickstart Record 0 (KR0):** This field specifies the first kickstart record as defined in Figure NEW.B.
...	...	DATA	...
(((N + 1) * 290) – 1) + 16:(N * 290) + 16	(((N + 1) * 290) – 1) + 12:(N * 290) + 12	DATA	**Kickstart Record N (KRN):** This field specifies the Nth kickstart record as defined in Figure NEW.B (if present).
DDGSTF=1	DDGSTF=0		
M + 3:M	Not Present	DDGST	**DDGST:** If DDGSTF is set to '1' in the FLAGS field, this field is valid and contains the data digest (refer to section 3.3.1.1). If the DDGSTF bit is cleared to '0', then this field is not present.

source: NVM Express Specification

Kickstart Records

Bytes	Description
00	**Transport Type (TRTYPE):** This field specifies the transport type as defined in the Transport Type (TRTYPE) field in the Discovery Log Page Entry data structure in the NVMe Base Specification.
01	**Address Family (ADRFAM):** This field specifies the address family as defined in the Address Family (ADRFAM) field in the Discovery Log Page Entry data structure in the NVMe Base Specification.
33:02	**Transport Service Identifier (TRSVCID):** This field specifies the NVMe Transport service identifier as an ASCII string as defined in the Transport Service Identifier (TRSVCID) field of the Discovery Log Page Entry data structure in the NVMe Base Specification.
289:34	**Transport Address (TRADDR):** This field specifies the address of the DDC that may be used for a Connect command as an ASCII string as defined in the Transport Address (TRADDR) field of the Discovery Log Page Entry data structure in the NVMe Base Specification.

source: NVM Express Specification

For kickstart discovery, the CDC acts as the passive side of the TCP connection and is set to "listen" for DDC-initiated TCP connection establishment requests. The IP address used by the DDC to establish the TCP connection with the CDC may be obtained from the A record provided in an mDNS response from the CDC.

Kamal Bakshi, Cisco

PDU Type (0A) KDReq -example

	CH Common Header				PSH			PDU Data		
Enet hdr	IP hdr	TCP hdr	PDU Opcode (**0A**)	Flags	hlen	PDU offset	PDU len	# of Kickstart Records	NUMKR	HDGST	KRN/Data	DDGST	Enet FCS
			00	01	02	03	04	08	10	12.	16		

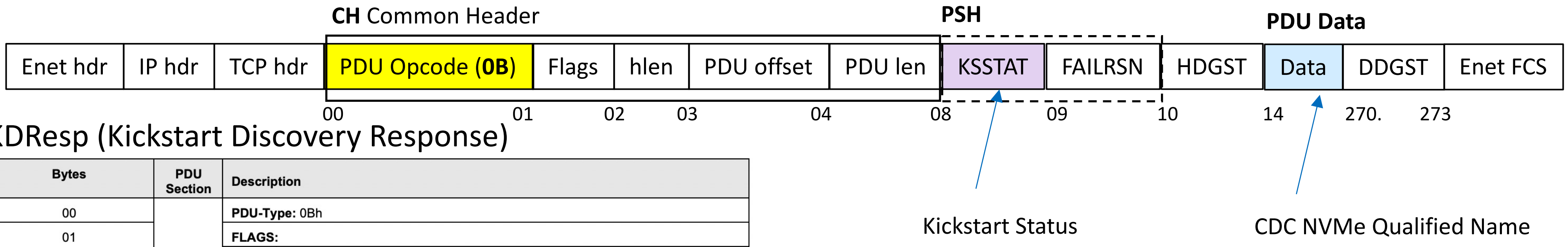
```
Ethernet II, Src: 6a:3e:af:cd:f8:a7 (6a:3e:af:cd:f8:a7), Dst: VMware_bf:62:41 (00:50:56:bf:62:41)
802.1Q Virtual LAN, PRI: 0, DEI: 0, ID: 900
Internet Protocol Version 4, Src: 9.1.1.9, Dst: 9.1.1.55
Transmission Control Protocol, Src Port: 58174, Dst Port: 8009, Seq: 129, Ack: 129, Len: 302
NVM Express Fabrics TCP Discovery Controller
  Pdu Type: Kickstart Discovery Request (10)
  Pdu Specific Flags: 0x00
  Pdu Specific Flags: 0x00
  Pdu Header Length: 12
  Pdu Data Offset: 12
  Packet Length: 302
  KDReq
    Number of Kickstart Records (NUMKR): 1
    Reserved: 0x0001
    Kickstart Record (KR): 03013830303900000000000000000000000000000000000000000000000000000000392e.
      Transport Type (TRTYPE): 3
      Address Family (ADRFAM): 1
      Transport Service Identifier (TRSVCID): 8009
      Transport Address (TRADDR):: 9.1.1.9
```

Kickstart Discovery Request

```
0000  00 50 56 bf 62 41 6a 3e  af cd f8 a7 81 00 03 84
0010  08 00 45 00 01 62 cc 04  40 00 40 06 59 50 09 01
0020  01 09 09 01 01 37 e3 3e  1f 49 59 b7 9a f8 62 76
0030  35 9f 80 18 00 3e 72 b1  00 00 01 01 08 0a 7f ea
0040  73 a7 da fa 0a 7a 0a 00  0c 0c 2e 01 00 00 01 00
0050  01 00 03 01 38 30 30 39  00 00 00 00 00 00 00 00
0060  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0070  00 00 00 00 39 2e 31 2e  31 2e 39 00 00 00 00 00
0080  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0090  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00a0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00b0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00c0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00d0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00e0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
```


Kamal Bakshi, Cisco

PDU Type (0B) KDResp -Kick Start Discovery Response



KDResp (Kickstart Discovery Response)

Bytes	PDU Section	Bits	Description
00	CH		**PDU-Type:** 0Bh
01	CH		**FLAGS:**
		Bits	**Description**
		7:2	Reserved
		1	**DDGSTF:** If set to ‘1’, then the DDGST field follows the PDU Data and contains a valid value. If cleared to ‘0’, then the DDGST field is not present
		0	**HDGSTF:** If set to ‘1’, then the HDGST field follows the PDU Header and contains a valid value. If cleared to ‘0’, then the HDGST field is not present.
02	CH		**HLEN:** Fixed length of 10 bytes (i.e., Ah).
03	CH		**PDO:** Data offset within PDU. This value shall be a multiple of the data alignment specified by the HPDA field in the ICReq PDU (refer to section 3.6.2.2) that was previously sent by the DDC on this TCP connection.
07:04	CH		**PLEN:** Fixed length of 274 bytes (i.e., 112h).
08	PSH		**Kickstart Status (KSSTAT):**
		Bits	**Description**
		7:3	Reserved
		1	**FAILURE:** If set to ‘1’, then the CDC shall not perform a pull registration due to the reason indicated in the Failure Reason (FAILRSN) field.
		0	**SUCCESS:** If set to ‘1’, then the CDC shall perform a pull registration.
09	PSH		**Failure Reason (FAILRSN):**
		Bits	**Description**
		7	Reserved
		6	Insufficient Discovery Resources
		5	TRSVCID does not match TRTYPE
		4	TRADDR does not match ADRFAM
		3	Invalid ADRFAM
		2	Invalid TRTYPE
		1	No additional information
		0	Reserved

source: NVM Express Specification

Bytes		PDU Section	Description
HDGSTF=1	HDGSTF=0		
13:10	Not Present	HDGST	**HDGST:** If HDGSTF is set to ‘1’ in the FLAGS field, this field is valid and contains the header digest (refer to section 3.3.1.1). If the HDGSTF bit is cleared to ‘0’, then this field is not present.
269:14	265:10	DATA	**CDC NVMe Qualified Name (CDCNQN):** This field indicates the NVMe Qualified Name (NQN) that uniquely identifies the CDC. Refer to the NVMe Qualified Names section in the NVMe Base Specification for the formatting requirements of NQNs.
DDGSTF=1	DDGSTF=0		
273:270	Not Present	DDGST	**DDGST:** If DDGSTF is set to ‘1’ in the FLAGS field, this field is valid and contains the data digest (refer to section 3.3.1.1). If the DDGSTF bit is cleared to ‘0’, then this field is not present.

source: NVM Express Specification

Kamal Bakshi, Cisco

PDU Type (0B) KDResp -example

CH Common Header | **PSH** | **PDU Data**

Enet hdr	IP hdr	TCP hdr	PDU Opcode (**0B**)	Flags	hlen	PDU offset	PDU len	KSSTAT	FAILRSN	HDGST	Data	DDGST	Enet FCS
			00	01	02	03	04	08	09	10	14	270.	273

KSSTAT: KickStart Status

Kickstart Discovery Response

```
0040  0a 99 7f ea 73 a7 0b 00  0a 0c 0c 01 00 00 01 00   ····s·.. ........
0050  00 00 6e 71 6e 2e 31 39  38 38 2d 31 31 2e 63 6f   ··nqn.19 88-11.co
0060  6d 2e 64 65 6c 6c 3a 53  46 53 53 3a 39 3a 32 30   m.dell:S FSS:9:20
0070  32 32 30 38 32 34 32 32  33 30 35 38 65 38 00 00   22082422 3058e8··
0080  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00   ········ ········
```

KSSTAT 01:(0000 00 01)

Kickstart Status (KSSTAT):

Bits	Description
7:3	Reserved
1	**FAILURE:** If set to '1', then the CDC shall not perform a pull registration due to the reason indicated in the Failure Reason (FAILRSN) field.
0	**SUCCESS:** If set to '1', then the CDC shall perform a pull registration.

Success
(CDC will perform PULL Registration)

```
▷ Ethernet II, Src: VMware_bf:37:26 (00:50:56:bf:37:26), Dst: 6a:3e:af:cd:f8:a7 (6a:3e:af:cd:f8:a7)
▷ 802.1Q Virtual LAN, PRI: 0, DEI: 0, ID: 900
▷ Internet Protocol Version 4, Src: 9.1.1.55, Dst: 9.1.1.9
▷ Transmission Control Protocol, Src Port: 8009, Dst Port: 58174, Seq: 129, Ack: 431, Len: 268
◢ NVM Express Fabrics TCP Discovery Controller
    Pdu Type: Kickstart Discovery Response (11)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 10
    Pdu Data Offset: 12
    Packet Length: 268
  ◢ KDResp
      Kickstart Status: 1 (SUCCESS)
      Failure Reason: 0 (NO FAILURE)
      CDC NVM Qualified Name (CDCNQN): nqn.1988-11.com.dell:SFSS:9:20220824223058e8
```


Kamal Bakshi, Cisco

PDU Type (04) CapsuleCmd -Capsule Command/SQE

CH Common Header | Command Capsule

Enet hdr	IP hdr	TCP hdr	PDU Opcode (**04**)	Flags	hlen	PDU offset	PDU len	SQE Header (64 bytes)	*HDGST*	PAD	In-Capsule Data	*DDGST*	Enet FCS
			00	01	02	03	04	08 **PSH**	72		N:	M:	M+3

A capsule is an NVMe unit of information exchange used in NVMe over Fabrics.
A command capsule contains a command (formatted as a Submission Queue Entry (SQE)) and may optionally include SGLs or data.

A capsule is independent of any underlying NVMe Transport unit (e.g., packet, message, or frame and associated headers and footers) and may consist of multiple such units.

Command capsules are transferred from a host to an NVM subsystem. The SQE contains an Admin command, an I/O command, or a Fabrics command.

Opcode: 04 Command Capsule

Bytes		PDU Section	Description
00		CH	**PDU-Type:** 04h
01		CH	**FLAGS:** (see Bits table below)
02		CH	**HLEN:** Fixed length of 72 bytes (i.e., 48h).
03		CH	**PDO:** Data offset within PDU (i.e., the offset from byte 0 to the CCICD field; the value of 'N'). This value shall be a multiple of the data alignment specified by the CPDA field in the ICResp PDU (refer to section 3.6.2.3) that was previously sent by the controller on this TCP connection.
07:04		CH	**PLEN:** Total length of PDU (including CH, PSH, HDGST, PAD, DATA, and DDGST) in bytes.
71:08		PSH	**NVMe-oF Command Capsule SQE (CCSQE):** Command Capsule SQE.
HDGSTF=1	HDGSTF=0		
75:72	Not present	HDGST	**HDGST:** If the HDGSTF bit is set to '1' in the FLAGS field, this field is present and contains a valid header digest (refer to section 3.3.1.1). If the HDGSTF bit is cleared to '0', then this field is not present.
N - 1:76	N - 1:72	PAD	**PAD:** If in-capsule data is present, the length of this field shall be the necessary number of bytes required to achieve the alignment specified by the CPDA field (refer to section 3.6.2.3).
M - 1:N		DATA	**NVMe-oF In-Capsule Data (CCICD):** This field contains the in-capsule data, if any, of the NVMe-oF Command Capsule.
DDGSTF=1	DDGSTF=0		
M + 3:M	Not present	DDGST	**Data Digest (DDGST):** If the DDGSTF bit is set to '1' in the FLAGS field, and the CCICD field is present, then this field contains the data digest (refer to section 3.3.1.1) of the CCICD field (i.e., the in-capsule data). If the DDGSTF bit is cleared to '0', then this field is not present.

FLAGS:

Bits	Description
7:2	Reserved
1	**DDGSTF:** If set to '1', then the DDGST field follows the PDU Data and contains a valid value. If cleared to '0', then the DDGST field is not present.
0	**HDGSTF:** If set to '1', then the HDGST field follows the PDU Header and contians a valid value. If cleared to '0', then the HDGST field is not present.

SQE
Submission Queue Entry

source: NVM Express TCP Transport Specification 1.0b

Kamal Bakshi, Cisco

PDU Type (04) CapsuleCmd/SQE

CH Common Header | Command Capsule

Enet hdr	IP hdr	TCP hdr	PDU Opcode (**04**)	Flags	hlen	PDU offset	PDU len	SQE Header (64 bytes)	*HDGST*	PAD	In-Capsule Data	*DDGST*	Enet FCS
			00 (04)CapsuleCmd	01	02	03	04	08 **PSH**	72		N:	M:	M+3

SQE Header (64bytes)

bytes

0					4	8	12	16	24	40	44	48	52	56	60 63
SQE Opcode	FUSE	reserved	PRP or SGL (PSDT)	Command ID	NSID	CDW2	3	Meta Data PTR	Data PTR	CDW10	11	12	13	14	15
00	08	10	14	16.	31										

bits

CDWs are command specific

I/O Queue Commands

01 Write
04 Write Uncorrectable
08 Write Zeroes
02 Read
00 Flush
0C Verify
05 Compare
19 Copy
09 Dataset Mgmt.
0D Resv. Register
0E Resv. Report
11 Resv. Acquire
15 Resv. Release

Admin Queue Commands

01 Create I/O SQ
00 Delete I/O SQ
05 Create I/O CQ
04 Delete I/O CQ

02 Get Log Page
06 Identify

09 Set Feature
0A Get Feature

0C AER
18 Keepalive

0D Namespace Mgmt.
15 NS Attachment

1C Virtualization Mgmt.
20 Capacity Mgmt.

19 Directive Send
1A Directive Receive

81 Security Send
82 Security Receive

1D NVMe-MI Send
1E NVMe-MI Receive

7F Fabric Commands

80 Format NVM
84 Sanitize
86 Get LBA Status
08 Abort
10 Firmware commit
11 Firmware download
14 Device Self Test
24 Lockdown
7C Doorbell Buffer Config

New Admin Cmds

Fabric Commands

01 Connect
08 Disconnect
00 Property Set
04 Property Get
05 Authentication Send
06 Authentication Receive

New Admin Cmds (CDC)

21 Discovery Info. Mgmt.
22 FZ Receive
25 FZ Lookup
29 FZ Send

Kamal Bakshi, Cisco

PDU Type (04) CapsuleCmd/SQE -example

Command Capsule

CH Common Header

Enet hdr	IP hdr	TCP hdr	PDU Opcode (**04**)	Flags	hlen	PDU offset	PDU len	SQE Header (64 bytes)	*HDGST*	PAD	In-Capsule Data	*DDGST*	Enet FCS
			00 (04)CapsuleCmd	01	02	03	04	08 **PSH**	72		N:	M:	M+3

SQE Header (64bytes)

bytes

0					4	8	12	16	24	40	44	48	52	56	60 63
SQE Opcode	FUSE	reserved	PRP or SGL (PSDT)	Command ID	NSID	CDW2	3	Meta Data PTR	Data PTR	CDW10	11	12	13	14	15
00 bits	08	10	14	16.	31										

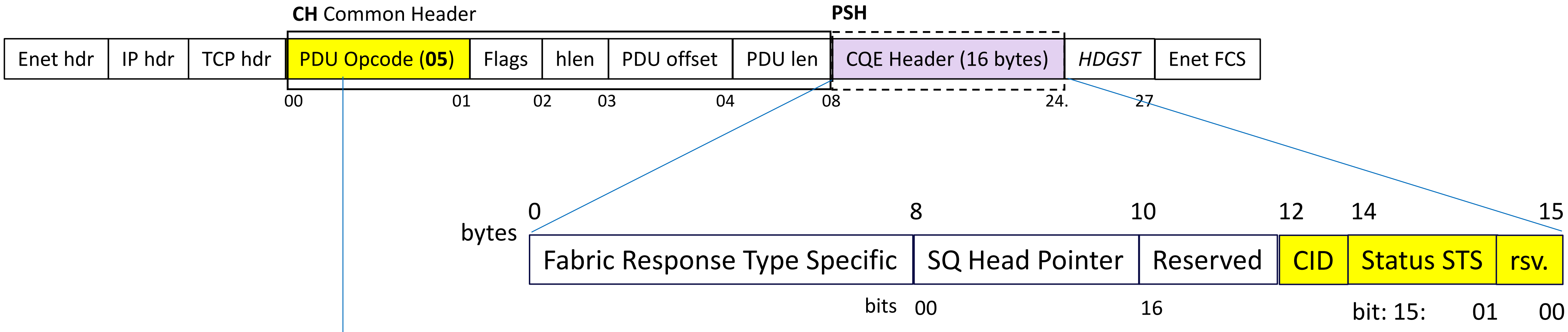
Identify Command

```
0040  d2 3c 04 00 48 00 48 00  00 00 06 40 02 20 00 00
0050  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0060  00 00 00 00 00 00 00 00  00 00 00 10 00 00 00 00
0070  00 5a 06 00 00 00 00 00  00 00 00 00 00 00 00 00
0080  00 00 00 00 00 00 00 00  00 00 ec 6e 79 e9
```

```
NVM Express Fabrics TCP Discovery Controller, NVMe Opcode: Identify (0x06) Cmd ID: 0x2002
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
NVM Express (Cmd)
    Opcode: 0x06 Identify
    [Cqe in: 175]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x2002
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  > SGL1
    Controller or Namespace Structure (CNS): 0x0006
    Reserved: 0000
    Controller Identifier (CNTID): 00000000
```


Kamal Bakshi, Cisco

PDU Type (05) CapsuleResp -Capsule Response/CQE



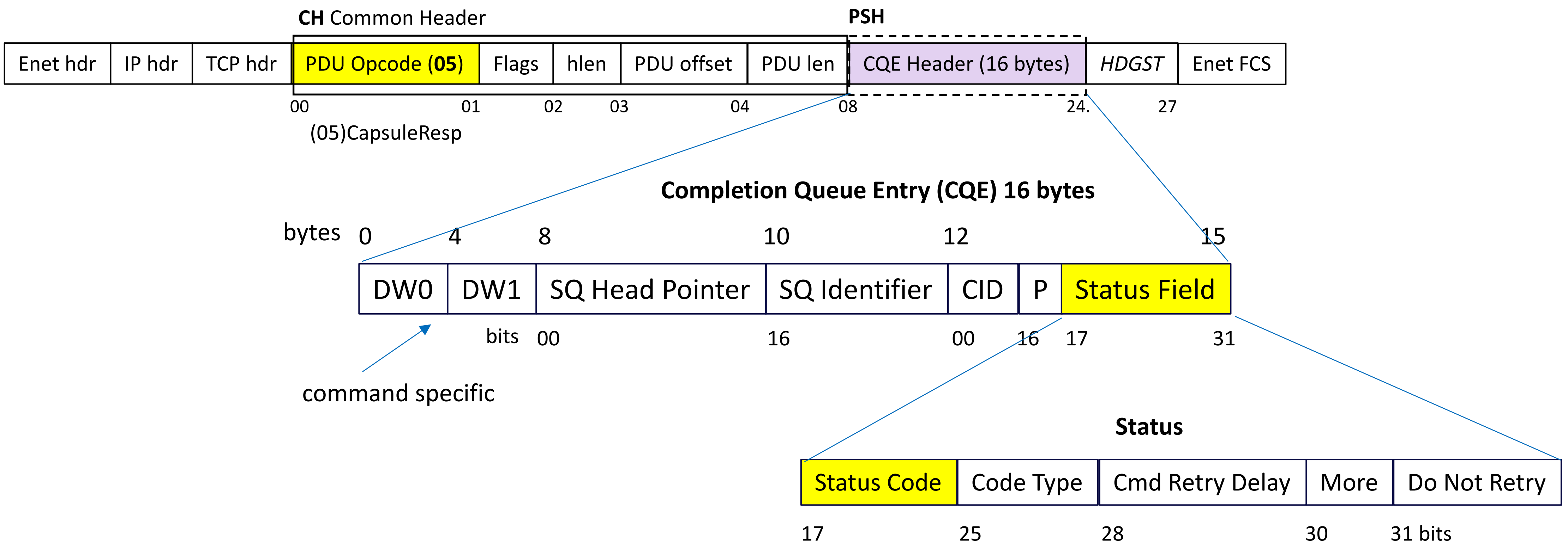
Opcode: 05 Response Capsule

<table>
<tr><th colspan="2">Bytes</th><th>PDU Section</th><th>Description</th></tr>
<tr><td colspan="2">00</td><td rowspan="5">CH</td><td>PDU-Type: 05h</td></tr>
<tr><td colspan="2">01</td><td>FLAGS:
<table><tr><th>Bits</th><th>Description</th></tr><tr><td>7:1</td><td>Reserved</td></tr><tr><td>0</td><td>HDGSTF: If set to ‘1’, then a valid HDGST value follows the PDU Header. If cleared to ‘0’, then the HDGST field is not present.</td></tr></table></td></tr>
<tr><td colspan="2">02</td><td>HLEN: Fixed length of 24 bytes (i.e., 18h).</td></tr>
<tr><td colspan="2">03</td><td>PDO: Reserved</td></tr>
<tr><td colspan="2">07:04</td><td>PLEN: Length of CH, PSH, and HDGST, if present, in bytes.</td></tr>
<tr><td colspan="2">23:08</td><td>PSH</td><td>NVMe-oF Response Capsule CQE (RCCQE): Response Capsule CQE.</td></tr>
<tr><td>HDGSTF=1</td><td>HDGSTF=0</td><td></td><td></td></tr>
<tr><td>27:24</td><td>Not present</td><td>HDGST</td><td>HDGST: If the HDGSTF bit is set to ‘1’ in the FLAGS field, this field is present and contains a valid header digest (refer to section 3.3.1.1). If the HDGSTF bit is cleared to ‘0’, then this field is not present.</td></tr>
</table>

source: NVM Express TCP Transport Specification 1.0b

Kamal Bakshi, Cisco

PDU Type (05) CapsuleResp/CQE



Status Codes

00 Successful Completion
01 Invalid Command Opcode
02 Invalid Field in command
03 Command ID conflict
04 Data Transfer Error
05 Commands Aborted, power loss
06 Internal Error
07 Cmd Abort Req.

08 Cmd Aborted, SQ deletion
09 Cmd Aborted, Failed Fused
0A Cmd Aborted, missing Fused
0B Invalid Namespace
0C Cmd Sequence Error
0D Invalid SGL Descriptor
0E Invalid Number of SGL
0F Data SGL length invalid

10 Metadata SGL length
11 SGL type invalid
12 Invalid use of CMB
13 PRP Offset Invalid
14 Atomic Write exceeded
15 Operation Denied
16 SGL Offset Invalid
18 Host ID Inconsistent

19 Keep Alive Timer Expired
1A Keep Alive Timeout Invalid
1B Cmd Aborted / Abort
1C Sanitize Failed
1D Sanitize in Progress
1E SGL Data Block invalid
1F Cmd not supported/CMB
20 Namespace in write protect

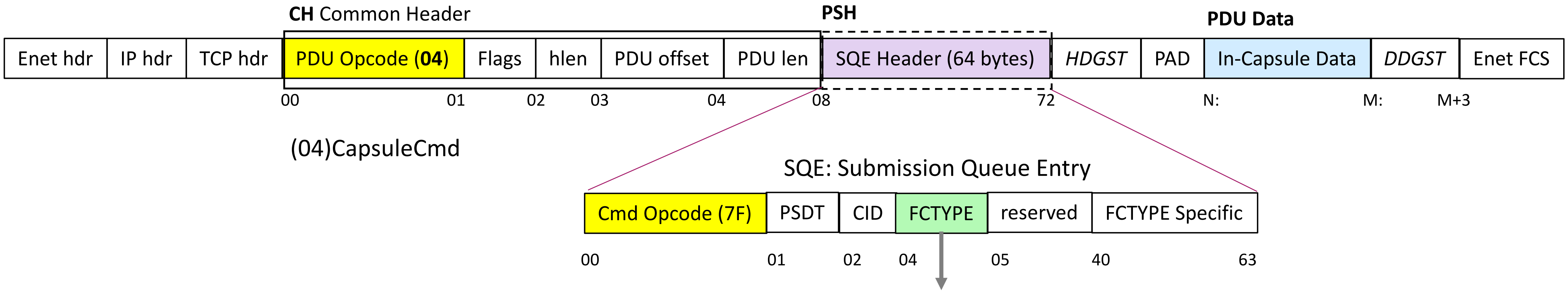
21 Command Interrupted
22 Transient Transport Error
23 Cmd Prohibited by feature
24 Admin Cmd Media not ready
80 LBA Out of Range
81 Capacity Exceeded
82 Namespace not ready
83 Reservation Conflict

84 Format in Progress
85 Invalid Value Size
86 Invalid Key Size
87 KV Key Does not exist
88 Unrecovered Error
89 Key Exists

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

PDU Type (04) CapsuleCmd(7F) -Fabric Commands



Fabrics commands are used to create queues and initialize a controller.

-01 Connect
-08 Disconnect

-04 Property Get
-00 Property Set

-05 Auth. Send
-06 Auth. Receive

Fabric Command Types

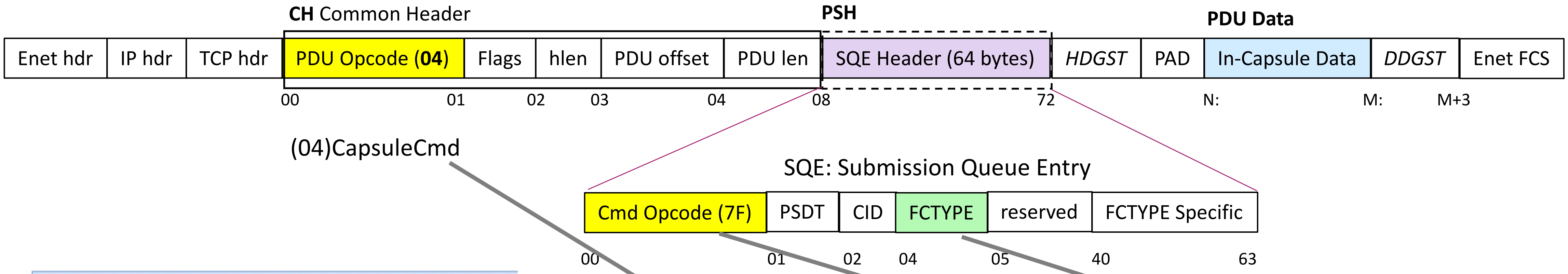
Command Type by Field			Combined Command Type[2]	O/M[1]	I/O Queue[3]	Command
(07) Generic Command	(06:02) Function	(01:00) Data Transfer[4]				
0b	000 00b	00b	00h	M	No	Property Set
0b	000 00b	01b	01h	M	Yes	Connect[5]
0b	000 01b	00b	04h	M	No	Property Get
0b	000 01b	01b	05h	O	Yes	Authentication Send
0b	000 01b	10b	06h	O	Yes	Authentication Receive
0b	000 10b	00b	08h	O	Yes	Disconnect
	Vendor Specific					
1b	na	na	C0h to FFh	O		Vendor specific

NOTES:
1. O/M definition: O = Optional, M = Mandatory.
2. Opcodes not listed are reserved.
3. All Fabrics commands, other than the Disconnect command, may be submitted on the Admin Queue. The I/O Queue supports Fabrics commands as specified in this column. If a Fabrics command that is not supported on an I/O Queue is sent on an I/O Queue, that command shall be aborted with a status code of Invalid Field in Command.
4. 00b = no data transfer; 01b = host to controller; 10b = controller to host; 11b = reserved
5. The Connect command is submitted and completed on the same queue that the Connect command creates. Refer to section 1.5.7.

source: NVMe-over-Fabrics -1.1a

Kamal Bakshi, Cisco

PDU Type (04) CapsuleCmd(7F) -Fabric Commands -example




```
◢ NVM Express Fabrics TCP Discovery Controller,
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 72
    Packet Length: 1096
  ◢ NVM Express Fabrics (Cmd)
      Opcode: 0x7f Fabric Cmd
      [Fabric Cqe in: 156]
      Reserved: 0x40
      Command ID: 0x1000
      Fabric Cmd Type: Connect (0x01)
      Reserved: 000000000000000000000000000000
    ▷ SGL1
      Record Format: 0x0000
      Queue ID: 0x0000
      SQ Size: 0x001f
```

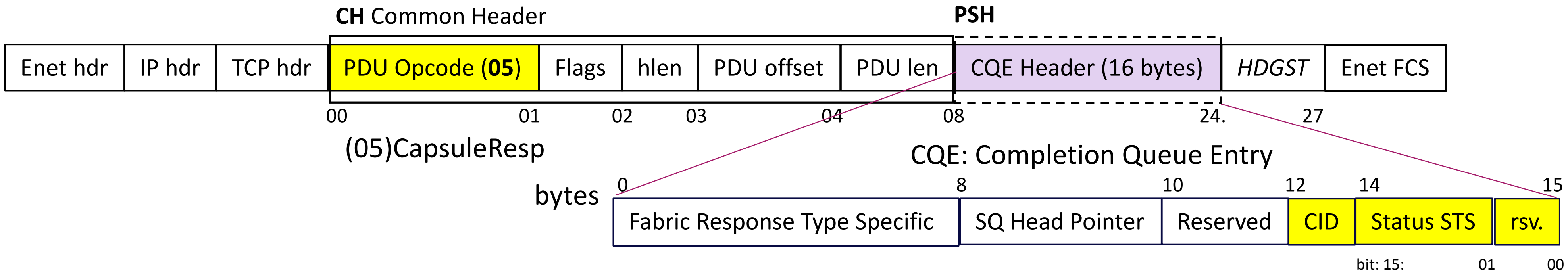
Fabric Connect

```
0040  d2 23 04 00 48 48 48 04  00 00 7f 40 00 10 01 00
0050  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
0060  00 00 00 00 00 00 00 00  00 00 00 04 00 00 00 00
0070  00 01 00 00 00 00 1f 00  00 00 30 75 00 00 00 00
0080  00 00 00 00 00 00 00 00  00 00 ae cb 70 db d2 0d
0090  45 ae b9 1f 74 05 b9 32  ae 19 ff ff 00 00 00 00
00a0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
```


Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

PDU Type (05) CapsuleResp -(Fabric Cmd) Response/CQE



Fabric Response Capsule -Completion Queue Entry (CQE)

Bytes	Description
07:00	The definition of this field is Fabrics response type specific.
09:08	**SQ Head Pointer (SQHD):** Indicates the current Submission Queue Head pointer for the associated Submission Queue[1].
11:10	Reserved
13:12	**Command Identifier (CID):** Indicates the identifier of the command that is being completed.
15:14	**Status (STS):** Specifies status for the associated Fabrics command. **Bits** / **Definition** 15:01 — Status Field as defined in section 4.6.1 of the NVMe Base specification. 00 — Reserved

NOTES:
1. The SQHD field is reserved if SQ flow control is disabled for the queue pair, refer to section 2.4 and to section 3.3.

source: NVMe-over-Fabrics -1.1a

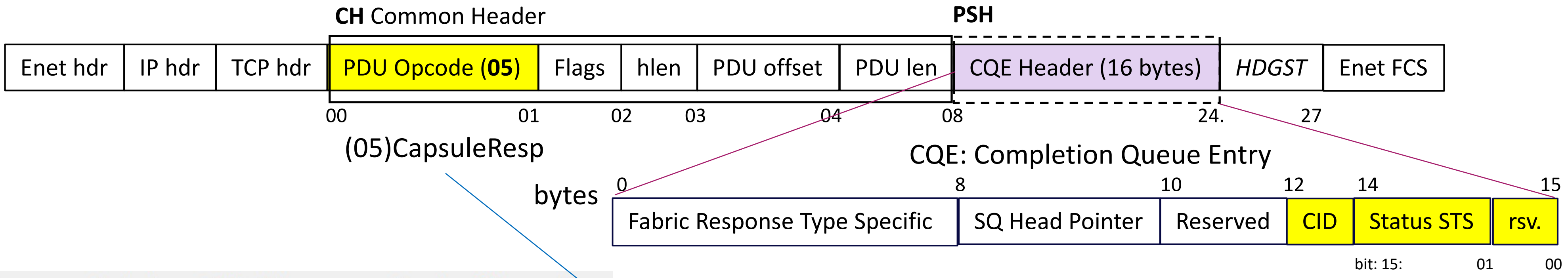
Fabrics commands use the status for commands defined in the NVMe Base specification.
Fabrics commands use an allocation of command specific status values from 80h to BFh

Value	Description	Commands Affected
80h	**Incompatible Format:** The NVM subsystem does not support the record format specified by the host.	Connect, Disconnect
81h	**Controller Busy:** The controller is already associated with a host (Connect command). This value is also returned if there is no available controller (Connect command). The controller is not able to disconnect the I/O Queue at the current time (Disconnect command).	Connect, Disconnect
82h	**Connect Invalid Parameters:** One or more of the command parameters (e.g., Host NQN, Subsystem NQN, Host Identifier, Controller ID, Queue ID) specified are not valid.	Connect
83h	**Connect Restart Discovery:** The NVM subsystem requested is not available. The host should restart the discovery process.	Connect
84h	**Connect Invalid Host:** The host is not allowed to establish an association to any controller in the NVM subsystem or the host is not allowed to establish an association to the specified controller.	Connect
85h	**Invalid Queue Type:** The command was sent on the wrong queue type (e.g., a Disconnect command was sent on the Admin queue).	Disconnect
86h to 8Fh	Reserved	
90h	**Discover Restart:** The snapshot of the records is now invalid or out of date. The host should re-read the Discovery Log Page.	Get Log Page
91h	**Authentication Required:** NVMe in-band authentication is required and the queue has not yet been authenticated.	NOTE 1
92h to AFh	Reserved	
B0h to BFh	**Transport Specific:** The status values in this range are NVMe Transport specific. Refer to the appropriate NVMe Transport binding specification for the definition of these status values.	

NOTES:
1. All commands other than Connect, Authenticate Send, and Authenticate Receive.

Kamal Bakshi, Cisco

PDU Type (05) CapsuleResp -(Fabric Cmd) Response/CQE




```
◢ NVM Express Fabrics TCP Discovery Controller, Cqe
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
  ◢ Cqe (For Cmd: Connect)
      [Fabric Cmd in: 155]
      [Cmd Latency: 20.695 ms]
      Controller ID: 0x0402
      Authentication Required: 0x0000
      Reserved: 00000000
      SQ Head Pointer: 0x0000
      Reserved: 0x0000
      Command ID: 0x1000
      0000 0000 0000 000. = Status: 0x0000
      .... .... .... ...0 = Reserved: 0x0
```

Connect Response

```
0000  50 6b 4b 4b df 3a 00 50   56 bf 37 26 08 00 45 00
0010  00 4c 33 54 40 00 3e 06   f4 f3 09 01 01 37 09 01
0020  01 2c 1f 49 db cc 02 84   20 66 01 78 d5 96 80 18
0030  01 f5 50 d5 00 00 01 01   08 0a a1 ec d2 38 31 46
0040  3d e0 05 00 18 00 18 00   00 00 02 04 00 00 00 00
0050  00 00 00 00 00 00 00 10   00 00 21 3a 84 2b
```

Successful

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

PDU Type (06) H2CData -Host to Controller Data

CH Common Header								PSH			PDU Data		
Enet hdr	IP hdr	TCP hdr	PDU Opcode (**06**)	Flags	hlen	PDU offset	PDU len	Packet Specific Header	*HDGST*	PAD	Command Data Buffer	*DDGST*	Enet FCS
			00	01	02	03	04	08	24		N:	M:	M+3

Opcode: 06 Host to Controller Data Transfer

Bytes		PDU Section	Description
00		CH	**PDU-Type:** 06h
01		CH	**FLAGS:** **Bits** / **Description** 7:3 / Reserved 2 / **LAST_PDU:** If set to '1', indicates the PDU is the last in the set of H2CData PDUs that correspond to the same R2T PDU. 1 / **DDGSTF:** If set to '1', then the DDGST field follows the PDU Data and contains a valid value. If cleared to '0', then the DDGST field is not present. 0 / **HDGSTF:** If set to '1', then the HDGST field follows the PDU Header and contains a valid value. If cleared to '0', then the HDGST field is not present.
02		CH	**HLEN:** Fixed length of 24 bytes (i.e., 18h).
03		CH	**PDO:** Data Offset within PDU (i.e., the offset from byte 0 to the PDU-Data field; the value of 'N'). This value shall be a multiple of the data alignment specified by the CPDA field in the ICResp PDU (refer to section 3.6.2.3) that was previously sent by the controller on this TCP connection.
07:04		CH	**PLEN:** Total length of PDU (including CH, PSH, HDGST, PAD, DATA, and DDGST) in bytes.
09:08		PSH	**Command Capsule CID (CCCID):** This field contains the SQE.CID value of the Command Capsule PDU associated with the Command Data Buffer.
11:10		PSH	**Transfer Tag (TTAG):** This field contains the Transfer Tag of the corresponding R2T received by the host.
15:12		PSH	**Data Offset (DATAO):** Byte offset from start of Command Data Buffer to the first byte to transfer. This value shall be a multiple of dwords.
19:16		PSH	**Data Length (DATAL):** PDU-Data field length in bytes (i.e., the value of M-N). This value shall be a multiple of dwords.
23:20		PSH	Reserved
HDGSTF=1	HDGSTF=0		
27:24	Not present	HDGST	**HDGST:** If the HDGSTF bit is set to '1' in the FLAGS field, this field is present and contains a valid header digest (refer to section 3.3.1.1). If the HDGSTF bit is cleared to '0', then this field is not present.
N - 1:28	N - 1:24	PAD	**PAD:** The length of this field shall be the necessary number of bytes required to achieve the alignment specified by the CPDA field (refer to section 3.6.2.3).
M - 1:N		DATA	**PDU-Data:** This field contains the contents of the Command Data Buffer being transferred. The length of this field is a multiple of dwords.
DDGSTF=1	DDGSTF=0		
M + 3:M	Not present	DDGST	**Data Digest (DDGST):** If the DDGSTF bit is set to '1' in the FLAGS field, this field is present and contains a valid data digest (refer to section 3.3.1.1). If the DDGSTF bit is cleared to '0', then this field is not present.

source: NVM Express TCP Transport Specification 1.0b

Kamal Bakshi, Cisco

PDU Type (07) C2HData -Controller to Host Data

			CH Common Header					PSH			PDU Data		
Enet hdr	IP hdr	TCP hdr	PDU Opcode (**07**)	Flags	hlen	PDU offset	PDU len	Packet Specific Header	*HDGST*	PAD	Host-resident data	*DDGST*	Enet FCS
			00	01	02	03	04	08	24		N:	M:	M+3

Opcode: 07 Controller To Host Data Transfer

<table>
<tr><th colspan="2">Bytes</th><th>PDU Section</th><th>Description</th></tr>
<tr><td colspan="2">00</td><td rowspan="6">CH</td><td>PDU-Type: 07h</td></tr>
<tr><td colspan="2">01</td><td>FLAGS:
<table>
<tr><th>Bits</th><th>Description</th></tr>
<tr><td>7:4</td><td>Reserved</td></tr>
<tr><td>3</td><td>SUCCESS: If set to ‘1’, indicates that the command referenced by CCCID was completed successfully with no other information and that no Response Capsule PDU is sent by the Controller.</td></tr>
<tr><td>2</td><td>LAST_PDU: If set to ‘1’, indicates the PDU is the last C2HData PDU sent in response to a Command Capsule PDU.</td></tr>
<tr><td>1</td><td>DDGSTF: If set to ‘1’, then the DDGST field follows the PDU Data and contains a valid value. If cleared to ‘0’, then the DDGST field is not present.</td></tr>
<tr><td>0</td><td>HDGSTF: If set to ‘1’, then the HDGST field follows the PDU Header and contains a valid value. If cleared to ‘0’, then the HDGST field is not present.</td></tr>
</table></td></tr>
<tr><td colspan="2">02</td><td>HLEN: Fixed length of 24 bytes (i.e., 18h).</td></tr>
<tr><td colspan="2">03</td><td>PDO: Data offset within PDU (i.e., the offset from byte 0 to the PDU-Data field; the value of ‘N’). This value shall be a multiple of the data alignment specified by the HPDA field in the ICReq PDU (refer to section 3.6.2.2) that was previously sent by the host on this TCP connection.</td></tr>
<tr><td colspan="2">07:04</td><td>PLEN: Total length of PDU (i.e., including CH, PSH, HDGST, PAD, DATA, and DDGST) in bytes.</td></tr>
<tr><td colspan="2">09:08</td><td rowspan="5">PSH</td><td>Command Capsule CID (CCCID): This field contains the SQE.CID value of the Command Capsule PDU associated with the host-resident data.</td></tr>
<tr><td colspan="2">11:10</td><td>Reserved</td></tr>
<tr><td colspan="2">15:12</td><td>Data Offset (DATAO): Byte offset from start of host-resident data to the first byte to transfer. This value shall be dword aligned.</td></tr>
<tr><td colspan="2">19:16</td><td>Data Length (DATAL): PDU-Data field length in bytes (i.e., the value of M-N). This value shall be dword aligned.</td></tr>
<tr><td colspan="2">23:20</td><td>Reserved</td></tr>
<tr><td>HDGSTF=1</td><td>HDGSTF=0</td><td colspan="2"></td></tr>
<tr><td>27:24</td><td>Not present</td><td>HDGST</td><td>HDGST: If the HDGSTF bit is set to ‘1’ in the FLAGS field, this field is present and contains a valid header digest (refer to section 3.3.1.1). If the HDGSTF bit is cleared to ‘0’, then this field is not present.</td></tr>
<tr><td>N - 1:28</td><td>N - 1:24</td><td>PAD</td><td>PAD: If the HPDA field (refer to section 3.6.2.2) is set to a non-zero value, then the length of this field shall be the necessary number of bytes required to achieve the alignment specified by the HPDA field.</td></tr>
<tr><td colspan="2">M - 1:N</td><td>DATA</td><td>PDU-Data: This field contains the host-resident data being transferred. The length of this field is a multiple of dwords.</td></tr>
<tr><td>DDGSTF=1</td><td>DDGSTF=0</td><td colspan="2"></td></tr>
<tr><td>M + 3:M</td><td>Not present</td><td>DDGST</td><td>Data Digest (DDGST): If the DDGSTF bit is set to ‘1’ in the FLAGS field, this field is present and contains a valid data digest (refer to section 3.3.1.1) of the PDU-Data field. If the DDGSTF bit is cleared to ‘0’, then this field is not present.</td></tr>
</table>

source: NVM Express TCP Transport Specification 1.0b

Kamal Bakshi, Cisco

PDU Type (09) R2T -Ready To Transfer

CH Common Header | **PSH**

Enet hdr	IP hdr	TCP hdr	PDU Opcode (**09**)	Flags	hlen	PDU offset	PDU len	Packet Specific Header	*HDGST*	Enet FCS
			00	01	02	03	04	08	24	27

Opcode: 09 Ready to Transfer

Bytes		PDU Section	Description
00		CH	**PDU-Type:** 09h
01		CH	**FLAGS:** (see Bits table below)
02		CH	**HLEN:** Fixed length of 24 bytes (i.e., 18h).
03		CH	**PDO:** Reserved
07:04		CH	**PLEN:** Length of CH, PSH, and HDGST, if present, in bytes.
09:08		PSH	**Command Capsule CID (CCCID):** This field contains the SQE.CID value of the Command Capsule PDU associated with the host-resident data.
11:10		PSH	**Transfer Tag (TTAG):** This field contains a controller generated tag. The rules of the tag generation are outside the scope of this specification.
15:12		PSH	**Requested Data Offset (R2TO):** Byte offset from the start of the host-resident data to the first byte to transfer. This value shall be dword aligned.
19:16		PSH	**Requested Data Length (R2TL):** Number of bytes of Command Data Buffer requested by the controller. This value shall be dword aligned.
23:20		PSH	Reserved
HDGSTF=1	HDGSTF=0		
27:24	Not present	HDGST	**HDGST:** If the HDGSTF bit is set to '1' in the FLAGS field, this field is present and contains a valid header digest (refer to section 3.3.1.1). If the HDGSTF bit is cleared to '0', then this field is not present.

FLAGS:

Bits	Description
7:1	Reserved
0	**HDGSTF:** If set to '1', then the HDGST field follows the PDU Header and contains a valid value. If cleared to '0', then the HDGST field is not present.

source: NVM Express TCP Transport Specification 1.0b

Kamal Bakshi, Cisco

PDU Type (02) H2CTermReq -Host To Controller Termination Request

			CH Common Header					**PSH**		**PDU Data**	
Enet hdr	IP hdr	TCP hdr	PDU Opcode (**02**)	Flags	hlen	PDU offset	PDU len	Fatal Error Status/Info	*reserved*	PDU header	Enet FCS
			00	01	02	03	04	08	14	24:	N-1

Opcode: 02 H2CTermReq -Host to Controller Terminate Connection Request

<table>
<tr><th>Bytes</th><th>PDU Section</th><th>Description</th></tr>
<tr><td>00</td><td rowspan="5">CH</td><td>PDU-Type: 02h</td></tr>
<tr><td>01</td><td>FLAGS: Reserved</td></tr>
<tr><td>02</td><td>HLEN: Fixed length of 24 bytes (18h).</td></tr>
<tr><td>03</td><td>PDO: Reserved</td></tr>
<tr><td>07:04</td><td>PLEN: Total length of PDU (including PDU Header and DATA) in bytes. This value shall not exceed a limit of 152 bytes.</td></tr>
<tr><td>09:08</td><td rowspan="3">PSH</td><td>Fatal Error Status (FES): Indicates the fatal error information.
<table>
<tr><th>Value</th><th>Description</th></tr>
<tr><td>00h</td><td>Reserved</td></tr>
<tr><td>01h</td><td>Invalid PDU Header Field: An invalid field in the transport header was detected by the host.</td></tr>
<tr><td>02h</td><td>PDU Sequence Error: An unexpected protocol sequence was detected by the host.</td></tr>
<tr><td>03h</td><td>Header Digest Error: An HDGST error was detected by the host.</td></tr>
<tr><td>04h</td><td>Data Transfer Out of Range: A C2HData PDU with data offset or data offset plus data length is out of its associated Command Data Buffer range.</td></tr>
<tr><td>05h</td><td>R2T Limit Exceeded: An R2T PDU that exceeds MAXR2T was received by the host.</td></tr>
<tr><td>06h</td><td>Unsupported Parameter: An unsupported parameter was received by the host.</td></tr>
<tr><td>07h to FFFFh</td><td>Reserved</td></tr>
</table></td></tr>
<tr><td>13:10</td><td>Fatal Error Information (FEI): Provides additional information based on the Fatal Error Status field.
<table>
<tr><th>Fatal Error Status</th><th>Contents of Error Specific Information</th></tr>
<tr><td>00</td><td>Reserved</td></tr>
<tr><td>01h</td><td>PDU Header Field Offset: This field indicates the offset in bytes from the start of the PDU Header to the start of the field that has an error. If multiple errors exist, then this field indicates the lowest offset that has an error.</td></tr>
<tr><td>02h</td><td>Reserved</td></tr>
<tr><td>03h</td><td>PDU Header Digest: This field indicates the HDGST that was received by the host caused a header digest verification error.</td></tr>
<tr><td>04h to 05h</td><td>Reserved</td></tr>
<tr><td>06h</td><td>Unsupported Parameter Field Offset: This field indicates the offset in bytes from the start of the PDU Header to the start of the field that has an error. If multiple errors exist, then this field indicates the lowest offset that has an error.</td></tr>
<tr><td>07h to FFFFh</td><td>Reserved</td></tr>
</table></td></tr>
<tr><td>23:14</td><td>Reserved</td></tr>
<tr><td>N - 1:24</td><td>DATA</td><td>Data: This field contains the PDU Header that was being processed by the host while the fatal error was detected.</td></tr>
</table>

source: NVM Express TCP Transport Specification 1.0b

Kamal Bakshi, Cisco

PDU Type (03) C2HTermReq -Controller to Host Termination Request

			CH Common Header					PSH		PDU Data	
Enet hdr	IP hdr	TCP hdr	PDU Opcode (**03**)	Flags	hlen	PDU offset	PDU len	Fatal Error Status/Info	*reserved*	PDU header	Enet FCS
			00	01	02	03	04	08	14	24:	N-1

Opcode: 03 C2HTermReq -Controller to Host Terminate Connection Request

<table>
<tr><th>Bytes</th><th>PDU Section</th><th>Description</th></tr>
<tr><td>00</td><td rowspan="5">CH</td><td>PDU-Type: 03h</td></tr>
<tr><td>01</td><td>FLAGS: Reserved</td></tr>
<tr><td>02</td><td>HLEN: Fixed length of 24 bytes (18h).</td></tr>
<tr><td>03</td><td>PDO: Reserved</td></tr>
<tr><td>07:04</td><td>PLEN: Total length of PDU (including PDU Header and DATA) in bytes. This value shall not exceed a limit of 152 bytes.</td></tr>
<tr><td>09:08</td><td rowspan="3">PSH</td><td>Fatal Error Status (FES): Indicates the fatal error information.
<table>
<tr><th>Value</th><th>Description</th></tr>
<tr><td>00h</td><td>Reserved</td></tr>
<tr><td>01h</td><td>Invalid PDU Header Field: An invalid field in the transport header was detected by the controller.</td></tr>
<tr><td>02h</td><td>PDU Sequence Error: An unexpected protocol sequence was detected by the controller.</td></tr>
<tr><td>03h</td><td>Header Digest Error: An HDGST error was detected by the controller.</td></tr>
<tr><td>04h</td><td>Data Transfer Out of Range: An H2CData PDU with data offset or data offset plus data length is out of its associated R2T range.</td></tr>
<tr><td>05h</td><td>Data Transfer Limit Exceeded: An H2CData PDU with data length that exceeds MAXH2CDATA was received by the controller.</td></tr>
<tr><td>06h</td><td>Unsupported Parameter: An unsupported parameter was received by the controller.</td></tr>
<tr><td>07h to FFFFh</td><td>Reserved</td></tr>
</table></td></tr>
<tr><td>13:10</td><td>Fatal Error Information (FEI): Provides additional information based on the Fatal Error Status field.
<table>
<tr><th>Fatal Error Status</th><th>Contents of Error Specific Information</th></tr>
<tr><td>00h</td><td>Reserved</td></tr>
<tr><td>01h</td><td>PDU Header Field Offset: This field indicates the offset in bytes from the start of the PDU Header to the start of the field that has an error. If multiple errors exist, then this field indicates the lowest offset that has an error.</td></tr>
<tr><td>02h</td><td>Reserved</td></tr>
<tr><td>03h</td><td>PDU Header Digest: This field indicates the HDGST that was received by the controller caused a header digest verification error.</td></tr>
<tr><td>04h to 05h</td><td>Reserved</td></tr>
<tr><td>06h</td><td>Unsupported Parameter Field Offset: This field indicates the offset in bytes from the start of the PDU Header to the start of the field that has an error. If multiple errors exist, then this field indicates the lowest offset that has an error.</td></tr>
<tr><td>07h to FFFFh</td><td>Reserved</td></tr>
</table></td></tr>
<tr><td>23:14</td><td>Reserved</td></tr>
<tr><td>N – 1:24</td><td>DATA</td><td>Data: This field contains the PDU Header that was being processed by the controller while the fatal error was detected.</td></tr>
</table>

source: NVM Express TCP Transport Specification 1.0b

Kamal Bakshi, Cisco

mDNS Query (_nvme-disc._tcp.local)

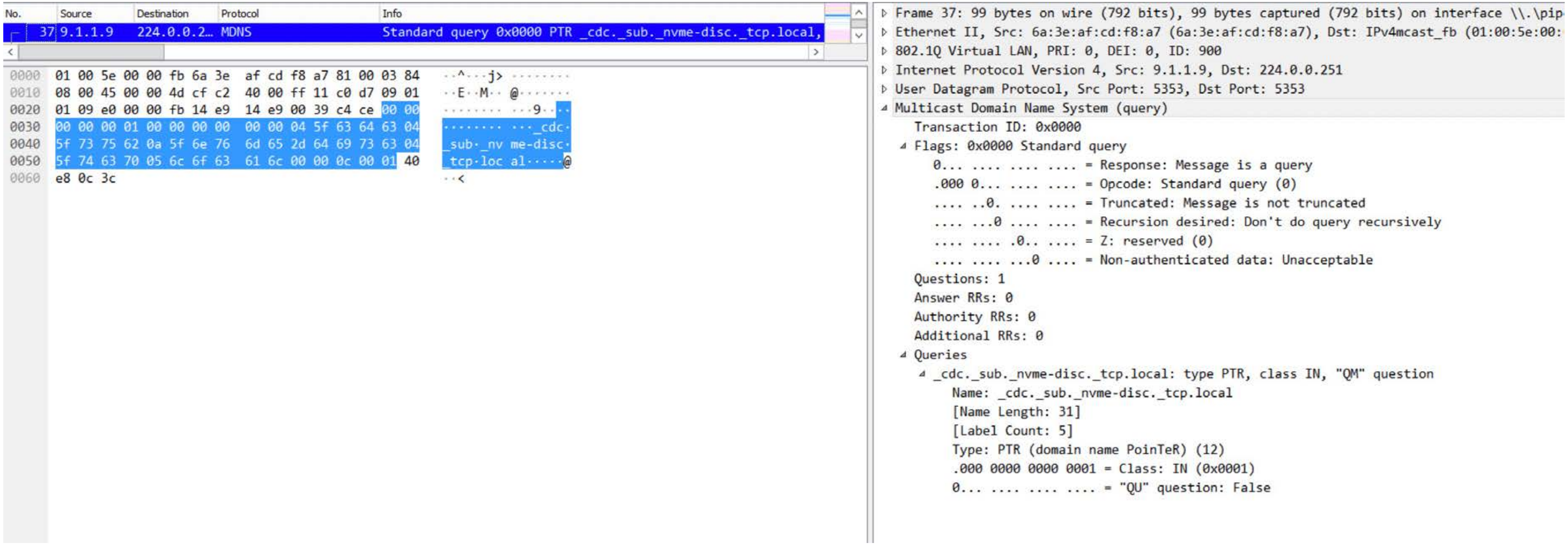
No.	Source	Destination	Protocol	Info
24	9.1.1.55	224.0.0.251	MDNS	Standard query 0x0000 PTR _nvme-disc._tcp.local, "QM" question

```
0000  01 00 5e 00 00 fb 00 50  56 bf 37 26 81 00 03 84   ··^····P V·7&····
0010  08 00 45 00 00 43 5a 92  40 00 ff 11 35 e4 09 01   ··E··CZ· @···5···
0020  01 37 e0 00 00 fb 14 e9  14 e9 00 2f 64 52 00 00   ·7······ ···/dR··
0030  00 00 00 01 00 00 00 00  00 00 0a 5f 6e 76 6d 65   ········ ···_nvme
0040  2d 64 69 73 63 04 5f 74  63 70 05 6c 6f 63 61 6c   -disc·_t cp·local
0050  00 00 0c 00 01 7a b1 b3  41                        ·····z·· A
```

```
> Frame 24: 89 bytes on wire (712 bits), 89 bytes captured (712 bits) on interface \\.\pipe\view_
> Ethernet II, Src: VMware_bf:37:26 (00:50:56:bf:37:26), Dst: IPv4mcast_fb (01:00:5e:00:00:fb)
> 802.1Q Virtual LAN, PRI: 0, DEI: 0, ID: 900
> Internet Protocol Version 4, Src: 9.1.1.55, 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: 1
    Answer RRs: 0
    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 (domain name PoinTeR) (12)
        .000 0000 0000 0001 = Class: IN (0x0001)
        0... .... .... .... = "QU" question: False
    [Response In: 25]
```


Kamal Bakshi, Cisco

mDNS Query (_cdc._sub._nvme-disc._tcp.local)



Kamal Bakshi, Cisco

mDNS Response

No.	Source	Destination	Protocol	Info
7	9.1.1.55	224.0.0.251	MDNS	Standard query response 0x0000 PTR 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local TXT, cache flush SRV, cache flush 0 0 80

```
0000  01 00 5e 00 00 fb 00 50  56 bf 37 26 81 00 03 84   ··^····P V·7&····
0010  08 00 45 00 00 d6 5f 3d  40 00 ff 11 30 a6 09 01   ··E···_= @···0···
0020  01 37 e0 00 00 fb 14 e9  14 e9 00 c2 76 ad 00 00   ·7······ ····v···
0030  84 00 00 00 00 04 00 00  00 00 0a 5f 6e 76 6d 65   ········ ···_nvme
0040  2d 64 69 73 63 04 5f 74  63 70 05 6c 6f 63 61 6c   -disc·_t cp·local
0050  00 00 0c 00 01 00 00 11  94 00 1d 1a 39 2d 31 2d   ········ ····9-1-
0060  31 2d 35 35 3a 30 38 2f  32 37 2f 32 32 3a 30 31   1-55:08/ 27/22:01
0070  3a 35 33 3a 30 35 c0 0c  c0 2d 00 10 80 01 00 00   :53:05·· ·-······
0080  11 94 00 37 05 70 3d 74  63 70 30 4e 51 4e 3d 6e   ···7·p=t cp0NQN=n
0090  71 6e 2e 31 39 38 38 2d  31 31 2e 63 6f 6d 2e 64   qn.1988- 11.com.d
00a0  65 6c 6c 3a 53 46 53 53  3a 39 3a 32 30 32 32 30   ell:SFSS :9:20220
00b0  38 32 34 32 32 33 30 35  38 65 38 c0 2d 00 21 80   82422305 8e8·-·!·
00c0  01 00 00 00 78 00 11 00  00 00 00 1f 49 08 39 2d   ····x··· ····I·9-
00d0  31 2d 31 2d 35 35 c0 1c  c0 9f 00 01 80 01 00 00   1-1-55·· ········
00e0  00 78 00 04 09 01 01 37  32 3c 6d 51               ·x·····7 2<mQ
```

```
> Frame 7: 236 bytes on wire (1888 bits), 236 bytes captured (1888 bits) on interface \\.\pipe\v:
> Ethernet II, Src: VMware_bf:37:26 (00:50:56:bf:37:26), Dst: IPv4mcast_fb (01:00:5e:00:00:fb)
> 802.1Q Virtual LAN, PRI: 0, DEI: 0, ID: 900
> Internet Protocol Version 4, Src: 9.1.1.55, Dst: 224.0.0.251
> User Datagram Protocol, Src Port: 5353, Dst Port: 5353
v Multicast Domain Name System (response)
    Transaction ID: 0x0000
  > Flags: 0x8400 Standard query response, No error
    Questions: 0
    Answer RRs: 4
    Authority RRs: 0
    Additional RRs: 0
  v Answers
    > _nvme-disc._tcp.local: type PTR, class IN, 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.loca
    v 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type TXT, class IN, cache flush
        Name: 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local
        Type: TXT (Text strings) (16)
        .000 0000 0000 0001 = Class: IN (0x0001)
        1... .... .... .... = Cache flush: True
        Time to live: 4500 (1 hour, 15 minutes)
        Data length: 55
        TXT Length: 5
        TXT: p=tcp
        TXT Length: 48
        TXT: NQN=nqn.1988-11.com.dell:SFSS:9:20220824223058e8
    v 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type SRV, class IN, cache flush, priorit
        Service: 9-1-1-55:08/27/22:01:53:05
        Protocol: _nvme-disc
        Name: _tcp.local
        Type: SRV (Server Selection) (33)
        .000 0000 0000 0001 = Class: IN (0x0001)
        1... .... .... .... = Cache flush: True
        Time to live: 120 (2 minutes)
        Data length: 17
        Priority: 0
        Weight: 0
        Port: 8009
        Target: 9-1-1-55.local
    > 9-1-1-55.local: type A, class IN, cache flush, addr 9.1.1.55
```

mDNS Announcement

No.	Source	Destination	Protocol	Info
10	9.1.1.44	224.0.0.251	MDNS	Standard query 0x0000 SRV 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local, "QM" question AAAA 9-1-1-55.local, "QM" question

```
0000  01 00 5e 00 00 fb 50 6b  4b 4b df 3a 08 00 45 00   ··^···Pk KK·:··E·
0010  00 e0 0c 21 40 00 ff 11  83 c3 09 01 01 2c e0 00   ···!@··· ·····,··
0020  00 fb 14 e9 14 e9 00 cc  26 fc 00 00 00 00 00 04   ········ &·······
0030  00 03 00 00 00 00 1a 39  2d 31 2d 31 2d 35 35 3a   ·······9 -1-1-55:
0040  30 38 2f 32 37 2f 32 32  3a 30 31 3a 35 33 3a 30   08/27/22 :01:53:0
0050  35 0a 5f 6e 76 6d 65 2d  64 69 73 63 04 5f 74 63   5·_nvme- disc·_tc
0060  70 05 6c 6f 63 61 6c 00  00 21 00 01 08 39 2d 31   p·local· ·!···9-1
0070  2d 31 2d 35 35 c0 37 00  1c 00 01 c0 42 00 01 00   -1-55·7· ····B···
0080  01 c0 0c 00 10 00 01 c0  0c 00 10 00 01 00 00 11   ········ ········
0090  94 00 37 05 70 3d 74 63  70 30 4e 51 4e 3d 6e 71   ··7·p=tc p0NQN=nq
00a0  6e 2e 31 39 38 38 2d 31  31 2e 63 6f 6d 2e 64 65   n.1988-1 1.com.de
00b0  6c 6c 3a 53 46 53 53 3a  39 3a 32 30 32 32 30 38   ll:SFSS: 9:202208
00c0  32 34 32 32 33 30 35 38  65 38 c0 42 00 01 00 01   24223058 e8·B····
00d0  00 00 00 78 00 04 09 01  01 37 c0 0c 00 21 00 01   ···x···· ·7···!··
00e0  00 00 00 78 00 08 00 00  00 00 1f 49 c0 42 0f 87   ···x···· ···I·B··
00f0  12 4c                                              ·L
```

```
> Frame 10: 242 bytes on wire (1936 bits), 242 bytes captured (1936 bits) on interface \\.\pipe\vi
> Ethernet II, Src: Mellanox_4b:df:3a (50:6b:4b:4b:df:3a), Dst: IPv4mcast_fb (01:00:5e:00:00:fb)
> Internet Protocol Version 4, Src: 9.1.1.44, 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: 4
     Answer RRs: 3
     Authority RRs: 0
     Additional RRs: 0
   > Queries
   v Answers
      > 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type TXT, class IN
      > 9-1-1-55.local: type A, class IN, addr 9.1.1.55
      v 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type SRV, class IN, priority 0, weight 0,
           Service: 9-1-1-55:08/27/22:01:53:05
           Protocol: _nvme-disc
           Name: _tcp.local
           Type: SRV (Server Selection) (33)
           .000 0000 0000 0001 = Class: IN (0x0001)
           0... .... .... .... = Cache flush: False
           Time to live: 120 (2 minutes)
           Data length: 8
           Priority: 0
           Weight: 0
           Port: 8009
           Target: 9-1-1-55.local
     [Retransmitted request. Original request in: 4]
     [Retransmission: True]
```


Kamal Bakshi / Cisco Systems

Kamal Bakshi, Cisco

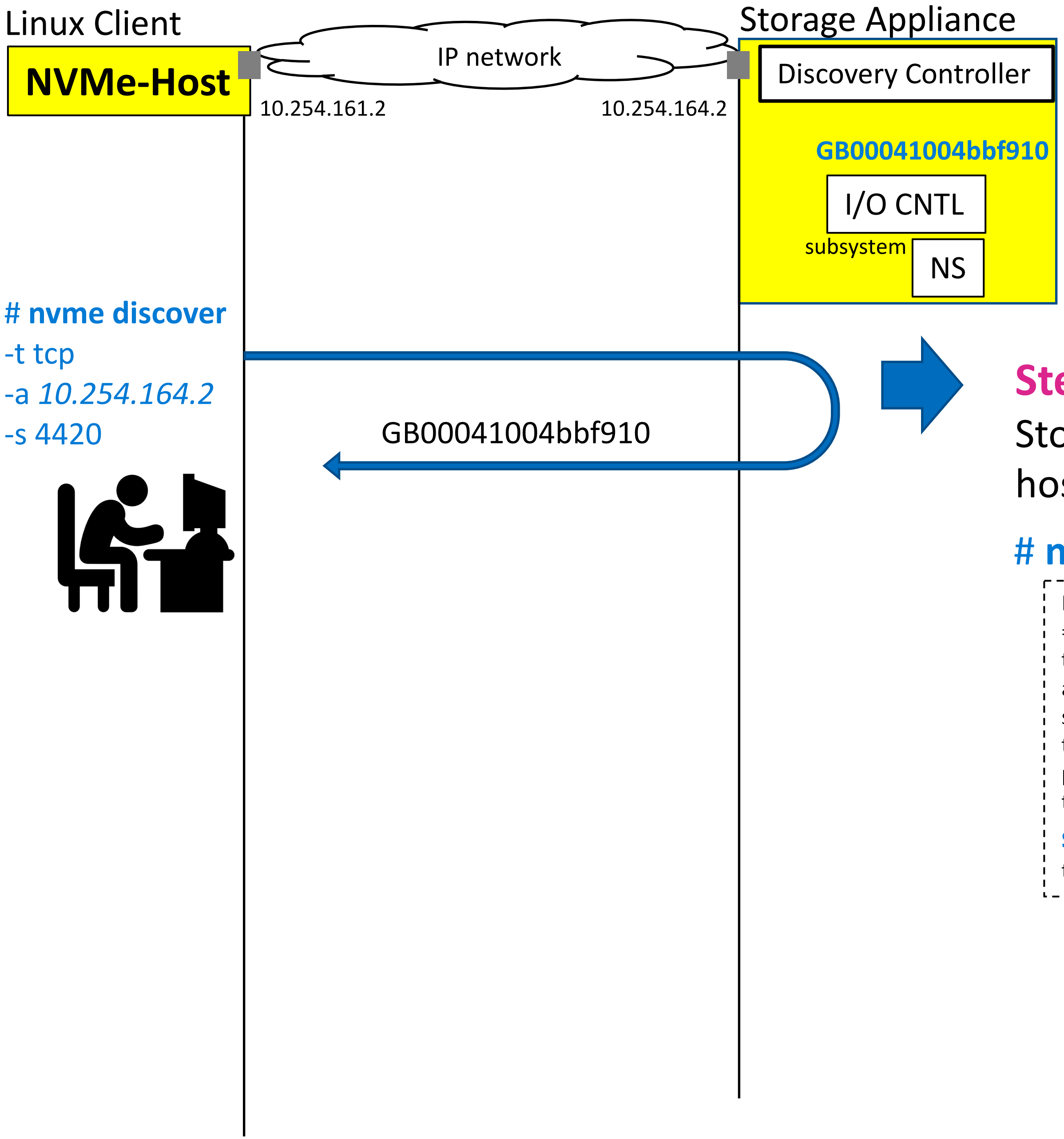
TCP Sync

No.	Source	Destination	Protocol	Info
22	9.1.1.44	9.1.1.55	TCP	33434 → 8009 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 SACK_PERM=1 TSval=666101698 TSecr=0 WS=128

```
0000  00 50 56 bf 37 26 50 6b  4b 4b df 3a 08 00 45 00   ·PV·7&Pk KK·:··E·
0010  00 3c a5 c8 40 00 40 06  80 8f 09 01 01 2c 09 01   ·<··@·@· ·····,··
0020  01 37 82 9a 1f 49 41 78  fb 50 00 00 00 00 a0 02   ·7···IAx ·P······
0030  fa f0 4a 88 00 00 02 04  05 b4 04 02 08 0a 27 b3   ··J····· ······'·
0040  e7 c2 00 00 00 00 01 03  03 07 9c 64 24 ee         ········ ··d$·
```

```
> Frame 22: 78 bytes on wire (624 bits), 78 bytes captured (624 bits) on interface \\.\pipe\view_
> Ethernet II, Src: Mellanox_4b:df:3a (50:6b:4b:4b:df:3a), Dst: VMware_bf:37:26 (00:50:56:bf:37:2
> Internet Protocol Version 4, Src: 9.1.1.44, Dst: 9.1.1.55
v Transmission Control Protocol, Src Port: 33434, Dst Port: 8009, Seq: 0, Len: 0
    Source Port: 33434
    Destination Port: 8009
    [Stream index: 0]
    [Conversation completeness: Incomplete, DATA (15)]
    [TCP Segment Len: 0]
    Sequence Number: 0    (relative sequence number)
    Sequence Number (raw): 1098447696
    [Next Sequence Number: 1    (relative sequence number)]
    Acknowledgment Number: 0
    Acknowledgment number (raw): 0
    1010 .... = Header Length: 40 bytes (10)
  v Flags: 0x002 (SYN)
      000. .... .... = Reserved: Not set
      ...0 .... .... = Nonce: Not set
      .... 0... .... = Congestion Window Reduced: Not set
      .... .0.. .... = ECN-Echo: Not set
      .... ..0. .... = Urgent: Not set
      .... ...0 .... = Acknowledgment: Not set
      .... .... 0... = Push: Not set
      .... .... .0.. = Reset: Not set
    > .... .... ..1. = Syn: Set
      .... .... ...0 = Fin: Not set
      [TCP Flags: ··········S·]
    Window: 64240
    [Calculated window size: 64240]
    Checksum: 0x4a88 [unverified]
    [Checksum Status: Unverified]
    Urgent Pointer: 0
  > Options: (20 bytes), Maximum segment size, SACK permitted, Timestamps, No-Operation (NOP), W
  > [Timestamps]
v TRANSUM RTE Data
    [RTE Status: OK]
    [Req First Seg: 22]
    [Req Last Seg: 22]
    [Rsp First Seg: 23]
```

Step-1 Discover the Storage Appliance



Step-1 (Discover the Storage Appliance)

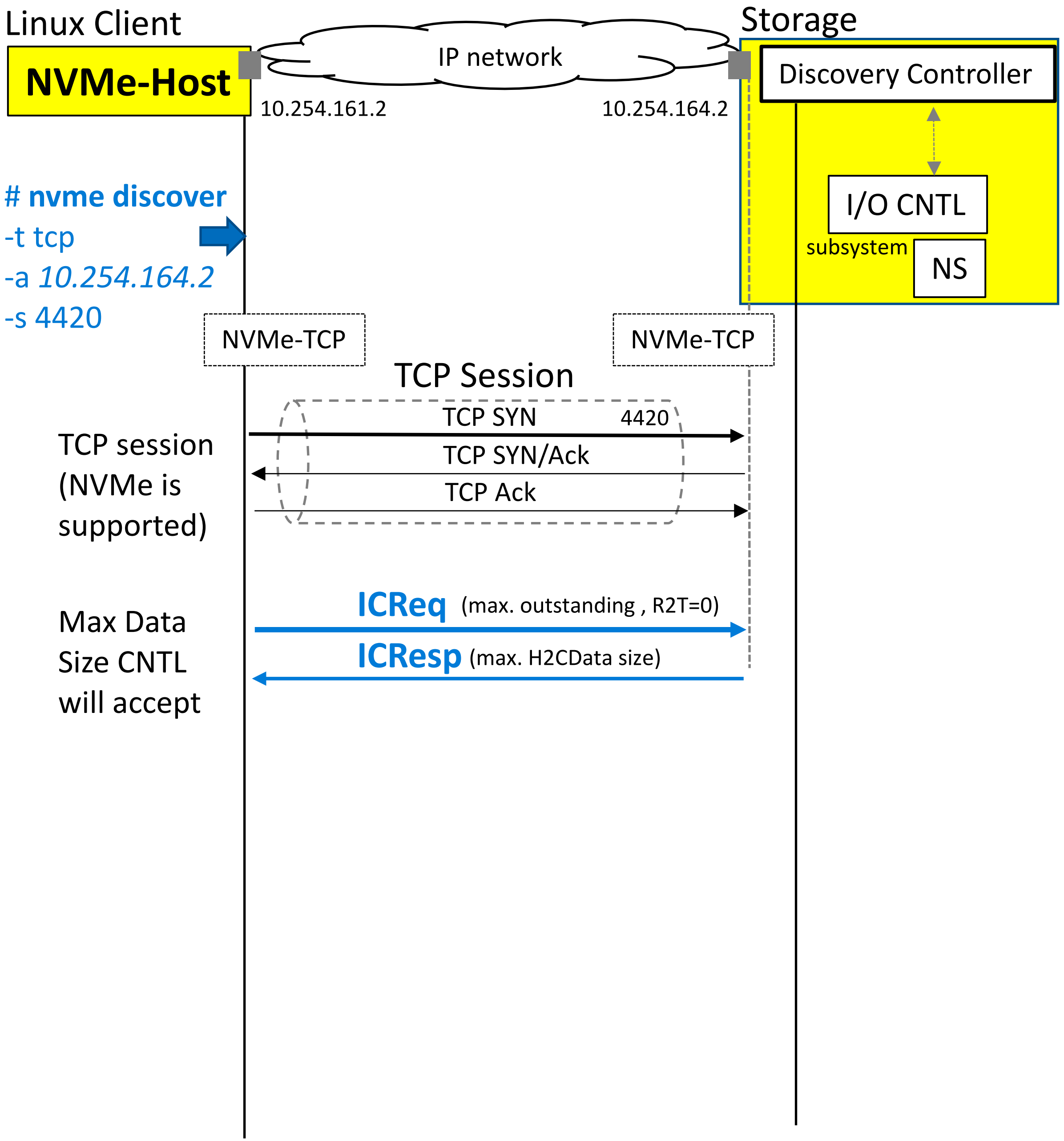
Storage Admin will issue a "NVMe discover" CLI command at the host to retrieve the Storage Appliance Subsystem.

 # nvme discover -t tcp -a *10.254.164.2* -s 4420

```
Discovery Log Number of Records 1, Generation counter 1
=====Discovery Log Entry 0======
trtype:  unrecognized
adrfam: ipv4
subtype: nvme subsystem
treq:    not specified
portid:  28
trsvcid: 4420
subnqn:  GB00041004bbf910
traddr:  10.254.164.2
```


Kamal Bakshi, Cisco

Initiate Connection Request to Discovery Cntl



ICReq

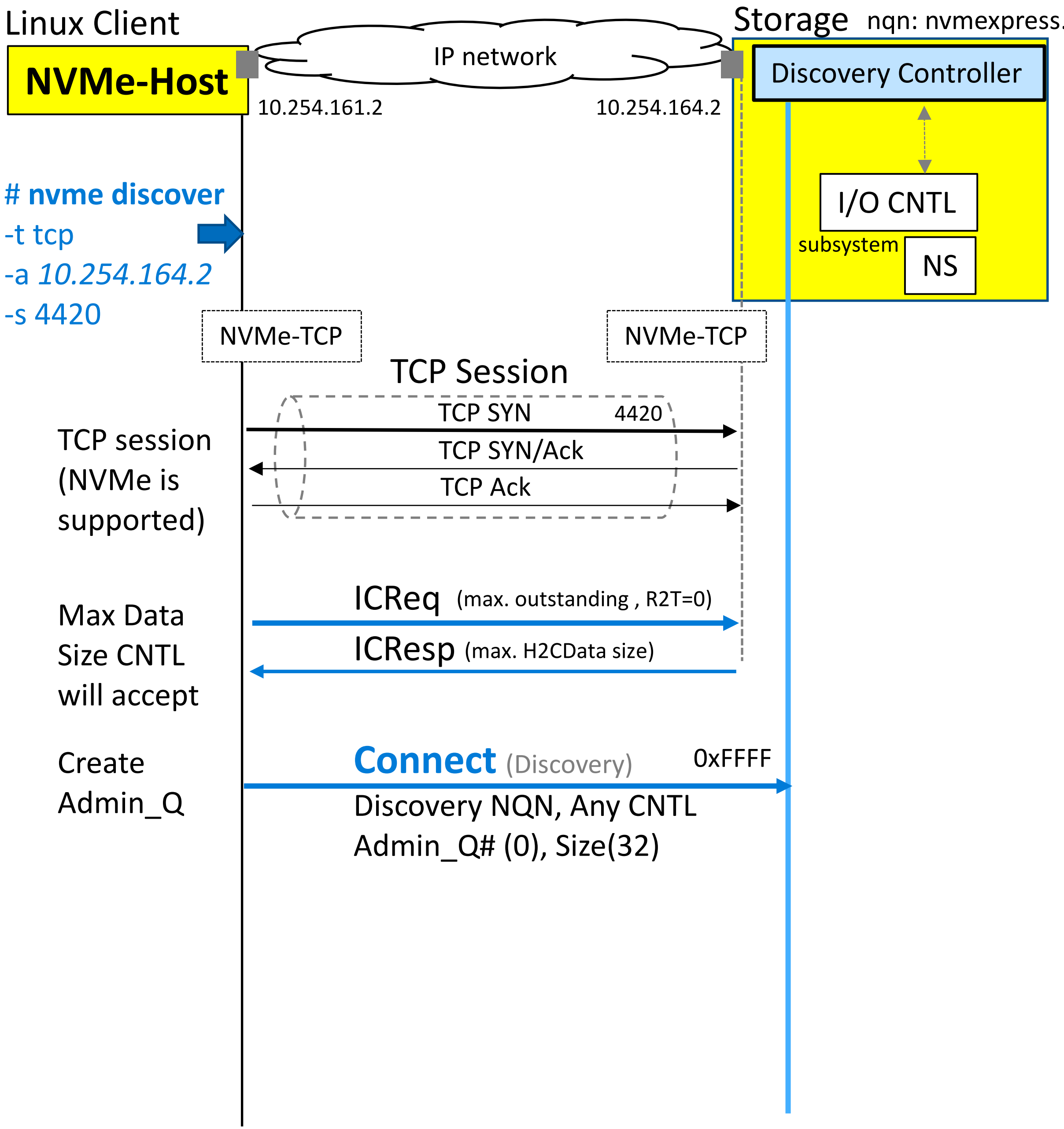
```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43944, Dst Port: 4420, Seq: 1, Ack: 1,
v NVM Express Fabrics TCP
    [Cmd Qid: 0 (AQ)]
    Pdu Type: ICReq (0)
  v Pdu Specific Flags: 0x00
      .... ...0 = PDU Header Digest: Not set
      .... ..0. = PDU Data Digest: Not set
      .... .0.. = PDU Data Last: Not set
      .... 0... = PDU Data Success: Not set
    Pdu Header Length: 128
    Pdu Data Offset: 0
    Packet Length: 128
  v ICReq
      Pdu Version Format: 0
      Host Pdu data alignment: 0
      Digest Types Enabled: 0
      Maximum r2ts per request: 0
```

ICResp

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43944, Seq: 1, Ack: 129,
v NVM Express Fabrics TCP
    [Cmd Qid: 0 (AQ)]
    Pdu Type: ICResp (1)
  v Pdu Specific Flags: 0x00
      .... ...0 = PDU Header Digest: Not set
      .... ..0. = PDU Data Digest: Not set
      .... .0.. = PDU Data Last: Not set
      .... 0... = PDU Data Success: Not set
    Pdu Header Length: 128
    Pdu Data Offset: 0
    Packet Length: 128
  v ICResp
      Pdu Version Format: 0
      Controller Pdu data alignment: 0
      Digest types enabled: 0
      Maximum data capsules per r2t supported: 65535
```


Kamal Bakshi, Cisco

Create NVMe Admin Queue

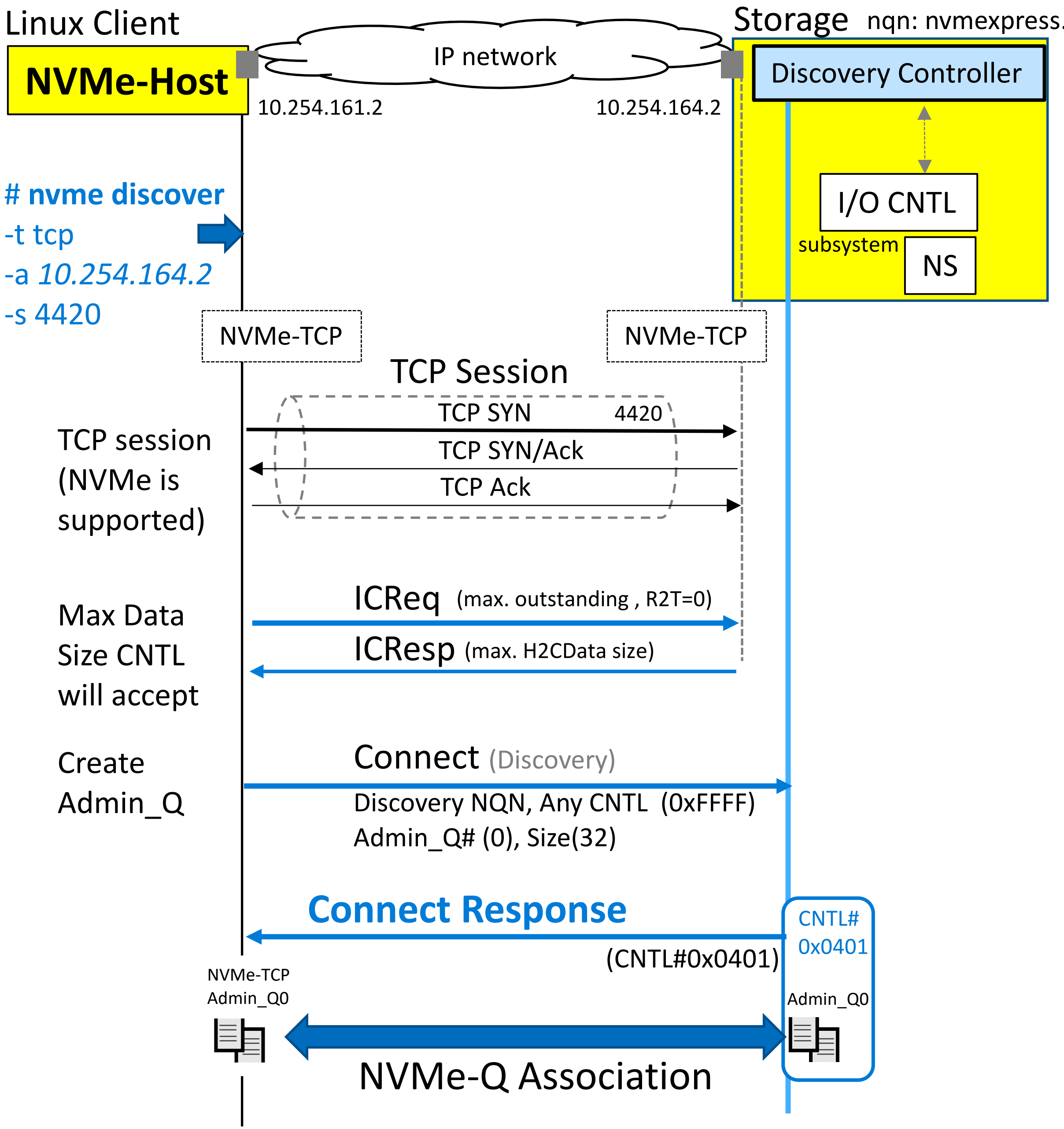


Connect Request

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43944, Dst Port: 4420, Seq: 129, Ack: 129, L
v NVM Express Fabrics TCP, Fabrics Type: Connect (0x01) Cmd ID: 0x0000
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 72
    Packet Length: 1096
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [Fabric Cqe in: 9]
      Reserved: 0x40
      Command Identifier: 0x0000
      Fabric Command Type: Connect (0x01)
      Reserved: 00000000000000000000000000000000000
    v SGL1
        0000 .... = Descriptor Type: 0x0 Data Block
        .... 0001 = Descriptor Sub Type: 0x1 Offset
        Address: 0x0000000000000000
        Length: 1024
        Reserved: 000000
      Record Format: 0
      Queue ID: 0 (AQ)
      Submission Queue Size: 32
    v Connect Attributes: 0x00
        .... ..00 = Priority Class: Urgent (0x0)
        .... .0.. = Disable SQ Flow Control: 0x0
        .... 0... = Support Deletion of IO Queues: 0x0
        0000 .... = Reserved: 0x0
      Reserved: 00
      Keep Alive Timeout: 0ms
      Reserved: 000000000000000000000000
  v Data
      Host Identifier: 741558b25cc841de933e4de51278ef74
      Controller ID: 0xffff
      Reserved: 00000000000000000000000000000000000000000000000000000000000000000000
      Subsystem NQN: nqn.2014-08.org.nvmexpress.discovery
      Host NQN: nqn.2014-08.org.nvmexpress:uuid:55e23c7b-7431-439b-81d7-430e286fbb98
      Reserved: 00000000000000000000000000000000000000000000000000000000000000000000
```


Kamal Bakshi, Cisco

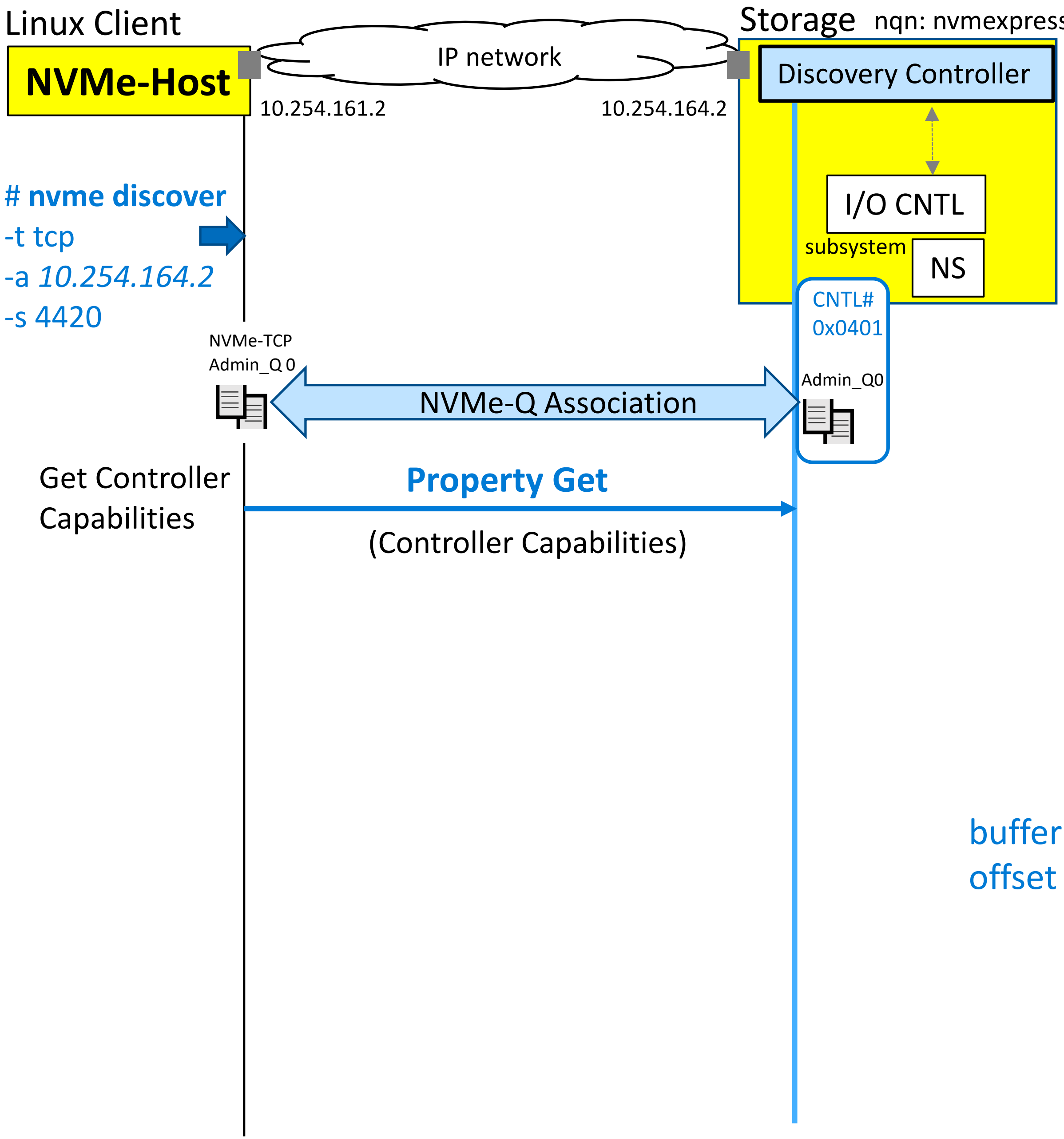
Create NVMe Admin Queue



Connect Response

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43944, Seq: 129, Ack:
v NVM Express Fabrics TCP, Cqe Fabrics Cmd: Connect (0x01) Cmd ID: 0x0000
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
  v Pdu Specific Flags: 0x00
      .... ...0 = PDU Header Digest: Not set
      .... ..0. = PDU Data Digest: Not set
      .... .0.. = PDU Data Last: Not set
      .... 0... = PDU Data Success: Not set
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
  v Cqe (For Cmd: Connect)
      [Fabric Cmd in: 8]
      [Cmd Latency: 24.476 ms]
      Controller ID: 0x0401
      Authentication Required: 0x0000
      Reserved: 00000000
      SQ Head Pointer: 0x0000
      SQ Identifier: 0x0000
      Command Identifier: 0x0000
    v Status Field: 0x0000
        .... .... .... ...0 = Reserved: 0x0
        .... ...0 0000 000. = Status Code: 0x00 (Successful Completion)
        .... 000. .... .... = Status Code Type: Generic Command Status (0x0)
        ..00 .... .... .... = Command Retry Delay: 0x0
        .0.. .... .... .... = More Information in Log Page: False
        0... .... .... .... = Do not Retry: False
```

Get Controller's Capabilities



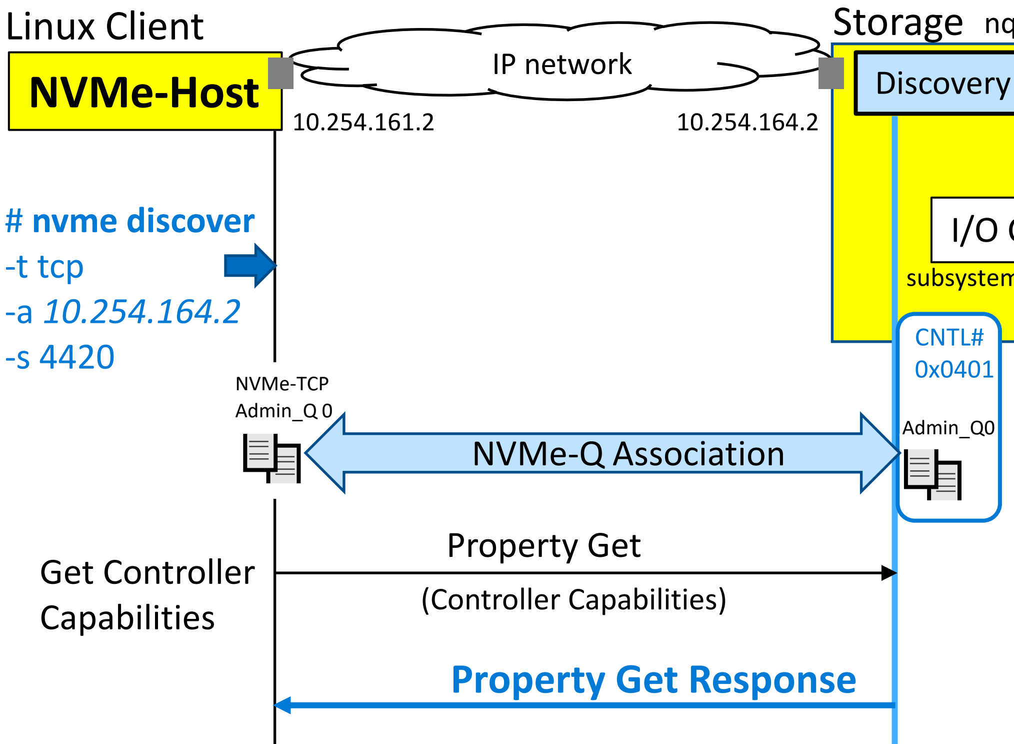
Property Get

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43944, Dst Port: 4420, Seq: 1225,
v NVM Express Fabrics TCP, Fabrics Type: Property Get (0x04) Cmd ID: 0x0012
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
  v Pdu Specific Flags: 0x00
      .... ...0 = PDU Header Digest: Not set
      .... ..0. = PDU Data Digest: Not set
      .... .0.. = PDU Data Last: Not set
      .... 0... = PDU Data Success: Not set
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [Fabric Cqe in: 12]
      Reserved: 0x40
      Command Identifier: 0x0012
      Fabric Command Type: Property Get (0x04)
      Reserved: 000000000000000000000000000000000000000000000000000000000000
    v .... .001 = Attributes: 0x1
        .... .001 = Property Size: 8 bytes (0x1)
        0000 0... = Reserved: 0x00
      Reserved: 000000
      Offset: Controller Capabilities (0x00000000)
      Reserved: 00000000000000000000000000000000
```

buffer
offset

Kamal Bakshi, Cisco

Get Controller's Capabilities



Controller Capabilities

- -Maximum Queue Entry =128
- -Arbitration Mechanism
- -Timeout
- -Command Sets Supported

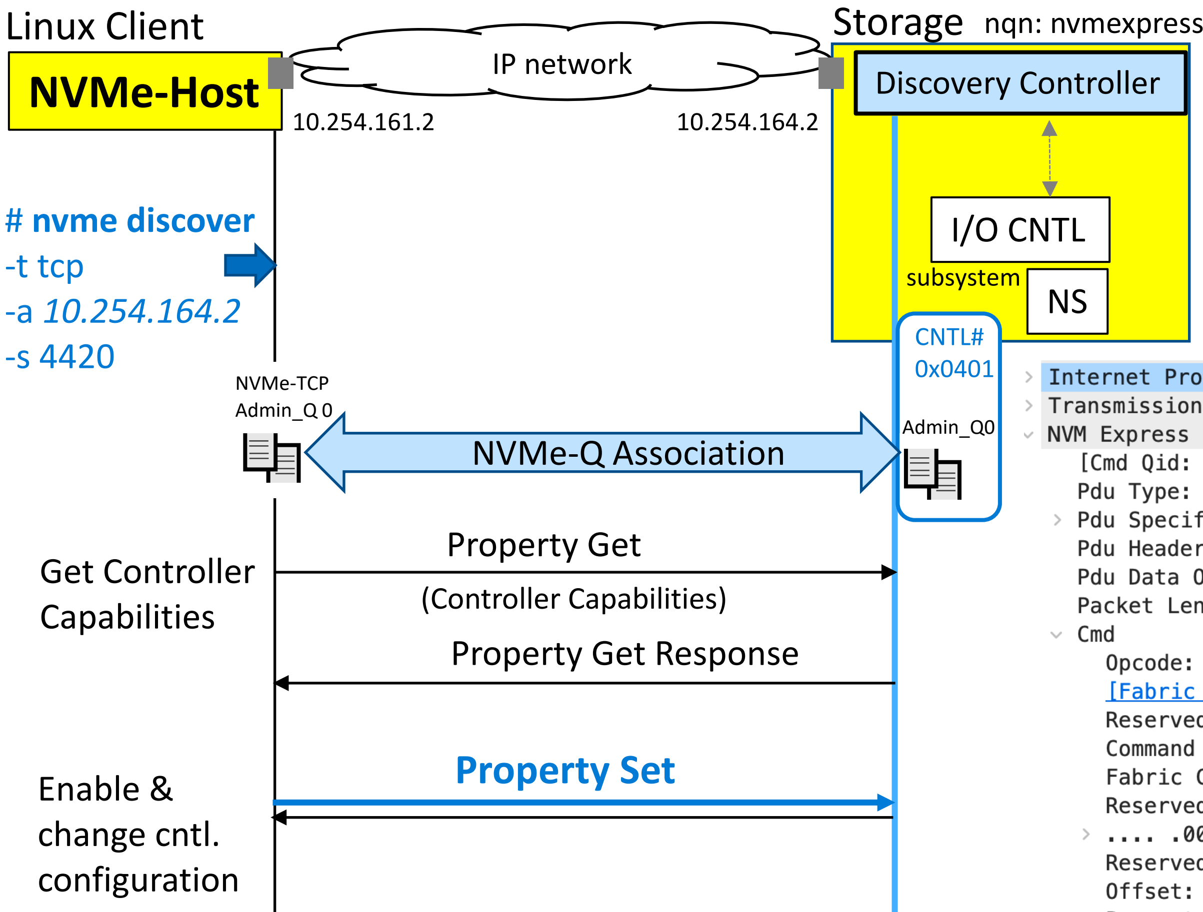
Property Get Response

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43944, Seq: 153, Ack: 1297, Len: 24
v NVM Express Fabrics TCP, Cqe Fabrics Cmd: Property Get (0x04) Cmd ID: 0x0012
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
  v Cqe (For Cmd: Property Get)
      [Fabric Cmd in: 11]
      [Cmd Latency: 0.018 ms]
    v Property Data: 7f00000f20000000
      v Controller Capabilities: 0x000000200f00007f
          .... .... .... .... .... .... .... .... .... .... .... .... 0000 0000 0111 1111 = Maximum Queue Entries Supported: 128
          .... .... .... .... .... .... .... .... .... .... .... ...0 .... .... .... .... = Contiguous Queues Required: False
          .... .... .... .... .... .... .... .... .... .... .... ..0. .... .... .... .... = Supports Arbitration Mechanism with Weighted Round Robin
          .... .... .... .... .... .... .... .... .... .... .... .0.. .... .... .... .... = Supports Arbitration Mechanism Vendor Specific: False
          .... .... .... .... .... .... .... .... .... .... 0000 0... .... .... .... .... = Reserved: 0x00
          .... .... .... .... .... .... .... .... 0000 1111 .... .... .... .... .... .... = Timeout (to ready status): f (7500 ms)
          .... .... .... .... .... .... .... 0000 .... .... .... .... .... .... .... .... = Doorbell Stride: 0x0 (4 bytes)
          .... .... .... .... .... .... ...0 .... .... .... .... .... .... .... .... .... = NVM Subsystem Reset Supported: False
          .... .... .... .... ...0 0000 001. .... .... .... .... .... .... .... .... .... = Command Sets Supported: 1 (NVM IO Command Set)
          .... .... .... .... ..0. .... .... .... .... .... .... .... .... .... .... .... = Boot Partition Support: False
          .... .... .... .... 00.. .... .... .... .... .... .... .... .... .... .... .... = Reserved: 0x0
          .... .... .... 0000 .... .... .... .... .... .... .... .... .... .... .... .... = Memory Page Size Minimum: 0x0 (4096 bytes)
          .... .... 0000 .... .... .... .... .... .... .... .... .... .... .... .... .... = Memory Page Size Maximum: 0x0 (4096 bytes)
          .... ...0 .... .... .... .... .... .... .... .... .... .... .... .... .... .... = Persistent Memory Region Supported: False
          .... ..0. .... .... .... .... .... .... .... .... .... .... .... .... .... .... = Controller Memory Buffer Supported: False
          0000 00.. .... .... .... .... .... .... .... .... .... .... .... .... .... .... = Reserved: 0x00
      SQ Head Pointer: 0x0000
      SQ Identifier: 0x0000
      Command Identifier: 0x0012
    v Status Field: 0x0000
        .... .... .... ...0 = Reserved: 0x0
        .... ...0 0000 000. = Status Code: 0x00 (Successful Completion)
        .... 000. .... .... = Status Code Type: Generic Command Status (0x0)
        ..00 .... .... .... = Command Retry Delay: 0x0
        .0.. .... .... .... = More Information in Log Page: False
        0... .... .... .... = Do not Retry: False
```

Controller Capabilities

Kamal Bakshi, Cisco

SET Controller's Configuration



Set Controller Configuration

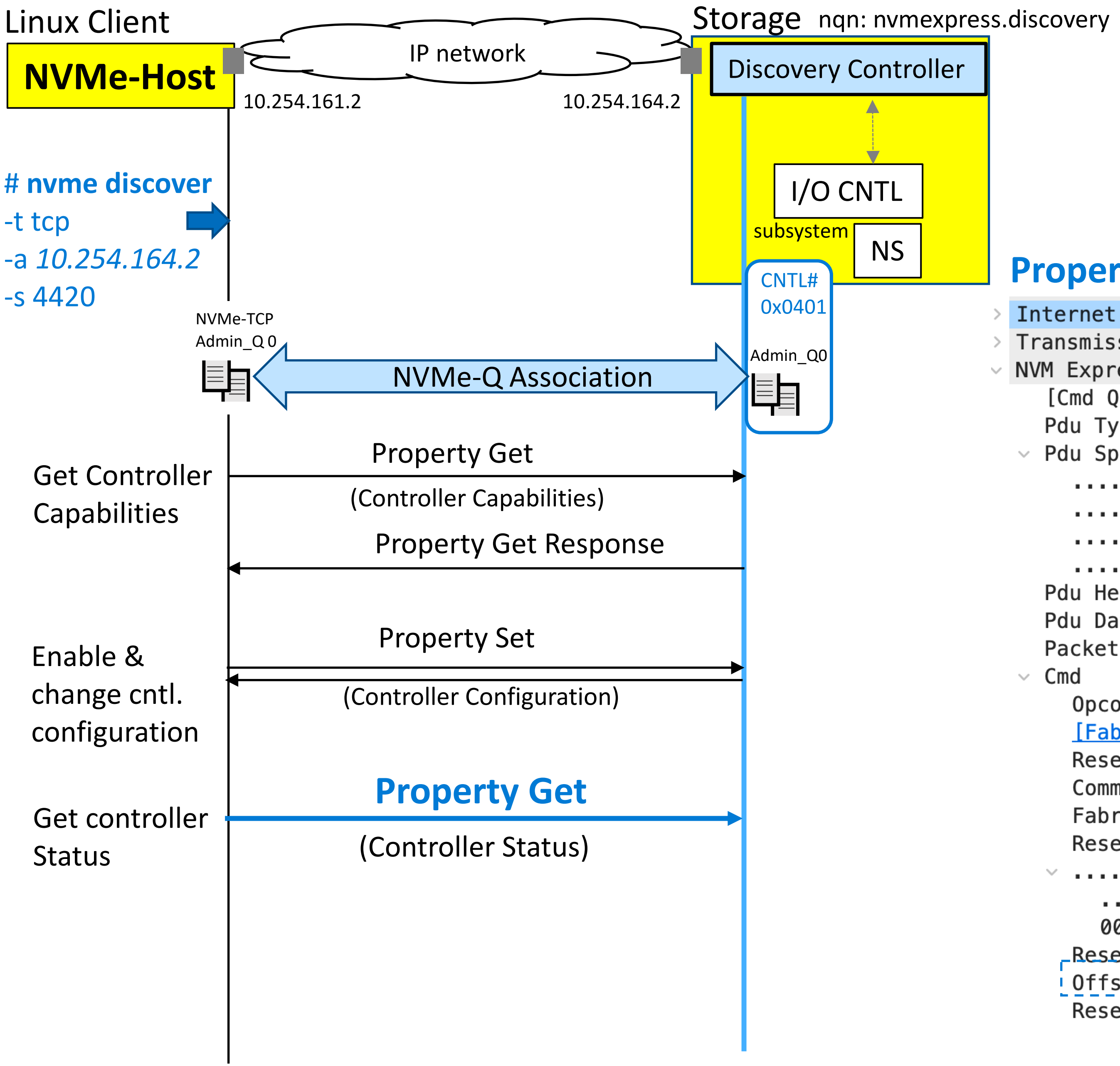
- -Enable
- -NVM command set =0
- -Memory Page size: 4096
- -Arbitration =Round Robin
- -I/O SQ/CQ Size

Property Set Request

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43944, Dst Port: 4420, Seq: 1297, Ack: 177, Len: 72
v NVM Express Fabrics TCP, Fabrics Type: Property Set (0x00) Cmd ID: 0x0013
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [Fabric Cqe in: 15]
      Reserved: 0x40
      Command Identifier: 0x0013
      Fabric Command Type: Property Set (0x00)
      Reserved: 0000000000000000000000000000000000000000000000000000000000000000005a
    > .... .000 = Attributes: 0x0
      Reserved: 000000
      Offset: Controller Configuration (0x00000014)
    v Property Data: 0100460000000000
      v Controller Configuration: 0x00460001
          .... .... .... .... .... .... .... ...1 = Enable: 0x1
          .... .... .... .... .... .... .... 000. = Reserved: 0x0
          .... .... .... .... .... .... .000 .... = IO Command Set Selected: NVM IO Command Set (0x0)
          .... .... .... .... .... .000 0... .... = Memory Page Size: 0x0 (4096 bytes)
          .... .... .... .... ..00 0... .... .... = Arbitration Mechanism Selected: Round Robin (0x0)
          .... .... .... .... 00.. .... .... .... = Shutdown Notification: No Shutdown (0x0)
          .... .... .... 0110 .... .... .... .... = IO Submission Queue Entry Size: 0x6 (64 bytes)
          .... .... 0100 .... .... .... .... .... = IO Completion Queue Entry Size: 0x4 (16 bytes)
          0000 0000 .... .... .... .... .... .... = Reserved: 0x00
        Reserved: 0x00000000
      Reserved: 0000000000000000
```


Kamal Bakshi, Cisco

Get Controller’s Status

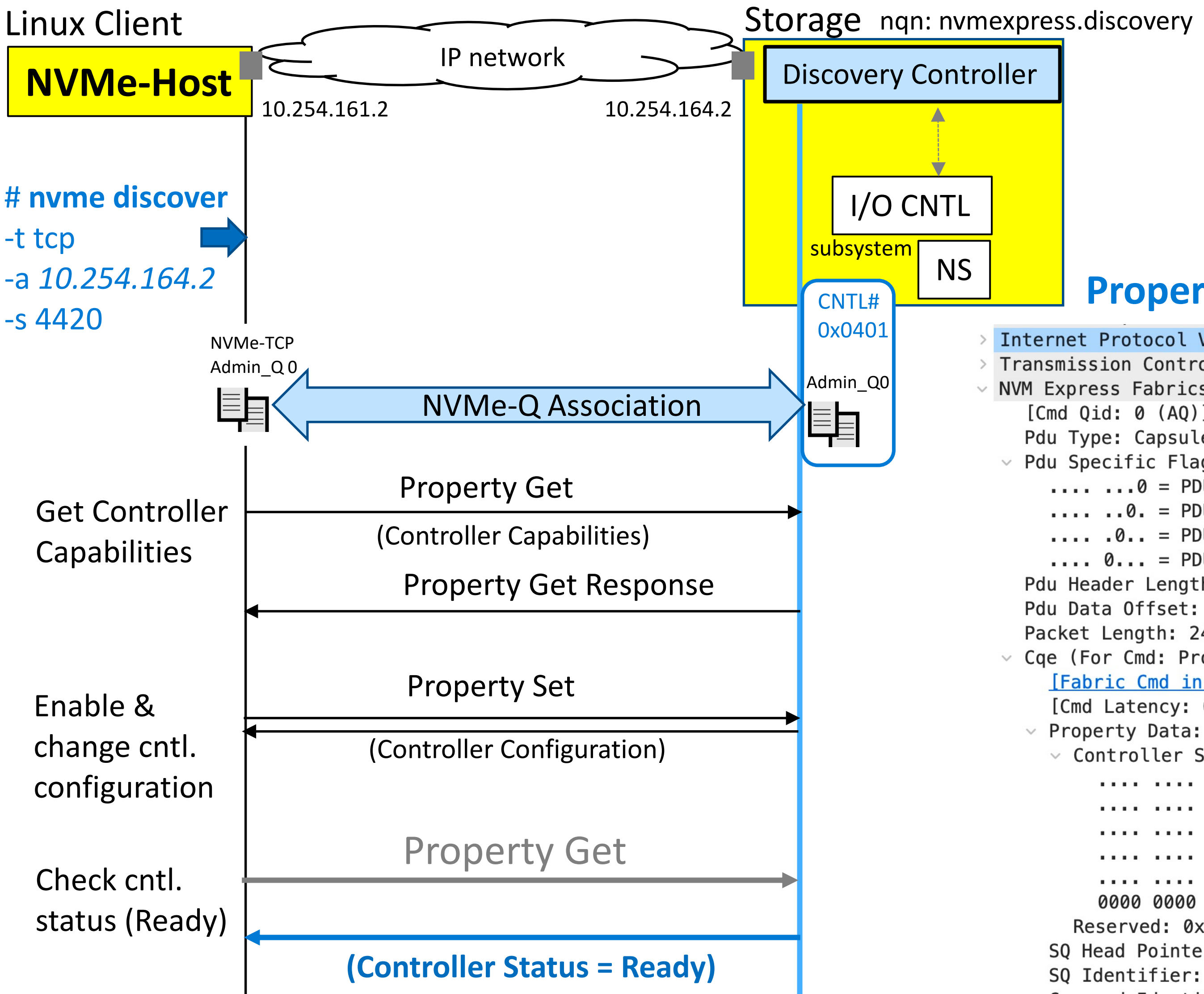


Property Get Request

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43944, Dst Port: 4420, Seq: 1369,
v NVM Express Fabrics TCP, Fabrics Type: Property Get (0x04) Cmd ID: 0x0014
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
  v Pdu Specific Flags: 0x00
      .... ...0 = PDU Header Digest: Not set
      .... ..0. = PDU Data Digest: Not set
      .... .0.. = PDU Data Last: Not set
      .... 0... = PDU Data Success: Not set
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [Fabric Cqe in: 18]
      Reserved: 0x40
      Command Identifier: 0x0014
      Fabric Command Type: Property Get (0x04)
      Reserved: 00000000000000000000000000000000000000000000000000000000000
    v .... .000 = Attributes: 0x0
        .... .000 = Property Size: 4 bytes (0x0)
        0000 0... = Reserved: 0x00
      Reserved: 000000
      Offset: Controller Status (0x0000001c)
      Reserved: 00000000000000000000000000000000
```


Kamal Bakshi, Cisco

Get Controller’s Status



Property Get Response

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43944, Seq: 201, Ack: 1441, Le
v NVM Express Fabrics TCP, Cqe Fabrics Cmd: Property Get (0x04) Cmd ID: 0x0014
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
  v Pdu Specific Flags: 0x00
      .... ...0 = PDU Header Digest: Not set
      .... ..0. = PDU Data Digest: Not set
      .... .0.. = PDU Data Last: Not set
      .... 0... = PDU Data Success: Not set
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
  v Cqe (For Cmd: Property Get)
      [Fabric Cmd in: 17]
      [Cmd Latency: 0.014 ms]
    v Property Data: 0100000000000000
      v Controller Status: 0x00000001
          .... .... .... .... .... .... .... ...1 = Ready: 0x1
          .... .... .... .... .... .... .... ..0. = Controller Fatal Status: 0x0
          .... .... .... .... .... .... .... 00.. = Shutdown Status: No Shutdown (0x0)
          .... .... .... .... .... .... ...0 .... = NVM Subsystem Reset Occurred: 0x0
          .... .... .... .... .... .... ..0. .... = Processing Paused: 0x0
          0000 0000 0000 0000 0000 0000 00.. .... = Reserved: 0x0000000
        Reserved: 0x00000000
      SQ Head Pointer: 0x0000
      SQ Identifier: 0x0000
      Command Identifier: 0x0014
    v Status Field: 0x0000
        .... .... .... ...0 = Reserved: 0x0
        .... ...0 0000 000. = Status Code: 0x00 (Successful Completion)
        .... 000. .... .... = Status Code Type: Generic Command Status (0x0)
        ..00 .... .... .... = Command Retry Delay: 0x0
        .0.. .... .... .... = More Information in Log Page: False
        0... .... .... .... = Do not Retry: False
```

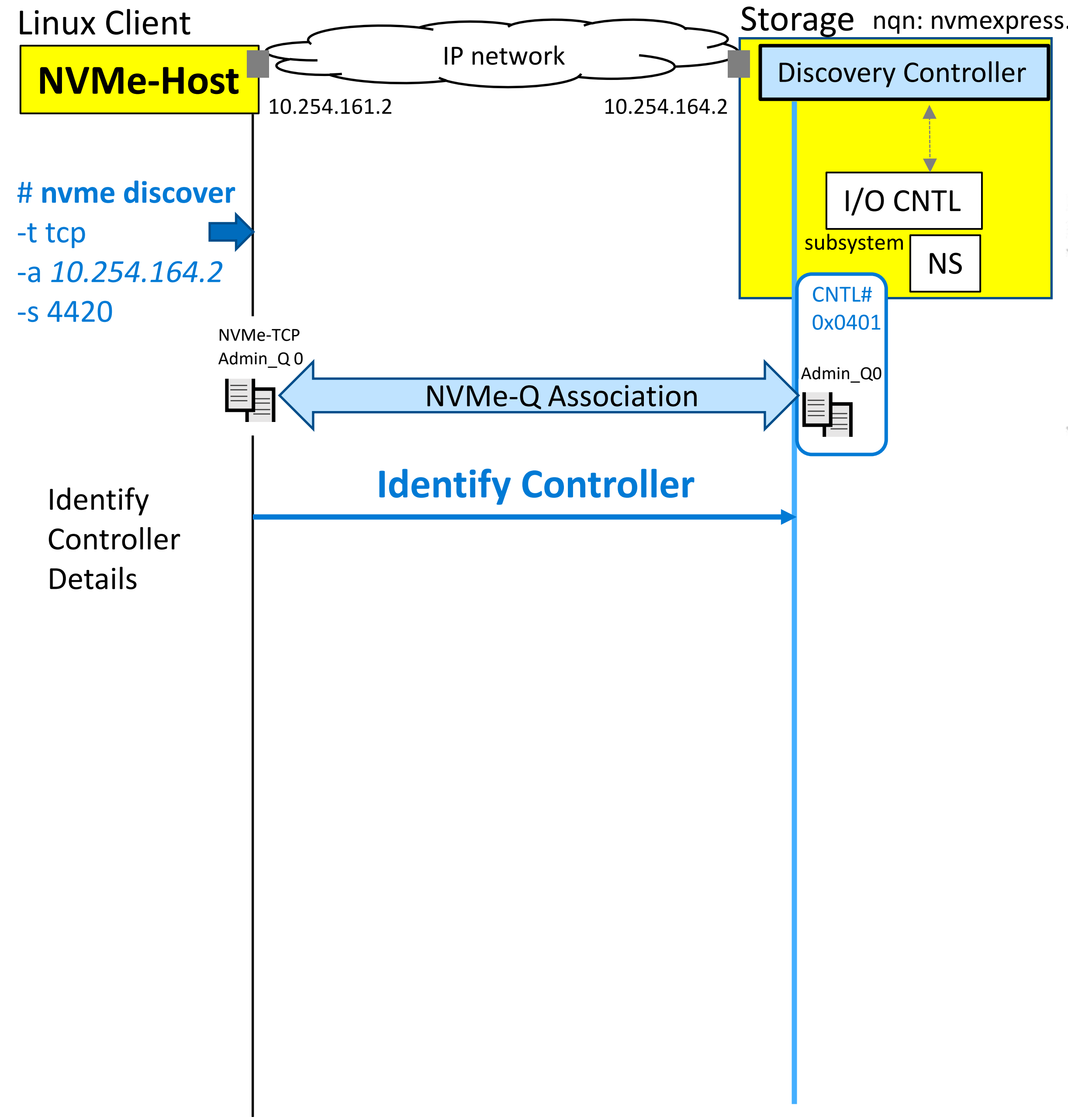
Controller Status

-Ready

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

Identify Controller CNS-01

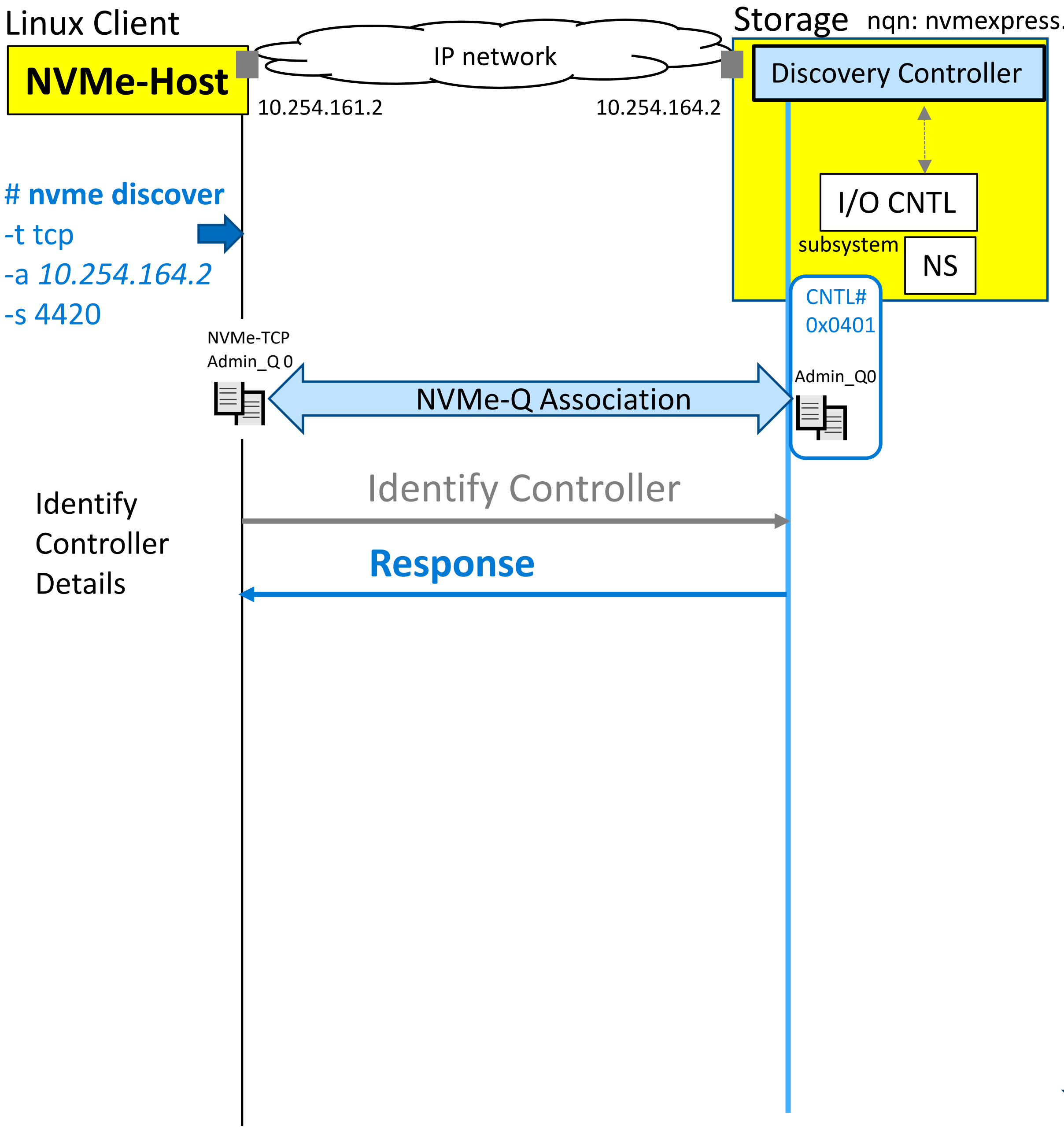


Identify Request

```
Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
Transmission Control Protocol, Src Port: 43944, Dst Port: 4420, Seq: 1585, Ack: 273, Le
NVM Express Fabrics TCP, NVMe Opcode: Identify (0x06) Cmd ID: 0x0013
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
NVM Express (Cmd)
    Opcode: 0x06 Identify
    [DATA Transfer 0: 27]
    [Cqe in: 27]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x0013
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
    SGL1
        0101 .... = Descriptor Type: 0x5 Reserved
        .... 1010 = Descriptor Sub Type: 0xa Reserved
    DWORD10: 0x00000001
        .... .... .... .... .... .... 0000 0001 = Controller or Namespace Structure (CNS):
        .... .... .... .... 0000 0000 .... .... = Reserved: 0x00
        0000 0000 0000 0000 .... .... .... .... = Controller Identifier (CNTID): 0x0000
    DWORD11: 0x00000000
        .... .... .... .... 0000 0000 0000 0000 = NVM Set Identifier (NVMSETID): 0x0000
        0000 0000 0000 0000 .... .... .... .... = Reserved: 0x0000
    DWORD12: 0x00000000
    DWORD13: 0x00000000
    DWORD14: 0x00000000
        .... .... .... .... .... .... .000 0000 = UUID Index: 0x00
        0000 0000 0000 0000 0000 0000 0... .... = UUID Index: 0x0000000
    DWORD15: 0x00000000
```


Kamal Bakshi, Cisco

Identify Data Buffer



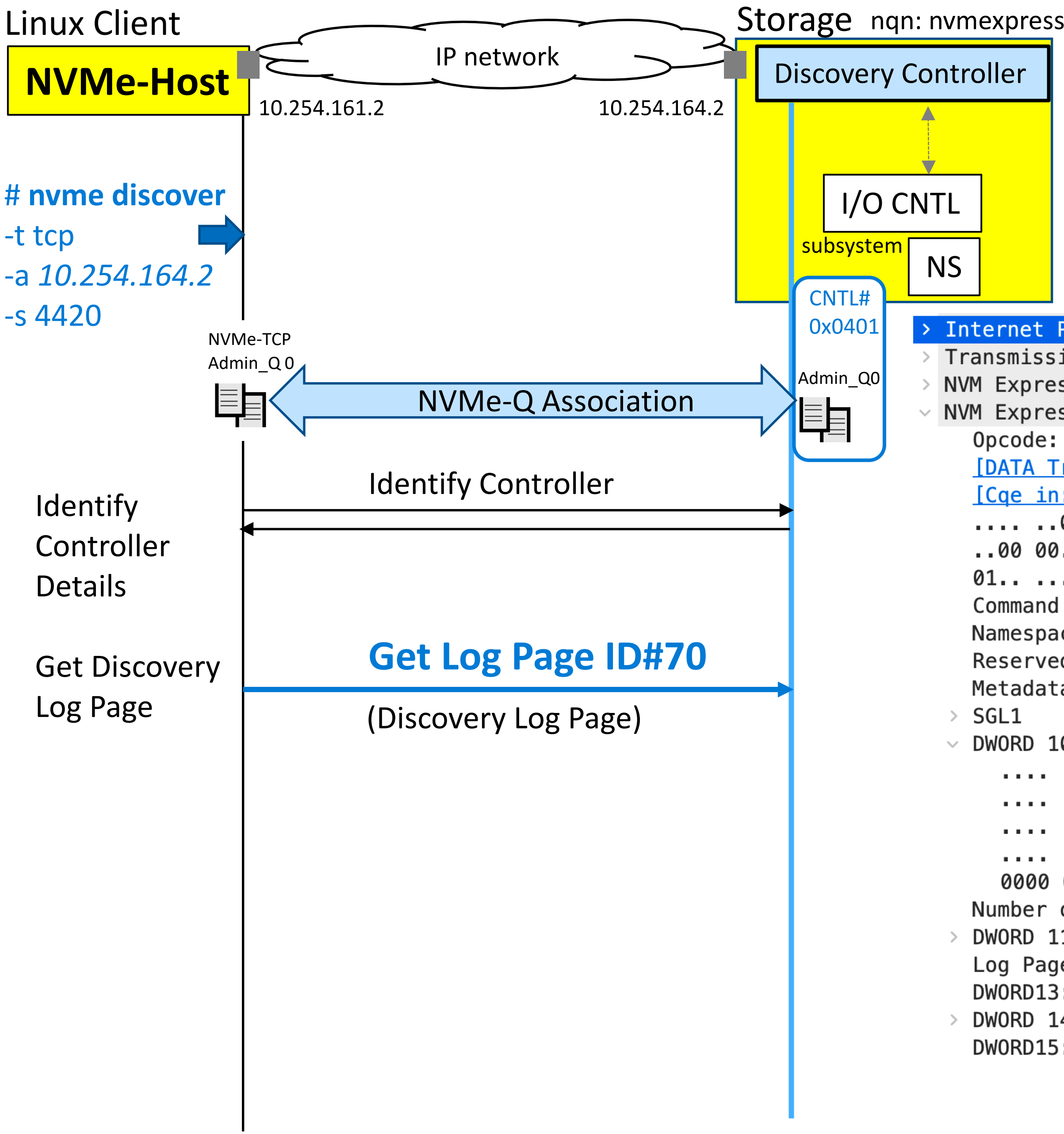
Identify Response

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43944, Seq: 273, Ack:
> NVM Express Fabrics TCP, C2HData Opcode: Identify (0x06), Cmd ID: 0x0013, Len:
v NVM Express
    PCI Vendor ID (VID): 0x0000
    PCI Subsystem Vendor ID (SSVID): 0x0000
    Serial Number (SN):
    Model Number (MN):
    Firmware Revision (FR): 342e372e302b0000
    Recommended Arbitration Burst (RAB): 0x0 (1 command)
    IEEE OUI Identifier (IEEE): 0x00000000
  > Controller Multi-Path I/O and Namespace Sharing Capabilities (CMIC): 0x00
    Maximum Data Transfer Size (MDTS): 0x4 (16 pages)
    Controller ID (CNTLID): 0x0401
  > Version (VER): 0x00010200
    RTD3 Resume Latency (RTD3R): 0 (not reported)
    RTD3 Entry Latency (RTD3E): 0 (not reported)
  > Optional Asynchronous Events Supported (OAES): 0x00000000
  > Controller Attributes (CTRATT): 0x00000000
  > Read Recovery Levels Support (RRLS): 0x0000
    Reserved: 0000000000000000
    Controller Type (CNTRLTYPE): Reserved (not reported) (0x00)
  > FRU Globally Unique Identifier (FGUID): 00000000000000000000000000000000
    Command Retry Delay Time 1: 0 (0 ms)
    Command Retry Delay Time 2: 0 (0 ms)
    Command Retry Delay Time 3: 0 (0 ms)
    Reserved: 000000000000000000000000000000000000000000000000000000000000000000000000
  > NVMe Management Interface: 00000000000000000000000000000000
  > Optional Admin Command Support (OACS): 0x0000
    Abort Command Limit (ACL): 0x0: (1 command)
    Asynchronous Event Request Limit (AERL): 0x0: (1 event)
  > Firmware Updates (FRMW): 0x00
  > Log Page Attributes (LPA): True
    Error Log Page Entries (ELPE): 0x0: (1 entry)
    Number of Power States Supported (NPSS): 0x0: (1 state)
  > Admin Vendor Specific Command Configuration (AVSCC): 0x00
  > Autonomous Power State Transition Attributes (APSTA): 0x00
```

multiple pages (not shown)

Kamal Bakshi, Cisco

Get Discovery Log Page

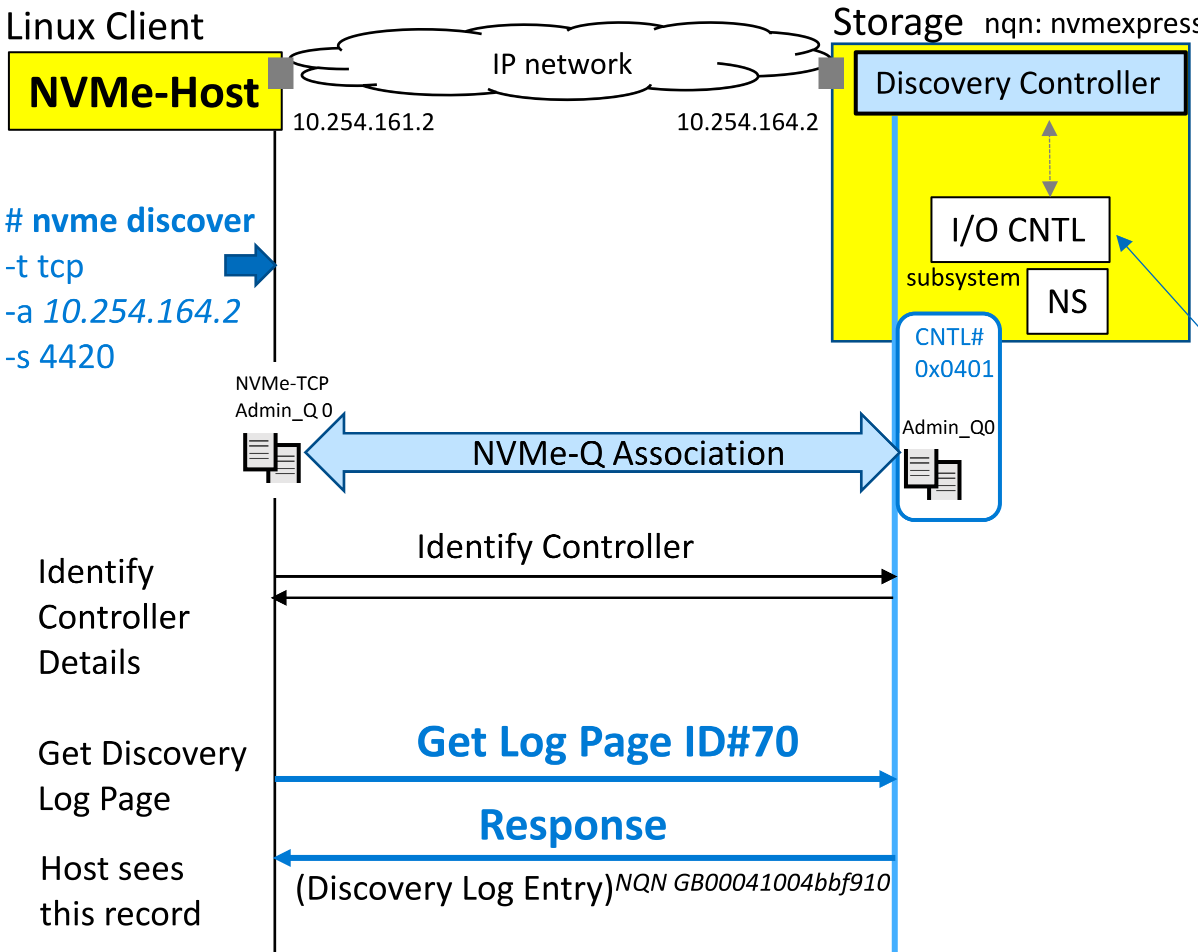


Get Log Page

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43944, Dst Port: 4420, Seq: 1729, Ack: 4481, Len: 72
> NVM Express Fabrics TCP, NVMe Opcode: Get Log Page (0x02) Cmd ID: 0x0015
v NVM Express (Cmd)
    Opcode: 0x02 Get Log Page
    [DATA Transfer 0: 33]
    [Cqe in: 33]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x0015
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  > SGL1
  v DWORD 10: 0x01ff0070
      .... .... .... .... .... .... 0111 0000 = Log Page Identifier (LID): NVMeOF Discovery (0x70)
      .... .... .... .... ...0 0000 .... .... = Log Specific Field (LSP): 0x00
      .... .... .... .... .00. .... .... .... = Reserved: 0x0
      .... .... .... .... 0... .... .... .... = Retain Asynchronous Event (RAE): False
      0000 0001 1111 1111 .... .... .... .... = Number of Dwords Lower (NUMDL): 0x01ff
    Number of Dwords: 511 (2048 bytes)
  > DWORD 11: 0x00000000
    Log Page Offset (DWORD 12 and DWORD 13): 0
    DWORD13: 0x00000000
  > DWORD 14: 0x00000000
    DWORD15: 0x00000000
```


Kamal Bakshi, Cisco

Discovery Log Entry



Response

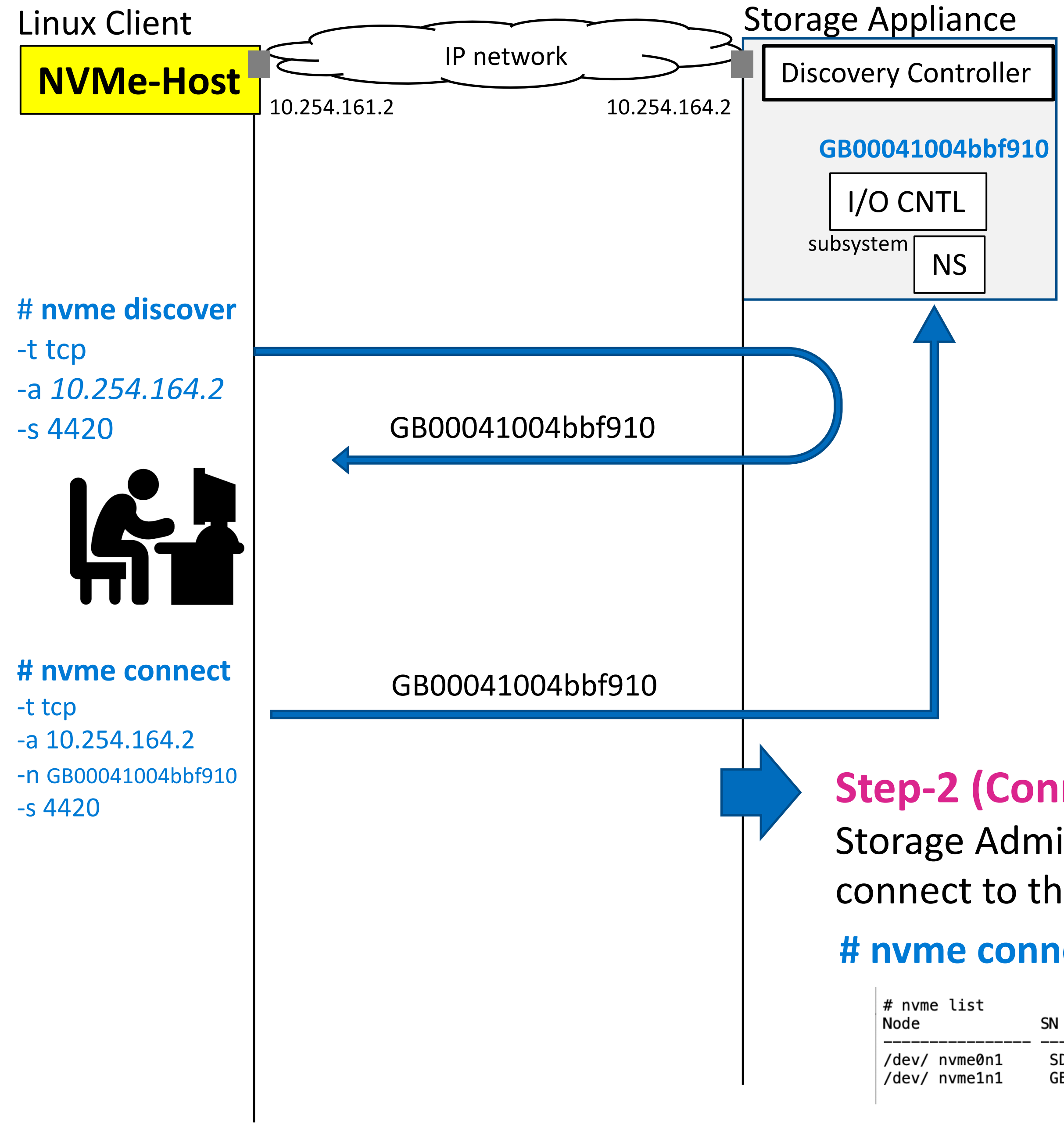
```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43944, Seq: 4481, Ack: 1801, Len: 2096
v NVM Express Fabrics TCP, C2HData Opcode: Get Log Page (0x02), Cmd ID: 0x0015, Len: 2048
    [Cmd Qid: 0 (AQ)]
    Pdu Type: C2HData (7)
  > Pdu Specific Flags: 0x04, PDU Data Last
    Pdu Header Length: 24
    Pdu Data Offset: 24
    Packet Length: 2072
  > NVMe/TCP Data PDU
    [Cmd in: 32]
v NVM Express
  v Nvme Data: NVMe Get Log Page (NVMeOF Discovery)
      Generation Counter (GENCTR): 1
      Number of Records (NUMREC): 1
      Record Format (RECFMT): 0
      Reserved: 0000000000000000000000000000000000000000000000000000000000000000…
    v Discovery Log Entry 0 (DLE0)
        Transport Type (TRTYPE): TCP Transport (0x03)
        Address Family (ADRFAM): AF_INET (0x01)
        Subsystem Type (SUBTYPE): NVM System with IO controllers (0x02)
      v Transport Requirements (TREQ): 0x00
          .... ...0 = Secure Channel Connection Requirement: False
          .... ..0. = Disable SQ Flow Control Support: False
          0000 0... = Reserved: 0x00
        Port ID (PORTID): 0x001c
        Controller ID (CNTLID): 0xffff
        Admin Max SQ Size (ASQSZ): 32
        Reserved: 0000000000000000000000000000000000000000000
        Transport Service Identifier (TRSVCID): 4420
        Reserved: 0000000000000000000000000000000000000000000000000000000000000000…
        NVM Subsystem Qualified Name (SUBNQN): GB00041004bbf910
        Transport Address (TRADDR): 10.254.164.2
      v Transport Specific Address Subtype (TSAS): 000000000000000000000000000000000000000000000000000
          Security Type (SECTYPE): No security (0x00)
          Reserved: 0000000000000000000000000000000000000000000000000000000000000000…
> NVM Express Fabrics TCP, Cqe NVMe Cmd: Get Log Page (0x02) Cmd ID: 0x0015
> NVM Express (Cqe)
```

I/O Subsystem NQN

Discovery Log Number of Records 1, Generation counter 1
=====Discovery Log Entry 0======
trtype: unrecognized
adrfam: ipv4
subtype: nvme subsystem
treq: not specified
portid: 28
trsvcid: 4420
subnqn: GB00041004bbf910
traddr: 10.254.164.2

Kamal Bakshi, Cisco

Step-2 Connect to I/O Subsystem



Step-1 (Discover the Storage Appliance)

Storage Admin will issue a “NVMe discover” CLI command at the host to retrieve the Storage Appliance Subsystem.

 # nvme discover -t tcp -a *10.254.164.2* -s 4420

```
Discovery Log Number of Records 1, Generation counter 1
=====Discovery Log Entry 0======
trtype:  unrecognized
adrfam: ipv4
subtype: nvme subsystem
treq:    not specified
portid:  28
trsvcid: 4420
subnqn:  GB00041004bbf910
traddr:  10.254.164.2
```

Step-2 (Connect to I/O Subsystem)

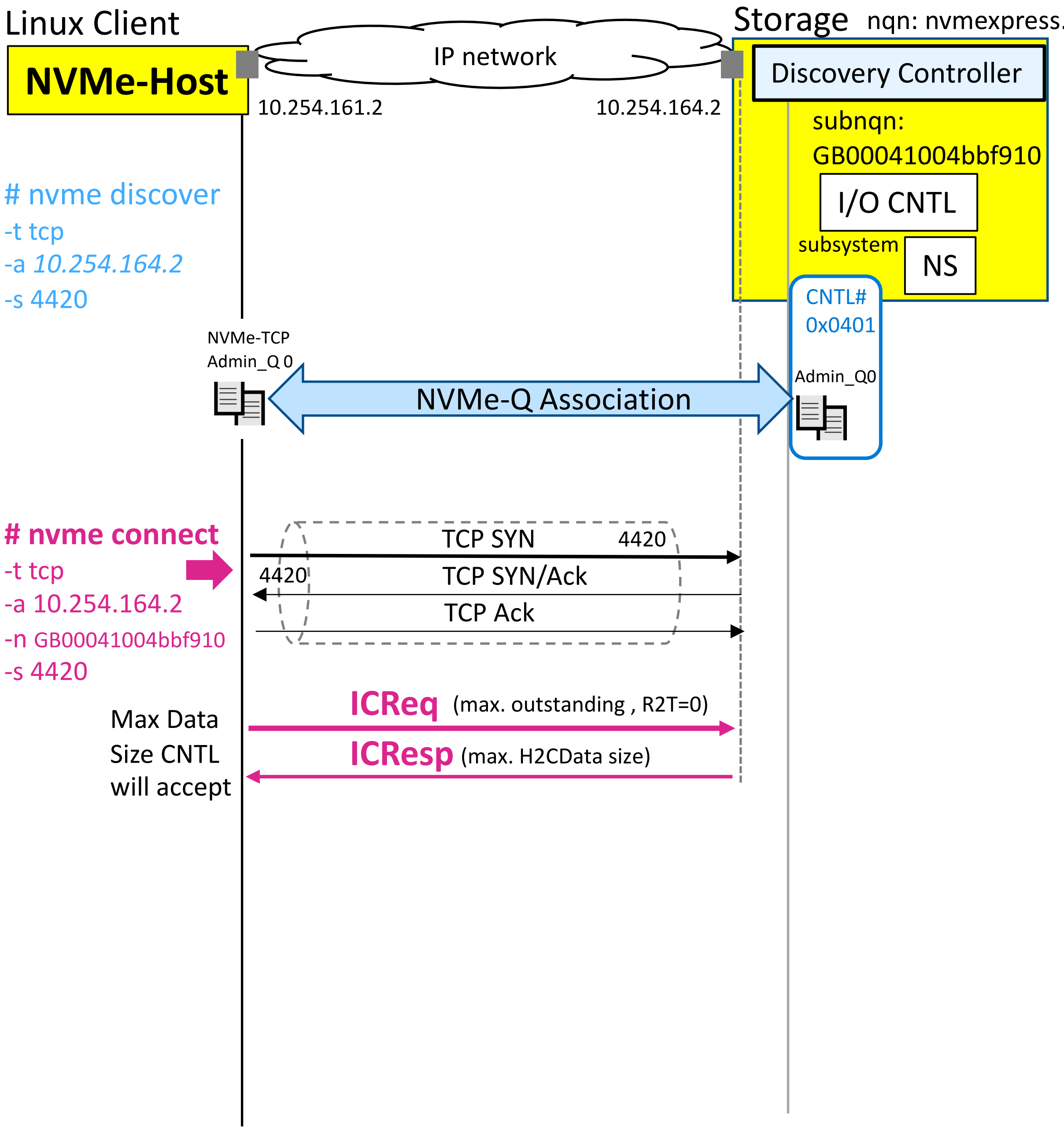
Storage Admin will issue a “NVMe connect” CLI command at the host to connect to the Storage Appliance Subsystem.

 # nvme connect -t tcp -a 10.254.164.2 -n GB00041004bbf910 -s 4420

```
# nvme list
Node             SN                   Model                                    Namespace Usage                      Format           FW Rev
---------------- -------------------- ---------------------------------------- --------- -------------------------- ---------------- --------
/dev/ nvme0n1     SDM00000EC75         UCSC-F-H16003                            1           1.60  TB /   1.60  TB    512   B +  0 B   KNCCP100
/dev/ nvme1n1     GB00041004bbf910     PVL-MX18S0P2L2C1-F100TP0TY1              1           2.15  TB /   2.15  TB      4 KiB +  0 B   22139242
```


Kamal Bakshi, Cisco

Initiate Connection to I/O Subsys



ICReq

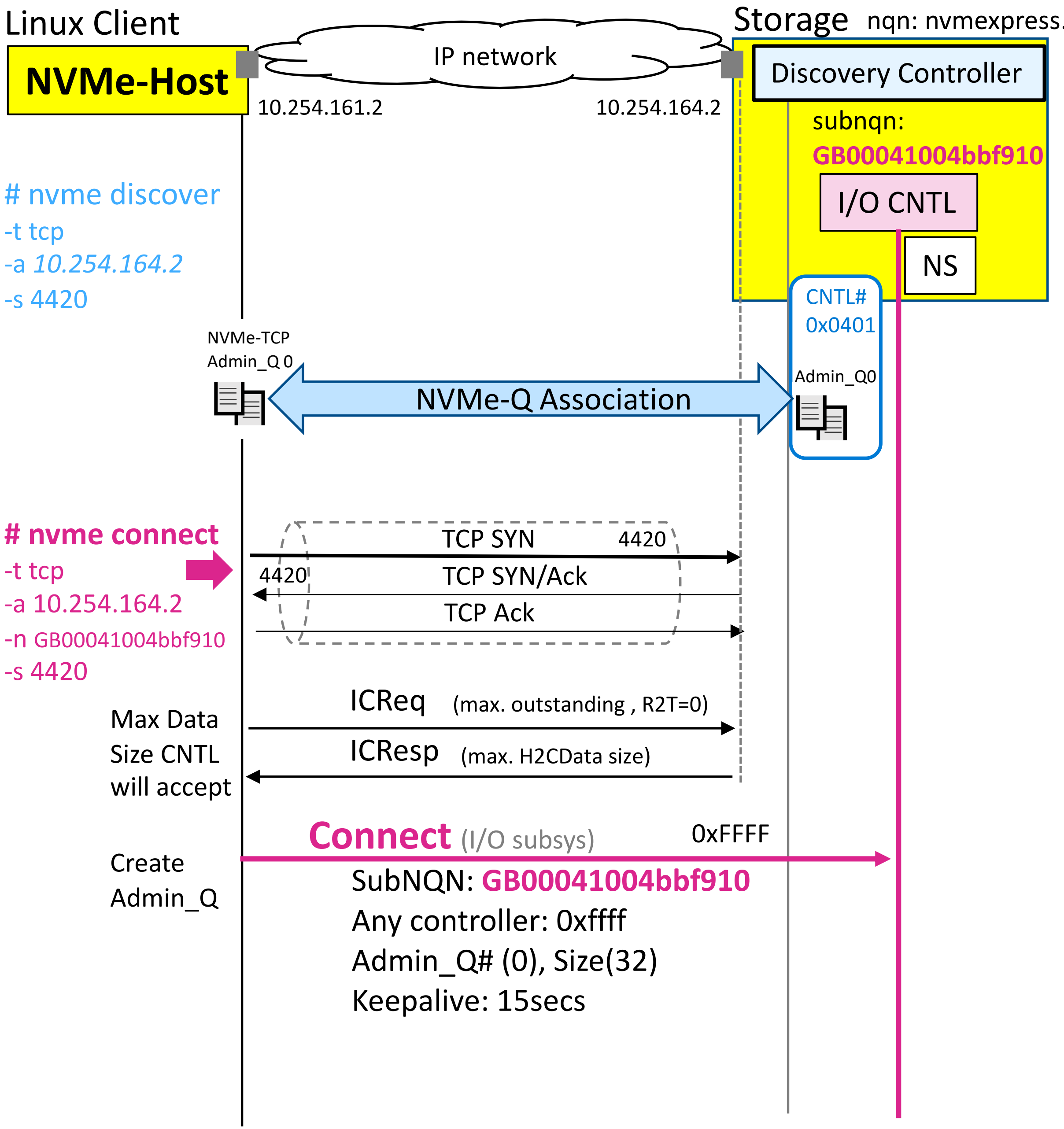
```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq: 1,
v NVM Express Fabrics TCP
    [Cmd Qid: 0 (AQ)]
    Pdu Type: ICReq (0)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 128
    Pdu Data Offset: 0
    Packet Length: 128
  v ICReq
      Pdu Version Format: 0
      Host Pdu data alignment: 0
      Digest Types Enabled: 0
      Maximum r2ts per request: 0
```

ICResp

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43946, Seq: 1,
v NVM Express Fabrics TCP
    [Cmd Qid: 0 (AQ)]
    Pdu Type: ICResp (1)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 128
    Pdu Data Offset: 0
    Packet Length: 128
  v ICResp
      Pdu Version Format: 0
      Controller Pdu data alignment: 0
      Digest types enabled: 0
      Maximum data capsules per r2t supported: 65535
```


Kamal Bakshi, Cisco

Create NVMe Admin Q

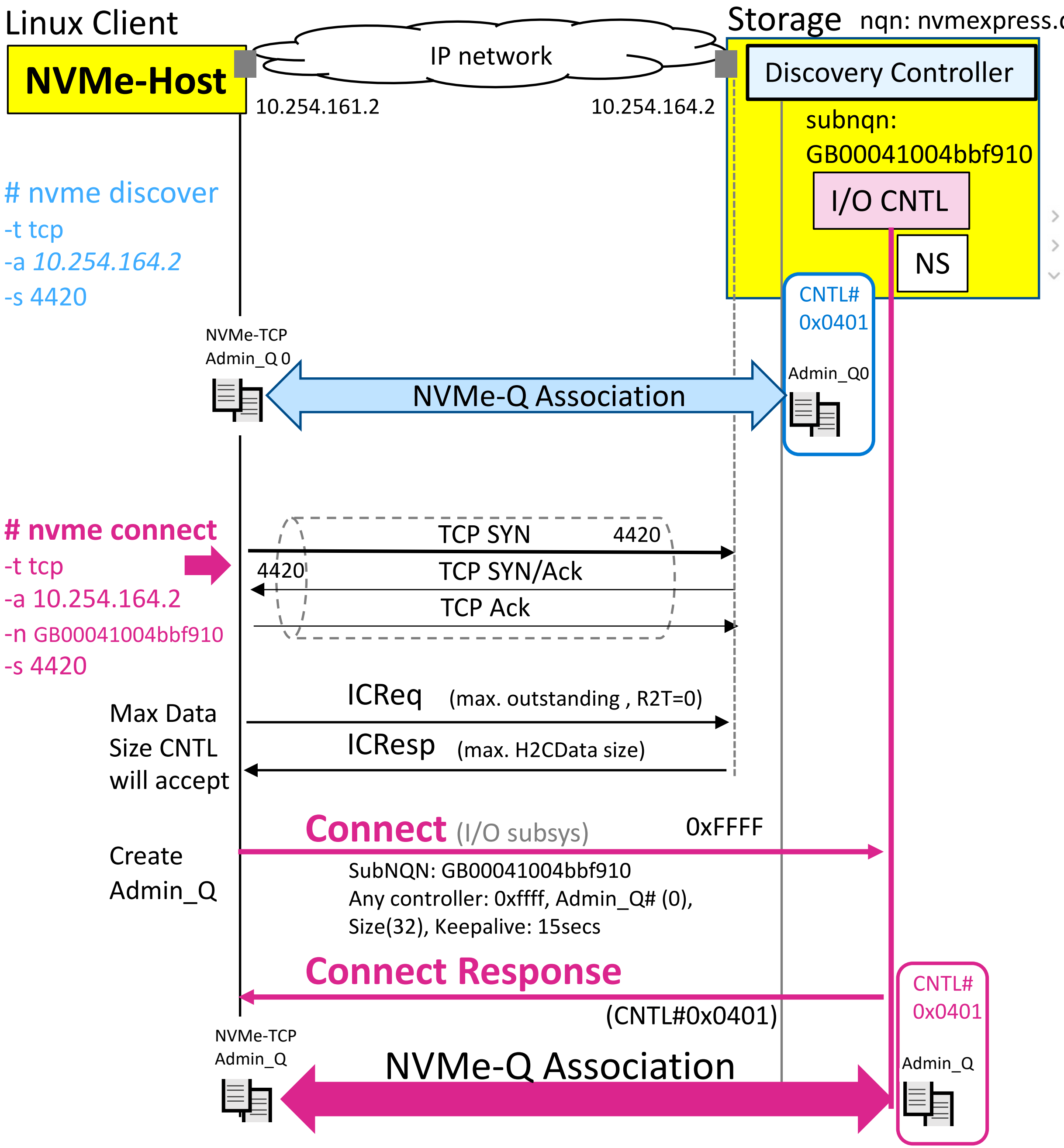


Connect Request

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq: 129, Ack: 129, Len:
v NVM Express Fabrics TCP, Fabrics Type: Connect (0x01) Cmd ID: 0x0000
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 72
    Packet Length: 1096
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [Fabric Cqe in: 10]
      Reserved: 0x40
      Command Identifier: 0x0000
      Fabric Command Type: Connect (0x01)
      Reserved: 00000000000000000000000000000000000
    > SGL1
      Record Format: 0
      Queue ID: 0 (AQ)
      Submission Queue Size: 32
    v Connect Attributes: 0x00
        .... ..00 = Priority Class: Urgent (0x0)
        .... .0.. = Disable SQ Flow Control: 0x0
        .... 0... = Support Deletion of IO Queues: 0x0
        0000 .... = Reserved: 0x0
      Reserved: 00
      Keep Alive Timeout: 15000ms
      Reserved: 000000000000000000000000
  v Data
      Host Identifier: 24756c3c1bd24e9d987da87397e55aa3
      Controller ID: 0xffff
      Reserved: 0000000000000000000000000000000000000000000000000000000000000000…
      Subsystem NQN: GB00041004bbf910
      Host NQN: nqn.2014-08.org.nvmexpress:uuid:55e23c7b-7431-439b-81d7-430e286fbb98
      Reserved: 0000000000000000000000000000000000000000000000000000000000000000…
```


Kamal Bakshi, Cisco

Create NVMe Admin Q



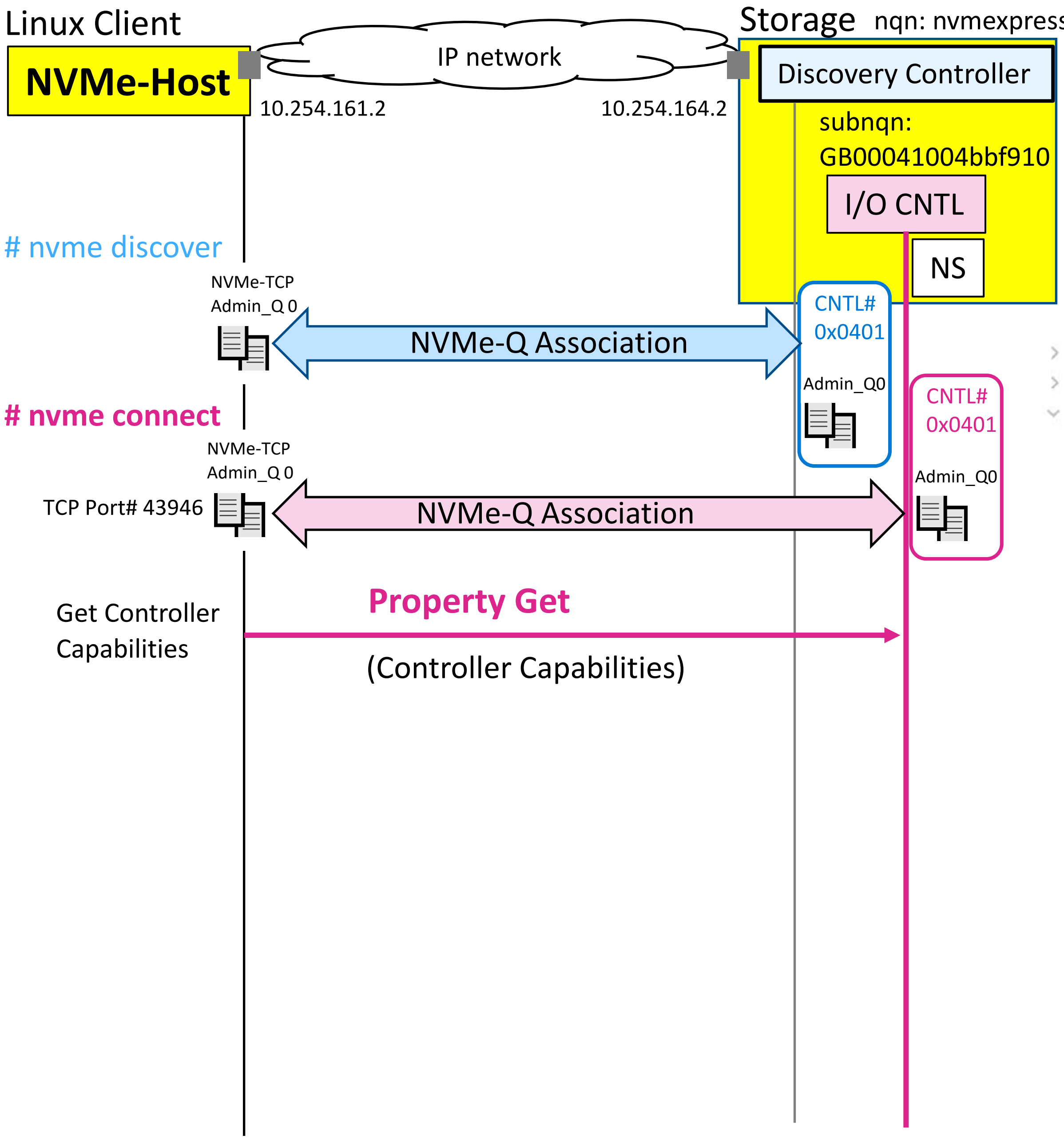
Connect Response

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43946, Seq: 129, Ack: 1225,
v NVM Express Fabrics TCP, Cqe Fabrics Cmd: Connect (0x01) Cmd ID: 0x0000
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
  v Cqe (For Cmd: Connect)
      [Fabric Cmd in: 9]
      [Cmd Latency: 22.519 ms]
      Controller ID: 0x0401
      Authentication Required: 0x0000
      Reserved: 00000000
      SQ Head Pointer: 0x0000
      SQ Identifier: 0x0000
      Command Identifier: 0x0000
    v Status Field: 0x0000
        .... .... .... ...0 = Reserved: 0x0
        .... ...0 0000 000. = Status Code: 0x00 (Successful Completion)
        .... 000. .... .... = Status Code Type: Generic Command Status (0x0)
        ..00 .... .... .... = Command Retry Delay: 0x0
        .0.. .... .... .... = More Information in Log Page: False
        0... .... .... .... = Do not Retry: False
```


Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

Get I/O Cntl. Capabilities

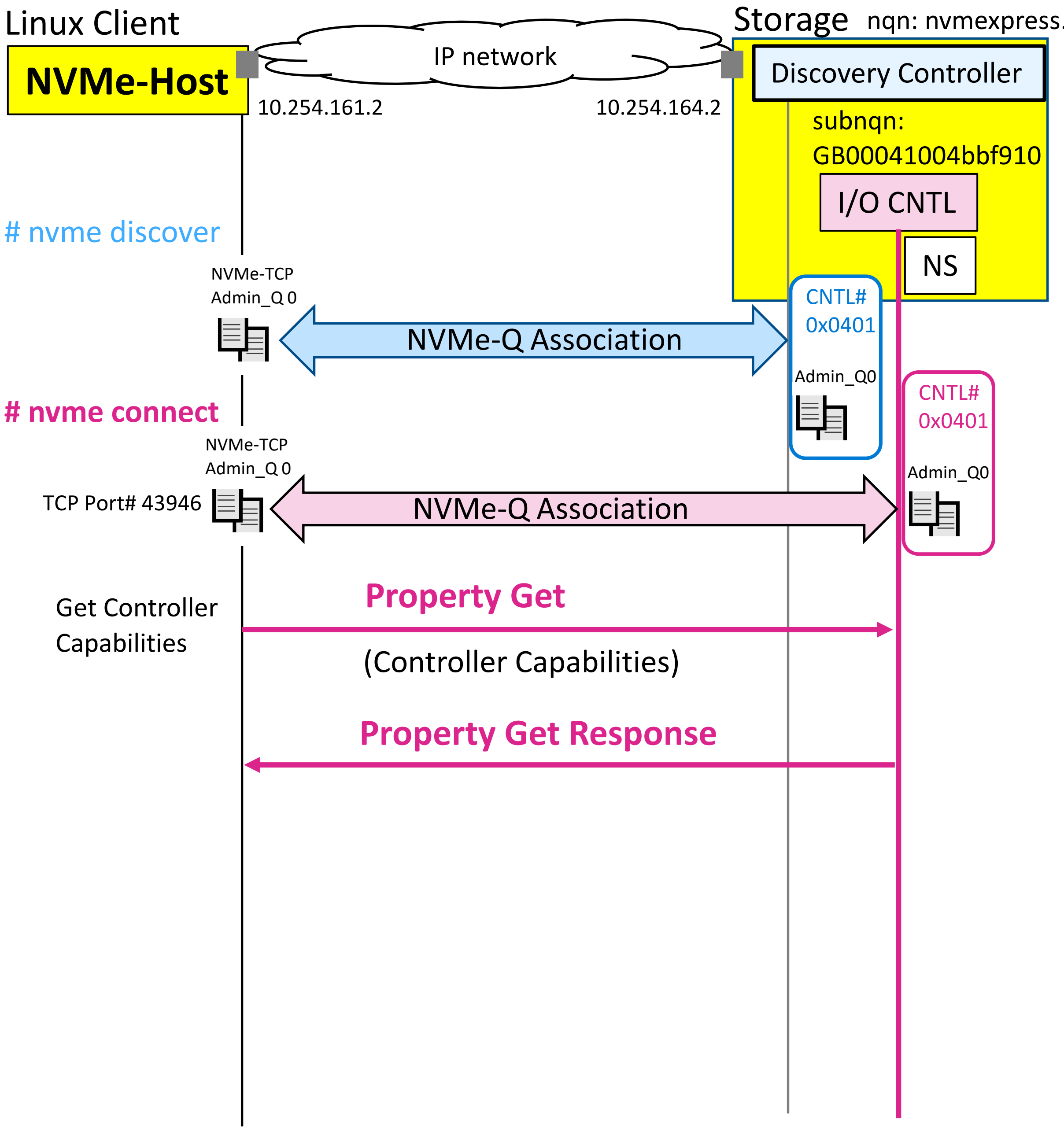


Property Get

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq: 1225, Ack: 153,
v NVM Express Fabrics TCP, Fabrics Type: Property Get (0x04) Cmd ID: 0x000e
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [Fabric Cqe in: 13]
      Reserved: 0x40
      Command Identifier: 0x000e
      Fabric Command Type: Property Get (0x04)
      Reserved: 000000000000000000000000000000000000000000000000000000000000000000000005a
    v .... .001 = Attributes: 0x1
        .... .001 = Property Size: 8 bytes (0x1)
        0000 0... = Reserved: 0x00
      Reserved: 000000
      Offset: Controller Capabilities (0x00000000)
      Reserved: 00000000000000000000000000000000
```


Kamal Bakshi, Cisco

Get I/O Cntl. Capabilities



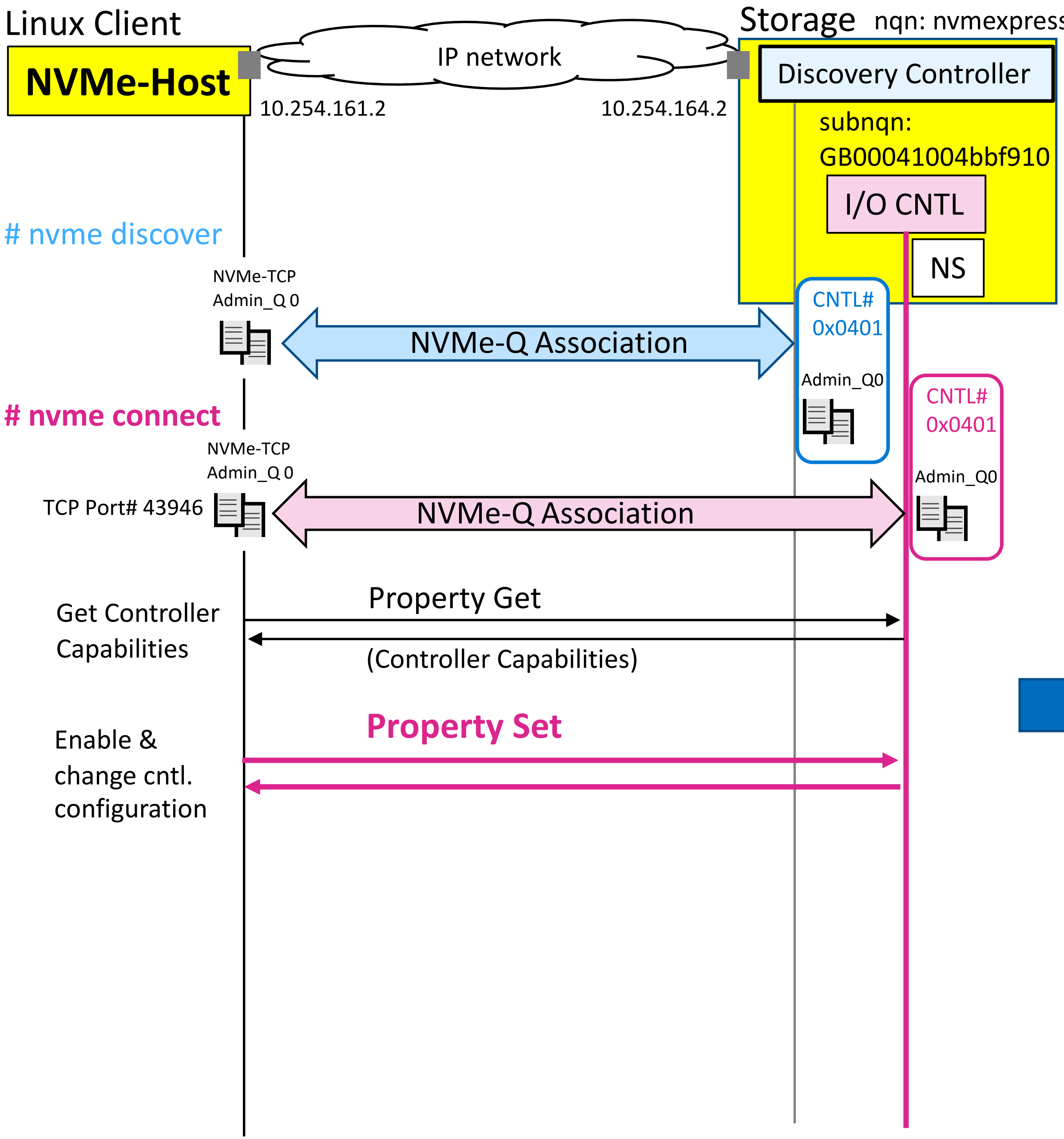
Property Get Response

```
Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
Transmission Control Protocol, Src Port: 4420, Dst Port: 43946, Seq: 153, Ack: 1297, Len: 24
NVM Express Fabrics TCP, Cqe Fabrics Cmd: Property Get (0x04) Cmd ID: 0x000e
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
    Pdu Specific Flags: 0x00
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
    Cqe (For Cmd: Property Get)
        [Fabric Cmd in: 12]
        [Cmd Latency: 0.018 ms]
        Property Data: 7f00000f20000000
            Controller Capabilities: 0x000000200f00007f
                = Maximum Queue Entries Supported: 128
                = Contiguous Queues Required: False
                = Supports Arbitration Mechanism with Weighted Round Robin
                = Supports Arbitration Mechanism Vendor Specific: False
                = Reserved: 0x00
                = Timeout (to ready status): f (7500 ms)
                = Doorbell Stride: 0x0 (4 bytes)
                = NVM Subsystem Reset Supported: False
                = Command Sets Supported: 1 (NVM IO Command Set)
                = Boot Partition Support: False
                = Reserved: 0x0
                = Memory Page Size Minimum: 0x0 (4096 bytes)
                = Memory Page Size Maximum: 0x0 (4096 bytes)
                = Persistent Memory Region Supported: False
                = Controller Memory Buffer Supported: False
                = Reserved: 0x00
        SQ Head Pointer: 0x0000
        SQ Identifier: 0x0000
        Command Identifier: 0x000e
        Status Field: 0x0000
            .... .... .... ...0 = Reserved: 0x0
            .... ...0 0000 000. = Status Code: 0x00 (Successful Completion)
            .... 000. .... .... = Status Code Type: Generic Command Status (0x0)
            ..00 .... .... .... = Command Retry Delay: 0x0
            .0.. .... .... .... = More Information in Log Page: False
            0... .... .... .... = Do not Retry: False
```

Controller Capability Data

Kamal Bakshi, Cisco

Set I/O Cntl. Configuration



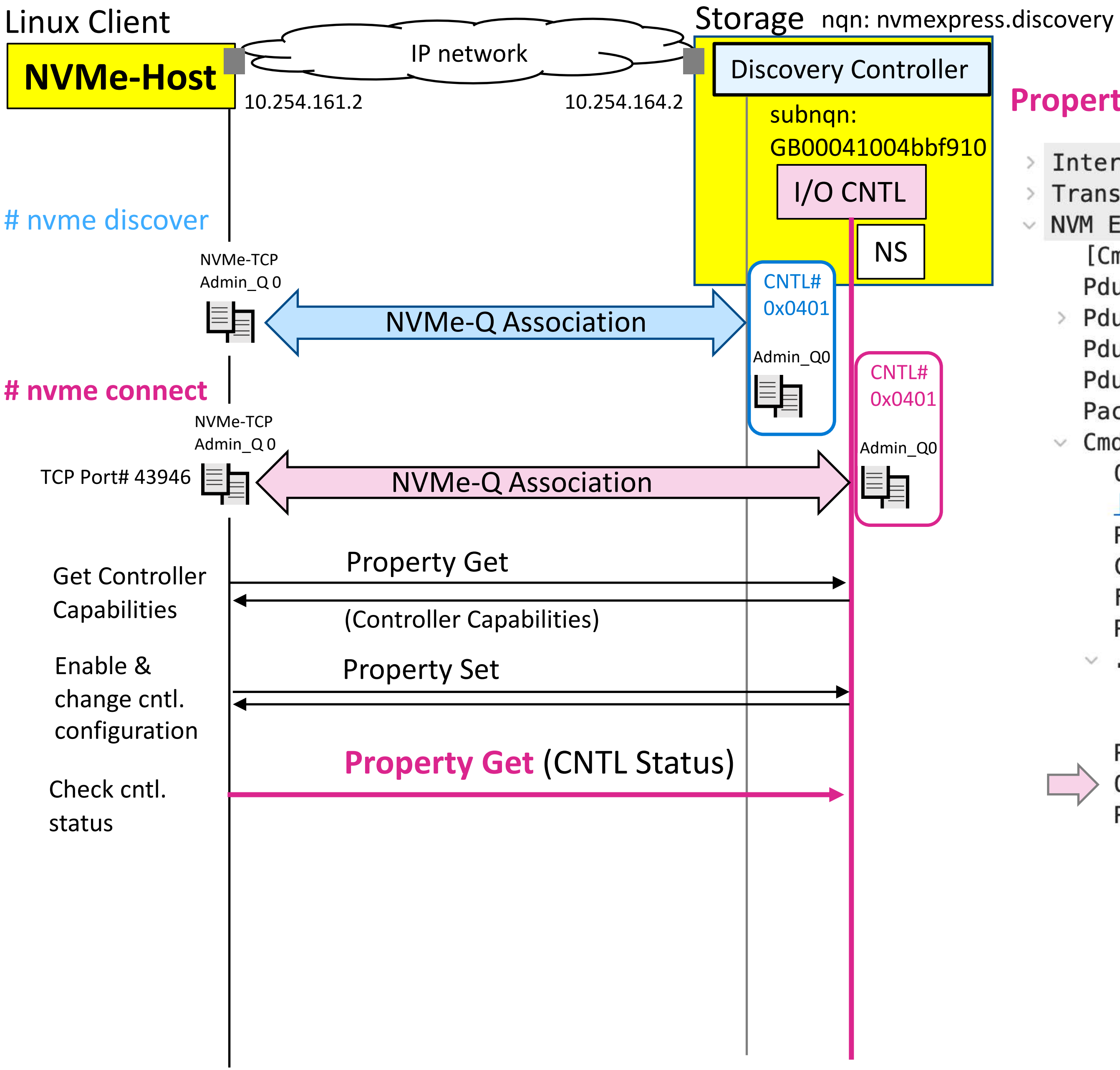
Property Set

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq: 1297,
v NVM Express Fabrics TCP, Fabrics Type: Property Set (0x00) Cmd ID: 0x000f
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [Fabric Cqe in: 16]
      Reserved: 0x40
      Command Identifier: 0x000f
      Fabric Command Type: Property Set (0x00)
      Reserved: 00000000000000000000000000000000000000000000000000000000000
    v .... .000 = Attributes: 0x0
        .... .000 = Property Size: 4 bytes (0x0)
        0000 0... = Reserved: 0x00
      Reserved: 000000
      Offset: Controller Configuration (0x00000014)
    v Property Data: 0100460000000000
      v Controller Configuration: 0x00460001
            = Enable: 0x1
            = Reserved: 0x0
            = IO Command Set Selected: NVM IO Command Set (0x0)
            = Memory Page Size: 0x0 (4096 bytes)
            = Arbitration Mechanism Selected: Round Robin (0x0)
            = Shutdown Notification: No Shutdown (0x0)
            = IO Submission Queue Entry Size: 0x6 (64 bytes)
            = IO Completion Queue Entry Size: 0x4 (16 bytes)
            = Reserved: 0x00
        Reserved: 0x00000000
      Reserved: 0000000000000000
```

 |

Kamal Bakshi, Cisco

Get I/O Cntl. Status

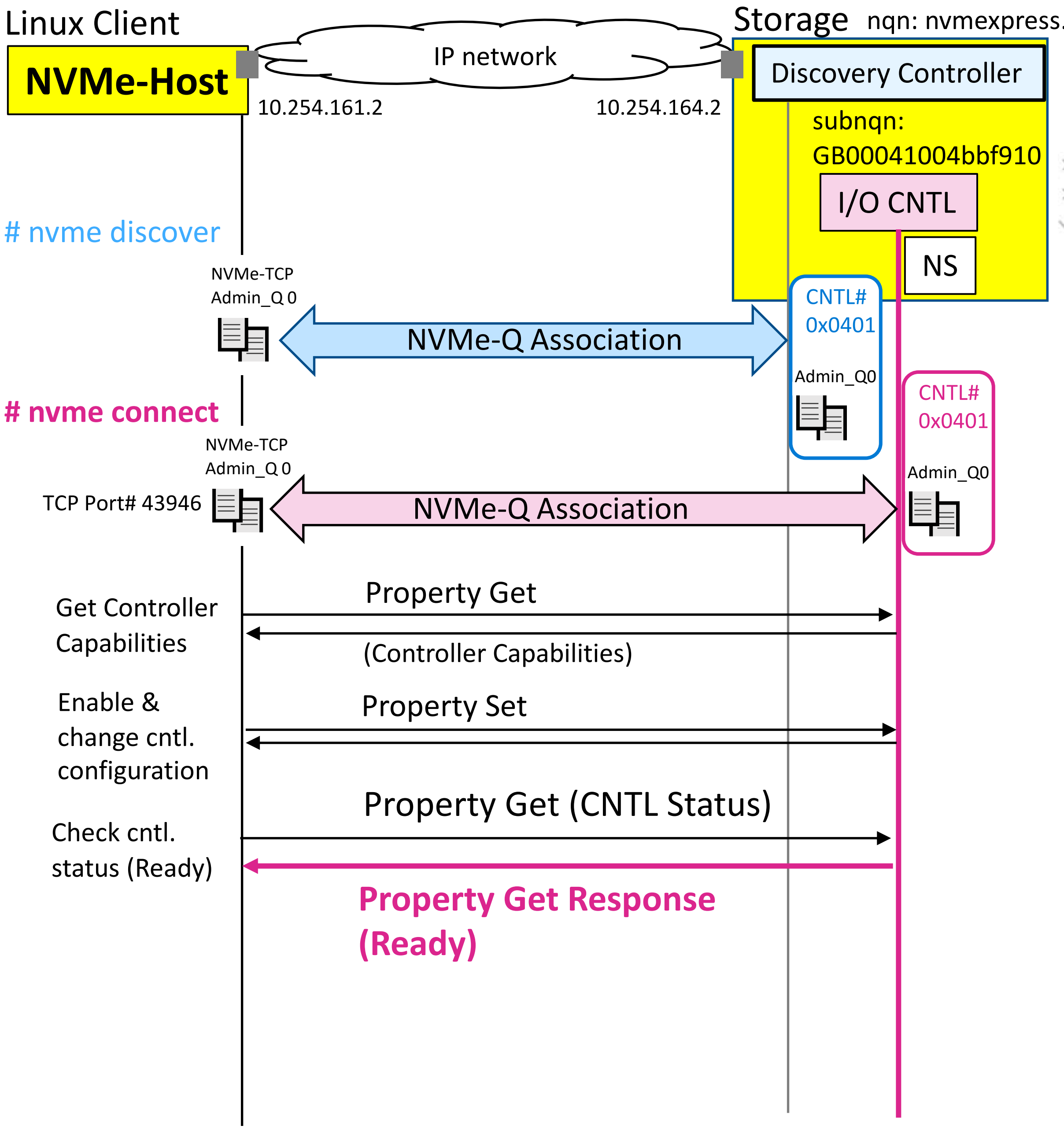


Property Get (Controller Status)

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq: 1369,
v NVM Express Fabrics TCP, Fabrics Type: Property Get (0x04) Cmd ID: 0x0010
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [Fabric Cqe in: 19]
      Reserved: 0x40
      Command Identifier: 0x0010
      Fabric Command Type: Property Get (0x04)
      Reserved: 0000000000000000000000000000000000000000000000000000000
    v .... .000 = Attributes: 0x0
        .... .000 = Property Size: 4 bytes (0x0)
        0000 0... = Reserved: 0x00
      Reserved: 000000
      Offset: Controller Status (0x0000001c)
      Reserved: 00000000000000000000000000000000
```


Kamal Bakshi, Cisco

Get I/O Cntl. Status

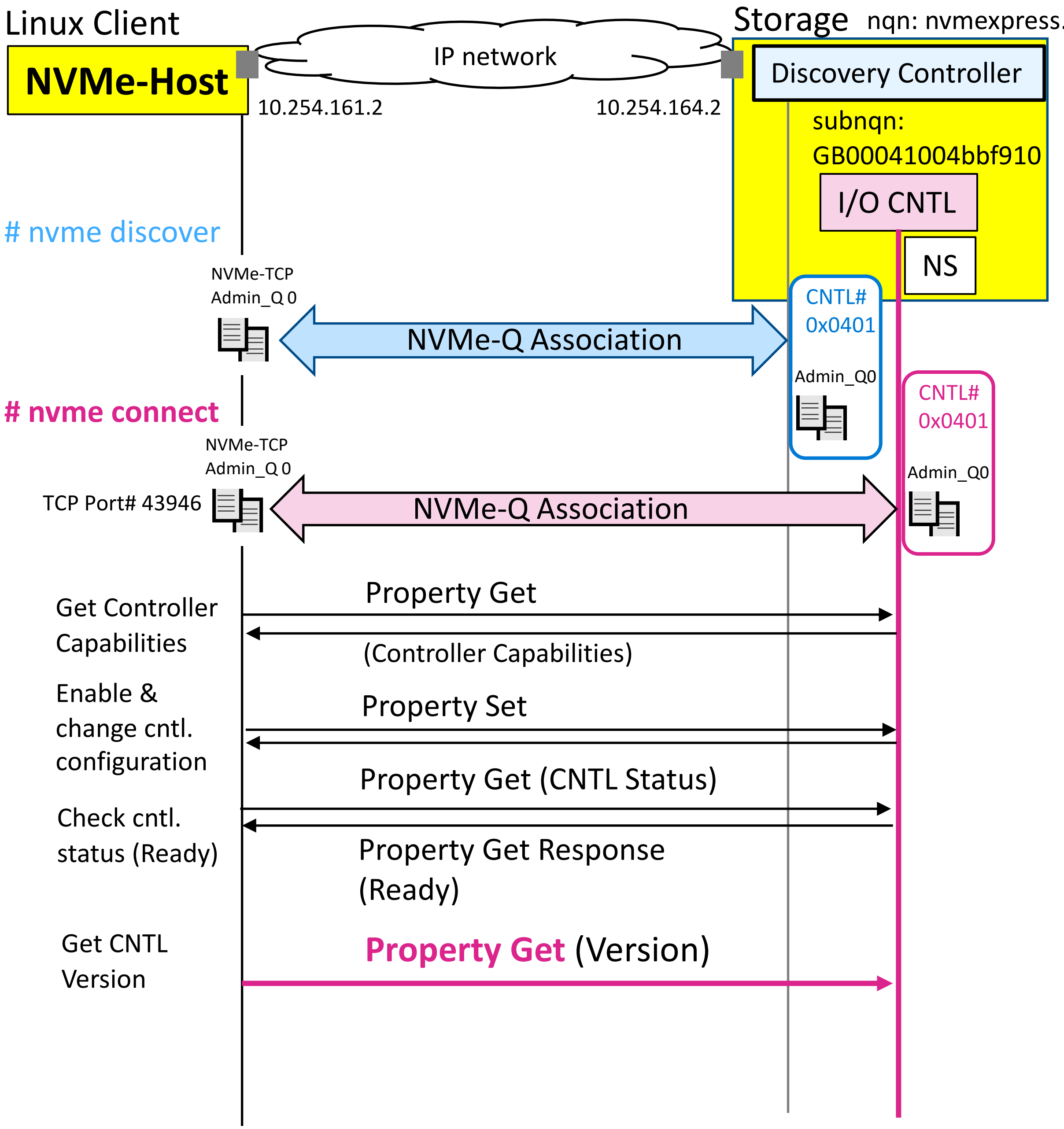


Property Get Response

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43946, Seq: 201, Ack: 1441,
v NVM Express Fabrics TCP, Cqe Fabrics Cmd: Property Get (0x04) Cmd ID: 0x0010
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
  v Cqe (For Cmd: Property Get)
      [Fabric Cmd in: 18]
      [Cmd Latency: 0.016 ms]
    v Property Data: 0100000000000000
      v Controller Status: 0x00000001
          .... .... .... .... .... .... .... ...1 = Ready: 0x1
          .... .... .... .... .... .... .... ..0. = Controller Fatal Status: 0x0
          .... .... .... .... .... .... .... 00.. = Shutdown Status: No Shutdown (0x0)
          .... .... .... .... .... .... ...0 .... = NVM Subsystem Reset Occurred: 0x0
          .... .... .... .... .... .... ..0. .... = Processing Paused: 0x0
          0000 0000 0000 0000 0000 0000 00.. .... = Reserved: 0x0000000
        Reserved: 0x00000000
      SQ Head Pointer: 0x0000
      SQ Identifier: 0x0000
      Command Identifier: 0x0010
    v Status Field: 0x0000
        .... .... .... ...0 = Reserved: 0x0
        .... ...0 0000 000. = Status Code: 0x00 (Successful Completion)
        .... 000. .... .... = Status Code Type: Generic Command Status (0x0)
        ..00 .... .... .... = Command Retry Delay: 0x0
        .0.. .... .... .... = More Information in Log Page: False
        0... .... .... .... = Do not Retry: False
```


Kamal Bakshi, Cisco

Get I/O Cntl. Version

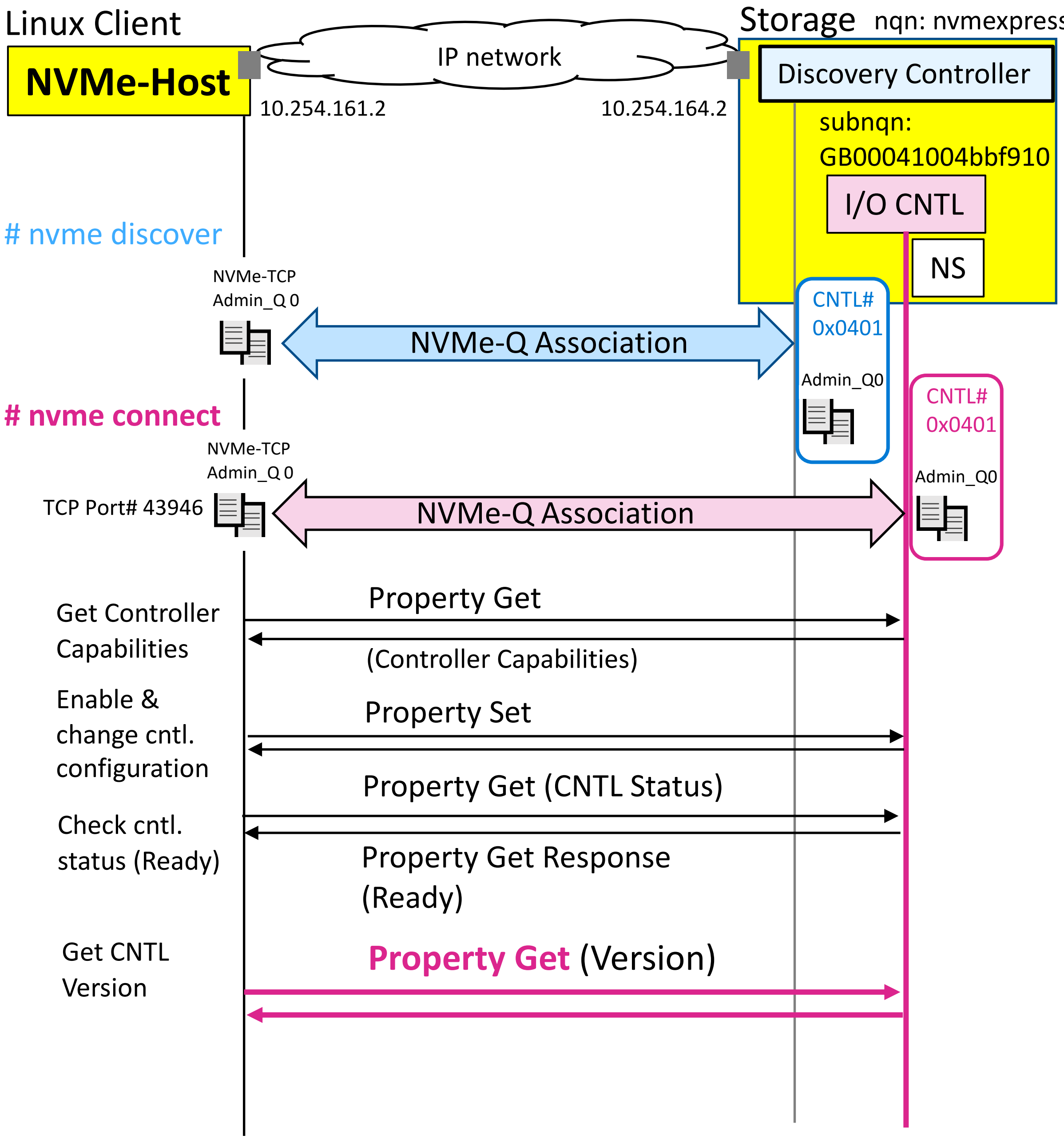


Property Get (Version)

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq
v NVM Express Fabrics TCP, Fabrics Type: Property Get (0x04) Cmd ID:
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [Fabric Cqe in: 22]
      Reserved: 0x40
      Command Identifier: 0x0011
      Fabric Command Type: Property Get (0x04)
      Reserved: 0000000000000000000000000000000000000000000000000
    v .... .000 = Attributes: 0x0
        .... .000 = Property Size: 4 bytes (0x0)
        0000 0... = Reserved: 0x00
      Reserved: 000000
      Offset: Version (0x00000008)
      Reserved: 00000000000000000000000000000000
```


Kamal Bakshi, Cisco

Get I/O Cntl. Version

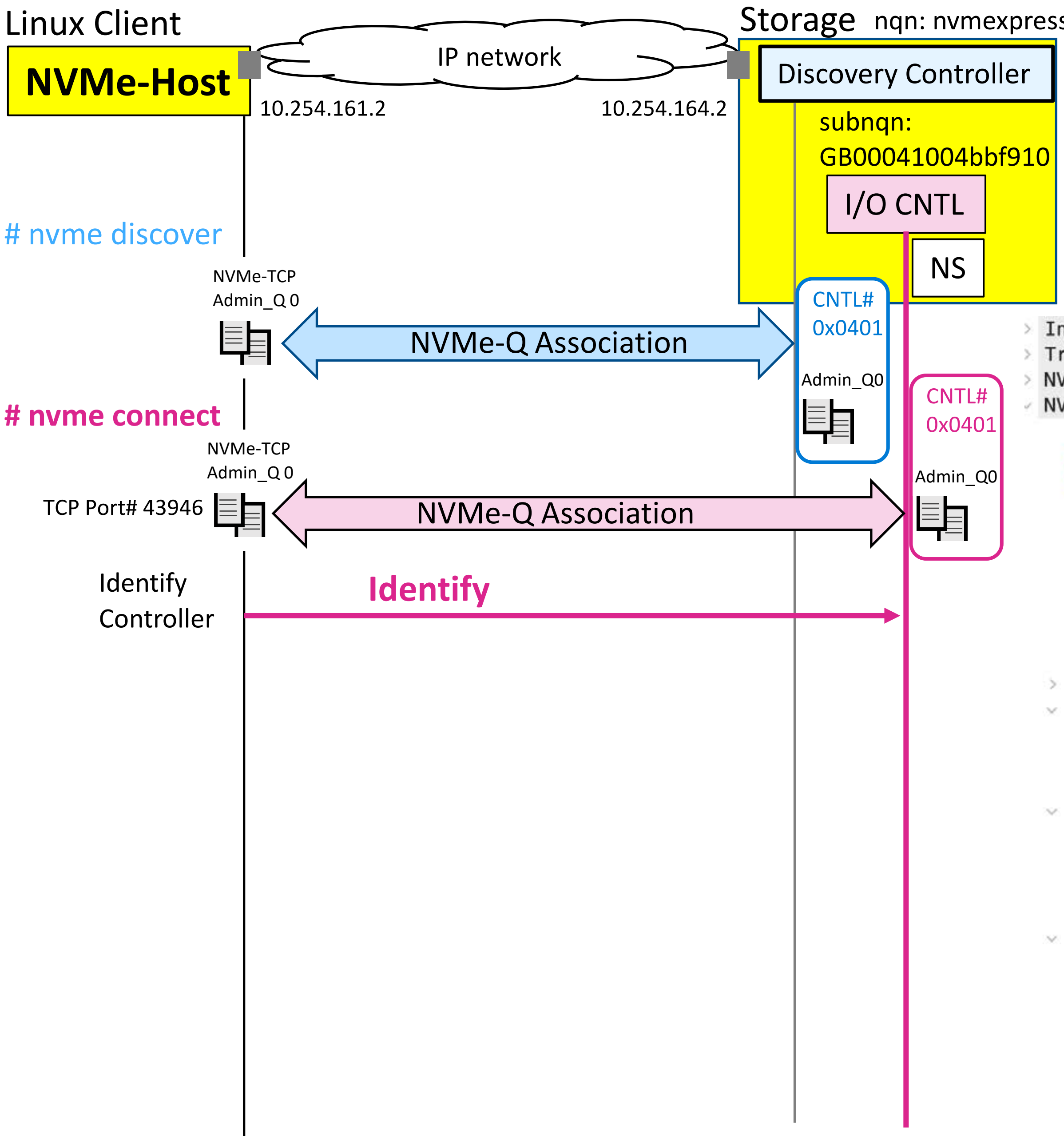


Property Get Response (Version#)

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43946, Seq: 225, Ack:
v NVM Express Fabrics TCP, Cqe Fabrics Cmd: Property Get (0x04) Cmd ID: 0x0011
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
  v Cqe (For Cmd: Property Get)
      [Fabric Cmd in: 21]
      [Cmd Latency: 0.017 ms]
    v Property Data: 0002010000000000
      v Version: 0x00010200
          .... .... .... .... .... .... 0000 0000 = Tertiary Version: 0
          .... .... .... .... 0000 0010 .... .... = Minor Version: 2
          0000 0000 0000 0001 .... .... .... .... = Major Version: 1
        Reserved: 0x00000000
      SQ Head Pointer: 0x0000
      SQ Identifier: 0x0000
      Command Identifier: 0x0011
    v Status Field: 0x0000
        .... .... .... ...0 = Reserved: 0x0
        .... ...0 0000 000. = Status Code: 0x00 (Successful Completion)
        .... 000. .... .... = Status Code Type: Generic Command Status (0x0)
        ..00 .... .... .... = Command Retry Delay: 0x0
        .0.. .... .... .... = More Information in Log Page: False
        0... .... .... .... = Do not Retry: False
```


Kamal Bakshi, Cisco

Identify I/O Cntl. CNS-01

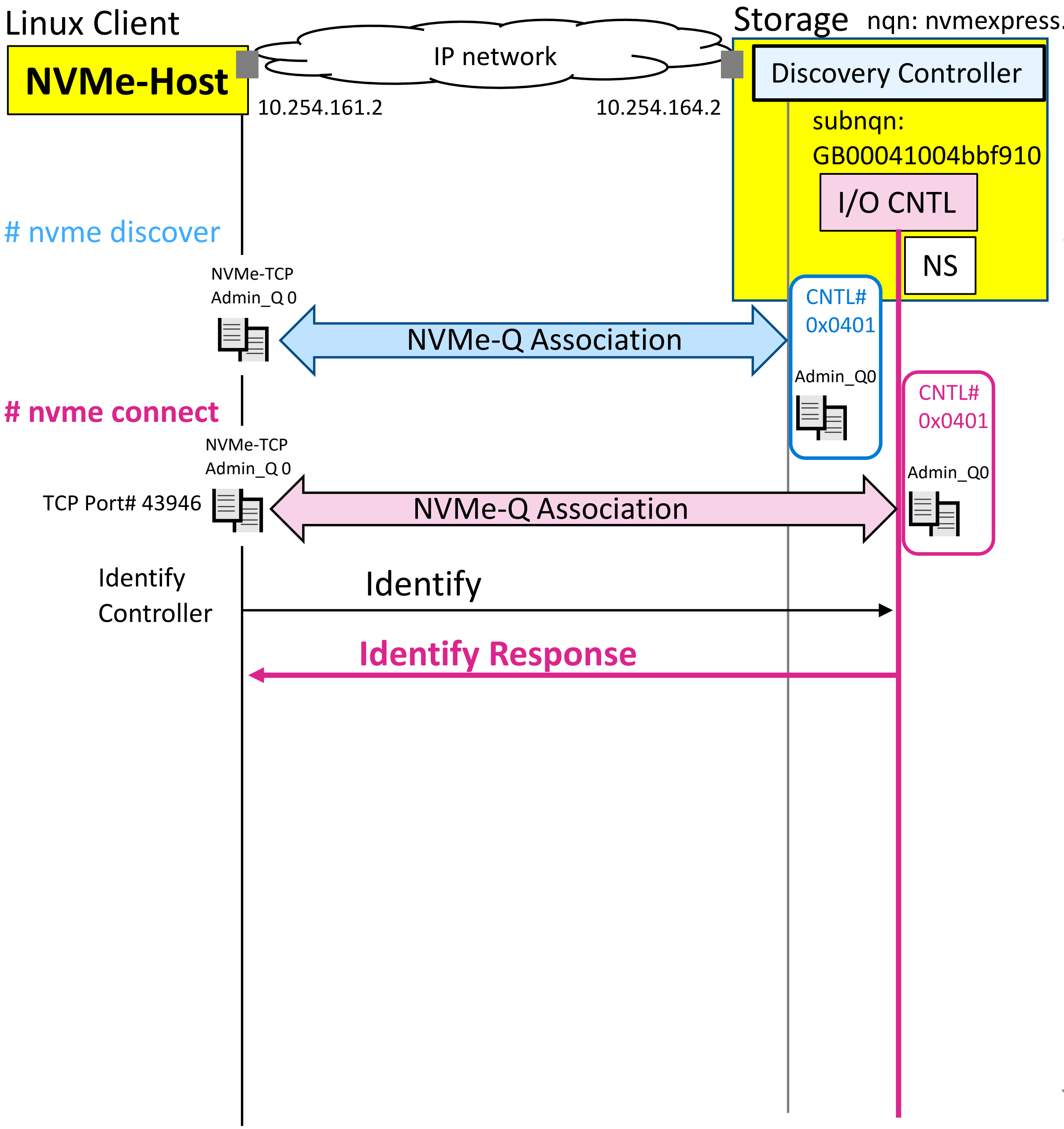


Identify Request

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq: 1585, Ack: 273, Len: 72
> NVM Express Fabrics TCP, NVMe Opcode: Identify (0x06) Cmd ID: 0x000f
v NVM Express (Cmd)
    Opcode: 0x06 Identify
    [DATA Transfer 0: 28]
    [Cqe in: 28]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x000f
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  > SGL1
  v DWORD10: 0x00000001
      .... .... .... .... .... .... 0000 0001 = Controller or Namespace Structure (CNS): Controller (0x01)
      .... .... .... .... 0000 0000 .... .... = Reserved: 0x00
      0000 0000 0000 0000 .... .... .... .... = Controller Identifier (CNTID): 0x0000
  v DWORD11: 0x00000000
      .... .... .... .... 0000 0000 0000 0000 = NVM Set Identifier (NVMSETID): 0x0000
      0000 0000 0000 0000 .... .... .... .... = Reserved: 0x0000
    DWORD12: 0x00000000
    DWORD13: 0x00000000
  v DWORD14: 0x00000000
      .... .... .... .... .... .... .000 0000 = UUID Index: 0x00
      0000 0000 0000 0000 0000 0000 0... .... = UUID Index: 0x0000000
    DWORD15: 0x00000000
```


Kamal Bakshi, Cisco

I/O Controller Details



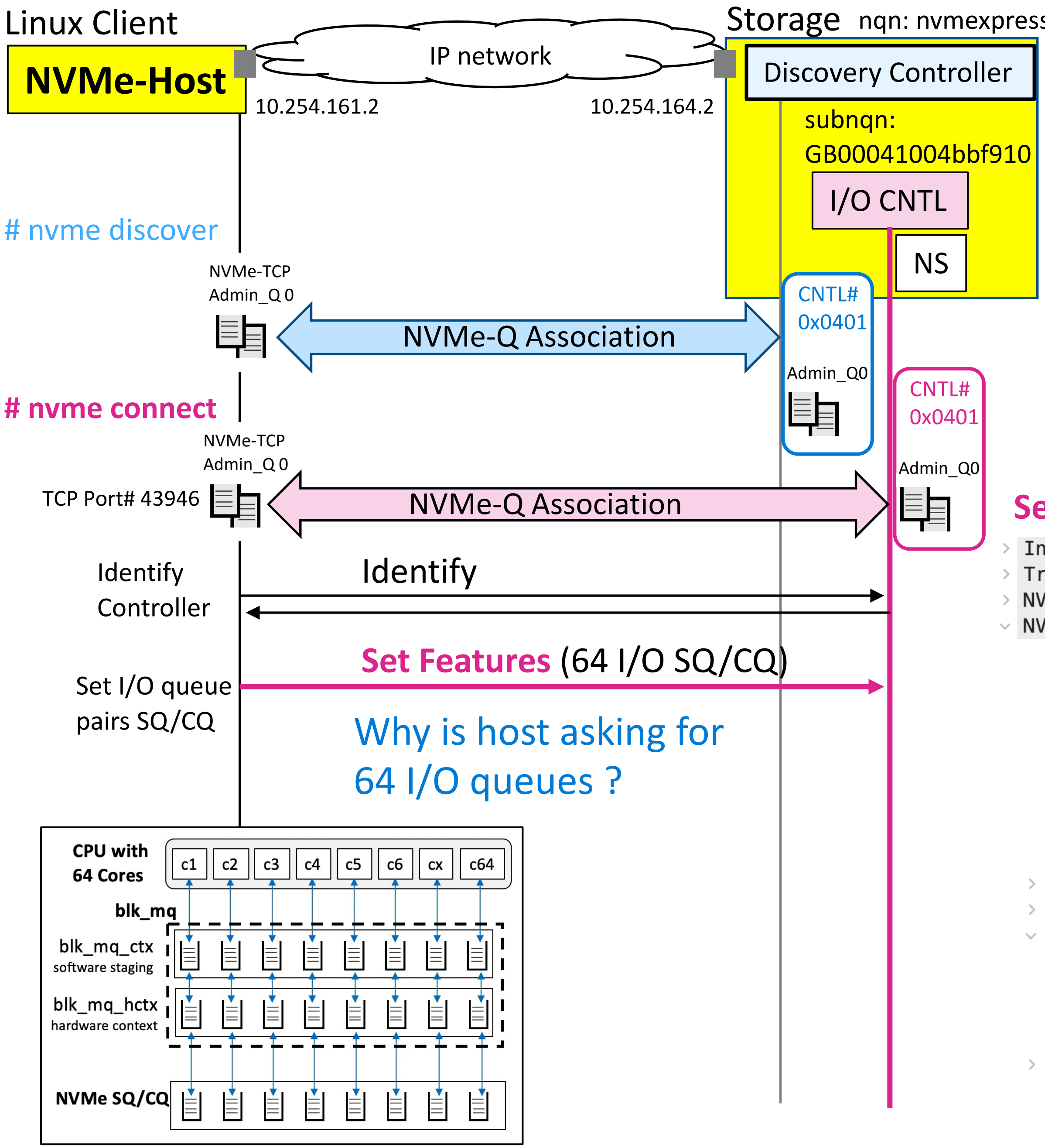
Identify Response

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43946, Seq: 273, Ack: 1657, Len: 4144
> NVM Express Fabrics TCP, C2HData Opcode: Identify (0x06), Cmd ID: 0x000f, Len: 4096
v NVM Express
    PCI Vendor ID (VID): 0x0000
    PCI Subsystem Vendor ID (SSVID): 0x0000
    Serial Number (SN): GB00041004bbf910
    Model Number (MN): PVL-MX18S0P2L2C1-F100TP0TY1
    Firmware Revision (FR): 3232313339323432
    Recommended Arbitration Burst (RAB): 0x6 (64 commands)
    IEEE OUI Identifier (IEEE): 0x00000000
  > Controller Multi-Path I/O and Namespace Sharing Capabilities (CMIC): 0x0b
    Maximum Data Transfer Size (MDTS): 0x4 (16 pages)
    Controller ID (CNTLID): 0x0401
  > Version (VER): 0x00010200
    RTD3 Resume Latency (RTD3R): 0 (not reported)
    RTD3 Entry Latency (RTD3E): 0 (not reported)
  > Optional Asynchronous Events Supported (OAES): 0x00000900
  > Controller Attributes (CTRATT): 0x00000001
  > Read Recovery Levels Support (RRLS): 0x0000
    Reserved: 0000000000000000
    Controller Type (CNTRLTYPE): Reserved (not reported) (0x00)
  > FRU Globally Unique Identifier (FGUID): 00000000000000000000000000000000
    Command Retry Delay Time 1: 0 (0 ms)
    Command Retry Delay Time 2: 0 (0 ms)
    Command Retry Delay Time 3: 0 (0 ms)
    Reserved: 0000000000000000000000000000000000000000000000000000000000000000…
  > NVMe Management Interface: 00000000000000000000000000000000
  > Optional Admin Command Support (OACS): 0x0000
    Abort Command Limit (ACL): 0x3: (4 commands)
    Asynchronous Event Request Limit (AERL): 0x3: (4 events)
  > Firmware Updates (FRMW): 0x03
  > Log Page Attributes (LPA): True
    Error Log Page Entries (ELPE): 0x7f: (128 entries)
    Number of Power States Supported (NPSS): 0x0: (1 state)
  > Admin Vendor Specific Command Configuration (AVSCC): 0x00
  > Autonomous Power State Transition Attributes (APSTA): 0x00
```

multiple pages (not shown)

Kamal Bakshi, Cisco

Create 64 I/O queues request



In linux, multi-queue block I/O layer uses two levels of queues to improve scalability. The software queues “blk_mq_ctx” removes the lock contention problem in multi-core setup and the hardware queues “blk_mq_hctx” maps to the device driver multiple dispatch queues like NVMe SQ/CQ.

By default NVMe driver will map each core to one SQ/CQ pair

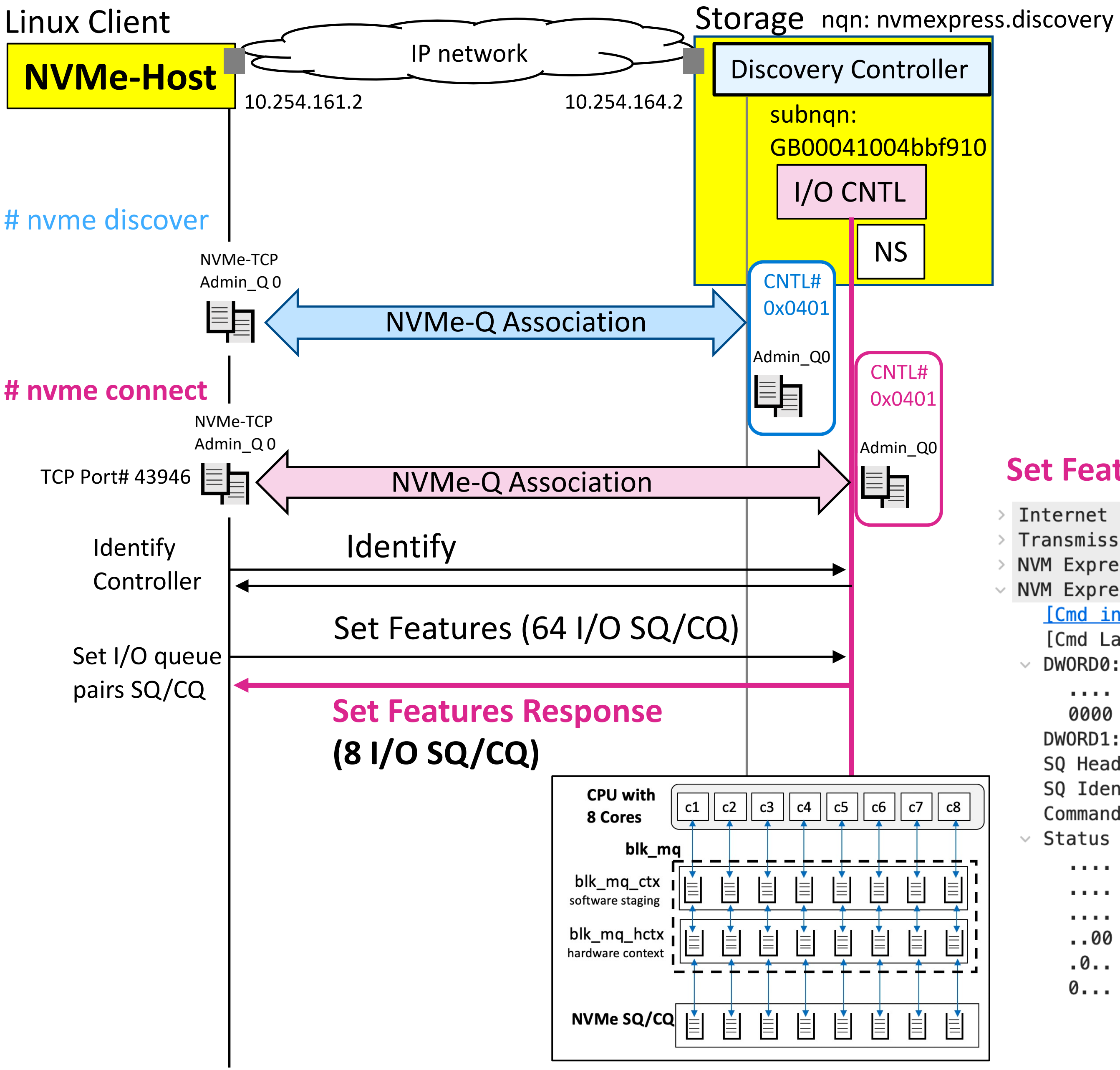
Set Feature (Host requests 64 I/O SQ/CQ)

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq: 1657, Ack: 4417, Len: 72
> NVM Express Fabrics TCP, NVMe Opcode: Set Features (0x09) Cmd ID: 0x0010
v NVM Express (Cmd)
    Opcode: 0x09 Set Features
    [Cqe in: 31]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x0010
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  > SGL1
  > DWORD 10: 0x00000007
  v DWORD11: 0x003f003f
      .... .... .... .... 0000 0000 0011 1111 = Number of IO Submission Queues Requested: 3f (64)
      0000 0000 0011 1111 .... .... .... .... = Number of IO Completion Queues Requested: 3f (64)
    DWORD12: 0x00000000
    DWORD13: 0x00000000
  > DWORD 14: 0x00000000
    DWORD15: 0x00000000
```

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

Accepted 8 I/O queues only

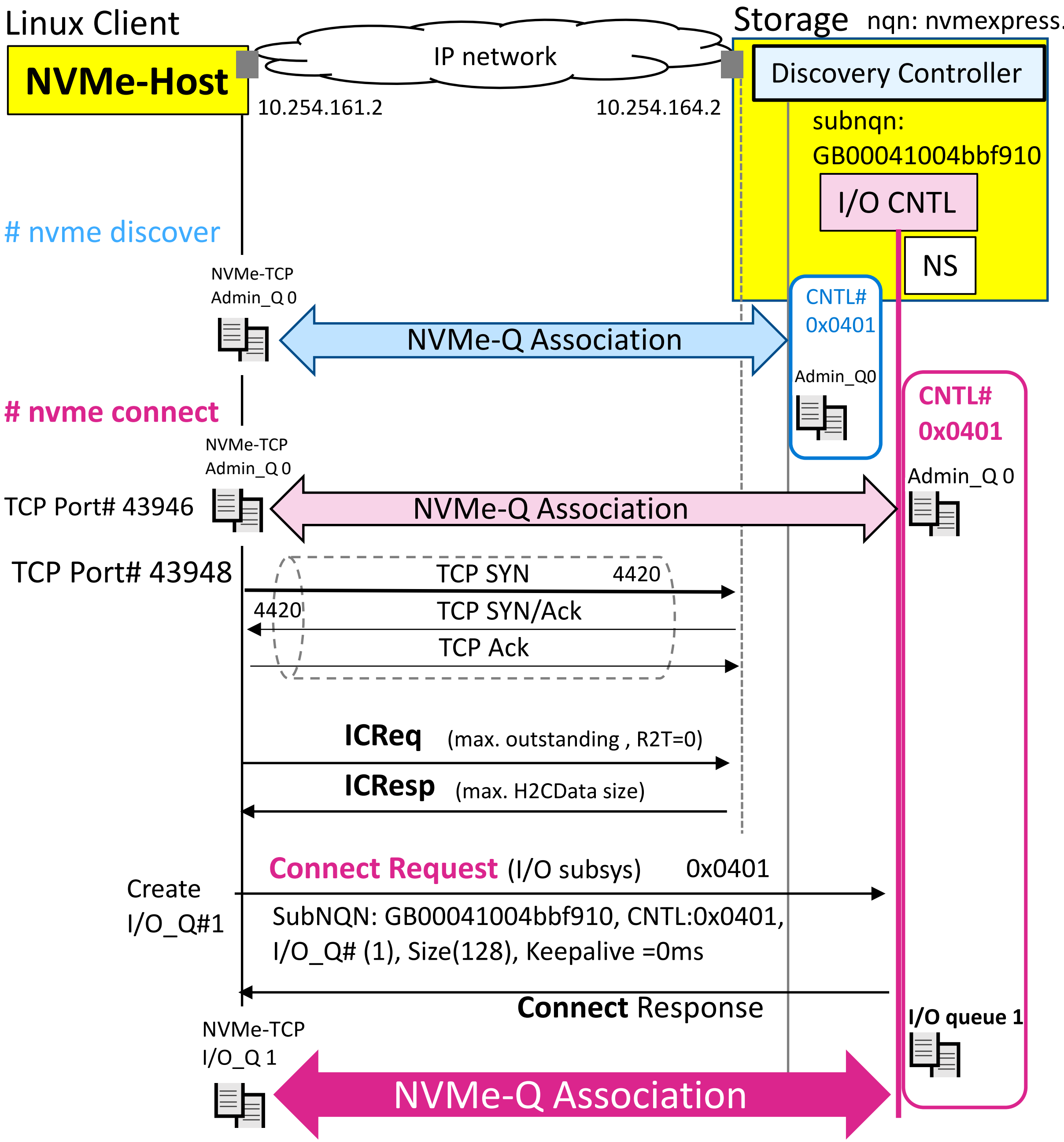


Based on the number of I/O queues response, Host will initiate the same number of NVMe association creation process (see next page)

Set Feature Response (Controller accepts only 8 I/O SQ/CQ pairs)

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43946, Seq: 4417, Ack: 1729, Len: 24
> NVM Express Fabrics TCP, Cqe NVMe Cmd: Set Features (0x09) Cmd ID: 0x0010
v NVM Express (Cqe)
    [Cmd in: 30]
    [Cmd Latency: 0.016 ms]
  v DWORD0: Set Feature Number of Queues Result: 0x00070007
      .... .... .... .... 0000 0000 0000 0111 = Number of IO Submission Queues Allocated: 7 (8)
      0000 0000 0000 0111 .... .... .... .... = Number of IO Completion Queues Allocated: 7 (8)
    DWORD1: 0x00000000
    SQ Head Pointer: 0x0000
    SQ Identifier: 0x0000
    Command Identifier: 0x0010
  v Status Field: 0x0000
      .... .... .... ...0 = Phase Tag: 0x0
      .... ...0 0000 000. = Status Code: 0x00 (Successful Completion)
      .... 000. .... .... = Status Code Type: Generic Command Status (0x0)
      ..00 .... .... .... = Command Retry Delay: 0x0
      .0.. .... .... .... = More Information in Log Page: False
      0... .... .... .... = Do not Retry: False
```


Kamal Bakshi, Cisco



Create 8 I/O queues

e.g. Connect Request for 1st I/O queue

Source TCP port# 43948

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43948, Dst Port: 4420, Seq: 129, Ack: 129, Len: 1096
v NVM Express Fabrics TCP, Fabrics Type: Connect (0x01) Cmd ID: 0x0000
    [Cmd Qid: 1 (IOQ)]
    Pdu Type: CapsuleCommand (4)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 72
    Packet Length: 1096
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [Fabric Cqe in: 90]
      Reserved: 0x40
      Command Identifier: 0x0000
      Fabric Command Type: Connect (0x01)
      Reserved: 0000000000000000000000000000000000000
    > SGL1
      Record Format: 0
      Queue ID: 1 (IOQ)
      Submission Queue Size: 128
    > Connect Attributes: 0x00
      Reserved: 00
      Keep Alive Timeout: 0ms
      Reserved: 000000000000000000000000
  v Data
      Host Identifier: 24756c3c1bd24e9d987da87397e55aa3
      Controller ID: 0x0401
      Reserved: 000000000000000000000000000000000000000000000000000000000000000…
      Subsystem NQN: GB00041004bbf910
      Host NQN: nqn.2014-08.org.nvmexpress:uuid:55e23c7b-7431-439b-81d7-430e286fbb98
      Reserved: 000000000000000000000000000000000000000000000000000000000000000…
```

I/O queue ID (1)

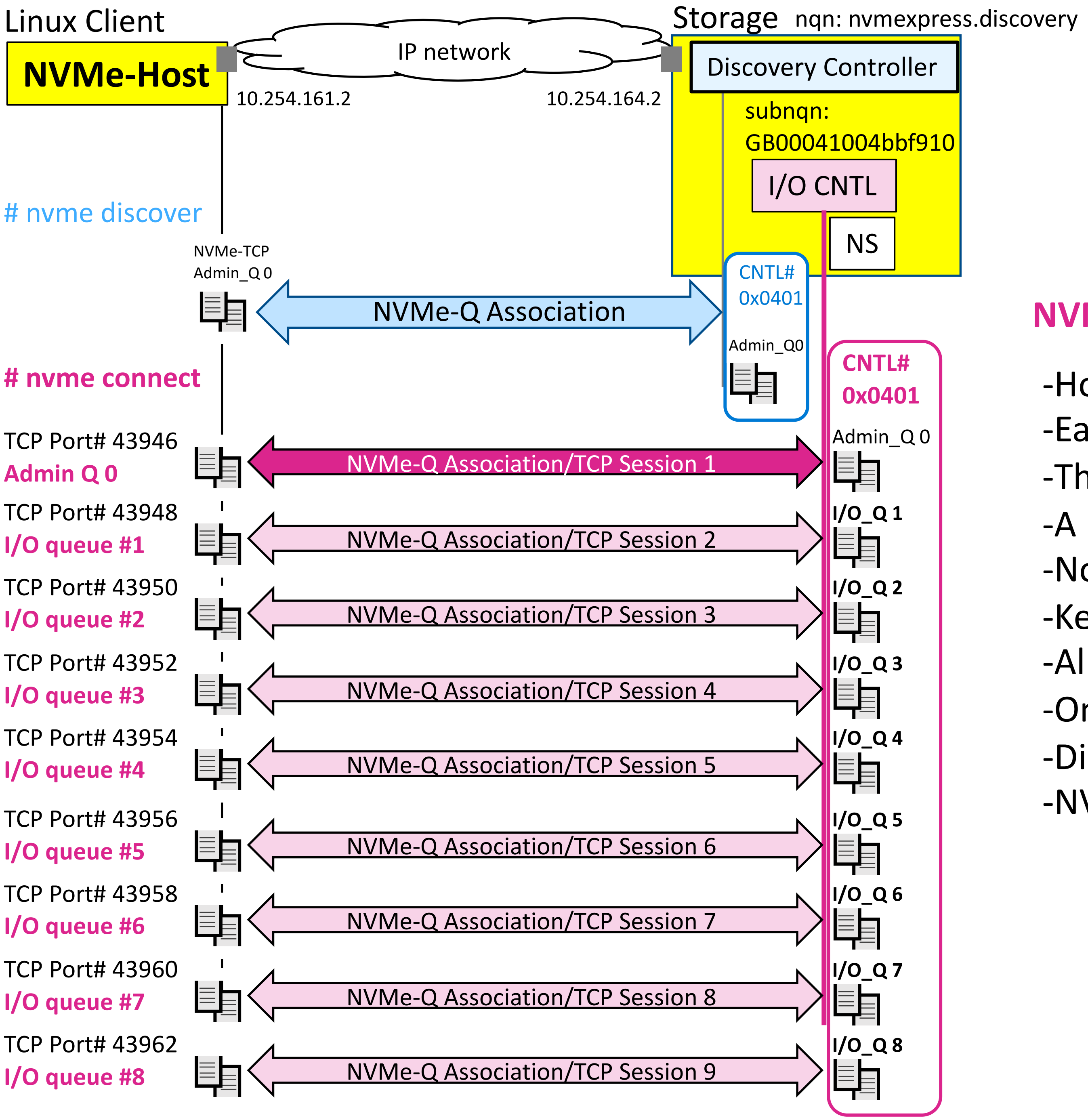
No keep alive for I/O queues

same controller

same Subsystem

Kamal Bakshi, Cisco

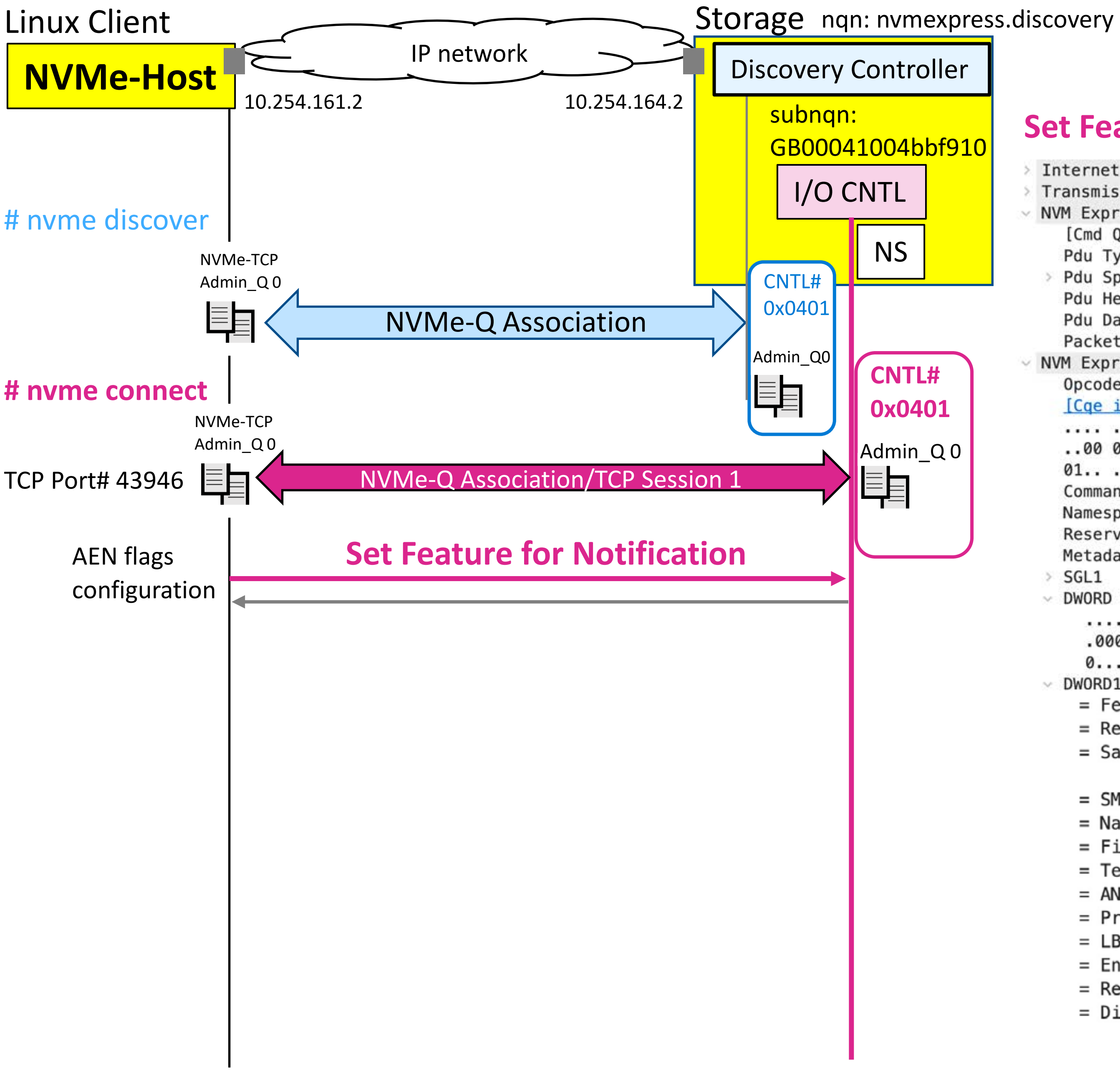
NVMe I/O Data Queues



NVMe Multi Queue Architecture

-Host creates the 8 I/O queues on the same controller
-Each I/O queue is mapped into a unique TCP session
-The NVMe/ TCP port id is 4420
-A unique source TCP port# is provided for each session
-No Keep Alive are maintained for I/O queues session
-Keep Alive is only maintained for the Admin_Q 0
-All NVMe commands are processed on Admin queue
-Only I/O commands, like read/write go over I/O queues
-Discovery controller does not have any I/O queues
-NVMe architecture allows up to 64k I/O queues

Set Notification Flag



Set Feature (Notification Flags)

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq: 1729, Ack: 4441, Len: 72
v NVM Express Fabrics TCP, NVMe Opcode: Set Features (0x09) Cmd ID: 0x0016
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
v NVM Express (Cmd)
    Opcode: 0x09 Set Features
    [Cqe in: 115]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x0016
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  > SGL1
  v DWORD 10: 0x0000000b
      .... .... .... .... .... .... 0000 1011 = Feature Identifier: Asynchronous Event Configuration (0x0b)
      .000 0000 0000 0000 0000 0000 .... .... = Reserved: 0x000000
      0... .... .... .... .... .... .... .... = Save: 0x0
  v DWORD11: 0x00000900
      = Feature Identifier: Asynchronous Event Configuration (0x0b)
      = Reserved: 0x000000
      = Save: 0x0

      = SMART and Health Critical Warnings Bitmask: 0x00
      = Namespace Attribute Notices: True
      = Firmware Activation Notices: False
      = Telemetry Log Notices: False
      = ANA Change Notices: True
      = Predictable Latency Event Aggregate Log Change Notices: False
      = LBA Status Information Notices: False
      = Endurance Group Event Aggregate Log Change Notices: False
      = Reserved: 0x0000
      = Discovery Log Page Change Notification: False
```

Set Async. Event Notification Flags

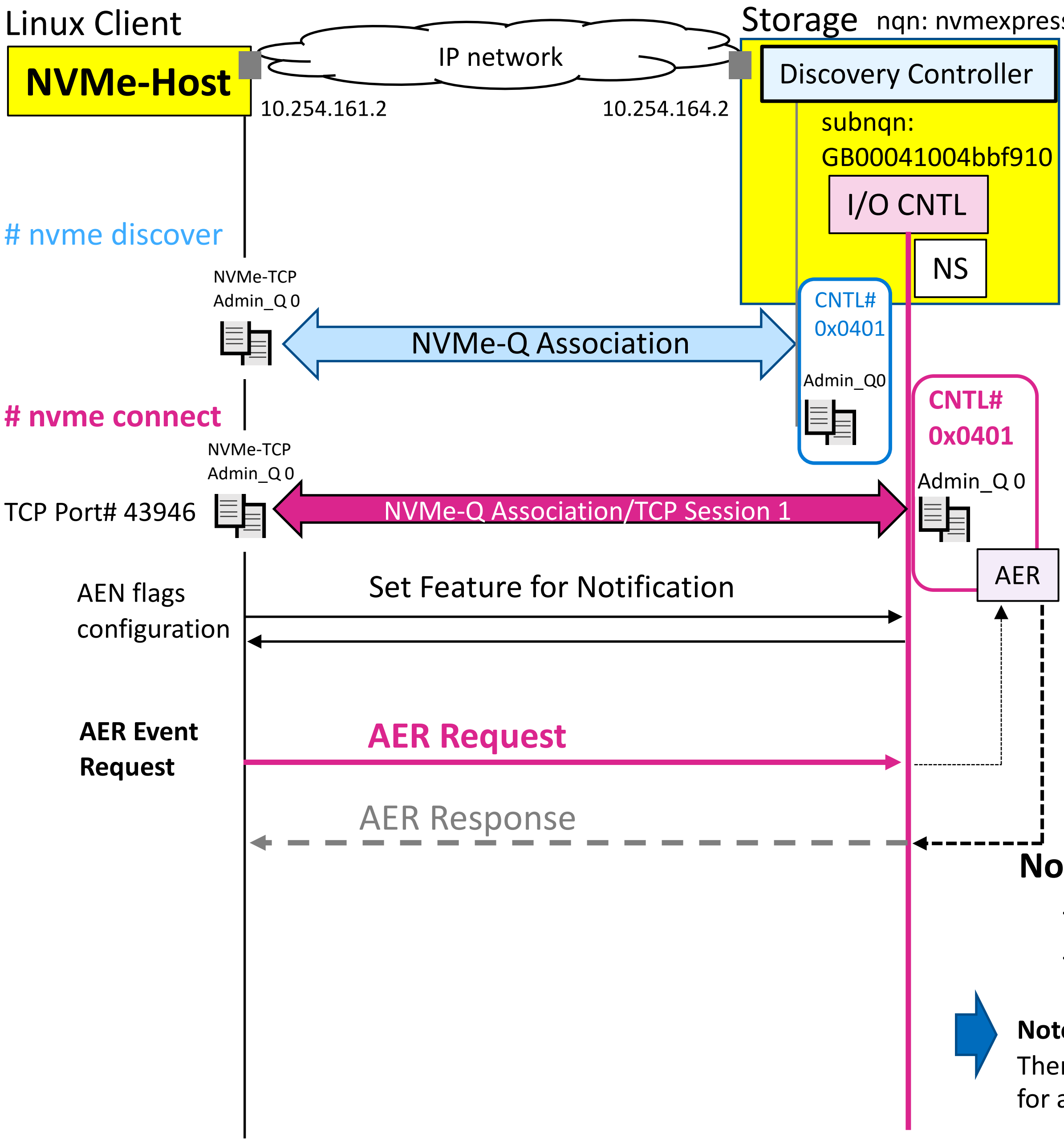
Notify for any changes in

-Namespace attribute

-ANA changes

Kamal Bakshi, Cisco

Send Notification Request



Async Event Request

```
Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq: 1945, Ack: 8609, Len: 72
NVM Express Fabrics TCP, NVMe Opcode: Async Event Request (0x0c) Cmd ID: 0x001f
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
NVM Express (Cmd)
    Opcode: 0x0c Async Event Request
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x001f
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
    SGL1
    DWORD10: 0x00000000
    DWORD11: 0x00000000
    DWORD12: 0x00000000
    DWORD13: 0x00000000
    DWORD14: 0x00000000
    DWORD15: 0x00000000
```

Previous Request: Async Event Config. (Notify Flags)

```
= Feature Identifier: Asynchronous Event Configuration (0x0b)
= Reserved: 0x000000
= Save: 0x0

= SMART and Health Critical Warnings Bitmask: 0x00
= Namespace Attribute Notices: True
= Firmware Activation Notices: False
= Telemetry Log Notices: False
= ANA Change Notices: True
= Predictable Latency Event Aggregate Log Change Notices: False
= LBA Status Information Notices: False
= Endurance Group Event Aggregate Log Change Notices: False
= Reserved: 0x0000
= Discovery Log Page Change Notification: False
```

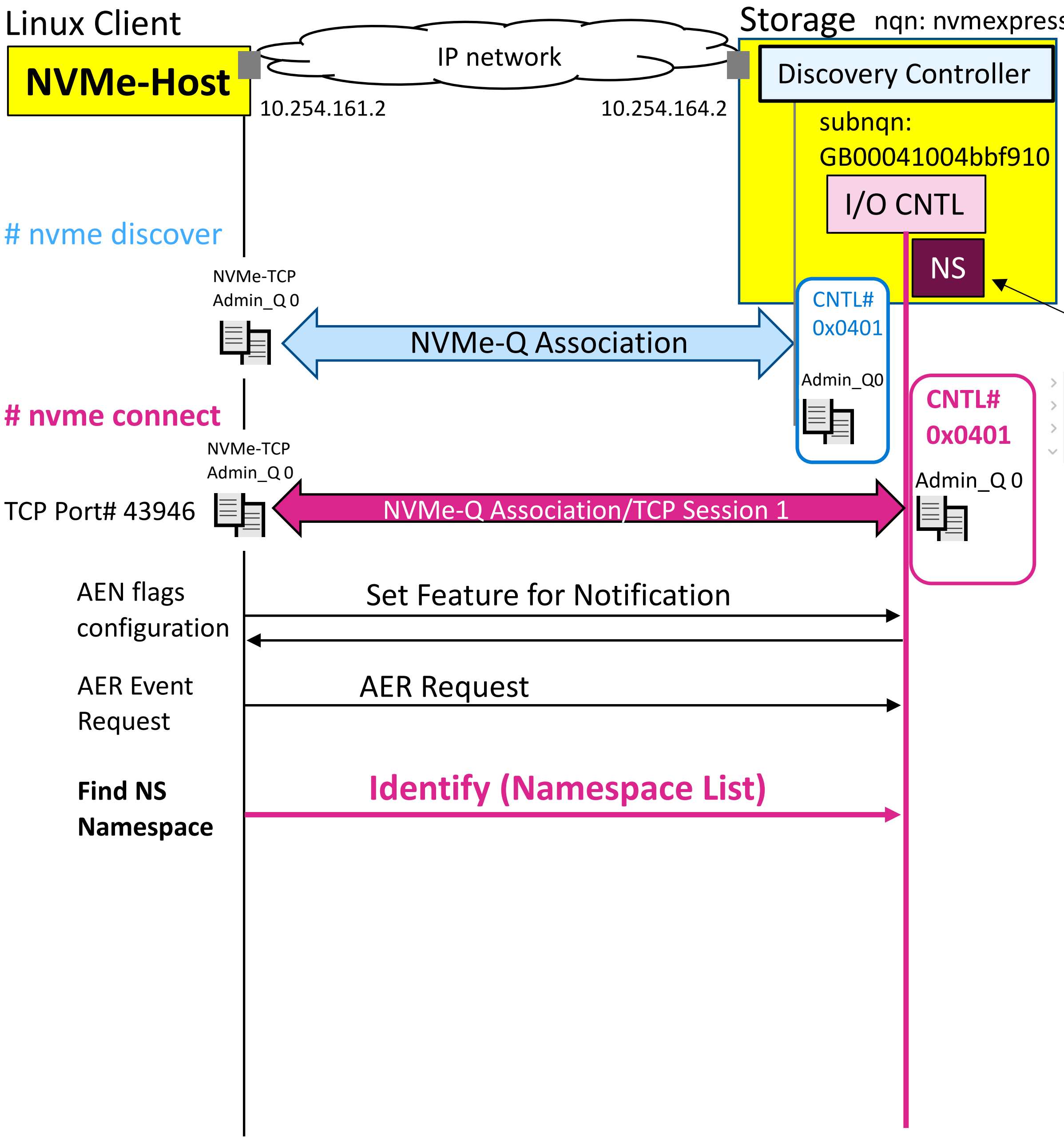
Notify for any changes in

- -Namespace attribute
- -ANA changes

Note:
There is no timeout value
for an outstanding AER request

Kamal Bakshi, Cisco

Identify Active Namespaces



Active Namespace List

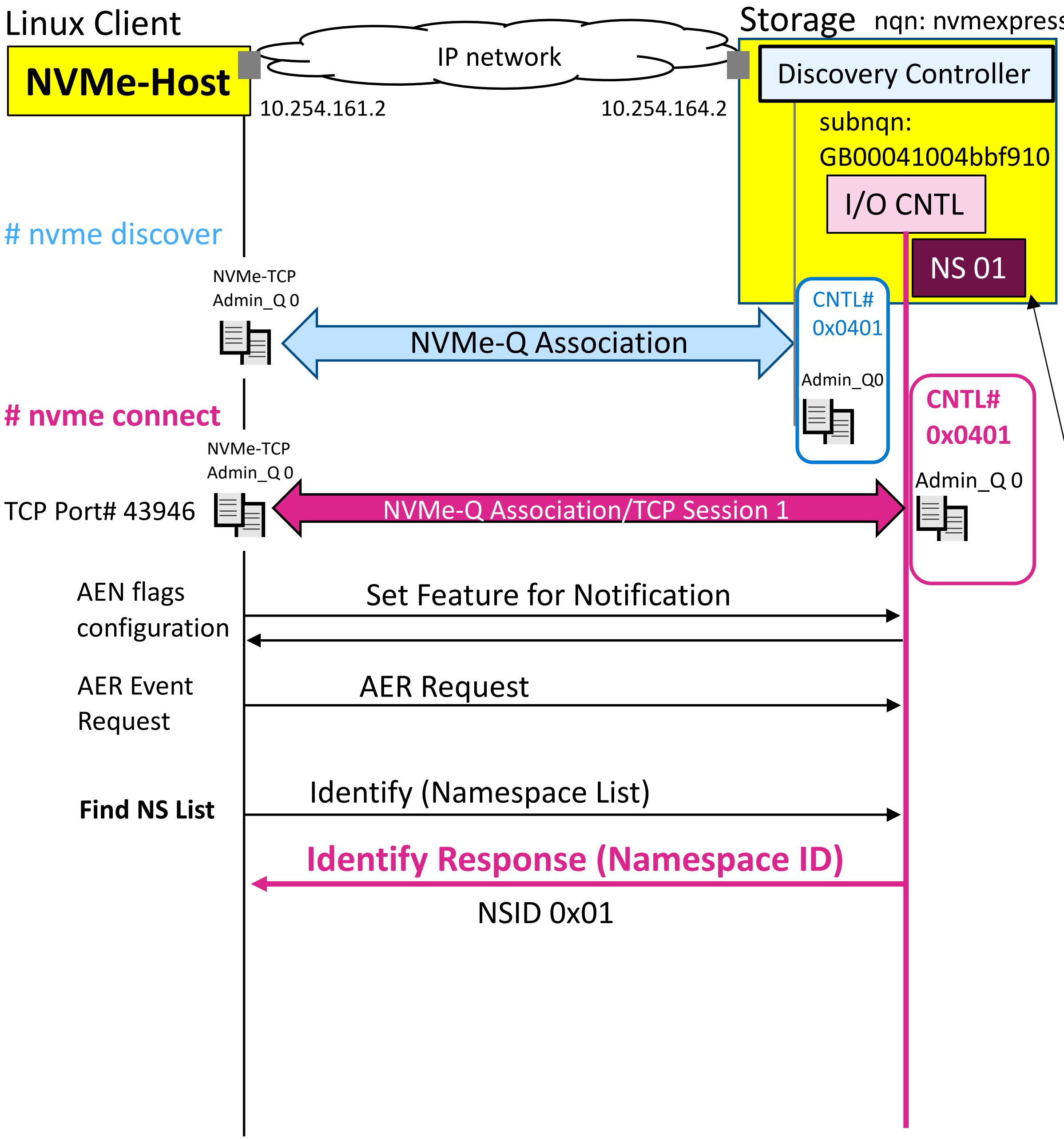
Identify the Namespace List

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq: 1873, Ack: 8609, Len: 72
> NVM Express Fabrics TCP, NVMe Opcode: Identify (0x06) Cmd ID: 0x000b
v NVM Express (Cmd)
    Opcode: 0x06 Identify
    [DATA Transfer 0: 120]
    [Cqe in: 120]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x000b
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  v SGL1
      0101 .... = Descriptor Type: 0x5 Reserved
      .... 1010 = Descriptor Sub Type: 0xa Reserved
  v DWORD10: 0x00000002
      .... .... .... .... .... .... 0000 0010 = Controller or Namespace Structure (CNS): Active Namespace List (0x02)
      .... .... .... .... 0000 0000 .... .... = Reserved: 0x00
      0000 0000 0000 0000 .... .... .... .... = Controller Identifier (CNTID): 0x0000
  > DWORD11: 0x00000000
    DWORD12: 0x00000000
    DWORD13: 0x00000000
  > DWORD14: 0x00000000
    DWORD15: 0x00000000
```


Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE

Active Namespace List



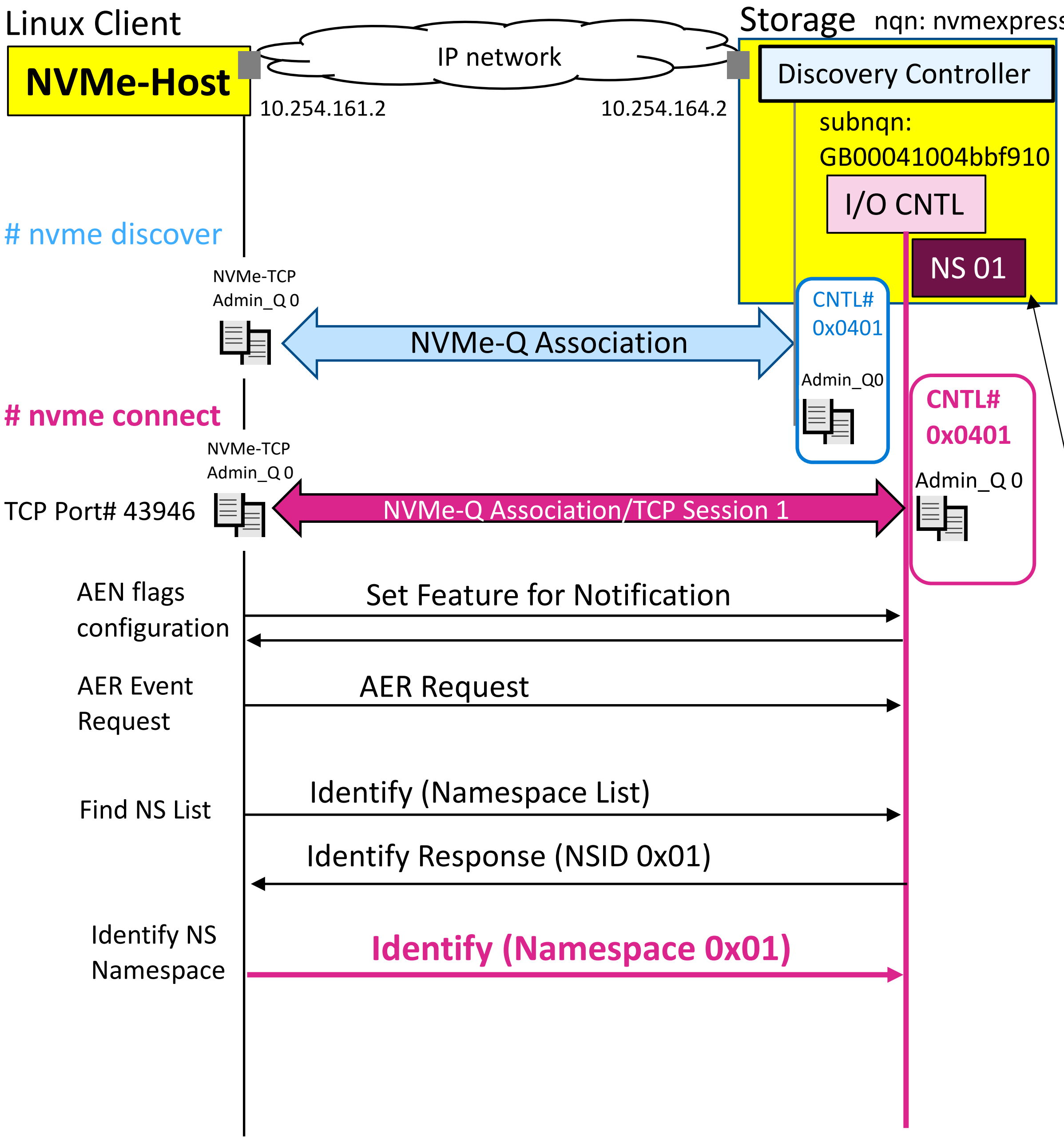
Identify Response (Namespace ID)

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43946, Seq: 8609, Ack: 1945,
v NVM Express Fabrics TCP, C2HData Opcode: Identify (0x06), Cmd ID: 0x000b, Len: 4096
    [Cmd Qid: 0 (AQ)]
    Pdu Type: C2HData (7)
  > Pdu Specific Flags: 0x04, PDU Data Last
    Pdu Header Length: 24
    Pdu Data Offset: 24
    Packet Length: 4120
  > NVMe/TCP Data PDU
    [Cmd in: 118]
v NVM Express
    nsid[0]: 1
v NVM Express Fabrics TCP, Cqe NVMe Cmd: Identify (0x06) Cmd ID: 0x000b
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
> NVM Express (Cqe)
```

Name space ID 0x01

Kamal Bakshi, Cisco

Identify Namespace CNS-0



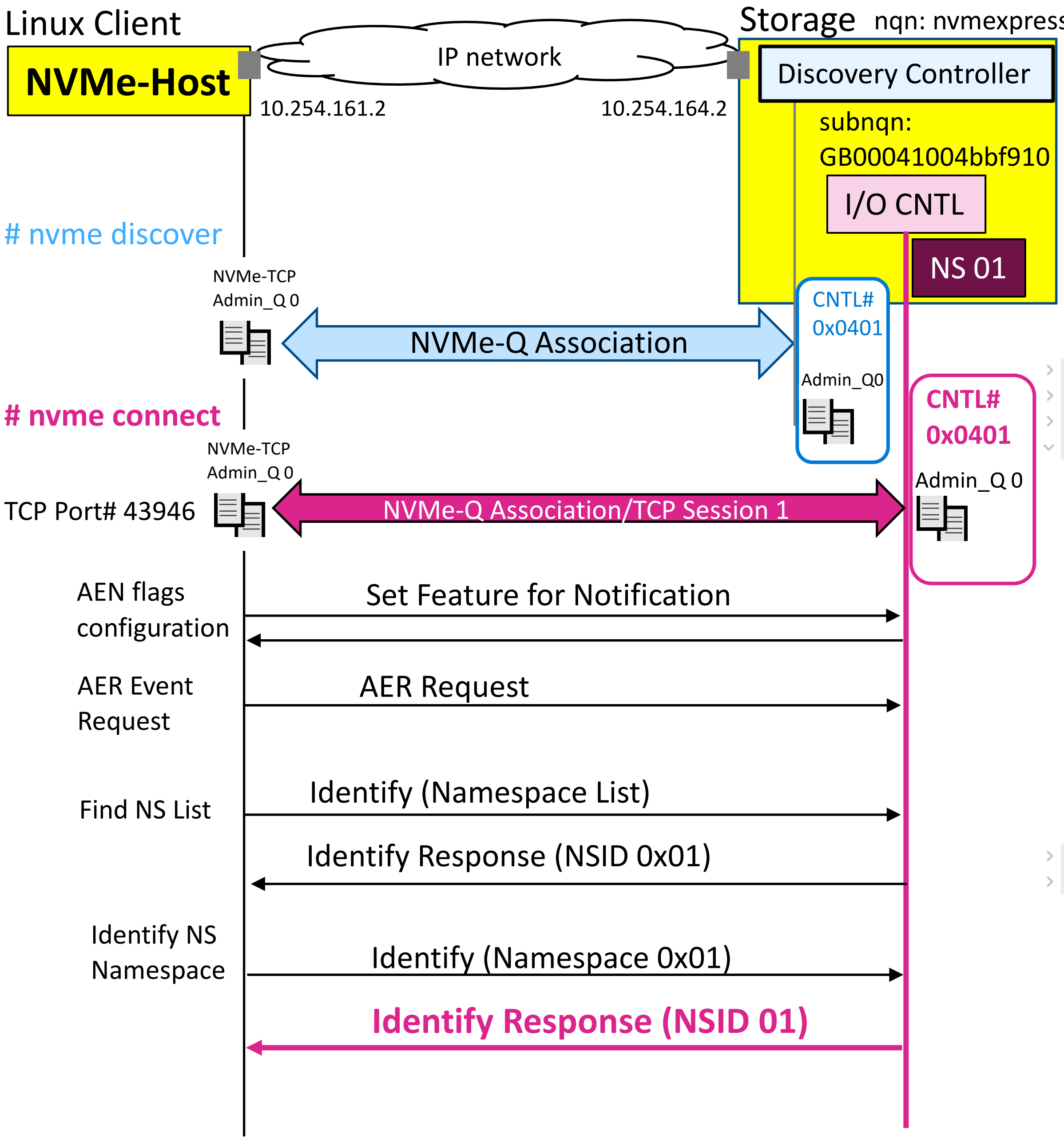
Identify (Namespace NSID 0x01)

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43946, Dst Port: 4420, Seq: 2017, Ack: 12753, Len: 72
v NVM Express Fabrics TCP, NVMe Opcode: Identify (0x06) Cmd ID: 0x000c
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
v NVM Express (Cmd)
    Opcode: 0x06 Identify
    [DATA Transfer 0: 123]
    [Cqe in: 123]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x000c
    Namespace Id: 0x00000001
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  > SGL1
  v DWORD10: 0x00000000
      .... .... .... .... .... .... 0000 0000 = Controller or Namespace Structure (CNS): Namespace (0x00)
      .... .... .... .... 0000 0000 .... .... = Reserved: 0x00
      0000 0000 0000 0000 .... .... .... .... = Controller Identifier (CNTID): 0x0000
  > DWORD11: 0x00000000
    DWORD12: 0x00000000
    DWORD13: 0x00000000
  > DWORD14: 0x00000000
    DWORD15: 0x00000000
```

NSID 0x01

Kamal Bakshi, Cisco

Namespace Info returned



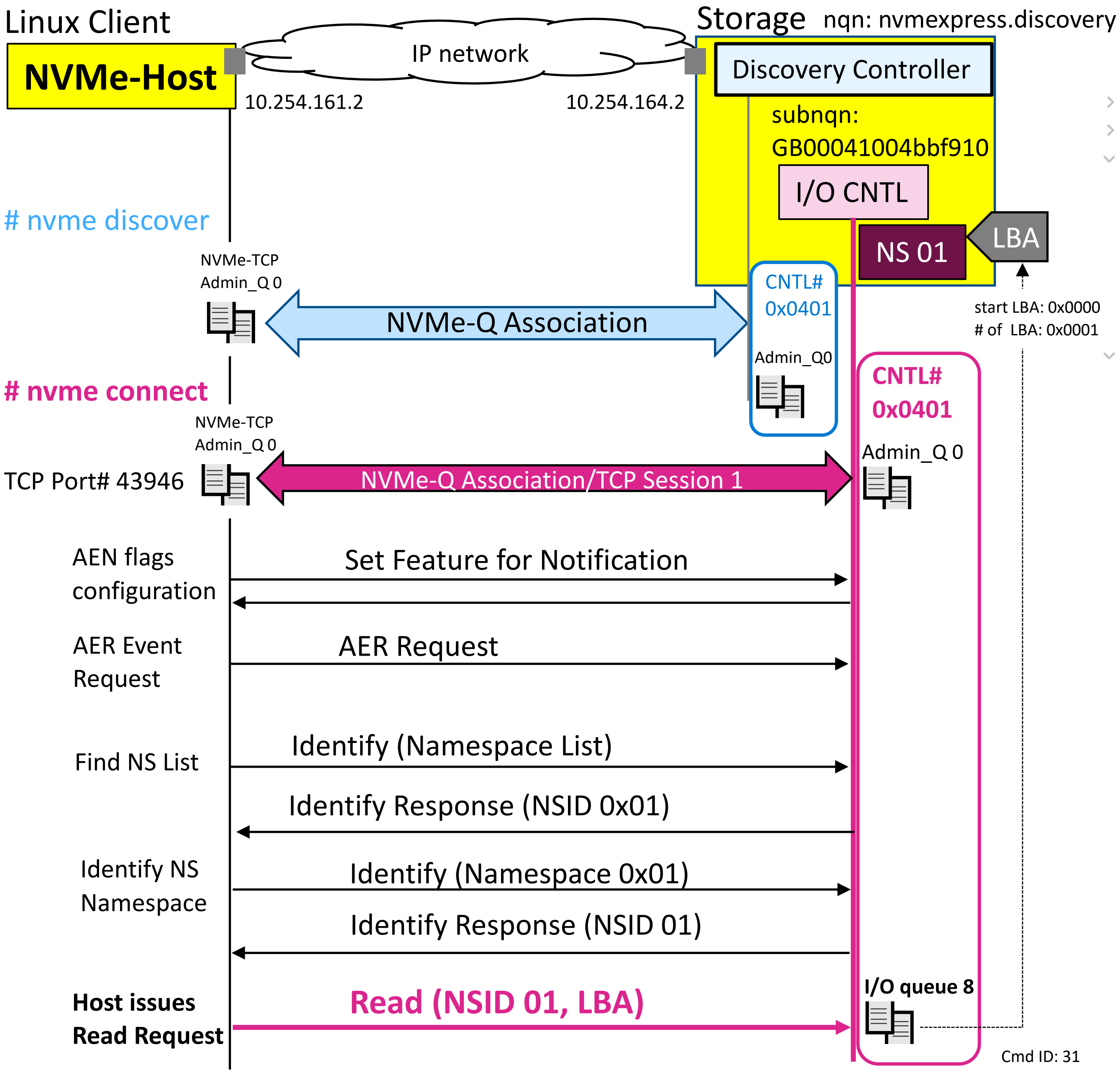
Identify NSID Response (detailed info regarding namespace)

```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43946, Seq: 12753, Ack: 2089, Len: 4144
> NVM Express Fabrics TCP, C2HData Opcode: Identify (0x06), Cmd ID: 0x000c, Len: 4096
v NVM Express
    Namespace Size (NSZE): 524288000 (0x000000001f400000)
    Namespace Capacity (NCAP): 0x000000001f400000
    Namespace Utilization (NUSE): 0x000000001f400000
    Namespace Features (NSFEAT): 0x00
    Number of LBA Formats (NLBAF): 0 (0x00)
    Formatted LBA Size (FLBAS): 0 (0x00)
    Metadata Capabilities (MC): 0x00
    End-to-end Data Protection Capabilities (DPC): 0x00
    End-to-end Data Protection Type Settings (DPS): 0x00
    Namespace Multi-path I/O and Namespace Sharing Capabilities (NMIC): 0x00
    Namespace Globally Unique Identifier (NGUID): 63393539333565622d303537392d3466
    IEEE Extended Unique Identifier (EUI64): 0000000000000000
  > LBA Formats
    Reserved: 0000000000000000000000000000000000000000000000000000000000000000…
    Vendor Specific: 0000000000000000000000000000000000000000000000000000000000000000… (offset
> NVM Express Fabrics TCP, Cqe NVMe Cmd: Identify (0x06) Cmd ID: 0x000c
> NVM Express (Cqe)
```

Namespace Info

Kamal Bakshi, Cisco

Issue NVMe Read



Read

QID #08 (Src Port 43962)

```
> Internet Protocol Version 4, Src: 10.254.161.2, Dst: 10.254.164.2
> Transmission Control Protocol, Src Port: 43962, Dst Port: 4420, Seq:
v NVM Express Fabrics TCP, NVMe Opcode: Read (0x02) Cmd ID: 0x0031
    [Cmd Qid: 8 (IOQ)]
    Pdu Type: CapsuleCommand (4)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
v NVM Express (Cmd)
    Opcode: 0x02 Read
    [DATA Transfer 0: 127]
    [Cqe in: 127]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x0031
    Namespace Id: 0x00000001
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  v SGL1
      0101 .... = Descriptor Type: 0x5 Reserved
      .... 1010 = Descriptor Sub Type: 0xa Reserved
    Start LBA: 0x0000000000000000
    Absolute Number of Logical Blocks: 1 (0x0001)
    .... ..00 0000 0000 = Reserved: 0x000
  > .... .0.. .... .... = Protection info fields: 0x0
    .0.. .... .... .... = Force Unit Access: 0x0
    0... .... .... .... = Limited Retry: 0x0
    Expected Initial Logical Block Reference Tag: 0x00000000
    Expected Logical Block Application Tag Mask: 0x0000
    Expected Logical Block Application Tag: 0x0000
  > DSM Flags
    Reserved: 000000
```

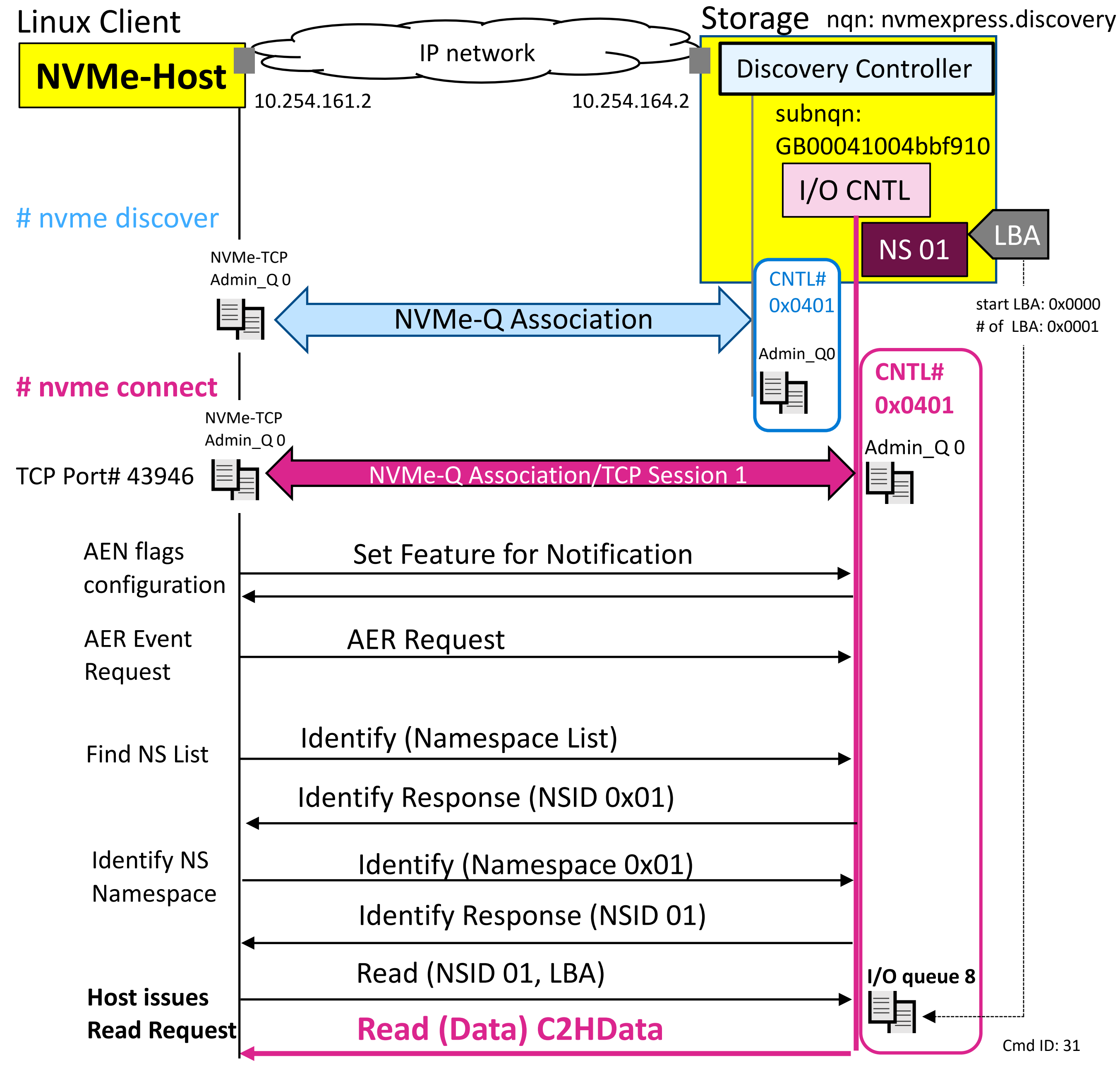
Namespace ID

Starting LBA/total blocks

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

NVMe Read Data



Read (Data)

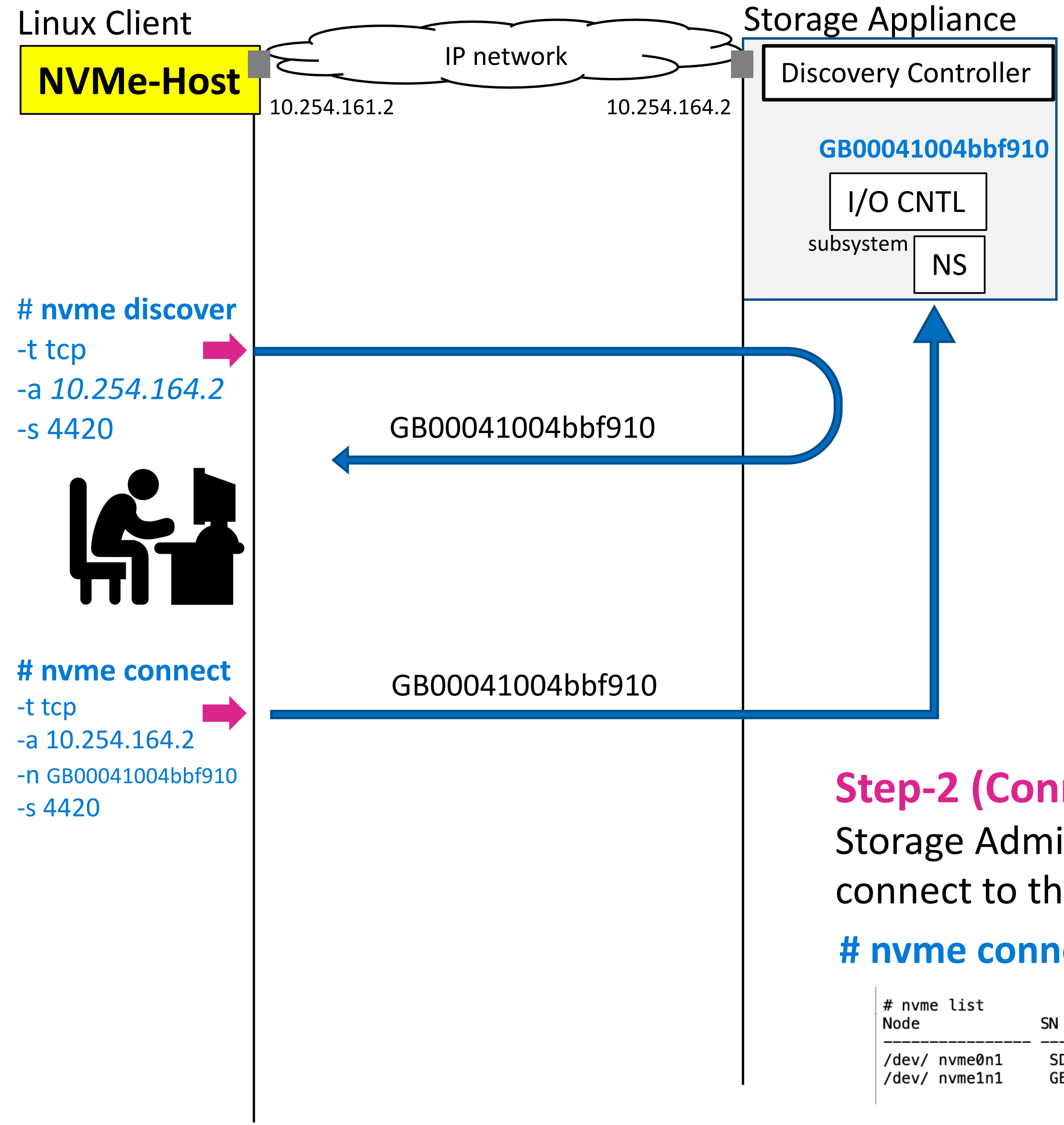
```
> Internet Protocol Version 4, Src: 10.254.164.2, Dst: 10.254.161.2
> Transmission Control Protocol, Src Port: 4420, Dst Port: 43962, Seq: 153, Ack: 1297,
v NVM Express Fabrics TCP, C2HData Opcode: Read (0x02), Cmd ID: 0x0031, Len: 4096
    [Cmd Qid: 8 (IOQ)]
    Pdu Type: C2HData (7)
  > Pdu Specific Flags: 0x04, PDU Data Last
    Pdu Header Length: 24
    Pdu Data Offset: 24
    Packet Length: 4120
  v NVMe/TCP Data PDU
      Command ID: 0x0031
      Transfer Tag: 0x0000
      Data Offset: 0
      Data Length: 4096
      Reserved: 00000000
    [Cmd in: 126]
  NVM Express
v NVM Express Fabrics TCP, Cqe NVMe Cmd: Read (0x02) Cmd ID: 0x0031
    [Cmd Qid: 8 (IOQ)]
    Pdu Type: CapsuleResponse (5)
  > Pdu Specific Flags: 0x00
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
v NVM Express (Cqe)
    [Cmd in: 126]
    [Cmd Latency: 0.239 ms]
    DWORD0: 0x00000000
    DWORD1: 0x00000000
    SQ Head Pointer: 0x0000
    SQ Identifier: 0x0008
    Command Identifier: 0x0031
  > Status Field: 0x0000
```


Kamal Bakshi, Cisco

NVMe/TCP Flows with CDC

Kamal Bakshi, Cisco

Manual Storage Discovery & Connect



Step-1 (Find Storage Appliance)

Storage Admin will issue a “NVMe discover” CLI command at the host to retrieve the Storage Appliance Subsystem.

 # nvme discover -t tcp -a *10.254.164.2* -s 4420

```
Discovery Log Number of Records 1, Generation counter 1
=====Discovery Log Entry 0======
trtype:  unrecognized
adrfam: ipv4
subtype: nvme subsystem
treq:    not specified
portid:  28
trsvcid: 4420
subnqn:  GB00041004bbf910
traddr:  10.254.164.2
```

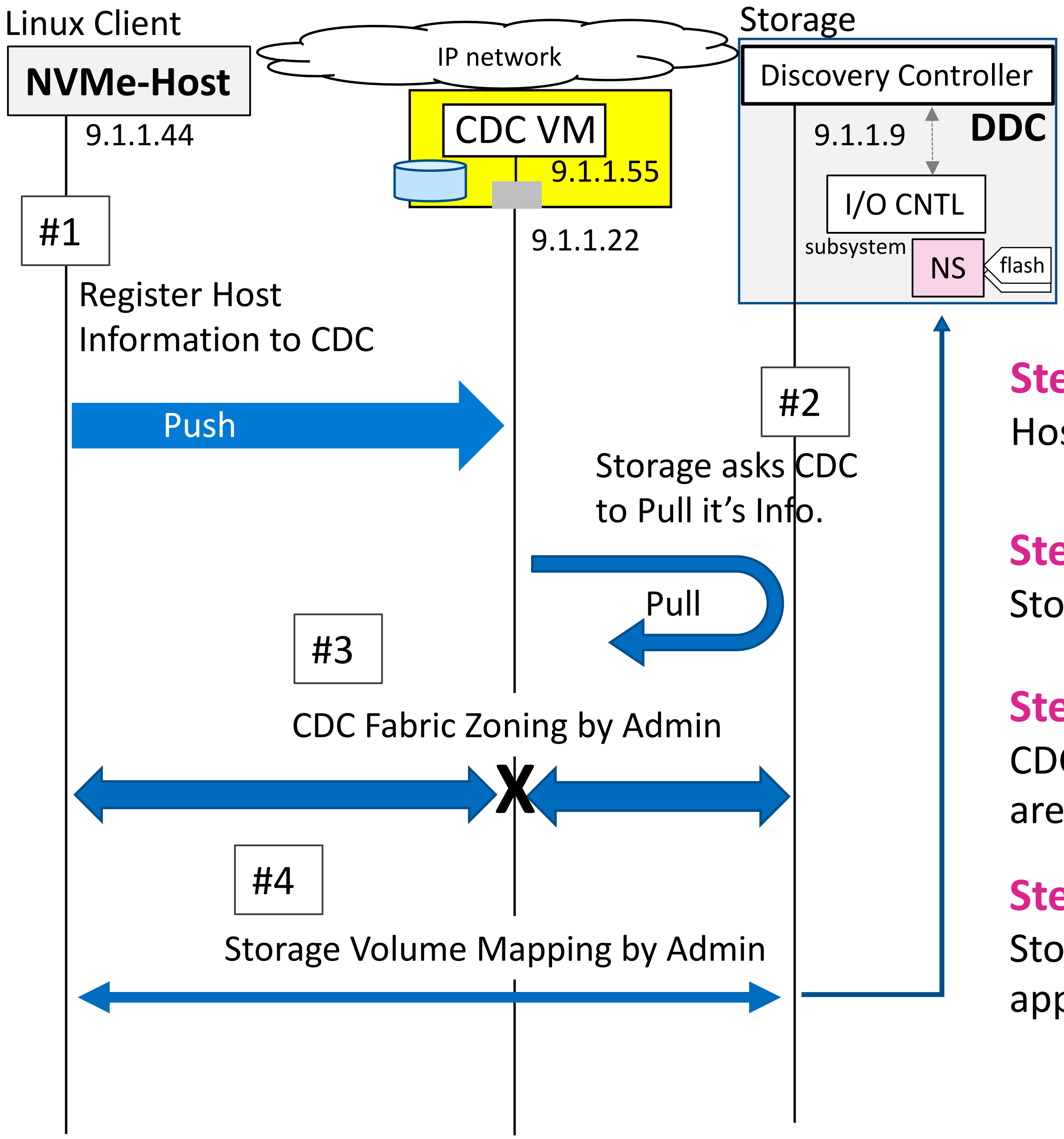
Step-2 (Connect to Storage Appliance)

Storage Admin will issue a “NVMe connect” CLI command at the host to connect to the Storage Appliance Subsystem.

 # nvme connect -t tcp -a 10.254.164.2 -n GB00041004bbf910 -s 4420

```
# nvme list
Node             SN                   Model                                    Namespace Usage                      Format           FW Rev
---------------- -------------------- ---------------------------------------- --------- -------------------------- ---------------- --------
/dev/ nvme0n1     SDM00000EC75         UCSC-F-H16003                              1          1.60  TB /   1.60  TB    512   B +  0 B   KNCCP100
/dev/ nvme1n1     GB00041004bbf910     PVL-MX18S0P2L2C1-F100TP0TY1                1          2.15  TB /   2.15  TB      4 KiB +  0 B   22139242
```


Kamal Bakshi, Cisco



Automatic Storage Discovery & Connect

CDC: New NVMe standard TP8009, TP8010

Step-1 Auto (Host Registration with CDC)
Host automatically finds CDC and pushes it's information to it.

Step-2 Auto (Storage Registration with CDC)
Storage automatically finds CDC and asks CDC to pull it's information.

Step-3 CDC Admin (Zoning [Host, Subsystem])
CDC admin configures the Host NQN and Subsystem's NQN (both NQNs are automatically discovered) into a same zone.

Step-4 Storage Admin (Mapping [Host, Volume])
Storage admin maps the Host NQN (automatically discovered) to the appropriate volume.

CDC: Centralized Discovery Controller

Kamal Bakshi, Cisco

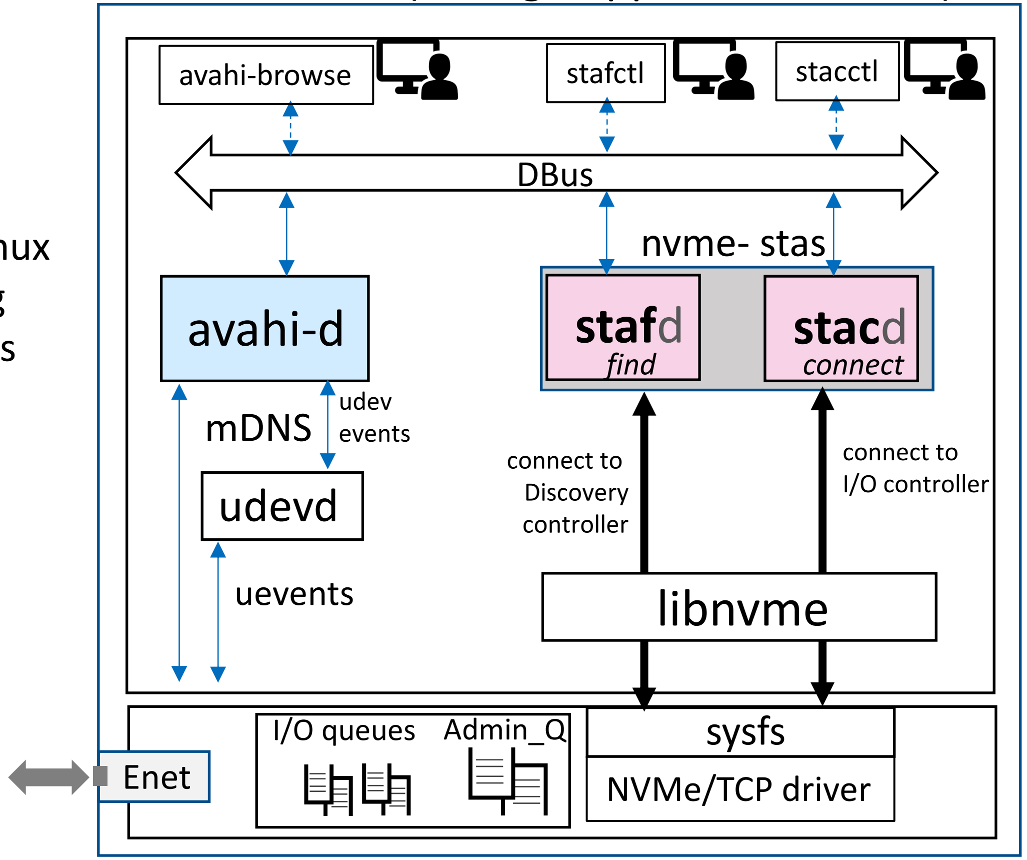
NVMe/TCP CDC Client (NVMe-STAS)

What does nvme-stas provide?

- A Central Discovery Controller (CDC) client for Linux
- Asynchronous Event Notifications (AEN) handling
- Automated NVMe subsystem connection controls
- Error handling and reporting
- Automatic (zeroconf) and Manual configuration

Avahi is a system which facilitates service discovery on a local network via mDNS/DNS-SD protocol suite.

NVMe-STAS (Storage Appliance Services)



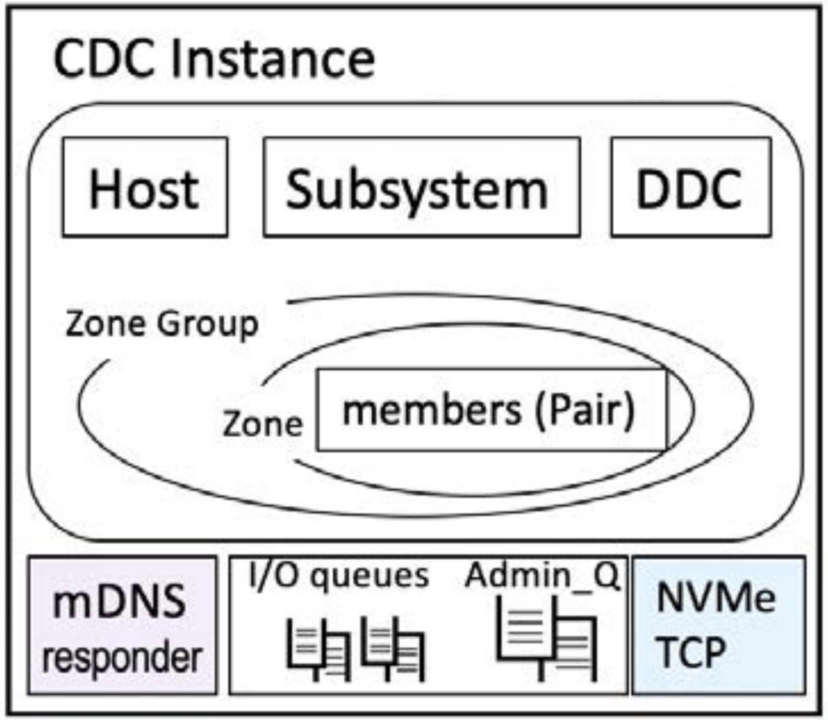
STAF
STorage Appliance Finder

STAC
STorage Appliance Connector

libnvme
Make all NVMe Linux features conveniently reachable to developers

Kamal Bakshi, Cisco

NVMe/TCP CDC Controller



Host Table
-Native CDC Host Info
-Discovered Host Info

Subsystem
-NVMe I/O Controller Info

DDC
-Direct Discovery Controller Info

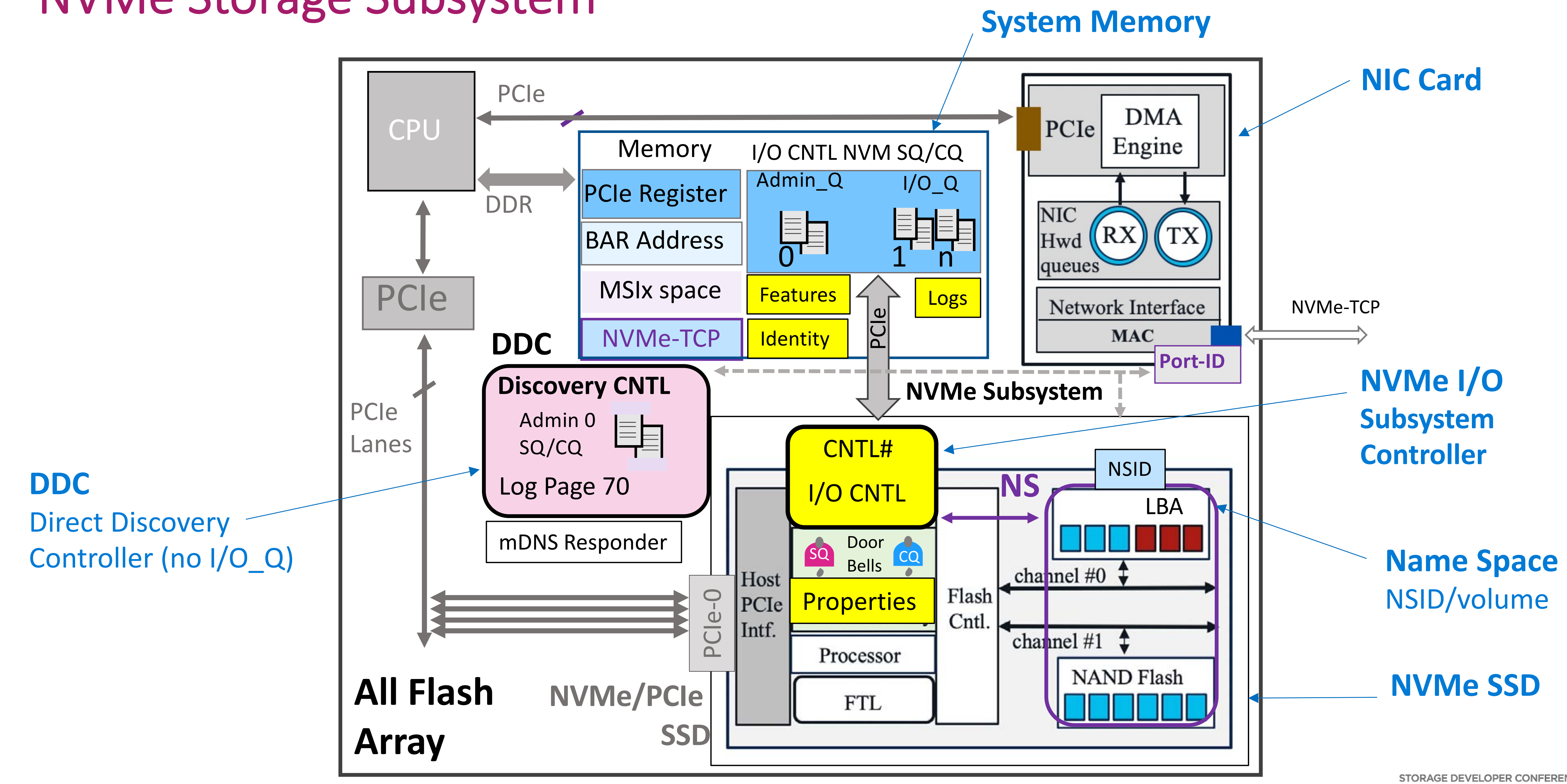
Zoning -Access Control
Which Host has access to what Subsystem?
-Zone Group, Zones
-Members (Pair)
(e.g. Host1, Subsystem3)

mDNS Packet

UDP:dst-224.0.0.251, port-5353			
DNS Query ID (set to 0)			
QR	**Opcode**	Flags	**RCODE**
QDCOUNT (# of questions)			
ANCOUNT (# of answers)			
NSCOUNT			
ARCOUNT			
QNAME (question)			
QTYPE			
QCLASS			
NAME (answer)			
RR TYPE			
CLASS			
TTL			
RDLENGTH			
RDATA			

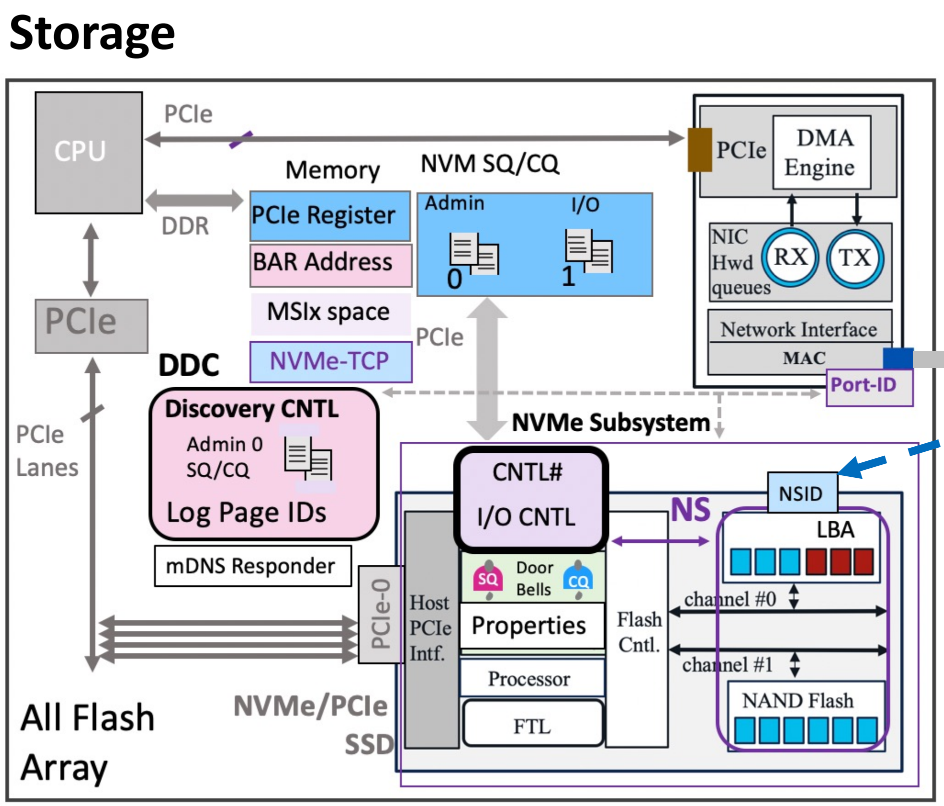
Kamal Bakshi, Cisco

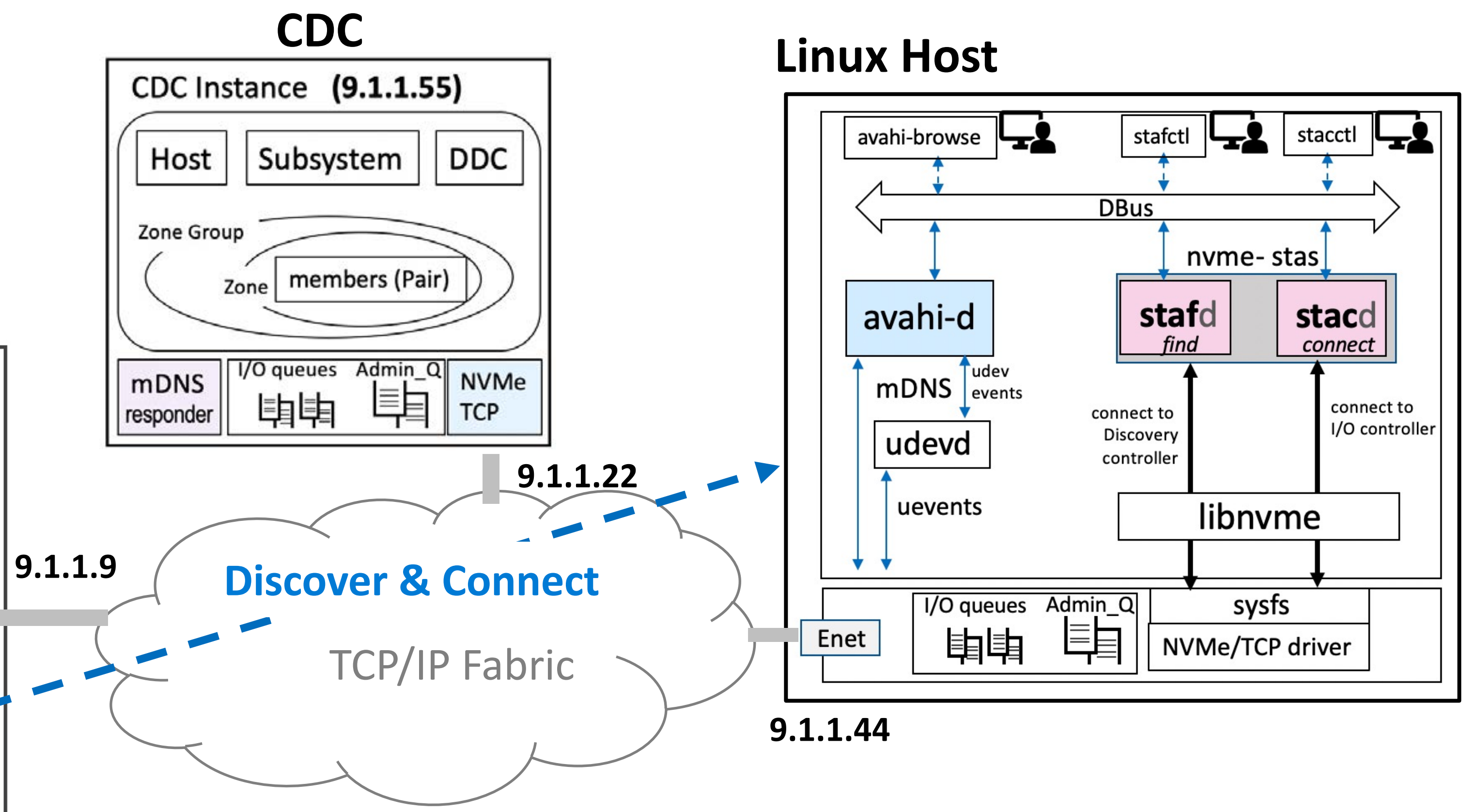
NVMe Storage Subsystem



Kamal Bakshi, Cisco

NVMe/TCP Fabric with CDC



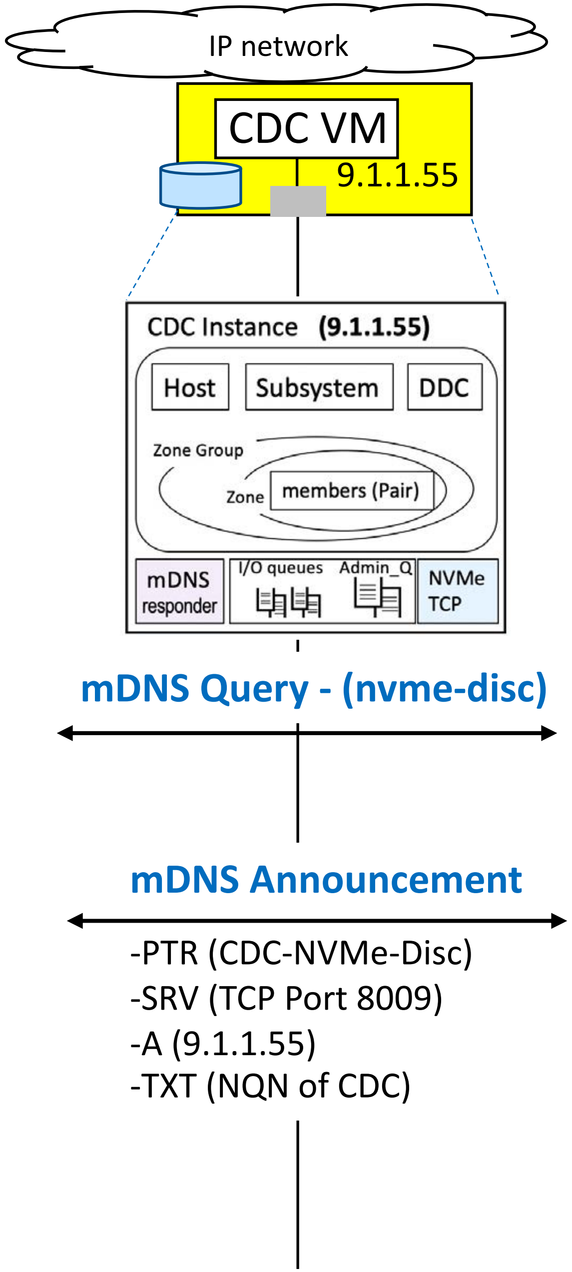


How does “Linux Host” automatically discovers the Storage subsystem and connects to it?

-Detailed flows are explained in next few pages

Kamal Bakshi, Cisco

CDC Initialization



During initialization (e.g., following a link transition or power cycle), before the CDC's mDNS responder function is enabled, the CDC shall probe to ensure the unique resource records the CDC are responsible for are unique on the local link. Upon successful completion of the probe, the CDC shall announce its newly registered resource records. Upon announcing its resource records, the CDC's mDNS responder function may be enabled and respond to queries for the service name of "_nvme-disc.<protocol>.local" or "_cdc_sub._nvme-disc.<protocol>.local"

A CDC may query for both CDC and DDC instances.

mDNS Query

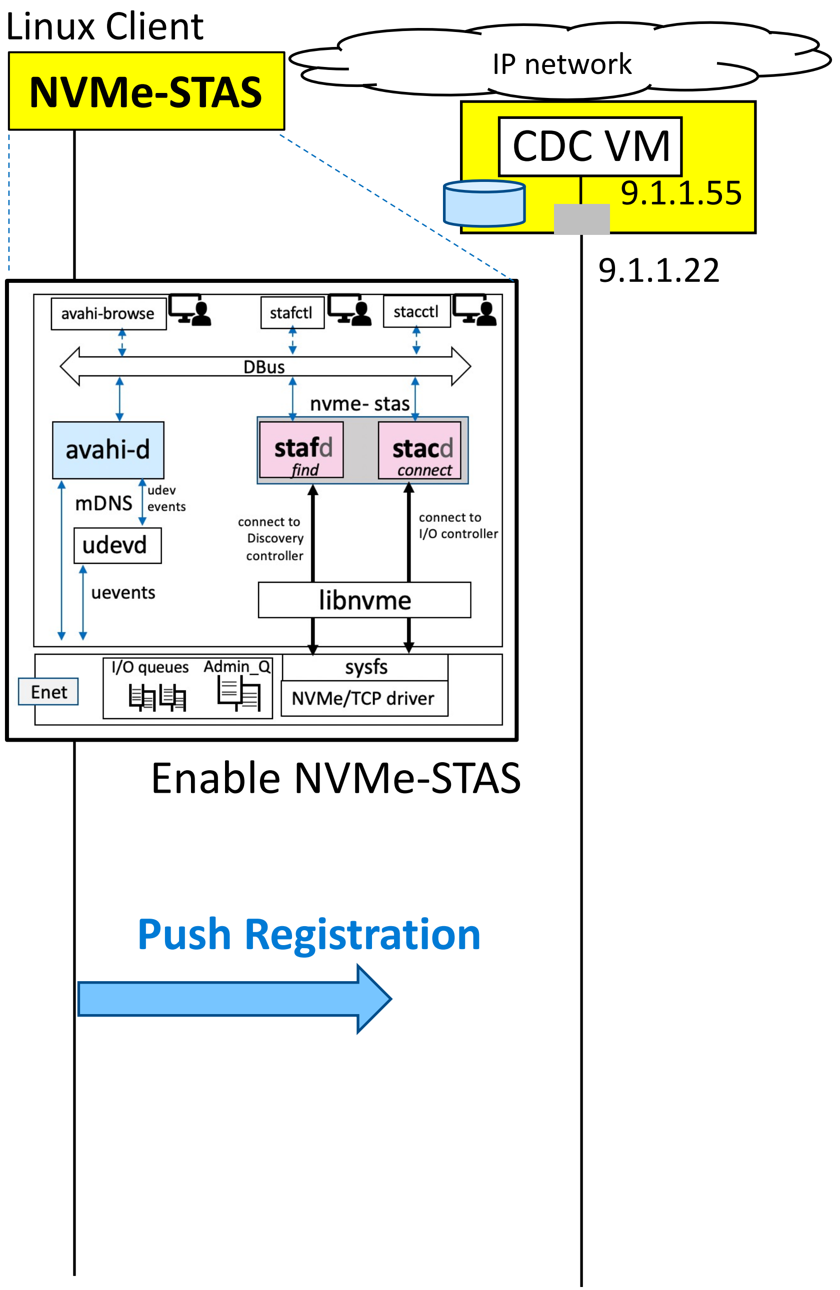
```
⌄ _nvme-disc._tcp.local: type PTR, class IN, "QM" question
      Name: _nvme-disc._tcp.local
      [Name Length: 21]
      [Label Count: 3]
      Type: PTR (domain name PoinTeR) (12)
      .000 0000 0000 0001 = Class: IN (0x0001)
      0... .... .... .... = "QU" question: False
```

mDNS Announcement

```
⌄ Answers
   > 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type TXT, class IN, cache flush
   > _nvme-disc._tcp.local: type PTR, class IN, 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local
   > 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type SRV, class IN, cache flush, priority
   > _services._dns-sd._udp.local: type PTR, class IN, _nvme-disc._tcp.local
   > _cdc._sub._nvme-disc._tcp.local: type PTR, class IN, 9-1-1-55:08/27/22:01:53:05._nvme-disc._
```

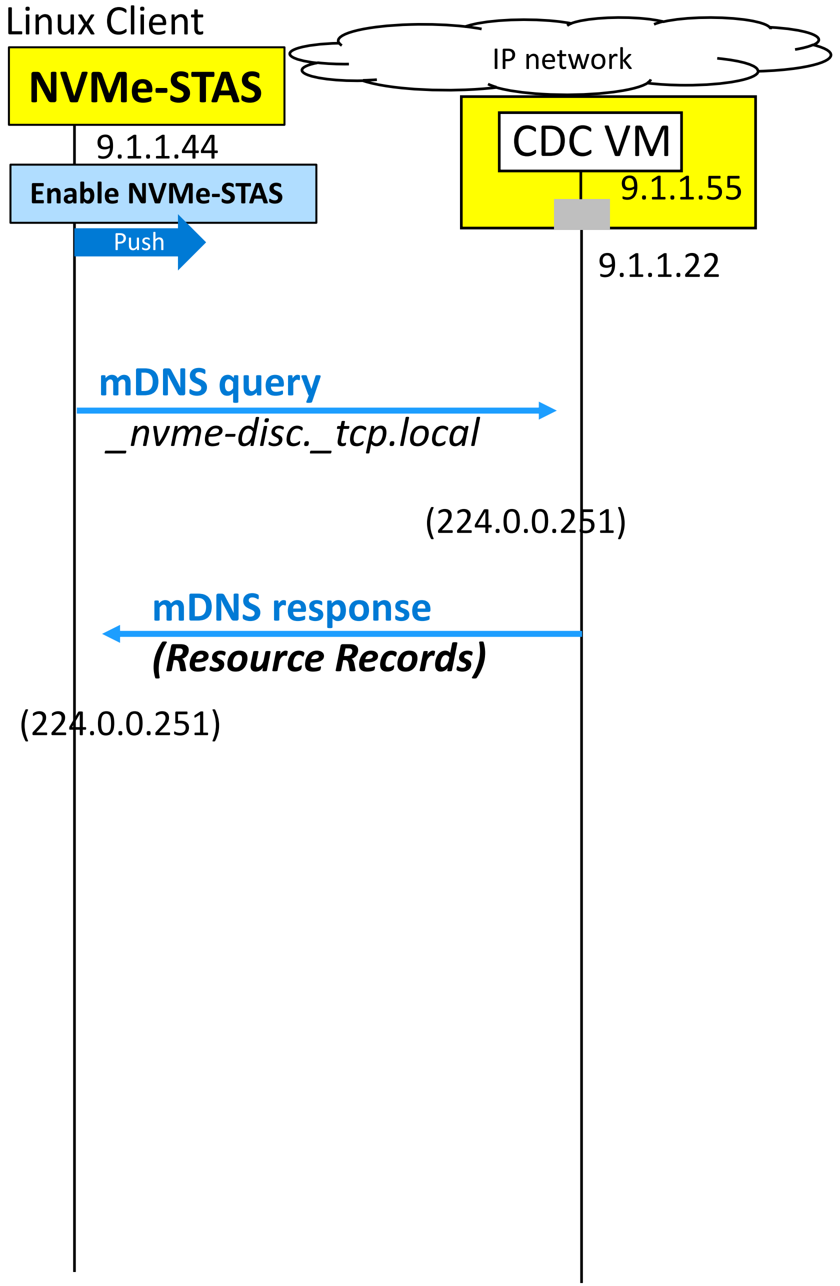

Kamal Bakshi, Cisco

How does Host Discovers CDC?



Kamal Bakshi, Cisco

Host mDNS Query



mDNS Query

```
_nvme-disc._tcp.local: type PTR, class IN, "QM" question
    Name: _nvme-disc._tcp.local
    [Name Length: 21]
    [Label Count: 3]
    Type: PTR (domain name PoinTeR) (12)
    .000 0000 0000 0001 = Class: IN (0x0001)
    0... .... .... .... = "QU" question: False
```

mDNS Response

```
Answers
  _nvme-disc._tcp.local: type PTR, class IN, 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.loca
  9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type TXT, class IN, cache flush
    Name: 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local
    Type: TXT (Text strings) (16)
    .000 0000 0000 0001 = Class: IN (0x0001)
    1... .... .... .... = Cache flush: True
    Time to live: 4500 (1 hour, 15 minutes)
    Data length: 55
    TXT Length: 5
    TXT: p=tcp
    TXT Length: 48
    TXT: NQN=nqn.1988-11.com.dell:SFSS:9:20220824223058e8
  9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type SRV, class IN, cache flush, priorit
    Service: 9-1-1-55:08/27/22:01:53:05
    Protocol: _nvme-disc
    Name: _tcp.local
    Type: SRV (Server Selection) (33)
    .000 0000 0000 0001 = Class: IN (0x0001)
    1... .... .... .... = Cache flush: True
    Time to live: 120 (2 minutes)
    Data length: 17
    Priority: 0
    Weight: 0
    Port: 8009
    Target: 9-1-1-55.local
  9-1-1-55.local: type A, class IN, cache flush, addr 9.1.1.55
```

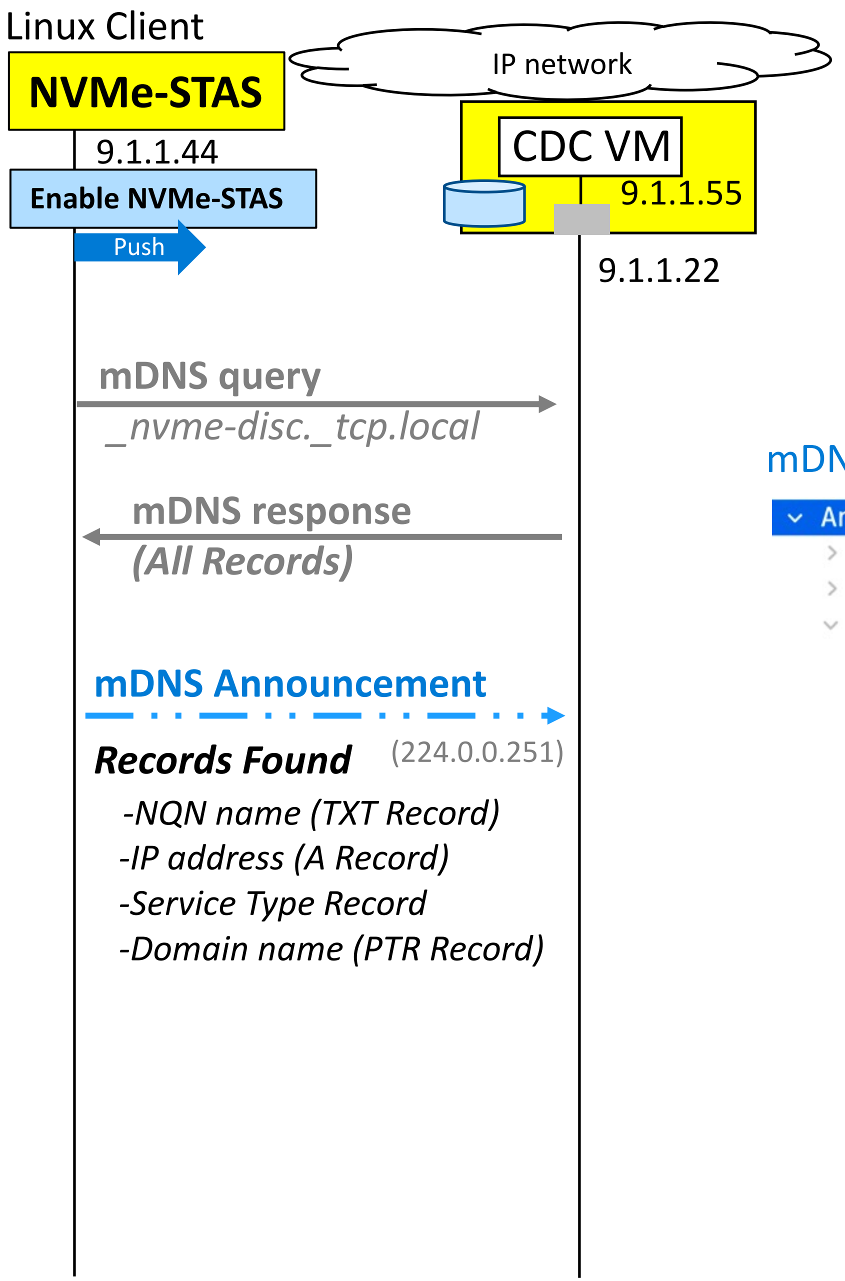
NQN of the CDC

IP address of CDC

 |

Kamal Bakshi, Cisco

Host mDNS Announcement

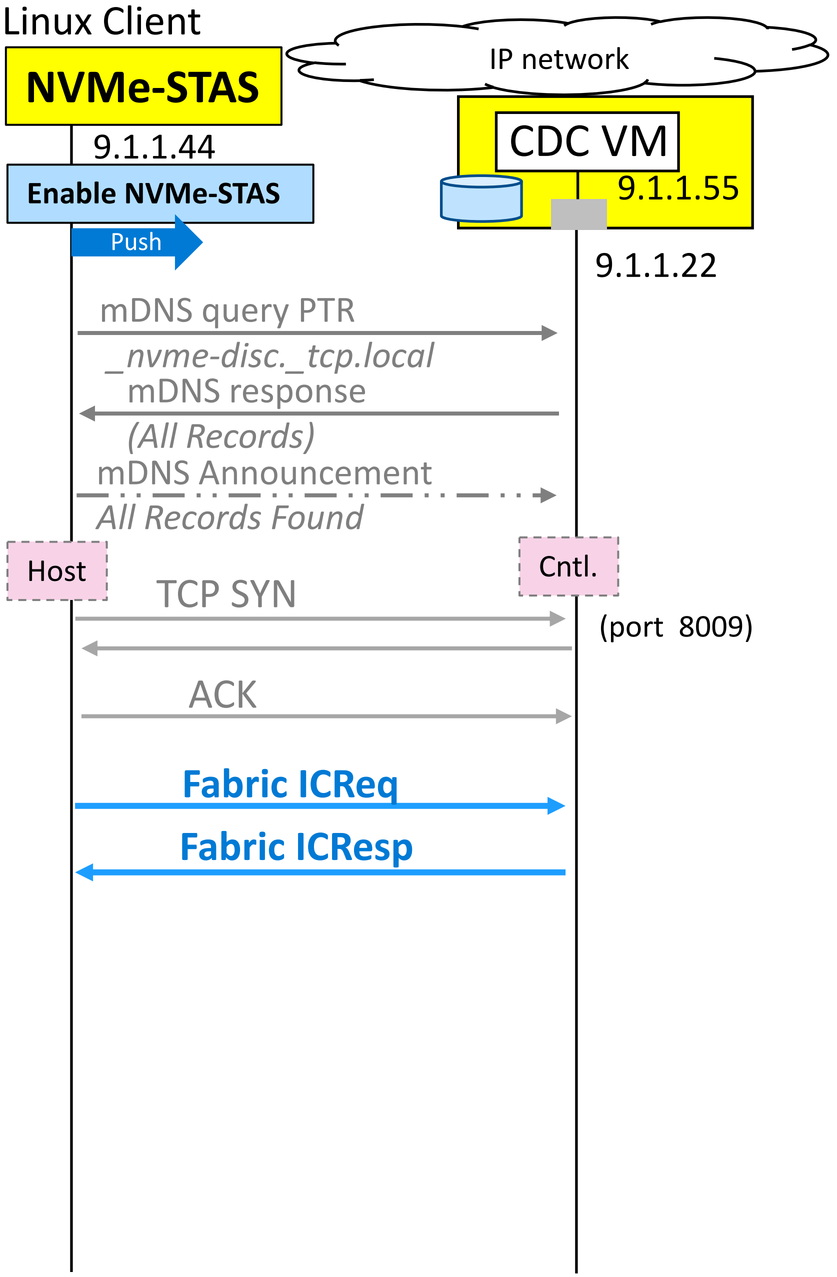


mDNS Announcement

```
⌄ Answers
  › 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type TXT, class IN
  › 9-1-1-55.local: type A, class IN, addr 9.1.1.55
  ⌄ 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type SRV, class IN, priority 0, weight 0,
      Service: 9-1-1-55:08/27/22:01:53:05
      Protocol: _nvme-disc
      Name: _tcp.local
      Type: SRV (Server Selection) (33)
      .000 0000 0000 0001 = Class: IN (0x0001)
      0... .... .... .... = Cache flush: False
      Time to live: 120 (2 minutes)
      Data length: 8
      Priority: 0
      Weight: 0
      Port: 8009
      Target: 9-1-1-55.local
```


Kamal Bakshi, Cisco

Host Initiates NVMe Connection to CDC



ICReq

```
NVM Express Fabrics TCP Discovery Controller
    [Cmd Qid: 0 (AQ)]
    Pdu Type: ICReq (0)
    Pdu Specific Flags: 0x00 Non-Kickstart discovery NVMe/TCP connection
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 128
    Pdu Data Offset: 0
    Packet Length: 128
    ICReq
      Pdu Version Format: 0
      Host Pdu data alignment: 0
      Digest Types Enabled: 0
      Maximum r2ts per request: 0
```

Maximum Number of Outstanding R2T (MAXR2T): Specifies the maximum number of outstanding R2T PDUs for a command at any point in time on the connection. This is a 0's based value.

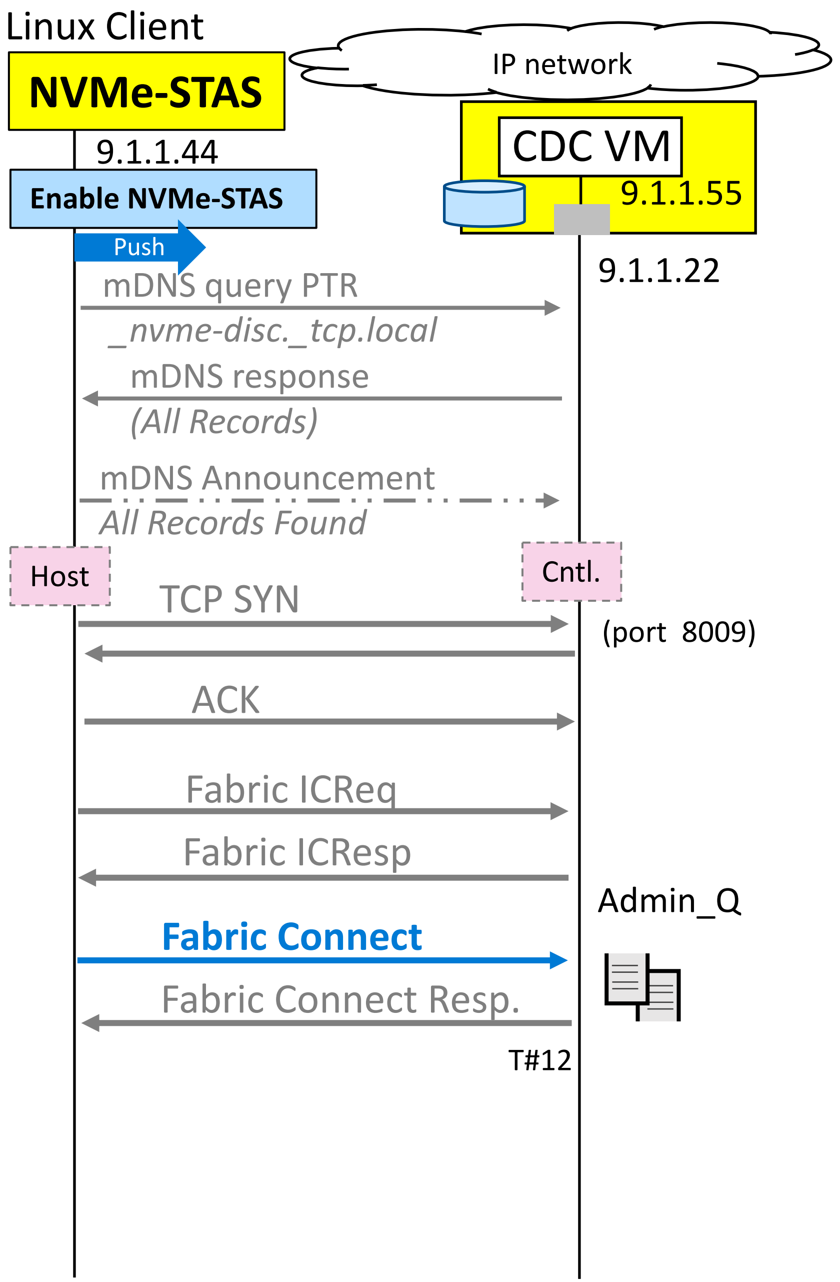
ICResp

```
NVM Express Fabrics TCP Discovery Controller
    [Cmd Qid: 0 (AQ)]
    Pdu Type: ICResp (1)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 128
    Pdu Data Offset: 0
    Packet Length: 128
    ICResp
      Pdu Version Format: 0
      Controller Pdu data alignment: 0
      Digest types enabled: 0
      Maximum data capsules per r2t supported: 1048576
```

Maximum Host to Controller Data length (MAXH2CDATA): Specifies the maximum number of PDU-Data bytes per H2CData PDU in bytes. This value is a multiple of dwords and should be no less than 4,096.

Kamal Bakshi, Cisco

Create Admin Queue at CDC



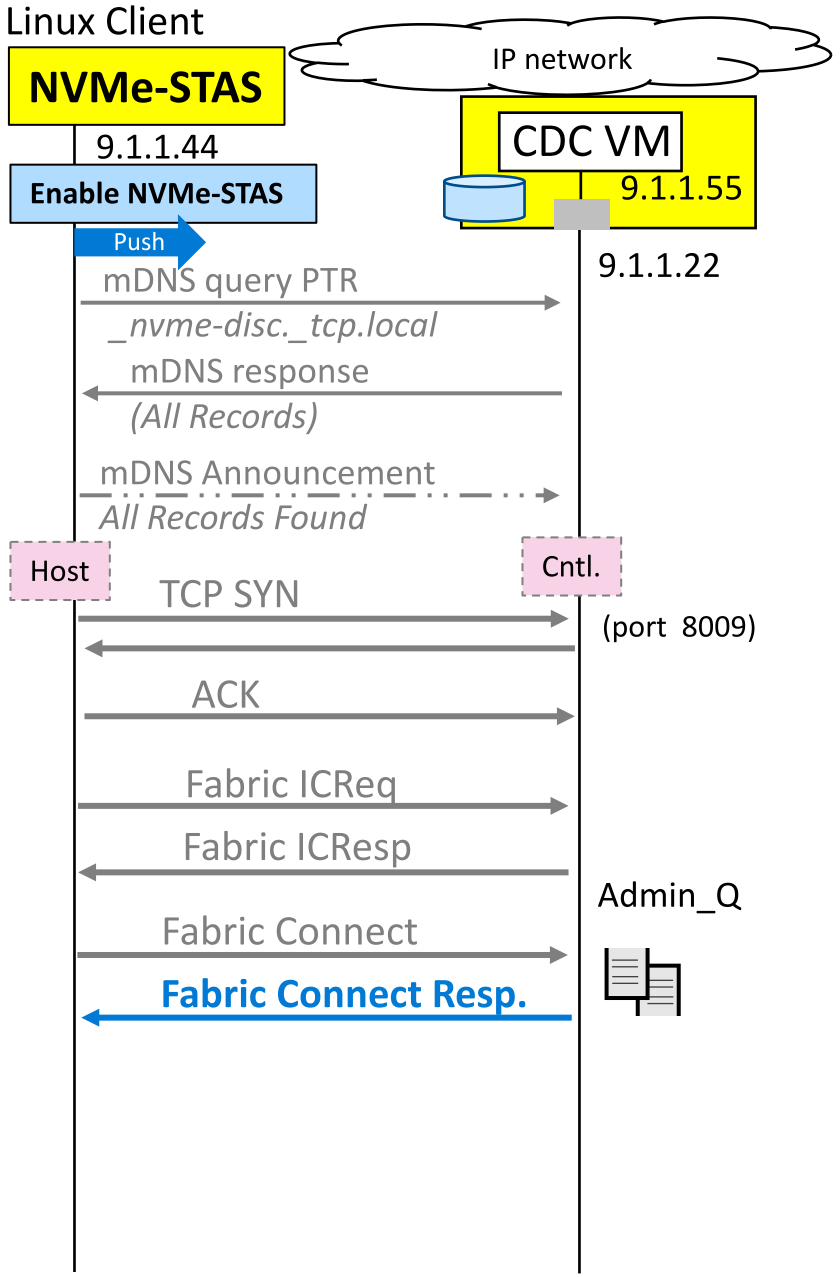

```
◢ NVM Express Fabrics TCP Discovery Controller, Fabrics Type: Connect (0x01) Cmd ID: 0x1000
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 72
    Packet Length: 1096
  ◢ NVM Express Fabrics (Cmd)
      Opcode: 0x7f Fabric Cmd

      Reserved: 0x40
      Command ID: 0x1000
      Fabric Cmd Type: Connect (0x01)
      Reserved: 00000000000000000000000000000000000000
    ▷ SGL1
      Record Format: 0x0000
      Queue ID: 0x0000
      SQ Size: 0x001f
    ▷ Connect Attributes: 0x00
      Reserved: 00
      Keep Alive Timeout: 30000ms
      Reserved: 000000000000000000000000
  ◢ Data
      Host Identifier: aecb70db-d20d-45ae-b91f-7405b932ae19
      Controller ID: 0xffff
      Reserved: 00000000000000000000000000000000000000000000000000000000000000000000000…
      Subsystem NQN: nqn.1988-11.com.dell:SFSS:9:20220824223058e8
      Host NQN: nqn.2014-08.org.nvmexpress:uuid:77b1468d-39f7-483b-b4e2-de789e5048f9
      Reserved: 00000000000000000000000000000000000000000000000000000000000000000000000…
```

Queue ID = 0 (Admin Queue)
Queue Size = 31

Keep Alive Time Out (KATO)
(30sec)

Kamal Bakshi, Cisco

Create Admin Queue at CDC

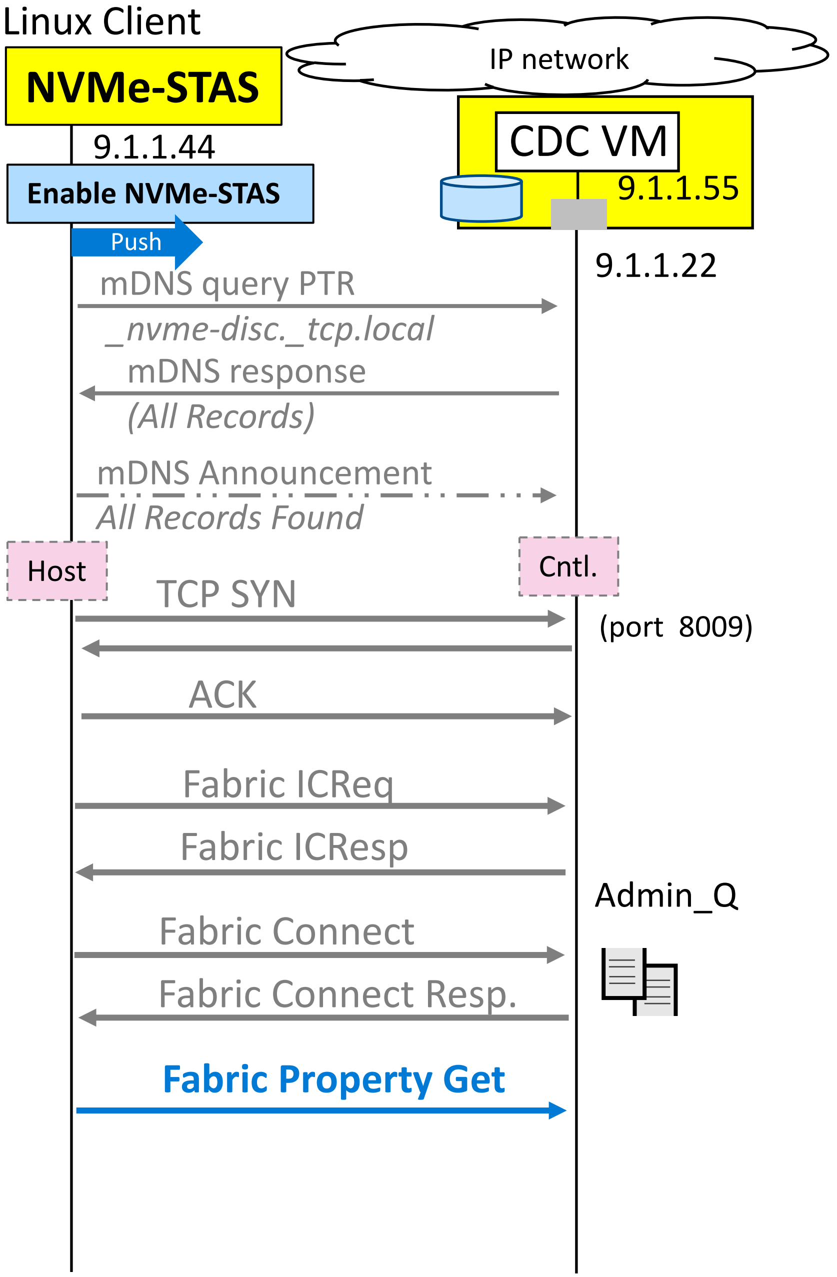



```
◢ NVM Express Fabrics TCP Discovery Controller, Cqe Fabrics Cmd: Connect (0x01) Cmd ID: 0x1000
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
  ◢ Cqe (For Cmd: Connect)

      [Cmd Latency: 15.413 ms]
      Controller ID: 0x0402
      Authentication Required: 0x0000
      Reserved: 00000000
      SQ Head Pointer: 0x0000
      Reserved: 0x0000
      Command ID: 0x1000
      0000 0000 0000 000. = Status: 0x0000
      .... .... .... ...0 = Reserved: 0x0
```

Controller ID

Kamal Bakshi, Cisco

Gets Properties of the CDC Controller

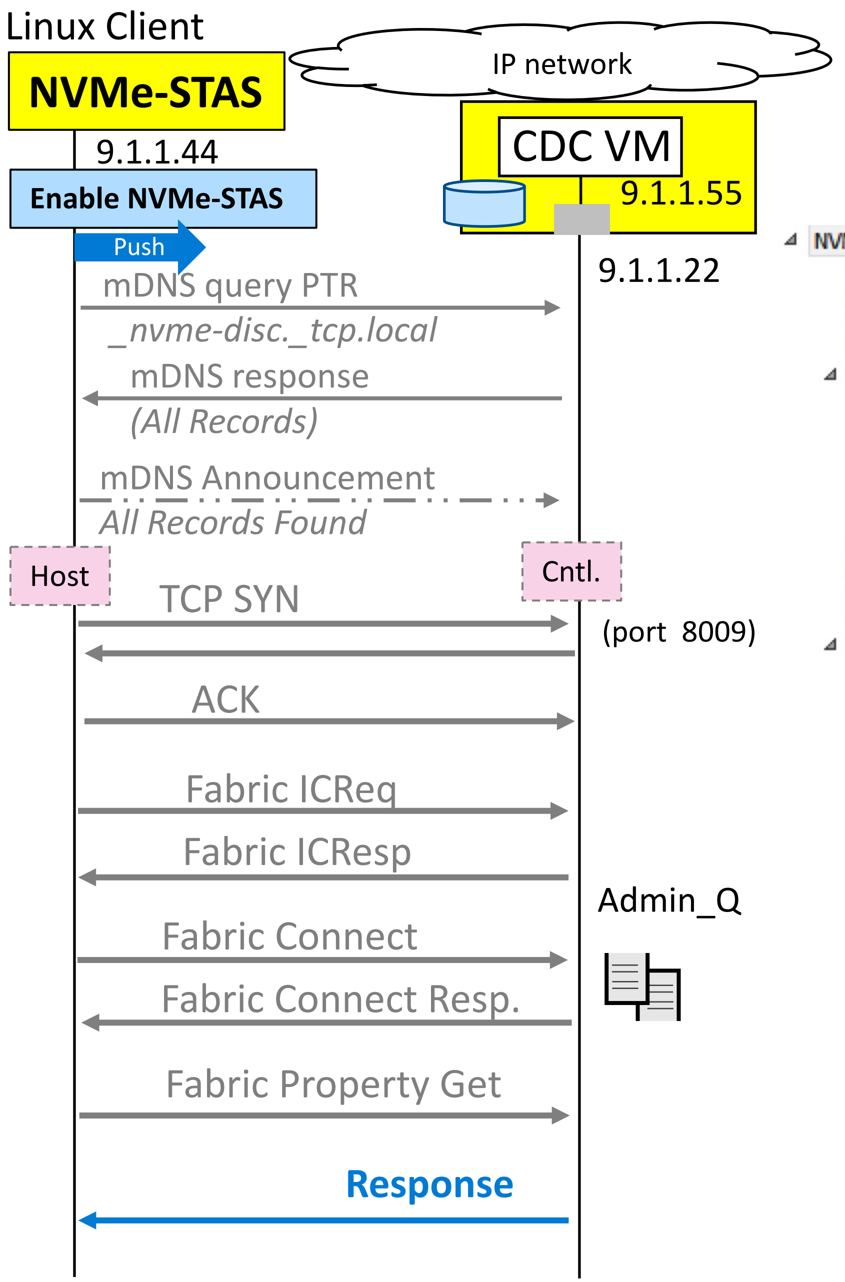



```
NVM Express Fabrics TCP Discovery Controller, Fabrics Type: Property Get (0x04) Cmd ID: 0x1009
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
  ◢ NVM Express Fabrics (Cmd)
       Opcode: 0x7f Fabric Cmd

       Reserved: 0x40
       Command ID: 0x1009
       Fabric Cmd Type: Property Get (0x04)
       Reserved: 000000000000000000000000000000000000000000000000000000000000000000005a
       0000 0... = Reserved: 0x00
       .... .001 = Property Size: 0x18 bytes
       Reserved: 000000
       Offset: 0x00000000Controller Capabilities
       Reserved: 00000000000000000000000000000000
```

Property requested “Controller Capabilities”

Kamal Bakshi, Cisco

Property Buffer returned

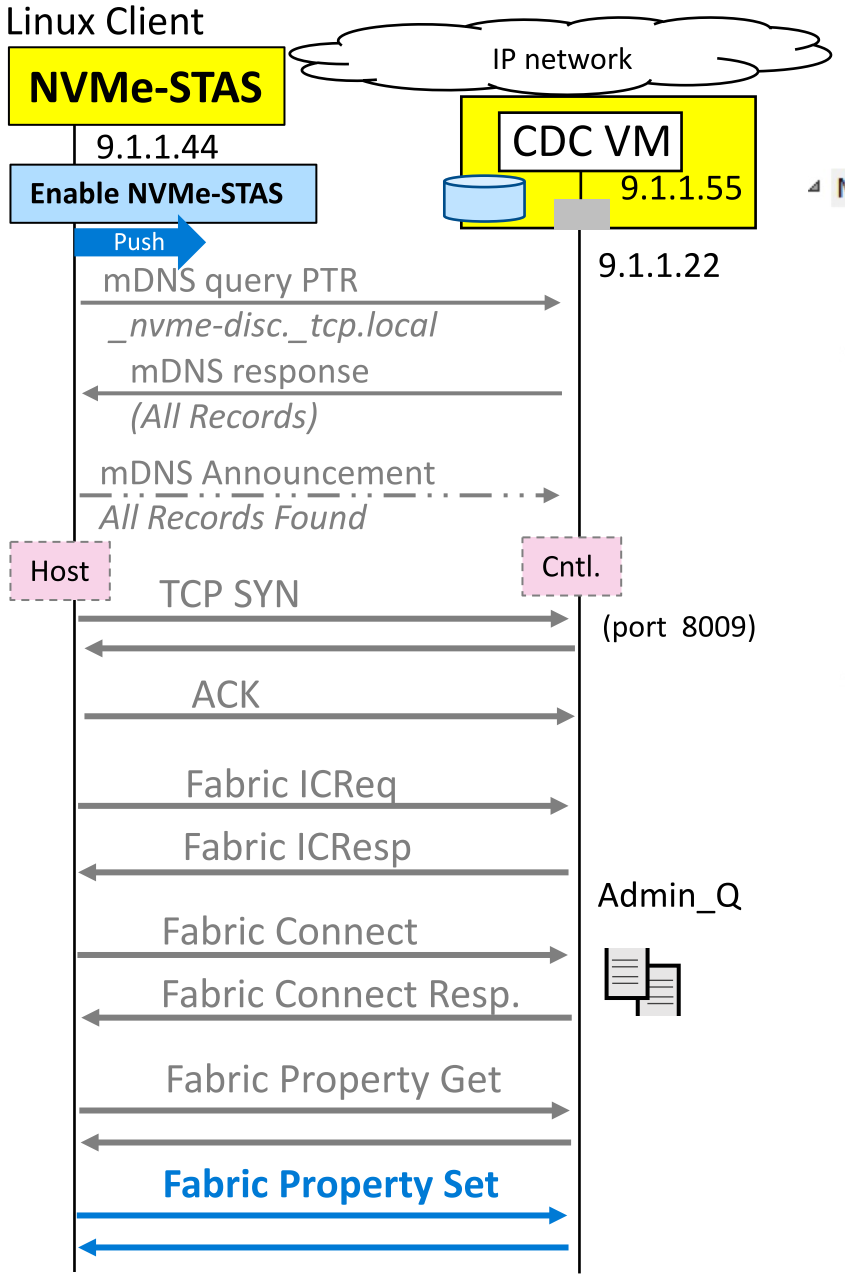



```
NVM Express Fabrics TCP Discovery Controller, Cqe Fabrics Cmd: Property Get (0x04) Cmd ID: 0x1009
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
    Pdu Specific Flags: 0x00
    Pdu Specific Flags: 0x00
        .... ...0 = PDU Header Digest: Not set
        .... ..0. = PDU Data Digest: Not set
        .... .0.. = PDU Data Last: Not set
        .... 0... = PDU Data Success: Not set
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
    Cqe (For Cmd: Property Get)

        [Cmd Latency: 0.534 ms]
        Cmd specific Status: 0x000000000101001f
            Continquous Queues Required (CQR): Continguous I/O Queues Requires
            Boot Partition Supoort (BPS):  Boot Partition Not Supported
            Persistent Memory Region Supported (PMRS): Persistent Memory Region NOt Supported
            Controller Memory Buffer Suported (CMBS): Controller Memory Buffer Not Supported

            Timeout: 0x01
            ....: 0x00  Doorbell Stride (DSTRD)   0, [NVM Subsystem Reset Feature Not Supported], [I/O Command Set Supported]
            Memory Page Size: 0x00  Memory Page Size Minimum (MPSMIN)   14  , Memory Page Size Maximum (MPSMAX)   14
        SQ Head Pointer: 0x0000
        Reserved: 0x0000
        Command ID: 0x1009
        0000 0000 0000 000. = Status: 0x0000
        .... .... .... ...0 = Reserved: 0x0
```

Controller Capabilities

Kamal Bakshi, Cisco

Sets CDC Cntl. Configuration



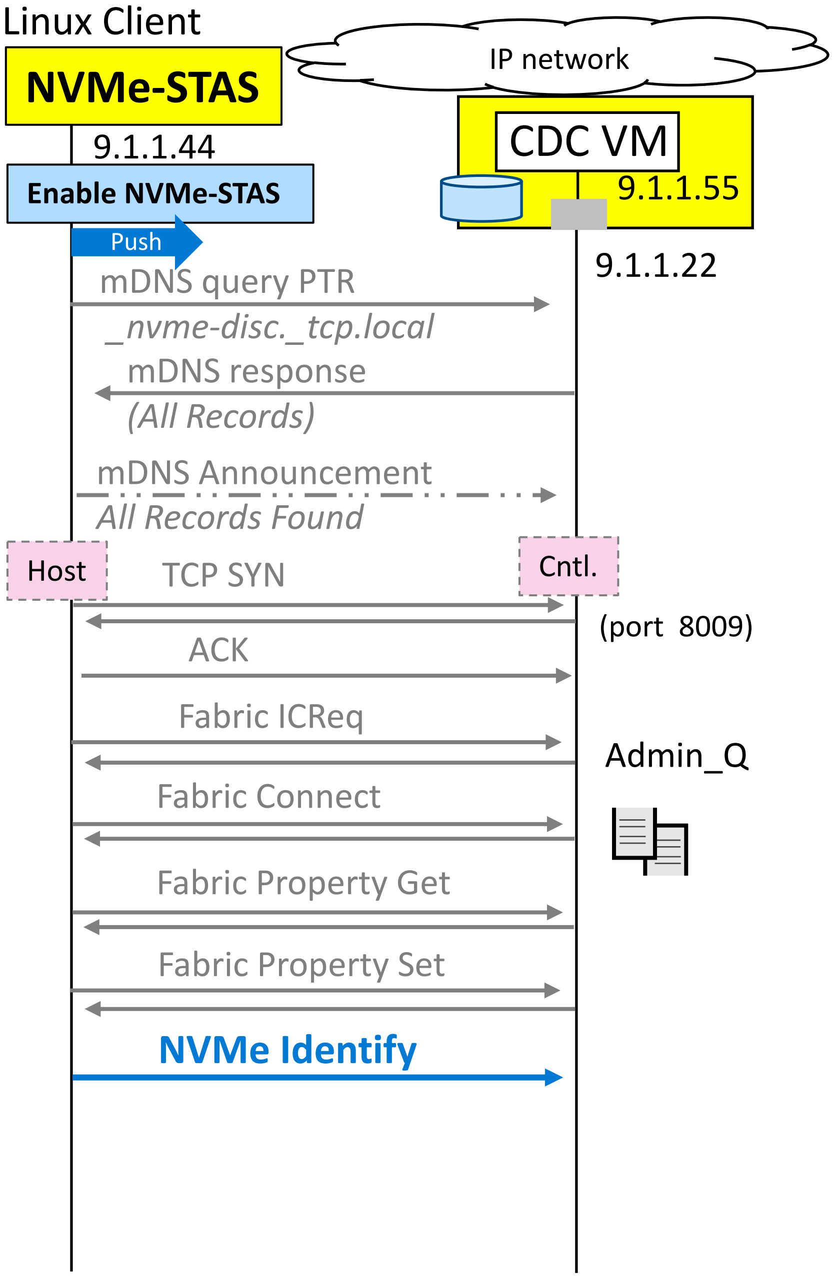

```
◢ NVM Express Fabrics TCP Discovery Controller, Fabrics Type: Property Set (0x00) Cmd ID: 0x100a
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
  ◢ Pdu Specific Flags: 0x00
      .... ...0 = PDU Header Digest: Not set
      .... ..0. = PDU Data Digest: Not set
      .... .0.. = PDU Data Last: Not set
      .... 0... = PDU Data Success: Not set
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
  ◢ NVM Express Fabrics (Cmd)
      Opcode: 0x7f Fabric Cmd

      Reserved: 0x40
      Command ID: 0x100a
      Fabric Cmd Type: Property Set (0x00)
      Reserved: 000000000000000000000000000000000000000000000000000000000000000000005a
      0000 0... = Reserved: 0x00
      .... .000 = Property Size: 0x04 bytes
      Reserved: 000000
      Offset: 0x00000014 Controller Configuration
      Value: 0x00460001
      Reserved: 0x00000000
      Reserved: 0000000000000000
```

Value set: 0x00460001

```
⌄ Controller Configuration: 0x00460001
    .... .... .... .... .... .... .... ...1 = Enable: 0x1
    .... .... .... .... .... .... .... 000. = Reserved: 0x0
    .... .... .... .... .... .... .000 .... = IO Command Set Selected: NVM IO Command Set (0x0)
    .... .... .... .... .... .000 0... .... = Memory Page Size: 0x0 (4096 bytes)
    .... .... .... .... ..00 0... .... .... = Arbitration Mechanism Selected: Round Robin (0x0)
    .... .... .... .... 00.. .... .... .... = Shutdown Notification: No Shutdown (0x0)
    .... .... .... 0110 .... .... .... .... = IO Submission Queue Entry Size: 0x6 (64 bytes)
    .... .... 0100 .... .... .... .... .... = IO Completion Queue Entry Size: 0x4 (16 bytes)
    0000 0000 .... .... .... .... .... .... = Reserved: 0x00
```


Kamal Bakshi, Cisco

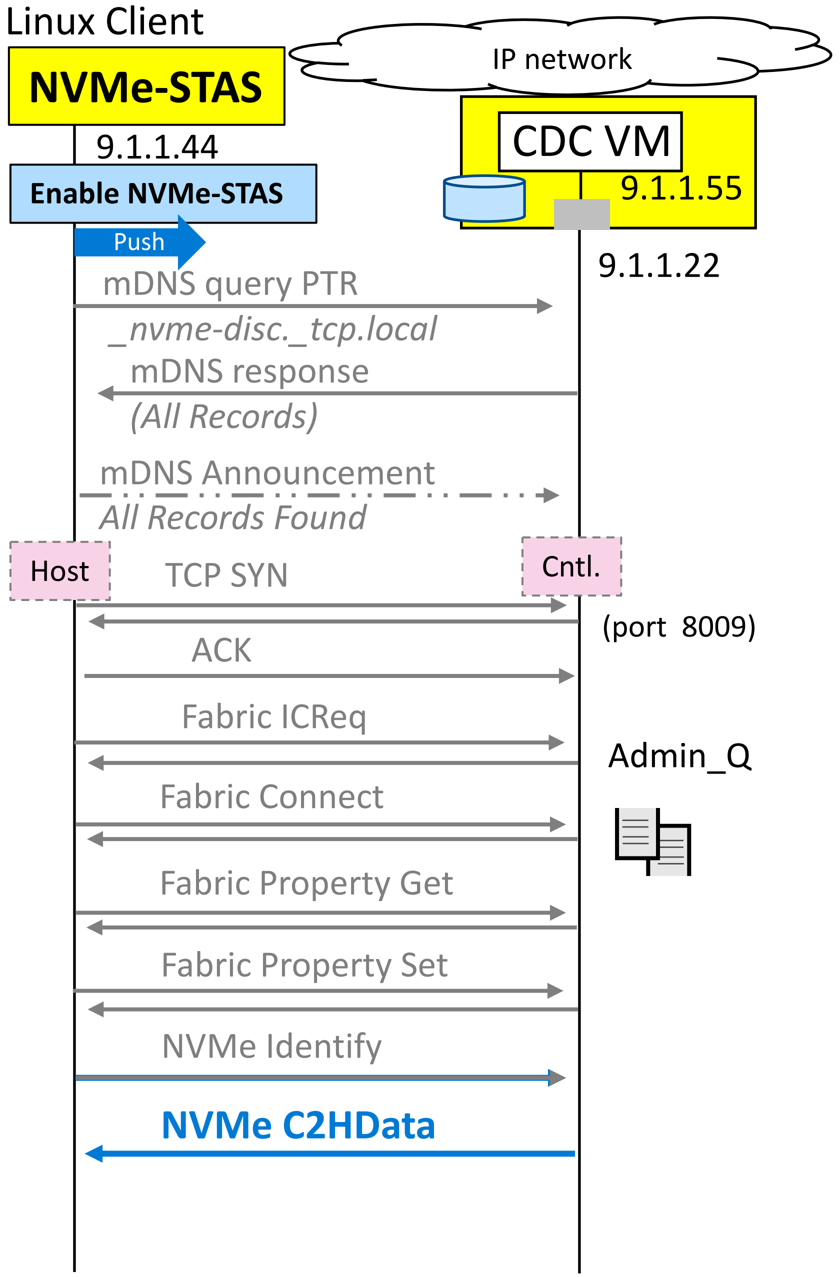
Identify CDC Controller CNS-01




```
◢ NVM Express Fabrics TCP Discovery Controller, NVMe Opcode: Identify (0x06) Cmd ID: 0x2009
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
  ◢ Pdu Specific Flags: 0x00
      .... ...0 = PDU Header Digest: Not set
      .... ..0. = PDU Data Digest: Not set
      .... .0.. = PDU Data Last: Not set
      .... 0... = PDU Data Success: Not set
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
◢ NVM Express (Cmd)
    Opcode: 0x06 Identify
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x2009
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  ◢ SGL1
      0101 .... = Descriptor Type: 0x5 Transport Data Block
      .... 1010 = Descriptor Sub Type: 0xa Command Data Buffer
      Reserved: 0x0000000000000000
      Length: 4096
      Reserved: 000000
    Controller or Namespace Structure (CNS): 0x0001
    Reserved: 0000
    Controller Identifier (CNTID): 00000000
```


Kamal Bakshi, Cisco

Identify Buffer Returned



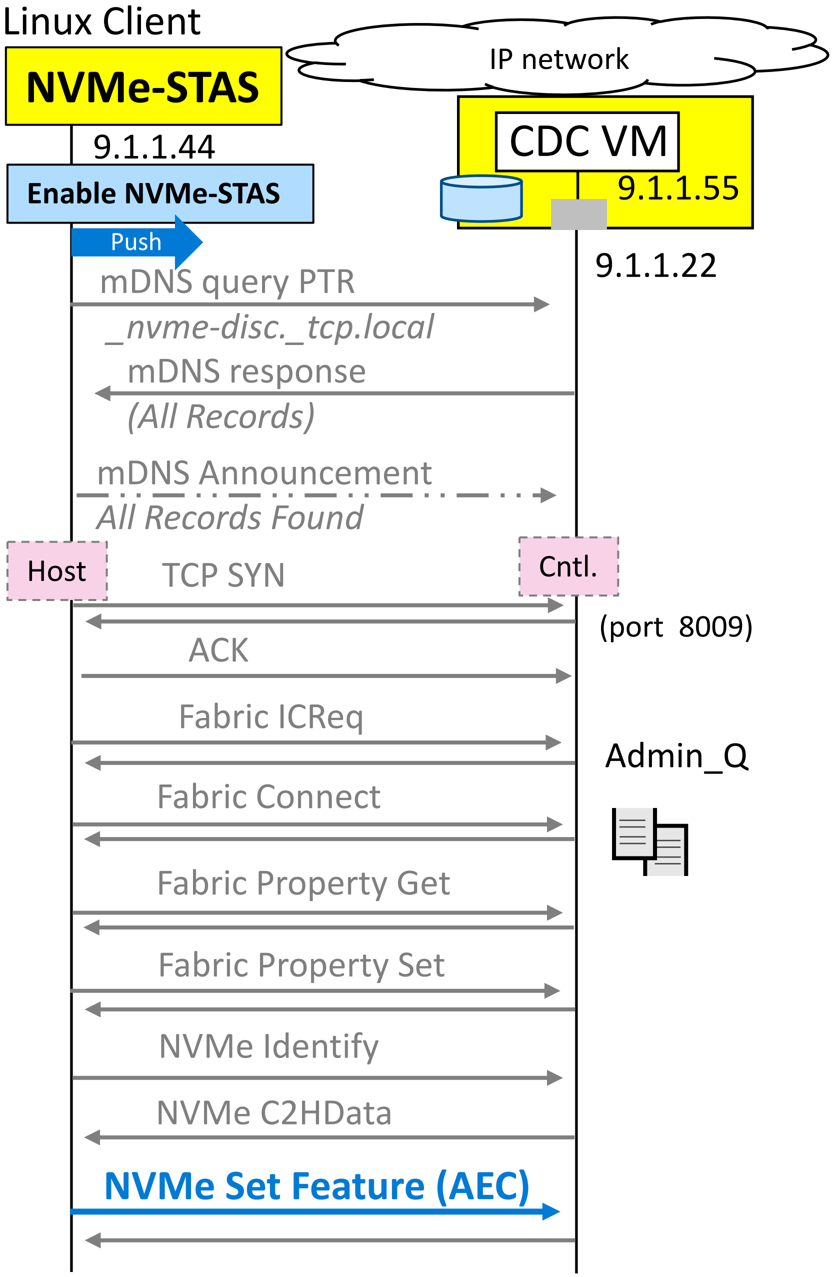

```
◢ NVM Express Fabrics TCP Discovery Controller, C2HData Opcode: Identify (0x06), Cmd ID: 0x2009, Len: 4096
    [Cmd Qid: 0 (AQ)]
    Pdu Type: C2HData (7)
    Pdu Specific Flags: 0x0c
  ▷ Pdu Specific Flags: 0x0c, PDU Data Last, PDU Data Success
    Pdu Header Length: 24
    Pdu Data Offset: 24
    Packet Length: 4120
  ◢ NVMe/TCP Data PDU
      Command ID: 0x2009
      Transfer Tag: 0x0000
      Data Offset: 0
      Data Length: 4096
      Reserved: 00000000
```

Controller Identity Details

```
◢ NVM Express
    PCI Vendor ID (VID): 0x0000
    PCI Subsystem Vendor ID (SSVID): 0x0000
    Serial Number (SN): 20220824223058e8
    Model Number (MN): DELL-SFSS
    Maximum Data Transfer Size (MDTS): 0x00
    Version (VER): 0x00020000
    Optional Asynchronous Events Supported (OAES): 0x80000000
    Optional Admin Command Support (OACS): 0x0000[]
    Abort Command Limit (ACL): 0x00
    Asynchronous Event Request Limit (AERL): 0x00
    Log Page Attributes: 0x04[Extended log page supported, ]
    Keep Alive Support (KAS): 0x0000
    Submission Queue Entry Size (SQES): 0x00
    Completion Queue Entry Size (CQES): 0x00
    Maximum Outstanding Commands (MAXCMD): 0x001f
    Number of Namespaces (NN): 0x00000000
    Optional NVM Command Support (ONCS): 0x0000
    SGL Support (SGLS): 0x00300002
    NVM Subsystem NVMe Qualified Name (SUBNQN): nqn.1988-11.com.dell:SFSS:9:20220824223058e8
    I/O Queue Command Capsule Supported Size (IOCCSZ): 0x00000000
    I/O Queue Response Capsule Supported Size (IORCSZ): 0x00000000
    In Capsule Data Offset : 0x0000
    Fabrics Controller Attributes  : 0 (NVM subsystem uses a dynamic controller model)
    Maximum SGL Data Block Descriptors: 0
    Optional Fabric Commands Support : 0 (Controller does not support the Disconnect command)
    Discovery Controller Type  : 2  Centralized Discovery controller
    Reserved: 0000000000000000000000000000000000000000000000000000000000000000000…
```


Kamal Bakshi, Cisco

Host Sets the Notification Flag at the Controller



Set AEC

Set Feature

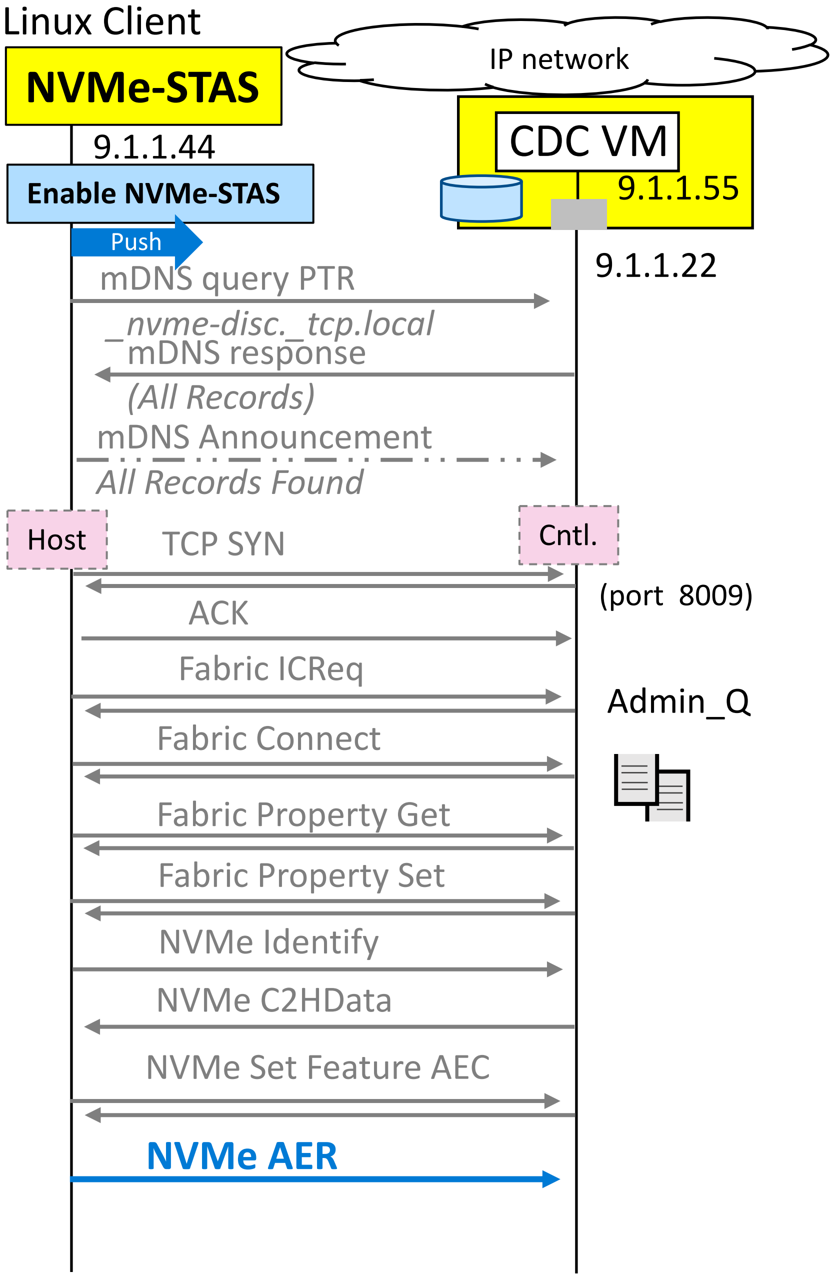
```
00 50 56 bf 37 26 50 6b  4b 4b df 3a 08 00 45 00
00 7c a5 d9 40 00 40 06  80 3e 09 01 01 2c 09 01
01 37 82 9a 1f 49 41 79  01 c9 3d 20 c8 89 80 18
01 f5 d6 94 00 00 01 01  08 0a 27 b3 e7 d8 15 db
c5 0d 04 00 48 00 48 00  00 00 09 40 0b 20 00 00
00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00
00 5a 0b 00 00 00 00 00  00 80 00 00 00 00 00 00
00 00 00 00 00 00 00 00  00 00 fa c7 53 74
```

Asynchronous Event Configuration

This Feature controls the events that trigger an asynchronous event notification to the host. If the condition for an event is true when the corresponding notice is enabled, then an event is sent to the host.

Kamal Bakshi, Cisco

Async. Event Registration at Controller

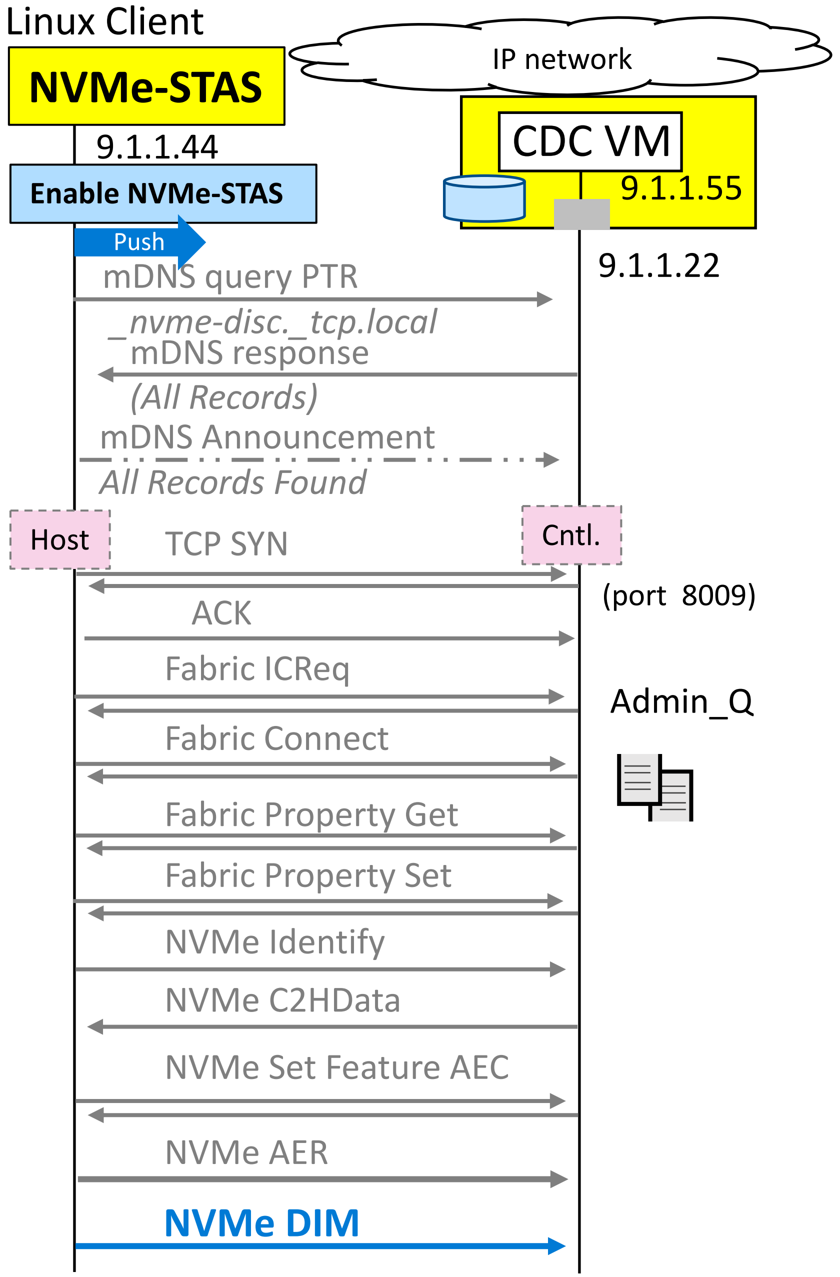


The Asynchronous Event Request (AER) command is submitted by host software to enable the reporting of asynchronous events from the controller. This command has no timeout. The controller posts a completion queue entry for this command when there is an asynchronous event to report to the host.

```
◢ NVM Express (Cmd)
    Opcode: 0x0c Async Event Request
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x001f
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  ◢ SGL1
      0101 .... = Descriptor Type: 0x5 Transport Data Block
      .... 1010 = Descriptor Sub Type: 0xa Command Data Buffer
      Reserved: 0x0000000000000000
      Length: 0
      Reserved: 000000
```


Kamal Bakshi, Cisco

Host Pushes DIM info to the CDC Controller



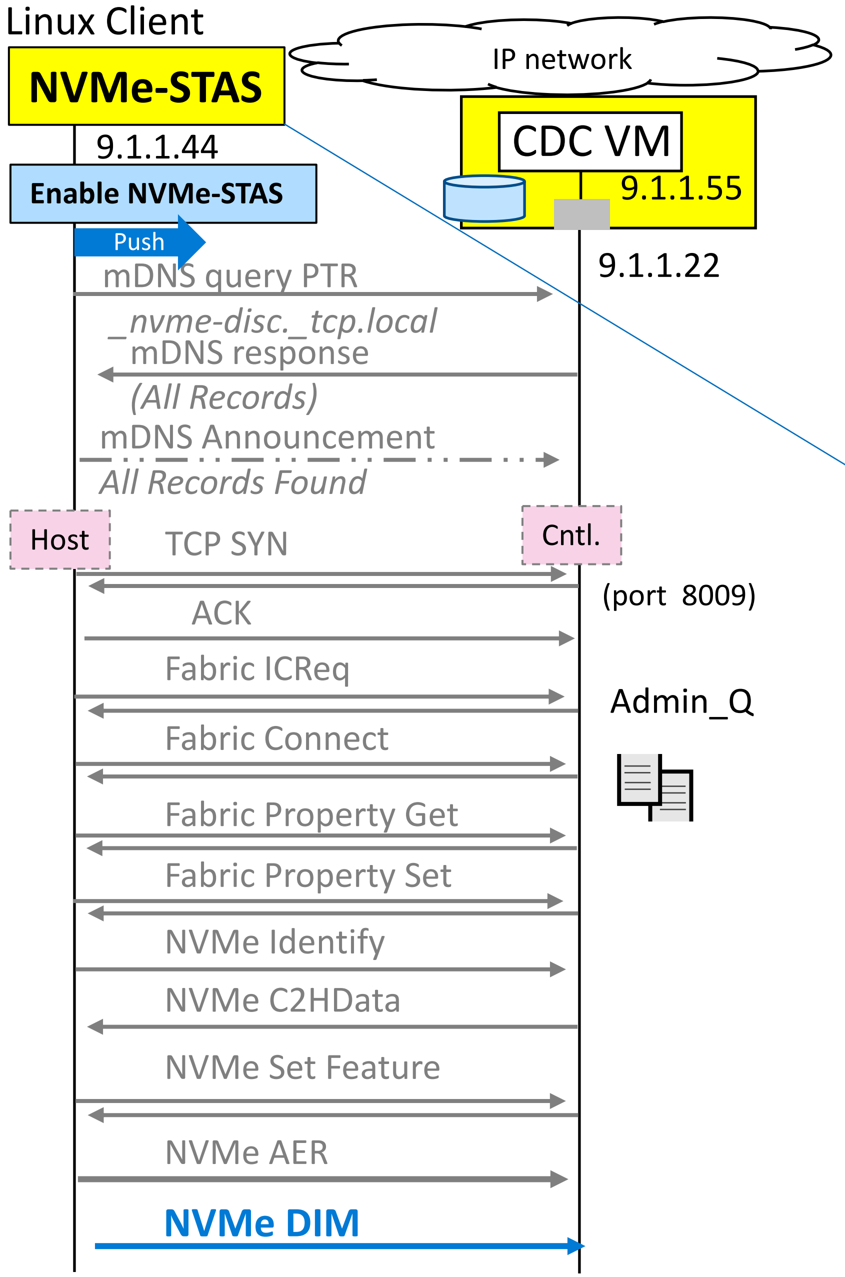

```
▷ NVM Express Fabrics TCP Discovery Controller, NVMe Opcode: Discovery Information Management (0x21) Cr
◢ NVM Express (Cmd)
    Opcode: 0x21 Discovery Information Management
    [Cqe in: 55]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x200c
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  ▷ SGL1
    Task: 0 Register
    Reserved: 0x00000000
◢ NVM Express
    Total Data Length (TDL): 2076
    Reserved: 00000000
    Number of Entries (NUMENT): 1
    Entry Format(ENTFMT): 0x0002 Extended Discovery Information Entry
    Entry Type(ETYPE): 0x0001 Host
    Port Local(PORTLCL): 0x00
    Reserved: 00
    Entry Key Type (EKTYPE): 0x005f TRADDR Based
    Entity Identifier(EID): nqn.2014-08.org.nvmexpress:uuid:77b1468d-39f7-483b-b4e2-de789e5048f9
    Entity Name(ENAME): UCS162  (Host Name)
    Entity Version(EVER): Linux 5.14.21-150400.22-default SLED 15.4  (Host Name & Version)
    Reserved: 00000000000000000000000000000000000000000000000000000000000000000000…
  ◢ Discovery Log Page Entry: 030100000000000000000000000000000000000000000000000000000000000000000
      Transport Type(TRTYPE): 3 [TCP Transport]
      Address Family(ADRFAM): 1 [AF_INET]
      Subsystem Type (SUBTYPE): 0 [Reserved]
      Transport Requirements (TREQ): 0 [Controller requires use of SQ flow control ,Connection over F
      PORT ID: 0
      Controller ID: 0
      Admin Max SQ Size: 0
      Reserved: 0000000000000000000000000000000000000000000
      Transport Service Identifier(TRSVCID):
      Reserved: 0000000000000000000000000000000000000000000000000000000000000000000…
      NVMe Qualified Name(NQN): nqn.2014-08.org.nvmexpress:uuid:77b1468d-39f7-483b-b4e2-de789e5048f9
      Transport Address(TRADDR):
      Transport Specific Address SubType (TSAS): 000000000000000000000000000000000000000000000000000
      Total Entry Length(TEL): 1052
      Number of Extended Attributes(NUMEXAT): 1
      Reserved: 0000
    ▷ Extended Attribute(EXAT): 01001000aecb70dbd20d45aeb91f7405b932ae19[Extended Attribute:  1]
```

Discovery Information

Host NQN

Kamal Bakshi, Cisco

DIM info is stored by the CDC Controller



Host Details is added in the CDC

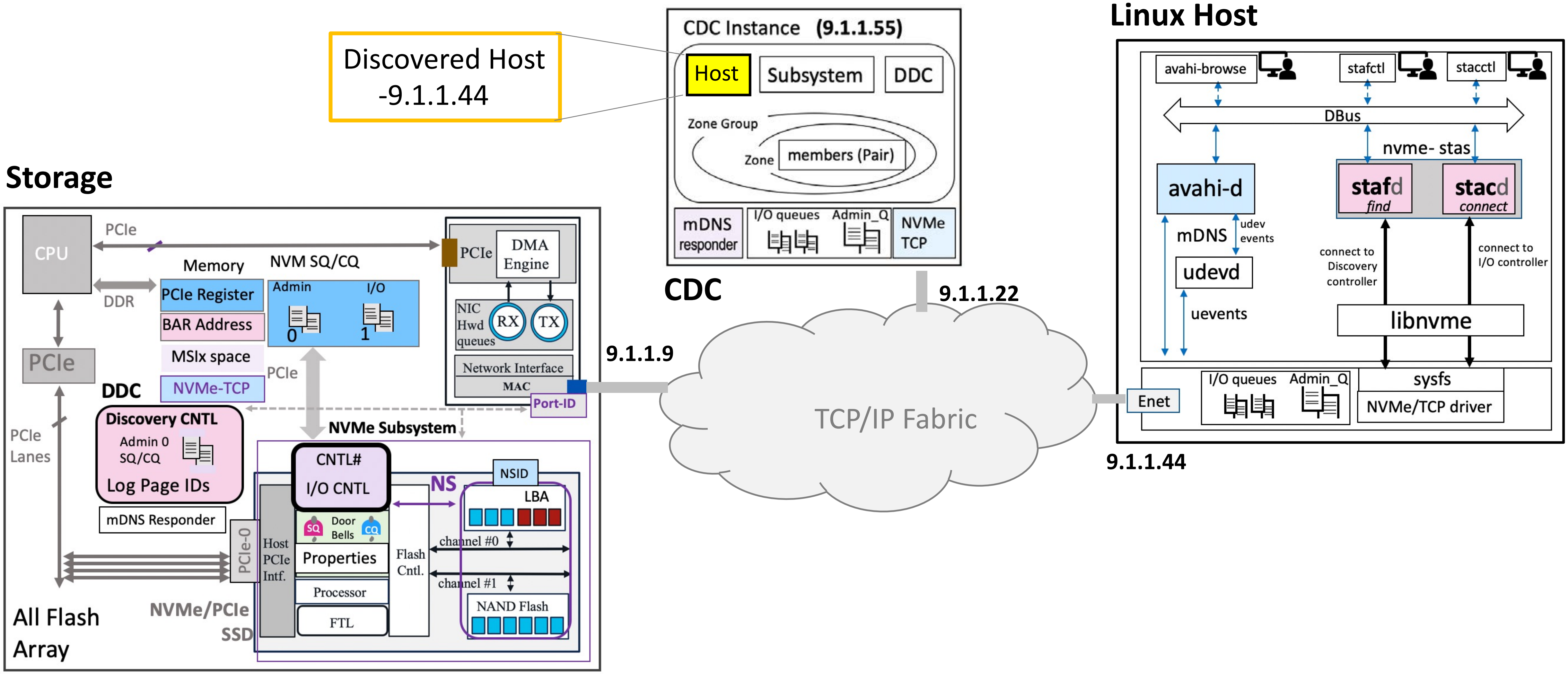
> General Information

Connection Status	Online
Entity Key Type	TRADDR
EName	sles15
EVersion	Linux 5.14.21-150400.22-default SLED 15.4
Host Identifier	aecb70dbd20d45aeb91f7405b932ae19
HostInterface	nqn.2014-08.org.nvmexpress:uuid:77b1468d-39f7-483b-b4e2-de789e5048f9@9.1.1.44:V4::0:57442:TCP
NQN	nqn.2014-08.org.nvmexpress:uuid:77b1468d-39f7-483b-b4e2-de789e5048f9
NodeName	stfs-cdcproxy-deployment-3-1
Registration Type	Explicit
Transport Requirements (TREQ)	Secure channel Not specified
Transport Specific Address Subtype (TSAS)	No Security
Transport Address	9.1.1.44
Transport Address Family	IPV4
Transport Type	TCP

source: Dell SFSS/CDC

Kamal Bakshi, Cisco

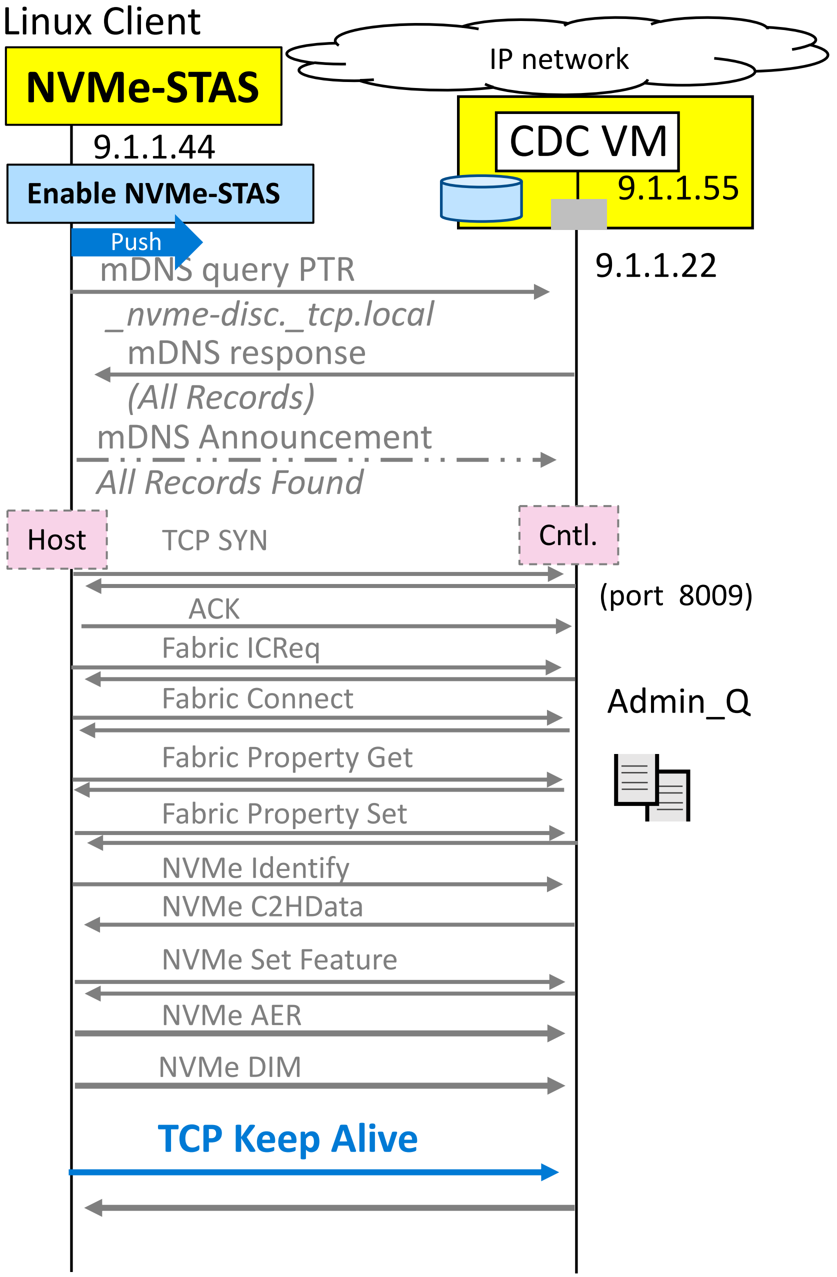
Host Infor in CDC database (NQN, IP, etc.)



 |

Kamal Bakshi, Cisco

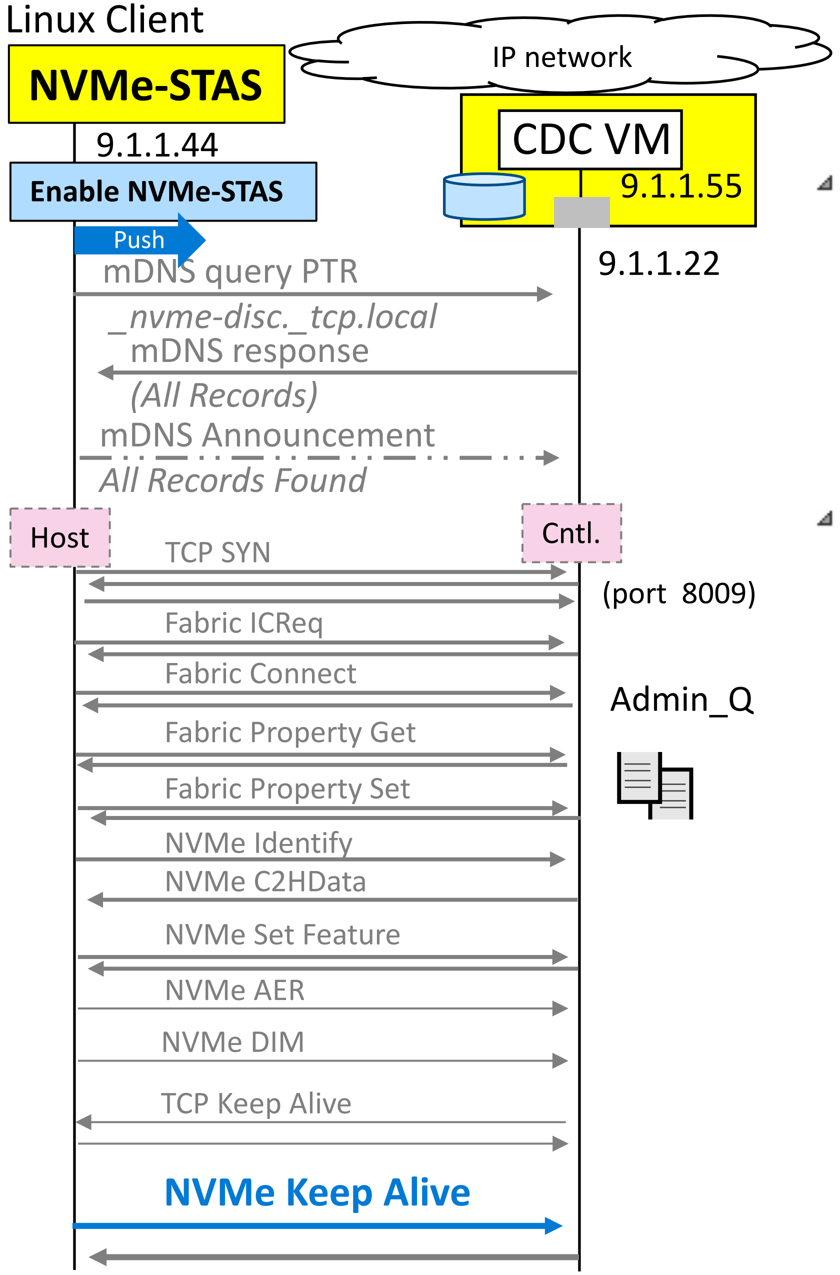
TCP Keep Alive (15 seconds)




```
◢ Transmission Control Protocol, Src Port: 33434, Dst Port: 8009, Seq: 6529, Ack: 11265, Len: 0
    Source Port: 33434
    Destination Port: 8009
    [Stream index: 0]
    [Conversation completeness: Incomplete, DATA (15)]
    [TCP Segment Len: 0]
    Sequence Number: 6529    (relative sequence number)
    Sequence Number (raw): 1098454225
    [Next Sequence Number: 6529    (relative sequence number)]
    Acknowledgment Number: 11265    (relative ack number)
    Acknowledgment number (raw): 1025565537
    1000 .... = Header Length: 32 bytes (8)
  ◢ Flags: 0x010 (ACK)
      000. .... .... = Reserved: Not set
      ...0 .... .... = Nonce: Not set
      .... 0... .... = Congestion Window Reduced (CWR): Not set
      .... .0.. .... = ECN-Echo: Not set
      .... ..0. .... = Urgent: Not set
      .... ...1 .... = Acknowledgment: Set
      .... .... 0... = Push: Not set
      .... .... .0.. = Reset: Not set
      .... .... ..0. = Syn: Not set
      .... .... ...0 = Fin: Not set
      [TCP Flags: ·······A····]
    Window: 501
    [Calculated window size: 64128]
    [Window size scaling factor: 128]
    Checksum: 0x2246 [unverified]
    [Checksum Status: Unverified]
    Urgent Pointer: 0
  ▷ Options: (12 bytes), No-Operation (NOP), No-Operation (NOP), Timestamps
  ▷ [Timestamps]
  ◢ [SEQ/ACK analysis]
      [iRTT: 0.000265000 seconds]
    ◢ [TCP Analysis Flags]
      ◢ [Expert Info (Note/Sequence): ACK to a TCP keep-alive segment]
          [ACK to a TCP keep-alive segment]
          [Severity level: Note]
          [Group: Sequence]
```

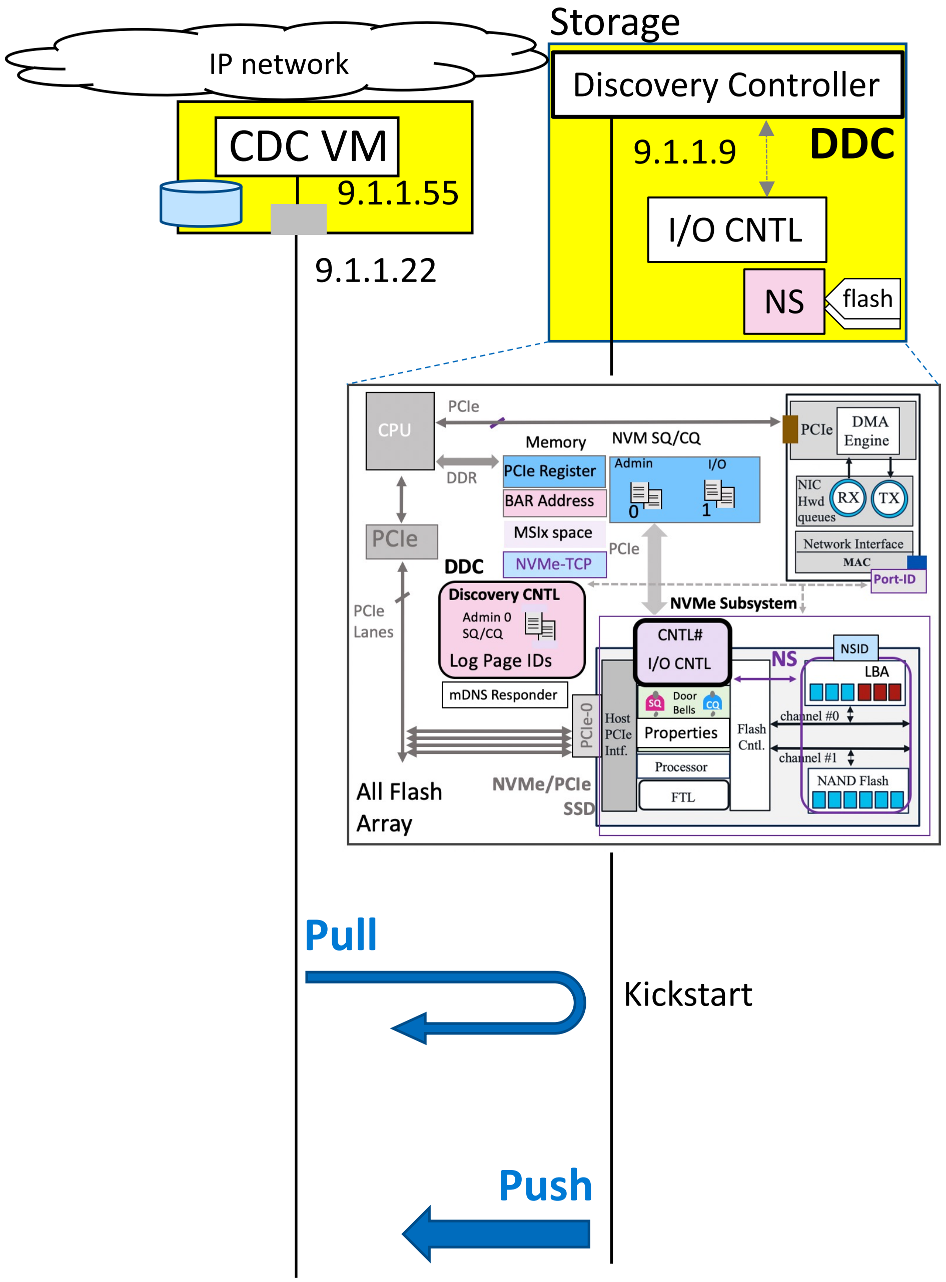

Kamal Bakshi, Cisco

NVMe Keep Alive (30 seconds)




```
◢ NVM Express Fabrics TCP Discovery Controller, NVMe Opcode: Keep Alive (0x18) Cmd ID: 0x2000
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
◢ NVM Express (Cmd)
    Opcode: 0x18 Keep Alive
    [Cqe in: 95]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x2000
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  ◢ SGL1
      0101 .... = Descriptor Type: 0x5 Transport Data Block
      .... 1010 = Descriptor Sub Type: 0xa Command Data Buffer
      Reserved: 0x0000000000000000
      Length: 0
      Reserved: 000000
```


Kamal Bakshi, Cisco



How does Storage Discovers CDC?

DDC Registration -Pull or Push ?

A DDC may determine if a CDC is present by transmitting a query that includes a DNS PTR record with the name in the form of:

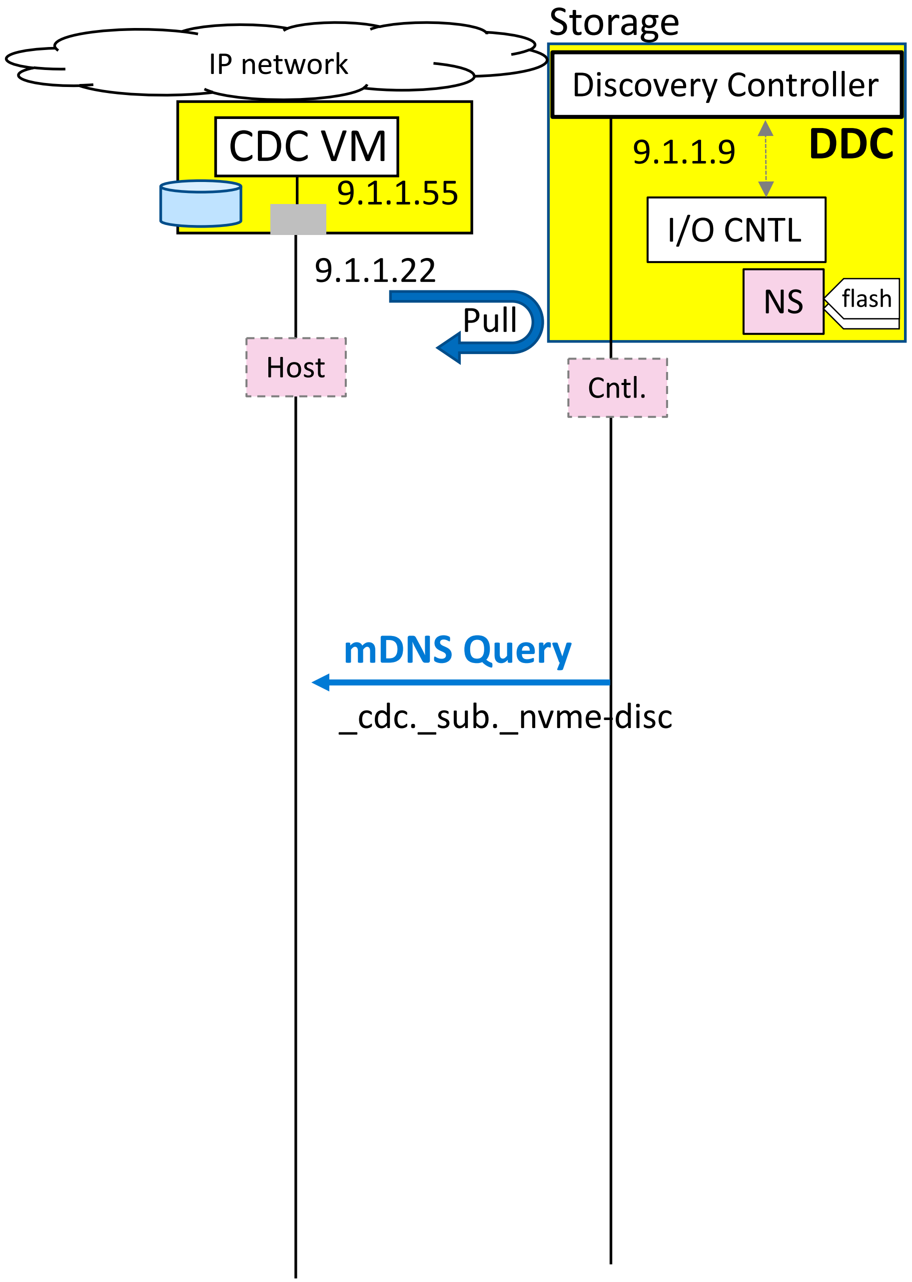
“_cdc._sub._nvme-disc._tcp.local”

Upon reception of an mDNS response that contains a DNS SRV record with the service name set to “_cdc._sub._nvme-disc”, the DDC may

a. Perform push registration with the CDC; or
b. Request a pull registration from the CDC (e.g., using Kickstart Discovery Request PDU (KDReq)

Kamal Bakshi, Cisco

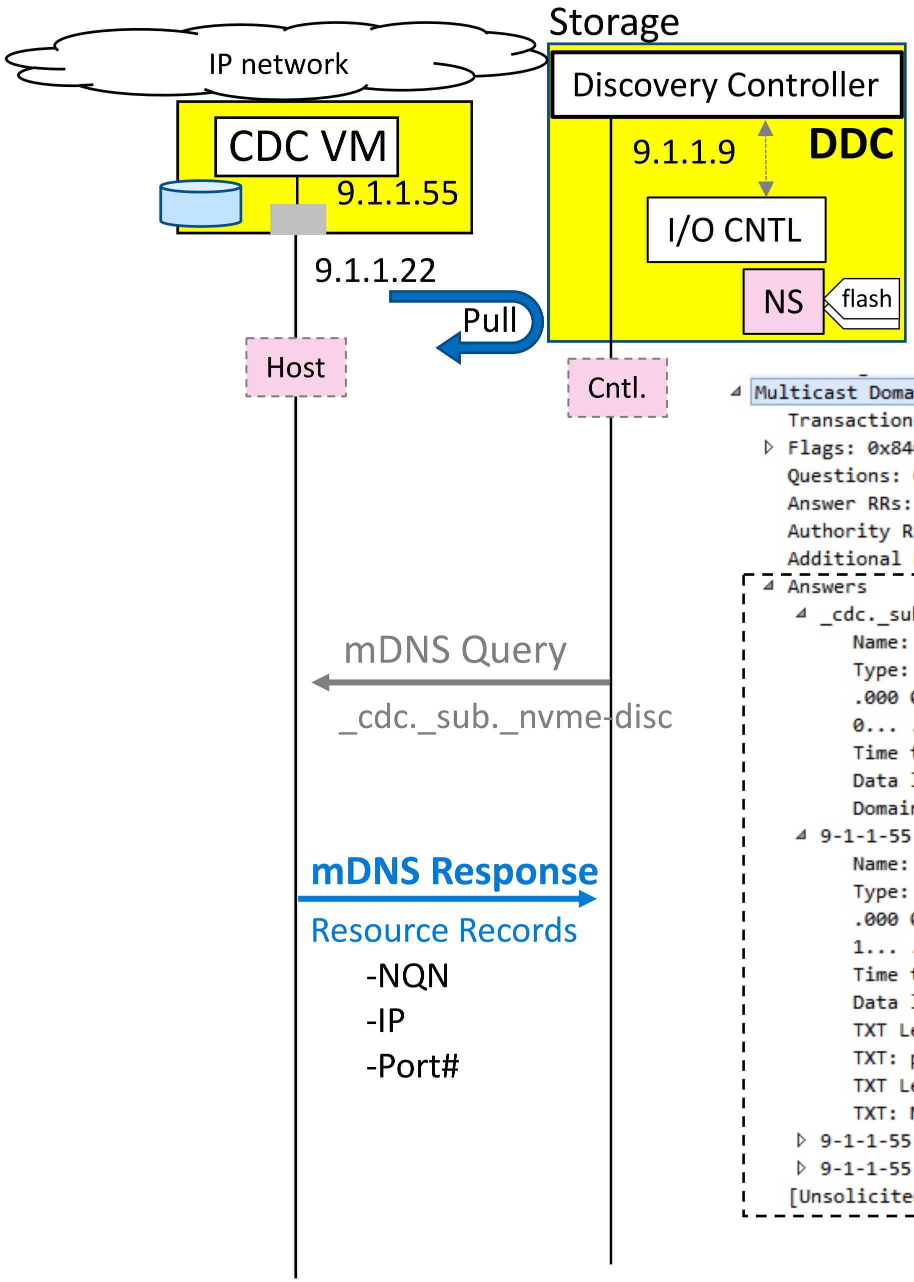
DDC requests pull Registration




```
◢ Multicast Domain Name System (query)
    Transaction ID: 0x0000
  ◢ Flags: 0x0000 Standard query
      0... .... .... .... = Response: Message is a query
      .000 0... .... .... = Opcode: Standard query (0)
      .... ..0. .... .... = Truncated: Message is not truncated
      .... ...0 .... .... = Recursion desired: Don't do query recursively
      .... .... .0.. .... = Z: reserved (0)
      .... .... ...0 .... = Non-authenticated data: Unacceptable
    Questions: 1
    Answer RRs: 0
    Authority RRs: 0
    Additional RRs: 0
  ◢ Queries
    ◢ _cdc._sub._nvme-disc._tcp.local: type PTR, class IN, "QM" question
        Name: _cdc._sub._nvme-disc._tcp.local
        [Name Length: 31]
        [Label Count: 5]
        Type: PTR (domain name PoinTeR) (12)
        .000 0000 0000 0001 = Class: IN (0x0001)
        0... .... .... .... = "QU" question: False
```


Kamal Bakshi, Cisco

CDC provides Resource Records



CDC Resource Records (4)

```
◢ Multicast Domain Name System (response)
    Transaction ID: 0x0000
  ▷ Flags: 0x8400 Standard query response, No error
    Questions: 0
    Answer RRs: 4
    Authority RRs: 0
    Additional RRs: 0
  ◢ Answers
    ◢ _cdc._sub._nvme-disc._tcp.local: type PTR, class IN, 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local
        Name: _cdc._sub._nvme-disc._tcp.local
        Type: PTR (domain name PoinTeR) (12)
        .000 0000 0000 0001 = Class: IN (0x0001)
        0... .... .... .... = Cache flush: False
        Time to live: 4500 (1 hour, 15 minutes)
        Data length: 29
        Domain Name: 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local
    ◢ 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type TXT, class IN, cache flush
        Name: 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local
        Type: TXT (Text strings) (16)
        .000 0000 0000 0001 = Class: IN (0x0001)
        1... .... .... .... = Cache flush: True
        Time to live: 4500 (1 hour, 15 minutes)
        Data length: 55
        TXT Length: 5
        TXT: p=tcp
        TXT Length: 48
        TXT: NQN=nqn.1988-11.com.dell:SFSS:9:20220824223058e8
    ▷ 9-1-1-55:08/27/22:01:53:05._nvme-disc._tcp.local: type SRV, class IN, cache flush, priority 0, weight 0, port 8009,
    ▷ 9-1-1-55.local: type A, class IN, cache flush, addr 9.1.1.55
    [Unsolicited: True]
```

NQN of CDC

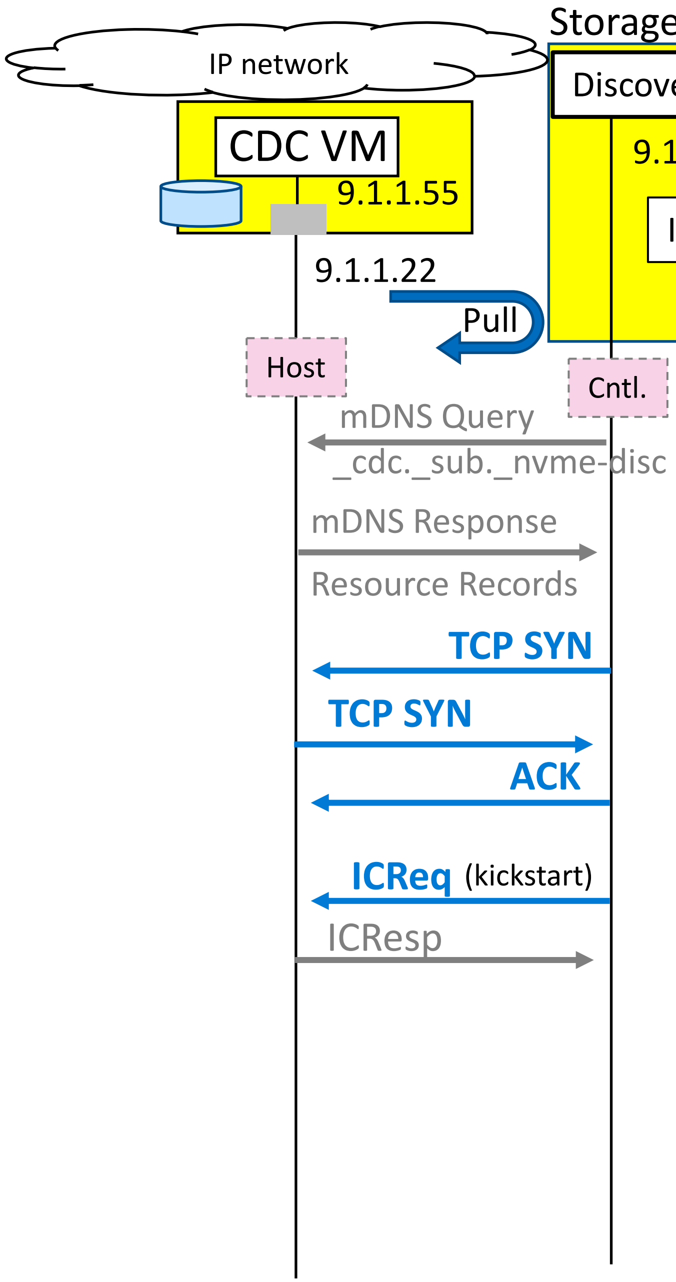
TCP port

IP address

 |

Kamal Bakshi, Cisco

DDC Sets Kickstart Discovery Flag in ICReq



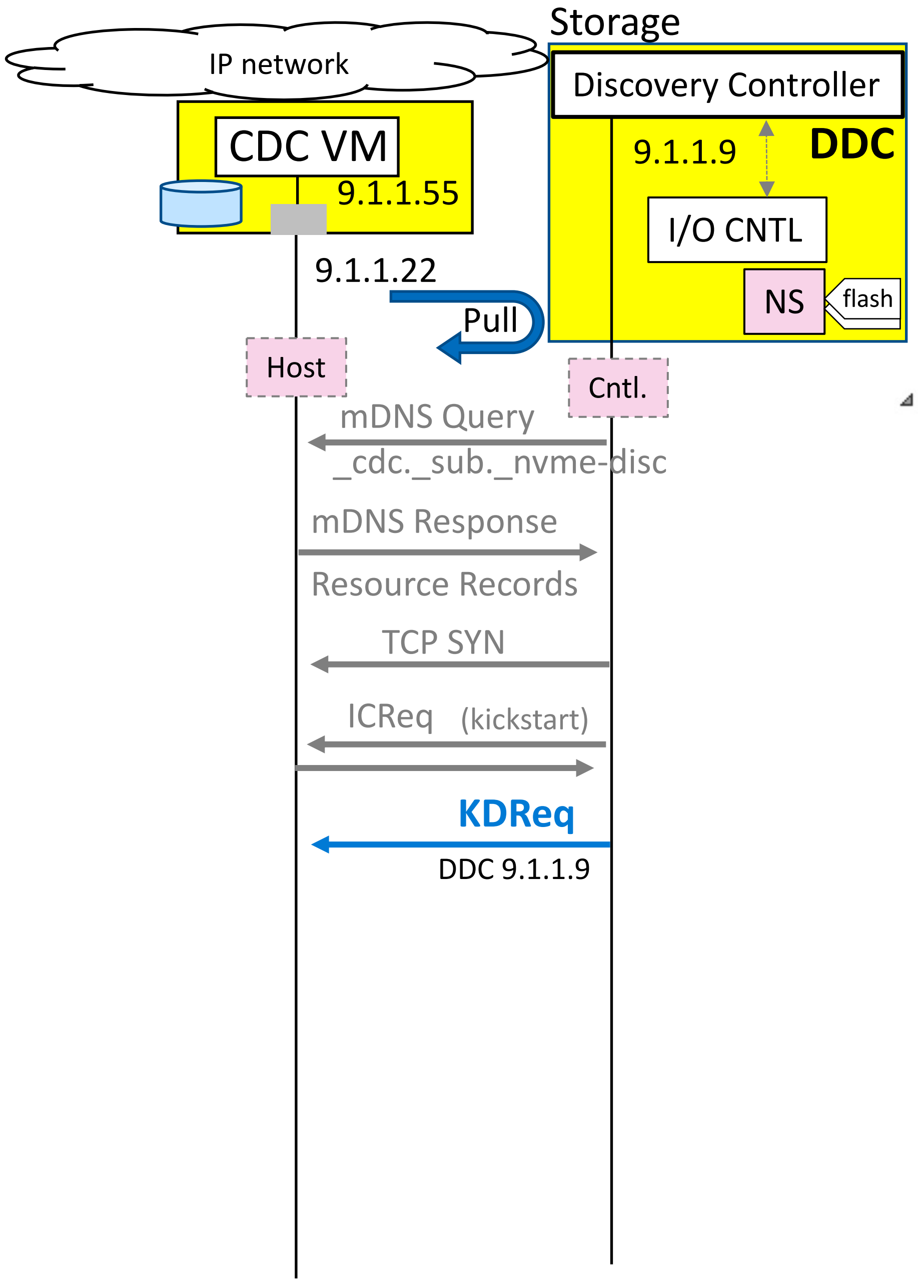
Kickstart Request Flag (asking for Pull)

ICReq

```
◢ NVM Express Fabrics TCP Discovery Controller
    Pdu Type: ICReq (0)
    Pdu Specific Flags: 0x80 Kickstart discovery NVMe/TCP connection
  ◢ Pdu Specific Flags: 0x80
        .... ...0 = PDU Header Digest: Not set
        .... ..0. = PDU Data Digest: Not set
        .... .0.. = PDU Data Last: Not set
        .... 0... = PDU Data Success: Not set
    Pdu Header Length: 128
    Pdu Data Offset: 0
    Packet Length: 128
  ◢ ICReq
      Pdu Version Format: 0
      Host Pdu data alignment: 0
      Digest Types Enabled: 0
      Maximum r2ts per request: 0
```


Kamal Bakshi, Cisco

DDC Requests for Kickstart



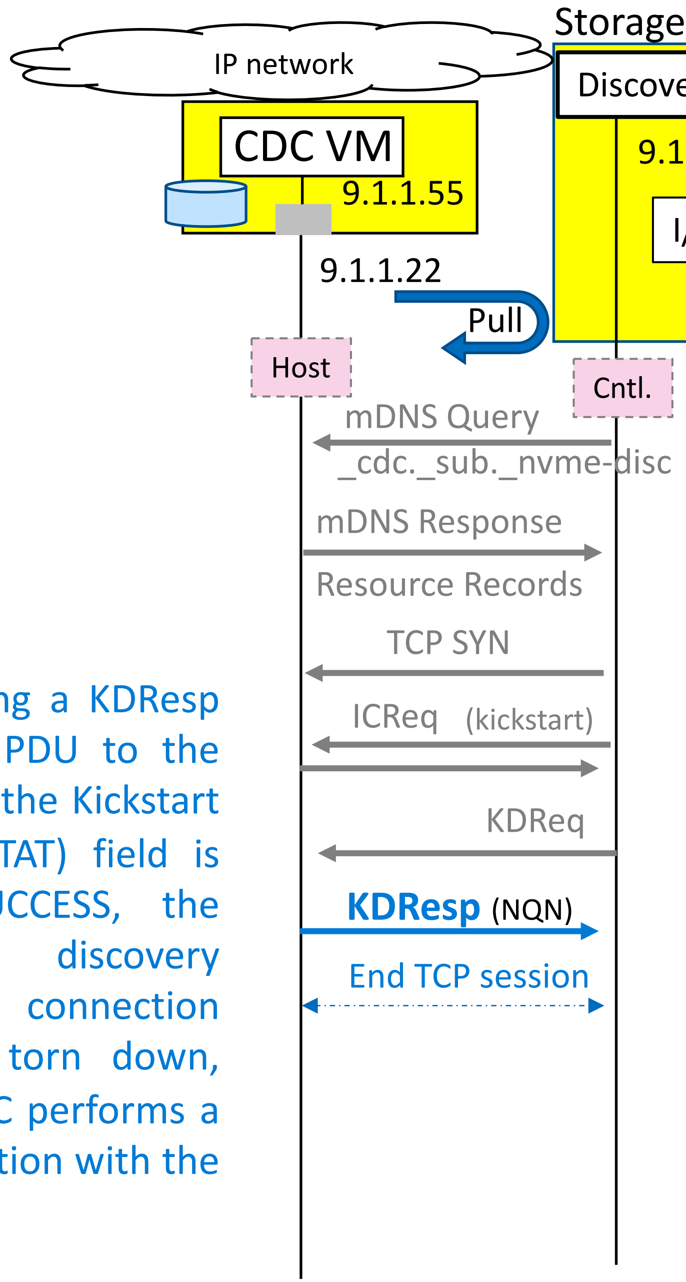
KDReq

```
NVM Express Fabrics TCP Discovery Controller
  Pdu Type: Kickstart Discovery Request (10)
  Pdu Specific Flags: 0x00
  Pdu Specific Flags: 0x00
    .... ...0 = PDU Header Digest: Not set
    .... ..0. = PDU Data Digest: Not set
    .... .0.. = PDU Data Last: Not set
    .... 0... = PDU Data Success: Not set
  Pdu Header Length: 12
  Pdu Data Offset: 12
  Packet Length: 302
  KDReq
    Number of Kickstart Records (NUMKR): 1
    Reserved: 0x0001
    Kickstart Record (KR): 0301383030390000000000000000000000000000
      Transport Type (TRTYPE): 3
      Address Family (ADRFAM): 1
      Transport Service Identifier (TRSVCID): 8009
      Transport Address (TRADDR):: 9.1.1.9
```

Kickstart Record

Kamal Bakshi, Cisco

CDC provides Kickstart Status



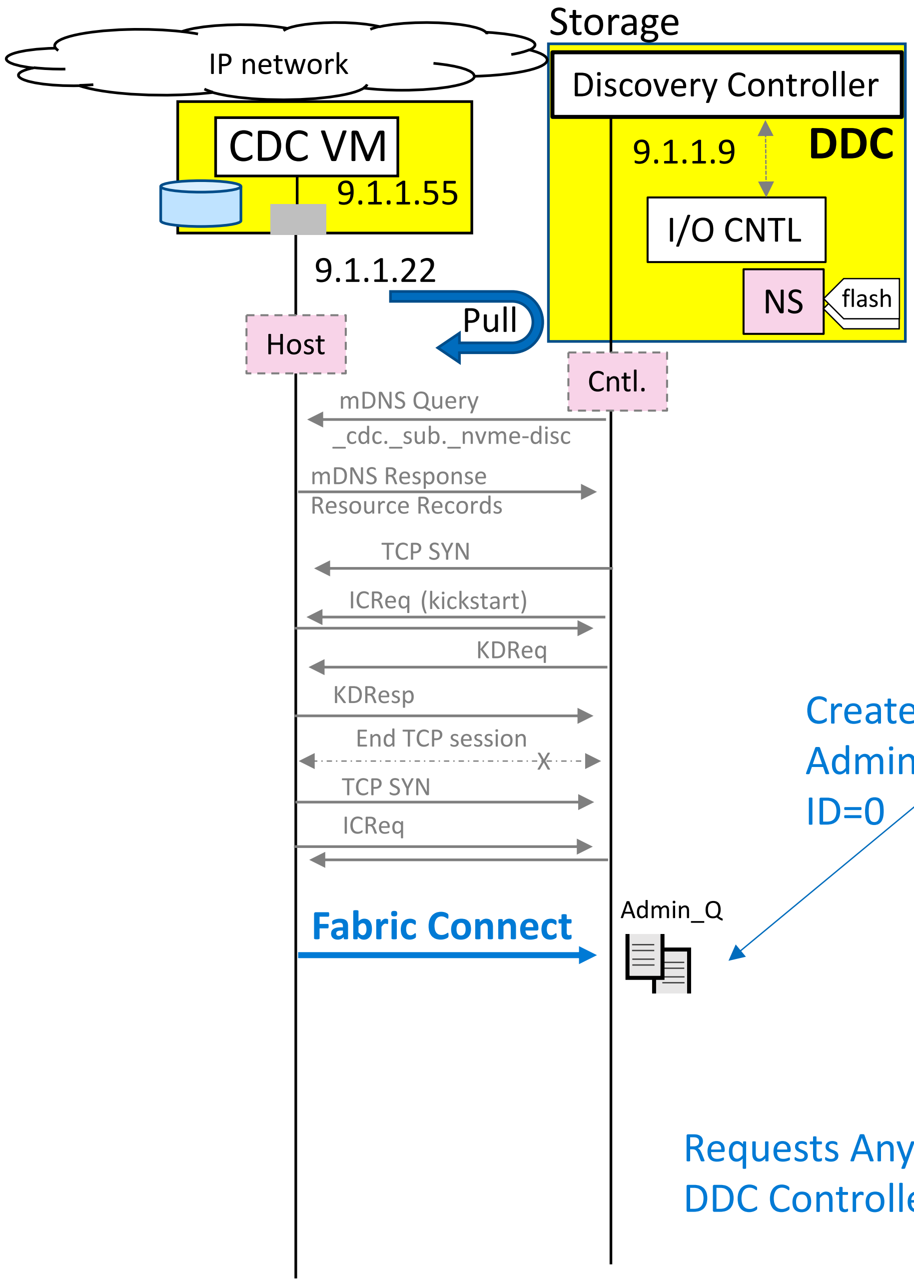
After sending a KDResp NVMe/TCP PDU to the DDC where the Kickstart Status (KSSTAT) field is set to SUCCESS, the kickstart discovery NVMe/TCP connection should be torn down, and the CDC performs a pull registration with the DDC.

```
NVM Express Fabrics TCP Discovery Controller
  Pdu Type: Kickstart Discovery Response (11)
  Pdu Specific Flags: 0x00
  Pdu Specific Flags: 0x00
    .... ...0 = PDU Header Digest: Not set
    .... ..0. = PDU Data Digest: Not set
    .... .0.. = PDU Data Last: Not set
    .... 0... = PDU Data Success: Not set
  Pdu Header Length: 10
  Pdu Data Offset: 12
  Packet Length: 268
  KDResp
    Kickstart Status: 1 (SUCCESS)
    Failure Reason: 0 (NO FAILURE)
    CDC NVM Qualified Name (CDCNQN): nqn.1988-11.com.dell:SFSS:9:20220824223058e8
```

Kickstart Response

Kamal Bakshi, Cisco

CDC Starts Pull Registration Process



Create
Admin Q
ID=0

NVMe Connect

```
◢ NVM Express Fabrics TCP Discovery Controller, Fabrics Type: Connect (0x01) Cmd ID: 0x0001
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 72
    Packet Length: 1096
  ◢ NVM Express Fabrics (Cmd)
      Opcode: 0x7f Fabric Cmd

      Reserved: 0x40
      Command ID: 0x0001
      Fabric Cmd Type: Connect (0x01)
      Reserved: 00000000000000000000000000000000000000
    ▷ SGL1
      Record Format: 0x0000
      Queue ID: 0x0000
      SQ Size: 0x0020
    ▷ Connect Attributes: 0x00
      Reserved: 00
      Keep Alive Timeout: 120000ms
      Reserved: 000000000000000000000000
  ◢ Data
      Host Identifier: 32313437-3438-3436-3733-000000000000
      Controller ID: 0xffff
      Reserved: 000000000000000000000000000000000000000000000000000000000000000000000000…
      Subsystem NQN: nqn.2014-08.org.nvmexpress.discovery
      Host NQN: nqn.1988-11.com.dell:SFSS:9:20220824223058e8
      Reserved: 000000000000000000000000000000000000000000000000000000000000000000000000…
```

Admin/SQ size 32

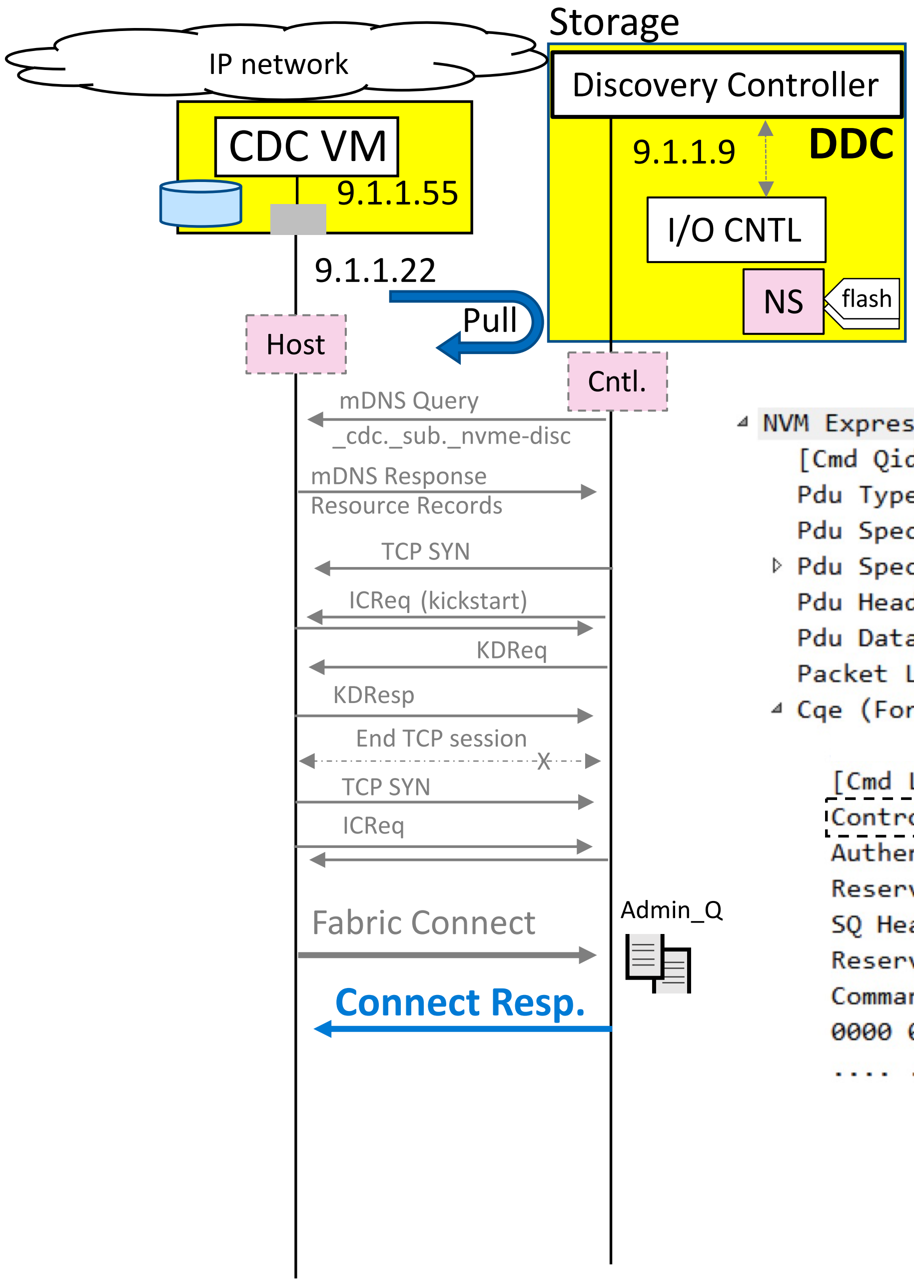
DDC NQN

Requests Any
DDC Controller ID

NQN
CDC (Host)

Kamal Bakshi, Cisco

CDC establishes NVMe session with DDC

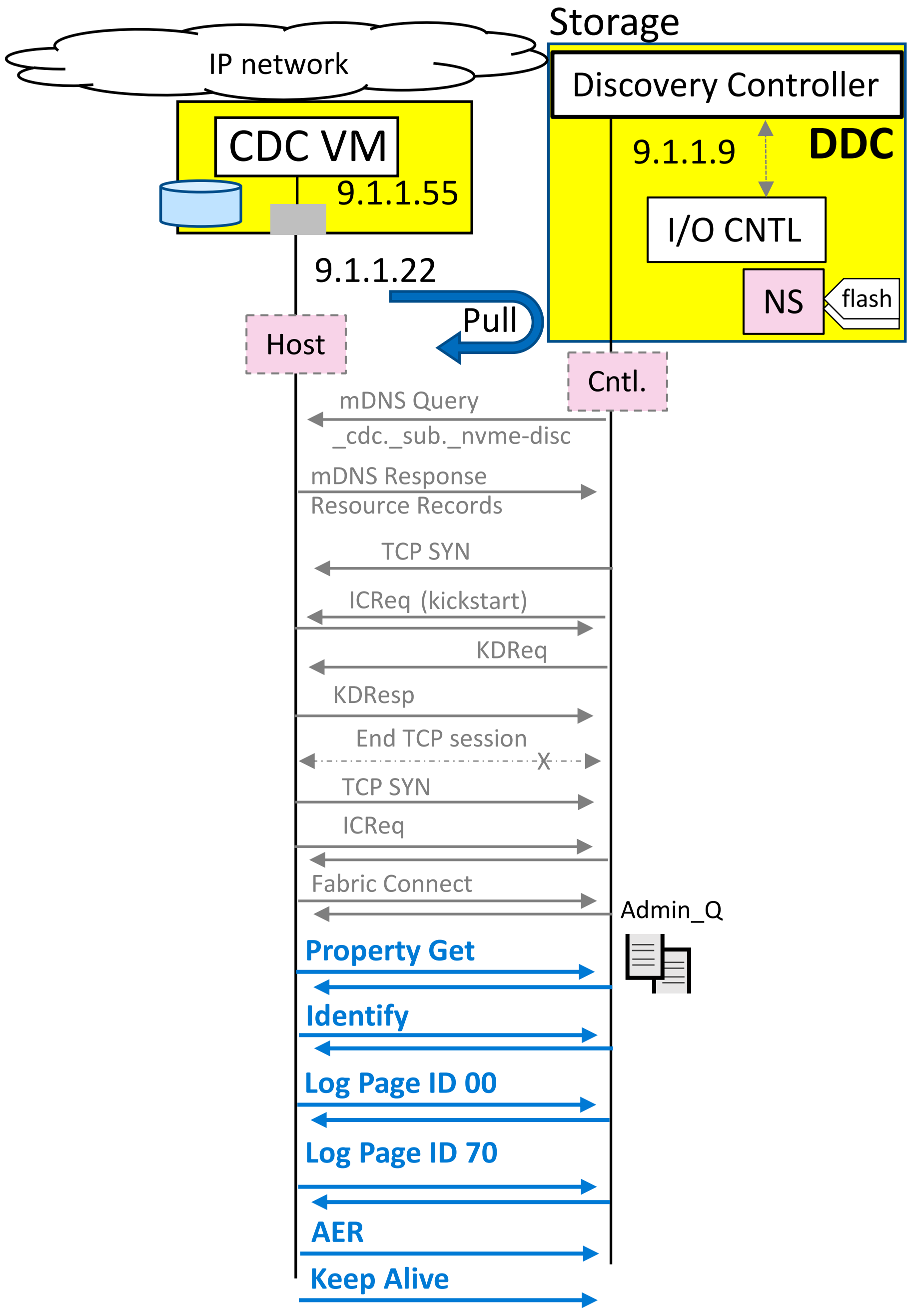



```
NVM Express Fabrics TCP Discovery Controller, Cqe Fabrics Cmd: Connect (0x01) Cmd ID: 0x0001
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
    Pdu Specific Flags: 0x00
    Pdu Specific Flags: 0x00
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
    Cqe (For Cmd: Connect)
      [Cmd Latency: 0.179 ms]
      Controller ID: 0x1000
      Authentication Required: 0x0000
      Reserved: 00000000
      SQ Head Pointer: 0x0100
      Reserved: 0x0000
      Command ID: 0x0001
      0000 0000 0000 000. = Status: 0x0000
      .... .... .... ...0 = Reserved: 0x0
```

DDC Dynamic Controller ID

Kamal Bakshi, Cisco

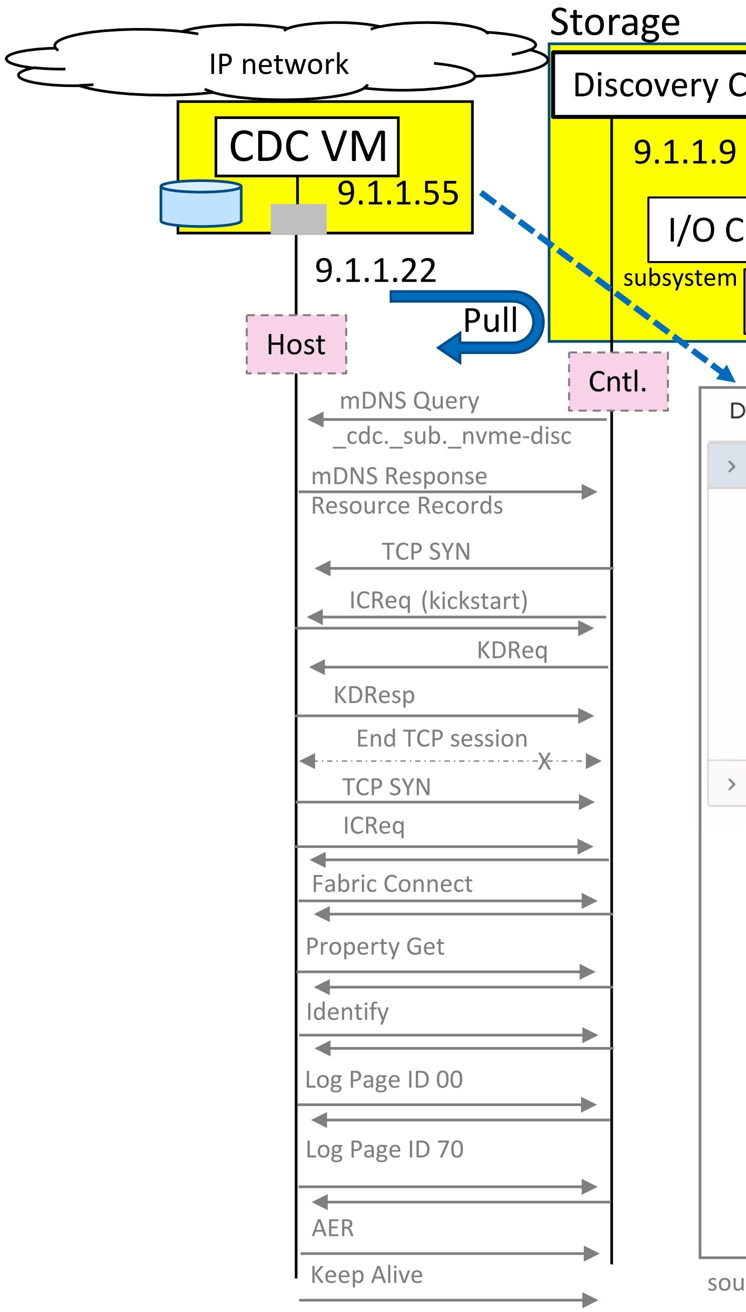
CDC gets the Log Page 70 from DDC




```
◢ Discovery Log Page Entry: 030102004009ffff200000000000000000000000000000000000000000000000003434
    Transport Type(TRTYPE): 3 [TCP Transport]
    Address Family(ADRFAM): 1 [AF_INET]
    Subsystem Type (SUBTYPE): 2 [NVM subsystem not associated with Discovery controllers and its
    Transport Requirements (TREQ): 0 [Controller requires use of SQ flow control ,Connection ove
    PORT ID: 2368
    Controller ID: 65535
    Admin Max SQ Size: 32
  ▷ Entry Flags: 0x00
    Reserved: 000000000000000000000000000000000000000000
    Transport Service Identifier(TRSVCID): 4420
    Reserved: 000000000000000000000000000000000000000000000000000000000000000000000000…
    NVM Subsytem Qualified Name: nqn.1988-11.com.dell:powerstore:00:35ddf55037ad6C7593F9
    Transport Address(TRADDR): 9.1.1.9
    Transport Specific Address SubType (TSAS): 0000000000000000000000000000000000000000000000000
  ◢ NVM Express Fabrics TCP Discovery Controller, Cqe NVMe Cmd: Get Log Page (0x02) Cmd ID: 0x000e
      [Cmd Qid: 0 (AQ)]
      Pdu Type: CapsuleResponse (5)
      Pdu Specific Flags: 0x00
    ◢ Pdu Specific Flags: 0x00
        .... ...0 = PDU Header Digest: Not set
        .... ..0. = PDU Data Digest: Not set
        .... .0.. = PDU Data Last: Not set
        .... 0... = PDU Data Success: Not set
      Pdu Header Length: 24
      Pdu Data Offset: 0
      Packet Length: 24
  ◢ NVM Express (Cqe)
      [Cmd in: 160]
      [Cmd Latency: 0.066 ms]
      Cmd specific Status: 0x0000000000000000
      SQ Head Pointer: 0x000b
      Reserved: 0x0000
      Command ID: 0x000e
      0000 0000 0000 000. = Status: 0x0000
      .... .... .... ...0 = Reserved: 0x0
```


Kamal Bakshi, Cisco

CDC extracts DDC and Subsystem Info



From Kickstart Discovery
(DDC controller)

From Log Page 70
(Subsystem I/O controller)

CDC Database

DDC Details

› General Information

Activate	true
Config Type	KickStart
Connection Status	Online
Port ID	8009
Transport Address	9.1.1.9
Transport Address Family	IPV4
Transport Type	TCP

› Subsystems

Subsystem Details

› General Information

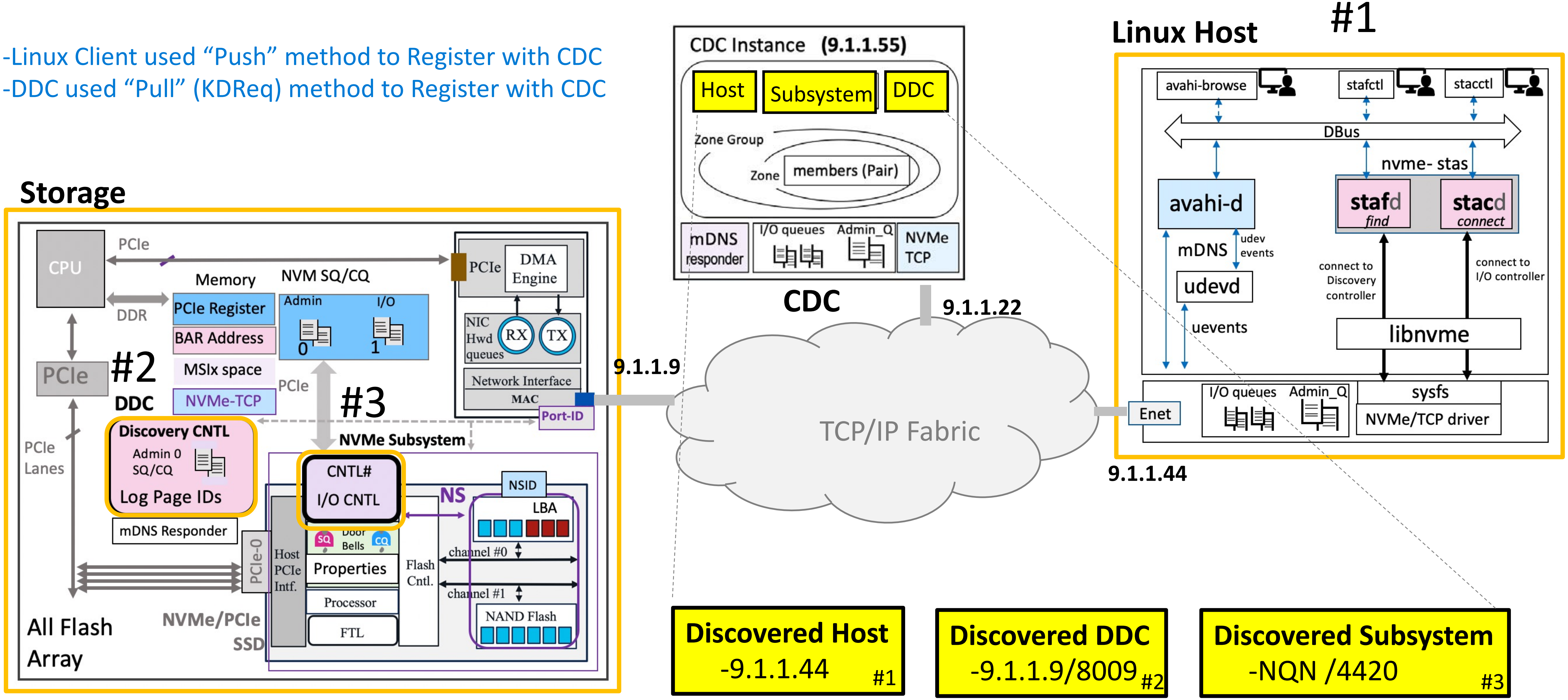
Admin Max SQ Size (ASQSZ)	32
Connection Status	Online
Controller Id	65535
DDC ID	@9.1.1.9:V4::0:8009:TCP
Entry Key Type	TRADDR
NQN	nqn.1988-11.com.dell:powerstore:00:3 5ddf55037ad6C7593F9
Node Name	stfs-cdcproxy-deployment-9-0
Port ID	2368
Generation Counter (GENCTR)	37
Registration Type	Pull
Sub Type	NVM Subsystem
Transport Requirements (TREQ)	Secure channel Not specified
Transport Specific Address Subtype (TSAS)	No Security
Transport Address	9.1.1.9
Transport Address Family	IPV4
Transport Service ID	4420
Transport Type	TCP

source: Dell SFSS/CDC

Kamal Bakshi, Cisco

CDC discovered Host, DDC & Subsystem (I/O controller)

-Linux Client used "Push" method to Register with CDC
-DDC used "Pull" (KDReq) method to Register with CDC

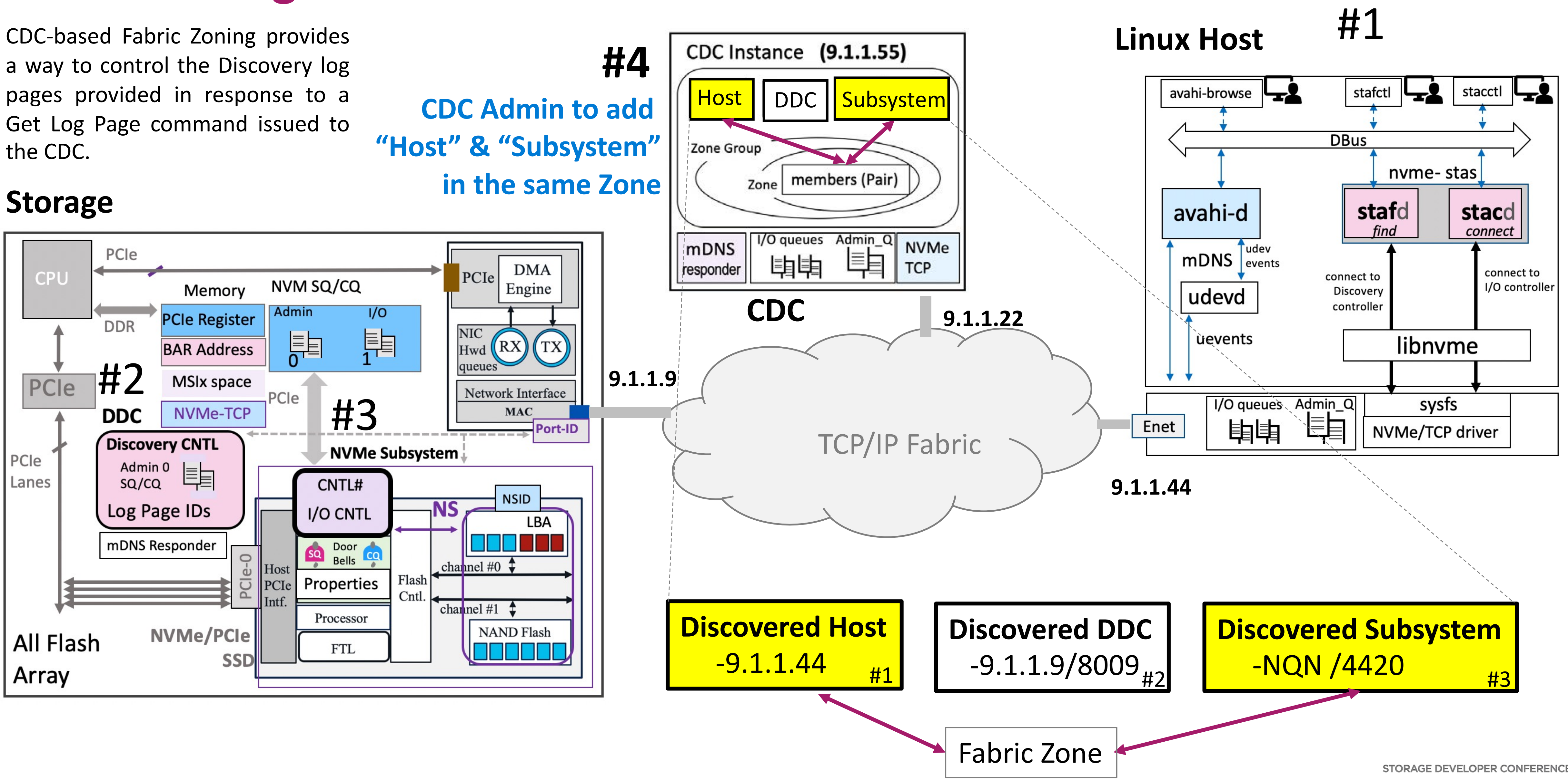


Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

Fabric Zoning at CDC

CDC-based Fabric Zoning provides a way to control the Discovery log pages provided in response to a Get Log Page command issued to the CDC.

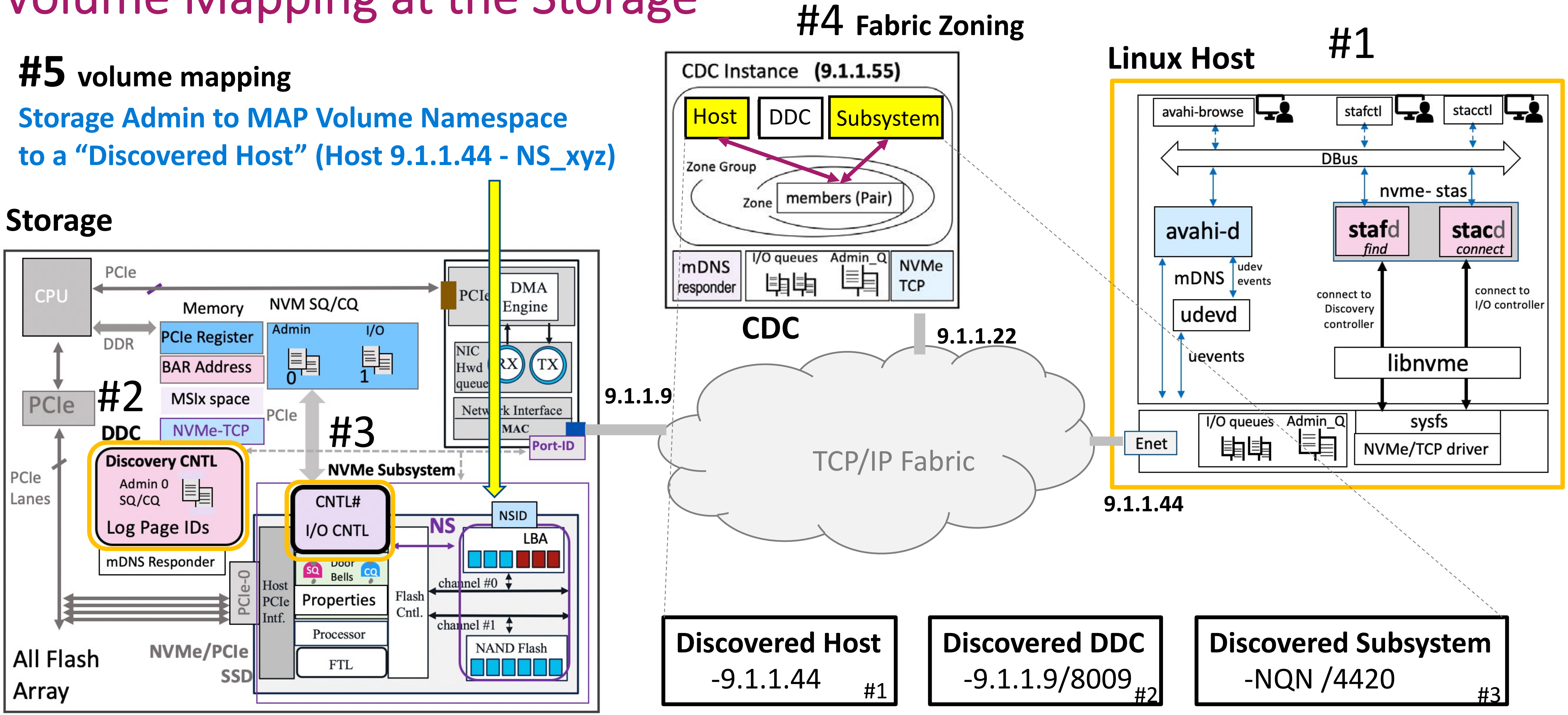


Kamal Bakshi, Cisco

Volume Mapping at the Storage

#5 volume mapping

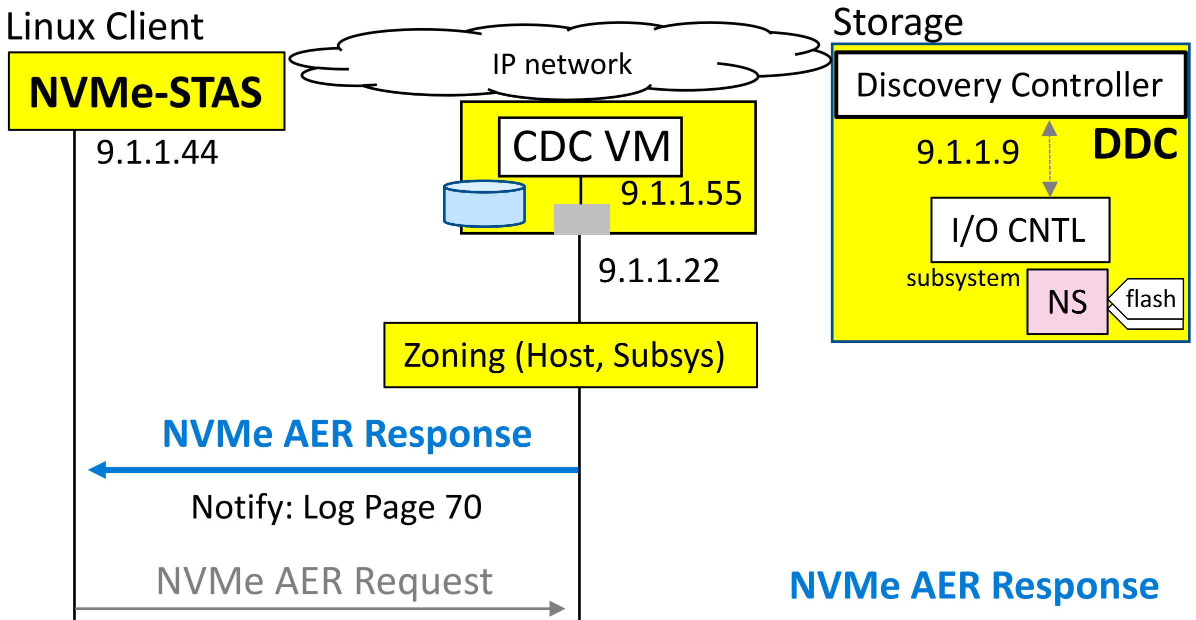
Storage Admin to MAP Volume Namespace to a "Discovered Host" (Host 9.1.1.44 - NS_xyz)



 |

Kamal Bakshi, Cisco

Activate the Fabric Zone at CDC



Zone Members

X REMOVE

CDC Admin enters the Zone members

☑	Role	Type	Id
☑	Host	NQN	nqn.2014-08.org.nvmexpress:uuid:77b1468d-39 483b-b4e2-de789e5048f9
☑	Subsystem	FullQualifiedName	nqn.1988-11.com.dell:powerstore:00:35ddf5503 6C7593F9@9.1.1.9:V4:4420:0:0:TCP

☑ 2 Members per page 10 1 - 2 of 2 members

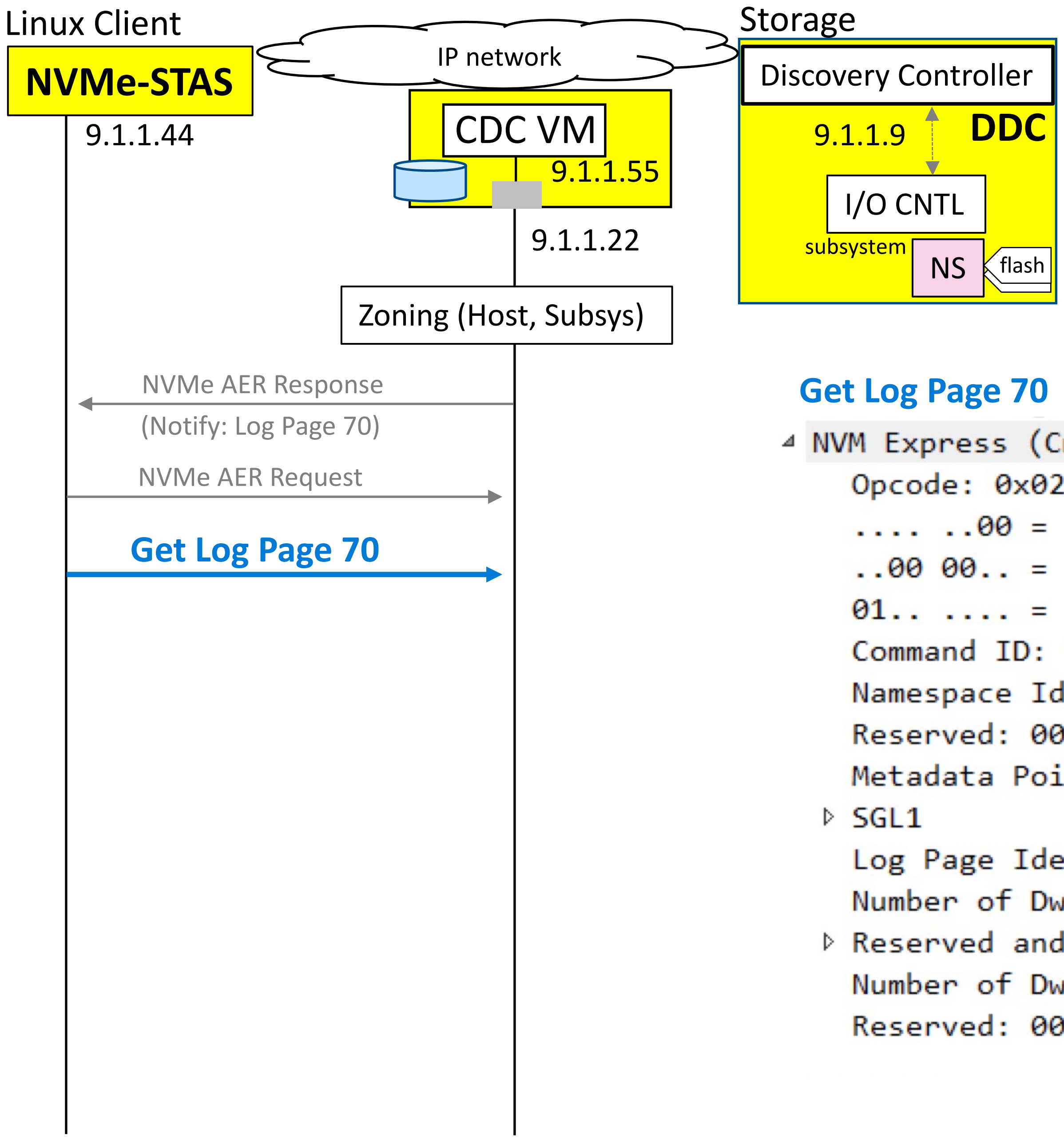
NVMe AER Response

```
◢ NVM Express Fabrics TCP Discovery Controller, Cqe NVMe Cmd: Async Event Request (0x0c)
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleResponse (5)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 24
    Pdu Data Offset: 0
    Packet Length: 24
◢ NVM Express (Cqe)
    [Cmd in: 178]
    [Cmd Latency: 4291220.323 ms]
    Cmd specific Status: 0x000000000070f002
    Asynchronous Event type: 2 Notice
    Asynchronous Event Information: 0xf0 Reserved for NVMe OF
    Log Page Identifier: 0x70 Discovery
    SQ Head Pointer: 0x0000
    Reserved: 0x0000
    Command ID: 0x001f
    0000 0000 0000 000. = Status: 0x0000
    .... .... .... ...0 = Reserved: 0x0
```

log page 70 change notification

Kamal Bakshi, Cisco

Host issues Get Log Page 70

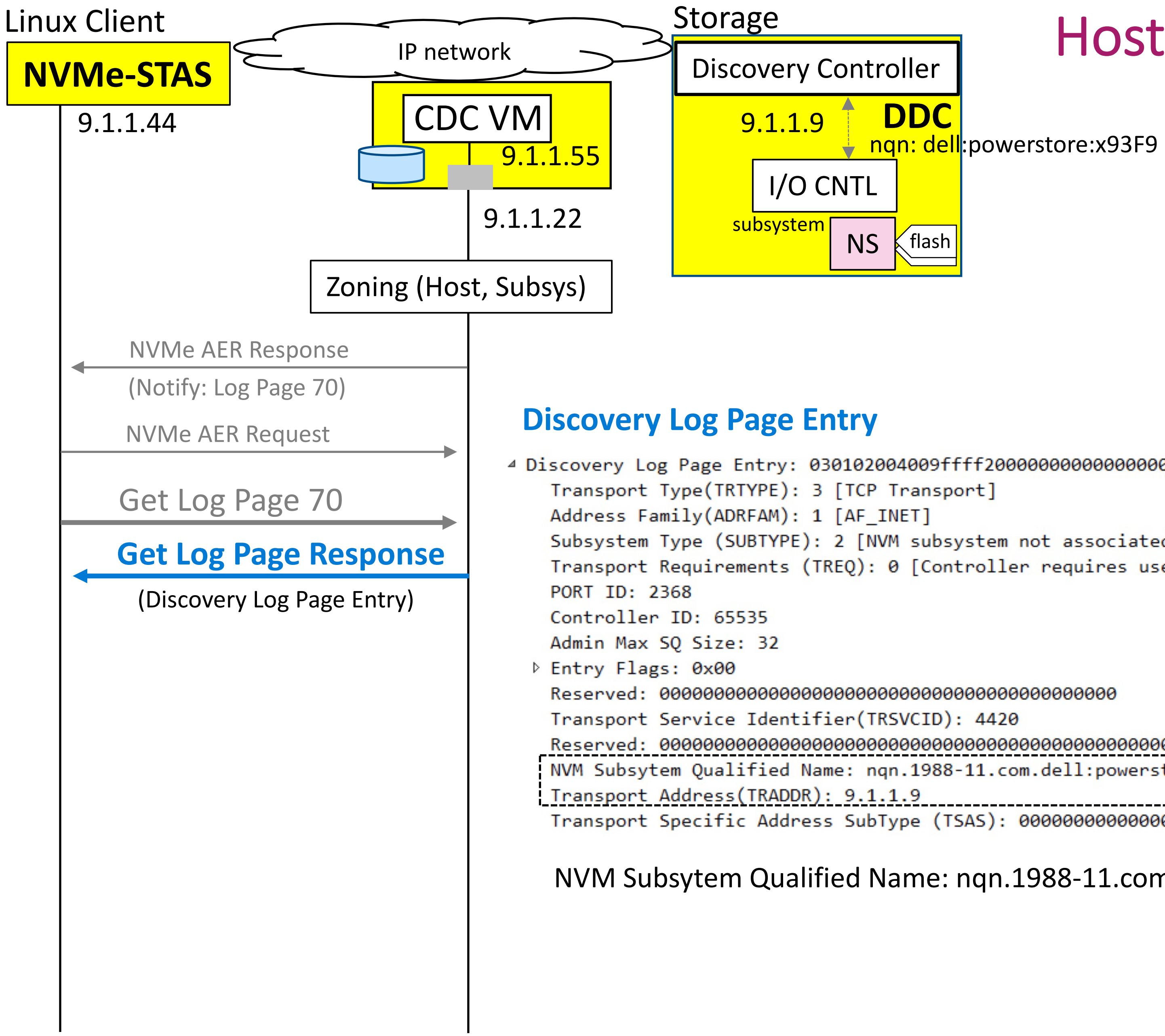


Get Log Page 70

```
◢ NVM Express (Cmd)
    Opcode: 0x02 Get Log Page
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x4004
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  ▷ SGL1
    Log Page Identifier: 0x70: Discovery
    Number of Dwords Lower: 0x003f: [0x3f 63]
  ▷ Reserved and RAE bits: 0x80, Retain Asynchronous Event
    Number of Dwords Upper: 0x0000: [0x3f 63]
    Reserved: 000000000000000000000000000000000000
```


Kamal Bakshi, Cisco

Host receives DLP with powerstore NQN



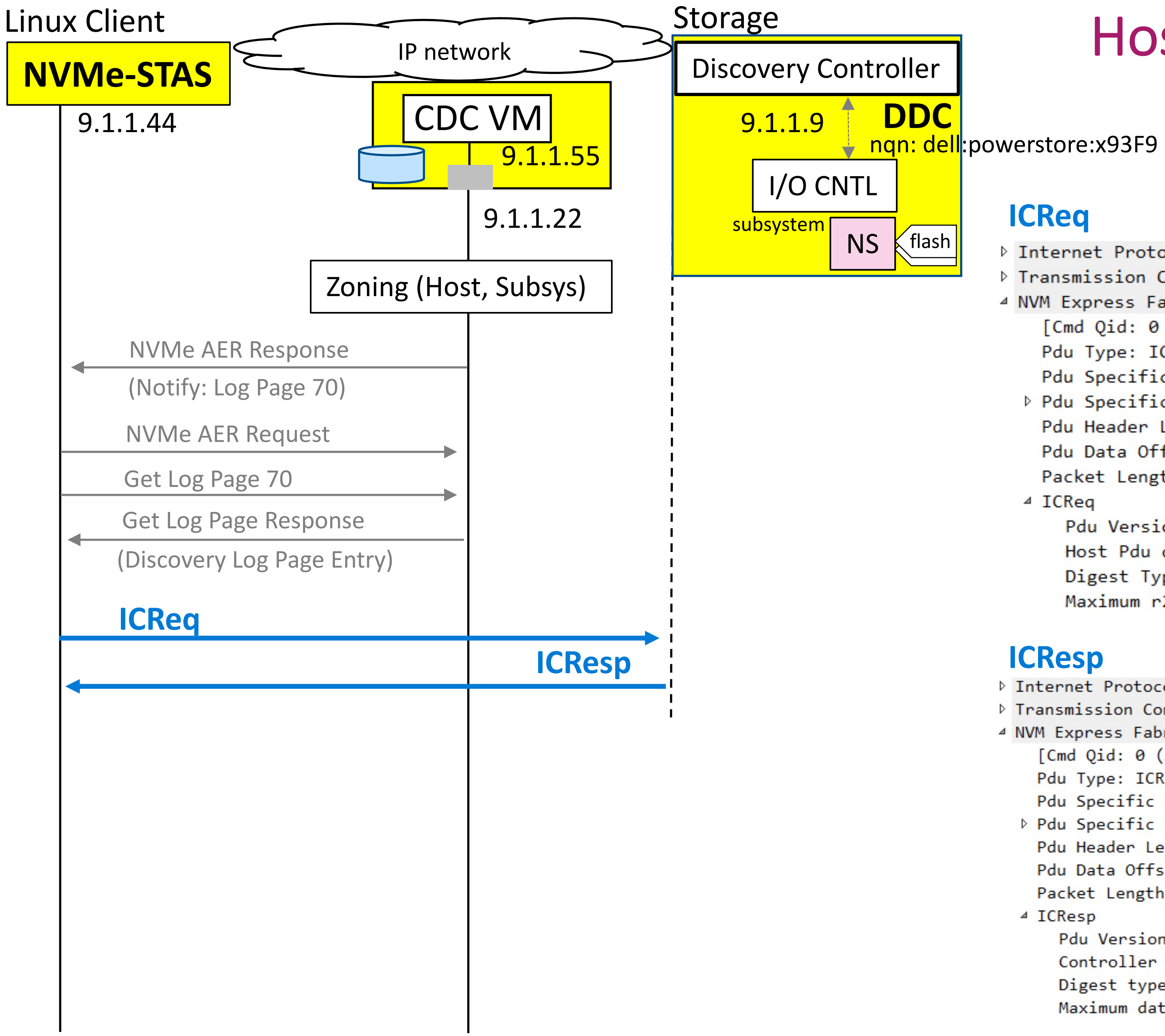
Discovery Log Page Entry

```
◢ Discovery Log Page Entry: 030102004009ffff20000000000000000000000000000000000000000000000034343230…[Page Entry:  0]
    Transport Type(TRTYPE): 3 [TCP Transport]
    Address Family(ADRFAM): 1 [AF_INET]
    Subsystem Type (SUBTYPE): 2 [NVM subsystem not associated with Discovery controllers and its controllers may have attached namespace]
    Transport Requirements (TREQ): 0 [Controller requires use of SQ flow control ,Connection over Fibre secure channel not specified]
    PORT ID: 2368
    Controller ID: 65535
    Admin Max SQ Size: 32
  ▷ Entry Flags: 0x00
    Reserved: 000000000000000000000000000000000000000000
    Transport Service Identifier(TRSVCID): 4420
    Reserved: 0000000000000000000000000000000000000000000000000000000000000000000000000…
    NVM Subsytem Qualified Name: nqn.1988-11.com.dell:powerstore:00:35ddf55037ad6C7593F9
    Transport Address(TRADDR): 9.1.1.9
    Transport Specific Address SubType (TSAS): 0000000000000000000000000000000000000000000000000000000000000000000000000…
```

NVM Subsytem Qualified Name: nqn.1988-11.com.dell:powerstore:00:35ddf55037ad6C7593F9

Kamal Bakshi, Cisco

Host initiates connection with storage



ICReq

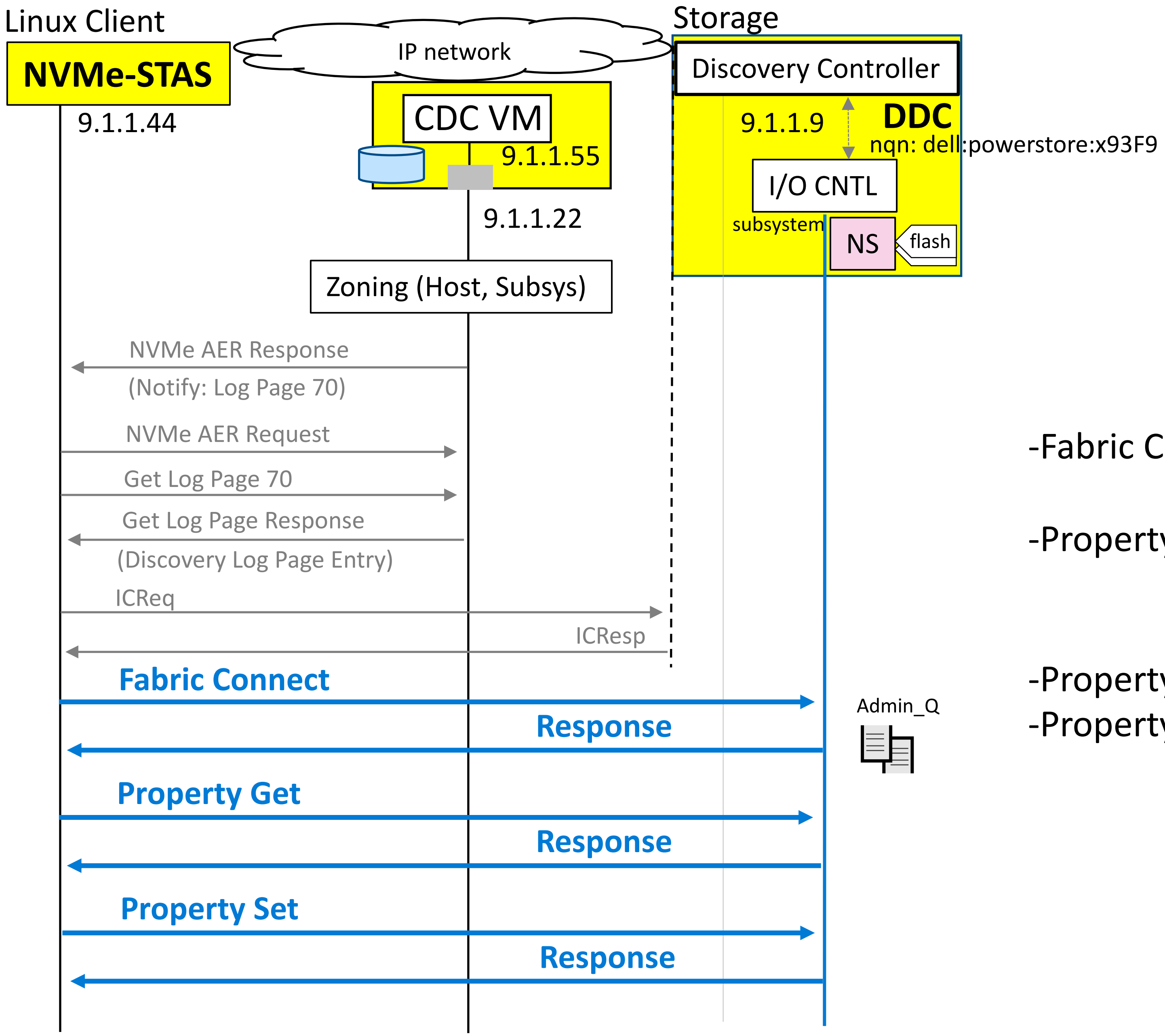
```
Internet Protocol Version 4, Src: 9.1.1.44, Dst: 9.1.1.9
Transmission Control Protocol, Src Port: 46472, Dst Port: 4420, Seq: 1, Ack: 1, Len: 128
NVM Express Fabrics TCP
  [Cmd Qid: 0 (AQ)]
  Pdu Type: ICReq (0)
  Pdu Specific Flags: 0x00Non-Kickstart discovery NVMe/TCP connection
  Pdu Specific Flags: 0x00
  Pdu Header Length: 128
  Pdu Data Offset: 0
  Packet Length: 128
  ICReq
    Pdu Version Format: 0
    Host Pdu data alignment: 0
    Digest Types Enabled: 0
    Maximum r2ts per request: 0
```

ICResp

```
Internet Protocol Version 4, Src: 9.1.1.9, Dst: 9.1.1.44
Transmission Control Protocol, Src Port: 4420, Dst Port: 46472,
NVM Express Fabrics TCP
  [Cmd Qid: 0 (AQ)]
  Pdu Type: ICResp (1)
  Pdu Specific Flags: 0x00
  Pdu Specific Flags: 0x00
  Pdu Header Length: 128
  Pdu Data Offset: 0
  Packet Length: 128
  ICResp
    Pdu Version Format: 0
    Controller Pdu data alignment: 0
    Digest types enabled: 0
    Maximum data capsules per r2t supported: 4194304
```


Kamal Bakshi, Cisco

Host setups NVMe connection with Storage I/O subsytem



-Fabric Connect creates Admin_Q id #0

-Property Get (Controller Capabilities)

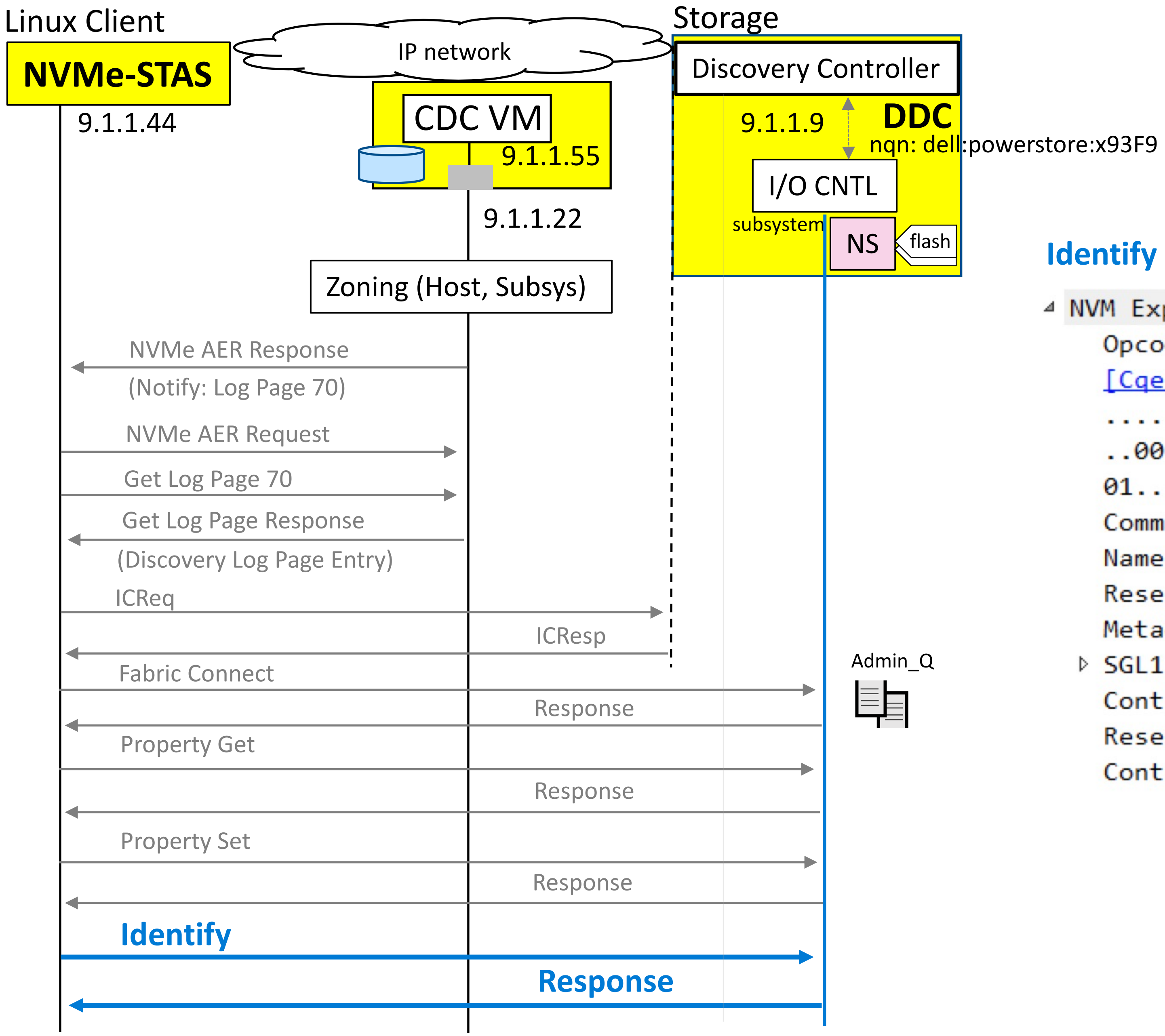
-Property Set (Controller Configuration)

-Property Get (Controller Status)

-Property Get (Version)

Kamal Bakshi, Cisco

Host gets Controller’s details

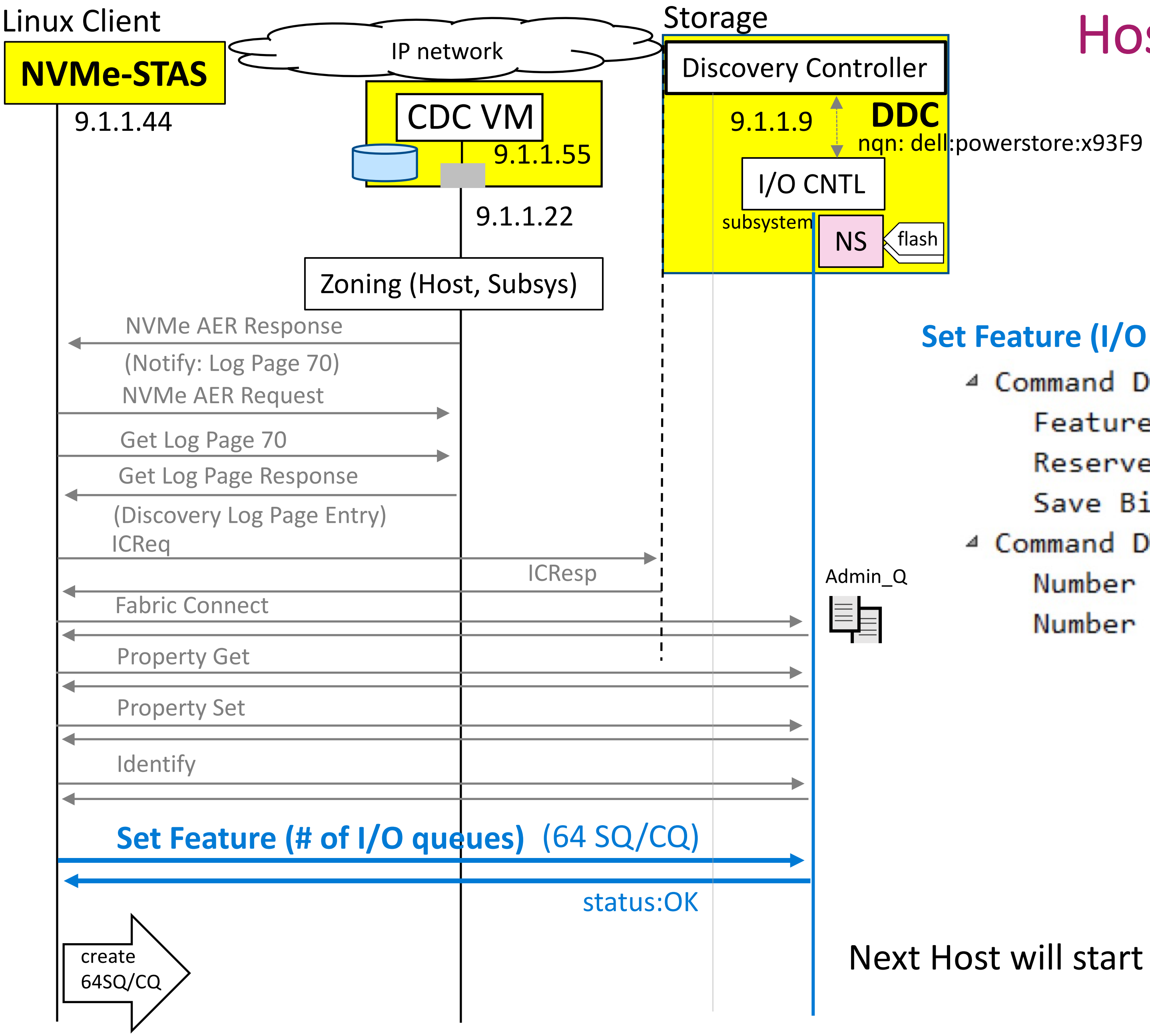


Identify

```
◢ NVM Express (Cmd)
    Opcode: 0x06 Identify
    [Cqe in: 1698]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x2011
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  ▷ SGL1
    Controller or Namespace Structure (CNS): 0x0001
    Reserved: 0000
    Controller Identifier (CNTID): 00000000
```


Kamal Bakshi, Cisco

Host requests to created 64 I/O SQ/CQ



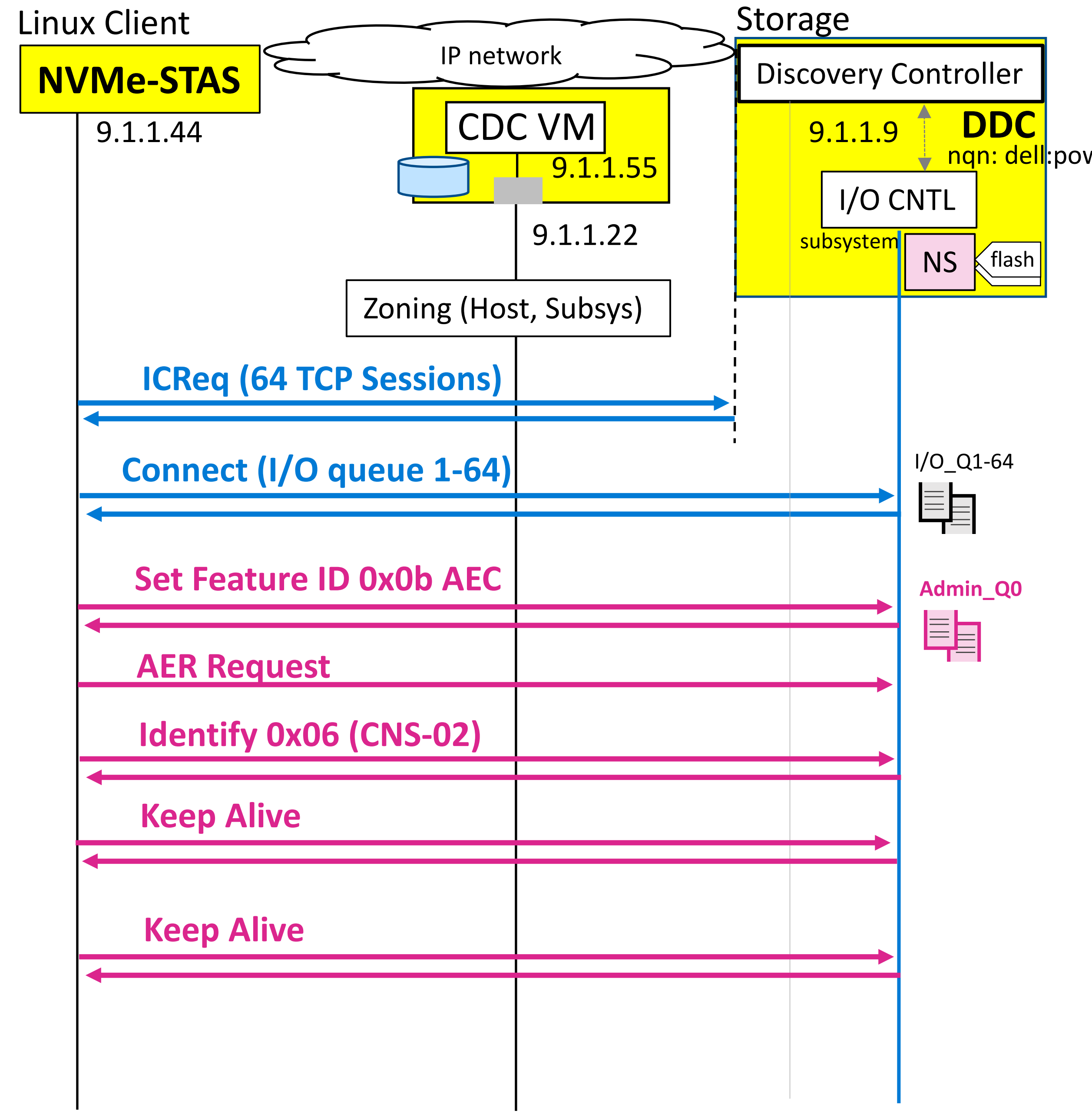
Set Feature (I/O queues)

- Command Dword 10: 0x00000007
  - Feature Identifier (FID): 0x07 Number of Queues
  - Reserved: 0x0000
  - Save Bit (SV): 0 Feature Identifier Saveable
- Command DWord 11: 0x003f003f
  - Number of I/O Submission Queue Requested: 0x003f
  - Number of I/O Completion Queue Requested: 0x003f

Next Host will start process of creating 64 I/O queues
(see next page)

Kamal Bakshi, Cisco

Host sends 64 Connects with QID#/Size



e.g. Connect Request for IOQ #64 (64 Connect Requests)

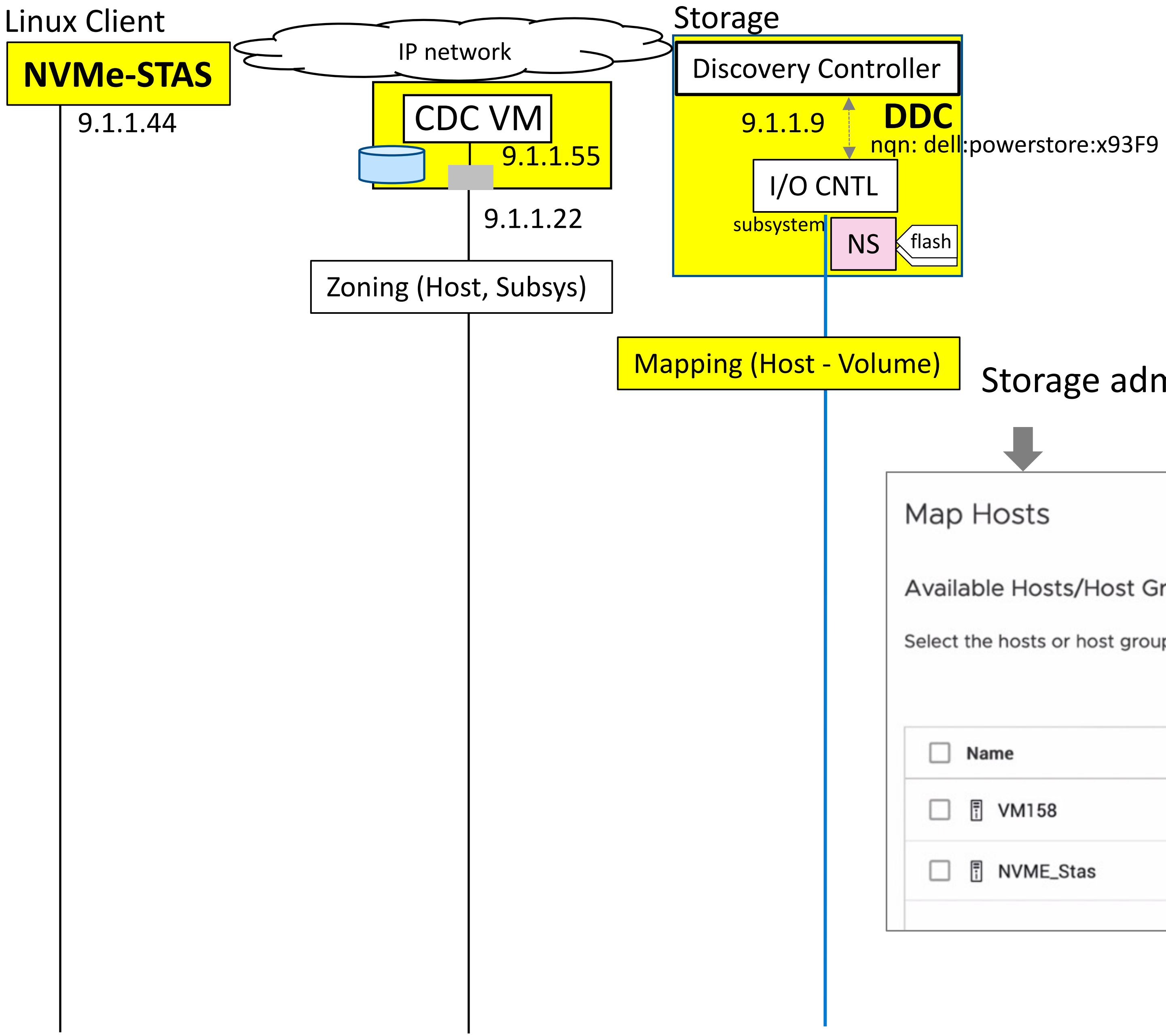
```
▷ Internet Protocol Version 4, Src: 9.1.1.44, Dst: 9.1.1.9
▷ Transmission Control Protocol, Src Port: 46614, Dst Port: 4420, Seq: 129, Ack: 129, Len:
◢ NVM Express Fabrics TCP, Fabrics Type: Connect (0x01) Cmd ID: 0x1000
    [Cmd Qid: 64 (IOQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 72
    Packet Length: 1096
  ◢ NVM Express Fabrics (Cmd)
      Opcode: 0x7f Fabric Cmd
      [Fabric Cqe in: 2615]
      Reserved: 0x40
      Command ID: 0x1000
      Fabric Cmd Type: Connect (0x01)
      Reserved: 00000000000000000000000000000000000000
    ▷ SGL1
      Record Format: 0x0000
      Queue ID: 0x0040
      SQ Size: 0x007f
      Connect Attributes: 0x00
      Reserved: 00
      Keep Alive Timeout: 0ms
      Reserved: 000000000000000000000000
  ◢ Data
      Host Identifier: aecb70db-d20d-45ae-b91f-7405b932ae19
      Controller ID: 0x0fff
      Reserved: 00000000000000000000000000000000000000000000000000000000000000000000…
      Subsystem NQN: nqn.1988-11.com.dell:powerstore:00:35ddf55037ad6C7593F9
      Host NQN: nqn.2014-08.org.nvmexpress:uuid:77b1468d-39f7-483b-b4e2-de789e5048f9
      Reserved: 00000000000000000000000000000000000000000000000000000000000000000000…
```

I/O QID #64

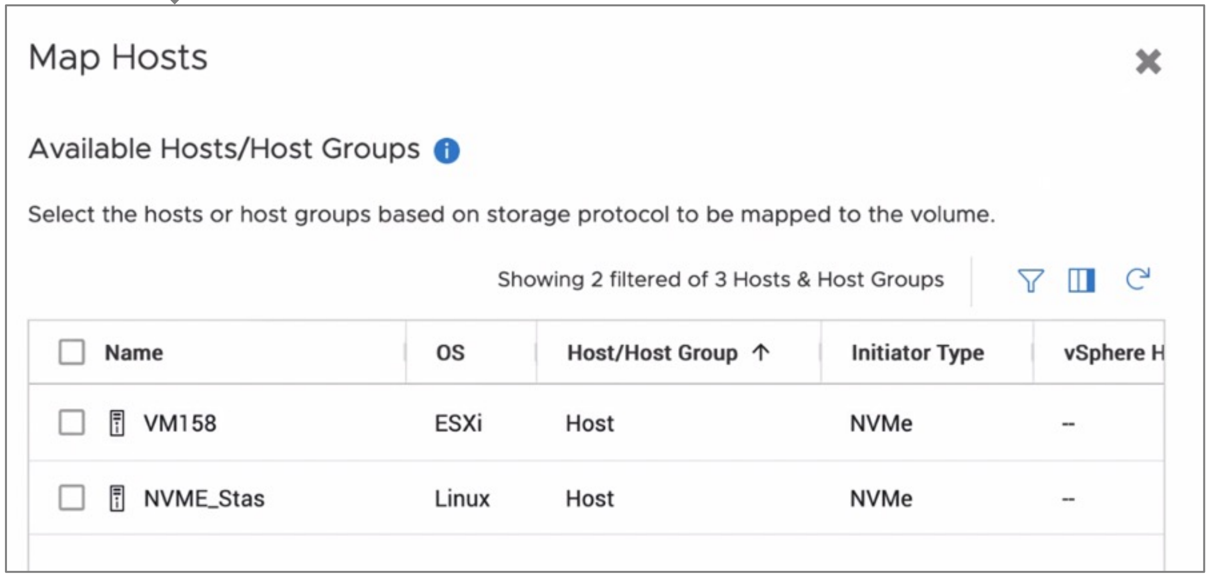
-Storage NQN
-Host NQN

Kamal Bakshi, Cisco

Volume-Host Mapping at Storage



Storage administrator maps a discovered Host NQN to a volume



Map Hosts

Available Hosts/Host Groups

Select the hosts or host groups based on storage protocol to be mapped to the volume.

Showing 2 filtered of 3 Hosts & Host Groups

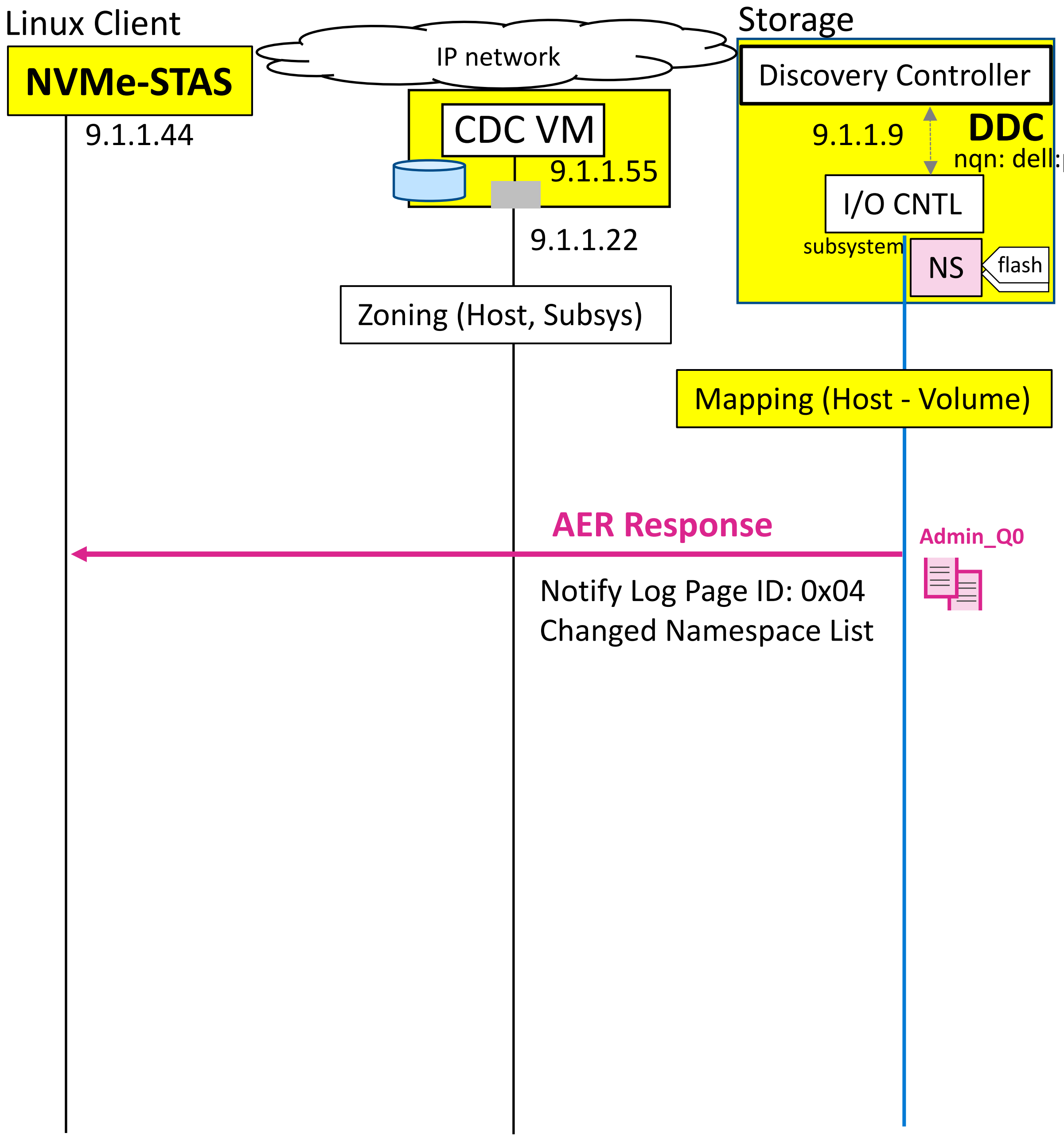
Name	OS	Host/Host Group ↑	Initiator Type	vSphere H
VM158	ESXi	Host	NVMe	--
NVME_Stas	Linux	Host	NVMe	--

source: Dell/PowerStore

 |

Kamal Bakshi, Cisco

Storage Notifies Host about changes

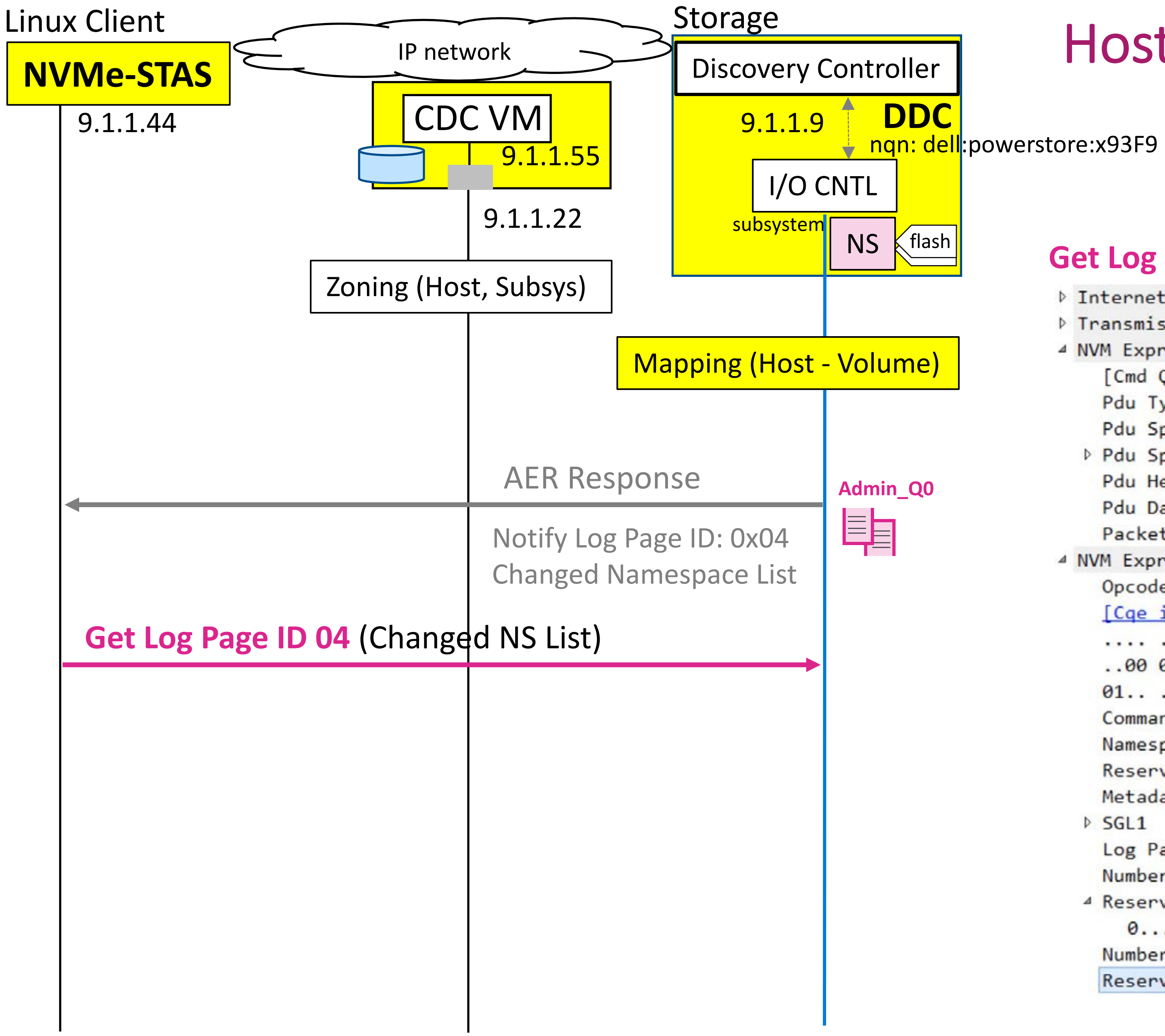


AER Response (Notification)

```
▷ Internet Protocol Version 4, Src: 9.1.1.9, Dst: 9.1.1.44
▷ Transmission Control Protocol, Src Port: 4420, Dst Port: 46472, Seq: 548273, Ack: 12097
▷ NVM Express Fabrics TCP, Cqe NVMe Cmd: Async Event Request (0x0c) Cmd ID: 0x001f
◢ NVM Express (Cqe)
    [Cmd in: 2873]
    [Cmd Latency: 189366.626 ms]
    Cmd specific Status: 0x0000000000040002
    Asynchronous Event type: 2 Notice
    Asynchronous Event Information: 0x00 Namespace Attribute Changed
    Log Page Identifier: 0x04 Changed Namespace List
    SQ Head Pointer: 0x0018
    Reserved: 0x0000
    Command ID: 0x001f
    0000 0000 0000 000. = Status: 0x0000
    .... .... .... ...0 = Reserved: 0x0
```


Kamal Bakshi, Cisco

Host requests changed Namespace List

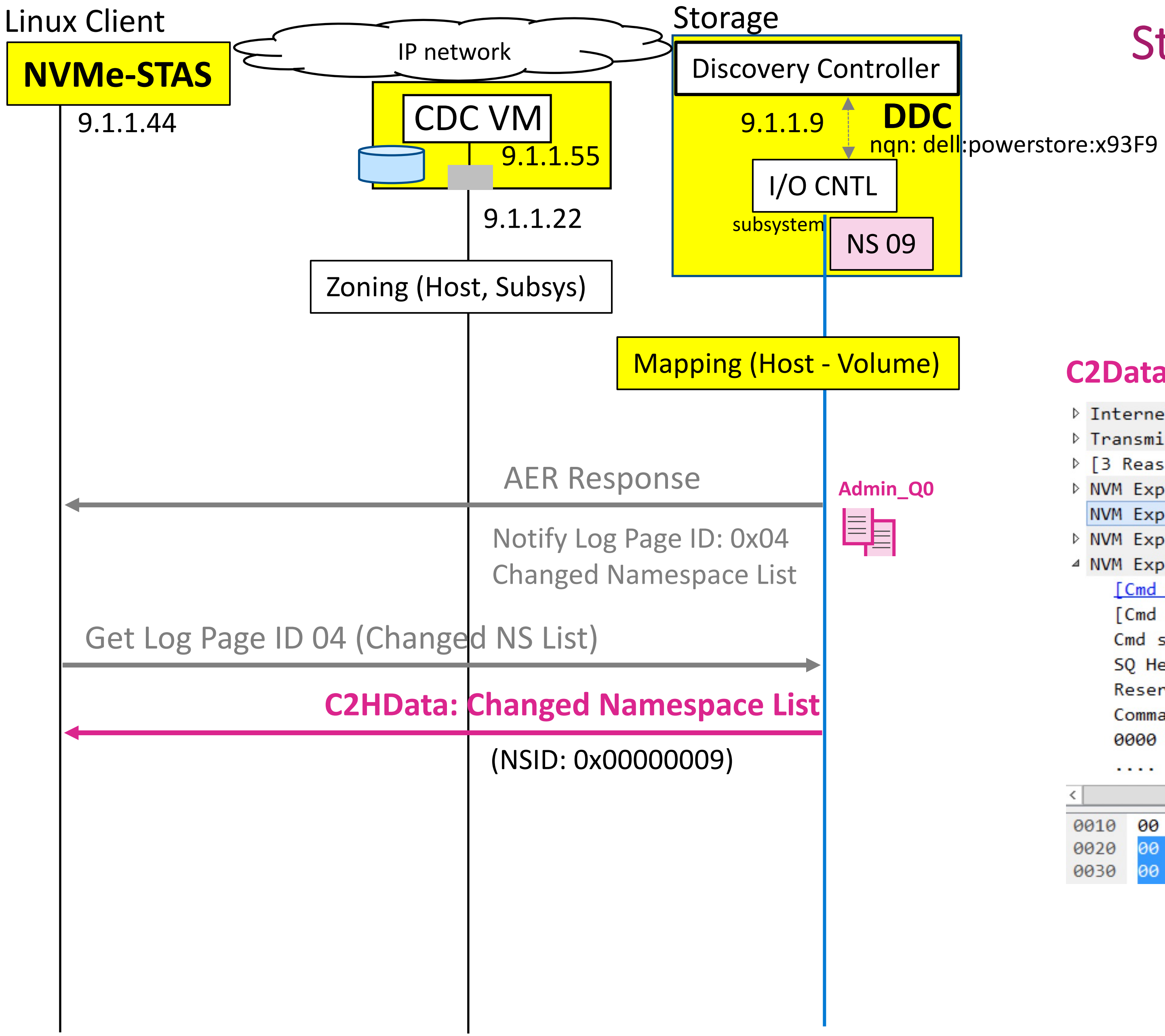


Get Log Page ID 04 (Changed NS List)

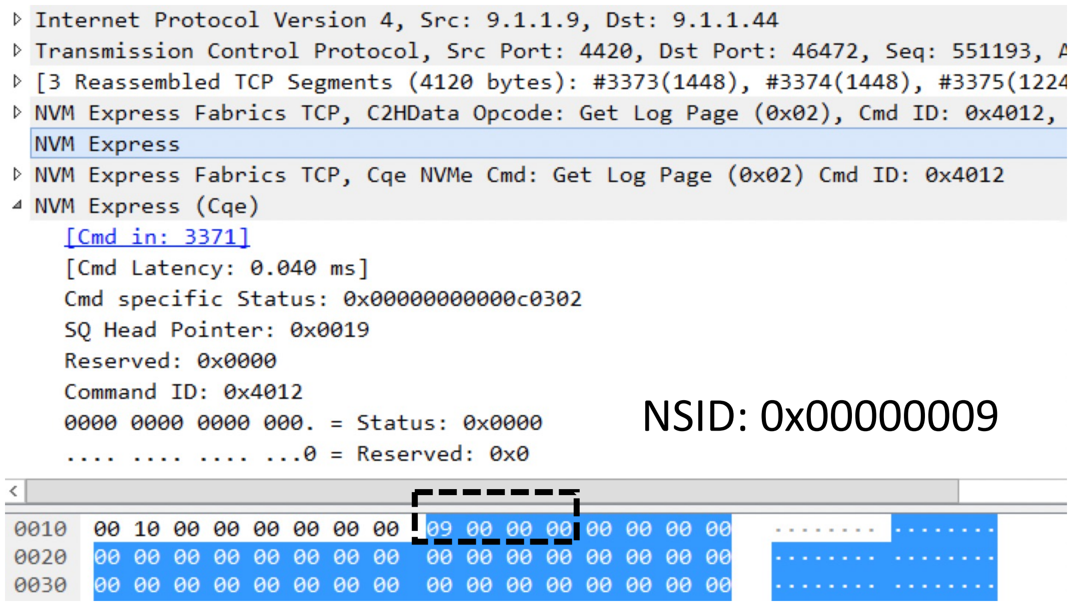
```
▷ Internet Protocol Version 4, Src: 9.1.1.44, Dst: 9.1.1.9
▷ Transmission Control Protocol, Src Port: 46472, Dst Port: 4420, Seq: 12169,
◢ NVM Express Fabrics TCP, NVMe Opcode: Get Log Page (0x02) Cmd ID: 0x4012
    [Cmd Qid: 0 (AQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
◢ NVM Express (Cmd)
    Opcode: 0x02 Get Log Page
    [Cqe in: 3375]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x4012
    Namespace Id: 0xffffffff
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  ▷ SGL1
    Log Page Identifier: 0x04: Changed Namespace List
    Number of Dwords Lower: 0x03ff: [0x3ff 1023]
  ◢ Reserved and RAE bits: 0x00
      0... .... = Retain Asynchronous Event: False
    Number of Dwords Upper: 0x0000: [0x3ff 1023]
    Reserved: 000000000000000000000000000000000000
```


Kamal Bakshi, Cisco

Storage returns Changed Namespace List



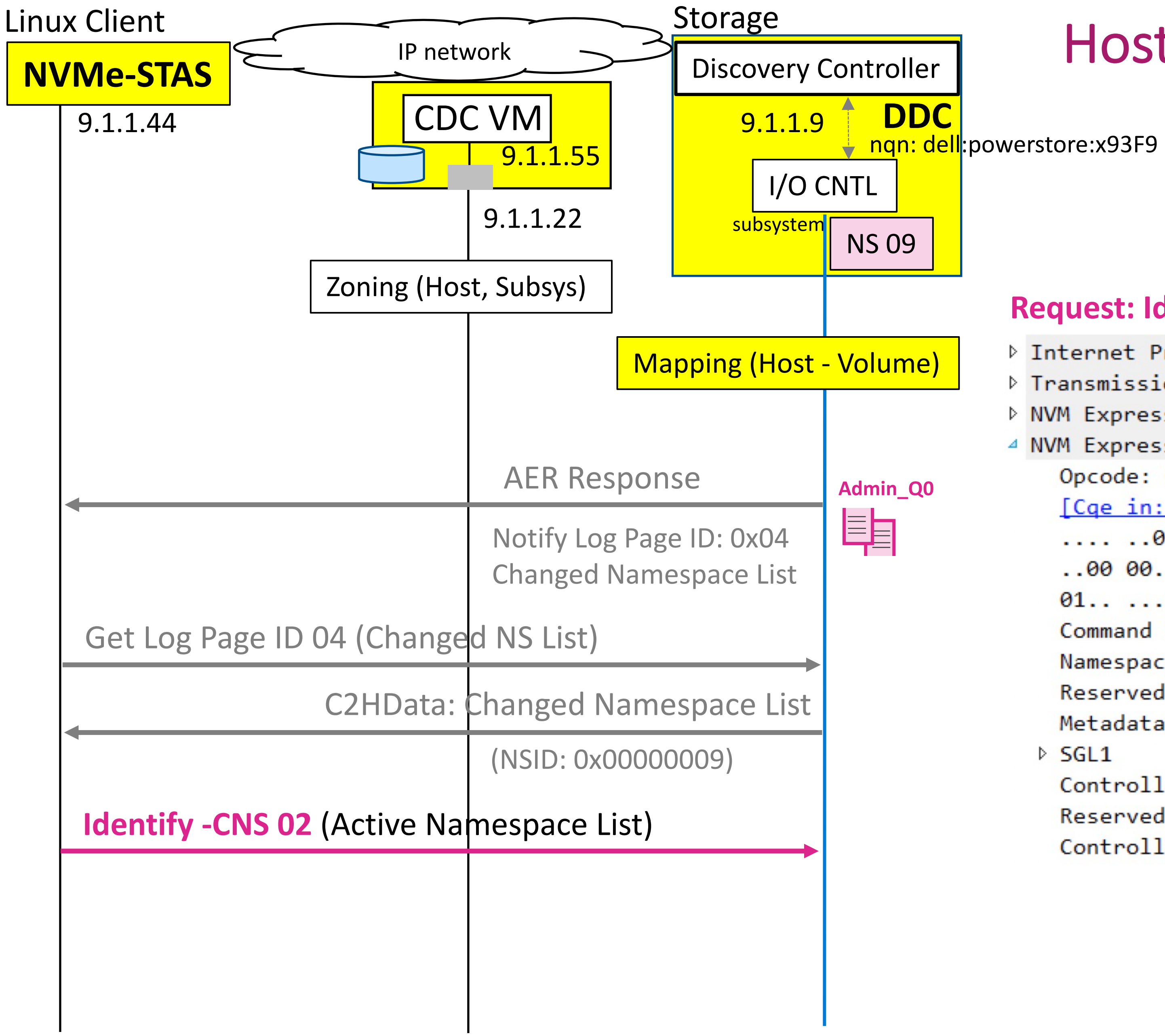
C2Data: Changed Namespace List



NSID: 0x00000009

Kamal Bakshi, Cisco

Host requests Active Namespace List



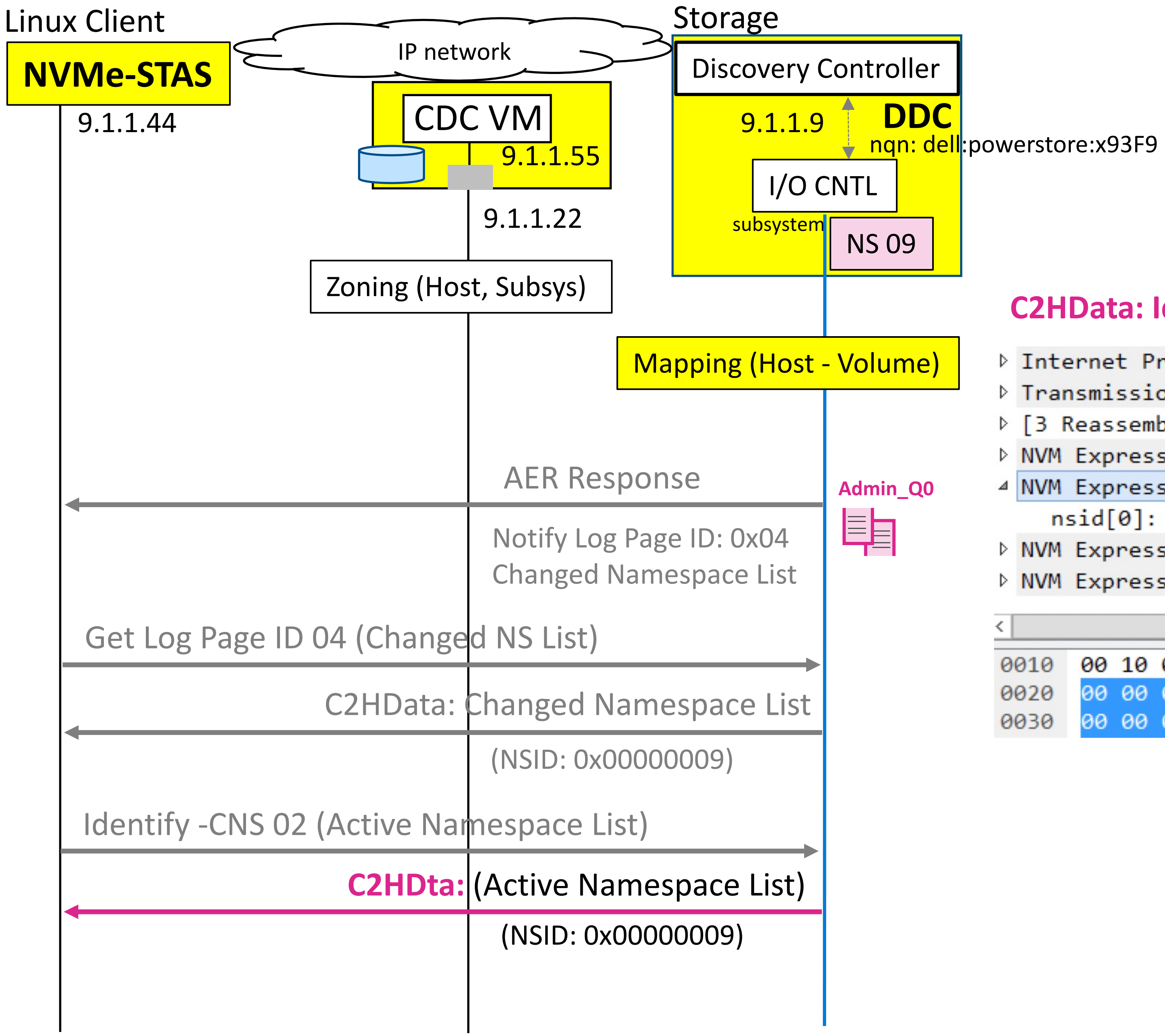
Request: Identify -CNS 02 (Active Namespace List)

```
▷ Internet Protocol Version 4, Src: 9.1.1.44, Dst: 9.1.1.9
▷ Transmission Control Protocol, Src Port: 46472, Dst Port: 4420, Seq: 12241,
▷ NVM Express Fabrics TCP, NVMe Opcode: Identify (0x06) Cmd ID: 0x4013
◢ NVM Express (Cmd)
    Opcode: 0x06 Identify
    [Cqe in: 3380]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x4013
    Namespace Id: 0x00000000
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  ▷ SGL1
    Controller or Namespace Structure (CNS): 0x0002
    Reserved: 0000
    Controller Identifier (CNTID): 00000000
```

Active Namespace List

Kamal Bakshi, Cisco

Active Namespace List NSID 0x09



C2HData: Identify -CNS 02 (Active Namespace List)

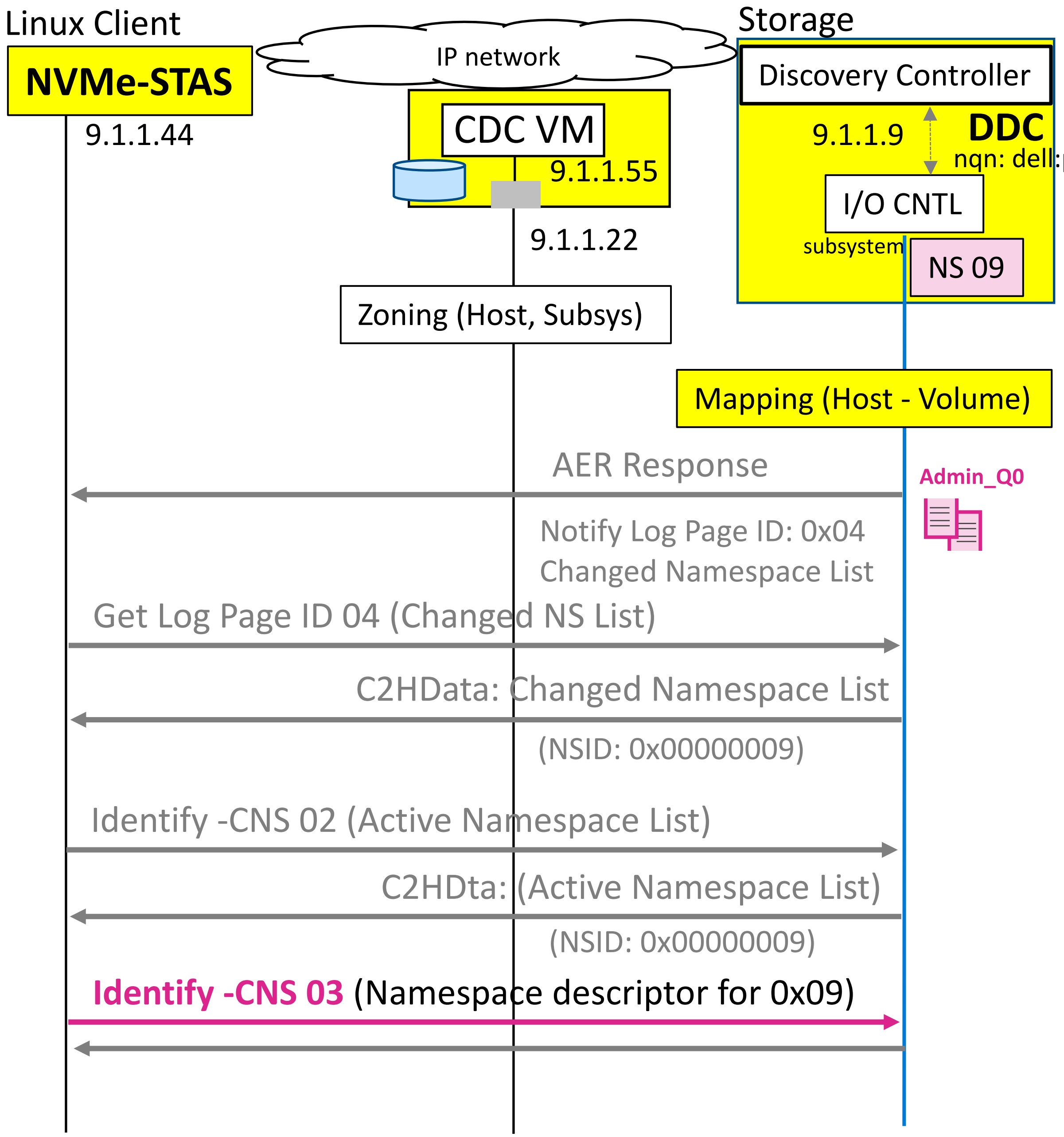
```
▷ Internet Protocol Version 4, Src: 9.1.1.9, Dst: 9.1.1.44
▷ Transmission Control Protocol, Src Port: 4420, Dst Port: 46472, Seq: 555337,
▷ [3 Reassembled TCP Segments (4120 bytes): #3378(1448), #3379(1448), #3380(12
▷ NVM Express Fabrics TCP, C2HData Opcode: Identify (0x06), Cmd ID: 0x4013, Le
◢ NVM Express
    nsid[0]: 9
▷ NVM Express Fabrics TCP, Cqe NVMe Cmd: Identify (0x06) Cmd ID: 0x4013
▷ NVM Express (Cqe)

0010  00 10 00 00 00 00 00 00  09 00 00 00 00 00 00 00   ........ ........
0020  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00   ........ ........
0030  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00   ........ ........
```

Active Namespace List

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE

Host requests Namespace Descriptor



Request: Identify -CNS 03 (Namespace descriptor for 0x09)

```
▷ Internet Protocol Version 4, Src: 9.1.1.44, Dst: 9.1.1.9
▷ Transmission Control Protocol, Src Port: 46472, Dst Port: 4420, Seq:
▷ NVM Express Fabrics TCP, NVMe Opcode: Identify (0x06) Cmd ID: 0x4014
◢ NVM Express (Cmd)
    Opcode: 0x06 Identify
    [Cqe in: 3385]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x4014
    Namespace Id: 0x00000009
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  ▷ SGL1
    Controller or Namespace Structure (CNS): 0x0003
    Reserved: 0000
    Controller Identifier (CNTID): 00000000
```

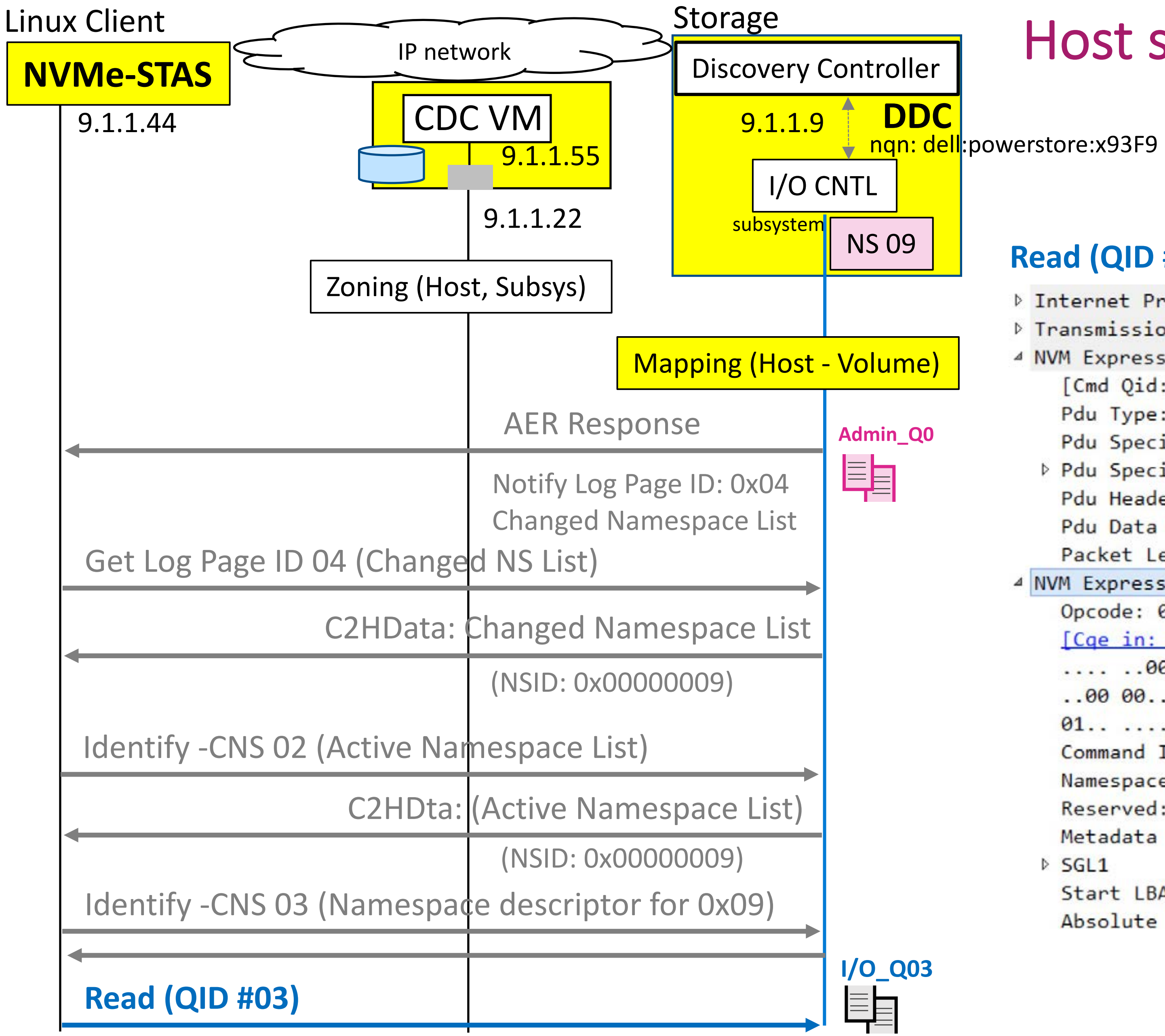
Namespace descriptor for NSID

NSID descriptor

 |

Kamal Bakshi, Cisco

Host sends Read Request over I/O queue

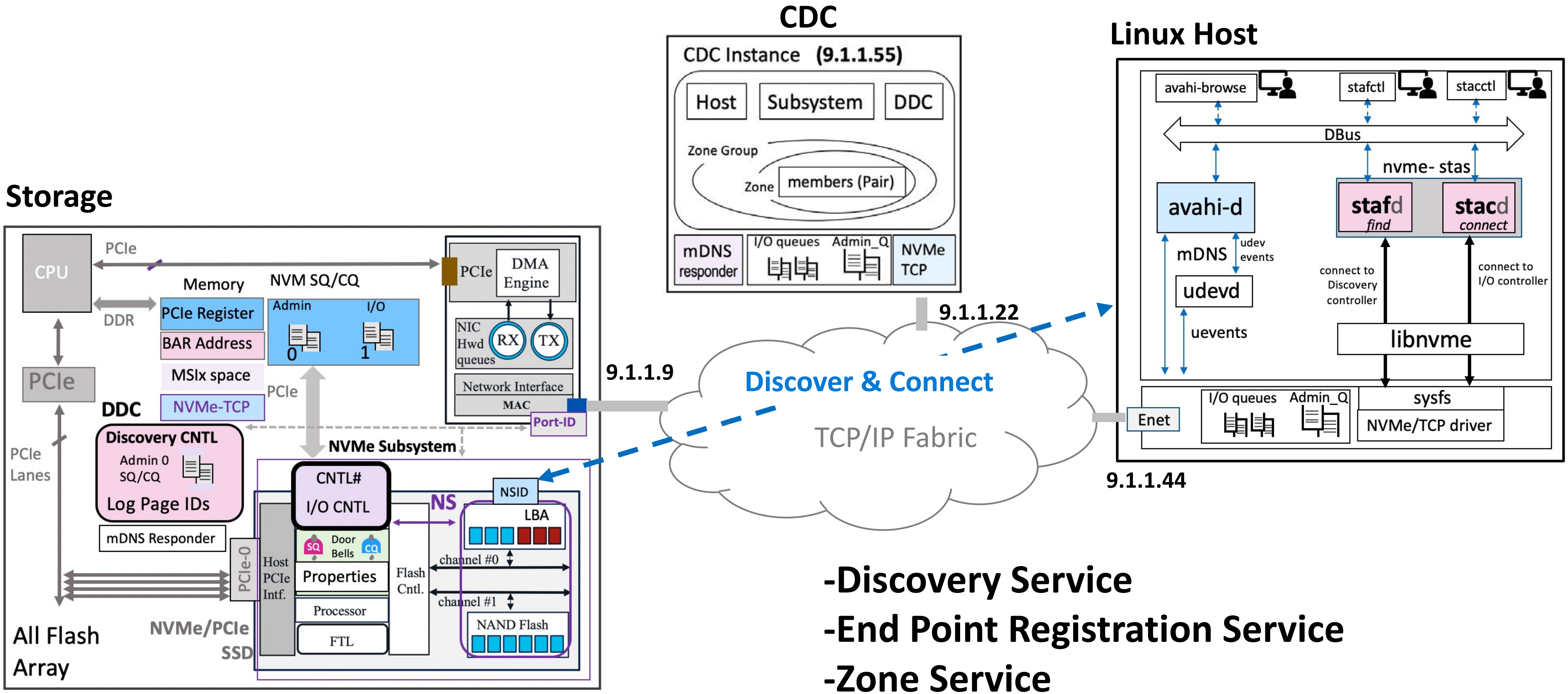


Read (QID #03)

```
▷ Internet Protocol Version 4, Src: 9.1.1.44, Dst: 9.1.1.9
▷ Transmission Control Protocol, Src Port: 46478, Dst Port: 4420, Seq: 1225
◢ NVM Express Fabrics TCP, NVMe Opcode: Read (0x02) Cmd ID: 0x1021
    [Cmd Qid: 3 (IOQ)]
    Pdu Type: CapsuleCommand (4)
    Pdu Specific Flags: 0x00
  ▷ Pdu Specific Flags: 0x00
    Pdu Header Length: 72
    Pdu Data Offset: 0
    Packet Length: 72
◢ NVM Express (Cmd)
    Opcode: 0x02 Read
    [Cqe in: 3588]
    .... ..00 = Fuse Operation: 0x0
    ..00 00.. = Reserved: 0x0
    01.. .... = PRP Or SGL: 0x1
    Command ID: 0x1021
    Namespace Id: 0x00000009
    Reserved: 0000000000000000
    Metadata Pointer: 0x0000000000000000
  ▷ SGL1
    Start LBA: 0x0000000000000000
    Absolute Number of Logical Blocks: 0x0008
```


Kamal Bakshi, Cisco

NVMe/TCP Fabric with CDC



- **-Discovery Service**
- **-End Point Registration Service**
- **-Zone Service**
- **-Asynchronous Notifications**

Kamal Bakshi, Cisco

Agenda

1. Data Center Storage Architecture
2. NVMe/FC Architecture
3. NVMe/TCP Architecture

Appendix

- NVMe Evolution
- NVMe/PCIe Architecture
- NVMe/FC Packets
- NVMe/RoCEv2 Architecture
- NVMe Advanced Features

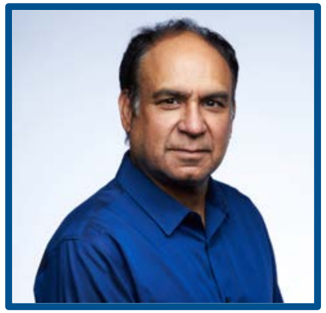
https://www.ciscolive.com/on-demand/on-demand-library.html?search=kamal%20bakshi#/

(Cisco Live video session that covers the above Appendix topics)

Kamal Bakshi, Cisco

Cisco NVMe-oF Research Center

For the past couple of years we have been extensively testing NVMe transports related technologies at Cisco DC Proof of Concept lab. For more information please reach out to Cisco.

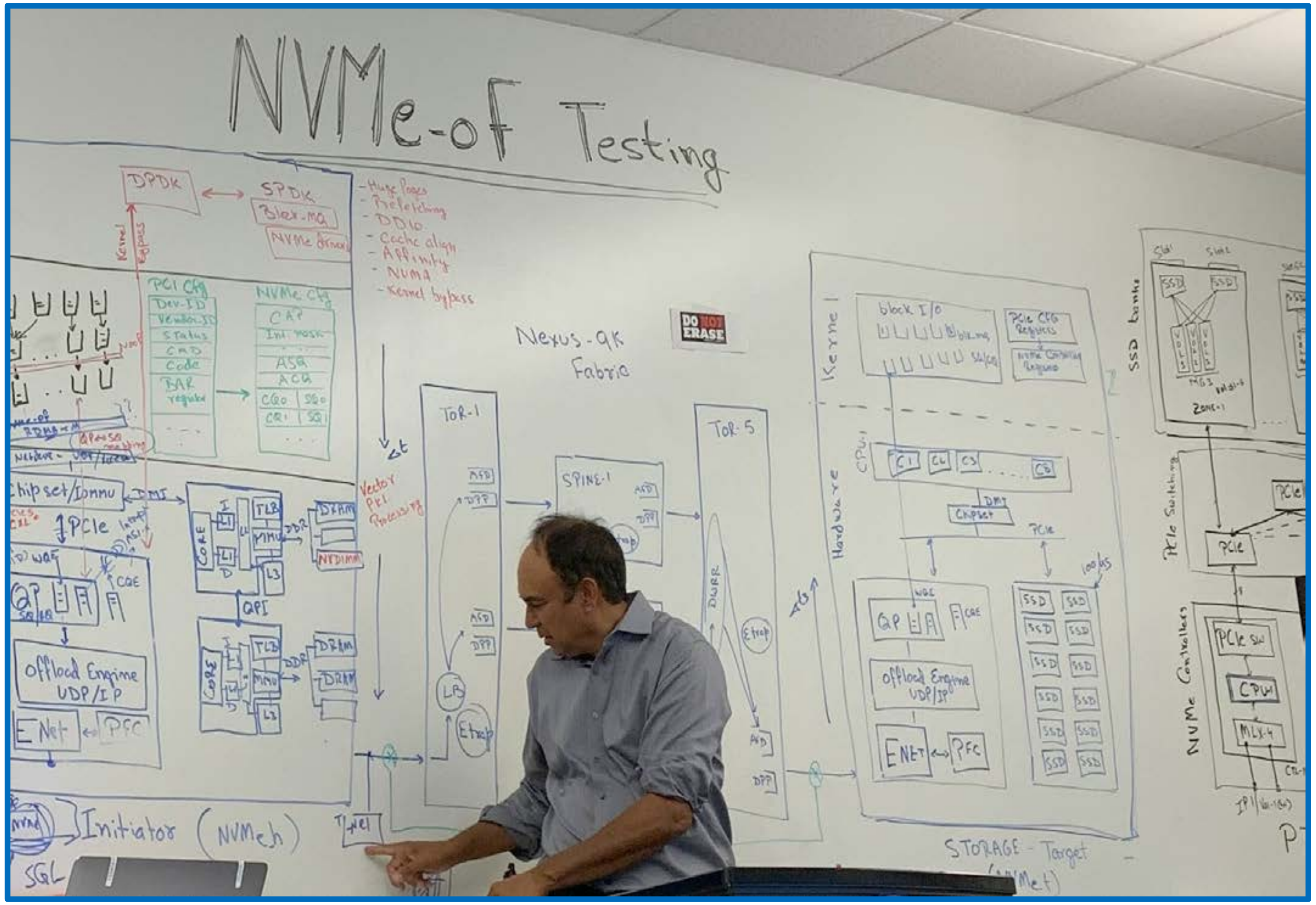



Kamal Bakshi
Director Storage

Dhans Kandhasamy
Director TME

Frank Wang
Technical Leader

Paresh Gupta
Technical Leader



Please send the corrections/feedback to me, kbakshi@cisco.com

STORAGE DEVELOPER CONFERENCE SDC 22

Other related sessions from the same presenter

https://www.youtube.com/watch?v=fl7eD3MLCK4

https://www.youtube.com/watch?v=flnsXmQiUHA

https://www.youtube.com/watch?v=WaeFo5L6lsw

****Cisco Storage Transport Fabric -NVMe Anywhere****

Learn how to deploy any NVMe-oF transport (NVMe/FC, NVMe/RoCEv2, NVMe/TCP) on Cisco networking fabric.

0-Cisco Nexus 9000 switch packet walk https://lnkd.in/gVqMm-TZ
1-Deploy NVMe Technology Anywhere https://lnkd.in/gV-g2kJ3
2-Cisco NVMe Storage Transport Solutions https://lnkd.in/gdb6SZ_M
3-Enabling NVMe-IP with Cisco Nexus 9k https://lnkd.in/gM_M8maw
4-Cisco ASIC On-Chip Smart Buffering https://lnkd.in/gHEsZmkv
5-Deep Dive Cisco MDS 64G ASIC https://lnkd.in/gth8VCnJ
6-Cisco NVMe-oF Demo: Overlay Transport https://lnkd.in/gcRcxi8
7-Cisco NVMe-oF Demo: Zero-Trust Security https://lnkd.in/gtNkQZ3
8-Cisco NVMe-oF Demo: Traffic Congestion https://lnkd.in/g9XSvYn
9-Cisco NVMe-oF Demo: Intent based NVMe https://lnkd.in/gpPFrHy
10-Cisco NVMe-oF Demo-Flow Analytics https://lnkd.in/gMagW3J
11-Cisco NVMe-oF Demo-DPP Prioritization https://lnkd.in/gPN9Kin
12-Cisco NVMe-oF Demo: RoCEv2 QoS: https://lnkd.in/gPiAbSR
13-Cisco NVMe-oF Demo: RoCEv2 PFC: https://lnkd.in/gnfaVaa
14-NVMe-RoCEv2 trouble shooting: Viavi https://lnkd.in/gcdX2BSS
15-NVMe packets analysis: Teledyne https://lnkd.in/gjxiizJ
16-Cisco Datacenter 400G packet walk https://lnkd.in/gD45_4Aa

Kamal Bakshi, Cisco

Please take a moment to rate this session.

Your feedback is important to us.

STORAGE DEVELOPER CONFERENCE

SDC 22

Additional Information

Appendix

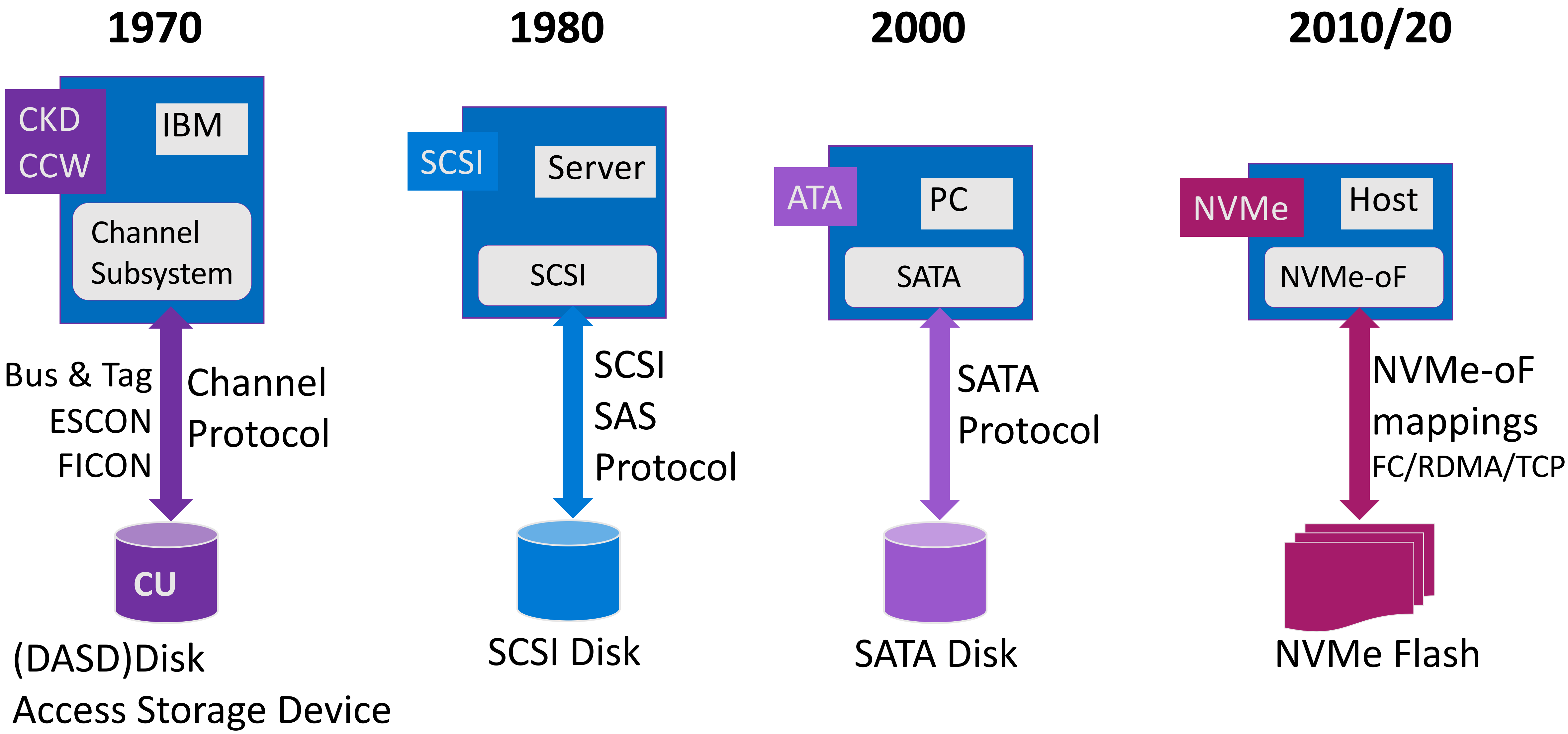
- NVMe Evolution
- NVMe/PCIe Architecture
- NVMe/FC Packets Example
- NVMe/RoCEv2 Architecture
- NVMe Advanced Features

Kamal Bakshi, Cisco

NVMe Evolution

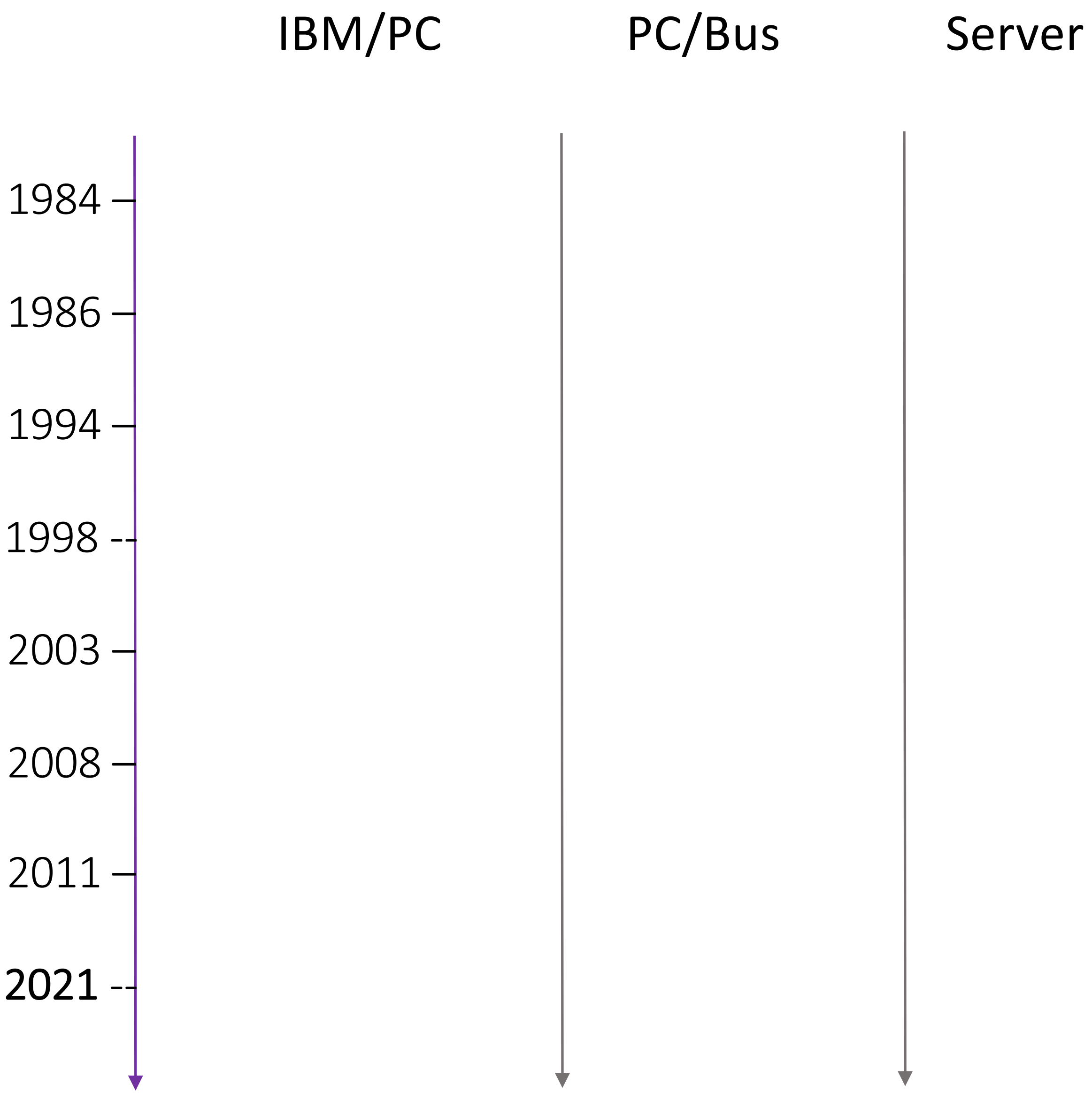
Kamal Bakshi, Cisco

50,000 feet view of NVMe

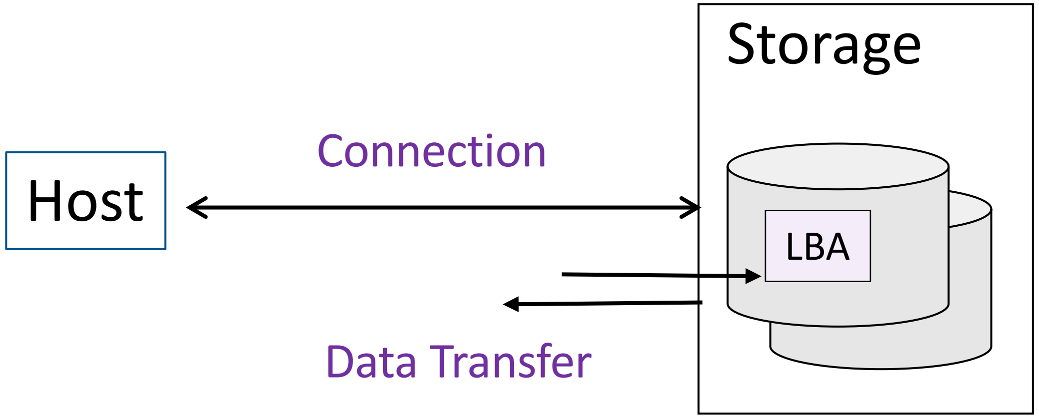


Kamal Bakshi, Cisco

Why NVMe ?

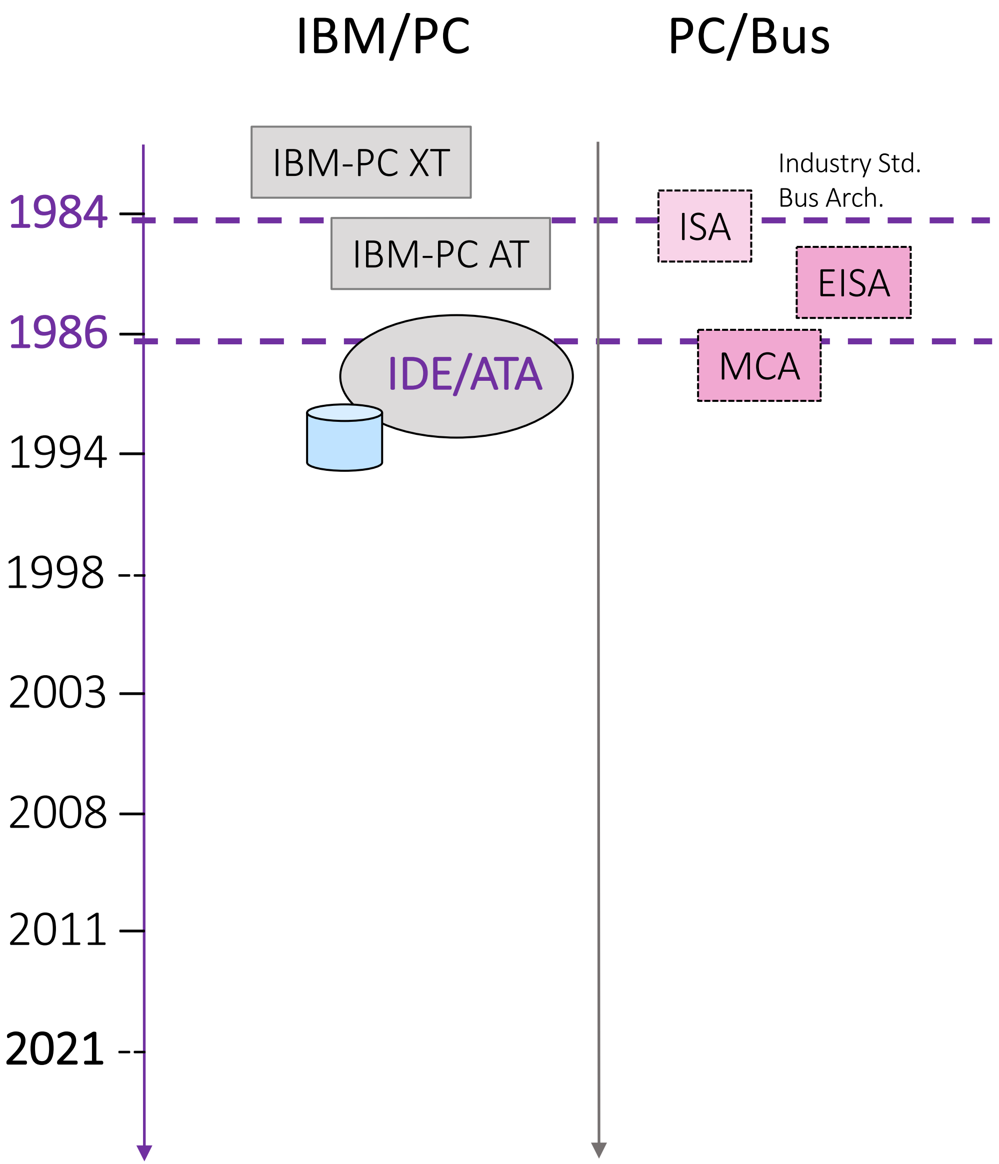


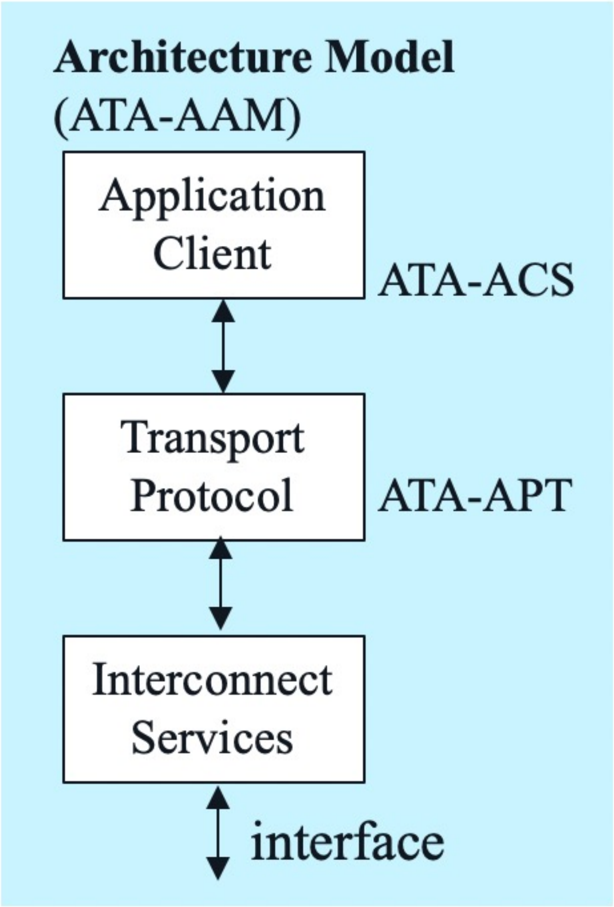
Problem Statement:
How to "connect", Host to the Storage, and do "Data Transfer" ?



Kamal Bakshi, Cisco

ATA (Advance Technology Attachment)



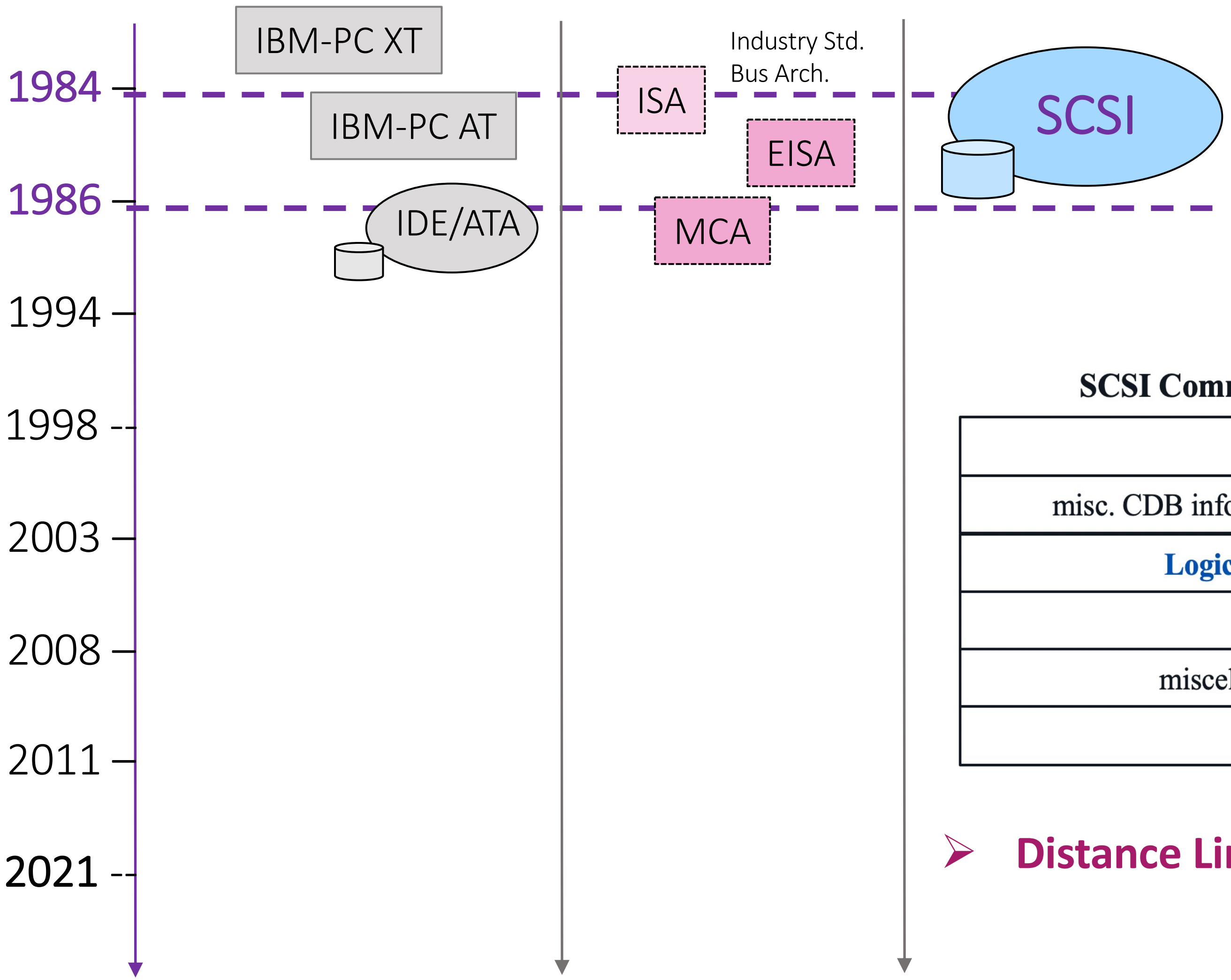


Write Sector Command Input

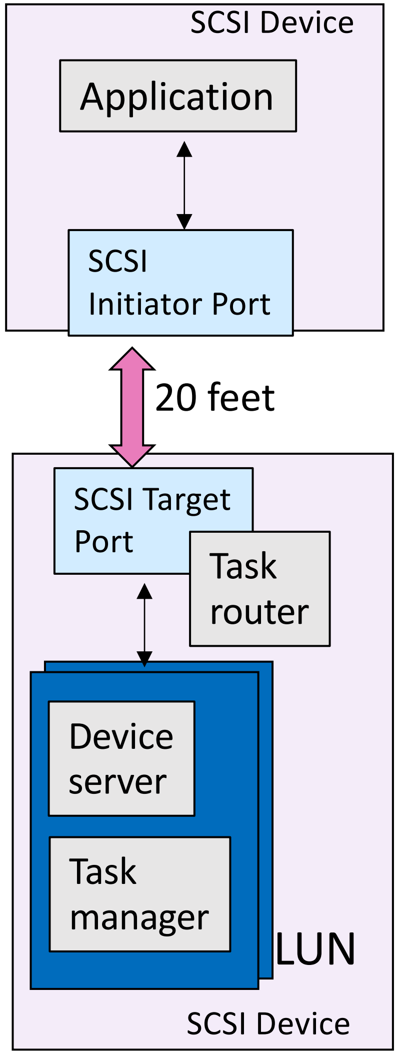
Field	Description
Feature	N/A
Count	# of Logical Sectors
LBA	Logical Block Address
Command	30h

Kamal Bakshi, Cisco

SCSI (Small Computer System Interface)



SCSI Architecture Model



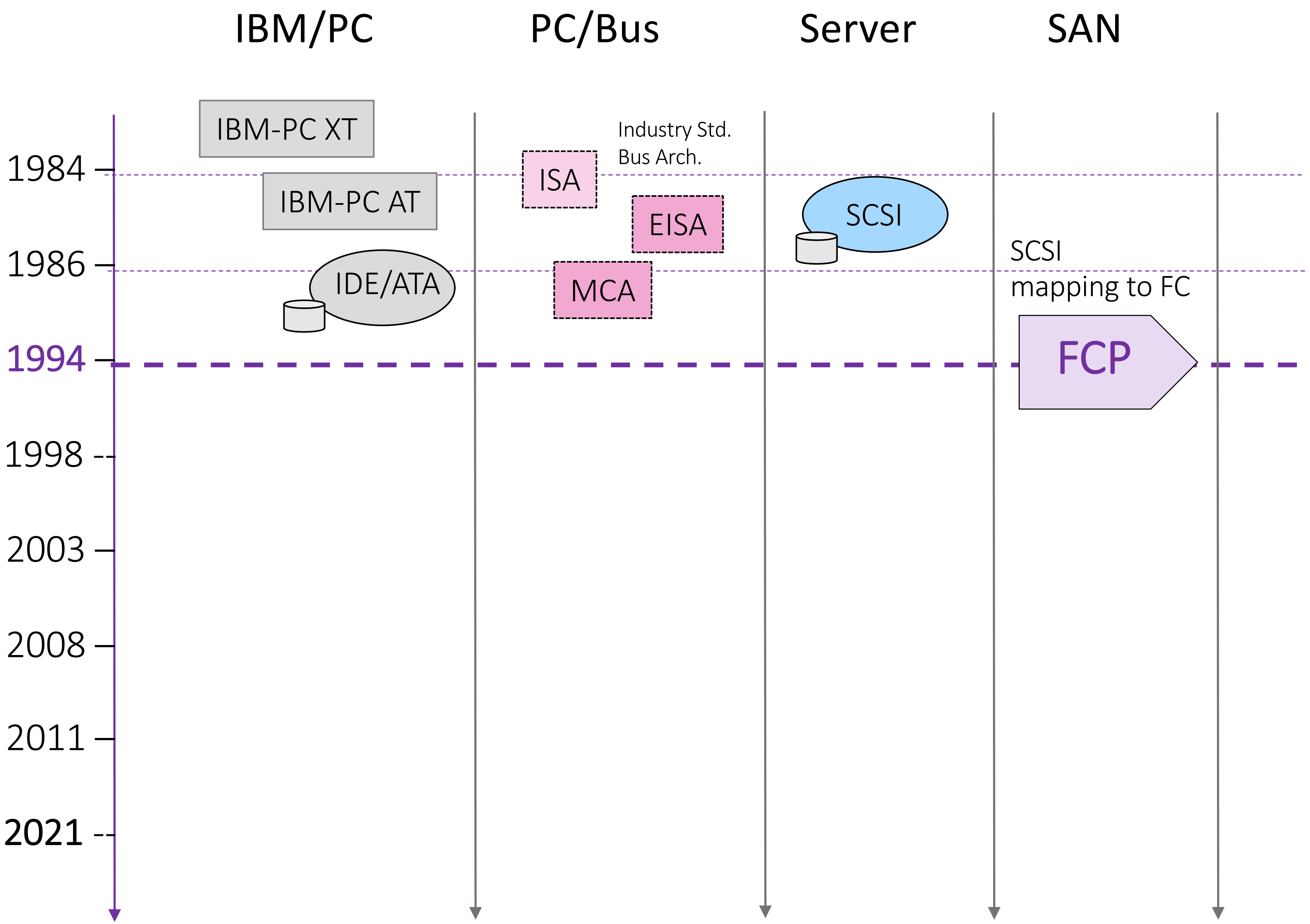
SCSI Command Descriptor Block -CDB

Operation Code	
misc. CDB information	Service Action (if any)
Logical Block Address (LBA)	
Transfer Length	
miscellaneous CDB information	
Control	

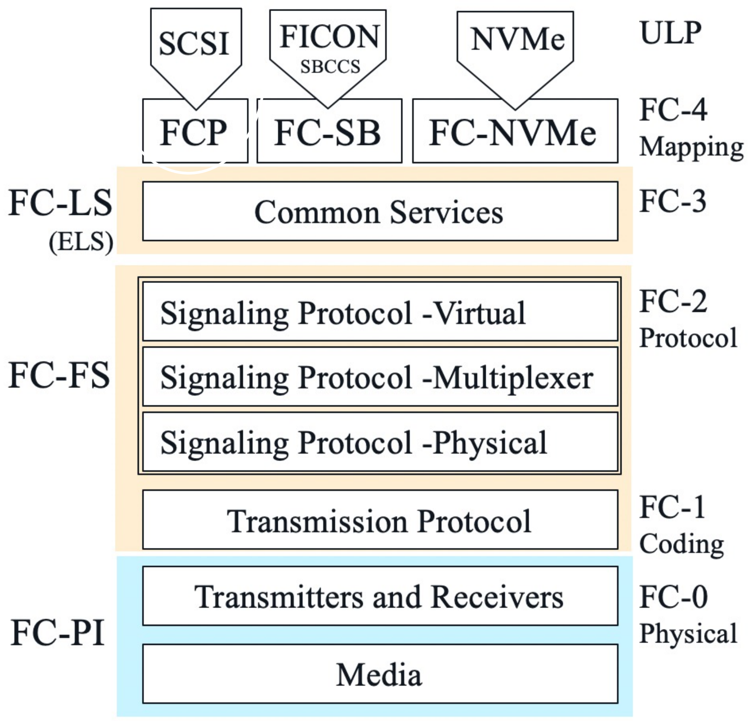
- **Distance Limitation with SCSI Cables**

Kamal Bakshi, Cisco

FCP (Fibre Channel Protocol)



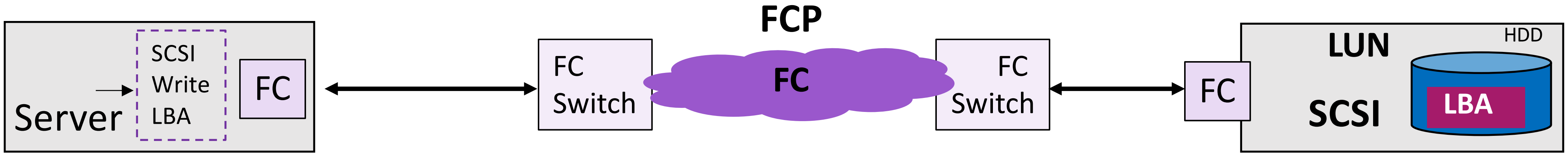
FCP -Fibre Channel Protocol



Fibre Channel Architecture

Kamal Bakshi, Cisco

FCP (SCSI Protocol mapped into Fibre Channel)



SOF	FC Frame header	Payload	CRC	EOF

SCSI WRITE (16) Command

Operation Code (8Ah)					
WRPROTECT	DPO	FUA	Rsvd	Obsolete	DLD2
Logical Block Address (LBA)					
Transfer Length					
DLD1	DLD0	Group Number			
Control					

FCP Command IU Payload

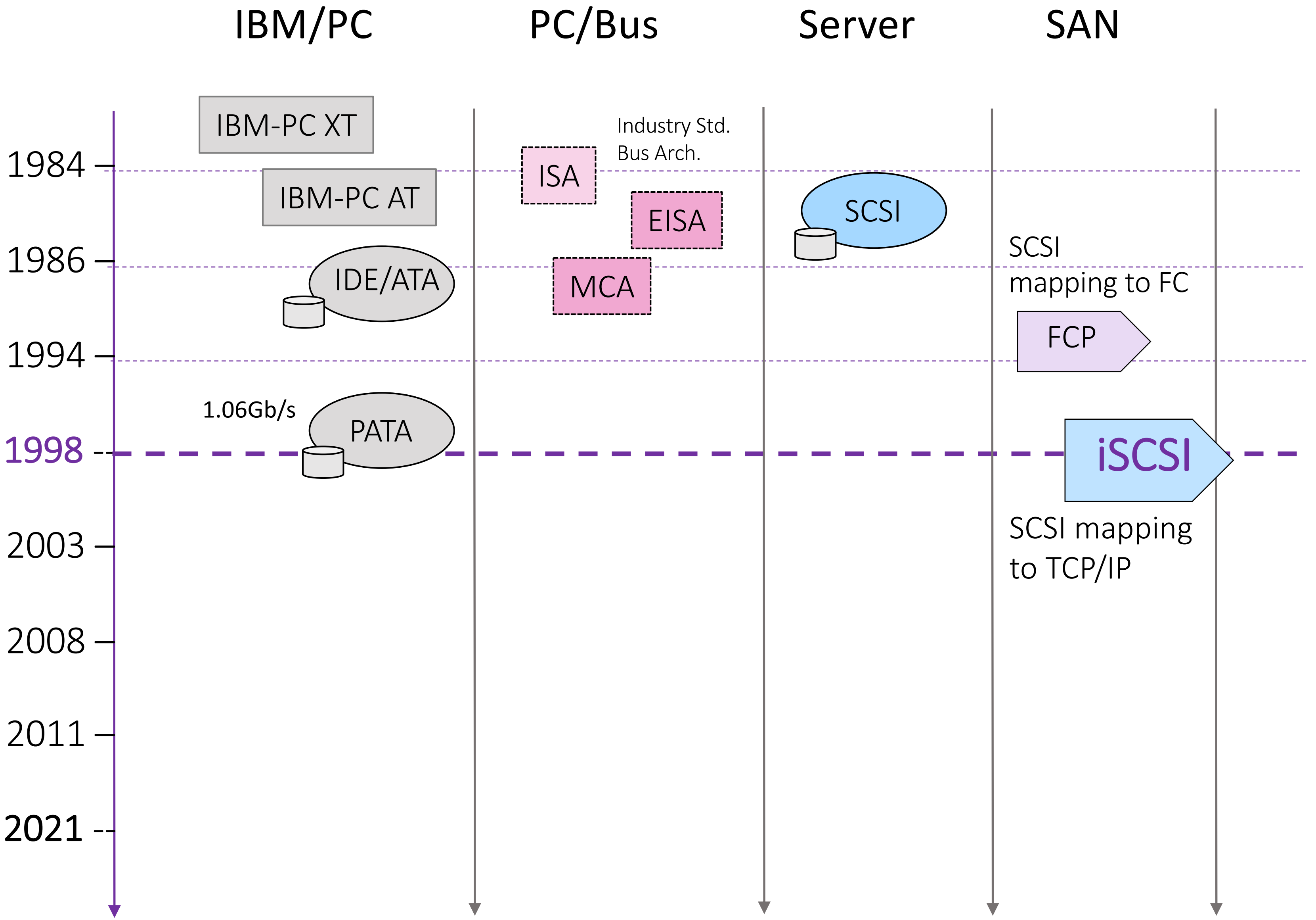
FCP_LUN		
Command Reference Number		
Rsvd	Command Priority	Task Attribute
Task Management Flags		
Additional FCP_CDB Length	RDDATA	WRDATA
FCP_CDB		
Additional FCP_CDB (if any)		
FCP_DL		
FCP_Bidirectional_Read_DL (if any)		

FC Frame Header

R_CTL	D_ID	
CS_CTL	S_ID	
TYPE	F_CTL	
SEQ_ID	DF_CTL	SEQ_CNT
OX_ID		RX_ID
Parameter		
FCP Payload		

Kamal Bakshi, Cisco

iSCSI (SCSI over TCP/IP)



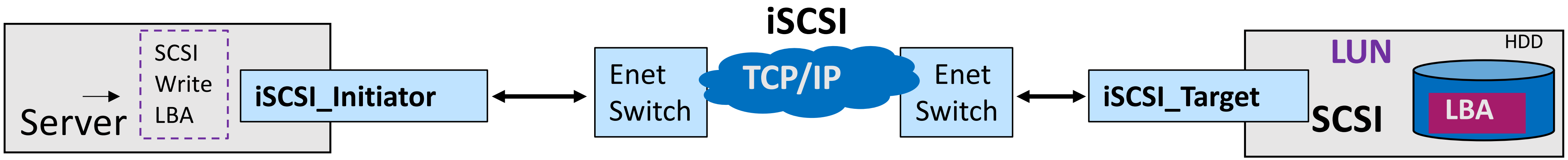
iSCSI Architecture

- iSCSI Initiator
- iSCSI Target
- "iqn" iSCSI Qualified Name
- Login/Logout
- Task Management
- iSNS Server (optional)
    - -Name Service
    - -Discovery Domain
    - -State Change Notification
- Single_queue / Multi_queue(recent)

Kamal Bakshi, Cisco

iSCSI (SCSI Protocol mapped into TCP/IP)

Issue: Limited Performance



Ethernet_hdr	IP_hdr	TCP_hdr	**iSCSI_PDU**	Ethernet_trailer

Port# 860,3260

SCSI WRITE (16) Command

Operation Code (8Ah)					
WRPROTECT	DPO	FUA	Rsvd	Obsolete	DLD2
Logical Block Address (LBA)					
Transfer Length					
DLD1	DLD0	Group Number			
Control					

SCSI Command PDU

Opcode (0x01)	Opcode specific flags
Total AHS length	Data Segment length
Logical Unit Number (LUN)	
Initiator Task Tag	
Expected Data Transfer Length	
Command Sequence Number	
ExpStatSN	
SCSI Command Descriptor Block (CDB)	

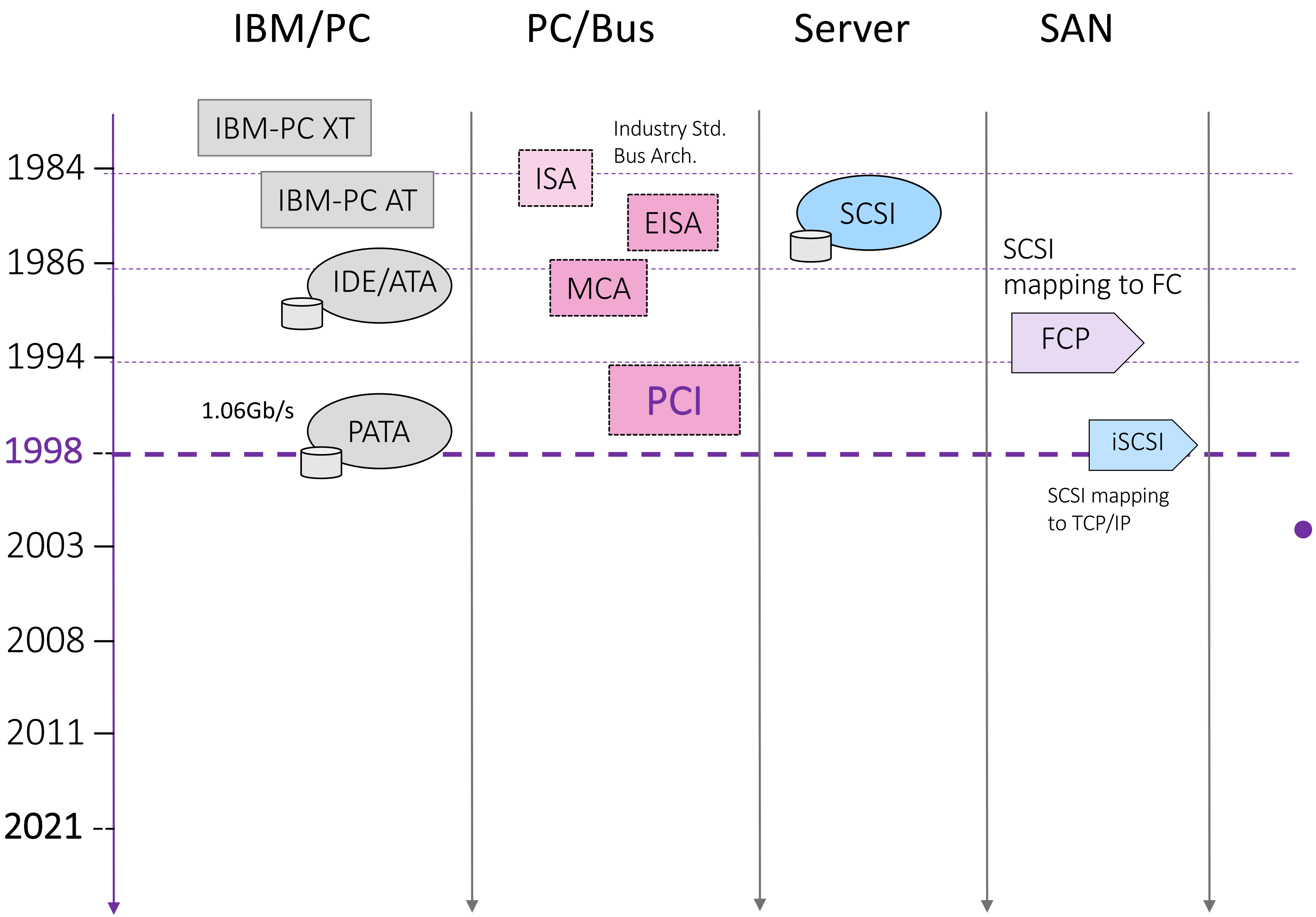
iSCSI PDU

Basic Header Segment (BHS)
Additional header Segments (AHS)*
Header-Digest*
Data Segment*
Data-Digest*

** Optional*

Kamal Bakshi, Cisco

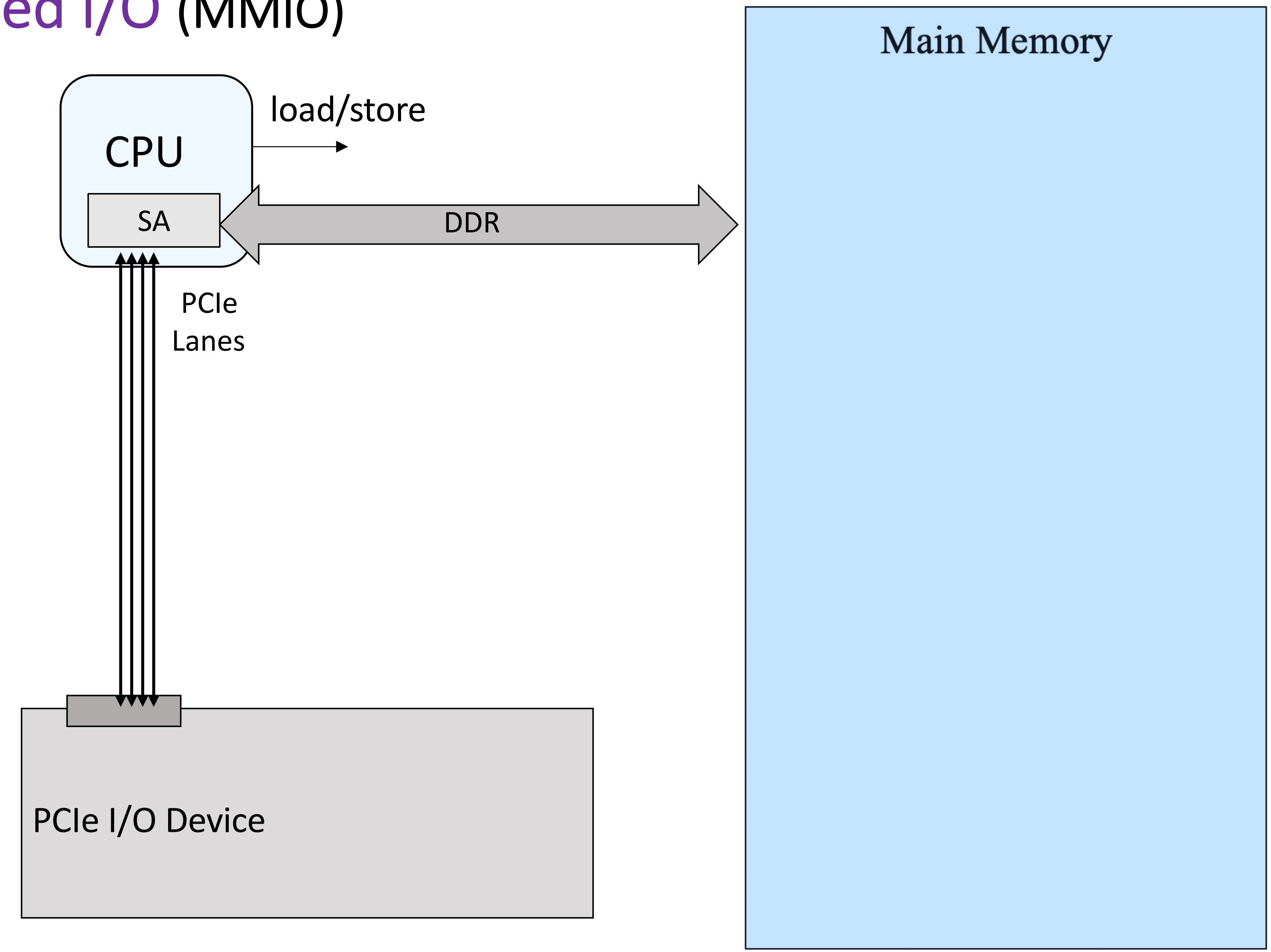
PCI (Peripheral Component Interconnect)



- **PCI Architecture**
  - **-Memory Mapped I/O**
  - **-PCI Config. Registers**
  - **-BAR space**
  - **-Capability Registers**
  - **-Message Signaled Interrupt**

PCIe Memory Mapped I/O (MMIO)

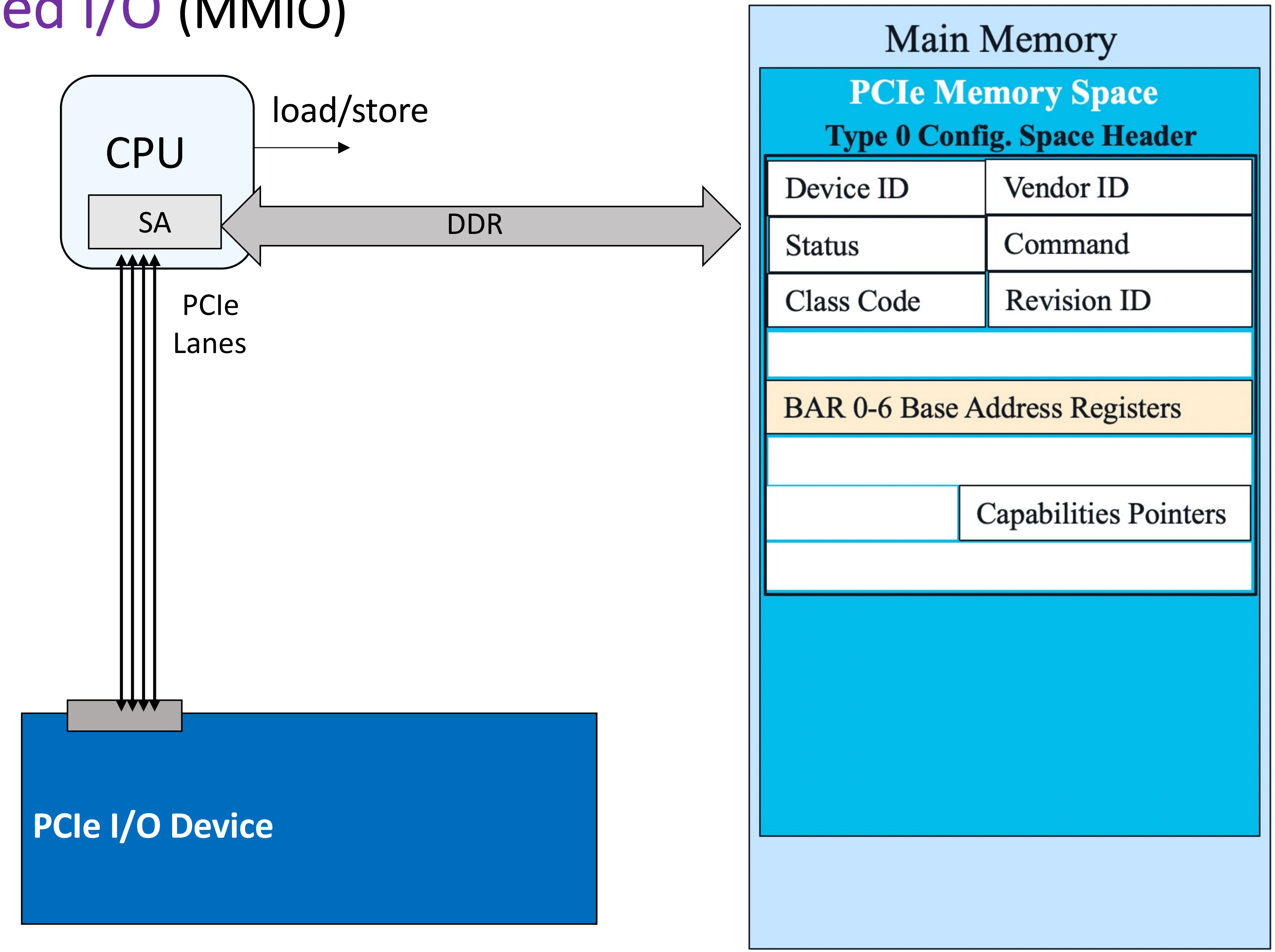
- **With MMIO I/O devices are directly mapped into CPU main memory.**
- **No special set of special CPU instructions needed.**



Kamal Bakshi, Cisco

PCIe Memory Mapped I/O (MMIO)

- With MMIO I/O devices are directly mapped into CPU main memory.
- No special set of special CPU instructions needed.
- **Each PCIe device has config. space in main memory.**

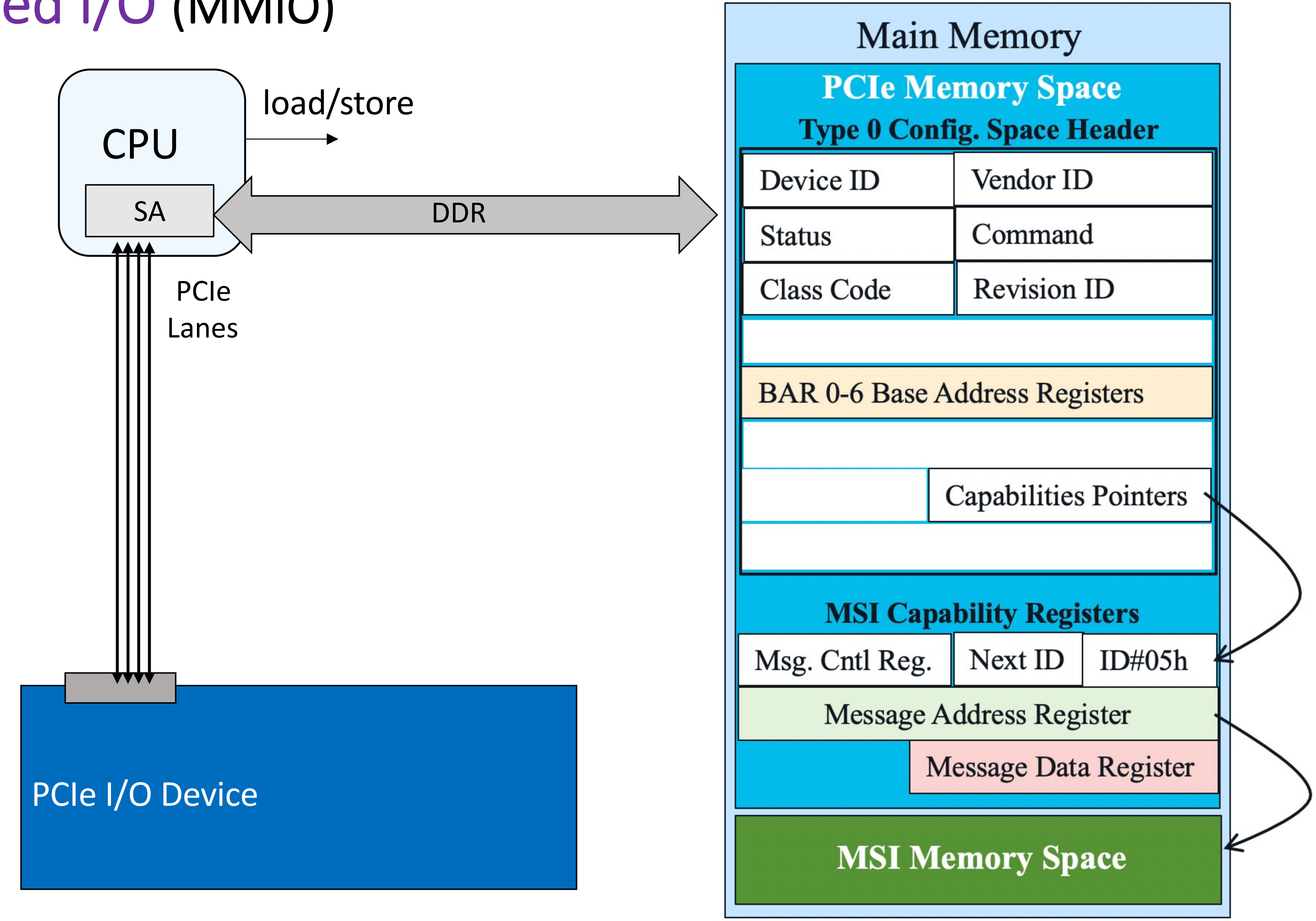


 |

Kamal Bakshi, Cisco

PCIe Memory Mapped I/O (MMIO)

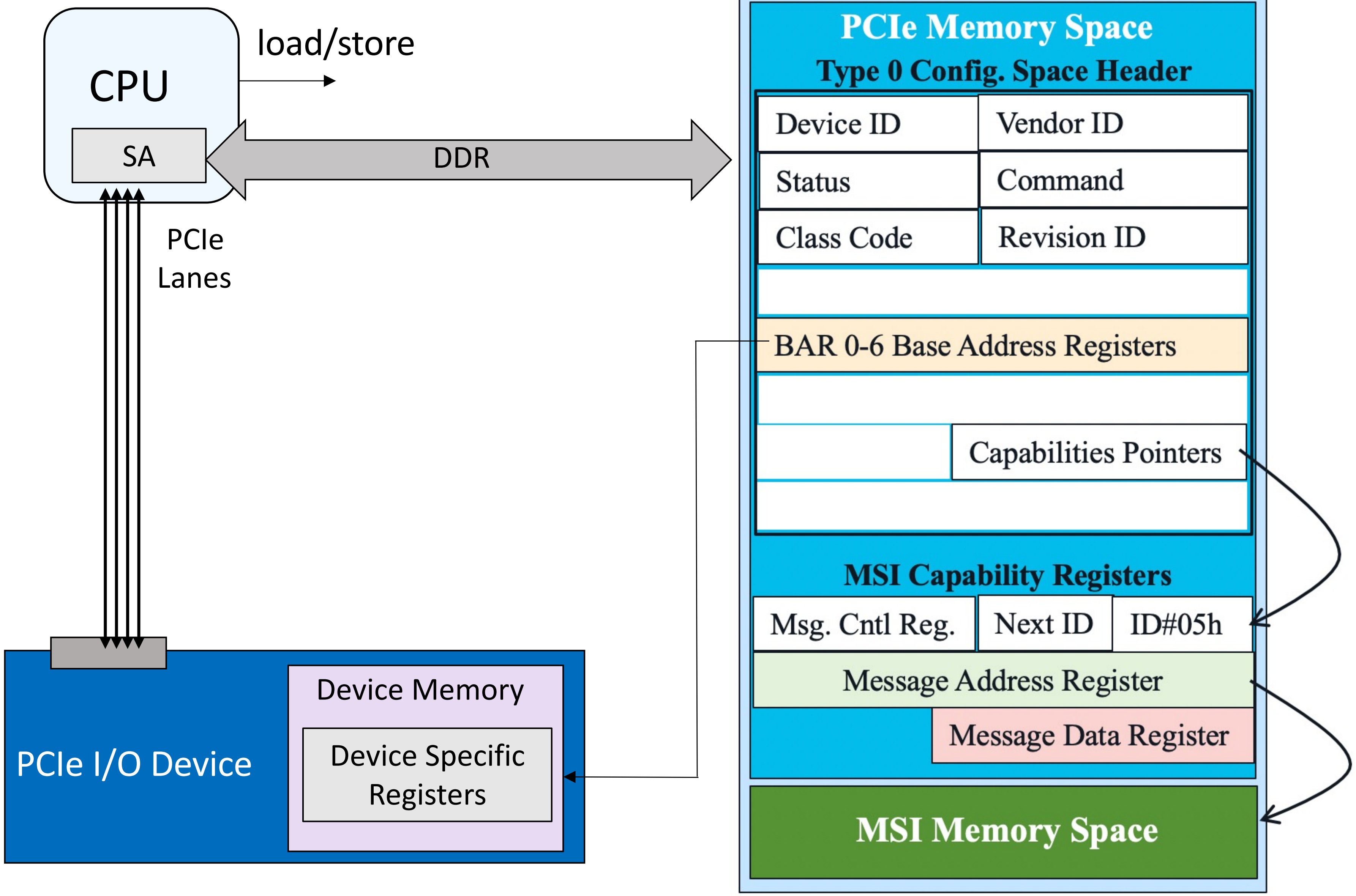
- With MMIO I/O devices are directly mapped into CPU main memory.
- No special set of special CPU instructions needed.
- **Each PCIe device has config. space in main memory.**



Kamal Bakshi, Cisco

PCIe Memory Mapped I/O (MMIO)

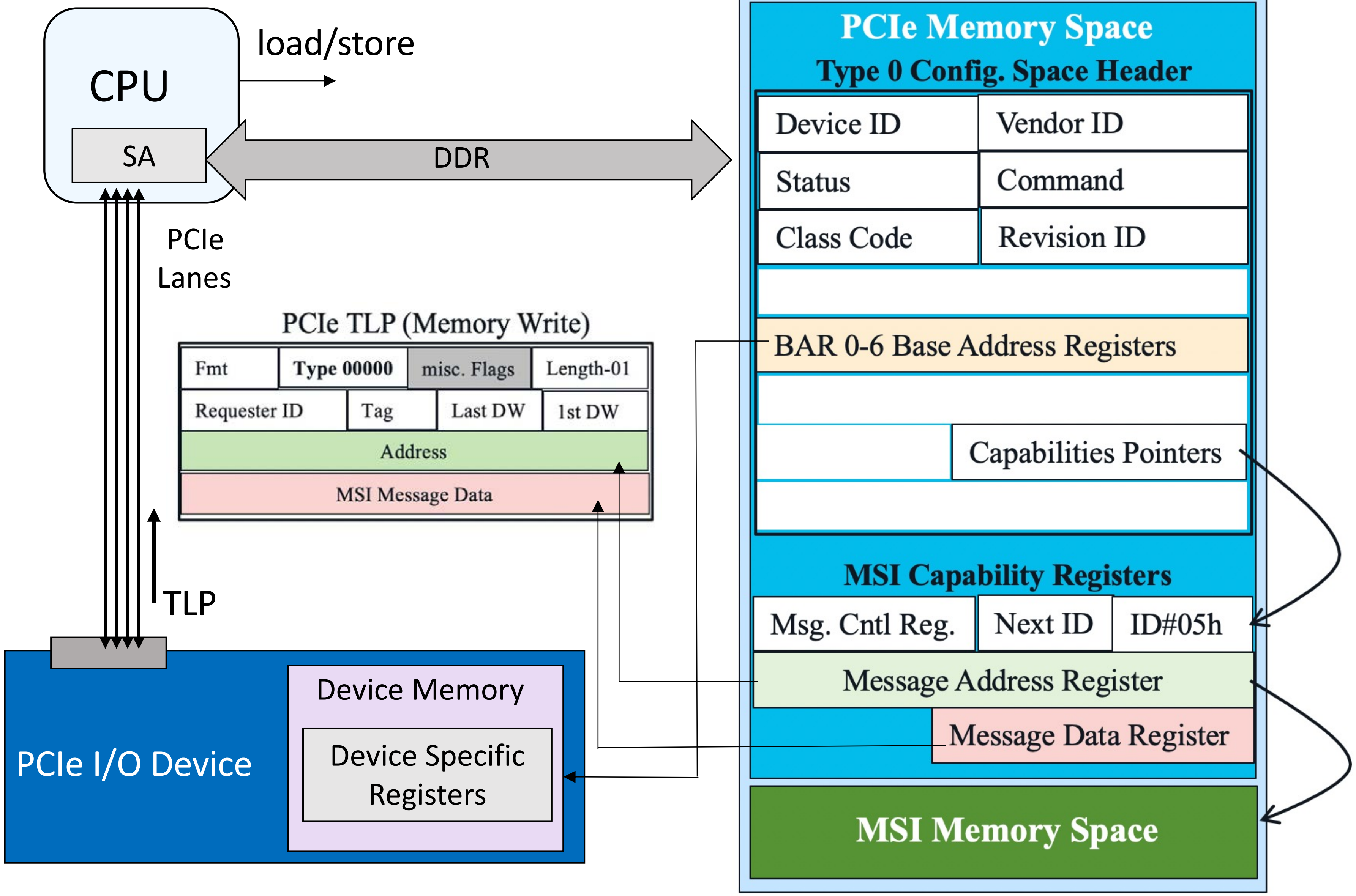
- With MMIO I/O devices are directly mapped into CPU main memory.
- No special set of special CPU instructions needed.
- Each PCIe device has config. space in main memory.
- **BAR registers map I/O device memory in the main memory**



Kamal Bakshi, Cisco

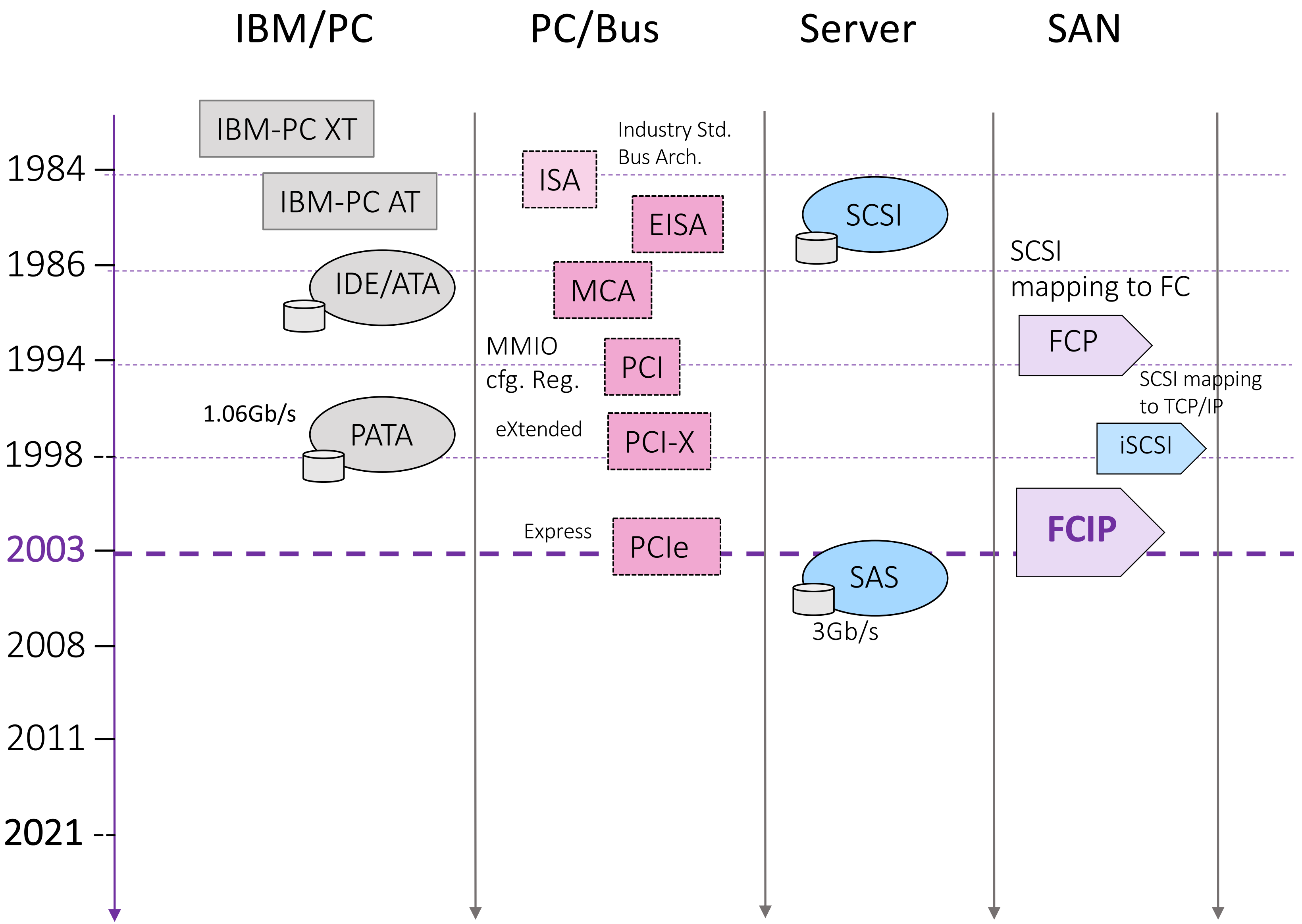
PCIe Memory Mapped I/O (MMIO)

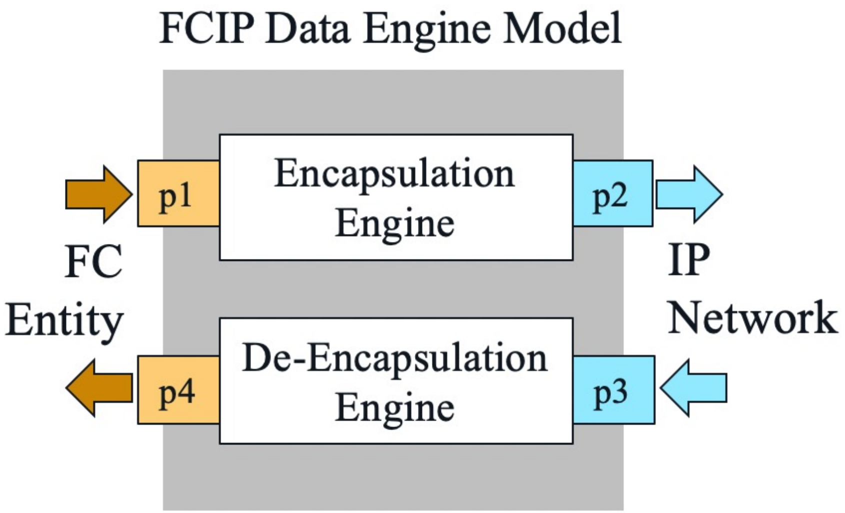
- With MMIO I/O devices are directly mapped into CPU main memory.
- No special set of special CPU instructions needed.
- Each PCIe device has config. space in main memory.
- BAR registers map I/O device memory in the main memory
- **MSI Message Signaled Interrupt**



Kamal Bakshi, Cisco

FCIP (Fibre Channel over IP)

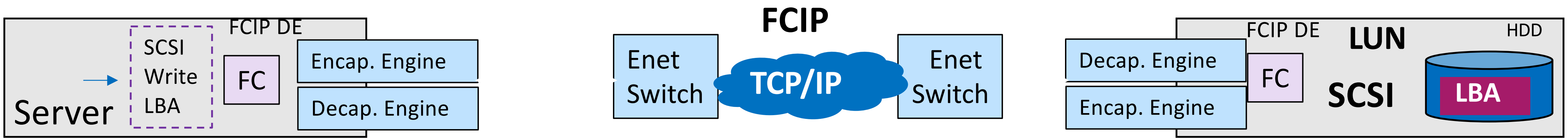




FCIP Architecture

Kamal Bakshi, Cisco

FCIP (FC Encapsulated inside TCP/IP Protocol)



Ethernet_hdr	IP_hdr	TCP_hdr	**FCIP_Header**	SOF	FC Frame header	Payload	CRC	EOF	Ethernet_trailer

Port# 3225

FCIP Header

DW0	Protocol #	Version	-Protocol #	-Version
DW1	Protocol #	Version	-Protocol #	-Version
DW2	pFlags	Reserved	-pFlags	-Reserved
DW3	Flags	Frame Length	-Flags	-Frame Length
DW4	Time Stamp (seconds)			
DW5	Time Stamp (second fraction)			
DW6	CRC			

FCP Header

R_CTL	D_ID	
CS_CTL	S_ID	
TYPE	F_CTL	
SEQ_ID	DF_CTL	SEQ_CNT
OX_ID		RX_ID
Parameter		

FCP Command IU Payload

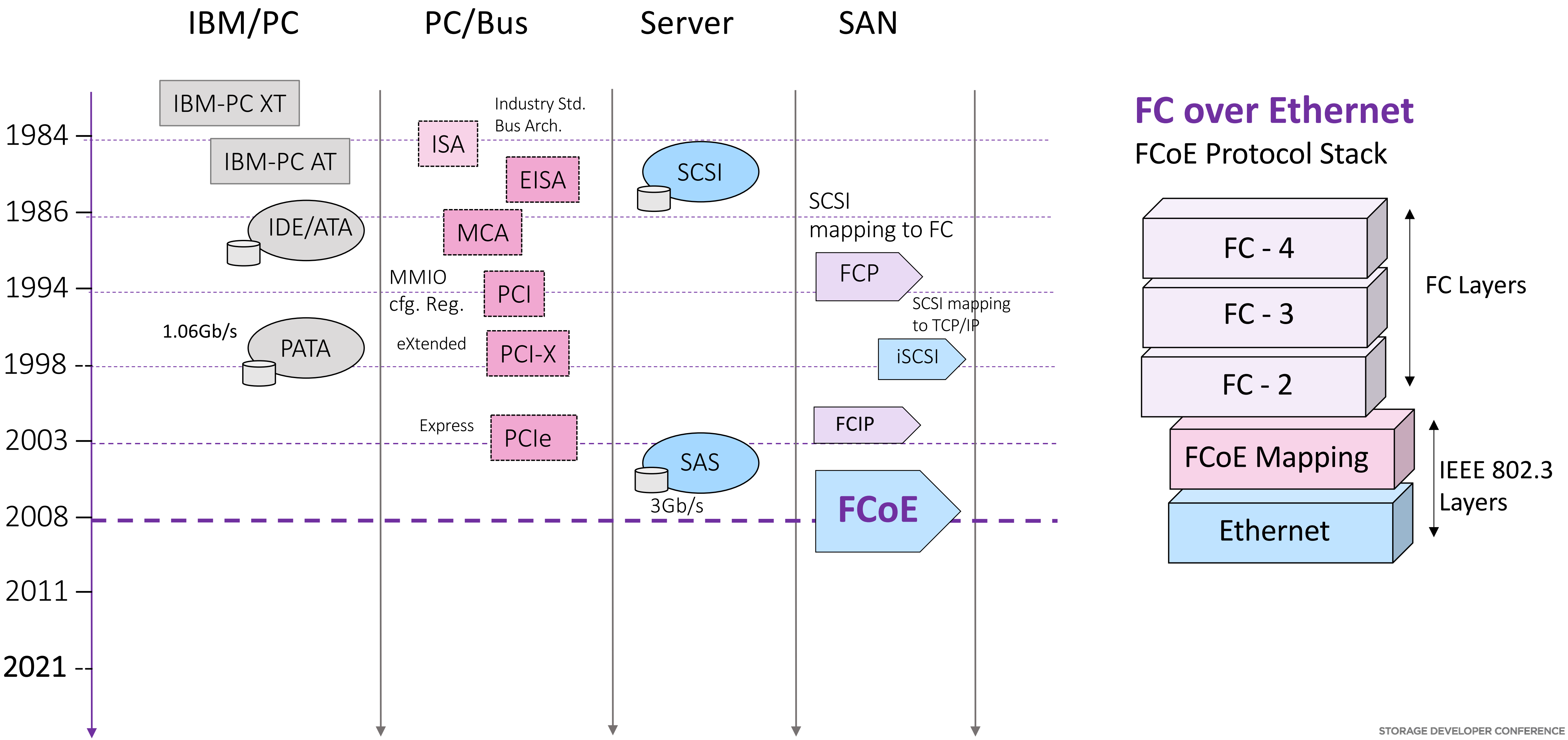
FCP_LUN		
Command Reference Number		
Rsvd	Command Priority	Task Attribute
Task Management Flags		
Additional FCP_CDB Length	RDDATA	WRDATA
FCP_CDB		
Additional FCP_CDB (if any)		
FCP_DL		
FCP_Bidirectional_Read_DL (if any)		

SCSI WRITE (16) Command

Operation Code (8Ah)					
WRPROTECT	DPO	FUA	Rsvd	Obsolete	DLD2
Logical Block Address (LBA)					
Transfer Length					
DLD1	DLD0	Group Number			
Control					

Kamal Bakshi, Cisco

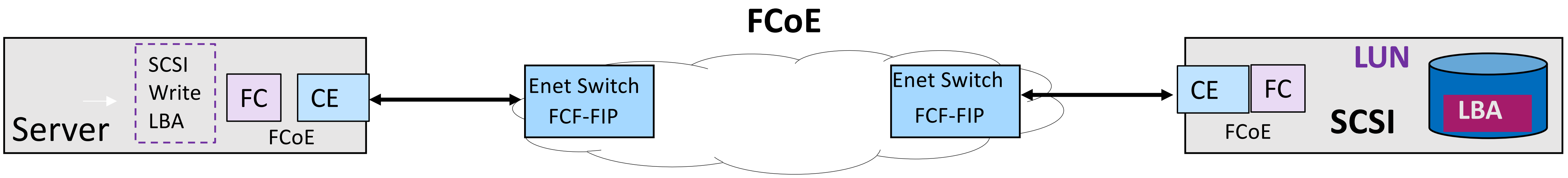
FCoE (Fibre Channel over Ethernet)



Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

FCoE (Fibre Channel over Ethernet)



FIP (FCoE Initialization Protocol) Frame Format

Enet_DA	Enet_SA	802.1Q	Type=FIP	version	rsvd	Encapsulated FIP operation (n words)	FCS
			0x8914				

FIP Protocol Code	Reserved	FIP Subcode

FIP Descriptor List Length	FP	SP	r	C	D	RP	A	S	F

FIP Descriptor List

FIP_Pad

FIP Operation (code/subcode)

0001/01h Discovery Solicitation
0001/02h Discovery Advertisement
0002/01h Virtual Link Inst. Request
0002/02h Virtual Link Inst. Reply
0003/01h FIP Keep Alive
0003/02h FIP Clear Virtual Links
0004/01h FIP VLAN Request
0004/02h FIP VLAN Notification
0004/03h FIP VN2VN VLAN
0005/01h N_Port_ID Probe Request
0005/02h N_Port_ID Probe Reply
0005/03h N_Port_ID Claim Notification
0005/04h N_Port_ID Claim Response
0005/05h N_Port_ID Beacon
FFF8h - FFFEh Vendor Specific

FIP Descriptor Types

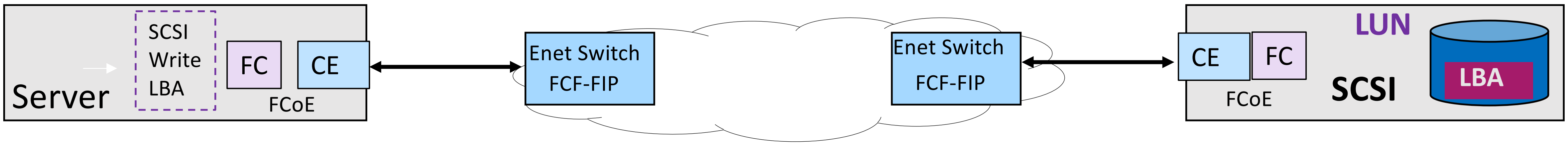
0-Reserved, 1-Priority, 2-MAC Address 3-FC_MAC, 4-Name_Identifier, 5-Fabric, 6-Max FCoE Size, 7-FLOGI, 8-NPIV FDISC, 9-LOGO, 10-ELP, 11-Vx_Port ID, 12-EKA_ADV_Period, 13-Vendor_ID, 14-VLAN, 15-VN2VN Attributes, 16-127 Reserved, 128-Clear Virtual Links Reason Code.

Kamal Bakshi, Cisco

FCoE (Fibre Channel over Ethernet)

Issue: Scaling of FCoE protocol to multi-hops

FCoE



FIP (FCoE Initialization Protocol) Frame Format

Enet_DA	Enet_SA	802.1Q	Type=FIP	version	rsvd	Encapsulated FIP operation (n words)	FCS
			0x8914				

FCoE Frame Format

Enet_DA	Enet_SA	802.1Q	Type=FCoE	version	rsvd	SOF	FC Frame header	Payload	CRC	EOF	rsvd	FCS
			0x8906									

SCSI WRITE (16) Command

Operation Code (8Ah)					
WRPROTECT	DPO	FUA	Rsvd	Obsolete	DLD2
Logical Block Address (LBA)					
Transfer Length					
DLD1	DLD0	Group Number			
Control					

FCP Command IU Payload

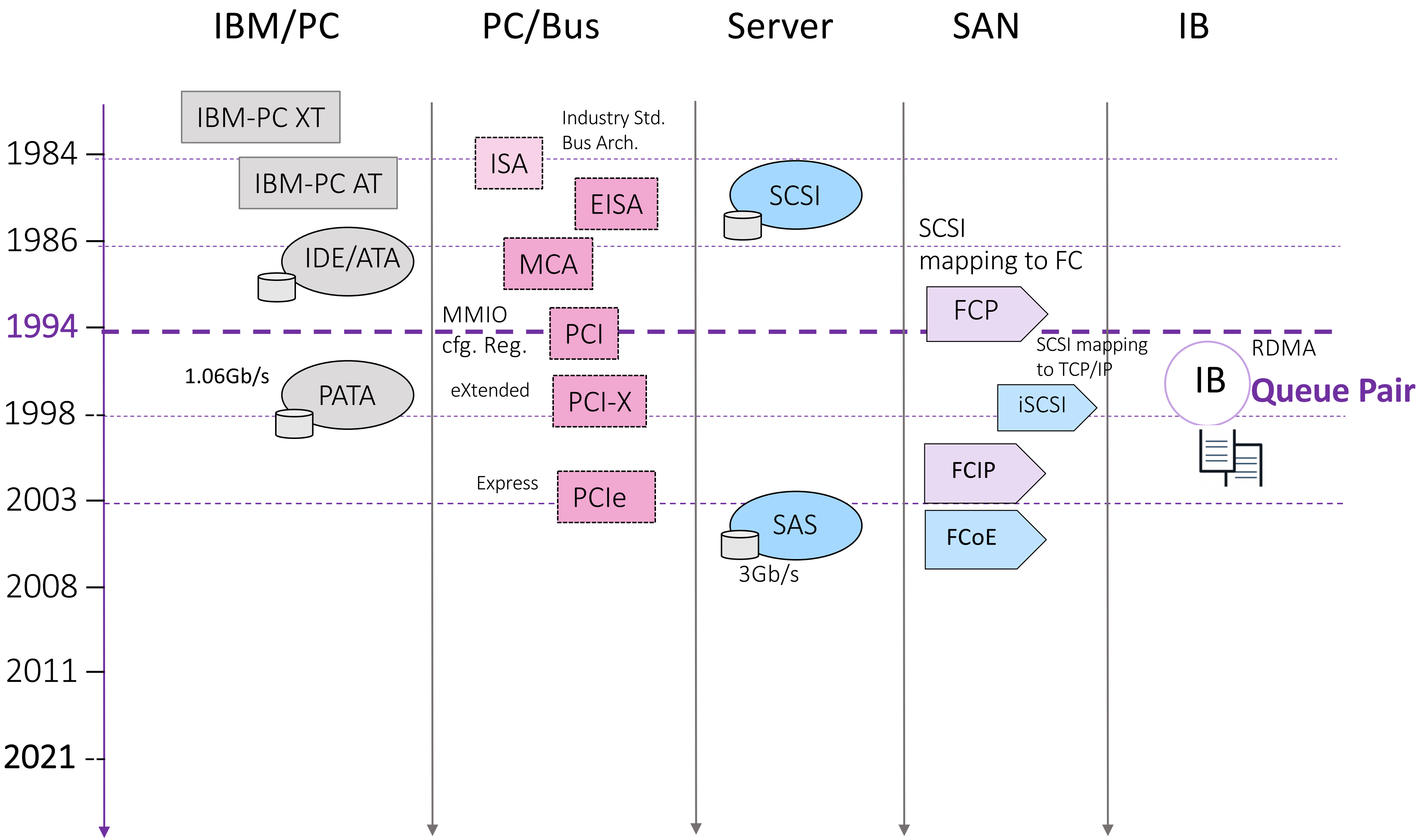
FCP_LUN		
Command Reference Number		
Rsvd	Command Priority	Task Attribute
Task Management Flags		
Additional FCP_CDB Length	RDDATA	WRDATA
FCP_CDB		
Additional FCP_CDB (if any)		
FCP_DL		
FCP_Bidirectional_Read_DL (if any)		

FC Frame Header

R_CTL	D_ID	
CS_CTL	S_ID	
TYPE	F_CTL	
SEQ_ID	DF_CTL	SEQ_CNT
OX_ID	RX_ID	
Parameter		
FCP Payload		

Kamal Bakshi, Cisco

IB (InfiniBand)



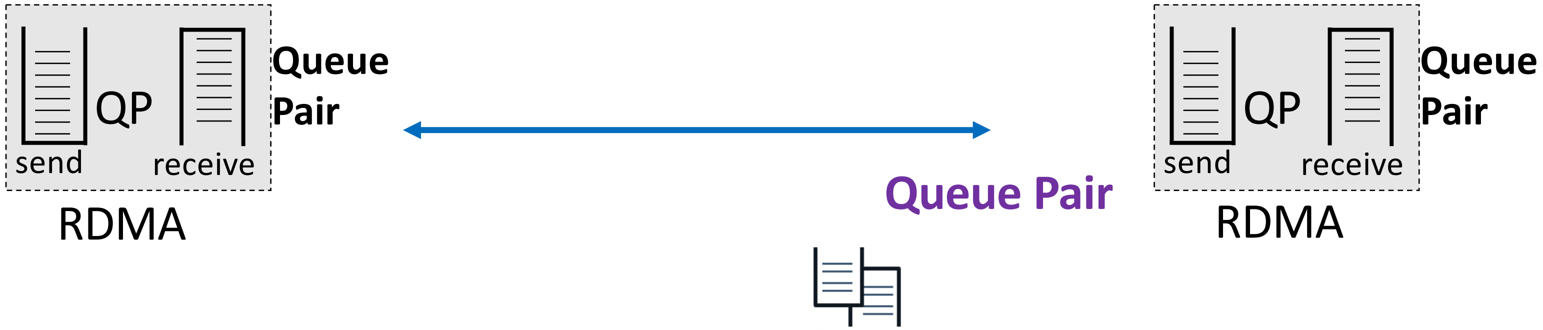
"Initially the IBTA vision for IB was simultaneously a replacement for PCI in I/O, Ethernet in the machine room, cluster interconnect and Fibre Channel."

...Wikipedia

 |

Kamal Bakshi, Cisco

InfiniBand (Queue Pair based Remote Direct Memory Access)

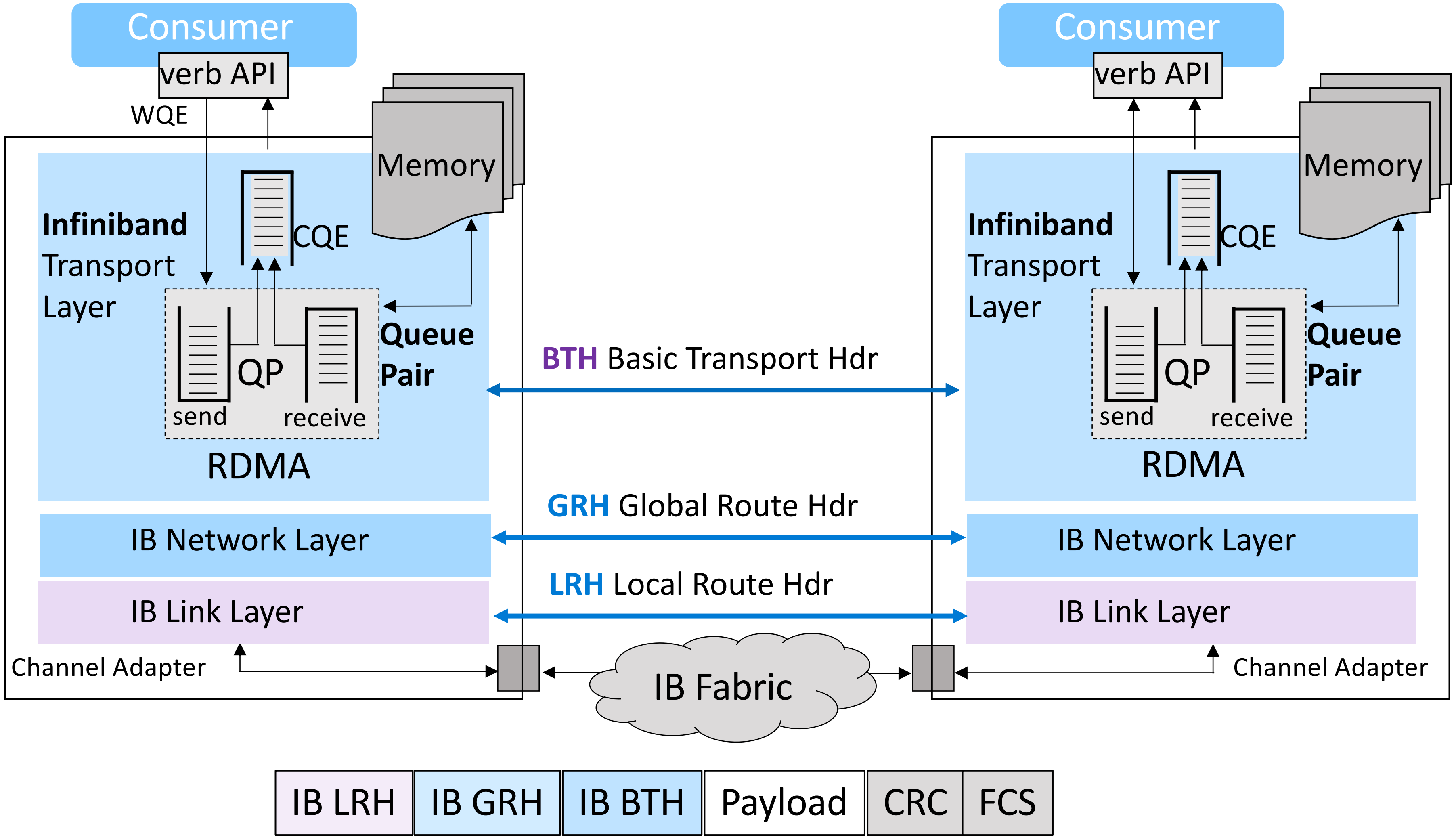


Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

InfiniBand (Queue Pair based Remote Direct Memory Access)

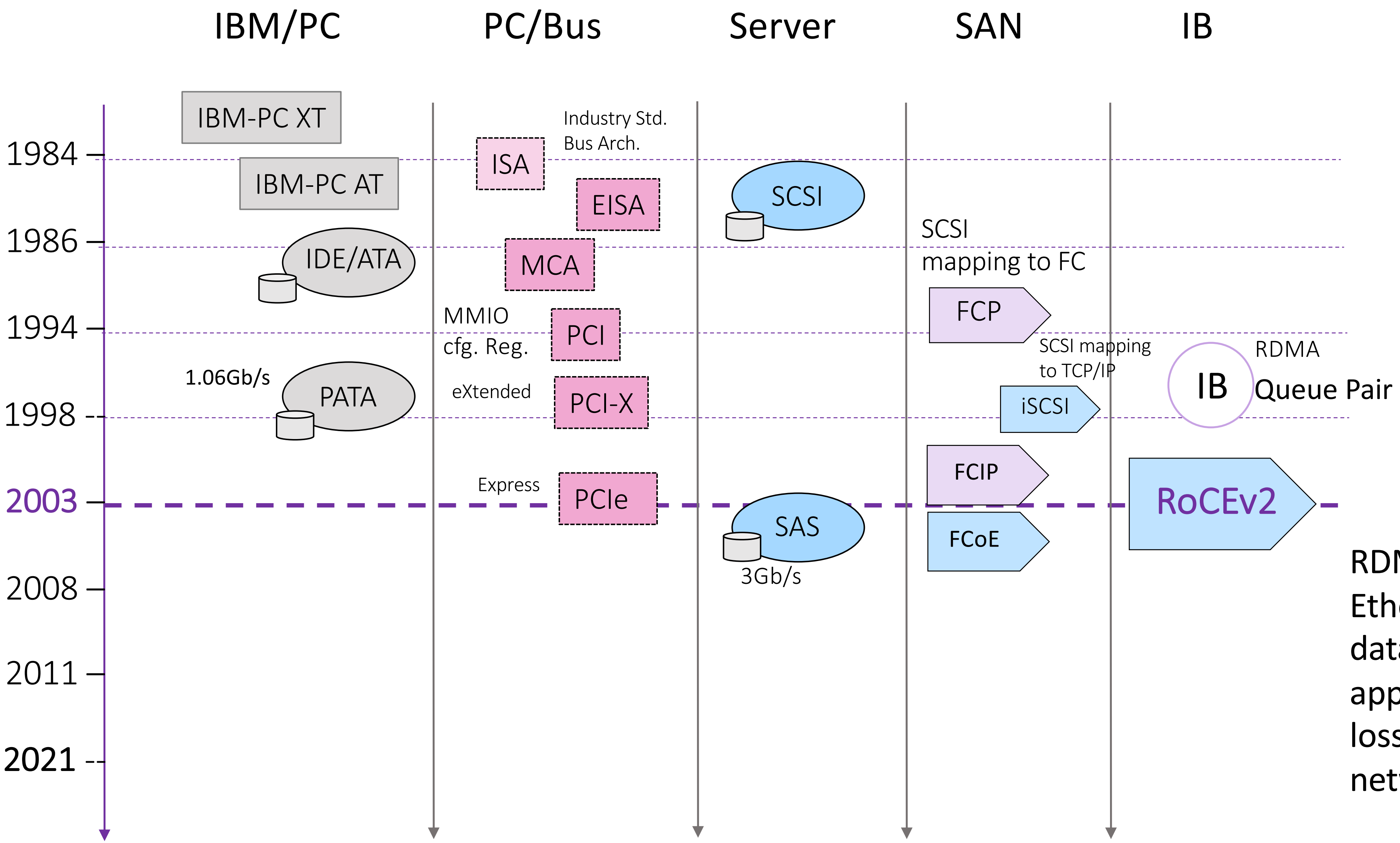
- -Verb API
- -RDMA Read/Write
- -Message Send/Receive
- -Kernel Bypass
- -Queue Pair
- -Completion Queue
- -Work Queue Element



Infiniband Packet

Kamal Bakshi, Cisco

RoCEv2 (RDMA over Converged Ethernet)

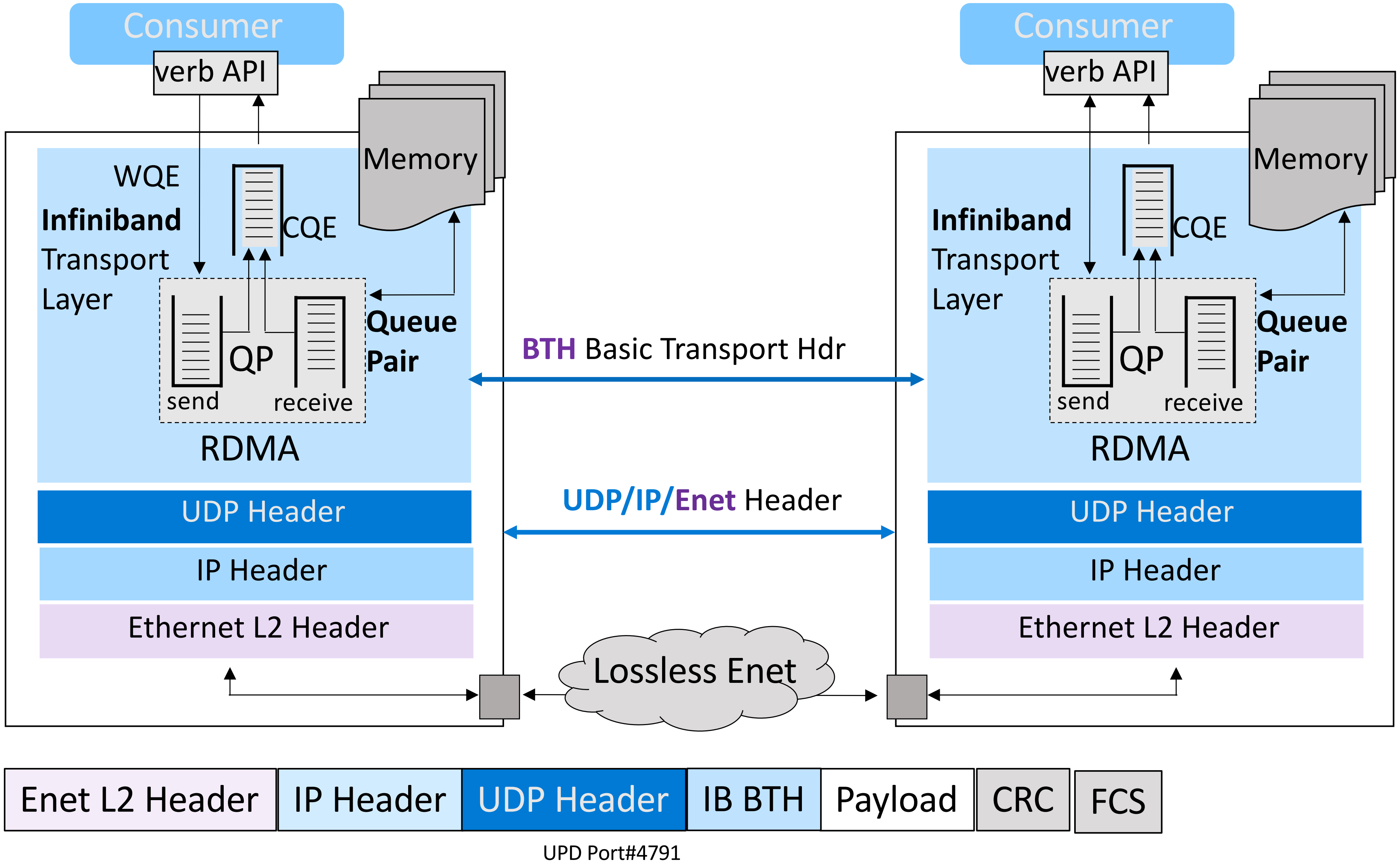


RDMA over converged Ethernet protocol allows data transfer between application memory over lossless Ethernet networks.

Kamal Bakshi, Cisco

RoCEv2 (Architecture)

- -Lossless Ethernet
- -PFC
- -ECN
- -DCQCN
- -CNP (IBTH)
- -Resilient RoCEv2

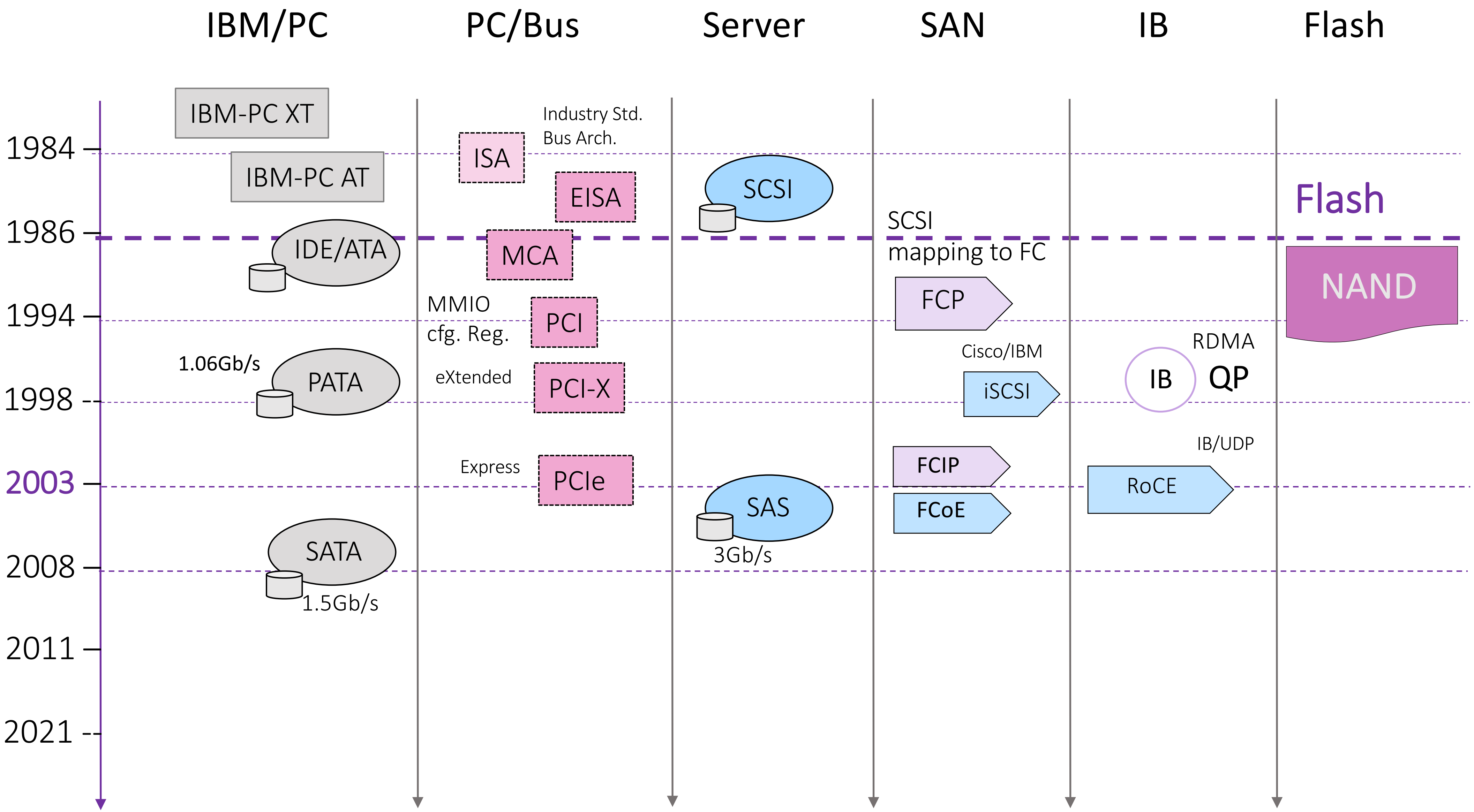


RoCEv2 Packet

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

Flash (Non Volatile Memory)



Kamal Bakshi, Cisco

Flash (Non Volatile Memory)

"Flash memory is an electronic non-volatile computer memory storage medium that can be electrically erased and reprogrammed. The two main types of flash memory, NOR flash and NAND flash, are named for the NOR and NAND logic gates." …Wikipedia

NOR vs NAND:

NOR flash is faster to read but takes longer to write or erase and is mostly used in consumer devices like smartphones. **NAND has higher capacity and is cheaper as compared to NOR.**

3D/V-NAND (Levels/Layers)

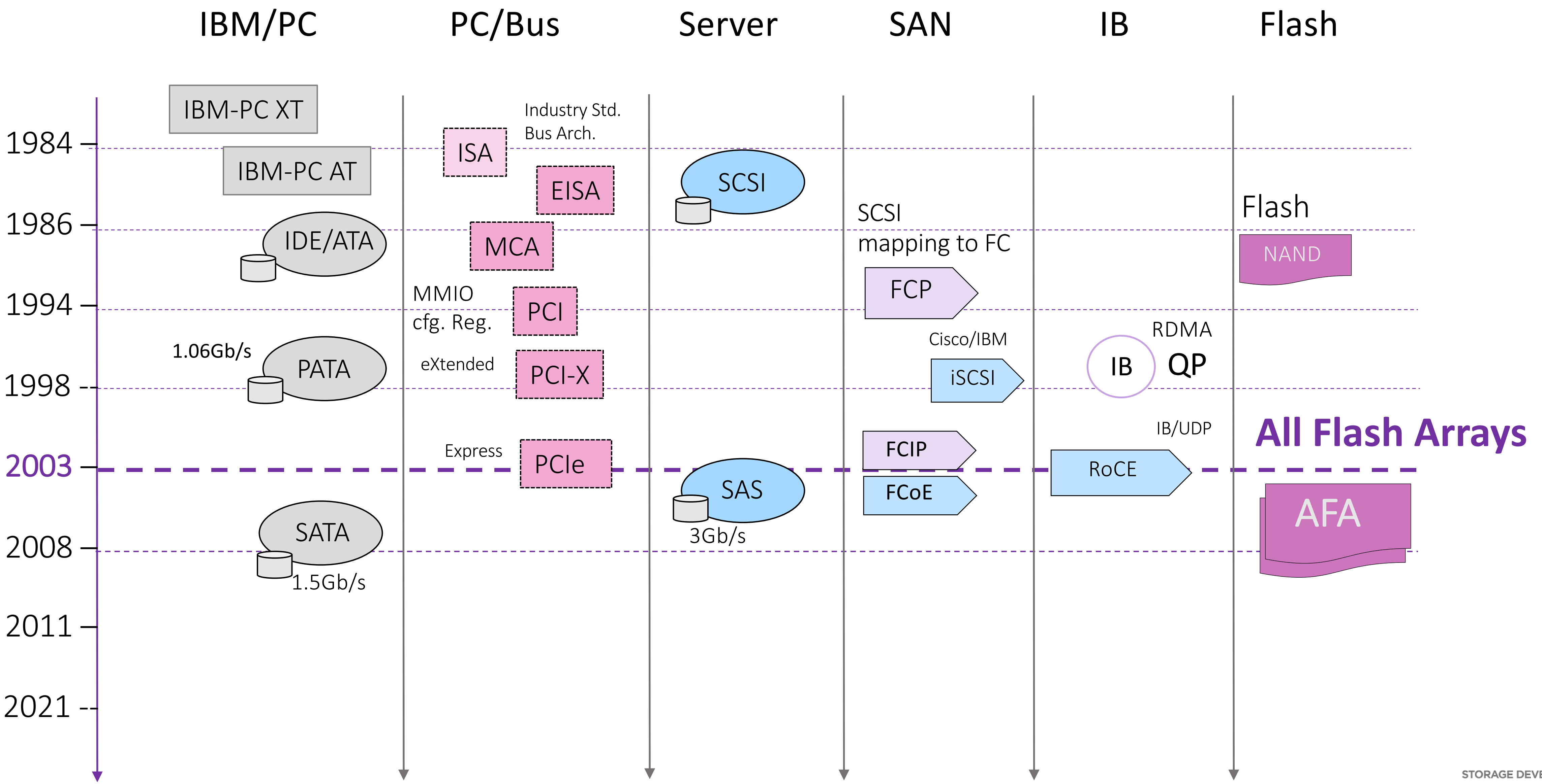
- SLC single level cell stores one bit per cell, MLC multi level cell stores two bits per cell, TLC triple level cell stores three bits per cell, QLC quad level cell stores 4 bits per cell.
- In 2D/planner NAND memory cells are connected in horizontal fashion but in 3D NAND they are stacked vertically in layers. (48, 64, 96, 128…**144-230**…256-layers…1000-layers!)

Storage Class Memory -SCM

- PCRAM: Phase Change Random Access Memory (Intel/Optane is based on PCRAM)
- ReRAM: Resistive Random-Access Memory
- MRAM: Magnetic Random-Access Memory
- STT-MRAM: Spin-Transfer Torque Magnetic Random-Access Memory
- Z-NAND: Samsung

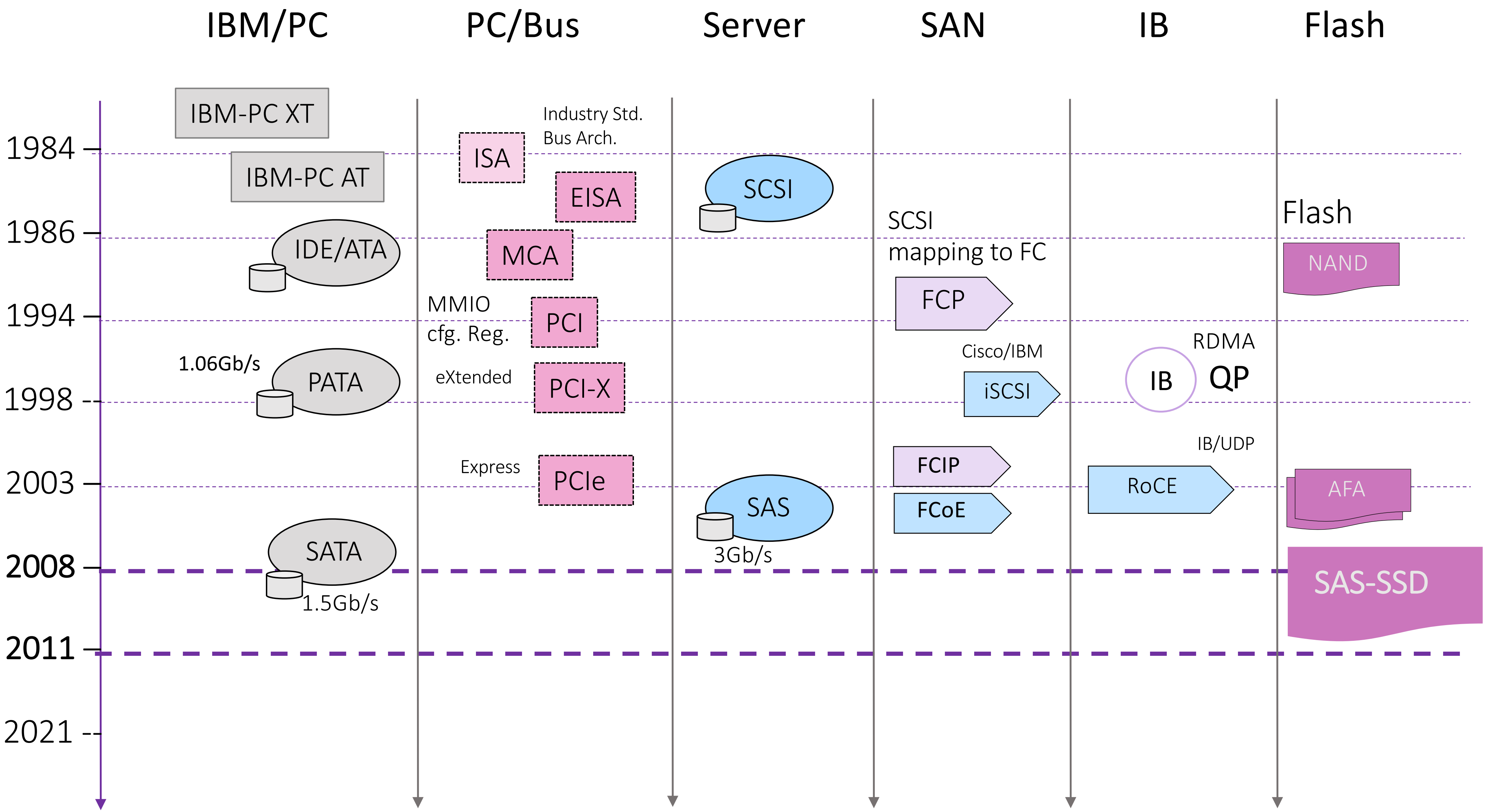
Kamal Bakshi, Cisco

SSD (Solid State Drive)



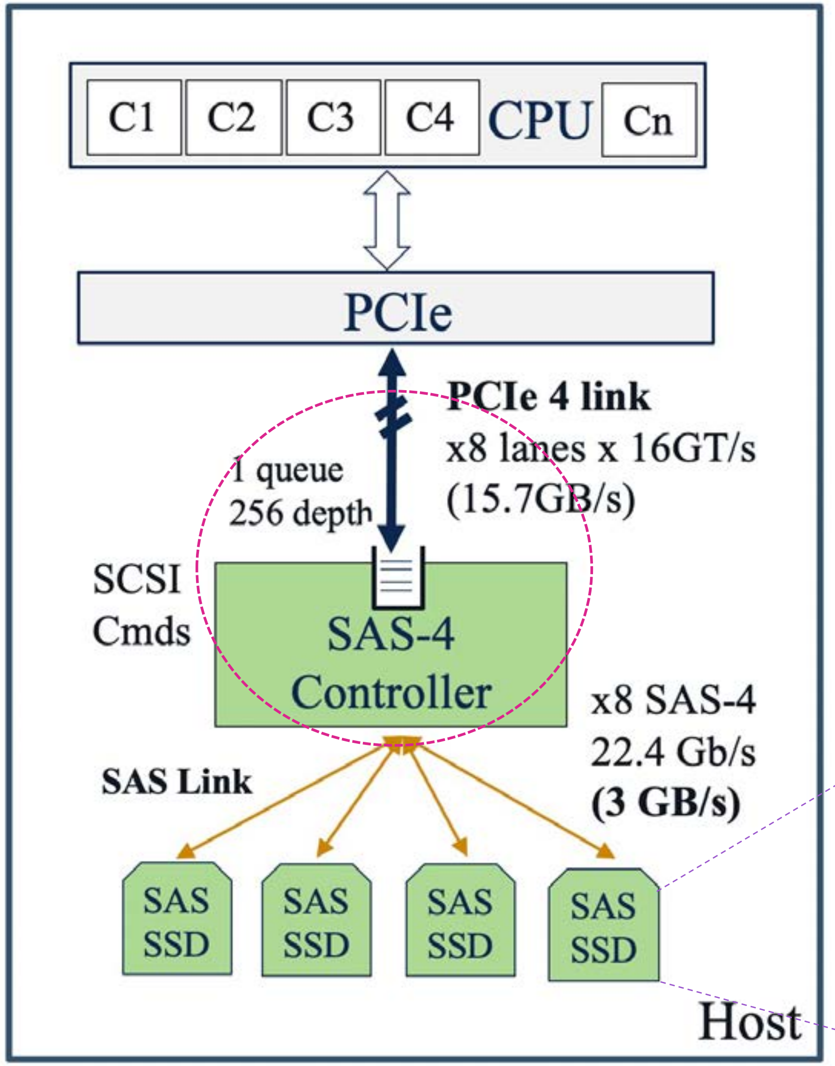
Kamal Bakshi, Cisco

SSD SAS (Serial Attached SCSI)



Kamal Bakshi, Cisco

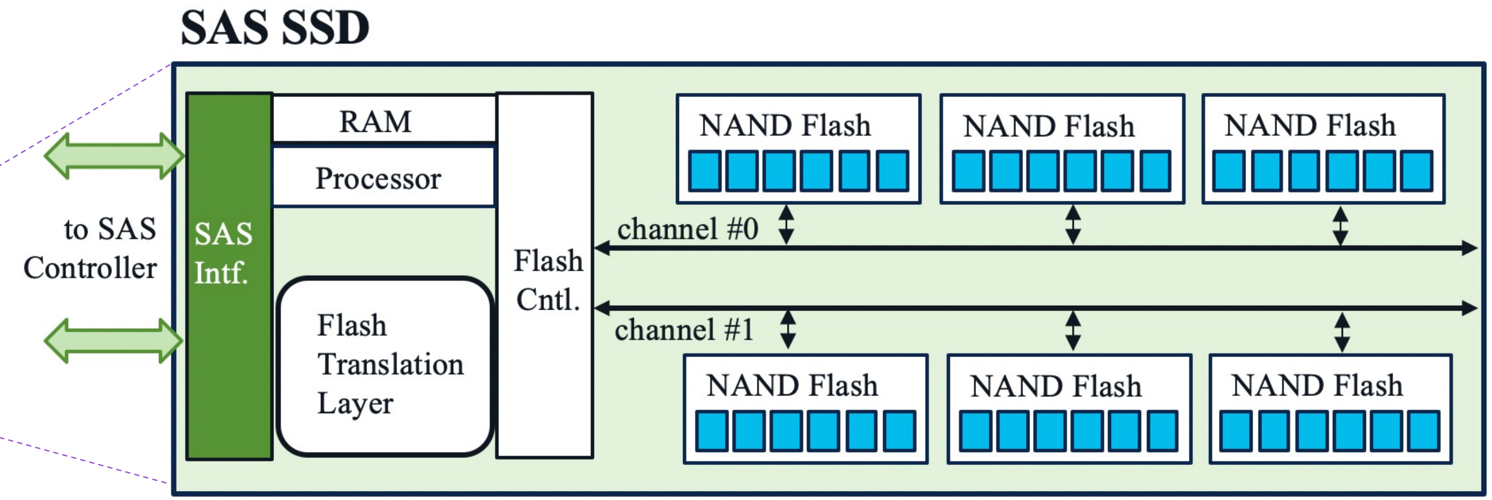
SAS-4 SSD (Maximum Throughput 3GB/s)



-SCSI Command Set

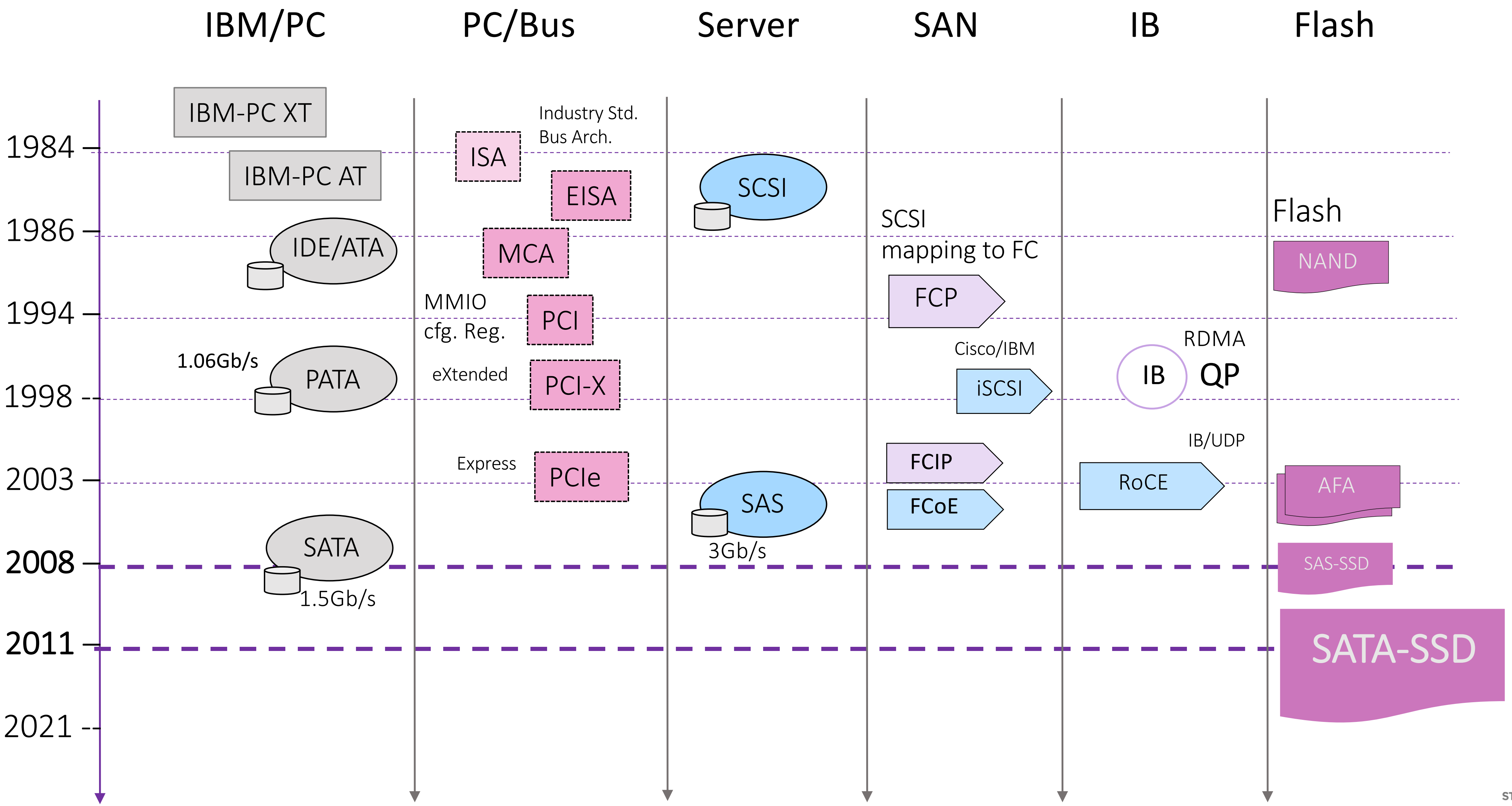
-SAS Controller/Interface

Limited max. speed → 24 Gb/s SAS-4 2017



Kamal Bakshi, Cisco

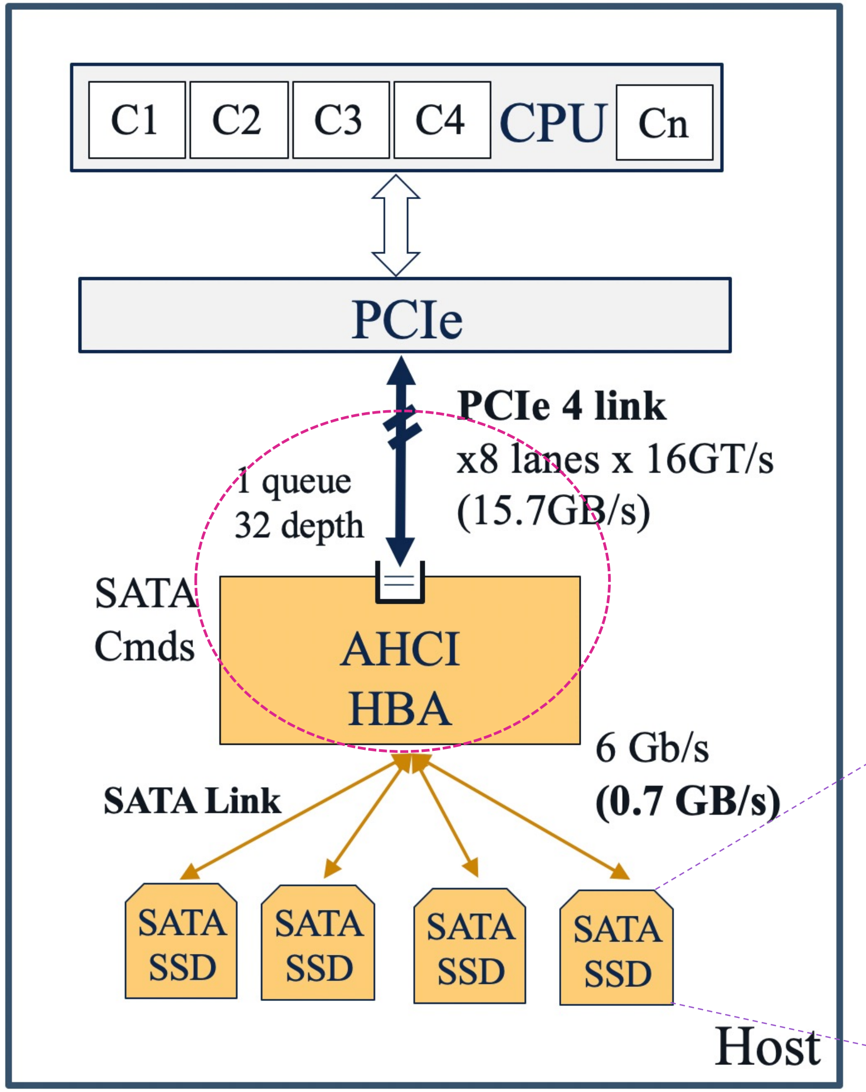
SSD SATA (Serial ATA)



Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

SATA SSD (Maximum Throughput 750MB/s)



-ATA Command Set

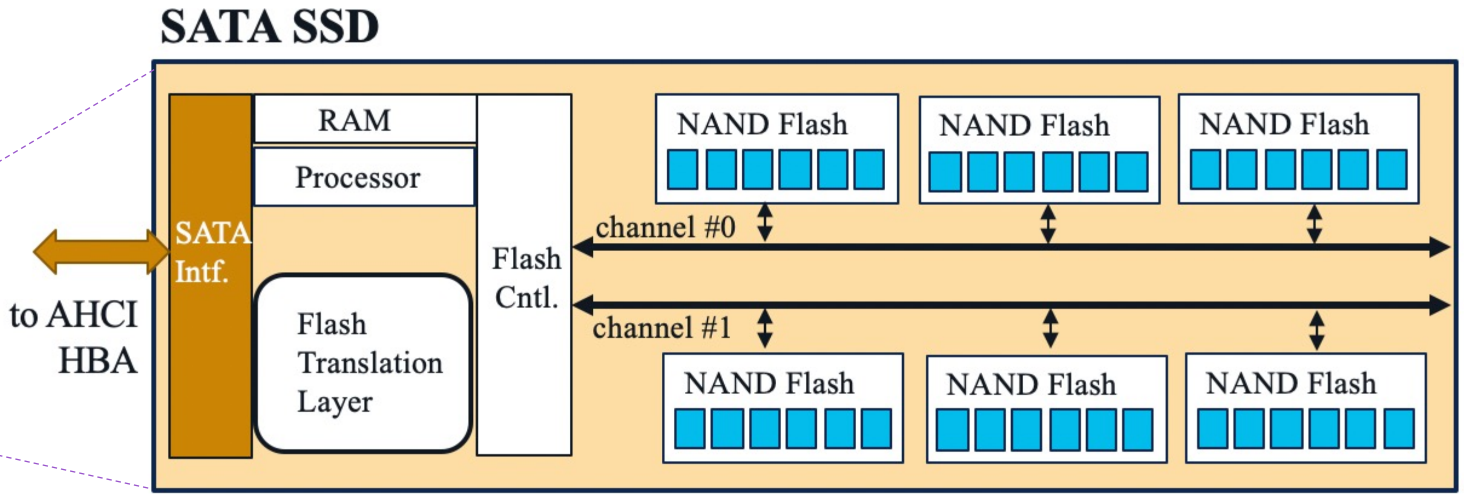
-SATA Controller/Interface

1.5 Gb/s SATA-1 2003

2.0 Gb/s SATA-2 2004

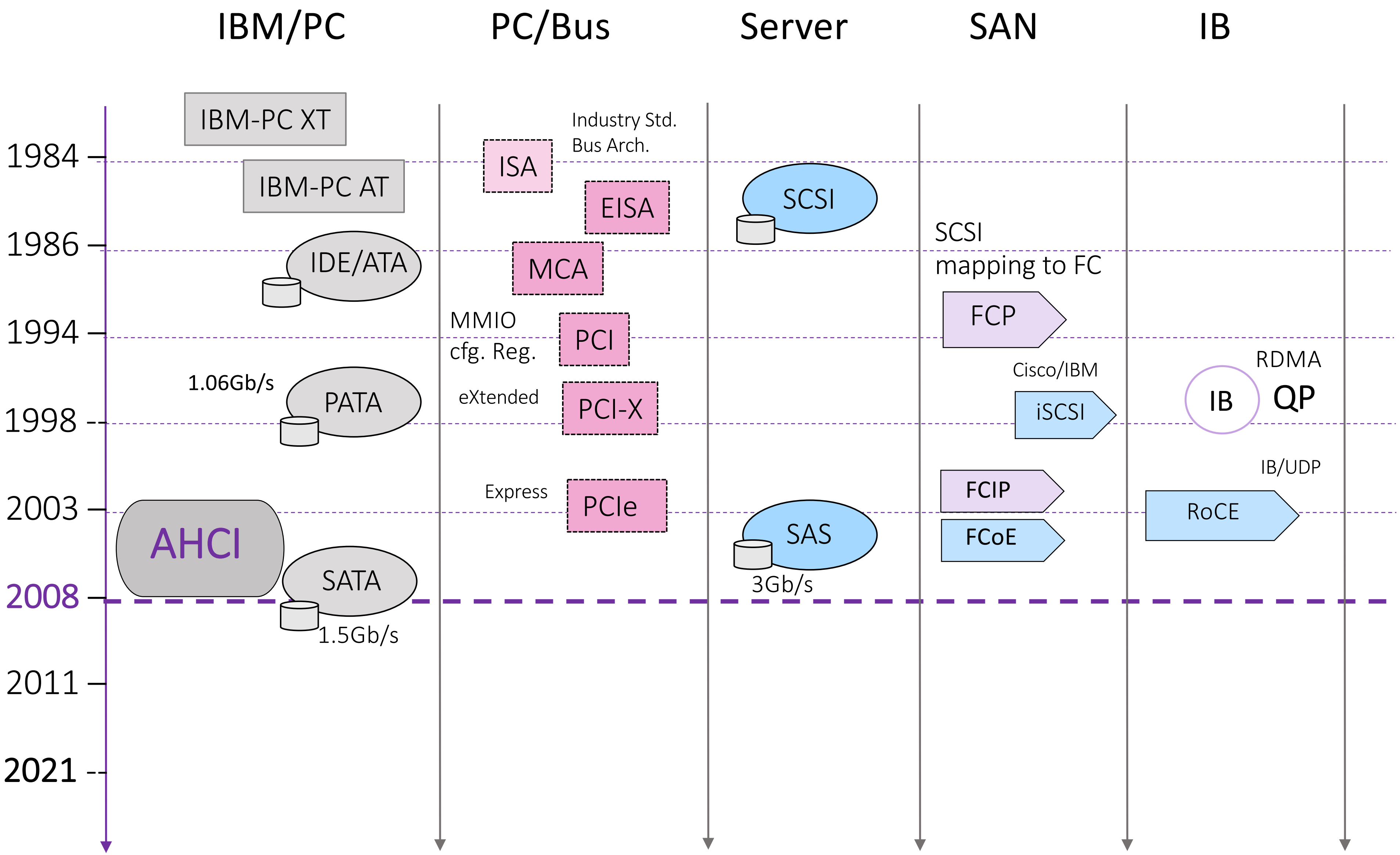
6.0 Gb/s SATA-3 2008 ← Limited max. speed

6+ Gb/s SATA-3.2 (SATA Express) 2013



Kamal Bakshi, Cisco

AHCI (Advance Host Controller Interface)

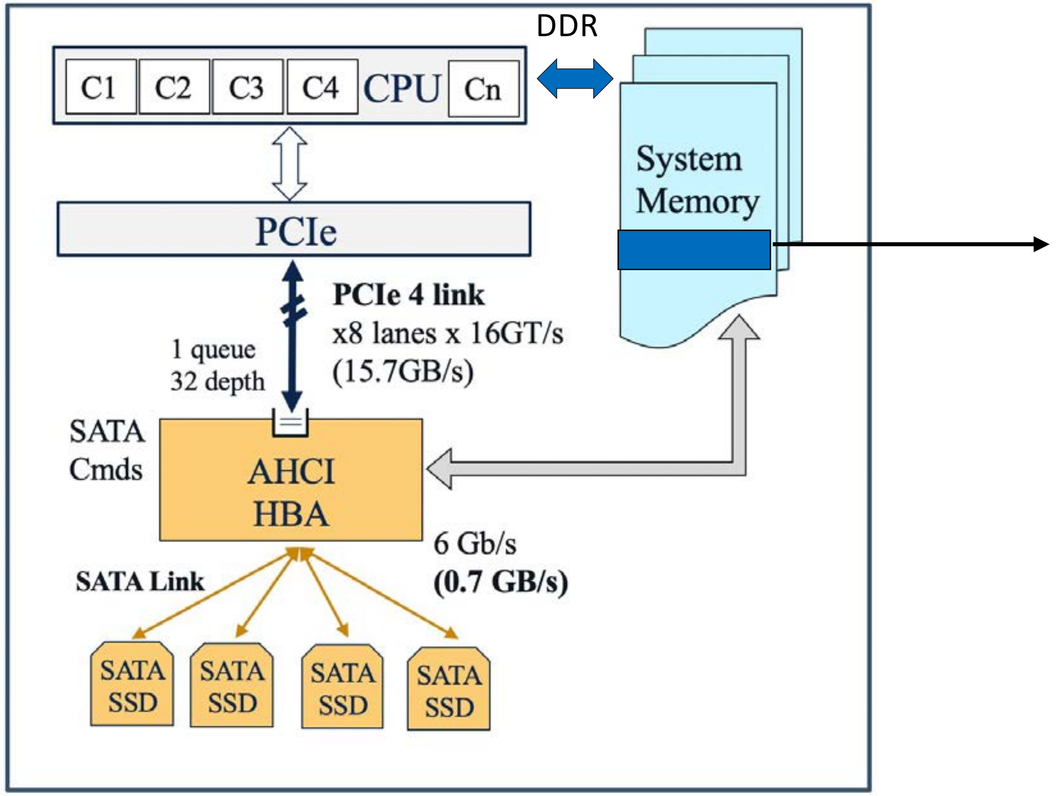


Advance Host Cntl. Intf.
"AHCI is a PCI class device that acts as a data movement engine between system memory and Serial ATA device."

Kamal Bakshi, Cisco

AHCI Advantages

-AHCI device allows **data movement between system memory and SATA device**
-It makes HBA implementation simpler as they are not required to parse ATA commands
-Data transfers between SATA device and system memory uses DMA thus offloading the CPU
-AHCI also enables hot-plugging



PCIe Register

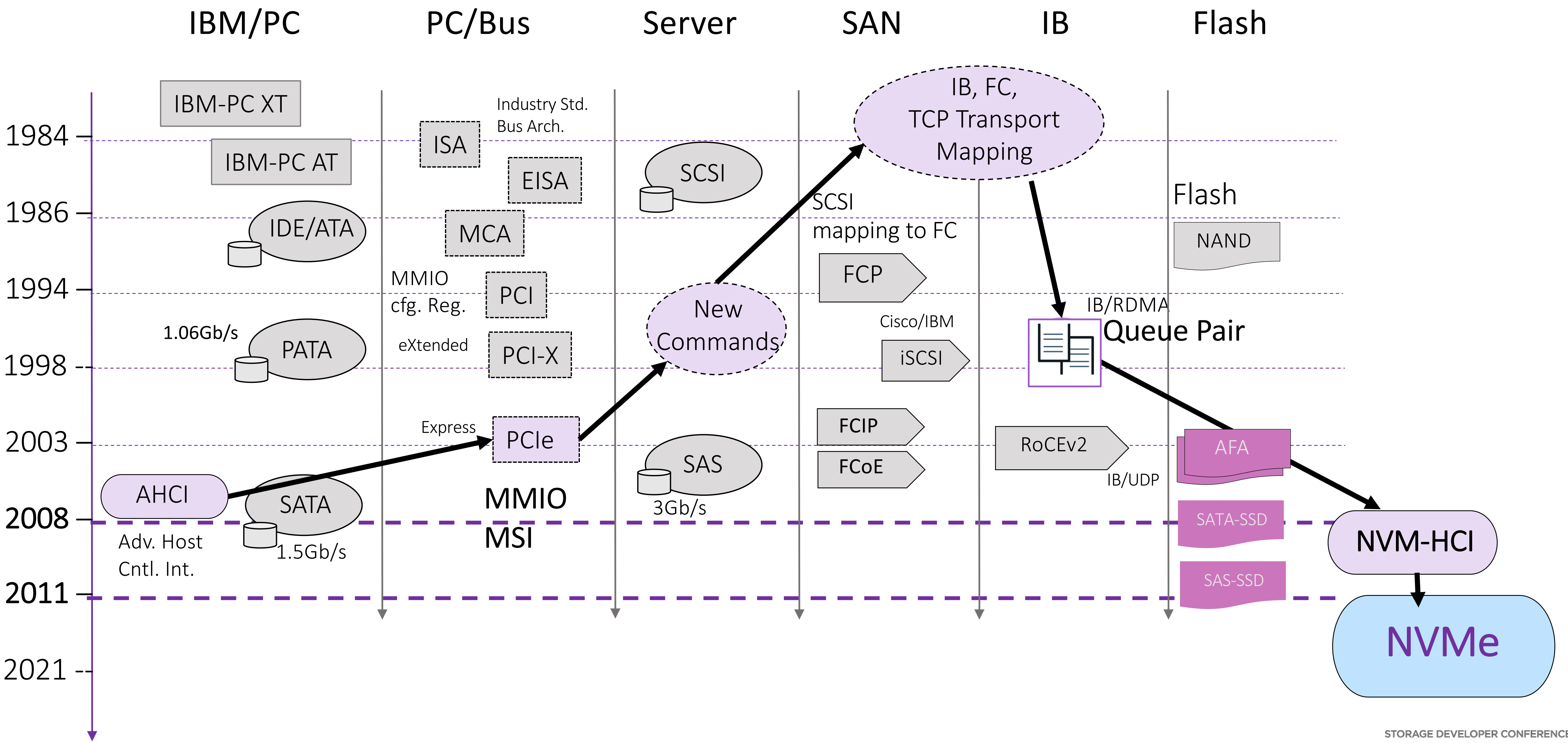
00h	3Fh	PCI Header
PMCAP	PMCAP+7	PCI Power Mgmt. Capability
MSICAP	MSICAP+9	Msg. Signaled Intr. Capability

PCIe Header

00-03h	ID	Identifier
04-05h	CMD	Command Register
06-07h	STS	Device Status
08-08h	RID	Revision ID
09-0Bh	CC	Class Code
10-23h	BARS	Base Address Registers 0-4
24-27h	ABAR	AHCI BAR - 05
2C-2Fh	SS	Subsystem Identifiers
34-34h	CAP	Capability Pointer

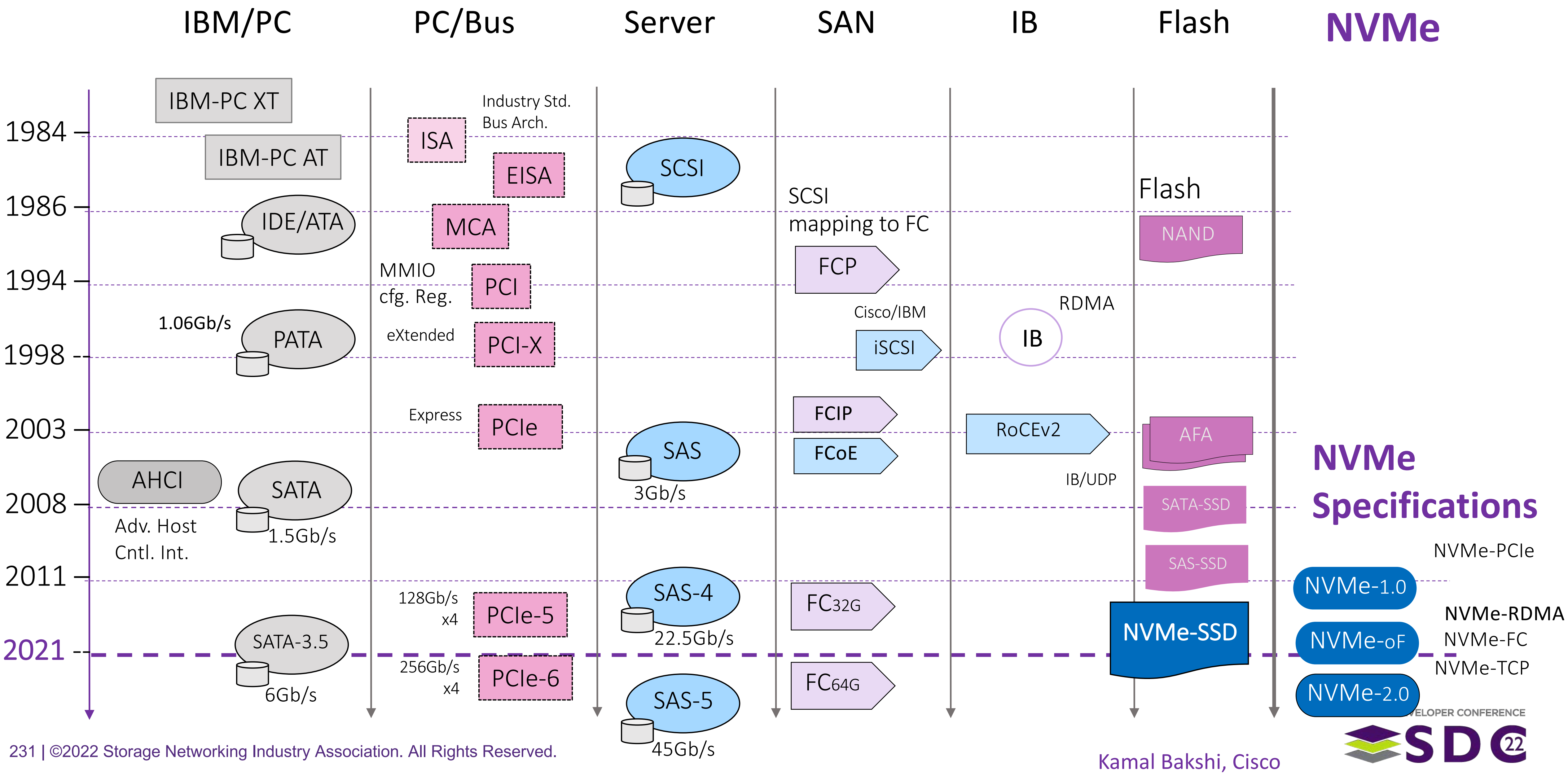
Kamal Bakshi, Cisco

Best of all worlds....

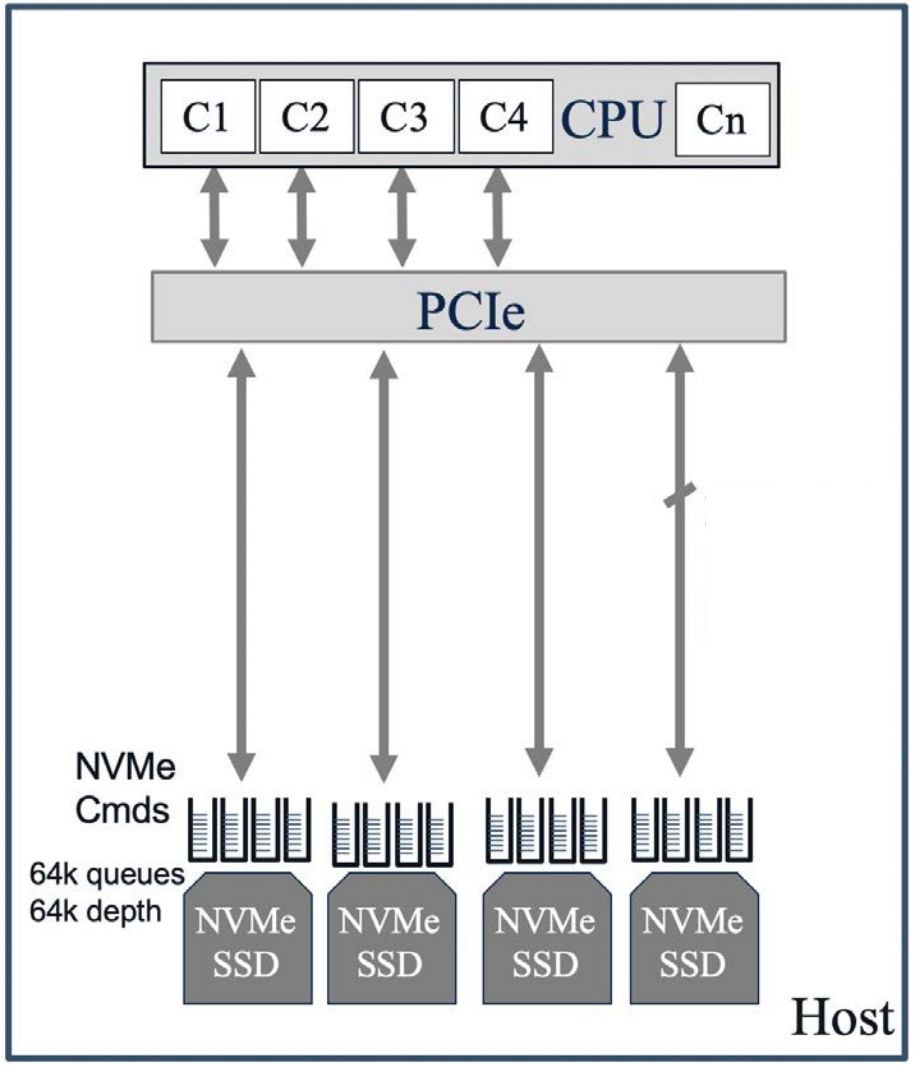


Kamal Bakshi, Cisco

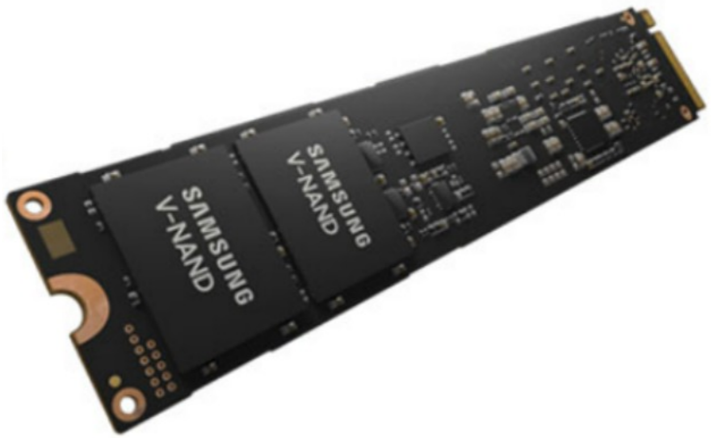
NVMe (Non Volatile Memory Express)



NVMe SSD Form Factors (M.2)



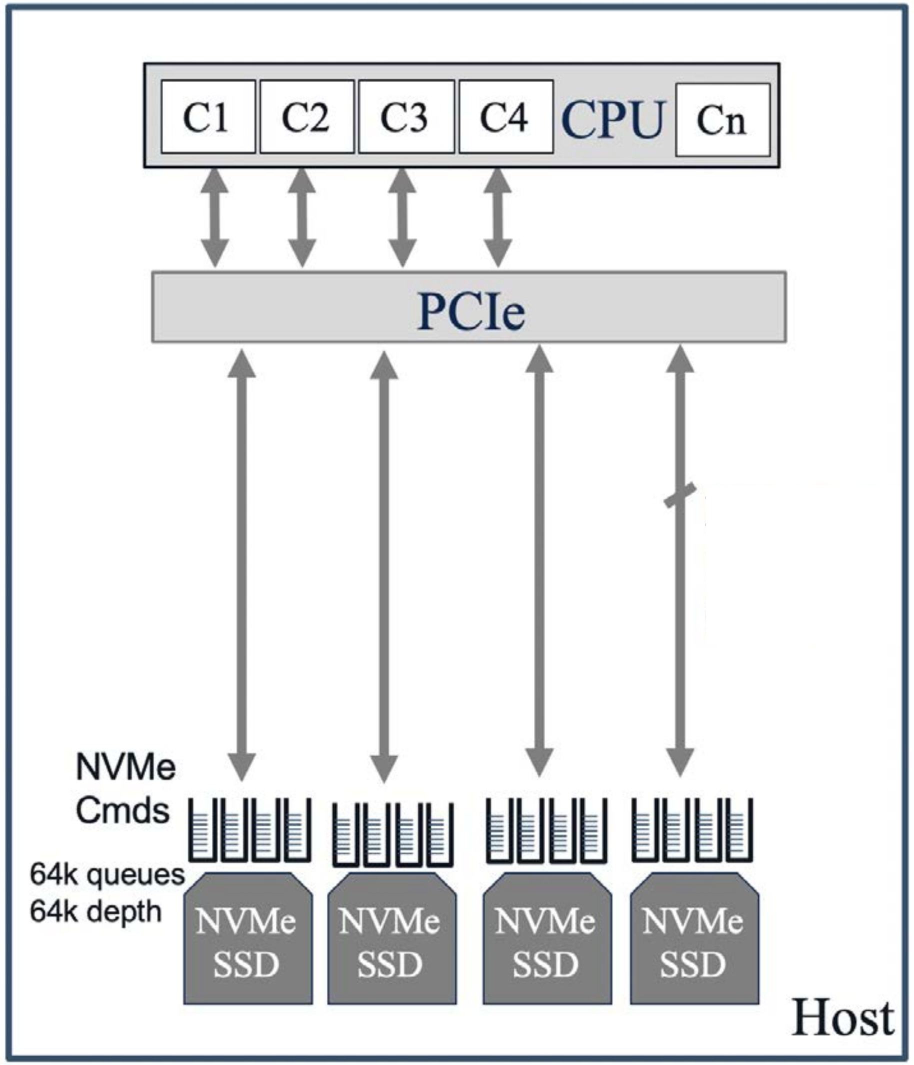
M.2 is a form factor specification for internally mounted SSDs. Formerly known as Next Generation Form Factor (NGFF) and comes in various widths and lengths.



Dimensions
16mm x 20mm
22mm x 30mm
22mm x 80mm
22mm x 110mm

Kamal Bakshi, Cisco

NVMe SSD Form Factors (U.2)

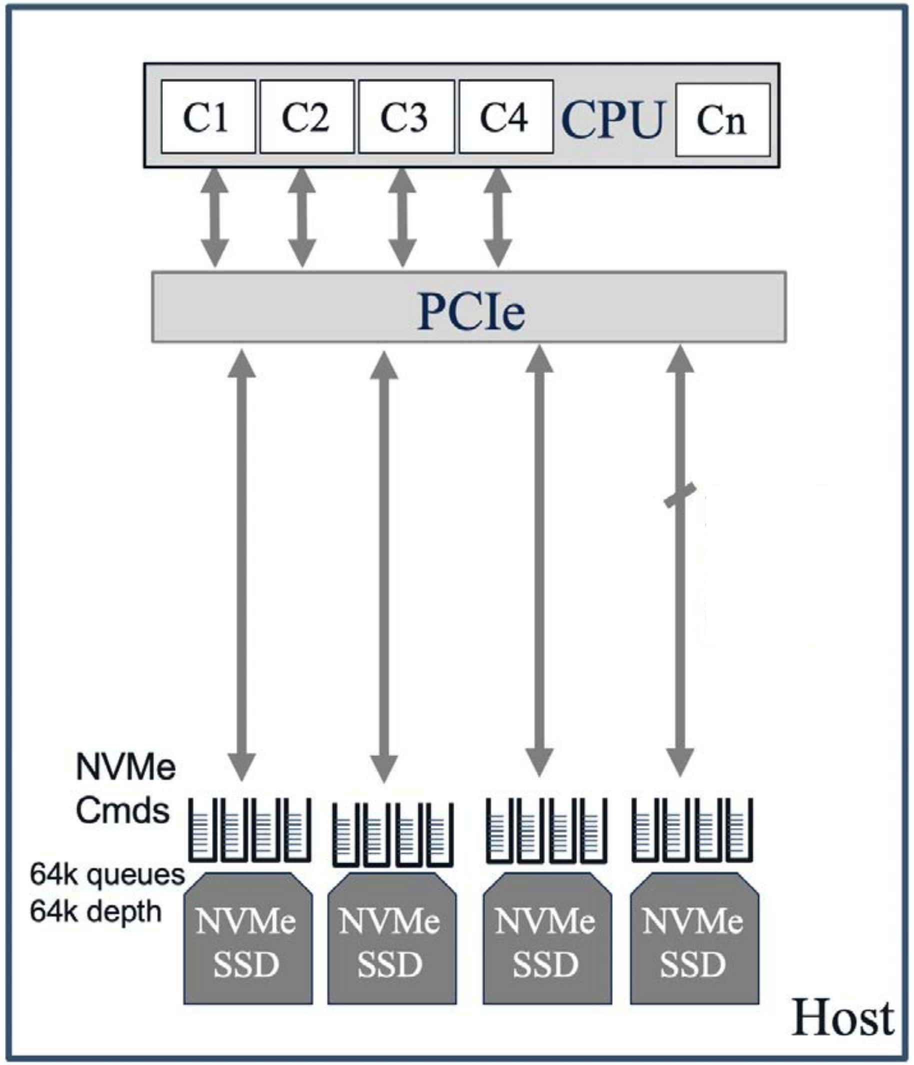


U.2 is defined as compliance with the PCI Express SFF-8639 Module specification, and no longer typically references SAS or SATA SSDs.

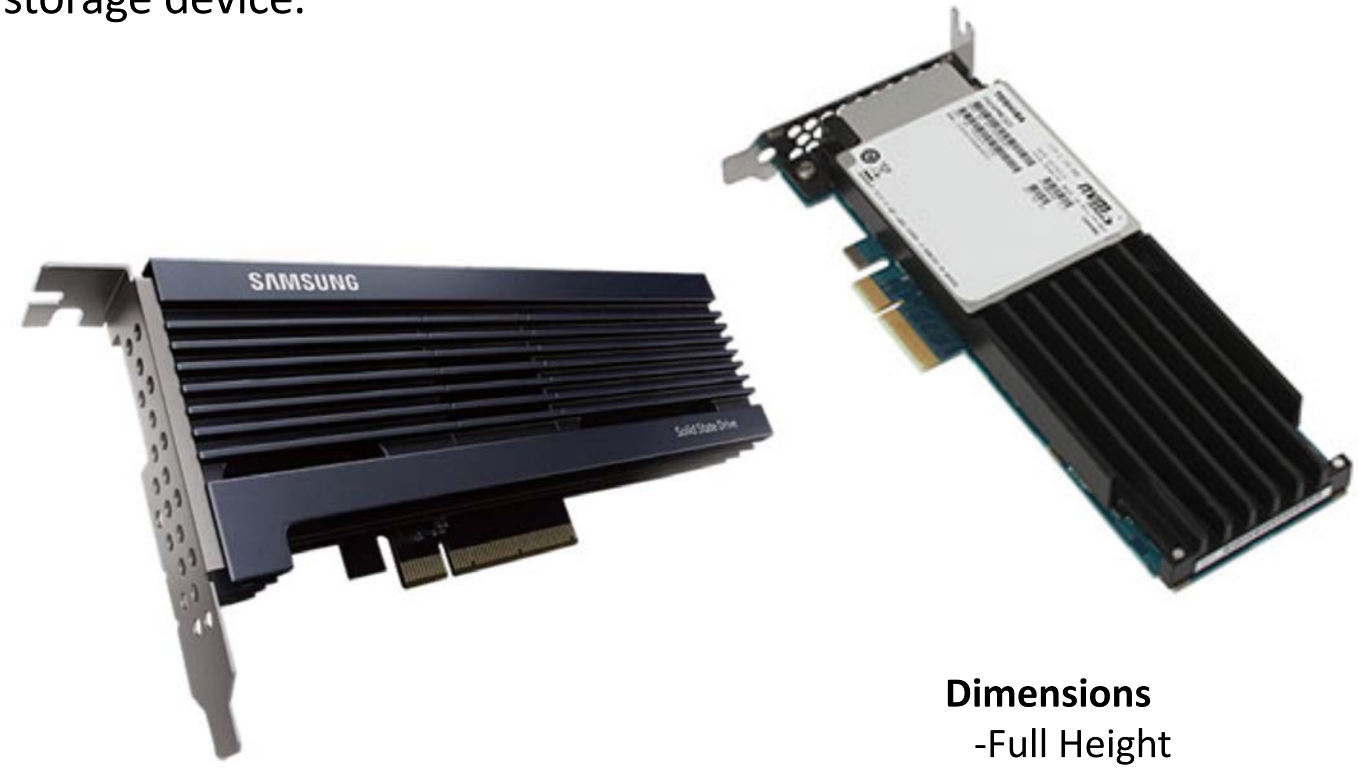
Dimensions
2.5-inch(7mm) [69.85x100x7 mm]
2.5-inch(15mm) [69.85x100x15mm]

Kamal Bakshi, Cisco

NVMe SSD Form Factors (AIC)



An Add-in Card (AIC) is a solid-state device that utilizes a standard card form factor such as a PCIe card. In addition, the larger size allows for the potential to add computational function to the storage device.



Dimensions

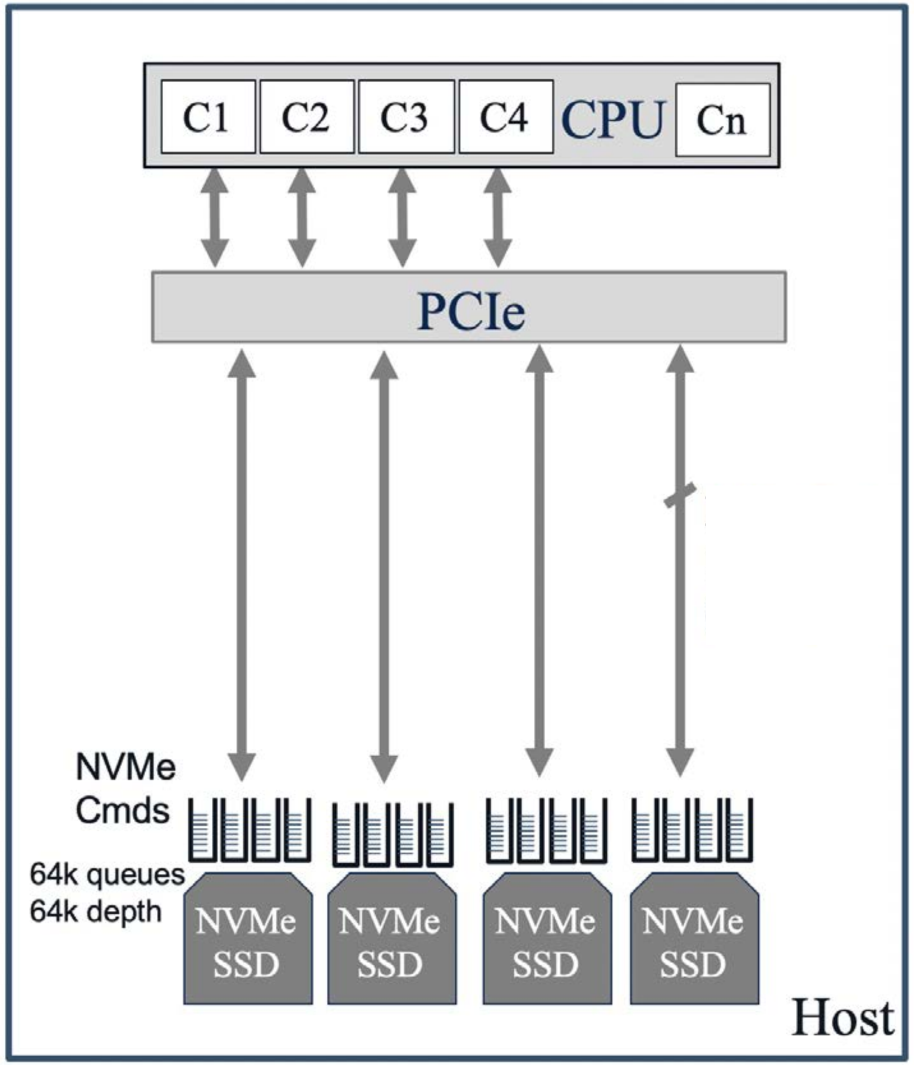
-Full Height

-Half Height

-Low Profile

Kamal Bakshi, Cisco

NVMe SSD Form Factors (EDSFF)

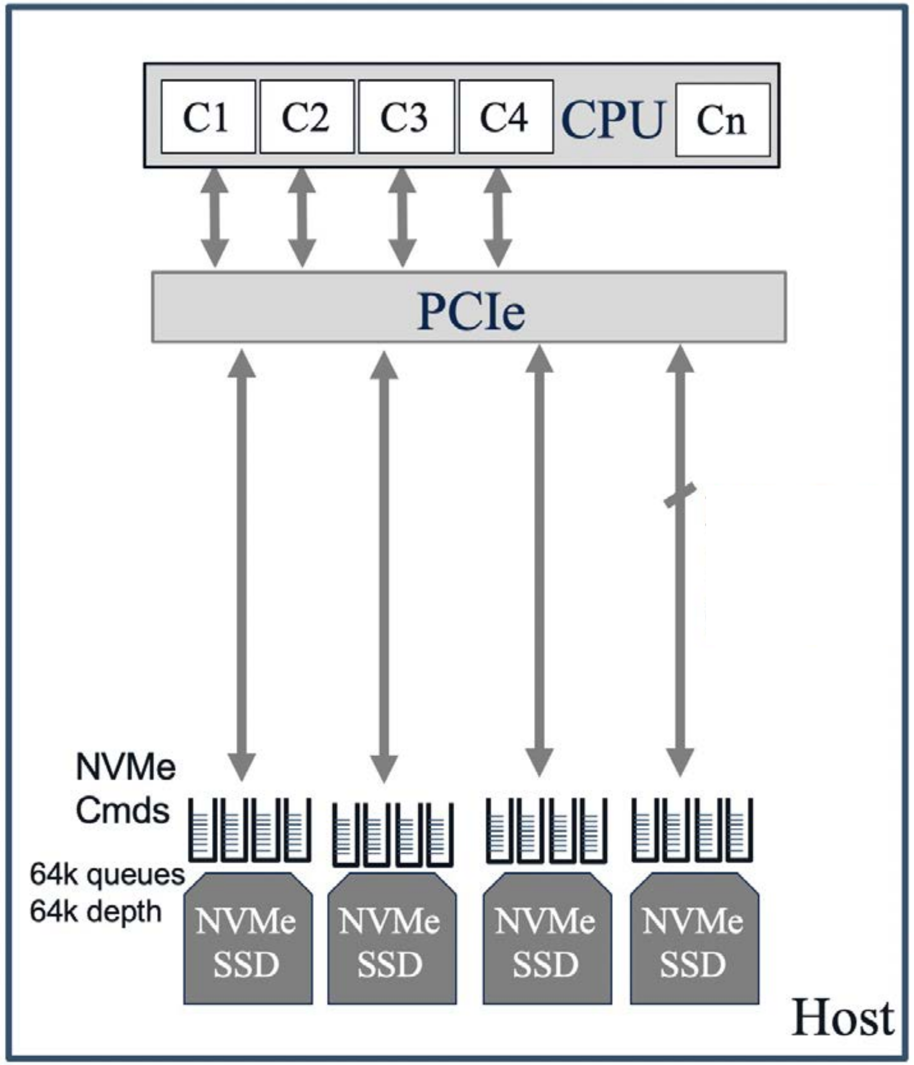


EDSFF stands for Enterprise and Data Center Standard Form Factor. The family of specifications were developed by a group of 15 companies working together to address the concerns of data center storage, and are now maintained by SNIA as part of the SFF Technology Affiliate Technical Work Group (SFF TA TWG).

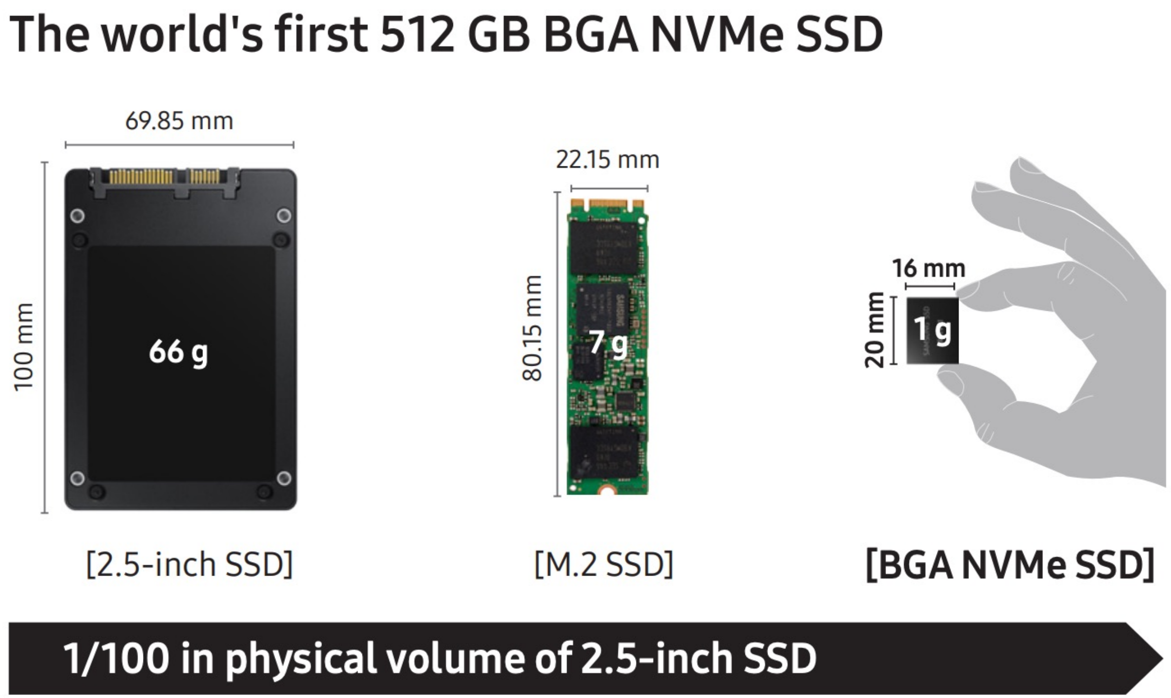
Dimensions (thickness)
E1.L (long) 9.5mm, 18mm
E1.S (short) 5.9mm, 8.01mm, 9.5mm, 15mm, 25mm
E3.S (short) 7.5mm, E3.S 2T 16.8mm
E3.L (long) 7.5mm, E3.L 2T 16.8mm

Kamal Bakshi, Cisco

NVMe SSD Form Factors (BGA)

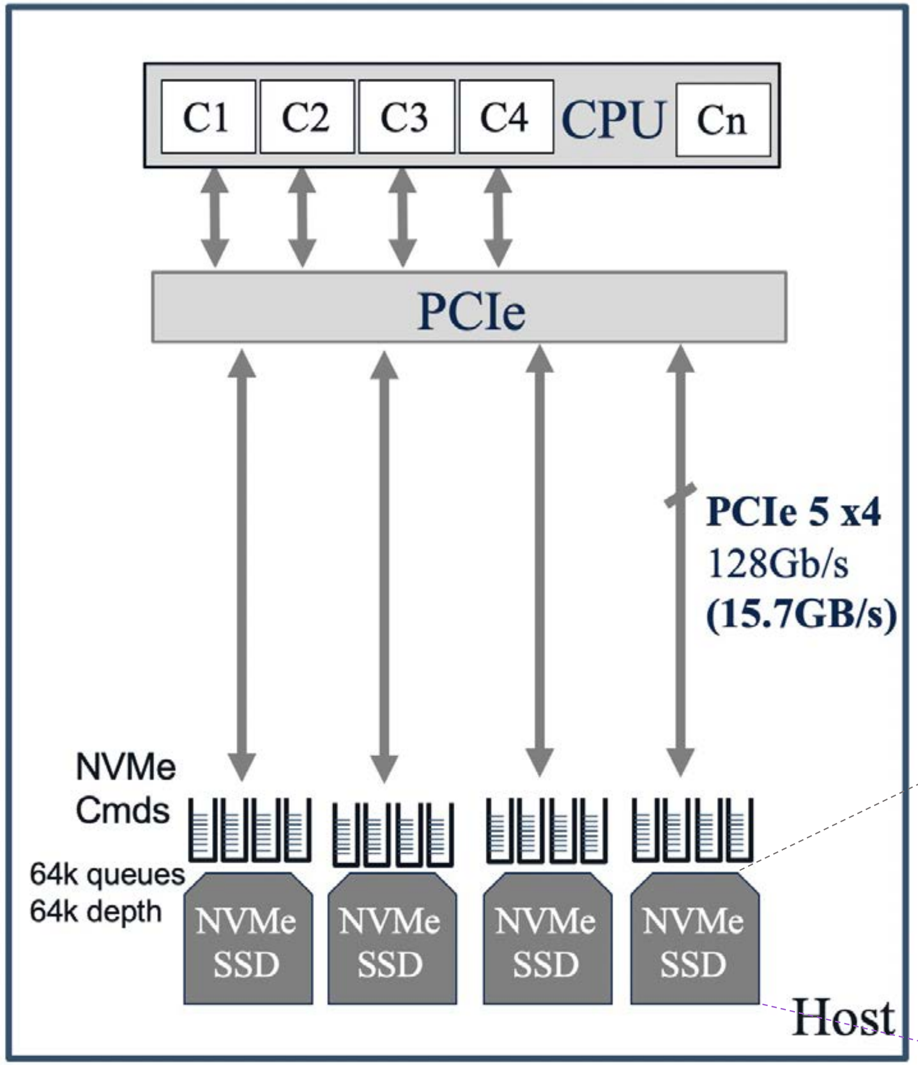


In 2016 Samsung started to mass produce the industry's first NVMe PCIe solid state drive (SSD) in a single ball grid array (BGA) package, for use in next-generation PCs and ultra-slim notebook PCs.



Kamal Bakshi, Cisco

NVMe SSD (15GB with PCIe-5)



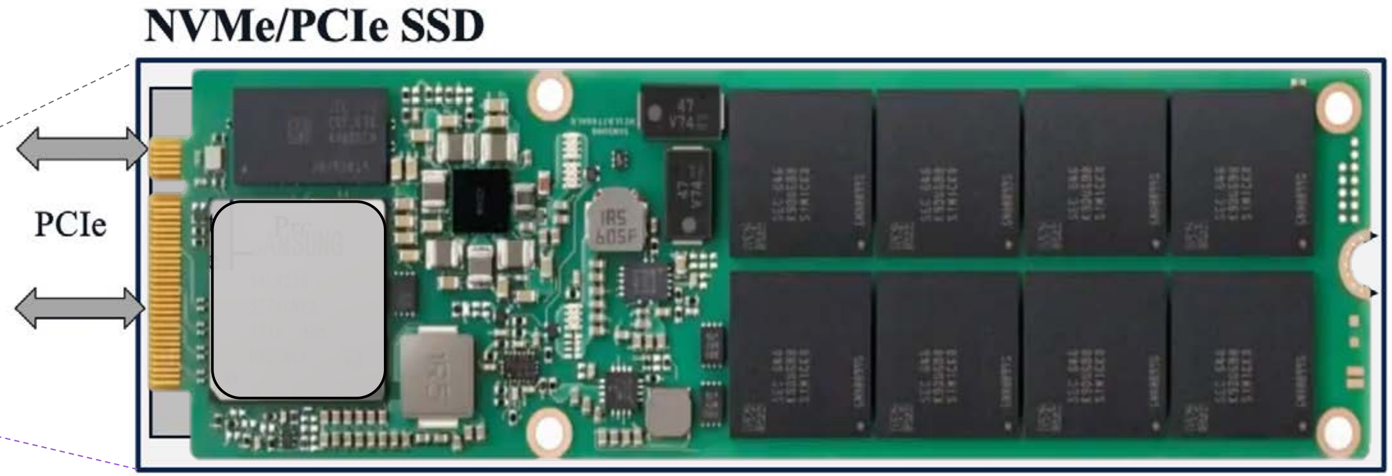
NVMe

- -New Block Storage Protocol for Flash
- -Maps directly into PCIe
- -Replaces SCSI commands
- -Transport mapping for RDMA/FC/TCP

Fabric Command
- -Connect/Disconnect
- -Set/Get Property

I/O Command
- -Read/Write
- -Flush

Admin Command
- -Create/Delete I/O SQ
- -Create/Delete I/O CQ
- -Get Log Page
- -Identify
- -Abort
- -Set/Get Feature
- -Async. Event Request

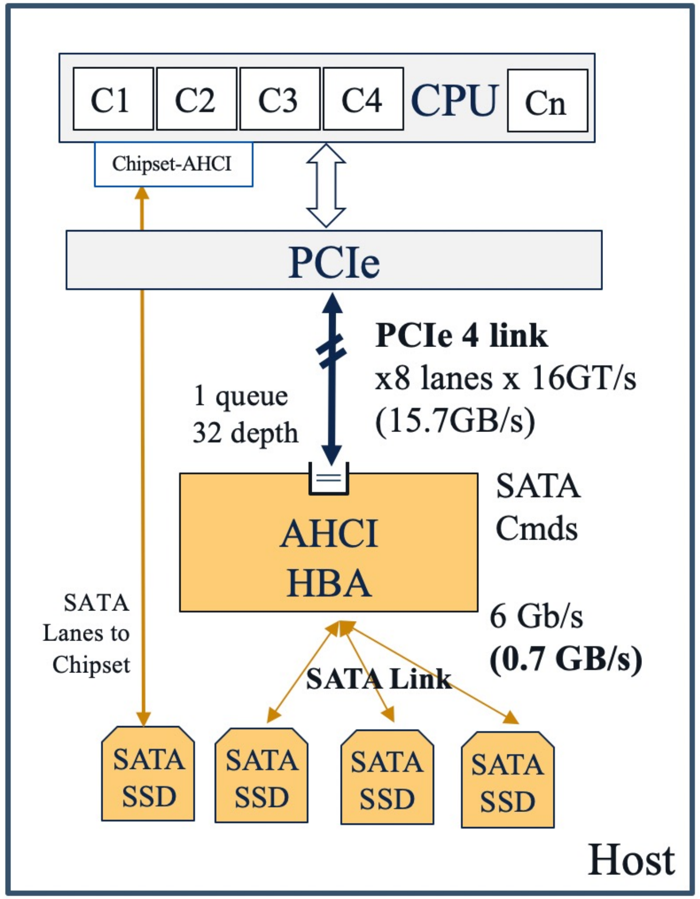


M.2 form factor

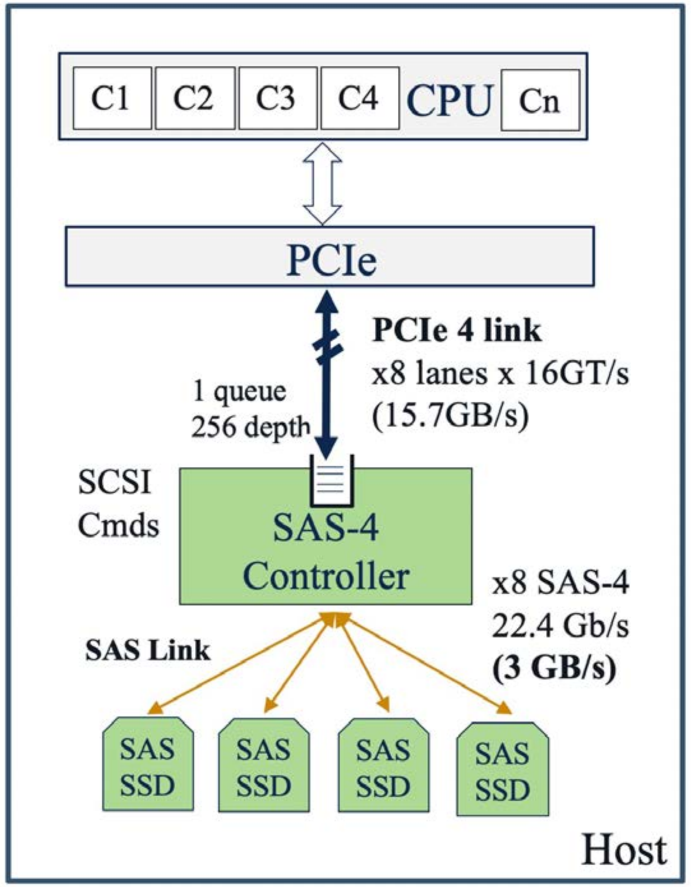
Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

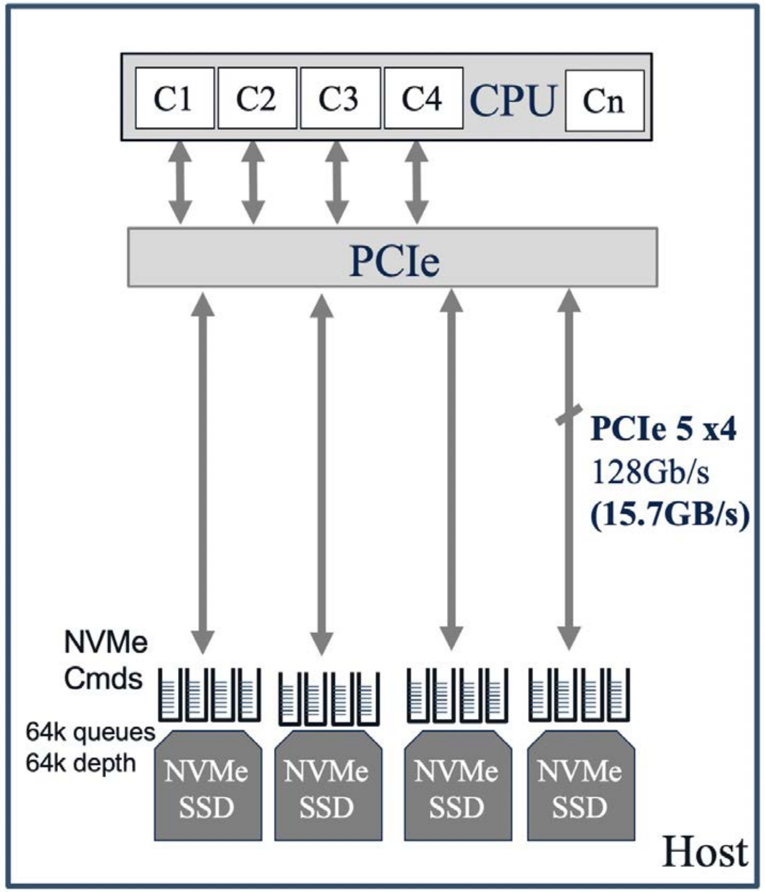
SATA, SAS, PCIe/NVMe



SATA-SSD



SAS-SSD



PCIe/NVMe-SSD

Kamal Bakshi, Cisco

NVMe-PCI Architecture

Kamal Bakshi, Cisco

NVMe-PCIe Transport

I — **Connect to NVMe Cntl.**

-Discover the NVMe I/O controller
-Initiate the NVMe connection over PCIe

II — **Create NVMe Queues**

-Create Admin Queue for NVMe commands
-Create I/O Queues for Data Transfer

III — **Discover Namespaces**

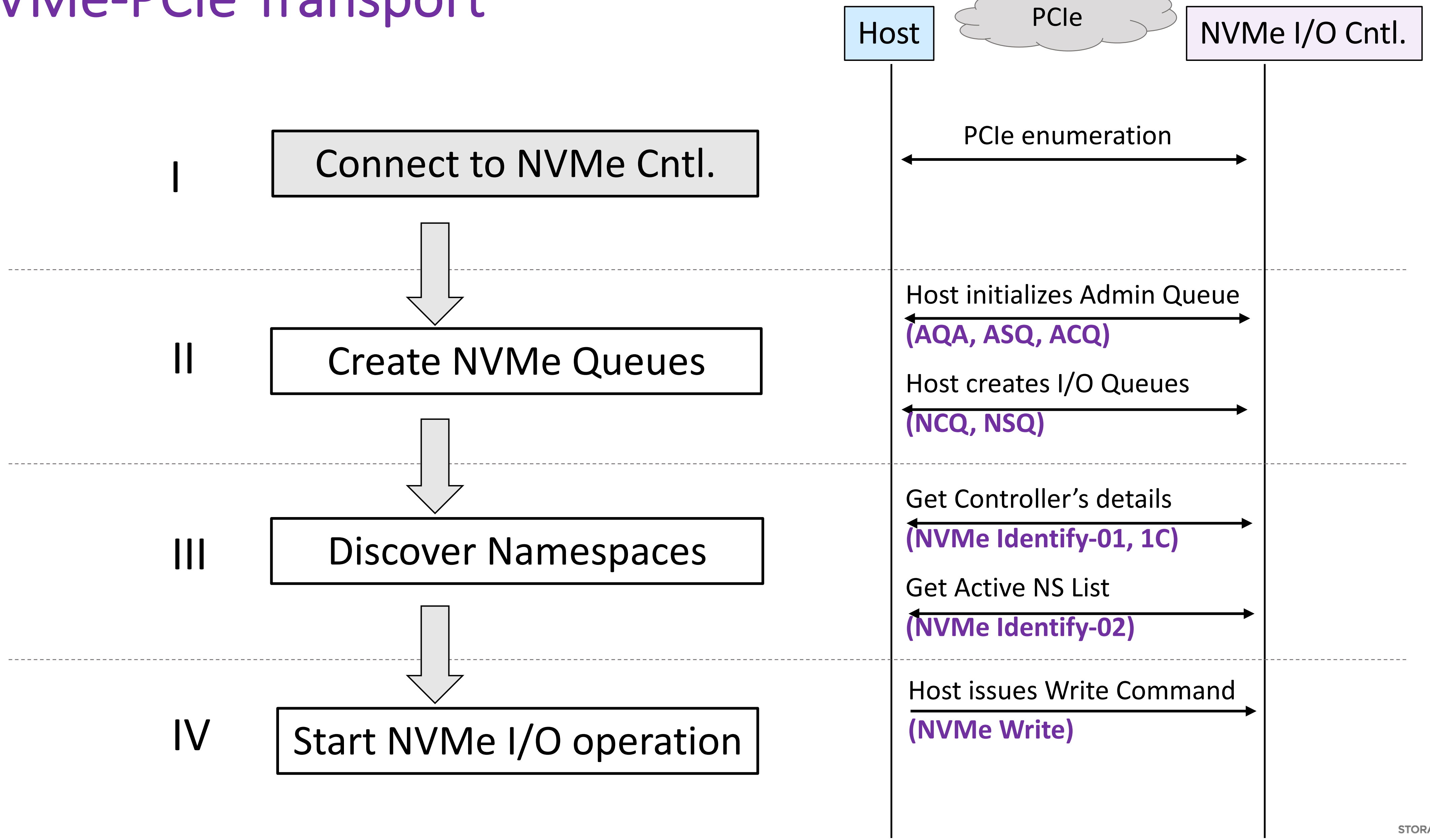
-Get a list of Active available Namespaces

IV — **Start NVMe I/O operation**

-Start I/O transfer with NVMe Write/Read
(I/O transfer uses I/O queues)

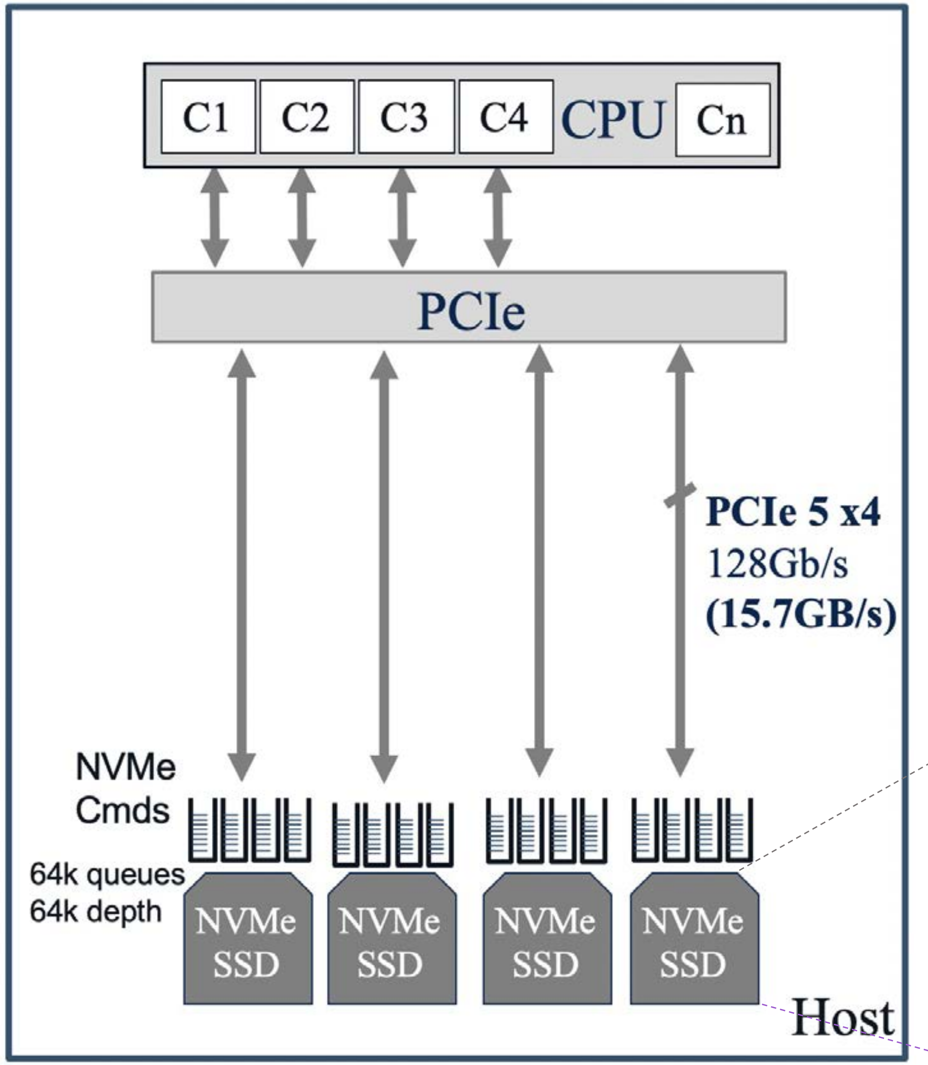
Kamal Bakshi, Cisco

NVMe-PCIe Transport



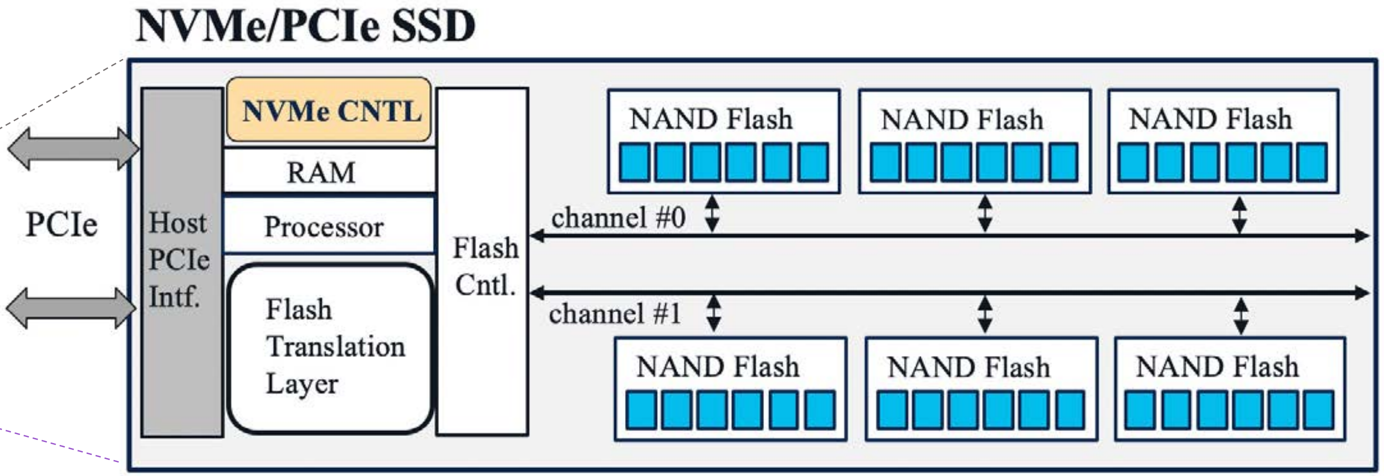
Kamal Bakshi, Cisco

NVMe-PCIe Transport



The **PCIe transport** provides reliable mechanisms for memory mapped data transfer of Admin and I/O command data through memory mapped I/O transactions.

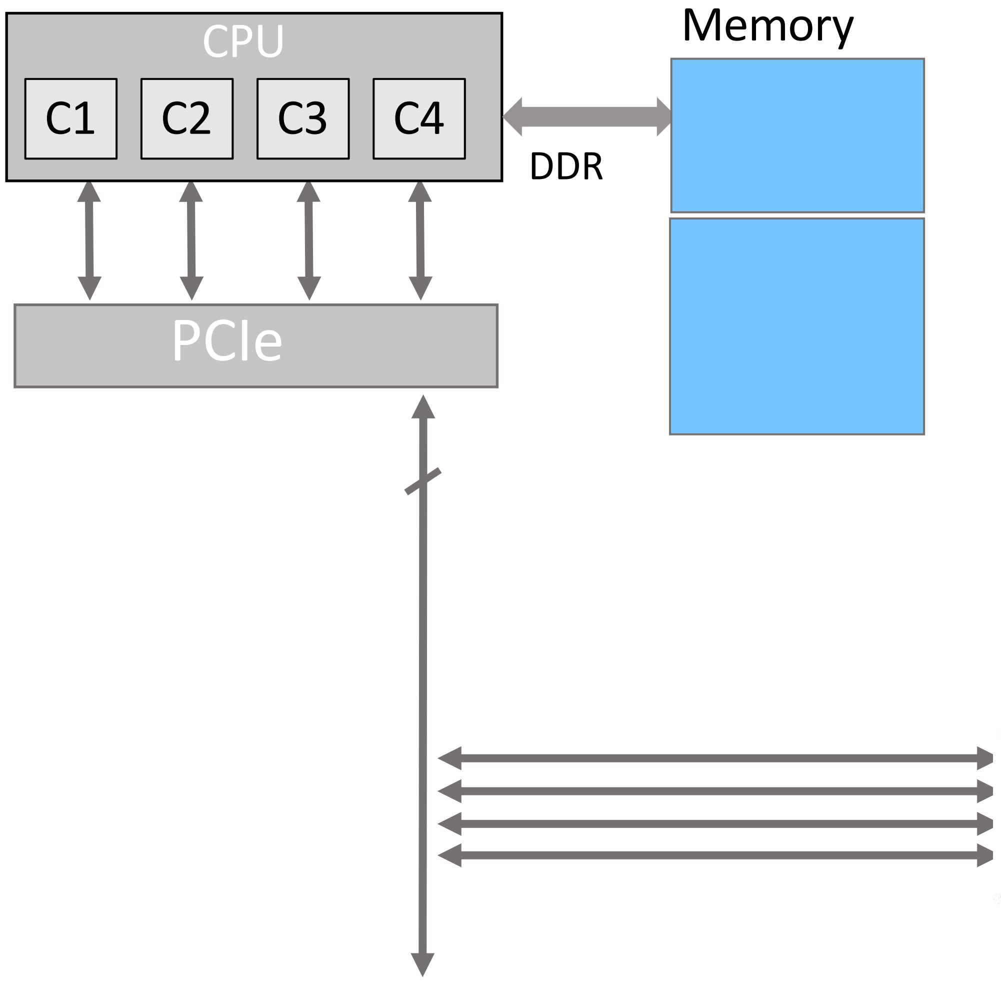
….NVMe-PCIe spec. 1.0



Kamal Bakshi, Cisco

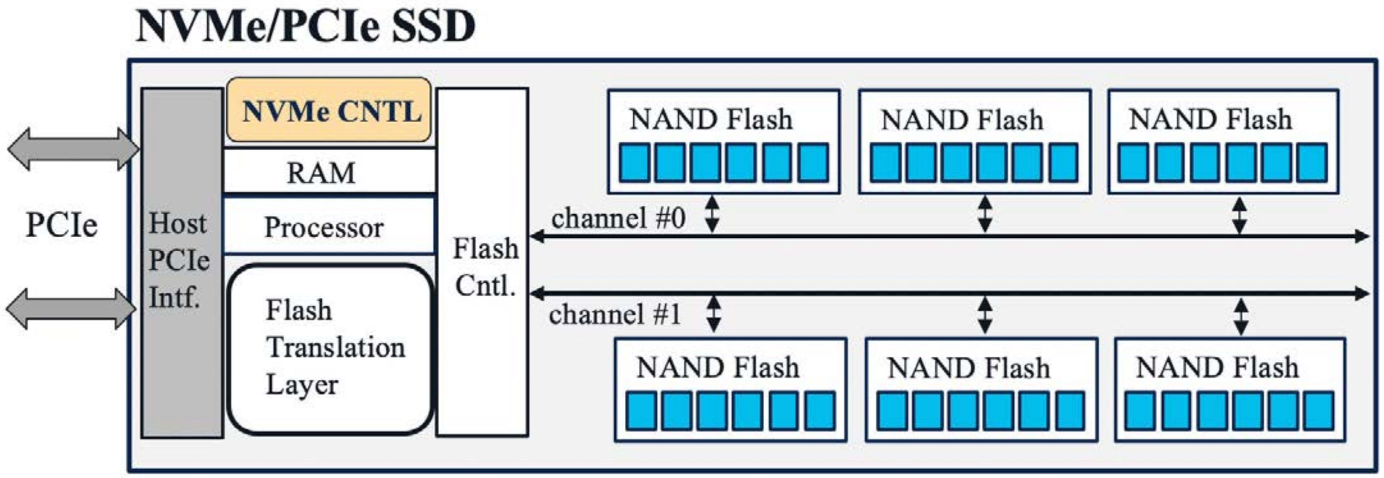
NVMe-PCIe Transport

A memory model is one in which commands, responses and data are transferred between fabric nodes by performing explicit memory read and write operations



Building Blocks of NVMe

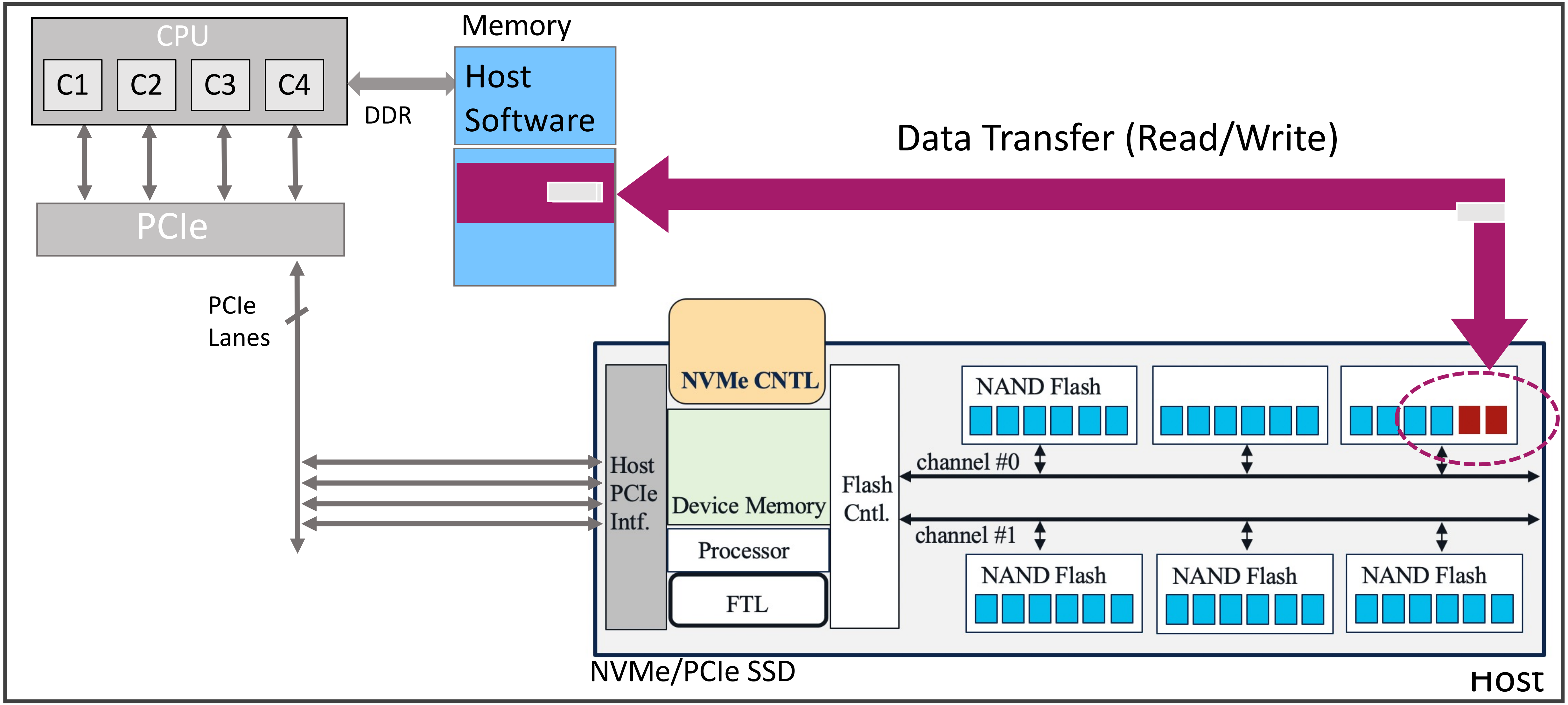
- PCIe Config Registers
- BAR Address, Capability Pointers
- Messaged Signaled Interrupt, Doorbell
- NVMe Queues, Admin/IO (SQ/CQ)
- NVMe Subsystem/Controller



Kamal Bakshi, Cisco

NVMe-PCIe Transport

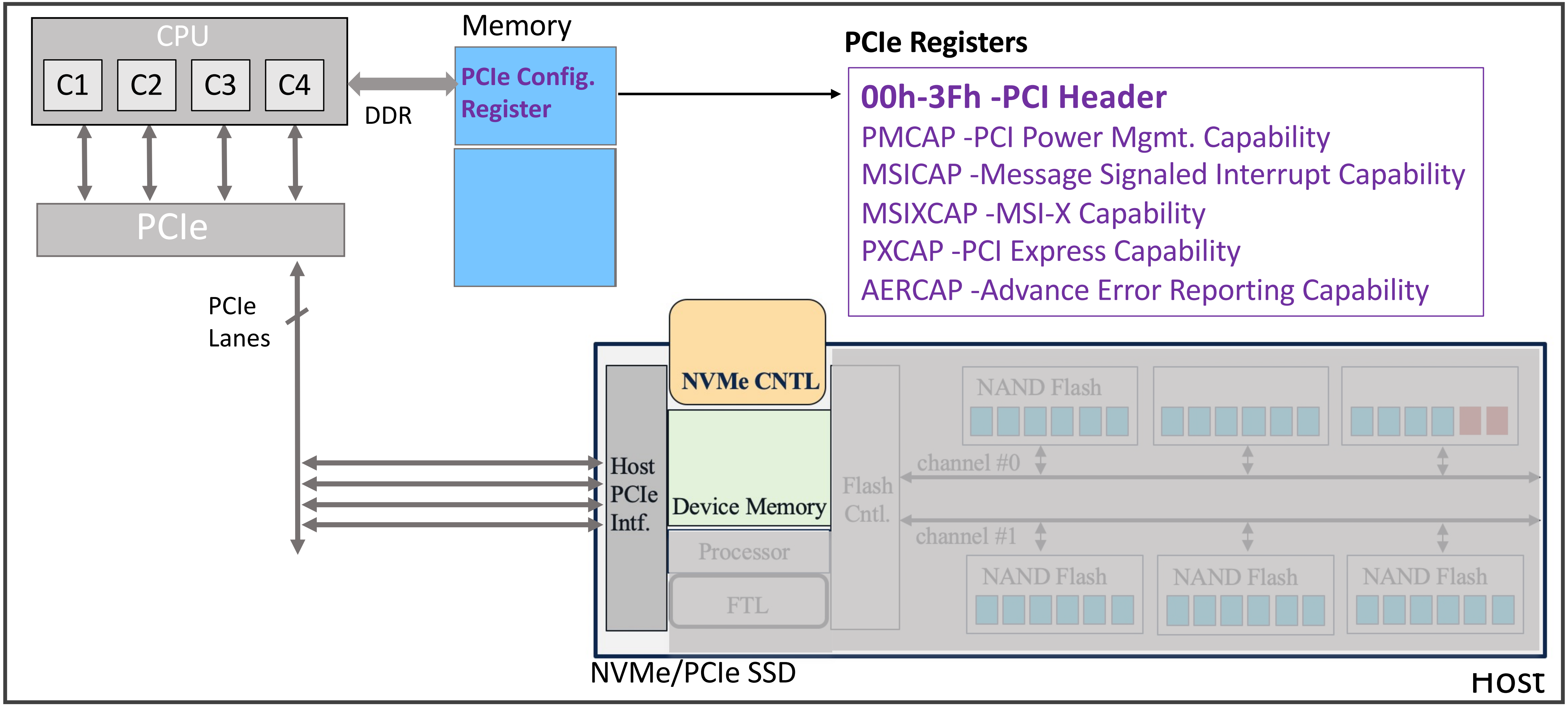
How does Data Transfer Works?



Kamal Bakshi, Cisco

NVMe-PCIe (Registers)

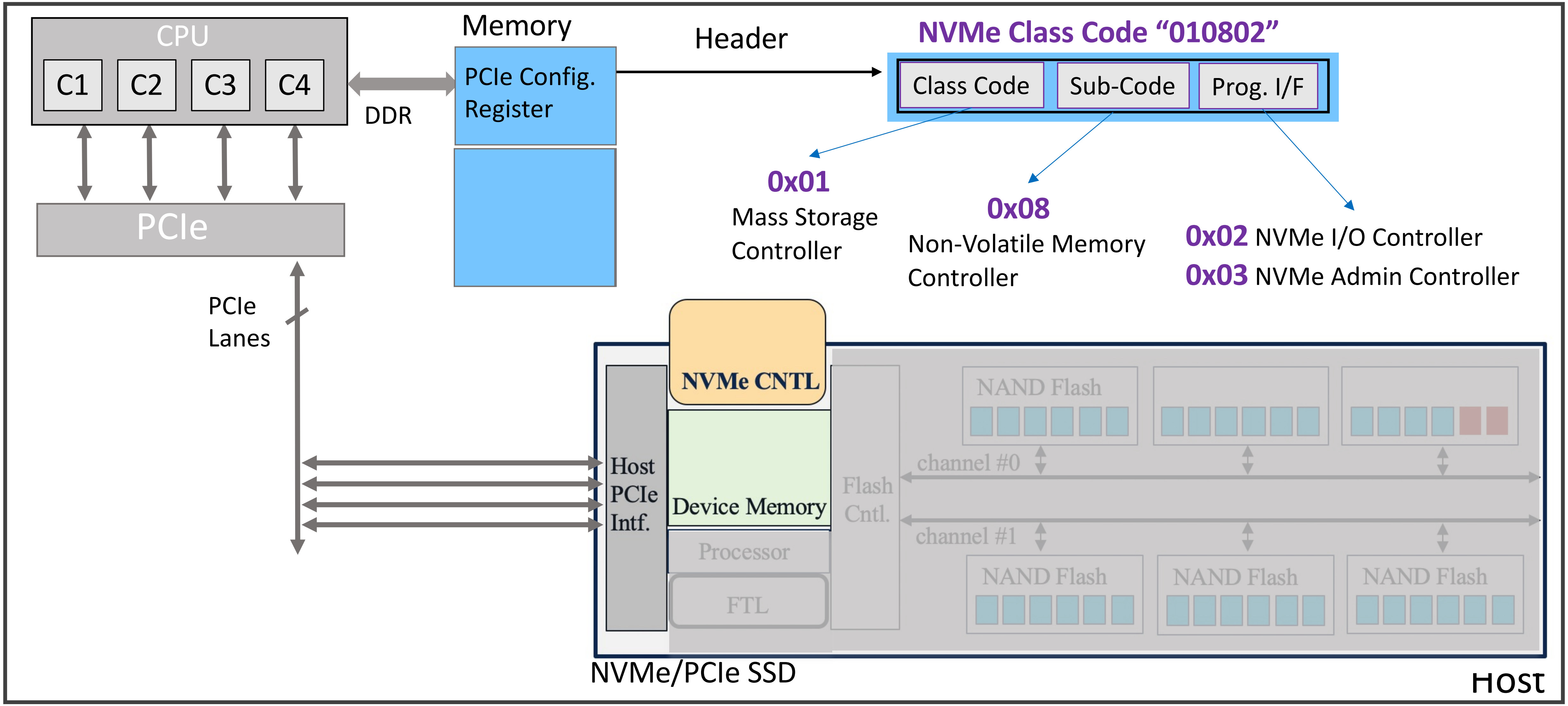
PCIe devices have set of registers that are mapped to memory locations



Kamal Bakshi, Cisco

NVMe-PCIe (Registers)

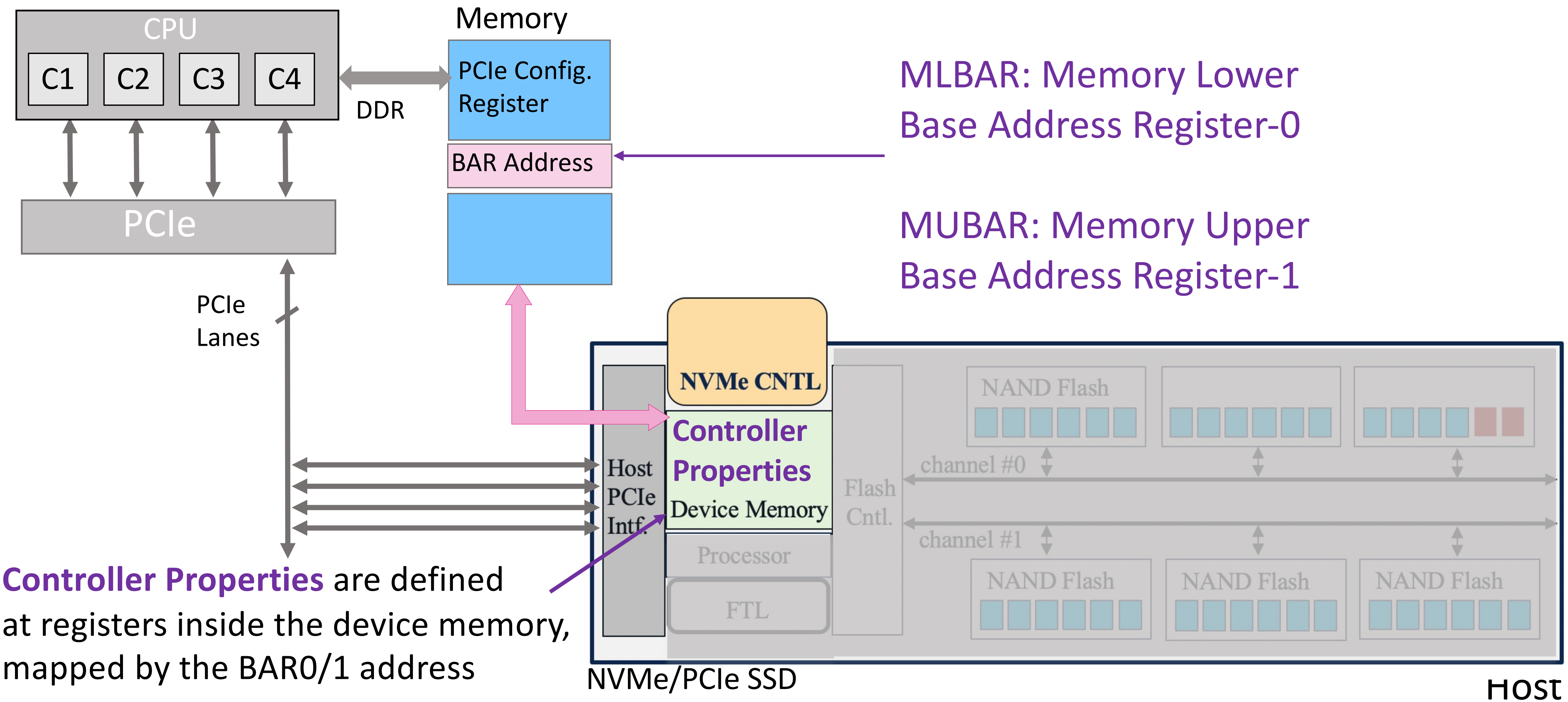
During PCIe enumeration "Class Code" is read



Kamal Bakshi, Cisco

NVMe-PCIe (Registers)

BAR registers maps Device Memory Registers into CPU memory



MLBAR: Memory Lower Base Address Register-0

MUBAR: Memory Upper Base Address Register-1

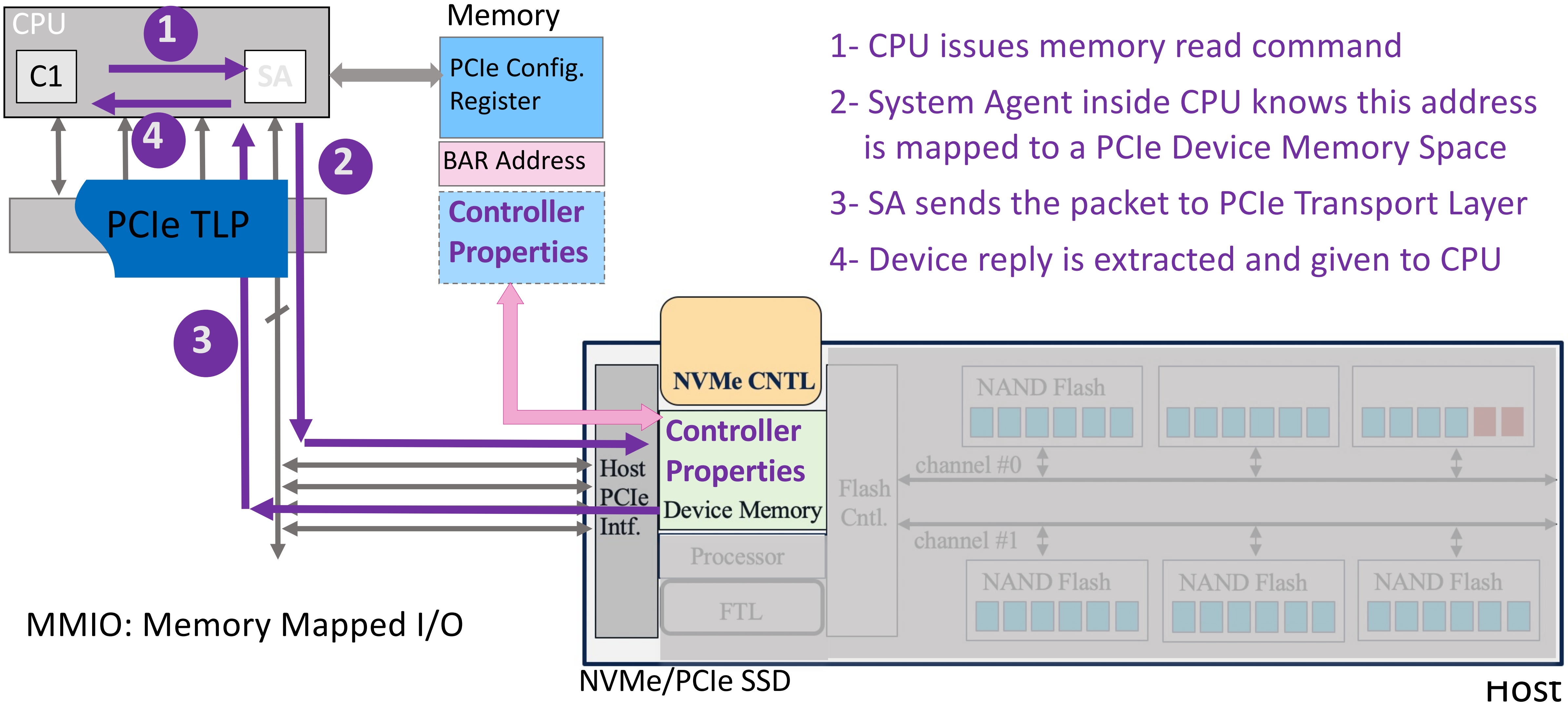
Controller Properties are defined at registers inside the device memory, mapped by the BAR0/1 address

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

NVMe-PCIe (Properties)

How does CPU reads the Controller Properties Register ?



1- CPU issues memory read command

2- System Agent inside CPU knows this address is mapped to a PCIe Device Memory Space

3- SA sends the packet to PCIe Transport Layer

4- Device reply is extracted and given to CPU

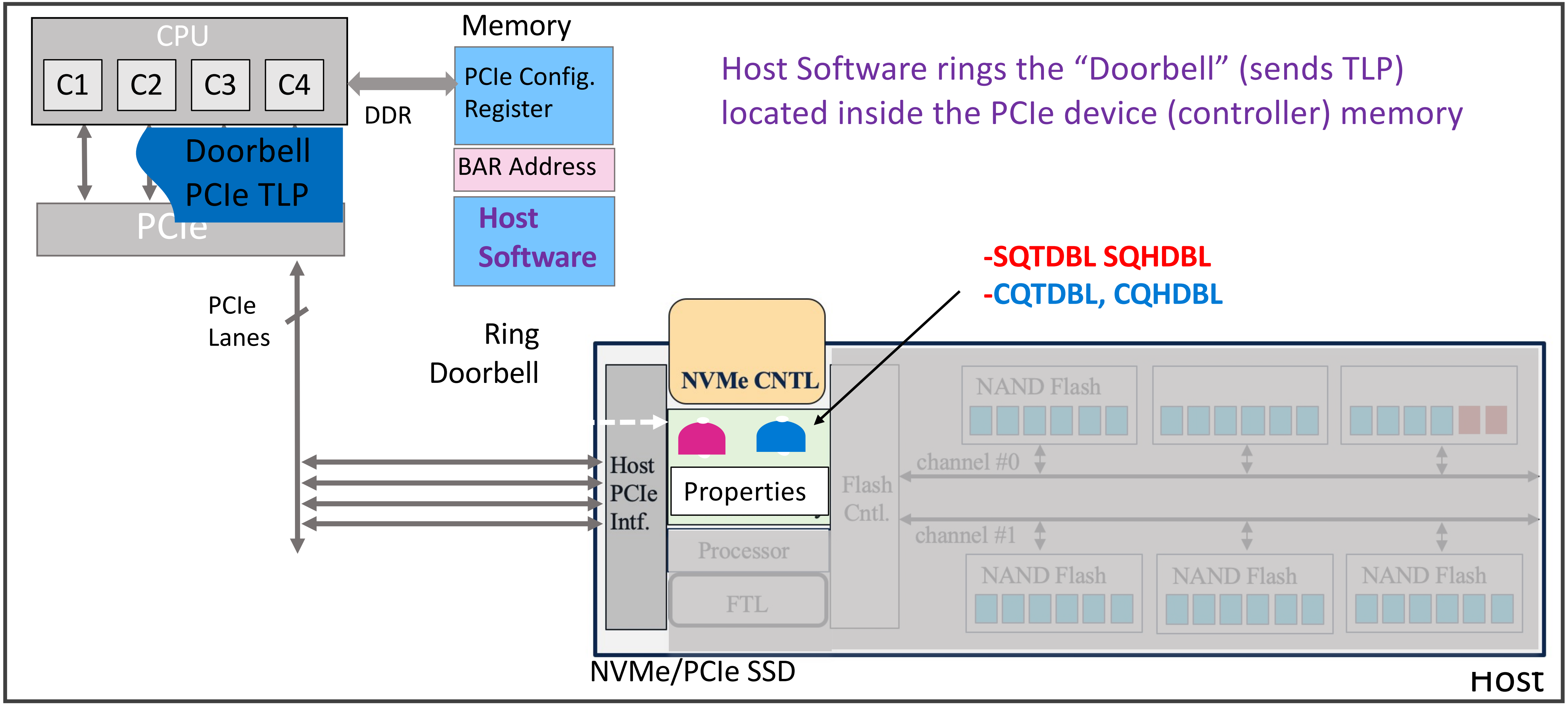
MMIO: Memory Mapped I/O

Host

Kamal Bakshi, Cisco

NVMe-PCIe (Doorbell)

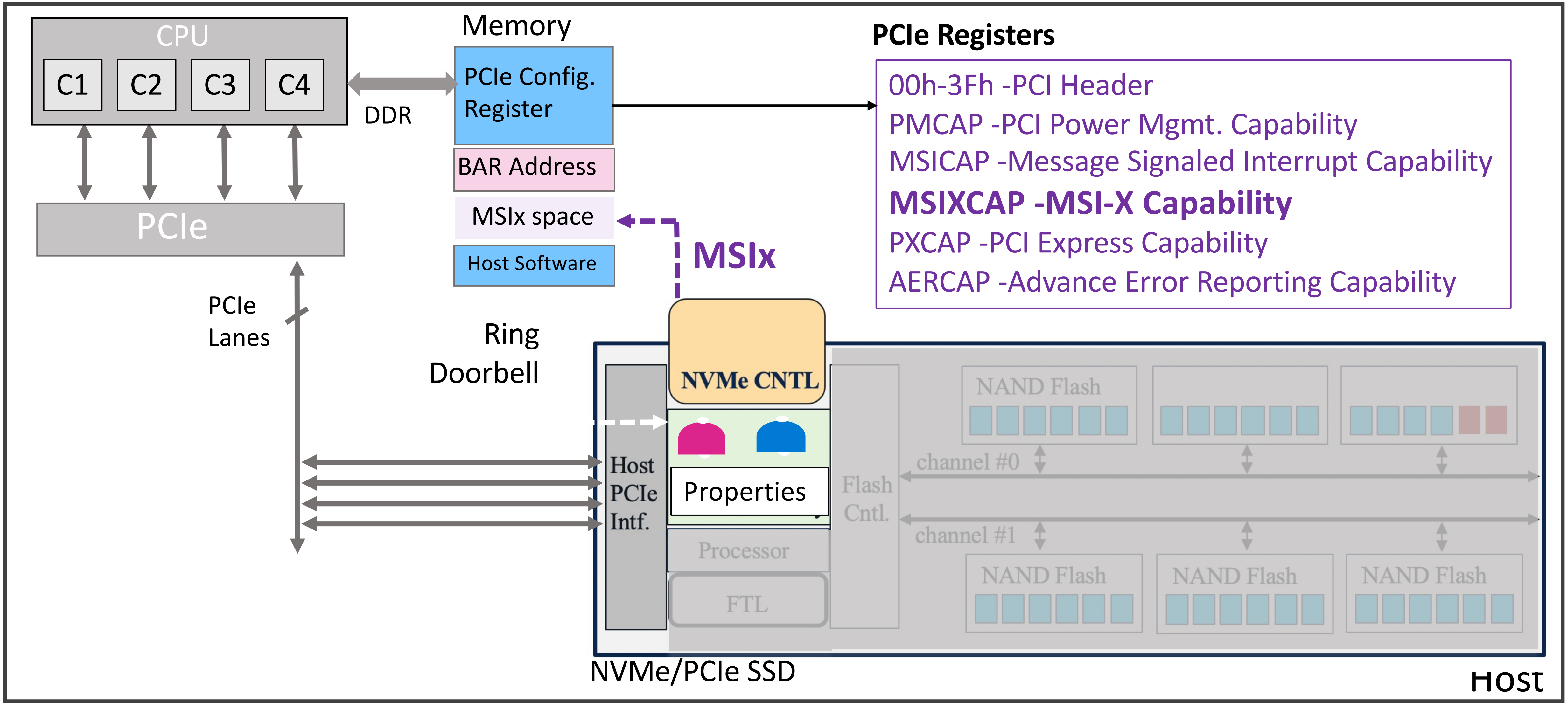
How does "Host Software" informs Controller about pending tasks?



Kamal Bakshi, Cisco

NVMe-PCIe (MSIx)

How does "Controller" informs Host about pending tasks ?



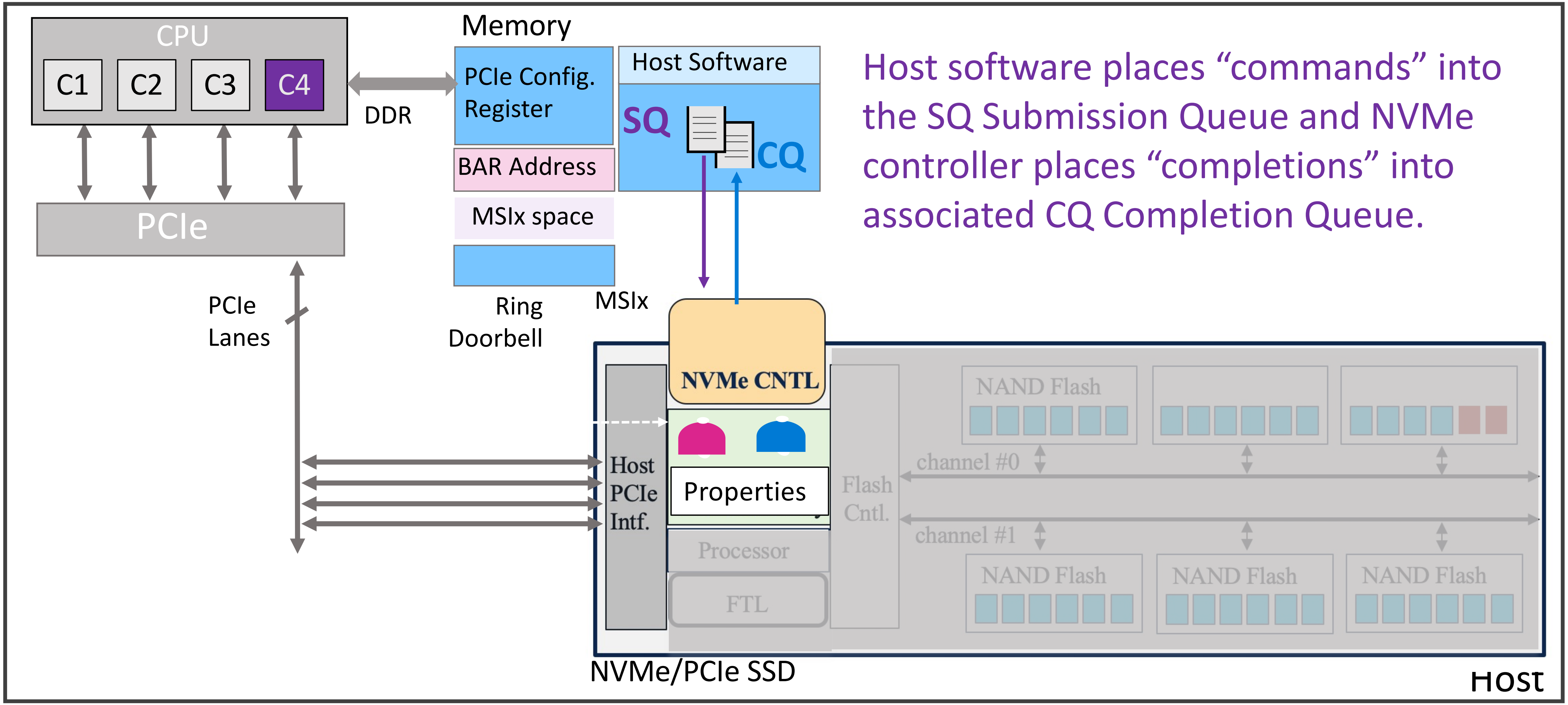
PCIe Registers

- 00h-3Fh -PCI Header
- PMCAP -PCI Power Mgmt. Capability
- MSICAP -Message Signaled Interrupt Capability
- **MSIXCAP -MSI-X Capability**
- PXCAP -PCI Express Capability
- AERCAP -Advance Error Reporting Capability

Kamal Bakshi, Cisco

NVMe-PCIe (SQ/CQ Pair)

NVMe is based on a paired Submission and Completion Queue mechanism.



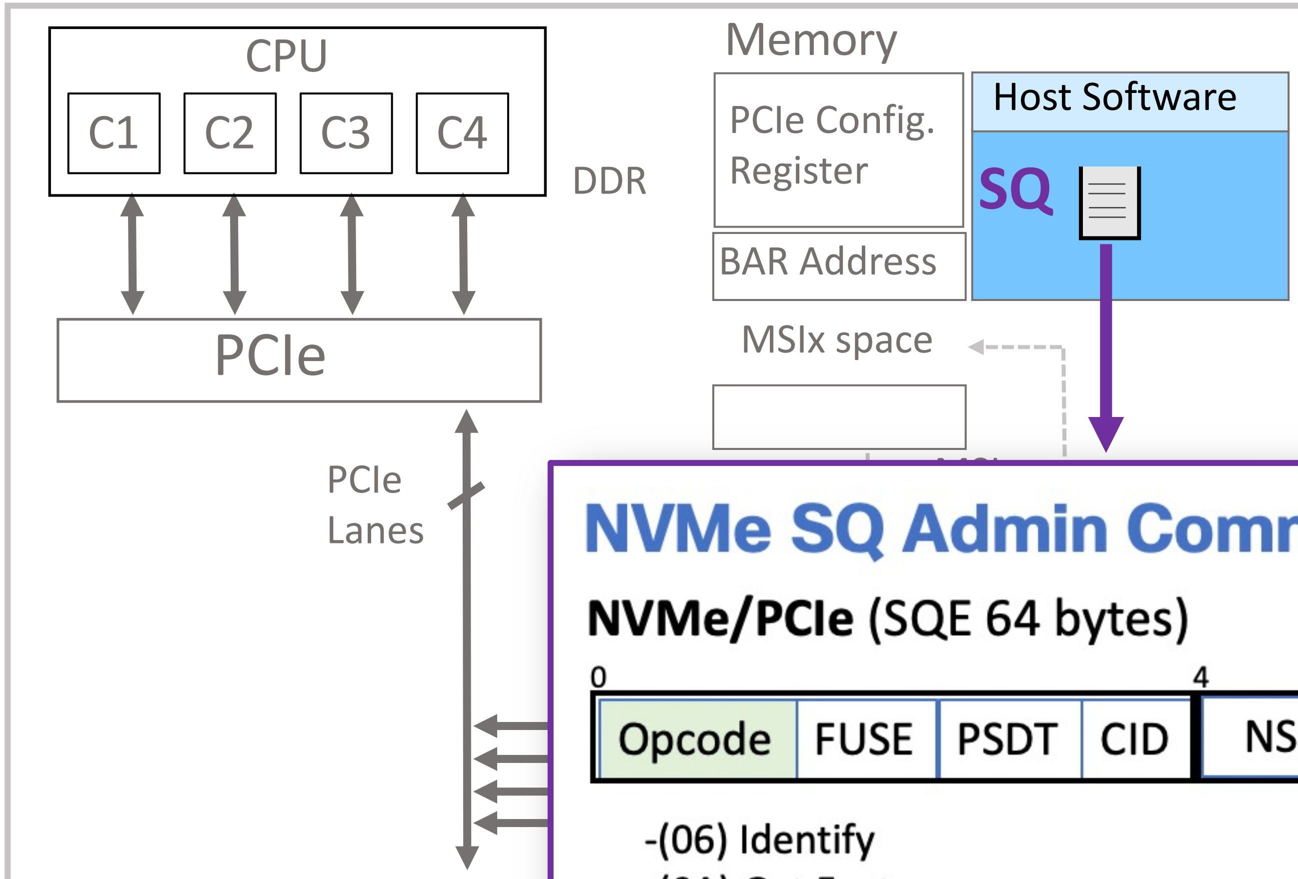
Host software places “commands” into the SQ Submission Queue and NVMe controller places “completions” into associated CQ Completion Queue.

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

NVMe-PCIe (SQE)

Submission Queue Entry



NVMe SQ Admin Commands **64B** -Queue #0

NVMe/PCIe (SQE 64 bytes)

0				4	8	16	24	40–63
Opcode	FUSE	PSDT	CID	NSID	DW2-3	MPTR	DPTR/SGL/PRP	DW 10-15

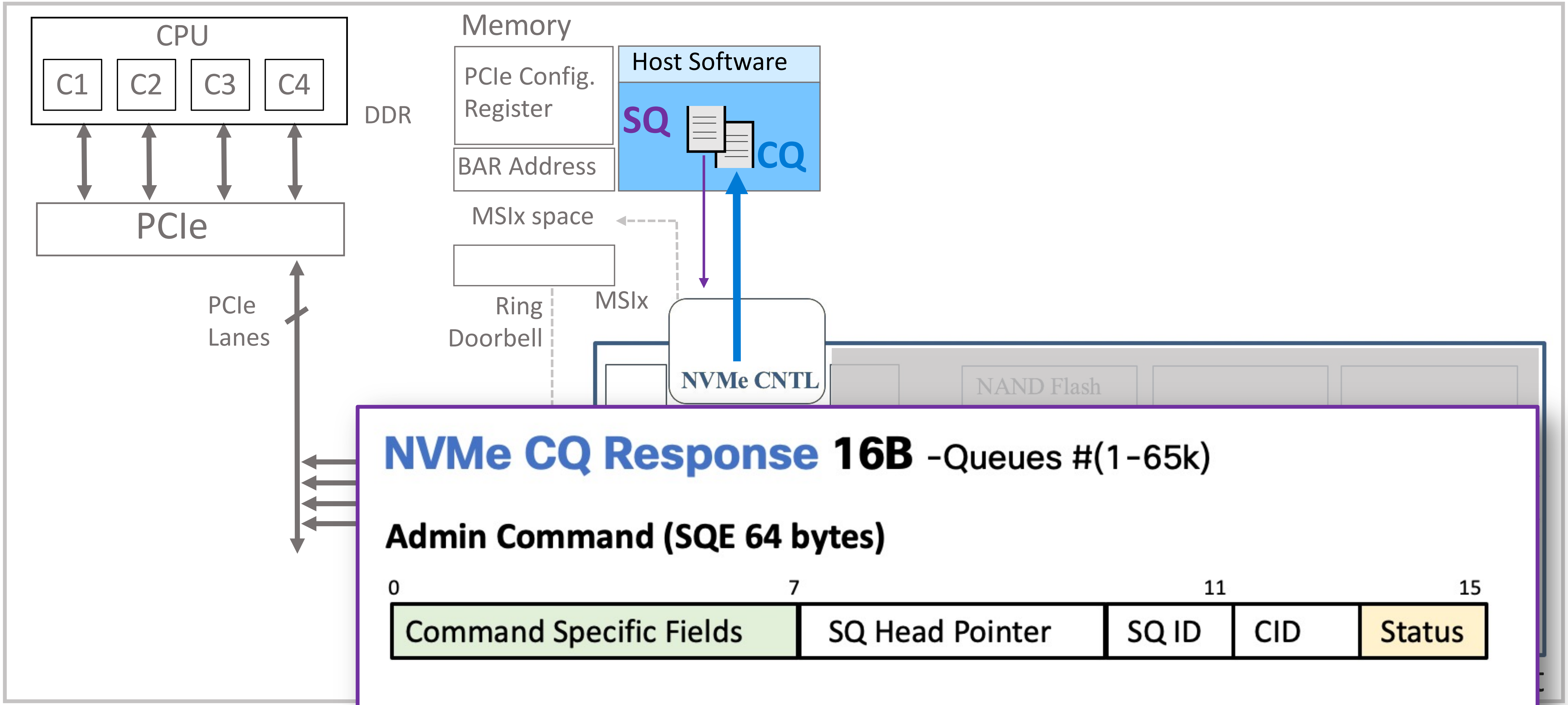
- -(06) Identify
- -(0A) Get Feature
- -(09) Set Feature
- -(02) Get Log Page
- -(0C) Async. Event Req.
- -(08) Abort

- -(01) Create I/O SQ
- -(00) Delete I/O SQ
- -(05) Create I/O CQ
- -(04) Delete I/O CQ

Kamal Bakshi, Cisco

NVMe-PCIe (CQE)

Completion Queue Entry



NVMe CQ Response 16B -Queues #(1-65k)

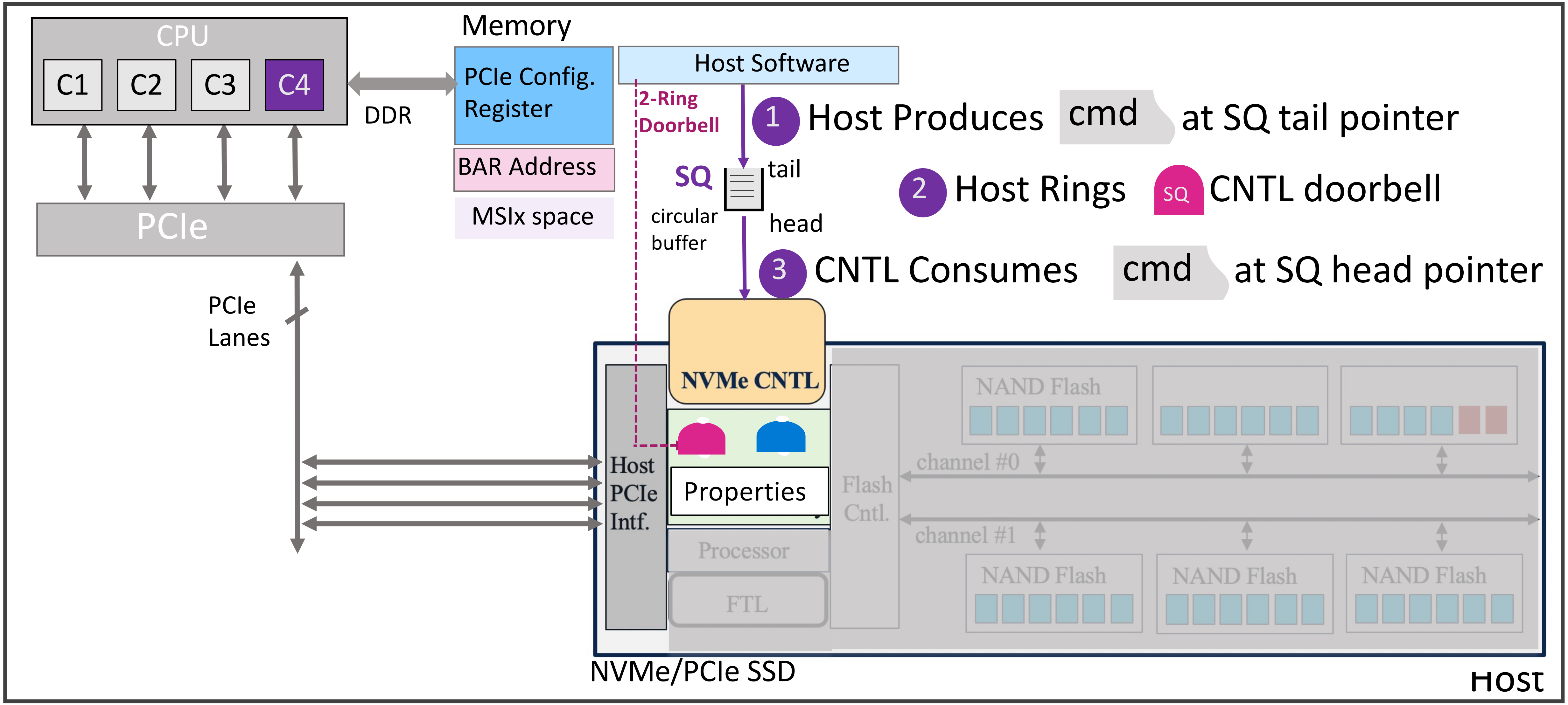
Admin Command (SQE 64 bytes)

0–7		11		15
Command Specific Fields	SQ Head Pointer	SQ ID	CID	Status

Kamal Bakshi, Cisco

NVMe-PCIe (Host to CNTL)

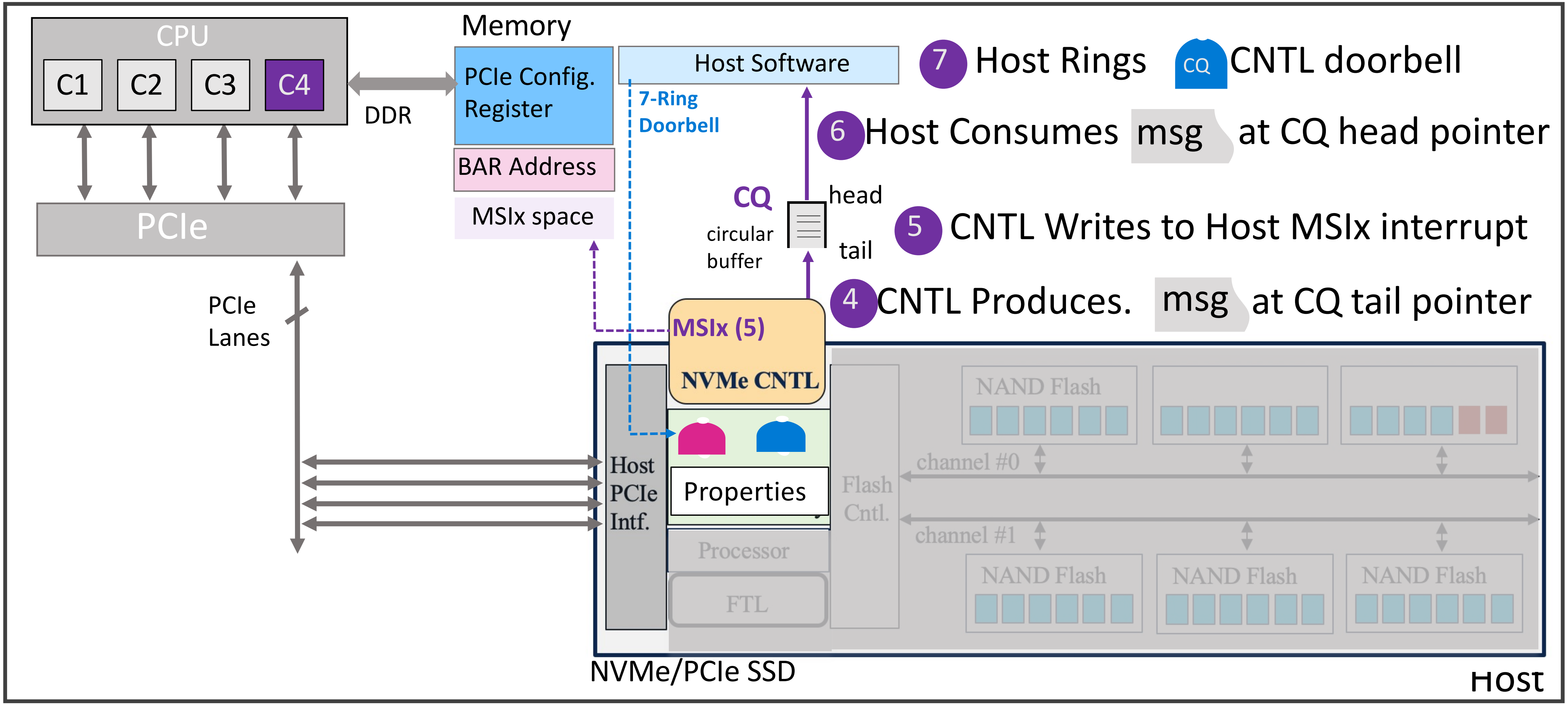
NVMe Queuing mechanism details



Kamal Bakshi, Cisco

NVMe-PCIe (CNTL to Host)

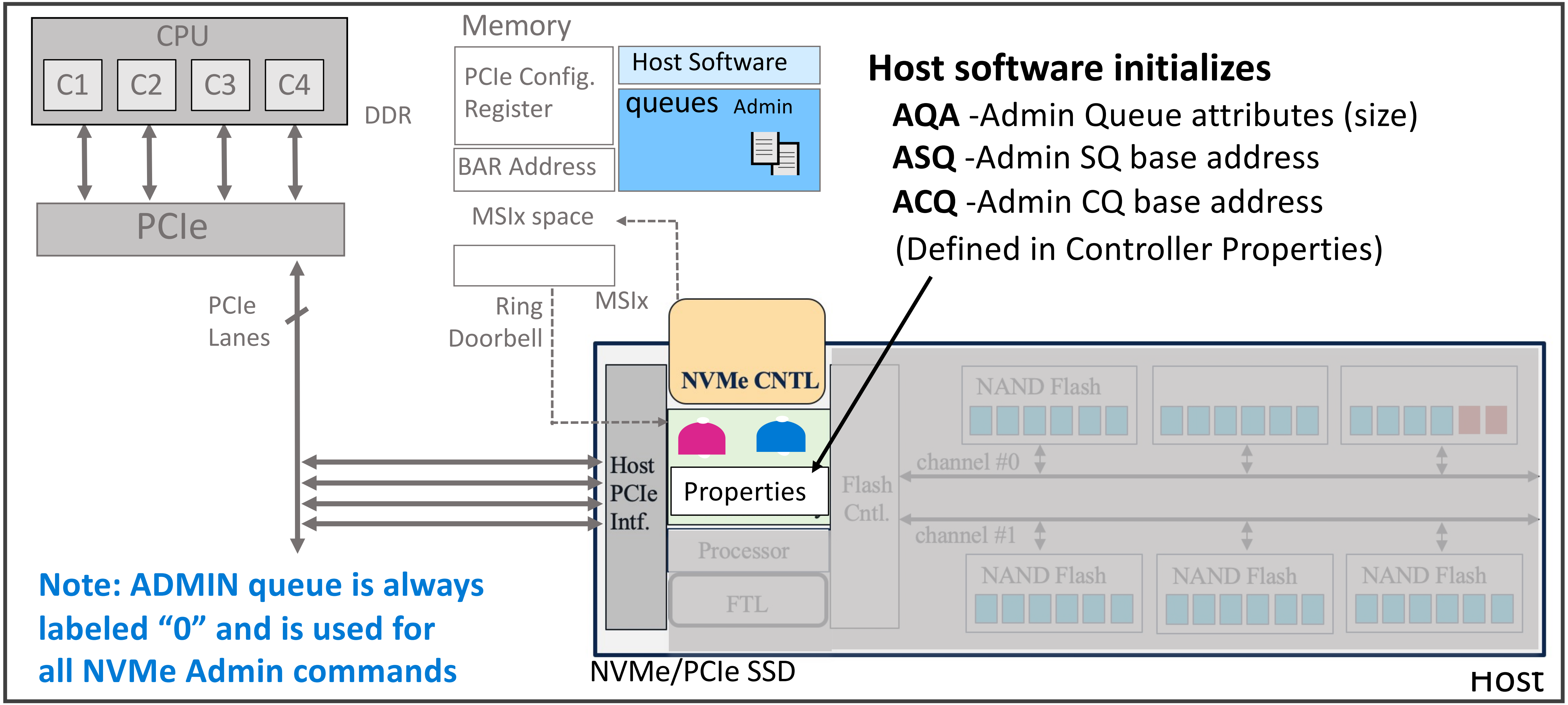
NVMe Queuing mechanism details



Kamal Bakshi, Cisco

NVMe-PCIe (Admin_Q)

Admin queues are created first and are used for "Administrative Tasks"



Host software initializes

AQA -Admin Queue attributes (size)

ASQ -Admin SQ base address

ACQ -Admin CQ base address

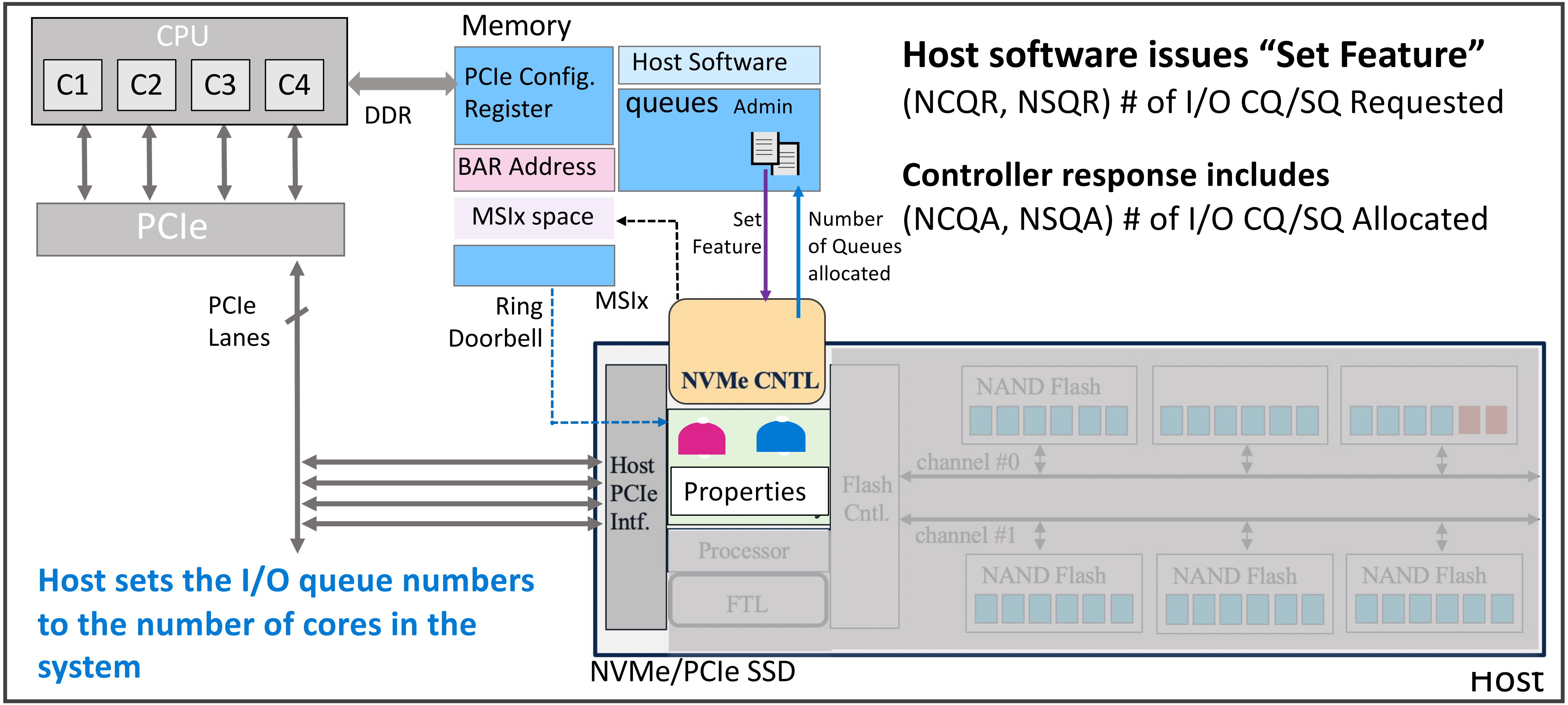
(Defined in Controller Properties)

Note: ADMIN queue is always labeled "0" and is used for all NVMe Admin commands

Kamal Bakshi, Cisco

NVMe-PCIe (I/O queues)

Using Admin Queues Host starts the I/O queues creation process



Host software issues "Set Feature"

(NCQR, NSQR) # of I/O CQ/SQ Requested

Controller response includes

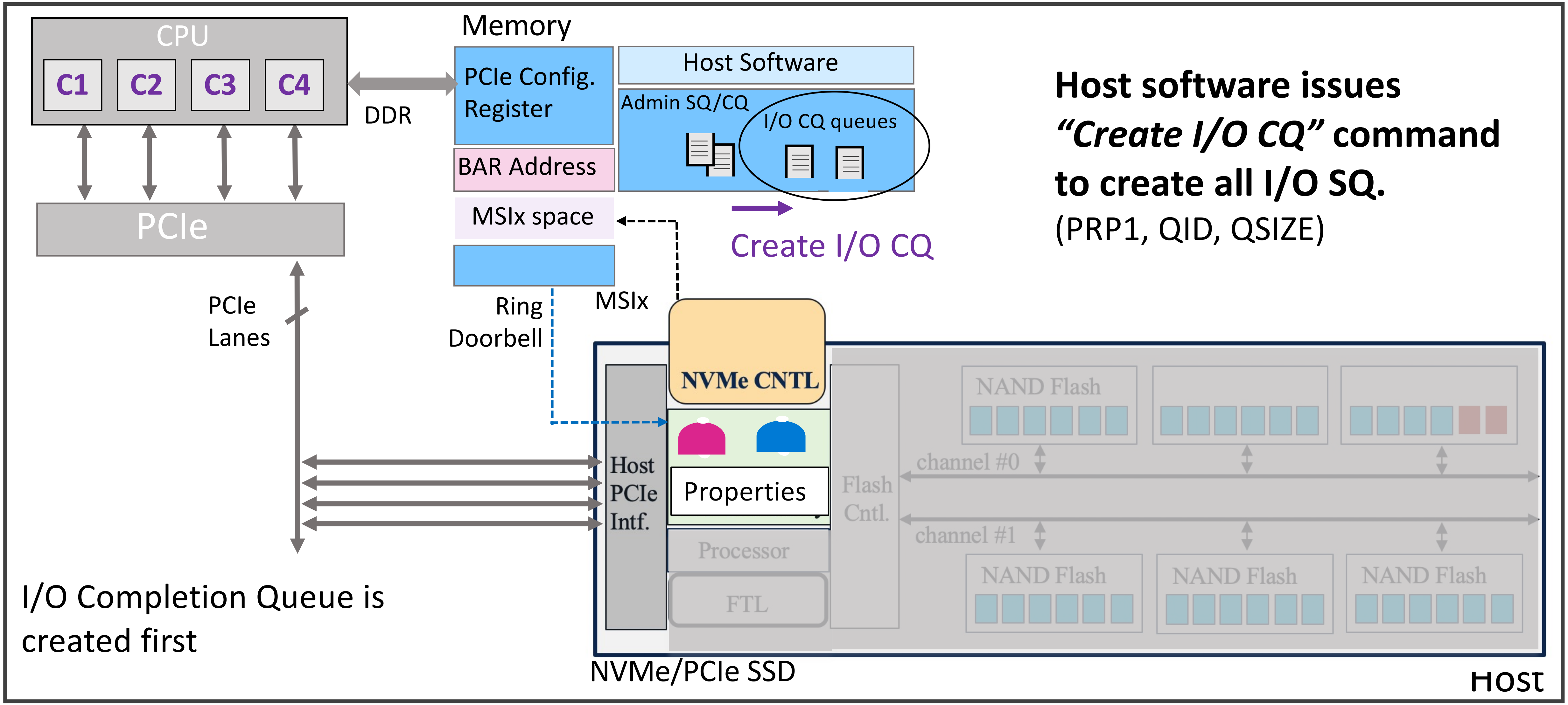
(NCQA, NSQA) # of I/O CQ/SQ Allocated

Host sets the I/O queue numbers to the number of cores in the system

Kamal Bakshi, Cisco

NVMe-PCIe (I/O queues)

Primary purpose for I/O queues is to transfer data (read/write)



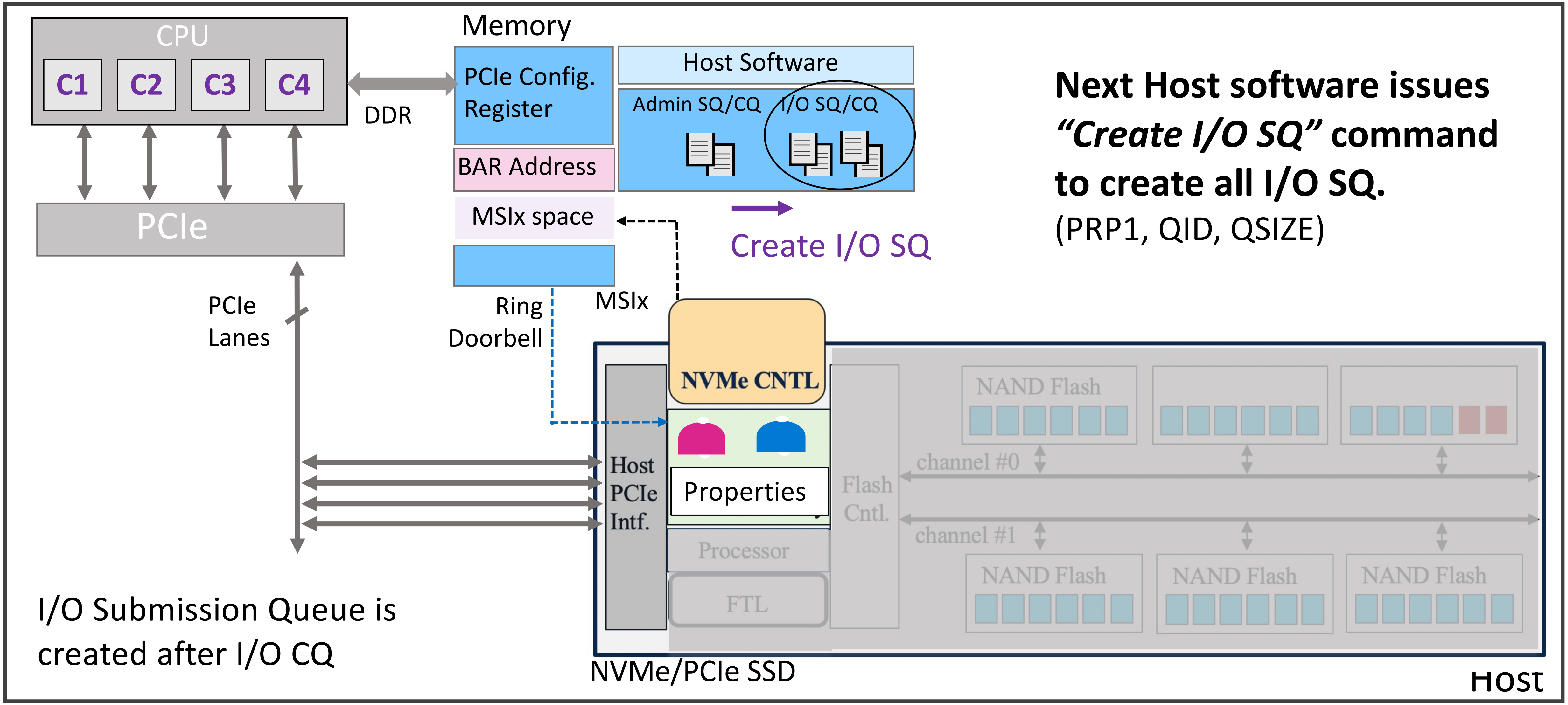
Host software issues *"Create I/O CQ"* command to create all I/O SQ.
(PRP1, QID, QSIZE)

I/O Completion Queue is created first

Kamal Bakshi, Cisco

NVMe-PCIe (I/O queues)

Primary purpose for I/O queues is to transfer data (read/write)



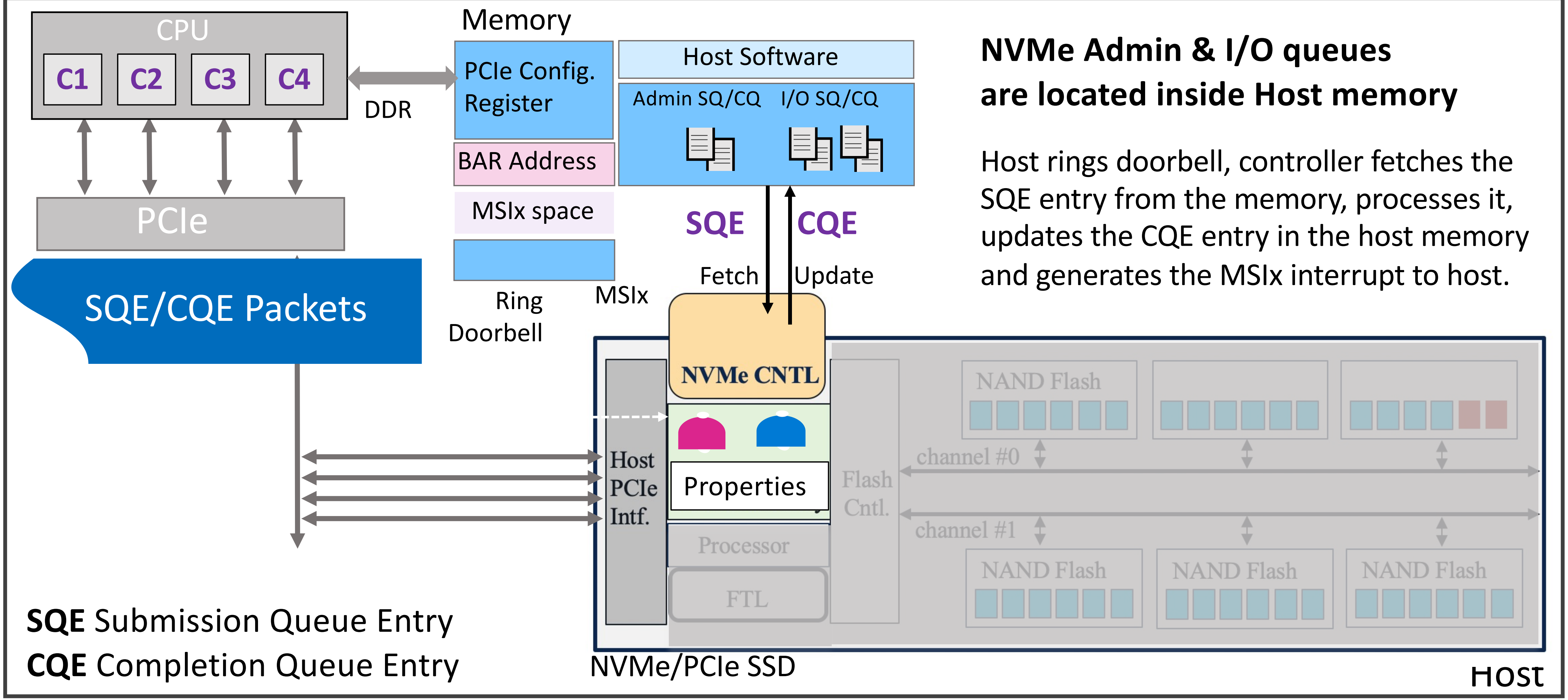
Next Host software issues *"Create I/O SQ"* command to create all I/O SQ.
(PRP1, QID, QSIZE)

I/O Submission Queue is created after I/O CQ

Kamal Bakshi, Cisco

NVMe-PCIe (MMIO)

Data transfer mechanism for Admin and I/O command data through memory MMIO



NVMe Admin & I/O queues are located inside Host memory

Host rings doorbell, controller fetches the SQE entry from the memory, processes it, updates the CQE entry in the host memory and generates the MSIx interrupt to host.

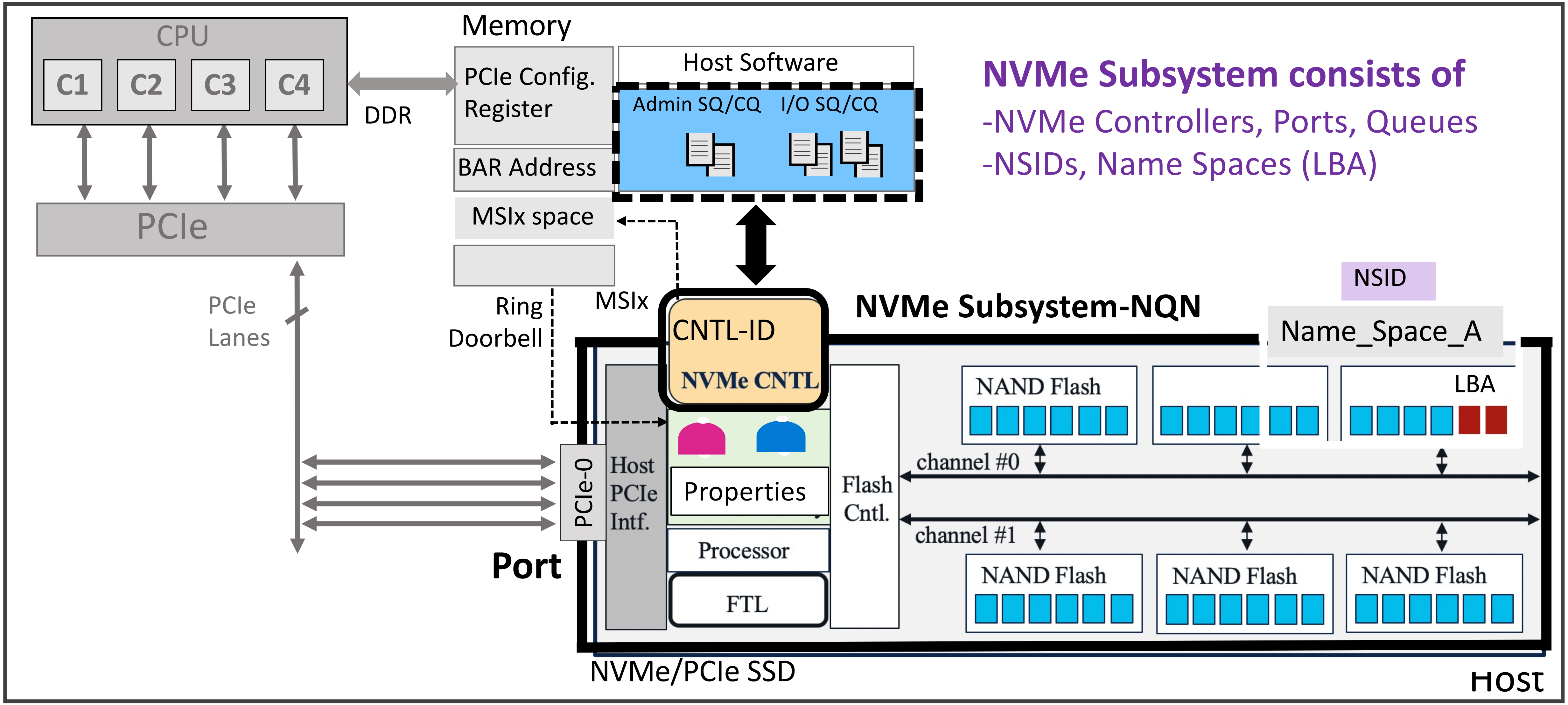
SQE Submission Queue Entry
CQE Completion Queue Entry

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

NVMe-PCIe (NVMe Subsystem)

What is NVMe Subsystem?



NVMe Subsystem consists of

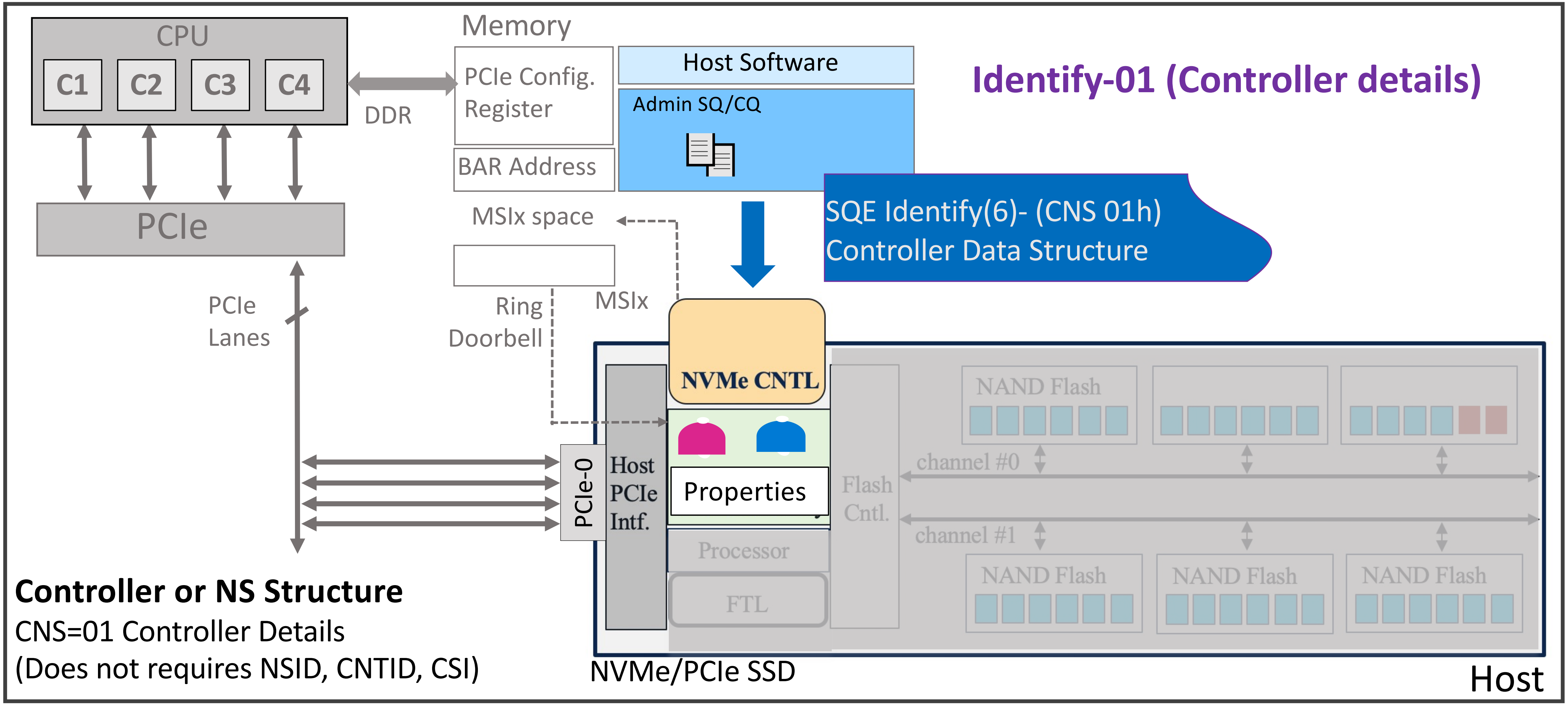
-NVMe Controllers, Ports, Queues

-NSIDs, Name Spaces (LBA)

Kamal Bakshi, Cisco

NVMe-PCIe (Identify-01)

Host issues series of "Identify" commands to get NVMe Subsystem details



Identify-01 (Controller details)

SQE Identify(6)- (CNS 01h)
Controller Data Structure

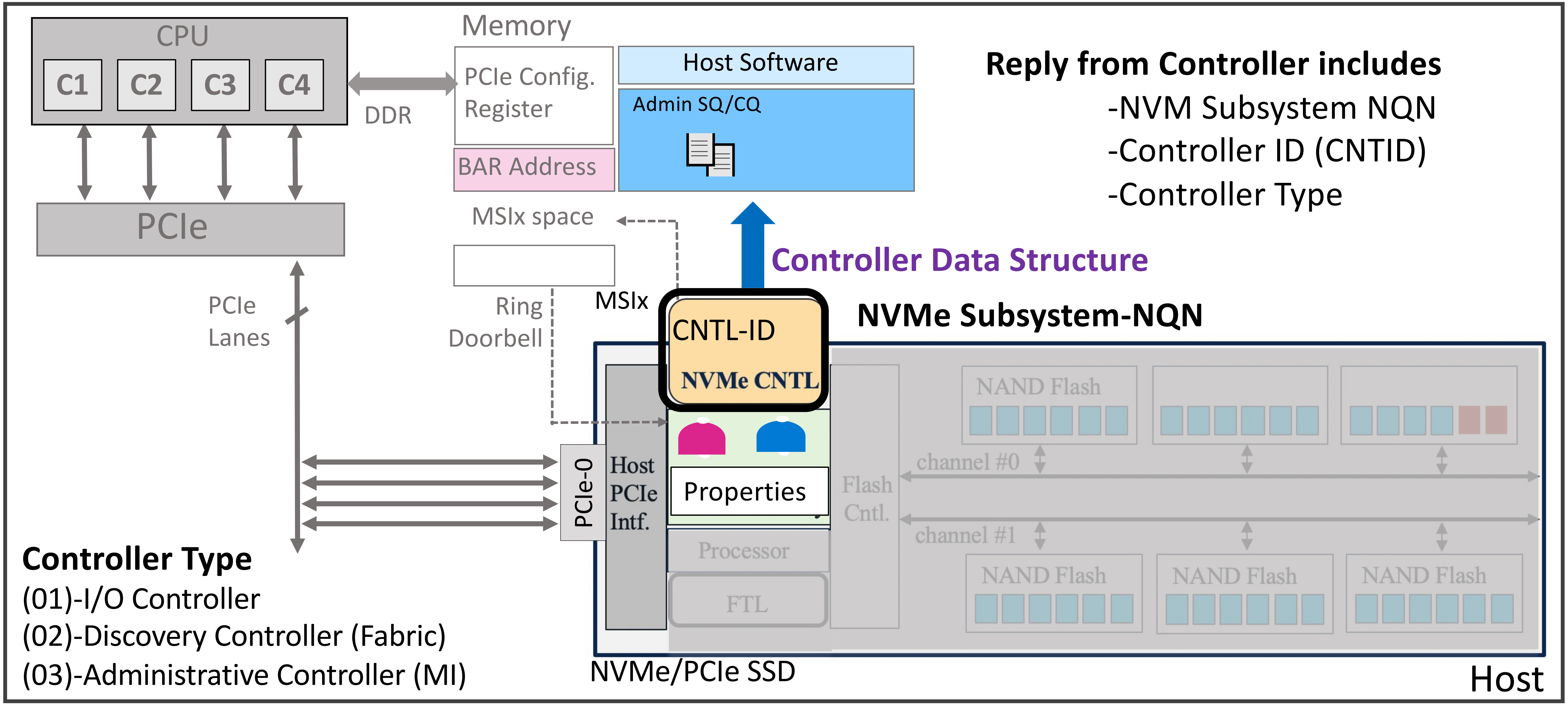
Controller or NS Structure

CNS=01 Controller Details

(Does not requires NSID, CNTID, CSI)

Kamal Bakshi, Cisco

NVMe-PCIe (Identify Reply-01)



Reply from Controller includes

- -NVM Subsystem NQN
- -Controller ID (CNTID)
- -Controller Type

Controller Type

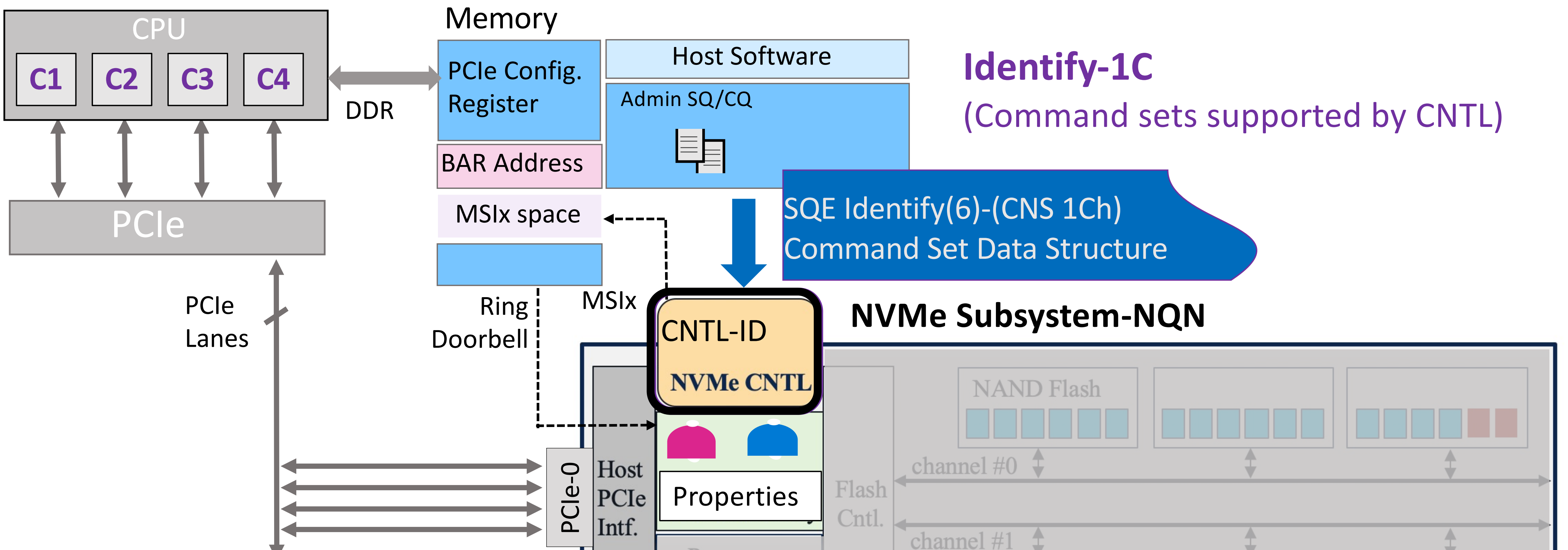
(01)-I/O Controller

(02)-Discovery Controller (Fabric)

(03)-Administrative Controller (MI)

Kamal Bakshi, Cisco

NVMe-PCIe (Identify-1C)



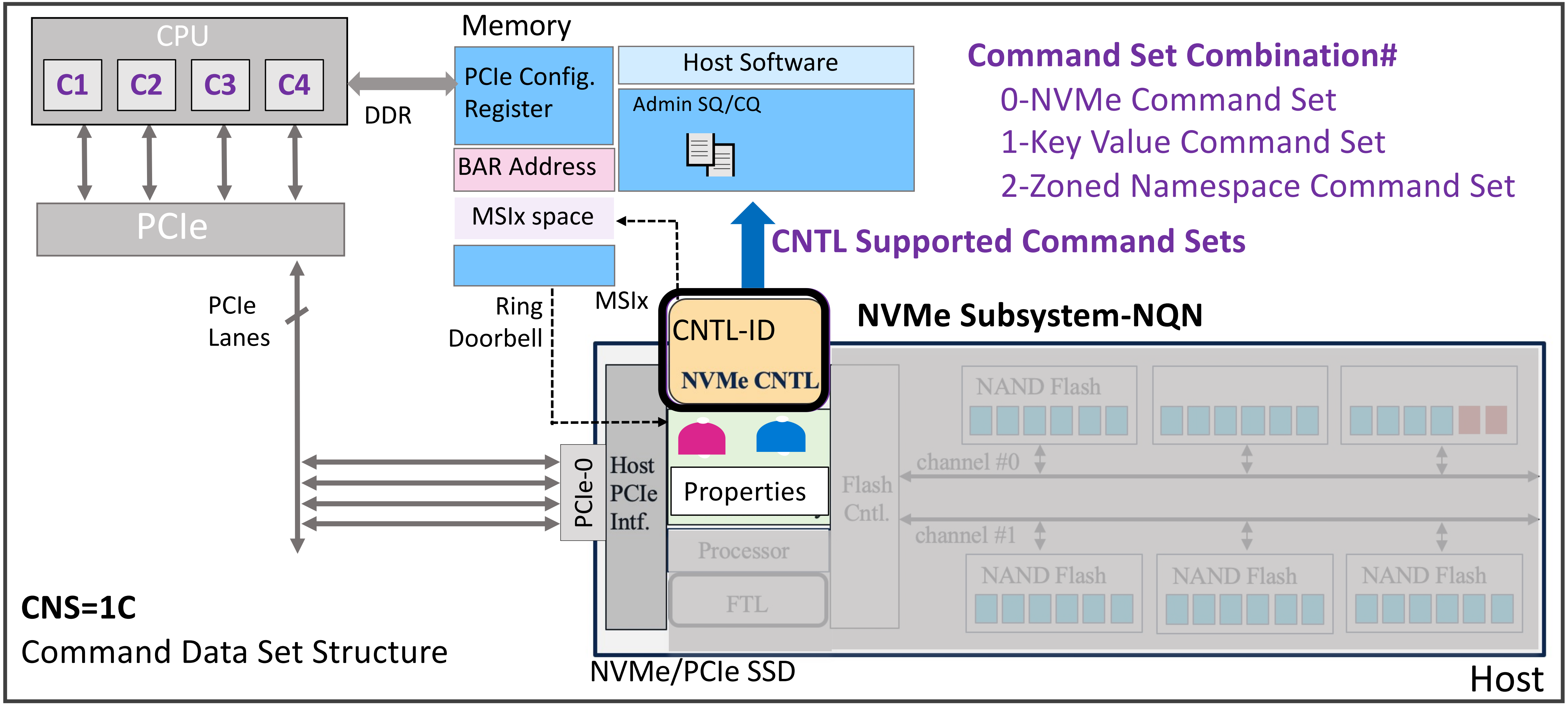
Identify-1C

(Command sets supported by CNTL)

A Controller may support one or more I/O command sets but each namespace, when created, is associated with exactly one I/O command set.

Kamal Bakshi, Cisco

NVMe-PCIe (Identify-1C Reply)



Command Set Combination#

- 0-NVMe Command Set
- 1-Key Value Command Set
- 2-Zoned Namespace Command Set

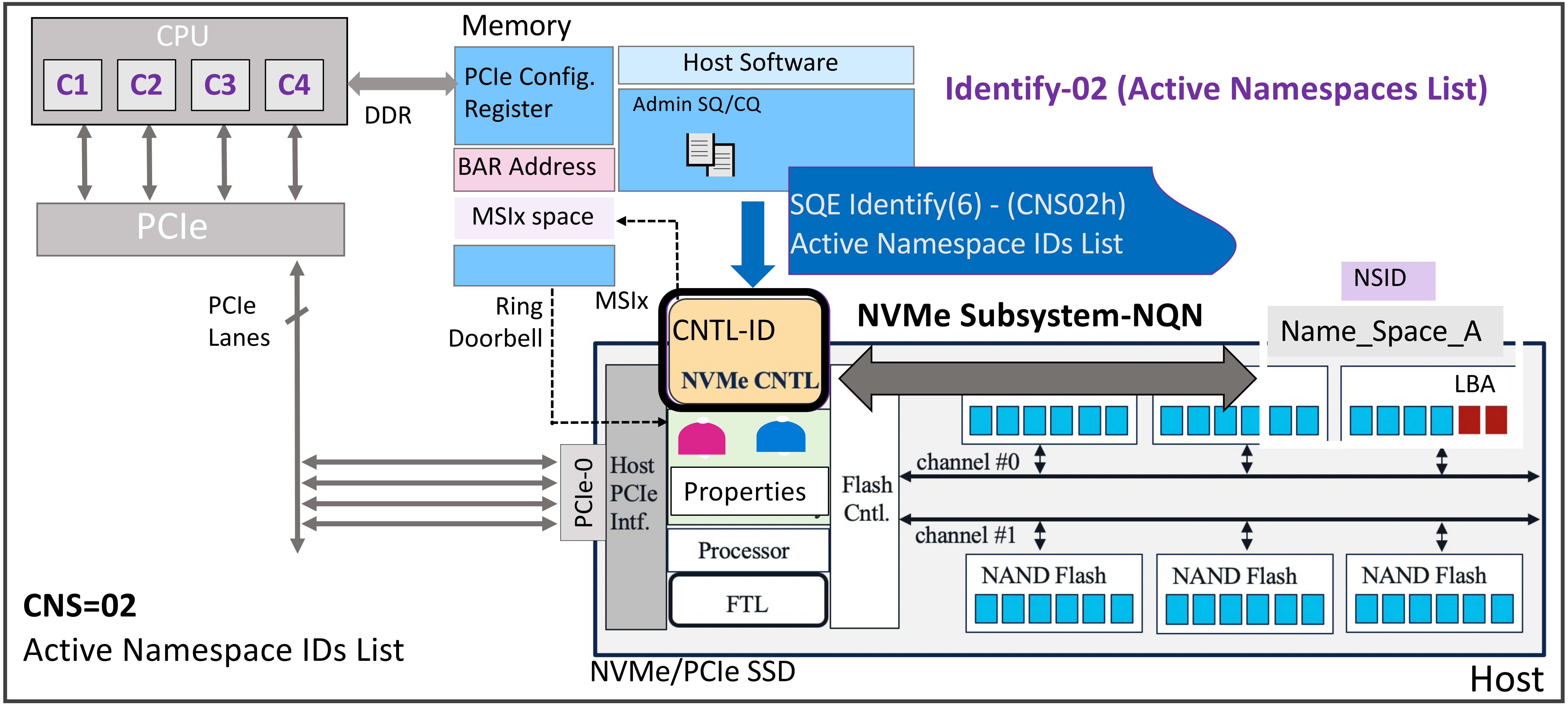
CNTL Supported Command Sets

CNS=1C

Command Data Set Structure

Kamal Bakshi, Cisco

NVMe-PCIe (Identify-07)

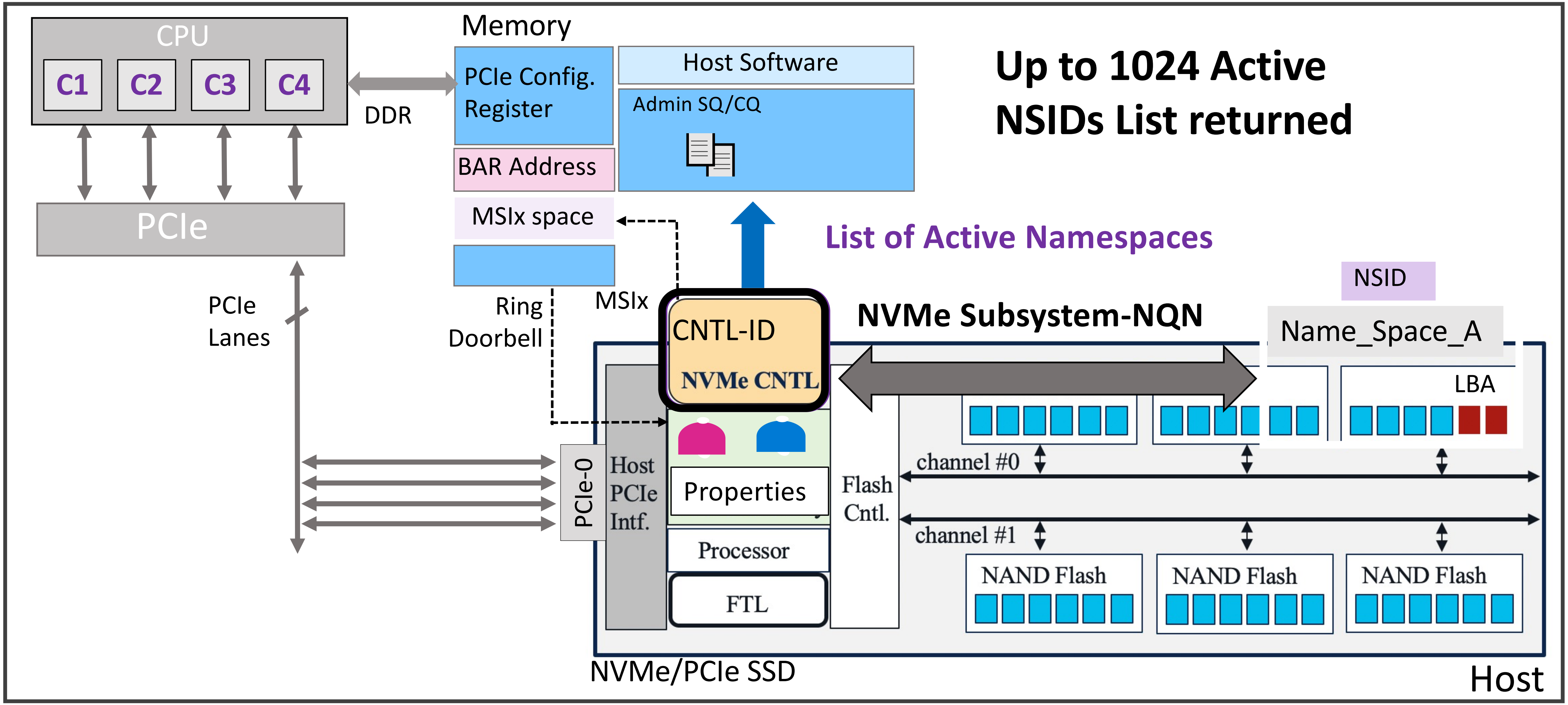


CNS=02

Active Namespace IDs List

Kamal Bakshi, Cisco

NVMe-PCIe (Identify-07 Reply)

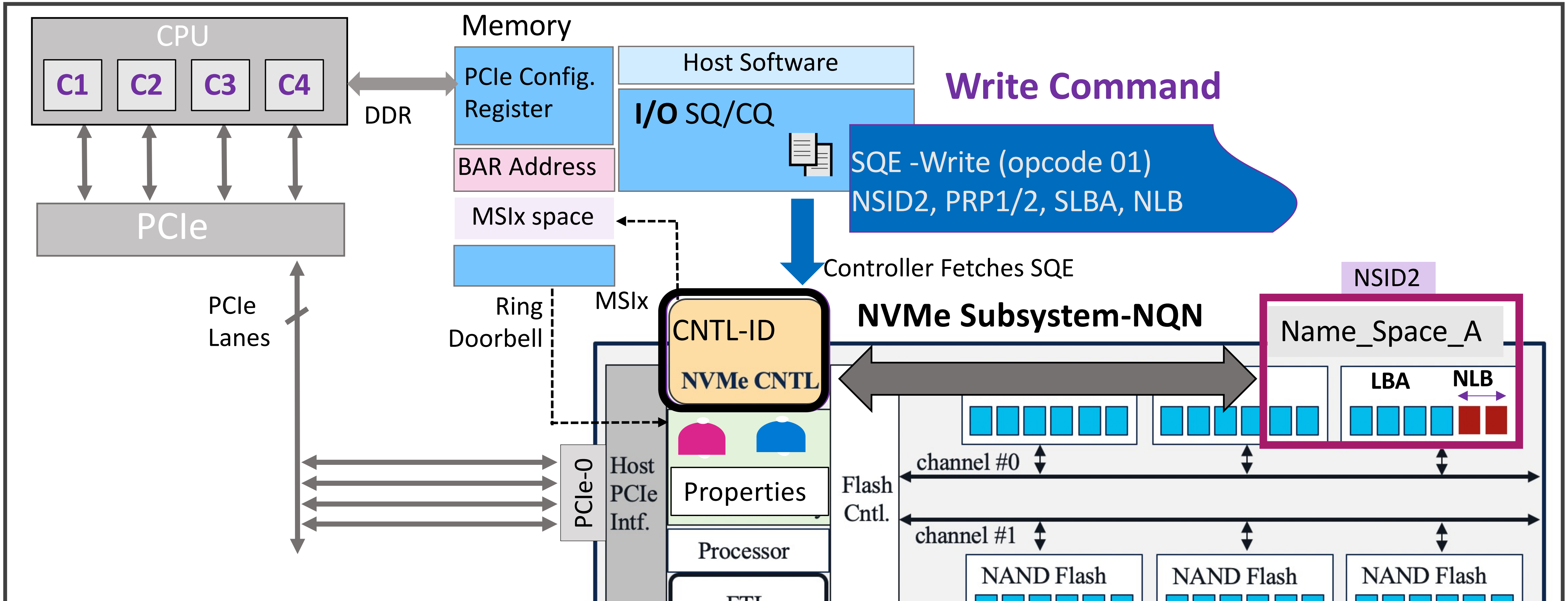


Up to 1024 Active NSIDs List returned

List of Active Namespaces

Kamal Bakshi, Cisco

NVMe-PCIe (SQE-Write)

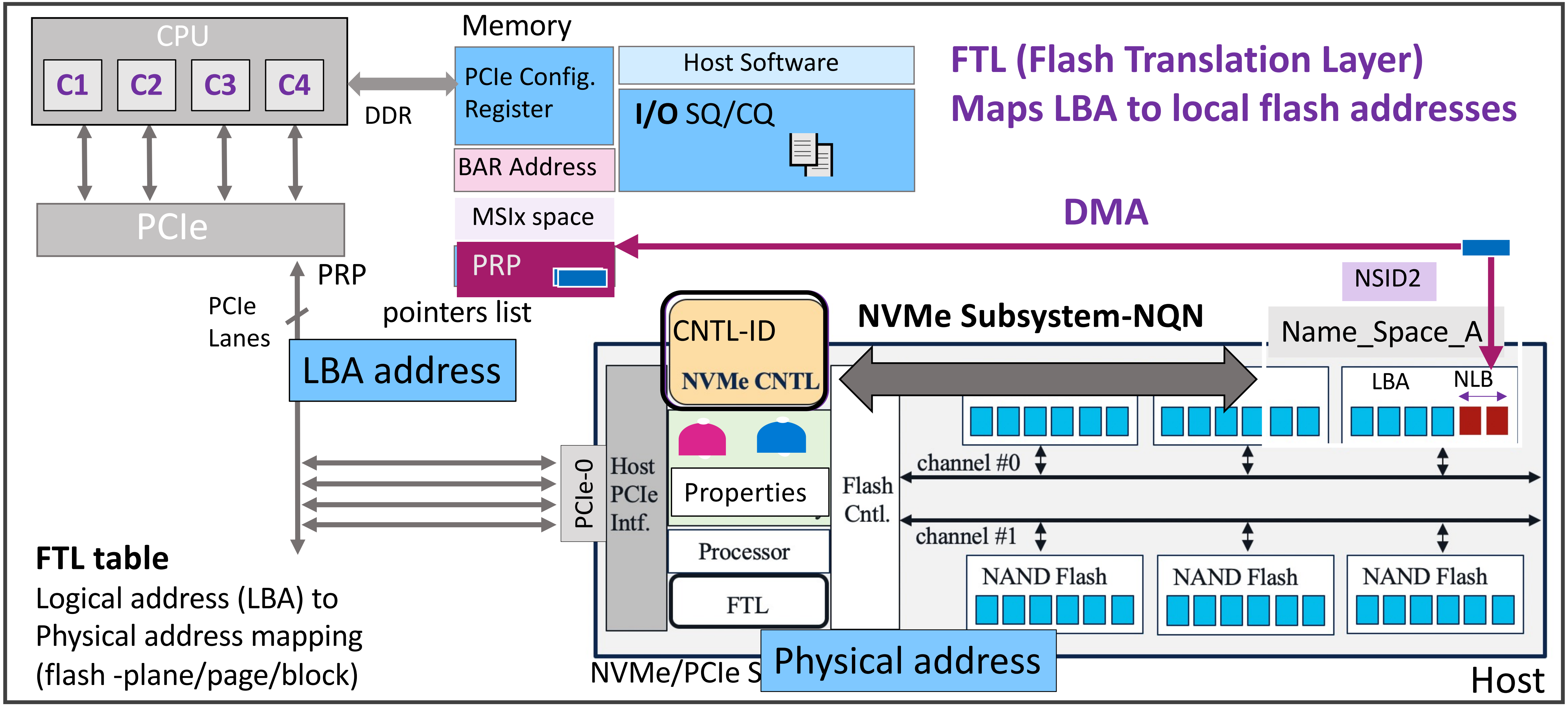


PRP: Physical Region Page

NLB: number of logical blocks

NVMe-PCIe (DMA Data)

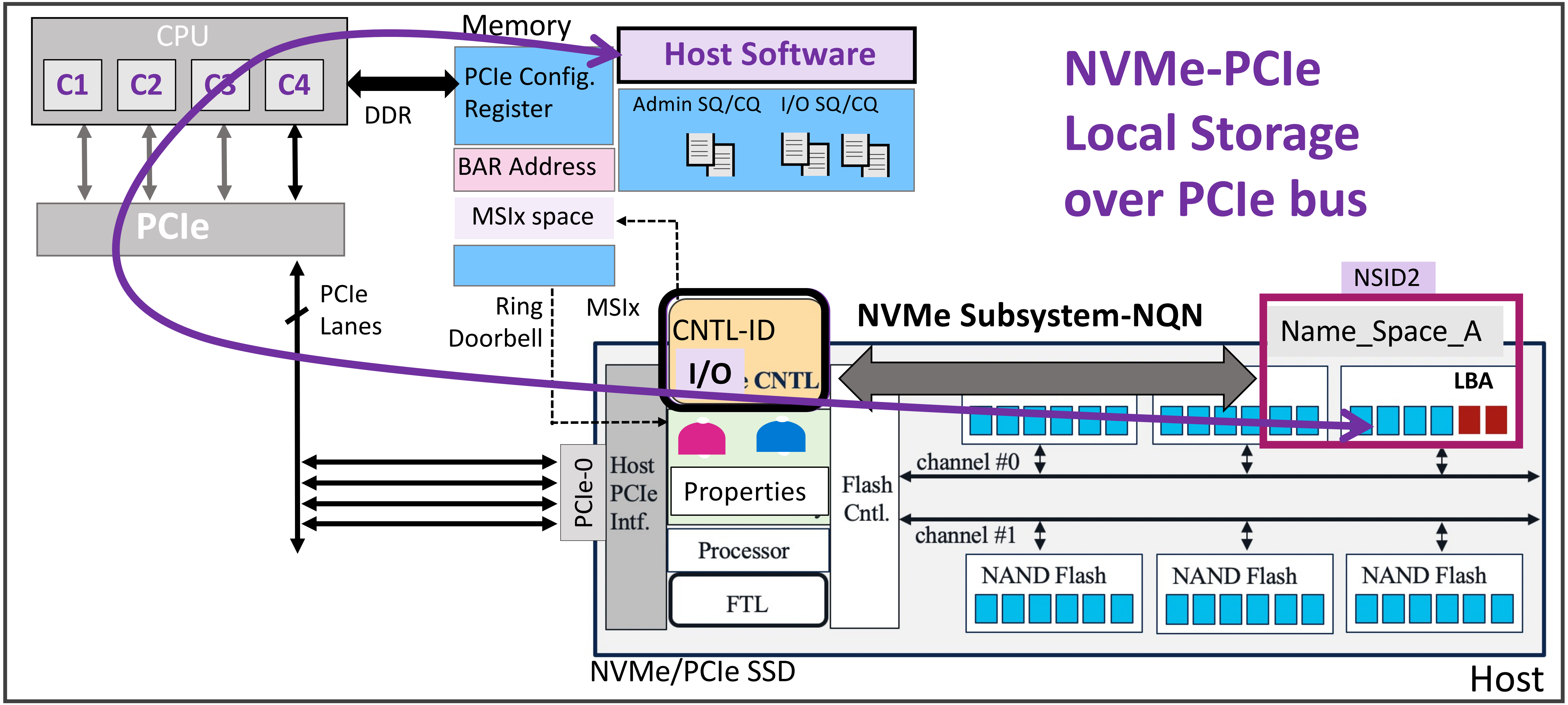
Controller Reads Data from PRP buffers and writes it to the Flash



FTL table
Logical address (LBA) to Physical address mapping (flash -plane/page/block)

Kamal Bakshi, Cisco

NVMe-PCIe

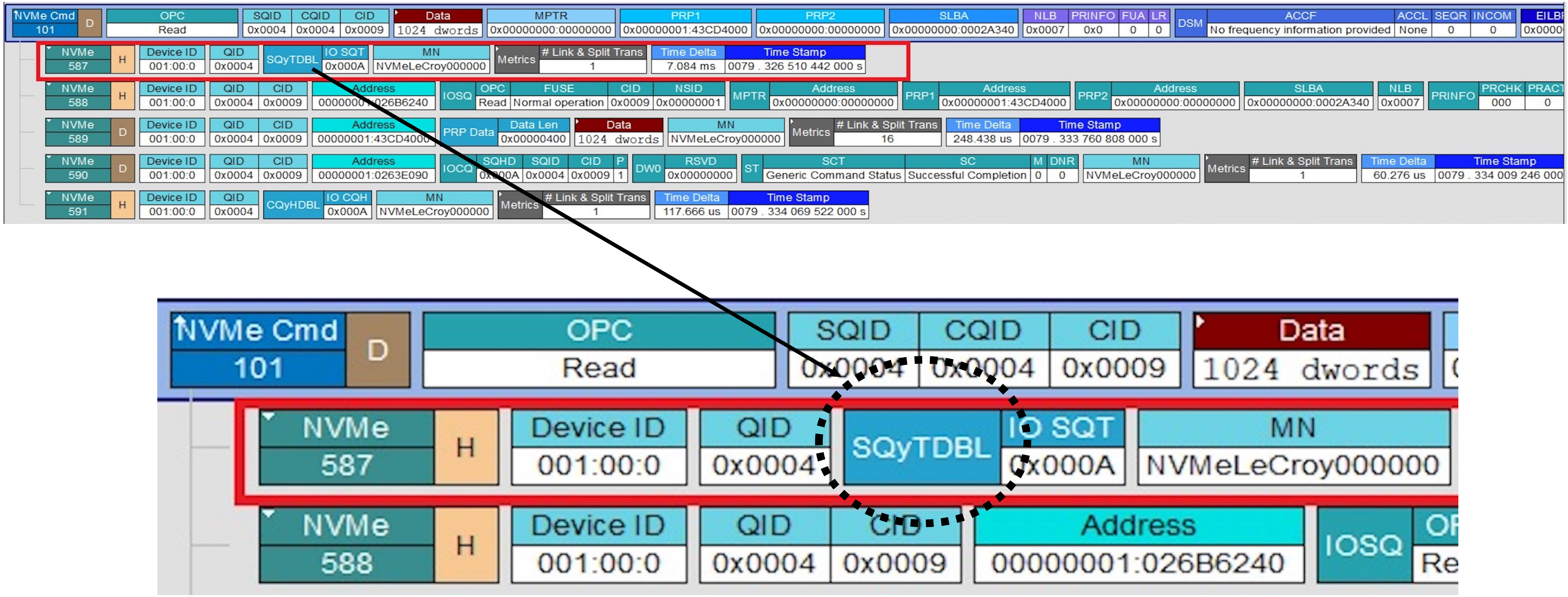


NVMe-PCIe
Local Storage
over PCIe bus

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

NVMe-PCIe Trace of a Doorbell Message



SQ Tail Doorbell

Trace: Courtesy of
Teledyne Technologies

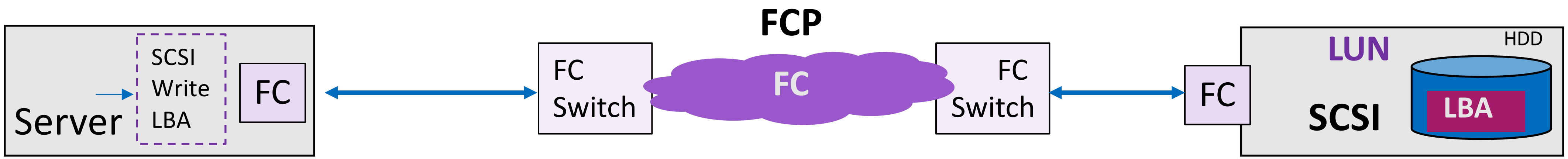
Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE

NVMe-FC Packet Examples

Kamal Bakshi, Cisco

FCP (SCSI Protocol mapped into Fibre Channel)



SOF	**FC Frame header**	**SCSI Payload**	CRC	EOF

SCSI WRITE (16) Command

Operation Code (8Ah)					
WRPROTECT	DPO	FUA	Rsvd	Obsolete	DLD2
Logical Block Address (LBA)					
Transfer Length					
DLD1	DLD0	Group Number			
Control					

FCP Command IU Payload

FCP_LUN		
Command Reference Number		
Rsvd	Command Priority	Task Attribute
Task Management Flags		
Additional FCP_CDB Length	RDDATA	WRDATA
FCP_CDB		
Additional FCP_CDB (if any)		
FCP_DL		
FCP_Bidirectional_Read_DL (if any)		

FC Frame Header

R_CTL	D_ID	
CS_CTL	S_ID	
TYPE	F_CTL	
SEQ_ID	DF_CTL	SEQ_CNT
OX_ID	RX_ID	
Parameter		
FCP Payload		

Kamal Bakshi, Cisco

FCP Protocol

SOF	Frame Header	Payload	CRC	EOF

00

R_CTL	D_ID	
CS_CTL/Pri	S_ID	
TYPE	F_CTL	
SEQ_ID	DF_CTL	SEQ_CNT
OX_ID		RX_ID
Parameter		
Payload		

Type of Payload Data

Type of Payload Data	
00	BLS
01	ELS
08	FCP
18	FC-SB
1B	CH-CU FC-SB
1C	CU-CH FC-SB
28	NVMe/FC

FC4 Device Data

R-CTL		Description
ROUTING	INFORMATION	
0h	0h	Uncategorized information
	1h	Solicited Data
	2h	Unsolicited Control
	3h	Solicited Control
	4h	Unsolicited Data
	5h	Data Descriptor
	6h	Unsolicited Command
	7h	Command Status
	8h	Extended Command Status
	Others	Reserved

FC4 Link Data

R-CTL		Description
ROUTING	INFORMATION	
3h	0h	Uncategorized information
	1h	Solicited Data
	2h	Unsolicited Control
	3h	Solicited Control
	4h	Unsolicited Data
	5h	Data Descriptor
	6h	Unsolicited Command
	7h	Command Status
	Others	Reserved

Link Control

ROUTING	INFORMATION	Description	Abbr.
Ch	0h	Acknowledge_1	ACK_1
	1h	Acknowledge_0	ACK_0
	2h	Nx_Port Reject	P_RJT
	3h	Fabric Reject	F_RJT
	4h	Nx_Port Busy	P_BSY
	5h	Fabric Busy to Data frame	F_BSY
	6h	Fabric Busy to Link_Control frame	F_BSY
	7h	Link Credit Reset	LCR
	8h	Notify - obsolete	NTY
	9h	End - Obsolete	END
	others	reserved	

BLS Basic Link Service

R_CTL		Description	
ROUTING	INFORMATION		
8h	0h	No Operation	NOP
	1h	Abort Sequence	ABTS
	2h	Obsolete	
	4h	Basic_Accept	BA_CC
	5h	Basic_Reject	BA_RJT
	6h	Obsolete	
	Others	Reserved	

ELS Extended Link Service

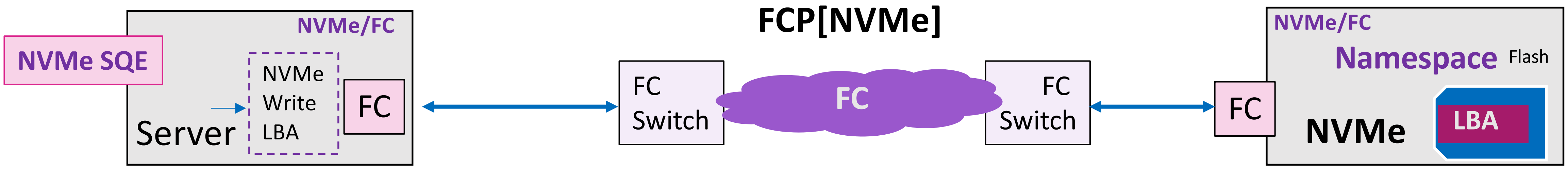
R_CTL		Description
ROUTING	INFORMATION	
0010b	0001b	Solicited Data[a]
	0010b	Request
	0011b	Reply
	Others	Reserved

[a] This value is only used by the Clock Synchronization Update (CSU) ELS.

FLOGI, PLOGI, PRLI, PRLO, FPIN...

Kamal Bakshi, Cisco

NVMe/FC Packets



SOF	FC Frame header	NVMe-FC header	NVMe Payload	CRC	EOF

SQE

FC Header

R_CTL(06)	D_ID	
CS_CTL/Pri	S_ID	
TYPE (8)	F_CTL	
SEQ_ID	DF_CTL	SEQ_CNT
OX_ID		RX_ID
Parameter		

NVMe/FC Header

Format ID (FD)	FC ID (28)	CMND IU Length
Reserved	Category	Flags
Connection ID		
Command Sequence Number		
Data Length		
NVMe SQE (64 bytes)		
DPS	LBADS	MS
Reserved		

NVMe Command

Bytes	DWORD	Byte 3 (31–24)	Byte 2 (23–16)	Byte 1 (15–8)	Byte 0 (7–0)
3-0	DW 0	Command Identifier (CID)		PSDT (15–14) / *reserved* (13–10) / FUSE (9–8)	Opcode (OPC)
7-4	DW 1	Namespace Identifier (NSID)			
11-8	DW 2	Command Specific DWORD 2			
15-12	DW 3	Command Specific DWORD 3			
19-16	DW 4	Meta Data Pointer (MPTR)			
23-20	DW 5				
27-24	DW 6	Data Pointer (DPTR) *based on PSDT* *PSDT-00* [PRP] *PSDT-01/10* [SGL]			
31-28	DW 7				
35-32	DW 8				
39-36	DW 9				
43-40	DW 10	Command Specific DWORD 10			
47-44	DW 11	Command Specific DWORD 11			
51-48	DW 12	Command Specific DWORD 12			
55-52	DW 13	Command Specific DWORD 13			
59-56	DW 14	Command Specific DWORD 14			
63-60	DW 15	Command Specific DWORD 15			

Kamal Bakshi, Cisco

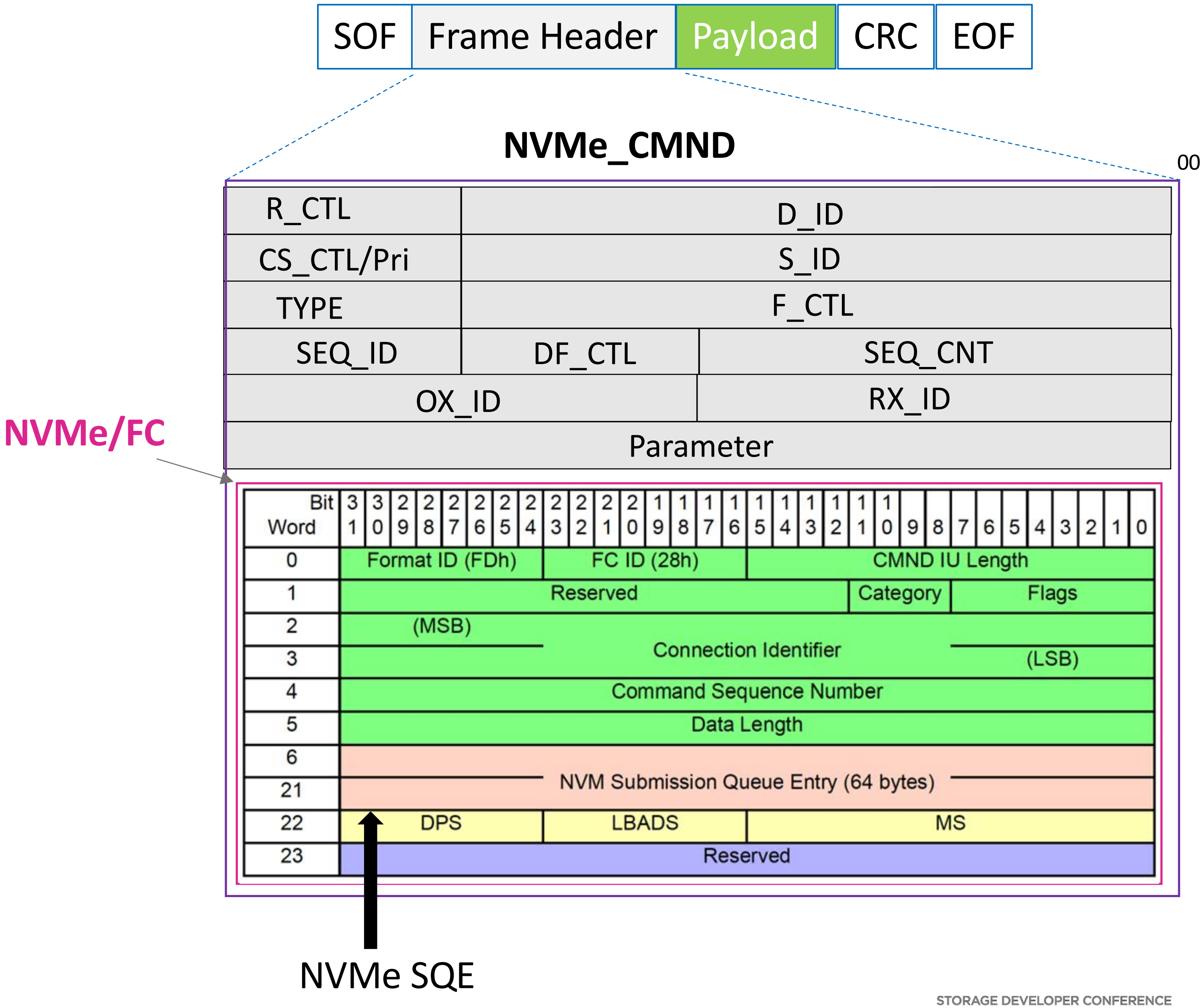
NVMe/FC Packets

Initiator

Description	Data block	
	R_CTL field	**Content**
Command request	06h	NVMe_CMND
Command request	06h	NVMe_CMND
Data-Out action	01h	NVMe_DATA
Confirm	03h	NVMe_CONF
Sequence Retransmission request	09h	NVMe_SR

Target

Description	Data block	
	R_CTL field	**Content**
Data-Out delivery request	05h	NVMe_XFER_RDY (Write)
Data-In action	01h	NVMe_DATA
Command response	07h	NVMe_RSP
Command response (NVMe_CONF IU request)	07h	NVMe_RSP
Extended response	08h	NVMe_ERSP
Extended response (NVMe_CONF IU request)	08h	NVMe_ERSP
Sequence Retransmission response	0Ah	NVMe_SR_RSP



Kamal Bakshi, Cisco

NVMe/FC Packets

SOF	Frame Header	Payload	CRC	EOF

NVMe_SR Request (Sequence Retry)

Bit / Word	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	FC ID (28h)								Opcode (01h)								Reserved								Retry R_CTL							
1	Reserved																															

NVMe_SR Response

Bit / Word	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	FC ID (28h)								Opcode (01h)								Reserved								Status							
1	Reserved																															

NVMe_XFER_RDY (Transfer Ready)

Bit / Word	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	Relative Offset																															
1	Burst Length																															
2	Reserved																															

NVMe_ERSP (Extended Response)

00

R_CTL	D_ID	
CS_CTL/Pri	S_ID	
TYPE	F_CTL	
SEQ_ID	DF_CTL	SEQ_CNT
OX_ID		RX_ID
Parameter		

Bit / Word	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	ERSP Result								Reserved								ERSP IU Length															
1	Response Sequence Number																															
2	Transferred Data Length																															
3	Reserved																															
4	NVM Completion Queue Entry (16 bytes)																															
5																																
6																																
7																																

NVMe CQE
Response

NVMe/FC

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

NVMe-FC Data Transfer

NVMe-Data

FCP Dataset

Bit set to 1 indicating Offset present

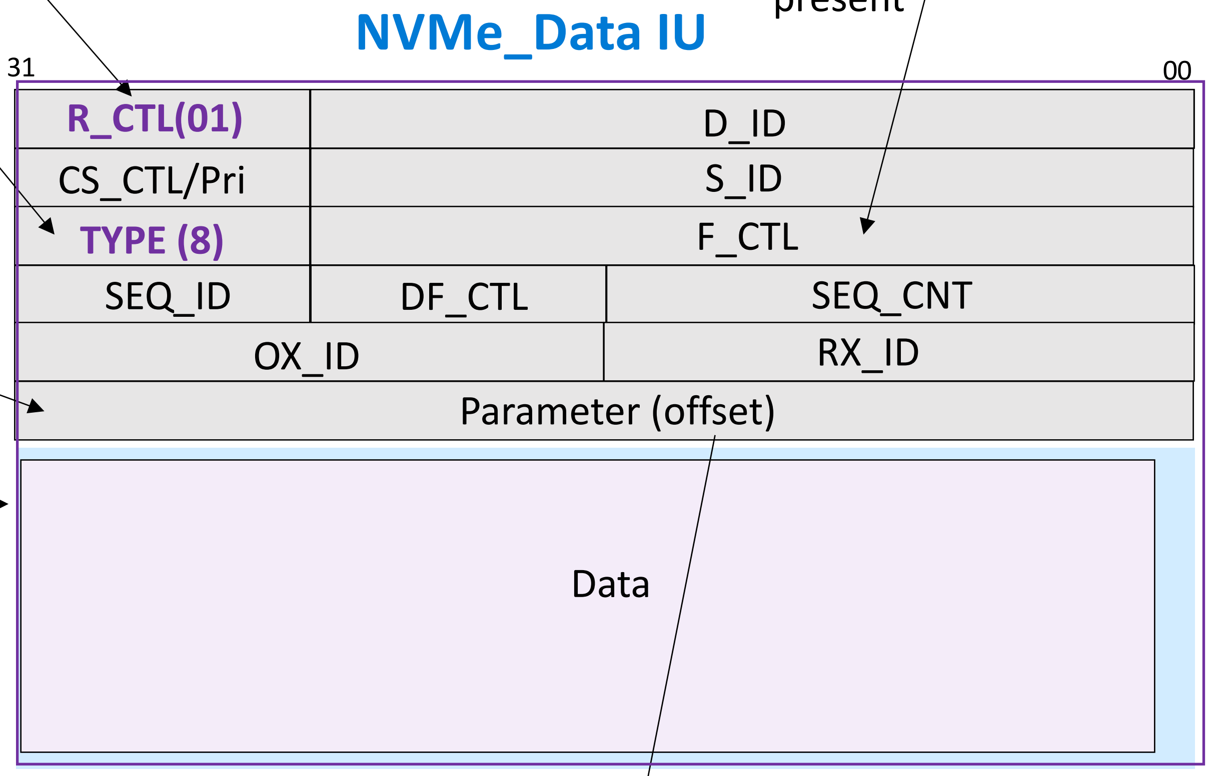
The start of the range is indicated by the Parameter field in the first frame of the Sequence. Relative offset value multiple of x4

Data Series

-Each frame in the Sequence is a continually increasing portion of the Data Series range.

-The length of the range is the Sequence payload length.

-If more than one NVMe_DATA IU is used to transfer the data, the relative offset value in the Parameter field is used to ensure that the NVM data is reassembled in the proper order.

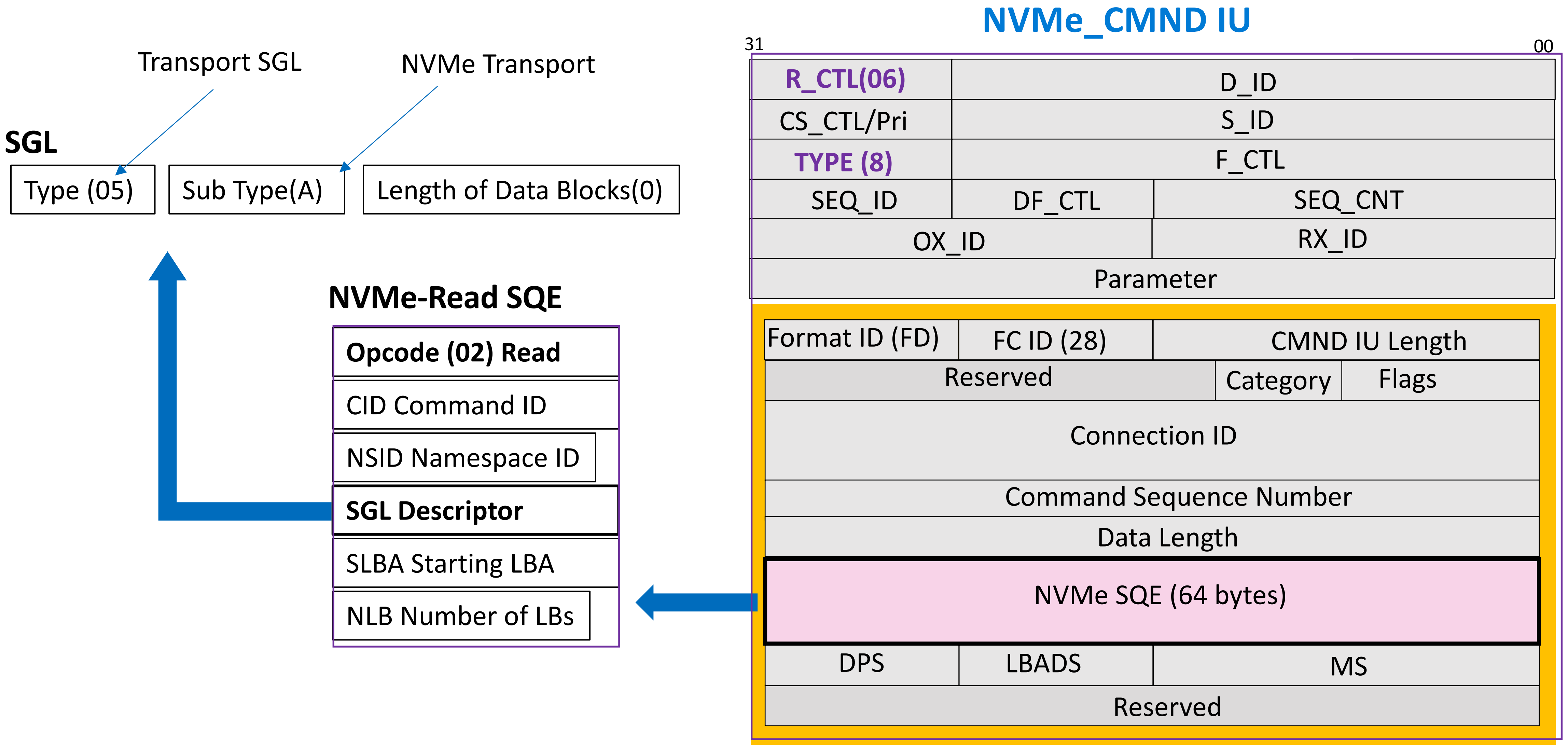


NVMe Data is transferred as FCP Data

Port(1,1,1)	FCP	FC4SData		FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0800;	2084	023E
Port(1,1,1)	FCP	FC4SData		FC4SData; SCSI FCP; Offset = 0x00000800; Len = 0x0800;	2084	023E
Port(1,1,1)	FCP	FC4ExtStatus		Success;	68	023E
Port(1,1,2)	FCP		FC4Cmd	Read; NSID = 0x00000004; LBA = 0x00000000; NbBlocks =	132	023F
Port(1,1,1)	FCP	FC4SData		FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0800;	2084	023F
Port(1,1,1)	FCP	FC4SData		FC4SData; SCSI FCP; Offset = 0x00000800; Len = 0x0800;	2084	023F
Port(1,1,1)	FCP	FC4Status		Good Status;	48	023F
Port(1,1,2)	FCP		FC4Cmd	Read; NSID = 0x00000004; LBA = 0x00000008; NbBlocks =	132	0240
Port(1,1,1)	FCP	FC4SData		FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0800;	2084	0240
Port(1,1,1)	FCP	FC4SData		FC4SData; SCSI FCP; Offset = 0x00000800; Len = 0x0800;	2084	0240
Port(1,1,1)	FCP	FC4Status		Good Status;	48	0240

Kamal Bakshi, Cisco

NVMe-FC Read Command IU



Kamal Bakshi, Cisco

NVMe-FC (PRLI -Process Log In)

```
SOF = SOFi3;
RCtl = ExtLinkReq; D_Id = 0xAE00C1;
CS_CTL = 0x00 [DSCP = 0x00]; S_Id = 0xAE00E0;
Type = EX_LNK_SRV; F_Ctl [Exchange Context = Originator; First_Sequence; End_Sequence; Sequence Initiative = Transfer Sequence Initiative];
SEQ_Id = 0x00; DF_Ctl = 0x00; SEQ_Cnt = 0x0000;
OX_Id = 0x0203; RX_Id = 0xFFFF;
PARA = 0x00000000;
Command Code = PRLI  (Interesting Event Found); Page Length = 16  Bytes; Payload Length = 20  Bytes;
Type Code = SCSI FCP; Flags [Established Image Pair];
Originator Process_Associator = 0x00000000;
Responder Process_Associator = 0x00000000;
Service Parameters [Rec_Support; Task Retry Identification Requested; Retry; Confirm Completion Allowed; Initiator Function; RXferRdyDisabled];
CRC = 0x028125FE    (Correct);
EOF = EOFt;
```

PRLI

Service Parameter
(Initiator Function = NVMe-FC/reply)

Kamal Bakshi, Cisco

NVMe-FC (Create Association)

```
SOF = SOFi3;
RCtl = FC4LinkUCtl; D_Id = 0xAE00C3;
CS_CTL = 0x00 [DSCP = 0x00]; S_Id = 0xAE00E0;
Type = FC-NVMe; F_Ctl [Exchange Context = Originator; First_Sequence; End_Sequence; Sequence Initiative = Transfer Sequence Initiative];
SEQ_Id = 0x00; DF_Ctl = 0x00; SEQ_Cnt = 0x0000;
OX_Id = 0x021D; RX_Id = 0xFFFF;
PARA = 0x00000000;
Command Code = NVMe Create Association;
Descriptor list length = 1016  Bytes;
Descriptor tag = NVMe Create Association;
Descriptor length = 1008  Bytes;
NVMe_ERSP Ratio = 0x0008;

Controller ID = 0xFFFF; Admin Submission Queue Size = 0x001F;

Host Identifier = ED9D0705 6B4F425D A99B99E8 FF67FC80;

Host NVMe Qualified Name = nqn.2014-08.org.nvmexpress:uuid:290ecc27-d30e-4f08-9a73-474e3802c9d8;
```

Command
(NVMe Create Association)

Controller ID
(Dynamic)

Admin Queue
depth

Host NQN

Kamal Bakshi, Cisco

NVMe-FC (Accept Create Association)

```
SOF = SOFi3;
RCtl = FC4LinkSCtl; D_Id = 0xAE00E0;
CS_CTL = 0x00 [DSCP = 0x00]; S_Id = 0xAE00C3;
Type = FC-NVMe; F_Ctl [Exchange Context = Responder; Last_Sequence; End_Sequence];
SEQ_Id = 0x00; DF_Ctl = 0x00; SEQ_Cnt = 0x0000;
OX_Id = 0x021D; RX_Id = 0x00D1;
PARA = 0x00000000;
Command Code = Accept;
Descriptor list length = 48  Bytes;
Descriptor tag = NVMe Link Service Request Information;
Descriptor length = 8  Bytes;
Accepted Command Code = NVMe Create Association;

Descriptor tag = NVMe Association Identifier;
Descriptor length = 8  Bytes;
NVMe Association Identifier = 0x5FBF79822FA30000;

Descriptor tag = NVMe Connection Identifier;
Descriptor length = 8  Bytes;
NVMe Connection Identifier = 0x5FBF79822FA30000;

CRC = 0x268BD28B    (Correct);
EOF = EOFt;
```

Accept

NVMe Association ID

NVMe Connection ID

Kamal Bakshi, Cisco

NVMe-FC (Connect)

SOF = SOFi3;
RCtl = FC4Cmd; D_Id = 0xAE00C3;
CS_CTL = 0x00 [DSCP = 0x00]; S_Id = 0xAE00E0;
Type = SCSI FCP; F_Ctl [Exchange Context = Originator; First_Sequence; End_Sequence; Sequence Initiative = Transfer Sequence Initiative];
SEQ_Id = 0x01; DF_Ctl = 0x00; SEQ_Cnt = 0x0000;
OX_Id = 0x020E; RX_Id = 0xFFFF;
PARA = 0x00000000;
Differentiator = FC-NVMe Cmd IU; CMD IU Length = 24 Words;
Flags [Write = ->Data];
NVMe Connection Identifier = 0xE5B420ADBB500000;

Command Sequence Number = 0x00000001;
Data Length = 0x00000400;
Opcode = Fabrics Cmd; Reserved = 0x40 (Unexpected Value Found); CID = 0x0000;
Fabrics Cmd = Connect;

SGL Entry 1 [

Length = 0x00000400 Bytes;
SGL Descriptor Type = Transport SGL Data Block descripter; SGL Descriptor SubType = 0xA Reserved (Unexpected Value Found)];
Record Format = NVMe 1.2.1; Queue ID = 0x0000;
Subm Queue Size = 32; Connect Attributes [Priority Class = Urgent];
Keep Alive Timeout = 0 ms;

Fabric Command = Connect

default queue size = 32

Queue ID = 0 (Admin)

Kamal Bakshi, Cisco

NVMe-FC (Reply Identify Active Name Space List)

```
SOF = SOFi3;
RCtl = FC4SData; D_Id = 0xAE00E0;
CS_CTL = 0x00 [DSCP = 0x00]; S_Id = 0xAE00C3;
Type = SCSI FCP; F_Ctl [Exchange Context = Responder; RO];
SEQ_Id = 0x81; DF_Ctl = 0x00; SEQ_Cnt = 0x0000;
OX_Id = 0x0239; RX_Id = 0x0353;
PARA = 0x00000000; Pld bytes = 0x0800;
Pld = 04000000 05000000 00000000 00000000 00000000 00000000 00000000 00000000...;
```

NSID =04

NSID =05

Kamal Bakshi, Cisco

NVMe-FC (Read command)

SOF	FC headers	NVMe-FC	NVMe-CMD	Payload	CRC	EOF

Index	Hex	Interpretation
SOF 000000	FB B5 56 56	SOF = SOFi3;
FCH 000000	06 AE 00 C3	RCtl = FC4Cmd; D_Id = 0xAE00C3;
FCH 000001	00 AE 00 E0	CS_CTL = 0x00 [DSCP = 0x00]; S_Id = 0xAE00E0;
FCH 000002	08 29 00 00	Type = SCSI FCP; F_Ctl [Exchange Context = Originator; First_Sequence; End_Sequence; Sequence Initiative = Transfer Sequence Initiative];
FCH 000003	01 00 00 00	SEQ_Id = 0x01; DF_Ctl = 0x00; SEQ_Cnt = 0x0000;
FCH 000004	02 3F FF FF	OX_Id = 0x023F; RX_Id = 0xFFFF;
FCH 000005	00 00 00 00	PARA = 0x00000000;
FCP 000000	FD 28 00 18	Differentiator = FC-NVMe Cmd IU; CMD IU Length = 24 Words;
FCP 000001	00 00 00 02	Flags [Read = <-Data];
FCP 000002	5F BF 79 82	NVMe Connection Identifier = 0x5FBF79822FA30002;
FCP 000003	2F A3 00 02	
FCP 000004	00 00 00 02	Command Sequence Number = 0x00000002;
FCP 000005	00 00 10 00	Data Length = 0x00001000;
NVMe 00000	02 40 51 00	Opcode = Read; PRP or SGL = SGL; CID = 0x0051;
NVMe 00001	04 00 00 00	NSID = 0x00000004;
NVMe 00002	00 00 00 00	
NVMe 00003	00 00 00 00	
NVMe 00004	00 00 00 00	Metadata SGL Segment Pointer = 0x00000000;
NVMe 00005	00 00 00 00	
NVMe 00006	00 00 00 00	SGL Entry 1 [
NVMe 00007	00 00 00 00	
NVMe 00008	00 10 00 00	Length = 0x00001000 Bytes;
NVMe 00009	00 00 00 5A	SGL Descriptor Type = Transport SGL Data Block descriptor; SGL Descriptor SubType = 0xA Reserved (Unexpected Value Found)];
NVMe 00010	00 00 00 00	Starting LBA = 0x00000000;
NVMe 00011	00 00 00 00	
NVMe 00012	07 00 00 00	Number of Logical Blocks = 0x08; PRInfoAction = Pass;
NVMe 00013	00 00 00 00	Dataset Management [Access Latency = None; Access Frequency = Unknown];
NVMe 00014	00 00 00 00	Expected Initial Block Ref Tag = 0x00000000;
NVMe 00015	00 00 00 00	Expected Block App Tag = 0x0000; Expected Block App Tag Mask = 0x0000;
FCP 000000	00 00 00 00	
FCP 000001	00 00 00 00	
End 000000	6E 0E 7A 10	CRC = 0x6E0E7A10 (Correct);
End 000001	95 75 75 FD	EOF = EOFt;

NVMe-CMD “Read”

Connection-ID

NSID

SLB

NLB

Kamal Bakshi, Cisco

NVMe-FC (Read NSID)

```
SOF = SOFi3;
RCtl = FC4Cmd; D_Id = 0xAE00C3;
CS_CTL = 0x00 [DSCP = 0x00]; S_Id = 0xAE00E0;
Type = SCSI FCP; F_Ctl [Exchange Context = Originator; First_Sequence; End_Sequence; Sequence Initiative = Transfer Sequence Initiative];
SEQ_Id = 0x01; DF_Ctl = 0x00; SEQ_Cnt = 0x0000;
OX_Id = 0x0240; RX_Id = 0xFFFF;
PARA = 0x00000000;
Differentiator = FC-NVMe Cmd IU; CMD IU Length = 24  Words;
Flags [Read = <-Data];
NVMe Connection Identifier = 0x5FBF79822FA30002;

Command Sequence Number = 0x00000003;
Data Length = 0x00001000;
Opcode = Read; PRP or SGL = SGL; CID = 0x0052;
NSID = 0x00000004;

Metadata SGL Segment Pointer = 0x00000000;

SGL Entry 1 [

Length = 0x00001000  Bytes;
SGL Descriptor Type = Transport SGL Data Block descriptor; SGL Descriptor SubType = 0xA Reserved  (Unexpected Value Found)];
Starting LBA = 0x00000008;

Number of Logical Blocks = 0x08; PRInfoAction = Pass;
Dataset Management [Access Latency = None; Access Frequency = Unknown];
Expected Initial Block Ref Tag = 0x00000000;
Expected Block App Tag = 0x0000; Expected Block App Tag Mask = 0x0000;

CRC = 0x263126B3    (Correct);
EOF = EOFt;
```

Read = SGL

NSID

Starting LBA

Number of Logical Blocks

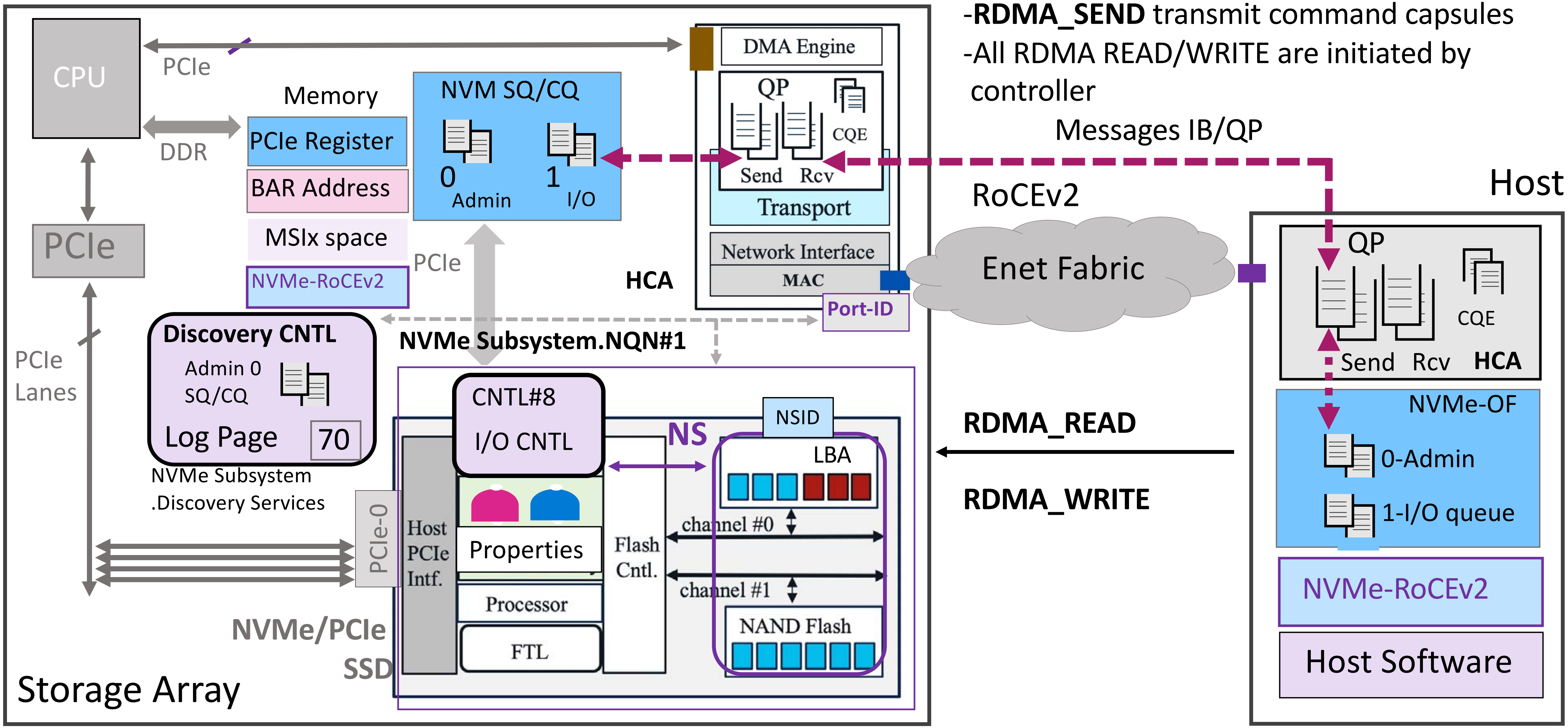
Kamal Bakshi, Cisco

NVMe/RoCEv2 Flows

Kamal Bakshi, Cisco

RoCEv2_Queues mapping to NVMe_Queues

- **RDMA_SEND** transmit command capsules
- All RDMA READ/WRITE are initiated by controller



Kamal Bakshi, Cisco

NVMe-RoCEv2 Transport

I — Connect Host to the NVMe Controller "Transport Binding"
- -Discover the NVMe I/O controller
- -Initiate the NVMe connection over RoCEv2 IB/QP

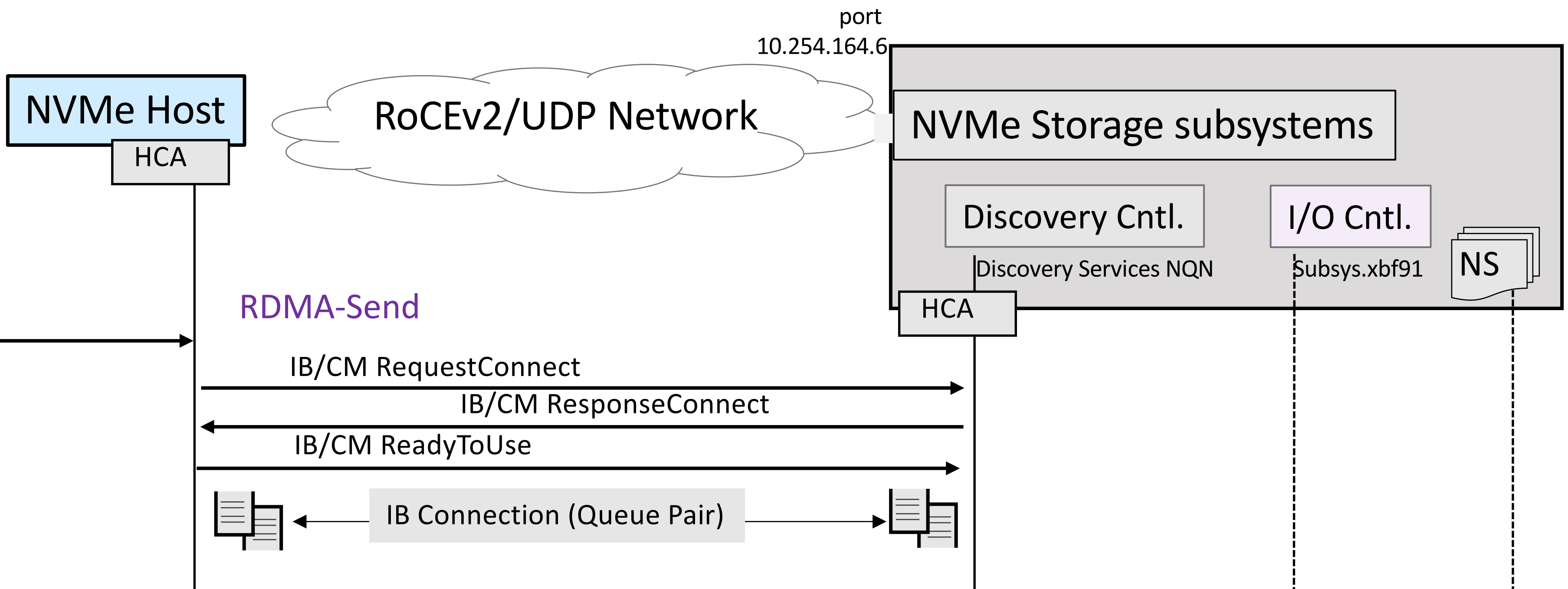
II — Create NVMe Admin & I/O Queues for Controller
- -Create Admin Queue for NVMe commands
- -Create I/O Queues for Data Transfer
  (Each NVMe queue pair is mapped to IB/QP queue pair)

III — Find the list of Active Namespaces
- -Get a list of Active available Namespaces
  (Controlled by the Storage Admin)

IV — Start NVMe I/O operation
- -Start I/O transfer with NVMe Write/Read
  (I/O transfer uses I/O queues)

Kamal Bakshi, Cisco

NVMe-RoCEv2 (Transport Binding)



 # nvme discover -t rdma
-a 10.254.164.6

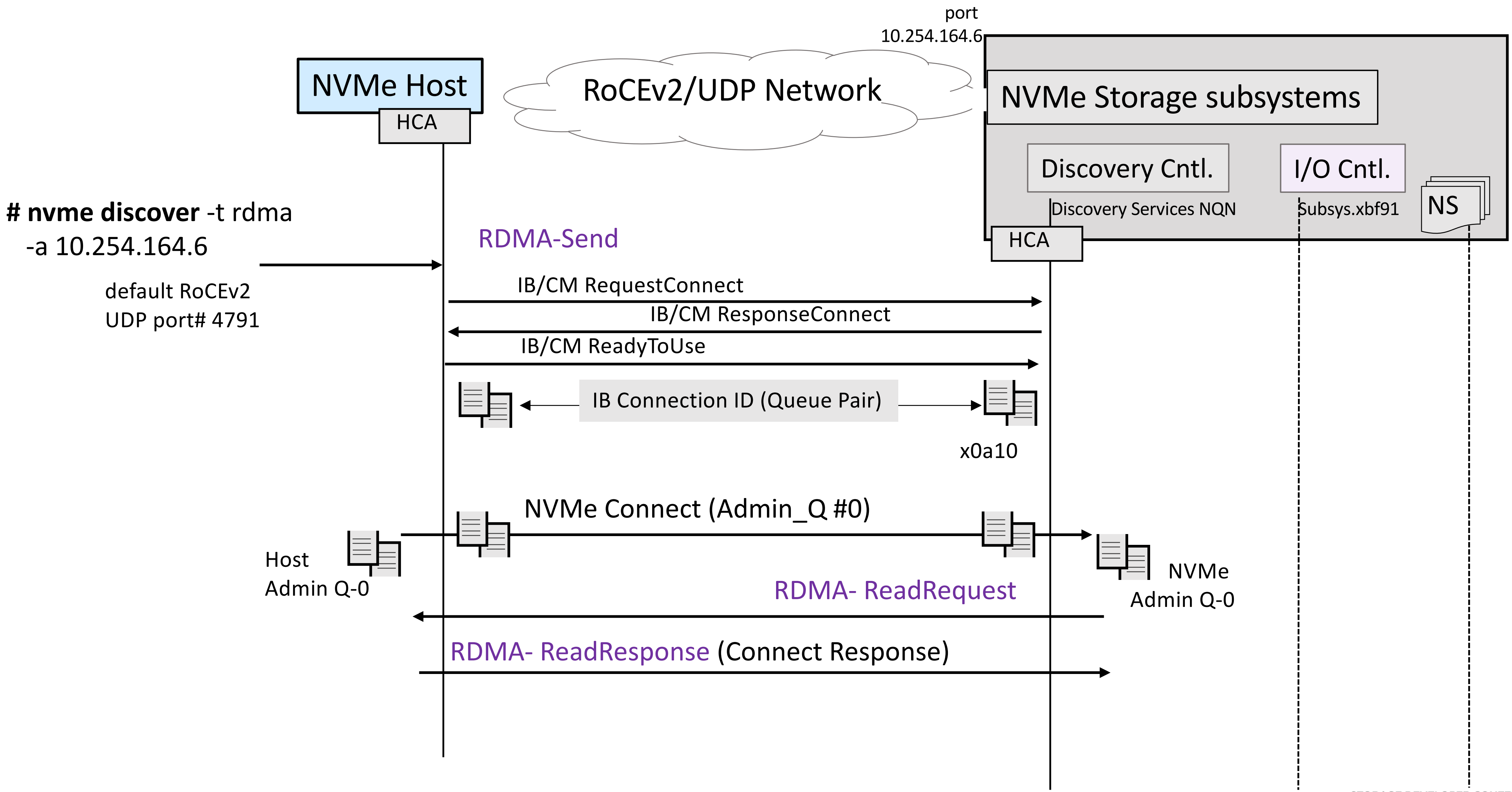
default RoCEv2
UDP port# 4791

Setup
IB/Queue Pair
connection

```
> Internet Protocol Version 4, Src: 10.254.85.2, Dst: 10.254.164.6
> User Datagram Protocol, Src Port: 49153, Dst Port: 4791
v InfiniBand
    > Base Transport Header
    v DETH - Datagram Extended Transport Header
        Queue Key: 0x0000000080010000
        Reserved: 00
        Source Queue Pair: 0x00000001
    > MAD Header - Common Management Datagram
    > CM ConnectRequest
      Invariant CRC: 0x8962f2c1
> NVM Express Fabrics RDMA
```

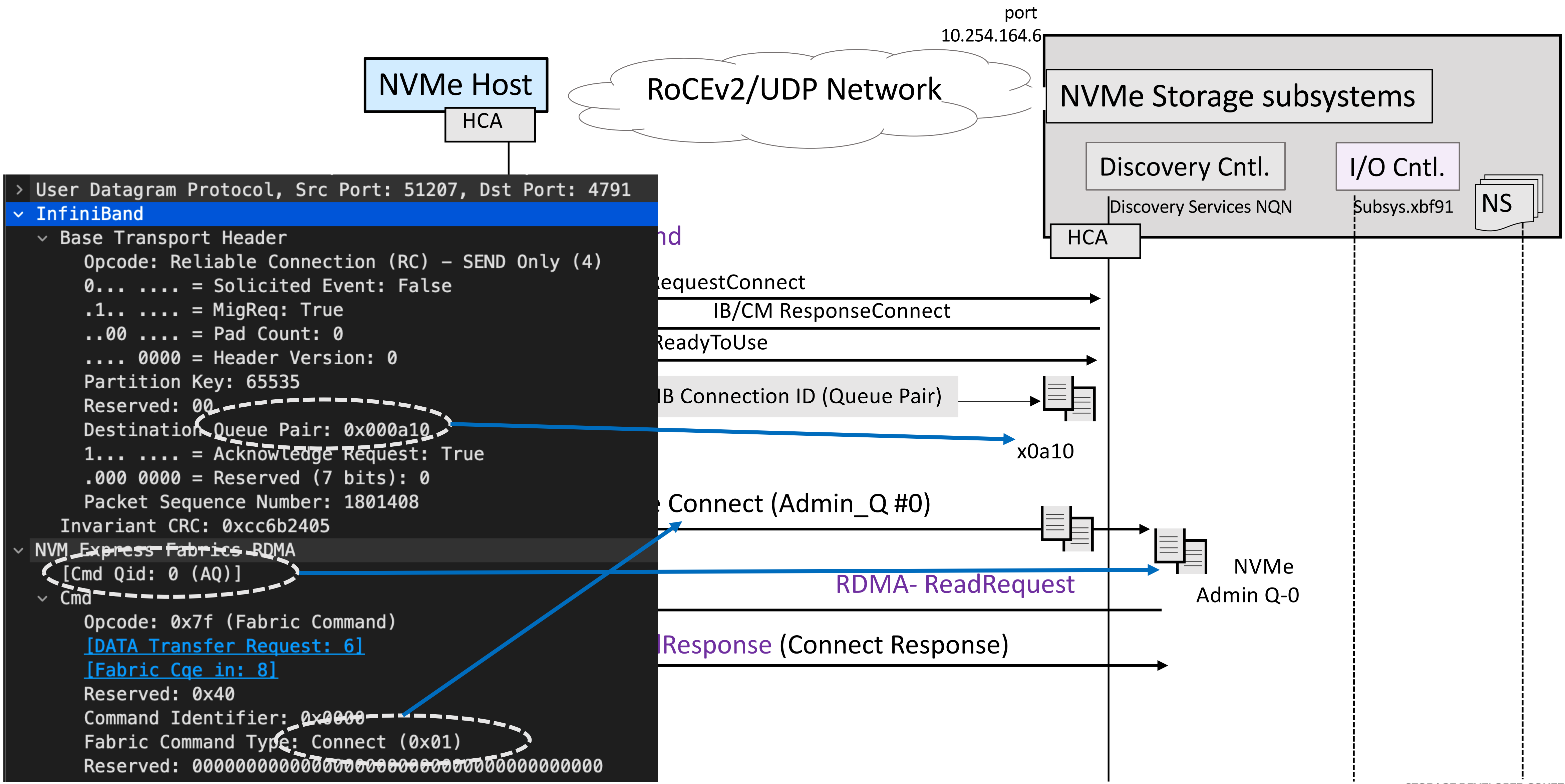

Kamal Bakshi, Cisco

NVMe-RoCEv2 (Transport Binding)



Kamal Bakshi, Cisco

NVMe-RoCEv2 (Transport Binding)

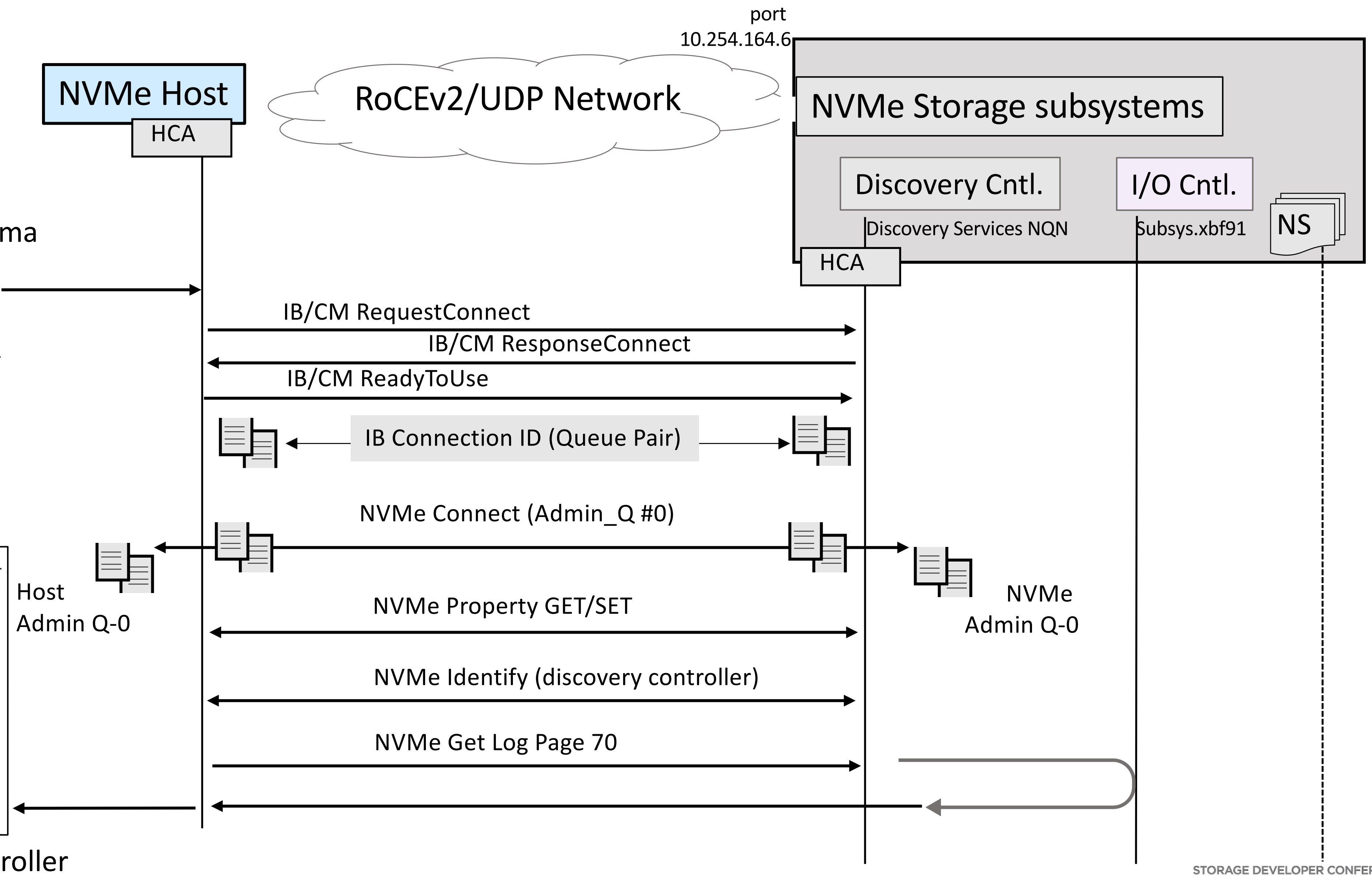


Kamal Bakshi, Cisco

NVMe-RoCEv2 (Transport Binding)

 # nvme discover -t rdma
-a 10.254.164.6

default RoCEv2
UDP port# 4791



Direct Discovery Controller Log

```
Discovery Log Number of Records 1
=====Discovery Log Entry 0======
trtype:   rdma
adrfam:   ipv4
subtype:  nvme subsystem
treq:     not specified
portid:   20
trsvcid:  4420
subnqn:   GB00041004bbf91
traddr:   10.254.164.6
rdma_prtype: unrecogni ed
rdma_qptype: unrecogni ed
rdma_cms:    unrecogni ed
rdma_pkey: 0x0000
```

I/O controller

Kamal Bakshi, Cisco

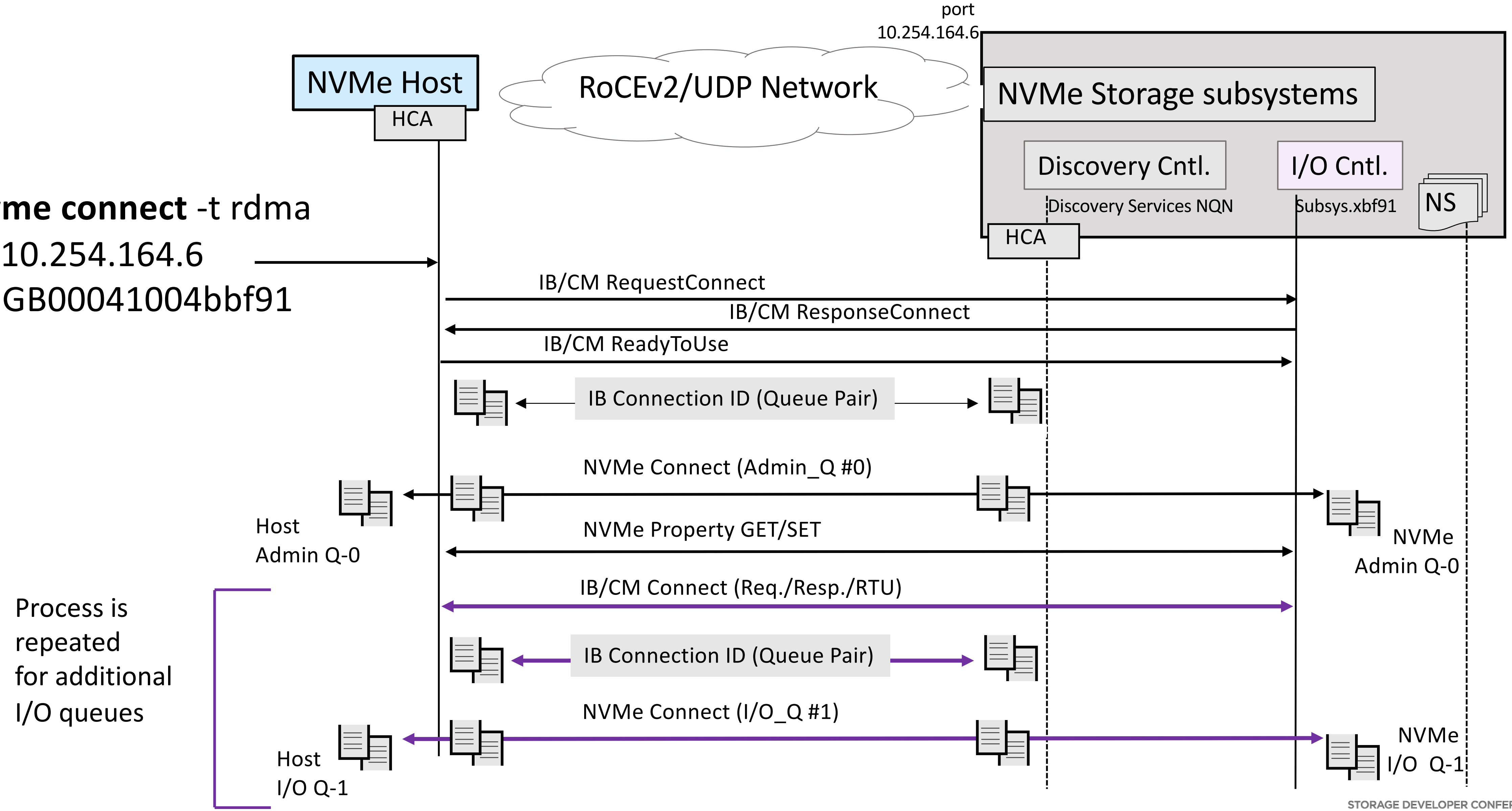
NVMe-RoCEv2 Transport

I	Connect Host to the NVMe Controller "Transport Binding"	-Discover the NVMe I/O controller -Initiate the NVMe connection over RoCEv2 IB/QP
II	Create NVMe Admin & I/O Queues for Controller	-Create Admin Queue for NVMe commands -Create I/O Queues for Data Transfer (Each NVMe queue pair is mapped to IB/QP queue pair)
III	Find the list of Active Namespaces	-Get a list of Active available Namespaces (Controlled by the Storage Admin)
IV	Start NVMe I/O operation	-Start I/O transfer with NVMe Write/Read (I/O transfer uses I/O queues)

Kamal Bakshi, Cisco

NVMe-RoCEv2 (Transport Binding)

nvme connect -t rdma
-a 10.254.164.6
-n GB00041004bbf91



Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

NVMe-RoCEv2 Transport

I — **Connect Host to the NVMe Controller "Transport Binding"**

- -Discover the NVMe I/O controller
- -Initiate the NVMe connection over RoCEv2 IB/QP

↓

II — **Create NVMe Admin & I/O Queues for Controller**

- -Create Admin Queue for NVMe commands
- -Create I/O Queues for Data Transfer
  (Each NVMe queue pair is mapped to IB/QP queue pair)

↓

III — **Find the list of Active Namespaces**

- -Get a list of Active available Namespaces
  (Controlled by the Storage Admin)

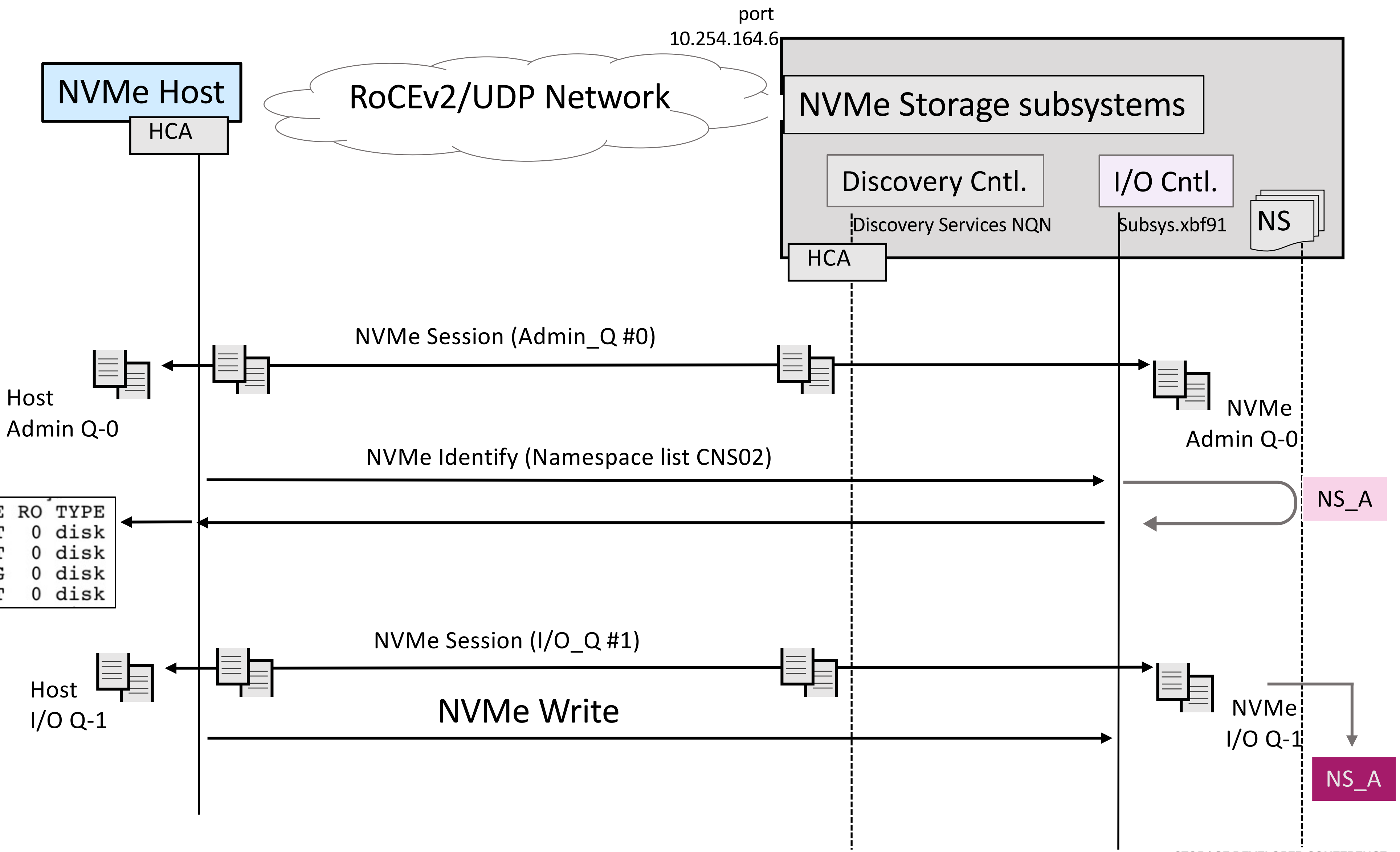
↓

IV — **Start NVMe I/O operation**

- -Start I/O transfer with NVMe Write/Read
  (I/O transfer uses I/O queues)

Kamal Bakshi, Cisco

NVMe-RoCEv2 (Transport Binding)



\# lsblk

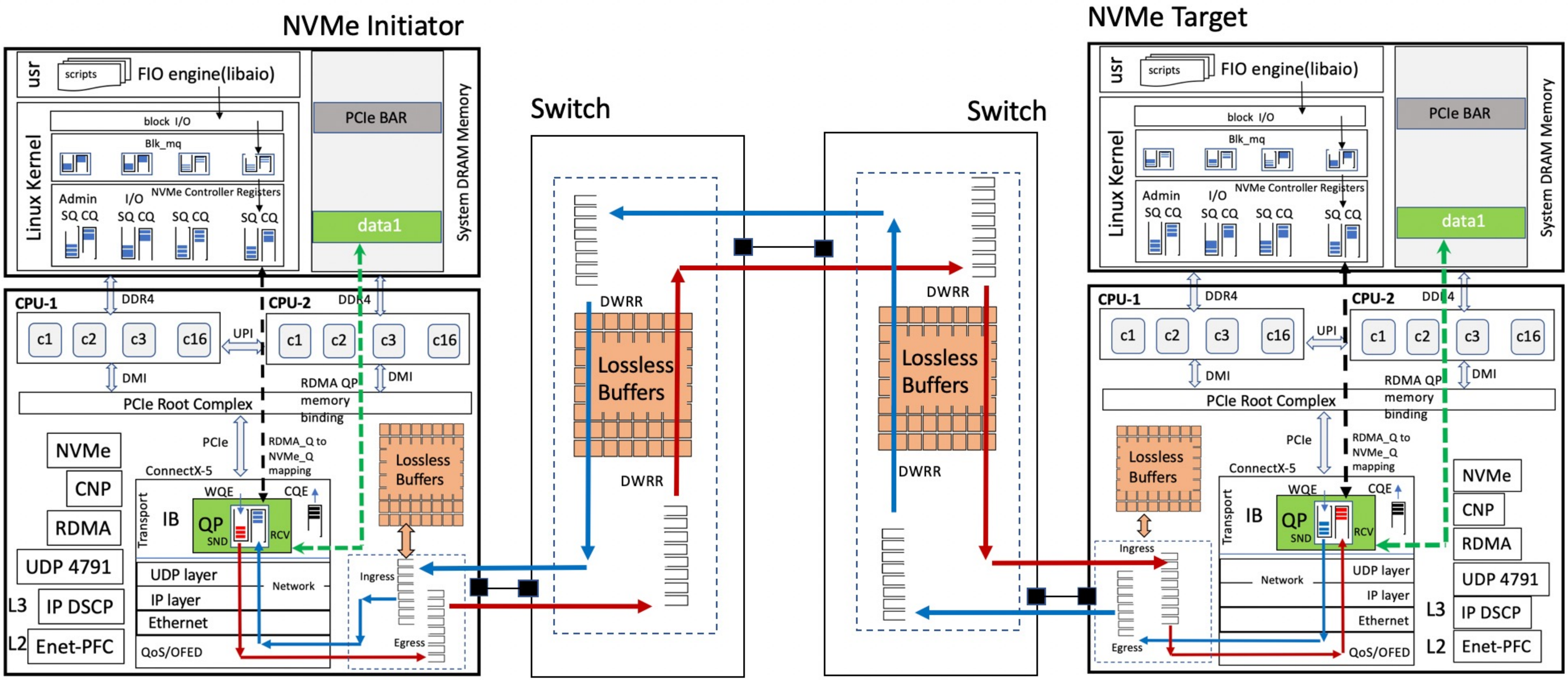
NAME	MAJ:MIN	RM	SIZE	RO	TYPE
nvme0n1	259:0	0	1.5T	0	disk
sdb	8:16	0	1.7T	0	disk
sdc	8:32	1	14.8G	0	disk
nvme1n1	259:1	0	2T	0	disk

nvme storage

Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE

NVMe-RoCEv2 Traffic Engineering



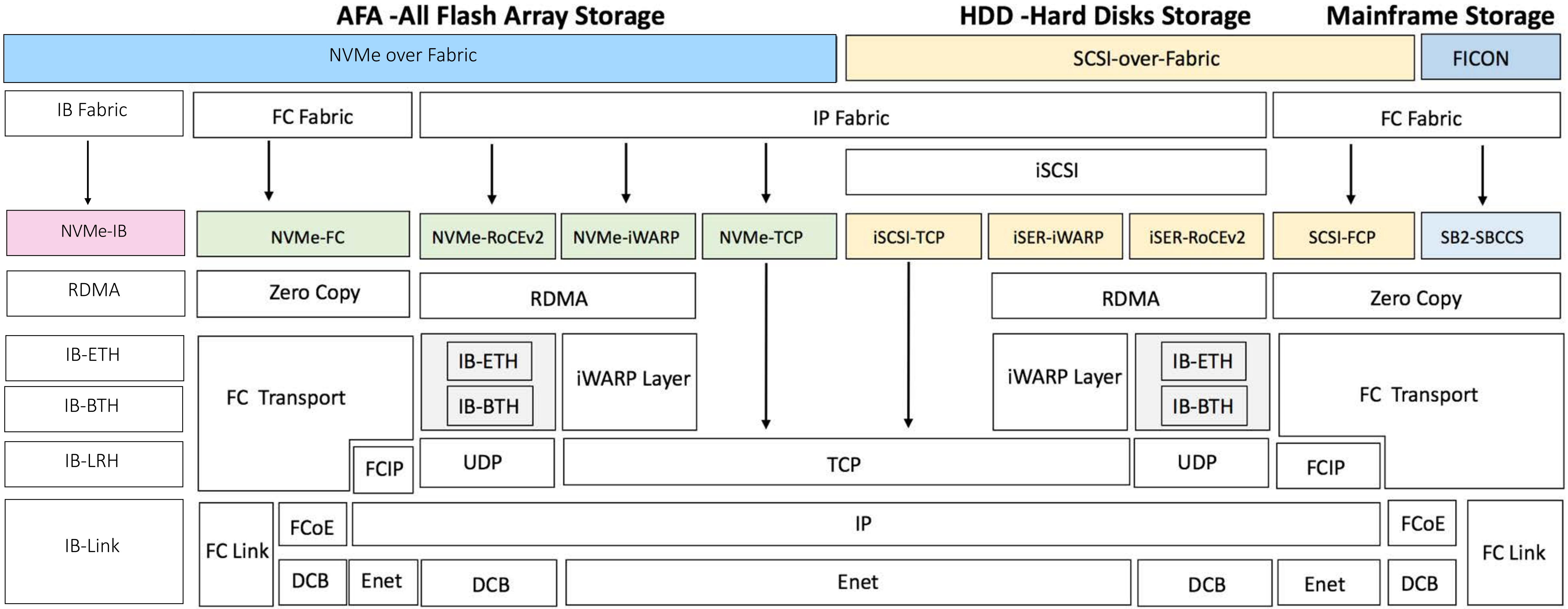
IB/CNP, DSCP, PFC, ECN

Kamal Bakshi, Cisco

NVMe Advanced Features

Kamal Bakshi, Cisco

Storage Protocols Stack



Kamal Bakshi, Cisco

Traditional SAN
1994-SCSI

1994 FC-SCSI

1986 SCSI
1994 FC Standard

[Enet[IP[TCP [iSCSI[**SCSI**]]]]] 2000 iSCSI

2001 FICON

2002 SRP

2007 iSER

[Enet[FCoE[FC[**SCSI**]]]] 2009 FCoE

NVMe-SSD 2013

Next-Gen SAN
2020-NVMe Fabric

NVMe/oF -2016

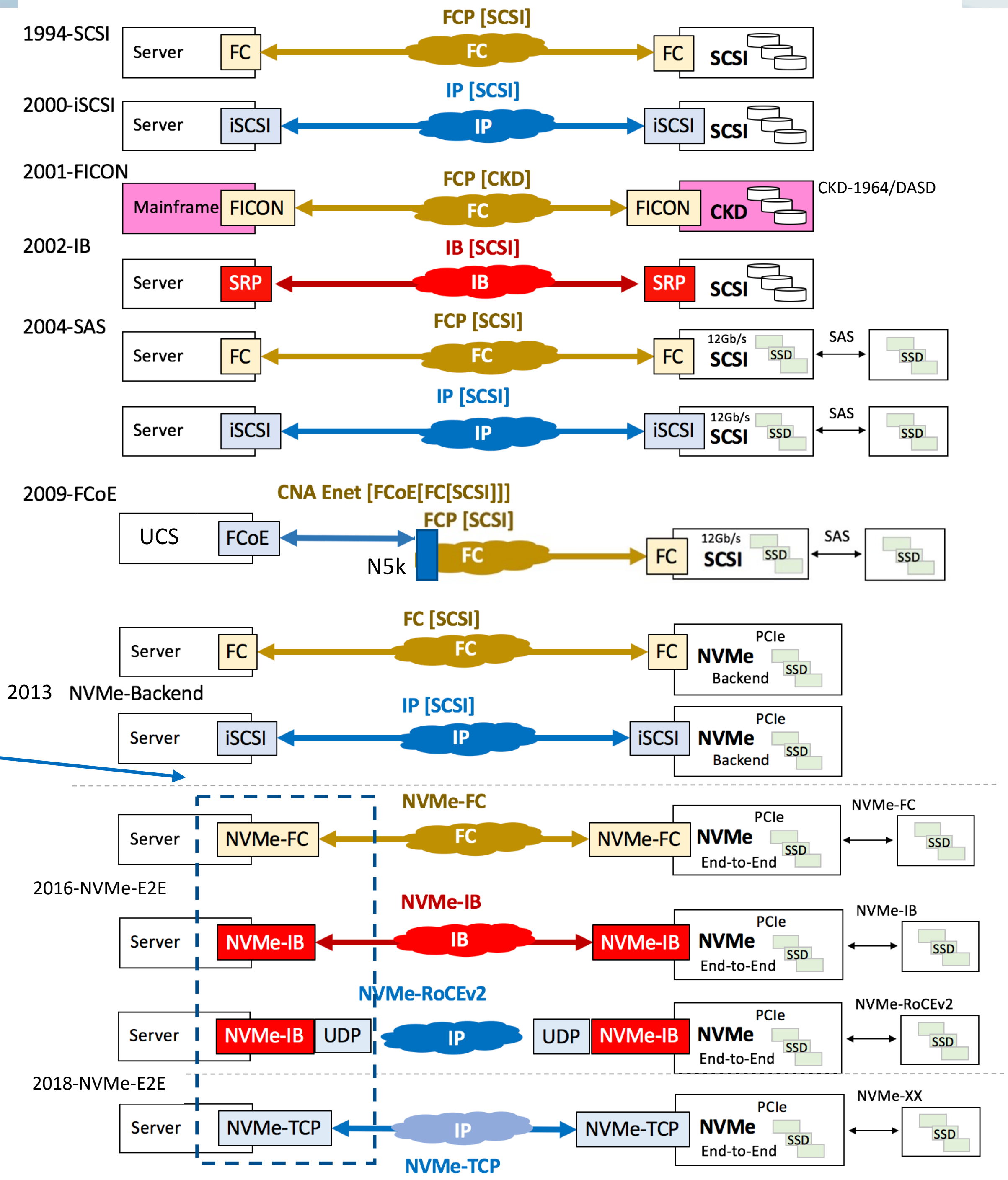
2016 NVMe-FC

2016 NVMe-IB

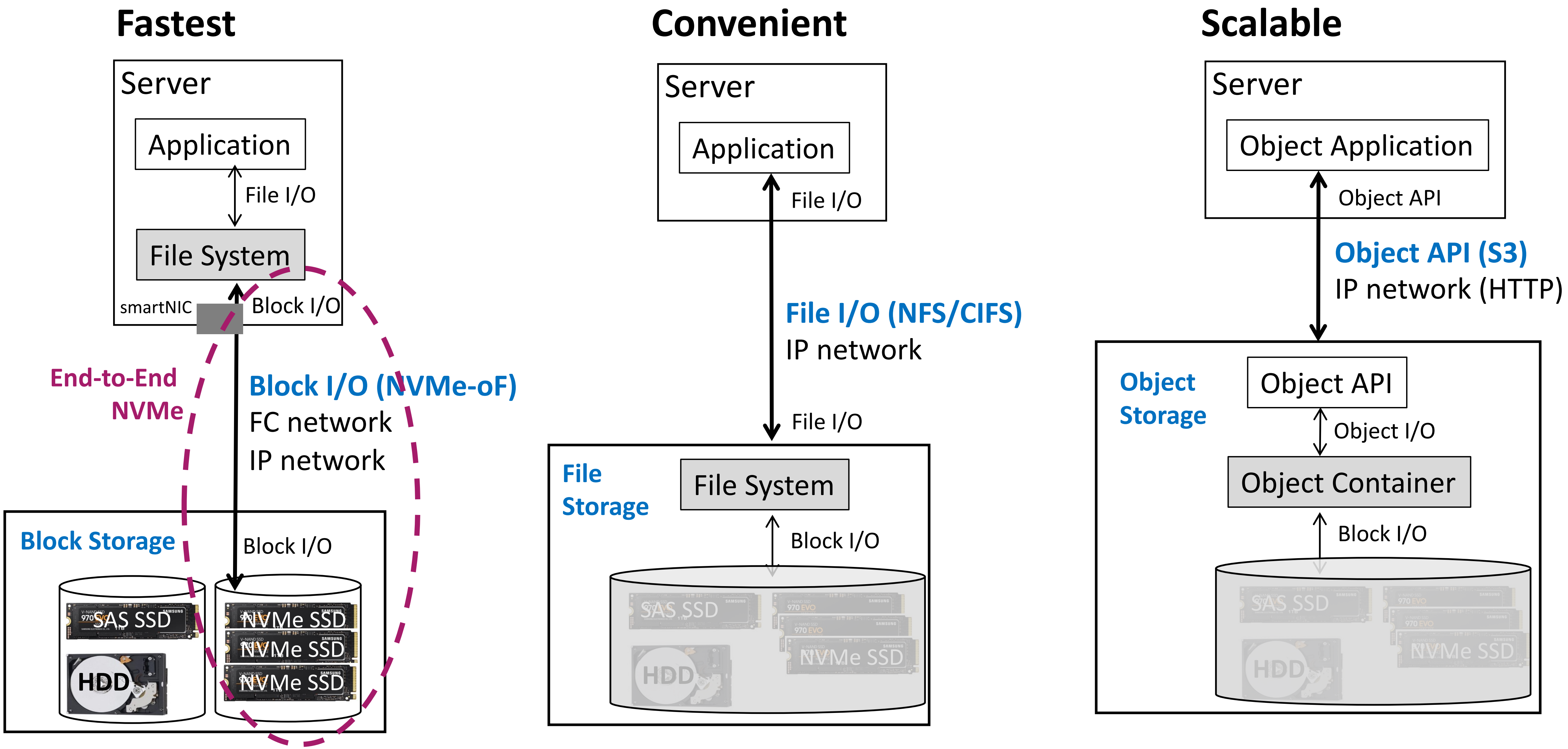
2016 NVMe-RoCEv2

2014-RoCEv2

[Enet[IP[TCP [NVMe]]]] 2018 NVMe-TCP

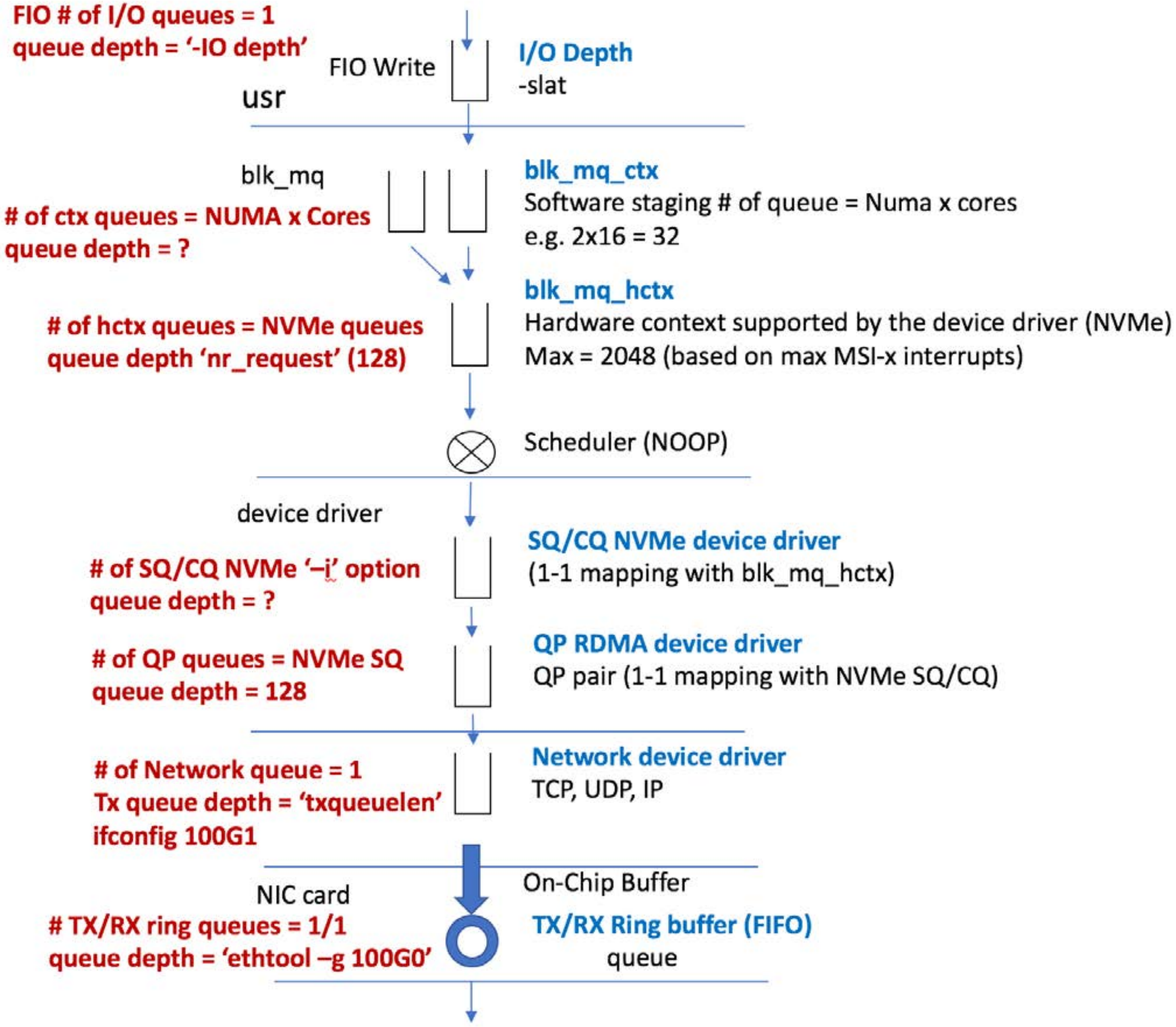


Block Storage is getting faster with NVMe



Kamal Bakshi, Cisco

Queues Mapping



Kamal Bakshi, Cisco

NVMe CLI Commands (debian)

- nvme-admin-passthru(1)
- nvme-ana-log(1)
- nvme-attach-ns(1)
- nvme-boot-part-log(1)
- nvme-capacity-mgmt(1)
- nvme-changed-ns-list-log(1)
- nvme-cmdset-ind-id-ns(1)
- nvme-compare(1)
- nvme-connect-all(1)
- nvme-connect(1)
- nvme-copy(1)
- nvme-create-ns(1)
- nvme-delete-ns(1)
- nvme-dera-stat(1)
- nvme-detach-ns(1)
- nvme-device-self-test(1)
- nvme-dir-receive(1)
- nvme-dir-send(1)
- nvme-disconnect-all(1)
- nvme-disconnect(1)
- nvme-discover(1)
- nvme-dsm(1)
- nvme-effects-log(1)
- nvme-endurance-event-agg-log(1)
- nvme-endurance-log(1)
- nvme-error-log(1)
- nvme-fid-support-effects-log(1)
- nvme-flush(1)
- nvme-format(1)
- nvme-fw-activate(1)
- nvme-fw-commit(1)
- nvme-fw-download(1)
- nvme-fw-log(1)
- nvme-gen-hostnqn(1)
- nvme-get-feature(1)
- nvme-get-lba-status(1)
- nvme-get-log(1)
- nvme-get-ns-id(1)
- nvme-get-property(1)
- nvme-help(1)
- nvme-huawei-id-ctrl(1)
- nvme-huawei-list(1)
- nvme-id-ctrl(1)
- nvme-id-domain(1)
- nvme-id-iocs(1)
- nvme-id-ns(1)
- nvme-id-nvmset(1)
- nvme-intel-id-ctrl(1)
- nvme-intel-internal-log(1)
- nvme-intel-lat-stats(1)
- nvme-intel-market-name(1)
- nvme-intel-smart-log-add(1)
- nvme-intel-temp-stats(1)
- nvme-io-passthru(1)
- nvme-lba-status-log(1)
- nvme-list-ctrl(1)
- nvme-list-endgrp(1)
- nvme-list-ns(1)
- nvme-list-subsys(1)
- nvme-list(1)
- nvme-lnvm-create(1)
- nvme-lnvm-diag-bbtbl(1)
- nvme-lnvm-diag-set-bbtbl(1)
- nvme-lnvm-factory(1)
- nvme-lnvm-id-ns(1)
- nvme-lnvm-info(1)
- nvme-lnvm-init(1)
- nvme-lnvm-list(1)
- nvme-lnvm-remove(1)
- nvme-lockdown(1)
- nvme-micron-clear-pcie-errors(1)
- nvme-micron-internal-log(1)
- nvme-micron-nand-stats(1)
- nvme-micron-pcie-stats(1)
- nvme-micron-selective-download(1)
- nvme-micron-smart-add-log(1)
- nvme-micron-temperature-stats(1)
- nvme-netapp-ontapdevices(1)
- nvme-netapp-smdevices(1)
- nvme-ns-descs(1)
- nvme-ns-rescan(1)
- nvme-nvm-id-ctrl(1)
- nvme-persistent-event-log(1)
- nvme-pred-lat-event-agg-log(1)
- nvme-predictable-lat-log(1)
- nvme-primary-ctrl-caps(1)
- nvme-read(1)
- nvme-reset(1)
- nvme-resv-acquire(1)
- nvme-resv-notif-log(1)
- nvme-resv-register(1)
- nvme-resv-release(1)
- nvme-resv-report(1)
- nvme-rpmb(1)
- nvme-sanitize-log(1)
- nvme-sanitize(1)
- nvme-security-recv(1)
- nvme-security-send(1)
- nvme-self-test-log(1)
- nvme-set-feature(1)
- nvme-set-property(1)
- nvme-show-hostnqn(1)
- nvme-show-regs(1)
- nvme-smart-log(1)
- nvme-subsystem-reset(1)
- nvme-supported-log-pages(1)
- nvme-telemetry-log(1)
- nvme-toshiba-clear-pcie-correctable-errors(1)
- nvme-toshiba-vs-internal-log(1)
- nvme-toshiba-vs-smart-add-log(1)
- nvme-transcend-badblock(1)
- nvme-transcend-healthvalue(1)
- nvme-verify(1)
- nvme-virtium-save-smart-to-vtview-log(1)
- nvme-virtium-show-identify(1)
- nvme-wdc-cap-diag(1)
- nvme-wdc-capabilities(1)
- nvme-wdc-clear-assert-dump(1)
- nvme-wdc-clear-fw-activate-history(1)
- nvme-wdc-clear-pcie-corr(1)
- nvme-wdc-clear-pcie-correctable-errors(1)
- nvme-wdc-cloud-SSD-plugin-version(1)
- nvme-wdc-drive-essentials(1)
- nvme-wdc-drive-log(1)
- nvme-wdc-drive-resize(1)
- nvme-wdc-enc-get-log(1)
- nvme-wdc-get-crash-dump(1)
- nvme-wdc-get-drive-status(1)
- nvme-wdc-get-latency-monitor-log(1)
- nvme-wdc-get-pfail-dump(1)
- nvme-wdc-id-ctrl(1)
- nvme-wdc-log-page-directory(1)
- nvme-wdc-namespace-resize(1)
- nvme-wdc-purge-monitor(1)
- nvme-wdc-purge(1)
- nvme-wdc-smart-add-log(1)
- nvme-wdc-smart-log-add(1)
- nvme-wdc-vs-drive-info(1)
- nvme-wdc-vs-error-reason-identifier(1)
- nvme-wdc-vs-fw-activate-history(1)
- nvme-wdc-vs-internal-log(1)
- nvme-wdc-vs-nand-stats(1)
- nvme-wdc-vs-smart-add-log(1)
- nvme-wdc-vs-telemetry-controller-option(1)
- nvme-wdc-vs-temperature-stats(1)
- nvme-write-uncor(1)
- nvme-write-zeroes(1)
- nvme-write(1)
- nvme-zns-changed-zone-list(1)
- nvme-zns-close-zone(1)
- nvme-zns-finish-zone(1)
- nvme-zns-id-ctrl(1)
- nvme-zns-id-ns(1)
- nvme-zns-offline-zone(1)
- nvme-zns-open-zone(1)
- nvme-zns-report-zones(1)
- nvme-zns-reset-zone(1)
- nvme-zns-set-zone-desc(1)
- nvme-zns-zone-append(1)
- nvme-zns-zone-mgmt-recv(1)
- nvme-zns-zone-mgmt-send(1)
- nvme(1)

Kamal Bakshi, Cisco

NVMe-oF Comparison

FC-SB/CKD FICON (FC)

(Not NVMe)

IBM Z mainframes process 30 billion transactions each day, including 87% of all credit card transactions on the planet.

-96 of the world's top 100 banks and 9 out of 10 of the world's biggest insurance companies still depend on mainframes (source google)

-Mainframe storage standard

FC-SCSI (FCP)

-120millions* FC ports shipped
-46millions* in use (FCIA* website)
-Dedicated purpose built Storage Network
-Built in Discovery & Name services
-Zoning & Security
-Lossless Fabric/Zero Copy
-Certified designs
-Gold standard in Enterprise storage

NVMe-FC

-Faster than FC-SCSI
-Advance Error detection & recovery
-Same FC transport

32G/64G
Fibre Channel Transport

NVMe-IB

-Infiniband Transport
-Lossless Infiniband Links
-HPC supercomputer
-RDMA, Zero Copy
-Low Latency
-IB stack offload

200G Infiniband Transport

NVMe-UPD/RoCEv2

-Infiniband Transport
-Lossless Ethernet Links
-RDMA, Zero Copy
-Low Latency
-IB stack offload
-Higher Performance than TCP

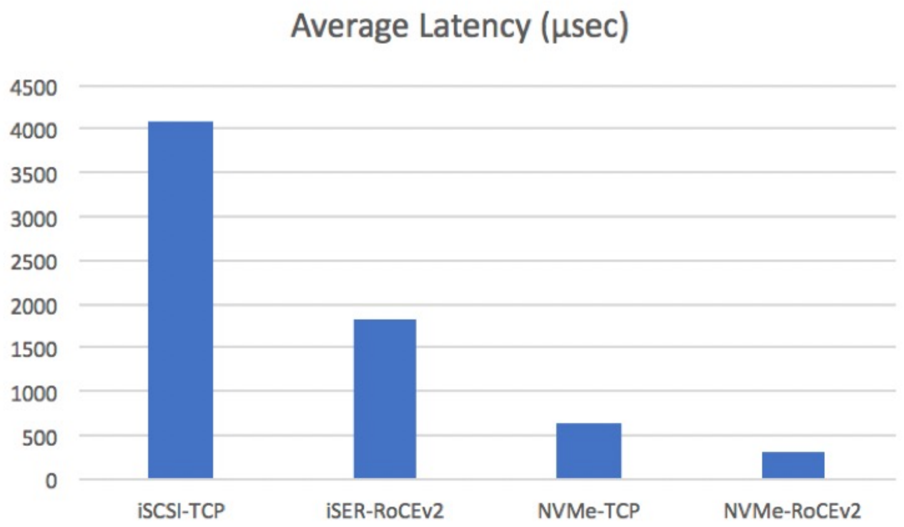
NVMe-TCP

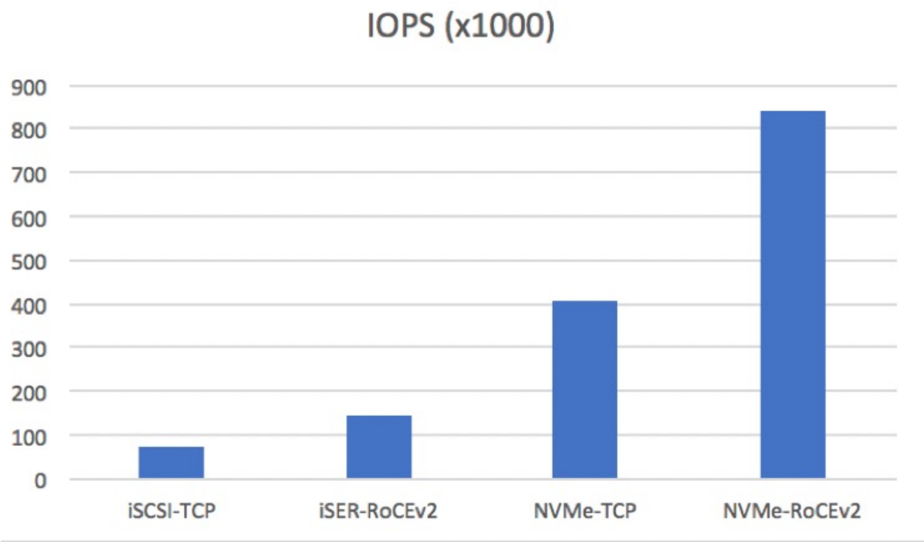
-Ubiquitous
-Scalable, simpler
-Price/Performance benefit
-Ample skillset
-(Faster than iSCSI)

400G Ethernet Transport

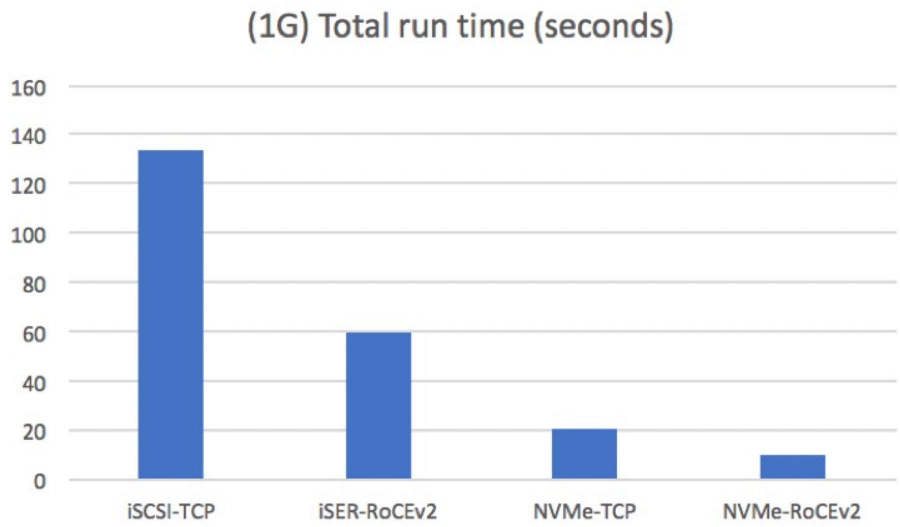
Kamal Bakshi, Cisco

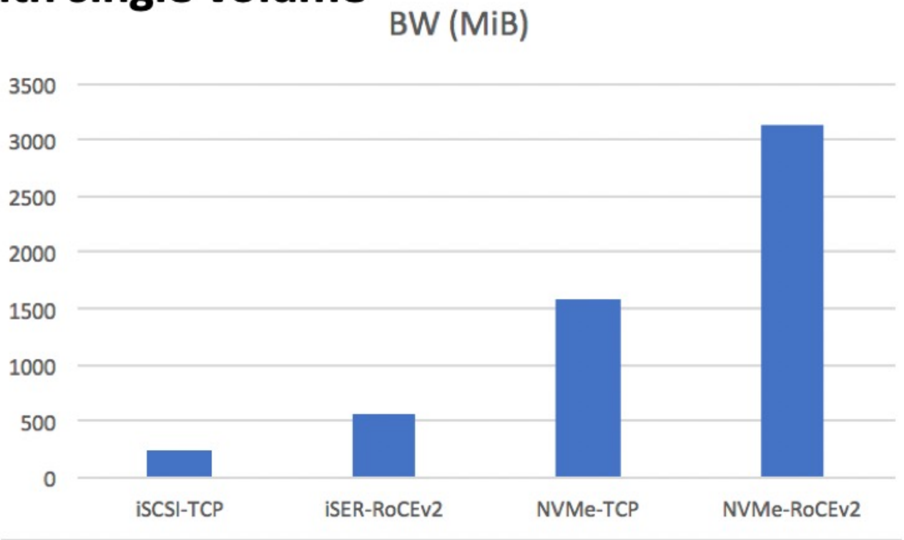
iSCSI vs NVMe-IP





4KB Random Reads 1G with single volume





Kamal Bakshi, Cisco

NVMe Commands

Control Plane

NVMe-Admin

-Create I/O SQ
-Delete I/O SQ
-Create I/O CQ
-Delete I/O CQ
-Get Features
-Set Features
-Keep Alive
-Identify
-Get Log Pages
-Abort
-Directive Send
-Directive Receive
-Async. Event Req.
-Namespace Mgmt.
-Namespace Attachment
-Virtualization Mgmt.

-Firmware Image Download
-Firmware Commit
-Device Self test
-NVMe-MI Send
-NVMe-MI Receive
-Door bell Buffer Config.

-Format NVM
-Sanitize
-Get LBA Status
-Security Send
-Security Receive

Transport over Fabric

NVMe-oF

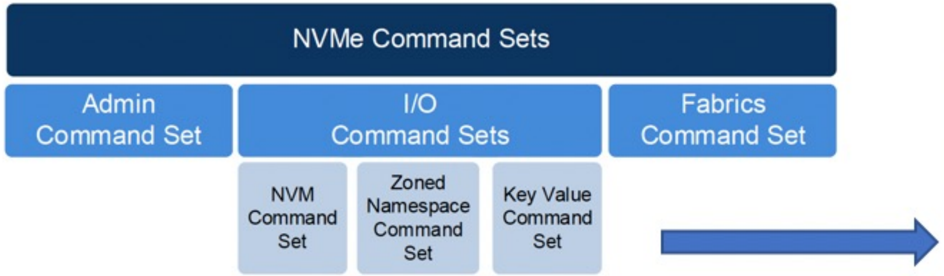
-Connect
-Disconnect
-Authentication Send
-Authentication Receive
-Property Get
-Property Set

Data Exchange

NVMe-I/O

-Write
-Write Uncorrectable
-Write Zeroes
-Flush
-Read
-Compare
-Verify
-Dataset Management
-Reservation Report
-Reservation Acquire
-Reservation Release

NVMe 2.0



NVMe SSD

ZNS NVMe

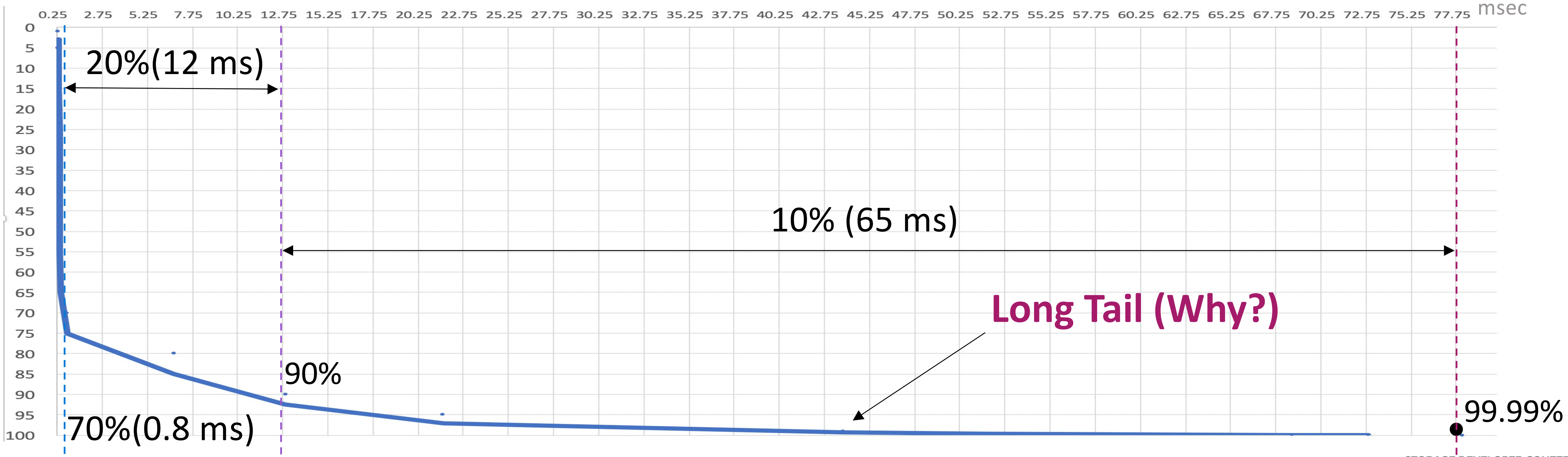
Key Value NVMe

Kamal Bakshi, Cisco

Tail Latency (Long Tail)

I/O Completion Latency

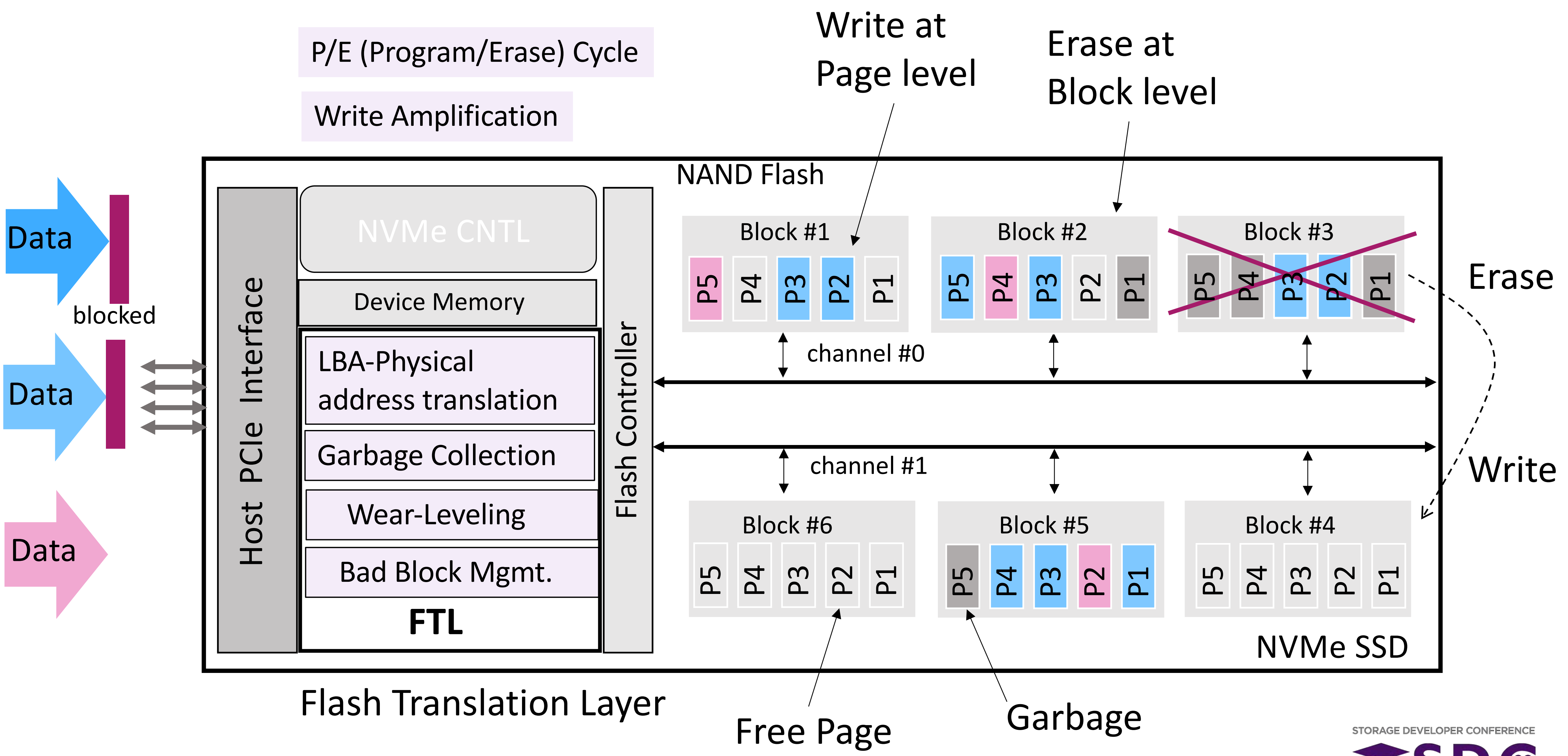
```
clat percentiles (usec):
 |  1.00th=[  302],  5.00th=[  326], 10.00th=[  343], 20.00th=[  363],
 | 30.00th=[  392], 40.00th=[  404], 50.00th=[  416], 60.00th=[  445],
 | 70.00th=[  816], 80.00th=[ 6718], 90.00th=[12911], 95.00th=[21627],
 | 99.00th=[43779], 99.50th=[51643], 99.90th=[68682], 99.95th=[72877],
 | 99.99th=[78119]
```



Kamal Bakshi, Cisco

STORAGE DEVELOPER CONFERENCE SDC 22

Flash Internals



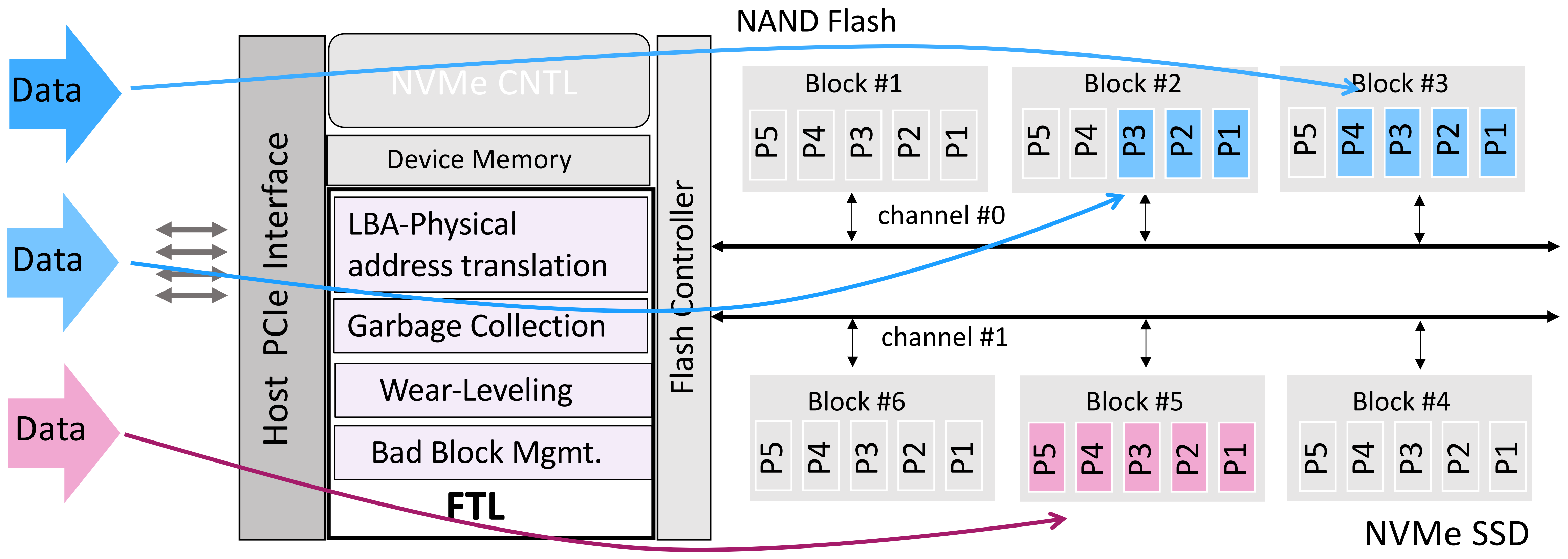
Kamal Bakshi, Cisco

Predictable Latency

I/O Determinism
- -NVMe Streams
- -NVMe Sets

Open Channel
- -Parallel Units/Chunks
- -LightNVM

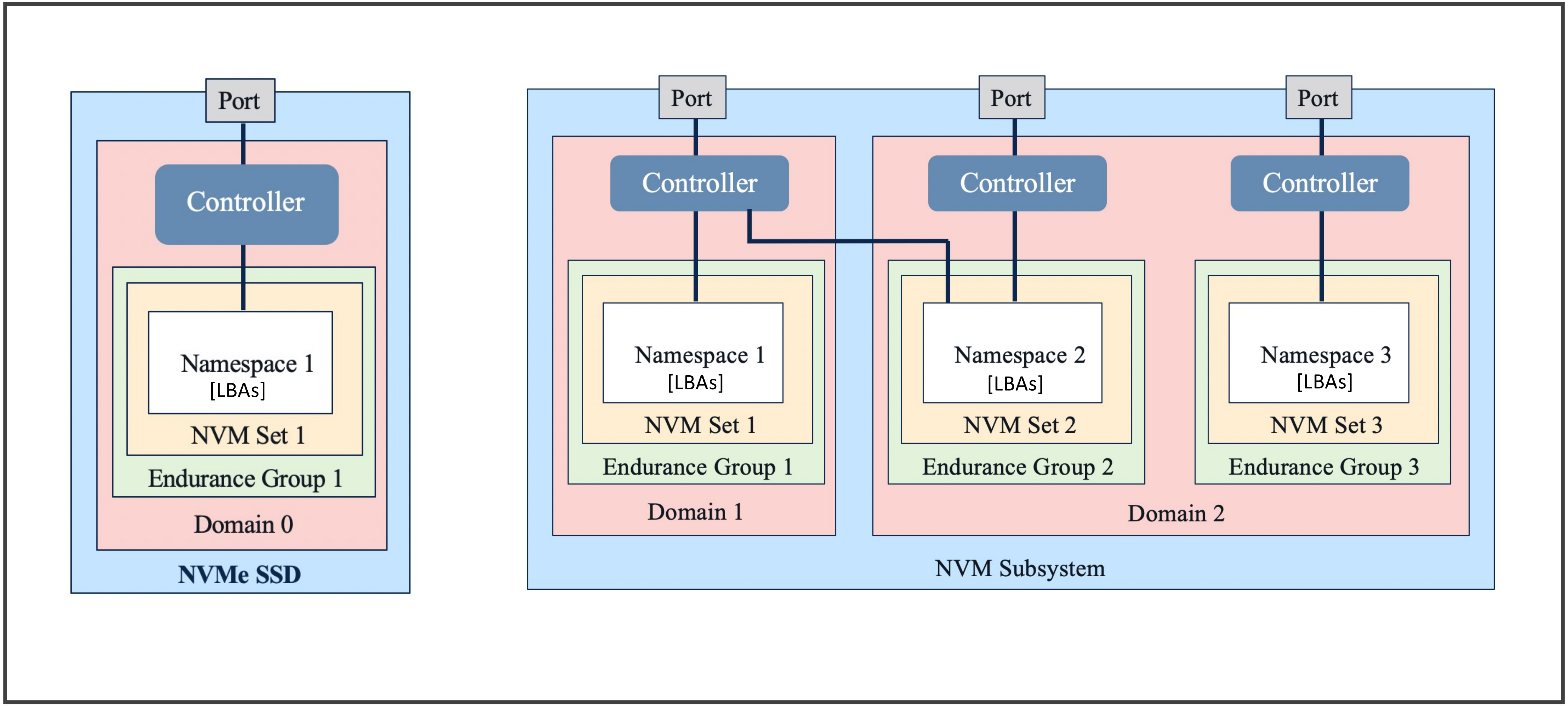
-NVMe ZNS
(Zoned Namespaces)



Flash Translation Layer

Kamal Bakshi, Cisco

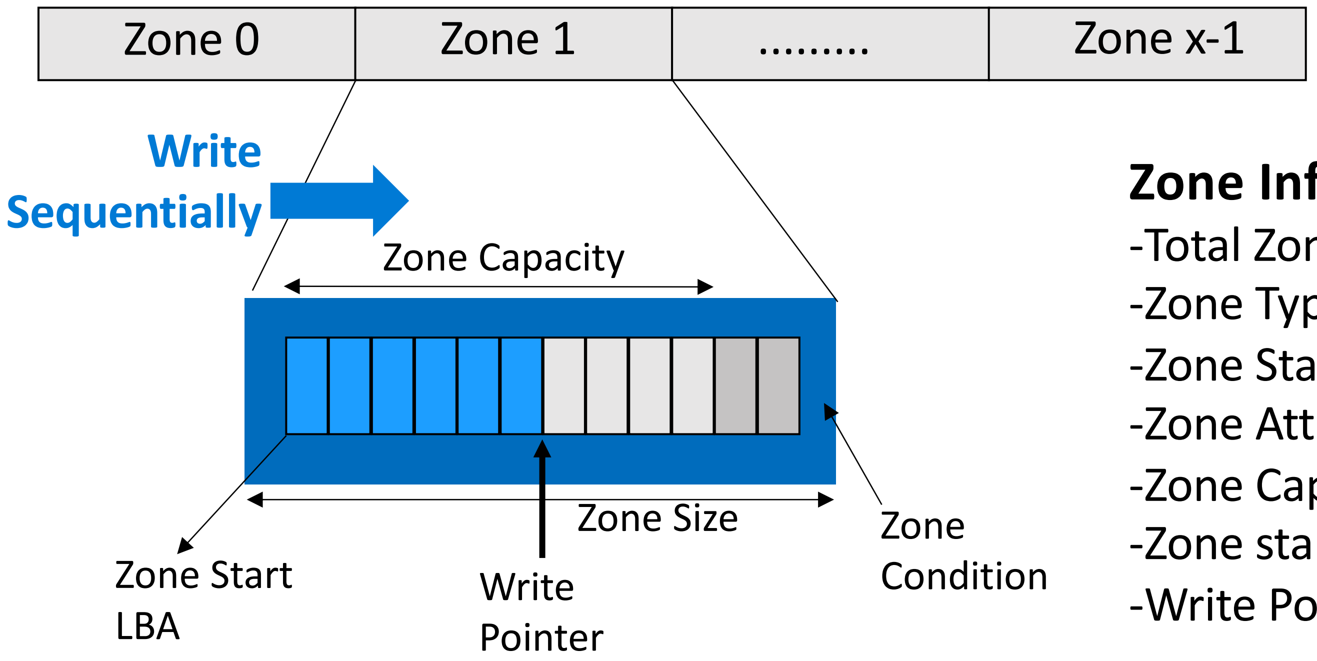
I/O Determinism (NVM Set)



Kamal Bakshi, Cisco

ZNS Zone Namespace

-Lower Write Amplification
-Lower P/E cycle (increased SSD life)
-Predictable Latency



Zone Info

-Total Zone
-Zone Type
-Zone State
-Zone Attribute
-Zone Capacity
-Zone start LBA
-Write Pointer

Host:

- Address Mapping
- Data Placement
- Garbage Collection
- Wear Leveling
- NS, Streams

SSD:

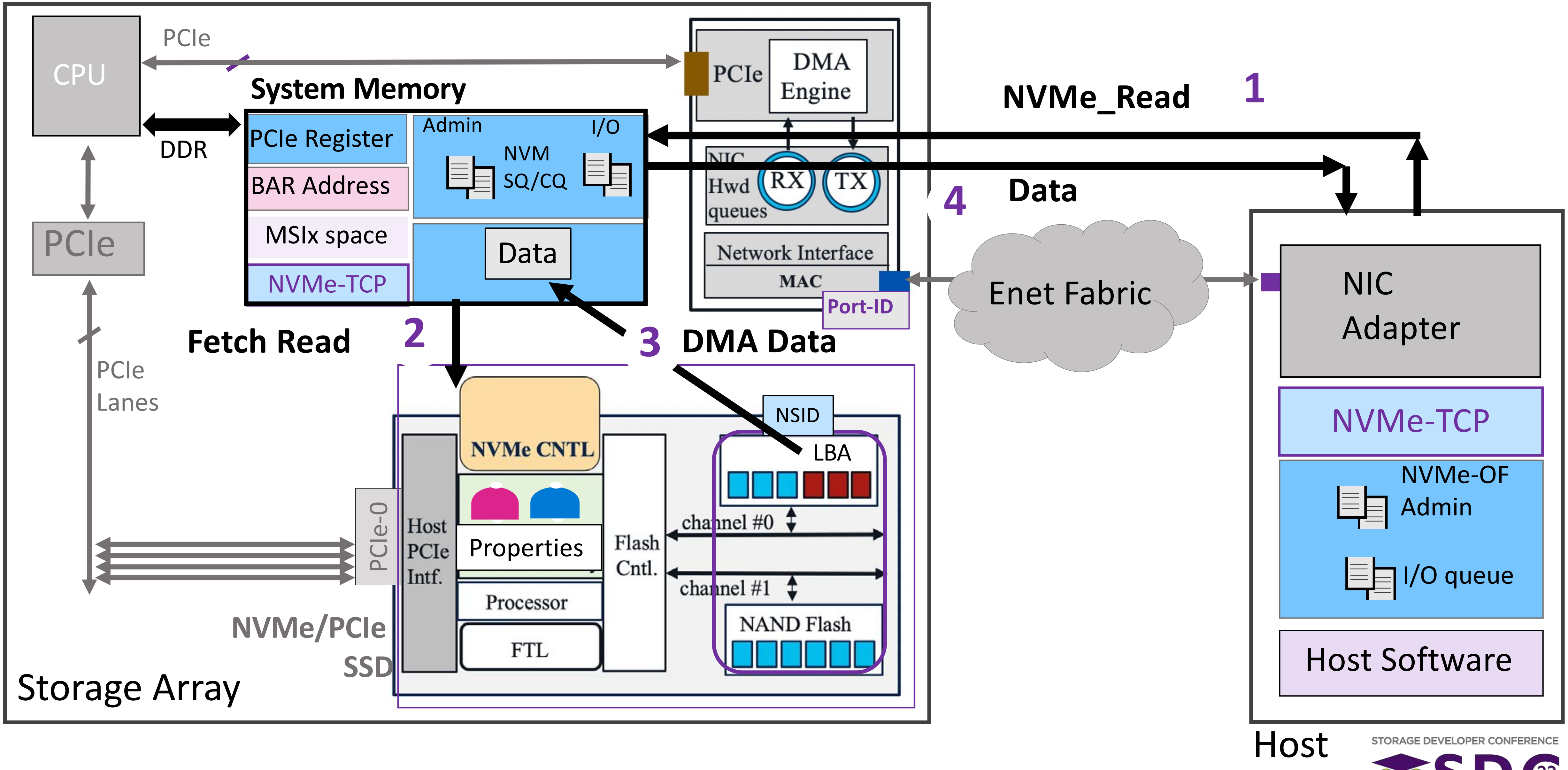
- Error retry handling
- Bad Block Mgmt.

New NVMe Commands

-Zone Mgmt. Send/Rcv
-Zone Append
-Zone Copy
-Zone Commit

Kamal Bakshi, Cisco

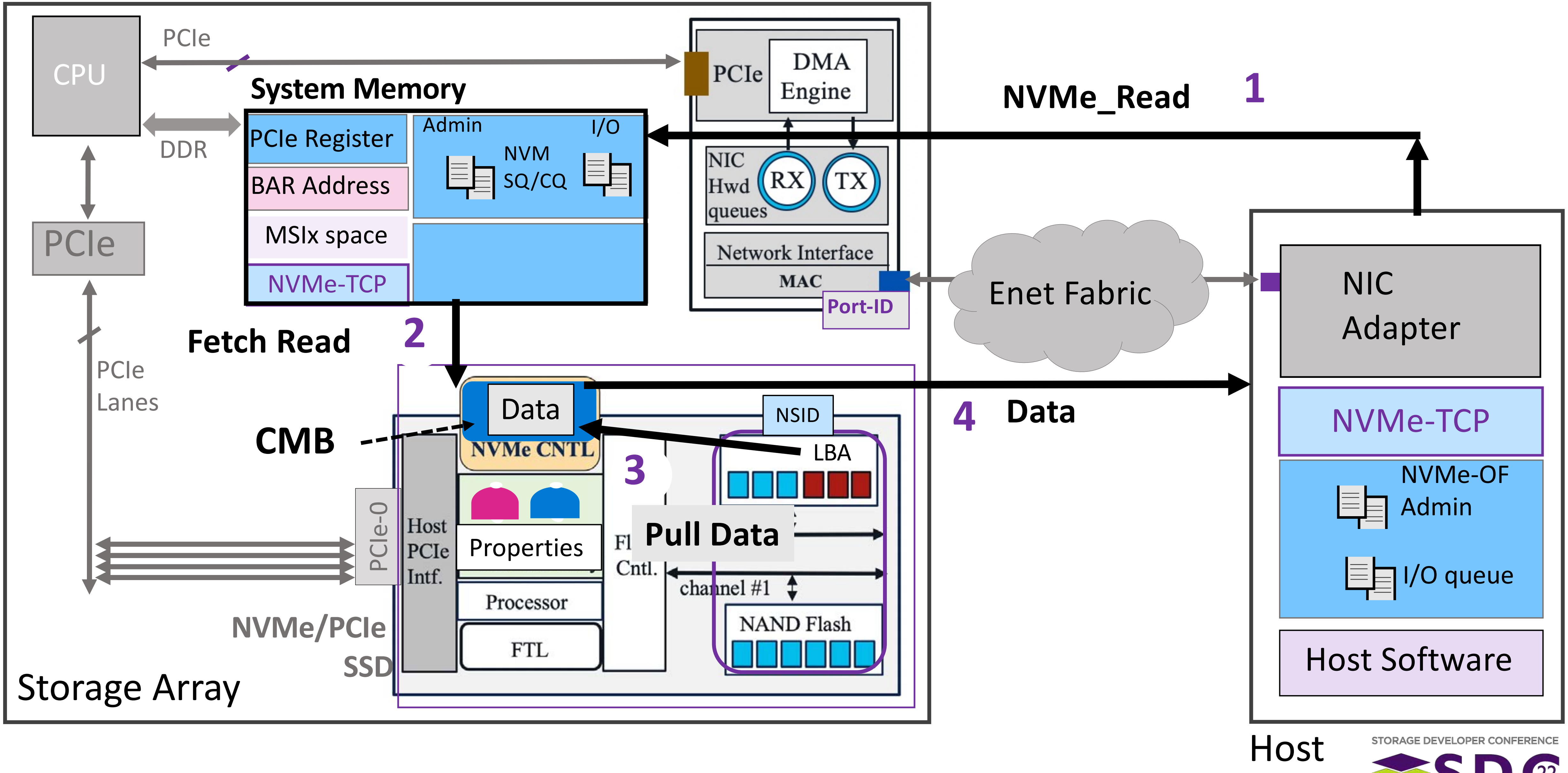
System Memory / CPU Intensive



Kamal Bakshi, Cisco

CMB / CPU Offloaded

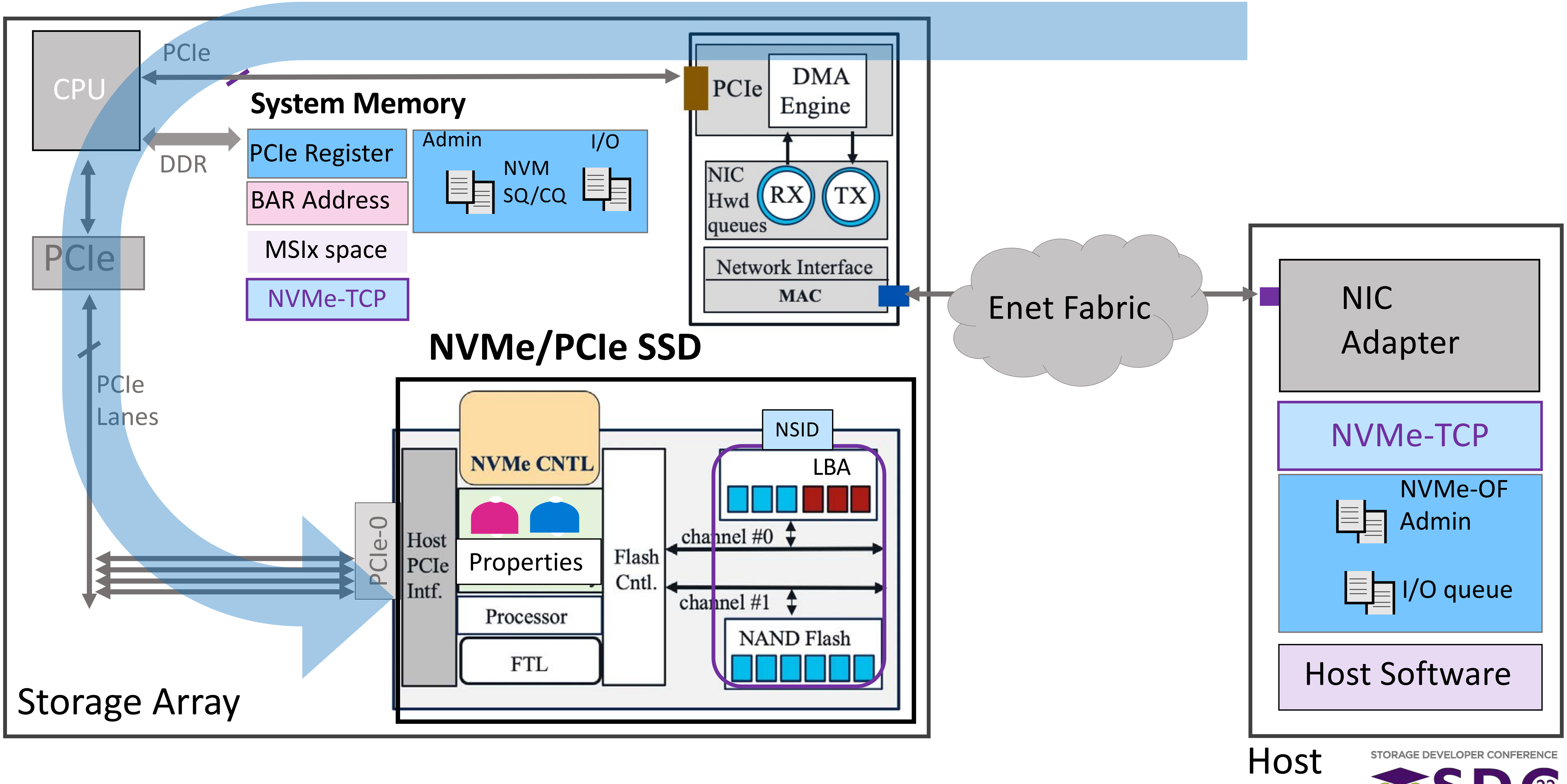
CMB Controller Memory Buffer



 |

Kamal Bakshi, Cisco

NVMe PCIe SSD

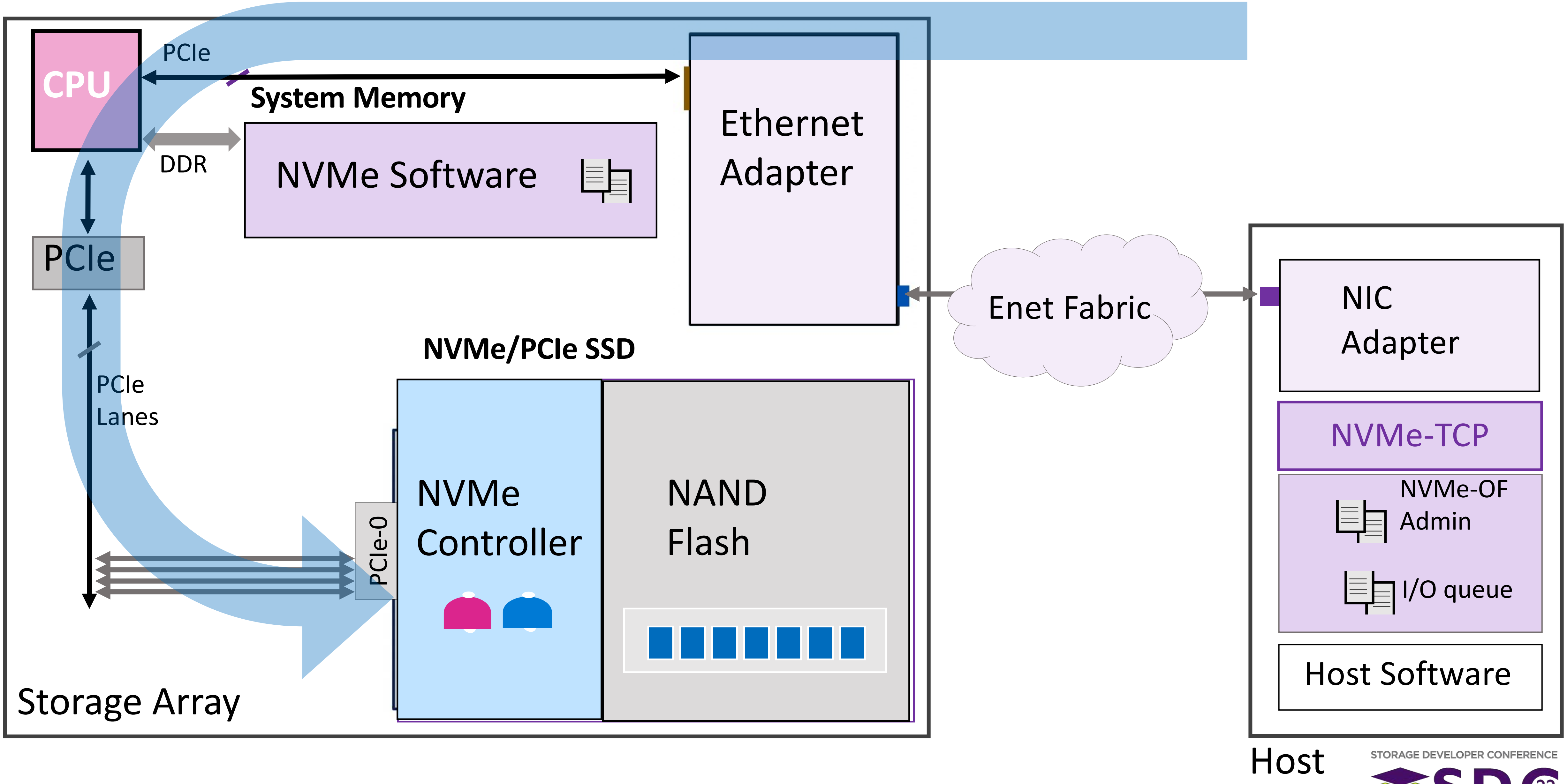


 |

Kamal Bakshi, Cisco

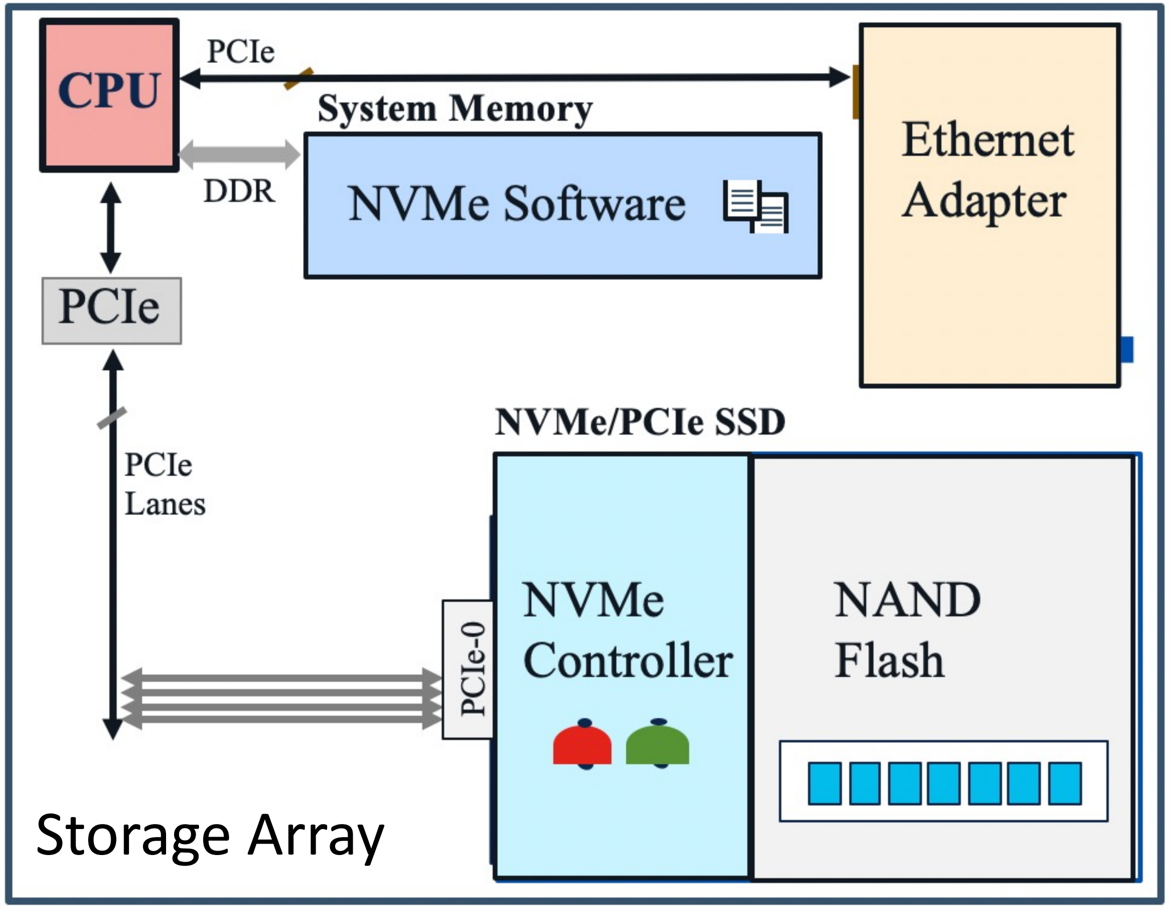
STORAGE DEVELOPER CONFERENCE SDC 22

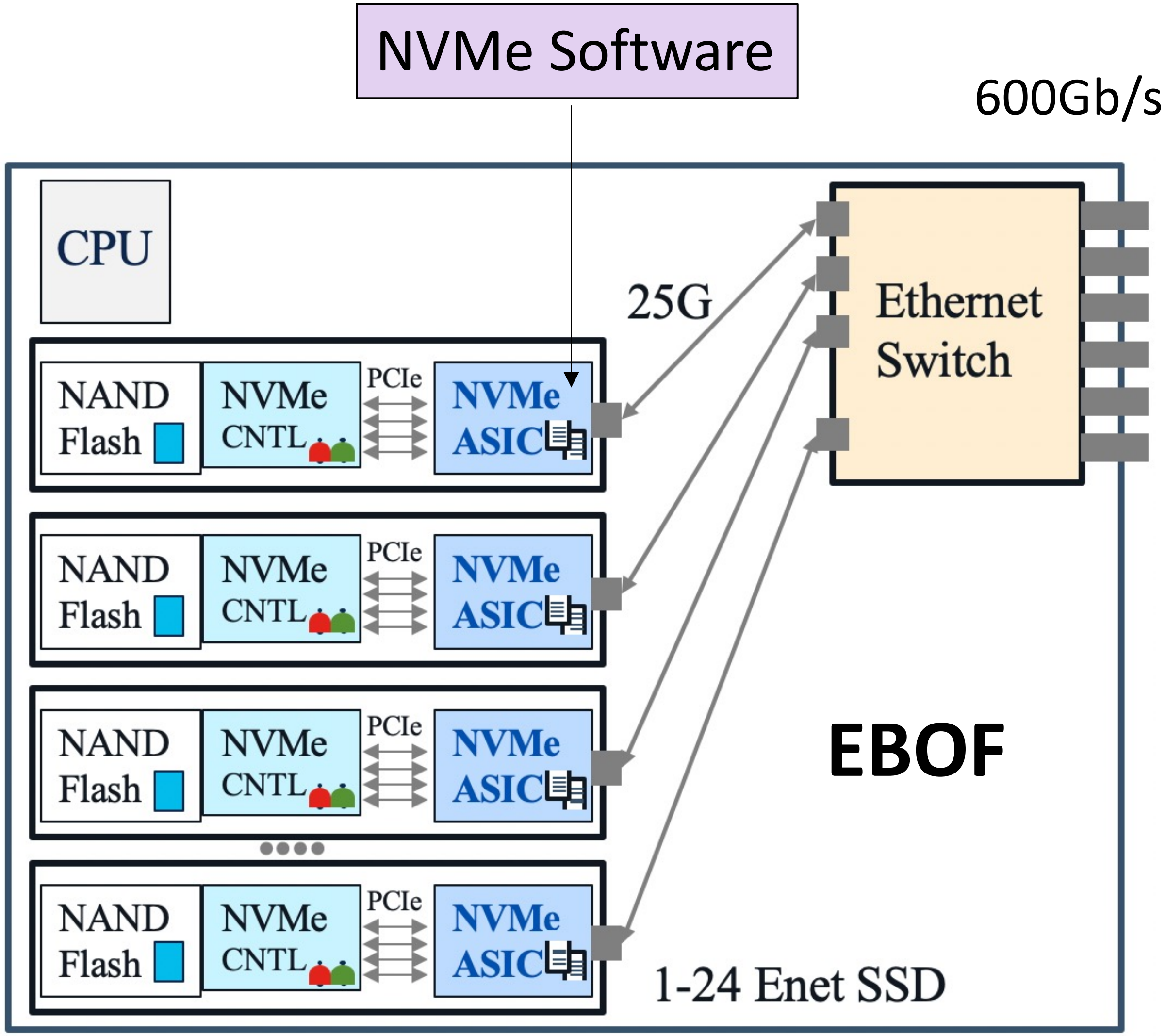
NVMe PCIe SSD (CPU is the Bottleneck)



Kamal Bakshi, Cisco

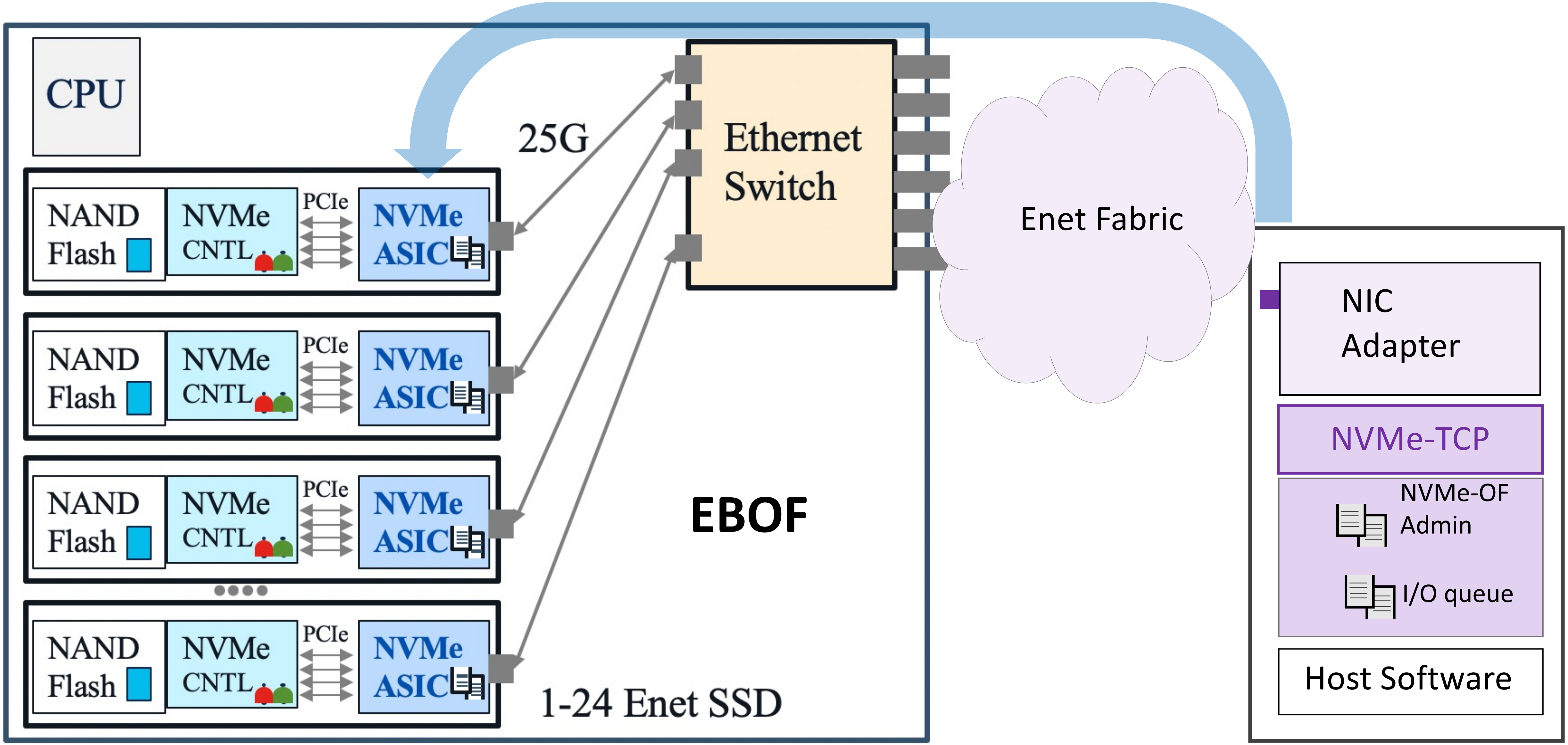
NVMe-oF Drive (Ethernet SSD)





Kamal Bakshi, Cisco

NVMe-oF Drive (Ethernet SSD)



Kamal Bakshi, Cisco

NVMe Key Value SSD

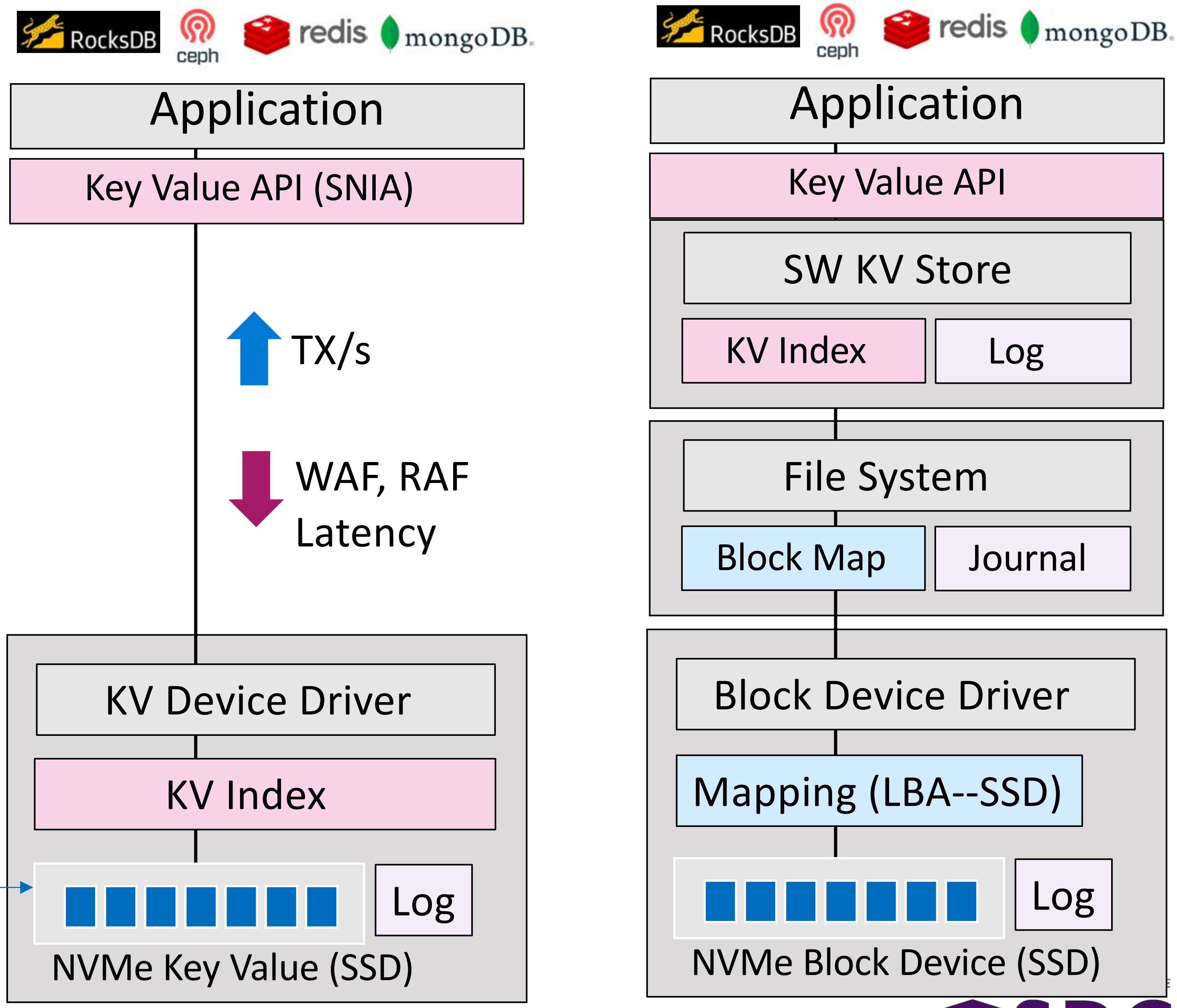
Today all storage protocols (Block, NFS or Object) uses LBA block addressing scheme.

KV protocol maps an address (Key, 32 bytes max.) to a physical location where (Value, 4GB max) is storage. No LBA, hence no translation in FTL.

Key Value API (SNIA)

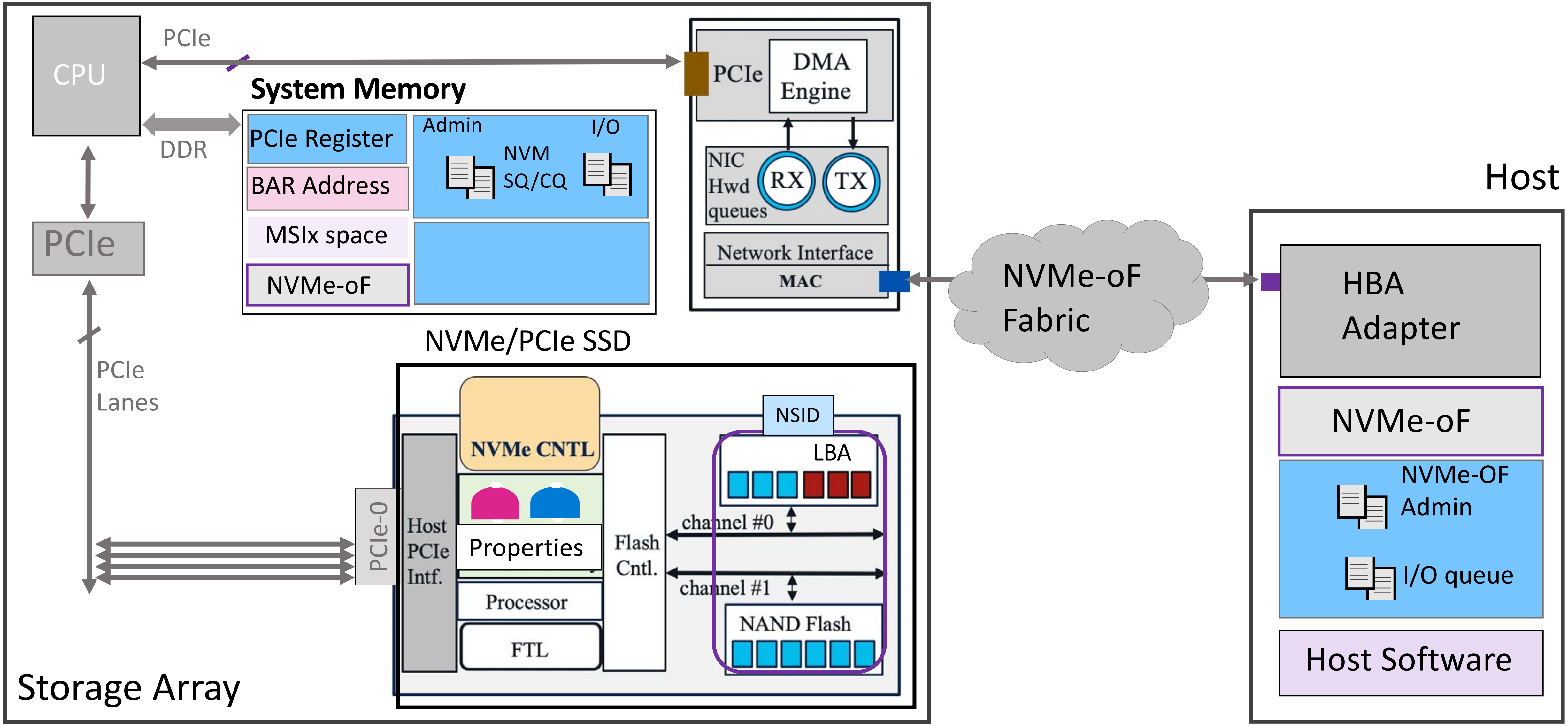
-Open/Retrieve Device
-Create/Delete Key Space
-Store, Retrieve, Delete,
-List, Delete Group

NVMe KV I/O Commands
(Store, Retrieve, List, Exist, Delete)



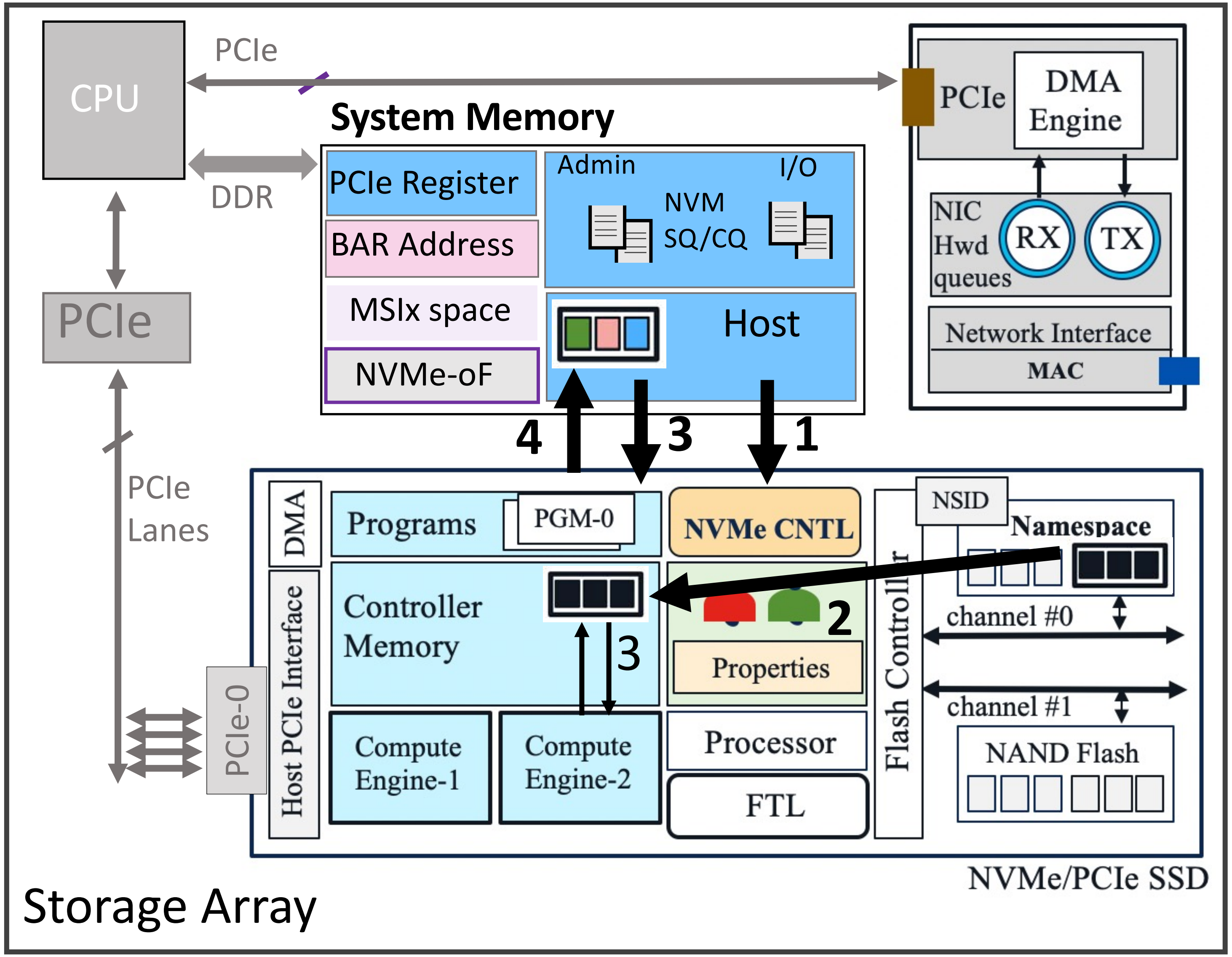
Kamal Bakshi, Cisco

NVMe Computational Storage



Kamal Bakshi, Cisco

NVMe Computational Storage



1 NVMe Read (NS) is issued to CNTL

2 CNTL moves the (NS) data to CM

3 Execute PGM-0 on CE-2

4 Read CM Output Data back to Host

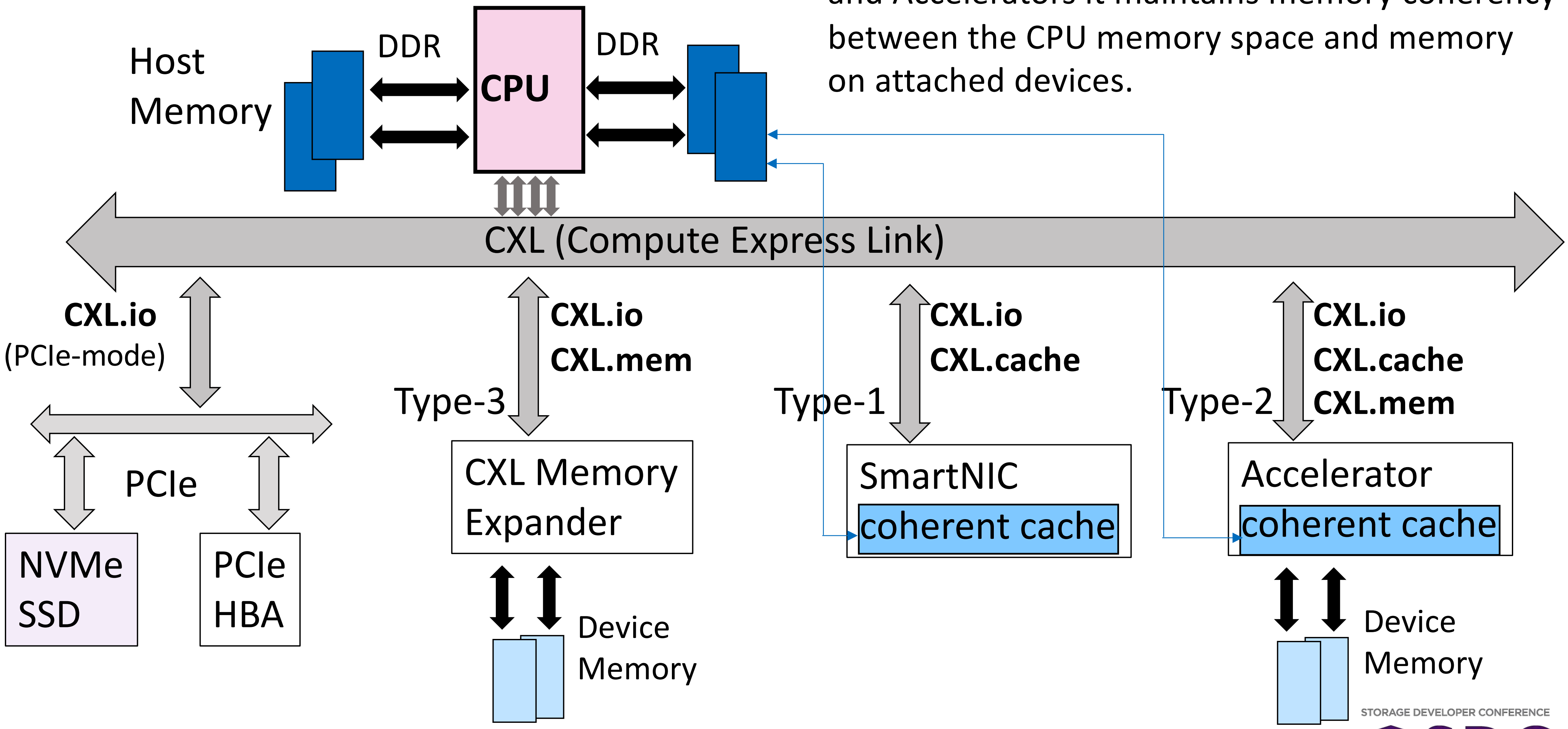
TP4091 Computational Programs
I/O Command Set
-Execute Program
-Load Program
-Activate Program

TP4131 Controller Local Memory

Kamal Bakshi, Cisco

CXL

CLX is an industry-supported Cache-Coherent Interconnect for Processors, Memory Expansion and Accelerators it maintains memory coherency between the CPU memory space and memory on attached devices.

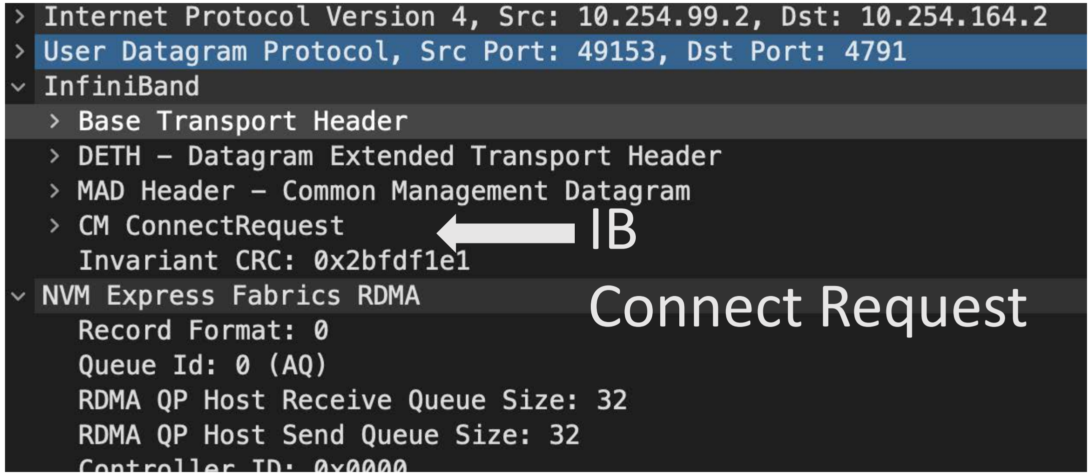


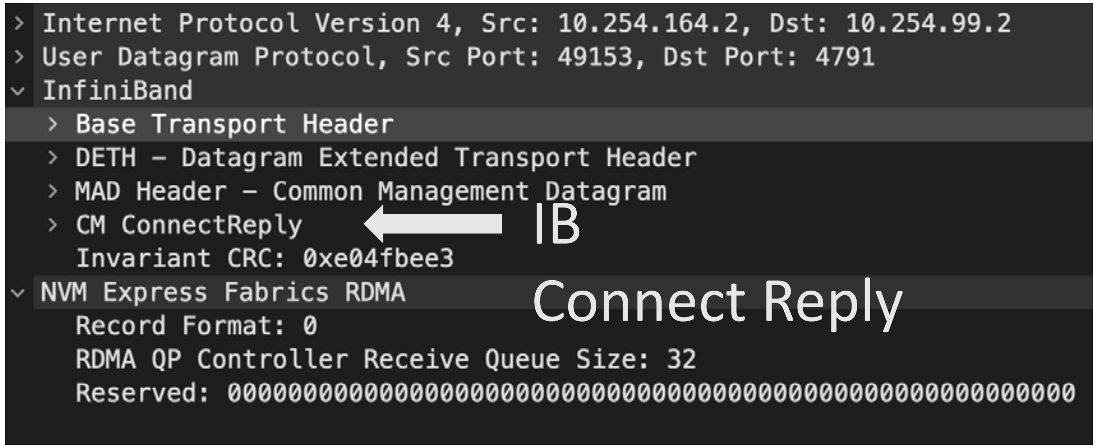
Kamal Bakshi, Cisco

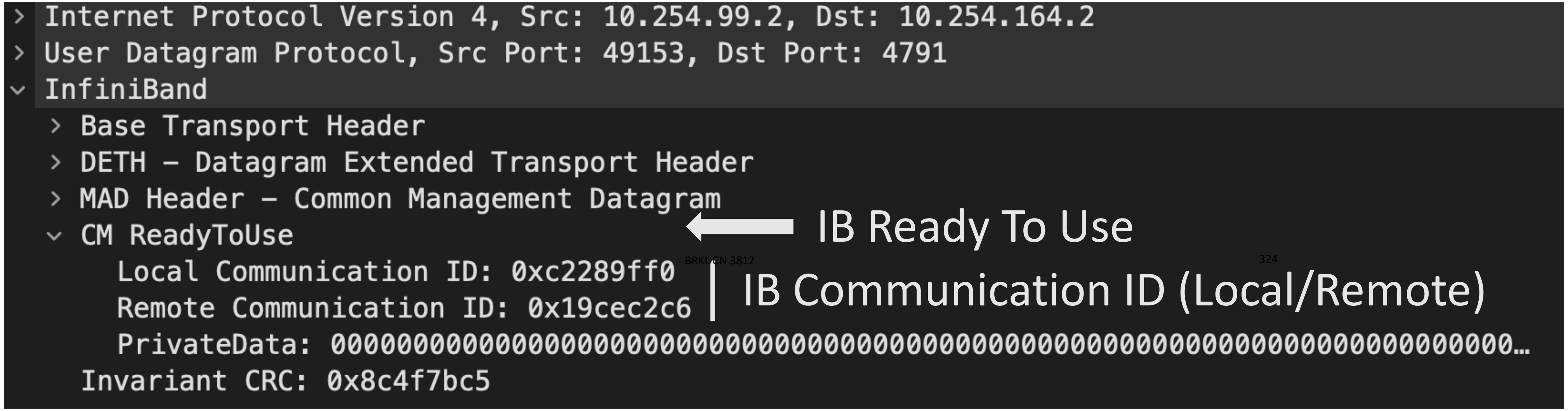
NVMe RoCEv2 examples

Kamal Bakshi, Cisco

NVMe-RoCEv2 (IB -CM Connection Manager)







Kamal Bakshi, Cisco

NVMe-RoCEv2 (NVMe-Connect)

```
> Internet Protocol Version 4, Src: 10.254.99.2, Dst: 10.254.164.2
> User Datagram Protocol, Src Port: 52195, Dst Port: 4791
v InfiniBand
  v Base Transport Header
      Opcode: Reliable Connection (RC) - SEND Only (4)
      0... .... = Solicited Event: False
      .1.. .... = MigReq: True
      ..00 .... = Pad Count: 0
      .... 0000 = Header Version: 0
      Partition Key: 65535
      Reserved: 00
      Destination Queue Pair: 0x000a52
      1... .... = Acknowledge Request: True
      .000 0000 = Reserved (7 bits): 0
      Packet Sequence Number: 9402734
    Invariant CRC: 0x6cfefa9e
v NVM Express Fabrics RDMA
    [Cmd Qid: 0 (AQ)]
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [DATA Transfer Request: 6]
      [Fabric Cqe in: 8]
      Reserved: 0x40
      Command Identifier: 0x0001
      Fabric Command Type: Connect (0x01)
      Reserved: 0000000000000000000000000000000000000
    > SGL1
      Record Format: 0
      Queue ID: 0 (AQ)
      Submission Queue Size: 32
    > Connect Attributes: 0x00
      Reserved: 00
      Keep Alive Timeout: 0ms
      Reserved: 000000000000000000000000
```

NVMe Connect Request

BRKDCN 3812

325

Kamal Bakshi, Cisco

NVMe-RoCEv2 (NVMe Get Property)

```
> Internet Protocol Version 4, Src: 10.254.99.2, Dst: 10.254.164.2
> User Datagram Protocol, Src Port: 52195, Dst Port: 4791
v InfiniBand
  v Base Transport Header
      Opcode: Reliable Connection (RC) - SEND Only (4)
      0... .... = Solicited Event: False
      .1.. .... = MigReq: True
      ..00 .... = Pad Count: 0
      .... 0000 = Header Version: 0
      Partition Key: 65535
      Reserved: 00
      Destination Queue Pair: 0x000a52
      1... .... = Acknowledge Request: True
      .000 0000 = Reserved (7 bits): 0
      Packet Sequence Number: 9402735
    Invariant CRC: 0x79327656
v NVM Express Fabrics RDMA
    [Cmd Qid: 0 (AQ)]
  v Cmd
      Opcode: 0x7f (Fabric Command)
      [Fabric Cqe in: 12]
      Reserved: 0x40
      Command Identifier: 0x001c
      Fabric Command Type: Property Get (0x04)
      Reserved: 0000000000000000000000000000000000000000000000000000000000000000000040
    > .... .001 = Attributes: 0x1
      Reserved: 000000
      Offset: Controller Capabilities (0x00000000)
      Reserved: 00000000000000000000000000000000
```

IB Queue#

NVMe Admin Queue#0

NVMe Get Property

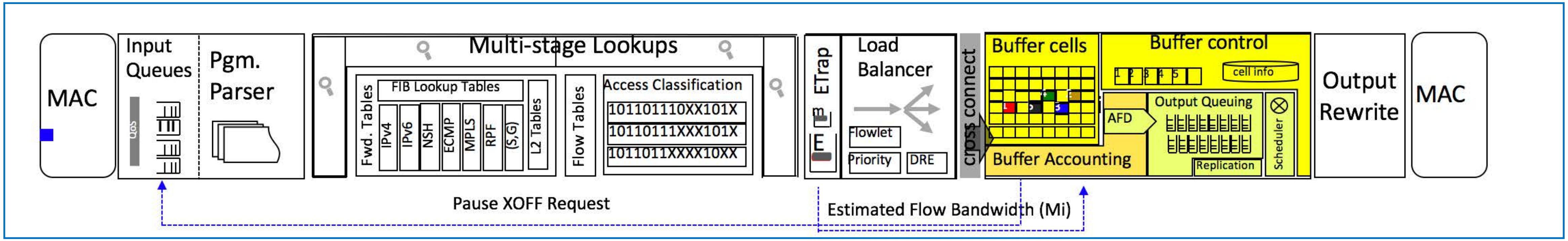
Kamal Bakshi, Cisco

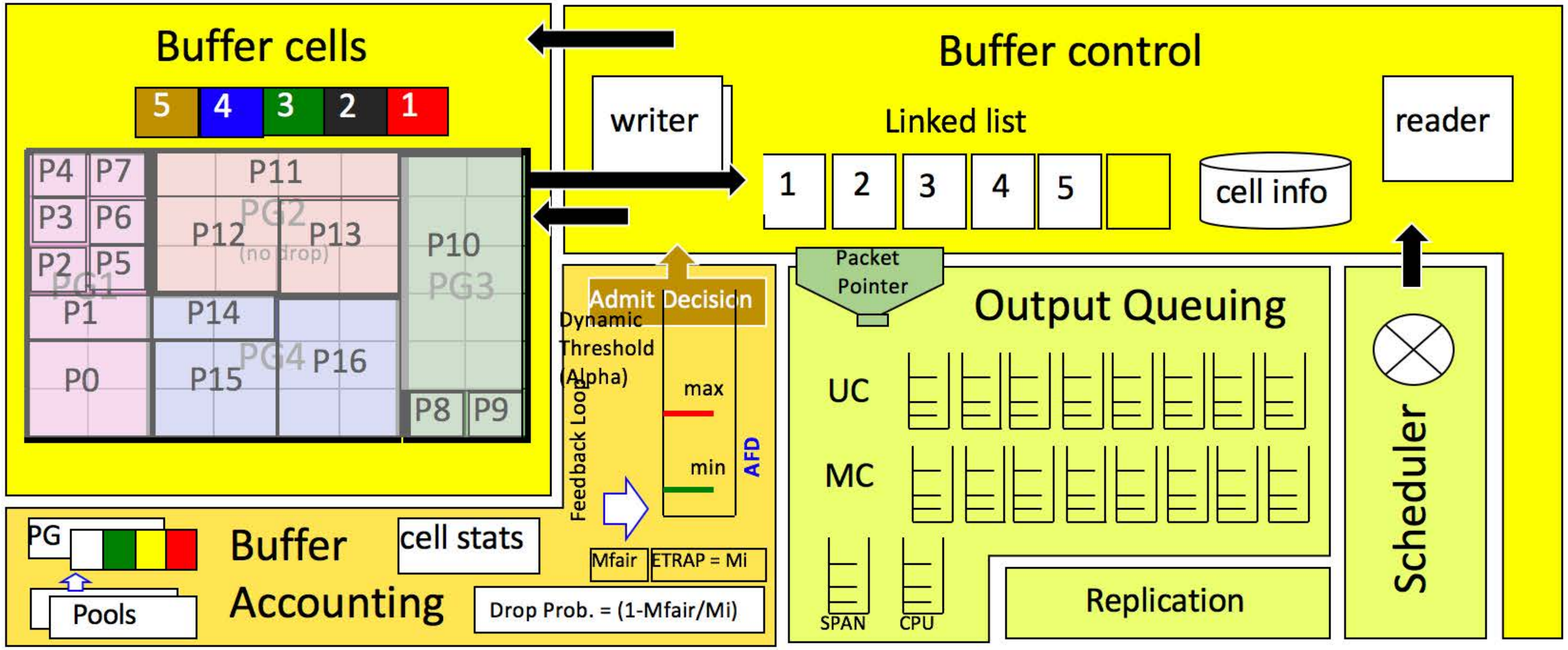
NVMe Misc. Slides

Kamal Bakshi, Cisco

Cisco N9k –Smart Buffering for NVMe/TCP

Cisco Cloud Scale ASIC - Pipeline



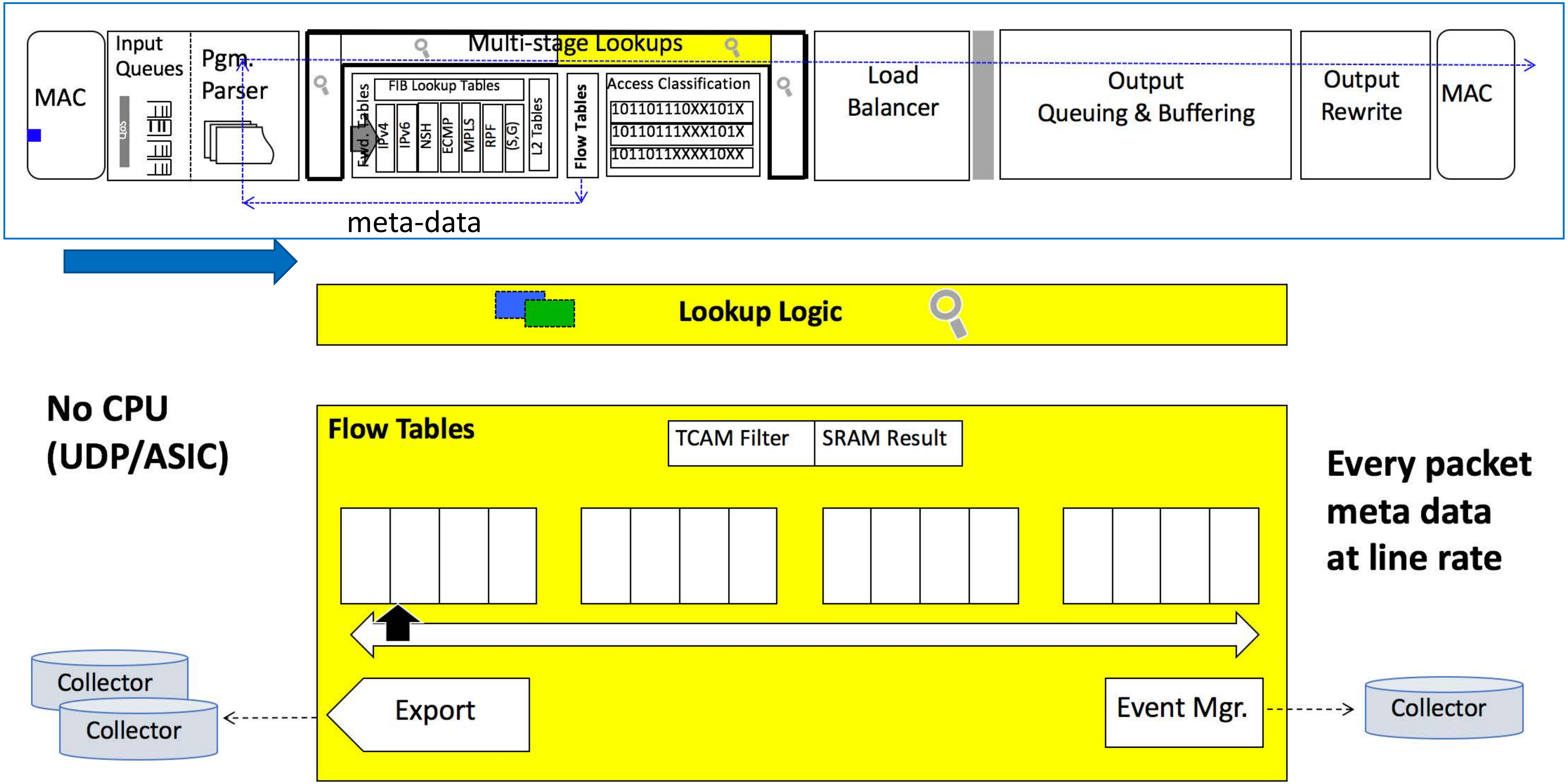


Approximate Fair Discard
Dynamic Packet Processing
On-Chip Memory

Kamal Bakshi, Cisco

Cisco N9k ASIC driven Analytics (TCP)

Cisco Cloud Scale ASIC - Pipeline



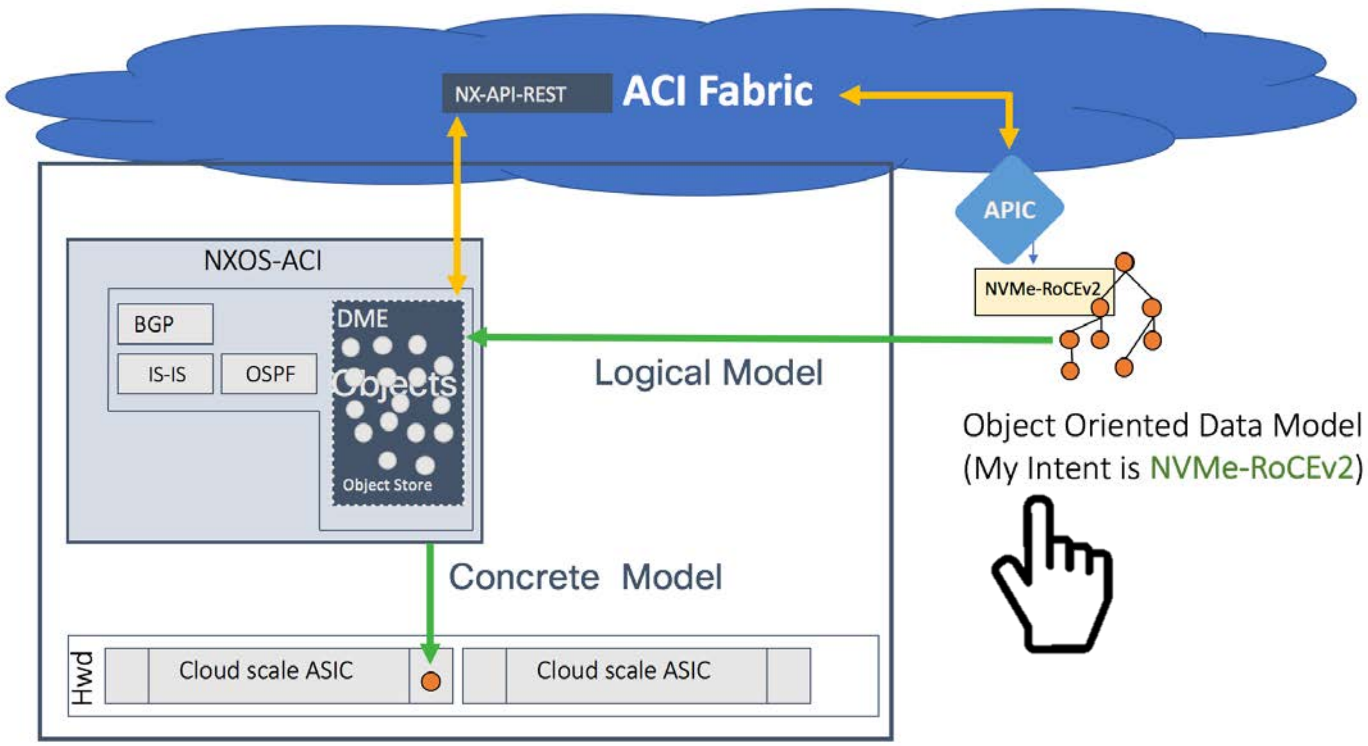
No CPU (UDP/ASIC)

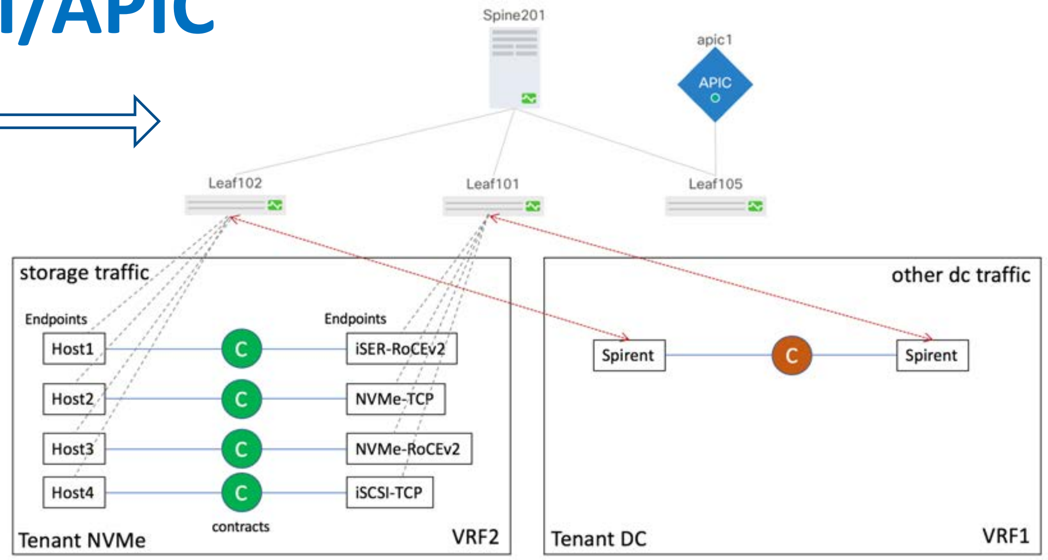
Every packet meta data at line rate

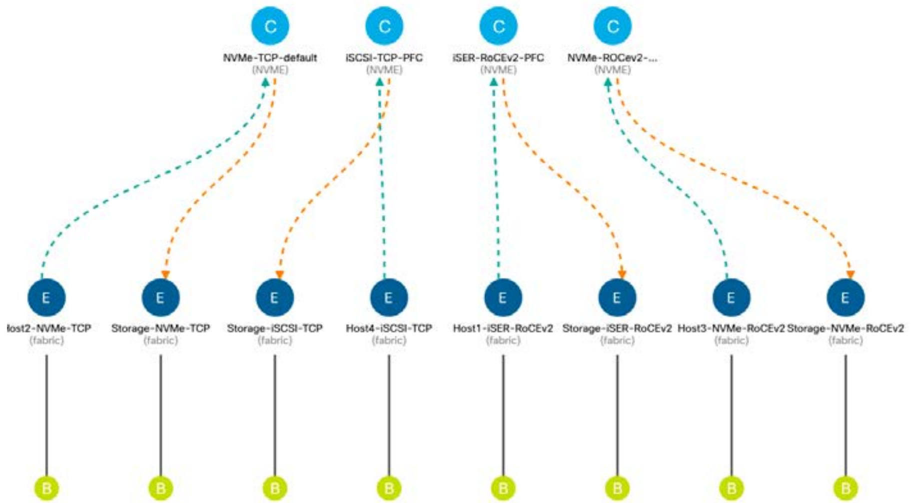
Cisco Network Insights

Kamal Bakshi, Cisco

NVMe Storage Automation with Cisco ACI/APIC

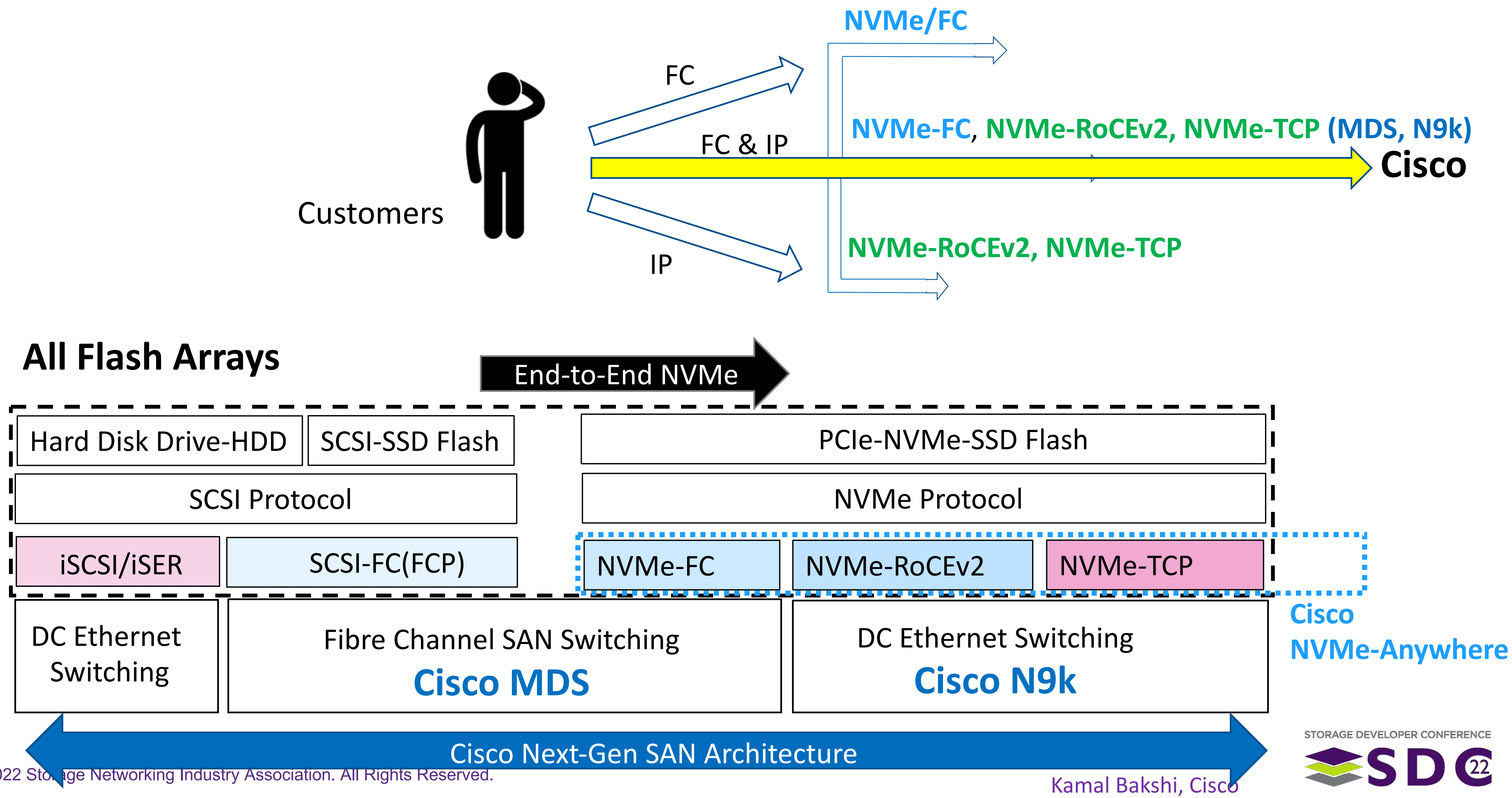




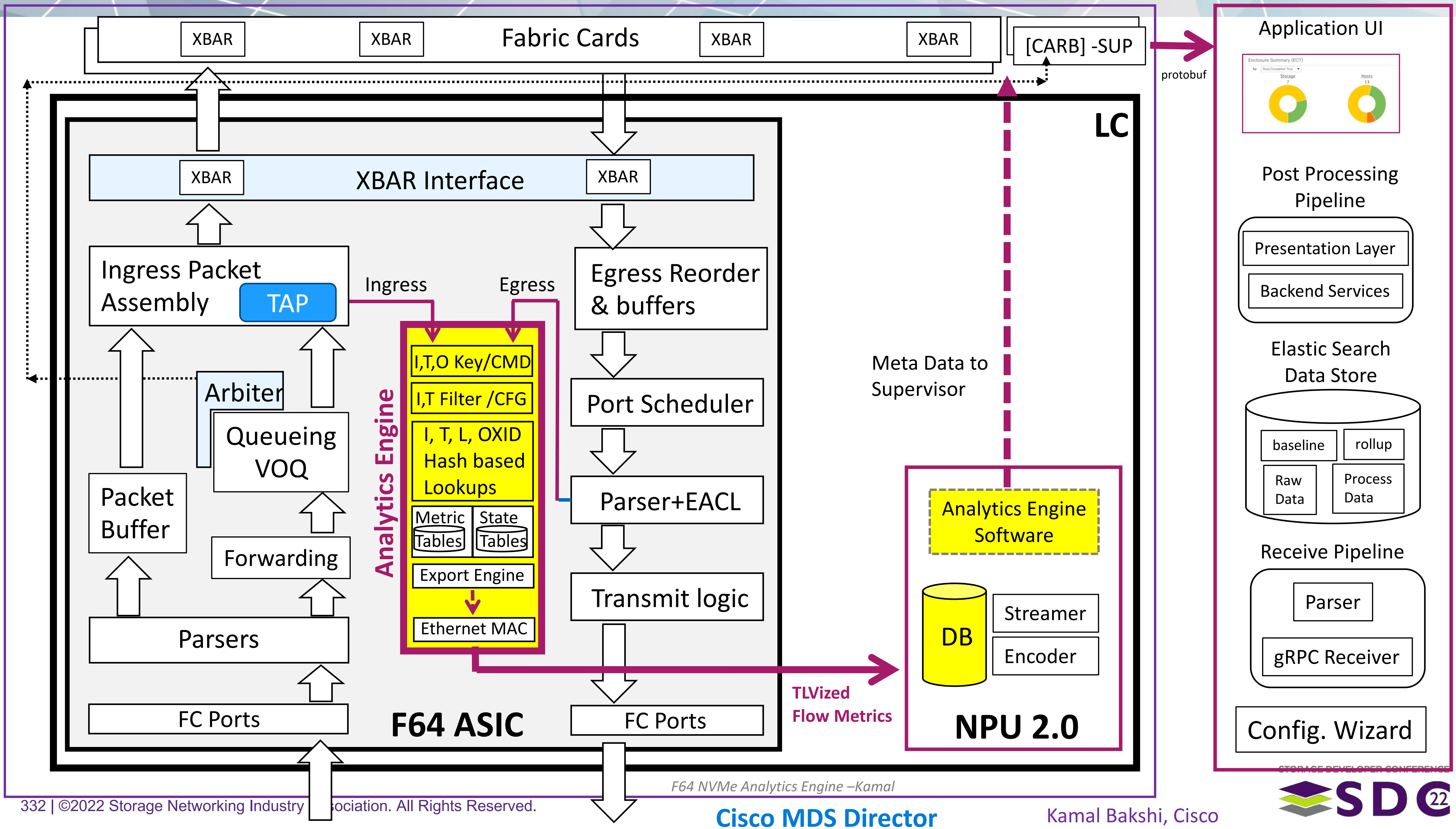


Kamal Bakshi, Cisco

End to End NVMe Enterprise Storage with Cisco networking



Kamal Bakshi, Cisco



Cisco MDS Director

Kamal Bakshi, Cisco

Thank You