FCIA
FIBRE CHANNEL INDUSTRY ASSOCIATION

NVMe® over FC: Deep Dive in Protocol, Architecture and Use Cases

Live Webcast

November 30, 2023

10:00 AM PT/1:00 PM ET

Today's Speakers

David Rodgers
Technical Business Development Manager
EXFO

Kamal Bakshi
Technical Director
Cisco

Nishant Lodha
Director, Emerging Technologies
Marvell Semiconductor

Mark Jones
Director Technical Marketing, Broadcom
FCIA Board of Directors

About the FCIA

"The Fibre Channel Industry Association (FCIA) is a mutual benefit, non-profit, international organization of manufacturers, system integrators, developers, vendors, industry professionals, and end users."

FCIA
FIBRE CHANNEL INDUSTRY ASSOCIATION

About the Fibre Channel Industry Association (FCIA)

25+ Years

Promoting Fibre Channel Technology

Industry Leading

Member Companies

160M FC Ports

Shipped Since 2001

FCIA
FIBRE CHANNEL INDUSTRY ASSOCIATION

2023 Fibre Channel Market Milestones

$50B FC Revenue

Total Revenue of Adapter and Switch products

35M+ FC Ports In Service

Adapter and switch ports estimated in service

160M FC Total Ports Shipped

Adapter and switch ports

*Quillin Research 2023 SAN forecast – cumulative adapter and switch ports from 1998 through 2022

FCIA
FIBRE CHANNEL INDUSTRY ASSOCIATION

Agenda

- NVMe® Overview
- NVMe/FC Architecture
- Adoption and Virtualization Use Cases
- Performance

NVMe Overview & NVMe/FC Architecture

Kamal Bakshi
Technical Director
Cisco

NVMe® Nomenclature & Trademarks

Kamal Bakshi

NVM Express® (Incorporated in March 2014)

NVM Express Logo®

NVMe®

NVMe®/PCIe® specification

NVM Express® Management Interface

NVMe-MI™

NVM Express® over Fabrics

NVMe-oF™

NVMe®/TCP protocol

NVMe®/RDMA protocol

NVMe®/IB transport

NVMe®/RoCE transport

NVMe®/iWARP transport

NVMe®/FC protocol

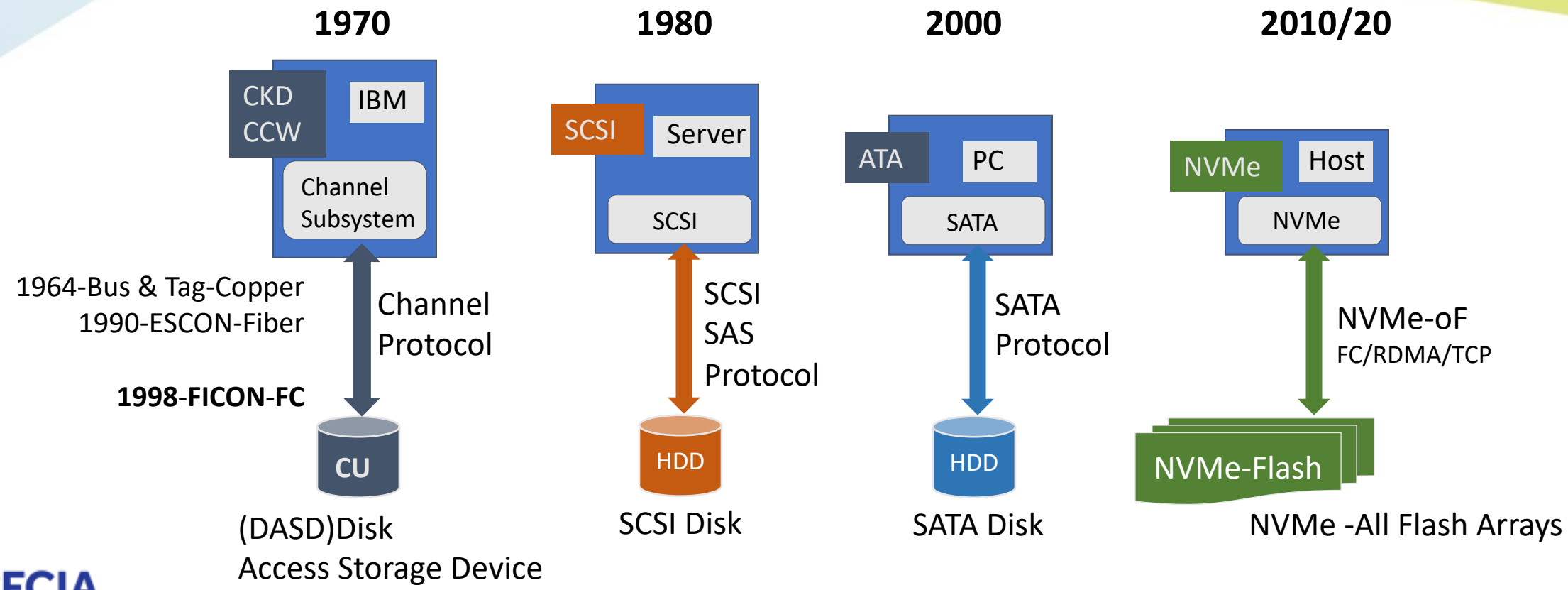
(**FC-NVMe-2**) INCITS/T11 standard

This standard describes the frame format and protocol definitions required to transfer commands and data between a NVM Express host and NVM Express subsystem using the Fibre Channel family of standards.

The appropriate registered trademark designations (™ or ®) must be applied.

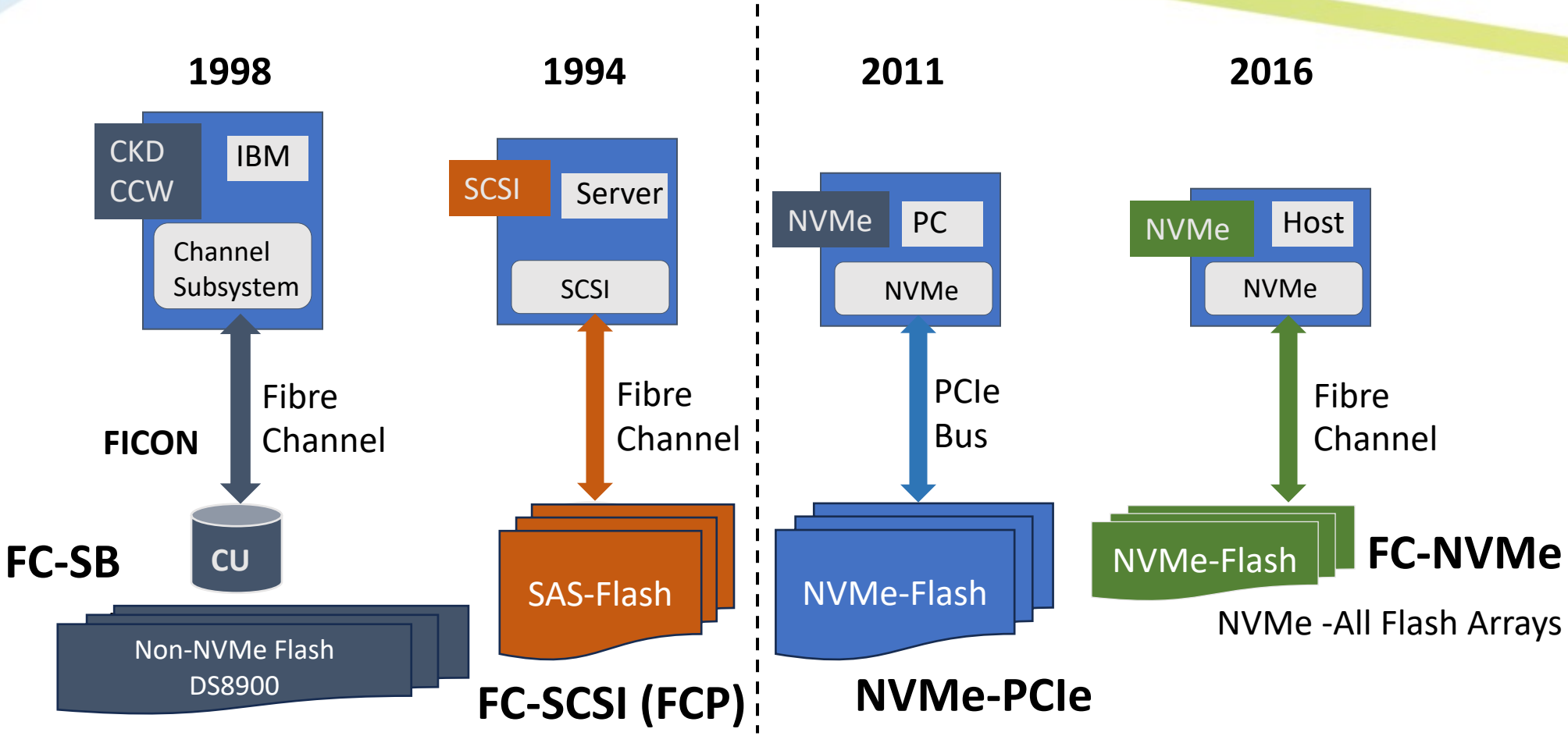
50,000 feet view of NVMe

Kamal Bakshi



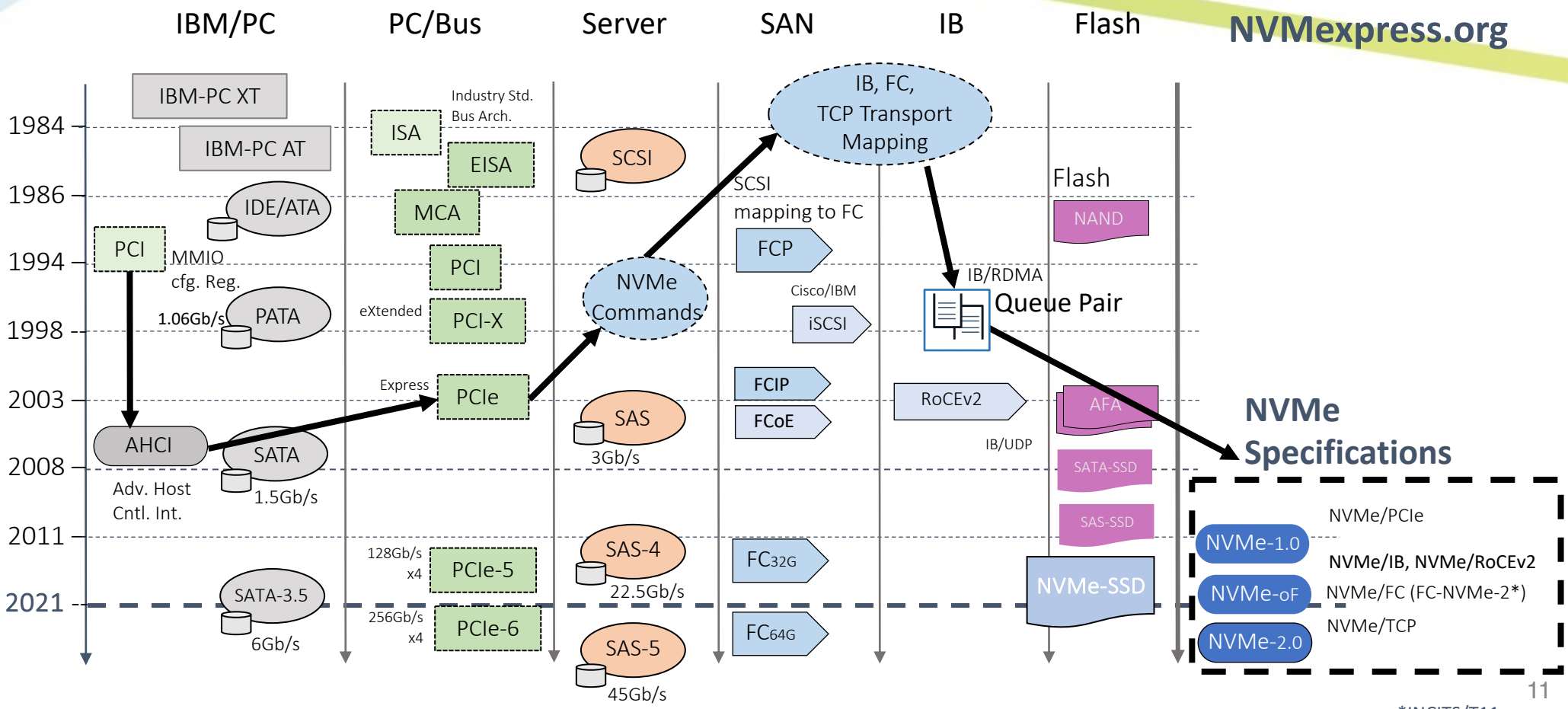
50,000 feet view of NVMe

Kamal Bakshi



NVMe 2.0 Journey

Kamal Bakshi



*INCITS/T11

FCIA
FIBRE CHANNEL INDUSTRY ASSOCIATION

NVMe 2.0 Specifications (4 Categories, 10 Docs., 1032 pages)

https://nvmexpress.org/specifications/

Kamal Bakshi

I-Mgmt.

- (9) Management Interface Specification
- (10) NVMe Network Boot Specification

II-Base Specification

(1) NVMe Base Specification + Admin Cmd, Fabric Cmd, Common I/O Cmd

III-NVMe Specific I/O command sets

- (2) NVM Command Sets
- (3) Zoned NS Command Sets
- (4) Key Value Command Sets

IV-Transport Mapping Specifications (PCIe, IB, Enet, FC)

Message/Capsule Based Model

Shared Memory Based Model

- (5) NVMe/PCIe
- (6) RDMA Spec.
  - NVMe/IB
  - NVMe/RoCEv2

Fibre Channel

- (7) FC-NVMe-2 (T11)
  - *+ FC-FS-6 (T11)*
- (8) NVMe/TCP

(NVMe/TCP)
Ethernet
- -iSCSI
- -FCoE
- -FCIP
- -IPFC
- -iFCP

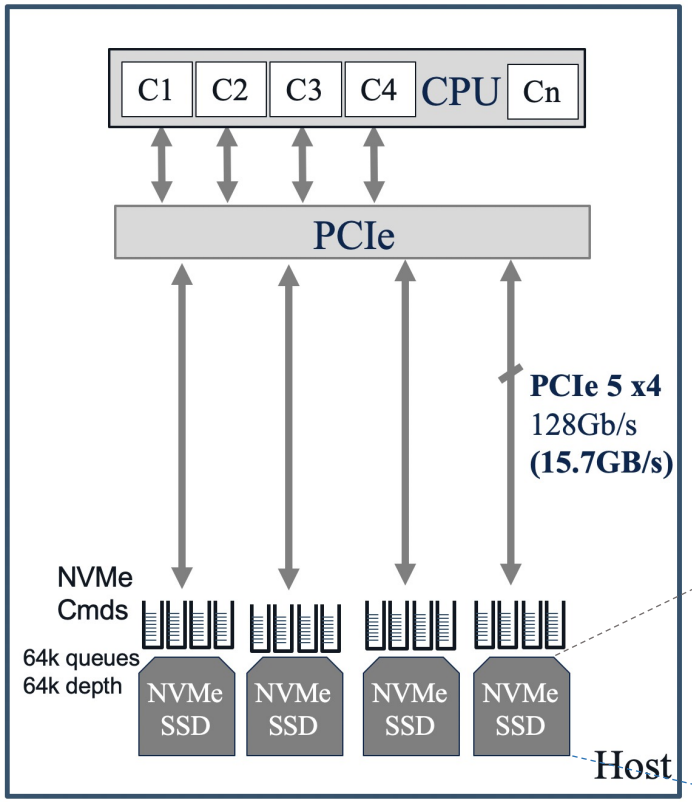
(NVMe/IB, RoCEv2)
RDMA
- -SRP
- -iSER

(NVMe/FC)
Fibre Channel
- -FCP/SCSI

FCIA FIBRE CHANNEL INDUSTRY ASSOCIATION

NVMe SSD (Local Storage)

Kamal Bakshi



NVMe

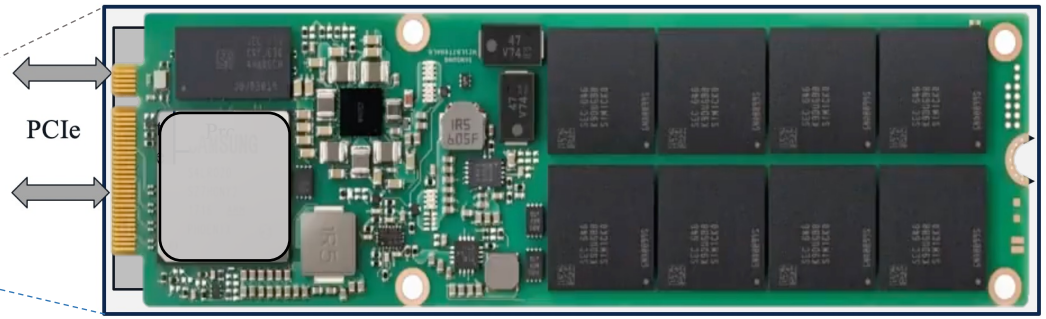
-New Block Storage Protocol for Flash
-Maps directly into PCIe
-Replaces SCSI commands
-Transport mapping for RDMA/FC/TCP

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

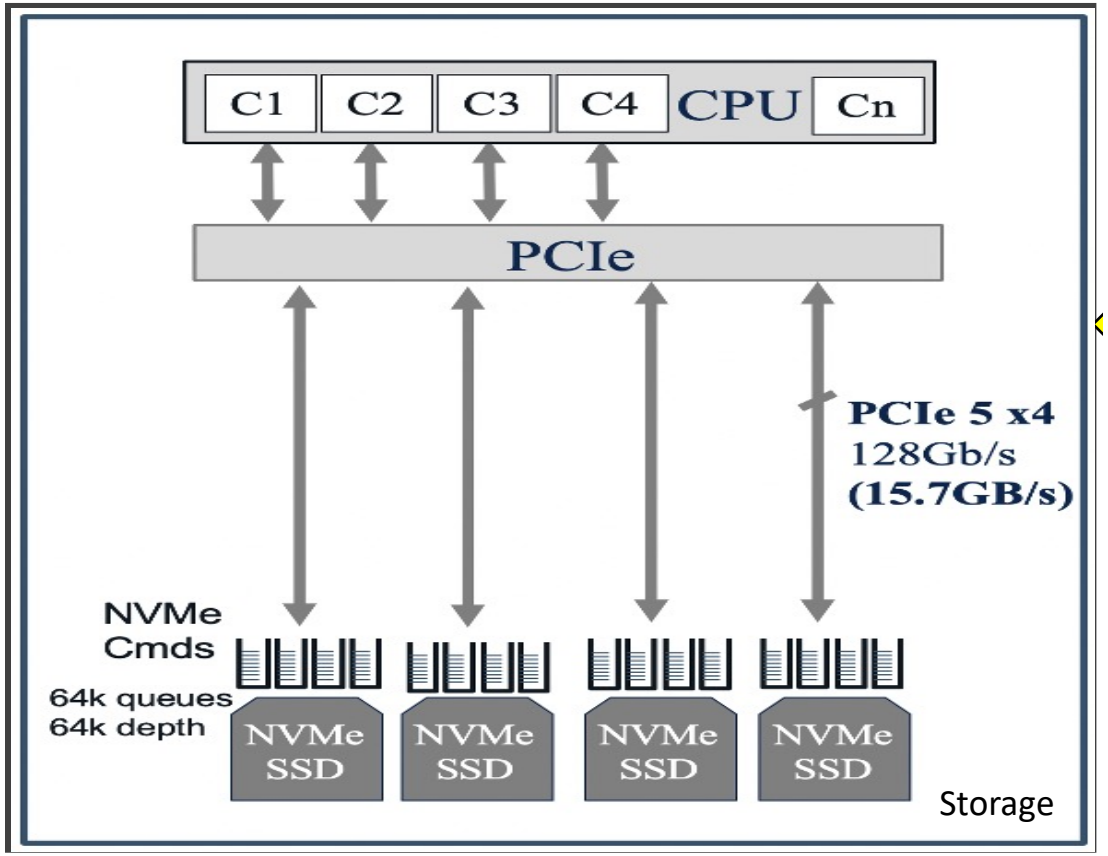
NVMe/PCIe SSD



M.2 form factor

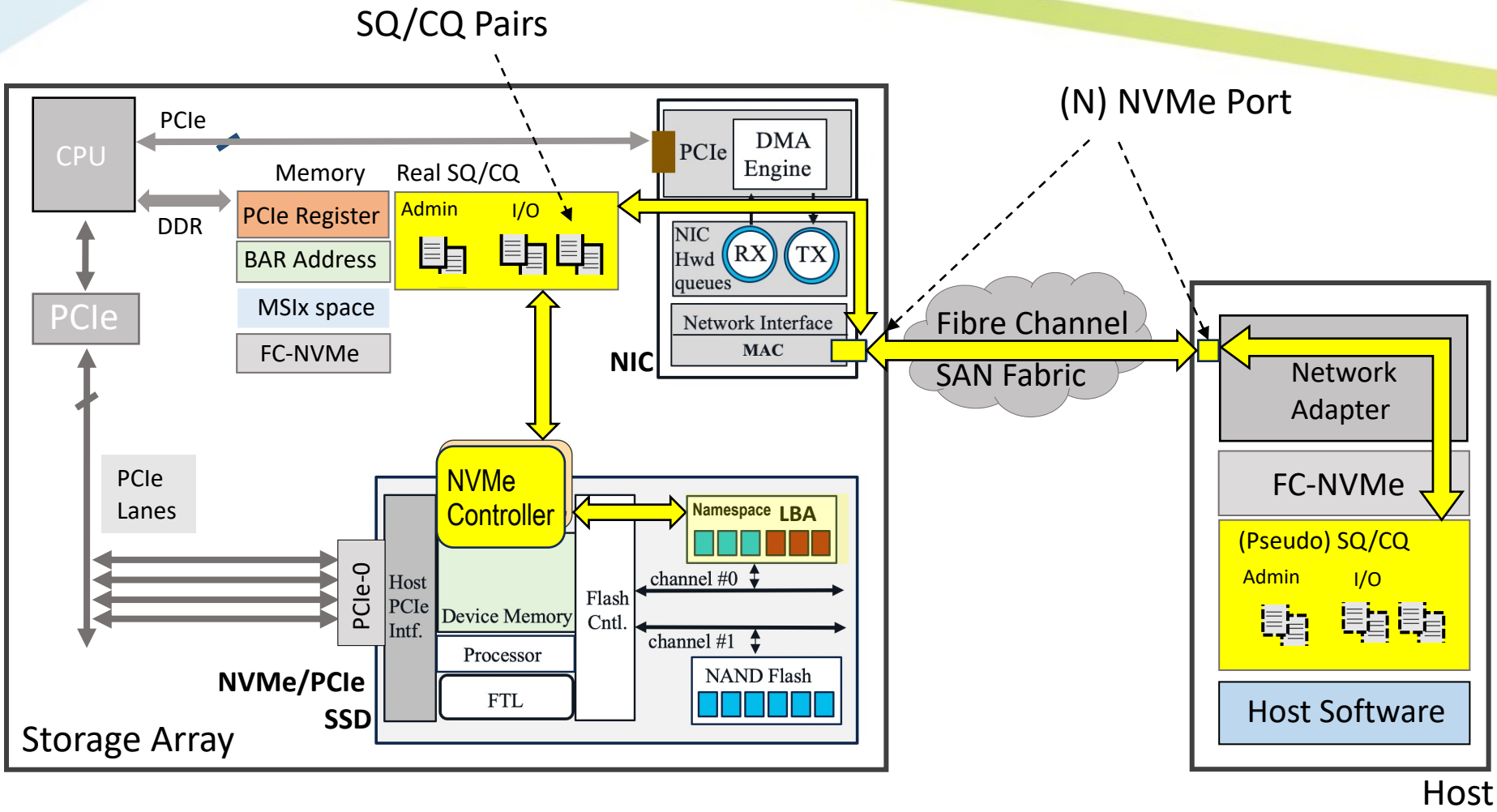
FC-NVMe (Remote Storage)

Kamal Bakshi



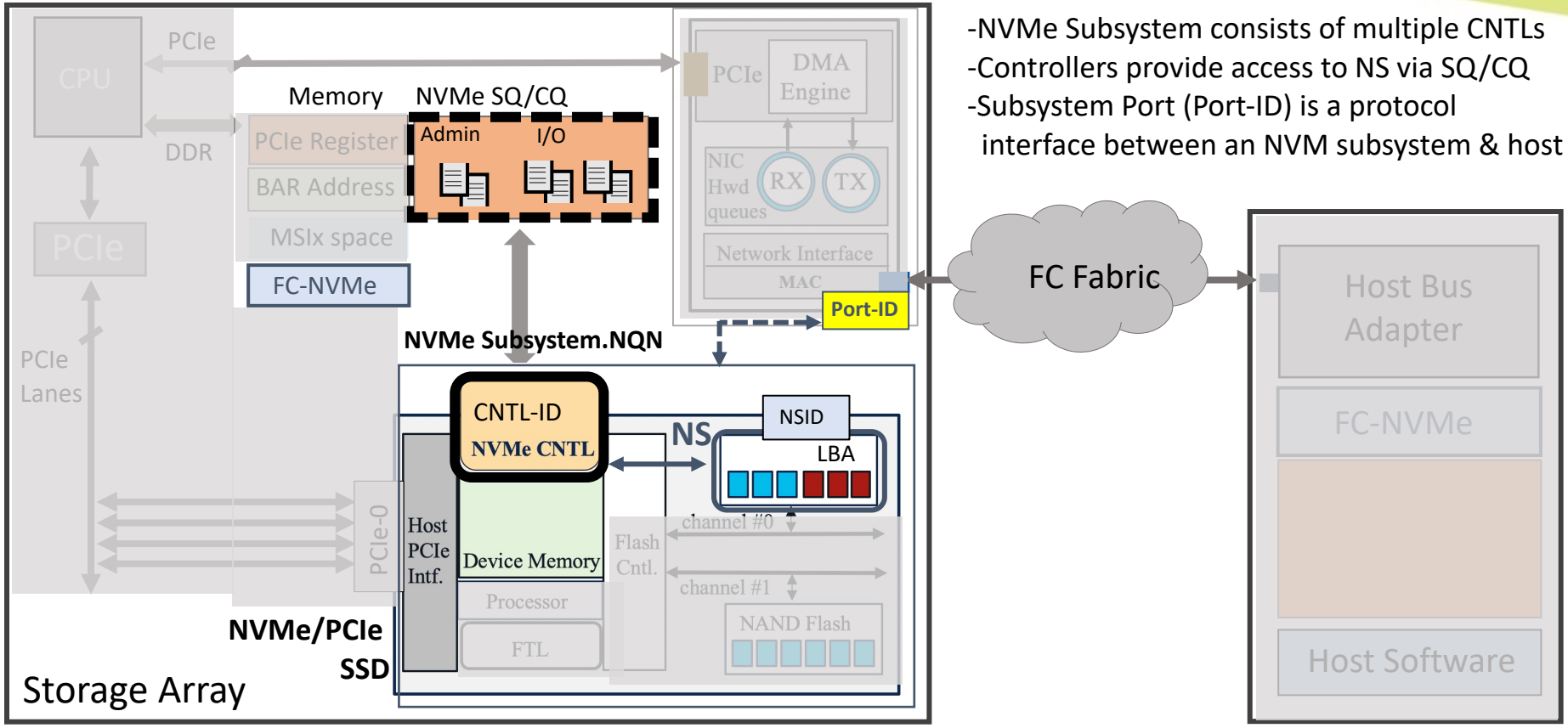
FC-NVMe (Remote Storage)

Kamal Bakshi



FC-NVMe (NVMe Subsystem)

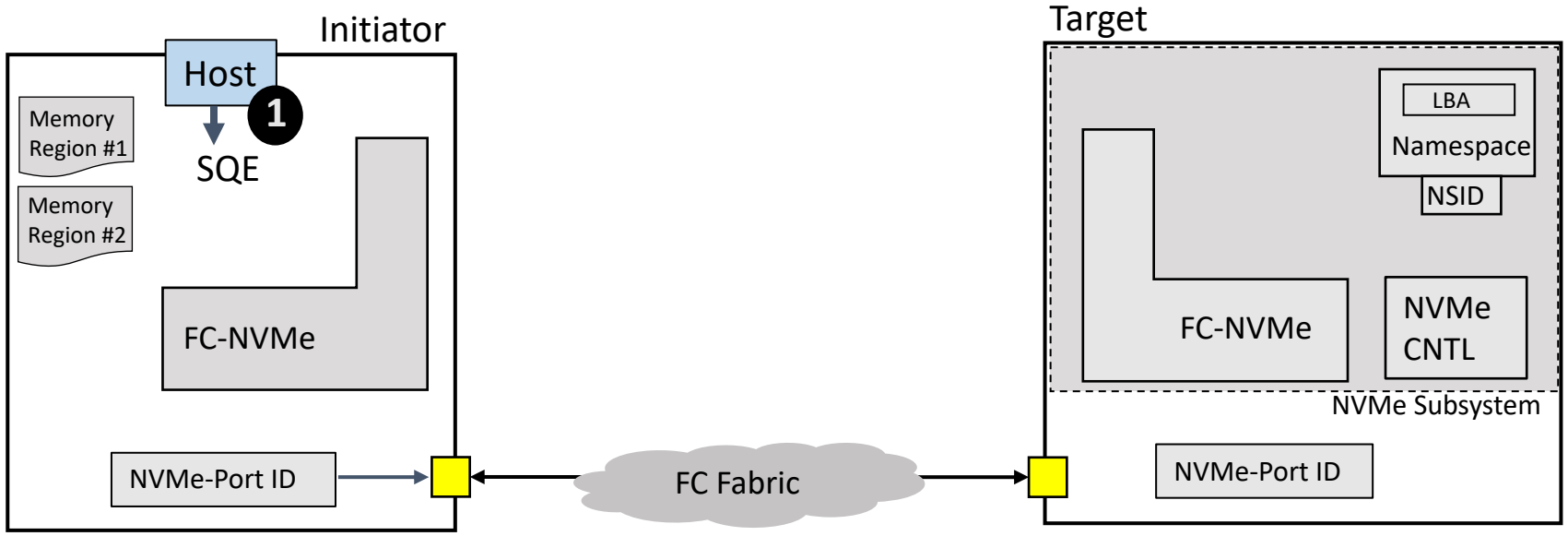
Kamal Bakshi



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

FC-NVMe (FC Mapping Abstractions)

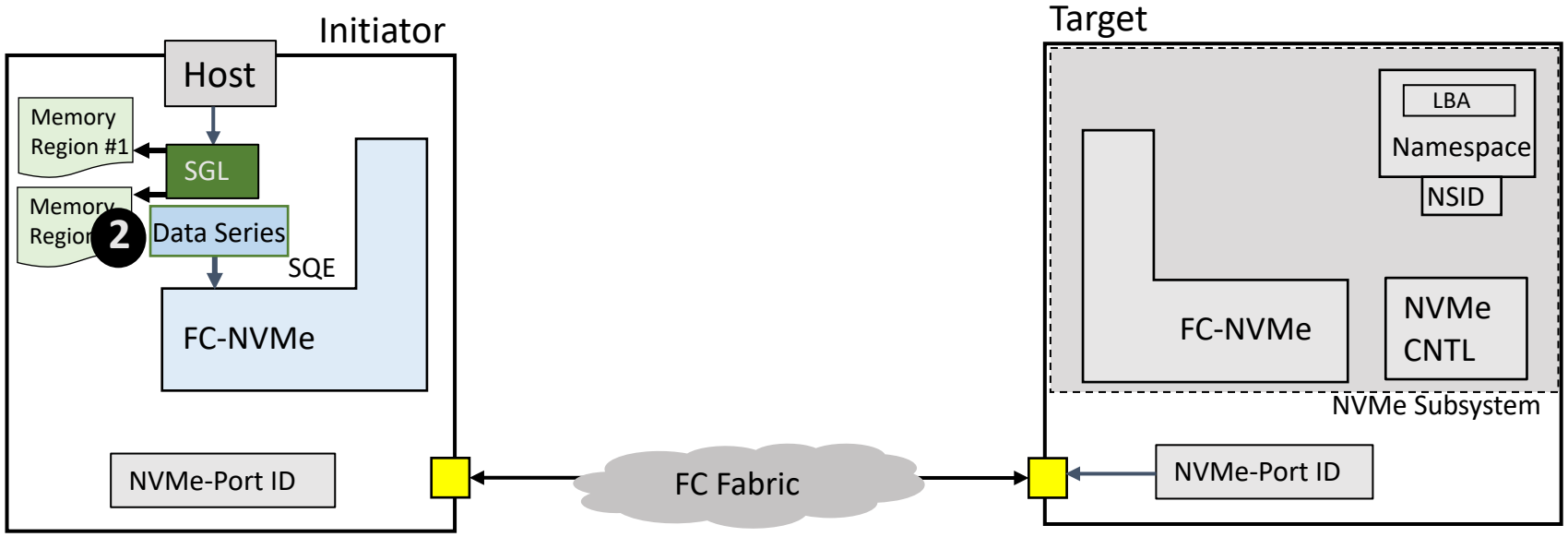
Kamal Bakshi



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

FC-NVMe (FC Mapping Abstractions)

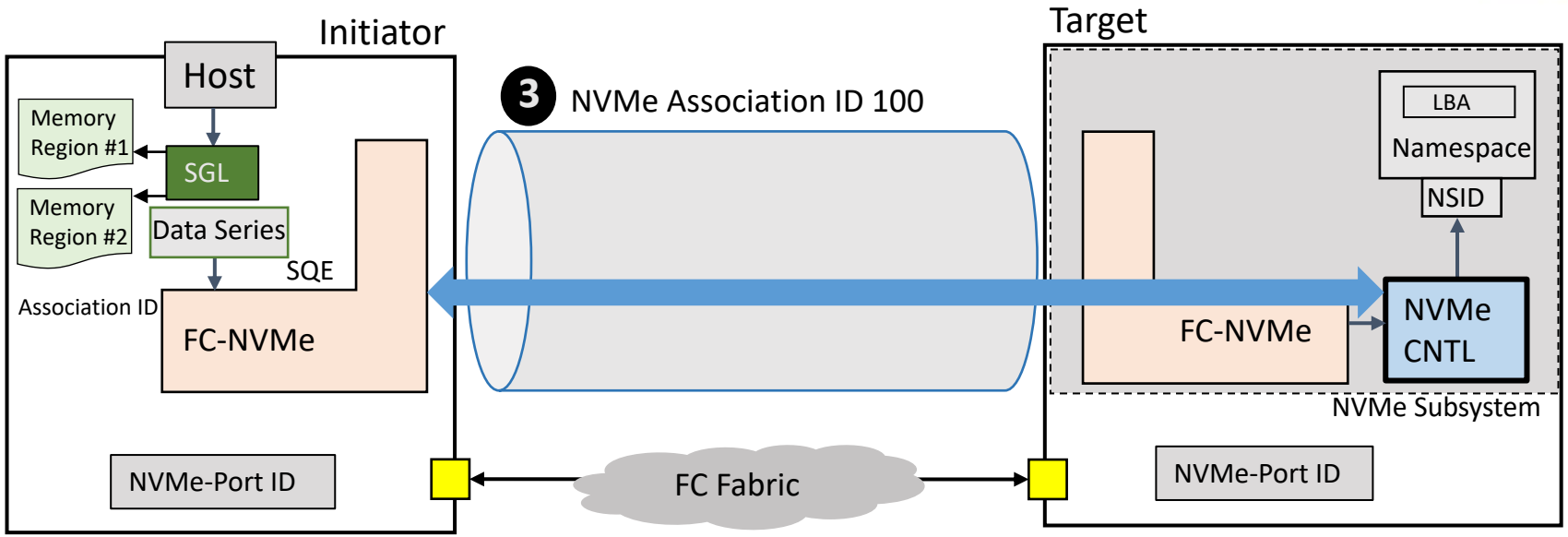
Kamal Bakshi



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

FC-NVMe (FC Mapping Abstractions) -Association ID

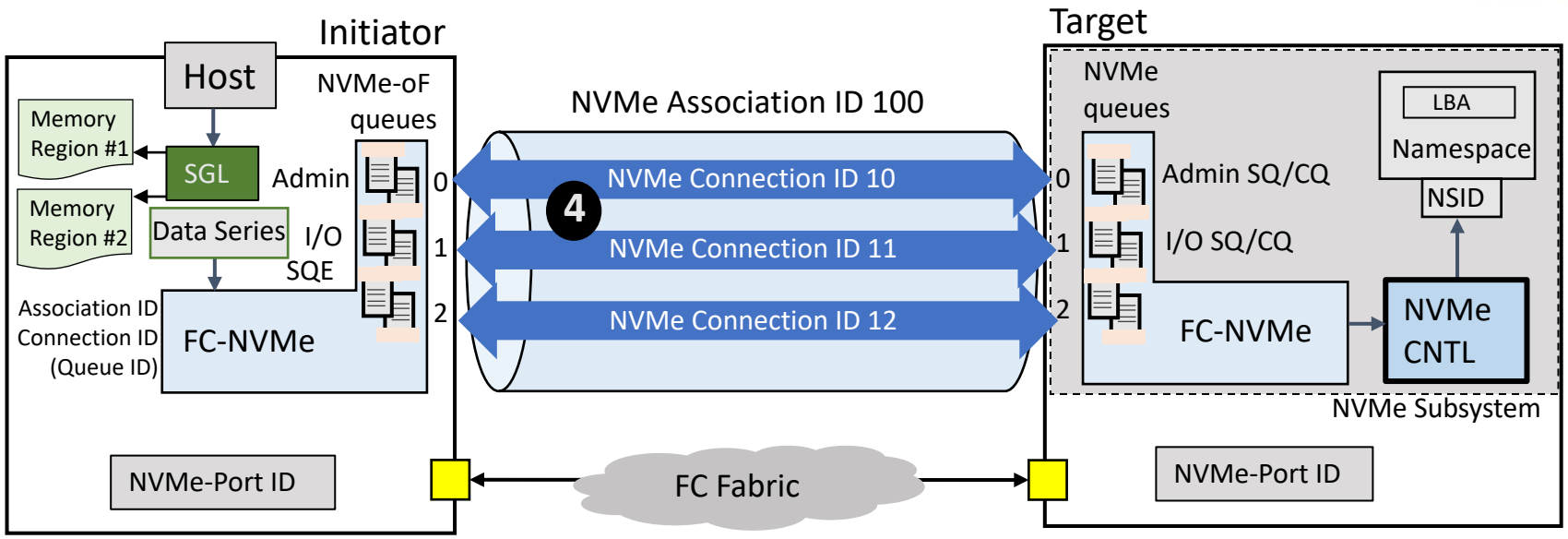
Kamal Bakshi



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

FC-NVMe (FC Mapping Abstractions) -Connection IDs

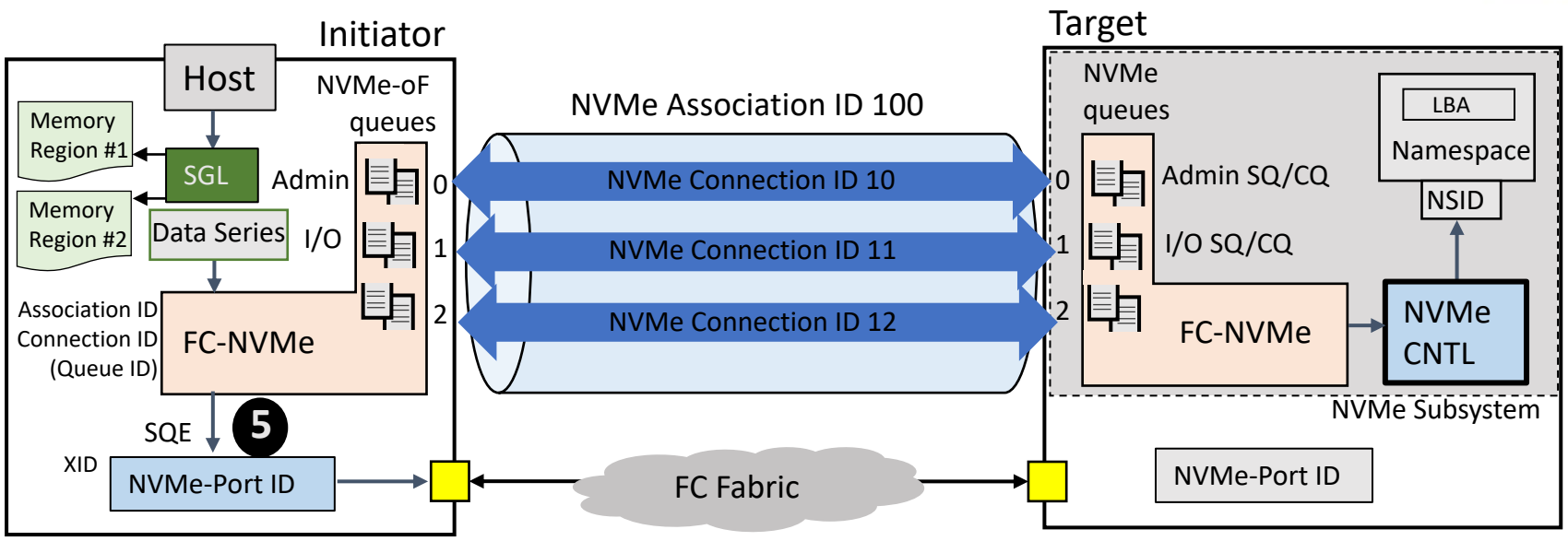
Kamal Bakshi



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.

FC-NVMe (FC Mapping Abstractions) -Exchange IDs

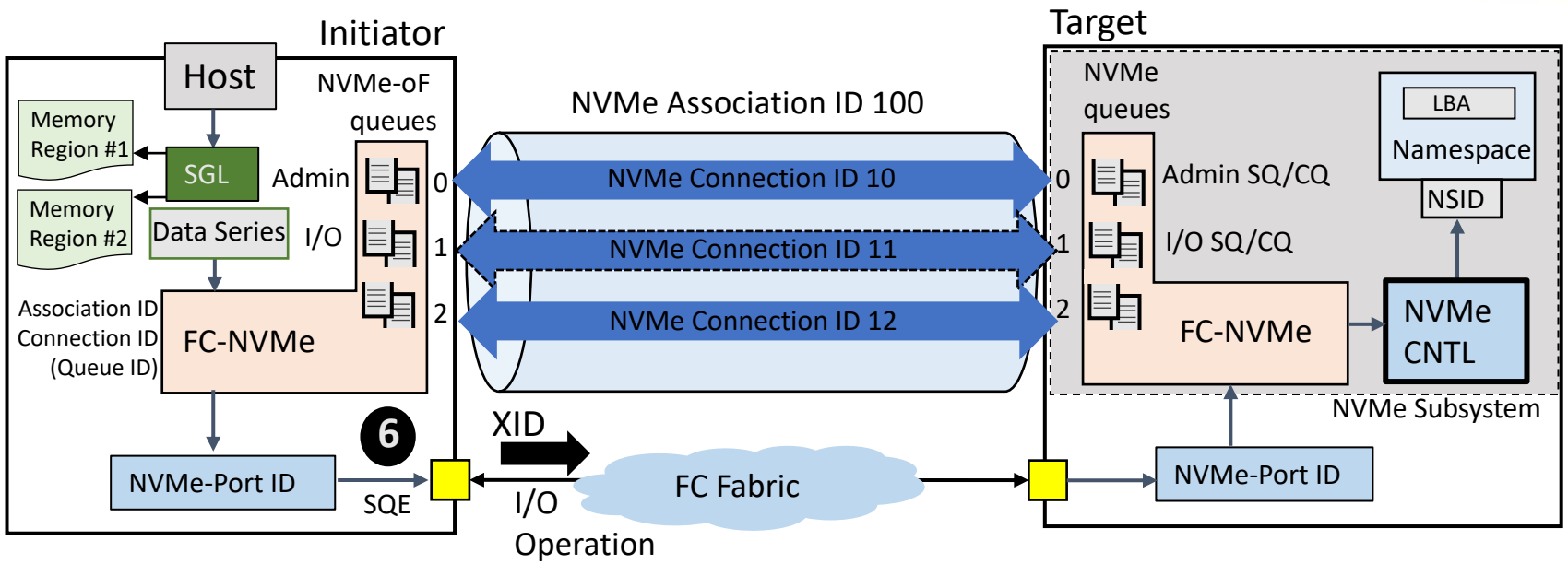
Kamal Bakshi



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.

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

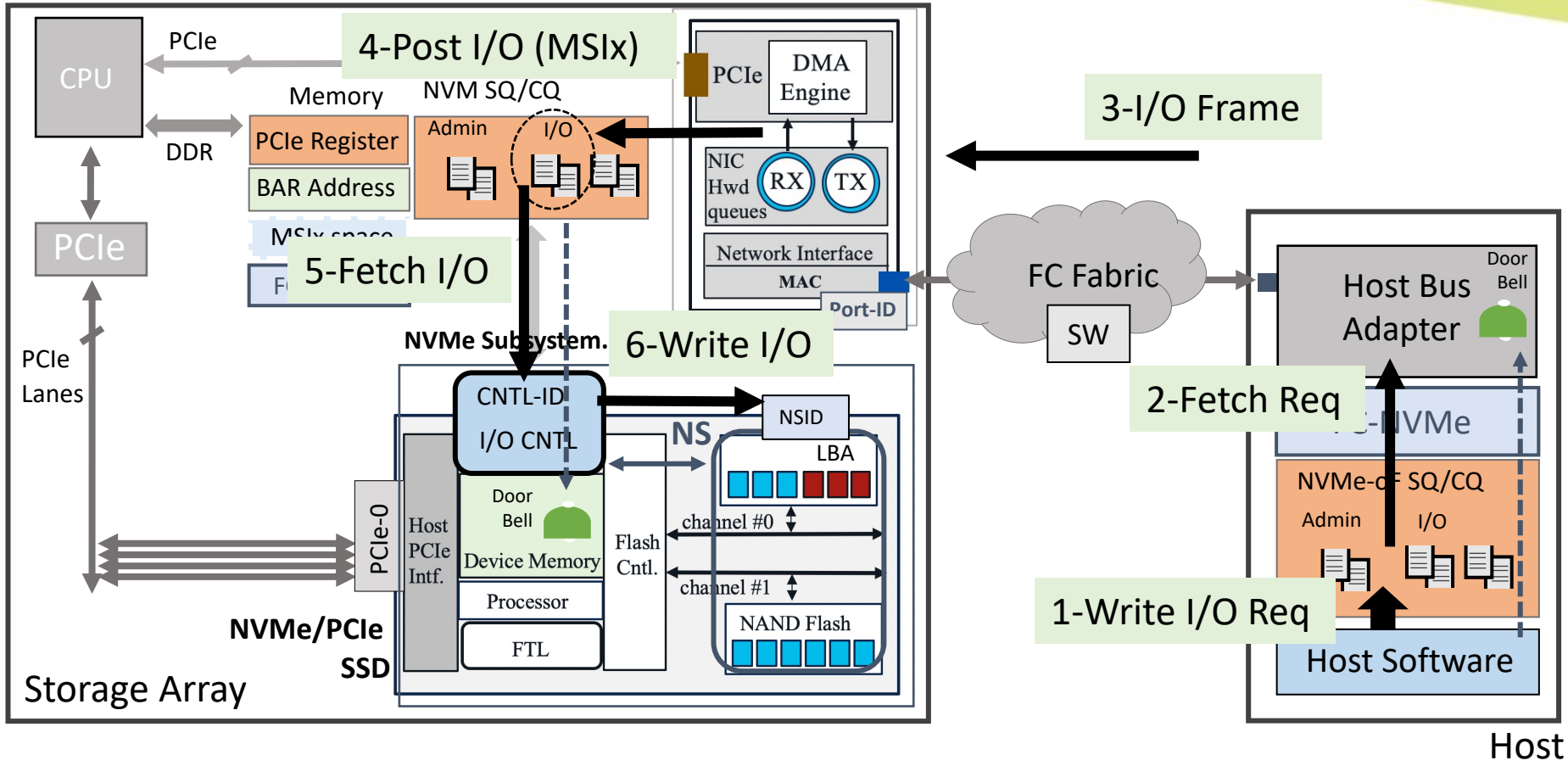
Kamal Bakshi



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

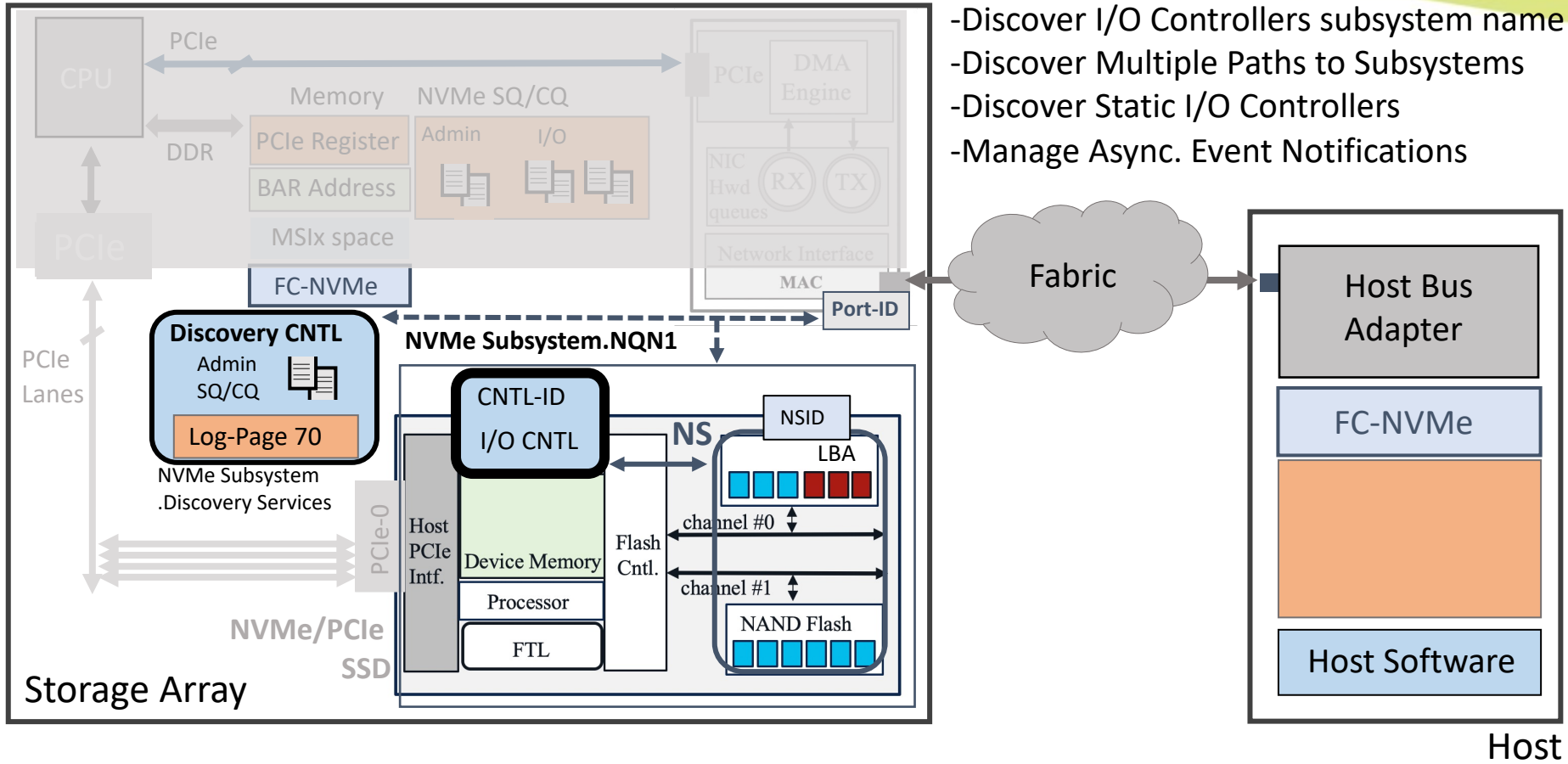
FC-NVMe (HBA/MSIx Interrupts)

Kamal Bakshi



FC-NVMe (Discovery Services Subsystem)

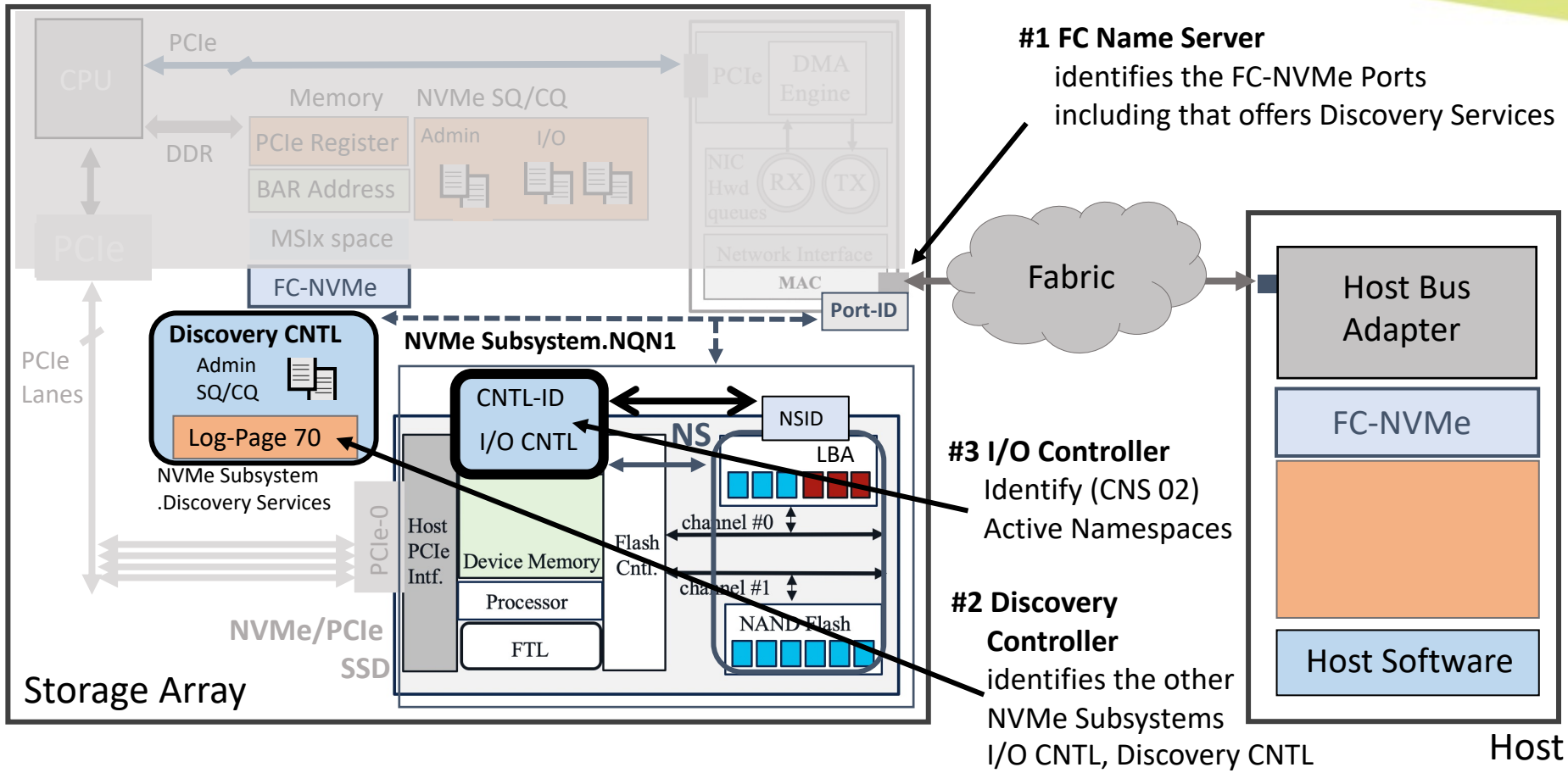
Kamal Bakshi



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

FC-NVMe (Discovery Services Subsystem)

Kamal Bakshi



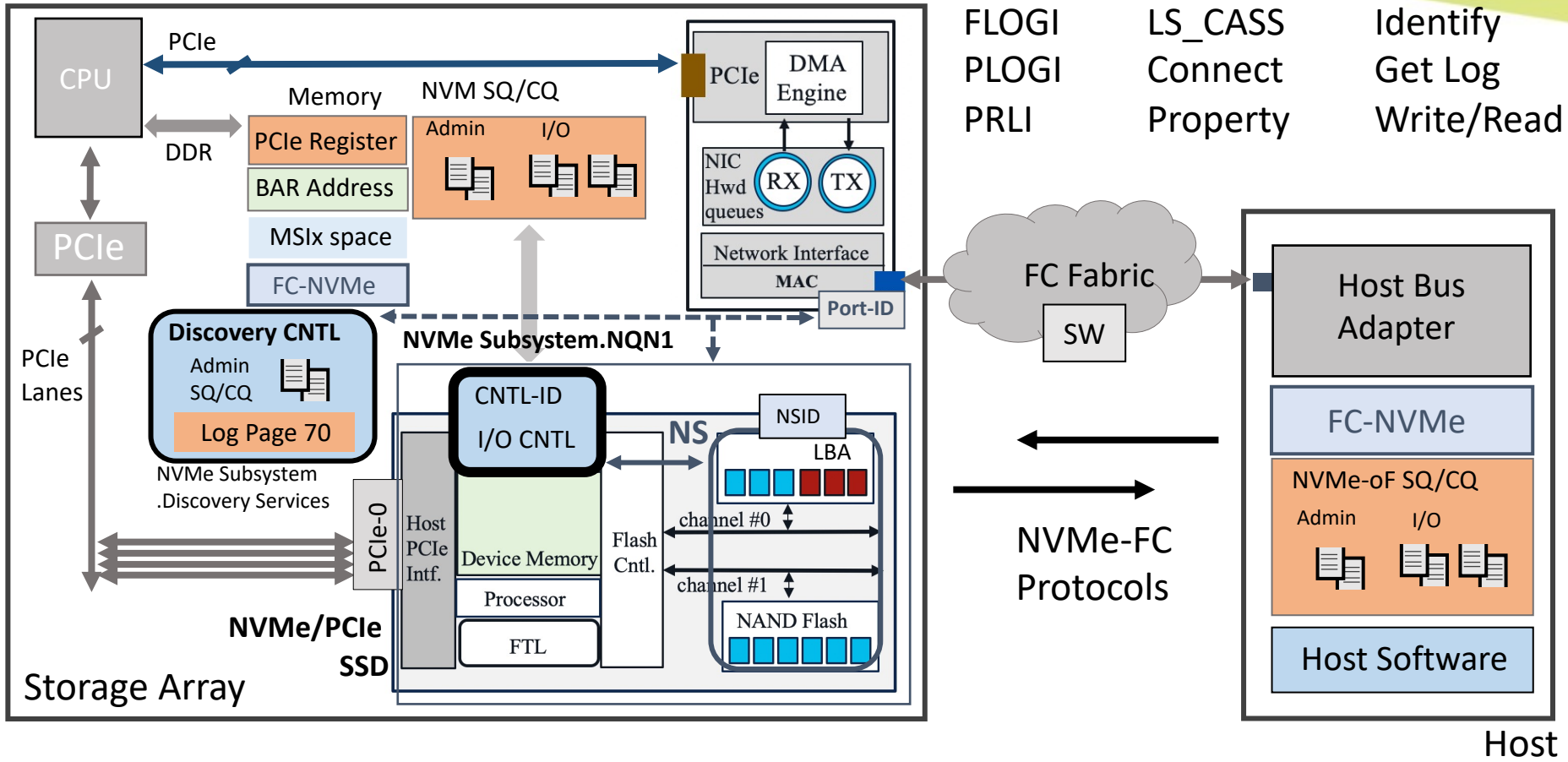
#1 FC Name Server
identifies the FC-NVMe Ports including that offers Discovery Services

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

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

FC-NVMe Protocol Flows

Kamal Bakshi



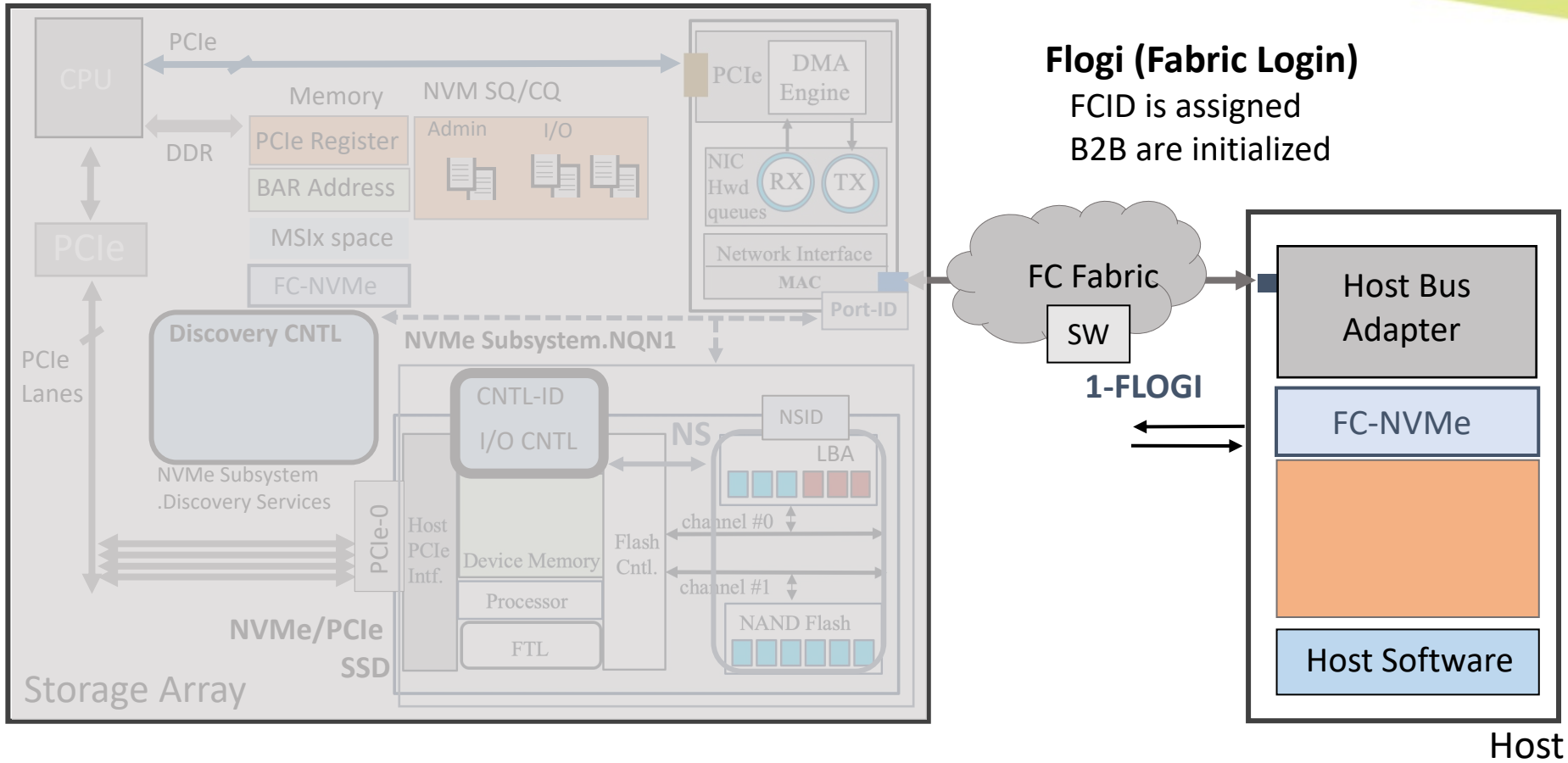
FLOGI
PLOGI
PRLI

LS_CASS
Connect
Property

Identify
Get Log
Write/Read

FC-NVMe Protocol Flows (FLOGI)

Kamal Bakshi



Flogi (Fabric Login)

FCID is assigned

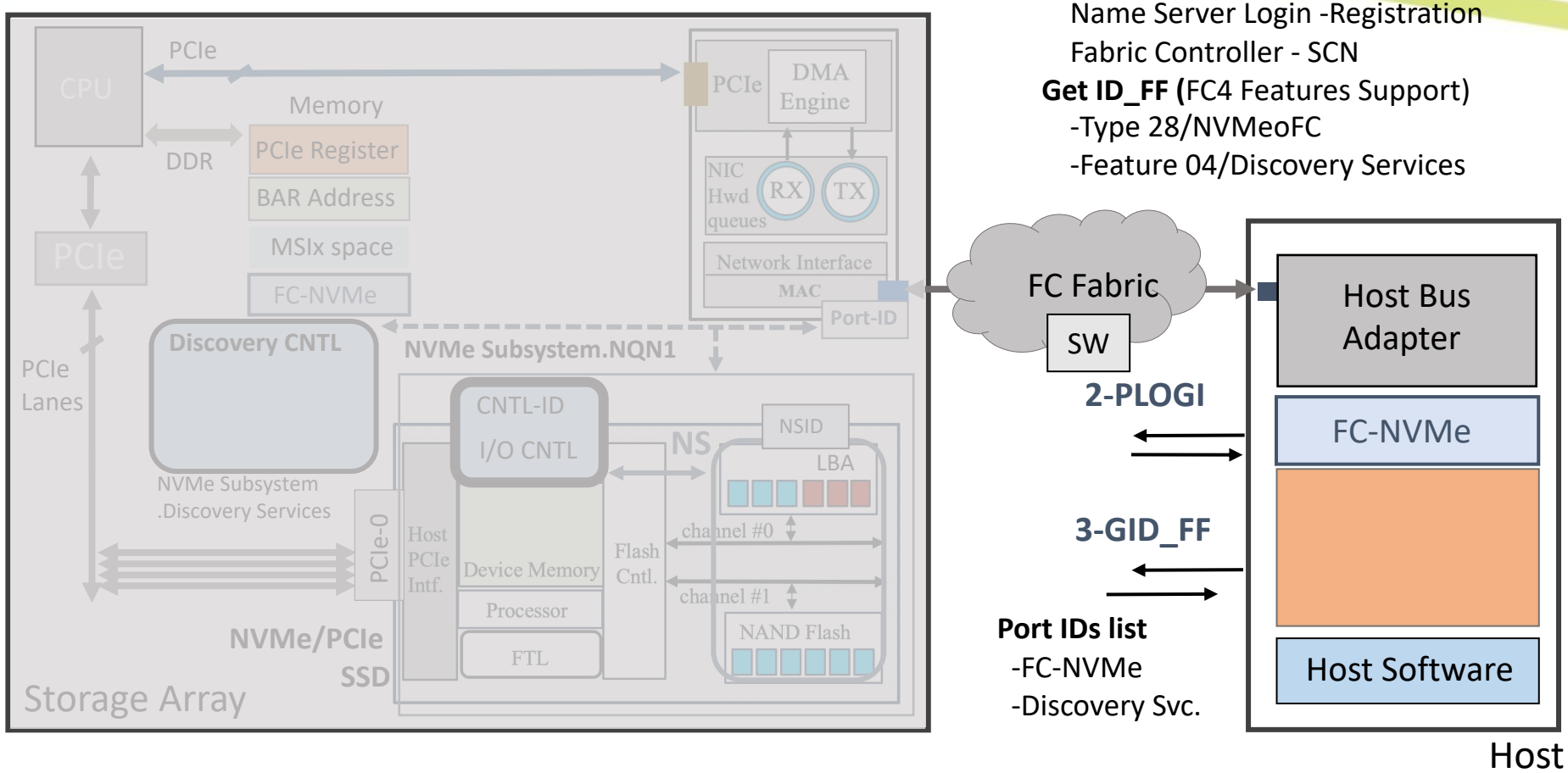
B2B are initialized

FC-NVMe Protocol Flows (PLOGI)

Kamal Bakshi

Port Logins
- Name Server Login -Registration
- Fabric Controller - SCN

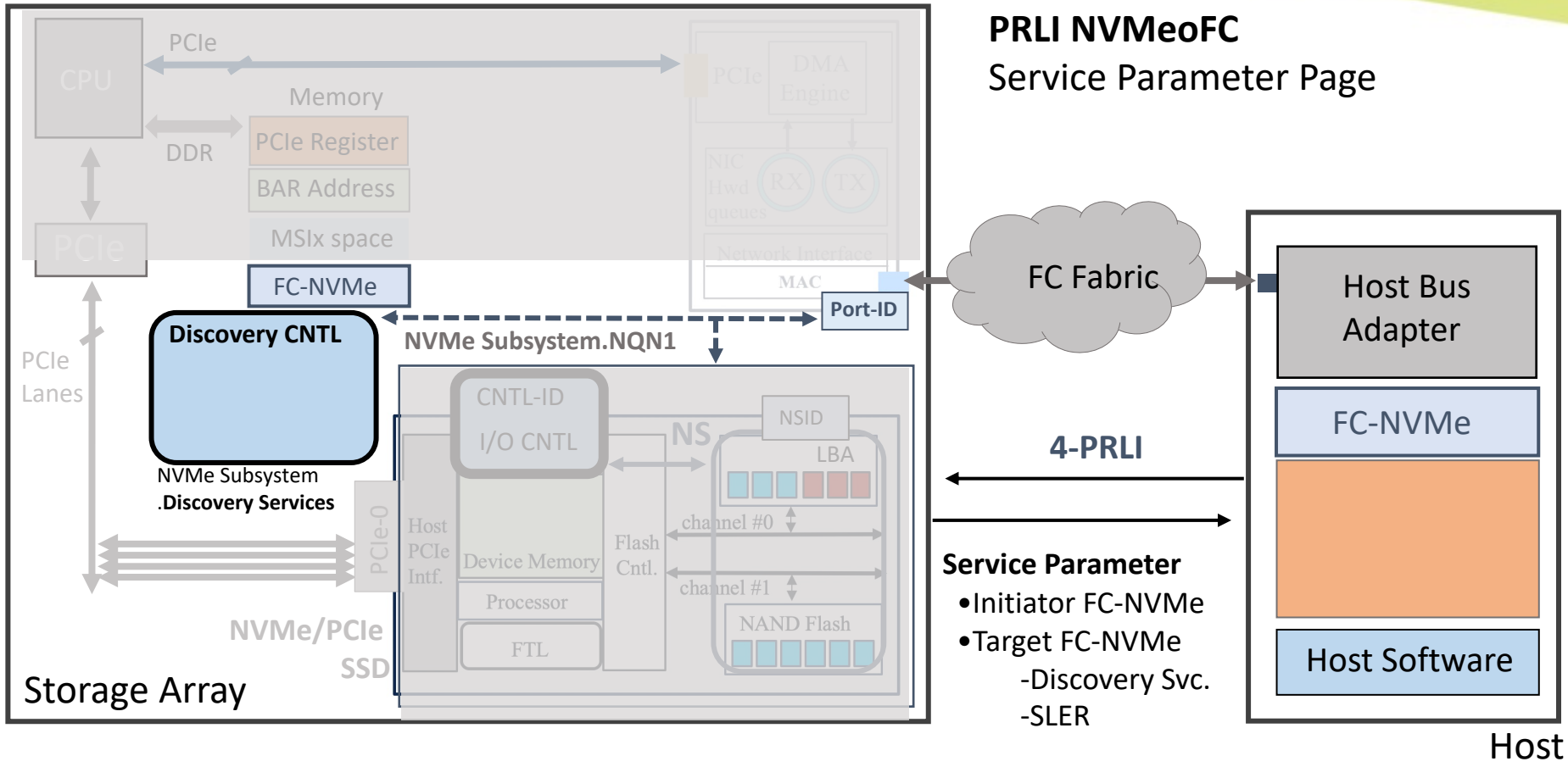
Get ID_FF (FC4 Features Support)
- -Type 28/NVMeoFC
- -Feature 04/Discovery Services



Port IDs list
- -FC-NVMe
- -Discovery Svc.

FC-NVMe Protocol Flows (PRLI)

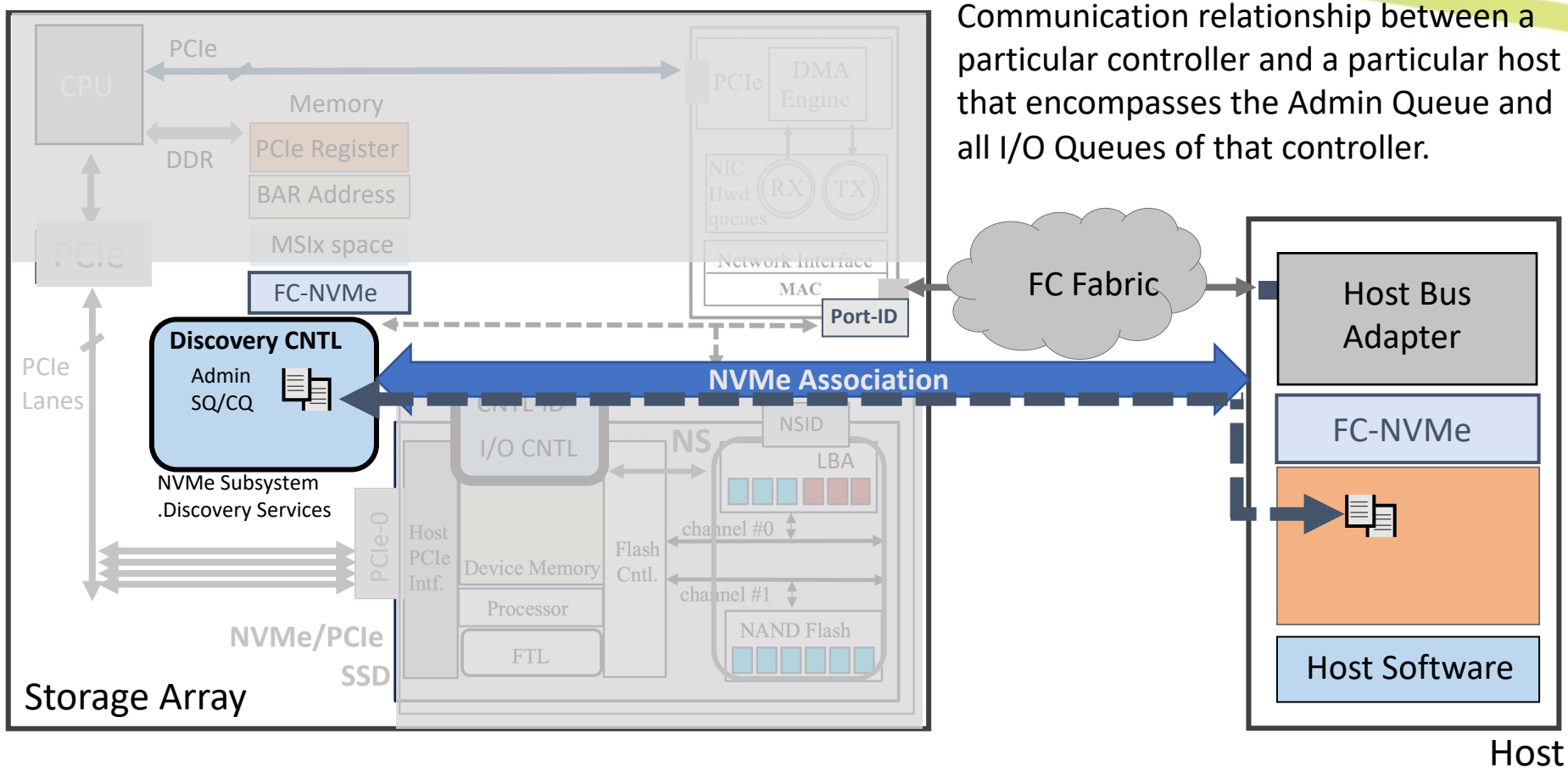
Kamal Bakshi



FC-NVMe Protocol Flows (NVMe Association)

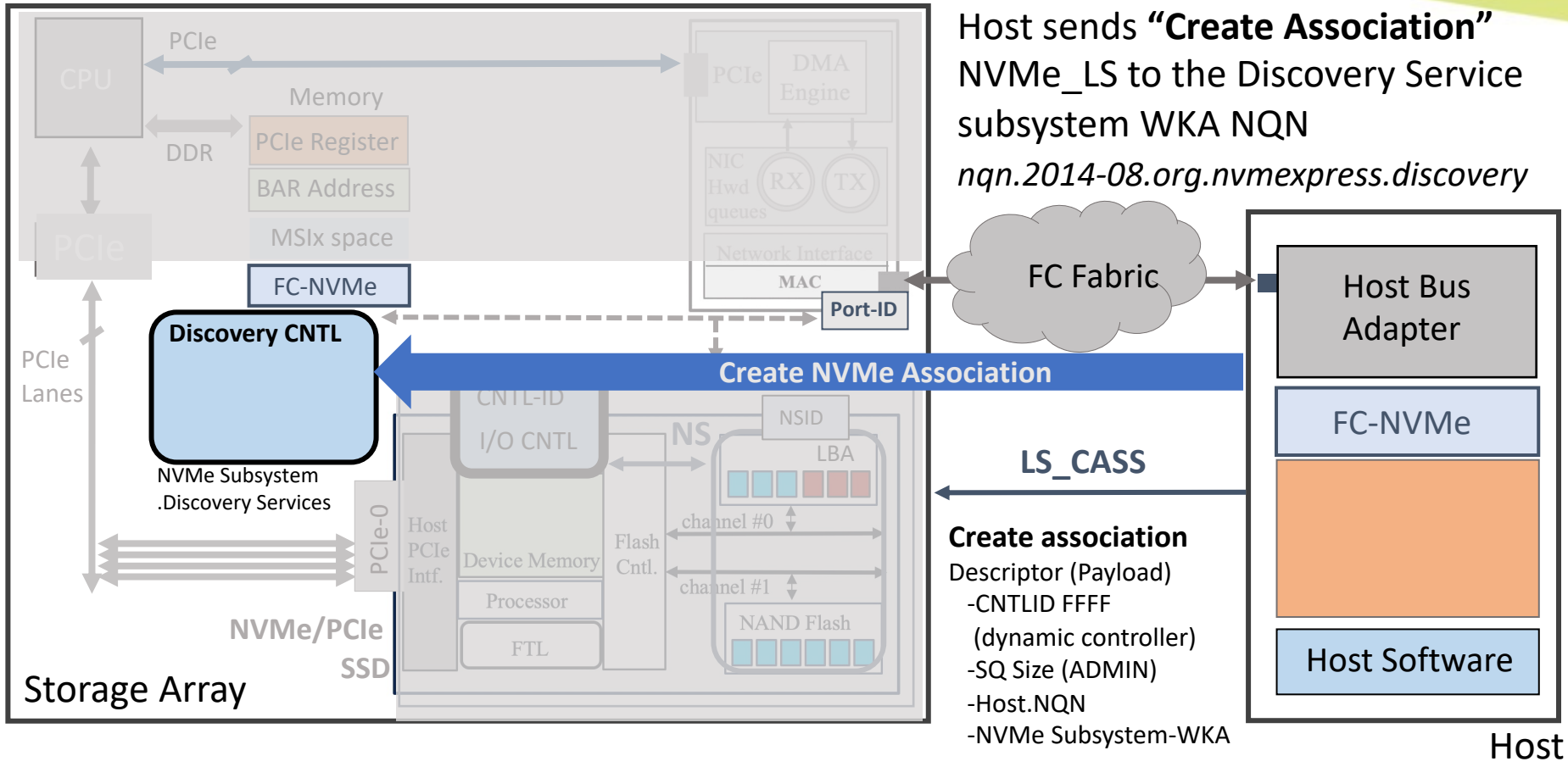
Kamal Bakshi

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



FC-NVMe Protocol Flows (LS _CASS)

Kamal Bakshi



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

FC-NVMe (Create Association)

Kamal Bakshi

```
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)

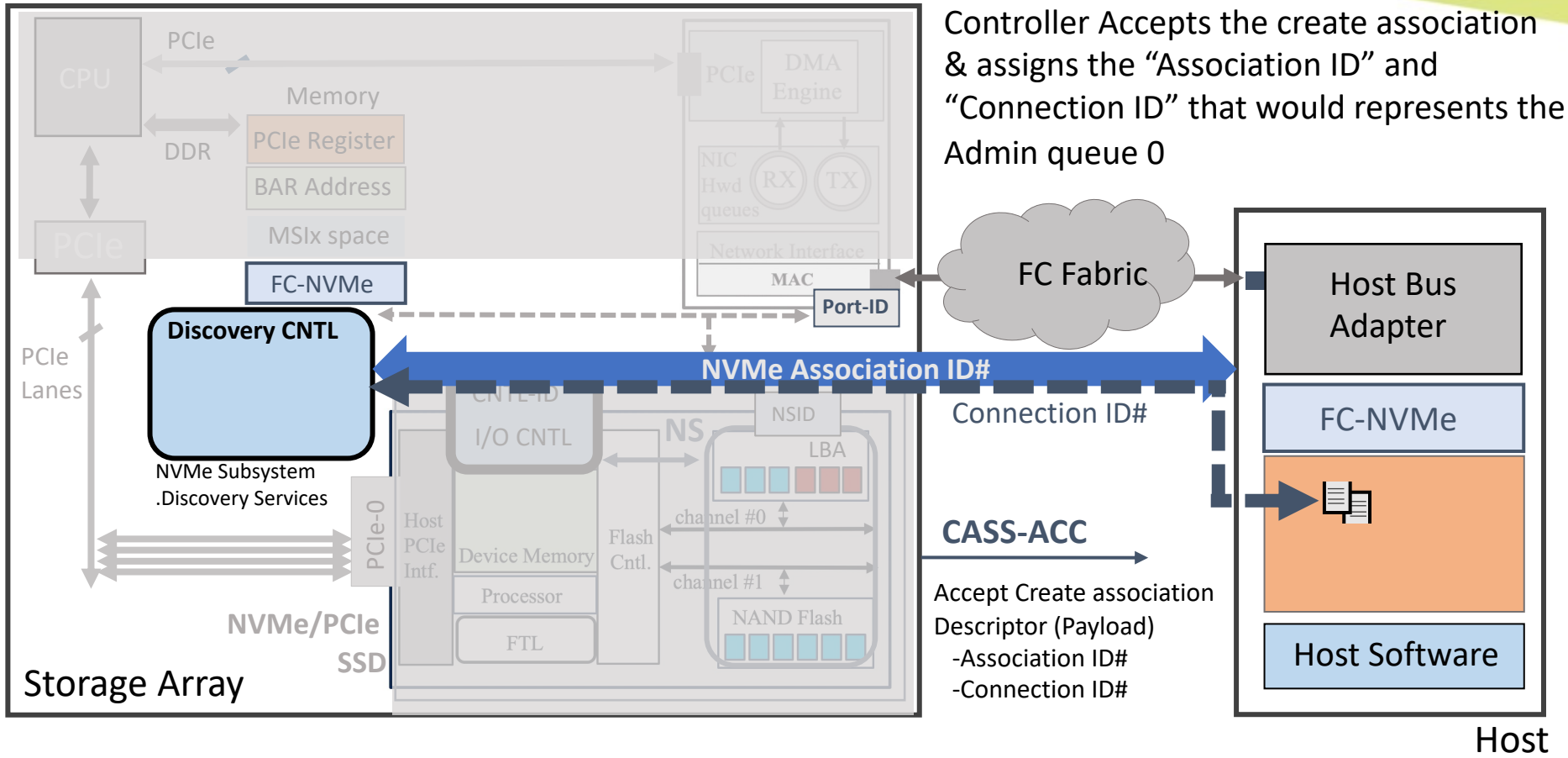
Admin Queue
depth

Host NQN

FC-NVMe Protocol Flows (LS CASS_ACC)

Kamal Bakshi

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



CASS-ACC

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

FC-NVMe (Accept Create Association)

Kamal Bakshi

```
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

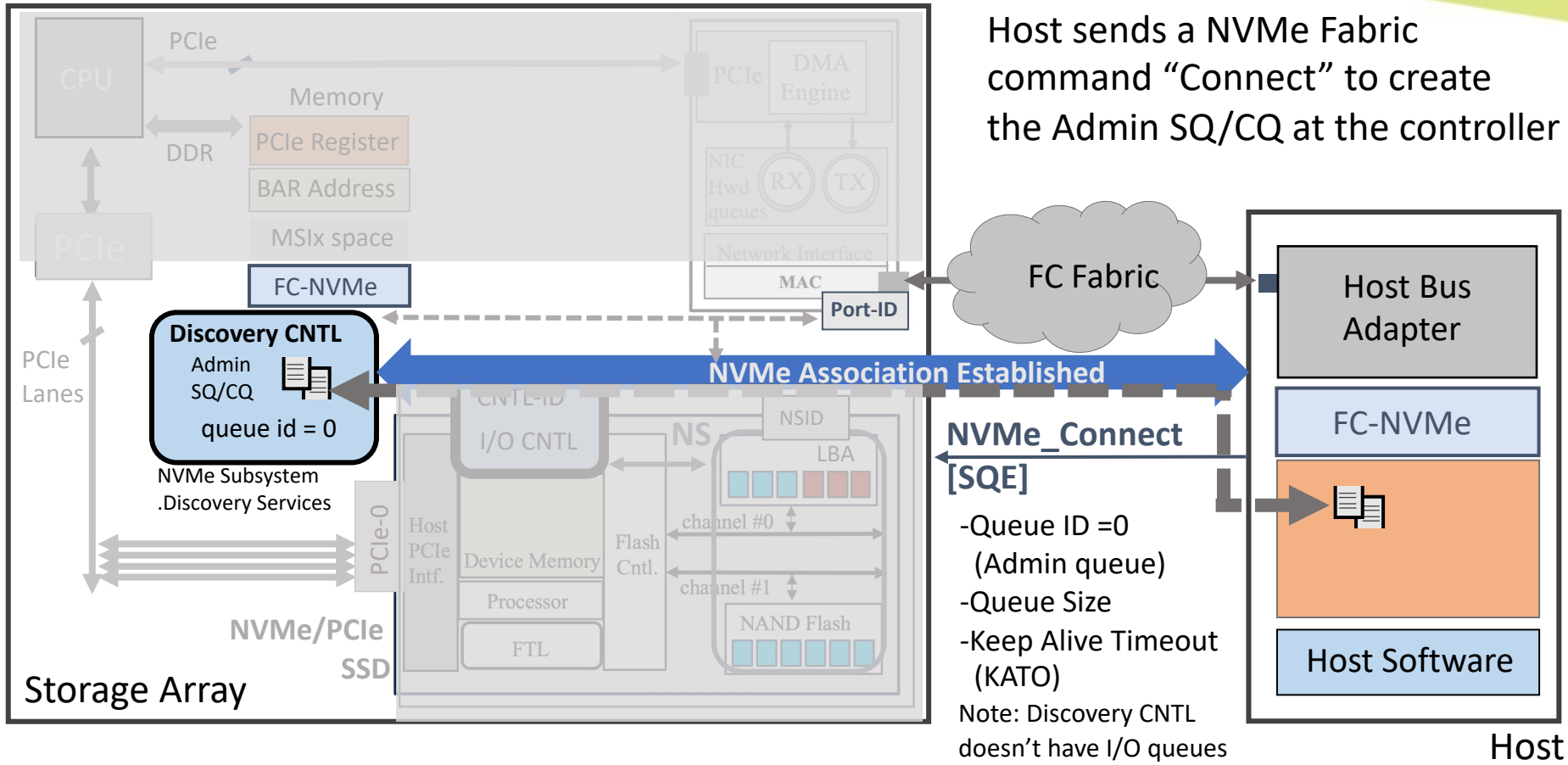
NVMe Association ID

NVMe Connection ID

FC-NVMe Protocol Flows (Connect Command SQE)

Kamal Bakshi

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



FC-NVMe (Connect)

Kamal Bakshi

```
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 descriptor; 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

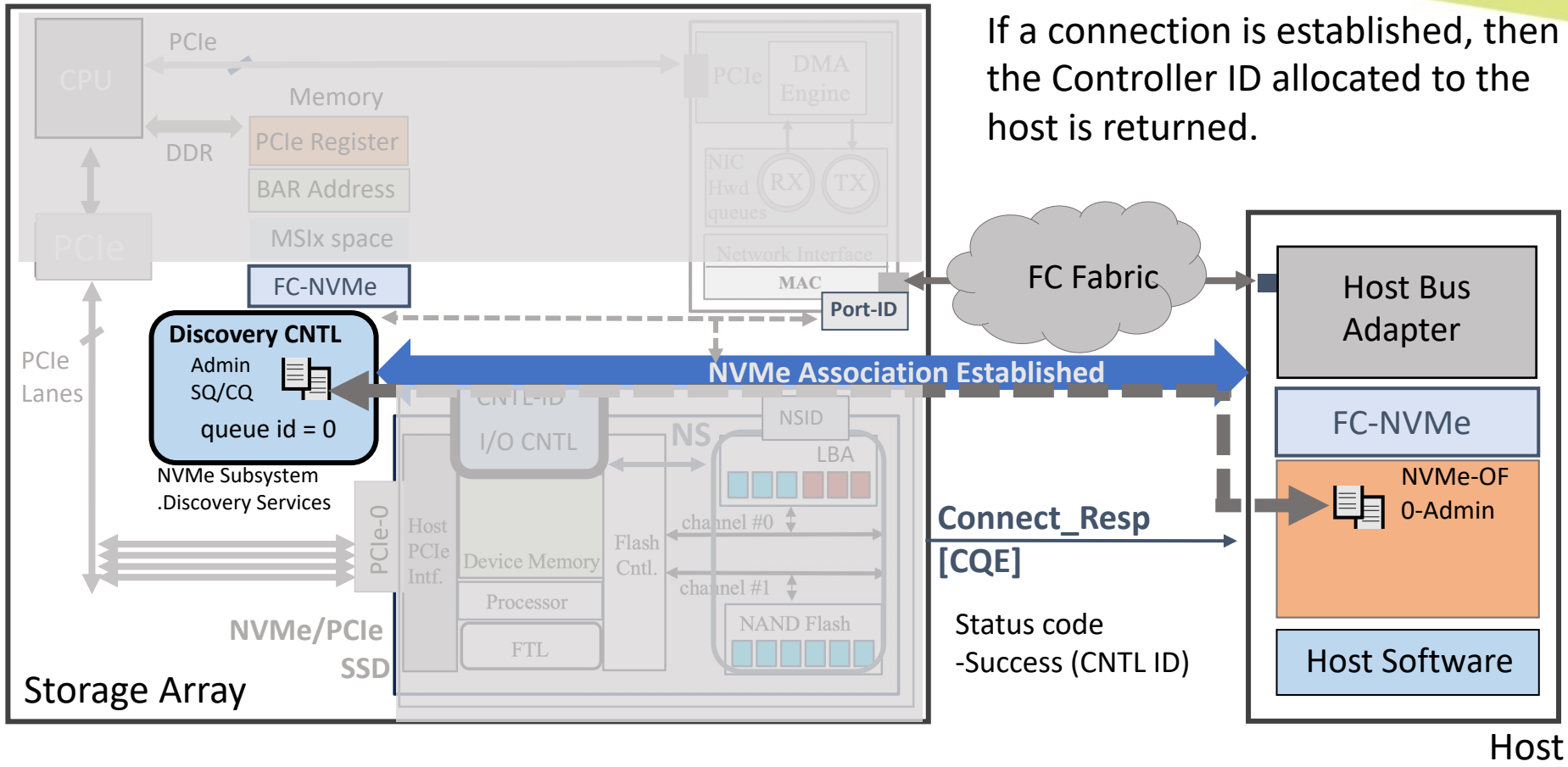
default queue
size = 32

Queue ID = 0 (Admin)

FC-NVMe Protocol Flows (Connect Response CQE)

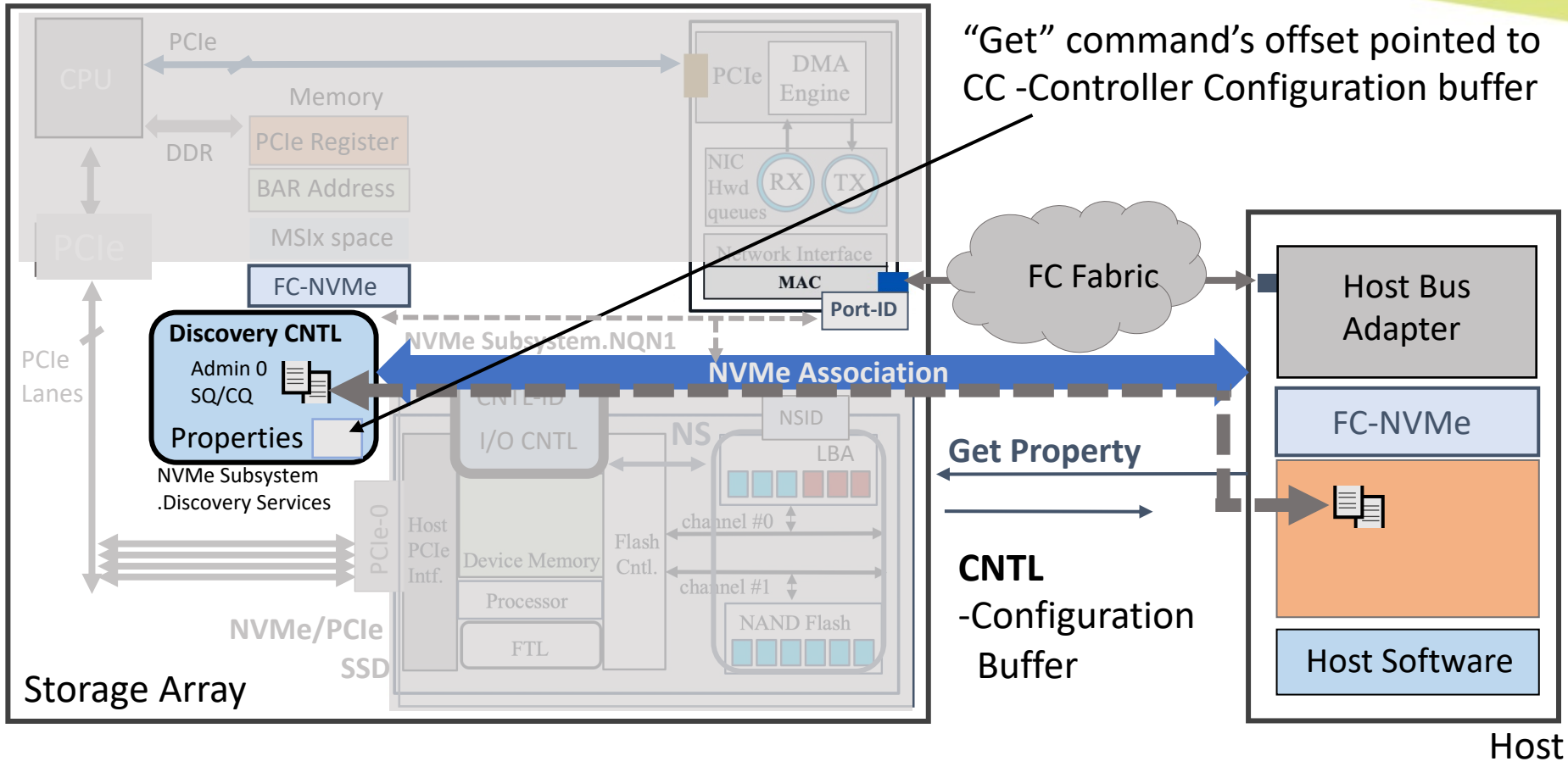
Kamal Bakshi

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



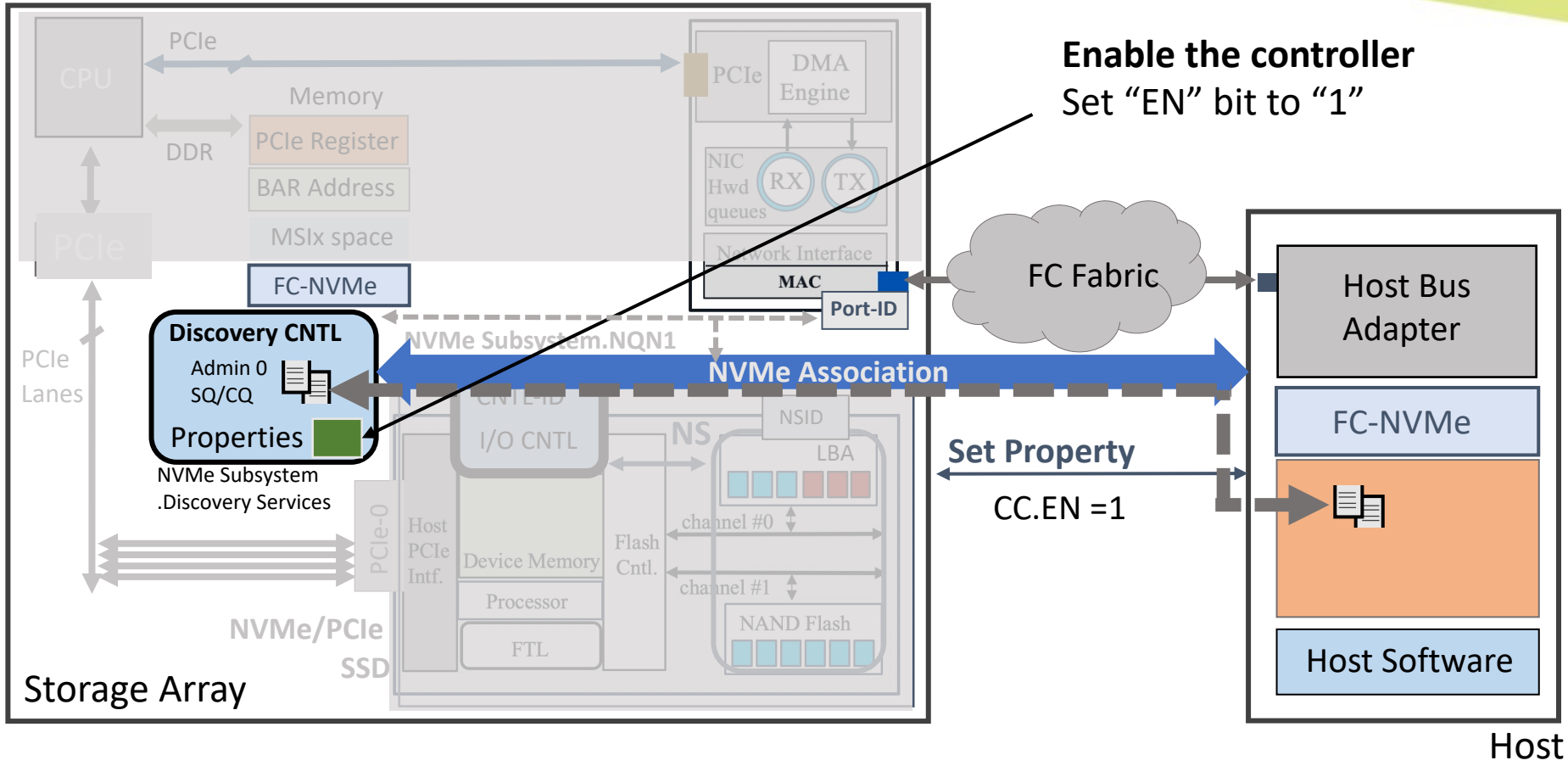
FC-NVMe Protocol Flows (Get Property CC)

Kamal Bakshi



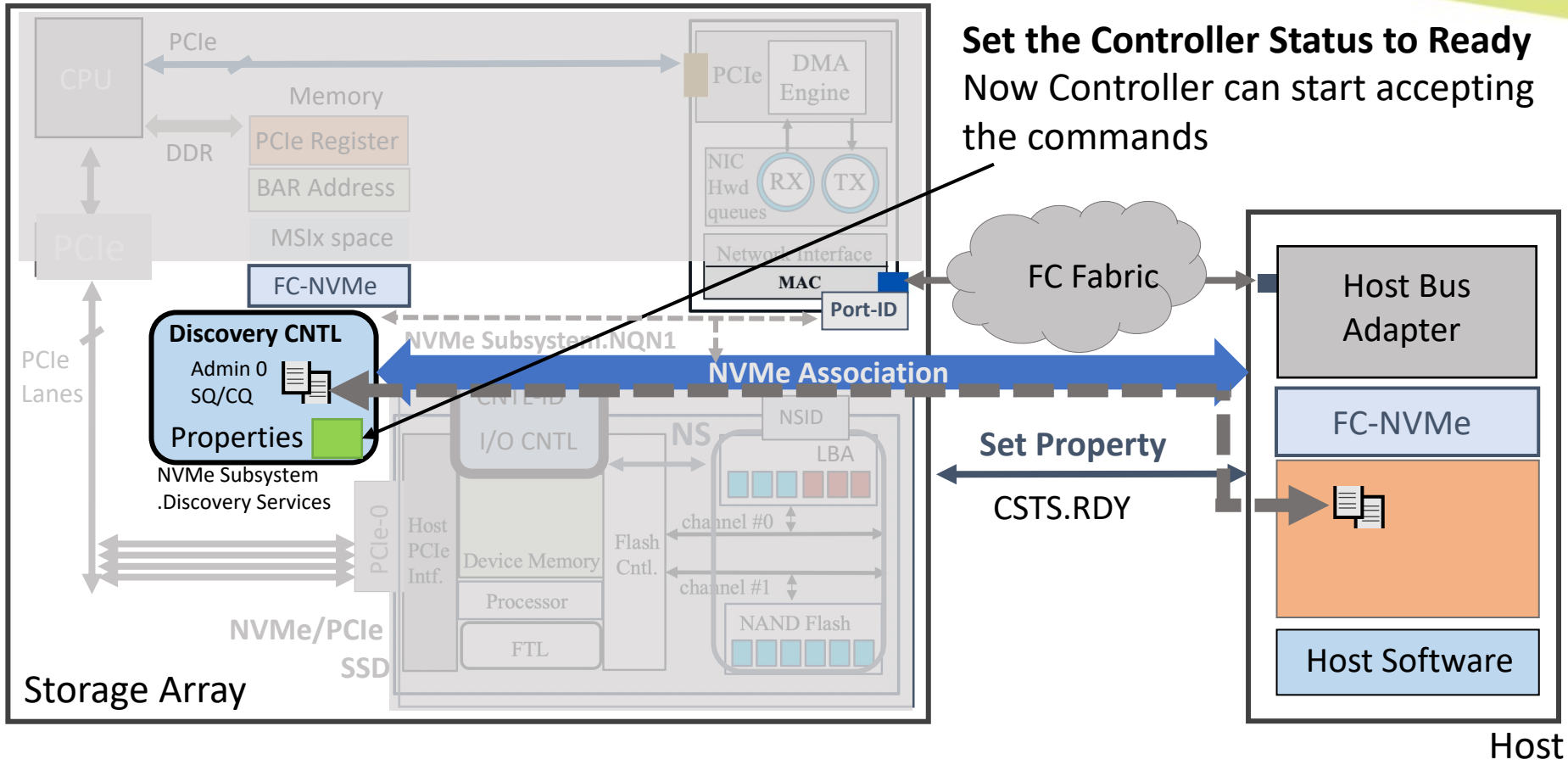
FC-NVMe Protocol Flows (Set Property CC.EN)

Kamal Bakshi



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

Kamal Bakshi

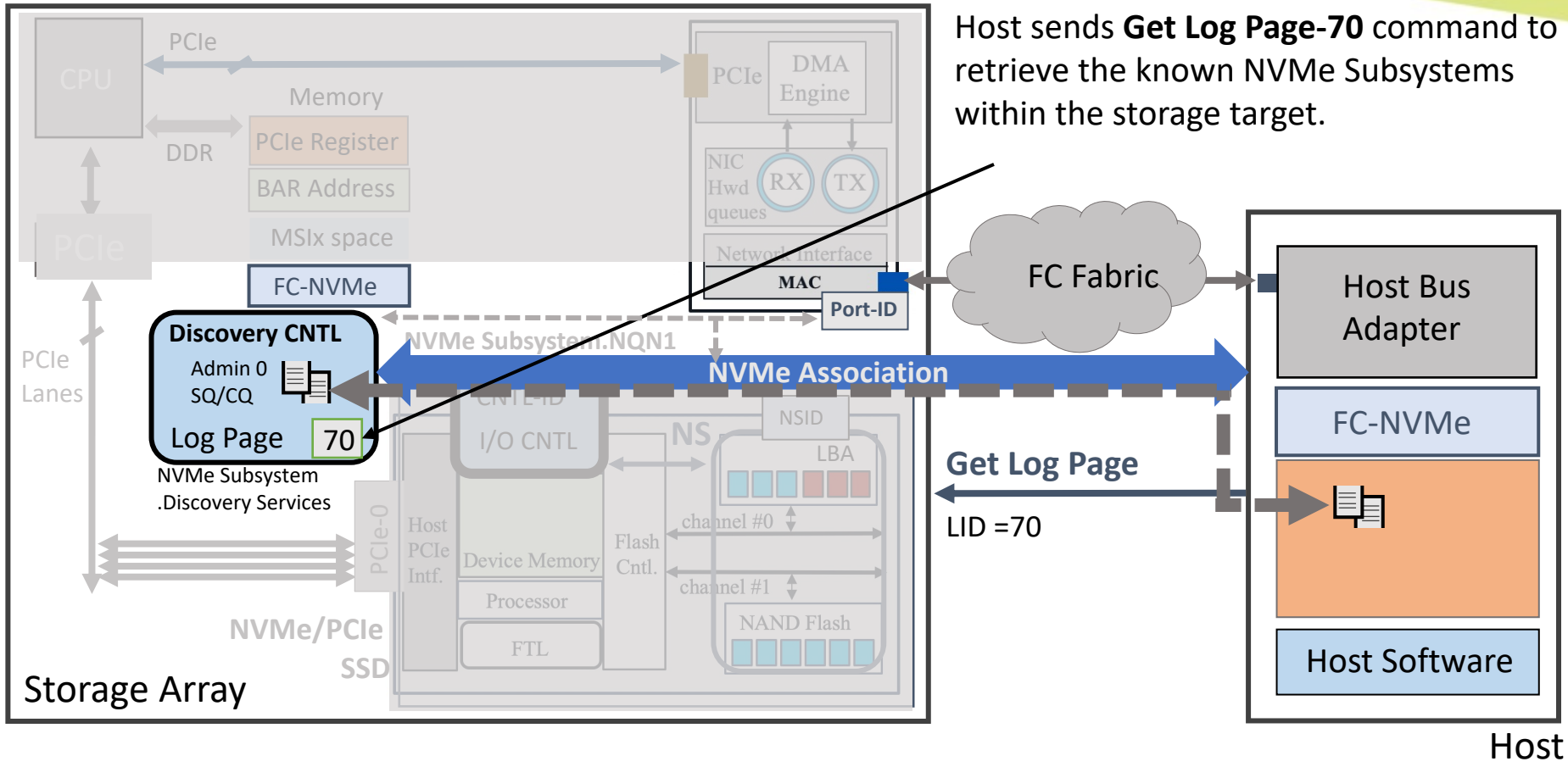


Set the Controller Status to Ready
Now Controller can start accepting the commands

FC-NVMe Protocol Flows (Get Log Page)

Kamal Bakshi

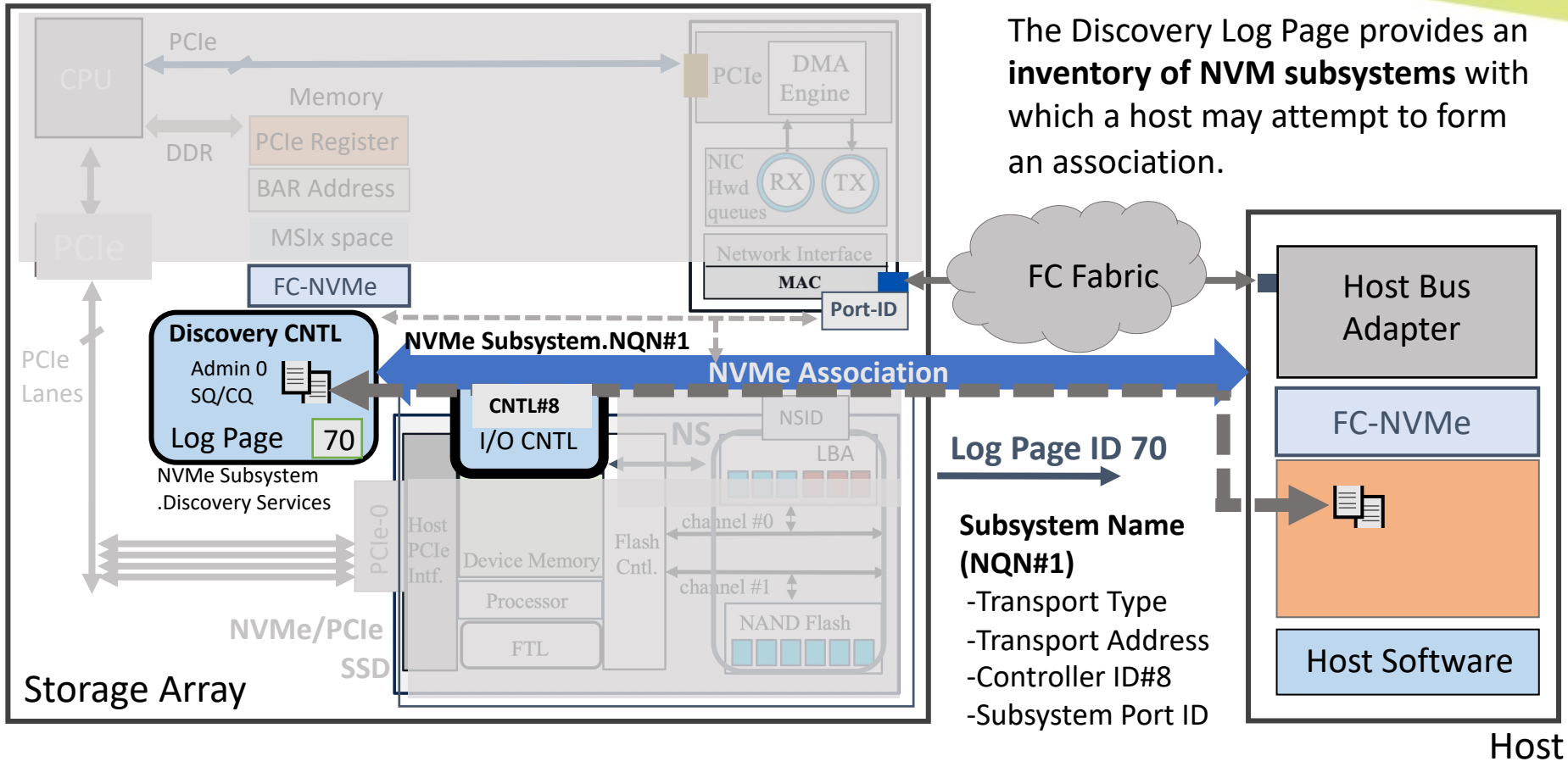
Host sends **Get Log Page-70** command to retrieve the known NVMe Subsystems within the storage target.



FC-NVMe Protocol Flows (Get Log Page)

Kamal Bakshi

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

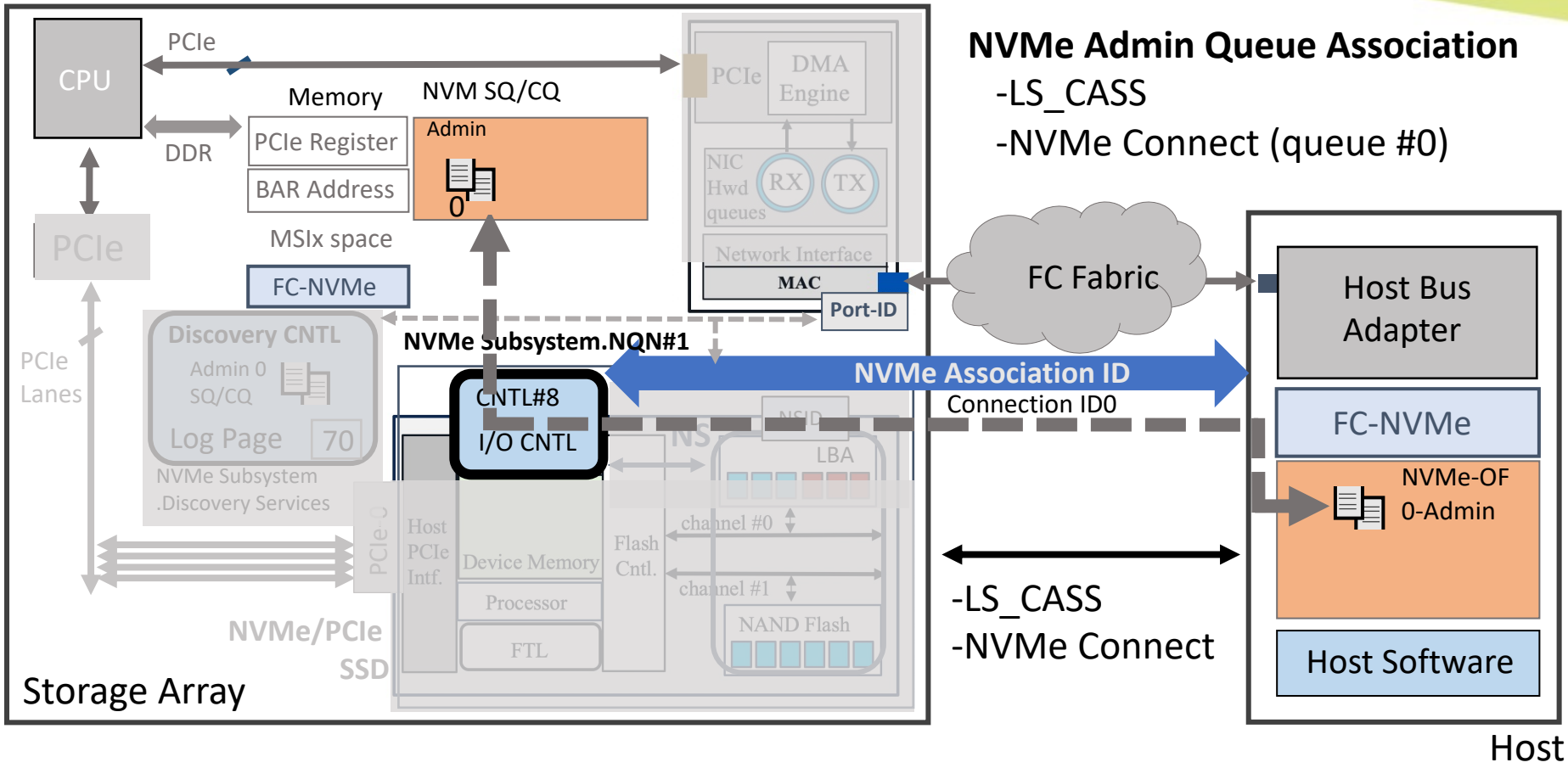


Log Page ID 70

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

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

Kamal Bakshi



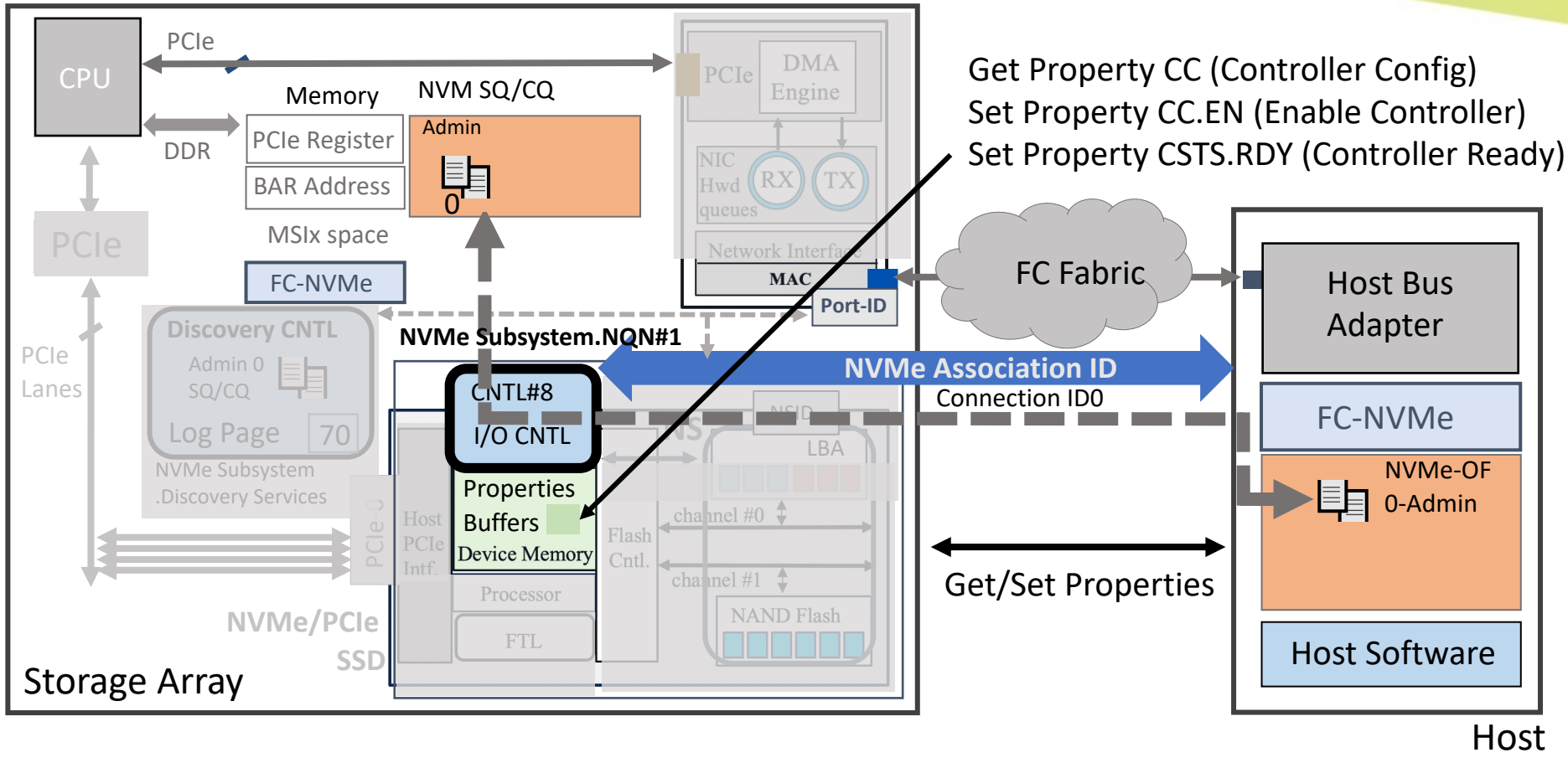
NVMe Admin Queue Association

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

-LS_CASS
-NVMe Connect

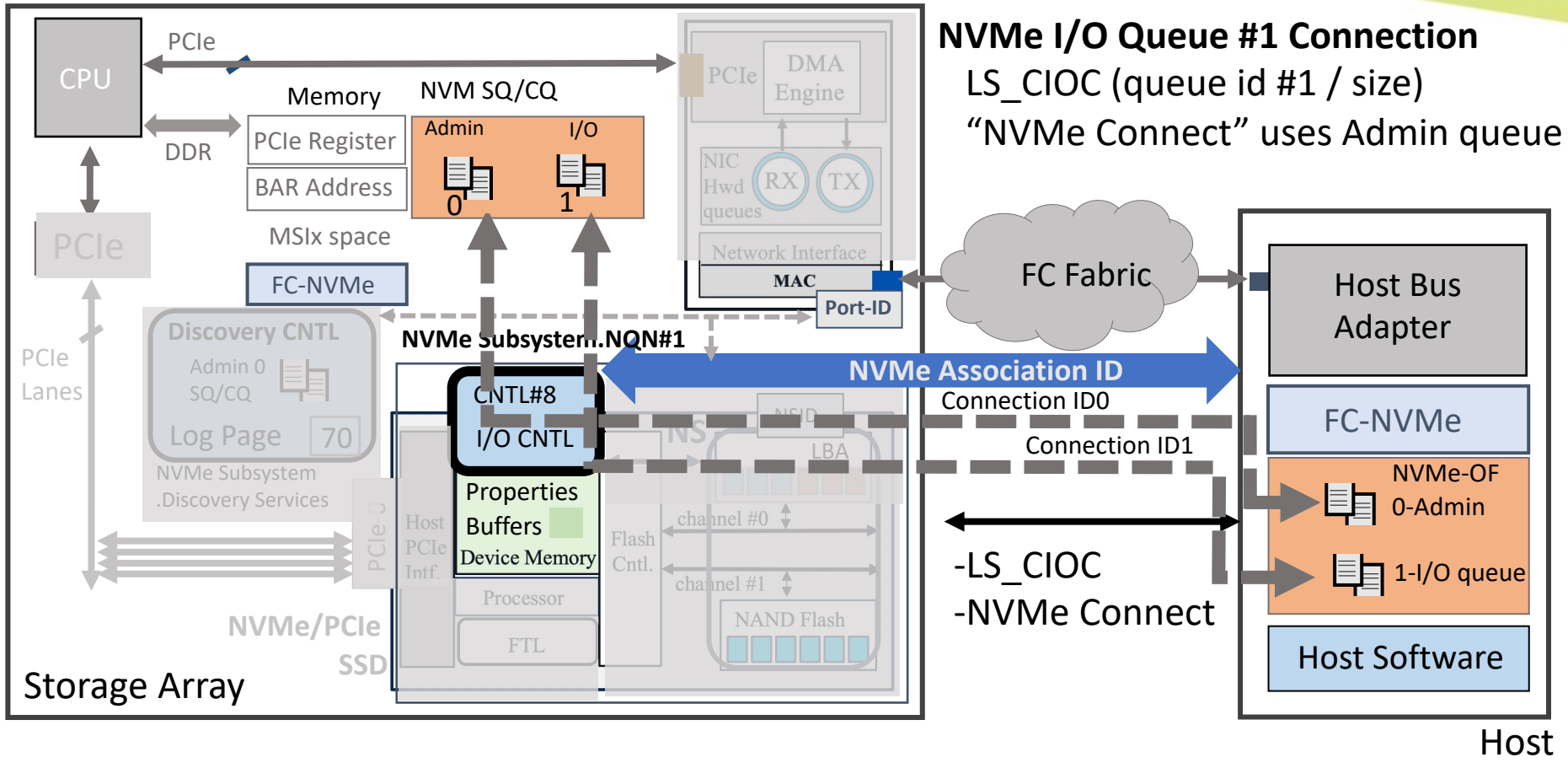
FC-NVMe Protocol Flows (I/O CNTL Ready to accept commands)

Kamal Bakshi



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

Kamal Bakshi



NVMe I/O Queue #1 Connection

LS_CIOC (queue id #1 / size)

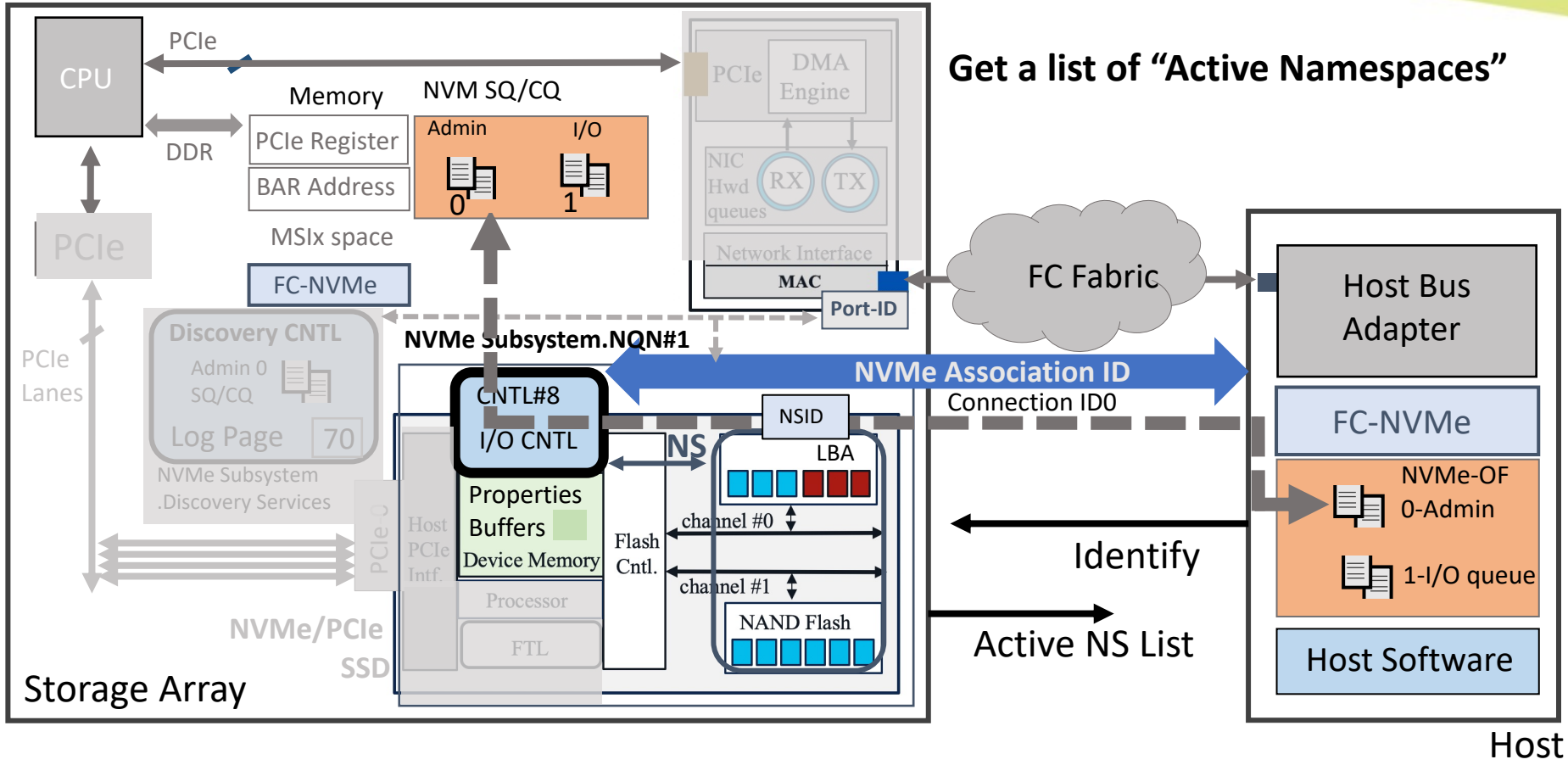
"NVMe Connect" uses Admin queue

-LS_CIOC

-NVMe Connect

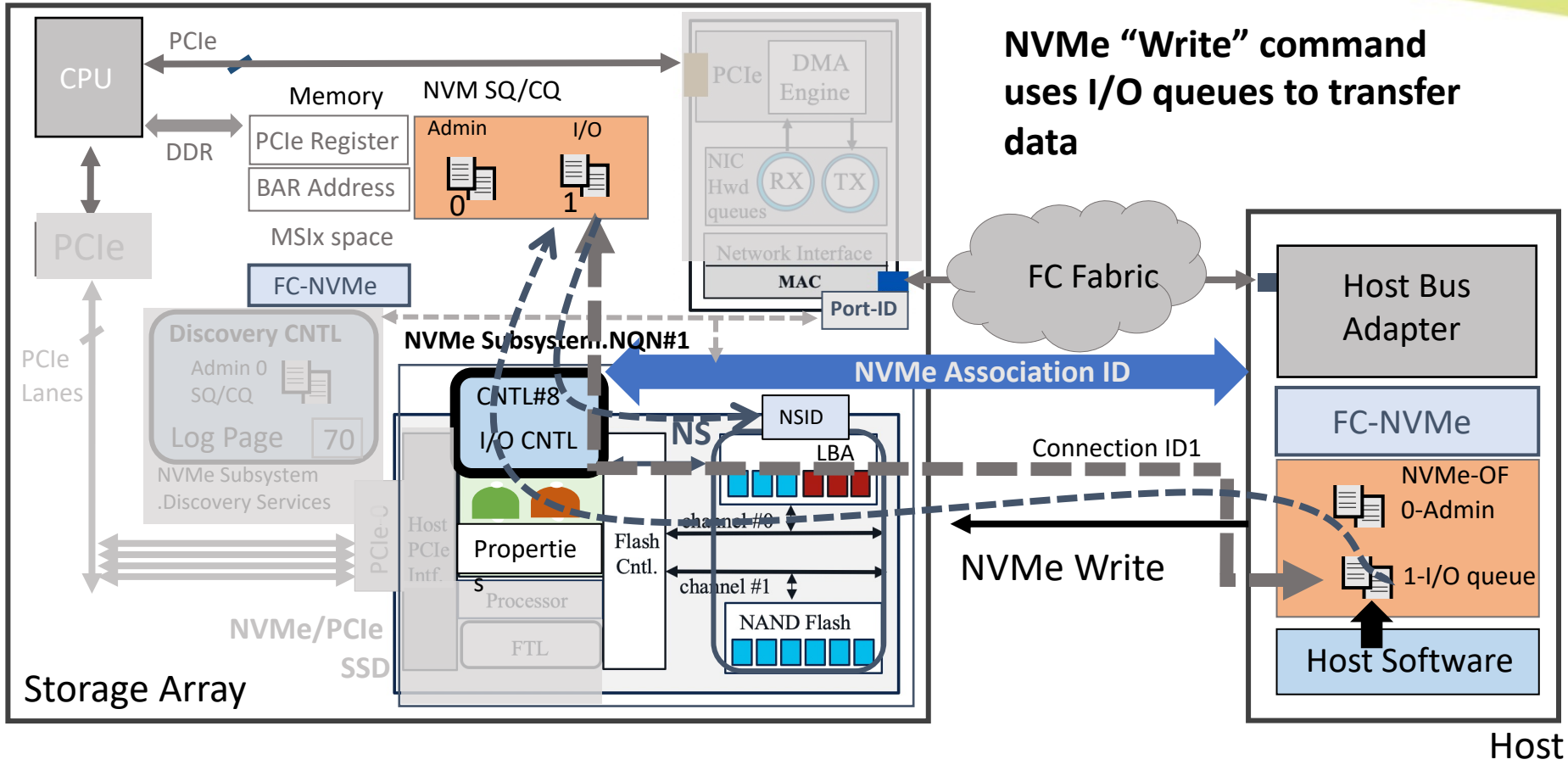
FC-NVMe Protocol Flows (NVMe Identify CNS 02)

Kamal Bakshi



FC-NVMe Protocol Flows (NVMe Write)

Kamal Bakshi



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

FC-NVMe (Read command)

Kamal Bakshi

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

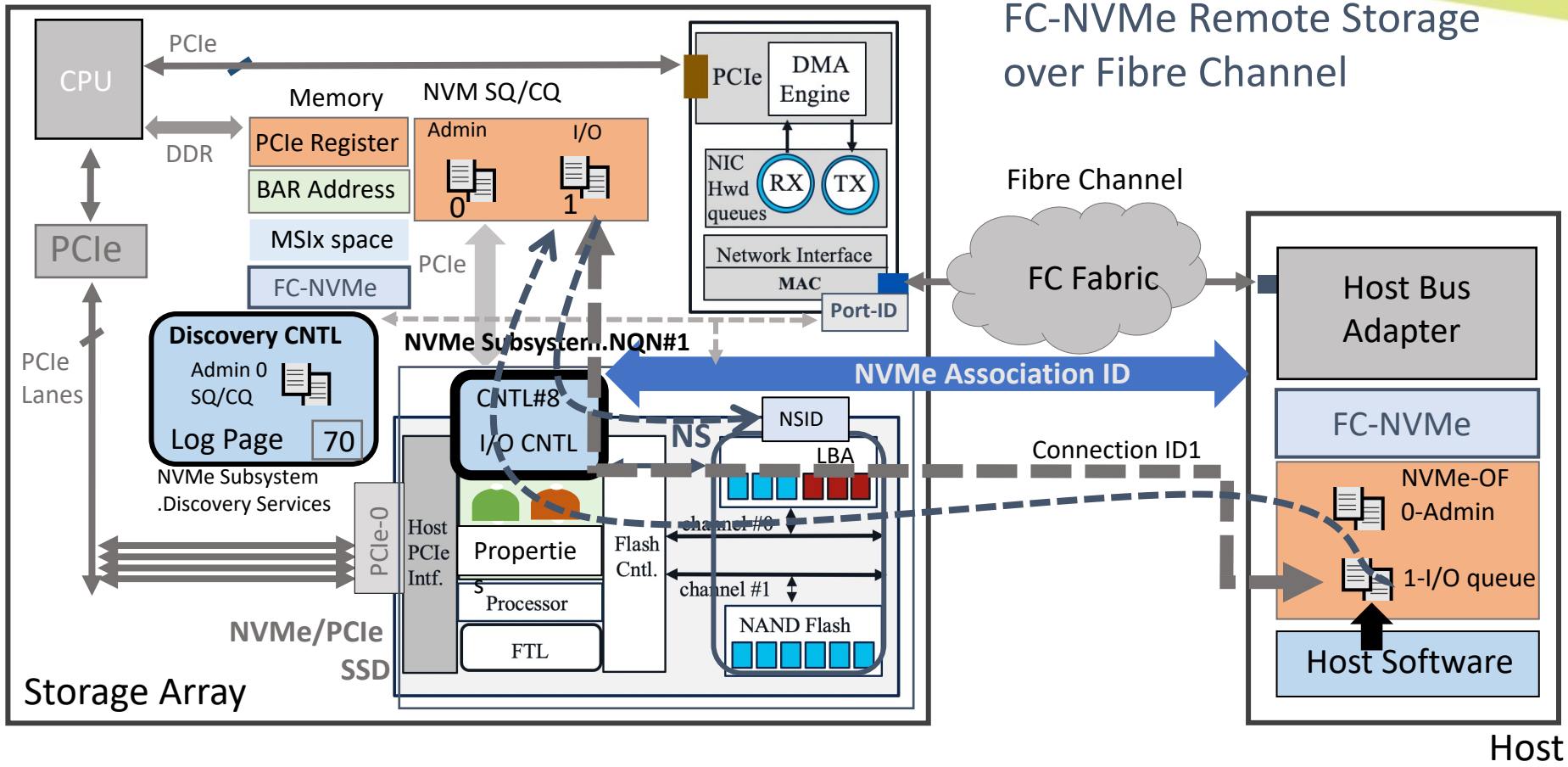
NSID

SLB

NLB

FC-NVMe Architecture

Kamal Bakshi



Summary FC-NVMe

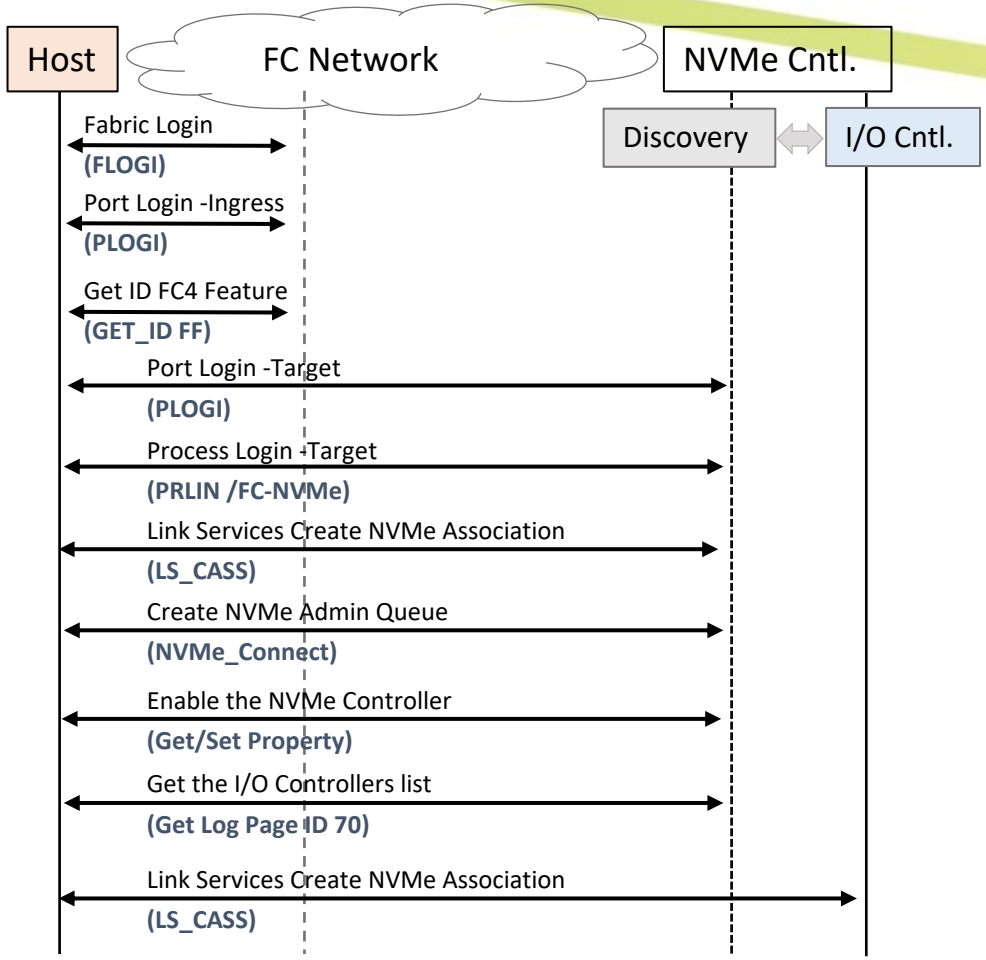
Kamal Bakshi

I Connect to NVMe Cntl.

II Create NVMe Queues

III Discover Namespaces

IV Start NVMe I/O operation



Summary FC-NVMe

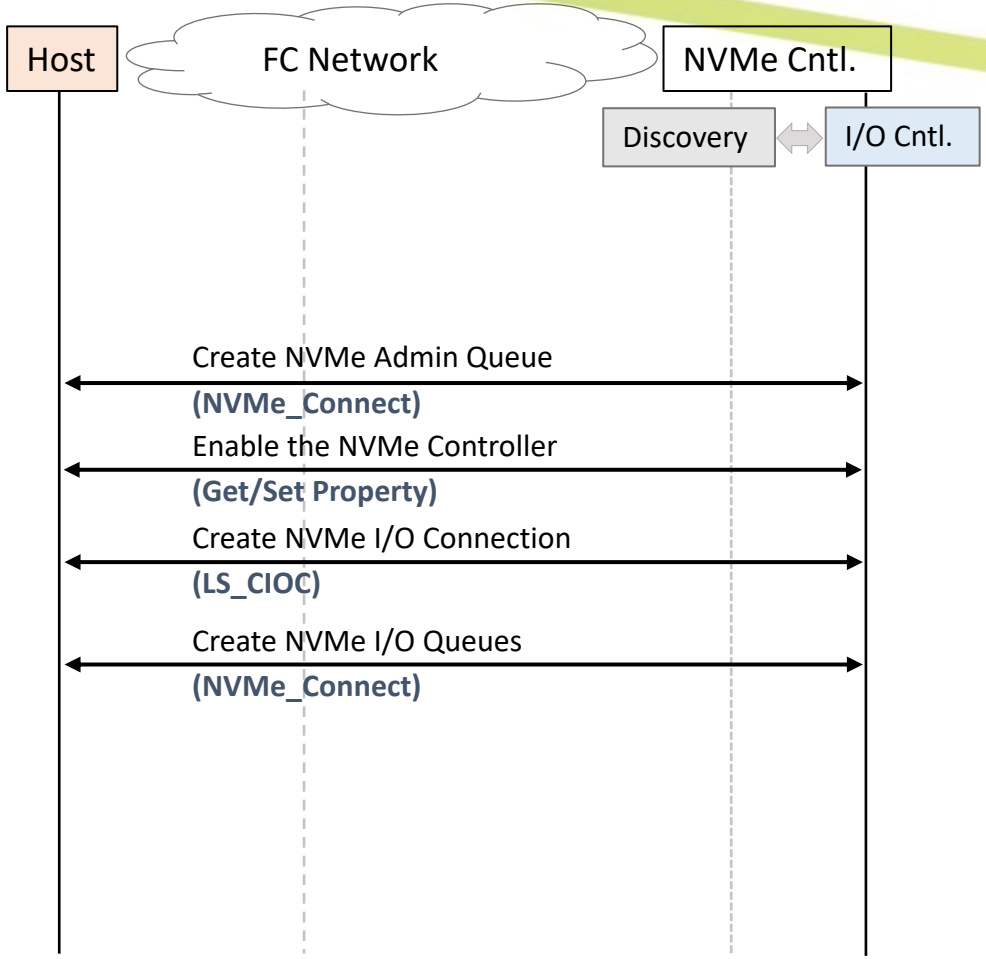
Kamal Bakshi

I Connect to NVMe Cntl.

II Create NVMe Queues

III Discover Namespaces

IV Start NVMe I/O operation



Summary FC-NVMe

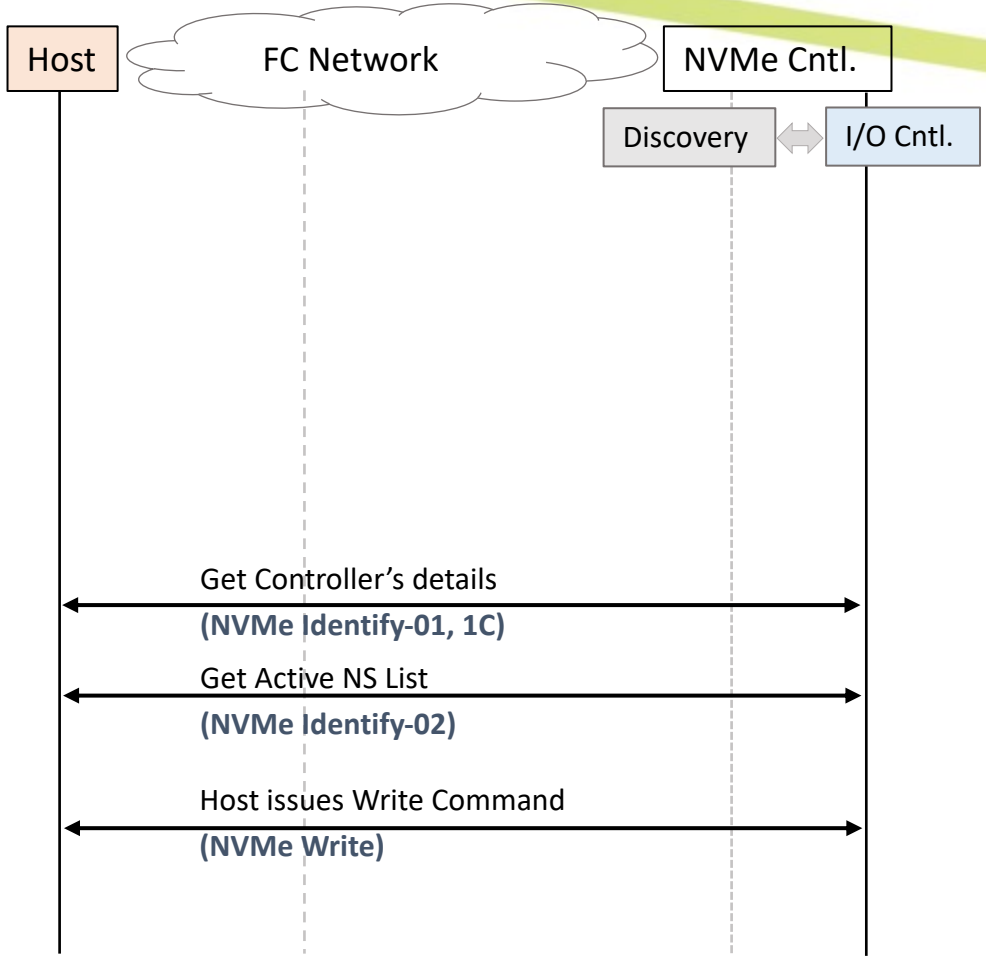
Kamal Bakshi

I Connect to NVMe Cntl.

II Create NVMe Queues

III Discover Namespaces

IV Start NVMe I/O operation



FC-NVMe mapping -Command IU

Kamal Bakshi

R_CTL (Routing)

To Target Port	To Initiator Port
01 NVMe_DATA	01 NVMe_DATA
03 NVMe_CONF	05 NVMe_XFER_RDY
06 NVMe_CMND	07 NVMe_RSP
09 NVMe_SR (Seq. Retry)	08 NVMe_ERSP
	0A NVMe_SR_RSP

FC ID = (28) NVMe/FC

Category
0001b Admin SQE
0xxxb I/O SQE

Flag
0 Write, 1 Read

Connection ID
Host queues mapping to the Controller's NVMe queues

DPS
Data Protection Type Setting

LBADS
LBA Data Size

MS
Meta Data Size

0001b for Admin queue
1xxxb for I/O queue
xxx = CSS

write
read(bit)

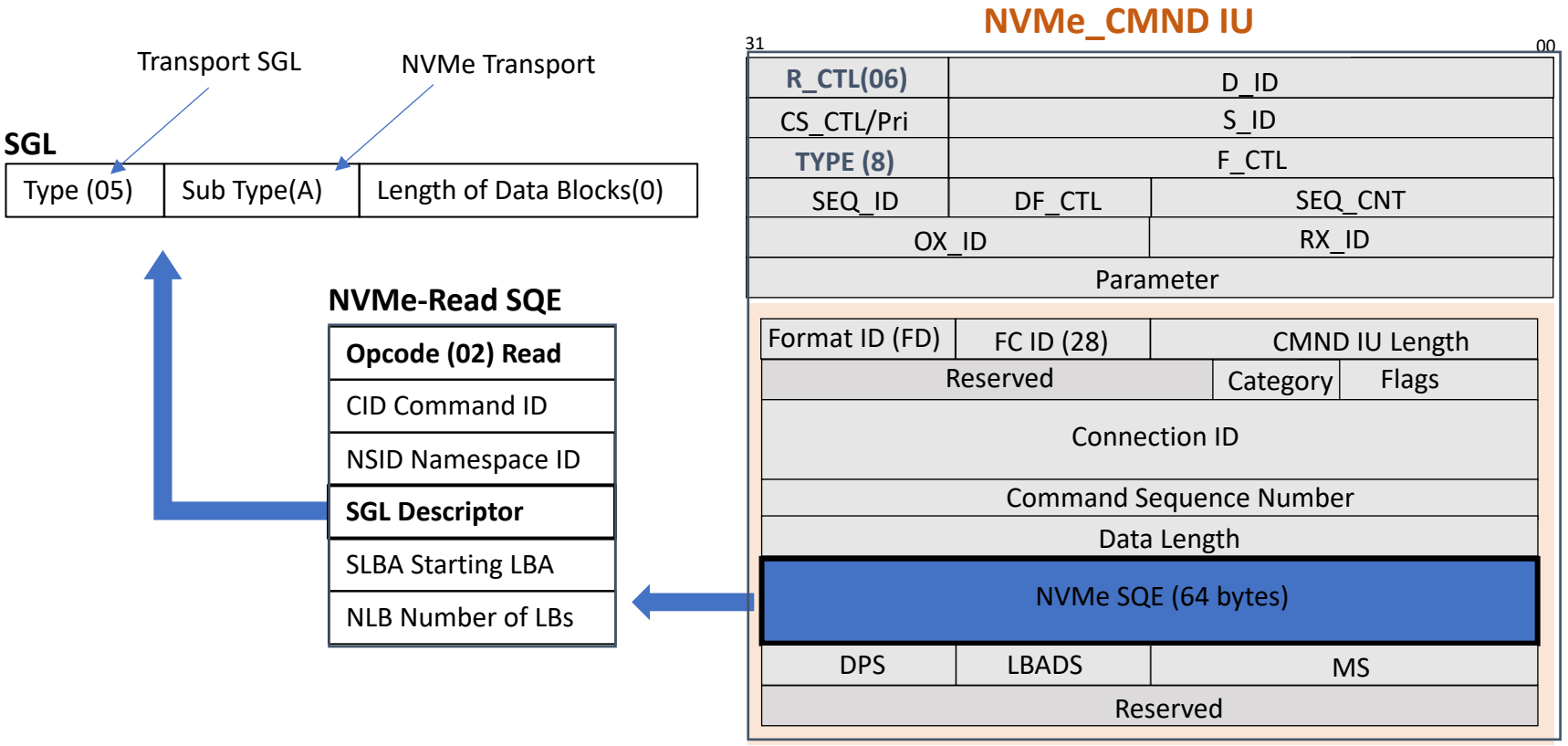
NVMe_CMND IU

31 … 00

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		
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		

FC-NVMe mapping -Read Command IU

Kamal Bakshi



FC-NVMe mapping -Data Transfer

Kamal Bakshi

NVMe-Data

FCP

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 IU

31		00
R_CTL(01)	D_ID	
CS_CTL/Pri	S_ID	
TYPE (8)	F_CTL	
SEQ_ID	DF_CTL	SEQ_CNT
OX_ID		RX_ID
Parameter (offset)		
Data		

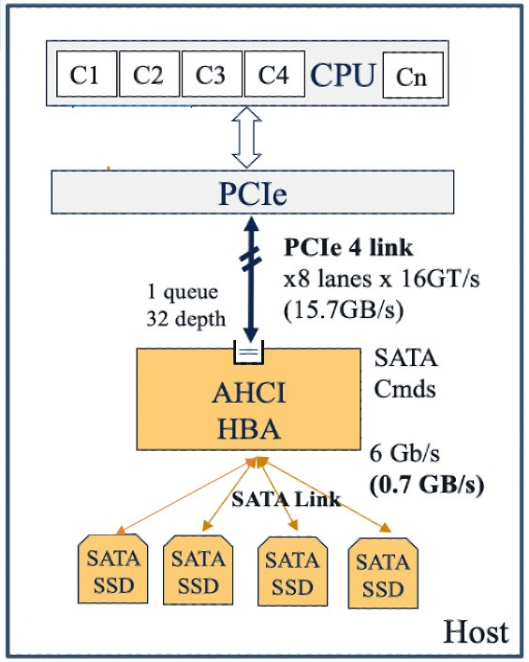
NVMe Data is transferred as FCP Data

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

NVMe Advantage#1: Higher Throughput/IOPS

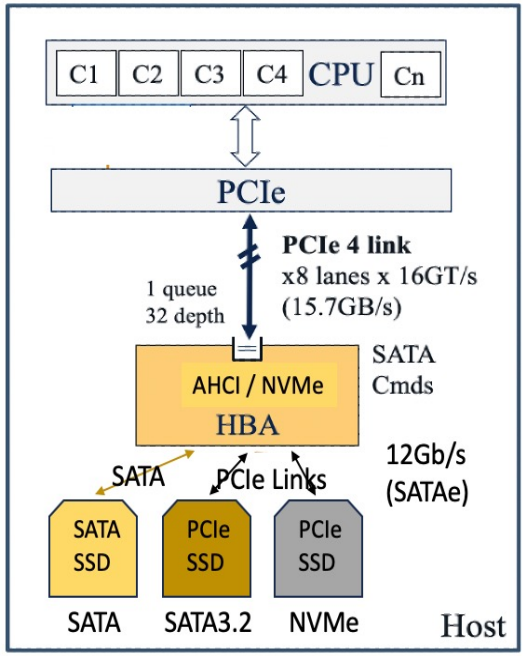
Kamal Bakshi

6Gb/s



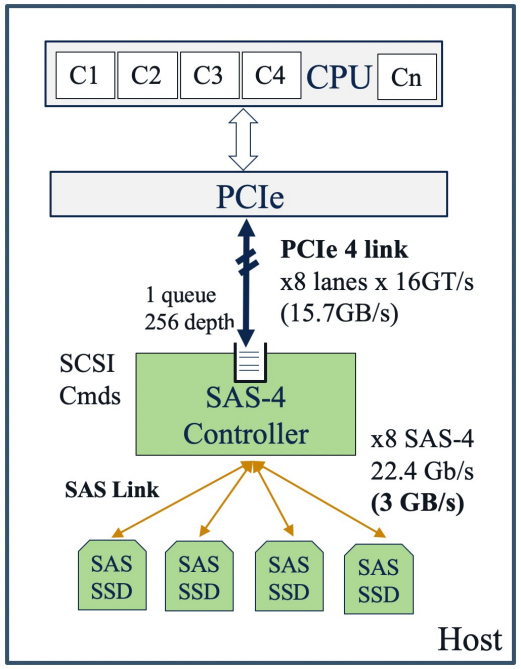
Legacy SATA

12Gb/s



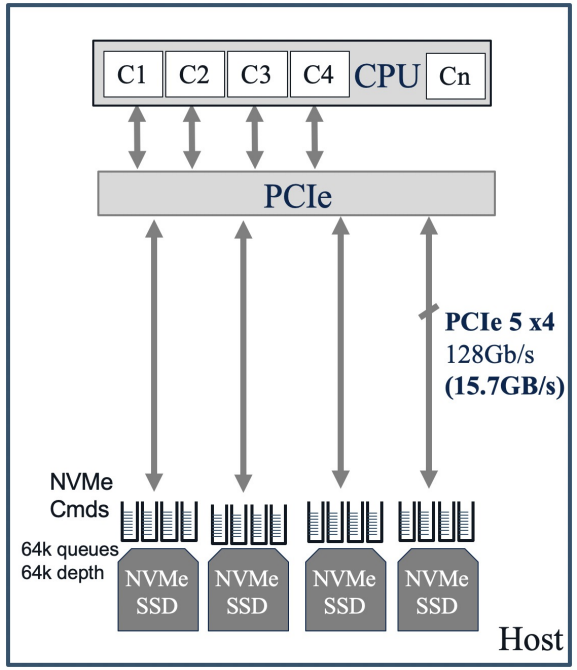
SATAexpress 3.2

22.4Gb/s



SAS-4 SSD

100Gb/s



PCIe/NVMe-SSD

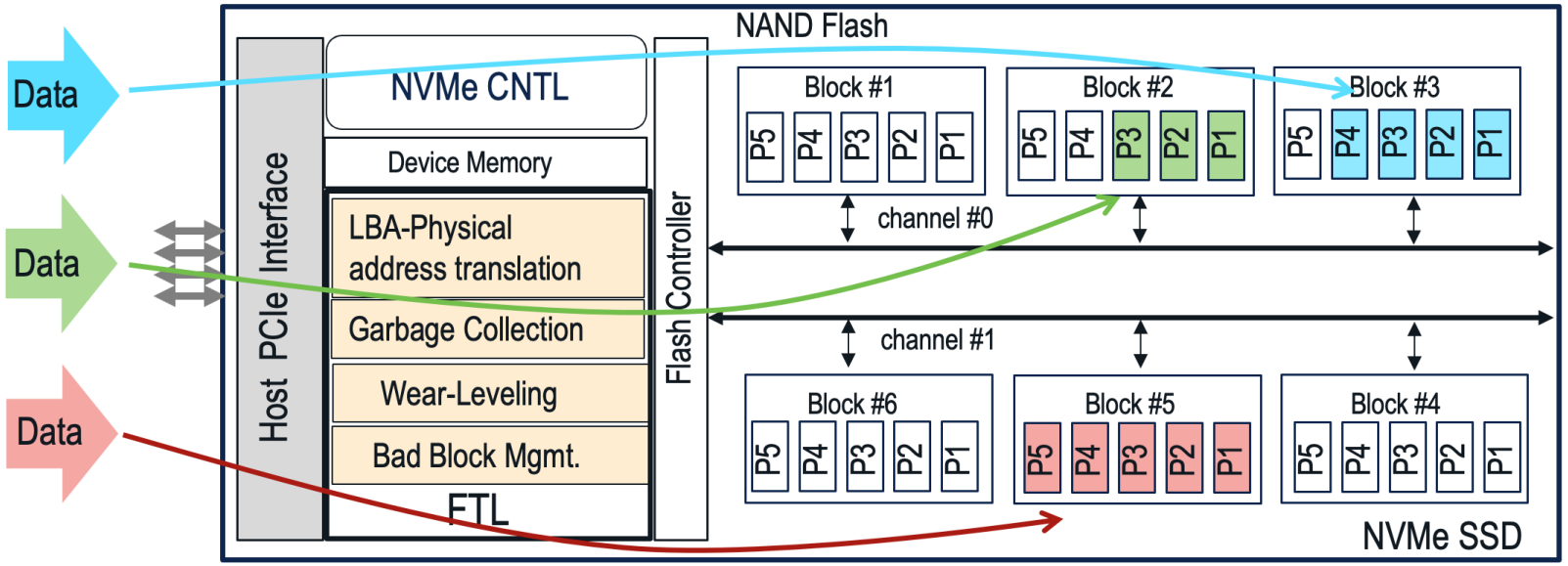
NVMe Advantage#2: Lower Tail Latency (ZNS SSD)

Kamal Bakshi

I/O Determinism
-NVMe Streams
-NVMe Sets

Open Channel
-Parallel Units/Chunks
-LightNVM

-NVMe ZNS
(Zoned Namespaces)



Flash Translation Layer

By dividing an NVMe namespace into zones, which are required to be sequentially written, ZNS offers essential benefits to hyper-scale organizations, all-flash array vendors and large storage-system vendors wishing to take advantage of storage devices optimized for sequential write workloads. ZNS reduces device-side write amplification, over-provisioning and DRAM while improving tail latency, throughput and drive capacity.

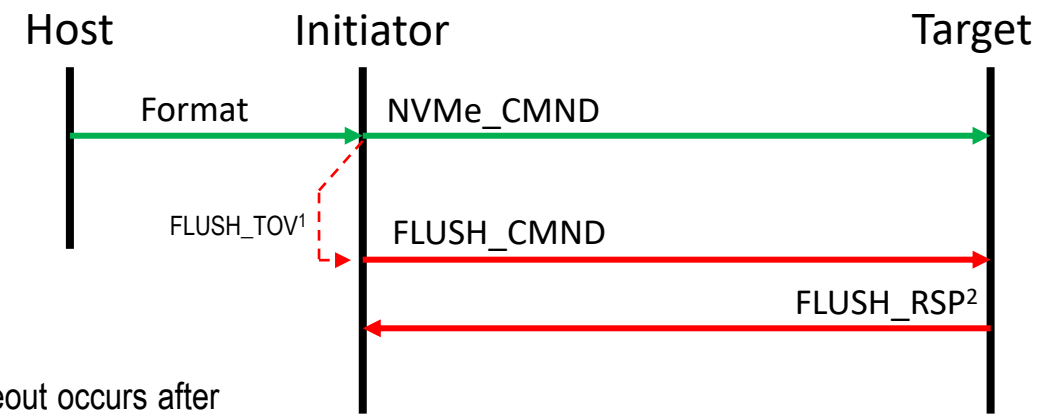
(FDP Flexible Data Placement –TP4146)*

NVMe Advantage#3: Error Detection & Recovery (SLER)

FC-NVMe-2 (Sequence Level Error Recovery)

Kamal Bakshi

Long Running Command



*1 FLUSH timeout occurs after transmitting the NVMe_CMND and a FLUSH BLS is sent to determine the status of the Exchange.

*2 The FLUSH_RSP indicates the Exchange is open and the target NVMe_Port holds Sequence Initiative.

No error recovery is required.

FLUSH_CMND

Bits / Word	31	30 .. 24	23 .. 16	15 .. 08	07 .. 00
0	HT	Reserved		Count	

HT: Halt Transmission bit

HT='1'

1-halt transmit operations on the Exchange;
2-transmit a FLUSH_RSP indicating the state of the Exchange; and
3-wait for an FC-4 specific event before further processing of the Exchange.

HT='0'

1-complete transmission of any active Sequence on the Exchange; and
2-transmit a FLUSH_RSP indicating the state of the Exchange.

FLUSH_RSP

Bits / Word	31 .. 24	23 .. 16	15 .. 08	07 .. 00
0	Flags	FC4INFO	Count	
1	FC4VALUE			
2	Reserved			
3	Reserved			

Flags

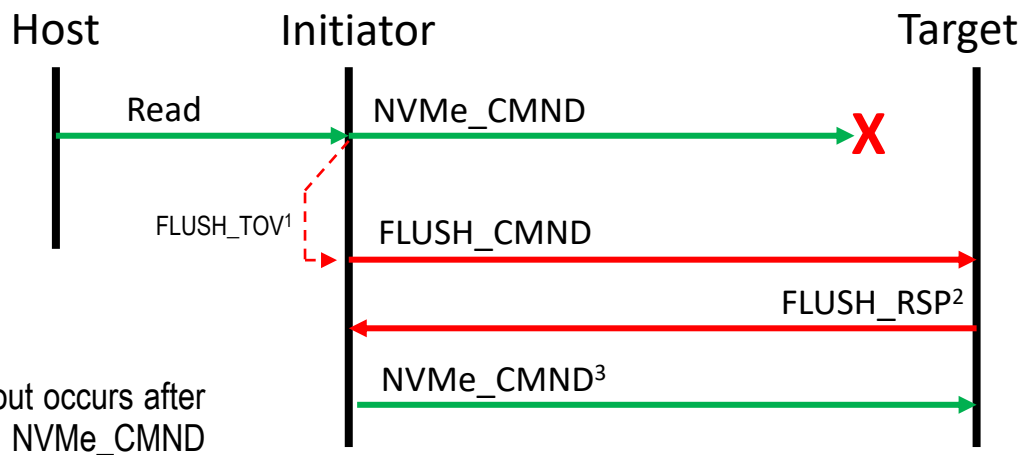
bit 30: OE (Open Exchange)
'0' Exchange is closed, '1' Exchange is open

NVMe Advantage#3: Error Detection & Recovery (SLER)

FC-NVMe-2 (Sequence Level Error Recovery)

Kamal Bakshi

Lost Command



*1 FLUSH timeout occurs after transmitting the NVMe_CMND and a FLUSH BLS is sent to determine the status of the Exchange.

*2 The FLUSH_RSP indicates the Exchange is not open at the target, and the Exchange is closed.

***3 The NVMe_CMND IU is retransmitted using the same OX_ID, SLER qualifier, and CSN.**

FLUSH_CMND

Bits / Word	31	30 .. 24	23 .. 16	15 .. 08	07 .. 00
0	HT	Reserved		Count	

HT: Halt Transmission bit

HT='1'

1-halt transmit operations on the Exchange;
2-transmit a FLUSH_RSP indicating the state of the Exchange; and
3-wait for an FC-4 specific event before further processing of the Exchange.

HT='0'

1-complete transmission of any active Sequence on the Exchange; and
2-transmit a FLUSH_RSP indicating the state of the Exchange.

FLUSH_RSP

Bits / Word	31 .. 24	23 .. 16	15 .. 08	07 .. 00
0	Flags	FC4INFO	Count	
1	FC4VALUE			
2	Reserved			
3	Reserved			

Flags

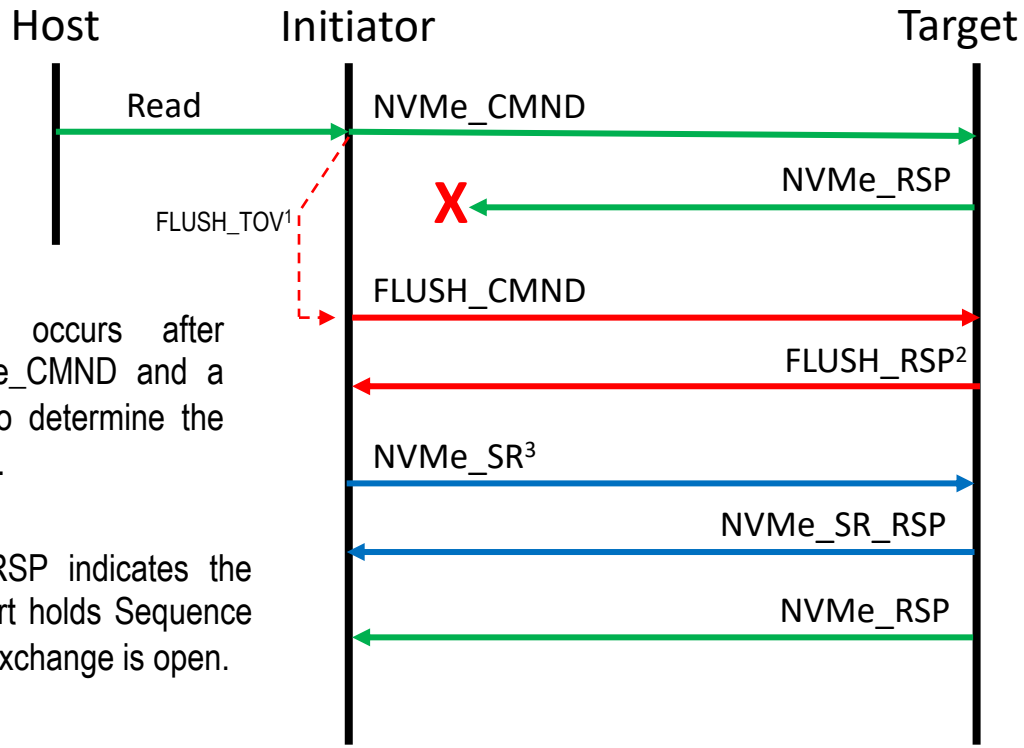
bit 30: OE (Open Exchange)
'0' Exchange is closed, '1' Exchange is open

NVMe Advantage#3: Error Detection & Recovery (SLER)

FC-NVMe-2 (Sequence Level Error Recovery)

Kamal Bakshi

Lost Response



*1 FLUSH timeout occurs after transmitting the NVMe_CMND and a FLUSH BLS is sent to determine the status of the Exchange.

*2 The FLUSH_RSP indicates the initiator NVMe_Port holds Sequence Initiative and the Exchange is open.

NVMe_SR

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																															

The NVMe_SR IU contains a request for Sequence retransmission from the initiator NVMe_Port to the target NVMe_Port to recover from an Exchange error.

NVMe_SR_RSP

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																															

00h: Accepted
01h: Invalid FC ID
02h: Invalid Opcode
03h: Logical error
04h: Invalid qualifiers
09h: Unable to perform retransmission request

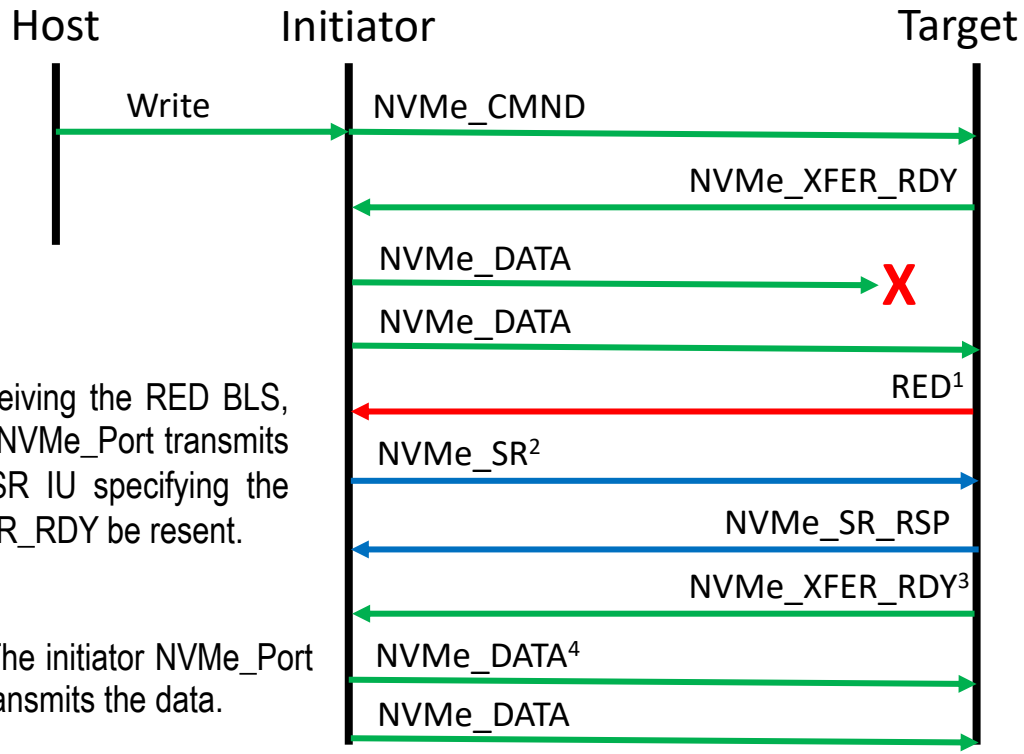
***3 The initiator NVMe_Port transmits an NVMe_SR IU specifying the NVMe_RSP be resent.**

NVMe Advantage#3: Error Detection & Recovery (SLER)

FC-NVMe-2 (Sequence Level Error Recovery)

Kamal Bakshi

RED: Responder Error Detected



The RED command may be transmitted by an Exchange Responder to indicate to the Exchange Originator that a Sequence error was detected on an open Exchange.

*1 After detecting a Sequence error the target transmits a RED BLS.

*2 After receiving the RED BLS, the initiator NVMe_Port transmits an NVMe_SR IU specifying the NVMe_XFER_RDY be resent.

*3 The target NVMe_Port retransmits the NVMe_XFER_RDY with Relative Offset set to zero.

*4 The initiator NVMe_Port retransmits the data.

NVMe Advantage#4: Performance Advantage for KV Databases (KV SSD)

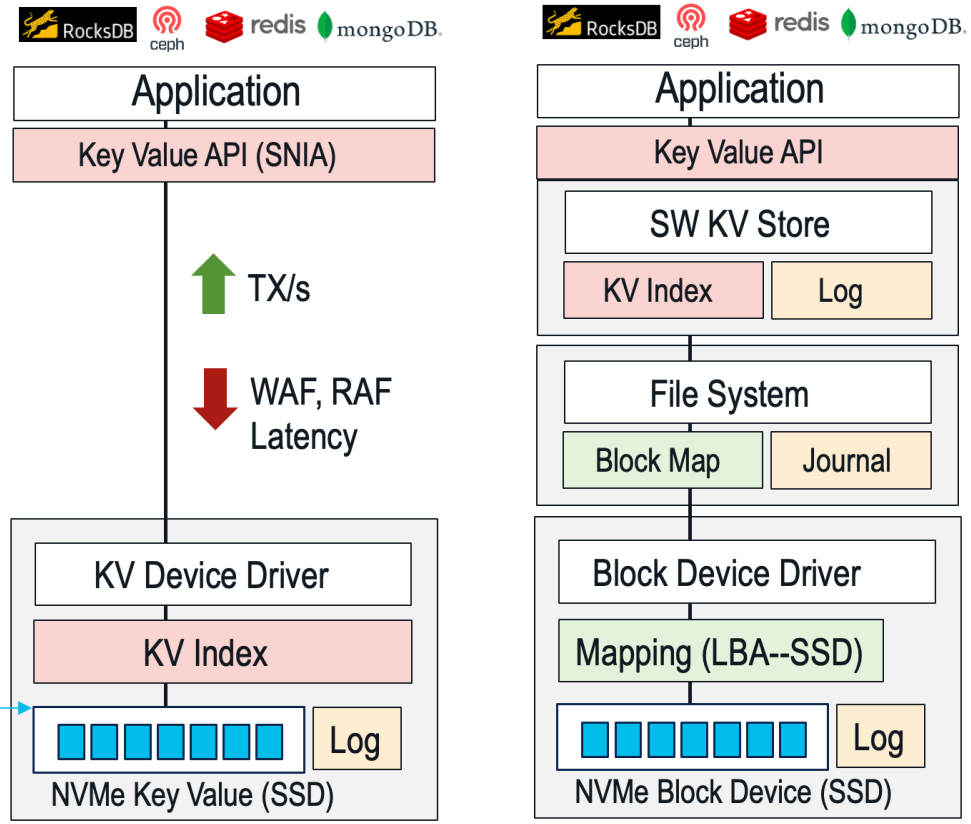
Kamal Bakshi

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.

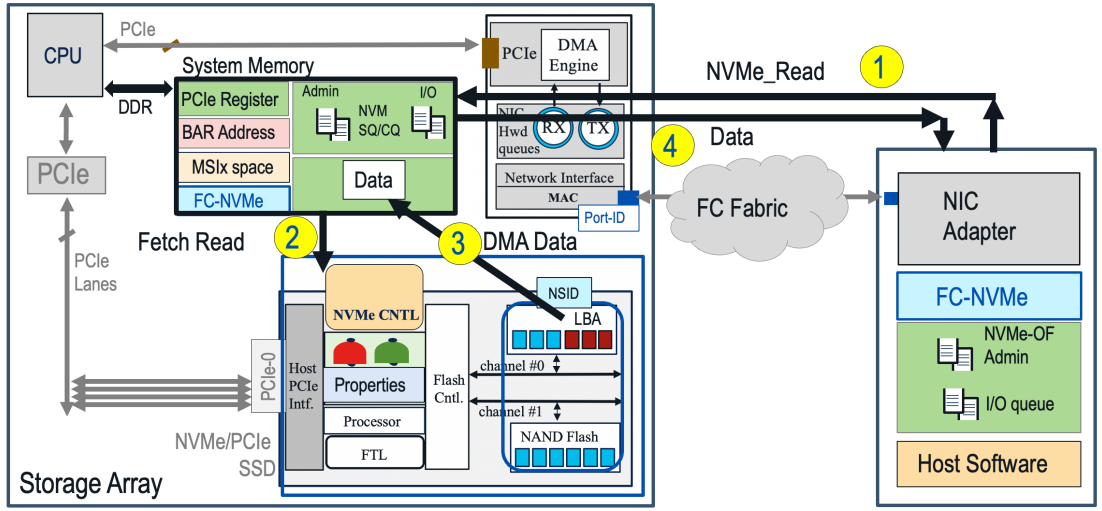
Key Value API (SNIA)
- -Open/Retrieve Device
- -Create/Delete Key Space
- -Store, Retrieve, Delete,
- -List, Delete Group

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

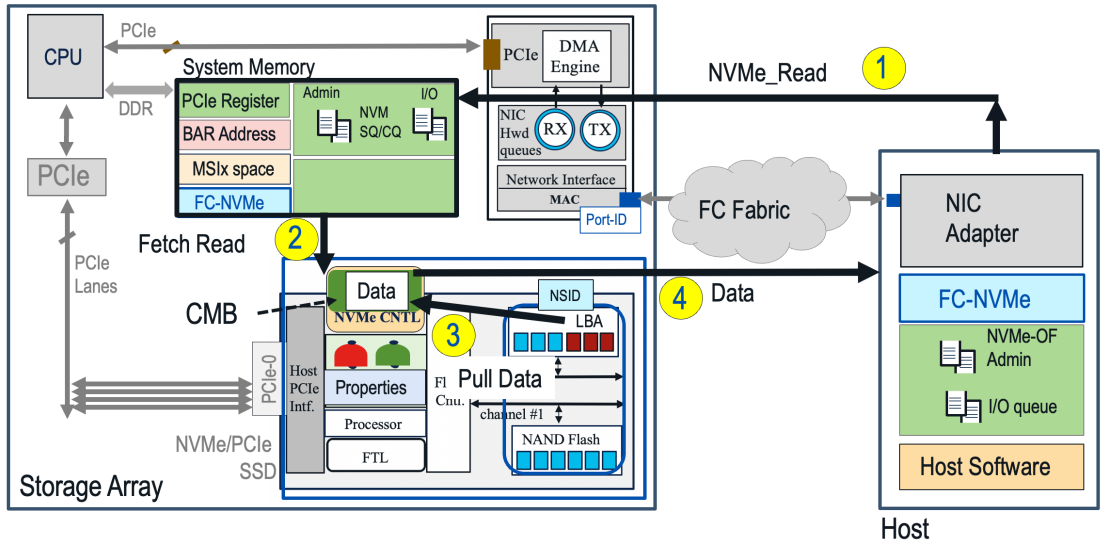


NVMe Advantage#5: Lower CPU utilization (CPU Offload with CMB)

Kamal Bakshi

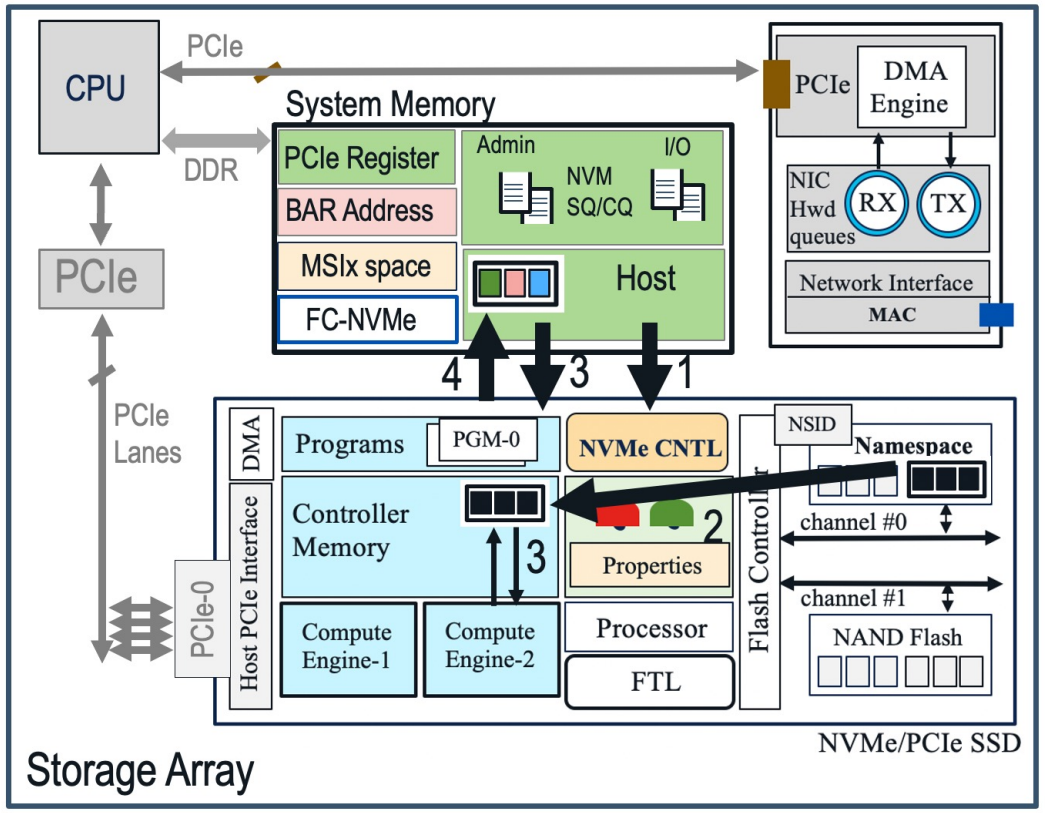


CPU Offload with NVMe Controller Memory Buffer (CMB)



NVMe Advantage#6: Computational Storage (Future)

Kamal Bakshi



1 NVMe Read (NS) is issued to CNTL

2 CNTL moves the (NS) data to CM

3 Execute PGM-0 on Compute E.-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

Adoption and Use Cases

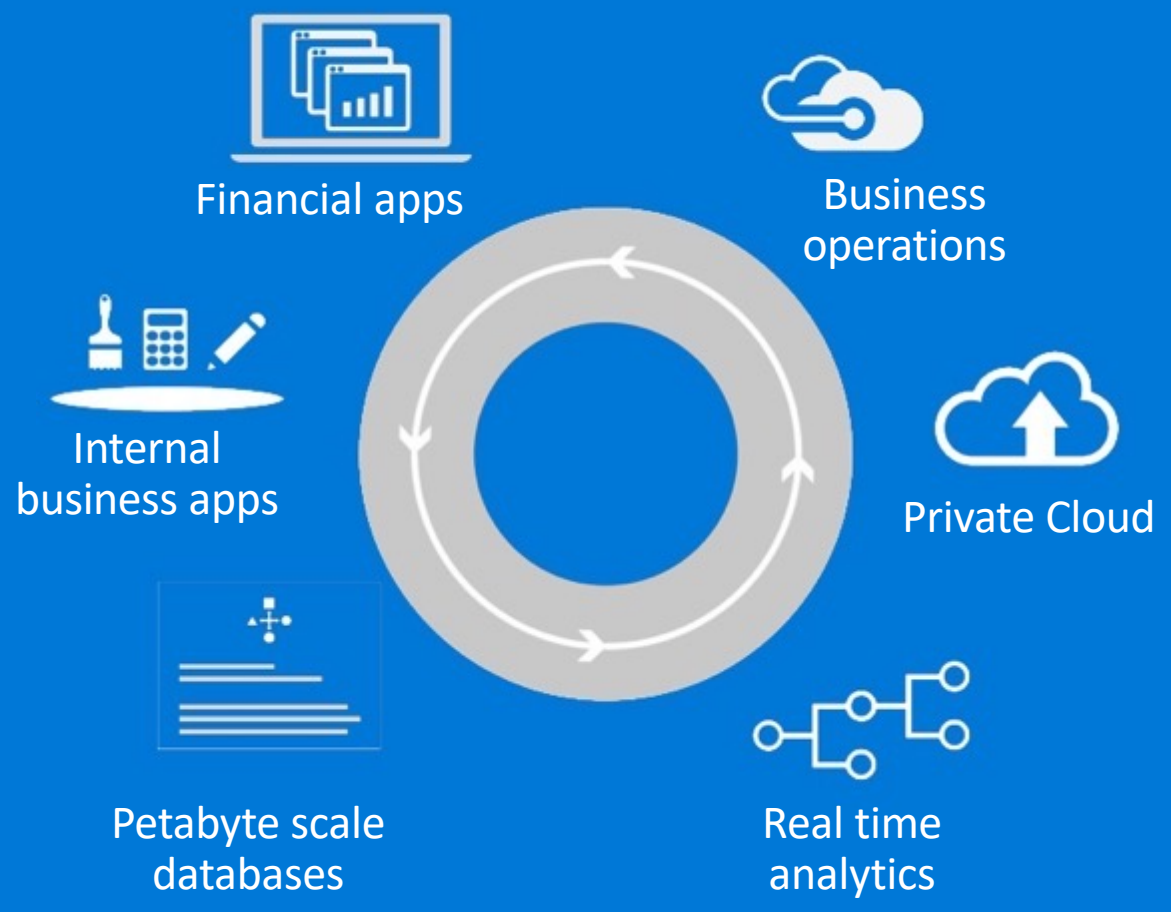
Nishant Lodha
Director, Emerging Technologies
Marvell Semiconductor

FC-NVMe Adoption - on the rise

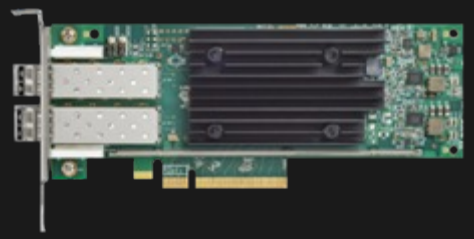
Low-latency business-critical applications that transact with all-flash arrays are driving the adoption of FC-NVMe

Industry estimates that ~10% of Fibre Channel deployments leverage FC-NVMe

Preserve and leverage existing investments with *full backward compatibility* with FCP-SCSI



FC-NVMe Hardware Ecosystem



HBAs

2/2

Broadcom Emulex
Marvell QLogic

Switches

2/2

Cisco MDS
Broadcom Brocade

Storage Array

5/5

DELLEMC, NetApp, HPE, Pure Storage, IBM, others

FC-NVMe: Operating System Support

Linux

Fully supported

RHEL 7.8 / 8.0 onwards

SLES 12 SP4 / SLES 15 SP2 onwards

Multipathing and Boot

VMware ESXi

Fully supported

VMware ESXi 7.0 onwards

VMware ESXi 8.0 adds vVols support

Multipathing and Boot

Microsoft Windows

Emulation Mode Support

No native support for FC-NVMe

HBA Vendors provide emulated drivers

Does not impact Windows VMs on VMware ESXi or KVM

FC-NVMe adoption is growing, join in!

Performance

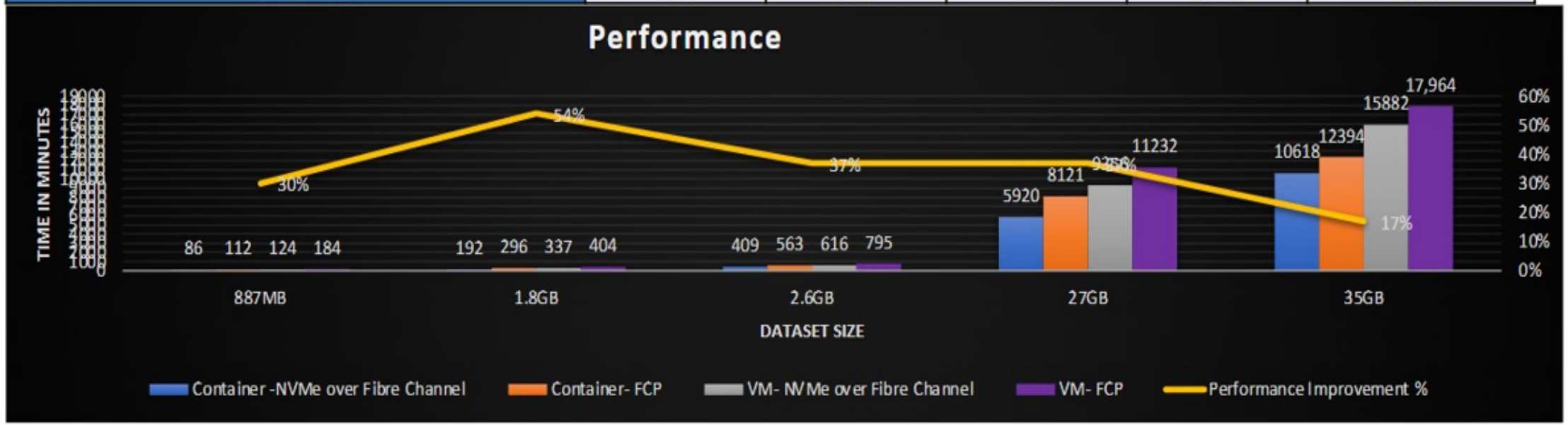
Mark Jones
Director Technical Marketing, Broadcom
FCIA Board of Directors

NVMe-FC Performance Example

Improves Performance of Containerized AI/ML Models

- From Feb 2023 FCIA Webcast https://www.brighttalk.com/webcast/14967/57217

Time in minutes					
Dataset size	887MB	1.8GB	2.6GB	27GB	35GB
Container -NVMe over Fibre Channel	86	192	409	6920	10618
Container- FCP	112	296	563	8121	12394
VM- NVMe over Fibre Channel	124	337	616	9356	15882
VM- FCP	184	404	795	11232	17,964
Performance Improvement with container on NVMe over Fibre Channel %	30%	54%	37%	37%	17%



FCIA
FIBRE CHANNEL INDUSTRY ASSOCIATION

NVMe/FC vs FCP - Database OLTP Performance

64GFC/NVMe-FC Database Server

All Flash Storage Array with 4x 32GFC NVMe/FC

OLTP TPM

53%

Faster Transactions than FCP/SCSI Protocol

HammerDB TPROC-C OLTP TPM

Stored Procedure LATENCY

58%

Lower Latency than FCP/SCSI Protocol

HammerDB TPROC-C Time Profile 95th percentile Stored Procedure latency

CPU Efficiency

106%

Better Server CPU Efficiency than FCP/SCSI Protocol

Greater HammerDB TPROC-C TPM per Server CPU Utilization

After this Webcast

- Please rate this event – we value your feedback
- We will post a Q&A blog at http://fibrechannel.org/ with answers to today's questions
- Our library of FCIA on-demand webcasts http://fibrechannel.org/webcasts/ includes:
  - Fibre Channel Fundamentals
  - FC-NVMe
  - FC-NVMe for ML
  - Long Distance Fibre Channel
  - Fibre Channel Speedmap
  - FCIP (Extension): Data Protection and Business Continuity
  - Fibre Channel Performance
  - FICON
  - Fibre Channel Cabling
  - 64GFC and 128GFC
  - FC Zoning Basics
- Follow us @FCIAnews for updates on future FCIA webcasts

Thank You