# FC-NVMe Deep Dive

Curt Beckmann, Principal Architect, Brocade
Craig W. Carlson, Chair, T11 FC-NVMe Committee, Cavium
J Metz, FCIA Board of Directors ,Cisco

FIBRE CHANNEL INDUSTRY ASSOCIATION

# Agenda

- Introduction
- Fibre Channel terms and background
- NVMe terms and background
- Deep dive on FC-NVMe
- FC-NVMe use cases
- Summary

# Today's Presenters

**Curt Beckmann**
**Principal Architect**
**Brocade**

**Craig W. Carlson**
**Chair, T11 FC-NVMe Committee**
**FCIA Board of Directors**
**Cavium**

**J Metz**
**FCIA Board of Directors**
**Cisco**

# What This Presentation Is

- A deeper dive into how FC-NVMe works
- A look at FC-NVMe use cases in the data center

# What This Presentation Is *Not*

- **Not a tutorial on Fibre Channel or NVMe over Fabrics**
  - For an introduction to FC-NVMe see the FCIA webcast “Introducing Fibre Channel NVMe” https://www.brighttalk.com/webcast/14967/242341
  - Comprehensive (no boiling the ocean)
- **A comparison between FC and other NVMe over Fabrics methods**

# Fibre Channel Background

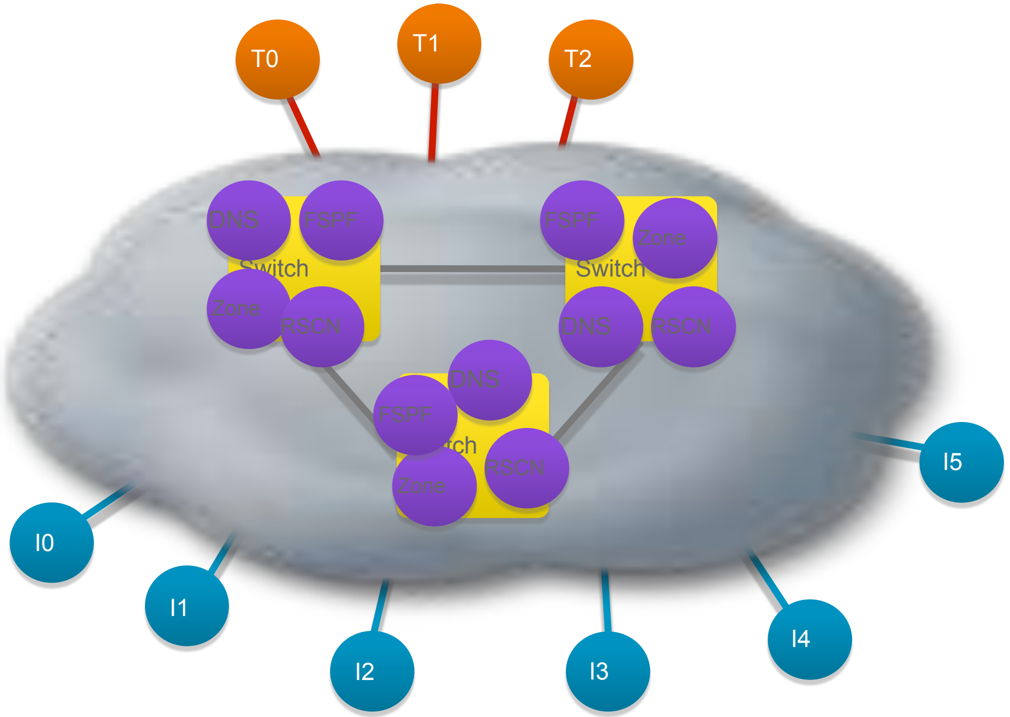
FIBRE CHANNEL INDUSTRY ASSOCIATION

# What is Fibre Channel?

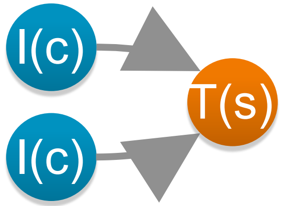
- A network purpose-built for storage
- A physical connection between a host and its storage
- A logical (protocol) connection between a host and its storage
- A collection of Fabric based Services to support host and storage

# Design Requirements

- **Fibre Channel Storage Area Network (SAN)**
  - Goal: Provide one-to-one connectivity
  - Transport and Services are on same layer in same devices
  - Well-defined end-device relationships (initiators and targets)
  - Does not tolerate packet drop – requires lossless transport
  - Only north-south traffic, east-west traffic mostly irrelevant
- **Network designs optimized for Scale and Availability**
  - High availability of network services provided through dual fabric architecture
  - Edge/Core vs. Edge/Core/Edge
  - Service deployment



Fabric topology, services and traffic flows are structured



Client/Server Relationships are pre-defined

# FC Basics and Terminology

- **Fibre Channel was traditionally defined with 3 classes of service**
  - Class 1 – Worked like a telephone cross bar switch
    - This Class has been deprecated and is no longer used
  - Class 2 – Acknowledged datagram service
  - Class 3 – Unacknowledged datagram service
- **Class 3 is the service used by FC-NVMe**

# FC Basics and Terminology (cont.)

- **Each unit of transmission is called a “Frame”**
  - A frame can be up to 2112 bytes
- **Multiple Frames can be bundled into a “Sequence”**
  - A Sequence can be used to transfer a large amounts of data – possibly up to multi-megabytes (instead of 2112 bytes for a single frame)
- **An interaction between two Fibre Channel ports is termed an “Exchange”**
  - An Exchange consists of a “Request” Sequence
  - And, a “Reply” Sequence
  - Many protocols (including SCSI and FC-NVMe) use an Exchange as a single command/response
  - Individual frames within the same Exchange are guaranteed to be delivered in-order
  - Each Exchange may be sent on a different path through the fabric
    - Different exchanges have no order guarantee
    - Allows Switches to pick most efficient route in Fabric

# FC Basics and Terminology (cont.)

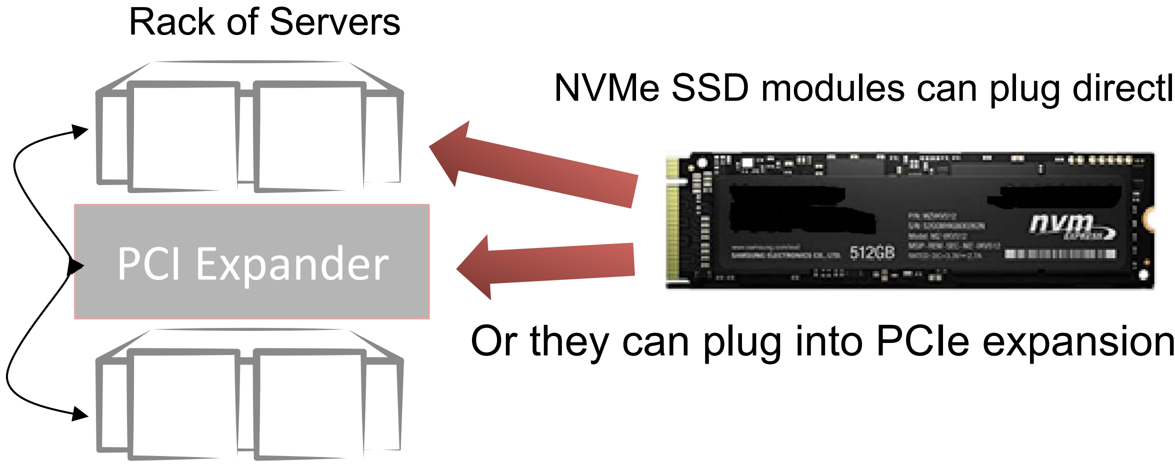
- **FCP (Fibre Channel Protocol) Data Transfer**
  - FCP is the Upper Layer Protocol originally specified to transport SCSI over Fibre Channel
  - The FCP Data Transfer protocol has since been adapted to transport FC-SB (FICON) and now FC-NVMe
  - Provides a high speed low latency transport

# NVMe Background

FIBRE CHANNEL INDUSTRY ASSOCIATION

# Three NVMe phases

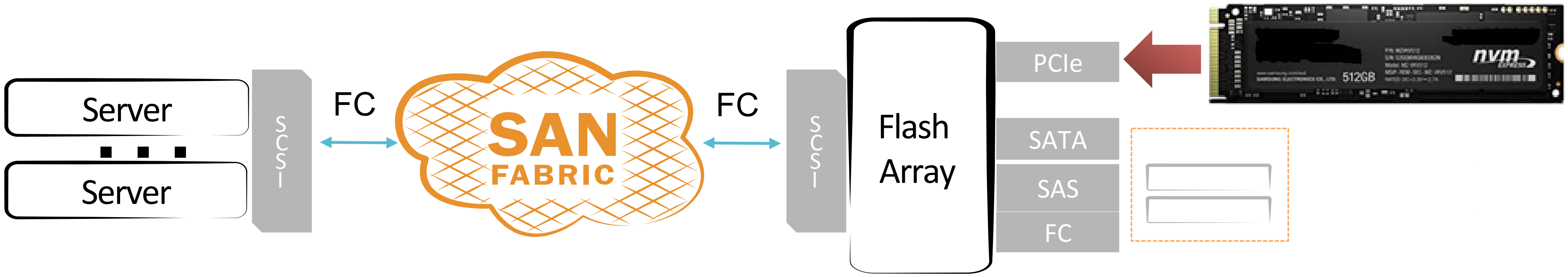
1 Basic (PCIe-based) NVMe in servers (last few years)



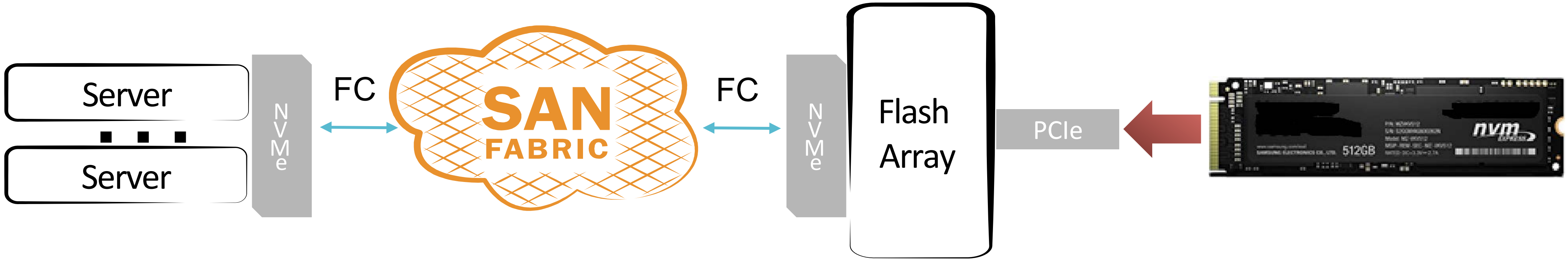
NVMe SSD modules can plug directly into servers

Or they can plug into PCIe expansion slots

2 Basic NVMe in storage backend (production!)



3 NVMe over Fabrics (demoing now)



# NVMe Basics and Terminology

- NVMe-oF – Shorthand for NVMe over Fabrics
- Host – The system which issues I/O commands to a Subsystem
- Subsystem – A non-volatile memory storage device
- Capsule – A unit of information exchange used in NVMe over Fabrics.  Contains NVMe commands and responses.
- Discovery Controller – The controller which contains NVMe-oF specific discovery information

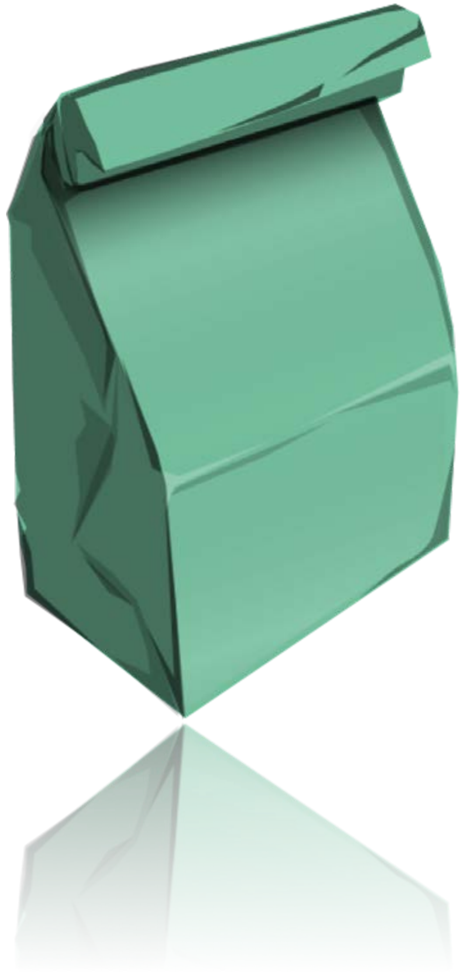
# NVMe Basics and Terminology (cont.)

- SQ (Submission Queue) – The queue used to submit I/O commands to the controller
- CQ (Completion Queue) – The queue used to indicate command completions for any return data and completion status by the controller
- Admin Queue – The queue used to submit Admin commands to the controller
- SQE (Submission Queue Entry) – A submission to the Submission or Admin Queue – Contains the command to be performed
- CQE (Completion Queue Entry) – A submission to the Completion Queue containing any returned data and completion status

# FC-NVMe

FIBRE CHANNEL INDUSTRY ASSOCIATION

# Take away from this section?

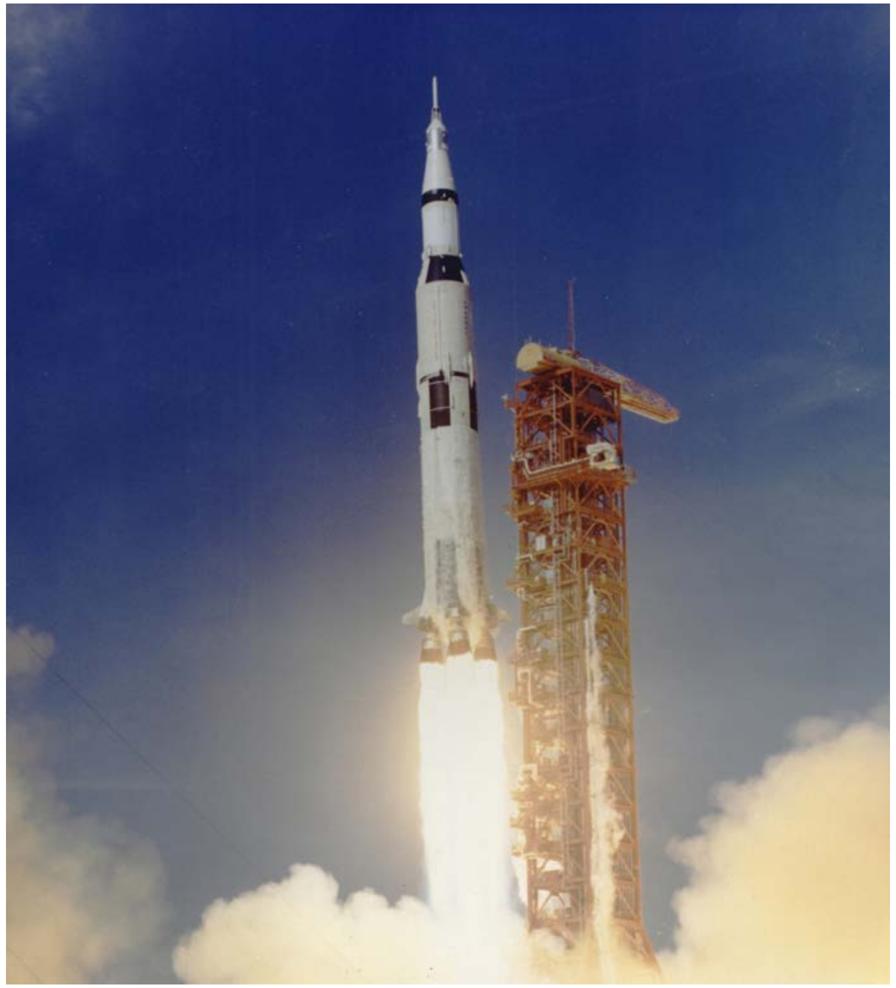


- **Most important part**
  - Deep Dive into how NVMe over Fabrics is mapped onto Fibre Channel
  - In depth look at how discovery is done in FC-NVMe
- **Next Section**
  - FC-NVMe use cases in the data center

# FC-NVMe

- **Goals**
  - Comply with NVMe over Fabrics Spec
  - High performance/low latency
  - Use existing HBA and switch hardware
    - Don’t want to require new ASICs to be spun to support FC-NVMe
  - Fit into the existing FC infrastructure as much as possible, with very little real-time software management
    - Pass NVMe SQE and CQE entries with no or little interaction from the FC layer
  - Maintain Fibre Channel metaphor for transportability
    - Name Server
    - Zoning
    - Management

# Performance

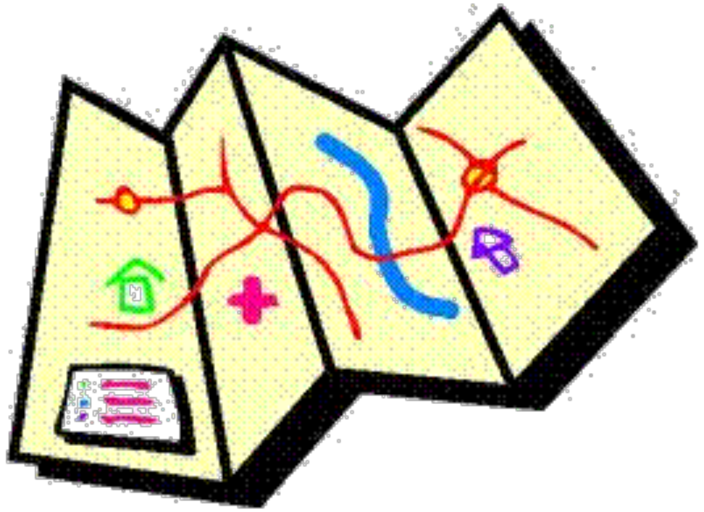


- **The Goal of High Performance/Low Latency**
  - Means that FC–NVMe needs to use an existing hardware accelerated data transfer protocol
  - FC does not have an RDMA protocol so FC-NVMe uses FCP for the data transfer protocol
    - Currently both SCSI and FC-SB (FICON) use FCP for data transfers
    - FCP is deployed as hardware accelerated in most (if not all) HBAs

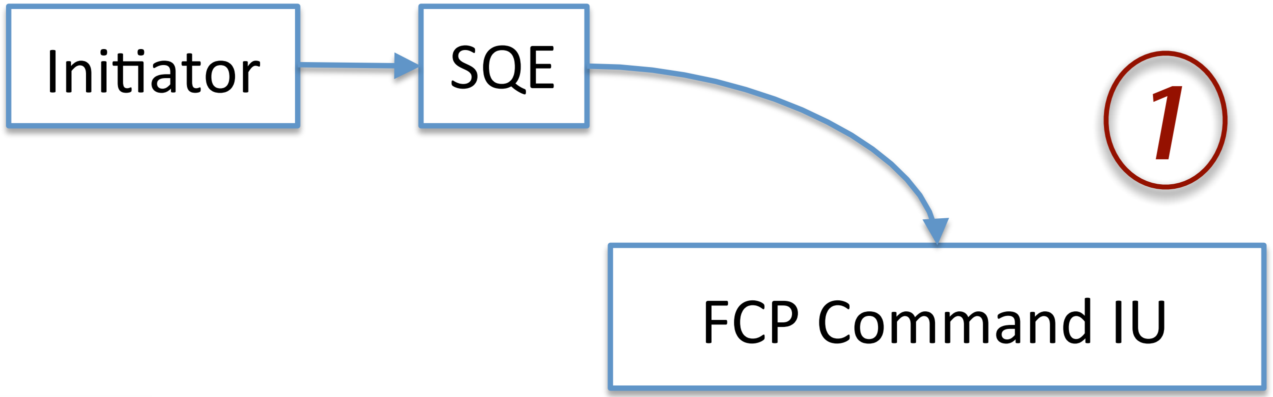
# FCP Data Transfer Mapping

- NVMe-oF capsules (i.e., commands, and responses) are directly mapped into FCP Information Units (IUs)

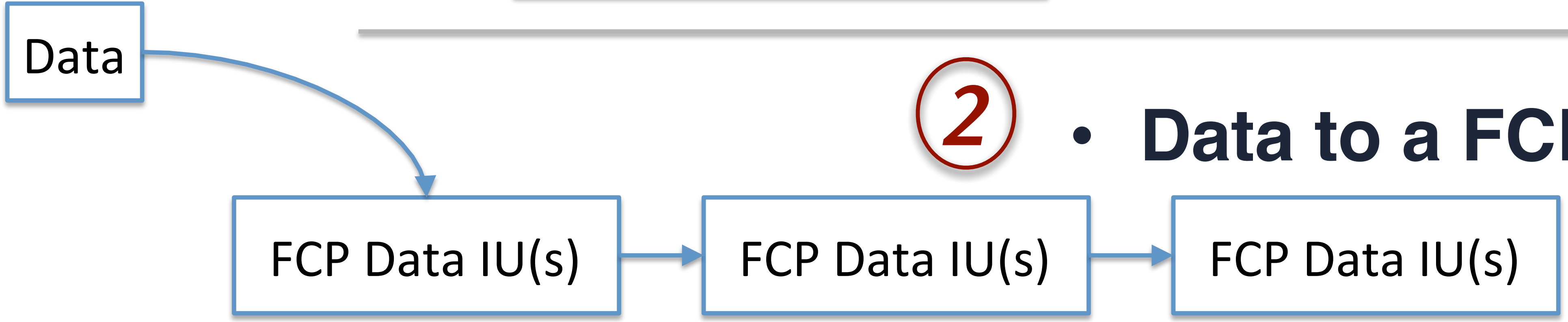
- NVMe-oF I/O operations are directly mapped to Fibre Channel Exchanges



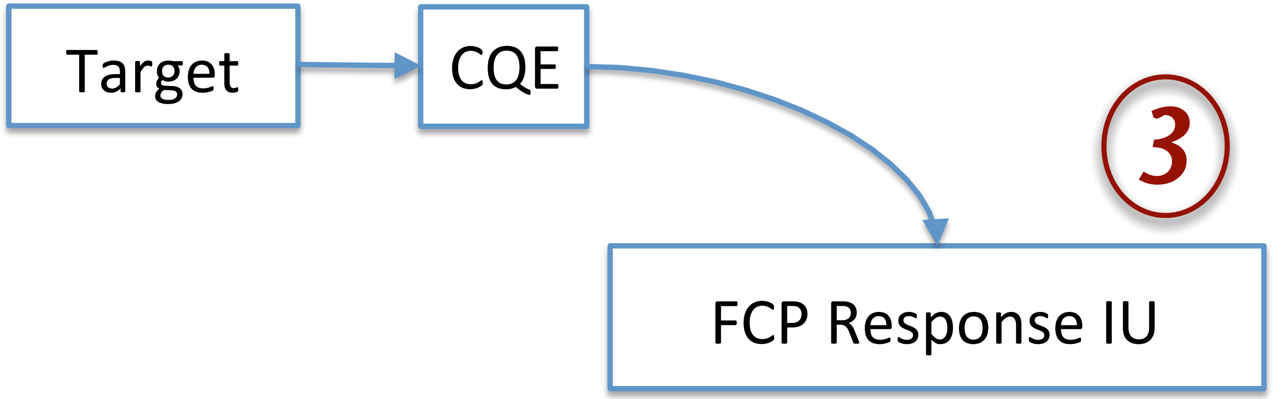
# FC-NVMe Information Units (IUs)



- **NVMe Submission Queue Entry (SQE) is mapped to a FCP Command IU**



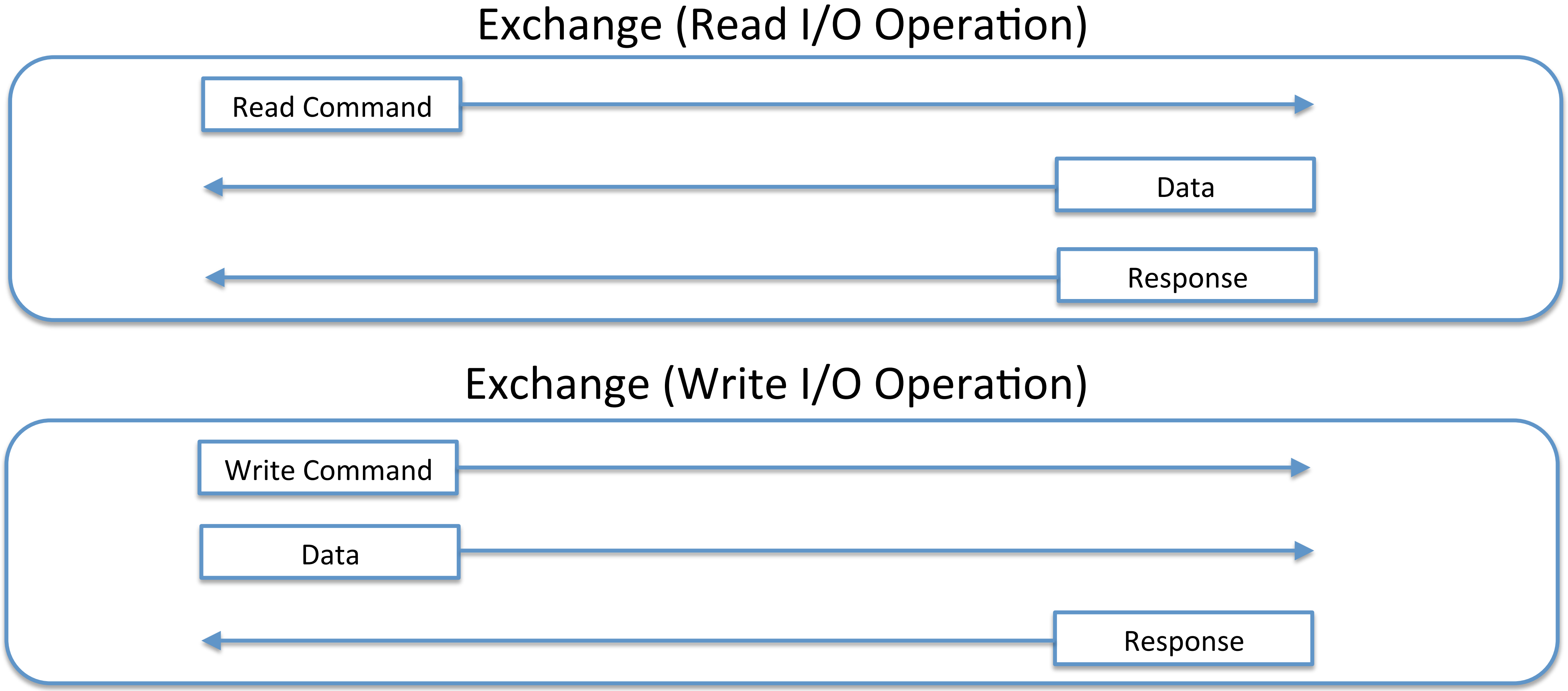
- **Data to a FCP Data IU**



- **NVMe Completion Queue Entry (CQE) to a FCP Response IU**

# I/O Operation

- **Transactions for a particular I/O Operation are bundled into an FC Exchange**



# FC-NVMe CMND IU Fields

- Format ID (FDh) and FC ID (28h)
  - uniquely identifies a FC-NVMe (vs. SCSI or FC-SB) CMND IU
- Flags – Indicates Read/Write
- NVMe Connection Identifier
  - Uniquely identifies FC-NVMe connection – Assigned during FC-NVMe connect
- Data Length – The total length of the data to be read or written
- Command Sequence Number – Incremented by one for each command sent – Allows commands to be processed in order they were sent (for those commands that require it)
- NVMe SQE – NVMe Submission Queue Entry sent down from NVMe layer

| Byte<br>Word | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | Format ID (FDh) | FC ID (28h) | CMND IU Length | |
| 1 | Reserved | | | Flags |
| 2 | NVMe Connection Identifier | | | |
| 3 | | | | |
| 4 | Command Sequence Number | | | |
| 5 | Data Length | | | |
| 6 – 21 | NVMe SQE (64 Bytes) | | | |
| 22 | Reserved | | | |
| 23 | Reserved | | | |

# FC-NVMe Responses

- **For FC-NVMe we have two types of responses**
  - Fast-path response – RSP_IU
    - Returned only for a good response
    - Allows for fast response processing
    - No ordering requirements
  - Extended response – ERSP_IU
    - Contains full NVMe response CQE
    - Potential slower processing time
    - Processed in-order as sent by Target/Subsystem as specified in NVMe-oF

# FC-NVMe Fast-path Response Wire Format

- **Response IU format required for FCP acceleration engine highest performance**
  - All zero response allows hardware accelerated path for completion

| Byte<br>Word | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | 00h | 00h | 00h | 00h |
| 1 | 00h | 00h | 00h | 00h |
| 2 | 00h | 00h | 00h | 00h |

# FC-NVMe Extended Response (ERSP) IU Fields

- ERSP Result – The response code indicating success or failure of the command
- Response Sequence Number – Incremented by one for each response sent – Allows responses to be processed in order they were sent
- Transferred Data Length – Indicates the total amount of data read or written by the command
- NVMe CQE – NVMe Completion Queue Entry sent down from NVMe layer

| Byte<br>Word | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | ERSP Result | Reserved | ESRP IU Length | |
| 1 | Response Sequence Number | | | |
| 2 | Transferred Data Length | | | |
| 3 | Reserved | | | |
| 4 – 7 | NVMe CQE (16 Bytes) | | | |

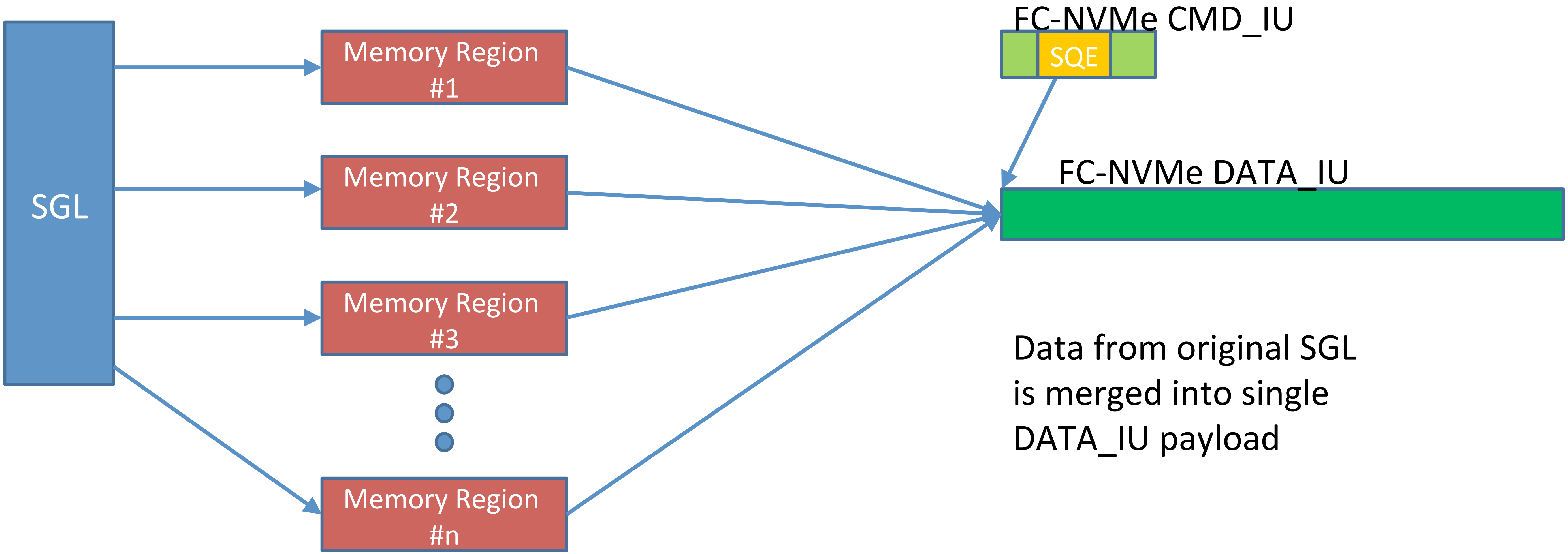
# Response rules

- **Response processing rules**
  - If the fast-path response is used
    - An Extended Response shall be sent for every n responses, where n is 1/10 of the SQ size – This allows the queue position pointers to be updated in a regular fashion
    - An Extended Response shall be sent when the SQ is 90% or more full – This prevents the SQ from not overwritten
    - An Extended Response shall be sent when the response comes from the last SQ entry (i.e., the SQ just became empty)
    - An Extended Response shall be sent if the completion status is non-zero (i.e., non-good completion status)
    - For the Extended Responses, Response Sequence Number shall be incremented for each Host<->(Subsystem,Controller,CQ)
    - All Extended Responses shall be processed in-order by the Initiator/Host
  - Fast-path responses are processed in the order they are received from the fabric (i.e., no reordering)

# Data transfer

- **All NVMe-oF data transfer are performed via SGL (Scatter Gather Lists)**
  - Scatter Gather List provide a list of memory locations for the HBA to read/write
- **For SGL data transfers**
  - The SGL list shall be converted into a single Data Block SGL within the NVMe Command/Response
    - Data is then transmitted via a DATA_IU
    - For writes, the local Host HBA may use the original SGL as a DMA gather list
    - For reads, the local Host HBA may retain the original SGL in order to direct place the received data

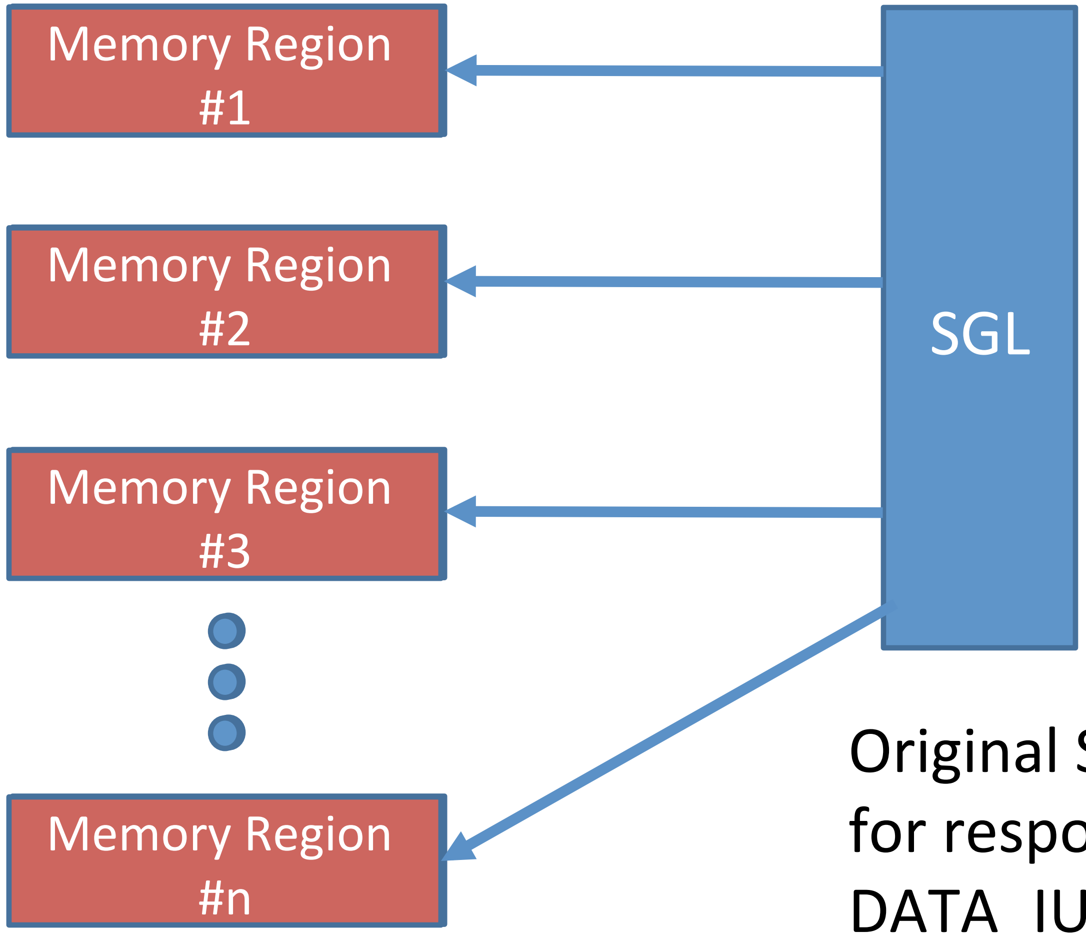
# SGL write example



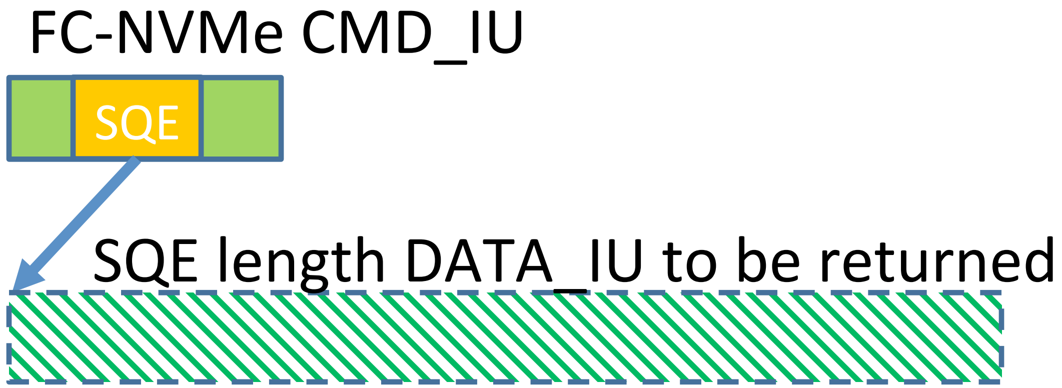
Data from original SGL is merged into single DATA_IU payload

# SGL read example

NVMe SQE command contains SGL pointing to local destination data buffers



Original SGL saved for response DATA_IU data

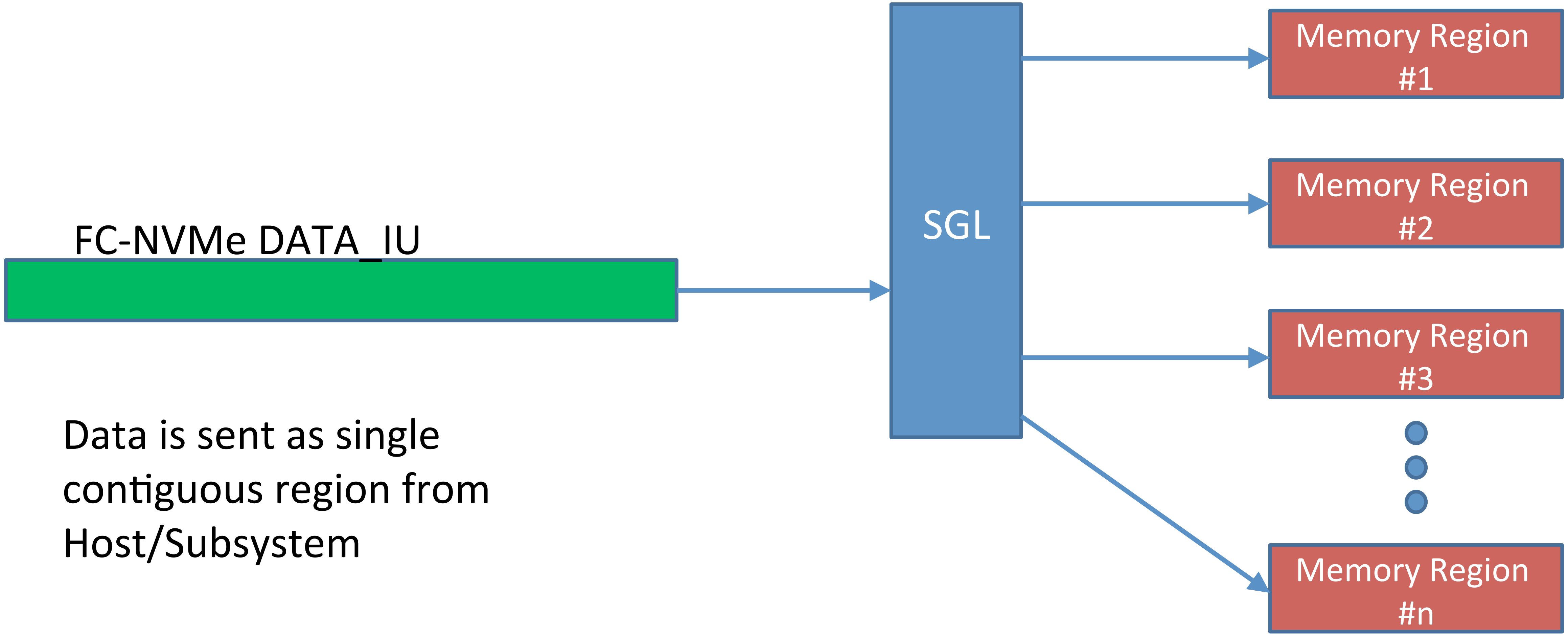


SGL List converted to single entry SGL for placement into CMD_IU

- Points to offset into response DATA_IU

# SGL read example (cont.)

In this example, local host HBA retains the SGL sent with original command for direct data placement

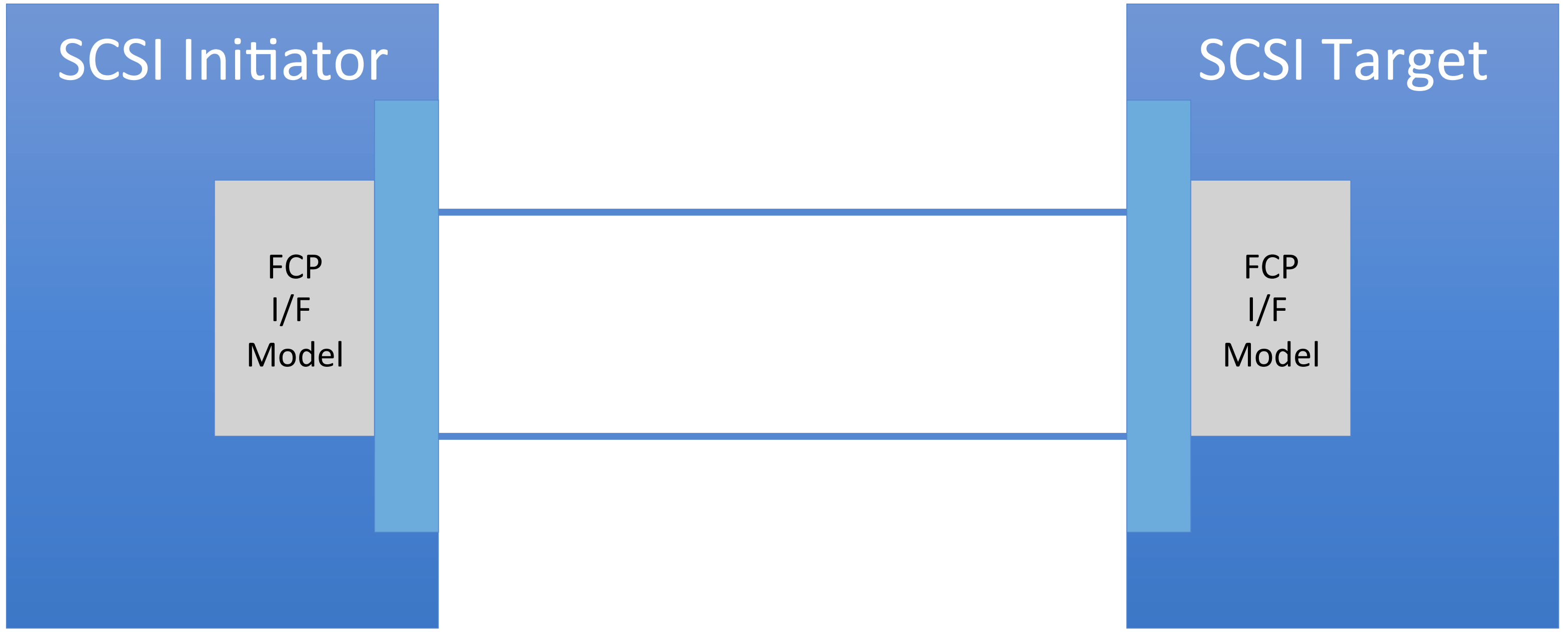


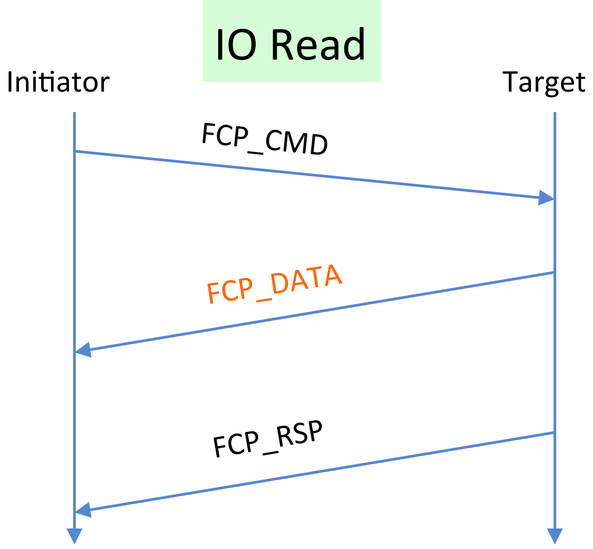
Data is sent as single contiguous region from Host/Subsystem

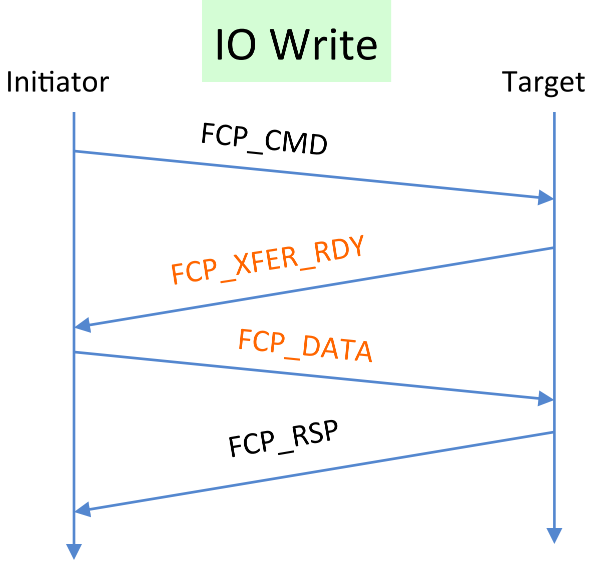
# Zero Copy

- **Zero-copy**
  - Allows data to be sent to user application with minimal copies
- **RDMA is a semantic which encourages more efficient data handling, but you don't need it to get efficiency**
- **FC has had zero-copy years before there was RDMA**
  - Data is DMA'd straight from HBA to buffers passed to user
- **Difference between RDMA and FC is the APIs**
  - RDMA does a lot more to enforce a zero-copy mechanism, but it is not required to use RDMA to get zero-copy

# FCP Transactions

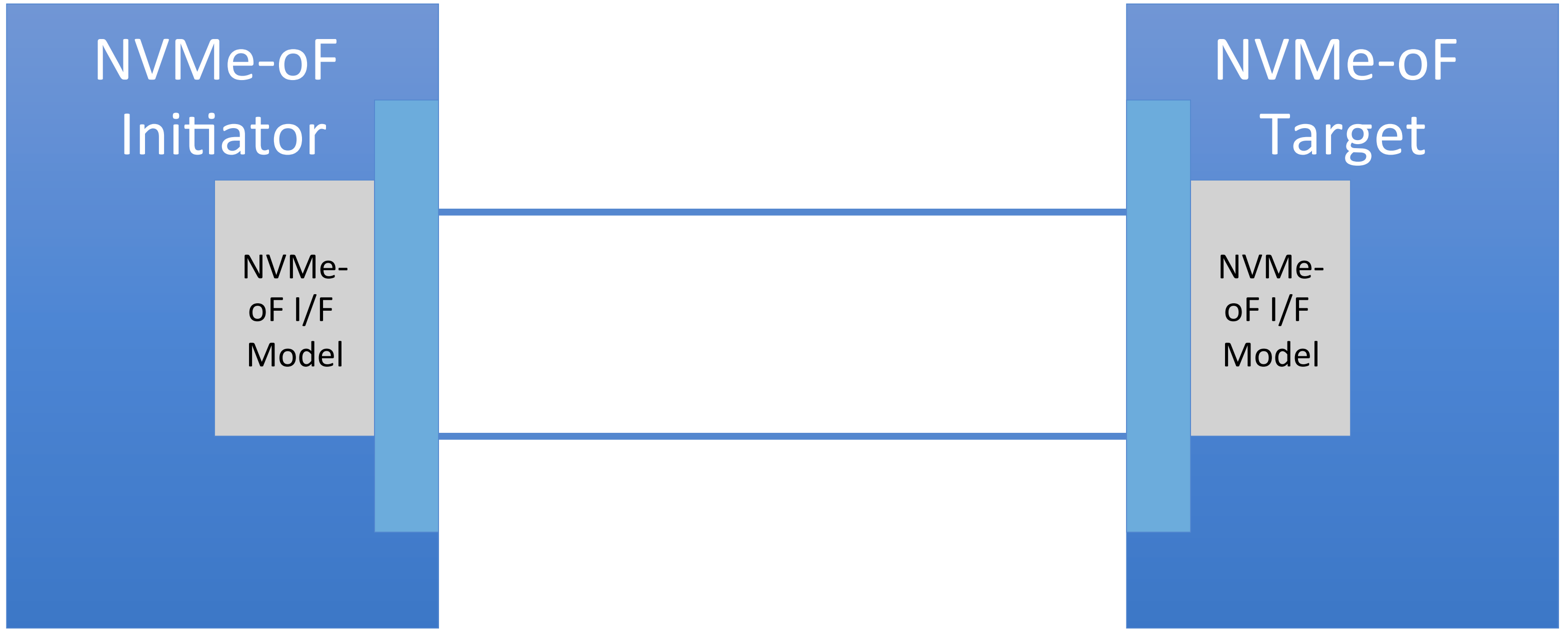


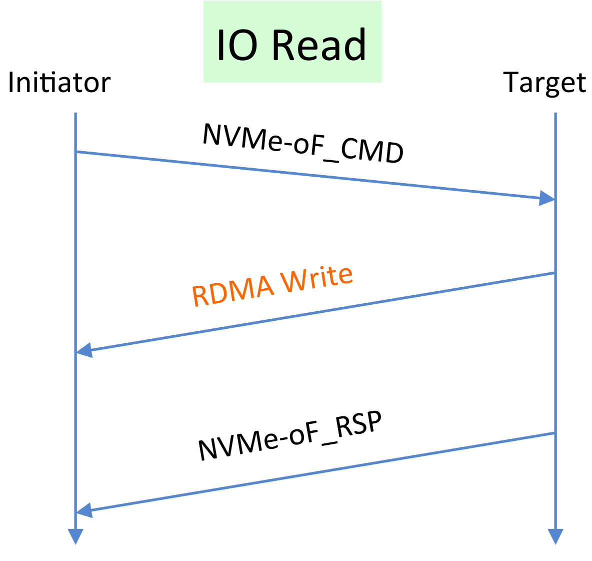


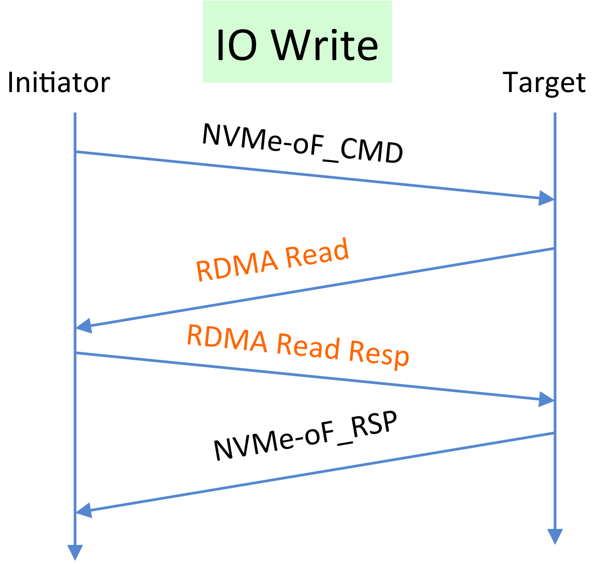


- **FCP Transactions look similar to RDMA**
  - – For Read
    - FCP_DATA from Target
  - – For Write
    - Transfer Ready and then DATA to Target

# NVMe-oF Protocol Transactions







- **NVMe-oF over RDMA protocol transactions**
  - RDMA Write
  - RDMA Read with RDMA Read Response

# Discovery

- **Two main types of FC topologies**
  - Point-to-point
    - No FC switches/Fabric
    - No FC Name Server
    - No Zoning or other Fabric services
    - Typically smaller configurations
  - Fabric based
    - 1 or more FC switches
    - Discovery based around the FC Name Server
    - Zoning in effect
    - Can be very large configurations (thousands of ports)

# Point-to-point discovery

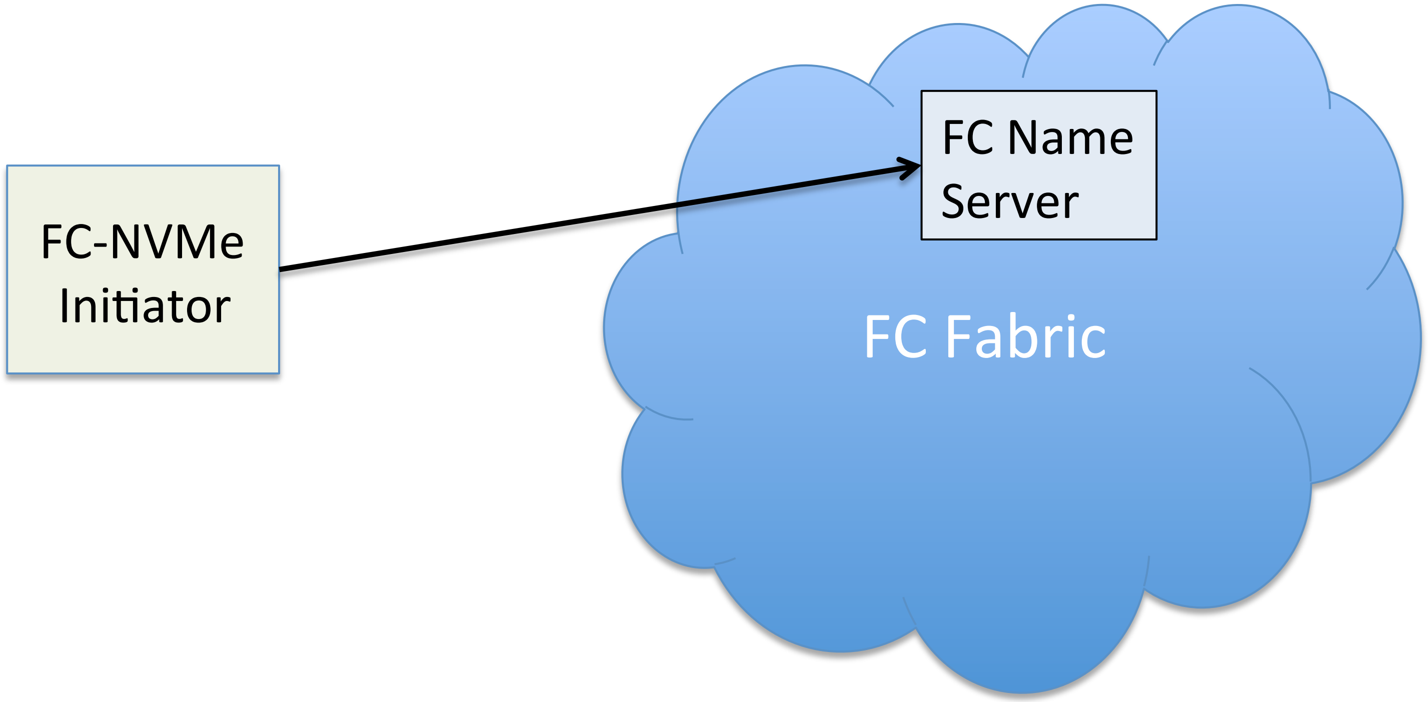
- **No FC Name Server**
  - Configurations are usually either self discovered or statically configured
  - NVMe Discovery Controller may be used for FC-NVMe
- **Note: Less common topology**

# Fabric Discovery

- **FC-NVMe Fabric Discovery uses both**
  - FC Name Server to identify FC-NVMe ports
  - NVMe Discovery Service to disclose NVMe Subsystem information for those ports
- **This dual approach allows each component to manage the area it knows about**
  - FC Name Server knows all the ports on the fabric and the type(s) of protocols they support
  - NVMe Discovery Service knows all the particulars about NVMe Subsystems

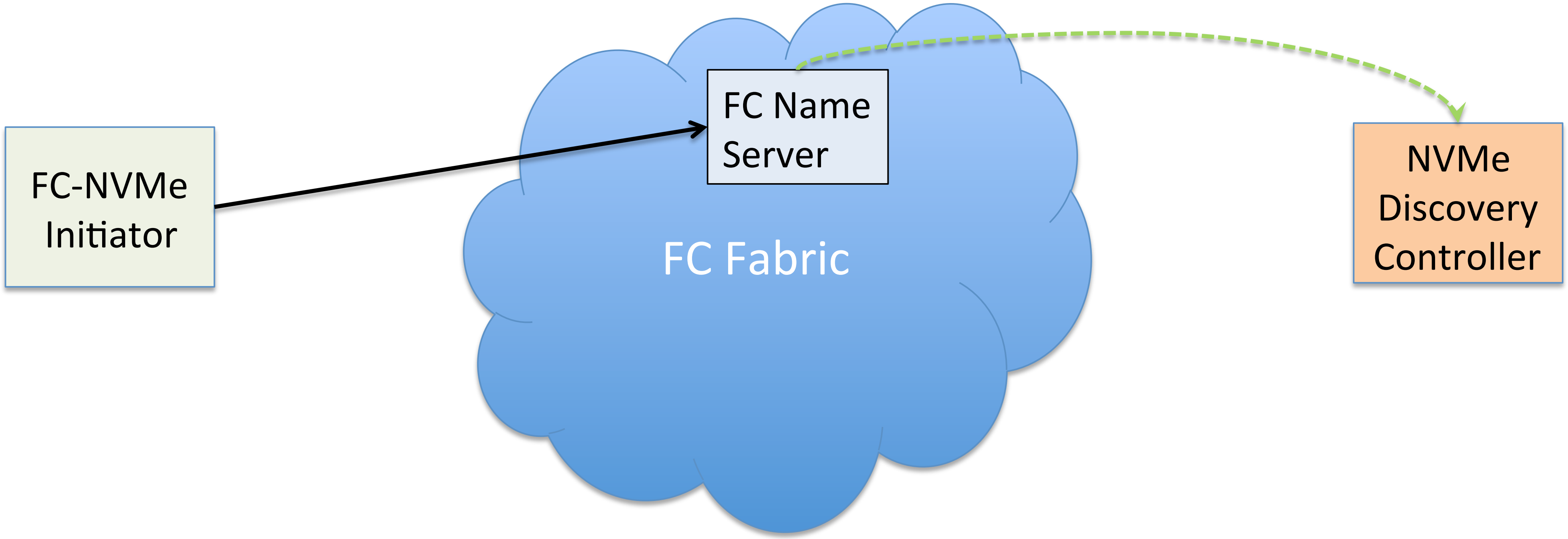
# FC-NVMe Discovery Example

- FC-NVMe Initiator connects to FC Name Server



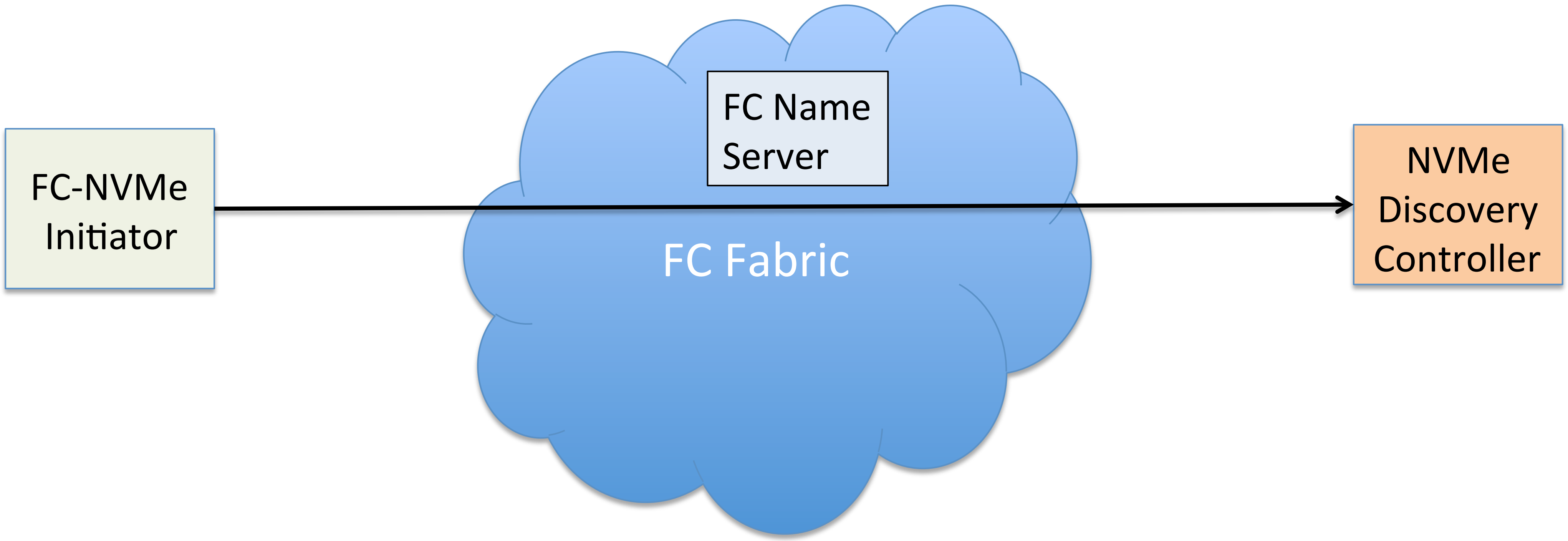
# FC-NVMe Discovery Example

- FC Name Server points to NVMe Discovery Controller(s)



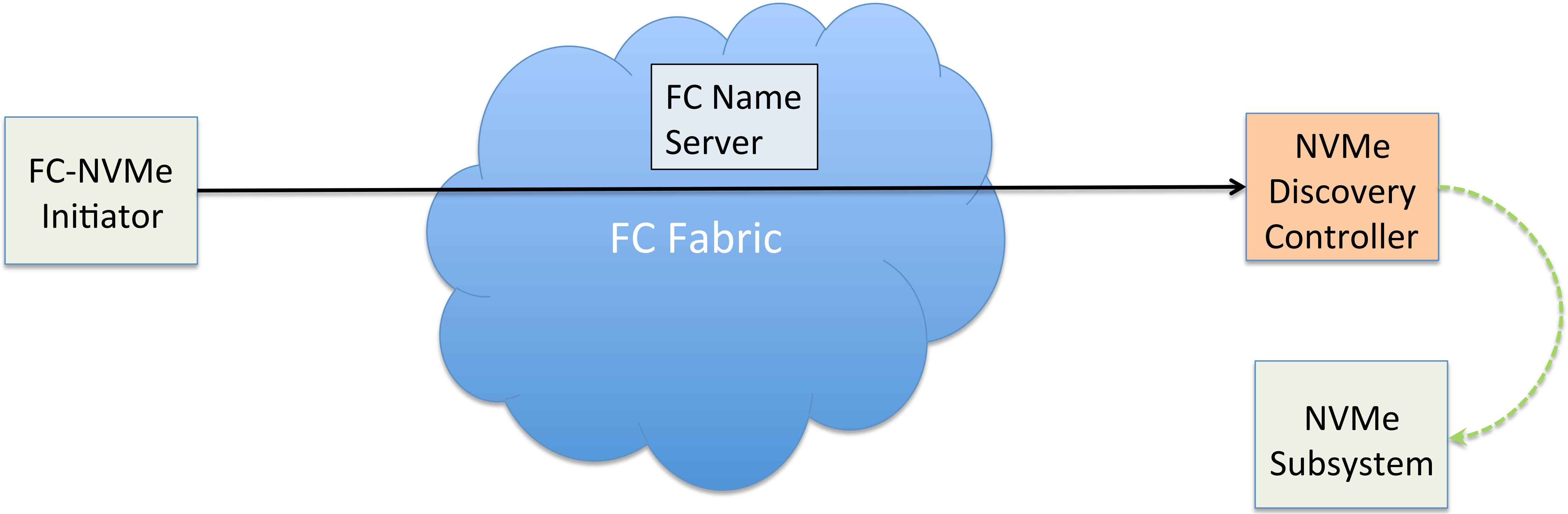
# FC-NVMe Discovery Example

- FC-NVMe Initiator connects to NVMe Discovery Controller(s)



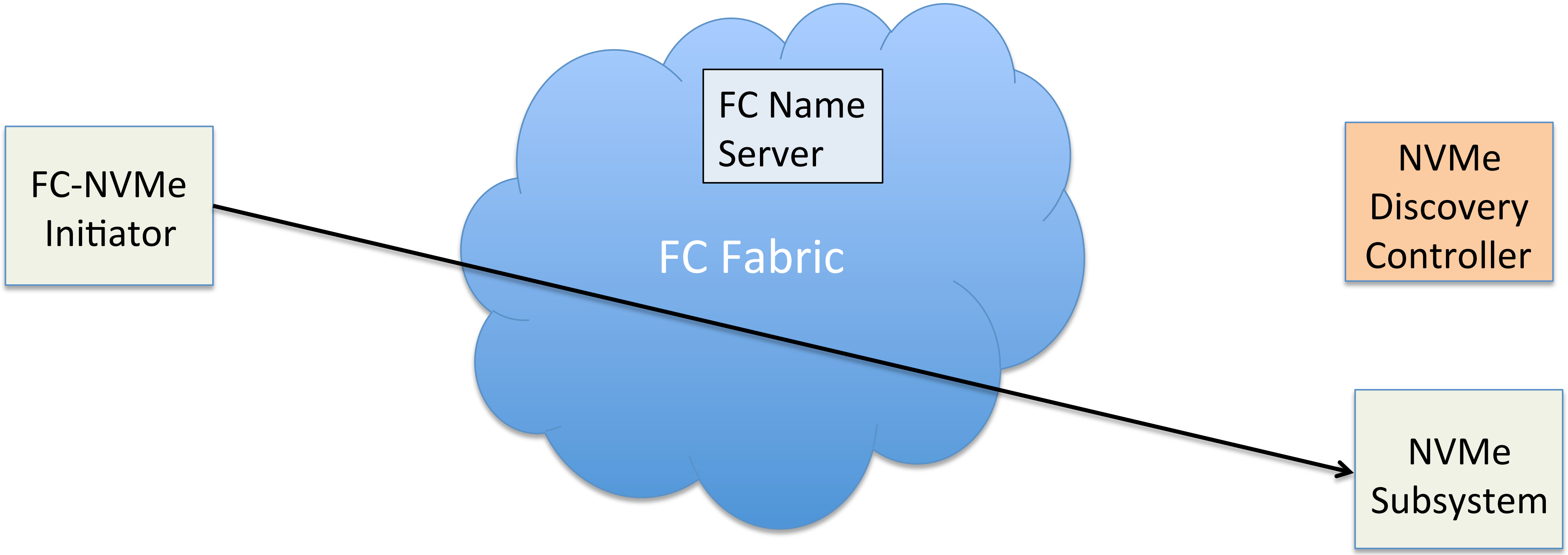
# FC-NVMe Discovery Example

- NVMe Discovery Controller(s) identify available NVMe Subsystems



# FC-NVMe Discovery Example

- FC-NVMe Initiator connects to NVMe Subsystem(s) to begin data transfers



# Zoning and Management

- **Of course, FC-NVMe also works with**
  - FC Zoning
  - FC Management Server and other FC Services

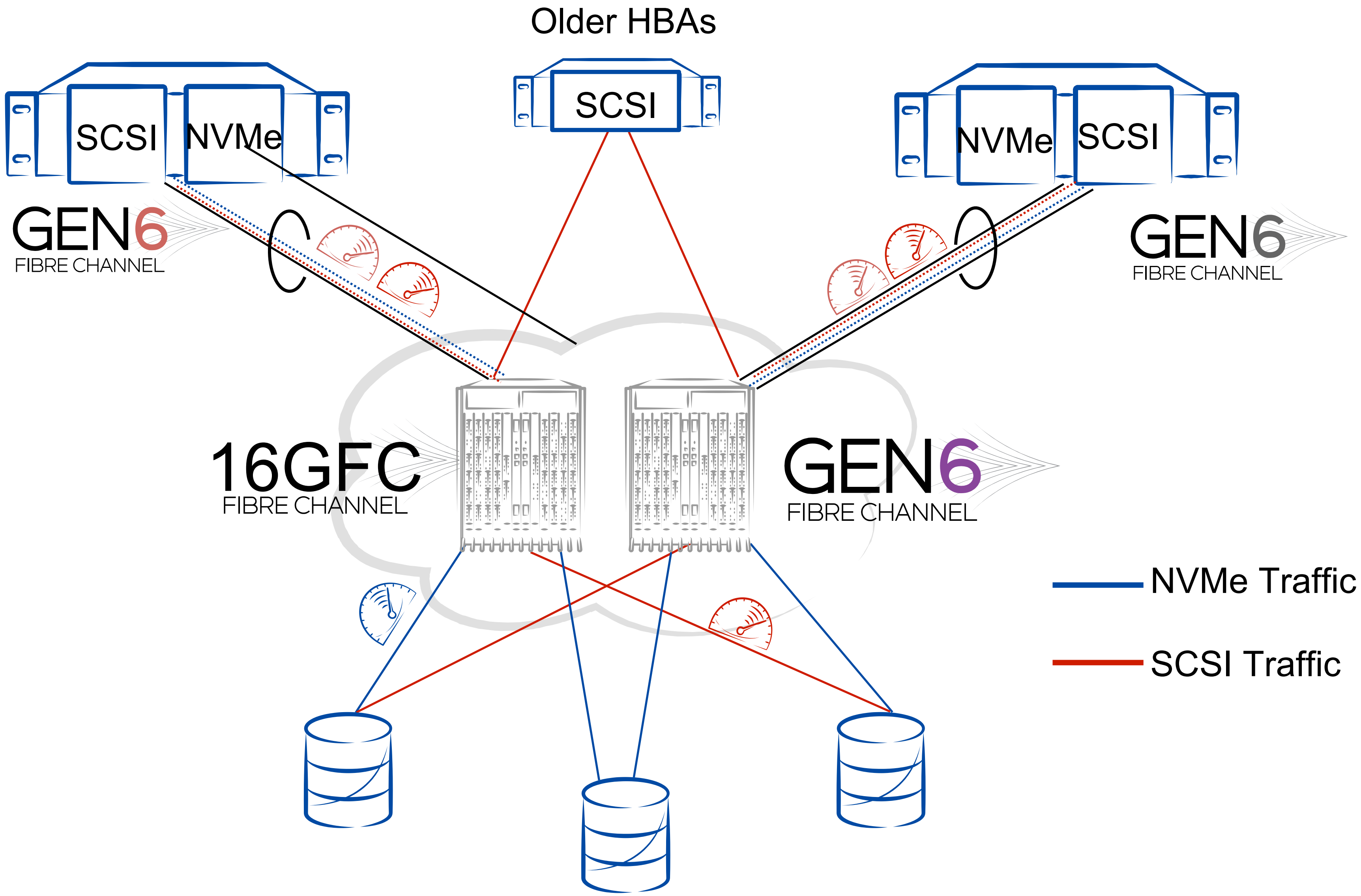
# Implications

FIBRE CHANNEL INDUSTRY ASSOCIATION

# Investment Protection: No Rip Out and Replace

Seamlessly add NVMe storage to existing SANs

- NVMe and SCSI coexist in the same server and SAN
- Dynamically migrate to NVMe on demand
- Transition applications and infrastructure at your own pace
- Existing 16G and Gen6 SANs can run NVMe without disruption

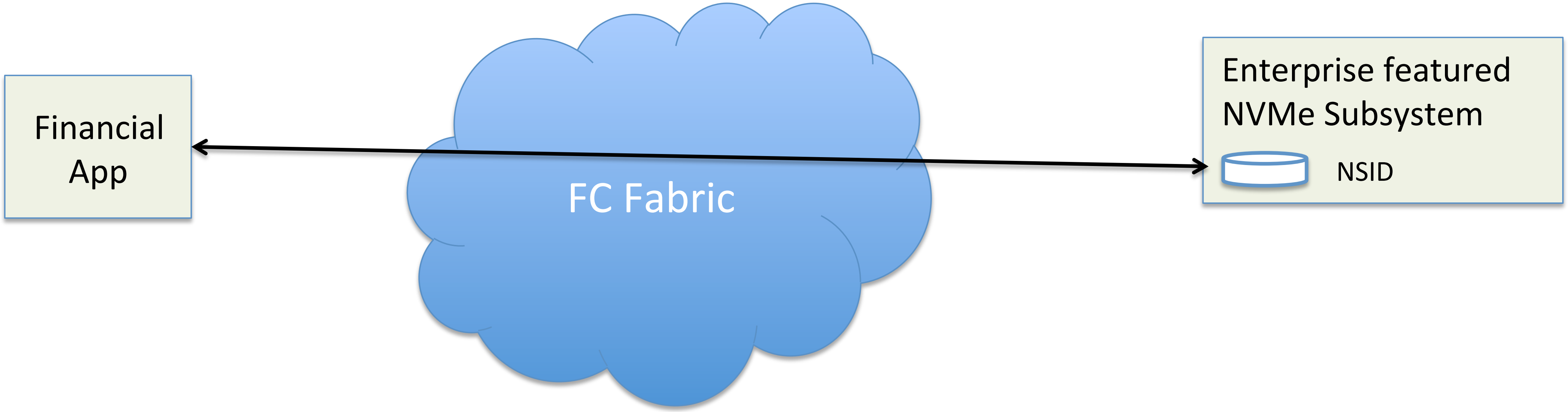


# Use Cases

FIBRE CHANNEL INDUSTRY ASSOCIATION

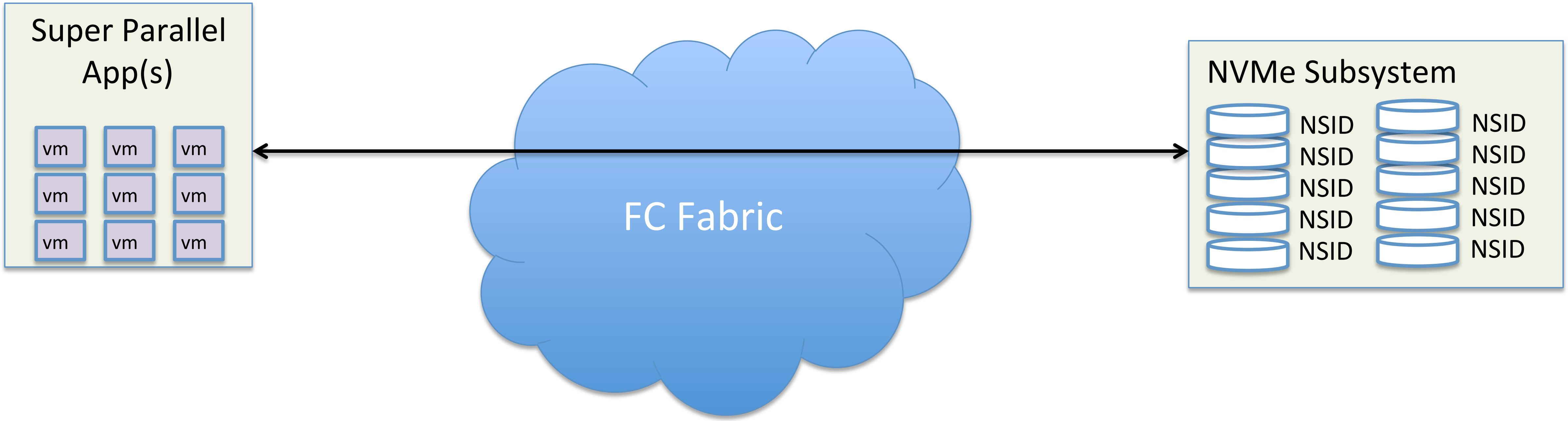
# Traditional latency sensitive storage apps

- NVMe reduces latency of enterprise storage features declines, the (server side) latency savings of the NVMe will motivate latency sensitive apps, like financial apps, to move to NVMe over FC



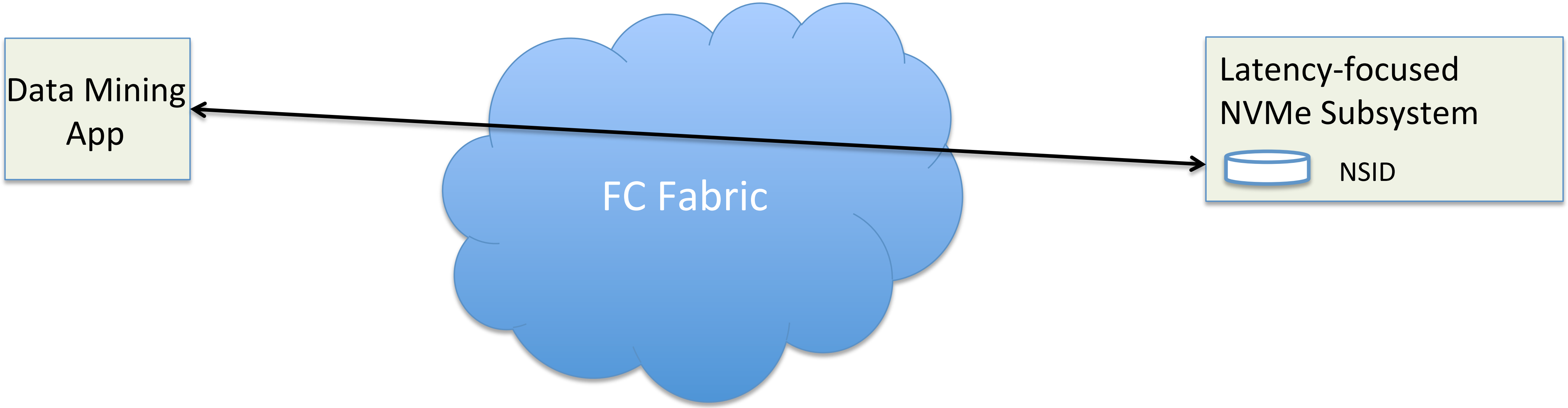
# Traditional IOPS sensitive storage apps

- The enhanced queuing capabilities of NVMe enable much more parallelism, even for existing apps. As the number of server cores and threads grows, and the number of VMs explodes, there is increasing need to exploit the potential parallelism in solid state targets. NVMe is designed to enable that.



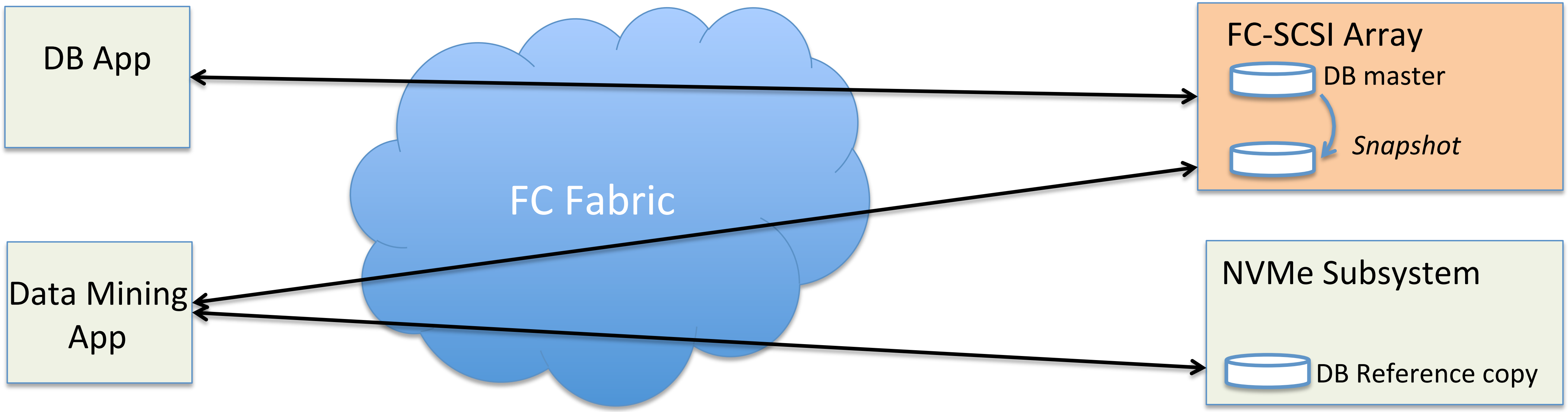
# New latency sensitive “memory” apps

- The ultra-low latency of SSDs is corresponding with new apps, such as data mining or machine learning.  These apps have voracious appetites for low latency persistent memory, beyond what fits in a server, and yet they often do not require the “enterprise features” of traditional enterprise storage.  We might call these “memory-oriented” applications.  They can leverage the low latency of Fibre Channel to access massive low latency NVMe arrays



# FC-NVMe / FC-SCSI dual protocol usage

- Database app maintains high value database on high SLA legacy array
- Data mining app requires super low latency reference image of DB
- Regularly Snapshot DB in legacy array
- Use Data mining server to copy snapshot to Ultra-low latency NSID
- Run Data mining application using low latency NSID reference copy



# Wrapping it up

FIBRE CHANNEL INDUSTRY ASSOCIATION

# FC-NVMe

- Wicked Fast!
- Builds on 20 years of the most robust storage network experience
- Can be run side-by-side with existing SCSI-based Fibre Channel storage environments
- Inherits all the benefits of Discovery and Name Services from Fibre Channel
- Capitalizes on trusted, end-to-end Qualification and Interoperability matrices in the industry

# Milestone

- **FC-NVMe completed 1st round of approval within the T11.3 committee in August!**
  - Ratification of document as technically stable

# 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 all the great questions we received today
- Follow us on Twitter @FCIAnews
- Join us for our next live FCIA webcast:

**Long-Distance Fibre Channel**

**October 10, 2017**

**10:00 am PT**

Register at … https://www.brighttalk.com/webcast/14967/277327

# Thank you!

FIBRE CHANNEL INDUSTRY ASSOCIATION