FC-NVMe Tutorial

About the presenter

- Presented by: Craig W. Carlson
  - Senior Technologist, Cavium
  - Member of SNIA Technical Council
  - Chair of FC-NVMe working group within T11
  - Chair T11.3 Committee on Fibre Channel Protocols
  - FCIA Board Member
- Thanks also to J. Metz of Cisco for contributing content

Agenda

- FC Refresher
- NVMe Refresher
- FC-NVMe
- Why Use FC-NVMe?
- Summary

What This Presentation Is

- A reminder of how Fibre Channel works
- A reminder of how NVMe over Fabrics work
- A high-level overview of Fibre Channel and NVMe, especially how they work together

What This Presentation Is *Not*

- A technical deep-dive on either Fibre Channel or NVMe over Fabrics
- Comprehensive (no boiling the ocean)
- A comparison between FC and other NVMe over Fabrics methods

Fibre Channel Refresher

What is Fibre Channel?

SNIA Global Education

- 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

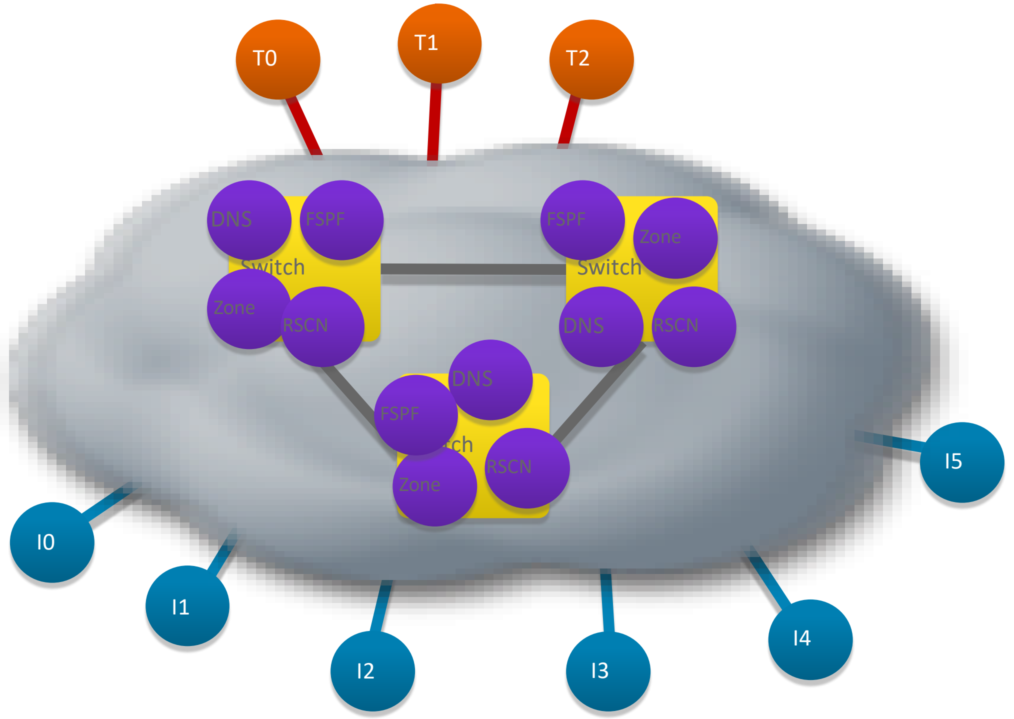
Design Requirements

Fibre Channel Storage Area Network (SAN)

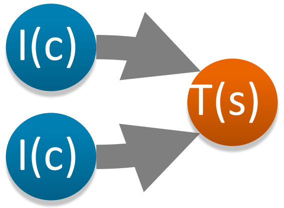
- Goal: Provide one-to-one connectivity
- Transport and Services are on same layer in same devices
- Well-defined end-device relationships (initiators and targets)
- Does not tolerate packet drop – requires lossless transport
- Only north-south traffic, east-west traffic mostly irrelevant

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

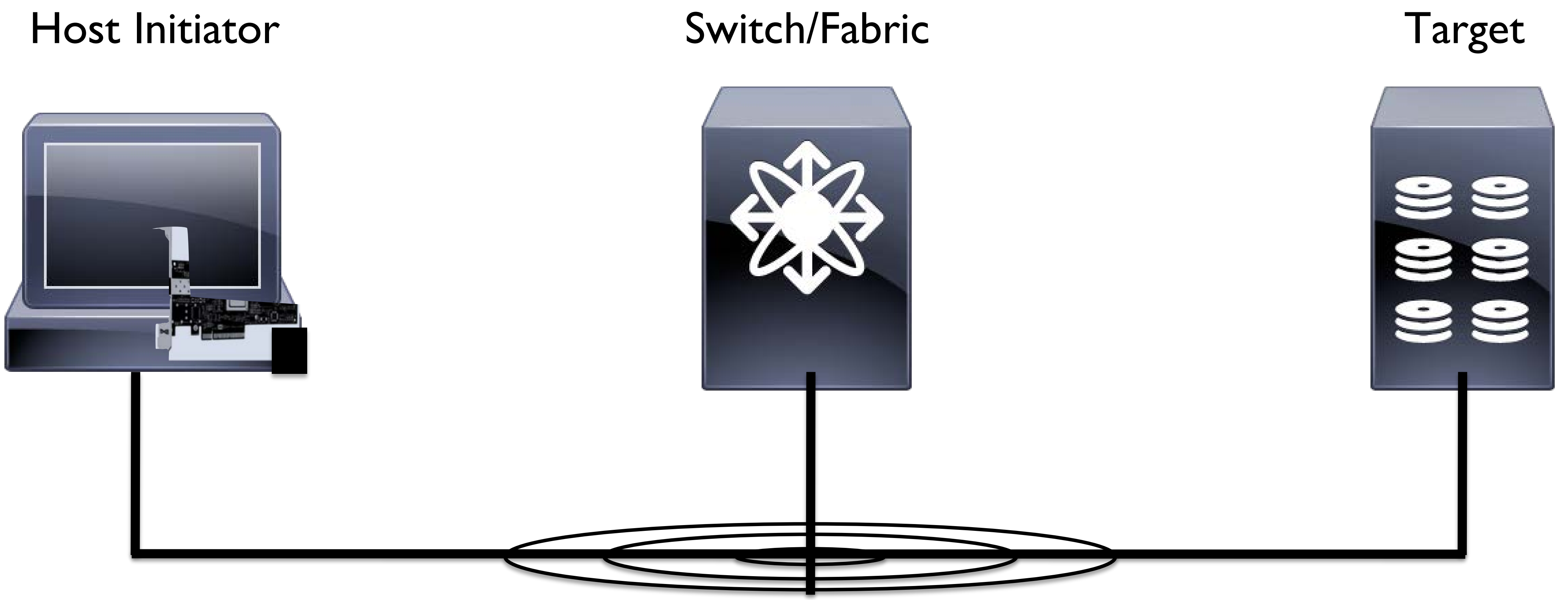


Fabric topology, services and traffic flows are structured



Client/Server Relationships are pre-defined

Design Elements



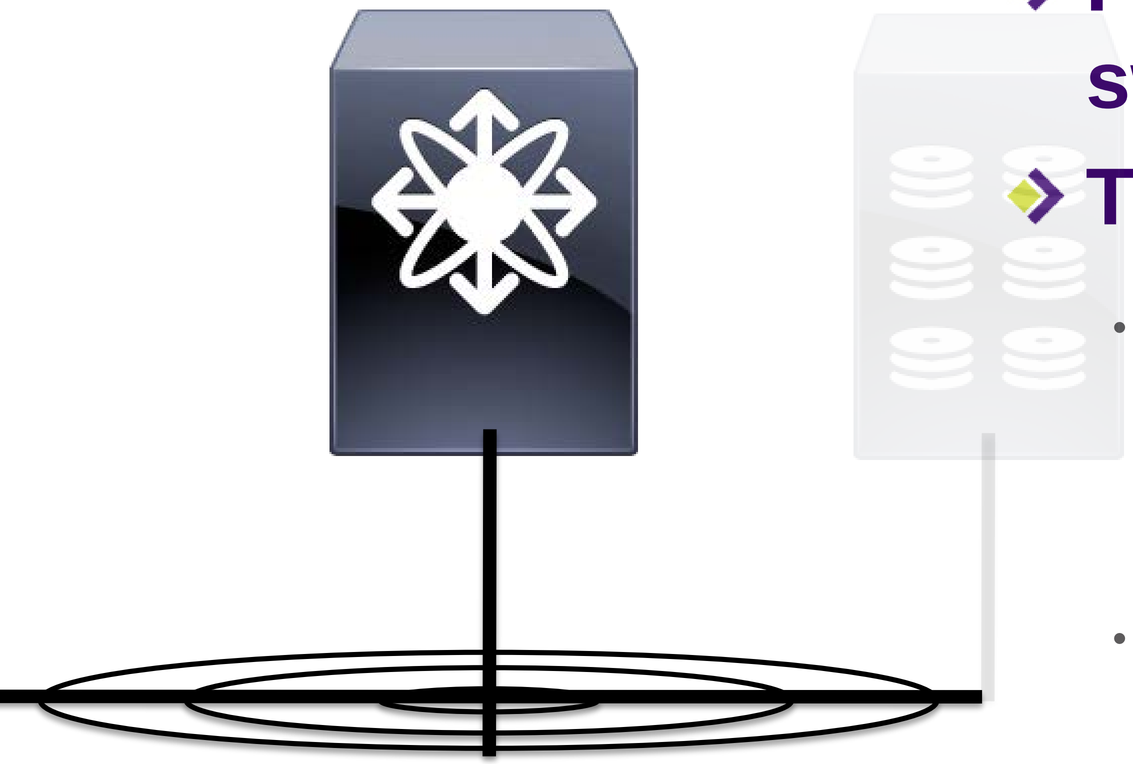
- Terminology that covers components or parts of the system
- Terminology that talks about the end-to-end system

Design Elements

- For FC the adapter which sits in a Host is called an HBA (Host Bus Adapter)
  - Equivalent to a NIC for Ethernet
- Where protocols such as NVMe or SCSI get encapsulated into a Fibre Channel Frame

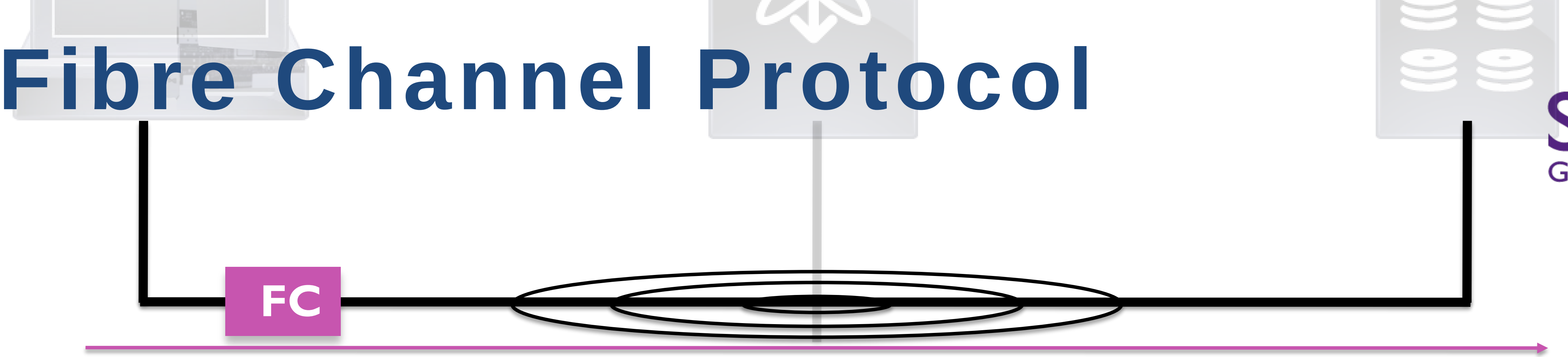
Design Elements

Switch/Fabric



- **Fabric intelligence is most often kept in the switch**
- **The Name Server**
  - Repository of information regarding the components that make up the Fibre Channel network
  - Name Server is implemented in the Fabric as a distributed redundant database
  - Components, like HBAs, can register their characteristics with the Name Server
  - Name server knows *everything* that goes on in the Fabric

The Fibre Channel Protocol



- **Fibre Channel typically uses an Unacknowledged Datagram Service**
  - Known as “Class 3”
  - Defined as a reliable datagram (connectionless) service
    - A class 3 frame will not be dropped unless an error occurs (i.e., bit error, or other unrecoverable error)

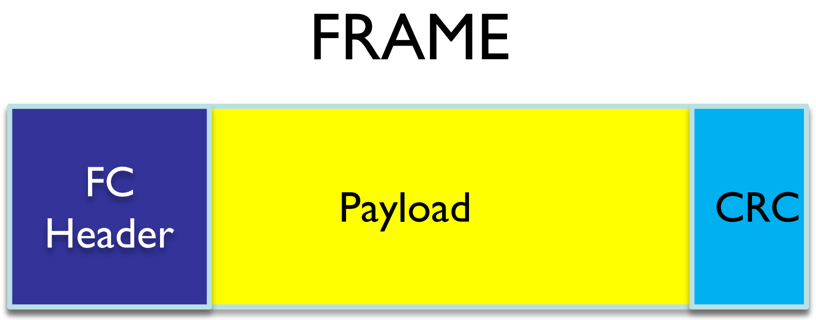
Frames, Sequences, and Exchanges

- **Fibre Channel data transfer has 3 fundamental constructs**
  - Frames – A “packet” of data
  - Sequences – A set of frames for larger data transfers
  - Exchanges – An associated set of commands and responses that make up a single command

Frames

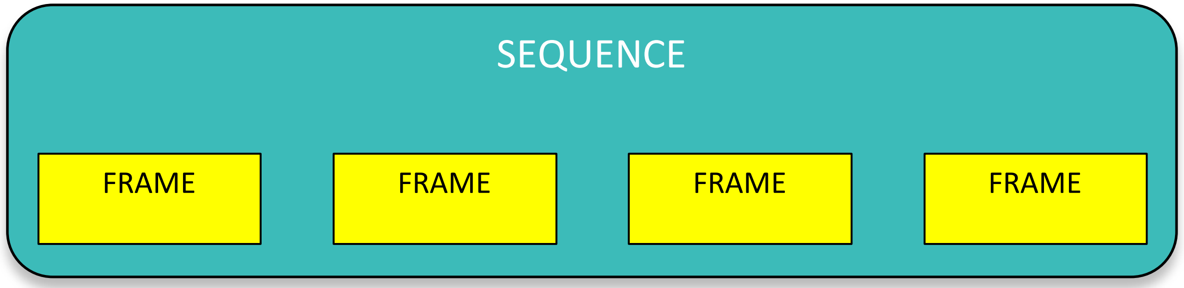
Each unit of transmission is called a “frame”

- A frame can be up to 2112 bytes
- Each frame consists of a FC Header, payload, and CRC



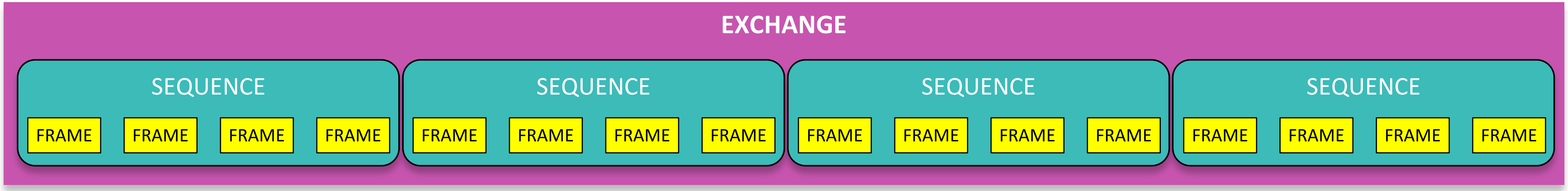
Sequences

- **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)



Exchanges

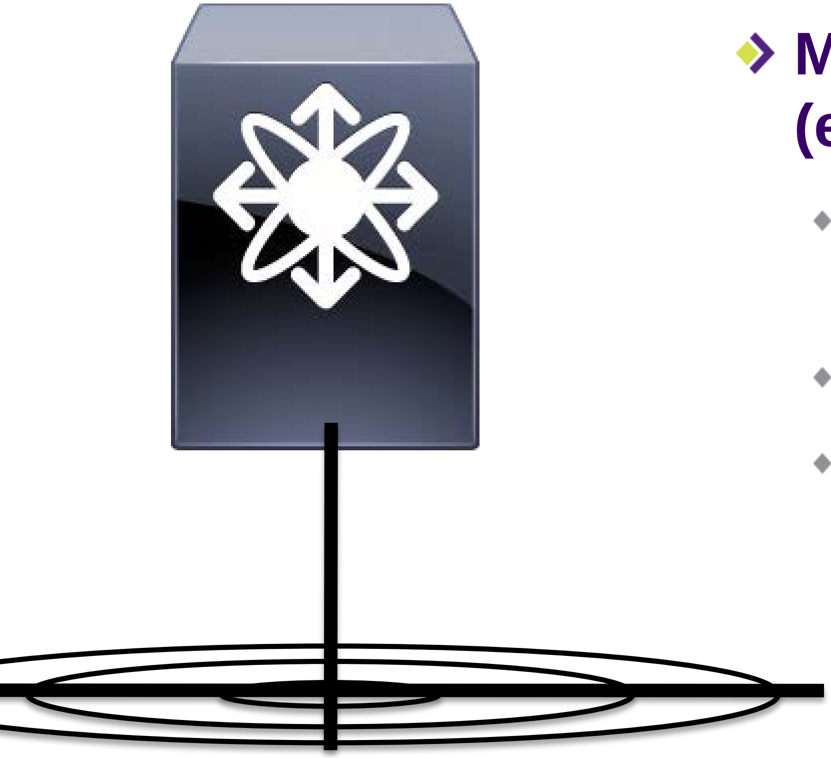
- **An interaction between two Fibre Channel ports is termed an "Exchange"**
  - Many protocols (including SCSI and FC-NVMe) use an Exchange as a single command/response
  - Individual frames within the same Exchange are guaranteed to be delivered in-order
  - Individual exchanges may take different routes through the fabric
    - This allows the Fabric to make efficient use of multiple paths between individual Fabric switches



*not to scale

Discovery in a FC Network

Switch/Fabric



- **Handled through the FC Name Server**
- **Many port attributes are automatically registered to the FC Name Server (e.g., Node WWN, Port WWN, Protocol types, etc.)**
  - Every Fibre Channel port and node has a hard-coded address called ***World Wide Name*** (WWN)
  - WWNN uniquely identify ***devices***
  - WWPN uniquely identify each ***port*** in a device

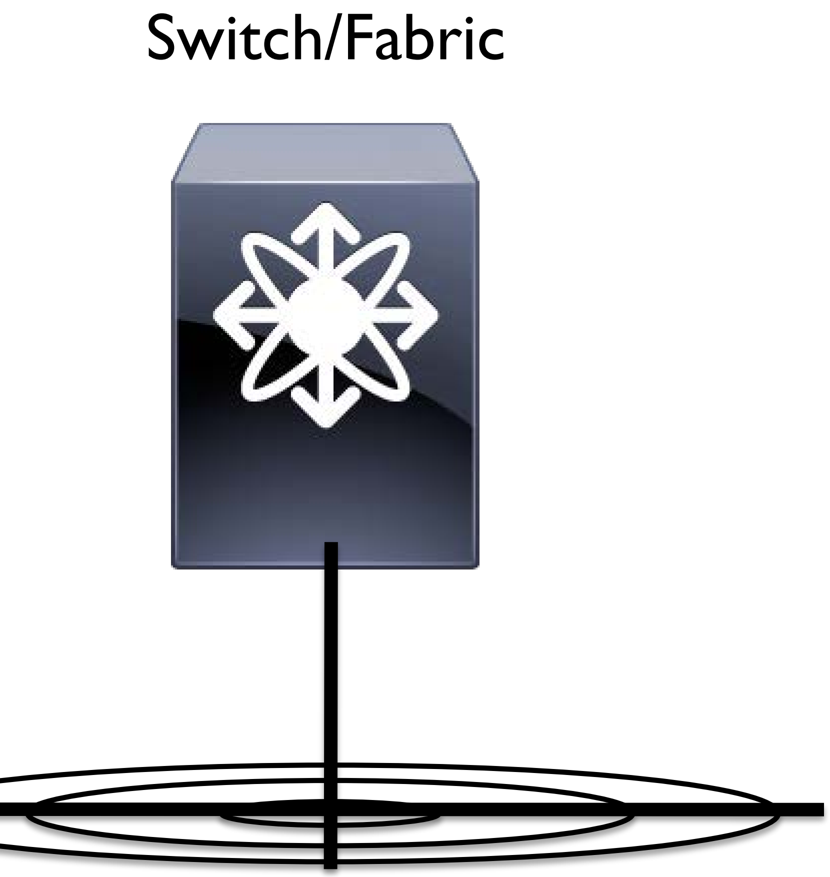
Example WWN

WWN	20:00:00:45:68:01:EF:25

Example WWNs from a Dual-Ported Device

WWNN	20:00:00:45:68:01:EF:25
WWPN A	21:00:00:45:68:01:EF:25
WWPN B	22:00:00:45:68:01:EF:25

Zones/Zoning



- **Zones provide added security and allow sharing of device ports**
- **Zoning allows a FC Fabric to control which ports get to see each other**
  - Zones can change frequently (e.g. backup)
- **Zoning is implemented by the switches in a Fabric**
  - Similar to ACLs in Ethernet switches
  - Central point of authority
  - Zoning information is distributed to all switches in the fabric
    - Thus all switches have the same zoning configuration
- **Standardized**

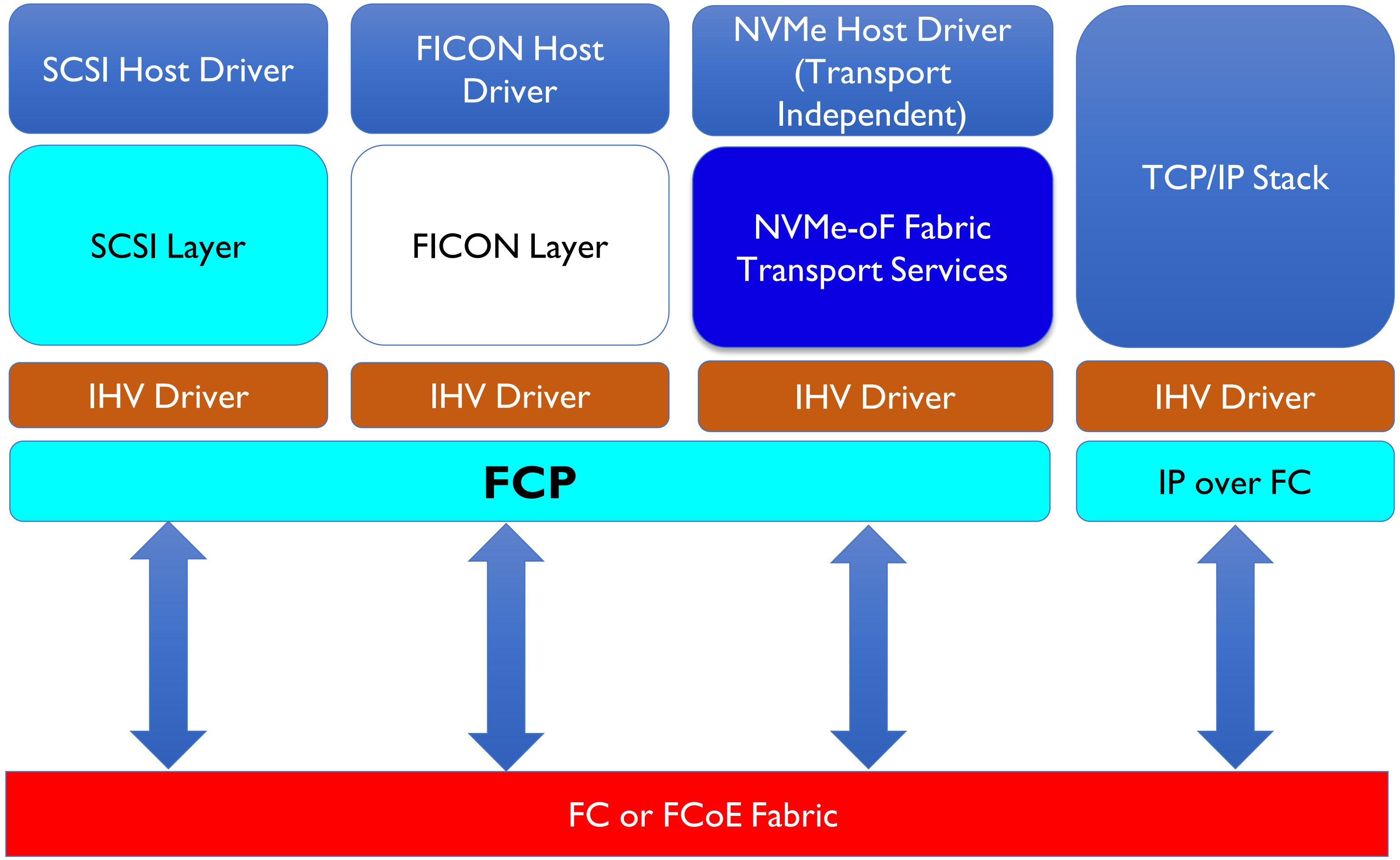
Fibre Channel Protocol

FC-4	Upper Layer Protocol Interface
FC-3	Common Services
FC-2	Framing and Flow Control
FC-1	Byte Encoding
FC-0	Physical Interface

- **Fibre Channel has layers, just like OSI and TCP**
- **At the top level is the Fibre Channel Protocol (FCP)**
  - Integrates with upper layer protocols, such as SCSI, FICON, and NVMe

What Is FCP?

- **What's the difference between FCP and "FCP"?**
  - FCP is a data transfer protocol that carries other upper-level transport protocols (e.g., FICON, SCSI, NVMe)
  - Historically FCP meant SCSI FCP, but other protocols exist now
- **NVMe "hooks" into FCP**
  - Seamless transport of NVMe traffic
  - Allows high performance HBA's to work with FC-NVMe



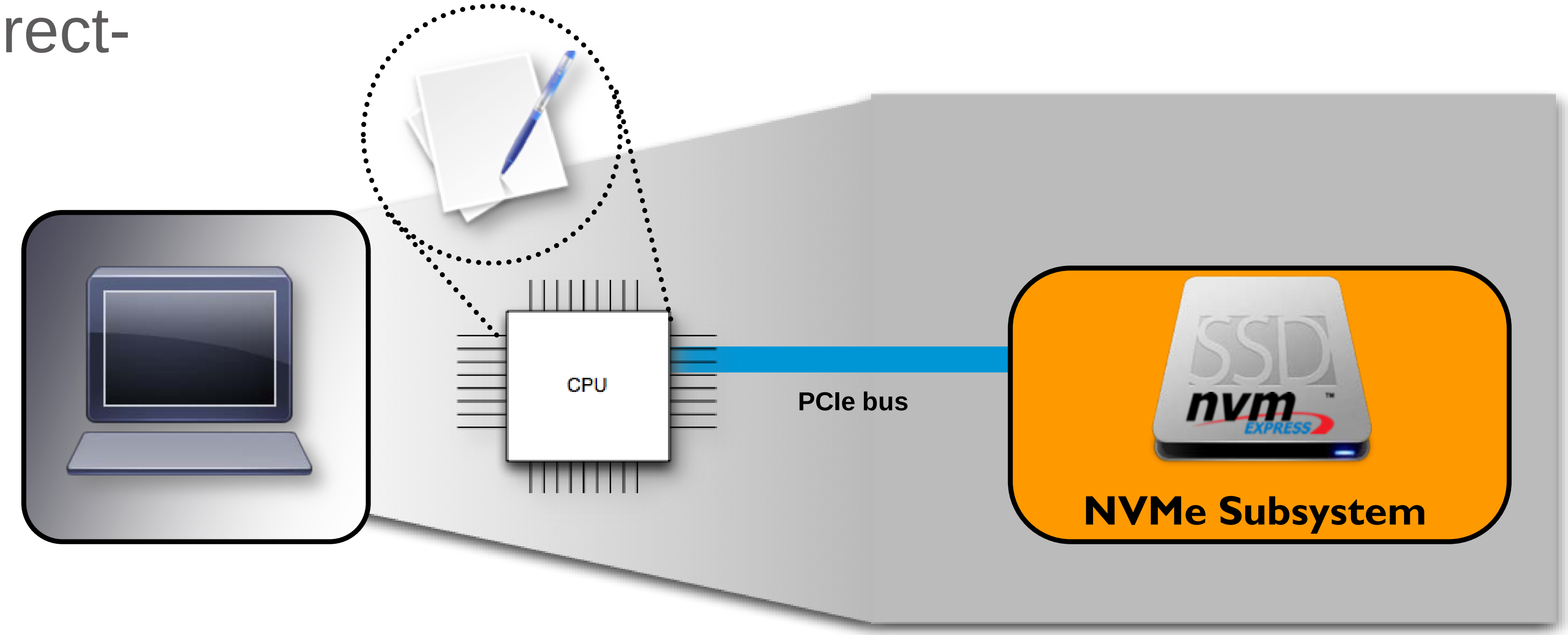
NVMe Refresher

What is Non-Volatile Memory Express (NVMe) and NVMe over Fabrics (NVMe-oF)?

SNIA Global Education

Non-Volatile Memory Express (NVMe)

- Began as an industry standard solution for efficient PCIe attached non-volatile memory storage (e.g., NVMe PCIe SSDs)
- Low latency and high IOPS direct-attached NVM storage



What is Non-Volatile Memory Express (NVMe) and NVMe over Fabrics (NVMe-oF)?

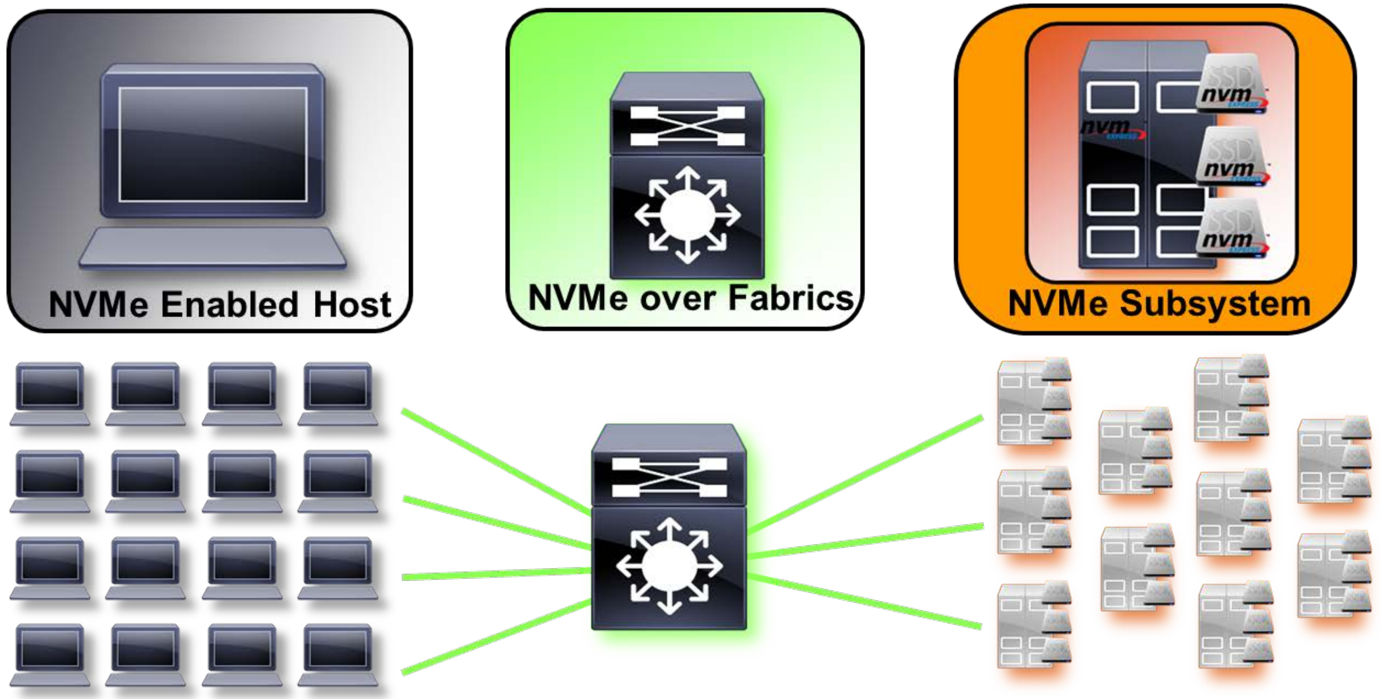
SNIA Global Education

Non-Volatile Memory Express (NVMe)

- Began as an industry standard solution for efficient PCIe attached non-volatile memory storage (e.g., NVMe PCIe SSDs)
- Low latency and high IOPS direct-attached NVM storage

NVMe over Fabrics (NVMe-oF)

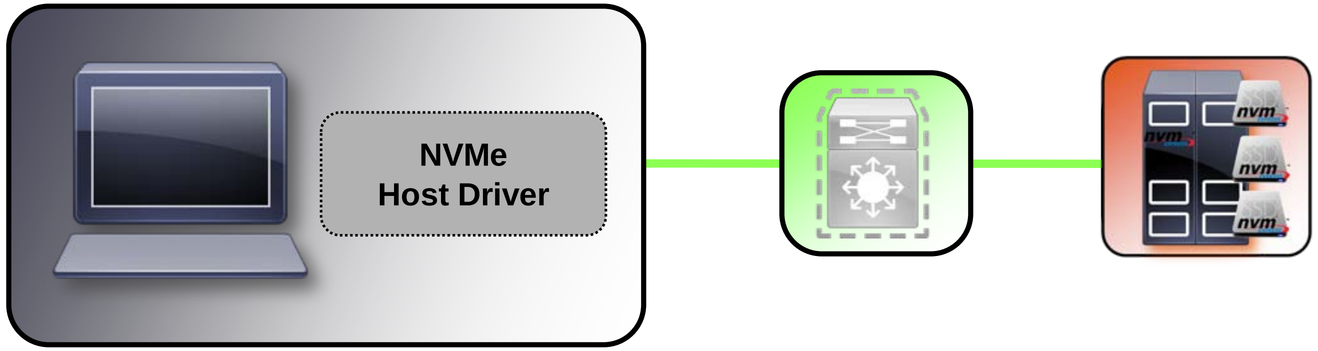
- Built on common NVMe architecture with additional definitions to support message-based NVMe operations
- Standardization of NVMe over a range Fabric types
  - Initial fabrics; RDMA(RoCE, iWARP, InfiniBand™) and Fibre Channel



NVMe Basics

- NVMe Drivers
- NVMe Subsystem
- NVMe Controller
- NVMe Namespaces & Media
- Queue Pairs

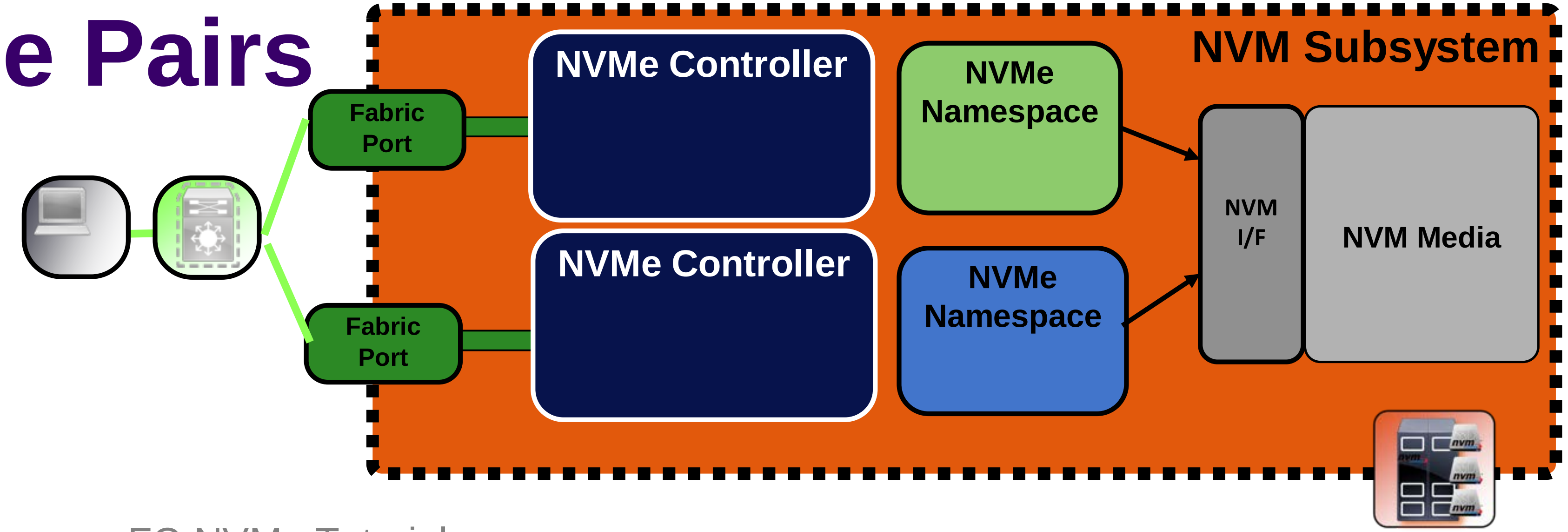
- In-box PCIe NVMe drivers in all major operating systems
- NVMe-oF will require specific drivers
  - FC-NVMe drivers will be provided by Fibre Channel vendors like always



NVMe Basics

- NVMe Drivers
- NVMe Subsystem
- NVMe Controller
- NVMe Namespaces & Media
- Queue Pairs

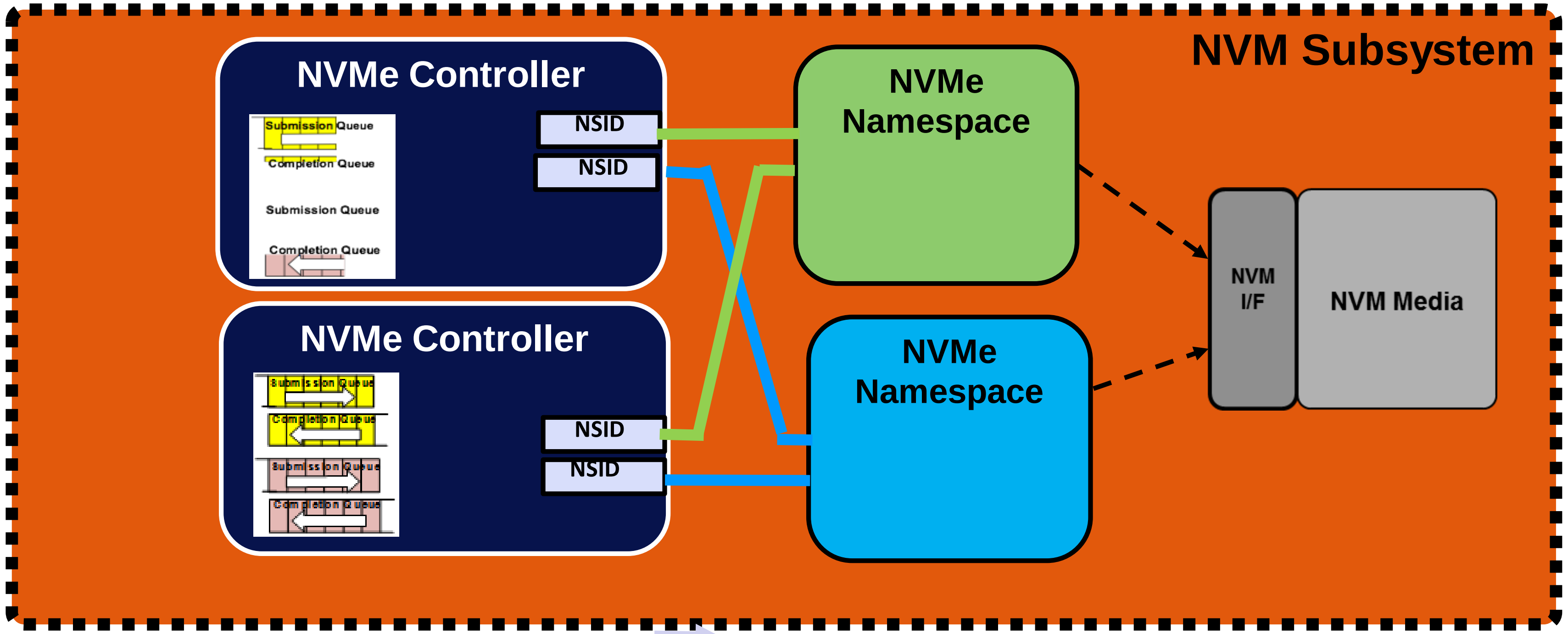
- Contains the architectural elements for NVMe targets
  - NVMe Controller
  - NVM Media
  - NVMe Namespaces
  - Interfaces



NVMe Basics

- NVMe Drivers
- NVMe Subsystem
- NVMe Controller
- NVMe Namespaces & Media
- Queue Pairs

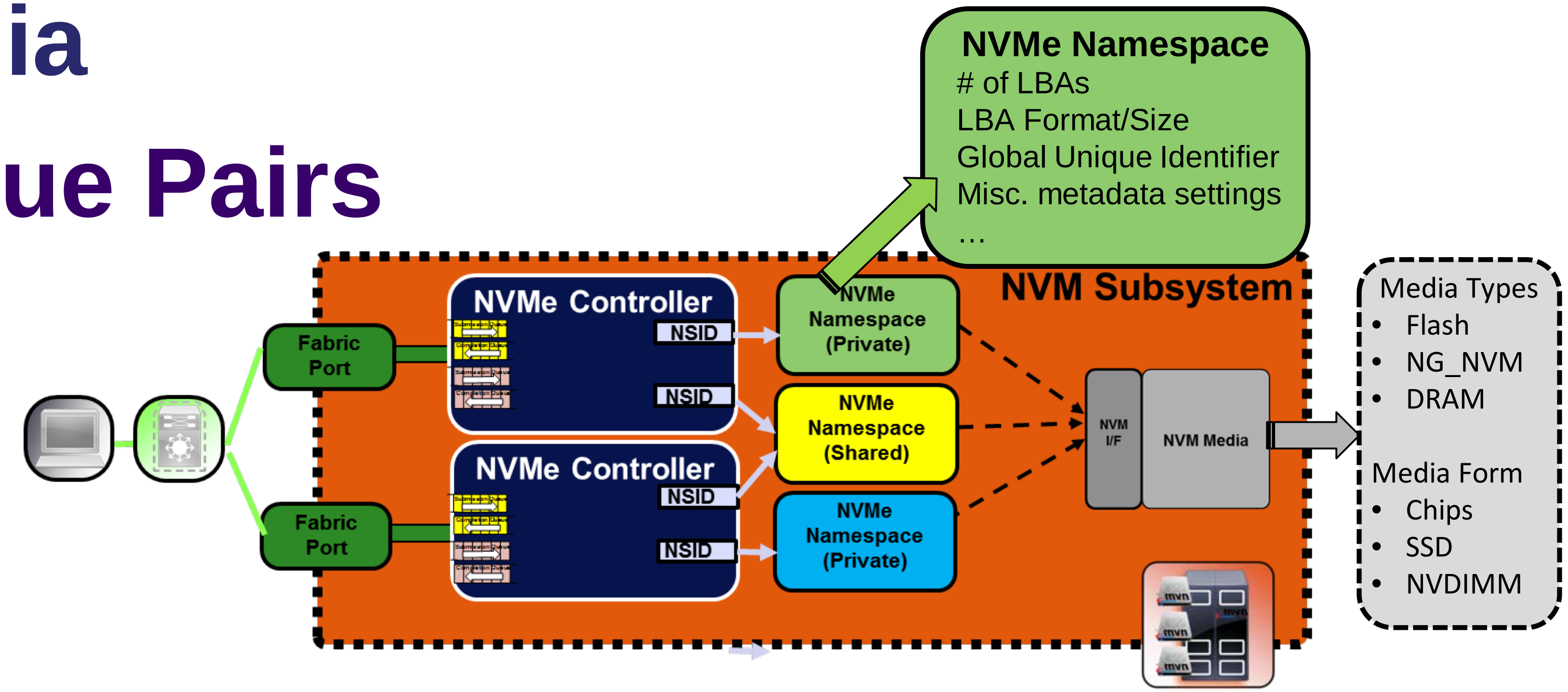
- NVMe Command Processing
- Access to NVMe Namespaces
  - Namespace ID (NSID) associates a Controller to Namespaces(s)



NVMe Basics

- NVMe Drivers
- NVMe Subsystem
- NVMe Controller
- NVMe Namespaces & Media
- Queue Pairs

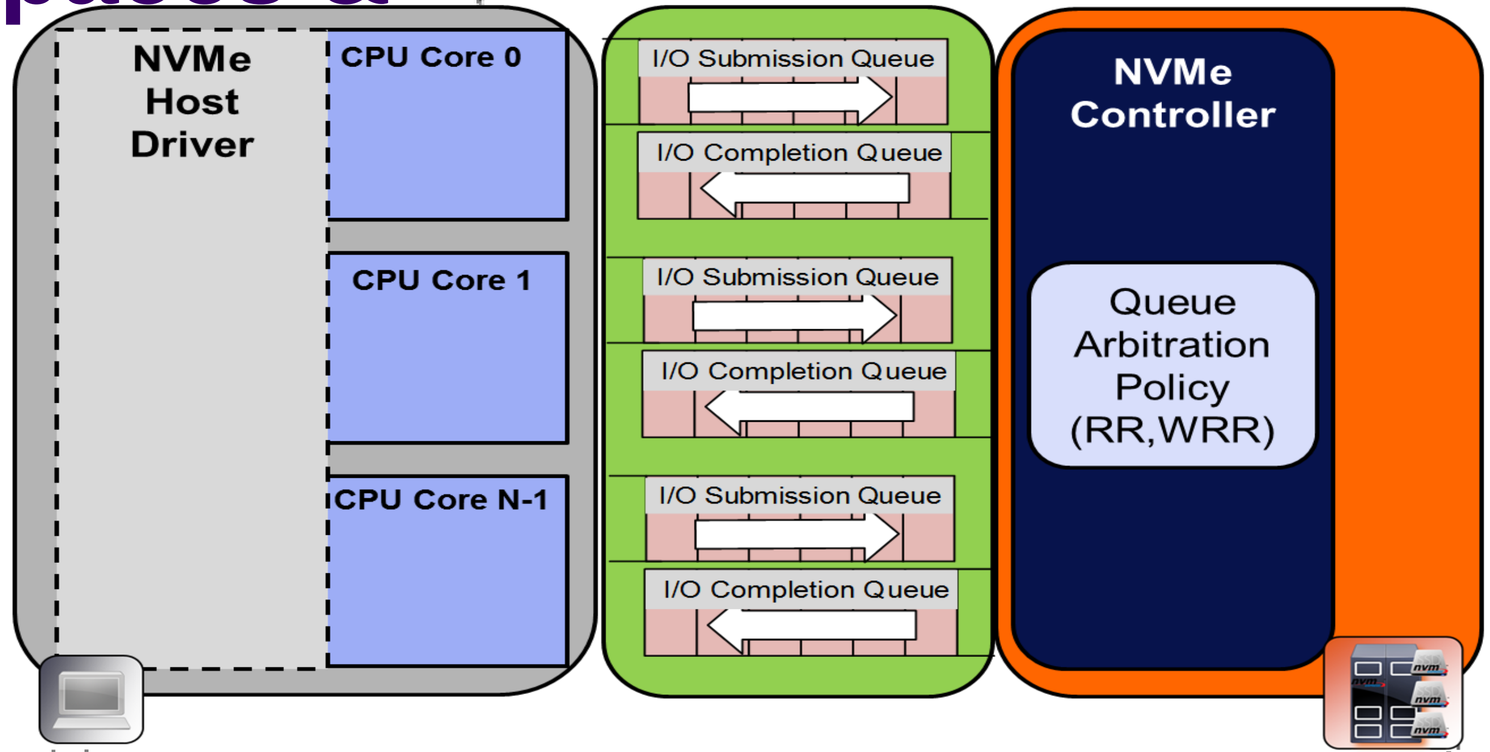
- Defines the mapping of NVM Media to a formatted LBA range
  - NVM Subsystem may have multiple Namespaces



NVMe Basics

- NVMe Drivers
- NVMe Subsystem
- NVMe Controller
- NVMe Namespaces & Media
- Queue Pairs

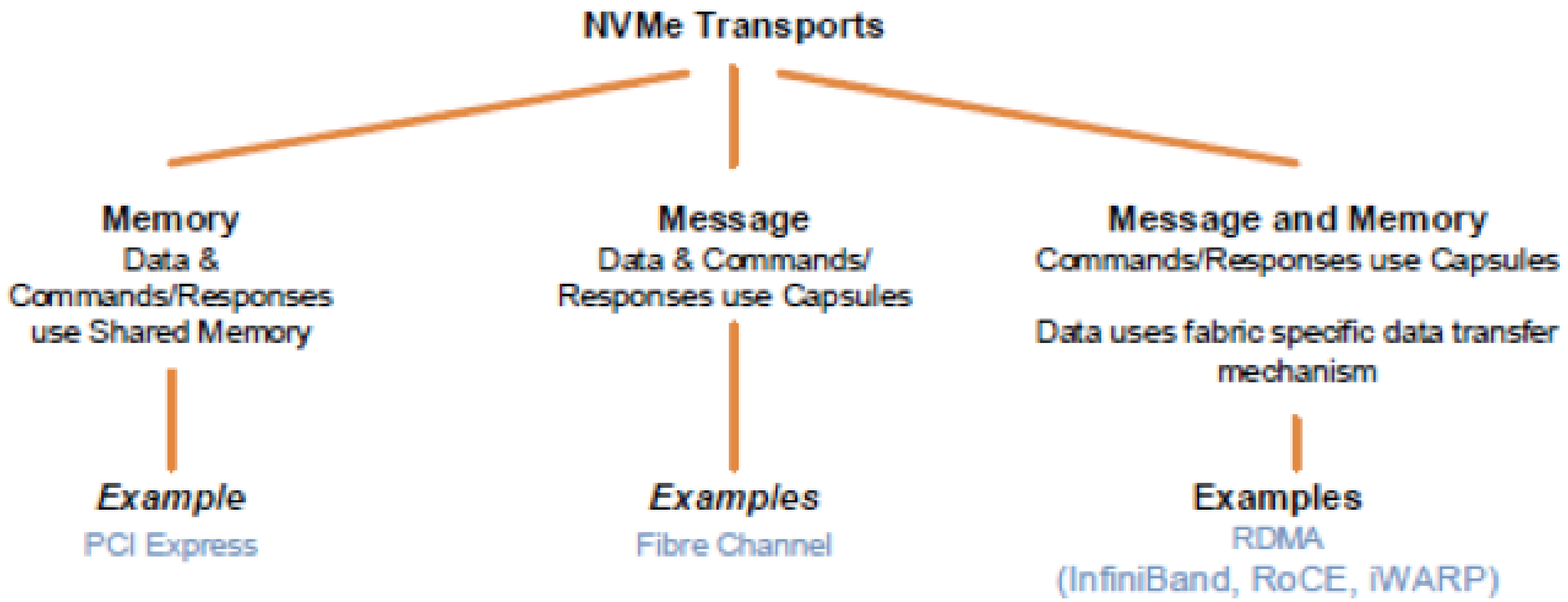
- I/O Submission and Completion Queue Pairs are aligned to Host CPU Cores
  - Independent per queue operations
- Transport type-dependent interfaces facilitate the queue operations and NVMe Command Data transfers



NVMe over Fabrics (NVMe-oF)

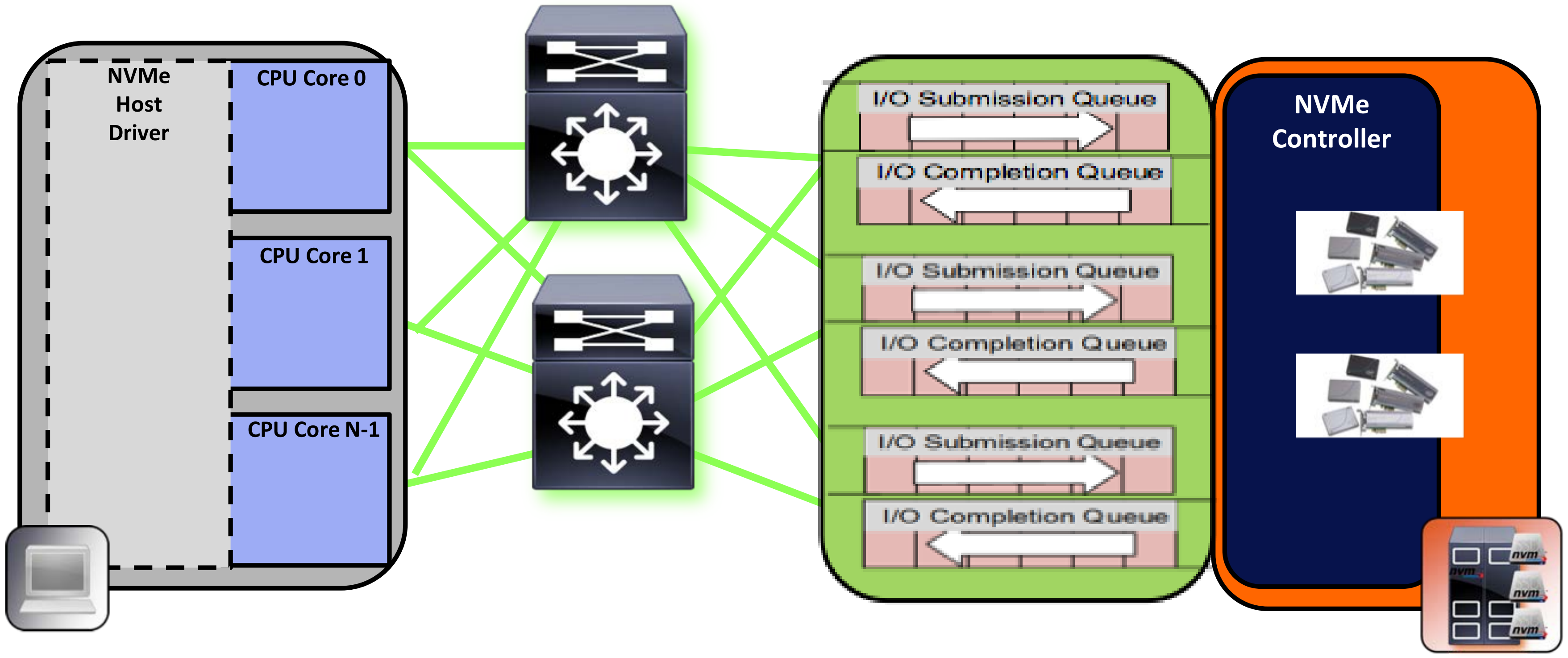
- NVMe is a Memory-Mapped, PCIe Model
- Fabrics is a message-based transport; no shared memory
- Fibre Channel uses capsules for both Data and Commands

Figure 1: Taxonomy of Transports



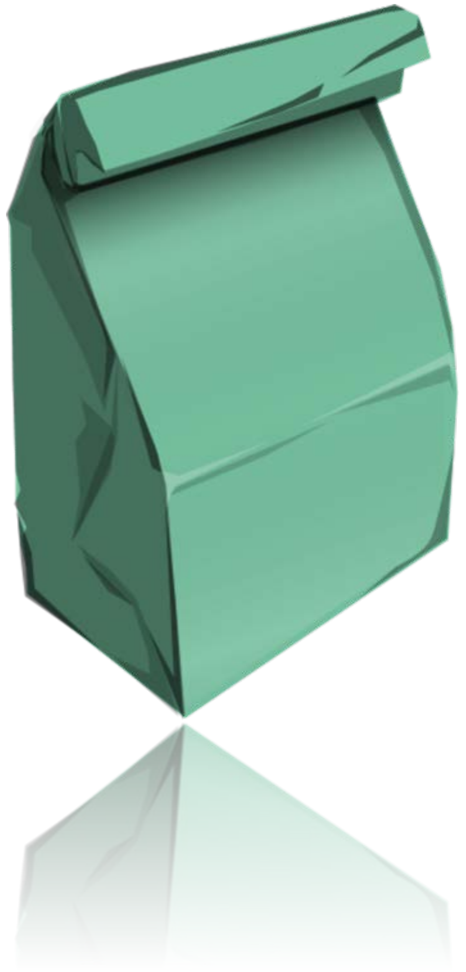
Extending Queue-Pairs over a Network

- **Each Host/Controller Pair have an independent set of NVMe queues**
- **Queue Pairs scale across Fabric**
  - Maintain consistency to multiple Subsystems
  - Each controller provides a separate set of queues, versus other models where single set of queues is used for multiple controllers



FC-NVMe

Take away from this section?



Most important part

- High level understanding of how FC-NVMe works
- Understand how FCP can be used to map NVMe to Fibre Channel

Next Section

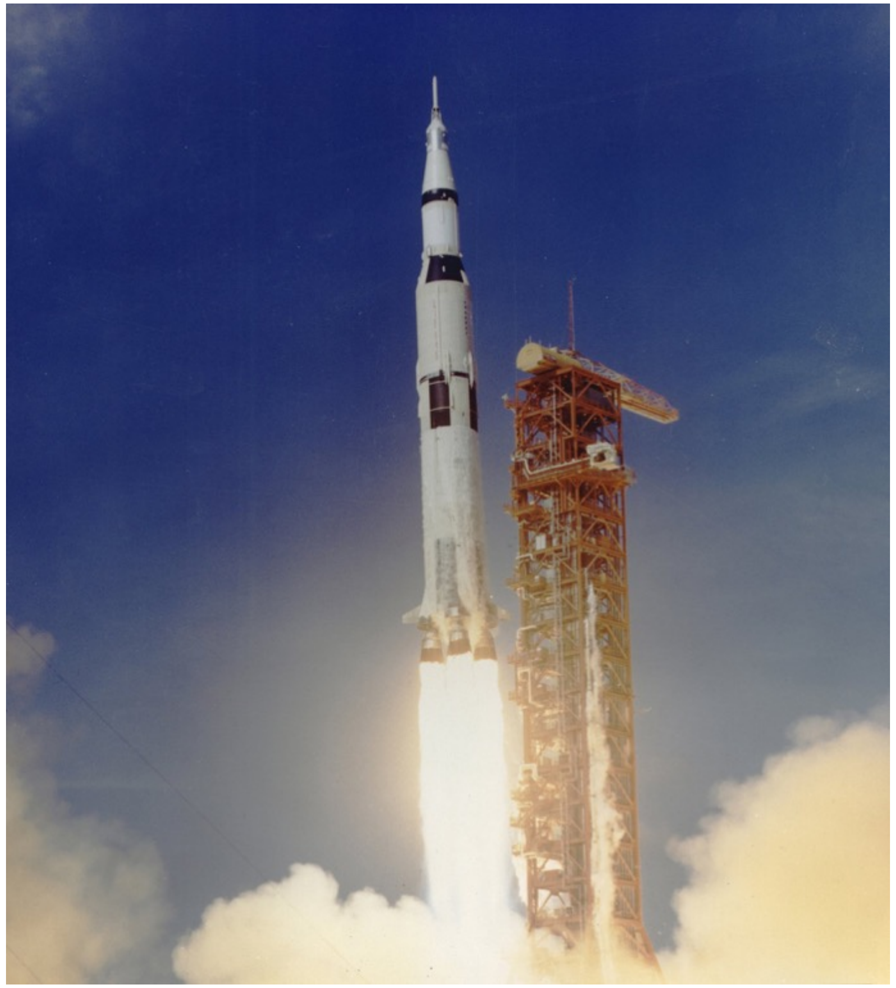
- Why use FC-NVMe?

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 Service Layer
  - Name Server
  - Zoning
  - Management

Performance

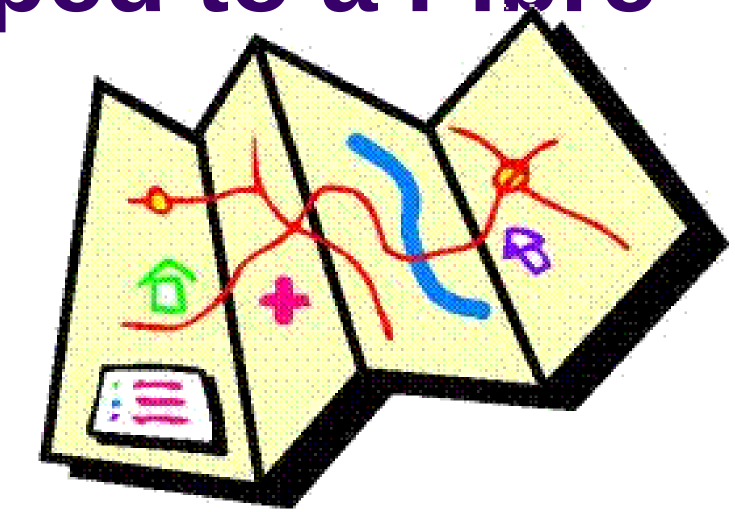


- **The Goal of High Performance/Low Latency**
  - Means that FC–NVMe needs to use an existing hardware accelerated data transfer protocol
  - FC does not have an RDMA protocol so FC-NVMe uses FCP as the data transfer protocol
    - Currently both SCSI and FC-SB (FICON) use FCP for data transfers
    - FCP is deployed as hardware accelerated in most (if not all) HBAs
    - Like FC, FCP is a connectionless protocol
      - Any FCP based protocols provide a way of creating a “connection”, or association between participating ports

FCP Mapping

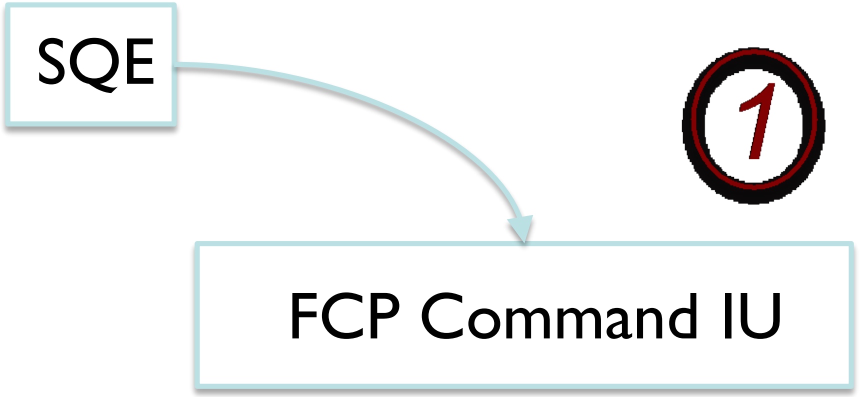
- The NVMe Command/Response capsules, and for some commands, data transfer, are directly mapped into FCP Information Units (IUs)

- A NVMe I/O operation is directly mapped to a Fibre Channel Exchange

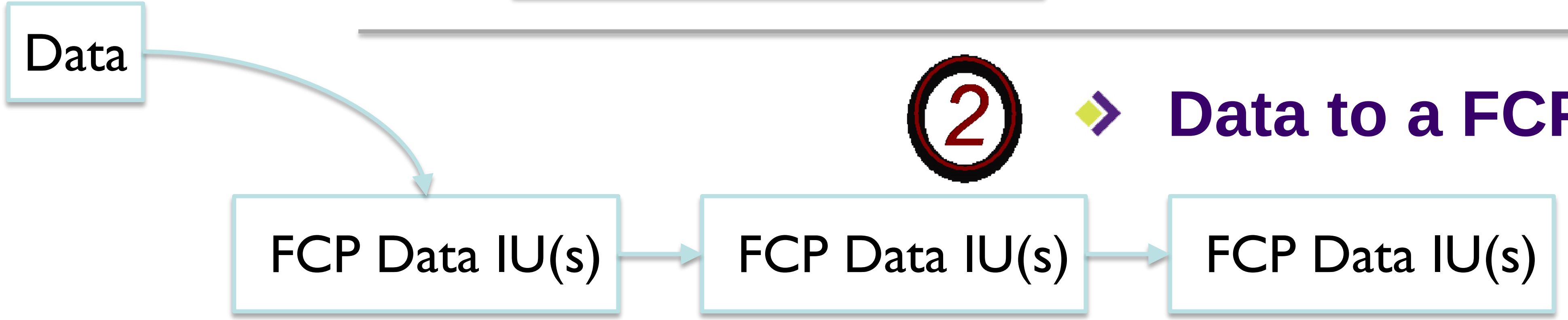


FC-NVMe Information Units (IUs)

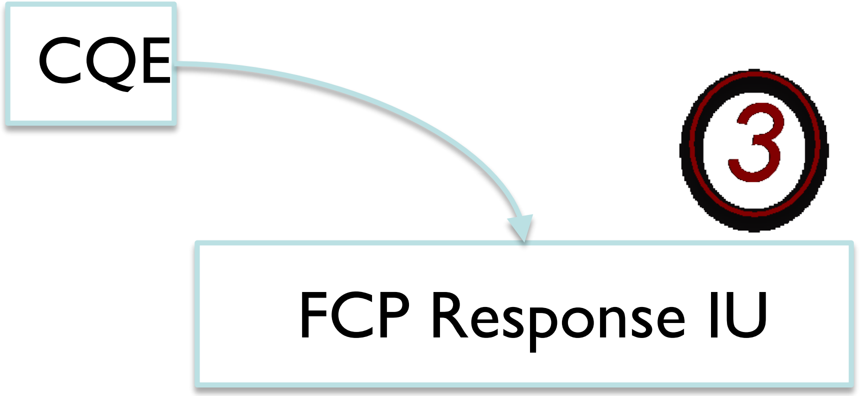
SNIA Global Education



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



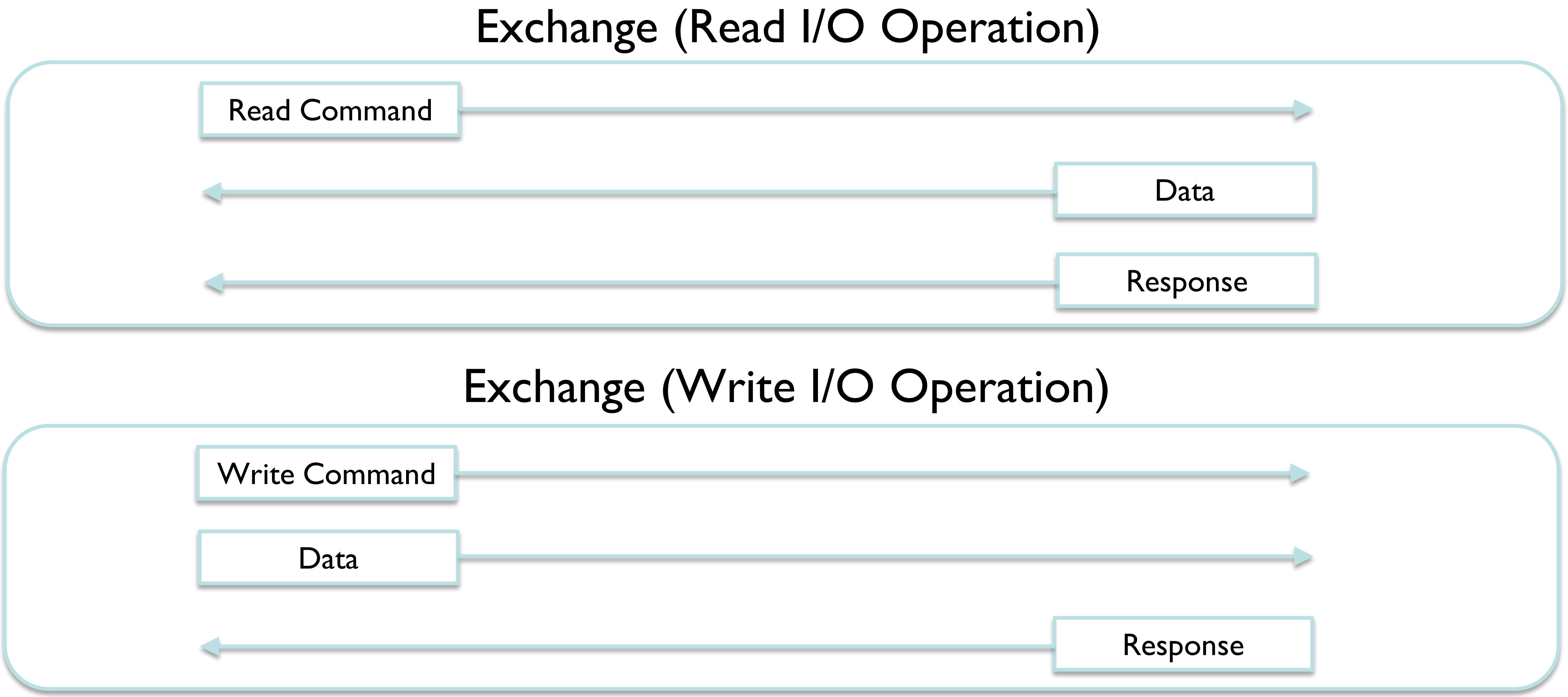
- Data to a FCP Data IU



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

I/O Operation

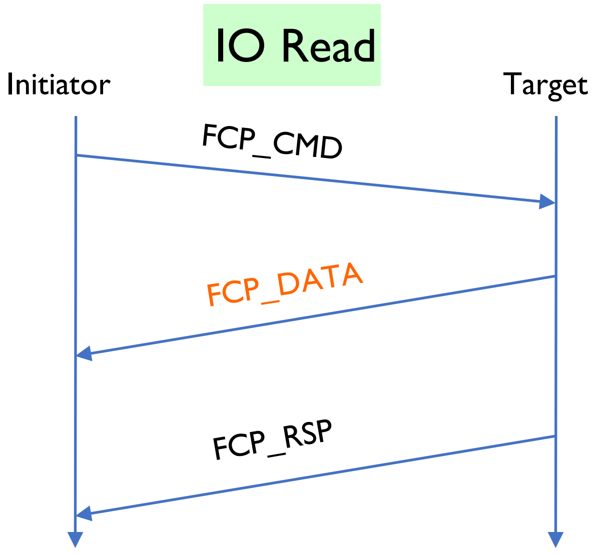
- **Transactions for a particular I/O Operation are bundled into an FC Exchange**

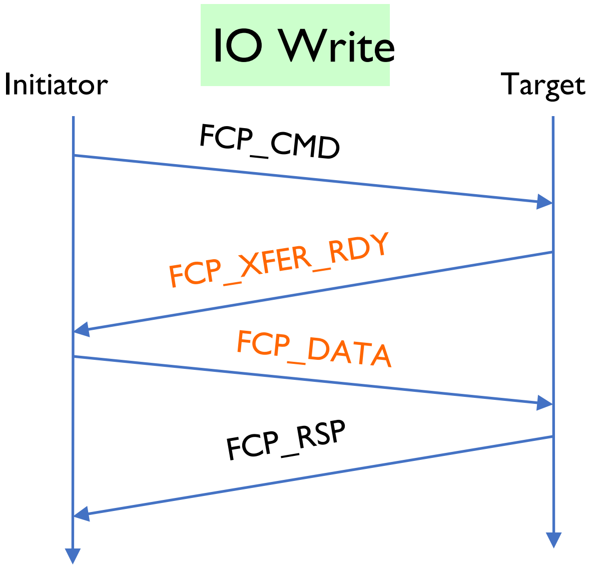


Zero Copy

- **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

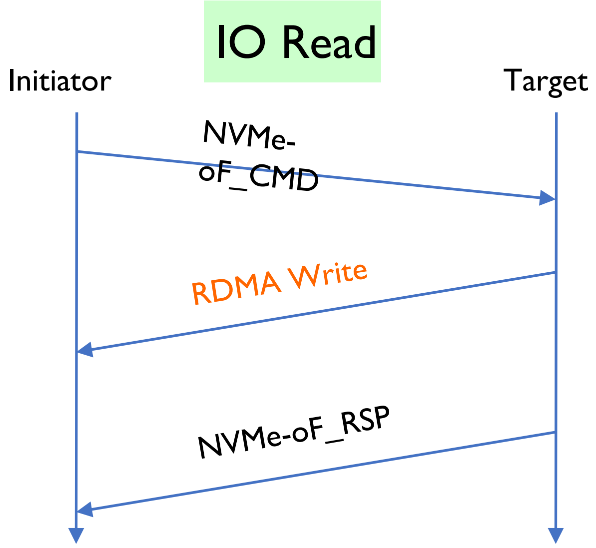
FCP Transactions

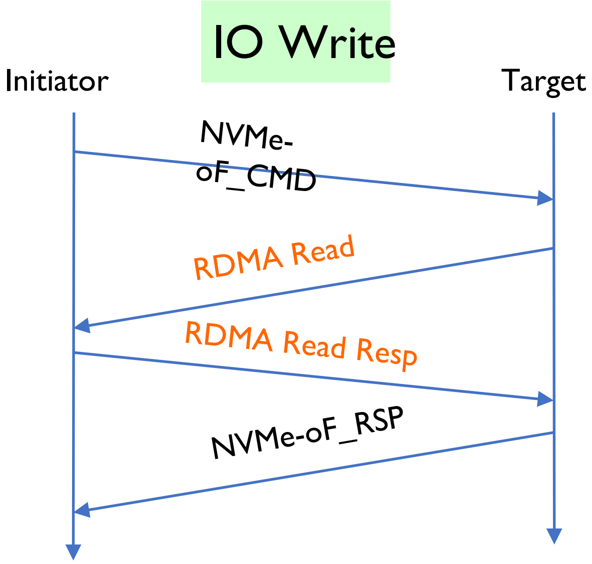




- **FCP Transactions look similar to RDMA**
  - For Read
    - FCP_DATA from Target
  - For Write
    - Transfer Ready and then DATA to Target

NVMe-oF Protocol Transactions





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

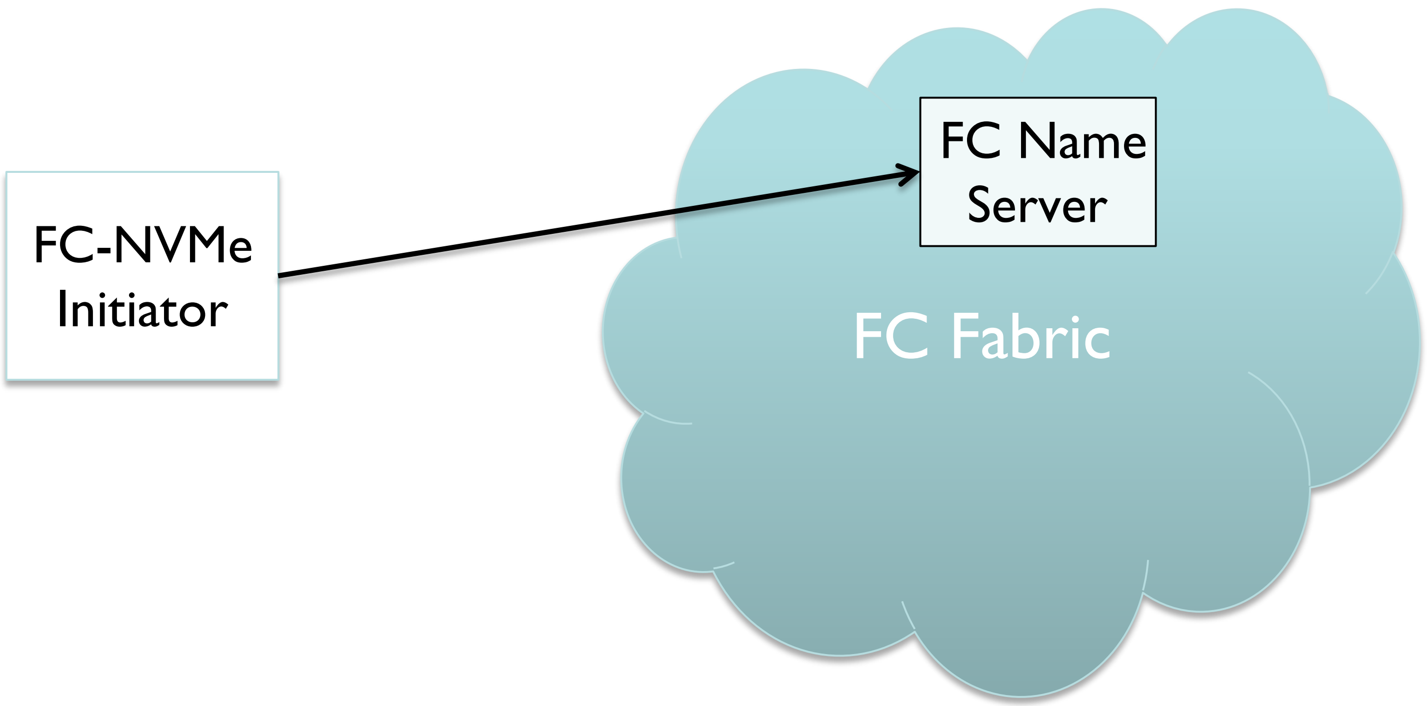
FC-NVMe Discovery

- **FC-NVMe 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

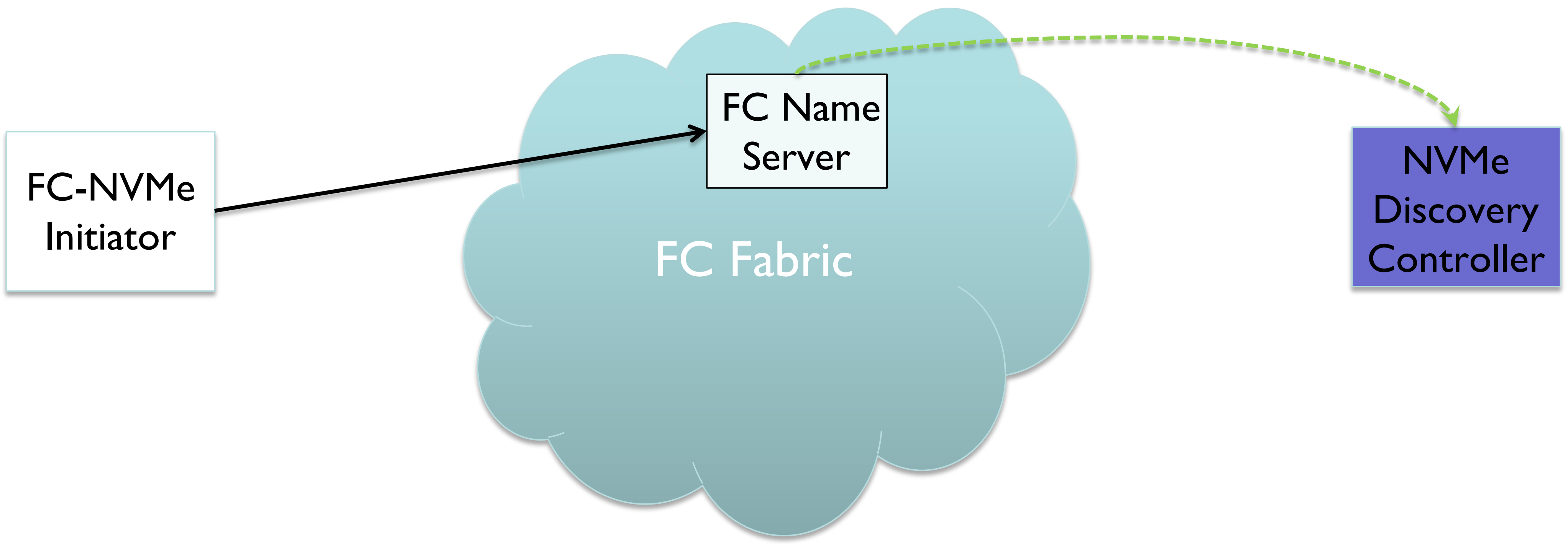
FC-NVMe Discovery Example

- FC-NVMe Initiator connects to FC Name Server



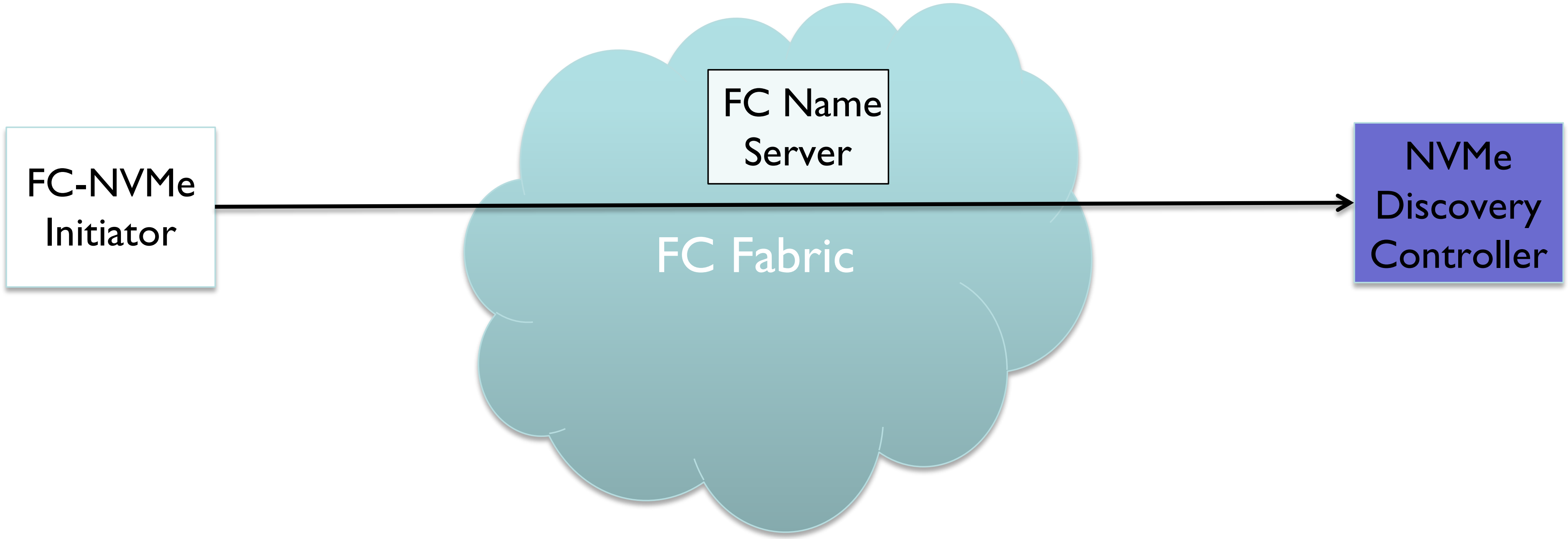
FC-NVMe Discovery Example

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



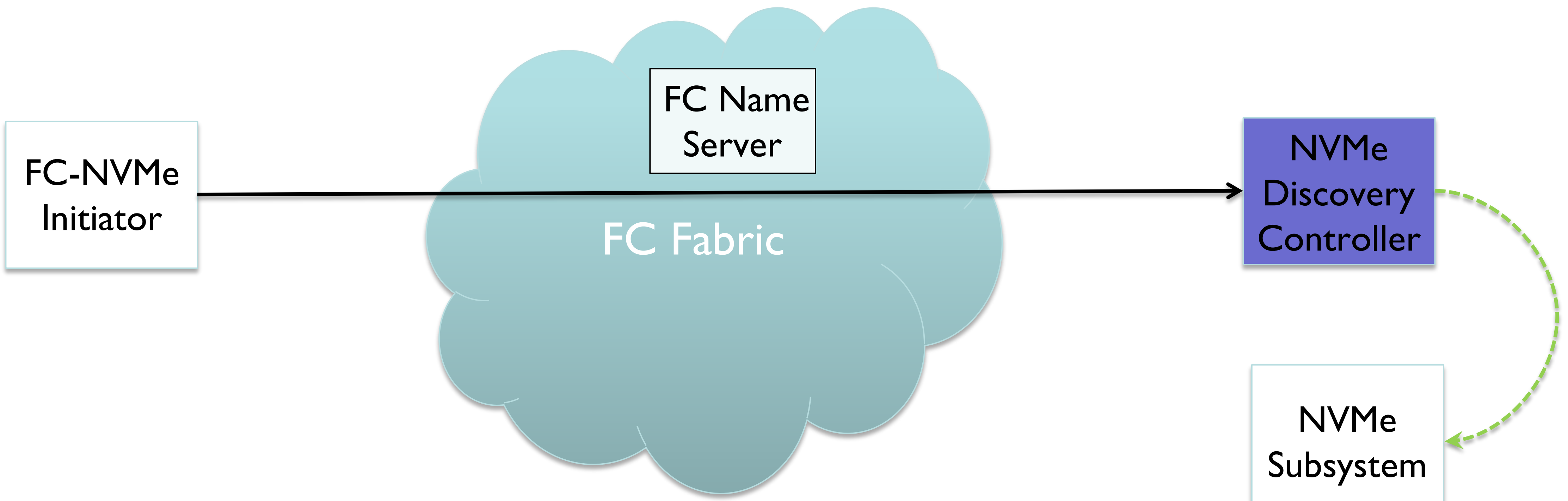
FC-NVMe Discovery Example

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



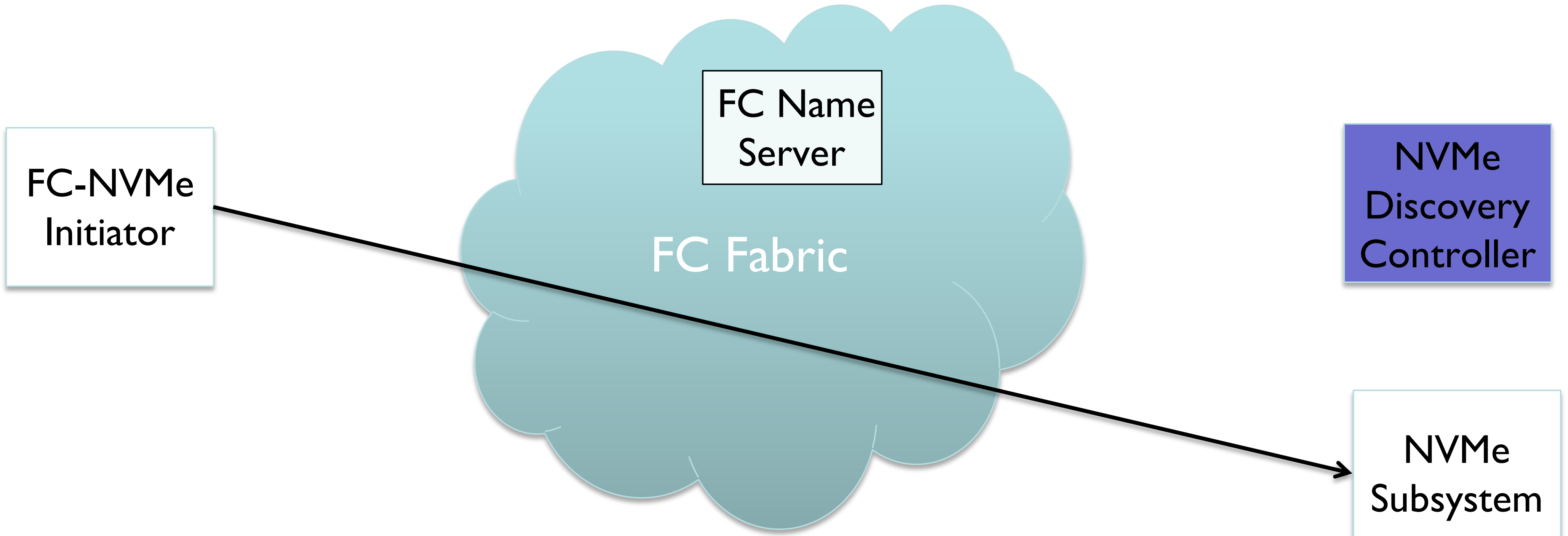
FC-NVMe Discovery Example

- NVMe Discovery Controller(s) identify available NVMe Subsystems



FC-NVMe Discovery Example

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



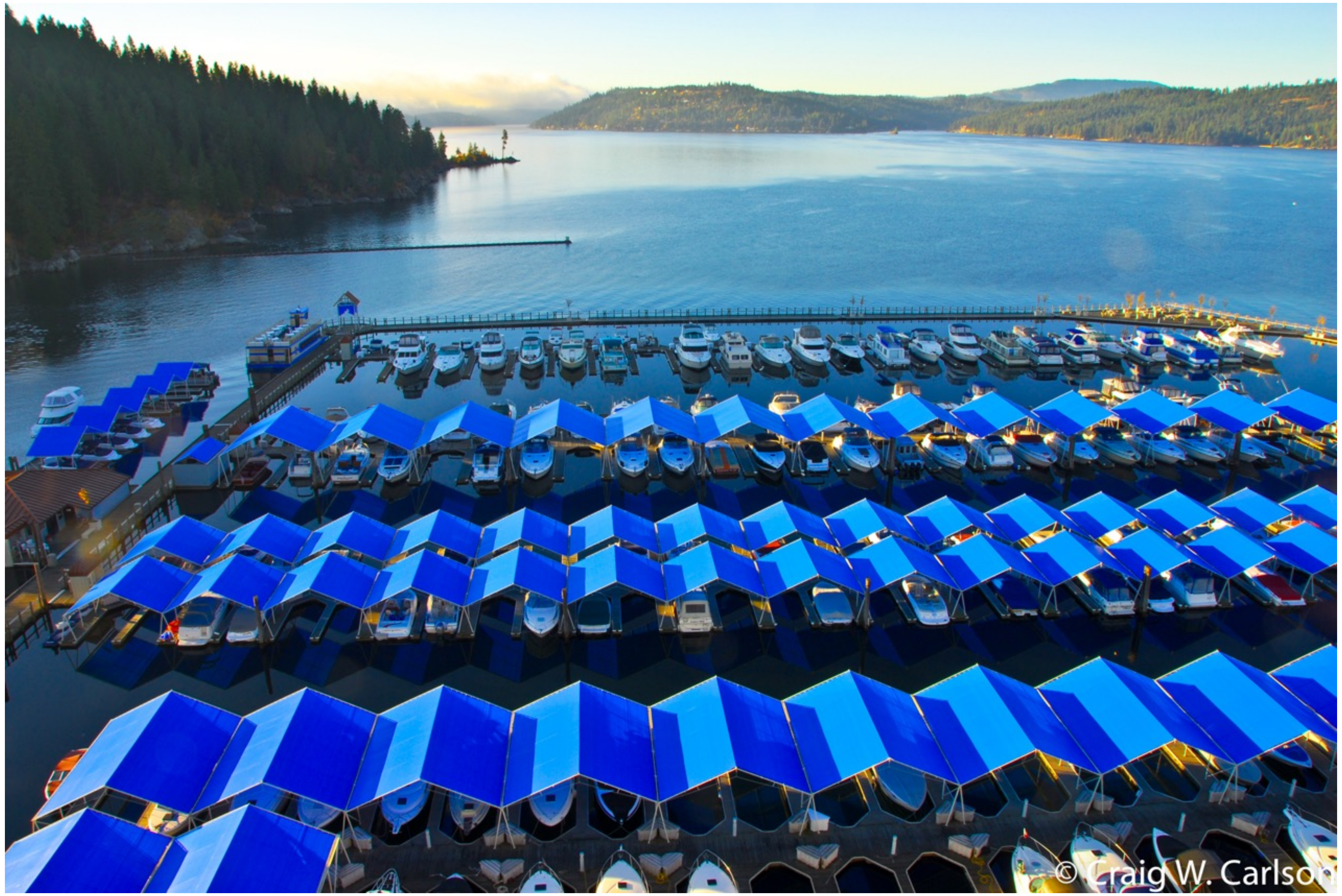
Zoning and Management

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

Why Use FC-NVMe?

Top 5 Reasons FC-NVMe Might Be The Right Choice

- **1) Dedicated Storage Network**



Top 5 Reasons FC-NVMe Might Be The Right Choice

- 1) Dedicated Storage Network
- 2) Run NVMe and SCSI Side-by-Side

Top 5 Reasons FC-NVMe Might Be The Right Choice

- 1) Dedicated Storage Network
- 2) Run NVMe and SCSI Side-by-Side
- 3) Robust and battle-hardened discovery and name service

Top 5 Reasons FC-NVMe Might Be The Right Choice

- 1) Dedicated Storage Network
- 2) Run NVMe and SCSI Side-by-Side
- 3) Robust and battle-hardened discovery and name service
- 4) Zoning and Security

Top 5 Reasons FC-NVMe Might Be The Right Choice

- 1) Dedicated Storage Network
- 2) Run NVMe and SCSI Side-by-Side
- 3) Robust and battle-hardened discovery and name service
- 4) Zoning and Security
- 5) Integrated Qualification and Support

Summary

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

More Info

- **FCIA**
  - www.fibrechannel.org
- **My contact**
  - craig.carlson@cavium.com

Thank you!

FIBRE CHANNEL INDUSTRY ASSOCIATION