FC-NVMe

About the presenter

- Presented by: Craig W. Carlson
  - Senior Technologist, Marvell
  - 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
- FC-NVMe Update
- FC-NVMe-2
- Why Use FC-NVMe?
- Summary

What This Presentation Is

- A reminder of how Fibre Channel works
- A reminder of how NVMe over Fabrics work
- An overview of FC-NVMe
- Update on FC-NVMe-2 (the new stuff)

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

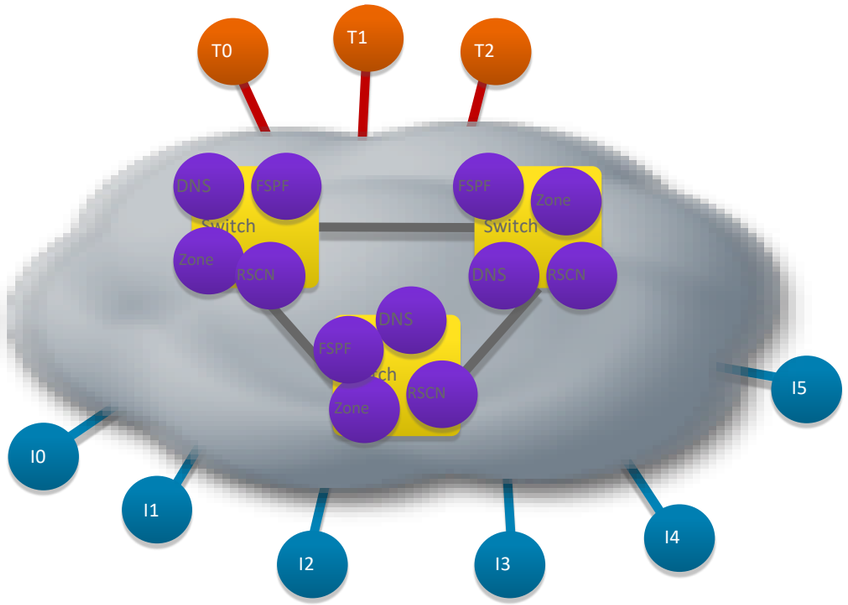
FIBRE CHANNEL REFRESHER

What is Fibre Channel?

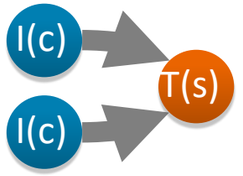
- A network purpose-built for storage
- A physical connection between a host and its storage
- A logical (protocol) connection between a host and its storage

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

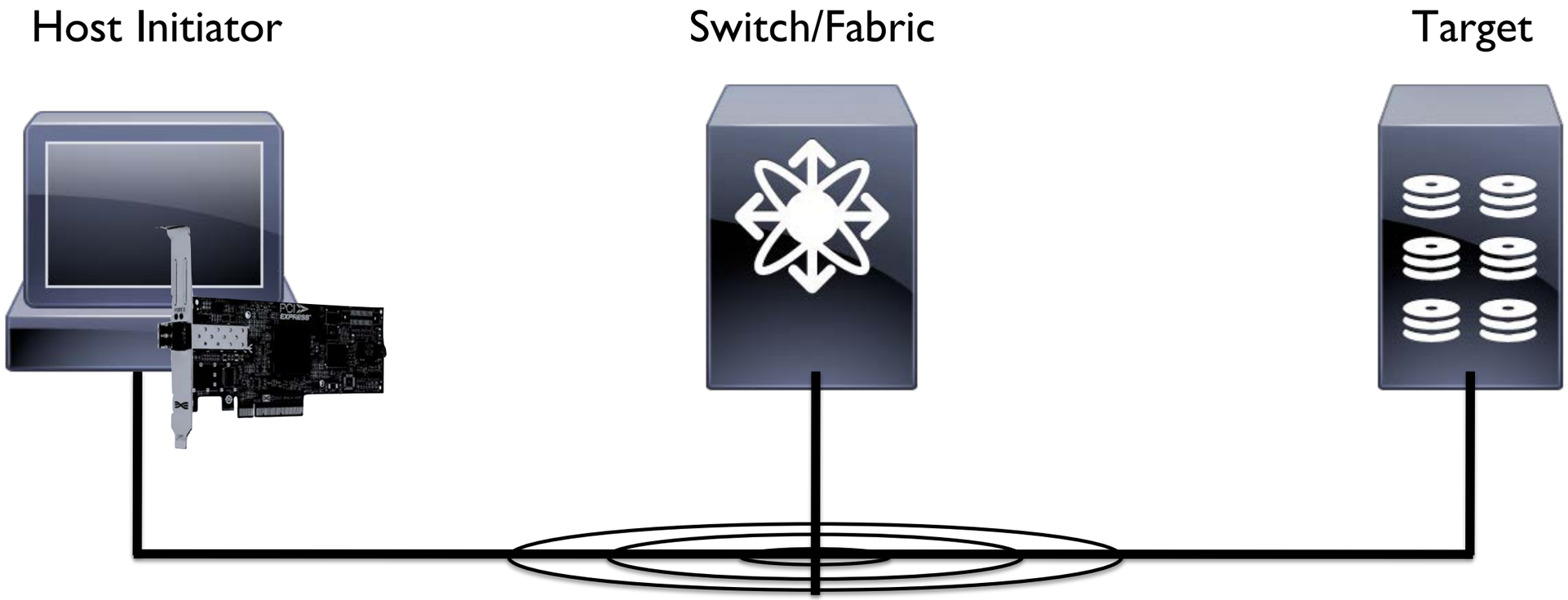


Fabric topology, services and traffic flows are structured



Client/Server Relationships are pre-defined

Design Elements



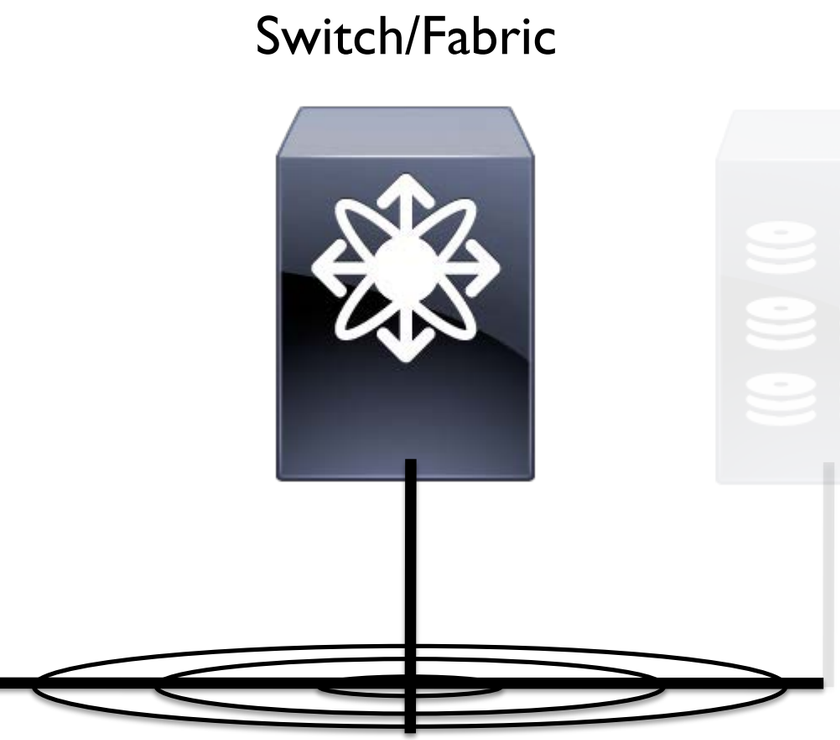
- Terminology that covers components or parts of the system
- Terminology that talks about the end-to-end system

Design Elements

Host Initiator

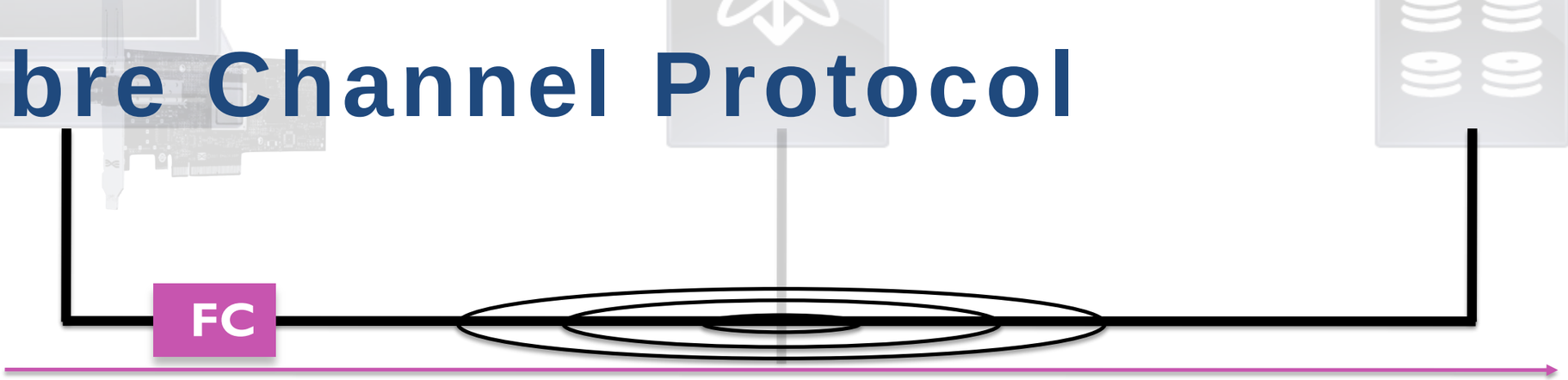
- **For FC the adapter which sits in a Host is called an HBA (Host Bus Adapter)**
  - Equivalent to a NIC for Ethernet
- **Where protocols such as NVMe or SCSI get encapsulated into a Fibre Channel Frame**

Design Elements



- **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



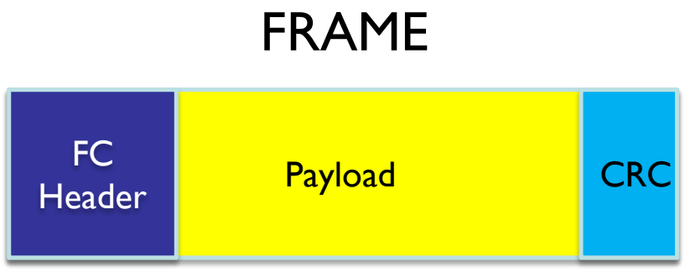
- **Fibre Channel typically uses an Unacknowledged Datagram Service**
  - Known as “Class 3”
  - Defined as a reliable datagram (connectionless) service
    - A class 3 frame will not be dropped unless an error occurs (i.e., bit error, or other unrecoverable error)

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

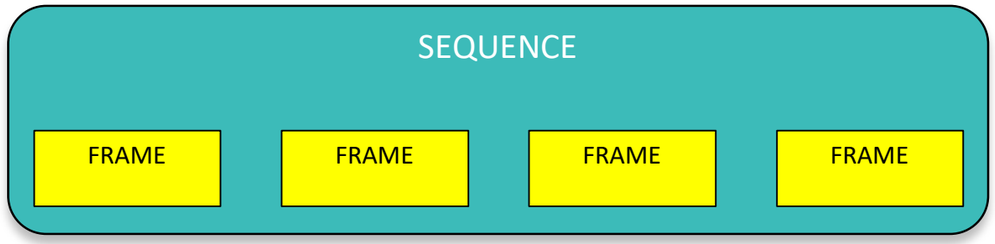
Frames

- **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



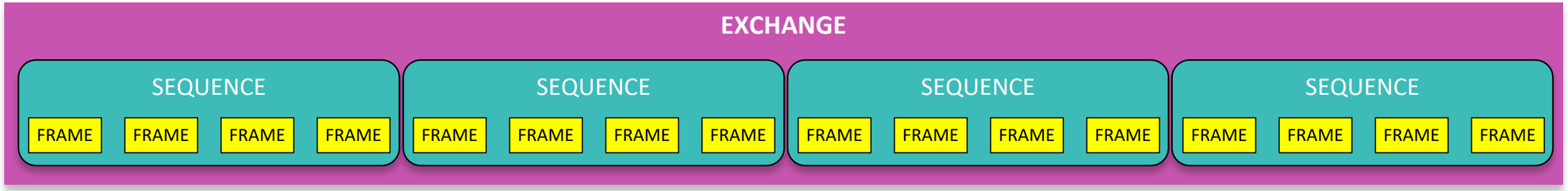
Sequences

- **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)



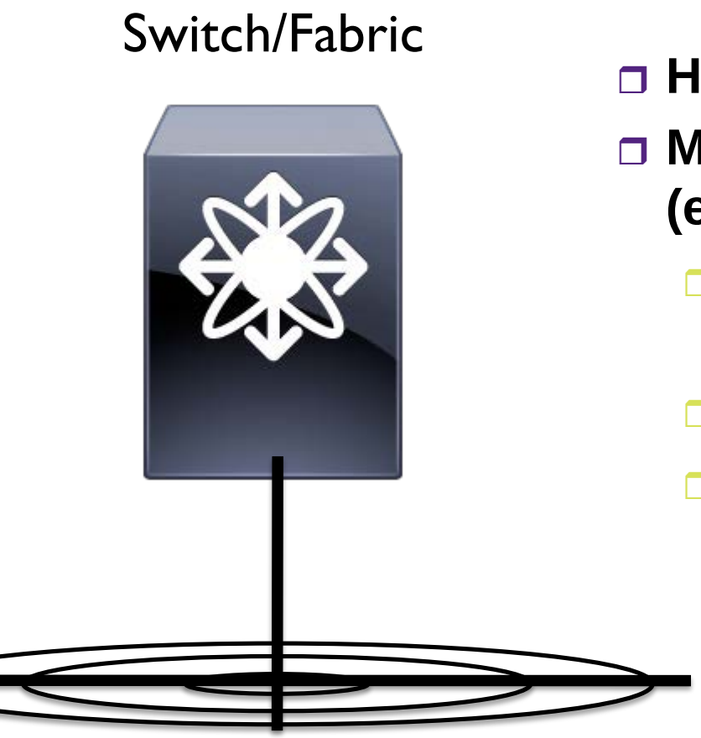
Exchanges

- **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

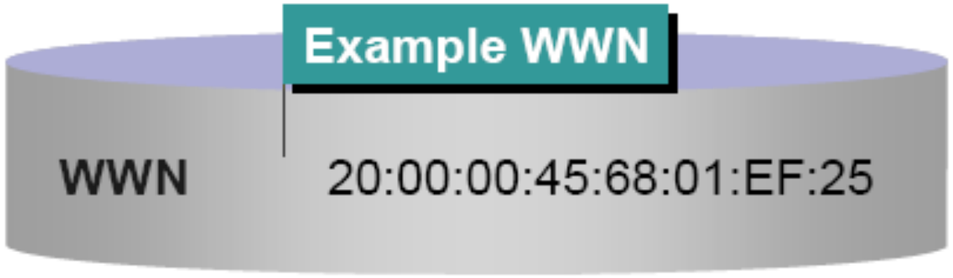


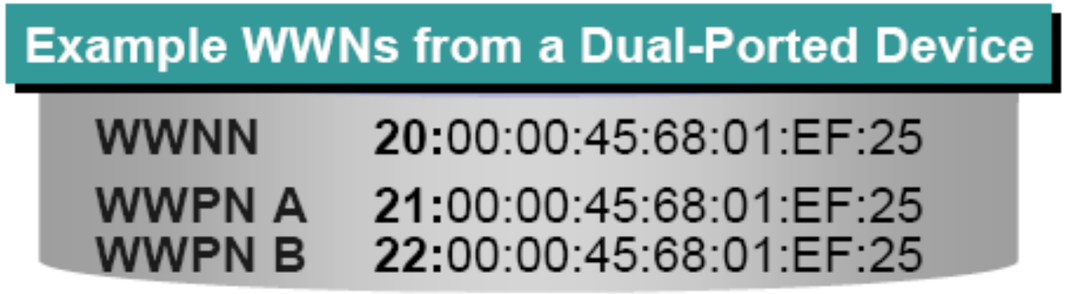
*not to scale

Discovery in a FC Network

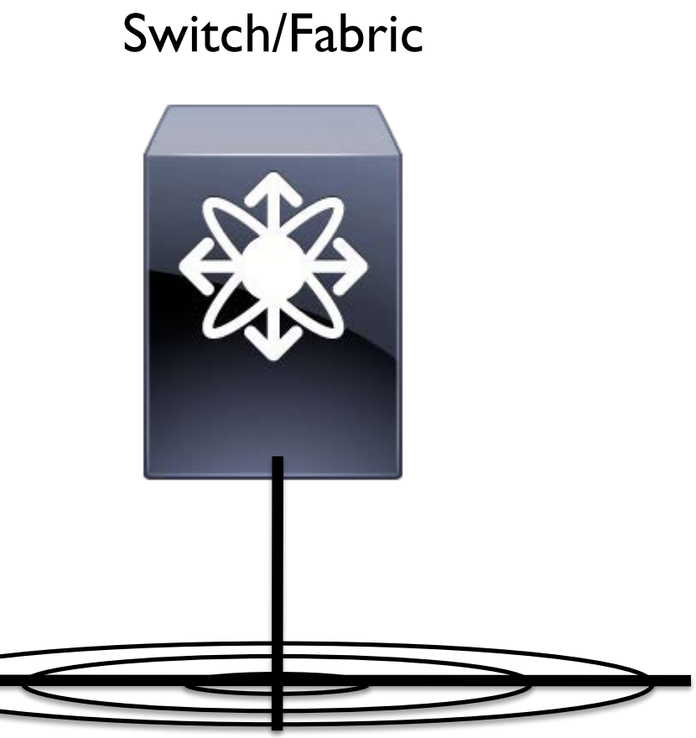


- **Handled through the FC Name Server**
- **Many port attributes are automatically registered to the FC Name Server (e.g., Node WWN, Port WWN, Protocol types, etc.)**
  - Every Fibre Channel port and node has a hard-coded address called ***World Wide Name*** (WWN)
  - WWNN uniquely identify ***devices***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



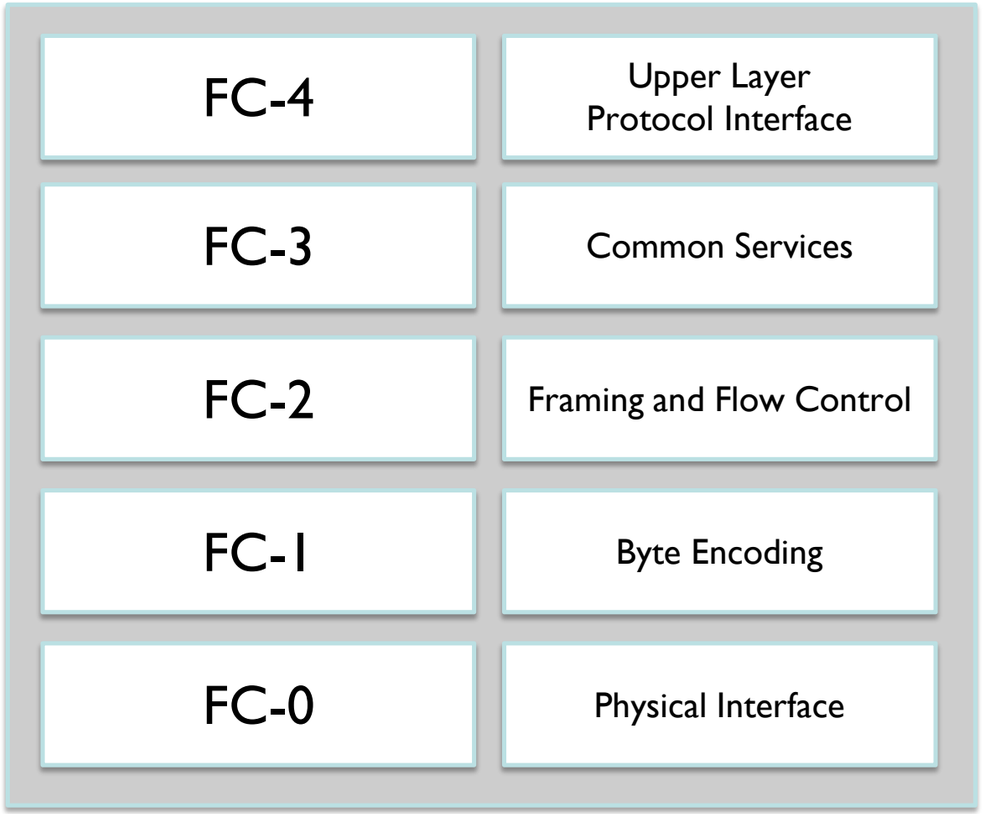


Zones/Zoning



- **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**

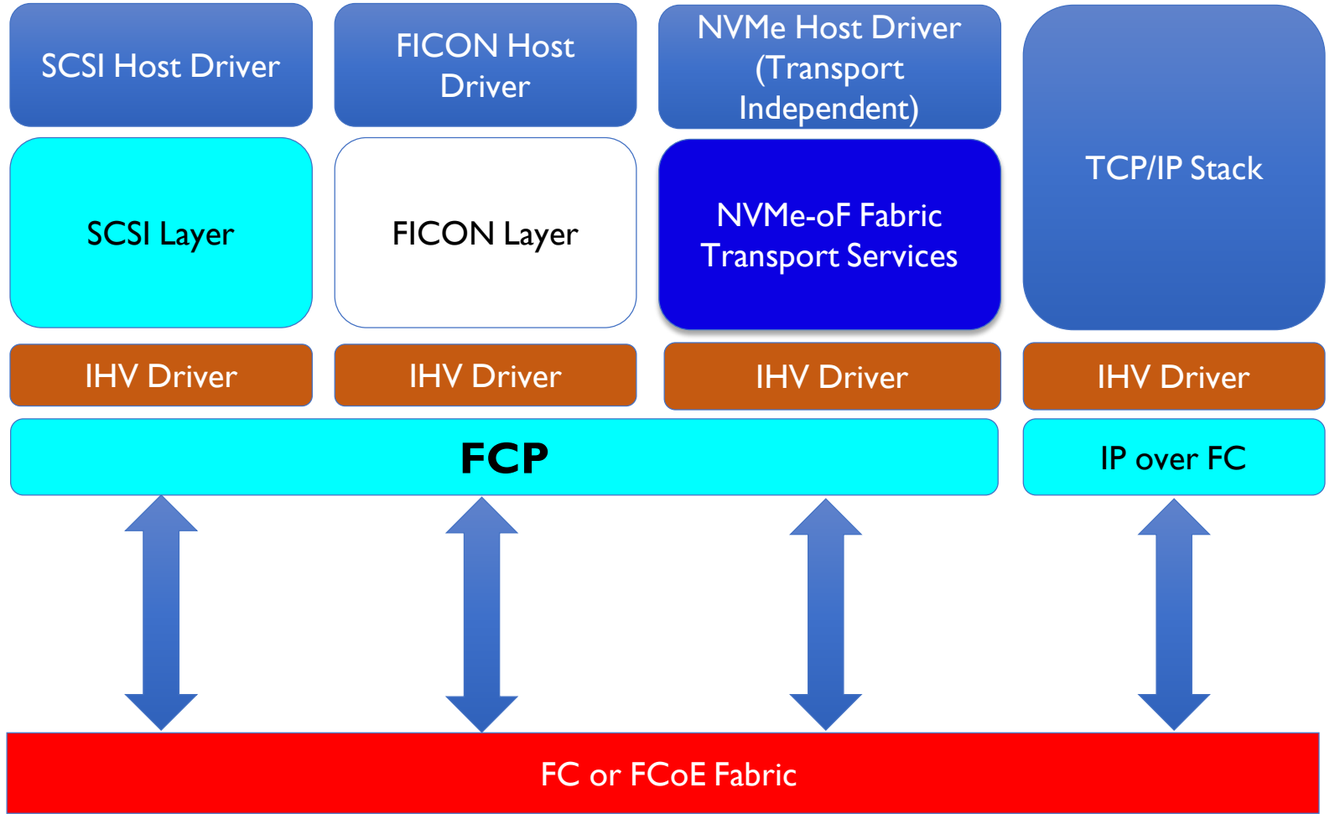
Fibre Channel Protocol



- **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?

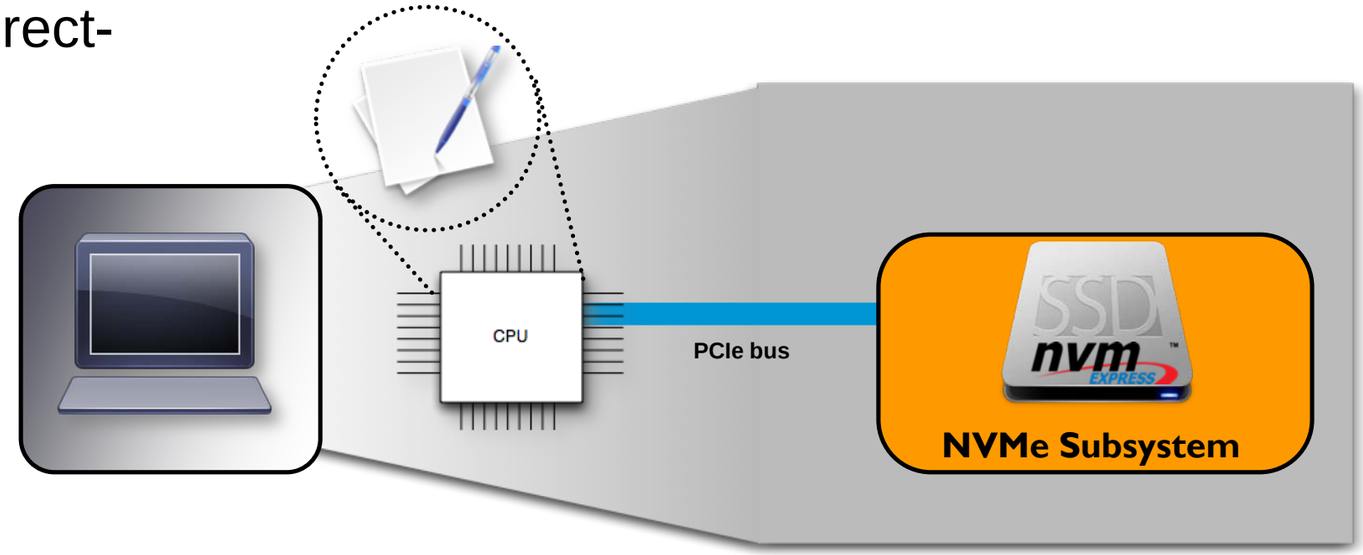
- **What's the difference between FCP and "FCP"?**
  - FCP is a data transfer protocol that carries other upper-level transport protocols (e.g., FICON, SCSI, NVMe)
  - Historically FCP meant SCSI FCP, but other protocols exist now
- **NVMe "hooks" into FCP**
  - Seamless transport of NVMe traffic
  -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



NVMe REFRESHER

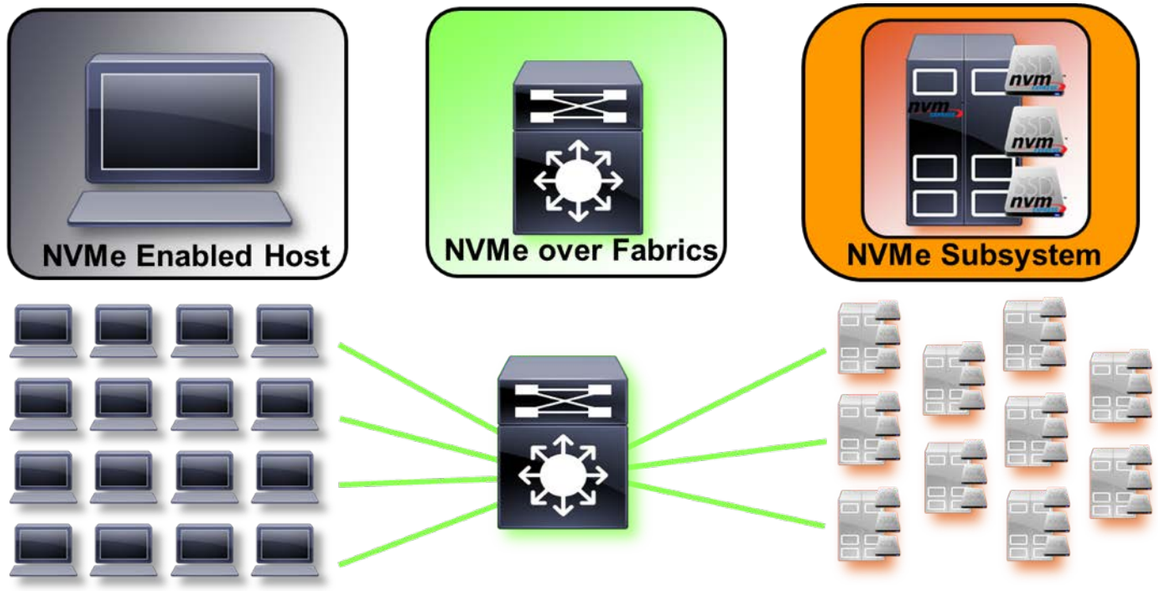
What is Non-Volatile Memory Express (NVMe) and NVMe over Fabrics (NVMe-oF)?

- **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)?

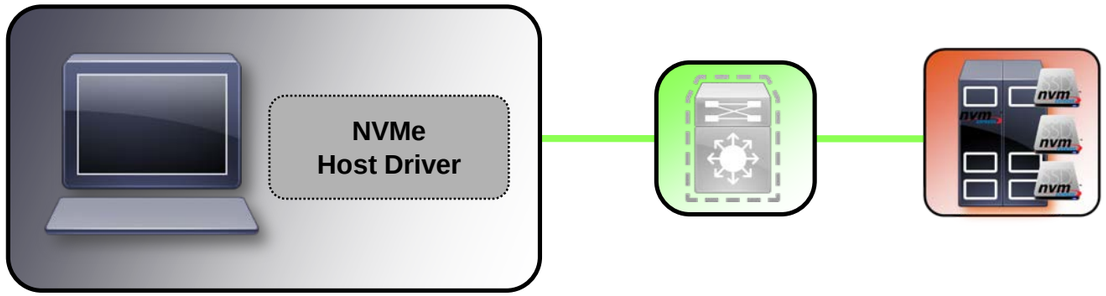
- **Non-Volatile Memory Express (NVMe)**
  - Began as an industry standard solution for efficient PCIe attached non-volatile memory storage (e.g., NVMe PCIe SSDs)
  - Low latency and high IOPS direct-attached NVM storage
- **NVMe over Fabrics (NVMe-oF)**
  - Built on common NVMe architecture with additional definitions to support message-based NVMe operations
  - Standardization of NVMe over a range Fabric types
    - Initial fabrics; RDMA(RoCE, iWARP, InfiniBand™) and Fibre Channel
    - TCP more recent addition



NVMe Basics

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

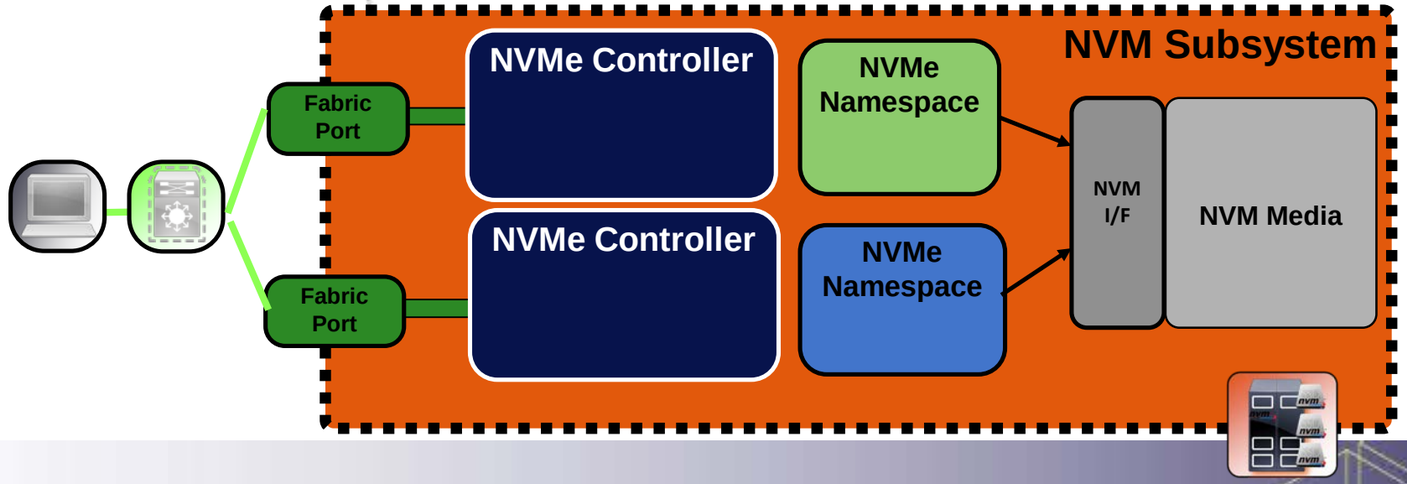
- In-box PCIe NVMe drivers in all major operating systems
- NVMe-oF requires specific drivers
  - FC-NVMe drivers will be provided by Fibre Channel vendors like always



NVMe Basics

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

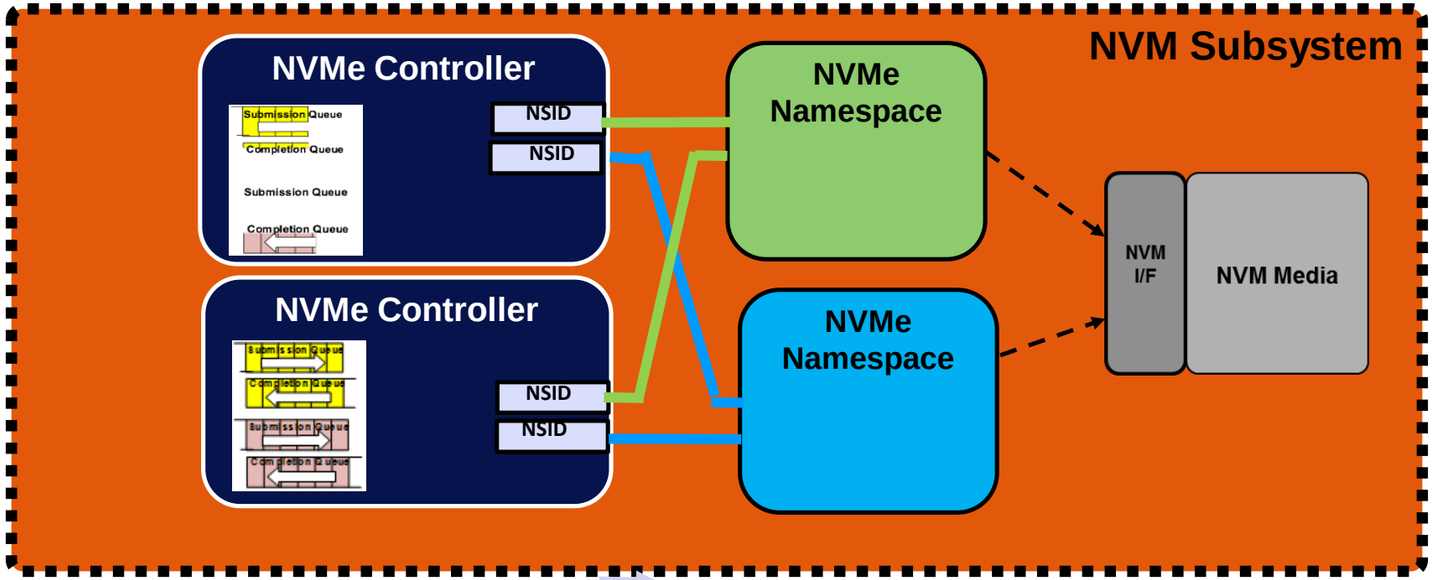
- Contains the architectural elements for NVMe targets
  - NVMe Controller
  - NVM Media
  - NVMe Namespaces
  - Interfaces



NVMe Basics

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

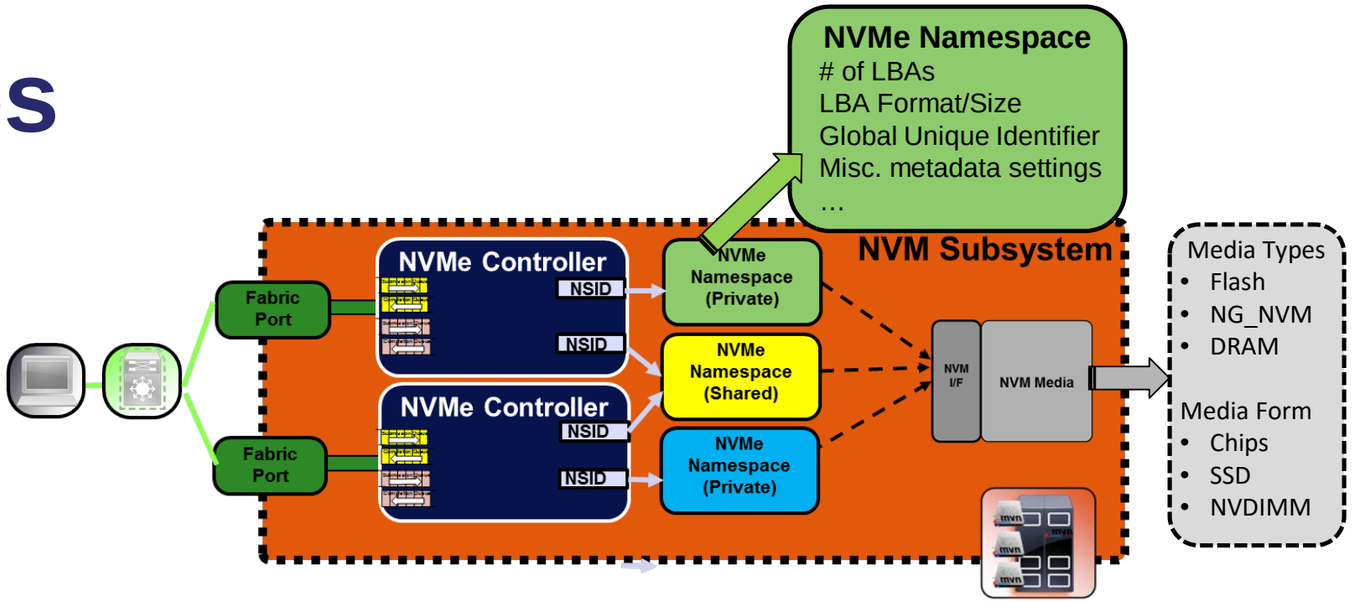
- NVMe Command Processing
- Access to NVMe Namespaces
  - Namespace ID (NSID) associates a Controller to Namespaces(s)



NVMe Basics

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

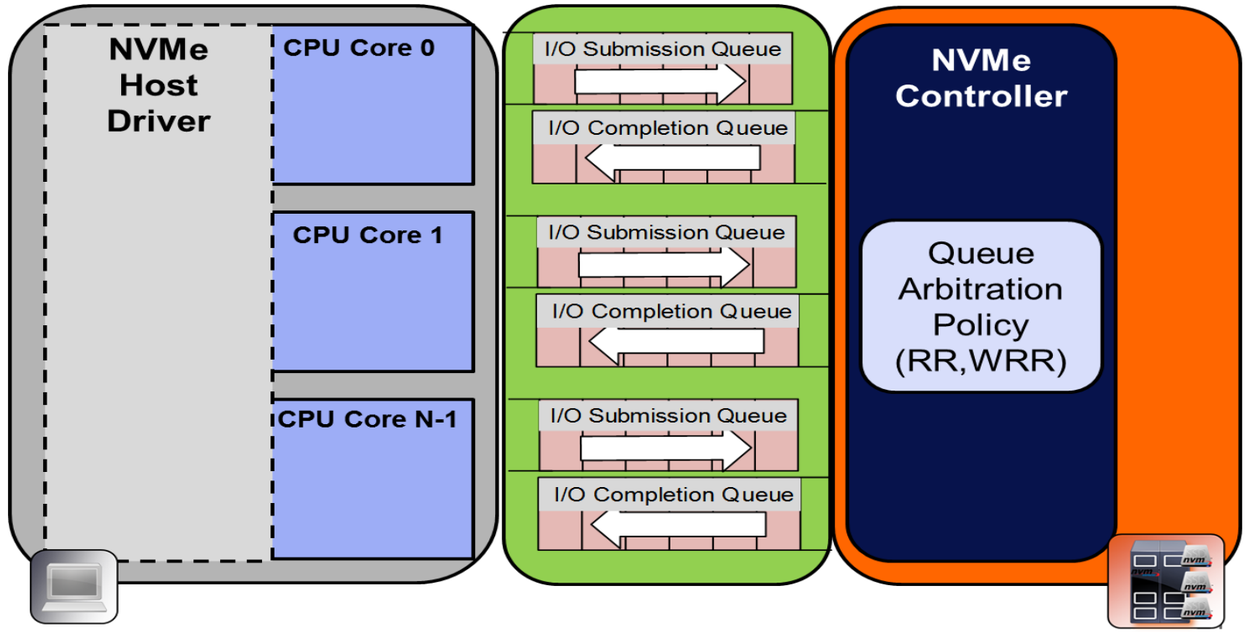
- Defines the mapping of NVM Media to a formatted LBA range
  - NVM Subsystem may have multiple Namespaces



NVMe Basics

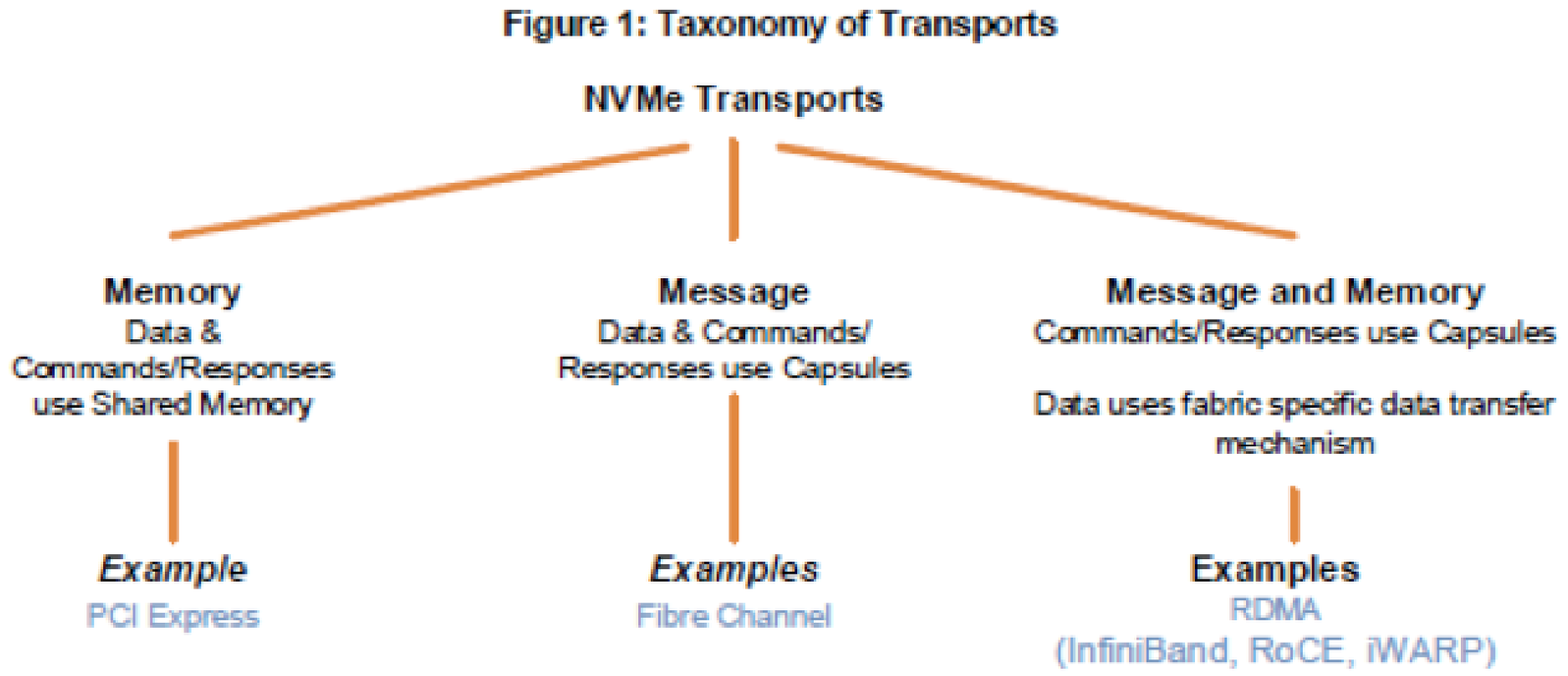
- NVMe Drivers
- NVMe Subsystem
- NVMe Controller
- NVMe Namespaces & Media
- Queue Pairs

- 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



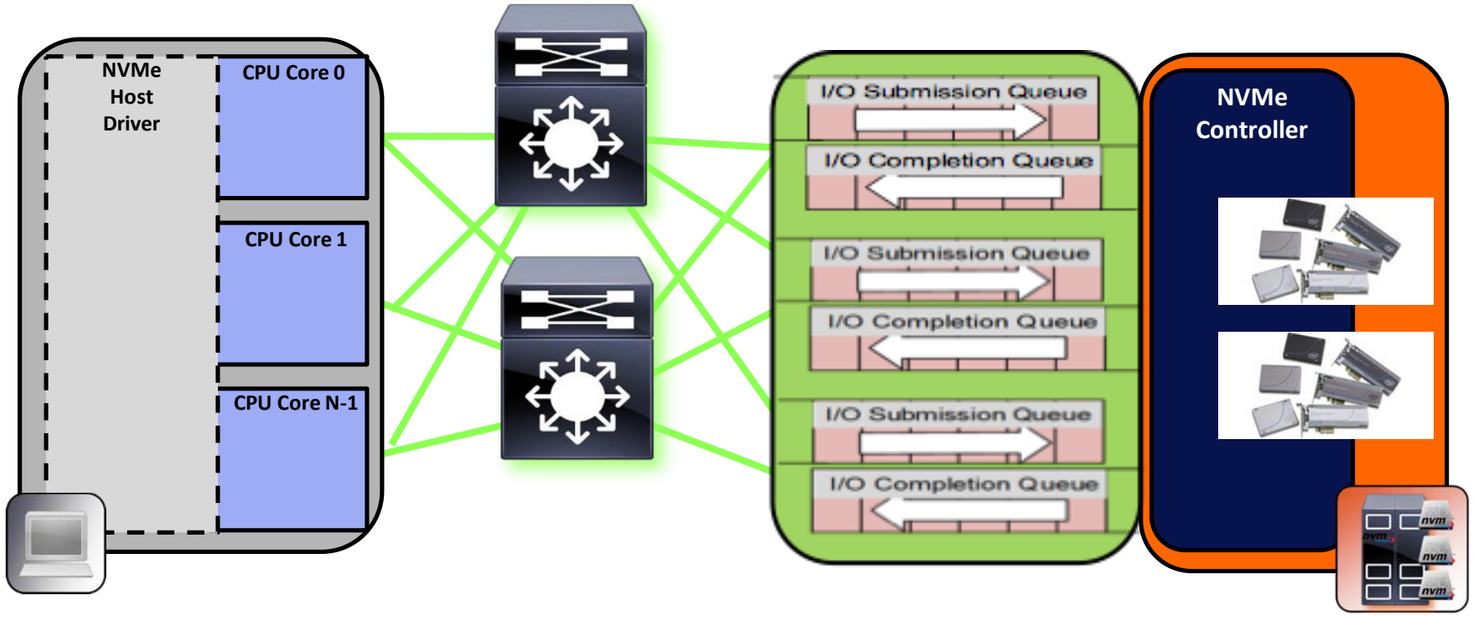
NVMe over Fabrics (NVMe-oF)

- 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



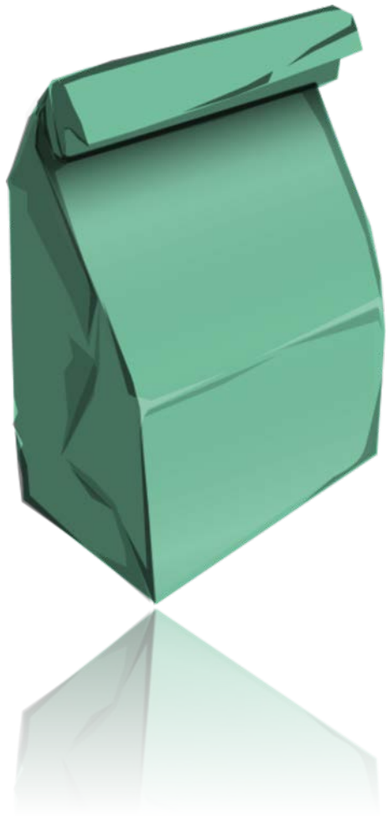
Extending Queue-Pairs over a Network

- **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?

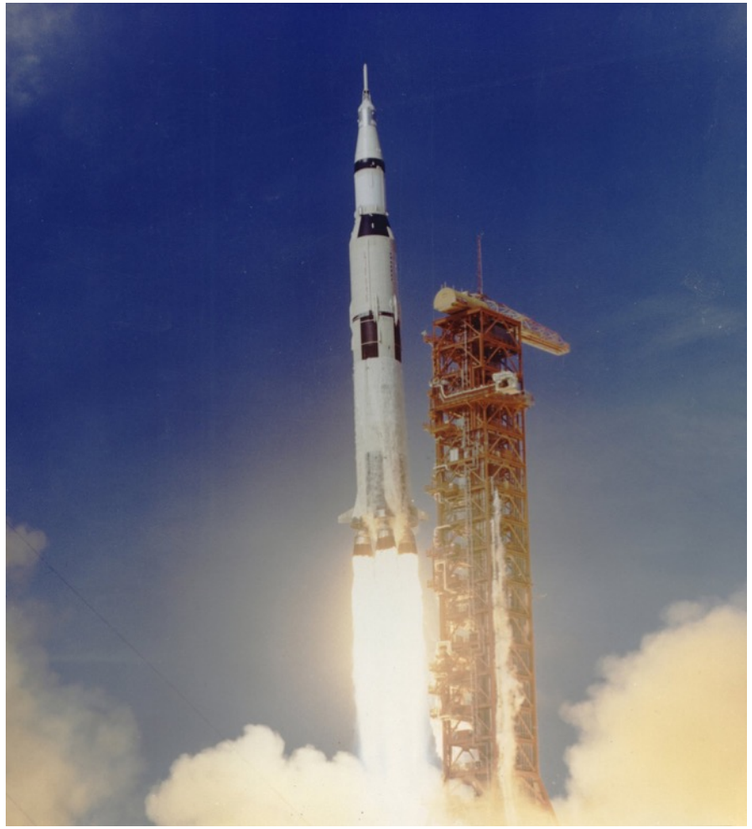


- **Most important part**
  - High level understanding of how FC-NVMe works
  - Update on FC-NVMe-2
- **Next Section**
  - Why use FC-NVMe?

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

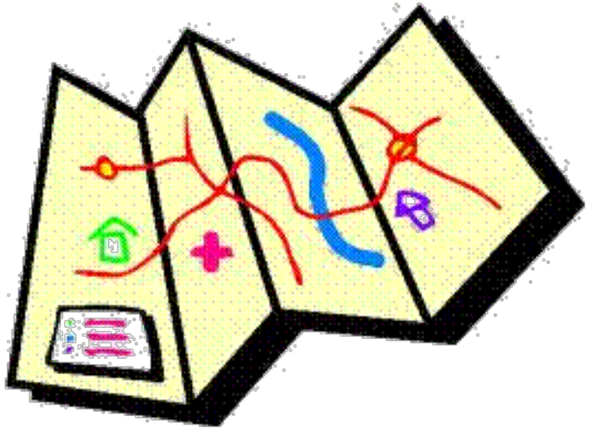
Performance



- **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

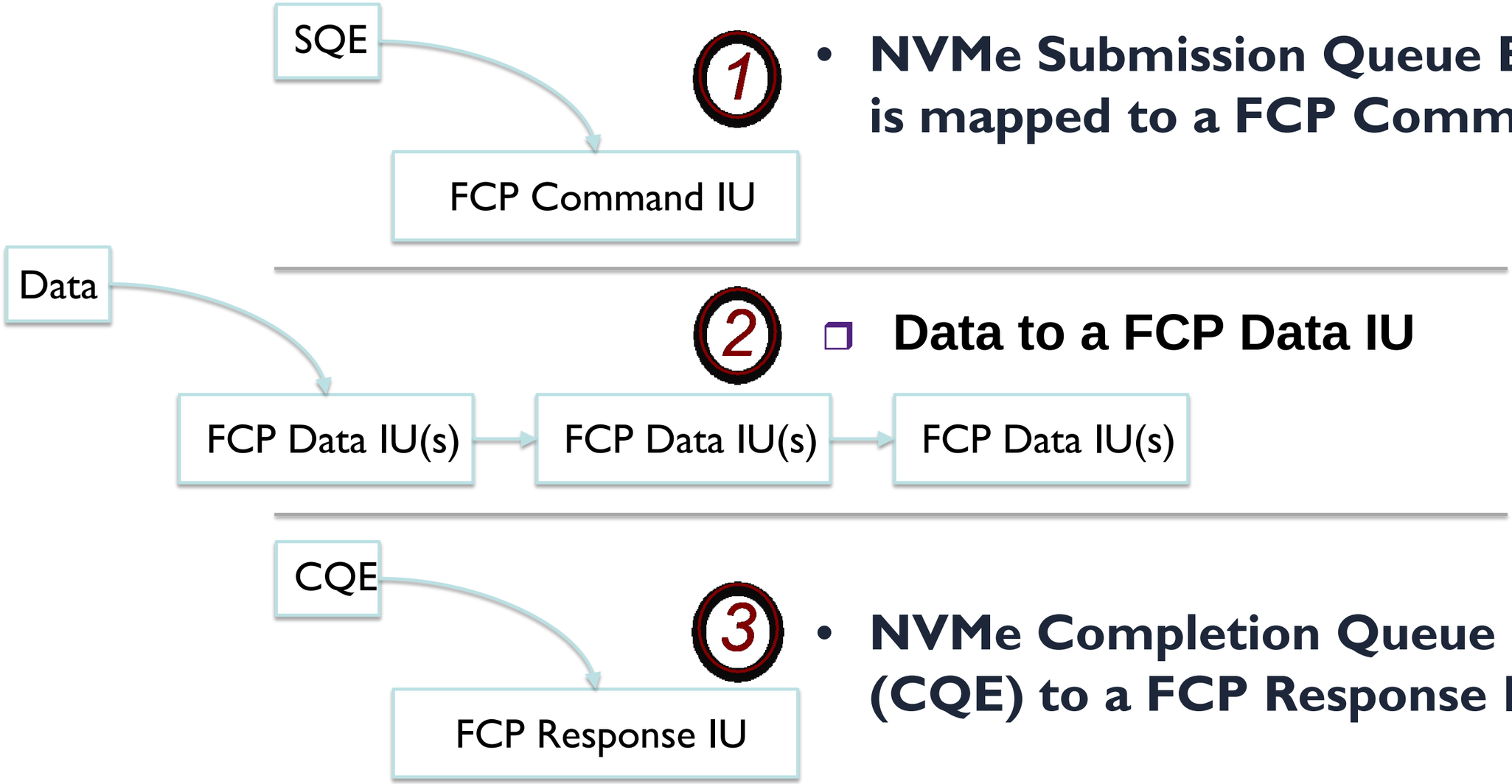
FCP Mapping

- 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

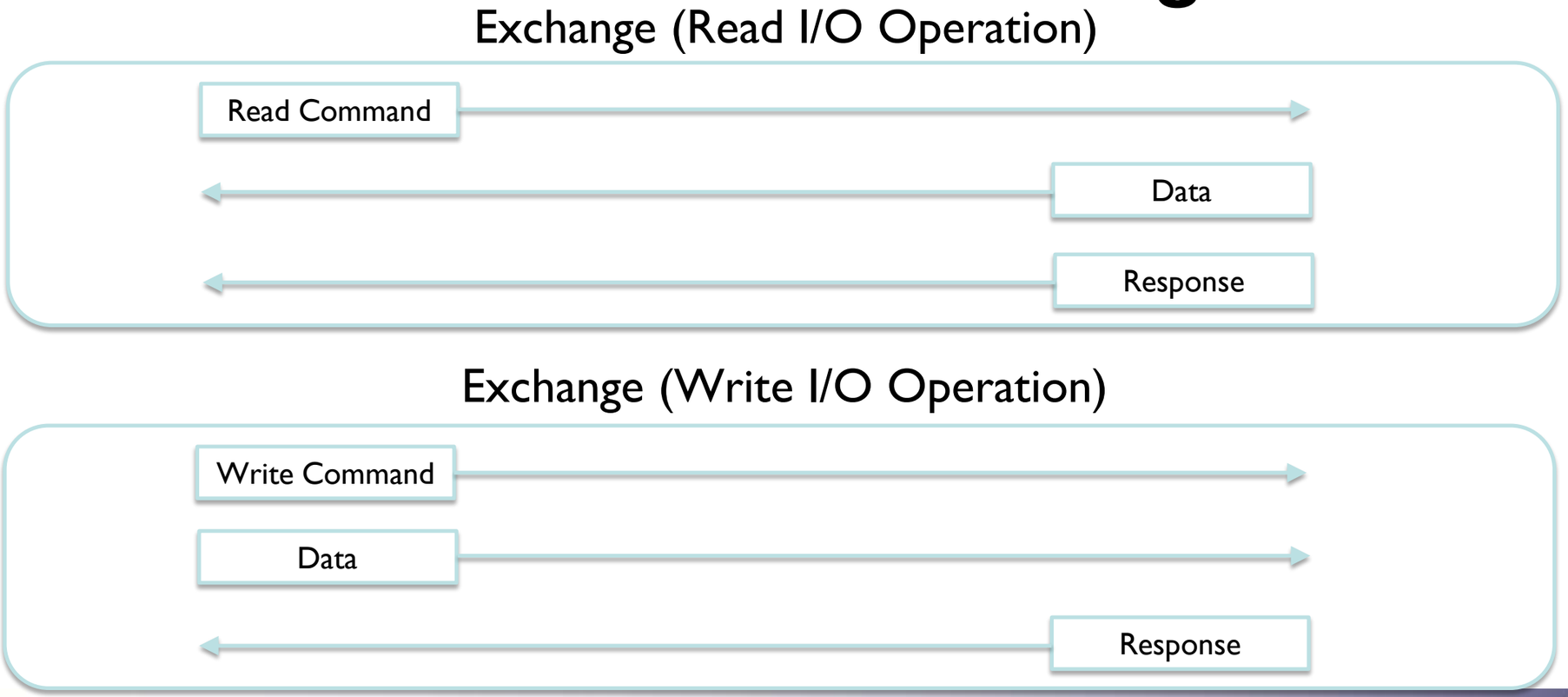
FC-NVMe Information Units (IUs)



- **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

FC-NVMe Discovery

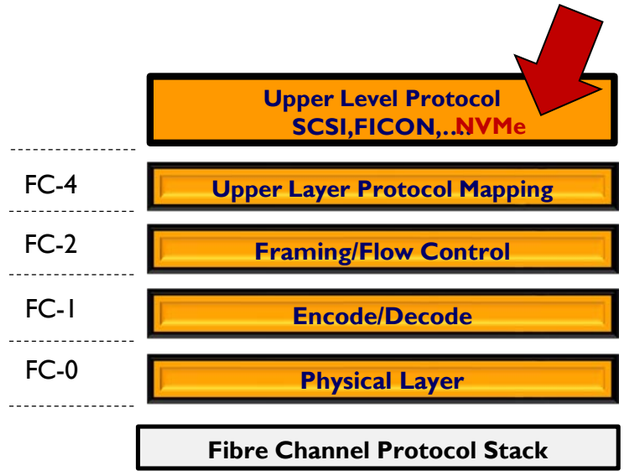
- **FC-NVMe Discovery uses both**
  - FC Name Server to identify FC-NVMe ports
  - NVMe Discovery Service to disclose NVMe Subsystem information for those ports
- **This dual approach allows each component to manage the area it knows about**
  - FC Name Server knows all the ports on the fabric and the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

Zoning and Management

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

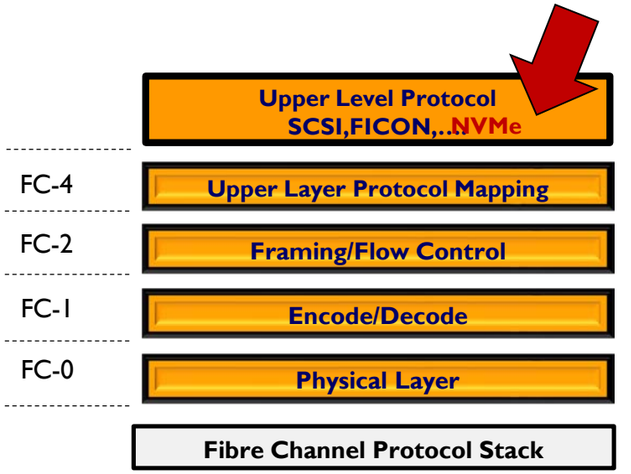
FC-NVMe UPDATE

FC-NVMe Update



- FC-NVMe Standard (T11) ratified on 8/10/2017
  - FC-NVMe Discovery/MPIO in development by Linux community
- FC-NVMe Linux community driver update
  - FC-NVMe drivers upstream in Linux 4.13 kernel
  - FC as a transport support will be available in Unified Host/Target SPDK
- Use existing FC HBA and FC switch hardware
  - Co-existence of FCP SCSI and FC-NVMe traffic
  - Currently shipping HW will support FC-NVMe

FC-NVMe Update



- Performance:
  - Demonstrated low latency high performance
- Availability
  - Linux based host drivers available now
- FC-NVMe-2 Started development spring ’18
  - Focusing on Enhanced Error Recovery

FC-NVMe Ecosystem Readiness

Element	FC-NVMe Support Status
FC Switches	Available today NOTE: Switches don't have to do anything new to support FC-NVMe
HBAs	**Host Side:** • Linux Unified Driver available for download today **Target Side:** • User mode (SPDK), Kernel mode - alpha drivers available today • FW available today
Operating Systems	Linux Community and OS Vendors: • SLES12SP3, RHEL7.5 support FC-NVMe (Tech Preview) today • SLES12SP4/SLES15, RHEL7.6 support FC-NVMe GA Q3/Q4,2018 • VMware and Microsoft – engaged
Storage	Multiple vendors to support 2H 2018

FC-NVMe-2

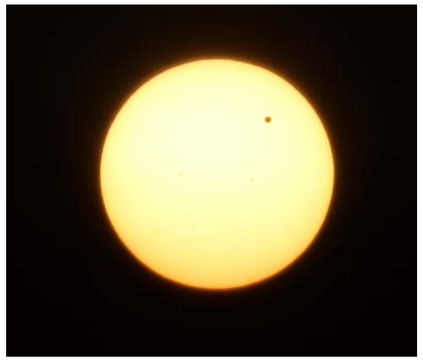
The next step

- The big new item in FC-NVMe-2 is Enhanced Error Recovery
- Allows errors (missing or corrupt frames) to be detected and recovered at the transport layer before the protocol layer knows anything was amiss

Buy Why?

- I thought it was reliable?
  - Bit errors do happen
    - Actual bit errors tend to be much lower than theoretical occurrences
  - Software/hardware errors can also lead to frame loss

What causes Bit Errors



Cosmic Rays from the sun and other sources.

Studies by IBM in the 1990s suggest that computers typically experience about one cosmic-ray-induced error per 256 megabytes of RAM per month.

Radiation from local environment

For modern chips care must be taken to minimize radiation from components

RF and power line noise from local equipment

Even changing generators at local power company can induce low frequency noise

What causes Bit Errors

Software/hardware bugs

Need I say more?

Common specified Bit Error Rate is 10^{-12} to 10^{-15}

Actual bit error rate is often much better, but with theoretical rate, bits could occur multiple times per hour

How did this work before?

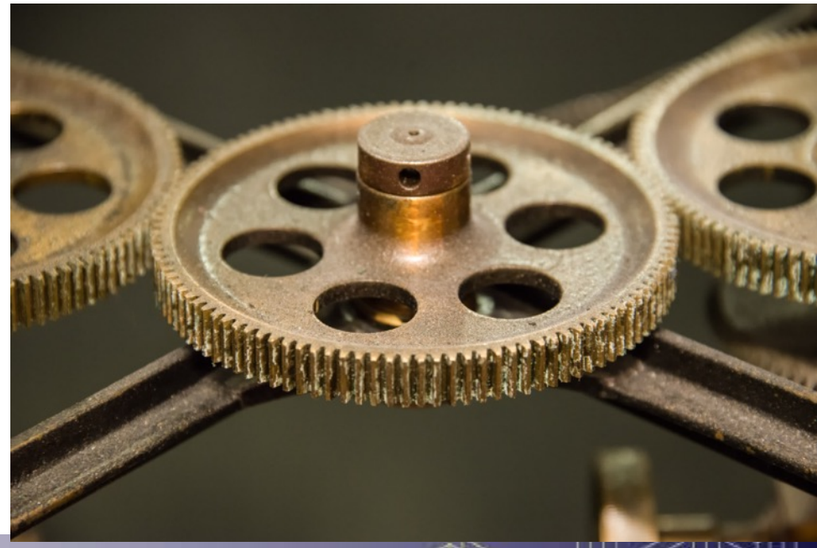
- Limited Error Recovery on the link
  - Low level error detection
  - FEC (Forward Error Correction) on some high speed links
- Protocol Level Error Recovery
  - Both SCSI and NVMe have their own recovery mechanisms

Enhanced Error Recovery

- Goal
  - Don’t let the protocol layer see any errors
    - Don’t want to rely on protocol level error recovery
- Enhanced Error Recovery
  - Detect and recover from errors before they reach the protocol layer
  - Protocol layer doesn’t even know anything happend

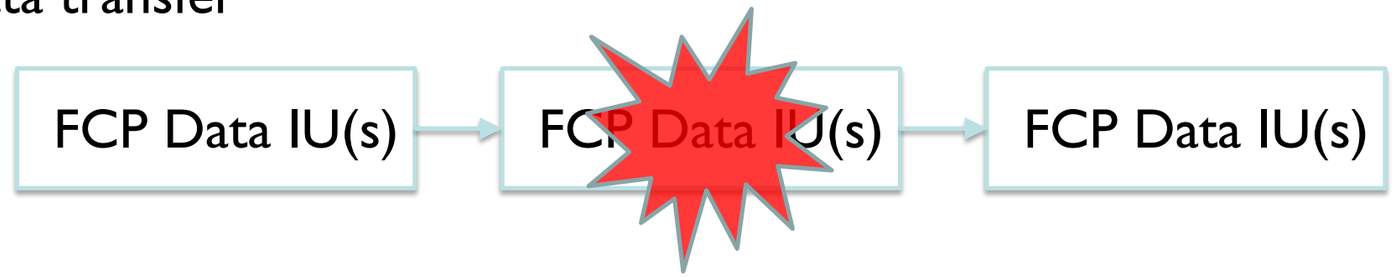
More details on Enhanced Error Recovery

- Error recovery takes place at FC Frame level
  - Missing frames timeout and are retransmitted
  - Defined new FC Basic Link Services for fast recovery
  - Protocol layer does not know anything happened



Example*

FC-NVMe data transfer



Frame loss detected (can be detected in 2 seconds)

Error recovery verifies the frame is really lost though new FLUSH Basic Link Service

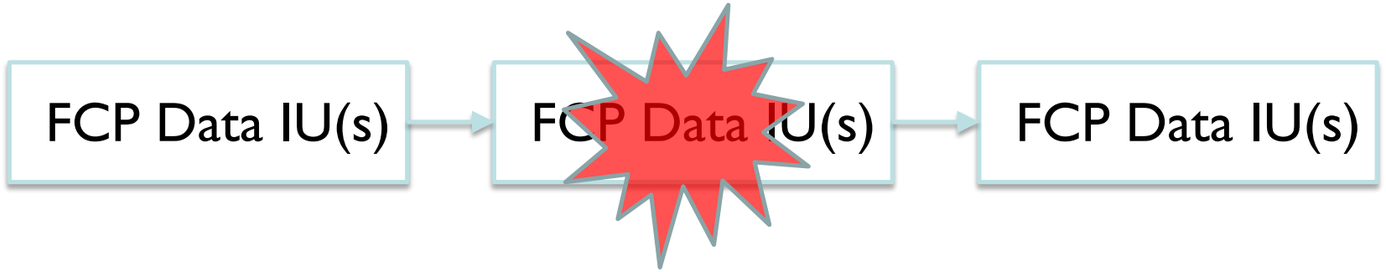
Data is retransmitted

Error is recovered without any knowledge/interaction from upper level protocol

* Actual error recovery process is more complex than shown here

Many different scenarios to cover

Lost FCP Command IU



Lost FCP Data IU

Lost FCP Response IU

Error Recovery Summary

- Goal is to recover from errors without upper level knowing anything happened
  - Recovery in 2 seconds or less
- This is going to be increasingly important as link speeds go up
- Starting with FC-NVMe and applying to SCSI FCP

WHY USE FC-NVMe?

Top 6 Reasons FC-NVMe Might Be The Right Choice

1. **Dedicated Storage Network**

Top 6 Reasons FC-NVMe Might Be The Right Choice

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

Top 6 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 6 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 6 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

Top 6 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
6. With FC-NVMe-2 Industry leading error detection/recovery

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!