ECE566
Enterprise Storage Architecture

Spring 2025

Storage Area Network (SAN)

Tyler Bletsch

Duke University

Adapted from the course "Information Storage and Management v2" (module 5-6), published by EMC corporation.

Includes additional content cited inline.

Classical Fibre Channel SAN

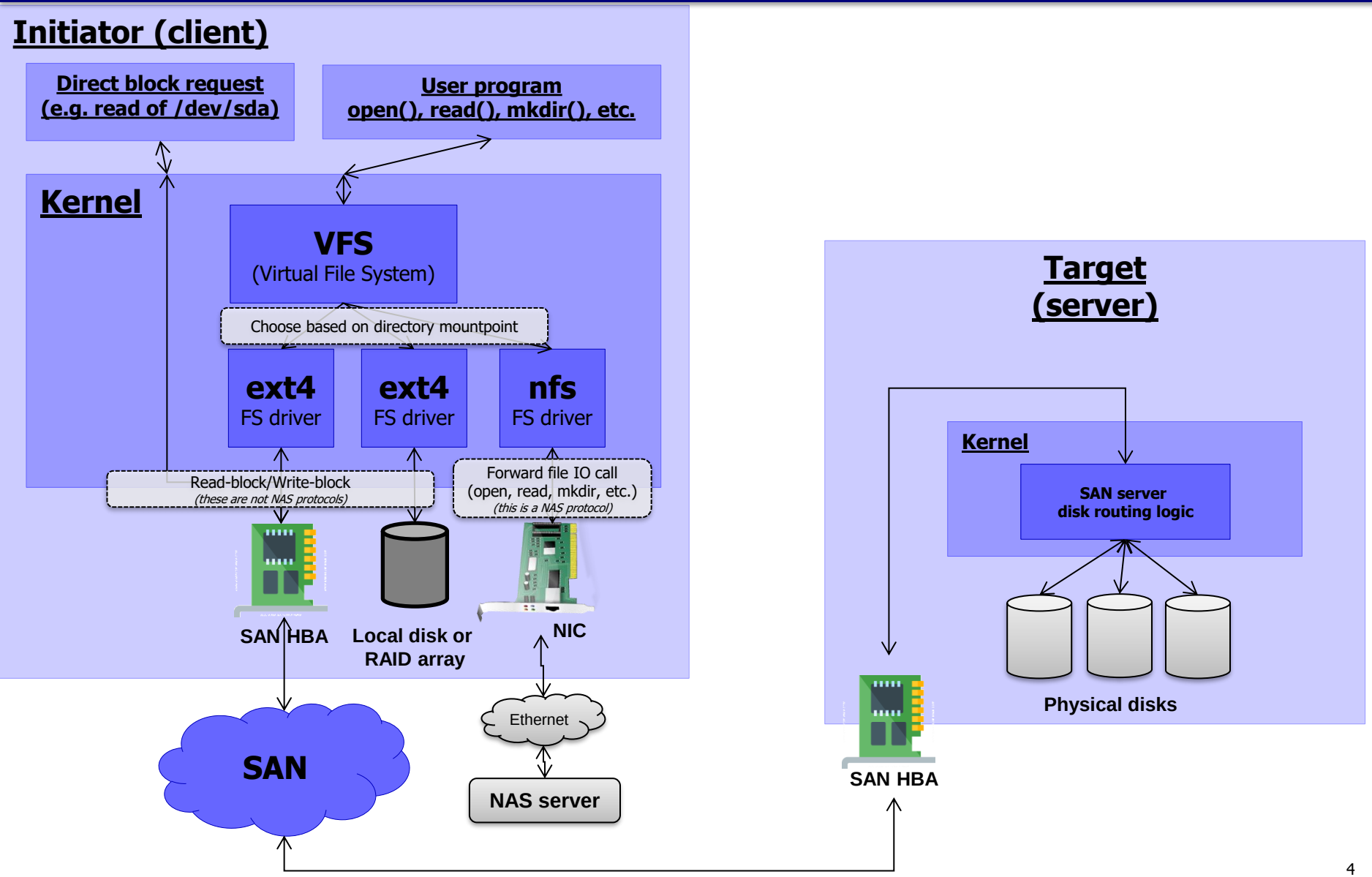
What is a SAN?

SAN

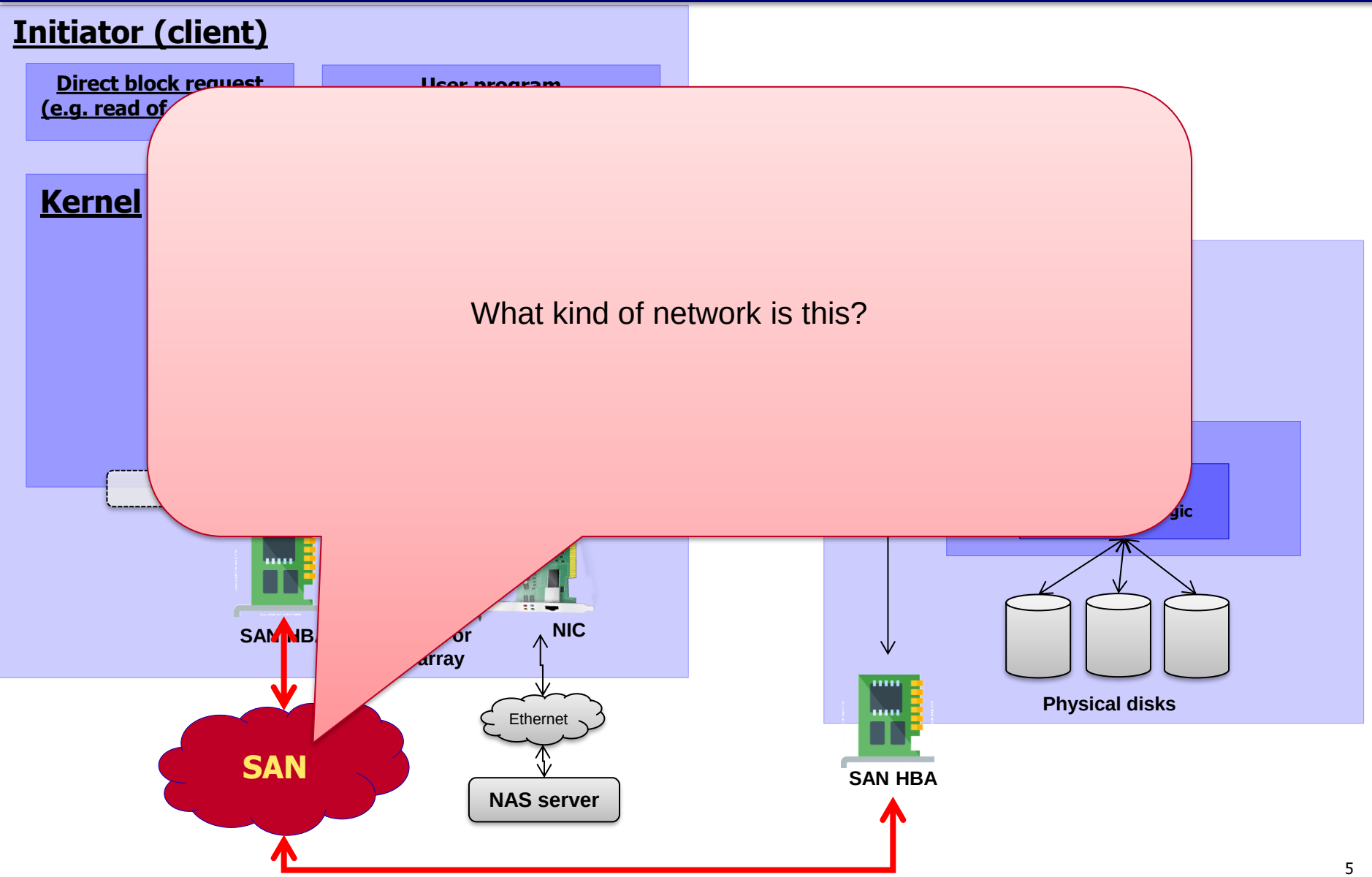
It is a high-speed, dedicated network of servers and shared storage devices.

- Centralizes storage and management
- Enables sharing of storage resources across multiple servers at block level
- Meets increasing storage demands efficiently with better economies of scale
- Common SAN deployments are:
  - Fibre Channel (FC) SAN: uses FC protocol for communication
  - IP SAN: uses IP-based protocols for communication

SAN block diagram



SAN block diagram



History and protocols of SAN: FCP

- Original SAN: **Fibre Channel Protocol (FCP)**
  - A network standard for sending SCSI frames, totally separate from Ethernet/IP
    - In fact, VERY different network design from Ethernet/IP
  - First on market in 1993, this protocol was the only SAN for many years
  - Successive technology versions labeled by data rate
    - 1Gb/s in 1997, 2Gb/s in 2001, 4Gb/s in 2004, … 256Gb/s in 2020
  - Expensive, single purpose network (only for SAN, nothing else)

  - Good news! It's finally dying (kinda)!
    - Despite 128Gb/s and 256Gb/s standards, little hardware for that
    - Some markets still love it (e.g. mainframe, some VM hosts, others)
    - Will probably never "fully" die, but will get smaller and more niche

  - So our coverage will be light

History and protocols of SAN: FCoE

- Weird off-shoot: **Fibre Channel over Ethernet (FCoE)**
  - I used to teach this a bunch – it was very popular for a while!
  - Now it's largely dead except for a few special areas, e.g. Cisco Unified Computing System (UCS)
  - I'm going to skip teaching much of it, except for this:
    - It promised Fibre Channel on the same network as Ethernet ☺
    - Because Fibre Channel was so different, you couldn't run it on 'normal' Ethernet – you needed special "Data Center Ethernet" (DCE), so all new hardware, thus ruining the potential benefit ☹
    - Meanwhile, if you wanted SAN on regular Ethernet, iSCSI came along (see next slide)…
    - Conclusion: why pay huge money when you could pay no money?

Source

History and protocols of SAN: iSCSI

- Cheap option: **iSCSI**
  - It's SCSI protocol over IP
    - The SCSI storage protocol that everything uses!
      +
    - The Ethernet/IP protocols that everything uses!
    - Result: Cheap! ☺
  - Developed in 1998, standardized in 2000
  - This is probably what you'd invent: "Just send SCSI over IP"

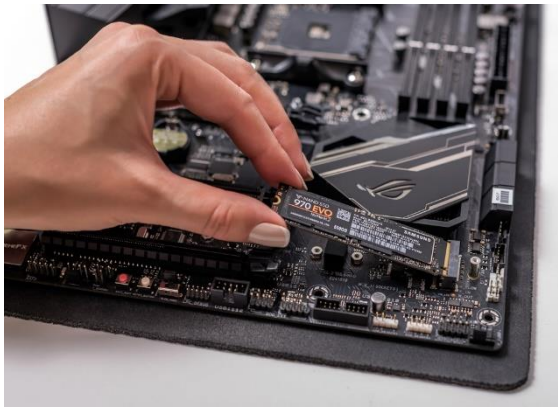
History and protocols of SAN: FCIP (lol)

- Comedy option: **FCIP**
  - It's FCP over IP? Invented after iSCSI existed? Why???
  - This thing died almost immediately upon launch
  - I've never taught it

  - Why bring it up?
    - These slides used to say "FCoE is the future, FCIP is dead"
    - Now they say "FCoE is dead, like FCIP"
    - Predicting the future is hard
    - The industry moves fast
    - Sometimes stuff dies

History and protocols of SAN: NVMe-OF

- New to this course as of 2025: **NVMe-OF** (NVMe-"Over Fabric")
  - Standardized in 2016, adoption is now becoming significant
  - SAN of the future? Who knows?
- NVMe-OF summary:
  - All past SAN protocols were SCSI requests over various networks
    - FCP is SCSI over FC fabric
    - iSCSI is SCSI over IP
    - FCoE was SCSI requests wrapped in FC over special ethernet
  - *Meanwhile*, when SSDs were getting faster, we invented a new drive protocol for them: NVMe
    - Supports SSD-specific operations like TRIM
    - Designed for low-latency, high-throughput (DMA is core to the protocol)
  - Insight: Make a SAN protocol that uses *NVMe* protocol over various fabrics!

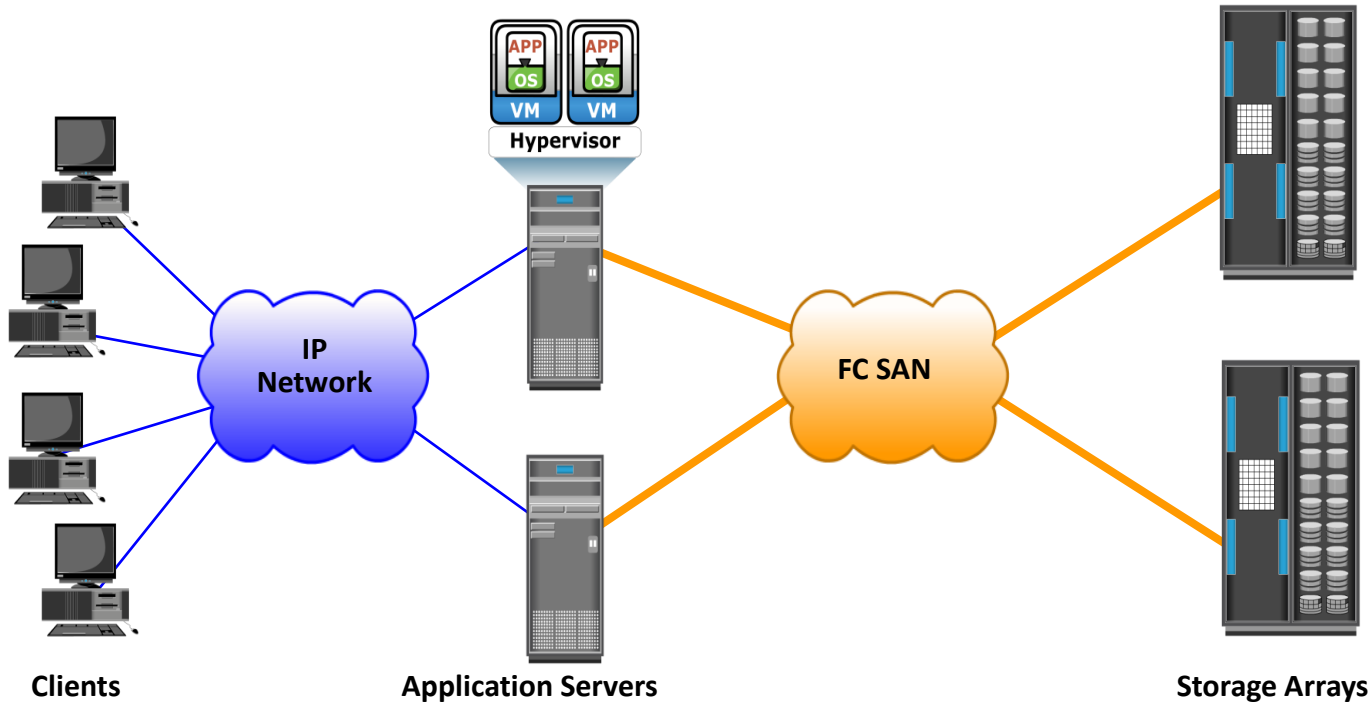


How we'll proceed

- I'm going to show some info about traditional FCP
  - It's still around
  - Most SAN protocols borrow terminology from it
  - We'll skip most of the nasty details
- Then we'll contrast with iSCSI and NVMe-OF

Understanding Fibre Channel

- High-speed network technology
  - Latest FC implementation supports speed up to 256 Gb/s
- Highly scalable
  - Theoretically, accommodate approximately 15 million devices

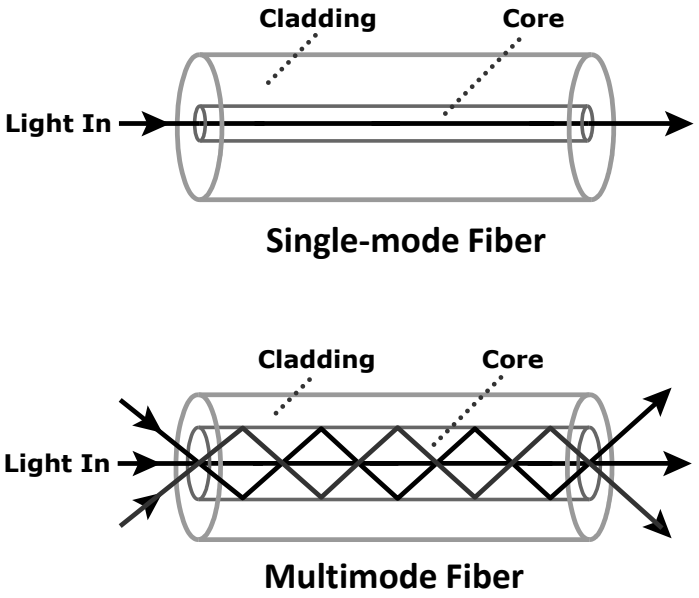


Cables

Note: this is also true for fiber optics used in Ethernet

- SAN implementation uses
  - Copper cables for short distance
  - Optical fiber cables for long distance
- Two types of optical cables: single-mode and multimode

Single-mode	Multimode
Carries single beam of light	Can carry multiple beams of light simultaneously
Distance up to 10km	Used for short distance (Modal dispersion weakens signal strength after certain distance)



Connectors

- Attached at the end of a cable
- Enable swift connection and disconnection of the cable to and from a port
- Commonly used connectors for fiber optic cables are:
  - Standard Connector (SC)
    - Duplex connectors
  - Lucent Connector (LC)
    - Duplex connectors
  - Straight Tip (ST)
    - Patch panel connectors
    - Simplex connectors

Note: this is also true for fiber optics used in Ethernet

Standard Connector

Lucent Connector

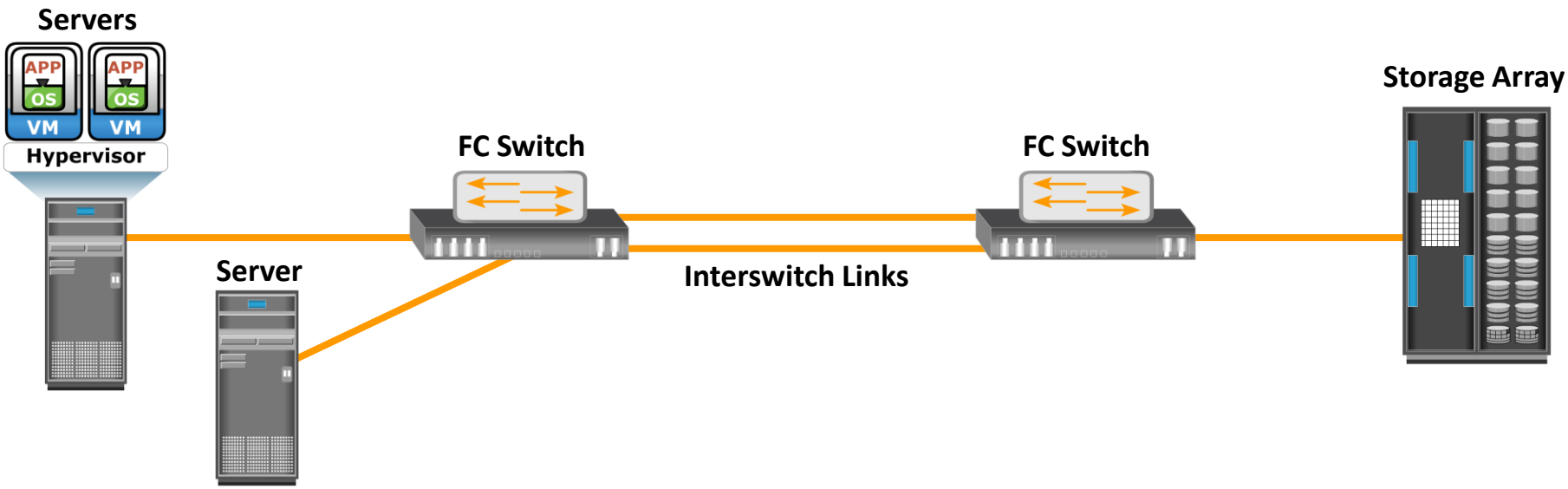
Straight Tip Connector

Interconnecting Devices

- Commonly used interconnecting devices in FC SAN are:
  - Hubs, switches, and directors
- Hubs provide limited connectivity and scalability
- Switches and directors are intelligent devices
  - Switches are available with fixed port count or modular design
  - Directors are always modular, and its port count can be increased by inserting additional 'line cards' or 'blades'
  - High-end switches and directors contain redundant components

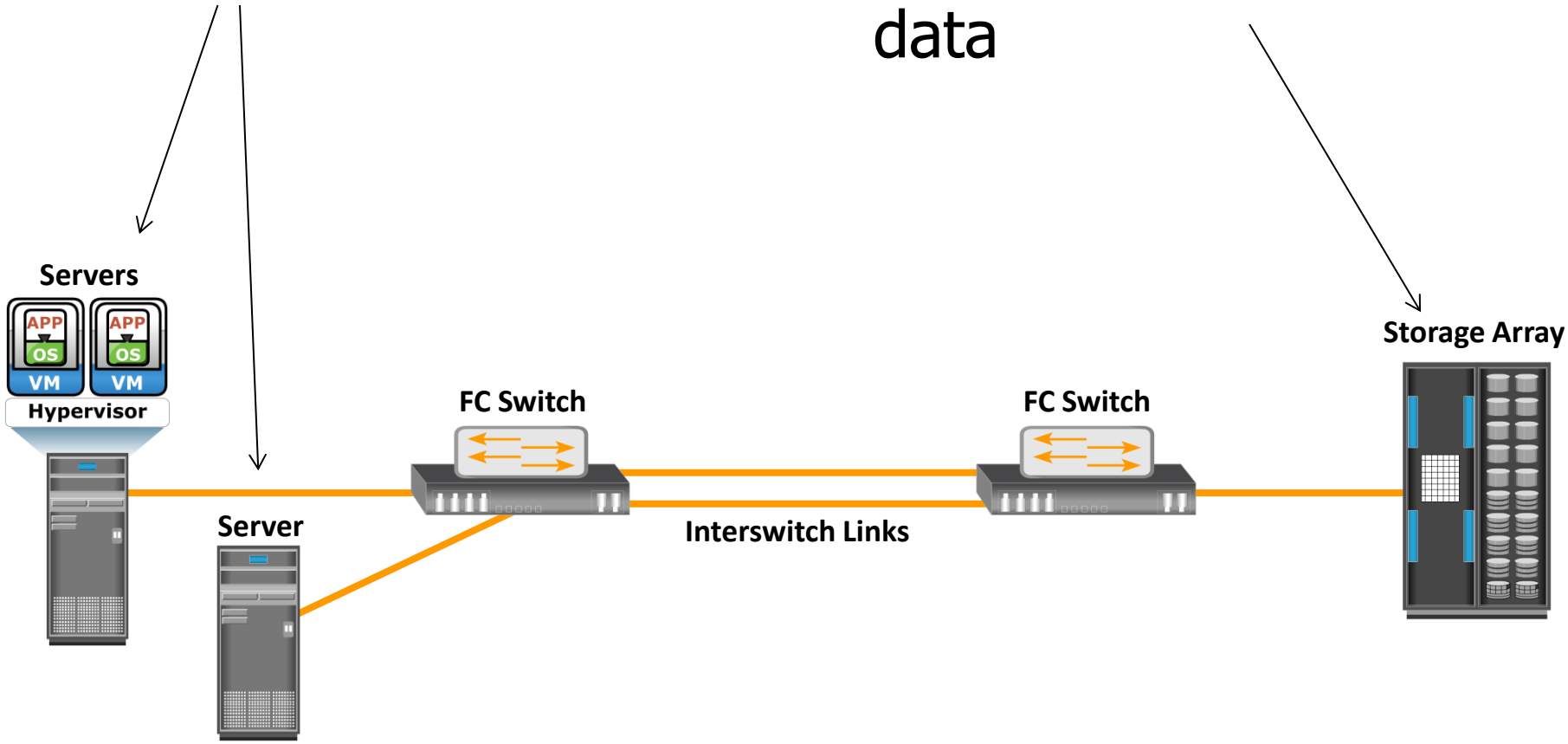
Fibre Channel Switch (FC-SW) Connectivity

- Creates a logical space (called fabric) in which all nodes communicate with one another using switches
  - Interswitch links (ISLs) enable switches to be connected together
- Provides dedicated path between nodes
- Addition/removal of node does not affect traffic of other nodes

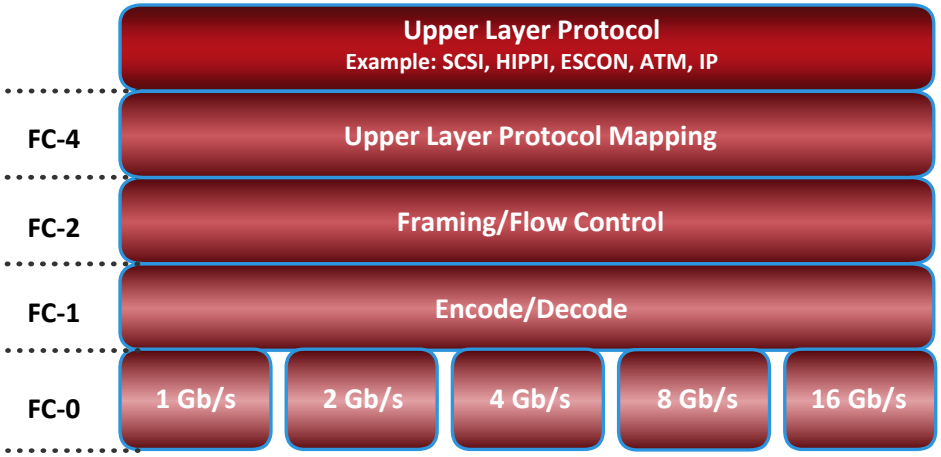


Terminology

- **Initiator**: Client using block devices
- **Target**: Storage controller with the data



Fibre Channel Protocol Stack



FC Layer	Function	Features Specified by FC Layer
FC-4	Mapping interface	Mapping upper layer protocol (e.g. SCSI) to lower FC layers
FC-3	Common services	Not implemented
FC-2	Routing, flow control	Frame structure, FC addressing, flow control
FC-1	Encode/decode	8b/10b or 64b/66b encoding, bit and frame synchronization
FC-0	Physical layer	Media, cables, connector

World Wide Name (WWN)

- Unique 64 bit identifier
- Static to node ports on an FC network
  - Similar to MAC address of NIC
  - World Wide Node Name (WWNN) and World Wide Port Name (WWPN) are used to uniquely identify nodes and ports respectively

World Wide Node Name (for the array)															
5	**0**	**0**	**6**	**0**	**1**	**6**	**0**	**0**	**0**	**6**	**0**	**0**	**1**	**B**	**2**
0101	**0000**	**0000**	**0110**	**0000**	**0001**	**0110**	**0000**	**0000**	**0000**	**0110**	**0000**	**0000**	**0001**	**1011**	**0010**
Format Type	**Company ID 24 bits**						**Port**	**Model Seed 32 bits**							

World Wide Port Name (for the HBA port)															
1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**c**	**9**	**2**	**0**	**d**	**c**	**4**	**0**
Format Type	**Reserved 12 bits**			**Company ID 24 bits**						**Company Specific 24 bits**					

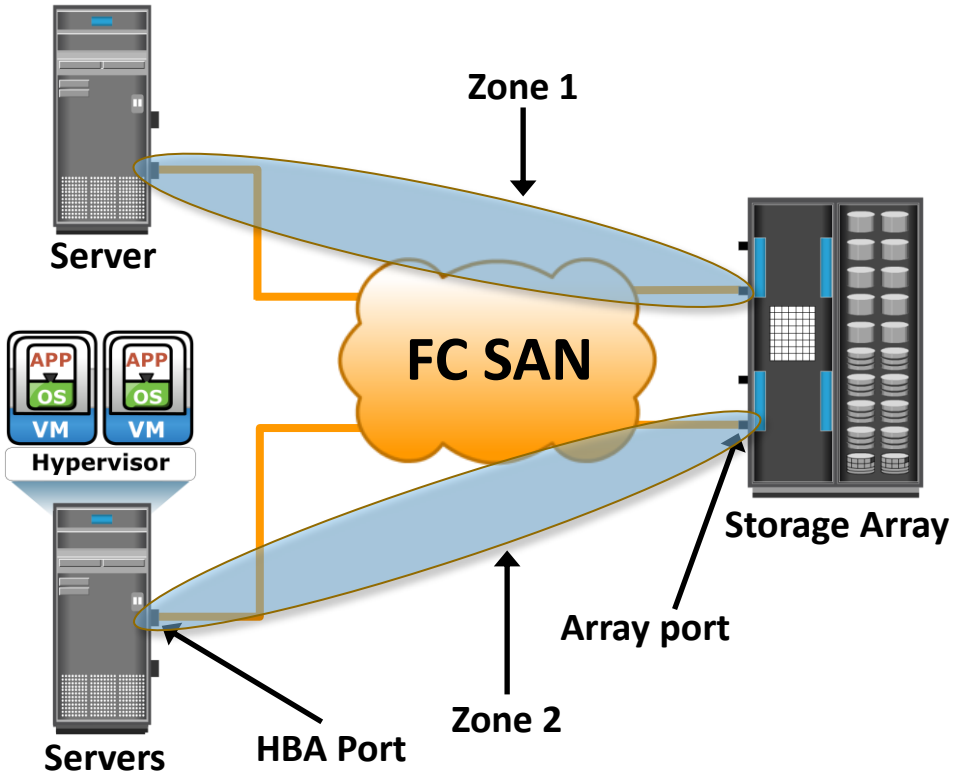
Zoning

Safe to ignore – included just to demonstrate complexity of FCP

Zoning

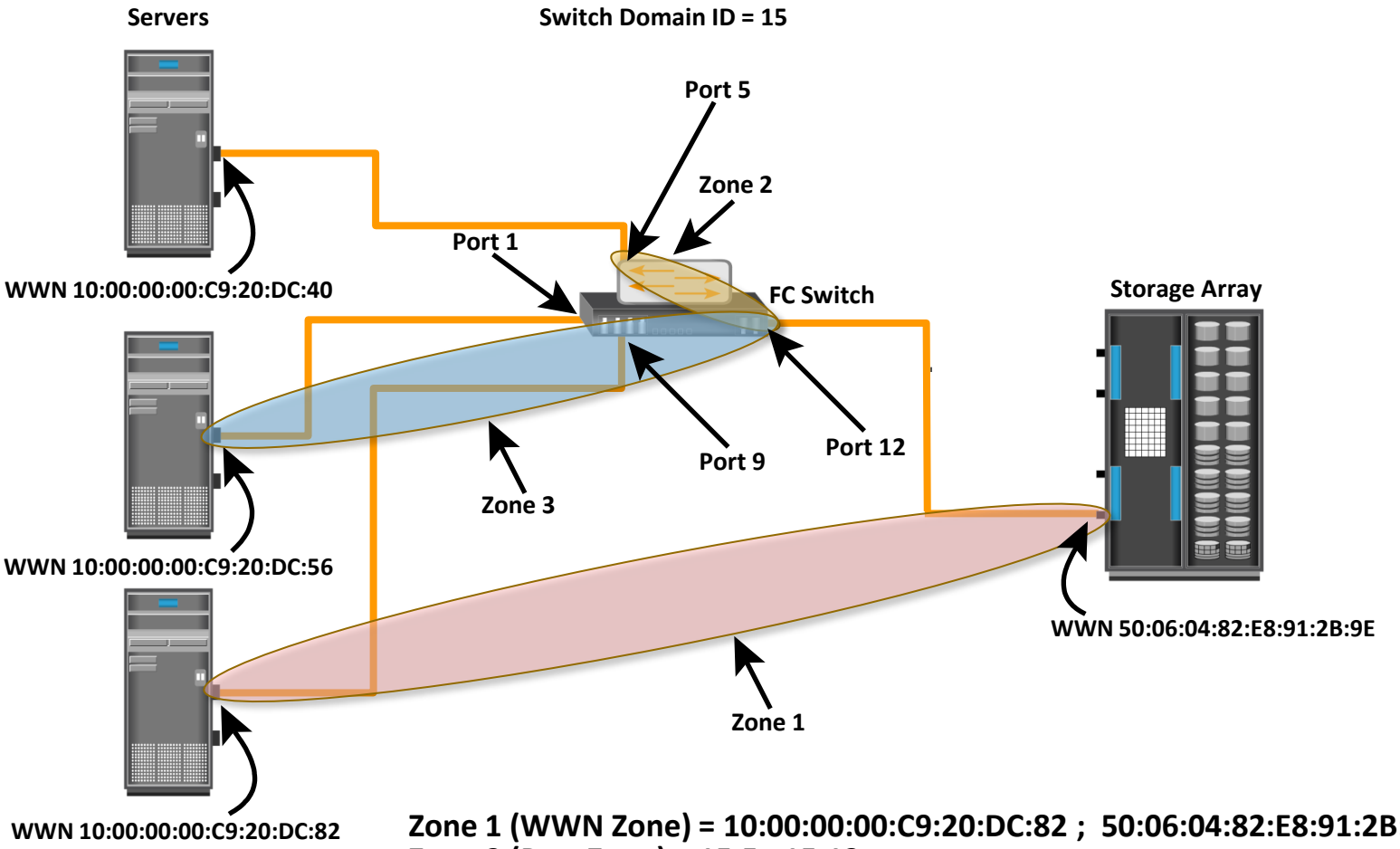
It is an FC switch function that enables node ports within the fabric to be logically segmented into groups, and communicate with each other within the group.

- Zone set comprises zones
- Each zone comprises zone members (HBA and array ports)
- Benefits
  - Restricts RSCN traffic
  - Provides access control



Types of Zoning

Safe to ignore – included just to demonstrate complexity of FCP



Zone 1 (WWN Zone) = 10:00:00:00:C9:20:DC:82 ; 50:06:04:82:E8:91:2B:9E
Zone 2 (Port Zone) = 15,5 ; 15,12
Zone 3 (Mixed Zone) = 10:00:00:00:C9:20:DC:56 ; 15,12

Ethernet SAN: iSCSI

Drivers for IP SAN

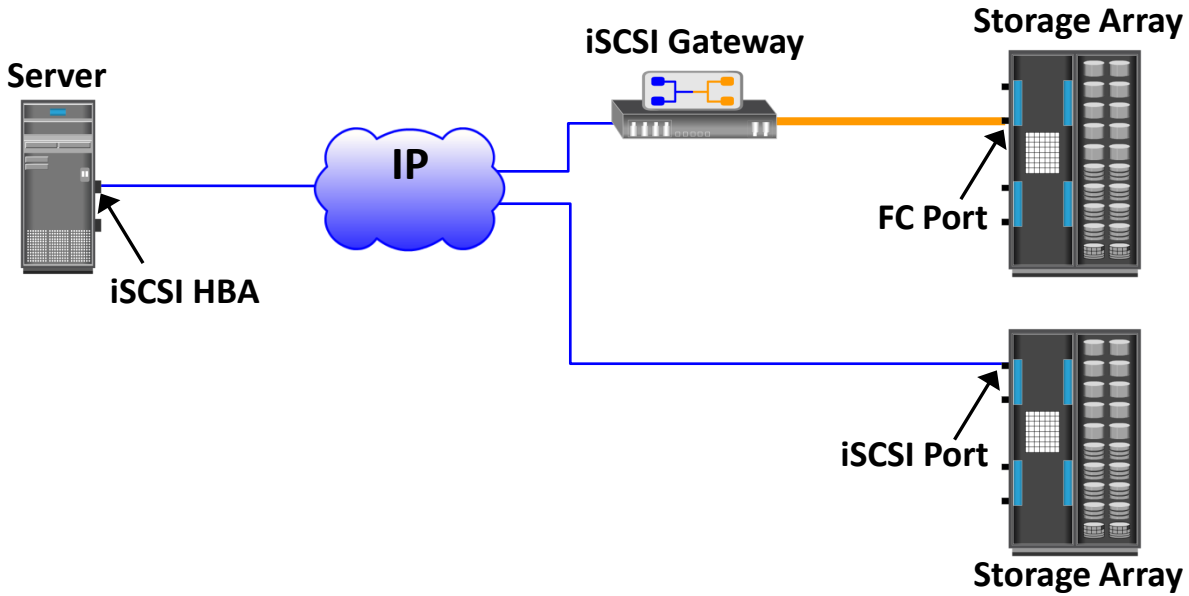
- IP SAN transports block-level data over IP network
- IP is being positioned as a storage networking option because:
  - Existing network infrastructure can be leveraged
  - Reduced cost compared to investing in new FC SAN hardware and software
  - Many long-distance disaster recovery solutions already leverage IP-based network
  - Many robust and mature security options are available for IP network

IP SAN Protocol: iSCSI

- IP based protocol that is used to connect host and storage
- Encapsulates SCSI commands and data into an IP packet and transports them using TCP/IP

Components of iSCSI

- iSCSI initiator
  - Example: iSCSI HBA (hardware, rare nowadays)
  - Example: iSCSI software initiator (pure software, common)
- iSCSI target
  - Storage array with iSCSI support
  - iSCSI gateway – enables communication with FC storage array
- IP network

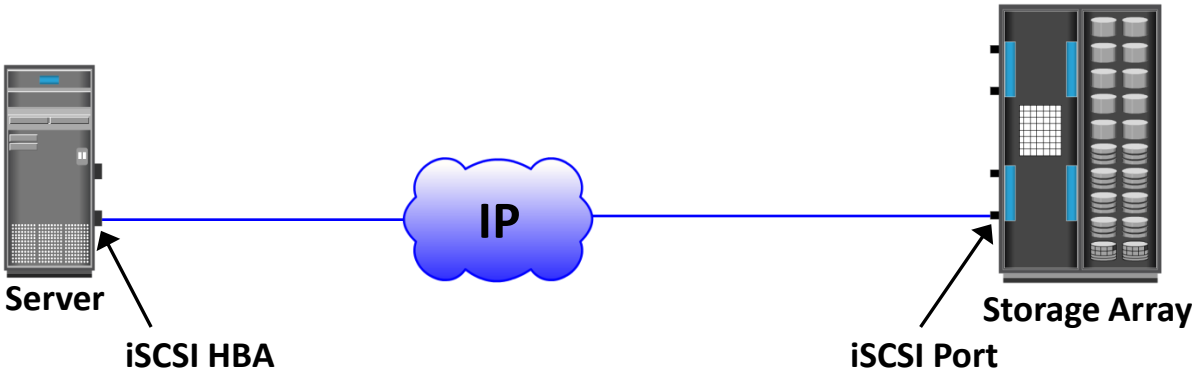


iSCSI Host Connectivity Options

- Standard NIC with software iSCSI initiator
  - NIC provides network interface
  - Software initiator provides iSCSI functionality
  - Requires host CPU cycles for iSCSI and TCP/IP processing
- TCP Offload Engine (TOE) NIC with software iSCSI initiator
  - Moves TCP processing load off the host CPU onto the NIC card
  - Software initiator provides iSCSI functionality
  - Requires host CPU cycles for iSCSI processing
- iSCSI HBA
  - Offloads both iSCSI and TCP/IP processing from host CPU
  - Simplest option for boot from SAN

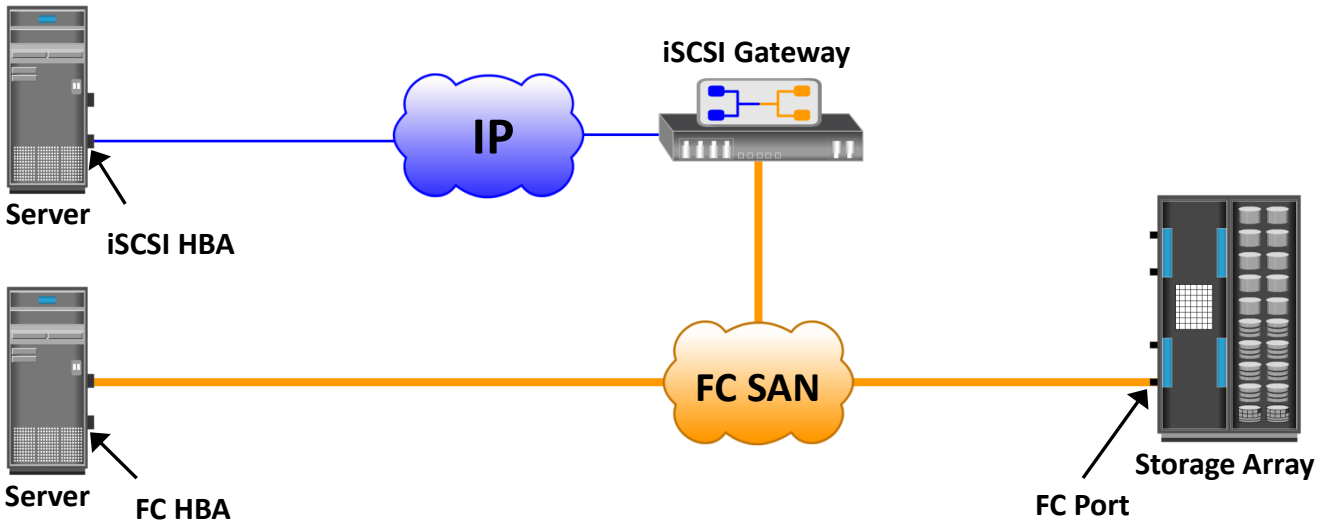
iSCSI Topologies: Native iSCSI

- iSCSI initiators are either directly attached to storage array or connected through IP network
  - No FC component
- Storage array has iSCSI port
- Each iSCSI port is configured with an IP address



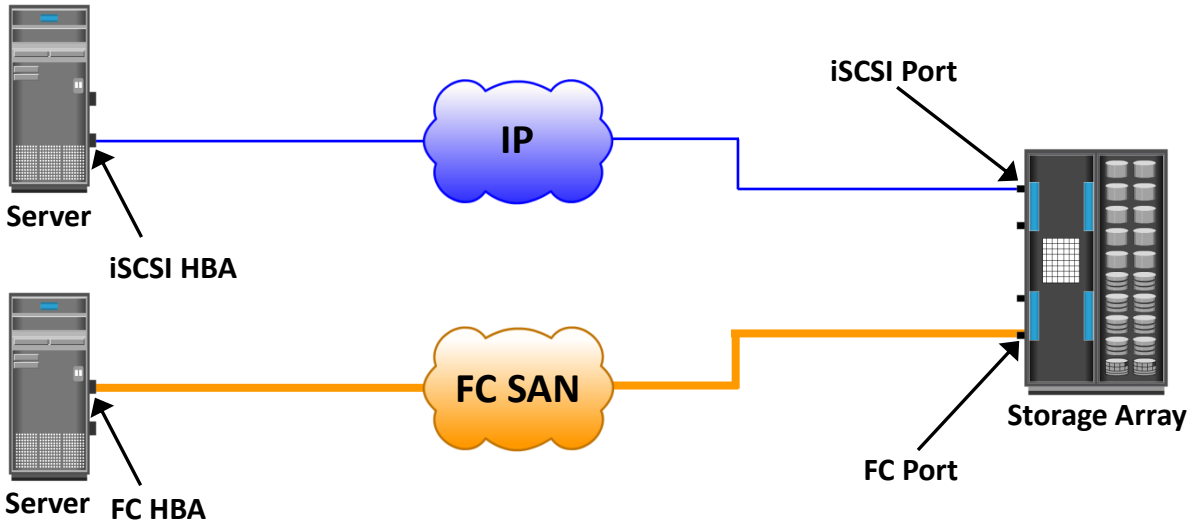
iSCSI Topologies: Bridged iSCSI

- iSCSI gateway is used to enable communication between iSCSI host and FC storage
- iSCSI gateway works as bridge between FC and IP network
  - Converts IP packets to FC frames and vice versa
- iSCSI initiator is configured with gateway's IP address as its target
- iSCSI gateway is configured as FC initiator to storage array

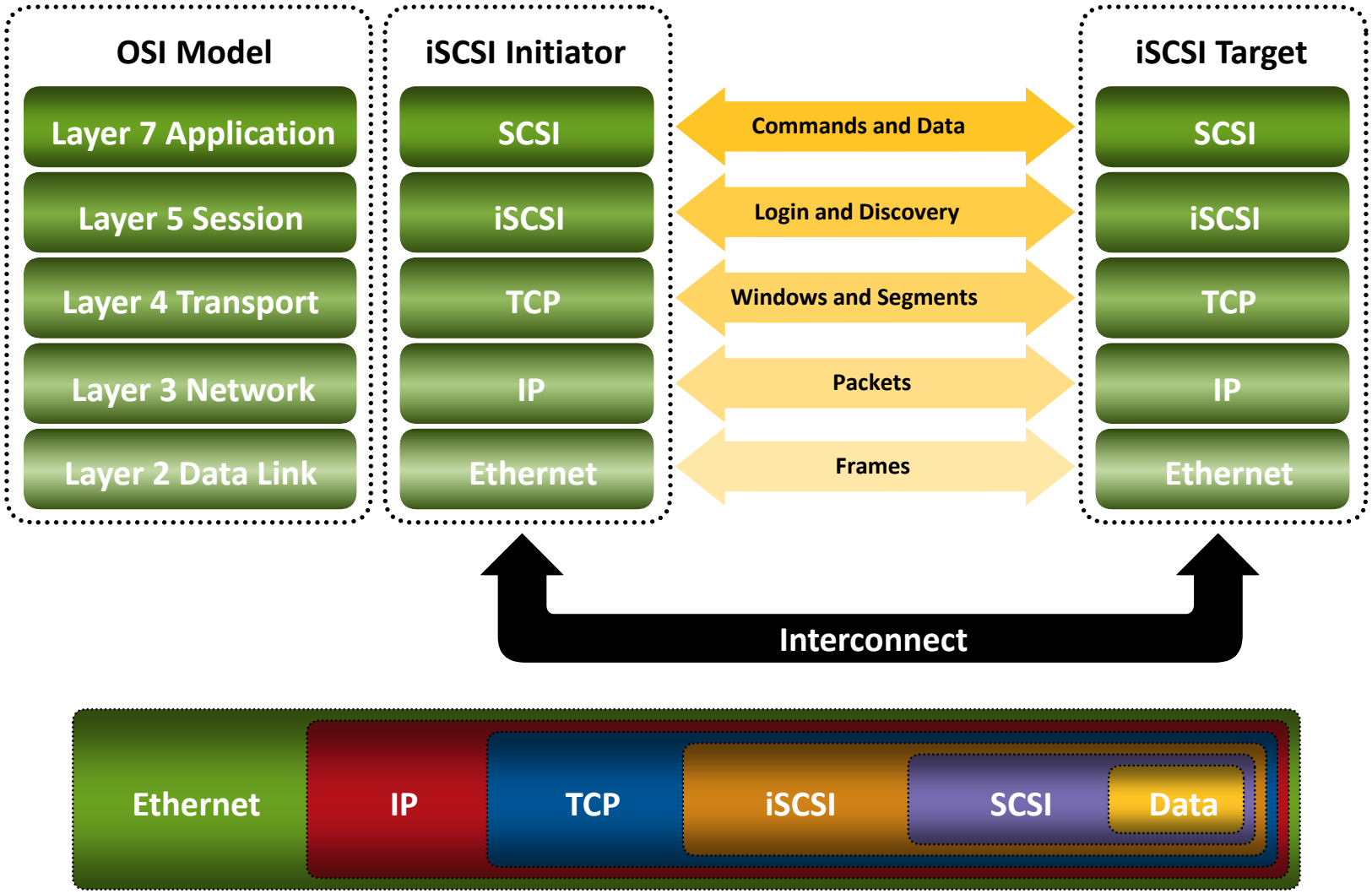


Combining FC and Native iSCSI Connectivity

- Array provides both FC and iSCSI ports
  - Enable iSCSI and FC connectivity in the same environment
  - No bridge devices needed



iSCSI Protocol Stack



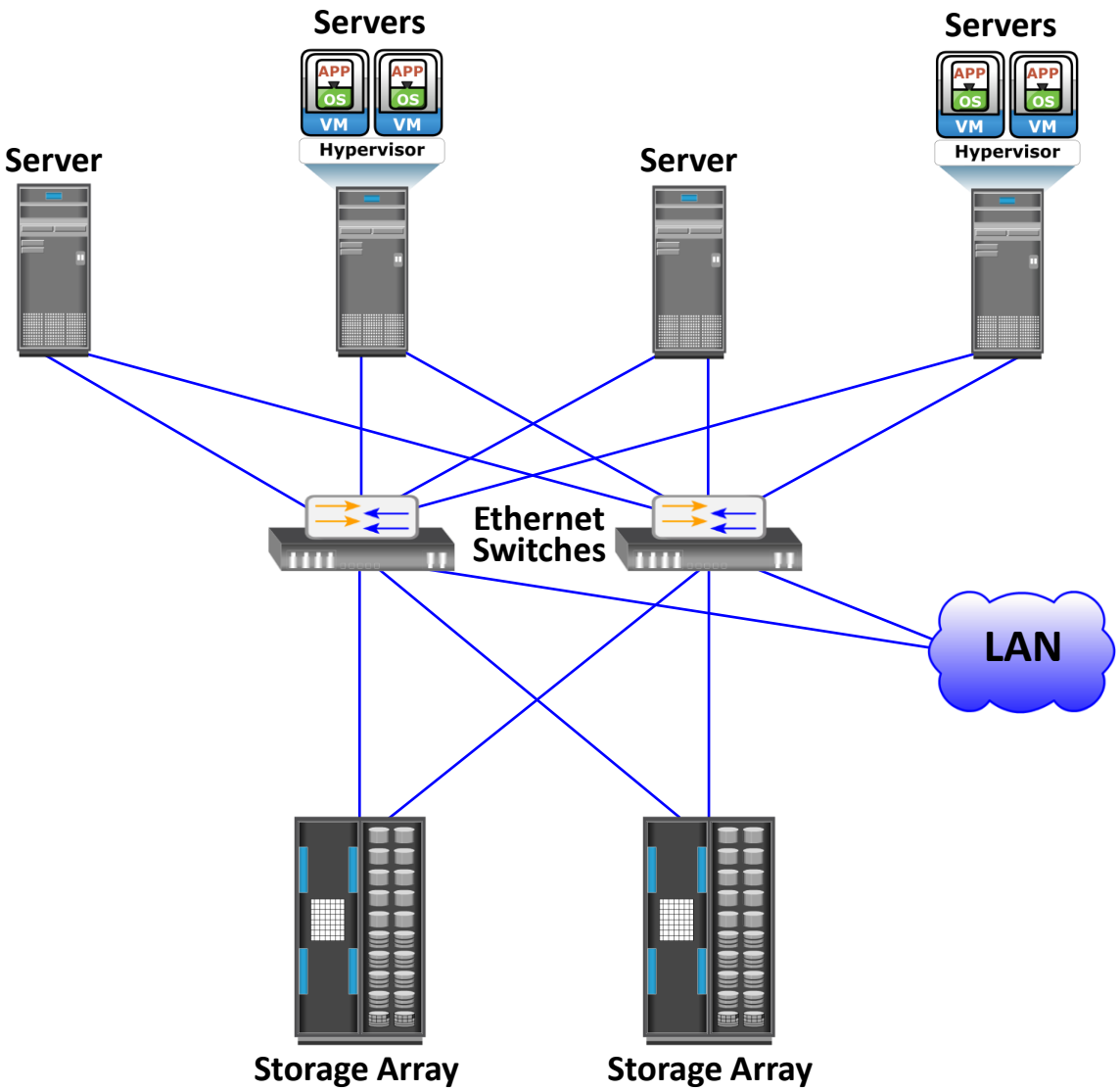
iSCSI Discovery

- For iSCSI communication, initiator must discover location and name of target on a network
- iSCSI discovery takes place in two ways:
  - SendTargets discovery
    - Initiator is manually configured with the target's network portal
    - Initiator issues SendTargets command; target responds with required parameters
  - Internet Storage Name Service (iSNS)
    - Initiators and targets register themselves with iSNS server
    - Initiator can query iSNS server for a list of available targets

iSCSI Name

- iSCSI name is a unique iSCSI identifier that is used to identify initiators and targets within an iSCSI network
- Two common types of iSCSI names are:
  - iqn: iSCSI Qualified Name
    - iqn.2008-02.com.example:optional_string
  - eui: Extended Unique Identifier
    - eui.0300732A32598D26

Data Center Infrastructure – with iSCSI

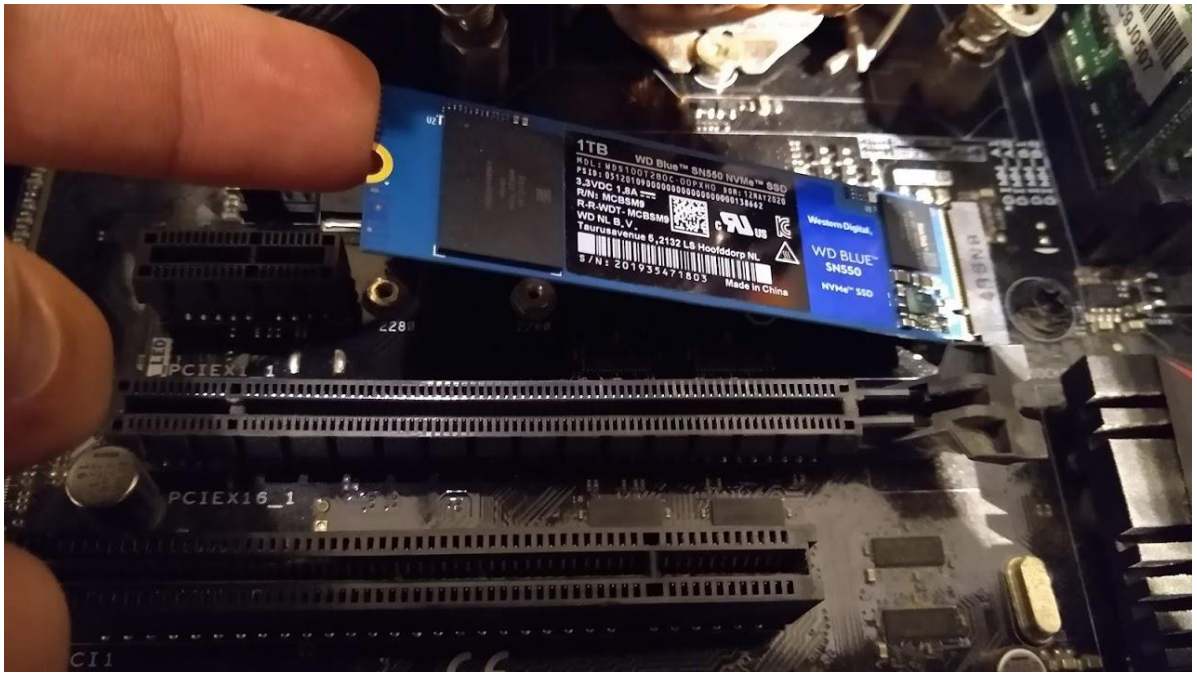


NVMe-OF

"Non-Volatile Memory Express – Over Fabric"

First, what's NVMe again?

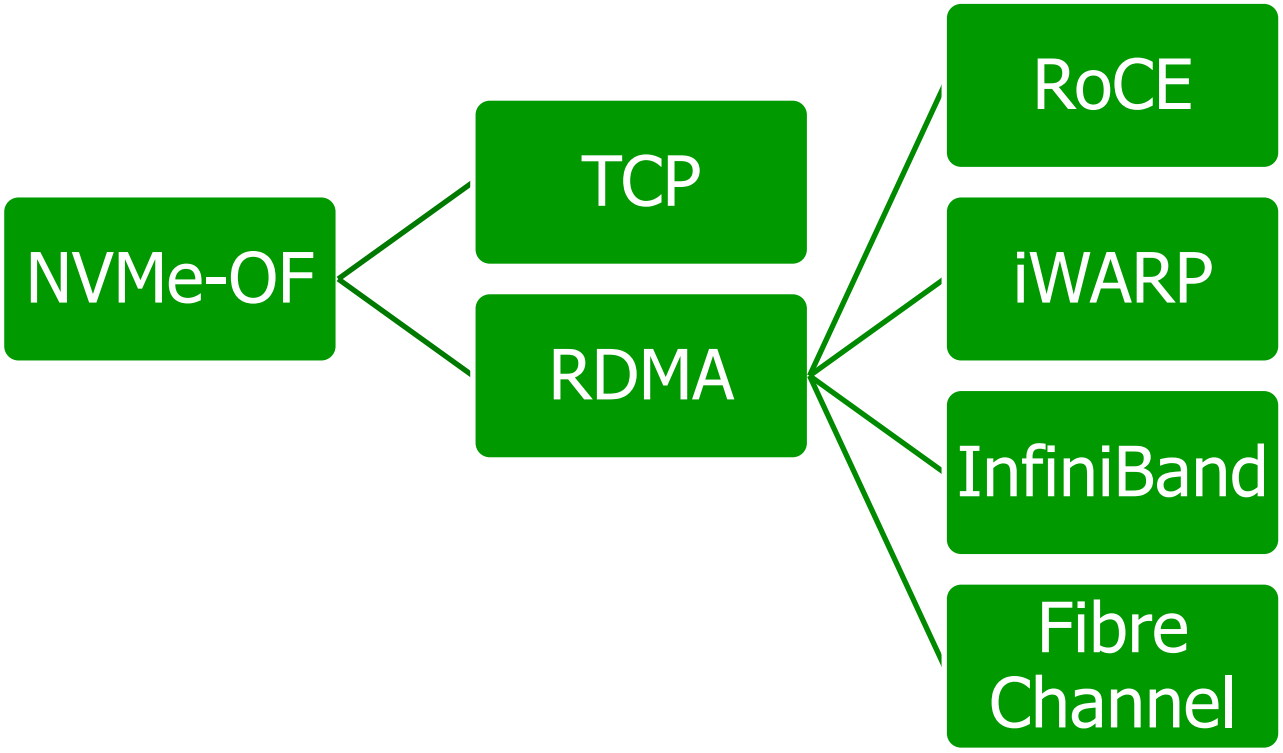
- NVMe (Non-Volatile Memory Express)
  - Newer standard for SSDs that don't have to pretend to be HDDs
  - Low-latency, high-throughput; directly hooks into the PCIe (Peripheral Component Interconnect Express) bus, the main system bus in most modern computers
  - Defines the physical interface (these little cards) as well as a communication protocol *that's not SCSI (wow!)*



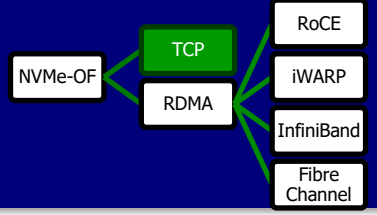
Introducing NVMe-OF

- **NVMe-OF:** NVMe-"Over Fabric"
  - Put the NVMe communication protocol over something other than a PCIe bus inside a computer – put it on a network or **"fabric"**

- But what kind of "fabric"?

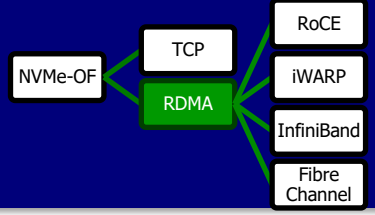


Choice of fabric: TCP



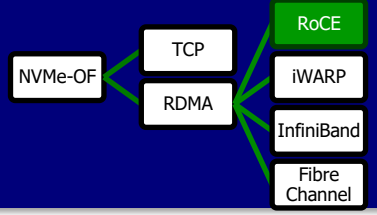
- TCP
  - "Regular" protocol, uses standard network cards on normal Ethernet
  - CPU is involved in sending receiving packets, so worse throughput/latency compared to RDMA (discussed next)
  - Common use cases:
    - Anyone with a large existing network: cloud providers, enterprise datacenters, virtual environments, big data analytics
    - We'll use NVMe-OF with TCP in an upcoming lab

Choice of fabric: RDMA in general



- First, recall DMA (Direct Memory Access)
  - Have IO device put stuff to/from memory without bothering the CPU
- This is RDMA (Remote Direct Memory Access)
  - Take this idea of over a whole network:
    - Data is read from RAM and put onto network by the sender's NIC
    - Data is received from network and put into RAM by the NIC
  - Result: "**Zero copy**" – CPU not involved all the way from end to end!
  - Several options:
    - **RoCE**: Needs special NICs, special network (but it's Ethernet-like)
    - **iWARP**: Needs special NIC, normal Ethernet network
    - **Infiniband**: Needs special NICs, special network, typically small scale and special-purpose
  - Use cases: *Performance critical*
    - High-Performance Computing (HPC), financial services, video production, high performance cloud infrastructure

Choice of fabric: RoCE



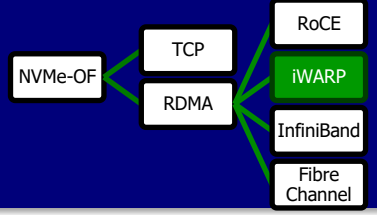
RoCE (RDMA over Converged Ethernet)

- Normal Ethernet is allowed to drop packets at any time, but "Converged Ethernet" has features to allow certain traffic to be **lossless** (guaranteed delivery)
  - Skipping details here – look up "Data Center Bridging" if interested
  - This is why we need special NICs and network hardware – every device must participate in these special protocols!

- RoCE traffic gets lower latency than normal Ethernet
- Can use same network for everything else, hence "converged"

- Many vendors make cards capable of RoCE (Mellanox, Broadcom, Intel)

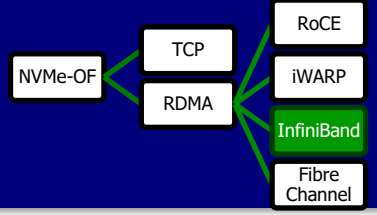
Choice of fabric: iWARP



iWARP (Internet Wide Area RDMA Protocol)

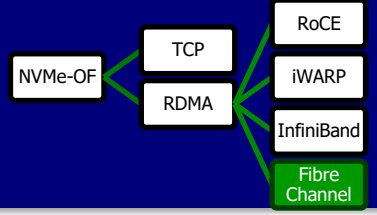
- Use special Ethernet cards that do RDMA (zero copy), but send the info over regular TCP network
- iWARP has cost and performance between plain TCP and RoCE
- Fewer vendors make cards capable for iWARP (Chelsio, Broadcom)

Choice of fabric: InfiniBand



- InfiniBand has been around for a while. The idea is: "I need low latency for a small network and I'm willing to spend money to get it!"
- Used in data centers, supercomputers, high-end cloud, high-frequency financial trading
  - In use long before NVMe-OF, but NVMe-OF is a good match for it
- Special *everything*, and you're likely to have it *in addition to* your traditional Ethernet network

Choice of fabric: Fibre Channel



- First, note that Fibre Channel is itself two things:
  - Fibre Channel *the fabric* (the network communications protocol)
  - Fibre Channel *the storage protocol* (the way it wraps SCSI requests)
- We can use Fibre Channel *the fabric* as the network for NVMe-OF!
  - This is called FC-NVMe

- Why?
  - If you already have an FC SAN, but you want to move toward NVMe
  - That's pretty much it – it has higher cost and complexity than RoCE or InfiniBand, so it's usually just done for migration purposes

Terminology

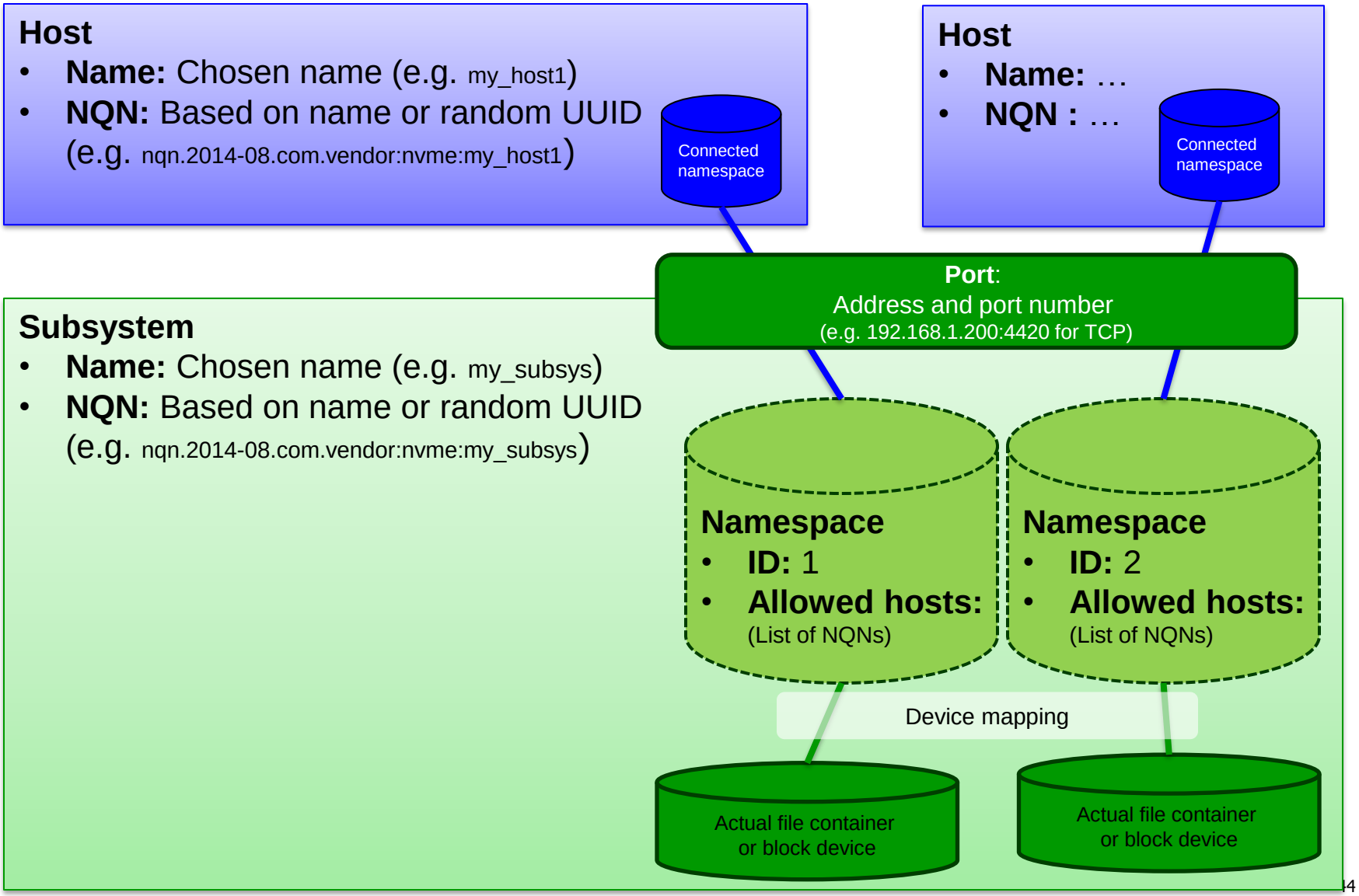
- Okay, I have my fabric! How do I configure NVMe-OF?
  - Because it's a new SAN protocol, all new weird and counterintuitive names were selected to help confuse you, sorry ☹
- To summarize all SAN terminology in one place:

FCP/iSCSI term	NVMe term	Definition
Initiator	**Host**	The machine that wants to connect to and make use of storage ("client")
Target	**Subsystem**	The machine that has the storage for others to use ("server")
LUN	**Namespace**	The block device being accessed via SAN
WWN / IQN	**NQN (NVMe Qualified Name)**	A unique identifier for the host or subsystem in question, similar to iSCSI

Setup of an NVMe-OF environment

Host
- **Name:** Chosen name (e.g. my_host1)
- **NQN:** Based on name or random UUID (e.g. nqn.2014-08.com.vendor:nvme:my_host1)

Host
- **Name:** …
- **NQN :** …



Summary

- Classic Fibre Channel Protocol (FCP):
  - Special-purpose network that wraps SCSI
  - Lossless circuit-switched network
  - Hosts talk via Host-Bus Adapters (HBAs)
  - Hosts labeled with World-wide Node Name (WWNN); each port has its own World-wide Port Name (WWPN)
- iSCSI
  - Send SCSI packets over plain TCP/IP Ethernet networks
  - Can be bridged to classic FCP
  - Endpoints talk via plain network card (in software) or HBA (hardware)
  - Endpoints labeled with iSCSI Qualified Names (IQNs)
- NVMe-OF
  - Newer SSD-focused SAN protocol
  - Available both via plain TCP or various high performance RDMA fabrics

SAN protocol tradeoffs

- Classic Fibre Channel (FC):
  - + Separate network no contention between storage and regular traffic
  - – Separate network means having to buy more hardware
  - ~ Available in physical links up to 64Gb/s
  - + Compatible with all the classic FC stuff you may already have; mature/reliable
- iSCSI:
  - + Uses commodity cheap Ethernet
  - – Storage traffic is in contention with regular traffic
  - ~ Ethernet available in 1Gb (very cheap), 10Gb (common in datacenter), 40Gb (somewhat common), 100Gb (expensive)
- NVMe-OF:
  - + Available both using commodity cheap Ethernet and high-performance fabrics
  - + Extract better performance if target storage is NVMe-based SSD
  - – More complex than iSCSI, high-perf fabrics cost comparable to FC

Questions?

Backup slides: FCoE

Dancing on the grave of

Lesson 2: Fibre Channel over Ethernet (FCoE)

During this lesson the following topics are covered:

- Drivers for FCoE
- Components of FCoE network
- FCoE frame mapping
- Converged Enhanced Ethernet (CEE)

FCoE Motivation, translated

(My personal take on the evolution of FCoE)

Pro-FCoE: "Oh my god, can we stop running two networks with separate hardware, terminology, cables, and teams? It's just data!"

Anti-FCoE: "But-but-but– my job security!! I mean, uh, your packets will be haunted if they don't pass through this special network that costs 10 times a commodity Ethernet network!"

Compromise: We'll just make a variant of Ethernet so the hardware still costs 10 times as much, then everyone's happy!

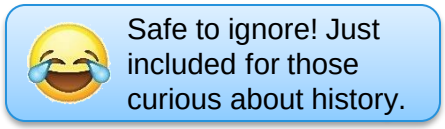
FCoE Motivation, translated

(My updated personal take on the evolution of FCoE)

Pro-FCoE: *doesn't exist*

Anti-FCoE: "lmao it's dead"

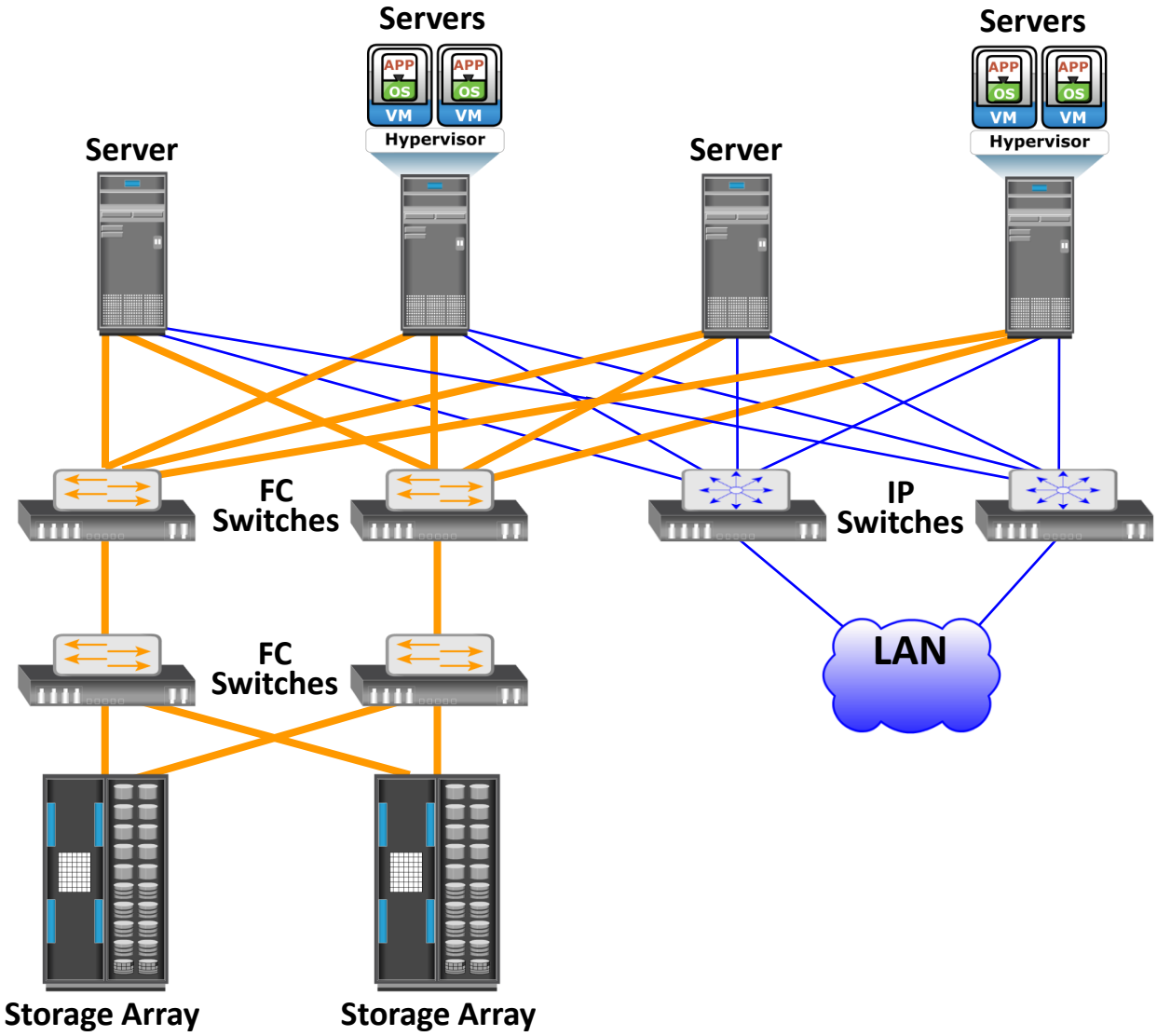
Drivers for FCoE



- FCoE is a protocol that transports FC data over Ethernet network (Converged Enhanced Ethernet)
- FCoE is being positioned as a storage networking option because:
  - Enables consolidation of FC SAN traffic and Ethernet traffic onto a common Ethernet infrastructure
  - Reduces the number of adapters, switch ports, and cables
  - Reduces cost and eases data center management
  - Reduces power and cooling cost, and floor space

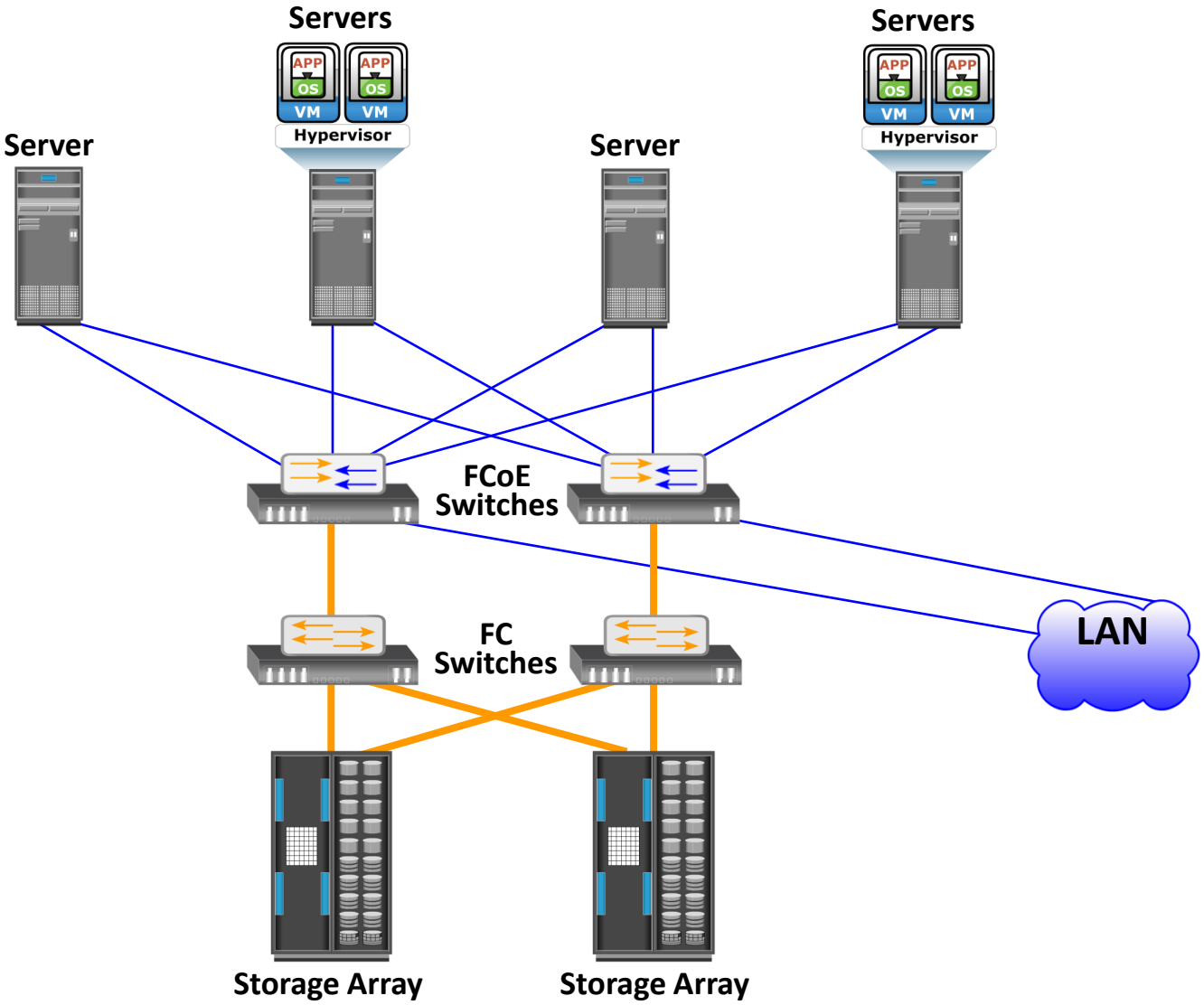
Data Center Infrastructure – Before Using FCoE

Safe to ignore! Just included for those curious about history.

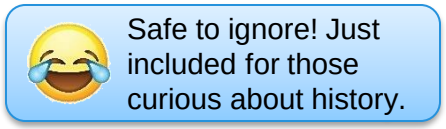


Data Center Infrastructure – After Using FCoE

Safe to ignore! Just included for those curious about history.

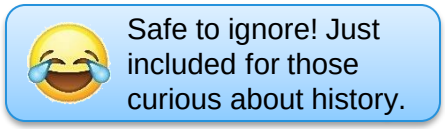


Components of an FCoE Network

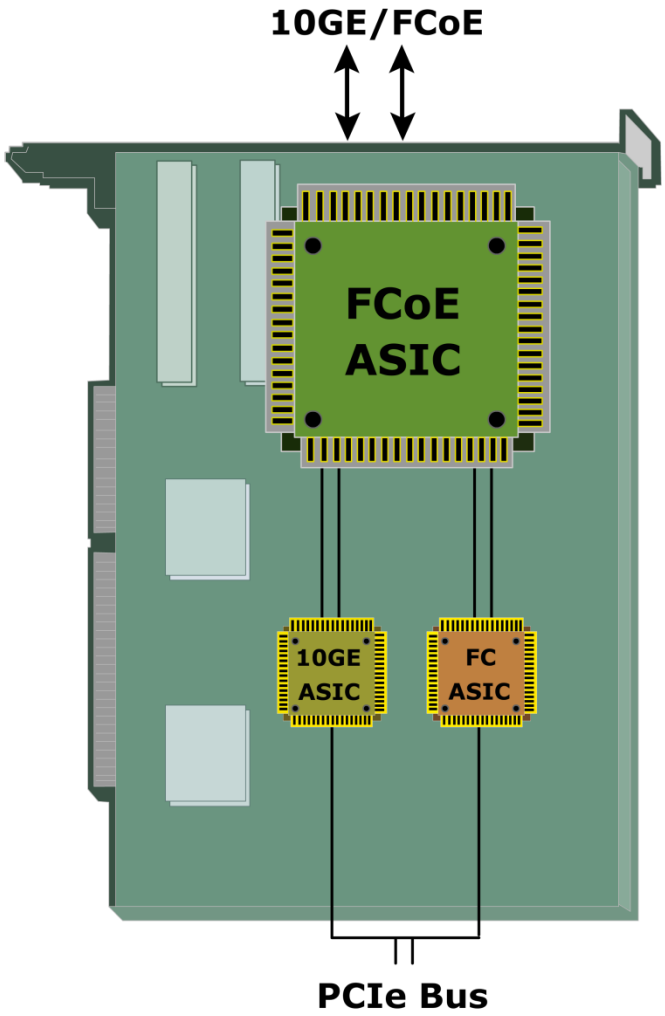


- Converged Network Adapter (CNA)
- Cable
- FCoE switch

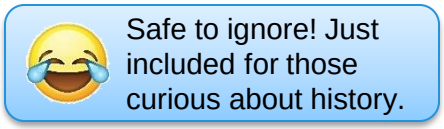
Converged Network Adapter (CNA)



- Provides functionality of both – a standard NIC and an FC HBA
  - Eliminates the need to deploy separate adapters and cables for FC and Ethernet communications
- Contains separate modules for 10 Gigabit Ethernet, FC, and FCoE ASICs
  - FCoE ASIC encapsulates FC frames into Ethernet frames



Cable



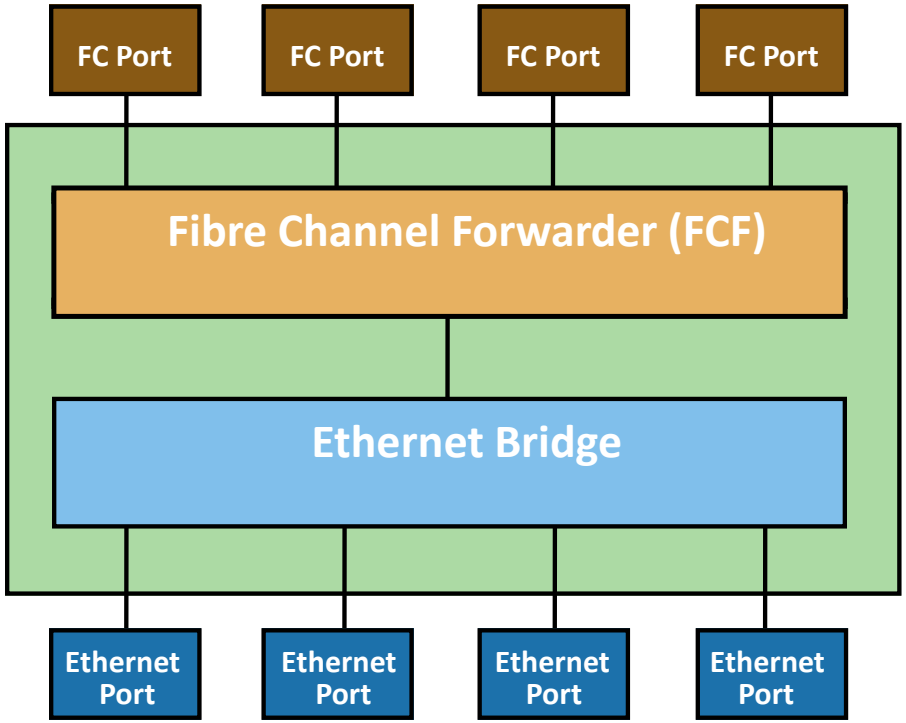
- Two options are available for FCoE cabling
  - Copper based Twinax cable
  - Standard fiber optical cable

Twinax Cable	Fiber Optical Cable
Suitable for shorter distances (up to 10 meters)	Can run over longer distances
Requires less power and are less expensive than fiber optical cable	Relatively more expensive than Twinax cables
Uses Small Form Factor Pluggable Plus (SFP+) connector	Uses Small Form Factor Pluggable Plus (SFP+) connector

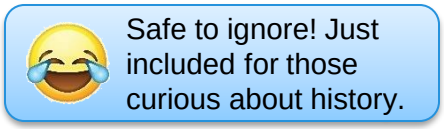
FCoE Switch

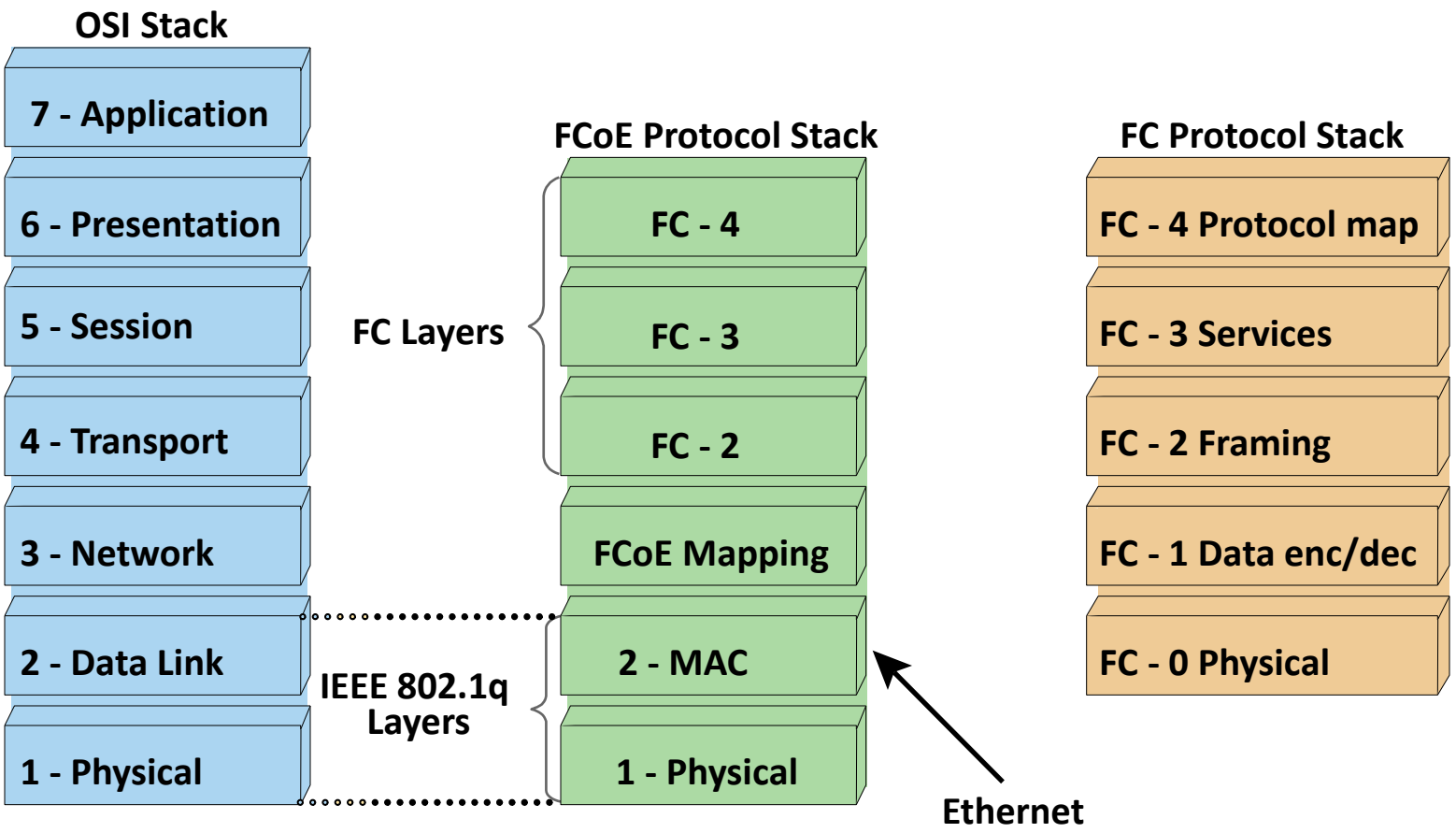
Safe to ignore! Just included for those curious about history.

- Provides both Ethernet and FC switch functionalities
- Consists of FCF, Ethernet bridge, and set of CEE ports and FC ports (optional)
  - FCF encapsulates and de-encapsulates FC frames
- Forwards frames based on Ethertype

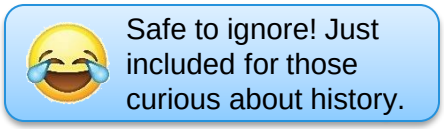


FCoE Frame Mapping





Converged Enhanced Ethernet



- Provides lossless Ethernet
- Lossless Ethernet requires following functionalities:
  - Priority-based flow control (PFC)
  - Enhanced transmission selection (ETS)
  - Congestion notification (CN)
  - Data center bridging exchange protocol (DCBX)

Reminder

- "Lossless Ethernet" is a contradiction
- Ethernet was fundamentally designed around packet dropping
- Higher-level protocols (e.g. TCP) interpret packet loss as a mechanism for flow control