# Fibre Channel SAN
# основы

Сергей Целиков
Системный инженер SAN,
+7 916 860-2579, sergey.tselikov@broadcom.com
Июнь 2020

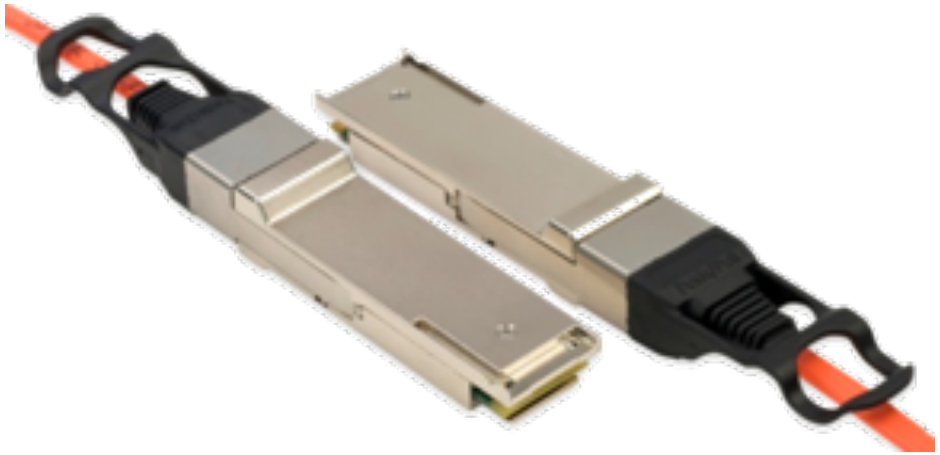
BROADCOM®

# Программа

- FC Fundamentals.
- Introduction to Fibre Channel Layers.
- Topologies, Terminology, and Addressing.
- Fibre Channel Theory.
- SAN Topologies and Topology Consideration.
- Availability.
- Zoning.
- FC Routing. Data Flow in FC Fabric.
- ISL Trunking.
- FC-to-FC routing.
- Virtual Fabrics.
- Access Gateway.
- FICON.
- Extension Solutions.
- Licensing.
- SAN Design Best Practices.

# Optics Overview

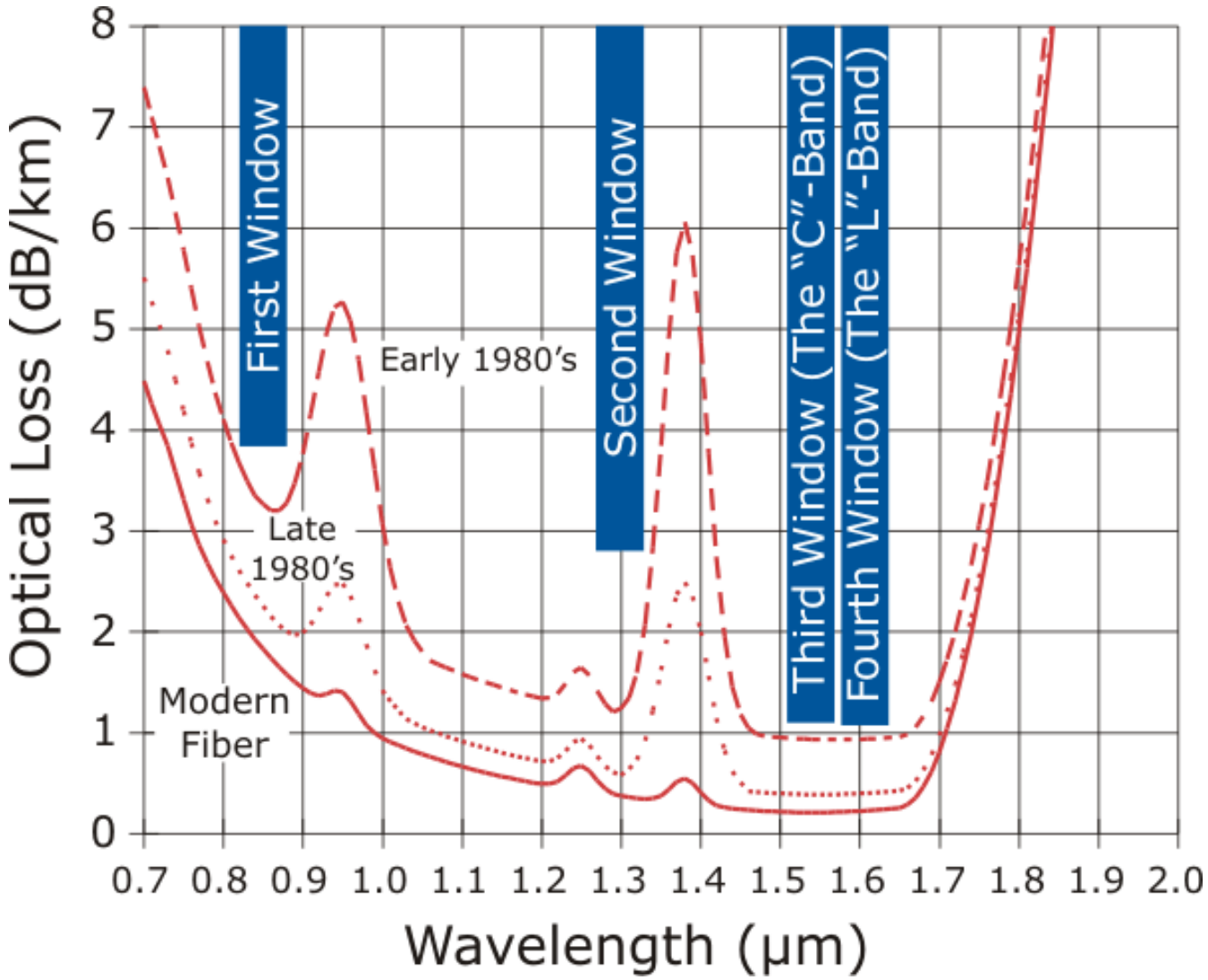
- Transceivers are used to transmit data over fiber or copper cabling
- Brocade switches and HBAs require Brocade-branded optics
- Most Fibre Channel transceivers are tri-mode:
  - 32 Gbps SFP+ supports 32, 16, and 8 Gbps speeds
  - 16 Gbps SFP+ supports 16, 8, and 4 Gbps speeds
  - 8 Gbps SFP+ supports 8, 4 and 2 Gbps speeds
  - 4 Gbps SFPs support 4, 2 and 1 Gbps speeds
- Exceptions are the 10 Gbps SFP+ and 4x32 Gbps QSFP which only synch at their respective speeds






BROADCOM

# Multi-Mode (MMF) and Single Mode Fiber (SMF)



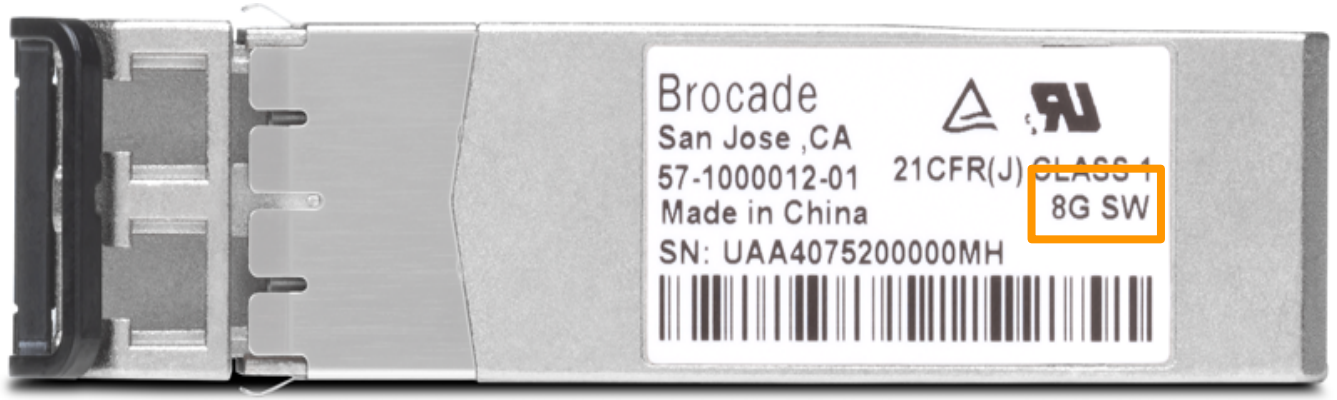




BROADCOM

# Optics Overview (cont.)

- Optics are divided into several categories depending on their transmission frequency and maximum distance:
  - Short Wave Length (SWL) (850 nm)
    - Uses multi-mode cable types (OM1, OM2, OM3, and OM4)
  - Long Wave Length (LWL) (1310 nm)
    - Uses single-mode cable types
  - Extended Long Wave Length (ELWL) (1310 nm)
    - Output power is boosted to extend over longer distances
    - Uses single-mode cable types
- Maximum transmission length is highly dependant on SFP, cable type, and speed



# Optics and Cable Comparison Chart

| Transceiver Type | Supported Speeds | Cable Type Used | Distance at Max Speed[1] |
|---|---|---|---|
| 4 Gbps SWL SFP | 4, 2, 1 Gbps | OM1/OM2/OM3 | 380 m |
| 4 Gbps LWL SFP | | Single-mode | 10 km |
| 4 Gbps ELWL SFP | | | 30 km |
| 8 Gbps SWL SFP | 8, 4, 2 Gbps | OM1/OM2/OM3 | 150 m |
| 8 Gbps LWL SFP | | Single-mode | 10 km |
| 8 Gbps ELWL SFP | | | 25 km |
| 10 Gbps SWL SFP+ | 10 Gbps | OM1/OM2/OM3 | 300 m |
| 10 Gbps LWL SFP+ | | Single-mode | 10 km |
| 16 Gbps SWL SFP+ | 16 Gbps | OM1/OM2/OM3 | 100 m |
| 16 Gbps LWL SFP+ | | Single-mode | 10 km |
| 4x16 Gbps SWL QSFP | | MPO 1x12 ribbon | 100 m |

# Optics Supported on Gen 6 Switches and Directors

| 32G | 16G | 10G FC | 4x16G QSFP | 4x32G QSFP |
|---|---|---|---|---|
| SWL, LWL | SWL, LWL, ELWL | SWL, LWL | SWL(100M), 2KM | SWL(100M), 2KM |

## 32GFC SWL Link distance

| 32GFC SWL SFP+ Speed | Distance (OM3 MMF) | Distance (OM4 MMF) |
|---|---|---|
| @ 32GFC | 70m | 100m |
| @ 16GFC | 100m | 125m |
| @ 8GFC | 150m | 190m |

## 128GFC SWL Link distance

| 128G SWL QSFP28 Speed | Distance (OM3 MMF) | Distance (OM4 MMF) |
|---|---|---|
| @ 4x32GFC | Target 70m (1.0dB connector loss) | Target 100m (1.0dB connector loss) |
| @ 4x16GFC | Target 100m | Target 125m |

Uses MTP/MPO 1x12 ribbon multimode fiber (OM3/OM4)

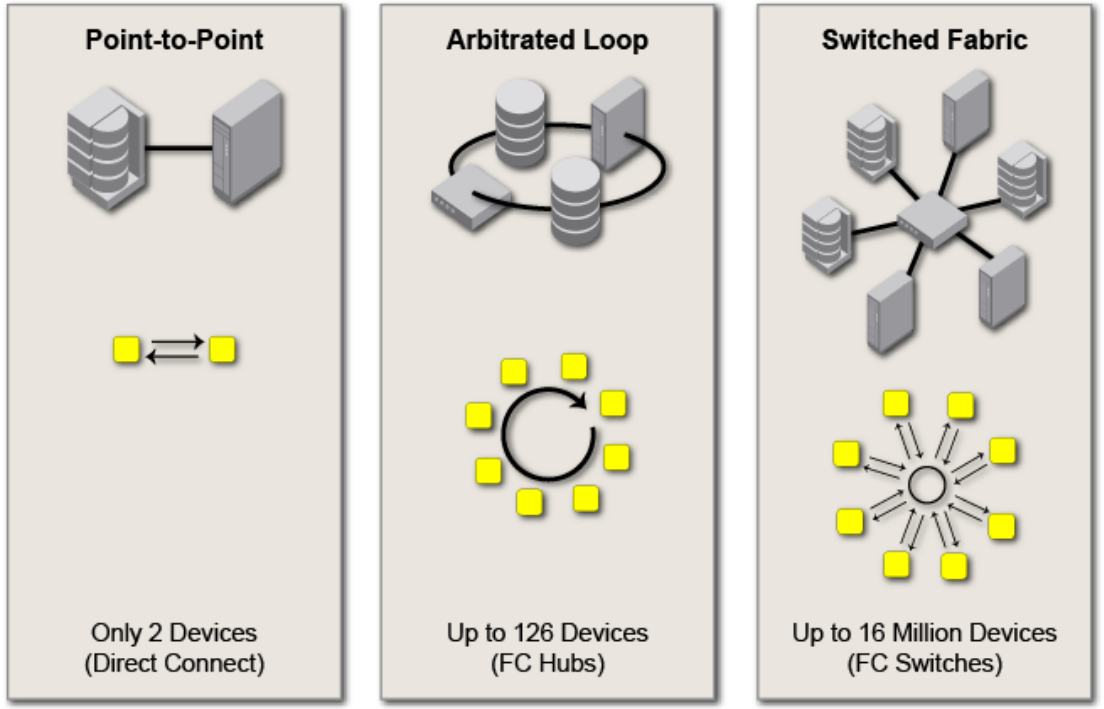



BROADCOM

# Fibre Channel Topologies, Terminology, and Addressing

# Fibre Channel Topologies

- Three kinds of Fibre Channel (FC) Topologies:
  - Point-to-Point (Pt to Pt) – Allows two devices to talk
  - Arbitrated Loop – Allows 126 devices to talk, Arbitrated Loop Physical Address (AL_PA) “00”, is reserved for the Fabric Loop Port (FL_Port)
  - Switched Fabric – Supports up to 16 Million devices and is the most commonly used topology







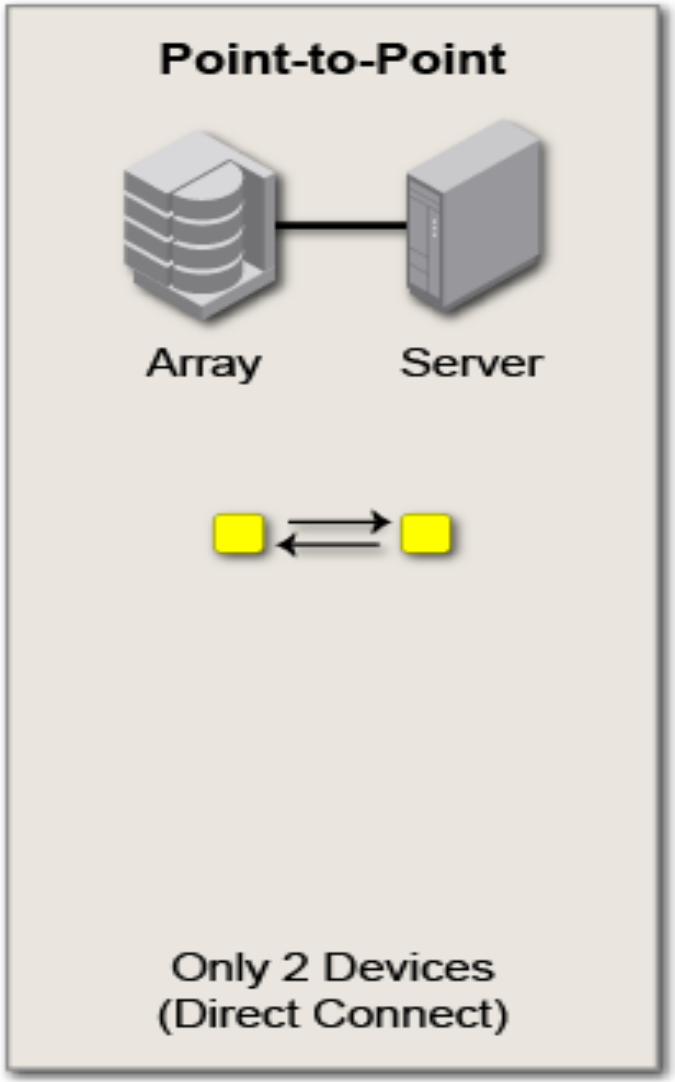
BROADCOM

# Point-to-Point

**Increase SCSI Distance**

**A storage array to a host**

**Two devices connected together**





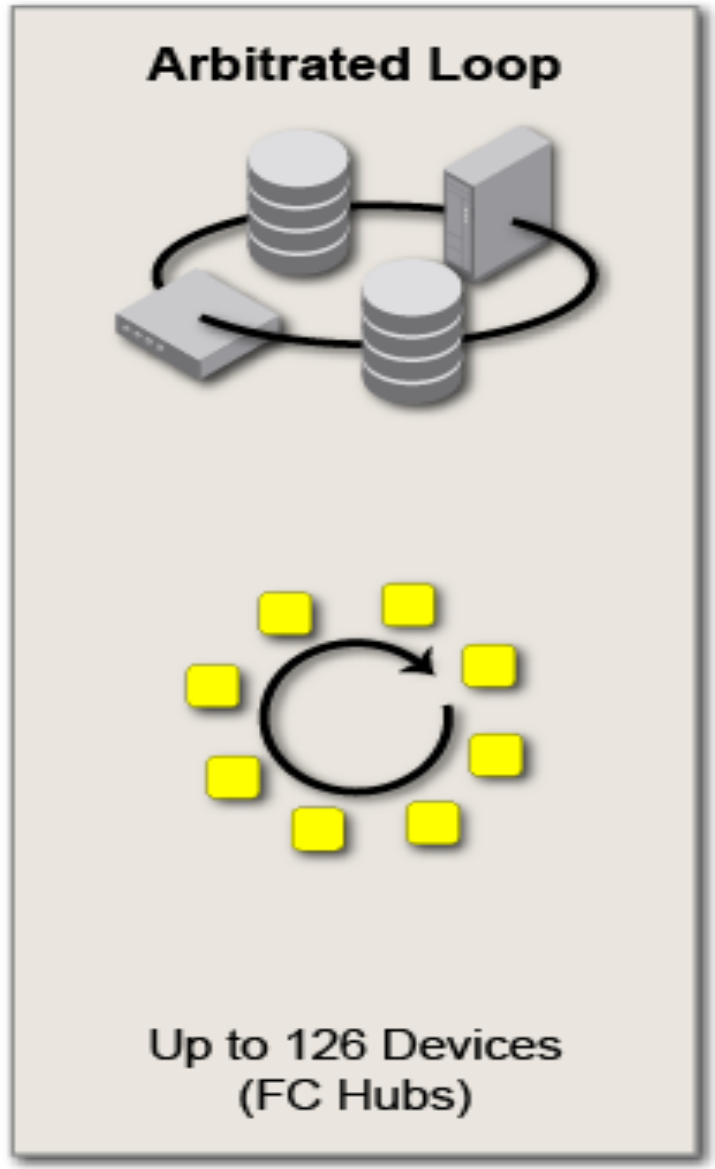


BROADCOM

# Fibre Channel Arbitrated Loop (FCAL)

- Theoretically up to 126 devices on a shared media  for small systems at reduced cost and reduced performance level

- Requires a port to successfully arbitrate prior to establishing a circuit to send and/or receive frames



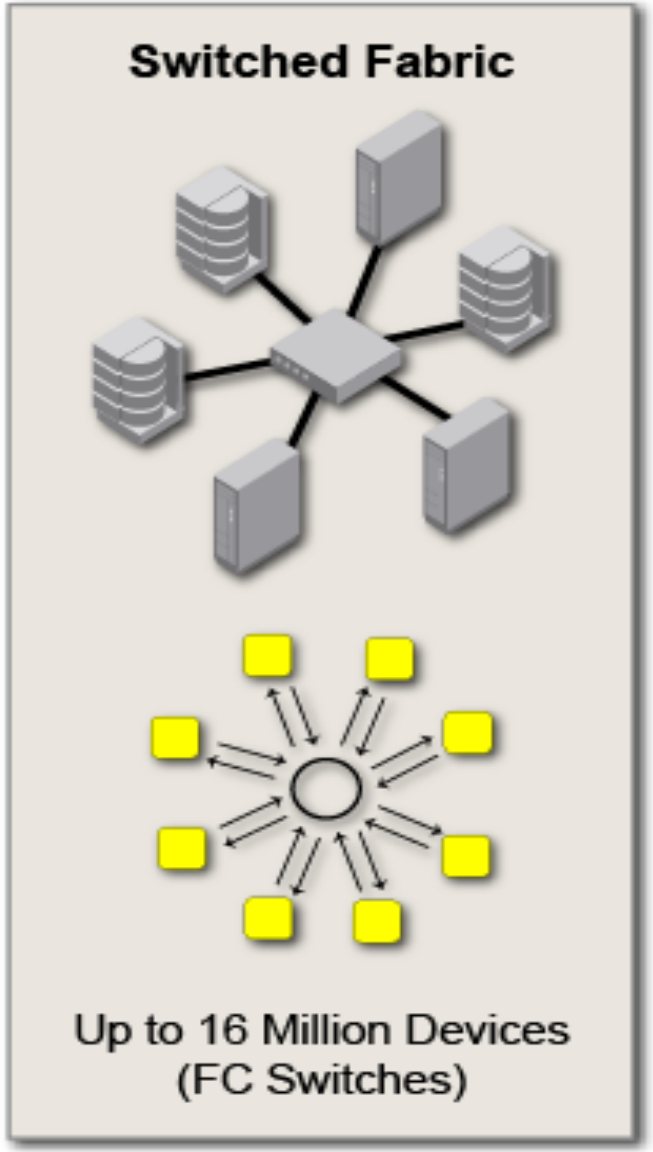





BROADCOM

# Switched Fabric

- Switch Fabric Topology:
  - Highest performance level
  - High scalability
  - Good fault isolation
  - Embedded management and services
  - Up to 239 unique domains (switches) with:
    - Unique switch names
    - Unique IP addresses



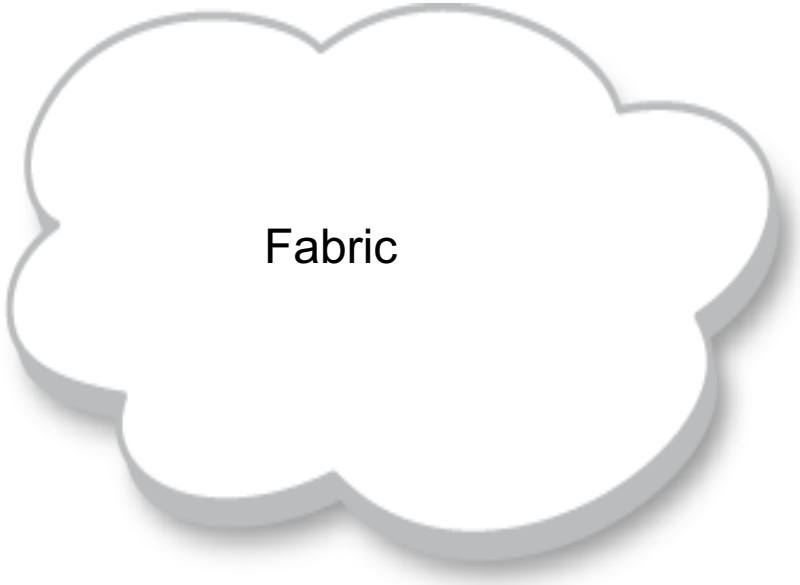





BROADCOM

# Switched Fabric (cont.)

- A **Fabric** is a connection of Fibre Channel switches and/or devices capable of routing frames using only a destination identifier (`D_ID`)
  – A Fabric is commonly pictured as a cloud



# Fibre Channel Addressing

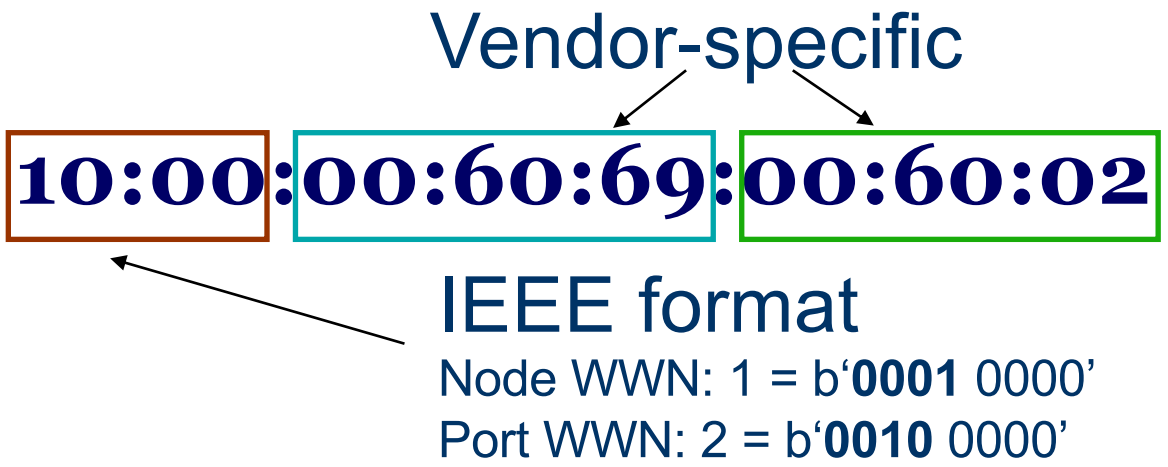
## WWN's and Port ID's

- FC has two types of addressing at Layer 2
  - Fixed – World Wide Name Burned in at factory
  - Dynamic – Port ID. A layer 2, 24-bit address that is assigned at fabric login.
- Note: There is no concept of ISO layer 3 or Layer 4 in FC.
  - BTW - "FC routing" is equivalent to L2 NAT, not a true ISO L3 protocol.

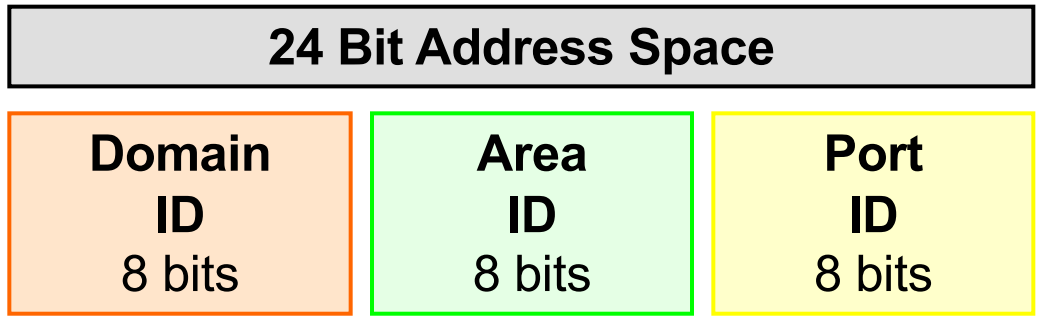
<table>
<tr><th>OSI Model</th><th>Ethernet & TCP/IP</th><th>Fibre Channel</th></tr>
<tr><td>Application</td><td rowspan="3">Application Layers (POP3, SMTP, DNS, DHCP,FTP, WWW protocols)</td><td>Upper Layer Protocols (ULP)<br>[FCP=SCSI] [FICON] [NVMe]</td></tr>
<tr><td>Presentation</td><td rowspan="2">FC-4: ULP Mapping</td></tr>
<tr><td>Session</td></tr>
<tr><td>Transport</td><td>TCP / UDP</td><td rowspan="3">Not Applicable</td></tr>
<tr><td>Network</td><td>Dynamic IP Address<br>10.77.77.77</td></tr>
<tr><td>Data Link</td><td rowspan="2">Fixed MAC Address<br>x'00-00-0E-21-17-6B'</td></tr>
<tr><td></td><td>FC-3: Common Services</td></tr>
<tr><td></td><td></td><td>Dynamic Native Address (8/24-bit)<br>Fixed World-Wide Name (64-bit)</td></tr>
<tr><td></td><td></td><td>FC-1: 8b/10b or 64b/66b Encoding</td></tr>
<tr><td>Physical</td><td>Physical Interface</td><td>FC-0: Physical Interface</td></tr>
</table>

# Fixed & Dynamic Address Formatting

## WWN's and Port ID's (PID's)




Every fabric device (HBA, switch, director, storage device) has one or more 64-bit WWN addresses.
Uses an IEEE-assigned addressing scheme.




Dynamic address (24-bit)
Assigned dynamically when logging into the Fibre Channel network

24-bit = 16 million fabric addresses

N_Port/F_Port usable range:

**x'010000'** to **x'EFEFFF'**




BROADCOM

# SAN Basics

## LAN / SAN Comparison

### LAN

- Nodes use Network Interface Cards (NIC) with 48-bit MAC Addresses.

### SAN

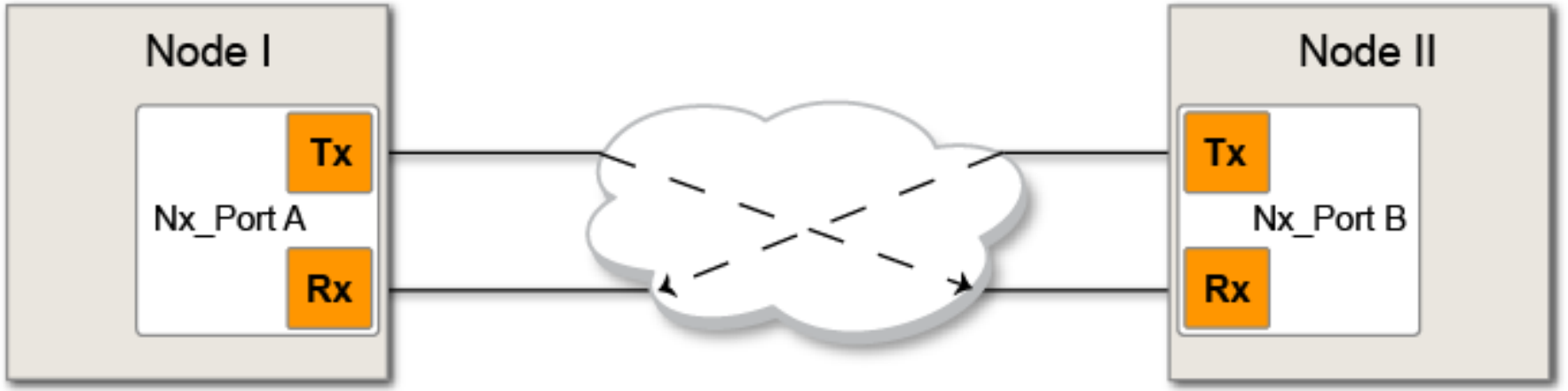
- Nodes use Host Bus Adapters (HBA) with 64-bit World Wide Names (WWN).




BROADCOM

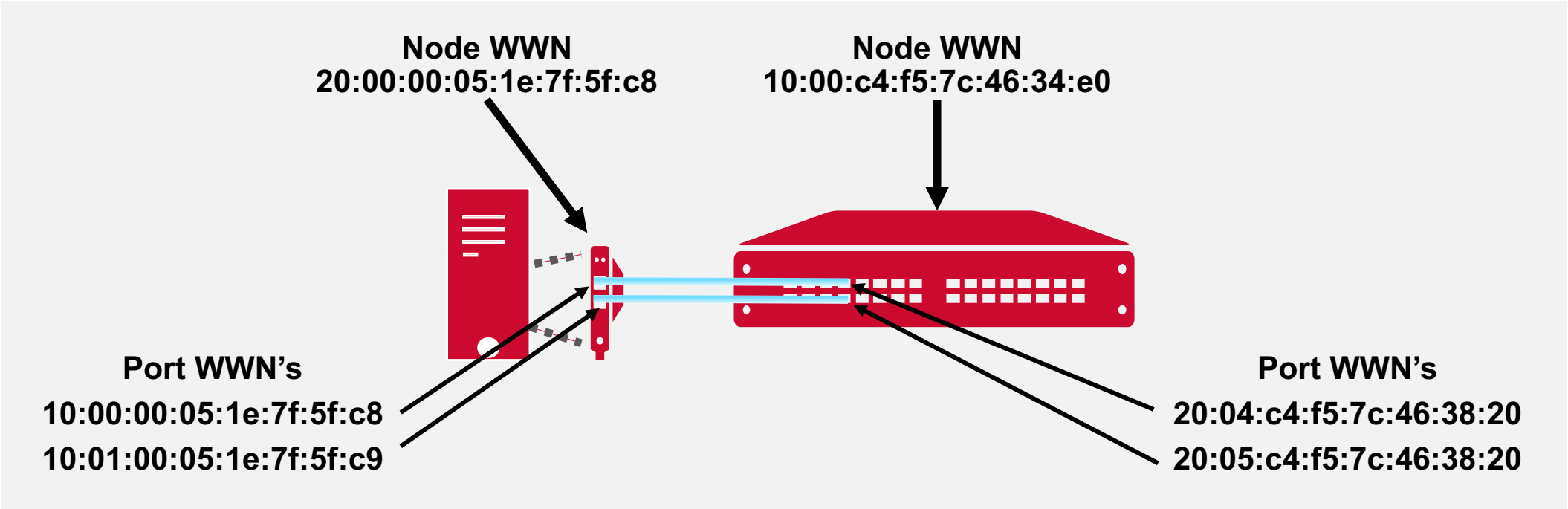
# Switched Fabric (cont.)

- An **FC Port** is a single physical connection that provides separate transmit (`tx`) and receive (`rx`) functions
  - `Tx` encodes and transforms data to serial format
  - `Rx` recovers clock from serial data received, decodes and de-serializes the data
- An **FC Node** is an FC device that transmits and receives information via one or more ports
- **Each Node** has a unique 64-bit address called “Node World Wide Name”. The format of this 64-bit identifier along with the format for the port on this nodes 64-bit identifier are specified by IEEE.
- **Each Port** also has a unique 64-bit address called “Port World Wide Name”.



# World Wide Name

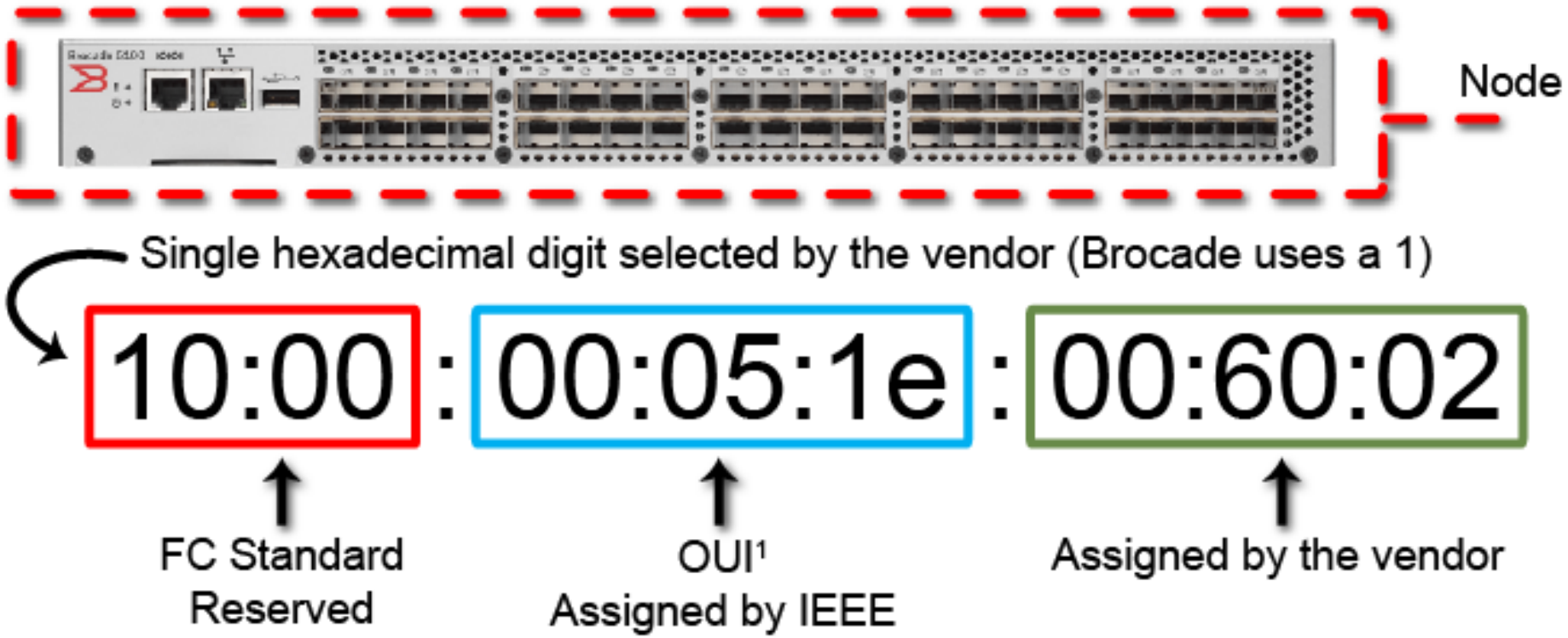
- A World Wide Name (WWN) is a globally unique address used to identify each Fibre Channel node and node port
  - A node is a switch, HBA or storage controller
    - Note – A storage can have one or more controllers and a server can have one or more HBAs.



BROADCOM

# Device Addressing: FC Node WWN (nWWN)

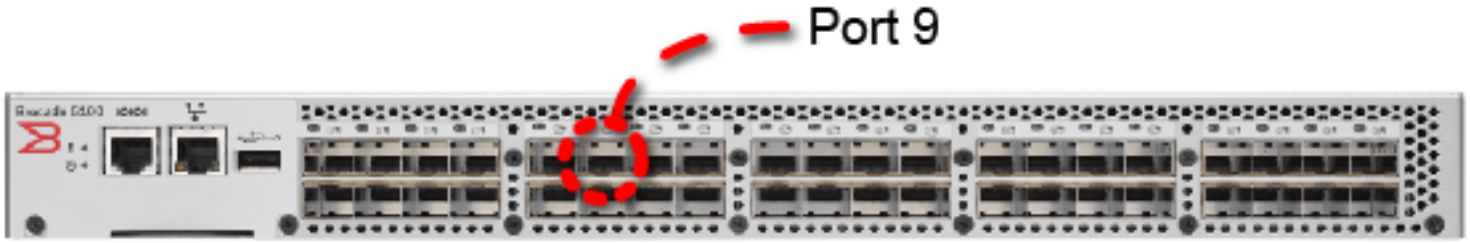



- FC node WWNs (nWWN) follow the format shown above
  - Applies to switches, storage, and HBAs



BROADCOM

# Device Addressing: FC Port WWN (pWWN)




Single hexadecimal digit selected by the vendor (Brocade uses a 2)

20:09 : 00:05:1e : 00:60:02

Set by the vendor to uniquely identify a port on a device or switch

OUI Assigned by IEEE

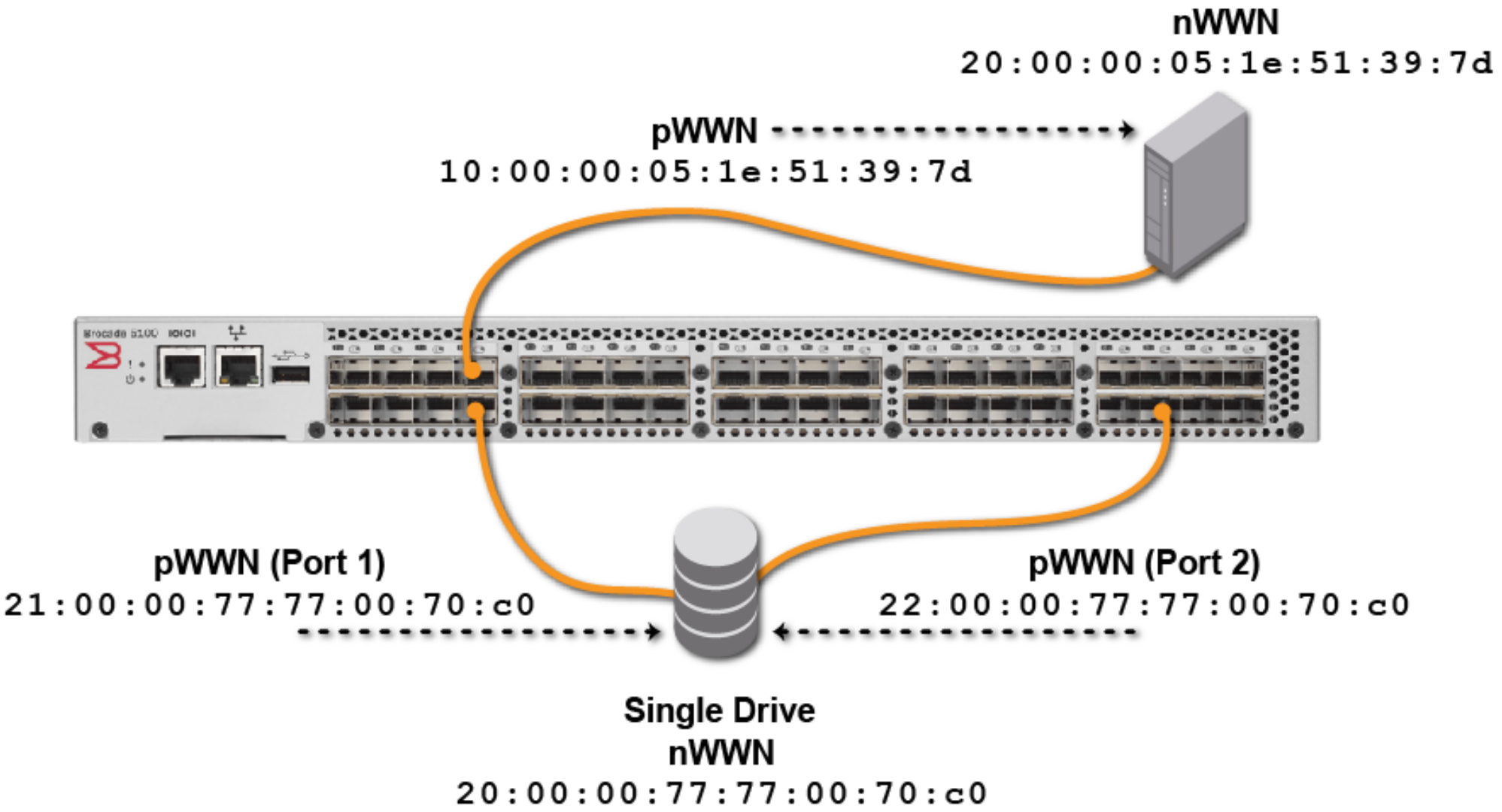
Assigned by the vendor

- FC port WWNs (pWWN) follow the format shown above
  - Applies to switches, storage, and HBAs
  - With NPIV, a physical server Host Bus Adapter (HBA) can provide up to 255 unique pWWN for use by virtual servers




BROADCOM

# Device Addressing: nWWN and pWWN Example







BROADCOM

# Device Addressing: FC Addresses

- Fabric addresses are 24-bits (3 bytes long)
- A device's Fabric address indicates:
  - The switch and port index to which the device is connected
- Fabric addresses are represented in hexadecimal format (`0x`) which often appear before the address

# SAN Basics

## LAN / SAN Comparison

### LAN

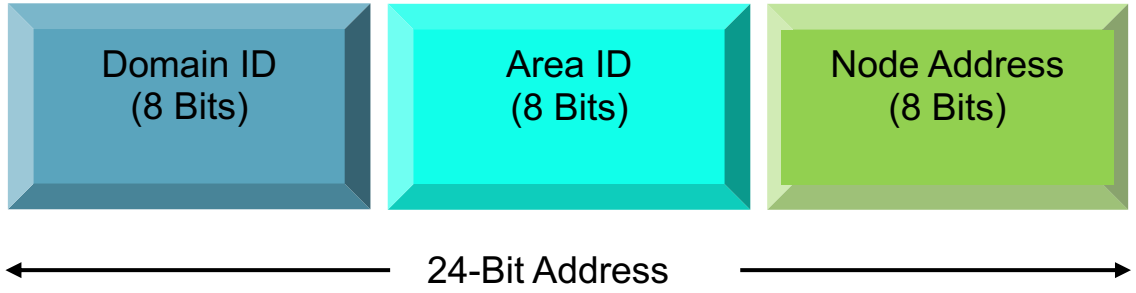
- Nodes use Dynamic IP address which is 32-bit (4 bytes) long.
- Address assignment is manual or requires special external service (DHCP)

### SAN

- Nodes use Dynamic Fabric address which is 24-bit (3 bytes) long.
- Automatic address assignment within fabric.

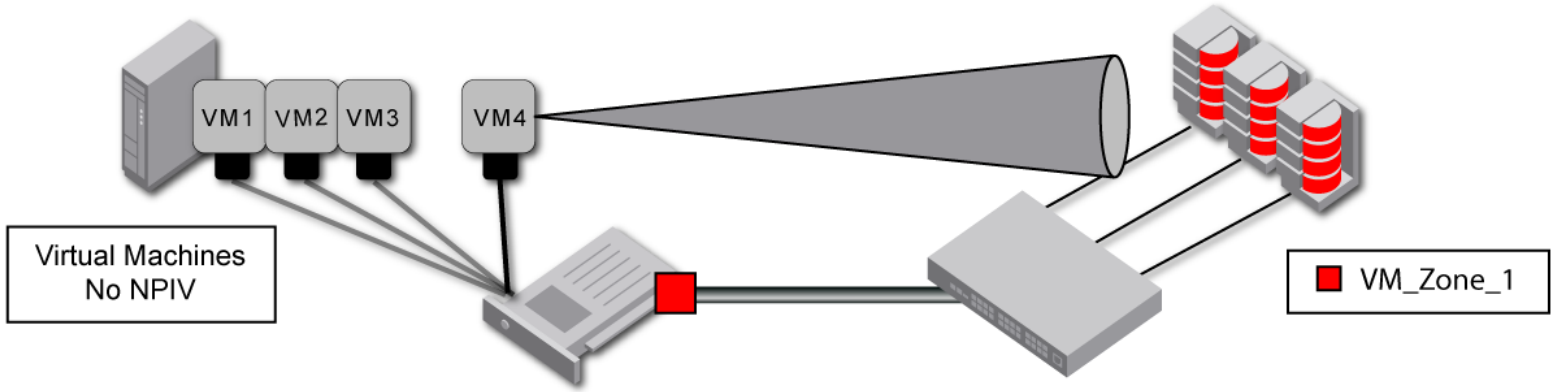
# Fibre Channel Network Addressing

- Each switch is responsible for assigning unique addresses
- In Fabric OS v6.0 and later, all FC addresses follow the *Core PID format*
  - Addresses are 24 bits:
    - Domain ID (8 bits) 0x01 - 0xEF
    - Area ID (8 bits) 0x00 - 0xFF
    - Node Address (8 bits) 0x00 - 0xFF
  - Address types:
    - Fabric DD AA XX
    - NPIV DD AA PP
    - Shared Area DD AA 00/40/80/C0



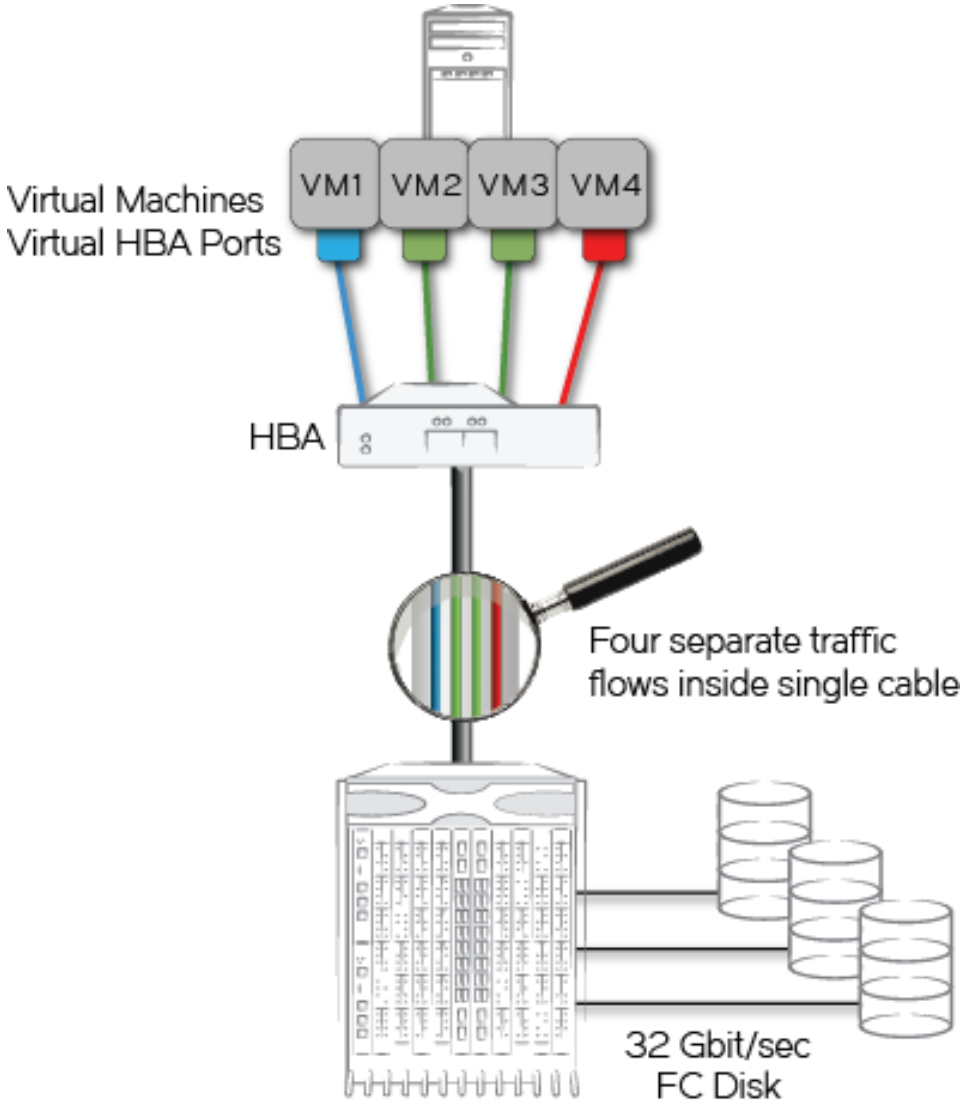
# N_Port ID Virtualization (NPIV)

- Virtual servers require secure access to storage the same way as physical servers
- Without NPIV, a single physical server connection is unable to provide independent storage access to individual virtual servers
  - Instead, all storage ports and Logical Unit Numbers (LUNs) are exposed to all virtual machines, reducing security and manageability
- NPIV is an ANSI standard designed to solve this problem



# Device Addressing: NPIV

- With NPIV, a physical server Host Bus Adapter (HBA) can provide up to 255 unique Port World Wide Names for use by virtual servers
- Fabric switches with NPIV support can then assign unique fabric IDs to each virtual server as they log in to the fabric
- With NPIV support, standard fabric zoning and storage LUN masking can be used with virtual machines to isolate storage ports and LUNs to the appropriate virtual server just as they are with physical servers



# NPIV Overview

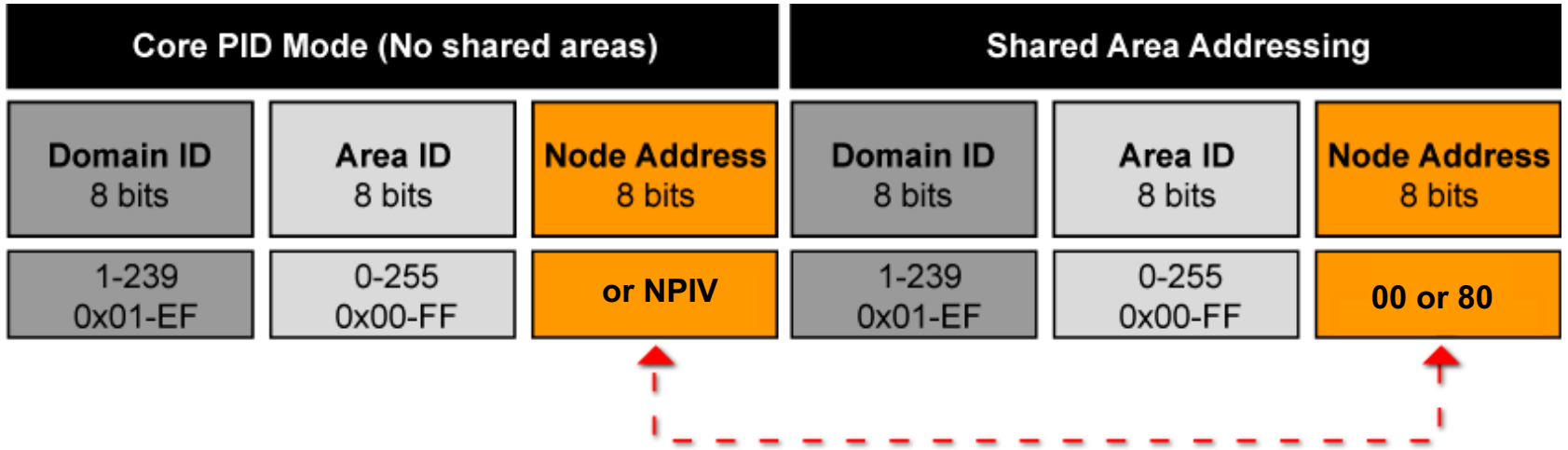
- Enabled by default on a per port basis
- Each NPIV device is assigned a unique device PID
- The end devices attached to the port provide unique WWNs for the following:
  - Port WWN
  - Node WWN
- To the fabric, the NPIV device acts the same as all other physical devices in the fabric
- NPIV is defined in the FC-LS T11 standard

# NPIV Scalability

- Each NPIV-enabled port on the switch can support up to 255 devices
  - The sum of all configured per-port login maximum values cannot exceed the total number of logins set for the switch with the `configure` command
- Shared area ports on high-density blades have varying limits
- Default value in Fabric OS is 126. To support more devices, the value must be changed using the `portcfgnpivport --setloginlimit` command

# 10 Bit Shared Area Addressing

- A shared area is an area ID that exists more than once in a single domain
  - These shared areas are differentiated by their node addresses
  - This allows for more than 256 ports in a single domain
- Director blades with more then 32 ports use shared areas
- Shared area PIDs use a node address of either 0x00, 0x40, 0x80, or 0xC0
  - Example of two shared areas on an FC32-48 blade in slot 1:

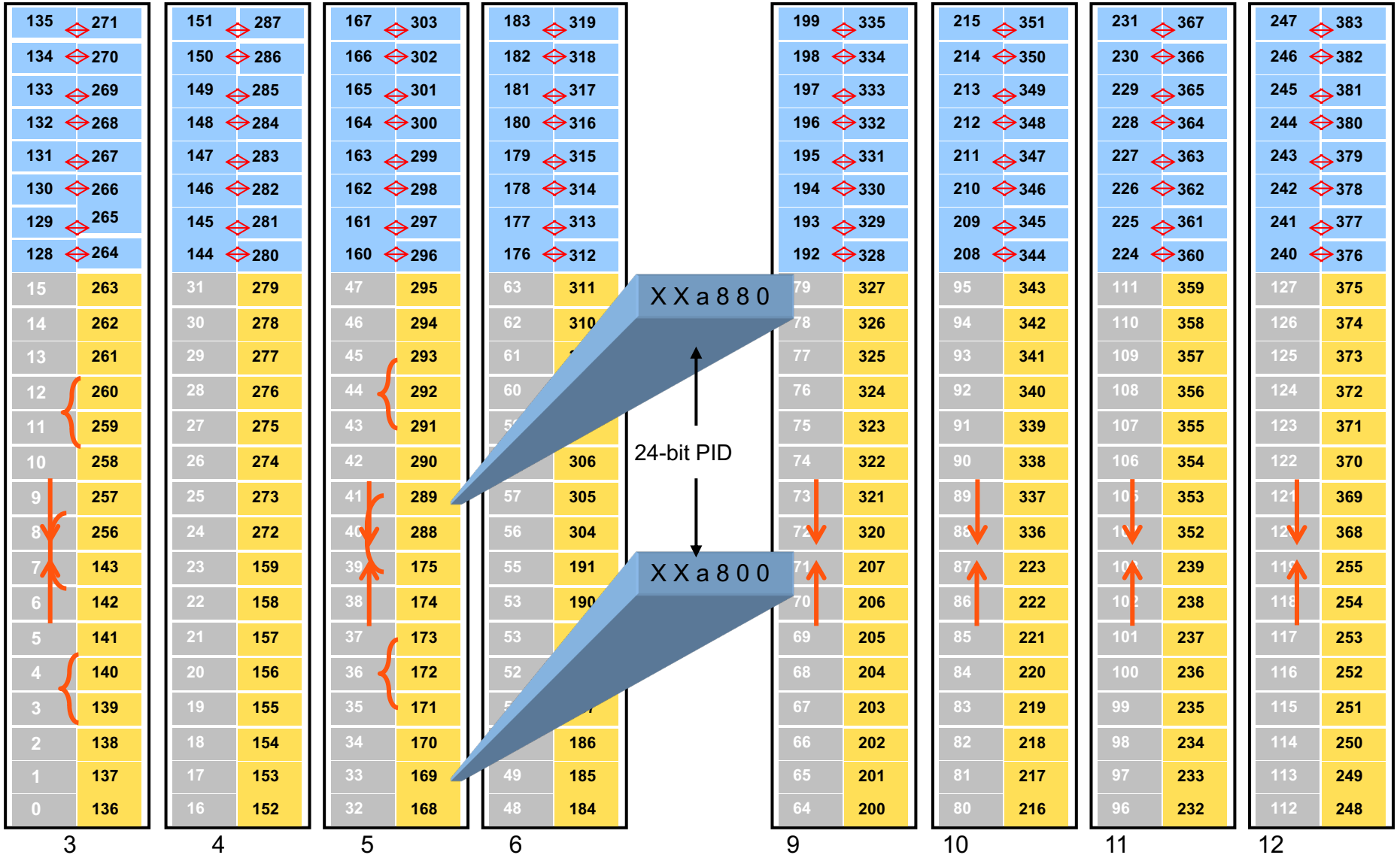






BROCADE
A Broadcom Limited Company

# Shared Area Addressing (cont.)







BROCADE
A Broadcom Limited Company

# Fibre Channel Theory

# SAN Basics

## LAN / SAN Comparison

### LAN

- For the most part, LAN's are painstakingly, manually configured…port by port.

### SAN

- In a Brocade SAN, the network ports configure themselves. It is plug and play.




BROADCOM

# Port Types

- Device ports (Nx_Ports):
  - N_Port – Node port, a fabric device directly attached
  - NL_Port – Node loop port, a device attached to a loop
- Switch ports:
  - U_Port – Universal port, a port waiting to become another port type
  - FL_Port – Fabric loop port, a port to which a loop attaches
  - G_Port – Generic port, a port waiting to be an F_Port or E_Port
  - F_Port – Fabric port, a port to which an N_Port attaches
  - E_Port – Expansion port, a port used for inter-switch links (ISLs)

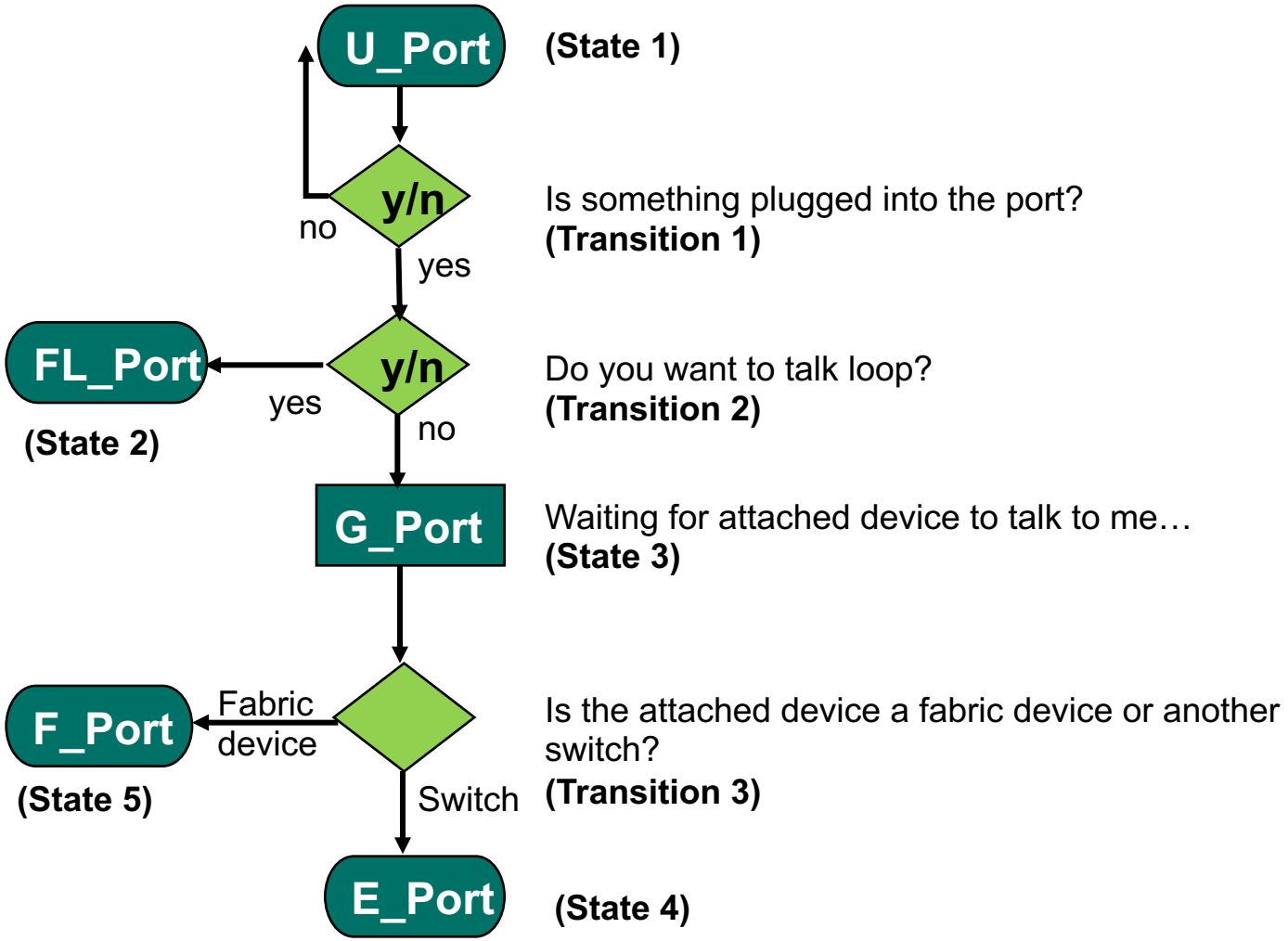
 

# Port Types (cont.)

- Configured ports:
  - EX_Port – A type of E_Port used to connect an FC router to a fabric
  - VE_Port – Virtual E_Port (used in FCIP fabrics)
  - VEX_Port – VEX_Ports are no different from EX_Ports, except underlying transport is IP rather than FC
  - D_Port – A configured port used to perform diagnostic tests on a link with another D_Port
  - M_Port – A configured port used by the Flow Mirror feature; frames sent between a pair of devise are mirrored to the M_Port
  - AE_Port – A type of E_Port used to connect an AMP to a fabric
  - AF_Port – A type of F_Port configured on an AMP used to receive monitored traffic flows

# Port Initialization Process (old)

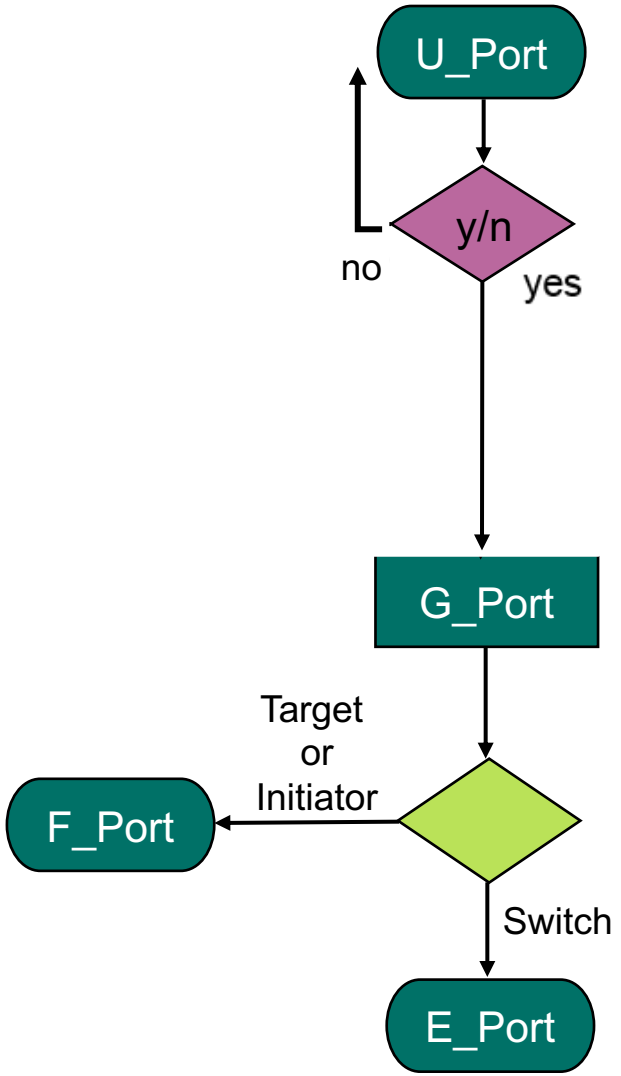






BROADCOM

# Port Initialization Process (modern)




Is something plugged into the port?

Waiting for attached device to talk to me…

Is the attached device a fabric device or another switch?

BROADCOM

# Port State Machine Processes
## Link Control Protocols – Port State Machine

- The Port State Machine (PSM) is used to manage the port connections once port speed has been negotiated. There are four states provided by the PSM:

1. Active State (AC) where the port is able to transmit and receive frames.
   - This is the expected state when an active device is attached
2. Link Reset State (LR): A primitive sequence used during link initialization between ports in fabric topology.
3. Offline State (OLS): Used during link initialization between ports in a fabric. It is sent to indicate that the transmitting port is attempting to initialize a link or is going offline.
4. Link Failure State (LF): User during link initialization between ports in a fabric. It is sent to indicate a broken link or inaccessible device.

# Port State Machine Processes
## Link Control Protocols – Port State Machine – (cont.)

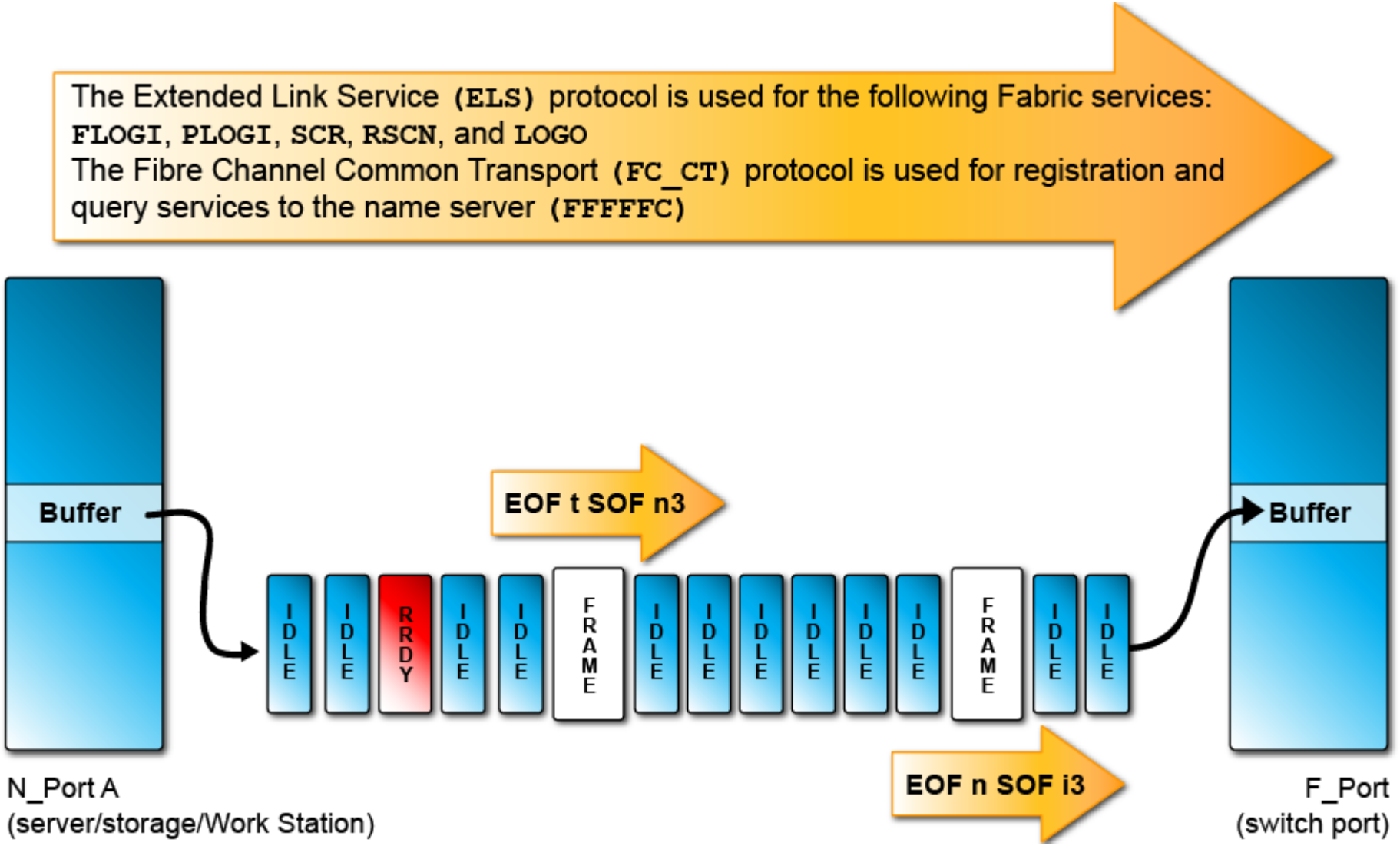
- The chain of events /flow → U_Port starting from power on:

1. NOS (Non Operational)
2. OLS (Off Line)
3. LR (Link Reset)
4. LRR (Link Reset Response)
5. AC (Active State)
6. IDLE (There are at least two idles between any two ordered sets)
7. IDLE

- The flow → switch portDisable to portEnable:

1. OLS (Off Line)
2. LR (Link Reset)
3. LRR(Link Reset Response)
4. AC (Active State)

# FC Communication Overview
## After Speed Negotiation and PSM ACTIVE



# SAN Basics

## LAN / SAN Comparison

### LAN

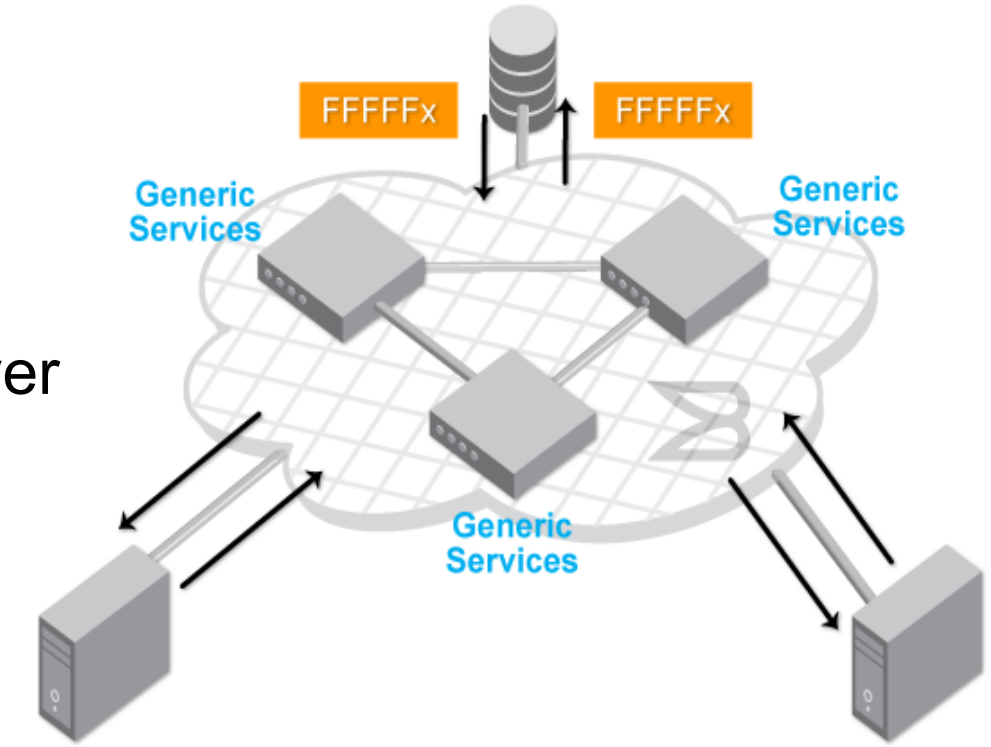
- Nodes communicate with nodes. They are unaware of switching infrastructure.

### SAN

- Nodes are intrinsically aware of the network infrastructure. They have conversations with the network; they rely on the network for device discovery, change notification, etc.

# Fabric Generic Services

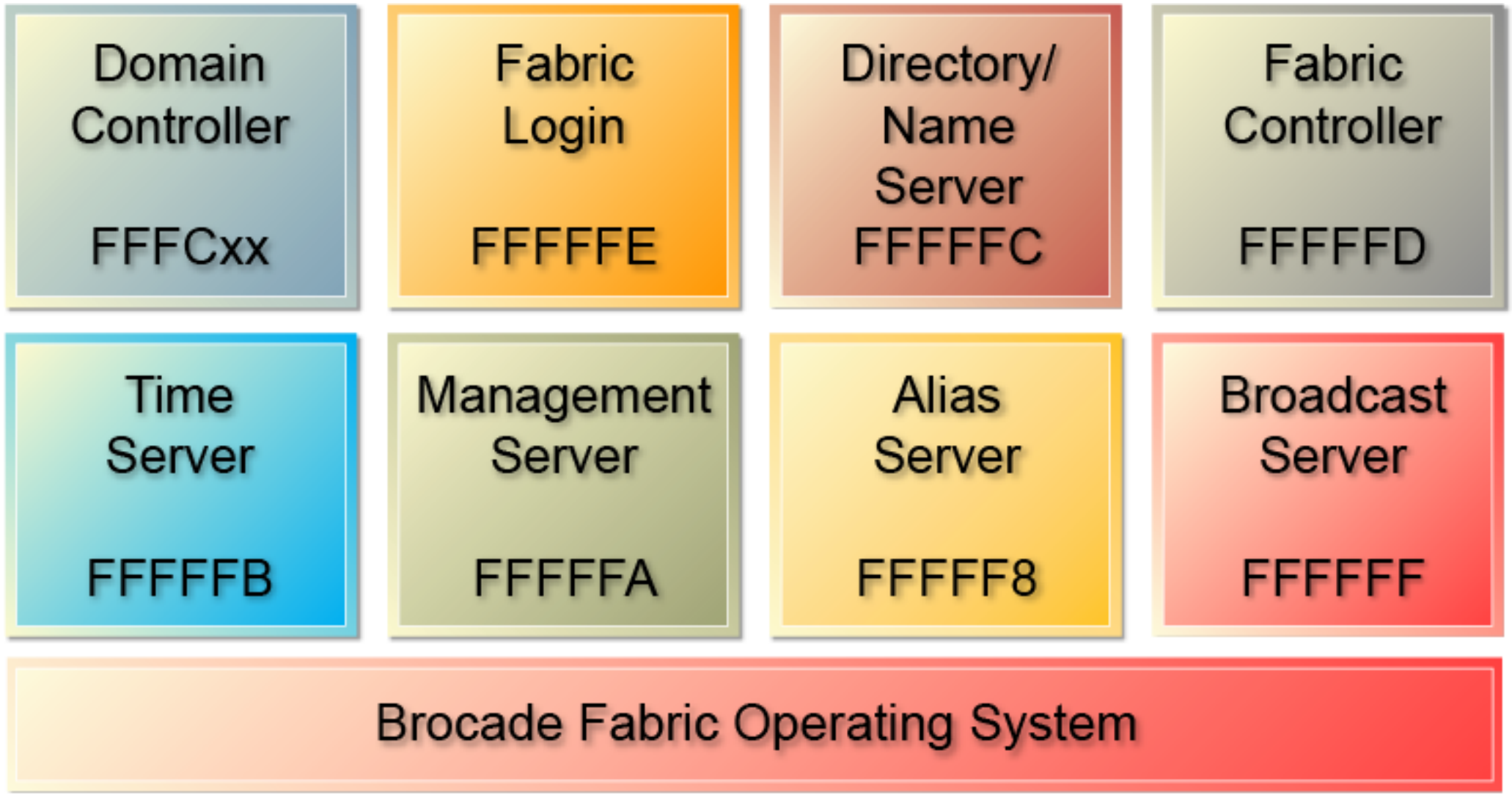
- All Fibre Channel fabrics provide fabric-wide generic services
- Each service is assigned a specific address referred to as its Well-Known Address (FFFFFx)
- Services used to manage a Fibre Channel network such as Fabric port logins, Name Server registration, etc.
- Usually only found in the switched fabric topology
- FC switches and devices communicate with these services via a set of well-known addresses



BROADCOM

# Well-Known Addresses of Fabric Services







BROADCOM

# Embedded Port Communication

- Embedded Port (Domain Controller) is responsible for communication between switches
  - The Fabric embedded port (domain controller) is assigned a Fabric address
  - Brocade uses `FFFCxx` where `xx` represents one of 239 possible domains
    - `FFFCxx` is used for `PLOGI` and `PRLI` to retrieve information to add to the name server data base
    - Embedded port probing (Fabric probing) is enabled by default thus allowing private targets that accept `PRLI` into Fabric

# SAN Basics
## LAN / SAN Comparison

### LAN

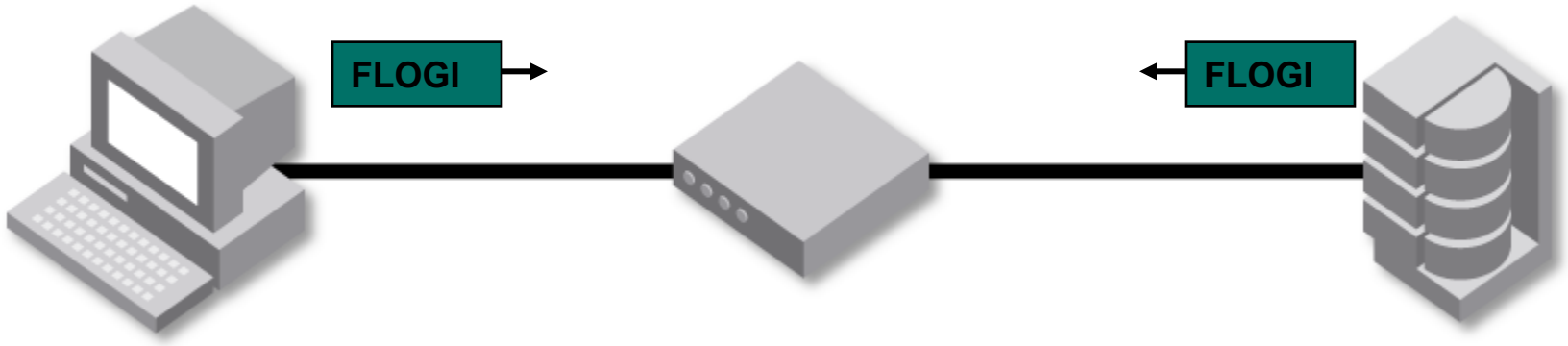
- A Layer 2 LAN uses broadcasts and must deal with unknown destination addresses.
- Must listen to traffic to learn MAC addresses

### SAN

- There are no broadcasts.
- Nodes must login to the fabric and register themselves before any traffic may flow, ergo – the SAN knows where everything is located.

# Fabric Login Server (0xFFFFFE)

- The **Fabric Login Server** manages the process of a FC node (device) joining the fabric
  - An FC node begins the process of joining a fabric by sending a fabric login (FLOGI) request to the Fabric Login Server
  - After the Fabric Login process completes, the Fabric Login Server provides the 24-bit FC address to the FC node

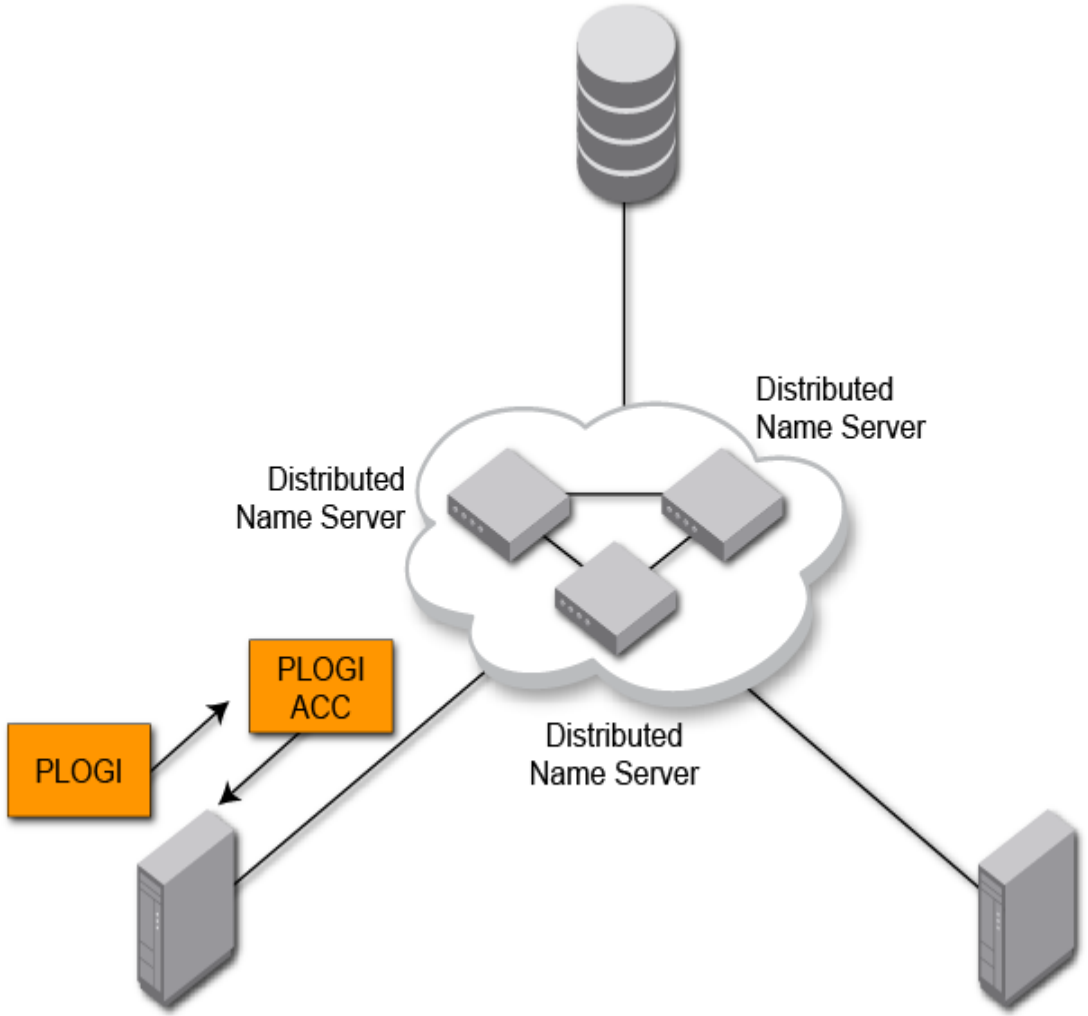


# Fabric Controller (0xFFFFFD)

- The **Fabric Controller** is responsible for managing the operation of the fabric, including all switch-to-switch traffic (Class F)
  - Fabric Initialization
  - Fabric Configuration
  - Generate links response
  - Start and stop connections
  - Frame Routing Management
  - Also manages the State Change Registration (SCR) process

# Name Server (0xFFFFFC)

- The **Name Server** provides a directory of all N/N_L Ports in the fabric
  - Each switch keeps a list of all locally-attached devices, updated automatically as devices connect/disconnect
  - The Name Server collects the local lists to form a global directory
  - Result: a distributed, transparent service that devices can use to discover other devices for upper layer protocol communication (PLOGI)



# Fabric Name Server versus TCP/IP DNS

- The Name Server can be compared with the TCP/IP DNS

| Functionality | Distributed Name Server (FFFFFC) | TCP/IP DNS Server (Socket 53 UDP) |
|---|---|---|
| Name server location and access | **Yes** (Well-Known) | **Yes** (via DHCP) |
| Dynamic member registration | **Yes** | **Yes** (Dynamic DNS) |
| Support detail member characteristics description | **Yes** | **Yes** |
| Support full or restricted queries | **Yes** | **No** (partial query only) |
| Automatic replication database | **Yes** | **No** |
| Multi-mastering (full read and write update) in the entire network | **Yes** | **No** (Primary & Secondary) |
| Automatic unify resources view | **Yes** | **No** (manual work) |

# Fabric Name Server (cont.)

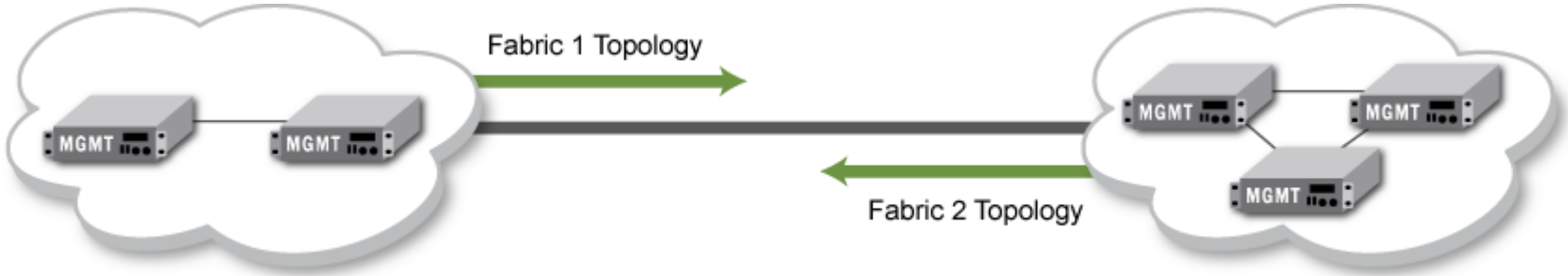
- FC devices may use Name Server (FFFFFC) to “discover” other devices in the Fabric
  - Devices always login to the Fabric, FLOGI to FFFFFE
  - Devices then login (PLOGI) to FFFFFC
  - After PLOGI, devices typically register and send queries to NS at FFFFFC, devices can also deregister
- When do devices communicate with FFFFFC
  - Upon initial connection, after FLOGI
  - After receiving an RSCN from the Fabric Controller

# FFFFFC Port and Node Attributes

- The Name Server database contains two record types:
  - Port Attributes
    - Port Identifier (Native port address ID)
    - Port Name (Port World Wide Name)
    - Class of Service (2, 3)
    - FC-4 Types (FCP, IP)
    - Port Type (N, NL)
    - Device Type (Initiator or Target)
    - Symbolic Port Name (free-form information)
  - Node Attributes
    - Node Name (Node World Wide Name)
    - Fibre Channel IP Address
    - Symbolic Node Name (free-form information)

# Management Server at FFFFF*FA*

- It is accessed by an external Fibre Channel node at the well-known address FFFFFA
  - An application can access information about the entire fabric management with minimal knowledge of the existing configuration
- It is replicated on every Brocade switch within a fabric.
- It provides an unzoned view of the overall fabric configuration
  - Assists in the autodiscovery of switch-based fabrics and their associated topologies, such as exchanging fabric names
- Adding or changing a fabric name does not produce an RSCN



# Communication Protocols
## Established Fabric Communication Processes Overview

- Extended Link Services include
  - FLOGI — Fabric Login
  - ACC — Accept
  - PLOGI — N_PORT Login
  - PRLI — Process Login
  - RSCN — Registered State Change Notification
  - SCR — State Change Registration
  - LOGO — Logout

- Fabric Devices typically
  - FLOGI → PLOGI to Name Server → SCR to Fabric Controller → Register & Query [using **Fibre Channel Common Transport** (FC_CT) Protocol] → LOGO

# Login Services

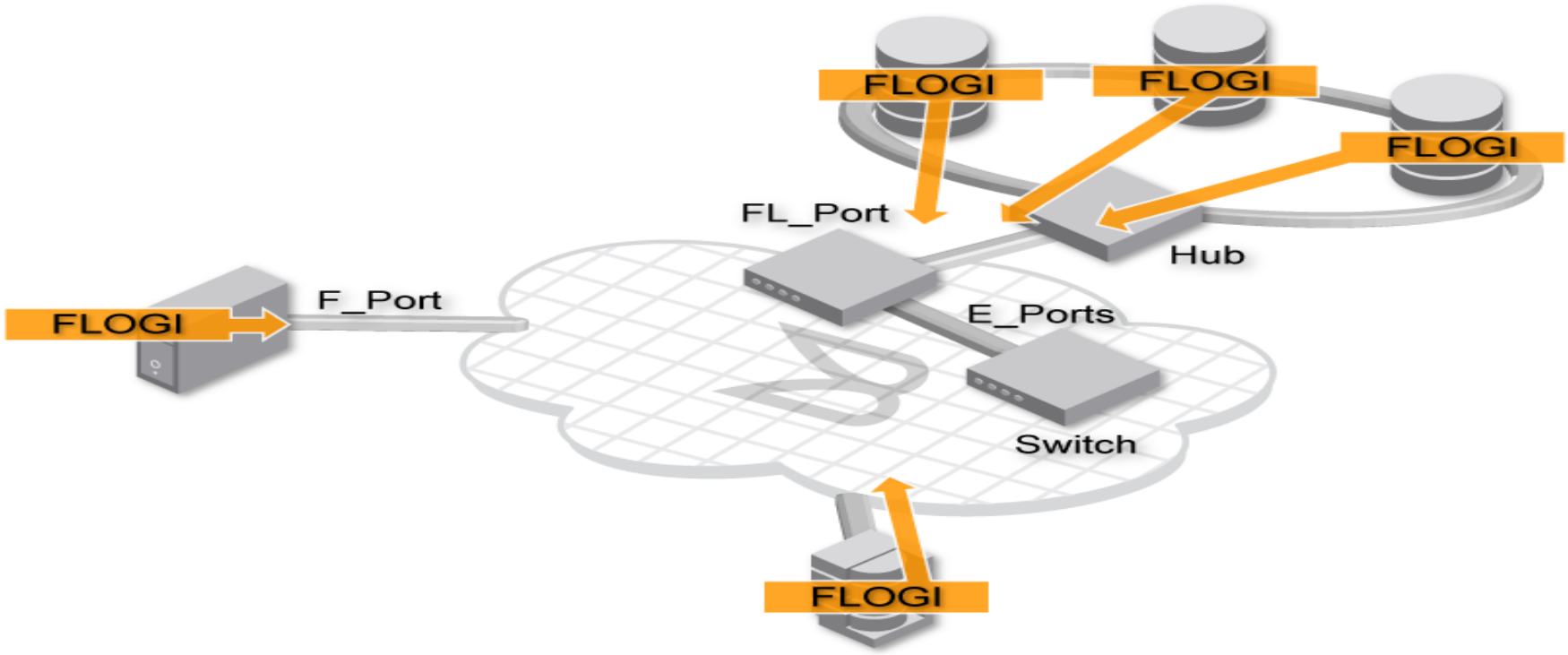
- For Fibre Channel, these are the login services:
  - Fabric Login (`FLOGI`) is used by an N_Port or NL_Port (Nx_Ports) to establish service parameters with the Fabric
    - The following information is implicitly captured and put into the Name Server during this process: type; COS; PID; port name (port WWN) ; and node name (node WWN)
  - N_Port Login (`PLOGI`) is used by one Nx_Ports to establish service parameters with another N_Port or NL_Port
  - Process Login (`PRLI`) is used by an upper-level process, such as SCSI, in one port to establish the upper-level process in the other port to enable SCSI command processing

# Fabric F_Port Service: Common Fabric Services

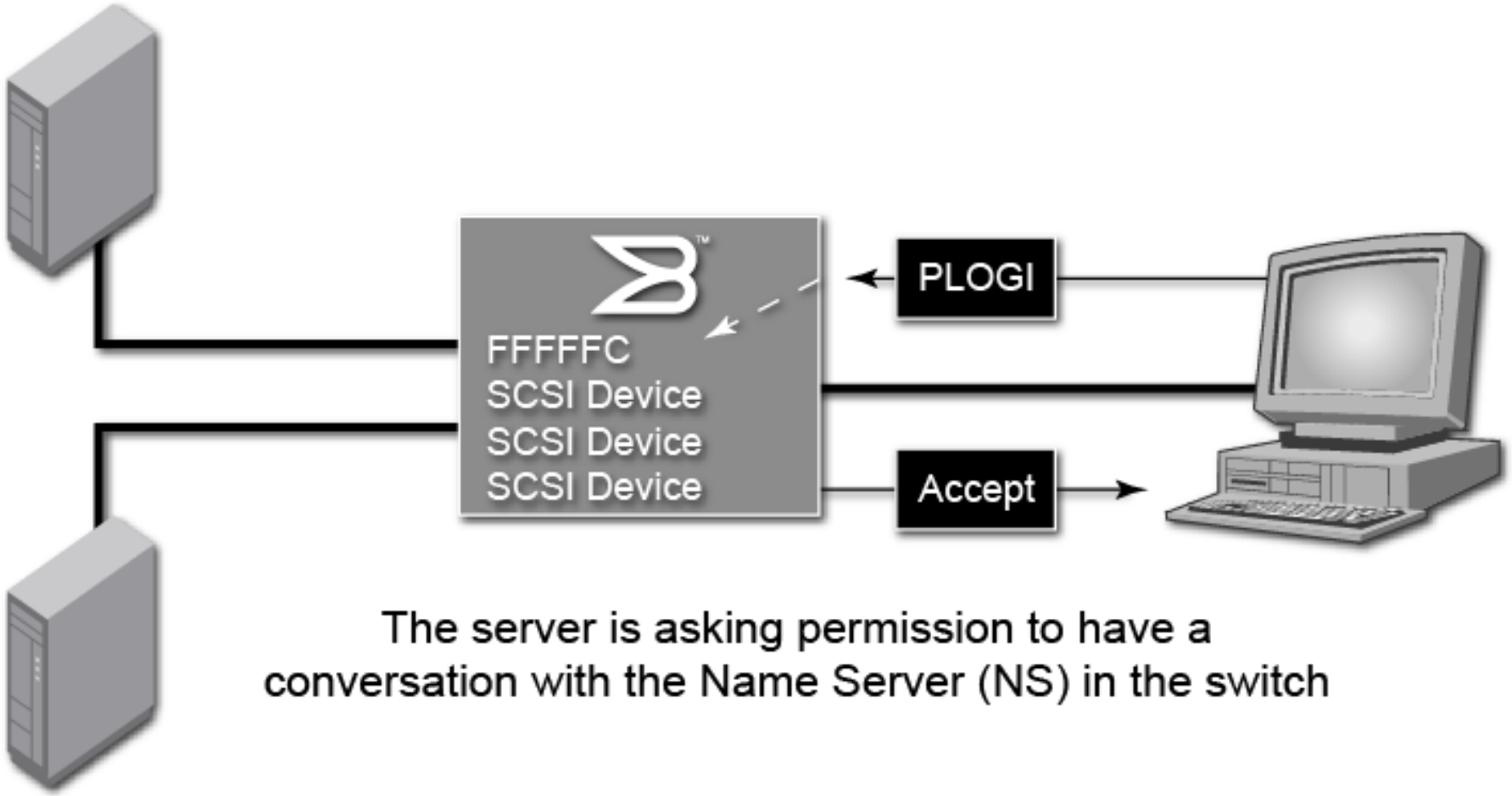
- Fabric “F_Port” Service is also referred to as Fabric Login
  - You need the Fabric F_Port to establish Fabric communication capability
- The Fabric F_Port accepts devices into the Fabric
  - FC devices must Fabric Login (`FLOGI`) to the Fabric F_Port to be part of the Fabric
  - This FLOGI is sent to Well-Known Address FFFFFE
- When do devices normally communicate with the Fabric F_Port?
  - Devices will `FLOGI` upon initial connection to the Fabric
  - Upon re-connection (possibly due to a link loss)

# Fabric Login (FLOGI)

- When devices first connect to the fabric, their address is 000000
- FLOGI is required before any frame can be sent through the fabric
- FLOGI is sent to well-known address FFFFFE (Fabric F_Port)



# Port Login (PLOGI)



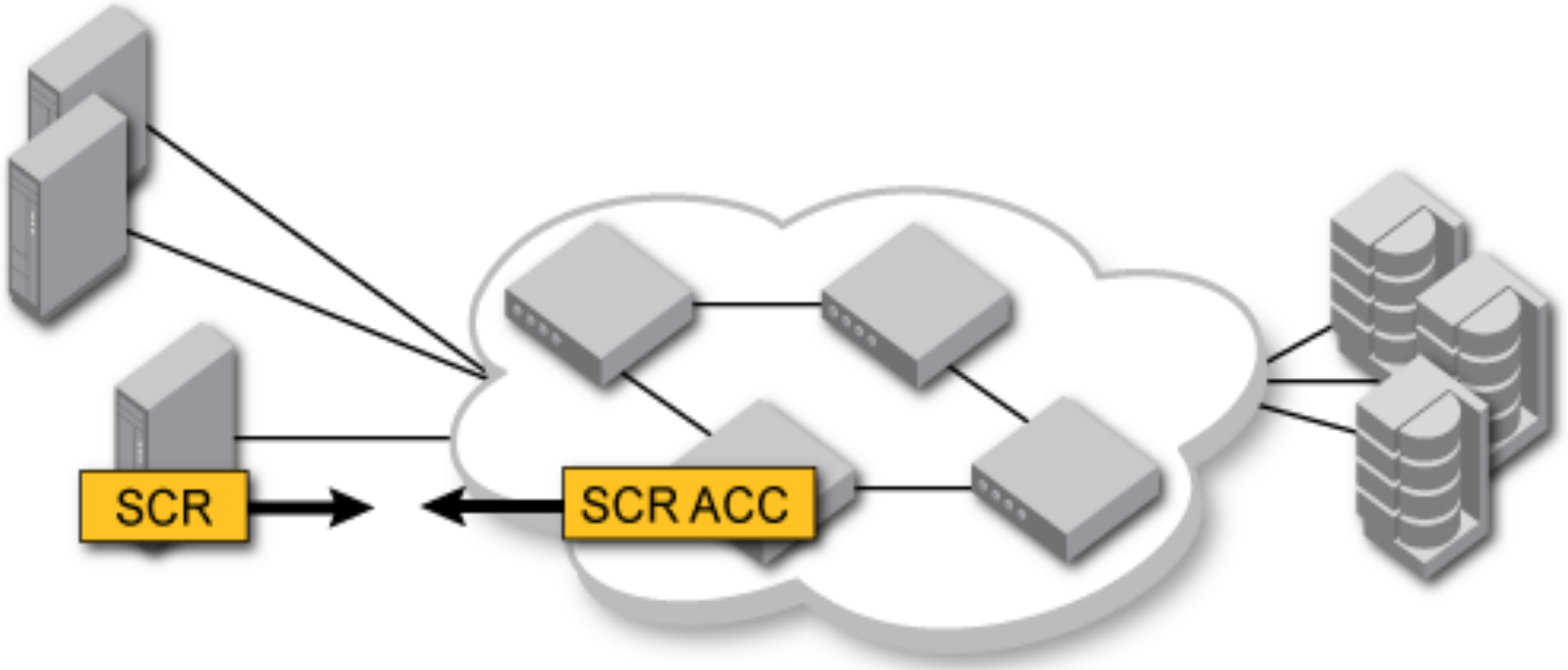

The server is asking permission to have a conversation with the Name Server (NS) in the switch

# State Change Notification Services

- **State Change Notification** (**SCN**) – State Change Notifications (SCN) are used for internal state change notifications, not external
  - This is the switch logging that the port is online or is an Fx_port
  - This is not sent from the switch to the Nx_ports!
- **State Change Register** (**SCR**) – Nx_Port request to receive notification when something in the Fabric changes
  - FC Devices that choose to receive RSCNs must register for this service
    - Devices send a State Change Registration (SCR) to the Fabric Controller
    - Registration indicates that the device wants to be notified of changes
  - Devices normally register immediately after a PLOGI to Name Server, but could do so at any time after PLOGI
  - Always optional; SCSI initiators usually register, SCSI targets do not
- **Registered State Change Notification** (**RSCN**) – Issued by the Fabric Controller or an Nx_Port to devices that registered (issued an SCR requesting this notification)
  - Brocade provides methods to help devices deal with RSCNs

# Fabric Controller at FFFF*FD*

- Responsible for operation of the fabric
  - Handles switch-to-switch (Class F) traffic
- Receives node requests for State Change Registration (SCR)





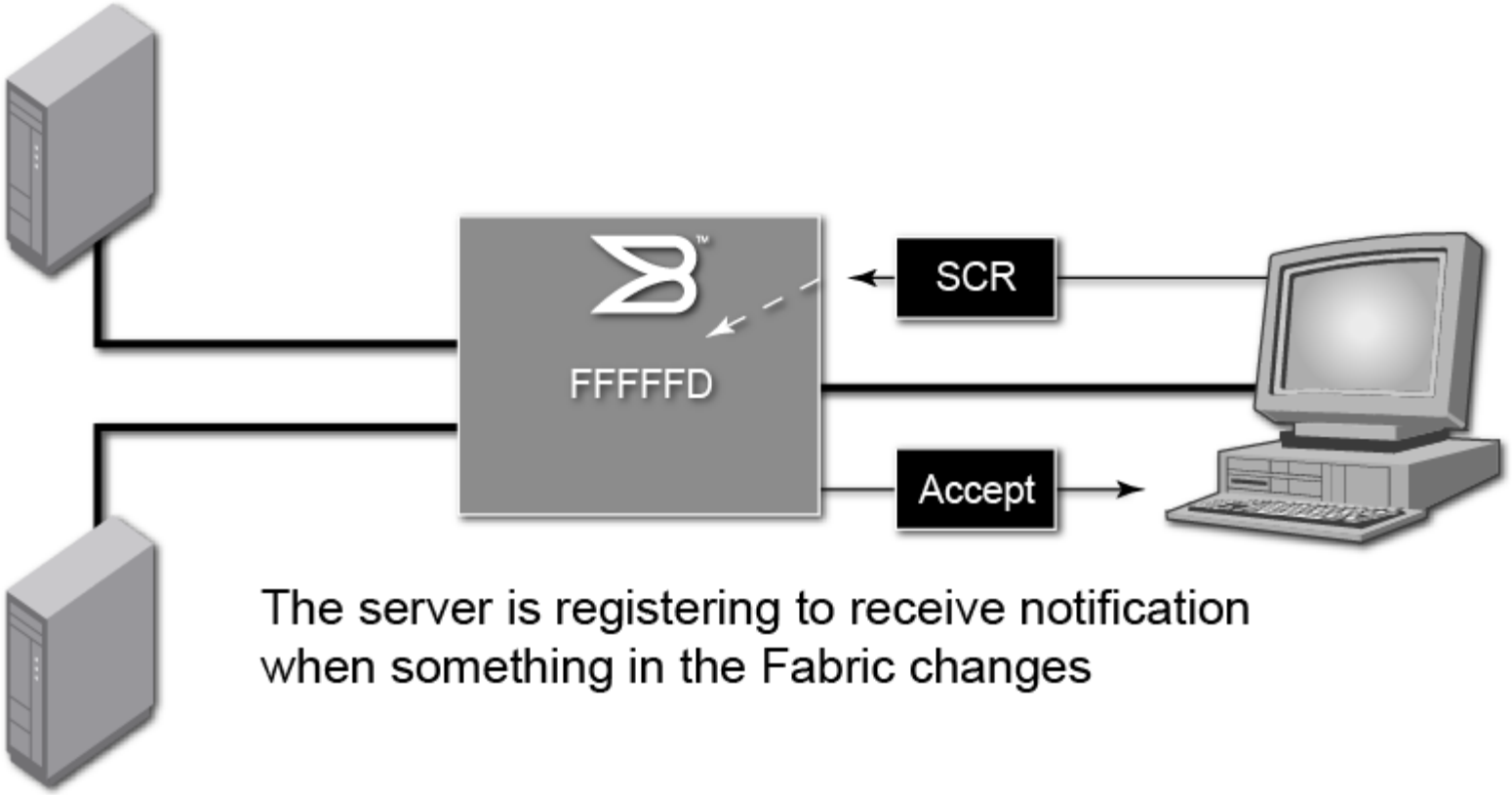


BROADCOM

# State Change Registration (SCR)

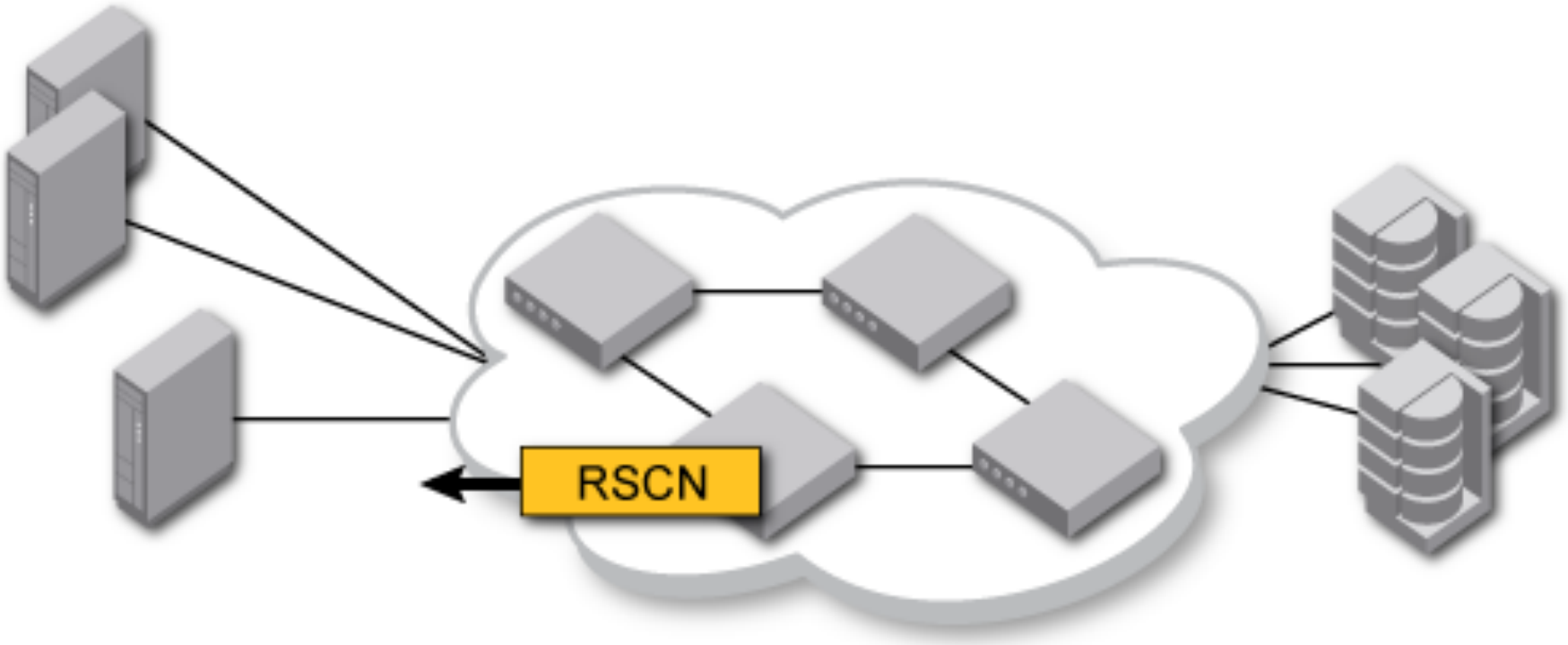
- Devices register to receive notifications for fabric changes with a State Change Registration (SCR)
- The Fabric Controller provides a a Registered State Change Notification (RSCN) when changes occur
- Automatic Network Notification: it eliminates unnecessary polling traffic. Fabrics use a notification mechanism to notify registered nodes of any changes in the Fabric

# State Change Registration (SCR) (cont.)



# Fabric Controller at FFFFFD (cont.)

- Distributes Registered State Change Notifications (RSCNs) to registered nodes
  - Sent when a device needs to know about a state change in the fabric or a device attached to it
  - A device must send an SCR to receive RSCNs

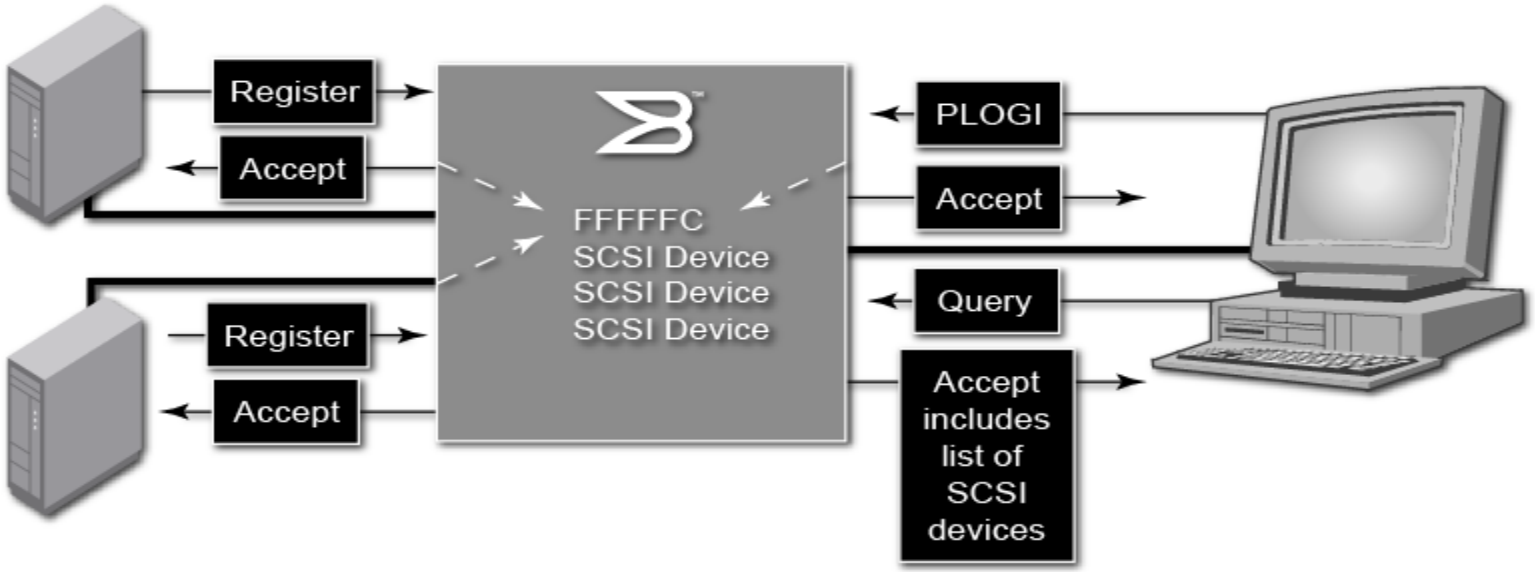


# Brocade RSCN Delivery

- Device power on or shutdown causes an aggregated RSCN to be sent to affected zoned devices only
- Zoning changes cause a fabric RSCN to be sent to affected zoned devices only
- Once devices receive an RSCN, they should query the Name Server to determine what devices they have access to
- Fabric reconfigurations with no domain change do not cause an RSCN to be sent

# Registration and Query Processes

- A PLOGI to the Name Server is required before devices can register or query
  - Typically, both SCSI initiators and targets register
  - Additionally, SCSI initiators usually query for a list of allowed communication devices so they can build or rebuild device tables.
- Recall that registration and query processes use the Fibre Channel Common Transport (FC_CT) protocol



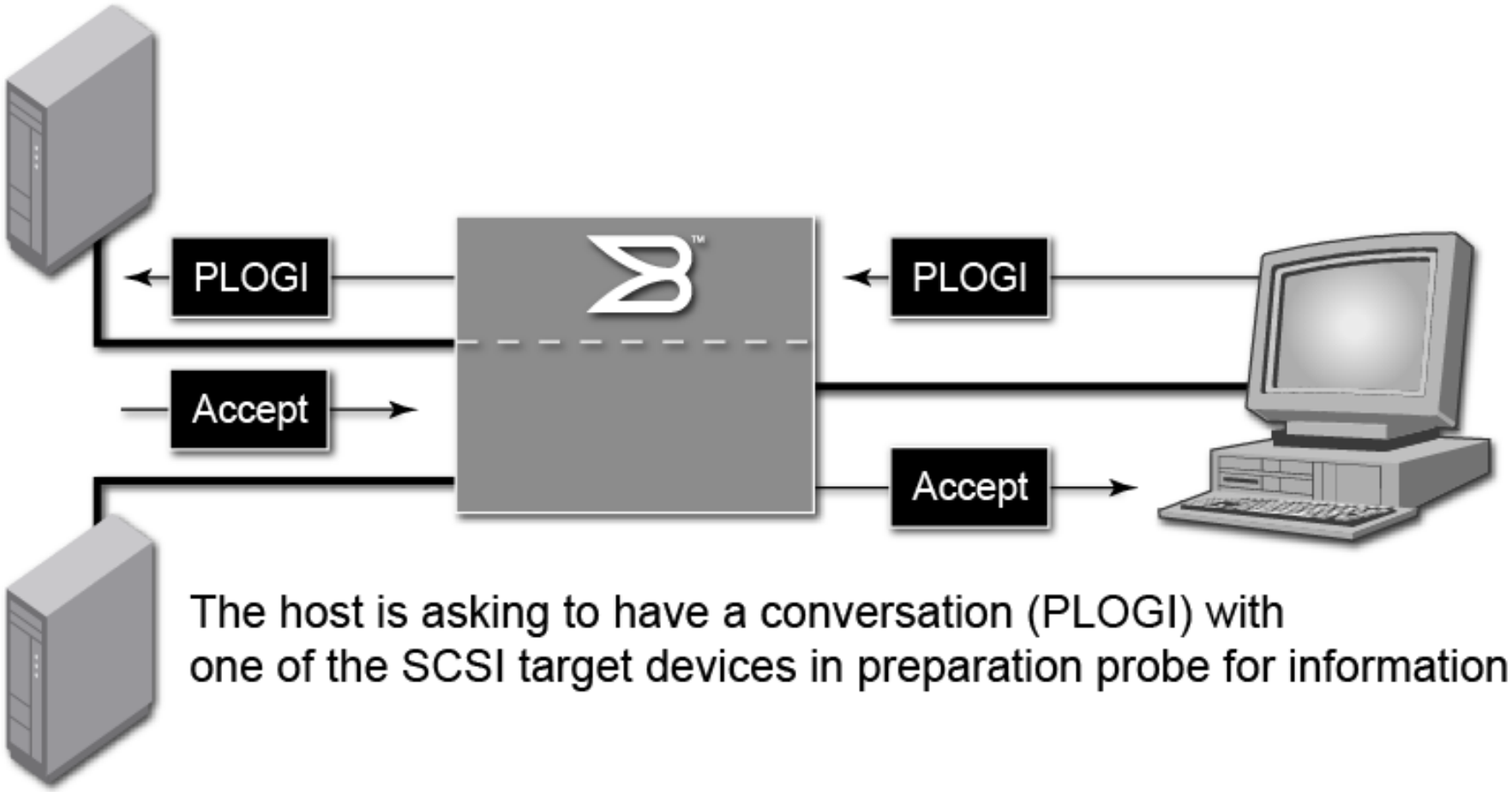
# SAN Attached Initiators and Target Initiator Communication Process Review

- Fabric Initiators are typically server HBAs that will:
  - FLOGI to FFFFFE and obtain 24-bit Fabric address
  - PLOGI to FFFFFC
  - SCR to FFFFFD
  - Register and query FFFFFC for database of Storage/Tape targets
    - FFFFFC matches HBA to zones (Fabric subsets containing devices you "allow" to communicate) and returns 24-bit addresses of targets that HBA requested within authorized zones
    - PLOGI and then sends SCSI probes to fabric destination addresses of targets returned from Name Server
  - Assign target ids (with server OS) to each node probed to build a device table that applications can use to store and retrieve data

# SAN Attached Initiators and Target (cont.)
## Target and Loop Communication Processes Review

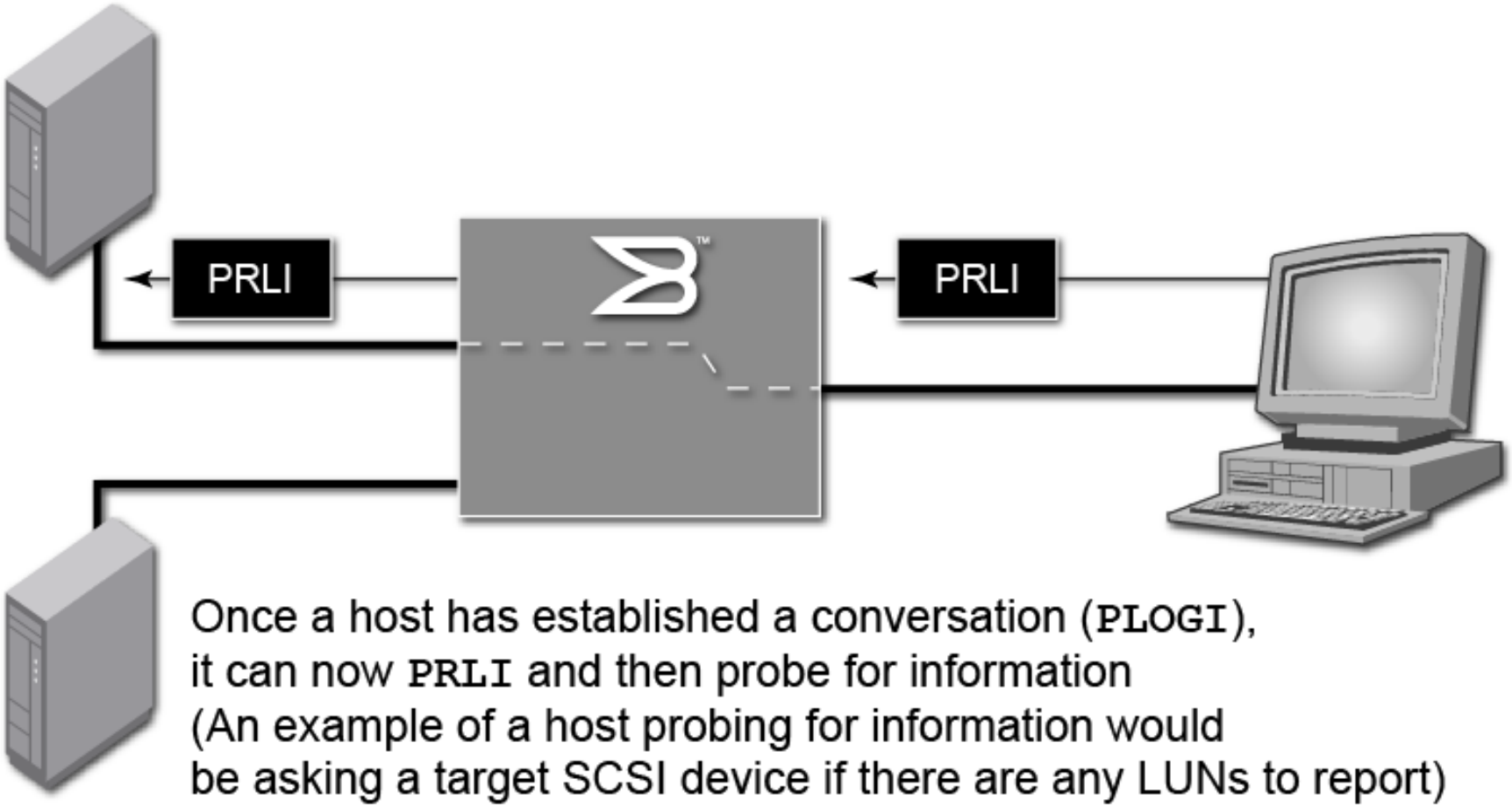
- SAN Targets are typically storage devices:  Fabric RAID controllers, JBODs (just a bunch of disks), and Tape including libraries
- Non-loop Fabric capable Targets  typically:
  - `FLOGI` to `FFFFFE` and obtain full 24-bit Fabric address
  - `PLOGI` to `FFFFFC`
  - Register  information with `FFFFFC`
    - Targets typically register symbolic node information that allows easy identification using Name Server commands (`nsshow`)

# Initiators PLOGI to Target(s)



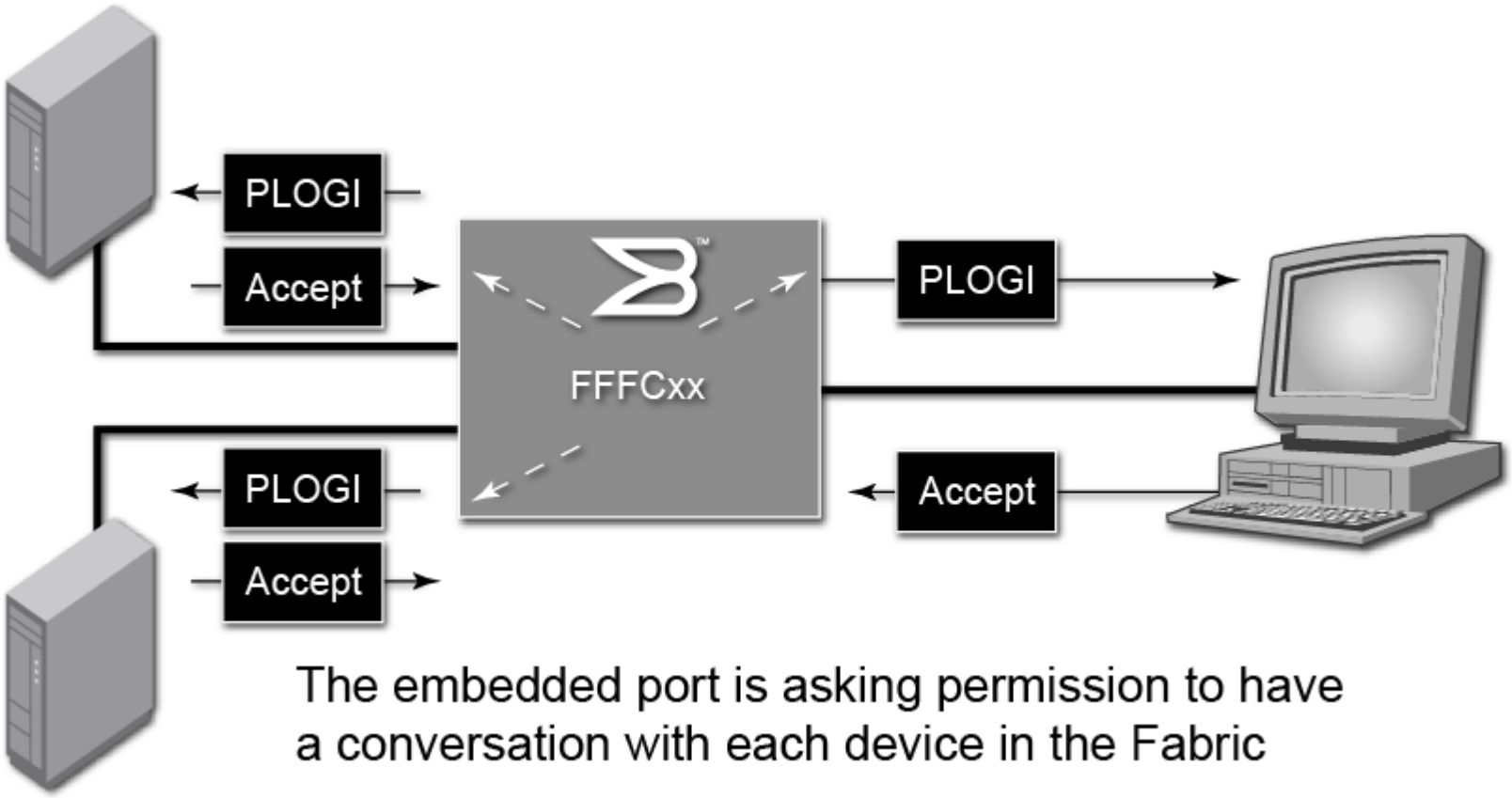

The host is asking to have a conversation (PLOGI) with one of the SCSI target devices in preparation probe for information

# Initiator PRLI to SCSI Probe / Inquiry



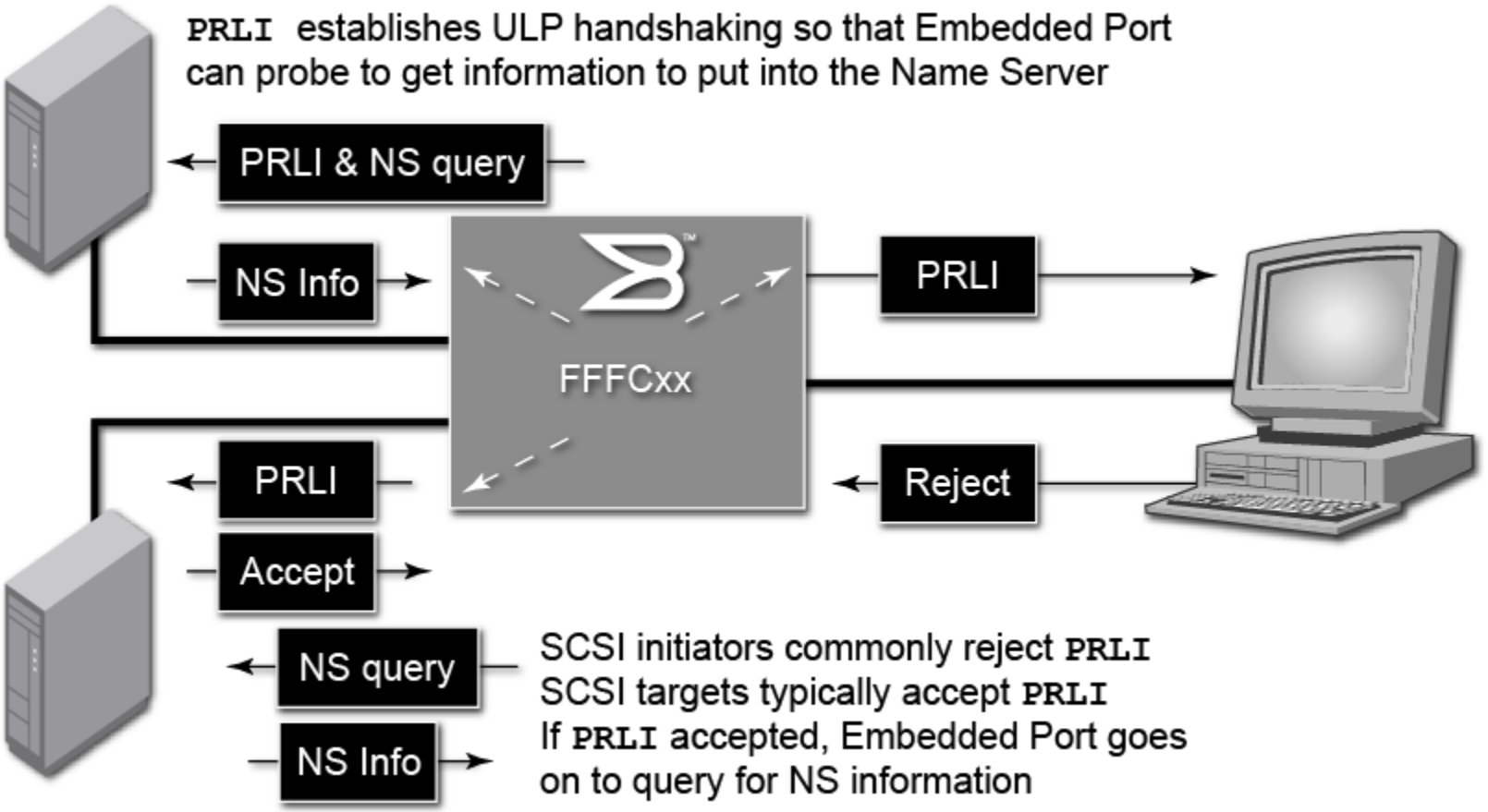

Once a host has established a conversation (PLOGI),
it can now PRLI and then probe for information
(An example of a host probing for information would
be asking a target SCSI device if there are any LUNs to report)

# Embedded Port Login (PLOGI)




The embedded port is asking permission to have a conversation with each device in the Fabric

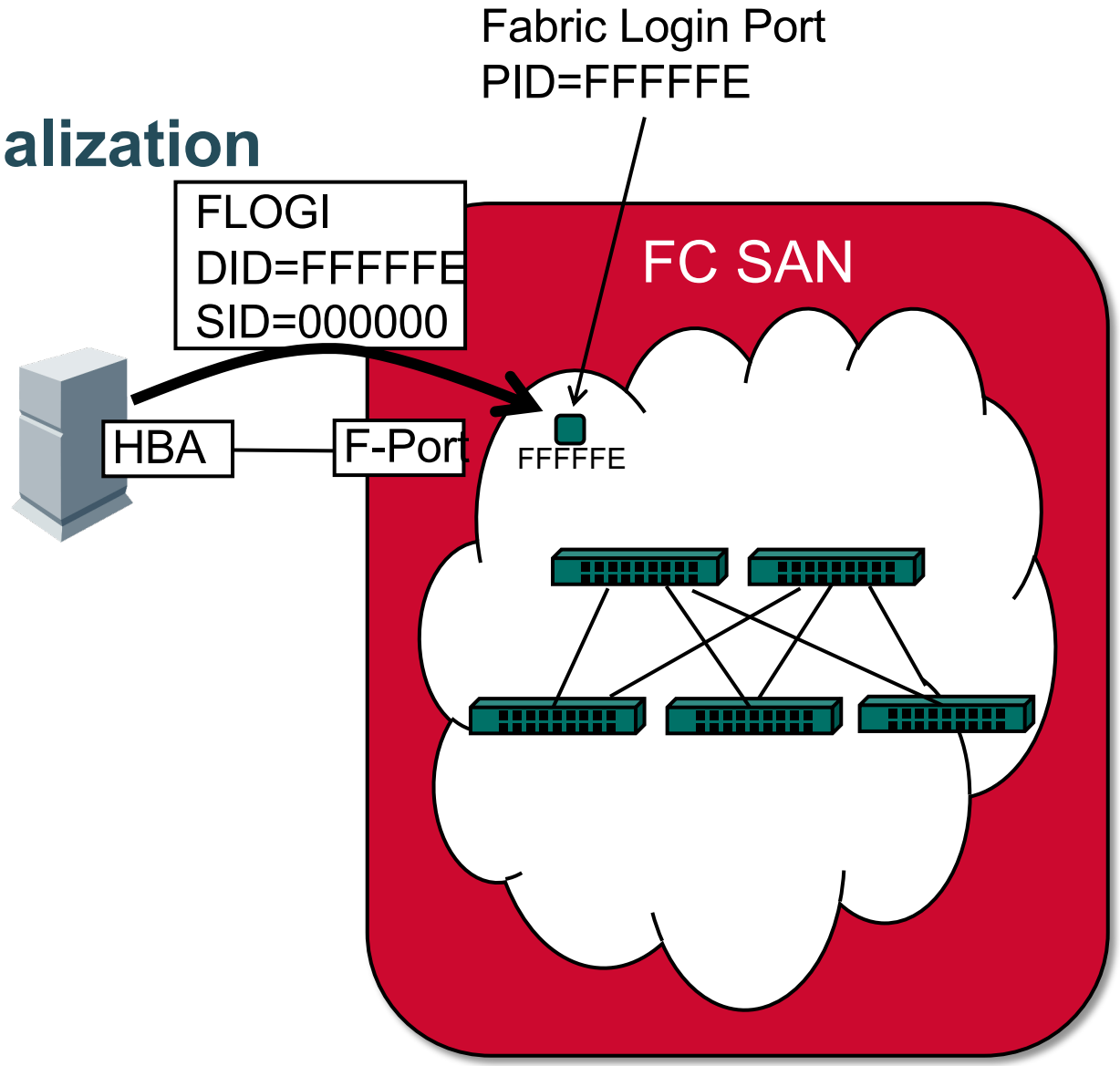
# Embedded Port Process Login (PRLI)



# Fabric Login

## The Process of Device Initialization

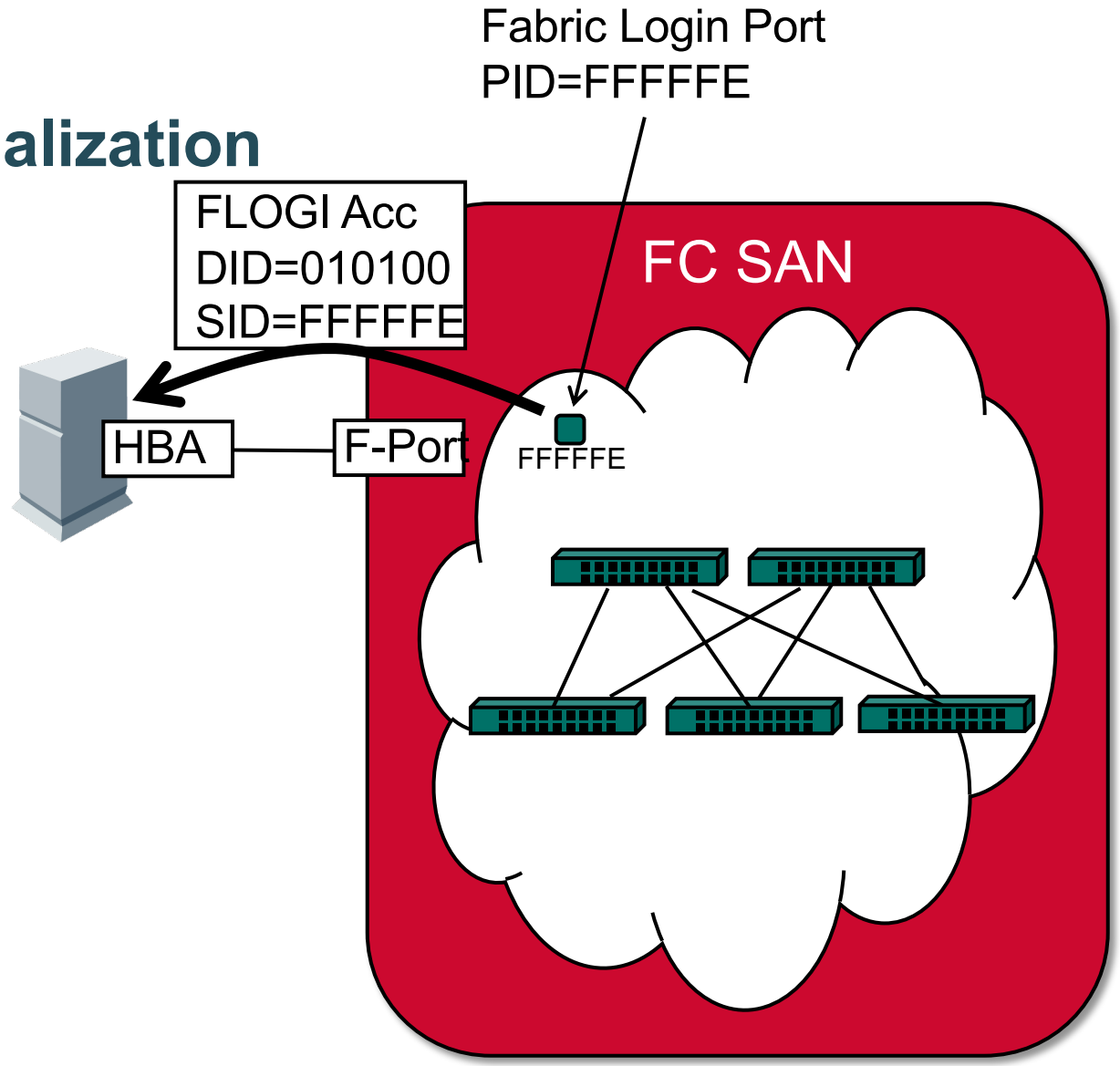
- First, the host sends a FLOGI (Fabric Login) to the Fabric Login Port's Well Known Address.
  - The destination address is the PID of the Fabric Login Port (xFFFFFE).
  - However, the source address on the FLOGI is "x000000". (The host doesn't know yet what its PID is.)



# Fabric Login

## The Process of Device Initialization

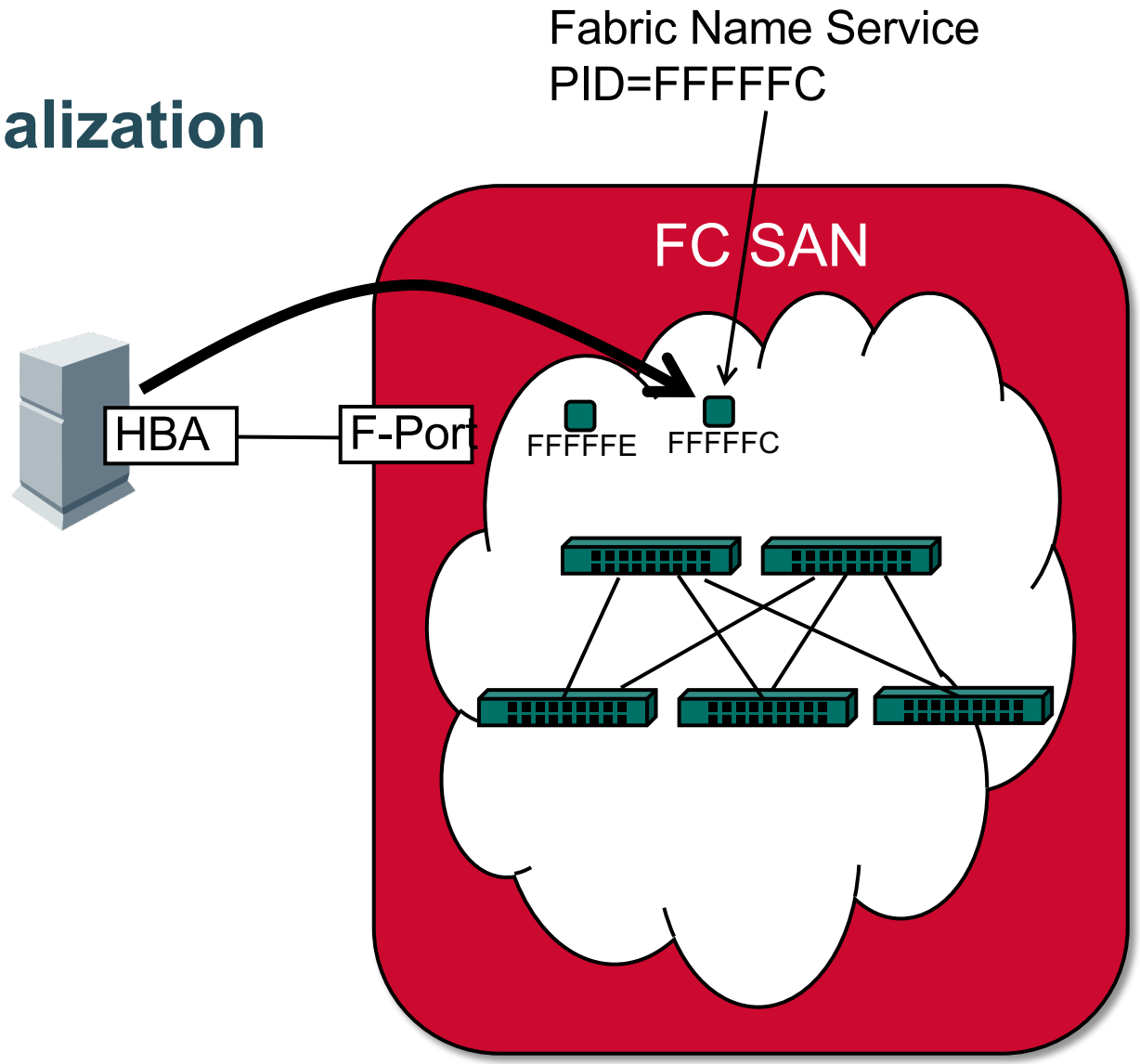
- The Fabric Login port responds with the hosts' newly-assigned PID.
  - Now the host has its PID! In this case it's x010100.
    - PID tells us:
      - Domain of the switch it's plugged into is "x01"
      - It's physical port number is "x01"
      - (Not hard and fast rule though)



# Name Service

## The Process of Device Initialization

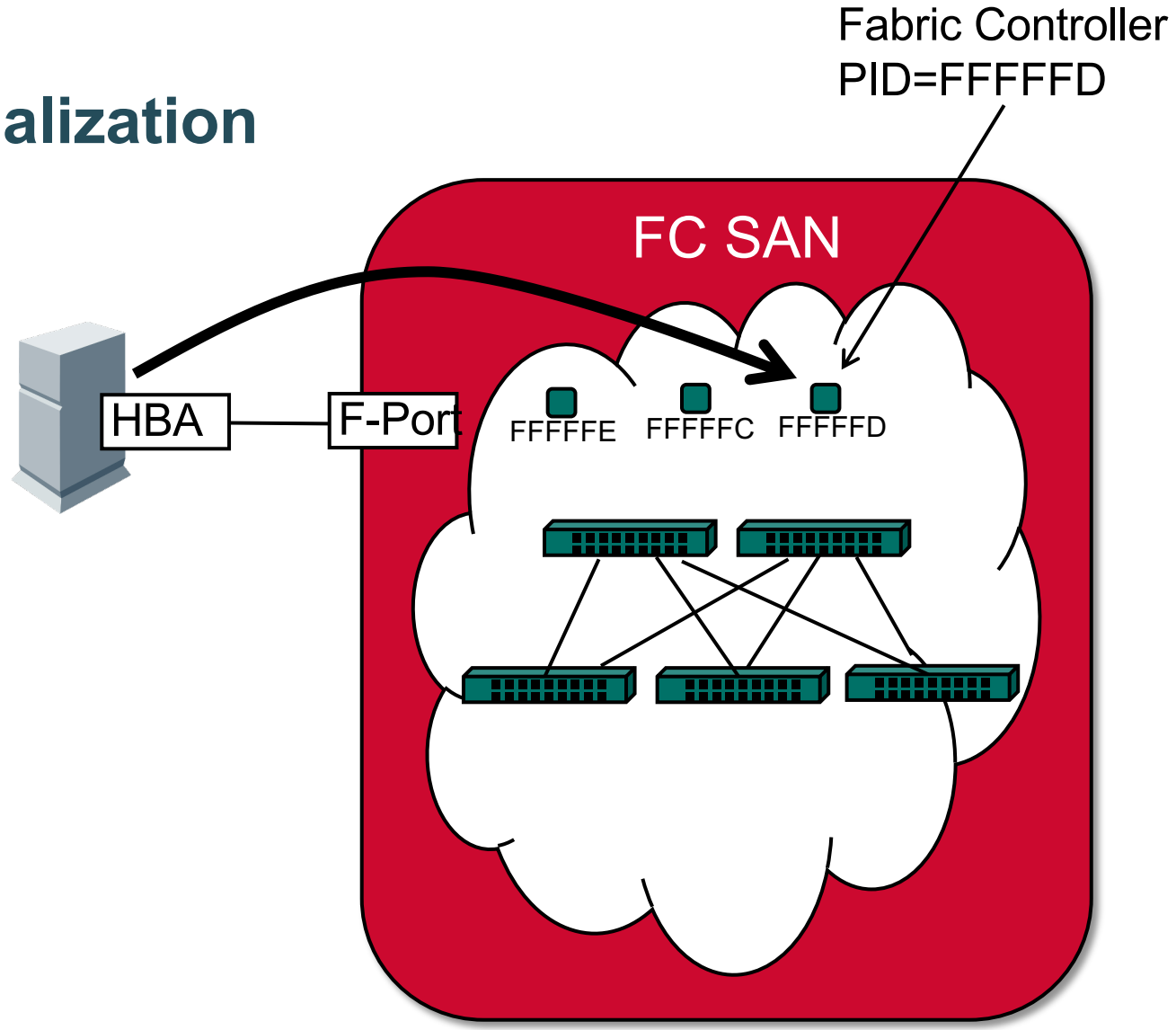
- Next, the host will login (PLOGI) to the port of the Name Service (FFFFFC).
  - It "registers" with the name service.
    - "Here's my PID, my WWNs, who made me (vendor specific information), more stuff, etc."



# Fabric Controller

## The Process of Device Initialization

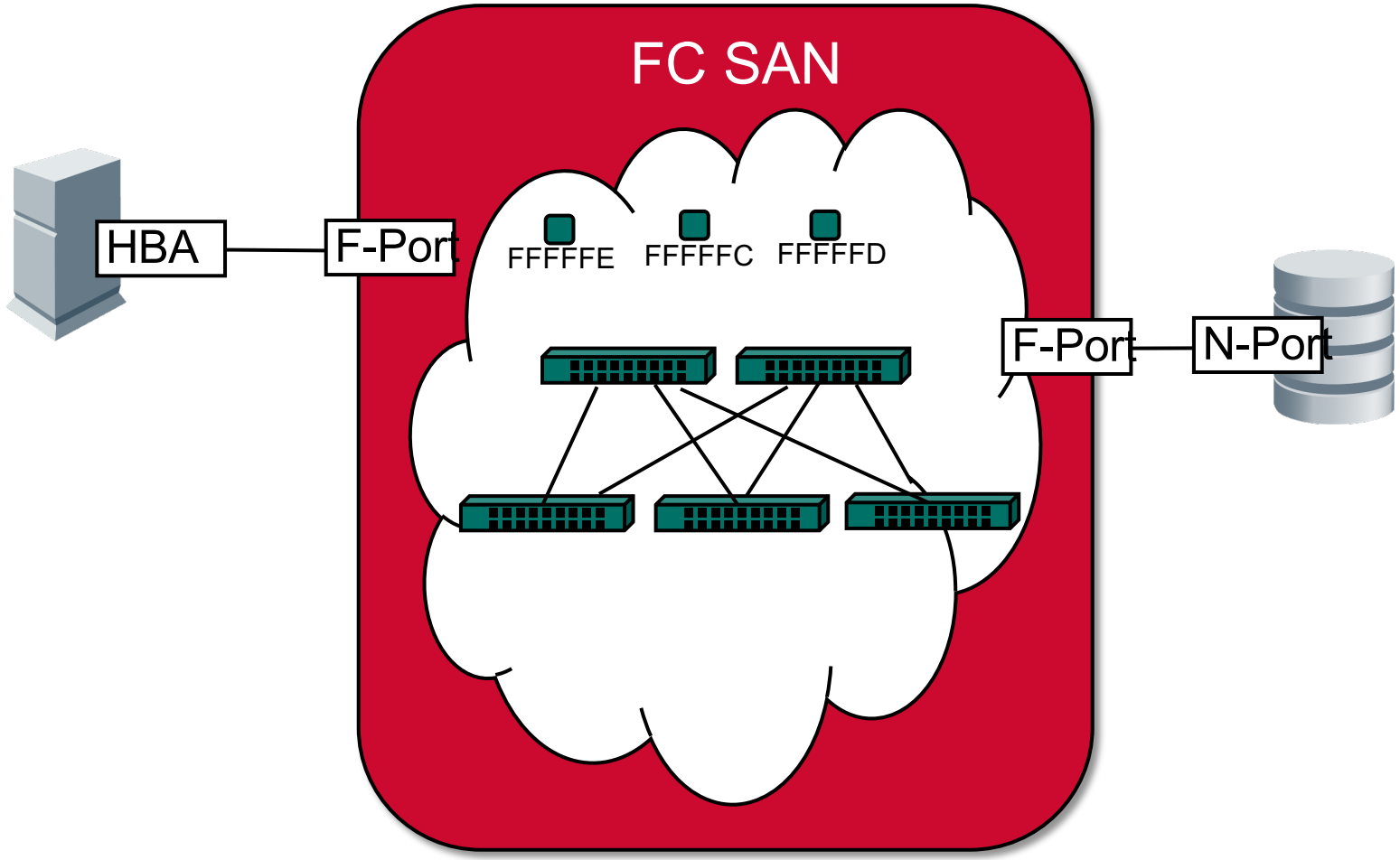
- Next, the host will PLOGI to the Fabric Controller (FFFFFD).
  - It "registers" for "State Change Notification".
    - "Mr. Fabric Controller, if anything changes in this network [that I need to know about] please notify me" I am Registering for "State Change Notification" (RSCN).



# Fabric Controller

## The Process of Device Initialization

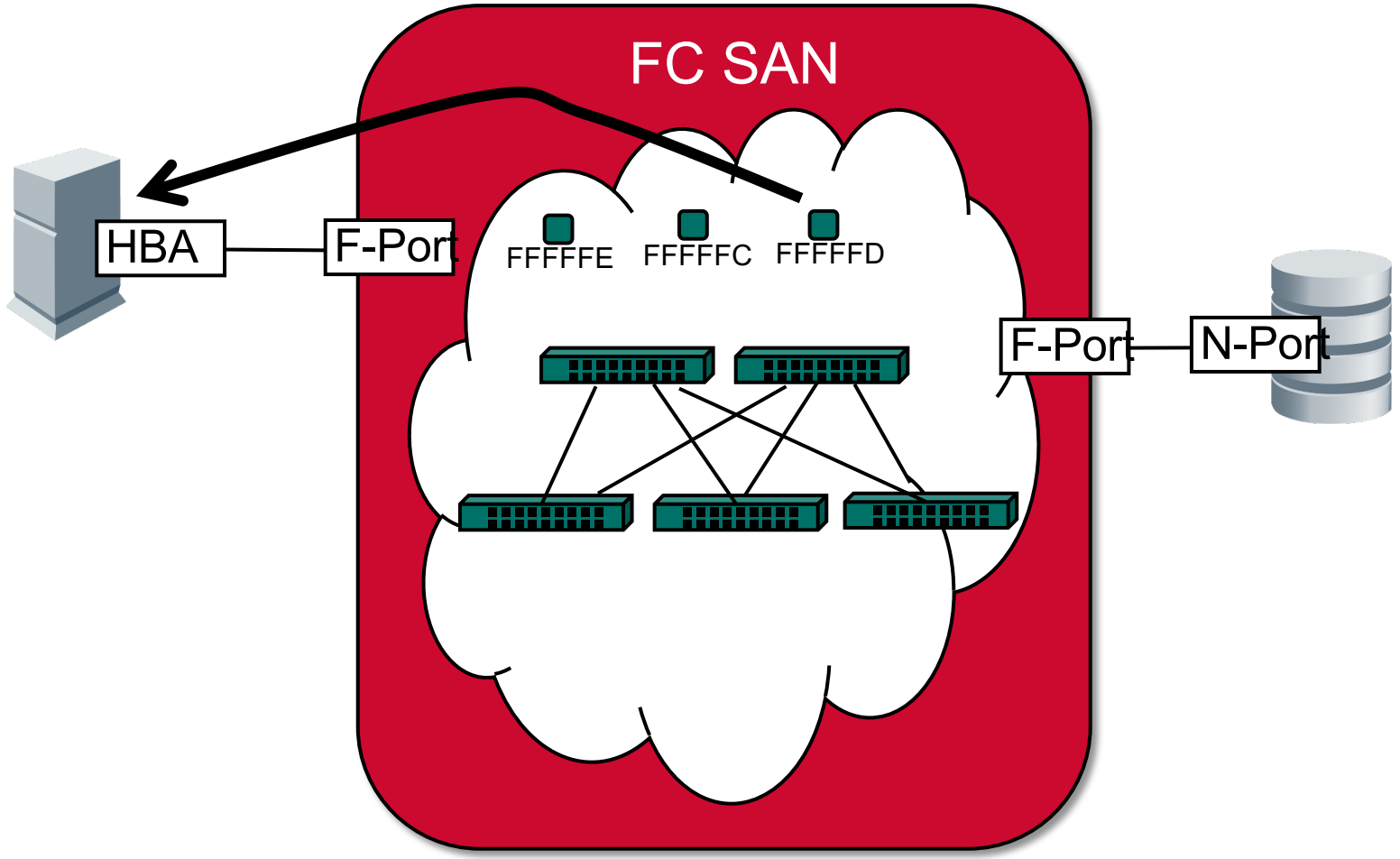
- Hey, what's that? A new storage device just came online, logged into the fabric and registered with the name service!



# Fabric Controller

## The Process of Device Initialization

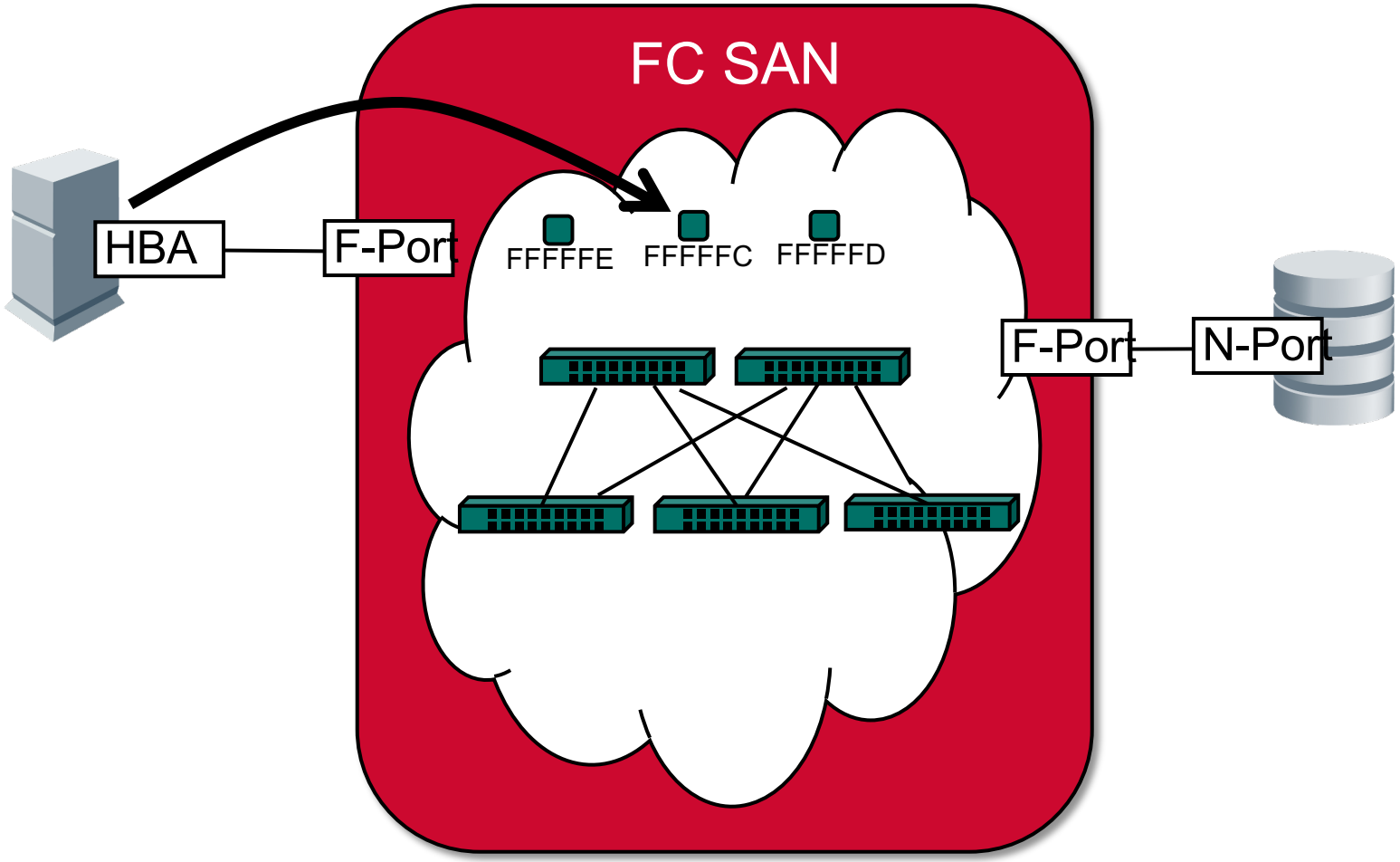
- If the new storage is in the same zone as the host, the fabric controller notifies the host of the change (State Change Notification).



# Fabric Controller

## The Process of Device Initialization

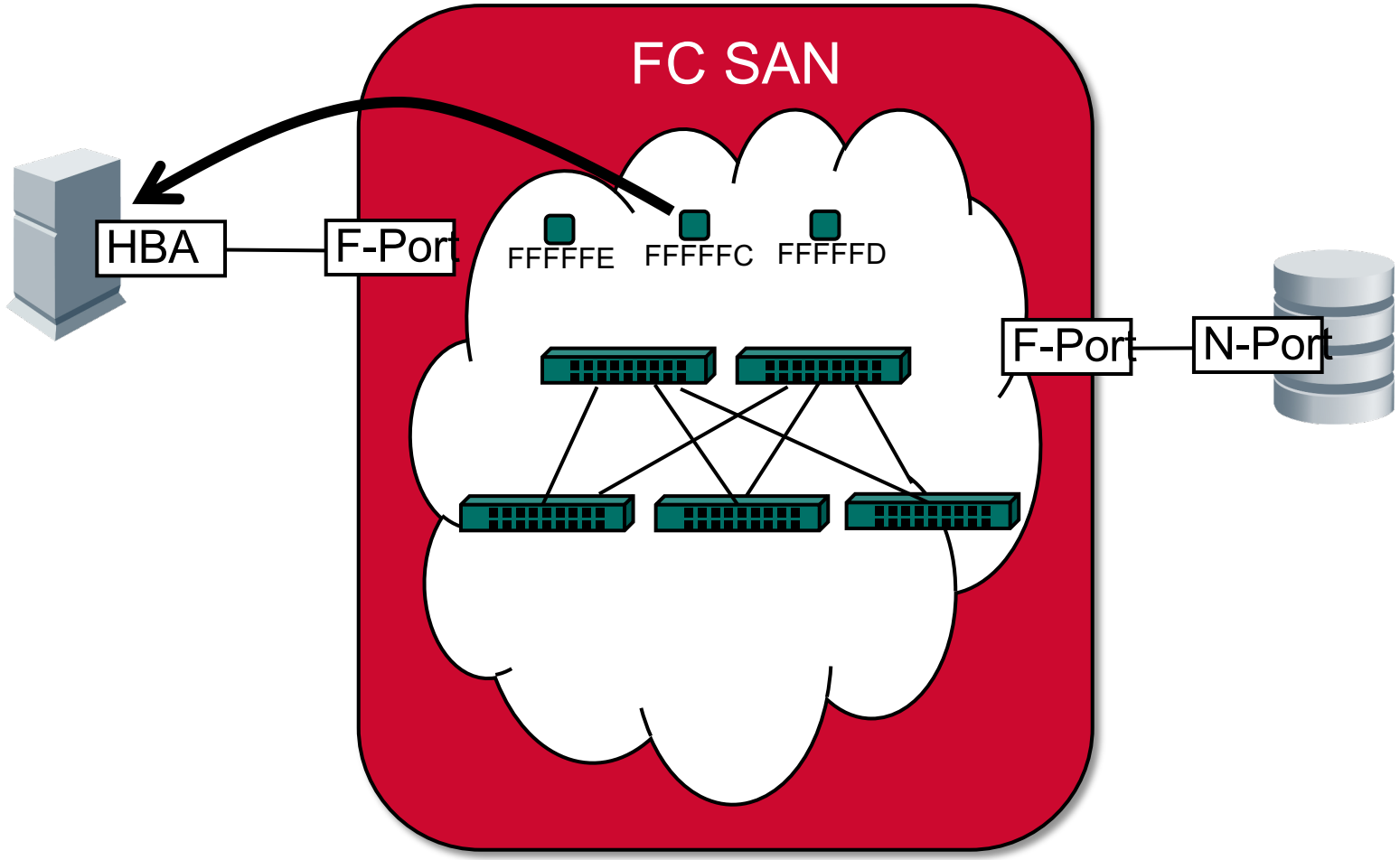
- The host PLOGI's back into the Name Service
  - Queries the NS for updated list of available devices



# Fabric Controller

## The Process of Device Initialization

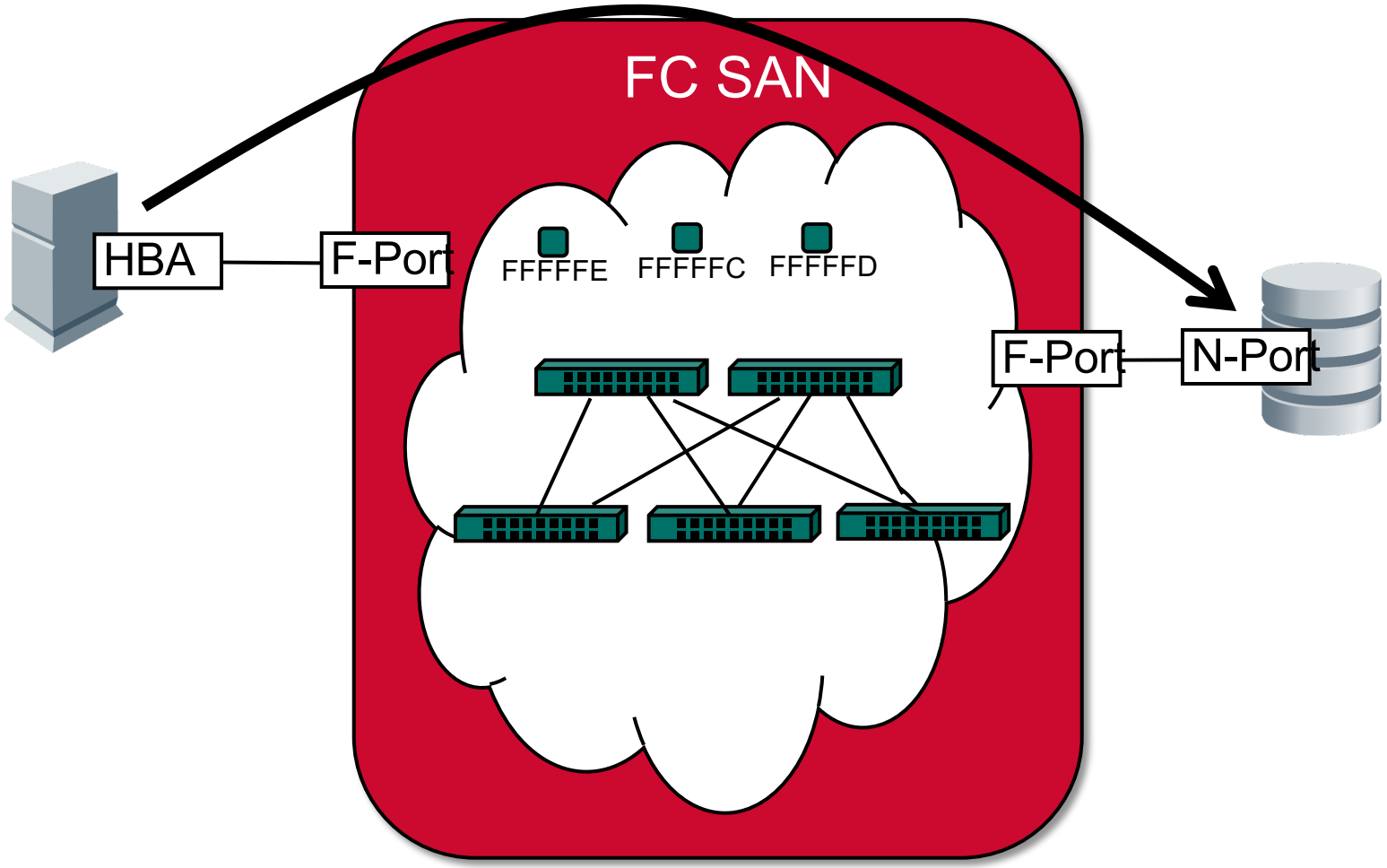
- Name service responds with a list of updated device PID's that are in the same zone as the host.



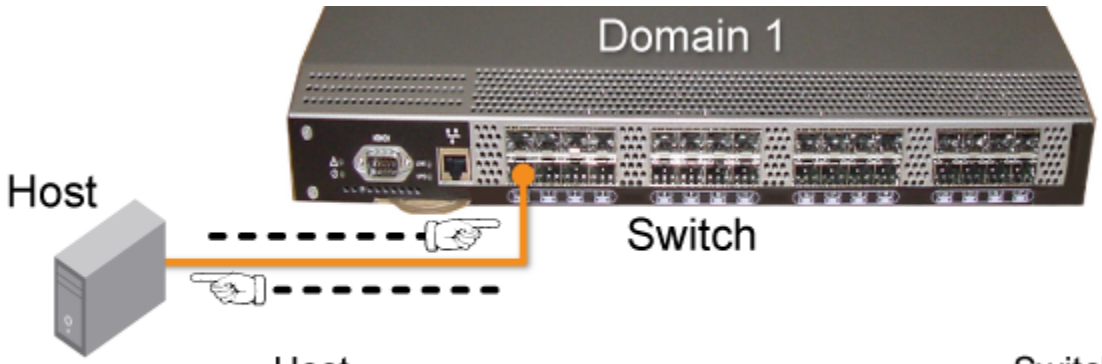
# Fabric Controller

## The Process of Device Initialization

- The host PLOGI's into the new storage device.
- Once logged in, the host can perform a SCSI probe to detect Logical Unit Numbers (LUN's).
- Once SCSI probe is complete, the storage can be formatted with a file system and mounted by the host's operating system.

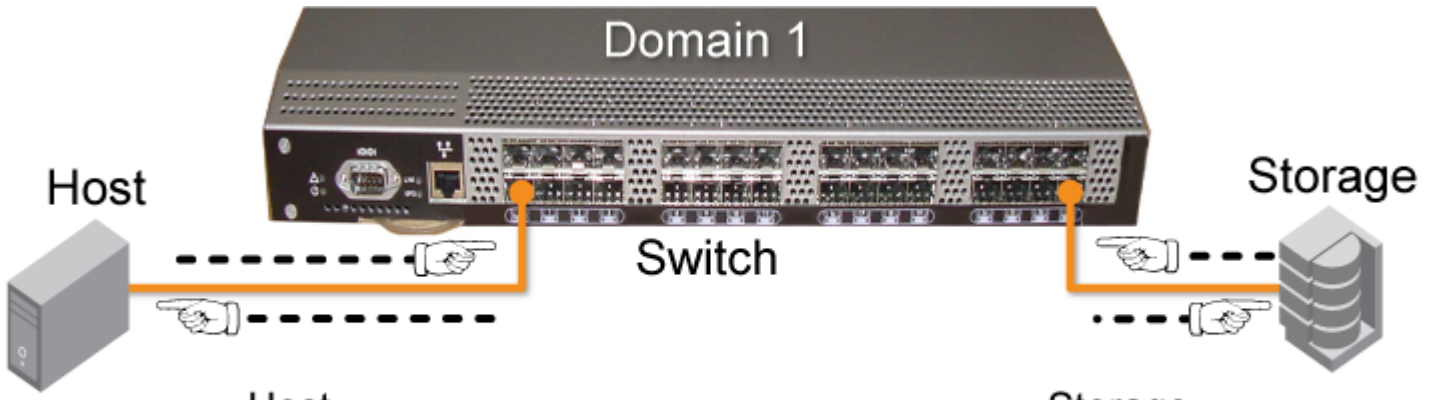


# Device Communication Example




| | Host | | | Switch | | |
|---|---|---|---|---|---|---|
| | DID | SID | REQUEST | DID | SID | REPLY |
| Fabric Login | FFFFFE | 000000 | FLOGI | 010400 | FFFFFE | ACCEPT |
| Port Login | FFFFFC | 010400 | PLOGI | 010400 | FFFFFC | ACCEPT |
| State Change Registration | FFFFFD | 010400 | SCR | 010400 | FFFFFD | ACCEPT |
| Register Attributes with Name Server | FFFFFC | 010400 | NS REG. | 010400 | FFFFFC | ACCEPT |
| Query for Devices | FFFFFC | 010400 | QUERY | 010400 | FFFFFC | ACCEPT |

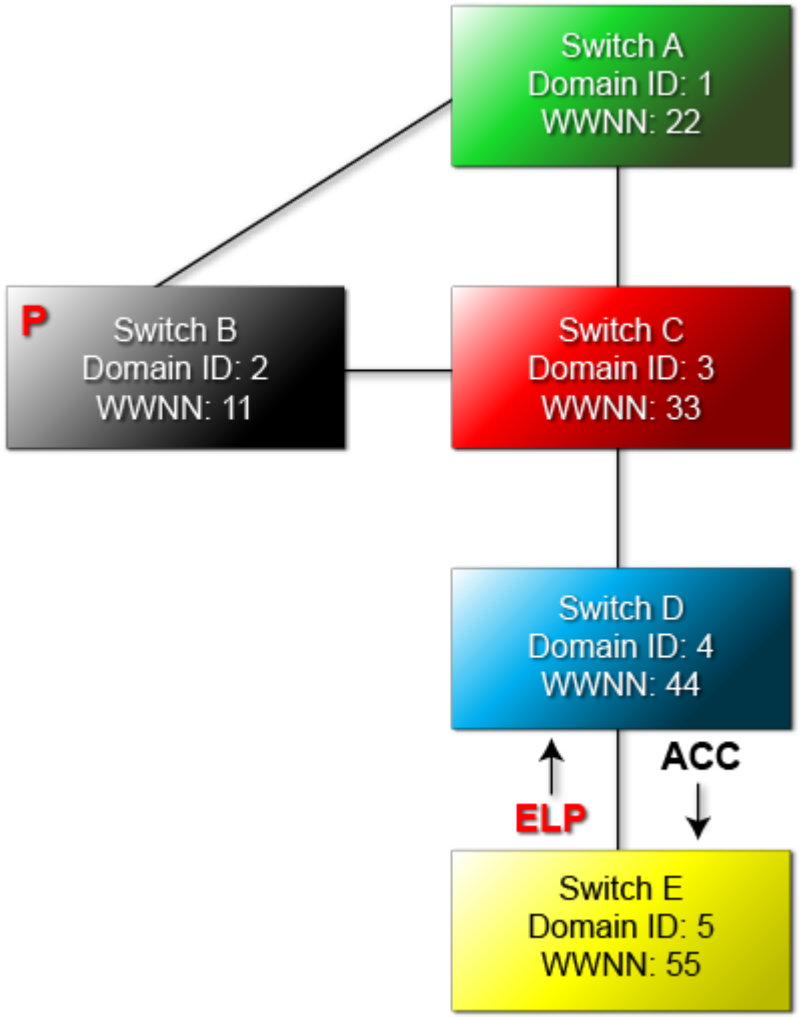
# Device Communication Example (cont.)




| | Host | | | Storage | | |
|---|---|---|---|---|---|---|
| | DID | SID | REQUEST | DID | SID | REPLY |
| Port Login | 011F00 | 010400 | PLOGI | 010400 | 011F00 | ACC |
| Process Login | 011F00 | 010400 | PRLI | 010400 | 011F00 | ACC |
| Inquiry | 011F00 | 010400 | INQ | 010400 | 011F00 | RSP |
| | | | | 010400 | 011F00 | SCSI STAT |

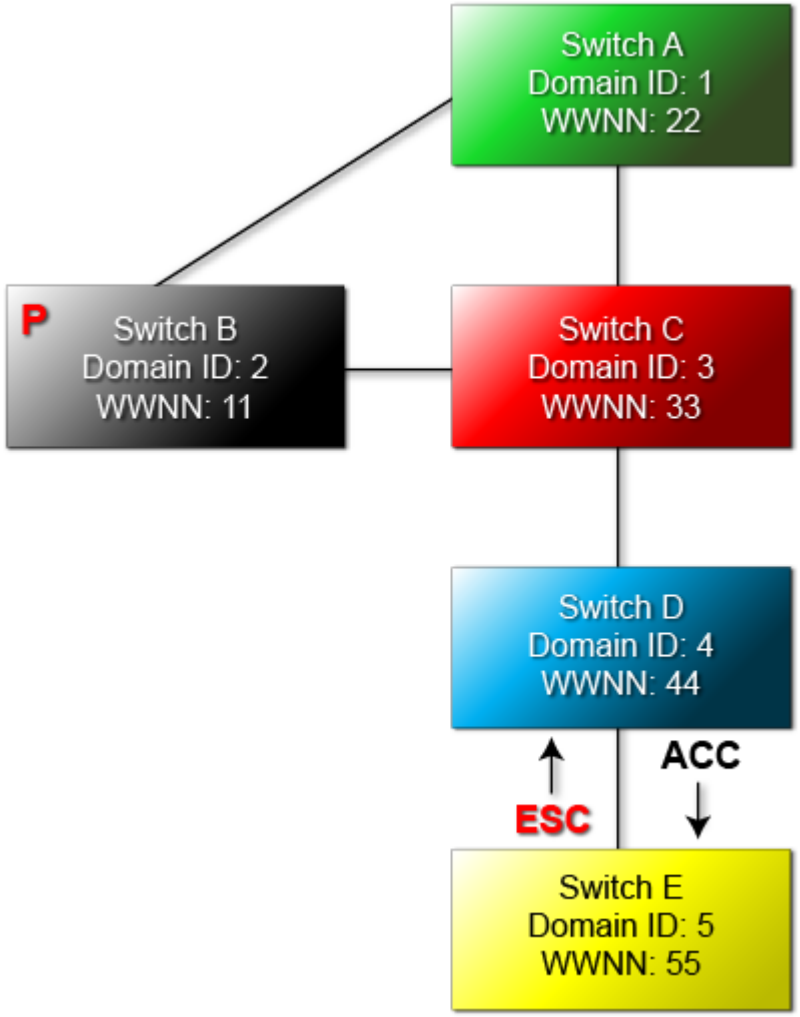
# Fabric Initialization Process

- ELP (Exchange Link Parameters)
  - Contains sender information
  - R_A_TOV / E_D_TOV
  - WWPN / Switch Name
  - Flow control used
  - more…
- ACC (accept frame)
  - Contains responder information
  - R_A_TOV / E_D_TOV
  - WWPN / Switch Name
  - Flow control used
  - more...
- Problems show up as:
  - Segmented / Isolated E_Port
- Reason / explanations
  - Incompatible Link Parameters
  - Incompatible flow control
  - Unauthorized switch name
  - ELP timeout



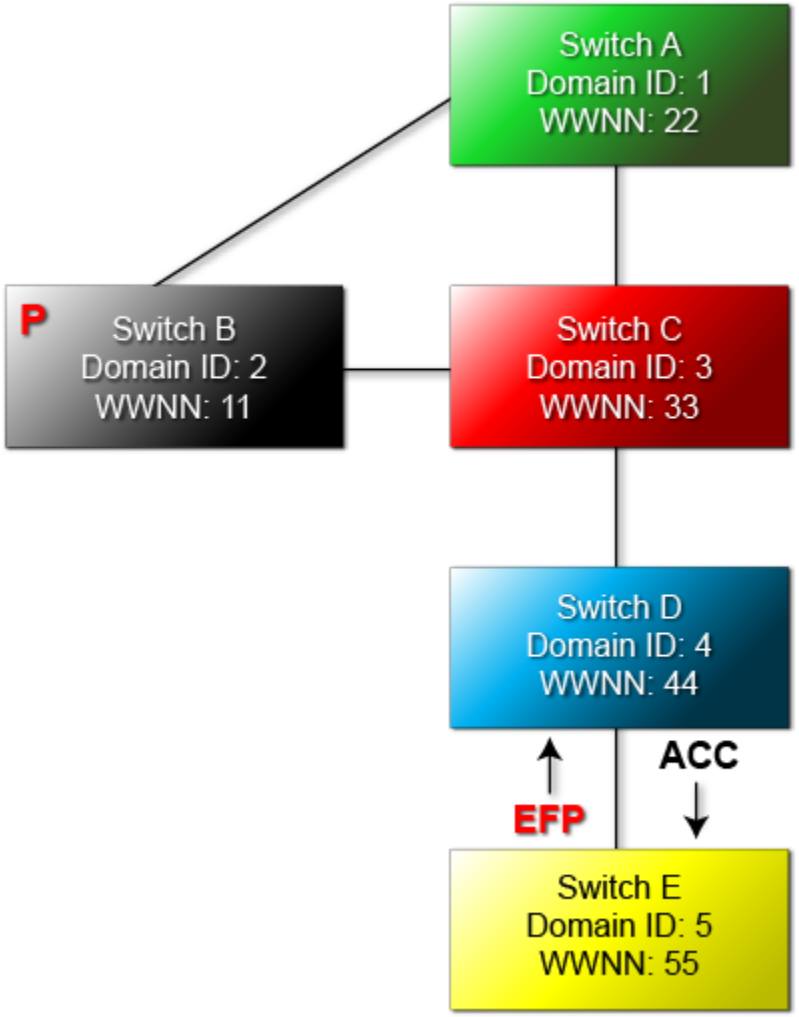
# Fabric Initialization Process (cont.)

- ESC (Exchange Switch Capabilities)
  - Vendor specific info
  - Virtual Fabric support
- ACC (accept frame)
  - Vendor specific info
  - Virtual Fabric support



# Fabric Initialization Process (cont.)

- EFP (Exchange Fabric Parameters)
  - Principal switch selection
    - Principal switch priority
    - Switch name
    - Domain ID list
  - Vendor specific info
  - Virtual Fabric support
- ACC (accept frame)
  - Principal switch selection
    - Principal switch priority
    - Switch name
    - Domain ID list
  - Vendor specific info
  - Virtual Fabric support



# SAN Topologies and Topology Consideration

# SAN Fabric Topologies

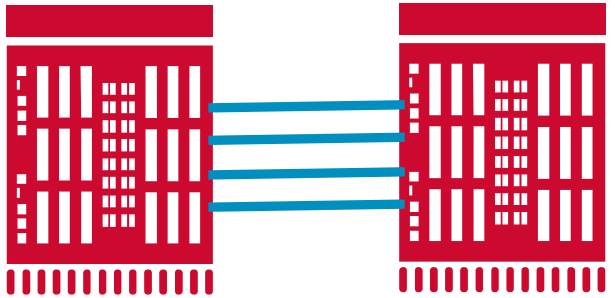
- **Single switch:** simplest SAN solution
  - Pro:
    - Locality between servers and storage
    - Well suited to higher port count switches
  - Con:
    - Scalability limited by switch port count
    - Not resilient and availability restricted to switch and director field replaceable components

➢Recommended solution for small deployments

  - Higher-port count switches and directors make single-switch topologies possible

- **Cascade:** switches/directors connected in series
  - Pro:
    - Lowest port count for ISLs
    - Straightforward way to simplify switch management
    - SAN starts small, stays small, and has high levels of locality
  - Con:
    - Availability and scalability
    - Lower performance with more then two switches/directors in a fabric

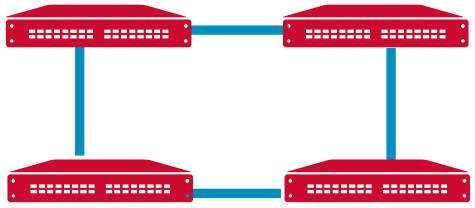
➢Recommended solution for two switch or director fabrics



# SAN Fabric Topologies (cont.)

- **Ring:** a variation on the Cascade topology
  - Pro:
    - Same as cascade but with better performance and availability
  - Con:
    - Poor scalability

➢Recommended solution for three to four switch/director fabrics

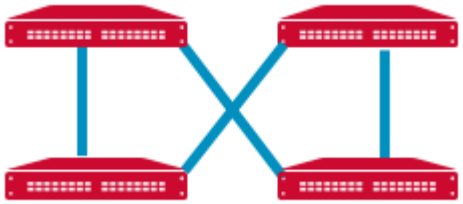


- **Mesh:** connects each switch/director to every other switch/director
  - Pro:
    - Provides greater fabric resiliency than Cascade or Ring
    - Maximum of one hop to any device
  - Con:
    - Does not scale well
    - Lower user port availability due to increased ISL count

➢Recommended for solution with no plans for growth and four switches per fabric
  - Higher-port count switches and directors make mesh-based topologies more scalable

**Full Mesh**

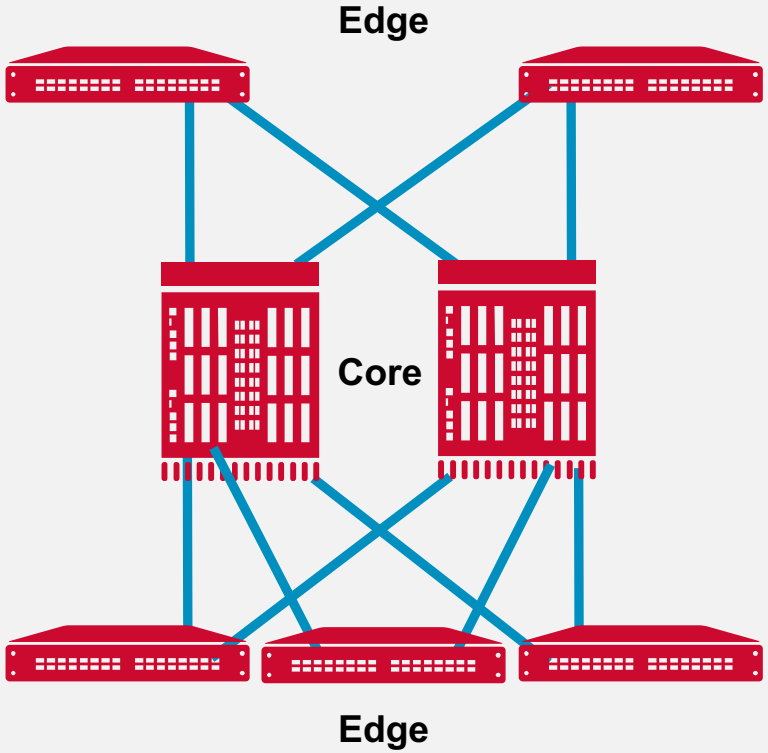
**Partial Mesh**



# SAN Fabric Topologies (cont.)

- **Core/Edge:** specializes the role of the switches and directors
  - Connect servers and storage to switches at the edge
  - Connect switches at the core to the edge switches
  - Hosts and storage may also be attached directly to the core, depending on design requirements
  - Pro:
    - Excellent scalability, availability, and performance
    - Adding a new Edge switch requires connections only to the Core switches, which makes this a highly scalable topology
    - Having two Core switches makes this is a highly resilient topology
  - Con:
    - High cost

➢Recommended for solutions with plans for growth and more than four switches per fabric
  - Provides the best mix of scalability, performance, and availability

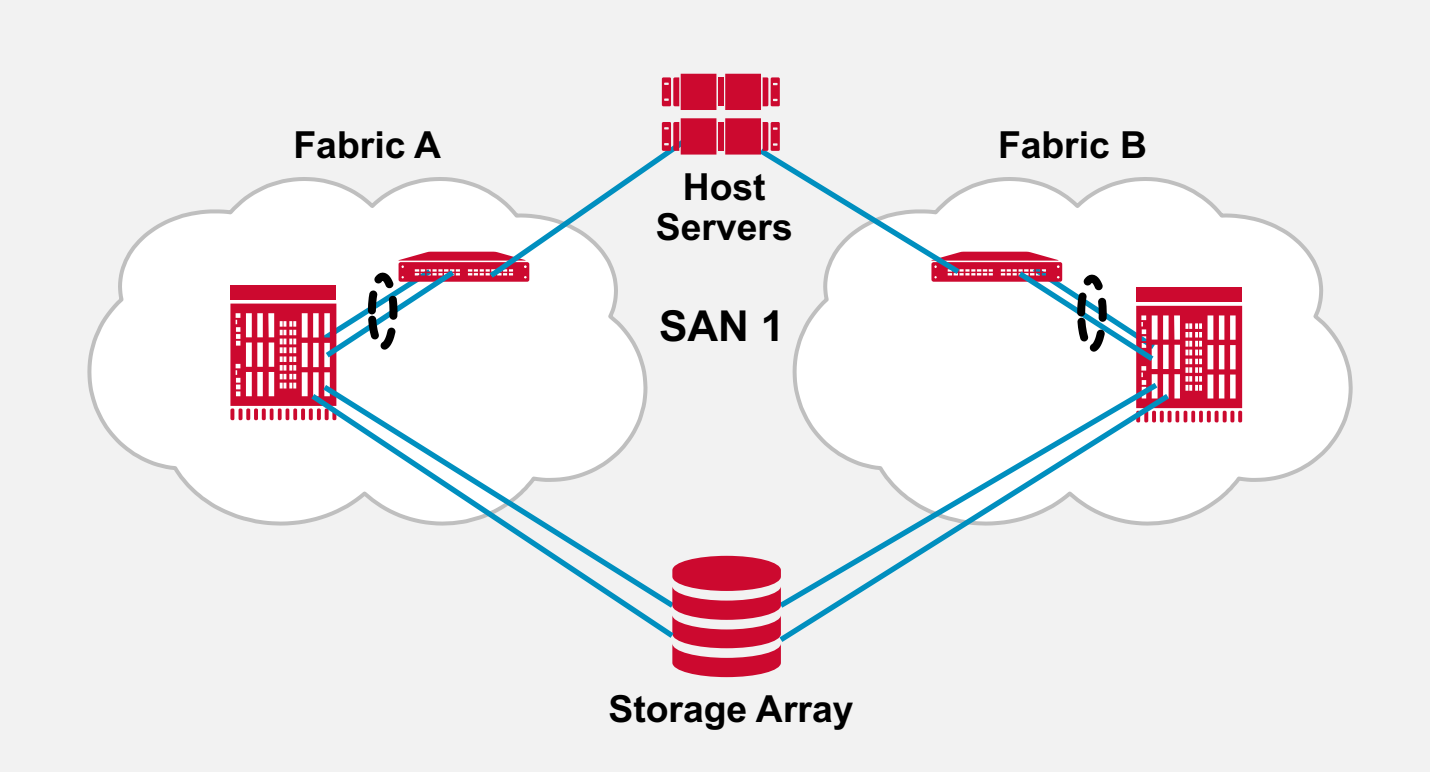





BROCADE
A Broadcom Company

# Fabric Redundancy and Resiliency

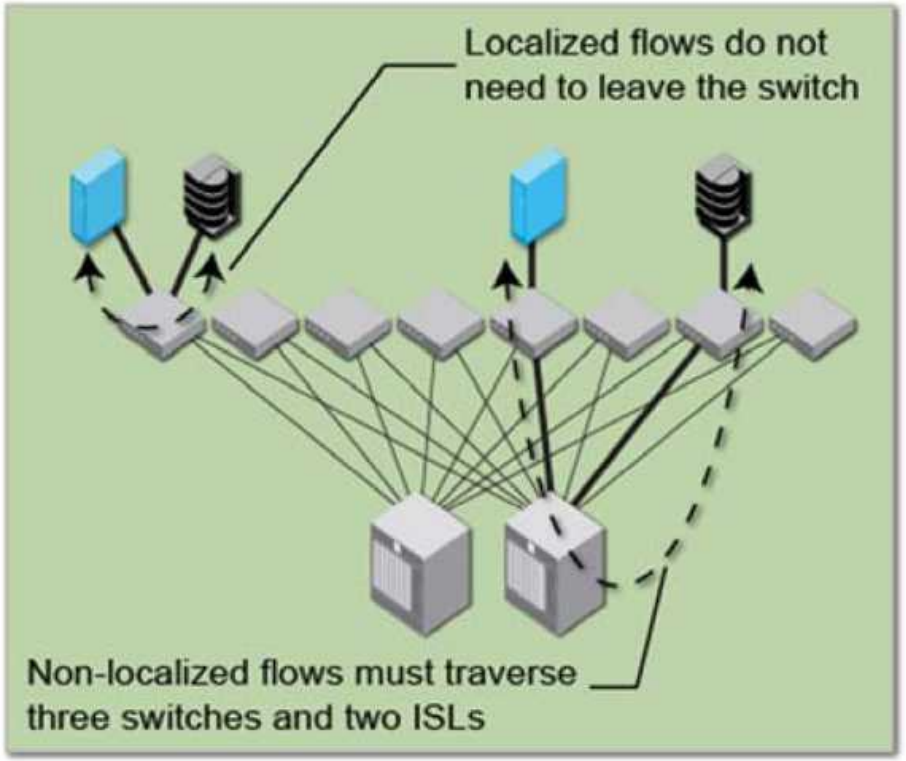
- A dual fabric design provides the highest level of redundancy



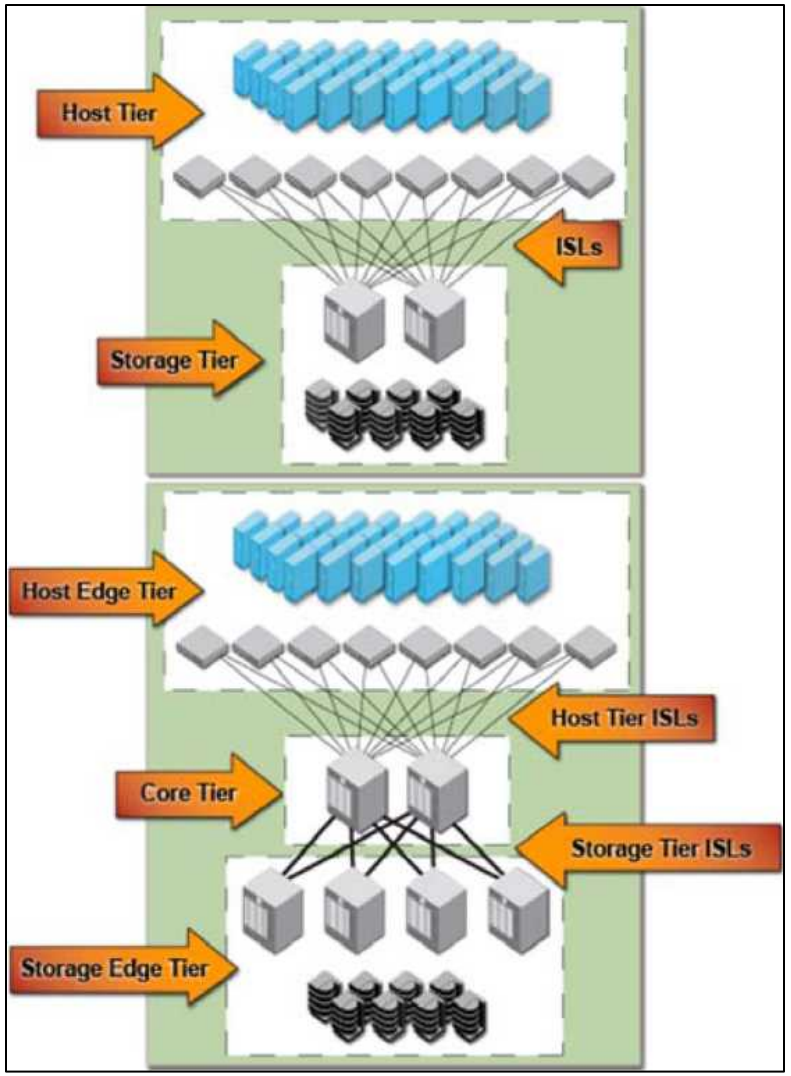



BROCADE
A Broadcom Company

# Localized vs. Tiered Designs




Tiered designs are easier to deploy and manage than localized designs



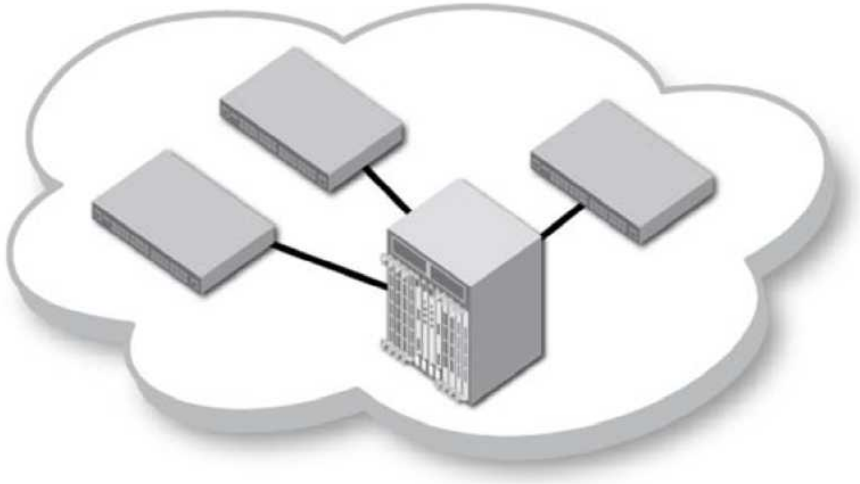
# Availability

BROCADE
A Broadcom Company

# Availability Levels

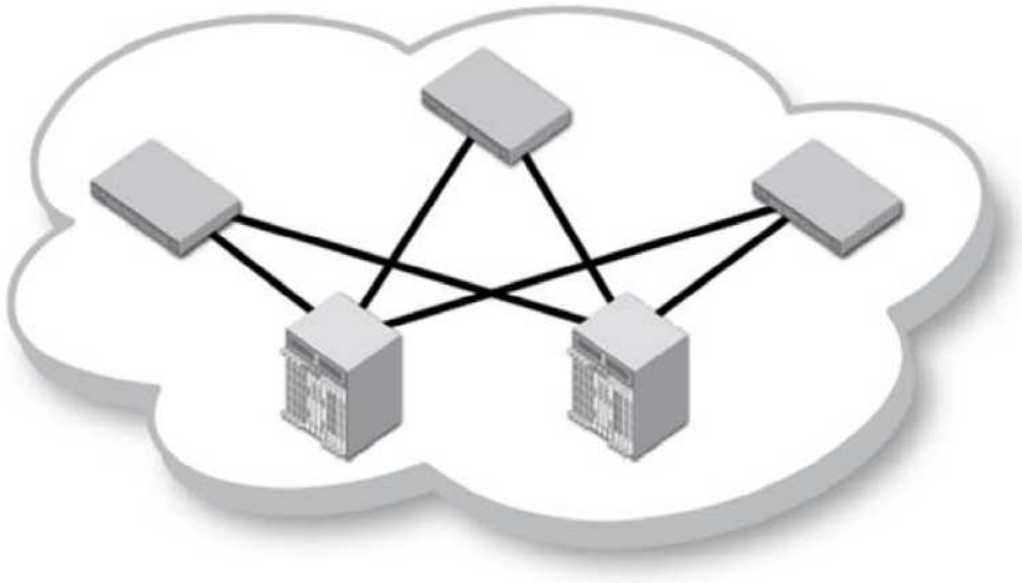
There are four basic levels of availability in a Data Center:

1. Single fabric, non-resilient: **Not Highly Available**
   - Fabric and Core are both single points of failure (SPOFs)



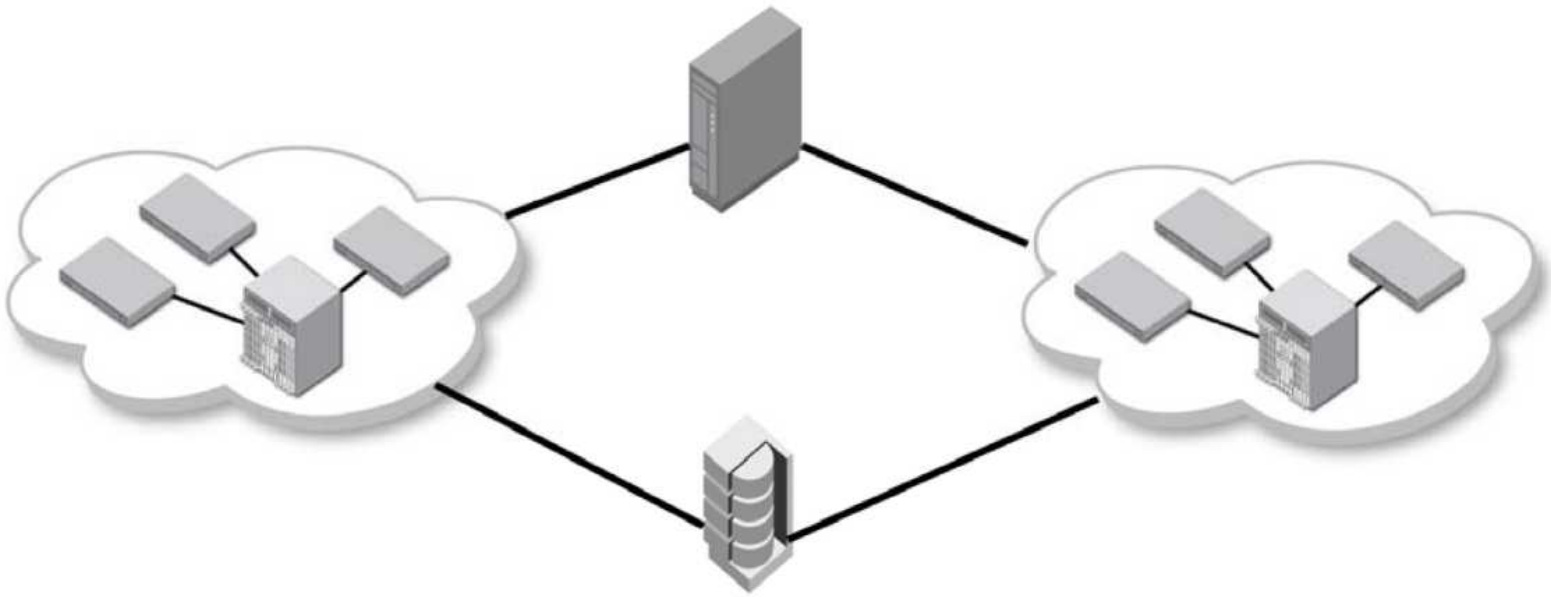
# Availability Levels (cont.)

2. Single fabric, resilient: **Not Highly Available**
   - Core is no longer an SPOF, but the fabric is an SPOF



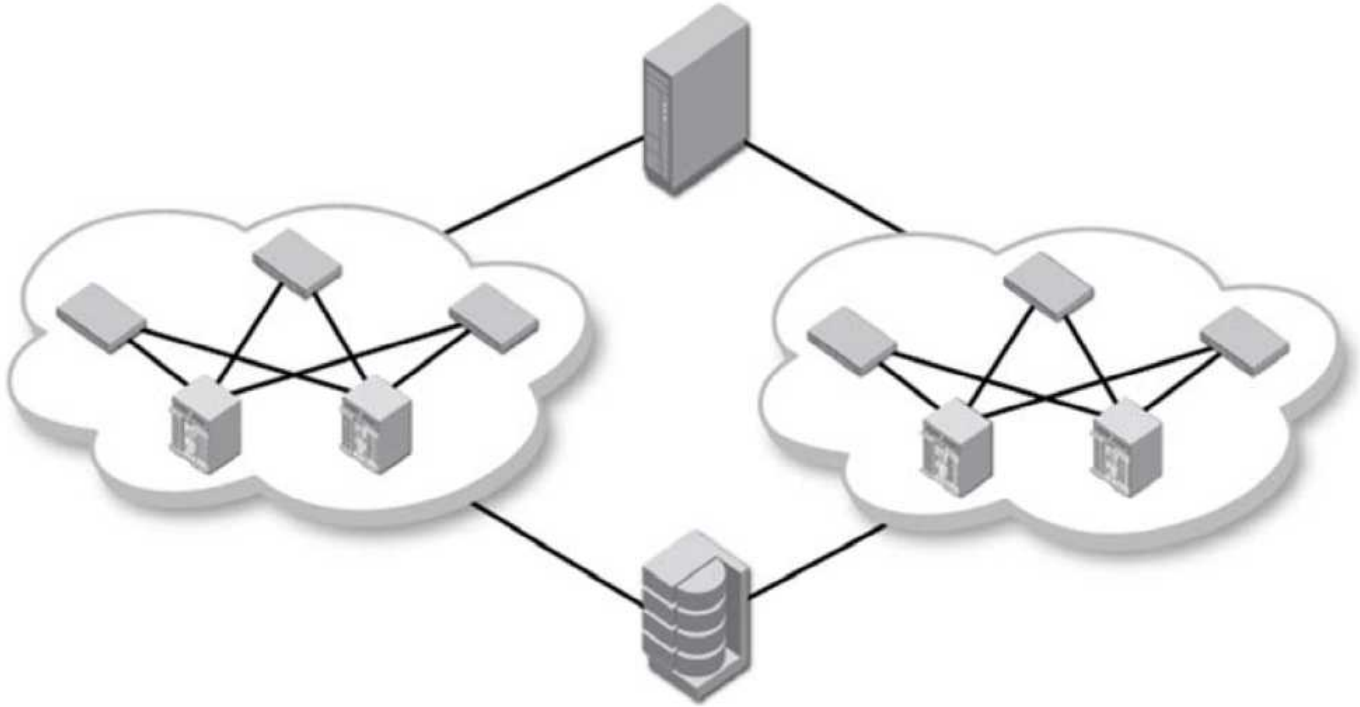
# Availability Levels (cont.)

3. Two fabrics, non-resilient: **Highly Available**
   (better if a Director is the Core)
   - No SPOF - not the best HA level



# Availability Levels (cont.)

4. Two fabrics, resilient: **Highest Availability**
   - No SPOFs; highest HA level



# How Does An Operating System Address Storage?

## Classic Device Hierarchy

- In Unix, physical devices are mapped to a device path
- For example, a disk (LUN) might be: c2t0l5 (or for Solaris – c2t0d5 – the d is for ‘disk’)

Controller 2 (HBA 2), Target 2 (SCSI Target 2), LUN 5 (Logical Unit Number 5 – think ‘partition’)

# How Does An Operating System Address Storage?

## Classic Device Hierarchy

- In Unix, physical devices are mapped to a device path
- For example, a disk (LUN) might be: c2t0l5

Controller 2 (HBA 2), Target 2 (SCSI Target 2), LUN 5 (Logical Unit Number 5 – think 'partition')

However, there is a very important problem that must be overcome:

- If the server "sees" a LUN down two separate paths, it will think there are two LUNs instead of one.
- **Multipath I/O drivers** MUST be installed on the host to correct this "double-vision" to avoid corruption.
- Most modern O/S's have MPIO by default.

# Zoning

# SAN Basics
## LAN / SAN Comparison

### LAN

- Nodes communicate with nodes.

### SAN

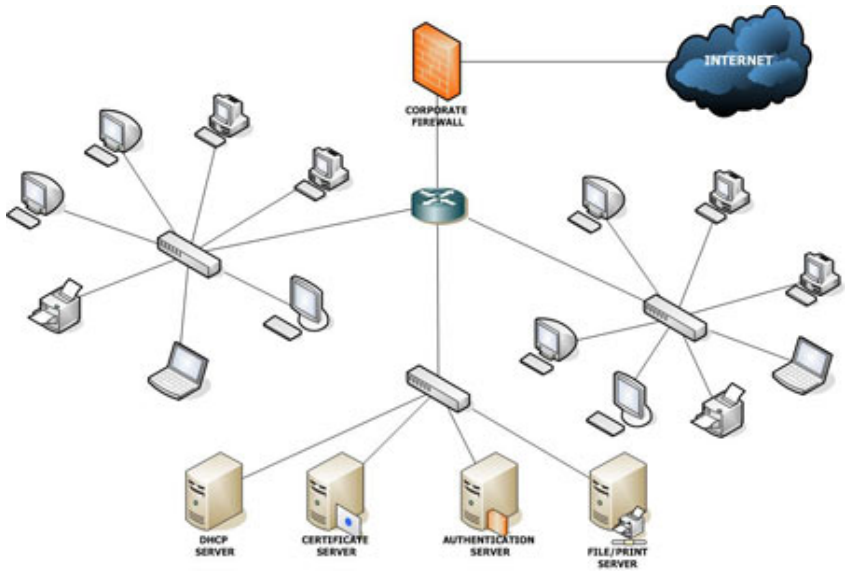
- Nodes distinctly categorized into two groups:
  - **Host (initiator)**
  - **Storage (target)**
- Hosts do not communicate with other hosts on a SAN…they only communicate with storage targets.

# SAN Basics

## LAN / SAN Comparison

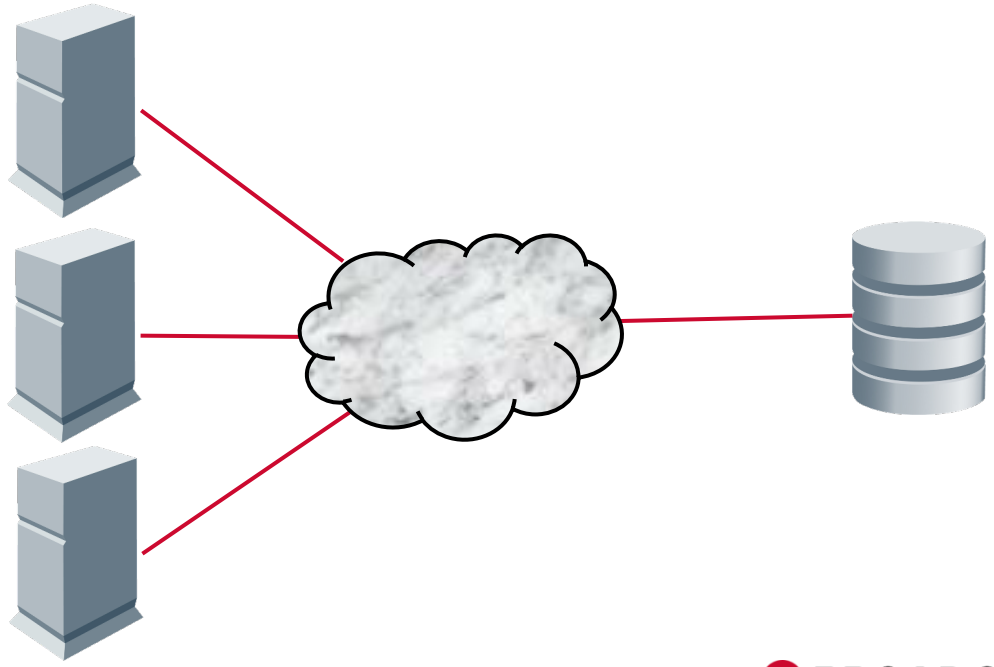
### LAN

- Networks provide ***any-to-any*** connectivity




### SAN

- Networks connect ***many-to-few.***



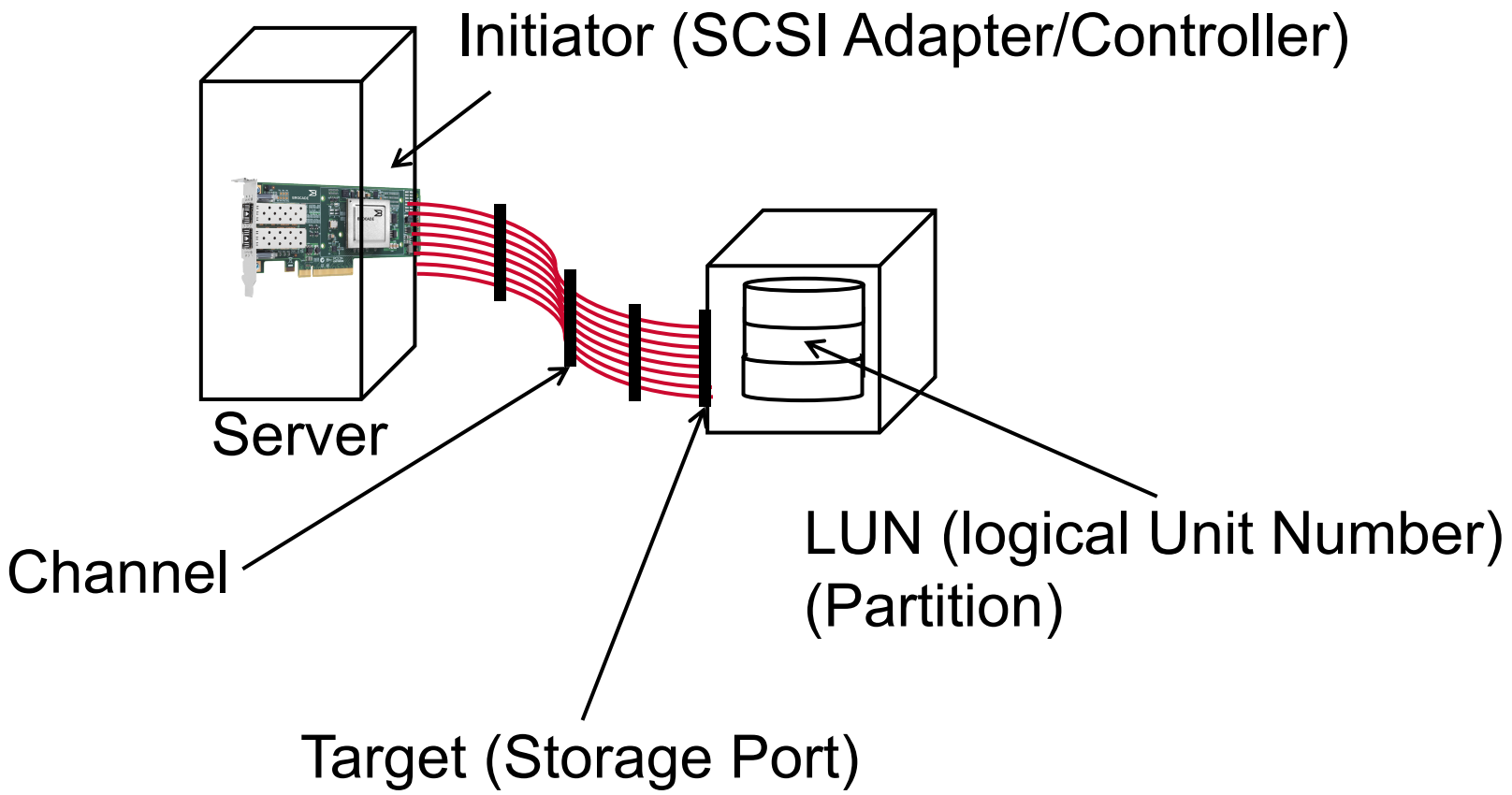



BROADCOM

# SAN

## The Basics

### SCSI Terminology







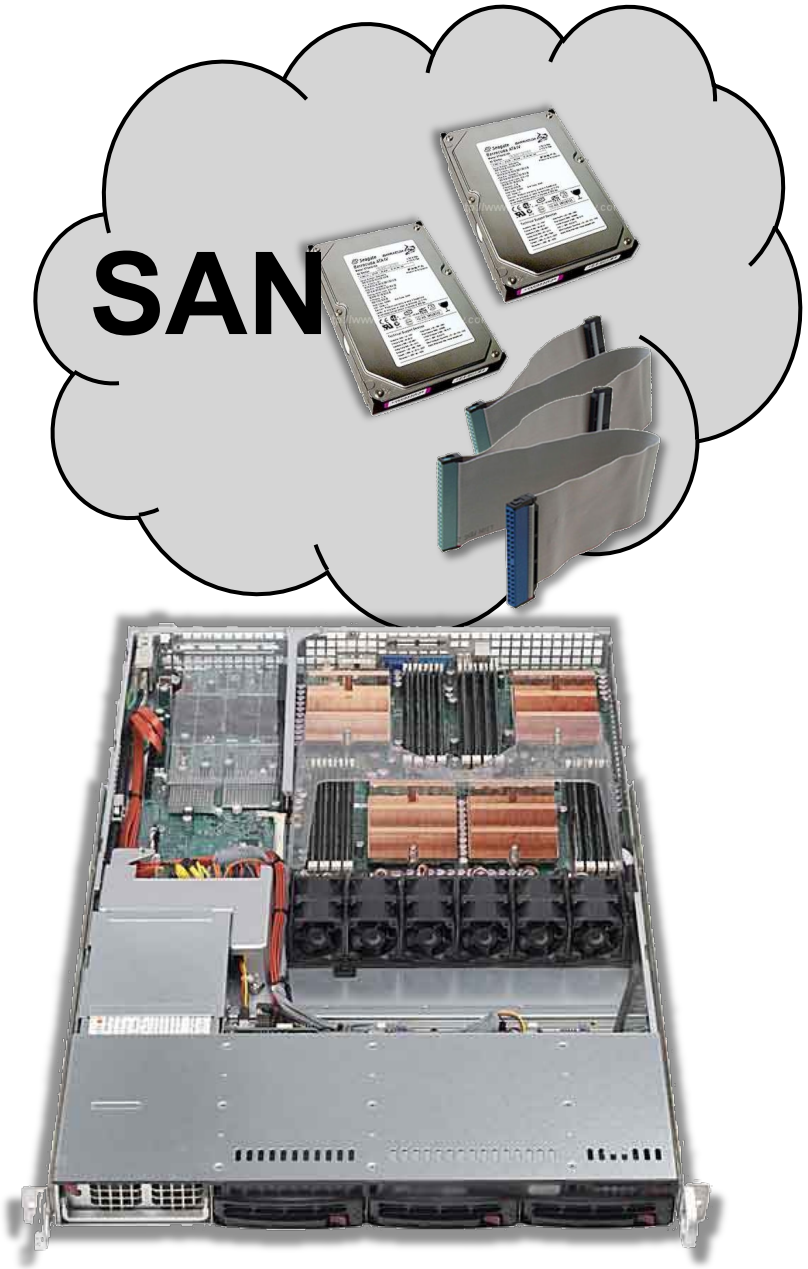
BROADCOM

# SAN

## The Basics

**SAN**

- Important zoning and masking tools in the SAN and target systems make certain that each host only sees what it thinks is a simple SCSI channel with a small number of attached drives (LUNs).





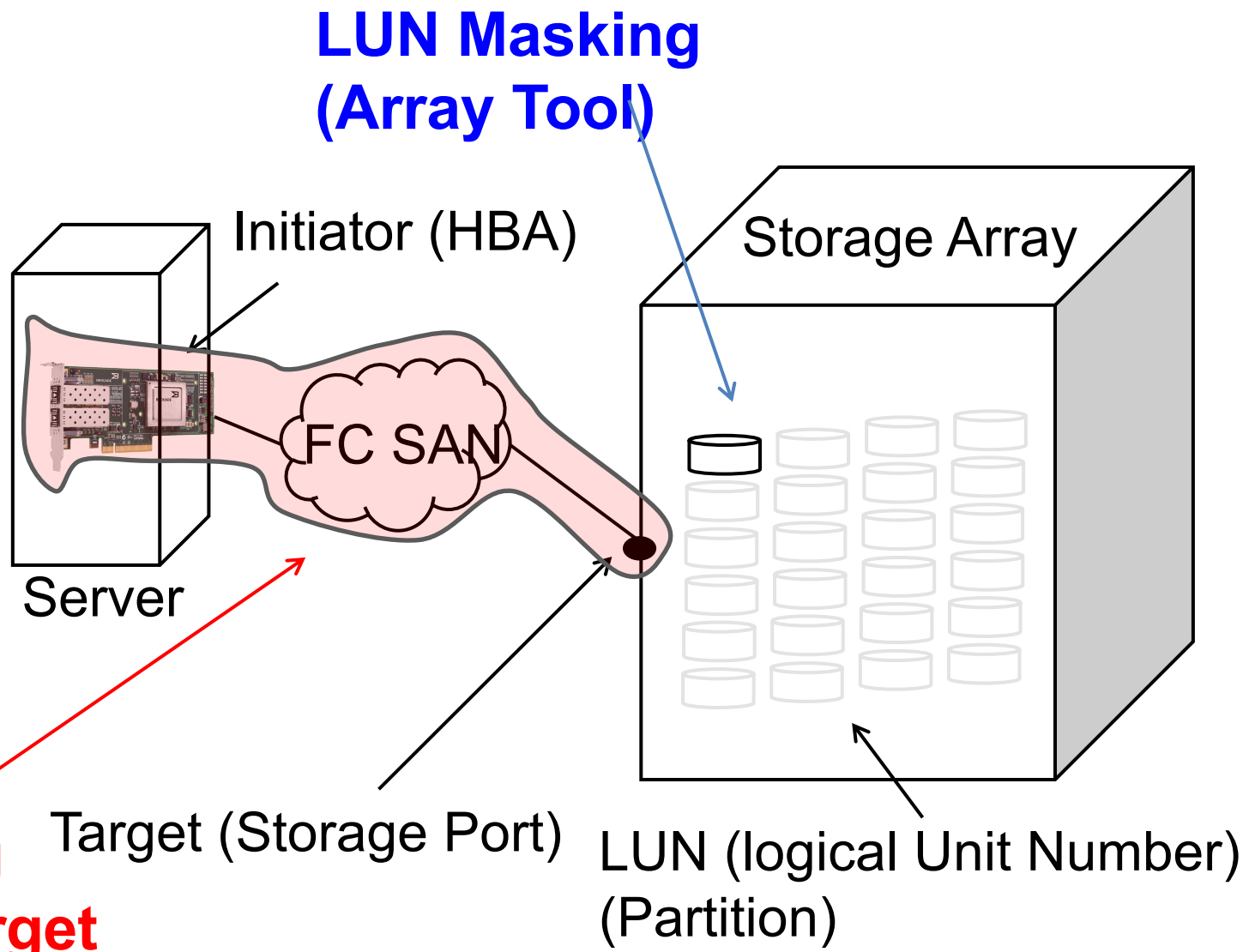


BROADCOM

# SAN

## The Basics

How does it look on a SAN?







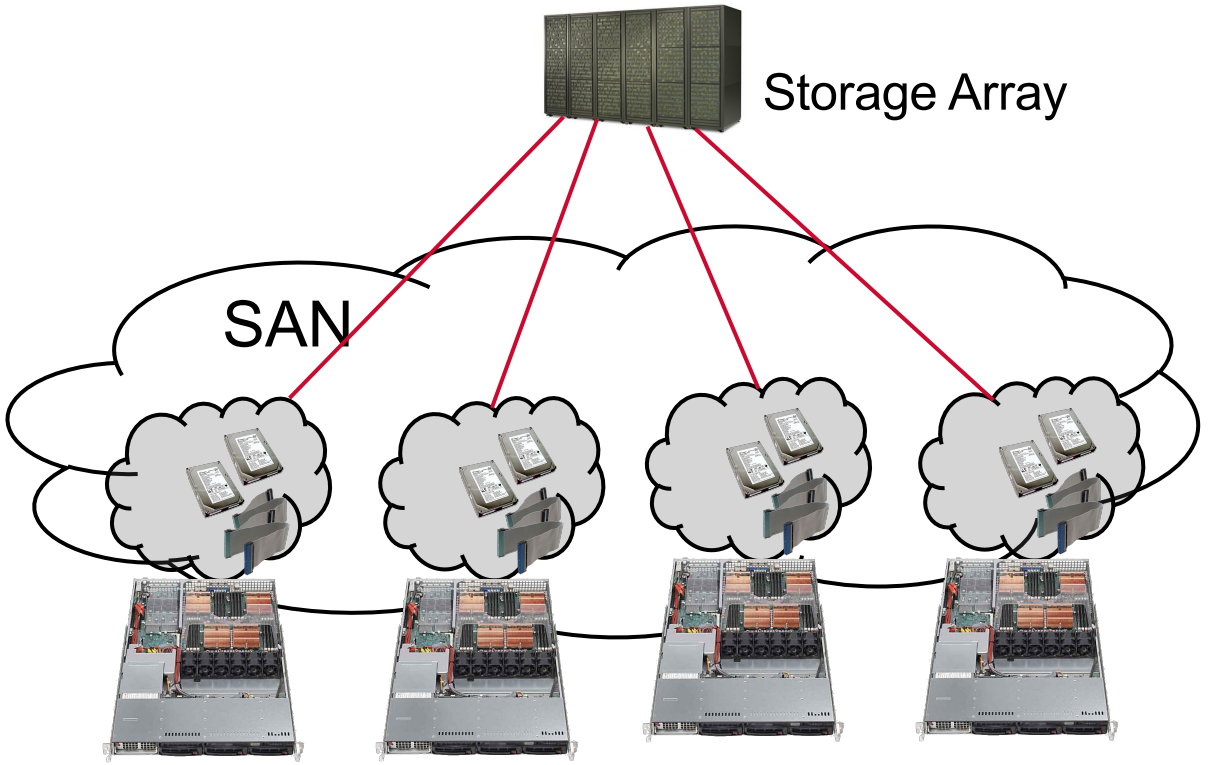
BROADCOM

# SAN
## The Basics

### SAN

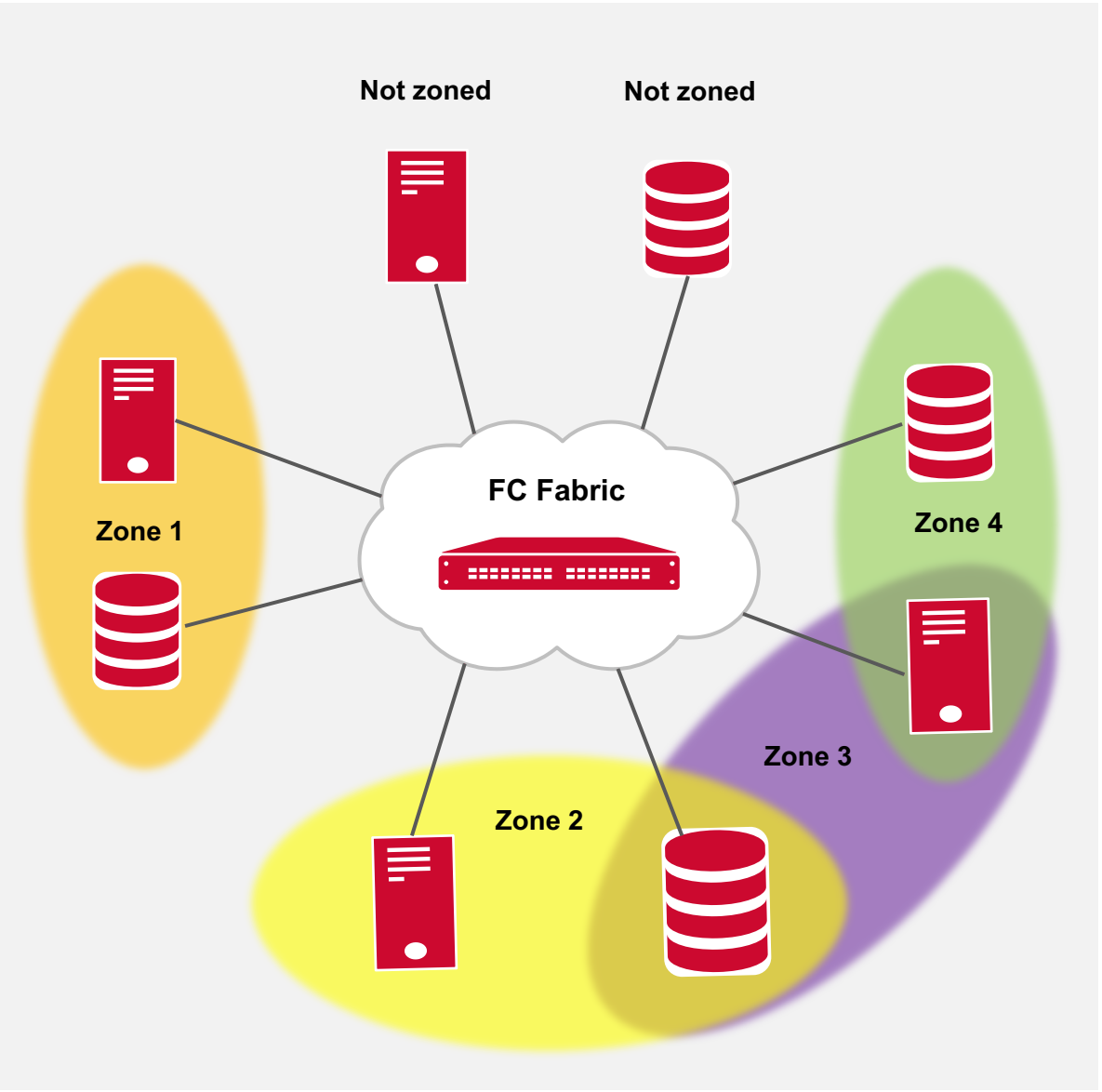
– Although a SAN may have many hundreds of initiators and targets, each host must only see its own storage, not other hosts or other systems' storage.

  – An exception is when certain server clustering tools are being use. In this case multiple servers may see the same storage pool.



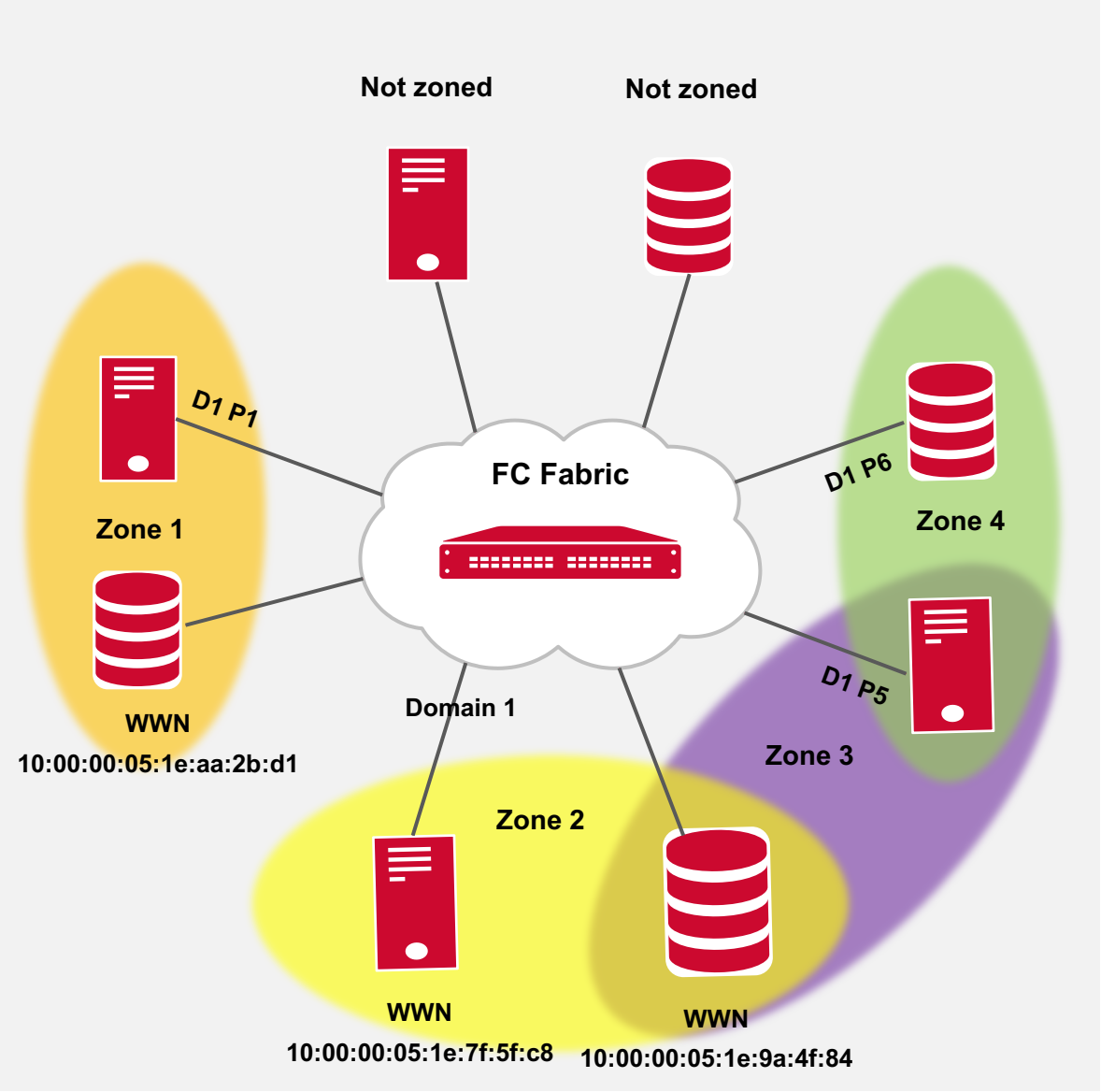
# Zoning

- Enables the partitioning of devices attached to a fabric into logical groups called zones
- When zoning is enabled:
  - Devices defined in the same zone are restricted to communicate only with devices in that zone
  - Devices that are not zoned are isolated from other devices in the fabric
  - A devices can be a member of multiple zones



# Zone Membership

- There are two ways to define a device as a member of a zone
  - WWN
    - Both Port WWN (WWPN) and Node WWN (WWNN) are supported
  - Domain/Port

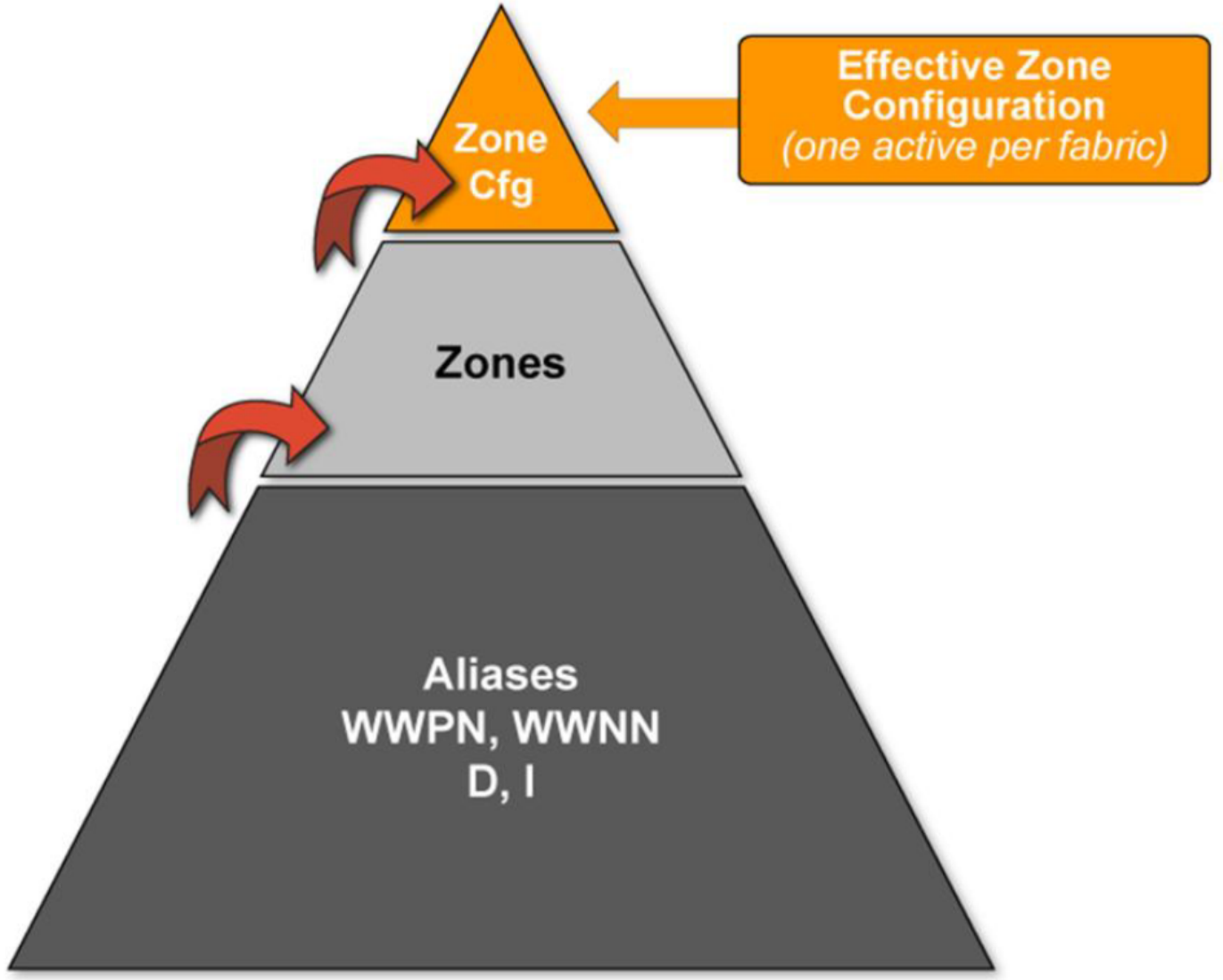





BROCADE
A Broadcom Company

# Zoning Brocade Fabrics

## Hierarchy of Objects




## Zoning can be managed with

- Command Line Interface (CLI)
- Web Tools
- SANnav



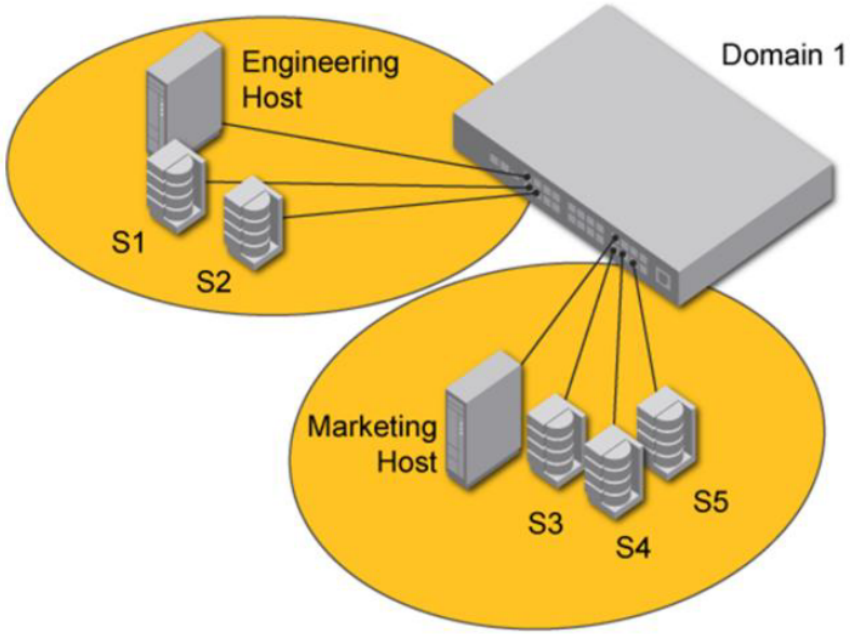

BROADCOM

# Zone Aliases

- The use of aliases is optional but aids in the management of the zoning structure and content:
  - Naming
    - Must begin with an alpha character
    - Can include numeric and underscore characters
    - Up to 64 characters
    - Case sensitive (DISK1 and Disk1 are unique names)
  - Members
    - Domain, Index
    - Node World Wide Name - from nsshow
    - Port World Wide Name - from nsshow, portloginshow or switchshow
  - Sample naming convention
    - Eng_Host1, Eng_Disk1, Eng_Disk2, Mkt_Host1, Mkt_Disk1, Mkt_Disk2
    - Zone_Eng, Zone_Mkt

# Zoning sequence

1. Plan zoning scheme to meet your objectives
2. Create aliases
3. Create zones
4. Create configuration
5. Enable configuration



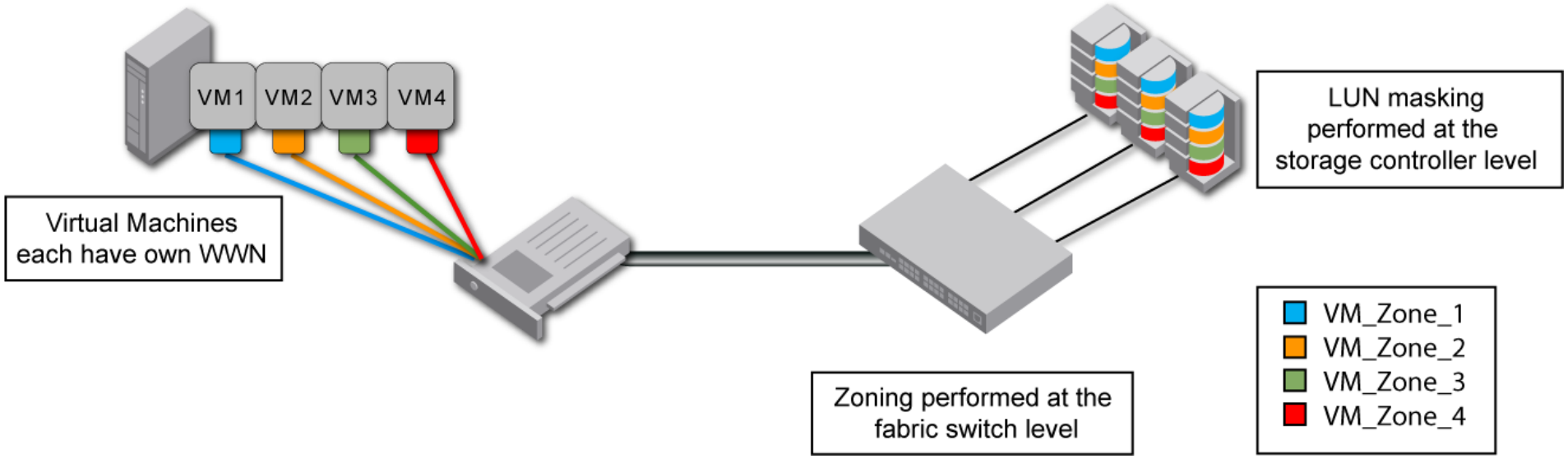
# Zoning Commands

| | Create | Delete | Add | Remove | Show |
|---|---|---|---|---|---|
| **Alias** | `aliCreate` | `aliDelete` | `aliAdd` | `aliRemove` | `aliShow` |
| **Zone** | `zoneCreate` | `zoneDelete` | `zoneAdd` | `zoneRemove` | `zoneShow` |
| **Config** | `cfgCreate` | `cfgDelete` | `cfgAdd` | `cfgRemove` | `cfgShow` |

- Use the `zonehelp` command to display help information
- Zoning has four more commands:
  - `cfgEnable`
  - `cfgDisable`
  - `cfgSave`
  - `cfgClear`

# NPIV Zoning

- Standard fabric zoning and storage LUN masking can be used with virtual machines to isolate storage ports and LUNs to the appropriate virtual server just as they are with physical servers
- To perform zoning to the granularity of the virtual N_Port IDs, WWN-based zoning must be used



# Implementation Considerations

- Define all members in a zone with <domain, index>
  - Provides frame-based hardware enforcement
  - Allows devices to communicate that are connected to the ports defined within the zone
  - Requires a zoning change if a device is moved to a port outside the zone
  - No zoning change if the device's WWN changes
- Define all members in a zone with their device WWN
  - Provides hardware enforcement
  - Allows devices to communicate that have their WWN in the same zone
  - Requires a zoning change if the device's WWN changes
  - No zoning change if a device is moved to another port in the fabric

# Zoning Best Practices

- Always implement zoning, even if LUN masking is used
- Use PWWN identification for all zoning configuration unless special circumstances require domain-index identification (for example, FICON)
- Make zone object names only as long as they need to be to be meaningful
- Define all zones so that frame-based enforcement is used
- Use single initiator zoning
- Use separate zones for tape and disk traffic if an HBA is carrying both types of traffic
- Implement `defzone --noaccess`

# EOF


**Спасибо!**

Узнайте больше:

www.brocade.com
http://www.linkedin.com/groups?gid=4246353
https://t.me/BrocadeRussiaSAN

BROADCOM®