

FCoE Commentary

EMC/HP/IBM

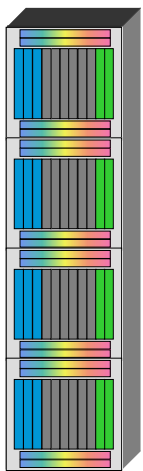
T11/07-408v0

July 13, 2007

Outline

- Background
 - Server rack density
- Frame format
 - FCoE header length field
 - Preventing OX_ID reuse
- Virtualization related items
 - (Virtual) Fabric to VLAN relationship
 - FC Zoning for FCoE
 - Direct N_Port to N_Port FCoE traffic
- Topology issues
 - Forwarding Loops
 - Failover
 - Management
- FLOGI and Security topics

Server Resource Density / Growth



Volume servers purchased by the rack – blade moving to volume
4+ blade enclosures / rack

8-16 blades / enclosure

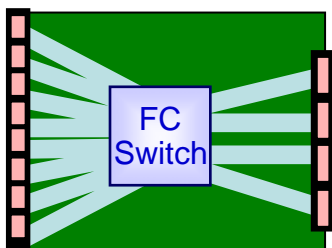
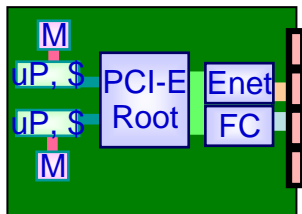
16-32 processors / blade

2 today -> 4 2007 -> 8 cores / processor 2010+

2x cores every 2-3 years

A 4 enclosure rack in 2010 with 1024 cores

4 Guest OS per core = 4096 guests per rack



Multiple switch fabric types per enclosure

Ethernet, Fibre Channel, InfiniBand, etc.

Multiple switch fabrics instances per enclosure

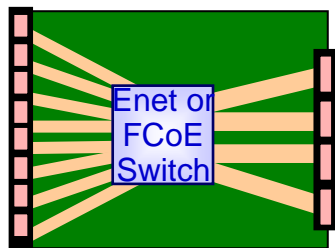
4-8 switch modules in range of topologies

Full mesh, full redundancy, etc.

Minimum 1 switch hop within enclosure

Minimum 1-2 switch hops per rack to fabric attached storage

At least one unique fabric address per fabric port per OS guest



Motivation for a single converged I/O fabric

FCoE Header – Length Field

- One standard FCoE frame format is required
- Need mechanism to identify padding when used
 - Use a length field, or
 - Use fixed padding, or
 - Use variable sized and/or not-always-present padding
- No length field enables cut-through or speed matching (vs store-and-forward) gateways
 - Only relevant for specific windows in time when FC and Ethernet link speeds are close.
- We prefer the flexibility provided by not having a length field

OX_ID reuse: Timestamp concerns

- Native FC N_Ports don't use timestamps
 - Native FC timestamp usage is switch-based
 - FCoE N_Port time synchronization - out-of-band wrt FC
 - SNTP or equivalent provides new ways to break FC
 - Synchronization scale increased by 10x-100x over current FC
- What timeout value is used for FCoE N_Port check?
 - Just use R_A_TOV: FCoE F_Port has to lie.
 - Q: What if only 3 sec of 10 sec R_A_TOV are left for FCoE?
 - A: FCoE F_Port timestamp is 7 sec before actual time, ouch!
 - New value (e.g., 3 sec): How to send value to FCoE N_Port
 - FLOGI ACCEPT is “right place”, but don't want to modify that ELS
- Where did the new value come from?
 - Site-specific (fabric-specific) time-budgeting exercise (OUCH!!)

OX_ID reuse: Increase R_A_TOV?

- Some Ethernet switches can match FC 0.5 sec max hold time
 - Allows use of current 10 sec R_A_TOV
 - Requires vendor-specific configuration of Ethernet switches
 - IEEE concern: (per-priority) PAUSE interaction with max hold time?
- Alternative: R_A_TOV increase
 - 30 seconds is a common OS timeout, R_A_TOV can't be larger
 - Increasing R_A_TOV can increase failover times
 - 8 hops @ 1 sec max hold time => 18 sec R_A_TOV
 - $(R_A_TOV - E_D_TOV [2 \text{ sec}]) / (2 * \text{hop count}) = \text{switch max hold time}$
 - Issue: How to verify switch max hold time values, configuration protocol?
 - Does not appear to risk OX_ID exhaustion
 - OX_ID for completed I/O is immediately reusable
- Bottom line: No FCoE header timestamp needed
 - End-to-end FC fabric R_A_TOV – more robust
 - Increase of R_A_TOV appears possible if necessary
- Prefer to work out design details of R_A_TOV-based solution

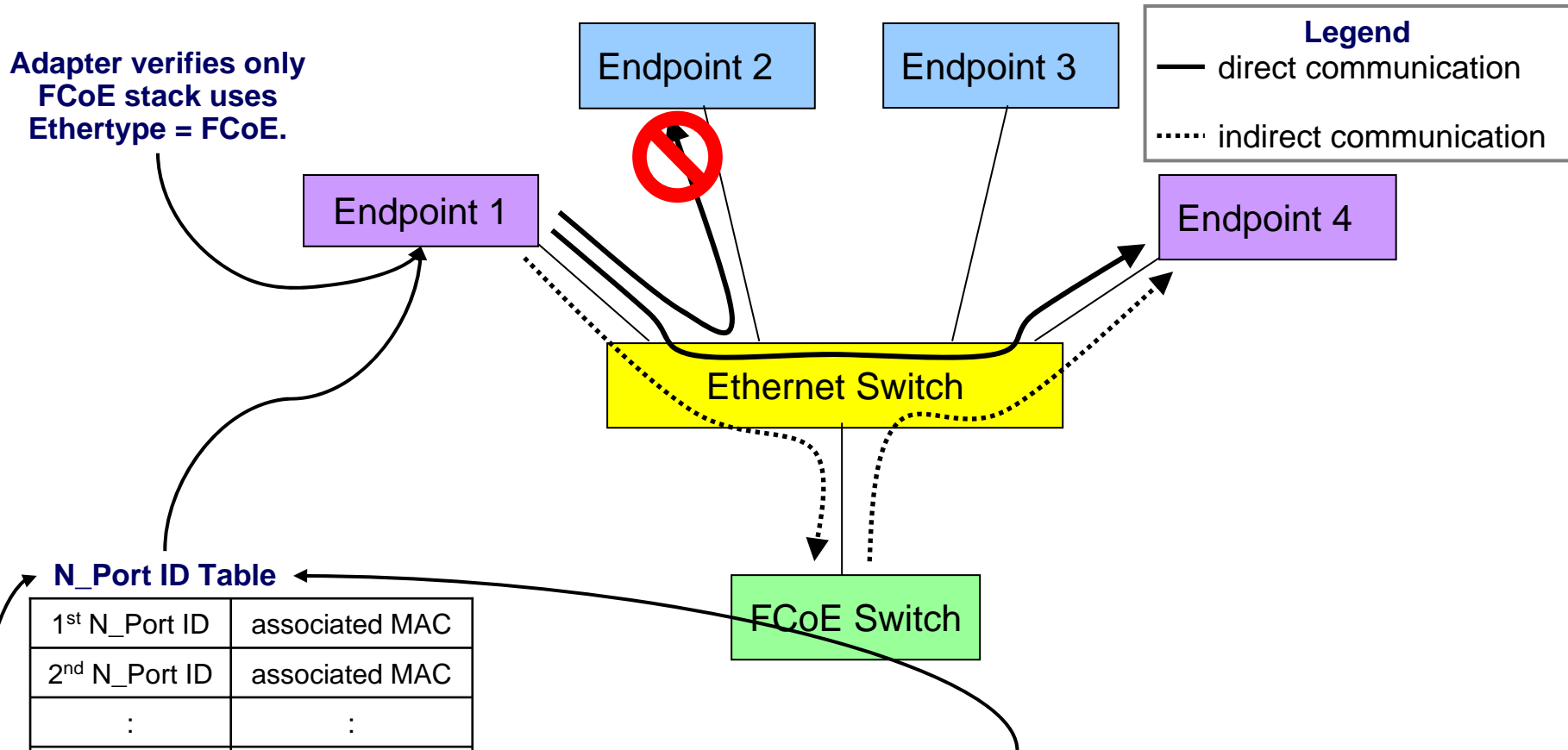
Virtual Fabric to VLAN relationship

- VLAN boundaries should be used to enforce FC (virtual) fabric boundaries (i.e., not zoning)
- Dual redundant FC fabrics are common
 - FCoE must not merge dual redundant FC fabrics!!
 - Same Virtual Fabric identifiers (tags) may be used in both fabrics
 - FCoE may result in single Ethernet infrastructure attached to both fabrics
 - 2 separate logical FC fabrics, single IP network
 - Virtual Fabric tags aren't enough to sort this out!!
- Recommendation: Flexibility is useful
 - Ability to have one (Virtual) Fabric correspond to multiple VLANs is more useful than 1-to-1 restriction

Ethernet and FC Zoning

- VLAN to Fabric mapping: multiple FC zones per VLAN
- Zoning for FCoE traffic should be as robust as native FC
 - Becomes important for native FCoE storage
 - If only HBAs are FCoE, FCoE-to-FC gateway can enforce zoning
- Two options for ACL check: NIC vs Switch
 - Could use Ethernet switch ACLs for Ethernet-only switches
 - ACL can be by MAC or MAC and EtherType
 - Need EtherType if MAC not exclusively used for FCoE traffic
 - HBA-based approach appears preferable (see next slides)
- ACL configuration must be automatable based on zoning information from FC fabric
 - Need standardized ACL interface
 - Probably not SNMP (not good for bulk config)
 - Alternatives: netconf? something new?

HBA Based MAC ACL Checks

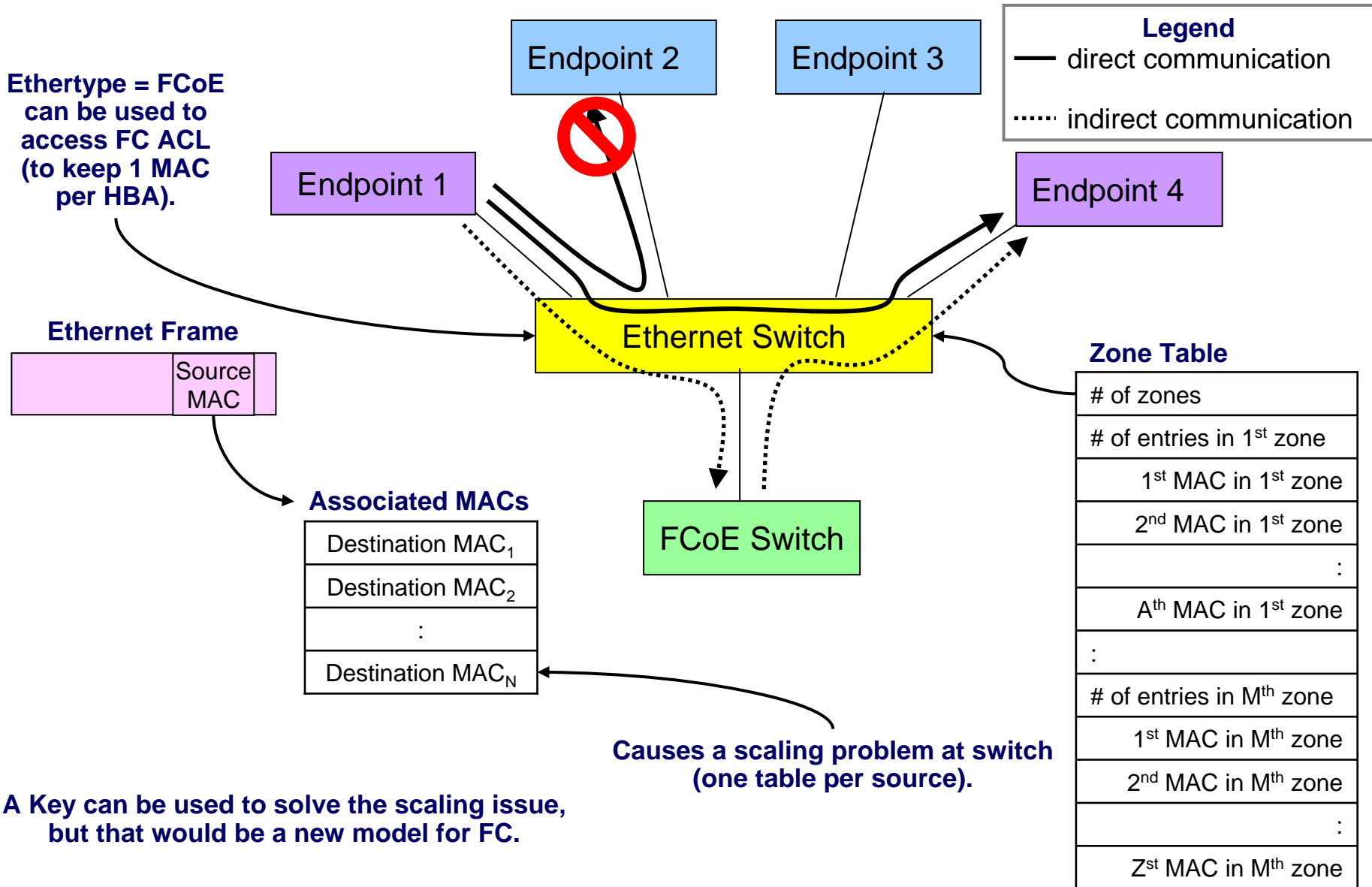


For adapters that support N_Port virtualization, each N_Port ID has a list of the N_Ports & MACs that N_Port ID can access.

Two HBA zone integrity options:

1. Hard endpoint zoning → semantically makes a requirement that HBA must prevent endpoint from manipulating N_Port ID table. Table is loaded by HBA HW as part of FCoE MAC translation service (*more later*). Endpoint can read the table, but not write into it.
2. Soft endpoint zoning → Endpoint is allowed to read/write into the table.

Switch Based MAC ACL Checks



FCoE Zoning Location

- FC Zones more complex than Ethernet ACLs
 - FC Zone to Ethernet ACL mapping requirements will exceed Ethernet switch ACL size limits
 - ACL has no notion of group, hence pairwise explosion
 - Consideration for FCoE N_Port to N_Port Ethernet-only traffic
 - Extensive use of this may require new Ethernet equipment
- Recommendation: Implement zoning at FCoE N_Ports (need to standardize)
 - Hardened implementations of soft endpoint zoning desired
- In all cases, FCoE N_Port must limit FC EtherType to FCoE traffic

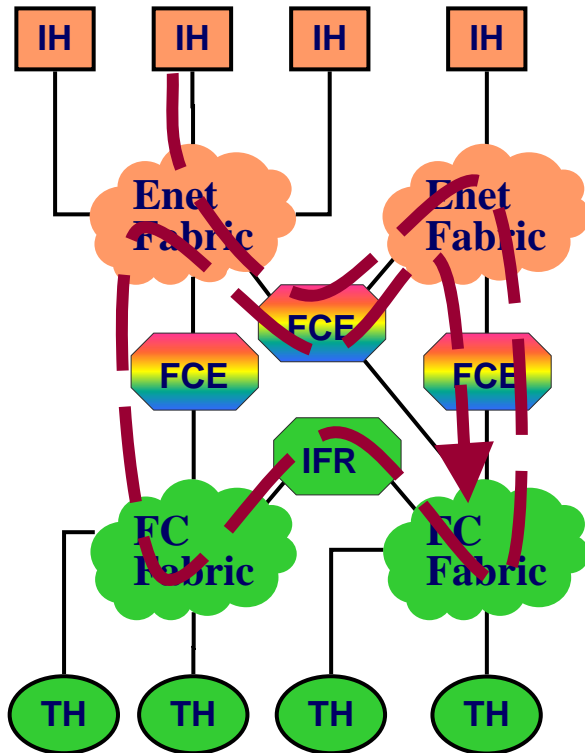
Direct N_Port to N_Port traffic

- FCoE N_Port to N_Port through Ethernet-only switches
 - Need to get permission for this traffic
 - Need to control this traffic (zoning, previous slides)
 - Need to map FCIDs to MAC addresses
- Mapping alternatives (diagrams in backup slides)
 - FCoE changes or adds FC nameserver responses
 - FCoE has new FCID to MAC translation service
 - Direct mapping (drop FCID into FC MAC template)
 - In all cases, have to deal with address response lifetime (caching) and RSCN behavior

Forwarding Loop Prevention

- Ethernet: Spanning tree removes loops
 - TRILL: additional possible improvement, when?
- Fibre Channel: FSPF prevents loops
 - Has to work for FC/FCoE combined fabric
 - Principle: FCoE E_Ports are E_Ports
- FCoE mapping of FC_IDs to MACs must not introduce new loop possibilities
 - Watch out for configuration changes

Forwarding Loop – Livelock



**The frame will live forever.
All scenarios of this sort must be
impossible!**

- Loop spans fabrics
 - Each fabric is loop-free
- Frame lives forever
 - No TTL or hop count
- Cause: misconfiguration
 - Human in the loop
- Mapping coordination is required
 - FCoE: FC_IDs to MACs
 - IFR across FC fabrics
- How is this handled?
 - Where is this specified?

Failover

- Standard must support failover across FC paths
 - I.e., don't break multipathing
- VRRP-like approach to MAC failover: promising
 - Failover of FCoE F_Port across Ethernet ports
 - VRRP = Virtual Router Redundancy Protocol
 - IETF RFC 3768
 - Requires additional work to adapt VRRP to FCoE
- VRRP is transparent to Fibre Channel
 - Transparency is a blessing and a curse
 - Can create commonality across what were supposed to be separate paths
 - See diagrams in backup slides

Management

- SMI-S is preferred standard management approach for Fibre Channel storage
 - How do we manage Ethernet switches in converged I/O environments (storage, networking, IPC)?
 - Existing Ethernet and FC management tools must work in a converged I/O environment.
- Troubleshooting
 - Ethernet and FC troubleshooting - independent by default
 - Minimum: propagate Ethernet problems to FC-visible effects
- Configuration of converged FC/Ethernet environment needs attention – somewhere
 - Network management view of all network traffic, plus
 - Storage management view of all things Fibre Channel
 - ... in a single management system!
 - Moving target due to ongoing Ethernet enhancements

FLOGI and Security Topics

- Need to use multicast for FLOGI
 - Unicast flooding is vulnerable to Ethernet bridges learning the MAC address they're not supposed to know
 - Provides an easy denial-of-service attack
 - Need to ensure only one response to any FLOGI
 - Should standardize at least one means of doing this
- Need to specify FCoE-to-FC gateway defenses against malicious FC frames
 - Much easier to put attack traffic on Ethernet than FC

Summary of Recommendations (1)

- FCoE header length field
 - Having no length field is the most flexible approach
- Preventing OX_ID reuse
 - R_A_TOV-based approach instead of timestamps
- (Virtual) Fabric to VLAN relationship
 - We prefer ability to map multiple VLANs per fabric
- FC Zoning for FCoE
 - FCoE must not reduce FC Zoning robustness
 - N_Port approach preferable to Ethernet switch ACLs
- Direct N_Port to N_Port FCoE traffic
 - Requires FCID to MAC mapping of some form
 - Multiple ways to do this, no recommendation yet

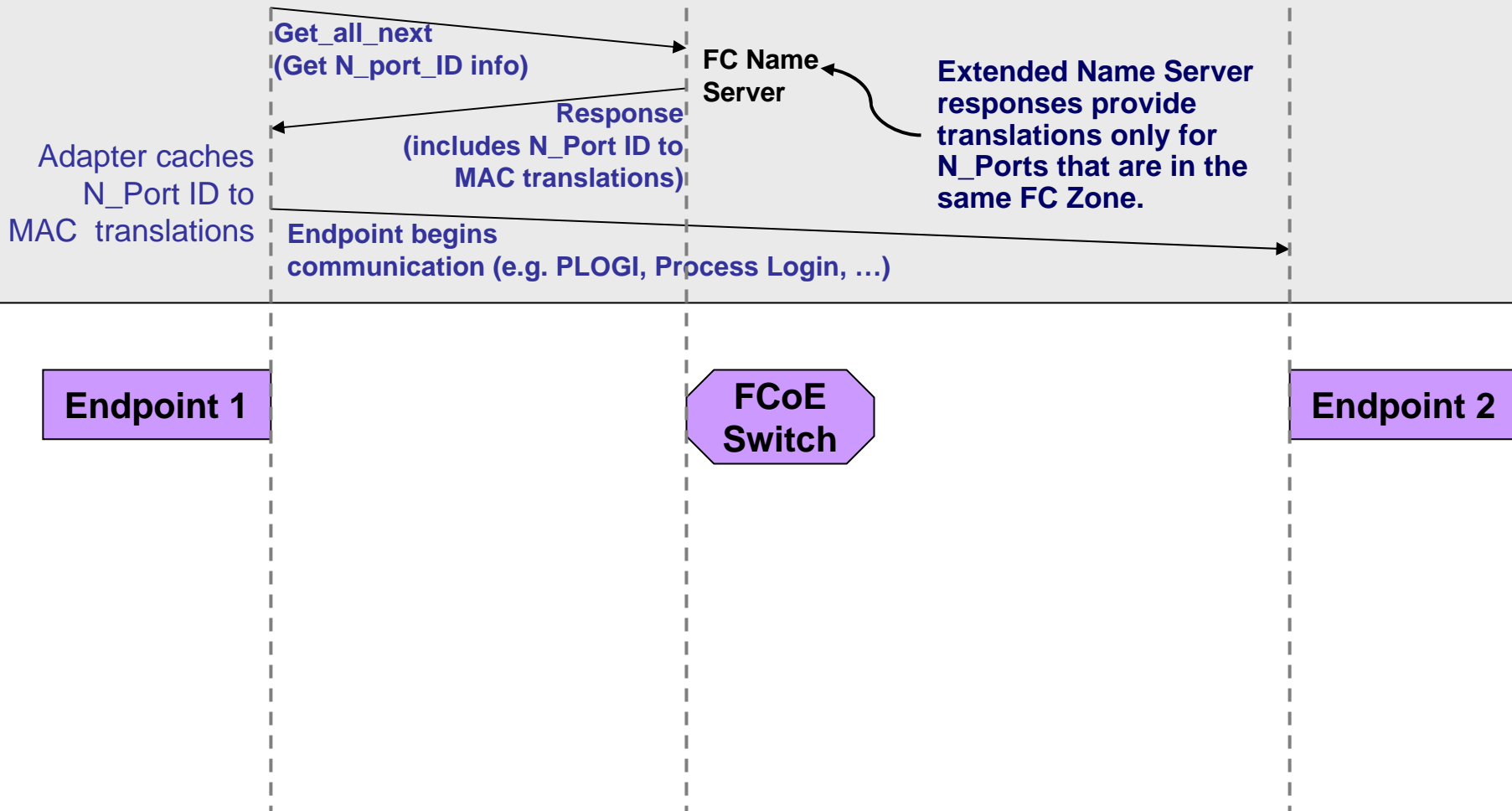
Summary of Recommendations (2)

- Forwarding Loops
 - Forwarding loops must be prevented
 - Configuration and reconfiguration details matter
- Failover
 - Don't break FC multipathing
 - VRRP is promising for FCoE F_Port failover
 - Must manage interactions with FC multipathing
 - Additional work needed to adapt VRRP to FCoE.
- Management
 - Lots of work needed here
- FLOGI and Security topics
 - Use multicast for FLOGI, not flooded unicast
 - Standardize at least one way of coordinating F_Ports
 - Specify FC fabric defenses against bad FCoE frames

Backup

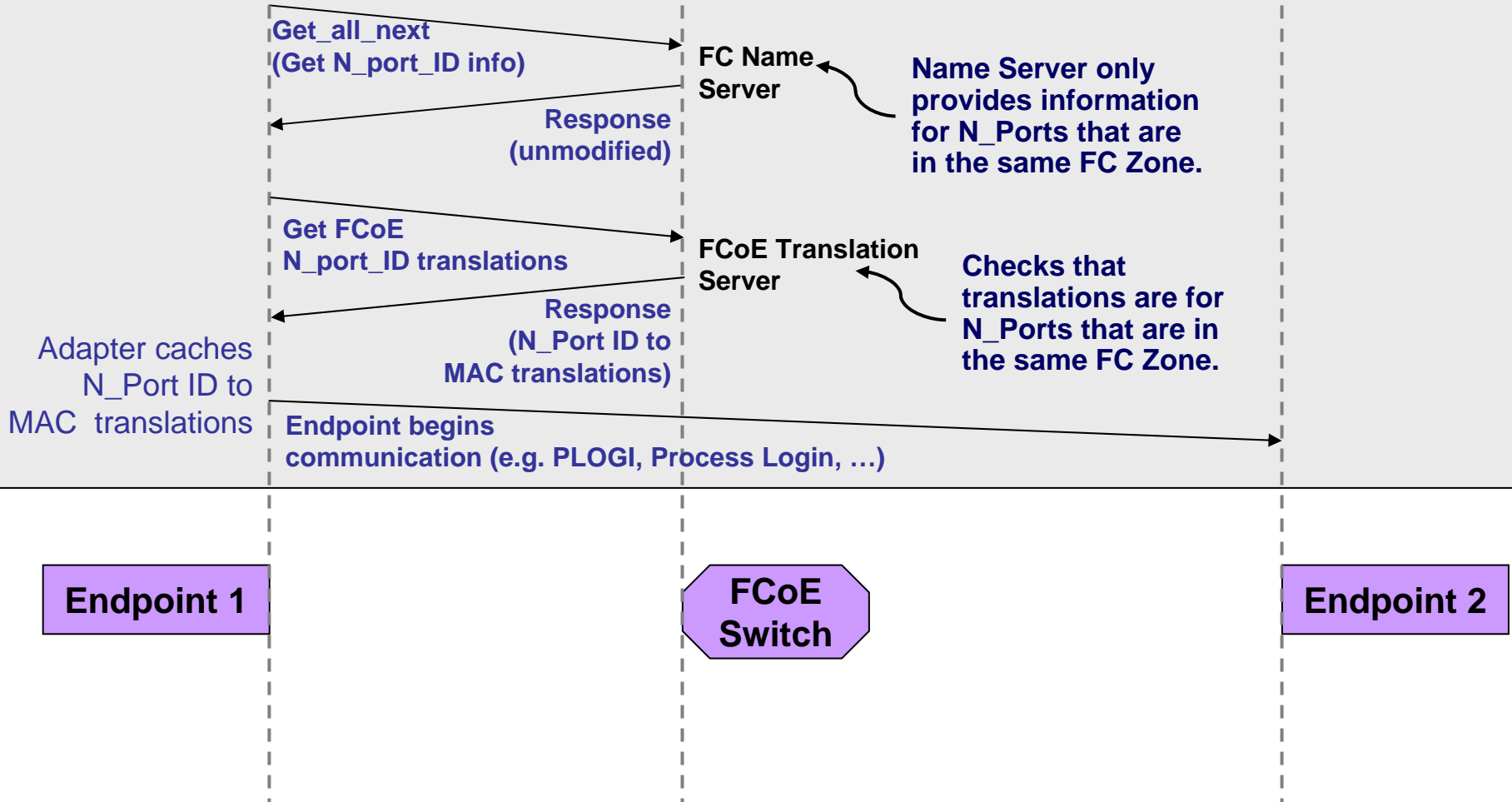
Extended FC Nameserver Option

FC Name Server responses include FCoE information



FCoE Mapping Service Option

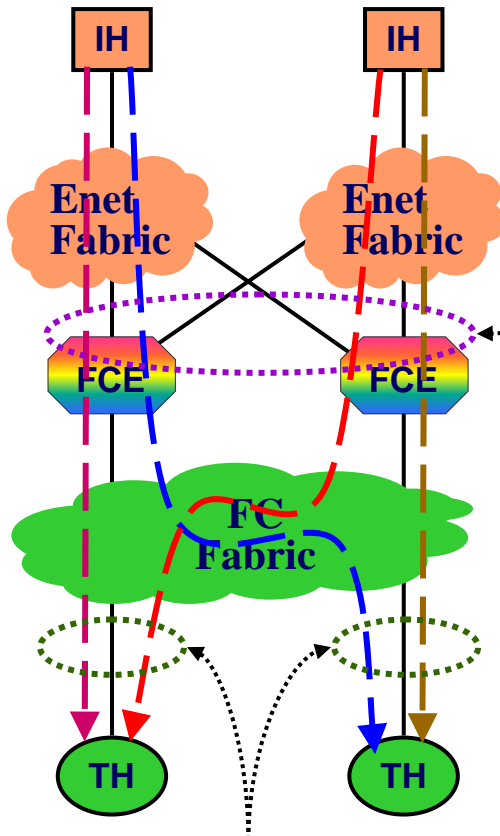
FCoE has new FCID to MAC translation service



Simple failover scenario

One host, one storage array, 2 ports each

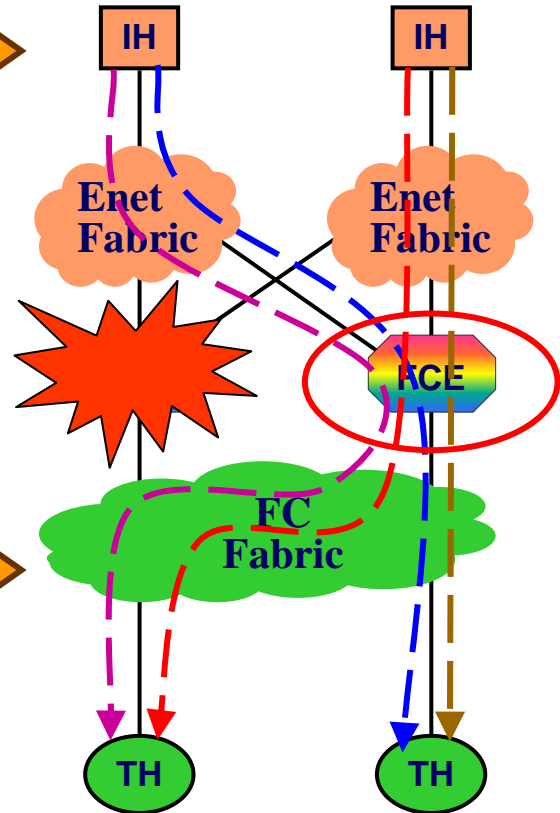
Before



FC Multipathing – 4 paths



After



Single Point of Failure:
FC doesn't see this!!

Legend

— Physical link

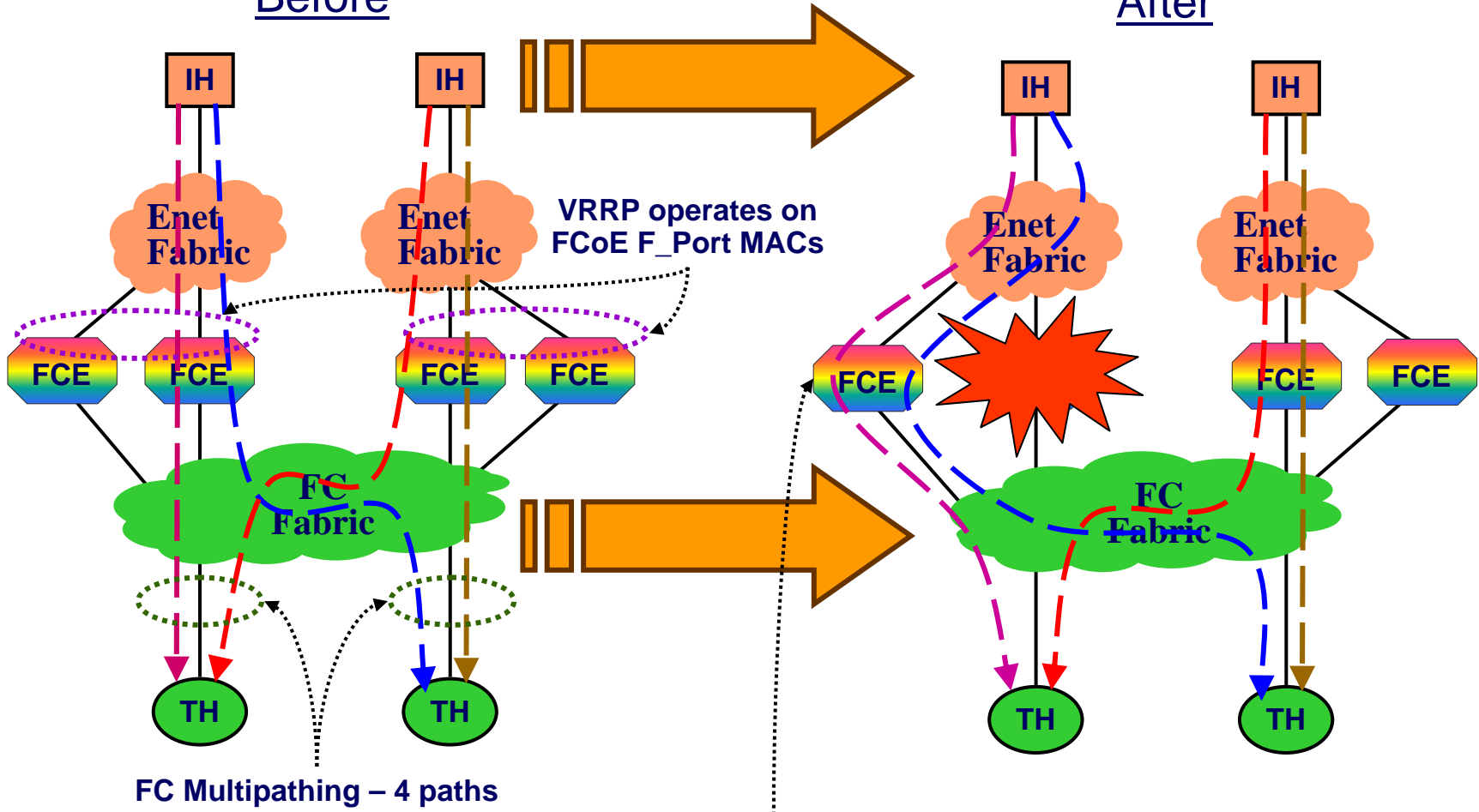
- - - Logical path

Robust failover scenario:

One host, one storage array, 2 ports each

Before

After



Legend

- Physical link
- - - Logical path

FCoE F_Port failover that preserves FC multipathing

FCoE Management Layers

