TWAREN年度教育訓練:
雲端網路規劃與設計

Bruce Wang

bruce_wang@ringline.com.tw

RingLine Corp.

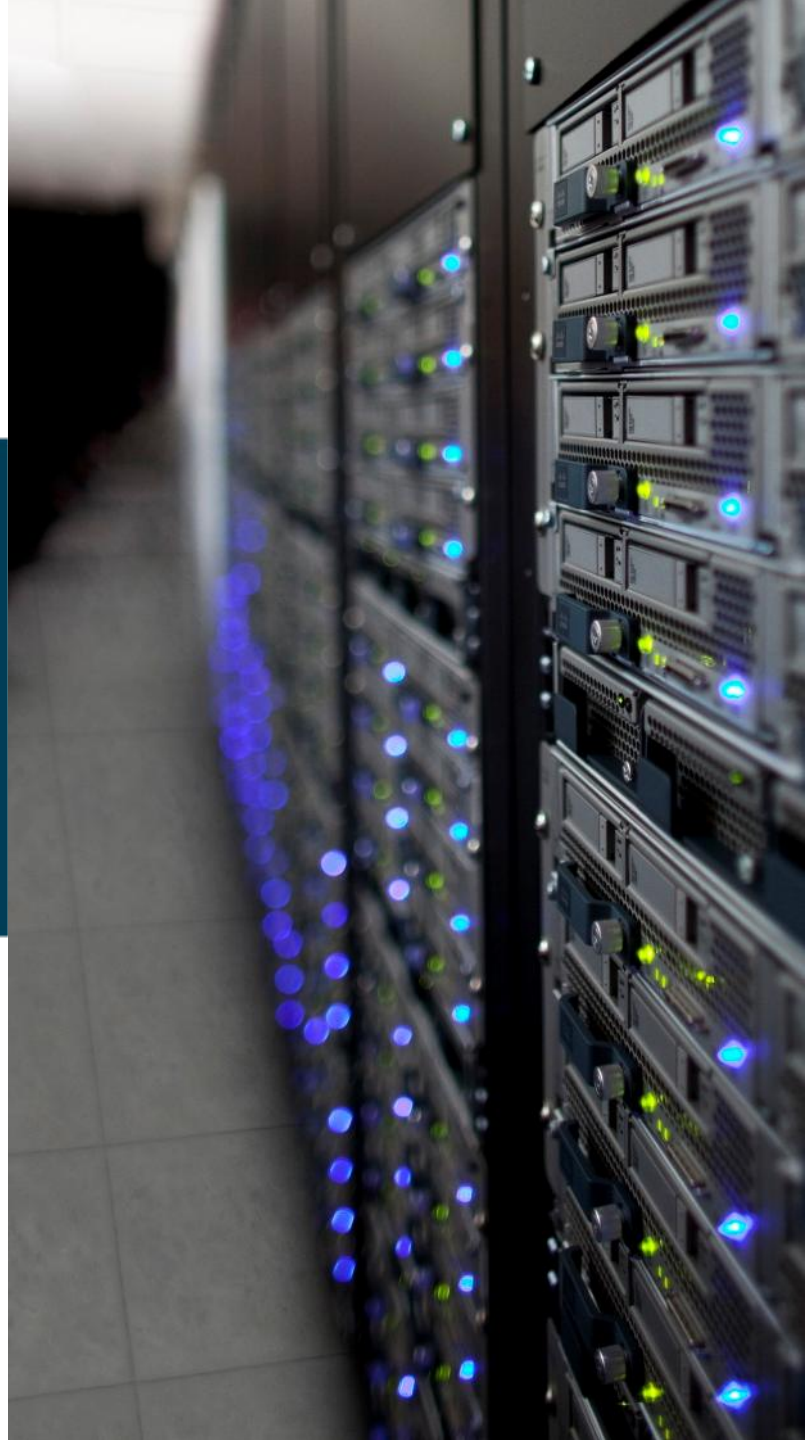
內容大綱

- 10G/40G/100G乙太網路標準
- 多鏈路互連交換技術FabricPath/TRILL
- 整合式網路傳輸技術FCoE
- 虛擬化網路交換技術802.1Qbh/802.1Qbg
- 雲端網路設計概念與範例

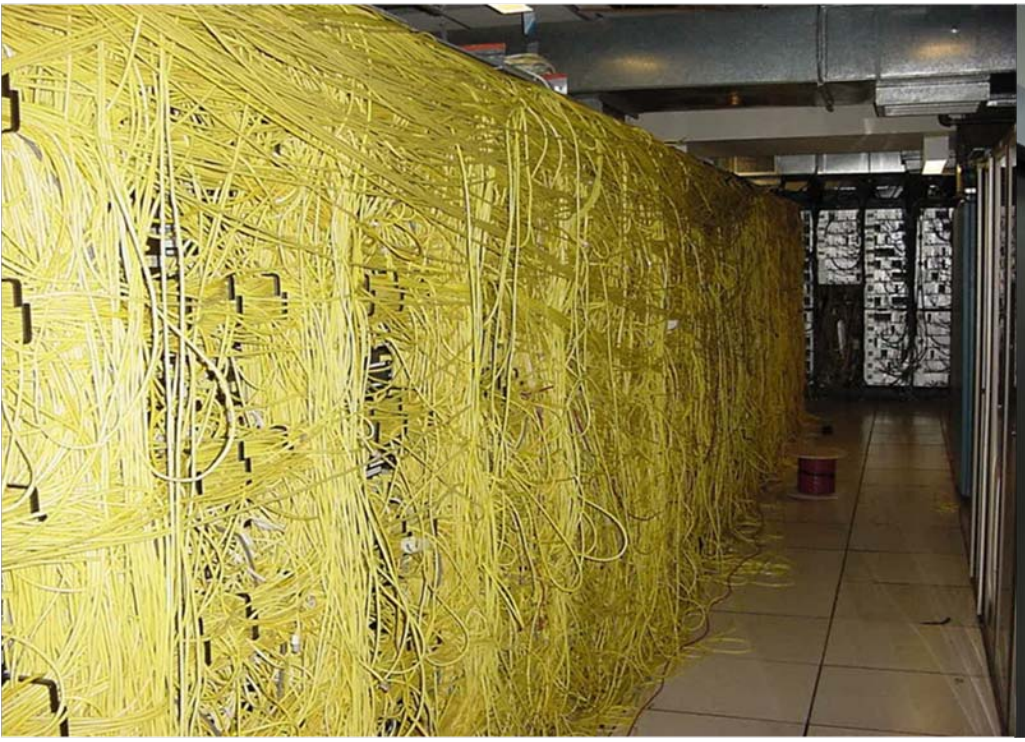
麟瑞科技
RING LINE CORPORATION

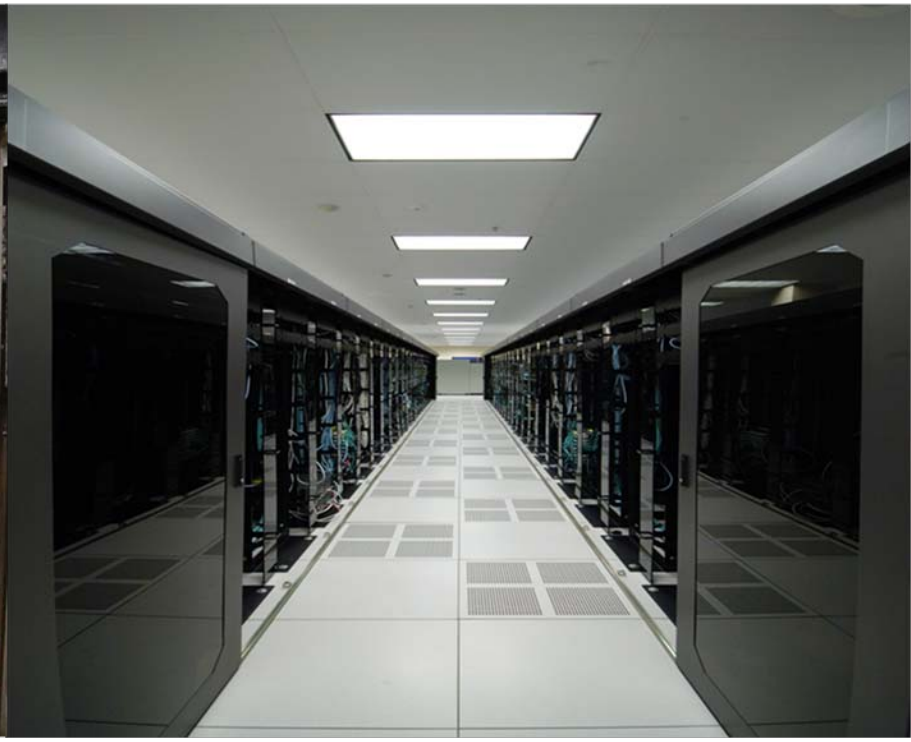
麟瑞科技
RING LINE CORPORATION

10G/40G/100G乙太網路



DC Facilities Top of Mind





Complexity, Cost, Power, Cooling

Standards Compliance

Reliability, Availability

Management, Security

Future Proofing

Increased Efficiency,

Simpler Operations

Scalability, Flexibility,

Technology adoption,

Modularity, Mobility

麟瑞科技 RING LINE CORPORATION

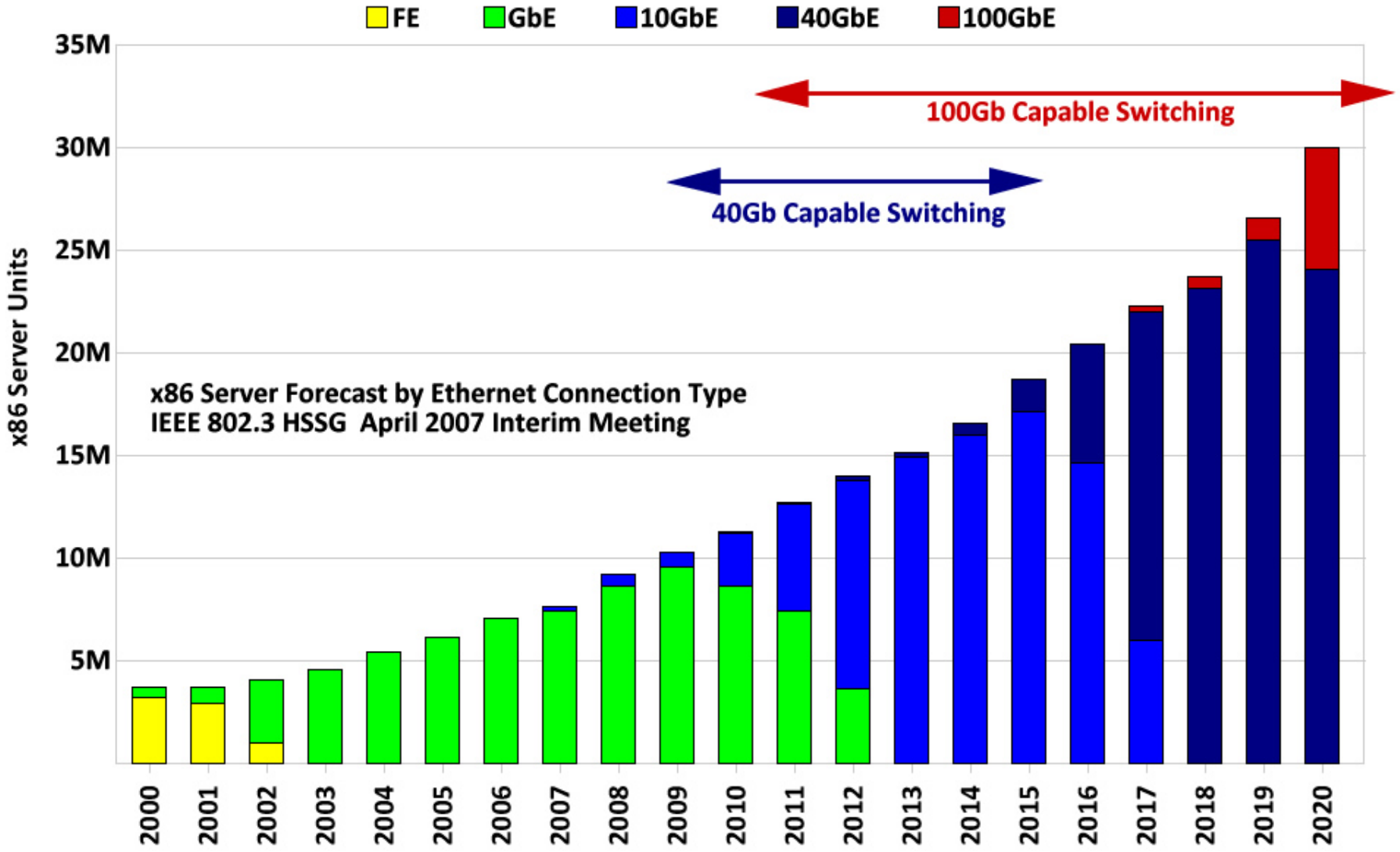
10 Gigabit Ethernet to the Server

Impacting DC access Layer Cabling Architecture

- Multi-core CPU architectures
- Virtual Machines driving Increased I/O bandwidth per server
  - increased business agility
- Increased network bandwidth demands
- Consolidation of Networks
  - Unified Fabrics / UIO
- Future Proofing - Network, Cable Plant, 10G/40G/100G

麟瑞科技

High Speed Ethernet Adoption on Servers



10GE Copper NIC Trend

10GBASE-T PHY from NIC to LOM – The Server View

- Dual port NICs on

  2008 1st gen silicon 90nm w/ ~10W

  2009 2nd gen silicon 90-65nm w/ ~6W

  2010/11 3rd gen silicon 65-40nm w/ ~4W peak, ~3W Avg w/EEE (802.3az)

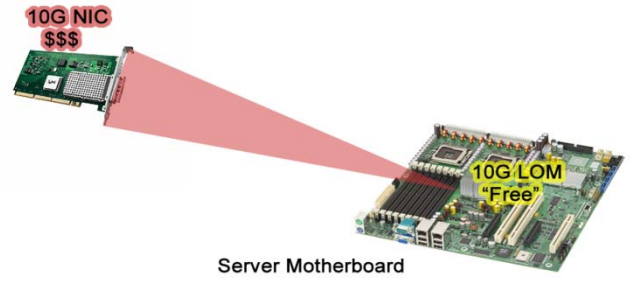
  2012 - 4th gen silicon 40nm w/ ~3W peak, <1W Avg w/EEE

- LOM removes the cost barrier to adopt 10G on servers
- Server vendors require LOM to be backward compatible, hence LOMs should support:

  interoperate with 100/1000/10000 switches

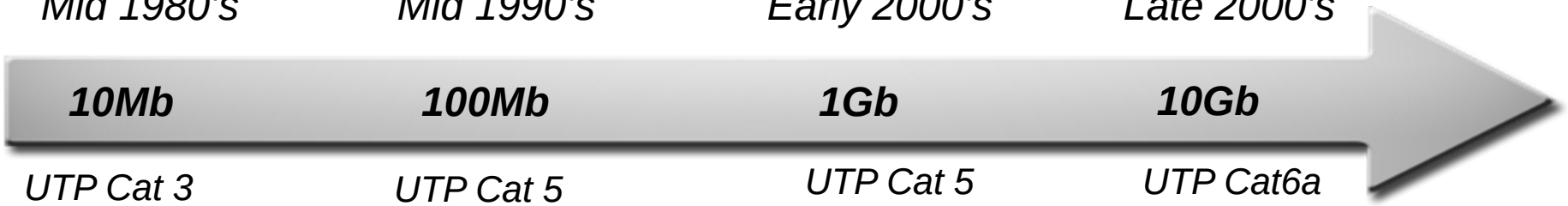
  support RJ45 cabling infrastructure

- PHY of choice for 10G LOM



10G Sever Media Option ?
- Fiber, Copper (CX1) , 6A in rack
- Fiber, CX1, 10GBASE-T (2011)

10 Gigabit Ethernet for Server Connectivity



10G Options

Connector (Media)	Cable	Distance	Transceiver (each side)	Transceiver Latency (link)	Standard
SFP+ CU* copper	Twinax	<7m	~ 0.1W	~ 0.1µs	SFF 8431**
X2 CX4 copper	Twinax	15m	4W	~ 0.1µs	IEEE 802.3ak
SFP+ USR MMF, ultra short reach	MM OM2 MM OM3	10m 100m		~ 0	ne
SFP+ SR MMF,short reach	MM OM2 MM OM3	82m 300m		~ 0	IEEE 802.3ae
RJ45 10GBASE-T copper	Cat6 Cat6a/7 Cat6a/7	55m 100m 30m	~ 6W*** ~ 6W*** ~ 4W***	2.5µs 2.5µs 1.5µs	IEEE 802.3an

In-rack X-rack

In-rack X-rack

~50% power savings with EEE

* Terminated cable

** Draft 3.0, not final

*** As of 2008; expected to decrease over time

10 Gigabit Transmissions

- Different Standards
  - 10GBase-T (IEEE 802.3an)
  - 10GBase-CX4 (IEEE 802.3ak)
  - 10GBase-R (IEEE 802.3xx)
    - LRM (802.3aq)
    - LR, ER, SR (802.3ae)
  - SFF 8431 (SFP+ Fiber & cu)
- Applications
  - Server Interconnects
  - Aggregation of Network Links
  - Switch to Switch Links
  - Storage Area Networks (SAN)

麟瑞科技

10GBase-T

- IEEE 802.3an
- Full duplex transmissions
- 100 meters on Class F (shielded) cabling
- 30-55 meters on Class E/Category 6 cabling
- 100m on Class EA/Category 6 augmented copper cabling
- Alien Cross-Talk suppression up to 500 MHz
- Cat 6 parameters extrapolated up to 500 MHz
- Cat 7 insertion loss characteristics

麟瑞科技

Twisted Pair Cabling For 10GBASE-T (IEEE 802.3an)

U/UTP (Old designation UTP)
Outer Unshielded/Inner Pairs Unschielded

Cat 6a:
*100m 10GBASE-T
**largest diameter up to 0.354 in

Cat 6:
*55m 10GBASE-T
**larger diameter than Cat5 (~0.3 in)

F/UTP (Old designation FTP)
Outer Foil Shielded/Inner Pairs Unshielded

Cat 6/6a:
*100m 10GBASE-T
**More flexible/easier to manage than Cat6a U/UTP
***Equivalent diameter to Cat6

S/FTP (Old designation S/STP)
Outer Foil Shielded/Inner Pairs Foil shielded

Cat 7:
*100m 10GBASE-T
**Most expensive
***Smaller diamter than Cat6a
****Not popular in North America

10G Copper Infiniband - 10GBase-CX4
10G Copper on Twin Axial copper

- IEEE 802.3ak
- Supports 10G up to 15 meters
- Quad 100 ohm twinax, Infiniband cable and connector
- Primarily for rack-to-rack links
- Low Latency
- Use in Infiinband environments

麟瑞科技
RING LINE CORPORATION

10G SPF+ Cu

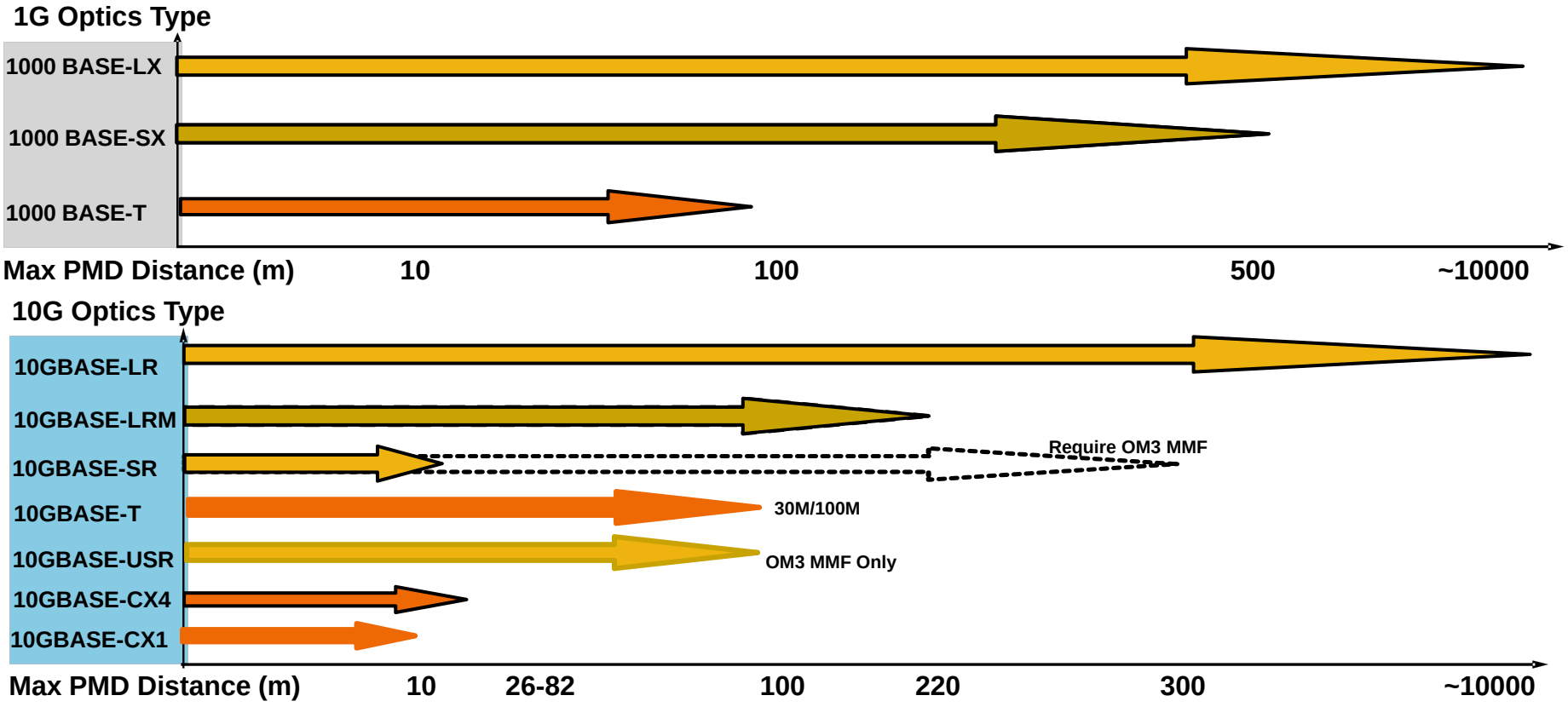
- SFF 8431
- Supports 10GE passive direct attached up to 7 meters
- Twinax with direct attached SFP+
- Primarily for in rack and rack-to-rack links
- Low Latency, low cost, low power

麟瑞科技

Media Comparison for 10G Data Center Ethernet: Present View

Copper Cable Media	Pros	Cons	Key Elements for Mass Adoption
Category 6A Copper Cable (U/UTP & S/STP)	Reach (100 Meters)	Power Requirements for Active Silicon (Currently 8-10 watts)	Lower Power Active Silicon (less than 3 watts/port)
	Structured Cabling (up to 4 connector models	Cooling Issues in the Data Center	Achieving low power levels within the next 3 yrs
	RJ45 Interface (TIA/ISO Industry Std.)	Size of the Horizontal & Patch Cables (introduces Cable Management issues)	48-port Switch Density (to achieve acceptable cost model)
	Easy migration path from 1G to 10G (Interoperability)	Availability	Smaller Horizontal & Patch Cables (less than 0.275 inches O.D.)
CX1 SFP+ Cable Assemblies	Low Power Requirements	Reach (10 meters)	Educating Customer on new connector interface
	Low Latency	Cost (compared to Cat 6A cabling)	Extending Cable Reach to at least 30 meter (w/cost effective solution)
	Migration path for 40/100 G	New connector interface (not traditional RJ45 interface)	Offering Structured Cabling
	Easy Cable Management	No structured cabling	Publishing an SFP+ Copper Standard

1GE-10GE Transceiver Perfomance



In Rack X-rack	Mid to End of Rack	Across Aisles	Across Sites
<10M	<100 M	<300 M	<10 KM

10GE (IEE 802.3ae) Optical Transmission

Media Options

The 802.3ae 10GbE standard defines 3 MM and 1 SM fiber category based on the maximum transmission reach as shown below (ISO 11801 Standard defines the following MM and SM fiber types):

SPEED	REACH		
	300m	500m	200m
100Mb/s	OM1	OM1	OM1
1,000Mb/s	OM1	OM2	OS1
10Gb/s	OM3*	OS1	OS1

OM1 is equivalent to standard 62.5/125µm MM fiber

OM2 is equivalent to standard 50/125µm fiber.

OM3 is laser enhanced 50/125µm fiber – 10gig

OS1 is equivalent to SM 8/125µm fiber.

* This refers to 10GBASE-SR, for LX4 & LRM you get 300 and 220m respectively irrespective of fiber type

Not all laser optimized 10Gig fiber cable is the same.		150M	300M	550M
	10Gig	OM2 Plus	OM3	OM3 Plus

10GE SFP+ Optical

- Smallest 10GE form factor
- Low Power
- Low Latency
- Hot swappable
- High density
- Optical SFP+ interoperates with other 10GE modules
  - XFP
  - XENPAK
  - X2

SFP+ Optical Module

Cost Effective 10G Server Connectivity

Today

SFP+ USR – ‘Ultra Short Reach’

- 100M on OM3 fiber, 30M on OM2 fiber
- Not a standard.

SFP+ Direct Attach

- 1, 3, 5 and 7M on Twinax
- 0.1W Power

Why 100GE ?

- **Ethernet Ubiquity**
  - Serial WAN technology not being developed beyond 40Gbps
  - Ethernet no longer just for LAN
  - High industry cooperation amongst IEEE, ITU, OIF to develop 100GE
- **Link Aggregation Inefficiencies**
  - Core networks typically need 4x-10x highest speed user interface
  - Lower speed interface bundling scales poorly, management challenges
- **Infrastructure Consolidation**
  - Expense reduction with saving multiple links, platforms and inter-connects
  - Peering Points, Cloud infrastructure demanding significant bandwidth today

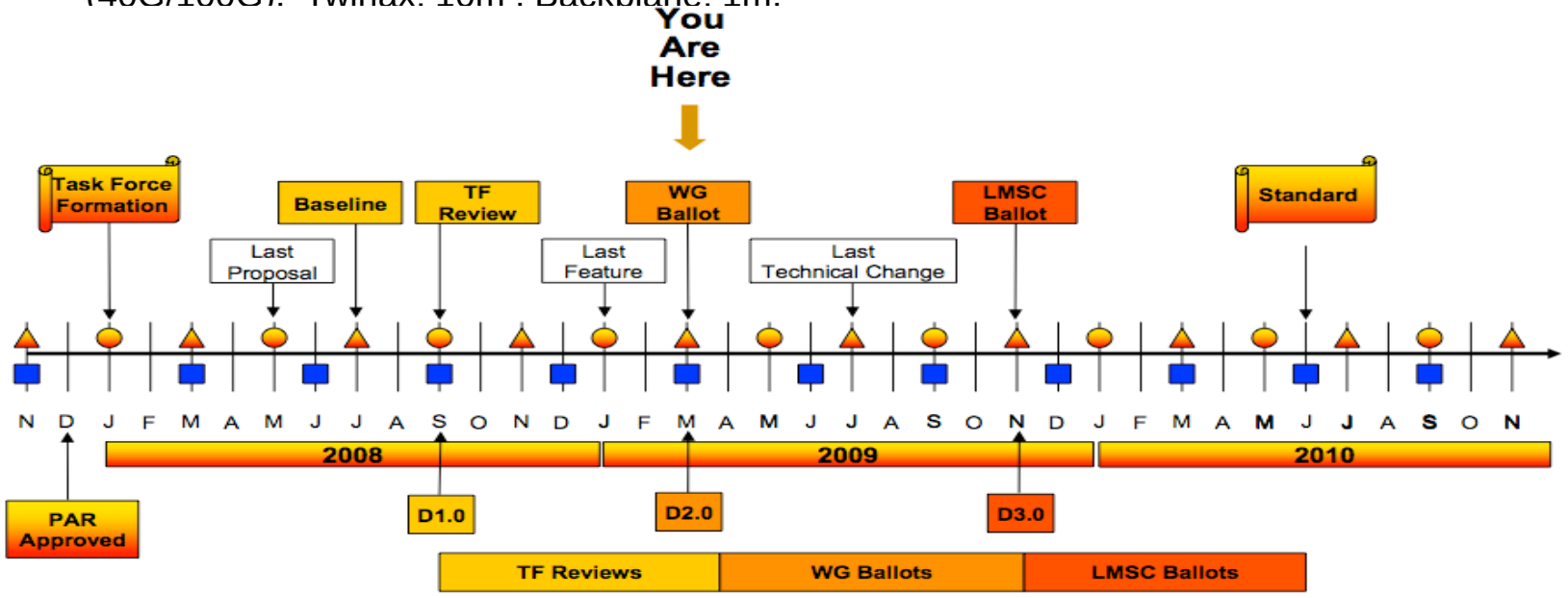
IEEE 802.3ba (40G/100G)

- Bandwidth requirements for computing, core and storage networking require different data rates for next generation Data Centers:

  40 Gb/s Ethernet interface - Servers, HPC clusters, blade servers, storage area networks and network attached storage

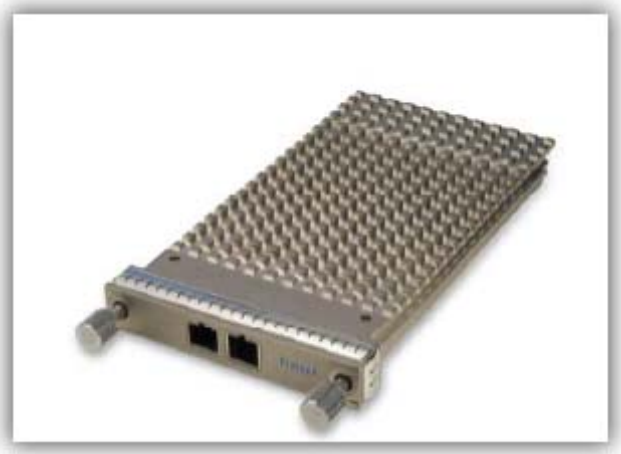
  100 Gb/s Ethernet interface - Core network switching, routing, and aggregation in DCs, internet exchanges and service provider peering points for high bandwidth applications such as video-on-demand

- Defined Channel Reach: SMF: 10 km (40G/100G), 40 km (100G), OM3: 100 m (40G/100G), Twinax: 10m , Backplane: 1m.



1st Gen 40GbE Transceivers

CFP



Applications:
Single Mode Fiber 10Km
Multi Mode OM-3 100m
Twinax Copper
“FourX” converter for 4x10GbE (SFP+)

Power Consumption:
Up to 8W @ 40GbE

QSFP

Applications:
Multimode Parallel Fiber
Twinax Copper
10 KM Single Mode (Future)

Power Consumption:
Up to 3.5W

麟瑞科技

1st Gen 100GbE Transceivers

100GbE CFP requires "Riding HeatSink" SMF optimized

CFP features a new concept known as the riding heat sink, in which the heat sink is attached to rails on the host card and "rides" on top of the CFP, which is flat topped.

Applications:

Single Mode Fiber 10Km and 40Km
Multi Mode Fiber OM-3 100m

Power Consumption:

Up to 25W

CXP MMF/Twinax optimized

CXP was created to satisfy the high-density requirements of the data center, targeting parallel interconnections for 12x QDR InfiniBand (120 Gbps), 100 GbE, and proprietary links between systems collocated in the same facility. The InfiniBand Trade Association is currently standardizing the CXP.

Applications:

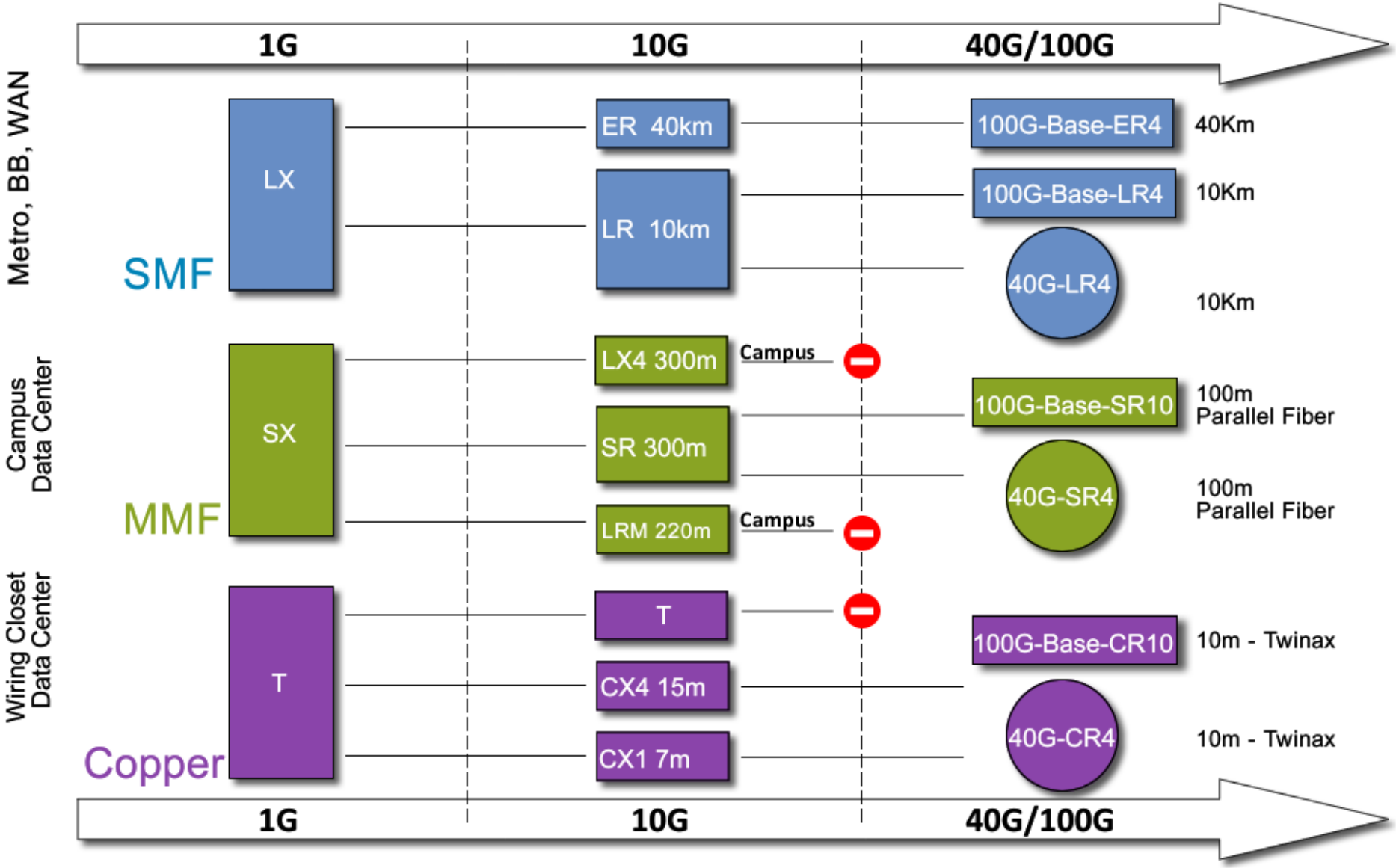
Multi Mode Fiber OM-3 100m
Twinax copper assembly 7m

Power Consumption:

Up to 3.5W

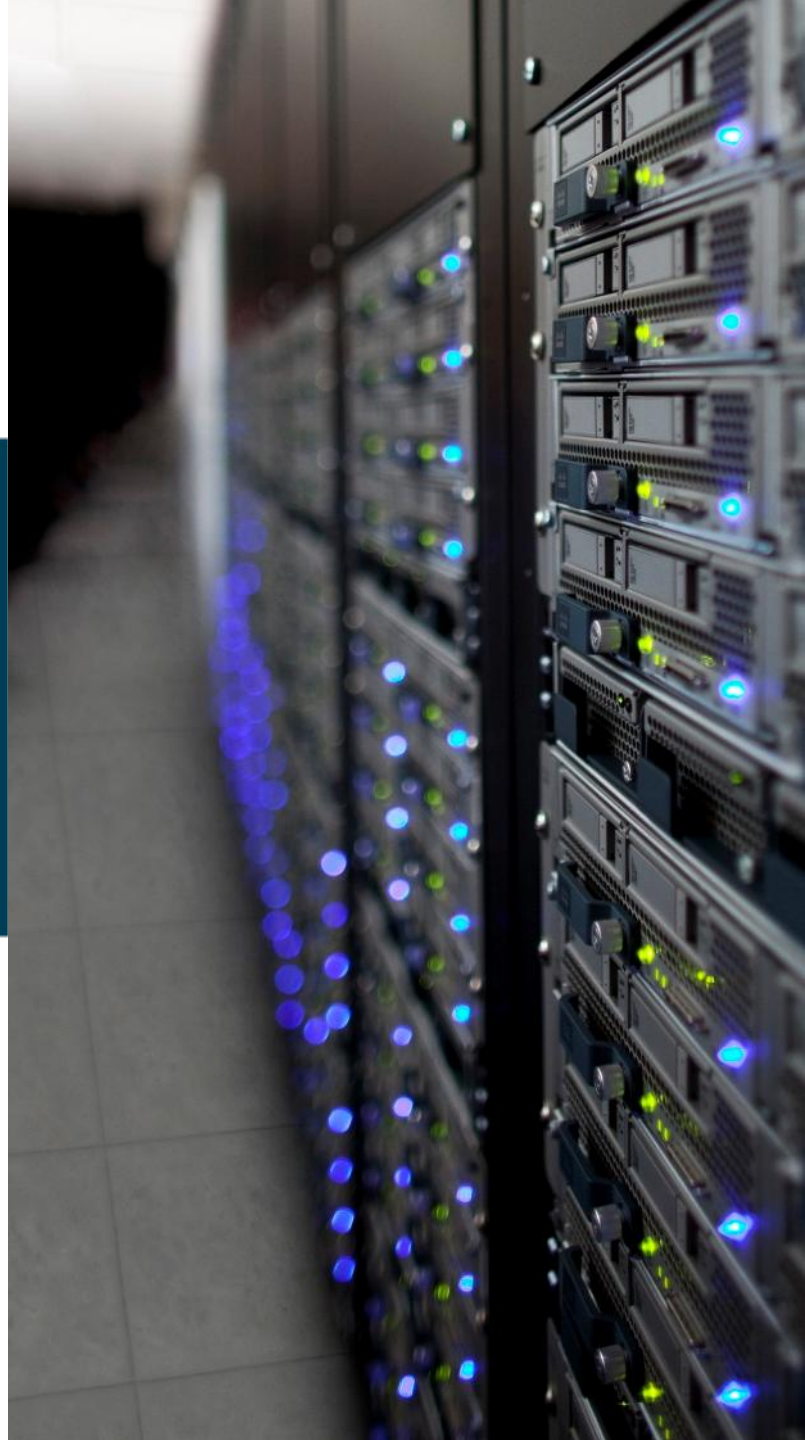
麟瑞科技

High Speed Ethernet Standard Interfaces



麟瑞科技
RING LINE CORPORATION

多鏈路互連交換技術
FabricPath/TRILL

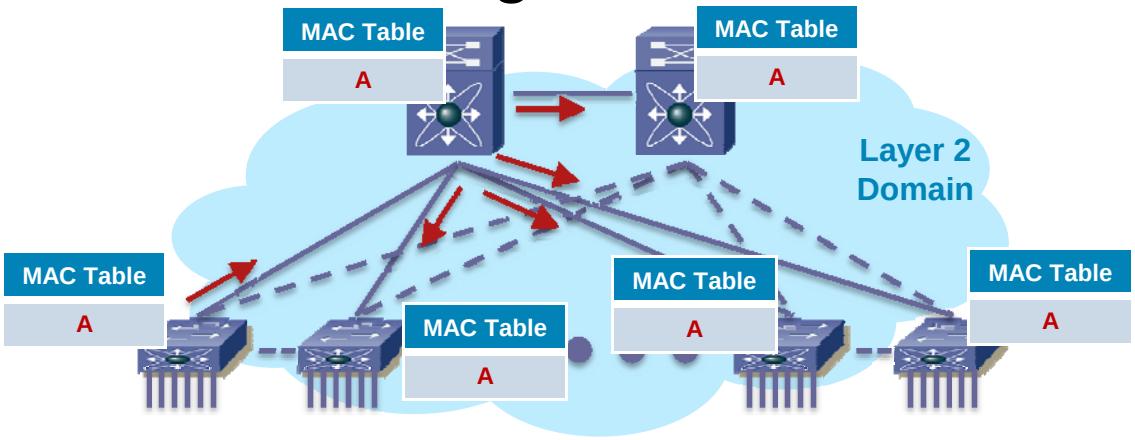


L2 Provides Flexibility in the Data Center

- Layer 2 is still required by some data center applications
- With Layer 2:
  - Server mobility does not require interaction between Network/Server teams
  - No physical constraint on server location
- Layer 2 is Layer 3 agnostic
- Layer 2 is “plug and play”

L2 Mac Address Scaling Issues

- Mac addresses facts:
  - They are billions
  - They have no location associated to them, no hierarchy
  - They are not “registered” by the hosts to the network
- A routing table is impossible at Layer 2:
  default forwarding behavior is **flooding**



- A filtering database is set up to limit flooding
- The whole mechanism is not scalable

L2的樹狀結構

Branches of trees never interconnect (no loop)

11 Physical Links

5 Logical Links

- The Spanning Tree Protocol (STP) is typically used to build this tree
- Tree topology implies:
  - Wasted bandwidth -> over-subscription exacerbated (E/W)
  - Sub-optimal paths
  - Conservative convergence -> failure catastrophic

RFC 5556 TRILL

- IETF standard **for Layer 2 Multipathing**
- Driven by multiple vendors, including Cisco
- Base protocol RFC ready for standardization but waiting on dependent standards
- Control-plane protocol RFCs still in process
- Target for standard completion is early CY2011

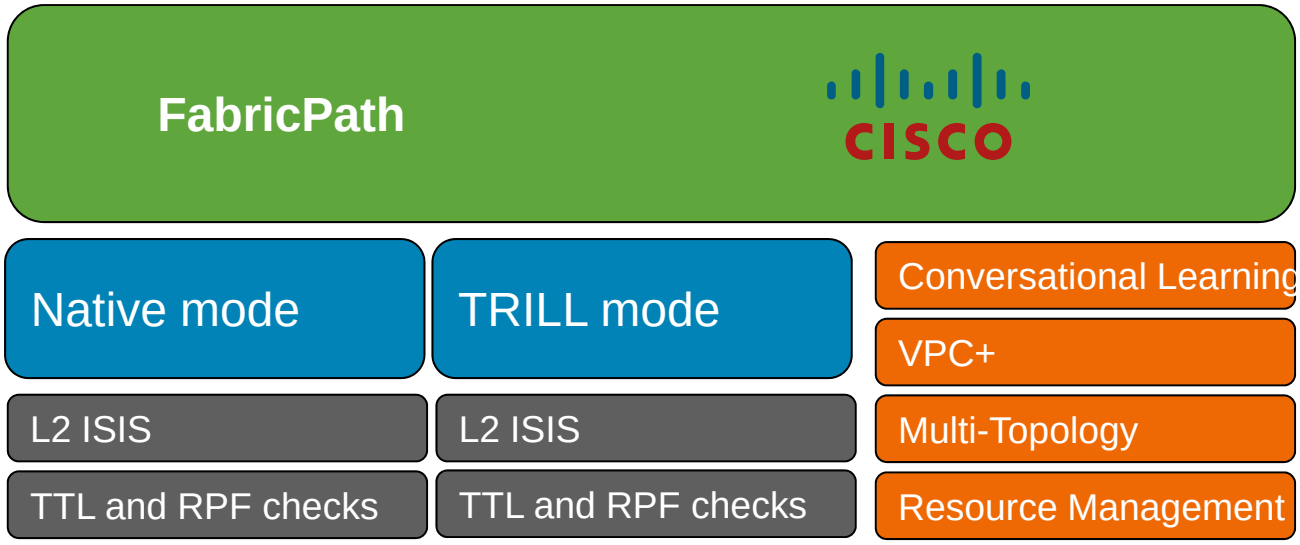
http://datatracker.ietf.org/wg/trill/

麟瑞科技

Cisco FabricPath 與 TRILL的關聯

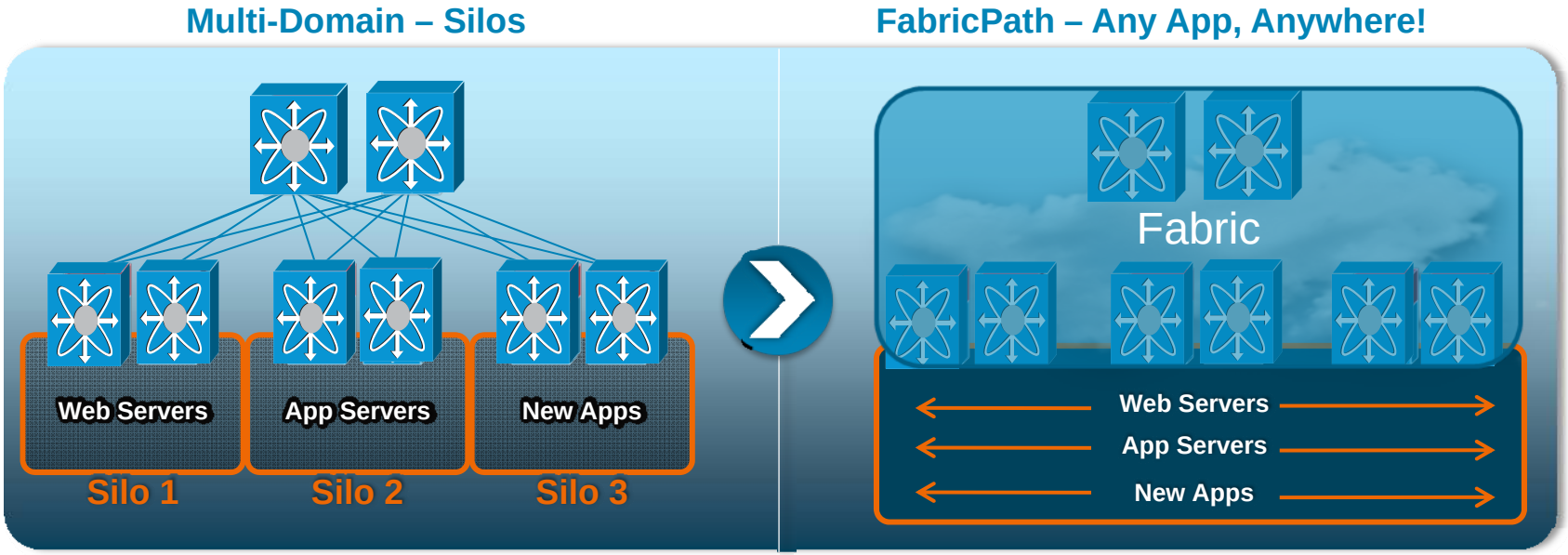
- **FabricPath** is an **umbrella term** for a set of Layer 2 multipathing technologies
- FabricPath initial release runs in a Native mode that is Cisco-specific, using proprietary encapsulation and control-plane elements
- Once TRILL standard complete, FabricPath will offer a TRILL-compliant mode for third-party interoperability. This will be achieved by a simple software upgrade.
- Nexus 7000 F1 I/O modules and Nexus 5500 HW are capable of running both FabricPath and TRILL modes

TRILL與FabricPath差異



FabricPath: Simple from the Outside

Benefits the Server Team



- FabricPath provides a Fabric that looks like a switch => No silos, workload mobility and maximum flexibility
- Lowers OPEX by simplifying server team operation (no disruption, no interaction with network team required)

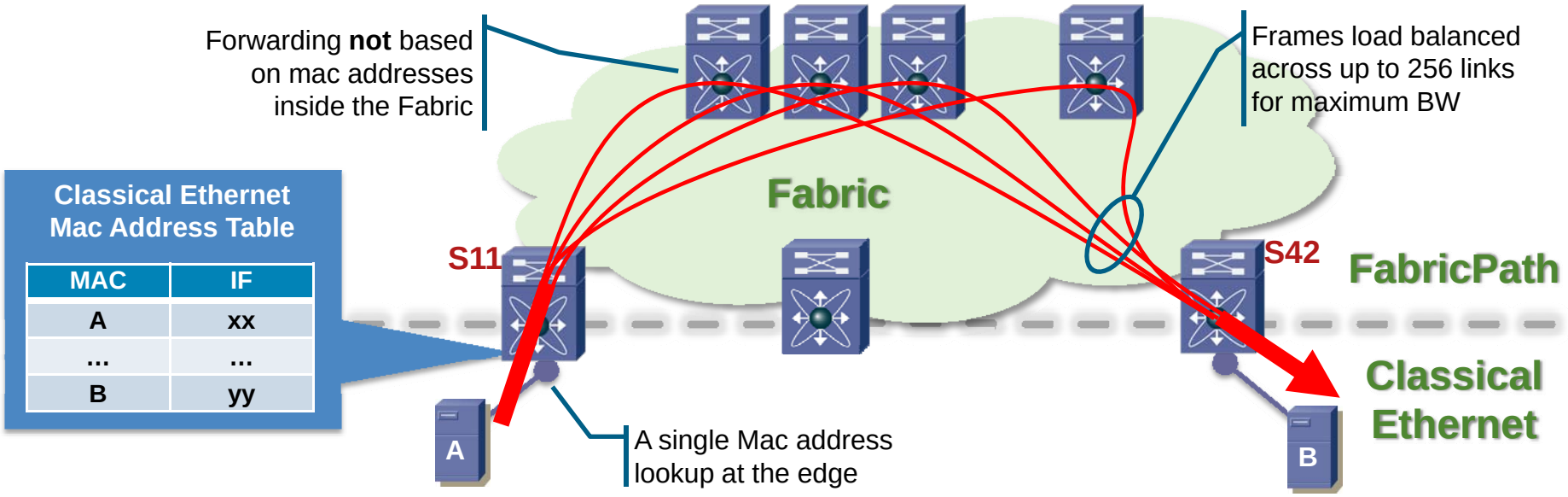
FabricPath: Simple from the Inside

Benefits the Network Team

- Reduces the number of switches required
  (higher port density possible without increasing oversubscription)
- Isolate the network from the users
  - No topology change propagation between inside/outside
  - Fabric can be upgraded/reconfigured live
- Open protocol, no secret sauce
  - Operates on a single control protocol (unicast, multicast, pruning)
  - Maintenance tools equivalent to those of L3 networks (ping, traceroute)
  - Little configuration (auto addressing)

麟瑞科技

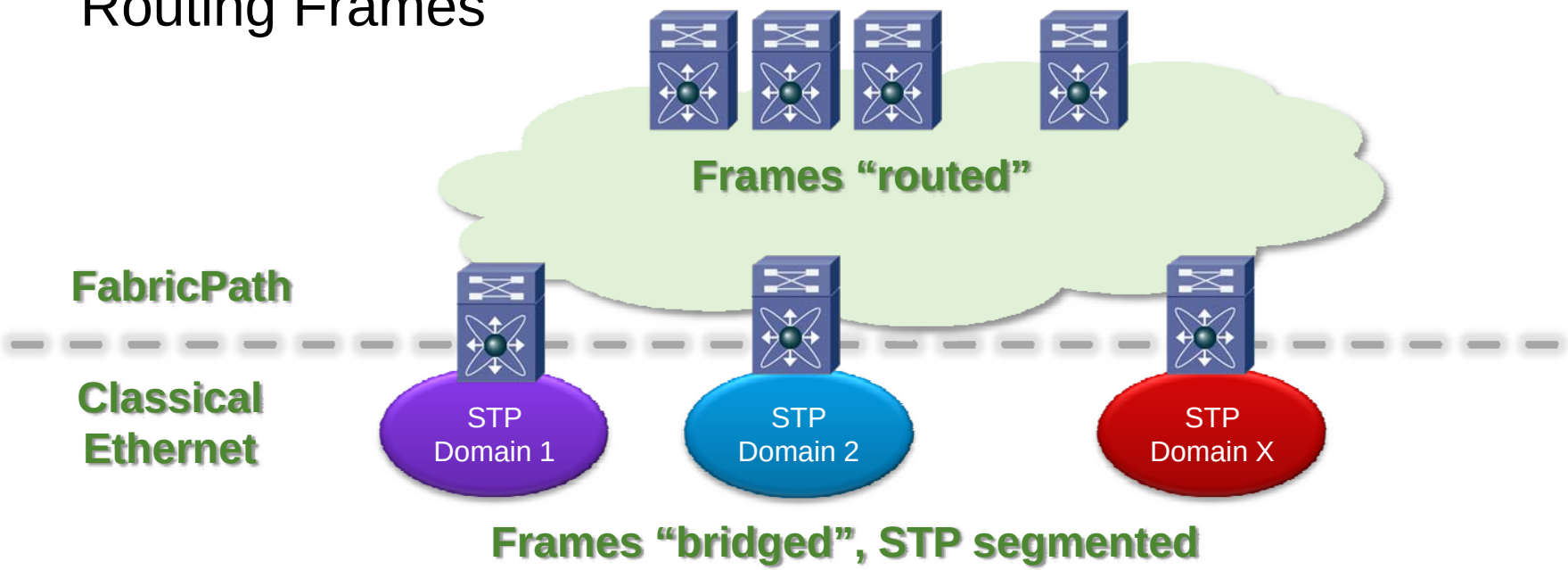
FabricPath is Efficient



- "Unlimited" bandwidth
  16 ways ECMP, up to 256x 10G links between 2 boxes, 160Tbps Fabric
- A single mac address lookup at the edge, then forwarding based on 12 bits up to the remote port
- Traffic goes across the shortest path
- Fast convergence, high resiliency

FabricPath is Scalable

Routing Frames

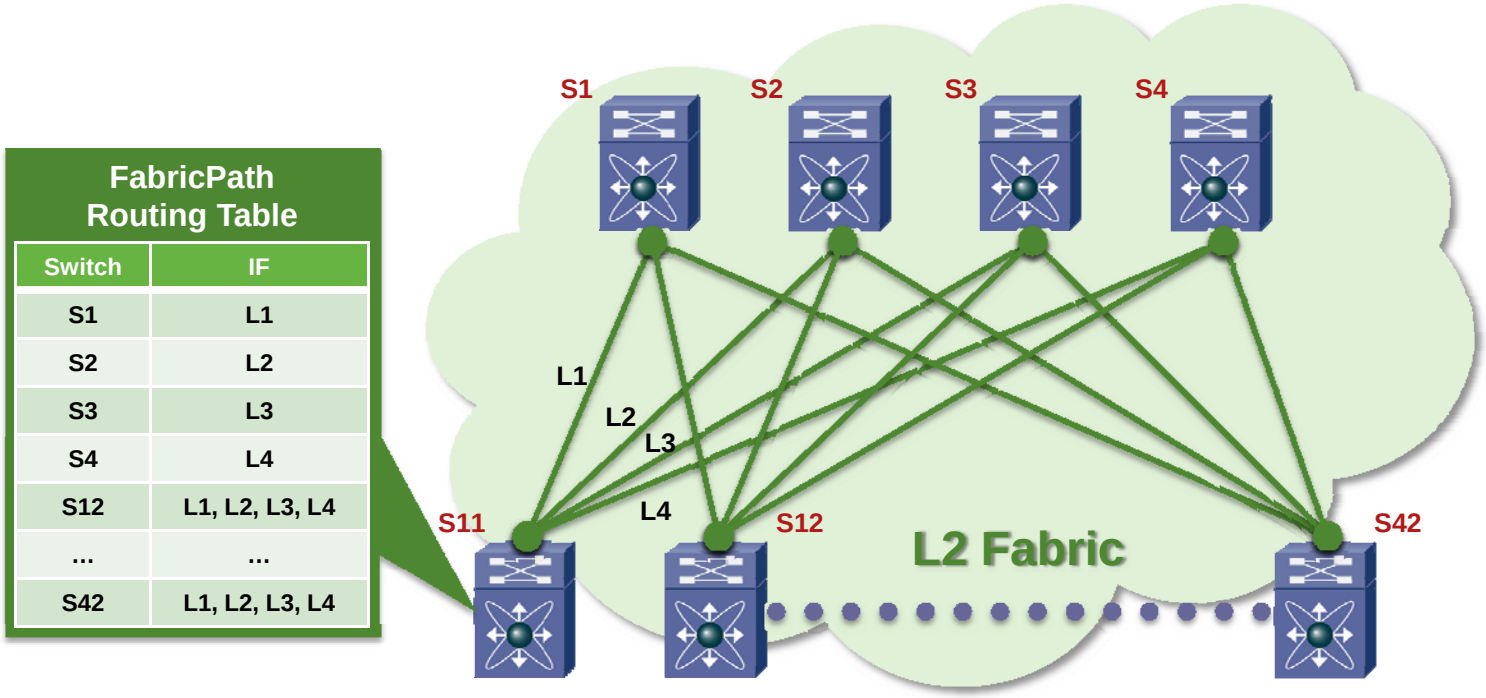


- Learning is only done at the edge, selectively
- Fabric does not rely on mac address tables for forwarding
- Growing the Fabric does not grow the risk (routing frames with TTL, RPF check, etc…)

Control Plane運作

Plug-N-Play L2 IS-IS is used to manage forwarding topology

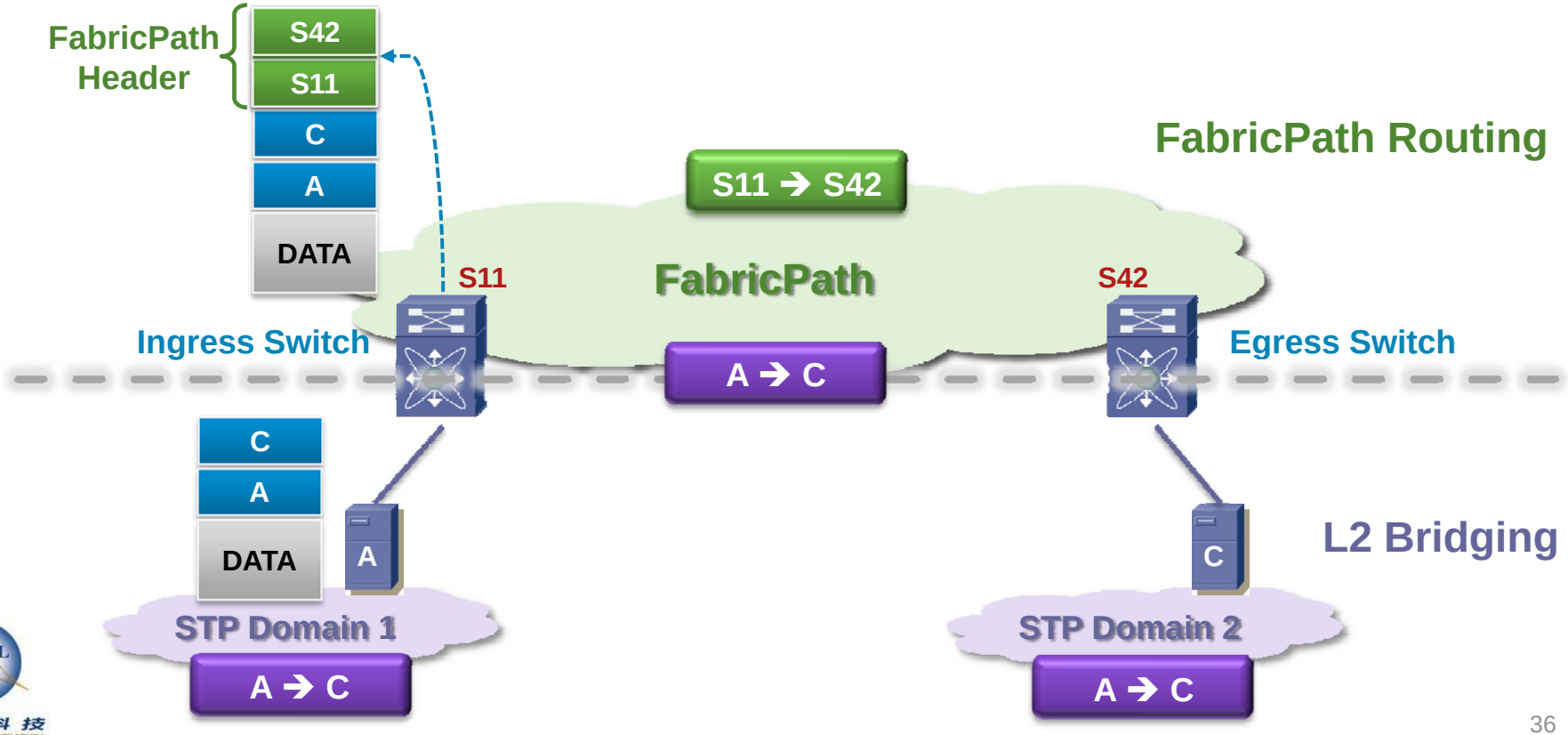
- Assigned switch addresses to all FabricPath enabled switches automatically (no user configuration required)
- Compute shortest, pair-wise paths
- Support equal-cost paths between any FabricPath switch pairs



Data Plane運作

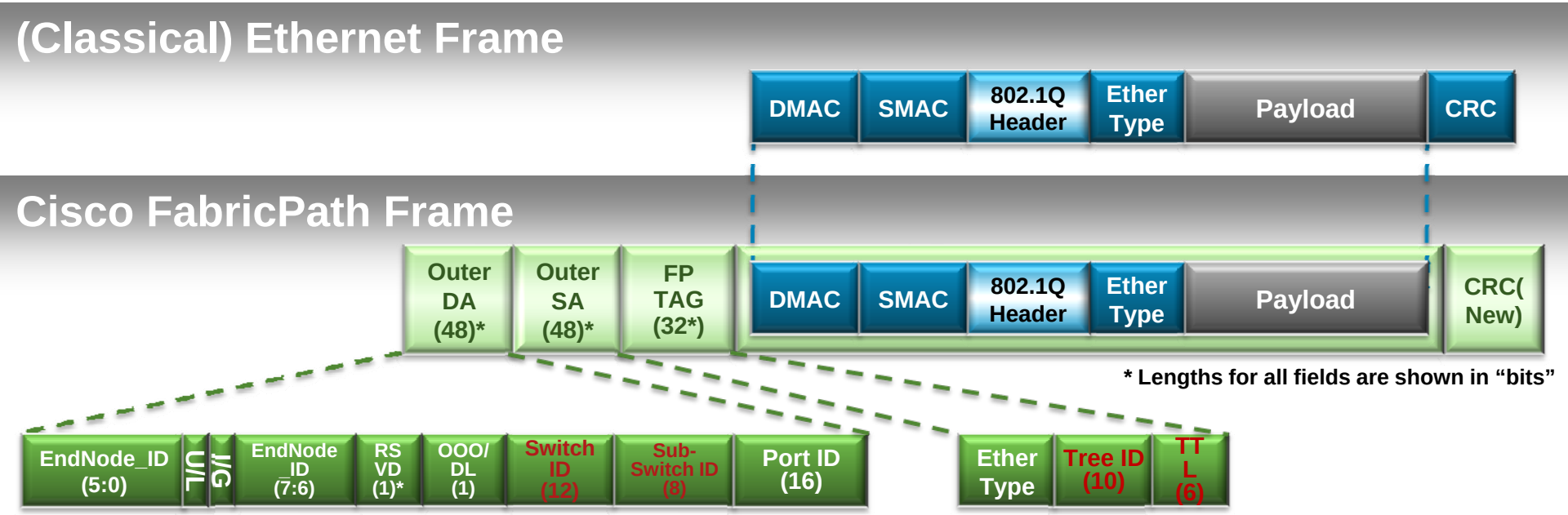
Encapsulation to creates hierarchical address scheme

- FabricPath header is imposed by ingress switch
- Ingress and egress switch addresses are used to make "Routing" decision
- No MAC learning required inside the L2 Fabric



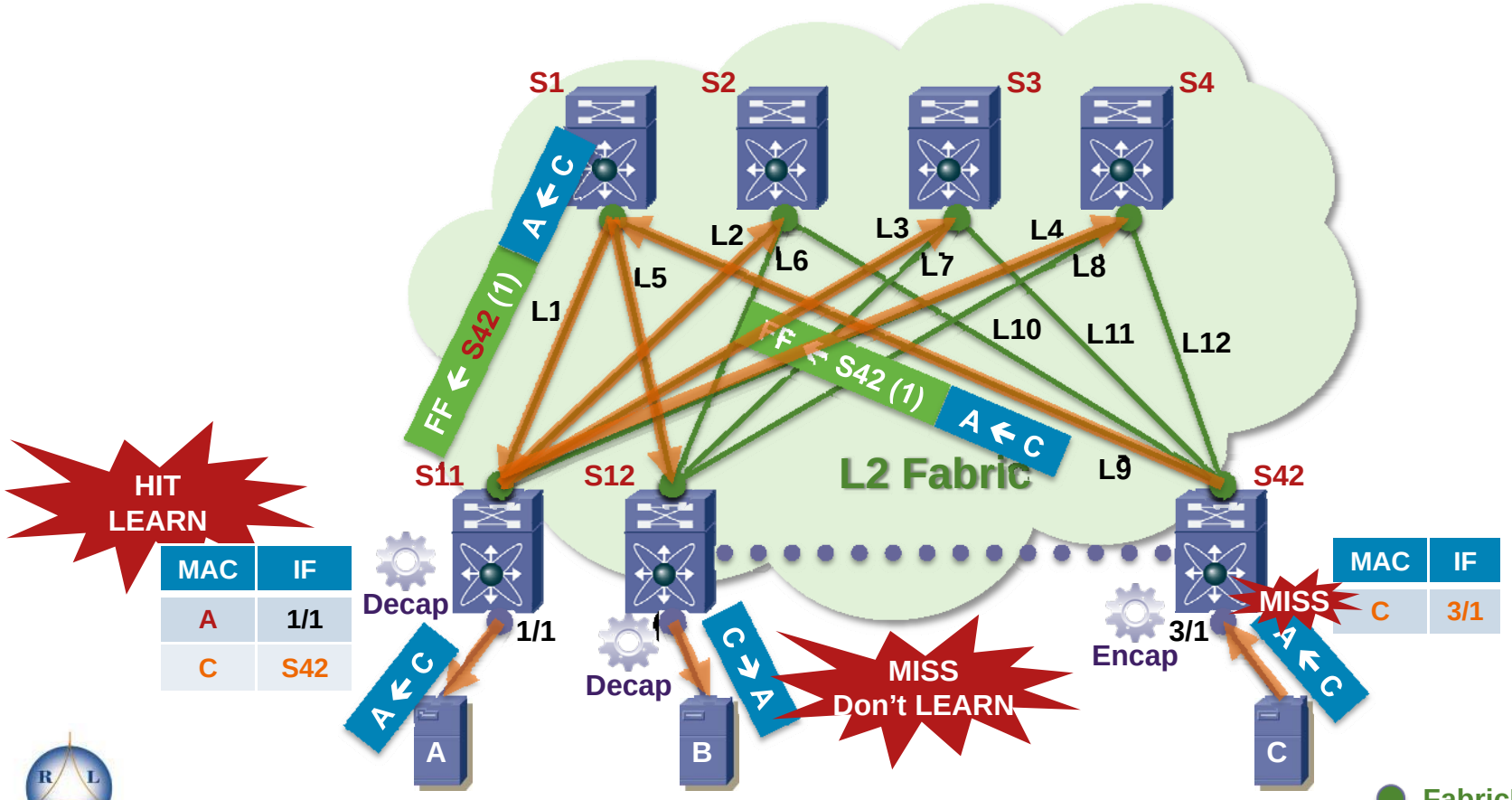
Native FabricPath Encapsulation

16-bytes header provide fields to help create hierarchical L2 address space and facilitate feature enhancements

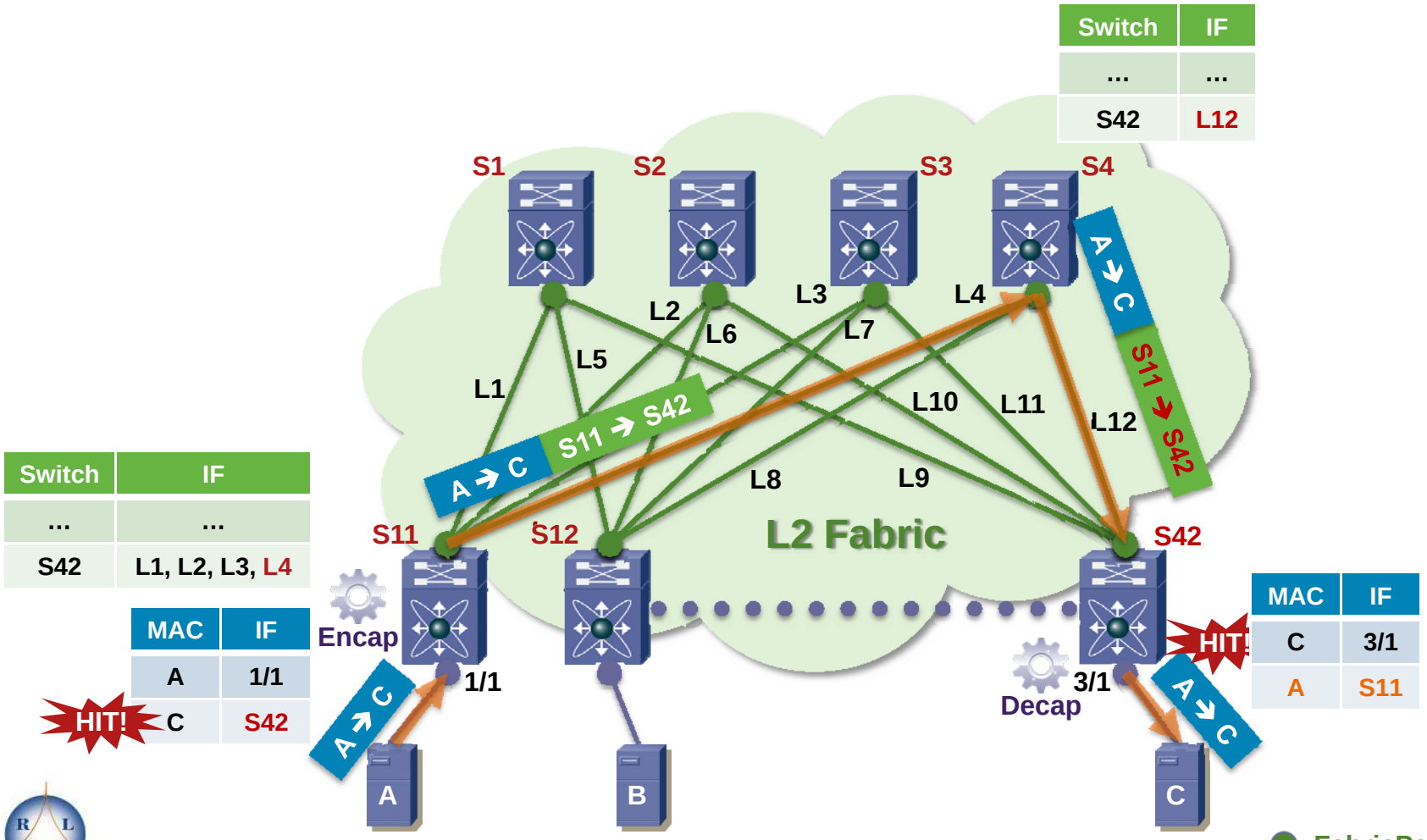


- ***Switch ID:*** 12-bit number identifying a particular device in the L2 fabric.
- ***Sub-Switch ID:*** Combined with Switch ID to identify vPC+ behind a pair of peer-switches
- ***Tree ID:*** Unique number assigned to help identify each distribution "Tree"
- ***Forwarding Tag (Ftag):*** mainly used to identify multicast trees
- ***TTL:*** Decremented at each hop, protection against temporary loops in the data plane

FabricPath Forwarding: Unknown Unicast



FabricPath Forwarding: Known Unicast



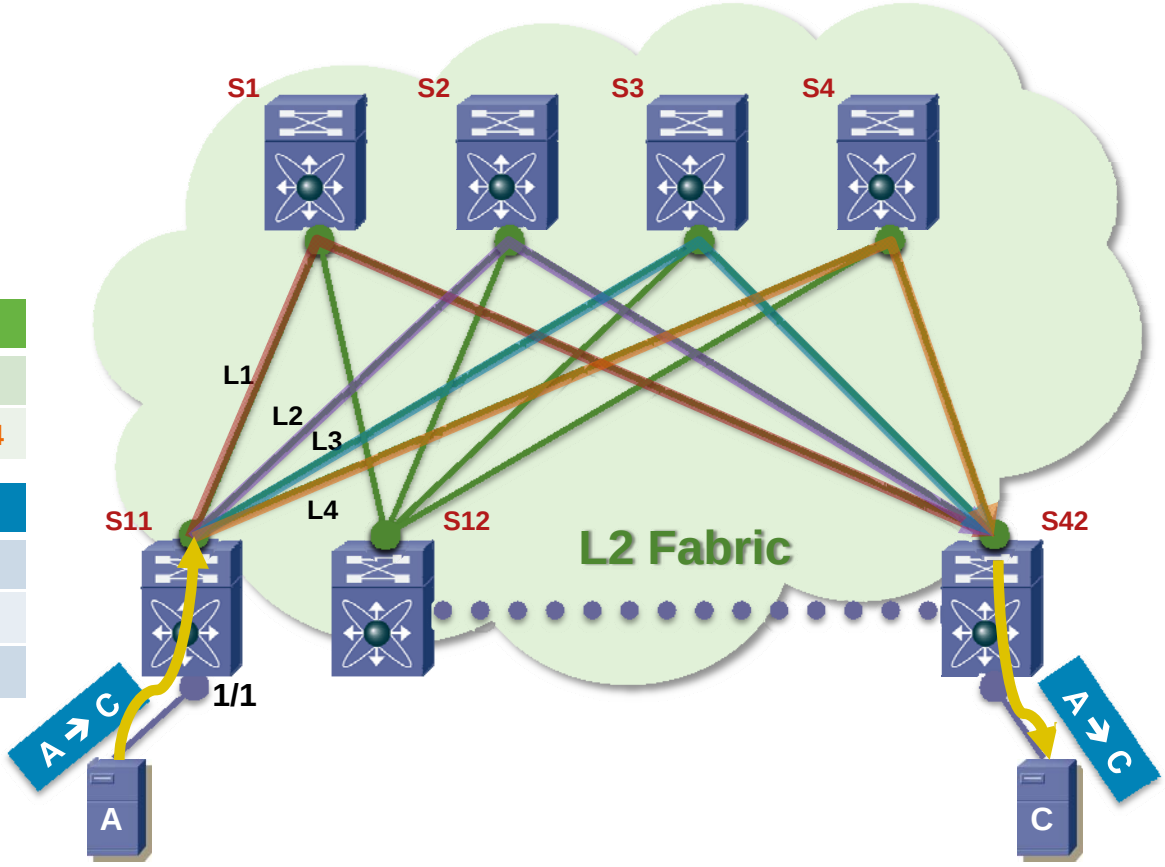
Unicast Equal Cost Multipathing

Forwarding decision based on ‘FabricPath Routing Table’

- Support more than 2 paths (16 way ECMP) across the Fabric
- Increase bi-sectional bandwidth beyond port-channel
- High availability with N+1 path redundancy

Switch	IF
…	…
S42	L1, L2, L3, L4

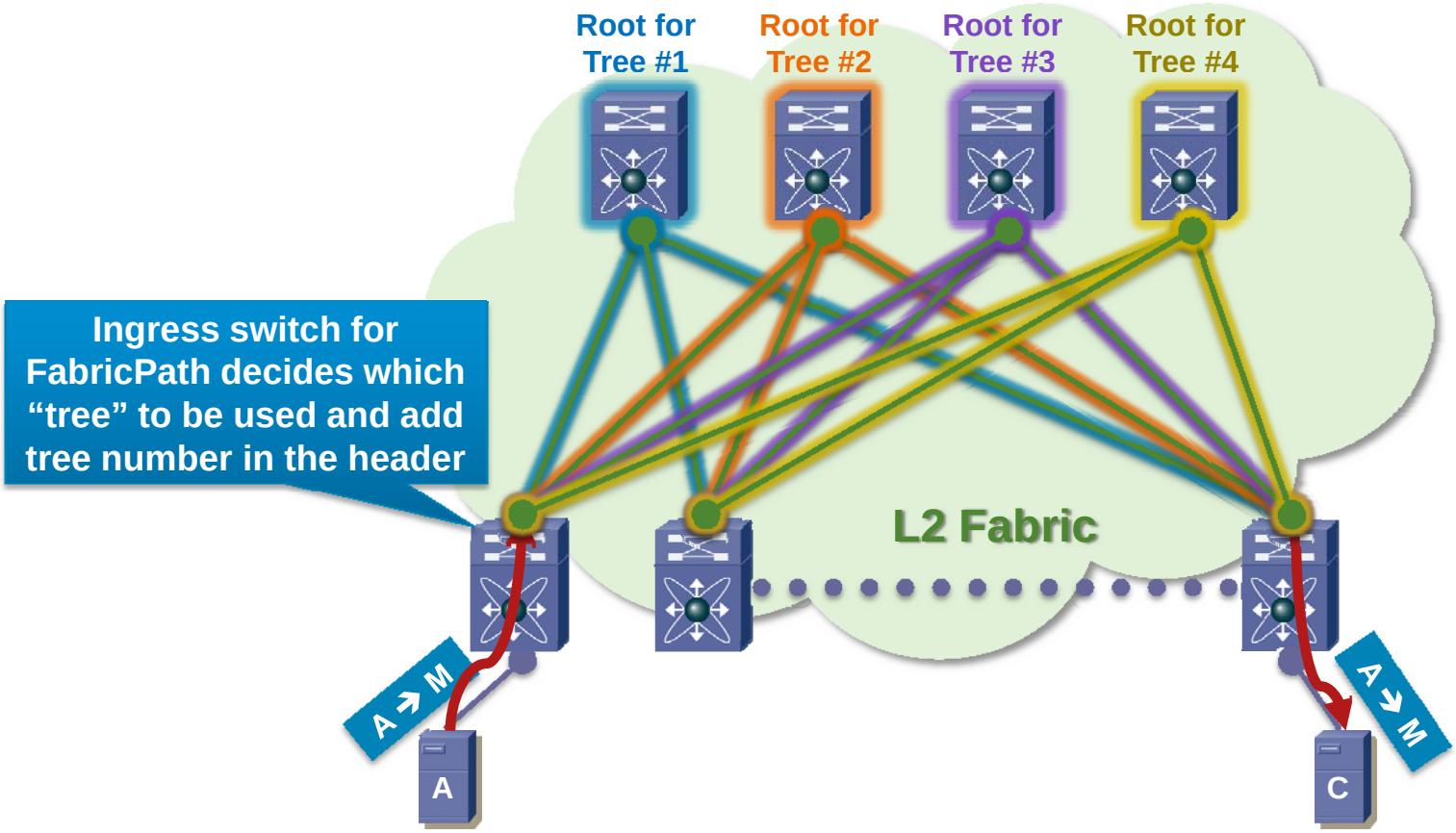
MAC	IF
A	1/1
…	…
C	S42



Multicast with FabricPath

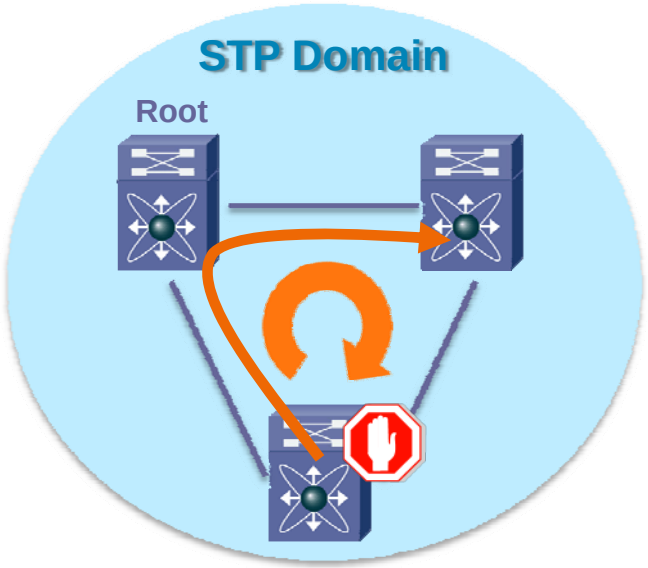
Forwarding through distinct ‘Trees’

- Several ‘Trees’ are rooted in key location inside the fabric
- All Switches in L2 Fabric share the same view for each ‘Tree’
- Multicast traffic load-balanced across these ‘Trees’

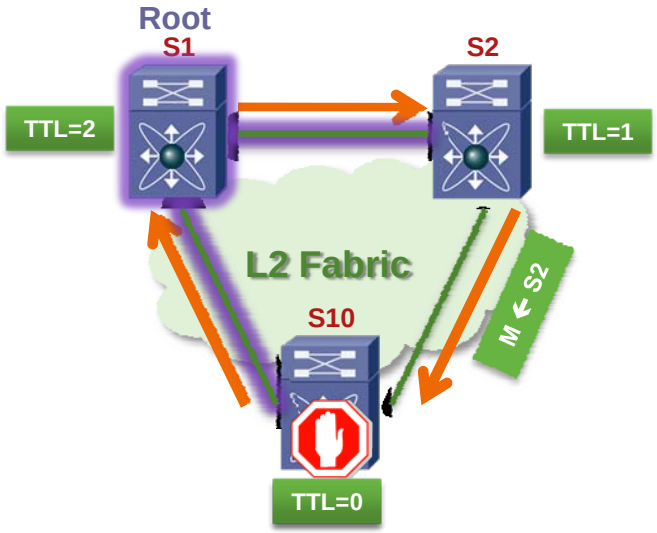


FabricPath避免迴圈機制

Time To Live (TTL) and Reverse Path Forwarding (RPF) Check



- Control protocol is the only mechanism preventing loops
- If STP fails -> infinite loop
  - •no backup mechanism in the data plane
  - •Complete network melt-down as the result of flooding



- TTL in FabricPath header
- Decrement by 1 at each hop
- Frames with TTL =0 are discarded
- RPF check for multicast based on “tree” info

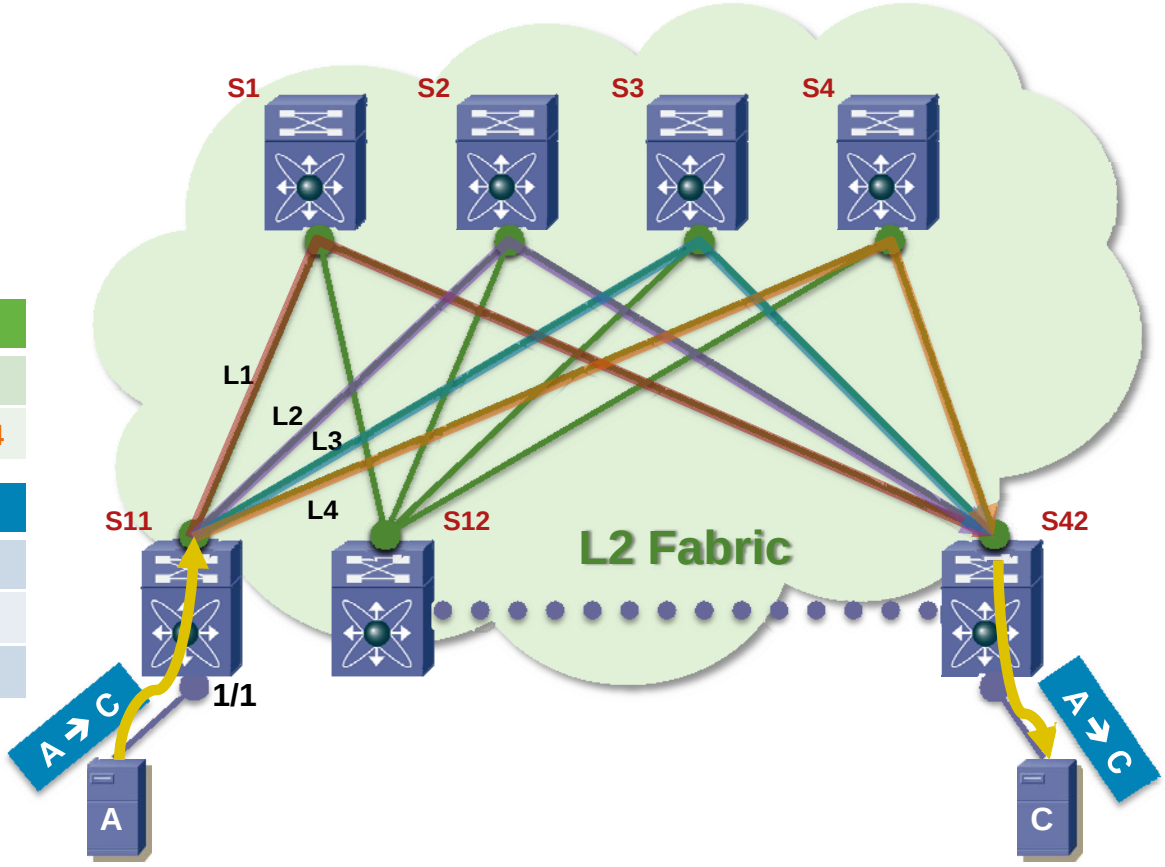
可同時支援16條路徑

Forwarding decision based on ‘FabricPath Routing Table’

- Support more than 2 paths (16 way ECMP) across the Fabric
- Increase bi-sectional bandwidth beyond port-channel
- High availability with N+1 path redundancy

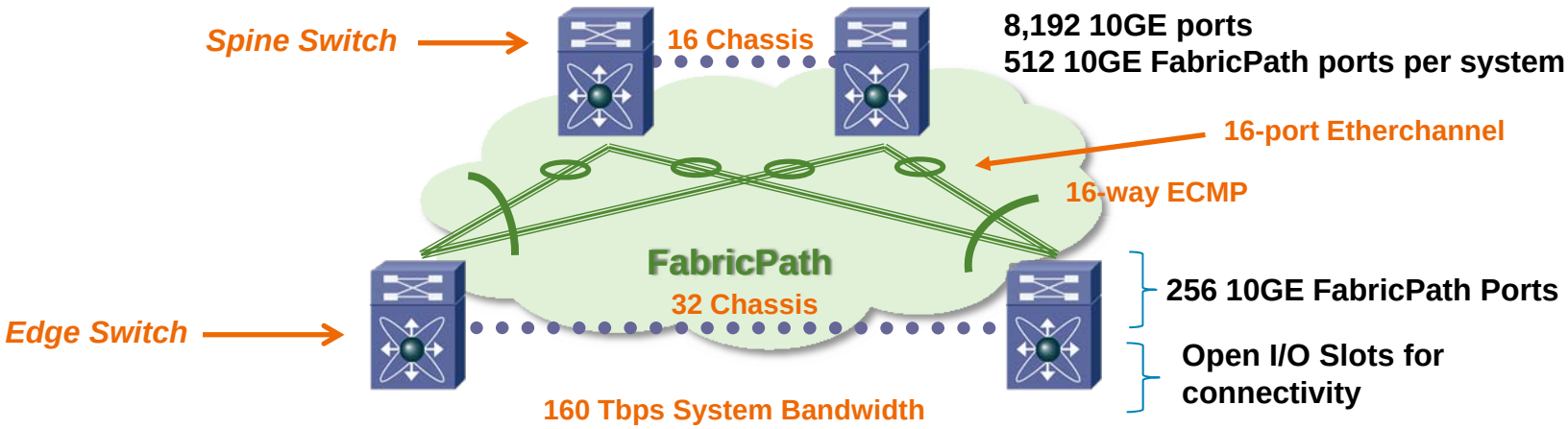
Switch	IF
…	…
S42	L1, L2, L3, L4

MAC	IF
A	1/1
…	…
C	S42



Use Case: High Performance Compute

Building Large Scalable Compute Clusters



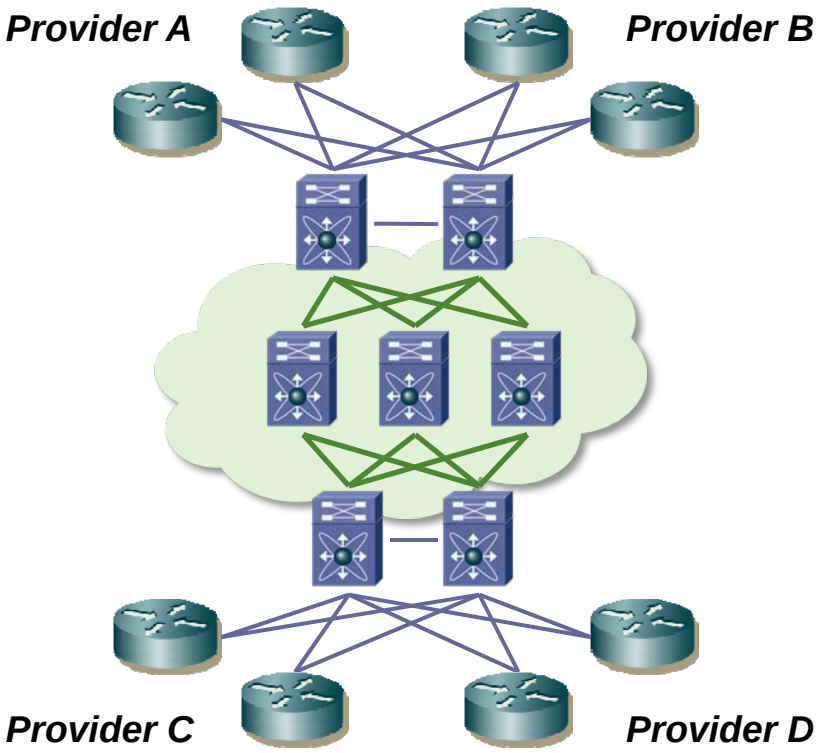
HPC Requirements

- HPC Clusters require high-density of compute nodes
- Minimal over-subscription
- Low server to server latency

FabricPath Benefits for HPC

- FabricPath enables building a high-density fat-tree network
- Fully non-blocking with FabricPath ECMP & port-channels
- Minimize switch hops to reduce server to server latencies

Use Case: L2 Internet Exchange Point



IXP Requirements

- Layer 2 Peering enables multiple providers to peer their internet routers with one another
- 10GE non-blocking fabric
- Scale to thousands of ports

FabricPath Benefits for IXP

- Transparent Layer 2 fabric
- Scalable to thousands of ports
- Bandwidth not limited by chassis / port-channel limitations
- Simple to manage, economical to build

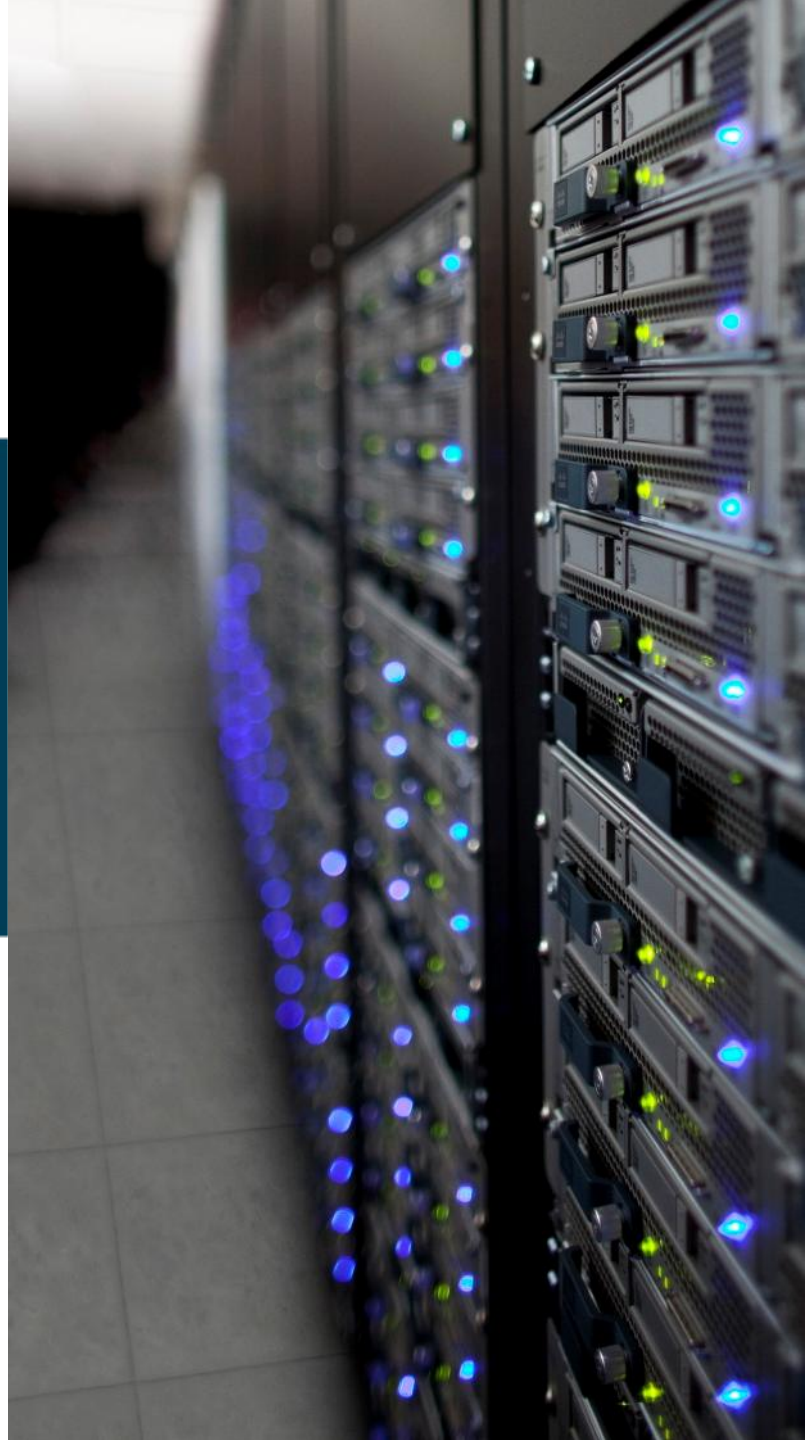
FabricPath Summary

- **FabricPath is simple**, keeps the attractive aspects of Layer 2
  - Transparent to L3 protocols
  - No addressing, simple configuration and deployment
  - A single control protocol for unicast, multicast and pruning
- **FabricPath is scalable**
  - Can extend a bridged domain without extending the risks generally associated to Layer 2 (frame routing, TTL, RPFC)
- **FabricPath is efficient**
  - High bi-sectional bandwidth (16 way ECMP with current HW)
  - Optimal path between any two nodes
  - Fast convergence

麟瑞科技

麟瑞科技
RING LINE CORPORATION

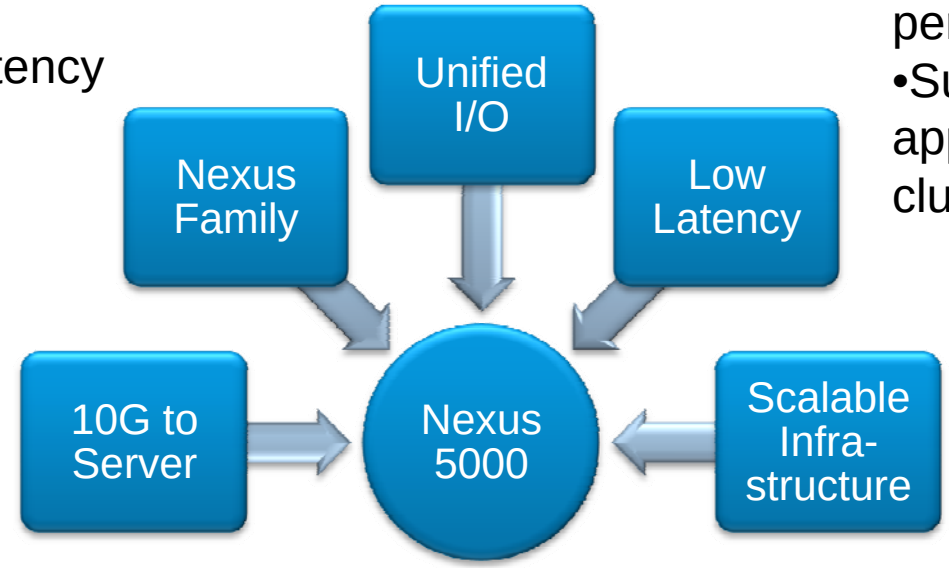
整合式網路傳輸技術
FibreChannel over Ethernet



Next-Gen Switch Design Goals

•Consolidate LAN & SAN infrasctucture
•Standards based solution
•Reduce total cost of ownership

•End-to-end data center architecture
•Operational consistency across platforms

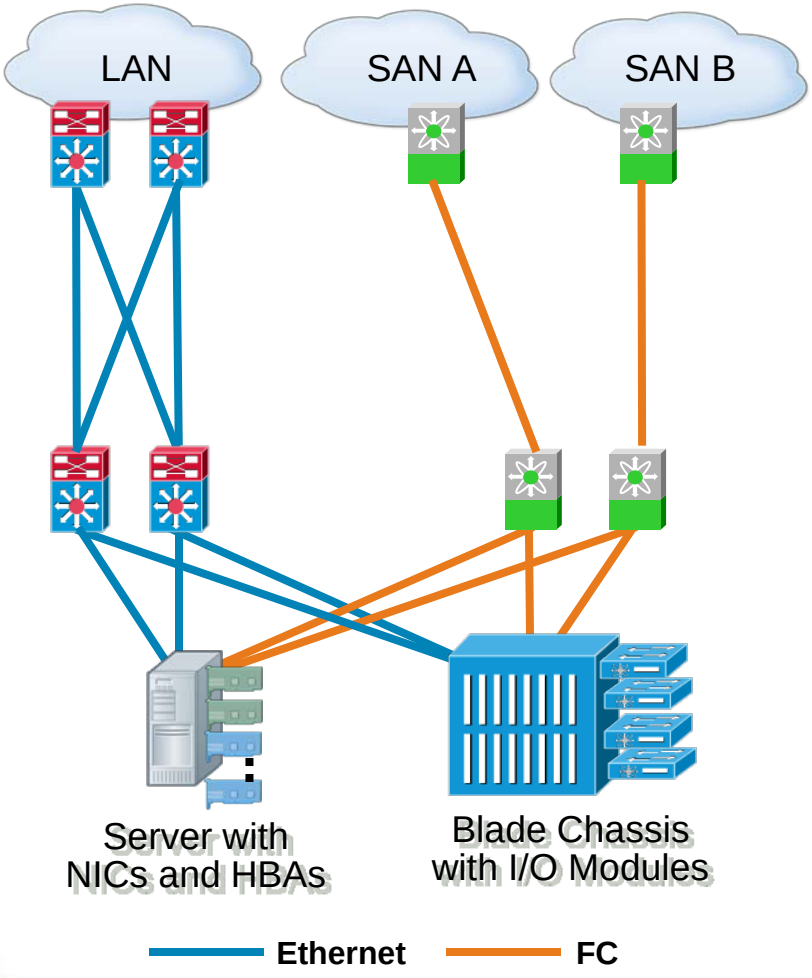


•Build with superior performance in mind
•Support low latency applications (e.g. HPC, clustered app's)

•Enable Virtualization
•Address increase in server processing power

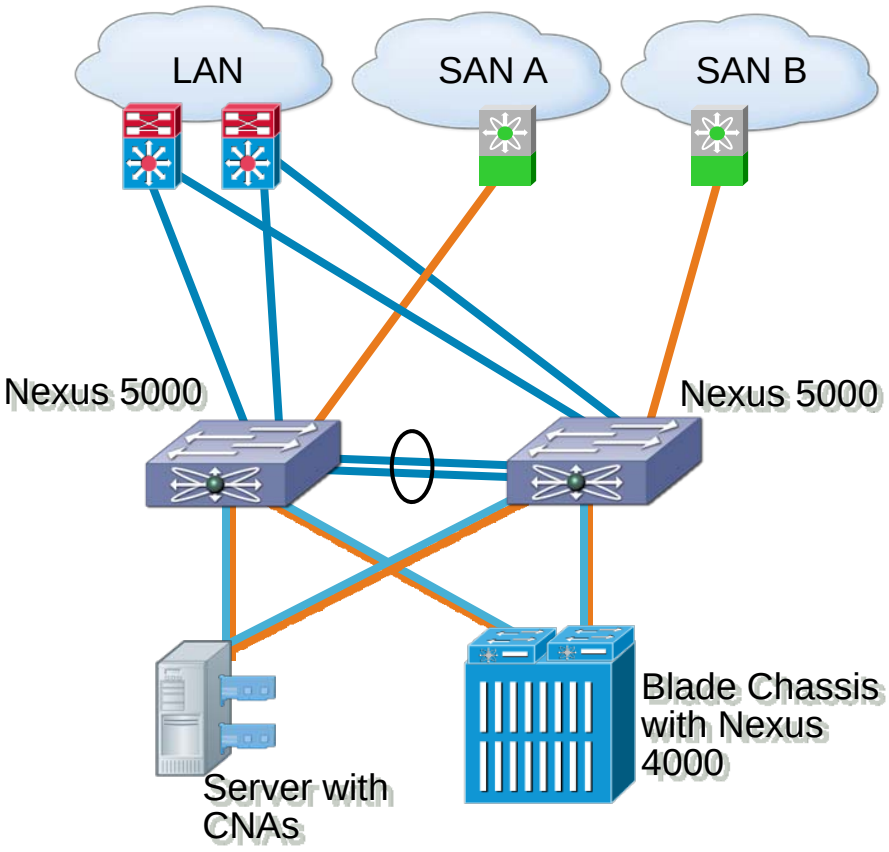
•Scale to 40G and 100G in future
•Increase feature velocity

Before I/O Consolidation



- Parallel LAN/SAN Infrastructure
- Inefficient use of Network Infrastructure
- 5+ connections per server – higher adapter and cabling costs
  - Adds downstream port costs; cap-ex and op-ex
  - Each connection adds additional points of failure in the fabric
- Multiple switching modules in Blade Chassis
- Longer lead time for server provisioning
- Multiple fault domains – complex diagnostics
- Management complexity

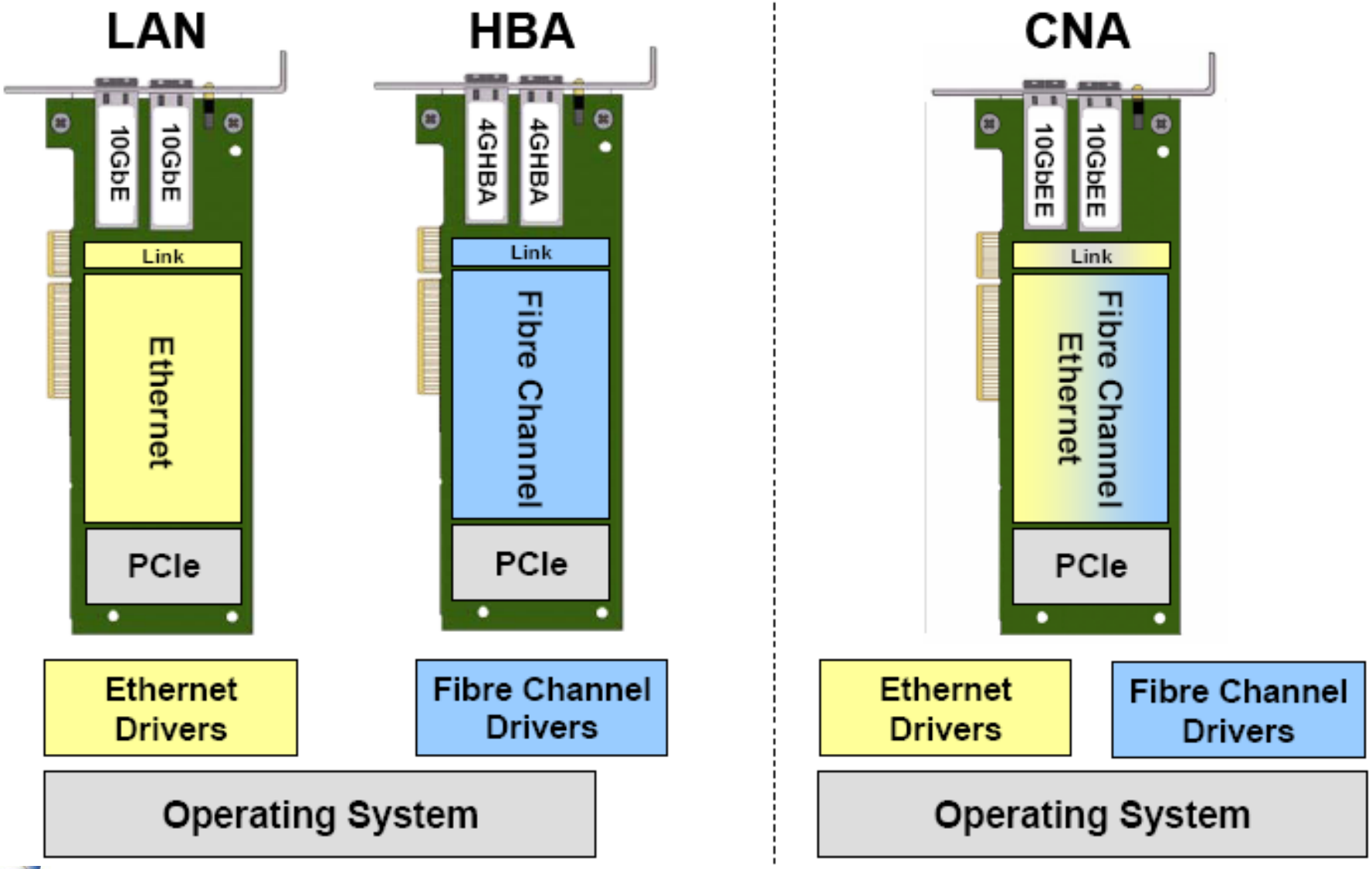
I/O Consolidation



- Reduction of server adapters
- Simplification of access layer and cabling
- Gateway free implementation – fits in installed base of existing LAN and SAN
- Lower Total Cost of Ownership
- Fewer Cables
- Investment Protection (LANs and SANs)
- Consistent Operational Model

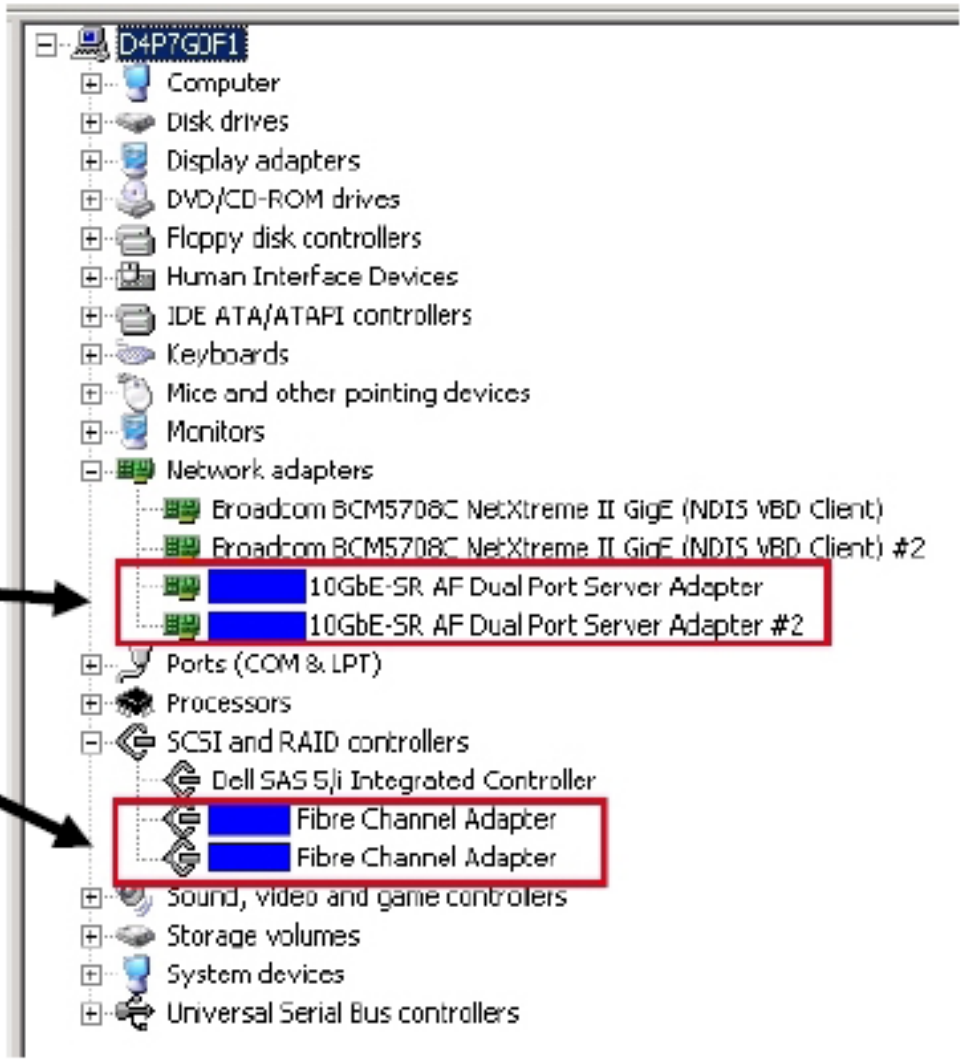
Data Center Bridging and FCoE — **Ethernet** — **Fibre Channel (FC)**

Adapter Evolution: Consolidation Network Adapter



Operating System View

- Standard drivers
- Same management
- Operating System sees:
  - Dual port 10 Gigabit Ethernet adapter
  - Dual Port 4 Gbps Fibre Channel HBAs



What is Fibre Channel over Ethernet?

- From a Fibre Channel standpoint it's

  FC connectivity over a new type of cable called… an Ethernet cloud

- From an Ethernet standpoints it's

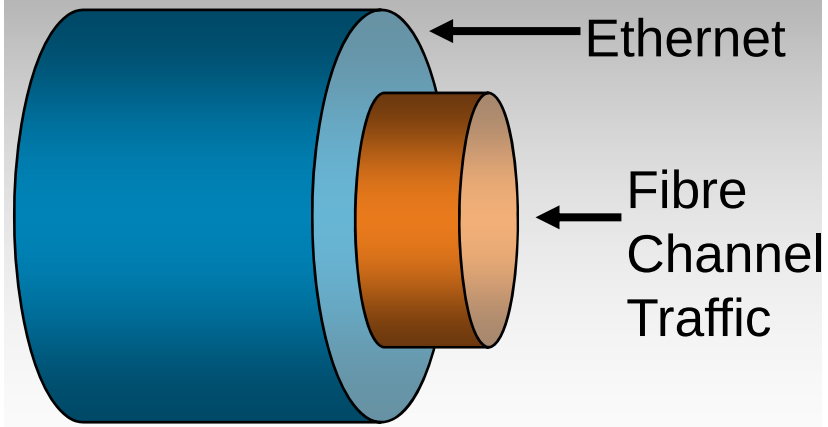
  Yet another ULP (Upper Layer Protocol) to be transported

FCoE is an extension of Fibre Channel
onto a Lossless Ethernet fabric

Unified Fabric Overview
Fibre Channel over Ethernet (FCoE)

FCoE

- Mapping of FC Frames over Ethernet
- Enables FC to Run on a Lossless Ethernet Network



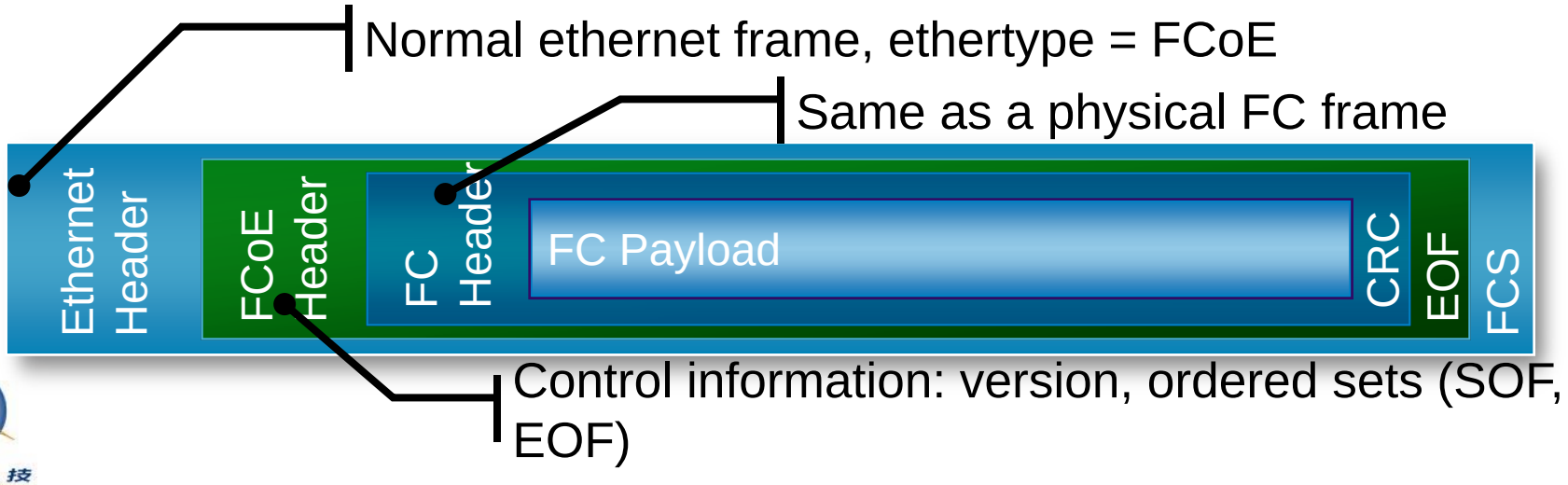
Benefits

- Fewer Cables
  - Both block I/O & Ethernet traffic co-exist on same cable
- Fewer adapters needed
- Overall less power
- Interoperates with existing SAN's
  - Management SAN's remains constant
- No Gateway

FCoE Enablers

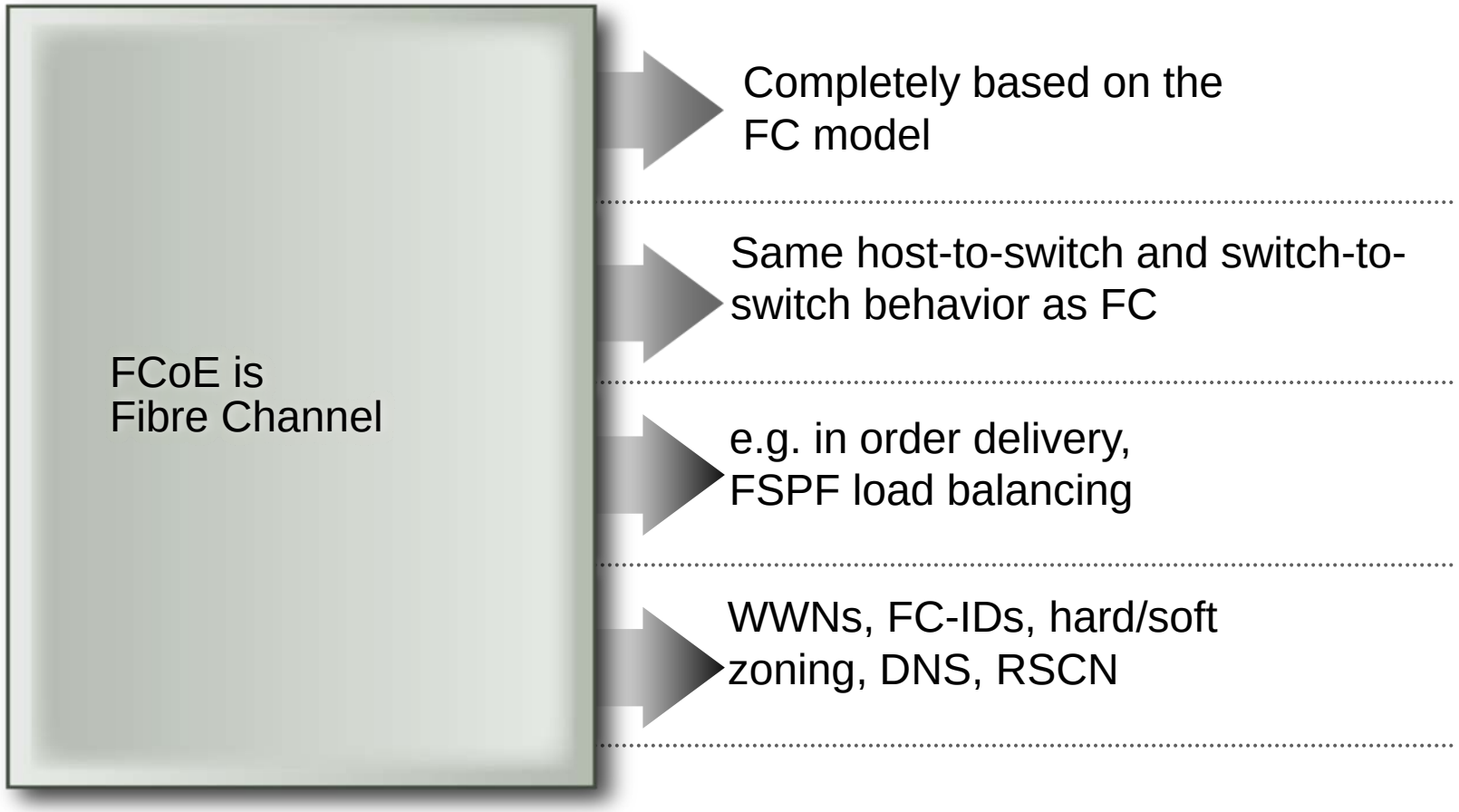
- 10Gbps Ethernet
- Lossless Ethernet

  Matches the lossless behavior guaranteed in FC by B2B credits
- Ethernet jumbo frames

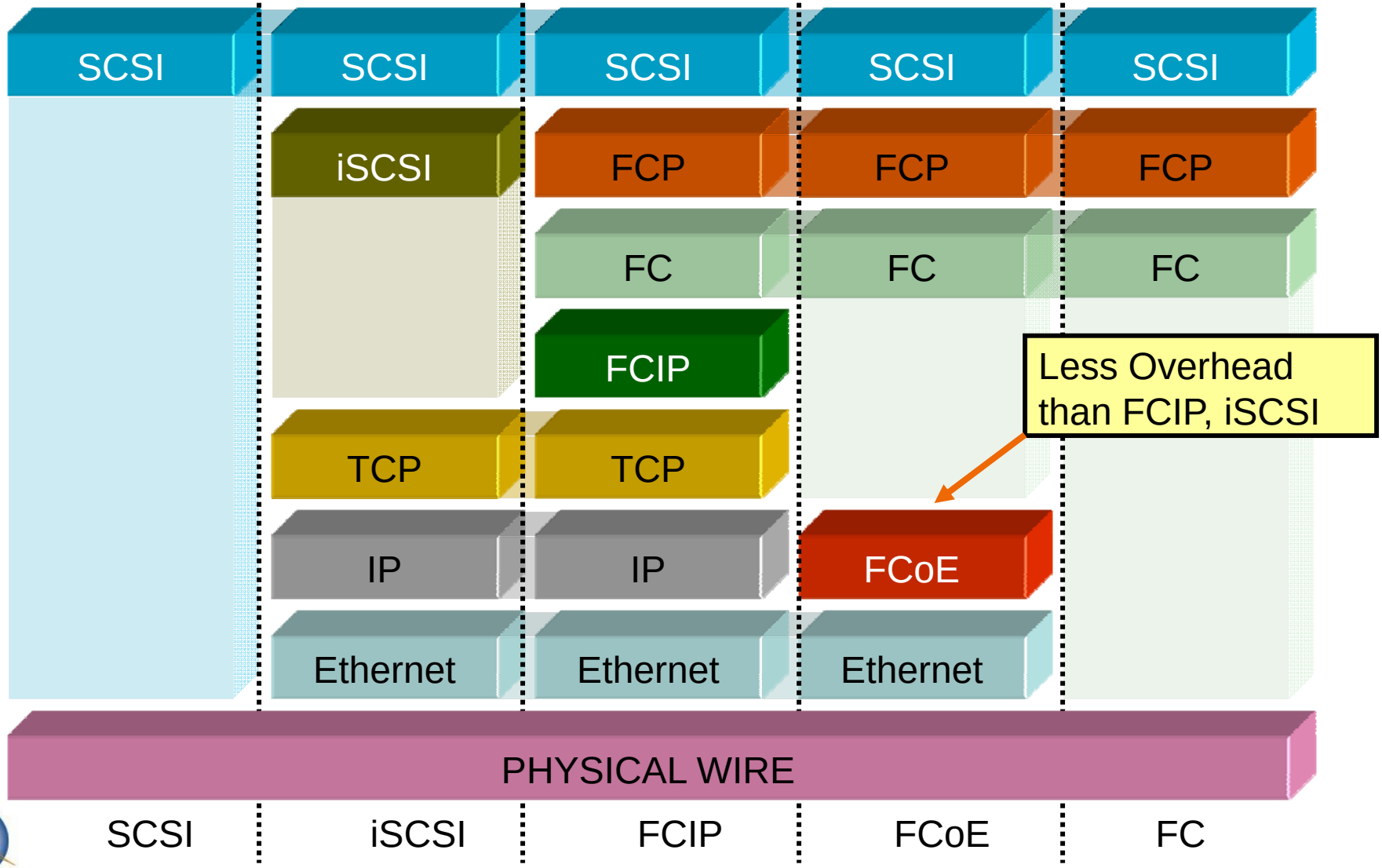


Unified I/O
Fibre Channel over Ethernet (FCoE)

FCoE is managed like FC at initiator, target, and switch level



Network Stack Comparison



A larger picture

- IEEE 802
  - •Evolution of Ethernet (10 GE, 40 GE, 100 GE, copper and fiber)
  - •Evolution of switching (Priority Flow Control, Enhanced Transmission, Congestion Management, Data Center Bridging eXchange)
- INCITS/T11
  - •Evolution of Fibre Channel (FC-BB-5)
  - •FCoE (Fibre Channel over Ethernet)
- IETF
  - •Layer 2 Multi-Path
    - •TRILL (Transparent Interconnection of Lots of Links)

麟瑞科技 RING LINE CORPORATION

What's FC-BB-5

- FC-BB-5 covers the majority of the FC features, using Ethernet
- From an Ethernet perspective, FC-BB-5 is
  - Ethernet control plane referred to as FIP (Fibre Channel over Ethernet Initiation Protocol)
    - discover and build virtual paths between end points
  - Ethernet data plane providing FCoE forwarding
    - including both FC control plane and FC data plane (FCF)

Protocol Organization

FCoE is really two different protocols:

FCoE itself …

- Is the data plane protocol
- It is used to carry most of the FC frames and all the SCSI traffic

FIP (FCoE initiation protocol)

- It is the control plane protocol
- It is used to discover the FC entities connected to an Ethernet cloud
- It is used to login to and logout from the FC fabric

The two protocols have:

- Two different Ethertypes
- Two different frame formats

麟瑞科技

IEEE DCB standards status

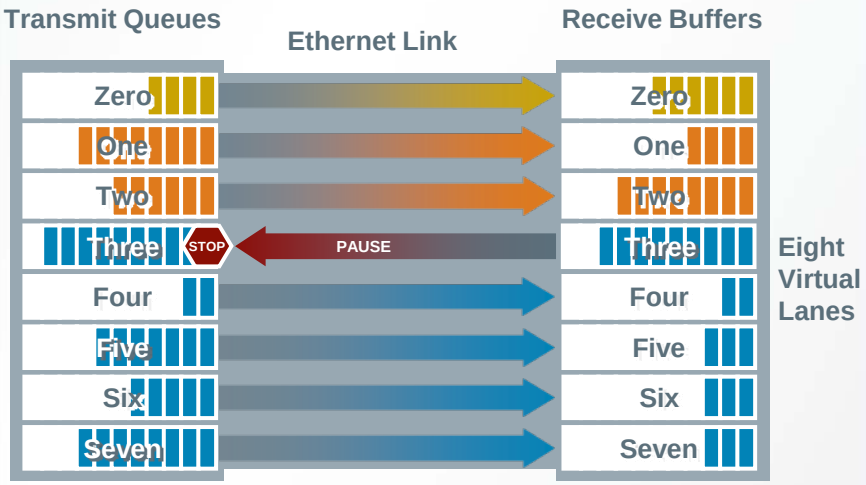
DCB technologies allow Ethernet to be lossless and to manage bandwidth allocation of SAN and LAN flows

Feature / Standard	Standards Status
IEEE 802.1Qbb Priority Flow Control (PFC) Enable multiple traffic types to share a common Ethernet link without interfering with each other	PAR approved 1.0 published
IEEE 802.1Qaz Bandwidth Management (ETS) Enable consistent management of QoS at the network level by providing consistent scheduling	PAR approved 1.0 published
Data Center Bridging Exchange Protocol (DCBX) Management protocol for enhanced Ethernet capabilities	This is part of IEEE 802.1Qaz

麟瑞科技

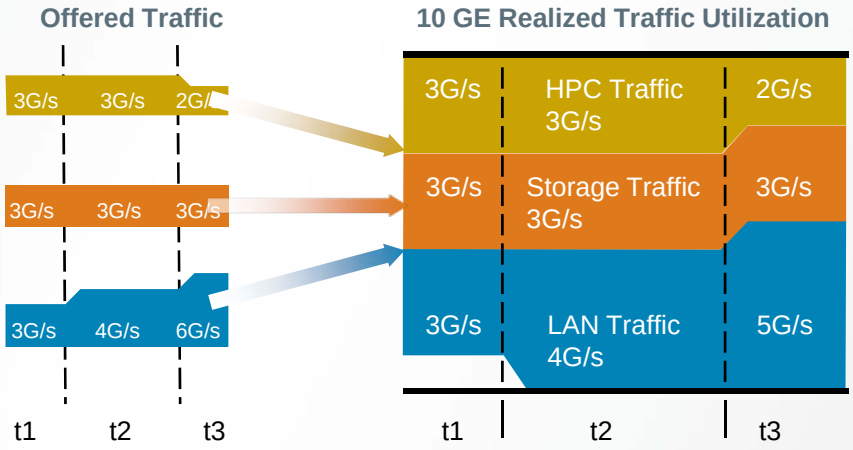
Data Center Ethernet: PFC & Bandwidth Management

Priority Flow Control



- *Enables lossless behavior for each class of service*
- *PAUSE sent per virtual lane when buffers limit exceeded*

CoS based Bandwidth Management



- *Enables Intelligent sharing of bandwidth between traffic classes control of bandwidth*
- *802.1Qaz Enhanced Transmission*

DCBX Overview

- Auto-negotiation of capability and configuration
  - Priority Flow Control capability and associated CoS values
- Allows one link peer to push config to other link peer
  - Link partners can choose supported features and willingness to accept
- Discovers FCoE Capabilities
- Responsible for Logical Link Up/Down signaling of Ethernet and FC
- DCBX negotiation failures will result in:
  - vfc not coming up
  - Per-priority-pause not enabled on CoS values with PFC configuration

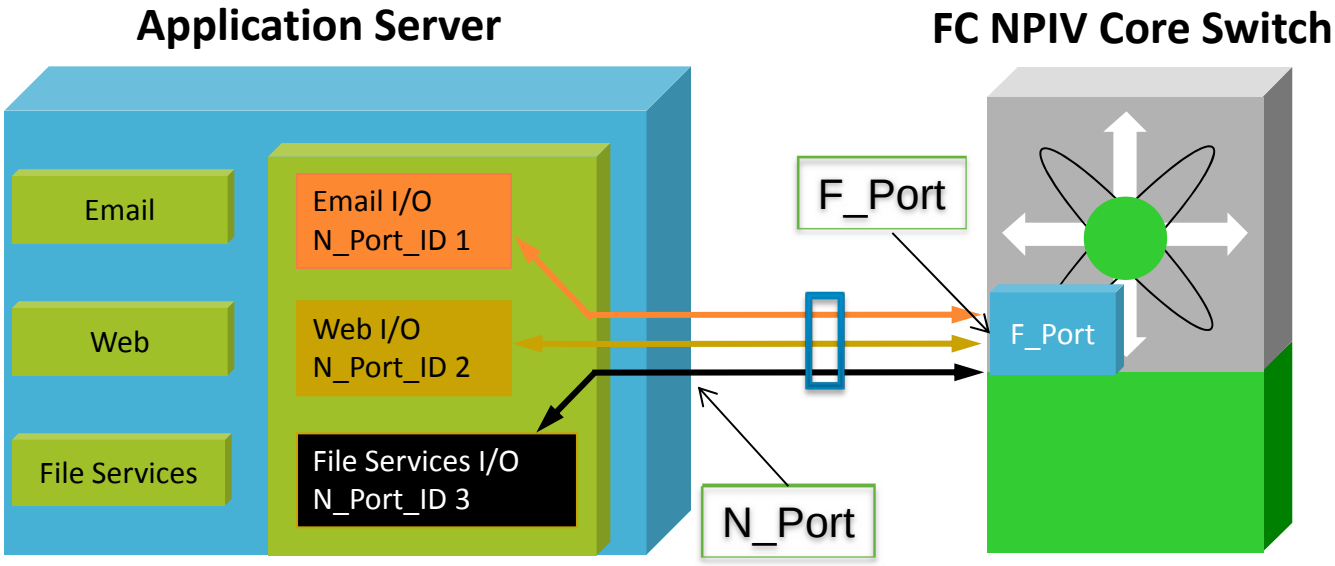
http://download.intel.com/technology/eedc/dcb_cep_spec.pdf
http://www.ieee802.org/1/files/public/docs2008/

What is NPIV? And Why?

- N-Port ID Virtualization (NPIV) provides a means to assign multiple FCIDs to a single N_Port

  Limitation exists in FC where only a single FCID can be handed out per F-port. Therefore and F-Port can only accept a single FLOGI

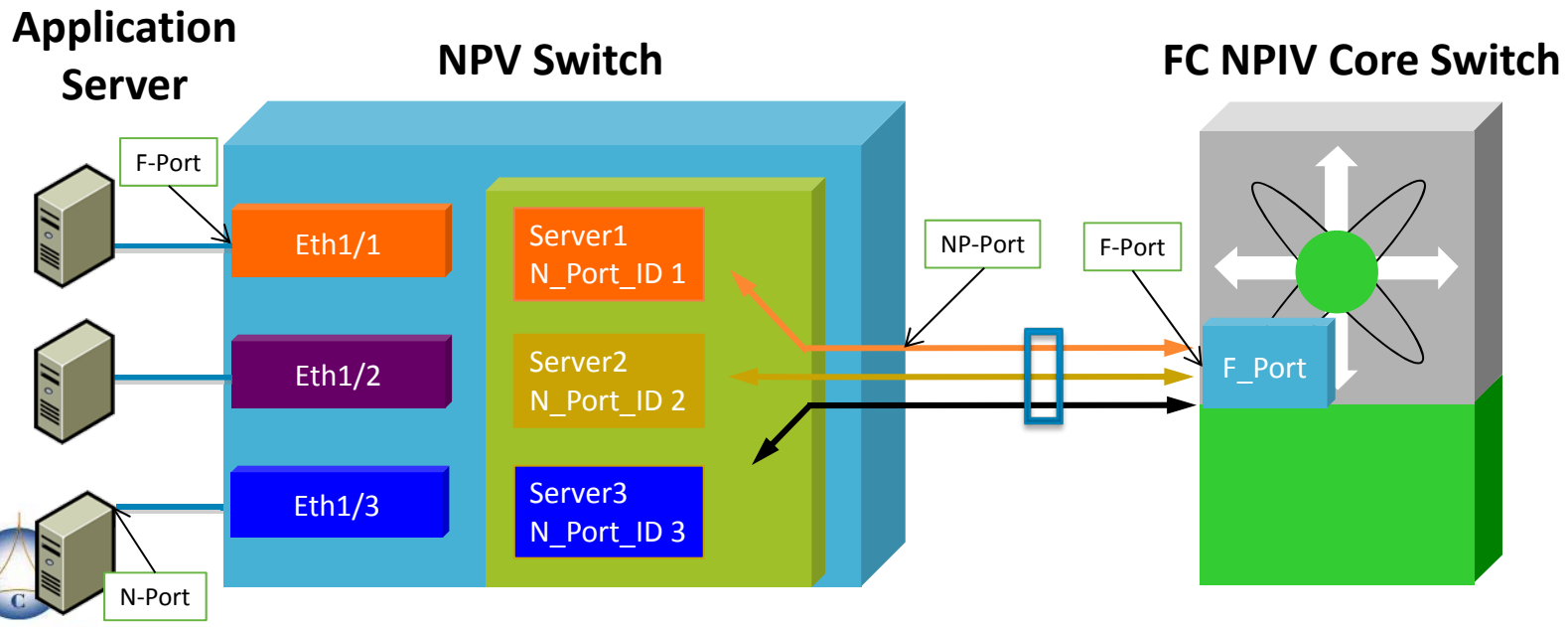
- allows multiple applications to share the same Fiber Channel adapter port
- usage applies to applications such as VMWare, MS Virtual Server and Citrix



What is NPV? And Why?

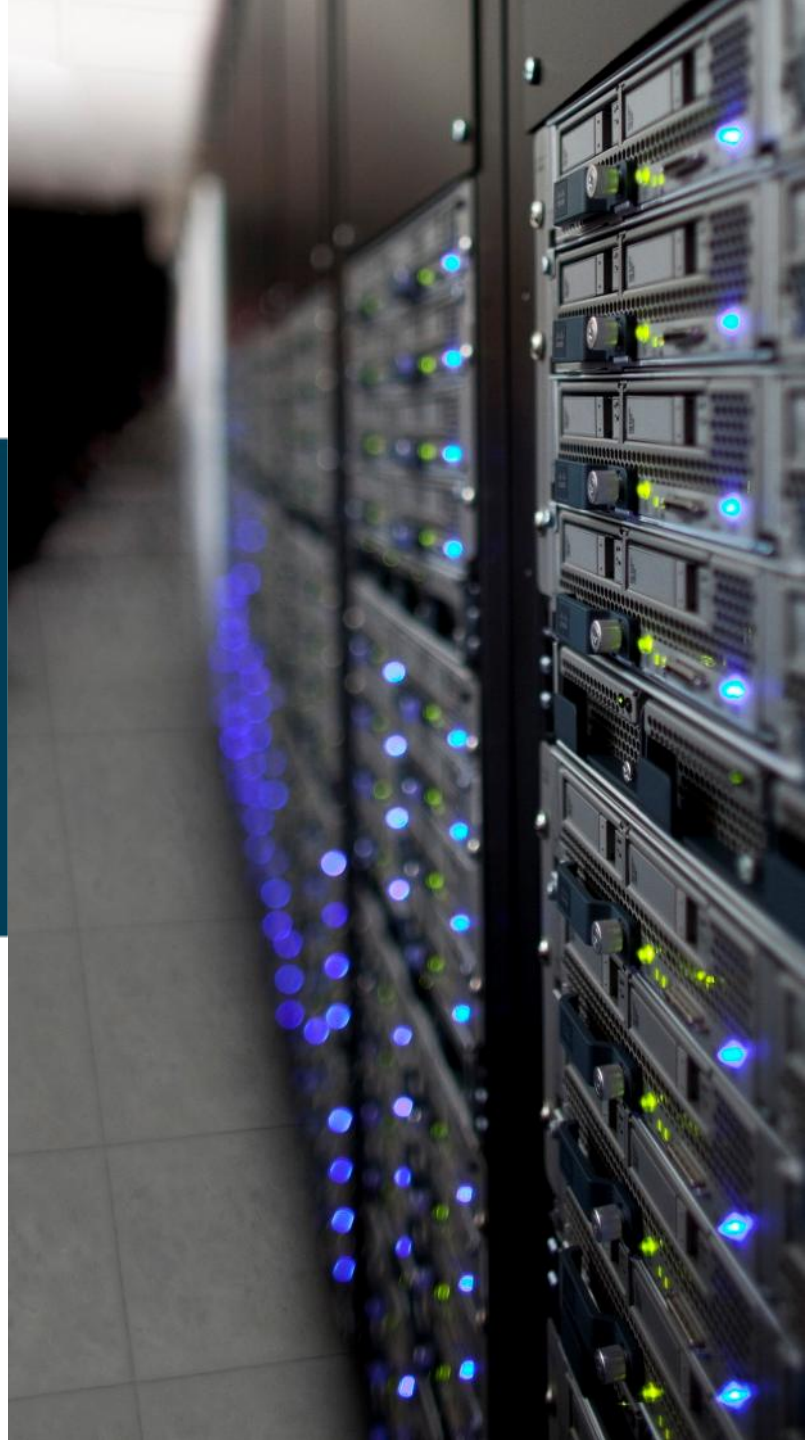
- N-Port Virtualizer (NPV) utilizes NPIV functionality to allow a “switch” to act like a server performing multiple logins through a single physical link
- Physical servers connected to the NPV switch login to the upstream NPIV core switch
- No local switching is done on an FC switch in NPV mode
- FC edge switch in NPV mode does not take up a domain ID

  Helps to alleviate domain ID exhaustion in large fabrics

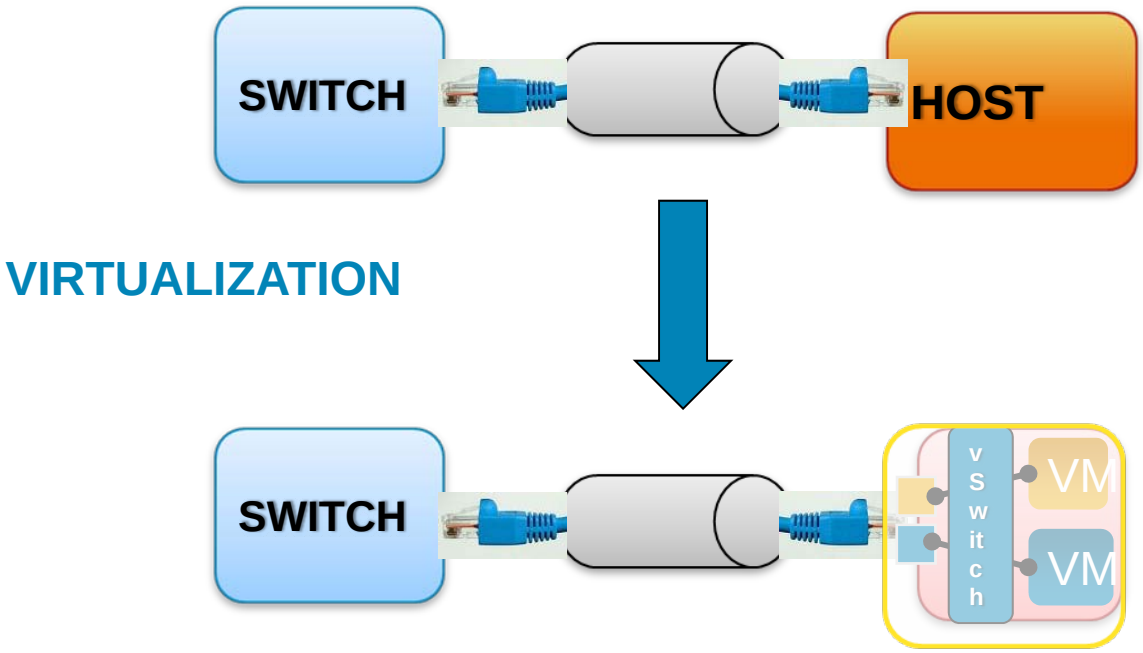


麟瑞科技
RING LINE CORPORATION

虛擬化網路交換技術 802.1Qbg與802.1Qbh



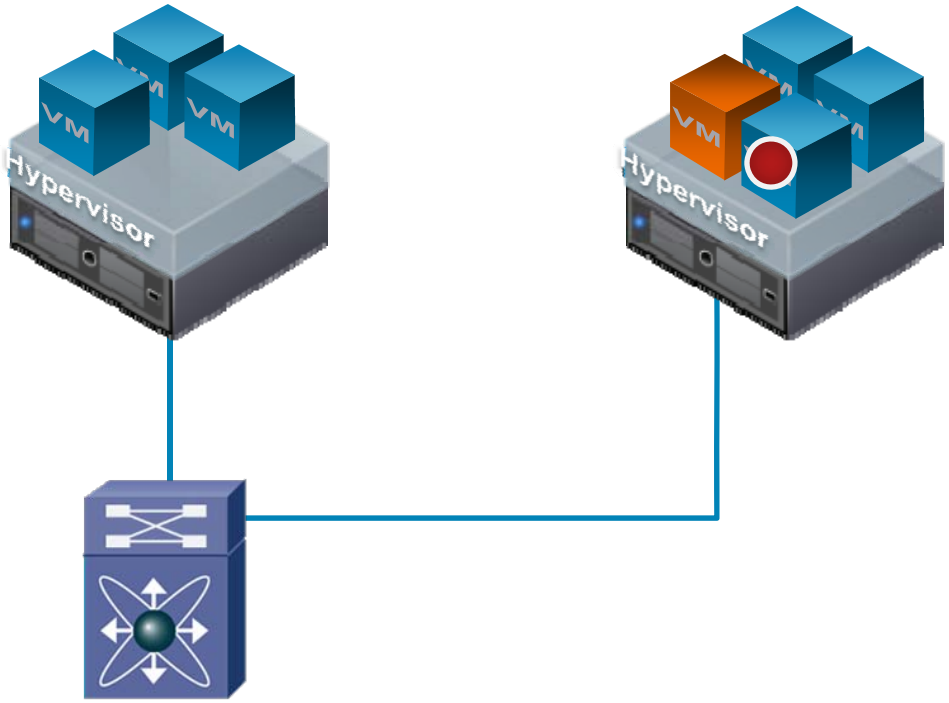
Datacenter Evolution



- Difficult to correlate network back to virtual machines
- Scaling globally depends on maintaining transparency while also providing operational consistency

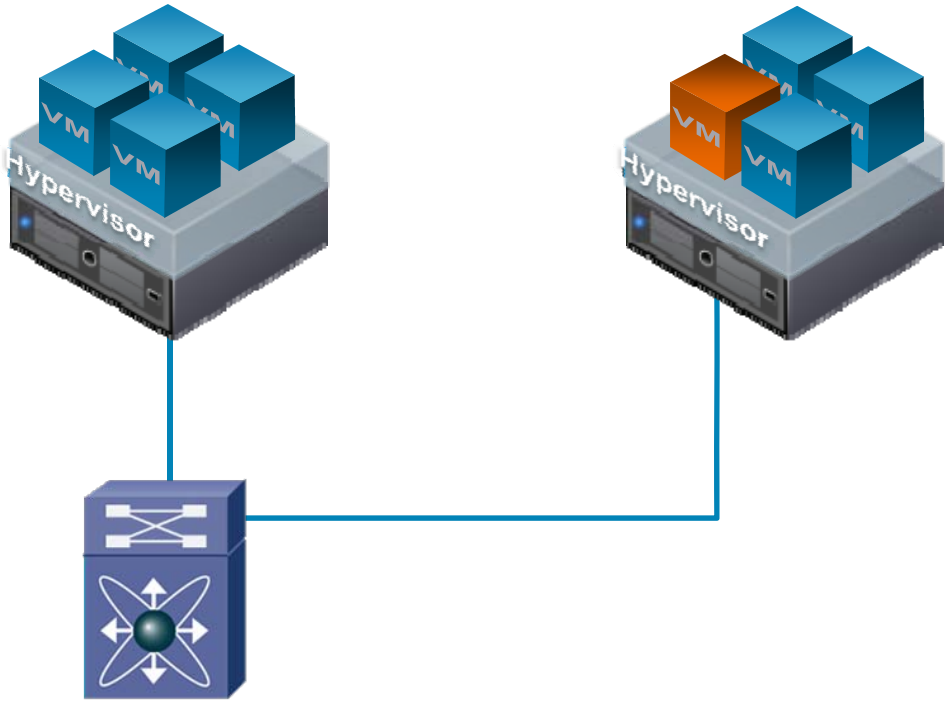
Server Virtualization Issues

1 Impossible to View or Apply Network Policy to Locally Switched Traffic

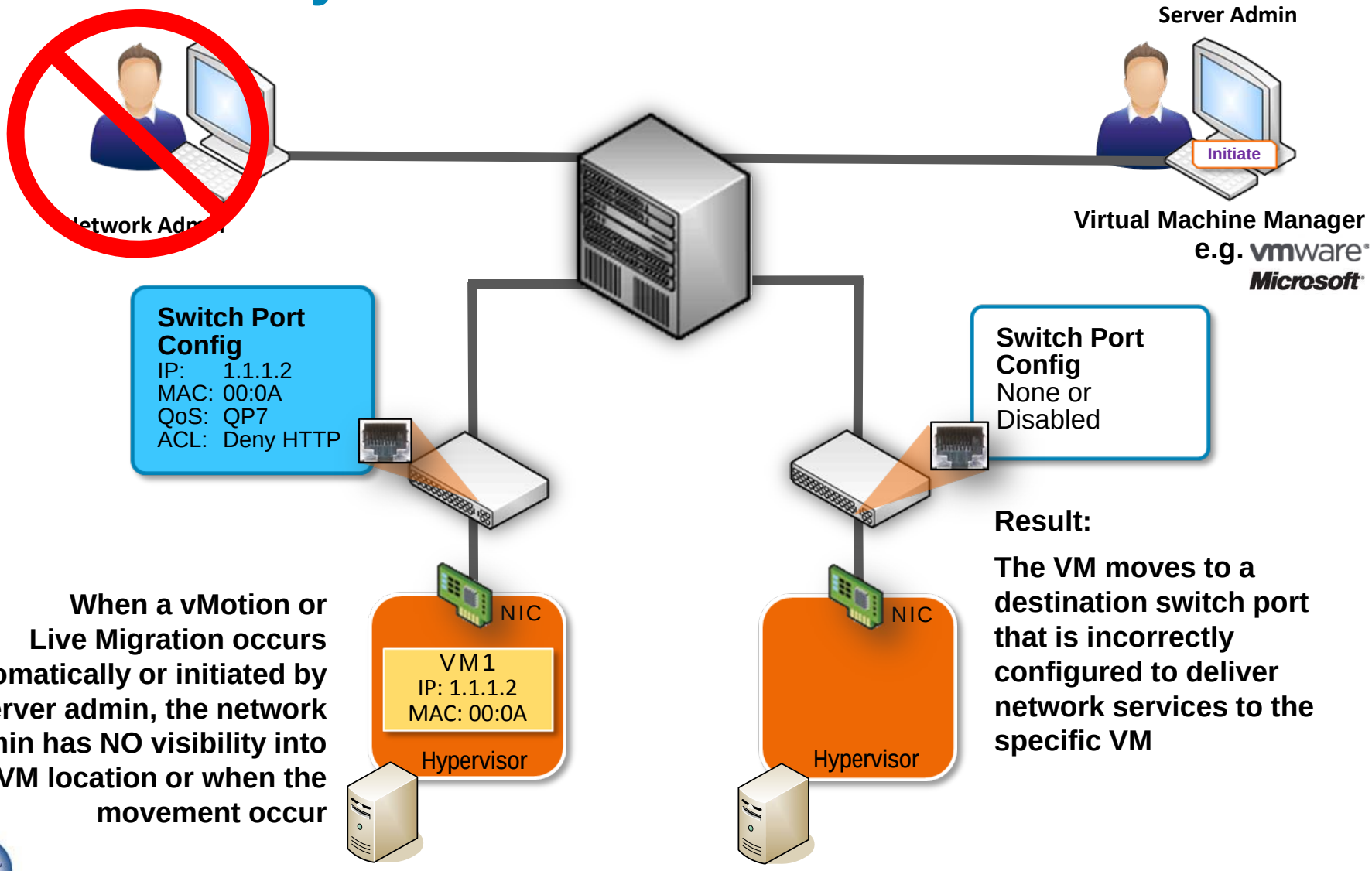


Server Virtualization Issues

2 vMotion Moves VMs Across Physical Ports—the Network Policy Should Follow



Today Network has Zero Visibility into VM Lifecycle



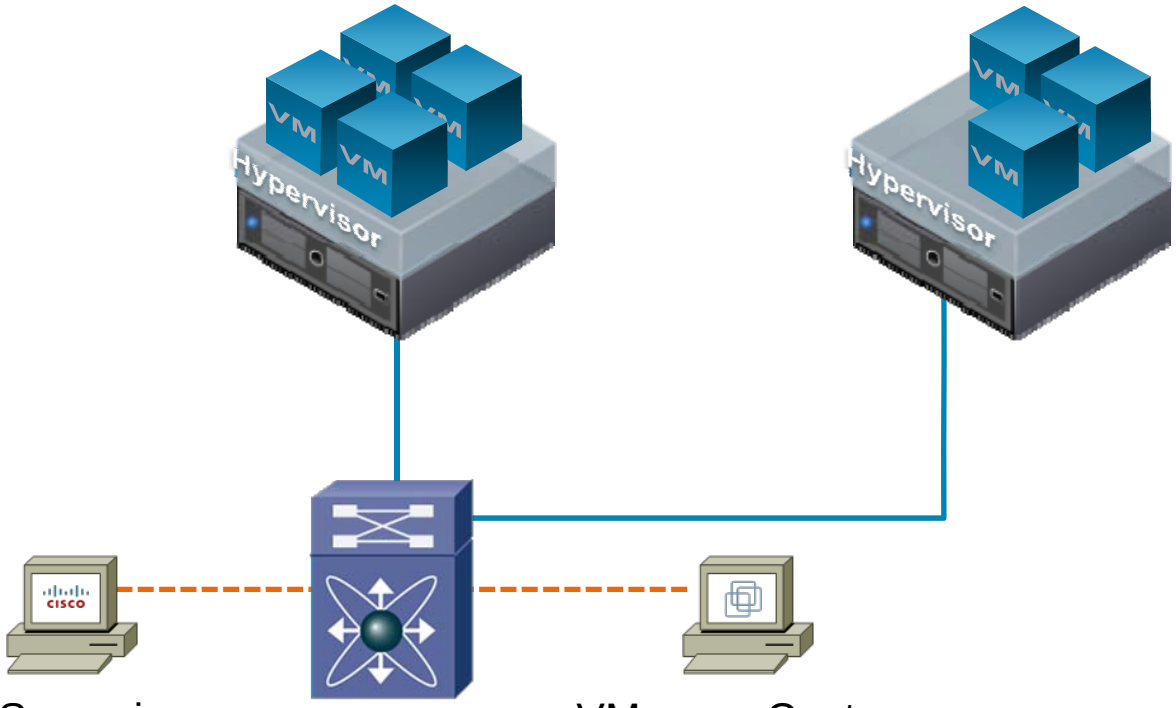
When a vMotion or Live Migration occurs automatically or initiated by server admin, the network admin has NO visibility into VM location or when the movement occur

Result:

The VM moves to a destination switch port that is incorrectly configured to deliver network services to the specific VM

Server Virtualization Issues

3 Need Shared Nomenclature Between Network Admin and Server Admin

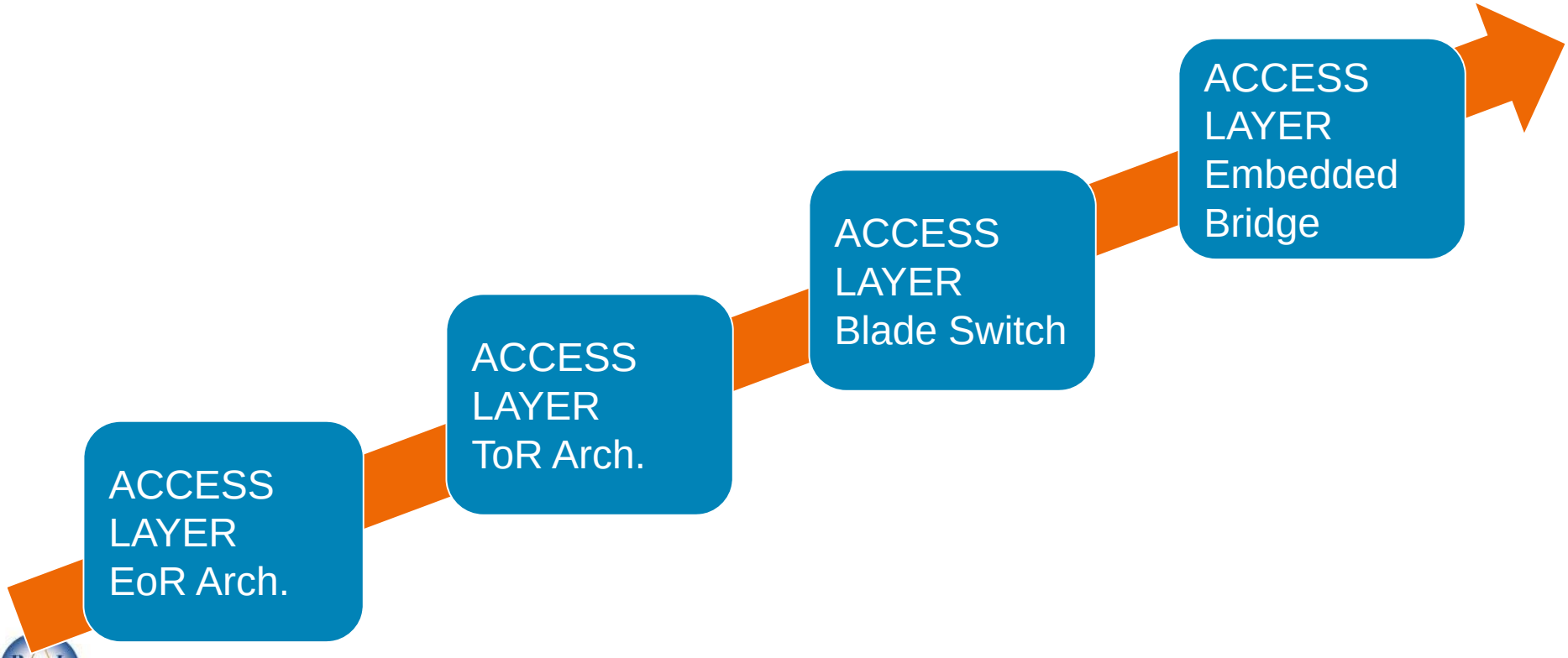


Switch Supervisor Interface

VMware vCenter Manager

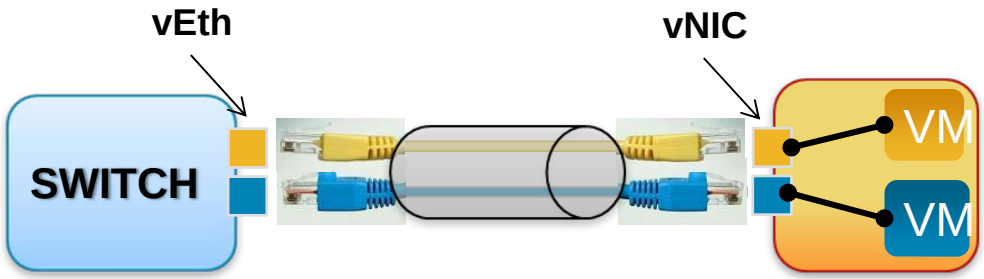
Server Virtualization Issues

4 Proliferation of Management Points & new network devices

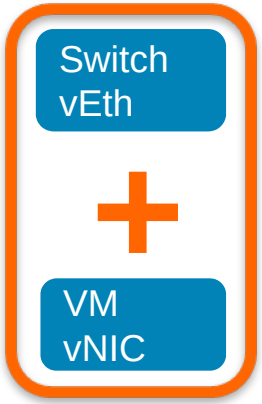


Cisco VN-Link Solution

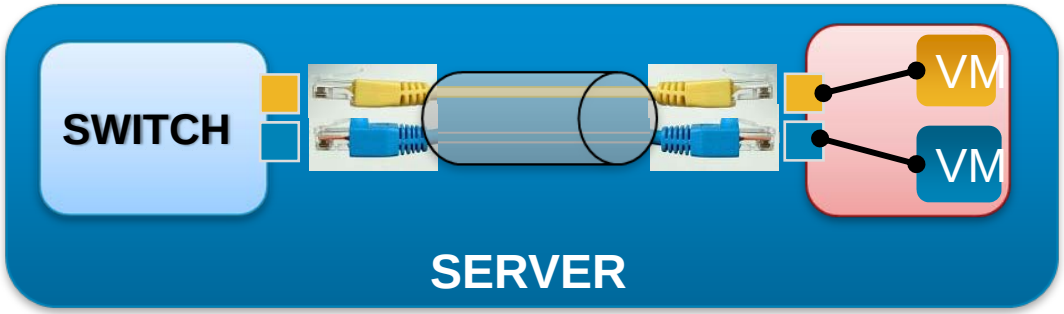
- A virtual network link between the switch and the VM
- Extends the network to the virtualization layer
- Enables:
  - Policy-Based VM Connectivity
  - Mobility of Network & Security Properties
  - Non-Disruptive Operational Model



The scope of the VN-Link

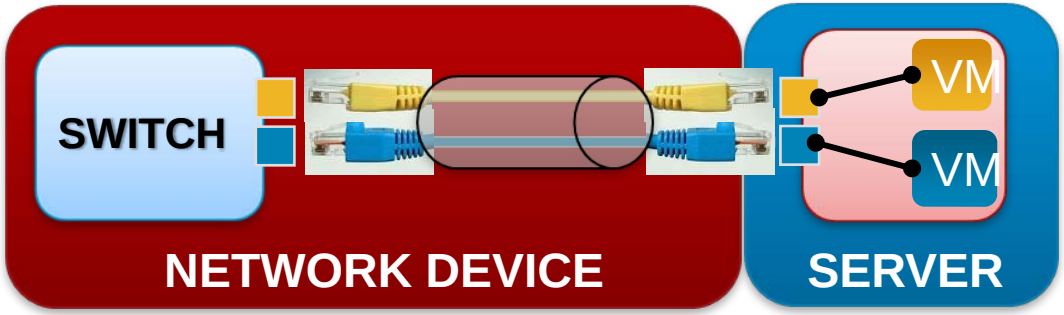


- Within the server (Hypervisor Switch)



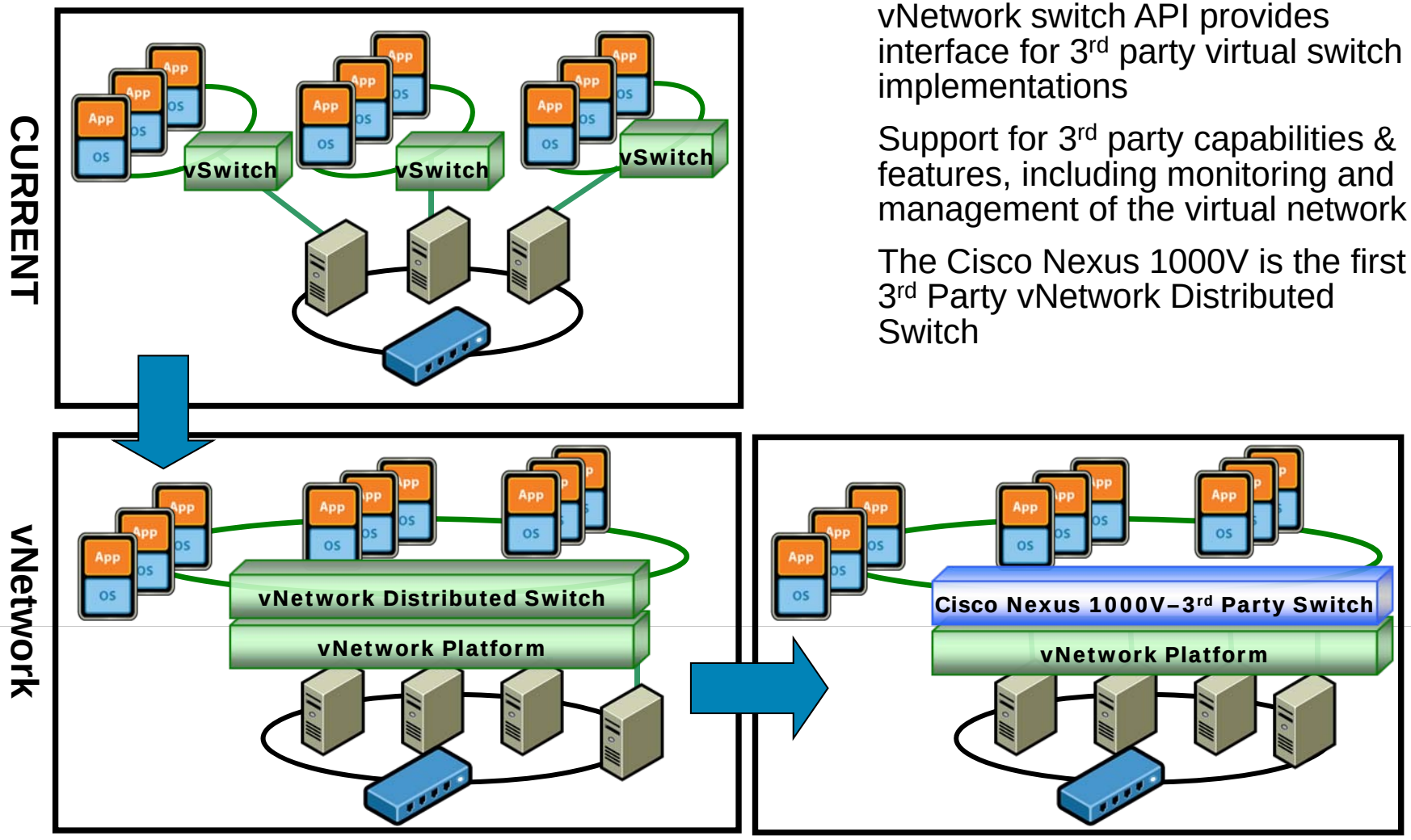
Nexus 1000V
- IEEE 802.1Q standard-based
- Rich NX-OS features

- Extending to physical upstream switch



Network Interface Virtualization (**VNTAG** Technology IEEE 802.1Qbh pre-standard)

VMware vNetwork Evolution



vNetwork switch API provides interface for 3rd party virtual switch implementations

Support for 3rd party capabilities & features, including monitoring and management of the virtual network

The Cisco Nexus 1000V is the first 3rd Party vNetwork Distributed Switch

Virtual Switch Options with vSphere 4

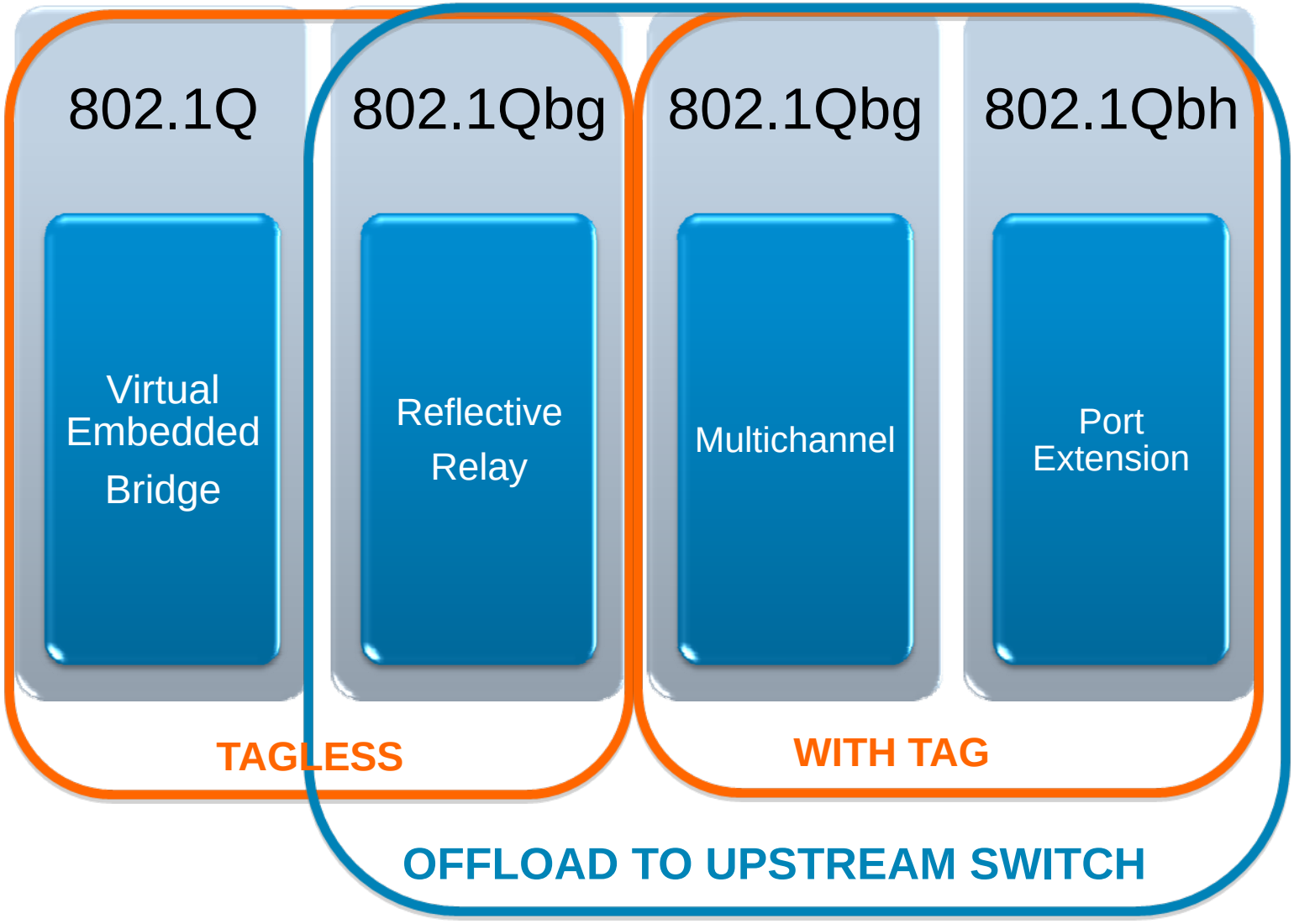
Virtual Switch	Model	Details
vNetwork Standard Switch	Host based: 1 or more per ESX host	- Same as vSwitch in ESX 3.5

vNetwork Distributed Switch	Distributed:
	1 or more per "Datacenter"
		- Expanded feature set
		  - Private VLANs
		  - Bi-directional traffic shaping
		  - Network VMotion
		- Simplified management

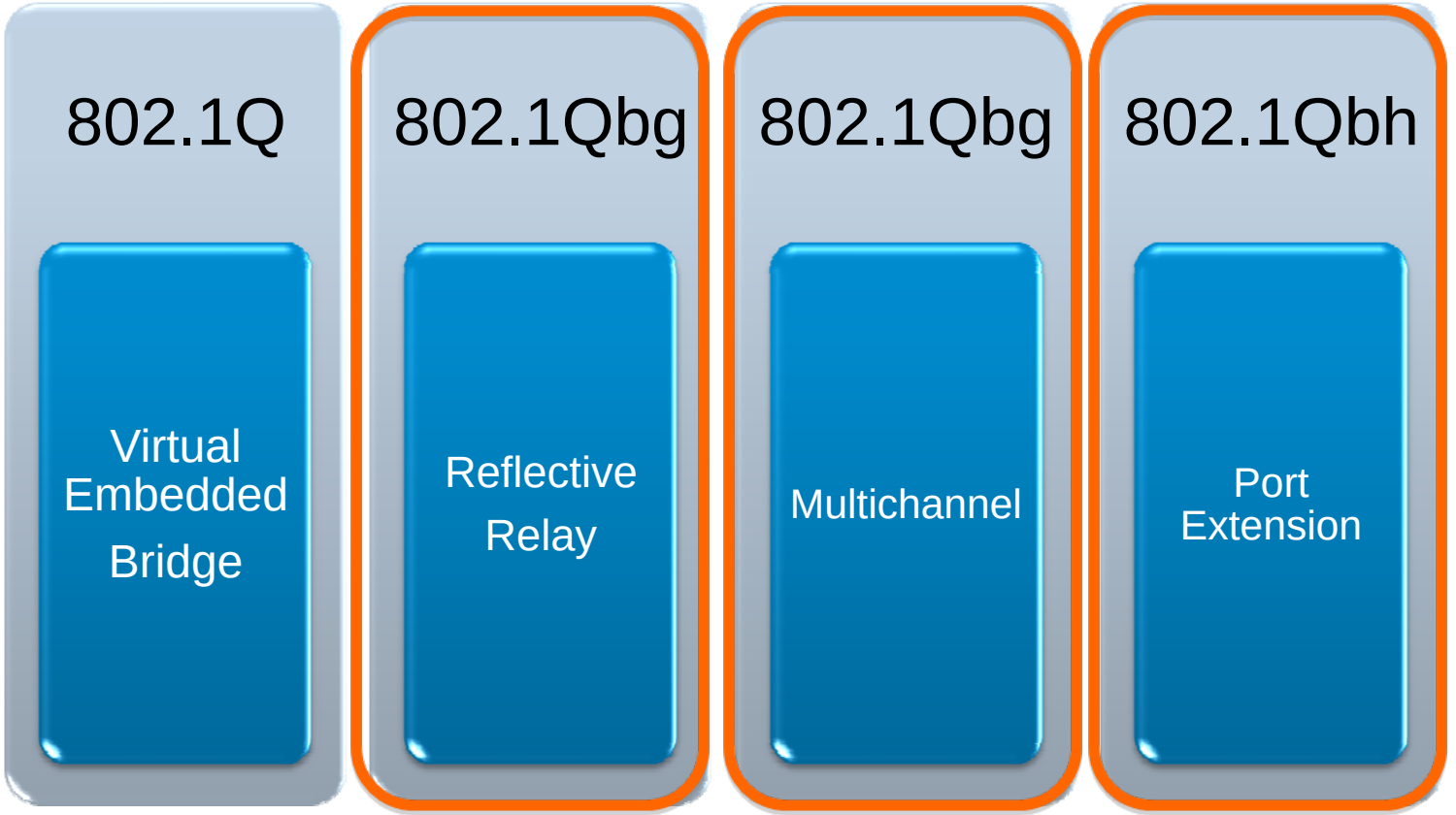
Cisco Nexus 1000V	Distributed: 1 or more per "Datacenter"	- Cisco Catalyst/Nexus feature set - Cisco IOS-like cli

Virtual networking concepts are similar with all virtual switch alternatives

Virtual Networking Standards Components



Virtual Networking Standards Components



HYPERVISOR-RESIDENT BRIDGE

NEW BEHAVIOR OF EXISTING BRIDGE

NEW BRIDGE

NEW BRIDGE NEW DEVICE

Standards-driven Approaches

	Tag-less (SW only)	Tag-based (New HW)
IEEE 802.1Q (standard)	Example: Nexus 1000V (works with existing physical servers and switches)	
IEEE 802.1Qbg (pre-standard)	Reflective Relay • New capability of external bridge (can be achieved via SW upgrade in certain existing switches)	Multi-channel (optional) • New Adapter HW • New switch HW
IEEE 802.1Qbh (pre-standard)		Port Extension • New Adapter HW • New switch HW

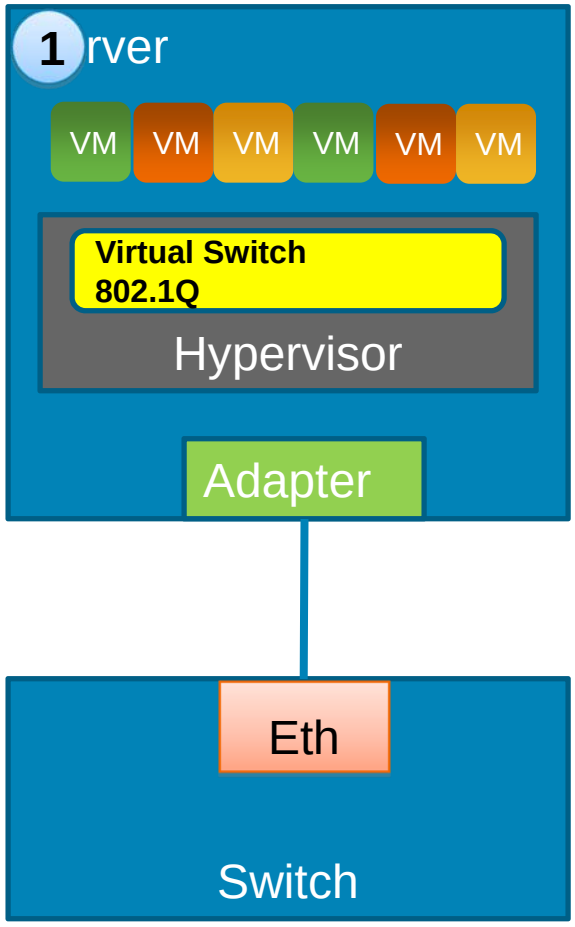
Reflective Relay is NOT "VEPA"

Basic VEPA = proprietary packet relay function in the server

Advanced VEPA = "Basic VEPA" + MC Tag function in server (Tag-based, new server & switch HW needed)

Virtual Access Layer

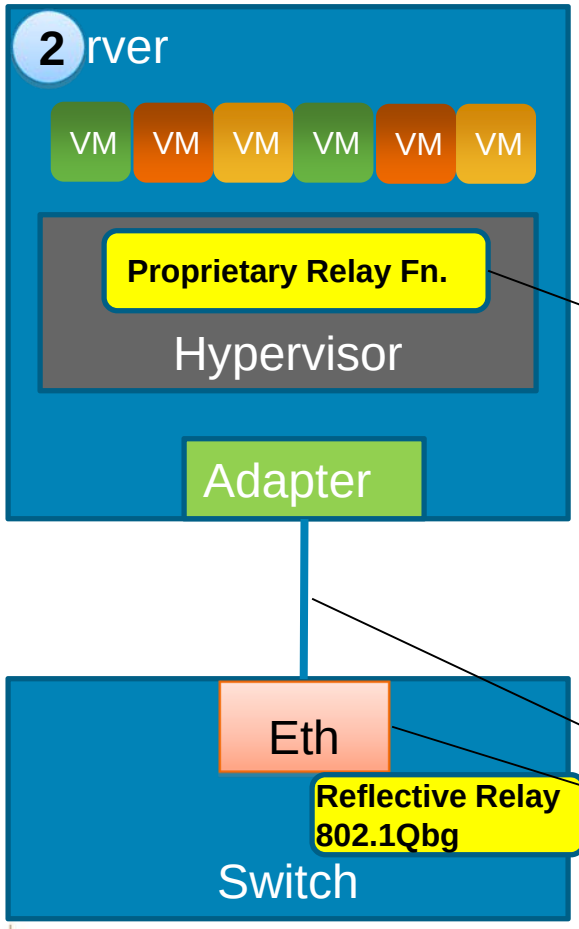
Option 1: Virtual Switch (standards based, 802.1Q)



Hypervisor Switching

	Hypervisor Vendor Switch	Nexus 1000V
Administrative Segregation	No	Yes
Visibility	No	Yes (Netflow, ERSPAN, …)
Security	No	Yes (PVLAN, ACL, …)
Mobility-transparent operation	Yes	Yes
Feature consistency	No	Yes NX-OS based
Feature richness	No	Yes NX-OS features
Policy-based provisioning	Yes (e.g. port group)	Yes (port profile)
Standards based (802.1Q)	No (e.g. SNMP MIBs, Mac learning)	Yes
Connects to any external 802.1Q bridge	Yes	Yes

Virtual Access Layer
Option 2: 802.1Qbg (Reflective Relay)



External Switching
802.1Qbg – Reflective Relay

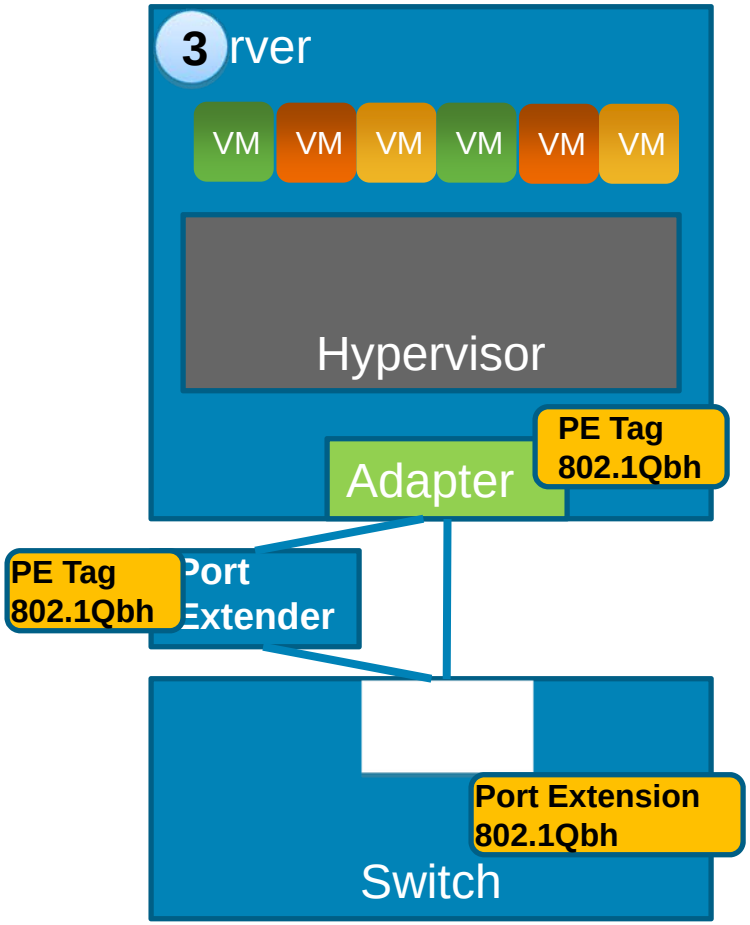
Claim:
- Leverage existing HW (with firmware upgrade)
- No impact to compute resources

Issues:
- Proprietary hypervisor module
  - Vendor dependent, Hypervisor dependent
- Classifies/switches packets – consumes CPU cycles
- Feature dispersion (across PRF and Switch)
- Increased mgmt complexity
- Mcast/Bcast replication – consumes CPU cycles
- Mcast source suppression – req. new adapter or per packet inspection by hypervisor (consumes CPU)

Issues:
- Extra utilization of link BW (out packets, in packets)
- HW/features may not scale beyond few VMs/server without HW modification of traditional switches (e.g. per VM features, such as ACLs)

PRF: Proprietary Relay Function (in hypervisor or adapter)

Virtual Access Layer
Option 3: 802.1Qbh (Port Extension)



External Switching
IEEE 802.Qbh

- **Benefits**
  - Scalable: HW supported mcast/bcast replication
  - Simplified: Manage thousands of ports centrally
  - Efficient: No impact to server CPU
- **Implications**
  - New adapter HW
  - New external switch
  - Feature velocity slower than vSwitch

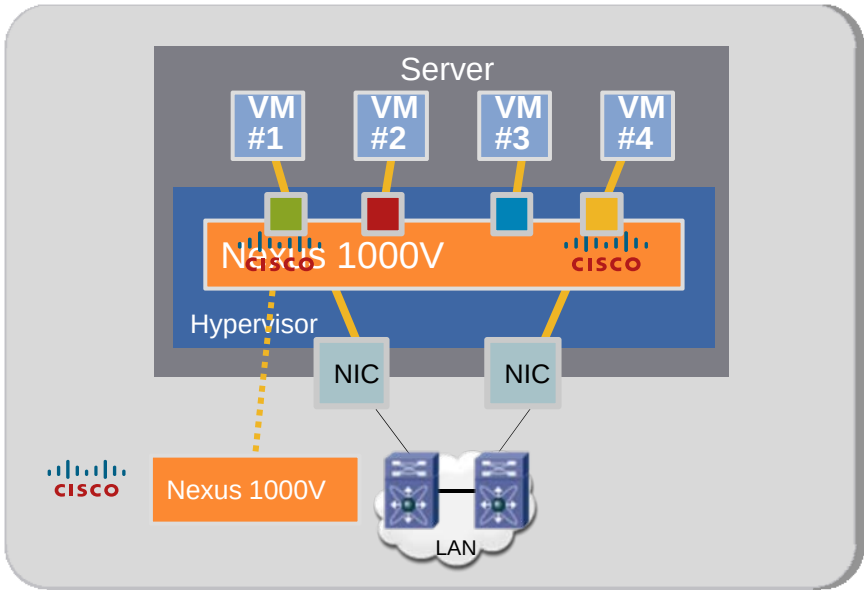
Hardware Virtualization Awareness: VN-Link for the Nexus 55x0

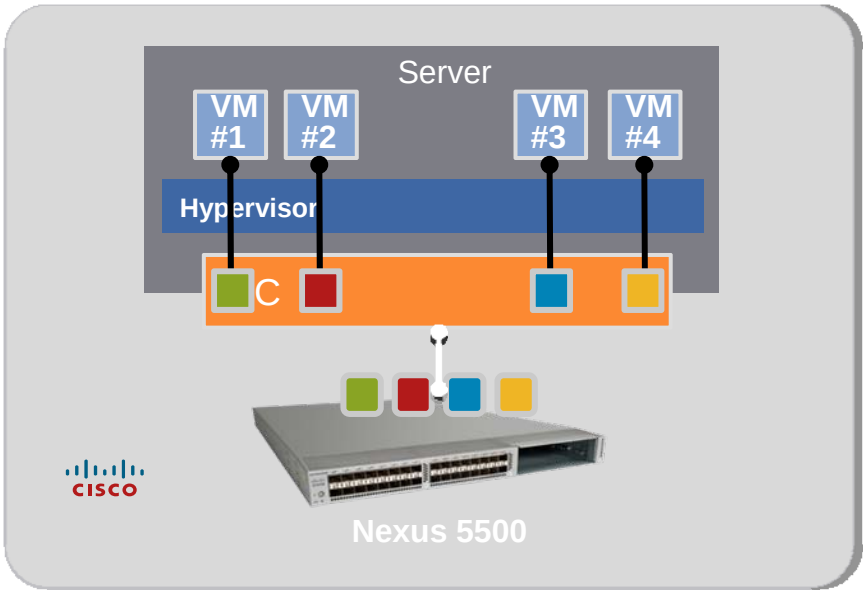
Nexus 1000V

- Software Hypervisor Switching
- Tagless (802.1Q)
- Feature set Flexibility

Nexus 55x0

- External Hardware Switching
- Tag-based (Pre-standard 802.1Qbh)
- Performance Consolidation





- Policy-Based VM Connectivity
- Mobility of Network and Security Properties
- Non-Disruptive Operational Model

互動與討論