The Evolution of Ethernet

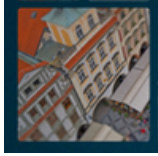
Developments, Trends and Predictions

Lincoln Dale ltd@cisco.com
Distinguished Engineer
Data Center Switching Technology Group

Cisco Systems Inc.

Agenda

- Evolving Network Topologies
  - Classical Ethernet / Spanning Tree and Link Aggregation
  - active/active infrastructure and why you should care
  - IETF TRILL (Transparent Interconnection of Lots of Links)
- Convergence of LAN and SAN
  - IEEE DCB (Data Centre Bridging)
  - INCITS T11 (ANSI) FCoE (Fibre Channel over Ethernet)
  - IEEE 802.1Qbb PFC (Priority Flow Control)
  - IEEE 802.1Qaz ETS (Enhanced Transmission Selection)
  - IEEE 802.1AB DCBXP (DCB Exchange Protocol)
- Server Virtualization / Virtual Machine awareness
  - 1G to 10G and Physical Cabling trends
  - Impact on network elements and implementing per-VM policies
  - Demands on the network for DR across data centers



cisco

Evolving Network Topologies

Classical Ethernet / Spanning Tree and Link Aggregation

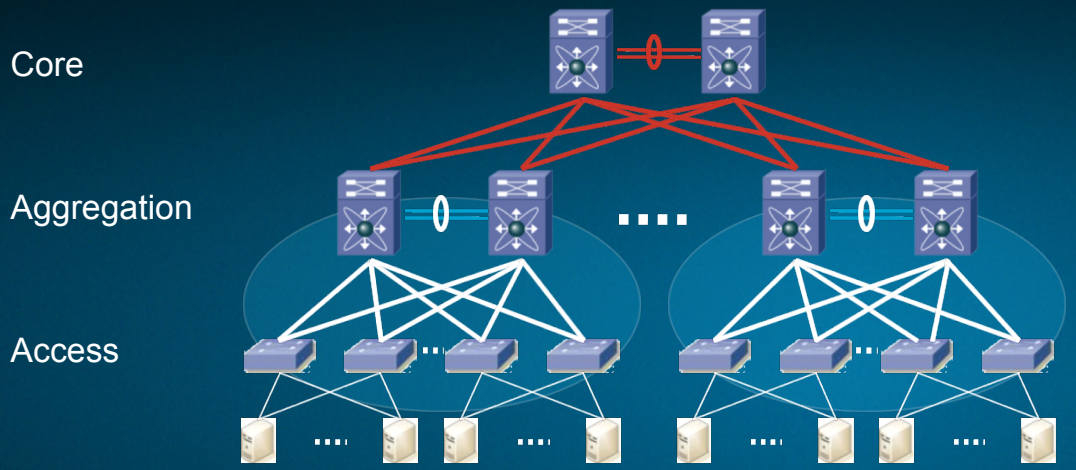
active/active infrastructure and why you should care

IETF TRILL (Transparent Interconnection of Lots of Links)

cisco

3-Tier Data Center Design

Not an L2 vs. L3 debate



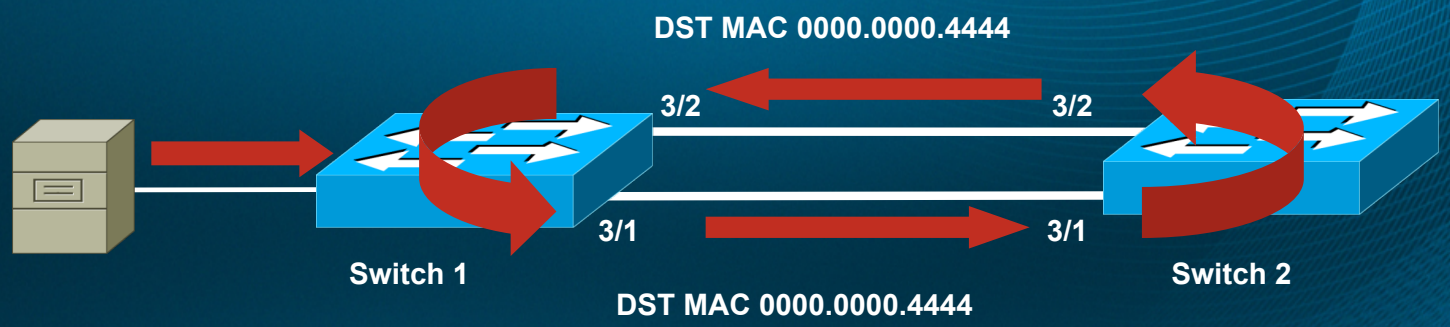
- L2 Access Layer enables higher scaleability and functionality than what extending L3 to the access edge can provide.
- 3 tiers can be 2 tiers depending on overall scale/size. Core/Agg can be combined, as can Agg/ Access

Layer	Switch Type	Port Speed	Configuration	Oversubscription	Other
Core	Modular	10GE	Layer 3 only	Low to medium	Campus hand off
Aggregation (Distribution)	Modular	10GE	L2/L3 boundary	Medium to high	Services (optional)
Access	Fixed or Modular	GE/10GE	Layer 2 only	Medium to high	ToR, MoR, Blade Switch

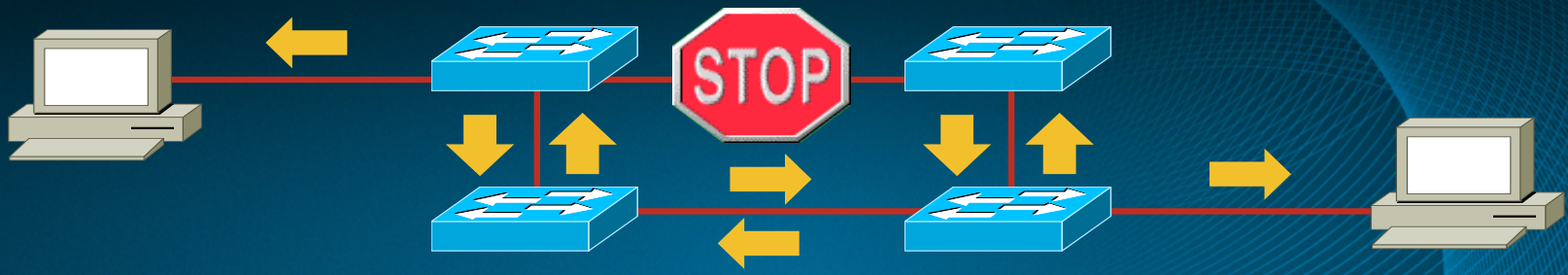
Table values are considered “typical” for a green field deployment

Spanning Tree (STP) – Why?
To prevent Loops at L2

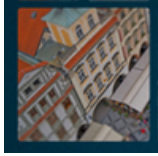
- Layer 2 topologies have sometimes proven a operational or design challenge
- Spanning tree protocol itself is not usually the problem, it's the external events that triggers the loop or flooding
- L2 has had no native mechanism to dampen down a problem and no solution to provide link redundancy other than STP
- STP is there to protect against loops in the network.



Spanning Tree Standards and Features
IEEE 802.1D, IEEE 802.1s, IEEE 802.1w



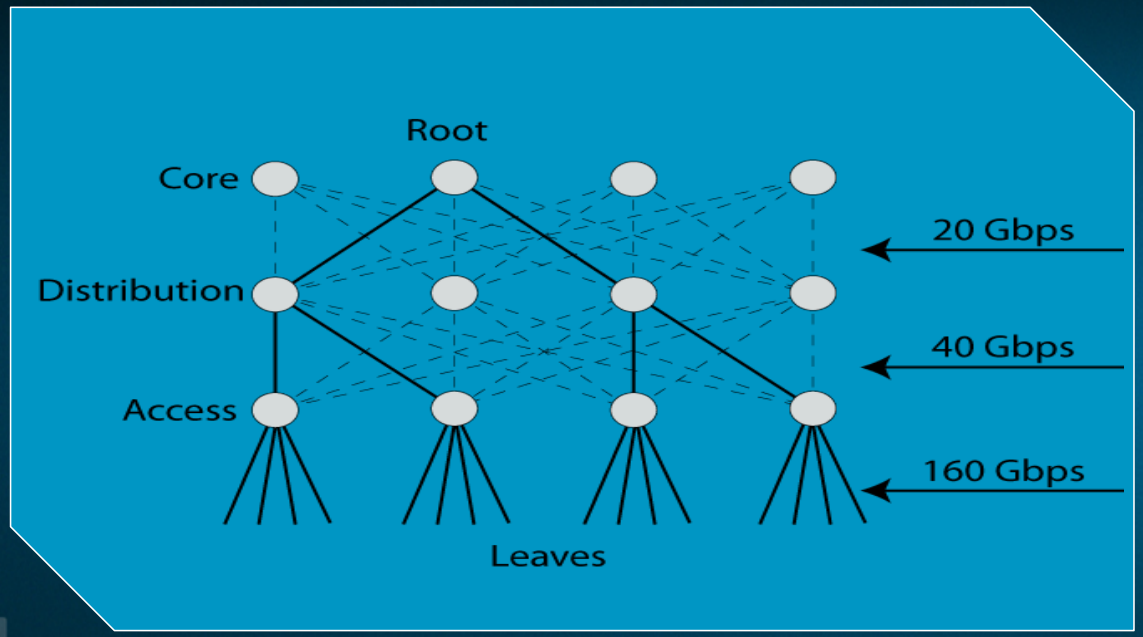
- **802.1D/1998**: legacy standard for bridging and Spanning Tree (STP)
- **802.1D/2004**: updated bridging and STP standard; includes 802.1s, 802.1t, and 802.1w
- **802.1s**: Multiple Spanning Tree Protocol (MSTP)—maps multiple VLANs into the same Spanning Tree instance
- **802.1t**: MAC address reduction/extended system ID—moves some BPDU bits to high-numbered VLANs from the priority field, which constrains the possible values for bridge priority; unique “MAC” per chassis not port
- **802.1w**: Rapid Spanning Tree Protocol (RSTP)—improved convergence over 1998 STP by adding roles to ports and enhancing BPDU exchanges
- **Cisco Features**: Per VLAN Spanning Tree (PVST), PVST+, UpLinkFast, BackboneFast, BPDU Guard, RootGuard, LoopGuard, Bridge Assurance, UDLD



cisco

Spanning Tree

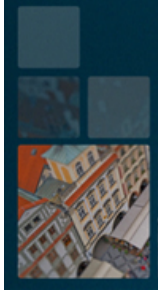
Network Reduced to a Simple Tree



Algorhyme

I think that I shall never see
a graph more lovely than a tree.
A tree whose crucial property
is loop-free connectivity.
A tree that must be sure to span
so packet can reach every LAN.
First, the root must be selected.
By ID, it is elected.
Least-cost paths from root are traced.
In the tree, these paths are placed.
A mesh is made by folks like me,
then bridges find a spanning tree.

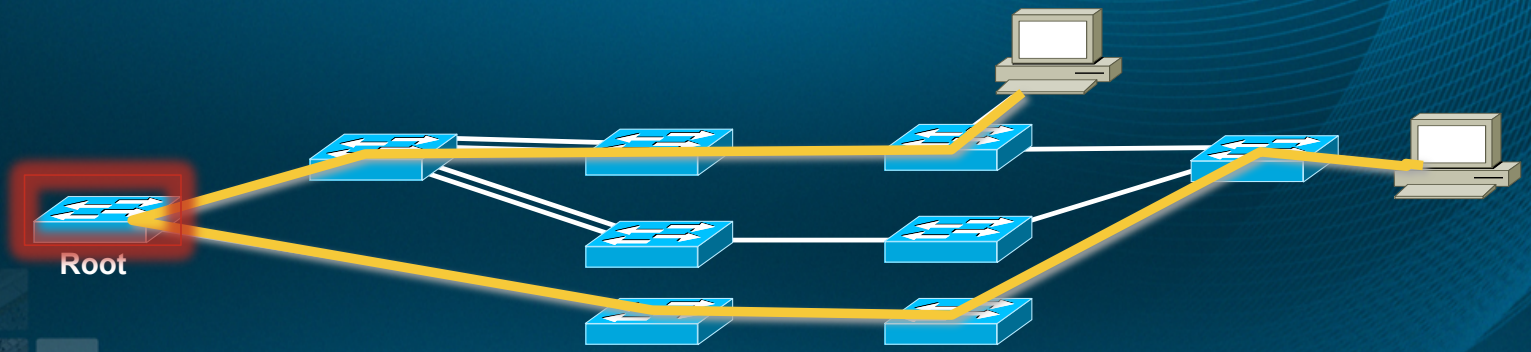
Radia Perlman



cisco

Spanning Tree (STP) –
Good at preventing loops, but no guarantees on optimal paths being used …

- Spanning Tree Protocol (STP) and its variants often have a bad reputation
  - Non-optimal forwarding
  - Parallel paths between two switches cannot be leveraged
  - Parallel paths in the network cannot be leveraged
- These problems can be solved at L3
  - But L3 cannot be deployed in many scenarios such as clusters, metro Ethernet, virtualized servers (VM's) or where physical flexibility is desired



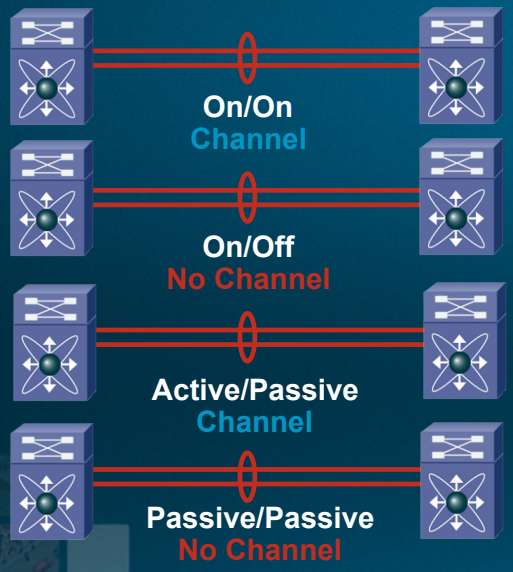
cisco

Link Aggregation

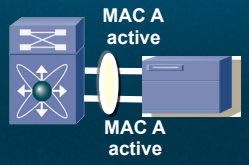
Increasing bandwidth point-to-point between two devices

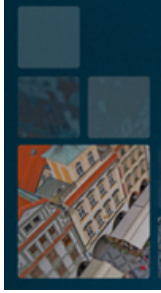
Port Channel aka EtherChannel

- **Standardized as IEEE 802.3ad link aggregation (LACP)**
- **Enables multiple physical parallel links between a pair of devices to be used as a single logical link for higher bisectional bandwidth.**
- **Can be used switch-to-switch, router-to-router, switch-to-host**



On: always be a channel/bundle member
Active: ask if the other side can/will
Passive: if the other side asks I will
Off: don't become a member of a channel/ bundle



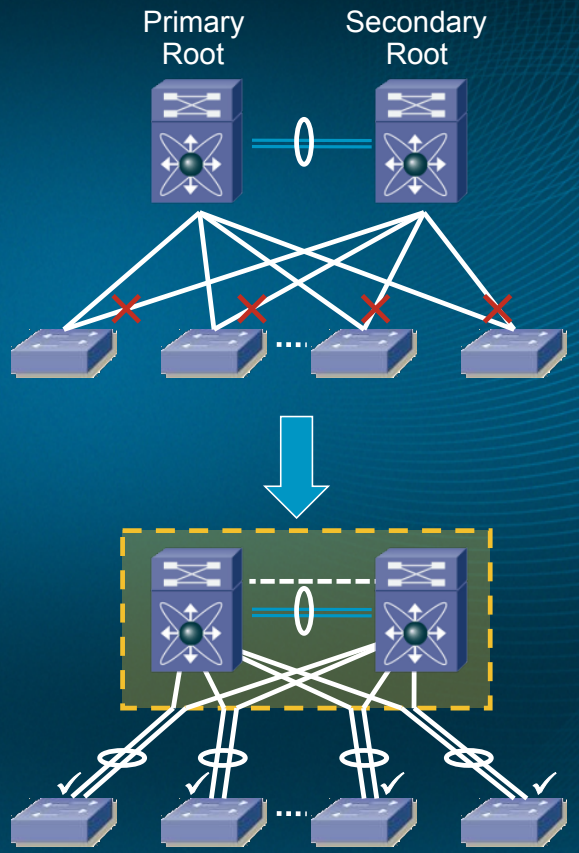


Virtual Port Chanel (vPC)

Turning Link Aggregation into point-to-multipoint

Evolutionary not Revolutionary

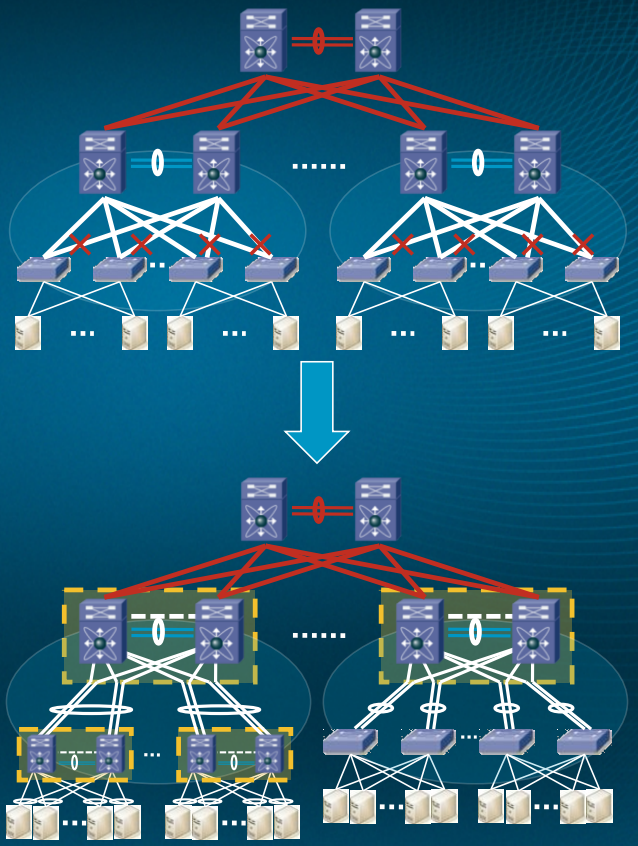
- Before vPC
  - STP blocks redundant uplinks
  - VLAN based load balancing
  - Re-convergence relies on STP
  - Protocol Failure → 💣
- With vPC
  - No blocked uplinks in STP
  - Lower oversubscription
  - EtherChannel load balancing (hash)
  - Convergence sub-second



cisco

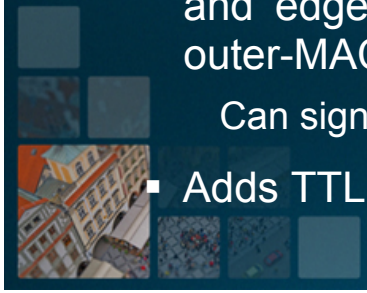
How does vPC help with STP?

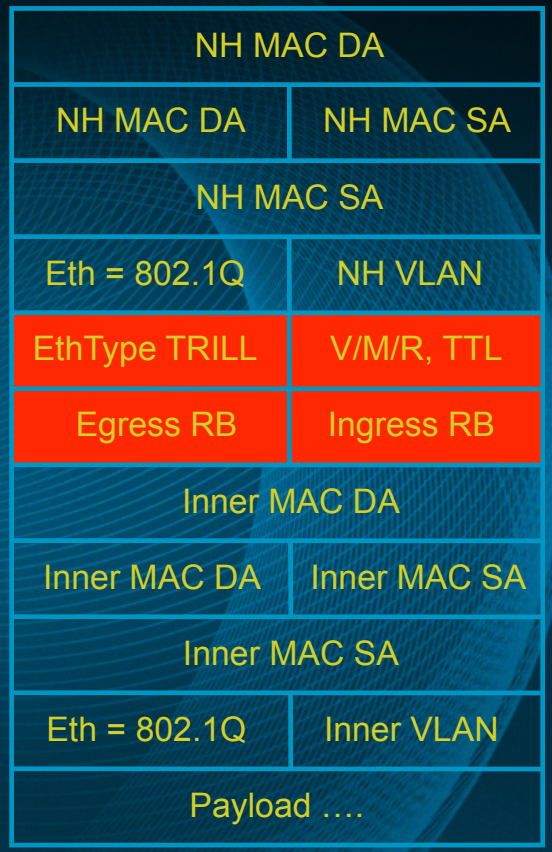
- Before vPC
  - STP blocks redundant uplinks
  - VLAN based load balancing
  - Re-convergence relies on STP
  - Protocol Failure → 💣
- With vPC
  - No blocked uplinks
  - Lower oversubscription
  - EtherChannel load balancing (hash)
  - Convergence sub-second



TRILL
Revolutionary, not Evolutionary

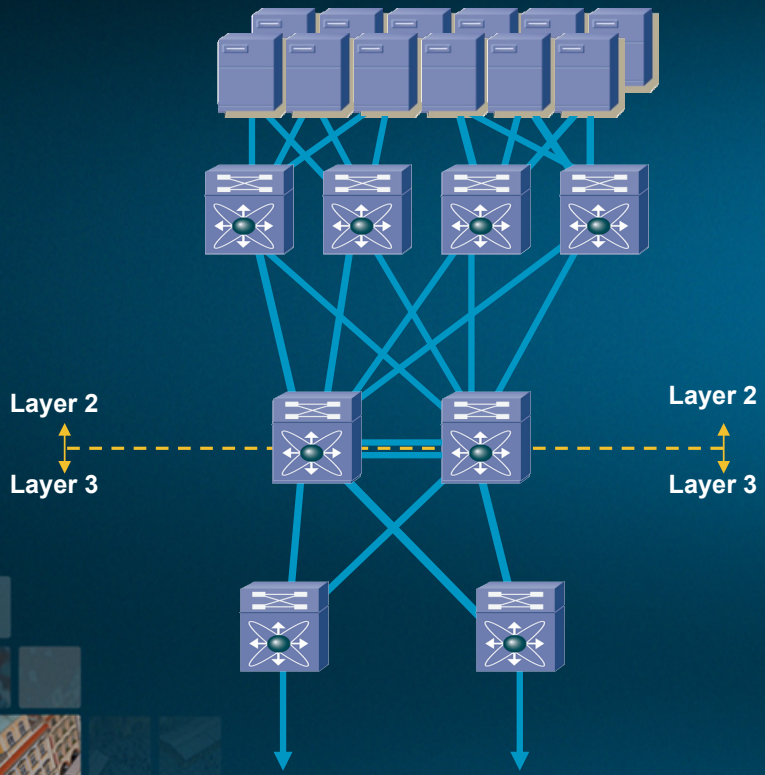
- TRILL is an IETF WG working on defining the behavior of a new kind of bridge: Rbridge (routing bridges)
- Transparent Interconnection of Lots of Links
  - Data format standard agreed upon already, final standard still a work in progress
- Use a routing protocol (IS-IS) to build a topology between switches
  - IS-IS is not an IP protocol
- Use all paths active end to end – no more blocked links
- Use hierarchical addressing (MAC-in-MAC) with concept of 'core' and 'edge' ports on L2 switches. Core ports forward based on outer-MAC, edge ports forward on inner-MAC
  - Can significantly shrink MAC table sizes on switches - $$$
- Adds TTL





cisco

What a network will look like in a IETF TRILL environment

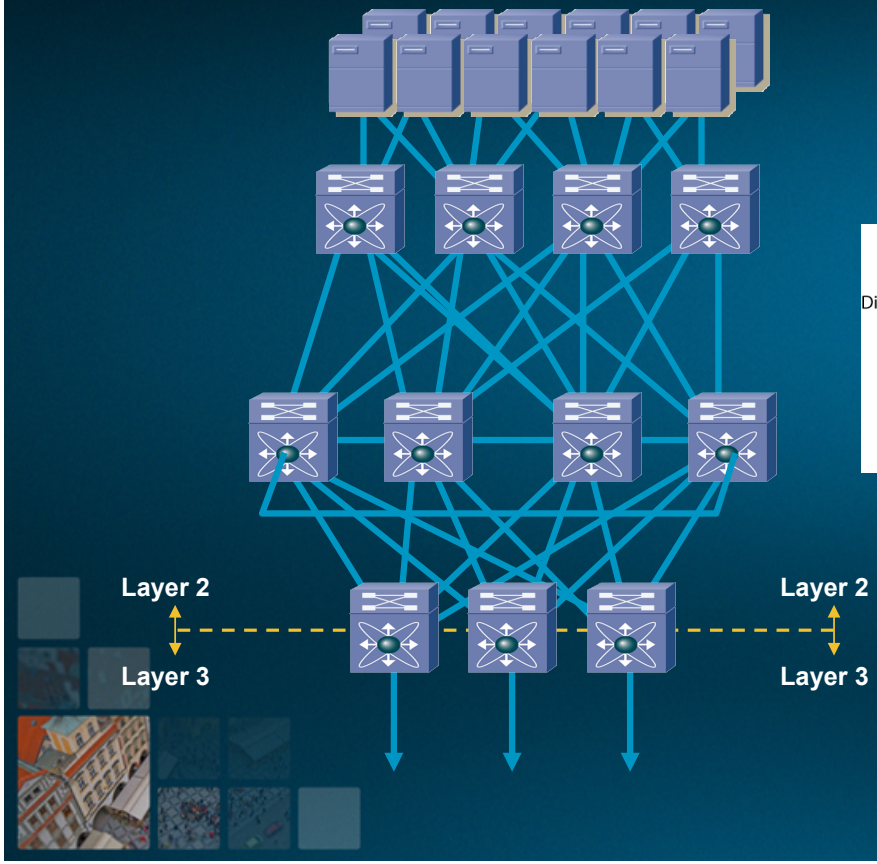


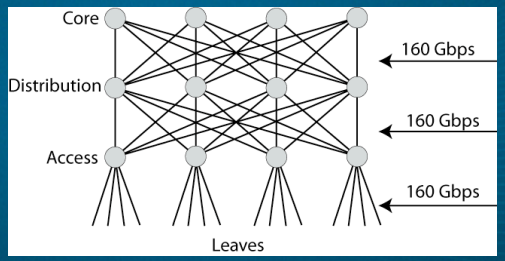
- Simplifies operations:
  - VLANs have edge significance only
  - Enables Arbitrary dual-homing of servers
  - Eliminates spanning tree
- Reduced state:
  - Access switch MAC learning limited to “interesting” flows only
  - Eliminates Logical Port limitations
- Enables graceful evolution

cisco

TRILL scale outs

- Flexible, arbitrary topologies that facilitate "any workload, any server"
- Scalable Bi-sectional bandwidth delivered using Fat Tree/CLOS network design





Algorhyme V2

I hope that we shall one day see
graph more lovely than a tree.

A graph to boost efficiency
While still configuration-free.

A network where RBridges can
Route packets to their target LAN.

The paths they find, to our elation,
Are least cost paths to destination!

With packet hop counts we now see,
The network need not be loop-free!

RBridges work transparently.
Without a common spanning tree.

Ray Perlner

CISCO

The Dilemma of Switch/Linecard Design

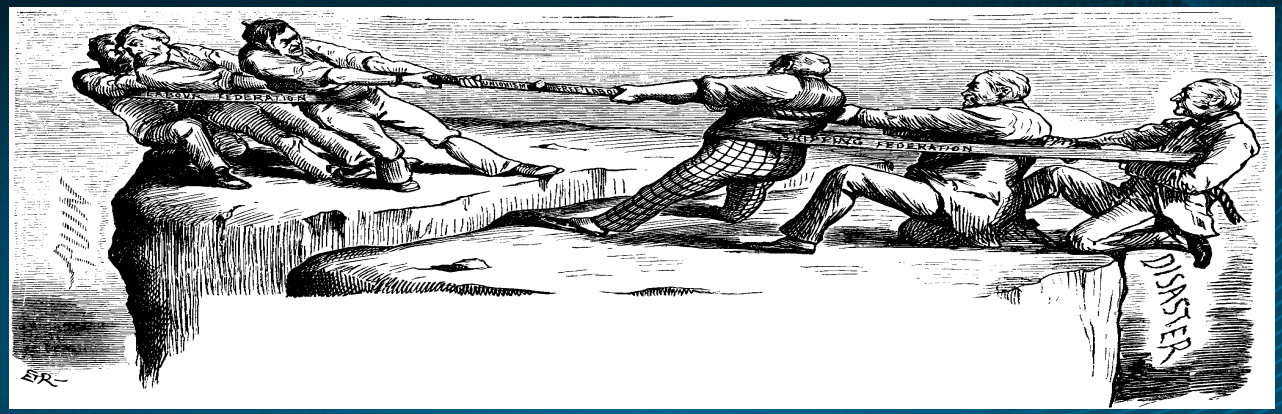
Optimize for functionality

Feature Rich

Large Buffers

Large Forwarding tables

Large CL-TCAM tables



Optimized for Cost

Integration improves Latency

High Performance and Density

Lower Latency
Higher Density
Lower Cost

High Degree → of integration → Required

Less Features

More Features
Bigger Buffers
Larger HW tables

→ Less Density →

Higher Cost per Port

cisco

Convergence of LAN and SAN

IEEE DCB (Data Centre Bridging)
INCITS T11 (ANSI) FC-BB-5 FCoE (Fibre Channel over Ethernet)
IEEE 802.1Qbb PFC (Priority Flow Control)
IEEE 802.1Qaz ETS (Enhanced Transmission Selection)
IEEE 802.1AB DCBXP (DCB Exchange Protocol)

FCoE is a standard

- From the FCIA announcement:

  "On June 3rd 2009, the FC-BB-5 working group of T11 completed its work and unanimously approved a final standard for FCoE.
  As a result, the plenary session of T11 approved forwarding the FC-BB-5 standard to INCITS for further processing as an ANSI standard.
  This is a major milestone in the final ratification of FCoE.
  For more details, you can click the link below for a copy of the standard.

  http://www.t11.org/ftp/t11/pub/fc/bb-5/09-056v5.pdf"

FCoE 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 Initialization Protocol)

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

The two protocols have:

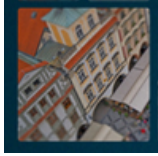
- Two different Ethertypes
- Two different frame formats

cisco

IEEE DCB (Data Center Bridging)

Feature / Standard	Standards Status
Priority Flow Control IEEE 802.1Qbb (PFC)	PAR approved, Editor Claudio DeSanti (Cisco), draft 1.0 published, expected WG ballot in 11/09
Bandwidth Management IEEE 802.1Qaz (ETS)	PAR approved, Editor Craig Carlson (Qlogic), draft 0.2 published, expected WG ballot in 11/09
Data Center Bridging Exchange Protocol (DCBX)	This is part of: Bandwidth Management IEEE 802.1Qaz

CEE (Converged Enhanced Ethernet) = IEEE DCB
Cisco DCE = IEEE DCB



cisco

Fibre Channel – Never Drop (but Block!)

Fibre Channel

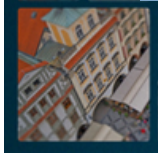
Blocks rather than drops

- ✓ Simplifies client (host) logic for high-speed send & receive
  - since network guarantees no dropped frames, no need for complex windowing protocols & retransmission mechanisms
- ✗ .. but at the cost that a mismatch in speeds between senders & receivers or a slow device can cause widespread blocking!
- ✗ limits the size of a network that can be built

Ethernet

Drops rather than blocks

- ✗ requires an upper-level protocol (TCP) to provide a reliable in-order guaranteed delivery
  - .. requires lots of buffering (RAM) & CPU cycles to provide high-throughput transport
- ✓ **Ubiquitous** – handles speed mismatches, self-paces to the rate required
- ✓ Ethernet + IP + TCP: **Proven to scale** (Internet)



cisco

802.1Qbb
Priority-based Flow Control

- 802.1Qbb PAR now approved

  http://www.ieee802.org/1/pages/802.1bb.html

  Cisco editor: Claudio DeSanti

  "… protocols, procedures and managed objects that enable flow control per traffic class on IEEE 802 full-duplex links… Priority-based Flow Control (PFC) is intended to eliminate frame loss due to congestion. This is achieved by a mechanism similar to the IEEE 802.3x PAUSE, but operating on individual priorities. … enables support for higher layer protocols that are highly loss sensitive while not affecting the operation of traditional LAN protocols utilizing other priorities… "

- Scope

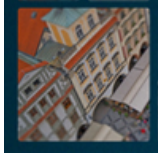
  Address no packet drop behavior

  Per-priority Pause to extend 802.3x PAUSE mechanism to accommodate different priority classes

  Selective Pausing

  Priority-based flow control to storage protocols over TCP/IP

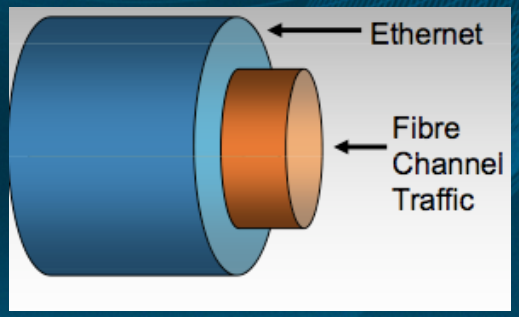
- Current Name: Priority-based Flow Control
- Status: Draft 1.0 released→CY2010/11

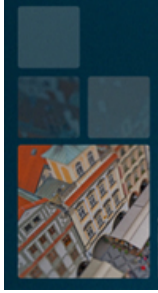


cisco

802.1Qaz – *Enhanced Transmission Selection*

- 802.1Qaz is now an amendment of 802.1Q
  - "Enhanced transmission Selection for bandwidth sharing between traffic classes"
  - Allows time sensitive flows (AVB) + PFC based priorities + 'normal' traffic to coexist on the wire
- DCBX (Data Center Discovery and Capability Exchange Protocol)
  - DCBX uses LLDP to exchange parameter between two link peers
- Scope
  - When the offered load in a traffic class doesn't use its allocated bandwidth, enhanced transmission selection will allow other traffic classes to use the available bandwidth. The bandwidth allocation priorities will coexist with strict priorities.
- Sounds a lot like the stuff we already have on routers (CBQ+LLQ) ☺
- Status: Draft 0.4 released→ Expected Approval 2010
- http://www.ieee802.org/1/pages/802.1az.html





cisco

802.1Qau – *Congestion Notification*

- Specifies protocols, procedures and managed objects that support congestion management of long-lived data flows within network domains of limited bandwidth delay product
- Bridges signal congestion information to end stations capable of rate limiting to avoid frame loss
- Latest Name: Quantized Congestion Notification (QCN)
- Technology applicable to multiple environments but particularly interesting in data center environment:
  - Server-to-server communication
  - Influences switch architecture: fabric and I/O modules
- Status: Expected Approval 2010
  - Frame format has not been defined
  - Editor Norman Finn
  - http://www.ieee802.org/1/pages/802.1au.html

Data Center Bridging – "Converged Ethernet"
Architectural Collection of Ethernet Extensions

Feature and Standard	Benefit
Priority-based Flow Control (PFC) IEEE 802.1Qbb	Enables multiple traffic types to share a common Ethernet link without interfering with each other Ensures ability to support FC traffic over Ethernet
Enhanced Transmission Selection (ETS) IEEE 802.1Qaz	Grouping classes of traffic into "Service Lanes" enables consistent management of QoS at the network level through consistent scheduling
Congestion Notification (QCN) IEEE 802.1Qau	Provides end to end management of sustained congestion for L2 networks
Data Center Bridging Exchange Protocol (DCBX) IEEE 802.1AB	Management protocol for auto-negotiation of DCB Ethernet capabilities over LLDP (Switch to Switch and Switch to NIC)
L2 Multi-Pathing IETF TRILL	Up to 16 way ECMP for full utilization of bi-sectional bandwidth. Eliminate Spanning Tree Protocol by using L2 IS-IS for topology convergence
Lossless Service	Allows the creation of a guaranteed delivery service within a switch for applications that require it.

Server Virtualization / Virtual Machine awareness

1G to 10G and Physical Cabling trends

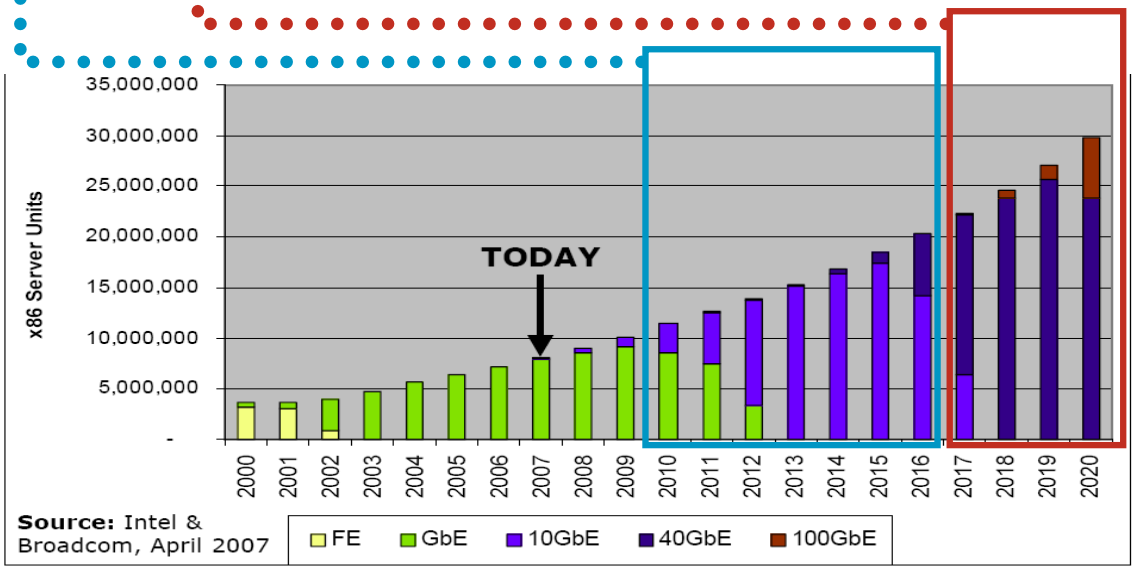
Impact on network elements and implementing per-VM policies

Demands on the network for DR across data centers

cisco

Server Ethernet Connection Evolution – Estimates from NIC vendors

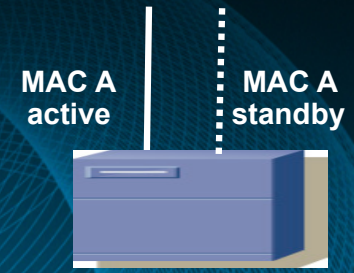
- 10GE NIC and LOM → 40G and 100G aggregation
- 40GE server connectivity → 100G uplinks



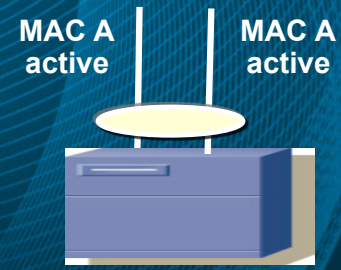
Base Server Connection: 10 year transition GbE | 5 years 10GbE | 5 years 40GbE

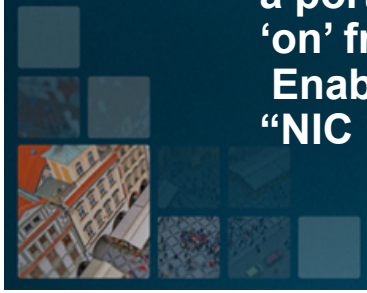
Active/Active Host Connectivity

- **Active/Standby Host Connection**
  - A host with more than 1 physical network connection sharing a common MAC address, but with only one link active at a time.
  - If the active link fails, the host will use the same MAC address on a standby link.
  - Enabled through “NIC Teaming” or “NIC Bonding” on the host.



- **Active/Active Host Port Channel**
  - A host with more than 1 physical network connection sharing a common MAC address, with more than one link active.
  - From the perspective of the switch, this is configured as a PortChannel, with the host either actively participating in a port channel protocol (e.g. LACP), or manually configured ‘on’ from the switch side.
  - Enabled through ‘active/active’ “NIC Teaming” or “NIC Bonding” on the host.





cisco

Physical Cabling Trends and Options – 10GbE

Typically <100m

In-Rack Cabling

- 10GBase-CX1 (aka Twinax)
- Up to 5m passvve
- Up to 25m active**

In-Row and X-Row Cabling

- 10GBase-USR**
  (30m using OM3 fiber)
- 10GBase-SR
  (300m using OM3 fiber)
- 10GBase-LR
  (10km using SMF)

cisco

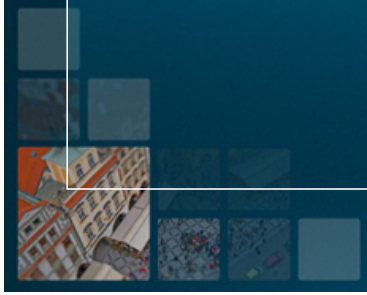
Data Center Access switch architecture choices

Top of Rack (ToR) Architecture benefits

- Flexible and scalable POD design
- Ease in replication of racks
- Shorter server-to-access switch cabling
- Fewer across-rack cables
- Lower cabling costs

End of Row (EoR) Architecture benefits

- Fewer configuration and management points in the network
- Fewer devices; require less power
- Lower CapEx and OpEx
- Ease in rolling out services and software upgrades
- Allows high-density server aggregation at access layer



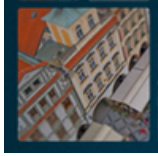
cisco

Hybrid ToR and EoR approaches

Combines benefits of Top of Rack (ToR) and End of Row (EoR) network architectures

- Physically resides on the top of each server rack
- Logically acts like an end of access row device

- **Reduces cable runs**
  Majority of physical cabling is within the rack, ≤5 foot cable
- **Reduce management points in the network**
  e.g.
  In a 4,000 port network design with traditional 48-port ToR access switches there would be 84 management points
  With a hybrid ToR/EoR approach, this could be reduced to a single management point without compromising any redundancy or resiliency
- **Ensures feature consistency across hundreds or thousands of server ports**



cisco

Server to Access Cabling reduction through convergence

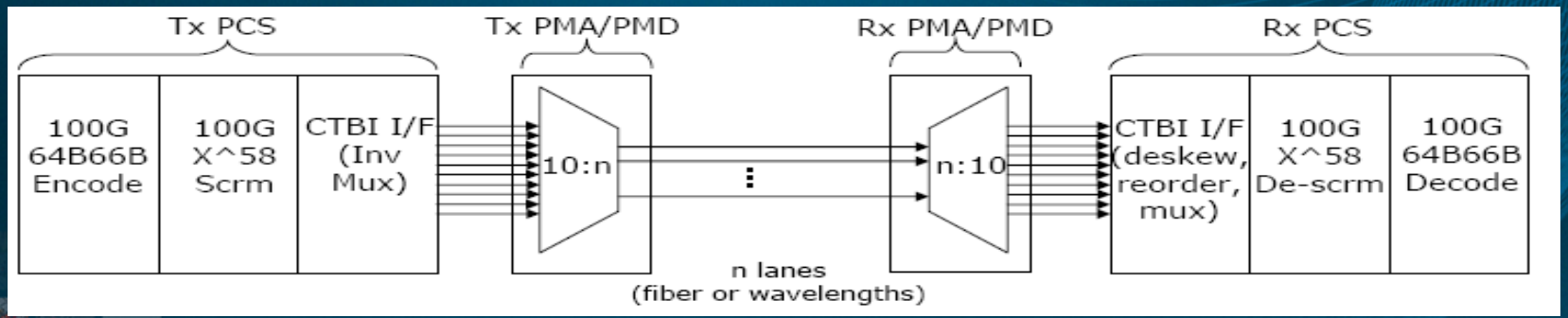
'Unified' Compute blades

86% cable reduction

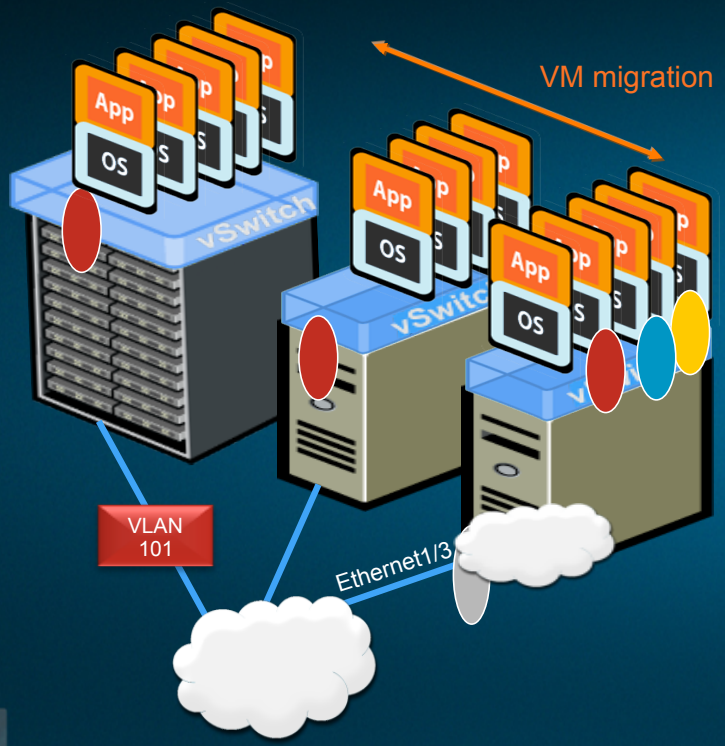
Non Unified Compute Infrastructure

100GbE and 40GbE: complete 2010

	40GbE	100GbE
At least 1m backplane	✓	
At least 10m cu cable	✓	✓
At least 100m OM3 MMF	✓	✓
At least 10km SMF		✓
At least 40km SMF		✓



Virtual Machine Visibility



Challenges:

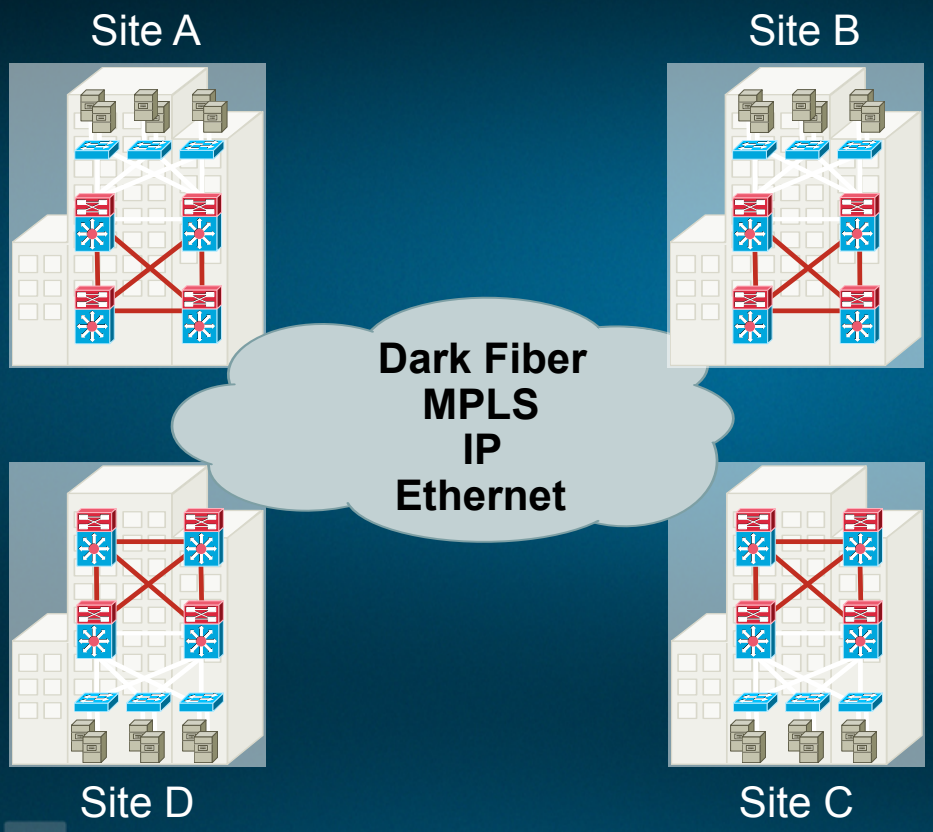
- VM migration (e.g. vMotion) may move VMs across physical ports — any ‘policy’ (ACLs, Policing, Accounting etc.) must migrate with the VM
- Not possible to view or apply policy to locally switched traffic – particularly problematic for compromised hosts
- Cannot correlate traffic on physical links — from multiple VMs

Future Trends:

- Extends network to the VM
- Consistent services
- Coordinated, coherent management

cisco

Extending L2 outside the data centre



Prediction

Choices of transport technology in the Data Center over the last 10 years?

Fibre Channel

iSCSI

Ethernet

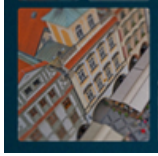
Infiniband

ATM

FDDI

Token Ring

What will be the technology choice for the Data Center in the next 5 years?



cisco

Q and A