



Data Centre Networking— Architecture and Design Guidelines

BRKDCT-2001



Ian Bond

**Cisco Networkers
2007**

HOUSEKEEPING

- We value your feedback, don't forget to complete your online session evaluations after each session and complete the Overall Conference Evaluation which will be available online from Friday.
- Visit the World of Solutions on Level -01!
- Please remember this is a 'No Smoking' venue!
- Please switch off your mobile phones!
- Please remember to wear your badge at all times including the Party!
- Do you have a question? Feel free to ask them during the Q&A section or write your question on the Question form given to you and hand it to the Room Monitor when you see them holding up the Q&A sign.

Before We Get Started:

- Q and A at end of session
- Intermediate level session focused on data centre front end architecture for the transactional model
- Other recommended sessions:
 - BRKDCT-2002 : Data Centre IP Front-End : Solution for Business Continuance
 - BRKDCT-2003 : High-Density Server Farms
 - BRKDCT-2004 : Data Centre Fibre Channel Back End Infrastructure: Solutions for Disaster Recovery
 - BRKDCT-2005 : Design and Deployment of Layer 2 Clusters and Geoclusters
 - BRKDCT-2006 : Introduction to High-Performance Computing
 - BRKDCT-2007 : Data Centre Optical Infrastructure for the Enterprise
 - BRKDCT-3008 : Advanced SAN Fabrics and Storage Virtualisation
 - BRKDCT-3009 : Advanced Understanding of Infiniband Technology
 - TECDCT-1001 : Architecting Data Centre Infrastructure and Services

Agenda

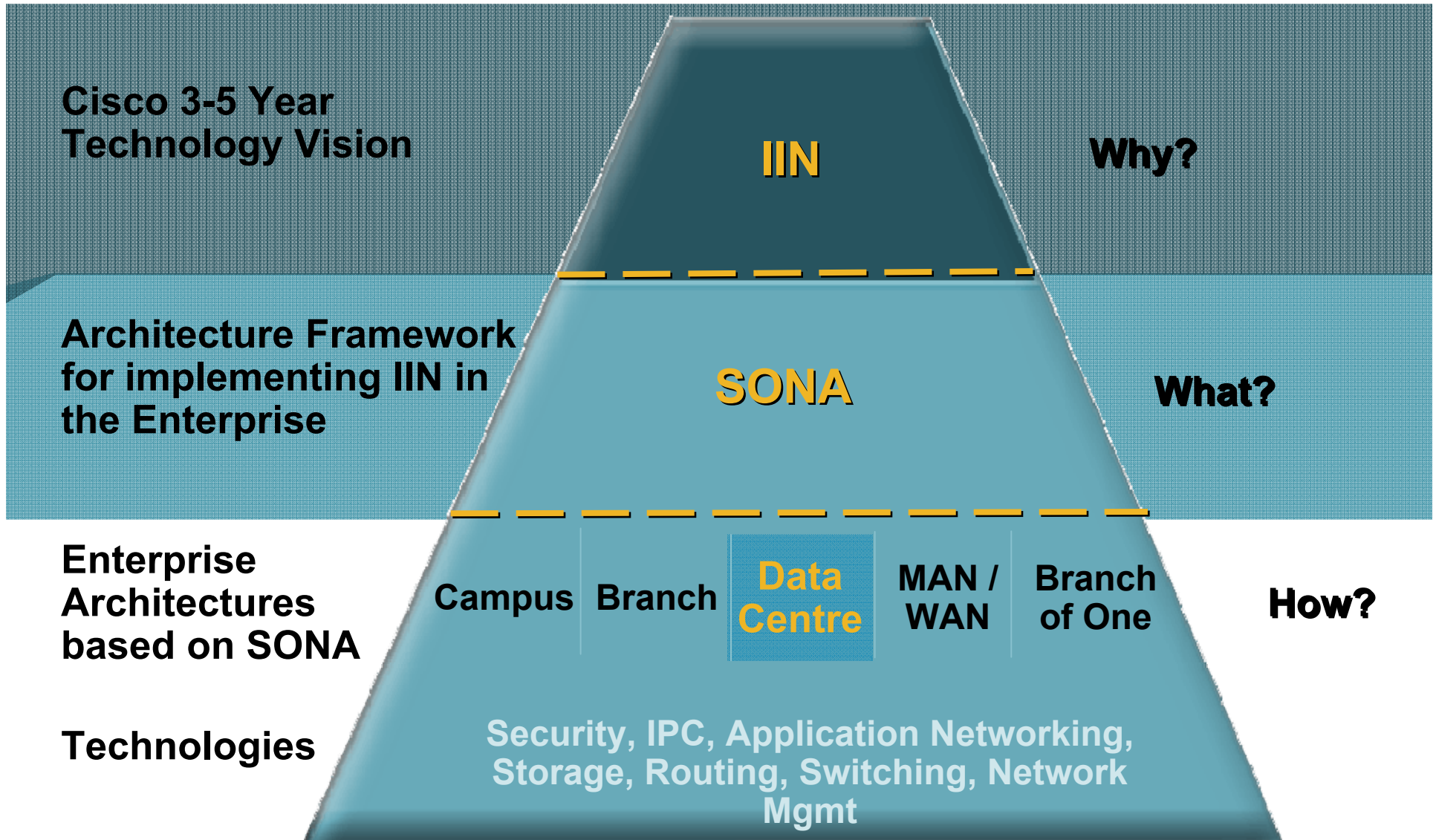
Data Centre Infrastructure

- Overview of Cisco Data Centre Network Architecture
- Core Layer Design
- Aggregation Layer Design
- Access Layer Design
- Density and Scalability Implications
- Scaling Bandwidth and Density
- Spanning Tree Design and Scalability
- Increasing HA in the DC
- A look at the Future...

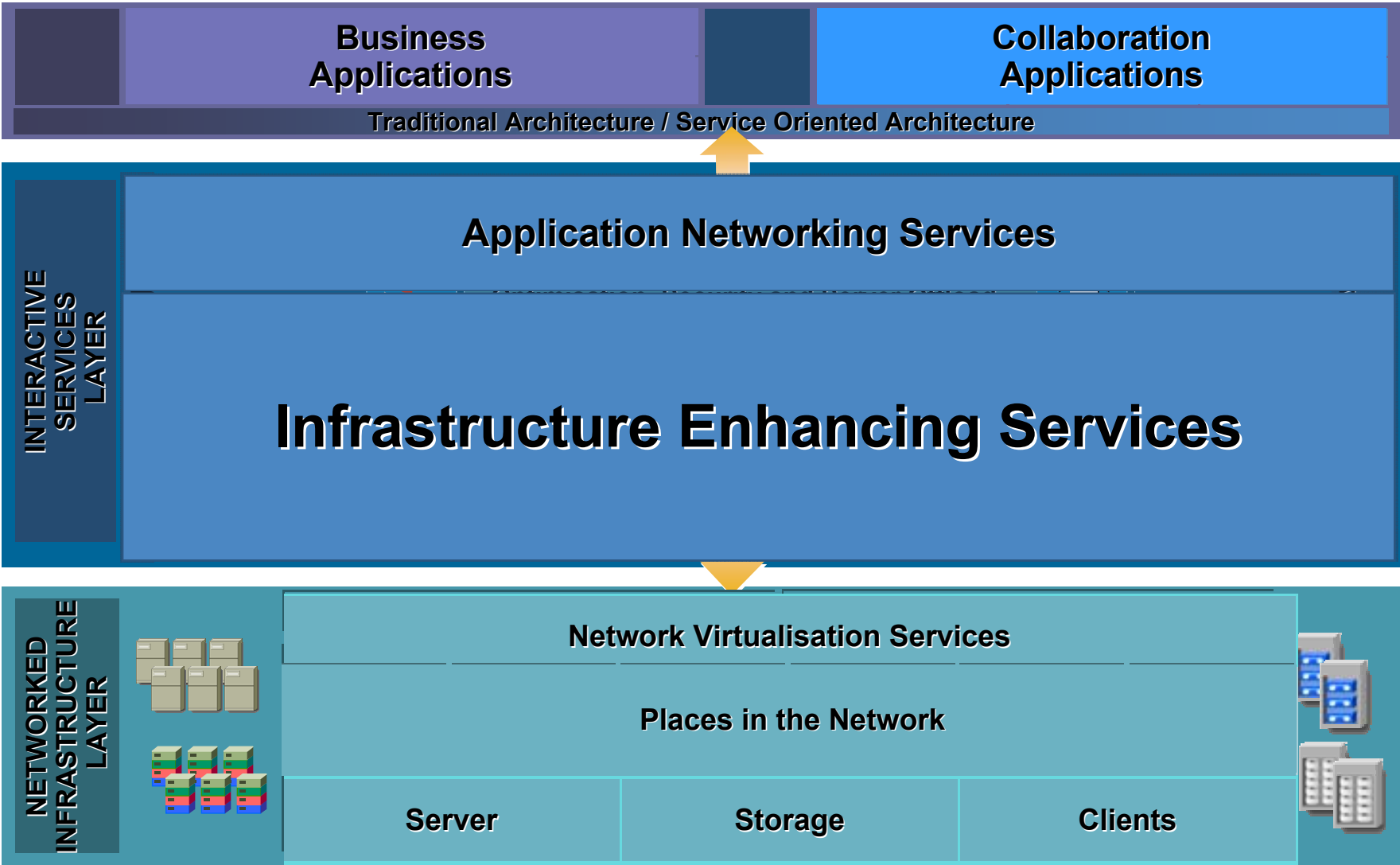
Cisco Data Centre Networking Architecture



Cisco Data Centre Network Architecture – Intelligent Information Network and Service Oriented Network Architecture



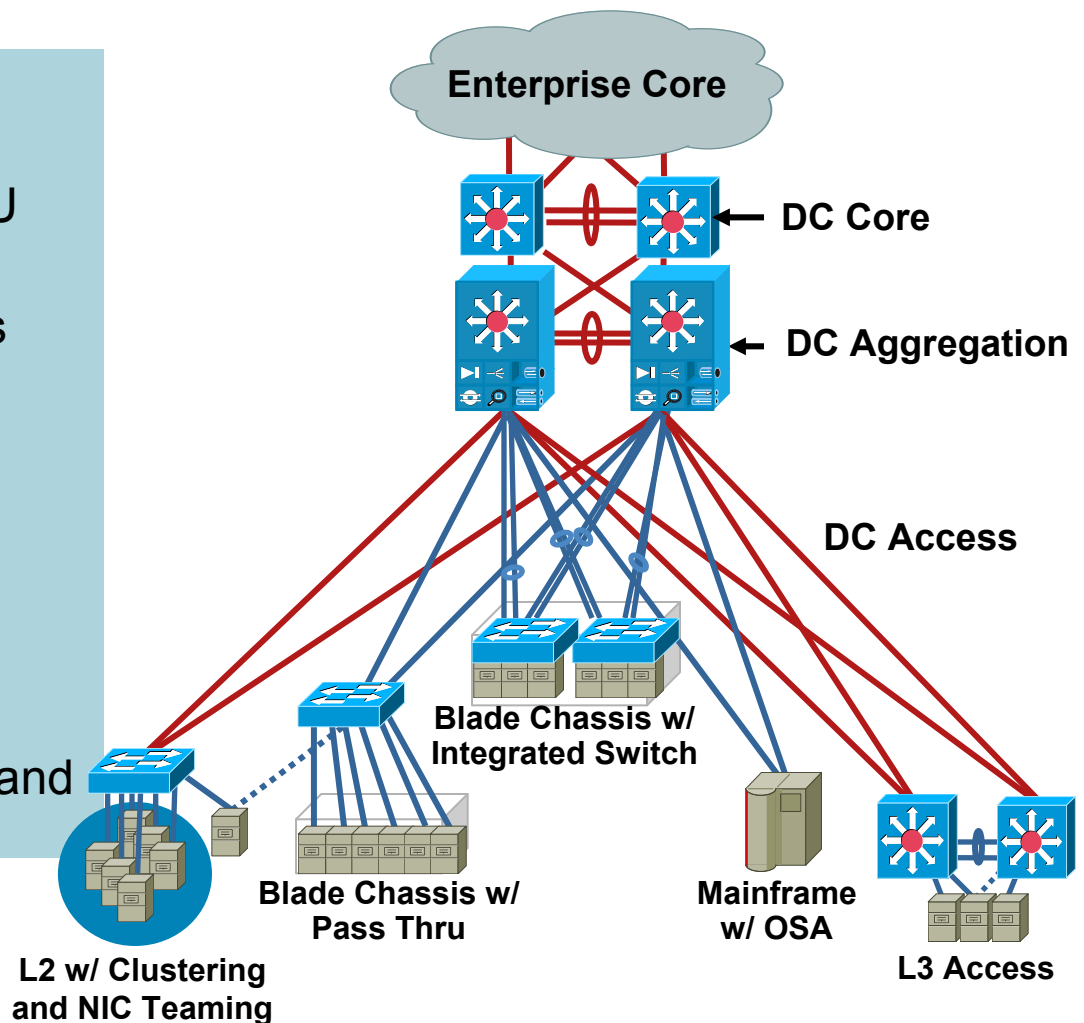
Cisco Data Centre Network Architecture



Data Centre Architecture Overview

Layers of the Enterprise Multi-Tier Model

- Layer 2 and layer 3 access topologies
- Dual and single attached 1RU and blade servers
- Multiple aggregation modules
- Web/app/database multi-tier environments
- L2 adjacency requirements
- Mix of over-subscription requirements
- Environmental implications
- Stateful services for security and load balancing

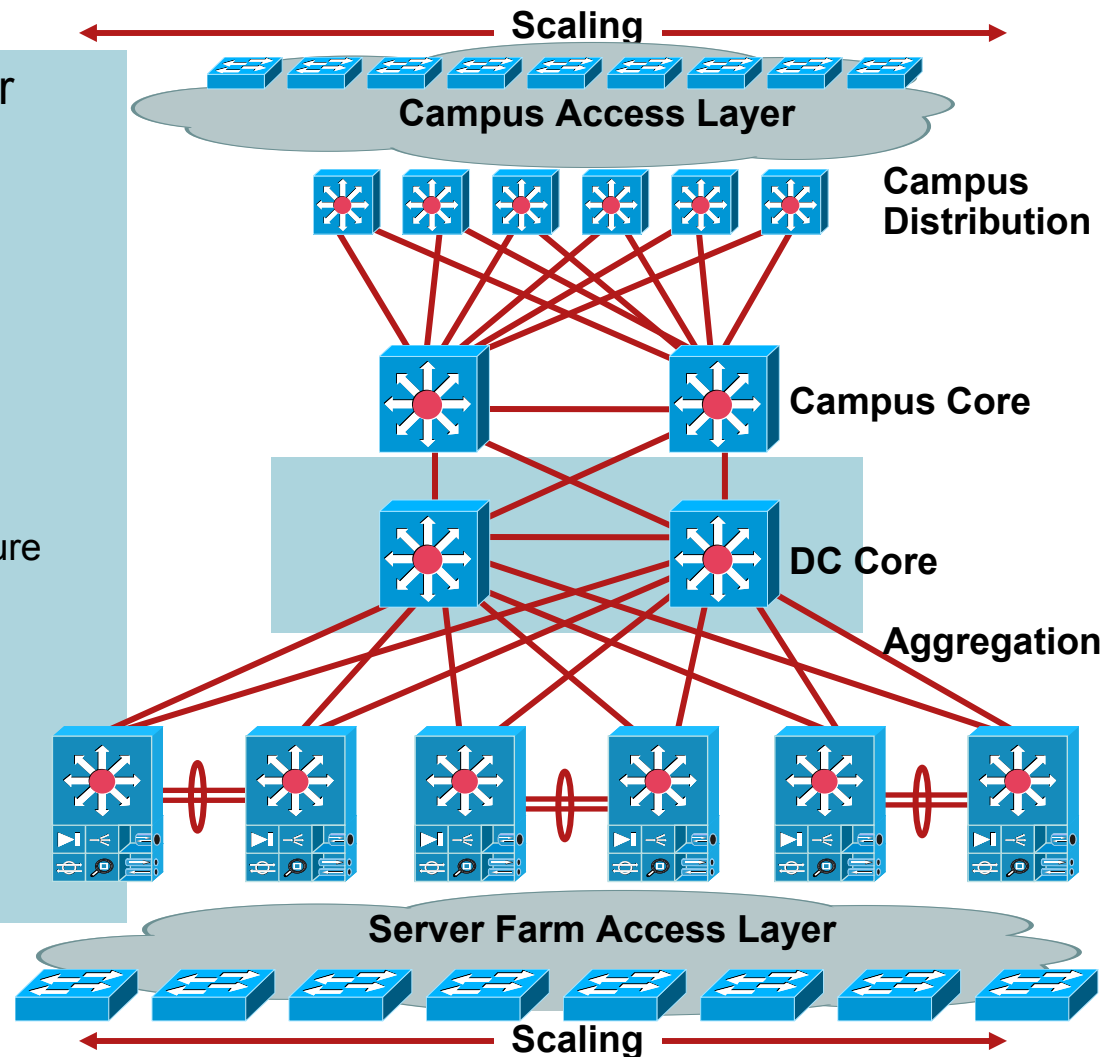


Core Layer Design



Core Layer Design Requirements

- Is a separate DC Core Layer required?
- Consider:
 - 10GigE port density
 - Administrative domains
 - Anticipate future requirements
- Key core characteristics
 - Distributed forwarding architecture
 - Low latency switching
 - 10GE scalability
 - Advanced ECMP
 - Advanced hashing algorithms
 - Scalable IP multicast support



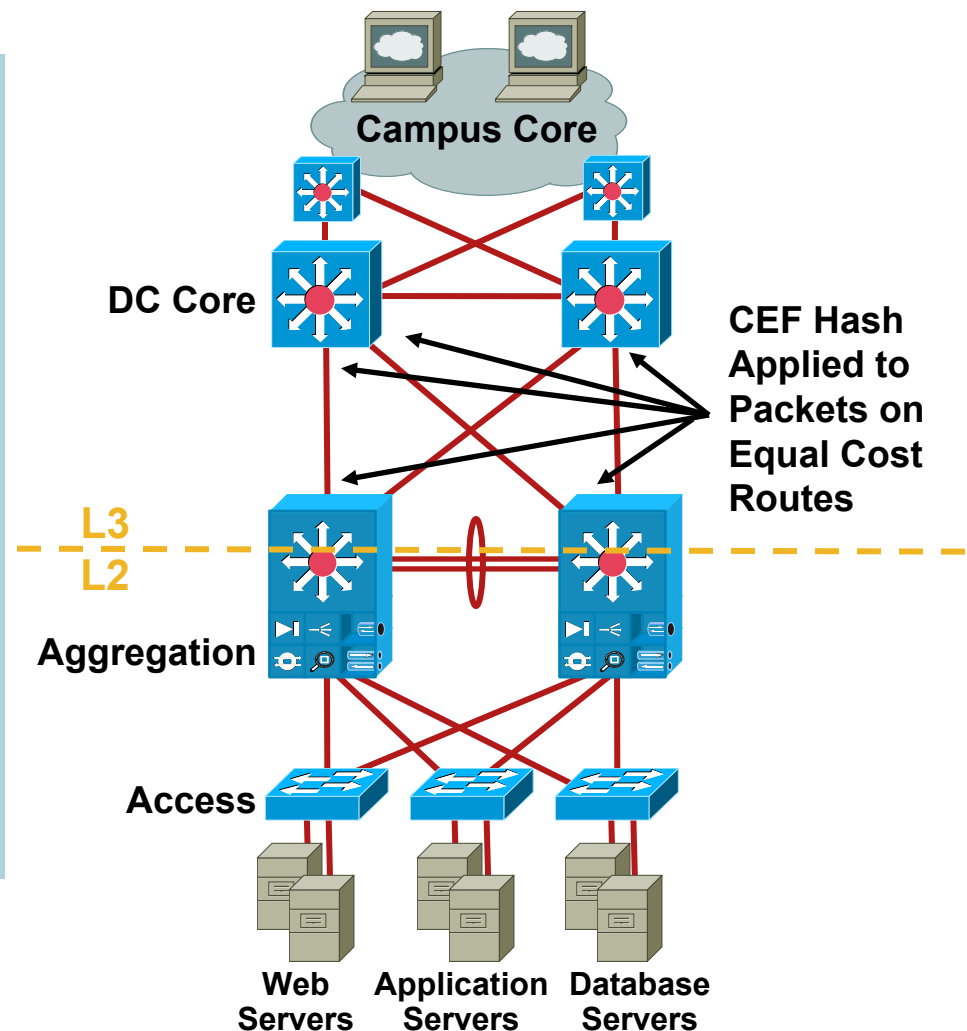
Core Layer Design

L2/L3 Characteristics

- Layer 2/3 boundaries:
 - All Layer 3 links at core, L2 contained at/below aggregation module
 - L2 extension through core is not recommended
- CEF hashing algorithm
 - Default hash is on L3 IP addresses only
 - L3 + L4 port is optional and may improve load distribution

CORE1(config)#mls ip cef load full

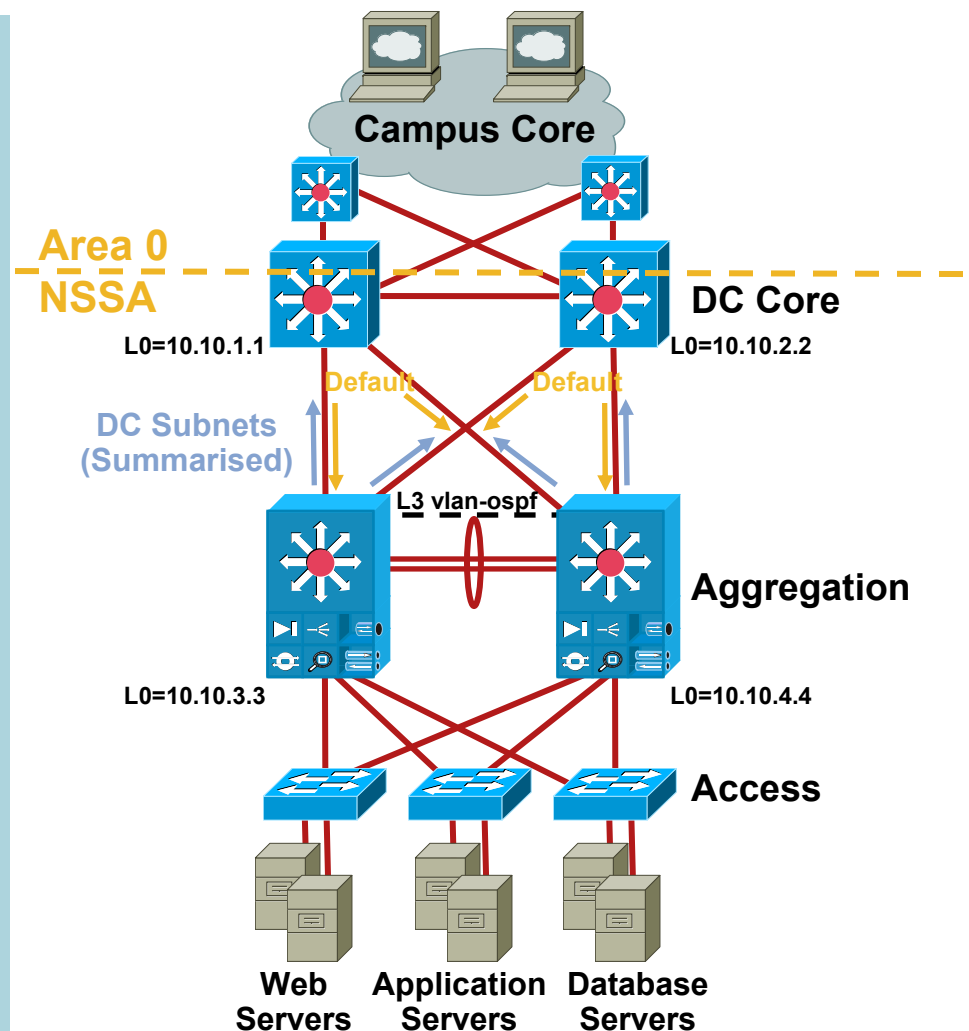
Leverages automatic source port randomisation in client TCP stack



Core Layer Design

Routing Protocol Design: OSPF

- NSSA helps to limit LSA propagation, but permits route redistribution (RHI)
- Advertise default into NSSA, summarise routes out
- OSPF default reference b/w is 100M, use “auto-cost reference-bandwidth” set to 10G value
- VLANs on 10GE trunks have OSPF cost = 1G (cost 1000), adjust bandwidth value to reflect 10GE for interswitch L3 vlan
- Loopback interfaces simplify troubleshooting (neighbor ID)
- Use passive-network default: open up only links to allow
- Use authentication: more secure and avoids undesired adjacencies
- Timers spf 1/1, interface hello-dead = 1/3



Core Layer Design

Routing Protocol Design: EIGRP

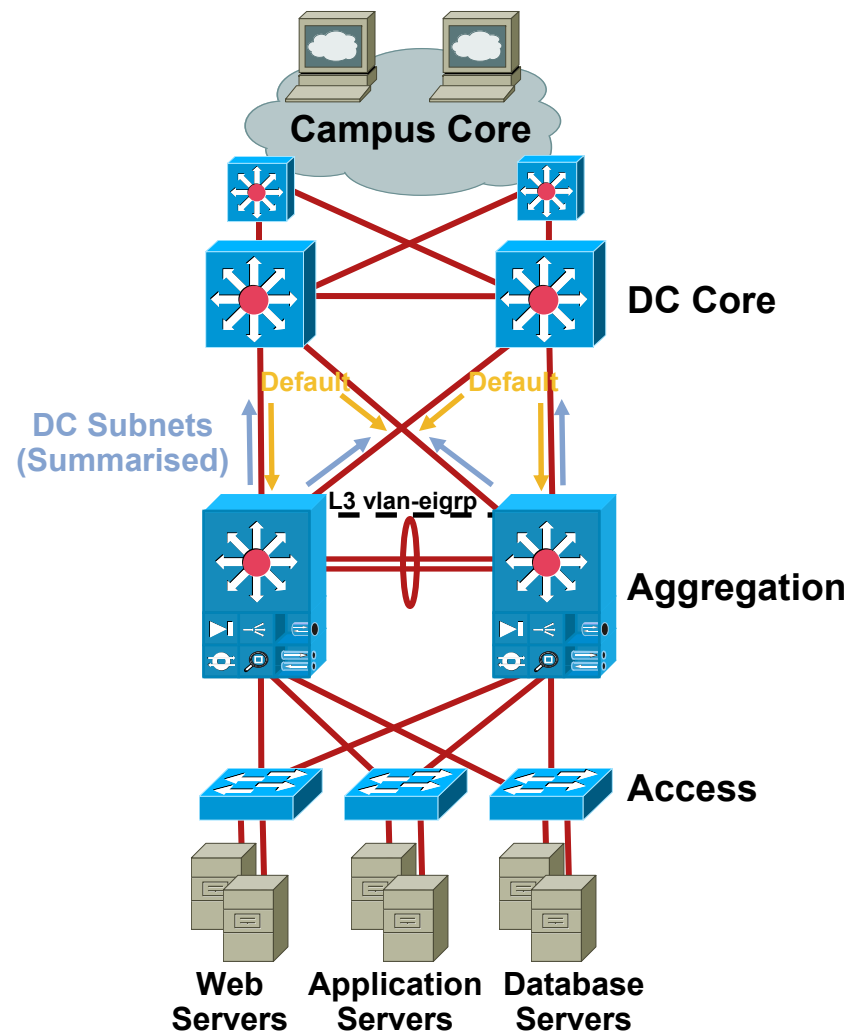
- Advertise default into DC with interface command on core:

```
ip summary-address eigrp 20 0.0.0.0 0.0.0.0 200
```

Cost of 200 required to be preferred route over the NULL0 route installed by EIGRP

- If other default routes exist (from internet edge for example), may need to use distribute lists to filter out
- Use passive-network default
- Summarise DC subnets to core with interface command on agg:

```
ip summary-address eigrp 20 10.20.0.0 255.255.0.0
```



Aggregation Layer Design



Aggregation Layer Design

Spanning Tree Design

- Rapid-PVST+ (802.1w) or MIST (802.1s),
- Choice of .1w/.1s based on scale of logical+virtual ports required
- R-PVST+ is recommended and best replacement for 802.1d

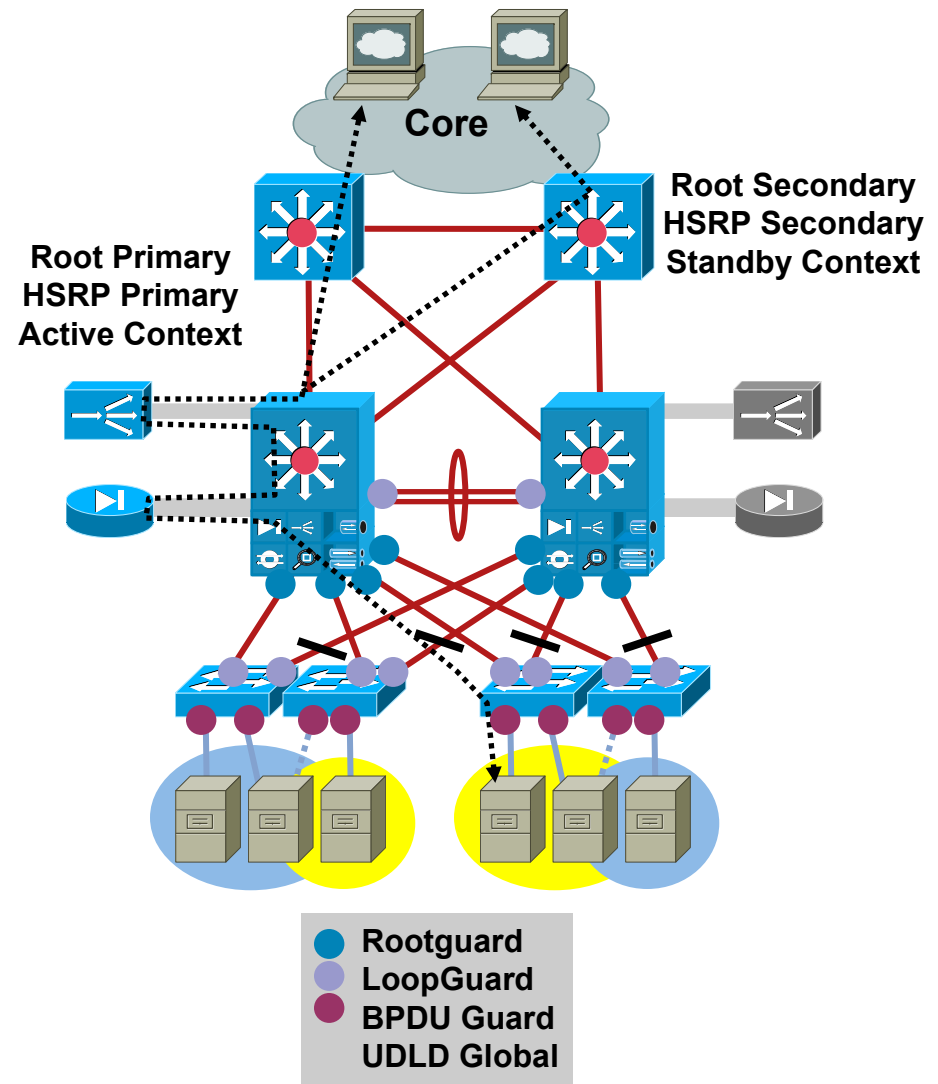
Fast converging: inherits proprietary enhancements (Uplink-fast, Backbone-fast)

Access layer uplink failures: ~300ms – 2sec

Most flexible design options

Combined with RootGuard, BPDUGuard, LoopGuard, and UDLD achieves most stable STP environment

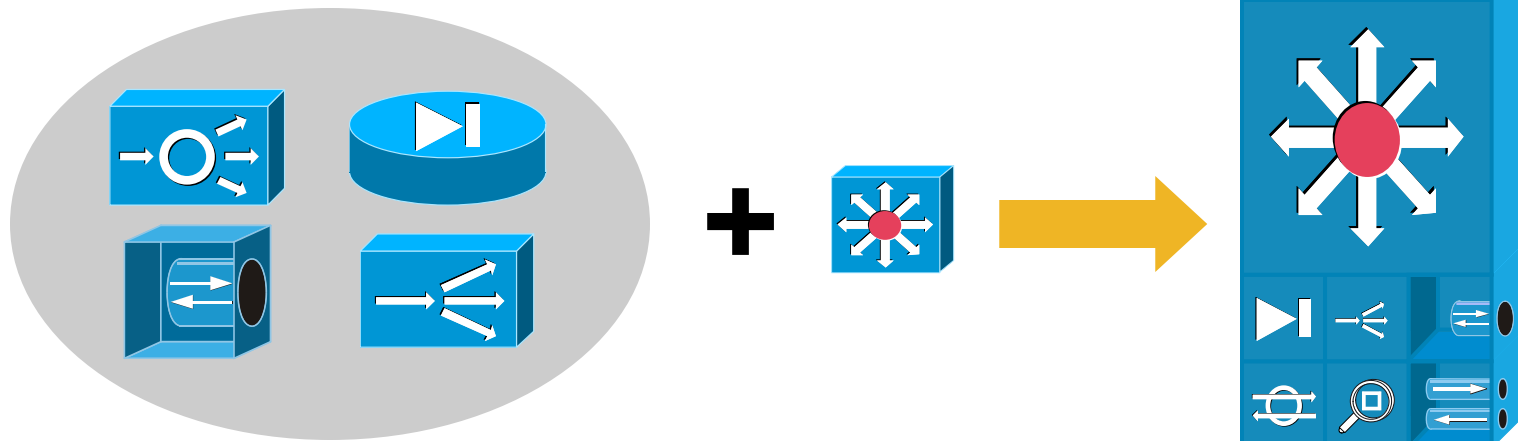
UDLD global only enables on Fiber ports, must enable manually on copper ports



Aggregation Layer Design

Integrated Services

Services: Firewall, Load Balancing, SSL Encryption/Decryption



- L4-L7 services integrated in Cisco Catalyst® 6500
- Server load balancing, firewall and SSL services may be deployed in:
 - Active-standby pairs (CSM, FWSM 2.X)
 - Active-active pairs (ACE, FWSM 3.1)
- Integrated blades optimise rack space, cabling, mgmt, providing flexibility and economies of scale
- Influences many aspects of overall design

Aggregation Layer Design

Service Module Placement Consideration

Cisco Catalyst 6500 Switch Fabric Channels

6-Slot	9-Slot	13-Slot
Dual	Dual	Single
Dual	Dual	Single
Dual	Dual	Single
Dual	Dual	Single
Dual	Dual	Single
* Dual	* Dual	Single
	Dual	Single
	Dual	* Single
	Dual	Dual
		Dual
		Dual
		Dual
		Dual
		Dual

Cisco Catalyst 6513

Single Channel Fabric Attached Modules
 Sup720, ACE, 6724, 6516
 FWSM, SSLSM, NAM-2, IDSM-2

Classic Bus Modules (No Channel)
 CSM, IDSM-1, NAM-1, 61xx-64xx series

(6704, 6708, 6748 Not Permitted in These Slots)

Dual Channel Fabric Attached Modules
 6748, 6704, 6708

(Supports All Single Channel and Classic Bus Modules Also)

The Choice Between 6509 and 6513 Usually Comes Down to:

6513: Supports Larger Number of Service Modules

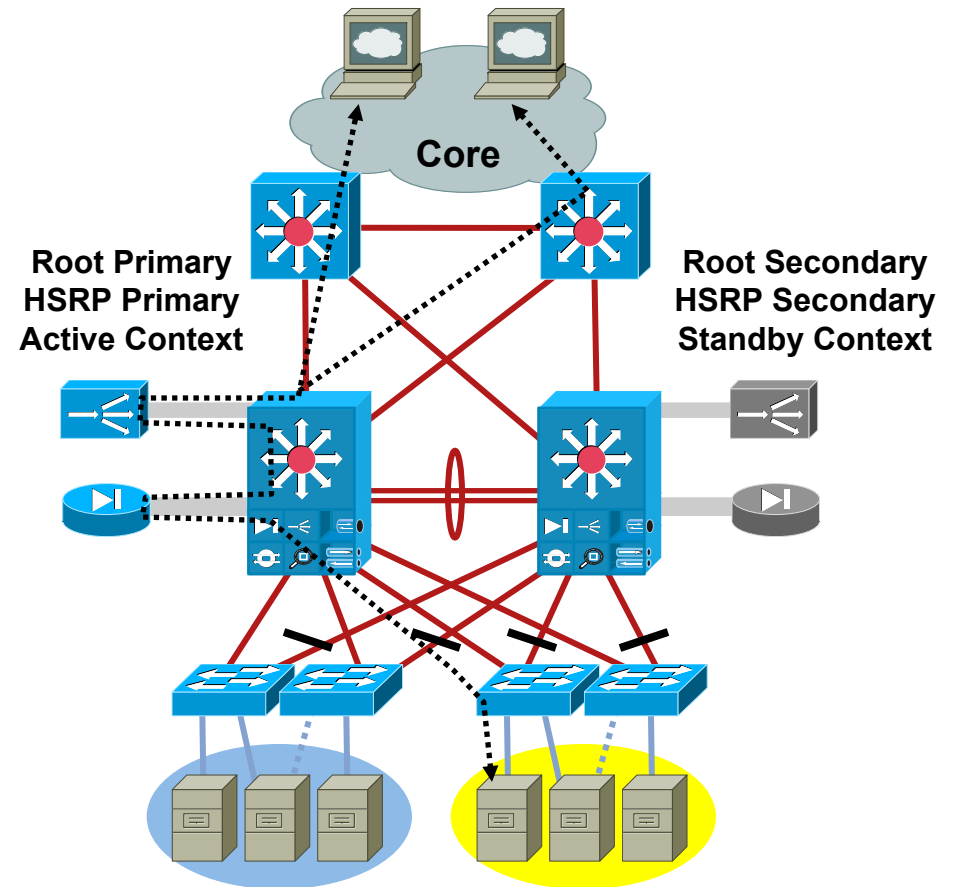
6509: Supports Larger Number of 10GE Ports

* Primary Sup720 Placement

Aggregation Layer Design

Active-Standby Service Design

- Active-standby services
 - Content Switching Module
 - Firewall Service Module (2.x)
 - SSL Module
- Under utilises access layer uplinks
- Under utilises service modules and switch fabrics
- Multiple service module pairs does permit active-active design but....



Aggregation Layer Design

Active-Active Service Design

- Active-Active Service Design

- Application Control Engine

- SLB+SSL+FW

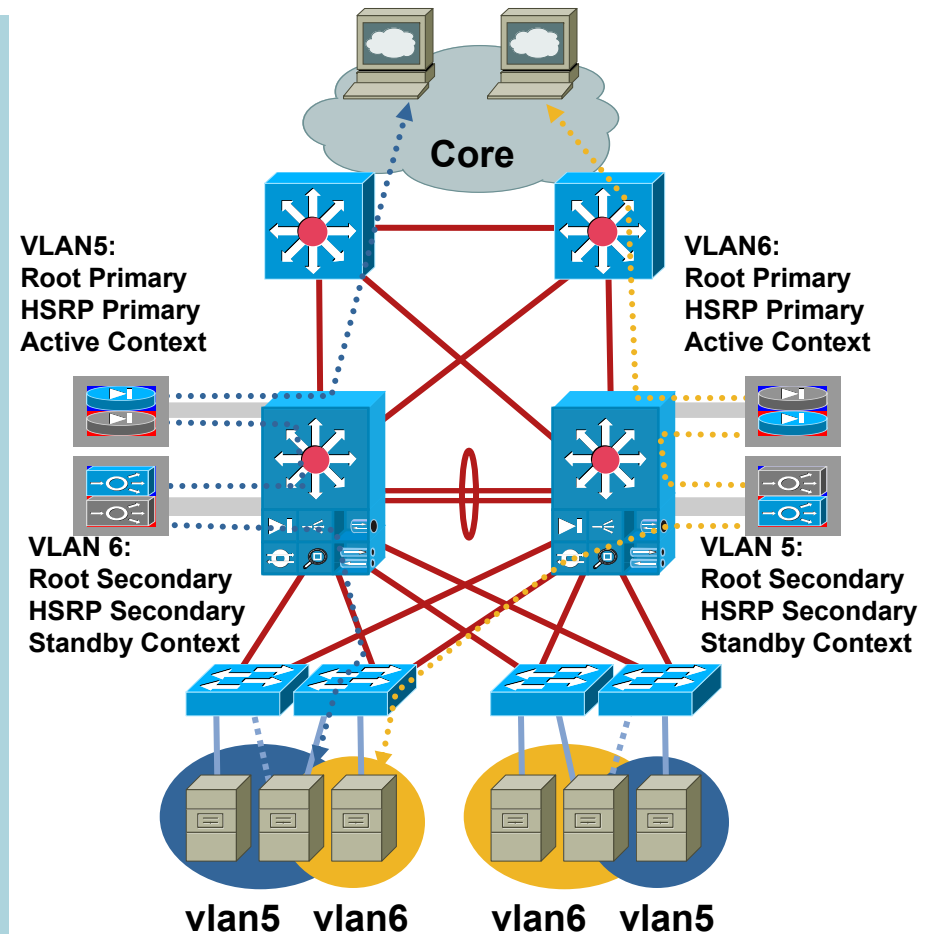
- 4G/8G/16G switch fabric options

- Active-standby distribution per context

- Firewall Service Module (3.x)

- Two active-standby groups permit distribution of contexts across two FWSM's (not per context)

- Permits uplink load balancing while having services applied
- Increases overall service performance

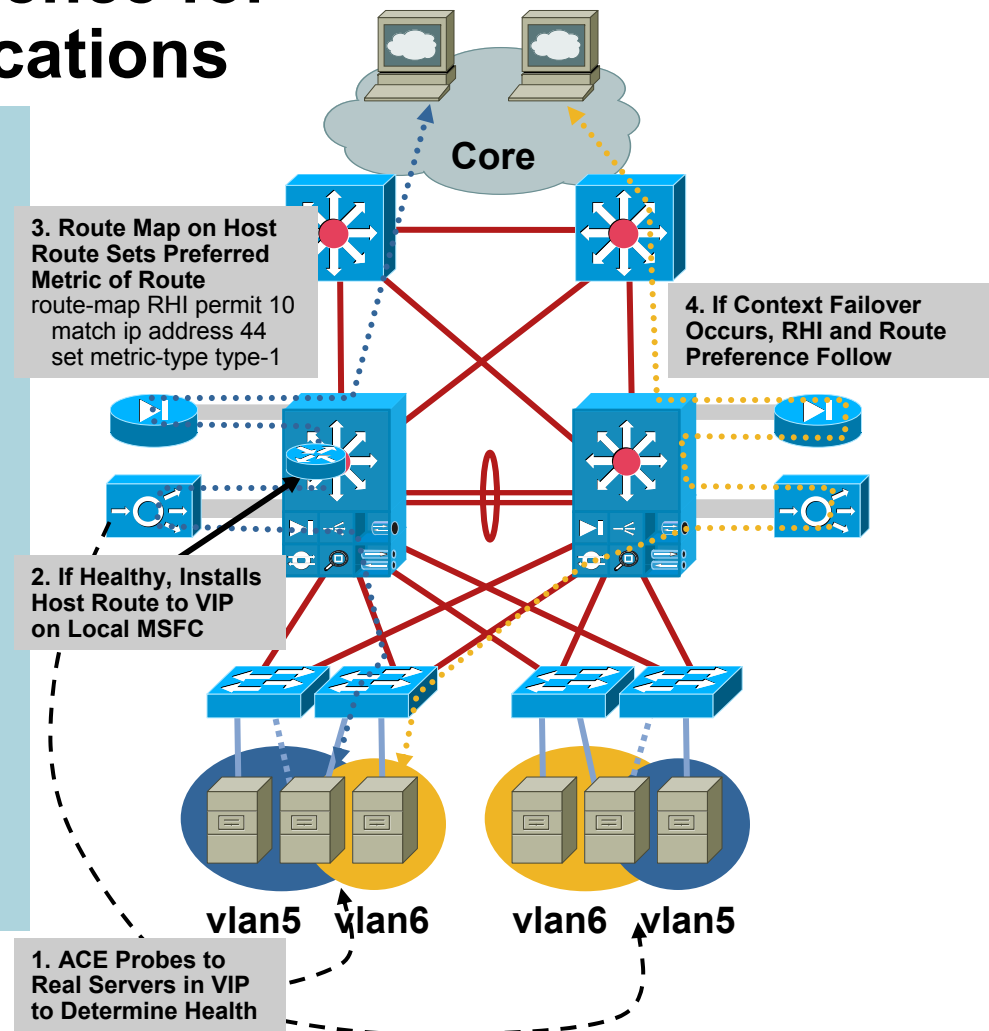


Aggregation Layer Design

Establishing Path Preference

Establish Route Preference for Service Enabled Applications

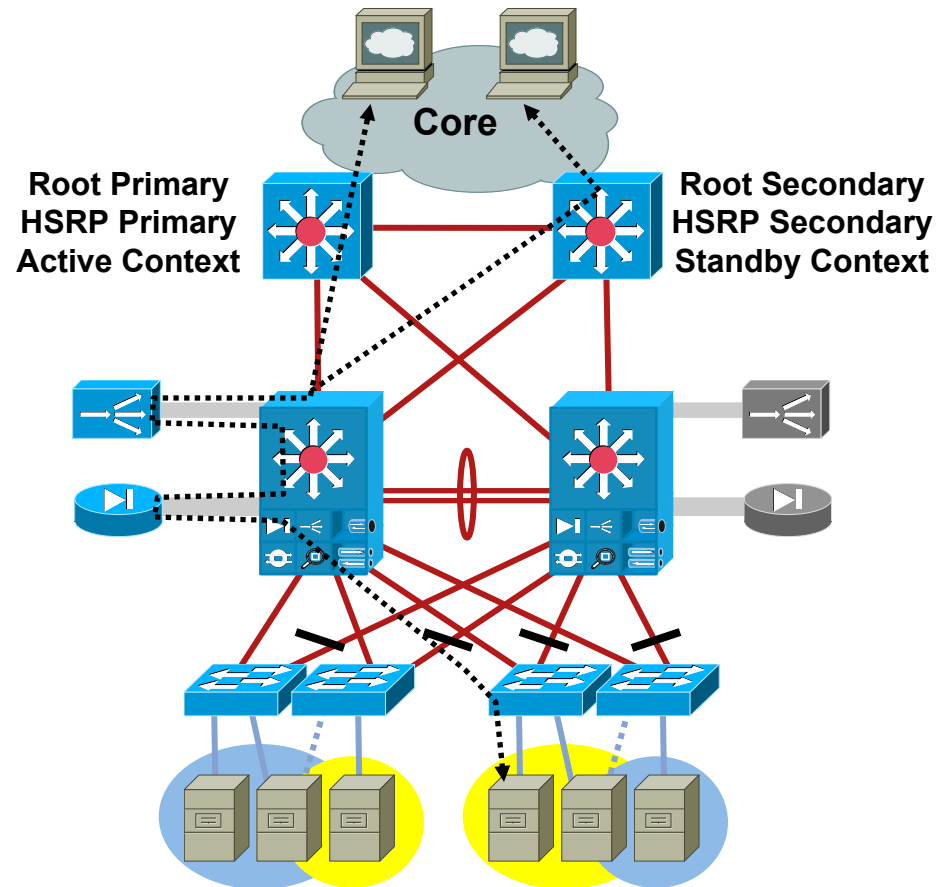
- Use Route Health Injection feature of ACE and CSM
- Introduce route-map to RHI injected route to set desired metric
- Aligns advertised route of VIP with active context on ACE, CSM, FWASM and SSL service modules
- Avoids unnecessary use of inter-switch link and asymmetrical flows



Core and Aggregation Layer Design

STP, HSRP and Service Context Alignment

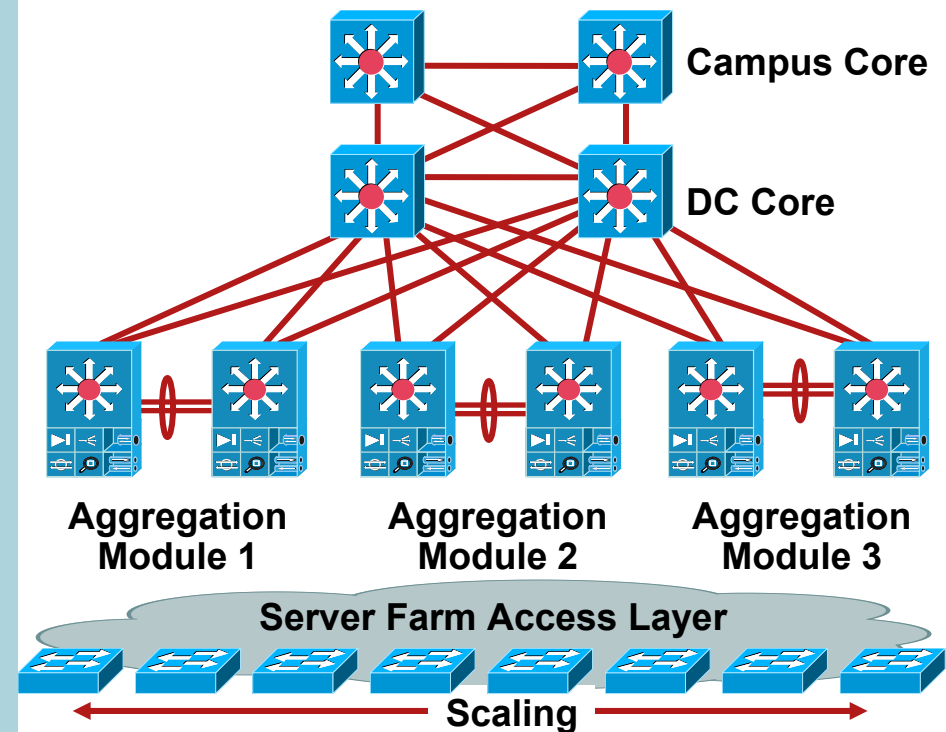
- Align server access to primary components in aggregation layer:
 - vlan root
 - primary HSRP instance
 - active service context
- Provides more predictable design
 - Simplifies troubleshooting
- More efficient traffic flow
 - Decreases chance of flow ping-pong across inter-switch link



Aggregation Layer Design

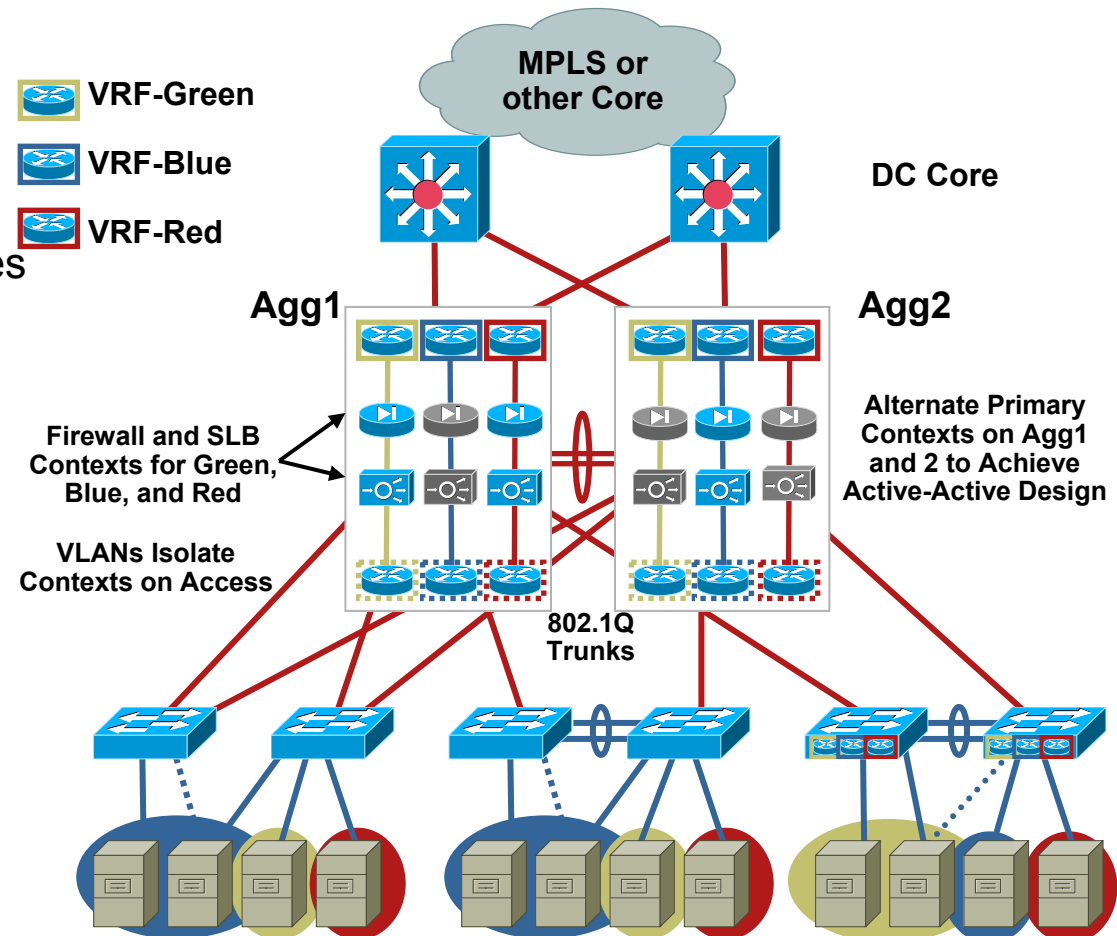
Scaling the Aggregation Layer

- Aggregation modules provide:
 - Spanning tree scaling
 - HSRP Scaling
 - Access layer density
 - 10GE/GEC uplinks
 - Application services scaling
 - SLB/firewall
 - Fault domain sizing
- Core layer provides inter-agg module transport:
 - Low latency distributed forwarding architecture (use DFC's)
 - 100,000's PPS forwarding rate
 - Provides inter-agg module transport medium in multi-tier model



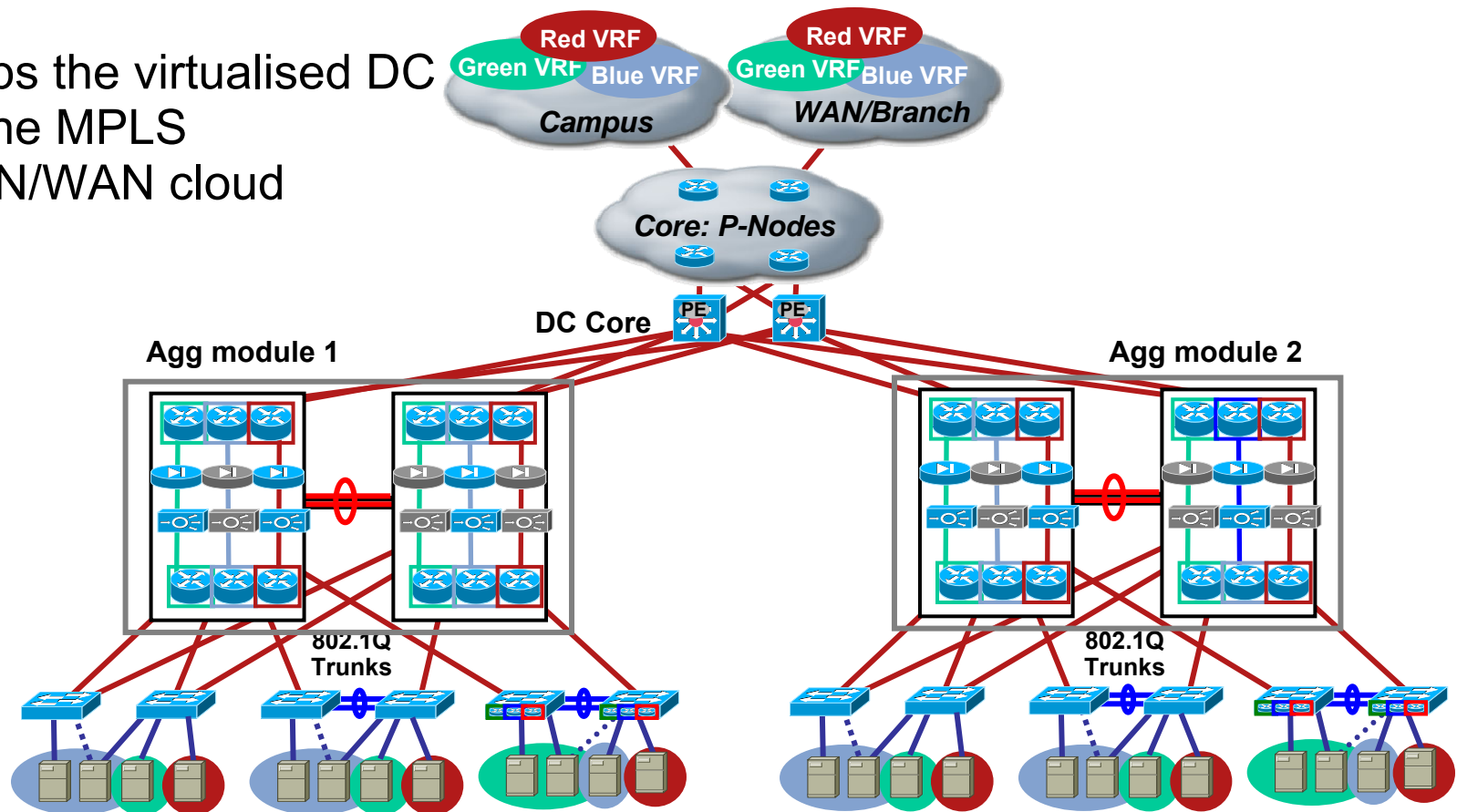
Aggregation Layer Design Using VRFs in the DC (1)

- Enables virtualisation/partitioning of network resources (MSFC, ACE, FWSM)
- Permits use of application services with multiple access topologies
- Maps well to path isolation MAN/WAN designs such as with MPLS
- Security policy management and deployment by user group/vrf



Aggregation Layer Design Using VRFs in the DC (2)

- Maps the virtualised DC to the MPLS MAN/WAN cloud



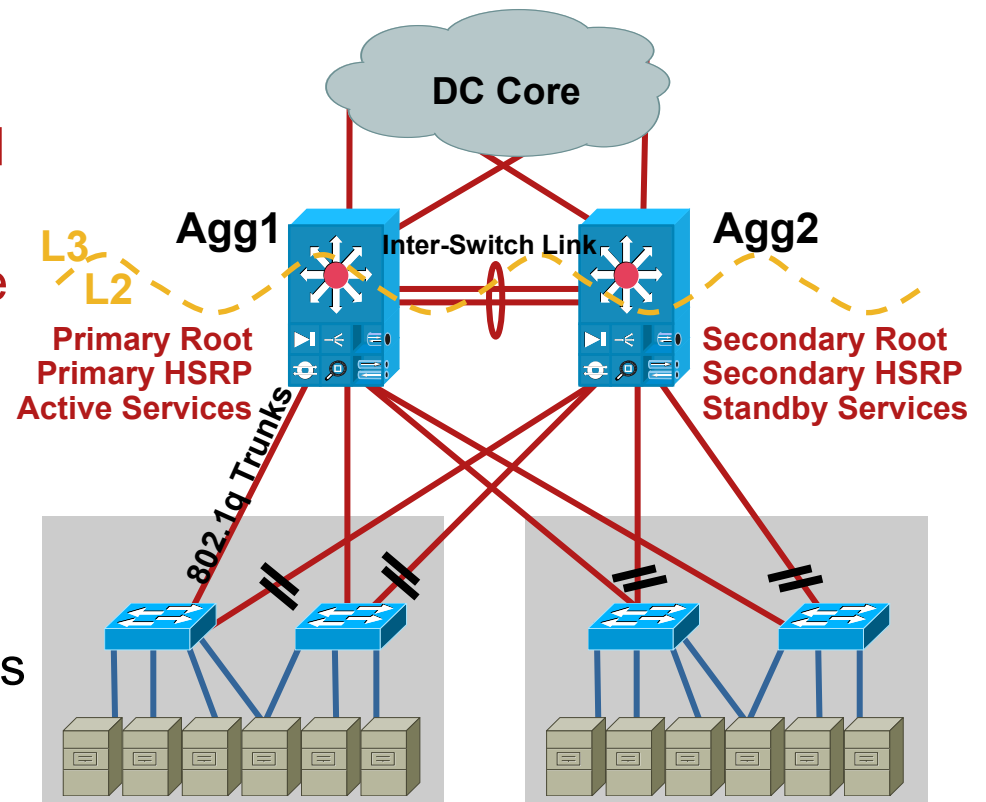
Access Layer Design



Access Layer Design

Defining Layer 2 Access

- L3 routing is first performed in the aggregation layer
- L2 topologies consist of **looped and loop-free models**
- **Stateful services at Agg can be provided across the L2 access (FW, SLB, SSL, etc.)**
- VLANs are extended across inter-switch link trunk for looped access
- VLANs **are not** extended across inter-switch link trunk for loop-free access



Access Layer Design

Establish a Deterministic Model

- Align active components in traffic path on common Agg switch:

Primary STP root

```
Aggregation-1(config)#spanning-tree vlan 1-10  
root primary
```

Primary HSRP (outbound)

```
standby 1 priority X
```

Active Service modules/contexts

Path Preference (inbound)

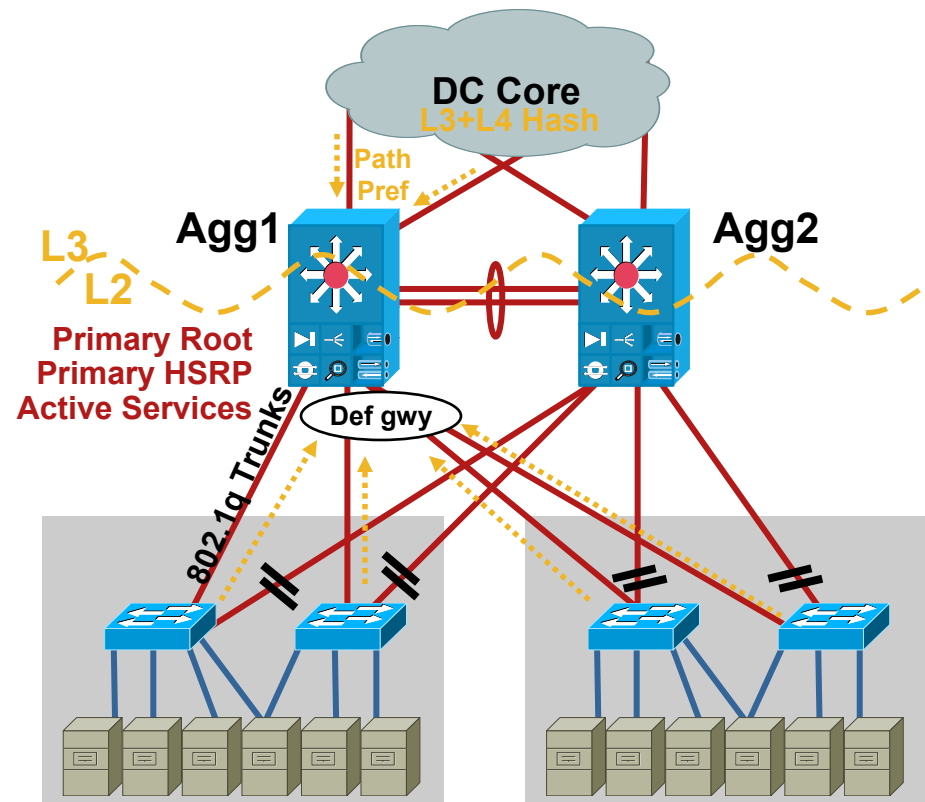
Use RHI & Route map

```
Route-map preferred-path
```

```
match ip address x.x.x.x
```

```
set metric -30
```

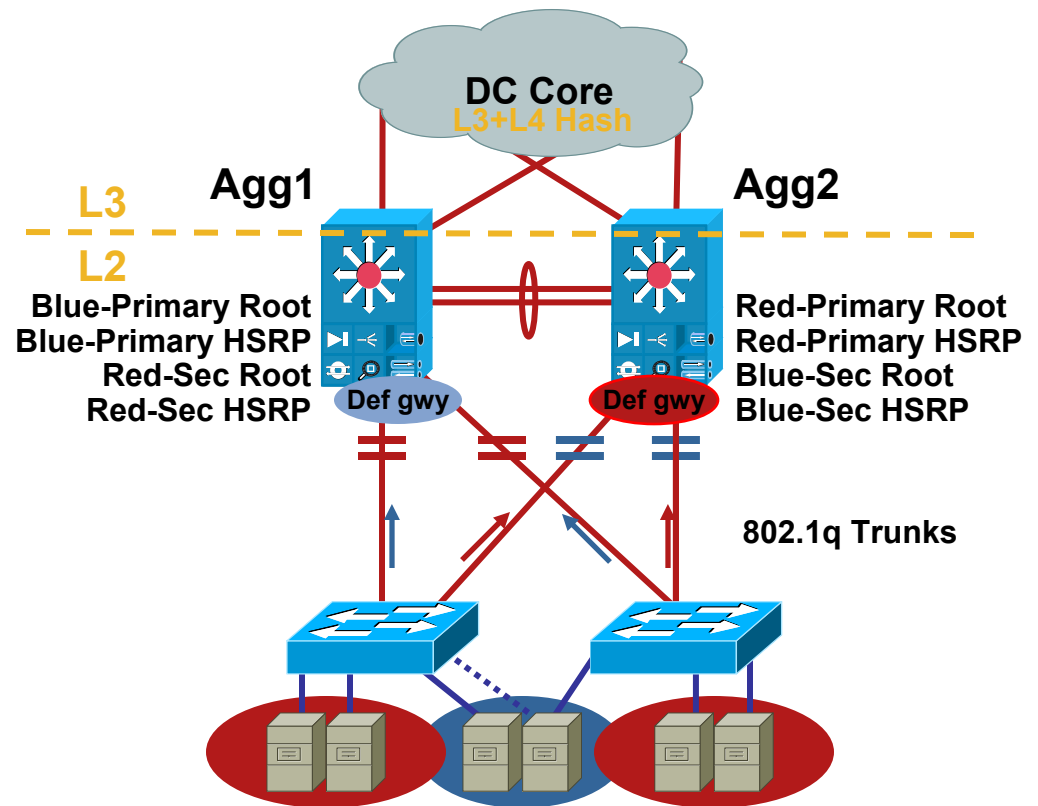
(also see slide 15)



Access Layer Design

Balancing VLANs on Uplinks: L2 Looped Access

- Distributes traffic load across uplinks
- STP blocks uplink path for VLANs to secondary root switch
- If active/standby service modules are used; consider inter-switch link utilisation to reach active instance
 - Multiple service modules may be distributed on both agg switches to achieve balance
 - Consider b/w of inter-switch link in a failure situation
- If active/active service modules are used;
 - (ACE and FWSM3.1)
 - Balance contexts+hsrp across agg switches
 - Consider establishing inbound path preference



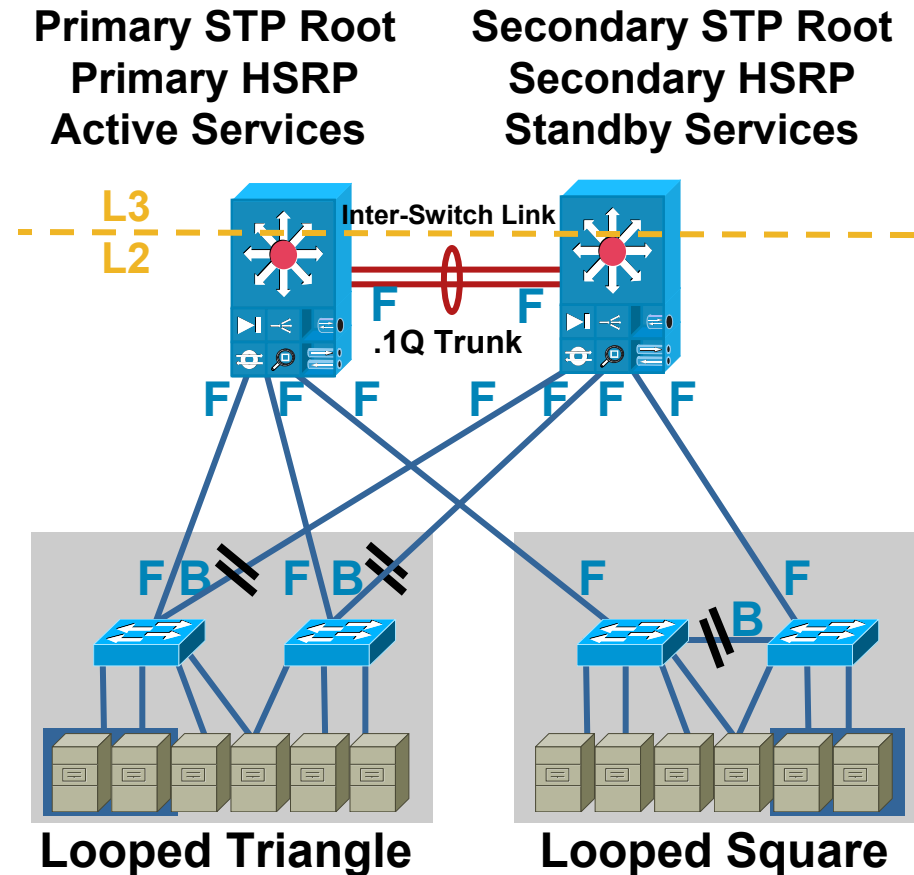
Access Layer Design: L2 Looped Design Model



Access Layer Design

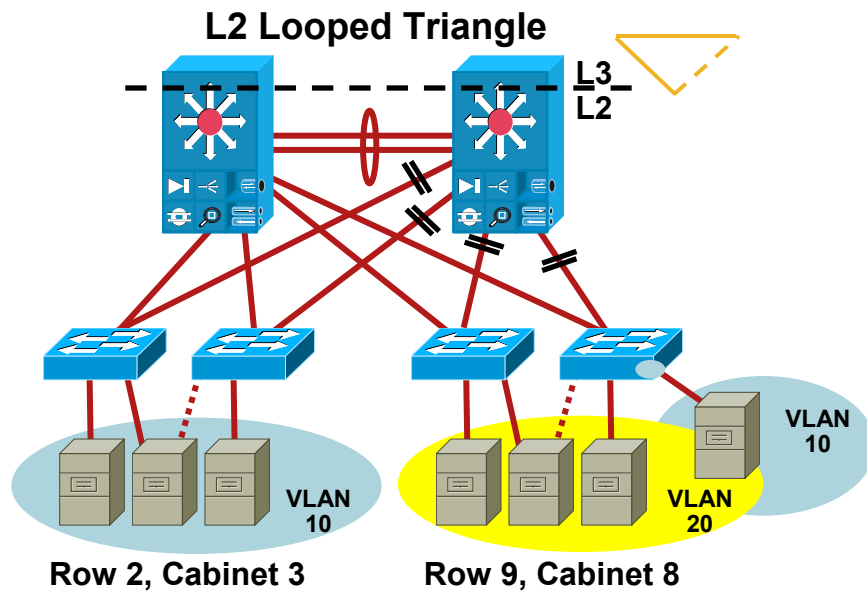
Looped Design Model

- VLANs are extended between aggregation switches, creating the looped topology
- Spanning Tree is used to prevent actual loops (Rapid PVST+, MST)
- Redundant path exists through a second path that is blocking
- **Two looped topology designs:**
Triangle and square
- VLANs may be load balanced across access layer uplinks
- Inter-switch link utilisation must be considered as this may be used to reach active services



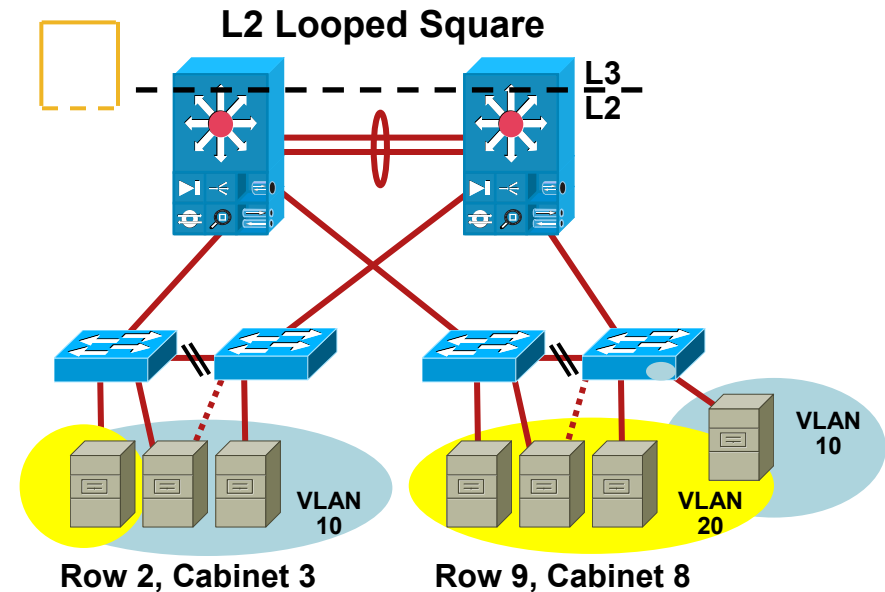
Access Layer Design

L2 Looped Topologies



Looped Triangle Access

- Supports VLAN extension/L2 adjacency across access layer
- Resiliency achieved with dual homing and STP
- Quick convergence with 802.1W/S
- Supports stateful services at aggregation layer
- Proven and widely used



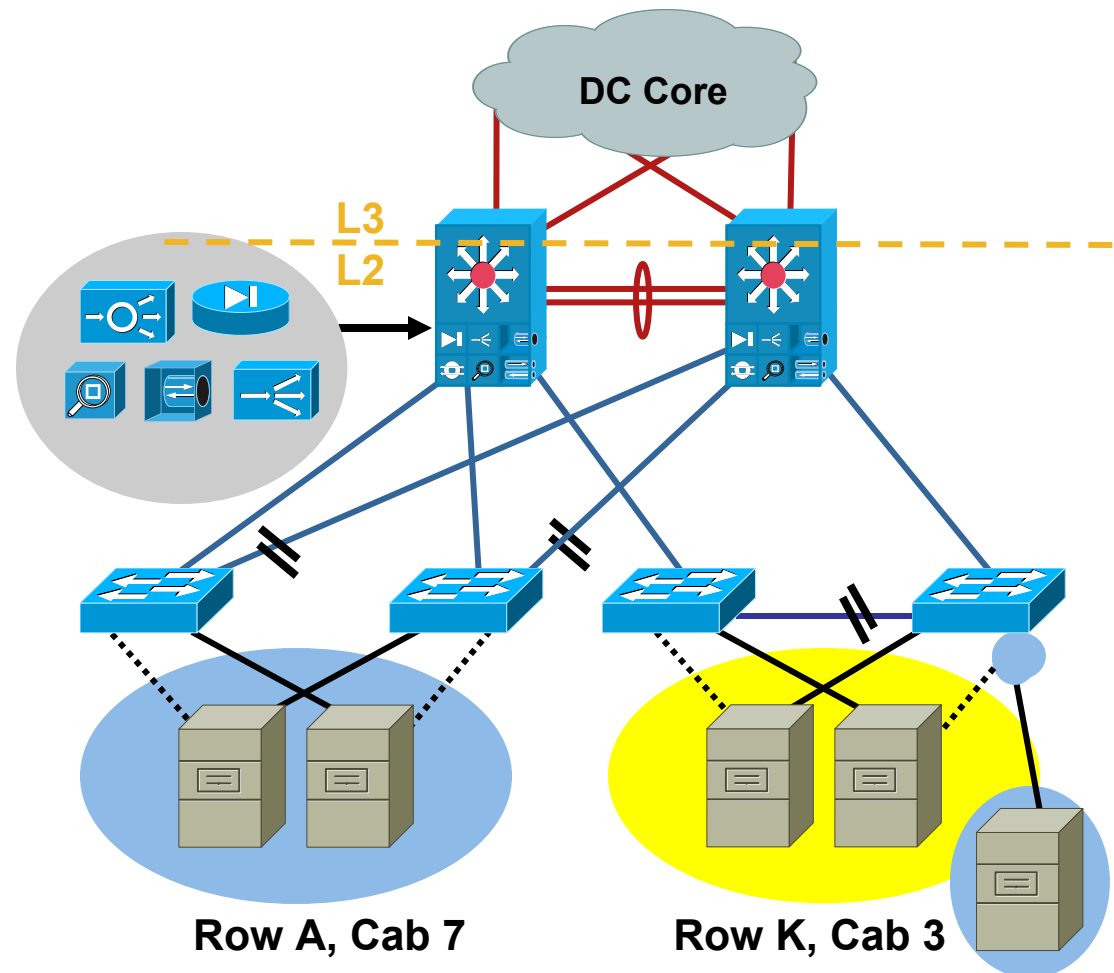
Looped Square Access

- Supports VLAN extension/L2 adjacency across access layer
- Resiliency achieved with dual homing and STP
- Quick convergence with 802.1W/S
- Supports stateful services at aggregation layer
- Active-active uplinks align well to ACE/FWSM 3.1
- Achieves higher density access layer, optimising 10GE aggregation layer density

Access Layer Design

Benefits of L2 Looped Design

- Services like firewall and load balancing can easily be deployed at the aggregation layer and shared across multiple access layer switches
- VLANs are primarily contained between **pairs** of access switches but—
- VLANs may be extended to different access switches to support
 - NIC teaming
 - Clustering L2 adjacency
 - Administrative reasons
 - Geographical challenges



Access Layer Design

Drawbacks of Layer 2 Looped Design

- Main drawback: if a loop occurs the network may become unmanageable due to the infinite replication of frames
- 802.1w Rapid PVST+ combined with STP related features and best practices improve stability and help to prevent loop conditions

UDLD

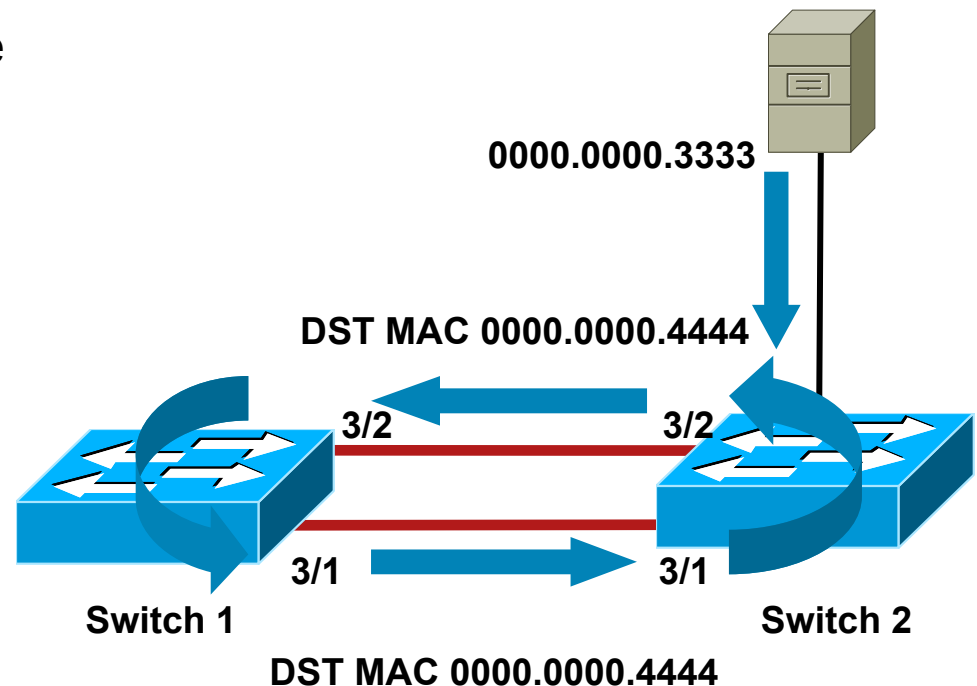
Loopguard

Rootguard

BPDUGuard

Limited domain size

Stay under STP watermarks for logical and virtual ports



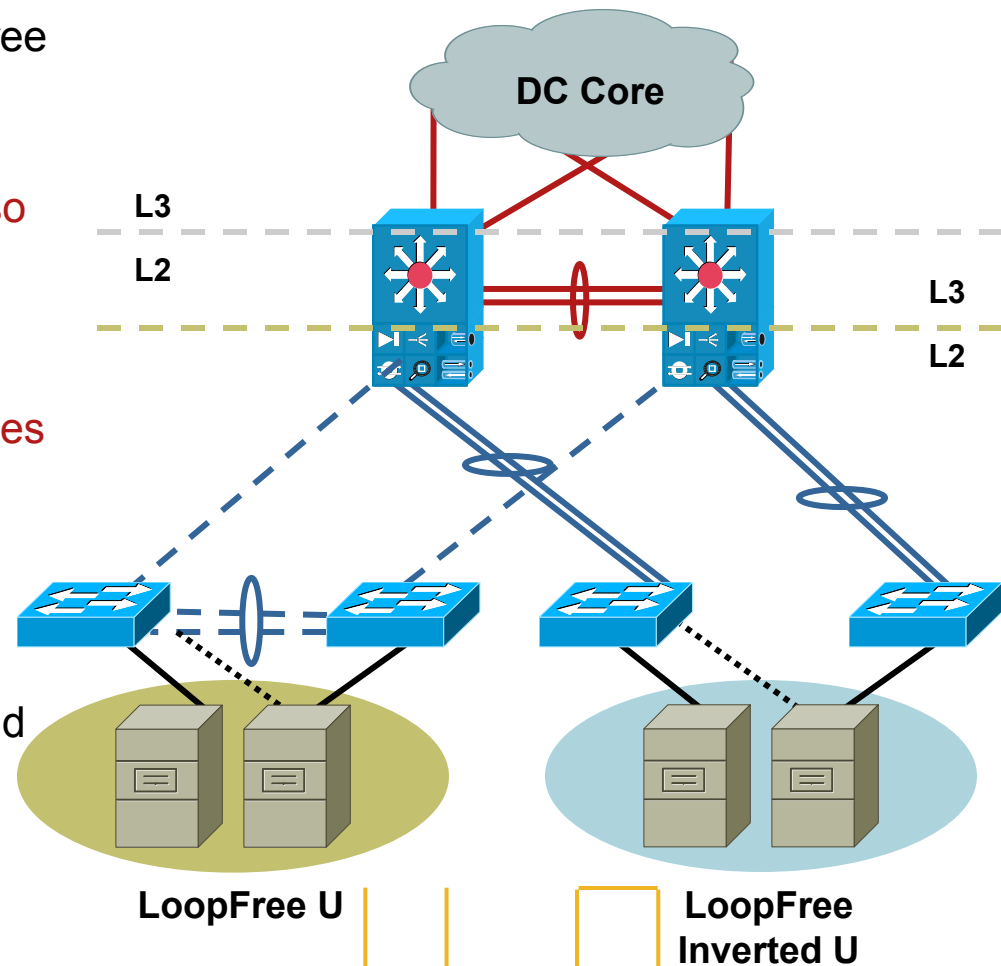
Access Layer Design: L2 LoopFree Design Model



Access Layer Design

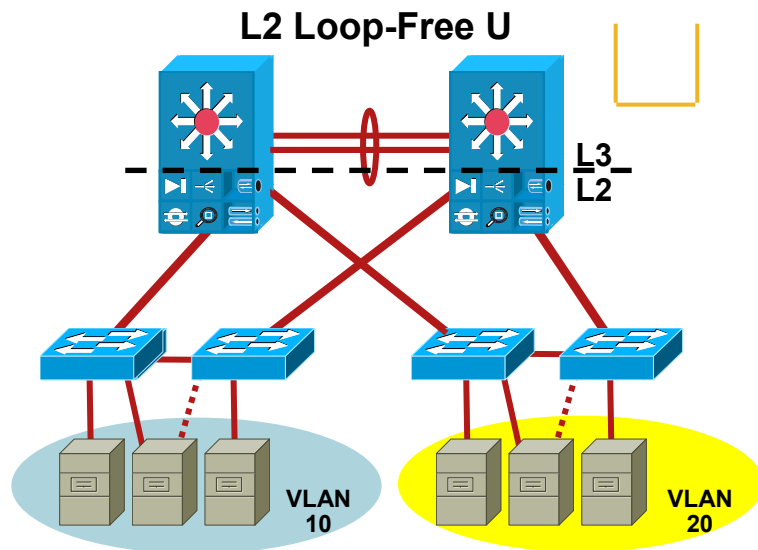
Loop-Free Design

- Alternative to looped design
- 2 Models: LoopFree U and LoopFree Inverted U
- **Benefit: Spanning Tree is enabled but no port is blocking so all links are forwarding**
- **Benefit: less chance of loop condition due to misconfigurations or other anomalies**
- L2-L3 boundary varies by loop-free model used: U or Inverted-U
- Implication considerations with service modules, L2 adjacency, and single attached servers



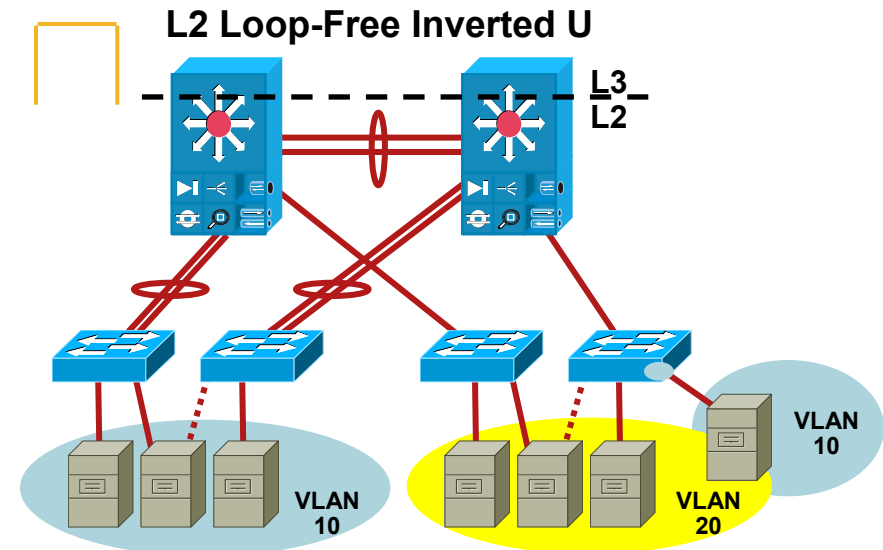
Access Layer Design

Loop-Free Topologies



Loop-Free U Access

- VLANs contained in switch pairs (no extension outside of switch pairs)
- No STP blocking; all uplinks active
- Autostate implications for certain service modules (CSM)
- ACE supports autostate and per context failover

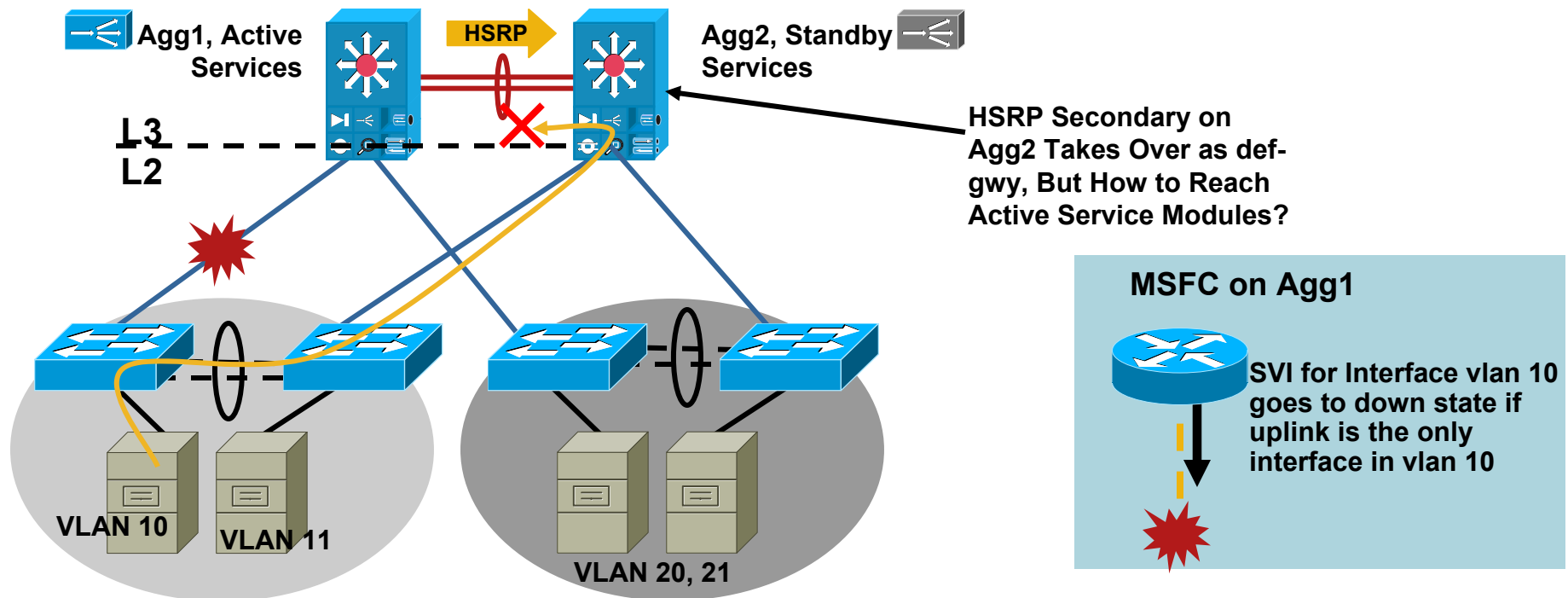


Loop-Free Inverted U Access

- Supports VLAN extension
- No STP blocking; all uplinks active
- Access switch uplink failure black holes single attached servers
- Supports all service module implementations

Access Layer Design

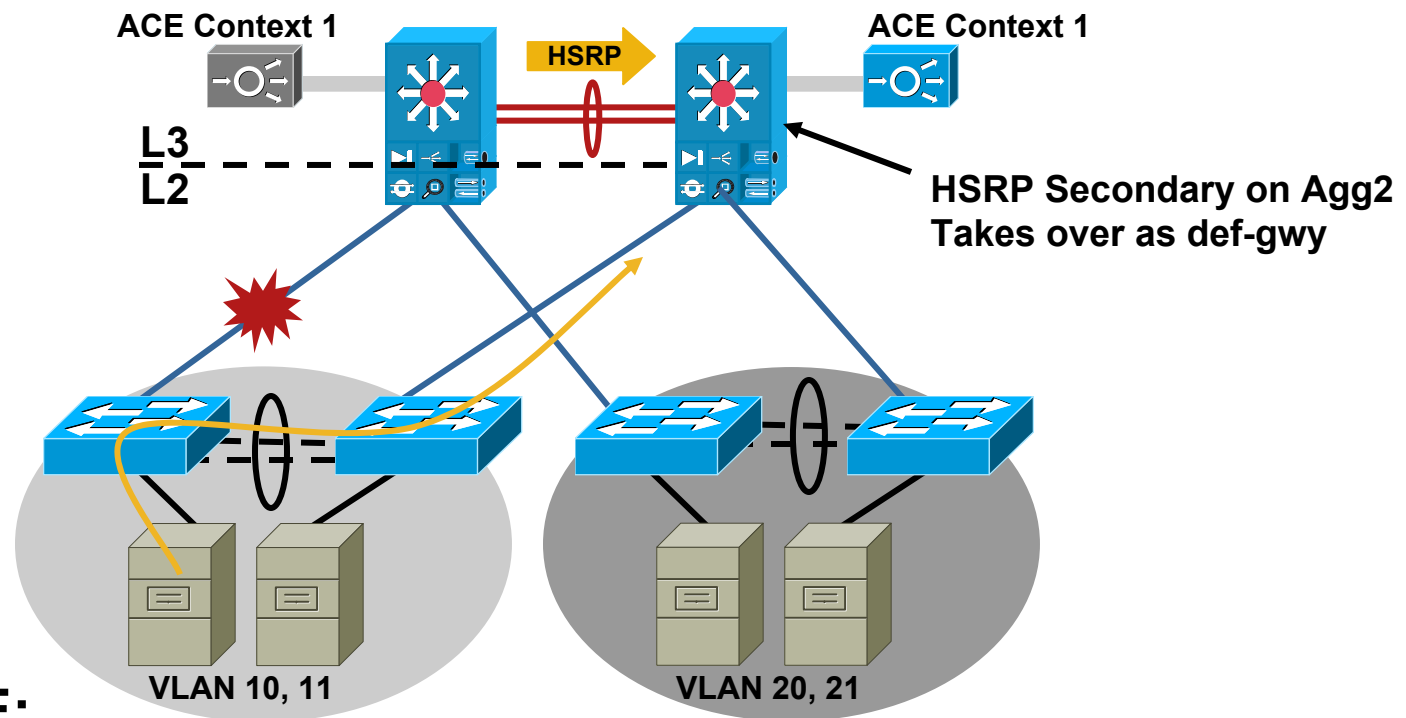
Loop-Free U Design and Service Modules (1)



- If the uplink connecting access and aggregation goes down, the VLAN interface on the MSFC goes down as well due to the way autostate works
- CSM and FWSM has implications as autostate is not conveyed (leaving black hole)
- Tracking and monitoring features may be used to allow failover of service modules based on uplink failure but would you want a service module failover for one access switch uplink failure?
- Not recommended to use loop-free L2 access with active-standby service module implementations. (See slide on ACE next)

Access Layer Design

Loop-Free U Design and Service Modules (2)



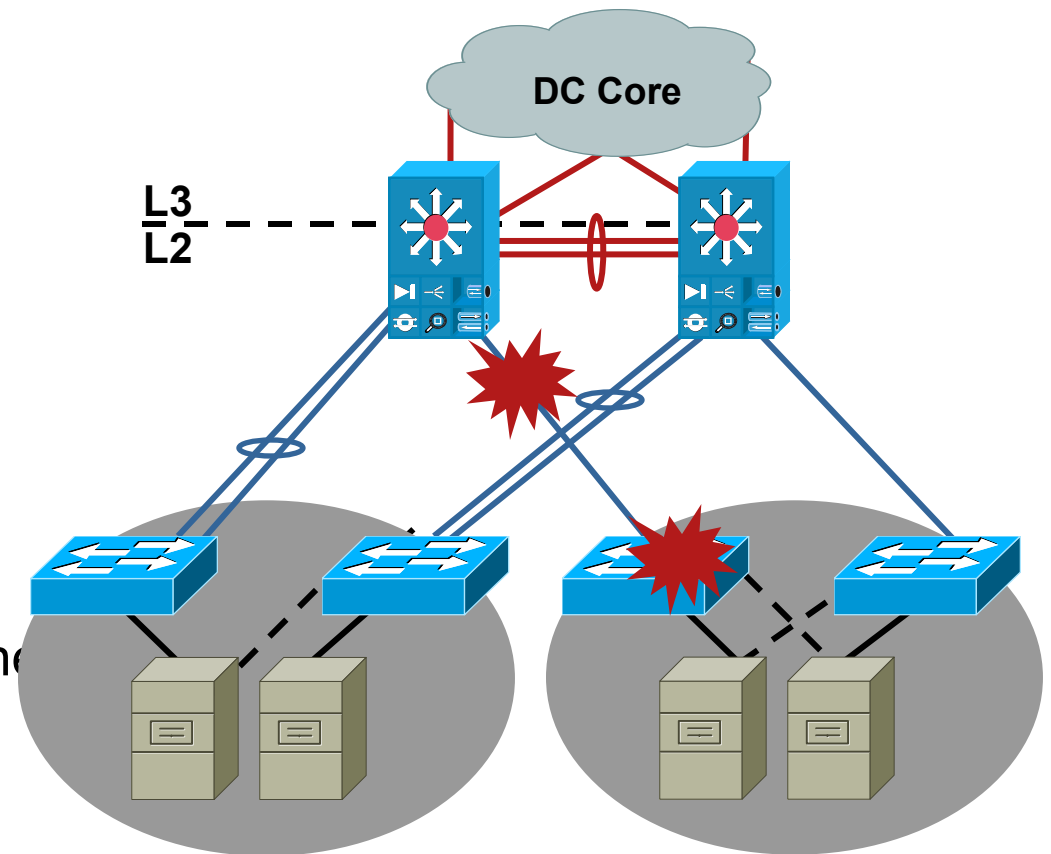
With ACE:

- Per context failover with autostate
- If uplink fails to Agg1, ACE can switchover to Agg2 (under 1sec)
- Requires ACE on access trunk side for autostate failover
- May be combined with FWSM3.1 for active-active design

Access Layer Design

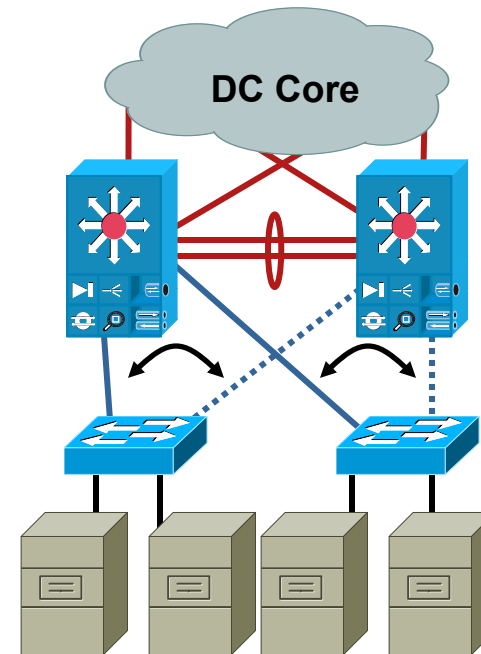
Drawbacks of Loop-Free Inverted-U Design

- Single attached servers are black-holed if access switch uplink fails
- Distributed EtherChannel[®] can reduce chance of black holing
- NIC teaming improves resiliency in this design
- Inter-switch link scaling needs to be considered when using active-standby service modules



Access Layer Design Using FlexLinks in the Data Centre

- Flexlinks are an active-standby pair defined on a common access switch
- Flexlink pairs have STP turned “off” so no BPDU’s are propagated
- Failover in 1-2 second range
- An interface can belong to only 1 FlexLink pair of same or different interface types (GE, 10GE, port-channel, etc.)
- Agg switch is not aware of FlexLinks configured on access switch, links are up and STP logical/virtual ports are active/allocated
- Supported as of 12.2.18SXF
- An alternative to triangle loop design **but must consider risk of possible loop** (see next slides)



```
Router# configure terminal
Router(conf)# interface Gigabit1/1
Router(conf-if)# switchport backup interface Gigabit1/2
Router(conf-if)# exit
```

```
Router# show interface switchport backup
Router Backup Interface Pairs:
Active Interface Backup Interface State
```

```
-----
Gigabit1/1 Gigabit1/2 Active Up/Backup Standby
Port-channel1 Gigabit7/1 Active Up/Backup Standby
Gigabit7/2 Gigabit7/3 Active Up/Backup Standby
```

Access Layer Design

Considerations when using FlexLinks (1)

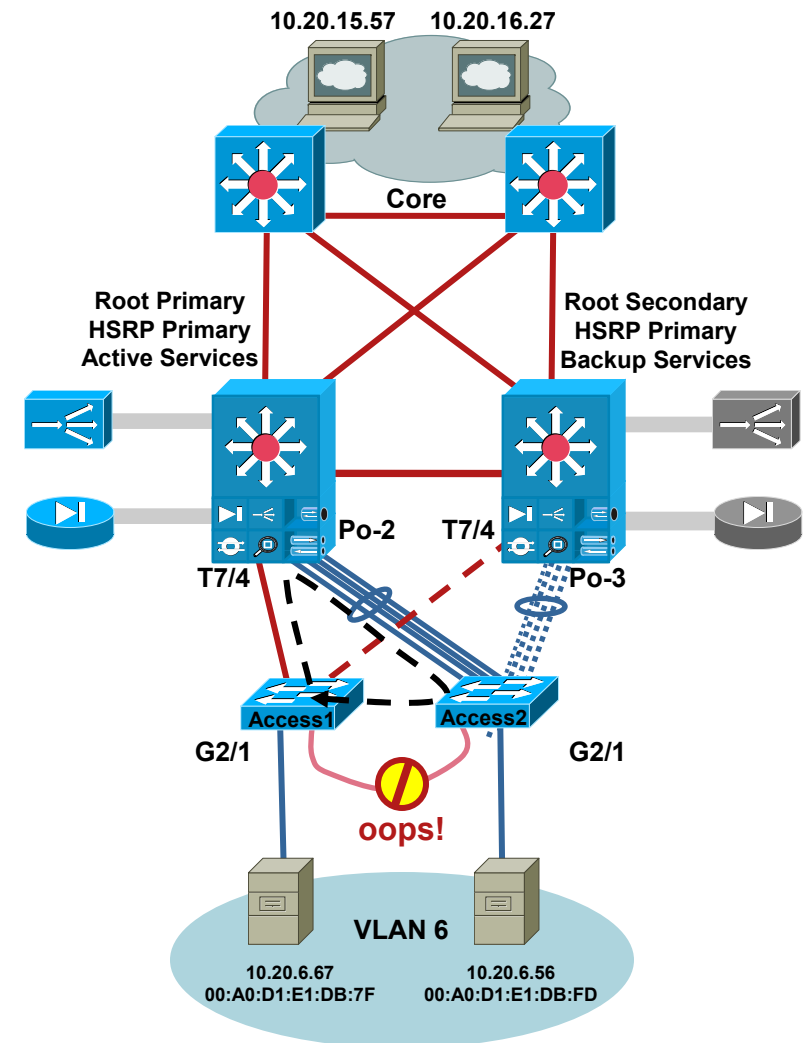
Loop Considerations:

STP is disabled so it is not a backup safety mechanism for possible loop conditions

If a loop is accidentally placed between access switch ports in same vlan—BPDUGuard will catch:

Apr 13 16:07:33: %PM-SP-4-ERR_DISABLE: bpduguard error detected on Gi2/2, putting Gi2/2 in err-disable state

NOTE: Without BPDU Guard on access ports, the Aggregation switch will be negatively affected with high CPU condition, hsrp flapping and other negative conditions



Access Layer Design Considerations when using FlexLinks (2)

- Loop Considerations:

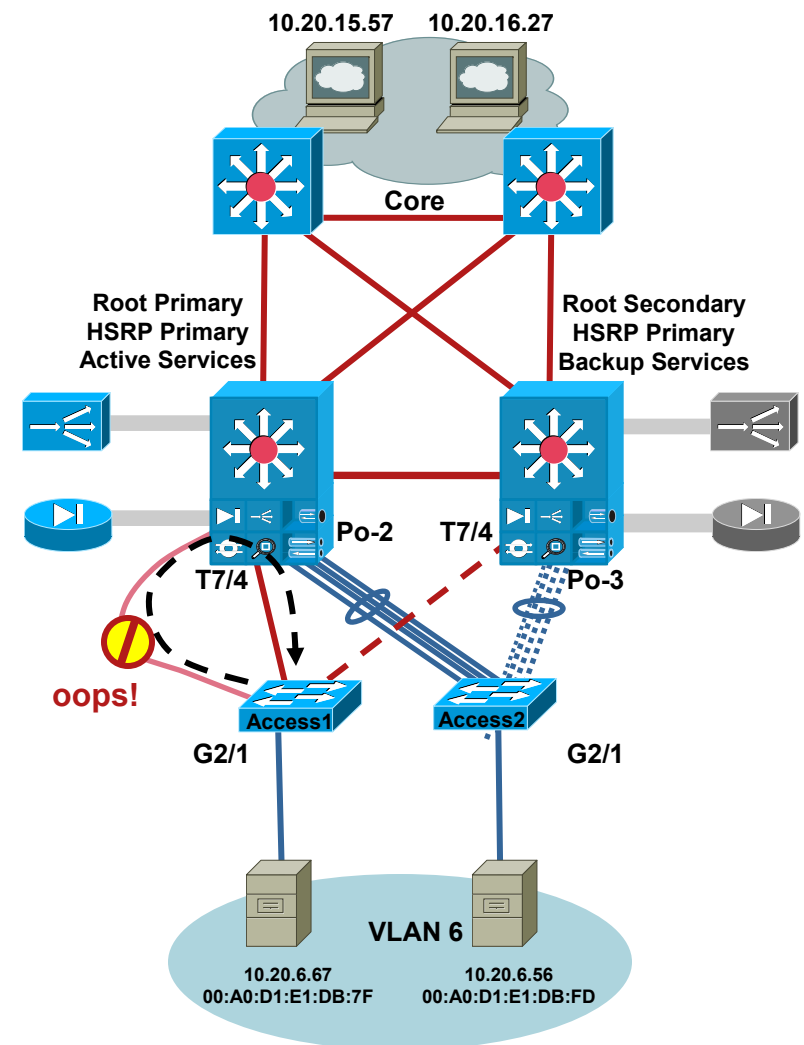
If a loop is accidentally placed between the access switch and the aggregation switch a loop condition will occur.

NOTE: This will create negative affects with high CPU condition, hsrp flapping and other conditions

- Other Considerations






No preempt: need to consider inter-switch link b/w for failover situations

Backup link is unused and in standby state which doesn't align to active/active service module designs



Access Layer Design

Comparing Looped, Loop-Free and Flexlinks

	Uplink vlans on Agg Switch in Blocking or Standby State	VLAN Extension Supported Across Access Layer	Service Module Black-Holing on Uplink Failure (5)	Single Attached Server Black-Holing on Uplink Failure	Access Switch Density per Agg Module	Must Consider Inter-Switch Link Scaling
 Looped Triangle	-	+	+	+	-	(3) +
 Looped Square	+	+	+	+	+	-
 Loop-Free U	+	-	(4) -	+	+	+
 Loop-Free Inverted U	+	+	+	(1,2) +/-	+	-
 FlexLinks	-	+	+	+	-	+

1. Use of Distributed EtherChannel Greatly Reduces Chances of Black Holing Condition
2. NIC Teaming Can Eliminate Black Holing Condition
3. When Service Modules Are Used and Active Service Modules Are Aligned to Agg1
4. ACE Module Permits L2 Loopfree Access with per Context Switchover on Uplink failure
5. Applies to when using CSM or FWSM in active/standby arrangement

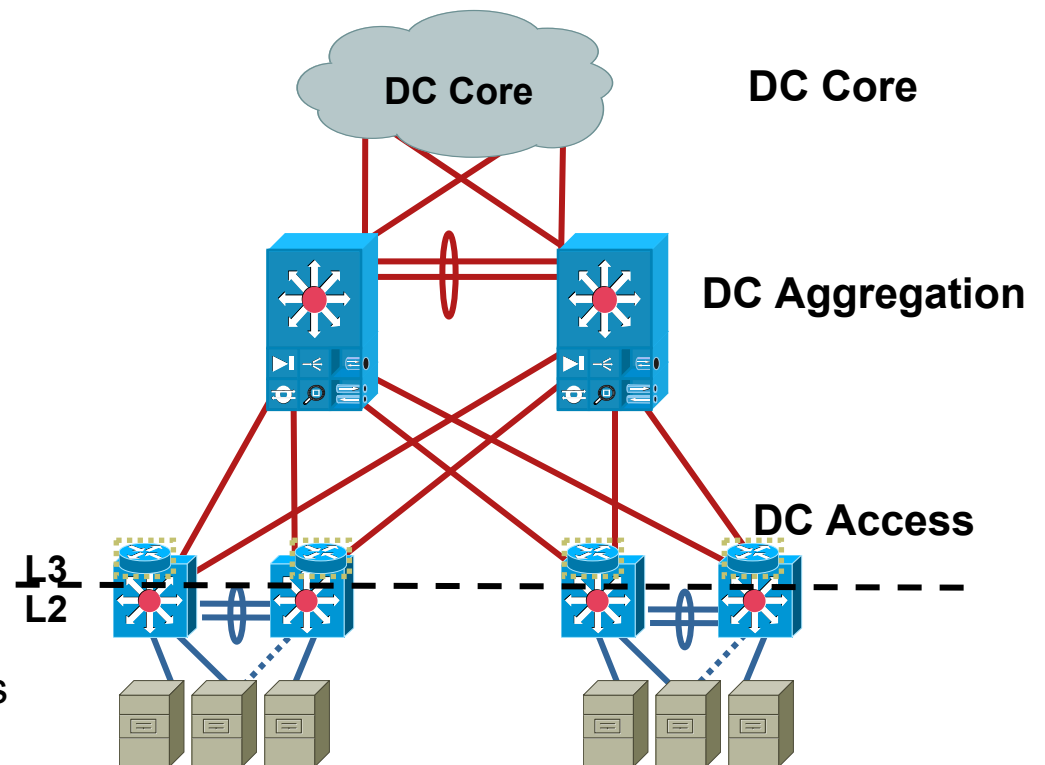
Access Layer Design: L3 Design Model



Access Layer Design

Defining Layer 3 Access

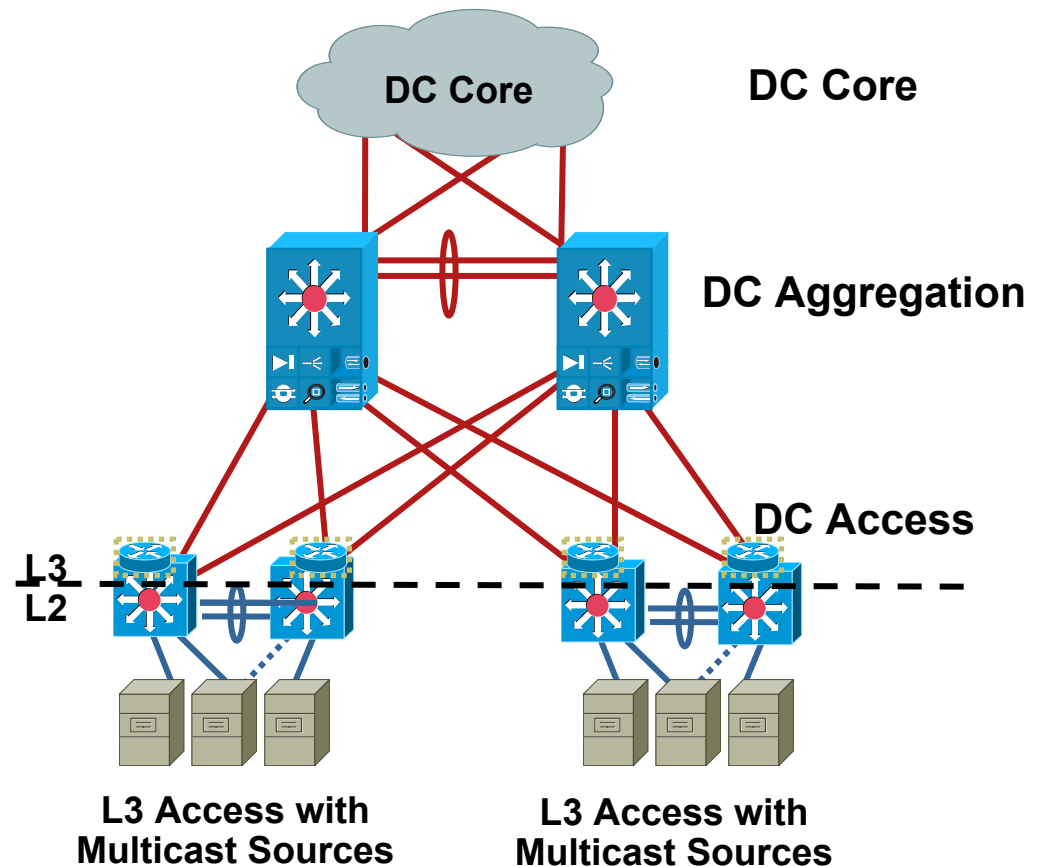
- L3 access switches connect to aggregation with a dedicated subnet
- L3 routing is first performed in the access switch itself
- .1Q trunks between pairs of L3 access switches support **L2 adjacency** requirements (limited to access switch pairs)
- All uplinks are active up to ECMP maximum of 8, no spanning tree blocking occurs
- Convergence time is usually better than Spanning Tree (Rapid PVST+ is close)
- Provides isolation/shelter for hosts affected by broadcasts



Access Layer Design

Need L3 for Multicast Sources?

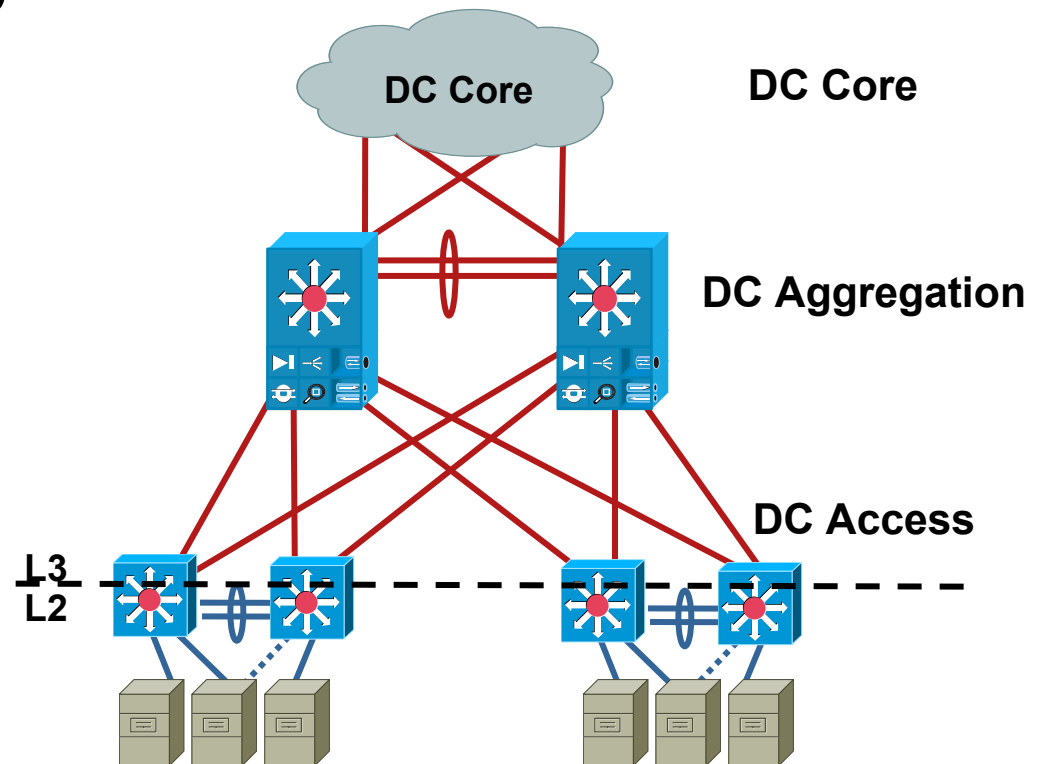
- Multicast sources on L2 access works well when **IGMP snooping** is available
- IGMP snooping at access switch automatically limits multicast flow to interfaces with registered clients in VLAN
- Use L3 when IGMP snooping is not available or when particular L3 administrative functions are required



Access Layer Design

Benefits of L3 Access

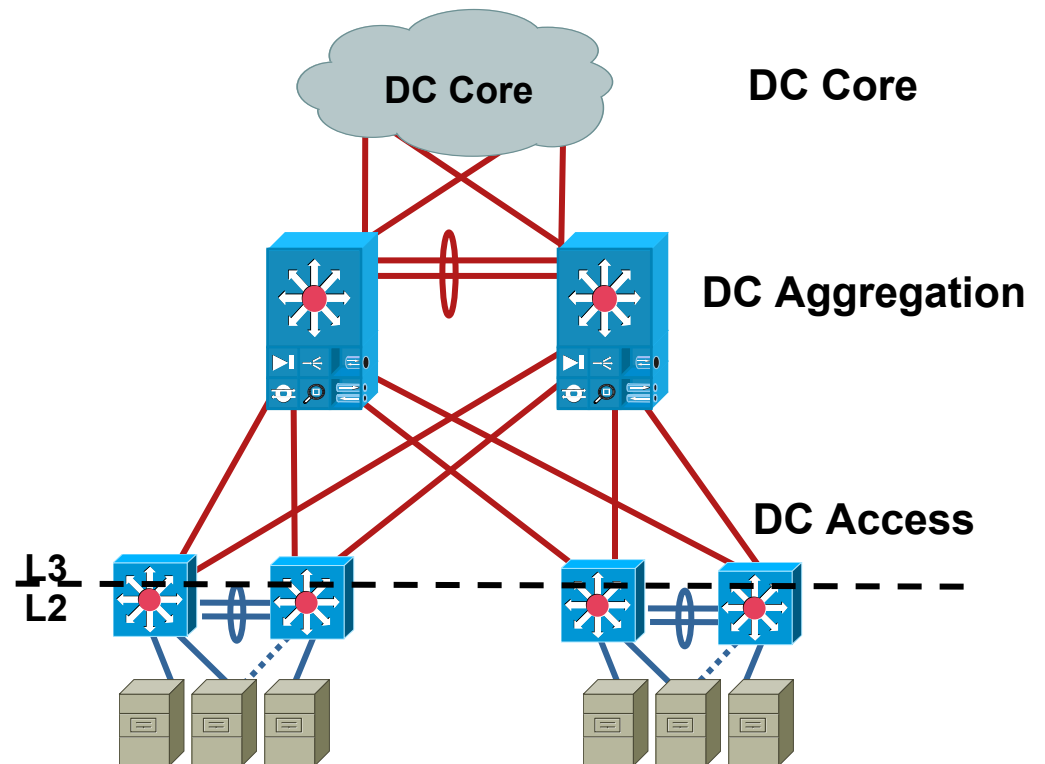
- Minimises broadcast domains attaining high level of stability
- Meet server stability requirements or isolate particular application environments
- Creates smaller failure domains, increasing stability
- All uplinks are available paths, no blocking (up to ECMP maximum)
- Fast uplink convergence: failover and fallback, no arp table to rebuild for agg switches



Access Layer Design

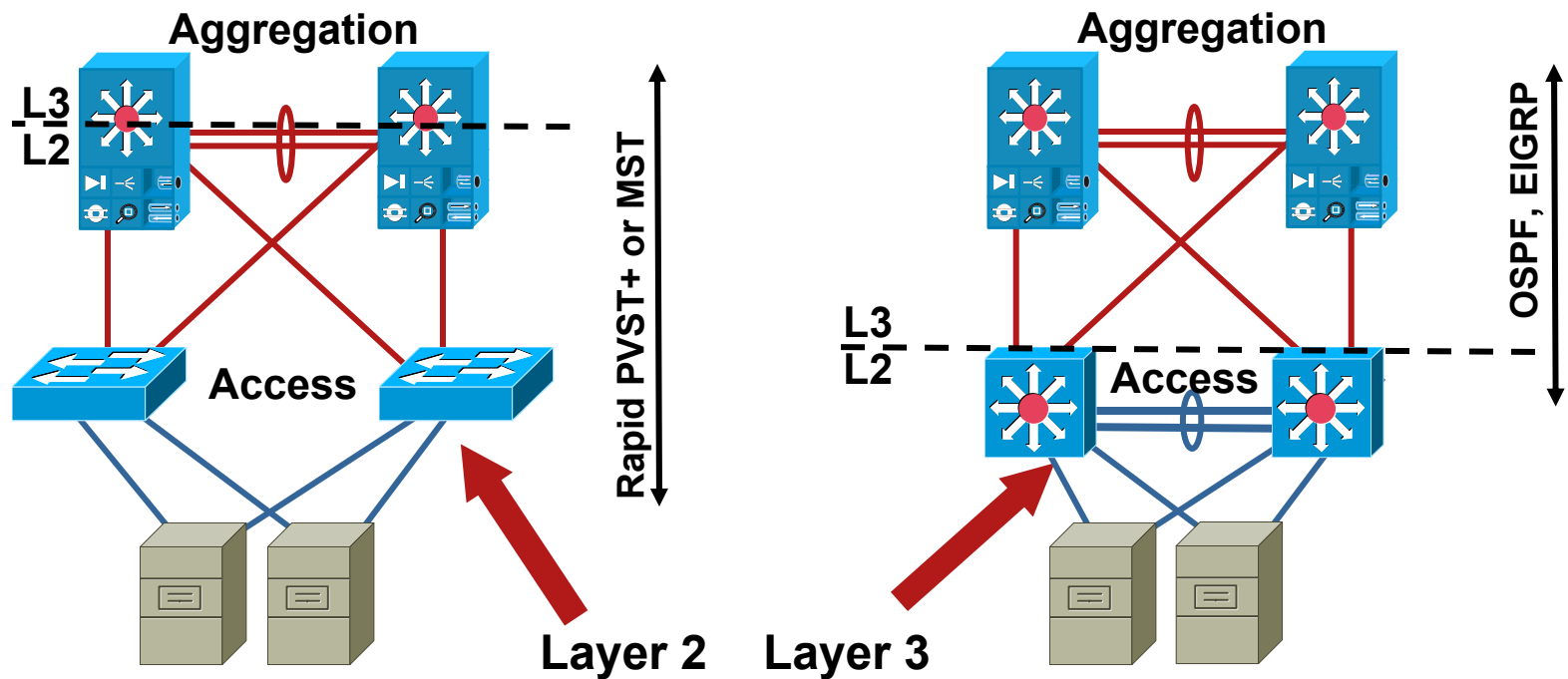
Drawbacks of Layer 3 Design

- L2 adjacency is limited to access pairs (clustering and NIC teaming limited)
- IP address space management is more difficult than L2 access
- If migrating to L3 access, IP address changes may be difficult on servers (may break apps)
- **Would normally require services to be deployed at each access layer pair to maintain L2 adjacency with server and provide stateful failover**



Access Layer Design

L2 or L3? What Are My Requirements?



The Choice of One Design Versus the Other One Has to Do With:

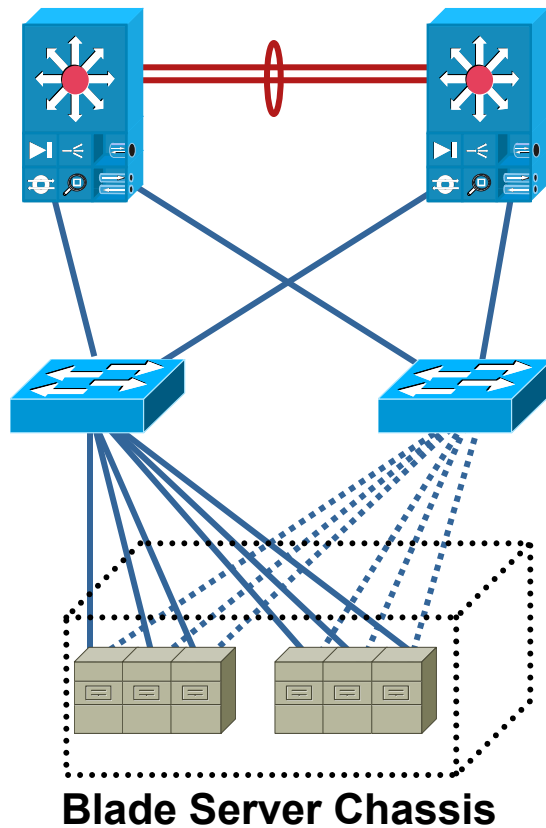
- Difficulties in managing loops
- Staff skillset; time to resolution
- Convergence properties
- NIC teaming; adjacency
- HA clustering; adjacency
- Ability to extend VLANs
- Specific application requirements
- Broadcast domain sizing
- Oversubscription requirements
- Link utilisation on uplinks
- Service module support/placement

Access Layer Design: BladeServers and VMs

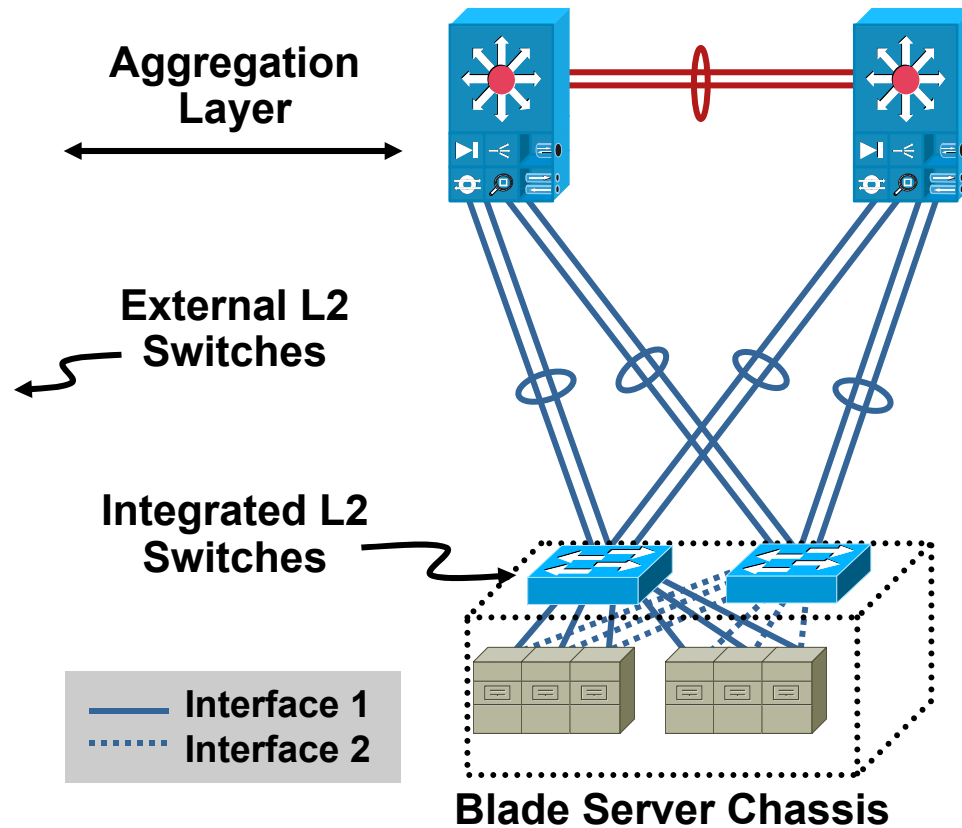


Blade Server Requirements Connectivity Options

Using Pass-Through Modules



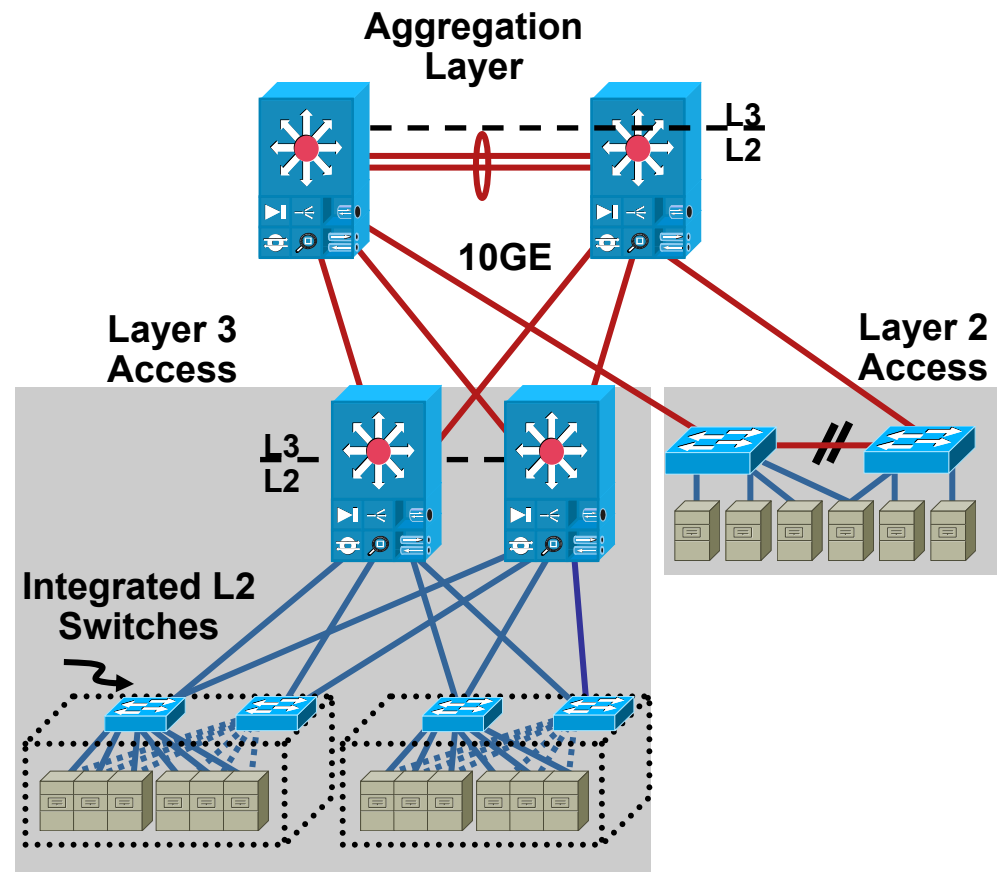
Using Integrated Ethernet Switches



Blade Server Requirements

Connectivity Options

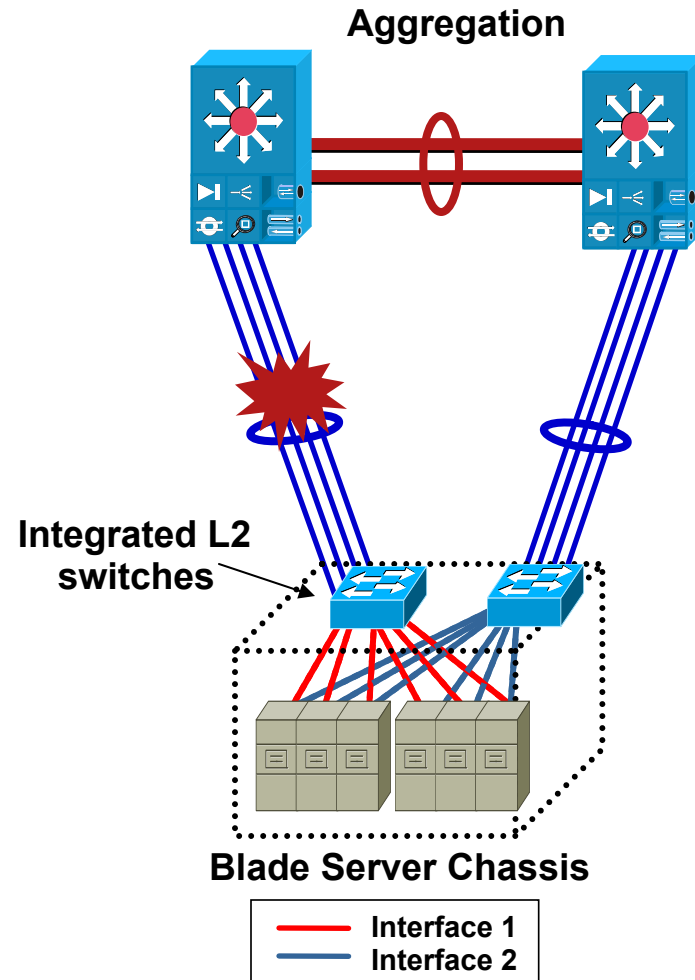
- Layer 3 access tier may be used to aggregate blade server uplinks
 - Permits using 10GE uplinks into agg layer
- Avoid dual tier Layer 2 access designs
 - STP blocking
 - Over-subscription
 - Larger failure domain
- Consider “Trunk Failover” feature of integrated switch



Blade Server Requirements

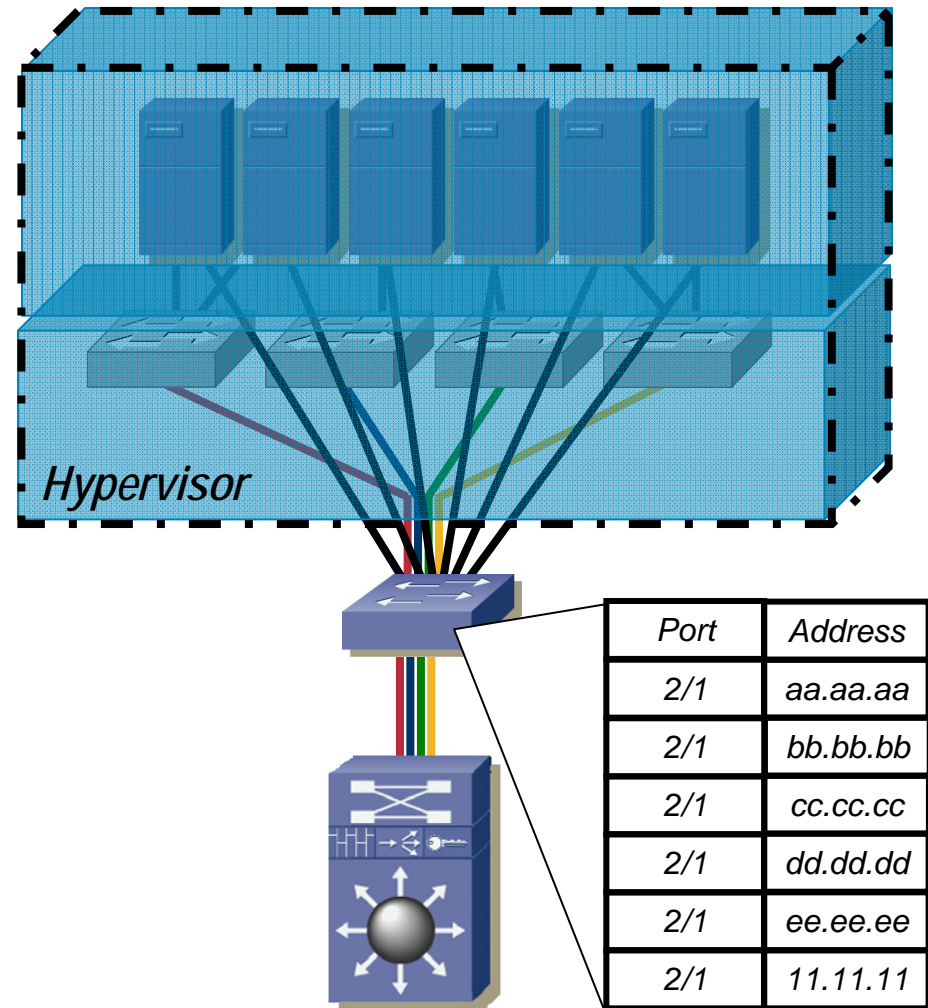
Trunk Failover Feature

- Switch takes down server interfaces if corresponding uplink fails, forcing NIC teaming failover
- Solves NIC teaming limitations; prevents black-holing of traffic
- Achieves maximum bandwidth utilisation:
 - No blocking by STP, but STP is enabled for loop protection
 - Can distribute trunk failover groups across switches
- Dependent upon the NIC feature set for NIC Teaming/failover



Virtual Machine Principles & Considerations

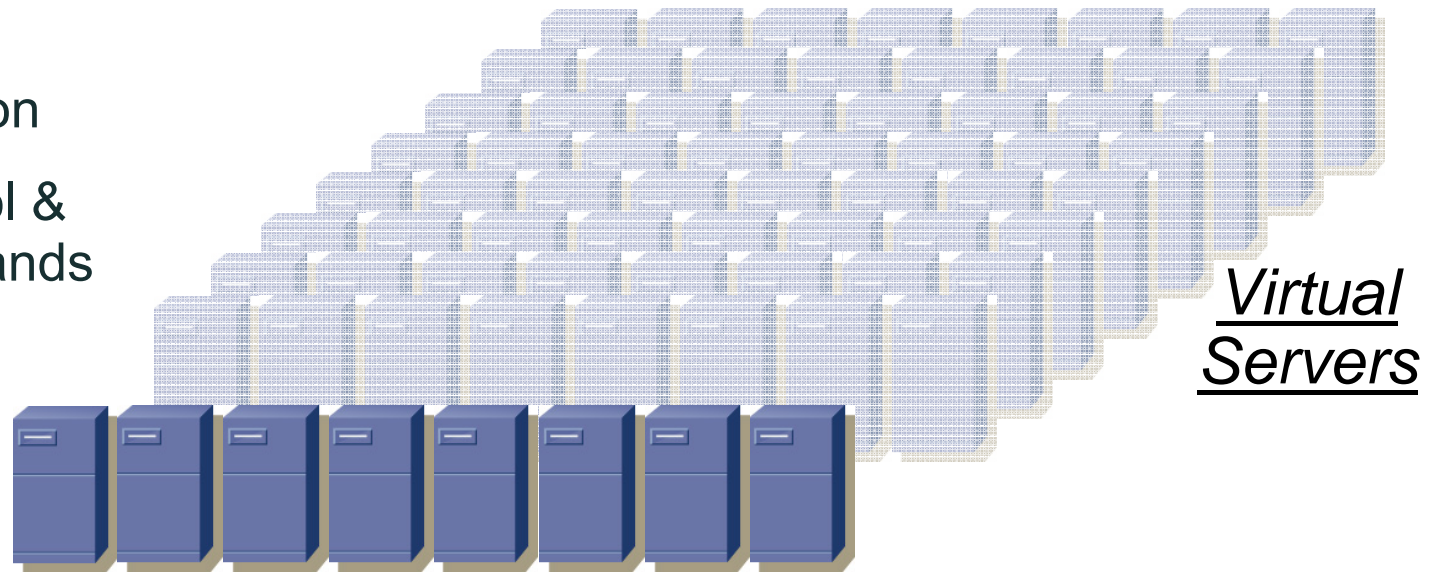
- Virtual Machines enable a single “real” server to host multiple O/S+applications
 - 1:1 → 1:n mapping
- Each Virtual Machine has a unique MAC + IP address network identity
- Rack-optimised servers, Blade Servers and Virtual Machines are driving very dense edge environments



Virtual Machine Considerations

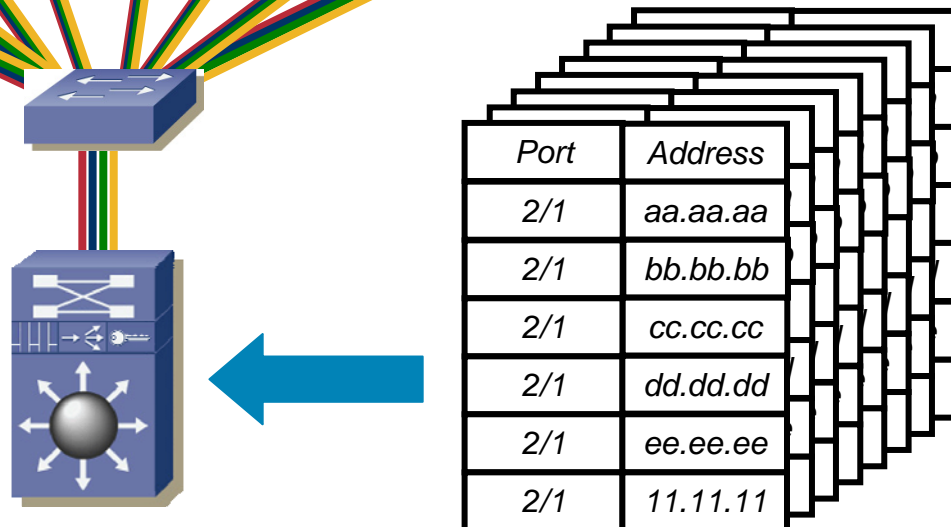
- Virtual Machines multiply network resource utilisation
- Increased Control & Data Plane demands

Real Servers



Virtual Servers

- Scalable Data & Control Plane Attributes...Hardware MAC learning, Dedicated Control buffers, Cache based forwarding, Broadcast Suppression, Layer-2 trace, etc



Density and Scalability Implications in the Data Centre



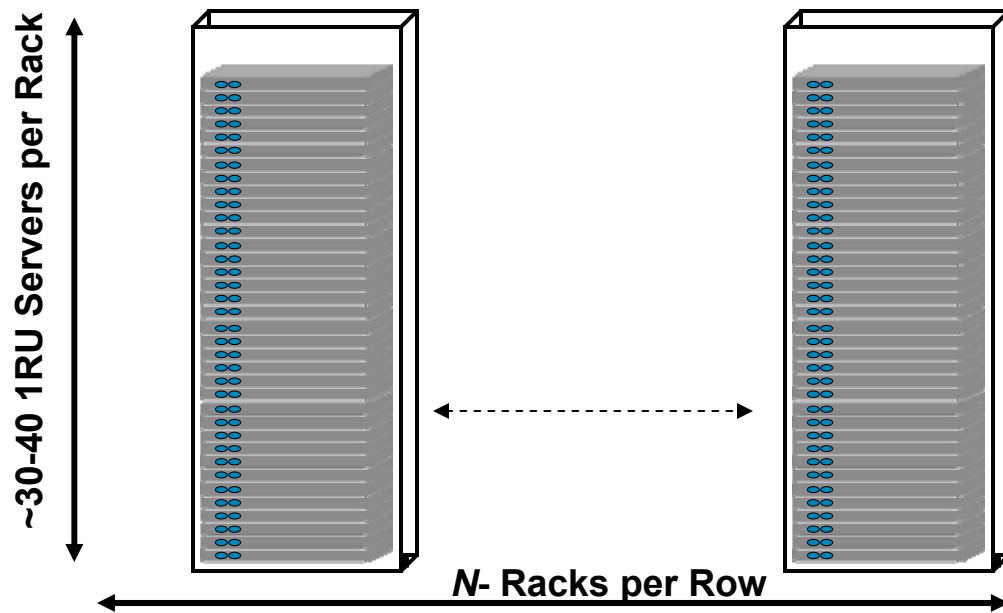
Density and Scalability Implications

Modular or 1RU Access Layer Switching Models

- Where are the issues?
 - Cabling
 - Power
 - Cooling
 - Spanning Tree Scalability
 - Management
 - Oversubscription
 - Sparing
 - Redundancy
- The right solution is usually based on business requirements
- Hybrid implementations can and do exist



Density and Scalability Implications Server Farm Cabinet Layout



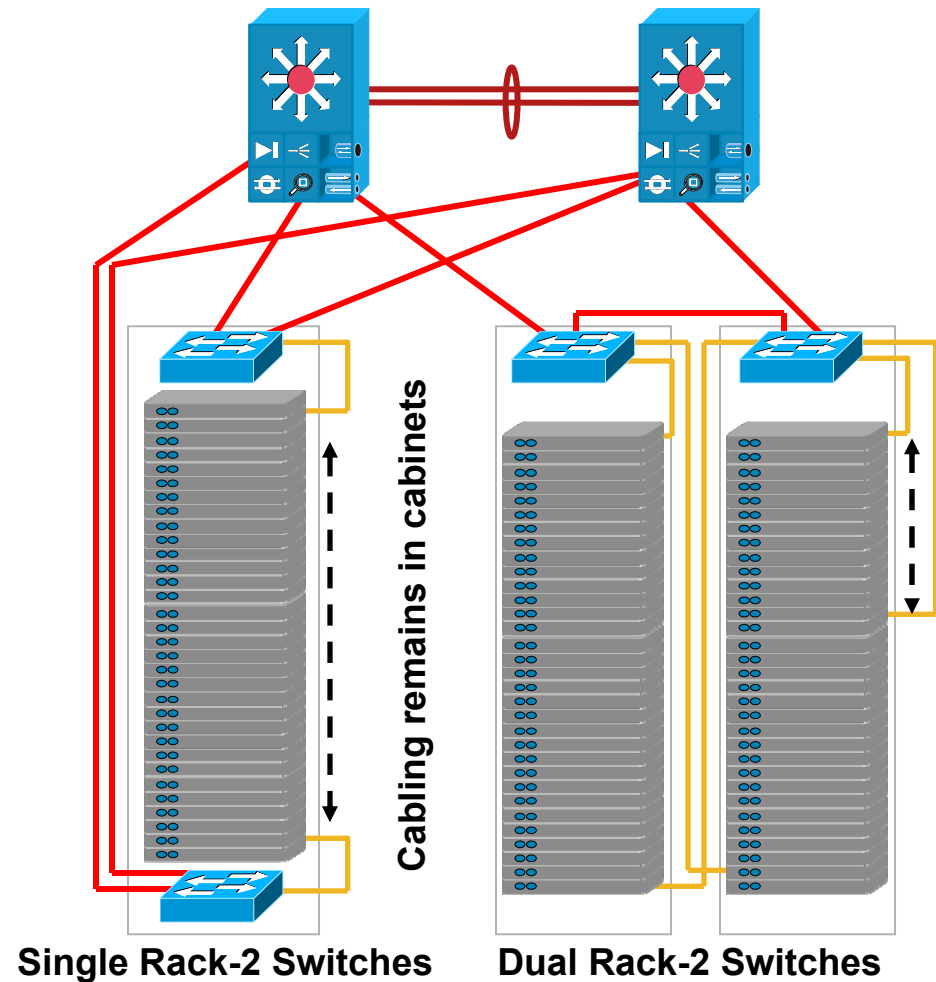
Considerations:

- How Many Interfaces per Server
- Top of Rack Switch placement or End of Row/Within the Row
- Separate Switch for OOB Network
- Cabling overhead vs under floor
- Patch systems
- Cooling capacity
- Power distribution/ Redundancy

Density and Scalability Implications Cabinet Design with 1RU Switching

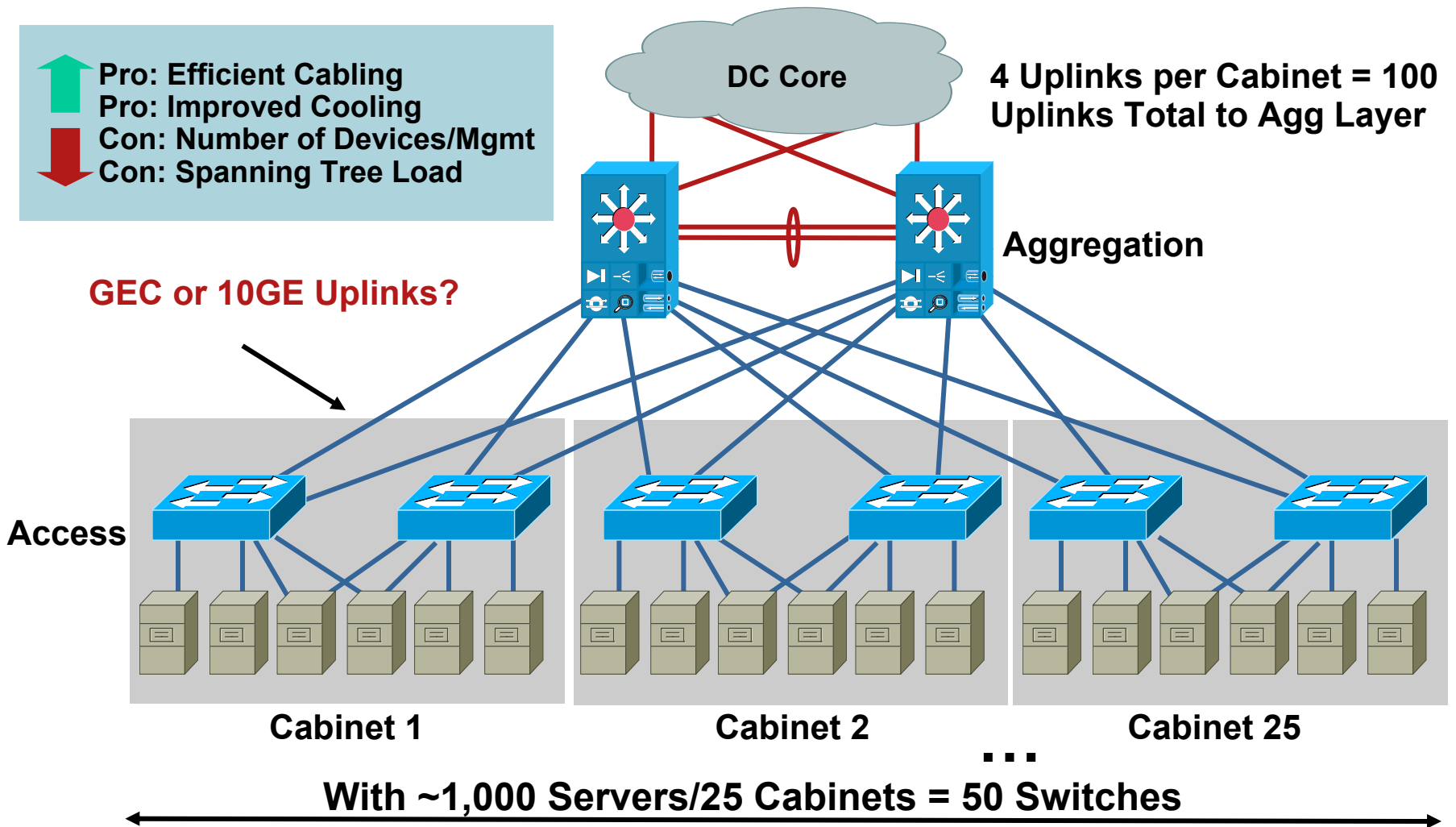
Servers Connect Directly to a 1RU Switch

- Minimises the number of cables to run from each cabinet/rack
- If NIC teaming support: two -1RU switches are required
- Will two 1RU switches provide enough port density?
- Cooling requirements usually do not permit a full rack of servers
- Redundant switch power supply are option
- Redundant switch CPU considerations: not an option
- GEC or 10GE Uplink Considerations



Density and Scalability Implications

Network Topology with 1RU Switching Model

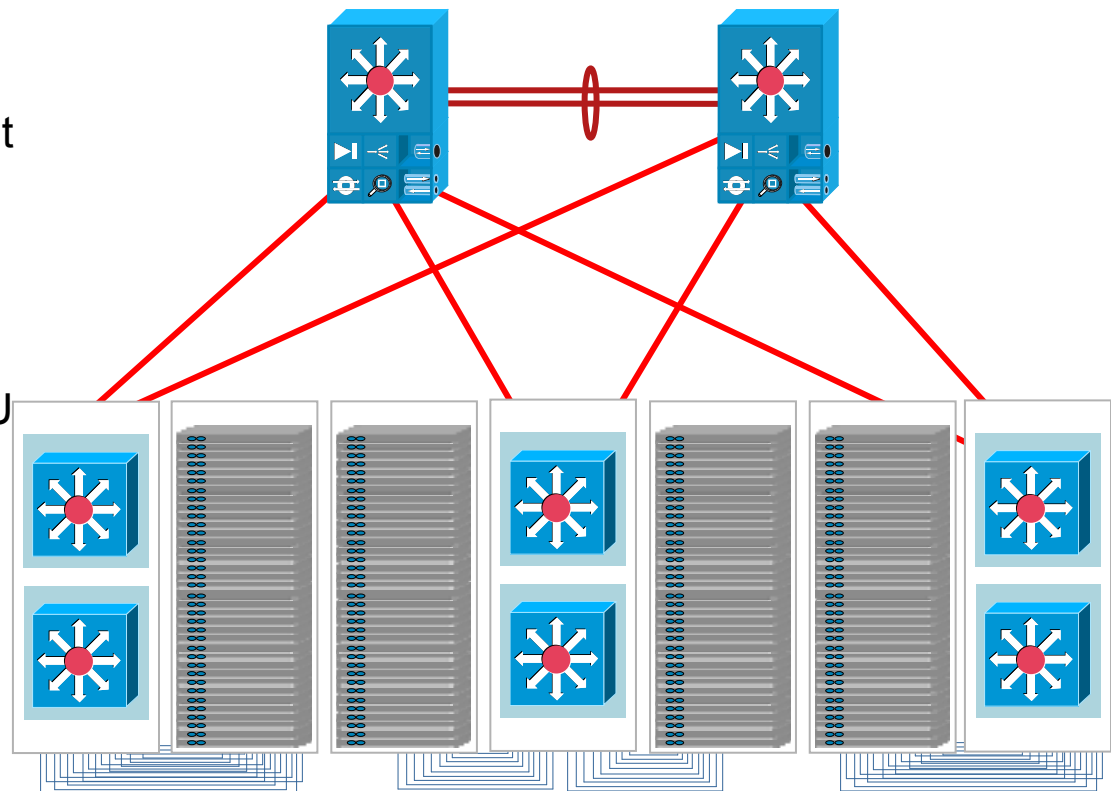


Density and Scalability Implications

Cabinet Design with Modular Access Switches

Servers Connect Directly to a Modular Switch

- Cable bulk at cabinet floor entry can be difficult to manage and block cool air flow
- Typically spaced out by placement at ends of row or within row
- Minimises cabling to Aggregation
- Reduces number of uplinks/aggregation ports
- Redundant switch power and CPU are options
- GEC or 10GE Uplink Considerations
- NEBS Considerations



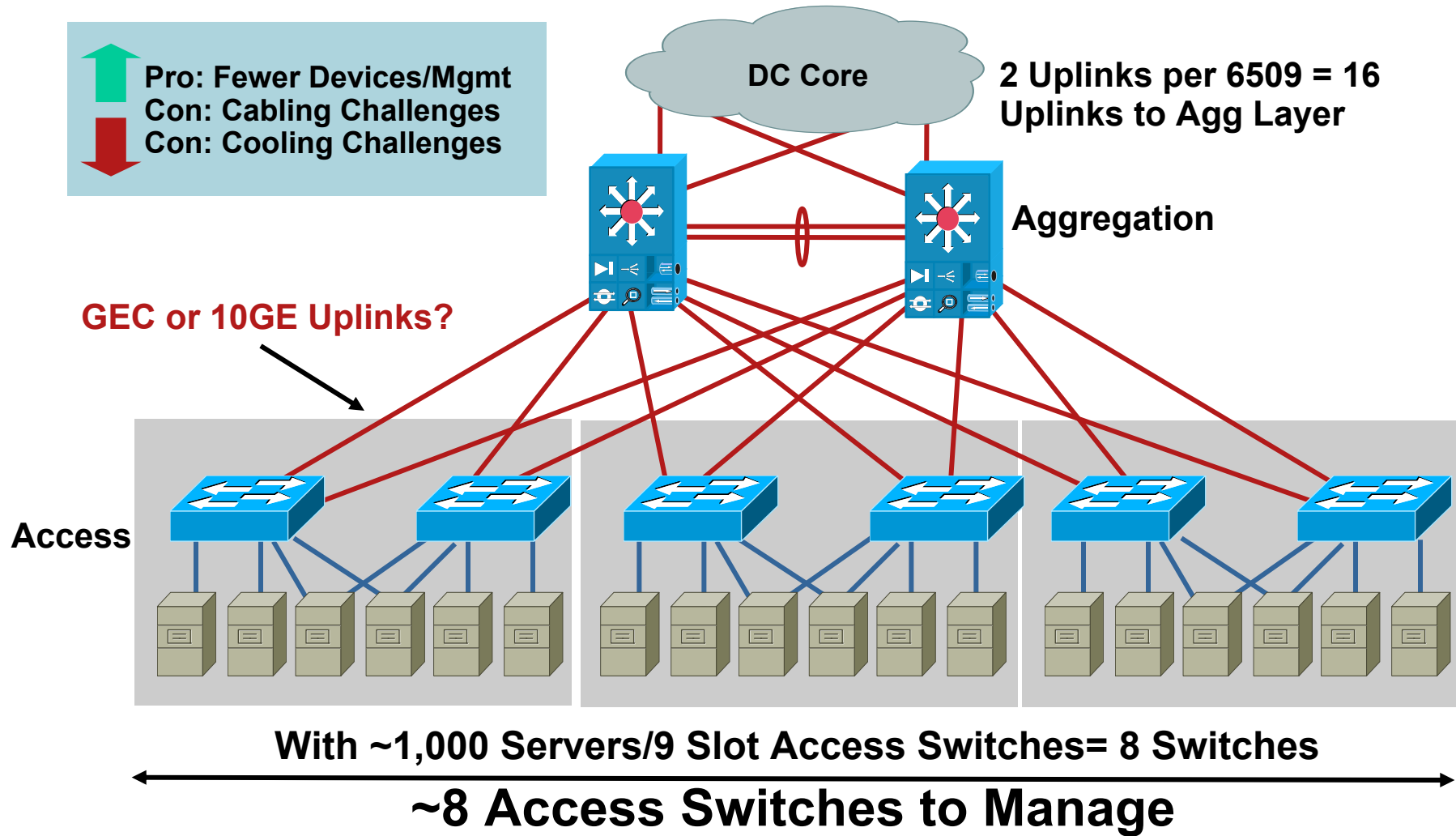
Cables route under raised floor or in overhead trays

Density and Scalability Implications

Network Topology with Modular Switches in the Access

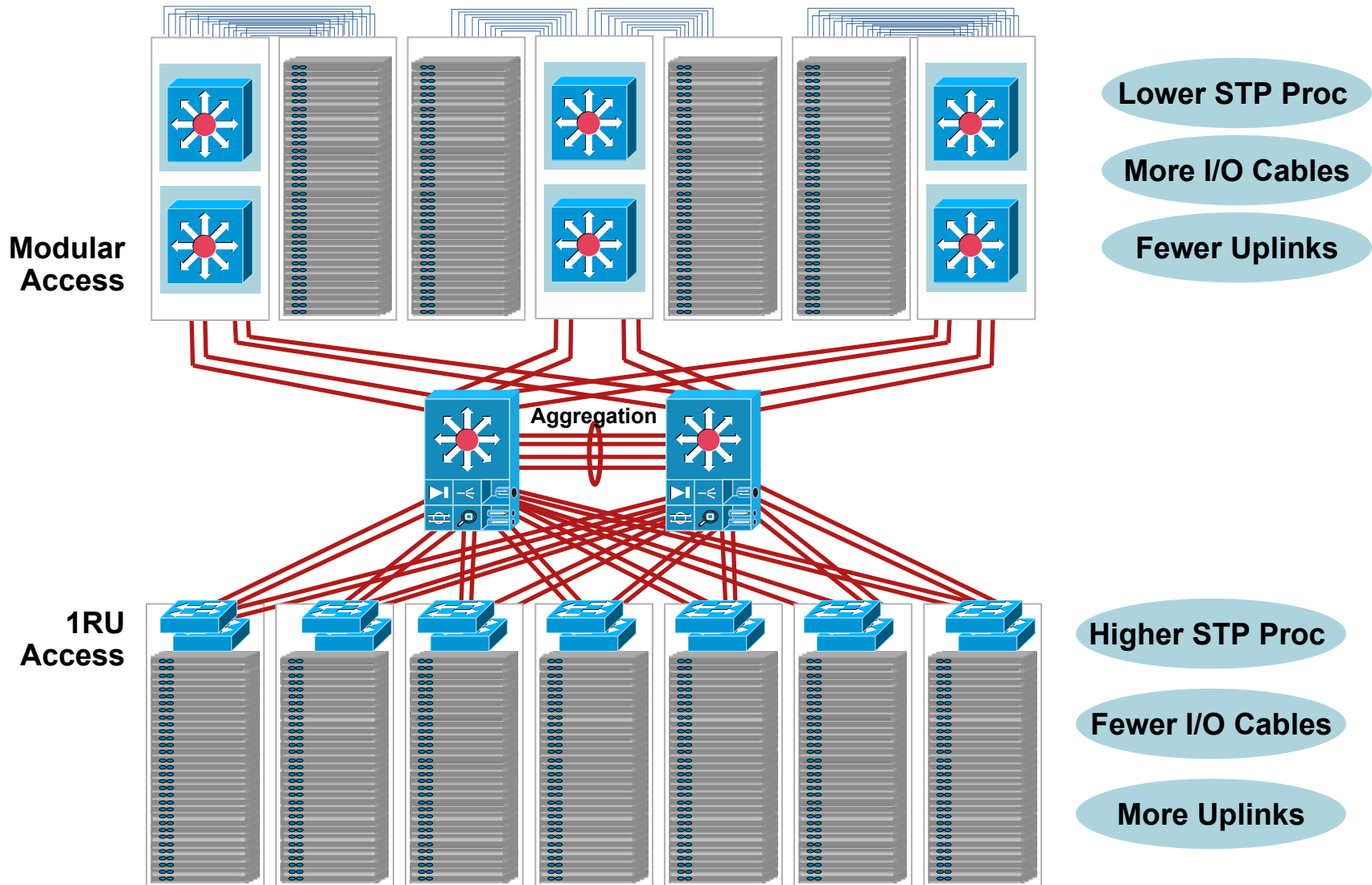


Pro: Fewer Devices/Mgmt
 Con: Cabling Challenges
 Con: Cooling Challenges



Density and Scalability Implications

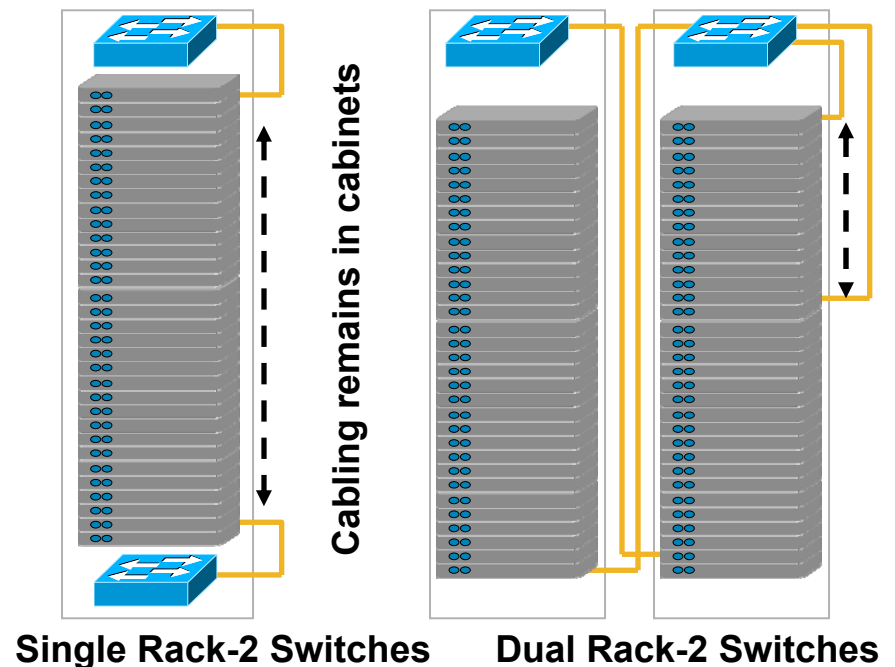
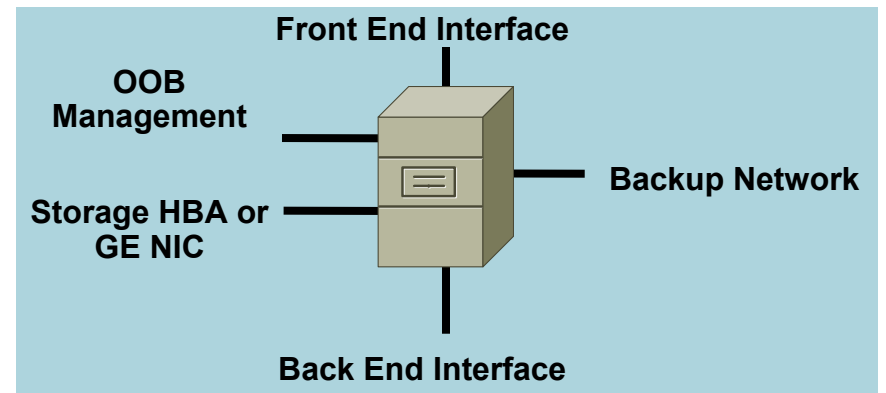
1RU and Modular Comparison



Density and Scalability Implications

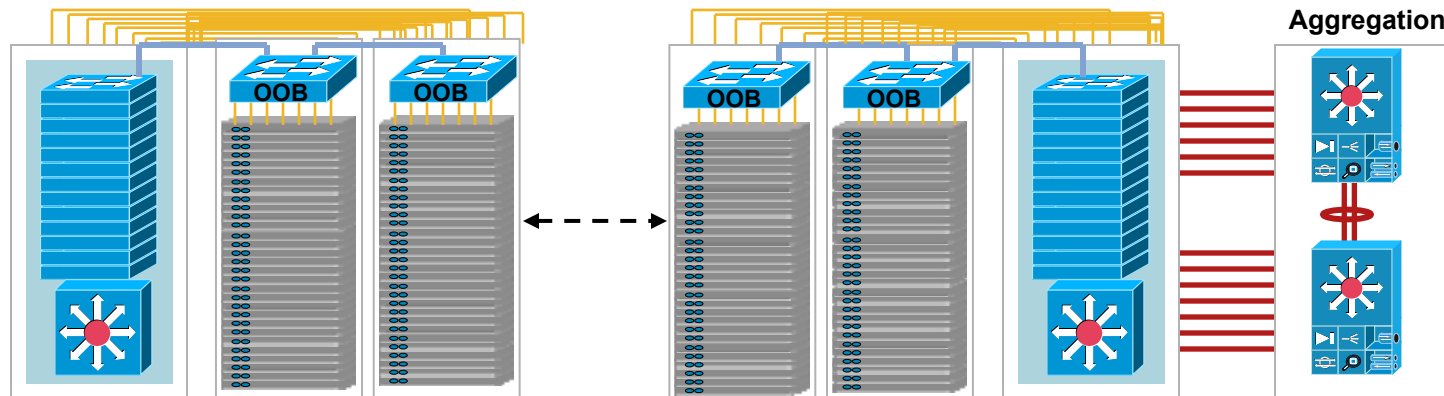
Density: How Many NICs to Plan For?

- Three to four NICs per server are common
 - Front end or public interface
 - Storage interface (GE, FC)
 - Backup interface
 - Back end or private interface
 - integrated Lights Out (iLO) for OOB mgmt
- May require more than two 1RU switches per rack
 - 30 servers@ 4 ports = 120 ports required in a single cabinet (3x48 port 1RU switches)
 - May need hard limits on cabling capacity
 - Avoid cross cabinet and other cabling nightmares



Density and Scalability Implications

Hybrid Example with Separate OOB

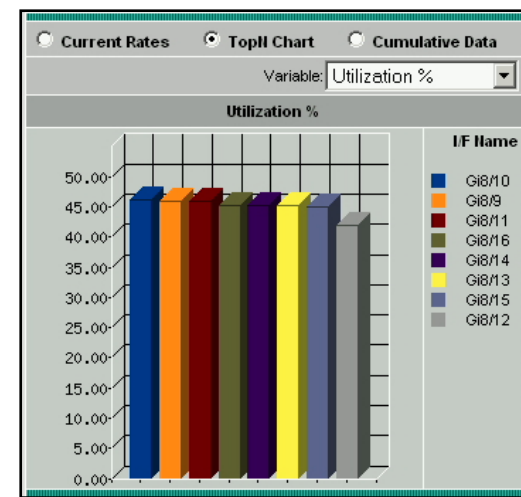
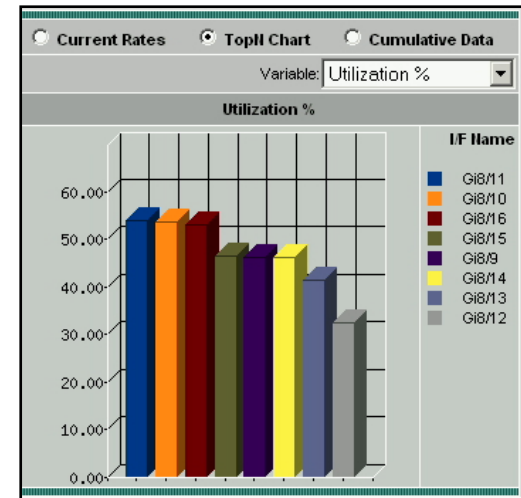


- 1RU End of Row Approach:
 - Smaller failure domains
 - Low power consumption (kw/hr)
 - GEC or 10GE uplinks
 - Permits server distribution via patch
 - Eliminates TOR port density issue
 - Requires solid patch & cabling system
- Separate OOB Switch:
 - Usually 10M ethernet (iLo)
 - Very low utilisation
 - Doesn't require high performance
 - Separate low end switch ok
 - Isolated from production network
- Hybrid Modular + 1RU
 - Provides flexibility in design
 - Dual CPU + Power for critical apps
 - Secondary port for NIC Teaming

Density and Scalability Implications

Oversubscription and Uplinks

- What is the oversubscription ratio per uplink?
 - Develop an oversubscription reference model
 - Identify by application, tier, or other means
- Considerations**
 - Future- true server capacity (PCI-X, PCI- Express)
 - Server platform upgrade cycle will increase levels of outbound traffic
 - Uplink choices available
 - Gigabit EtherChannel 10GE
 - 10Gig EtherChannel
 - Flexibility in adjusting oversubscription ratio
 - Can I upgrade to 10GE easily? 10G EtherChannel?
 - Upgrade CPU and switch fabric? (sup1-2-720-?)



Density and Scalability Implications

Spanning Tree

- 1RU switching increase chances of larger spanning tree diameter
- BladeServer switches are logically similar to adding 1RU switches into the access layer
- A higher number of trunks will increase STP logical port counts in aggregation layer
- Determine spanning tree logical and virtual interfaces before extending VLANs or adding trunks
- Use aggregation modules to scale STP and 10GE density



Scaling Bandwidth and Density



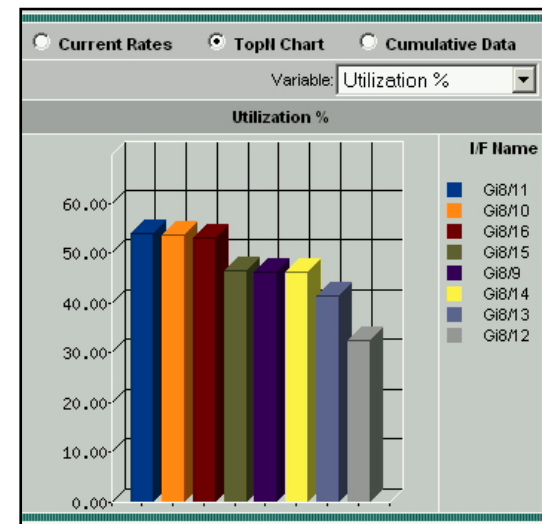
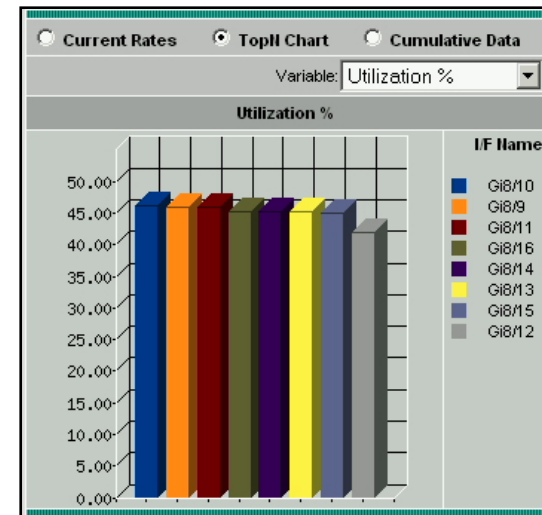
Scaling B/W with GEC and 10GE

Optimising EtherChannel Utilisation

- Ideal is graph on top right
- Bottom left graph more typical
- Analyze the traffic flows in and out of the server farm:
 - IP addresses (how many?)
 - L4 port numbers (randomised?)
- Default L3 hash may not be optimal for GEC: L4 hash may improve

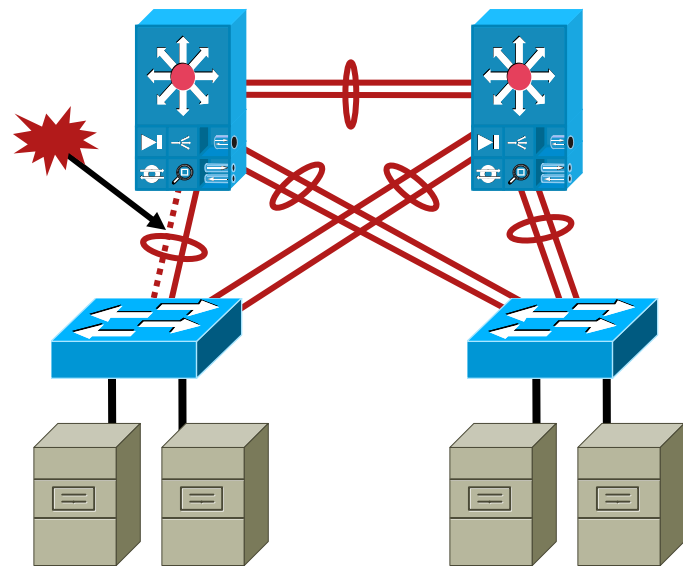
agg(config)# port-channel load balance src-dst-port

- 10 GigE gives you effectively the full bandwidth without hash implications



Scaling B/W with GEC and 10GE Using EtherChannel Min-Links

- Min-Links feature is available as of 12.2.18SXF
- Set the minimum number of member ports that must be in the link-up state or declare the link down
- Permits higher bandwidth alternate paths to be used or lower bandwidth paths to be avoided
- Locally significant configuration
- Only supported on LACP EtherChannels

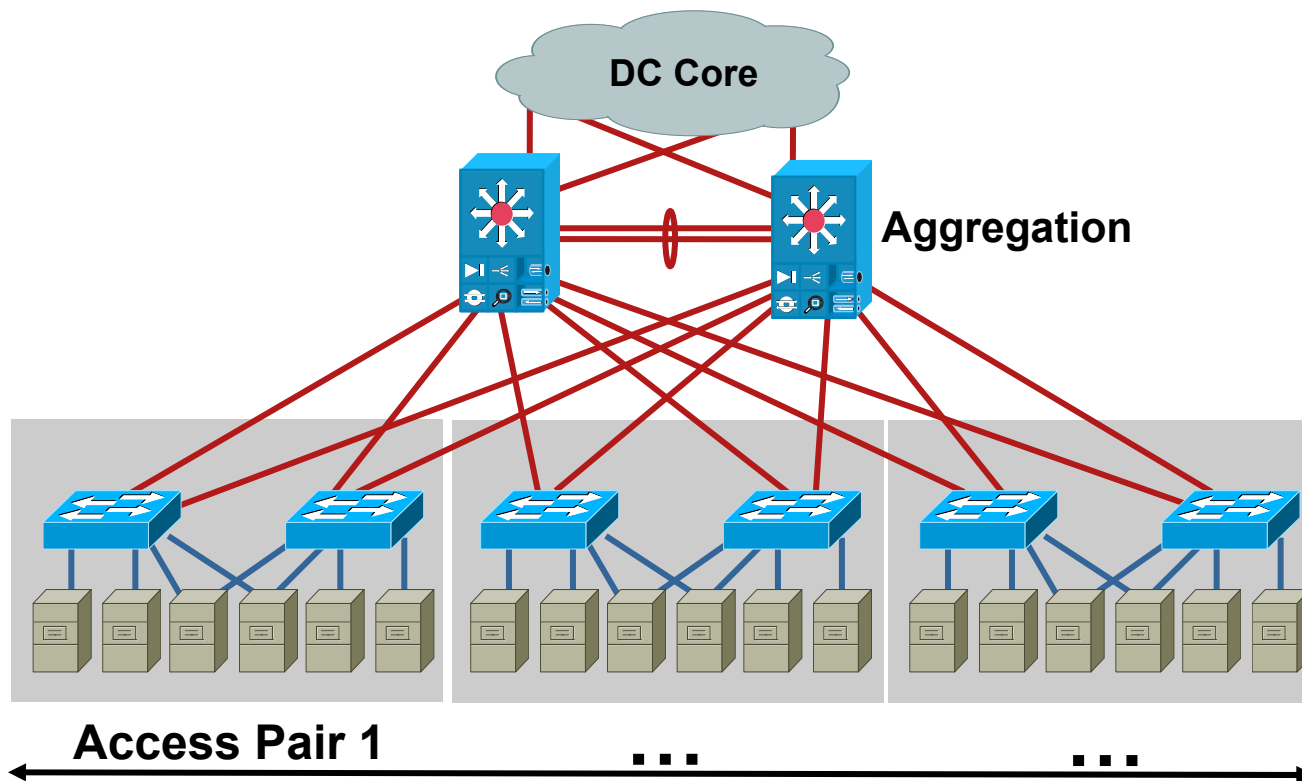


```
Router# configure terminal  
Router(config)# interface port-channel 1  
Router(config-if)# port-channel min-links 2  
Router(config-if)# end
```

Scaling B/W with GEC and 10GE

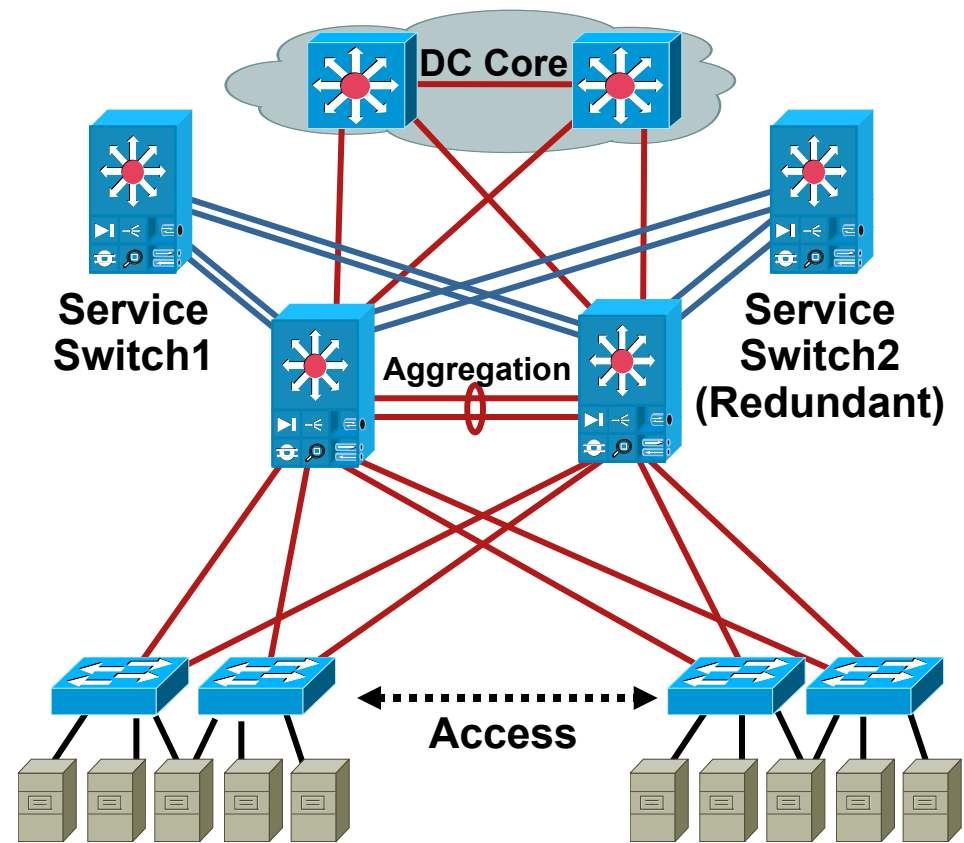
Migrating Access Layer Uplinks to 10GE

- How do I scale as I migrate from GEC to 10GE uplinks?
- How do I increase the 10GE port density at the agg layer?
- Is there a way to regain slots used by service modules?



Scaling B/W with GEC and 10GE Service Layer Switch

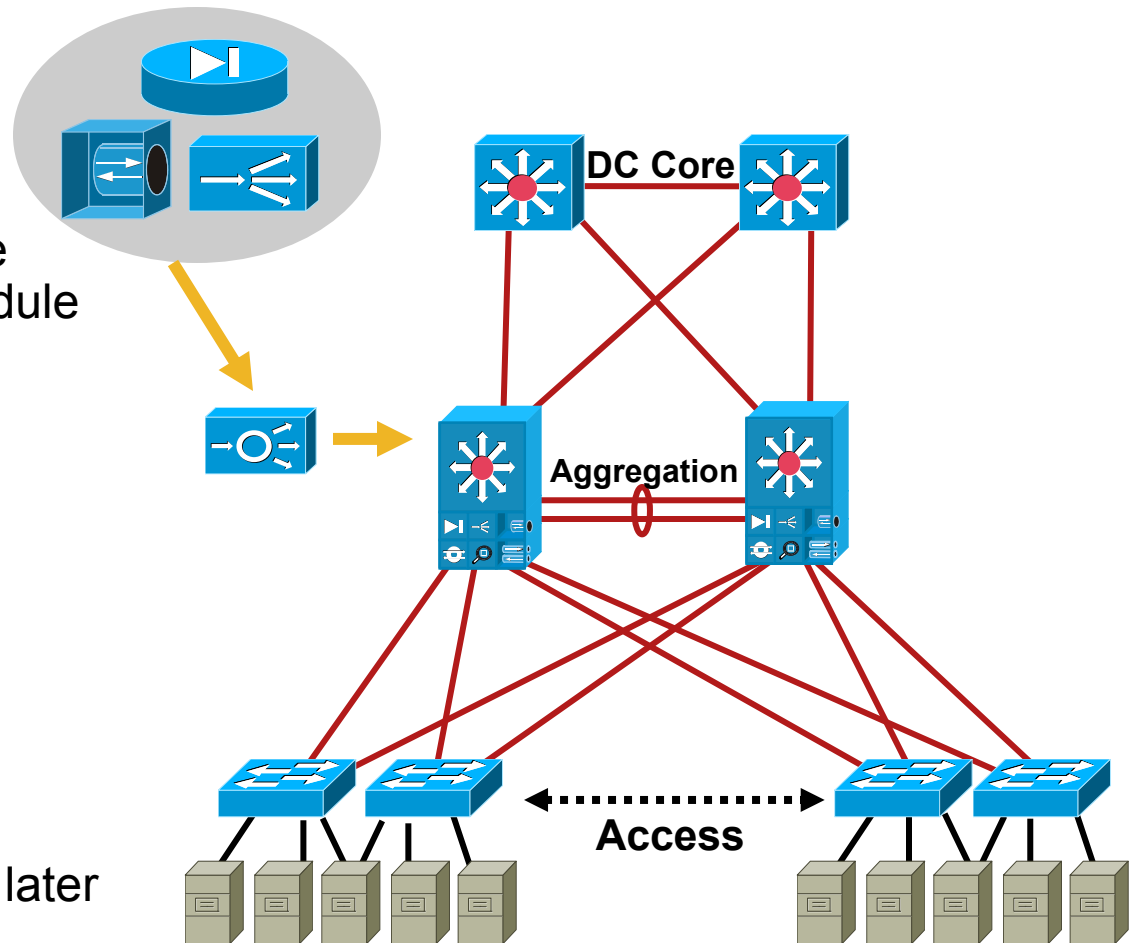
- Move certain services out of aggregation layer
- Ideal for CSM, SSL modules
- Opens slots in agg layer for 10GE ports
- May need QOS or separate links for FT paths
- Extend only necessary L2 VLANs to service switches via .1Q trunks (GEC/TenG)
- RHI installs route in local MSFC only, requiring L3 peering with aggregation



Scaling B/W with GEC and 10GE

Consolidate to ACE

- Consider consolidating multiple service modules onto ACE Module
 - SLB
 - Firewall
 - SSL
- 4/8/16G Fabric Connected
- Active-Active Designs
- Higher CPS + Concurrent CPS
- Single TCP termination, lower latency
- Feature gap may not permit till later release



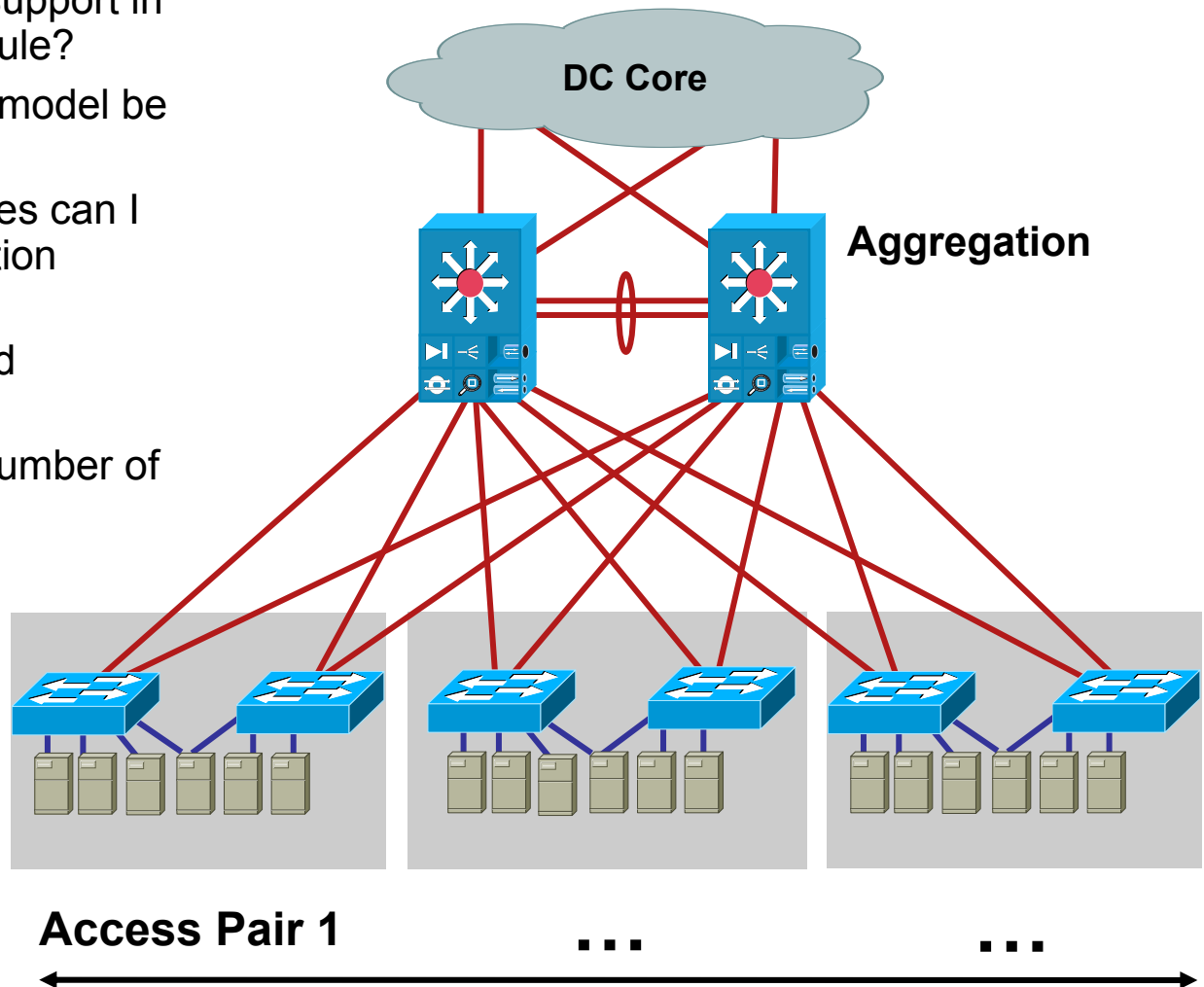
Spanning Tree Scalability



Spanning Tree Scalability

Common Questions

- How many VLANs can I support in a single aggregation module?
- Can a “VLAN Anywhere” model be supported?
- How many access switches can I support in each aggregation module?
- What is the recommended oversubscription rate?
- What are the maximum number of logical ports?
- Are there STP hardware restrictions?



Spanning Tree Scalability

Spanning Tree Protocols Used in the DC

- Rapid PVST+ (802.1w)

 - Most common in data centre today

 - Scales to large size (~10,000 logical ports)

 - Coupled with UDLD, Loopguard, RootGuard and BPDU Guard, provides a strong-stable L2 design solution

 - Easy to implement, proven, scales

- MST (802.1s)

 - Permits very large scale STP implementations (~30,000 logical ports)

 - Not as flexible as Rapid PVST+

 - Service module implications (FWSM transparent mode)

 - More common in service providers and ASPs

This Focuses on the Use of Rapid PVST+

Spanning Tree Scalability

Spanning Tree Protocol Scaling

	MST	RPVST+	PVST+
Total Active STP Logical Interfaces	50,000 Total 30,000 Total with Release 12.2(17b)SXA	10,000 Total	13,000 Total
Total Virtual Ports per LineCard	6,000 ² per Switching Module	1,800 ² per Switching Module(6700) 1200 for Earlier Modules	1,800 ² per Switching Module

1 CSCed33864 Is Resolved in Release 12.2(17d)SXB and Later Releases

2 10 Mbps, 10/100 Mbps, and 100 Mbps Switching Modules Support a Maximum of 1,200 Logical Interfaces per Module

http://www.cisco.com/univercd/cc/td/doc/product/lan/cat6000/122sx/ol_4164.htm#wp26366

Spanning Tree Scalability

Spanning Tree Protocol Scaling

Number of Total STP Active Logical Interfaces=

- Trunks on the switch * active VLANs on the trunks + number of non-trunking interfaces on the switch

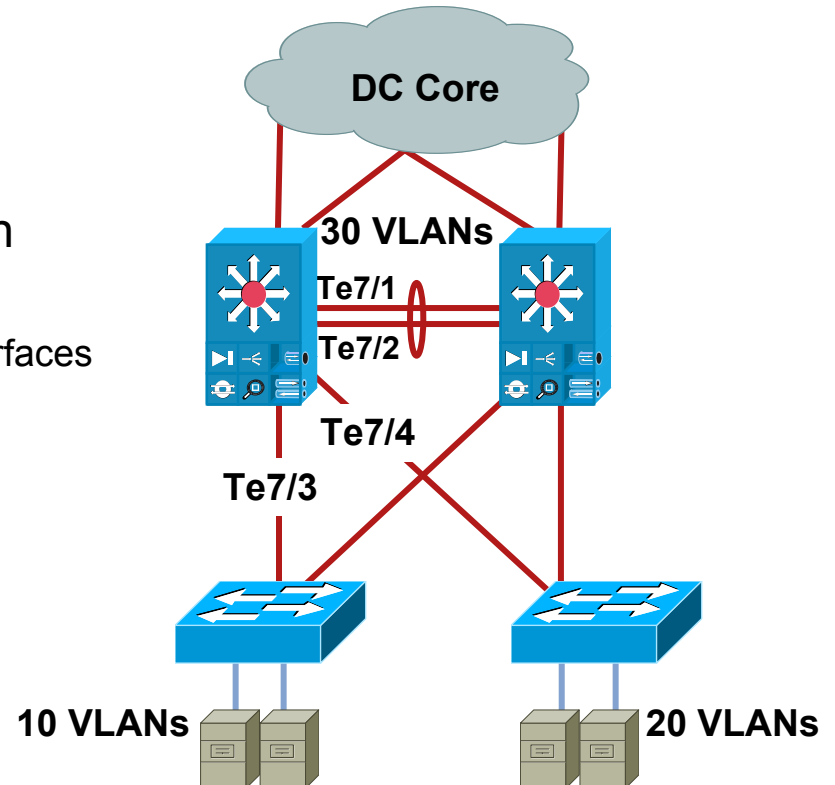
In this example, aggregation 1 will have:

$$10 + 20 + 30 = 60 \text{ STP active logical interfaces}$$

```
AGG1#sh spann summ tot
Switch is in rapid-pvst mode
Root bridge for: VLAN0010, VLAN0020, VLAN0030
EtherChannel misconfig guard is enabled
Extended system ID is enabled
Portfast Default is disabled
PortFast BPDU Guard Default is disabled
Portfast BPDU Filter Default is disabled
Loopguard Default is enabled
UplinkFast is disabled
BackboneFast is disabled
Pathcost method used is long
```

Name	Blocking	Listening	Learning	Forwarding	STP Active
30 VLANs	0	0	0	60	60
AGG1#					

STP Active Column = STP Total Active Logical Interfaces



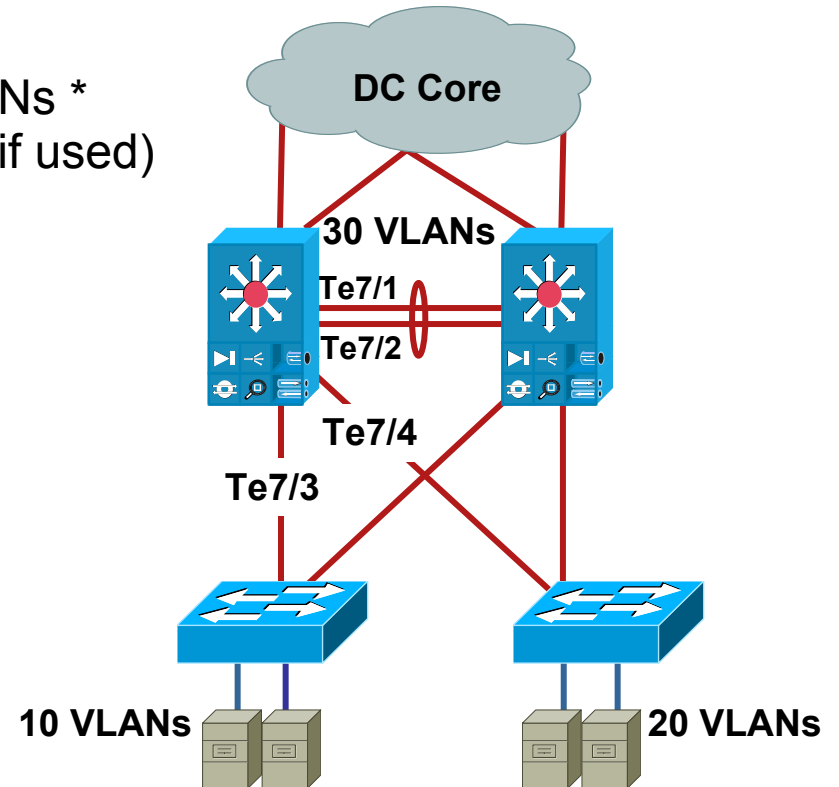
Spanning Tree Scalability

Spanning Tree Protocol Scaling

Number of Virtual Ports per Line Card=

- For line card x: sum of all trunks * VLANs * (the number of ports in a port-channel if used)
- $$10 + 20 + (30*2)$$
- $$=90 \text{ Virtual Port's on line card 7}$$

```
AGG1#sh vlan virtual-port slot 7
Slot 7
Port      Virtual-ports
-----
Te7/1    30   EtherChannel
Te7/2    30
Te7/3    10
Te7/4    20
Total virtual ports:90
AGG1#
```



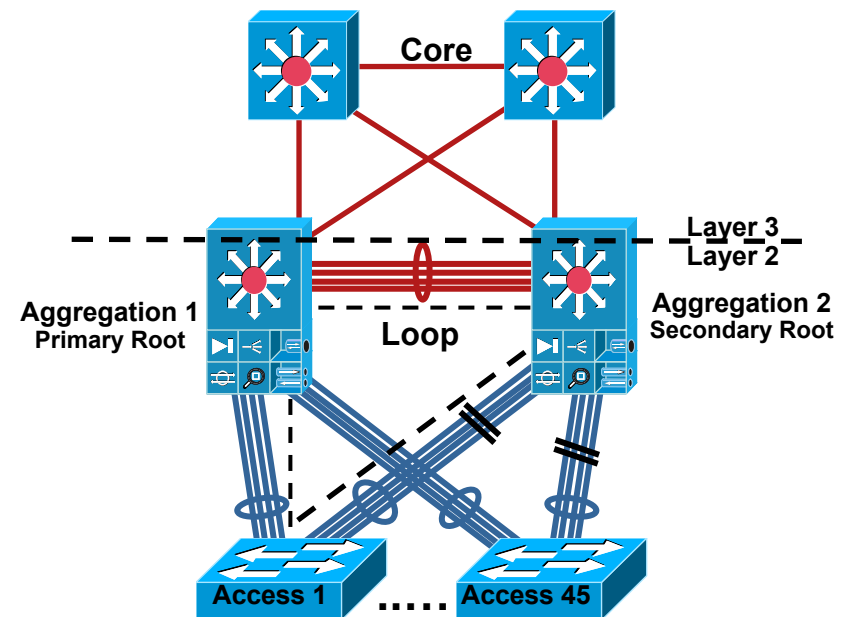
NOTE: VPs Are Calculated per Port in Channel Groups

Spanning Tree Scalability

Spanning Tree Protocol Scaling

Example: Calculating Total Active Logical Ports

- 120 VLANs system wide
- No manual pruning performed on trunks
- 1RU access layer environment
- 45 access switches each connected with 4GEC
- Dual homed, loop topology
 $(120 * 45 \text{ access links}) + 120$
instances on link to
agg2=5400+120=5520
- This is under the maximum recommendation of 10,000 when using Rapid PVST+

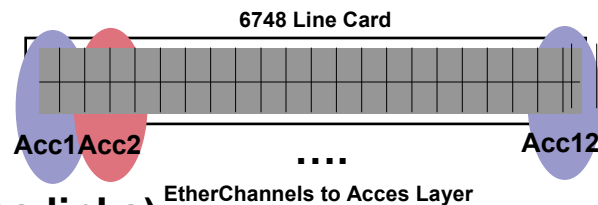


Spanning Tree Scalability

Spanning Tree Protocol Scaling

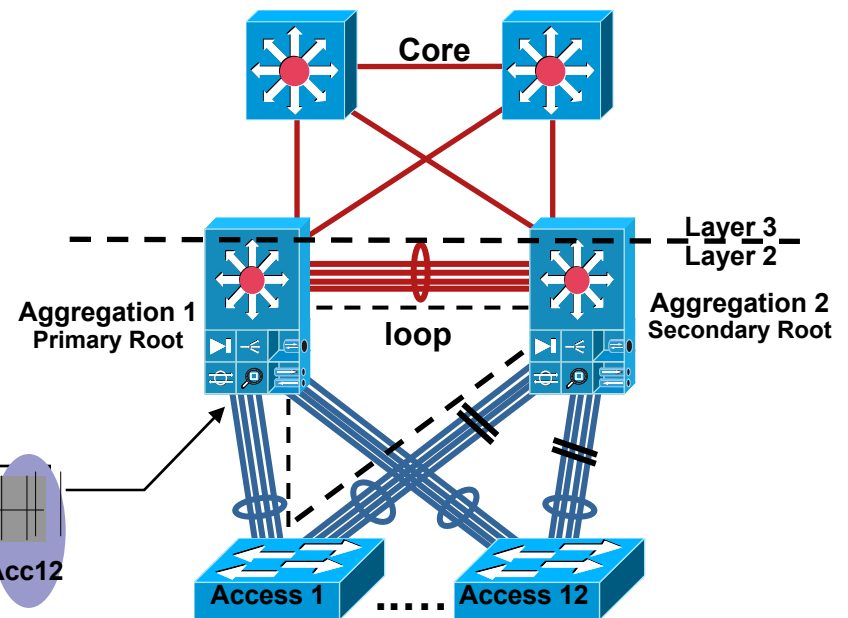
Example: Calculating Virtual Ports per Line Card

- 120 VLANs system wide
- no manual pruning performed on trunks
- 12 access switches, each connected with 4GEC across 6700 line card



(120 * 48 access links)
=5,760 virtual ports

- This is above the recommended watermark



~10 VLANs Used on Each Access Switch

Maximum Number of Supported VLANs in This Design Would Be $1800/48=37.5$

Spanning Tree Scalability

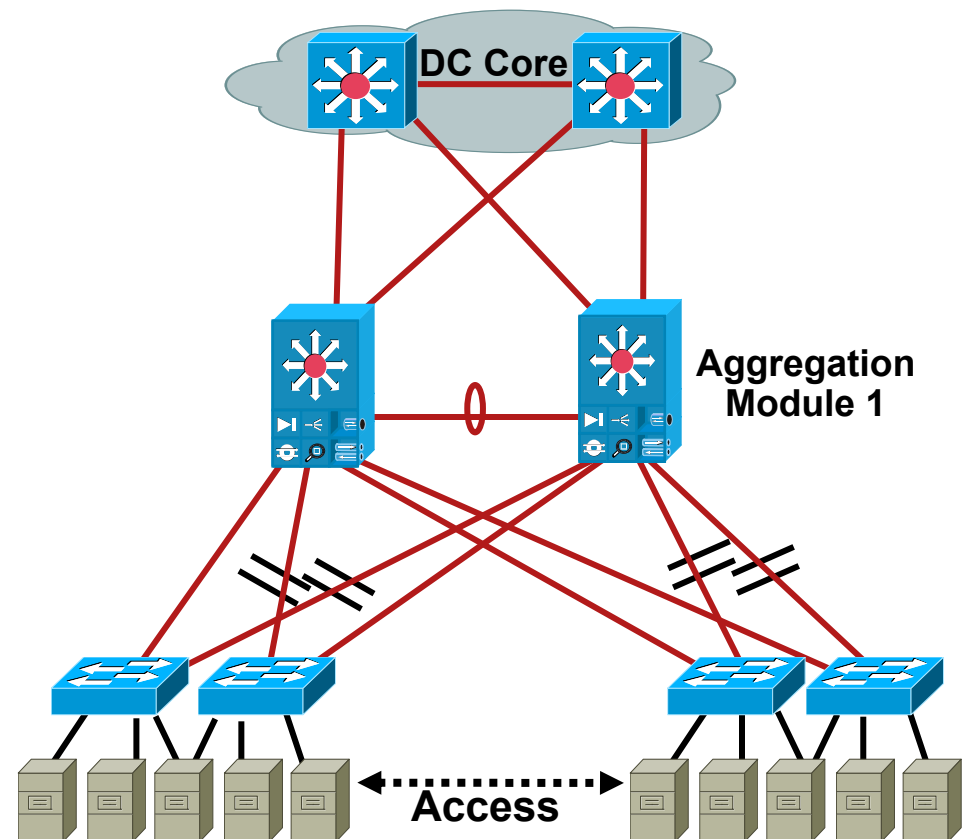
Why STP Watermarks Are Important

Watermarks Are Not Hard Limits But—

- If exceeded, performance is unpredictable
- Larger impact when interface flaps, or shut/no_shut
- Small networks may not see a problem
- Large networks will usually see problems
- Convergence time will be affected

Spanning Tree Scalability Design Guidelines

- Add aggregation modules to scale, dividing up the STP domain
- Maximum five hundred HSRP instances on Sup720 (depends on other cpu driven processes)
- If logical/virtual ports near upper limits perform:
 - Manual pruning on trunks
 - Add agg modules
 - Use MST if meets requirements

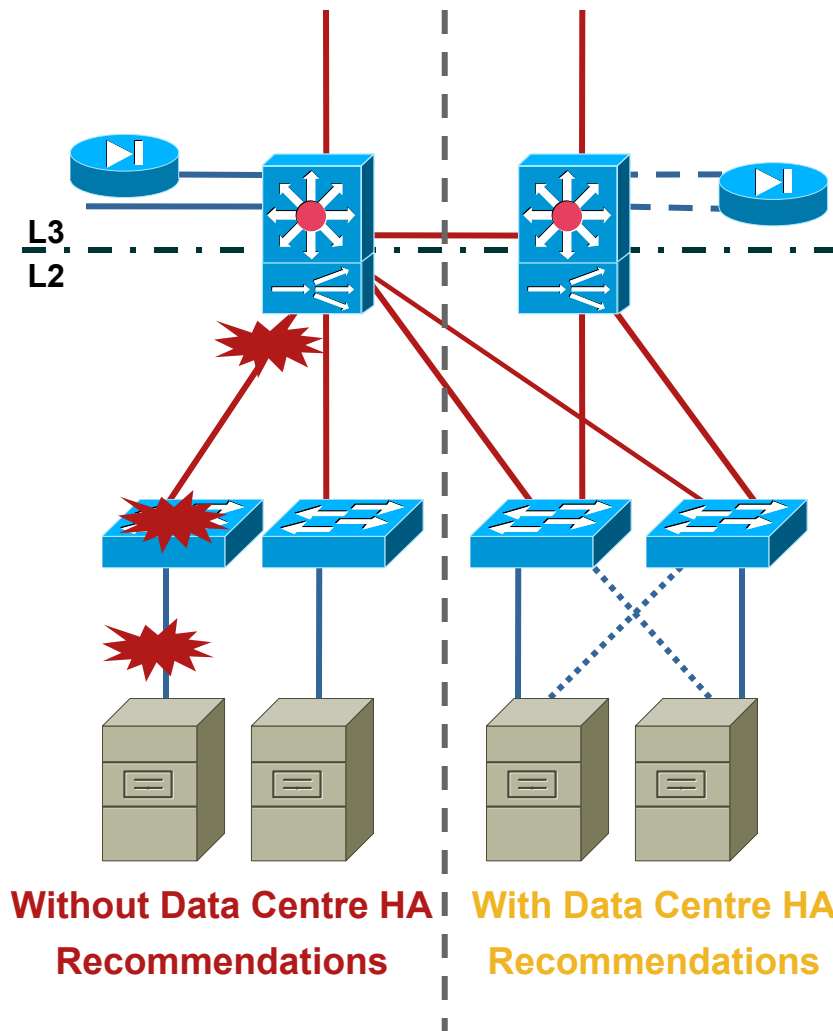


Increasing HA in the Data Centre



Increasing HA in the Data Centre Server High Availability

Common Points of Failure



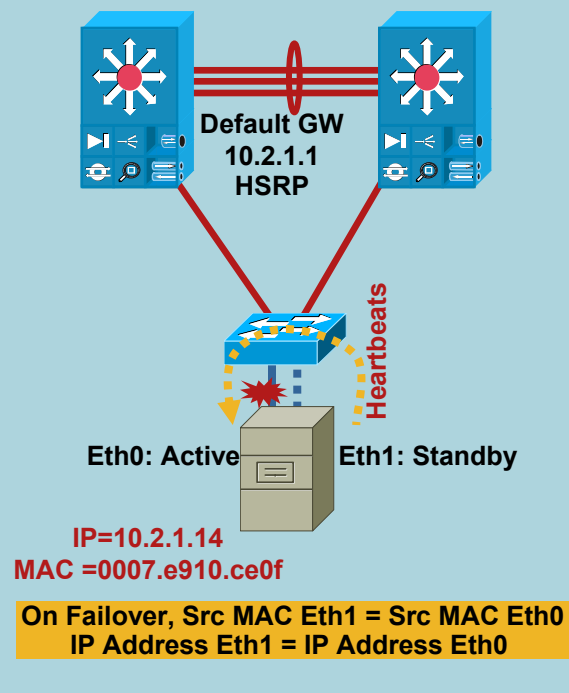
1. Server network adapter
2. Port on a multi-port server adapter
3. Network media (server access)
4. Network media (uplink)
5. Access switch port
6. Access switch module
7. Access switch

These Network Failure Issues Can Be Addressed by Deployment of Dual Attached Servers Using Network Adapter Teaming Software

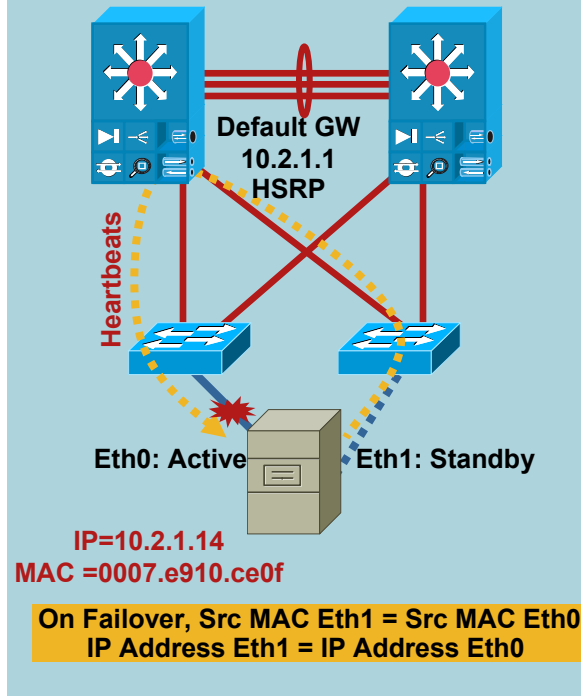
Increasing HA in the Data Centre

Common NIC Teaming Configurations

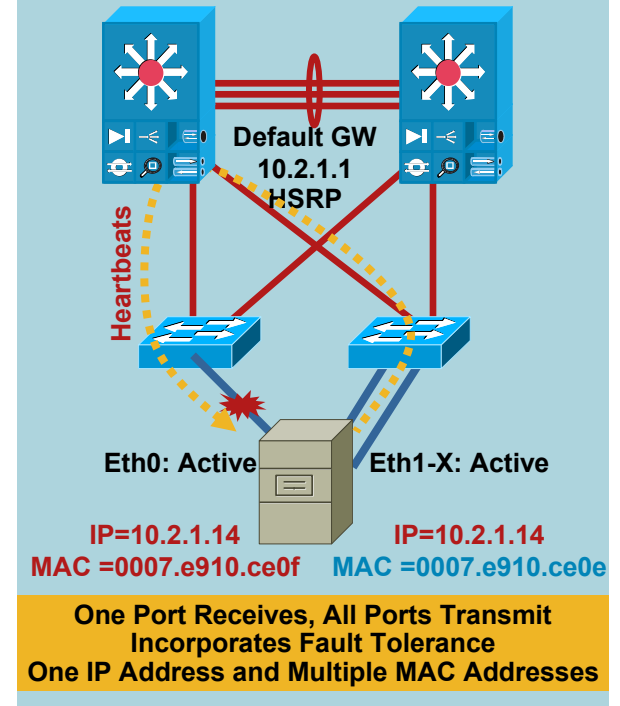
AFT—Adapter Fault Tolerance



SFT—Switch Fault Tolerance



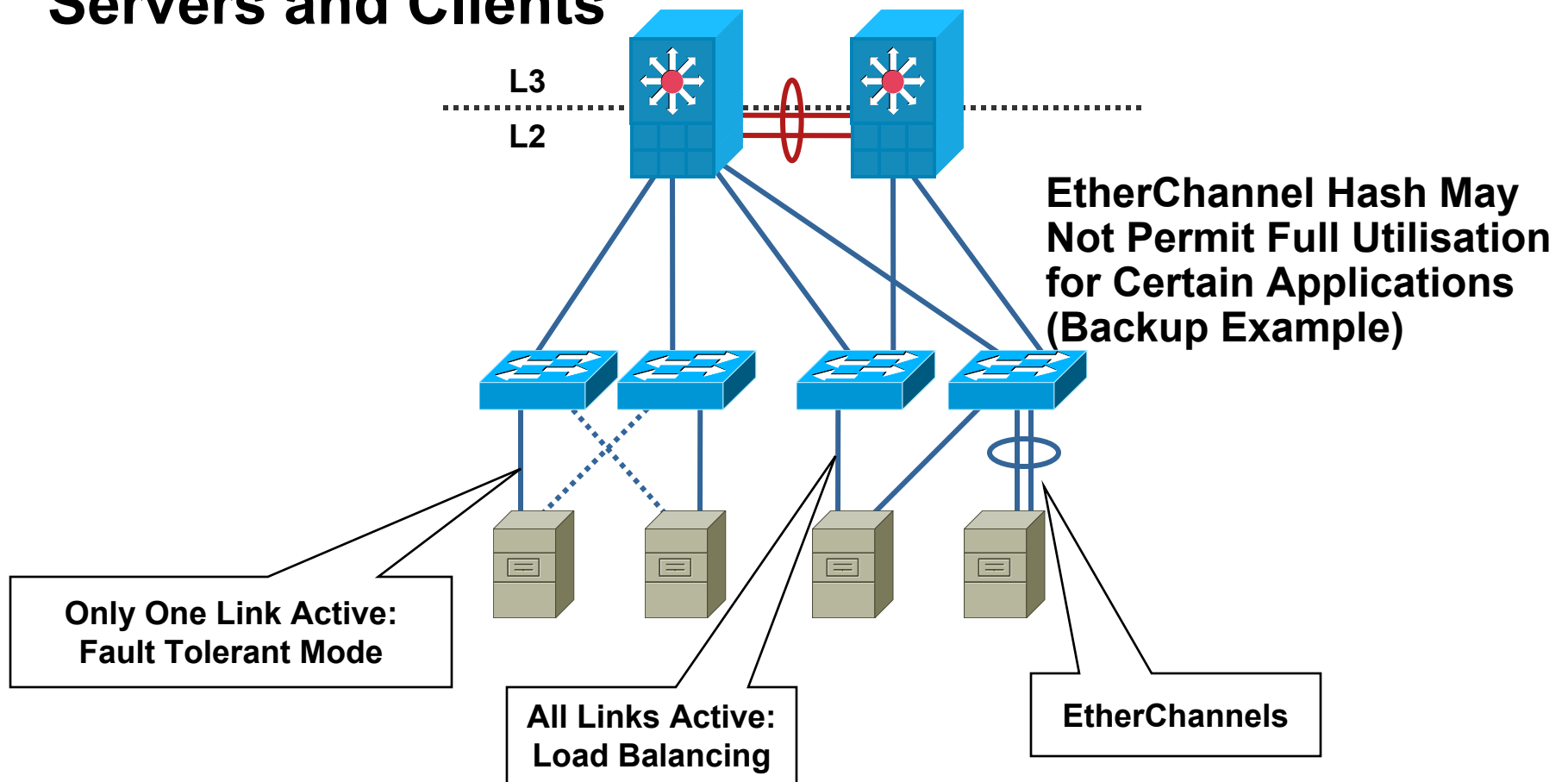
ALB—Adaptive Load Balancing



Note: NIC manufacturer drivers are changing and may operate differently. Also, server OS have started integrating NIC teaming drivers which may operate differently.

Increasing HA in the Data Centre Server Attachment: Multiple NICs

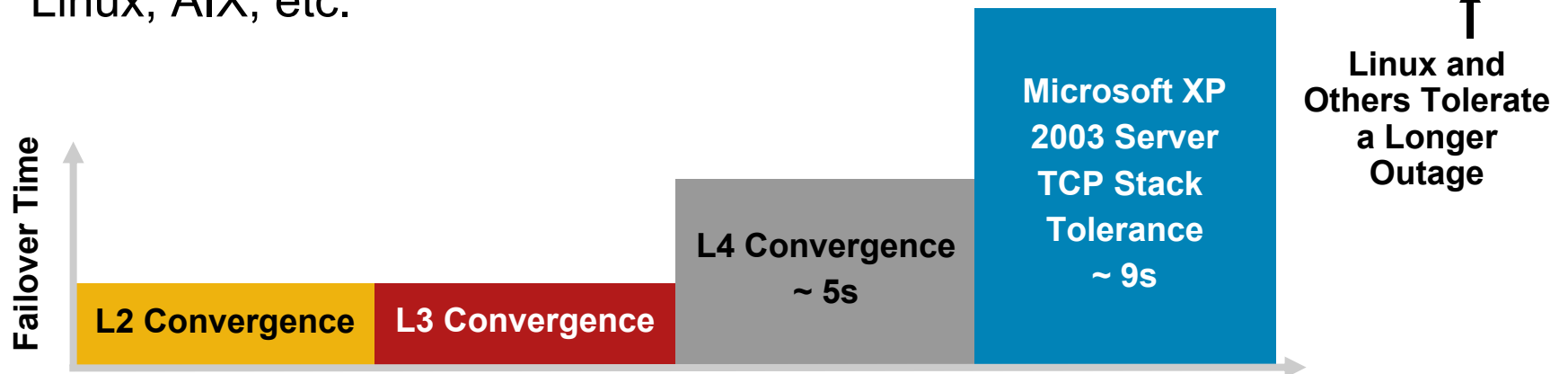
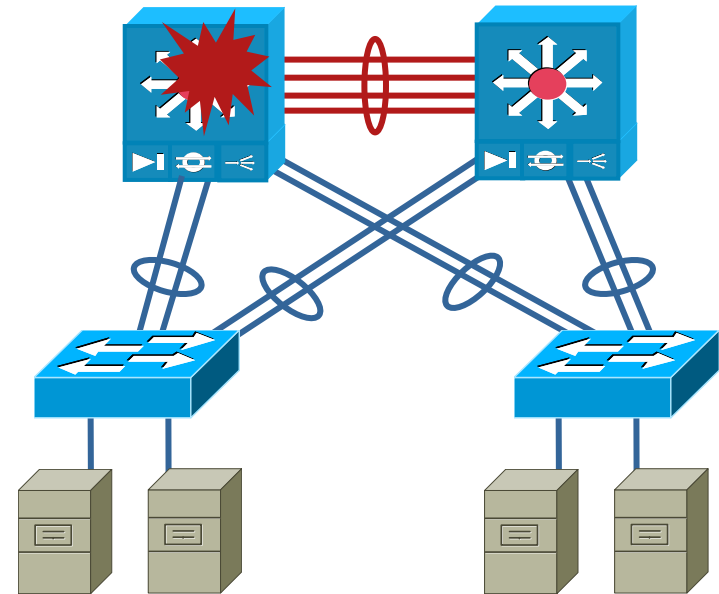
You Can Bundle Multiple Links to Allow
Generating Higher Throughputs Between
Servers and Clients



Increasing HA in the Data Centre

Failover: What Is the Time to Beat?

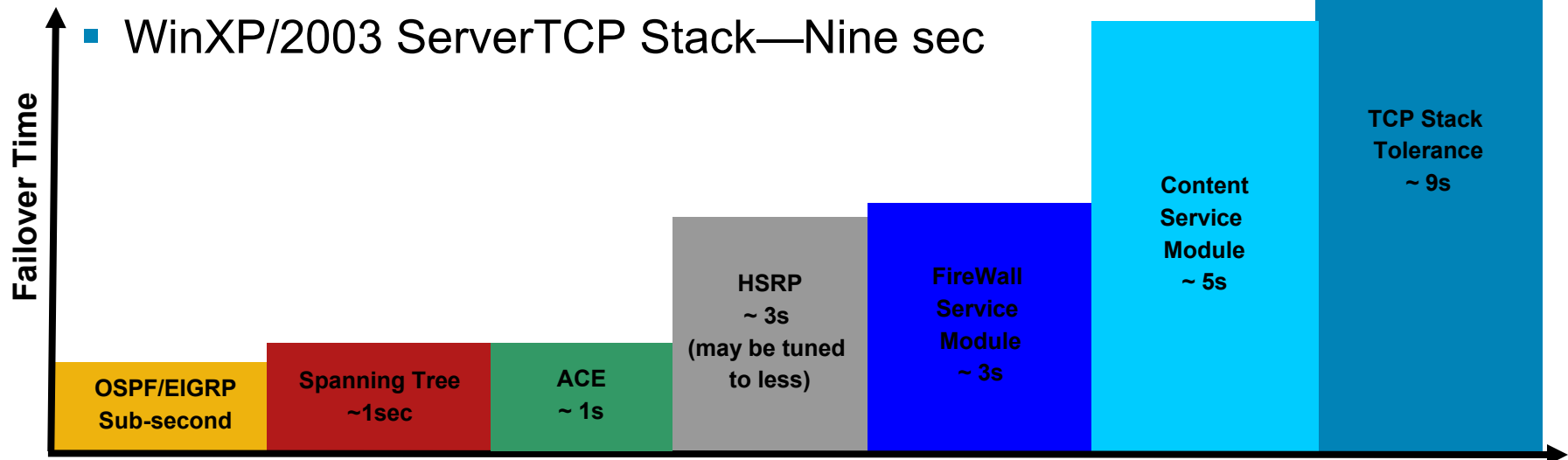
- The overall failover time is the combination of convergence at L2, L3, + L4 components
 - Stateful devices can replicate connection information and typically failover within 3-5sec
 - EtherChannels < 1sec
 - STP converges in ~1 sec (802.1w)
 - HSRP can be tuned to <1s
- Where does TCP break? Microsoft, Linux, AIX, etc.



Increasing HA in the Data Centre

Failover Time Comparison

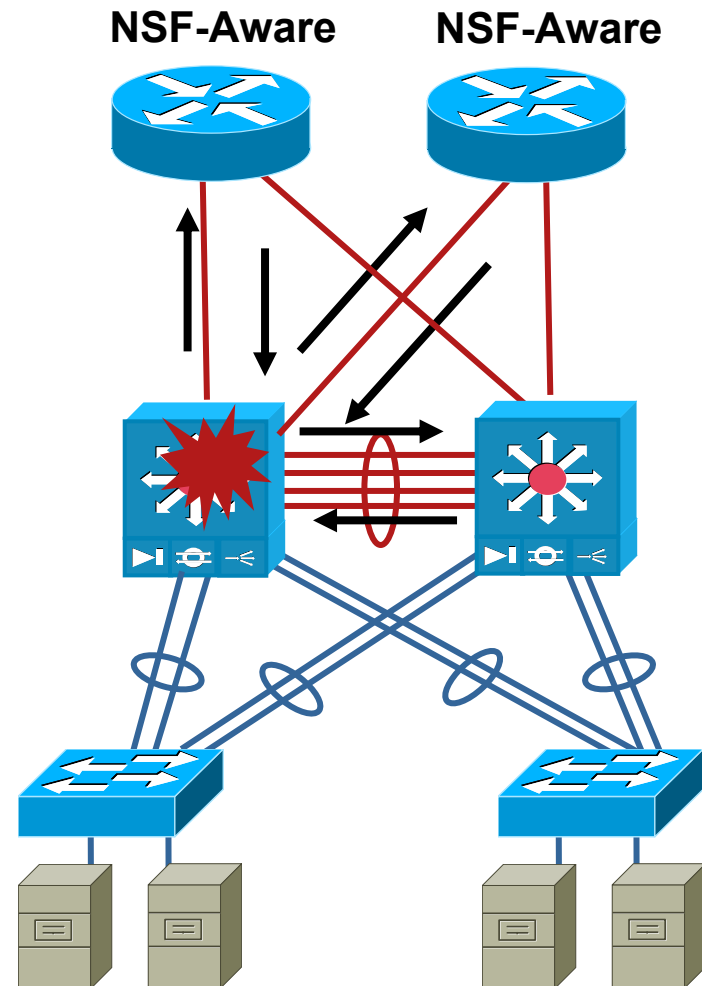
- STP-802.1w—One sec
- OSPF-EIGRP—One sec
- ACE Module with Autostate
- HSRP—Three sec (using 1/3)
- FWSM Module—Three sec
- CSM Module—Five sec
- WinXP/2003 ServerTCP Stack—Nine sec



Increasing HA in the Data Centre

Non-Stop Forwarding/Stateful Switch-Over

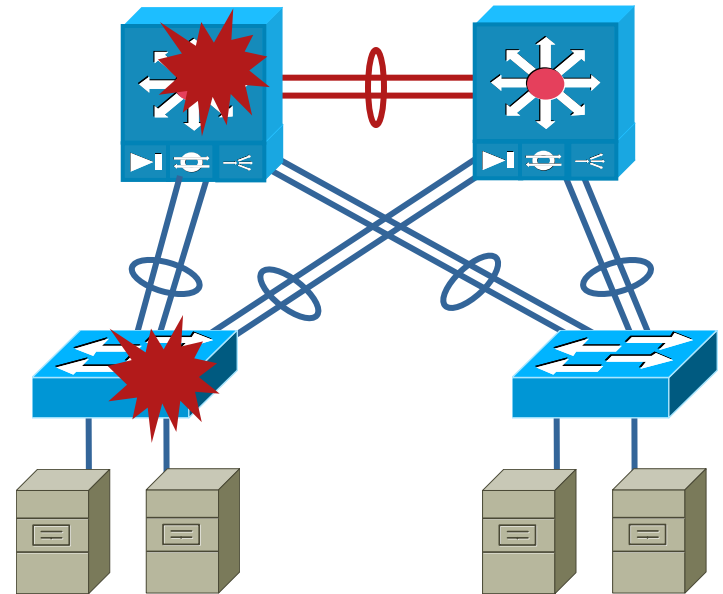
- NSF/SSO is a supervisor redundancy mechanism for intra-chassis supervisor failover
- SSO synchronises layer 2 protocol state, hardware L2/L3 tables (MAC, FIB, adjacency table), ACL and QoS tables
- SSO synchronises state for: trunks, interfaces, EtherChannels, port security, SPAN/RSPAN, STP, UDLD, VTP
- NSF with EIGRP, OSPF, IS-IS, BGP makes it possible to have no route flapping during the recovery
- Aggressive RP timers may not work in NSF/SSO environment



Increasing HA in the Data Centre

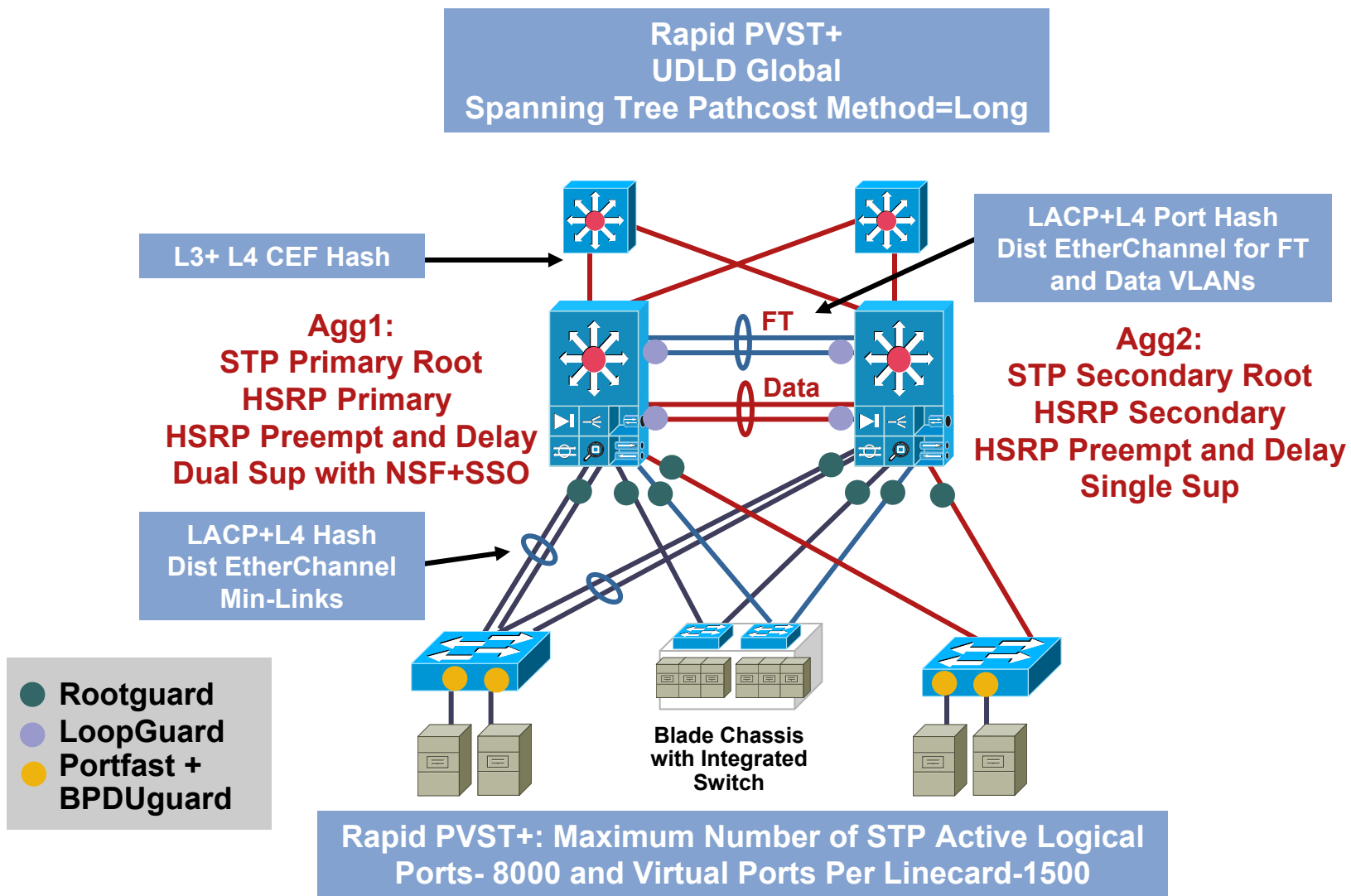
NSF/SSO in the Data Centre

- SSO in the access layer:
 - Improves availability for single attached servers
- SSO in the aggregation layer:
 - Consider in primary agg layer switch
 - Avoids rebuild of arp, igp, stp tables
 - Prevents service module switchover (can be up to ~6sec depending on which module)
 - SSO switchover time less than two sec
 - 12.2.18SXD3 or higher
- Possible implications
 - HSRP state between Agg switches is not tracked and will show switchover until control plane recovers
 - IGP Timers cannot be aggressive (tradeoff)



Increasing HA in the Data Centre

Best Practices: STP, HSRP, Other



A look at the Future...



Traffic Differentiation in Ethernet Networks

- IEEE work in progress
- Possible future consolidated I/O technology

Traffic Types and Requirements

- **Enable Ethernet to carry LAN, SAN and IPC traffic : I/O consolidation**
 - Eliminates multiple backplanes (eg Blade Servers)
 - Should support appropriate characteristics for each traffic type
- **LAN**
 - Large number of flows, not very sensitive to latency
 - E.g. dominant traffic type in Front End Servers
- **SAN**
 - Large packet sizes, sensitive to packet drops
- **IPC**
 - Mix of large & small messages, small messages latency sensitive
 - E.g. HPC Applications

Challenges in Traffic Differentiation

- **Link Sharing (Transmit)**

 - Different traffic types may share same queues/links

 - Large burst from one traffic should not affect other traffic types

- **Resource Sharing**

 - Different traffic types may share same resources (e.g. buffers)

 - Large queued traffic for one traffic type should not starve other traffic types out of resources

- **Receive Handling**

 - Different traffic types may need different Receive handling (e.g. interrupt moderation)

 - Optimisation for CPU utilisation for one traffic type should not create large latency for small message for other traffic types

IEEE Higher Speed Study Group

- **Objectives**

- Support full-duplex operation only**

- Preserve the 802.3/Ethernet frame format at the MAC Client service interface**

- Preserve minimum and maximum FrameSize of current 802.3 Std**

- Support a speed of 100 Gb/s at the MAC/PLS service interface**

- Support at least 10km on SMF.**

- Support at least 100 meters on OM3 MMF**

Meet the Experts

Data Centre

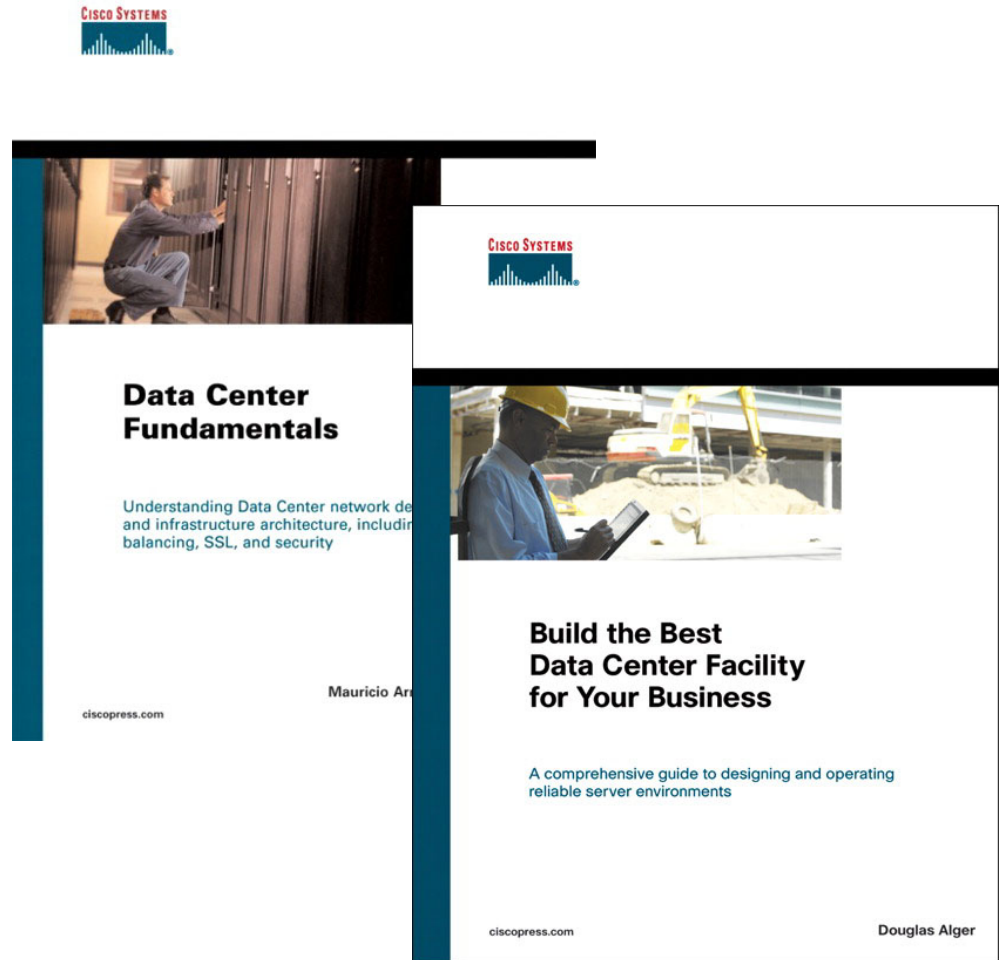
- Victor Moreno
Technical Leader



Recommended Reading

BRKDCT -2001

- Build the Best Data Center Facility for Your Business
- Data Center Fundamentals



Available in the Cisco Company Store

Q and A



Complete Your Online Session Evaluation



