数据中心网络架构和设计指南

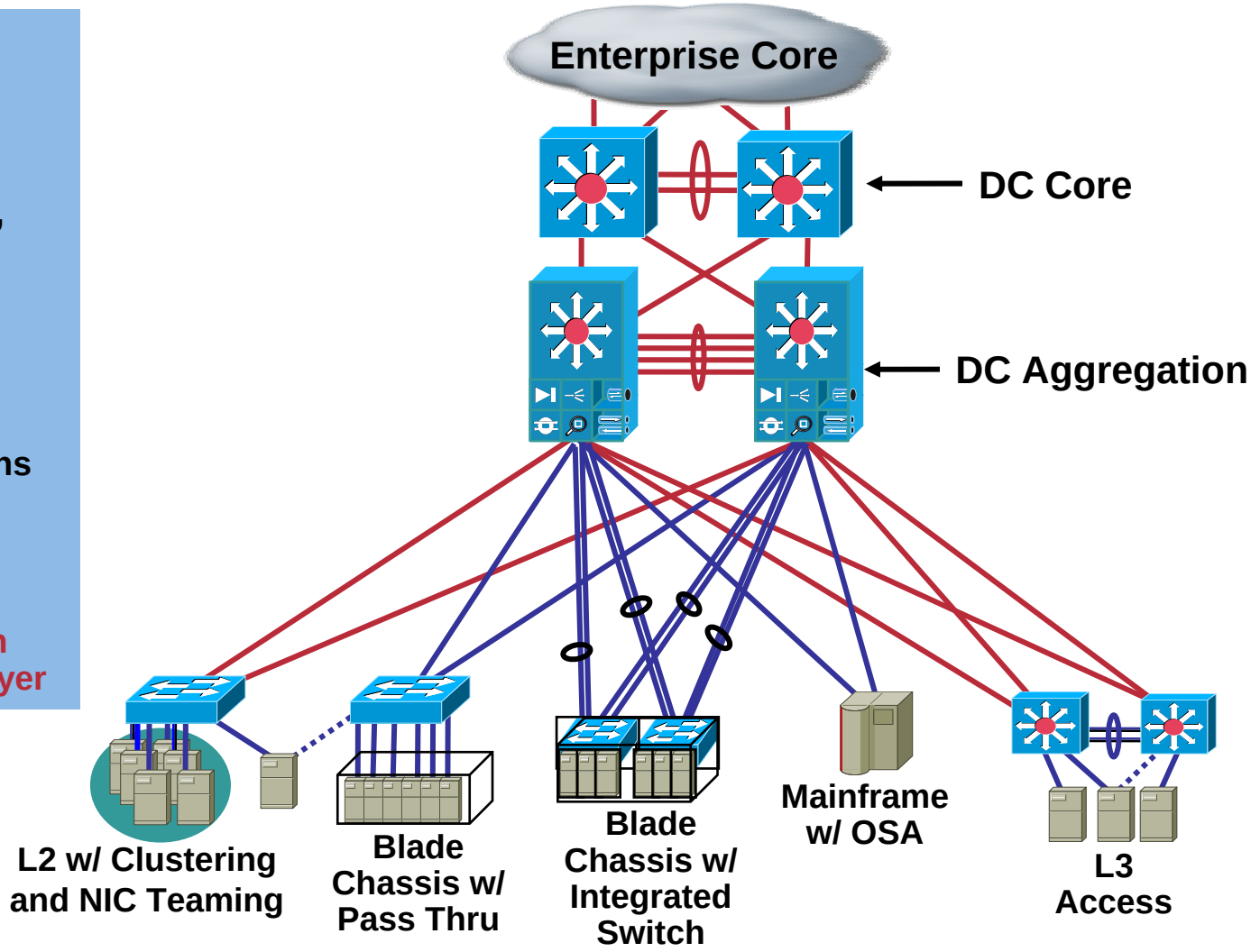
目录
数据中心网络架构

- Server Farm Architecture Overview
- Design Requirements in the Server Farm
- Access Layer Design Models
- Density and Scalability Implications
- Scaling B/W with Gigabit EtherChannel® and 10GE
- Spanning Tree Design and Scalability
- High Availability in the DC
- Summary

数据中心服务器群交换架构

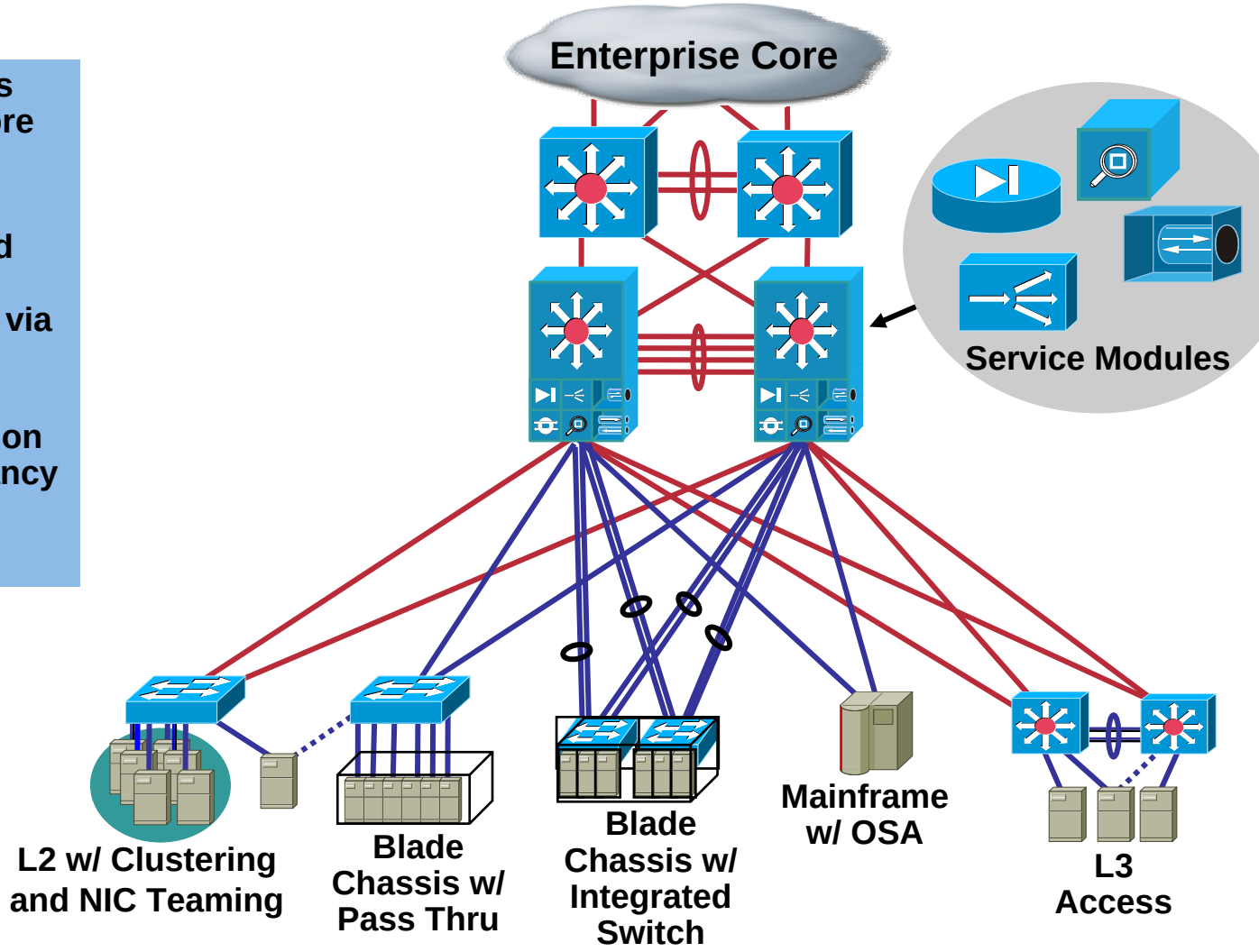
定义数据中心接入层
2层, 3层服务器和主机连接

- L2 and L3 requirements
- Dual and single attached
- High performance, low latency L2 switching
- Mix of over-subscription requirements
- Many uplink options
- STP processing for configured VLANs only
- Utilizes services in the aggregation layer



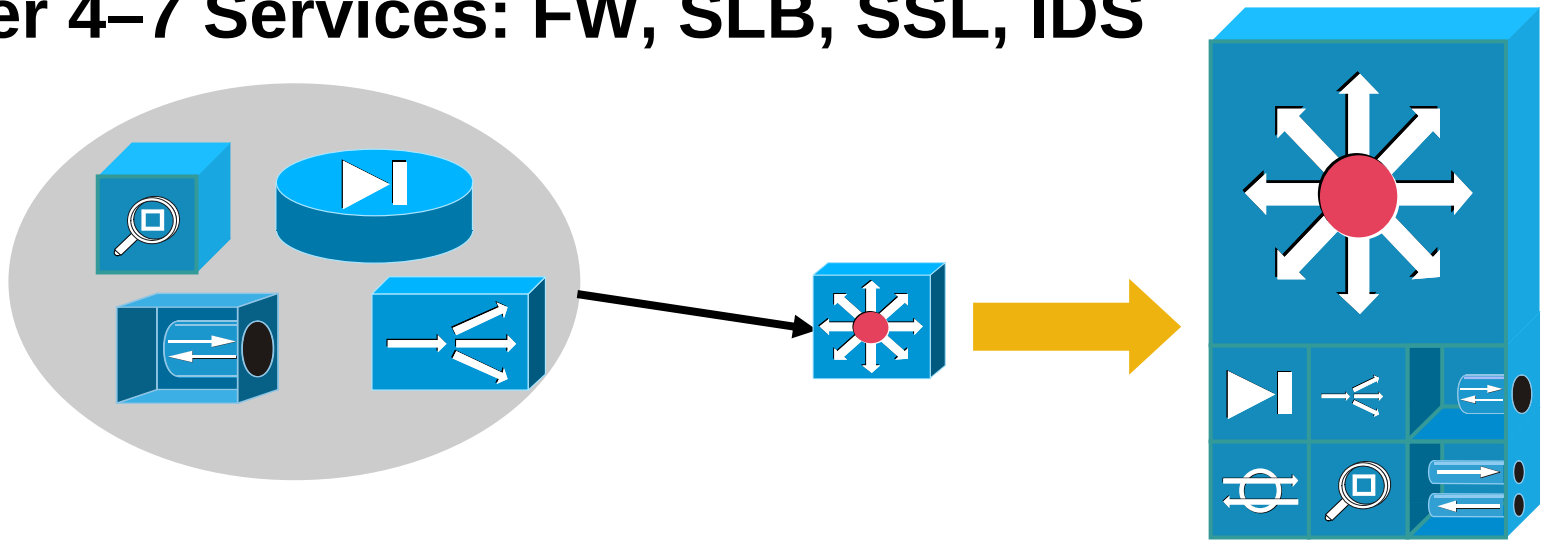
定义数据中心汇聚层
提供共享的应用/安全服务

- Aggregates access uplinks into DC core
- Large STP processing load
- Provides advanced application and security functions via service modules
- Maintains session state and connection tables for redundancy
- What are these services?



定义数据中心汇聚层
集成服务

Layer 4–7 Services: FW, SLB, SSL, IDS

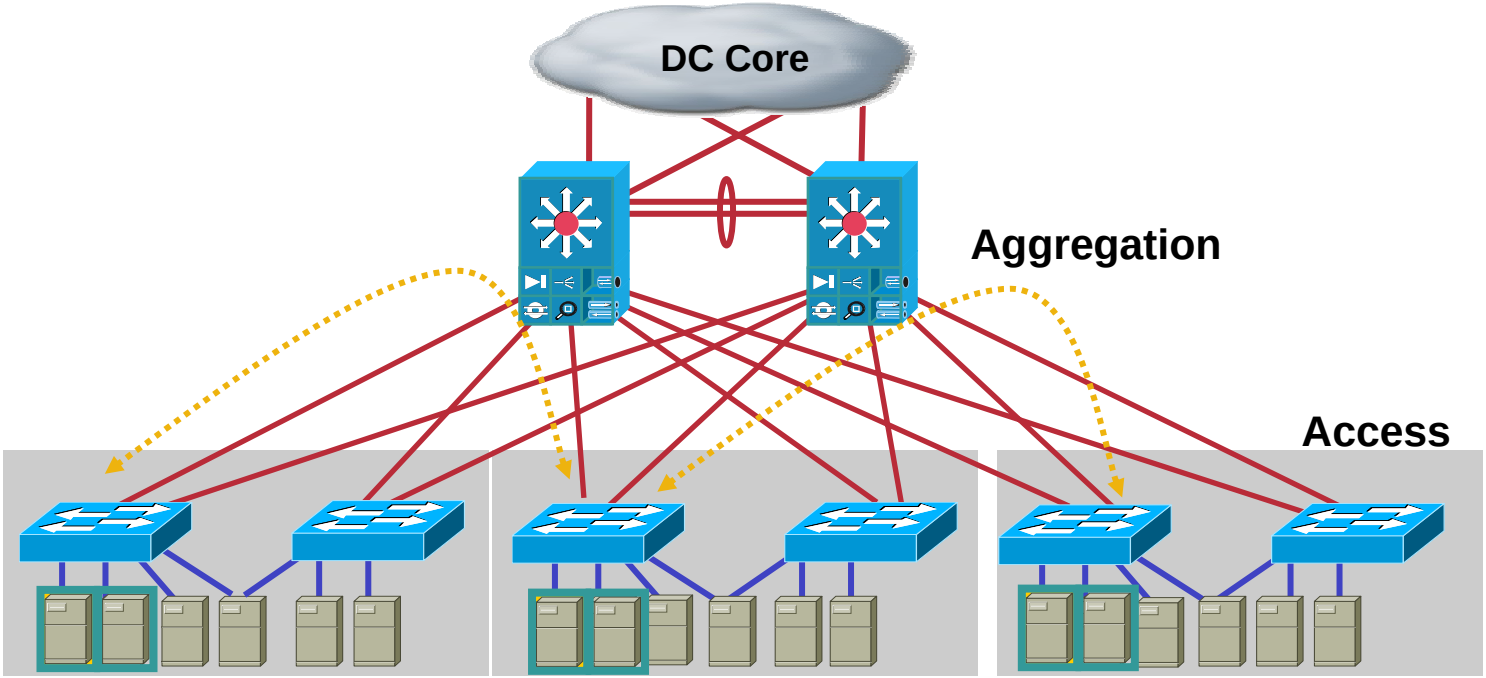


- **Application and security services can be deployed as:**
  - **Appliances**
  - **Service blades**
- **Service blades such as firewall blades and load balancing blades…provide hardware-based stateful functions**
- **Integrated blades optimize rack space, cabling and configuration mgmt**
- **Provide highest flexibility and economies of scale**

定义数据中心汇聚层
服务器之间的通信路径

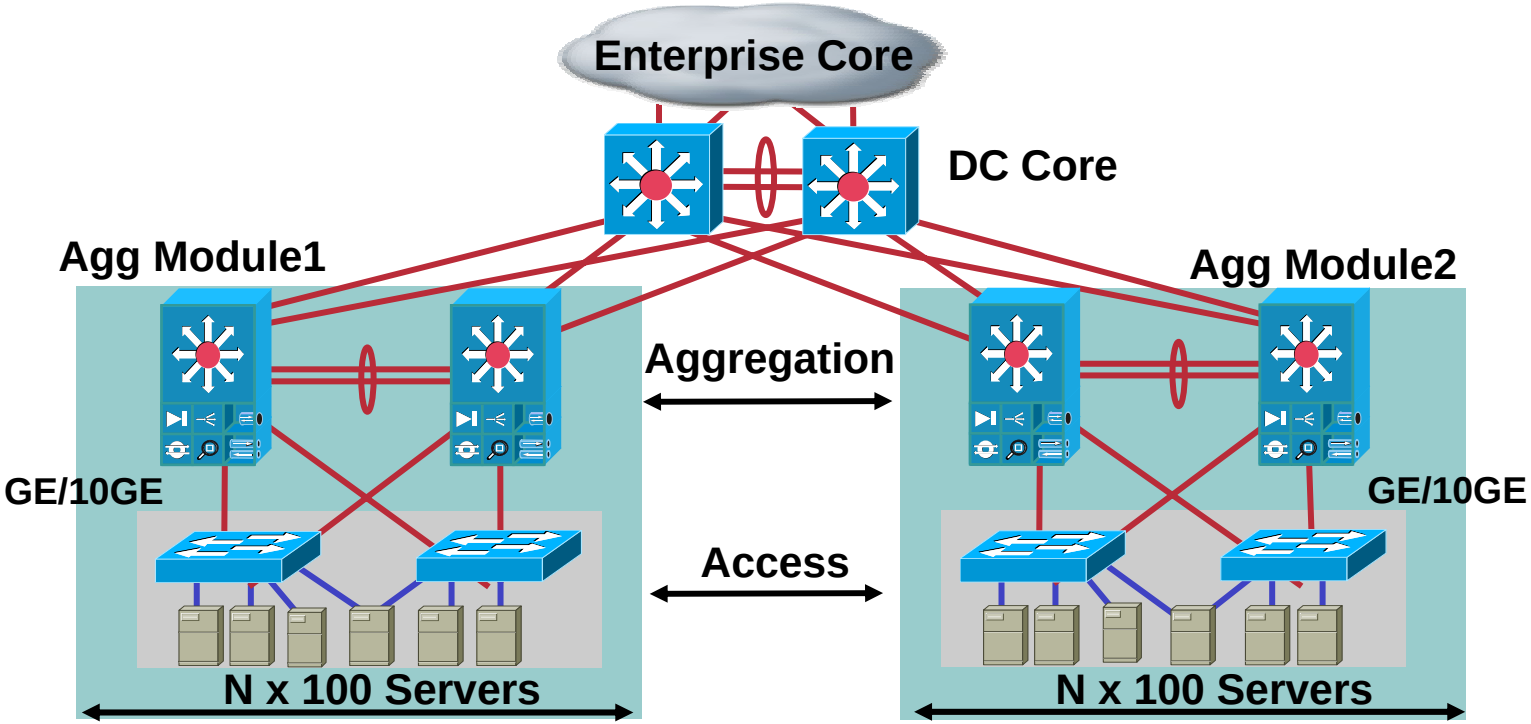
What Types of Server to Server Traffic Will Exist?

Multi-Tier Interaction, Backup, Replication, Cluster Messaging, Storage over IP



- The aggregation module may provide the primary communication path for server to server traffic
- Non traditional traffic emerging
- Driving lower oversubscription and 10GE uplinks
- Servers now ship with PCI-X NIC's and GE
- Plan bandwidth for future server true capacity

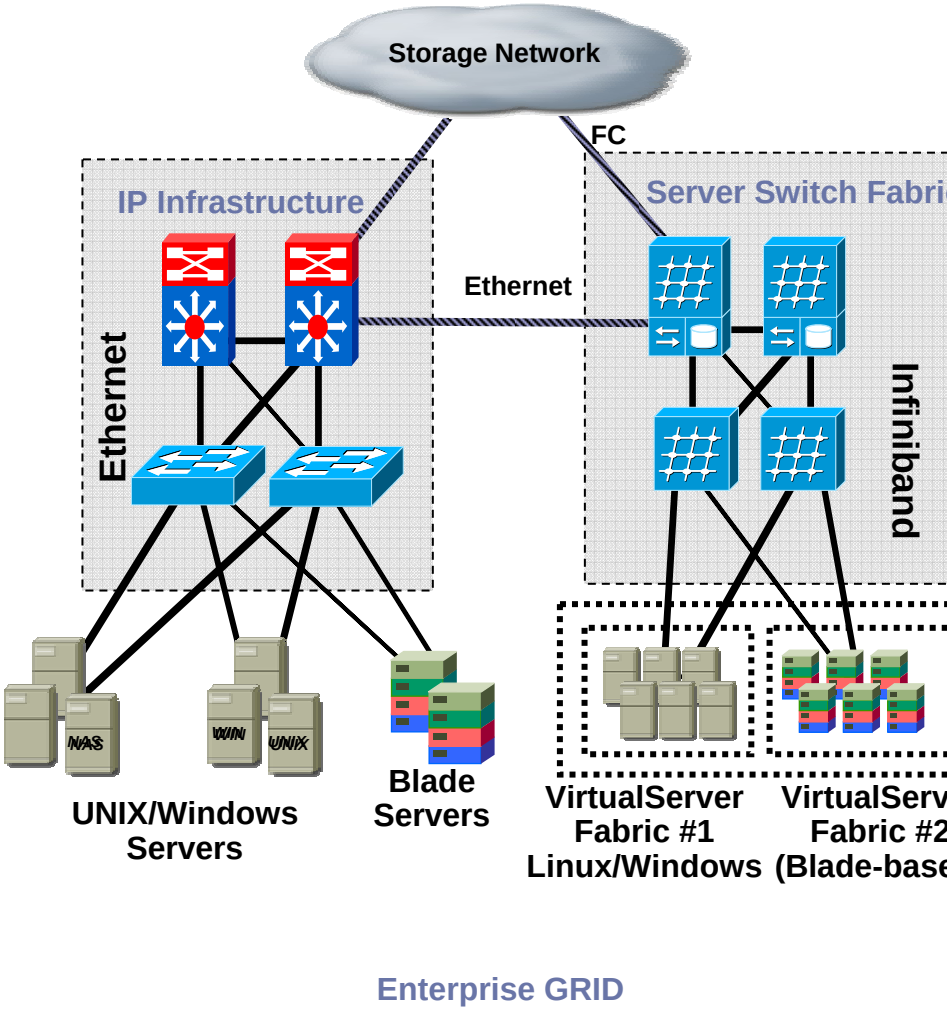
定义数据中心核心层
汇聚层之间的高速交换矩阵



- Interconnects AGG modules
- Isolates failure domains
- Scales large STP diameters
- Improves 10GE scaling
- Plan and build DC core up front

定义数据中心服务器交换矩阵
服务器间流量的高速交换

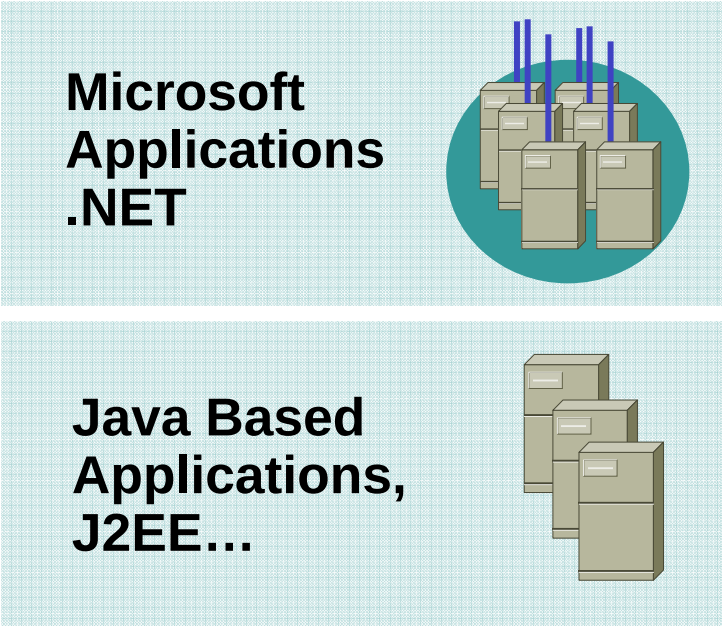
- **Purpose built server switching fabric enabling:**
  - **Low latency RDMA**
  - **Server virtualization**
  - **GRID/Utility computing**
- **Clustering environments**
  - **Database clustering**
  - **HA clustering**
  - **HPC clustering**
- **Gateway to IP switching and storage layers**
- **New, leading edge, still maturing**

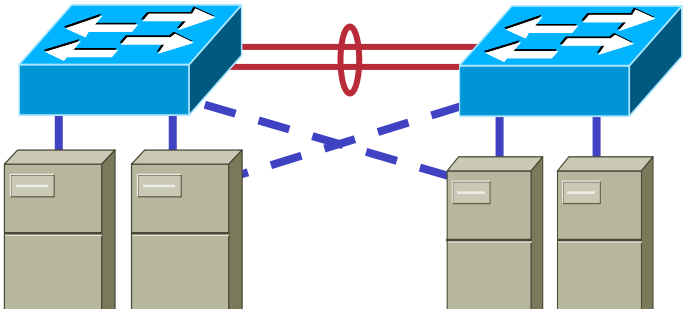


服务器群网络的设计需求

什么时候需要2层的网络连接
满足服务器群应用要求

- **Clustering: applications often execute on multiple servers clustered to appear as a single device; common for HA and load balancing requirements; (Windows MSCS and NLB)**
- **NIC teaming software requires layer 2 adjacency between teamed NICs**





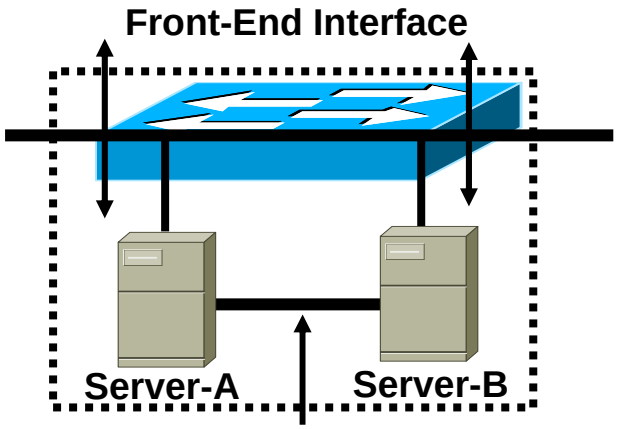
定义网络2层邻接关系

“Layer 2 adjacency between servers means that the servers are in the same broadcast domain. When servers are Layer 2 adjacent, each server receives all broadcasts and multicast packets from another server.”

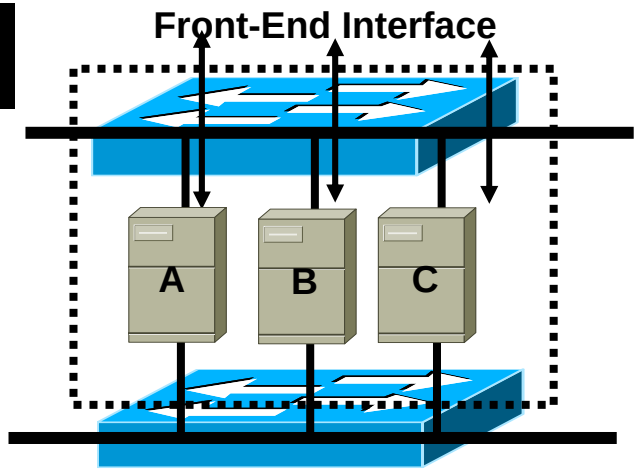
Packet Magazine: Second Quarter 2005
Designing the Data Center Access Layer

定义集群的服务器
高可用性集群

- Common goal: combine multiple servers to appear as a unified system through special s/w and network interconnects
- A 2 Node HA cluster can use a dedicated crossover cable for exchange of data, session state, monitoring…
- Two or more servers use a switch to provide the interconnect on an isolated layer 2 segment/VLAN
- Examples: MS-Windows 2003 Advanced Server 2003 Cluster Service (MSCS), for Exchange and SQL Servers (up to eight nodes)
- Veritas Clustering for HA
- L2 Adjacency is required



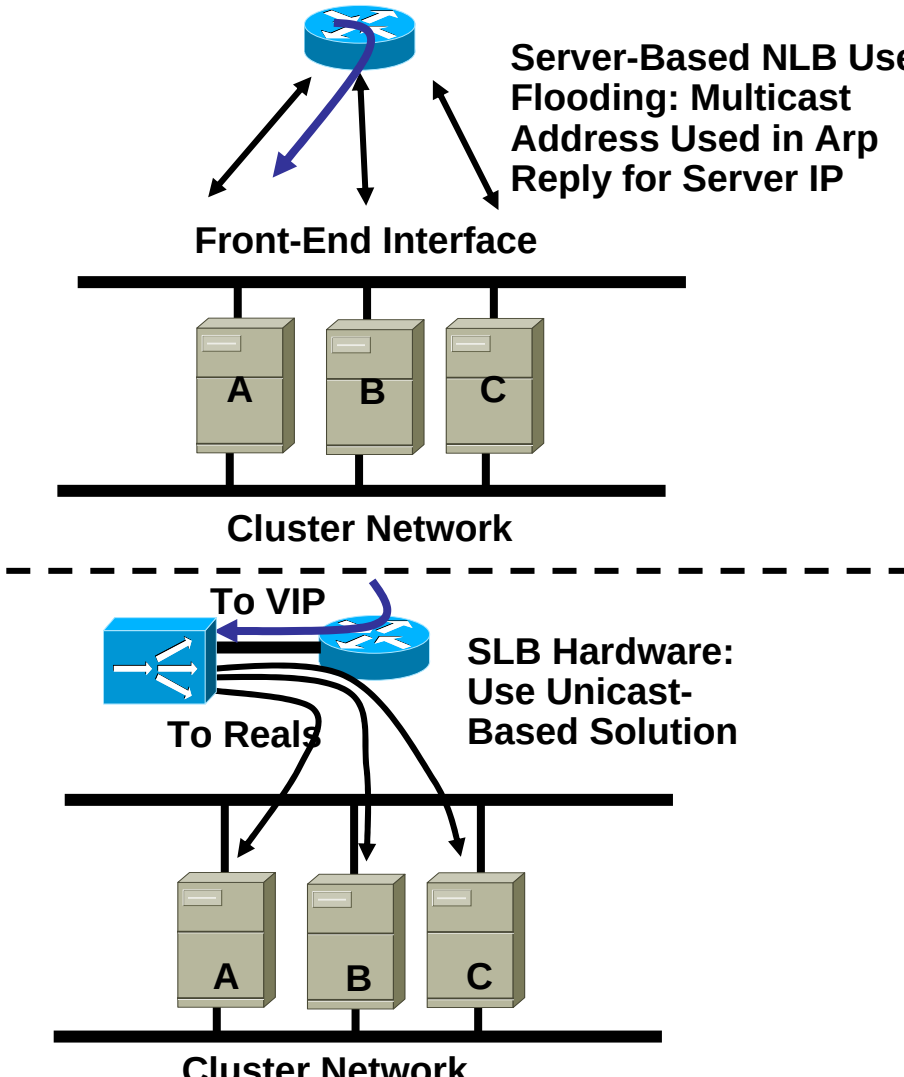
Cluster Network with 2 Nodes -Crossover Cable Connection



Cluster Network with 2 or More Nodes on Switch VLAN

定义集群的服务器
负载均衡集群

- Transactional/HTTP-based applications are “clustered” for scalability purposes (MSCS NLB example, up to 32 nodes)
- Layer 2 segment used to “multicast” all incoming packets to all hosts in cluster (L2 Adjacency is required)
- Single IP address associated with a multicast MAC address in the cluster arp reply (Windows)
- Purpose built load balancers provide a standard hardware based unicast solution supporting hundreds of nodes



定义集群服务器
数据库集群

Objective: improve DB lock times and enable more efficient parallel scans

Database Cluster Examples

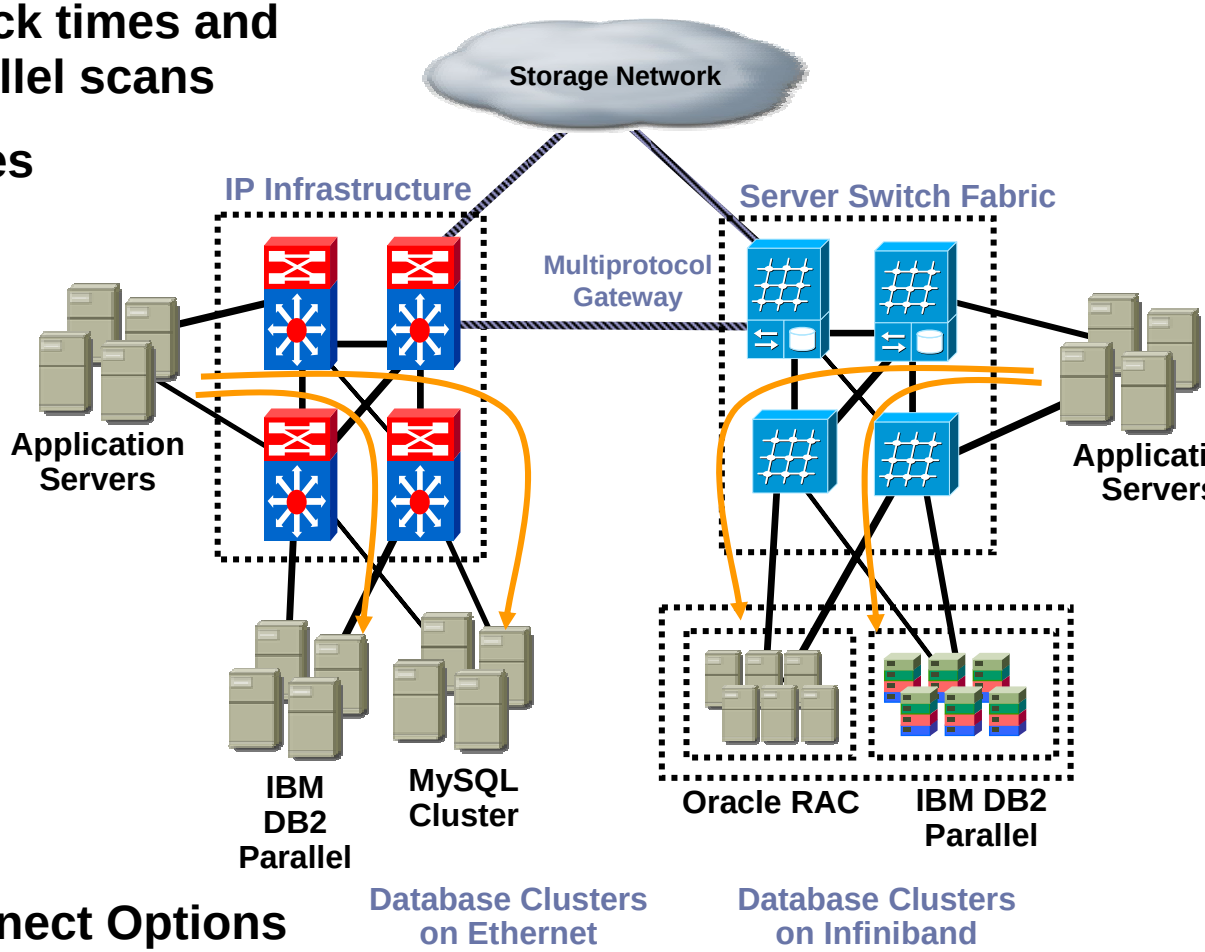
- Oracle 10g RAC
- IBM DB2 Parallel
- SQL Server
- MySQL Cluster

Storage Approaches

- Shared Everything
  - FS locking
- Shared Nothing
  - Slices up database

Standards based Interconnect Options

- Ethernet, Infiniband
- Low Latency + High B/W

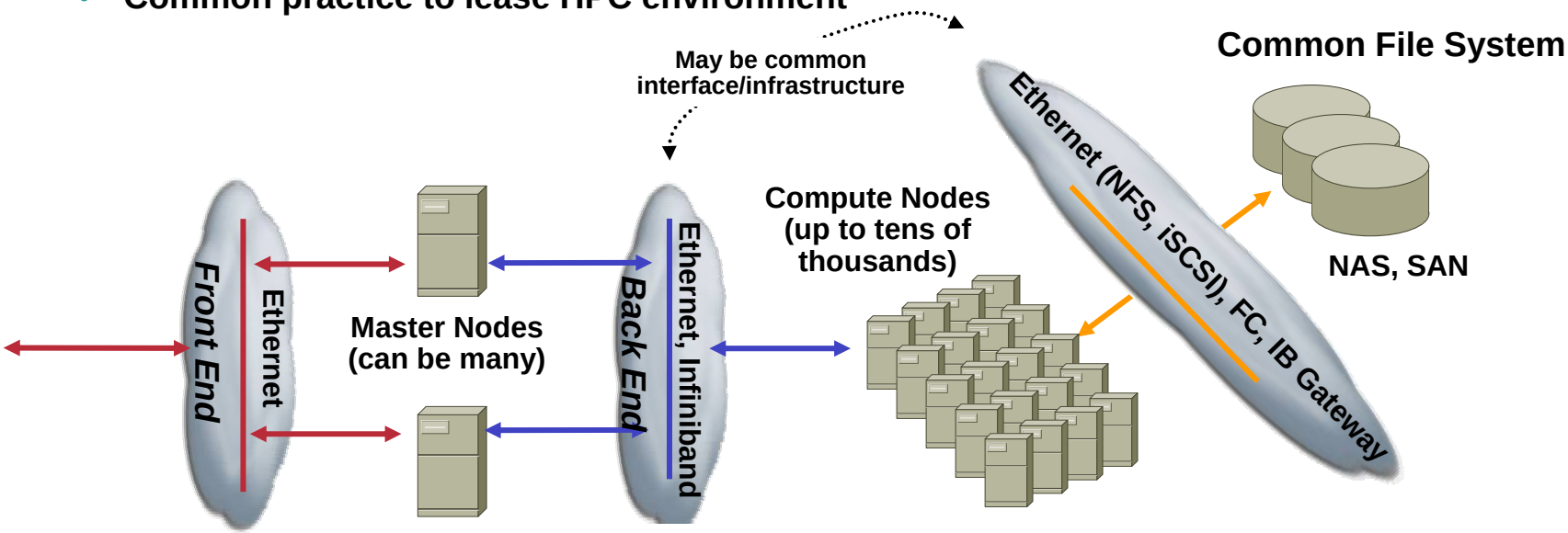


定义集群服务器
高性能计算集群

Specific Applications

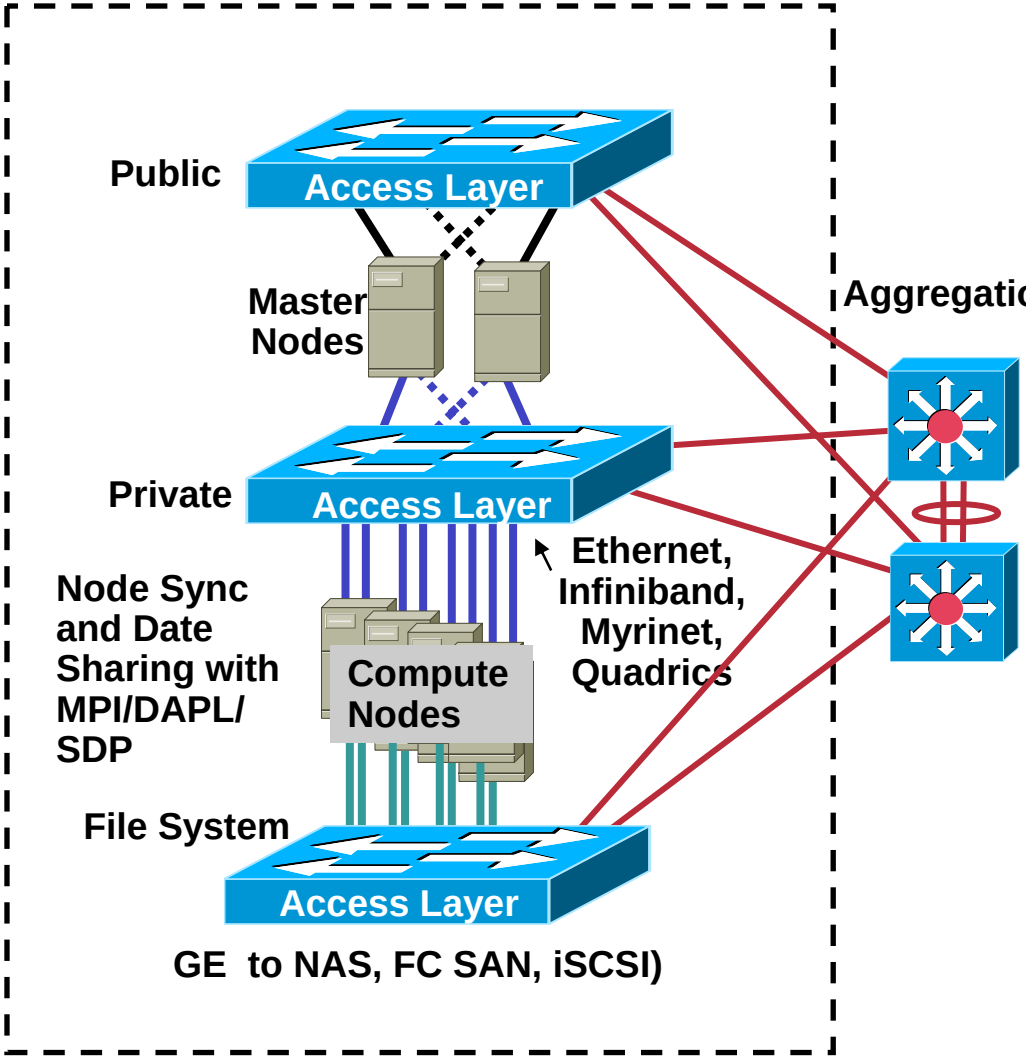
- Animation rendering
- Seismology
- Oil exploration
- Biochemistry
- Financial analysis
- Common practice to lease HPC environment

Implementations are very customized and rarely alike



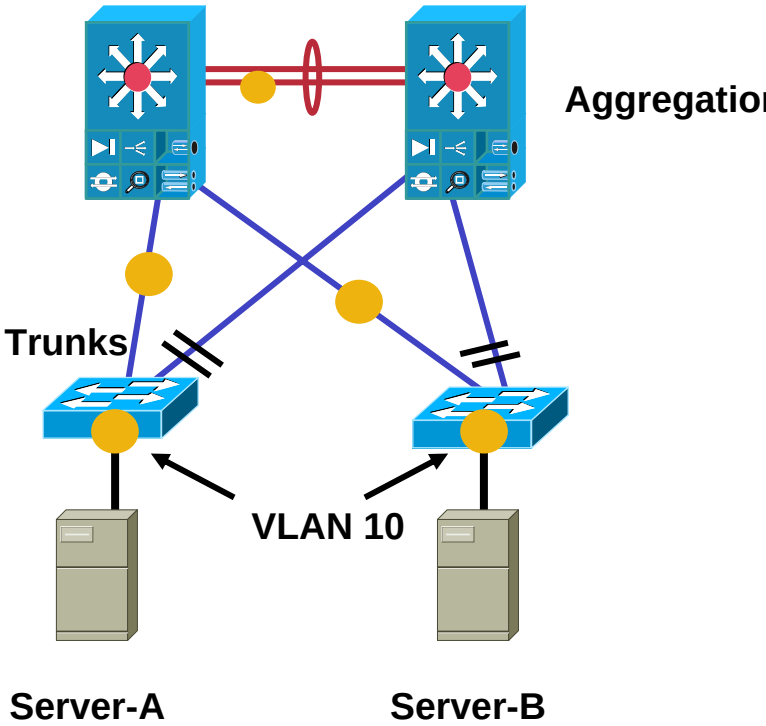
定义集群服务器
高性能计算集群需求和对网络的影响

- L2 adjacency between compute nodes may be required (application dependent)
- Common file system used by compute nodes
- Latency is critical to performance
- Network and system staff usually don't communicate clustering needs well
- Who determines which servers to include in a cluster? Same rack? row?
- Will there be an impact on access layer uplinks?



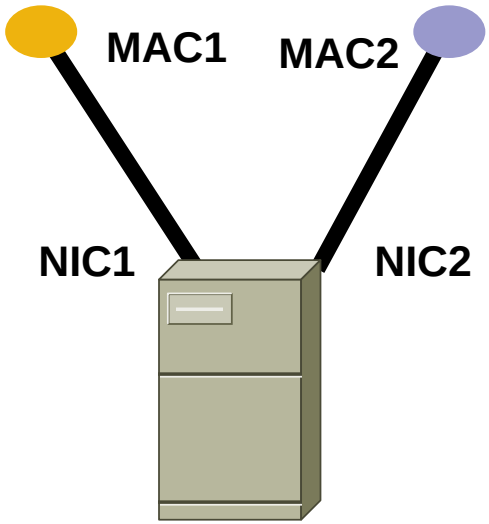
服务器集群
网络设计影响

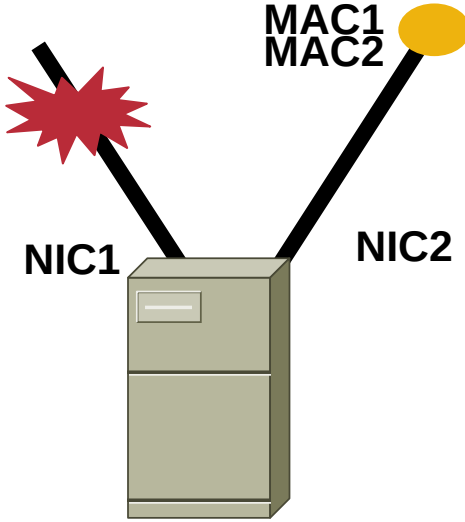
- **Server-A and server-B communicate at layer 2 to exchange state, session and other information**
- **Servers (2 or more) in cluster may be across different access switches—extending VLANs and Spanning Tree diameter**
- **Server to server cluster fabric may require higher b/w uplinks (GEC, 10GE)**



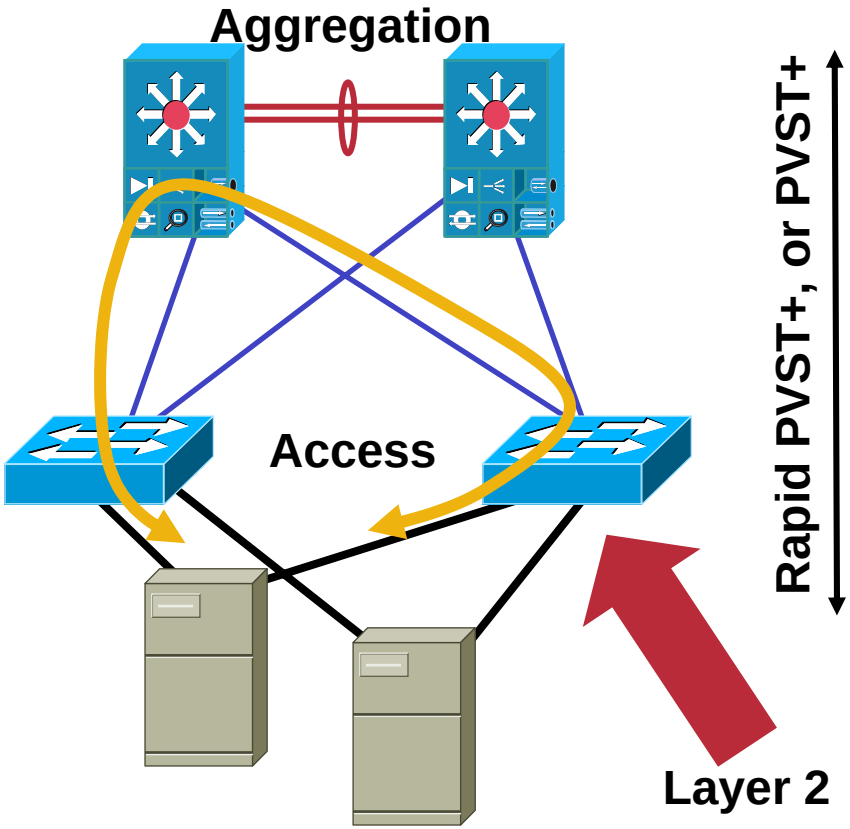
NIC Teaming 需求
Teaming 需要网卡在同一网段上

- Servers are often dual connected for high availability purposes
- The NIC driver bundles multiple NIC cards as if they were a single interface
- If one NIC loses connectivity the redundant NIC becomes active and inherits the same MAC address as the primary one
- The server is always reachable at the same IP address
- This means that both NIC's need to belong to the same BROADCAST domain—same subnet
- Optional probes/heatbeats for monitoring are multicast-based

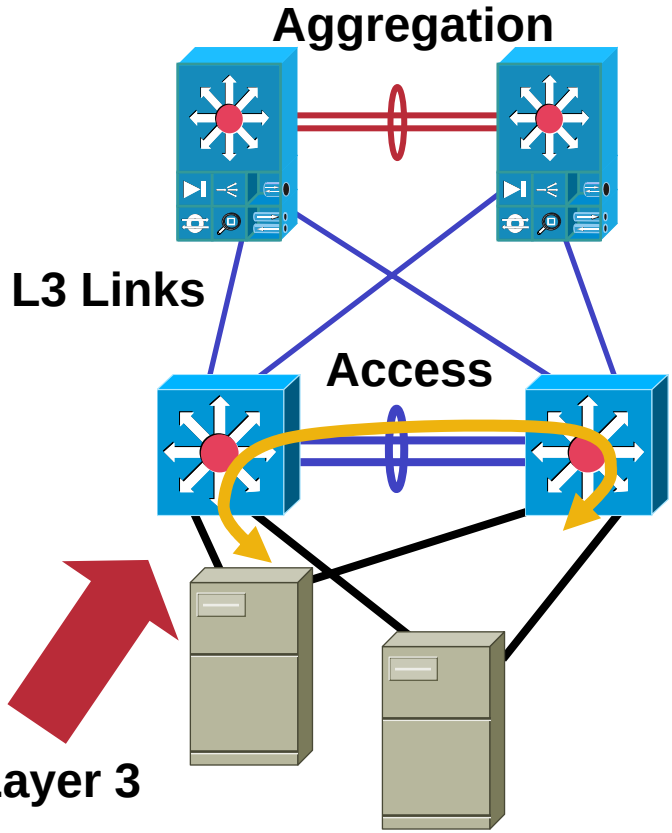




NIC Teaming 需求
数据中心中集成网卡Teaming支持



Rapid PVST+, or PVST+

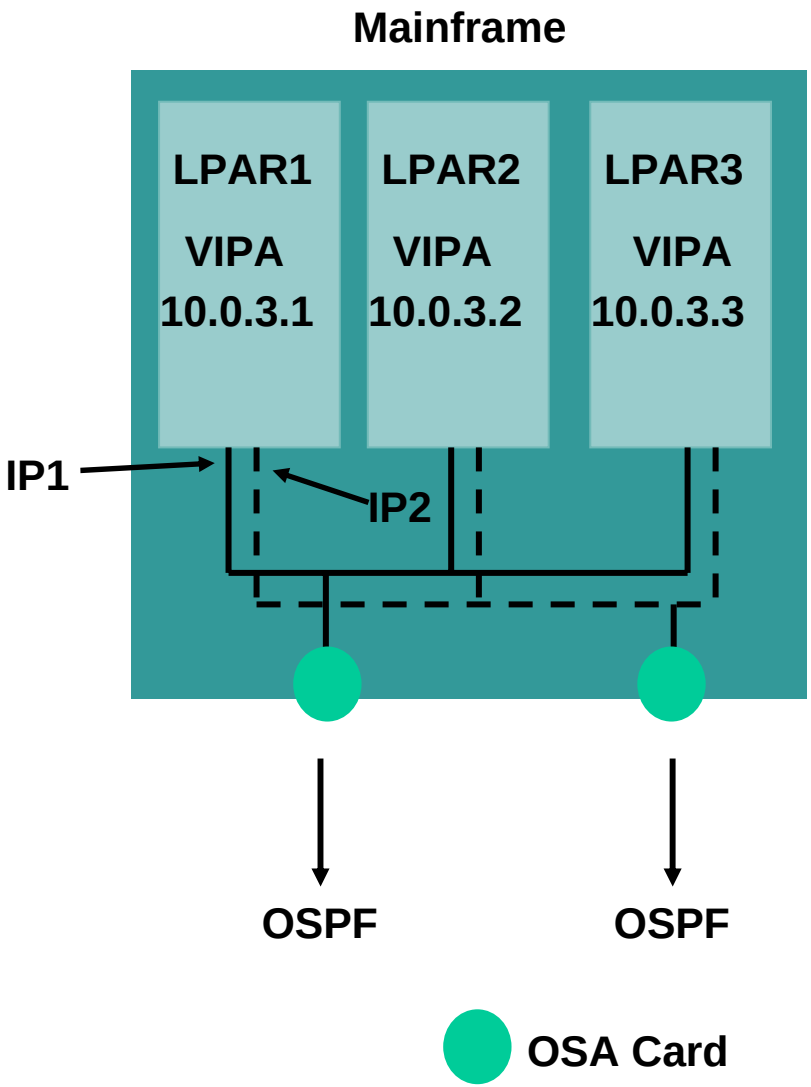


OSPF, EIGRP

- The access switches need to provide layer 2 adjacency between the NIC cards of servers with NIC teaming configured; a layer 2 path must exist between such servers

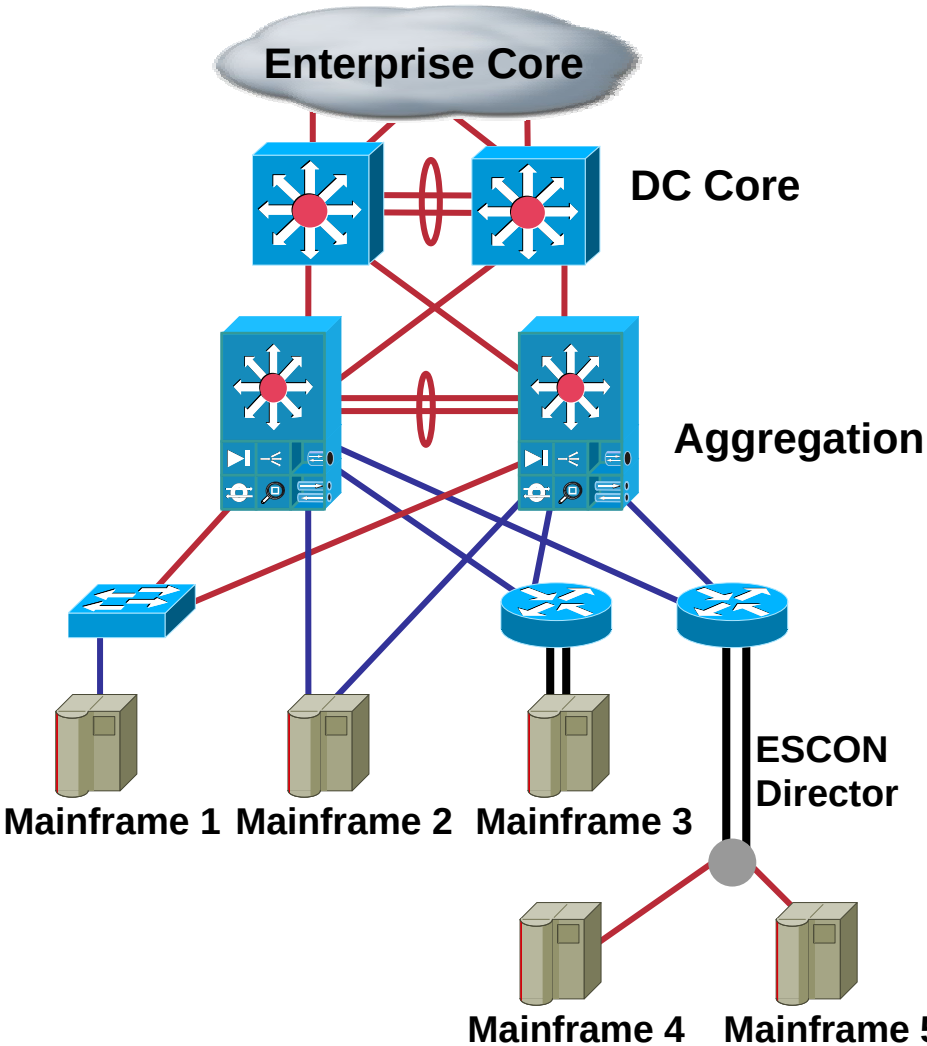
主机连接需求
IP 地址和主机

- Mainframes can be used to consolidate server farms by running Linux on several Logical Partitions (LPARs) (IBM virtual storage concept)
- Each LPAR has one IP address per network card and a static VIPA
- Mainframes with OSA cards can attach to Ethernet ports
- Mainframes use OSPF or RIP to advertise the internal IP address

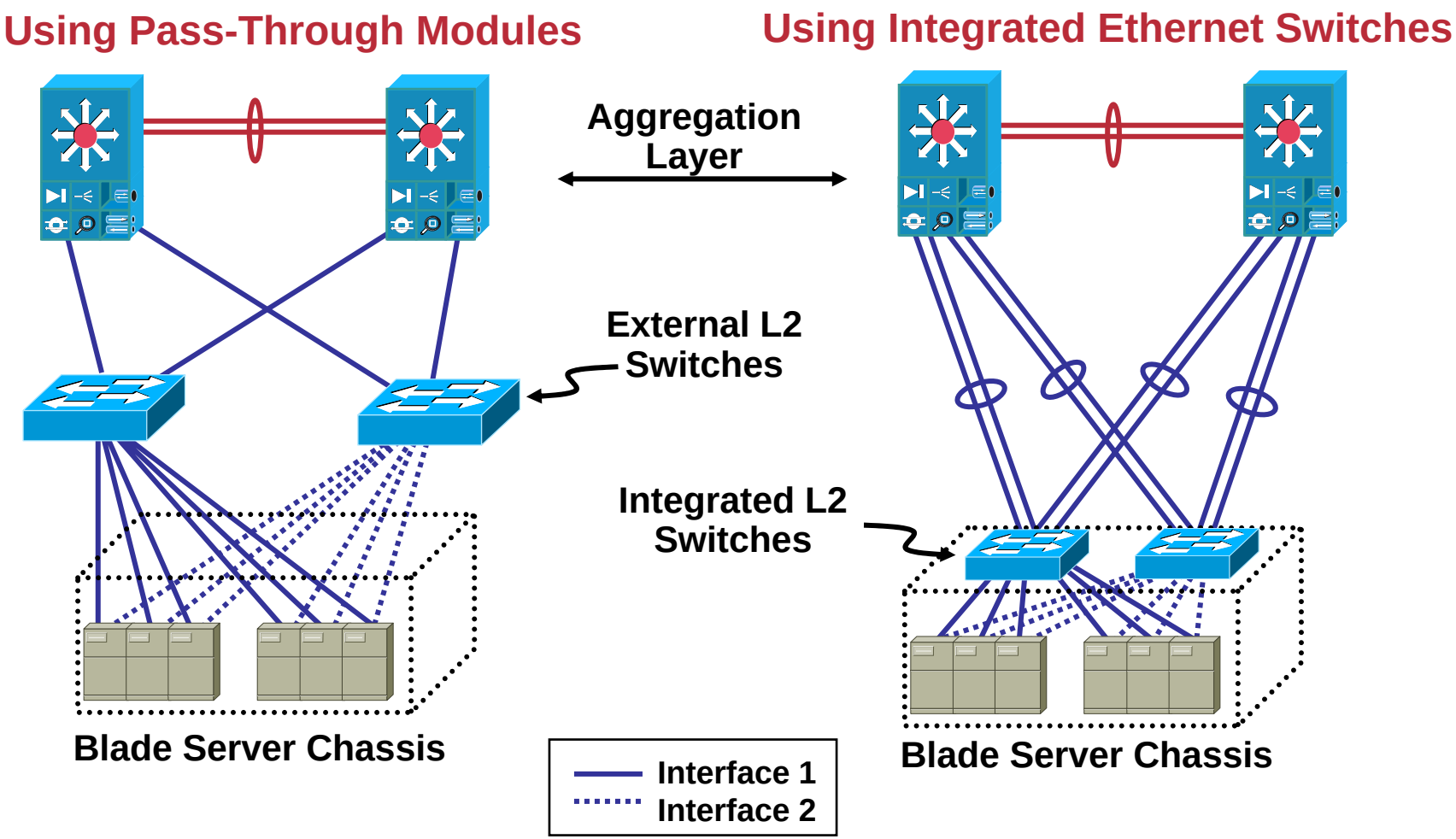


主机连接需求
3层的OSPF连接

- Mainframes with OSA cards attach to Ethernet ports (aggregation or access switches)
- Mainframes run OSPF
- Layer 3 links provide fast convergence times
- Mainframes can be attached to a 75xx/72xx router with ESCON connections



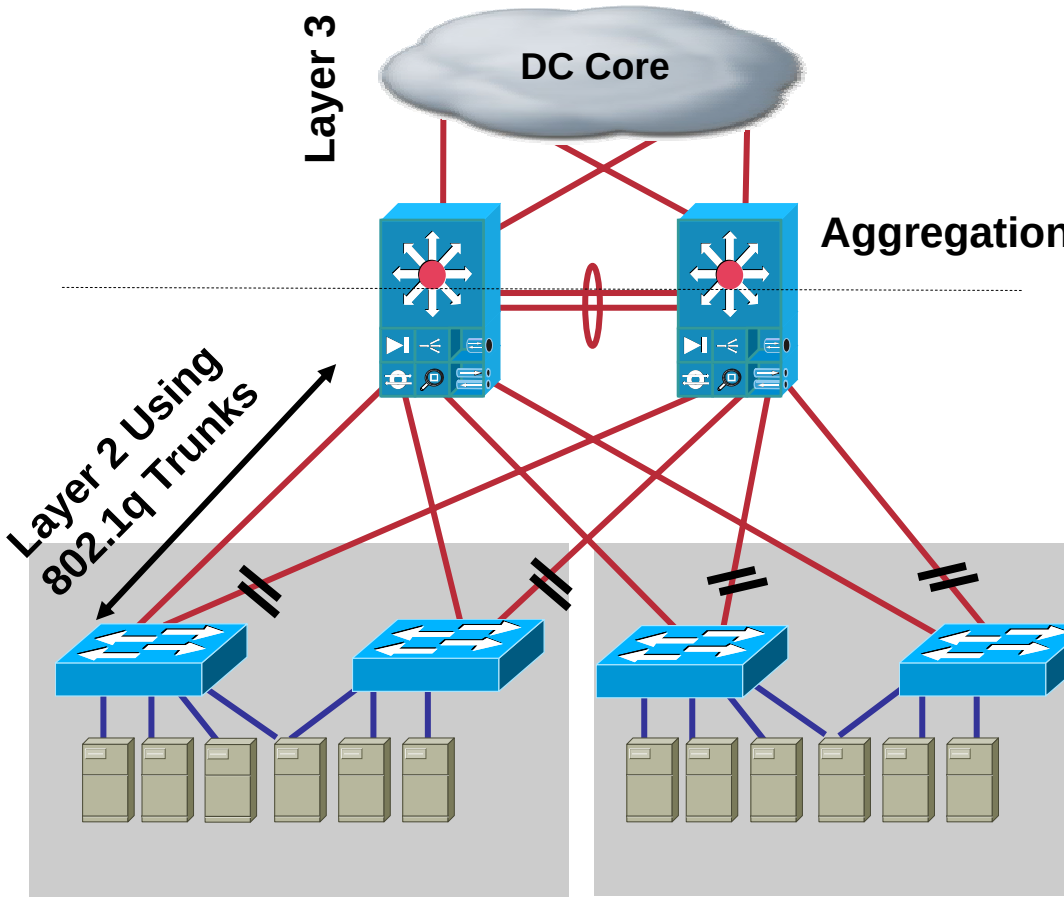
刀片服务器需求
连接选择



接入层网络设计模式

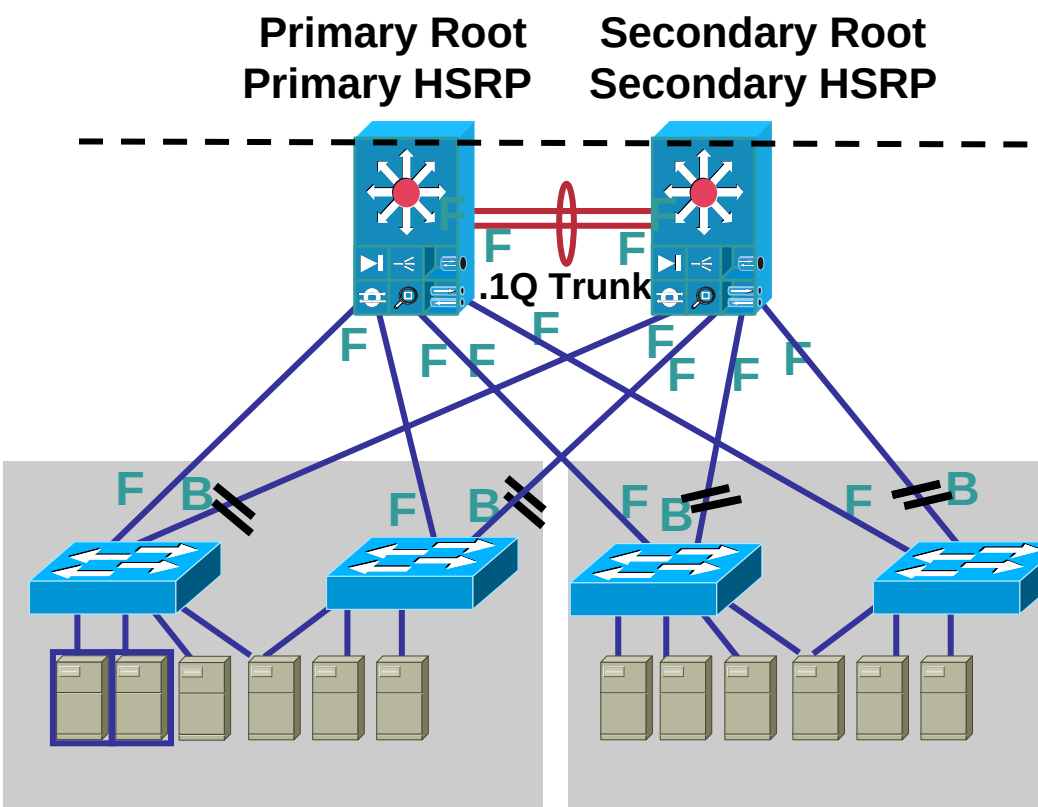
2层接入设计模式
Defining Layer 2 Access

- Layer 2 access provides layer 2 adjacency between servers in the access switches
- It DOESN'T mean carrying all VLANs unnecessarily across all access switches
- L3 processing is first performed in the aggregation layer
- L2 topologies consist of looped, loop free, and hub and spoke



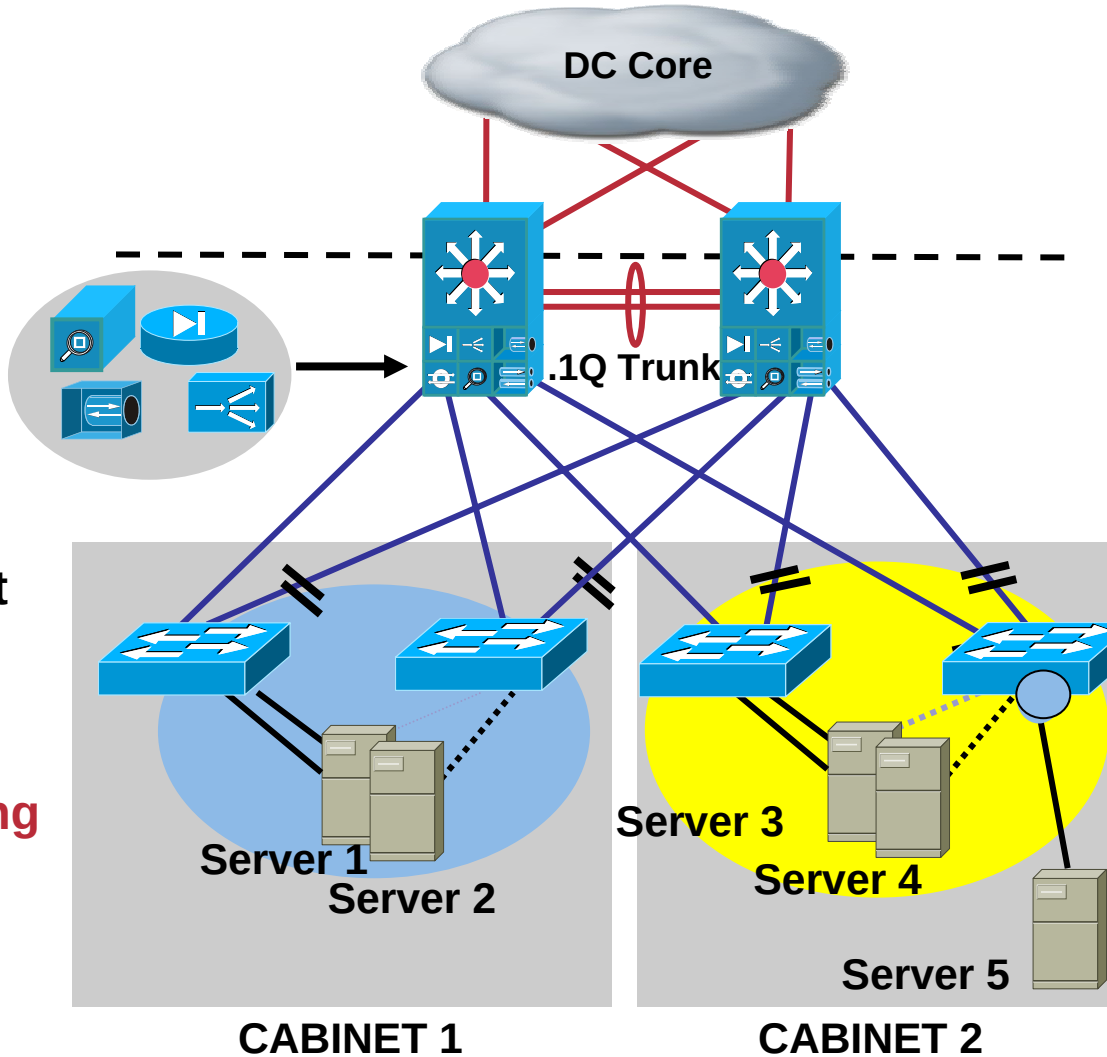
2层接入设计模式
环路设计

- 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 uplink that is blocking
- The backup link goes forwarding when the primary link is lost



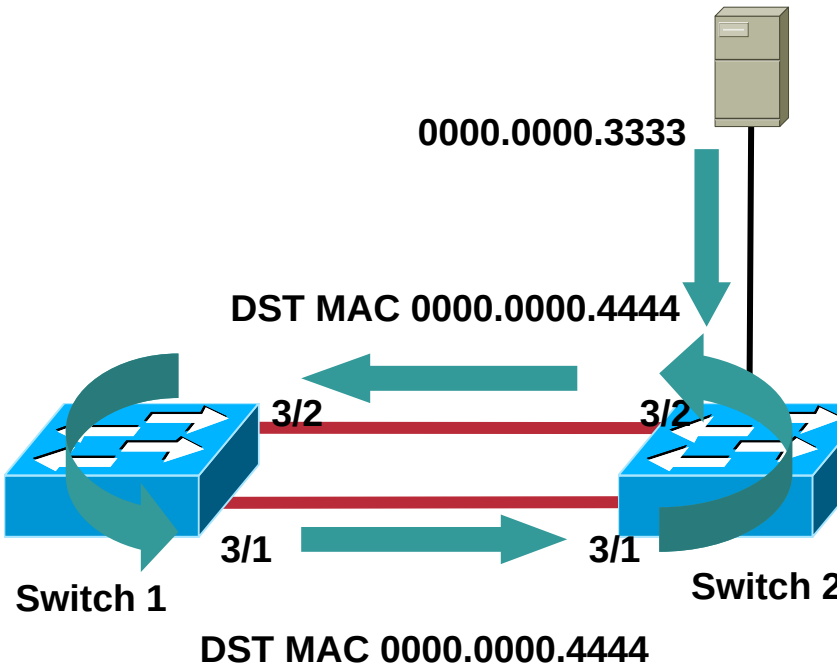
2层接入设计模式
环路设计

- Services like firewall and load balancing can be deployed at the aggregation layer and shared across multiple access layer switches
- VLANs are primarily contained between pairs of access switches
- A VLAN may be provisioned on a different access switch if administrative reasons require this
- NIC teaming and clustering can be supported across access layer modules



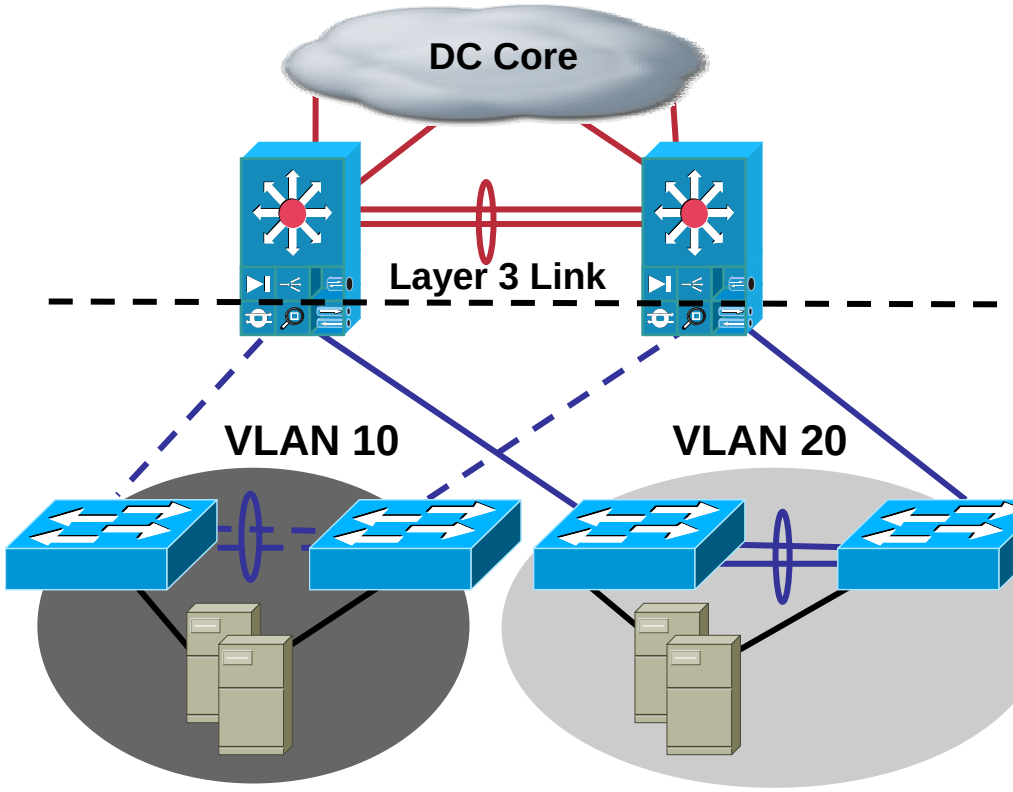
2层接入设计模式
环路设计的缺点

- **Main drawback: if a loop occurs the network may become unmanageable due to the infinite replication of frames**
- **New features plus best practices improve stability and prevent loop conditions**
  - **UDLD**
  - **Loopguard**
  - **Rootguard**
  - **BPDUguard**



2层接入设计模式
无环路设计

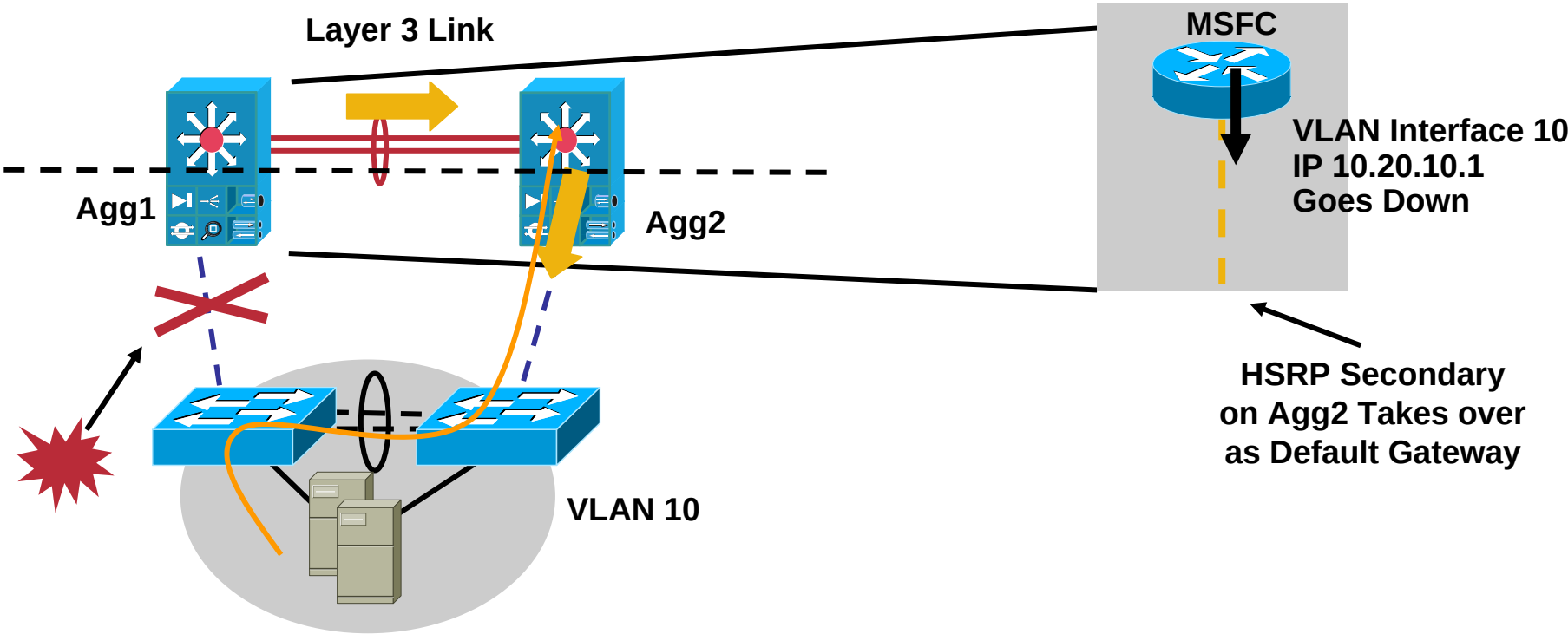
- Each pair of access switches are assigned a set of VLANs specific to that pair
- No VLANs are trunked between aggregation switches
- Spanning Tree is enabled but no port is blocking
- All links are forwarding
- NIC teaming and clustering can be supported within access layer modules



VLAN 10 Gateway 10.10.10.1

VLAN 20 Gateway 10.10.20.1

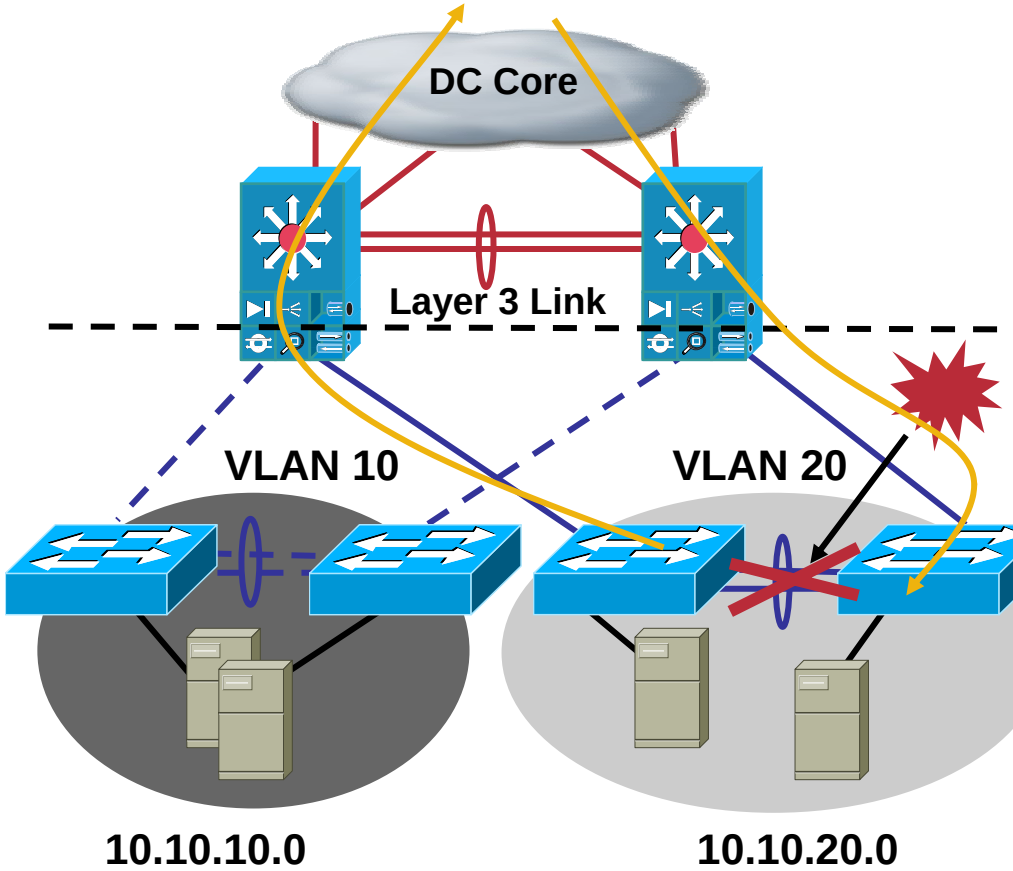
2层接入设计模式
无环路设计和Autostate



- If the uplink connecting access and aggregation goes down, the VLAN interface on the MSFC goes down as well (autostate, i.e. when there is no port forwarding on a given VLAN, the VLAN interface on the RP goes down)
- There can be service module implications as state is not conveyed
- See new tracking features for CSM and FWSM

2层接入设计模式
无环路设计缺点

- If the trunk between access switch pairs is broken, the return IP path may be broken
- VLANs must be restricted to access switch pairs
- If VLAN's are extended between access layer modules then STP blocking will occur

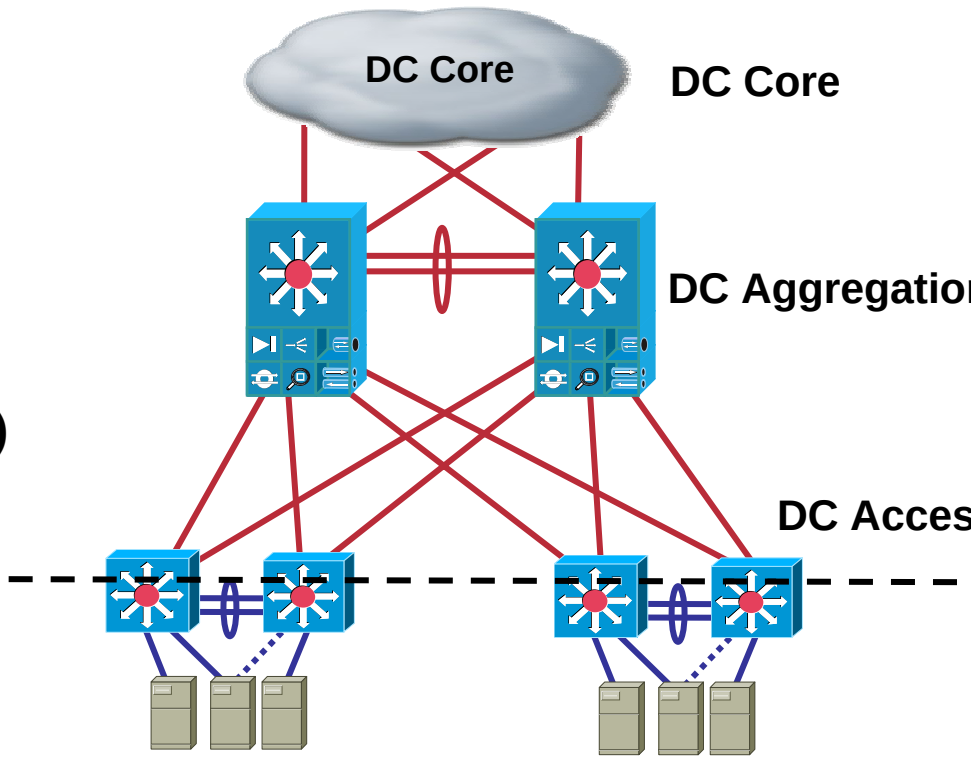


VLAN 10 Gateway 10.10.10.1

VLAN 20 Gateway 10.10.20.1

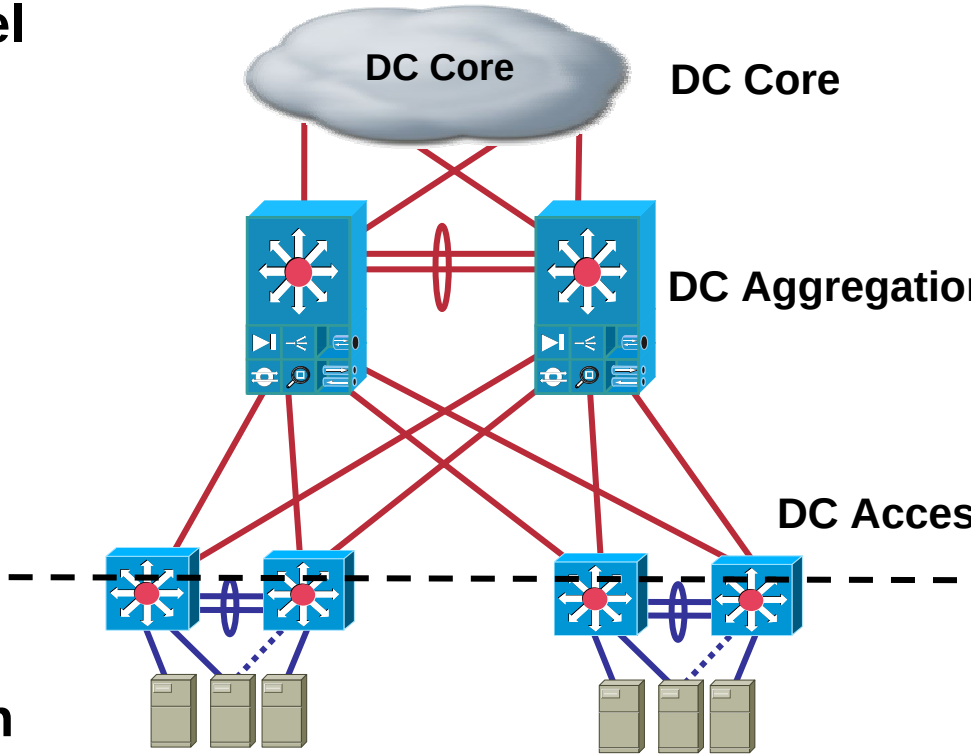
3层接入设计
定义3层接入

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



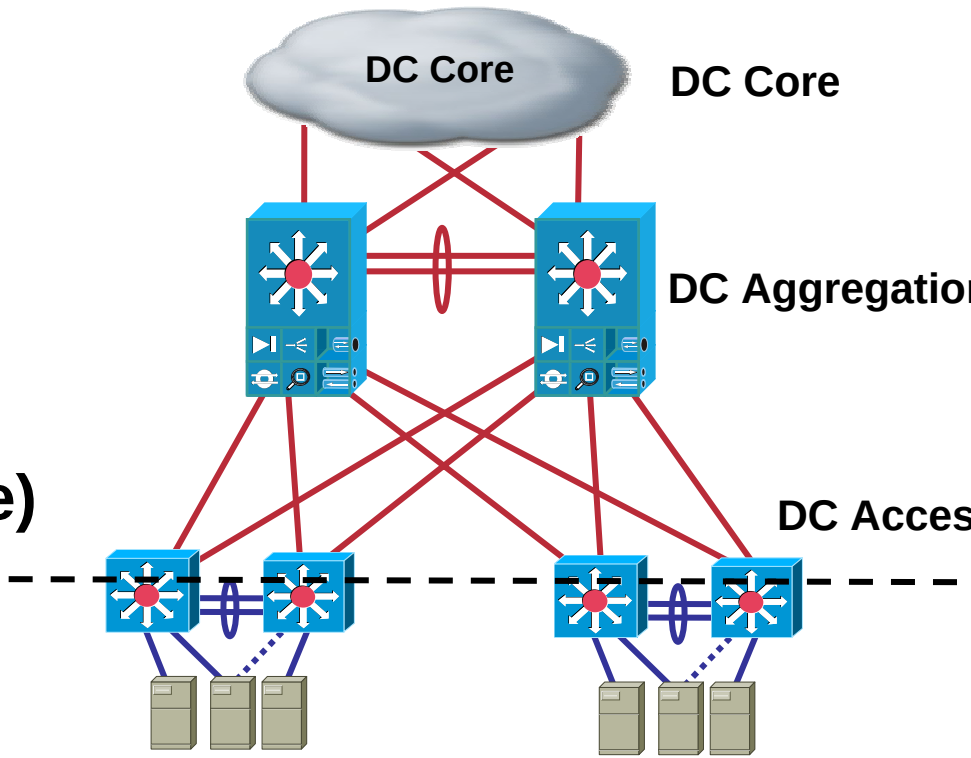
3层接入设计模式
3层接入的优点

- Minimizes broadcast domains attaining high level of stability
- Meet server stability requirements or isolate particular application environments
- All uplinks are available paths, no blocking (up to ECMP maximum)
- Load balance uplink path selection with GLBP or manual HSRP configuration
- Very good convergence time can be attained



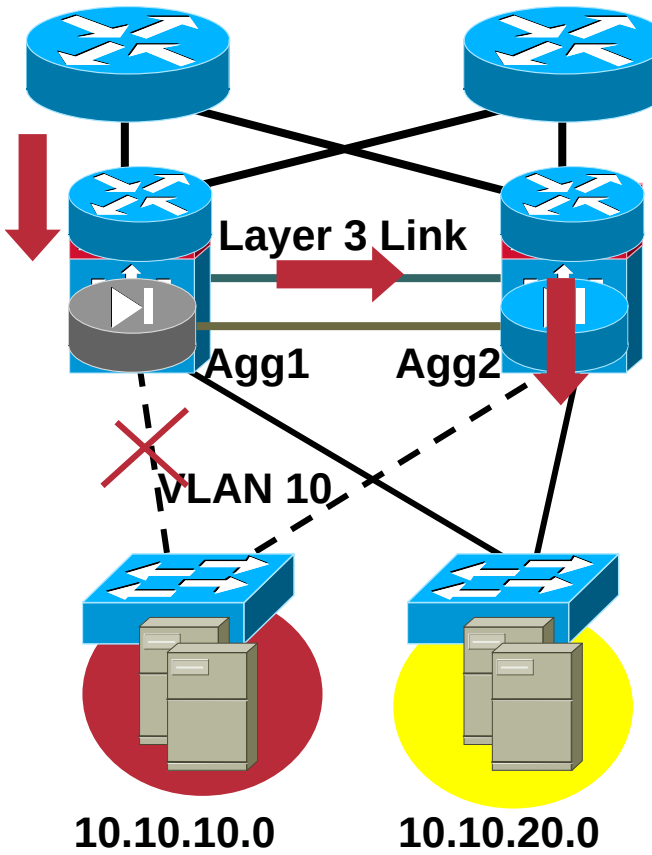
3层接入设计模式
3层接入设计的缺点

- Clustering and NIC teaming limited to access pairs
- IP address space management
- Service Module implications (next slide)

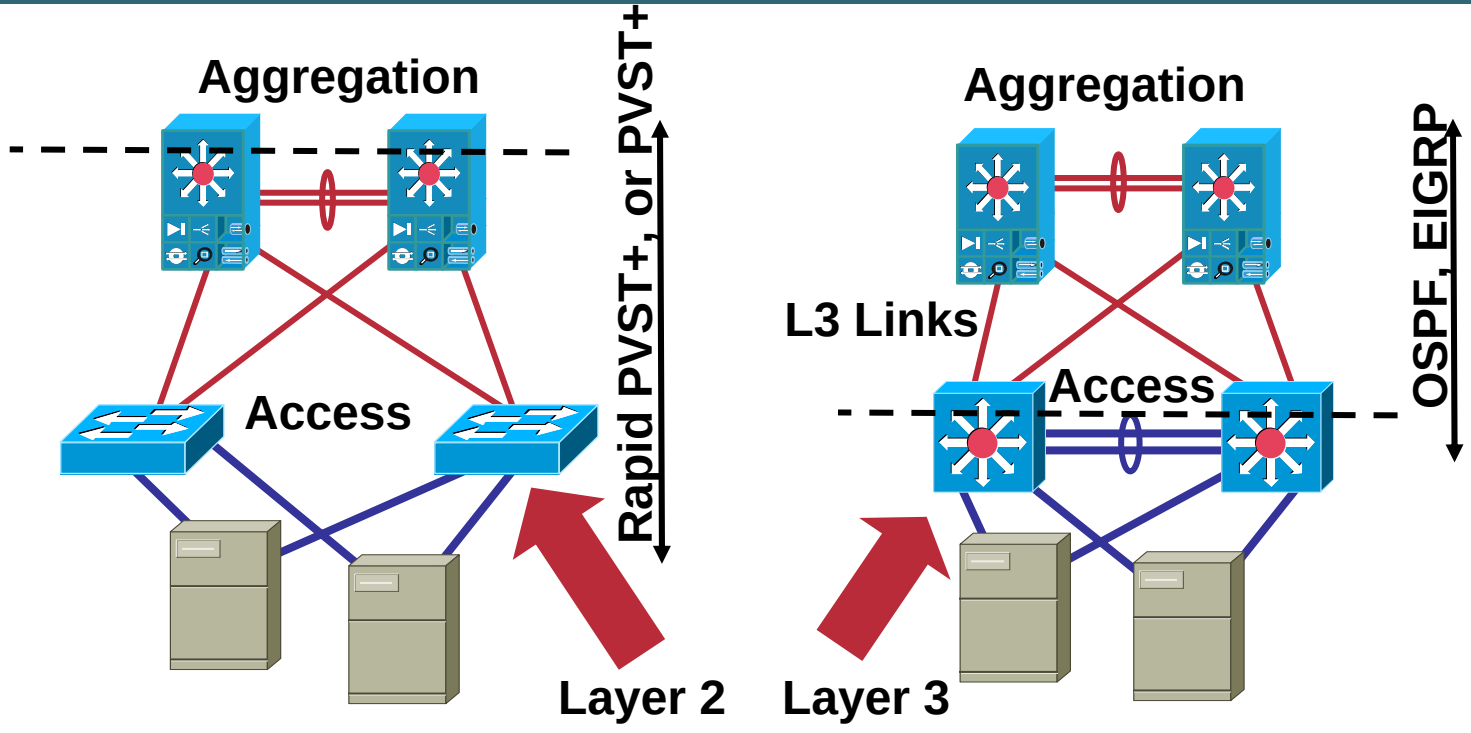


选择2层接入或3层接入
服务模块的需求

- L2 adjacency requirements between servers and service modules may be required
- Service module active/standby vs. active/active operation considerations
- Service modules require L2 adjacency for state and session synchronization
- Utilize service module interface tracking and monitor features
- TEC-DC102 or Bof-04 for more details



2层接入和3层接入比较
我们的需求是什么



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

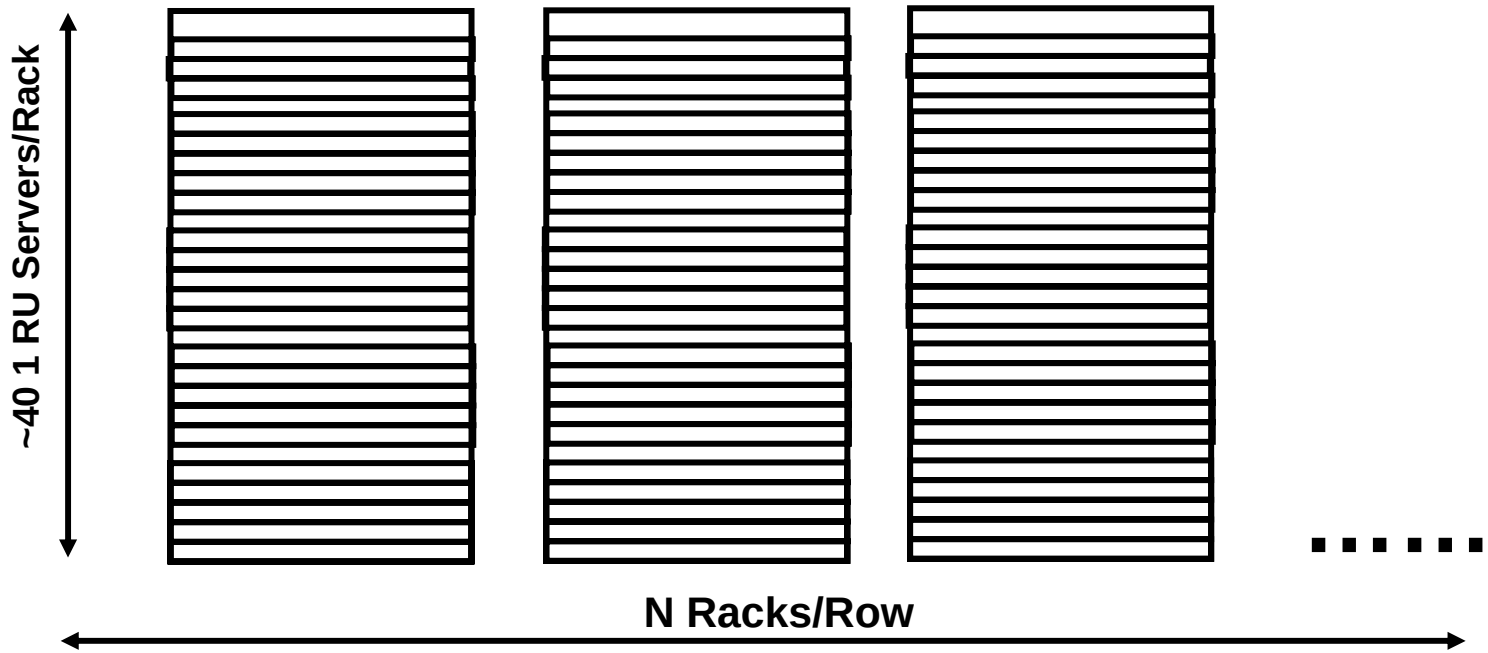
- Difficulties in managing loops
- Staff skillset—time to resolution
- Convergence properties
- NIC teaming—adjacency
- HA Clustering—adjacency
- Specific application requirements
- Broadcast domain sizing
- Oversubscription requirements
- Link utilization on uplinks
- Ability to extend VLANs

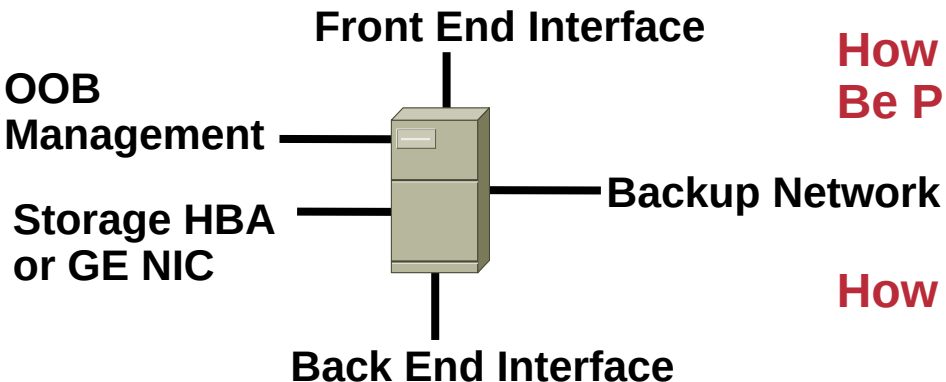
密度和可扩展性

密度和扩展性的含意
Modular and 1RU Access Layer Switching Models

- **What are the issues?**
  - Server density
  - Management
  - Oversubscription
  - Equipment sparing
  - Redundancy
  - Cabling
  - STP scalability
  - Environmentals
- **The right solution is completely based on business requirements**
- **Hybrid implementations can and do exist**

密度和扩展性的含意
Example Server Farm Cabinet Layout





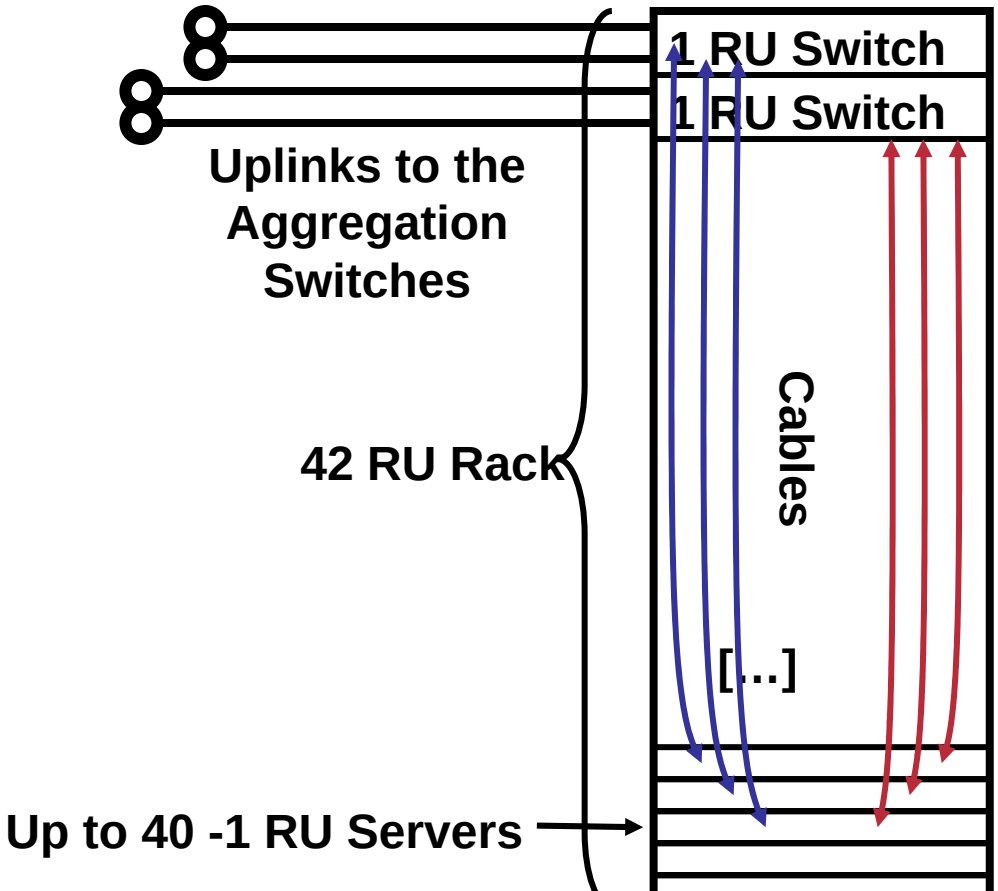
How Many Interfaces per Server Should Be Planned For?

How to Provide Network Connectivity?

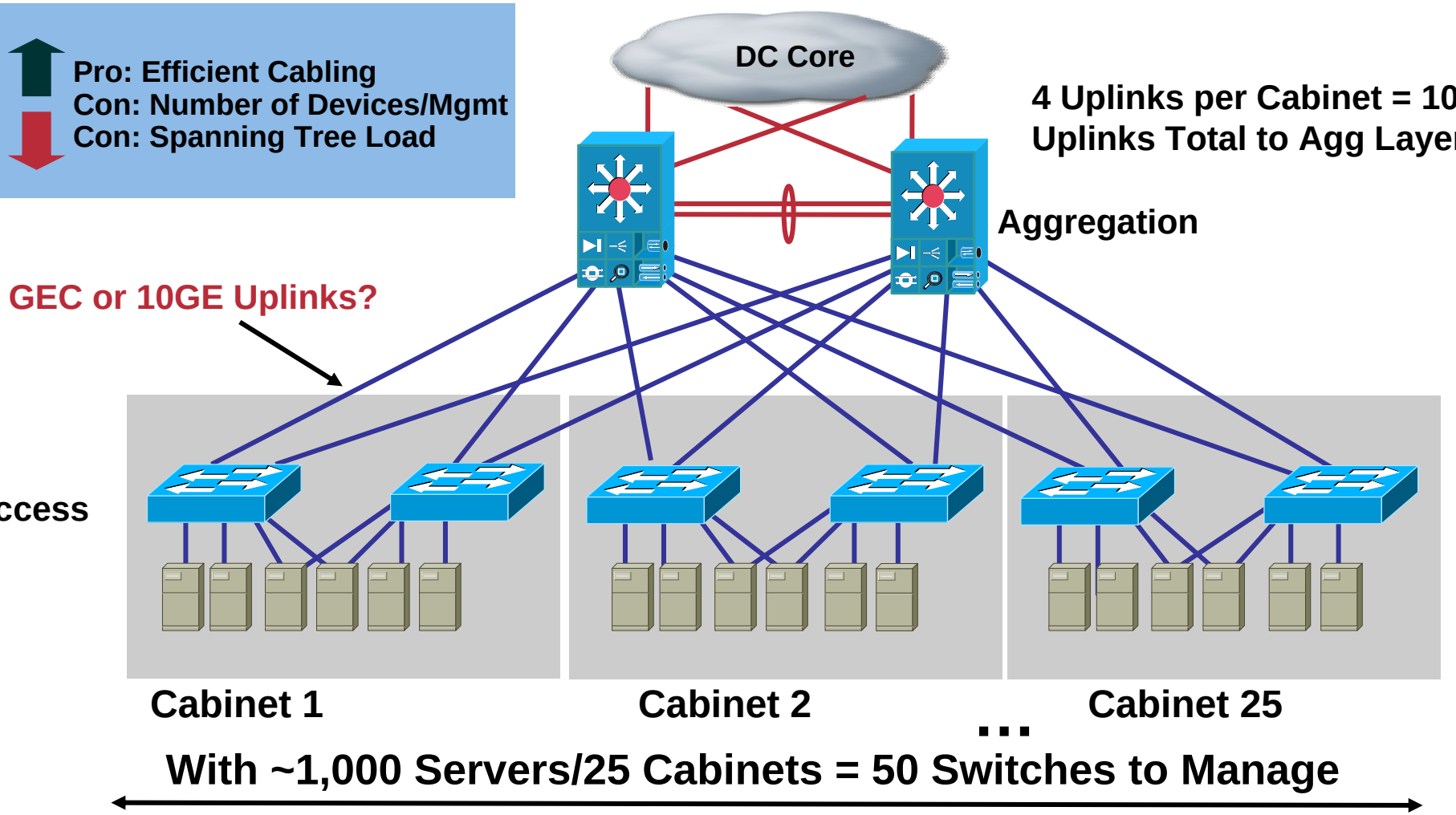
密度和扩展性的含意
Cabinet Design with 1RU Switching

Servers Connect Directly to a 1RU Switch in the Rack

- Minimizes the number of cables to run from each cabinet/rack
- All servers are dual homed: 2 -1RU switches per rack are required
- Will 2 1RU switches provide enough port density?
- Cooling requirements may not permit a full rack of servers

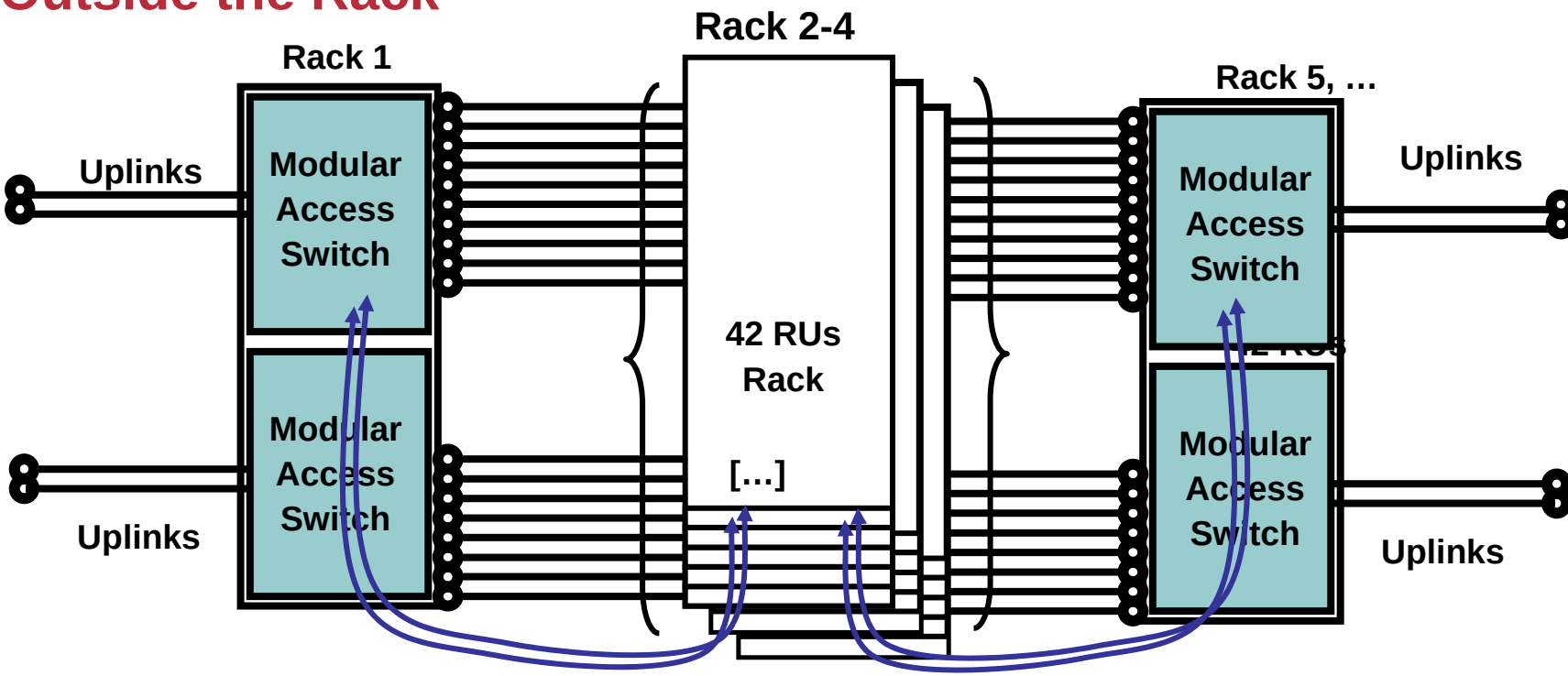


密度和扩展性的含意
Network Topology with 1RU Switching Model



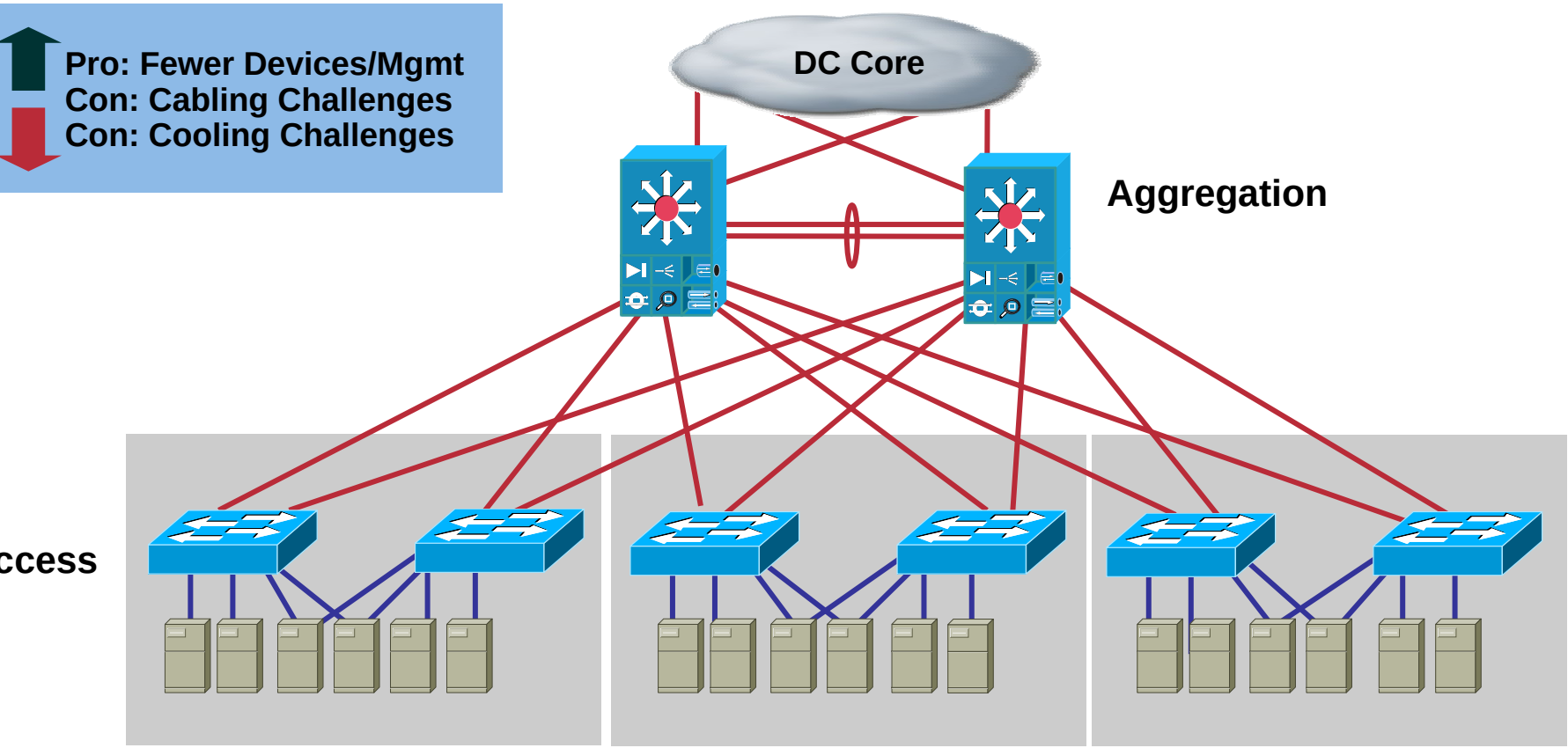
密度和扩展性的含意
Cabinet Design with Modular Access Switches

Servers Connect Directly to a Modular Switch Outside the Rack



- Cable bulk at cabinet floor entry can be difficult to manage
- Cable bulk can block air flow
- Typically placed at ends of cabinet row
- May need to space switches out within the row and at ends

密度和扩展性的含意
Network Topology with Modular Switches in the Access

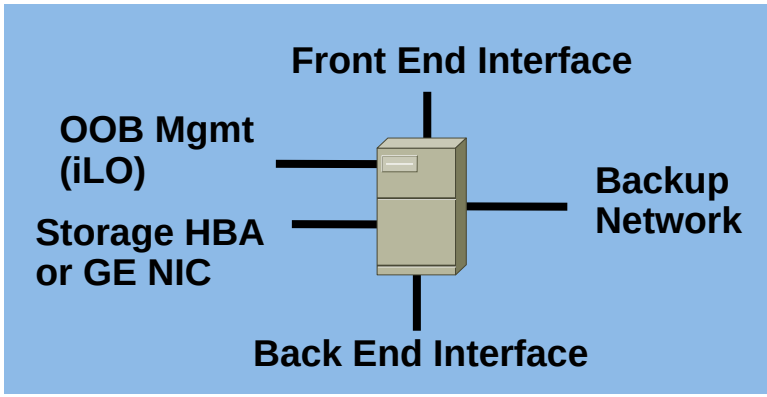


With ~1,000 Servers/9 Slot Access Switches= 8 Switches

~8 Access Switches to Manage

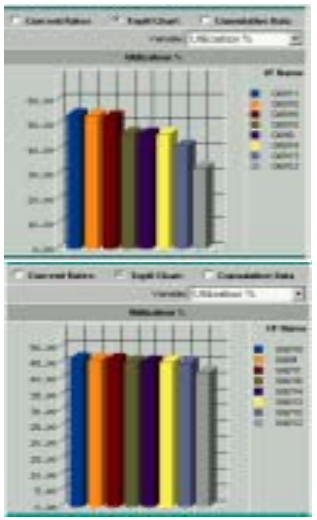
密度和扩展性的含意
Density: How Many NICs?

- **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 up to 4 1RU switches per rack to meet port density requirements**
  - 30 servers with 4 active ports = 120 ports required in a single cabinet (3x48 port 1RU switches)
  - May need hard limits on cabling capacity to
  - Avoid cross cabinet cabling



密度和扩展性的含意
Oversubscription and Uplinks

- **What is the oversubscription ratio per uplink?**
  - **Develop an oversubscription model**
  - **Identify by application, tier, or other means**
- **Consider future true server capacity (PCI-X, PCI- Express)**
  - **Server platform upgrade cycle will increase levels of outbound traffic**
- **Consider uplink choices that will scale with your business**
  - **Gigabit EtherChannel 10GE**
  - **10Gig EtherChannel**
- **Consider flexibility in adjusting oversubscription ratio**
  - **Can I upgrade to 10GE easily? 10G EtherChannel?**

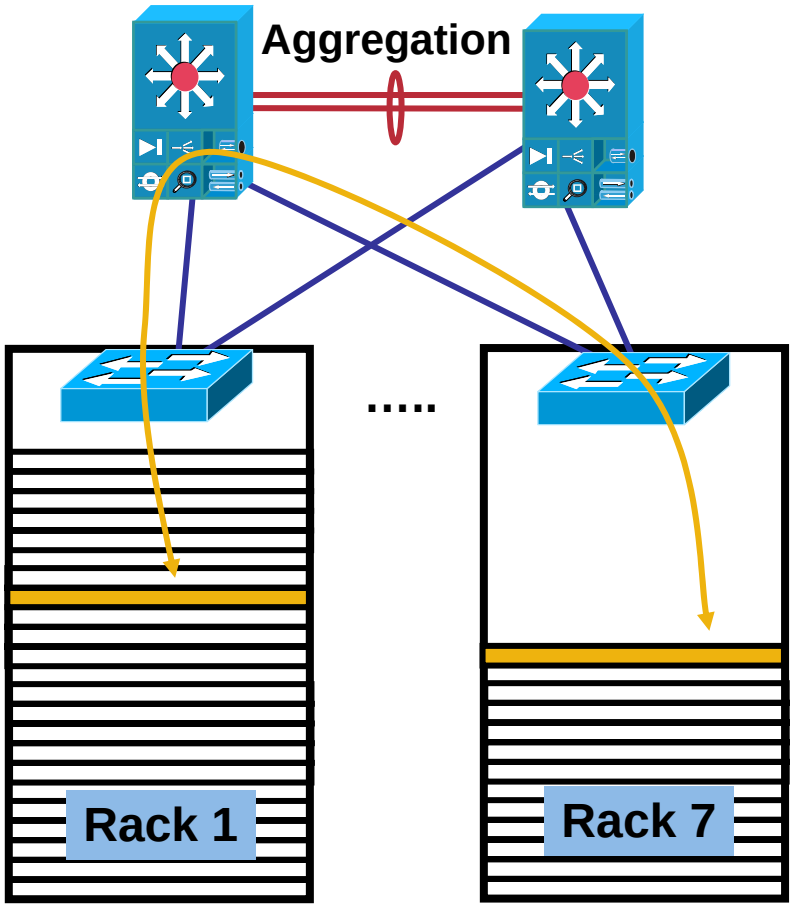


I1 Mark Noe 4/29/2005
fine graphic that shows gec
to 10GE

is this a request for CIR? Dana x65463

密度和扩展性的含意
L2 Adjacency Requirements

- What if NIC teaming or clustering requirements grow later?
- How far will the VLAN need to be extended?
- How many additional STP logical ports will be added at the agg layer?
- How to do this if a layer 3 access is used?



密度和扩展性的含意
Spanning Tree

- 1RU switching increase chances of larger spanning tree diameter
- A higher number of trunks will increase STP logical port counts in agg layer
- Determine spanning tree logical and virtual interfaces before extending VLANs or adding trunks
- Use aggregation modules to scale STP

  (Covered in More Detail Later in Presentation)

密度和扩展性的含意
Management, Sparing and Redundancy

- **More switches could result in:**
  - **More extended VLANs**
  - **Higher maintenance**
    - **Code upgrades**
    - **Configuration changes**
- **Use a network device to server ratio as a TCO guideline**
- **Consider sparing standards**
  - **Skill set requirements**
  - **Time to resolution**
- **Redundancy:**
  - **Failure exposure level per application**
  - **CPU, power, uplinks**

密度和扩展性的含意
Cabling

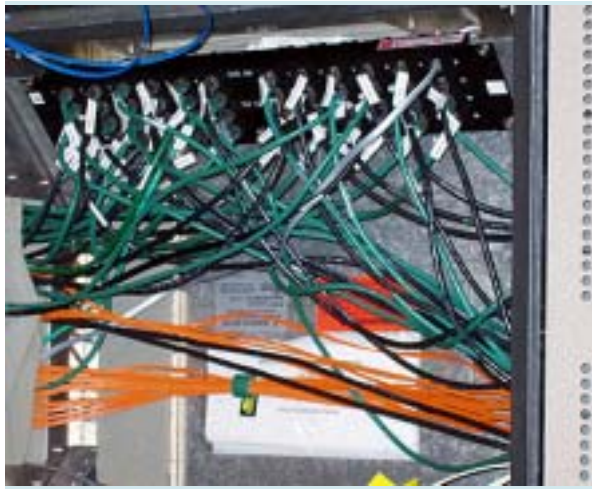
- Carefully examine cabling design
- Determine maximum cable bulk
- Determine air flow restrictions
- With modular access:

  Consider:

  Distribution within row approach

  Every other row approach

  Examine air flow of components in cabinet to avoid heat re-circulation



使用千兆以太网捆绑和万兆以太网扩展带宽

使用千兆以太网捆绑和万兆以太网扩展带宽
选择千兆或万兆上联

What Needs to Be Considered?

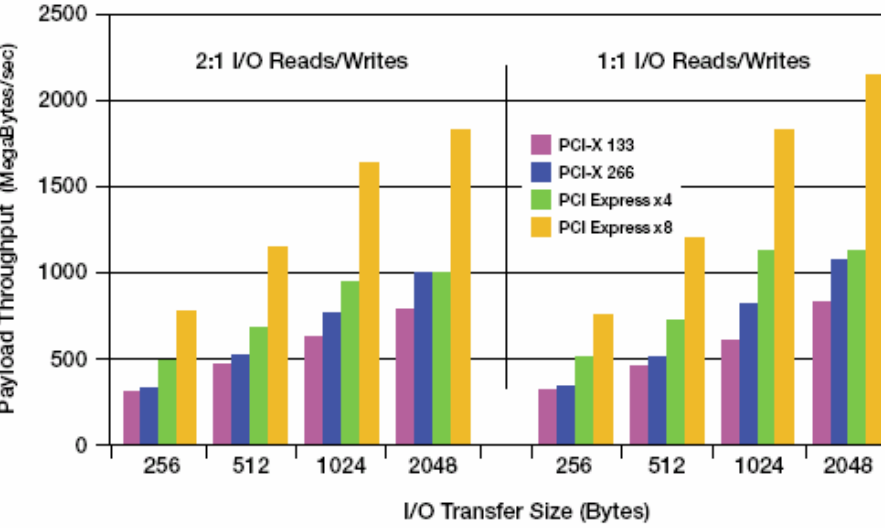
- **Server NIC Improvements**
  - **PCI-x and PCI-Express, RDMA, TOE**
- **EtherChannel Hashing Algorithm**
- **GLBP Designs**
- **ECMP Designs**
- **10GE NICs**
  - **Server consolidation efforts**
- **10GE Density**
  - **How to scale the aggregation layer**

使用千兆以太网捆绑和万兆以太网扩展带宽
服务器性能提高

- **PCI Improvements**

  PCI, PCI-x and PCI-Express

- **PCI-X now pervasive, PCI-Express now shipping**
- **PCI to PCI-X: 8:1, PCI-X to PCI-Express: 4:1 increase**



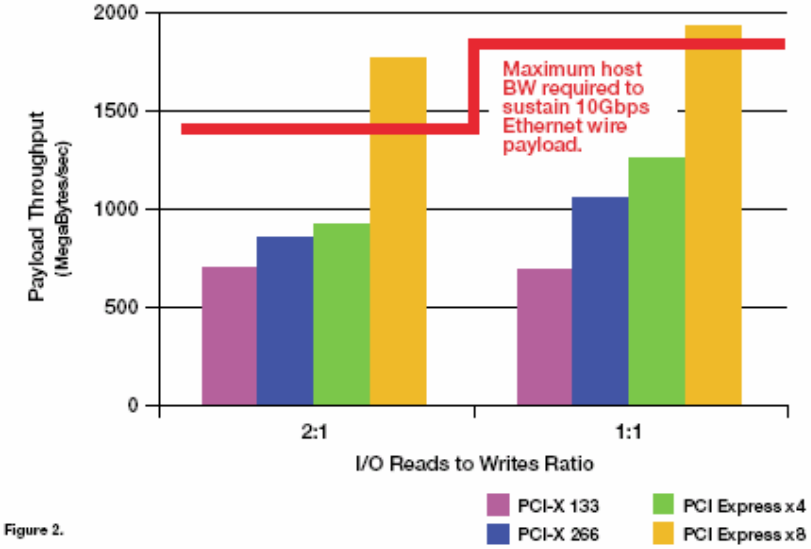
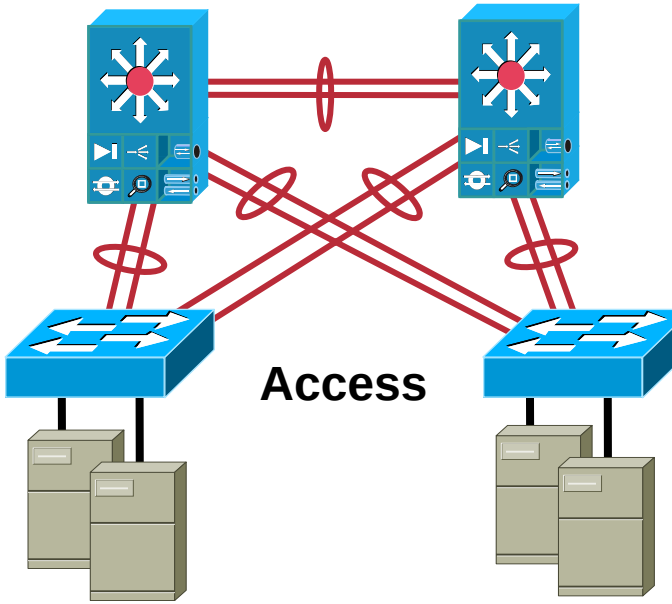
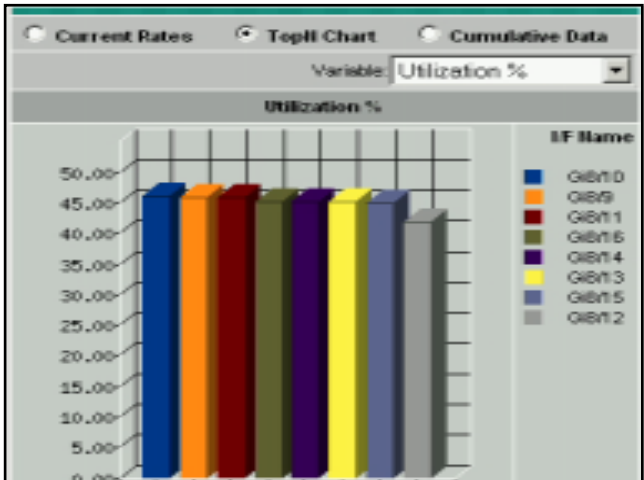


Figure 2.

使用千兆以太网捆绑和万兆以太网扩展带宽
EtherChannel Hash Algorithms

- EtherChannels provide link redundancy and increased bandwidth
- Can be on physically different line cards decreasing failure impact
- Uses two main aggregation protocols: PagP and LACP
- Optional hash algorithms available: default is L3 IP source/dest address
- The goal is to have link utilization equally distributed on all the links as in the picture on the right





使用千兆以太网捆绑和万兆以太网扩展带宽
EtherChannel vs. 10 Gigabit Ethernet

- **Analyze the traffic flows in and out of the server farm:**

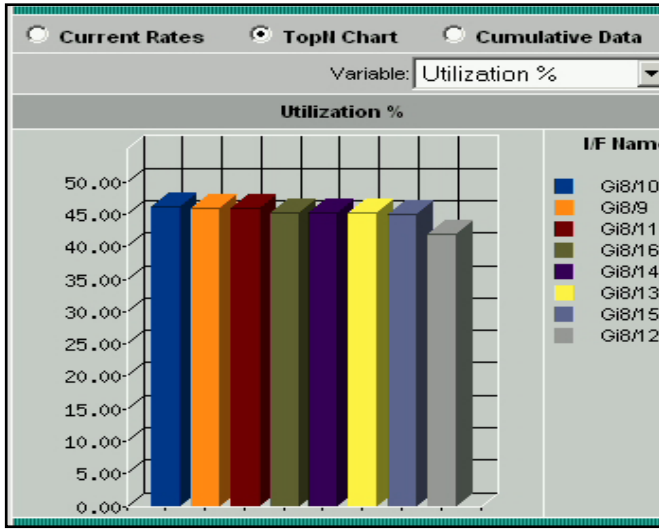
  IP addresses (how many)

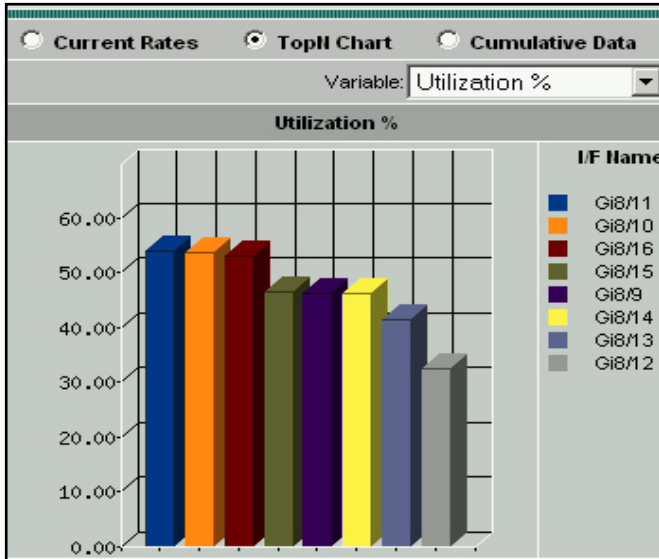
  L4 port numbers (randomized?)

- **Default L3 hash may not be optimal: look at L4 hash**

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

- **Ideal is graph on top right**
- **Bottom left graph more typical**
- **10 GigE gives you effectively the full bandwidth**



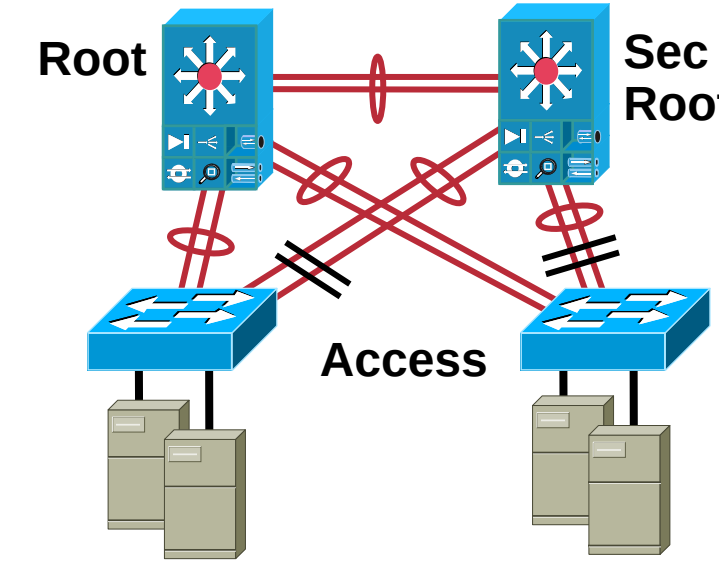


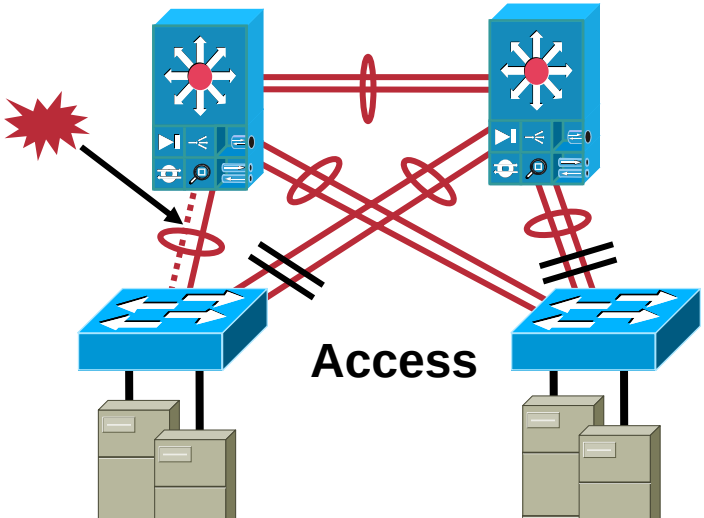
使用千兆以太网捆绑和万兆以太网扩展带宽
Optimizing EtherChannel Paths (1)

- **Under normal conditions:**

  etherchannel to the root is active
  etherchannel to the sec-root is blocking

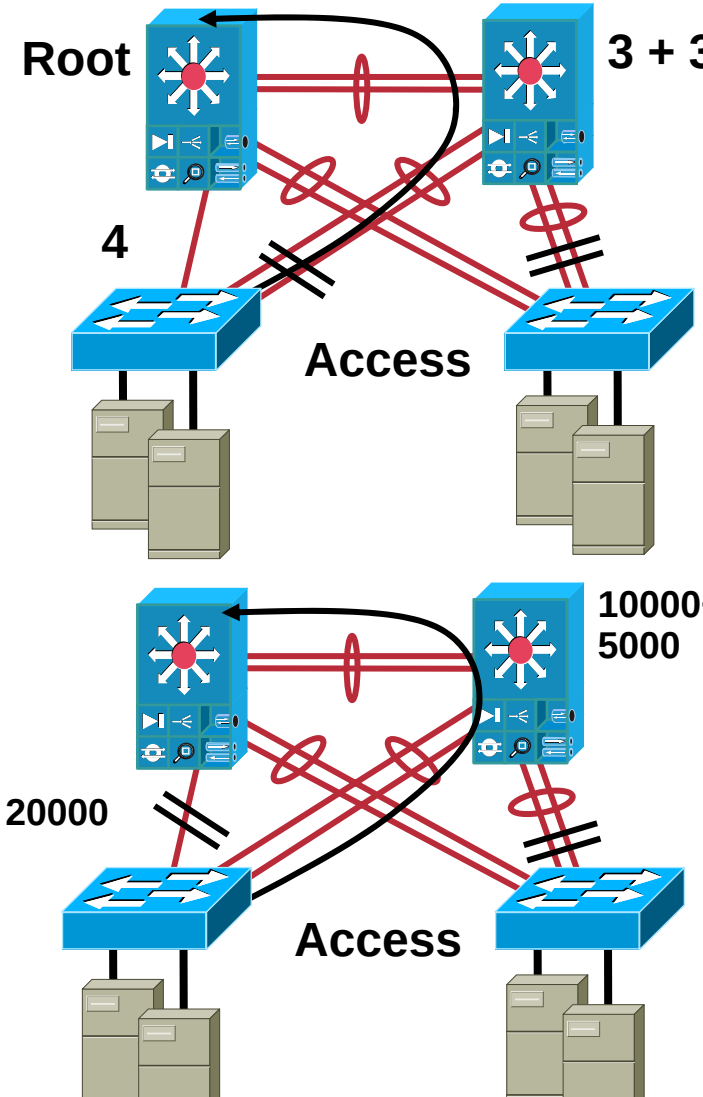
- **What if only one link in channel is broke?**
- **Will STP converge to higher b/w path?**
- **Is it possible to put the remaining link in blocking mode and the alternate port in forwarding (the 2G channel)?**





使用千兆以太网捆绑和万兆以太网扩展带宽
Optimizing EtherChannel Paths (2)

- LACP-configured channels change path cost when ports join or leave the channel
- The default STP cost of EtherChannels is as follows:
  - 1 Gb = 4
  - 2 Gb = 3
  - 3 Gb = 2
  - 4 Gb = 2
- Hence if one link fails the traffic takes the single Gigabit link
- If using *spanning-tree pathcost method long*
  - 1 Gig link = 20000
  - 2 Gig link = 10000
  - 3 Gig link = 6660
  - 4 Gig link = 5000



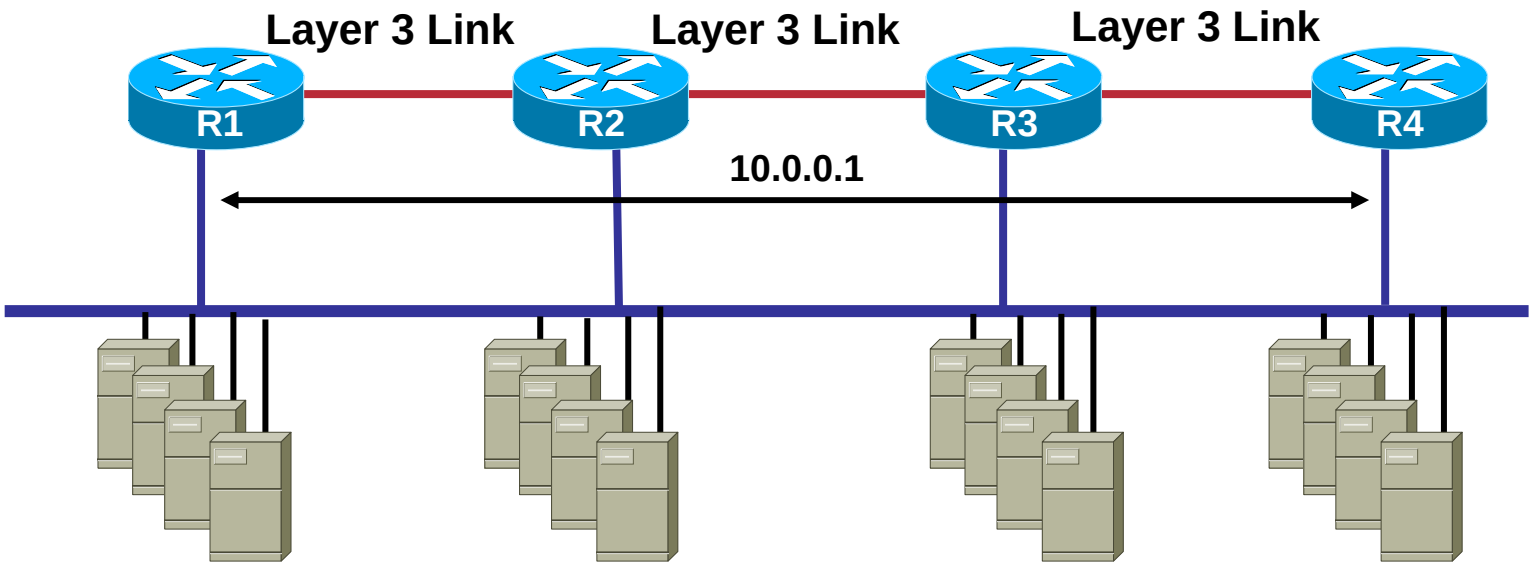
使用千兆以太网捆绑和万兆以太网扩展带宽
GLBP Refresher

AVG/AVF
IP: 10.0.0.251
MAC: MAC-R1
vIP: 10.0.0.1
vMAC: 0007.B400.0101

AVF
IP: 10.0.0.252
MAC: MAC-R2
vIP: 10.0.0.1
vMAC: 0007.B400.0102

AVF
IP: 10.0.0.253
MAC: MAC-R3
vIP: 10.0.0.1
vMAC: 0007.B400.0103

AVF
IP: 10.0.0.254
MAC: MAC-R4
vIP: 10.0.0.1
vMAC: 0007.B400.0103



IP: IP1, IP2, IP3, IP4
GW: 10.0.0.1
ARP: 0007.B400.0101

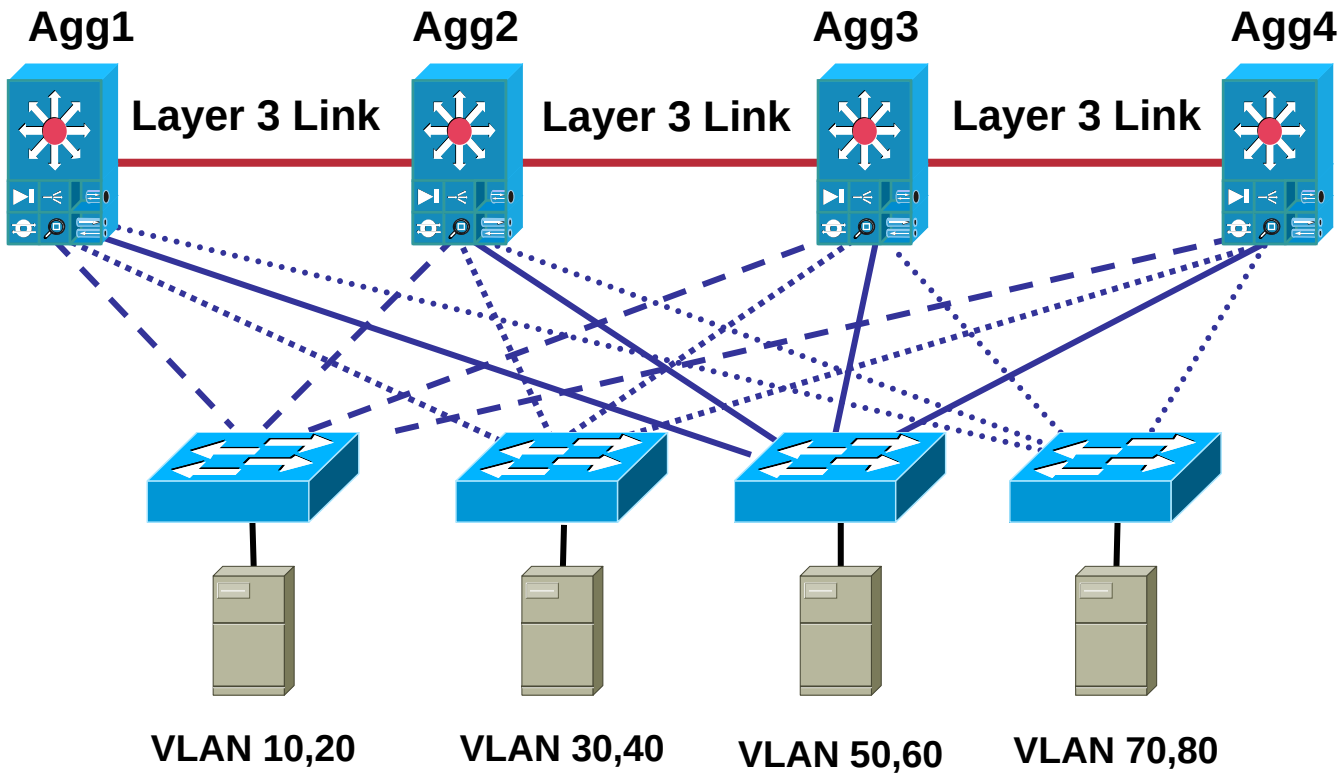
IP: IP5, IP6, IP7, IP8
GW: 10.0.0.1
ARP: 0007.B400.0102

IP: […]
GW: 10.0.0.1
ARP: 0007.B400.0103

IP: […]
GW: 10.0.0.1
ARP: 0007.B400.0104

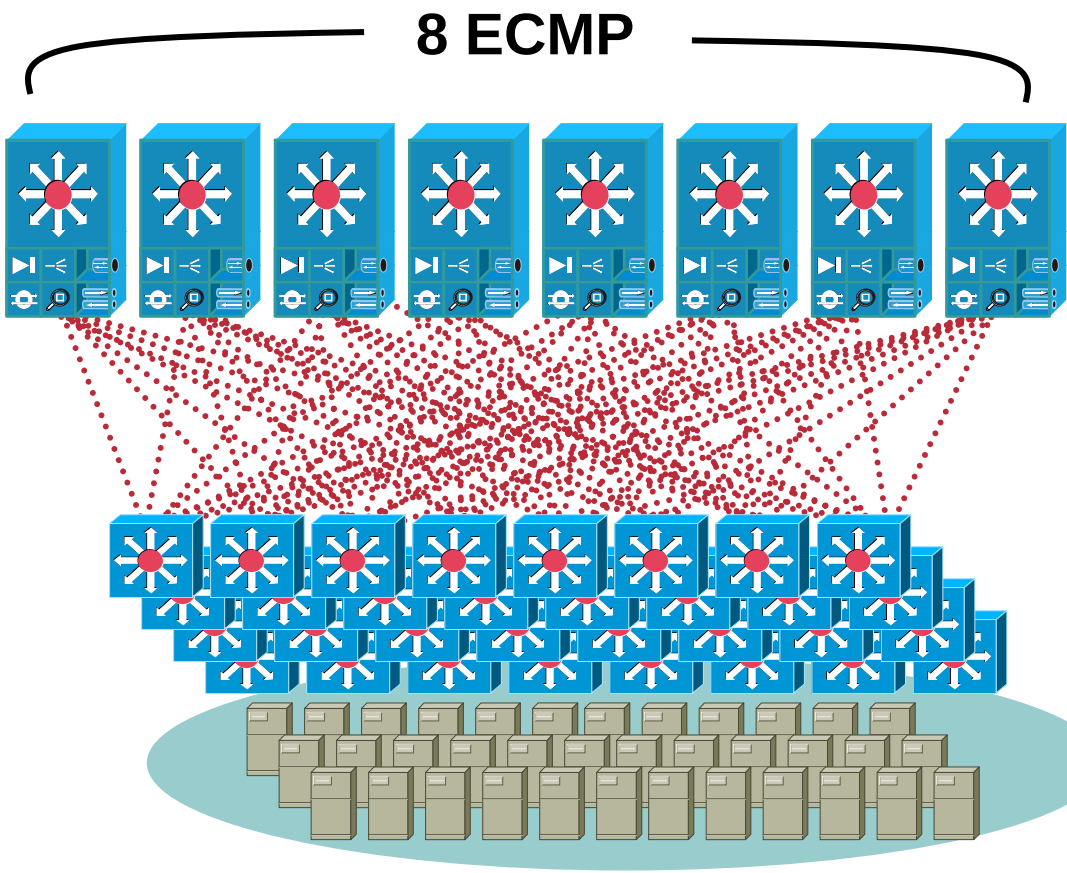
使用千兆以太网捆绑和万兆以太网扩展带宽
Using GLBP to Scale Bandwidth

- **Looped topology may not be desirable: only one uplink active**
- **Use GLBP to distribute default gateway: maximum four gateways**



使用千兆以太网捆绑和万兆以太网扩展带宽
Using Layer 3 ECMP to Scale B/W

- All links are layer 3 links in this picture
- Relies on Equal Cost Multiple Path (ECMP) with DCEF load balancing
- Scales to equal cost routes that the CEF hardware supports (currently eight)
- Permits very low oversubscribed designs
- Popular for large clusters and HPC designs



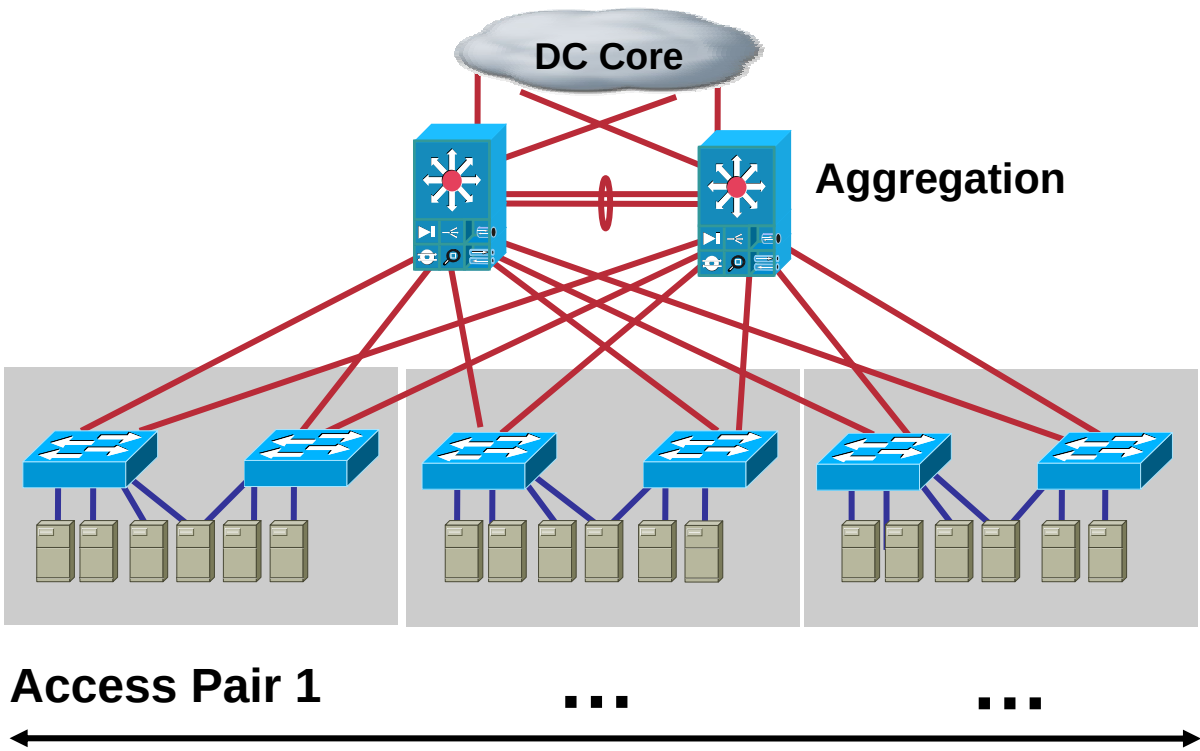
Example Above:
9,216 Servers Attached with GE
3.6:1 Oversubscription
278Mbps per Server
All 6700 dCEF Enabled Modules

使用千兆以太网捆绑和万兆以太网扩展带宽
Connecting Servers with 10GE

- **What is driving 10GE at the server level?**
  - **Storage and clustering requirements**
  - **Server consolidation efforts**
    - **Virtual machine solutions on large SMP's with 802.1Q trunks**
- **Influences access layer design**
  - **Oversubscription ratios are even more important**
- **Influences aggregation layer scaling**
  - **Port Density**

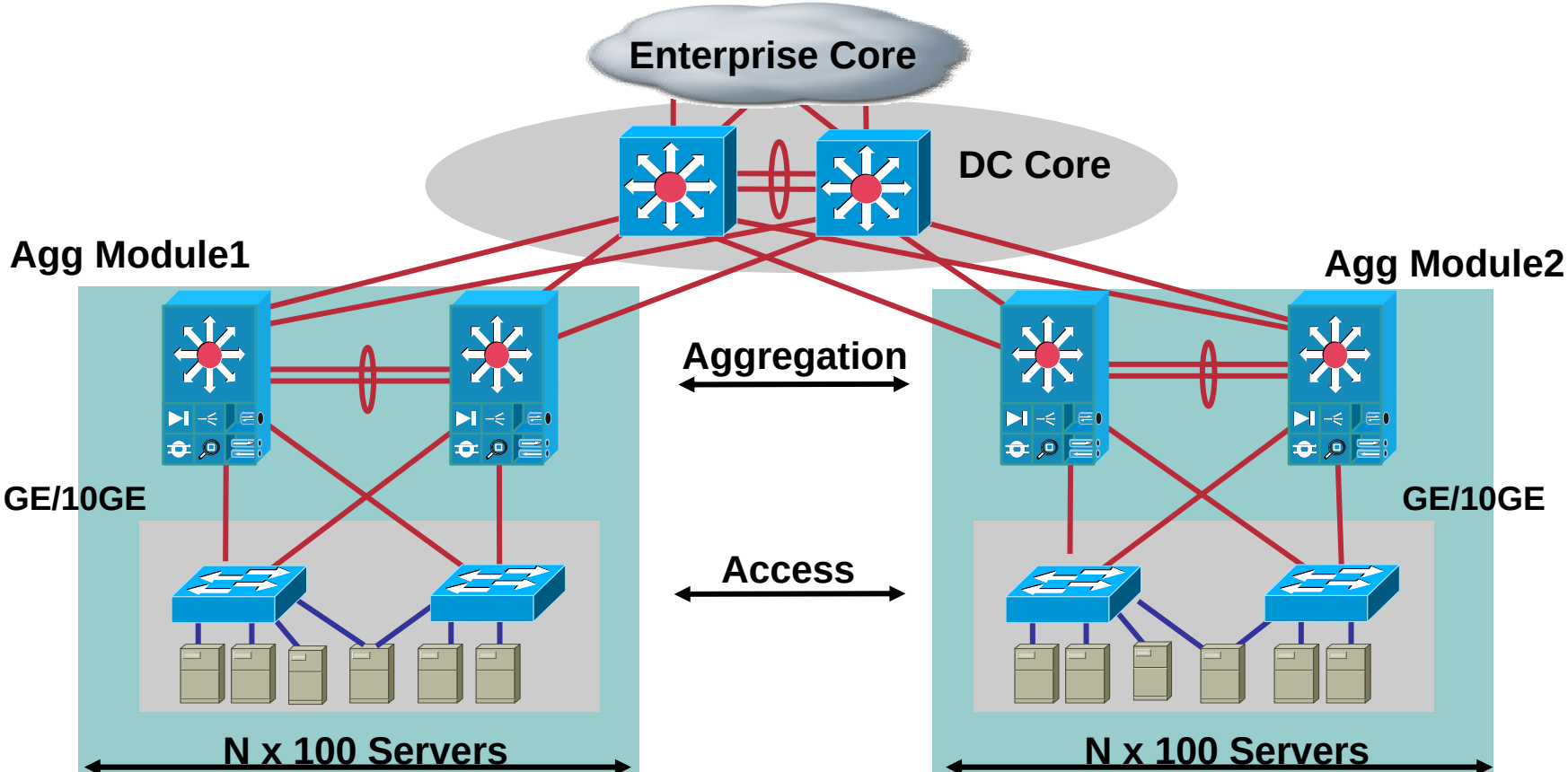
使用千兆以太网捆绑和万兆以太网扩展带宽
Migrating Access Layer Uplinks to 10GE

- 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?



使用千兆以太网捆绑和万兆以太网扩展带宽
Aggregation Modules: Scaling Horizontally

A Data Center Core Provides Interconnectivity Between Multiple Aggregation Modules Which Adds Ports Required to Support 10GE to the Access Layer



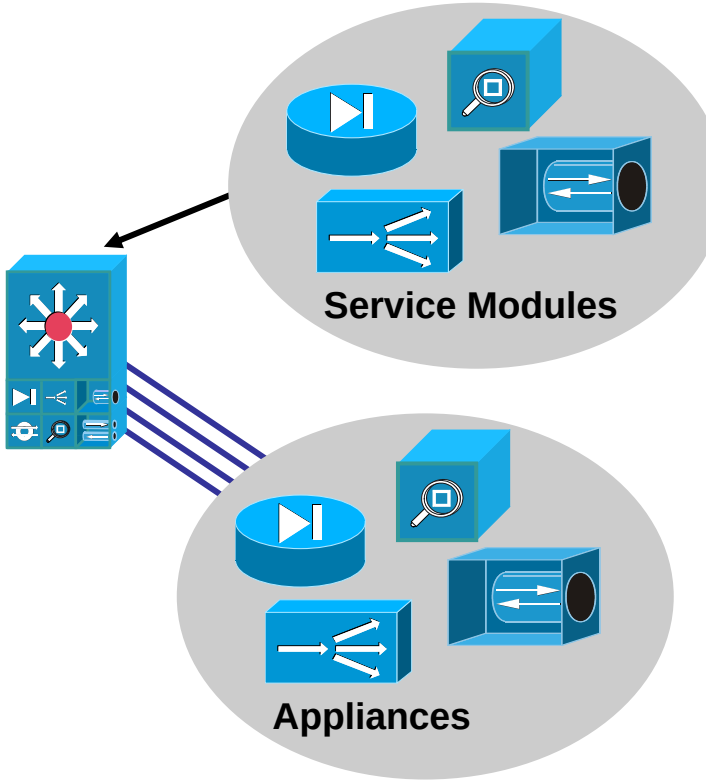
使用千兆以太网捆绑和万兆以太网扩展带宽
Service Layer Switch Introduction

- **Where to deploy DC service modules?**

  CSM, SSL, IDSM, VPNSM, etc…

- **Are all service modules required to be physically placed in the aggregation layer switches?**

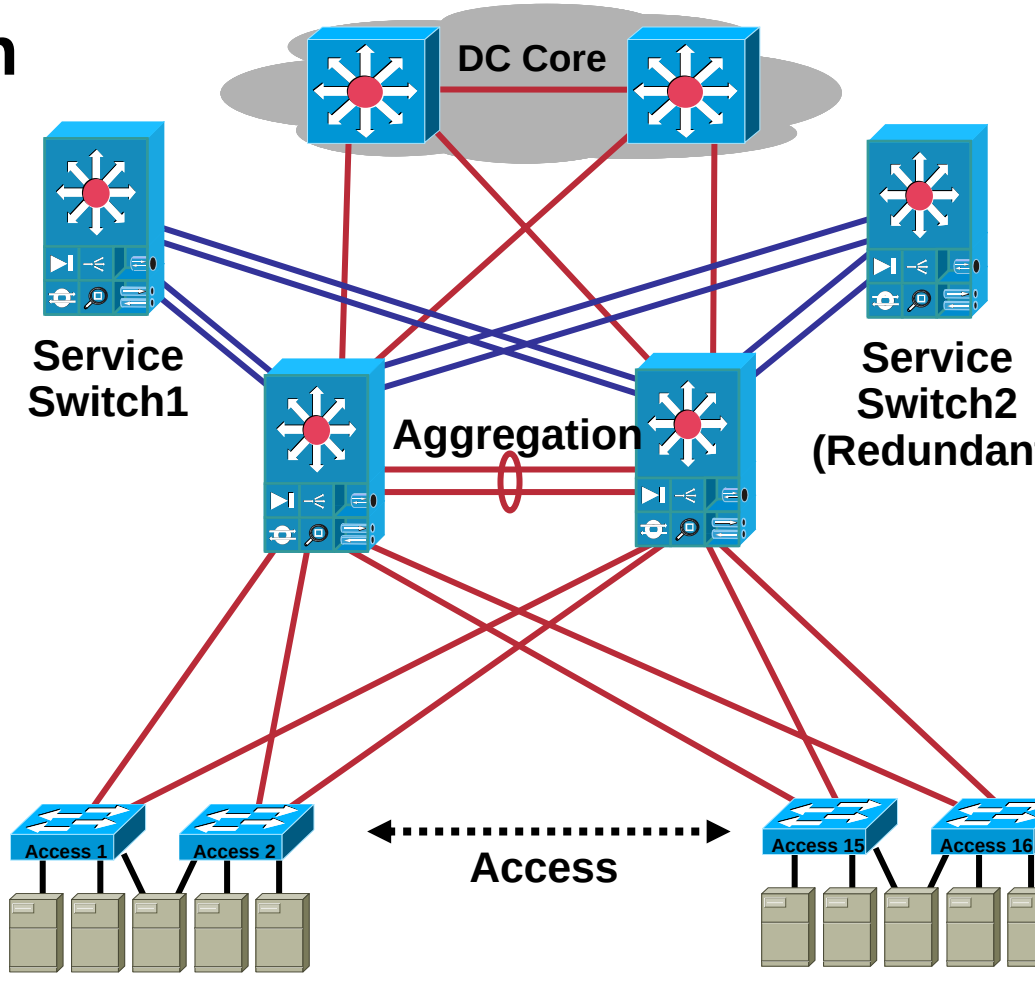
- **Can I move certain service modules to other locations?**



使用千兆以太网捆绑和万兆以太网扩展带宽
Service Layer Switch Introduction

Service Layer Switch

- Move certain services out of aggregation layer
- Particularly works well for CSM modules
- Extend only necessary L2 VLANs to service switches via .1Q trunks
- Opens slots in agg layer for 10GE ports
- Service Layer can utilize GEC uplinks
- Use QOS or separate GE links for FT paths

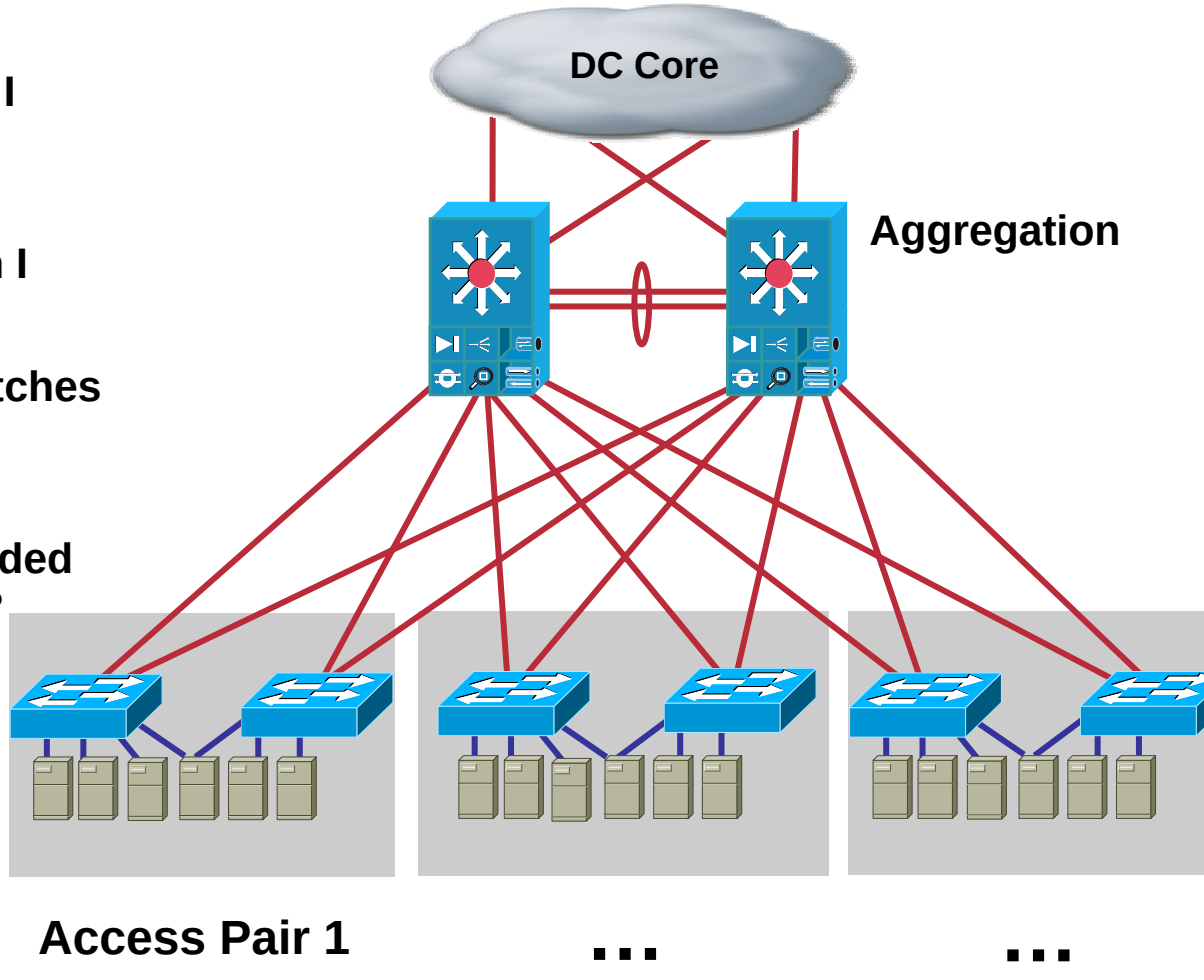


SPANNING TREE 设计和扩展

Spanning Tree 扩展
Common Questions

Which STP Protocol Should Be Used?

- How many VLANs can I support in a single aggregation module?
- How many servers can I support per complex?
- 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 any STP hardware restrictions?



Spanning Tree 扩展
Common Questions (Cont.)

- **Why are these questions important?**
  - Scalability
  - Convergence
  - Throughput/performance
  - Manageability
- **Datacenter L2 designs are getting larger**
  - NIC teaming/dual homing
  - Clustering
  - Applications requiring L2 adjacency
  - Server growth: adoption of blade and 1RU server technology

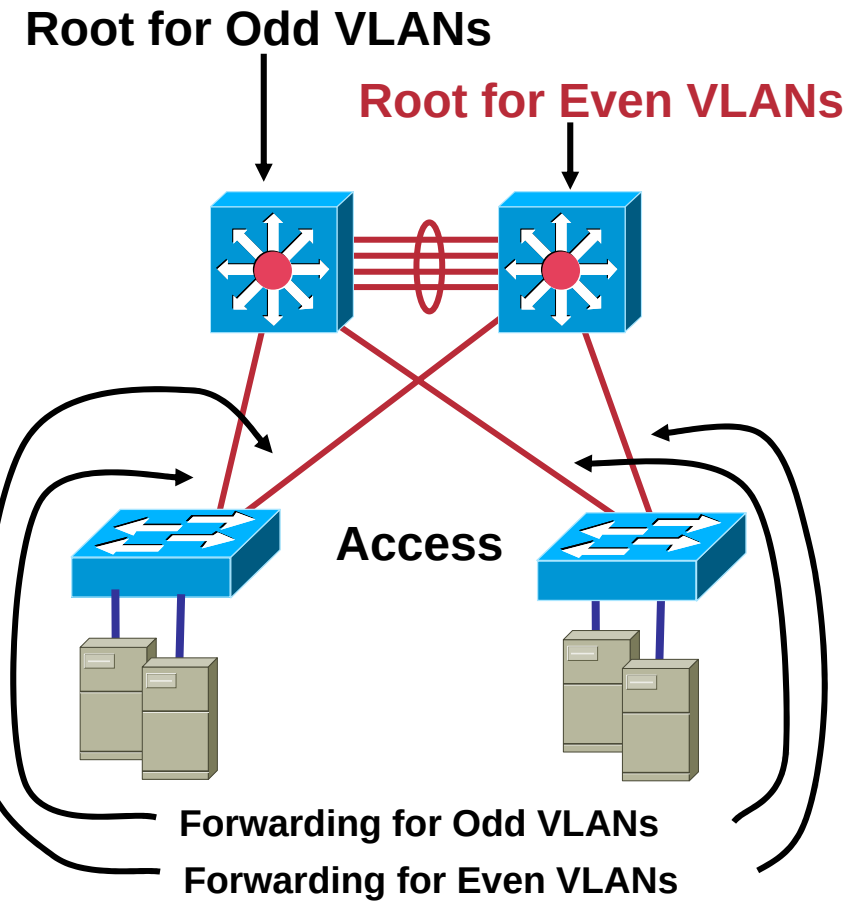
Spanning Tree 扩展
Spanning Tree Protocols Used in the DC

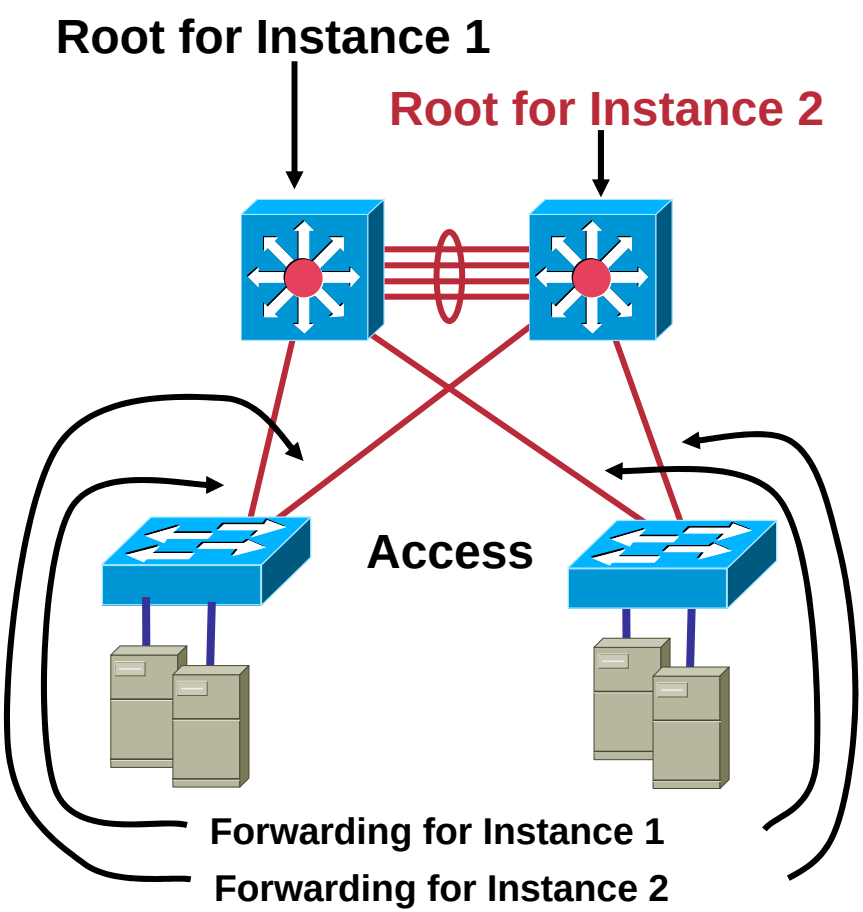
- **Rapid PVST+ (802.1w)**
  - Most common in data center today
  - Scales to large size (~10,000 logical ports)
  - Coupled with UDLD, Loopguard, RootGuard and BPDU Guard, provides a strong-stable L2 design solution
  - Easy to implement, proven, scales
- **MST (802.1s)**
  - Permits very large scale STP implementations (~30,000 logical ports)
  - Not as flexible as Rapid PVST+
  - More common in service providers and ASPs

This Focuses on the Use of Rapid PVST+

Spanning Tree 扩展

Comparing PVST+ and MST





- ↑ Load Distribution on Uplinks
- ↓ If You Have 100 Vlans, the Cpu Handles 100 Topologies

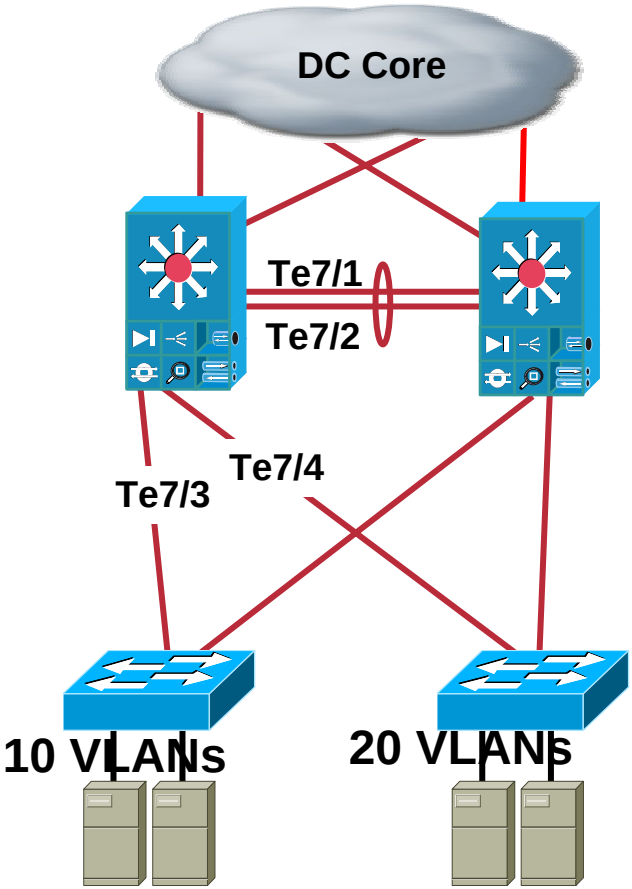
- ↑ If You Have 100 Vlans, the CPU Needs to Maintain Only Two Topologies
- ↓ Less Design Flexibility

Spanning Tree 扩展
Spanning Tree Protocol Scaling

- **Concern: how many access switches/VLANs can be connected to the aggregation layer?**
- **How will the number of VLANs affect convergence?**
- **Be conscious of:**

  Number of total STP active logical interfaces

  Number of virtual ports per LineCard



Spanning Tree 扩展
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	1,800 [2] per Switching Module

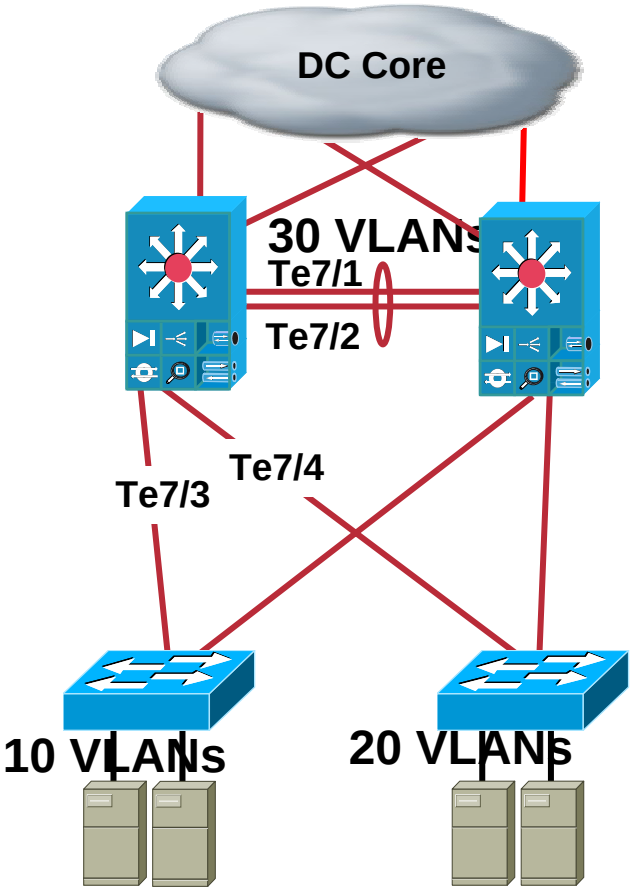
1 CSCed33864 Is Resolved in Release 12.2(17d)SXB and Later Releases

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

Spanning Tree 扩展
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 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 扩展
Spanning Tree Protocol Scaling

Number of Virtual Ports per Line Card=

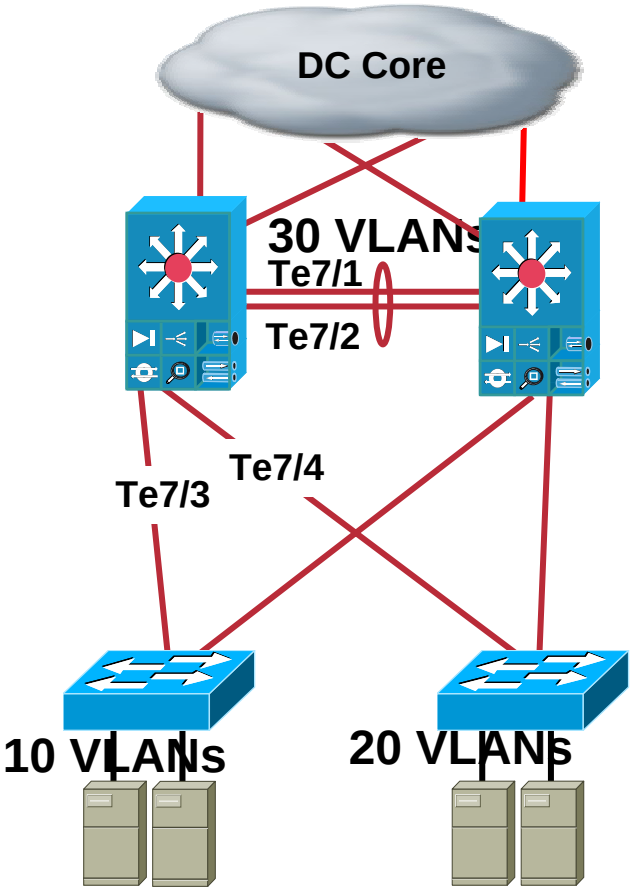
For line card x: sum of all trunks * VLANs * (the number of ports in a port-channel if used)

10 + 20 + (30*2)

=90 Virtual Port's on line card 7

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

EtherChannel

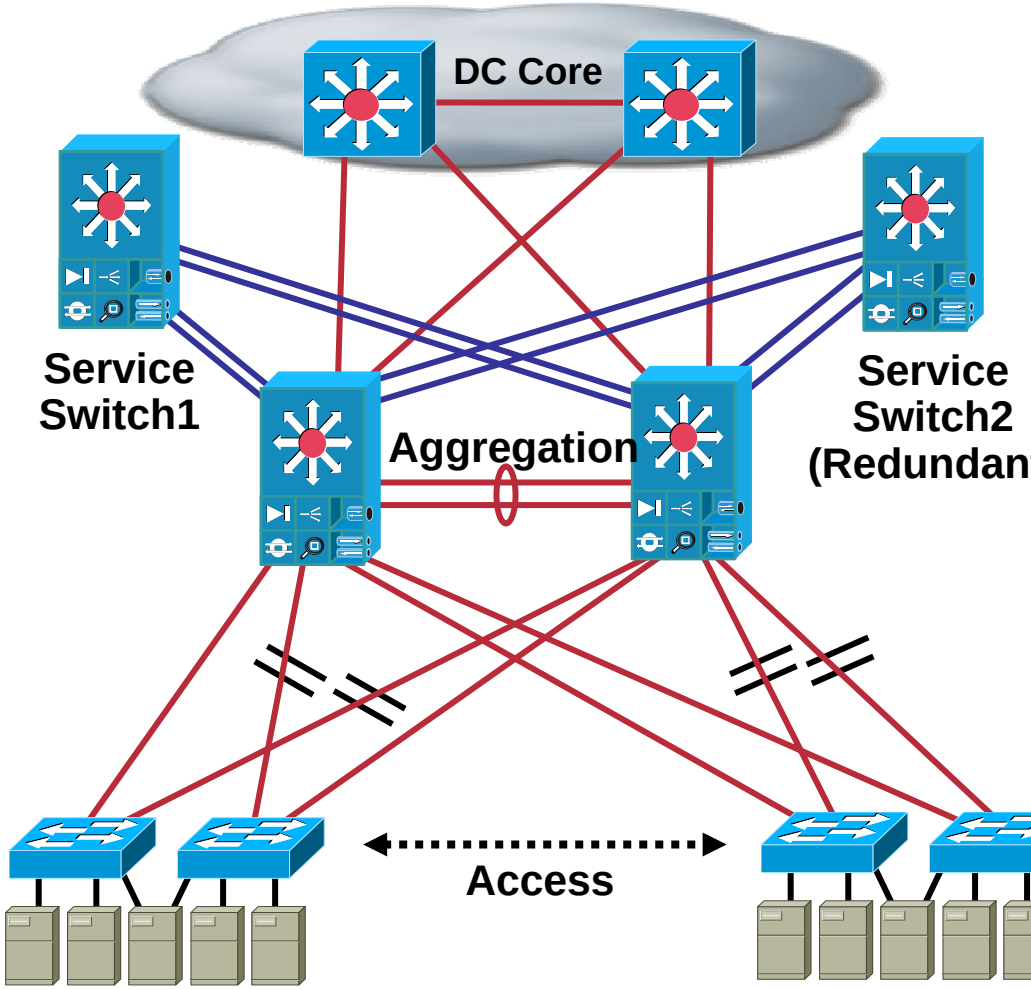


NOTE: VP's are calculated per port in channel groups

Spanning Tree 扩展
Design Guidelines

General Guideline for Rapid PVST+ Implementation Using Modular Access:

- Maximum eight access switches per agg module
- Maximum 100 VLANs per modular access switch
- Add aggregation modules to scale beyond this
- Does not consider HSRP/GLBP scalability

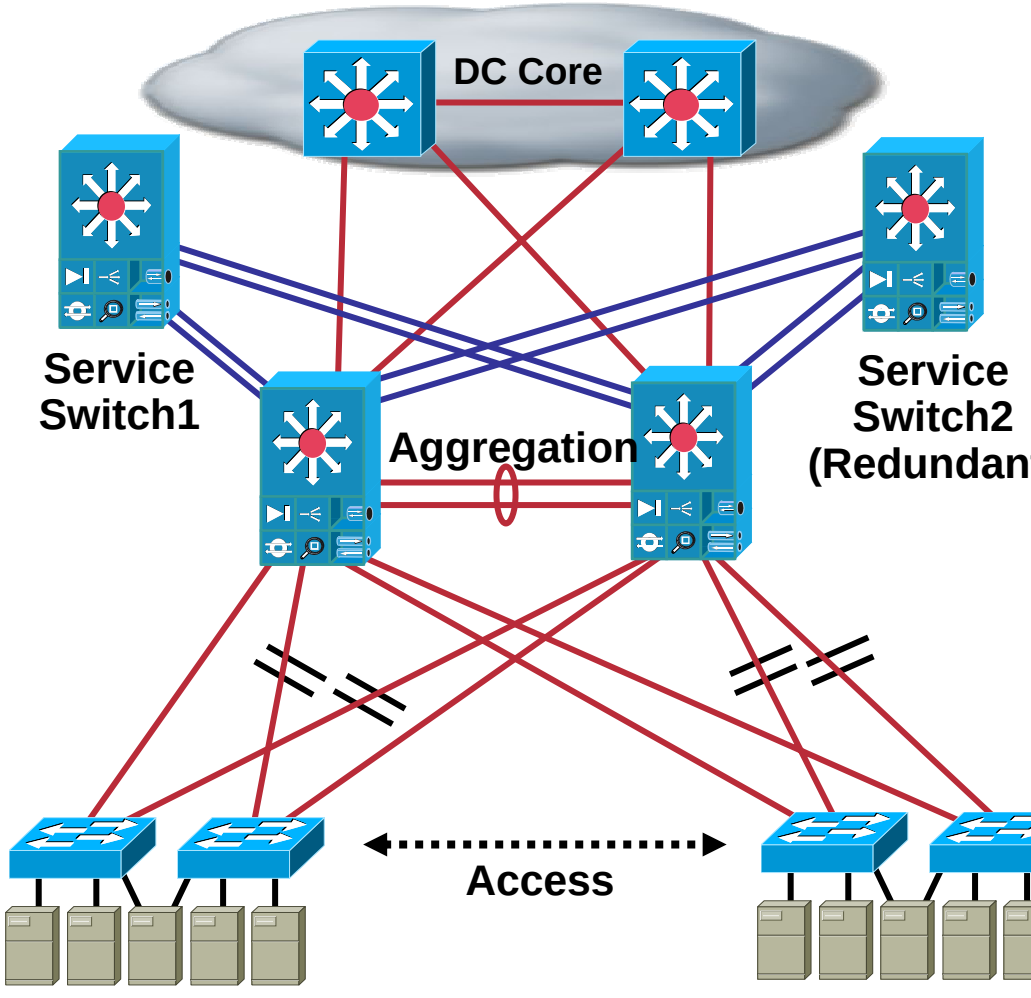


Eight Access Switches per Agg Module

Spanning Tree 扩展
Design Guidelines

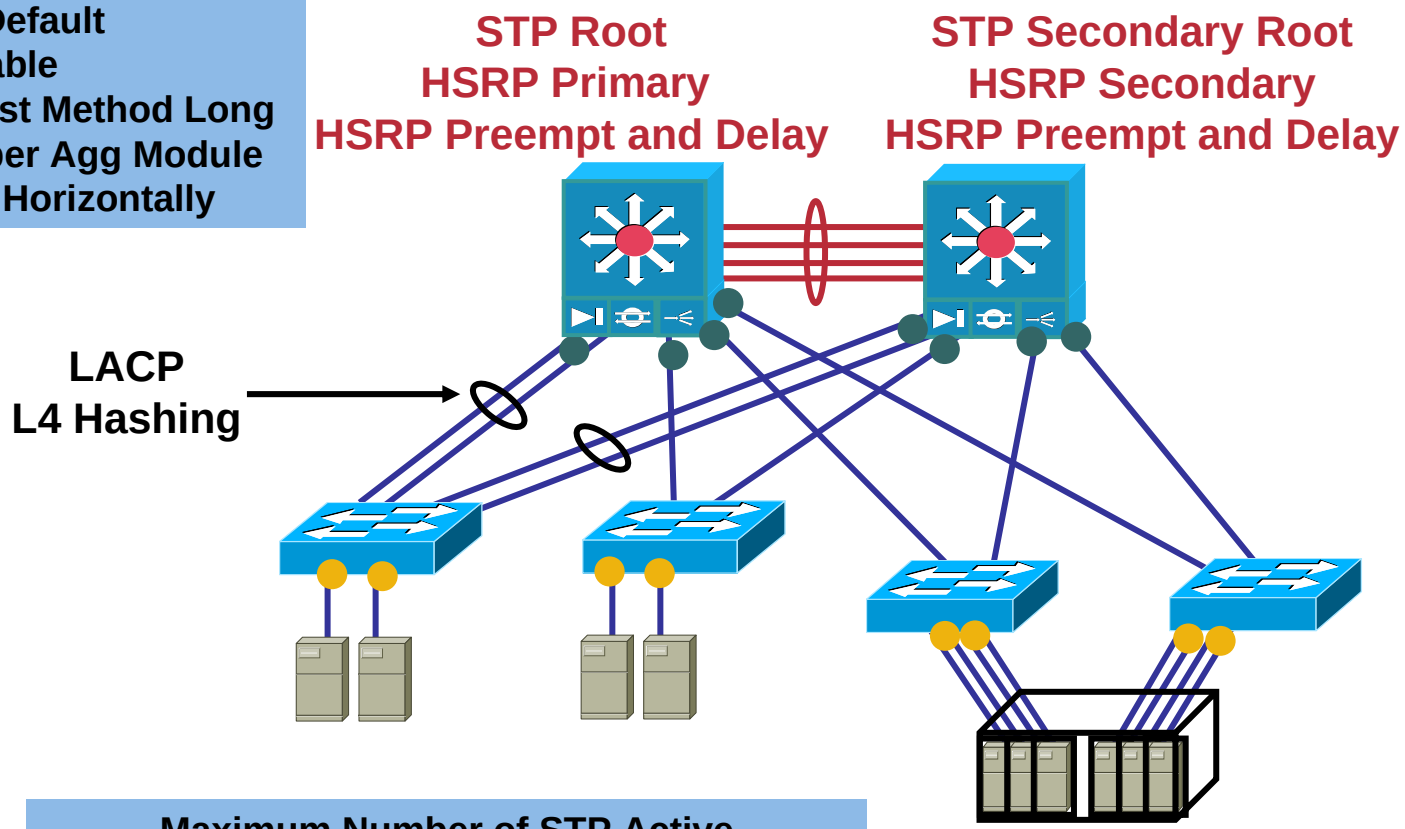
General Guideline for Rapid PVST+ Implementation Using 1RU Access:

- **Maximum 30 VLANs per access switch**
- **Maximum 20 access switches**
- **Leaves more buffer space as VLAN extension is more likely**
- **Does not consider HSRP/GLBP scalability**



Spanning Tree 扩展
Spanning Tree Best Practices Summary

Rapid PVST+
Loopguard Default
UDLD Enable
Spanning Tree Pathcost Method Long
Limit Access Nodes per Agg Module
Scale Aggregation Horizontally



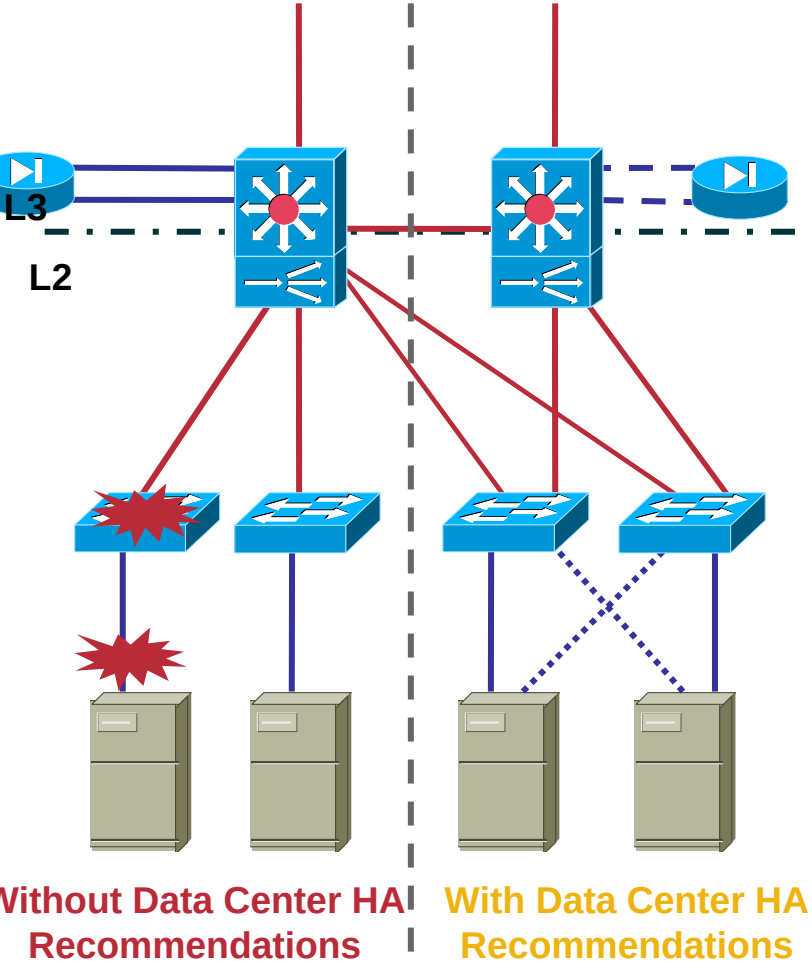
- Rootguard
- Portfast+ BPDUguard

Maximum Number of STP Active
Logical Ports
Virtual Ports Per Linecard Is 1800

数据中心高可用性

数据中心高可用性
Server High Availability

Common Points of Failure

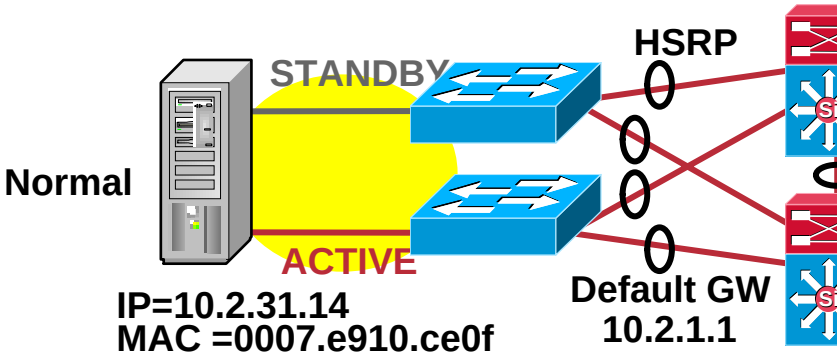


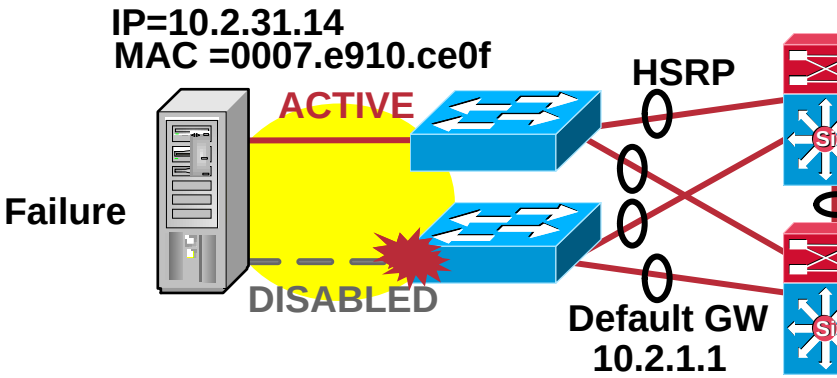
1. Server network adapter
2. Port on a multi-port server adapter
3. Network media (server access)
4. Network media (uplink)
5. Access switch port
6. Access switch module
7. Access switch

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

数据中心高可用性
Fault Tolerance Modes

- **ACTIVE/STANDBY**
  - Adapter Fault Tolerance (AFT)
  - Switch Fault Tolerance (SFT)
  - Network Fault Tolerance (NFT)
- **Single IP address and MAC address move from active to standby adapter in the event of network failure**
  - Failover occurs in less than one second with no TCP session loss





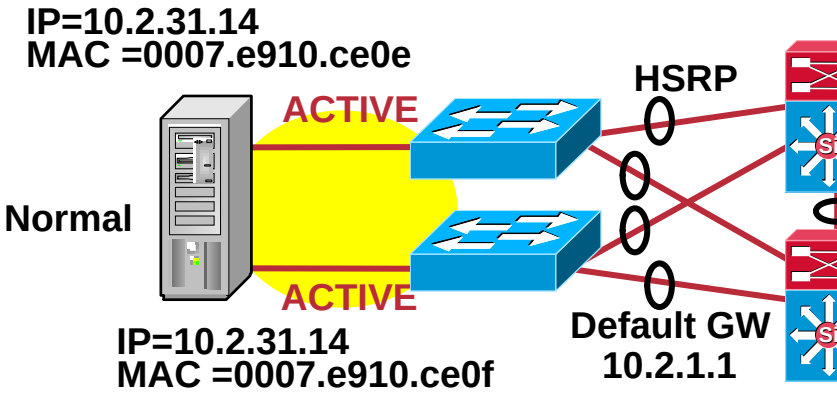
数据中心高可用性
Fault Tolerance Modes

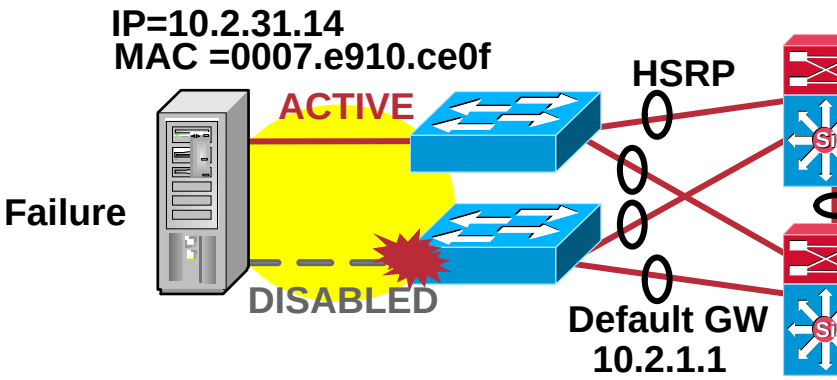
- **ACTIVE/ACTIVE**

  Adaptive load balancing

  (IP and IPX traffic only)

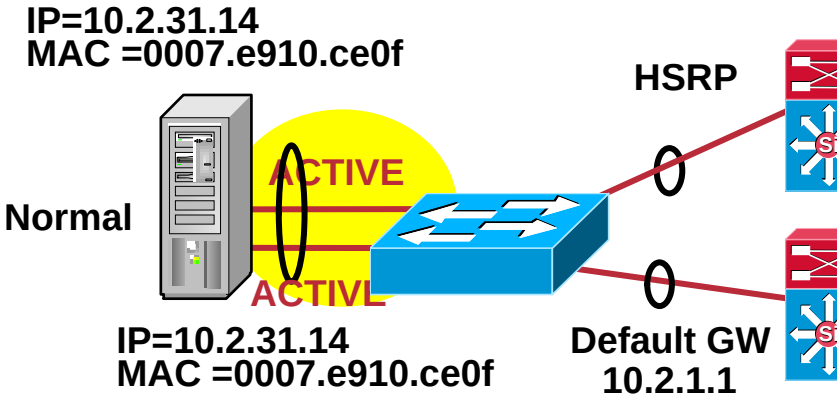
- **One port receives, all ports transmit based on source dest IP; incorporates fault tolerance**
- **Multiple MAC addresses exist for the same server IP address**
- **Failover occurs in less than one second with no TCP session loss**

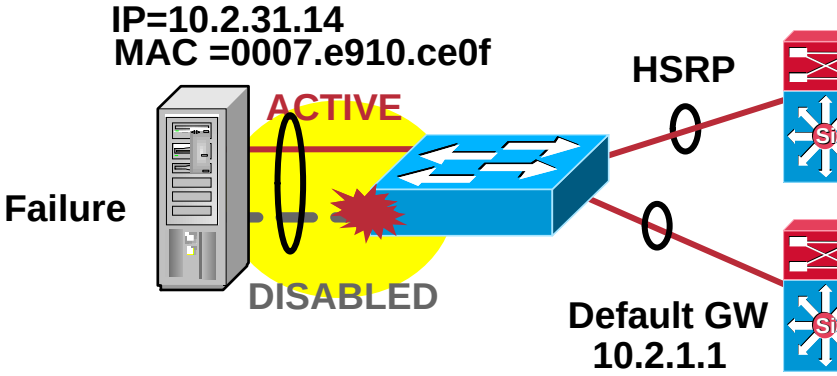




数据中心高可用性
Link Aggregation Modes

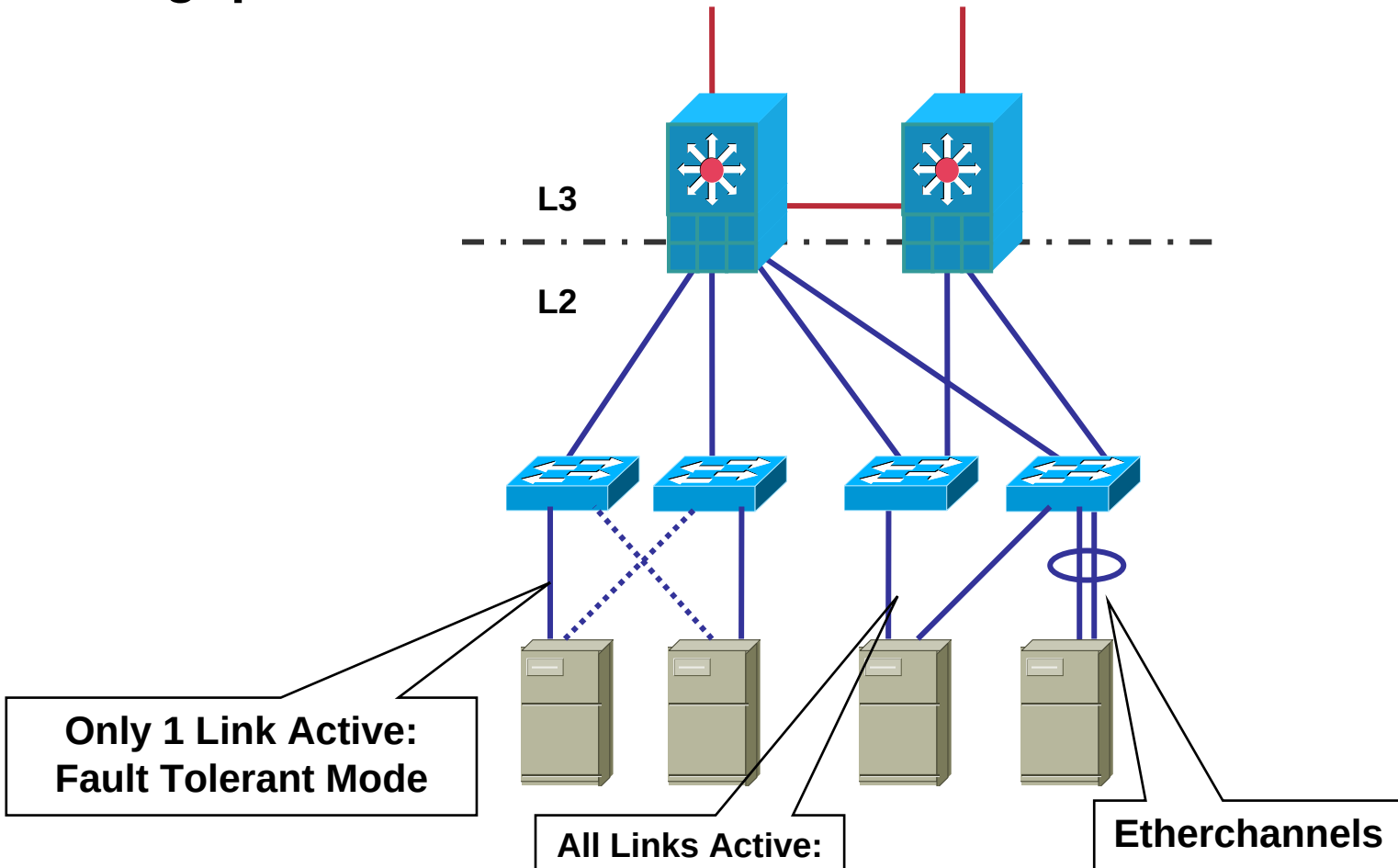
- **ACTIVE/ACTIVE**
  - EtherChannel
  - Fast EtherChannel
  - Gigabit EtherChannel
  - IEEE 802.3ad (LACP)
- **Multiple physical links operate as one logical link; incorporates fault tolerance and load balancing based on source Dest IP**
- **Failover occurs in less than one second with no TCP session loss**





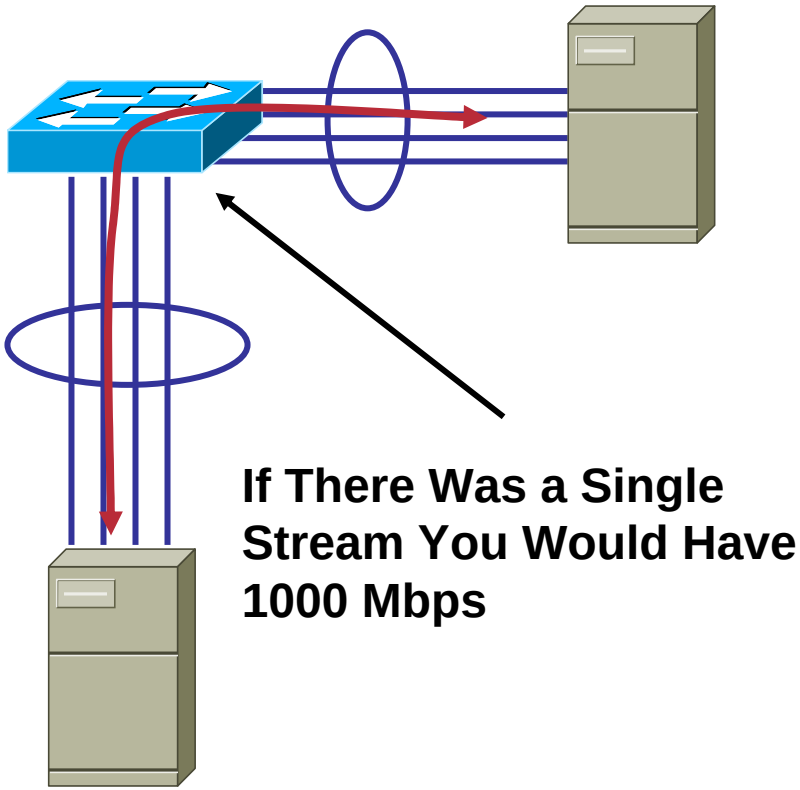
数据中心高可用性
Server Attachment: Multiple NICs

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



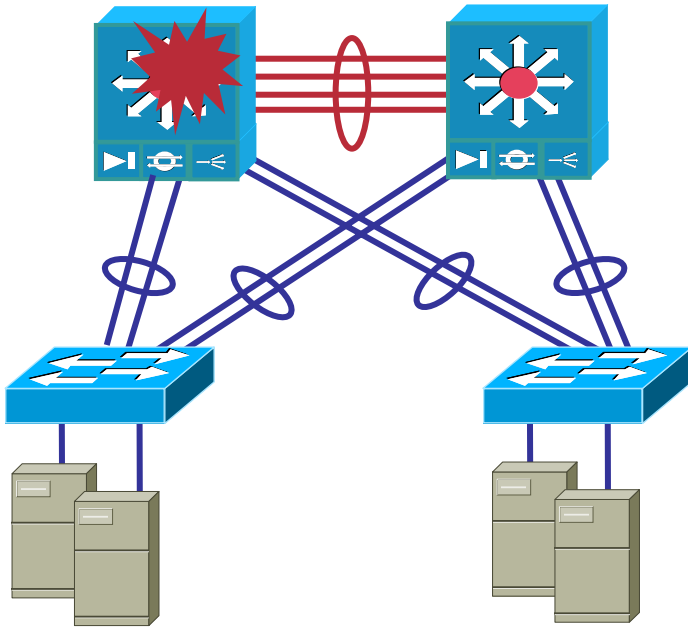
数据中心高可用性
EtherChannel

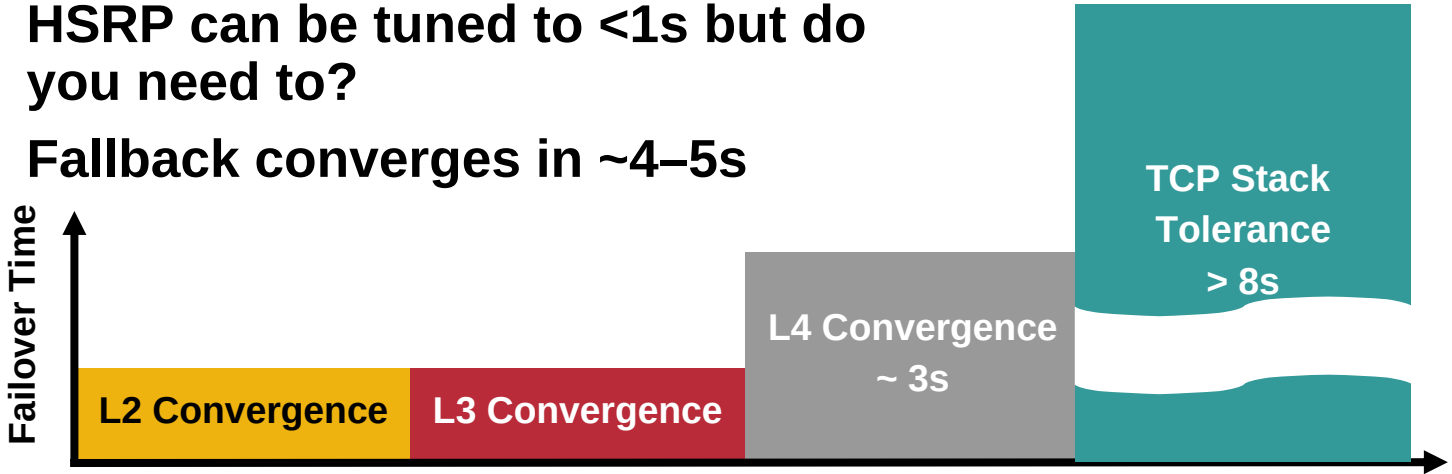
- EtherChannel load distribution works for many-to-many communications
- EtherChannels are ineffective for 1-to-1 communication
- Backup software creates multiple streams
- Using port aggregation of Ethernet links provides benefits for the aggregate traffic not for individual stream



数据中心高可用性
Failover Times

- The overall failover time is the combination of convergence at L2, L3, L4
- Stateful devices replicate connection information and typically failover within 3s
- EtherChannels << 1s
- STP converges in <1
- HSRP can be tuned to <1s but do you need to?
- Fallback converges in ~4–5s

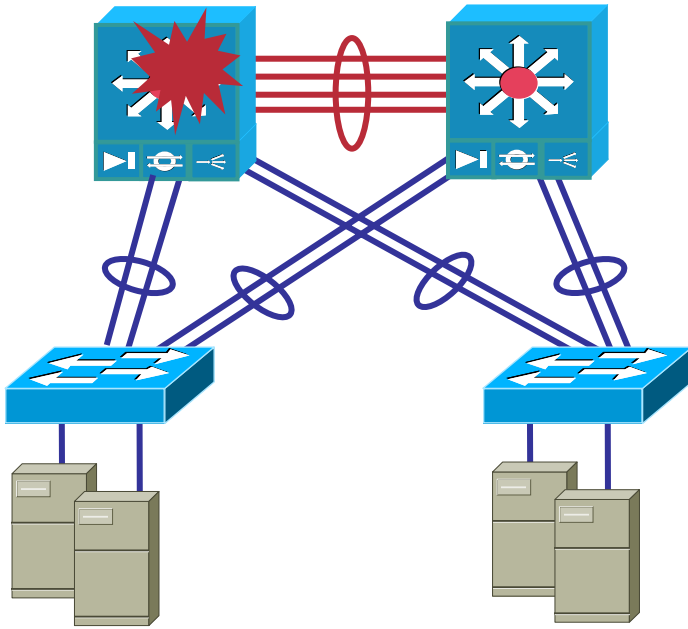


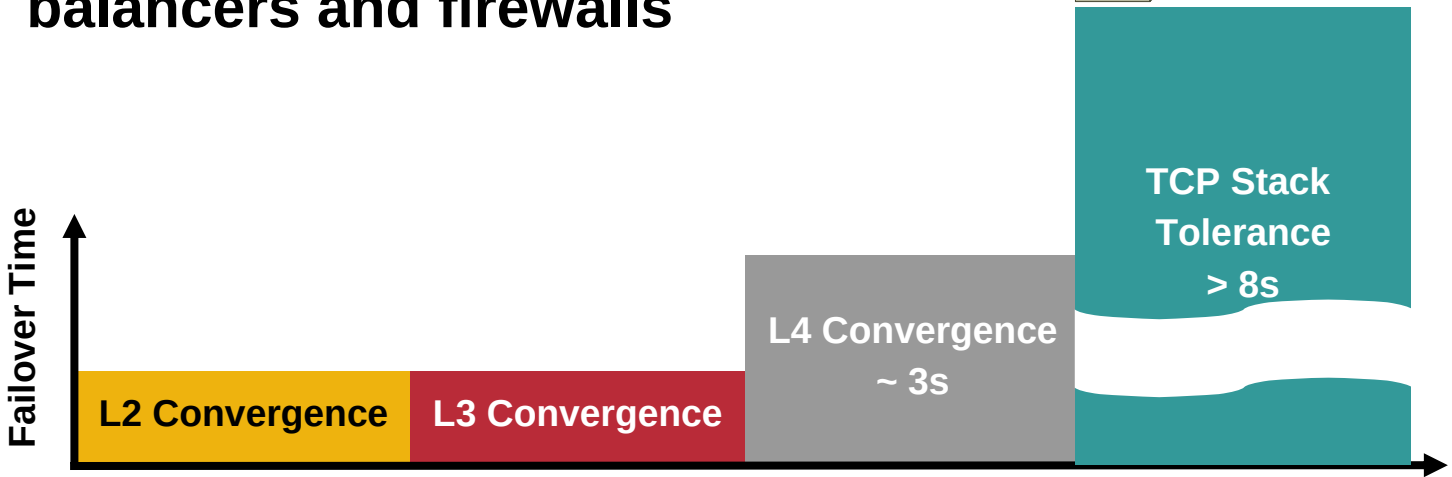


数据中心高可用性
Supervisor Failover in the Data Center

If the Supervisor in One Aggregation Switch Fails, the Failover Time Accounts For:

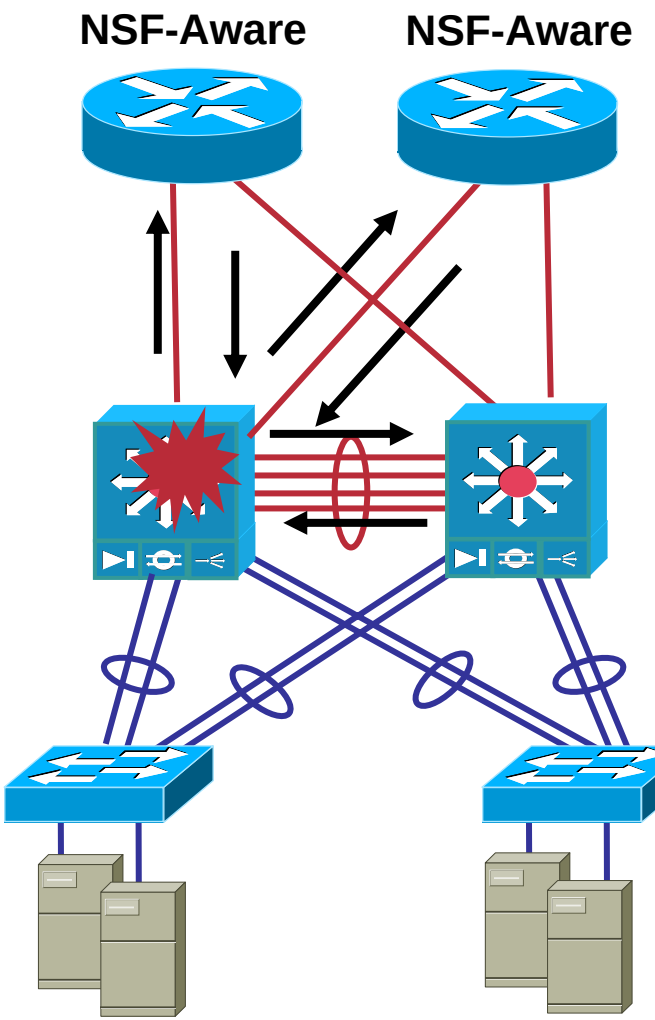
- **The failure of the spanning tree root**
- **The HSRP primary**
- **The recovery time of load balancers and firewalls**





数据中心高可用性
NSF/SSO

- NSF/SSO is a supervisor redundancy mechanism for intrachassis failover
- SSO synchronizes layer 2 protocol state, hardware L2/L3 tables (MAC, FIB, adjacency table), ACL and QoS tables
- SSO synchronizes 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



总结

总结

- **Understand the applications, clustering and server HA requirements to effectively design the DC access layer**
- **Communicate well with sys-admin staff on areas of STP, VLAN extension, cabling and 10GE requirements**
- **Understand the implications of 1RU and modular access layer switches**
- **NIC teaming can greatly enhance the HA architecture**

Q and A


CISCO SYSTEMS