Cisco Switching Portfolio

- **Features, Scalability, Longevity**
  - Datacenter Access
  - Distribution/Core
  - Wiring Closet

- **Number of Employees/Density**
  - Small
  - Medium-sized
  - Large

- **Catalyst 29xx**
  - Lite

- **Catalyst 3560**
  - E

- **Catalyst 3750**
  - E

- **Catalyst 4500 E-Series**

- **Catalyst 4900 Series**

- **Catalyst 6500**

- **Nexus 5000**

- **Nexus 7000**

- **Blade Switches**
Next Generation Campus Design
Unified Communications Evolution

- IP Telephony is now a mainstream technology
- Ongoing evolution to the full spectrum of Video and Collaboration technologies
- High Definition Executive Communication Applications require stringent Service-Level Agreement (SLA)
  - Reliable Service – Highly Available Infrastructure
  - Application Service Management - QoS
Next Generation Campus Design
Unified Communications Evolution

- Availability Requirements for UC are more than just five 9’s
- Also need to consider the subjective impact to real time communications

![Bar Chart]

- Minimal Impact to Video, none to Voice
- Minimal Impact to Voice
- User Hangs Up
- Phone Resets*

* Phone to reset time depends on the signaling protocol, SCCP or SIP, and call state; active, ringing, …
Cisco’s Campus Architecture
Hierarchical, Modular and Resilient

- Offers hierarchy for scalability
- Modular building blocks—Easy to grow, understand, and troubleshoot
- Predictable traffic patterns under normal and failure conditions
- Small fault domains to isolate problems
- Promotes load balancing and failover
- Can be applied to all campus designs; Multi-Layer L2/L3 Routed Access designs

Diagram:
- Access
- Distribution
- Core
- Building Blocks
- Redundant Switches
- Redundant L3 Links
- Redundant Supervisor
- Layer 2 or Layer 3
- Layer 3 Equal Cost Link’s
- Data Center
- Building Blocks
- Internet
- WAN
AGENDA

- Systems Level Resiliency
- Network Level Resiliency – Routing
- Campus Core and Foundation Services
- Emerging Campus Design
  - Routed Access
  - Virtual Switch Campus Design
System Level Resiliency
Comprehensive Physical Redundancy

• Nexus 7000, Catalyst 6500 and 4500 highly redundant Modular systems
  - Redundant hot swappable Supervisors
  - Redundant hot swappable Power Supplies
  - N+1 redundant fans with hot swappable fan trays
  - Hot swappable line cards
  - Passive data backplane
  - Redundant system clock modules

• Catalyst 3750/3750E StackwisePlus* technology
  - 1:N Master redundancy
  - Hot swappable stack members
  - Hot swappable Power Supplies*
Graceful Restart
Non-Stop Forwarding/Stateful Switch-Over

- NSF/SSO is a supervisor redundancy mechanism for intra-chassis supervisor 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
- Non-Stop Forwarding (NSF) provides the capability for the routing protocols to gracefully restart after an SSO fail-over
  - The newly active redundant supervisor continues forwarding traffic using the synchronized HW forwarding tables
  - The NSF capable Routing Protocol requests a graceful neighbor start
  - Routing neighbors reform with no loss of traffic
- Aggressive RP timers may not work in NSF/SSO environment

No Route Flaps During Recovery
Nexus 7000
Service Restart

- Stateful Restart with PSS
  - Checkpoints states to PSS
  - Recover states from PSS upon restart
- Stateful Restart with GR
  - Fresh start without traces from former instantiation.
  - Graceful Restart (NSF) for L3 Protocols
- Supervisor Switchover
- Non-disruptive In Service Software Upgrad
Nexus 7000
Stateful Fault Recovery Using PSS

- Multiple Service Instances
- Independent memory-protected re-startable processes
- Services checkpoint their runtime state to the PSS for recovery in the event of a failure
  - Layer2 Services
  - Layer3 Services
- Neighbors never see event occur

If a fault occurs in a process:
- “Sysmgr” determines best recovery action
  (restart process, switchover to redundant supervisor)
- Process restarts with no impact on data plane
State checkpointing (PSS) allows instant, stateful process recovery
Nexus 7000
Stateful Fault Recovery Using Graceful Restart [1]
Nexus 7000
Nexus 7000
Design Considerations for NSF/SSO
Single Points of Connectivity = SSO and NSF

- Access switch is the single point of failure even in HA campus design
- Business requirement driving new requirements
  - Unified Communications integration requires high uptime
  - Critical locations require continuous connectivity (eg. Hospital, Call Center)
  - Must protect for both planned and unplanned outages
- Supervisor disruption is most common cause of access switch outages
- Network outage until physical replacement or reload vs. 1 to 3 seconds
In Service Software Upgrade process
Catalyst 4500 and 6500

- ISSU upgrade is a 4 step process
- Possible to rollback (abort) up until you complete the 4\textsuperscript{th} step (commit to final state)
- Leverages NSF/SSO to implement supervisor transition
- Requires that the two images are compatible for upgrade/downgrade processing

* The [issu acceptversion] is an optional step during the ISSU procedure
In Service Software Upgrade process
Catalyst 6500 VSS System

1. Copy the new software image to Active and Standby supervisor flash memory

Old software image is represented with green color and new software image is represented with peach color

2. ISSU loadversion

3. ISSU runversion

4. ISSU acceptversion

5. ISSU commitversion
n7k# install all kickstart bootdisk:4.1-kickstart system bootdisk:4.1-system
n7k#
AGENDA

- Systems Level Resiliency
- Network Level Resiliency – Routing
- Campus Core and Foundation Services
- Emerging Campus Design
  - Routed Access
  - Virtual Switch Campus Design
Network Level Resiliency
Routing Convergence Improvements

• Network convergence is the time needed for traffic to be rerouted to the alternative path after the network event
• Network convergence requires all affected routers to process the event and update the appropriate data structures used for forwarding
• Network Convergence is the time required to:
  1 - Detect the event – Link Failure, BFD, IP Dampening
  2 - Propagate the event – LSA Backoff Timers
  3 - Process the event – SPF Backoff Timers
  4 - Update the routing table/FIB
• ISIS and OSPF have the same Fast Convergence behavior
Redundancy and Protocol Interaction
Link Neighbor Failure Detection

• Indirect link failures take time to detect
• With no direct HW notification of link or node loss, convergence times are dependent on Spanning Tree BPDUs or Routing Protocol Hellos
• Hardware detection and recovery is both faster and more deterministic
• Use point-to-point routed links in Campus Core!
Improving Indirect Layer 3 Neighbour Failure Detection

- EIGRP, OSPF, IS-IS, mBGP all have native hello/dead mechanisms

  ```
  !Send OSPF Hellos every 250ms (1sec/4)
  interface gigabitethernet 4/1
dampening
  ip address 10.122.0.26 255.255.255.254
  ip ospf dead-interval minimal hello-multiplier 4
  ip ospf network point-to-point
  ```

- Bidirectional Forwarding Detection (BFD)* provides a lightweight protocol independent mechanism

  ```
  !Send BFD Hellos every 100ms
  interface gigabitethernet 4/1
dampening
  ip address 10.122.0.26 255.255.255.254
  bfd interval 100 min_rx 100 multiplier 3
  bfd neighbor 10.122.0.27
  router eigrp 100
  bfd interface gigabitethernet4/1
  ```

- With point-to-point routed links, the costs of sub-second timers (processing load, complexity, ..) may outweigh the benefits.

*Verify Cisco IOS Release Availability, ESE does not yet have specific configuration guidance
**OSPF Design Rules for HA Campus**

**LSA/SPF Exponential Back-Off Throttle Mechanism**

- **Sub-second timers without risk**
  1. spf-start or initial hold timer controls how long to wait prior to starting the SPF calculation
  2. If a new topology change event is received during the hold interval, the SPF calculation is delayed until the hold interval expires and the hold interval is temporarily doubled
  3. The hold interval can grow until the maximum period configured is reached
  4. After the expiration of any hold interval, the timer is reset

```
router ospf 100
  timers throttle spf <spf-start> <spf-hold> <spf-max-wait>
  timers throttle lsa all <lsa-start> <lsa-hold> <lsa-max-wait>
  timers lsa arrival <lsa-hold>
```
Stable Convergence
IP Event Dampening

- Prevents routing protocol churn caused by constant interface state changes
- Dampening is applied on a system: nothing is exchanged between routing protocols
  - Static routing, RIP, EIGRP, OSPF, IS-IS, BGP
  - In addition, it supports HSRP and CLNS routing
  - Applies on physical interfaces and can’t be applied on subinterfaces individually

```yaml
interface GigabitEthernet1/1
  dampening
  ip address 10.120.0.205 255.255.255.254
```
Multicast Deployment – Best Practices

- Use IGMP Snooping capable hardware in the access
- Multicast Subsecond Convergence
  - Join/prune aggregation
  - PIM HELLO option
  - Triggered RPF
- Use PIM sparse mode
  - Enable PIM on ALL interfaces
  - Enable PIM sparse mode on routing nodes (Core, Distribution, and possibly Access)
  - Use Anycast RP & MSDP for RP redundancy and fast convergence
  - There are other combinations of RP redundancy, RP assignment and others options.
- Use PIM-SSM
  Eliminates need for RP
  Eliminate need for MSDP
  Helps prevent unknown sources
Multicast RP Engineering
Anycast RP Configuration

interface loopback 0
  ip address 10.0.0.2 255.255.255.255

interface loopback 1
  ip address 10.1.1.1 255.255.255.255
  ip msdp peer 10.0.0.3 connect-source loopback0
  ip msdp originator-id loopback 0

Interface loopback 0
  ip address 10.0.0.3 255.255.255.255

Interface loopback 1
  ip address 10.1.1.1 255.255.255.255
  ip msdp peer 10.0.0.2 connect-source loopback0
  ip msdp originator-id loopback 0

ip pim rp-address 10.1.1.1

For Your Reference
Moving to PIM Source Specific Mode

Receiver learns of source, group/port
Receiver sends IGMPv3 (S,G) Join
First-hop sends PIM (S,G) Join directly toward Source

(S, G) Join
IGMPv3 (S, G) Join

Out-of-band source directory, example: web server
Moving to PIM Source Specific Mode

Result: Shortest path tree rooted at the source, with no shared tree.
SSM Mapping

- Use an external or internal database for Source to Group mapping
- Allows only for one source per Group
- Router maps group to a single source
- Uses either DNS or static internal database
- DNS method allows content providers to provide the mapping independent from network operators
SSM DNS Mapping – Configuration

```
ip igmp ssm-map enable
ip igmp ssm-map query dns

! ip pim ssm range SSM-GROUP
! ip access-list standard SSM-GROUP
  permit 239.0.0.0 0.255.255.255
!
ip name-server 10.151.1.103
ip domain multicast ssm.cisco.fr
ip domain-name cisco.fr
```

**Enabling SSM Mapping**

**Enabling DNS Mapping**

**Specific DNS Server zone for SSM**

**DNS zone for SSM**

```
cat-3#sh ip igmp ssm-mapping
SSM Mapping : Enabled
DNS Lookup : Enabled
Mcast domain : ssm.cisco.fr
Name servers : 10.151.1.103

cat-3#
cat-3#sh ip igmp ssm-mapping 239.1.1.2
Group address: 239.1.1.2
Database : DNS
DNS name : 2.1.1.239.ssm.cisco.fr
Expire time : 604623026
Source list : 10.151.1.102

cat-3#
cat-3#
```

```
1.1.1.239   IN    A     10.151.1.104
2.1.1.239   IN    A     10.151.1.102
```
AGENDA

- Systems Level Resiliency
- Network Level Resiliency – Routing
  - Campus Core and Foundation Services
- Emerging Campus Design
  - Routed Access
  - Virtual Switch Campus Design
Best Practices—Campus Core

- Applies to both **Multi-Layer** and **Routed Access** campus designs
- **Keep the Core simple** - higher throughput and fastest recovery around failures.
  - Use L3 redundant links between the Distribution and Core
    - Fast re-route around failures
    - Optimal load balancing
    - No black-holes during recovery
- Use **point-to-point** routed interfaces
  - No Layer 2 switches or VLANs (SVIs)
- Summarize routes into the Core
  - Fault isolation and faster failover

```plaintext
interface TenGigabitEthernet3/1
description 10GigE to Distribution
dampening
ip address 10.122.0.20 255.255.255.254
ip ospf network point-to-point
mls qos trust dscp
```
Best Practice—Build Triangles Not Squares
Deterministic vs. Non-Deterministic

- Redundant Layer 3 equal cost links provide fast convergence
- Hardware based—fast recovery to remaining path
- Convergence is extremely fast (dual equal-cost paths: no need for OSPF or EIGRP to recalculate a new path)
CEF Equal Cost Path Recovery
Redundancy and Protocol Interaction

- The recovery from most component failures is based on L3 CEF equal cost path recovery
- Time to restore traffic flows is based on
  - Time to detect link failure
  - Process the removal of the lost routes from the SW FIB
  - Update the HW FIB
- No dependence on external events (no routing protocol convergence required)
- Behavior is deterministic

**Equal Cost Links: Link/Box Failure**
Does Not Require Multi-Box Interaction
CEF Equal Cost Path Recovery
load-balancing Hashing Mechanism

IPv4 Lookup—10.100.20.199

Prefix Entries / FIB

- 172.20.45.1
- 10.100.20.100
- MASK (/32)
- ...
- 10.100.3.0
- 10.100.2.0
- MASK (/24)
- ...
- 10.100.0.0
- 172.16.0.0
- MASK (/16)

Result Memory

- Adjacency Entry #1
- Adjacency Entry #2
- ...
- Adjacency Entry #15
- Adjacency Entry #16
- ...
- Adj Idx 15 - Path Count 3
- Adjacency Entry #25

Source IP
Dest IP
Optional L4 Ports

IPv4 Lookup—10.100.20.199

Result Memory

- Adj Offset: 0
- Adj Offset: 1
- Adj Offset: 2

Adjacency Table

- New MAC and VLAN
- New MAC and VLAN
- New MAC and VLAN
- Adj Idx 15: Rewrite info
- Adj Idx 15+1: Rewrite info
- Adj Idx 15+2: Rewrite info
- New MAC and VLAN
- New MAC and VLAN
- New MAC and VLAN
- New MAC and VLAN
- New MAC and VLAN
- New MAC and VLAN
- New MAC and VLAN

Hash Result

Switch#show mls cef exact-route 10.77.17.8 10.100.20.199
Interface: Gi1/1, Next Hop: 10.10.1.2, Vlan: 1019, Destination Mac: 0030.f272.31fe
Switch#show mls cef exact-route 10.44.91.111 10.100.20.199
Interface: Gi2/2, Next Hop: 10.40.1.2, Vlan: 1018, Destination Mac: 000d.6550.a8ea
Layer 2 Loops and Spanning Tree
Spanning Tree Should Behave the Way You Expect

- The root bridge should stay where you put it
  - Loopguard and rootguard
  - UDLD
- Only end station traffic should be seen on an edge port
  - BPDU guard
  - Port-Security
- There is a reasonable limit to B-Cast and M-Cast traffic volumes
  - Configure storm control on backup links to aggressively rate limit B-Cast and M-Cast
  - Utilize Sup720 rate limiters or SupIV/V with HW queuing structure
Multilayer Campus Design
Layer 2 Access with Layer 3 Distribution

- Each access switch has unique VLAN’s
- No layer 2 loops
- Layer 3 link between distribution
- No blocked links

- At least some VLAN’s span multiple access switches
- Layer 2 loops
- Layer 2 and 3 running over link between distribution
- Blocked links
Multilayer Network Design

Good Solid Design Option, but ....

- Utilizes multiple Control Protocols
  - Spanning Tree (802.1w, ...), FHRP (HSRP, ...), Routing Protocol (EIGRP, ...)
- Convergence is dependent on multiple factors
  - FHRP - 900msec to 9 seconds
  - Spanning Tree - Upto 50 seconds
  - Poor load balancing – single uplink, asymmetric routing etc
- STP, if it breaks badly, no inherent mechanism to stop the loop
AGENDA

• Systems Level Resiliency
• Network Level Resiliency – Routing
• Campus Core and Foundation Services
  • Emerging Campus Design
    – Routed Access
    – Virtual Switch Campus Design
Routed Access
Layer 3 Distribution with Layer 3 Access

- Move the Layer 2/3 demarcation to the network edge
- Upstream convergence times triggered by hardware detection of light lost from upstream neighbor
- Beneficial for the right environment

EIGRP/OSPF

10.1.20.0 VLAN 20 Data 10.1.40.0 VLAN 40 Data
10.1.120.0 VLAN 120 Voice 10.1.140.0 VLAN 140 Voice
Routed Access Design Considerations

Design Motivations

• Simplified Control Plane
  – No STP feature placement (root bridge, loopguard, ...)
  – No default gateway redundancy setup/tuning
  – No matching of STP/HSRP priority
  – No L2/L3 multicast topology inconsistencies

• Ease of Troubleshooting (leverage well know toolset)
  – Show ip route
  – Traceroute
  – Ping and extended pings
  – Extensive protocol debugs
  – Consistent troubleshooting: access, dist, core

• Failure differences
  – Routed topologies fail closed—i.e. neighbor loss
  – Layer 2 topologies fail open—i.e. broadcast and unknowns flooded
Routed Access
Simplified Network Recovery

• Routed Access network recovery is dependent on L3 re-route
• Time to restore downstream flows is based on a full routing protocol re-route
  – Time to detect link failure
  – Time to determine new route
  – Process the update of the SW RIB & FIB
  – Update the HW FIB
• Time to restore upstream traffic flows is based on ECMP re-route
  – Time to detect link failure
  – Process the removal of the lost routes from the SW FIB
  – Update the HW FIB
Routed Access Design Considerations

Design Requirements

- VLANs are localized to a single wiring closet switch
- IP addressing—do you have an address allocation plan to support a routed access design?
- Platform requirements;
  - Requires a Cisco Catalyst 3560 or above
  - Cisco Catalyst IOS Feature Set considerations
    - IP Base for EIGRP-Stub and PIM*
    - IP Services for OSPF and PIM
Routed Access Design
Advantages, Yes in the Right Environment

• Ease of implementation, less to get right
  – No matching of STP/FHRP priority
  – No L2/L3 multicast topology inconsistencies
  – No STP configuration in Dist

• Single control plane and well known tool set
  – traceroute, show ip route, show ip eigrp neighbor, etc.

• Most Cisco Catalysts support L3 switching today

• EIGRP converges in <200 msec

• OSPF converges in <200 msec with tuning

• RPVST+ convergence times dependent on GLBP/HSRP tuning

Both L2 and L3 Can Provide Sub-Second Convergence
Current Network Scaling Challenges
Campus and Data Center

Traditional Data Center designs are requiring ever increasing Layer 2 adjacencies between servers due to applications, Virtualization technology and server growth. The size of Layer 2 networks is stretched, placing more burden on loop-avoidance protocols (Spanning Tree)

- **BGP, IGP, ECMP**
  - Policy Management

- **FHRP, Single active uplink per VLAN, L2 reconvergence, excessive BPDUs**

- **Dual-Homed Servers, Single active uplink per VLAN (PVST), L2 reconvergence**

![Diagram showing L2 Core, L2/L3 Aggregation, L2 Access, and DC Pod: L2 Domain with BGP, IGP, ECMP and FHRP details.](image-url)
Virtual Switch
Catalyst 6500 Virtual Switching System (VSS)

- Virtual Switching System consists of two Catalyst 6500’s defined as members of the same virtual switch domain running a VSL (Virtual Switch Link) between them
- Single Control Plane with Dual Active Forwarding Planes
- Extends NSF/SSO infrastructure to Two Switches

Virtual Switch Domain

Virtual Switch Link (VSL)

Switch 1 + Switch 2 = VSS
Virtual Switch – VSS

Two to One

Two switches look like one
  Two physical switches
  One virtual switch

Virtual Switch:
  All ports appear to be on the same physical switch
  Single point of management
  Single configuration
  Single IP/MAC
  Single control plane protocol instance

Benefits
  Simplify infrastructure management
  L2 DC Interconnect High Availability
Impact of VSS on the Campus Design

Control Plane Simplification

- Virtual Switch Design simplifies the topology
- Redundant supervisors provide resiliency via SSO
- No need for HSRP, GLBP or VRRP
- A single multicast router on the access subnets simplifies the multicast topology
- No L2 loops in the topology so no need for spanning tree to provide for link redundancy
- Do **NOT** disable spanning tree as it is still possible to create an external loop
- Catalyst 6500 Load-balancing scheme modified to keep traffic forwarding local
A Virtual Switch-enabled Campus/Datacenter allows for maximum scalability so bandwidth can be added when required, but still providing a larger Layer 2 hierarchical architecture free of reliance on Spanning Tree...

- **L3 Core**
  - Single router node, Fast L2 convergence, Scalable architecture

- **L2/L3 Aggregation**
  - Dual Active Uplinks, Fast L2 convergence, minimized L2 Control Plane, Scalable

- **L2 Access**
  - Dual-Homed Servers, Single active uplink per VLAN (PVST), Fast L2 convergence
Virtual Portchannels
Nexus 7000 vPC

Two Physical to a single logical
- Devices connect to a single “logical” switch
- Connections are treated as portchannel

Virtual PortChannel:
- Ports to virtual switch could form a cross-chassis portchannel
- virtual Portchannel behaves like a regular Etherchannel

Benefits
- Provide non-blocking L2 paths
- Lessen Reliance on STP
Virtual Portchannels
Nexus 7000 vPC

- What Is Virtual Port Channel
  - Eliminates STP blocked ports
  - Leverages all available uplink bandwidth
  - Eliminates active-standby mode on dual-homed servers
  - Provides fast, transparent convergence upon link/device failure
  - Works seamlessly with current network design/topology

- Downstream end:
  - Standard link load balancing protocols available
    (depends on downstream device; src/dst-mac, round-robin, etc.)
  - Works with LACP and manually configured links

- vPC end:
  - Same as above
  - Load-balancing scheme modified to keep traffic forwarding local (i.e., packet headed into the link aggregation group will use one of the local links rather than across the vPC peer-link)
Next Generation Campus Design
Evolving the Campus Foundation Architecture

• Traditional Layer 2 designs remain valid
• Evolving architectures provide
  – Simplified Control Plane: Remove dependence on STP
  – Increased Capacity: Provide flow-based load balancing
  – High Availability: 200 msec or better recovery
• Flexibility to provide for the right implementation for each network requirement
Campus Design Guidance
Where to go for more information
