Cisco Nexus 5000/5500 and 2000 Switch Architecture

Mike Herbert

Technical Marketing Engineer - SAVTG

Session Goal

- This session presents an in-depth study of the architecture of the Nexus 5000/5550 family of Data Center switches and the Nexus 2000 Fabric Extender. Topics include internal architecture of the Nexus 5000, 5500 and 2000, the architecture of fabric and port extenders as implemented in the Nexus 2000 and Adapter FEX, Unified I/O, and 10G cut-thru Layer 2 and Layer 3 Ethernet. This session is designed for network engineers involved in network switching design and Data Center architecture.
- Related sessions:

  BRKARC-3470 - Cisco Nexus 7000 Switch Architecture

  BRKCRS-3144 - Troubleshooting Cisco Nexus 7000 Series Switches

  BRKCRS-3145 - Troubleshooting Cisco Nexus 5000 / 2000 Series Switches

  BRKARC-3472 - NX-OS Routing & Layer 3 Switching

  BRKDCT-2023 - Evolution of the Data Centre Access Architecture *

  BRKSAN-2047 - FCoE Design, Operations and Management Best Practices *

** This session is focusing on the Hardware Architecture of the Nexus 5000, 5500 and 2000, please see the DC Access and FCoE design sessions for a detailed discussion of the best practices design options for N2K/N5K and FCoE*

Nexus 5000/5500 and 2000 Architecture

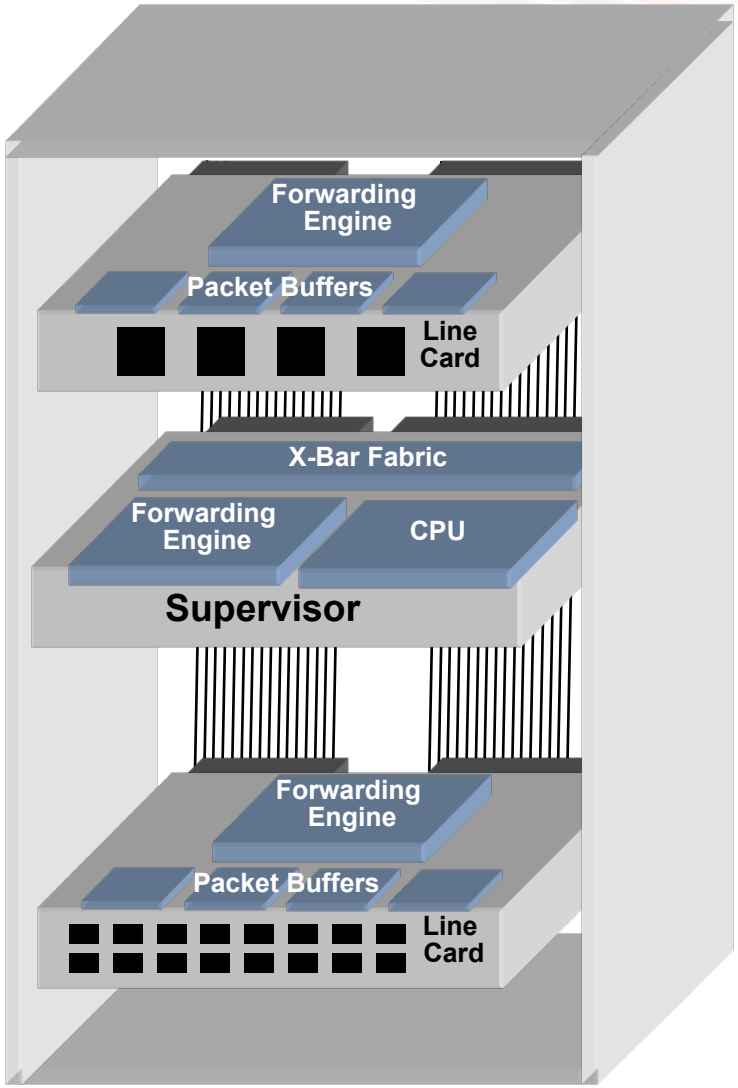
Agenda

- Nexus 5000/5500 Architecture
  - Hardware Architecture
  - Day in the Life of a Packet
  - Layer 3 Forwarding
- Nexus 2000 Architecture
  - FEXLink Architecture
  - FEX Forwarding
  - Extending FEXLink – Adapter FEX
- Nexus 5000/5500
  - Multicast
  - Port Channels
  - QoS

Nexus 5000/5500 and 2000 Architecture

Switch Morphology

- What's in a switch?
- Lookup/forwarding logic
  - L2/L3 forwarding, ACL, QoS TCAM
- CPU
  - Control and management
- Packet transport
  - Cross bar switching fabric
  - Component interconnects
- Ports (line cards)
  - Port and Fabric buffers
- Your Architectural Decisions
  - How to optimize the availability, functionality, operational and capital costs of the 'network fabric'
  - How to interconnect 'components' to meet these needs



Nexus 5000/5500 and 2000 Architecture

Virtualized Data Center Access

Nexus 5010

20 Fixed Ports 10G/FCoE/IEEE DCB
Line-rate, Non-blocking 10G
1 Expansion Module Slot
Redundant Fans & Power Supplies

Nexus 2000 Fabric Extender

48 Fixed Ports 100M/1G Ethernet (1000 BASE-T)
32 Fixed ports 1G/10G/FCoE/IEEE DCB
4-8 Fixed Port 10G Uplink
Distributed Virtual Line Card

Nexus 5020

40 Fixed Ports 10G/FCoE/IEEE DCB
Line-rate, Non-blocking 10G
2 Expansion Module Slots
Redundant Fans & Power Supplies

Nexus 5548UP

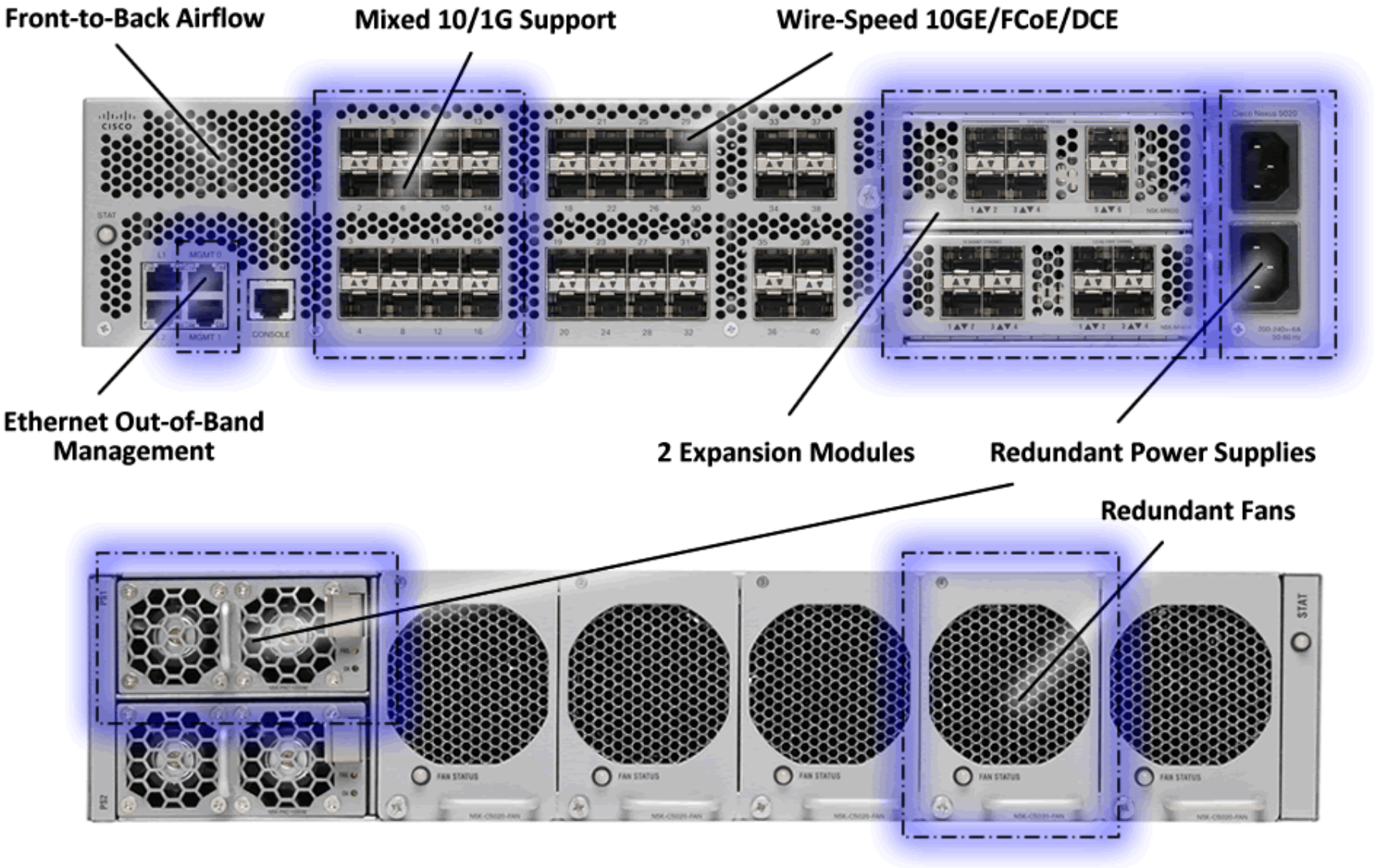
32 Fixed Ports 1/10G Ethernet or 1/2/4/8 FC
Line-rate, Non-blocking 10G FCoE/IEEE DCB
1 Expansion Module Slot
IEEE 1588, FabricPath & Layer 3 Capable
Redundant Fans & Power Supplies

Nexus 5596UP

48 Fixed Ports 1/10G Ethernet or 1/2/4/8 FC
Line-rate, Non-blocking 10G FCoE/IEEE DCB
3 Expansion Module Slot
IEEE 1588, FabricPath & Layer 3 Capable
Redundant Fans & Power Supplies

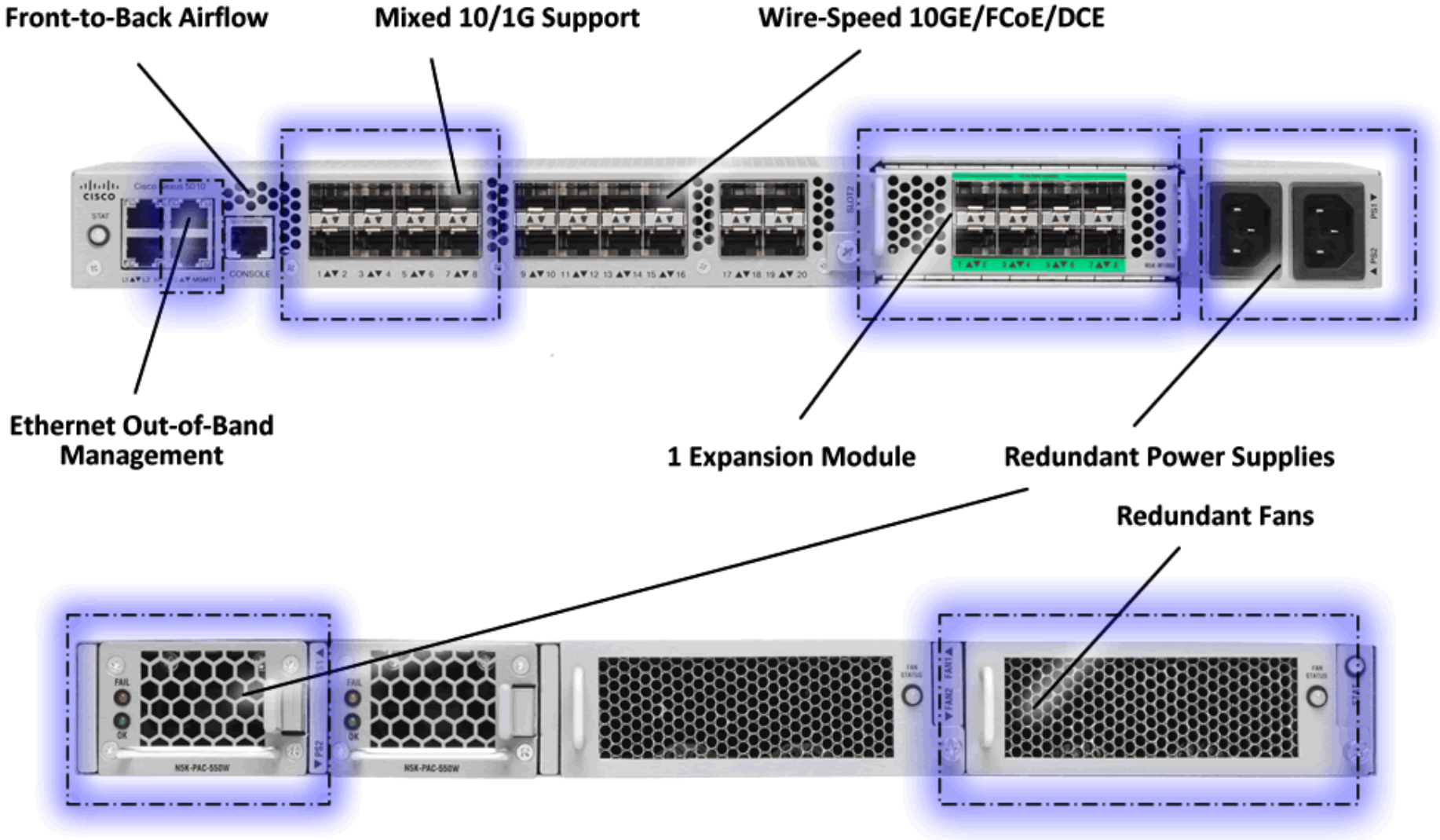
Nexus 5000 Hardware

Nexus 5020



Nexus 5000 Hardware

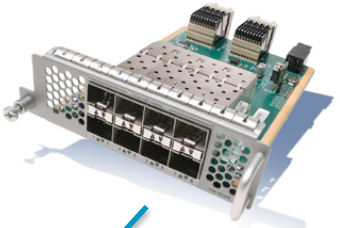
Nexus 5010

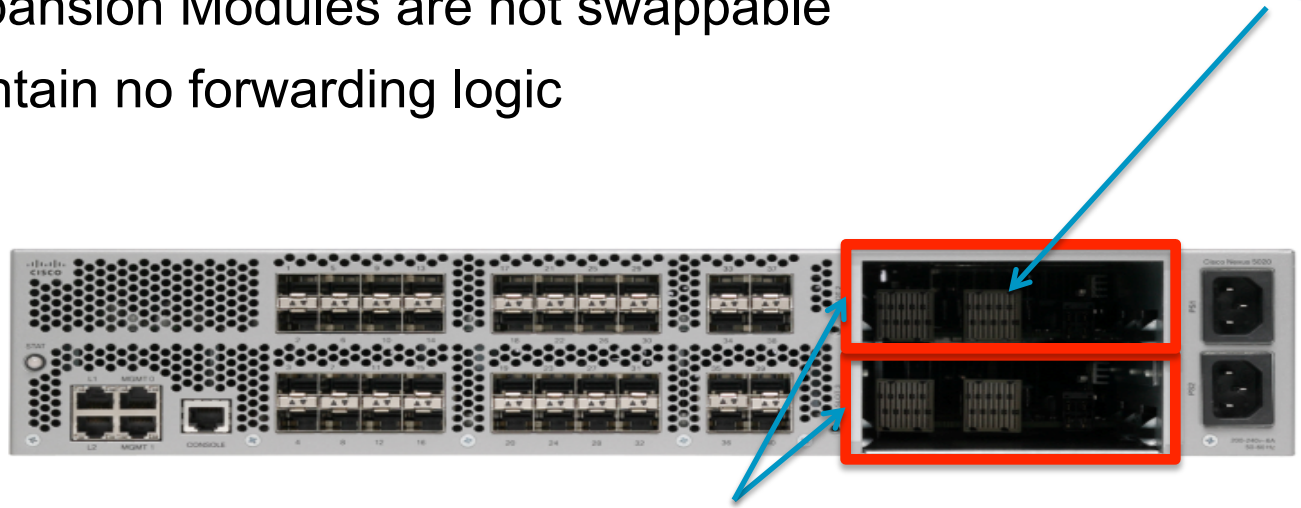


Nexus 5000 Architecture

Nexus 5000 Expansion Modules

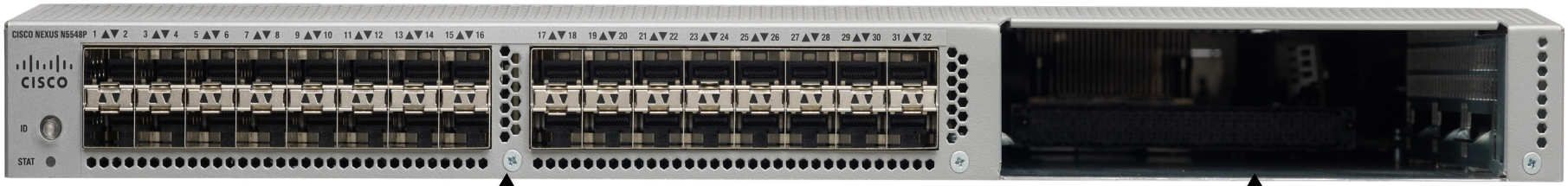
- Nexus 5000 utilizes expansion slots to provide flexibility of interface types
  - Additional 10GE DCB/FCoE compliant ports
  - 1/2/4/8G Fibre Channel ports
- Nexus 5020 has two expansion module slots
- Nexus 5010 has one expansion module slot
- Expansion Modules are hot swappable
- Contain no forwarding logic





Nexus 5500 Hardware

Nexus 5548 (5548P & 5548UP)



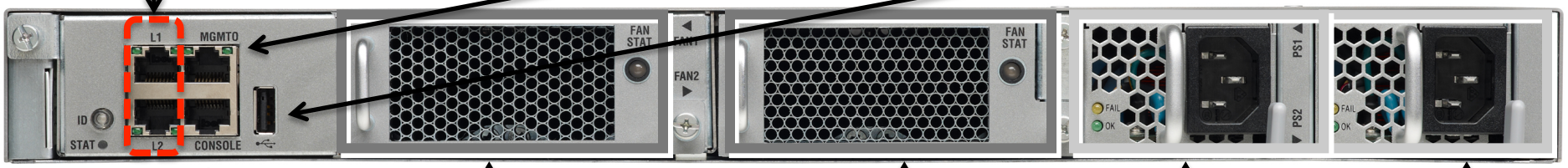
32 x Fixed Unified Ports 1/10 GE or 1/2/4/8 FC

Expansion Module

Fabric Interconnect Not Active on Nexus

Out of Band Mgmt 10/100/1000

USB Flash



Console

Fan Module

Fan Module

Power Entry

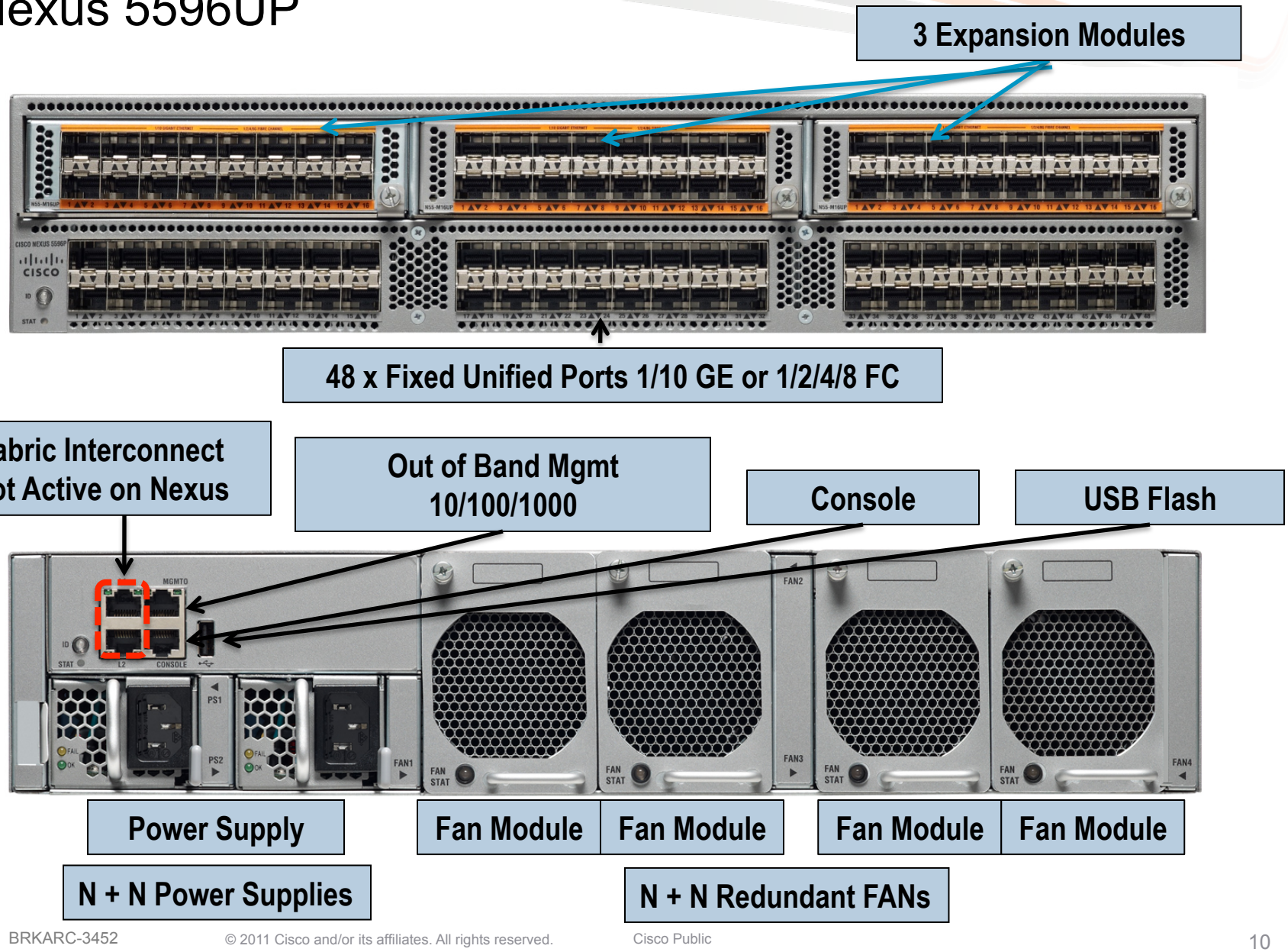
Power Entry

N + N Redundant FANs

N + N Power Supplies

Nexus 5500 Hardware

Nexus 5596UP



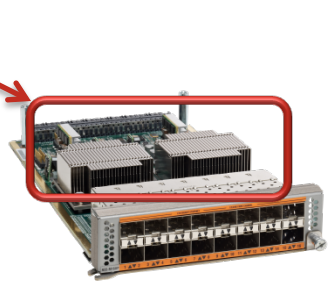
Nexus 5500 Hardware

Nexus 5500 Expansion Modules

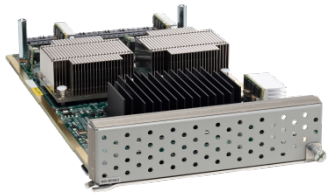
- Nexus 5500 expansion slots
  - Expansion Modules are hot swappable (Future support for L3 OIR)
  - Contain forwarding ASIC (UPC-2)

16 x 1/10GE

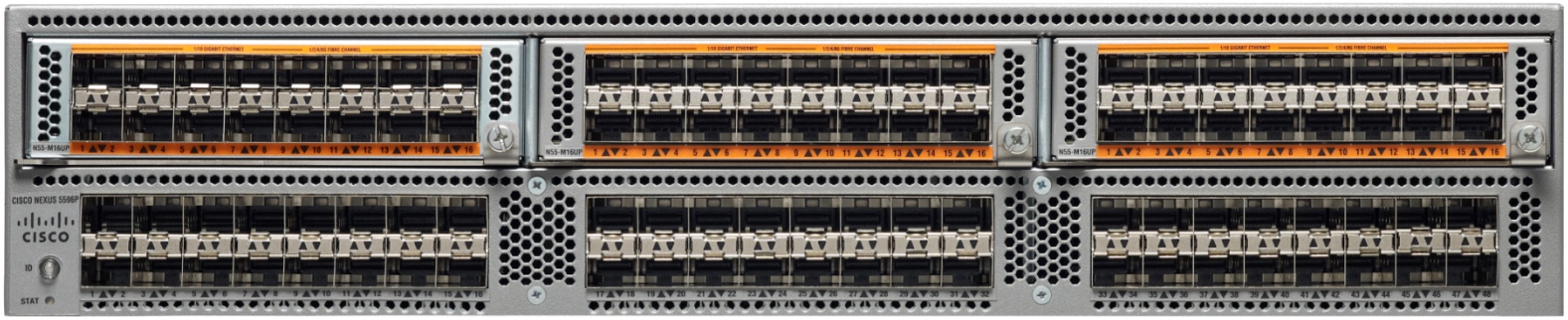
8 x 1/10GE + 8 x 1/2/4/8G FC



16 unified ports individually configurable as 1/10GE or 1/2/4/8G FC



L3 module for 160G of L3 I/O bandwidth



Nexus 5000 Hardware

Nexus 5000 Power Supplies

- Nexus 5020 power supplies
  - 1200 watt (N5K-PAC-1200W)
  - 750 watt (N5K-PAC-750W)
- Fully-loaded Nexus 5020 with 2 expansion modules and all links running at line rate only requires a ***single*** 750 watt power supply

```
dc11-5020-3# sh environment power

Power Supply:
Voltage: 12 Volts
-----------------------------------------------------
PS  Model                Power        Power     Status
                         (Watts)      (Amp)
-----------------------------------------------------
1   N5K-PAC-1200W        1200.00    100.00     ok
2   N5K-PAC-1200W        1200.00    100.00     ok

<snip>
Total Power Capacity                              2400.00 W

Power reserved for Supervisor(s)                   625.20 W
Power currently used by Modules                     72.00 W


                                                  ------------
Total Power Available                             1702.80 W
                                                  ------------
```

Nexus 5500 Hardware

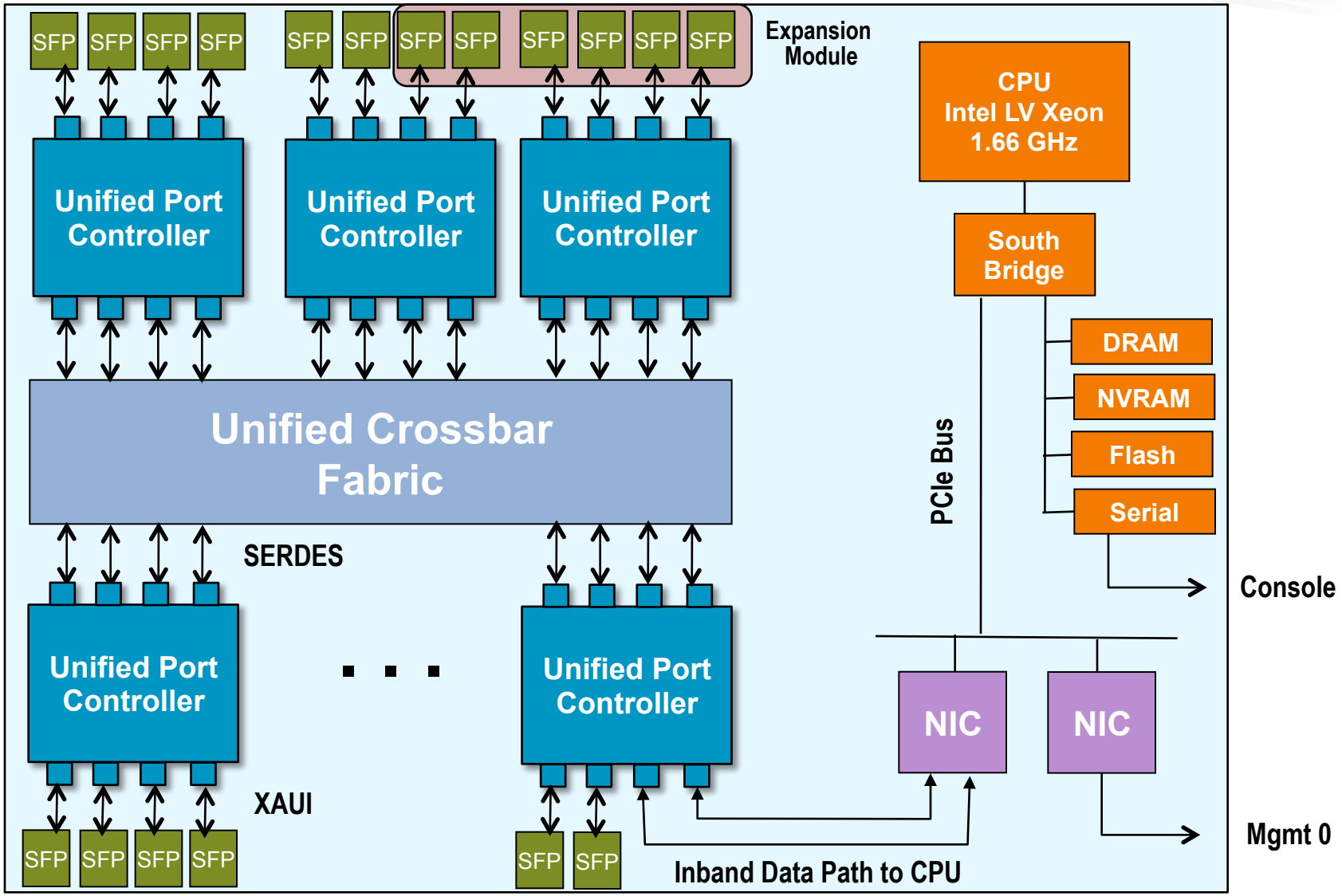
Nexus 5500 Reversible Air Flow and DC Power Supplies

- Nexus 2000, 5548UP and 5596UP will support reversible airflow (new PS and fans)
- Nexus 2000, 5548UP and 5596UP will support DC power supplies (not concurrent with reversible airflow)
- Note: 5548UP and 5596UP ***ONLY,*** not 5010/5020/5548P

	Nexus 2000	Hardware Availability	Nexus 5000	Hardware Availability
Front-to-Back Airflow, AC Power	Nexus 2148T Nexus 2200 Series	Today	Nexus 5010/5020 Nexus 5548P/ 5548UP/5596UP	Today
Back-to-Front Airflow, AC Power	Nexus 2200 Series	Q2CY11	Nexus 5548UP/ 5596UP	Nexus 5548UP Q2CY11, Nexus 5596UP (Future)
Front-to-Back Airflow, DC Power	Nexus 2200 Series	Q2CY11	Nexus 5548UP/ 5596UP	Nexus 5548UP Q3CY11, Nexus 5596UP (Future)
Back-to-Front Airflow, DC Power	N/A	N/A	N/A	N/A

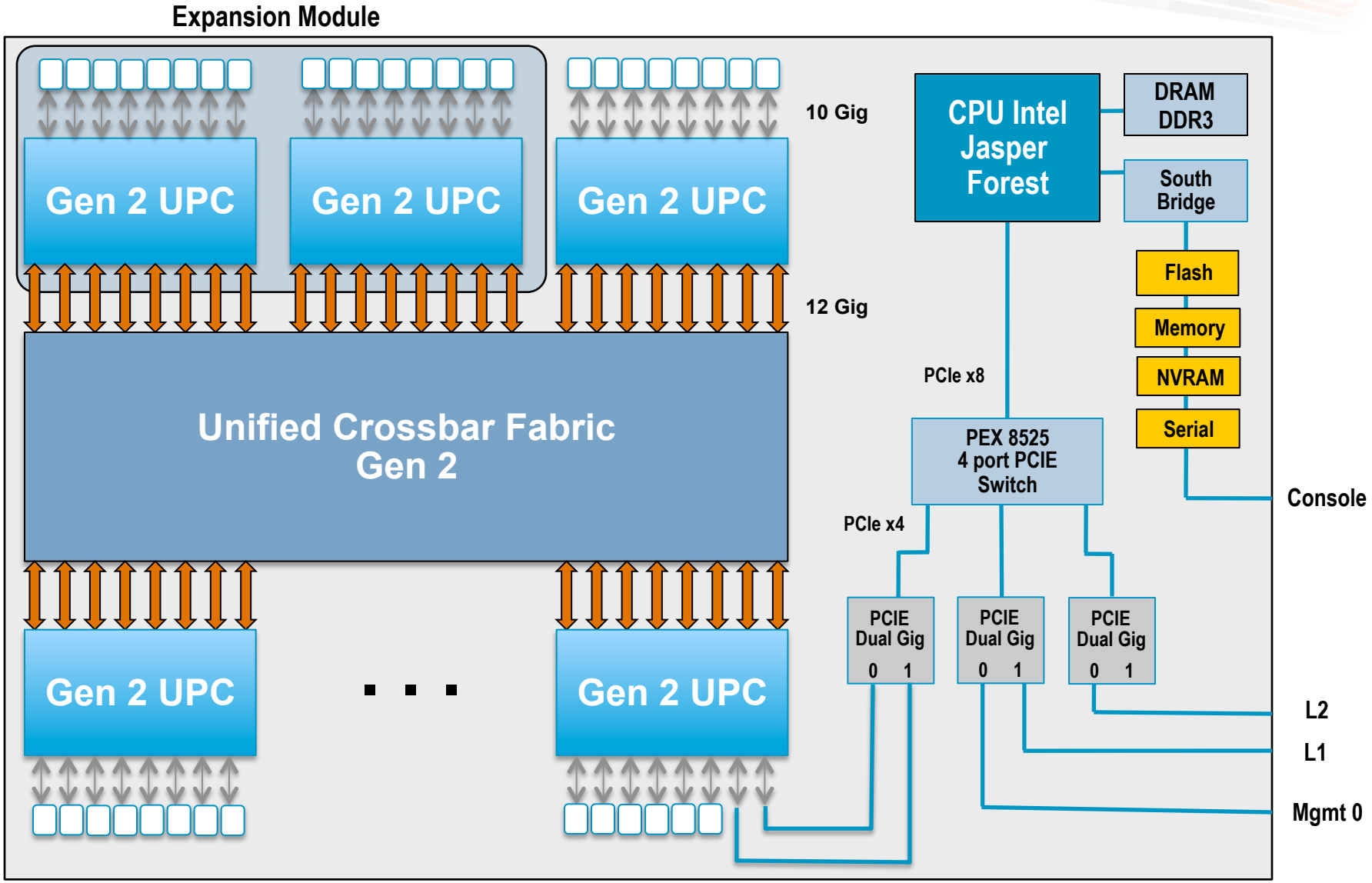
Nexus 5000 Hardware Overview

Data and Control Plane Elements



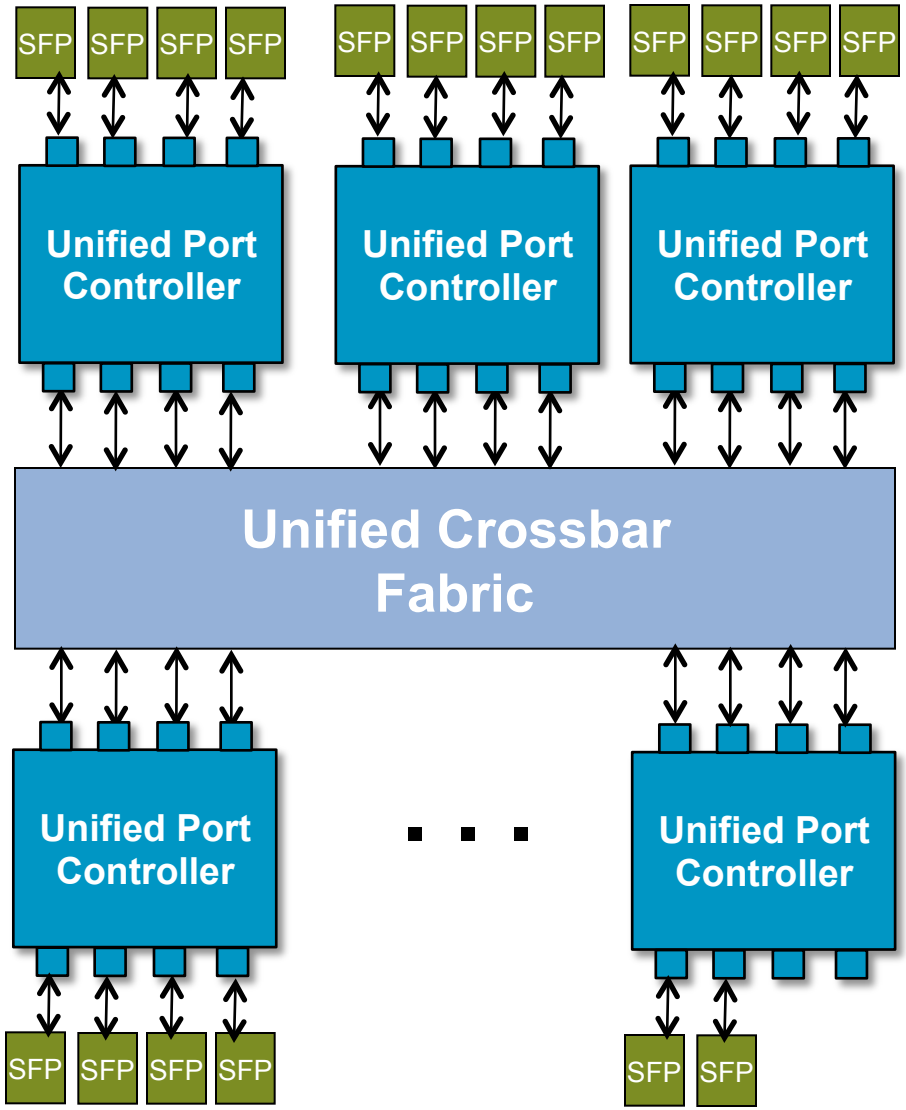
Nexus 5500 Hardware Overview

Data and Control Plane Elements



Nexus 5000/5500 Hardware Overview

Data Plane Elements – Distributed Forwarding



- Nexus 5000/5500 use a distributed forwarding architecture
- Unified Port Controller (UPC) ASIC interconnected by a single stage Unified Crossbar Fabric (UCF)
- Unified Port Controllers provide distributed packet forwarding capabilities
- ***All*** port to port traffic passes through the UCF (Fabric)
- Cisco Nexus 5020: Layer 2 hardware forwarding at 1.04 Tbps or 773.8 million packets per second (mpps)
- Cisco Nexus 5596: Layer 2 hardware forwarding at 1.92Tbps or 1428 mpps

Nexus 5000/5500 Hardware Overview

Data Plane Elements - Unified Crossbar Fabric

- Nexus 5000 (Gen-1)
  - 58-port packet based crossbar and scheduler
  - Three unicast and one multicast crosspoint per egress port
- Nexus 5550 (Gen-2)
  - 100-port packet based crossbar and new schedulers
  - 4 crosspoints per egress port dynamically configurable between multicast and unicast traffic
- Central tightly coupled scheduler

  Request, propose, accept, grant, and acknowledge semantics

  Packet enhanced iSLIP scheduler

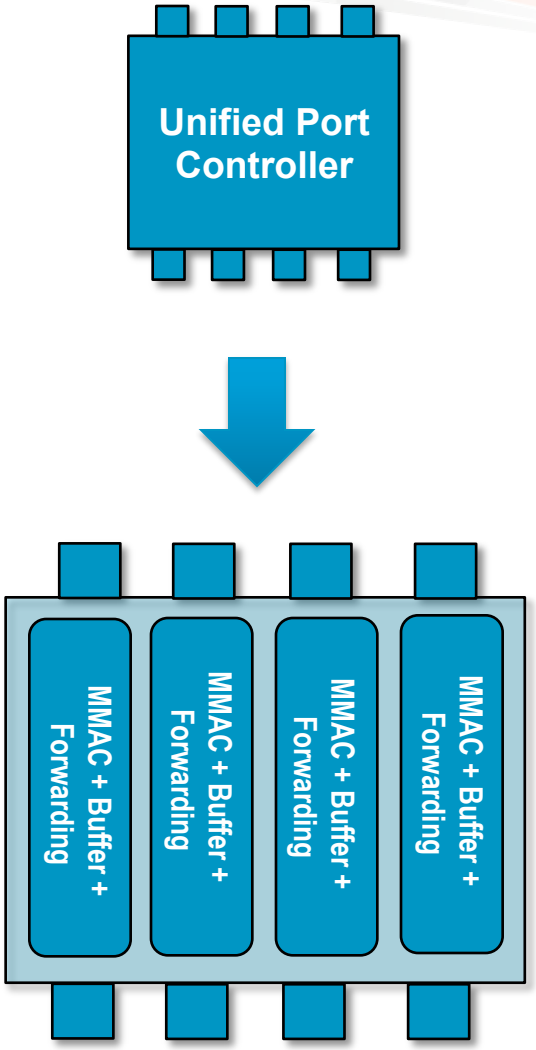
  Distinct unicast and multicast schedulers (see slides later for differences in Gen-1 vs. Gen-2 multicast schedulers)

  Eight classes of service within the Fabric

Nexus 5000 Hardware Overview

Data Plane Elements - Unified Port Controller (Gen 1)

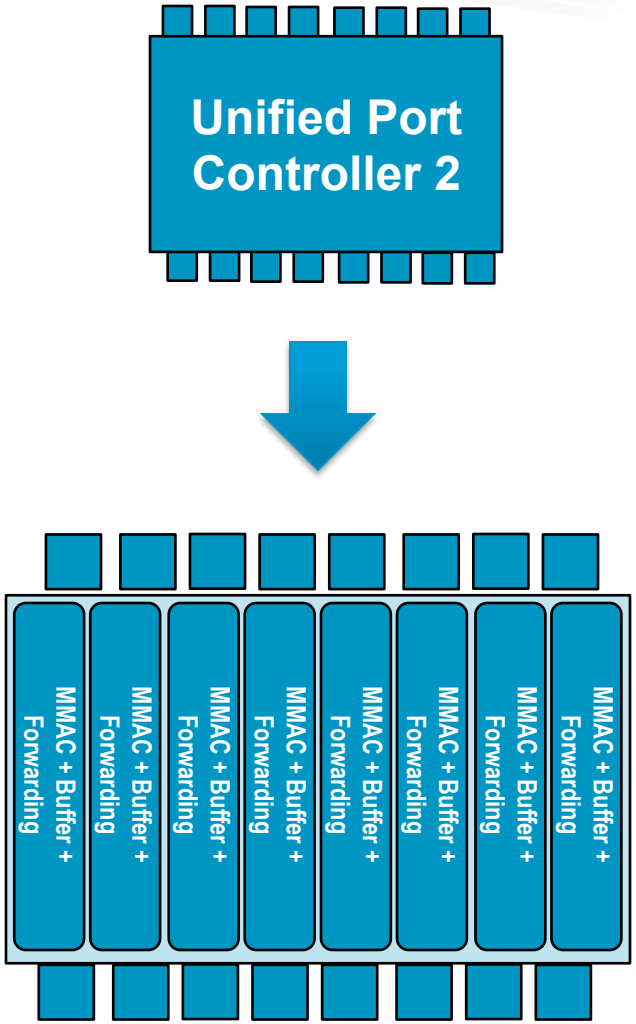
- Each UPC supports four ports and contains,
- Multimode Media access controllers (MAC)
  - Support 1/10 G Ethernet and 1/2/4 G Fibre Channel on the UPC + PHY
  - (2/4/8 G Fibre Channel MAC/PHY is located on the Expansion Module)
- Packet buffering and queuing
  - 480 KB of buffering per port
- Forwarding controller
  - Ethernet and Fibre Channel Forwarding and Policy



Nexus 5500 Hardware Overview

Data Plane Elements - Unified Port Controller (Gen 2)

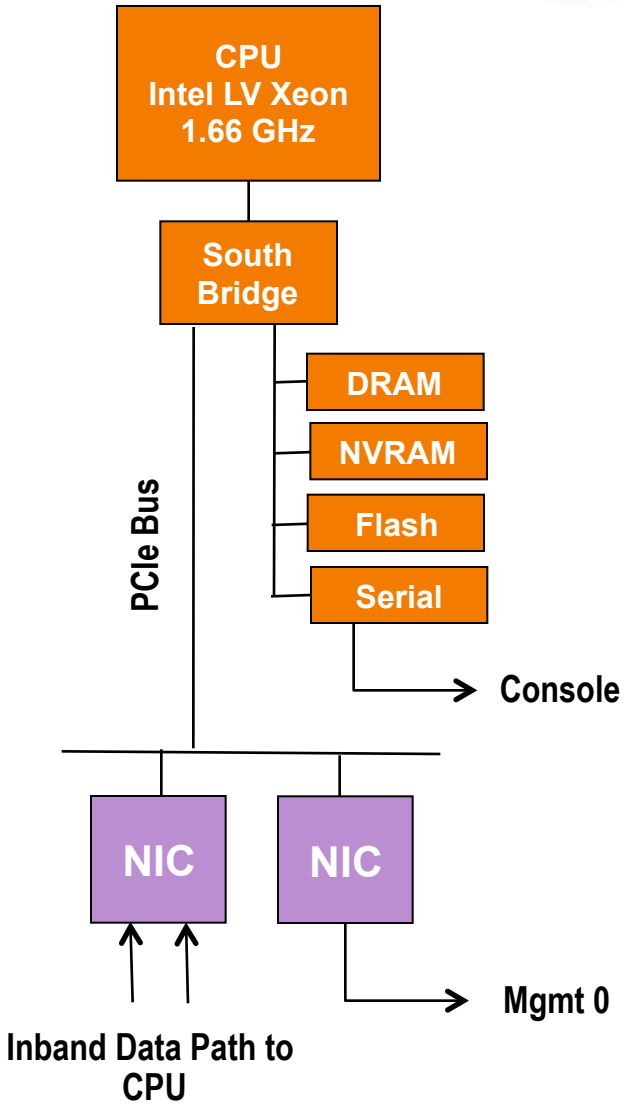
- Each UPC supports eight ports and contains,
- Multimode Media access controllers (MAC)
  - Support 1/10 G Ethernet and 1/2/4/8 G Fibre Channel
  - All MAC/PHY functions supported on the UPC (5548UP and 5596UP)
- Packet buffering and queuing
  - 640 KB of buffering per port
- Forwarding controller
  - Ethernet (Layer 2 and FabricPath) and Fibre Channel Forwarding and Policy (L2/ L3/L4 + all FC zoning)



Nexus 5000 Hardware Overview

Control Plane Elements – Nexus 5000

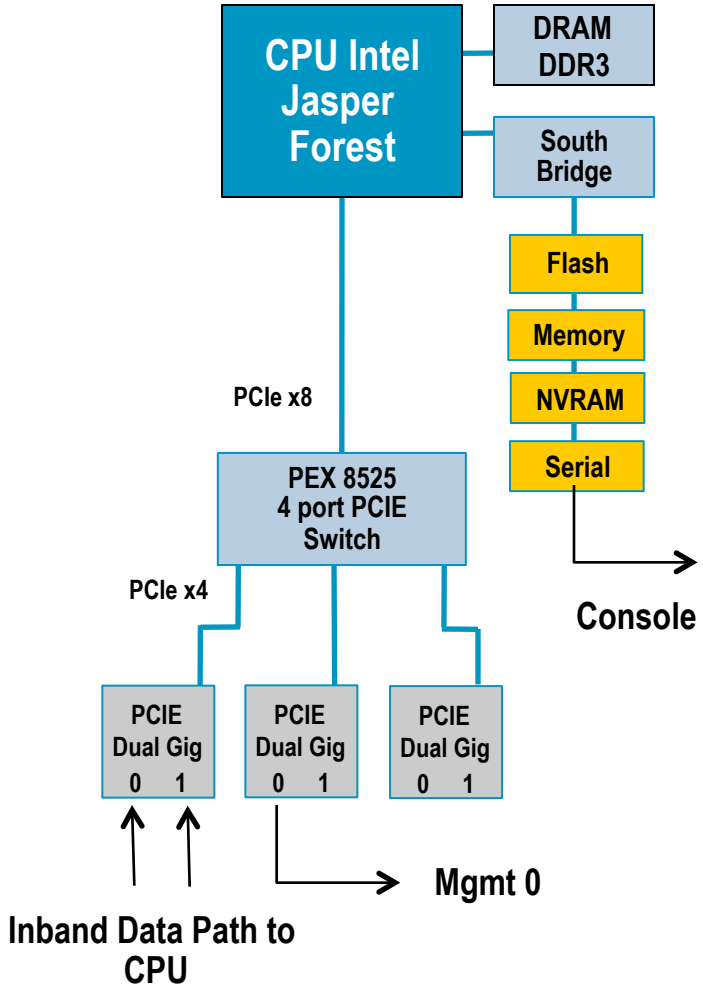
- CPU - 1.66 GHz Intel LV Xeon
- DRAM - 2 GB of DDR2 400 (PC2 3200) in two DIMM slots
- Program Store - 1 GB of USB-based (NAND) flash
- Boot/BIOS - 2 MB of EEPROM with locked recovery image
- On-Board Fault Log - 64 MB of flash for failure analysis
- NVRAM - 2 MB of SRAM: Syslog and licensing information
- Management Interfaces - RS-232 console port: console 0
- Mgmt 0 interface partitioned from in-band VLANs



Nexus 5500 Hardware Overview

Control Plane Elements – Nexus 5500

- CPU - 1.7 GHz Intel Jasper Forest (Dual Core)
- DRAM - 8 GB of DDR3 in two DIMM slots
- Program Store - 2 GB of eUSB flash for base system storage and partitioned to store image, configuration, log.
- Boot/BIOS Flash - 8 MB to store upgradable and golden version of (Bios + bootloader) image
- On-Board Fault Log (OBFL) - 64 MB of flash to store hardware related fault and reset reason
- NVRAM - 6 MB of SRAM to store Syslog and licensing information
- Management Interfaces
  - RS-232 console port: console0
  - 10/100/1000BASE-T: mgmt0 partitioned from inband VLANs



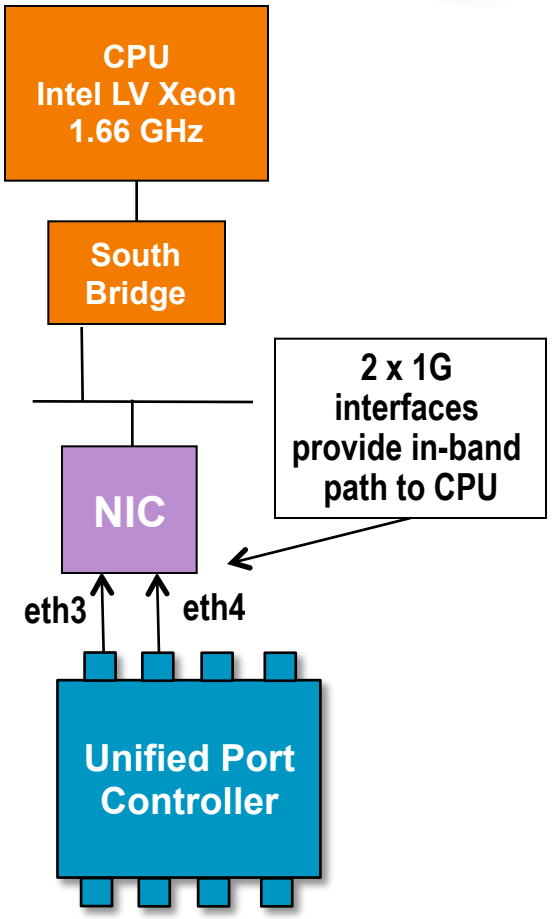
Nexus 5000/5500 Hardware Overview

Control Plane Elements - CoPP

- In-band traffic is identified by the UPC and punted to the CPU via two dedicated UPC interfaces, 5/0 and 5/1, which are in turn connected to eth3 and eth4 interfaces in the CPU complex
- Eth3 handles Rx and Tx of ***low*** priority control pkts

  IGMP, CDP, TCP/UDP/IP/ARP (for management purpose only)
- Eth4 handles Rx and Tx of ***high*** priority control pkts

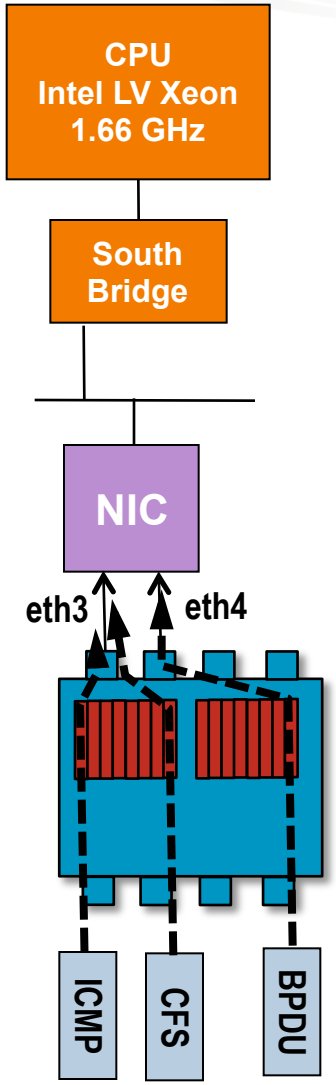
  STP, LACP, DCBX, FC and FCoE control frames (FC packets come to Switch CPU as FCoE packets)



Nexus 5000/5500 Hardware Overview

Control Plane Elements - CoPP

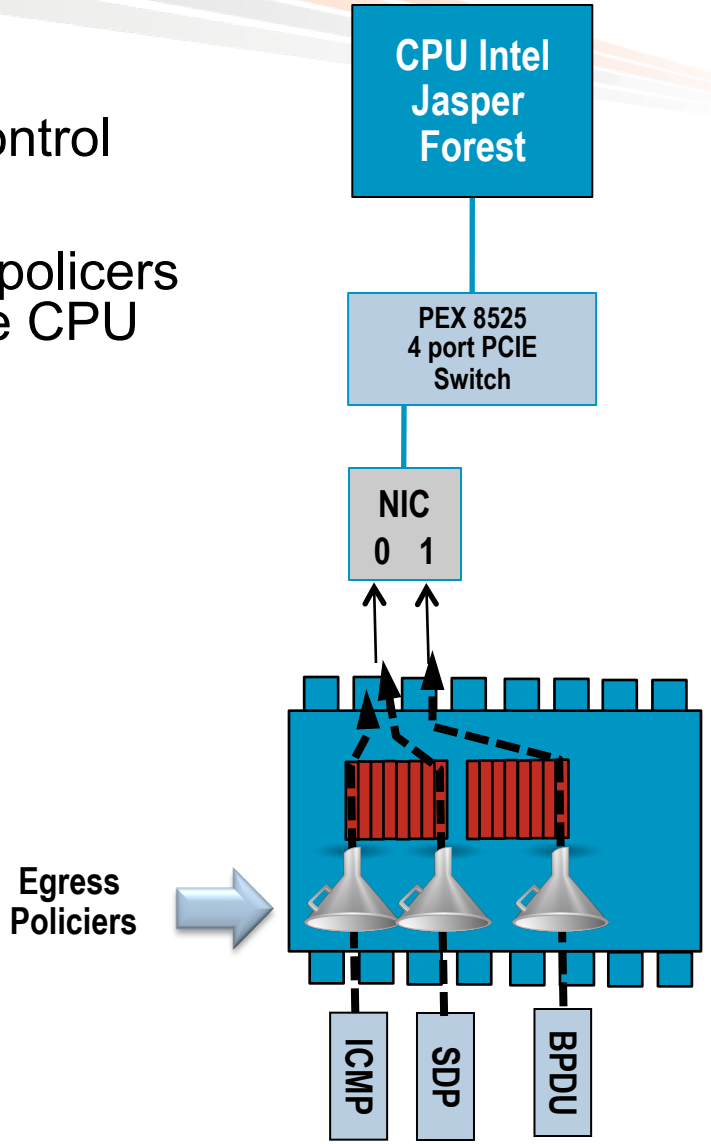
- CPU queuing structure provides strict protection and prioritization of inbound traffic
- Each of the two in-band ports has 8 queues and traffic is scheduled for those queues based on control plane priority (traffic CoS value)
- Prioritization of traffic between queues on each in-band interface
  - CLASS 7 is configured for strict priority scheduling (e.g. BPDU)
  - CLASS 6 is configured for DRR scheduling with 50% weight
  - Default classes (0 to 5) are configured for DRR scheduling with 10% weight
- Additionally each of the two in-band interfaces has a priority service order from the CPU
  - Eth 4 interface has high priority to service packets (no interrupt moderation)
  - Eth3 interface has low priority (interrupt moderation)



Nexus 5500 Hardware Overview

Control Plane Elements - CoPP

- On ***Nexus 5500*** an additional level of control invoked via policers on UPC-2
- Software programs a number of egress policers on the UPC-2 to avoid overwhelming the CPU (partial list)
  - STP: 20 Mbps
  - LACP: 1 Mbps
  - DCX: 2 Mbps
  - Satellite Discovery protocol: 2 Mbps
  - IGMP: 1 Mbps
  - DHCP: 1 Mbps
  - . . .
- CLI exposed to tune CoPP (Future)



Nexus 5000/5500 Hardware Overview

Control Plane Elements

- Monitoring of in-band traffic vis the NX-OS built-in ethanalyzer

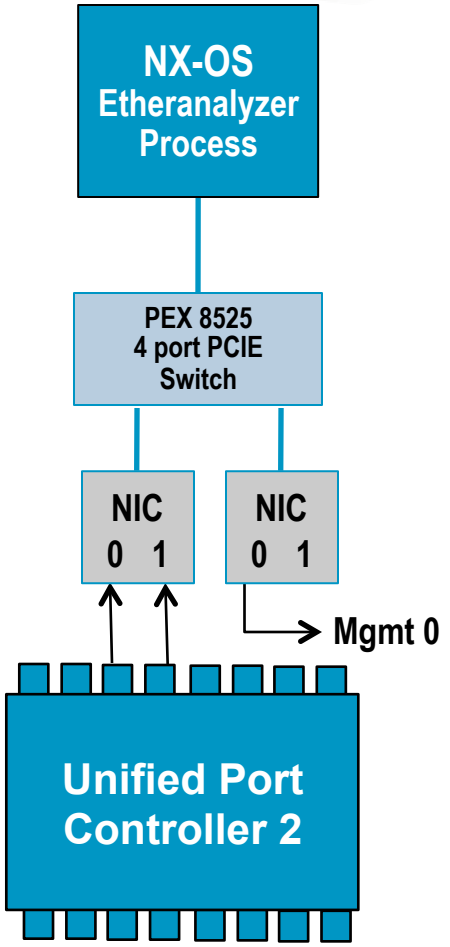
  Eth3 is equivalent to 'inbound-lo'

  Eth4 is equivalent to 'inbound-hi'

```
dc11-5020-3# ethanalyzer local sniff-interface ?
  inbound-hi    Inbound(high priority) interface
  inbound-low   Inbound(low priority) interface
  mgmt          Management interface
```

- CLI view of in-band control plane data

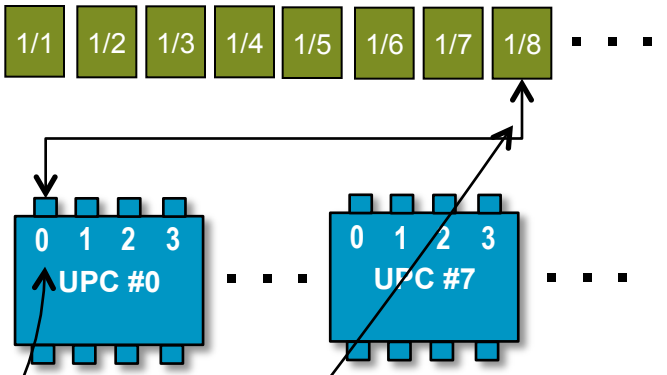
```
dc11-5020-4# sh hardware internal cpu-mac inband counters
eth3      Link encap:Ethernet  HWaddr 00:0D:EC:B2:0C:83
          UP BROADCAST RUNNING PROMISC ALLMULTI MULTICAST  MTU:2200  Metric:1
          RX packets:3 errors:0 dropped:0 overruns:0 frame:0
          TX packets:630 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:252 (252.0 b)  TX bytes:213773 (208.7 KiB)
          Base address:0x6020 Memory:fa4a0000-fa4c0000

eth4      Link encap:Ethernet  HWaddr 00:0D:EC:B2:0C:84
          UP BROADCAST RUNNING PROMISC ALLMULTI MULTICAST  MTU:2200  Metric:1
          RX packets:85379 errors:0 dropped:0 overruns:0 frame:0
          TX packets:92039 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:33960760 (32.3 MiB)  TX bytes:25825826 (24.6 MiB)
          Base address:0x6000 Memory:fa440000-fa460000
```



Nexus 5000 Hardware Overview

Nexus 5000 – UPC (Gen 1) and Port Mapping

- UPC interfaces are indirectly mapped to front panel ports
- Mapping of ports to UPC (Gatos) ASIC
  - The left column identifies the Ethernet interface identifier, xgb1/8 = e1/8
  - Column three and four reflect the UPC port that is associated with the physical Ethernet port

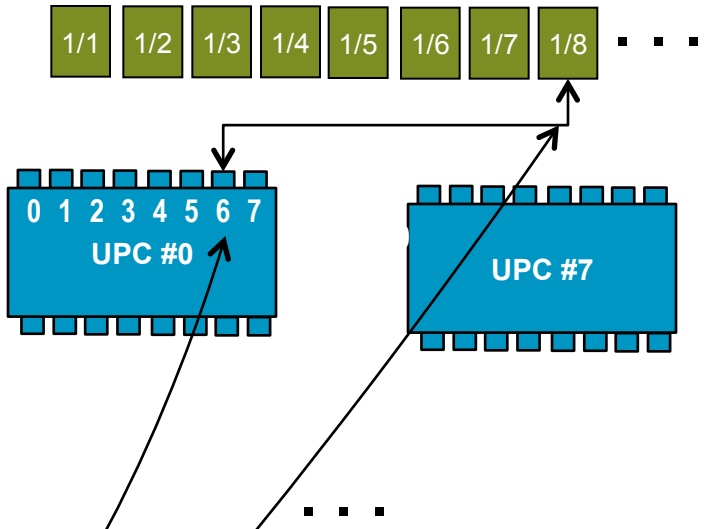



```
nexus-5020# show hardware internal gatos all-ports
<snip>

Gatos Port Info:
name    |log|gat|mac|flag|adm|opr|c:m:s:l|ipt|fab|xgat|xpt|if_index|diag
--------+---+---+---+----+---+---+-------+---+---+----+---+--------+----
1gb1/8  |7  |0  |0  |b7  |en |up |0:0:0:0|0  |55 |0   |2  |1a007000|pass
1gb1/7  |6  |0  |1  |b7  |dis|dn |0:1:1:0|1  |54 |0   |0  |1a006000|pass
1gb1/3  |2  |0  |2  |b7  |en |up |1:2:2:0|2  |56 |0   |4  |1a002000|pass
xgb1/4  |3  |0  |3  |b7  |dis|dn |1:3:3:f|3  |57 |0   |6  |1a003000|pass
<snip>
xgb1/1  |0  |7  |2  |b7  |dis|dn |1:2:2:f|2  |6  |7   |4  |1a000000|pass
xgb1/2  |1  |7  |3  |b7  |dis|dn |1:3:3:f|3  |7  |7   |6  |1a001000|pass
```

Nexus 5500 Hardware Overview

Nexus 5500 – UPC (Gen 2) and Port Mapping

- UPC-2 interfaces are indirectly mapped to front panel ports
- Mapping of ports to UPC-2 ASIC
  - The left column identifies the Ethernet interface identifier, xgb1/8 = e1/8
  - Column three and four reflect the UPC port that is associated with the physical Ethernet port



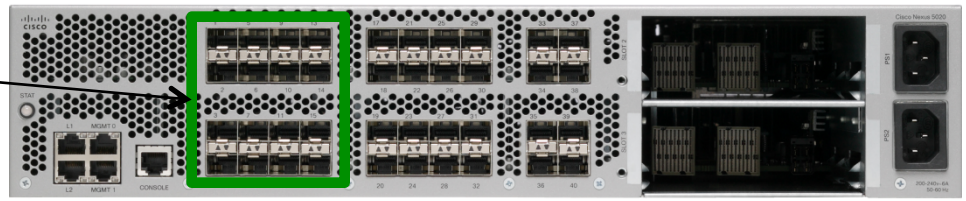

```
nexus-5548# show hardware internal carmel all-ports

Carmel Port Info:
name    |log|car|mac|flag|adm|opr|m:s:l|ipt|fab|xcar|xpt|if_index|diag|ucVer
--------+---+---+---+----+---+---+-----+---+---+----+---+--------+----+-----
xgb1/2 |1   |0   |0 -|b7   |dis|dn |0:0:f|0   |92 |0    |0   |1a001000|pass| 4.0b
xgb1/1 |0   |0   |1 -|b7   |dis|dn |1:1:f|1   |88 |0    |0   |1a000000|pass| 4.0b
xgb1/4 |3   |0   |2 -|b7   |dis|dn |2:2:f|2   |93 |0    |0   |1a003000|pass| 4.0b
xgb1/3 |2   |0   |3 -|b7   |dis|dn |3:3:f|3   |89 |0    |0   |1a002000|pass| 4.0b
xgb1/6 |5   |0   |4 -|b7   |dis|dn |4:4:f|4   |90 |0    |0   |1a005000|pass| 4.0b
xgb1/5 |4   |0   |5 -|b7   |dis|dn |5:5:f|5   |94 |0    |0   |1a004000|pass| 4.0b
xgb1/8 |7   |0   |6 -|b7   |dis|dn |6:6:f|6   |95 |0    |0   |1a007000|pass| 4.0b
<snip>
sup0    |32 |4   |4 -|b7   |en |dn |4:4:0|4   |62 |0    |0   |15020000|pass| 0.00
sup1    |33 |4   |5 -|b7   |en |dn |5:5:1|5   |59 |0    |0   |15010000|pass| 0.00
```

Nexus 5000 Hardware Overview

5010/5020 - UPC (Gen 1) and 1G Ethernet

- Support for 1G speed on first 16 ports of Nexus 5020 and first eight ports of Nexus 5010
- Need to explicitly specify that the port runs at 1G speed
- Requires the use of a standard 1G SFP
  - GLC-T, GLC-SX-MM, GLC-LH-SM, SFP-GE-T, SFP-GE-S, SFP-GE-L (DOM capable SFP are supported)
- Supports for all features at 1G speed other than Unified I/O
  - No FCoE (no 1G Converged Network Adapters are shipping)
  - No Priority Flow Control (standard Pause is available)

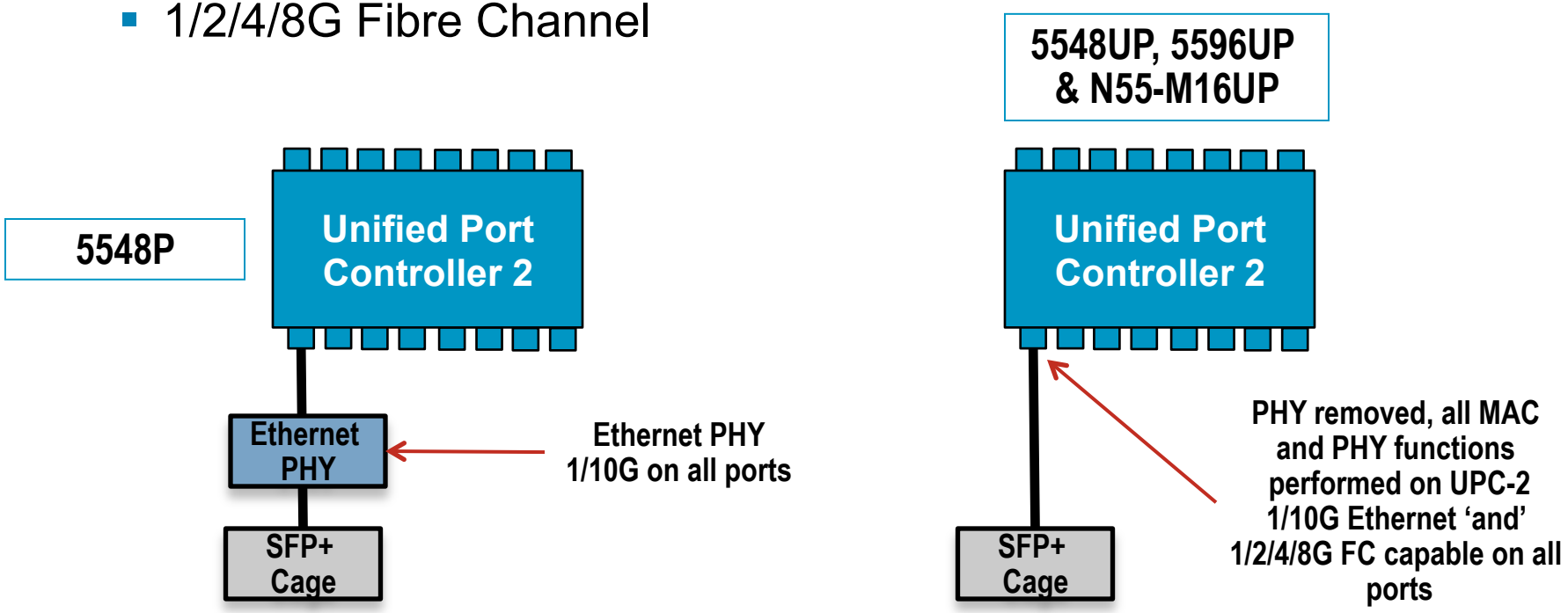
```
interface Ethernet1/3
  switchport access vlan 800
  speed 1000
  channel-group 800
```



Nexus 5500 Hardware Overview

5548UP/5596UP – UPC (Gen-2) and Unified Ports

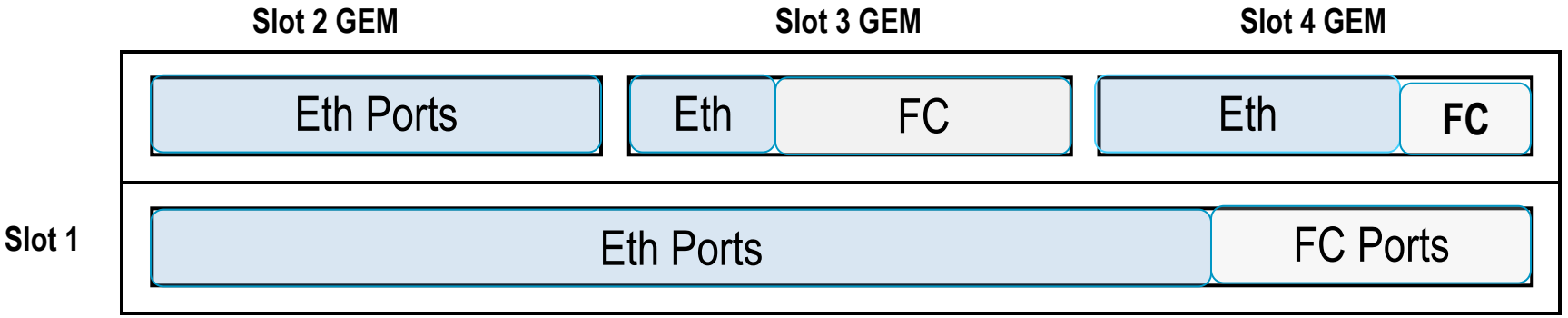
- All versions of 5500 support 1/10G on all ports
- 5548UP, 5596UP and N55-M16UP (Expansion Module) support Unified Port capability on all ports
  - 1G Ethernet Copper/Fibre
  - 10G DCB/FCoE Copper/Fibre
  - 1/2/4/8G Fibre Channel



Nexus 5500 Hardware Overview
5548UP/5596UP – UPC (Gen-2) and Unified Ports

- With the 5.0(3)N1 and later releases each module can define any number of ports as Fibre Channel (1/2/4/8 G) or Ethernet (either 1G or 10G)
- Initial SW releases supports only a continuous set of ports configured as Ethernet or FC within each 'slot'
  - Eth ports have to be the first set and they have to be one contiguous range
  - FC ports have to be second set and they have to be contiguous as well
- Future SW release will support per port dynamic configuration

```
n5k(config)# slot <slot-num>
n5k(config-slot)# port <port-range> type <fc | ethernet>
```



Nexus 5000 & 5500
Reference

For Your Reference

Product Features & Specs	Nexus 5010	Nexus 5020	Nexus 5548P	Nexus 5548UP	Nexus 5596UP
Switch Fabric Throughput	520Gbps	1.04Tbps	960Gbps	960Gbps	1.92Tbps
Switch Footprint	1RU	2RU	1RU	1RU	2RU
1 Gigabit Ethernet Port Density	8	16	48	48	96
10 Gigabit Ethernet Port Density	26	52	48	48	96
8G Native Fibre Channel Port Density	6	12	16	48	96
Port-to-Port Latency	~ 3.2us	~ 3.2us	~2.0us	~1.8us	~ 1.8us
No. of VLANs	512	512	4096	4096	4096
Layer 3 Capability			✔	✔	✔
1 Gigabit Ethernet FEX Port Scalability (L2 mode)	576	576	1152	1152	1152
10 Gigabit Ethernet FEX Port Scalability (L2 mode)	384	384	768	768	768
40 Gigabit Ethernet Capable			✔	✔	✔
Reversed Airflow				✔	✔

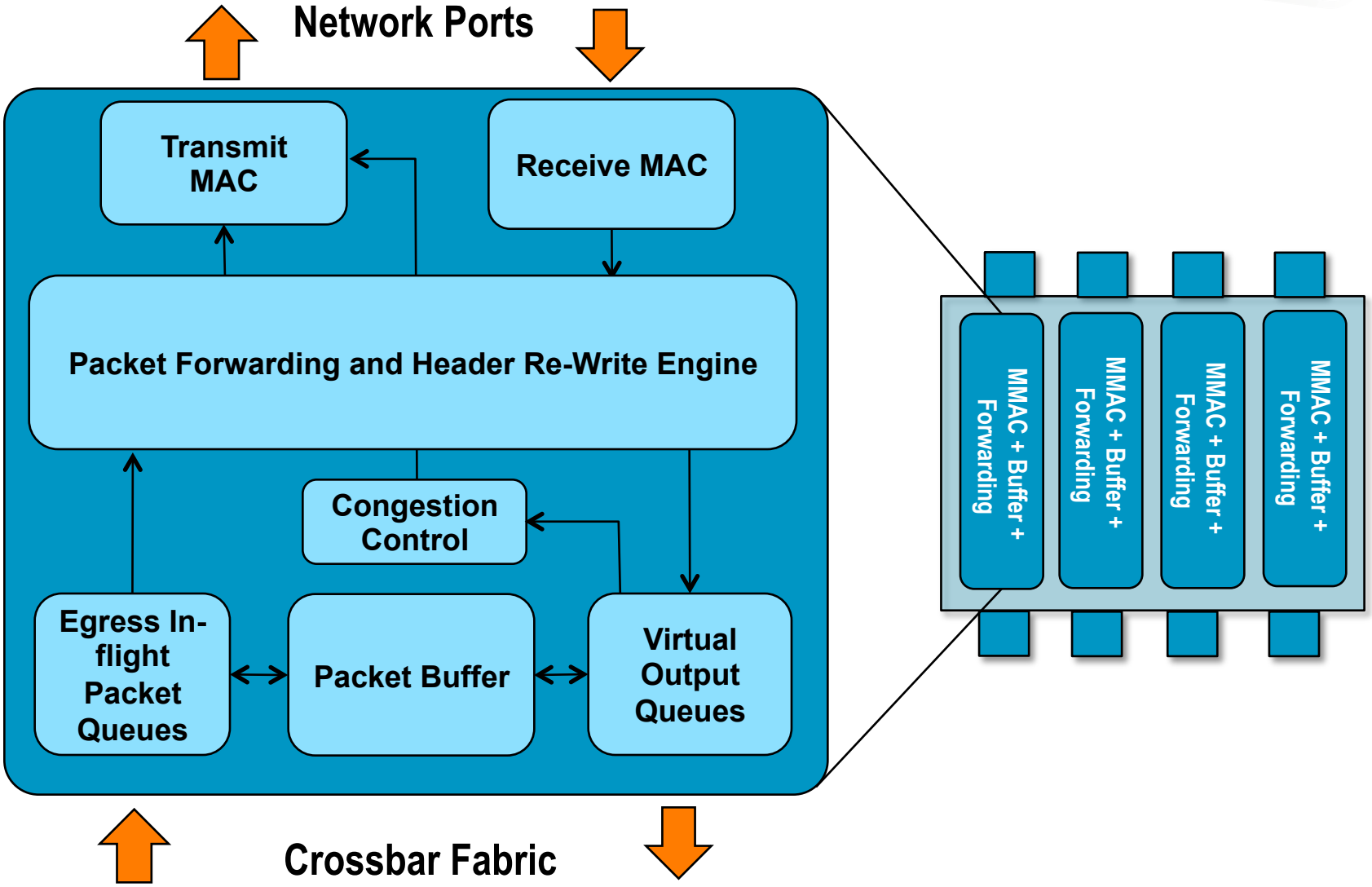
Nexus 5000/5500 and 2000 Architecture

Agenda

- Nexus 5000/5500 Architecture
  - Hardware Architecture
  - Day in the Life of a Packet
  - Layer 3 Forwarding
- Nexus 2000 Architecture
  - FEXLink Architecture
  - FEX Forwarding
  - Extending FEXLink – Adapter FEX
- Nexus 5000/5500
  - Multicast
  - Port Channels
  - QoS

Nexus 5000 & 5500 Packet Forwarding

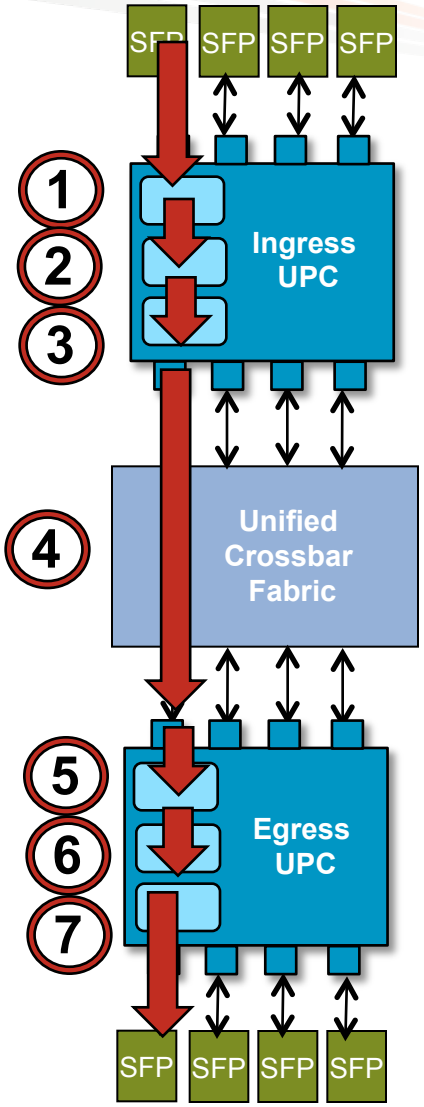
UPC Details



Nexus 5000 & 5500 Packet Forwarding

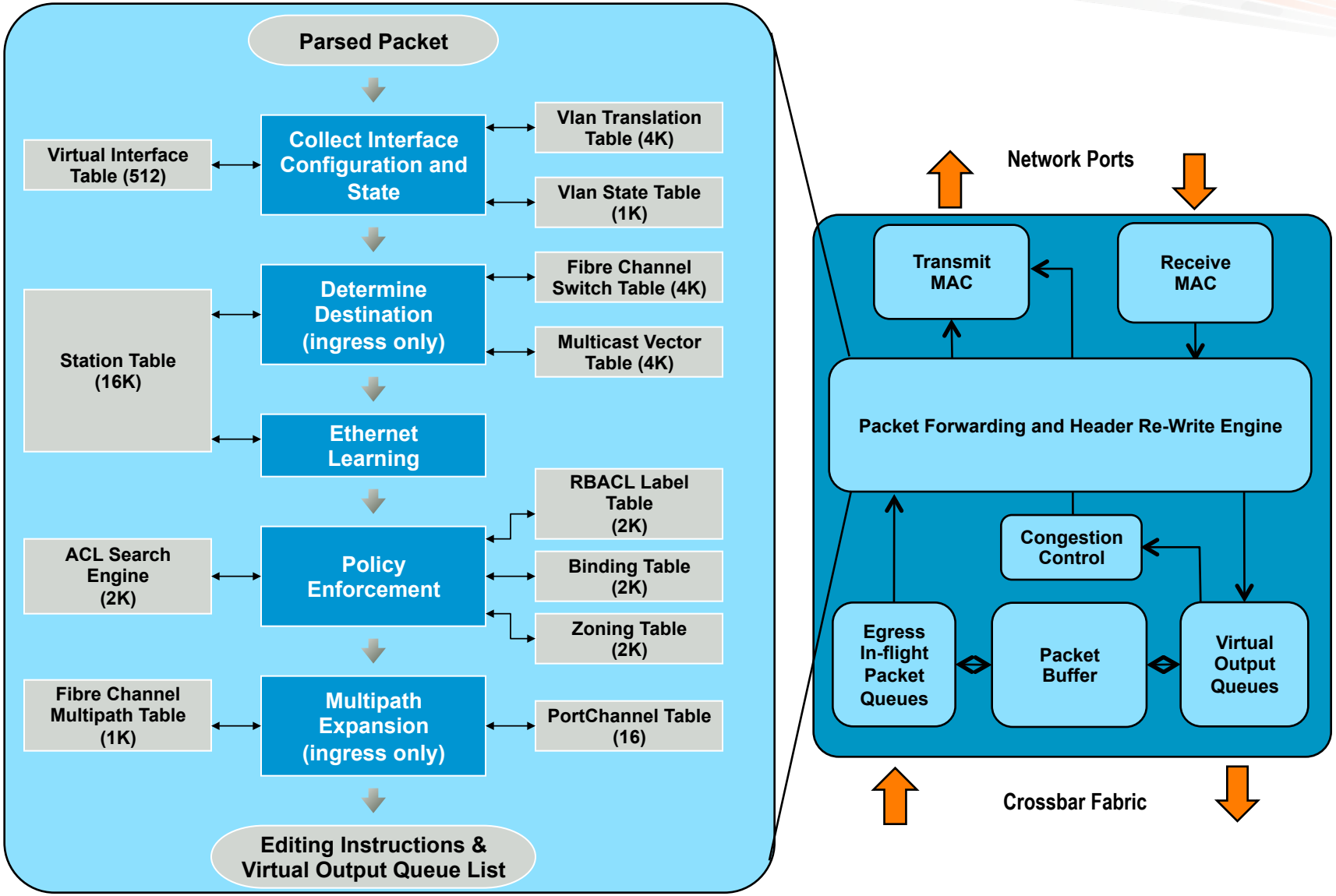
Packet Forwarding Overview

1. Ingress MAC - MAC decoding, MACSEC processing (not supported currently), synchronize bytes
2. Ingress Forwarding Logic - Parse frame and perform forwarding and filtering searches, perform learning apply internal DCE header
3. Ingress Buffer (VoQ) - Queue frames, request service of fabric, dequeue frames to fabric and monitor queue usage to trigger congestion control
4. Cross Bar Fabric - Scheduler determines fairness of access to fabric and determines when frame is de-queued across the fabric
5. Egress Buffers - Landing spot for frames in flight when egress is paused
6. Egress Forwarding Logic - Parse, extract fields, learning and filtering searches, perform learning and finally convert to desired egress format
7. Egress MAC - MAC encoding, pack, synchronize bytes and transmit



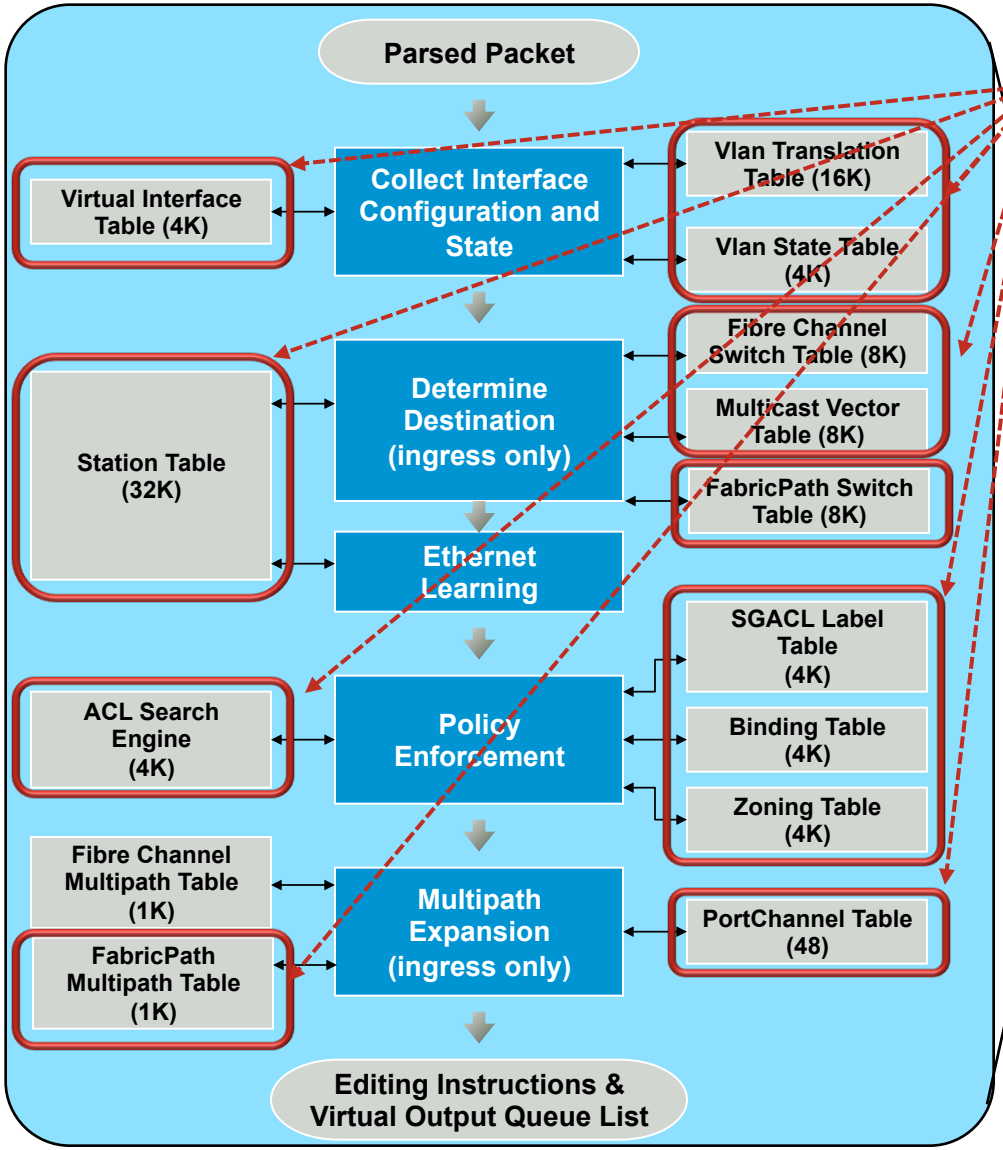
Nexus 5000 Hardware Overview

Nexus 5000 UPC (Gen 1) Forwarding Details

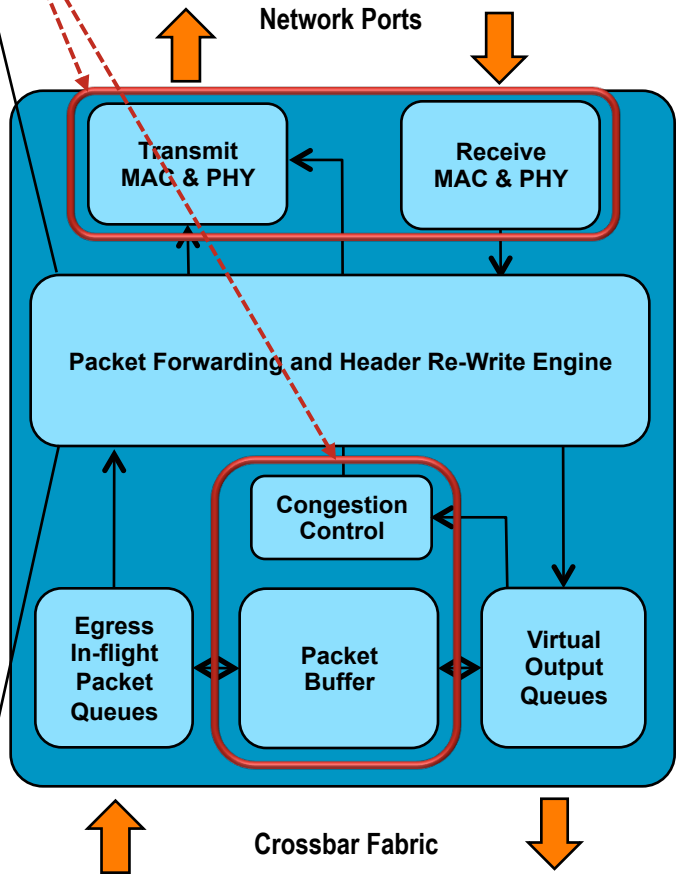


Nexus 5500 Hardware Overview

Nexus 5500 UPC (Gen-2) Forwarding Details

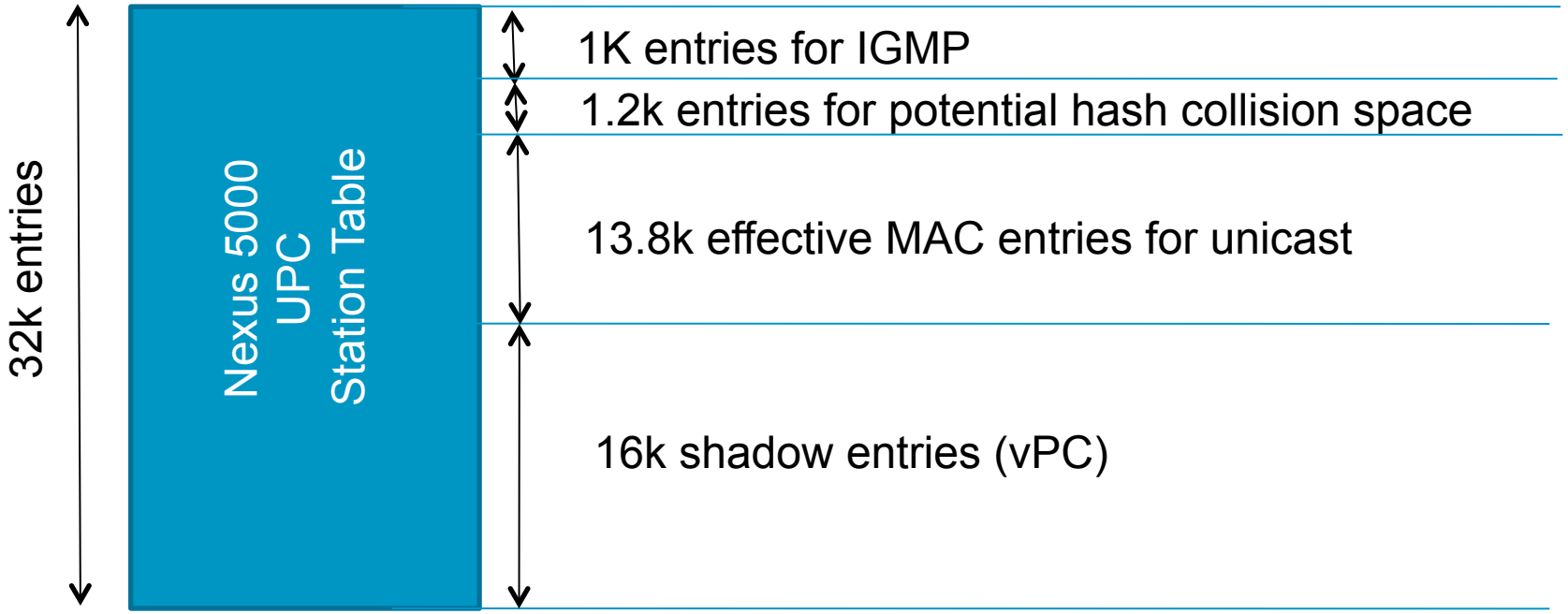


- Changes in the 5500 UPC-2
- Larger tables and more functions



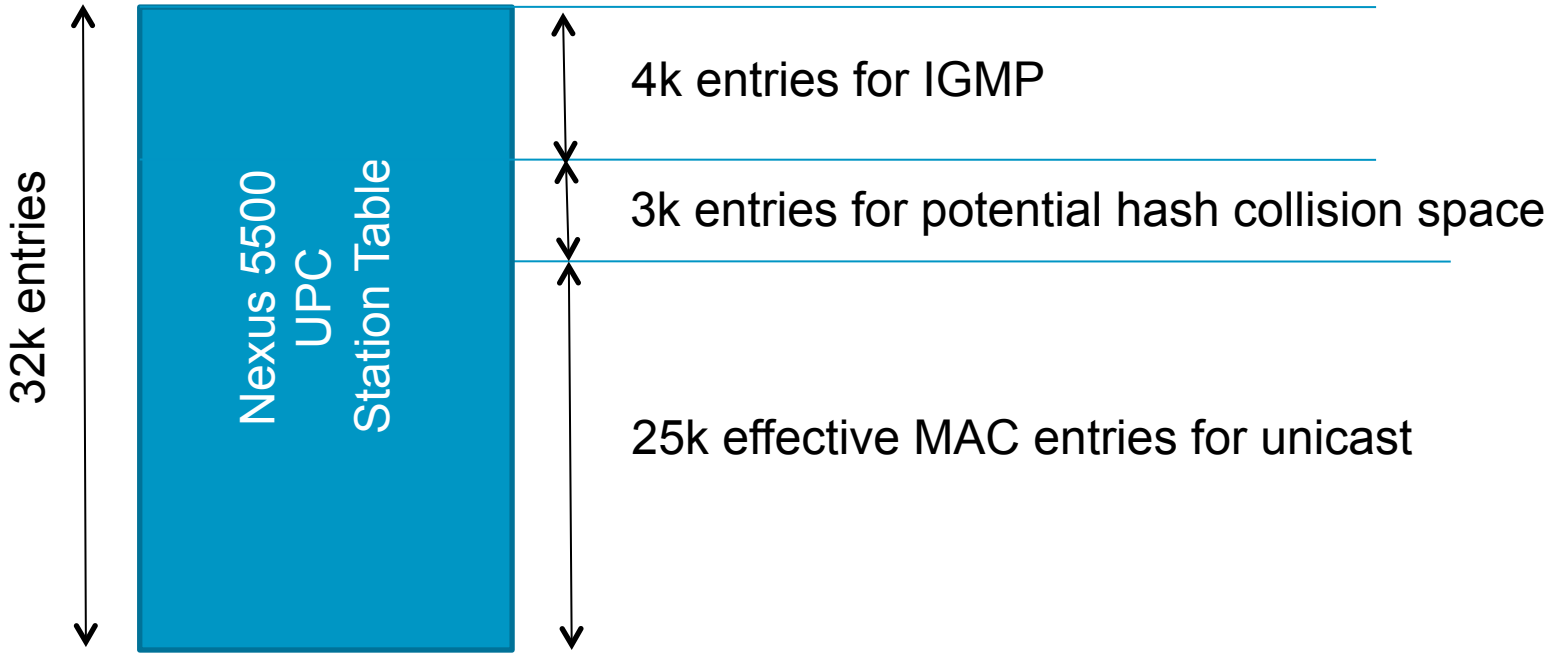
Nexus 5000
Station (MAC) Table allocation

- Nexus 5000 has a 32K Station table entries
  - 16K shadow entries used for vPC
  - 1k reserved for multicast (Multicast MAC addresses)
  - 1.2k assumed for hashing conflicts (conservative)
  - 13.8k effective Layer 2 unicast MAC address entries



Nexus 5500
Station (MAC) Table allocation

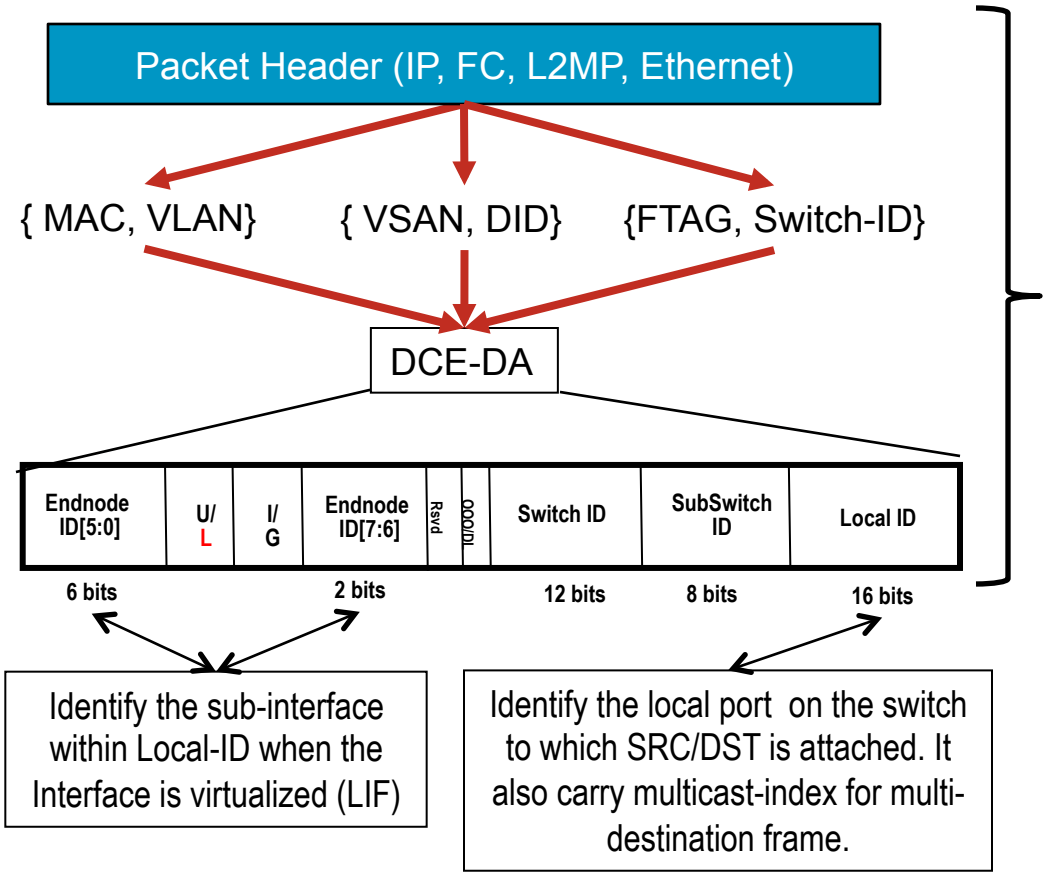
- Nexus 5500 has a 32K Station table entries
  - 4k reserved for multicast (Multicast MAC addresses)
  - 3k assumed for hashing conflicts (very conservative)
  - 25k effective Layer 2 unicast MAC address entries

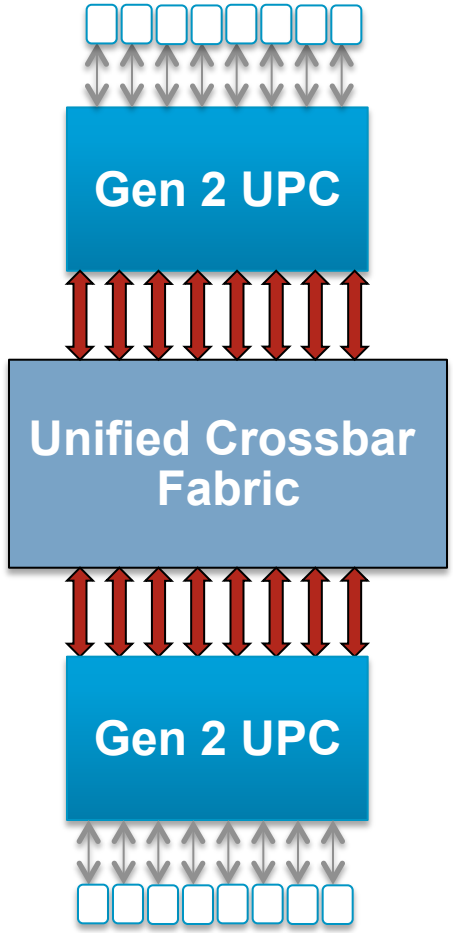


Nexus 5000 & 5500 Packet Forwarding

DCE – Internal Nexus 5000/5500 Forwarding Header

- All frames forwarded internally using Cisco DCE Header after parsing the packet header

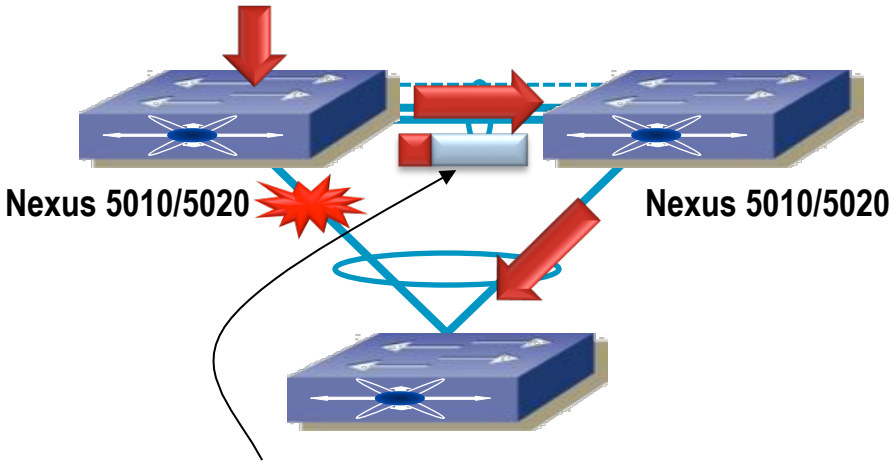




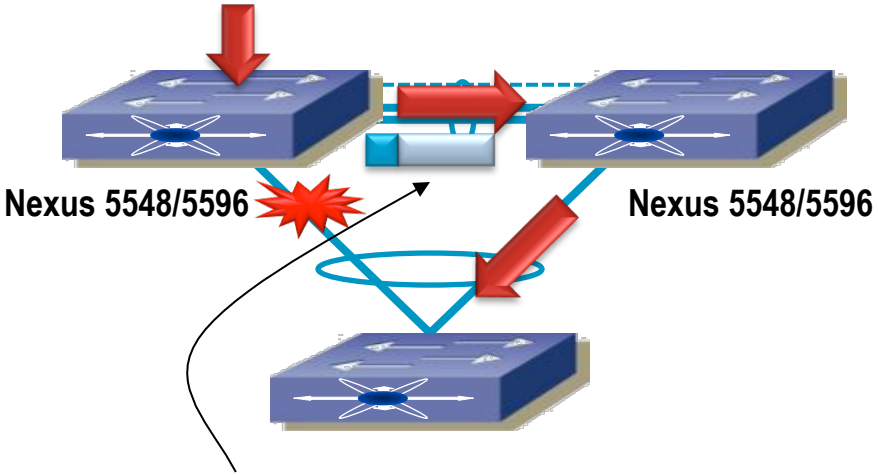
Nexus 5000 & 5500 Packet Forwarding

vPC peer-link 5000/5500 Forwarding

- Nexus 5000 uses a different mechanism to identify vPC forwarded frames sent across the vPC peer-link
- Nexus 5010/5020 leverages a shadow VLAN and MAC address to identify ‘vPC’ frames received on the peer switch to prevent looping frames
- Nexus 5548/5596 leverages a DCE header to identify the vPC topology for each frame to prevent looping frames
- Nexus 5000 and 5500 can ***not*** be configured as vPC peers



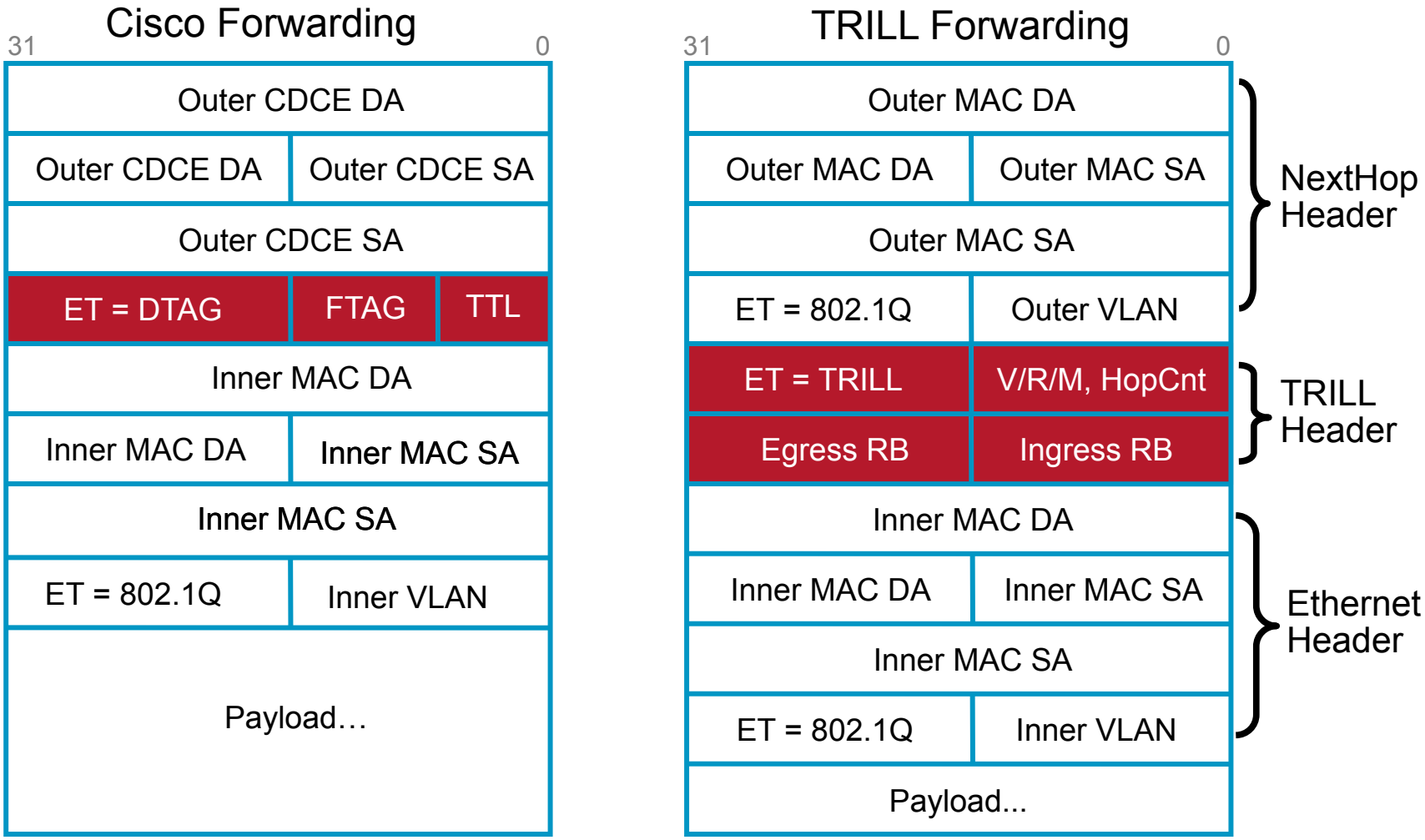
5000 leverages .1q tagging on vPC peer-link



5500 leverages DCE framing on vPC peer-link

Nexus 5500 FabricPath

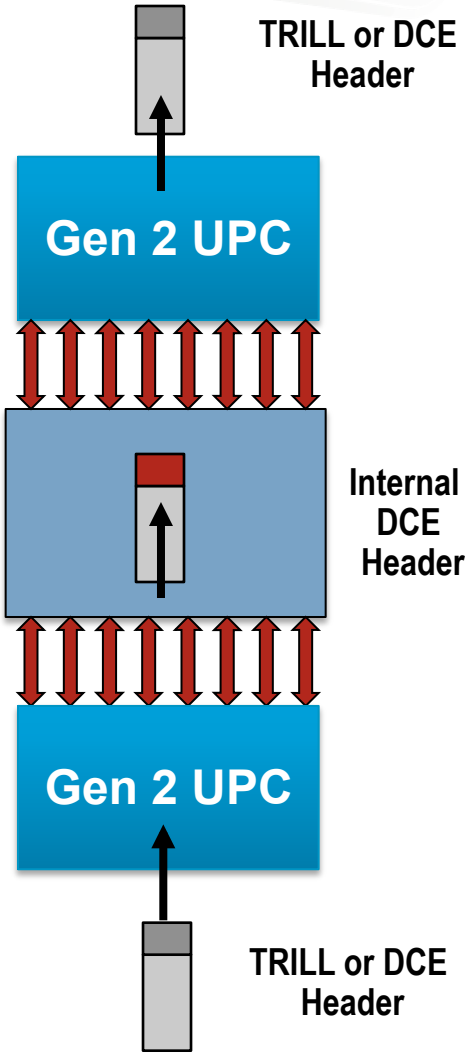
Standards Based + Cisco Extensions



Nexus 5500 FabricPath

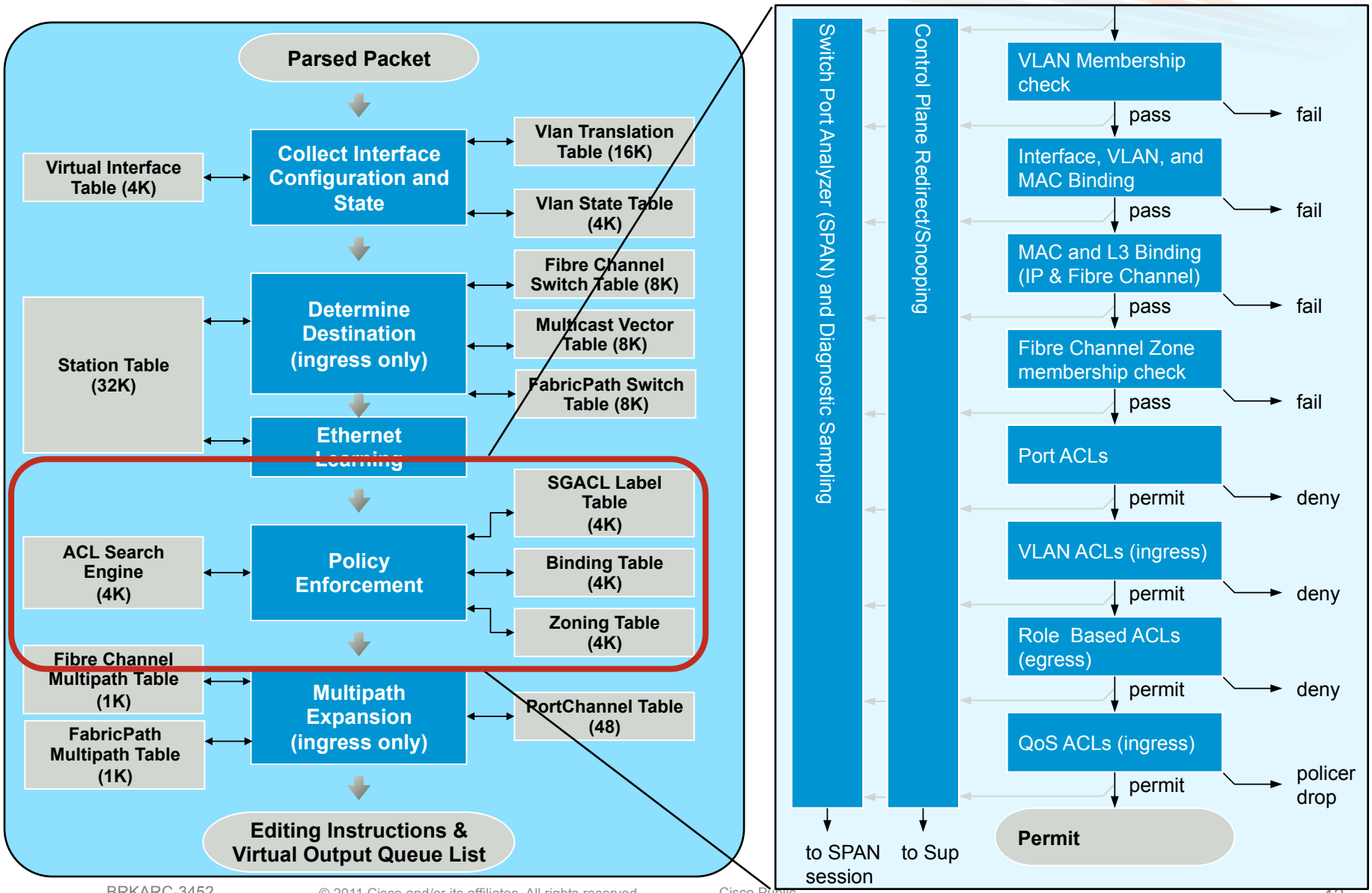
Standards Based + Cisco Extensions

- Nexus 5500 supports two modes of mac-in-mac encapsulation for FabricPath
  - TRILL (IETF standard)
  - DCE (Cisco Pre-Standard)
- Support for TRILL and CTRILL (Cisco extension to TRILL)
- Advertise up to 64 TRILL RBridge addresses ‘or’ 64 DCE Switch IDs
- Support 8K TRILL forwarding entries ‘or’ 8K DCE forwarding entries
- Support up to 16 equal cost forwarding path
- Support shared/source based multicast tree
- DCE Switch ID can be assigned at the LIF level
- Nexus 5500 can be configured to use either TRILL or DCE encapsulation mode on switch to switch links



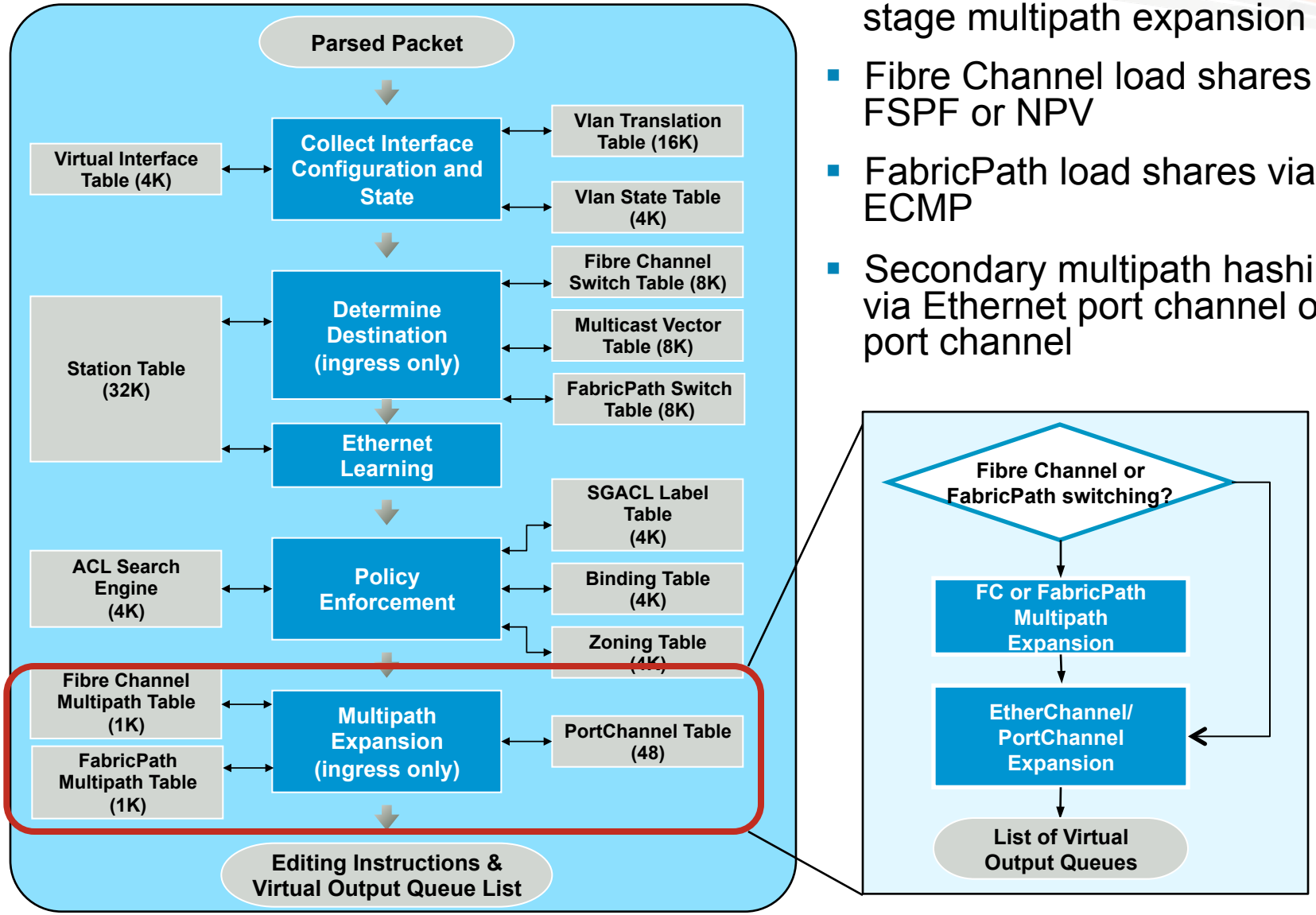
Nexus 5000 & 5500 Packet Forwarding

UPC Policy Enforcement



Nexus 5000 & 5500 Packet Forwarding

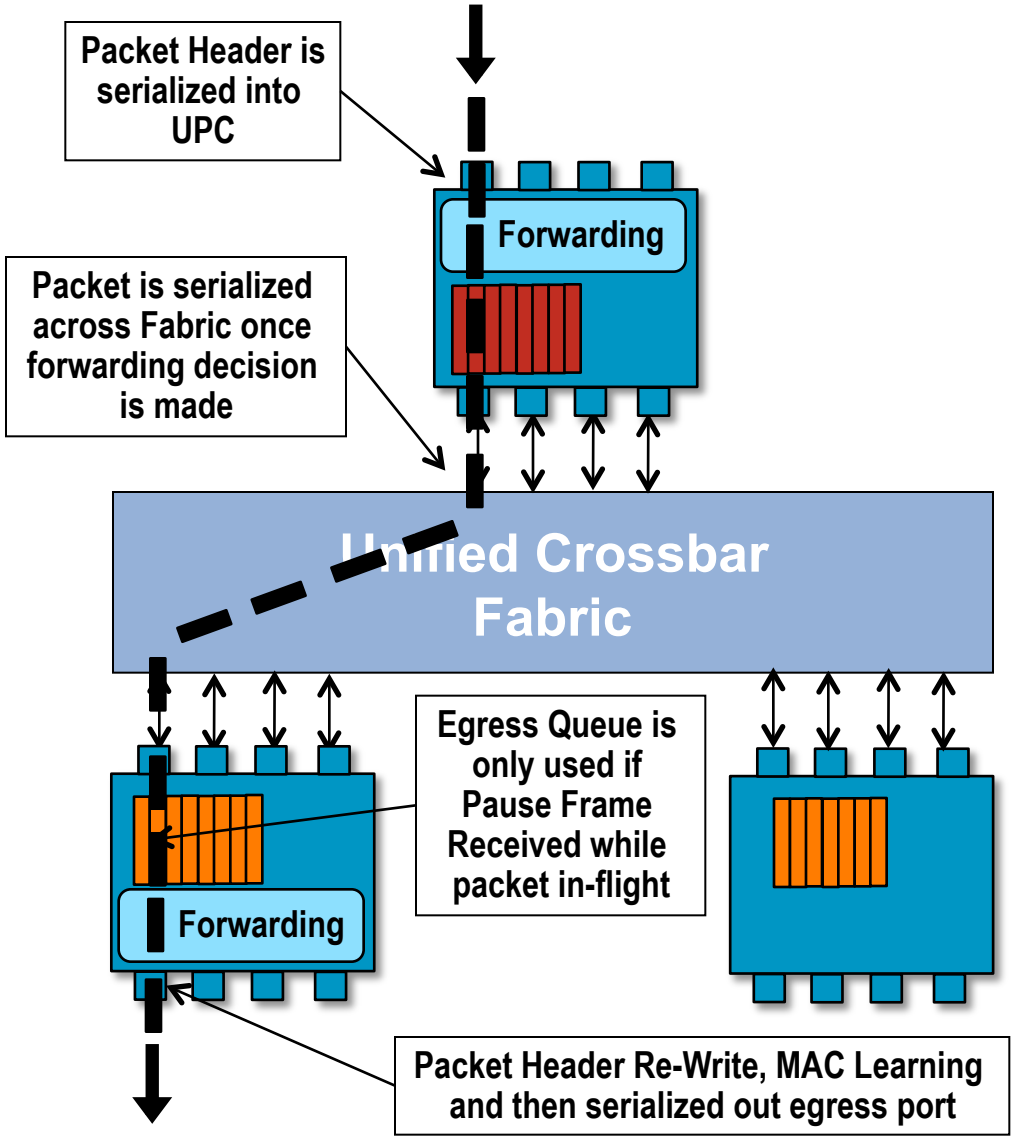
UPC Multipath Expansion



- Nexus 5000/5500 utilizes a two stage multipath expansion
- Fibre Channel load shares via FSPF or NPV
- FabricPath load shares via ECMP
- Secondary multipath hashing via Ethernet port channel or FC port channel

Nexus 5000 & 5500 Packet Forwarding

Packet Forwarding—Cut Thru Switching

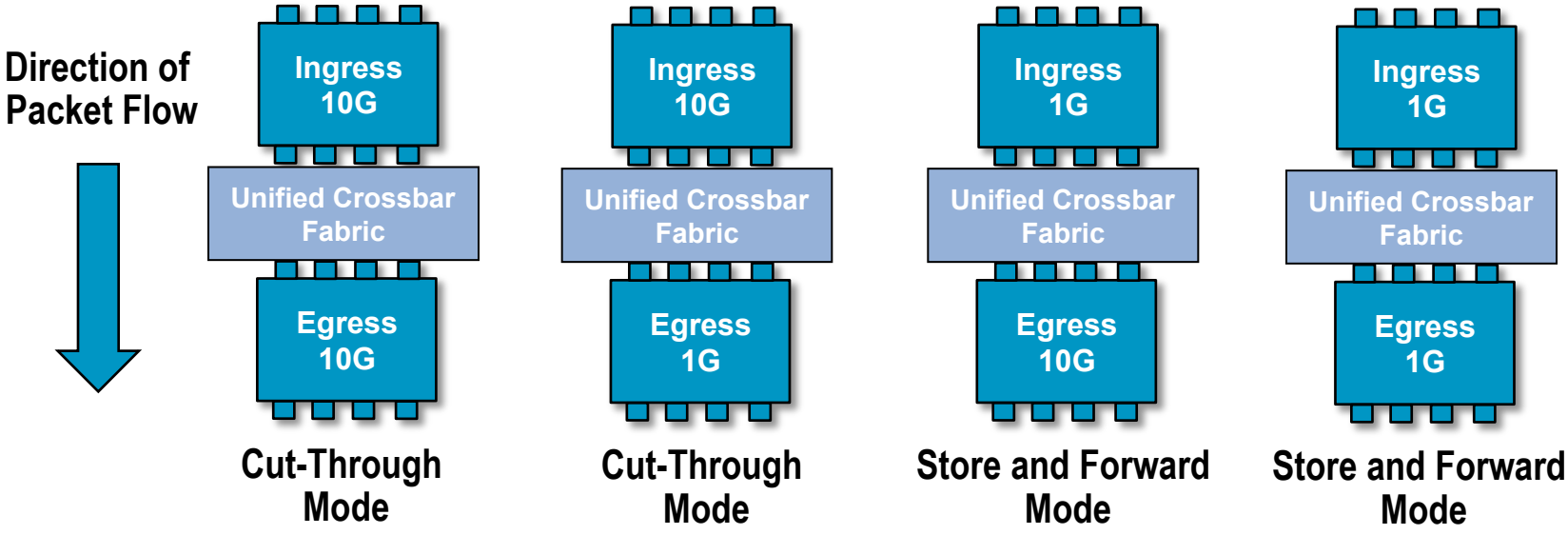


- Nexus 5000 & 5500 utilize a Cut Thru architecture when possible
- Bits are serialized in from the ingress port until enough of the packet header has been received to perform a forwarding and policy lookup
- Once a lookup decision has been made and the fabric has granted access to the egress port bits are forwarded through the fabric
- Egress port performs any header rewrite (e.g. CoS marking) and MAC begins serialization of bits out the egress port

Nexus 5000 & 5500 Packet Forwarding

Packet Forwarding—Cut-Through Switching

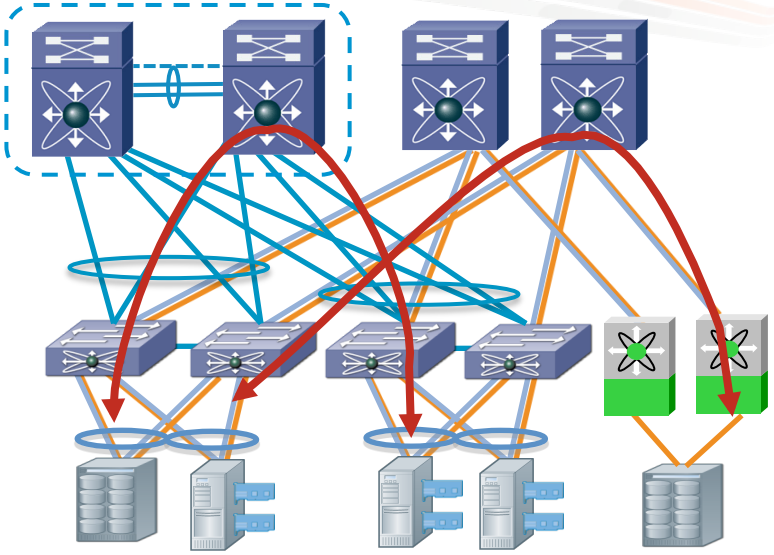
- Nexus 5000 and 5500 utilize both cut-through and store and forward switching
- Cut-through switching can only be performed when the ingress data rate is equivalent ***or*** faster than the egress data rate
- The X-bar fabric is designed to forward 10G packets in cut-through which requires that 1G to 1G switching also be performed in store and forward mode

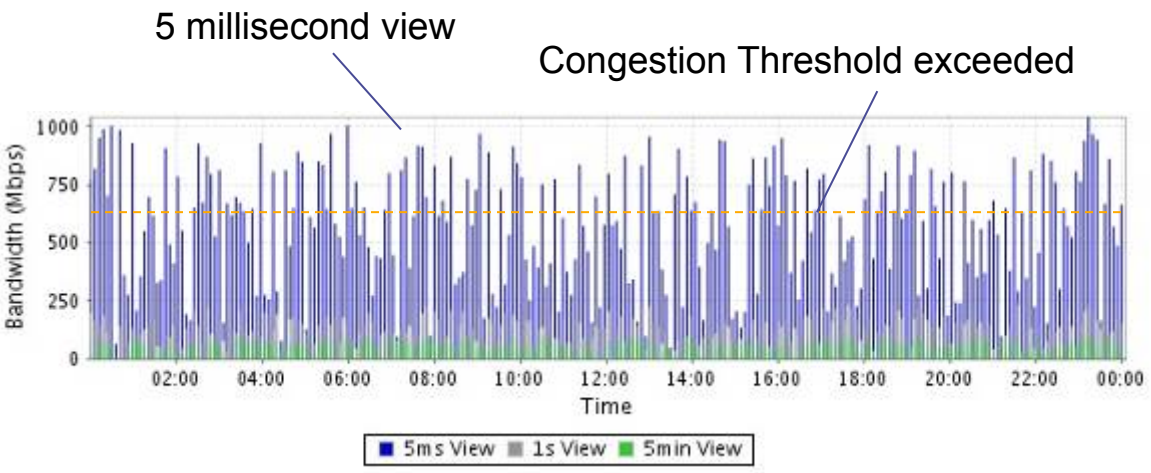


Data Center Architecture

Minimizing Latency ‘and’ Loss

- Why Cut-Through Switching?
- It is only one variable in overall fabric optimization
- Designs target consistency of performance under variable conditions
- A balanced fabric is a function of maximal throughput ‘and’ minimal loss => “Goodput”





Data Center Design Goal: Optimizing the balance of end to end fabric latency with the ability to absorb traffic peaks and prevent any associated traffic loss

Nexus 5000 & 5500 Packet Forwarding

For Your Reference

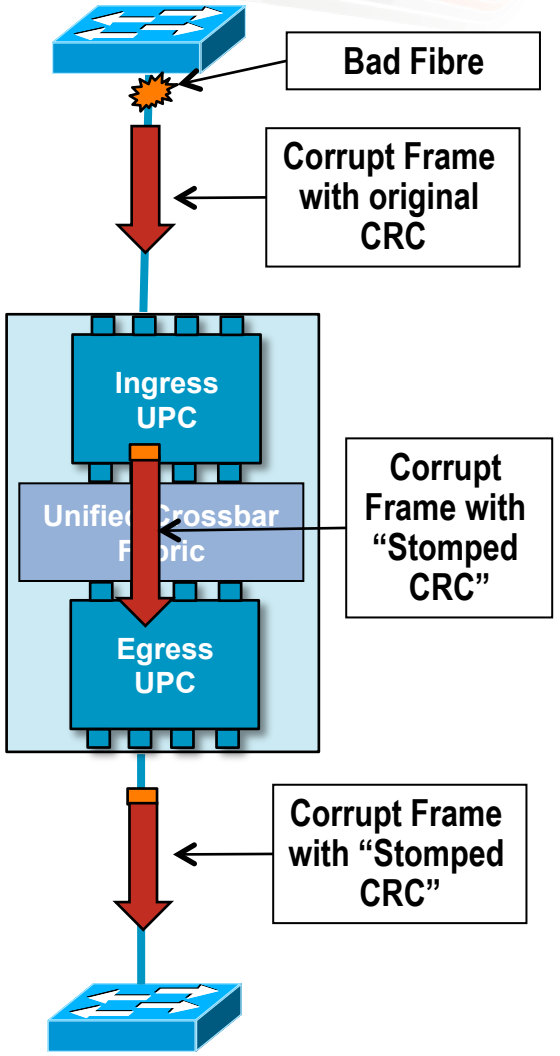
Forwarding Mode Behavior (Cut-Through or Store and Forward)

Source Interface	Destination Interface	Switching Mode
10 GigabitEthernet	10 GigabitEthernet	Cut-Through
10 GigabitEthernet	1 GigabitEthernet	Cut-Through
1 GigabitEthernet	1 GigabitEthernet	Store-and-Forward
1 GigabitEthernet	10 GigabitEthernet	Store-and-Forward
FCoE	Fibre Channel	Cut-Through
Fibre Channel	FCoE	Store-and-Forward
Fibre Channel	Fibre Channel	Store-and-Forward
FCoE	FCoE	Cut-Through

Nexus 5000 & 5500 Packet Forwarding

Packet Forwarding - Cut Through Switching

- In Cut-Through switching frames are not dropped due to bad CRC
- Nexus 5000/5500 implements a CRC ‘stomp’ mechanism to identify frames that have been detected with a bad CRC upstream
- A packet with a bad CRC is “stomped”, by replacing the “bad” CRC with the original CRC exclusive-OR’d with the STOMP value (a 1’s inverse operation on the CRC)
- In Cut Through switching frames with invalid MTU (frames with a larger MTU than allowed) are not dropped
- Frames with a “> MTU” length are truncated and have a stomped CRC included in the frame



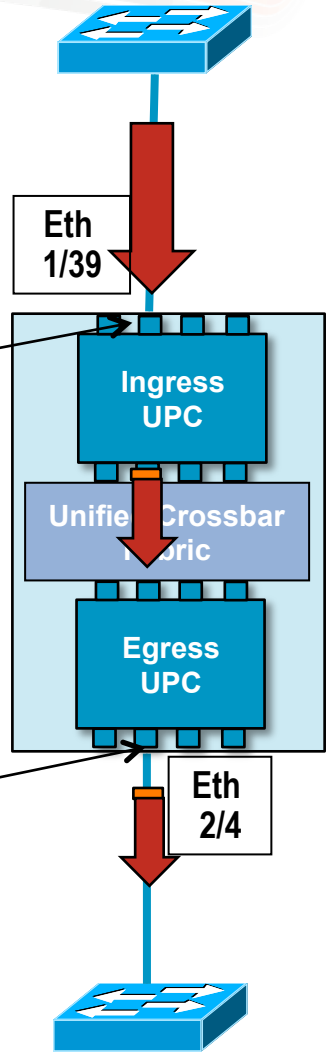
Nexus 5000 & 5500 Packet Forwarding

Packet Forwarding—Cut Through Switching

- Corrupt or Jumbo frames arriving inbound will count against the Rx Jumbo or CRC counters
- Corrupt or Jumbo frames will be identified via the Tx output error and Jumbo counters

```
dc11-5020-4# sh int eth 1/39
<snip>
 RX
    576 unicast packets  4813153 multicast packets  55273 broadcast packets
    4869002 input packets  313150983 bytes
    31 jumbo packets  0 storm suppression packets
    0 runts  0 giants  0 CRC  0 no buffer
    0 input error  0 short frame  0 overrun   0 underrun  0 ignored
    0 watchdog  0 bad etype drop  0 bad proto drop  0 if down drop
    0 input with dribble  0 input discard
    0 Rx pause
```

```
dc11-5020-4# sh int eth 2/4
<snip>
 TX
    112 unicast packets  349327 multicast packets  56083 broadcast packets
    405553 output packets  53600658 bytes
    31 jumbo packets
    31 output errors  0 collision  0 deferred  0 late collision
    0 lost carrier  0 no carrier  0 babble
    0 Tx pause
```



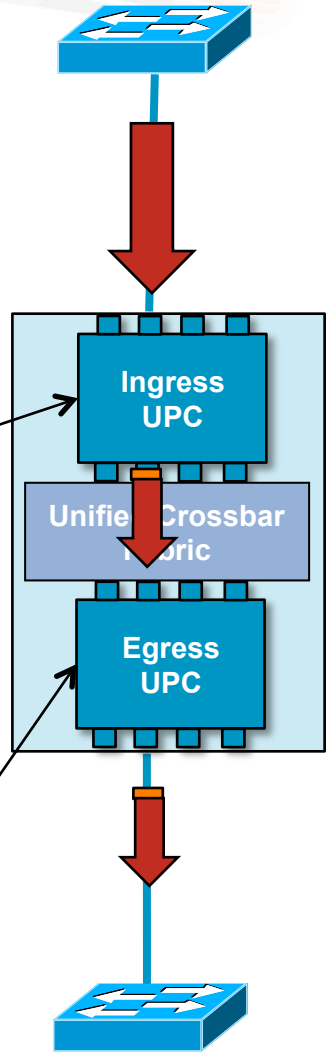
Nexus 5000 & 5500 Packet Forwarding

Packet Forwarding—Cut Thru Switching

- CRC and ‘stomped’ frames are tracked internally between ASIC’s within the switch as well as on the interface to determine internal HW errors are occurring

```
dc11-5020-4# sh hardware internal gatos asic 2 counters interrupt
<snip>
Gatos 2 interrupt statistics:
Interrupt name                                  |Count    |ThresRch|ThresCnt|Ivls
------------------------------------------------+--------+--------+--------+----
<snip>
gat_bm_port0_INT_err_ig_mtu_vio                 |1f       |0       |1f
<snip>
```

```
dc11-5020-4# sh hardware internal gatos asic 13 counters interrupt
<snip>
Gatos 13 interrupt statistics:
Interrupt name                                  |Count    |ThresRch|ThresCnt|Ivls
------------------------------------------------+--------+--------+--------+----
<snip>
gat_fw2_INT_eg_pkt_err_cb_bm_eof_err            |1f       |0       |1       |0
gat_fw2_INT_eg_pkt_err_eth_crc_stomp            |1f       |0       |1       |0
gat_fw2_INT_eg_pkt_err_ip_pyld_len_err          |1f       |0       |1       |0
gat_mm2_INT_rlp_tx_pkt_crc_err                  |1f       |0       |1       |0
<snip>
```



Note: Please see session BRKCRS-3145 (Troubleshooting the Cisco Nexus 5000 / 2000 Series Switches) for more information on this type of troubleshooting

Nexus 5000 & 5500 Packet Forwarding

For Your Reference

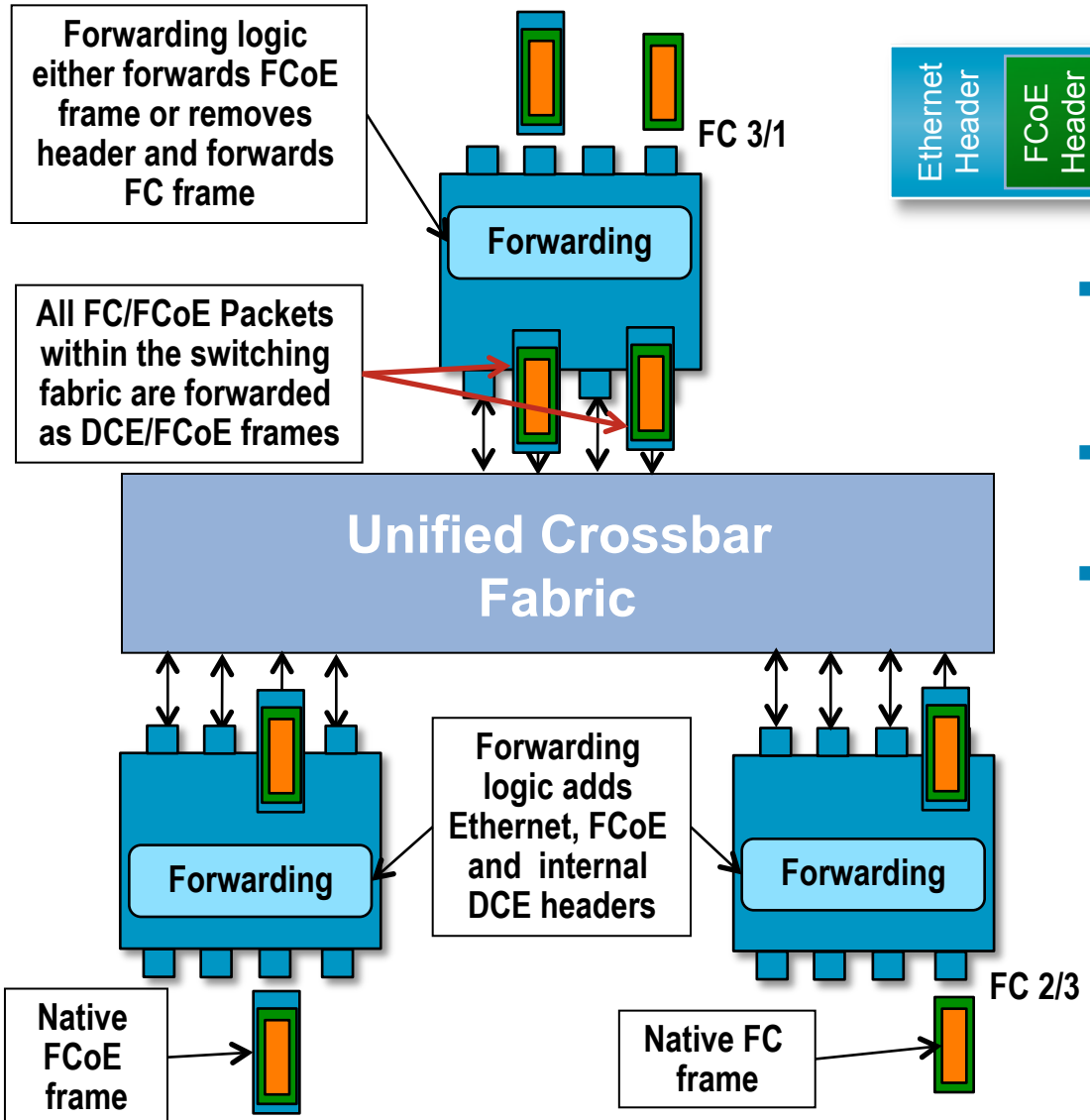
CRC Behavior for Cut-Thru Frames

- The table below indicates the forwarding behavior for a corrupt packet (CRC error) arriving on a port operating in cut-through mode

Source Interface Type	Destination Interface Type	Action
10GE/DCE/FCoE	10GE/DCE/FCoE	The CRC frame is transmitted as is
10GE/DCE/FCoE	Native Fibre Channel	The FC CRC is stomped. Also the frame is transmitted with EOFa
Native Fibre Channel	Native Fibre Channel	The FC CRC is stomped. Also the frame is transmitted with EOFa
Native Fibre Channel	10GE/DCE/FCoE	The FC CRC is stomped. Also the frame is transmitted with EOFa. Also the Ethernet CRC is stomped

Nexus 5000 & 5500 Packet Forwarding

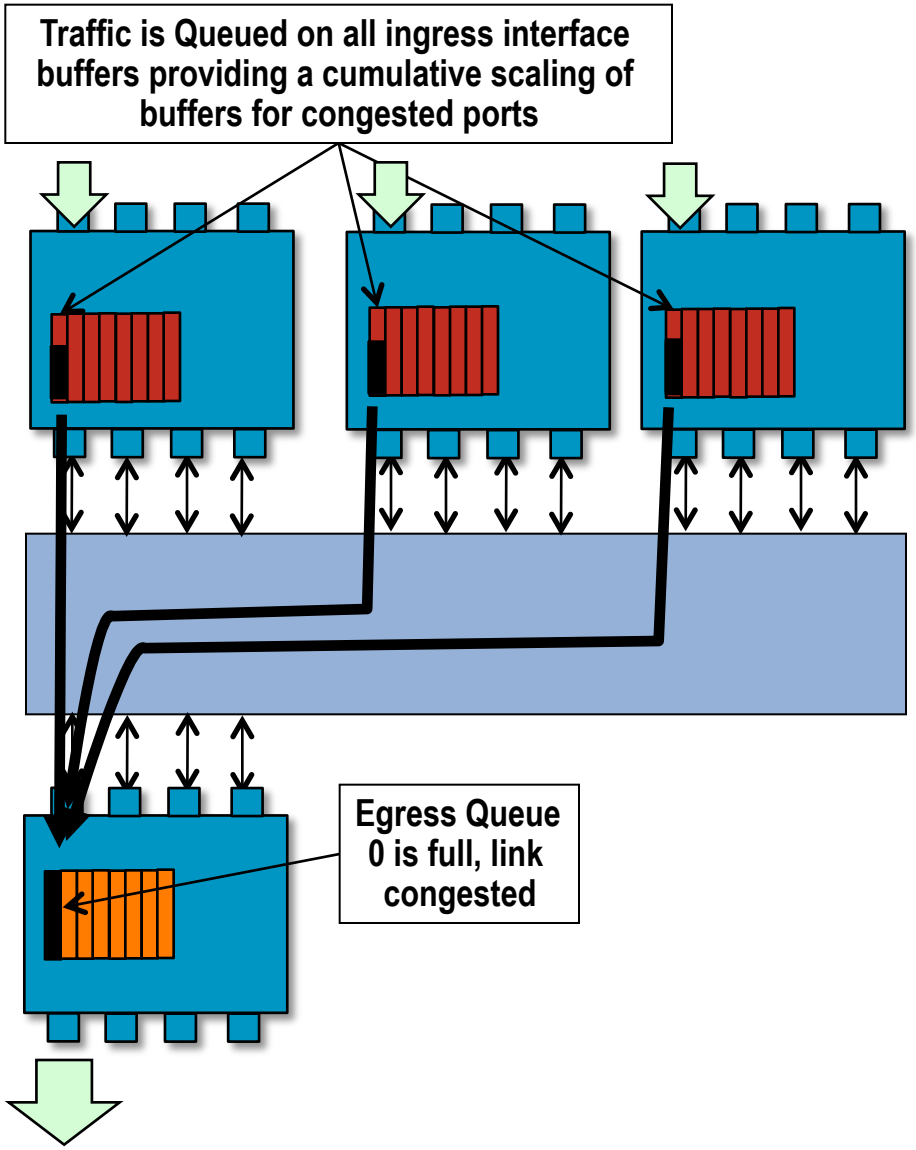
Packet Forwarding—Fibre Channel and FCoE



- Nexus 5000 and 5500 operate as both an Ethernet switch and a Fibre Channel switch
- Supports native FC as well as FCoE interfaces
- Internally within the switching fabric all Fibre Channel frames are forwarded as DCE/FCoE frames
  - FC to FCoE
  - FC to FC
  - FCoE to FC
  - FCoE to FCoE

Nexus 5000 & 5500 Packet Forwarding

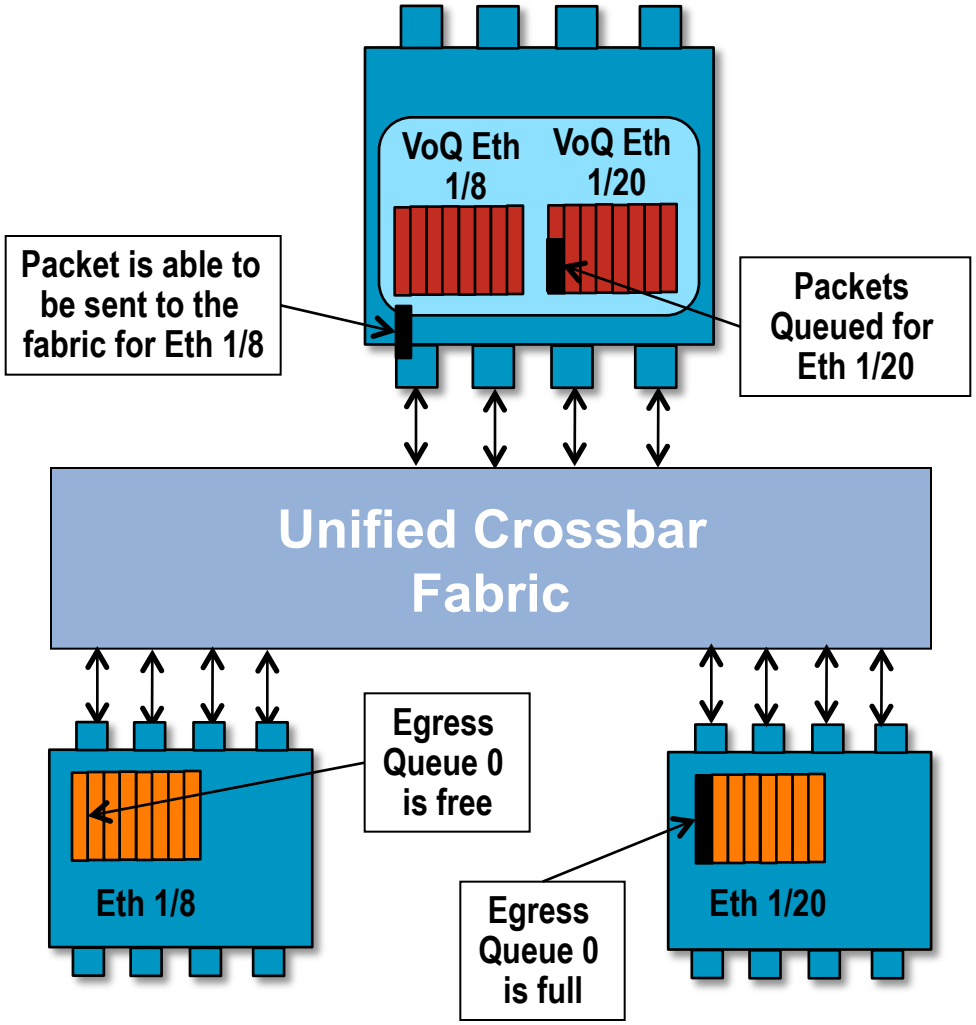
Packet Forwarding—Ingress Queuing



- In typical Data Center access designs multiple ingress access ports transmit to a few uplink ports
- Nexus 5000 and 5500 utilize an ***Ingress*** Queuing architecture
- Packets are stored in ingress buffers until egress port is free to transmit
- Ingress queuing provides an additive effective
- *The total queue size available is equal to [number of ingress ports x queue depth per port]*
- Statistically ingress queuing provides the same advantages as shared buffer memory architectures

Nexus 5000 & 5500 Packet Forwarding

Packet Forwarding—Virtual Output Queues



- Nexus 5000 and 5500 use an 8 Queue QoS model for unicast traffic
- Traffic is Queued on the Ingress buffer until the egress port is free to transmit the packet
- To prevent Head of Line Blocking (HOLB) Nexus 5000 and 5500 use a Virtual Output Queue (VoQ) Model
- Each ingress port has a unique set of 8 virtual output queues for every egress port (1024 Ingress VOQs = 128 destinations * 8 classes on every ingress port)
- If Queue 0 is congested for any port traffic in Queue 0 for all the other ports is still able to be transmitted
- Common shared buffer on ingress, VoQ are pointer lists and not physical buffers

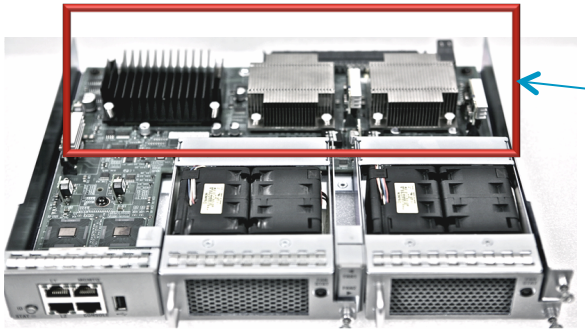
Nexus 5000/5500 and 2000 Architecture

Agenda

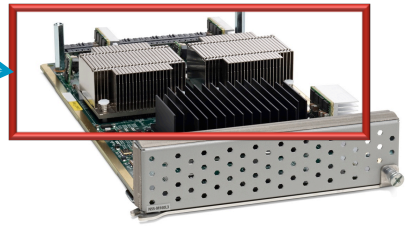
- Nexus 5000/5500 Architecture
  - Hardware Architecture
  - Day in the Life of a Packet
  - Layer 3 Forwarding
- Nexus 2000 Architecture
  - FEXLink Architecture
  - FEX Forwarding
  - Extending FEXLink – Adapter FEX
- Nexus 5000/5500
  - Multicast
  - Port Channels
  - QoS

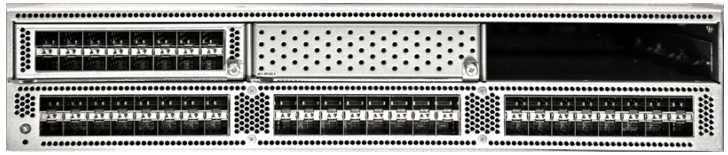
Nexus 5500 Series

Nexus 5500 with Layer 3 support



160Gbps (240Mpps)
Layer 3 processing





1) Remove Fans
2) Replace Daughtercard with L3 enabled daughtercard
3) Install License and enabled NX-OS features

1) Install L3 Expansion Module(s)
2) Install License and enabled NX-OS features

Nexus 5548P/UP

- Ordered with L3 daughtercard installed or order a FRU for an L2 5548
- Daughtercard can be replaced while in the track

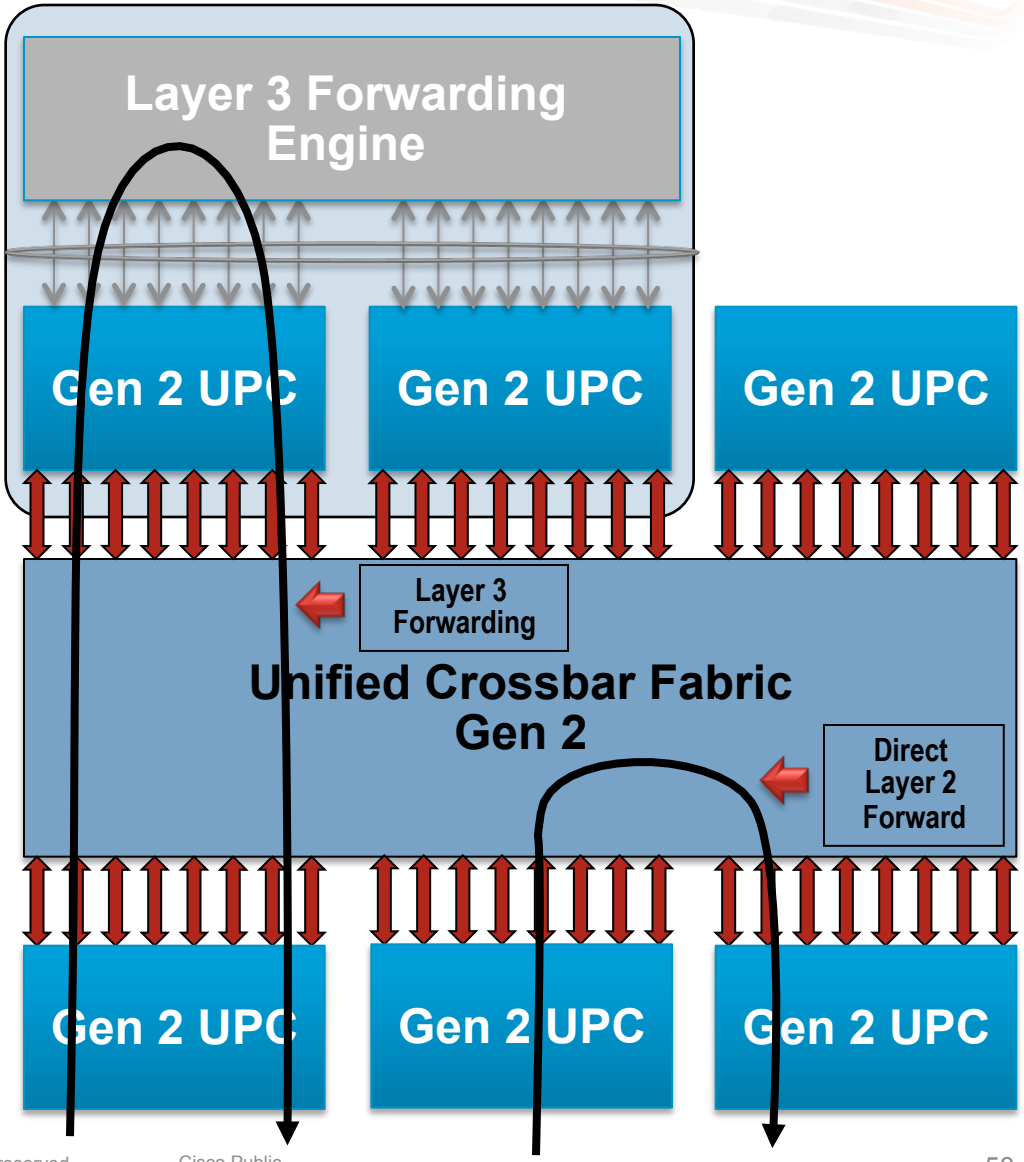
Nexus 5596UP

- At FCS one Layer 3 Expansion Module
- Support for OIR of Layer 3 Expansion Module and/or up to three Layer 3 Expansion Modules (Future)

Nexus 5500 Series

Nexus 5500 with Layer 3 support

- Layer 3 Forwarding Engine connects the the X-Bar via two UPC (160GBps)
- Optional two stage forwarding
- Stage 1 – Ingress UPC forwards to destination MAC address
- If MAC address is external packet directly forwarded to egress port across X-Bar fabric (single stage only)
- If MAC address is the router MAC address (e.g. HSRP vmac) packet is forwarded across fabric to layer 3 Engine
- Stage 2 – Layer 3 lookup occurs and packet is forwarded to egress port across X-Bar fabric
- Only 'routed' packets are forwarded through the Layer 3 engine



Nexus 5500 Series

Nexus 5500 with Layer 3 support

- A single NX-OS CLI is used to configure, manage and troubleshoot the 5500 for ***all*** protocols (vPC, STP, OSPF, FCoE, …)
- There is ***'NO'*** need to manage the Layer 3 ASIC directly (no 'session 15' interface is required)
- Routing Protocols are consistently configured across all layer 3 enabled NX-OS switches (Nexus 7000, Nexus 5500, Nexus 3000)
- Interfaces supported for Layer 3
  - L3 routed interface (non-FEX ports)
  - L3 sub-interface
  - SVI (FEX ports could be members of VLANs)
  - Port channels

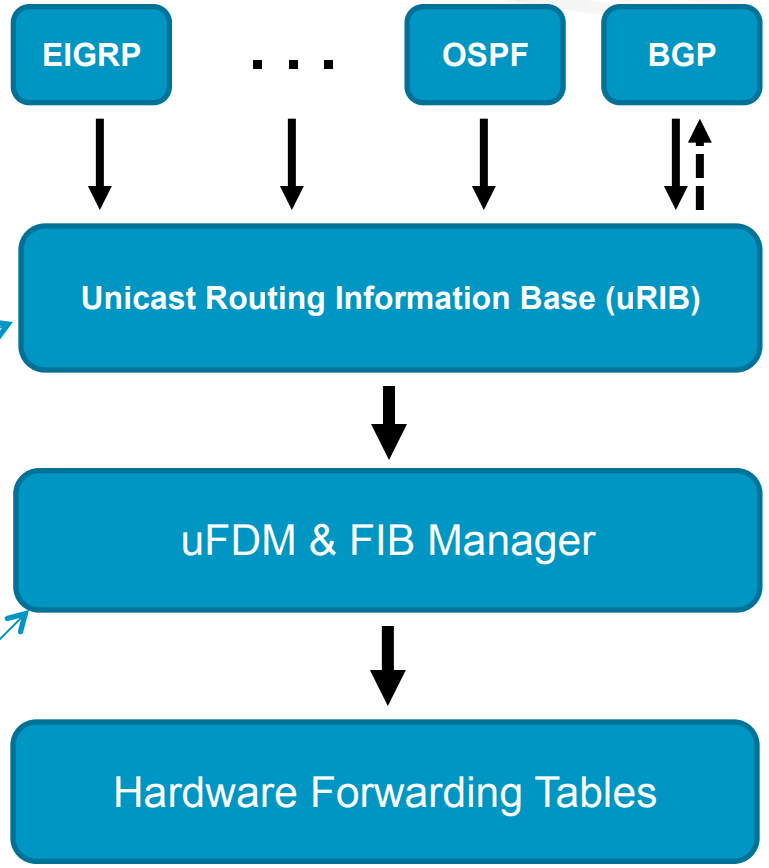
```
L3-5548-1# sh run ospf

!Command: show running-config ospf
!Time: Fri Mar 25 14:21:05 2011

version 5.0(3)N1(1)
feature ospf

router ospf 1
  router-id 100.100.100.1
  area 0.0.0.0 authentication message-digest
  log-adjacency-changes
router ospf 100
  graceful-restart helper-disable
router ospf 2

interface Vlan10
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0

interface Vlan20
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0

interface Vlan100
  ip ospf authentication-key 3 9125d59c18a9b015
  ip ospf cost 4
  ip ospf dead-interval 4
  ip ospf hello-interval 1
  ip router ospf 1 area 0.0.0.0
```

Nexus 5500 Series

Nexus Unicast Routing

- NX-OS software & hardware architecture consistent between Nexus 5500 and Nexus 7000

```
L3-5548-1# sh ip route
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]

10.1.1.0/24, ubest/mbest: 1/0, attached
    *via 10.1.1.1, Vlan10, [0/0], 3d00h, direct
10.1.1.1/32, ubest/mbest: 1/0, attached
    *via 10.1.1.1, Vlan10, [0/0], 3d00h, local

L3-5548-1# sh forwarding route

IPv4 routes for table default/base

------------------+------------------+---------------
Prefix            | Next-hop         | Interface
------------------+------------------+---------------
10.1.1.0/24         Attached           Vlan10
10.1.1.0/32         Drop               Null0
10.1.1.1/32         Receive            sup-eth1
10.1.1.2/32         10.1.1.2           Vlan10
10.1.1.255/32       Attached           Vlan10
```

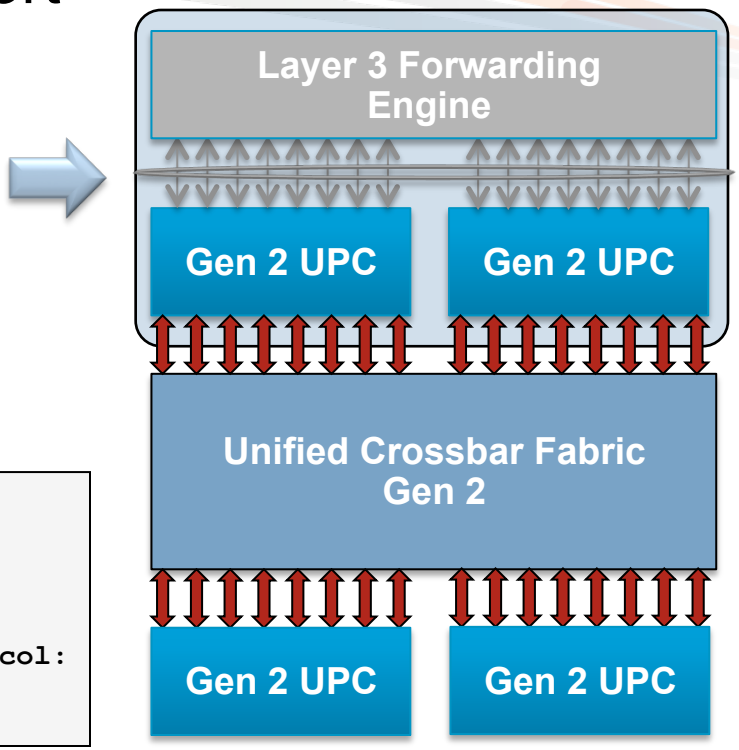


Note: Please see session BRKARC-3471 (Cisco NXOS Software - Architecture) for more information on NX-OS Software Architecture

Nexus 5500 Series

Nexus 5500 with Layer 3 support

- Layer 3 Forwarding Engine connects the the X-Bar via two UPC Gen-2 using a 16 x 10G internal port-channel (iPC)
- Traffic is load shared across the 16 fabric connections (iPorts)
- Recommendation configure L2/L3/L4 port channel hashing (global switch parameter)




```
L3-5548-1# sh port-channel load-balance

Port Channel Load-Balancing Configuration:
System: source-dest-port

Port Channel Load-Balancing Addresses Used Per-Protocol:
Non-IP: source-dest-mac
IP: source-dest-port source-dest-ip source-dest-mac
```

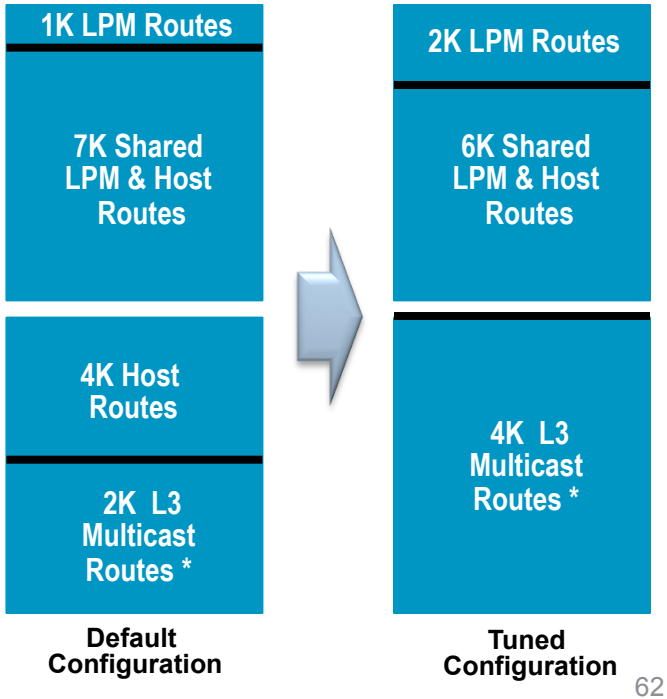
```
L3-5548-1# sh mod
Mod Ports  Module-Type                        Model                  Status
--- -----  -------------------------------- ---------------------- ------------
<snip>
3    0      O2 Daughter Card with L3 ASIC    N55-D160L3              ok


L3-5548-1# sh int port-channel 127
port-channel127 is up
<snip>
Members in this channel: Eth3/1, Eth3/2, Eth3/3, Eth3/4, Eth3/5, Eth3/6, Eth3/7, Eth3/8,
Eth3/9, Eth3/10, Eth3/11, Eth3/12, Eth3/13, Eth3/14, Eth3/15, Eth3/16
```

Nexus 5500 Series

Nexus 5500 with Layer 3 support

- Layer 3 Forwarding Tables can be tuned for specific design scenarios
- Similar to SDM templates used on Catalyst 3750/3650
- Three table space allocations
  - Host Routes (1 entry per /32) – Adjacent Hosts
  - LPM (1 entry per route) – Longest Prefix Match Routes
  - Multicast Routes (*2 entries per mcast route) – (S,G) and (*,G)

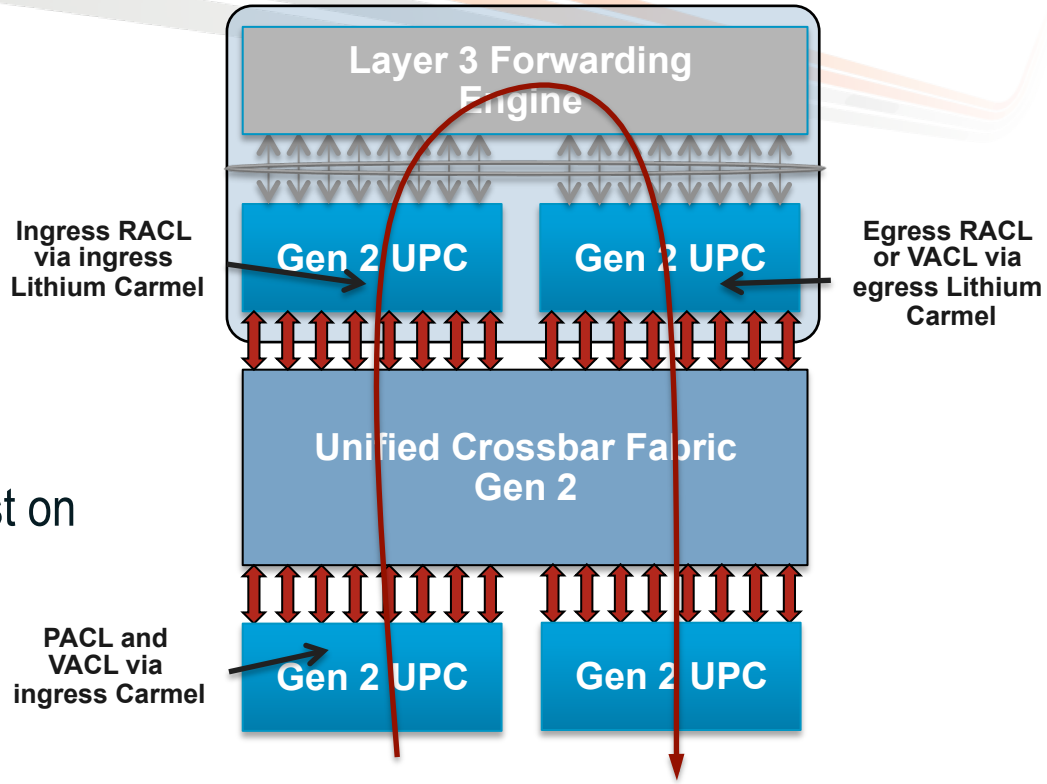
```
L3-5548-1# show hardware profile status
Reserved LPM Entries = 1024.
Reserved Host Entries = 4000.
Reserved Mcast Entries = 2048.
Used LPM Entries = 8.
Used Host Entries in LPM = 0.
Used Mcast Entries = 0.
Used Host Entries  in Host = 21.

L3-5548-1(config)# hardware profile module 3 lpm-entries 2048
L3-5548-1(config)# hardware profile multicast max-limit 4096

L3-5548-1# show hardware profile status
Reserved LPM Entries = 2048.
Reserved Host Entries = 4000.
Reserved Mcast Entries = 4096.
Used LPM Entries = 8.
Used Host Entries in LPM = 0.
Used Mcast Entries = 0.
Used Host Entries  in Host = 21.
```



Nexus 5500 Series

RACL Support

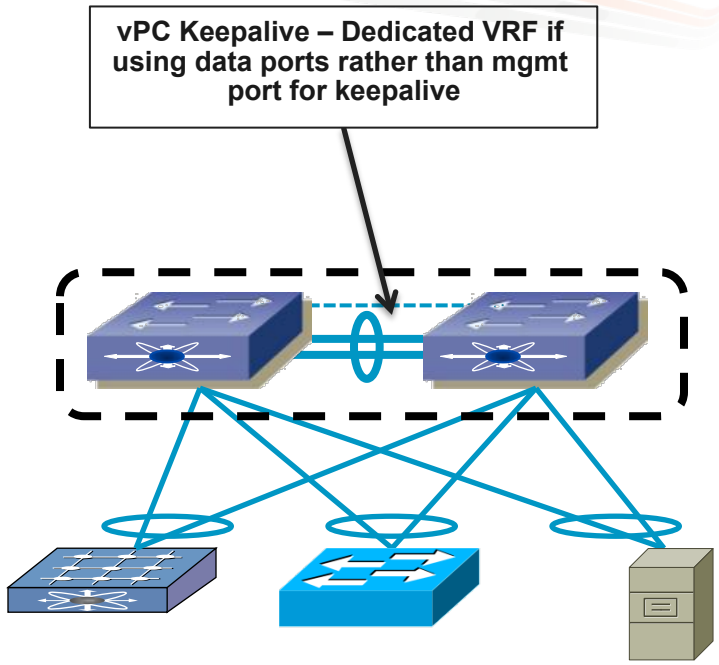
- RACLs can be configured on:
  - L3 Physical interface
  - L3 port-channel interface
  - L3 Sub-Interface
  - L3 Vlan Interface (SVI)
- RACLs and VACLs can not co-exist on the same SVI
  - First one configured is allowed
- Ingress - 1600 ACE supported
- Egress - 2048 ACE supported




```
L3-5548-1(config)# interface ethernet 1/17
L3-5548-1(config-if)# ip access-group acl01 in
L3-5548-1(config-if)# ip access-group acl01 out

Verifying the RACLs programming
L3-5548-1# show ip acc summ
IPV4 ACL acl01
    Total ACEs Configured: 1
    Configured on interfaces:
        Ethernet1/17 - ingress (Router ACL)
        Ethernet1/17 - egress (Router ACL)
<snip>
```

Nexus 5500 Series
VRF-Lite Support

- Prior to this release, N5k support two VRFs
  - VRF management & VRF default
- With 5.0(3)N1(1) user can create additional VRFs
  - VRF-lite,
  - VRF aware Unicast -BGP/OSPF/RIP
  - VRF Aware Multicast
- Hardware supports 1K VRF
- Current Solution testing limit – 64 VRF's
- Similar to N7K '*if*' user data ports are used as keepalive link, it is now recommended to create dedicate VRF for keepalive link




```
interface Vlan123
  vrf member vpc_keepalive
  ip address 123.1.1.2/30
  no shutdown
vpc domain 1
  peer-keepalive destination 123.1.1.1 source 123.1.1.2 vrf vpc_keepalive
```

Nexus 5000/5500 and 2000 Architecture

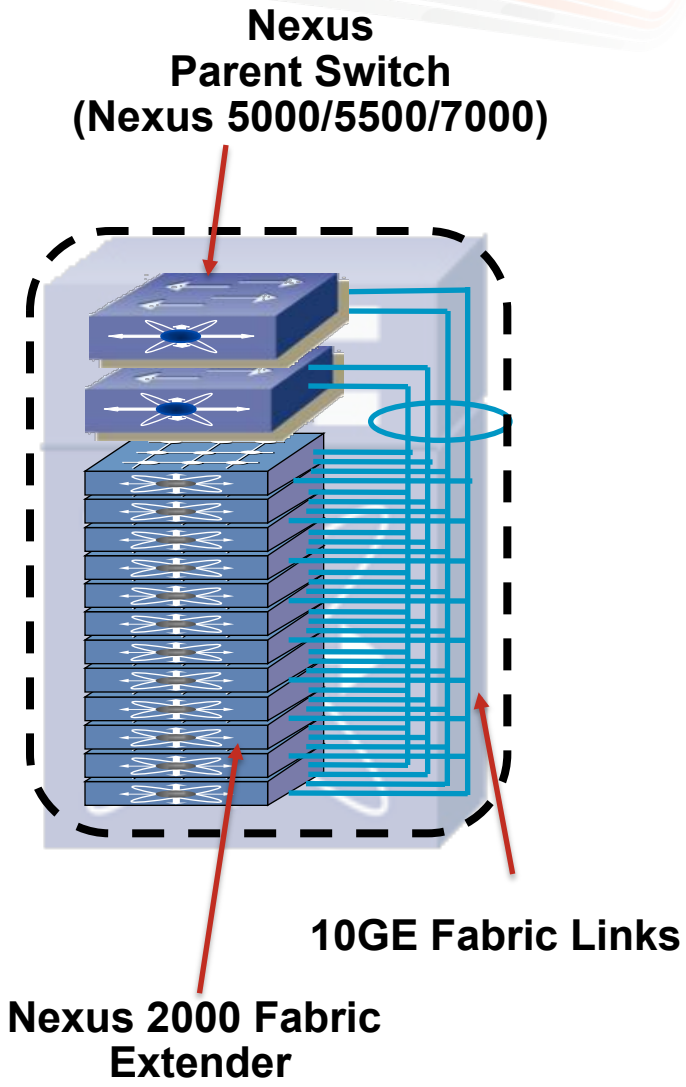
Agenda

- Nexus 5000/5500 Architecture
  - Hardware Architecture
  - Day in the Life of a Packet
  - Layer 3 Forwarding
- Nexus 2000 Architecture
  - FEXLink Architecture
  - FEX Forwarding
  - Extending FEXLink – Adapter FEX
- Nexus 5000/5500
  - Multicast
  - Port Channels
  - QoS

Cisco FEXlink: Virtualized Access Switch

Nexus 2000 Fabric Extender

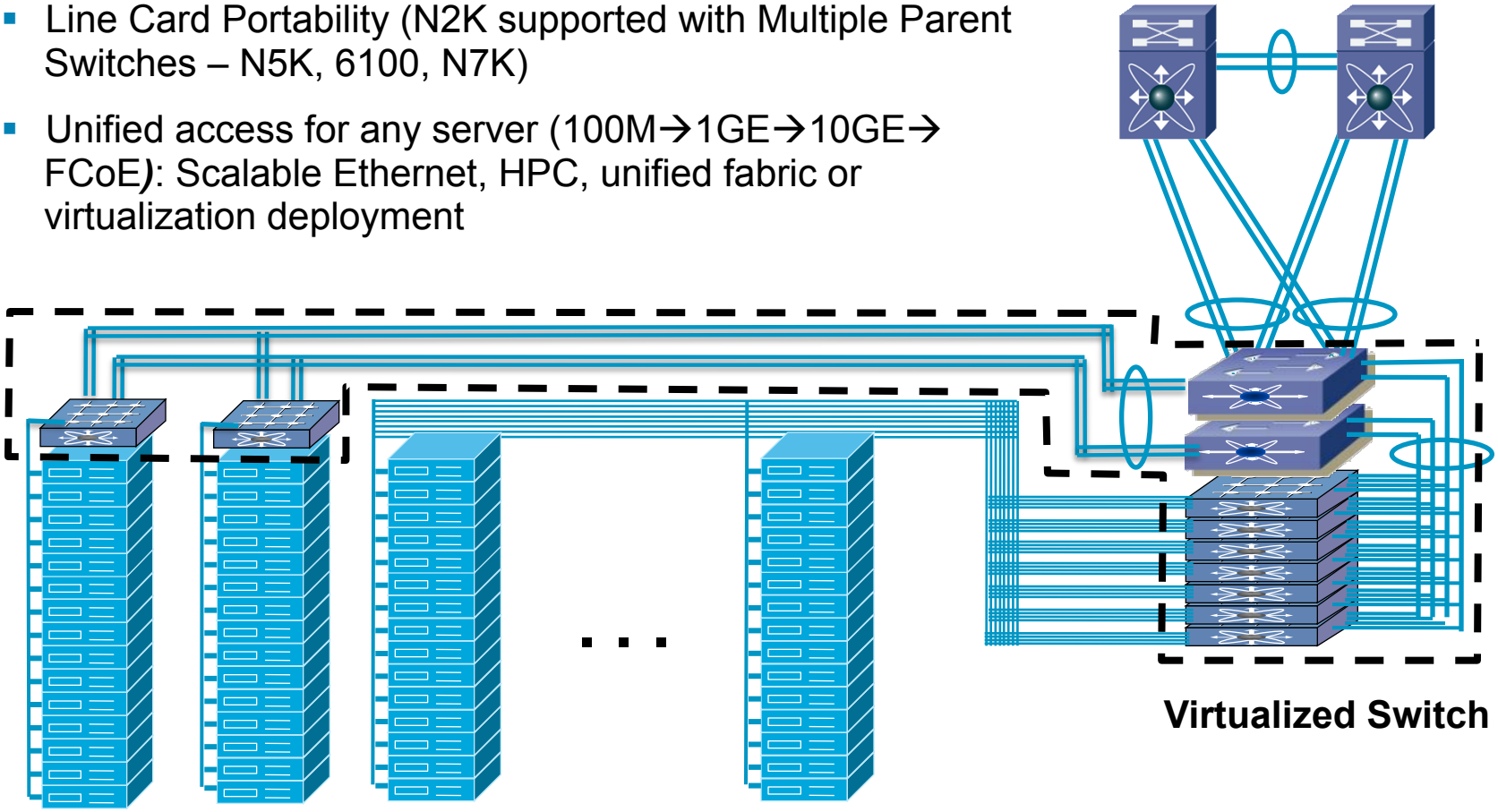
- FEXLink is an architectural approach to building a switching fabric
- Decoupled elements
  - Control Plane
  - Forwarding ASIC
  - Port ASIC
- Component upgrade path is modeled on a classical Modular architecture
  - Upgrade of components can be implemented separately
  - Replacing the Parent Switch is functionally equivalent to upgrading the Supervisor and Fabric ASIC's
- Single System software image and point of configuration and management



Cisco FEXlink: Virtualized Access Switch

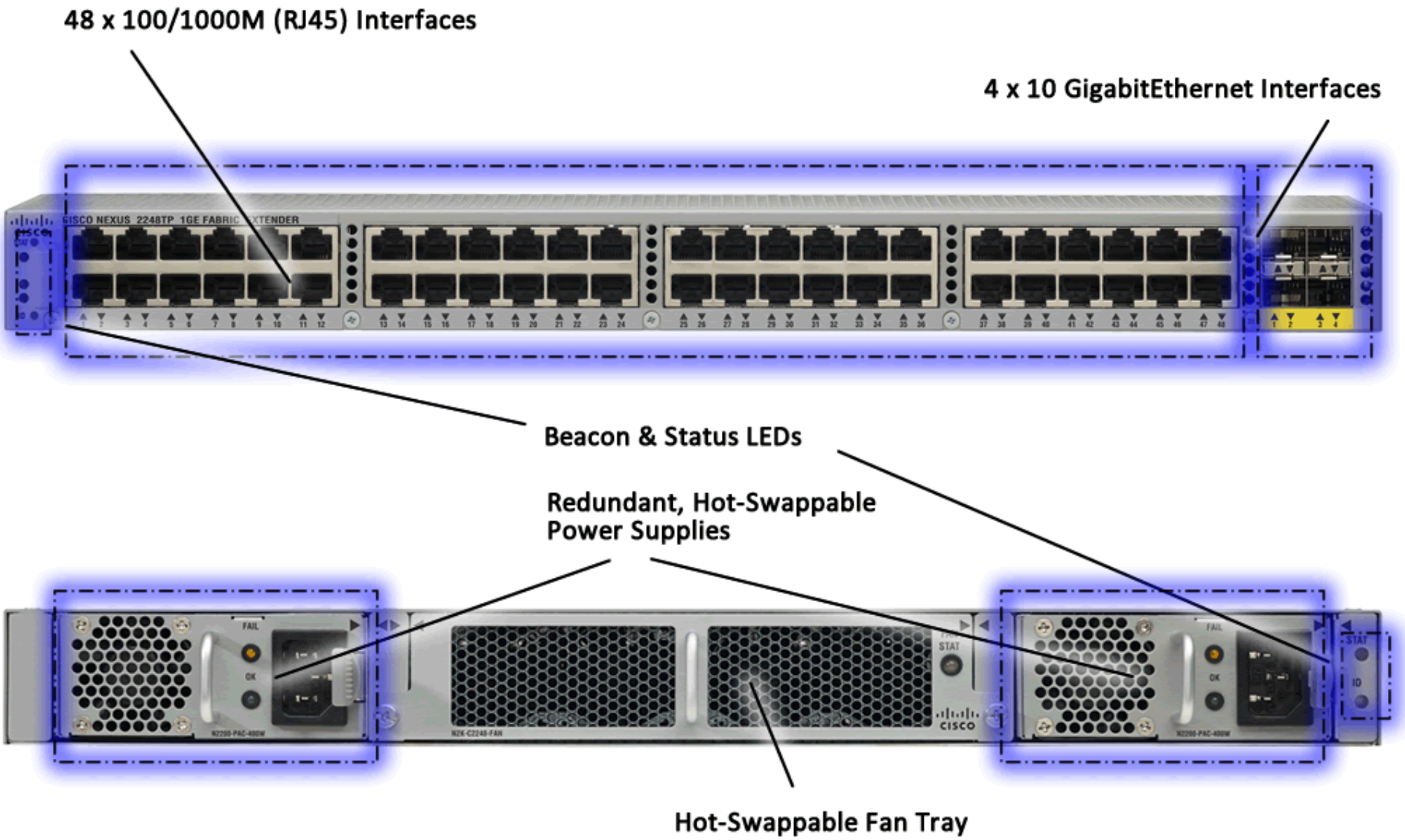
Changing the device paradigm

- De-Coupling of the Layer 1 and Layer 2 Topologies
- Simplified Management Model, plug and play provisioning, centralized configuration
- Line Card Portability (N2K supported with Multiple Parent Switches – N5K, 6100, N7K)
- Unified access for any server (100M→1GE→10GE→ FCoE*)*: Scalable Ethernet, HPC, unified fabric or virtualization deployment



Cisco Nexus 2248T Fabric Extender

Overview



Cisco Nexus 2200 Mechanicals

Reversible Airflow & DC Power Supplies

- Nexus 2200 chassis support front-to-back airflow and reversed airflow (Port side in hot aisle)
- Change of airflow achieved through new power supply/Fan Tray combination
- Nexus 2200 chassis support AC and DC Power Supplies
- DC Power Supply support for *front-to-back* airflow only
- Software availability: 5.0(3)N1, hardware FCS: Q2CY11

400W DC Power Supply
Front-to-back airflow
48 to -60 VDC

Front-to-back airflow – Power supply aligned with cold aisle

Reversed airflow
Black alignment indicator for
Back-to-front airflow

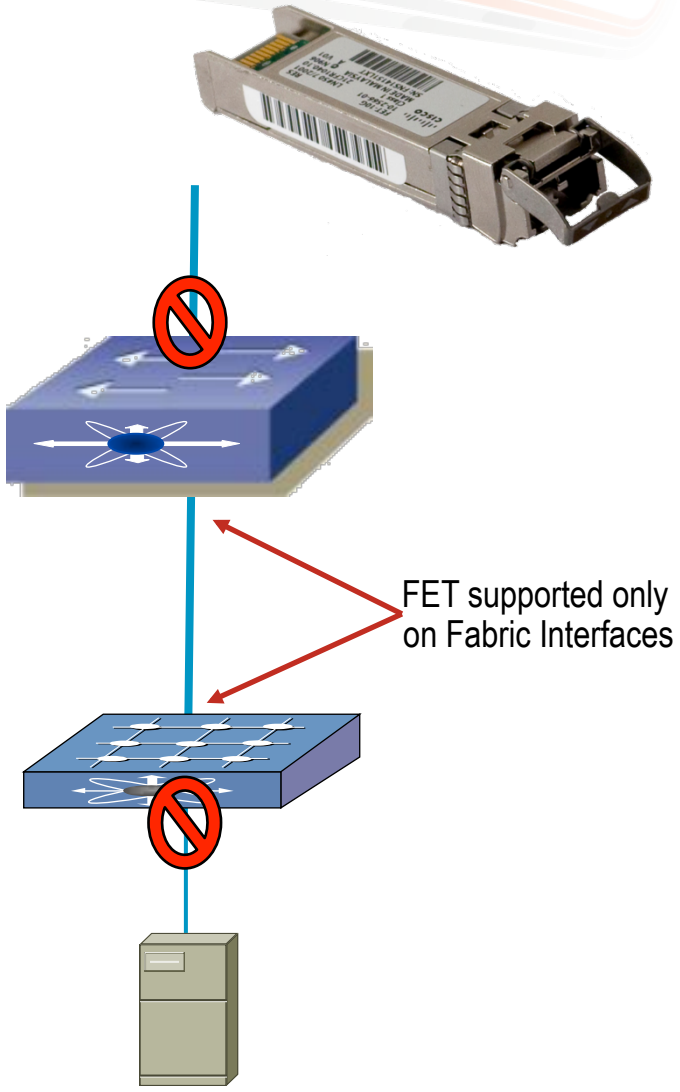
Nexus 2148T, 2248TP, 2224TP, 2232PP, 2232TM Capabilities

For Your Reference

Model	Nexus 2148T	Nexus 2224TP	Nexus 2248TP	Nexus 2232PP-10G	Nexus 2232TM-10G
Product Shipping	Yes (Q1CY09)	No (Q3CY10)	Yes (Q2CY10)	Yes (Q2CY10)	Target (Q2CY11)
Form Factor	1 RU	1 RU	1 RU	1 RU	1 RU
Uplink Ports	4 x 10GbE SFP+	2 x 10GbE SFP+	4 x 10GbE SFP+	8 x 10GbE SFP+	8 x 10GbE SFP+
Uplink Transceivers Supported	Copper CX-1 (passive): 1m, 3m, 5m. Copper CX1 (active): 7m, 10m Optical: FET (Nexus 2200 platforms), SR, LR				
Host Facing Ports	48 x 1GbE RJ45 (note: 1000BaseT only)	24 x 100/1000Base-T RJ45	48 x 100/1000Base-T RJ45	32 x SFP/SFP+ (1/10G)	32 x 1/10G Base-T RJ45
FCoE	N/A	N/A	N/A	Yes	No
Dimensions	1.72 x 17.3 x 20.0 in	1.72 x 17.3 x 17.7in	1.72 x 17.3 x 17.7in	1.72 x 17.3 x 17.7 in	1.72 x 17.3 x 17.7 in
Max Operational Power	165W	80-95W	95-110W	230-270W	300-400W
Supports FET	No	Yes	Yes	Yes	Yes
Multiple PortChannel member ports on a FEX	Not Supported	Yes	Yes	Yes	Yes
Scalability	1152 GbE Ports per N5K	576 GbE Ports per N5K	1152GbE Ports per N5K 1536 GbE Ports per N7K	768 1/10GbE Ports per N5K	768 1/10GbT Ports per N5K
Number of FEX	24 FEX per N5500	24 FEX per N5500	24 FEX per N5500 32 FEX per N7K	24 FEX per N5500	24 FEX per N5500

Virtualized Access Switch
Fabric Extender Transceiver (FET)

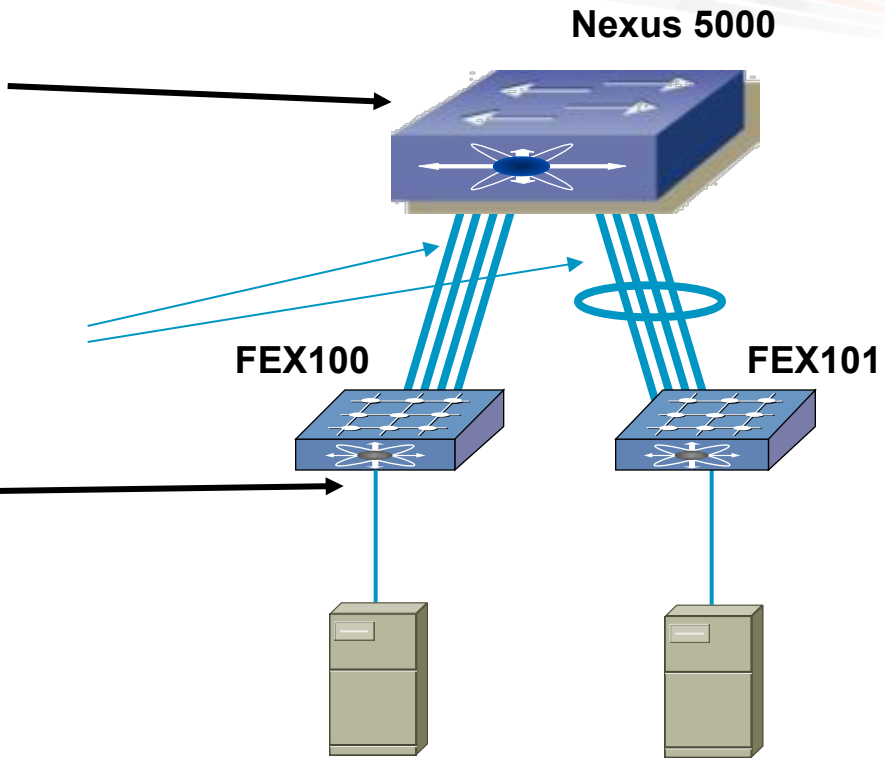
- Cost-effective transceiver to interconnect Nexus 2000 and Nexus 5000 and 7000 parent switch (only supported on FEX Fabric interfaces)
- SFP+ form-factor
- Multimode fiber (MMF)
- FET with OM3 MMF can operate up to 100m
- FET with OM2 MMF can operate up to 20m
- FET with 62.5/125um MMF can operate up to 10m
- Approximately 1 watt (W) per transceiver
- Incompatible with SR optics



Cisco Nexus 2000 Fabric Extender

Fabric Extender Terminology

- **Parent Switch:** Acts as the combined Supervisor and Switching Fabric for the virtual switch
- **Fabric Links**: Extends the Switching Fabric to the remote line card (Connects Nexus 5000 to Fabric Extender)
- **Host Interfaces** (HIF)
- Fabric connectivity between Nexus 5000 and Nexus 2000 (FEX) can leverage either **pinning** or **port-channels**



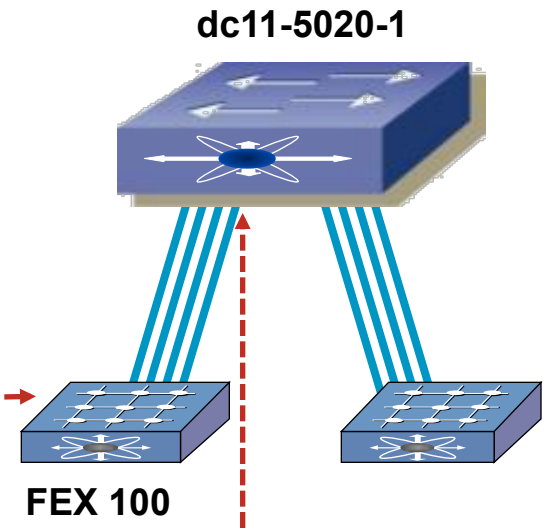

```
dc11-5020-1# show interface fex-fabric
     Fabric      Fabric       Fex                 FEX
Fex  Port      Port State    Uplink    Model         Serial
--------------------------------------------------------------
100   Eth1/17        Active      1    N2K-C2148T-1GE  JAF1311AFLL
100   Eth1/18        Active      2    N2K-C2148T-1GE  JAF1311AFLL
100   Eth1/19        Active      3    N2K-C2148T-1GE  JAF1311AFLL
100   Eth1/20        Active      4    N2K-C2148T-1GE  JAF1311AFLL
101   Eth1/21        Active      1    N2K-C2148T-1GE  JAF1311AFMT
101   Eth1/22        Active      2    N2K-C2148T-1GE  JAF1311AFMT
```

Cisco Nexus 2000 Fabric Extender

Configuring the Fabric

- Two step process
- Define the Fabric Extender (100–199) and the number of fabric uplinks to be used by that FEX (valid range: 1–4)

```
dc11-5020-1# switch# configure terminal
dc11-5020-1(config)# fex 100
dc11-5020-1(config-fex)# pinning max-links 4
```



- Configure Nexus 5000 ports as fabric ports and associate the desired FEX

```
dc11-5020-1# switch# switch# configure terminal
dc11-5020-1(config)# interface ethernet 1/1
dc11-5020-1(config-if)# switchport mode fex-fabric
dc11-5020-1(config-if)# fex associate 100
. . .
<repeat for all 4 interfaces used by this FEX>
```

Cisco Nexus 2000 Fabric Extender

Fabric Connectivity

Show the attached Fabric Extenders

```
dc11-5020-1# show fex
  FEX             FEX             FEX                 FEX
Number       Description         State      Model            Serial
------------------------------------------------------------------------------
100         FEX0100         Online      N2K-C2148T-1GE    JAF1311AFLL
101         FEX0101         Online      N2K-C2148T-1GE    JAF1311AFMT
```

Show the status of fabric link 'port-channel 100'

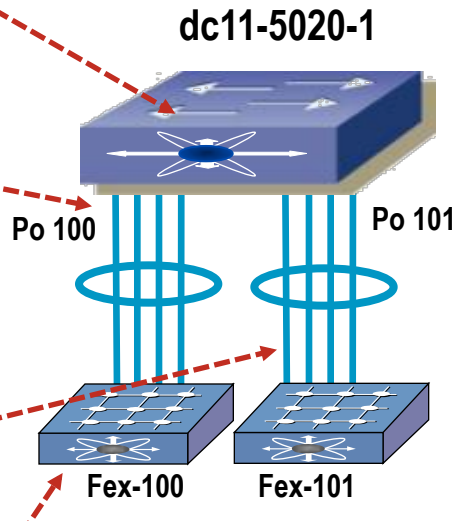
```
dc11-5020-1# show interface port-channel 100
port-channel100 is up
  Hardware: Port-Channel, address: 000d.eca4.5318 (bia 000d.eca4.5318)
<snip>
  Port mode is fex-fabric
```

Show the N2K interfaces carried over fabric link 'port-channel 101'

```
dc11-5020-1# sh int port-channel 101 fex-intf
Fabric            FEX
Interface         Interfaces
---------------------------------------------------
 Po101            Eth101/1/1    Eth101/1/2    Eth101/1/3    Eth101/1/4
                  Eth101/1/5    Eth101/1/6    Eth101/1/7    Eth101/1/8
<snip>
```

Show the interfaces themselves (N2K interfaces are N5K ports)

```
dc11-5020-1# sh int brief
<snip>
--------------------------------------------------------------------------------
Ethernet      VLAN    Type Mode   Status  Reason                   Speed     Port
Interface                                                                    Ch #
--------------------------------------------------------------------------------
Eth100/1/1    1       eth  access up      none                       1000(D) --
Eth100/1/2    1       eth  access down    Link not connected         1000(D) --
Eth100/1/3    1       eth  access up      none                       1000(D) --
```



Nexus 5000/5500 and 2000 Architecture

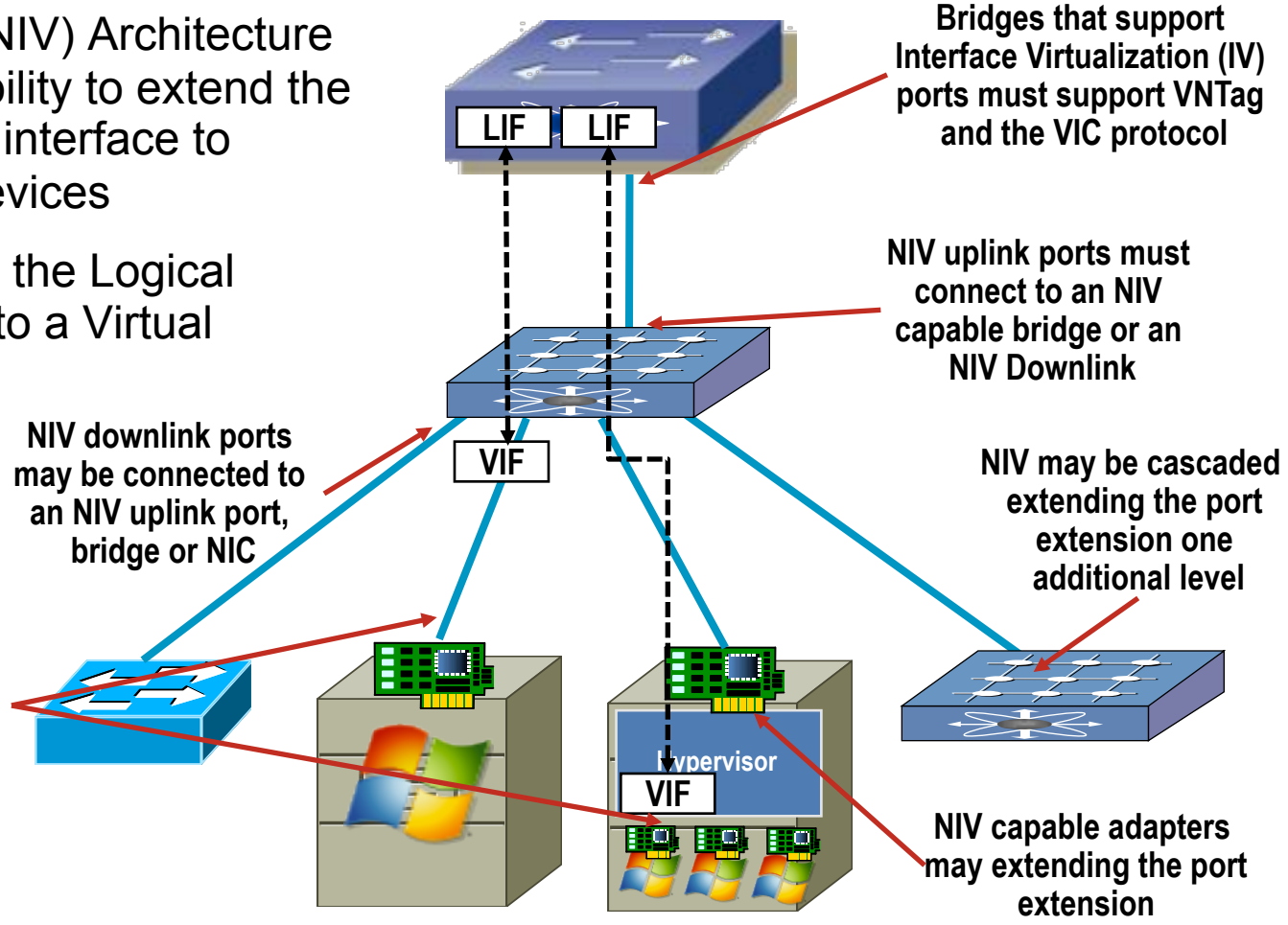
Agenda

- Nexus 5000/5500 Architecture
  - Hardware Architecture
  - Day in the Life of a Packet
  - Layer 3 Forwarding
- Nexus 2000 Architecture
  - FEXLink Architecture
  - FEX Forwarding
  - Extending FEXLink – Adapter FEX
- Nexus 5000/5500
  - Multicast
  - Port Channels
  - QoS

Nexus 2000 Fabric Extender

Network Interface Virtualization Architecture (NIV)

- The Network Interface Virtualization (NIV) Architecture provides the ability to extend the bridge (switch) interface to downstream devices
- NIV associates the Logical Interface (LIF) to a Virtual Interface (VIF)

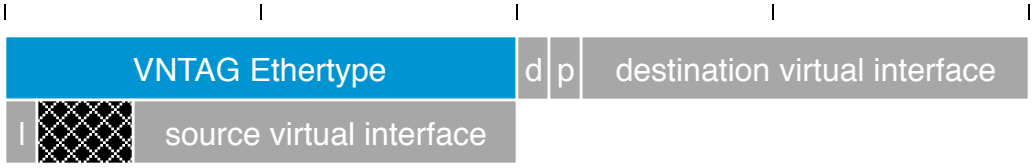


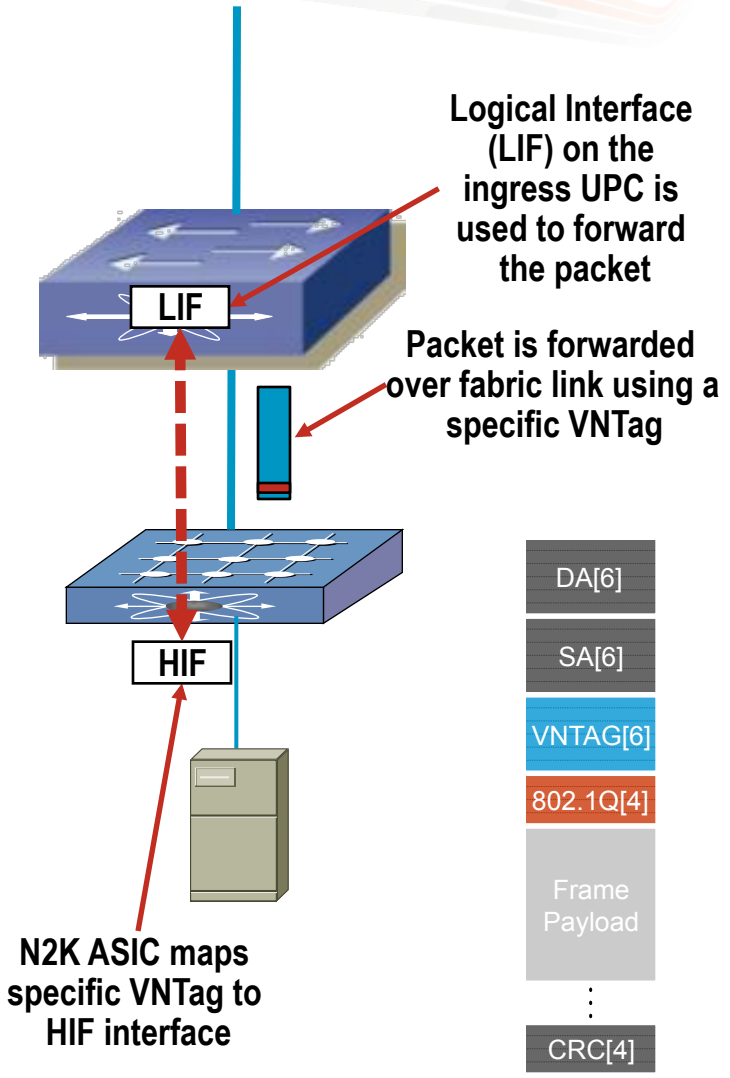
Note: Not All Designs Supported in the NIV Architecture Are Currently Implemented

Nexus 2000 Fabric Extender

VN-Tag Port Extension

- Nexus 2000 Fabric Extender operates as a remote line card and does ***not*** support local switching
- All forwarding is performed on the Nexus 5000/5500 UPC or Nexus 7000 EARL
- VNTag is a Network Interface Virtualization (NIV) technology that 'extends' the Nexus 5000/7000 port down (Logical Interface = LIF) to the Nexus 2000 VIF referred to as a Host Interface (HIF)
  - VNTag is added to the packet between Fabric Extender and Nexus 5000/5500/7000
  - VNTag is stripped before the packet is sent to hosts
- VNTag allows the Fabric Extender to act as a data path of Nexus 5000/5500/7000 for all policy and forwarding

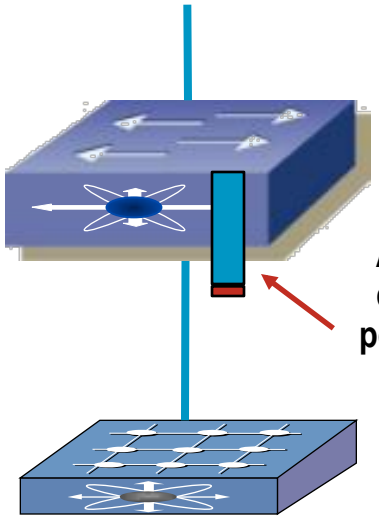




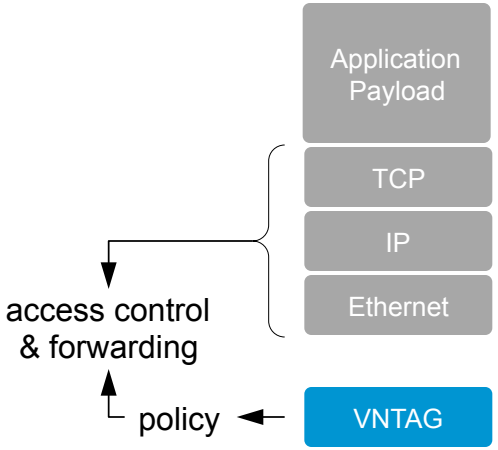
Nexus 2000 Fabric Extender

VN-Tag Port Extension

- Nexus 5000/5500/7000 ingress processing on fabric ports
- UPC extracts VNTAG which identifies the Logical Interface (LIF) corresponding to the physical HIF on the actual Nexus 2000
- Ingress policy based on physical Nexus 5000/5500/7000 port and LIF
  - Access control and forwarding based on frame fields and virtual interface (LIF) policy
  - Physical link level properties (e.g. MACSEC, …) are based on the Nexus 5000/5500/7000 port
- Forwarding selects destination port(s) and/or destination virtual interface(s)

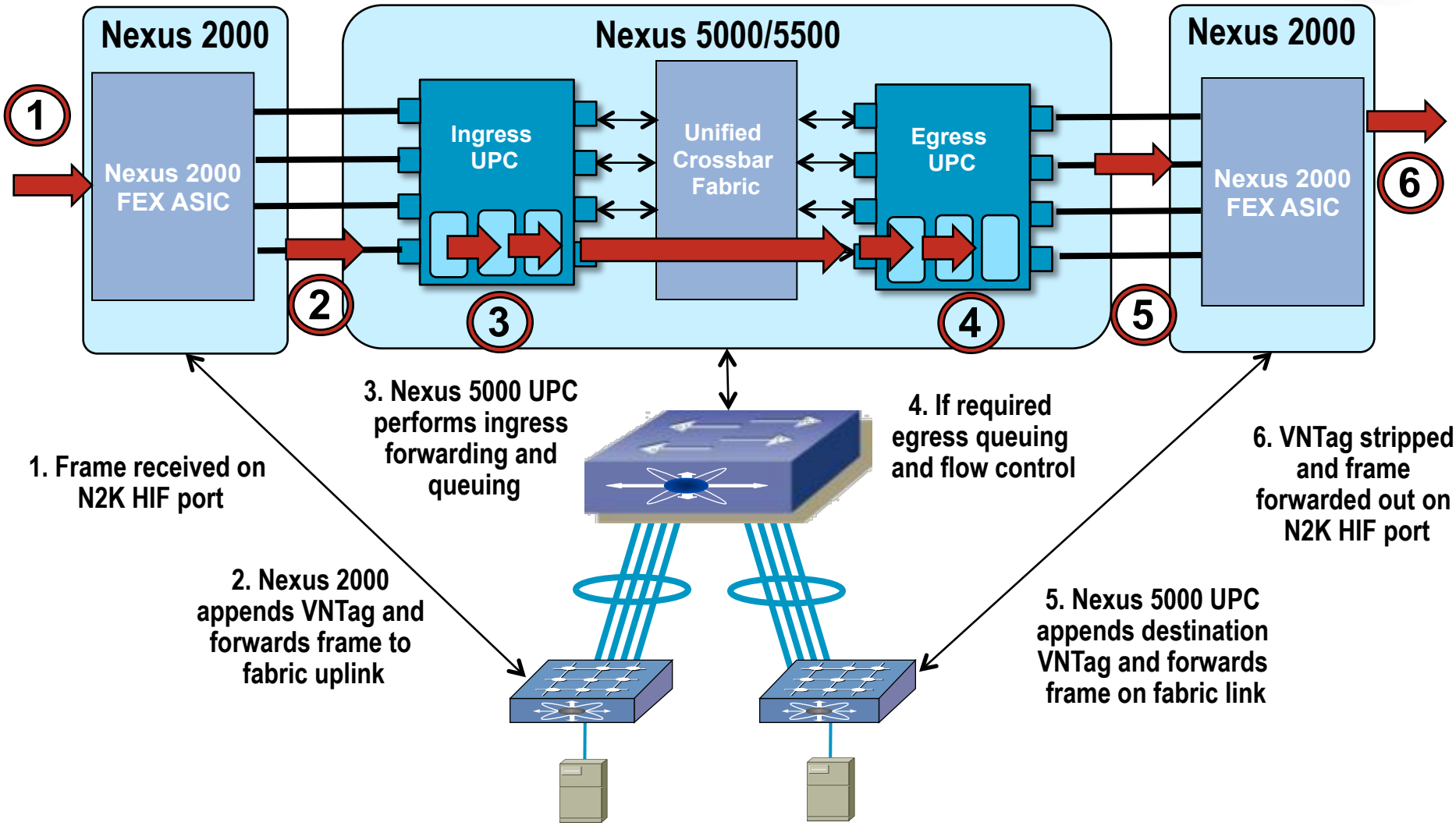


ASIC identifies the correct forwarding policy using network topology (packet header) + switch topology (VNTag)



Nexus 5000/5500 and 2000

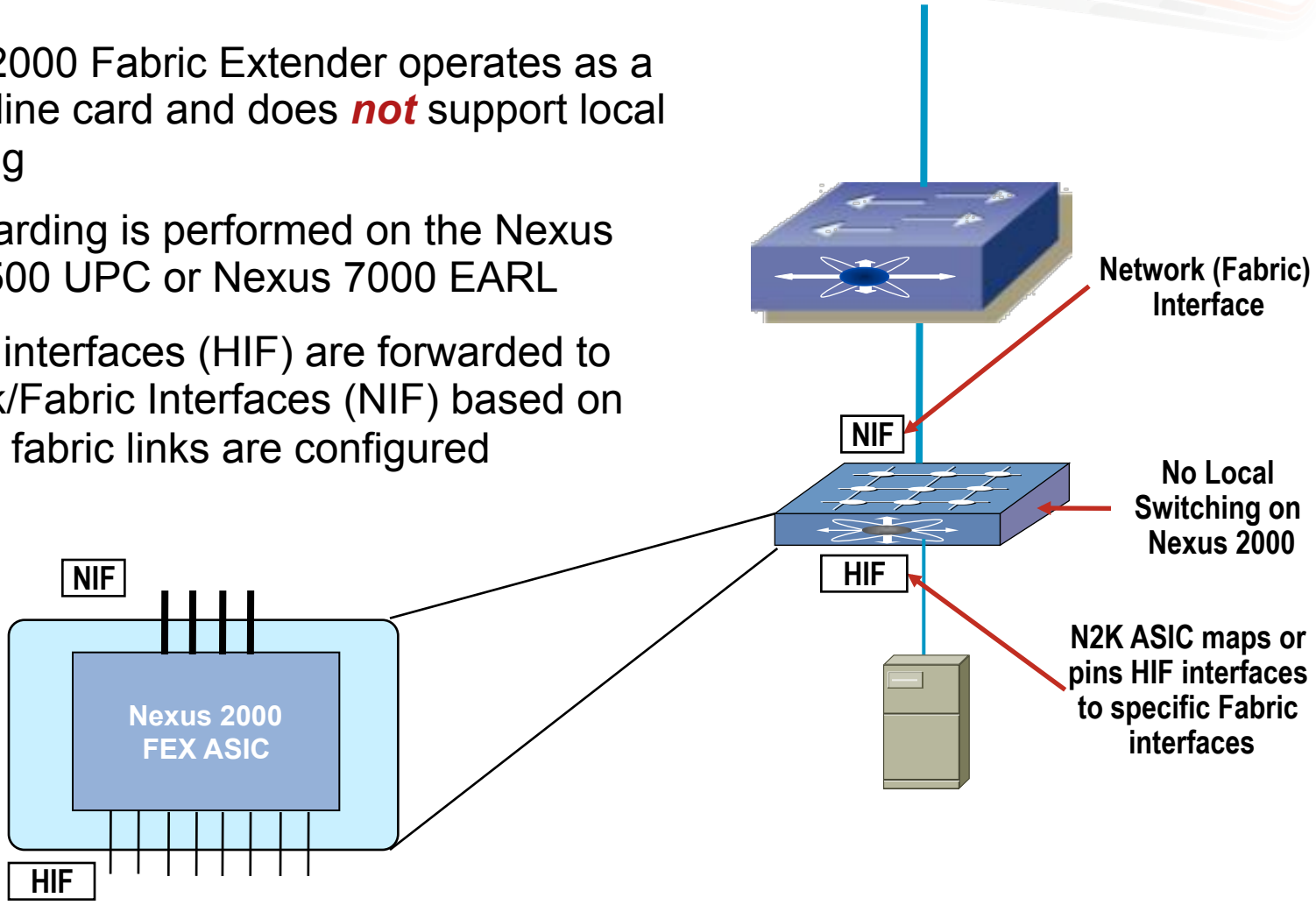
Packet Forwarding Overview



Nexus 2000 Fabric Extender

Nexus 2000 Packet Forwarding

- Nexus 2000 Fabric Extender operates as a remote line card and does ***not*** support local switching
- All forwarding is performed on the Nexus 5000/5500 UPC or Nexus 7000 EARL
- Ingress interfaces (HIF) are forwarded to Network/Fabric Interfaces (NIF) based on how the fabric links are configured



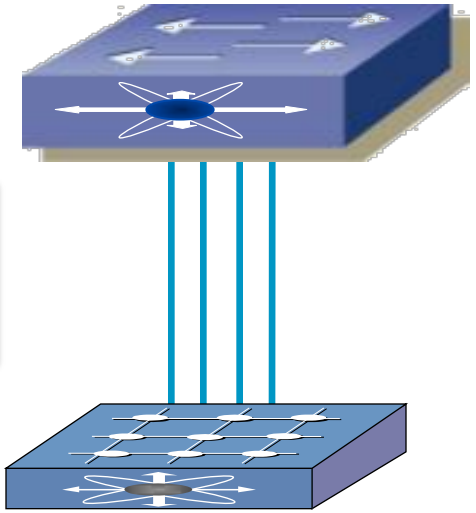
Nexus 2000 Fabric Extender

Fabric—Static Pinning

- Static Pinning associates (maps) specific server ports to specific fabric links
- Need to ensure that the ***same*** number of Ethernet ports are assigned as fex-fabric ports as defined in the ‘max-links’ parameter for that Fabric Extender

```
interface Ethernet1/1
  switchport mode fex-fabric
  fex associate 100

interface Ethernet1/2
  switchport mode fex-fabric
  fex associate 100

interface Ethernet1/3
  switchport mode fex-fabric
  fex associate 100

interface Ethernet1/4
  switchport mode fex-fabric
  fex associate 100

!
fex 100
  pinning max-links 4
  description Rack_100
```

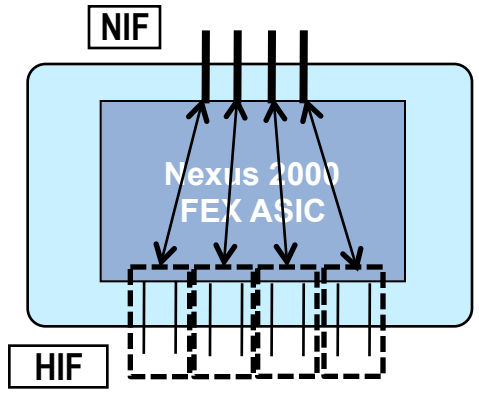
Ports Are Configured as Fabric and Associated with a Specific Fabric Extender



Nexus 2000 Fabric Extender

Fabric—Static Pinning

- Packets within the Nexus 2000 are ‘pinned’ or mapped from a specific ingress interface (HIF) to a specific fabric interface (NIF)
- When configured in ‘static pinning’ mode specific HIF are statically mapped to specific NIF
- Changing the number of fabric links requires the ASIC ‘pinning’ to be changed and is ***disruptive*** to traffic flows




```
dc11-5020-2# sh fex 150 detail
  FEX: 150 Description: FEX0150   state: Online
<snip>
  pinning-mode: static    Max-links: 2
  Fabric port for control traffic: Eth1/29
  Fabric interface state:
    Eth1/29 - Interface Up. State: Active
    Eth1/30 - Interface Up. State: Active
Fex Port        State  Fabric Port  Primary Fabric
       Eth150/1/1  Down     Eth1/29     Eth1/29
       Eth150/1/2  Down     Eth1/29     Eth1/29
<snip>
      Eth150/1/25  Down     Eth1/30     Eth1/30
      Eth150/1/26  Down     Eth1/30     Eth1/30
<snip>
```

Fabric Ports

Fabric Pinning

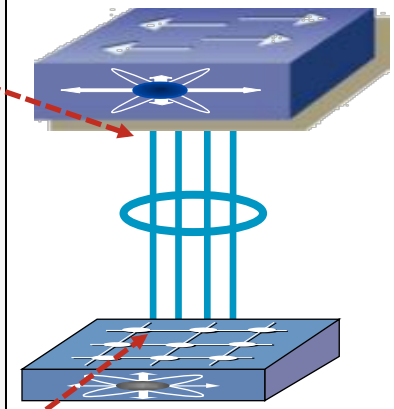
Nexus 2000 Fabric Extender

Fabric—Port Channel Configuration

```
interface port-channel1
  switchport mode fex-fabric
  description Fabric Extender 100
  fex associate 100

interface Ethernet1/1
  switchport mode fex-fabric
  description Fabric Extender 100 Etherchannel Link
  channel-group 1
  fex associate 100

interface Ethernet1/2
  switchport mode fex-fabric
  description Fabric Extender 100 Etherchannel Link
  channel-group 1
  fex associate 100

interface Ethernet1/3
  switchport mode fex-fabric
  description Fabric Extender 100 Etherchannel Link
  channel-group 1
  fex associate 100

interface Ethernet1/4
  switchport mode fex-fabric
  description Fabric Extender 100 Etherchannel Link
  channel-group 1
  fex associate 100

fex 100
  pinning max-links 1
  description Fabric Extender 100 - Using Etherchannel 1
```

Configure the Physical Ports as Members of the Fabric EtherChannel

Configure the Port Channel and Its Members to be Associated with a Specific Fabric Extender

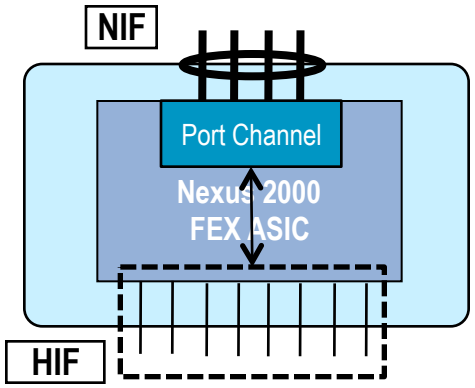


All server ports 'pinned' to a single logical fabric link

Nexus 2000 Fabric Extender

Fabric Port Channel Configuration

- In the fabric port channel configuration the internal forwarding within the Nexus 2000 ASIC is still ‘pinned’
- All HIF interfaces are pinned to an internal port channel NIF interface rather than to specific physical NIF interfaces
- Changing the number of fabric links does not require a changing in the internal forwarding mapping within the Nexus 2000 ASIC and is thus ***‘non-disruptive’***



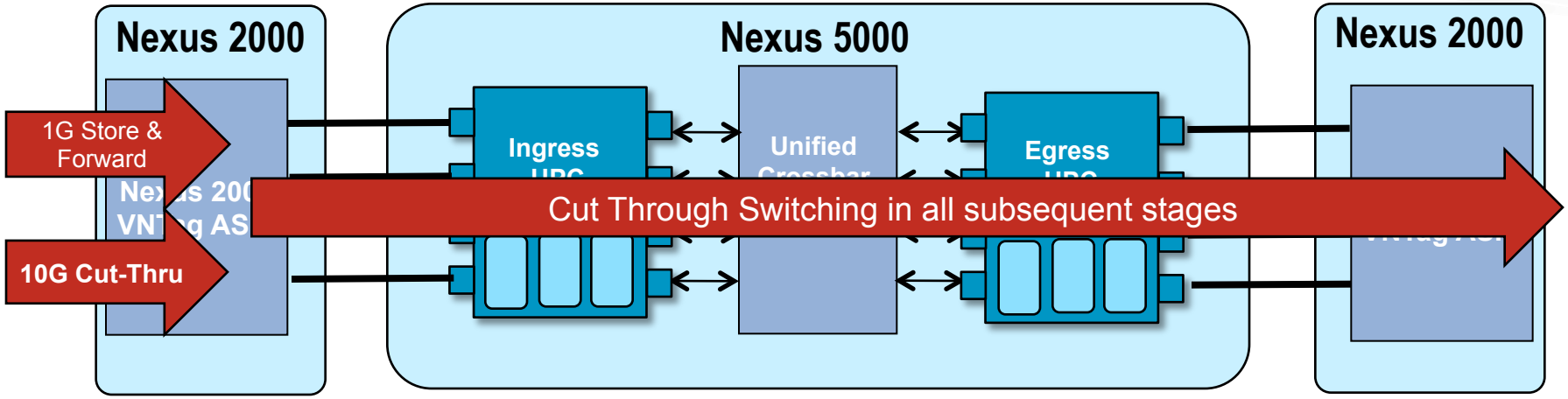

```
dc11-5020-1# sh fex 101 detail
  FEX: 101 Description: vPC-FEX   state: Online
<snip>
  pinning-mode: static    Max-links: 1
  Fabric port for control traffic: Eth1/21
  Fabric interface state:
    Po101 - Interface Up. State: Active
    Eth1/21 - Interface Up. State: Active
    Eth1/22 - Interface Up. State: Active
  Fex Port         State  Fabric Port  Primary Fabric
       Eth101/1/1    Up       Po101       Po101
       Eth101/1/2    Up       Po101       Po101
       Eth101/1/3  Down       Po101       Po101
<snip>
```

Fabric Ports

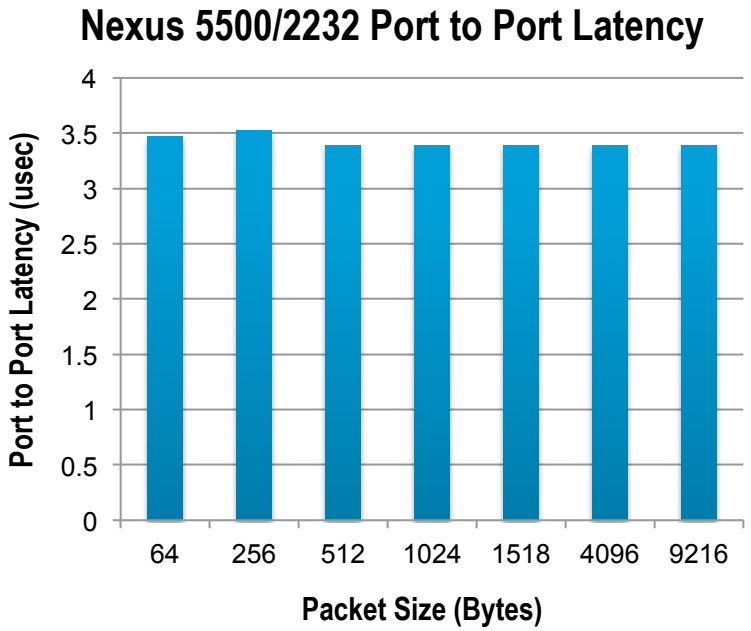
Fabric Pinning

Nexus 5000/5500 and 2000 Virtual Switch

Packet Forwarding Latency

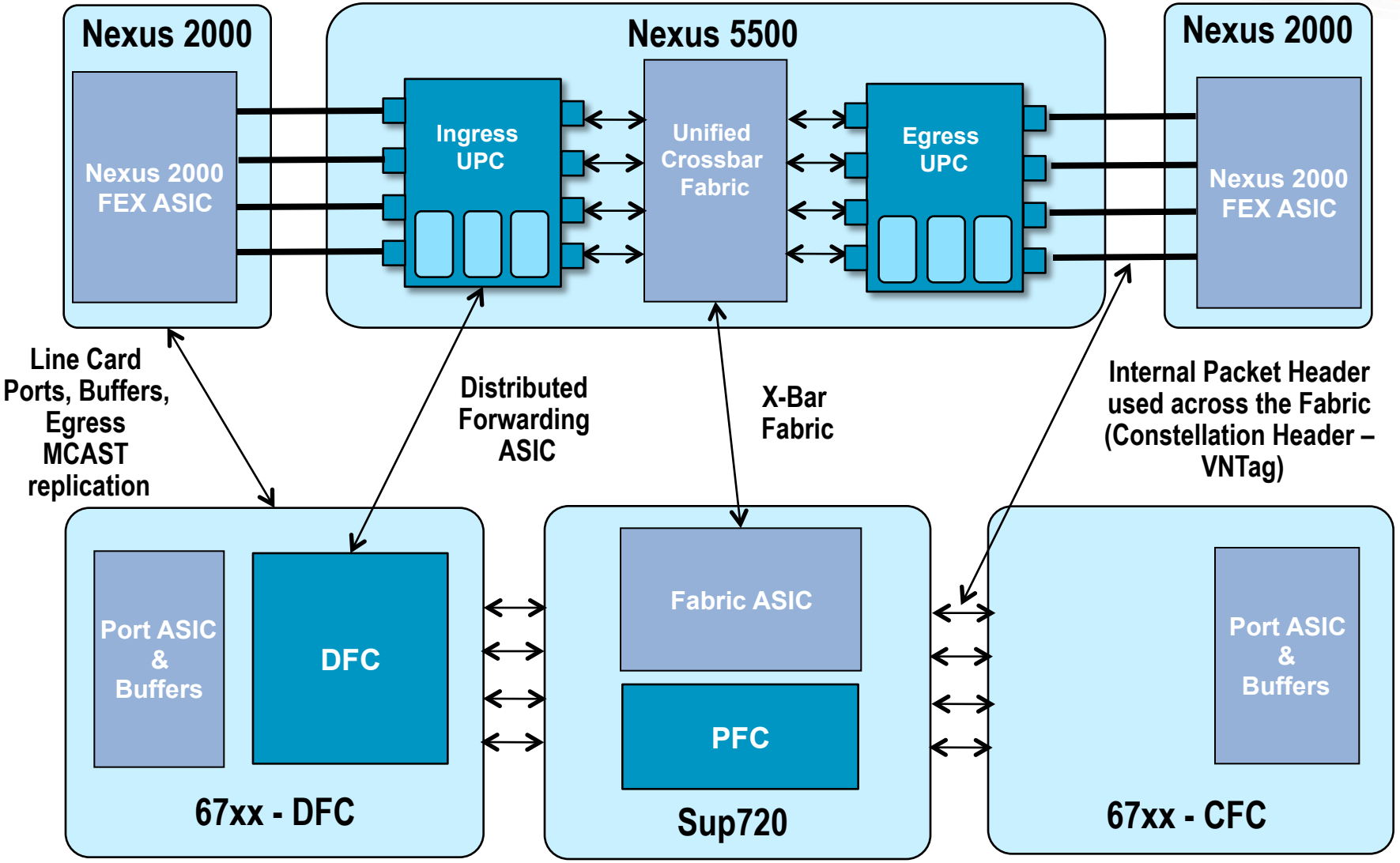


- Nexus 2000 also supports Cut -Through switching
  - 1GE to 10GE on first N2K ingress is store and forward
  - All other stages are Cut Through (10GE N2K port operates in end to end cut through)
- Port to Port latency is dependent on a single store and forward operation at most



Nexus 5000/5500 and 2000

Switching Morphology—Is this Really Different?



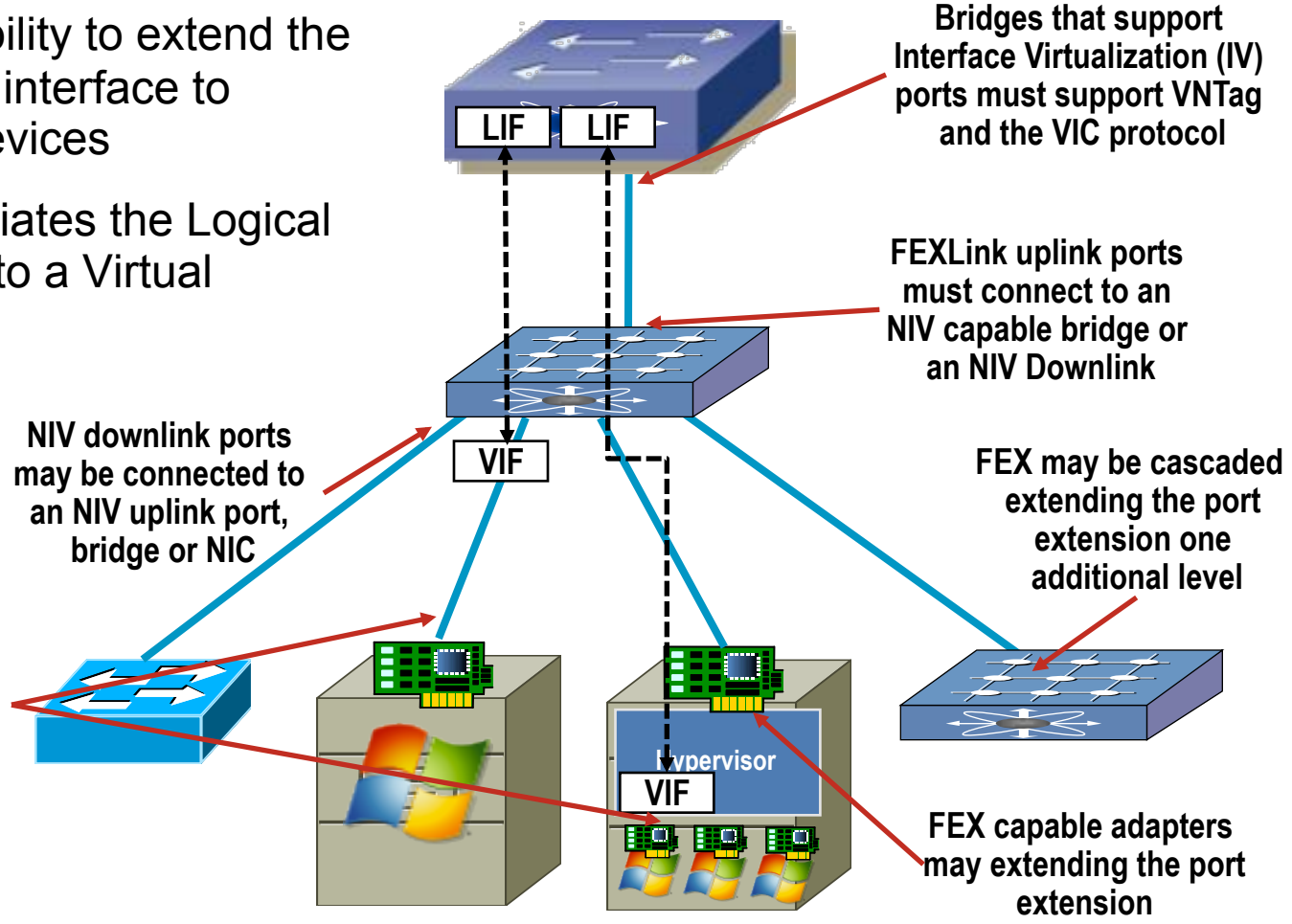
Nexus 5000/5500 and 2000 Architecture

Agenda

- Nexus 5000/5500 Architecture
  - Hardware Architecture
  - Day in the Life of a Packet
  - Layer 3 Forwarding
- Nexus 2000 Architecture
  - FEXLink Architecture
  - FEX Forwarding
  - Extending FEXLink – Adapter FEX
- Nexus 5000/5500
  - Multicast
  - Port Channels
  - QoS

FEX-Link
Extended Fabric, Ports and Virtualized Switching

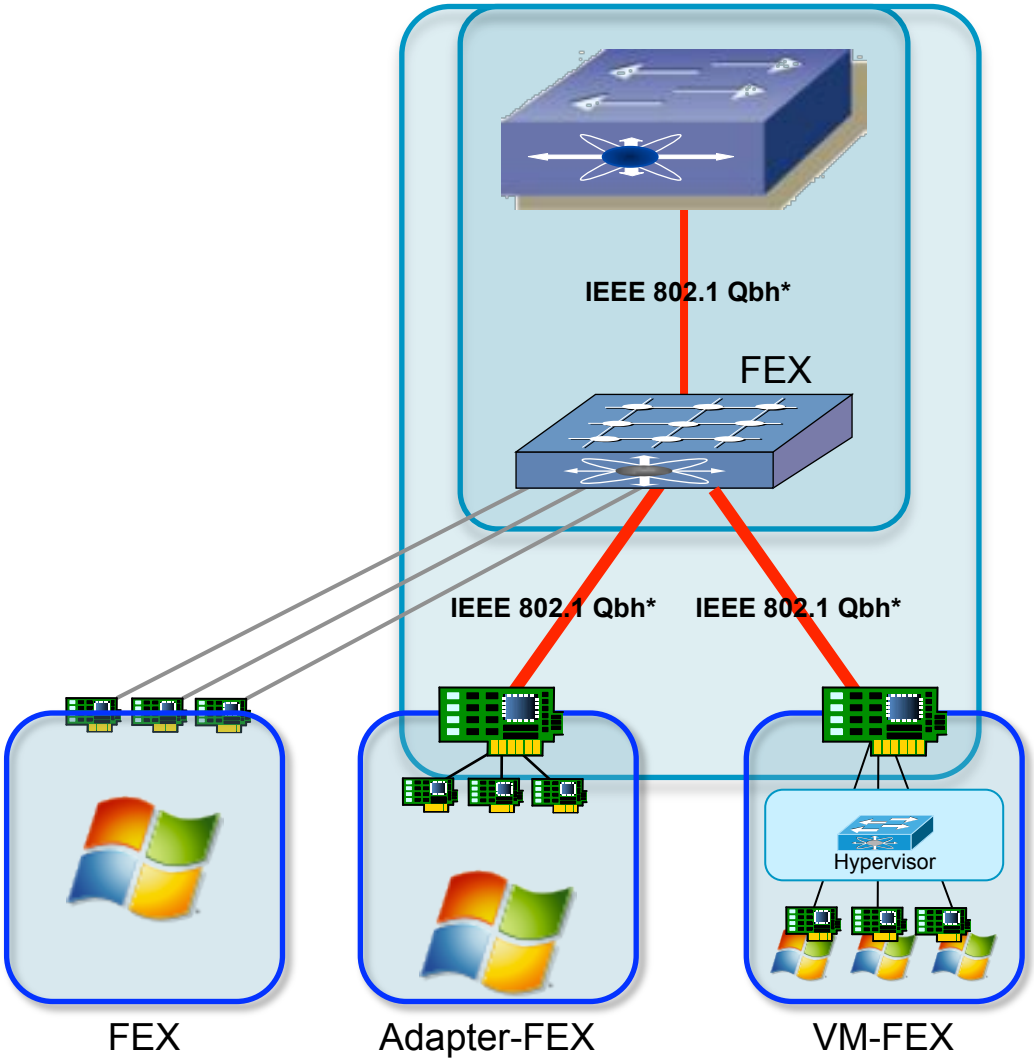
- The FEXLink Architecture provides the ability to extend the bridge (switch) interface to downstream devices
- FEXLink associates the Logical Interface (LIF) to a Virtual Interface (VIF)



Note: Not All Designs Supported in the FEX Architecture Are Currently Implemented

FEX-Link

Extender Ports and Virtualized Switching



- Adapter FEX leverages the foundational architecture to extend the VIF to server PCI bus
- O/S sees unique PCI addresses on bus and installs a unique device driver per address
- Each PCI address maps to a tag on the adapter uplink and is in turn mapped to the LIF on the Nexus 5500
- VM-FEX leverages Nexus 5500 as a vDC and VIF ports are seen as ‘veth’ ports for the VM’s

*IEEE 802.1Qbh is pre-standard

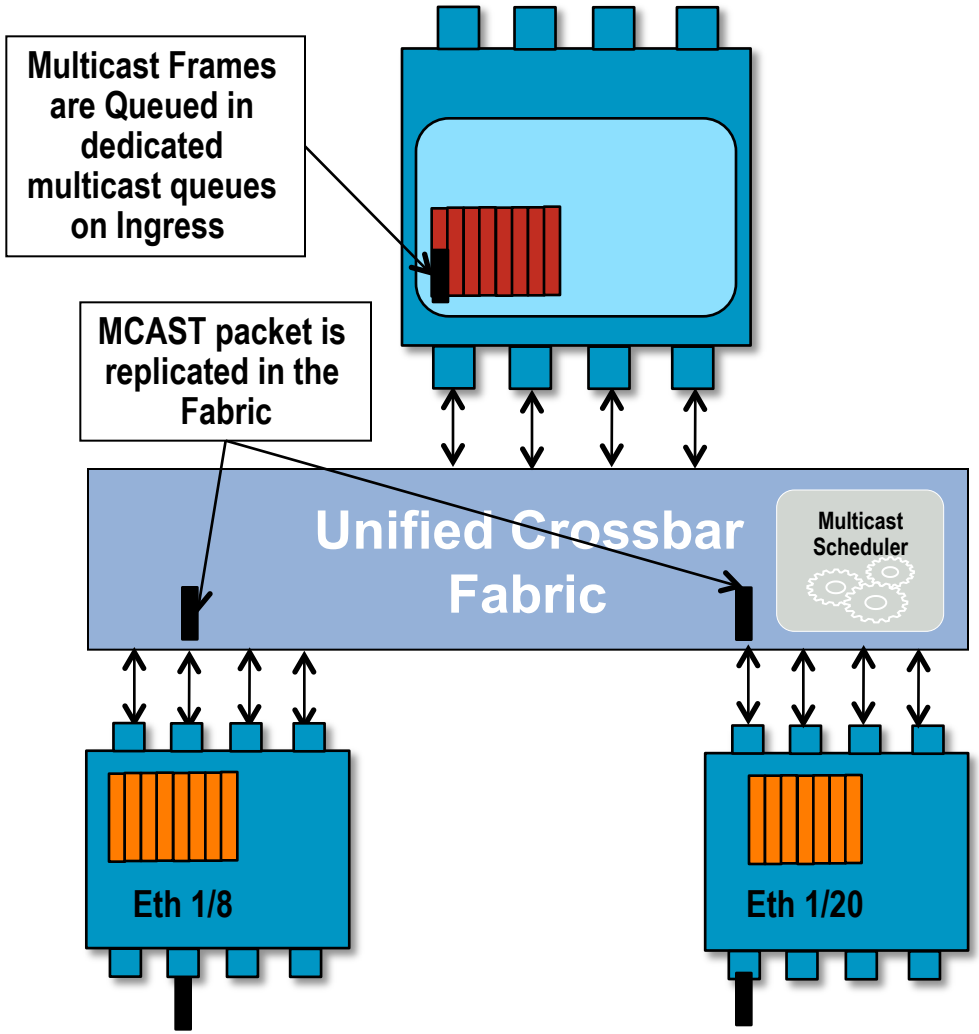
Nexus 5000/5500 and 2000 Architecture

Agenda

- Nexus 5000/5500 Architecture
  - Hardware Architecture
  - Day in the Life of a Packet
  - Layer 3 Forwarding
- Nexus 2000 Architecture
  - FEXLink Architecture
  - FEX Forwarding
  - Extending FEXLink – Adapter FEX
- Nexus 5000/5500
  - Multicast
  - Port Channels
  - QoS

Nexus 5000/5500 Multicast Forwarding

Fabric-Based Replication

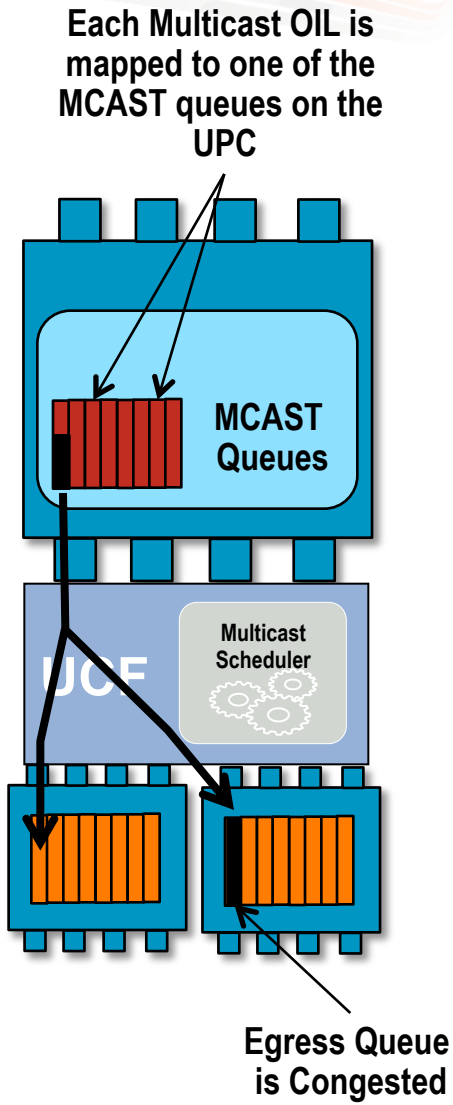


- Nexus 5000 and 5500 use fabric based egress replication
- Traffic is queued in the ingress UPC for each MCAST group
- When the scheduler permits the traffic if forwarded into the fabric and replicated to all egress ports
- When possible traffic is super-framed (multiple packets are sent with a single fabric scheduler grant) to improve throughput

Nexus 5000 Multicast Forwarding

Multicast Queues and Multicast Group Fan-Out

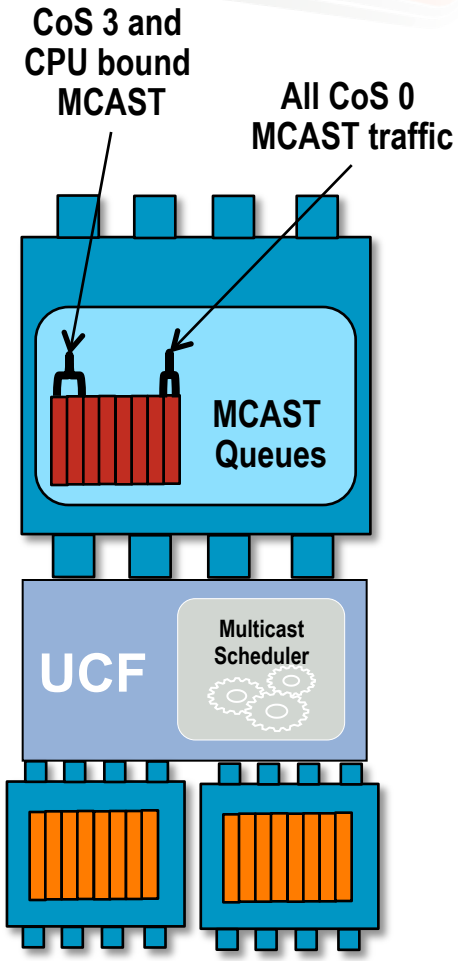
- A “FAN-OUT” = is an Output Interface List (OIL)
- The Nexus 5000 currently supports 1000 fan-outs and 4000 Multicast Groups
- The multicast groups need to be mapped to the 1000 fan-outs
- There are eight multicast queues per UPC forwarding engine (no VoQ for multicast)
- Hardware needs to map fan-outs to the eight queues
- Multicast scheduler waits until all egress queues are free to accept a frame before traffic in that queue is replicated across the fabric



Nexus 5000 Multicast Forwarding

Multicast Queues and Multicast Group Fan-Out

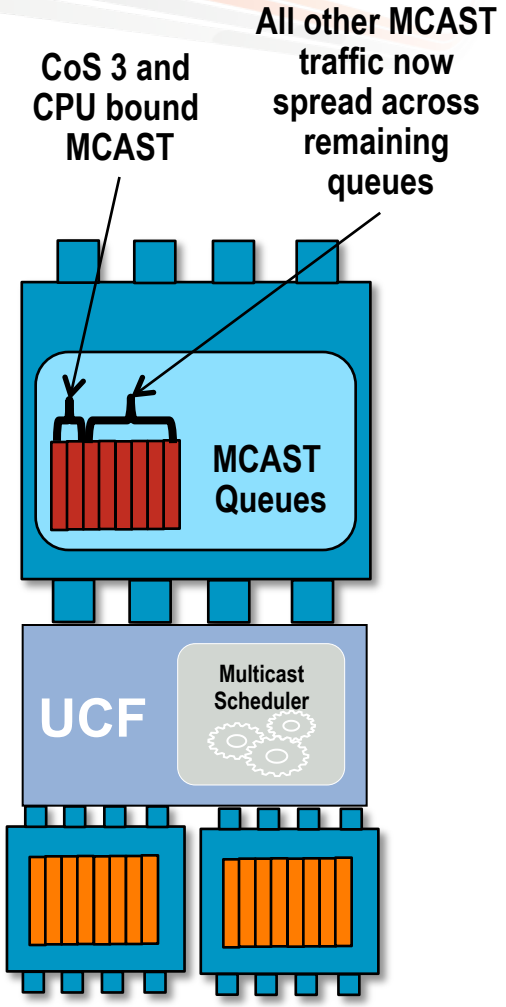
- Overlap of multicast groups to fan-outs to queues can result in contention for the fabric for a specific group
- Tuning of the multicast traffic and fan-out mapping to queues can be used to prioritize specific groups access to the fabric
- Of the eight queues available for multicast two are reserved (FCoE and sup-redirect multicast) leaving six for the remainder of the multicast traffic
- By default the switch uses the frame CoS to identify the multicast queue for a specific group
- If more groups are mapped to one CoS group than another the system queuing for multicast may be non-optimal



Nexus 5000 Multicast Forwarding

Multicast Queues and Multicast-optimization

- "Multicast-optimize" when enabled for a class of traffic assigns multicast fan-outs in that class to any unused CoS queues on a round robin basis
- With multicast optimization, you can assign these classes of traffic to the unused queues
  - One 'class of service' (CoS-based)
  - IP multicast (traffic-based)
  - All flood (traffic-based)

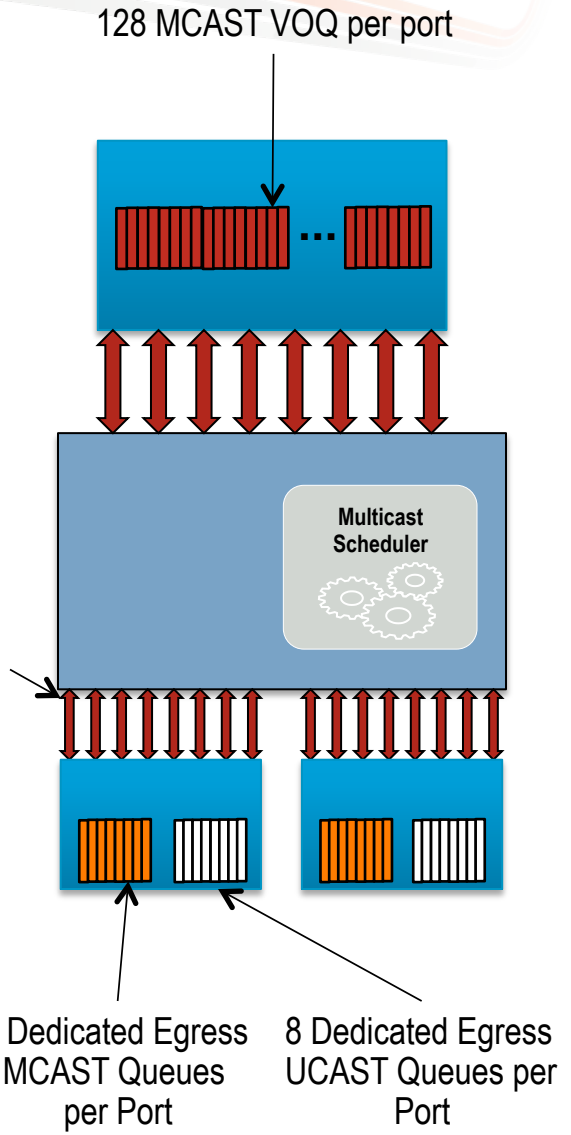



```
class-map type qos class-ip-multicast
policy-map type qos MULTICAST-OPTIMIZE
  class class-ip-multicast
    set qos-group 2
class-map type network-qos class-ip-multicast
  match qos-group 2
policy-map type network-qos MULTICAST-OPTIMIZE
  class type network-qos class-ip-multicast
    multicast-optimize
  class type network-qos class-default
system qos
  service-policy type qos input MULTICAST-OPTIMIZE
  service-policy type network-qos MULTICAST-OPTIMIZE
```

Nexus 5500 Multicast Forwarding

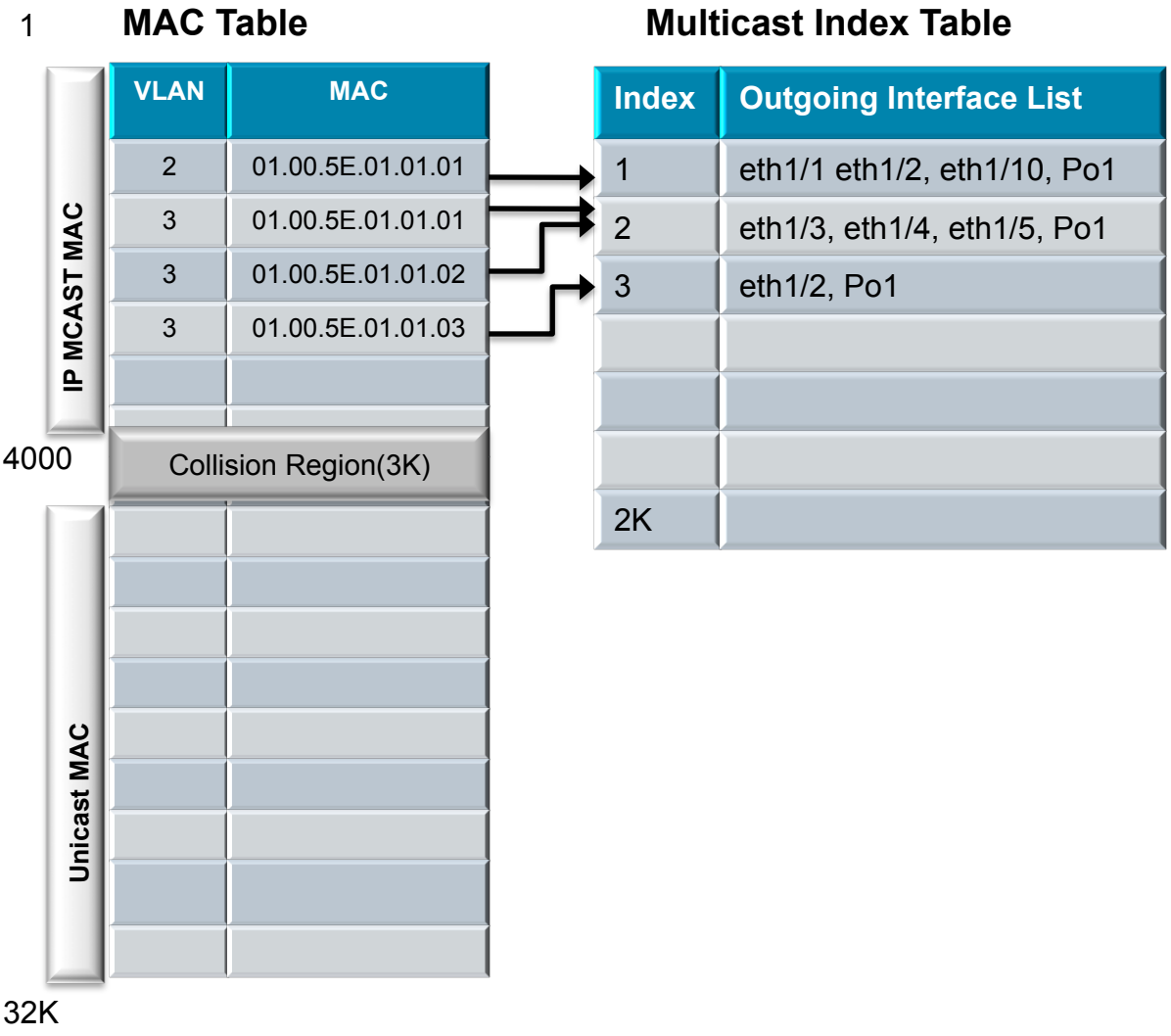
Nexus 5500 Data Plane Changes

- Nexus 5500 supports 4000 IGMP snooping entries
- Dedicated Unicast & Multicast Queuing and Scheduling Resources
  - 128 MCAST VOQ per port
  - 8 for egress queues for unicast and 8 for multicast
  - 4 Egress cross-points (fabric buffer) per egress port
  - Out of 4 fabric buffer, one is used for unicast, one for multicast and two are shared between unicast and multicast
- Two configurable Multicast scheduler modes
- Overloaded mode (Proxy Queue)
  - Congested egress ports are ignored
  - Multicast packets are sent to non-congested port only
- Reliable mode
  - Packets are sent to switch fabric when *all* OIF ports are ready, ie, have fabric buffer and egress buffer to accept the multicast packets



Nexus 5500 Multicast Forwarding

IP Multicast Forwarding Table

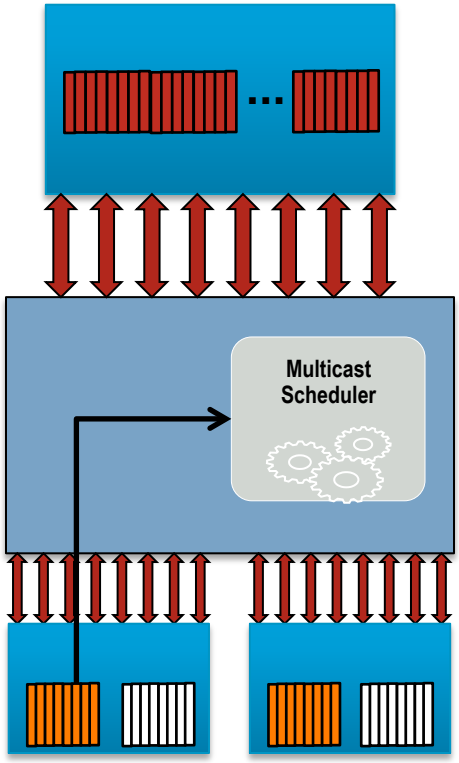


- Multicast IP address is mapped to multicast MAC address with prefix 01.00.5E
- Nexus 5500 checks the destination MAC against the multicast MAC address to make forwarding decision
- IP multicast MAC shares same 32K MAC address table as unicast MAC
- Support 4K groups at FCS
- Multicast Index Table keep tracks of the OIF (Outgoing Interface List) or fanout
- L3 and L4 headers are used for ACL and QoS processing

Nexus 5500 Multicast Forwarding

Nexus 5500 Data Plane Changes

- Proxy queues to detect congestion at egress
- One proxy queue for each hardware egress queue
- Bytes are added to proxy queue when packets arrive at egress hardware queue
- Proxy queues are drained at 98% of port speed using DWRR
- When proxy queue is full egress port sends “overload” message to central scheduler
- Central scheduler excludes the port in multicast scheduling calculation when overload bit is set AND there is no fabric buffer available. Multicast packet is sent over to non-congested port
- In case of congestion there is a delay for proxy queue to signal overload



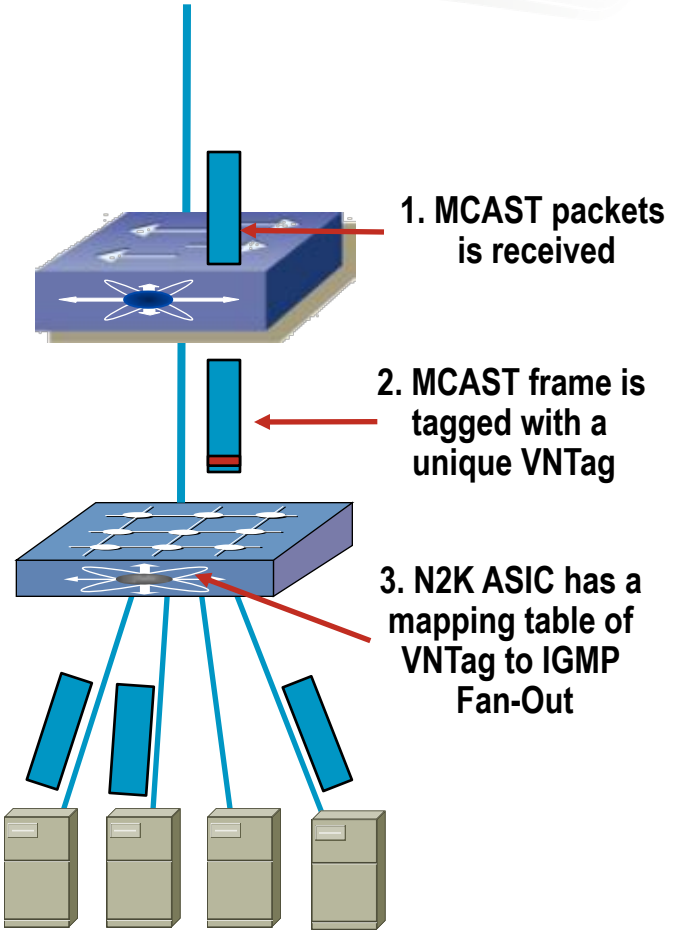
Proxy Queue sends overload signal to scheduler when port congested

```
N5k(config)#hardware multicast disable-slow-port-pruning
```

Nexus Virtualized Access Switch

Nexus 2000 Multicast Forwarding

- Nexus 2000 supports egress based Multicast replication
- Each fabric link has a list of VNTag's associated with each Multicast group
- A single copy of each multicast frame is sent down the fabric links to the Nexus 2000
- Extended Multicast VNTag has an associated flooding fan-out on the Nexus 2000 built via IGMP Snooping
- Nexus 2000 replicates and floods the multicast packet to the required interfaces
- Note: When the fabric links are configured using static pinning each fabric link needs a separate copy of the multicast packet (each pinned group on the Nexus 2000 replicates independently)
- Port Channel based fabric links only require a single copy of the multicast packet



Nexus 5000/5500 and 2000 Architecture

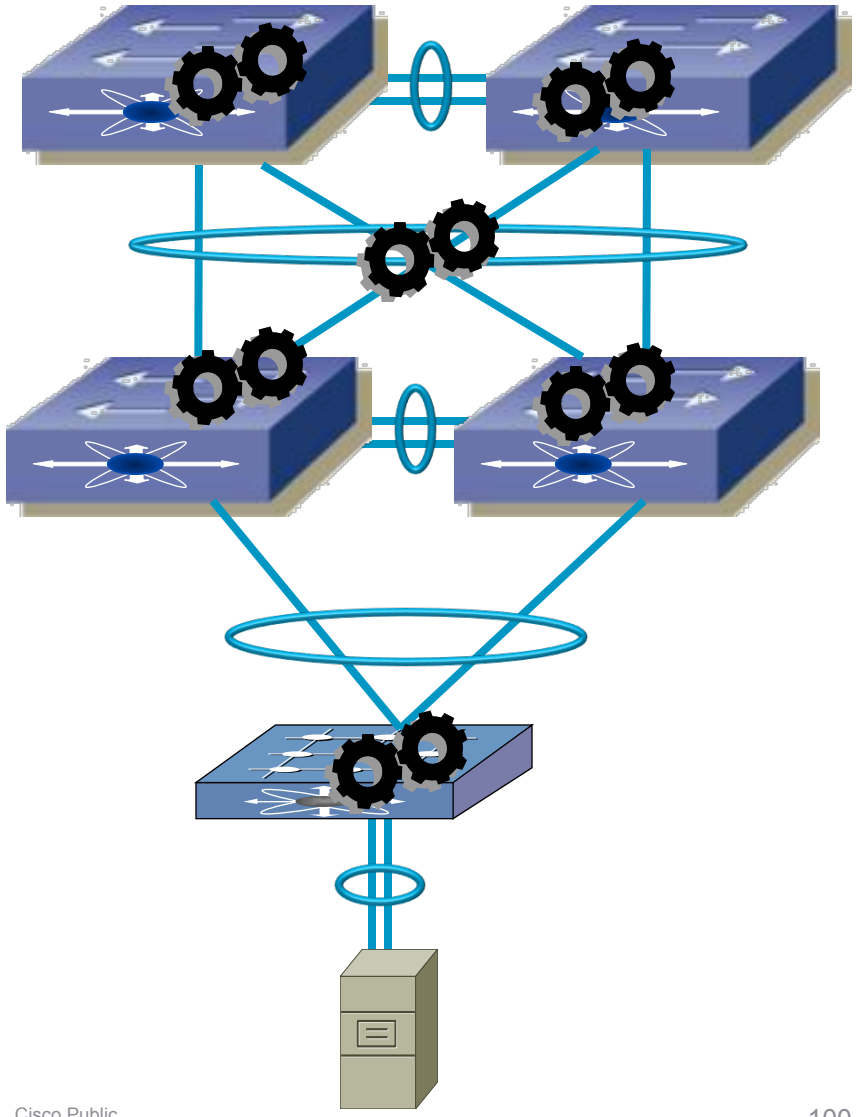
Agenda

- Nexus 5000/5500 Architecture
  - Hardware Architecture
  - Day in the Life of a Packet
  - Layer 3 Forwarding
- Nexus 2000 Architecture
  - FEXLink Architecture
  - FEX Forwarding
  - Extending FEXLink – Adapter FEX
- Nexus 5000/5500
  - Multicast
  - Port Channels
  - QoS

Nexus 5000/5500 Port Channels

Nexus 5000/5500 Port Channel Types

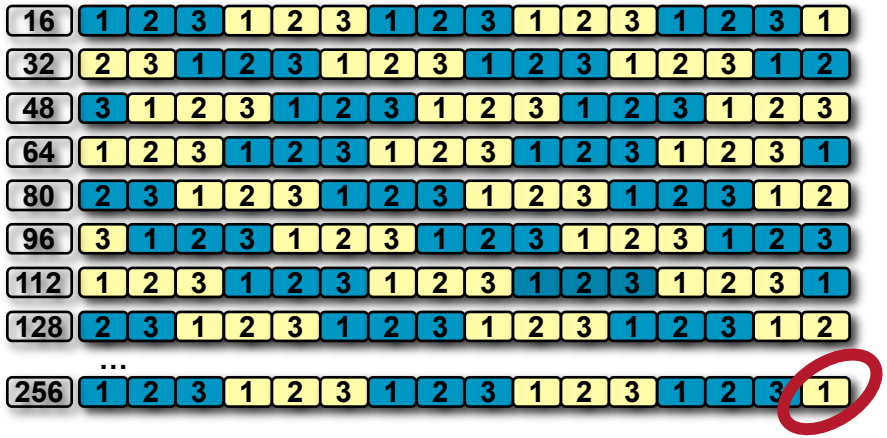
- Nexus 5010/5020 supports 16 port channels of up to 16 links each
- Nexus 5548/5596 support 48 port channels of up to 16 links each
- Nexus 2200 FEX supports 24 port channels of up to 8 links each
- Port channels configured on FEX do not take any resource from the Nexus 5000/5500 switch
- Nexus 5500 LIF port channels (MLID) do not consume a HW port channel resource
- Nexus 5548/5596 support up 768 vPC port channels



Nexus 5000/5500 Port Channels

Nexus 5000/5500 Port Channel Efficiency

- Prior generations of Etherchannel load sharing leveraged eight hash buckets
- Could lead to non optimal load sharing with an odd number of links
- Nexus 5000 and 2000 utilize 256 buckets
- Provides better load sharing in normal operation and avoids in-balancing of flows in any link failure cases

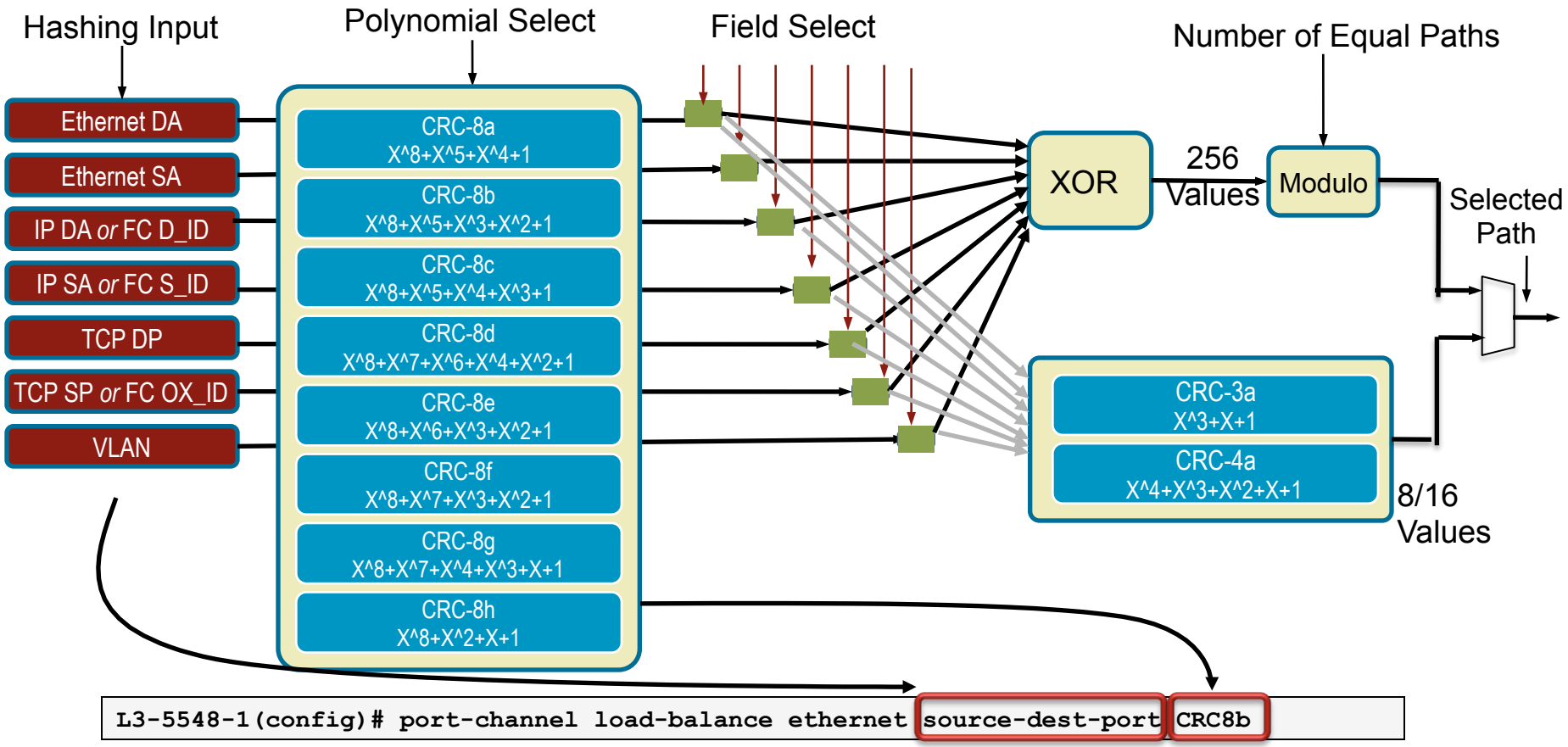



```
dc11-5020-3# sh port-channel load-balance forwarding-path interface port-channel 100
dst-ip 10.10.10.10 src-ip 11.11.11.11
Missing params will be substituted by 0's.
Load-balance Algorithm: source-dest-ip
crc8_hash: 24   Outgoing port id: Ethernet1/37
```

Nexus 5000/5500 Port Channels

Nexus 5000/5500 Port Channel Efficiency

- Nexus 5500 increases potential randomization to hashing
  - VLAN added to hash input
  - Increased number of polynomials and two stage hashing



Nexus 2000 Port Channels

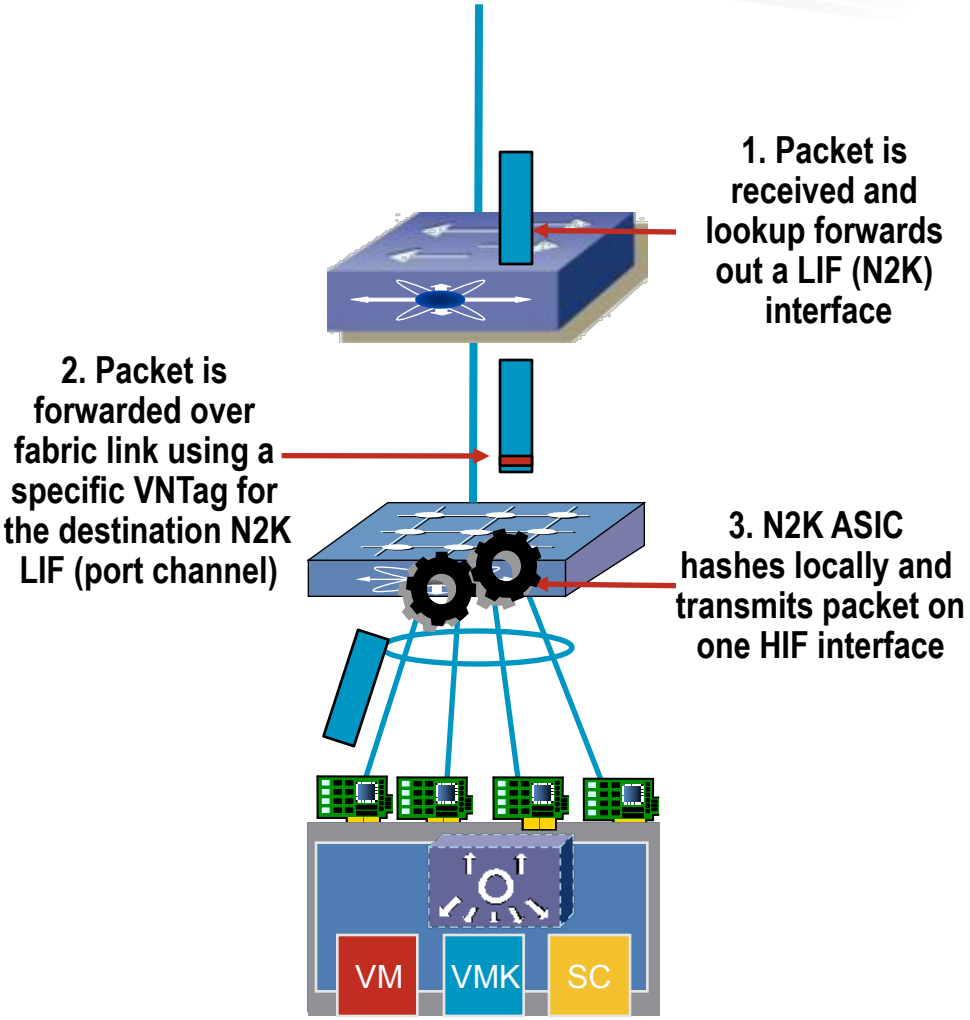
Nexus 2248/2232 Port Channels

- Nexus 2200 series FEX support local port channels
- All FEX ports are extended ports (Logical Interfaces = LIF)
- A local port channel on the N2K is still seen as a single extended port
- Extended ports are each mapped to a specific VNTag
- HW hashing occurs on the N2K ASIC
- Number of 'local' port channels on each N2K is based on the local ASIC

  2148T – 0

  2248T – 24

  2232 - 16



Nexus 5000/5500 and 2000 Architecture

Agenda

- Nexus 5000/5500 Architecture
  - Hardware Architecture
  - Day in the Life of a Packet
  - Layer 3 Forwarding
- Nexus 2000 Architecture
  - FEXLink Architecture
  - FEX Forwarding
  - Extending FEXLink – Adapter FEX
- Nexus 5000/5500
  - Multicast
  - Port Channels
  - QoS

Nexus 5000/5500 QoS

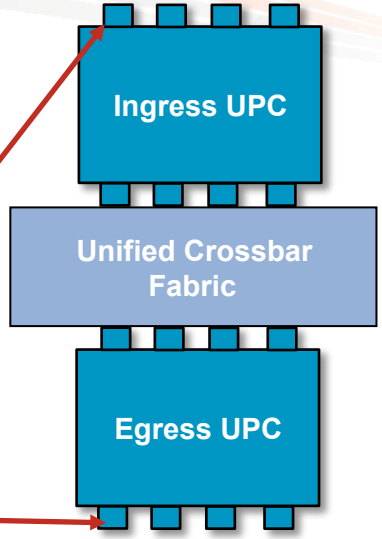
QoS Capabilities and Configuration

- Nexus 5000 supports a new set of QoS capabilities designed to provide per system class based traffic control
  - Lossless Ethernet—Priority Flow Control (IEEE 802.1Qbb)
  - Traffic Protection—Bandwidth Management (IEEE 802.1Qaz)
  - Configuration signaling to end points—DCBX (part of IEEE 802.1Qaz)
- These new capabilities are added to and managed by the common Cisco MQC (Modular QoS CLI) which defines a three-step configuration model
  - Define matching criteria via a *class-map*
  - Associate action with each defined class via a *policy-map*
  - Apply policy to entire system or an interface via a *service-policy*
- Nexus 5000/7000 leverage the MQC qos-group capabilities to identify and define traffic in policy configuration

Nexus 5000/5500 QoS

QoS Policy Types

- There are three QoS policy types used to define system behavior (qos, queuing, network-qos)
- There are three policy attachment points to apply these policies to
  - Ingress interface
  - System as a whole (defines global behavior)
  - Egress interface



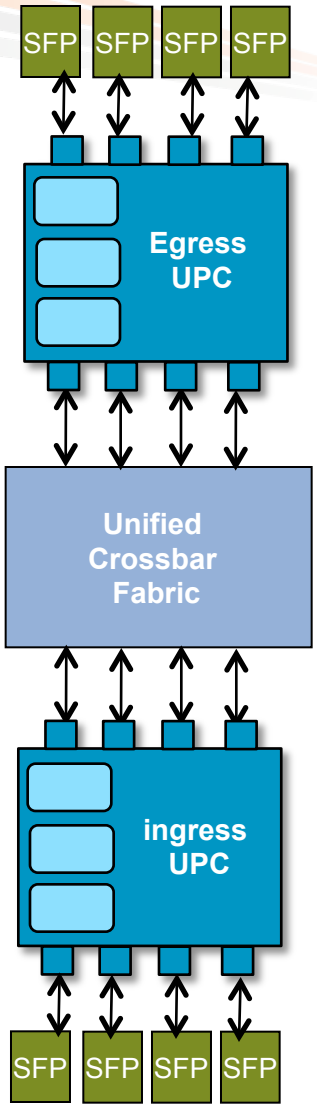
Policy Type	Function	Attach Point
qos	Define traffic classification rules	system qos ingress Interface
queuing	Strict Priority queue Deficit Weight Round Robin	system qos egress Interface ingress Interface
network-qos	System class characteristics (drop or no-drop, MTU), Buffer size, Marking	system qos

Nexus 5000 QoS

UPC (Gen 1) QoS Defaults

- QoS is enabled by default (not possible to turn it off)
- Four default class of services defined when system boots up
  - Two for control traffic (CoS 6 & 7)
  - One for FCoE traffic (class-fcoe – CoS 3)
  - Default Ethernet class (class-default – all others)
- You can define up to four additional system classes for Ethernet traffic.
- Control traffic is treated as strict priority and serviced ahead of data traffic
- The two base user classes (class-fcoe and class-default) get 50% of guaranteed bandwidth by default

```
dc11-5020-2# sh policy-map system type qos input
<snip>
    Class-map (qos):   class-fcoe (match-any)
      Match: cos 3
      set qos-group 1

    Class-map (qos):   class-default (match-any)
      Match: any
      set qos-group 0
```



Nexus 5000 QoS

UPC (Gen 1) Buffering

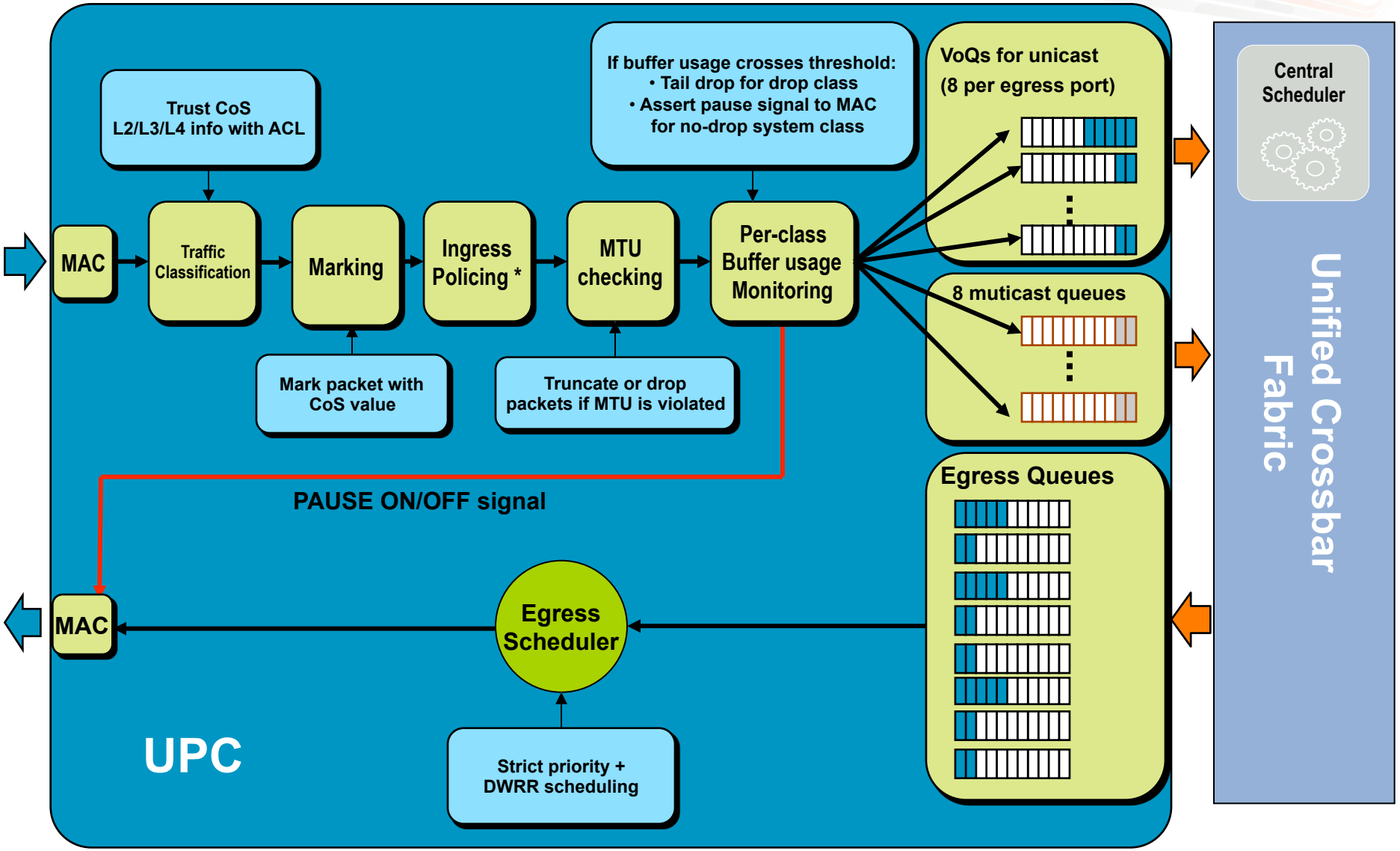
- 480KB dedicated packet buffer per one 10GE port or per two FC ports
- Buffer is shared between ingress and egress with majority of buffer being allocated for ingress
  - Ingress buffering model
  - Buffer is allocated per system class
  - Egress buffer only for in flight packet absorption
- Buffer size of ingress queues for drop class can be adjusted using *network-qos* policy

Class of Service	Ingress Buffer(KB)	Egress Buffer(KB)
Class-fcoe	76.8	18.8
Sup-Hi & Sup-Lo	18.0 & 18.0	9.6 & 9.6
User defined no-drop class of service with MTU<2240	76.8	18.8
User defined no-drop class of service with MTU>2240	81.9	18.8
Tail drop class of service	20.4	18.8
Class-default	All remaining buffer	18.8

Default Classes

Nexus 5000 QoS

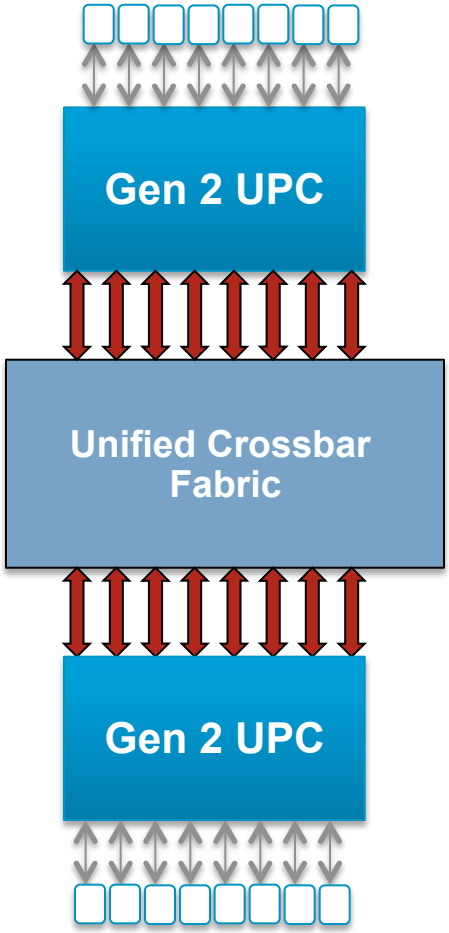
UPC (Gen 1) QoS Capabilities (*Not Currently Supported)



Nexus 5500 QoS
UPC (Gen 2) QoS Defaults

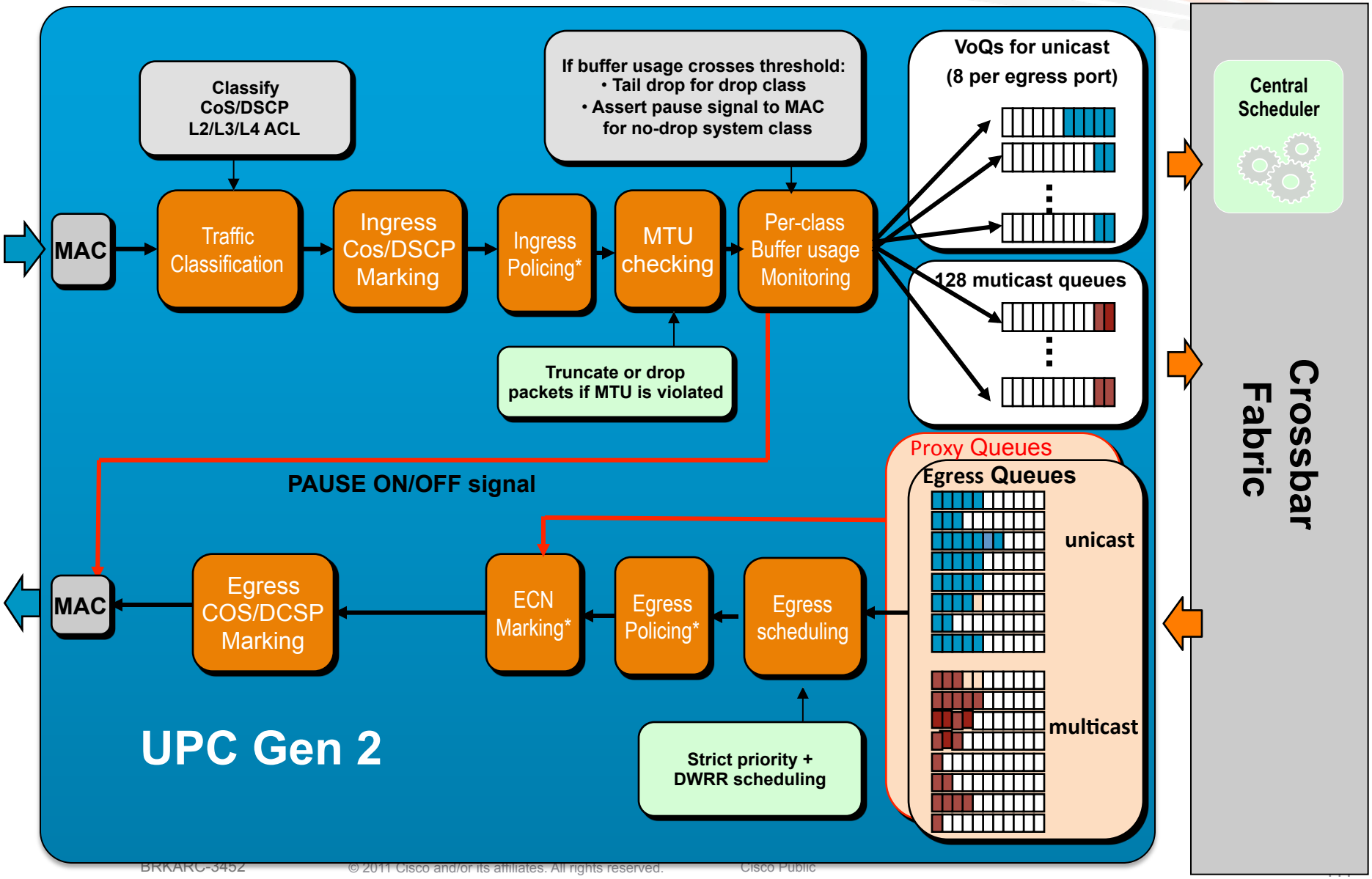
- QoS is enabled by default (not possible to turn it off)
- Three default class of services defined when system boots up
  - Two for control traffic (CoS 6 & 7)
  - Default Ethernet class (class-default – all others)
- Cisco Nexus 5500 switch supports five user-defined classes and the one default drop system class
- FCoE queues are ***'not'*** pre-allocated
- When configuring FCoE the predefined service policies must be added to existing QoS configurations

```
# Predefined FCoE service policies
service-policy type qos input fcoe-default-in-policy
service-policy type queuing input fcoe-default-in-policy
service-policy type queuing output fcoe-default-out-policy
service-policy type network-qos fcoe-default-nq-policy
```



Nexus 5500 QoS

UPC (Gen 2) QoS Capabilities (*Not Currently Supported)



Nexus 5500 QoS
UPC (Gen 2) Buffering

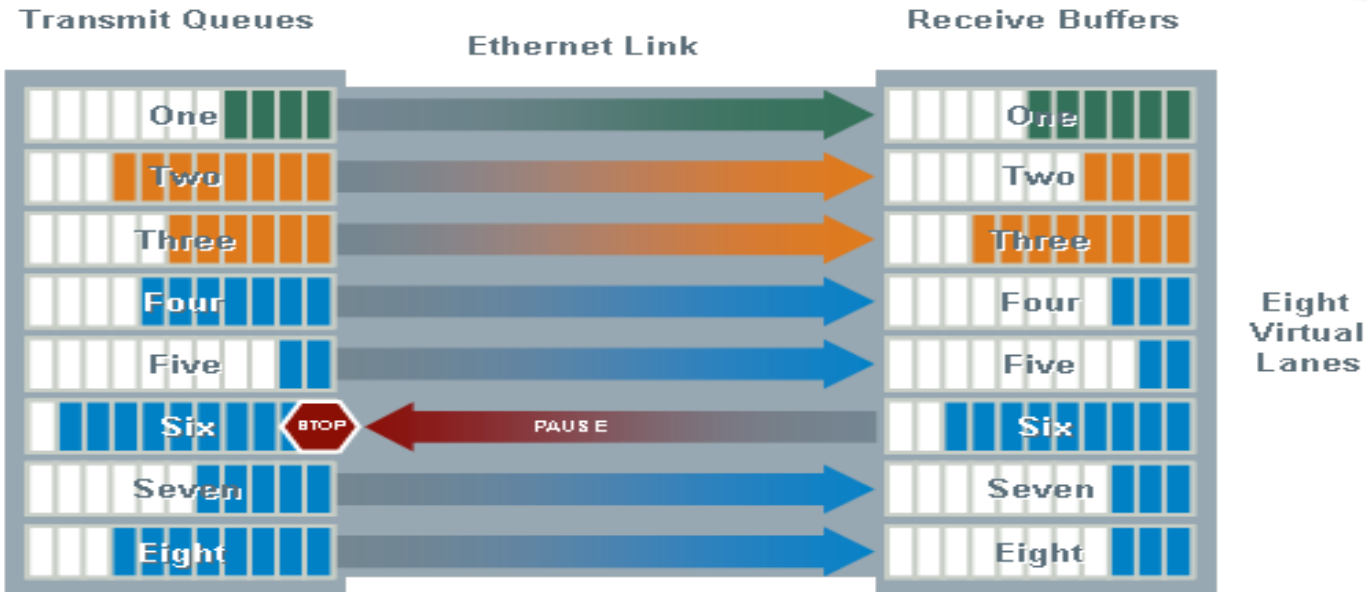
- 640KB dedicated packet buffer per one 10GE port
- Buffer is shared between ingress and egress with majority of buffer being allocated for ingress
  - Ingress buffering model
  - Buffer is allocated per system class
  - Egress buffer only for in flight packet absorption
- Buffer size of ingress queues for drop class can be adjusted using *network-qos* policy

Class of Service	Ingress Buffer(KB)	Egress Buffer(KB)
Class-fcoe	78	19
Sup-Hi & Sup-Lo	18.0 & 18.0	9.6 & 9.6
User defined no-drop class of service with MTU<2240	78	19
User defined no-drop class of service with MTU>2240	88	19
User defined tail drop class of service with MTU<2240	22	19
User defined tail drop class of service with MTU>2240	29	19
Class-default	All remaining buffer	19

Default Classes

Nexus 5000/5500 QoS

Priority Flow Control and No-Drop Queues



- Nexus 5000 supports a number of new QoS concepts and capabilities
- Priority Flow Control is an extension of standard 802.3x pause frames
- No-drop queues provide the ability to support loss-less Ethernet using PFC as a per queue congestion control signaling mechanism

Nexus 5000/5500 QoS

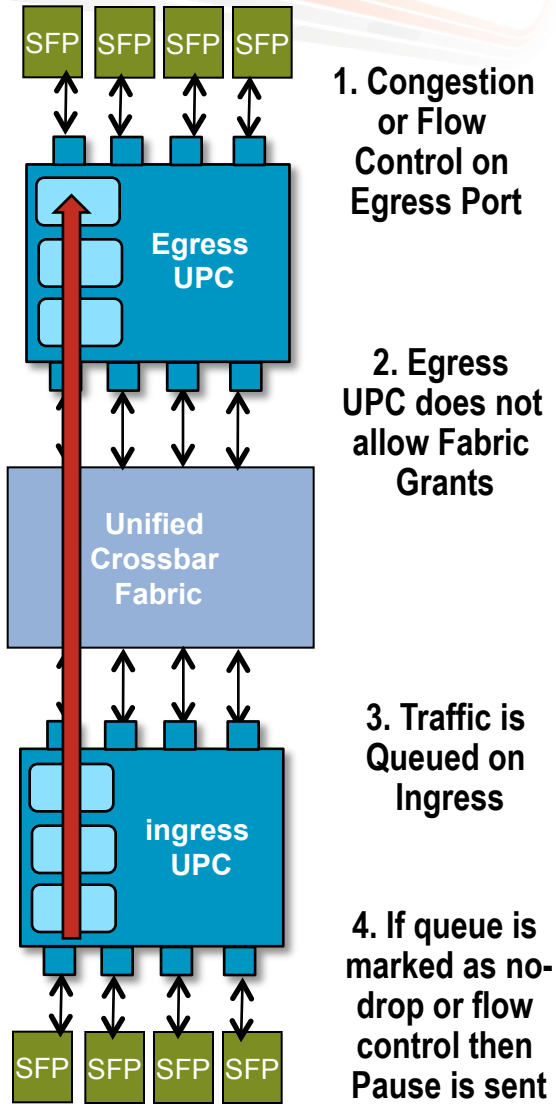
Priority Flow Control and No-Drop Queues

- Actions when congestion occurs depending on policy configuration

  PAUSE upstream transmitter for lossless traffic

  Tail drop for regular traffic when buffer is exhausted

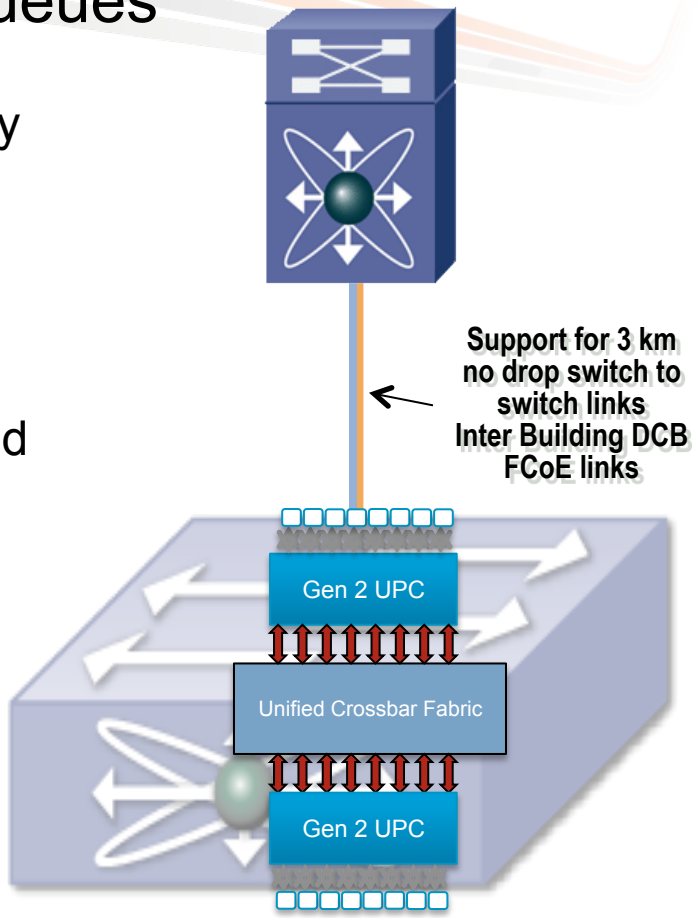
- Priority Flow Control (PFC) or 802.3X PAUSE can be deployed to ensure lossless for application that can't tolerate packet loss
- Buffer management module monitors buffer usage for no-drop class of service. It signals MAC to generate PFC (or link level PAUSE) when the buffer usage crosses threshold
- FCoE traffic is assigned to *class-fcoe*, which is a no-drop system class
- Other class of service by default have normal drop behavior (tail drop) but can be configured as no-drop



Nexus 5000/5500 QoS

Priority Flow Control and No-Drop Queues

- Tuning of the lossless queues to support a variety of use cases
- Extended switch to switch no drop traffic lanes
  - Support for 3km with Nexus 5000 and 5500
  - Increased number of no drop services lanes (4) for RDMA and other multi-queue HPC and compute applications



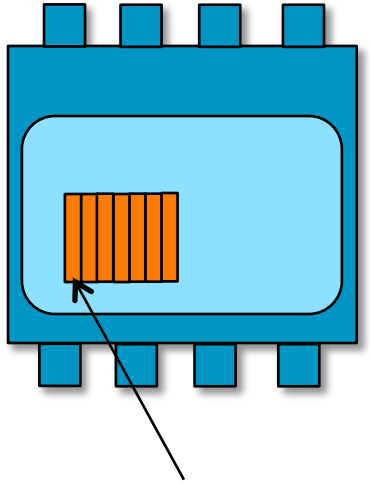
Configs for 3000m no-drop class	Buffer size	Pause Threshold (XOFF)	Resume Threshold (XON)
N5020	143680 bytes	58860 bytes	38400 bytes
N5548	152000 bytes	103360 bytes	83520 bytes

```
5548-FCoE(config)# policy-map type network-qos 3km-FCoE
5548-FCoE(config-pmap-nq)# class type network-qos 3km-FCoE
5548-FCoE(config-pmap-nq-c)# pause no-drop buffer-size 152000 pause-threshold 103360
resume-threshold 83520
```

Nexus 5000/5500 QoS
MTU per Class of Service (CoS Queue)

- MTU can be configured for each class of service (no interface level MTU)
- No fragmentation since Nexus 5000 is a L2 switch
- When forwarded using cut-through, frames are truncated if they are larger than MTU
- When forwarded using store-and-forward, frames are dropped if they are larger than MTU

```
class-map type qos iSCSI
  match cos 2
class-map type queuing iSCSI
match qos-group 2
policy-map type qos iSCSI
  class iSCSI
    set qos-group 2
class-map type network-qos iSCSI
  match qos-group 2
policy-map type network-qos iSCSI
  class type network-qos iSCSI
    mtu 9216
system qos
  service-policy type qos input iSCSI
  service-policy type network-qos iSCSI
```

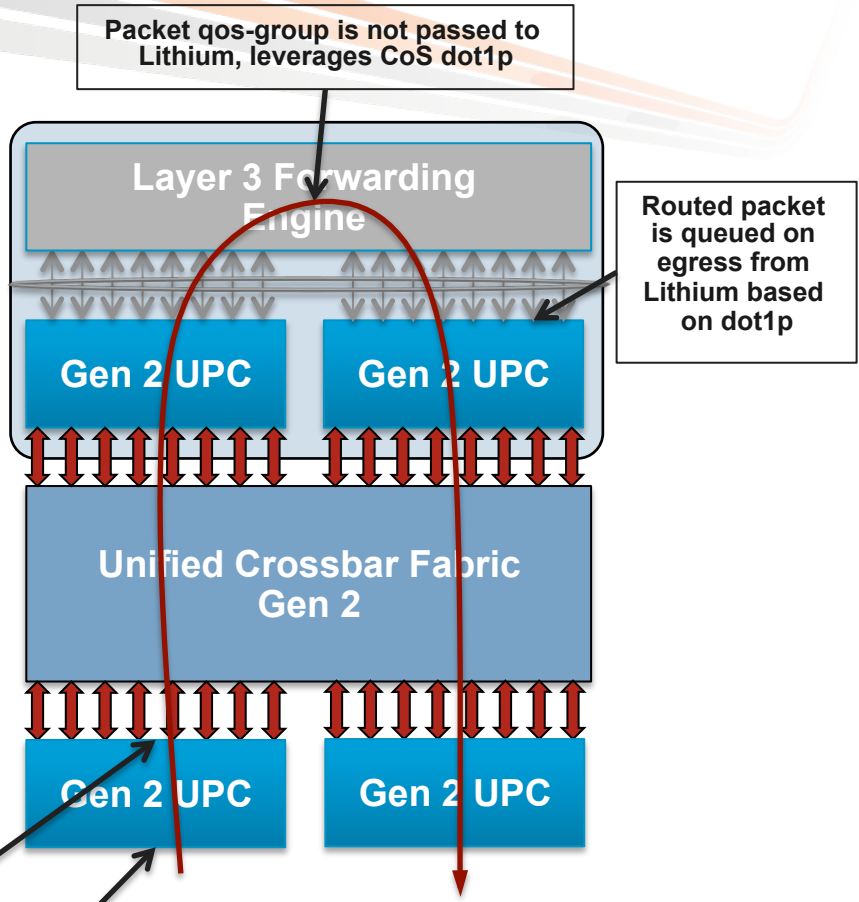


Each CoS queue on the Nexus 5000 supports a unique MTU

Nexus 5500 Series

Layer 3 QoS Configuration

- Internal QoS information determined by ingress Carmel (UPC) ASIC is 'not' passed to the Lithium L3 ASIC
- Need to mark all routed traffic with a dot1p CoS value used to:
  - Queue traffic to and from the Lithium L3 ASIC
  - Restore qos-group for egress forwarding
- Mandatory to setup CoS for the frame in the network-qos policy, one-to-one mapping between a qos-group and CoS value
- Classification can be applied to *physical interfaces* (L2 or L3, including L3 port-channels) not to SVIs



If traffic is congested on ingress to L3 ASIC it is queued on ingress UPC ASIC

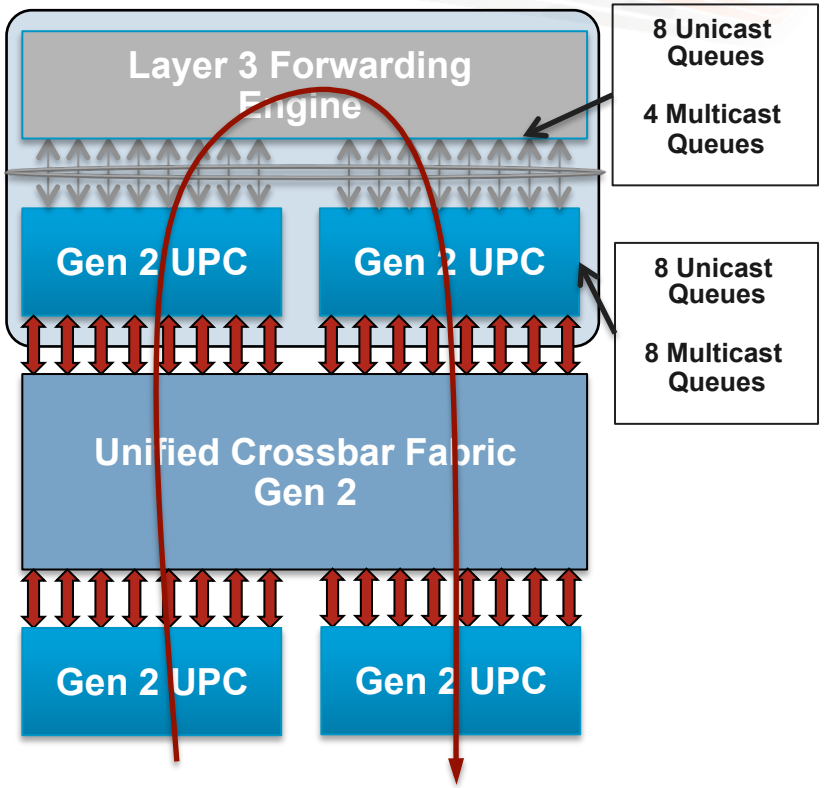
On initial ingress packet QoS matched and packet is associated with a qos-group for queuing and policy enforcement

```
class-map type network-qos nqcm-grp2
     match qos-group 2

class-map type network-qos nqcm-grp4
     match qos-group 4

policy-map type network-qos nqpm-grps
     class type network-qos nqcm-grp2
          set cos 4
     class type network-qos nqcm-grp4
          set cos 2
```

Nexus 5500 Series

Layer 3 QoS Configuration

- Apply “type qos” and network-qos policy for classification on the L3 interfaces and on the L2 interfaces (or simply system wide)
- Applying “type queuing” policy at system level in egress direction (output)
- Trident has CoS queues associated with every interface
  - 8 Unicast CoS queues
  - 4 Multicast CoS queues
- The individual dot1p priorities are mapped one-to-one to the Unicast CoS queues
  - This has the result of dedicating a queue for every traffic class
- With the availability of only 4 multicast queues the user would need to explicitly map dot1p priorities to the multicast queues
- wrr-queue cos-map <queue ID> <CoS Map>

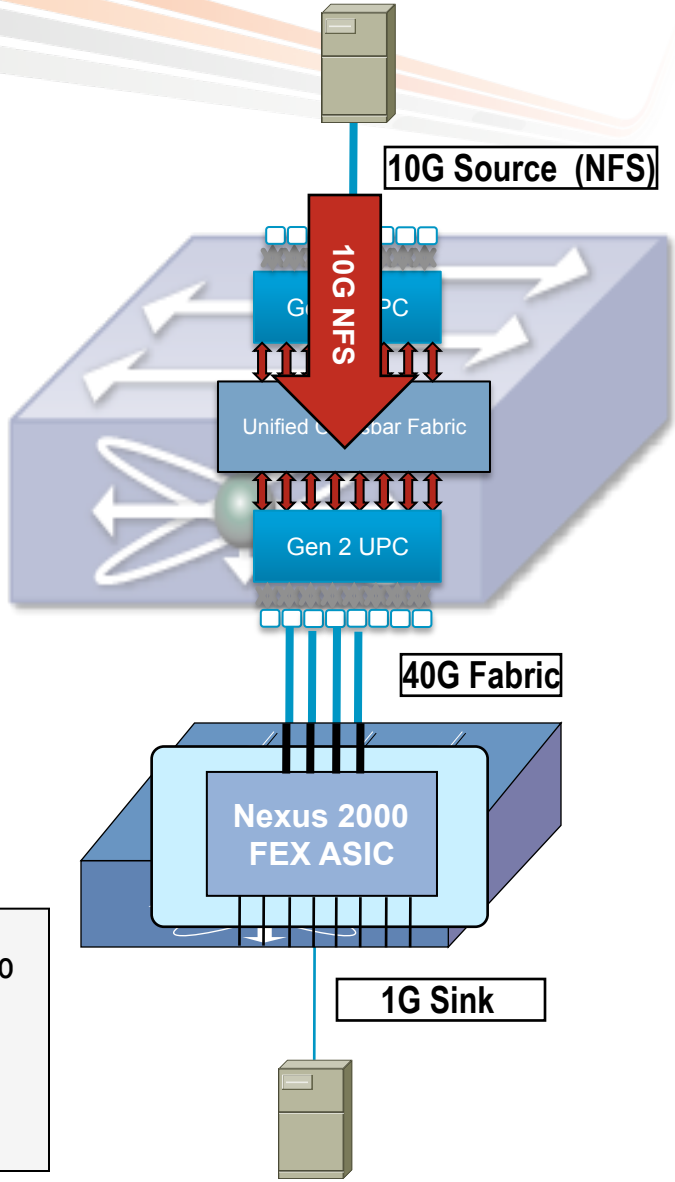



```
Nexus-5500(config)# wrr-queue cos-map 0 1 2 3

Nexus-5500(config)# sh wrr-queue cos-map
MCAST Queue ID            Cos Map
0                         0 1 2 3
1
2                         4 5
3                         6 7
```

Nexus 2000 QoS

Tuning the Port Buffers

- Each Fabric Extender (FEX) has local port buffers
- You can control the queue limit for a specified Fabric Extender for egress direction (from the network to the host)
- You can use a lower queue limit value on the Fabric Extender to prevent one blocked receiver from affecting traffic that is sent to other non-congested receivers ("head-of-line blocking")
- A higher queue limit provides better burst absorption and less head-of-line blocking protection



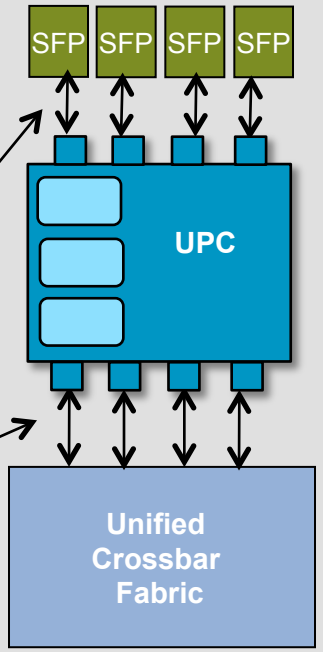

```
dc11-5020-3(config)# fex 100
dc11-5020-3(config-fex)# hardware N2248T queue-limit 356000


dc11-5020-3(config-fex)# hardware N2248T queue-limit ?
  <CR>
  <2560-652800>  Queue limit in bytes
```

Nexus 5000/5500 QoS

Mapping the Switch Architecture to 'show queuing'

```
dc11-5020-4# sh queuing int eth 1/39

Interface Ethernet1/39 TX Queuing
qos-group  sched-type  oper-bandwidth
    0       WRR             50
    1       WRR             50

Interface Ethernet1/39 RX Queuing
qos-group  0
    q-size: 243200, HW MTU: 1600 (1500 configured)
    drop-type: drop, xon: 0, xoff: 1520
    Statistics:
        Pkts received over the port             : 85257
        Ucast pkts sent to the cross-bar        : 930
        Mcast pkts sent to the cross-bar        : 84327
        Ucast pkts received from the cross-bar  : 249
        Pkts sent to the port                   : 133878
        Pkts discarded on ingress               : 0
        Per-priority-pause status               : Rx (Inactive), Tx (Inactive)

<snip - other classes repeated>

Total Multicast crossbar statistics:
    Mcast pkts received from the cross-bar      : 283558
```

Egress (Tx) Queuing Configuration

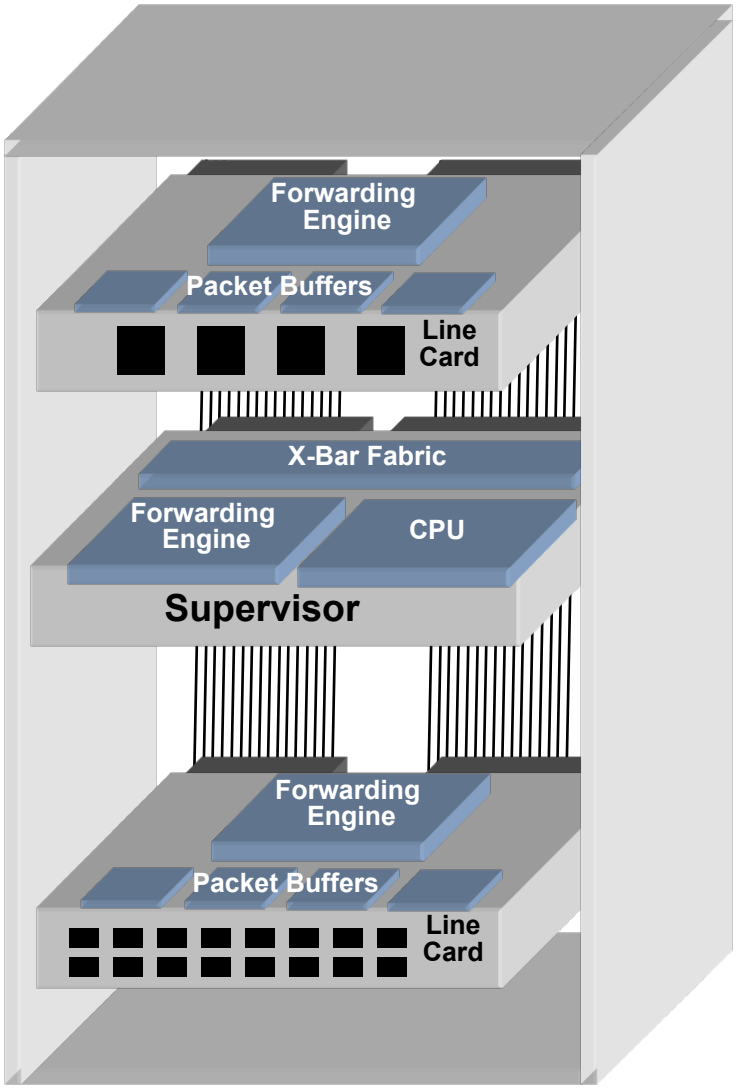


Packets Arriving on this port but dropped from ingress queue due to congestion on egress port

Nexus 5000/5500 and 2000 Architecture

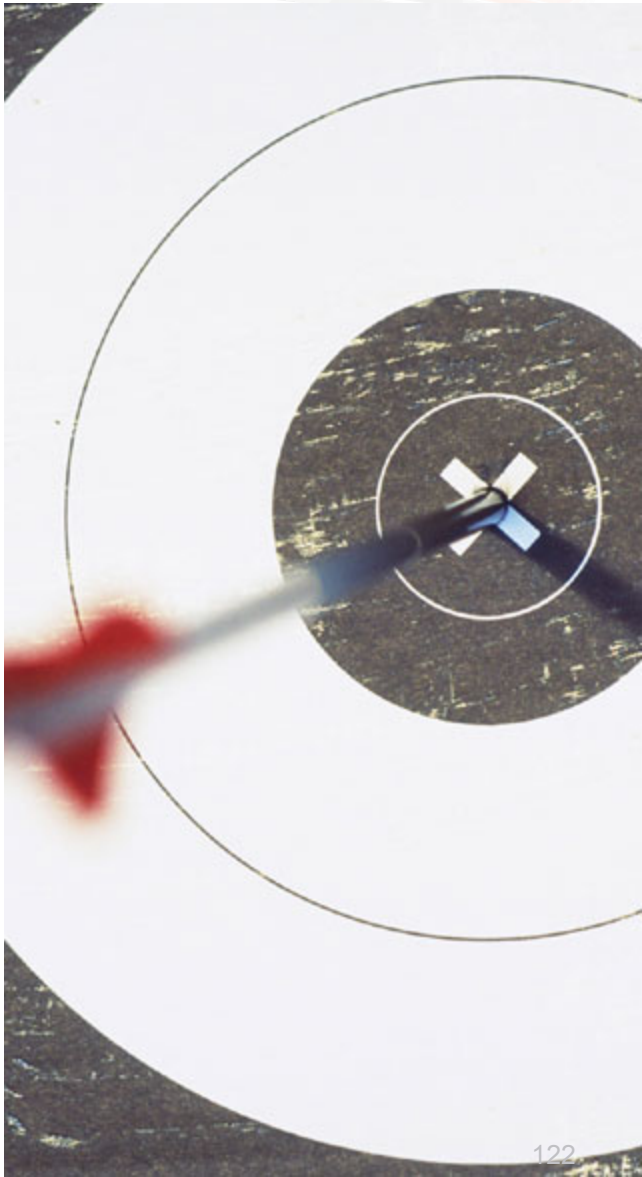
Data Center Switch

- The functional elements of the Nexus 5000/5500 and 2000 are familiar
  - Distributed forwarding—L2/L3 forwarding, ACL, QoS TCAM
  - Protected management and control plane
  - Non-blocking cross bar switching fabric
  - Flexible connectivity through multiple line cards
- Some new capabilities and physical form factor
  - QoS - DCB, per class MTU, no-drop queues and VoQ
  - Multiprotocol—Ethernet and FC/FCoE forwarding
  - Remote Line Cards (FEX & VNTag)



Conclusion

- You should now have a thorough understanding of the Nexus 5000/5500 Data Center switches and the Nexus 2000 Fabric Extender packet flows, and key forwarding engine functions…
- **Any questions?**



Complete Your Online Session Evaluation

- Receive 25 Cisco Preferred Access points for each session evaluation you complete.
- Give us your feedback and you could win fabulous prizes. Points are calculated on a daily basis. Winners will be notified by email after July 22nd.
- Complete your session evaluation online now (open a browser through our wireless network to access our portal) or visit one of the Internet stations throughout the Convention Center.
- Don't forget to activate your Cisco Live and Networkers Virtual account for access to all session materials, communities, and on-demand and live activities throughout the year. Activate your account at any internet station or visit www.ciscolivevirtual.com.

Learn. Connect. Collaborate. together.

Visit the Cisco Store for Related Titles

http://theciscostores.com

Thank you.