
TOMORROW
starts here.

Nexus 9000 Architecture

BRKDCT-3640

Mike Herbert

Principal Engineer

#clmel

Cisco live!

What is our goal today?

Agenda
Nexus 9000 Architecture

- Nexus 9000
  - Nexus 9000 Hardware
  - Nexus 9500 Chassis
  - Nexus 9500 Line Cards
  - Nexus 9500 Packet Forwarding
  - Nexus 9300
- Nexus 9000 and 40G
- Nexus 9000 Designs: FEX, vPC & VXLAN
- Nexus 9000 & Dev-Ops
- ACI & Nexus 9000

Cisco Nexus 9000 Series Switches

High Performance 10/40/100 Gbps Switch Family

SCALABLE 1 GE/10 Gbps/40 Gbps/100 GE PERFORMANCE

Nexus® 9300

- FCS Q12014 — 48 1/10G SFP+ & 12 QSFP+

- FCS Q1 2014 — 96 1/10G-T & 8 QSFP+

- FCS Q1 2014 — 12-port QSFP+ GEM

Nexus 9500

- FCS Q4 2013 — Aggregation line card 36 40G QSFP+

- FCS Q1 2014 — ACI Ready Leaf Line Card 48 1/10G-T & 4 QSFP+

- FCS Q1 2014 — ACI-ready Leaf line card 48 1/10G SFP+ & 4 QSFP+

- FCS Q4 2013 — C9500 8-Slot

FLEXIBLE FORM FACTORS CAN ENABLE VARIABLE DATA Centre DESIGN AND SCALING

PERFORMANCE | PORTS | PRICE | POWER | PROGRAMMABILITY

Nexus 9500 Platform Architecture

Overview

High port density

-- 288 x 40 Gbps/Nexus 9508 -- 1152 x 10 Gbps/Nexus 9508

Layer 2 and Layer 3 line-rate performance on all ports and all packet sizes

Low latency

- Up to 3.5 usec on the 36 x 40GE QSFP line card (N9K-X9636PQ)

Power efficiency

- Platinum-rated power supplies; 90-94% power efficiency across all workloads
- 3.5 W/10 Gbps port; 15 W/40 Gbps port

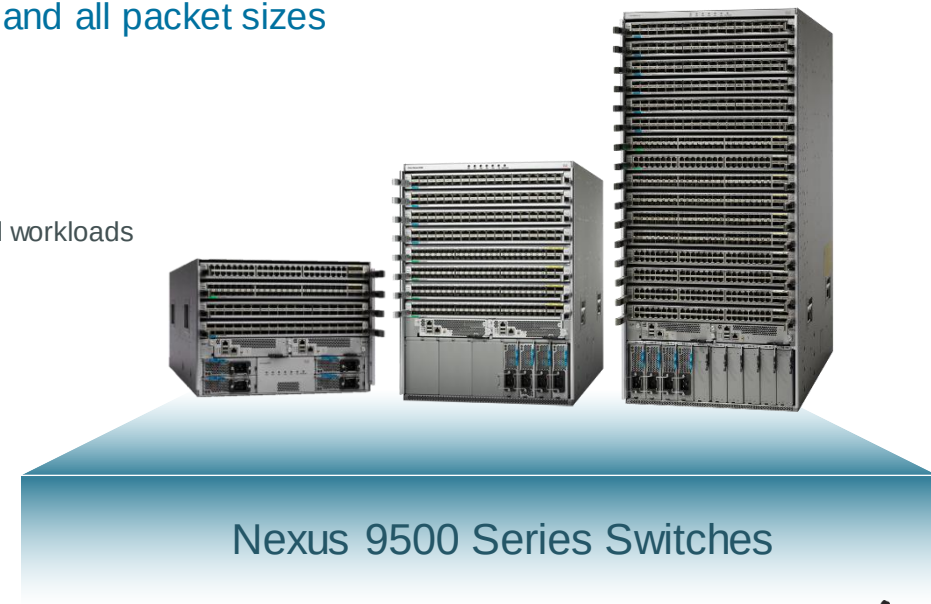
First modular chassis without a mid-plane

- Unobstructed front-to-back airflow

VXLAN bridging, gateway, routing

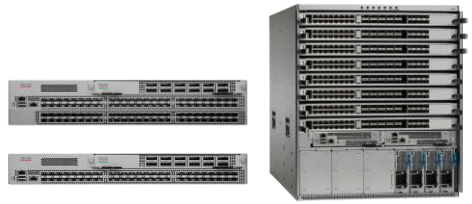
Highly integrated switch and buffer functionality

- Only 2 to 4 ASICs per line card
- No buffer bloat
- Mix of 28 nm Cisco® and 40 nm Broadcom ASICs



Merchant and Custom ASICs (Merchant+)

- ✧ Merchant+ Strategy
- ✧ Best Performance and Functionalities
- ✧ Optimal Pricing

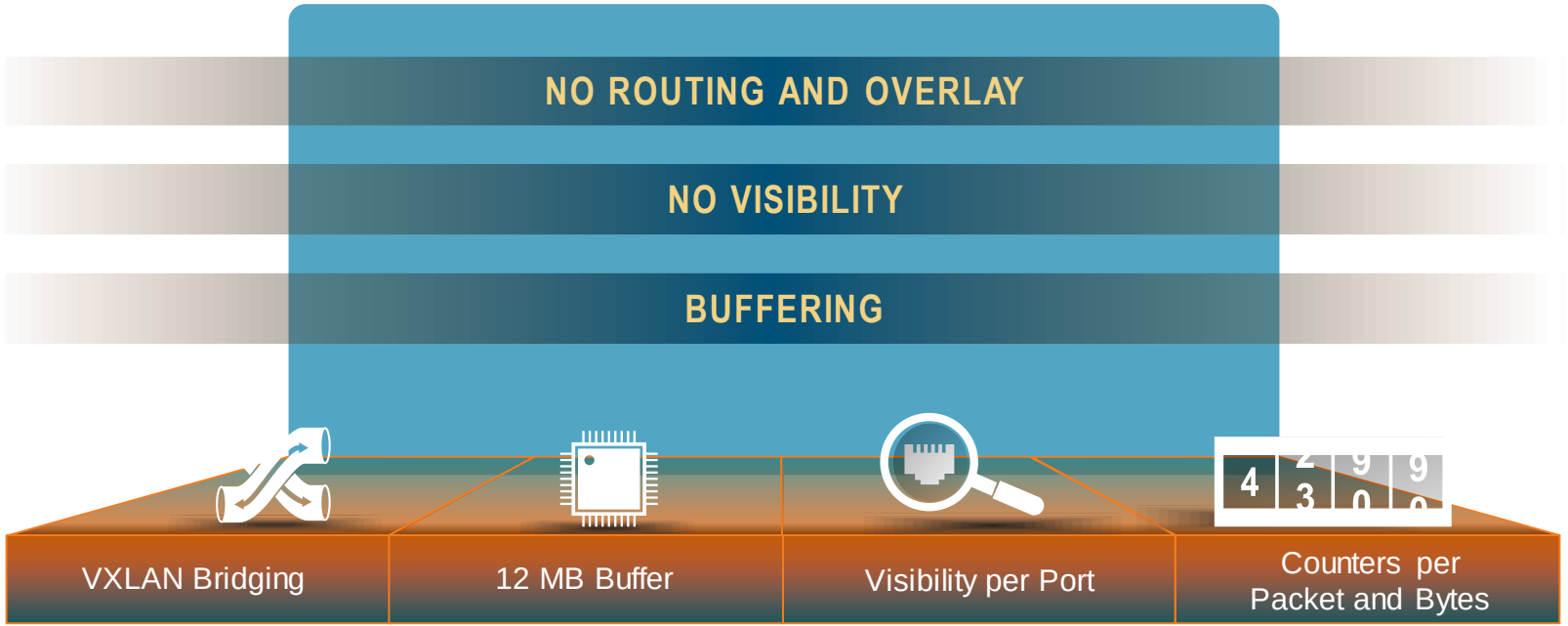


	NFE	ALE	ASE
ASIC Technology	40 nm	28 nm	28nm
40Gbps Ports	32 (24)	24 (24)	42(42)
Buffer (MB)	12 MB	40 MB	23 MB
L2/ L3	L2/ L3	L2/ L3	L3

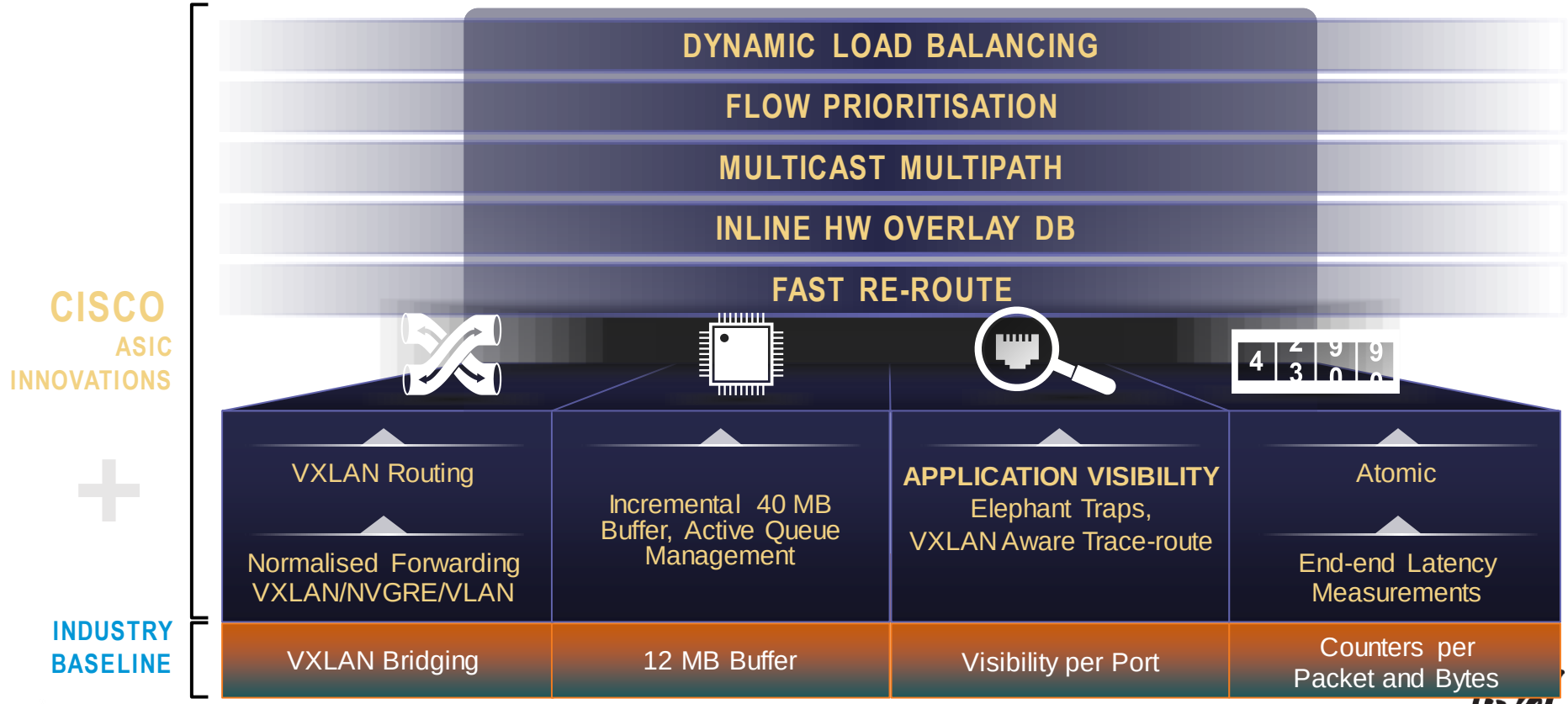
Merchant ASIC --- NFE (Broadcom T2)

Custom ASIC --- Cisco ALE (ACI Leaf Engine), ASE (ACI Spine Engine)

Merchant Silicon Alone Leaves Room for Improvement



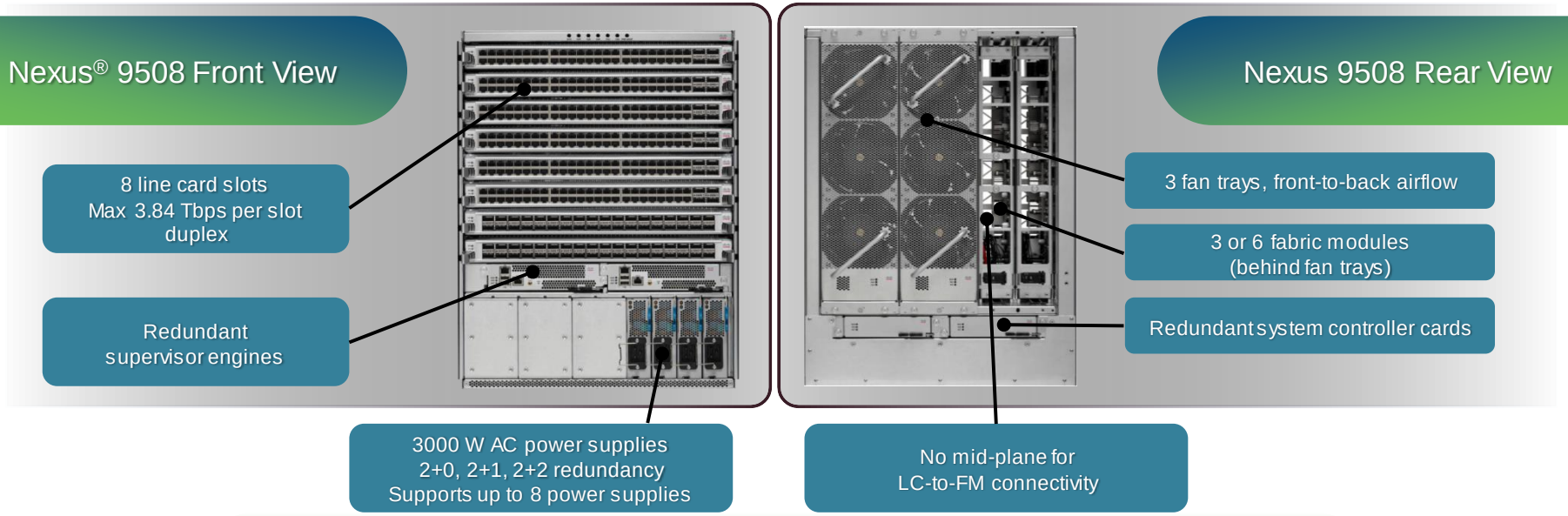
Merchant +



Agenda
Nexus 9000 Architecture

- Nexus 9000
  - Nexus 9000 Hardware
  - Nexus 9500 Chassis
  - Nexus 9500 Line Cards
  - Nexus 9500 Packet Forwarding
  - Nexus 9300
- Nexus 9000 and 40G
- Nexus 9000 Designs: FEX, vPC & VXLAN
- Nexus 9000 & Dev-Ops
- ACI & Nexus 9000

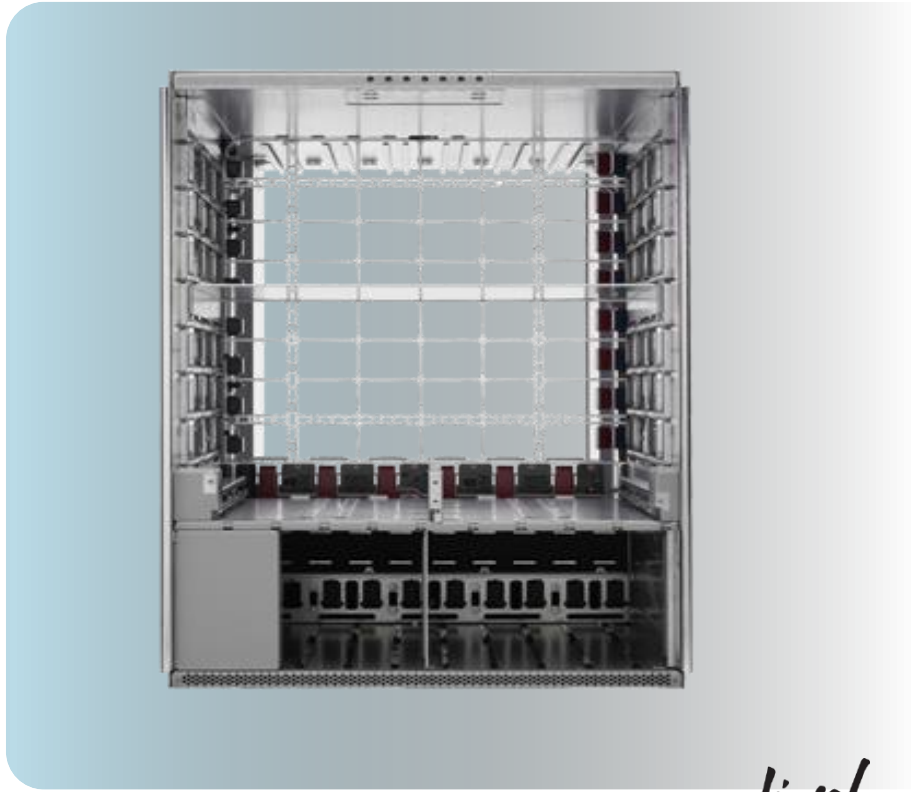
Nexus 9500 Platform Architecture



Nexus 9500 Platform Architecture

Chassis Design: No Mid-Plane

- Designed for:
  - Power & Cooling Efficiency
  - Designed for Reliability
  - Designed for Future Scale

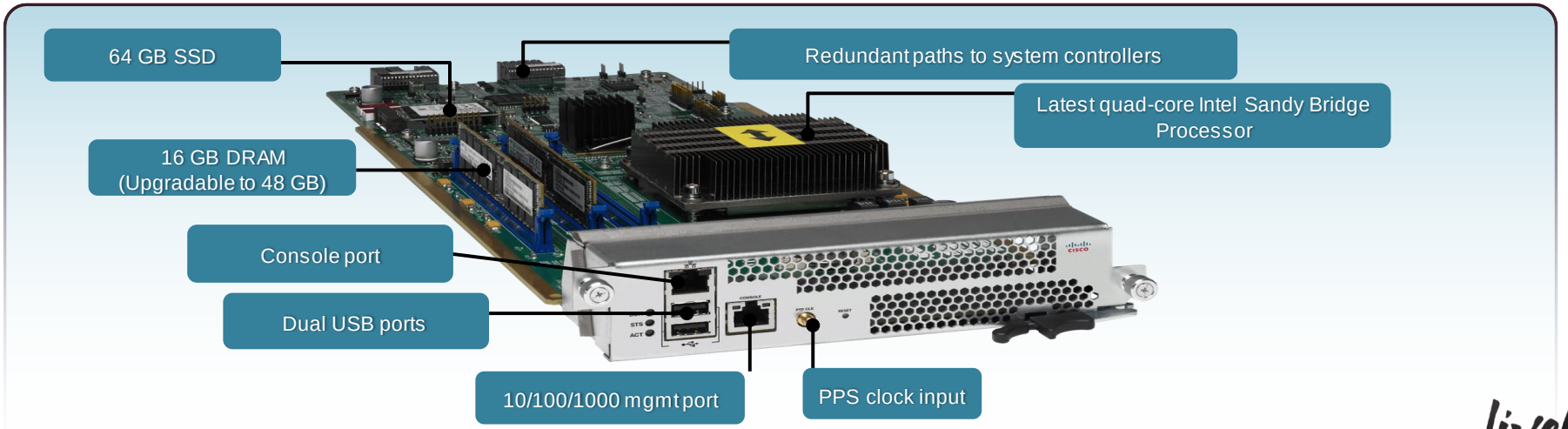


Nexus 9500 Platform Architecture

Supervisor Module

- Redundant half-width supervisor engine
- Performance- and scale-focused
- Range of management interfaces
- External clock input (PPS)

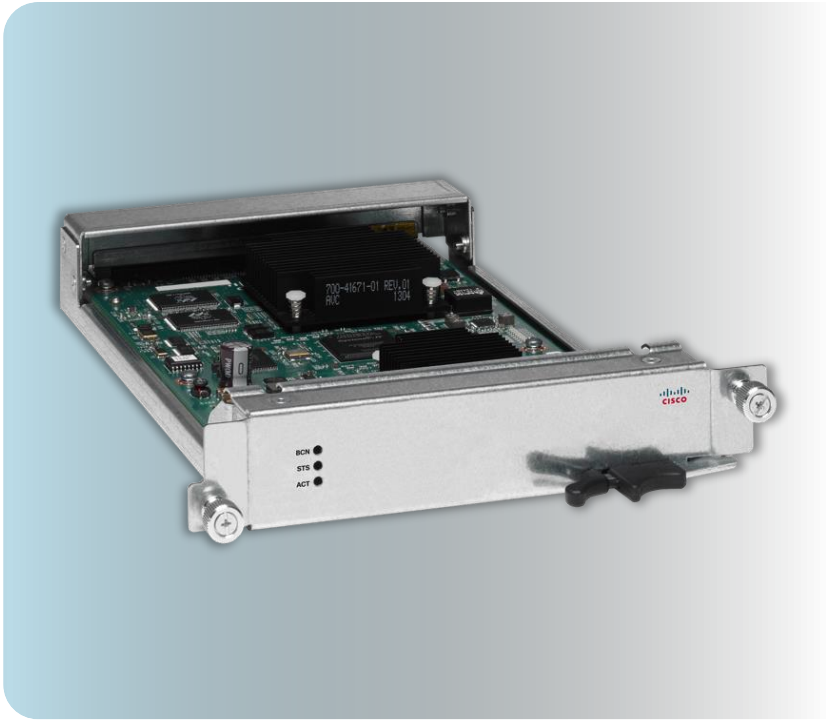
Supervisor Module	
Processor	Romley, 1.8 GHz, 4 core
System Memory	16 GB, upgradable to 48 GB
RS-232 Serial Ports	One (RJ-45)
10/100/1000 Management Ports	One (RJ-45)
USB 2.0 Interface	Two
SSD Storage	64 GB



Nexus 9500 Platform Architecture

System Controller Module

- Redundant half-width system controller
- Offloads supervisor from device management tasks
  - Increased system resiliency
  - Increased scale
- Performance- and scale-focused
  - Dual core ARM processor, 1.3 GHz
- Central point-of-chassis control
- Ethernet Out of Band Channel (EOBC) switch:
  - 1 Gbps switch for intra-node control plane communication (device management)
- Ethernet Protocol Channel (EPC) switch:
  - 1 Gbps switch for intra-node data plane communication (protocol packets)
- Power supplies through system management bus (SMB)
- Fan trays



Nexus 9500 Platform Architecture

Power Supplies

3000W AC PSU

- Single 20A input – 220V
- Support for range of international cabling options
- 92%+ Efficiency
- Range of PS configurations
  - Minimum 1 PS
  - 2 PS for fully loaded chassis
  - N+1 redundancy
  - N+N grid redundancy
- Up to 2x head room for future port densities, bandwidth, and optics

** 80 Plus Platinum is equivalent to Climate Saver/ Green Grid Platinum rating*

Nexus 9500 Platform Architecture

Power Efficiency by Design

- 1st modular chassis w/o a mid-plane --- Unobstructed front-back airflow
- Platinum rated PS --- 90%-94% power efficiency across all work loads
- Highly integrated switch and buffer functionality --- Only 2 to 4 ASICs per line card

Traffic type	Power (watts)	Fan Speed
No traffic	3233	10%
100% line-rate with IMIX packets	4746	20%
100% line-rate with 64 byte packets	5470	25%

Test Results on a fully loaded Nexus 9508 switch with 288 40GE ports:

Nexus 9500 Platform Architecture

Fan Trays and Fabric Modules

- Up to 6 Fabric Modules
  - Different cost points for 1/10G access and 40G aggregation
  - Flexibility for future generation of fabric modules
  - Quad Core ARM CPU 1.3 GHz for Supervisor offload
- Smooth degradation during replacement

- 3 Fan Trays
  - 3 dual fans per tray
  - Dynamic speed control driven by temperature sensors
  - Straight Airflow across LC and FM
- N+1 Redundancy per Tray

Nexus 9500 Platform Architecture

Fan Tray

Fan trays are installed after the Fabric Module.

To service a FM, the fan tray must be removed first.

1) If one fan tray is removed, the other two fan trays will speed up 100% to compensate for the loss of cooling power

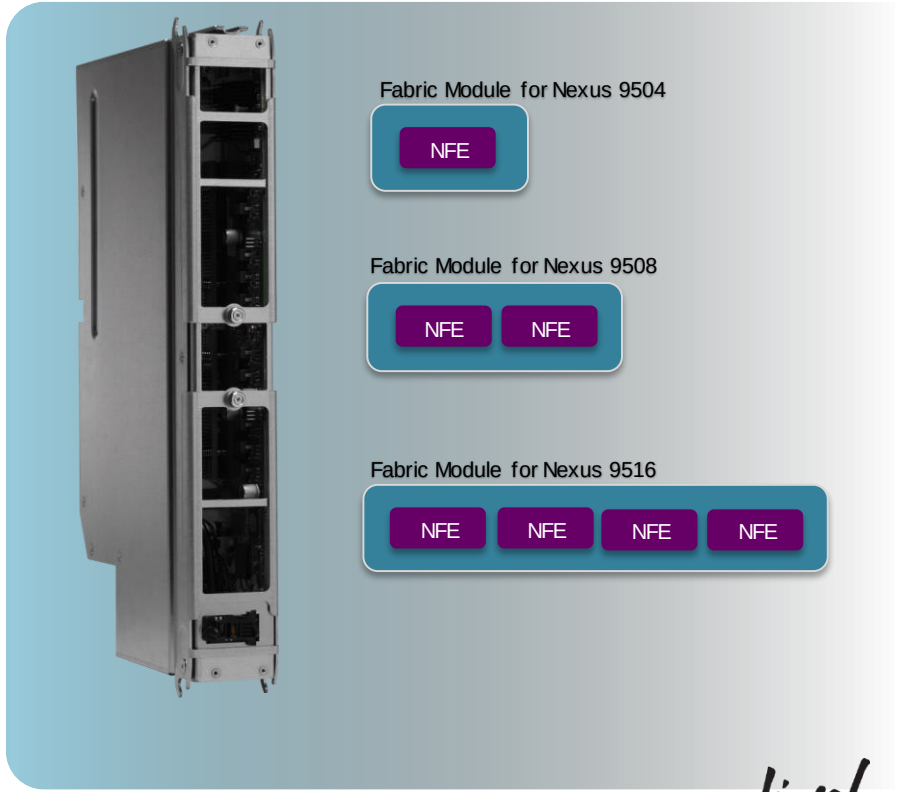
1) Temperature Sensors in the chassis will shut down components once max temp is reached.

Nexus 9500 Platform Architecture

Fabric Modules

- Interconnect I/O module slots
- Installed at the rear of the chassis
- Uses Broadcom T2 as the network forwarding engine (NFE)
- Up to 3.84 Tbps duplex per line card slot
- All fabric cards are active and carry traffic

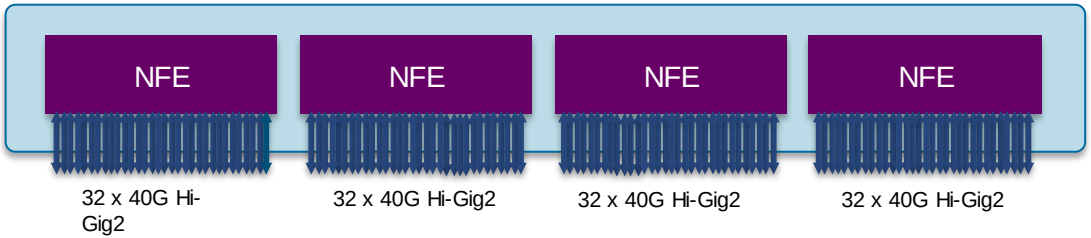
Chassis Type	Nexus 9504	Nexus 9508	Nexus 9516
NFEs per Fabric Module	1	2	4



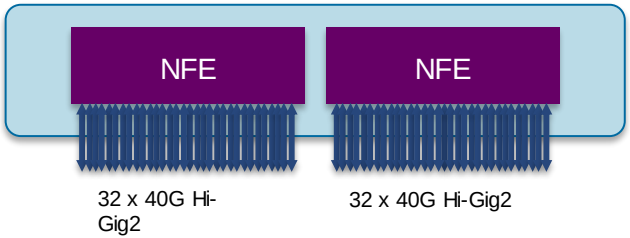
Nexus 9500 Platform Architecture

Fabric Modules

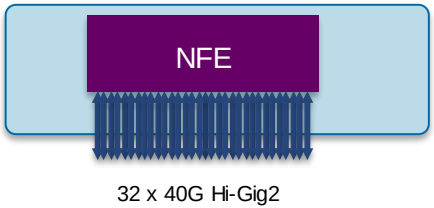
Fabric Module for Nexus 9516 Switches



Fabric Module for Nexus 9508 Switches



Fabric Module for Nexus 9504 Switches

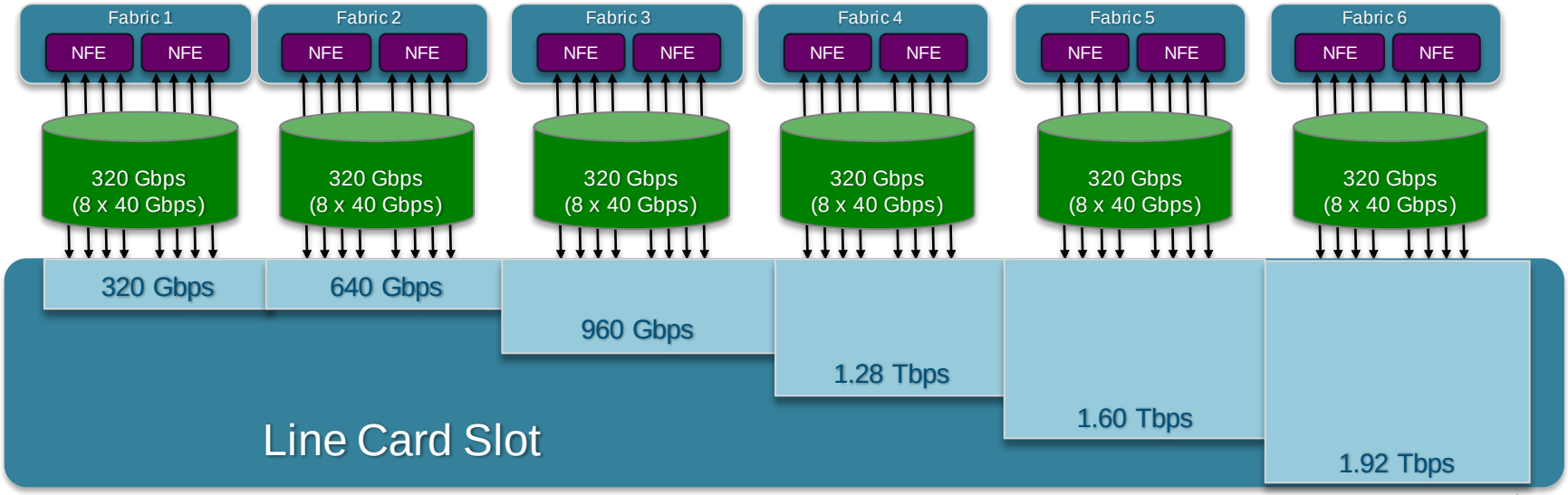


* The 40G Hi-Gig2 links can be clocked at 42G rate

Nexus 9500 Platform Architecture

Data Plane Scaling with Fabric Modules (Nexus 9500 as an Example)

- Nexus 9508 Fabric Module can provide up to 320 Gbps to each I/O module slot
- With 6 fabric modules, each I/O module slot can have up to 1.92 Tbps forwarding bandwidth in each direction



Agenda
Nexus 9000 Architecture

- Nexus 9000
  - Nexus 9000 Hardware
  - Nexus 9500 Chassis
  - Nexus 9500 Line Cards
  - Nexus 9500 Packet Forwarding
  - Nexus 9300
- Nexus 9000 and 40G
- Nexus 9000 Designs: FEX, vPC & VXLAN
- Nexus 9000 & Dev-Ops
- ACI & Nexus 9000

Nexus 9500 Switch Line Cards

X9600 Series Line Cards
High Performance 40G Aggregation

- N9K-X9636PQ
- 12x 100GE Port Line Card

Merchant only

NX-OS Mode Only

X9500 Series Line Cards
Performance 10G/40G Access/Aggregation

- N9K-X9564PX
- N9K-X9564TX
- N9K-X9536PQ

Merchant +

NX-OS Mode & ACI Leaf Ready

X9700 Series Line Cards
High Performance 40G ACI Spine

- N9K-X9736PQ

Merchant +

ACI Spine Only

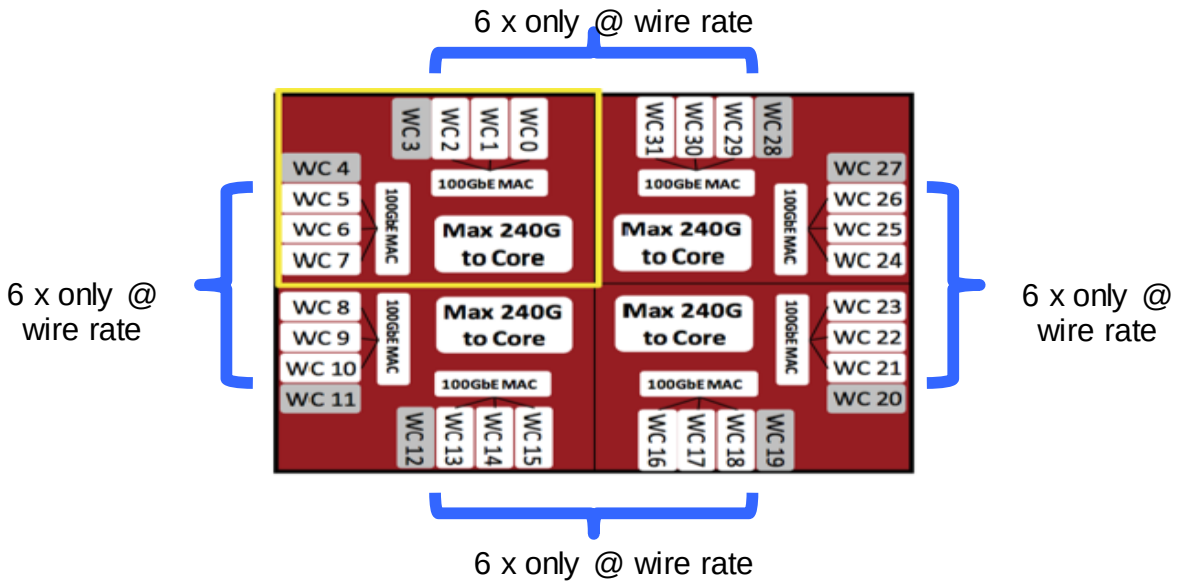
Nexus 9500 N9K-X9600 Series Line Cards

N9K-X9636PQ

N9K-X9636PQ line card needs 6 fabric modules to operate at full line rate on all 36 ports and for all packet sizes.

- 36x 40 Gbps QSFP ports
- 2.88 Tbps full-duplex fabric connectivity
- Layer 2 and 3 line-rate performance on all ports for all packet sizes
- Supports 4x 10 Gbps break-out mode
- Works in 4-slot and 8-slot Nexus 9500 chassis

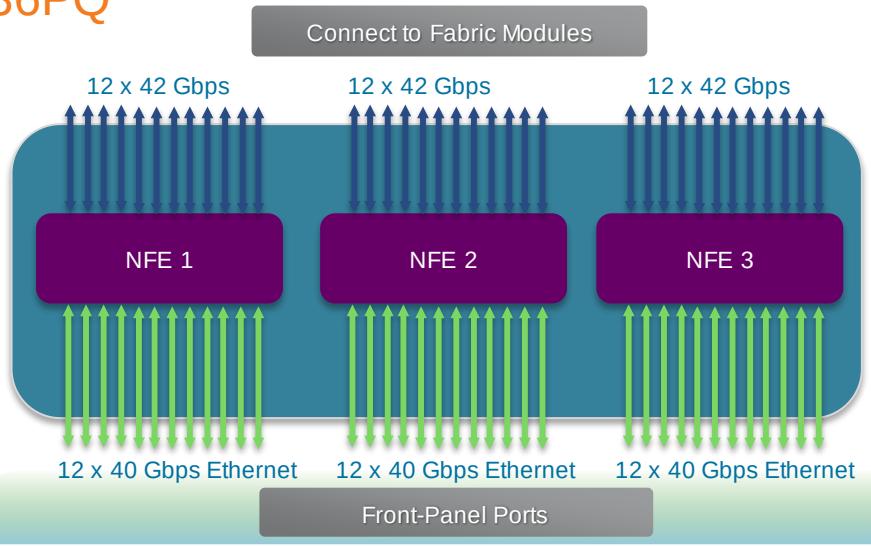
Broadcom Trident 2 Internals



- 24 of 32 ports are full line rate for all packet sizes
- Only 24 full-line-rate ports per T2 are used on N9K-X9636PQ line card

Nexus 9500 N9K-X9600 Series Line Cards

N9K-X9636PQ

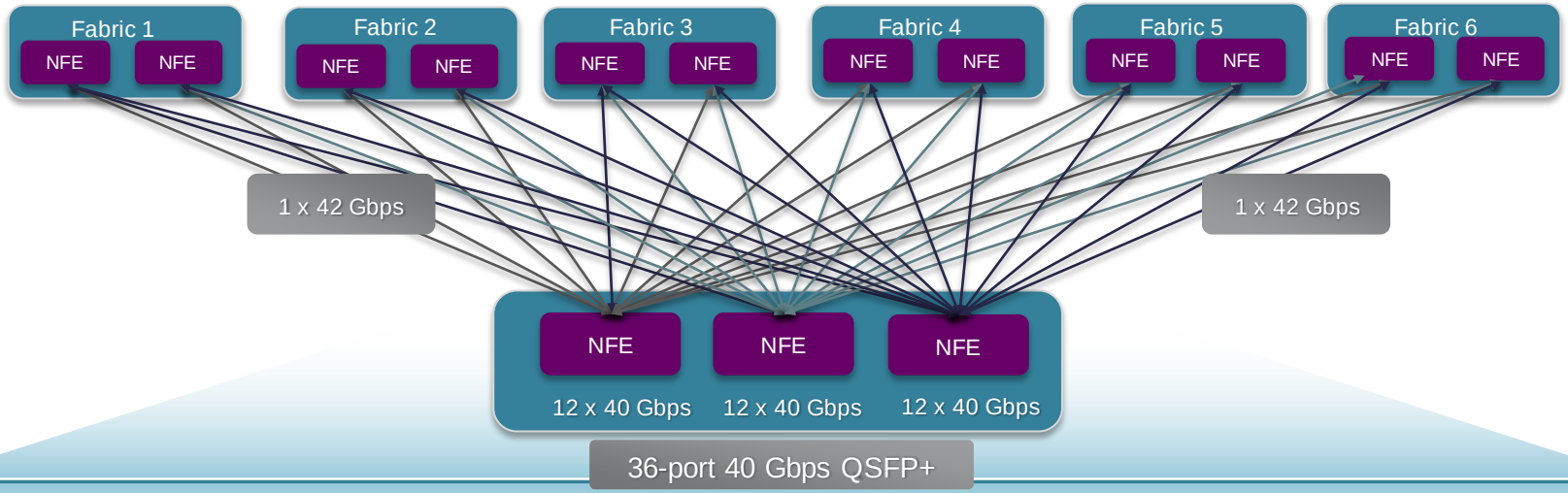


N9K-X9636PQ line card needs 6 fabric modules to operate at line rate on all 36 ports.

- 3 network forwarding engines (NFE)
- Each NFE runs in full-line-rate mode, providing 12 x 40 Gbps links to the front panel and 12 x 40 Gbps internal links to the fabric modules

Nexus 9500 N9K-X9600 Series Line Cards

N9K-X9636PQ Fabric Connectivity



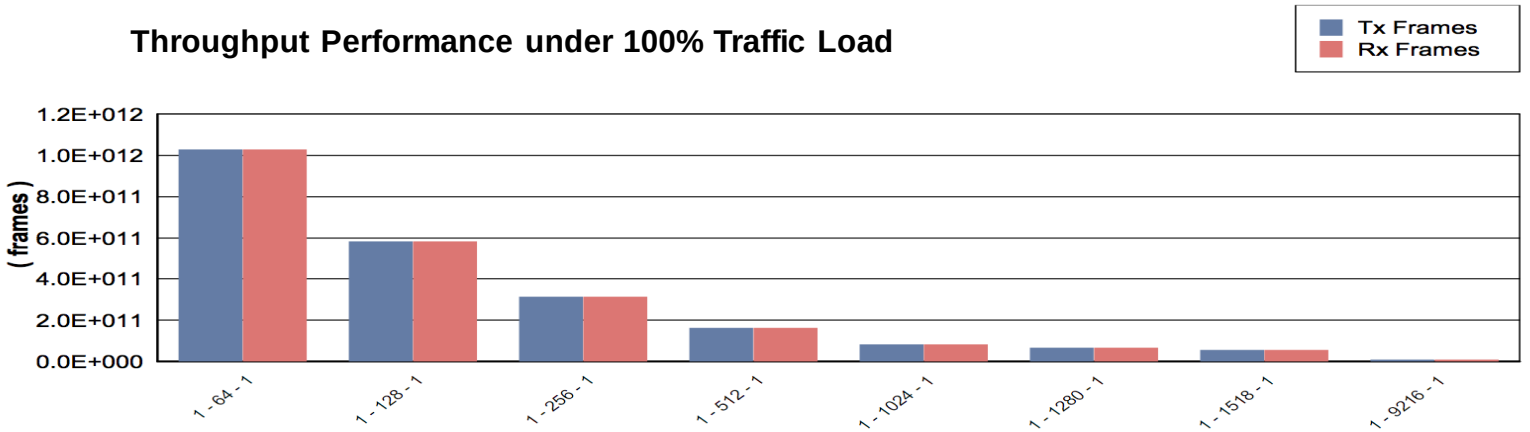
- All ports on the line card can operate at line rate for any packet sizes with 6 fabric modules
- Each NFE has 12 x 40 Gbps internal links to fabric modules - one to each Fabric NFE
- The Internal 40 Gbps links are running at 42 Gbps clock rate to compensate the internal overhead

Nexus 9500 N9K-X9600 Series Line Cards

Full Line Rate on All 40GE Ports

Unprecedented Full Line Rate Performance:

- Proved with RFC 2544/ RFC 2889/ RFC 3918 Throughput Test Results on a fully loaded Nexus 9508 switch with 288 40GE ports:
  - All ports are line rate at 100% unicast traffic load
  - All ports are line rate at 100% multicast traffic load
  - Full line rate for all packet sizes (64~9216 Bytes)

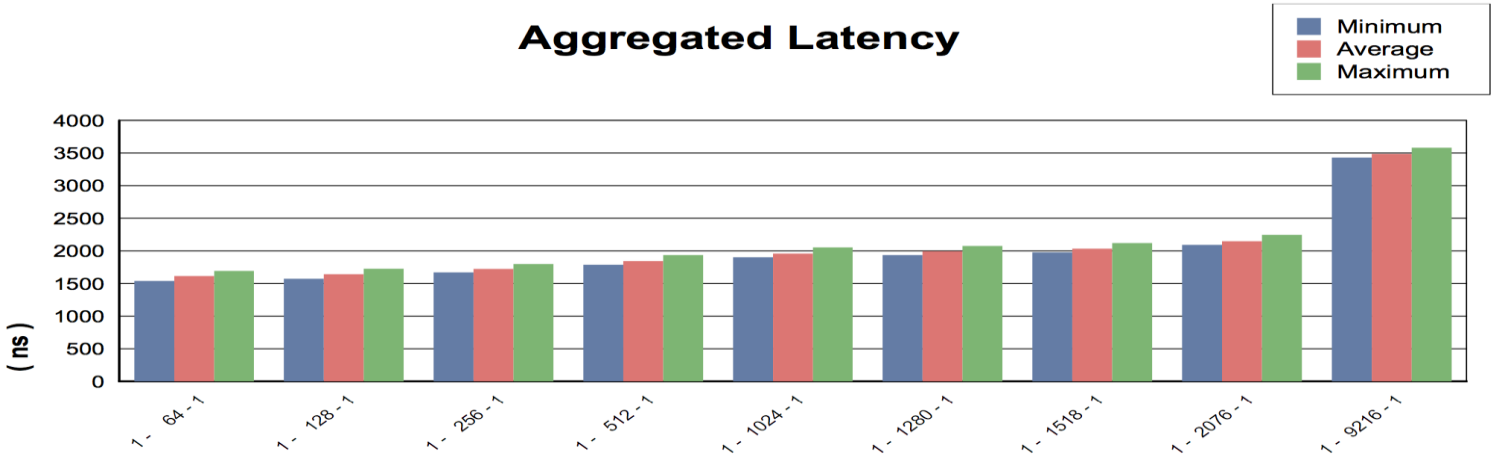


Nexus 9500 N9K-X9600 Series Line Cards

Low Latency Performance on N9K-X9636PQ

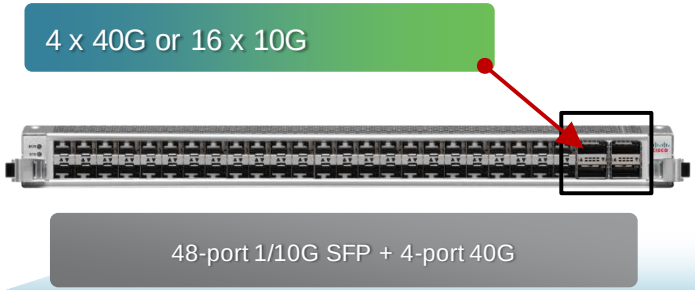
Low Latency (same for both Unicast and Multicast):

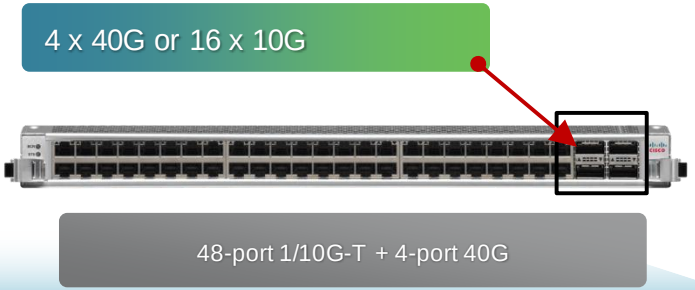
- Proved with RFC 2544/ RFC 2889/ RFC 3918 Throughput Test Results on a fully loaded Nexus 9508 switch with 288 40GE ports.
- Consistent unicast and multicast Latency at 100% traffic load:
  - 1.6 usec (64-Byte packets)
  - 3.5 usec (9216-Byte packets)



Nexus 9500 N9K-X9500 Series Line Cards

N9K-X9564PX & N9K-X9564TX





N9K-X9564PX --- 48 1/10G SFP+ ports + 4 40G QSFP+ ports

N9K-X9564TX --- 48 1/10GBase-T ports + 4 40G QSFP+ ports

1.92 Tbps duplex fabric connectivity

Layer 2 and 3 line-rate performance on all ports for all packet sizes

Cisco® NX-OS and Application Centric Infrastructure (ACI) mode

Works in 4, 8 and 16-slot Nexus 9500 chassis

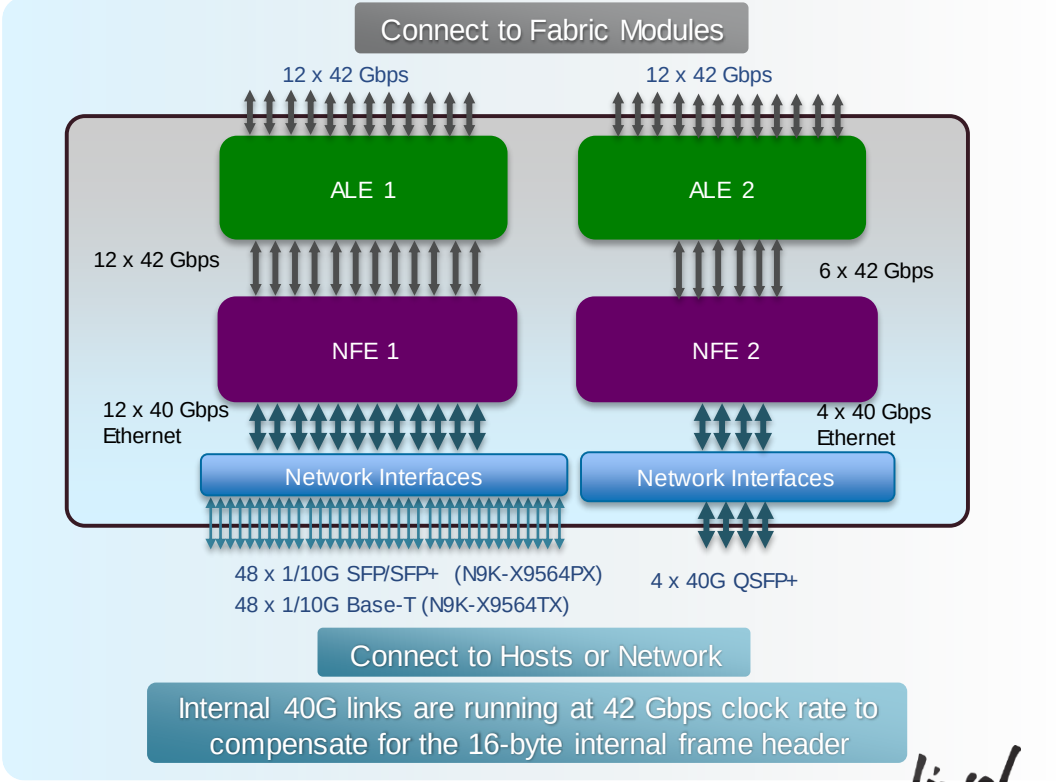
Nexus 9500 N9K-X9500 Series Line Cards

N9K-X9564PX & N9K-X9564TX (Cont'ed)

2 network forwarding engines (NFEs)

2 application leaf engines (ALEs) for additional buffering and packet handling

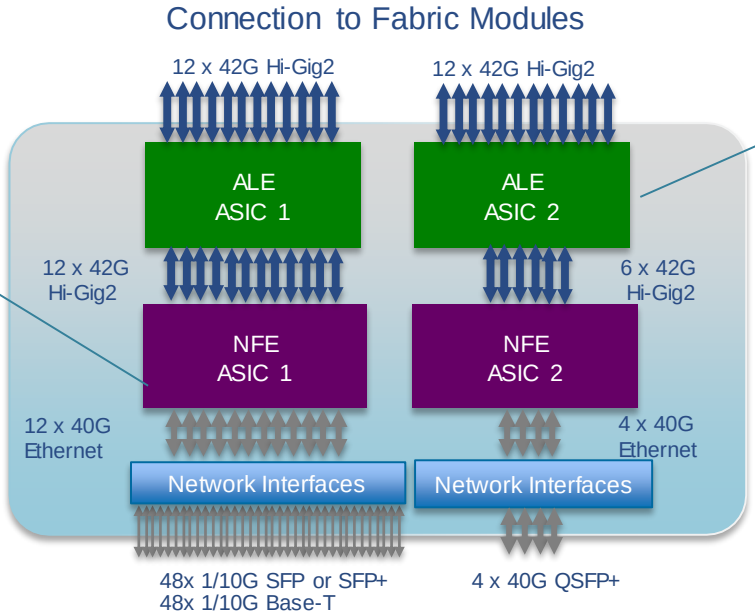
Works in 4, 8 and 16 slot chassis
Line rate performance on all ports and all packet sizes with 3 or 6 fabric modules



Nexus 9500 N9K-X9500 Series Line Cards

N9K-X9564PX & N9K-X9564TX (Cont'ed)

NFE ASICs act as main forwarding engines for standalone mode.



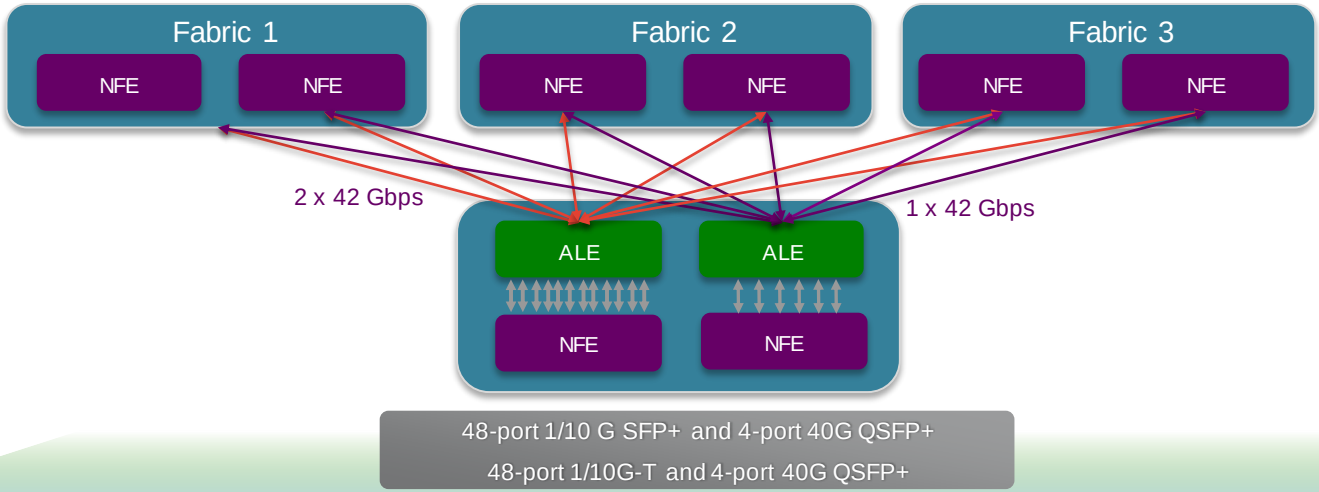
ALE ASICs perform additional packet processing and buffering for standalone mode.

NX-OS & Fabric Access LCs

Nexus 9500 N9K-X9500 Series Line Cards

N9K-X9564PX & N9K-X9564TX Fabric Connectivity

Fabric Connectivity with 3 Fabric Modules in Nexus 9508 Chassis

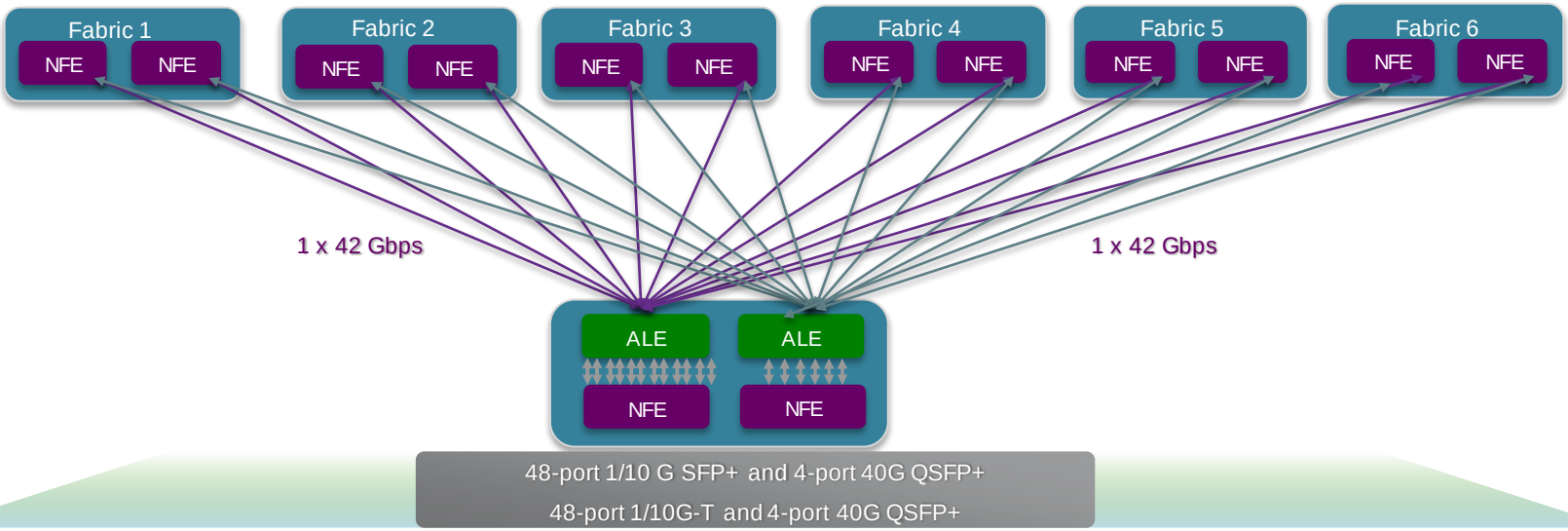


- Minimum of 3 fabric modules to get all line-rate ports
- When a chassis has only 3 fabric modules, they need to be positioned in fabric module slot 2, 4 and 6.

Nexus 9500 N9K-X9500 Series Line Cards

N9K-X9564PX & N9K-X9564TX Fabric Connectivity

Fabric Connectivity with 6 Fabric Modules in Nexus 9508 Chassis



When a chassis has 6 fabric modules, N9K-X9564PX and N9K-X9564TX are connected to and utilise all 6 fabric modules

The above diagram shows the line card fabric connectivity in a Nexus 9508 chassis as an example

Nexus 9500 N9K-X9500 Series Line Cards

N9K-X9536PQ

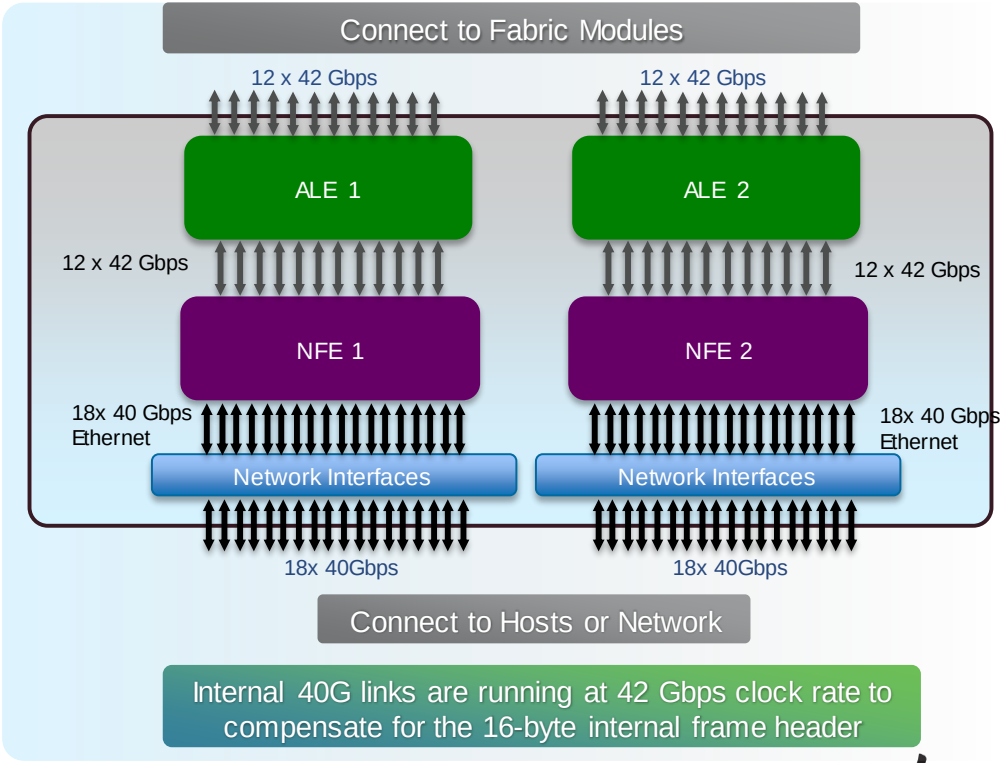
N9K-X9536PQ is supported in all Nexus 9500 Chassis

- 36x 40 Gbps QSFP ports, 1.5:1 oversubscribed
- 1.92 Tbps duplex fabric connectivity
- Full Layer-2 & Layer-3 functions
- Supports 4x 10 Gbps break-out mode
- Support Cisco® NX-OS mode and ACI ready
- Sported in all Nexus 9500 chassis types, including 4, 8 and 16-slot

Nexus 9500 N9K-X9500 Series Line Cards

N9K-X9536PQ Series Line Cards

- 2 network forwarding engines (NFEs)
- 2 application leaf engines (ALEs) for additional buffering and packet handling
- Need 3 fabric modules, can work with 6

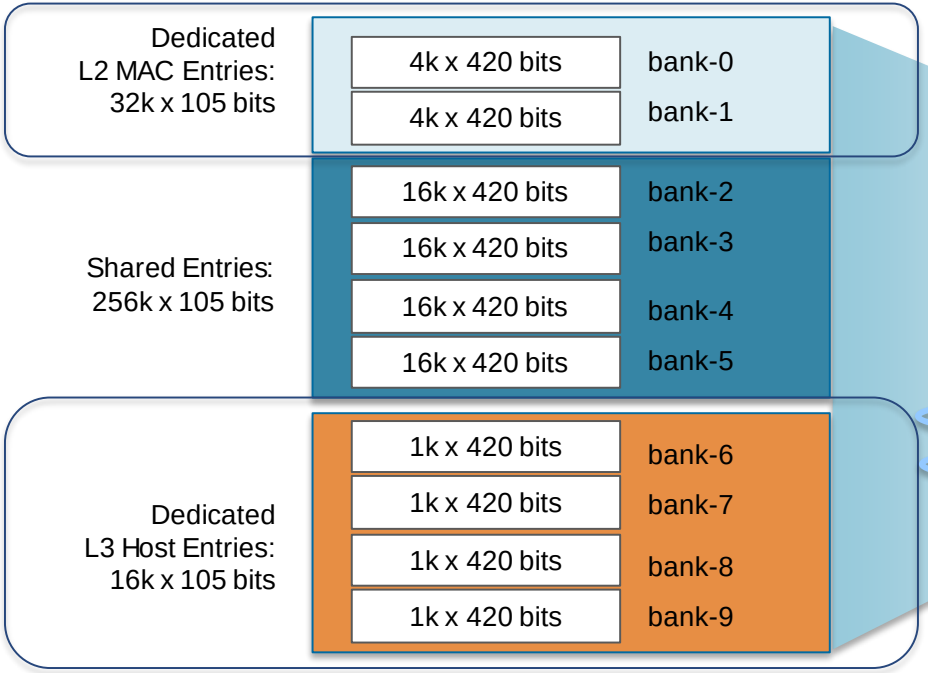


Agenda
Nexus 9000 Architecture

- Nexus 9000
  - Nexus 9000 Hardware
  - Nexus 9500 Chassis
  - Nexus 9500 Line Cards
  - Nexus 9500 Packet Forwarding
  - Nexus 9300
- Nexus 9000 and 40G
- Nexus 9000 Designs: FEX, vPC & VXLAN
- Nexus 9000 & Dev-Ops
- ACI & Nexus 9000

NFE Unified Forwarding Table

- NFE has a 16K traditional LPM TCAM table.
- Additionally NFE has the following Unified Forwarding Table for ALPM (Algorithm LPM) Mode

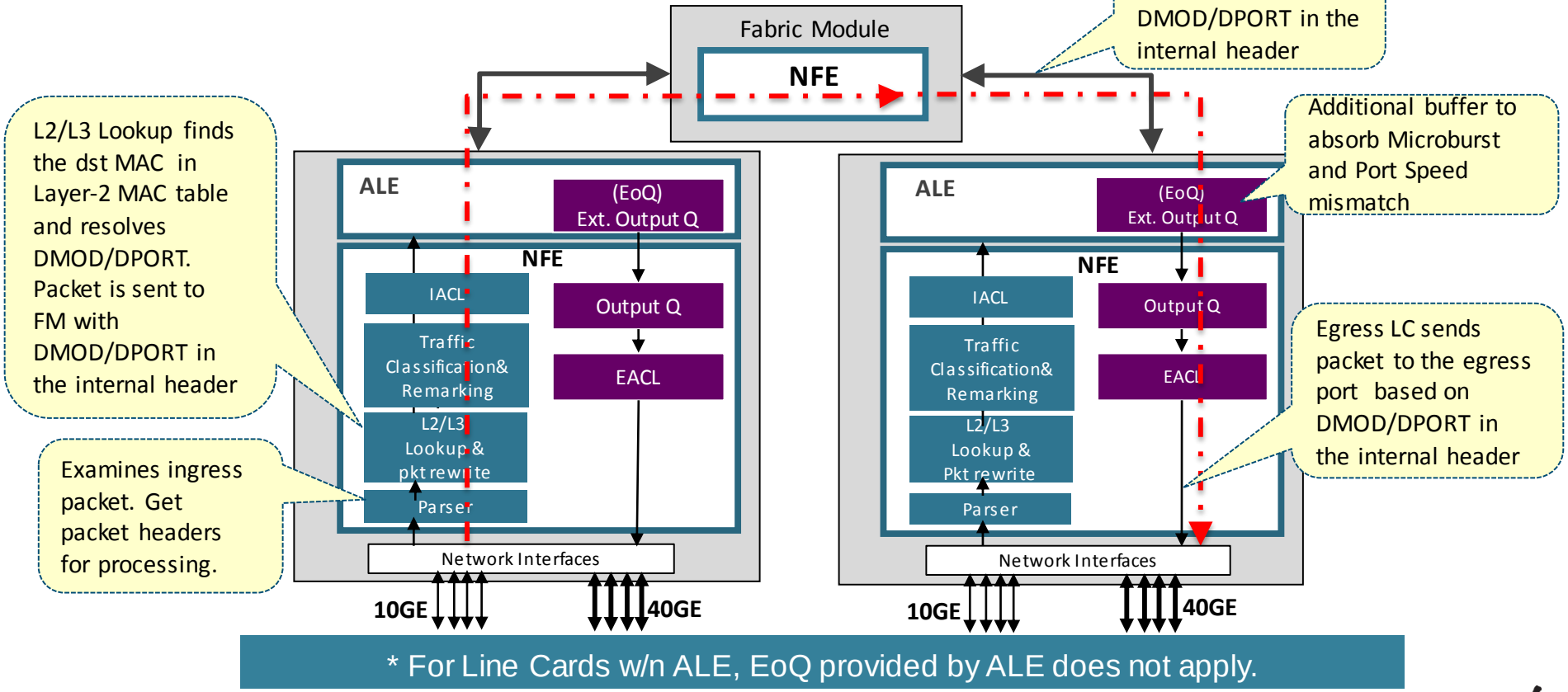


SUPPORTED COMBINATIONS

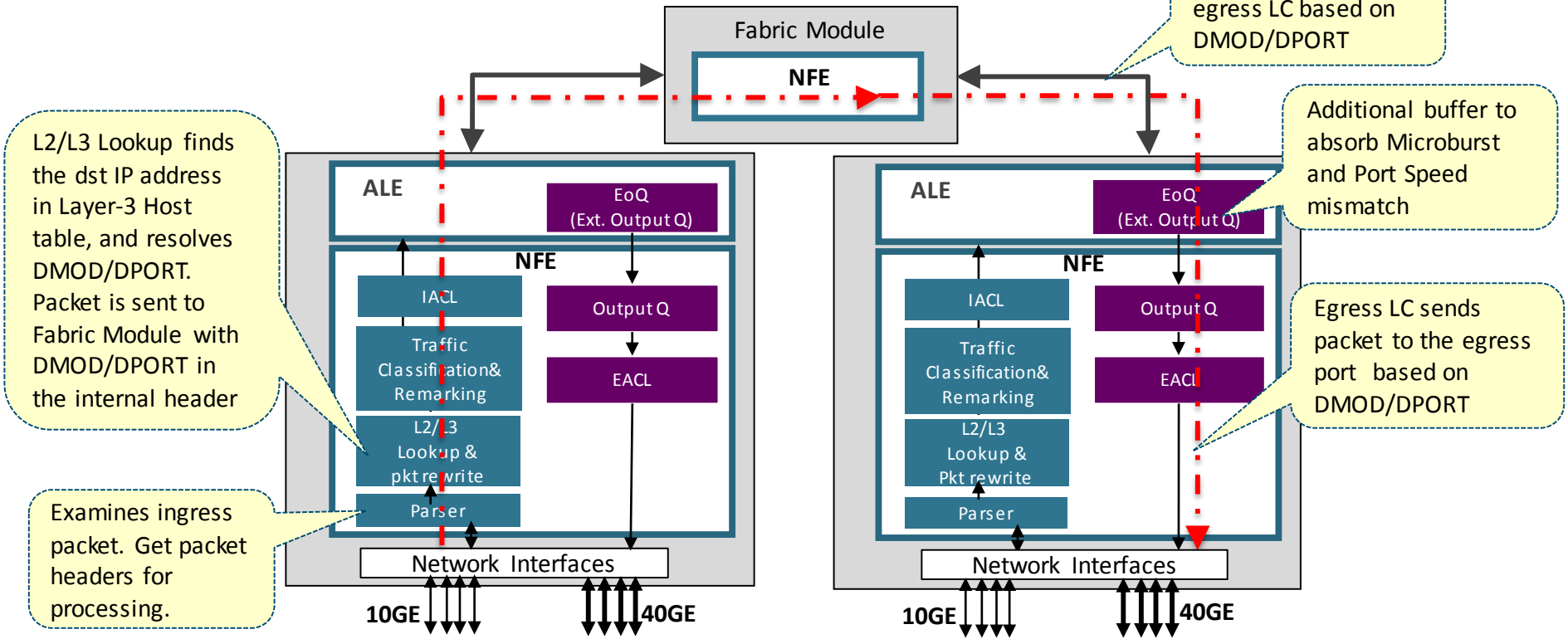
Mode	L2	L3 Hosts	LPM
0	288K	16K	0
1	224K	56K	0
2	160K	88K	0
3	96K	120K	0
4	32K	16K	128K

In default setting, N9500 line card NFE uses Mode 3, fabric module NFE uses Mode 4.

Layer-2 Unicast Packet Walk

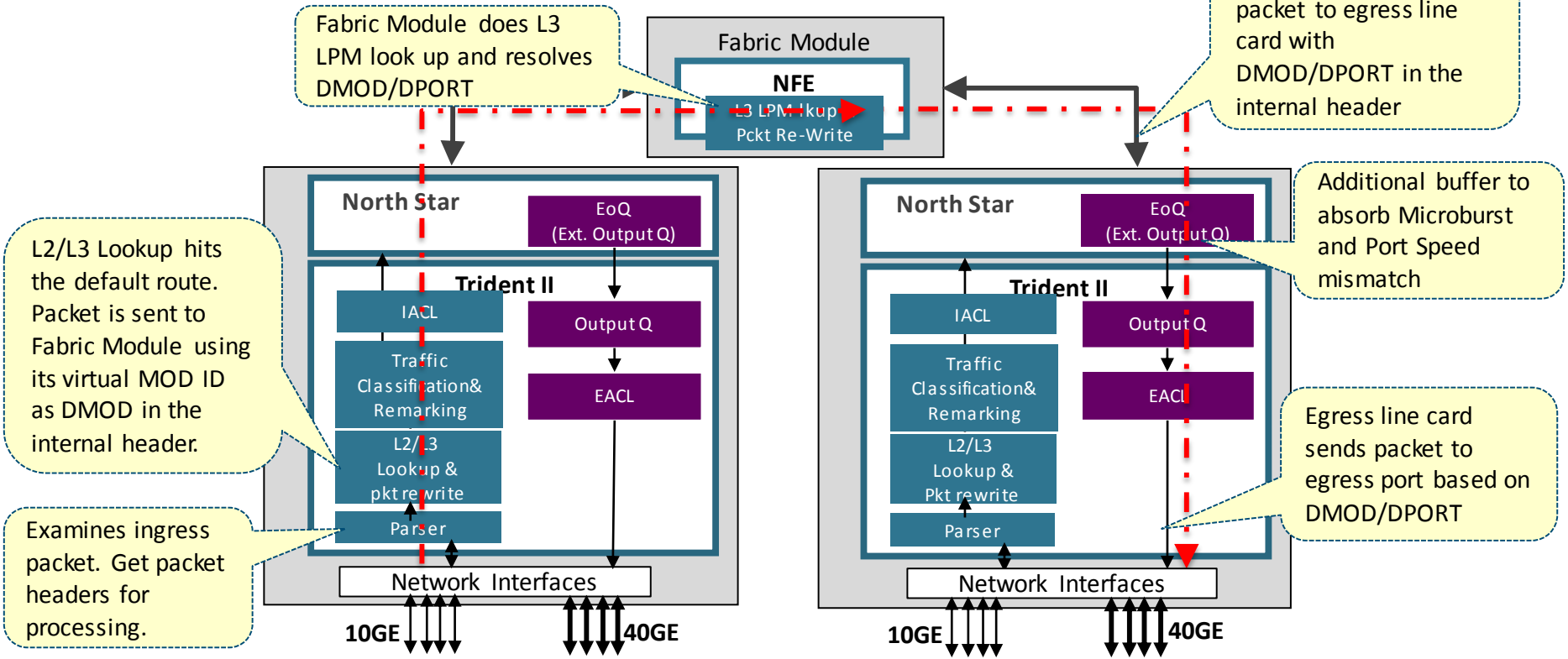


Layer-3 Host Unicast Packet Walk



* * For Line Cards w/n ALE, EoQ provided by ALE does not apply.

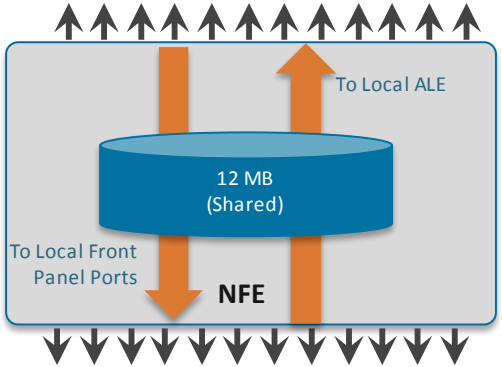
Layer-3 LPM Unicast Packet Walk



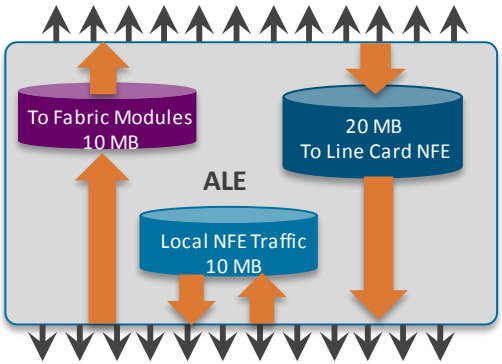
* For Line Cards w/n ALE, EoQ provided by NorthStar does not apply.

Nexus 9500 N9K-X9500 Series Line Cards

Buffer Allocation on N9K-X9500 Series Line Cards



- Shared 12MB buffer

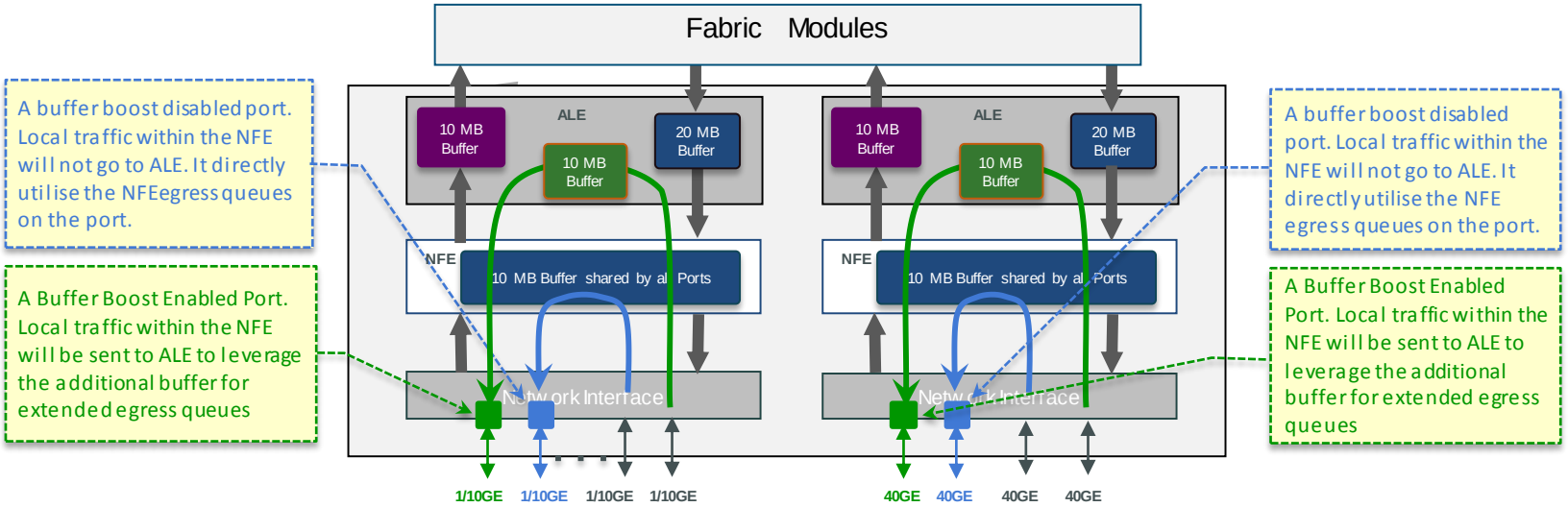


- Shared 10MB ingress buffer for fabric bound traffic;
- Shared 20MB egress buffer for traffic coming from fabric and going to front panel port.
- Shared 10MB egress buffer for local NFE traffic

Module Type	Buffer Size
N9K-X9564TX (48-Port 1/10G T + 4-Port 40G)	104 MB
N9K-X9564PX (48-Port 1/10G F + 4-Port 40G)	104 MB
N9K-X9536PX (36-Port 40G)	104 MB

Nexus 9500 N9K-X9500 Series Line Cards

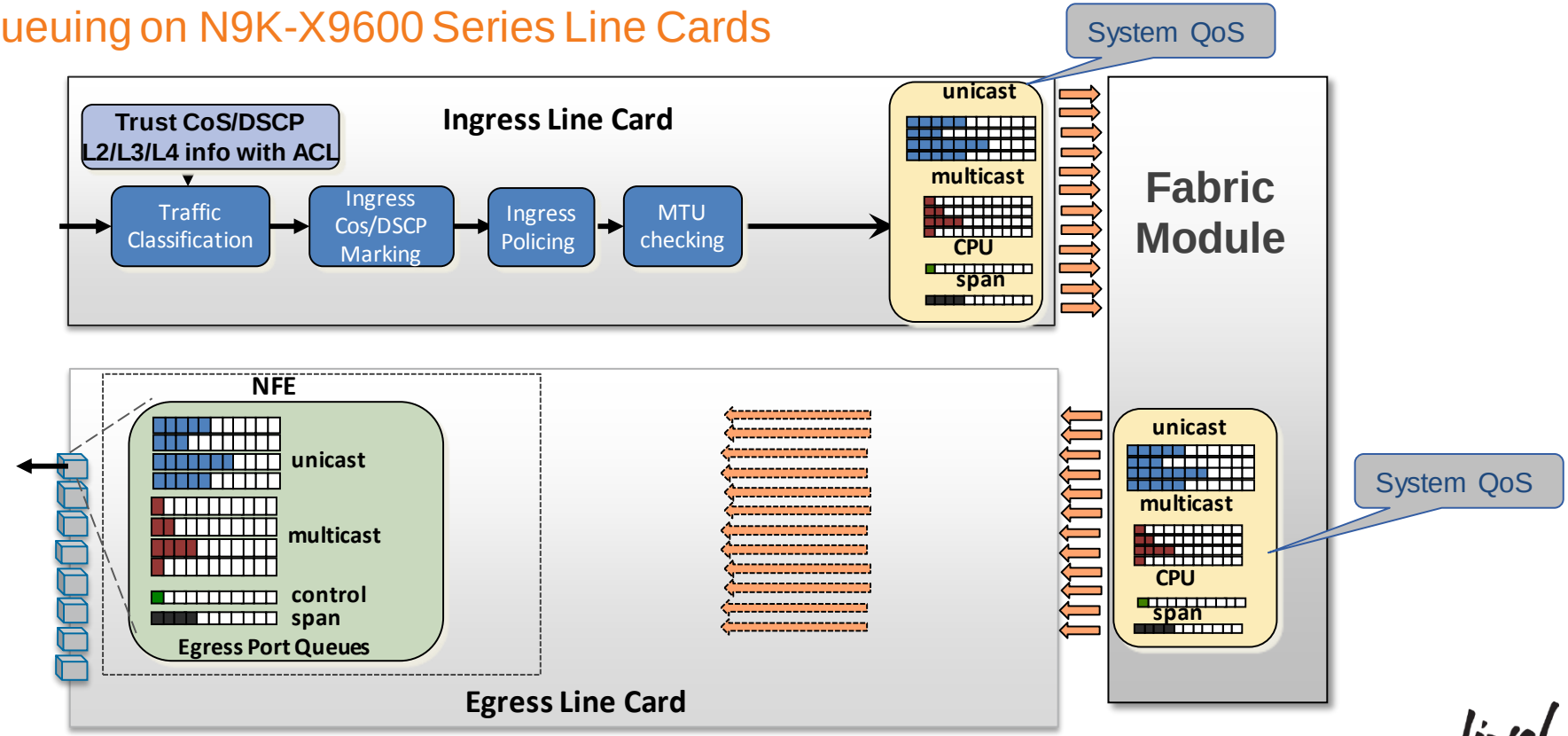
Buffer Boost Function on N9K-X9500 Series Line Cards



- Buffer boost function is introduced to allow line cards with ALEs to leverage the additional buffer on ALE for NFE local traffic.
- When buffer boost is enabled on a port, NFE local traffic to this port is sent to ALE for additional buffer space.
- When buffer boots is disabled on a port, NFE local traffic to this port remains local on this NFE.
- Buffer boots is enabled on ports of line cards with ALEs by default. It is a port property and can be disabled on a per-port basis.

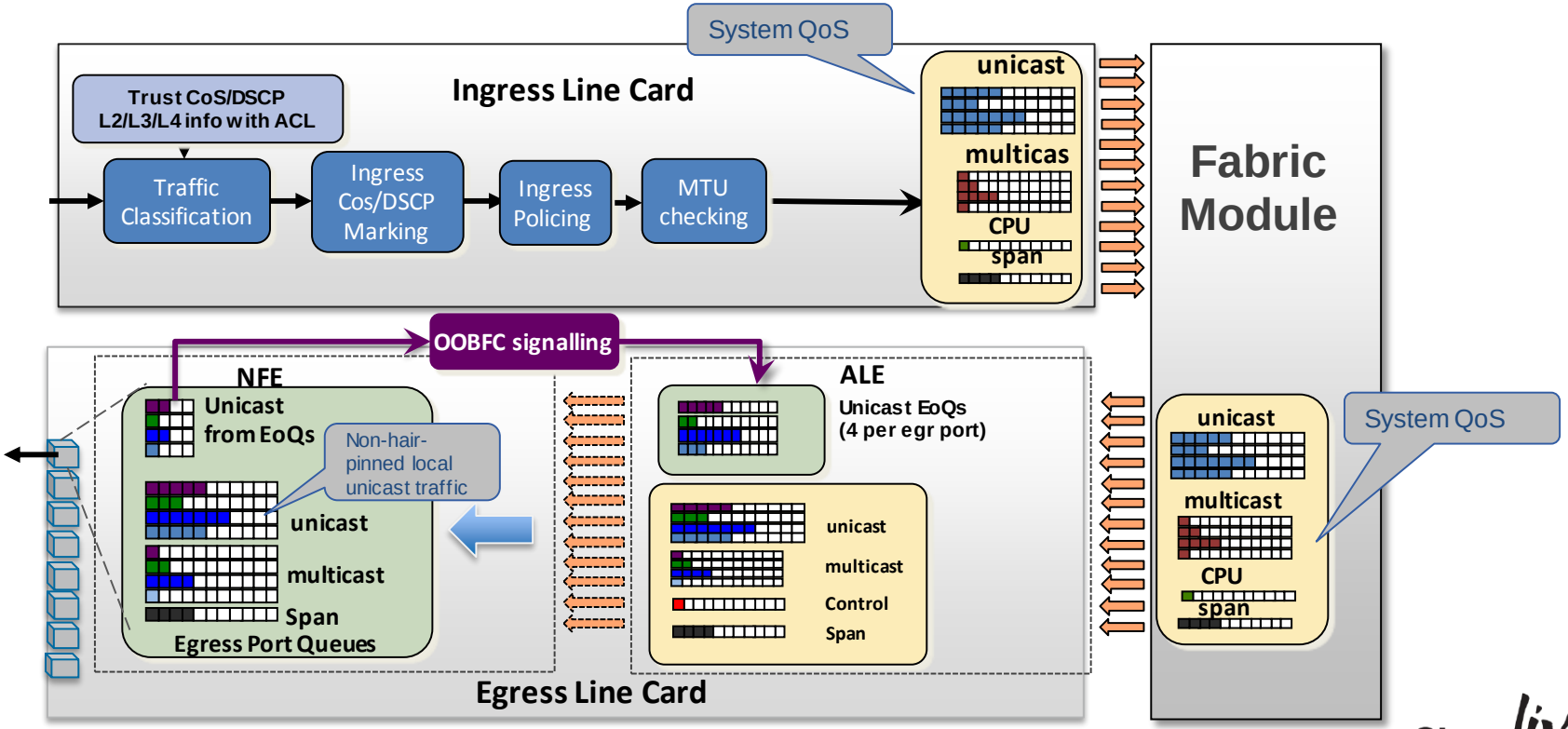
Queuing on Nexus 9500 Switch Line Cards

Queuing on N9K-X9600 Series Line Cards



Queuing & Scheduling on Nexus 9500 Line Cards

Queuing on N9K-9500 Series Line Cards



Nexus 9500 Switch QoS

- **Ingress QoS Classification**
  - Policy-map type qos)
  - Match on CoS/ IP Precedence/ DSCP /ACL
  - Set qos-group
  - Remark CoS/ IP Precedence/ DSCP
  - Ingress policing

- **Network-QoS**
  - Policy-map type network-qos
  - Match on qos-group
  - Enable PFC

- **Egress Queuing and Shaping**
  - Policy-map type queueing
  - Enable WRED or ECN (default is tail drop)
  - 4 user-defined classes based on qos-group
  - 1 control class for CPU and 1 class for SPAN traffic
  - 3 PFC non-drop queues
  - 3 Priority queues

Nexus 9500 Switch Forwarding Tables

Default Forwarding Mode in NX-OS 6.1.2(I1)1

	Nexus® 9500
	Default
LPM Routes	128K (on Fabric Module)
IP Host Routes	88K (on Line Card)
MAC Address Entries	160K (On Line Card)
Multicast Routes	32K* (Hardware capable of 72K)

* Shared with IP hosts

This table shows system maximum scalability. For validated scale, refer to the Nexus 9000 verified scalability white papers

Nexus 9500 Switch Forwarding Tables

Default Forwarding Mode in NX-OS 6.1.2(I2)1

	Nexus® 9500
	Default
LPM Routes	128K (on Fabric Module)
IP Host Routes	120K (on Line Card)
MAC Address Entries	96K (On Line Card)
Multicast Routes	32K* (Hardware capable of 72K)

* Shared with IP hosts

This table shows system maximum scalability. For validated scale, refer to the Nexus 9000 verified scalability white papers

Nexus 9500 Switch Forwarding Tables

Non-Hierarchical Mode (NX-OS 6.1.2(I2)1)

Non-hierarchical mode moves LPM table from fabric modules to line cards so that line card NFE can locally forward LPM routed traffic, avoiding the latency of sending traffic to fabric module for LPM lookup.

Forwarding Tables in Non-Hierarchical Mode

- No forwarding on fabric modules
- Line cards are programmed symmetrically with IPv4 LPM + Host + Mac tables
- Two options for line card setting in this mode:
  - MAX LPM --- Line cards are programmed for maximum LPM scale
  - MAX Host --- Line cards are programmed for maximum IP Host Scale

	Nexus® 9500	
	Non-Hierarchical Mode	
	Max LPM Mode	Max IP Host Mode
LPM Routes	128K	16K
IP Host Entries	16K	120K
MAC Address Entries	32K	96K
Multicast Routes	8K*	32K*

* Shared with IP hosts

This table shows system maximum scalability. For validated scale, refer to the Nexus 9000 verified scalability white papers

Nexus 9500 Switch Forwarding Tables

Max-Host Mode

Max-Host mode is to support maximum IP host scalability, especially in IPv4 and IPv6 dual-stack mode.

Forwarding Tables in Max Host Mode:

- IPv4 Host Table and IPv4 LPM Table are programmed on fabric modules
- IPv6 Host Table, IPv6 LPM Table and MAC Table are programmed on line cards
- IP multicast routes share the same table as IP Hosts. It's programmed on both fabric modules and line cards.

	Nexus® 9500	
	Max-Host Mode	
	IPv4	IPv6
LPM Routes	16K	4K
IP Host Entries	120K	40K (can scale up in future release)
MAC Address Entries	---	160K
Multicast Routes	32K* (programmed on both FM and LC)	

* Shared with IP hosts

This table shows system maximum scalability. For validated scale, refer to the Nexus 9000 verified scalability white papers

Nexus 9500 System Scalabilities

	Nexus® 9500		
	Default	Non-Hierarchical Mode	MAX-Host Mode
LPM Routes	128K	16K	16K
IP Host Entries	120K	120K or 88K	120K IPv4 + 40K IPv6
MAC Address Entries	96K	96K or 160K	160K
Multicast Routes	32K* (hardware capable of 72K)	32K* (hardware capable of 72K)	32K* (hardware capable of 72K)
IGMP Snooping Groups	32K* (hardware capable of 72K)	32K* (hardware capable of 72K)	32K* (hardware capable of 72K)
VRF	1000	1000	1000
Maximum Links in Port Channel	32	32	32
Maximum ECMP Paths	64	64	64
Maximum vPC Port Channels	528	528	528
Active SPAN/ERSPAN Sessions	Minimum 4 (Up to 32)	Minimum 4 (Up to 32)	Minimum 4 (Up to 32)
Maximum RPVST Instances	507	507	507
Maximum HSRP Groups	490	490	490
Maximum VLANs	4K	4K	4K
Maximum MST Instances	64	64	64
Maximum RPVST Instances	500	500	500

* Shared with IP hosts

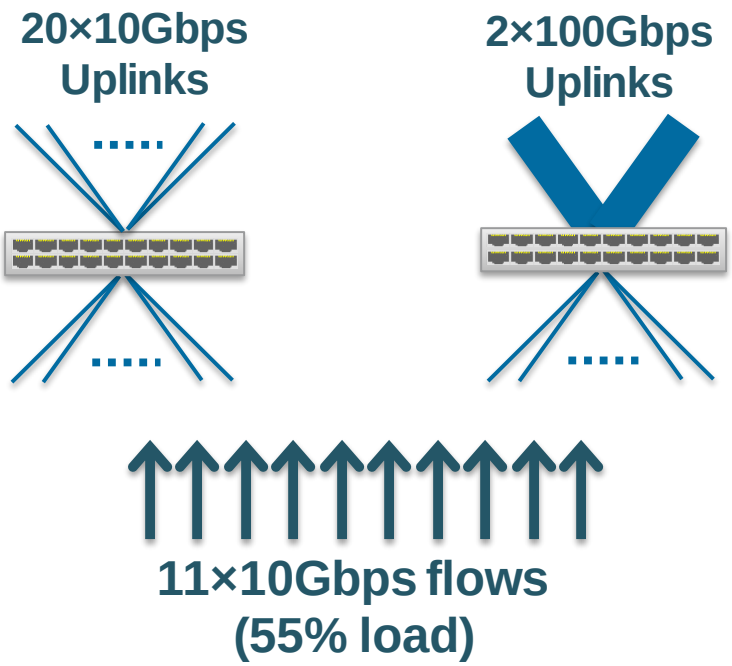
This table shows system maximum scalability. For validated scale, refer to the Nexus 9000 verified scalability white papers

Agenda
Nexus 9000 Architecture

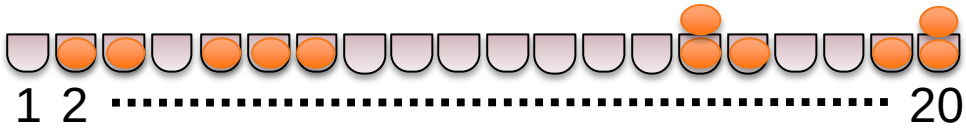
- Nexus 9000
  - Nexus 9000 Hardware
  - Nexus 9500 Chassis
  - Nexus 9500 Line Cards
  - Nexus 9500 Packet Forwarding
  - Nexus 9300
- Nexus 9000 and 40G
- Nexus 9000 Designs: FEX, vPC & VXLAN
- Nexus 9000 & Dev-Ops
- ACI & Nexus 9000

Why Speedup in Network Design

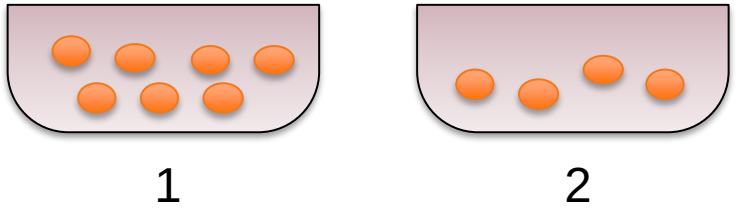
Higher speed links improve ECMP efficiency



Prob of 100% throughput = 3.27%

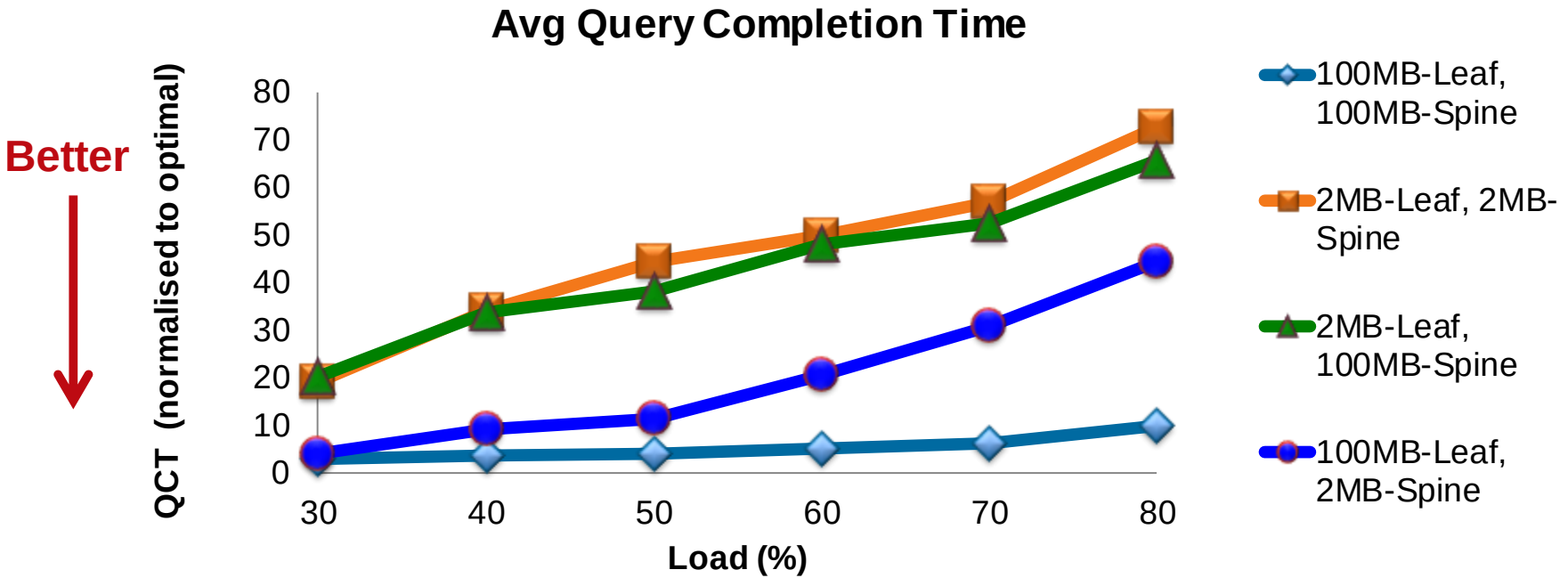


Prob of 100% throughput = 99.95%



Impact of Buffering

Where are large buffers more effective for Incast?



Nexus 9300 Platform Architecture

Uplink Module

- 12-port 40 Gb QSFP+
- Additional 40 MB buffer
- Full VXLAN gateway, bridging and routing capability

Nexus® 9396PQ

- 960G
- 48-port 1/10 Gb SFP+ and 12-port 40 Gb QSFP+
- 2 RU

Nexus 9396TX (Future)

- 960G
- 48-port 1/10 GBaseT & 12-port 40 Gb QSFP+
- 2 RU

Nexus 93128TX

- 1,280G
- 96-port 1/10 G-T and 8-port 40 Gb QSFP+
- 3 RU

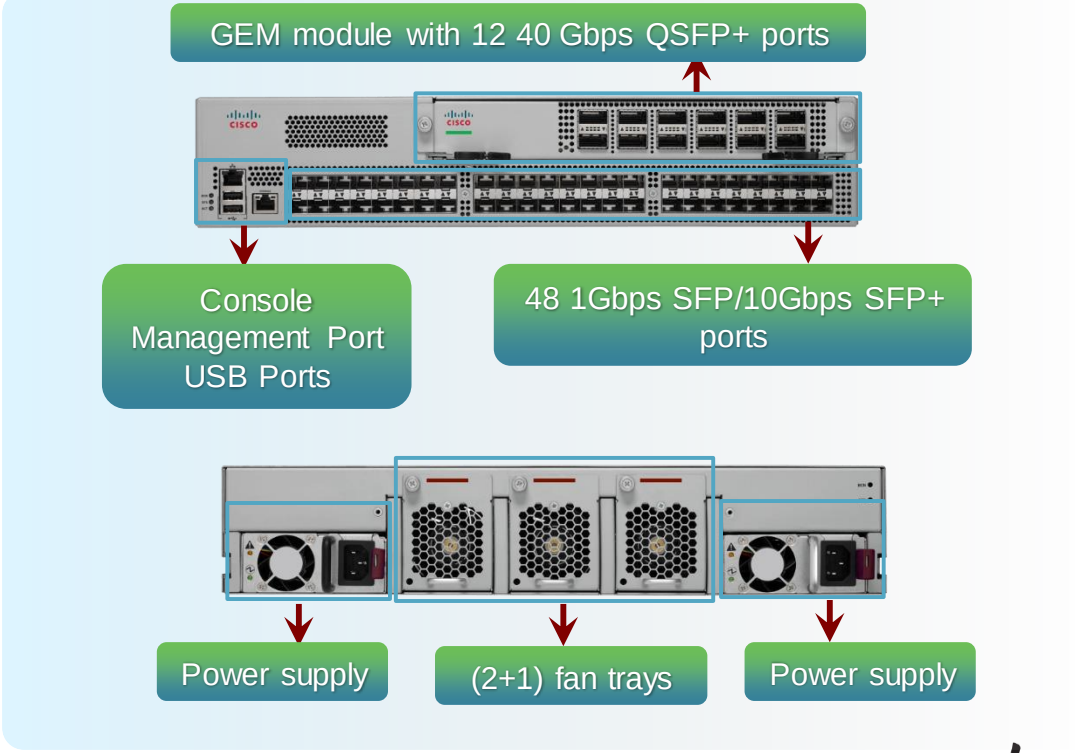
Nexus 9300 - Common

- Redundant fan and power supply
- Front-to-back and back-to-front airflow
- Dual- core CPU with default 64 GB SDD

Nexus 9300 Platform Architecture

Cisco Nexus® 9396PX

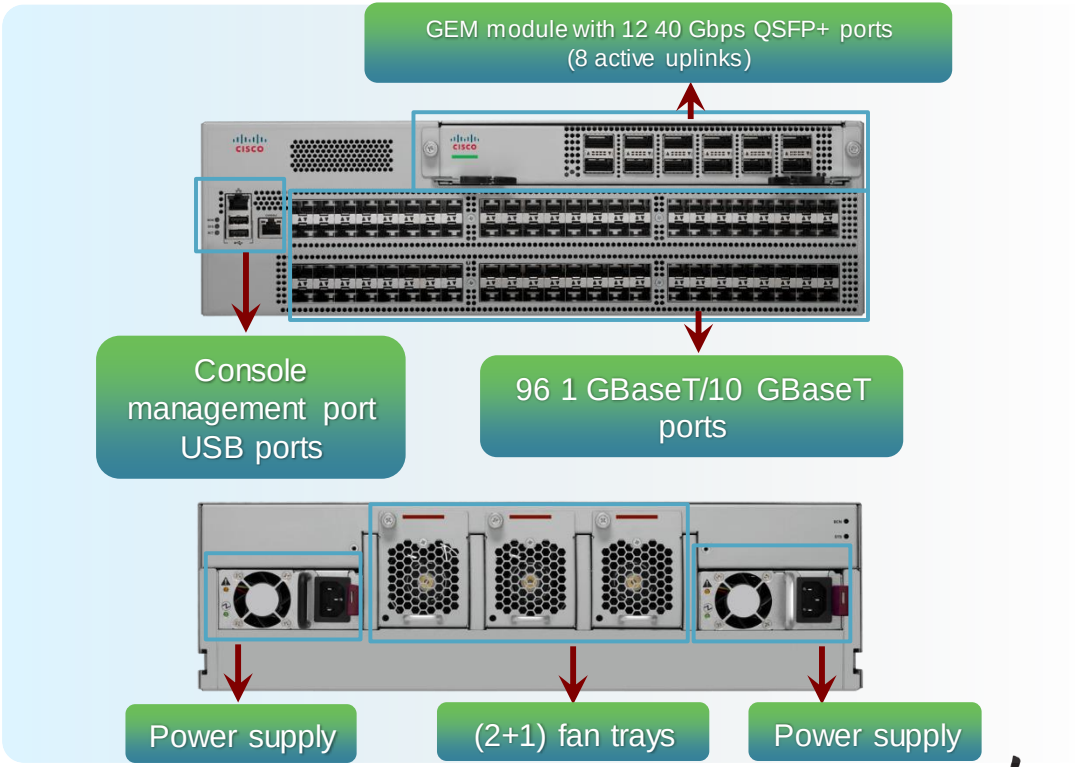
- 2 RU height
- 48 1 Gb SFP/10 Gbps SFP+ ports
- 12 40 Gbps QSFP ports (on GEM module)
- 1 100/1000baseT management port
- 1 RS232 console port
- 2 USB 2.0 ports
- Front-to-back and back-to-front airflow options
- 1+1 redundant power supply options
- 2+1 redundant fans
- No-blocking architecture with line-rate performance on all ports for all packet sizes



Nexus 9300 Platform Architecture

Cisco Nexus® 93128TX

- 3 RU height
- 96 1/10 Gbps BaseT ports
- 8 40 Gbps QSFP ports (on GEM module)
- 1 100/1000baseT management port
- 1 RS232 console port
- 2 USB 2.0 ports
- Front-to-back and back-to-front airflow options
- 1+1 redundant power supply options
- 2+1 redundant fans



Nexus 9300 Platform Architecture

- 12-port 40 Gbps QSFP (FCS)
- Additional 40 MB buffer (3.5 times of BCOM NFE)
- Full VXLAN gateway, bridging, and routing capability
- Common for Nexus® 9396 and Nexus 93128 Switches
  - Four ports will be disabled when installed in a Cisco® Nexus 93128 Switch.
  - A white LED under each QSFP port pair indicates port-pair availability.
  - The LED will be on if the port pair is available.

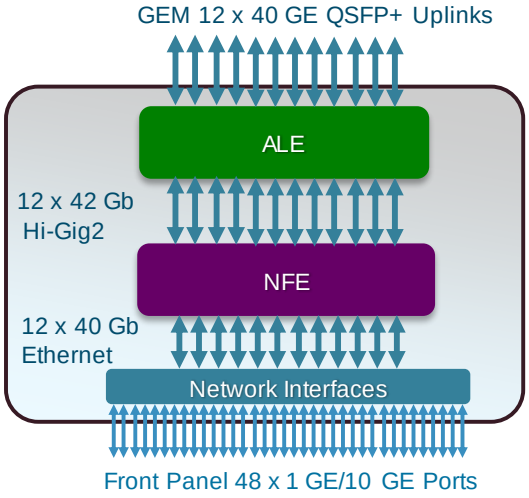
Generic Expansion Module

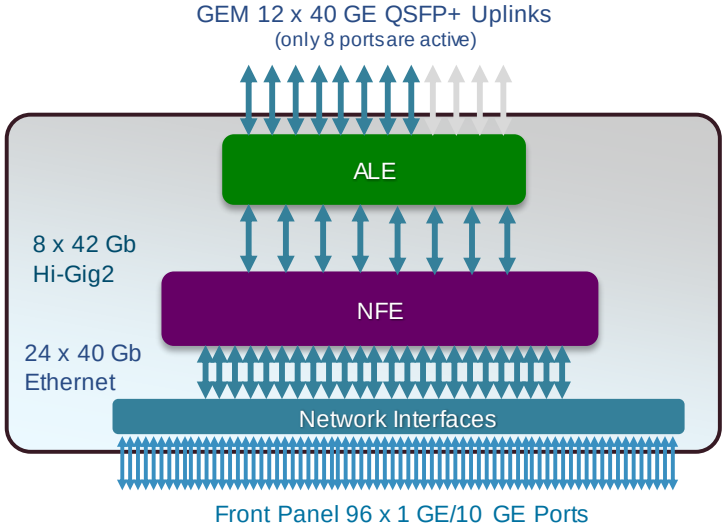
- Redundant (1+1) 650 W and 1200 W AC PS options
- 80-Plus-Platinum-certified power supplies*
- Redundant (2+1) hot-swappable fan trays
- Port side exhaust (blue) and port side intake (red) options to support front-to-back or back-to-front airflow

Fan and Power Supply

* 80 Plus Platinum is equivalent to a Climate Saver or Green Grid Platinum rating

Nexus 9300 System Block Diagram



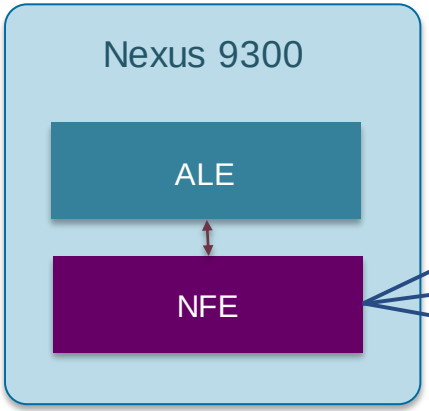


Nexus 9396PQ/ Nexus 9396TX

Nexus 93128TX

Nexus 9300 Unicast Forwarding

- In Nexus 9300 system there is no separate NFE ASIC that would distribute LPM Route learning from the rest of the system.
- As a result of this, the forwarding tables on a single NFE ASIC is completely responsible for LPM Routes, Host Routes, and MAC Address learning.
- However, it is possible to adjust the allocation of table space based on defined templates.



Feature	Option 1	Option 2
LPM Routes	16K	128K
Host Routes	120K	16K
MAC Addresses	96K	32K

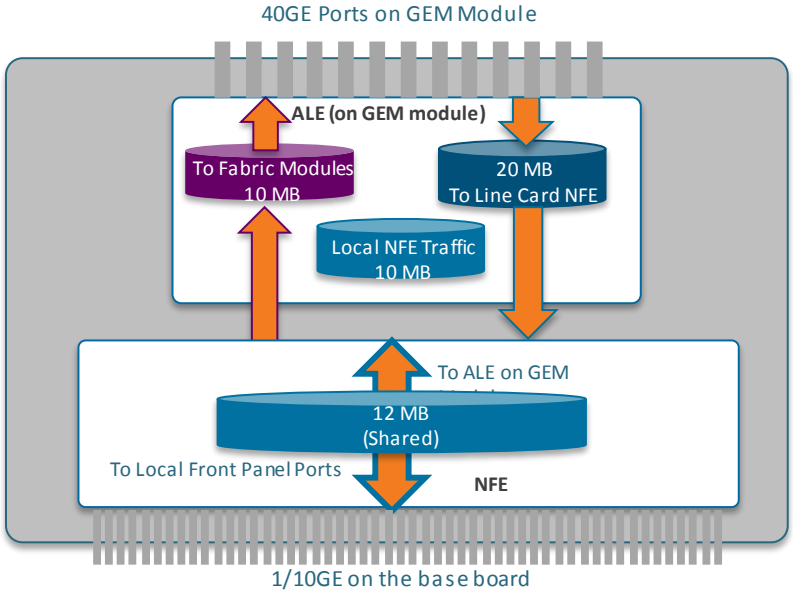
Nexus 9300 Unicast Forwarding

Buffer on NFE:

- Shared 12MB buffer

Buffer on ALE (on GEM module):

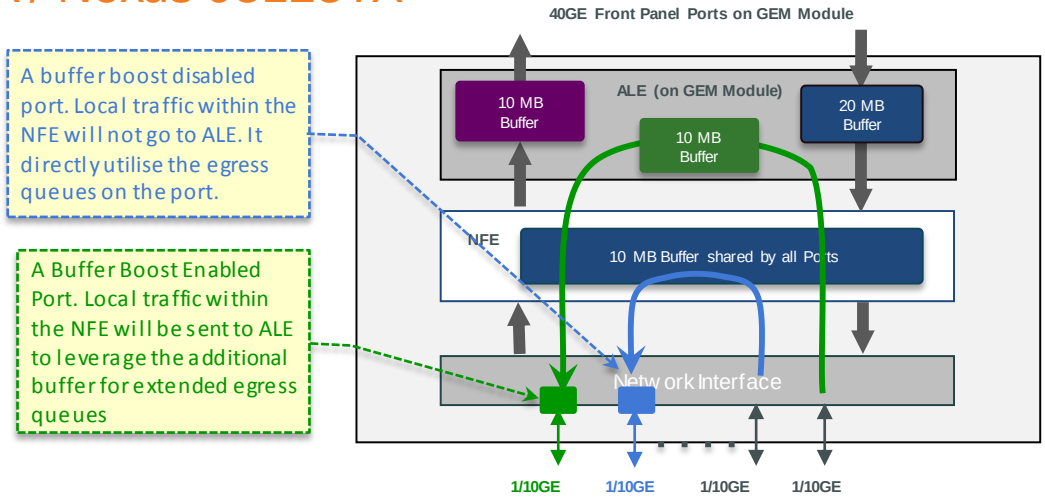
- Shared 10MB egress buffer for all 40GE ports on GEM module;
- Shared 20MB egress buffer for traffic coming from GEM 40GE ports and going to 1/10GE front panel port.
- Shared 10MB egress buffer for local NFE traffic



Nexus 9396PX / Nexus 93128TX

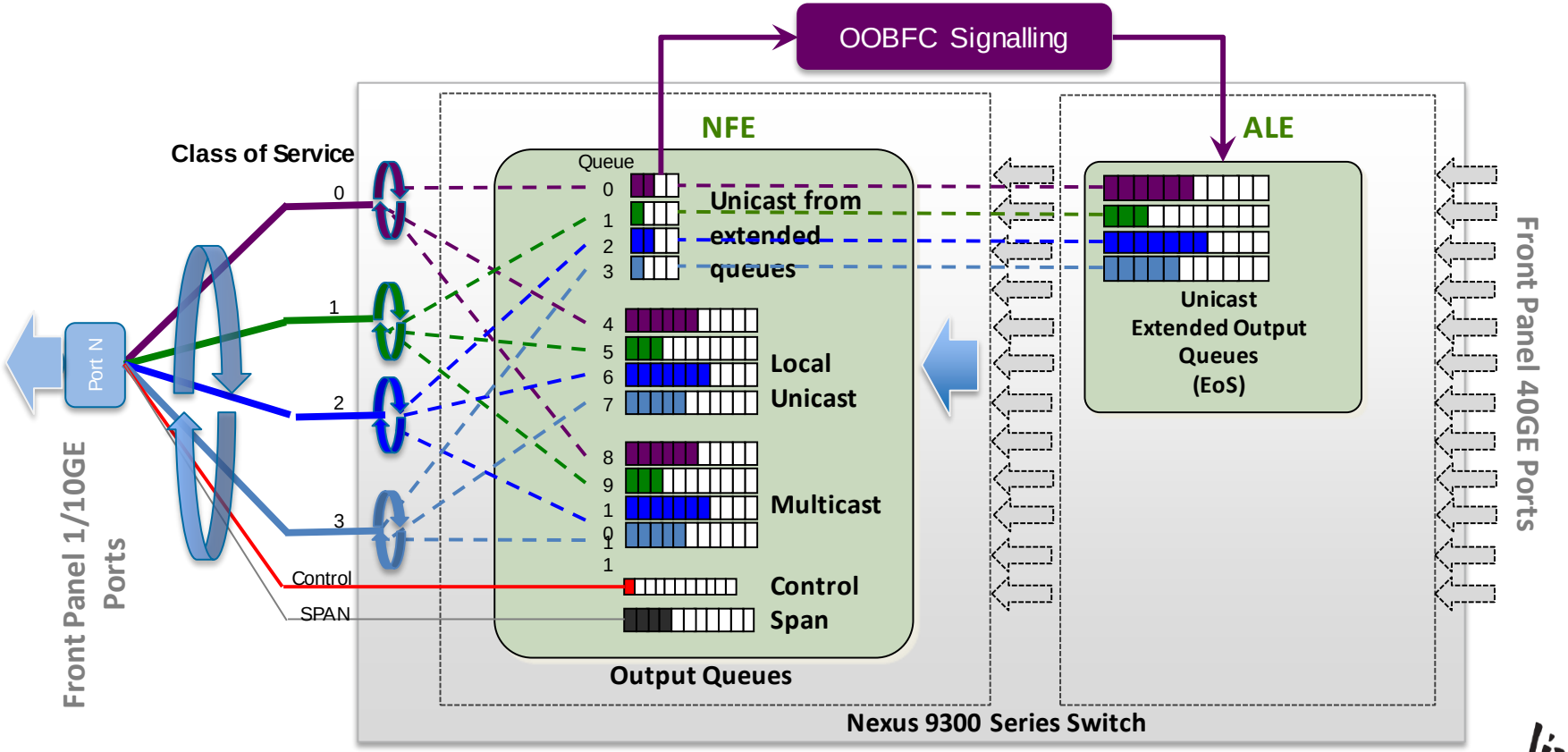
Buffer Boost Function Nexus 9300 Switches

Nexus 9396PX / Nexus 93128TX



- Buffer boost function is introduced on 1/10GE ports to leverage the additional buffer on ALE for local traffic.
- When buffer boost is enabled on a port, local traffic from other 1/10GE ports to this port is sent to ALE for additional buffer space.
- When buffer boots is disabled on a port, local traffic from other 1/10GE ports to this port remains local on NFE.
- Buffer boots is enabled on all 1/10GE ports by default. It is a port property and can be disabled on a per-port basis.

Queuing and Scheduling on Nexus 9300



Nexus 9300 System Scalabilities

In NX-OS Mode

	Nexus 9300	
	Option 1	Option 2
LPM Routes	16K	128K
IP Host Entries	120K	16K
MAC Address Entries	96K	32K
Multicast Routes	32K* (hardware capable of 72K)	8K*
Multicast Fan Outs	8K (no vPC)	8K (no vPC)
IGMP Snooping Groups	32K* (hardware capable of 72K)	8K*
ACL TCAM	Hardware: 4K ingress, 1K egress Available to user: 3K ingress, 768 egress	Hardware: 4K ingress, 1K egress Available to user: 3K ingress, 768 egress
VRF	1000	1000
Max Links in Port Channel	32	32
Max ECMP Paths	64	64
Max vPC Port Channels	528	528
Max Active SPAN Sessions	4	4
Max RPVST Instances	507	507
Max HSRP Groups	490	490
MAX VLANs	4K	4K
SPAN/ERSPAN	4 active sessions	4 active sessions

* Shared with IP hosts

Agenda
Nexus 9000 Architecture

- Nexus 9000
  - Nexus 9000 Hardware
  - Nexus 9500 Chassis
  - Nexus 9500 Line Cards
  - Nexus 9500 Packet Forwarding
  - Nexus 9300
- Nexus 9000 and 40G
- Nexus 9000 Designs: FEX, vPC & VXLAN
- Nexus 9000 & Dev-Ops
- ACI & Nexus 9000

Optical Innovation --- Removing 40 Gb Barriers

Problem

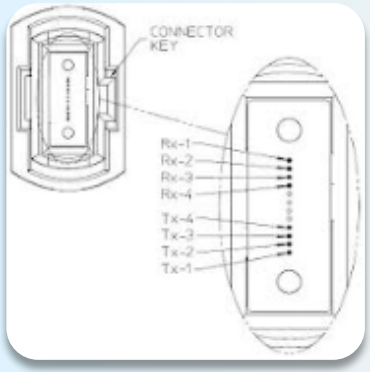
- 40 Gb optics are a significant portion of capital expenditures (CAPEX)
- 40 Gb optics require new cabling

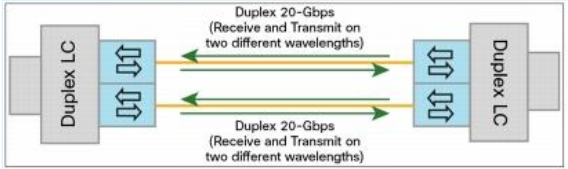
Solution

- Re-use existing 10 Gb MMF cabling infrastructure
- Re-use patch cables (same LC connector)

Cisco® 40 Gb SR-BiDi QSFP

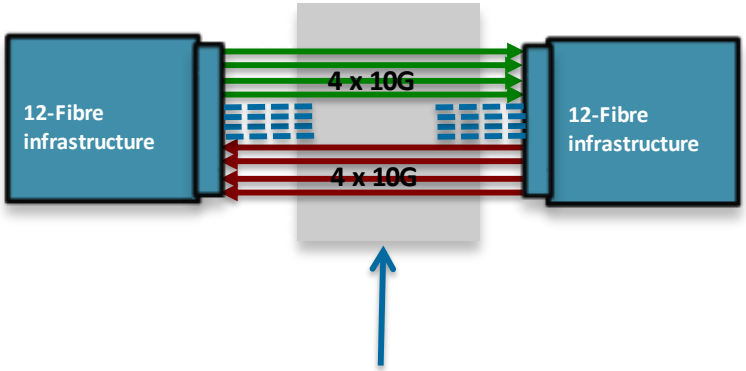
- QSFP, MSA-compliant
- Dual LC connector
- Support for 100 m on OM3 and upto 150m on OM4
- TX/RX on two wavelengths at 20 Gb each





Available end of CY13 and supported across all Cisco QSFP ports

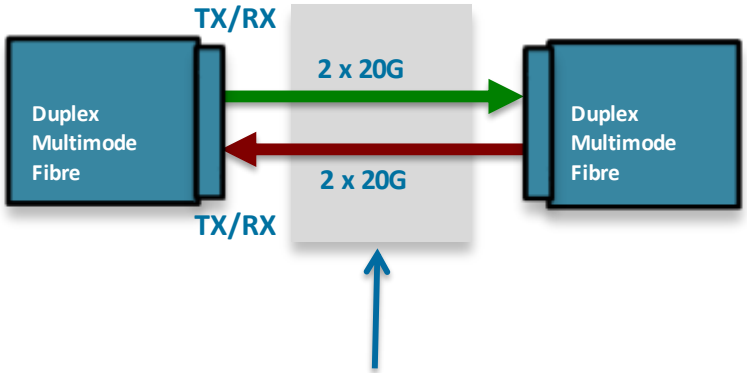
QSFP-BIDI vs. QSFP-40G-SR4



12-Fibre ribbon cable with MPO connectors at both ends

Higher cost to upgrade from 10G to 40G due to 12-Fibre infrastructure

QSFP-BIDI

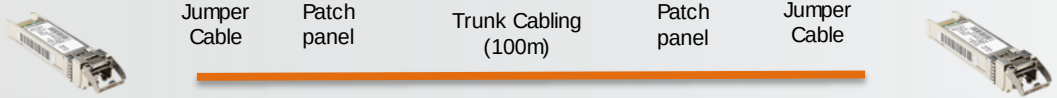


Duplex multimode fibre with Duplex LC connectors at both ends

Use of duplex multimode fibre lowers cost of upgrading from 10G to 40G by leveraging existing 10G multimode infrastructure

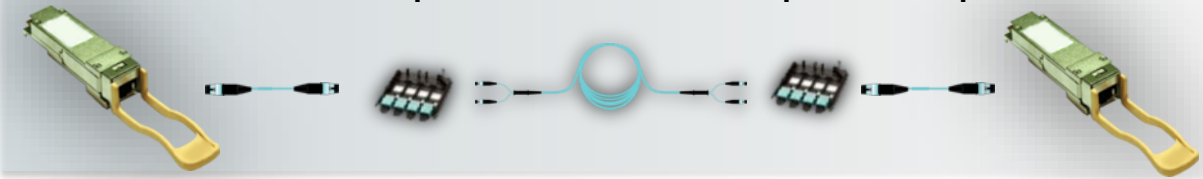
Significant Cabling and Transceiver Savings

10G Optical Link

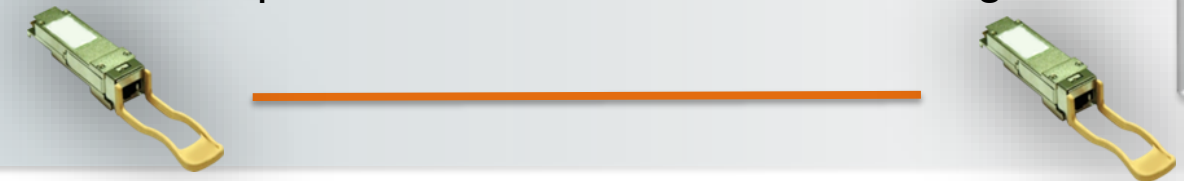


+$6,259*

Traditional 40G Optical Link—Complete Replacement



40G BiDi Optical Link—Reuse all 10G Cabling/Patch Panels



+$2,200*

Cost Avoidance Brownfield

Cost Savings Greenfield

$4,059
SAVINGS (LIST) PER 40G LINK

Source: Corning OM3 Cable & Patch Panel list prices, Cisco 40G BiDi list price, Competitors 40G SR4 list price

Agenda
Nexus 9000 Architecture

- Nexus 9000
  - Nexus 9000 Hardware
  - Nexus 9500 Chassis
  - Nexus 9500 Line Cards
  - Nexus 9500 Packet Forwarding
  - Nexus 9300
- Nexus 9000 and 40G
- Nexus 9000 Designs: FEX, vPC & VXLAN
- Nexus 9000 & Dev-Ops
- ACI & Nexus 9000

Nexus 9500/9300 FEX Support

Optimised Model for Long Term TCO during Evolution

Cisco Nexus® 5500

Cisco Nexus® 2000 FEX

Cisco Nexus® 9300

Cisco Nexus® 2000 FEX

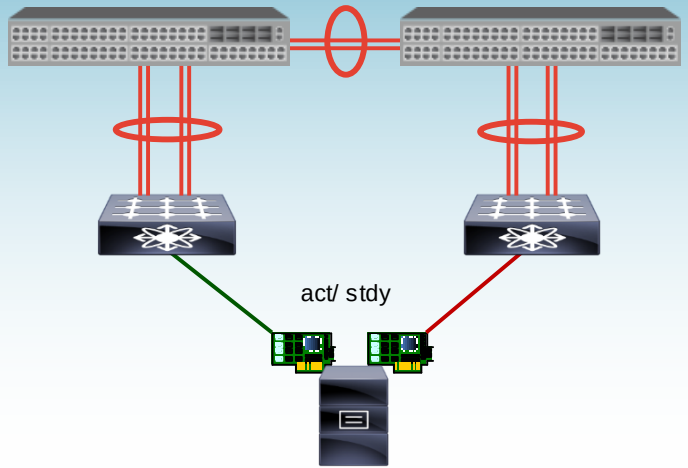
- Migration of Nexus 5500 to Nexus 9300 provides
- Increased scalability
  - 160K MAC
  - 16K Routes
  - 44K MRoutes
  - 160K IGMP Groups
- Addition of 40G uplinks for lower oversubscription
- Addition of VXLAN Bridging, Gateway and Routing capabilities
- Line Rate Layer 2 and Layer 3
- Reduction of Latency

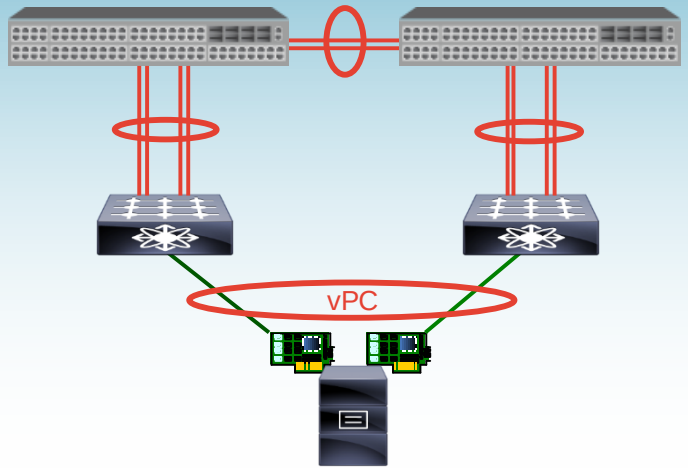
Nexus 9500/9300 FEX Support (Cont'ed)

Supported Topology

Supported FEX Topology:

- Single-homed FEX
- vPC port channel to hosts

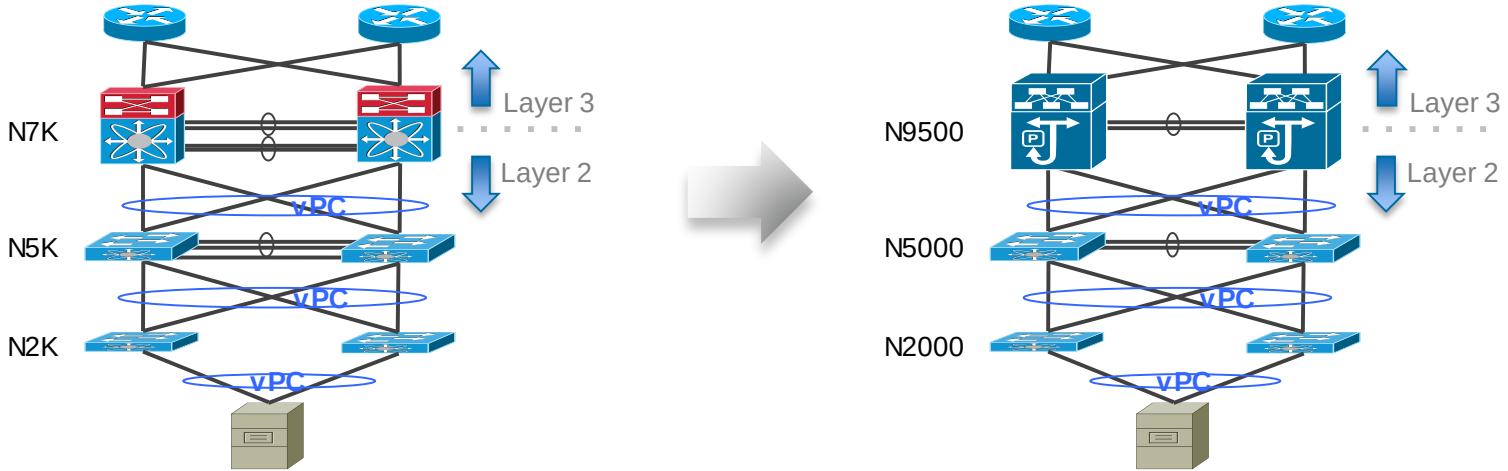




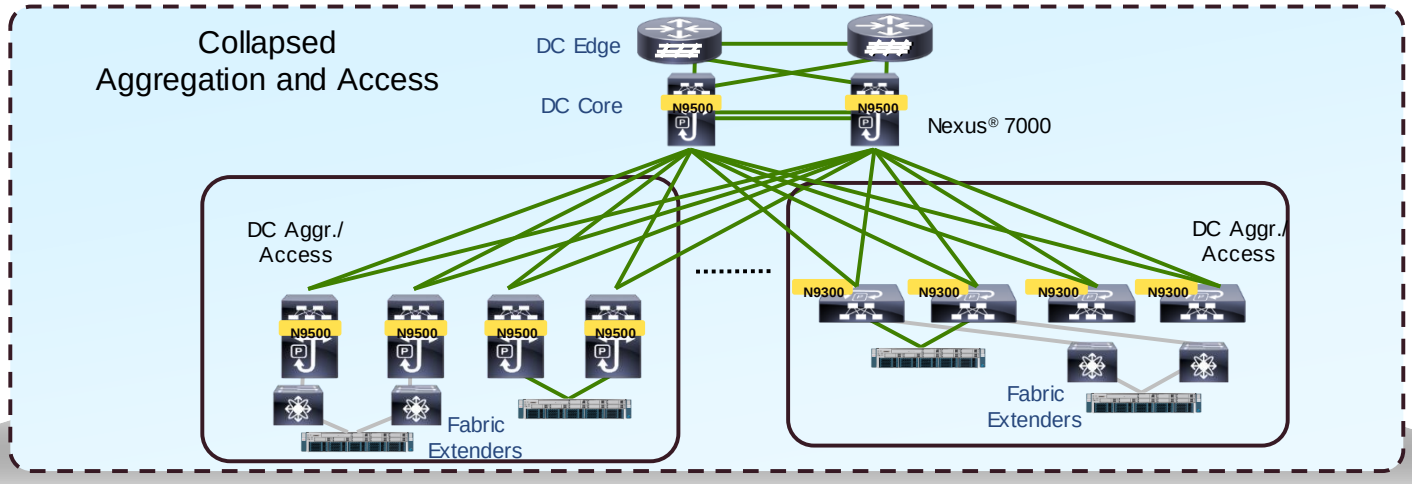
Migration and Interop with Existing Nexus Platforms

Pod Design Migration with vPC

- Nexus 9000 is fully compatible with all existing Nexus vPC & FEX designs
- When customer is looking at consolidation of multiple aggregation or high density 40G aggregation look to migrate to Nexus 9500



Collapsed Aggregation and Access



Configuration Options

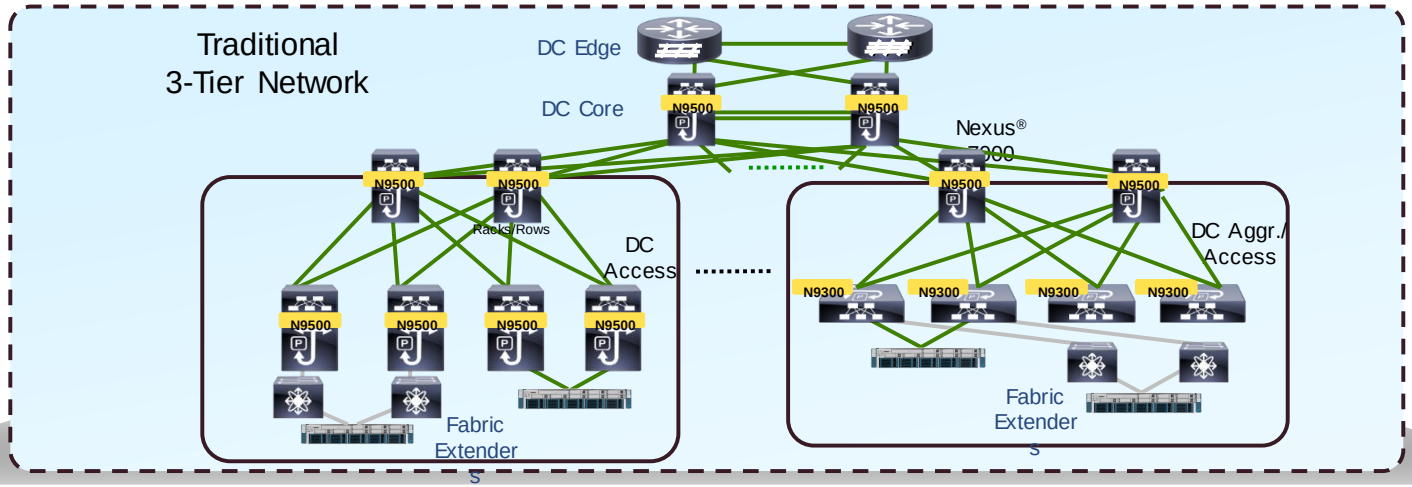
Option 1:

- Layer 2 and 3 boundary is on the core
- vPC between aggregation and core
- Nexus 9500 can be EoR with FEX in the racks
- Nexus 9300 can be MoR and EoR with FEX in the racks

Option 2:

- Layer 3 all the way to access
- ECMP between access and aggregation
- Nexus 9500 can be EoR with FEX in the racks
- Nexus 9300 can be MoR and EoR with FEX in the racks

Traditional 3-Tier Network Design



Configuration Options

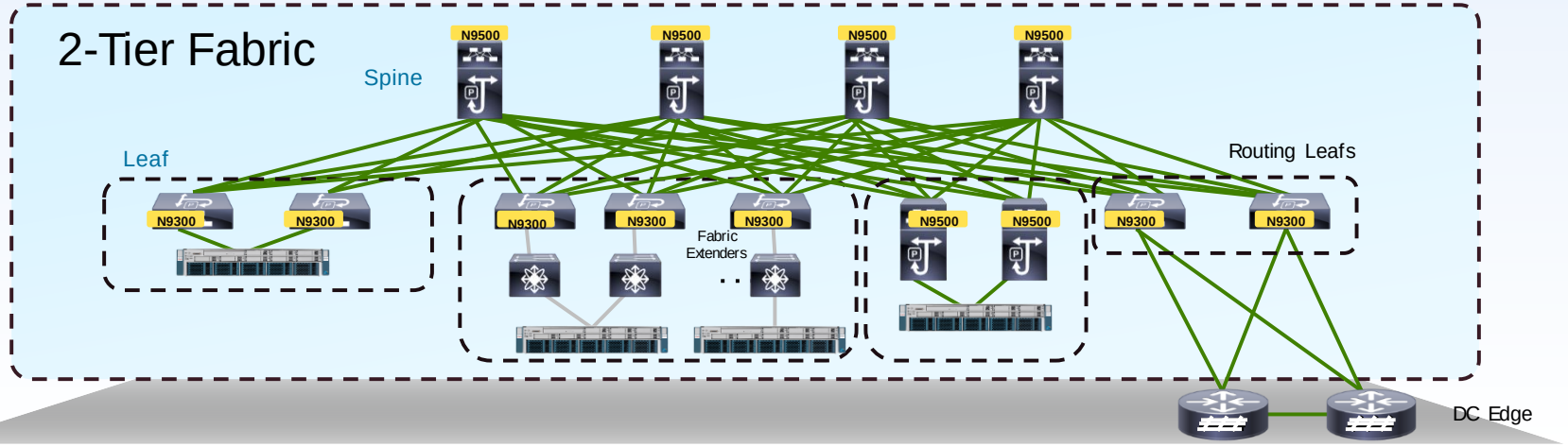
Option 1:

- Layer 2 or 3 boundary is on aggregation
- vPC between access and aggregation
- Nexus 9500 can be EoR
- Nexus 9300 can be ToR or MoR

Option 2:

- Layer 2 or 3 boundary is on aggregation
- vPC between access and aggregation
- Nexus 9500 can be EoR
- Nexus 9300 can be ToR or MoR

Spine-Leaf Fabric Design



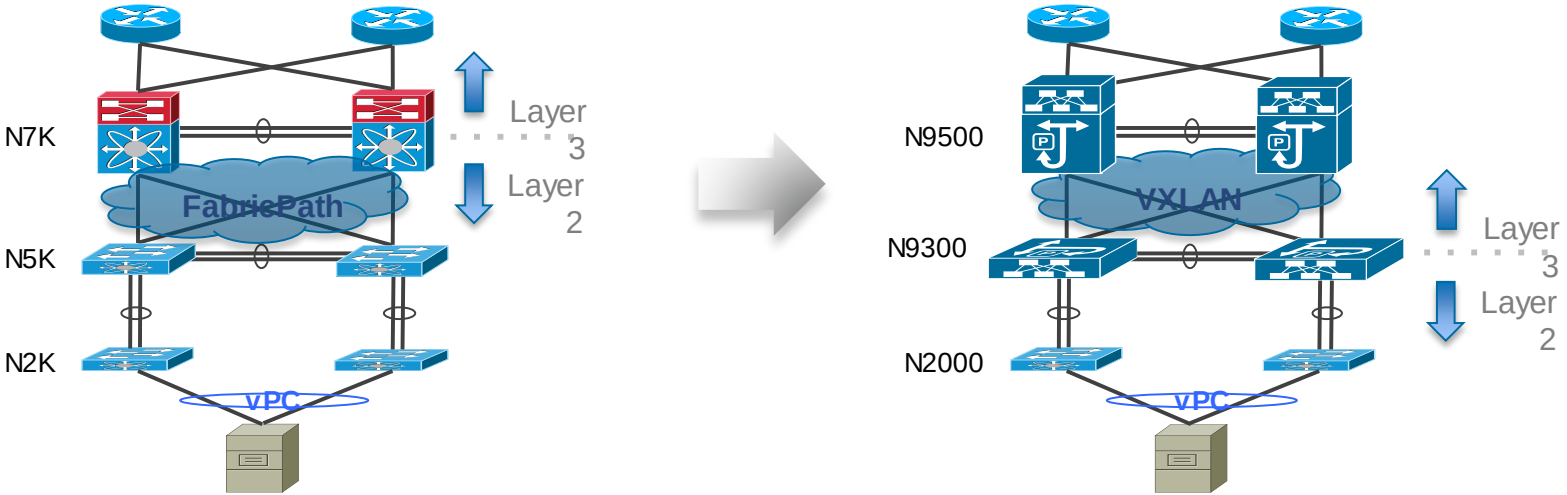
Configuration Options:

- Layer 3 IP fabric
- Layer 3 all the way to access
- Layer 2 extensibility can be provided by VXLAN
- Nexus® 9500 can be EoR
- Nexus 9300 can be ToR and MoR

Migration and Interop with Existing Nexus

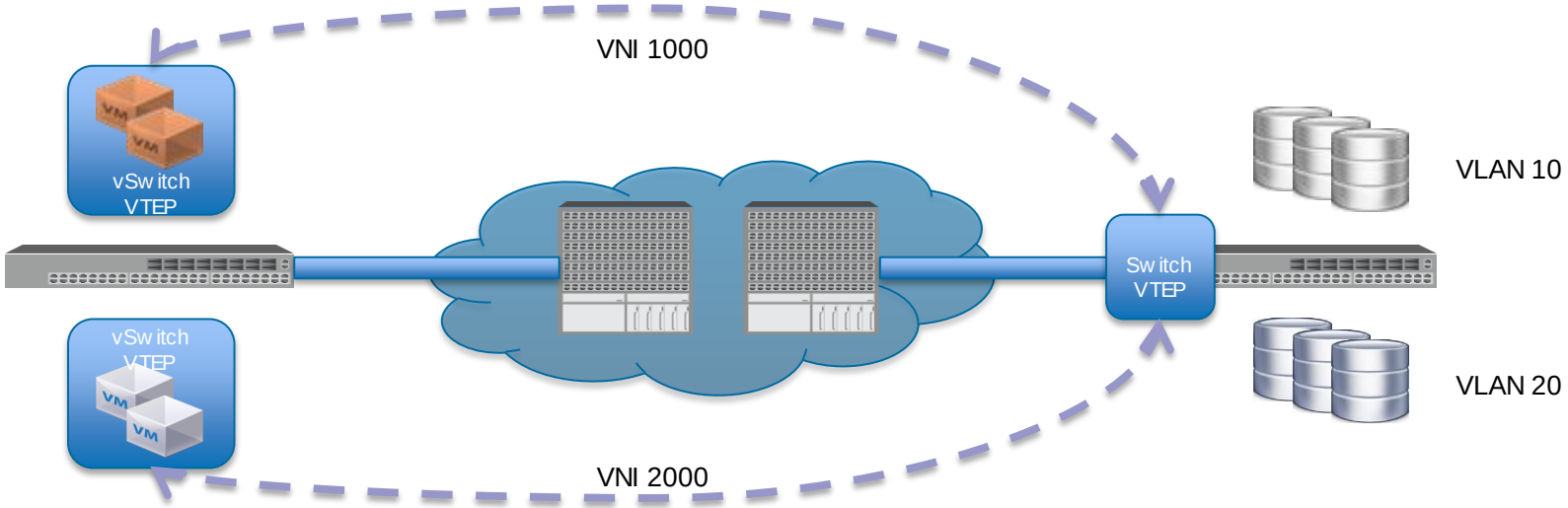
PoD Design Migration with VXLAN

- When customer is looking to migrate to a routed access (Layer 3 to the edge) design look to position Nexus 9300 and 9500 with integrated VXLAN capabilities
- When customer is looking to add VXLAN capabilities look to position Nexus 9300 for both VXLAN Gateway (P to V) and VXLAN bridging and routing capabilities



VXLAN Overview

VXLAN can be implemented on both Hypervisor-based Virtual Switches to allow for scalable VM deployments, as well as on Physical switches, which provides the ability to bridge VXLAN segments back into VLAN segments. In these cases, the Physical Switch instantiates a VTEP, and function as a VXLAN Gateway…



VXLAN: Bridging, Gateway and Routing

- Layer-2 Extension with VXLAN
  - VXLAN ← → VLAN (Gateway)
  - VXLAN ← → VXLAN (Bridging)
  - VLAN ← → VLAN (Bridging)

- Inter-VXLAN Routing
  - VXLAN A ← → VXLAN B (Routing)
    (IP Subnet A ← → IP Subnet B)

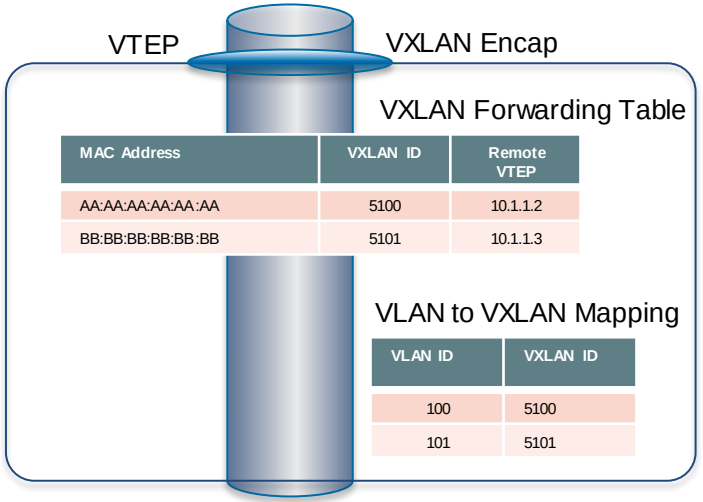
Nexus 9000 Series

VXLAN Support

- VXLAN bridging and gateway are supported on Nexus 9300 Series switches
- VXLAN bringing and gateway will be supported on Nexus 9500 line cards
- VXLAN routing will be supported on Nexus 9300 and Nexus 9500

VXLAN Gateway Function on Nexus 9000 Series

```
feature nv overlay
feature vn-segment-vlan-based

vlan 100
  vn-segment 5100

interface e1/1 ← local switch port
  switchport
  switchport mode access
  switchport access vlan 100
  no shut

interface nve1  ← VTEP tunnel interface
  no shutdown
  source-interface loopback0
  member vni 5100 mcast-group 239.1.1.1

interface loopback0
  ip address 10.1.1.2/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

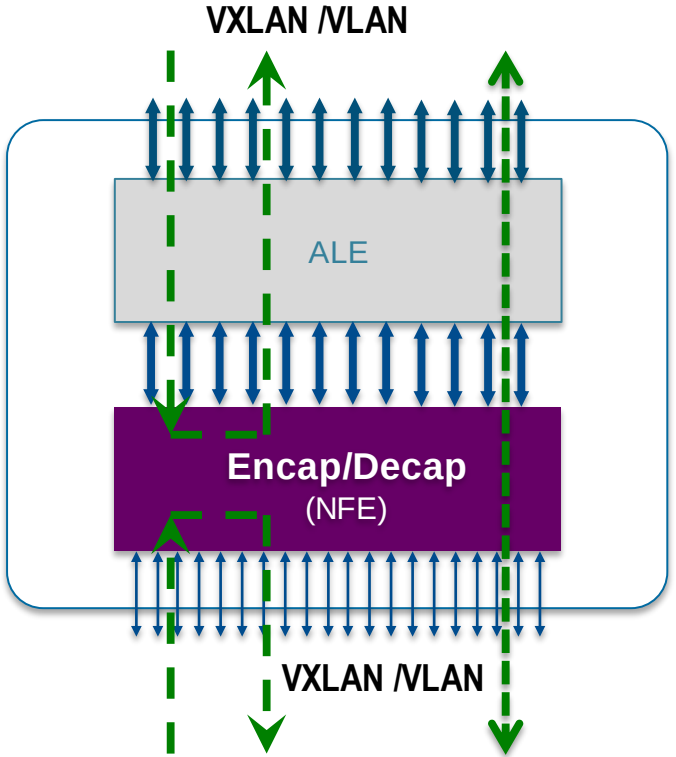



```
N9396-2# show nve vni
Interface          VNI        Multicast-group   VNI State
----------------   --------   ---------------   ---------
nve1               5100        239.1.1.1            up
nve1               5101        239.1.1.2            up

N9396-2# show nve peers
Interface           Peer-IP              VNI         Up Time
-------------       ----------------     --------    -------
nve1                10.1.1.3             5100       00:52:24
nve1                10.1.1.4             5101       00:41:16
N9396-2#
```

VXLAN Forwarding on Nexus 9000 NX-OS Mode

VXLAN Bridging and Gateway

- VXLAN Encapsulation and De-encapsulation occur on T2
- Bridging and Gateway are independent of the port type (1/10/40G ports)
- Encapsulation happens on the egress port
- Decapsulation happens on the ingress port

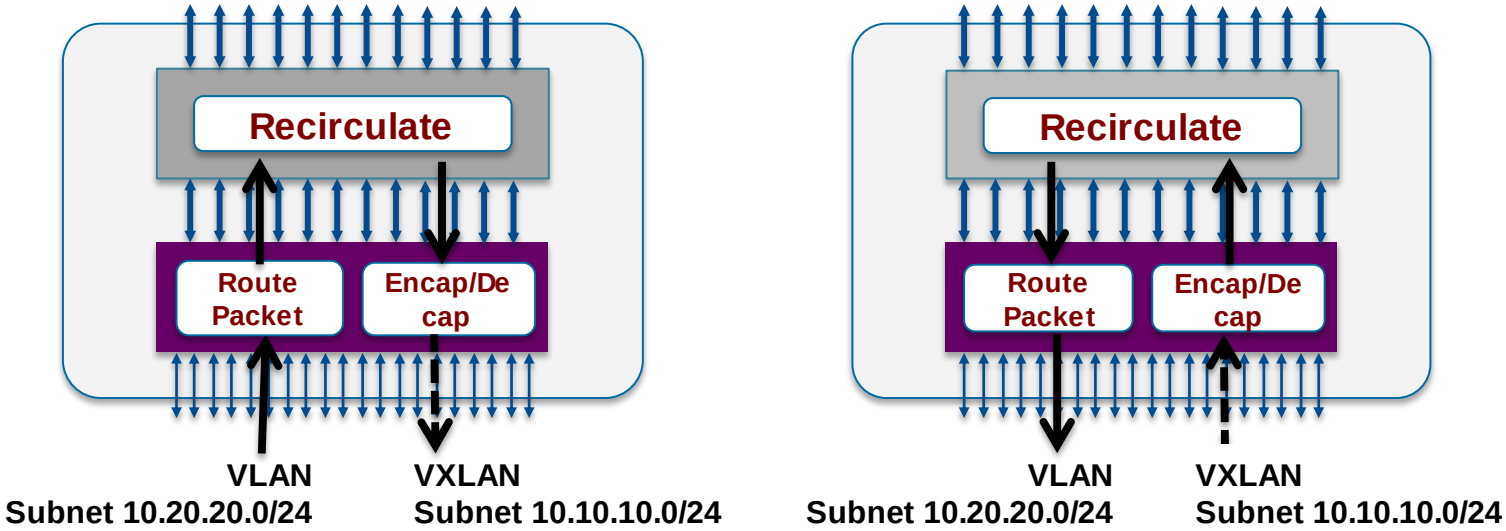


VXLAN Forwarding on Nexus 9000 NX-OS

Mode VXLAN Routing

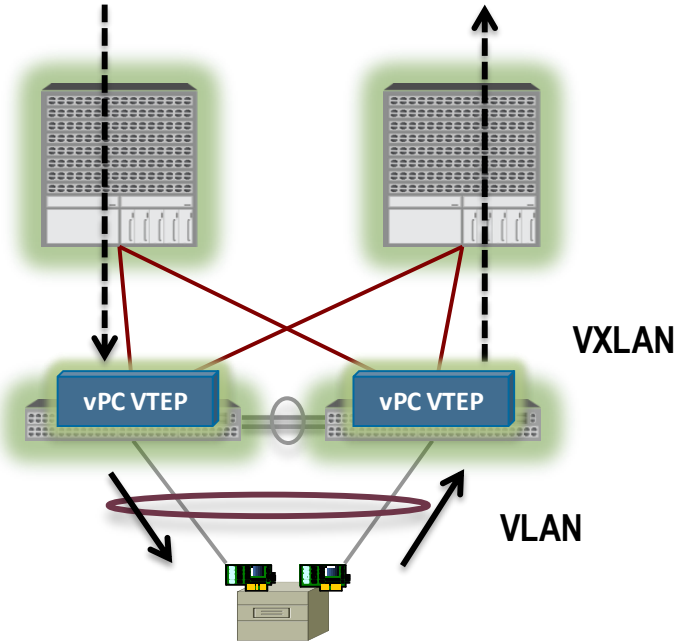
VXLAN Routing is not supported currently on Broadcom

- Additional recirculation required for VXLAN routing through NS



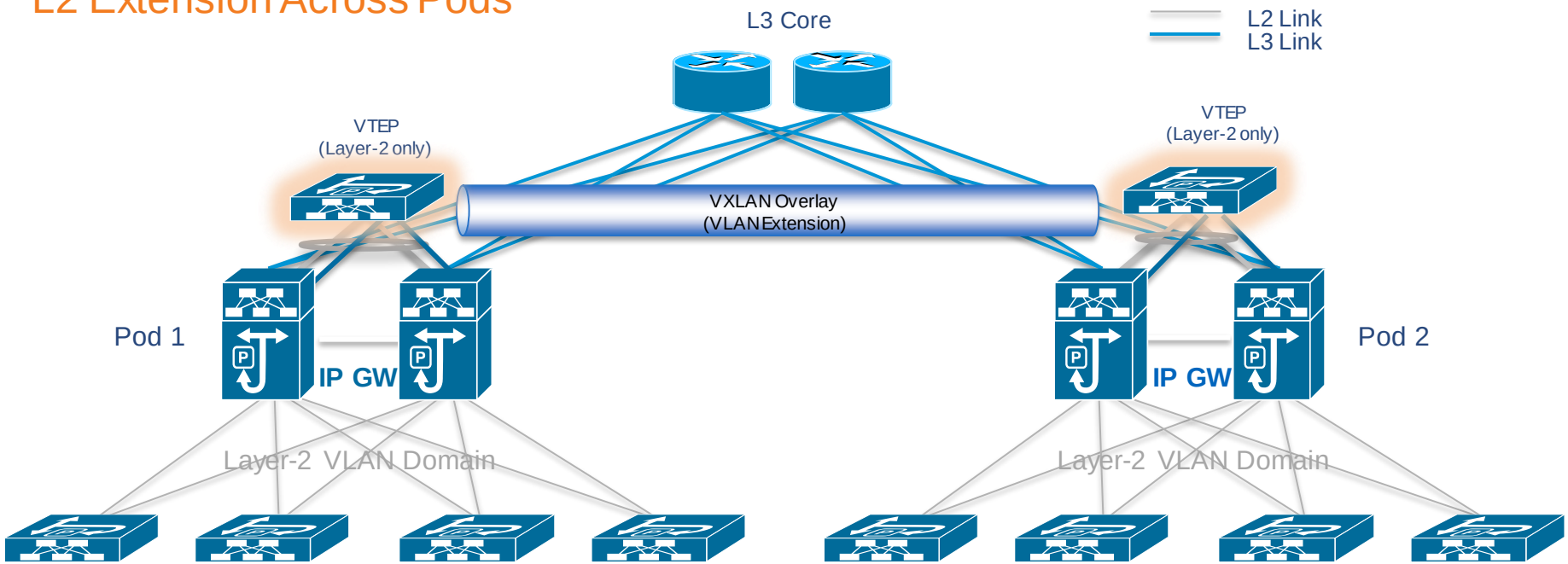
vPC VTEP with Nexus 9000 Series Switches

- When vPC is enabled an ‘anycast’ VTEP address is programmed on both vPC peers
- Symmetrical forwarding behaviour on both peers provides
- Multicast topology prevents BUM traffic being sent to the same IP address across the L3 network (prevents duplication of flooded packets)
- vPC peer-gateway feature must be enabled on both peers
- VXLAN header is ‘not’ carried on the vPC Peer link



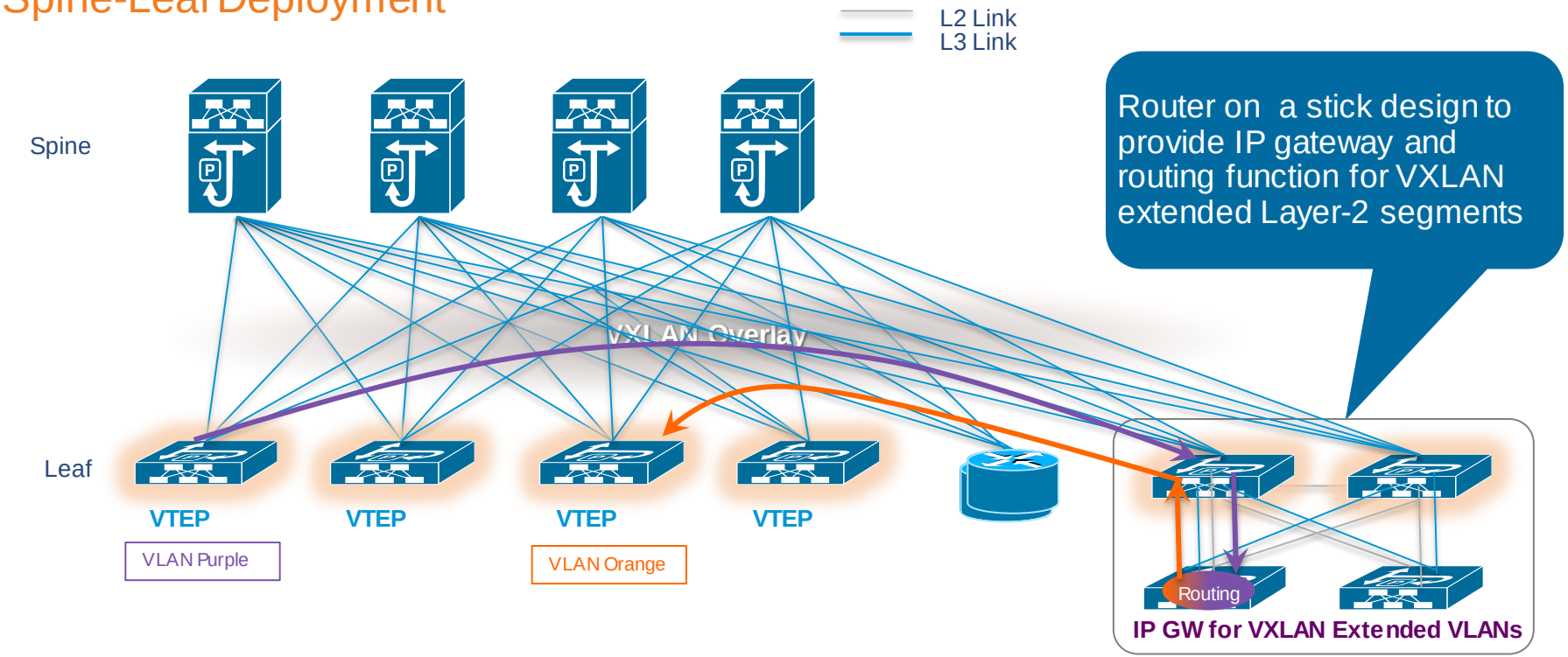
Design with VXLAN Bridging Only

L2 Extension Across Pods



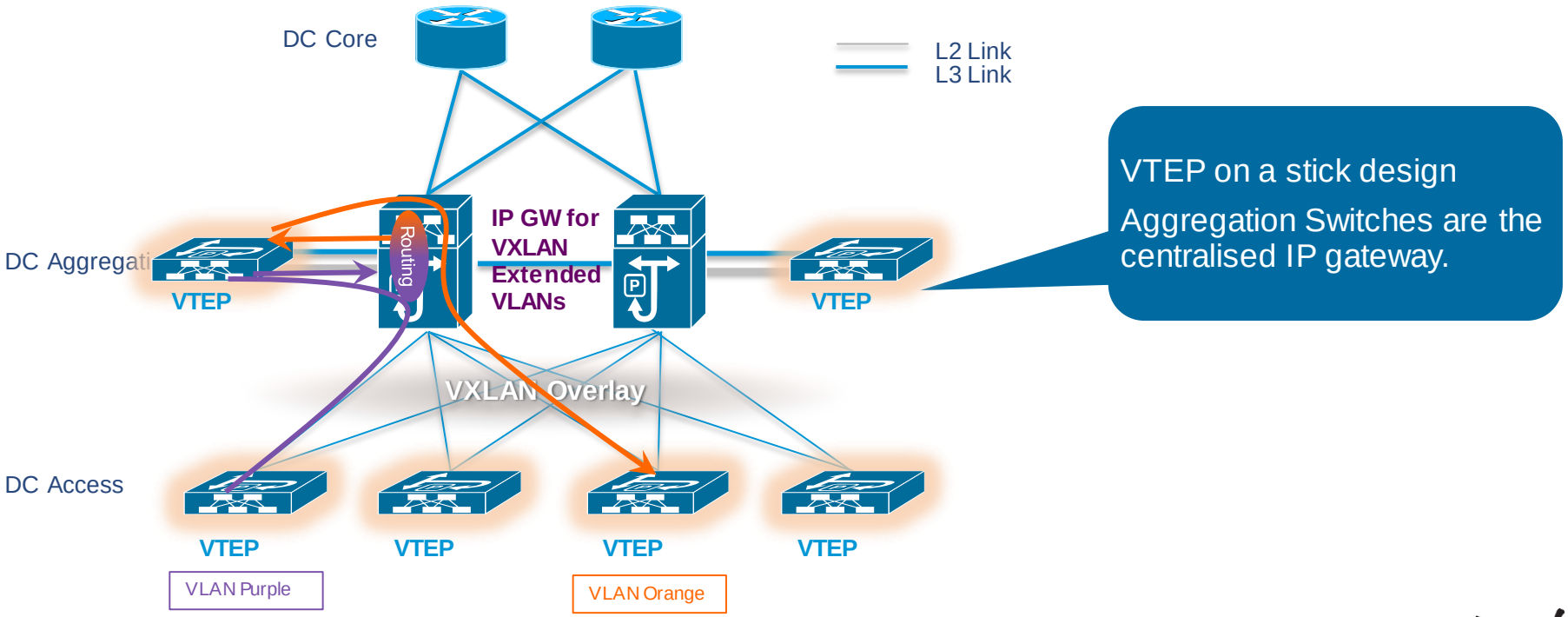
Design with VXLAN Bridging Only

Spine-Leaf Deployment



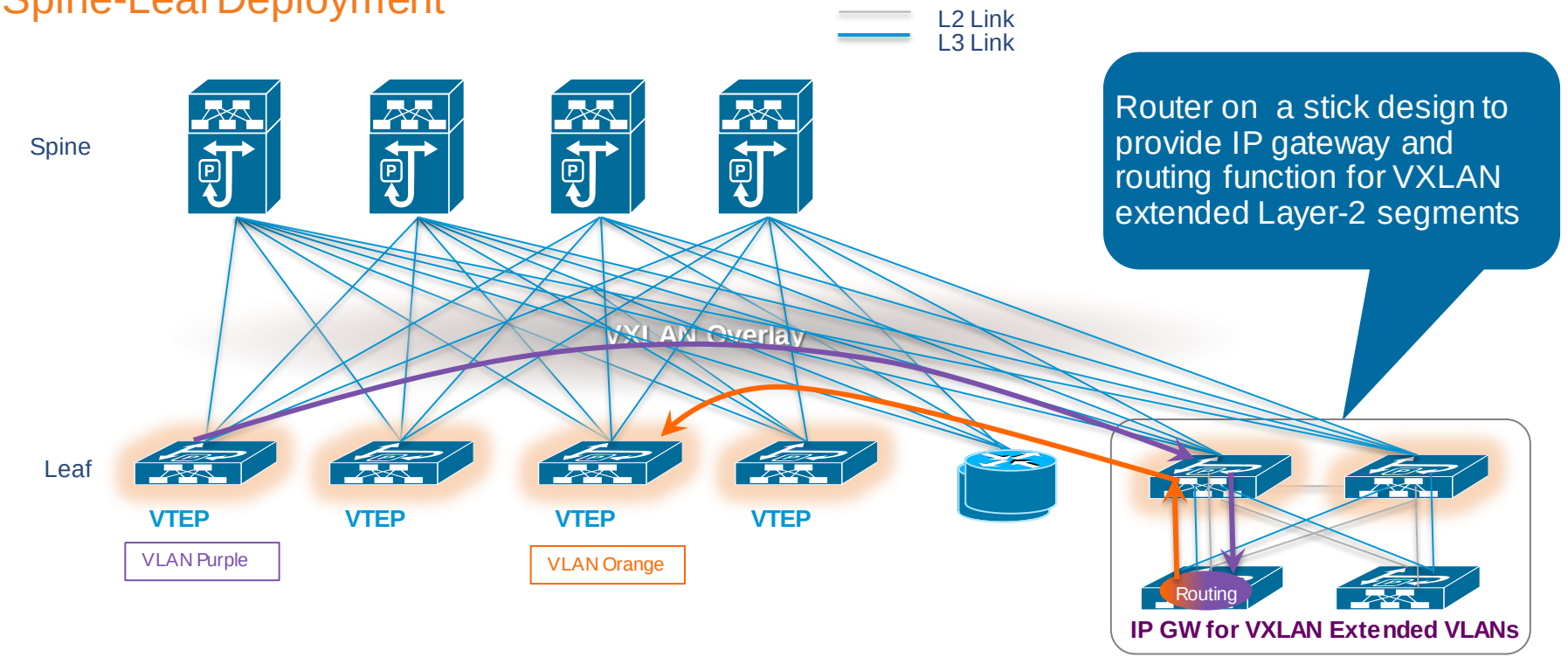
Design with VXLAN Bridging Only

Design Options for VXLAN Routing --- Option B: VTEP on a Stick



Design with VXLAN Bridging Only

Spine-Leaf Deployment



Agenda
Nexus 9000 Architecture

- Nexus 9000
  - Nexus 9000 Hardware
  - Nexus 9500 Chassis
  - Nexus 9500 Line Cards
  - Nexus 9500 Packet Forwarding
  - Nexus 9300
- Nexus 9000 and 40G
- Nexus 9000 Designs: FEX, vPC & VXLAN
- Nexus 9000 & Dev-Ops
- ACI & Nexus 9000

Optimised Cisco NX-OS

Purpose-Built Data Centre OS

- **Modern:** 64-bit Linux 3.4.10 Kernel; single image for Nexus® 9500 and 9300; combined kick-start and system image
- **Comprehensive:** Purpose-built data centre feature set, including Layer 2, Layer 3, and VXLAN
- **Modular:** Code runs in DRAM only when invoked
- **Fault containment:** Complete process isolation for both features and services
- **Resiliency:** Restartable, user-space network stack and drivers; support for ISSU (modular) and OS patchability
- **Management infrastructure:** CLI, SNMP, NetConf/XML, Cisco onePK™, Open Containers, JSON

Nexus 9000: Openness of Linux

Programmable

- NX-API
  - JSON-RPC
  - XML/JSON
- Python scripting
- Customisable CLIs
- BASH access
- Broadcom shell access
- Linux containers
- OpenFlow support
- Cisco onePK™

Automation and Orchestration

- OpenStack network plugin
- Chef
- Puppet
- XMPP support
- OpenDaylight integration

Visibility

- Dynamic buffer monitoring
- Enhanced Ethanalyser
- SMTP email “pipe” output
- Embedded Event Manager (EEM)
- Flow monitoring
- vTracker

SNMP (v1, v2, v3), Syslog, NETCONF, RMON, CLI

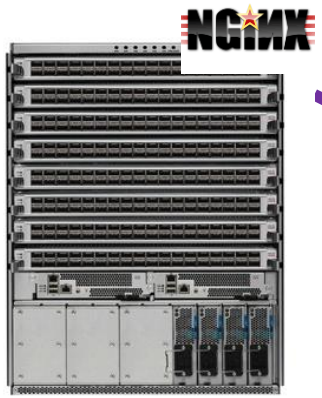
Python Scripting

- Built in Python Shell
- Can be used to execute CLI commands and reference Objects through Python interpreter
- Most commands can be executed to return the command output as a Python Dictionary
- Pass arguments to python scripts from CLI
- Libraries portable
- Integration with Embedded Event Manager (EEM)

Python Modules in the Cisco Package

- acl – IPv4 and IPv6 related access list classes
- bgp – BGPSession and BGPSession.BGPNeighbor classes
- cisco_secret – CiscoSecret classes used by BGPSession.BGPNeighbor.cfg_password()
- cisco_socket – Allows getting/setting the vrf on specific sockets
- feature – Inspect, enable and disable features
- interface – Interface related but works differently than you would expect probably
- key - used with the line parser to parse lines
- line_parser – used to parse lines of cli output
- md5sum – Get md5sum of of an image
- nxcli – command line parser related
- ospf – OSPFSession and OSPFSession.OSPFInterface classes
- routemap – manipulate routemap related objects
- routes – manipulate route related objects
- section_parser – Parses sections
- ssh – get, set and interact with secure shell related objects
- system – get and set system related info
- tacacs – tacacs+ related
- vrf – set and get VRF's

Enhanced NX-OS NX-API



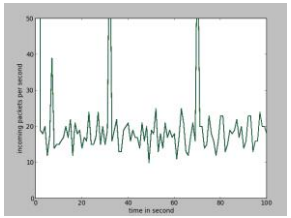
N9K


```
{
  "ins_api": {
    "type": "cli_show",
    "version": "0.1",
    "sid": "session1",
    "outputs": {
      "output": {
        "TABLE_interface": {
          "ROW_interface": [
            {
              "interface": "mgmt0",
              "state": "up",
              "ip_addr": "172.21.128.227",
              "speed": "1000",
              "mtu": "1500"
            },
            {
              "interface": "loopback0",
              "state": "up"
            }
          ]
        }
      }
    }
  }
}
```


HTTP

CLI Input


```
<?xml version="1.0" encoding="ISO-8859-1"?>
<ins_api>
  <type>cli_conf</type>
  <version>0.1</version>
  <sid>1</sid>
  <input>show interface brief </input>
  <output_format>xml</output_format>
</ins_api>
```

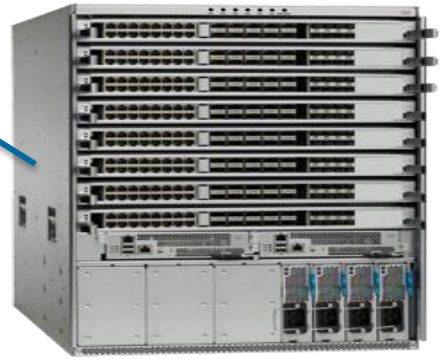
Programmability

Cisco live!

Bash Access and Linux Containers


```
TME-1-9508-1# run bash
bash-4.2$
bash-4.2$ ifconfig -a
dummy0    Link encap:Ethernet  HWaddr a6:9f:04:2b:d3:ef
          BROADCAST NOARP  MTU:1500  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:0 (0.0 B)  TX bytes:0 (0.0 B)

eth0      Link encap:Ethernet  HWaddr 00:00:00:00:1b:01
          inet6 addr: fe80::200:ff:fe00:1b01/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:477374330 errors:0 dropped:0 overruns:0 frame:0
          TX packets:272305025 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:79582379696 (74.1 GiB)  TX bytes:58519512337 (54.5 GiB)

eth1      Link encap:Ethernet  HWaddr c0:67:af:a0:de:2e
          inet6 addr: fe80::c267:afff:fea0:de2e/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:4813640 errors:0 dropped:0 overruns:0 frame:0
          TX packets:182072 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:706614536 (673.8 MiB)  TX bytes:91737078 (87.4 MiB)

eth2      Link encap:Ethernet  HWaddr 00:00:00:01:1b:01
          inet6 addr: fe80::200:ff:fe01:1b01/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:9400  Metric:1
```

- Issue a CLI to gain access to Linux Bash Shell
- Leverage favorite Linux commands like ps, grep etc. available and could be used for further monitoring and scripting
- Role-based access to Bash



BCM Shell Access

- Issue a CLI to get shell access to underlying BCM chips
- Direct read/write access to hardware tables
- Can Peek/Poke underlying registers
- Python wrapper to get BCM Shell output

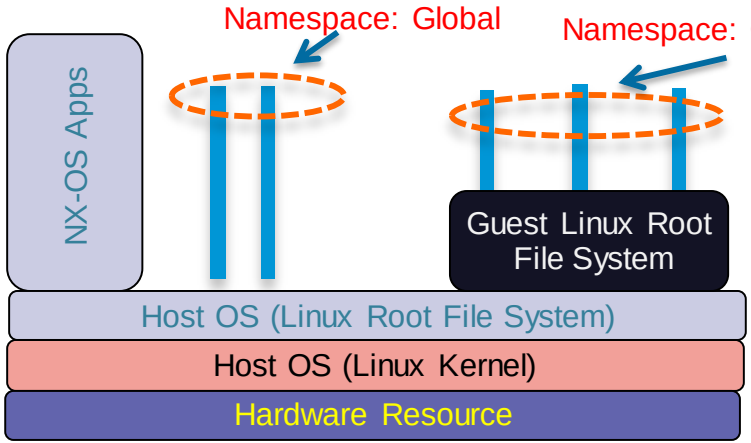

```
TME-1-9508-1# bcm-shell module 1
Warning: BCM shell access should be used with caution
Entering bcm shell on module 1
Available Unit Numbers: 0 1 2
bcm-shell.0> ^[[A^[[A^[[B

bcm-shell.0> l3 l3table show
l3 l3table show
Unit 0, free L3 table entries: 212960
Entry VRF IP address       Mac Address            INTF MOD PORT    CLASS HIT
147488 1    192.168.1.2       00:00:00:00:00:00  100006    0    0          0 y
149300 1    30.1.1.255        00:00:00:00:00:00  149150    0    0          0 n
150696 1    30.1.1.1          00:00:00:00:00:00  100012    0    0          0 n
152696 1    10.1.1.3          00:00:00:00:00:00  100007    0    0          0 y
154860 1    192.168.1.15      00:00:00:00:00:00  149150    0    0          0 n
156336 1    192.168.1.0       00:00:00:00:00:00  100000    0    0          0 n      (LOCAL ROUTE)
163452 1    192.168.1.13      00:00:00:00:00:00  149151    0    0          0 y      (LOCAL ROUTE)
165120 1    192.168.1.3       00:00:00:00:00:00  149150    0    0          0 n
166280 1    30.1.1.0          00:00:00:00:00:00  100000    0    0          0 n      (LOCAL ROUTE)
168280 1    10.1.1.2          00:00:00:00:00:00  100006    0    0          0 y
170444 1    192.168.1.14      00:00:00:00:00:00  100010    0    0          0 y
173968 1    192.168.1.1       00:00:00:00:00:00  149151    0    0          0 y      (LOCAL ROUTE)
174872 1    30.1.1.2          00:00:00:00:00:00  149151    0    0          0 n      (LOCAL ROUTE)
179036 1    192.168.1.12      00:00:00:00:00:00  100000    0    0          0 n      (LOCAL ROUTE)
183716 1    192.168.1.11      00:00:00:00:00:00  149150    0    0          0 n
184680 1    192.168.1.6       00:00:00:00:00:00  100007    0    0          0 y
186876 1    10.1.1.10         00:00:00:00:00:00  149151    0    0          0 n      (LOCAL ROUTE)
192308 1    192.168.1.9       00:00:00:00:00:00  149151    0    0          0 n      (LOCAL ROUTE)
193012 1    10.10.10.10       00:00:00:00:00:00  149151    0    0          0 y      (LOCAL ROUTE)
193528 1    192.168.1.4       00:00:00:00:00:00  100000    0    0          0 n      (LOCAL ROUTE)
201800 1    192.168.1.7       00:00:00:00:00:00  149150    0    0          0 n
```

Linux Containers

- Enable Nexus switches to host customer applications using LXC virtualisation technologies
- Customers can use ‘Application Development Toolkit’ to compile & package their applications into Cisco OVAs for deployment on Nexus switches

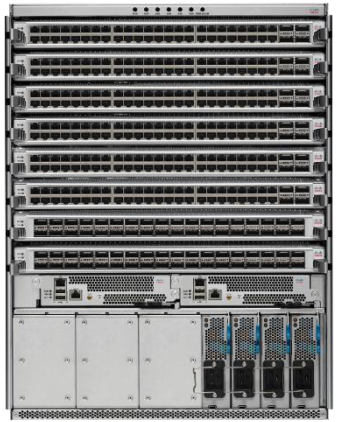
LXC

This is an operating system virtualisation technology that shares the host kernel with the guest but provides isolation through namespace extensions to the Linux kernel.
http://linuxcontainers.org/



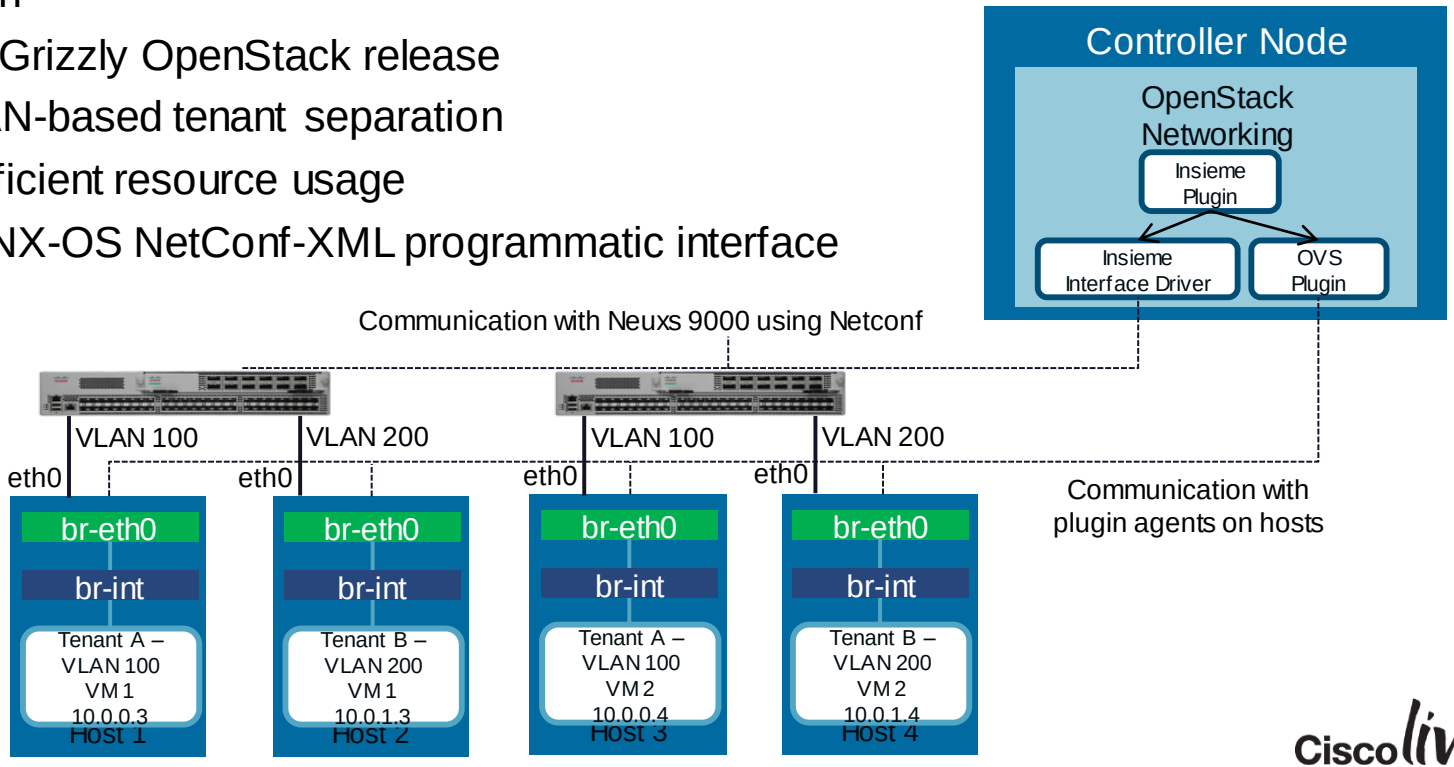
LinuX Containers (LXC) on Nexus 9000/3000

- Provides a secure and segregated operating environment for applications
- Can run either Cisco or Open Source applications
- Can use standard Linux distros
- OS Level Virtualisation
- Shared Kernel
- Shared physical resources
- Isolation through name spaces



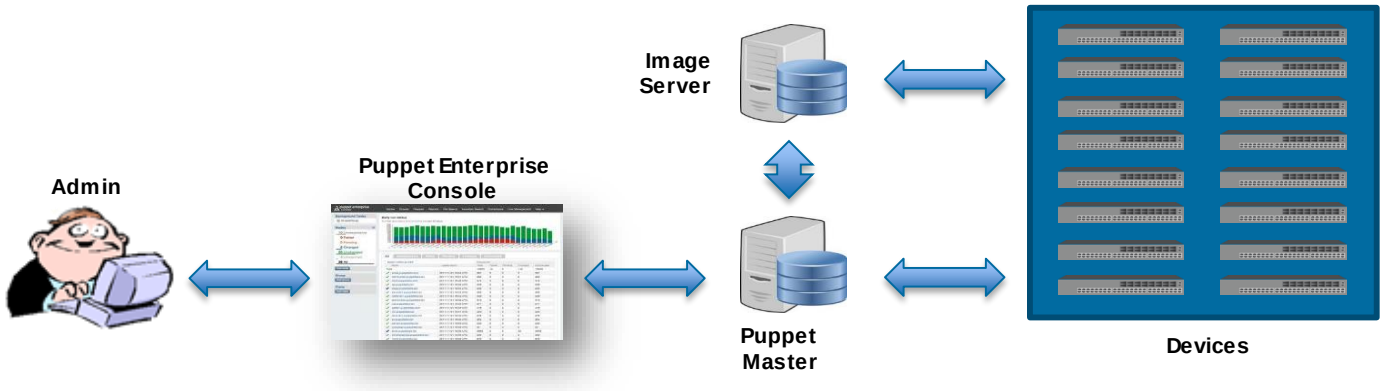
OpenStack Network (Neutron) Plugin

- Enables fully automated compute, storage and network resource orchestration
- Support for Grizzly OpenStack release
- Enable VLAN-based tenant separation
- Enhance efficient resource usage
- Leverages NX-OS NetConf-XML programmatic interface



NX-OS Image Patching

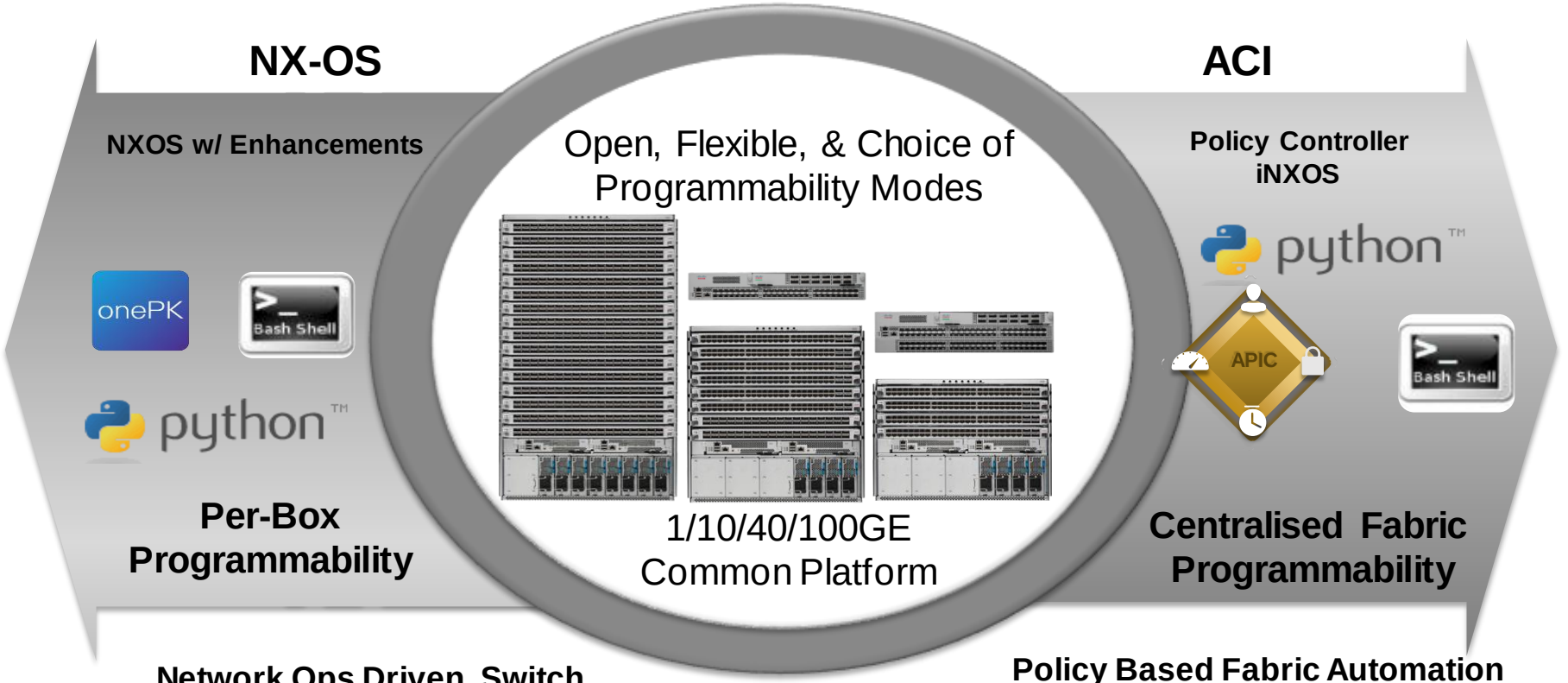
- Upgrade service executable or library in a NX-OS image
- Version and Compatibility control
- Allows Reverting a Patch
- Integration with server management tools like Puppet/Chef



Agenda
Nexus 9000 Architecture

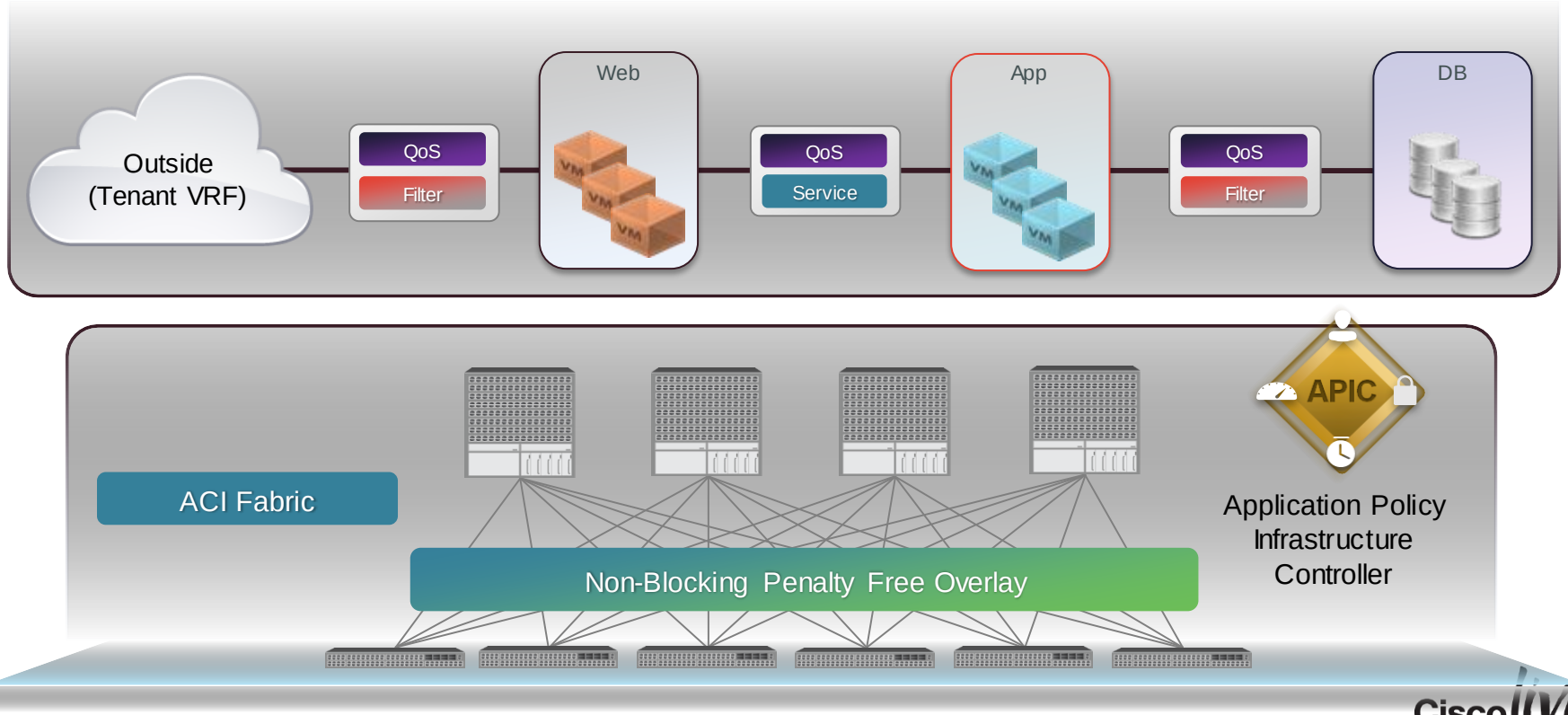
- Nexus 9000
  - Nexus 9000 Hardware
  - Nexus 9500 Chassis
  - Nexus 9500 Line Cards
  - Nexus 9500 Packet Forwarding
  - Nexus 9300
- Nexus 9000 and 40G
- Nexus 9000 Designs: FEX, vPC & VXLAN
- Nexus 9000 & Dev-Ops
- ACI & Nexus 9000

Common Platform: Two Modes of Operation



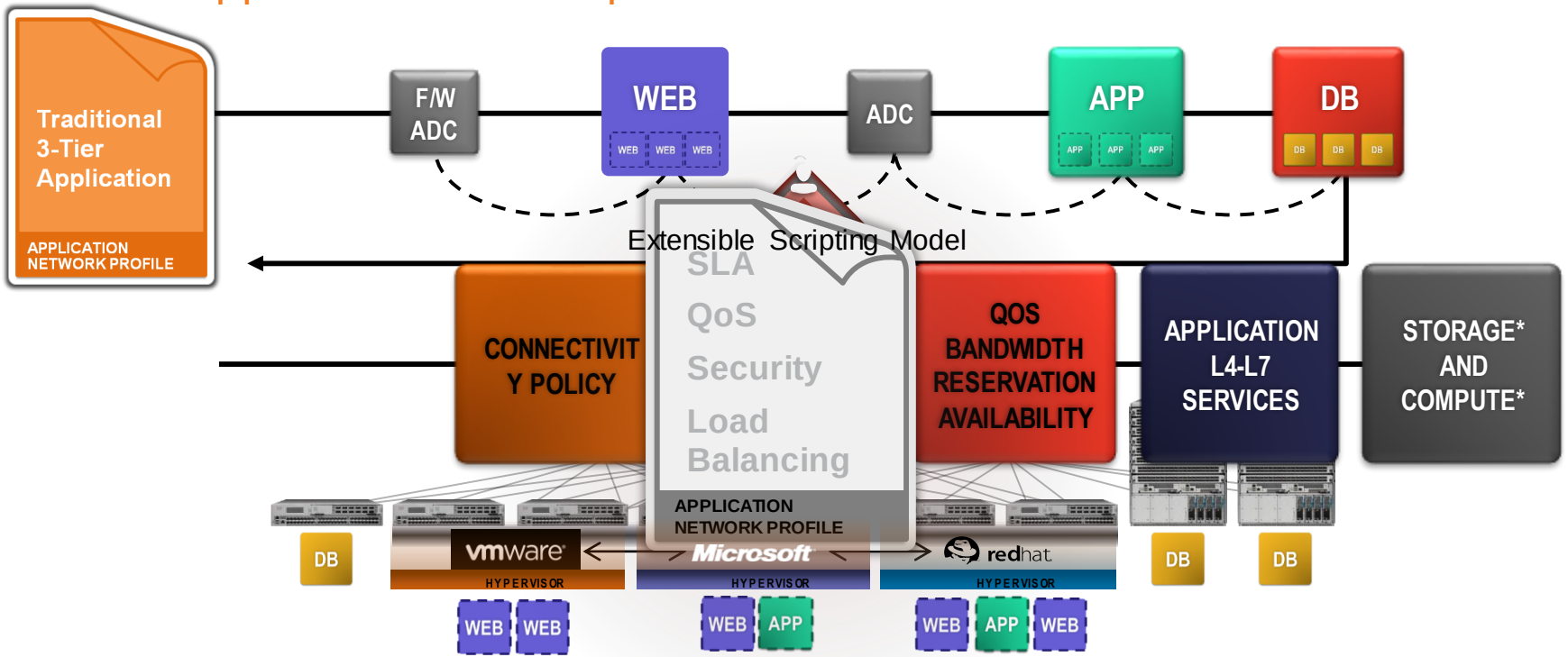
Cisco ACI

Logical Network Provisioning of Stateless Hardware

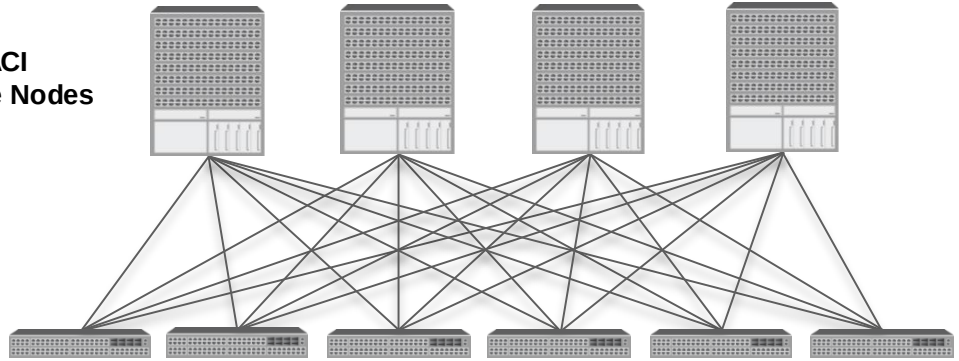


AGILITY: Any application, anywhere - Physical and virtual

Common application network profile



Cisco ACI Fabric



Application Infrastructure Policy Controller

- Decoupling of end point identity, location and associated policy all of which are independent from the underlying forwarding graph
- Full normalisation of the ingress encapsulation mechanism used, 802.1Q VLAN, IETF VXLAN, IETF NVGRE
- Distributed Layer 3 gateway ensures optimal forwarding for layer 3 and layer 2
- Support for standard bridging and routing semantics without standard location constraints (any IP address anywhere)
- Service insertion and redirection
- Removal of flooding requirements for IP control plane (ARP/GARP)

Community Code Development

- Visit us on GitHub: https://github.com/datacentre/nexus9000
- ACI and NX-OS code examples and libraries
- Open source and community developed tools by partners and 3rd party developers


Q & A

Complete Your Online Session Evaluation

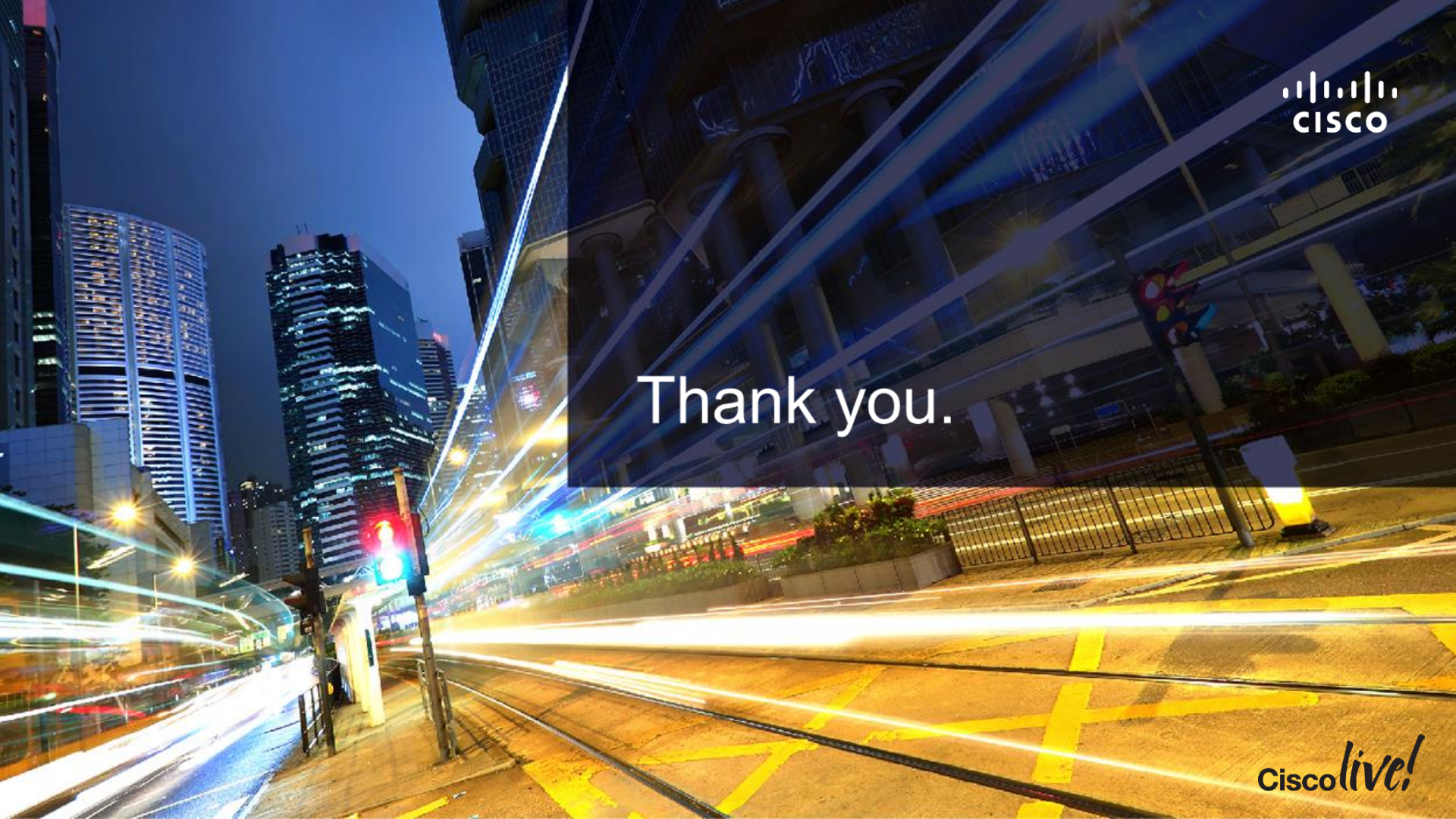
Give us your feedback and receive a Cisco Live 2015 T-Shirt!

Complete your Overall Event Survey and 5 Session Evaluations.

- Directly from your mobile device on the Cisco Live Mobile App
- By visiting the Cisco Live Mobile Site http://showcase.genie-connect.com/clmelbourne2015
- Visit any Cisco Live Internet Station located throughout the venue

T-Shirts can be collected in the World of Solutions on Friday 20 March 12:00pm - 2:00pm

Learn online with Cisco Live!
Visit us online after the conference for full access to session videos and presentations. www.CiscoLiveAPAC.com


Thank you.


CISCO