
TOMORROW starts here.
cisco
Cisco live!

Nexus 9000 Architecture

BRKDCT-3640

Mike Herbert

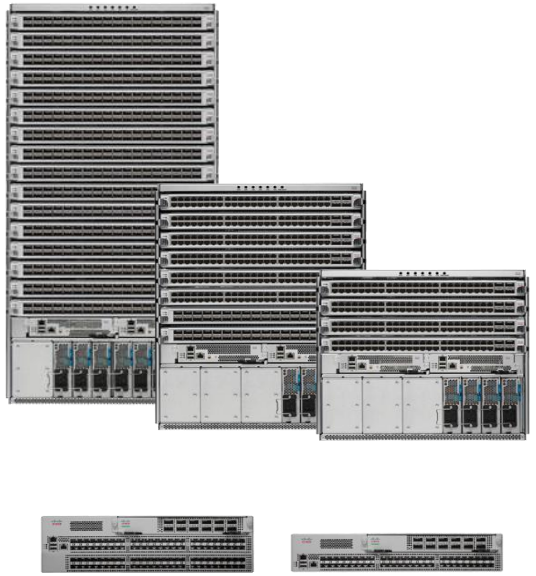
Principal Engineer INSBU

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 Gbps/40 Gbps/100 Gbps Switch Family

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

Nexus® 9300

FCS Q4 2013 — 48 1/10G SFP+ & 12 QSFP+

FCS Q1 2014 — 96 1/10G-T & 8 QSFP+

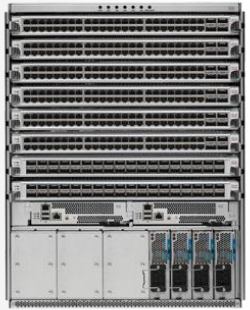
FCS Q1 2014 — 12-port QSFP+ GEM

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
- 14 W/40 Gbps port

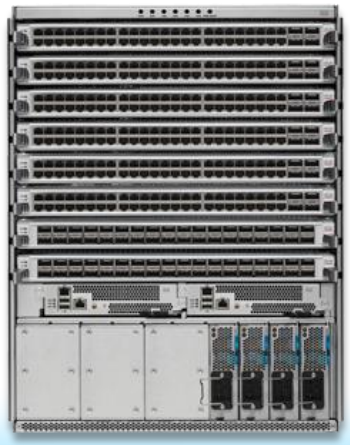
First modular chassis without a mid-plane

- 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



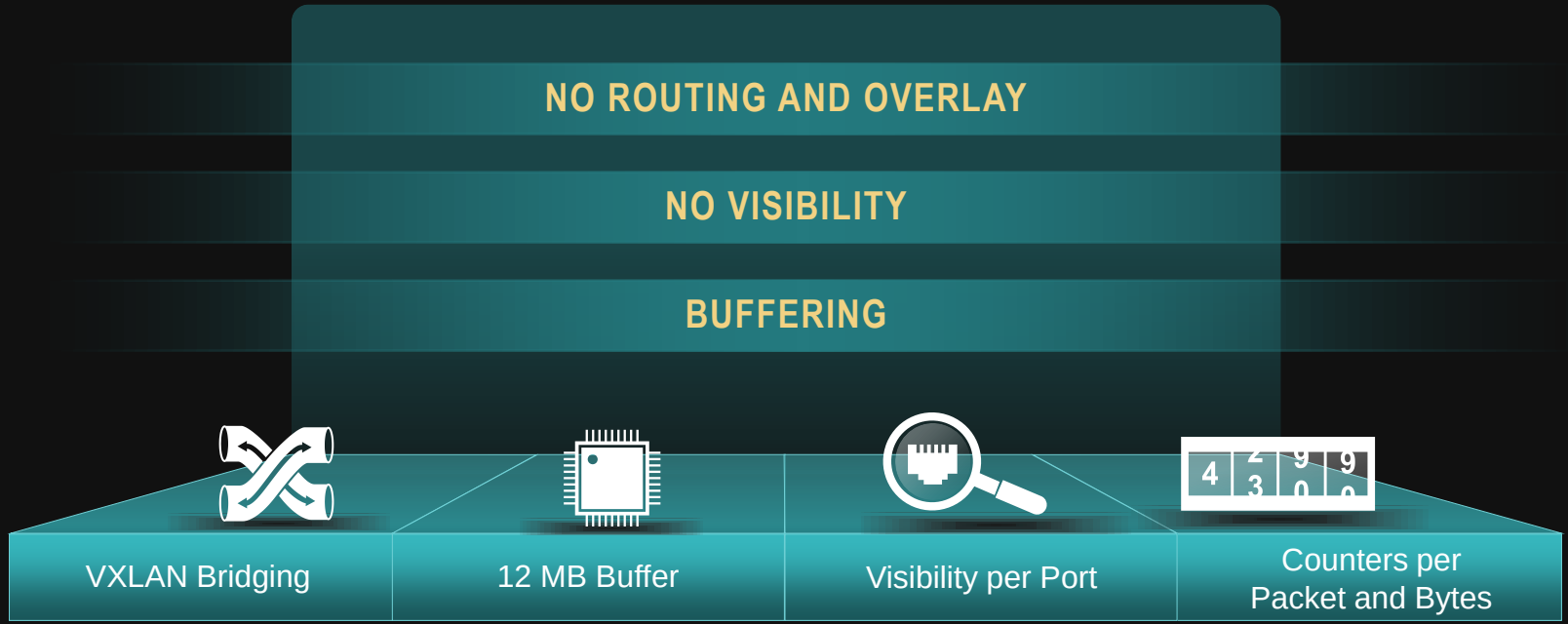
Merchant and Custom ASICs (Merchant+)

- ✧ Merchant+ Strategy BROADCOM + CISCO
- ✧ 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 Trident II)
- Custom ASIC --- Cisco ALE (ACI Leaf Engine), ASE (ACI Spine Engine)

MERCHANT SILICON ALONE LEAVES ROOM FOR IMPROVEMENT



MERCHANT +

CISCO ASIC INNOVATIONS

+

INDUSTRY BASELINE

Broadcom Trident 2

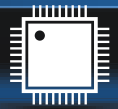
DYNAMIC LOAD BALANCING

FLOW PRIORITISATION

MULTICAST MULTIPATH

INLINE HW OVERLAY DB

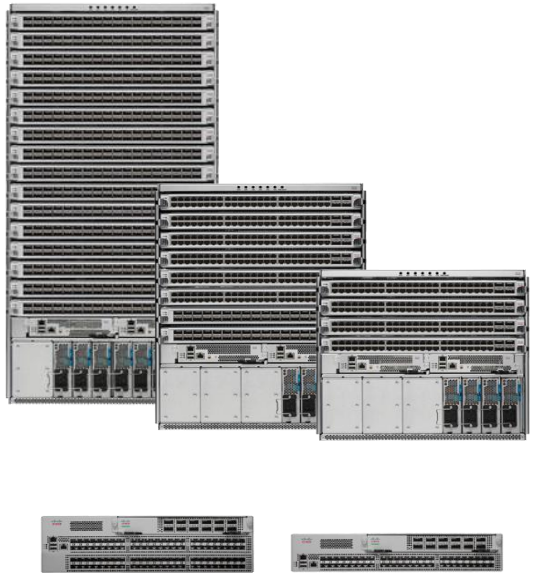
FAST RE-ROUTE



VXLAN Routing Normalised Forwarding VXLAN/NVGRE/VLAN	Incremental 40 MB Buffer, Active Queue Management	**APPLICATION VISIBILITY** Elephant Traps, VXLAN Aware Trace-route	Atomic End-end Latency Measurements
VXLAN Bridging	12 MB Buffer	Visibility per Port	Counters per Packet and Bytes

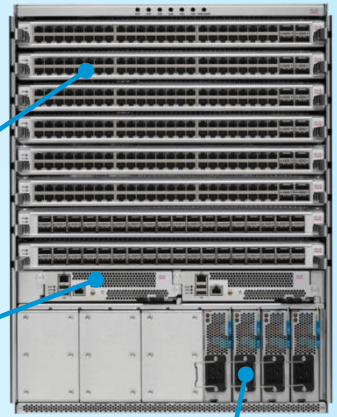
Agenda – Nexus 9000 Architecture

- 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
8-Slot Modular Chassis

Nexus® 9508 Front View

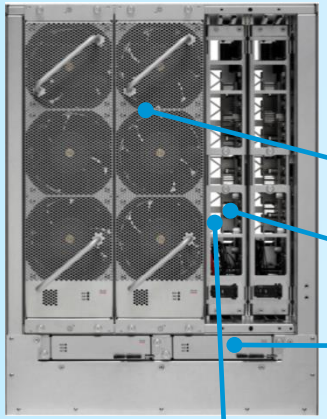


8 line card slots
Max 3.84 Tbps per slot
duplex

Redundant
supervisor engines

3000 W AC power supplies
2+0, 2+1, 2+2 redundancy
Supports up to 8 power supplies

Nexus 9508 Rear View



3 fan trays, front-to-back
airflow

3 or 6 fabric modules
(behind fan trays)

Redundant system controller
cards

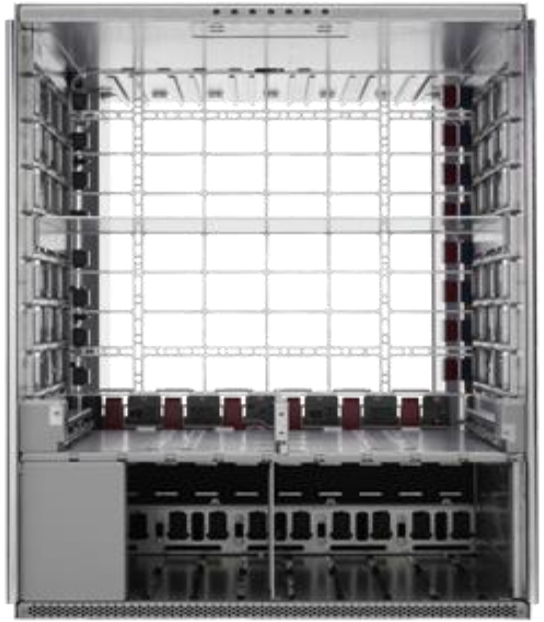
No mid-plane for
LC-to-FM connectivity

Chassis Dimensions: 13 RU x 30 in. x 17.5 in (HxWxD)

Designed for Power and Cooling Efficiency
Designed for Reliability
Designed for Future Scale

Chassis Design: No Mid-Plane

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

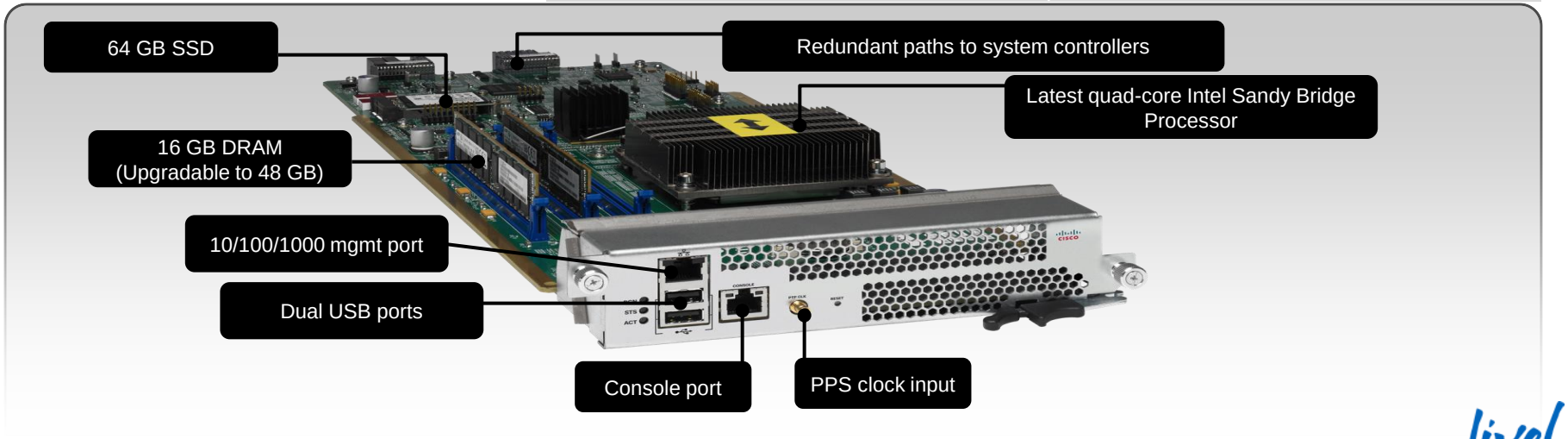


Nexus 9500 Platform Architecture

Supervisor Module

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

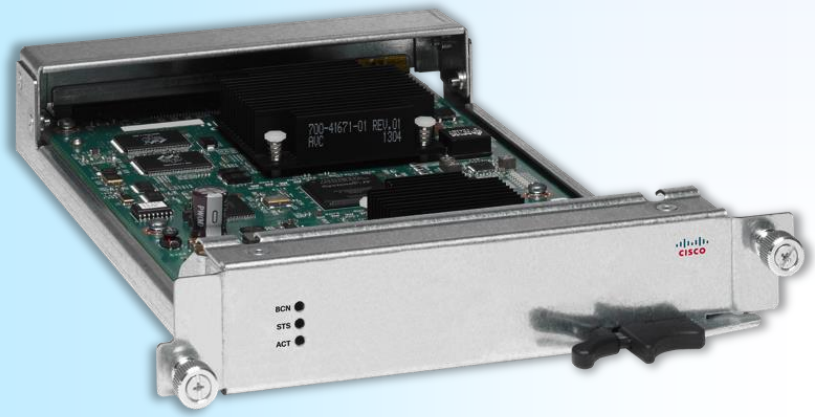
Supervisor Module	
Processor	Romley, 1.8 GHz, 4 core
System Memory	16 GB, upgradable to 48 GB
RS-232 Serial Ports	One (RJ-45)
10/100/1000 Management Ports	One (RJ-45)
USB 2.0 Interface	Two
SSD Storage	64 GB



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



Power Supplies

3000W AC PSU

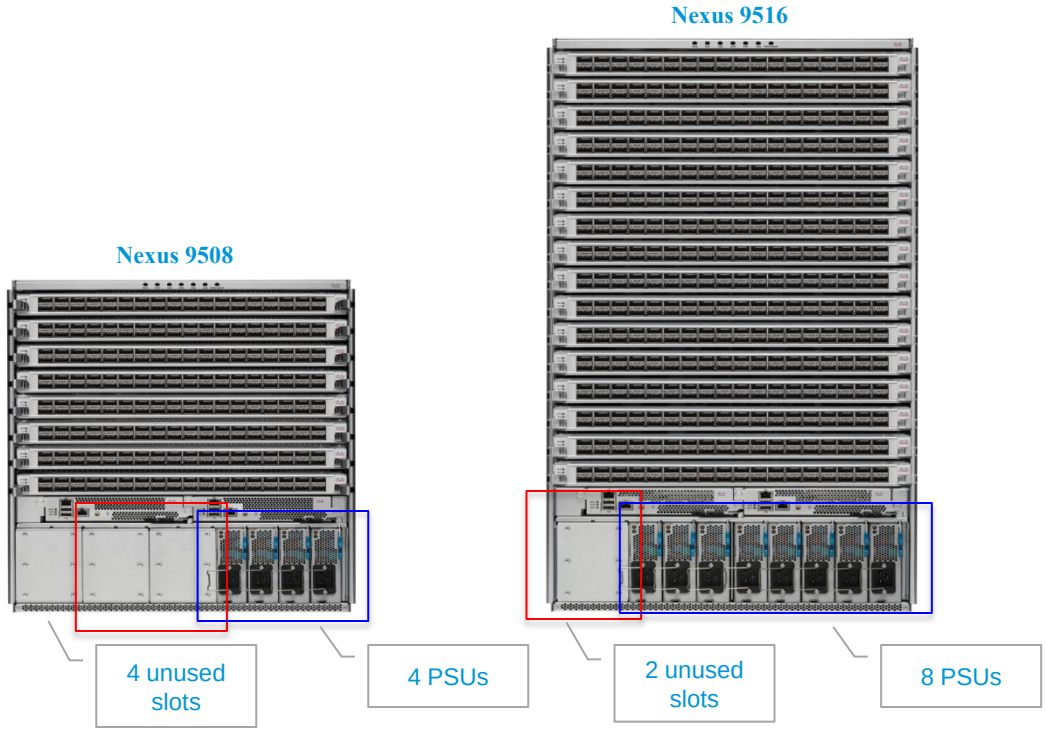
- Single 20A input – 220V
- Support for range of international cabling options
- 92%+ Efficiency
- Range of PS configurations
  - Minimum 1 PS
  - 2 PS for fully loaded chassis
  - N+1 redundancy
  - N+N grid redundancy
- 2x head room for future port densities, bandwidth, and optics
  - Up to 8 PS total

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

Power Headroom on N9508 and N9516

- Common PSU for all chassis types for all Nexus 9500 chassis, 4-slot, 8-slot and 16-slot.
- Power Headroom for future growth

Both Nexus 9508 and 9516 have spare PSU slots for power expansion to support future I/O modules that may requires more power.



Nexus 9500: Power Efficient 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	0%
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:

Fabric Modules and Fan Trays

- 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

Fan Tray

Fan trays are installed after the Fabric Module.

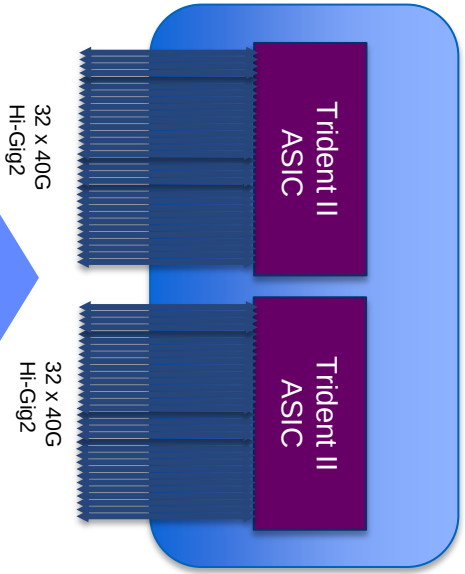
To service a FM, the fan tray must be removed first.

1) If multiple fan trays are removed, the switch shuts down after 2 minutes.

1) As soon as one fan is removed, other fans increase speed to 100% to prevent overheating.

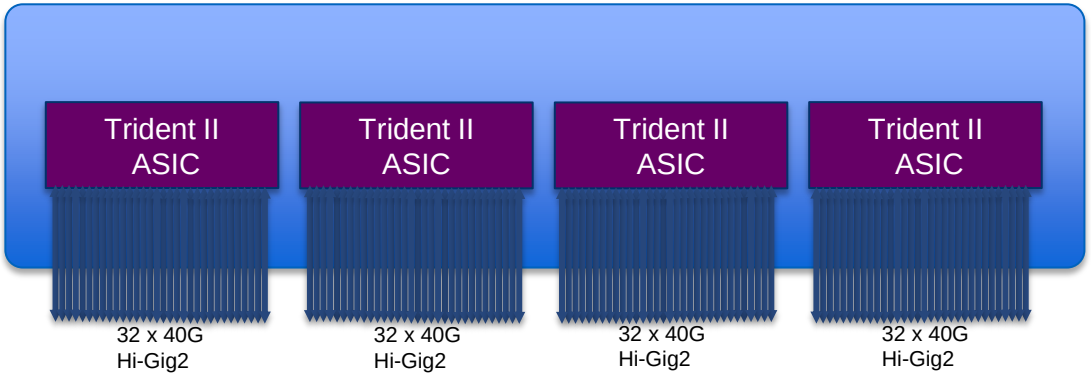
Fabric Modules

Each Fabric Module contains two Broadcom Trident II ASICs (Network Forwarding Engines)



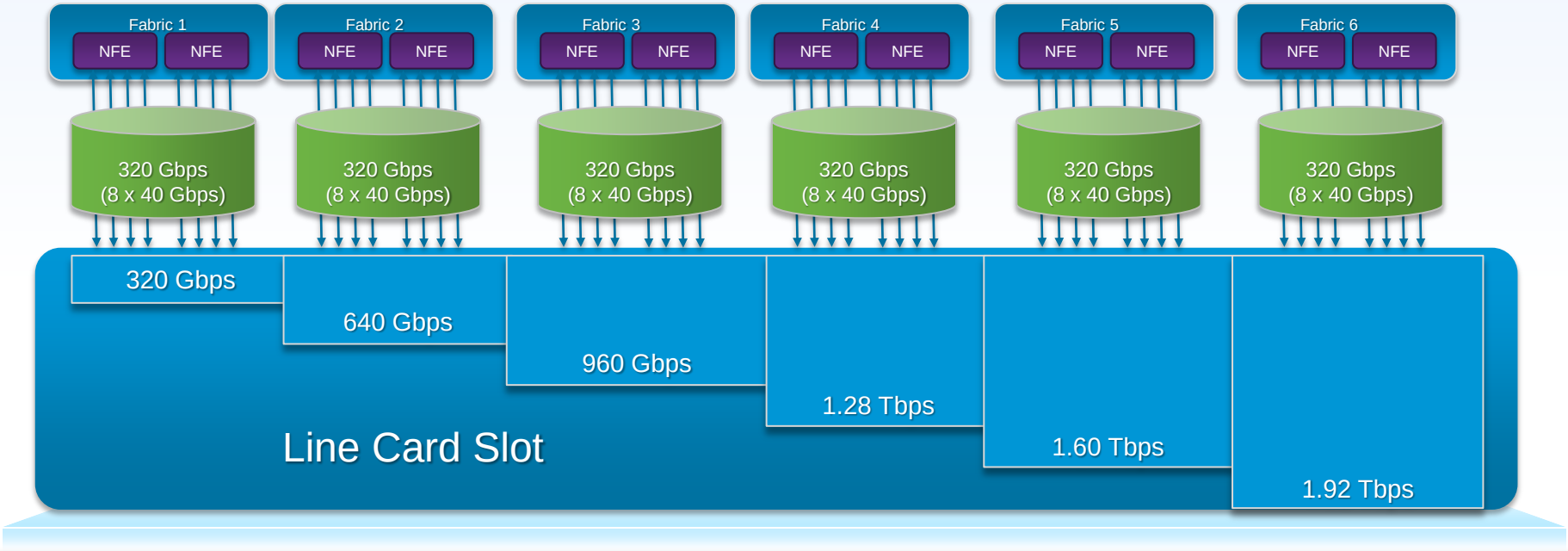
16 slots Fabric

Note that 16 slot Fabric Module will have four Trident II (NFEs)



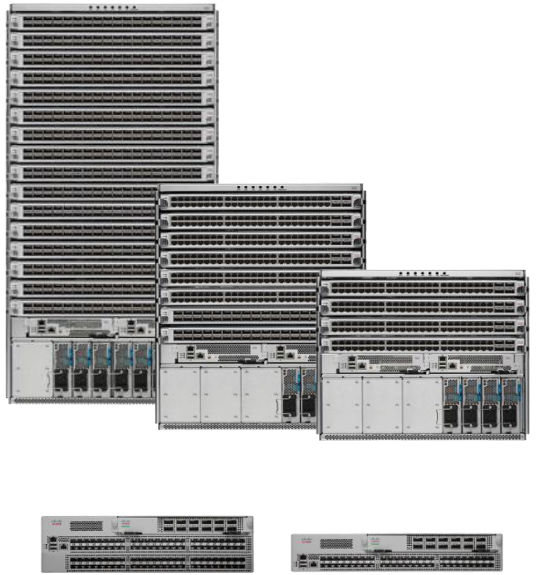
Nexus 9500 Fabric Module
Data Plane Scaling For 8-Slot Chassis

- Each fabric module for the 8-slot chassis 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

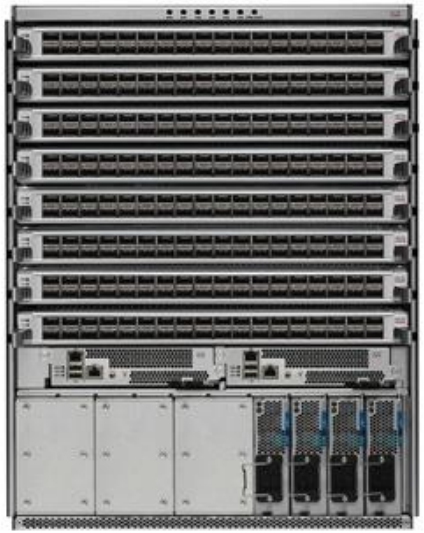


Agenda – Nexus 9000 Architecture

- 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 Line Cards



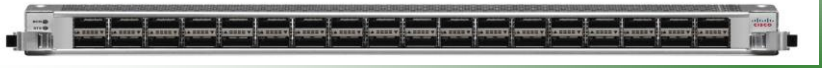
40G Aggregation

36 ports 40G QSFP+ (Non Blocking)

1/10G Access and 10/40G Aggregation

48 ports 10G SFP+ & 4 ports 40G QSFP+
48 ports 1/10G-T & 4 ports 40G QSFP+
(non blocking)

36 ports 40G QSFP+ ((1.5:1 oversubscribed)



ACI Access Ready

40G Fabric Spine

36 ports 40G QSFP+ (Non Blocking)

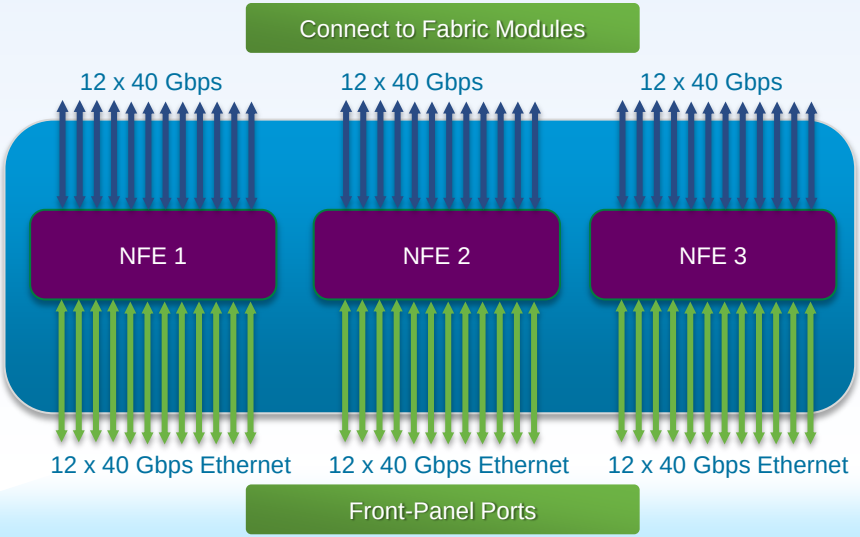
ACI Spine

Cisco Nexus 9500 Line Cards
N9K-X9636PQ

A 36-port, 40 Gbps QSFP+ line card needs 6 fabric modules to operate at 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

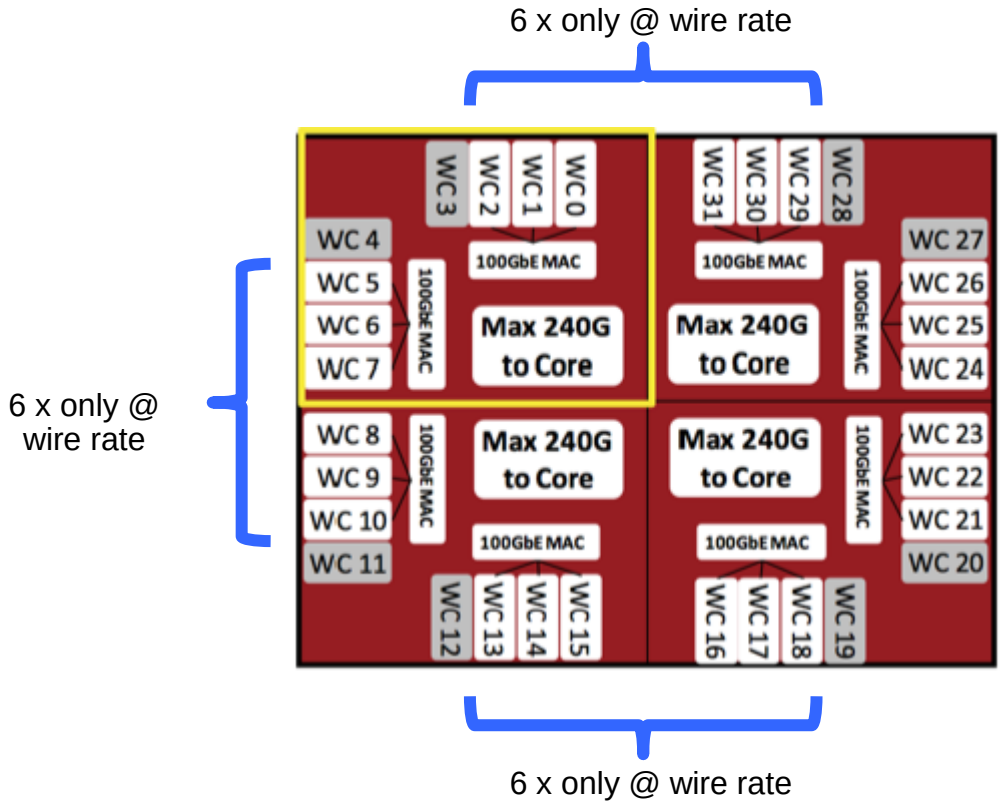
Nexus 9500 Line Card
N9K-X9636PQ



Each 36-port, 40 Gbps QSFP+ line card needs 6 fabric modules to operate at line rate on all 36 ports.

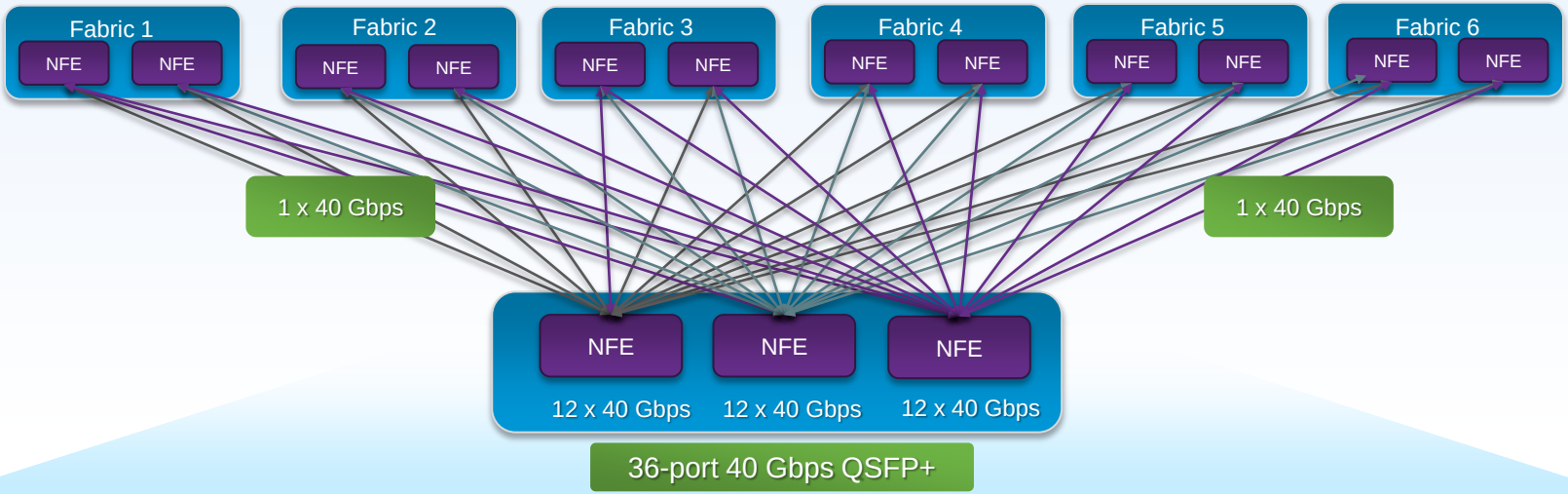
- Three network forwarding engines (NFE)
- Each NFE has 12 x 40 Gbps links to the front panel and 12 x 40 Gbps internal links to the fabric modules

Broadcom Trident 2 Internals



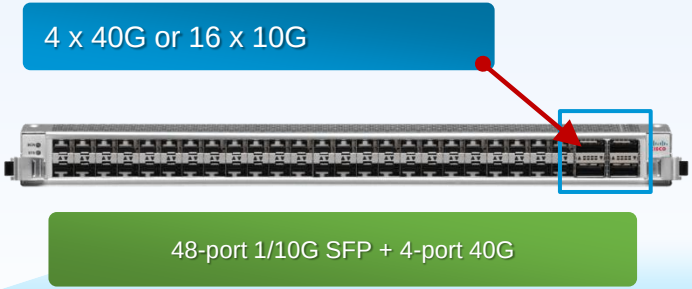
- In order to operate at line rate for all packet sizes only 24 of 32 ports are leveraged in the 9500 architecture
- 3 ASIC for 36 ports on the 36 x 40G line card
- 2 ASIC per Fabric Module – 240Gbps per line card (not an oversubscribed 320Gbps)

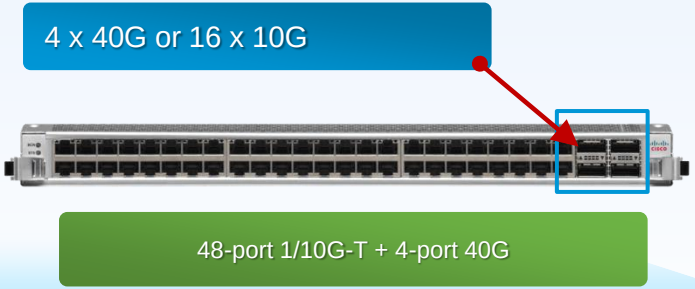
Nexus 9500 Line Cards
N9K-X9636PQ Fabric Connectivity



- All ports on the line card can operate at line rate for any packet sizes with 6 fabric modules
- Each of the 3 NFEs on the line card has 12 x 40 Gbps links to fabric modules - one to each Fabric Trident II ASIC

Nexus 9500 Line Cards
N9K-X9564PX And N9K-X9564TX



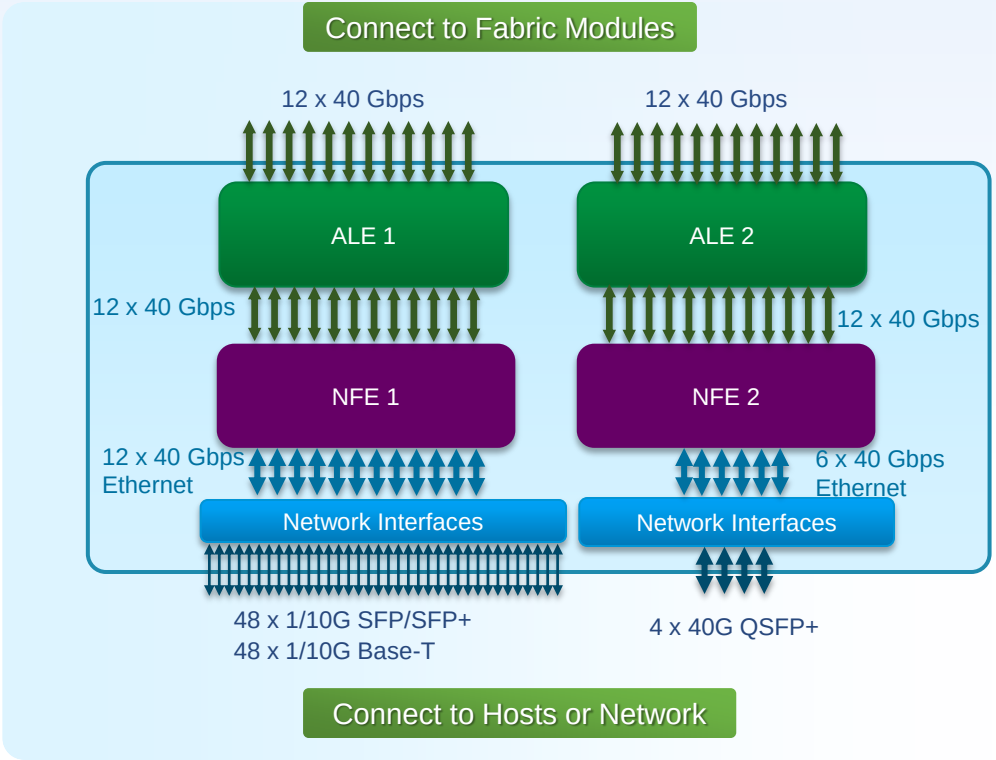


- N9K-X9564PX
  - 48 1/10G SFP+ ports + 4 40G QSFP+ ports
- N9K-X9564TX
  - 48 1/10GBase-T ports + 4 40G QSFP+ ports
- 1.92 Tbps duplex fabric connectivity
- Layer 2 and 3 line-rate performance on all ports for all packet sizes
- Cisco® NX-OS and Application Centric Infrastructure (ACI) mode

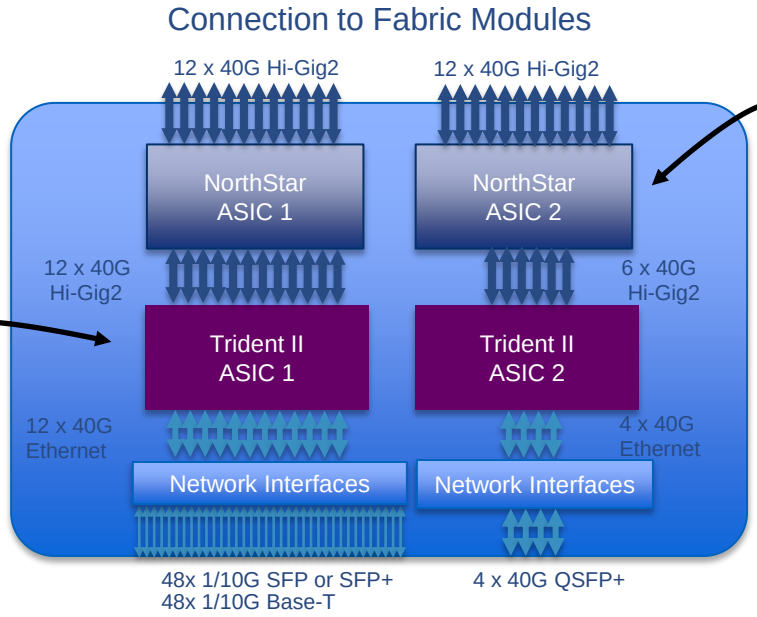
Nexus 9500 Line Cards
N9K-X9564PX And N9K-X9564TX

2 network forwarding engines (NFEs)

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



Details

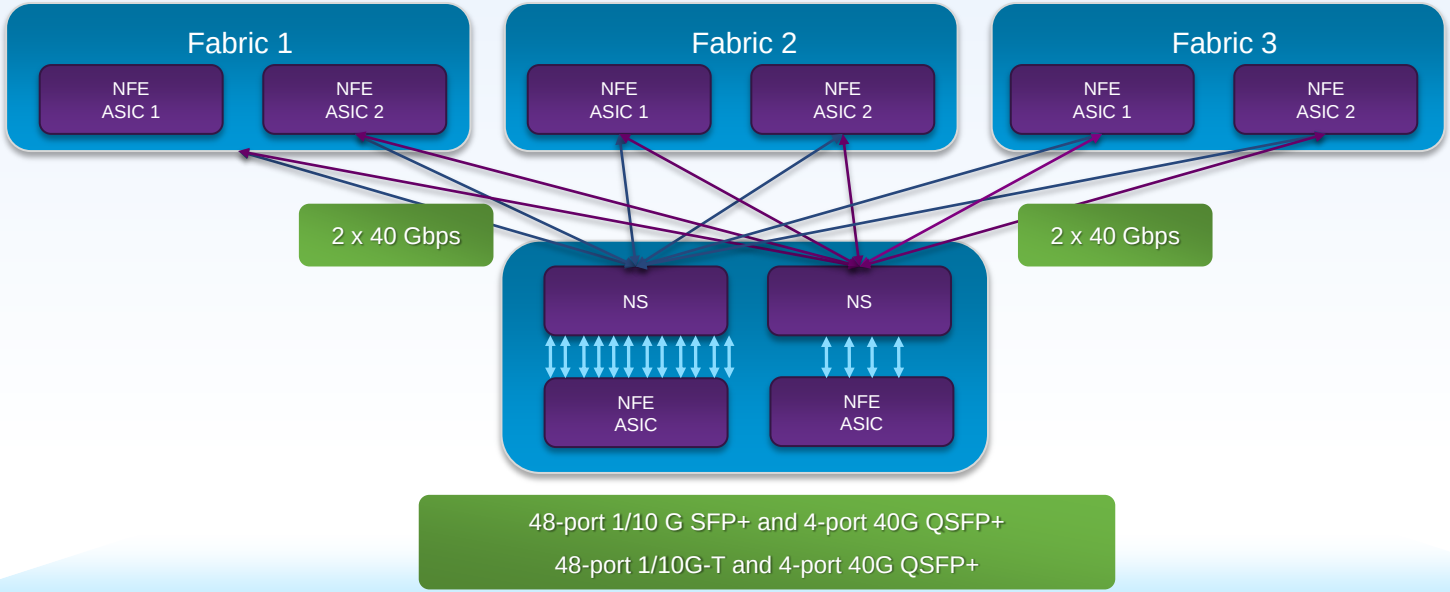


T2 ASICs act as main forwarding engines for standalone mode.

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

Standalone & Fabric Access LCs

Nexus 9500 Line Cards
N9K-X9564PX And N9K-X9564TX Fabric Connectivity



- Minimum of 3 fabric modules to get all line-rate ports

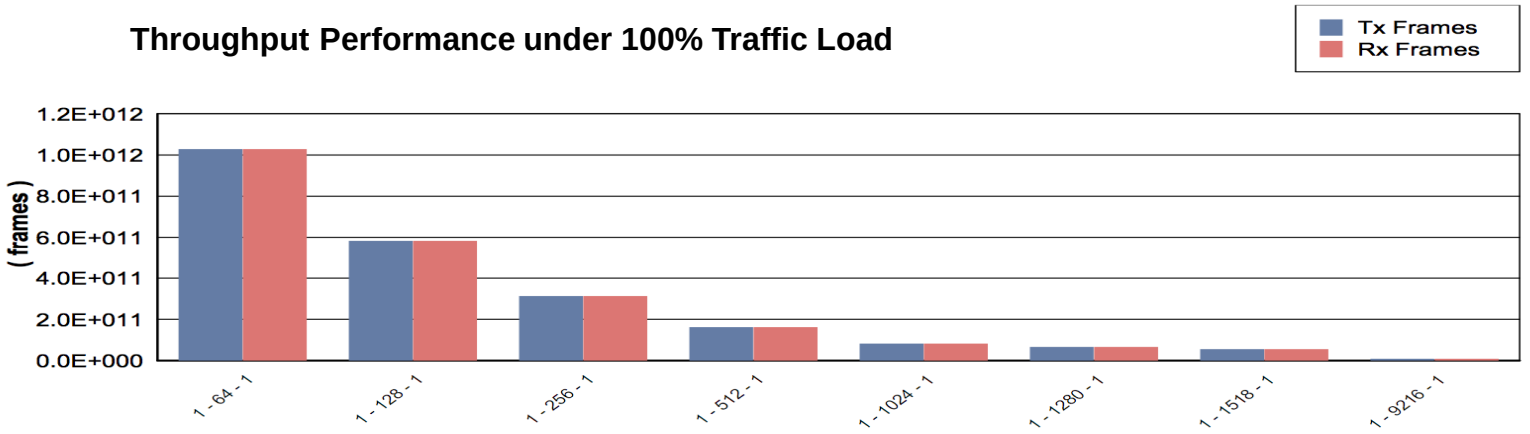
Nexus 9000 Series
Full Line Rate Throughput Performance

- **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)

Throughput Performance under 100% Traffic Load

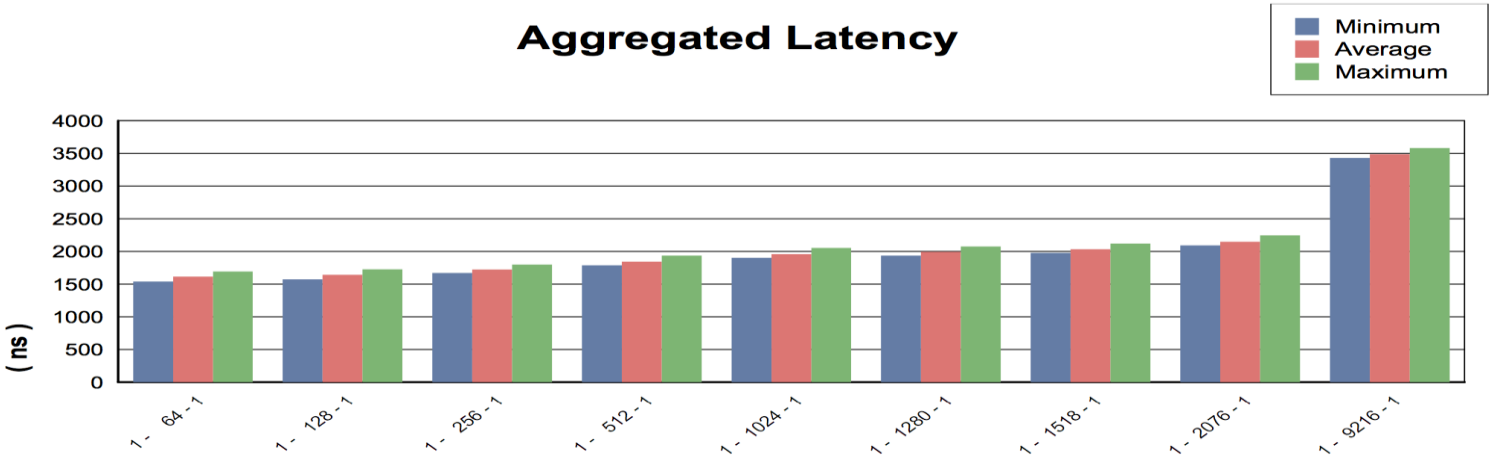


Nexus 9000 Series
Low Latency Performance

- **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:

- Unicast Latency at 100% traffic load:
  - 1.6 usec (64-Byte packets)
  - 3.5 usec (9216-Byte packets)
- Multicast latency at 100% traffic load:



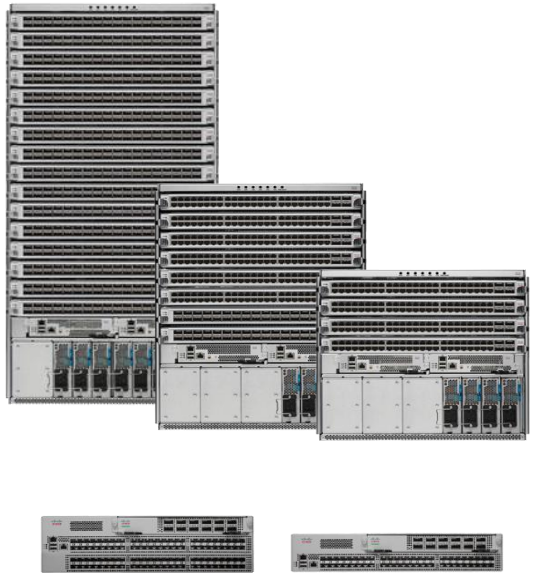
Nexus 9500 Series
Line Cards



	N9K-X9636PQ	N9K-X9564PX	N9K-X9564TX
1/10 Gb SFP/SFP+ Ports	--	48	--
1/10 Gb BaseT Ports	--	--	48
40 Gb QSFP Ports	36	4	4
Maximum Number of 1 Gb Ports	--	48	48
Maximum Number of 10 GE Ports	144	64	64
Maximum Number of 40 GE Ports	36	4	4
Minimum Number of Fabric Modules for Full Line-Rate Performance	6	3	3

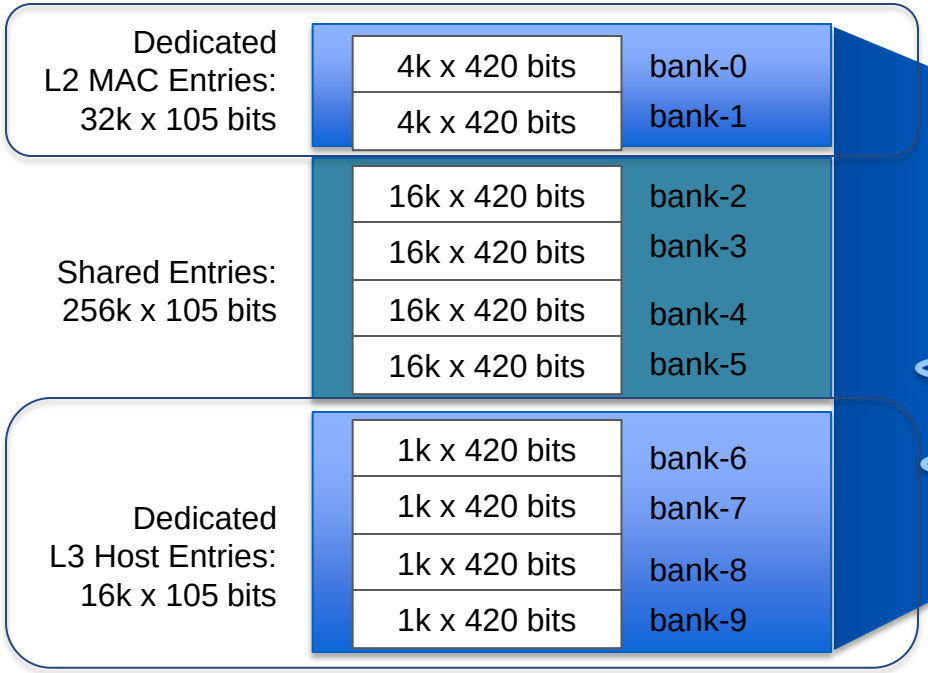
Agenda – Nexus 9000 Architecture

- 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



Trident 2 Unified Forwarding Table

T2 has a 16K traditional LPM TCAM. In addition to it, T2 has the following Unified Forwarding Table:

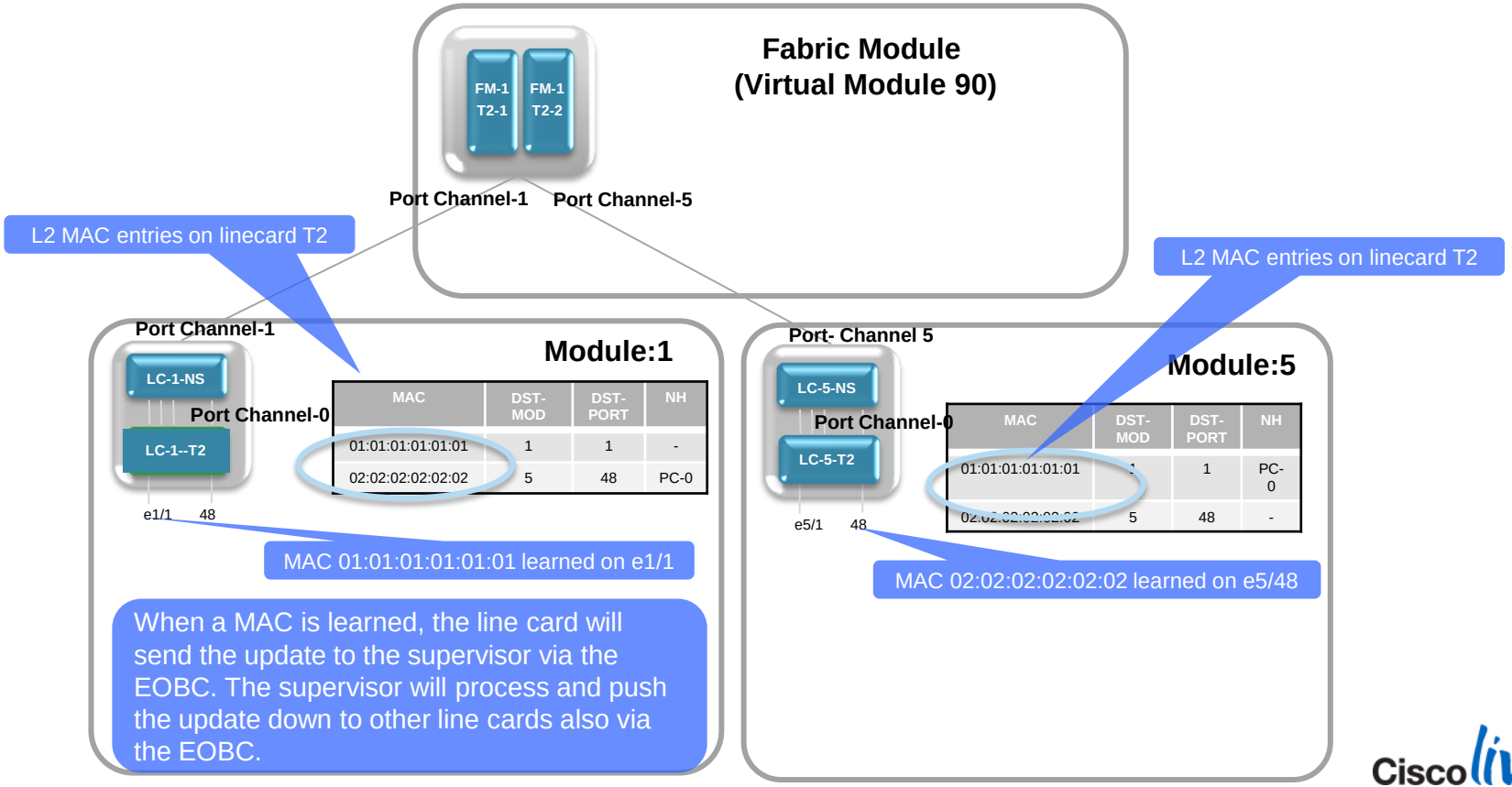


SUPPORTED COMBINATIONS

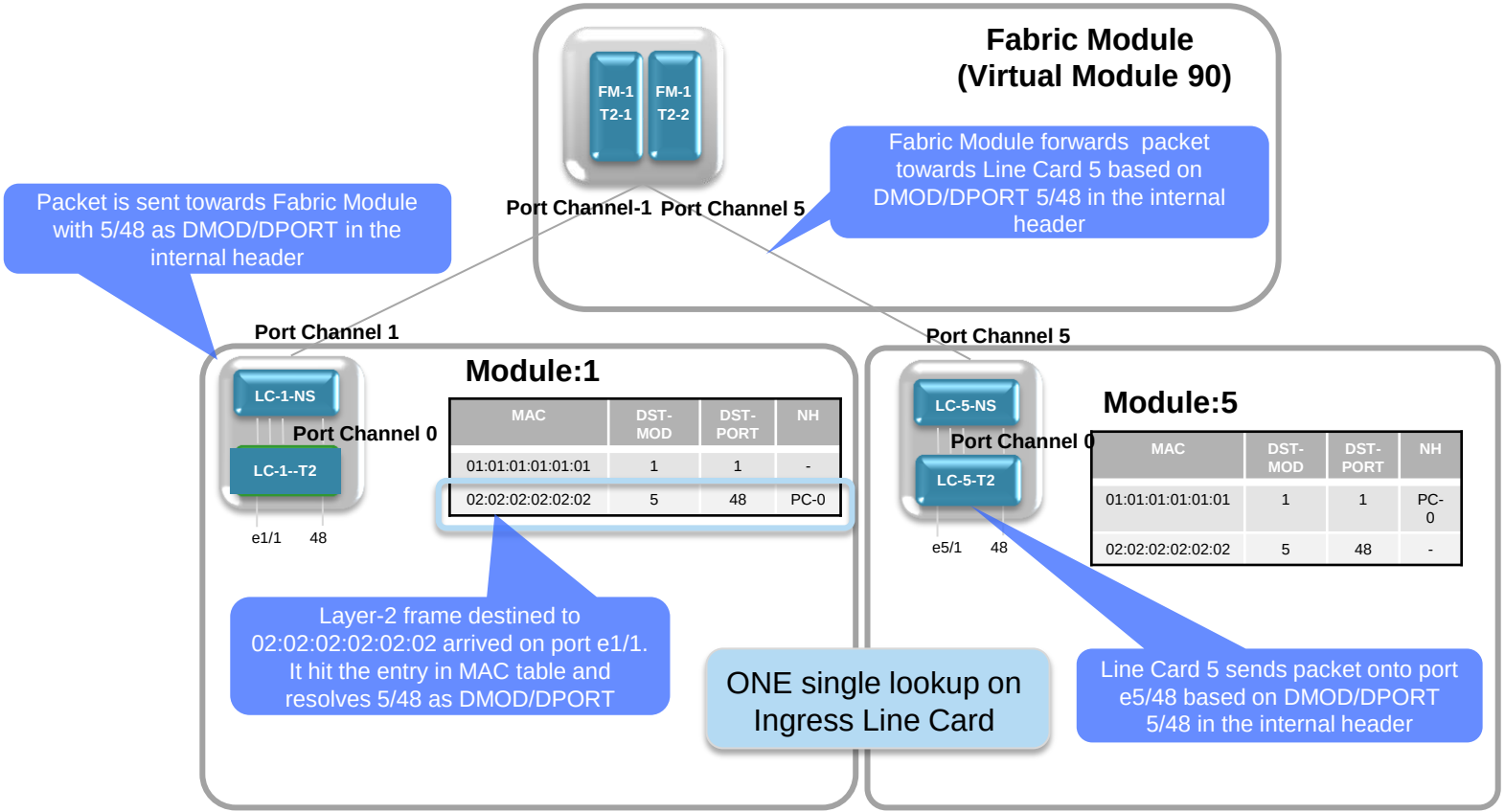
Mode	L2	L3 Hosts	LPM
0	288K	16K	0
1	224K	56K	0
2	160K	88K	0
3	96K	120K	0
4	32K	16K	128K

N9300 Trident 2 can be programmed in either mode 2 to support Layer-2 profile or mode 4 to support Layer-3 profile

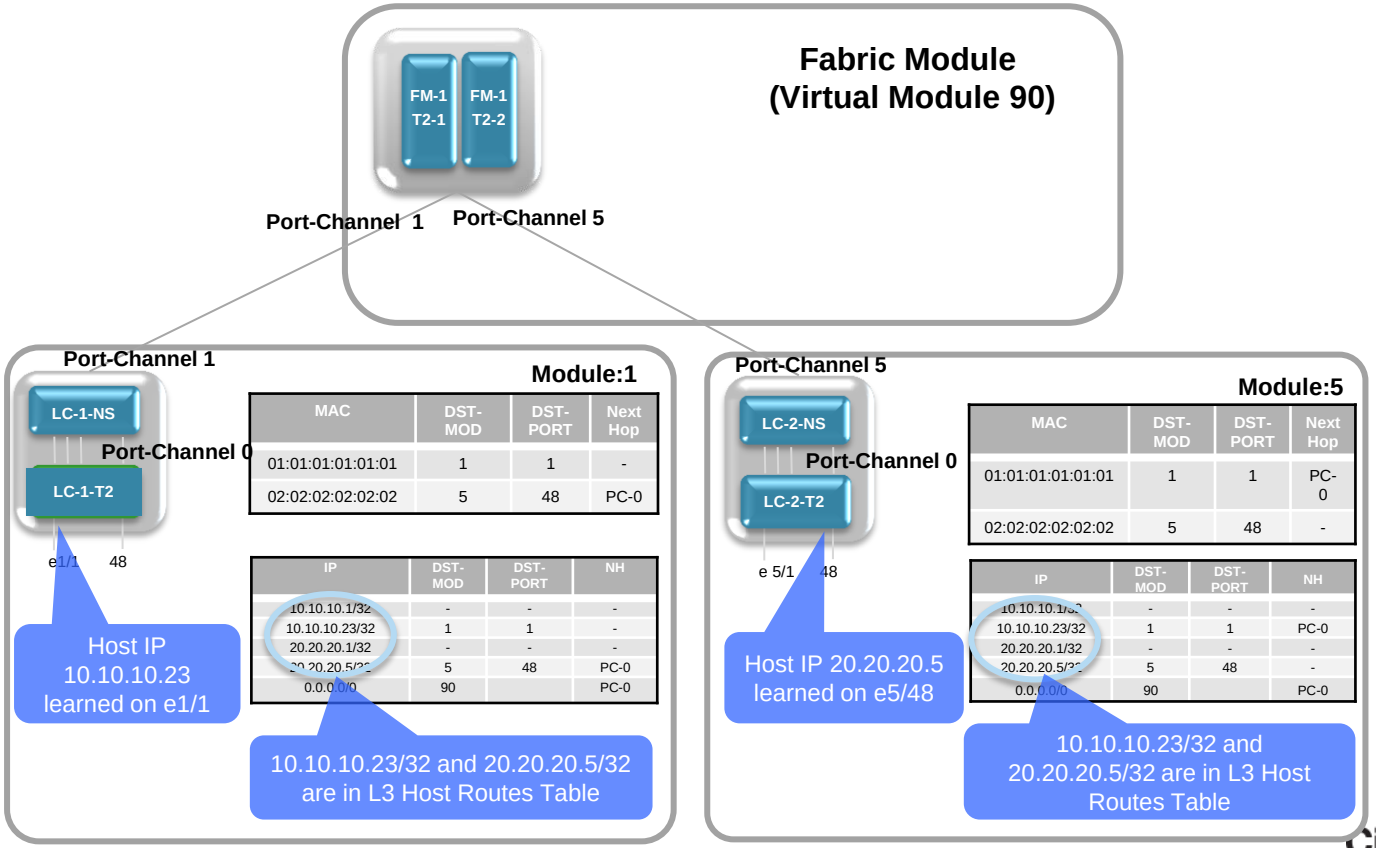
L2 Forwarding



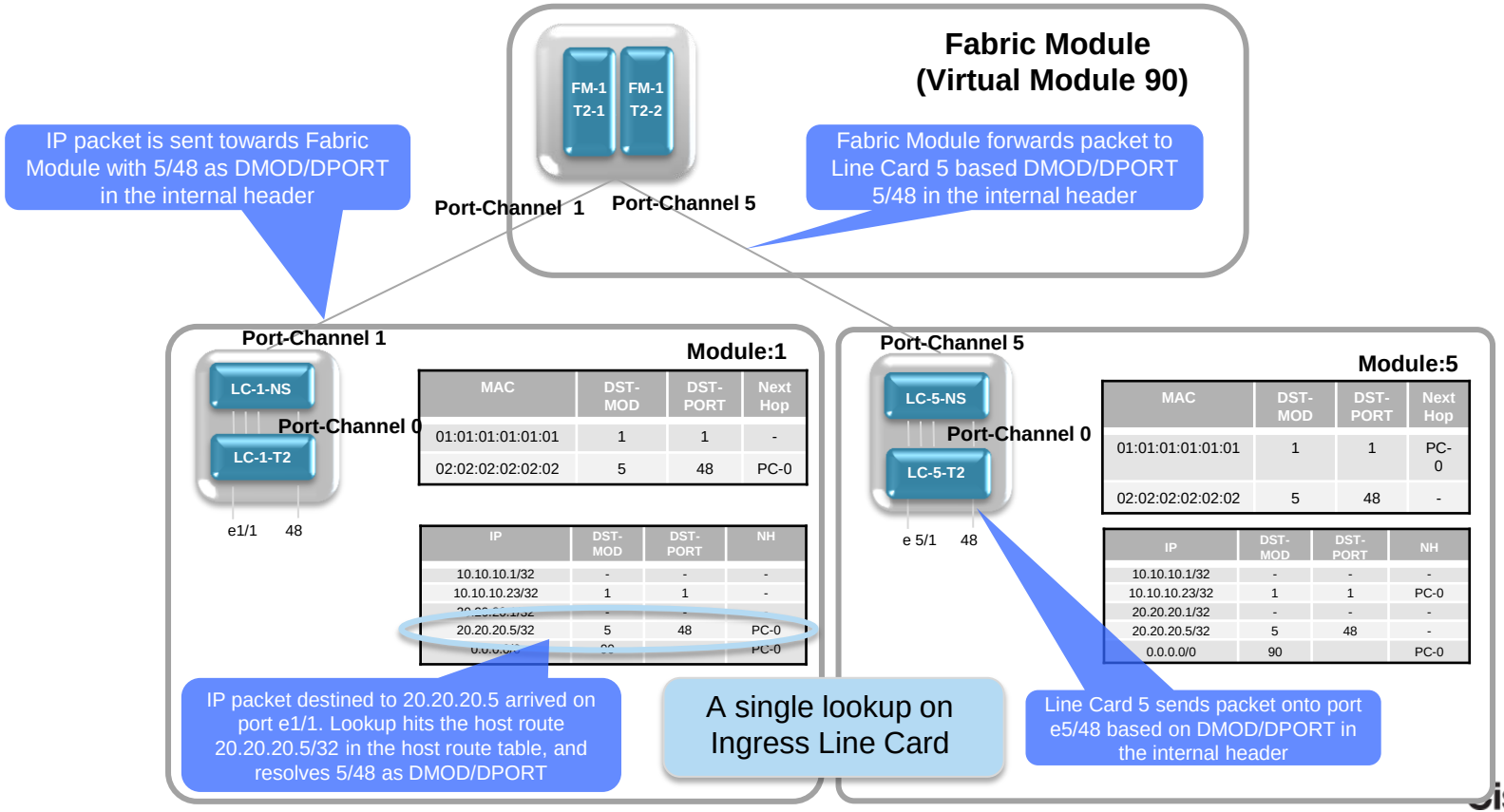
L2 Forwarding



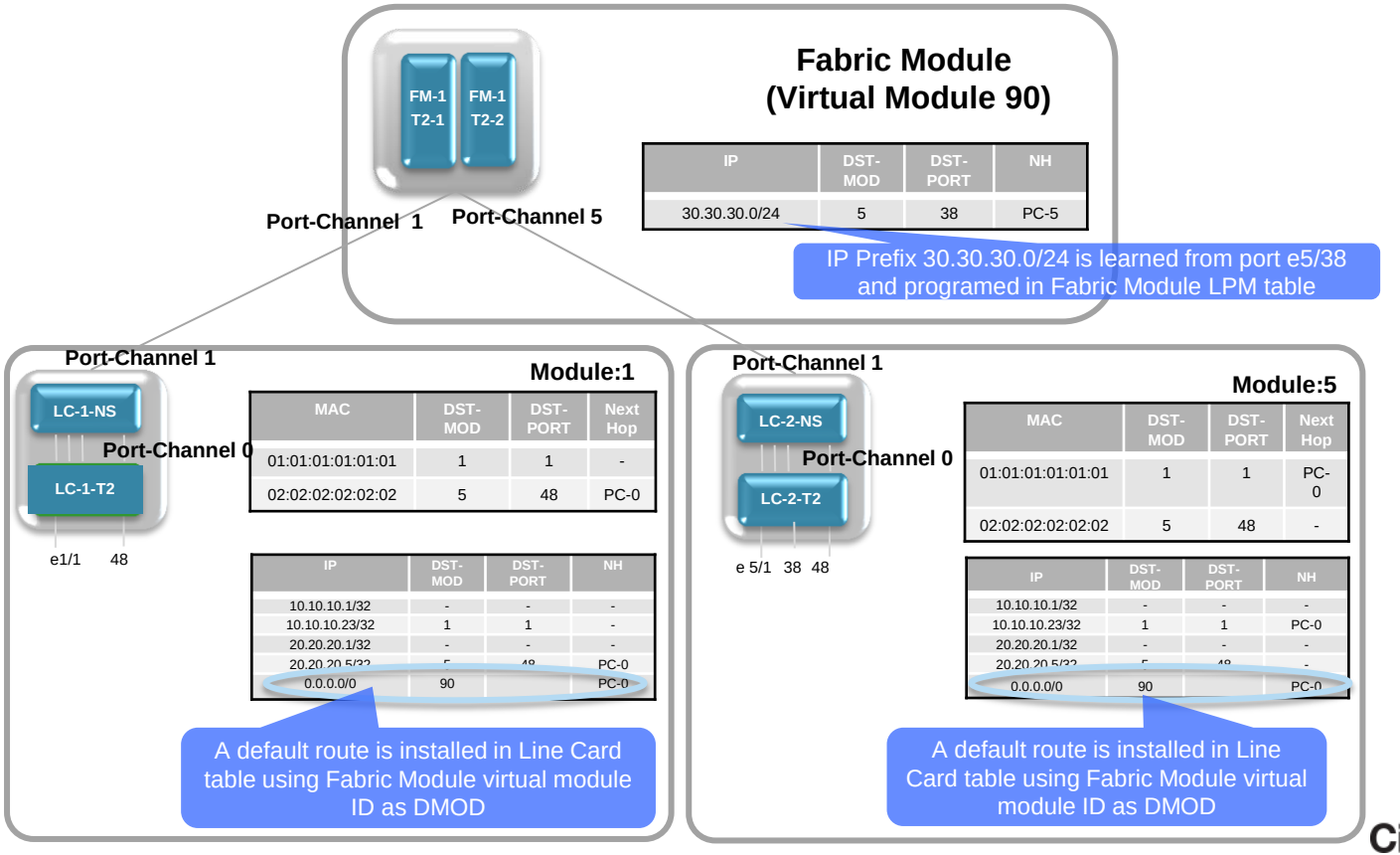
L3 Host Lookup



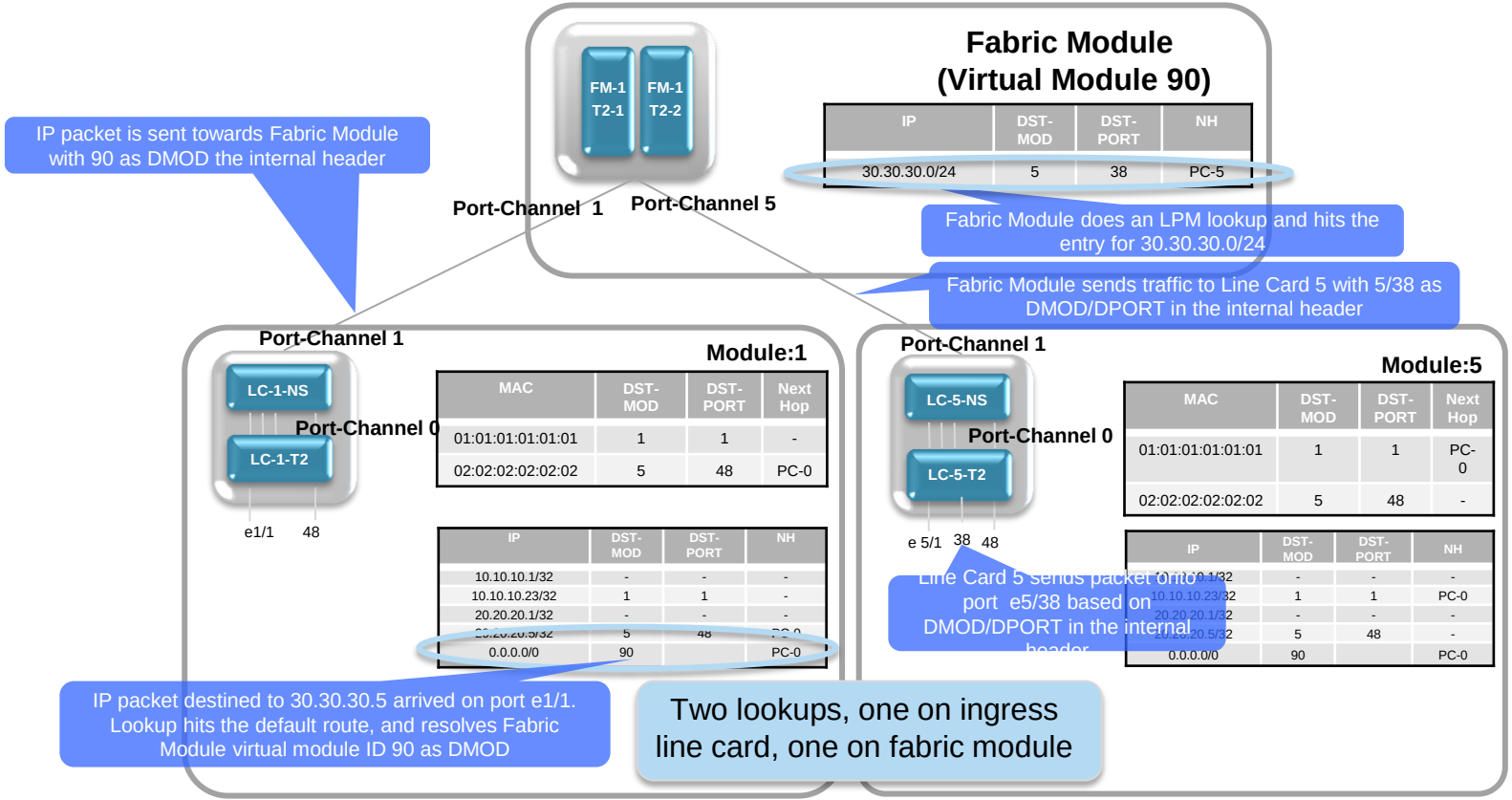
L3 Host Lookup



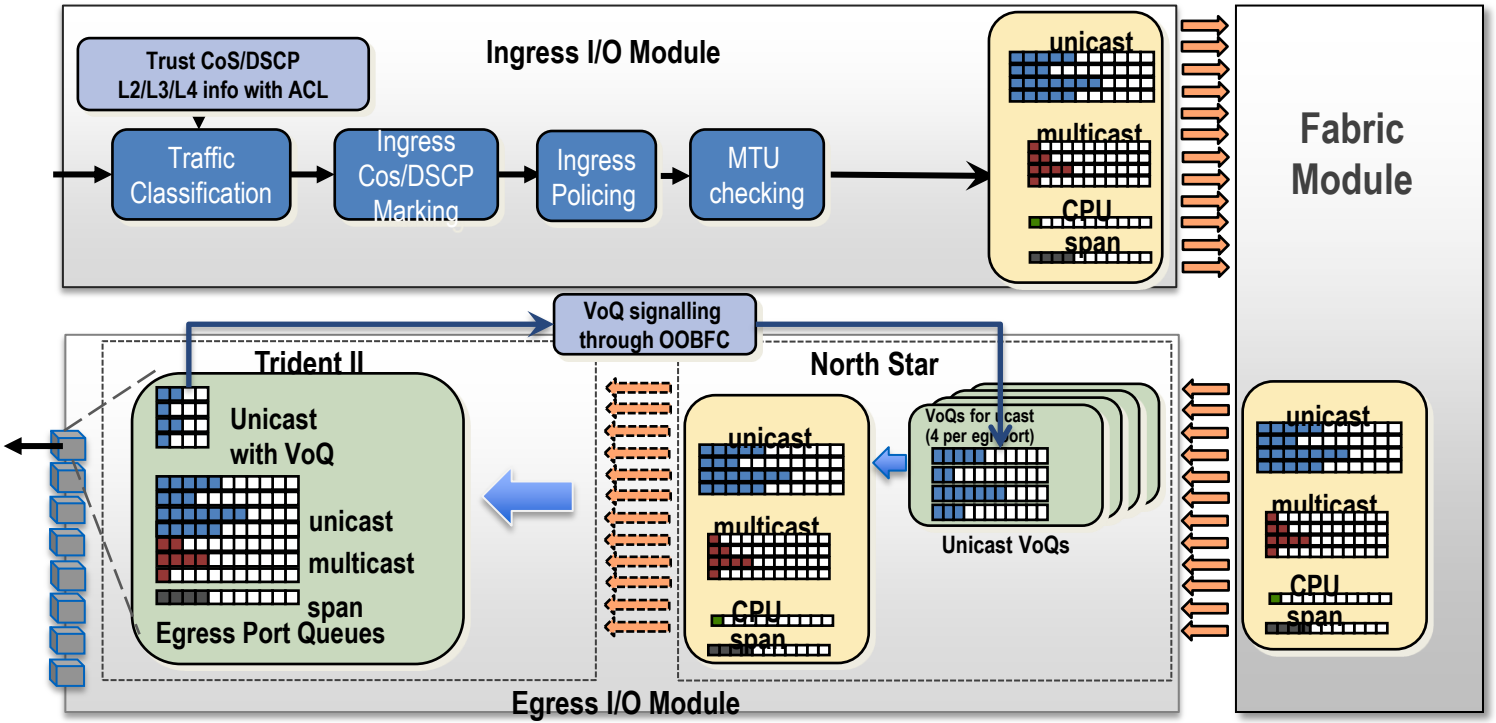
LPM Lookup



LPM Lookup



Nexus 9500 Queuing



Nexus 9300/9500 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

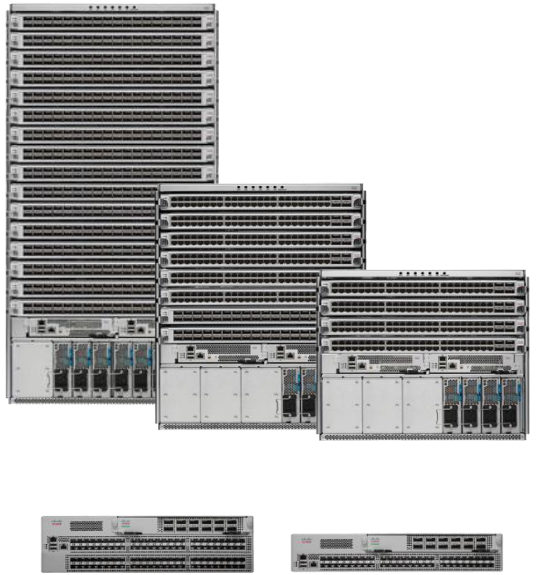
- **Egress Queuing and Shaping**
  - Policy-map type queueing
  - Enable WRED or ECN (default is tail drop)
  - 4 user-defined classes based on qos-group
  - 1 control class for CPU and 1 class for SPAN traffic
  - 3 PFC non-drop queues
  - 3 Priority queues

9500 Standalone Mode Scalability

	Nexus® 9500
LPM Routes	128K
IP Host Entries	88K
MAC Address Entries	160K
Multicast Routes	32K* (hardware capable of 72K)
Multicast Fan Outs	8K (no vPC)
IGMP Snooping Groups	32K* (hardware capable of 72K)
ACL TCAM	Hardware: Ingress: 4K per NFE; up to 96K per system Egress: 1K per NFE; up to 24K per system Available to users: Ingress: 3K per NFE; up to 72K per system Egress: 768 per NFE; up to 18.4K per system
VRF	1000
Maximum Links in Port Channel	32
Maximum ECMP Paths	64
Maximum vPC Port Channels	528
Maximum Active SPAN Sessions	4
Maximum RPVST Instances	507
Maximum HSRP Groups	490
Maximum VLANs	4K
SPAN / ERSPAN	Minimun4, up to 32 active sessions
VXLAN	10K local VXLAN hosts + VTEPs**

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 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

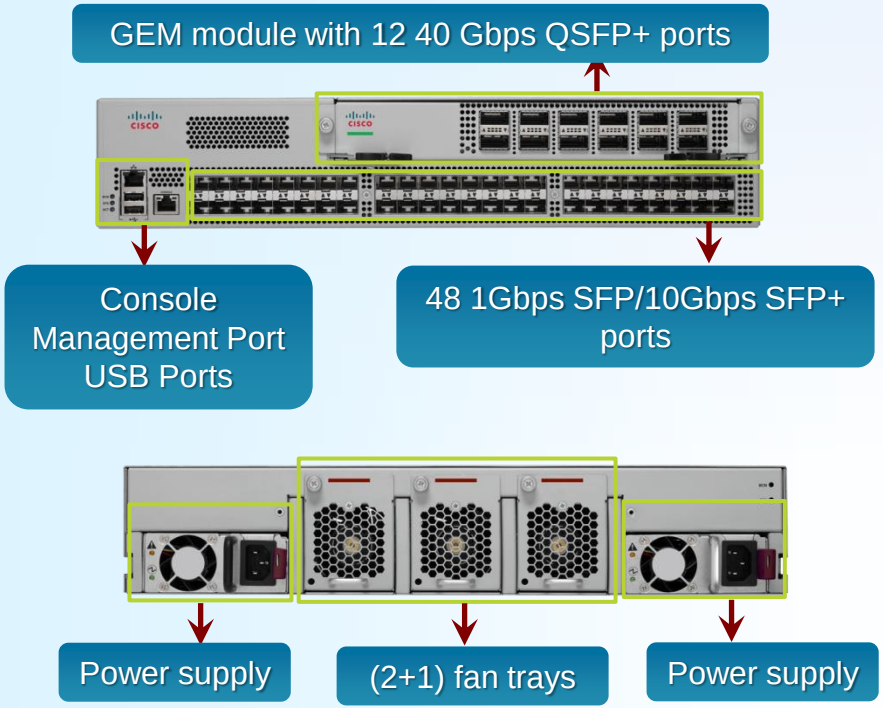
Nexus 9300 - Common

- 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

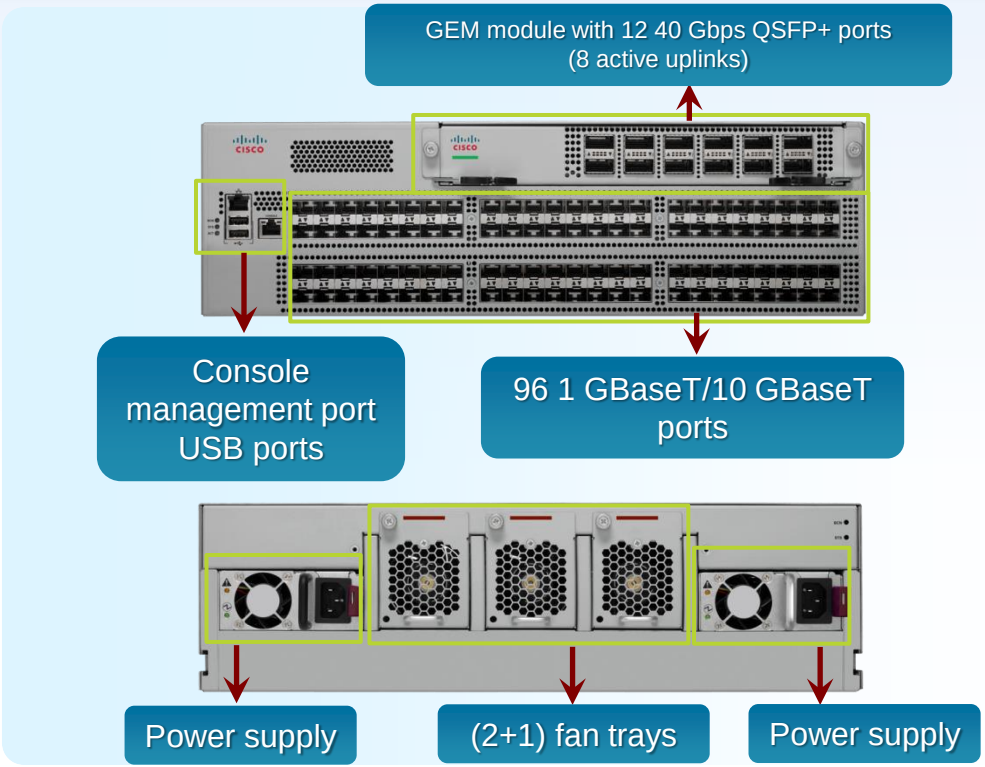
- 2 RU height
- 48 1 Gb SFP/10 Gbps SFP+ ports
- 12 40 Gbps QSFP ports (on GEM module)
- 1 100/1000baseT management port
- 1 RS232 console port
- 2 USB 2.0 ports
- Front-to-back and back-to-front airflow options
- 1+1 redundant power supply options
- 2+1 redundant fans
- No-blocking architecture with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

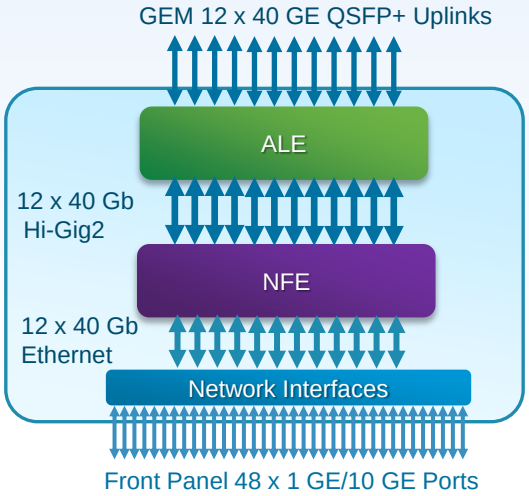
Generic Expansion Module

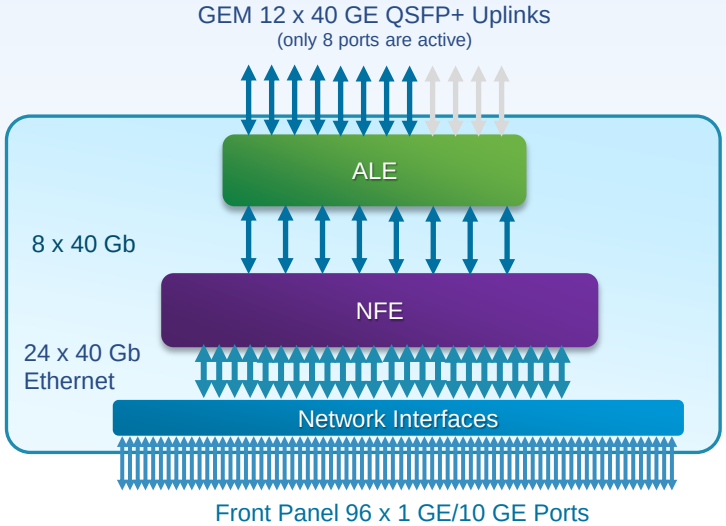
- Redundant (1+1) 650 W and 1200 W AC PS options
- 80-Plus-Platinum-certified power supplies*
- Redundant (2+1) hot-swappable fan trays
- Cold-air intake (blue) and hot-air exhaust (red) options to support front-to-back or back-to-front airflow

Fan and Power Supply

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

Nexus 9300 System Block Diagram



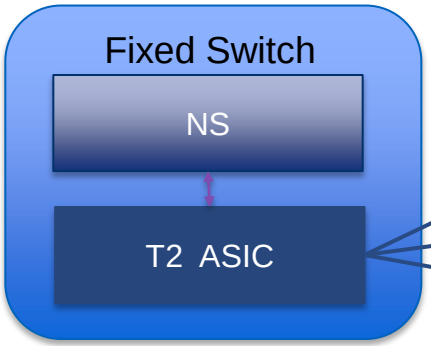


Nexus® 9396PQ/
Nexus 9396TX

Nexus
93128TX

Nexus 9300 Unicast Forwarding

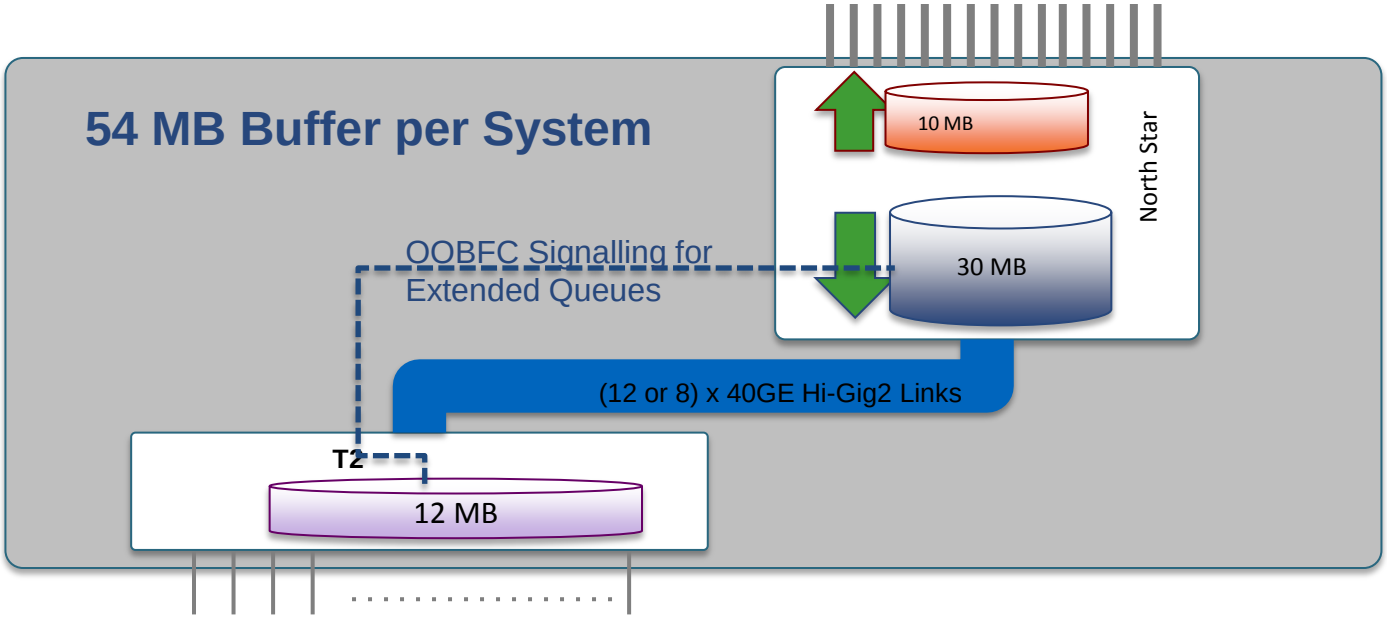
- In Nexus 9300 system there is no separate T2 ASIC that would distribute LPM Route learning from the rest of the system.
- As a result of this, the forwarding tables on a single T2 ASIC is completely responsible for LPM Routes, Host Routes, and MAC Address learning.
- However, it is possible to adjust the allocation of table space based on defined templates.



Feature	Scale (L2 Template)	Scale (L3 Template)
LPM Routes	16K	128K
Host Routes	88K	16K
MAC Addresses	160K	32K

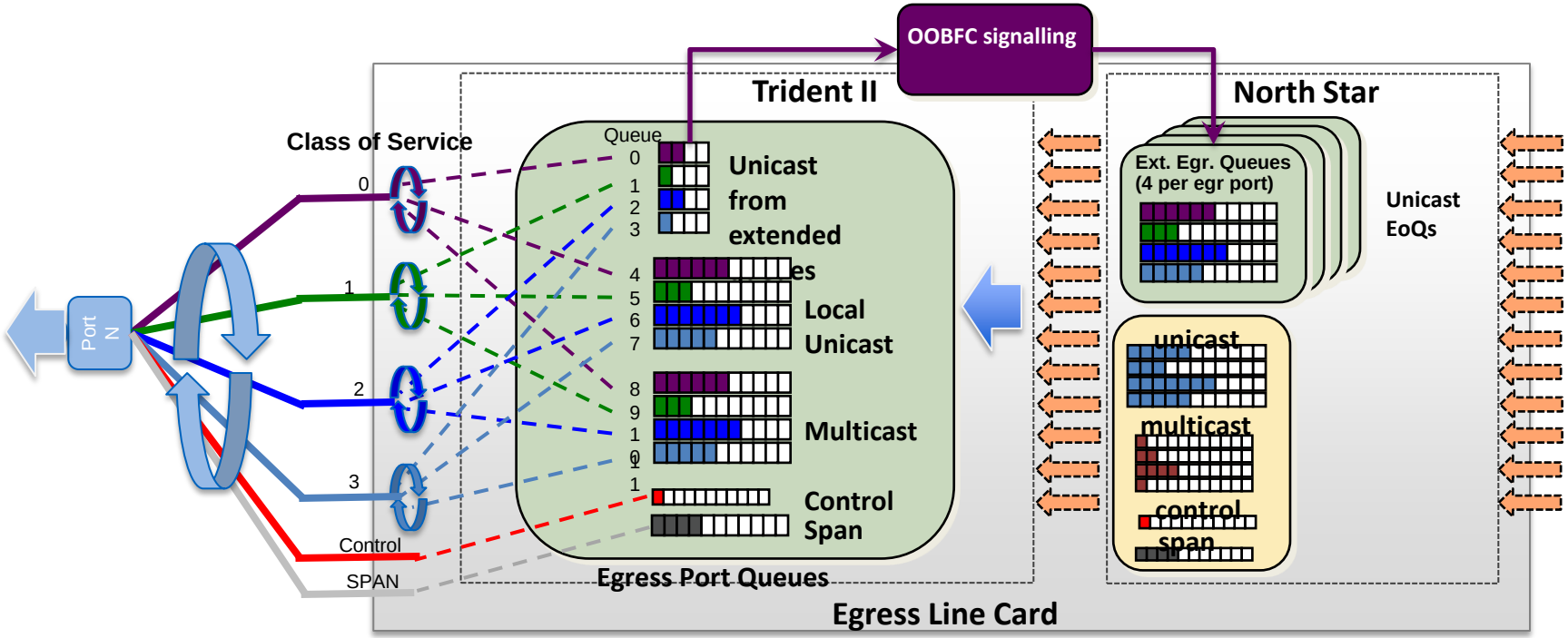
Nexus 9300 Buffer Structure

Port 49 – 60 on Nexus 9396
Port 96 – 103 on Nexus (40GE)



Port 0 – 47 on Nexus 9396
Port 0 – 95 on Nexus 93128 (10GE)

Queuing & Scheduling on Nexus 9300



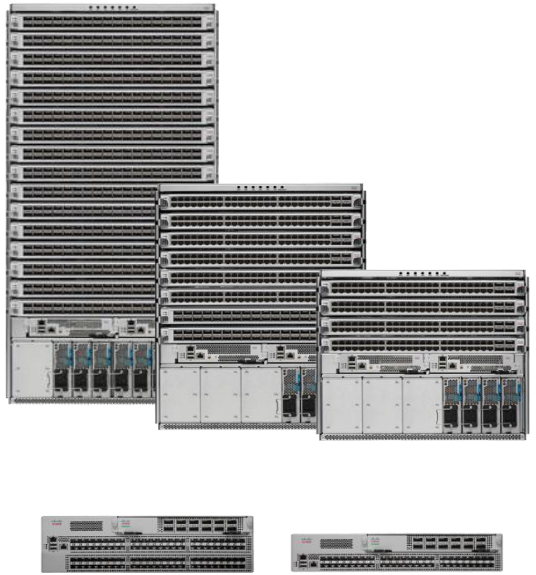
Nexus 9300 System Scalabilities
In Standalone Mode

	Nexus 9300	
	Layer 2 Template	Layer 3 Template
LPM Routes	16K	128K
IP Host Entries	88K	16K
MAC Address Entries	160K	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
VXLAN	10K (local VXLAN hosts + VTEPs)	10K (local VXLAN hosts + VTEPs)

* 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

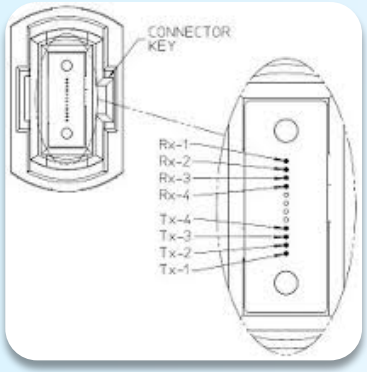
- 40 Gb optics are a significant portion of capital expenditures (CAPEX)
- 40 Gb optics require new cabling

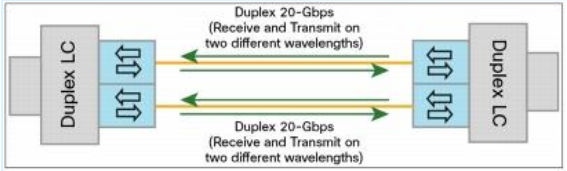
Solution

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

Cisco® 40 Gb SR-BiDi QSFP

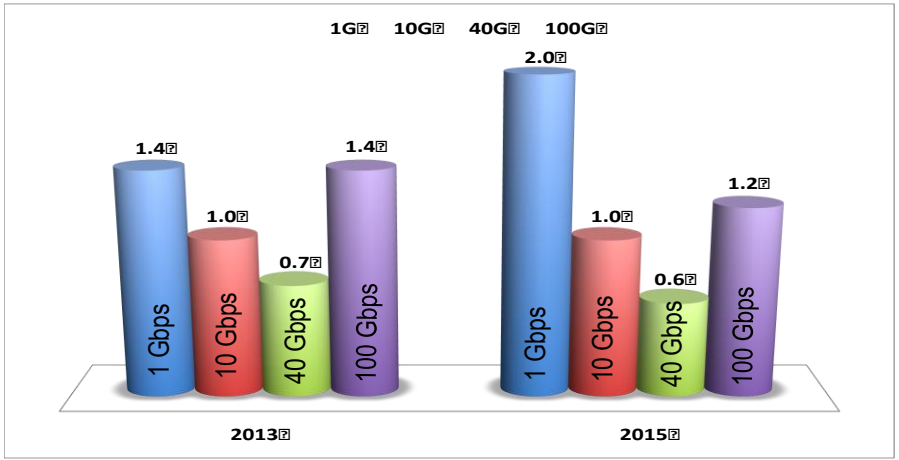
- QSFP, MSA-compliant
- Dual LC connector
- Support for 100 m on OM3 and upto 150m on OM4
- TX/RX on two wavelengths at 20 Gb each





Available end of CY13 and supported across all Cisco QSFP ports

Why a 40G Fabric?

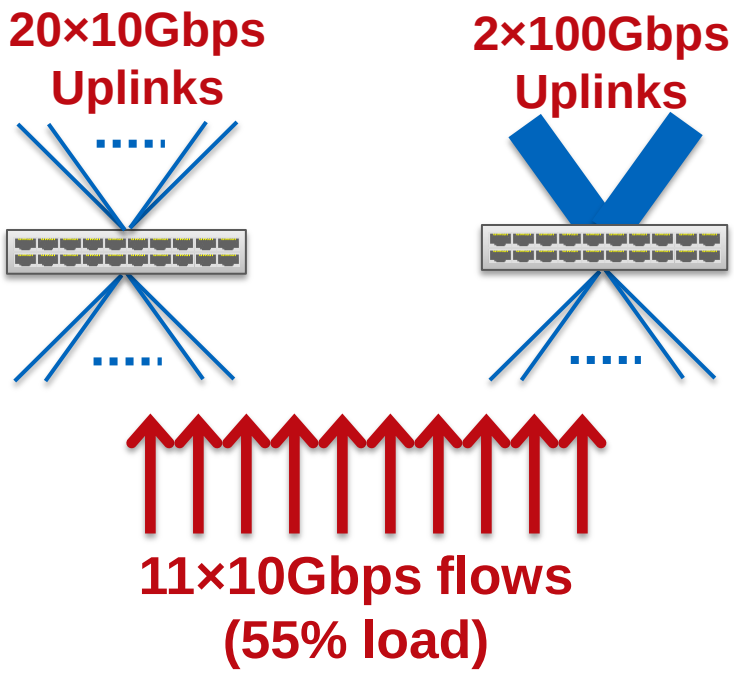


- Optimal Fabric Capacity and Cost
  - 40G provides the optimal cost point currently
  - Speed-up (higher speed transport than edge ports) necessary to achieve effective throughput in a switching network
  - 100G support (Future)

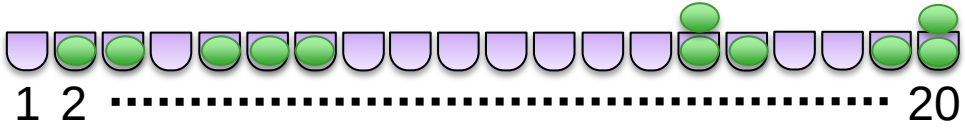
- 40G BiDi Optics
  - QSFP pluggable, MSA compliant
  - Dual LC Connector
  - Support for 100m on OM3 and 125m+ on OM4
  - TX/RX on 2 wavelength @ 20G each

Why Speedup in Network Design

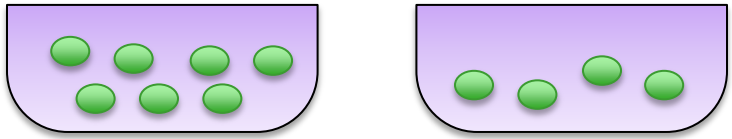
Higher speed links improve ECMP efficiency



Prob of 100% throughput = 3.27%

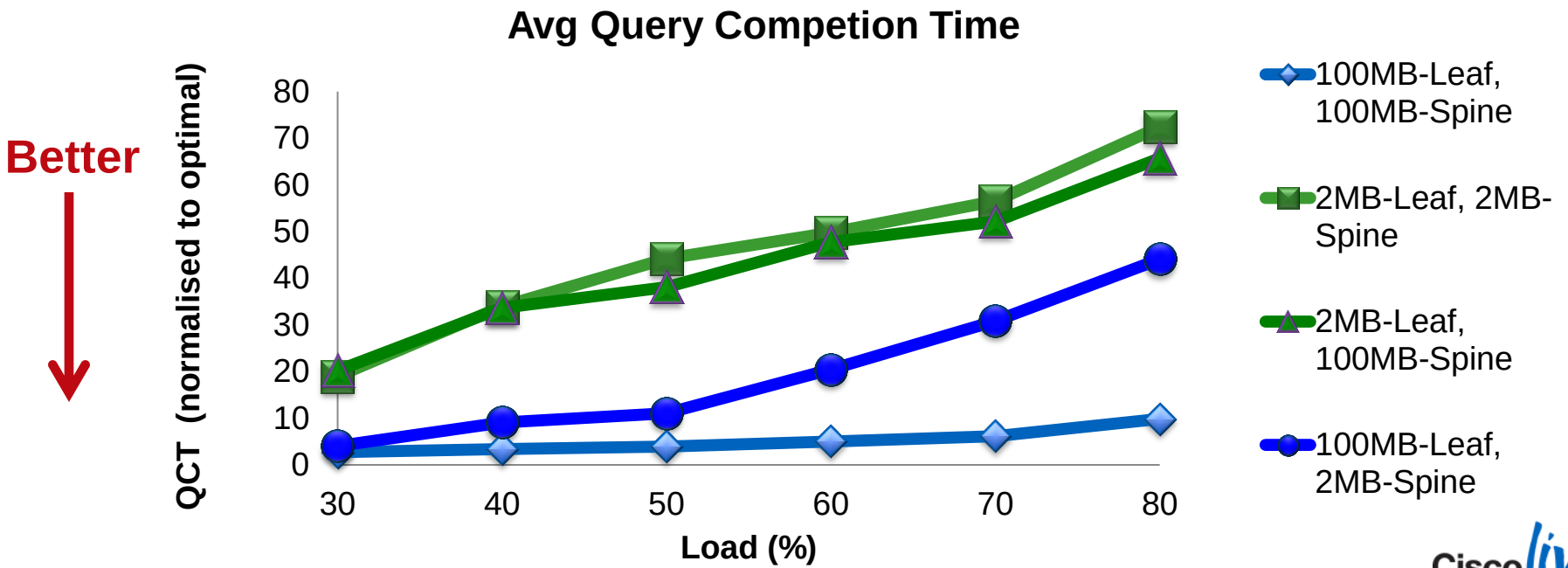


Prob of 100% throughput = 99.95%



Impact of Buffering

Where are large buffers more effective for Incast?

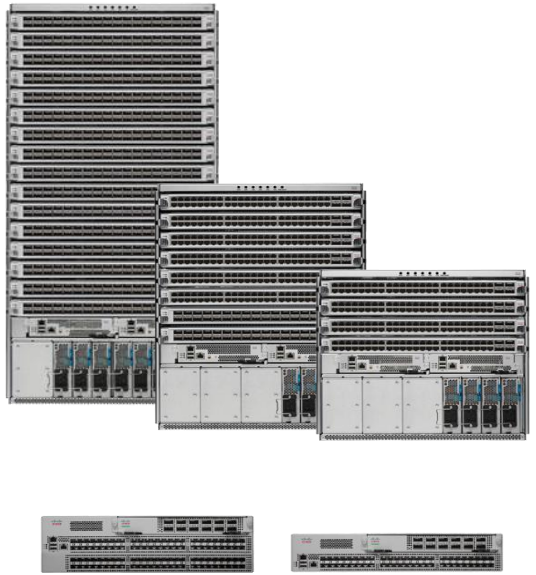


Nexus 9000 Supported Optics

- 10G SFP Transceivers – SR, LR
- 10G Direct Attached Cables – Passive Copper, Active Optical
- 10G FET (When supporting Nexus 2000)
- 40G QSFP Transceivers – SR4, CSR4, BiDi, LR4
- 40G Cables – Passive Copper, Active Optical
- 40G FET
- 1G Transceivers – SM, MM, GLC-T

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 FEXlink: Virtualised Access Switch
Optimised Model for long term TCO during evolution

Cisco Nexus® 5500

+

Cisco Nexus® 2000 FEX

Cisco Nexus® 9300

+

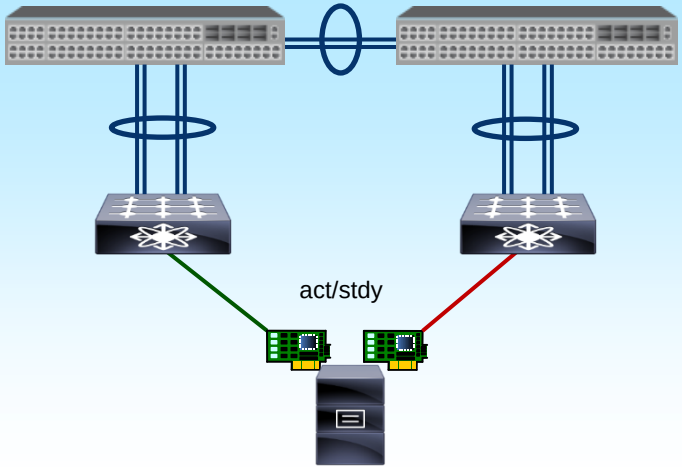
Cisco Nexus® 2000 FEX

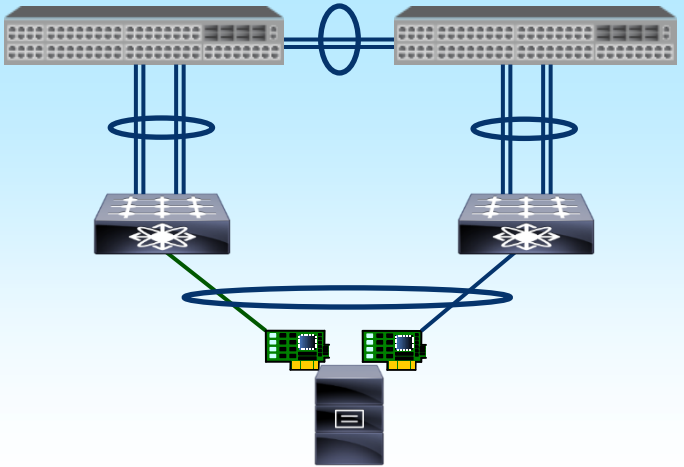
- Migration of Nexus 5500 to Nexus 9300 provides
- Increased scalability
  - 160K MAC
  - 16K Routes
  - 44K MRoutes
  - 160K IGMP Groups
- Addition of 40G uplinks for lower oversubscription
- Addition of VXLAN Bridging, Gateway and Routing capabilities
- Line Rate Layer 2 and Layer 3
- Reduction of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

Nexus 9500/9300
Nexus 2000 FEX Support

Supported FEX Topology:

- Single-homed FEX
- vPC port channel to hosts

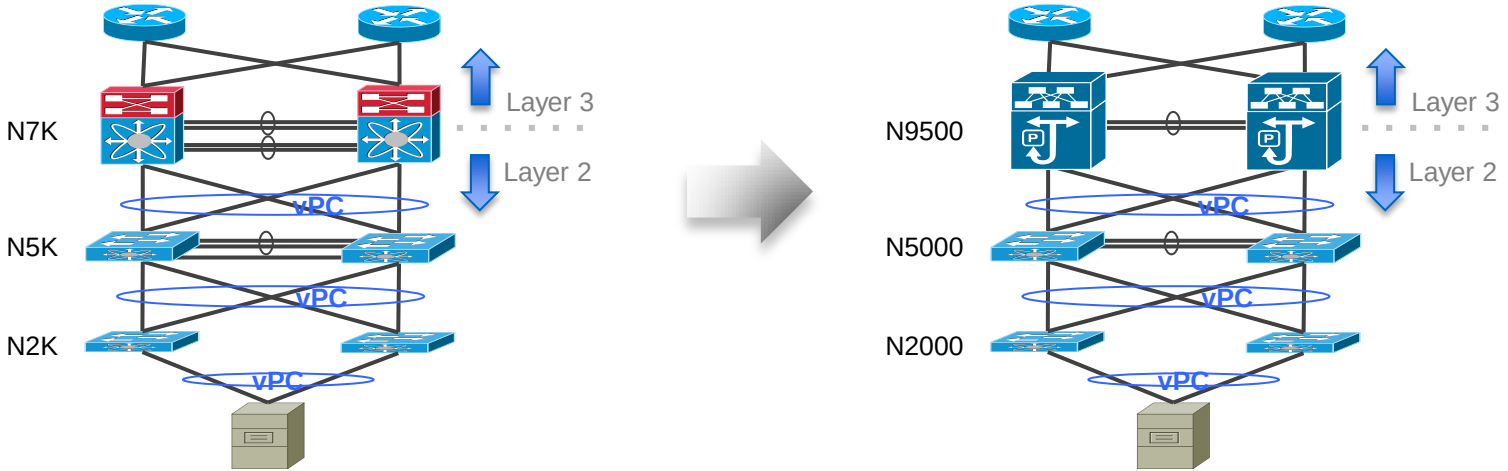




Migration and Interop with Existing Nexus

Pod Design Migration with VPC

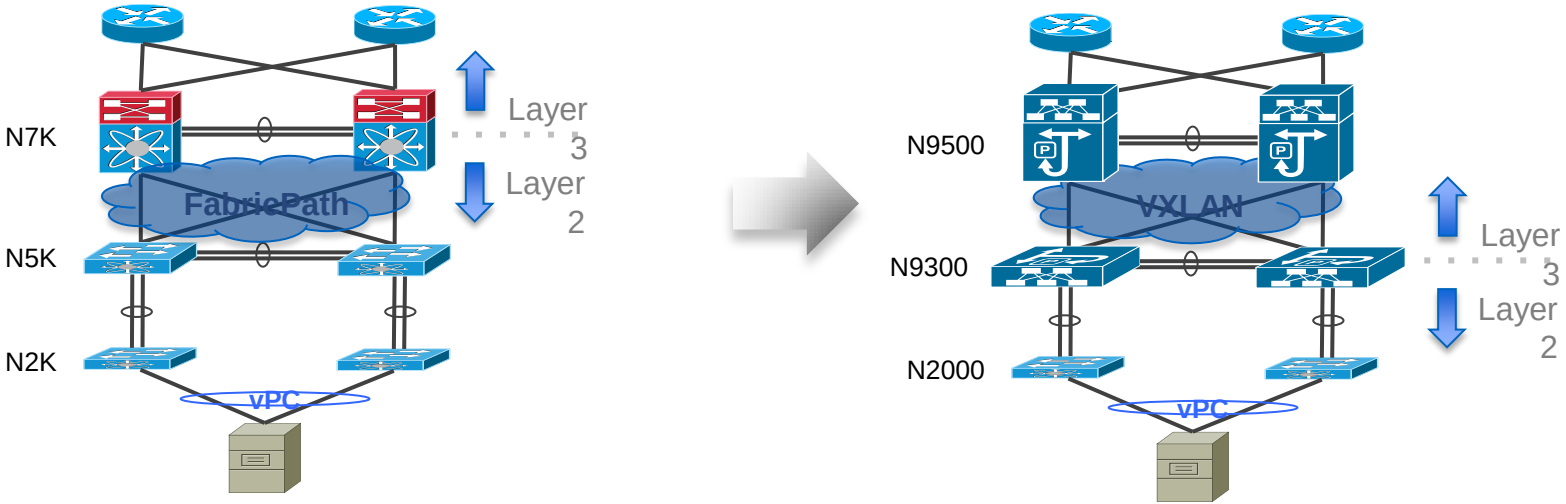
- Nexus 9000 is fully compatible with all existing Nexus vPC & FEX designs
- When customer is looking at consolidation of multiple aggregation or high density 40G aggregation look to migrate to Nexus 9500



Migration and Interop with Existing Nexus

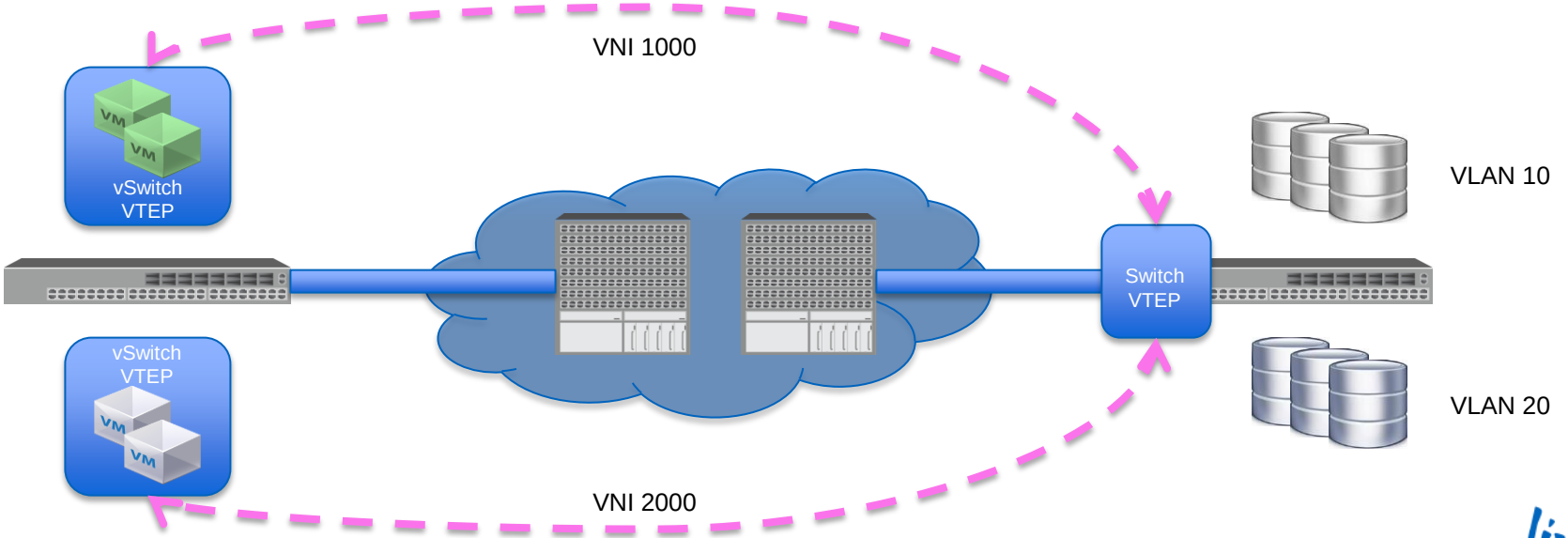
Pod Design Migration with VXLAN

- 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

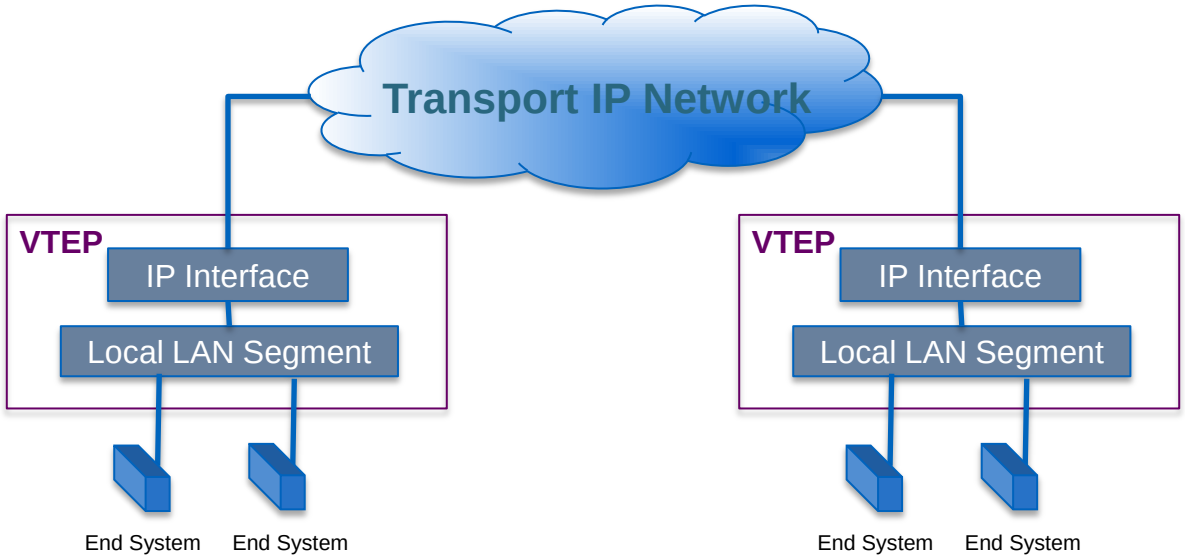
VXLAN can be implemented on both Hypervisor-based Virtual Switches to allow for scalable VM deployments, as well as on Physical switches, which provides the ability to bridge VXLAN segments back into VLAN segments. In these cases, the Physical Switch instantiates a VTEP, and function as a VXLAN Gateway…



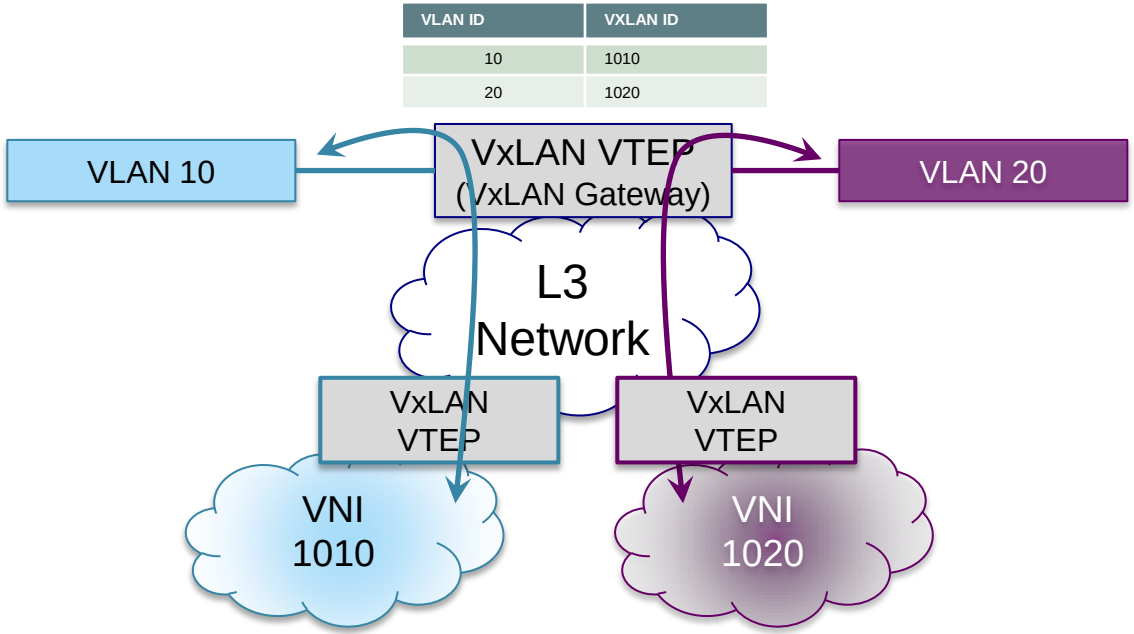
VXLAN VTEP

VXLAN terminates its tunnels on VTEPs (Virtual Tunnel End Point).

Each VTEP has two interfaces - one to provide bridging function for local hosts, the other has an IP identification in the core network for VxLAN encapsulation/de-encapsulation.



VXLAN gateway bridges traffic between VXLAN segment and another physical / logical layer 2 domain (such as a VLAN)…

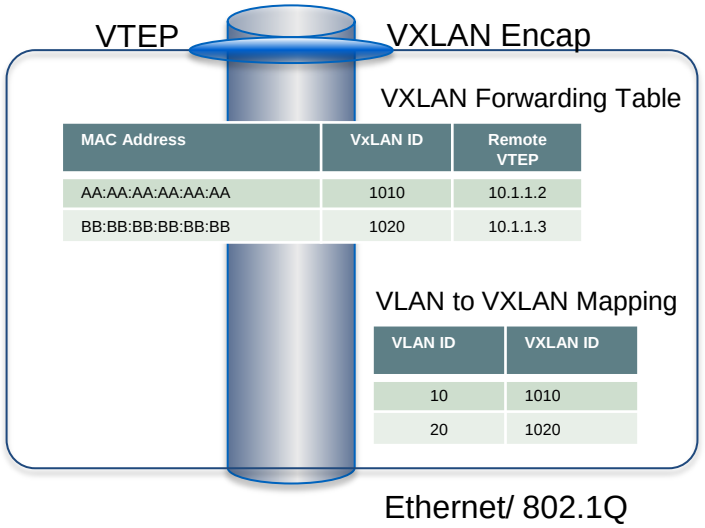


Nexus 9000 Series
VXLAN Gateway

The Nexus 9000 series supports VXLAN Gateway function, allowing VLANs to be bridged/mapped to VXLAN Segments and vice versa…

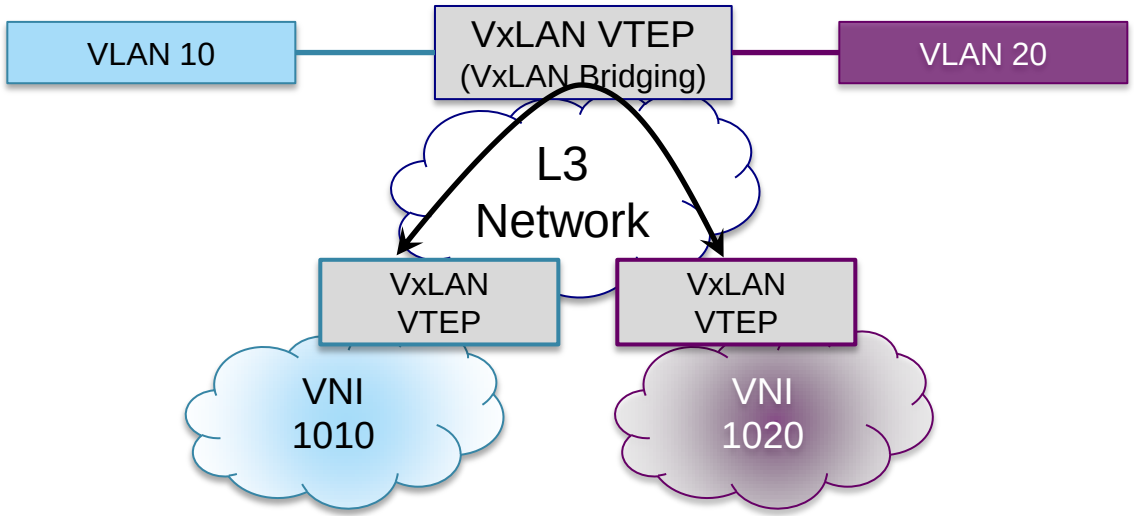
```
feature nv overlay
feature vn-segment-vlan-based

interface et4/13
  switchport
  switchport access vlan 10
  no shut
interface nve1
  no shutdown
  source-interface loopback0
  overlay-encapsulation vxlan
  member vni 1010 mcast-group 230.1.1.1

vlan 10
  vn-segment 1010
```

```
switch# show nve vni
Interface          VNI        Multicast-group   VNI State
-----------------  --------   ---------------   ---------
nve1               1010        230.1.1.1            up
switch# show nve peers
Interface           Peer-IP              VNI         Up Time
-------------       ----------------     --------    -------
nve1                10.1.1.2             1010       00:52:24
switch#
```



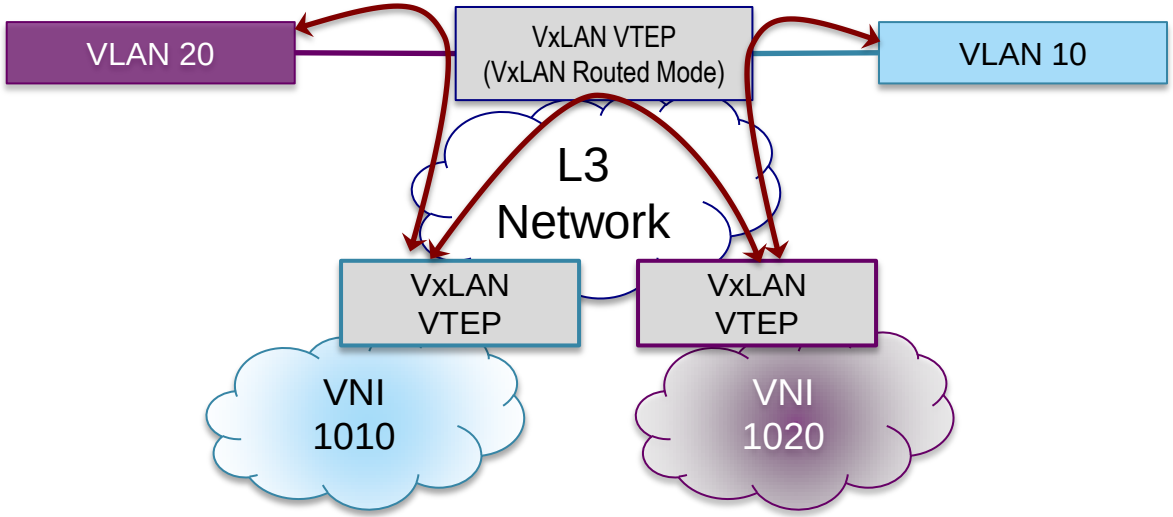
VXLAN “Bridging” bridges traffic between VXLAN segments

VLAN ID	VXLAN ID
10	1010
20	1020



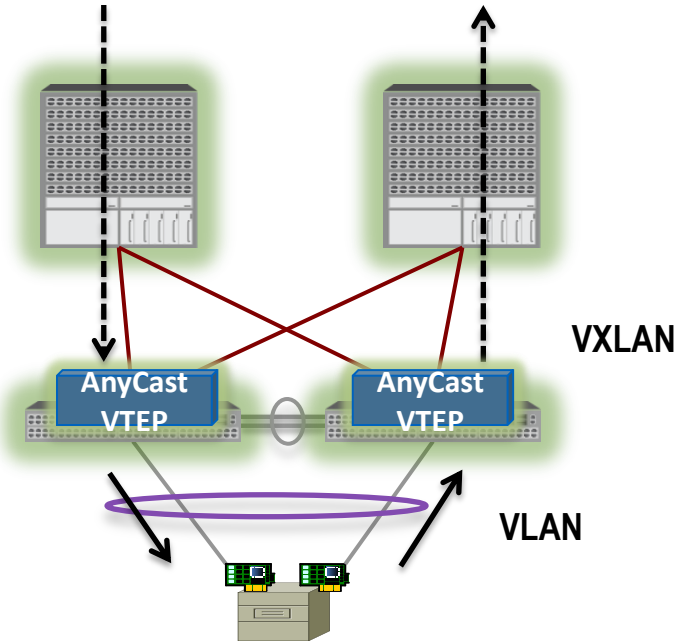
Nexus 9000 Series
VXLAN Routed Mode

VXLAN routed mode ‘routes’ traffic between VXLAN segments and between VXLAN another physical / logical layer 2 domain (such as a VLAN)…

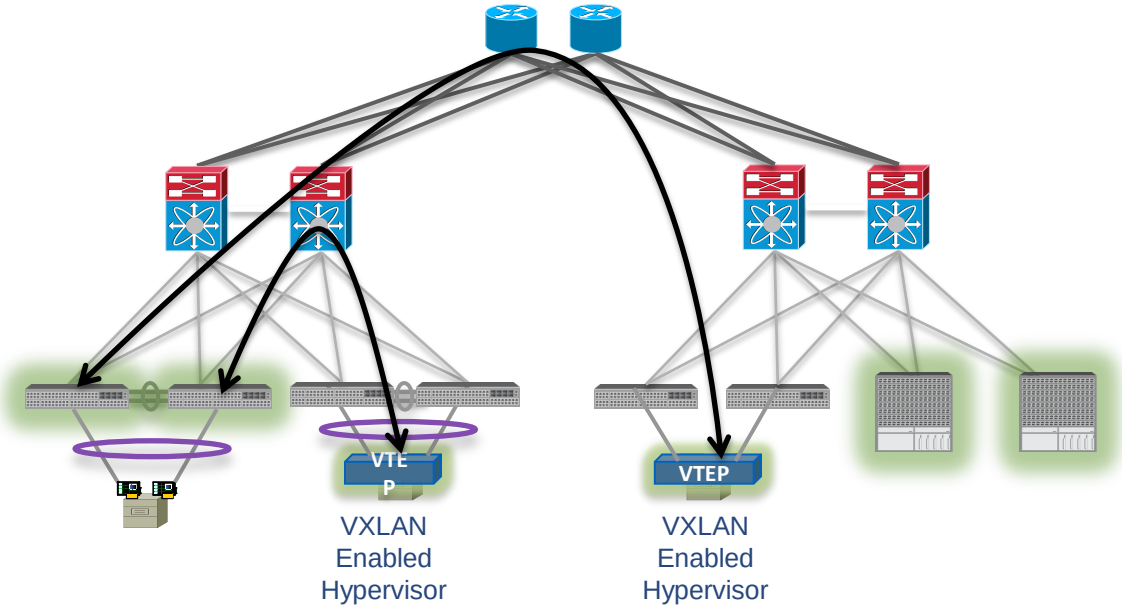


VXLAN Forwarding
vPC

- 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 (MCT link)

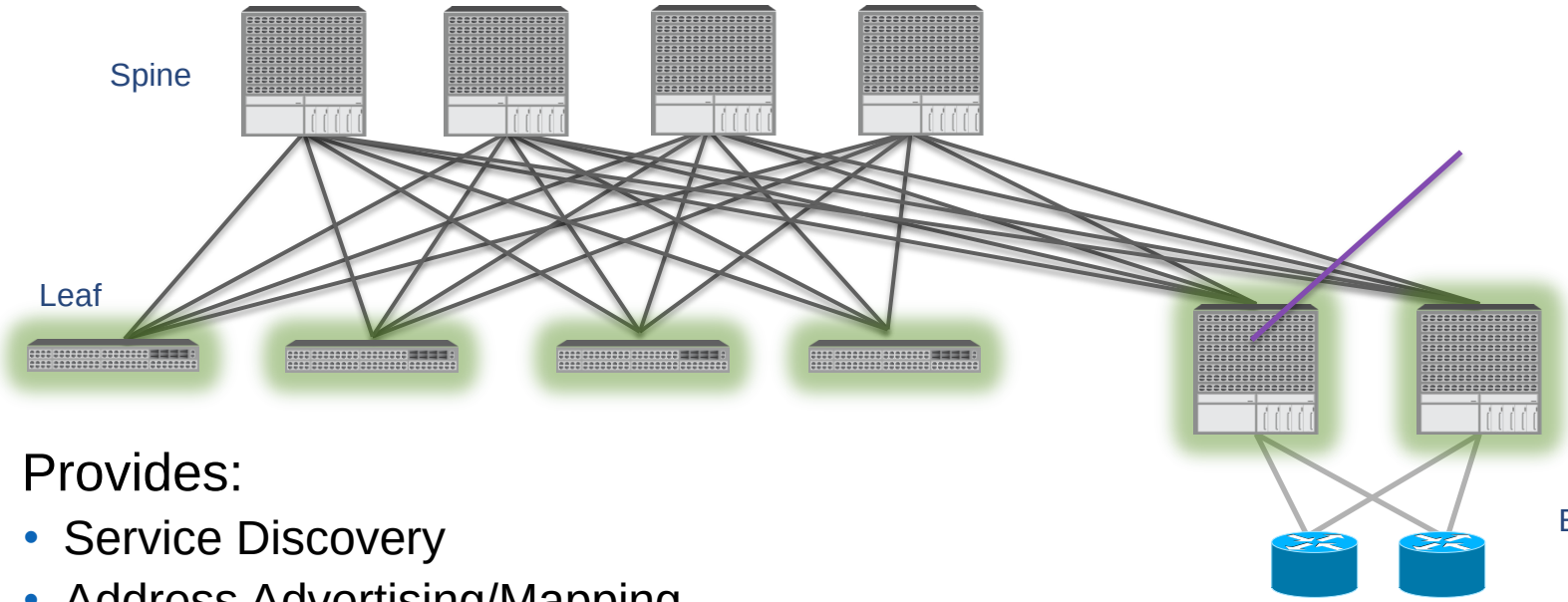


VXLAN Gateway – Routed Access + IP Mobility



- VXLAN Gateway defined at access layer (leaf) – Nexus 9000
- Multicast needs to be enabled for VXLAN to work on the source interface
- Next hop of VTEP needs to be Layer 3
- vPC needs peer gateway
- Only 1:1 mapping is allowed for VXLAN to VLAN
- Recommended N9K to be configured as STP root switch in each L2 network
- Link discovery protocols like CDP, LLDP will not discover neighbours on the remote VTEPs
- Virtual to physical migration (P2V)

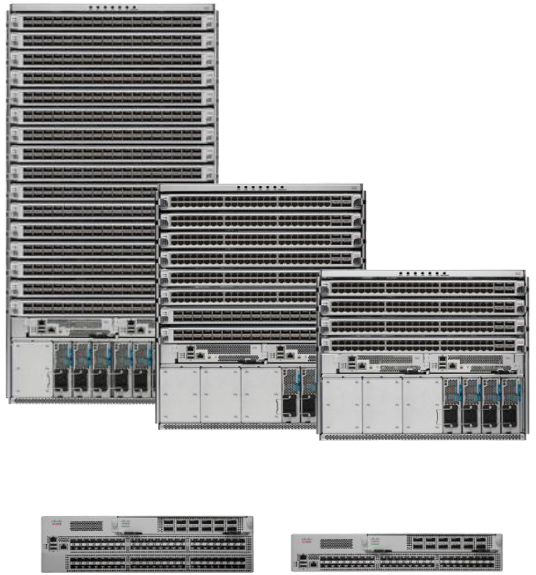
VXLAN EVPN + Anycast Gateway



- Provides:
  - Service Discovery
  - Address Advertising/Mapping
  - Tunnel Management
  - Extensions for multi-homing and advanced services can be provided

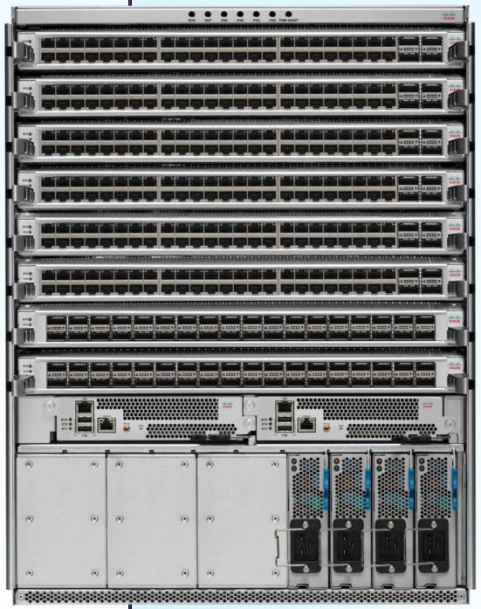
Agenda – Nexus 9000 Architecture

- 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, 9300 & 3x00; combined kick-start and system image (Trimmed base image, ~ 50%)
- Comprehensive: L2/L3/VXLAN, ISSU, Patching* (Cold, Hot), Online Diagnostics (GOLD) etc.
- 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 Ethanalyzer
- SMTP email “pipe” output
- Embedded Event Manager (EEM)
- Flow monitoring
- vTracker

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

Python Scripting

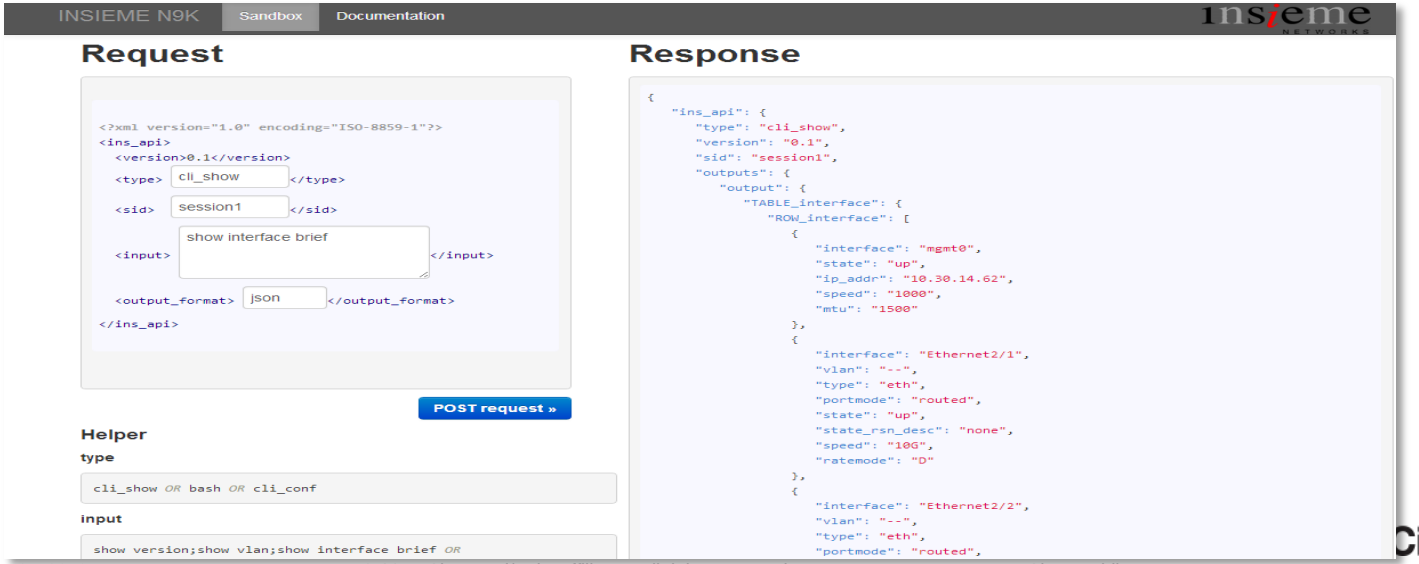
- Built in Python Shell
- Can be used to execute CLI commands and reference Objects through Python interpreter
- Most commands can be executed to return the command output as a Python Dictionary
- Pass arguments to python scripts from CLI
- Libraries portable
- Integration with Embedded Event Manager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

Nexus 9000 - NX-API

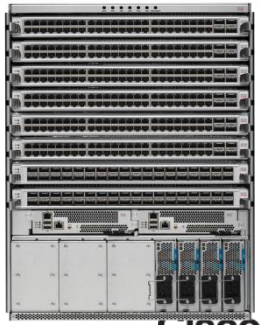
- Open RPC API – Extensible to support REST
- Universal Access: http or https based
- Programmability Oriented
- Ready for Integration: CLI based input and structured output (JSON/XML)



Bash Access

- 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
- Bash shell has non-root privileges to protect against unintended operator errors
- Role-based access to Bash

```
Insieme-N9K# bash
Insieme-N9K(shell)> for i in {1..5}
> do
> echo "RX Counters"
> date
> ifconfig eth0 | grep "RX packets" | cut -d ':' -f2 | cut -d ' ' -f1
> sleep 1
> done
RX Counters
Sun Feb  6 03:07:04 UTC 2011
763939
RX Counters
Sun Feb  6 03:07:05 UTC 2011
763944
RX Counters
Sun Feb  6 03:07:06 UTC 2011
763955
RX Counters
Sun Feb  6 03:07:07 UTC 2011
763964
RX Counters
Sun Feb  6 03:07:08 UTC 2011
763969
Insieme-N9K(shell)> ifconfig eth0 down
SIOCSIFFLAGS: Permission denied
Insieme-N9K(shell)>
```



BCM Shell Access

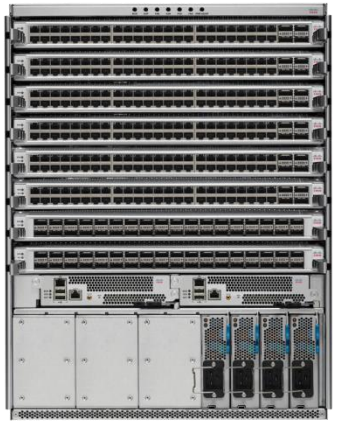
- Issue a CLI to get shell access to underlying BCM chips
- Direct read/write access to hardware tables
- Can Peek/Poke underlying registers
- Python wrapper to get BCM Shell output


```
module-22# test hardware internal bcm-usd bcm-diag-shell
Available Unit Numbers: 0 1
bcm-shell.0> vlan show
vlan 1  ports hg (0x00000000000000000000000000000000000000000000000000000001fffffffe), untagged none (0x00000000
00000000000000000000000000000000000000000000000000000000) MCAST_FLOOD_UNKNOWN
vlan 4032       ports hg (0x00000000000000000000000000000000000000000000000000000001fffffffe), untagged none (0x
00000000000000000000000000000000000000000000000000000000) MCAST_FLOOD_UNKNOWN
vlan 4033       ports hg (0x00000000000000000000000000000000000000000000000000000001fffffffe), untagged none (0x
00000000000000000000000000000000000000000000000000000000) MCAST_FLOOD_UNKNOWN
vlan 4034       ports hg (0x00000000000000000000000000000000000000000000000000000001fffffffe), untagged none (0x
00000000000000000000000000000000000000000000000000000000) MCAST_FLOOD_UNKNOWN
vlan 4035       ports hg (0x00000000000000000000000000000000000000000000000000000001fffffffe), untagged none (0x
00000000000000000000000000000000000000000000000000000000) MCAST_FLOOD_UNKNOWN
vlan 4042       ports none (0x00000000000000000000000000000000000000000000000000000000000000000), untagged none (
0x00000000000000000000000000000000000000000000000000000000) MCAST_FLOOD_UNKNOWN
vlan 4043       ports cpu,hg (0x00000000000000000000000000000000000000000000000000000001ffffffff), untagged none
 (0x00000000000000000000000000000000000000000000000000000000) MCAST_FLOOD_UNKNOWN
```

LinuX Containers (LXC) on Nexus 9000/3000

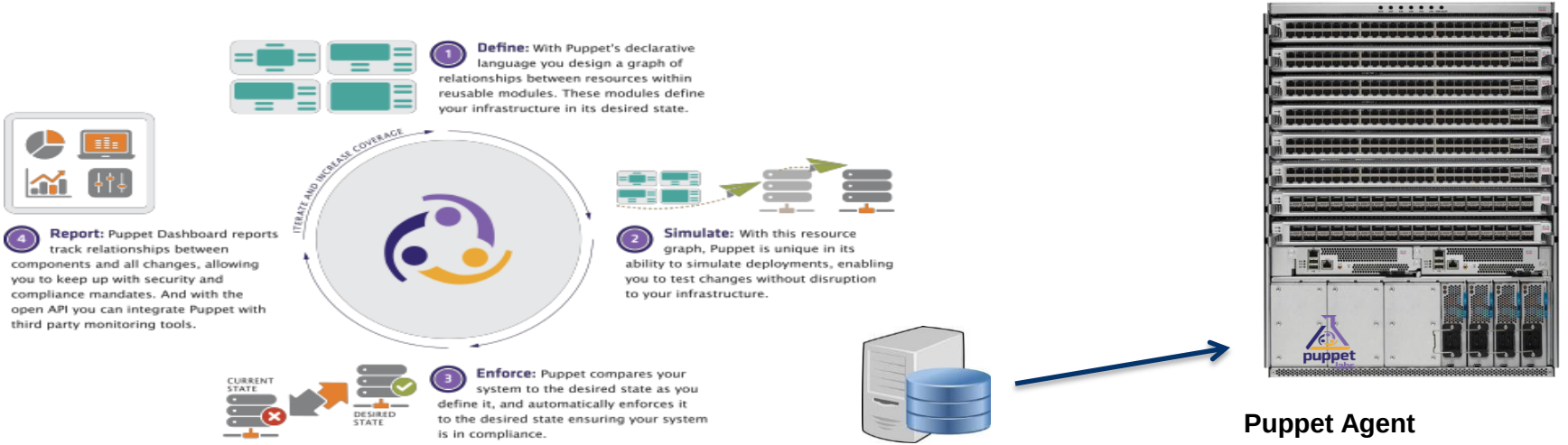
- Provides a secure and segregated operating environment for applications
- Can run either Cisco or Open Source applications
- Can use standard Linux distros
- OS Level Virtualisation
- Shared Kernel
- Shared physical resources
- Isolation through name spaces

NX-OS



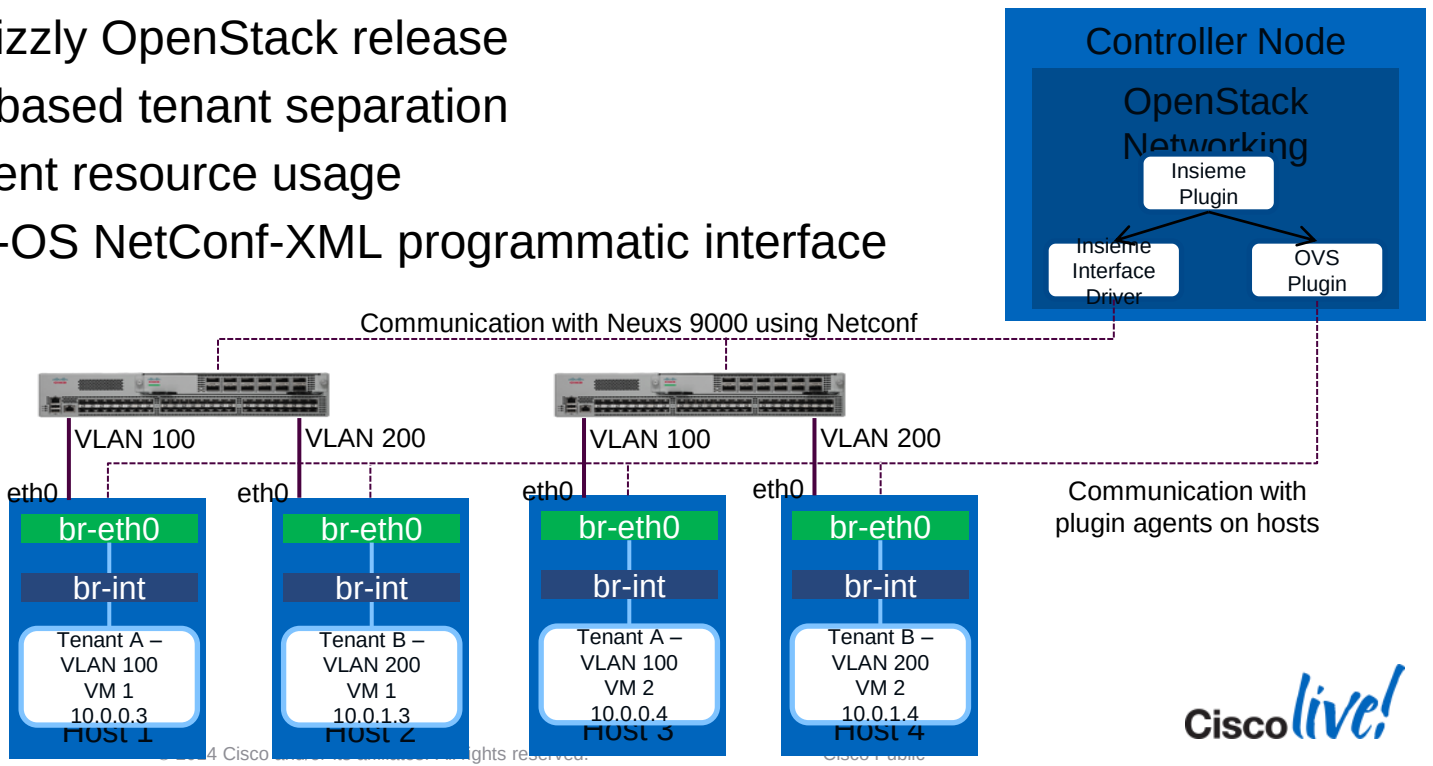
Puppet / Chef

- Cross-platform IT automation software leveraging declarative language to manage IT infrastructure lifecycle
- Allows for automation of config or patch roll-out



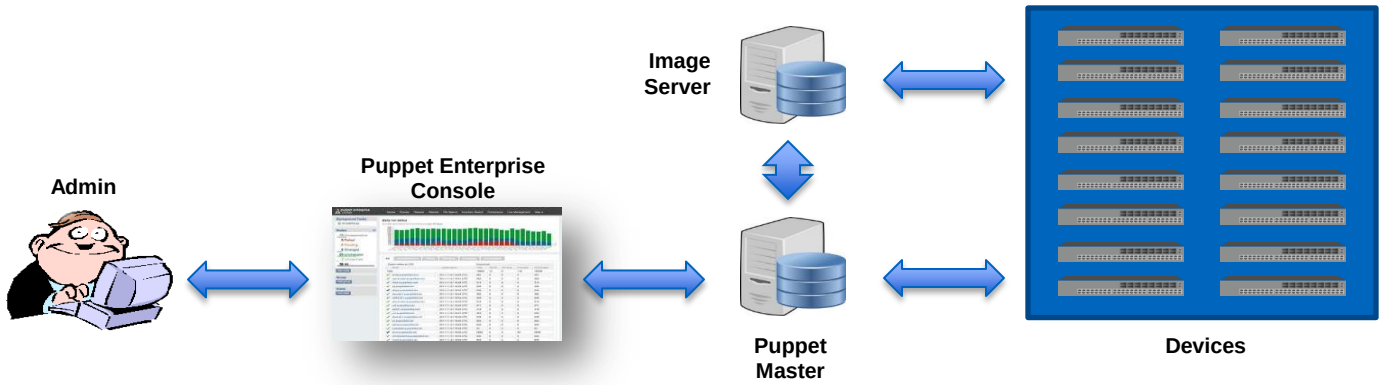
OpenStack Network (Neutron) Plugin

- 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



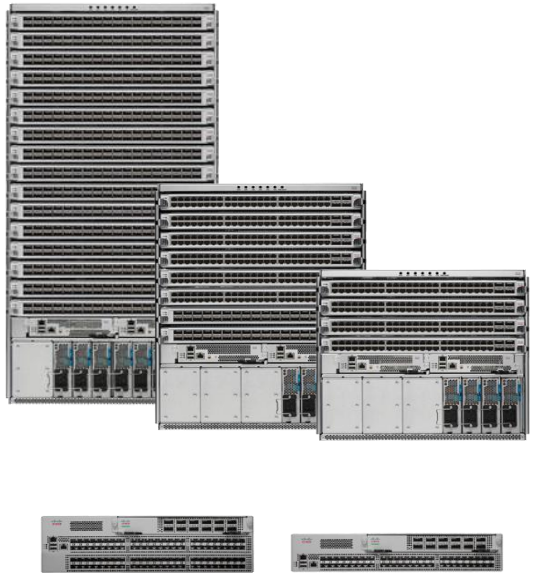
NX-OS Image Patching

- 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

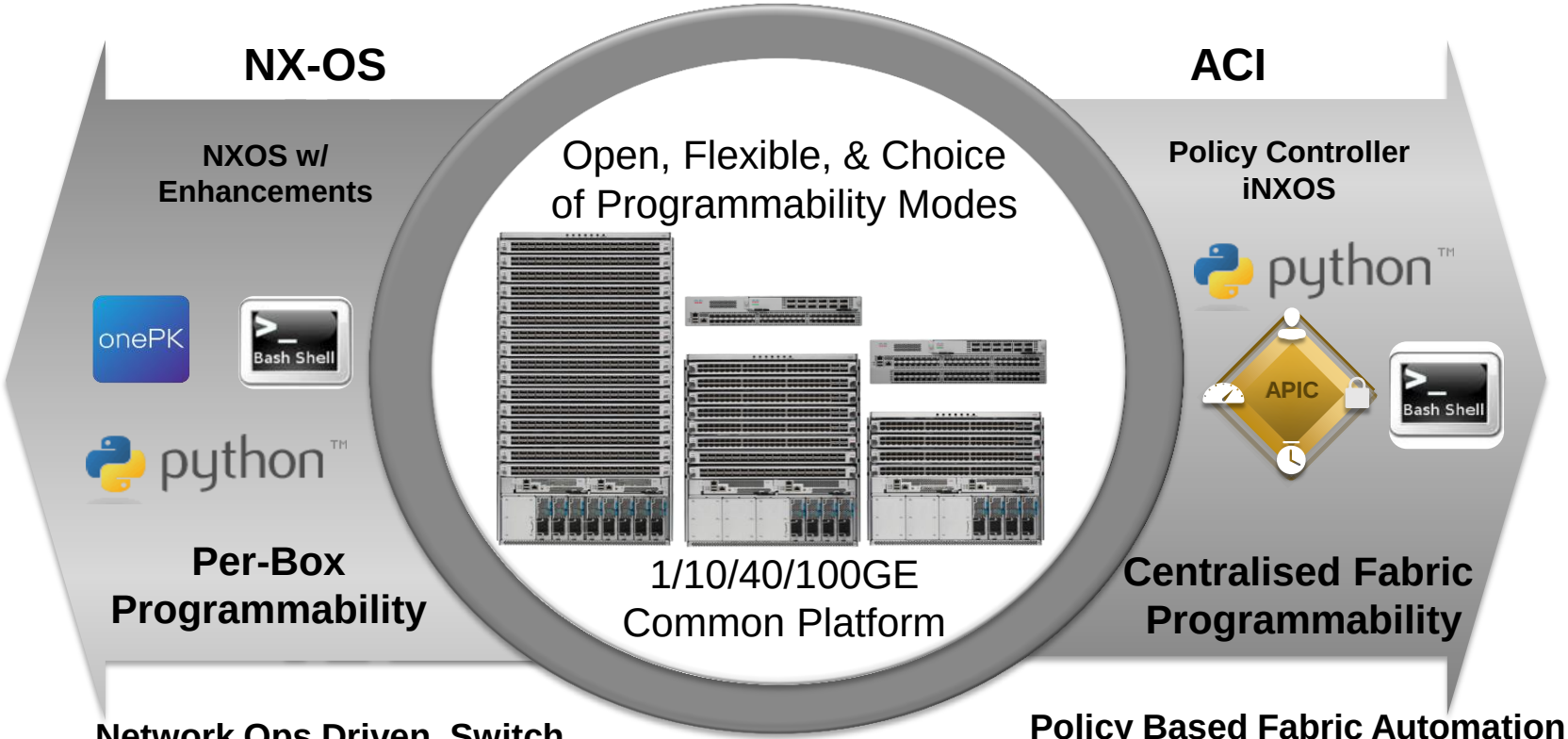


Agenda – Nexus 9000 Architecture

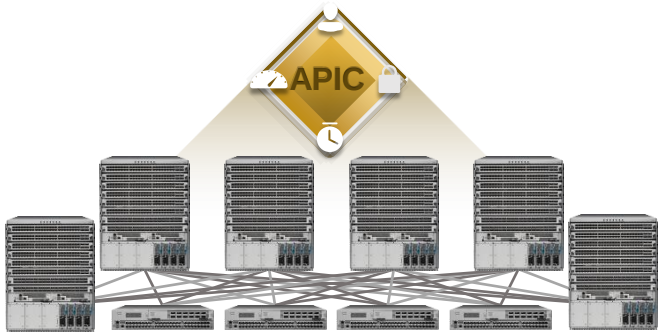
- Nexus 9000
  - Nexus 9000 Hardware
  - Nexus 9500 Chassis
  - Nexus 9500 Line Cards
  - Nexus 9500 Packet Forwarding
  - Nexus 9300
- Nexus 9000 and 40G
- Nexus 9000 Designs: FEX, vPC & VXLAN
- Nexus 9000 & Dev-Ops
- ACI & Nexus 9000



Common Platform: Two Modes of Operation

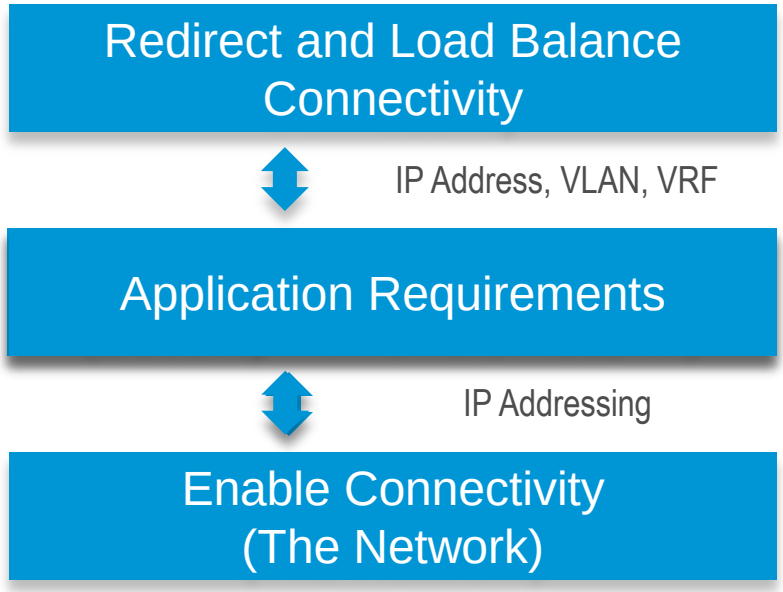


ACI Design Philosophy

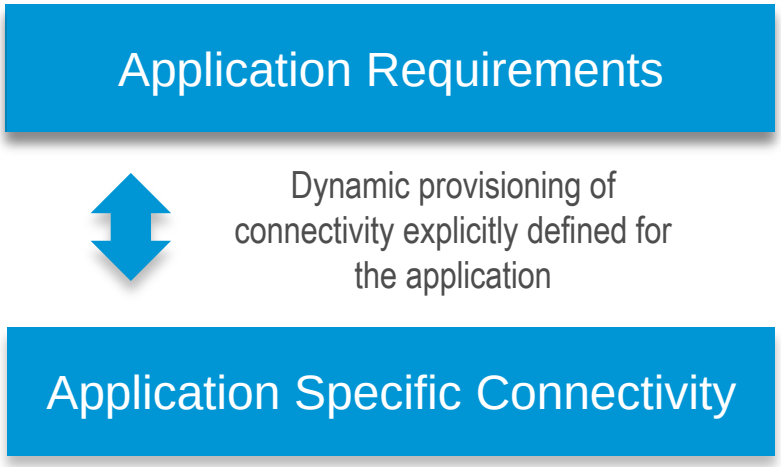


- System Architecture
  - Expand Networking From Boxes To Systems
- Open Source & Multi-vendor
  - Innovations Published to Open Source
- Physical & Virtual
  - Traditional, Virtualised, & Next-Generation Non Virtualised Applications
- Velocity
  - Abstraction, Abstraction, Abstraction
- Costs
  - Best of Merchant & Custom Silicon for CAPEX Unmatched Automation for OPEX

Overloaded Network Constructs

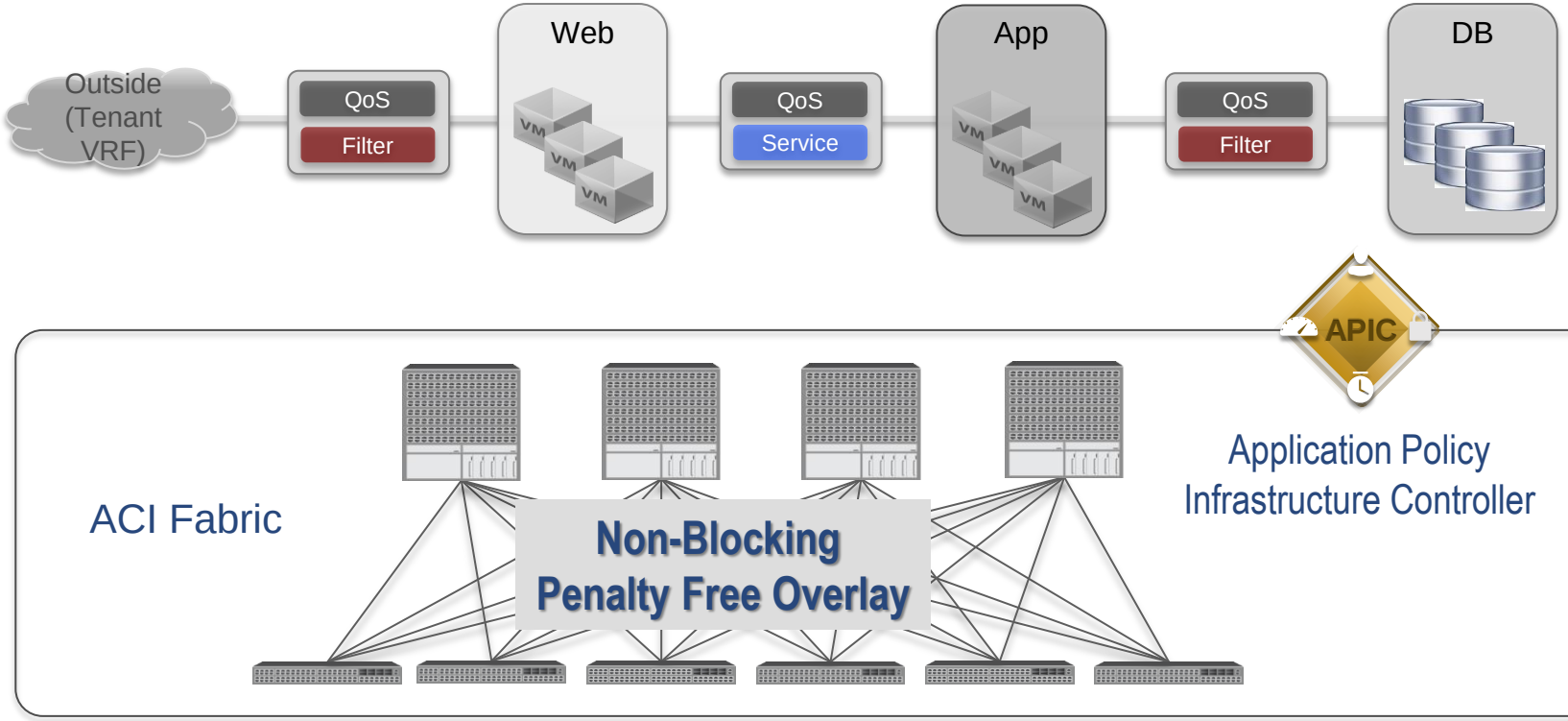


ACI directly maps the application connectivity requirements onto the network and services fabric



ACI Fabric

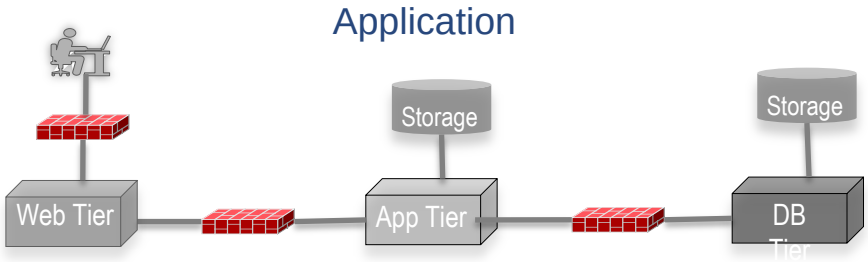
Logical Network Provisioning of Stateless Hardware



Application Network Profile

Policy Based Fabric Management

- Extend the principle of UCSM service profiles to the entire fabric
- Network Profile: Stateless Definition of Application Requirements
  - Application Tiers
  - Connectivity policies
  - L4 – L7 Services
  - XML/JSON Schema
- Fully Abstracted from the infrastructure implementation
  - Removes dependencies of the infrastructure
  - Portable across different Data centre fabrics



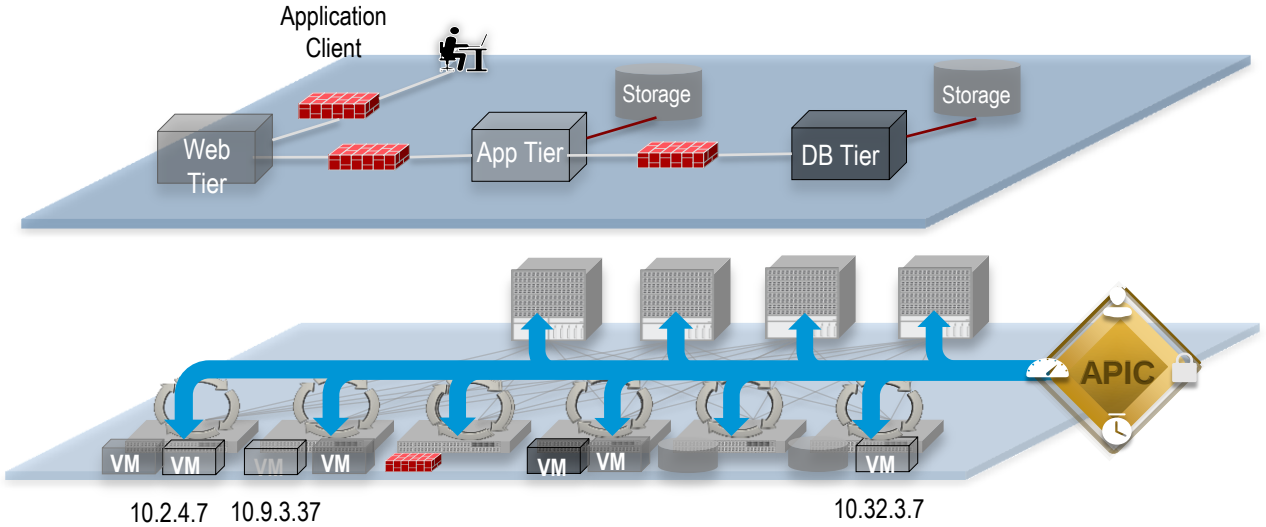
Network Profile fully describes the application connectivity requirements

```
## Network Profile: Defines Application Level Metadata (Pseudo Code Example)

<Network-Profile = Production_Web>
 <App-Tier = Web>
   <Connected-To = Application_Client>
      <Connection-Policy = Secure_Firewall_External>
  <Connected-To = Application_Tier>
      <Connection-Policy = Secure_Firewall_Internal & High_Priority>
. . .
 <App-Tier = DataBase>
   <Connected-To = Storage>
      <Connection-Policy = NFS_TCP & High_BW_Low_Latency>
. . .
```

Application Policy Model & Instantiation

Application Policy Model: Defines the application requirements (Application Network Profile)

Policy Instantiation: Each device dynamically instantiates the required changes based on the policies

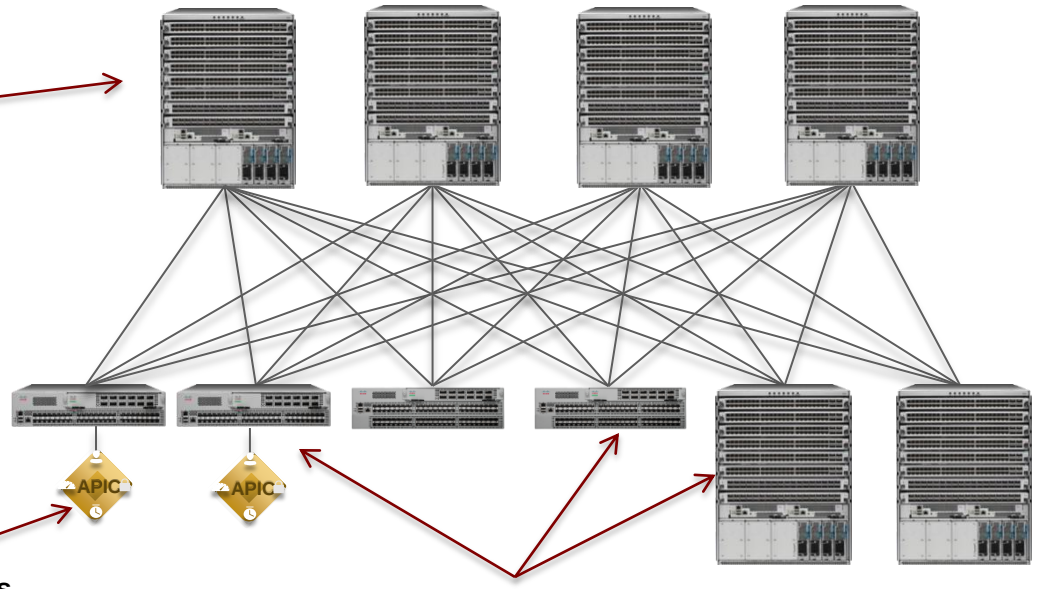


- All forwarding in the fabric is managed via the Application Network Profile
  - IP addresses are fully portable *anywhere* within the fabric
  - Security & Forwarding are fully *decoupled* from any physical or virtual network attributes
  - Devices autonomously update the state of the network based on configured policy requirements

ACI Fabric

Fabric Spine Nodes

- **16 Slot Modular**
- **8 Slot Modular**
- **4 Slot Modular**
- **Mini-Spine (36 ports)**

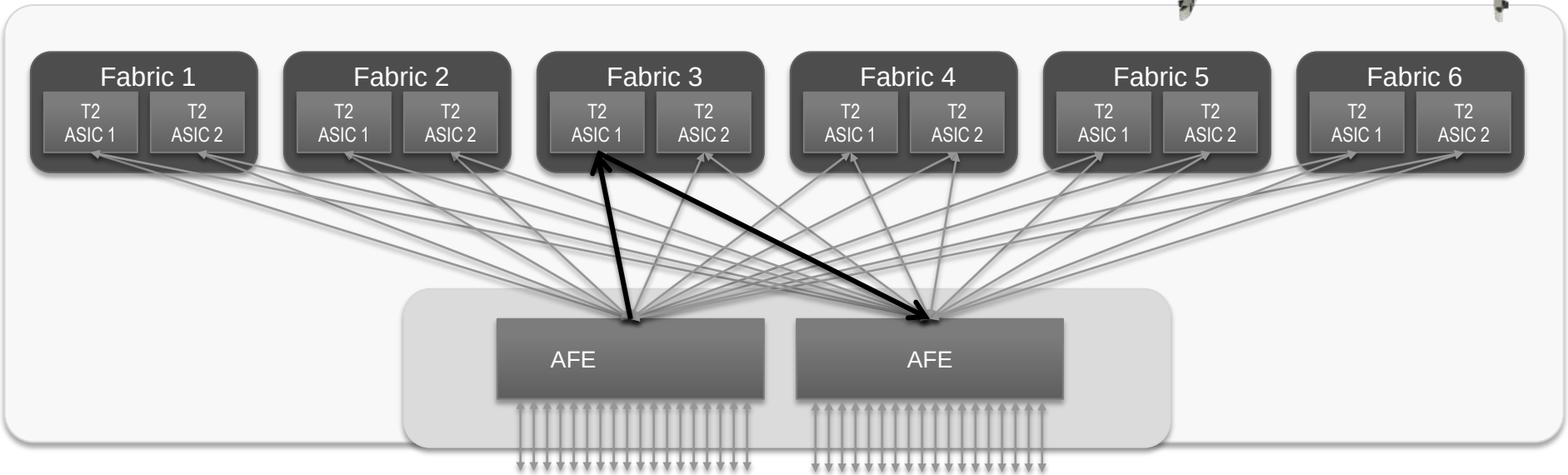


APIC Servers

UCS ‘C’ Series (Intel)

Fabric Leaf Nodes
4, 8 & 16 slot Modular (post FCS)
48 x 1/10 + 12 x 40G
96 x 1/10 + 8 x 40G
Variety of 1 & 2 RU form factors (post FCS)

Nexus 9736 series line card



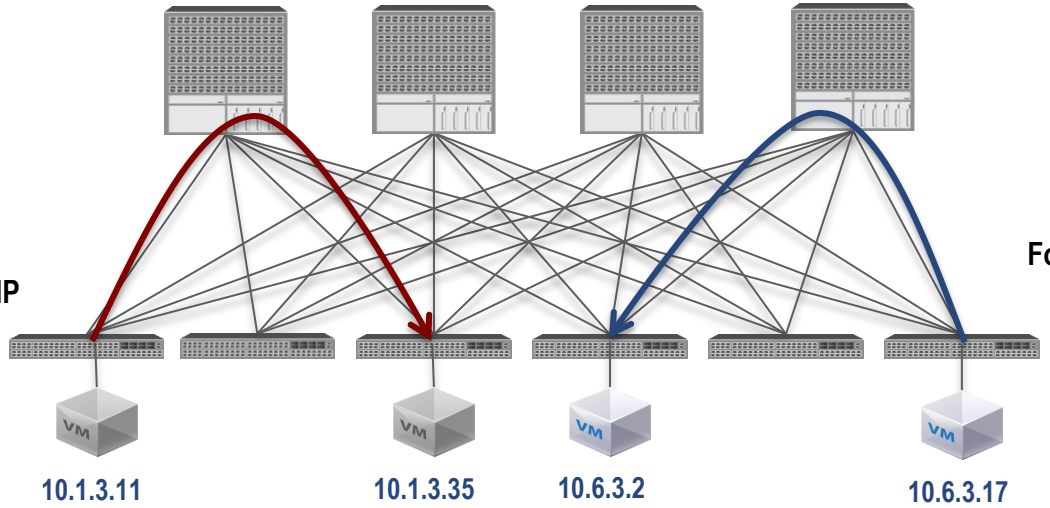
- Fabric Line Card leverage Cisco ASIC's (2 x 24nm ASIC for 36 x 40 line rate ports)
- Support > 1,000,000 IPv6 host routes (16 slot chassis)

ACI - Host Routed Fabric

Layer 2 and Layer 3

IP Forwarding:

Forwarded using DIPi address, HW learning of IP address



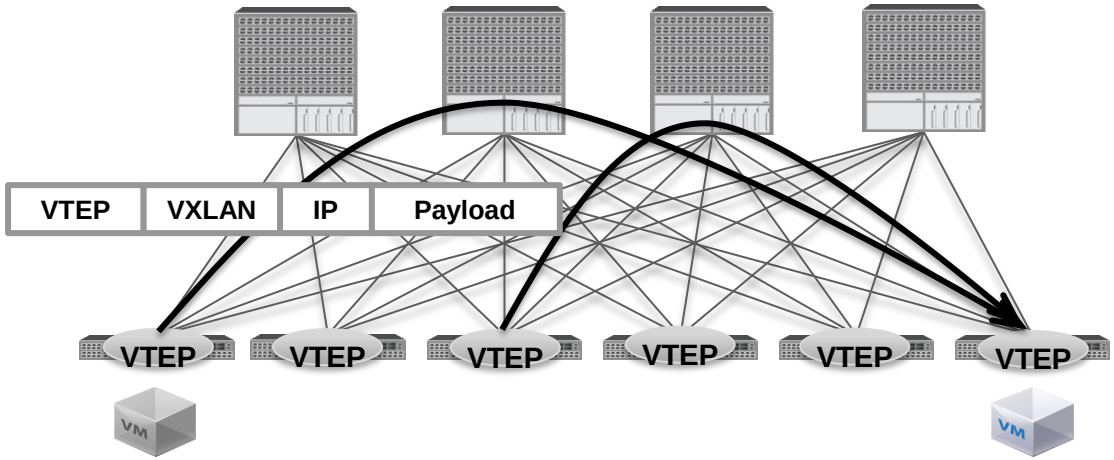
MAC Forwarding:

Forwarded using DMAC address, HW learning of MAC address

- Forward based on destination IP Address for intra and inter subnet (Default Mode)
  - Bridge semantics are preserved for intra subnet traffic (no TTL decrement, no MAC header rewrite, etc.)
  - Non-IP packets will be forwarded using MAC address. Fabric will learn MAC's for non-IP packets, IP address learning for all other packets
- Route if MAC is router-mac, otherwise bridge (standard L2/L3 behaviour)

ACI Fabric

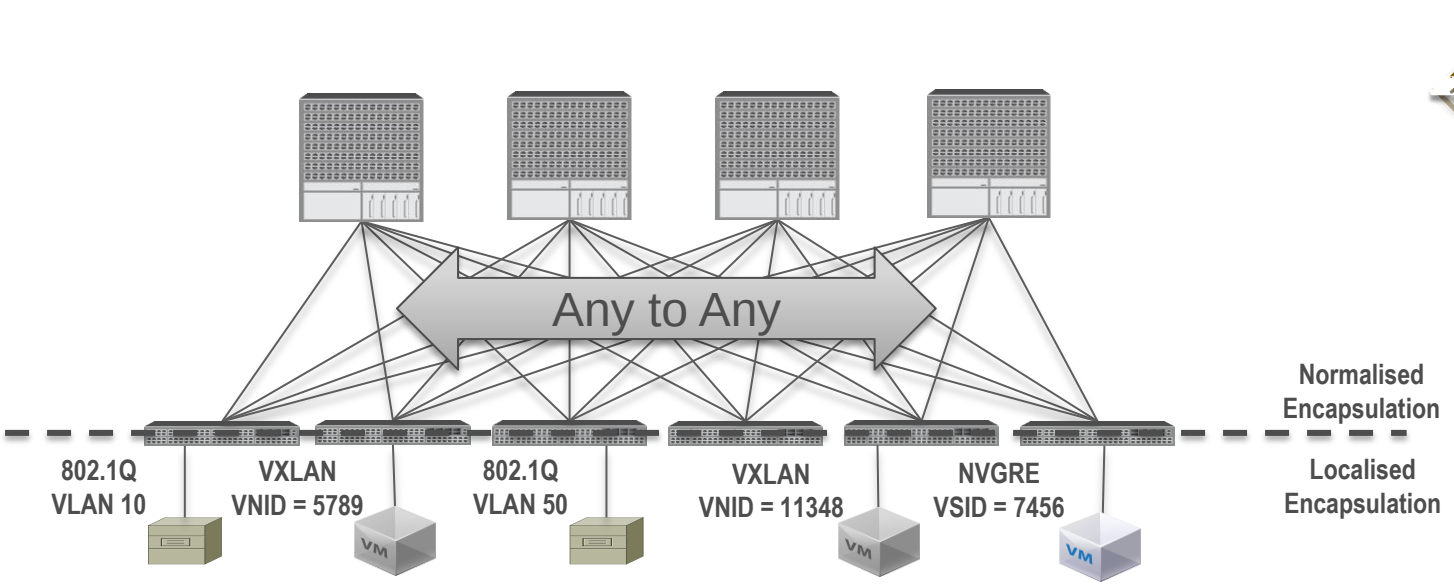
Decoupled Identity, Location & Policy



- ACI Fabric decouples the tenant end-point address, it's "identifier", from the location of that end-point which is defined by it's "locator" or VTEP address
- Forwarding within the Fabric is between VTEPs (VXLAN tunnel endpoints) and leverages an extender VXLAN header format referred to as the VXLAN policy header
- The mapping of the internal tenant MAC or IP address to location is performed by the VTEP using a distributed mapping database

Physical, Any Virtual and Distributed

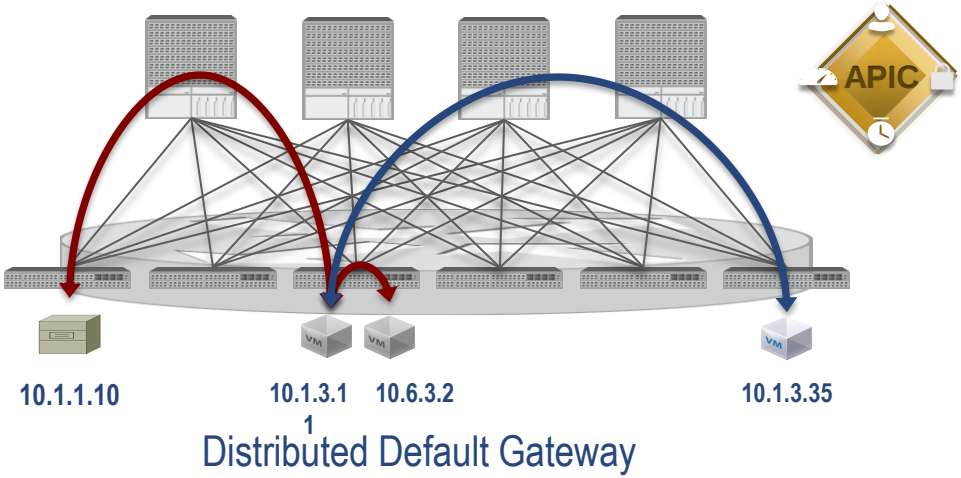
Encapsulation Normalisation

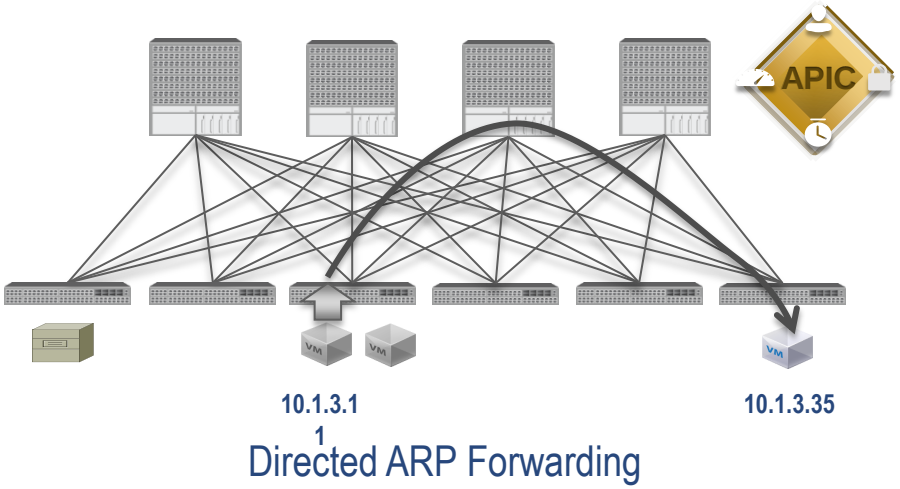


Forwarding is ‘not’ limited to nor constrained by the encapsulation type or encapsulation specific ‘overlay’ network

Location Independent Forwarding

Layer 2 and Layer 3

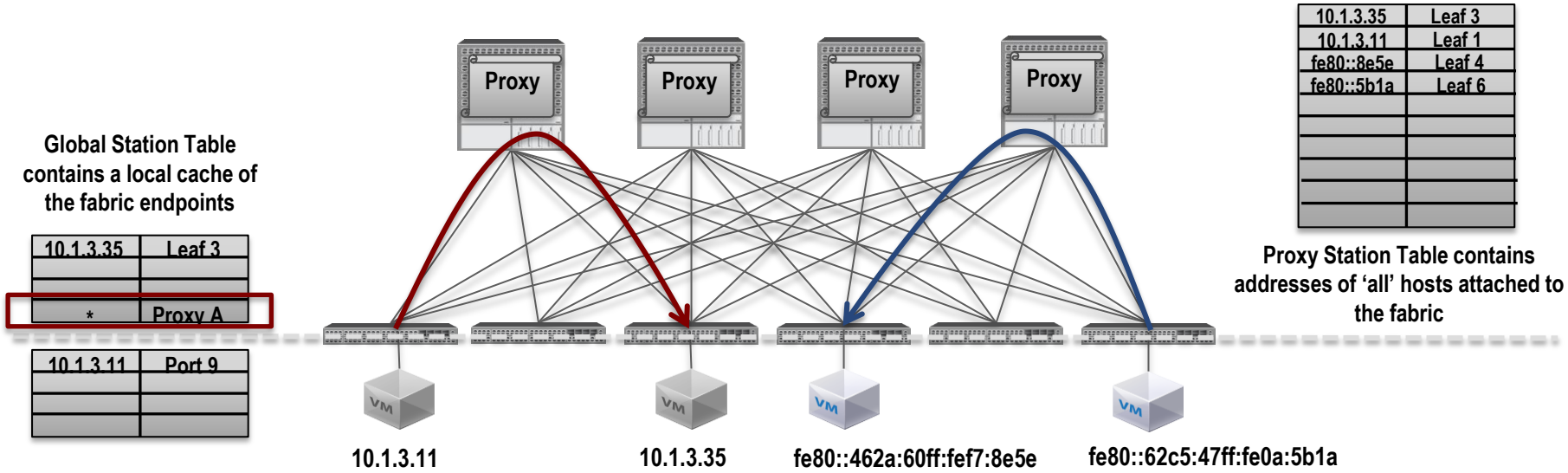




- ACI Fabric supports full layer 2 and layer 3 forwarding semantics, no changes required to applications or end point IP stacks
- ACI Fabric provides optimal forwarding for layer 2 and layer 3
  - Fabric provides a pervasive SVI which allows for a distributed default gateway
  - Layer 2 and layer 3 traffic is directly forwarded to destination end point
- IP ARP/GARP packets are forwarded directly to target end point address contained within ARP/GARP header (elimination of flooding)

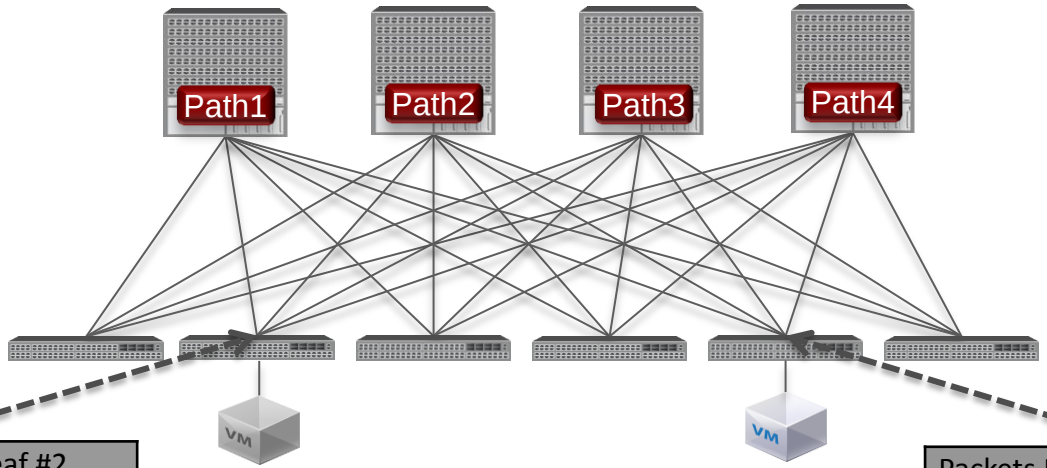
Host Routed Fabric

Inline Hardware Mapping DB - 1,000,000+ hosts



- The Forwarding Table on the Leaf Switch is divided between local (directly attached) and global entries
- The Leaf global table is a cached portion of the full global table
- If an endpoint is not found in the local cache the packet is forwarded to the 'default' forwarding table in the spine switches (1,000,000+ entries in the spine forwarding table)

Telemetry
Atomic Counters



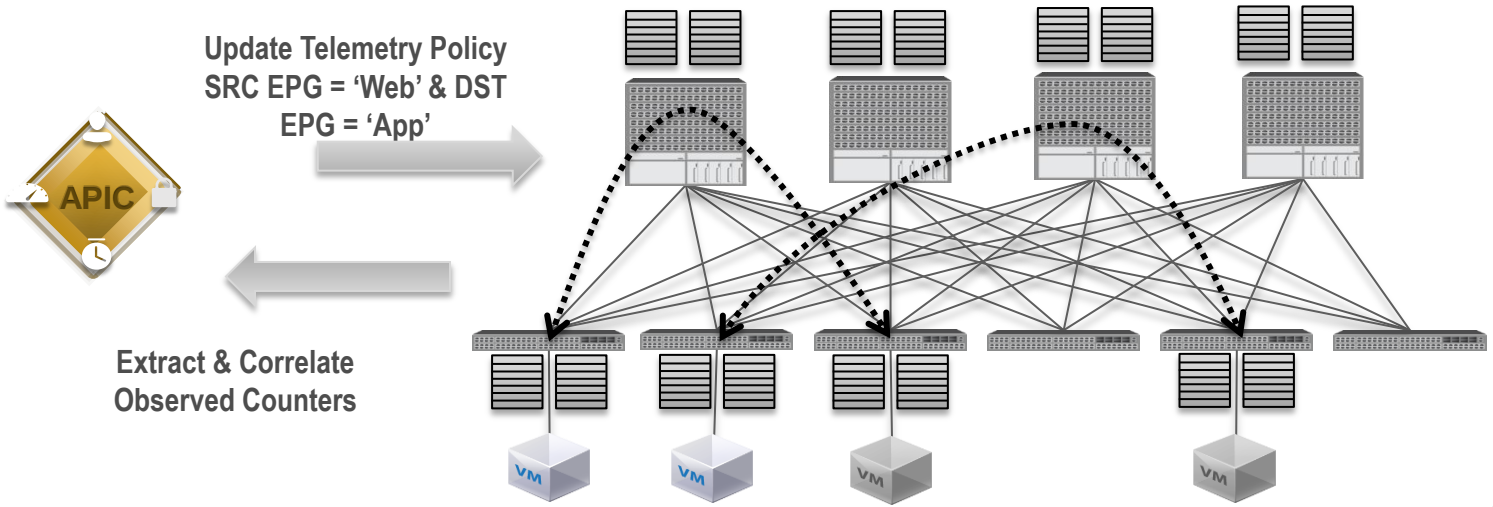
Packets sent from Leaf #2 to Leaf #5	
Path 1	2068
Path 2	2963
Path 3	2866
Path 4	2506

Difference	
Path 1	2
Path 2	0
Path 3	-3
Path 4	0

Packets Received on Leaf #5 sent from Leaf #2	
Path 1	2066
Path 2	2963
Path 3	2869
Path 4	2506

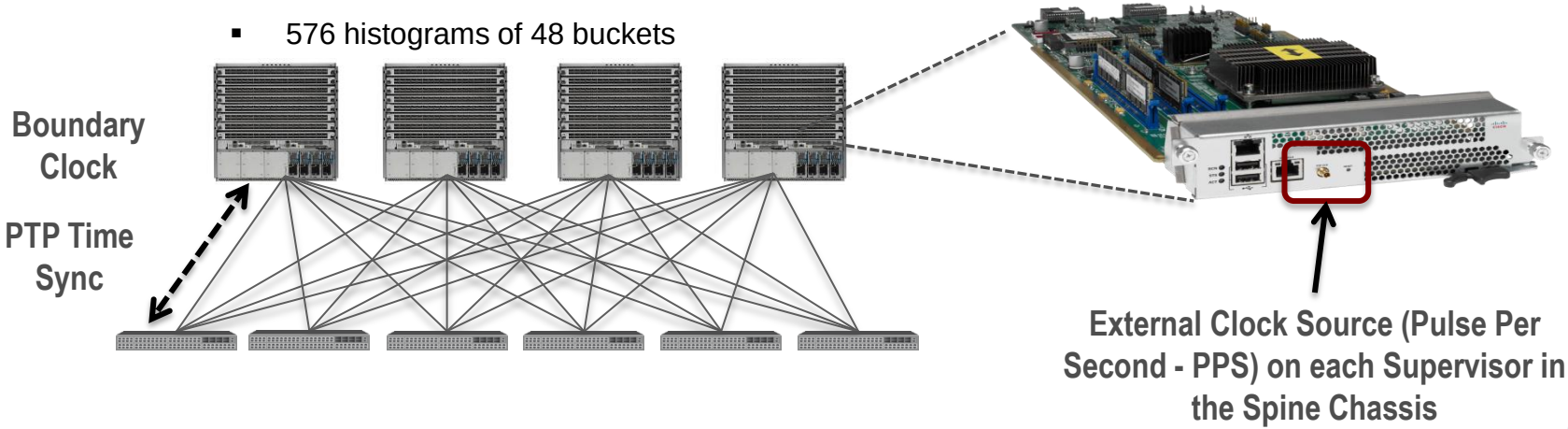
Telemetry
Filter Based Atomic Counters

- A second Bank of counters are used for on-demand monitoring
- Counters are incremented if a programmed TCAM entry is matched & the odd/even bit is set
- TCAM match is programmed via policy on the APIC and distributed to all nodes
  - Criteria to match against: EPG, IP Address, TCP/UDP port, Tenant VRF or Bridge Domain



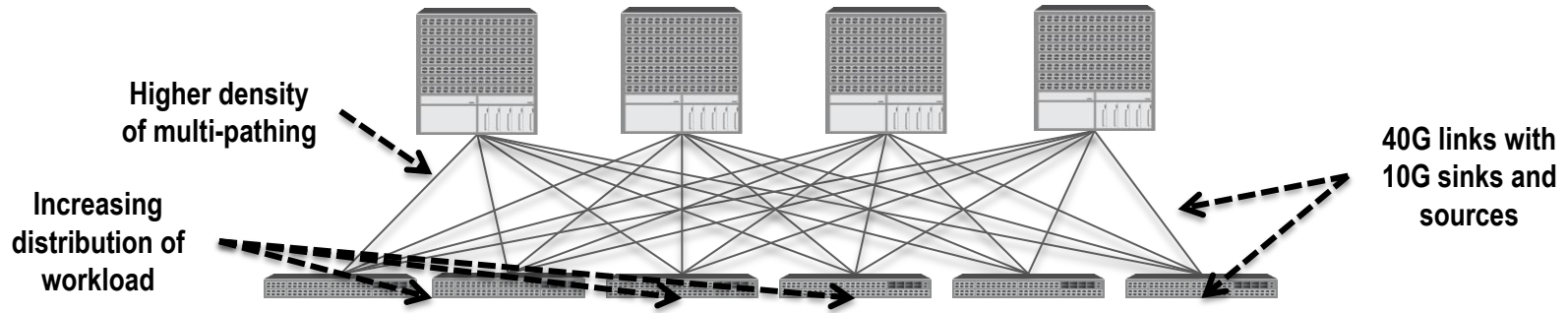
Telemetry
Fabric Latency Measurements

- Matrix of Latency Measurements between all Leaves
  - Per Port Average Latency & Variance to up to 576 other iLeaves
    - Maximum, Accumulation, Sum of Square and Packet Count
  - Per Port 99% Latency (recorded to up to 576 other iLeaves)
    - 99% of all packets have recorded latency less than this value
  - 48 bucket histogram
    - 576 histograms of 48 buckets



ACI Fabric

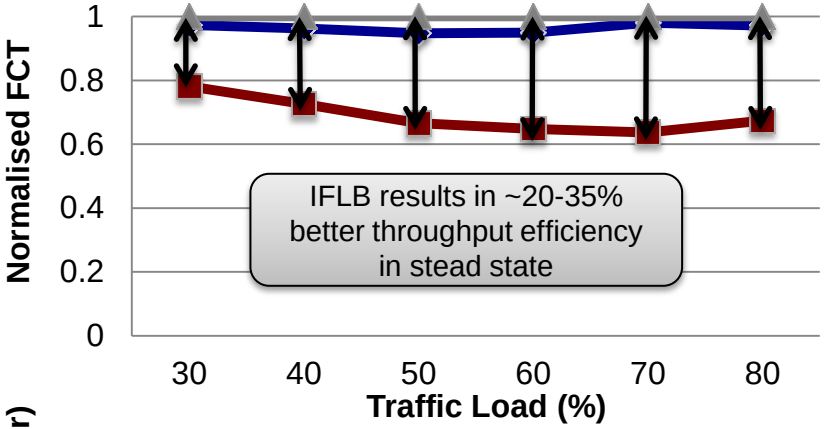
Why Focus on Next Generation DC QoS

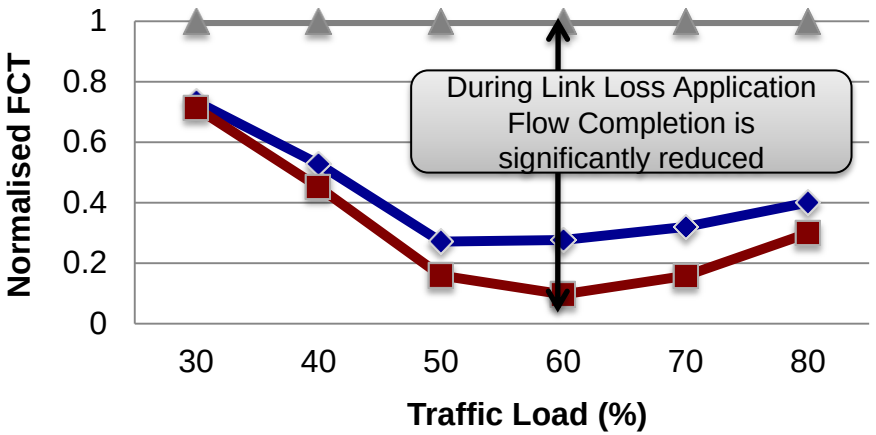


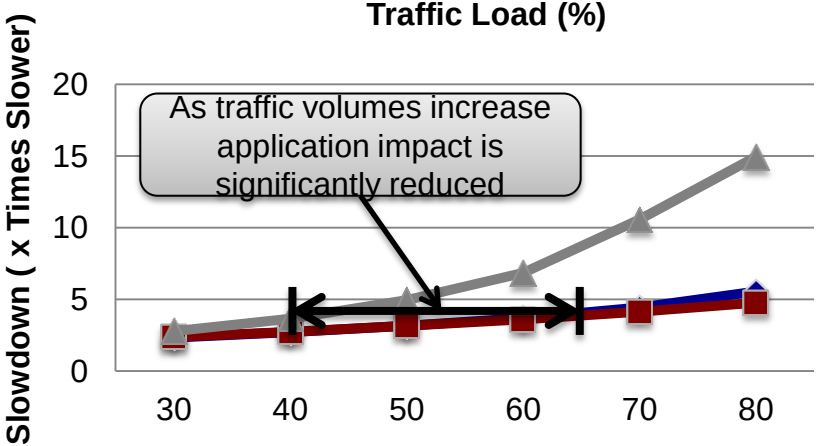
- Topology and traffic pattern changes require us to re-evaluate the assumptions of congestion management within the Data centre
  - Higher density of uplinks with greater multi-pathing ratio is resulting in more variability in congestion patterns
  - Distribution of workload is adding another dimension of traffic patterns
- Two options
  - Spend the time to statically engineering marking, queuing and traffic patterns to accommodate these new
  - Build a more systems based reactive approach to congestion management for traffic within the Data centre

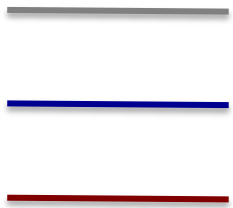
Application Performance Improvements

ACI Fabric Load Balancing









Standard ECMP with No Priority

ECMP 'with' Priority

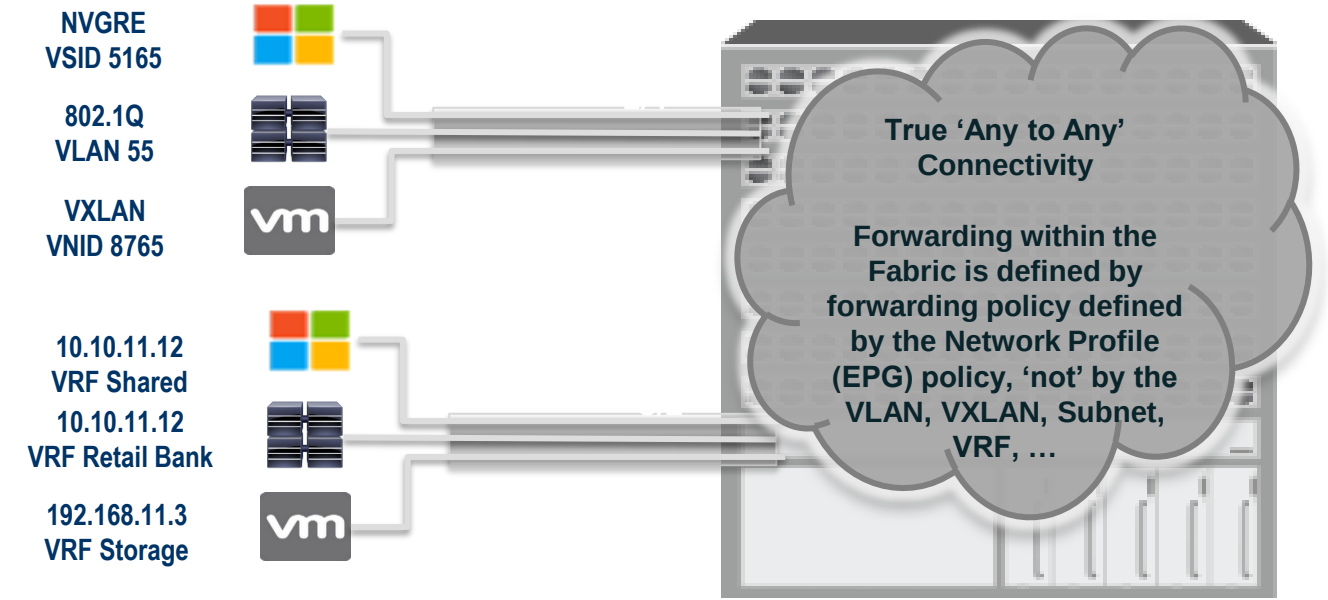
Dynamic Load Balancing with Priority

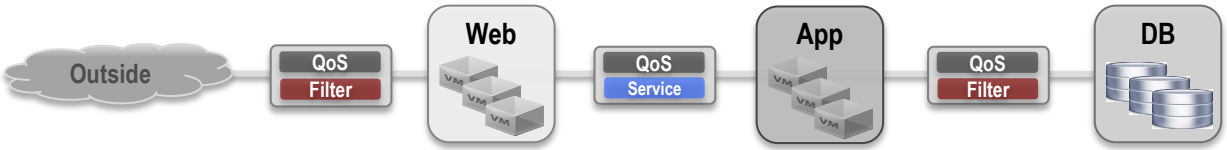
Fabric Infrastructure

Endpoint Based Forwarding with Distributed Policy

All single port can support all encapsulations simultaneously

Forwarding is defined by Policy EPG ‘Web’ can talk to EPG ‘DB’ independent of IP subnet, VLAN/VXLAN, VRF is Policy says it should

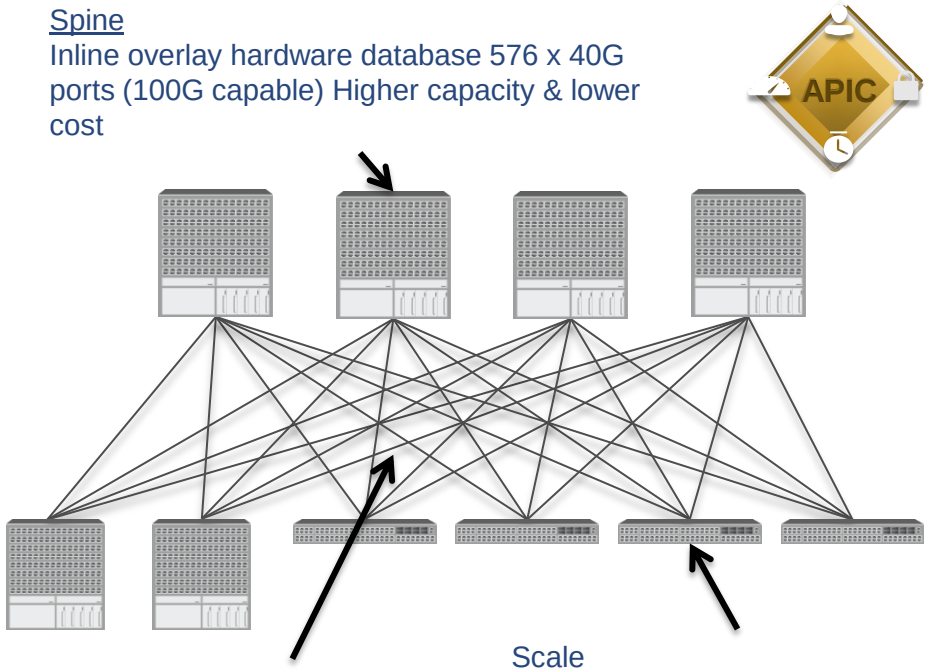




ACI - Based on a Better Network

ACI Fabric

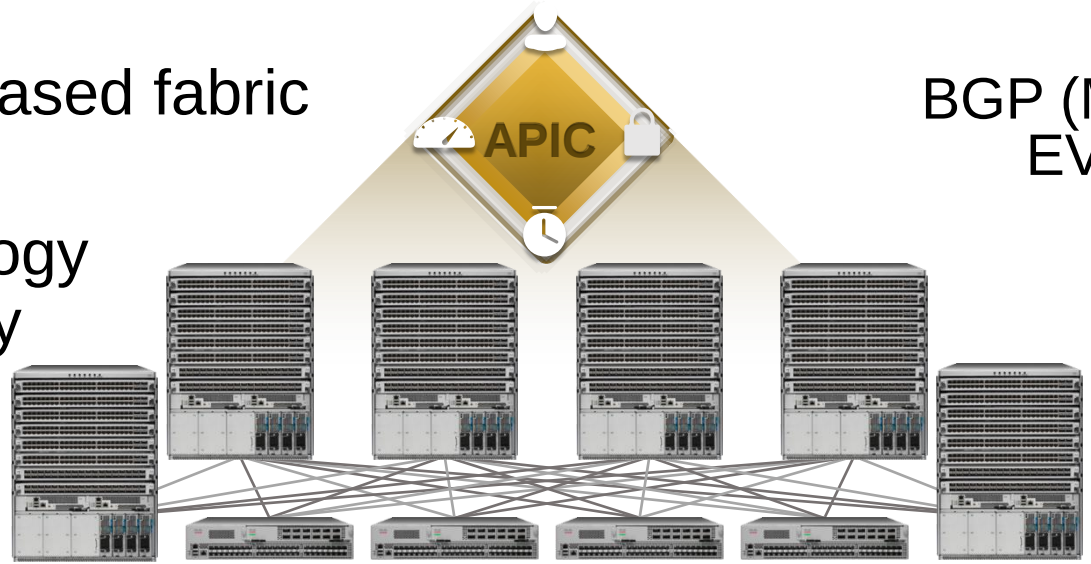
- Industry's most efficient fabric
  - 1/10G edge - High density 40G spine (100G capable)
  - 1M+ IPv4 & IPv6 endpoints
  - 64K+ Tenants
  - 55K+ 1/10G Hosts in a single tier 3:1 oversubscribed Fabric
- Routed fabric – Optimal IP Forwarding
  - Bridging (L2) *and* Routing (L3) of VXLAN/NVGRE/VLAN at scale
  - No x86 GW's – Physical & Virtual
  - Application Agility – Place & Join without limits in Fabric
- Full visibility into virtual and physical
- Common operations from Hypervisor to Compute, To Fabric, to WAN



Standards Based Architecture

VXLAN-based fabric

IS-IS Topology Discovery



BGP (MP-BGP EVPN)

ACI Fabric Attached Device Protocol*

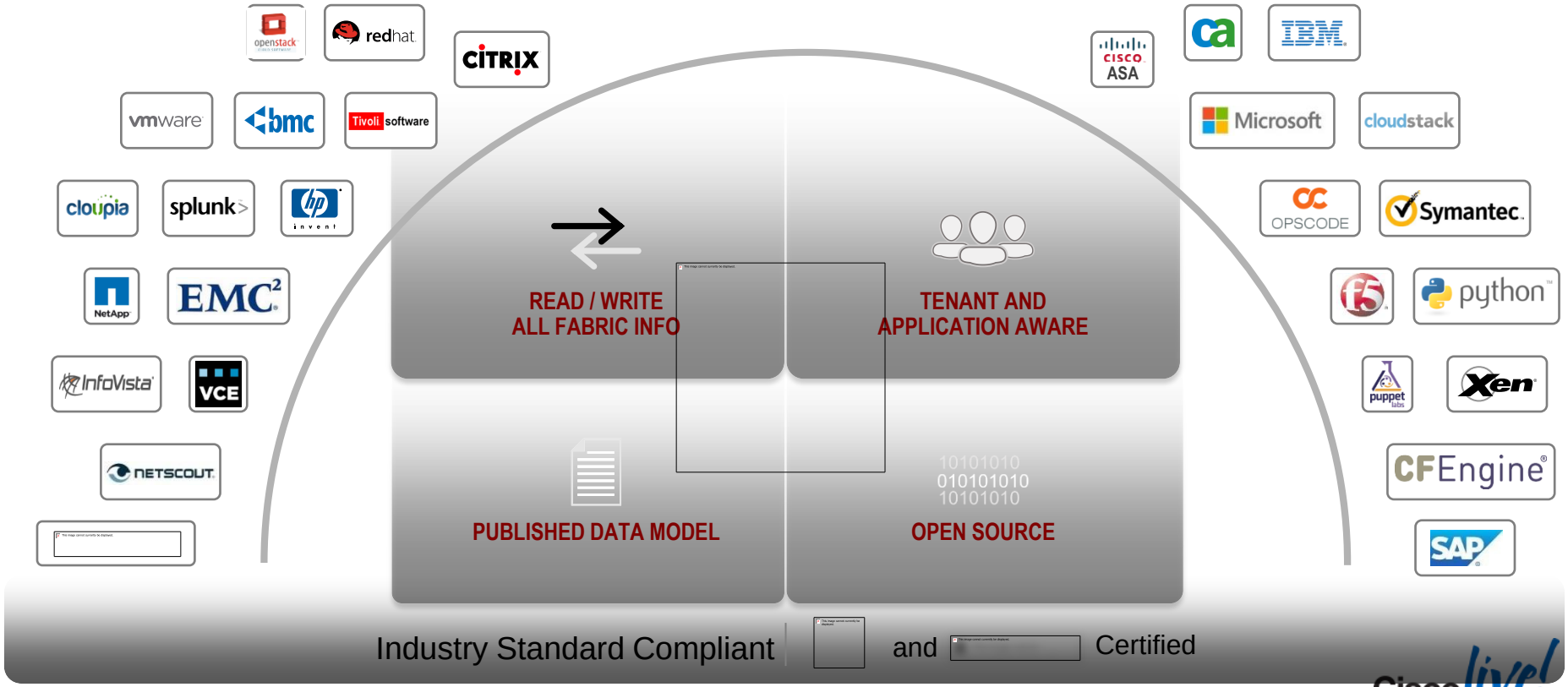
Network Service Header (NSH) Protocol*

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

Open Ecosystem, Open APIs, Open Source

Comprehensive access to underlying information model






Q & A

Complete Your Online Session Evaluation

Give us your feedback and receive a Cisco Live 2014 Polo 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 www.ciscoliveaustralia.com/mobile
- Visit any Cisco Live Internet Station located throughout the venue

Polo Shirts can be collected in the World of Solutions on Friday 21 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


CISCO
TM