We’re ready.
Are you?

Nexus 9000 Architecture

Mike Herbert - Principal Engineer, Cisco

Cisco live!

The Original Agenda

- This intermediate level session will describe the Cisco Nexus 9000 architecture and innovations in terms of hardware, software, mechanical design, optical advantages in the 40 GE environment and power budget. The unique combination of Merchant silicon combined with Cisco internally developed ASICs make this platform a leader in the Data Centre switch market. This session will also approach the Data Centre design aspect and describe the Spine-Leaf architecture advantages.

The new Agenda – It is still is the N9K Architecture Session but with details on next Generation as well

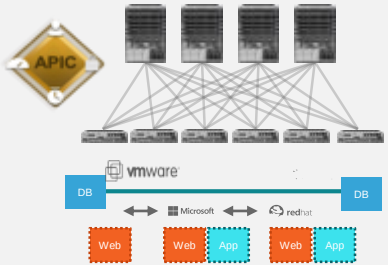
In the upcoming year, 2016, the industry will see a significant capacity, capability and cost point shift in Data Centre switching. The introduction of 25/100G supplementing the previous standard of 10/40G at the same cost points and power efficiency which represents a 250% increase in capacity for roughly the same capital costs is just one example of the scope of the change. These changes are occurring due to the introduction of new generations of ASICs leveraging improvements in semiconductor fabrication combined with innovative developments in network algorithms, SerDes capabilities and ASIC design approaches. This session will take a deep dive look at the technology changes enabling this shift and the architecture of the next generation nexus 9000 Data Centre switches enabled due to these changes. Topics will include a discussion of the introduction of 25/50/100G to compliment existing 10/40G, why next generation fabrication techniques enable much larger forwarding scale, more intelligent buffering and queuing algorithms and embedded telemetry enabling big data analytics based on network traffic.

Agenda

- Existing and New Nexus 9000 & 3000
- What's New
  - Moore's Law and 25G SerDes
  - The new building blocks (ASE-2, ASE-3, LSE)
  - Examples of the Next Gen Capabilities
- Nexus 9000 Switch Architecture
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)
- 100G Optics

Cisco Data Centre Networking Strategy:
Providing Choice in Automation and Programmability

Application Centric Infrastructure

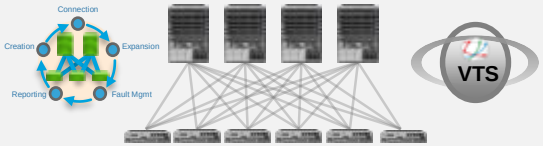


Turnkey integrated solution with security, centralised management, compliance and scale

Automated application centric-policy model with embedded security

Broad and deep ecosystem

Programmable Fabric

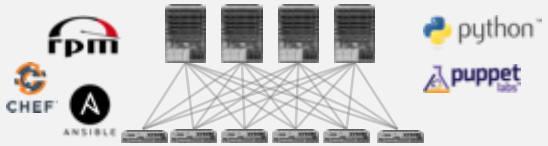


VxLAN-BGP EVPN standard-based

3rd party controller support

Cisco Controller for software overlay provisioning and management across N2K-N9K

Programmable Network



Modern NX-OS with enhanced NX-APIs

DevOps toolset used for Network Management
(Puppet, Chef, Ansible etc.)

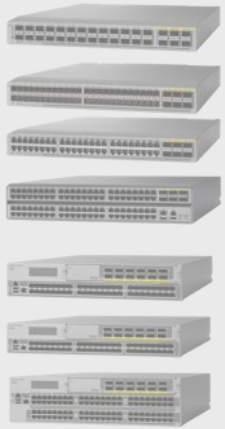
Nexus 9400 (line cards), 9200, 3100, 3200

Nexus 9700EX + 9300EX

Nexus 9000 Portfolio
10/25/40/50/100G on Merchant or Cisco Silicon

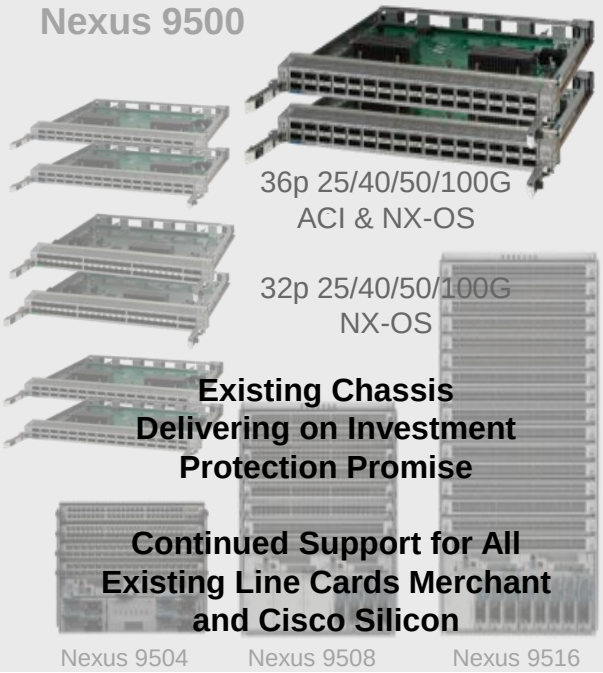
Over 6000 Nexus 9K Customers

Nexus 9300



48p 10G & 6p 40G
96p 10G & 6p 40G
32p 40G

Nexus 9500



36p 25/40/50/100G
ACI & NX-OS

32p 25/40/50/100G
NX-OS

Existing Chassis Delivering on Investment Protection Promise

Continued Support for All Existing Line Cards Merchant and Cisco Silicon

Nexus 9504 Nexus 9508 Nexus 9516

Nexus 9300EX

Industry Only 25G Native VXLAN

48p 10/25G SFP & 6p 40/50/100G
48p 10GT & 6p 40/50/100G

Nexus 9200

Industry Only 25G Native VXLAN

36p wire rate 40/50/100G
56p 40G + 8p 40/50/100G
72p 40G
48p 10/25G SFP & 4p 40/50/100G + 2p 40G

Continued Support of Broadcom Silicon
Nexus 3000: 10 Million Ports Shipped

Nexus 3100

64p 40G

32p 40G

48p 10G & 6p 40G

48p 1G & 4p 10G

Nexus 3100V

32p 40G

48p 10G & 6p 100G

VXLAN routing, 100G uplinks, No 25G

T2+

Nexus 3200

Shipping for 3+ months

32p 25/50/100G

64p 40G Single Chip

VXLAN bridging, **25/100G**

Tomahawk

Single NX-OS Image for Nexus 3000 & Nexus 9000

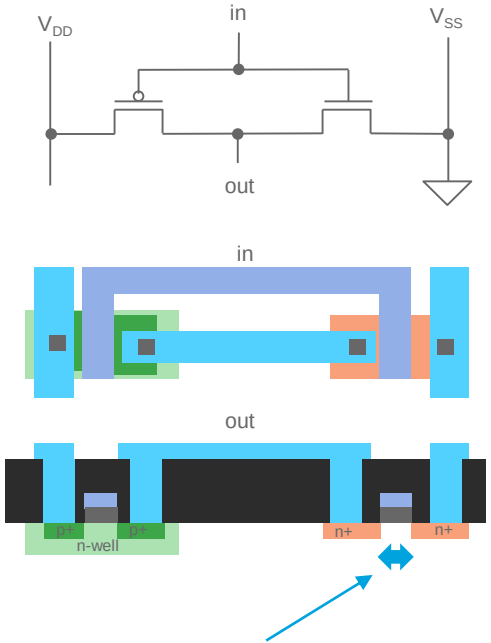
Agenda

- Existing and New Nexus 9000 & 3000
- What's New
  - Moore's Law and 25G SerDes
  - The new building blocks (ASE-2, ASE-3, LSE)
  - Examples of the Next Gen Capabilities
- Nexus 9000 Switch Architecture
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)
- 100G Optics

"The number of transistors incorporated into a chip will approximately double every 24 months …"

"Moore's Law" - 1975

Moore’s Law
CMOS



“Feature size”

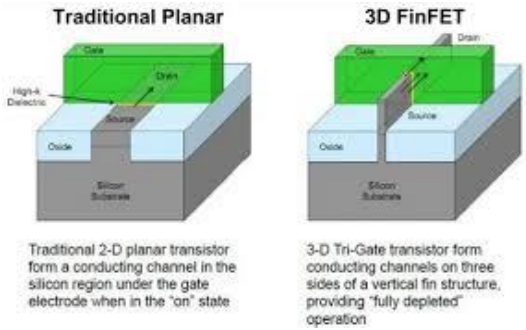
This dimension is what Moore’s Law is all about !!

Moore's Law

It's all about the Economics

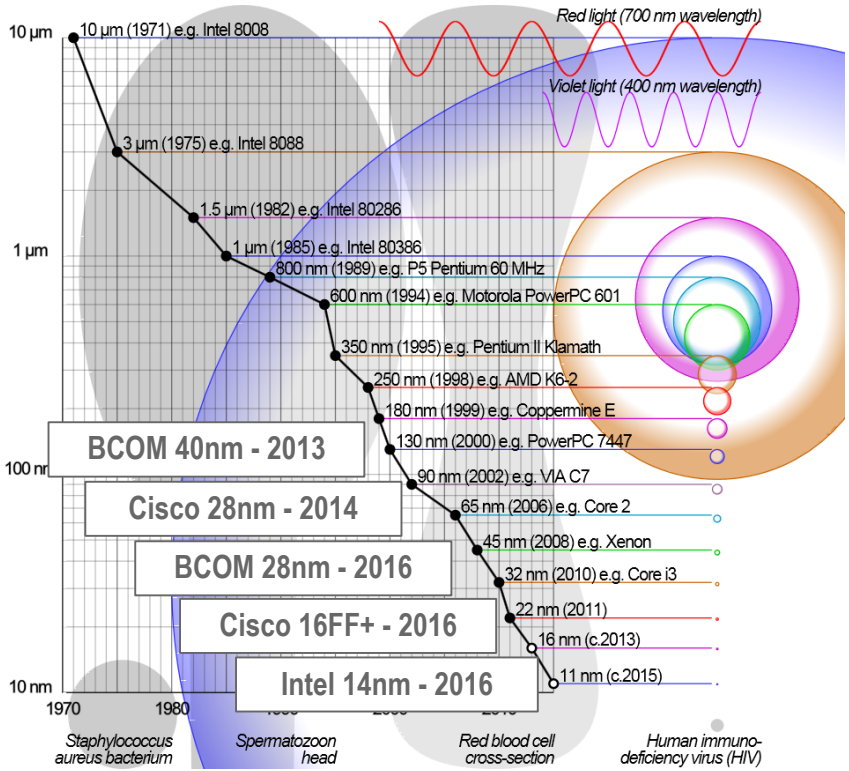
- Increased function, efficiency
- Reduced costs, power
- ~ 1.6 x increase in gates between process nodes

The new generation of Nexus 9000 is leveraging 16nm FF+ (FinFet)



Traditional 2-D planar transistor form a conducting channel in the silicon region under the gate electrode when in the "on" state

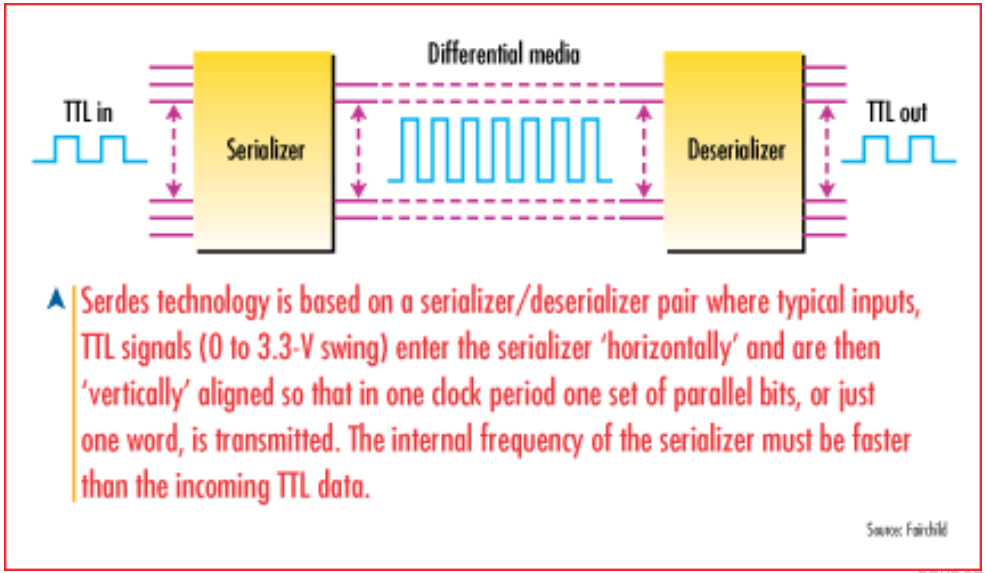
3-D Tri-Gate transistor form conducting channels on three sides of a vertical fin structure, providing "fully depleted" operation



http://en.wikipedia.org/wiki/Semiconductor_device_fabrication

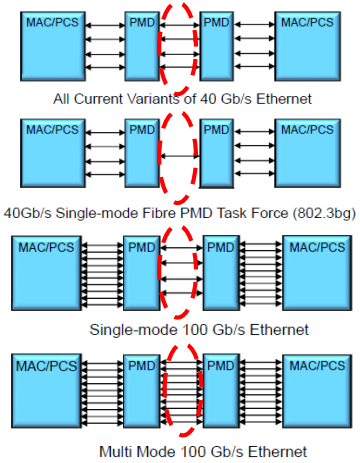
SerDes: Serializer + Deserializer

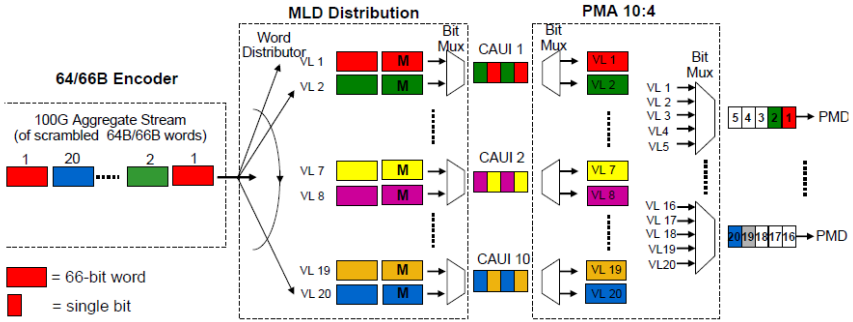
- SerDes Clocking Increases
  - 10.3125G (40G, 10G)
  - ***25.78125(25G/50G/100G) - 2016***



Multi Lane Distribution (MLD)

MLD (Multi Lane Distribution)

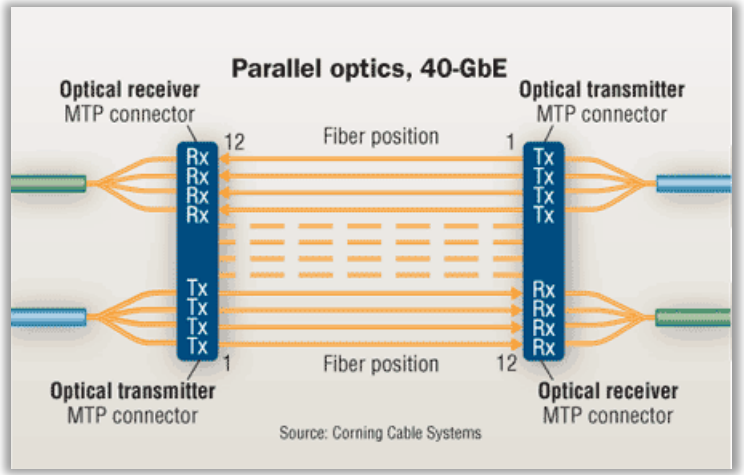




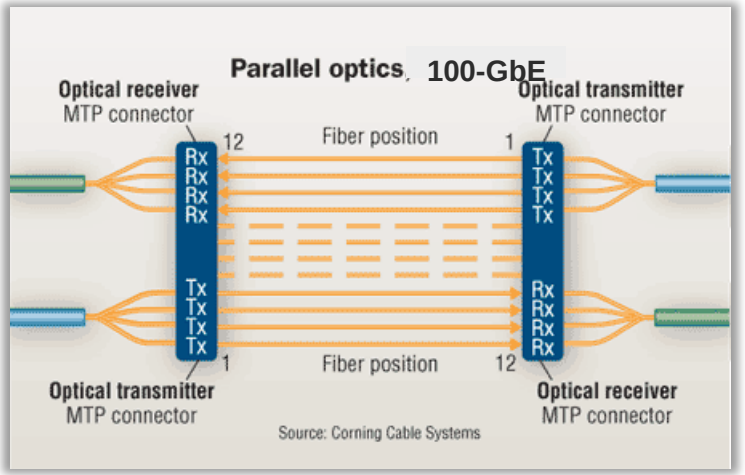
- 40GE/100GE interfaces have multiple lanes (coax cables, fibres, wavelengths)
- MLD provides a simple (common) way to map 40G/100G to physical interfaces of different lane widths

Parallel Lanes

4 x10 = 40G shifts to 4 x 25 = 100G

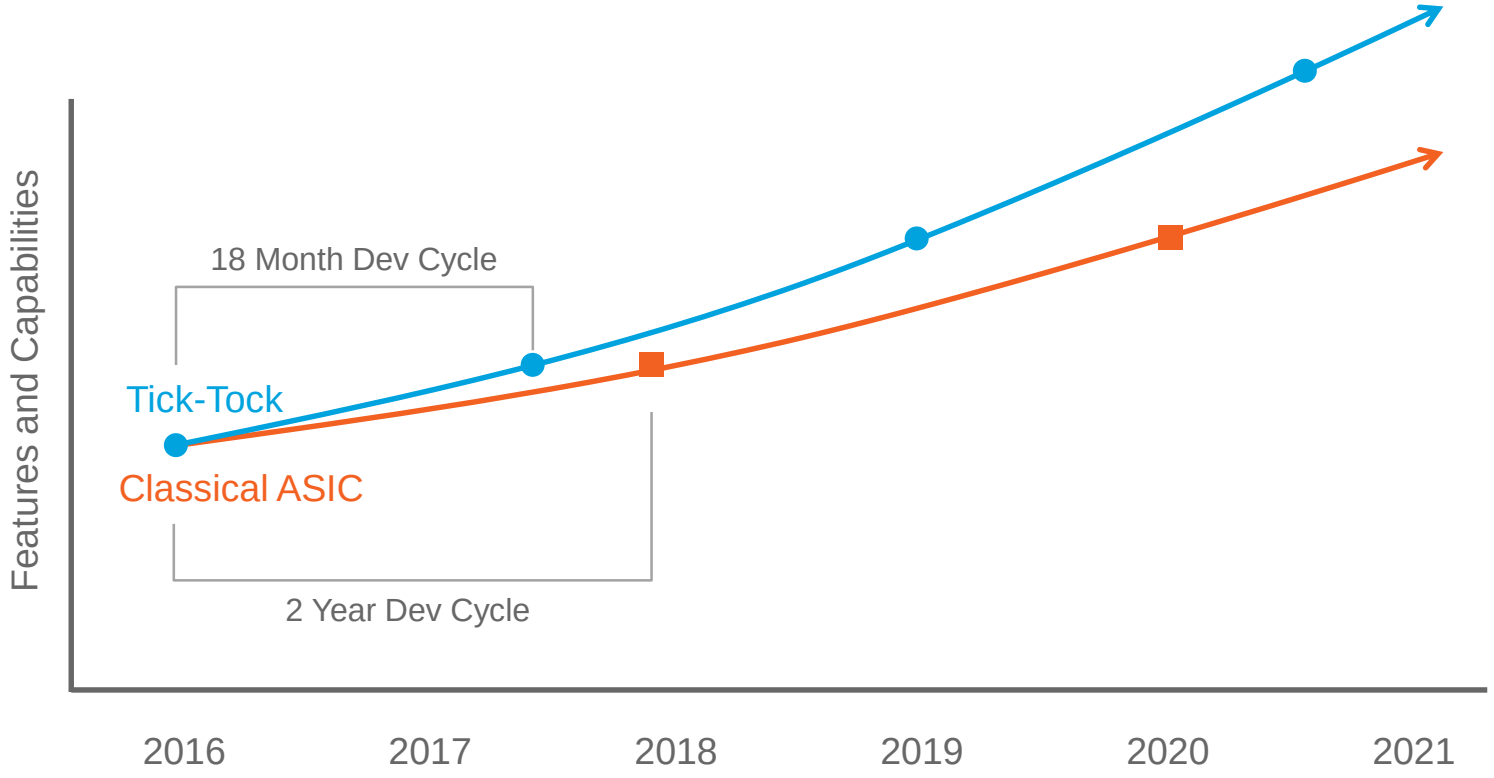


Backed by 10G SerDes



Backed by 25G SerDes

Development Cycle Decreasing
Time to Leverage Moore's Law is Reducing



ASIC Used by Nexus 3000/9000

Scale
- Route/ Host tables
- Sharding
- Encap normalisation
- EPG/ SGT/ NSH

Telemetry
- Analytics
- Netflow
- Atomic Counters

Optimisation
- Smart Buffers
- DLB/ Flow Prioritisation

ASIC Used by Nexus 3000/9000

16nm

- ASE2 – ACI Spine Engine 2
- 3.6 Tbps Forwarding (Line Rate for all packet sizes)
  - 36x100GE, 72x40GE, 144x25GE, ...

- ASE3 – ACI Spine Engine 3
- 1.6 Tbps Forwarding (Line Rate for all packet sizes)
- 16x100GE, 36x40GE, 74x25GE, ...
- Flow Table (Netflow, …)

- Standalone leaf and spine, ACI spine
- 16K VRF, 32 SPAN, 64K MCAST fan-outs, 4K NAT
- MPLS: Label Edge Router (LER), Label Switch Router (LSR), Fast Re-Route (FRR), Null-label, EXP QoS classification
- Push /Swap maximum of 5 VPN label + 2 FRR label
- 8 unicast + 8 Multicast
- Flexible DWRR scheduler across 16 queues
- Active Queue Management
  - AFD ,WRED, ECN Marking
- Flowlet Prioritisation & Elephant-Trap for trapping 5 tuple of large flows

ASIC Used by Nexus 3000/9000

16nm

- LSE – Leaf Spine Engine
- Standalone leaf & spine, ACI leaf and spine
- Flow Table (Netflow, …)
- ACI feature and service and security enhancement
- 40MB Buffer
- 32G fibre channel and 8 unified port
- 25G and 50G RS FEC (clause 91)
- Energy Enhancement Ethernet, IEEE 802.3az
- Port TX SPAN support for multicast
- MPLS: Label Edge Router (LER), Label Switch Router (LSR), Fast Re-Route (FRR), Null-label, EXP QoS classification
- Push /Swap maximum of 5 VPN label + 2 FRR label
- 16K VRF, 32 SPAN, 64K MCAST fan-outs, 50K NAT
- 8 unicast + 8 Multicast
- Flexible DWRR scheduler across 16 queues
- Active Queue Management
  - AFD ,WRED, ECN Marking
- Flowlet Prioritization, Elephant-Trap for trapping 5 tuple of large flows

ASIC Used by Nexus 3000/9000

Merchant
28nm

- Broadcom Tomahawk
- 3.2 Tbps I/O & 2.0 Tbps Core

  Tomahawk supports 3200 Gbps when average packet size is greater than 250 bytes. When all ports are receiving 64 byte packets, throughput is 2000 Gbps
- 32 x 100GE
- Standalone leaf and spine
- VXLAN Bridging

- Broadcom Trident 2+
- 1.28Tbps I/O & 0.96T Core (< 192B pkt)
  - 32 x 40GE (line rate for 24 x 40G)
- Standalone leaf and spine
- VXLAN Bridging & Routing (with-out recirculation)

Cisco Nexus 3000/9000 ASIC Mapping

ASIC	Fixed Platform	Modular Platform
ALE (ACI Leaf Engine)	GEM Module (ACI Leaf/NX-OS) N9K-M12PQ, N9K-M6PQ	(NX-OS) N9K-X9564PX, N9K-X9564TX, N9K-X9536PQ
ALE2	(ACI Leaf/NX-OS) N9K-C9372PX, N9K-C9372TX, N9K-C93120TX, N9K-C9332PQ	NA
ALE2	(ACI Leaf/NX-OS) N9K-C9372PX-E, N9K-C9372TX-E, GEM: N9K-M6PQ-E	NA
ASE (ACI Spine Engine)	(ACI Spine) N9K-C9336PQ	(ACI Spine) N9K-X9736PQ
ASE2	(NX-OS) N9K-C9236C, N9K-C92304QC, N9K-C9272Q	(ACI Spine/NX-OS) N9K-C9504-FM-E, N9K-C9508-FM-E
ASE3	(NX-OS) N9K-C92160YC-X	NA
LSE (Leaf Spine Engine)	(ACI Leaf/NX-OS) N9K-C93180YC-EX, N9K-C93108TC-EX (ACI Spine) N9K-C9372C-EX	(ACI Spine/NX-OS) N9K-X9736C-EX
NFE (Trident T2)	(ACI Leaf/NX-OS) N9K-C9372PX(E), N9K-C9372TX (E), N9K-C93120TX, N9K-C9332PQ, N9K-C9396PX, N9K-C9396TX, N9K-C93128TX GEM Module (NX-OS) N9K-M4PC-CFP2	(NX-OS) N9K-X9564PX, N9K-X9564TX, N9K-X9536PQ, N9K-X9464PX, N9K-X9464TX, N9K-X9432PQ, N9K-X9636PQ
NFE2 (Tomahawk)		(NX-OS) N9K-X9432C-S

Agenda

- Existing and New Nexus 9000 & 3000
- What's New
  - Moore's Law and 25G SerDes
  - The new building blocks (ASE-2, ASE-3, LSE)
  - Examples of the Next Gen Capabilities
- Nexus 9000 Switch Architecture
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)
- 100G Optics

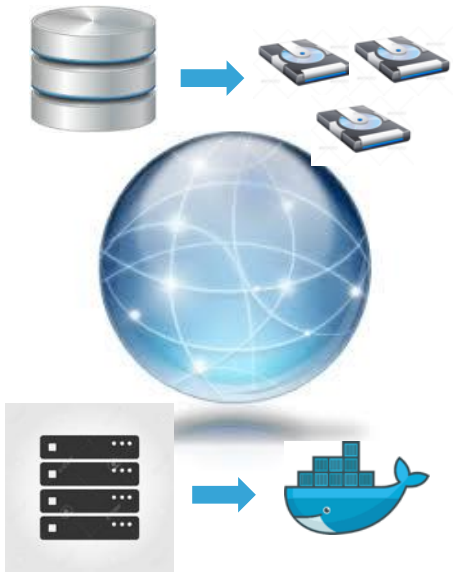
Hyper-Converged Fabrics

Containers, Scale-Out Storage mixed with existing VM and Bare Metal

Distributed IP storage for cloud apps and traditional storage (iSCSI/NAS,FC) for existing apps

Distributed Apps via Container based Micro-services

Inter-process Communication across fabric



Data Centre Implications

Order(s) of Magnitude increase in density of endpoints

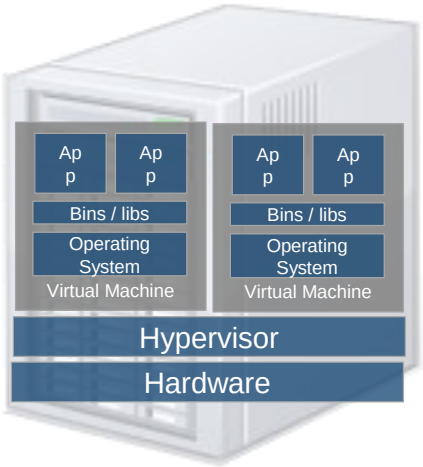
Increased I/O traffic drives larger Bandwidth

Mix of traffic types drives need for better queueing (not buffering)

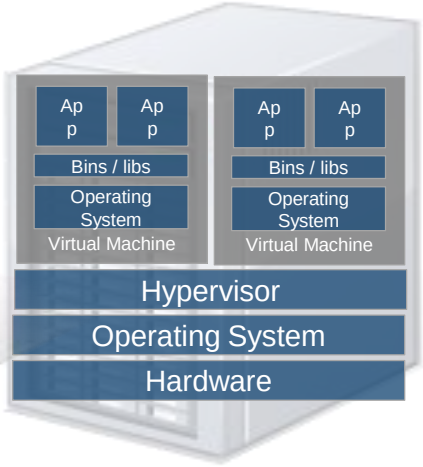
Security Density is Increasing as well

Hypervisors vs. Linux Containers

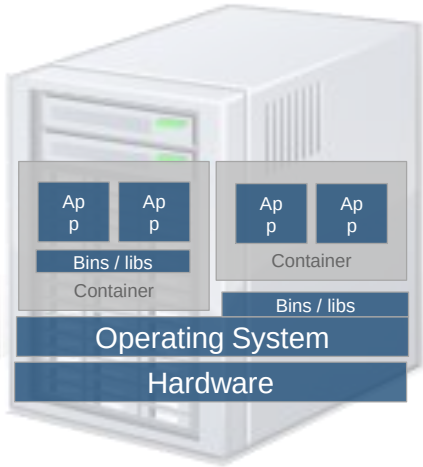
Containers share the OS kernel of the host and thus are lightweight.
However, each container must have the same OS kernel.



Type 1 Hypervisor



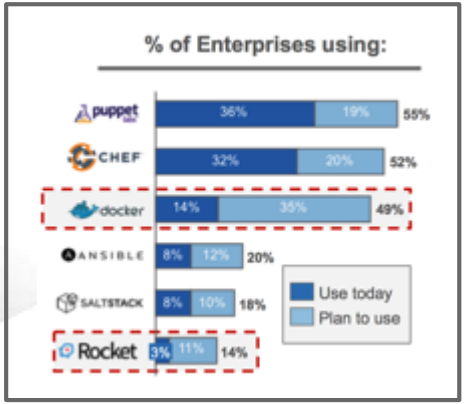
Type 2 Hypervisor



Linux Containers (LXC)

Containers are isolated, but share OS and, where appropriate, libs / bins.

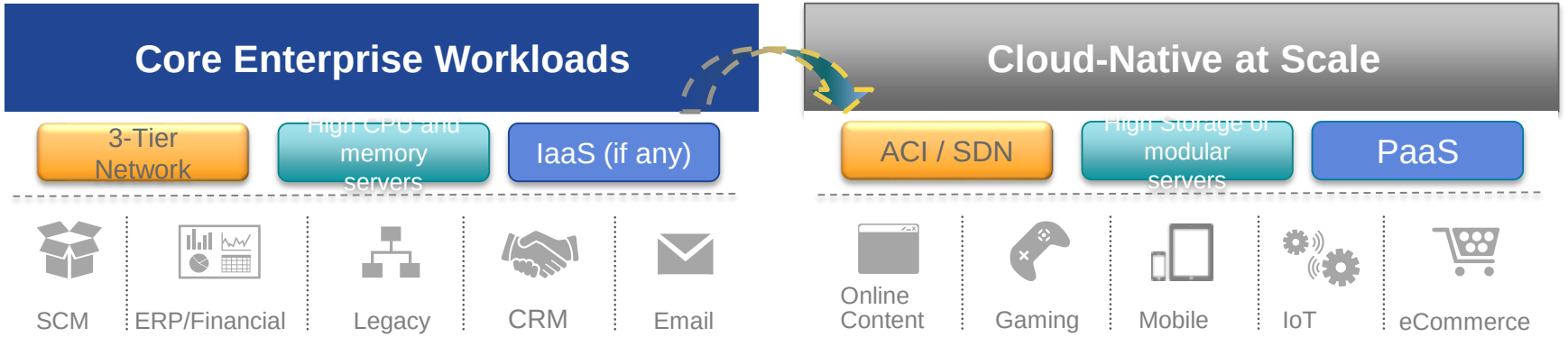
Adoption / plan to use

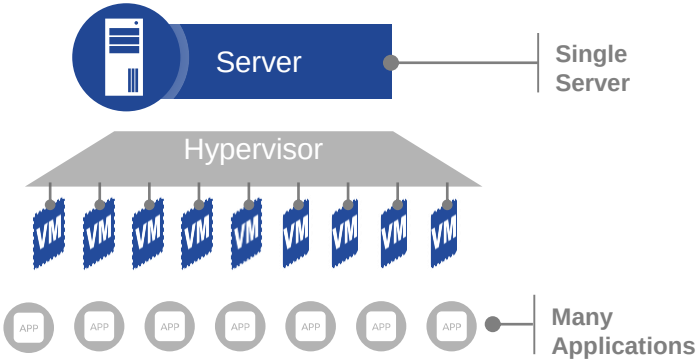


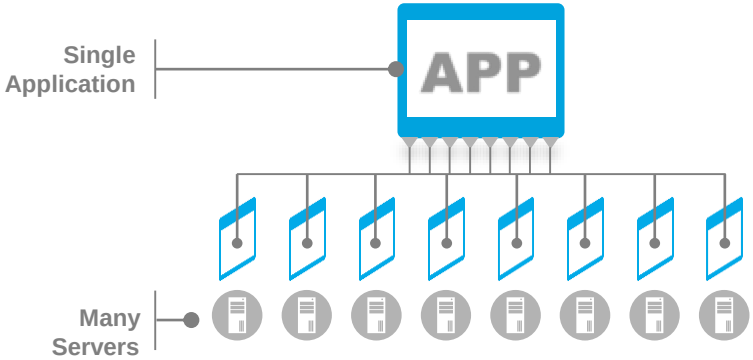
Source: RightScale 2015 State of the cloud report

Ex.: Applications & Software development

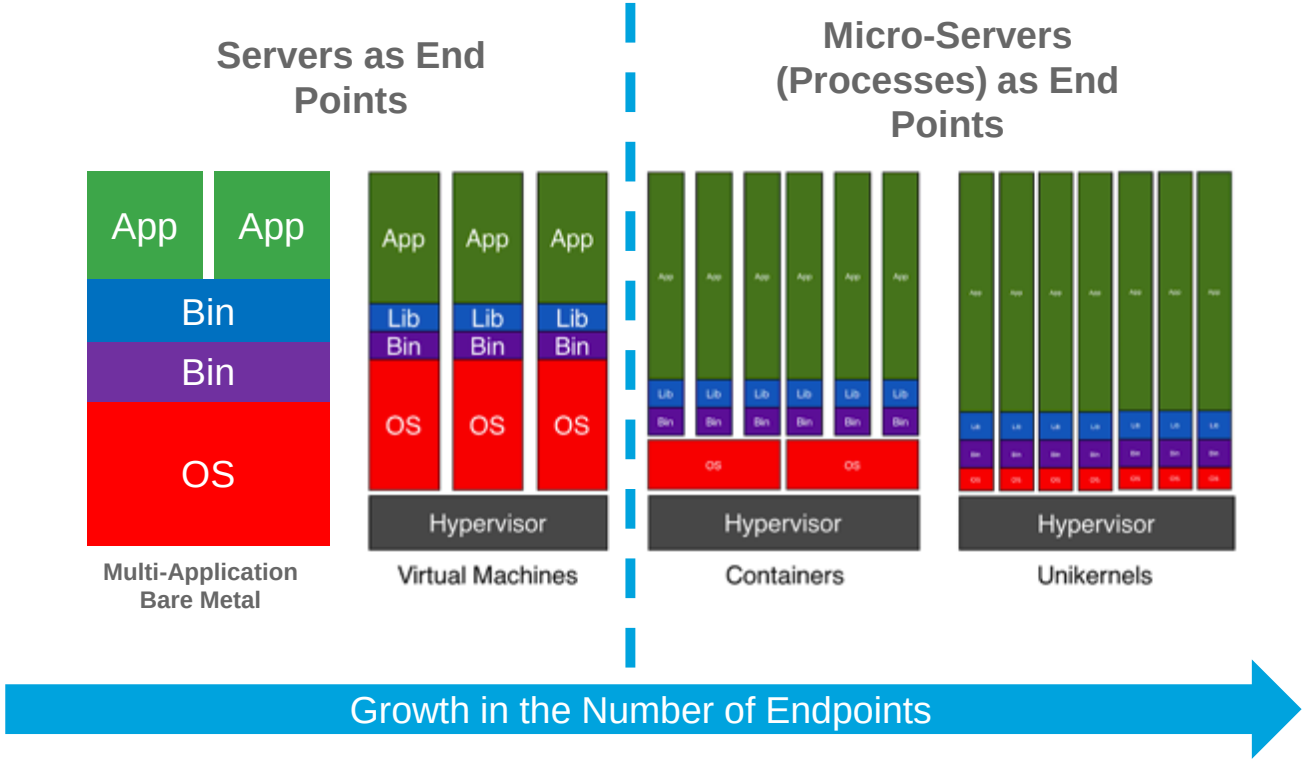
Monolithic Apps versus Cloud-Native App with Distributed Data







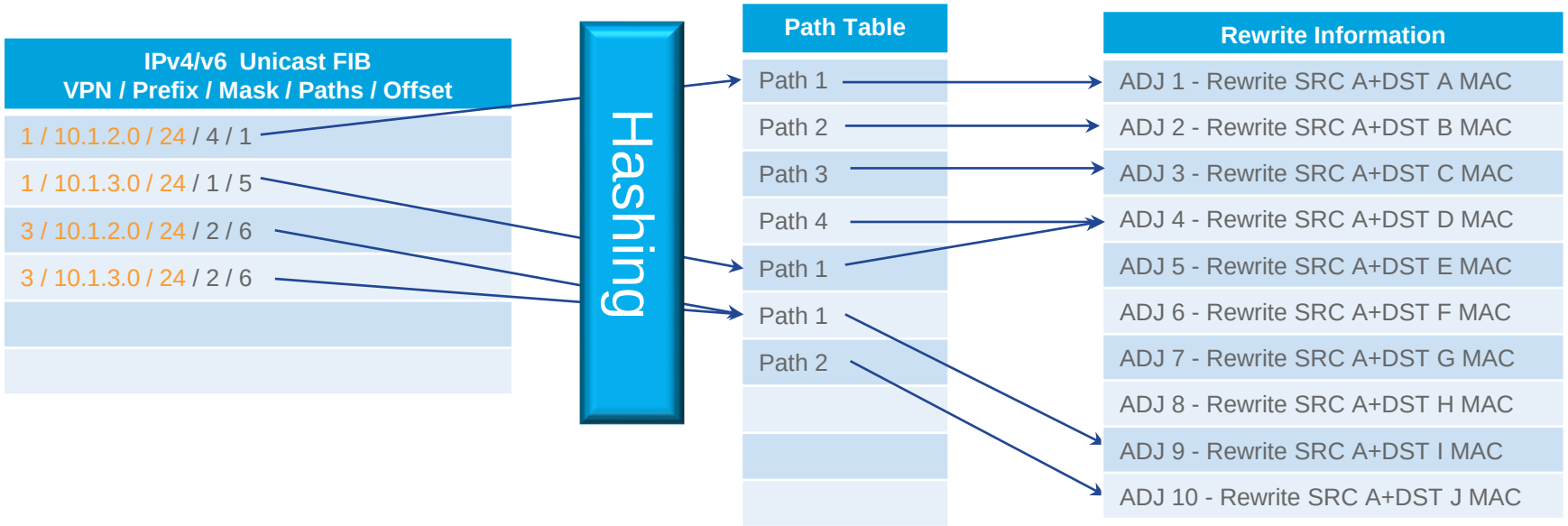
Bare Metal, Hypervisors, Containers & Unikernels Changes in End Point Density



Unikernels, also know as "virtual library operating system"

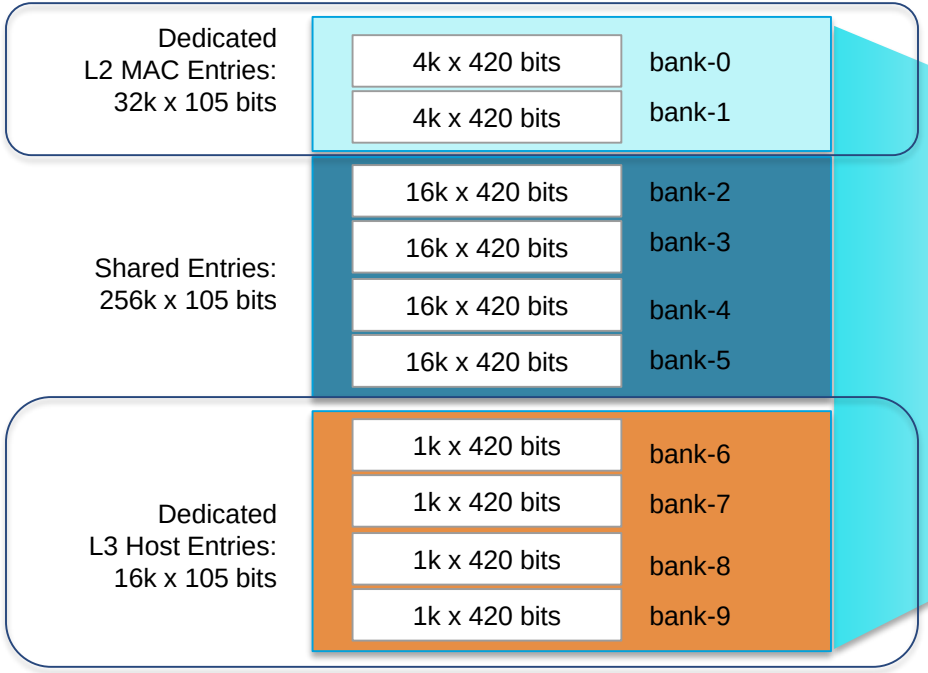
Why does the increase in endpoints matter?

Scale and Flexibility of Forwarding Tables will be stressed



NFE (Trident 2) Unified Forwarding Table

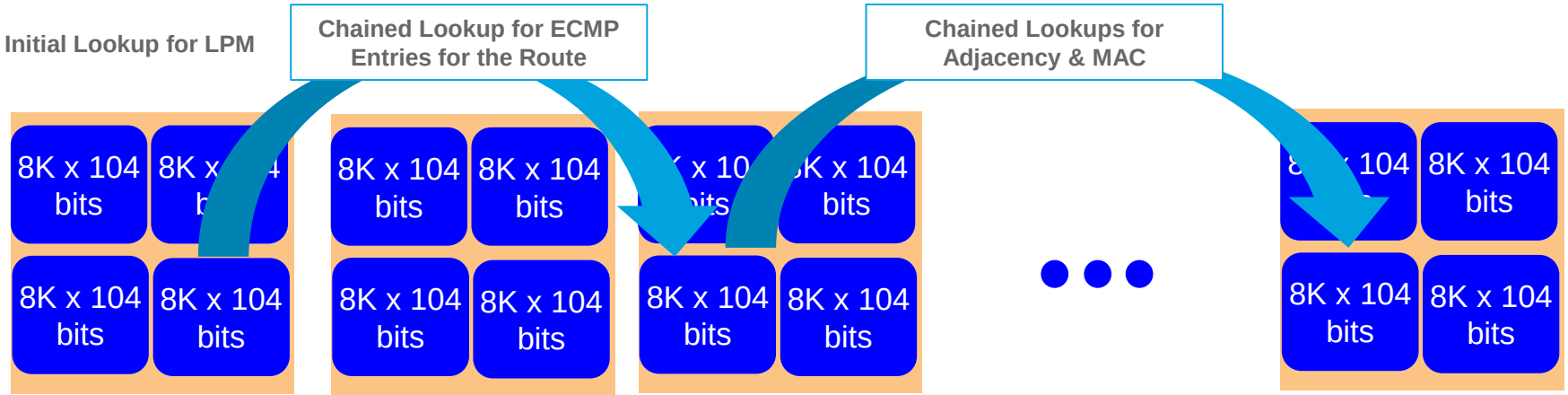
- NFE has a 16K traditional LPM TCAM table.
- Additionally NFE has the following Unified Forwarding Table for ALPM (Algorithm LPM) Mode
- NFE has dedicated adjacency table (48K)



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

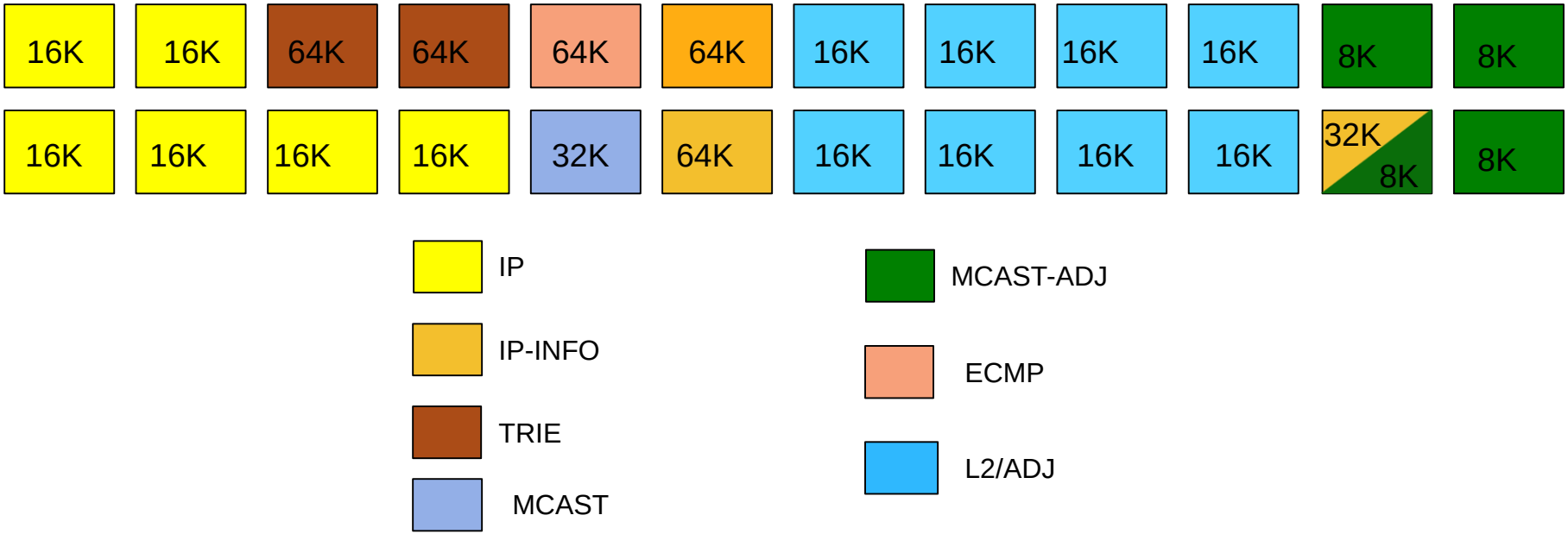
ASE2, ASE3 & LSE
Tile Based Forwarding Tables



- Improve flexibility by breaking the lookup table into small re-usable portions, “tiles”
- Chain lookups through the “tiles” allocated to the specific forwarding entry type
  - IP LPM, IP Host, ECMP, Adjacency, MAC, Multicast, Policy Entry
  - e.g. Network Prefix chained to ECMP lookup chained to Adjacency chained to MAC
- Re-allocation of forwarding table allows maximised utilisation for each node in the network
  - Templates will be supported initially

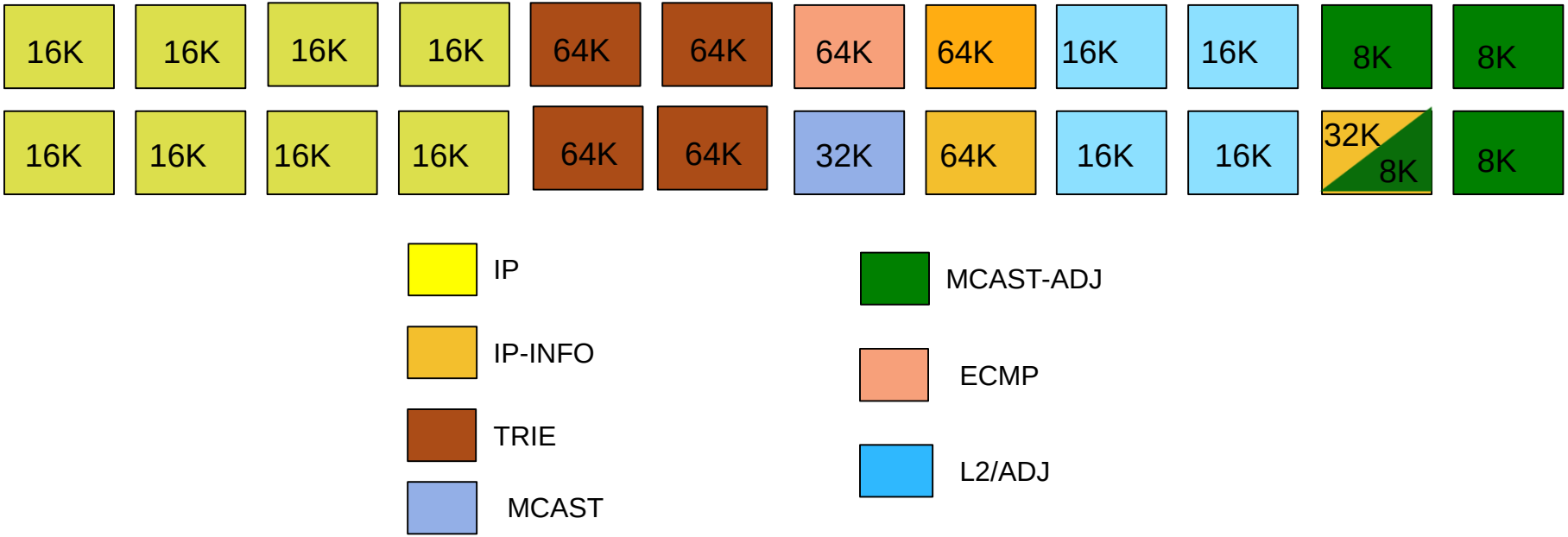
Example Templates – Nexus 9200 (ASE-2)

- Host Heavy Template
  - e.g. Aggregation for smaller network



Example Templates – Nexus 9200 (ASE-2)

- Balanced Host and Routes
  - e.g. Aggregation for Classical L2/L3 design



N9200 (ASE-2) Initial TCAM Templates

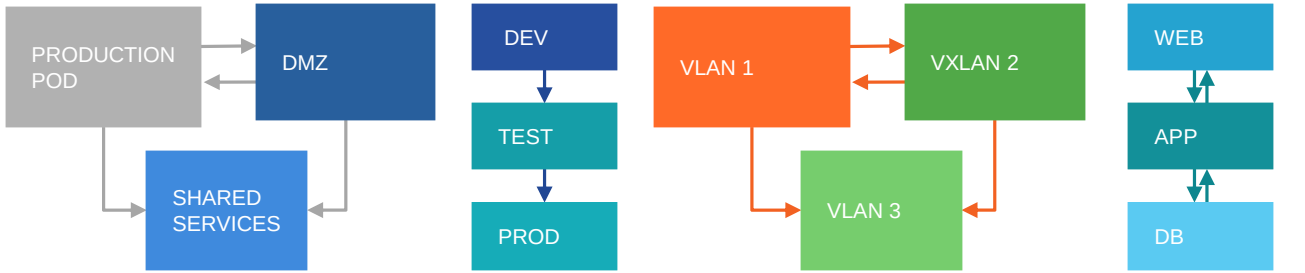
Forwarding Table	FCS Host / MAC balanced	Host Heavy	LPM Heavy	
IPv4 Prefix (LPM)	16K	16K	256K	LPM
IPv6/64 Prefix	16K	16K	256K	
IPv6 /128 Prefix	8K	8K	128K	
IPv4 host routes	112K	256K	32K	Host
IPv6 host routes	48K	192K	16K	
MAC	96K	16K	16K	

ASE2, ASE3 & LSE Optimisation

Different Lookup Table Layouts and Defaults

	ASE2 3.6T N9200	ASE3 1.6T N9200	LSE 1.8T N9300EX/X9700EX	T2 1.28T/ 1 slice N3100	Tomahawk 3.2T / 4 slices N3200	Jericho On Chip	
IPv4 Prefix (LPM)	256K*	256K*	750K*	192K*	128K*	192K	LPM
IPv6/64 Prefix (LPM)	256K*	256K*	750K*	84K*	84K*	64K	
IPv6 Prefix /128 (LPM)	128K*	128K*	384K*	20K*	20K*	64K	
IPv4 host routes	256K*	256K*	750K*	120K*	104K*	750K	Host
IPv6 host routes	128K*	128K*	384K*	20K*	20K*	64K*	
MAC	256K*	256K*	512K*	288K*	136K*	750K	

Hyper-Converged Fabrics Introduces the Same Scaling Problem for Segmentation and Security



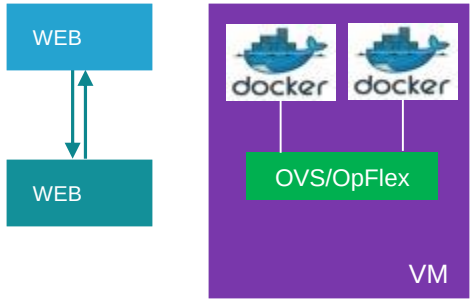
Basic DC Network Segmentation

Segment by Application Lifecycle

Network centric Segmentation

Per Application-tier / Service Level Micro-Segmentation

New



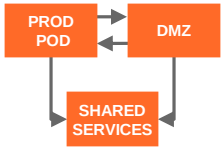
Intra-EPG Micro-Segmentation

Container Security

Level of Segmentation/Isolation/Visibility

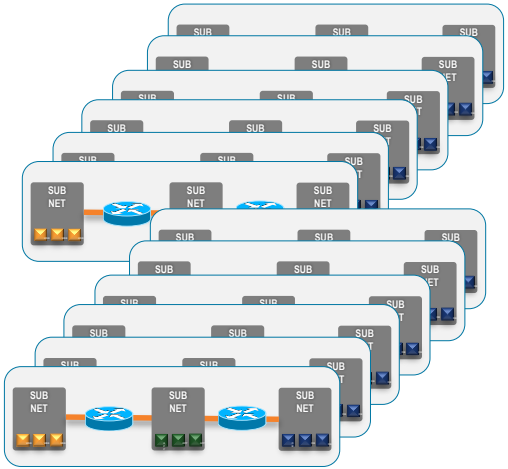
Fabric Wide Segmentation

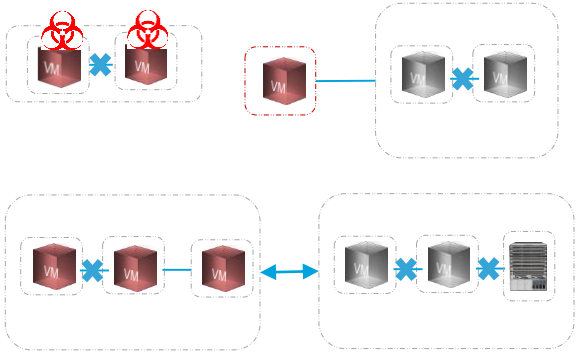
Multi-tenancy at Scale



Basic DC Segmentation

Application Lifecycle Segmentation





Macro Segmentation
2K VRF + **6K** TCAM

Macro Segmentation at Scale
16K VRF per switch

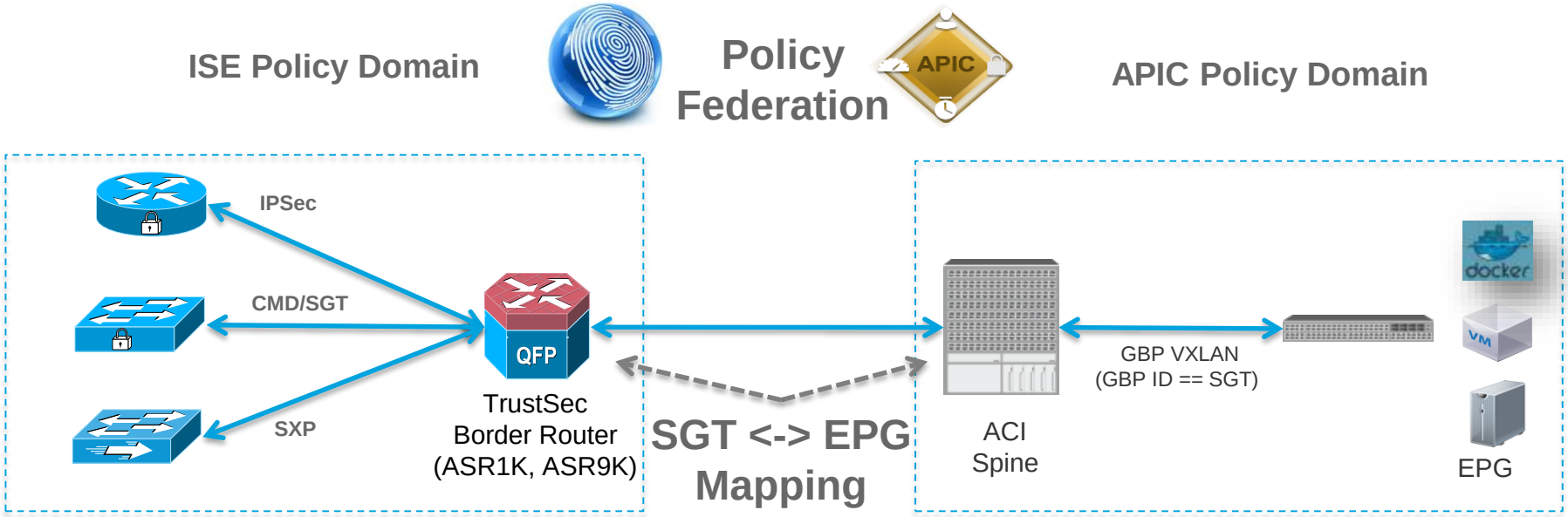
ASE-2, ASE-3, LSE

ACI
Micro-Segmentation at Scale

140K Security Policies per switch

LSE

Consistent Campus Security – DC Policy Enforcement



Real-time Flow Sensors ASE-3 & LSE

Hardware Sensors in ASICs

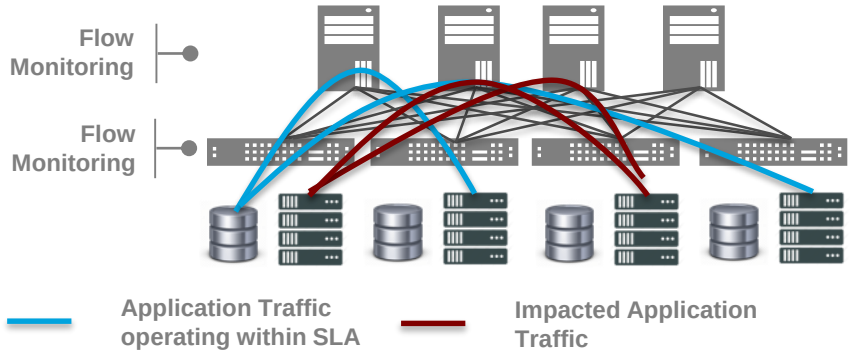
- Hash table to store flows (5-tuple) and the related stats
- Flow table is exported from the ASIC periodically through UDP tunnels
- Capable to capture all flows, selective capture possible
- Stats have two modes
  - Concise (64K flows/ slice): byte and packet count, start and end timestamps
  - Detailed (32K flow/ slice): concise and analytics information

Sensor Data (Examples)

- Capture predefined anomalies
  - TTL changed
  - Anomalous TCP flags seen (xmas flags, syn & rst, syn & fin,…)
  - Header fields inconsistencies
  - Anomalous seq/ack numbers
- Capture standard errors
  - IP length error, tiny fragment etc
- Measure burstiness in the flow
  - Capture the size and time window when the max burst was seen in the flow
  - Correlate burst arrival time across multiple flows in s/w to infer microburst

Fabric Wide Troubleshooting
Real Time Monitoring, Debugging and Analysis

Granular Fabric Wide Flow Monitoring Delivering Diagnostic Correlation



Debug
Understand ‘what’ and ‘where’ for drops and determine application impact

Monitor
Track Latency (avg/min/max), buffer utilisation, network events

Analyse
Specific events and suggest potential solution (e.g. trigger automatic rollback)

Pervasive NetFlow at Scale

‘If you can’t see it, you can’t secure it’

Customer Asks

Top Talker Analysis

Business Critical vs. Best Effort Security Telemetry

Fabric Wide Trouble-shooting

On demand & full history

Capacity planning

Hotspot Detection, Trending

10/25/40/100G Line rate

Cisco Solution

Collect all data everywhere in the network…every packet, every flow, every switch

Protects customers’ NetFlow investment:

solarwinds

FLUKE networks

InfoVista Orchestrating network performance

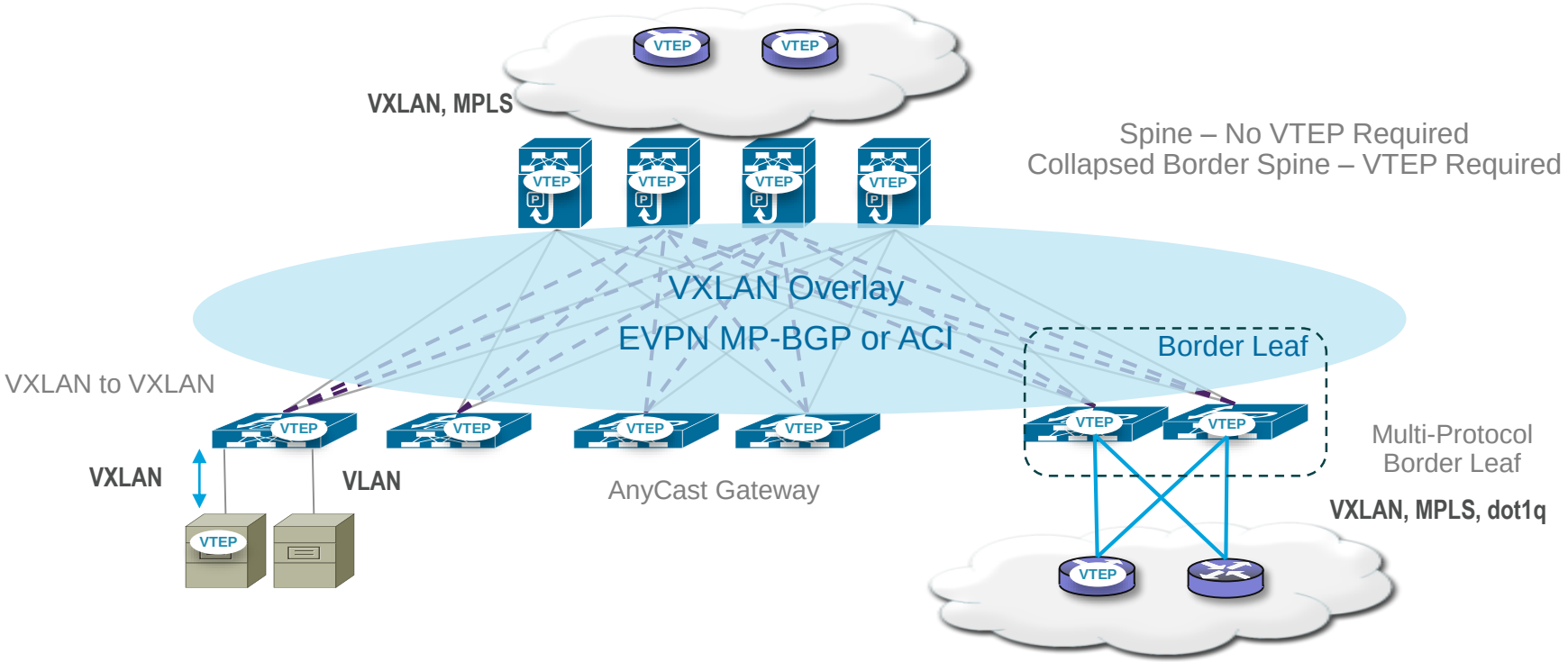
ca NetQoS

CISCO Lancope

CISCO DEVELOPER Registered

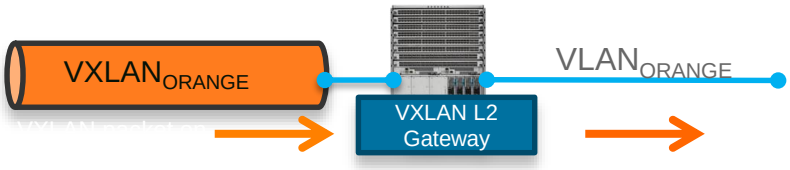
Industry First: Built-In NetFlow Capability across Leaf & Spine

VXLAN & Fabric Design Requirements
Host-based Forwarding

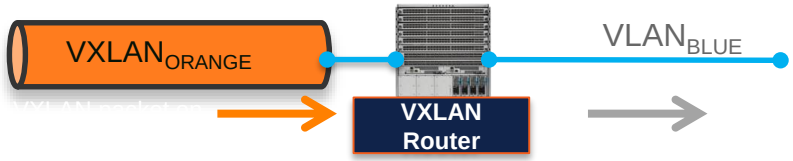


VXLAN Support
Gateway, Bridging, Routing*

VXLAN to VLAN Bridging
(L2 Gateway)



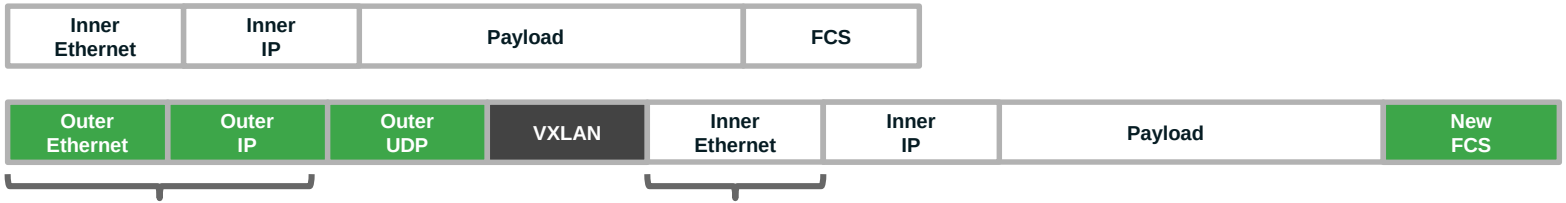
VXLAN to VLAN Routing
(L3 Gateway)



VXLAN to VXLAN Routing
(L3 Gateway)

VxLAN Routing – Trident 2

- VxLAN Routing (forwarding into, out of or between overlays) is not supported in the native pipeline on Trident 2
- During phase 1 of the pipeline lookup the packet the lookup leverages the same station table to identify if the packet is destined to the default GW MAC (switch MAC) or if the packet is an encapsulated packet with the local TEP as the terminating tunnel (either ‘or’ operation)
- If the packet is encapsulated and the tunnel terminates on the switch the phase 2 portion of the lookup the internal packet header can not be resolved via the FIB but only via the L2 station stable (limitation of T2 implementation)
- The internal packet can not be routed after de-encap, similar pipeline limitation prevents a packet that is routed then being encapsulated and have that encapsulated packet forwarded

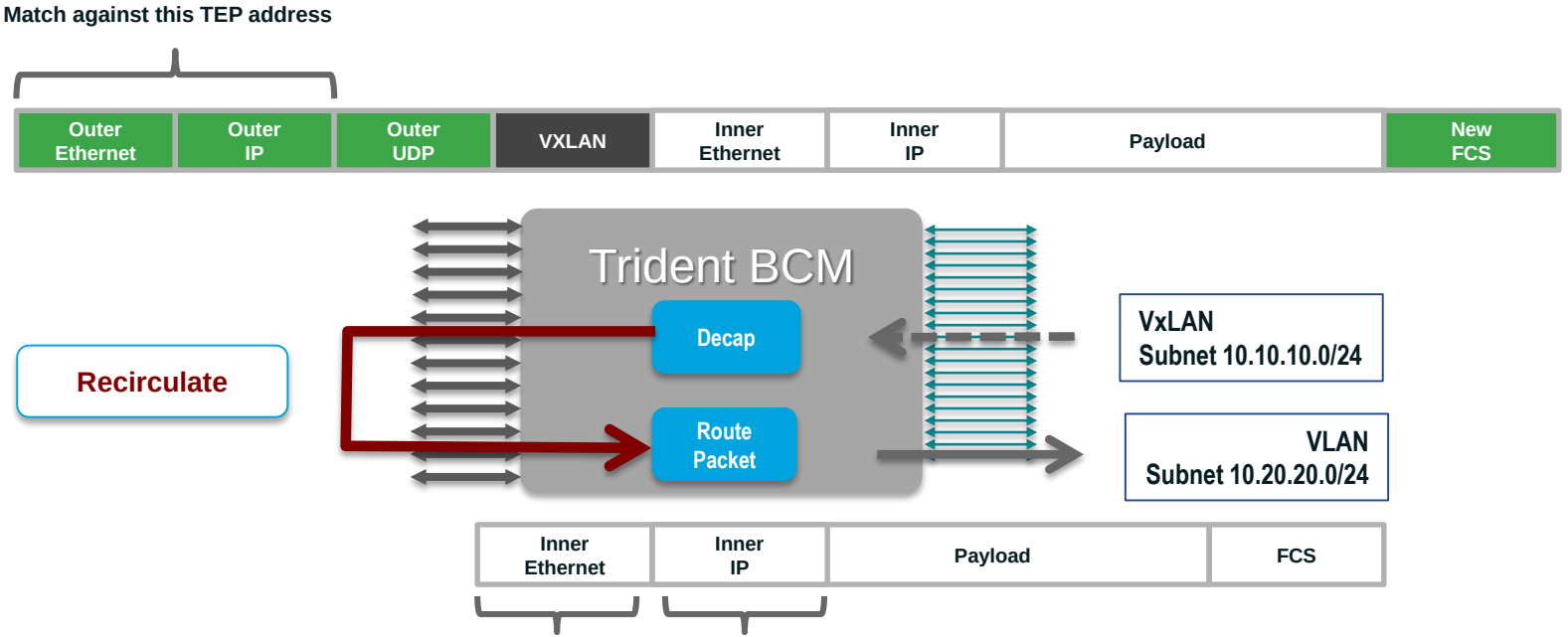


Initial pipeline can only resolve either this packet is destined to default GW MAC or destined to this tunnel endpoint

Second phase lookup can operate only against the L2 station table if tunnel terminates on this switch

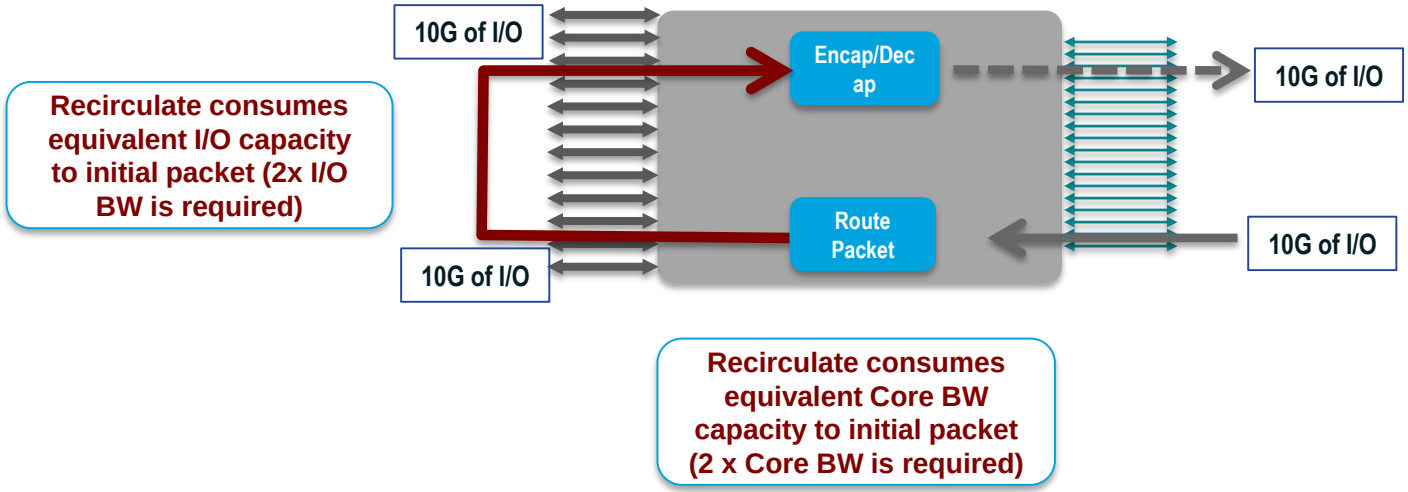
VxLAN to VLAN Routing – Trident 2

VxLAN routed mode via loopback is possible, packet is de-encapsulated, forwarded out through a loopback (either Tx/Rx loopback or via external component), on second pass the match for ‘my router’ MAC results in L3 lookup and subsequent forward via L2 VLAN



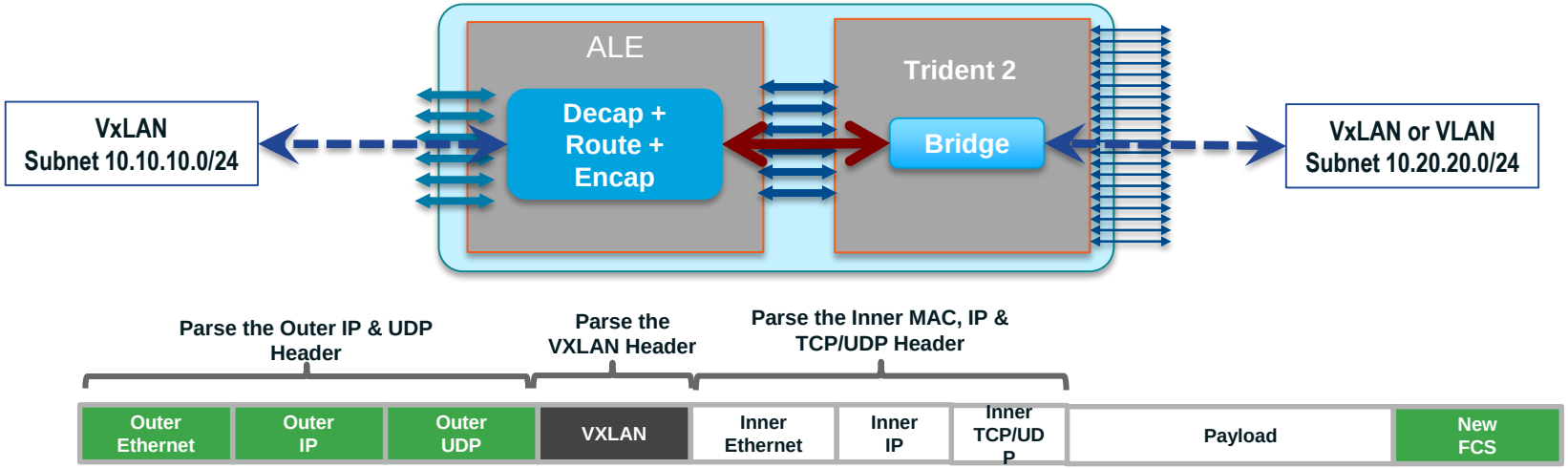
VxLAN Routing – Trident 2 Considerations

- Leveraging loopback through Trident 2 will consume twice the I/O and Core BW as packets are forwarded through the ASIC twice
- VXLAN to VXLAN routing will consume 3x the I/O and Core BW
- Need to understand the ratio of I/O to lookups in cases where recirculation is required



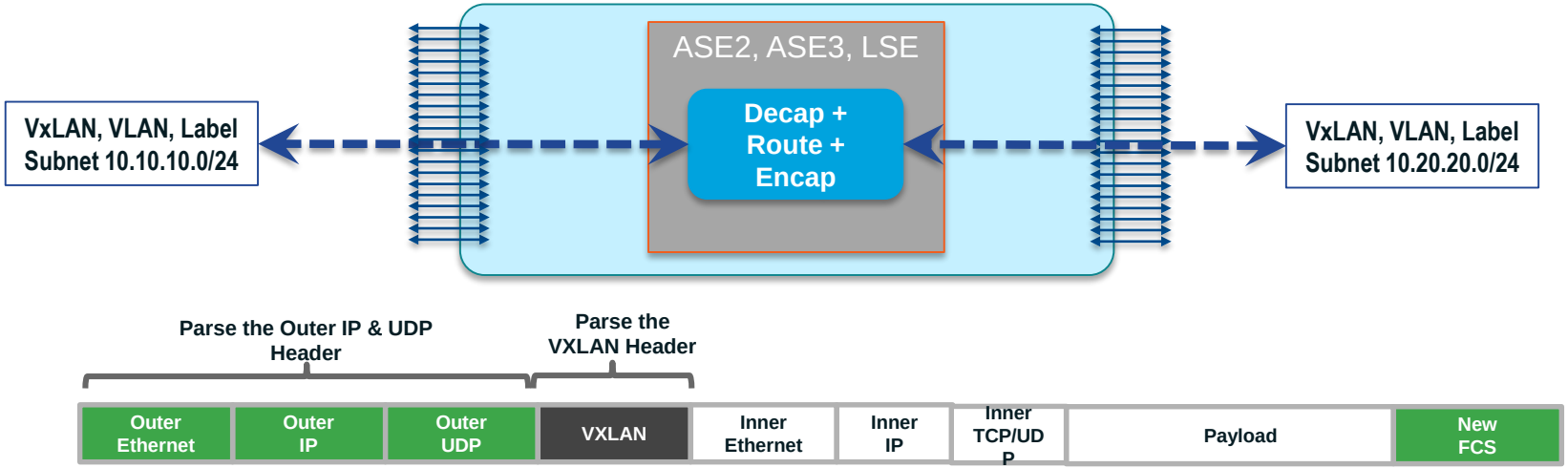
VLAN/VxLAN to VxLAN Routing
Nexus 9300 ACI Mode

- ALE (leaf) and ASE (Spine) ASIC parse the full outer MAC, IP/UDP header, VXLAN and inner MAC, IP & UDP/TCP header in one pipeline pass
- VLAN to VXLAN ‘and’ VXLAN to VXLAN routing is performed in a single pass
- Line rate performance for all encapsulations with all packet sizes



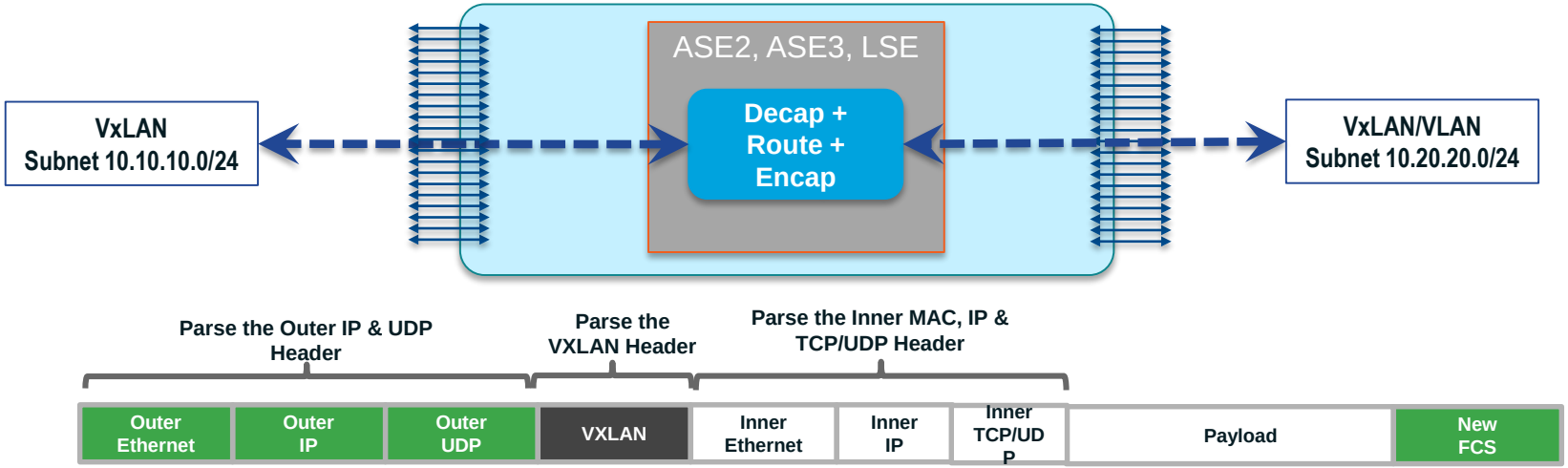
VLAN/VxLAN to VxLAN Routing
Nexus 9300EX, 9200 Standalone Mode

- ASE2, ASE3 & LSE ASIC parse the full outer MAC, IP/UDP header, VXLAN header in one pipeline pass
- VLAN to VXLAN ‘and’ VXLAN to VXLAN routing is performed in a single pass
- Line rate performance for all encapsulations with all packet sizes



VLAN/VxLAN to VxLAN Routing
Nexus 9300EX ACI Mode

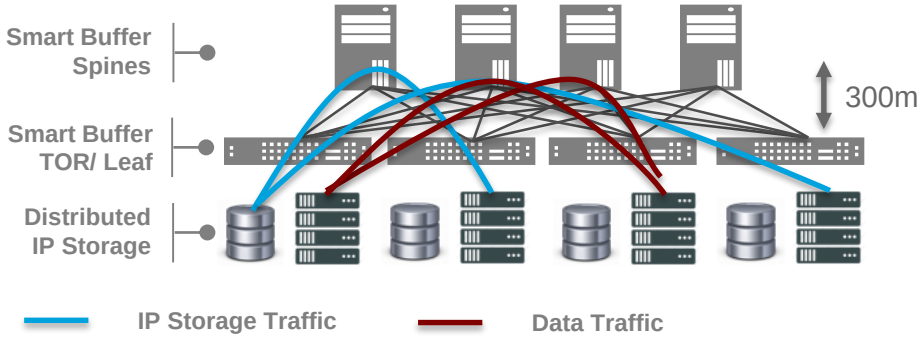
- LSE (Leaf and Spine) ASIC parse the full outer MAC, IP/UDP header, VXLAN and inner MAC, IP & UDP/TCP header in one pipeline pass
- VLAN to VXLAN ‘and’ VXLAN to VXLAN routing is performed in a single pass
- Line rate performance for all encapsulations with all packet sizes



Hyper-Converged Fabric
Distributed IPC & IP Storage

Smart Buffers in Leaf / Spine

Dedicated buffer for guaranteeing lossless traffic



Requirements

Mix of application workloads

Dynamic change of traffic profile
(Distributed IP storage, voice/video, big data, virtual network services, distributed micro-services – IPC)

Maximise application performance

Cisco Solution

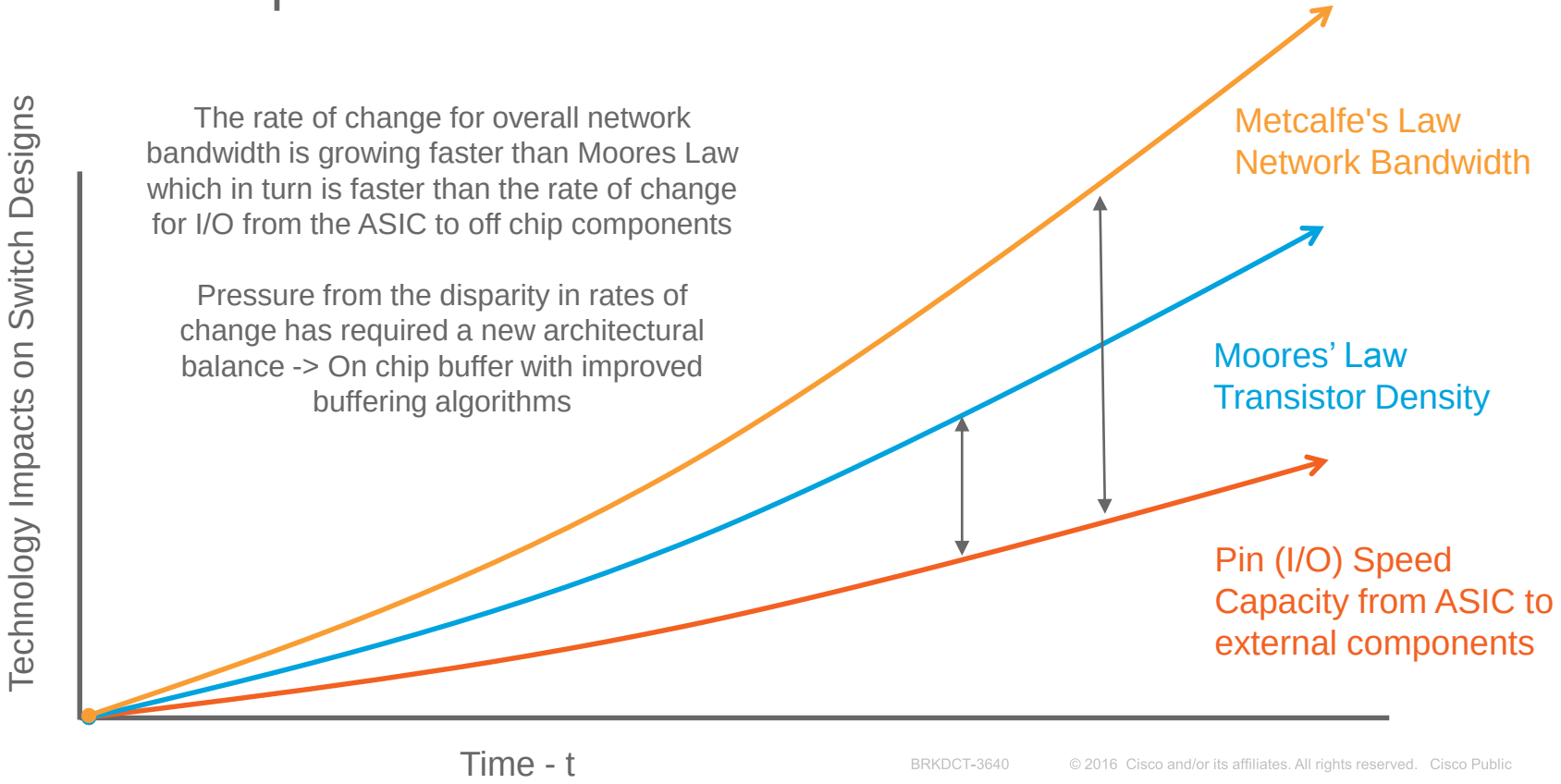
Flow-let switching across spine/ leaf fabric

Dynamic load balancing based on congestion/ latency

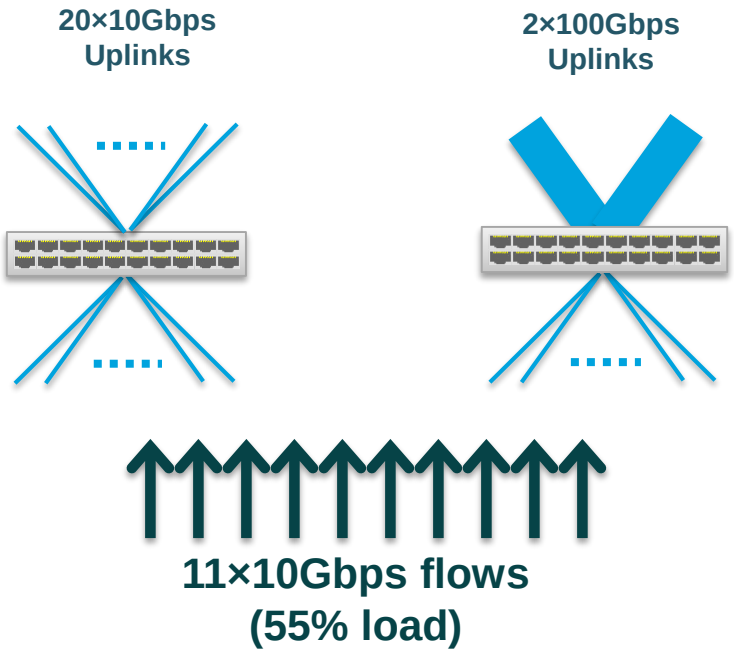
Adaptive scheduling

Improvement of Application Flow Completion Time

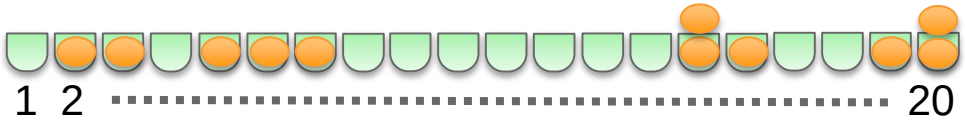
Metcalfe, Moore and ASIC Pin I/O Rates
The Off Chip Bandwidth Problem



Buffering in a Hyper Converged Data Centre
Higher speed links improve ECMP efficiency



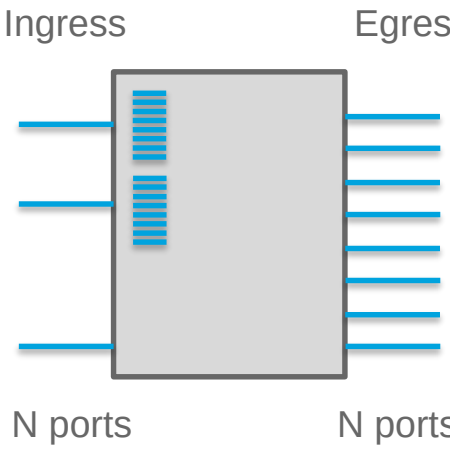
Prob of 100% throughput = 3.27%



Prob of 100% throughput = 99.95%

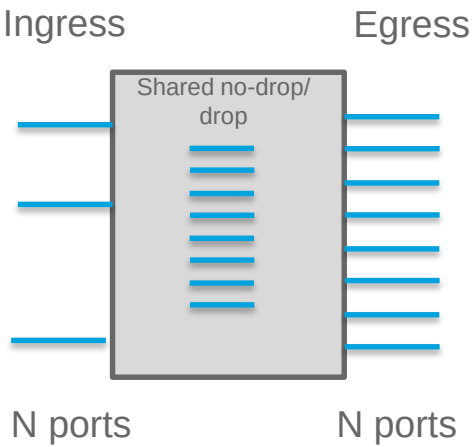
Buffering in a Hyper Converged Data Centre VoQ vs. Output Queue Design

VOQ
Virtual Output Queue



Input buffer for every egress port
NxN buffer size

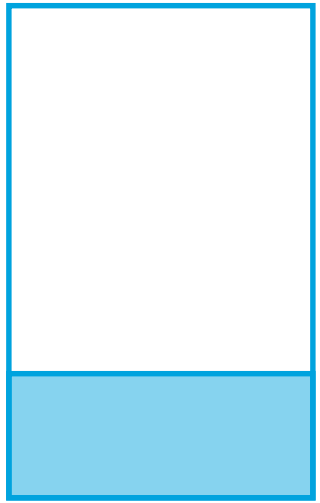
Output Queue Buffer



Shared buffer for N egress ports
N buffer size

Buffering in a Hyper Converged Data Centre
Two Requirements for Buffers

- Long Lived TCP flows
  - Maximise the utilisation of the available network capacity (ensure links are able to run at line rate)
  - Window Size Increases to probe the capacity of the network
  - Delay x Bandwidth Product (C x RTT)
    - e.g if your network had 100 Msec of latency with 10G interface, 125KBytes is required to keep the interface running at maximum capacity (line rate)
- Incast Scenarios
  - Headroom, how much space is available to absorb the burst of traffic (excess beyond the buffer required by long lived TCP flows)



Buffer Available for Burst Absorption

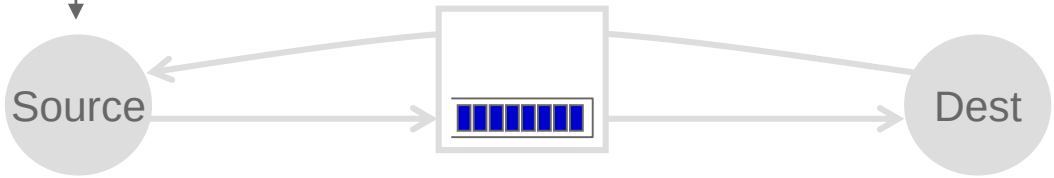
Buffer Required for Maximising Network Utilisation

Long Lived TCP Flows
TCP Congestion Control and Buffer Requirements

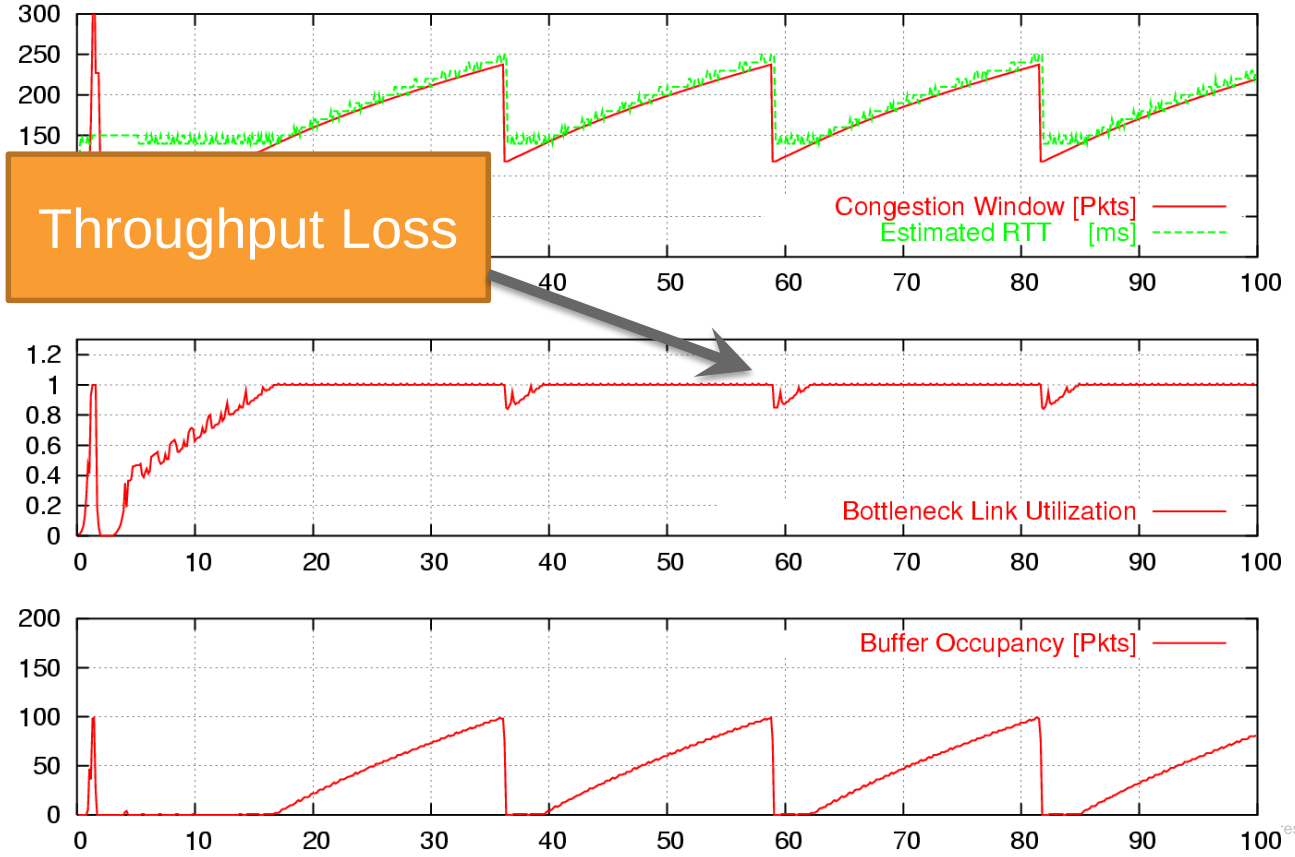
Rule for adjusting W

- If an ACK is received: W ← W+1/W
- If a packet is lost: W ← W/2

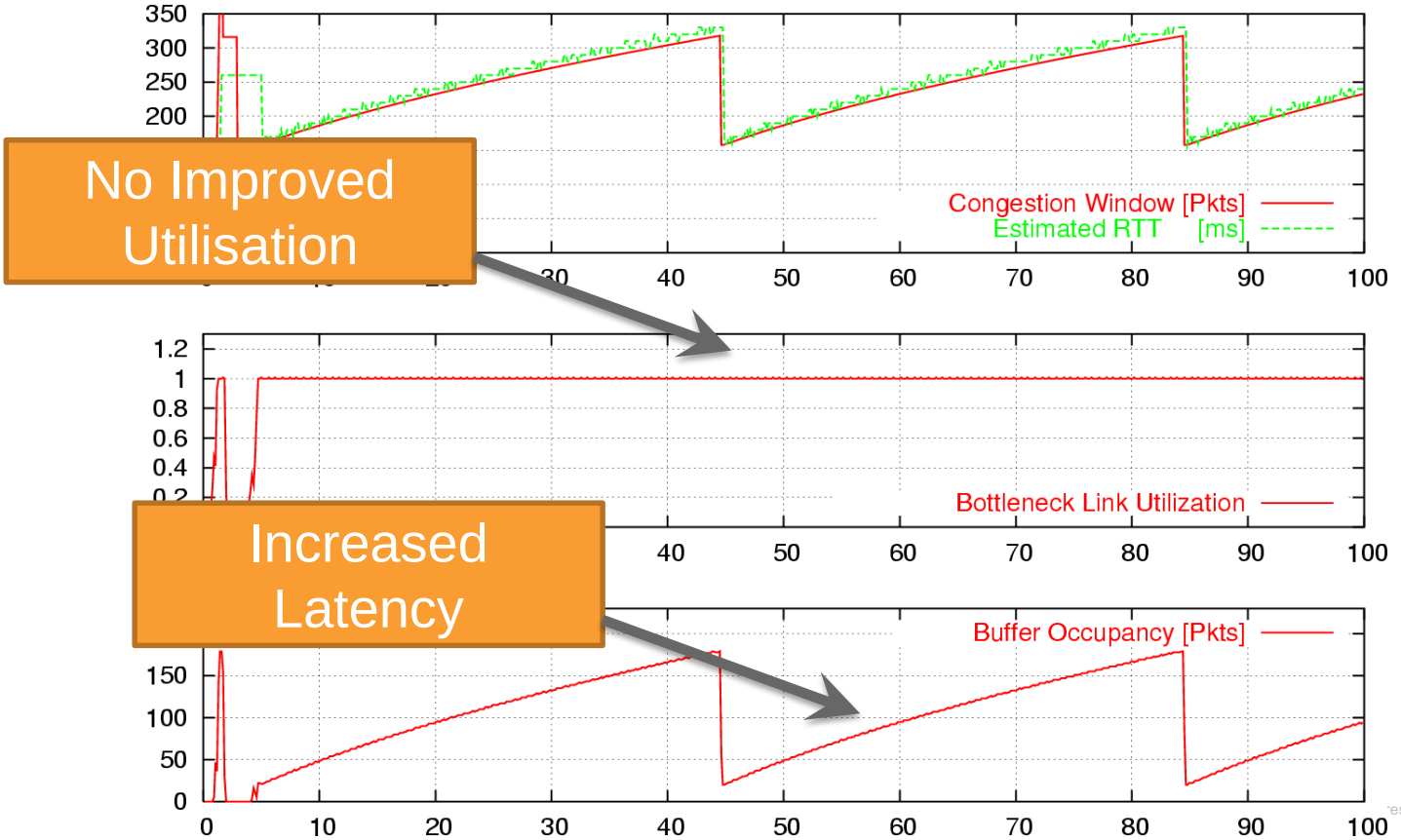
Only W packets may be outstanding



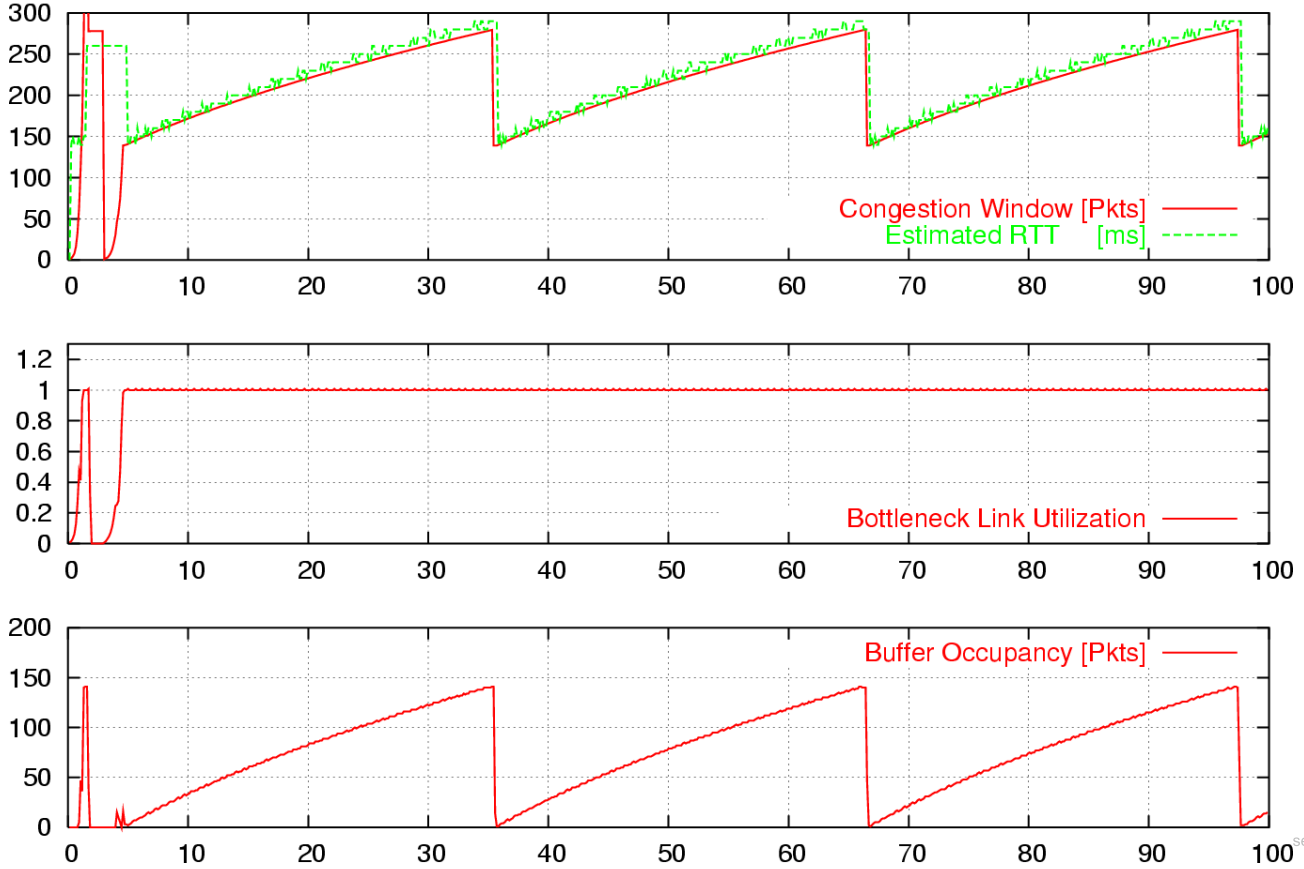
Goldilocks – Too Cold (not enough)
Buffer < C x RTT



Goldilocks – Too Hot (too much)
Buffer > C x RTT



Goldilocks – Just Right
Buffer = C x RTT



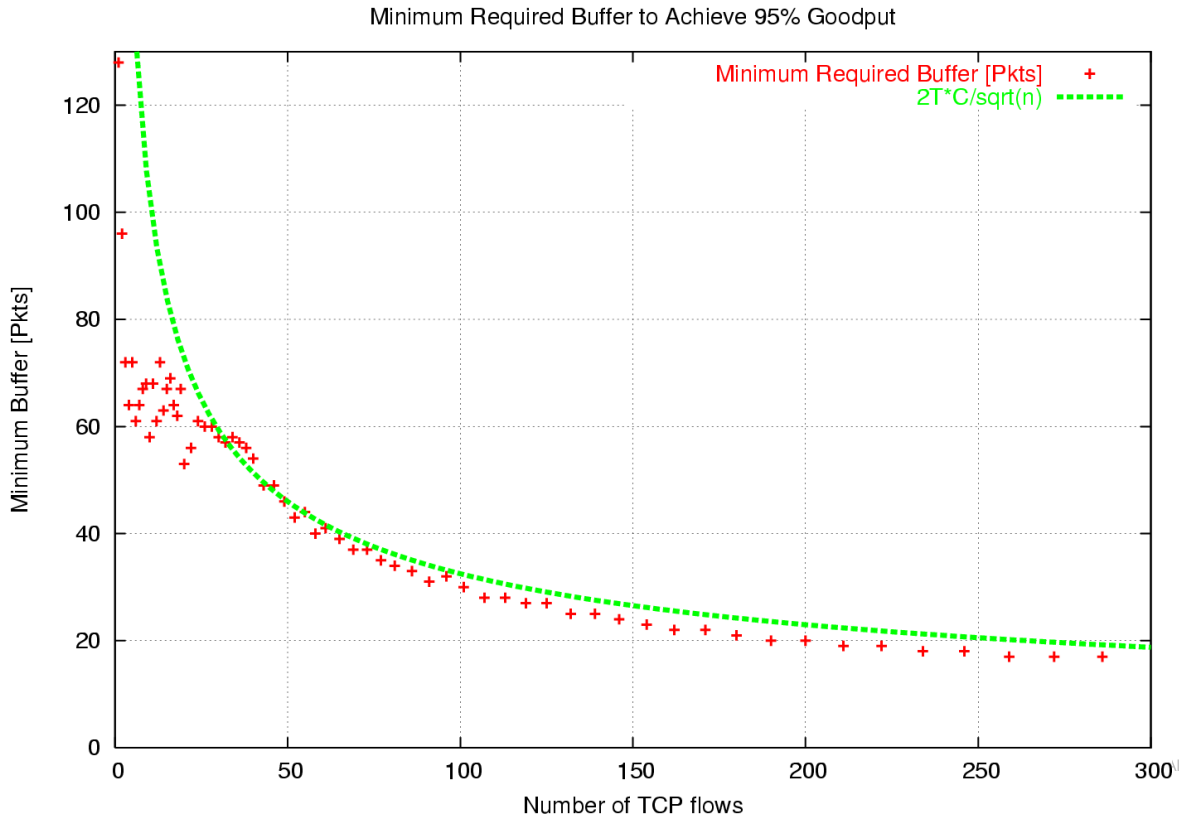
Long Lived TCP Flows
TCP Congestion Control and Buffer Requirements

- Rule of thumb is for one TCP flow, B = $C \times RTT$
- But, typical link carries 10's - 1000s of flows and it turns out that the actual buffer requirement is less than this

Required buffer is $\frac{C \times RTT}{\sqrt{n}}$ instead of $C \times RTT$

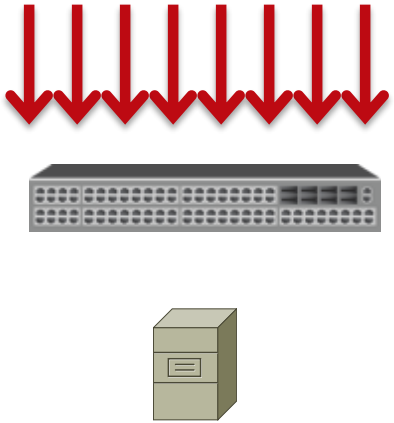
- Proven by theory and experiments in real operational networks
- For example, see Beheshti et al. 2008: "Experimental Study of Router Buffer Sizing"

Long Lived TCP Flows
TCP Congestion Control and Buffer Requirements



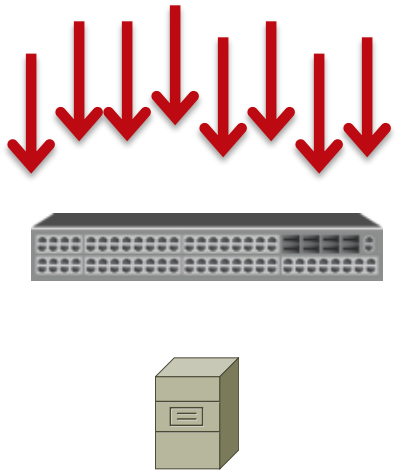
Understanding TCP Incast Collapse

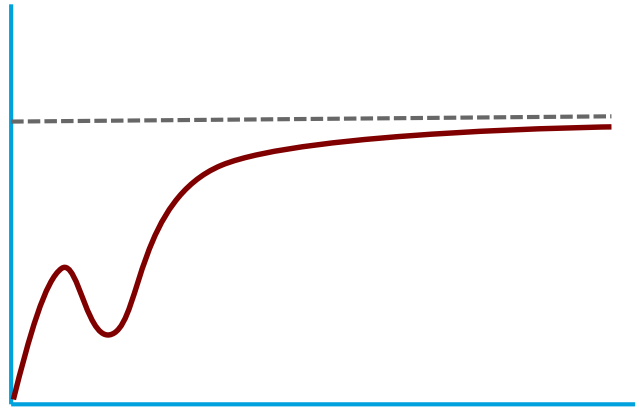
- Synchronised TCP sessions arriving at common congestion point (all sessions starting at the same time)
- Each TCP session will grow window until it detects indication of congestion (packet loss in normal TCP configuration)
- All TCP sessions back off at the same time



Understanding TCP Incast Collapse

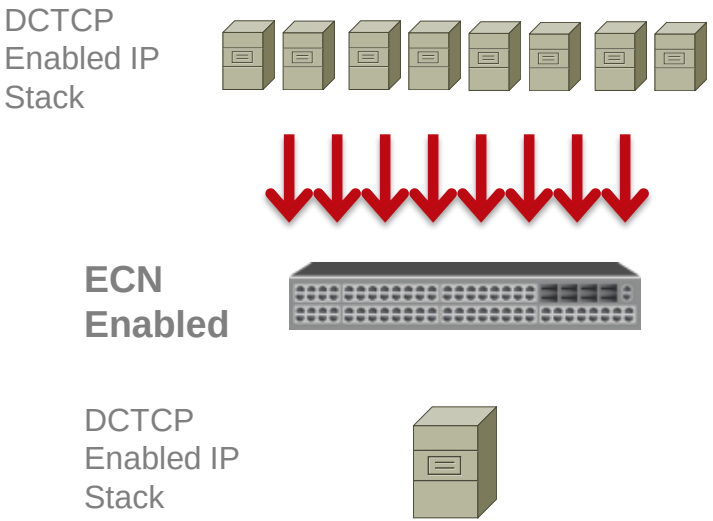
- TCP Incast Collapse requires sources processes to offer connections in a synchronised manner (process dispatch synchronised with no variability in disk seek times)
- TCP Incast Collapse impact is not a permanent condition, TCP will ramp traffic back to maximise link capacity with longer lived sessions

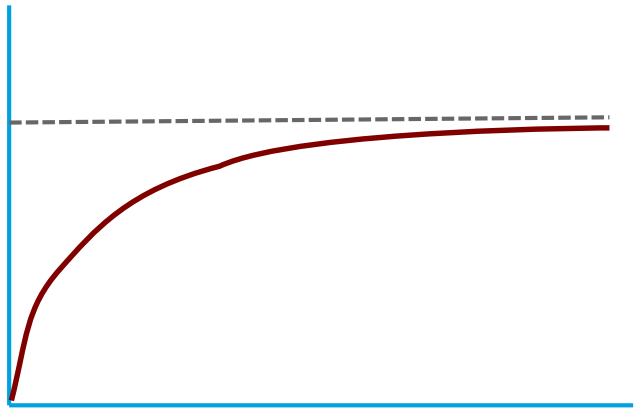




TCP Incast Collapse

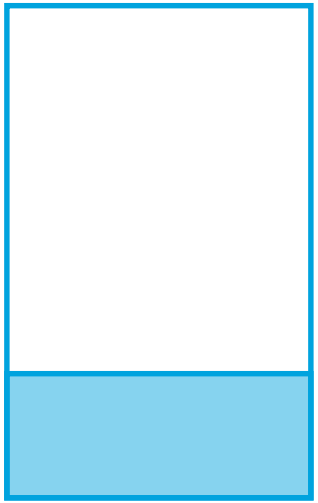
- DCTCP will prevent Incast Collapse for long lived flows
- Notification of congestion prior to packet loss





Buffering in a Hyper Converged Data Centre
Two Requirements for Buffers

- How to minimise the buffer used by long lived flows while ensuring maximal use of network capacity
  - Approximate Fair Drop (AFD) for active queue management
  - Computes a “fair” rate for each flow at the output queue and dropping flows in proportion to the amount they exceed the approximated fair rate
- How to ensure the incast flows are serviced as fast as possible to keep the buffer available
  - Dynamic Packet (Flow) Prioritisation (DPP)

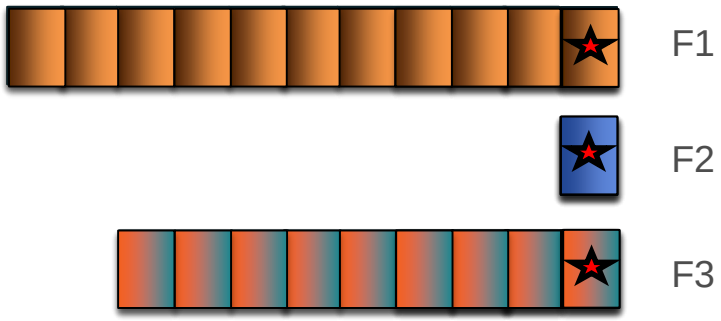


Buffer Available for Burst Absorption

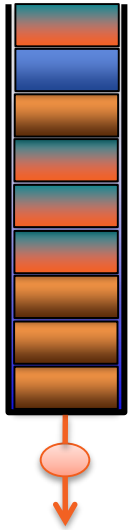
Buffer for minimising long lived TCP flow completion time

Buffering in a Hyper Converged Data Centre
Dynamic Flow Prioritisation

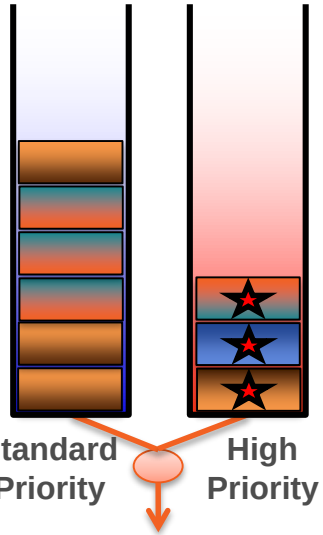
Real traffic is a mix of large (elephant) and small (mice) flows.



Key Idea:
Fabric detects initial few flowlets of each flow and assigns them to a high priority class.



Standard (single priority):
Large flows severely impact performance (latency & loss). for small flows



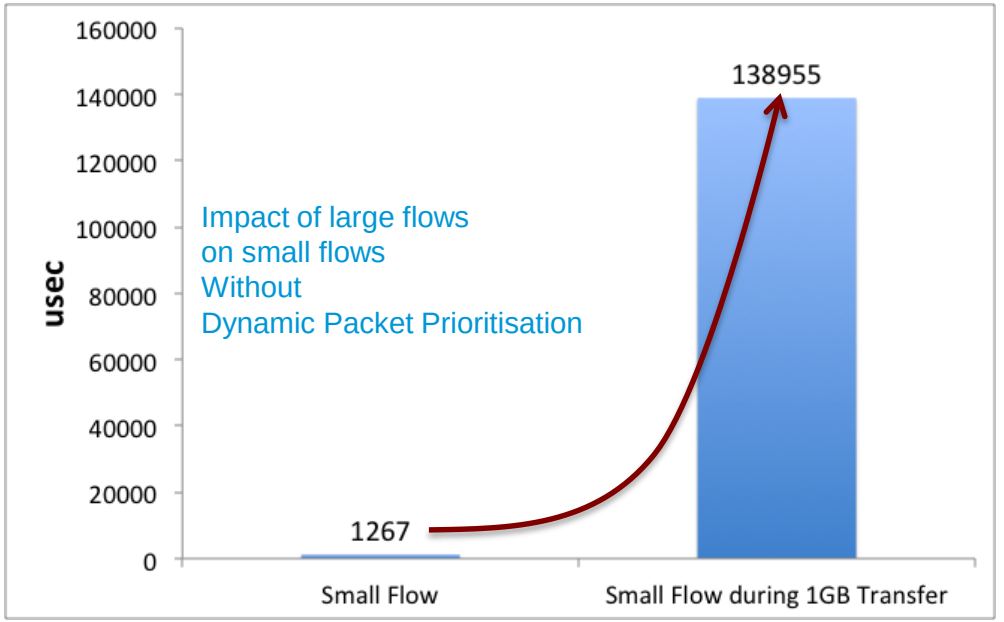
Dynamic Flow Prioritisation:
Fabric automatically gives a higher priority to small flows.

Dynamic Packet Prioritisation

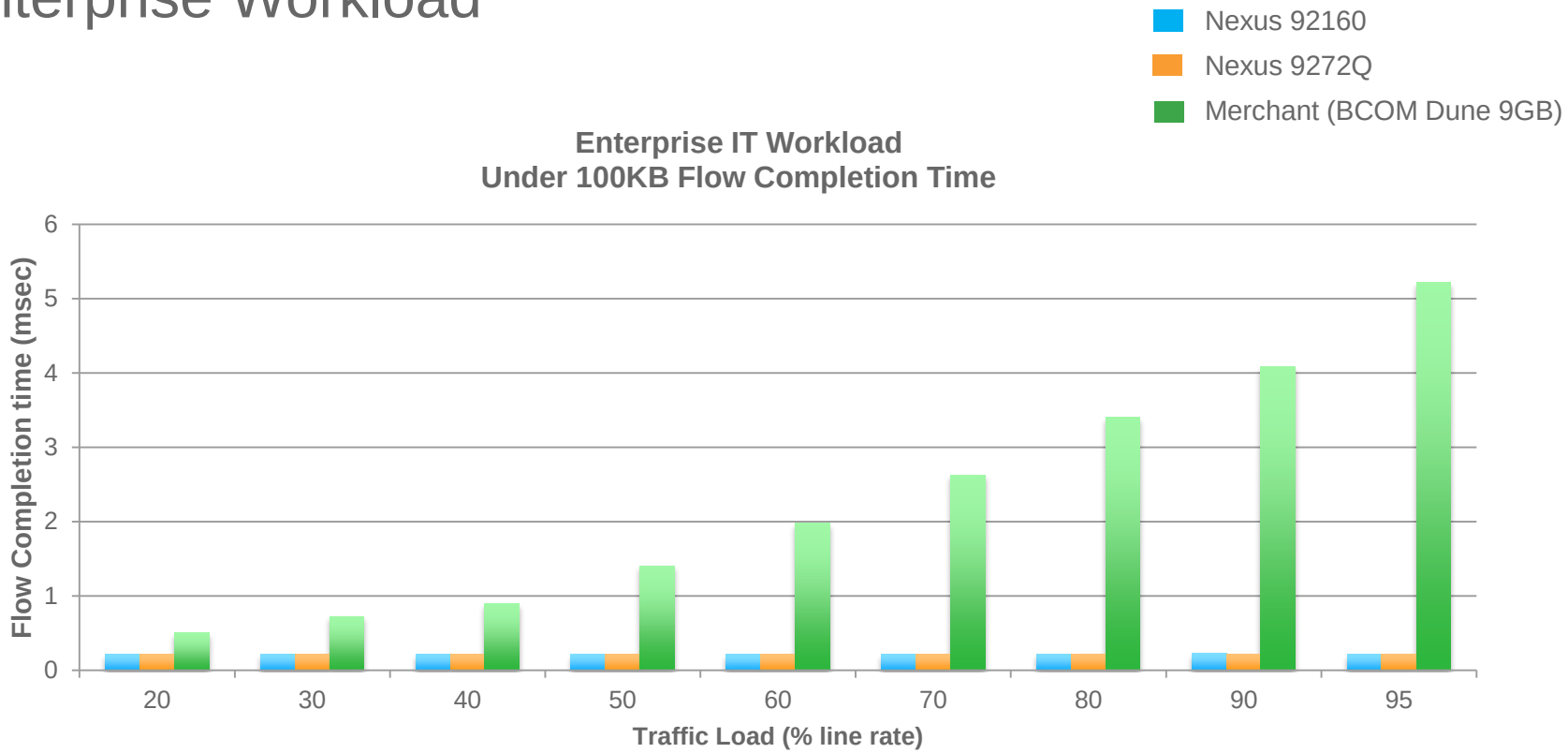
Large Flows tend to use up resources:

- Bandwidth
- Buffer

Unless smaller flows are prioritised – large flows could potentially have an adverse impact on the smaller flows

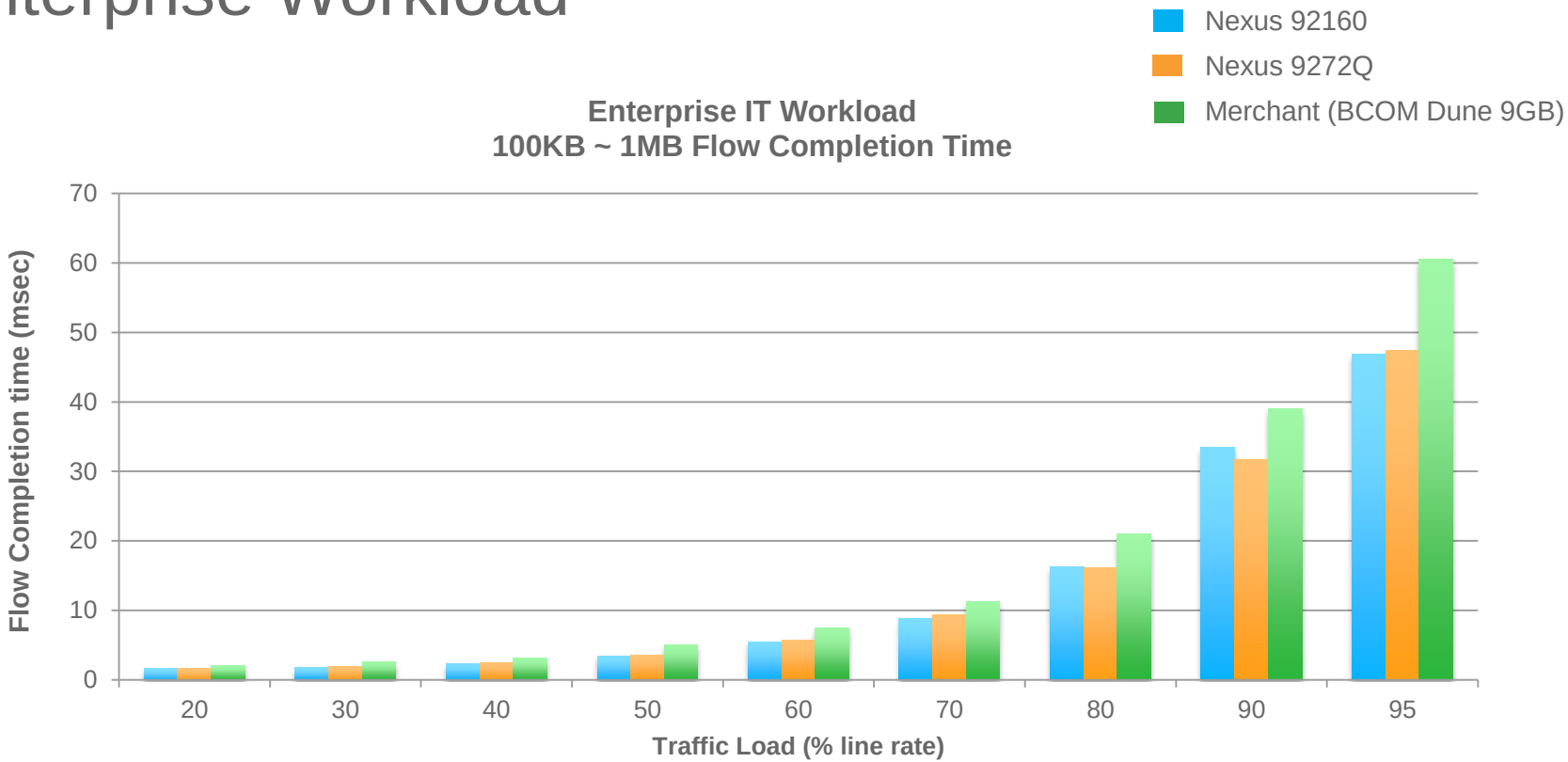


Impact of AFD/DPP
Enterprise Workload



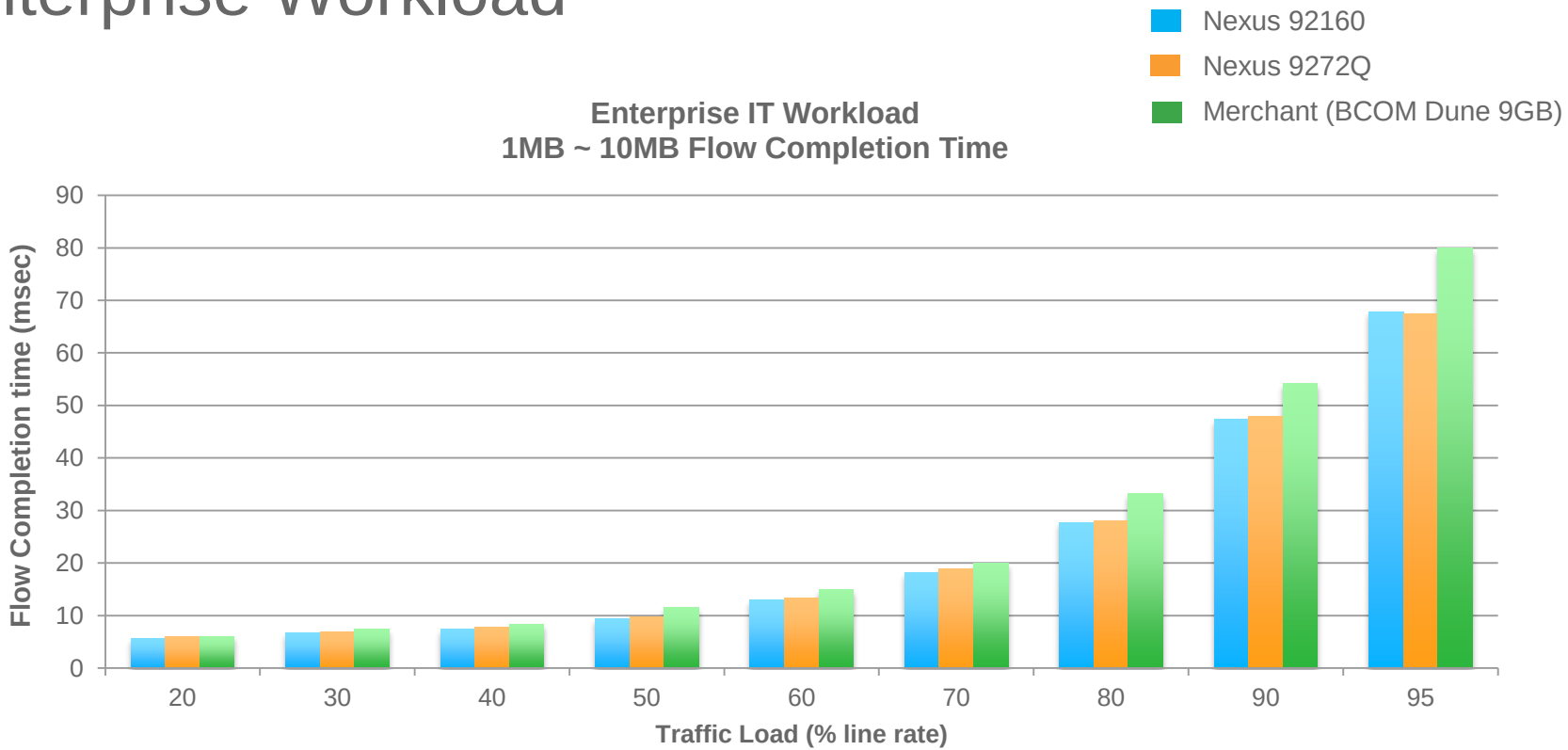
http://miercom.com/cisco-systems-speeding-applications-in-data-Centre-networks/

Impact of AFD/DPP Enterprise Workload



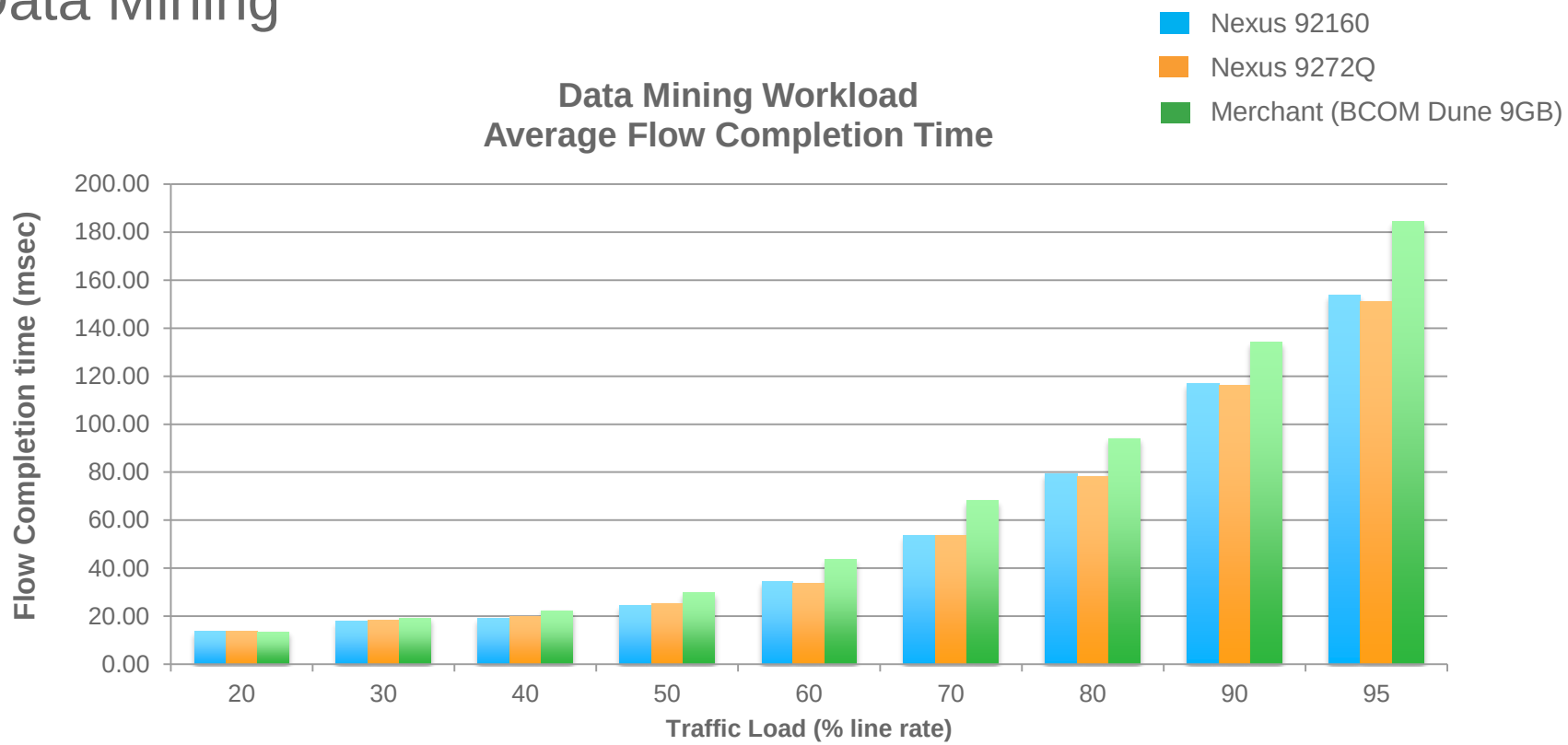
http://miercom.com/cisco-systems-speeding-applications-in-data-Centre-networks/

Impact of AFD/DPP Enterprise Workload



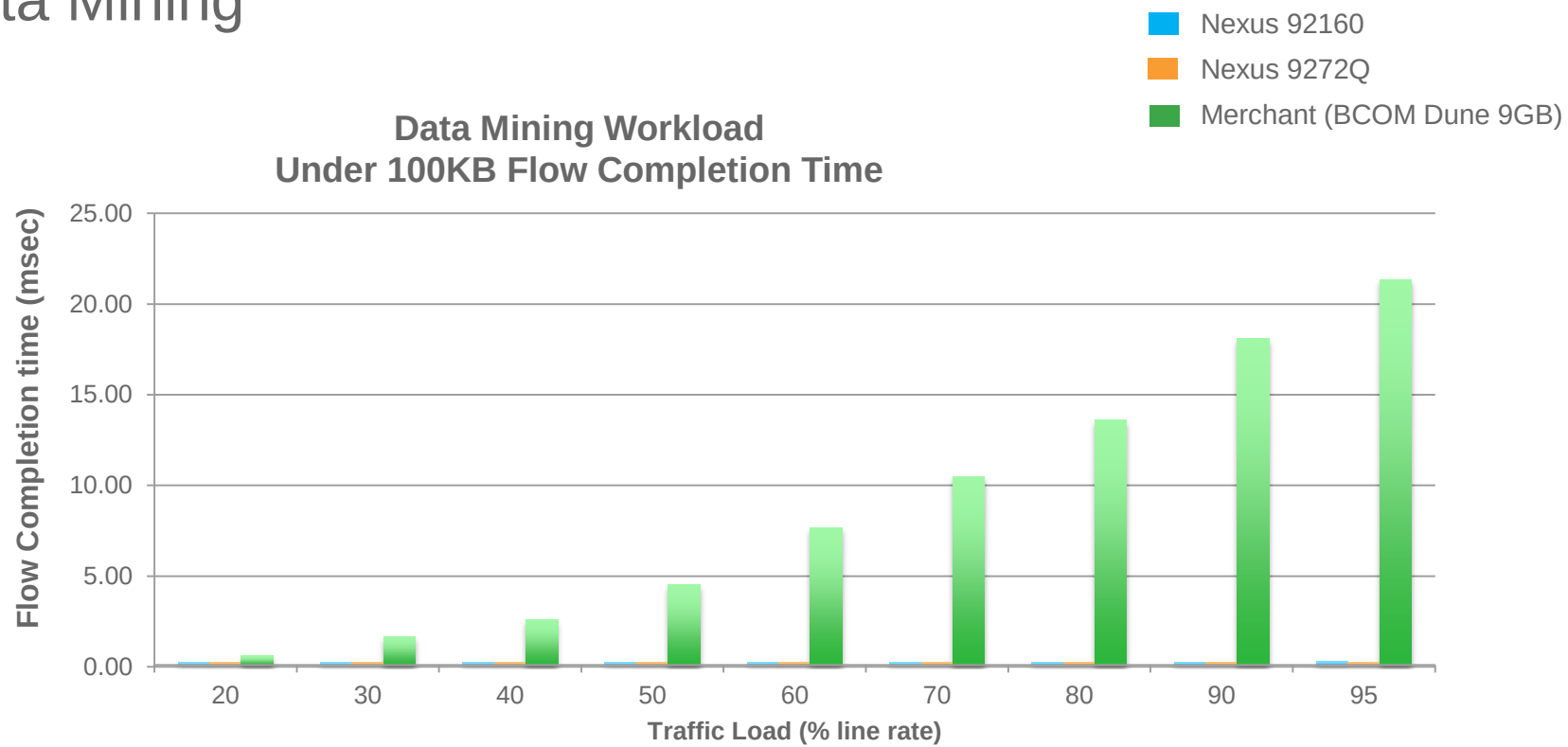
http://miercom.com/cisco-systems-speeding-applications-in-data-Centre-networks/

Impact of AFD/DPP on Incast Traffic
Data Mining



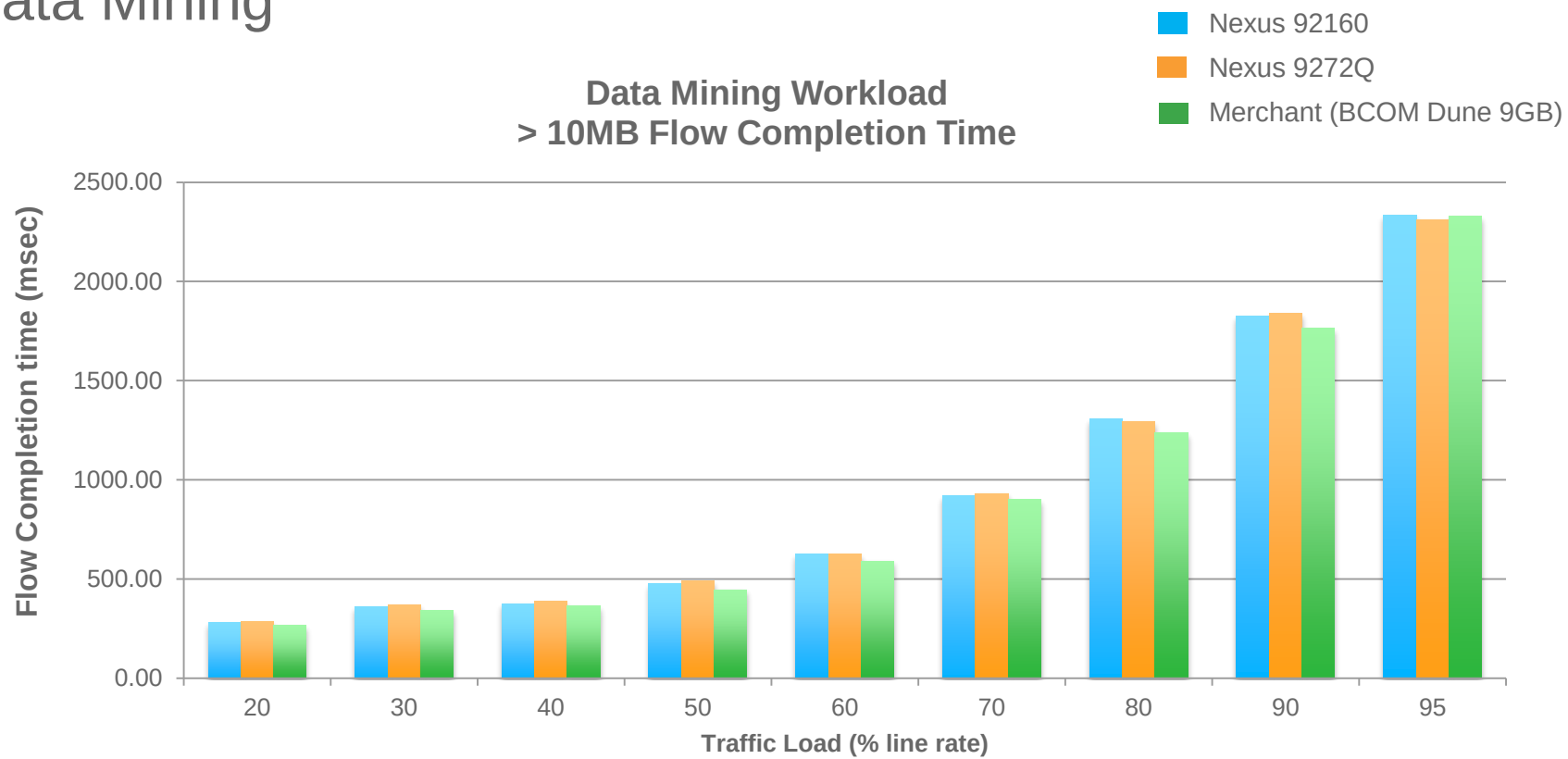
http://miercom.com/cisco-systems-speeding-applications-in-data-Centre-networks/

Impact of AFD/DPP on Incast Traffic
Data Mining



http://miercom.com/cisco-systems-speeding-applications-in-data-Centre-networks/

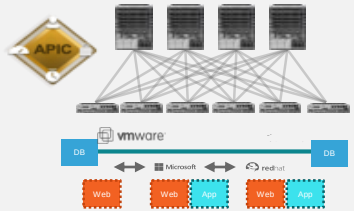
Impact of AFD/DPP on Incast Traffic
Data Mining



http://miercom.com/cisco-systems-speeding-applications-in-data-Centre-networks/

Why do we discuss automation so much?

Application Centric Infrastructure

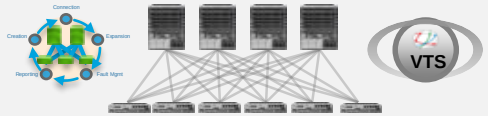


Turnkey integrated solution with security, centralised management, compliance and scale

Automated application centric-policy model with embedded security

Broad and deep ecosystem

Programmable Fabric

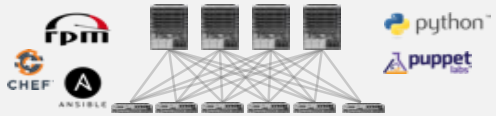


VxLAN-BGP EVPN standard-based

3rd party controller support

Cisco Controller for software overlay provisioning and management across N2K-N9K

Programmable Network



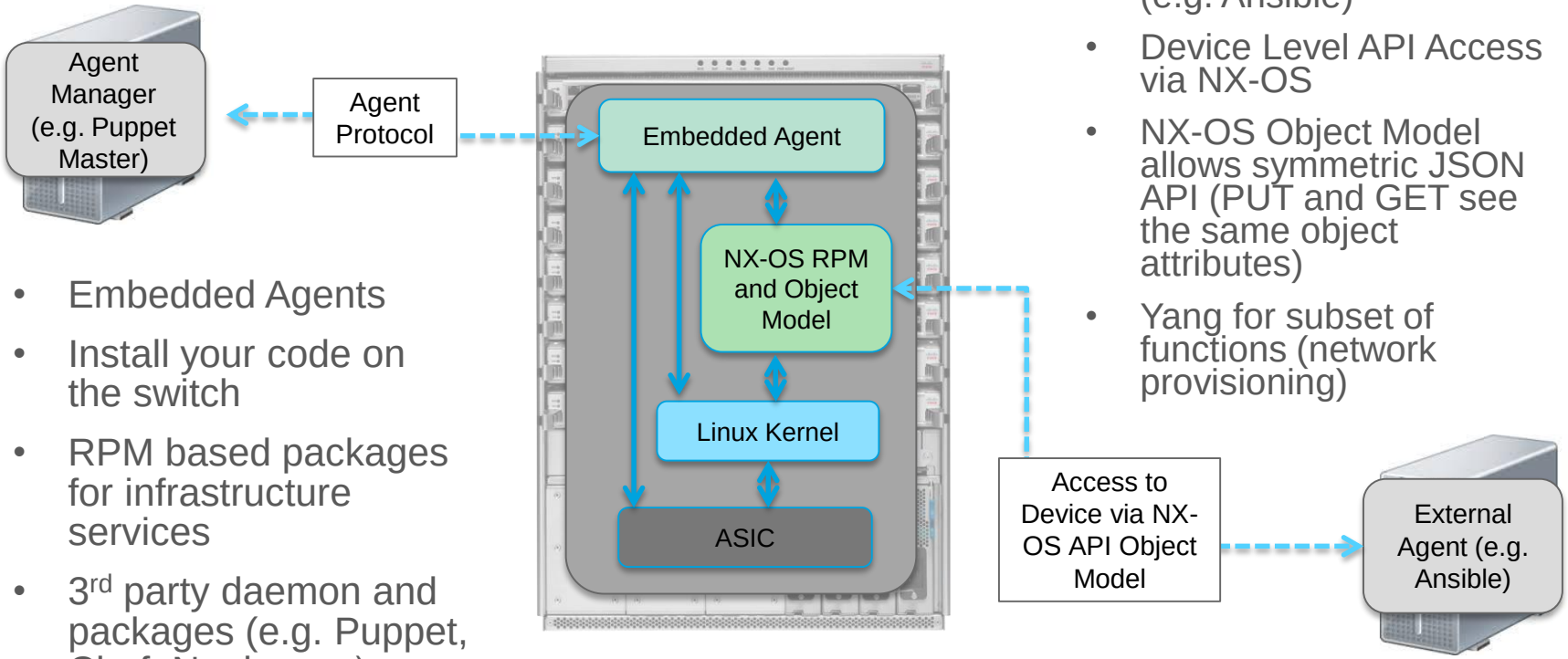
Modern NX-OS with enhanced NX-APIs

DevOps toolset used for Network Management
(Puppet, Chef, Ansible etc.)

Automation, API's, Controllers and Tool-chain's

When you take advantage of Moore's Law you need to shift to a server like operational models

Open NX-OS
Server Like Management Approach

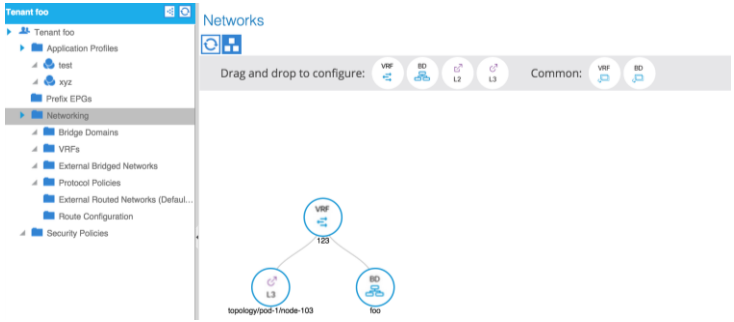


- Embedded Agents
- Install your code on the switch
- RPM based packages for infrastructure services
- 3rd party daemon and packages (e.g. Puppet, Chef, Nagios, …)

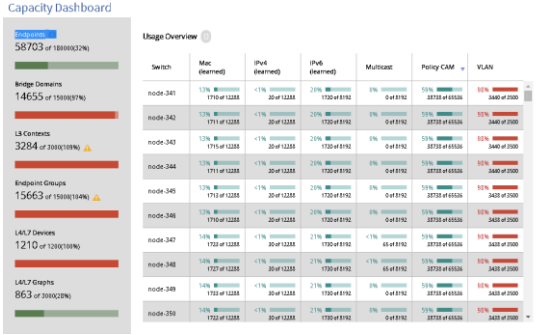
- External Agents and Tools (e.g. Ansible)
- Device Level API Access via NX-OS
- NX-OS Object Model allows symmetric JSON API (PUT and GET see the same object attributes)
- Yang for subset of functions (network provisioning)

APIC
Cloud Like Automation and Operations

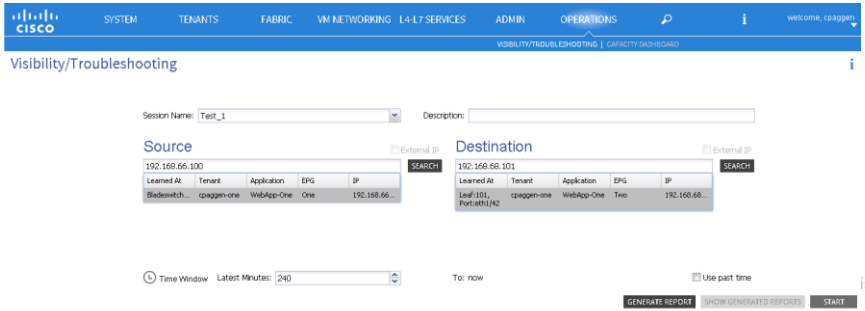
Drag and Drop Configuration



Capacity Dashboard



Troubleshooting Wizards



Agenda

- Existing and New Nexus 9000 & 3000
- What's New
  - Moore's Law and 25G SerDes
  - The new building blocks (ASE-2, ASE-3, LSE)
  - Examples of the Next Gen Capabilities
- Nexus 9000 Switch Architecture
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)
- 100G Optics

Nexus 9300 Series Switches Portfolio
First Generation

N9K-C93120TX

N9K-C9332PQ 

N9K-C9372PX 

N9K-C9372TX 

N9K-C9396PX 

N9K-C9396TX 

N9K-C93128TX 

Nexus® 9372PX/ 9372TX
- 1 RU w/n GEM module slot
- 720Gbps
- 6-port 40 Gb QSFP+
- 48-port 1/10 Gb SFP+ on Nexus 9372PX
- 48-port 1/10 G-T on Nexus 9372TX

Nexus 9332PQ
- 1 RU w/n GEM module slot
- 1,280Gbps
- 32-port 40 Gb QSFP+

Nexus 93120TX
- 2 RU w/n GEM module slot
- 1200Gbps
- 6-port 40 Gb QSFP+
- 96-port 1/10 G-T

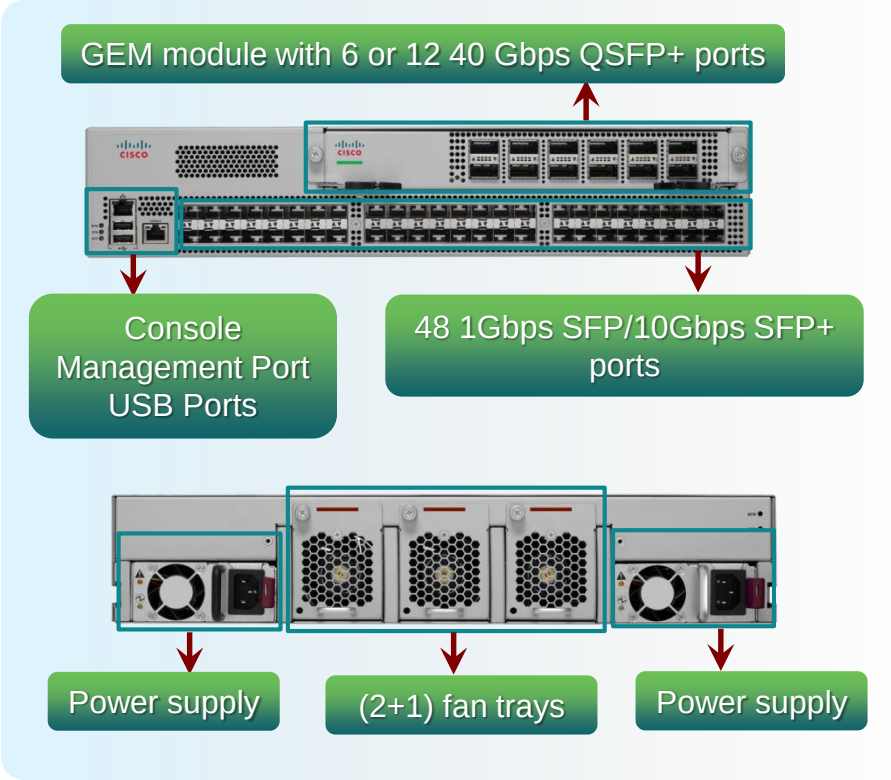
Nexus® 9396PX/ 9396TX
- 2 RU with 1 GEM module slot
- 960Gbps
- 48-port 1/10 Gb SFP+ on Nexus 9396PX
- 48-port 1/10 G-T on Nexus 9396TX
- 6 ports 40 Gb QSFP+ on N9K-M6PQ GEM module
- 12 ports 40 Gb QSFP+ on N9K-M12PQ GEM module
- 4 ports 100 Gb CFP2 on N9K-M4PC-CFP2 GEM module

Nexus 93128TX/ 93128PX
- 3 RU with 1 GEM module slot
- 1,280Gbps
- 96-port 1/10 G-T on Nexus 93128TX
- 96-port 1/10 SFP+ on Nexus 93128P
- 6 ports 40 Gb QSFP+ on N9K-M6PQ GEM module
- 8 ports 40 Gb QSFP+ on N9K-M12PQ GEM module
- 2 ports 100 Gb CFP2 on N9K-M4PC-CFP2 GEM module

Nexus 9300 Platform Architecture First Generation

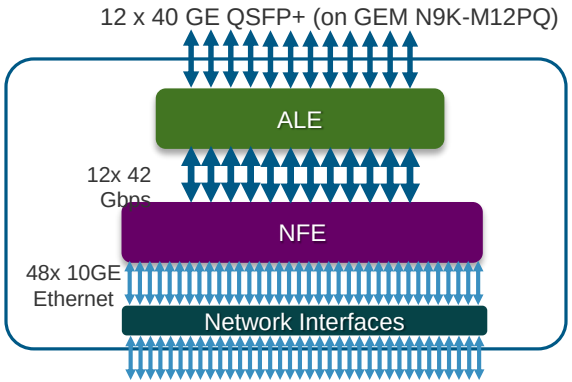
Cisco Nexus® 9396PX / 9396TX

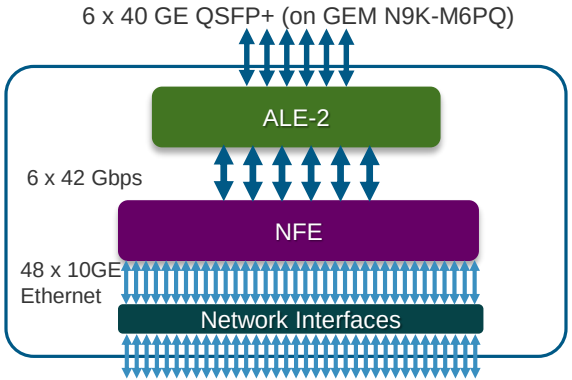
- 2 RU height
- 48x 1Gb SFP / 10 Gb SFP+ ports on Nexus 9396PX
- 48x 1/10 Gb Base-T ports on Nexus 9396TX
- 12x 40 Gb QSFP+ ports on N9K-M12PQ GEM module
- 6x 40 Gb QSFP+ ports on N9K-M6PQ GEM module
- 4x 100 Gb CFP2 ports on N9K-M4PC-CFP2 GEM module
- 1 100/1000baseT management port
- 1 RS232 console 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 full line-rate performance on all ports for all packet sizes
- VXLAN bridging & routing with N9K-M12PQ or N9K-M6PQ GEM module
- VXLAN bridging only with N9K-M4PC-CFP2 GEM module



First Gen Nexus 9300 Series Switch Architecture

Nexus 9396PX/TX Block Diagram with N9K-M12PQ or N9K-M6PQ GEM Module





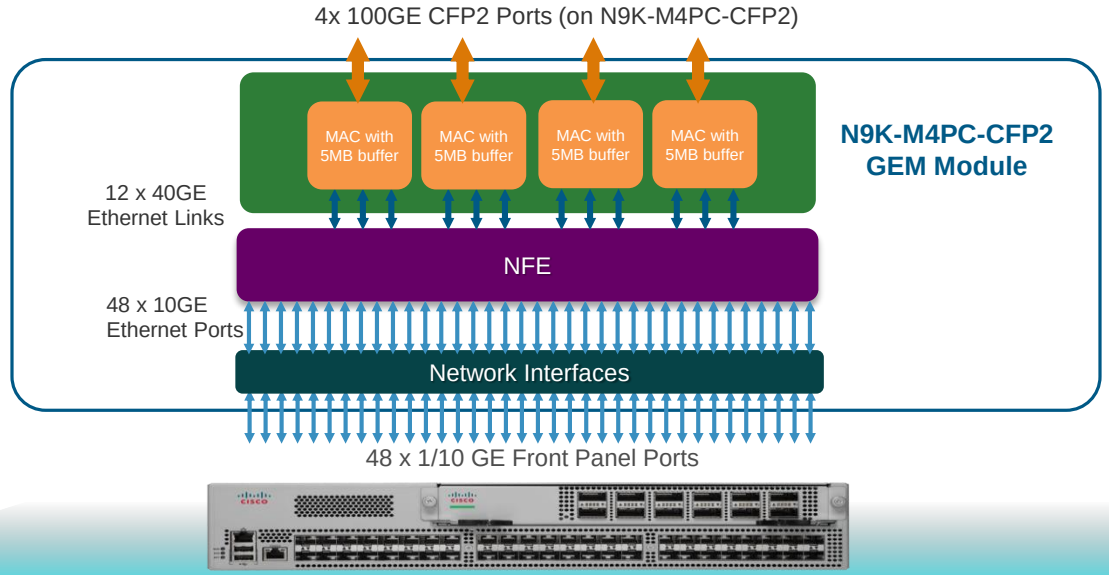
Nexus® 9396PX/TX with N9K-M12PQ GEM Module

Nexus® 9396PX/TX with N9K-M6PQ GEM Module

- Hardware is capable of VXLAN bridging and routing
- Hardware is capable of supporting both NX-OS and ACI
- Line rate performance for packet sizes > 200-Bytes

First Gen Nexus 9300 Series Switch Architecture

Nexus 9396PX/TX Block Diagram with N9K-M4PC-CFP2 GEM Module



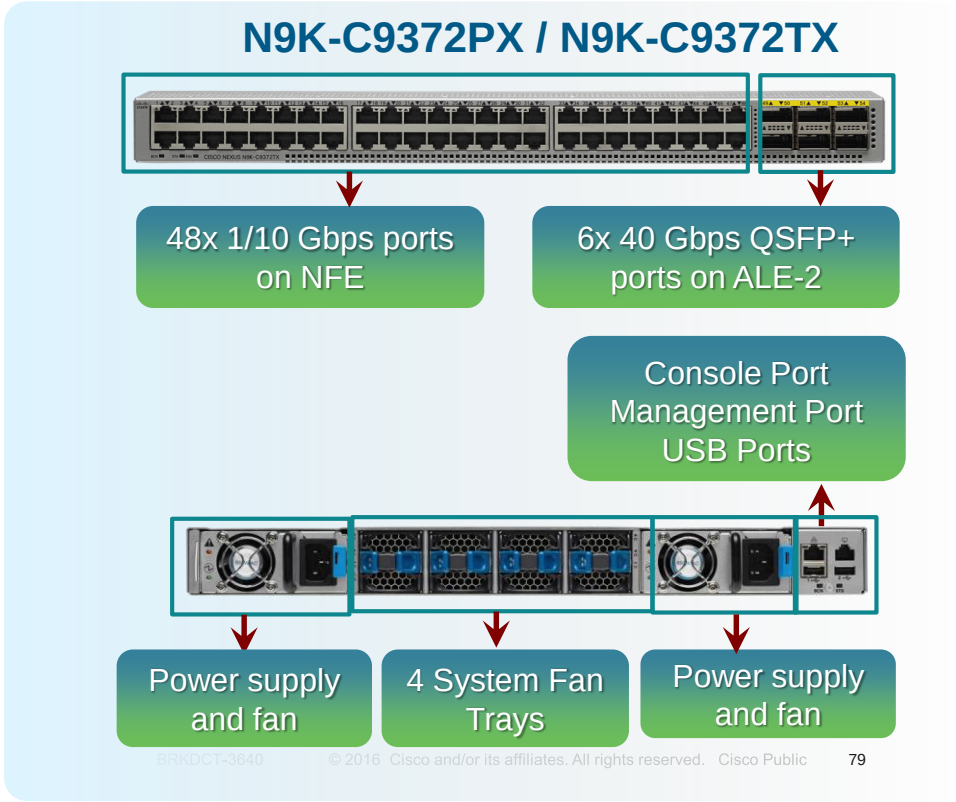
Nexus® 9396PX/TX with N9K-M4PC-CFP2 GEM Module

- Hardware is capable of VXLAN bridging only
- Hardware is capable of supporting NX-OS only
- Line rate performance for packet sizes > 200-Bytes

First Gen Nexus 9300 Series Switch Architecture

Cisco Nexus® 9372PX / 9372TX

- 1 RU height
- No GEM module
- 48x 1Gb SFP / 10 Gb SFP+ ports on Nexus 9372PX
- 48x 1/10 Gb Base-T ports on Nexus 9372TX
- 6x 40 Gb QSFP+ ports
- 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
- Full line rate performance for all packet sizes
- VXLAN bridging and routing
- Capable of supporting both NX-OS and ACI modes



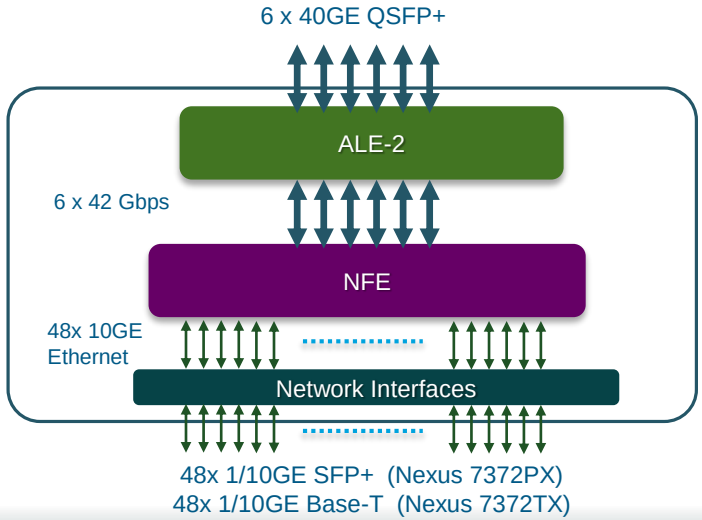
Nexus 9300 Series Switch Architecture

Nexus 9372PX/ Nexus 9372TX Block Diagram

1 application 1leaf engines (ALE-2)) for additional buffering and packet handling

1 network forwarding engine (NFE)

1 RU with redundant power supplies and fan.
6 QSFP+ 40GE ports and 40 SFP+ 10GE ports



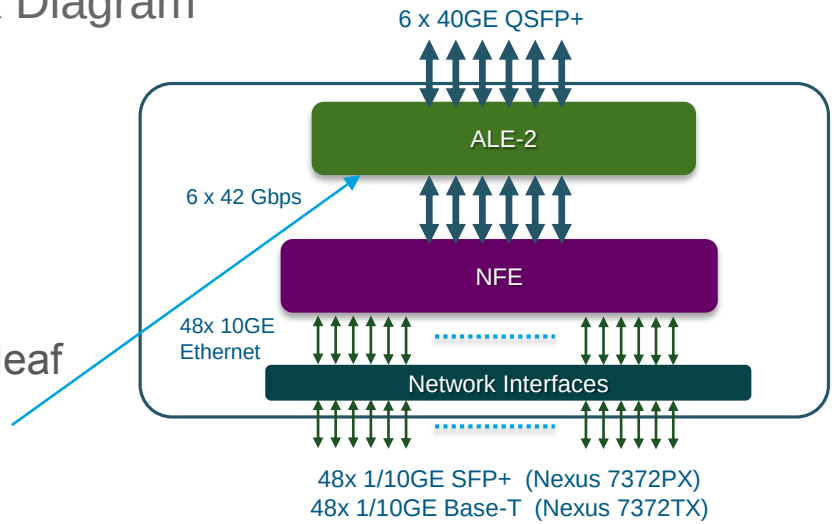
Nexus® 9372PX, Nexus 9372TX

- The 6 40GE links between NFE and ALE-2 run at 42Gbps clock rate to accommodate the internal packet header.
- Hardware is capable of VXLAN bridging and routing
- Hardware is capable of supporting both NX-OS and ACI modes
- Full line rate performance for all packet sizes

Nexus 9300 ’E’ Series

Nexus 9372PX-E/ Nexus 9372TX-E Block Diagram

- Support for IP and MAC based EPG in ACI mode for non VM’s
  - Support for VM Attribute including MAC/IP is supported on multiple vSwitches without the need for the ‘E’ leaf
- Allows static over-ride for the class-id (EPG) in the Local Station table



N9K-C9372TX
Show module information:

```
# sh mod
Mod Ports Module-Type                          Model      Status
--- ----- ---------------------------------- --------   --------

1   54    48x1/10G-T 6x40G Ethernet Modul  N9K-C9372TX  ctive *
```

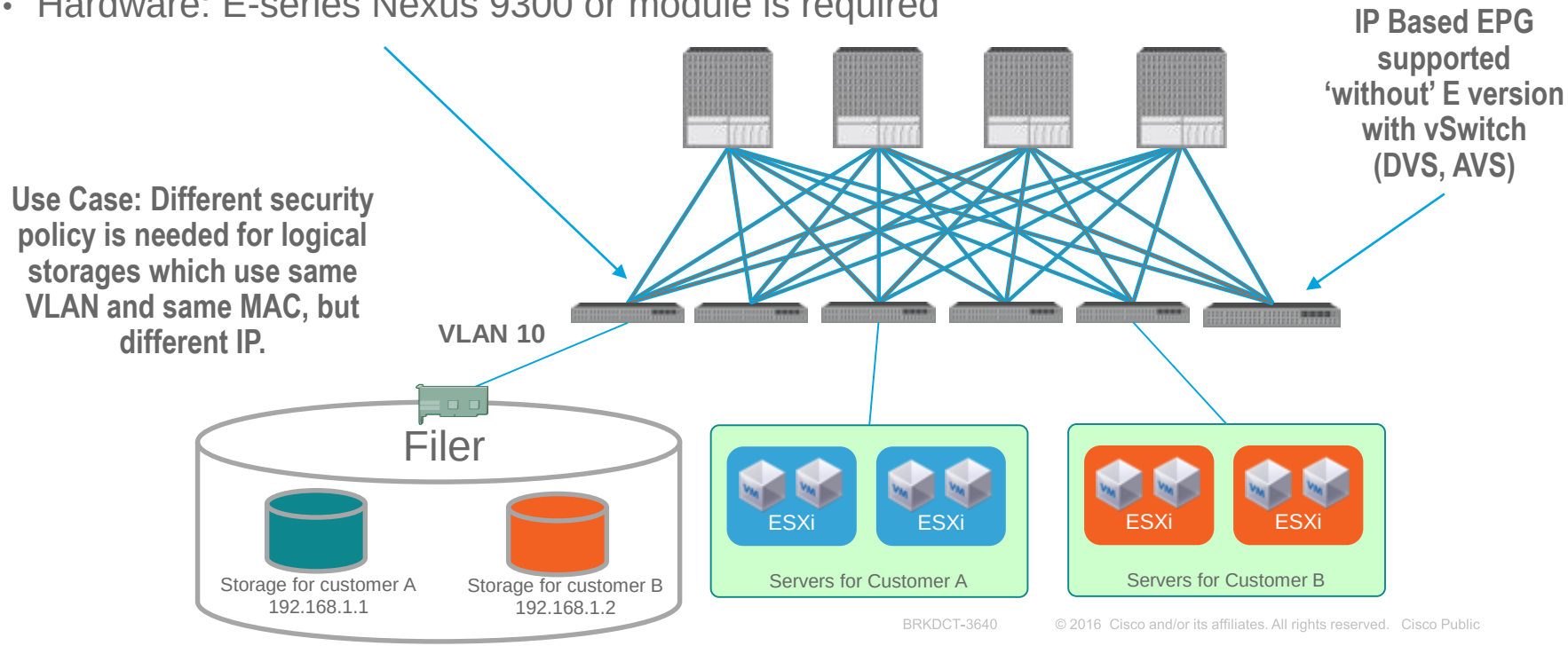
N9K-C9372TX-E
Show module information:

```
# sh mod
Mod Ports Module-Type                          Model          Status
--- ----- ---------------------------------- ---------------  --------

1   54    48x1/10G-T 6x40G Ethernet Module  N9K-C9372TX-E    active *
```

IP based EPG - Shared storage amongst customers

- With release 1.2(1), ACI provides IP based EPG classification on physical leaves for physical domain
- Hardware: E-series Nexus 9300 or module is required



Next Gen – 9200 & 9300EX
Q1CY16 – Q2CY16

Nexus 9300-EX

48p 10/25G SFP + 6p 40/100G QSFP
Nexus 93180YC-EX

48p 1/10GT + 6p 40/100G QSFP
Nexus 93108TC-EX

Dual personality – **ACI and NX-OS mode**
Industry's first native 25G VXLAN capable switch
Flexible port configurations – 1/10/25/40/50/100G
Up to 40 MB shared buffer
Native Netflow

Nexus 9200

36p 40/100G QSFP
Nexus 9236C

56p 40G + 8p 40/100G QSFP
Nexus 92304QC

72p 40G QSFP
Nexus 9272Q

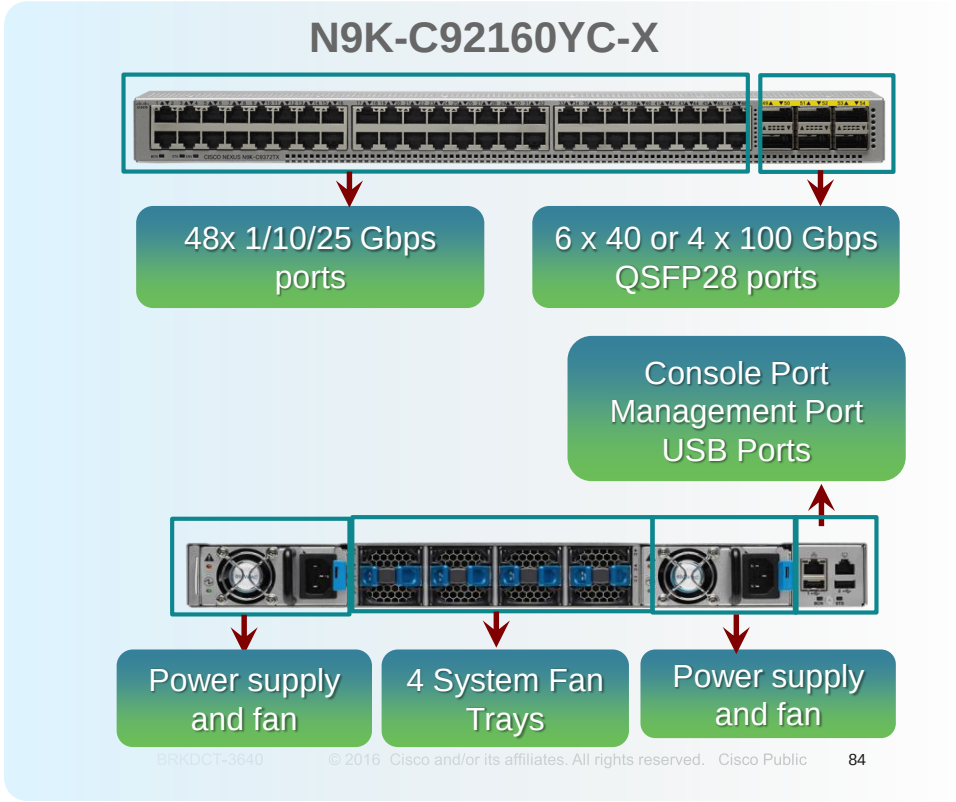
48p 10/25G SFP + 4p 100G/ 6p 40G QSFP
Nexus 92160YC-X

NX-OS switches
Industry's first 36p 100G 1RU switch
Industry's first native 25G VXLAN capable switch
Up to 30 MB shared buffer
High density compact 40/100G aggregation

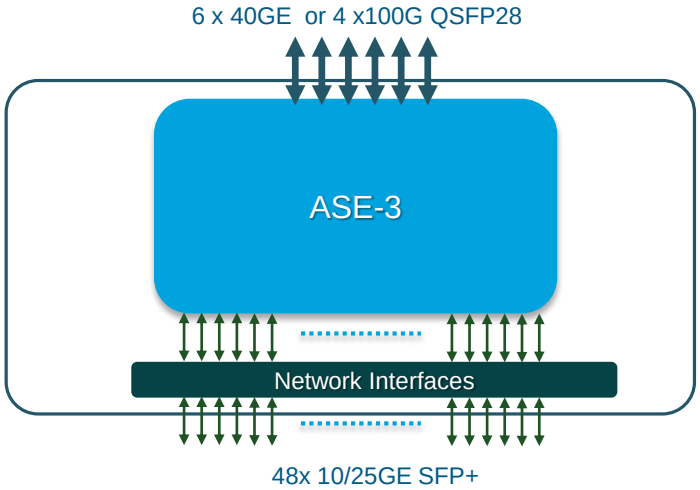
Nexus 9200 Series Switch Architecture
ASE3 Based

Cisco Nexus® Nexus 92160YC-X

- 1 RU height
- No GEM module
- 48 10/25 Gb SFP+ ports
- 6 x 40 Gbps or 4 x 100Gbps QSFP28 ports
- 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
- Full line rate performance for all packet sizes
- VXLAN bridging and routing
- Full Netflow
- Capable of supporting both NX-OS modes



Nexus 9200 Series
ASE3



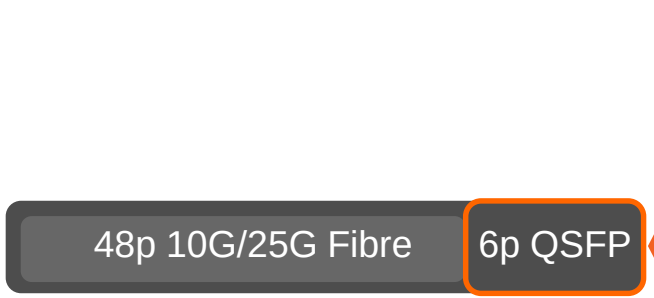
- ASIC: ASE3
- 1RU
- 2-core CPU (Intel Ivy Bridge Gladden)
- 2MB NVRAM
- 64MB
- Two Power supply (650W) 1 + 1 redundant
- Power consumption < 300 W
- Four Fans 3 + 1 redundant

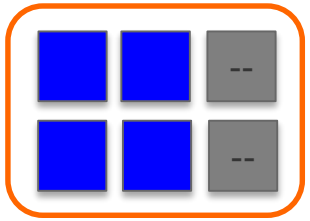
- 48 1/10/25GE SFP+ and 6 40GE QSFP or 4 100GE QSFP28 or 2 100GE QSFP28 + 4 40GE QSFP

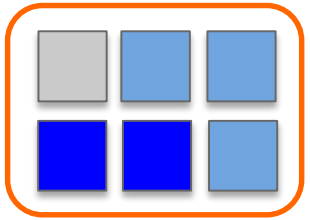
- Netflow and data analytics

Nexus 92160 Port Configuration

1RU 48 Port 10/25G Fibre + 6 Port 40G/ 4 Port 100G







Uplink Option 1 - 4p 100G w. dynamic break-outs

Port configuration supported:

- 4p 100G
- 16p 10G/25G
- 4p 40G
- QSA (roadmap)

Uplink Option 2 - 6p 40G w. 2p support 100G and dynamic break-outs

Port configuration supported:

- 6p 40G
- 2p 100G + 4p 40G
- 8p 10G/25G + 4p 40G
- QSA (roadmap)

48p individually configurable for 10 or 25G access

- Redundant 1+1 AC/DC Power supplies
- Redundant fan modules
- No ACI Support
- NEBS Certified

Nexus 92160 Port Configuration

1RU 48 Port 10/25G Fibre + 6 Port 40G/ 4 Port 100G

CLI to find the operation mode:

```
drvly15(config-if-range)# sh running-config | grep portmode
hardware profile portmode 48x25G+2x100G+4x40G
```

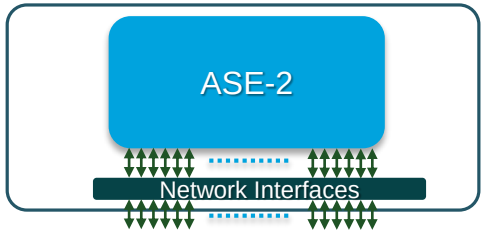
```
92160# sh mod
Mod Ports          Module-Type                    Model         Status
--- ----- ---------------------------------- --------------------- ---------
1    54   48x10/25G+(4x40G+2x100G or 4x100G) Et N9K-C92160YC         active *
```

- Breakout modes
- There are two breakout modes
  - 40G to 4x10G breakout.
    - This breaks out 40G ports into 4 X 10G ports
    - Cli command
- interface breakout module 1 port <x> map 10g-4x
  - 100G to 4x25G breakout.
    - This breaks out 100G ports into 4 X 25G ports
    - Cli command
- interface breakout module 1 port <x> map 25g-4x

ASIC Route Scale

	ASE2 3.6T / 6 slices N9200	ASE3 1.6T / 2 slices N9200	LSE 1.8T / 2 slices N9300EX/X9700EX	T2 1.28T/ 1 slice N3100	Tomahawk 3.2T / 4 slices N3200	Jericho ***	
IPv4 Prefix (LPM)	256K*	256K*	750K*	192K*	128K*	192K*	LPM
IPv6/64 Prefix (LPM)	256K*	256K*	750K*	84K*	84K*	64K*	
IPv6 Prefix /128 (LPM)	128K*	128K*	384K*	20K*	20K*	64K*	
IPv4 host routes	256K*	256K*	750K*	120K*	104K*	750K**	Host
IPv6 host routes	128K*	128K*	384K*	20K*	20K*	64K*	
MAC	256K*	256K*	512K*	288K*	136K*	750K**	
Flow Table	No	Yes	Yes	No	No	No	

Nexus 9200 Series Switch Architecture
ASE2 Based

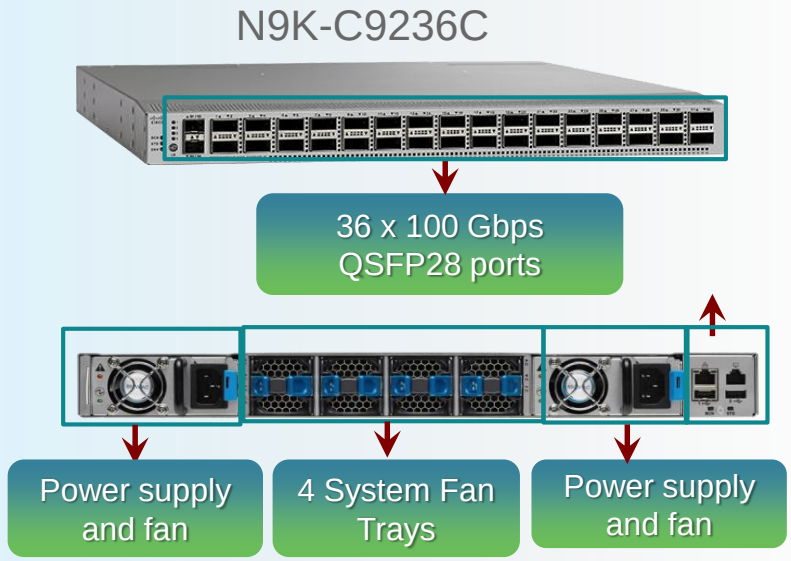


Cisco Nexus® Nexus N9K-C9236C

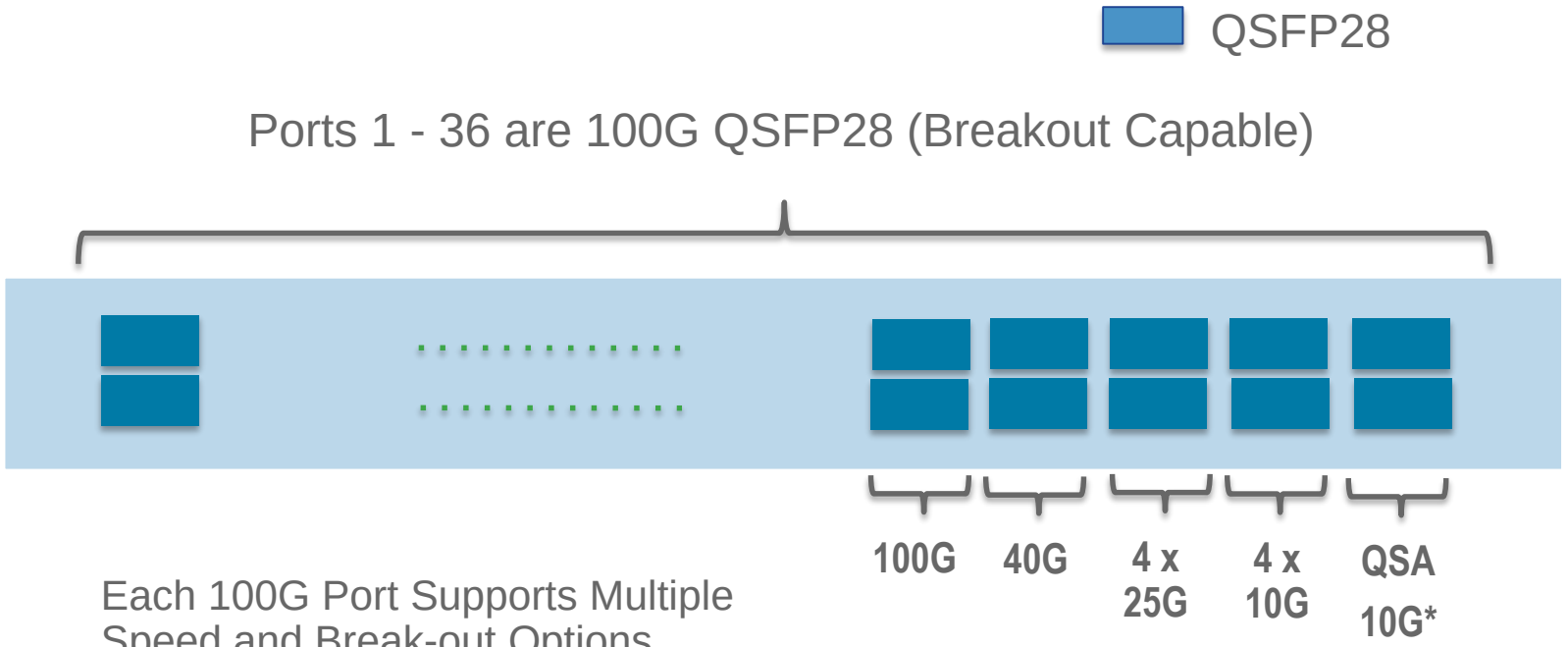
- ASIC: ASE2
- 4-core CPU (Intel Ivy Bridge Gladden 4 core at 1.8 GHz)
- 8G DIMM memory
- 2MB NVRAM
- Two Power supply (1200W) 1 + 1 redundant
- Power consumption 450 W
- Two Fans 3 + 1 redundant
- 36 x 40/100G ports
- 144 10/25G ports (when all ports in breakout mode

Each 100G Port Break-out Options:

- 1p 100G → SR, AOC
- 1p 40G
- 4p 10G
- 4p 25G → 1m, 3m
- QSA (roadmap)



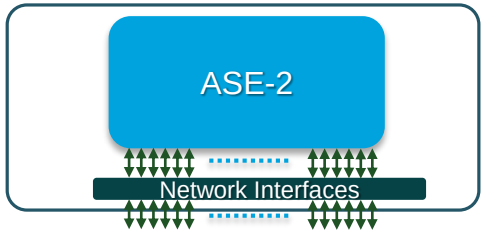
Nexus 9236C Port Configuration
1 RU 36 Port 100G Fibre



Each 100G Port Supports Multiple Speed and Break-out Options

* (QSA in a future SW release)

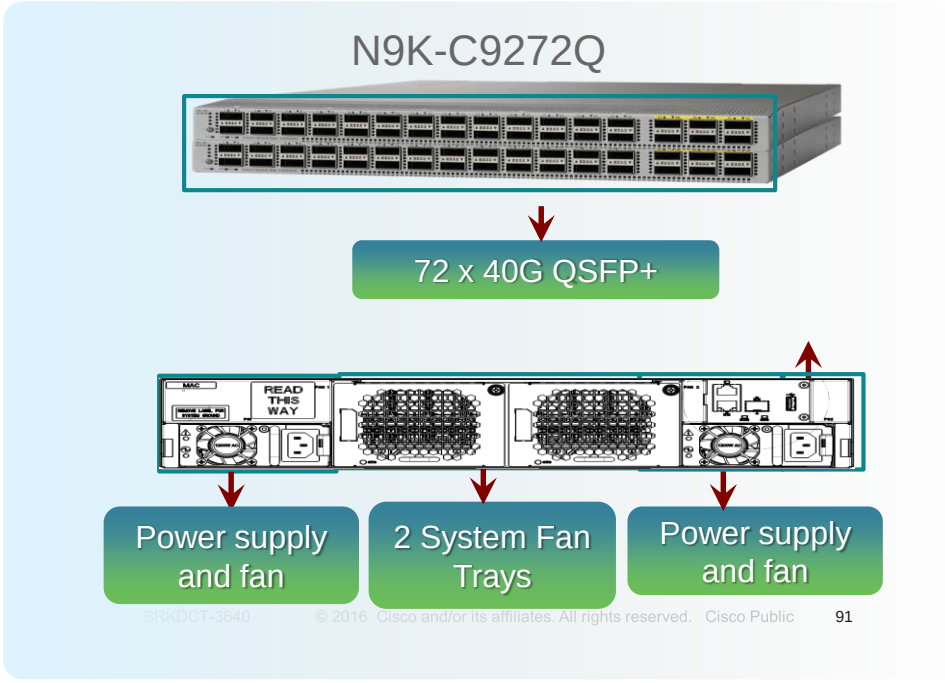
Nexus 9200 Series Switch Architecture
ASE2 Based



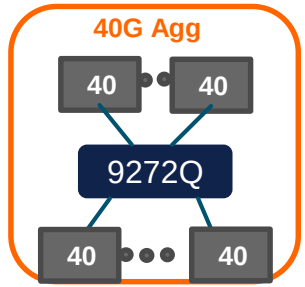
Cisco Nexus® Nexus N9K-C9272Q

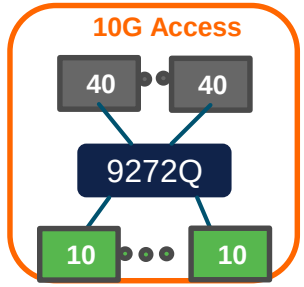
- ASIC: ASE2
- 2RU
- 4-core CPU (Intel Ivy Bridge Gladden 4 core at 1.8 GHz)
- 8G DIMM memory
- 2MB NVRAM
- 64MB
- Two Power supply (1200W) 1+1 redundant
- Power consumption 650 W
- Two Fans 1 + 1 redundant

- 72 QSFP ports
- The top 36 QSFP ports operate at 40GE only, not breakout capable
- The bottom 36 QSFP ports can be 4x 10GE



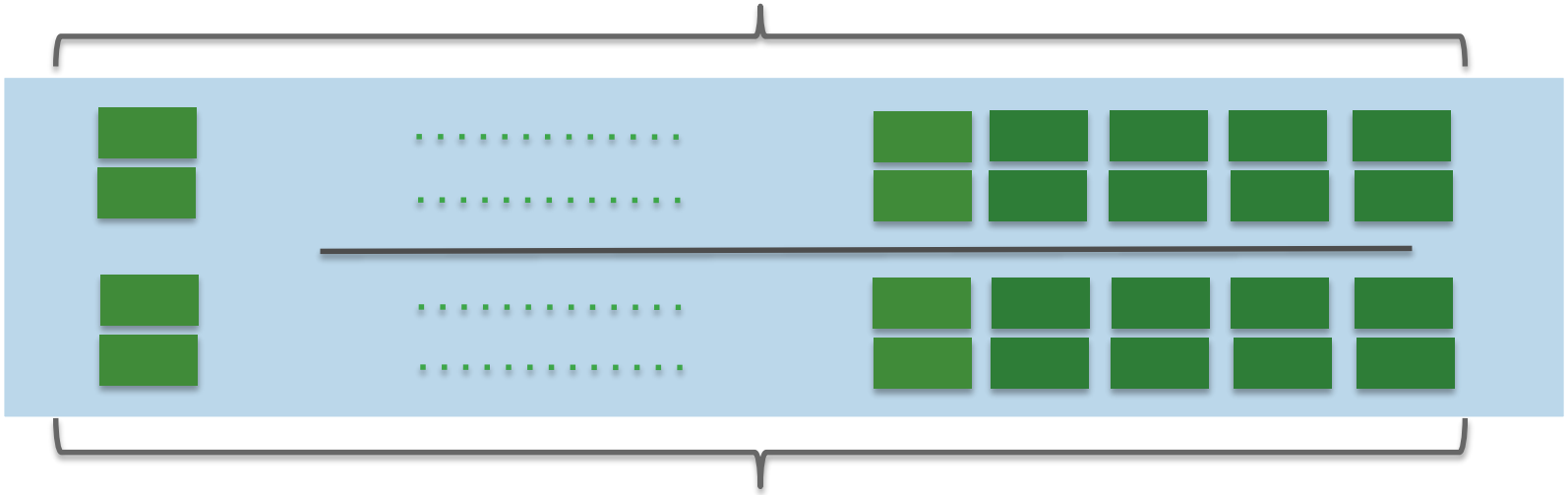
Nexus 9272Q Port Configuration
2RU 72 Port 40G Fibre





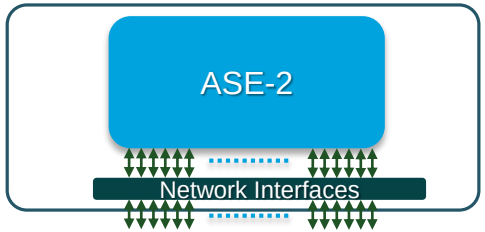
QSFP+

Ports 1 - 36 are 40G QSFP+



Ports 37 - 72 are 40G QSFP+ (Breakout Capable 144 x 10G)

Nexus 9200 Series Switch Architecture
ASE2 Based

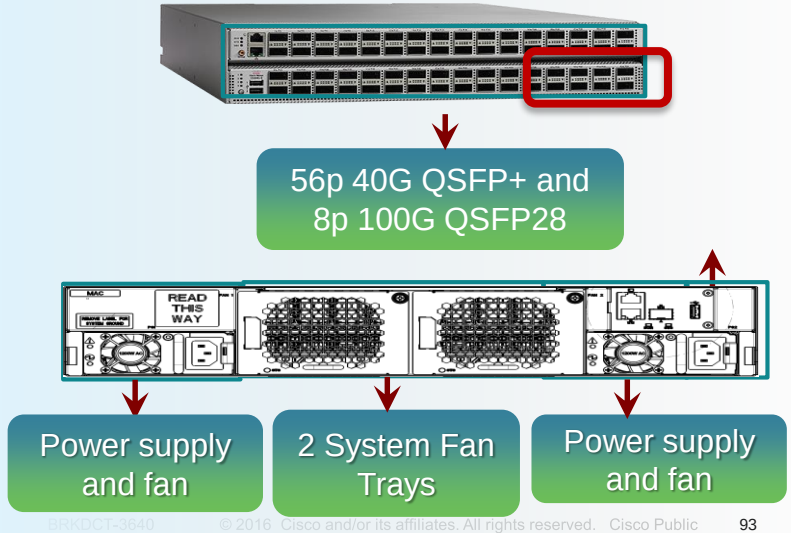


Cisco Nexus® Nexus N9K-C92304QC

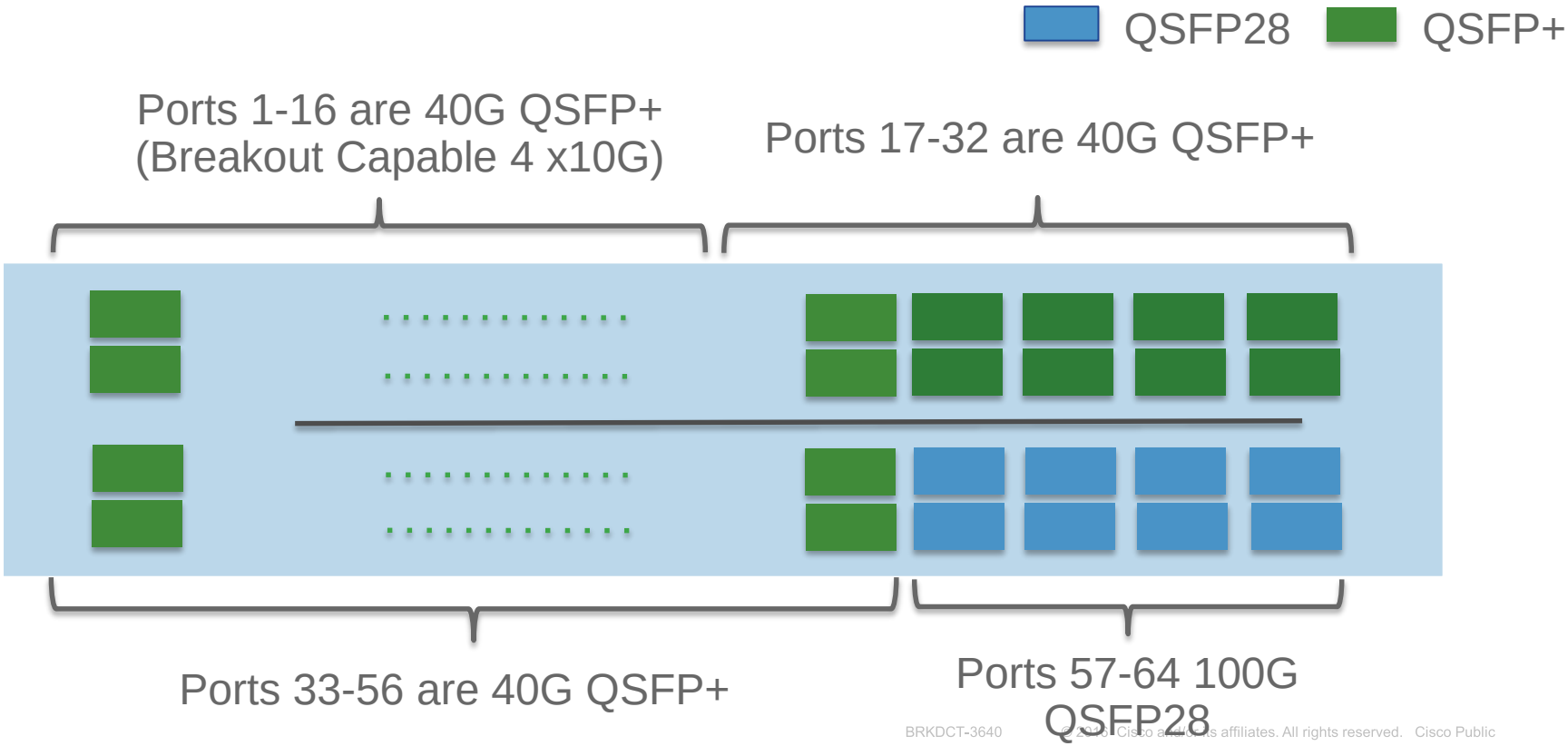
- ASIC: ASE2
- 2RU
- 4-core CPU (Intel Ivy Bridge Gladden 4 core at 1.8 GHz)
- 8G DIMM memory
- 2MB NVRAM
- 64MB
- Two Power supply (1200W) 1+1 redundant
- Power consumption 650 W
- Two Fans 1 + 1 redundant

- 56 40GE QSFP ports
- The first 16 QSFP ports is breakout capable to 4x 10GE
- 8 100GE QSFP28 ports

N9K-C92304QC



Nexus 92304QC Port Configuration
2RU 56p 40G Fibre + 8p 40G/00G



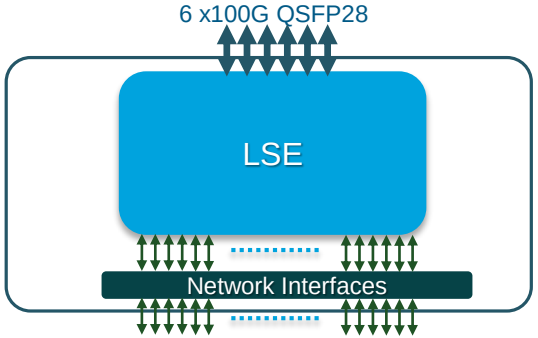
ASIC Route Scale

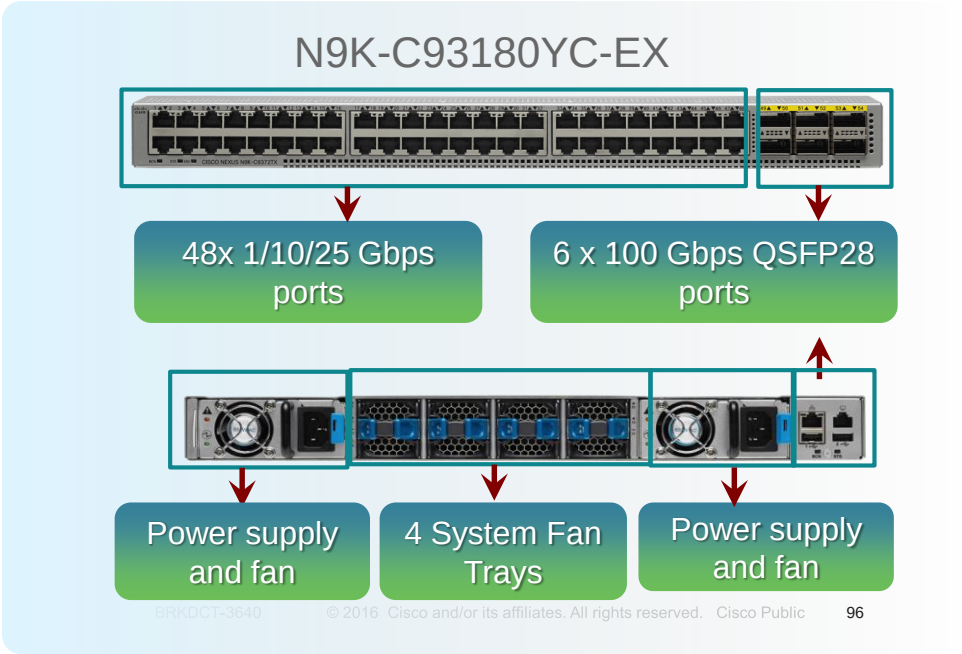
	ASE2 3.6T / 6 slices N9200	ASE3 1.6T / 2 slices N9200	LSE 1.8T / 2 slices N9300EX/X9700EX	T2 1.28T/ 1 slice N3100	Tomahawk 3.2T / 4 slices N3200	Jericho ***	
IPv4 Prefix (LPM)	256K*	256K*	750K*	192K*	128K*	192K*	LPM
IPv6/64 Prefix (LPM)	256K*	256K*	750K*	84K*	84K*	64K*	
IPv6 Prefix /128 (LPM)	128K*	128K*	384K*	20K*	20K*	64K*	
IPv4 host routes	256K*	256K*	750K*	120K*	104K*	750K**	Host
IPv6 host routes	128K*	128K*	384K*	20K*	20K*	64K*	
MAC	256K*	256K*	512K*	288K*	136K*	750K**	
Flow Table	No	Yes	Yes	No	No	No	

Nexus 9300EX Series
LSE Based

Cisco Nexus® Nexus N9K-C93180YC-EX

- ASIC: LSE
- 1RU
- 2-core CPU (Intel Ivy Bridge Gladden)
- 2MB NVRAM
- 64MB
- Two Power supply (650W) 1 + 1 redundant
- Power consumption 248 W
- Four Fans 3 + 1 redundant
- 48 x 10G/25G SFP28 and 6 x 40G/100G QSFP28
- Support both NX-OS mode and ACI mode (ACI leaf)
- Netflow and data analytics





ASIC Route Scale

	ASE2 3.6T / 6 slices N9200	ASE3 1.6T / 2 slices N9200	LSE 1.8T / 2 slices N9300EX/X9700EX	T2 1.28T/ 1 slice N3100	Tomahawk 3.2T / 4 slices N3200	Jericho ***	
IPv4 Prefix (LPM)	256K*	256K*	750K*	192K*	128K*	192K*	LPM
IPv6/64 Prefix (LPM)	256K*	256K*	750K*	84K*	84K*	64K*	
IPv6 Prefix /128 (LPM)	128K*	128K*	384K*	20K*	20K*	64K*	
IPv4 host routes	256K*	256K*	750K*	120K*	104K*	750K**	Host
IPv6 host routes	128K*	128K*	384K*	20K*	20K*	64K*	
MAC	256K*	256K*	512K*	288K*	136K*	750K**	
Flow Table	No	Yes	Yes	No	No	No	

Nexus 9300/ 9200/ 3000 Naming Convention

Nexus 9300

- 9300EX Cisco ASIC/ ACI Leaf/ VXLAN Routing/ Analytics/ 100G uplinks
- 9300 Cisco ASIC + BRCM T2/ ACI Leaf/ VXLAN Routing/ 40G uplinks

Nexus 9200

- 9200X Cisco ASIC/ VXLAN Routing/ Analytics
- 9200 Cisco ASIC/ VXLAN Routing

Nexus 3000

- 3100 Trident 2/ VXLAN bridging
- 3100-V Trident 2+/ VXLAN routing/ 100G uplinks
- 3200 Tomahawk/ VXLAN bridging

Naming Conventions

Nexus 9000

-E enhanced

-X analytics/ Netflow

-S 100G Merchant

Nexus 3000

-XL 8G DRAM

-V VXLAN Routing

Port Speed

PX: 1/10G SFP+

TX: 100M/1G/10GT

Y: 10/25G SFP+

Q: 40G QSFP+

C: 100G QSFP28

Aggregate Bandwidth

If same speed ports then its the # of ports

- N9K-C9232Q – 32p 40G
- N9K-C9236C – 36p 100G

If mix speed ports then its the sum across all ports

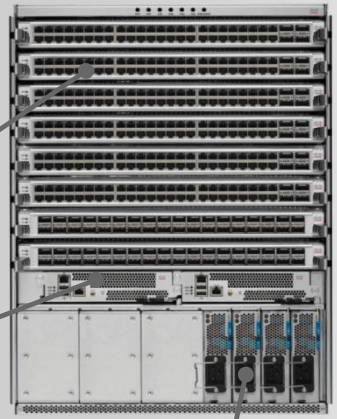
- N9K-C93180YC – 48p 25G + 6p 100G = 1800G
- N9K-C31108PC – 48p 10G + 6p 100G =1080G

Agenda

- Existing and New Nexus 9000 & 3000
- What's New
  - Moore's Law and 25G SerDes
  - The new building blocks (ASE-2, ASE-3, LSE)
  - Examples of the Next Gen Capabilities
- Nexus 9000 Switch Architecture
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)
- 100G Optics

Nexus 9500 Platform Architecture

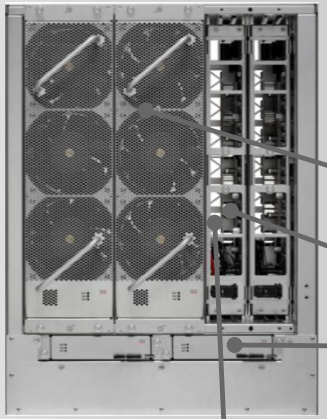
Nexus® 9508 Front View



8 line card slots
Max 3.84 Tbps per slot
duplex

Redundant
supervisor engines

Nexus 9508 Rear View



3 fan trays, front-to-back airflow

3 or 6 fabric modules
(behind fan trays)

Redundant system controller cards

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

No mid-plane for
LC-to-FM connectivity

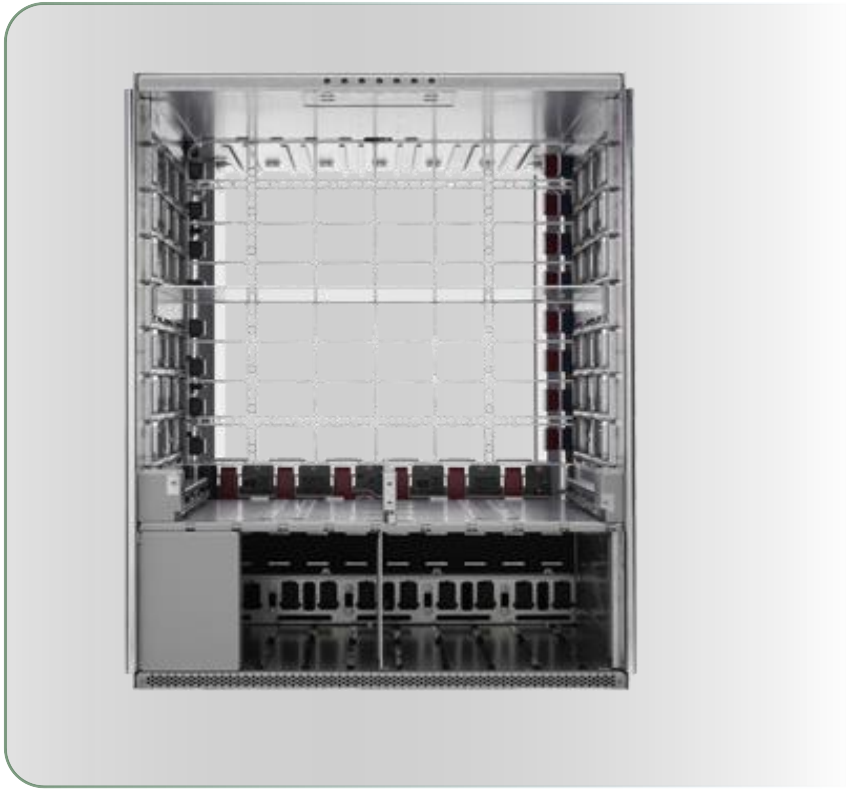
Chassis Dimensions: 13 RU x 30 in. x 17.5 in (HxWxD)

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

Nexus 9500 Platform Architecture

Chassis Design: No Mid-Plane

- Designed for:
  - Power & Cooling Efficiency
  - Designed for Reliability
  - Designed for Future Scale
- Current Chassis, Supervisor, Power Supply, Fan Trays, Systems Controller will support all New Line Cards and Fabric Modules

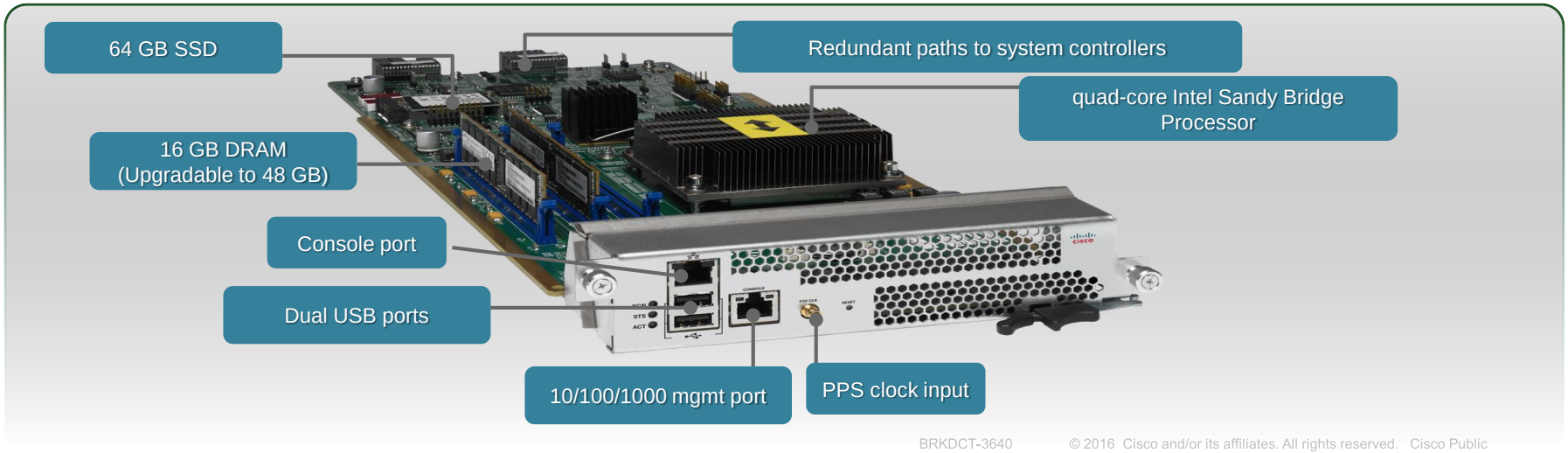


Nexus 9500 Platform Architecture

Supervisor Module Sup-A

- 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

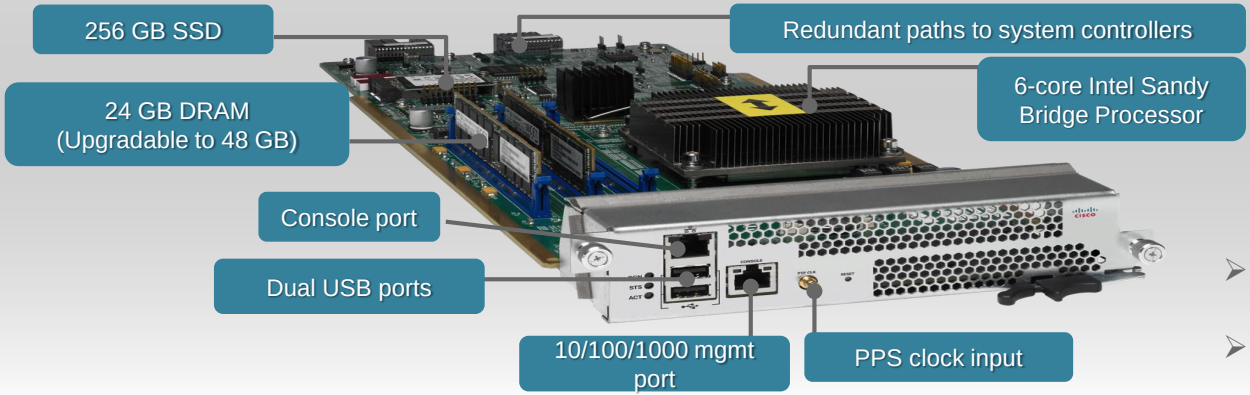


Nexus 9500 Platform Architecture

Supervisor Module Sup-B

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

Supervisor Module	
Processor	2.1 GHz, 6 cores 2.2GHz IVY Bridge
System Memory	24 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	256 GB



- 50% more CPU power
- 50% more memory space
- 300% more SSD storage

- Increase control protocols performance and convergence time.
- Ready for application intensive deployment

Nexus 9500 Platform Architecture System Controller

- 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
Fan Trays

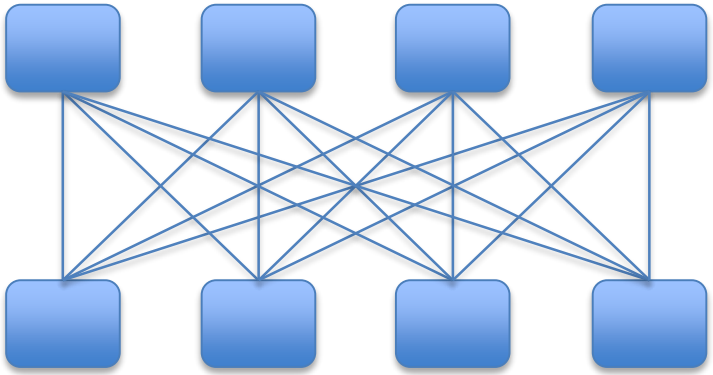
Fan trays are installed after the Fabric Module.

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

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

Modular Nexus 9500
A CLOS Based SOC Architecture (Leaf and Spine)



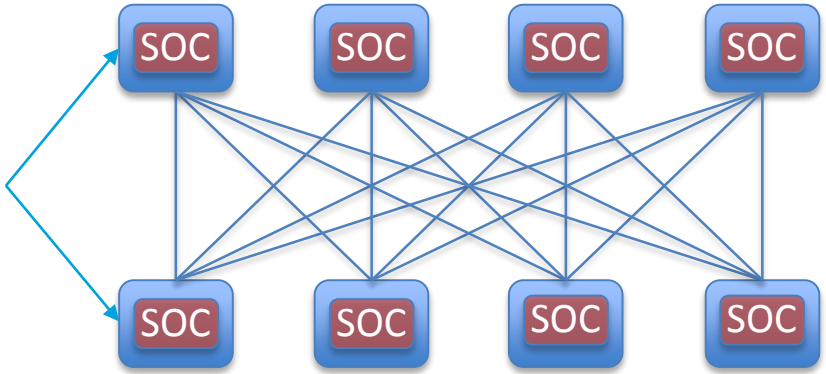
Spine

Leaf

Non Blocking CLOS

Modular Nexus 9500
A CLOS Based SOC Architecture (Leaf and Spine)

Leverage Switch on Chip (SOC) Based components



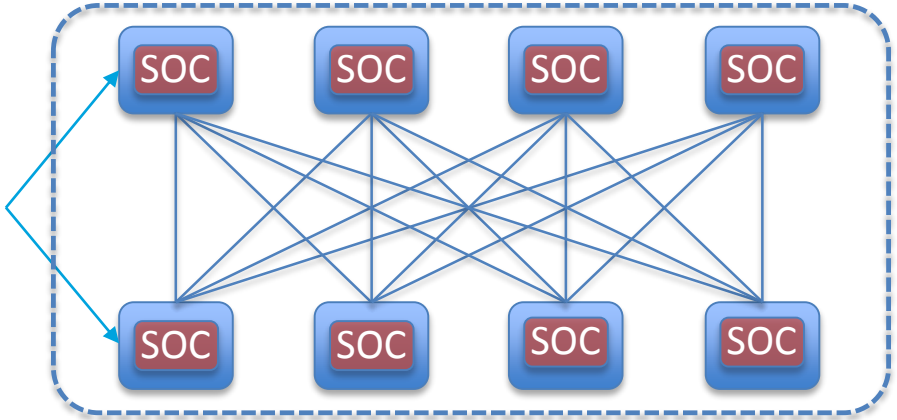
Spine

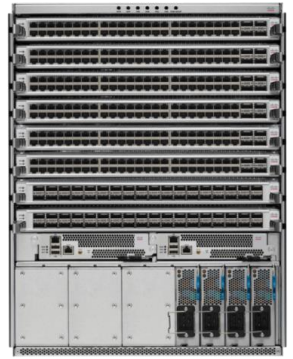
Leaf

Non Blocking CLOS

Modular Nexus 9500
A CLOS Based SOC Architecture (Leaf and Spine)

Leverage Switch on Chip (SOC) Based components



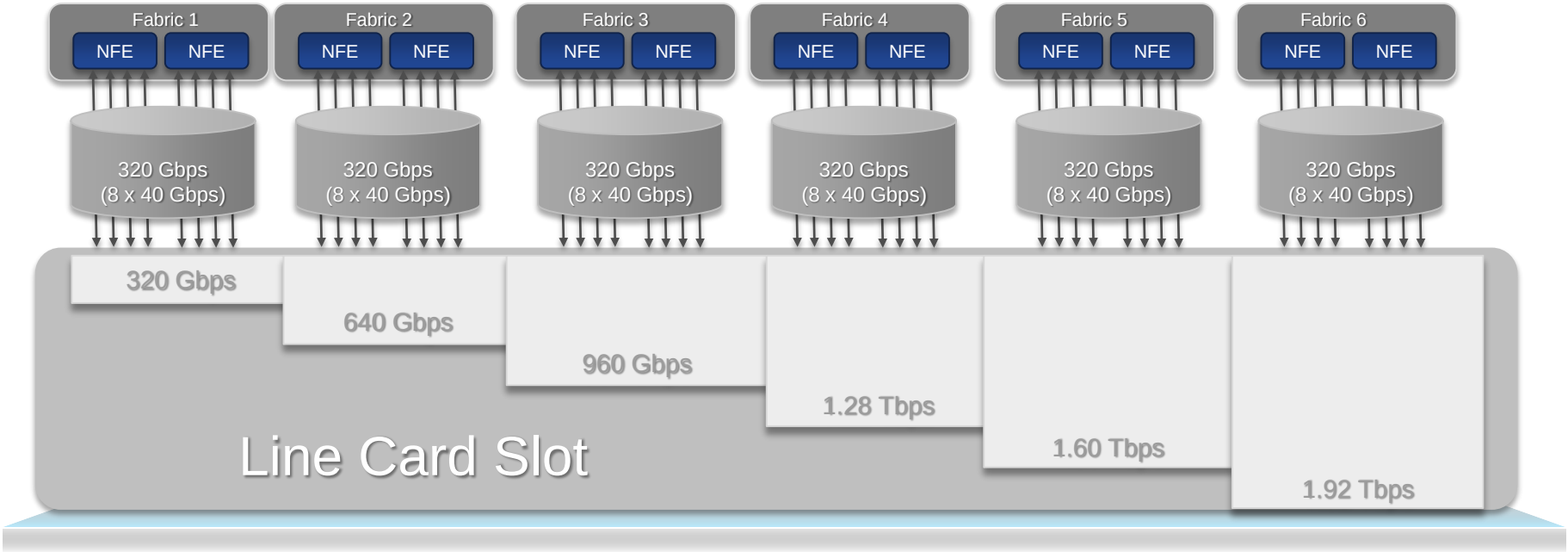


Non Blocking CLOS

First Gen Nexus 9500 Series Switch Fabric Module

Data Plane Scaling (Using Nexus 9508 as an example)

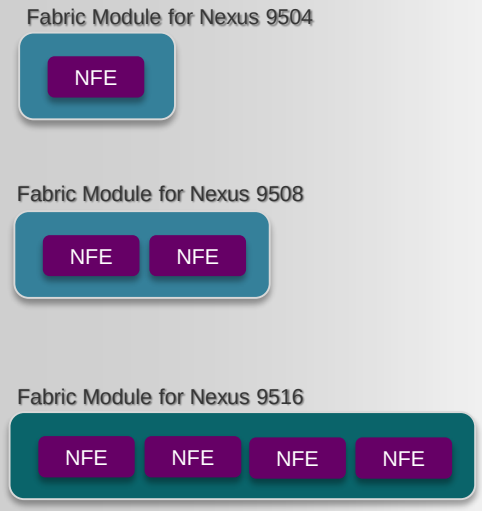
- Each fabric module can provide up to 320 Gbps to each line card slot
- With 6 fabric modules, each lie card slot can have up to 1.92 Tbps forwarding bandwidth in each direction.



First Gen Nexus 9500 Series Switch Fabric Module

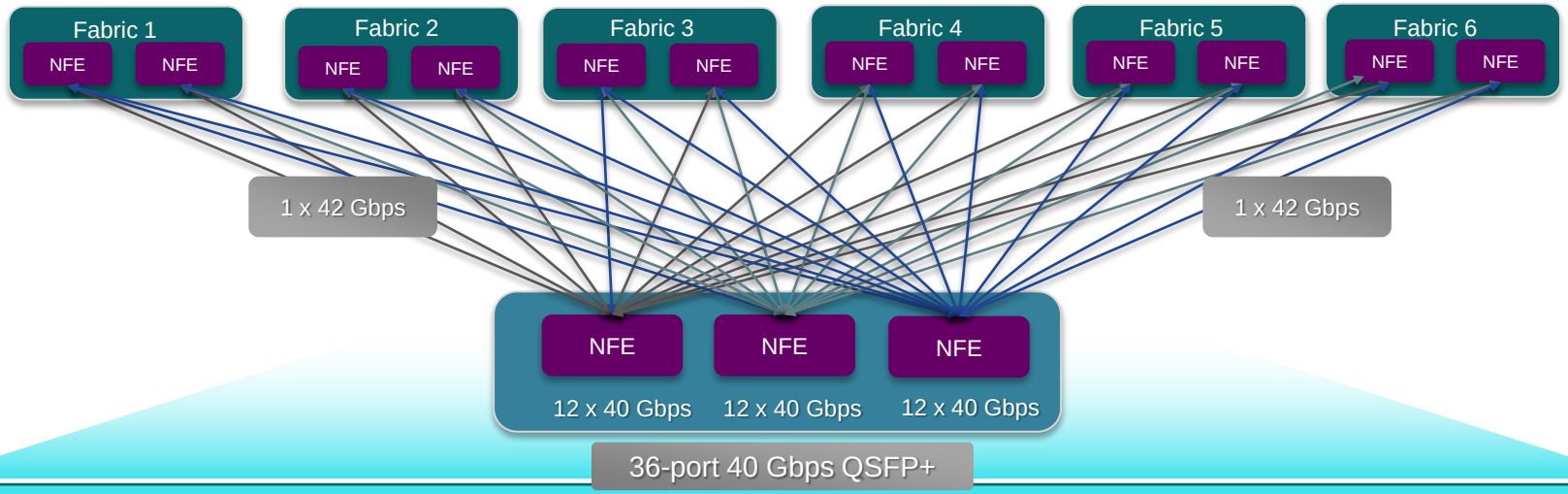
- Interconnect I/O module slots
- Installed at the rear of the chassis
- Uses Broadcom T2 as the network forwarding engine (NFE)
- Up to 3.84 Tbps duplex per line card slot
- All fabric cards are active and carry traffic

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



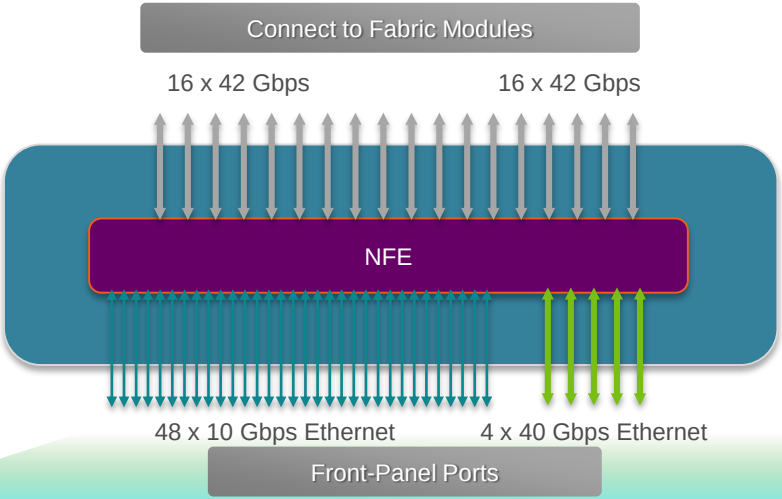
Nexus 9500 N9K-X9600 Series Line Cards

- 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-X9400 Series Line Cards



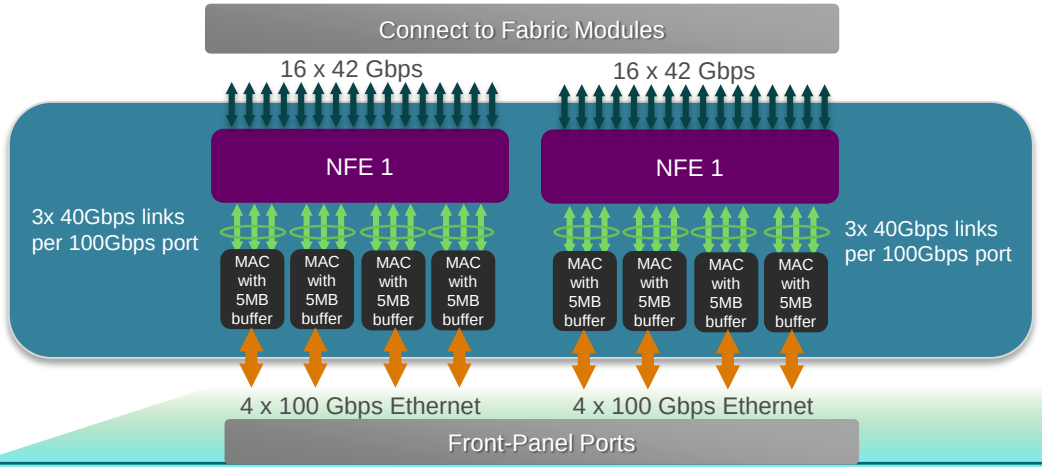
Internal 40G links are running at 42 Gbps clock rate to compensate for the 16-byte internal frame header

N9K-X9464PX/TX line cards are supported in all Nexus 9500 chassis types.

- 1.92 Tbps full-duplex fabric connectivity
- Full Layer-2 and Layer-3 feature sets
- Hardware supports 4x 10 Gbps break-out mode on 40 Gbps ports
- Cisco® NX-OS mode only
- Supported by all Nexus 9500 chassis, including Nexus 9504, 9508 and 9516
- Operate with 4 fabric modules for maximum performance (in fabric module slots 2, 3, 4 and 6)

Nexus 9500 N9K-X9400 Series Line Cards

N9K-X9408PC-CFP2



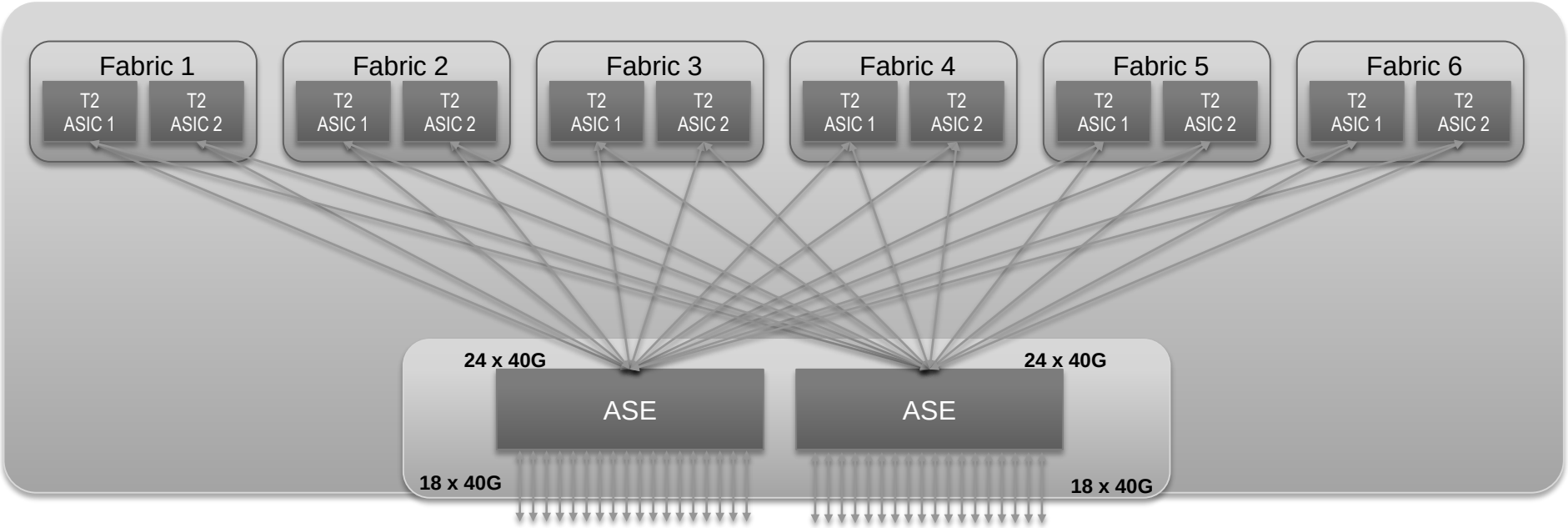
Internal links to fabric modules are running at 42 Gbps clock rate to compensate for the 16-byte internal frame header

N9K-X9408PC-CFP2 is supported in all Nexus 9500 chassis types.

- Two network forwarding engines (NFE)
- Each NFE supports 4x 100 Gbps front panel ports
- Oversubscribed for small packets (<193 Bytes)
- Line rate performance for larger packet sizes (> 193 Bytes)
- Each 100GE front panel port is essentially 3x 40GE ports on NFE
- Supports up to 40GE flows
- The 100GE MAC ASIC per front panel port has additional 5MB buffer
- Requires 4 fabric modules for maximum bandwidth (in fabric module slots 2, 3, 4 & 6)

Fabric Spine Switch Architecture

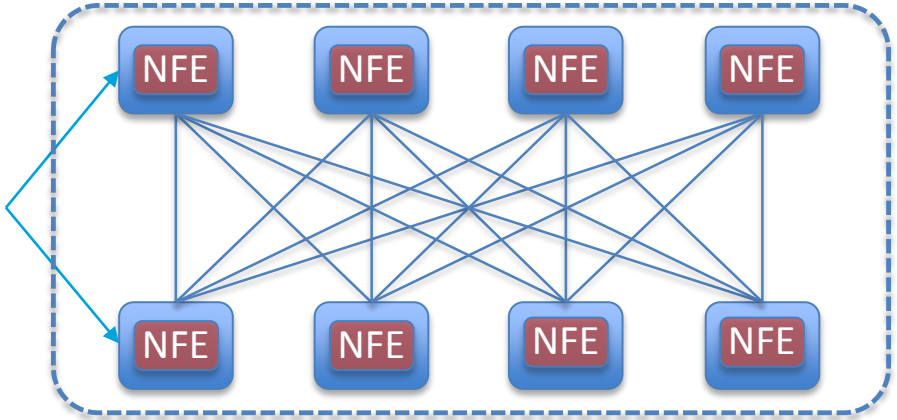
8-Slots (6 Fabrics) – 36 x 40G Line Card



36-port 40G QSFP
Line Rate
(FCS)

Modular Nexus 9500
Hierarchical Forwarding Lookups

Leverage Switch on Chip (SOC) Based components



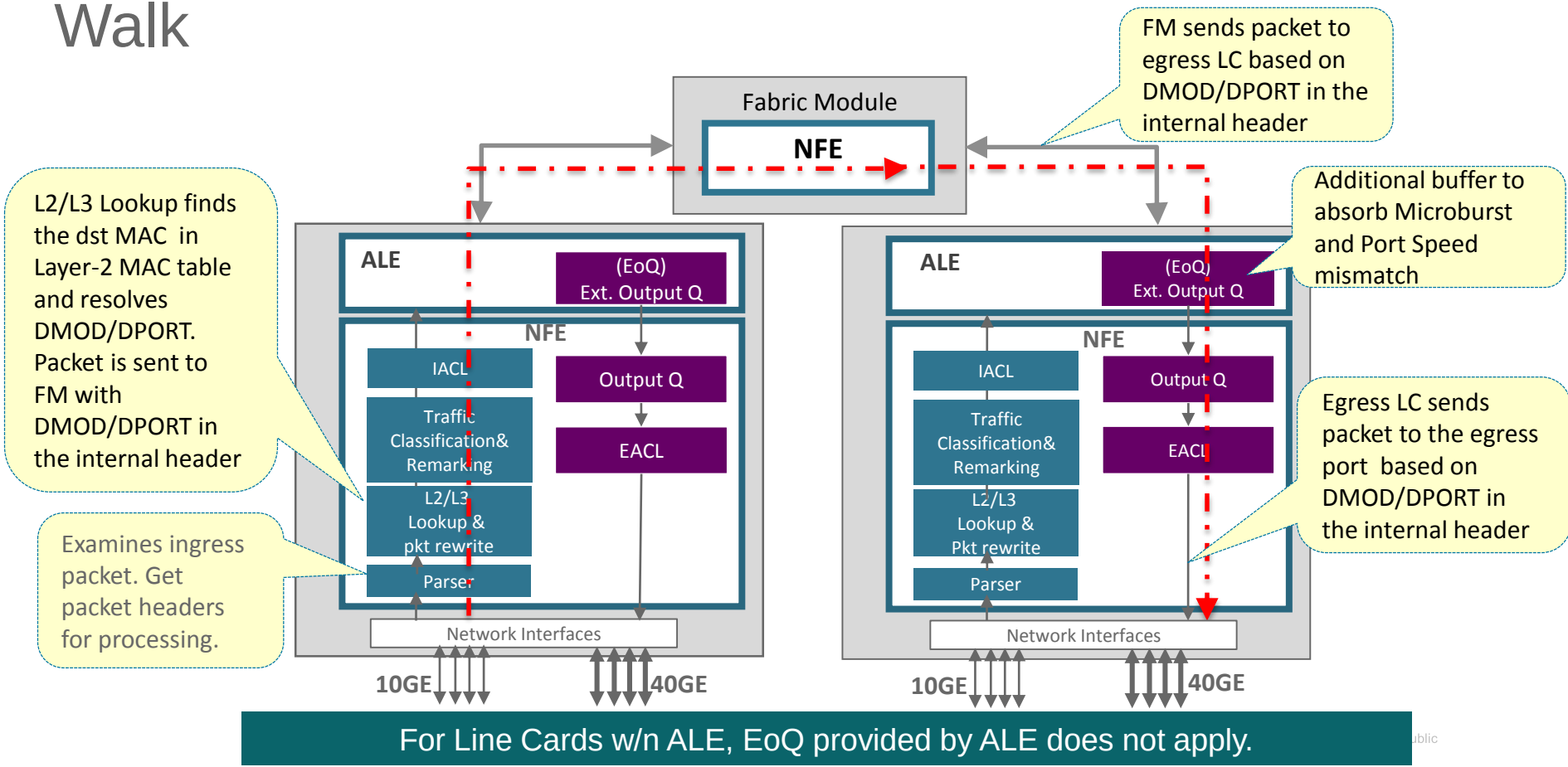
Network (LPM) Routes

Host Routes

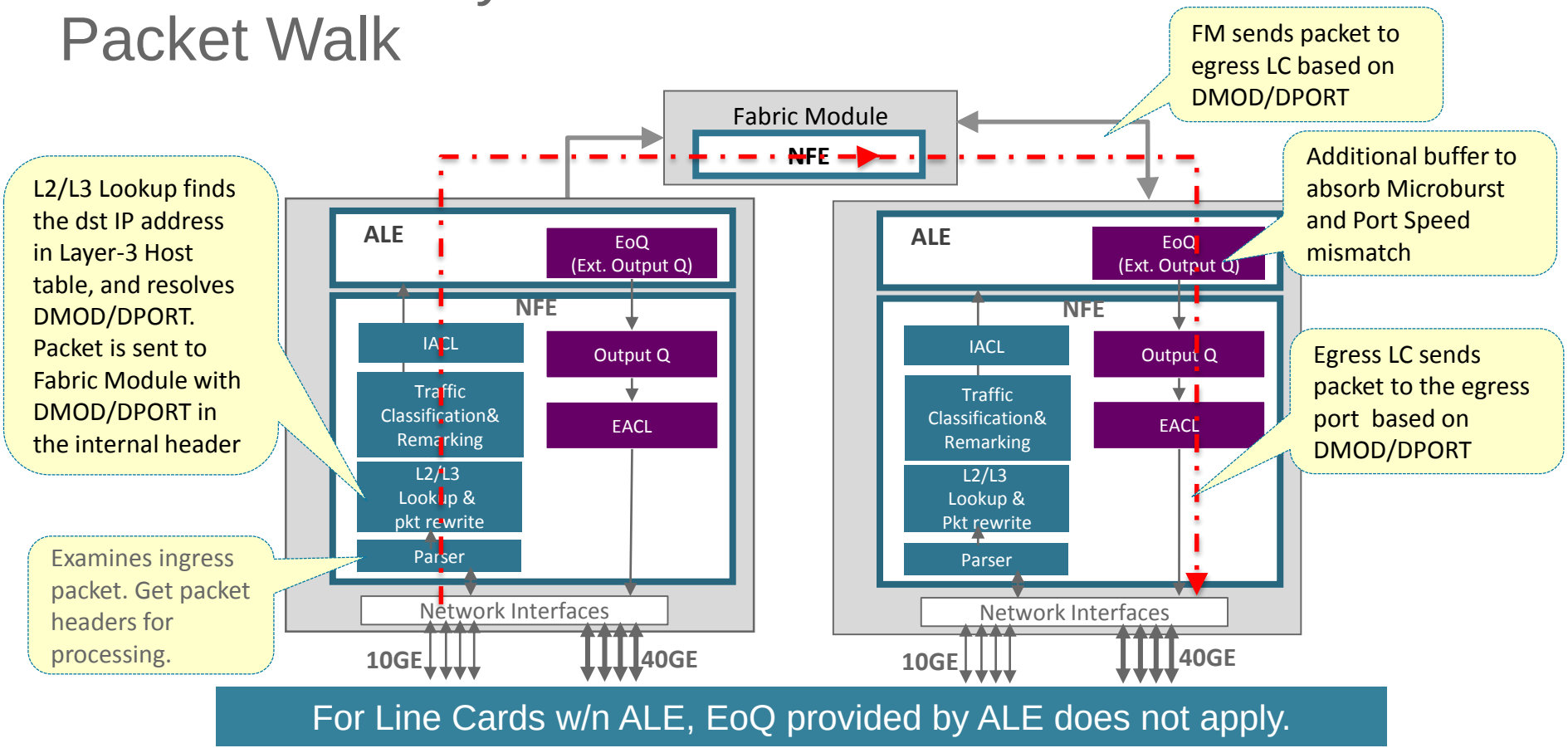
Non Blocking CLOS

Summarisation and Balance of Host and Network Routes Shift

Nexus 9500 Layer-2 Unicast Packet Walk

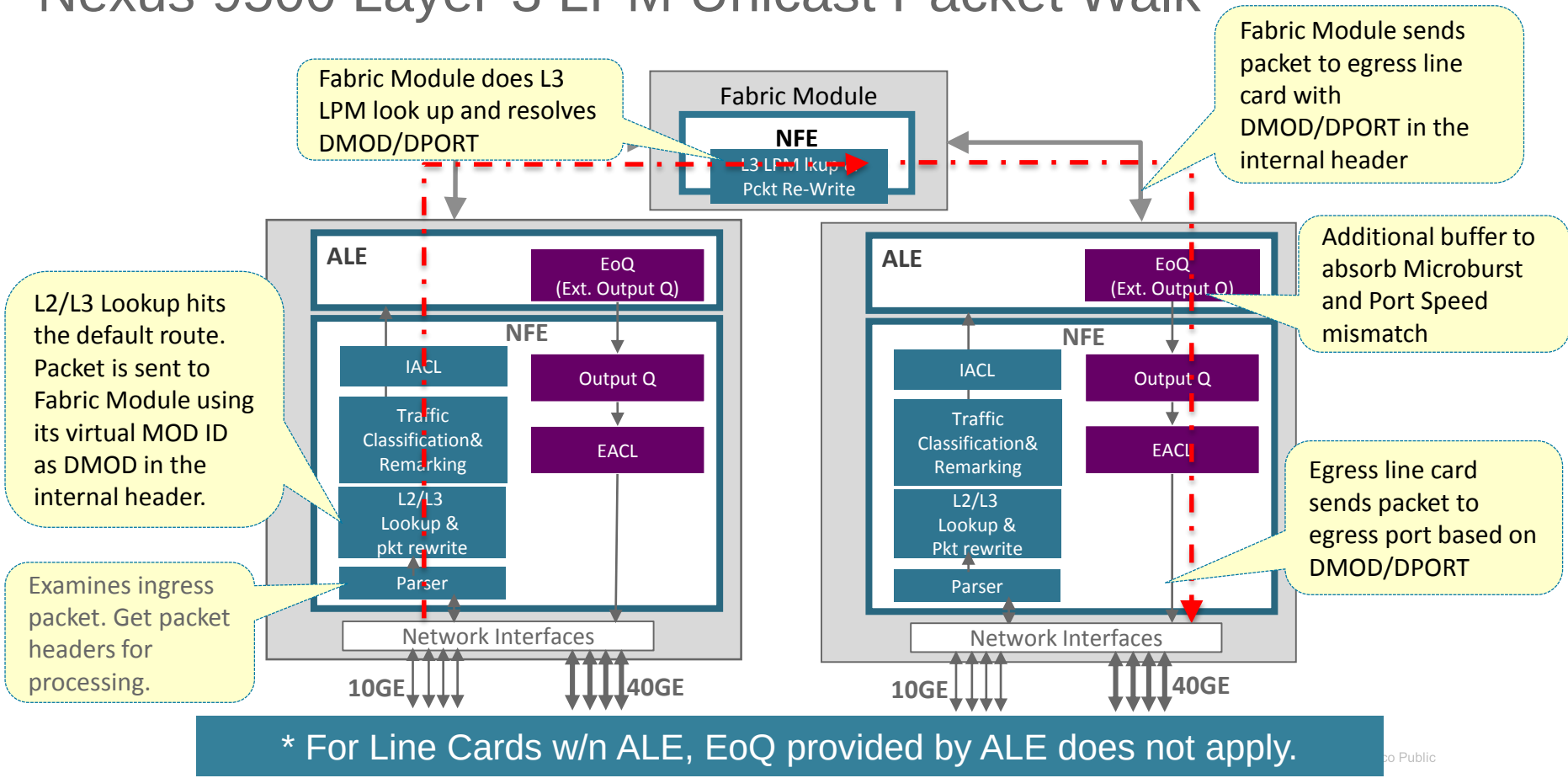


Nexus 9500 Layer-3 Host Unicast Packet Walk

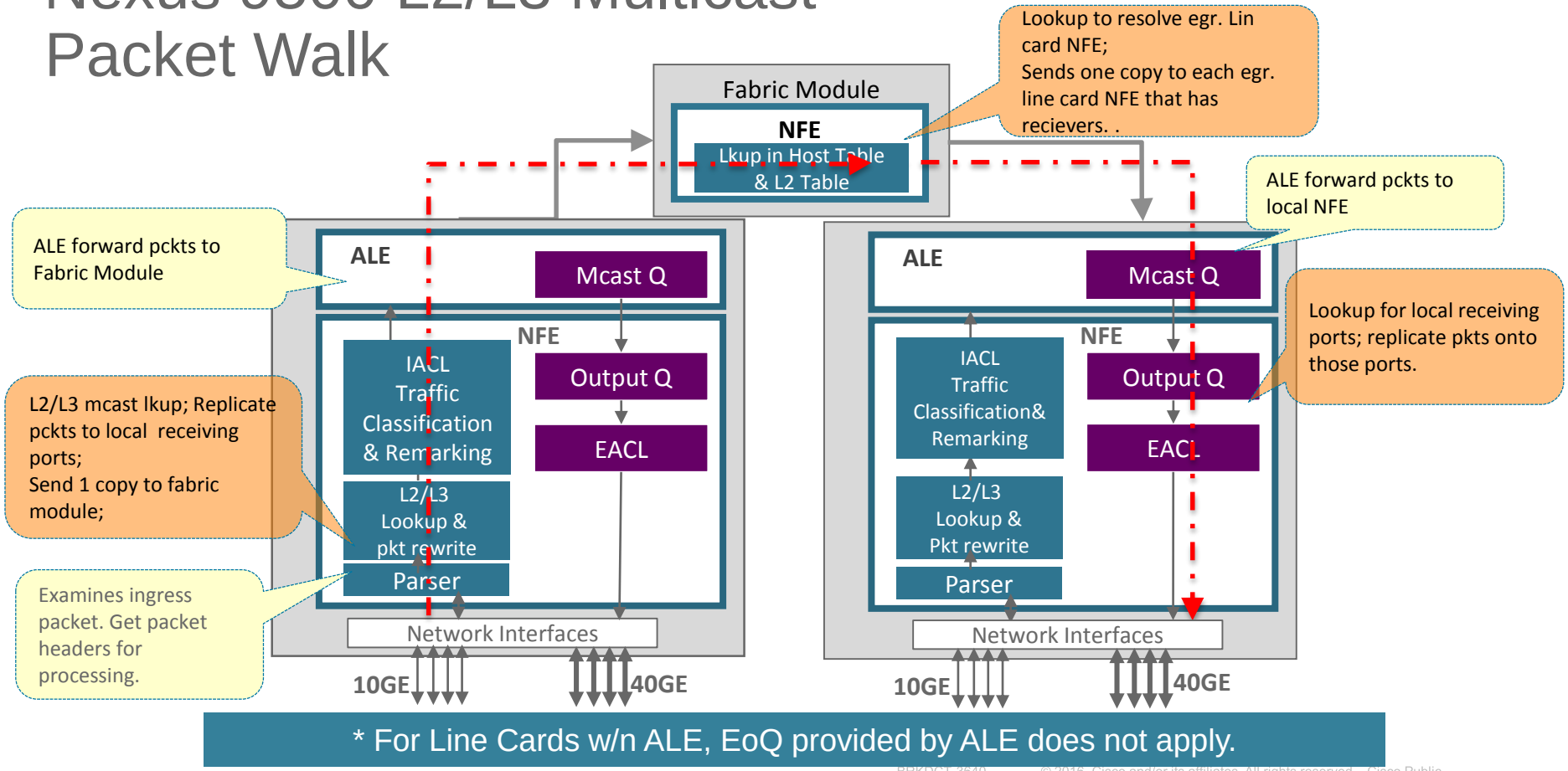


For Line Cards w/n ALE, EoQ provided by ALE does not apply.

Nexus 9500 Layer-3 LPM Unicast Packet Walk



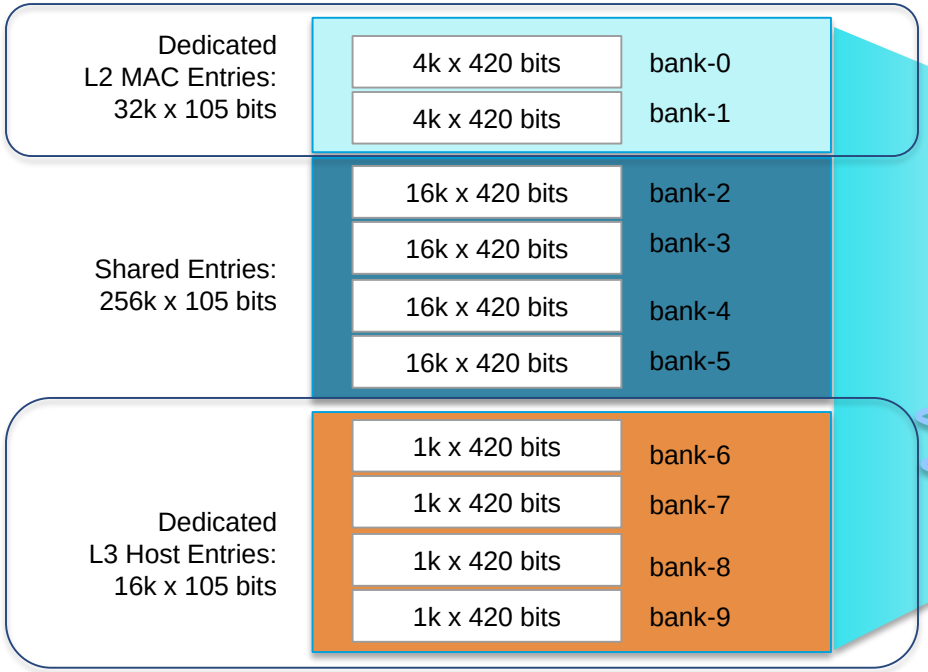
Nexus 9500 L2/L3 Multicast Packet Walk



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

Nexus 9500 Hierarchical Forwarding
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



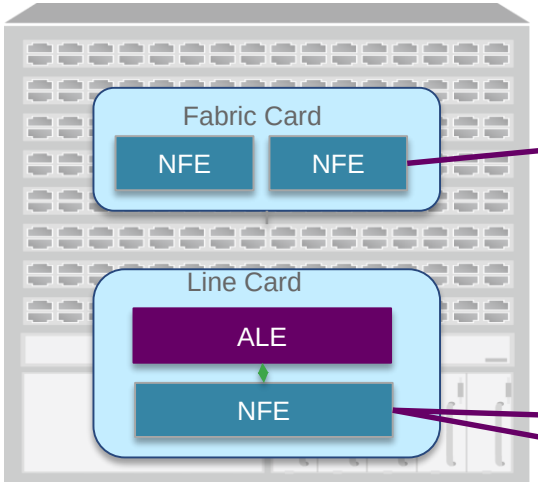
SUPPORTED COMBINATIONS

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.

Nexus 9500 Forwarding Programming Mode

Hierarchical Routing Mode (Default)



Feature	Scale	NFE Mode
IPv4/v6 LPM Routes	128K	4

Feature	Scale	NFE Mode
IPv4/v6 Host Routes*	120K*	3
MAC addresses	96K	

* Shares the same table with multicast routes

Nexus 9500 Forwarding Programming Mode

	MAC Table		IPv4/IPv6 Host Table		IPv4/IPv6 LPM Route Table		Multicast Route Table	
	Location	NFE Mode	Location	NFE Mode	Location	NFE Mode	Location	NFE Mode
Hierarchical routing mode (default)	LC	3	LC	3	FM	4	LC+FM	3
Hierarchical 64-bit ALPM mode	LC	3	LC	3	FM	4	LC+FM	3
Hierarchical Max-host routing mode	LC	2	IPv4 on FM	3	IPv4 on FM	3	LC+FM	
			IPv6 on LC	2	IPv6 on LC	2		
Non-hierarchical routing mode	LC	3	LC	3	LC	3	LC	3
Non-hierarchical routing Max-L3 mode	LC	4	LC	4	LC	4	LC	4

Forwarding Programming Mode	Configuration Command
Default Hierarchical routing mode	Default
Hierarchical 64-bit ALPM mode	9508(config)# system routing hierarchical max-mode l3 64b-alpm
Hierarchical Max-host routing mode	9508(config)# system routing max-mode host
Non-hierarchical routing mode	9508(config)# system routing non-hierarchical
Non-hierarchical routing Max-L3 mode	9508(config)# system routing non-hierarchical max-mode l3

CLI to Show Forwarding Programming Mode

```
9508# sh system routing mode
Configured System Routing Mode: Non-Hierarchical (Default)
Applied System Routing Mode: Hierarchical (Default)
Configured SVI post-routed unknown-unicast hardware flood mode: enabled
US-DUR-LC01-9508#
```

```
9508# show forwarding route summary module 1

Module Type                    : Line-Card
Module Mode                    : Mode-3
Module Route Download-type     : Host only
(IPv4+IPv6) (1)

IPv4 routes for table default/base

'**' denotes routes NOT programmed in hardware
due to hierarchical routing

Cumulative route updates: 1005038
Cumulative route inserts: 1005005
Cumulative route deletes: 143
Total number of routes: 24
Total number of paths : 25

Number of routes per mask-length:
  /32 : 24
```

```
9508# show forwarding route summary module 26

Module Type                    : Fabric-Module
Module Mode                    : ALPM (Mode-4)
Module Route Download-type     : LPM only
(IPv4+IPv6) (2)

IPv4 routes for table default/base

'**' denotes routes NOT programmed in hardware due
to hierarchical routing

Cumulative route updates: 1005043
Cumulative route inserts: 1004930
Cumulative route deletes: 54
Total number of routes: 8
Total number of paths : 8

Number of routes per mask-length:
  /8  : 1          /30 : 5

US-DUR-LC01-9508#
```

CLI to Check Forwarding Table Sizes

	Software	Hardware (BCM-shell)
MAC Table	show mac address-table count	l2 show
IP Host Table	show ip route sum sh forwarding route summary mod <#>	l3 l3table show [on LC]
IP LPM Table	show ip route sum show forwarding route sum mod <#>	l3 defip show [on FM]
egress next-hop table		l3 egress show [on both LC and FM]

Commands to check hardware table size:
Leverage NX-OS “| count” to get the account of the hardware entries. Example”

```
TME-1-9508-1# bcm-shell mod 1 "l3 l3table show" | count
```

Nexus 9500 – Moving Forward

9500 Series

9504 9508 9516

Existing 4-, 8-, 16- slot chassis

No mid-plane to update

Power and cooling within existing shipping system profile

Existing shipping Power Supply, Supervisor and System Controllers

X9700-EX (NX-OS and ACI)

Cisco ASIC

16nm Technology

32p 100G QSFP Line card

- 10/25/40/50/100G
- Analytics Readiness

+

Fabric Module

- Back-ward compatible w/ existing Nexus 9300 ACI Leafs (40G uplinks) in ACI mode

Migrate From NX-OS to ACI Spine with Just a Software Upgrade

X9400-S (NX-OS)

Merchant ASIC

28nm Technology

32p 100G QSFP Line card

- 10/25/40/50/100G

+

Fabric Module

- Back-ward compatible w/ existing Broadcom T2 based line cards

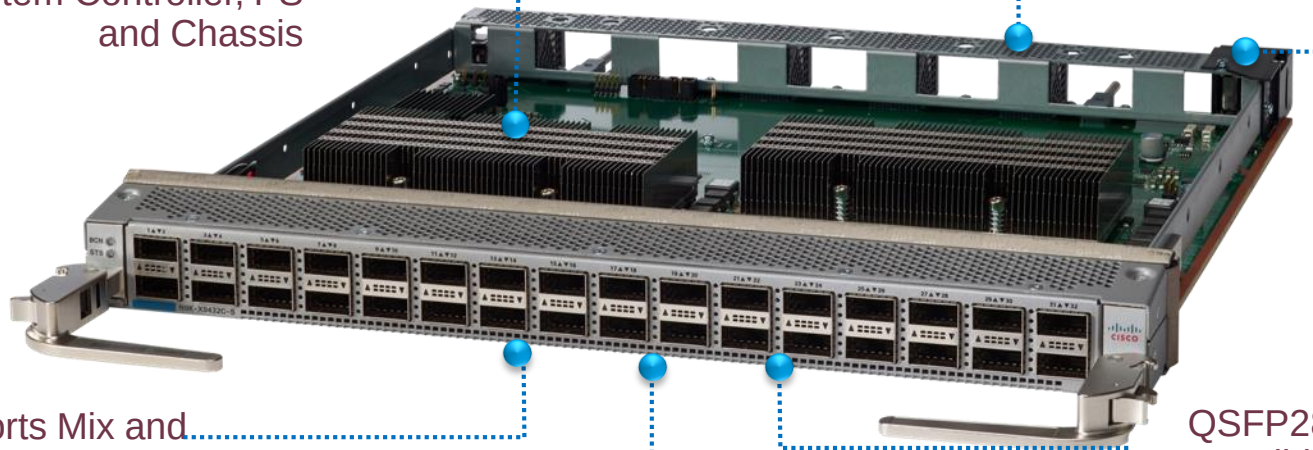
Upgrade to 100G Infrastructure While Reusing Existing Chassis

40/100G - Merchant
N9K-X9432C-S

Flexible Speed 10,25,40,50,100G

Investment Protection with Supervisors, System Controller, PS and Chassis

Supported in NX-OS mode



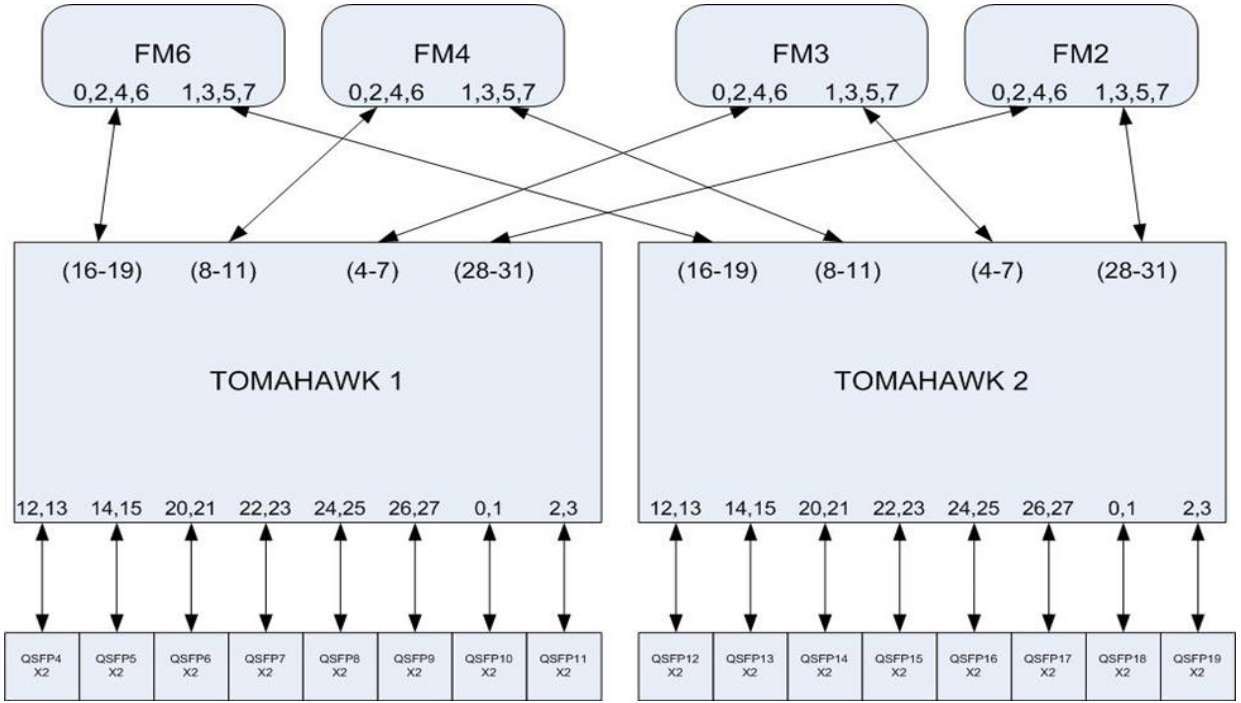
Supports Mix and Match Current Linecards*

QSFP28 Connector, Pin compatible with 40G QSFP+

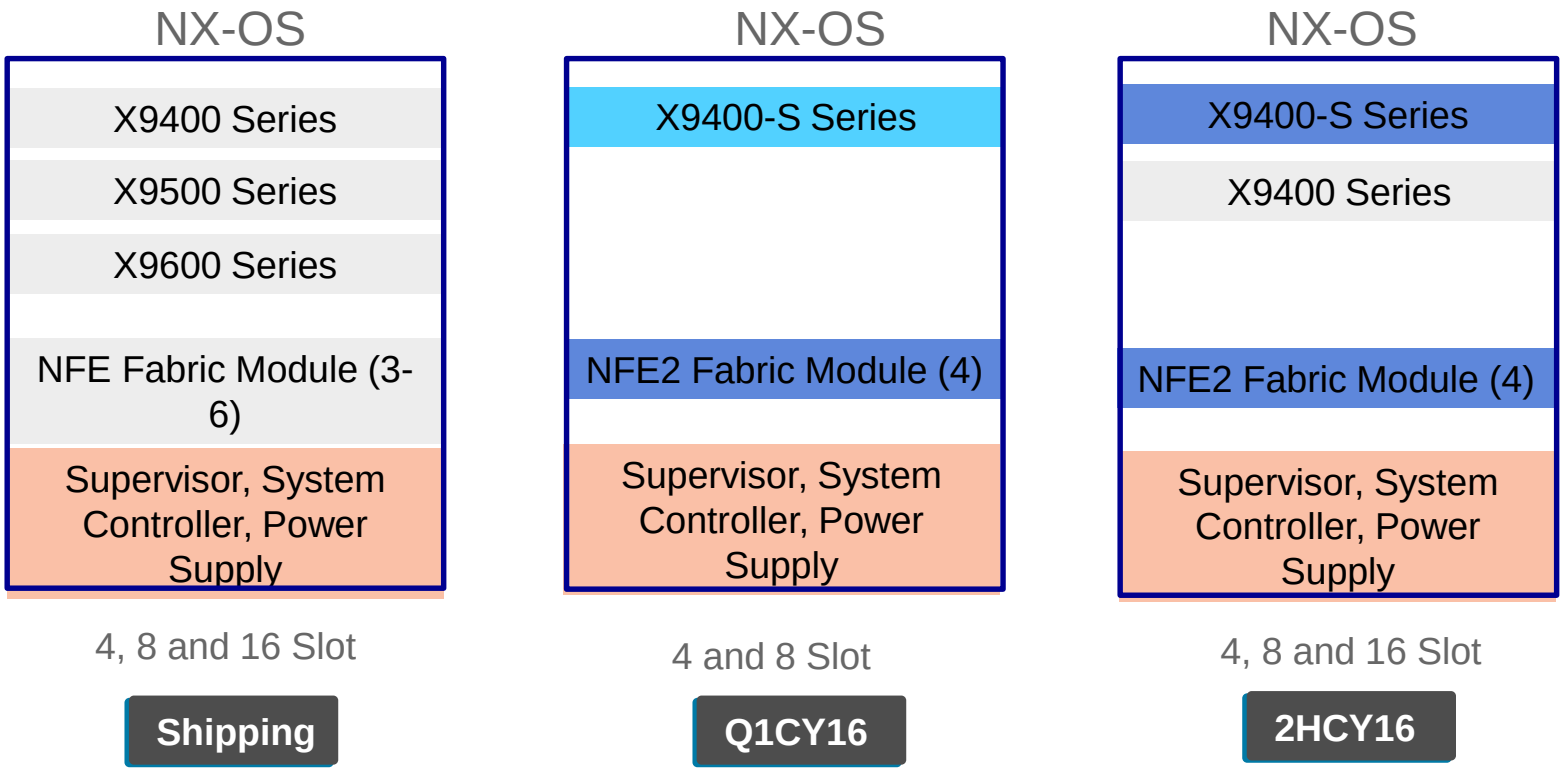
4, 8 and 16* Chassis

* future

40/100G - Merchant
N9K-X9432C-S



Nexus 9500 – LC vs Fabric Compatibility
Merchant ASIC

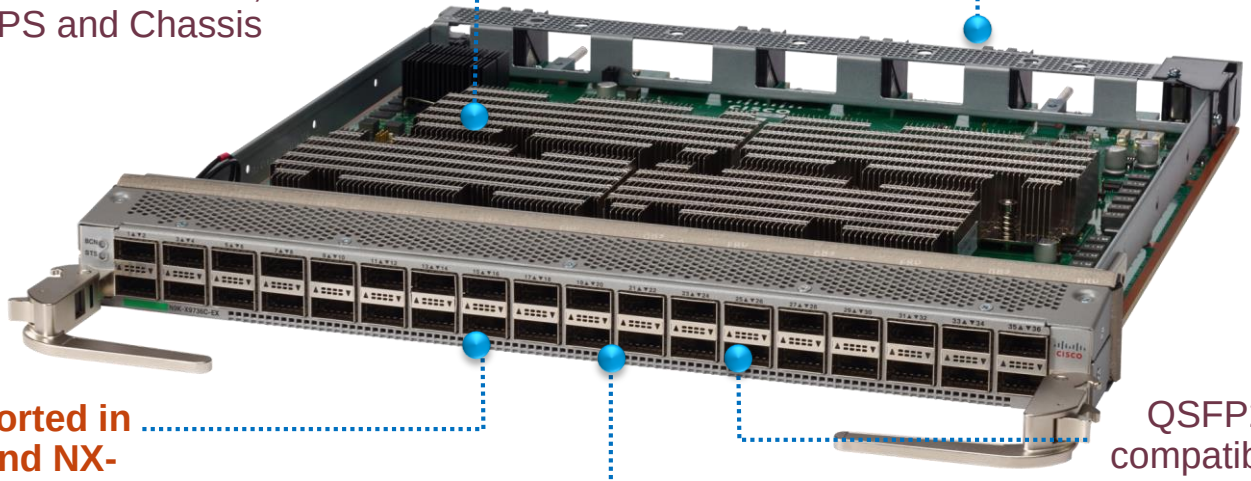


Note: 9636PQ works in 4 and 8 slot only

40/100G - LSE
N9K-X9736C-EX (Q2CY16)

Flexible Speed 10,25,40,50,100G

Investment Protection with Supervisors, System Controller, PS and Chassis



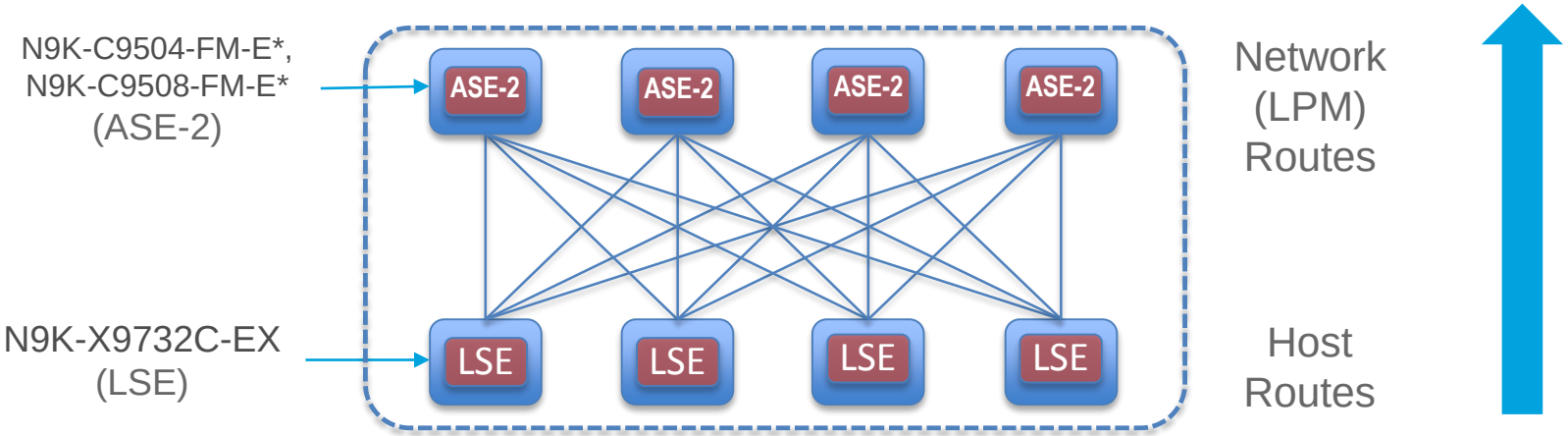
Supported in ACI and NX-OS mode

QSFP28 Connector, Pin compatible with 40G QSFP+

4, 8 and 16* Chassis

* future

Modular Nexus 9500
Generation 2 Line Cards and Fabric Modules



1. IPv4: 1M LPM+ host
2. IPv4: 750K LPM + host **AND** IPv6 /64: 256K

Summarisation and Balance of Host and Network Routes Shift

ASE2-Based New Fabric Module for Nexus 9500

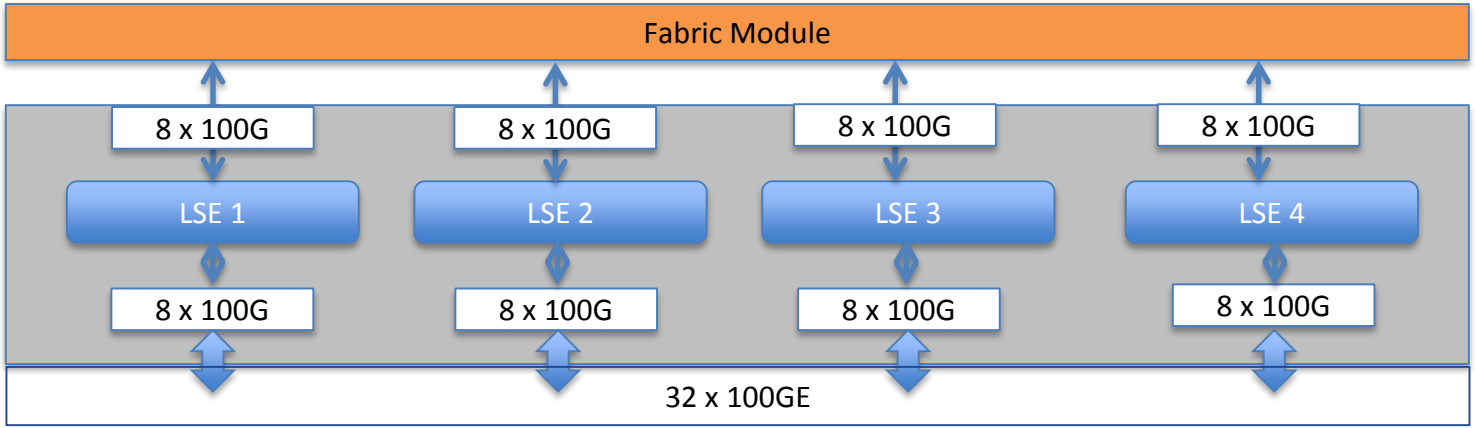
- The new fabric module is built with ASE2 ASICs
- Continue to use an internal CLOS architecture with fabric modules at the spine and line cards at the leaf
- Each Nexus 9500 switch needs up to 4 ASE2-based fabric modules
- Each ASE2 ASIC on a fabric module provides 32 x 100GE internal ports to interconnect line cards.

  (32 out of 36 100GE ports on an ASE2 are used to build the CLOS architecture with evenly distributed bandwidth between each line card and fabric module.)
- The number of ASE2 on a fabric module depends on the chassis types it supports:

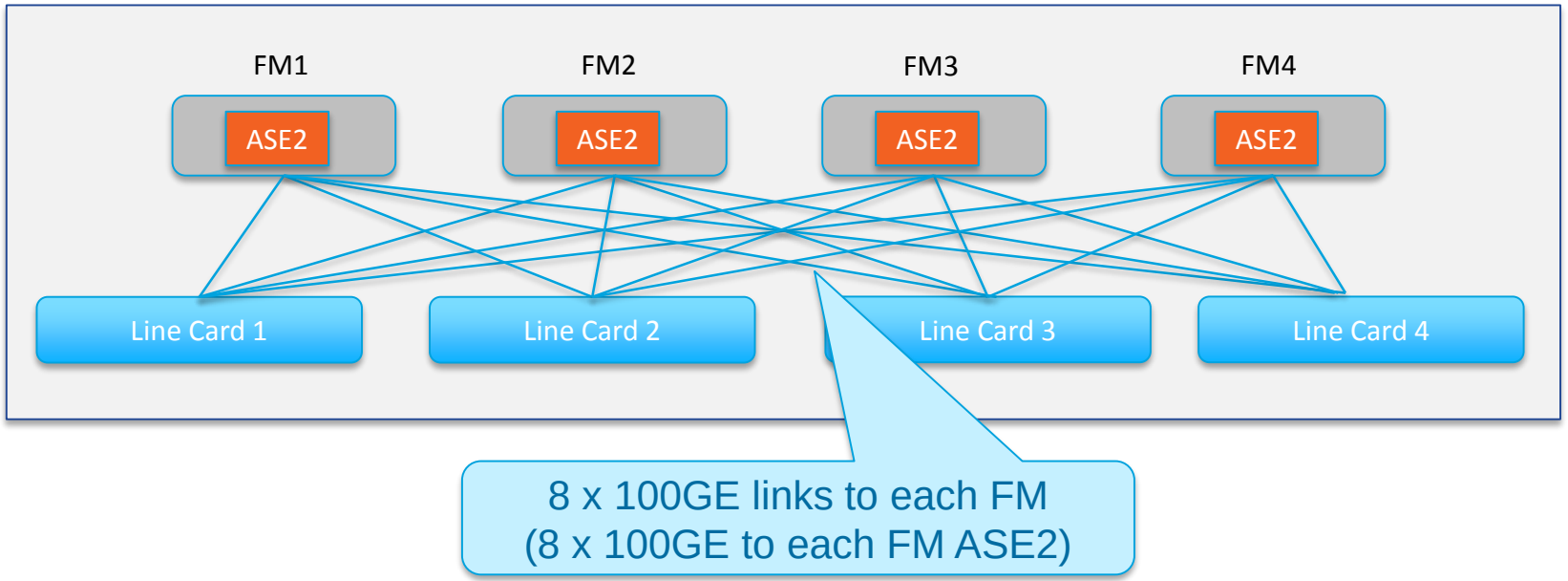
Chassis Type	Nexus 9504	Nexus 9508	Nexus 9516
# of ASE2 on FM	1	2	4

LSE-Based New Line Card for Nexus 9500
9700EX

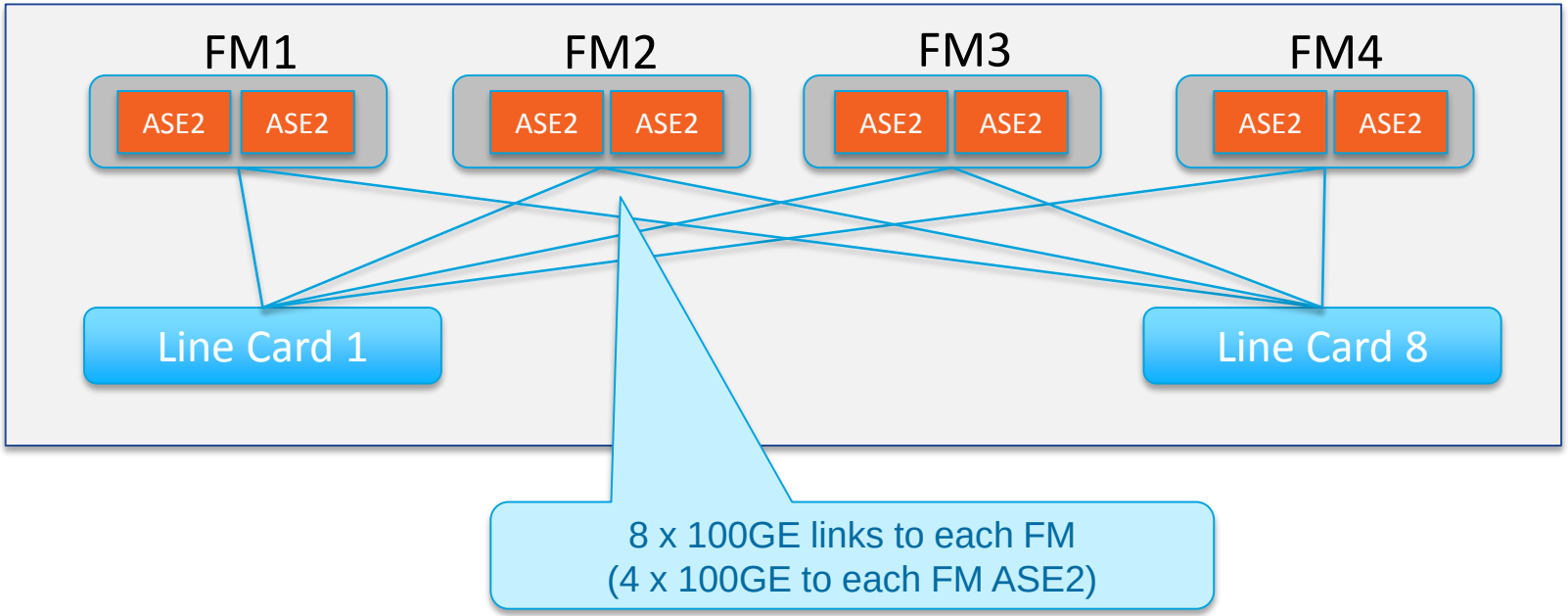
- Built with 4 LSE ASICs,
- 32 x 100GE front panel ports + 32x 100GE internal links to fabric modules
- Each LSE provides 8 x 100GE front panel ports
- Each LSE has 8 x 100GE internal links to fabric modules, evenly distributed among the 4 fabric modules



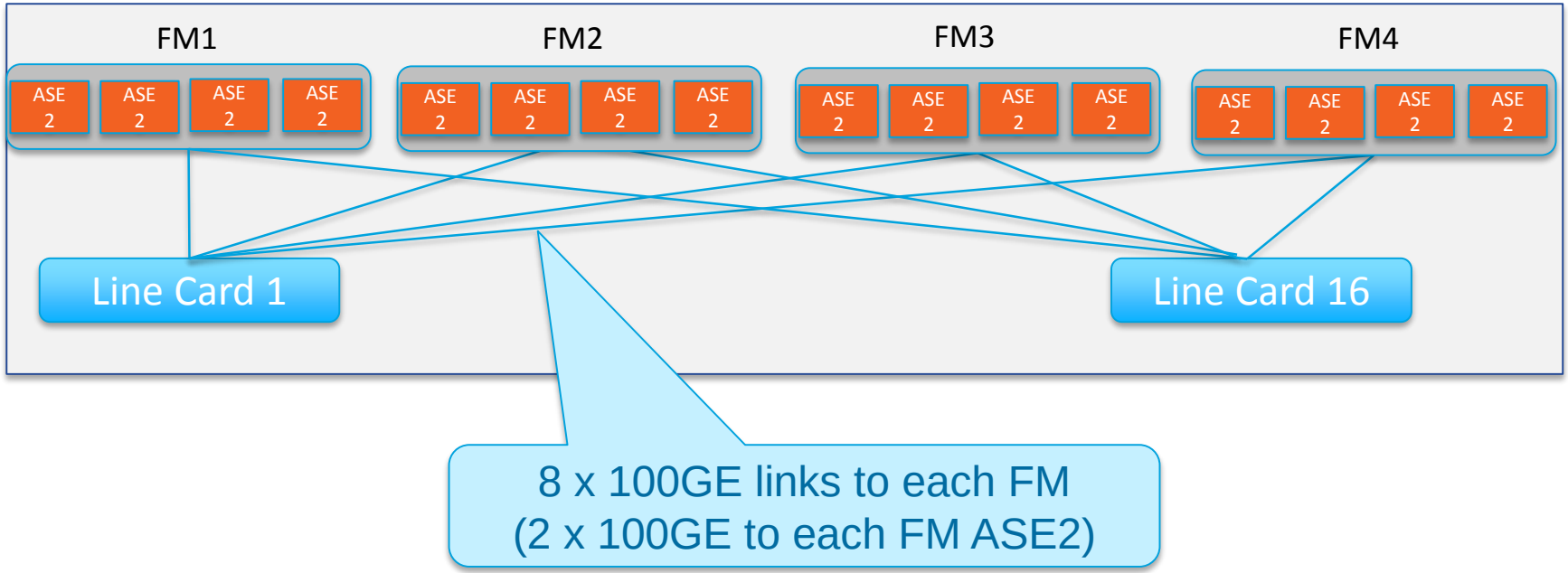
Internal Fabric Architecture of Nexus 9504



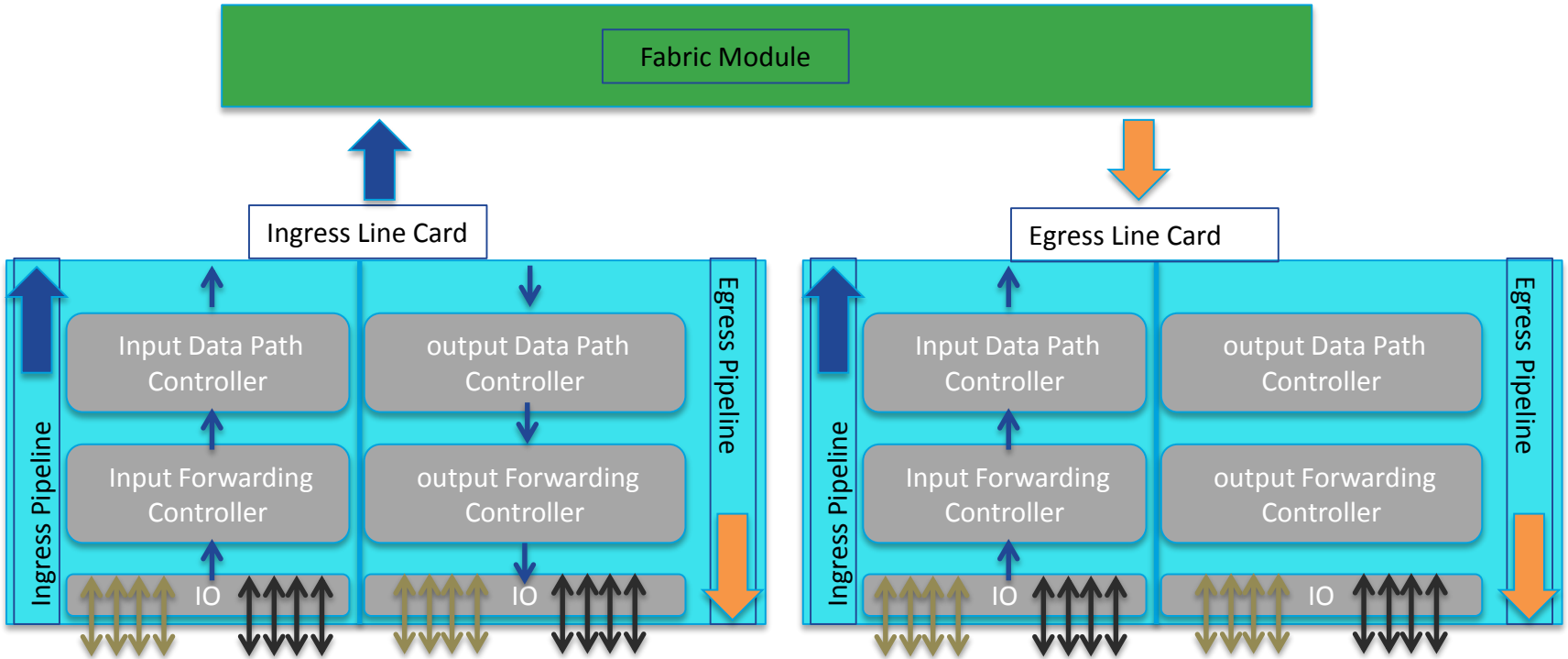
Internal Fabric Architecture of Nexus 9508



Internal Fabric Architecture of Nexus 9516



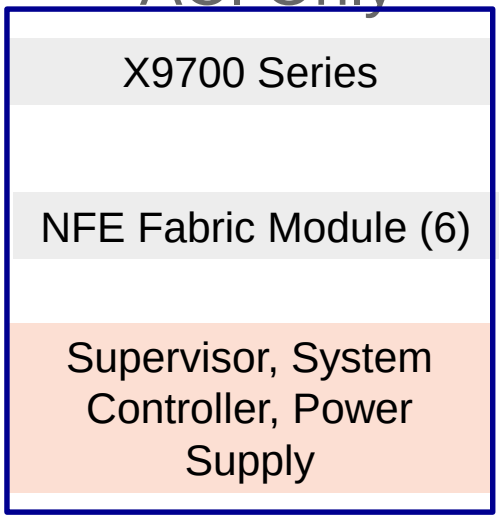
Packet Processing Pipeline



Nexus 9500 – LC and Fabric Compatibility

Cisco ASIC

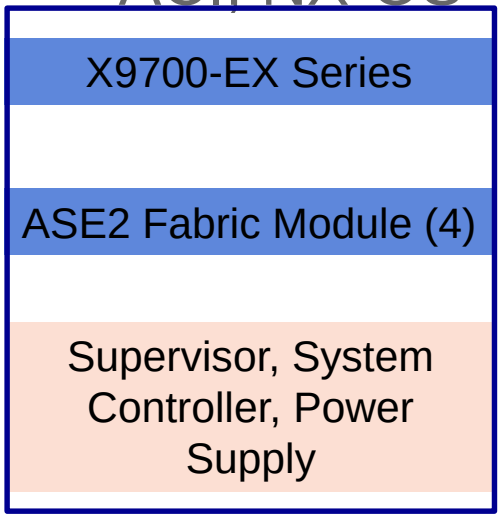
ACI Only



4, 8 and 16 Slot

Shipping

ACI, NX-OS



4, 8 and 16* Slot

Q2CY16

Nexus 9500 LC/FM Compatibility Matrix Supported Combination in same Chassis

NX-OS

Fabric Modules/ Line Cards	Gen1 (T2)	Gen2-E (ASE2)	Gen2-S (TH)
X9400	yes	no	Yes (post FCS)
X9500	yes	no	Yes (post FCS)
X9600	yes	no	Yes (post FCS)
X9400S	no	no	Yes
X9700E/EX	no	yes	no

ACI

Fabric Modules/ Line Cards	Gen1 (T2)	Gen2-E (ASE2)
X9700	yes	no
X9700E/EX	no	Yes

Agenda

- Existing and New Nexus 9000 & 3000
- What's New
  - Moore's Law and 25G SerDes
  - The new building blocks (ASE-2, ASE-3, LSE)
  - Examples of the Next Gen Capabilities
- Nexus 9000 Switch Architecture
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)
- 100G Optics

Optical Innovation --- Removing 40 Gb Barriers

Problem

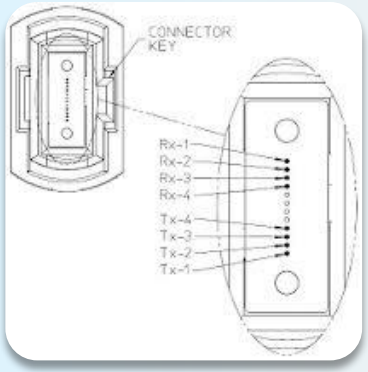
- 40 Gb optics are a significant portion of capital expenditures (CAPEX)
- 40 Gb optics require new cabling

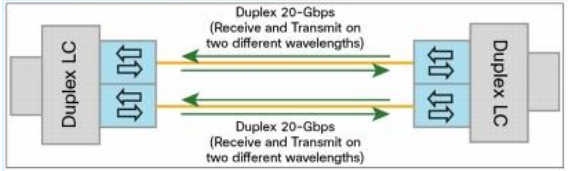
Solution

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

Cisco® 40 Gb SR-BiDi QSFP

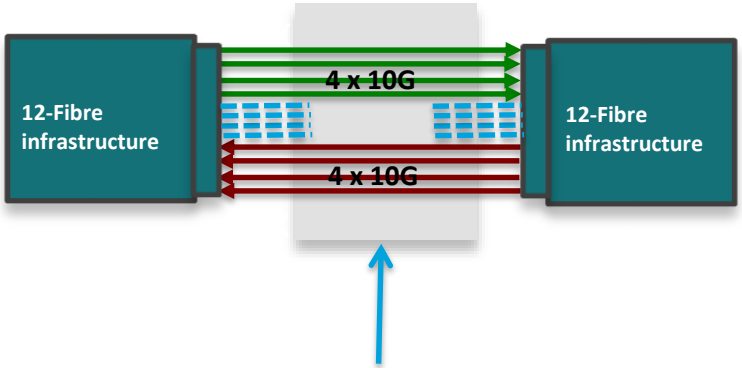
- QSFP, MSA-compliant
- Dual LC connector
- Support for 100 m on OM3 and upto 150m on OM4
- TX/RX on two wavelengths at 20 Gb each





Available end of CY13 and supported across all Cisco QSFP ports

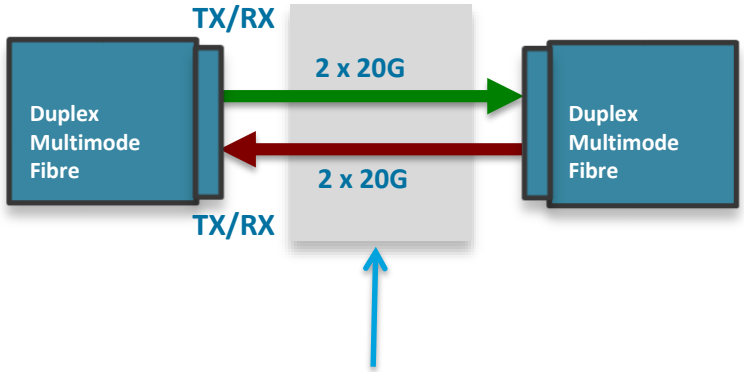
QSFP-BIDI vs. QSFP-40G-SR4



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

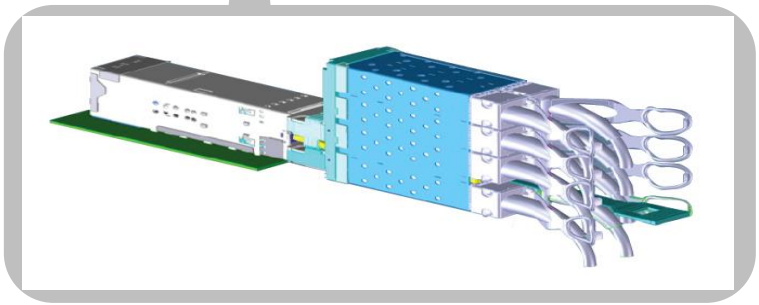
Use of duplex multimode fibre lowers cost of upgrading from 10G to 40G by leveraging existing 10G multimode infrastructure

Cisco 100G QSFP28 Optics Portfolio

MPO: 8 strands
LC: 2 strands
SMF: Single Mode Fibre
MMF: Multi Mode Fibre

Optics Type	Description	Connector	Availability
QSFP-100G-SR-BD	40/100G, 100m	MMF LC	1HCY16
QSFP-100G-SR4-S	100GBASE-SR4, 100m	MMF MPO	Q4CY15
QSFP-100G-LR4-S	100GBASE-LR4, 10km	SMF LC	Q4CY15
QSFP-100G-CWDM4-S	100GE CWDM4, 2km	SMF LC	Q4CY15
QSFP-100G-PSM4-S	100GBASE-PSM4, 2km	SMF MPO	Q4CY15
QSFP-100G-CU	100GBASE QSFP to QSFP copper direct-attach cables	Twinax	Q4CY15
QSFP-4SFP25G-CU	100GBASE QSFP to 4x25G SFP+ copper break-out cables	Twinax	Q4CY15
QSFP-100G-AOC	100GBASE QSFP to QSFP active optical cables	AOC (Active Optic Cable)	Q4CY15

Cisco QSFP-to-SFP Converters



2HCY15

2 QSFP to 8 SFP+

2x40G -> 8x10G/ 2x100G -> 8x 25G

2 QSFP to 4 QSFP

2x100G -> 4x 50G

Fit with 1 RU TOR switches only

Flexible conversion of ports on an as needed basis

32p 40G -> 96p 10G & 8p 40G

32p 100G -> 64p 25G & 16p 100G

32p 100G -> 48p 50G & 8p 100G

No break-out cable

Support for standard 10G/ 25G SFP and 40/50/100G QSFP

Optics Pluggable Multispeed Interfaces

Pluggable Options

- 1G SFP
- 10G SFP+, Twinax, AOC
- 25G SFP+, Twinax, AOC

Pluggable Options

- 1G SFP (via QSA)
- 10G SFP+, Twinax, AOC (via QSA)
- 25G SFP+, Twinax, AOC (via SLIC)
- 40G QSFP, Twinax, AOC
- 50G Twinax, AOC (via SLIC)
- 100G QSFP, Twinax, AOC

Optics 2 Fibre vs. 8 Fibre

2 fibre
LC connector

2 fibre
LC connector

8 fibre
MTP connector
Breakout possible

Examples

MMF: Bidi
SMF: LR4, CWDM4

MMF: SR4
SMF: pSMF4

Note: Trade-off between fibre cost and optics cost

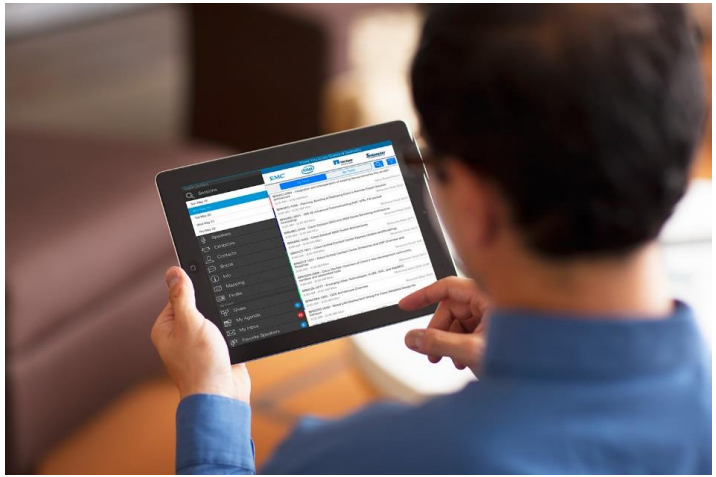
Q & A

Complete Your Online Session Evaluation

Give us your feedback and receive a **Cisco 2016 T-Shirt** by completing the 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/ciscolivemelbourne2016/
- Visit any Cisco Live Internet Station located throughout the venue

T-Shirts can be collected Friday 11 March at Registration



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