CISCO

Cisco live!

7–10 March 2017 • Melbourne, Australia

Your Time Is Now

Nexus 9000 Architecture

Mike Herbert, Principal Engineer, INSBU

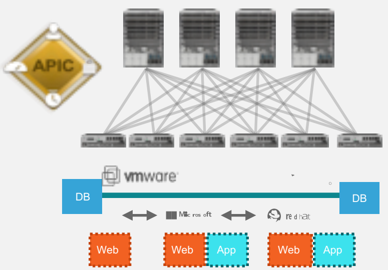
BRKDCT-3640

Agenda

- What’s New
  - 2nd Generation Nexus 9000
  - Moore’s Law and 25G SerDes
  - The new building blocks (ASE-2, ASE-3, ASE-4, LSE, LSE-2)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry, Encryption
- Design Impacts of 25G, 50G and 100G
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)

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

Application Centric Infrastructure

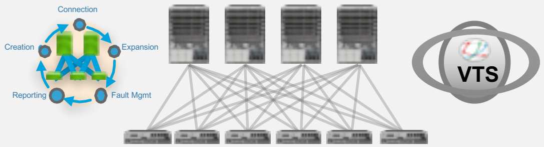


Turnkey integrated solution with security, centralised management, compliance and scale

Automated application centric-policy model with embedded security

Broad and deep ecosystem

Programmable Fabric

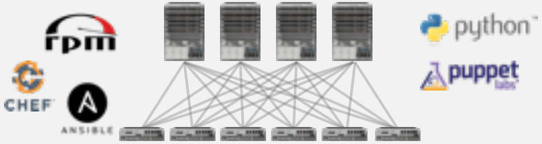


VxLAN-BGP EVPN standard-based

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 & 9600 (line cards), 9200, 3100, 3200

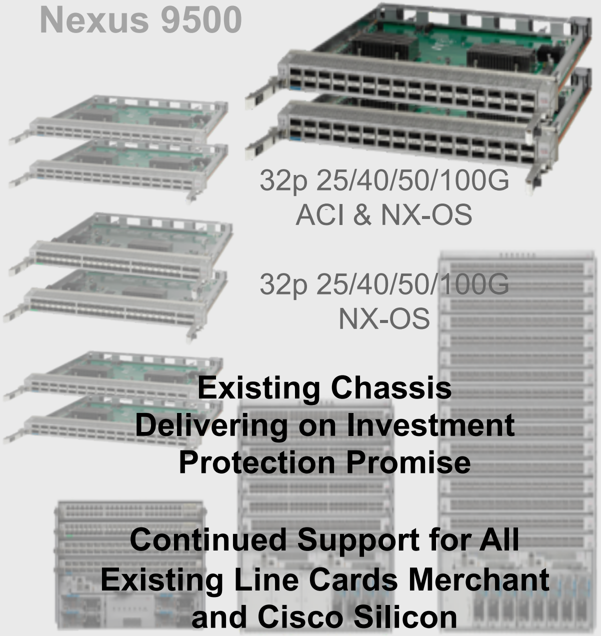
Nexus 9700EX + 9300EX

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

Nexus 9300

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

Nexus 9500



32p 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 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

Nexus 9K/3K Portfolio

Data Centre Deployment Options

Cloud Scale Switch on Chip

- Advanced Telemetry (Flow Cache, SSX, Triggered Events)
- Smart Buffering
- Rich Forwarding Feature Set
- Optimised Scale, Cost, Power

Cisco: Cloud Scale ASIC's

- High Speed Fabrics (ACI, VXLAN, Segment Routing, GRID, HPC)
- General Data Centre Design

Modular X9700EX
Fixed 9200 & 9300EX

BCOM Switch on Chip

- BCOM Switch On Chip solution
- Published SDK

Broadcom: Trident II+, Tomahawk

- Fabric Designs (customers specifically looking for BCOM based SOC)

Modular X9400S
N3x00

BCOM Cross Bar ASIC

- Off Chip Buffer and Forwarding Tables

Broadcom: Jericho

- Financial Multicast (UDP)
- Collapsed Core/ DC Edge (Large Routing Tables)

Modular X9600R Shipping
Fixed Q3CY17

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

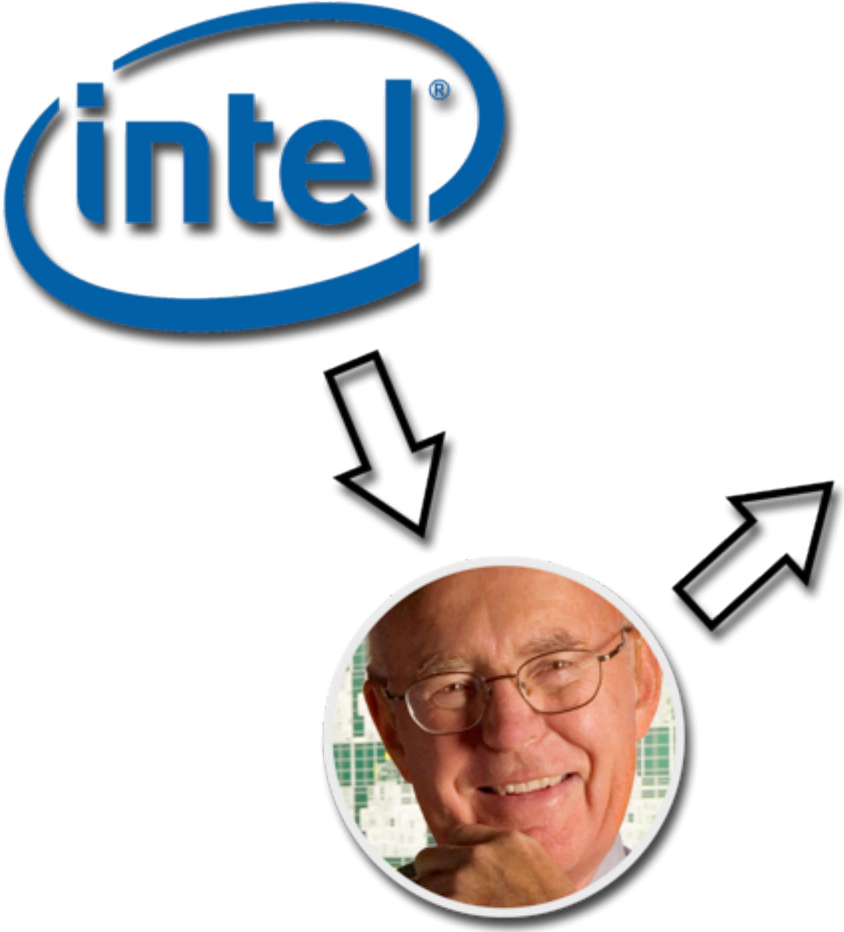
VXLAN bridging, **25/100G**

Tomahawk

Single NX-OS Image for Nexus 3000 & Nexus 9000

Agenda

- What's New
  - 2nd Generation Nexus 9000
  - Moore's Law and 25G SerDes
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- Design Impacts of 25G, 50G and 100G
- Nexus 9000 Switch Platforms
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)



Gordon Moore

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

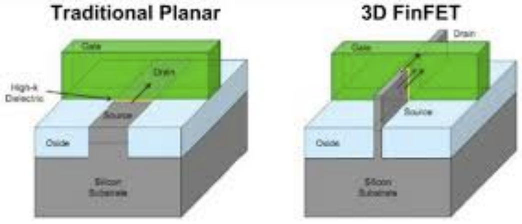
“Moore’s Law” - 1975

Moore’s Law

It’s all about the Economics

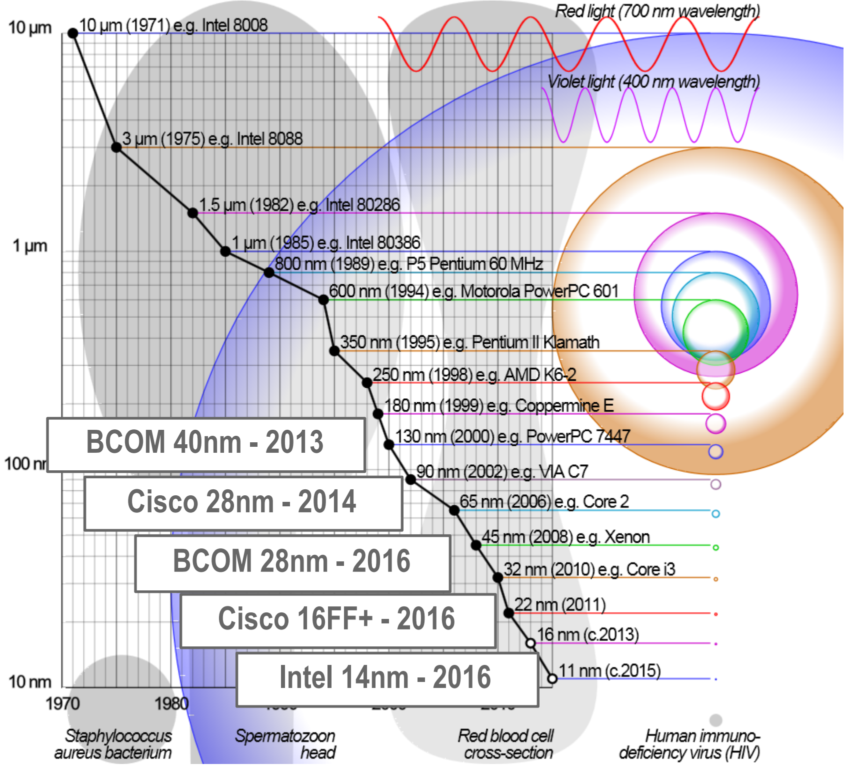
- Increased function, efficiency
- Reduced costs, power
- ~ 1.6 x increase in gates between process nodes

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



BCOM 40nm - 2013

Cisco 28nm - 2014

BCOM 28nm - 2016

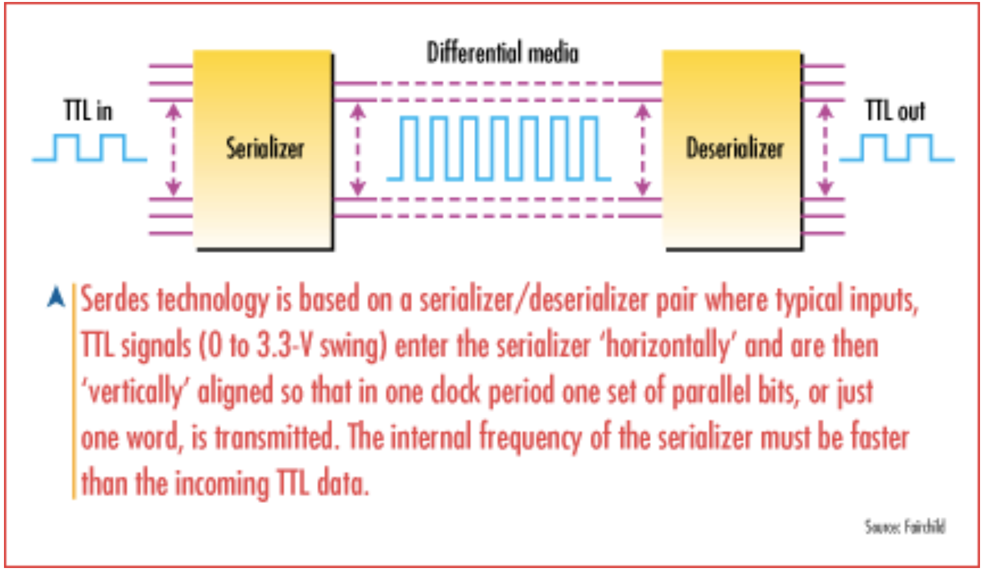
Cisco 16FF+ - 2016

Intel 14nm - 2016

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

SerDes: Serialiser + Deserialiser

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

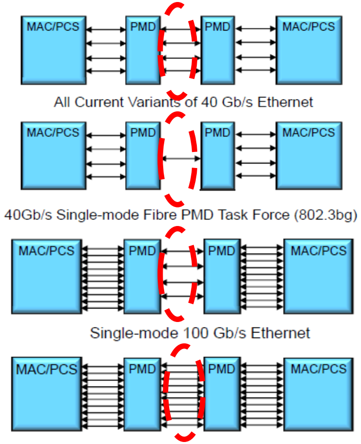


Serdes technology is based on a serializer/deserializer pair where typical inputs, TTL signals (0 to 3.3-V swing) enter the serializer 'horizontally' and are then 'vertically' aligned so that in one clock period one set of parallel bits, or just one word, is transmitted. The internal frequency of the serializer must be faster than the incoming TTL data.

Source: Fairchild

Multi Lane Distribution (MLD)

MLD (Multi Lane Distribution)

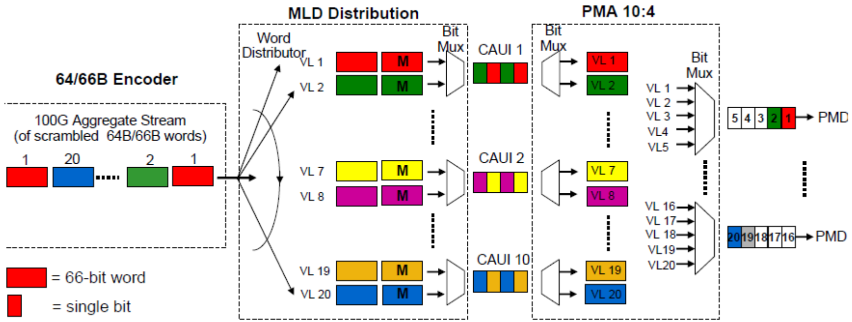


All Current Variants of 40 Gb/s Ethernet

40Gb/s Single-mode Fibre PMD Task Force (802.3bg)

Single-mode 100 Gb/s Ethernet

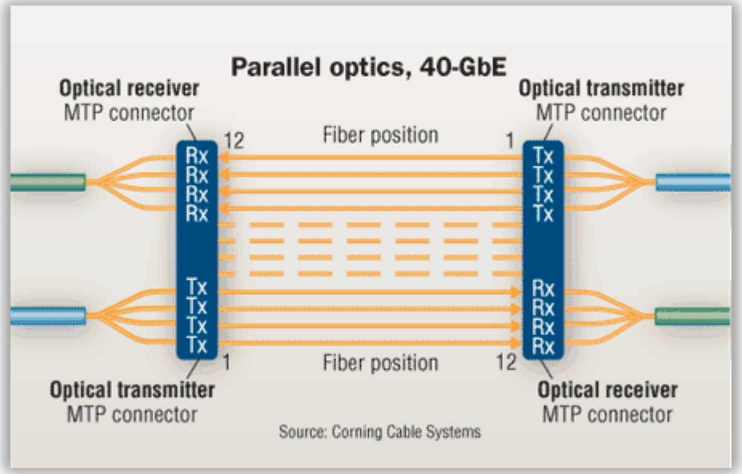
Multi Mode 100 Gb/s Ethernet



- 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

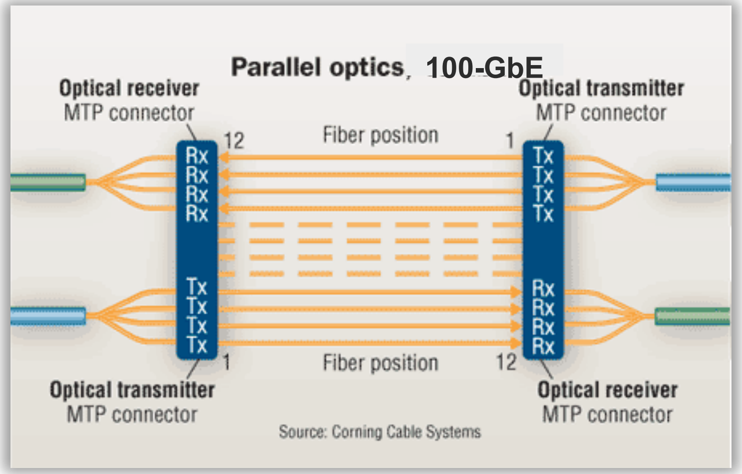
Parallel Lanes

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



Source: Corning Cable Systems

Backed by 10G SerDes



Source: Corning Cable Systems

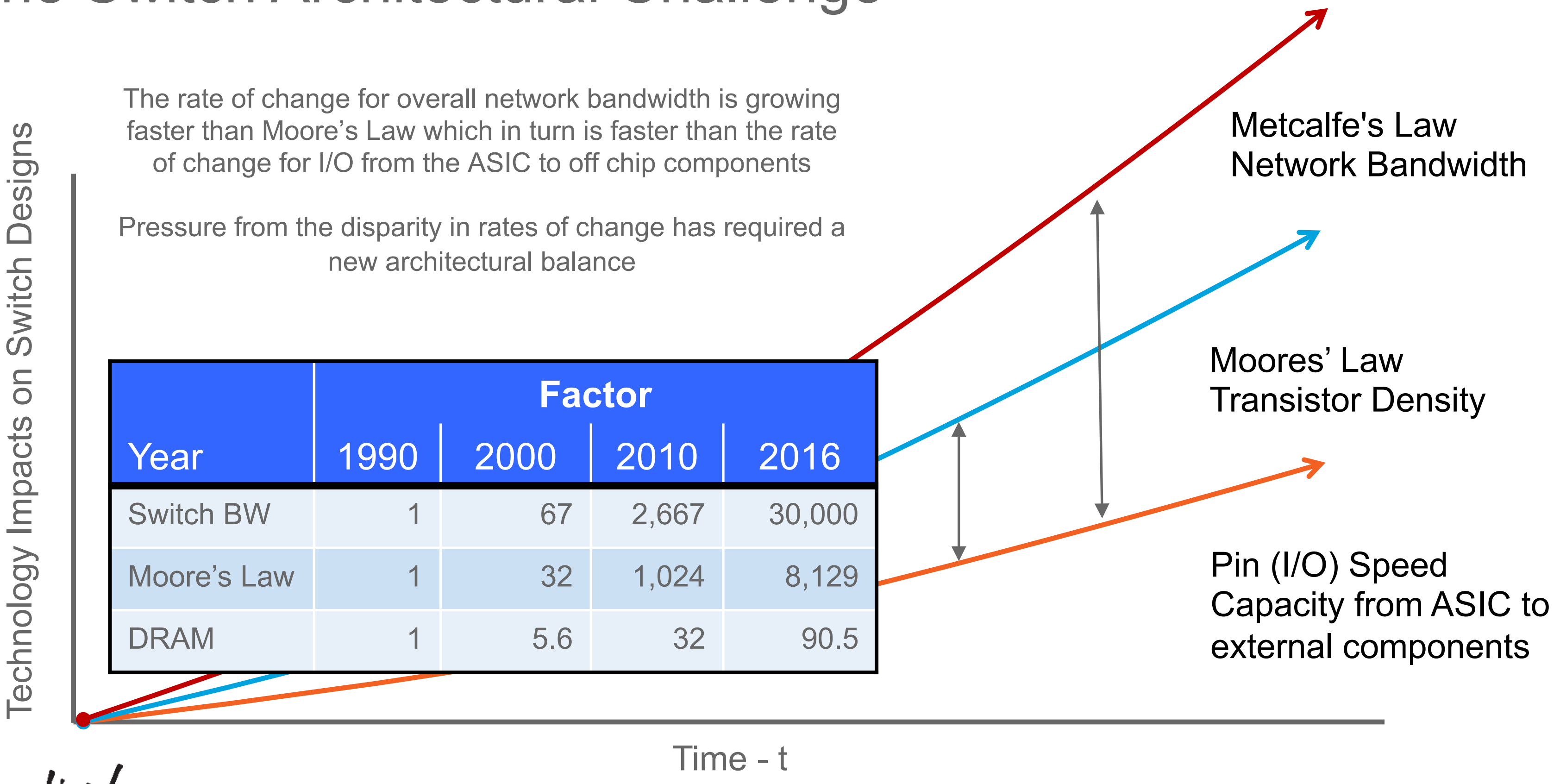
Backed by 25G SerDes

Metcalfe, Moore and ASIC Pin I/O Rates

The Switch Architectural Challenge

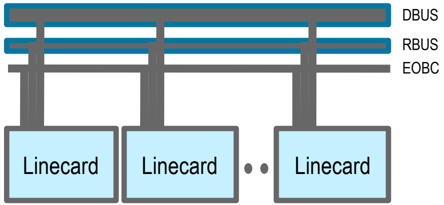
The rate of change for overall network bandwidth is growing faster than Moore's Law which in turn is faster than the rate of change for I/O from the ASIC to off chip components

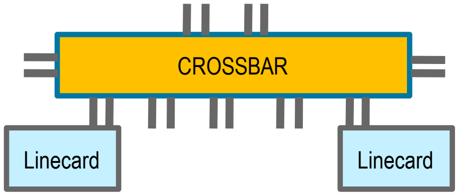
Pressure from the disparity in rates of change has required a new architectural balance

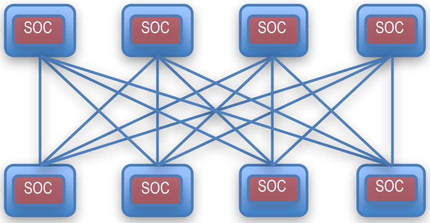


Year	Factor			
	1990	2000	2010	2016
Switch BW	1	67	2,667	30,000
Moore's Law	1	32	1,024	8,129
DRAM	1	5.6	32	90.5

Switching Architecture Changes
Shifting of Internal Architecture





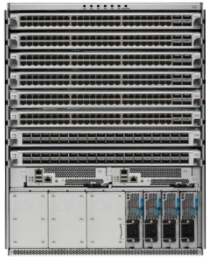


Design Shifts Resulting from Increasing Gate Density and Bandwidth

10/100M

100M/1G

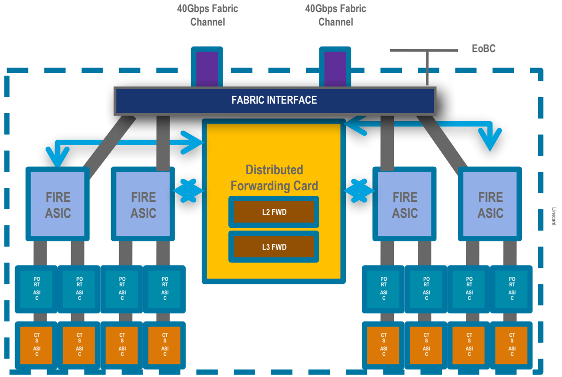
1G/10G



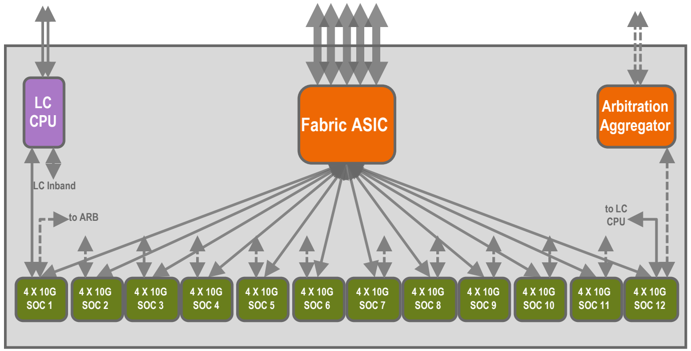
10G/100G

Switching Architecture Changes

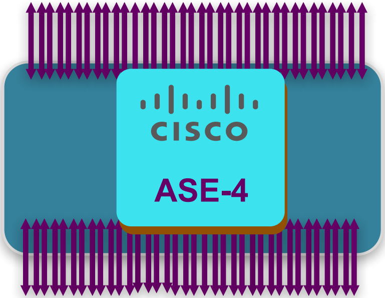
Consolidation of Functions onto fewer components



32 x 10G Ports

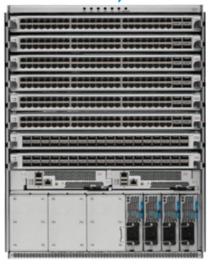


48 x 10G Ports



64 x 100G Ports

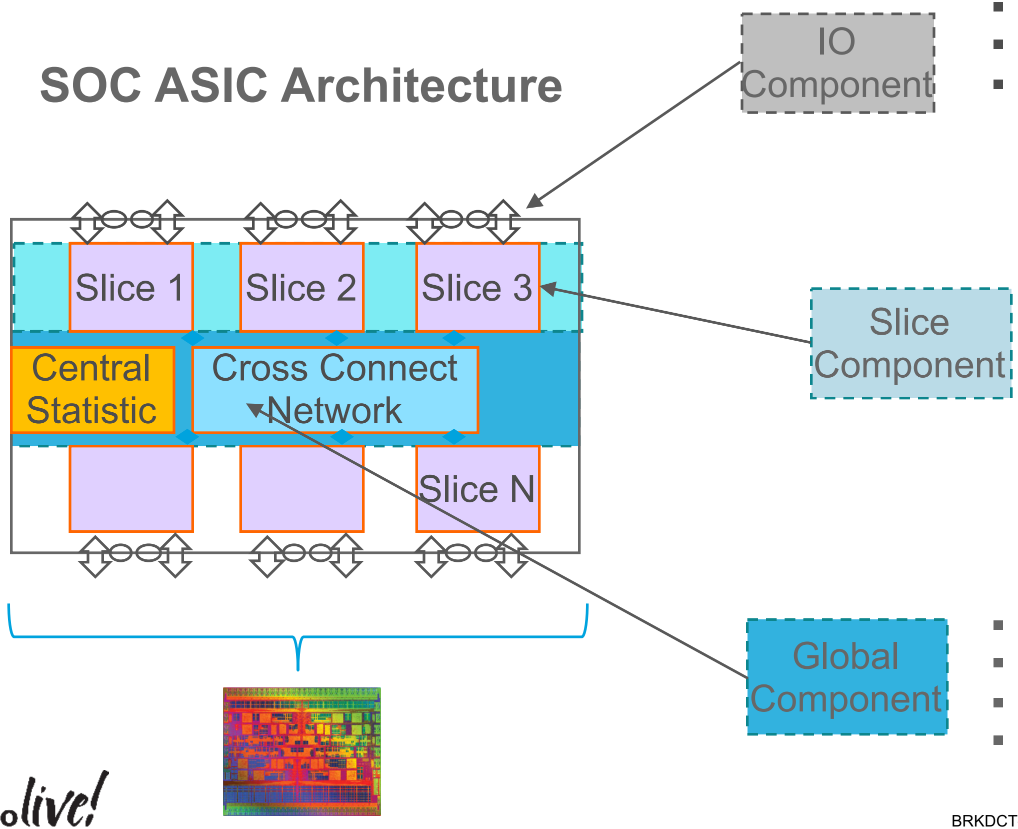
Design Shifts Resulting from Increasing Gate Density and Bandwidth



Switch On Chip (SOC)

It is a full multi-stage switch on an ASIC

SOC ASIC Architecture



- The IO components consists of high speed SerDes.
- They vary based on the total number of ports
- They determine the total bandwidth capacity of the ASIC

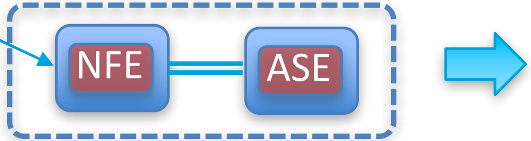
- Multi-mode MAC
- Packet parser
- Forwarding controller
- Input packet buffering for pause
- Output packet buffering
- Buffer accounting
- Output queuing and scheduling
- Output Rewrite

- Gen2 PCIe controller for register and eDMA access
- Broadcast network to connect all the slices together
- Counter modules to collect packet statistics
- PLL to generate core and MAC clocks

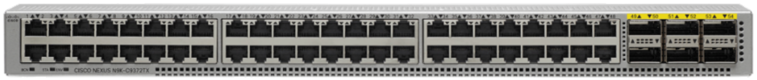
Fixed First Generation Nexus 9300

A Dual ASIC based Switch

Switch on Chip (SOC)



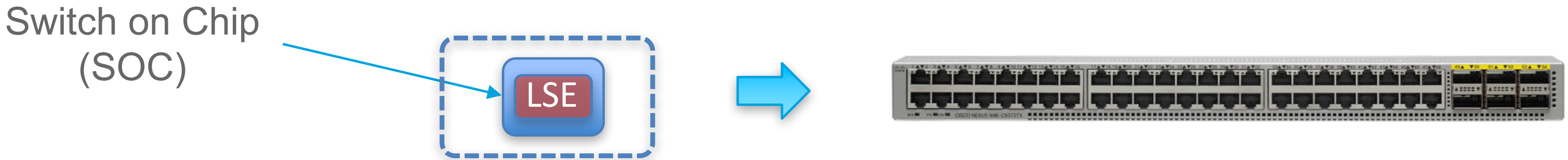
Nexus 9372E



Leverages Merchant (BCOM) + Cisco

Fixed Second Generation Nexus 9200 & 9300EX

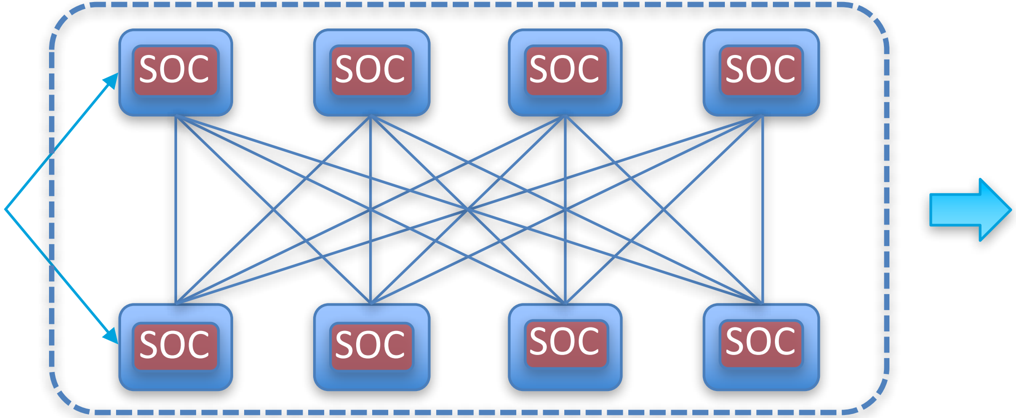
A Single ASIC based Switch

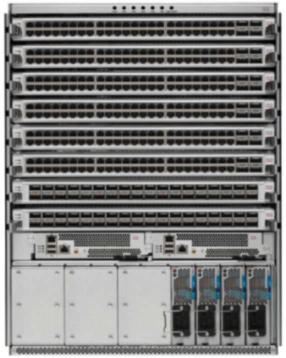


The Switch ‘is’ the ASIC

Modular Nexus 9500

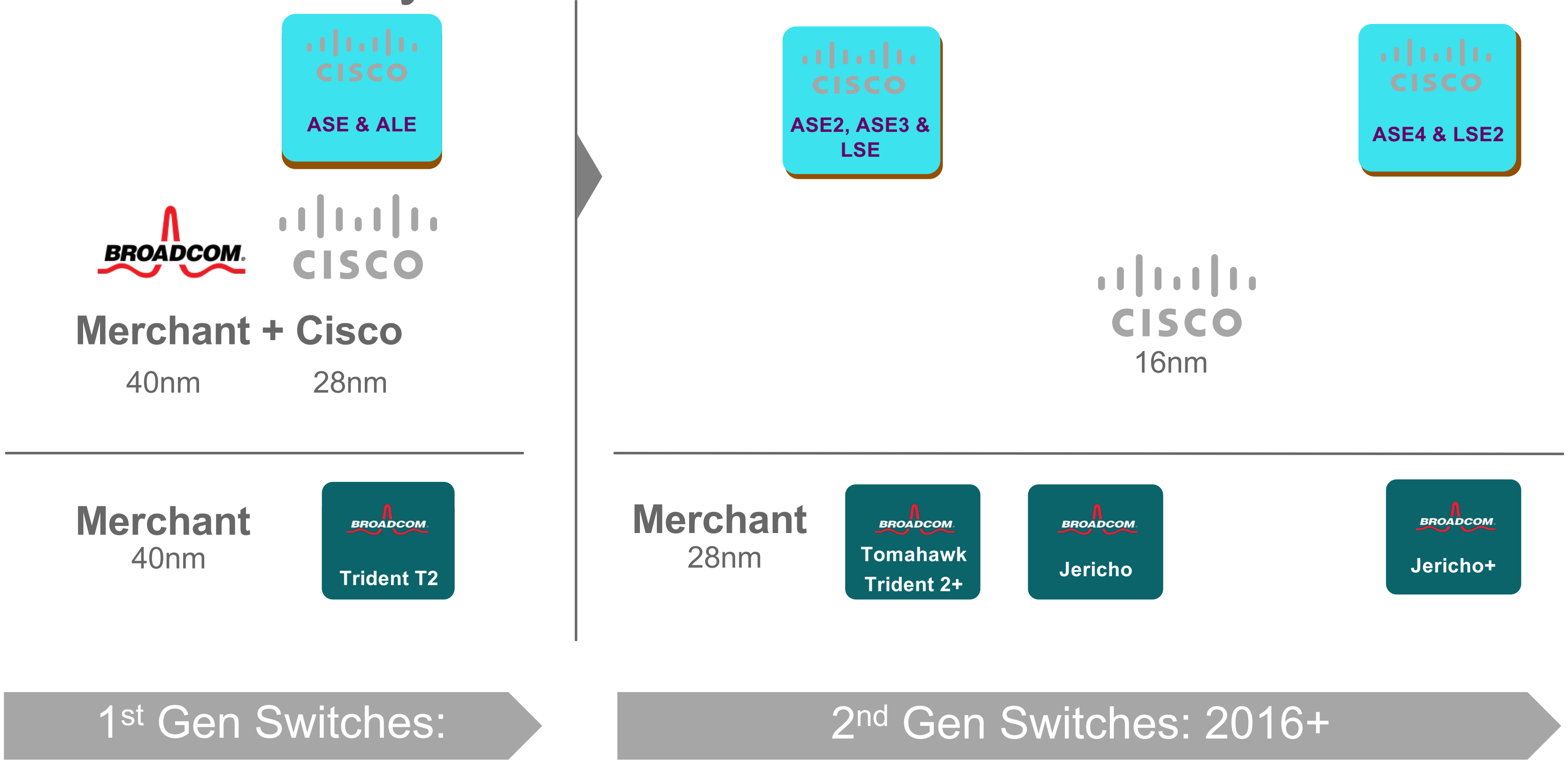
A CLOS Based SOC Architecture





Non Blocking Leaf and Spine based CLOS Network inside the Switch

ASIC Used by Nexus 3000/9000



ASE & ALE

Merchant + Cisco

40nm 28nm

Merchant
40nm

Trident T2

1st Gen Switches:

ASE2, ASE3 & LSE

ASE4 & LSE2

16nm

Merchant
28nm

Tomahawk
Trident 2+

Jericho

Jericho+

2nd Gen Switches: 2016+

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

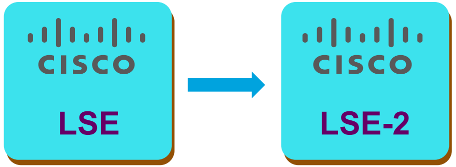
ASIC Used by Nexus 3000/9000

16nm

- LSE – Leaf Spine Engine
- Standalone leaf & spine, ACI leaf and spine
- Flow Table (Netflow, …)
- ACI feature and service and security enhancement
- 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 with flexible DWRR scheduler across 16 queues
- Active Queue Management
  - AFD ,WRED, ECN Marking
- Flowlet Prioritisation, Elephant-Trap for trapping 5 tuple of large flows

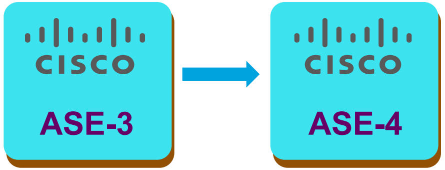
Evolving ASIC ‘Tick’ to the EX ‘Tock’

16nm



- LSE – Leaf Spine Engine
- Standalone leaf & spine, ACI leaf and spine
- **Larger Scale for Route and Policy Tiles**
- Flow Table (**Netflow**, …) + **Streaming HW Statistics**

- **Line Rate Hardware Encryption (MACSEC & CloudSEC)**
- Flowlet Prioritisation, Elephant-Trap for trapping 5 tuple of large flows



- LSE – Leaf Spine Engine
- Standalone leaf & spine, and ACI spine
- **Flow Table (Netflow, …) + Streaming HW Statistics**

- Flowlet Prioritisation, Elephant-Trap for trapping 5 tuple of large flows

Merchant
28nm

ASIC Used by Nexus 3000/9000

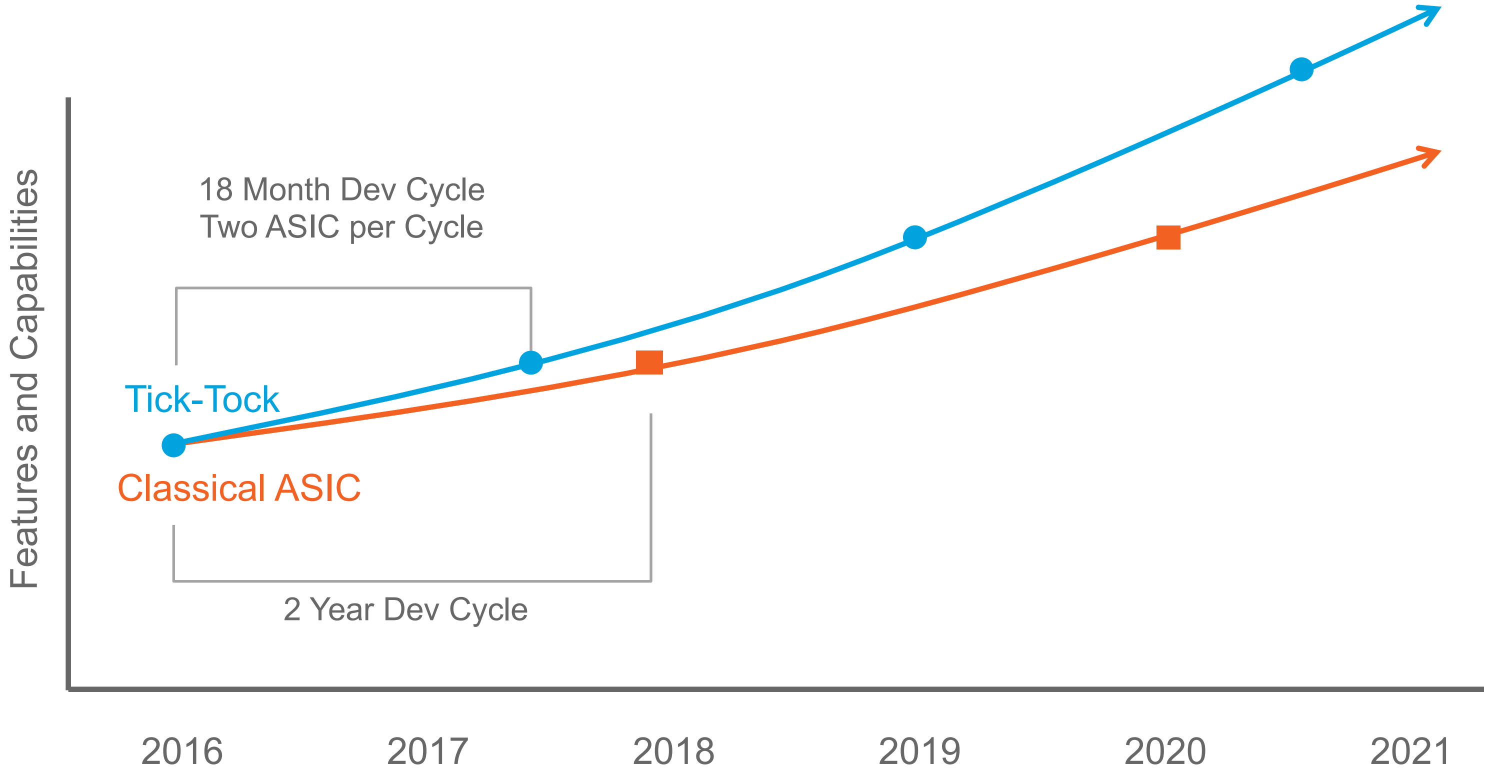
- Broadcom Tomahawk
- 3.2 Tbps I/O & 2.0 Tbps Core

  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)

Development Cycle Decreasing

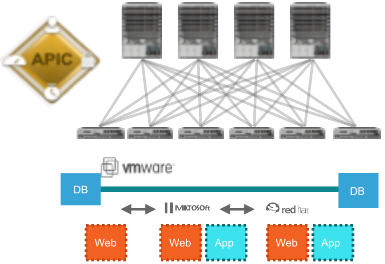
Time to Leverage Moore’s Law is Reducing

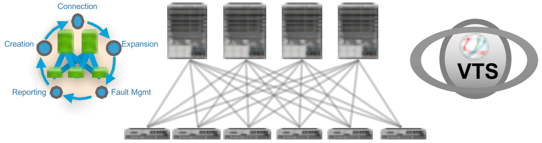


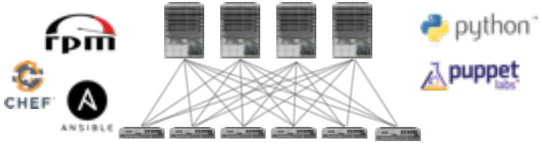
Responding to Fast Market Changes

Sharing Platforms Among Different Architectures

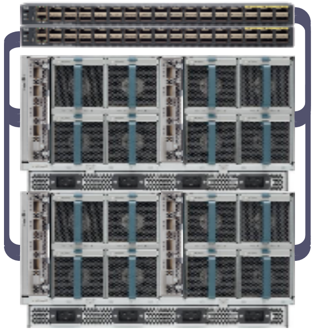
- Common hardware platforms for ACI and NX-OS fabric





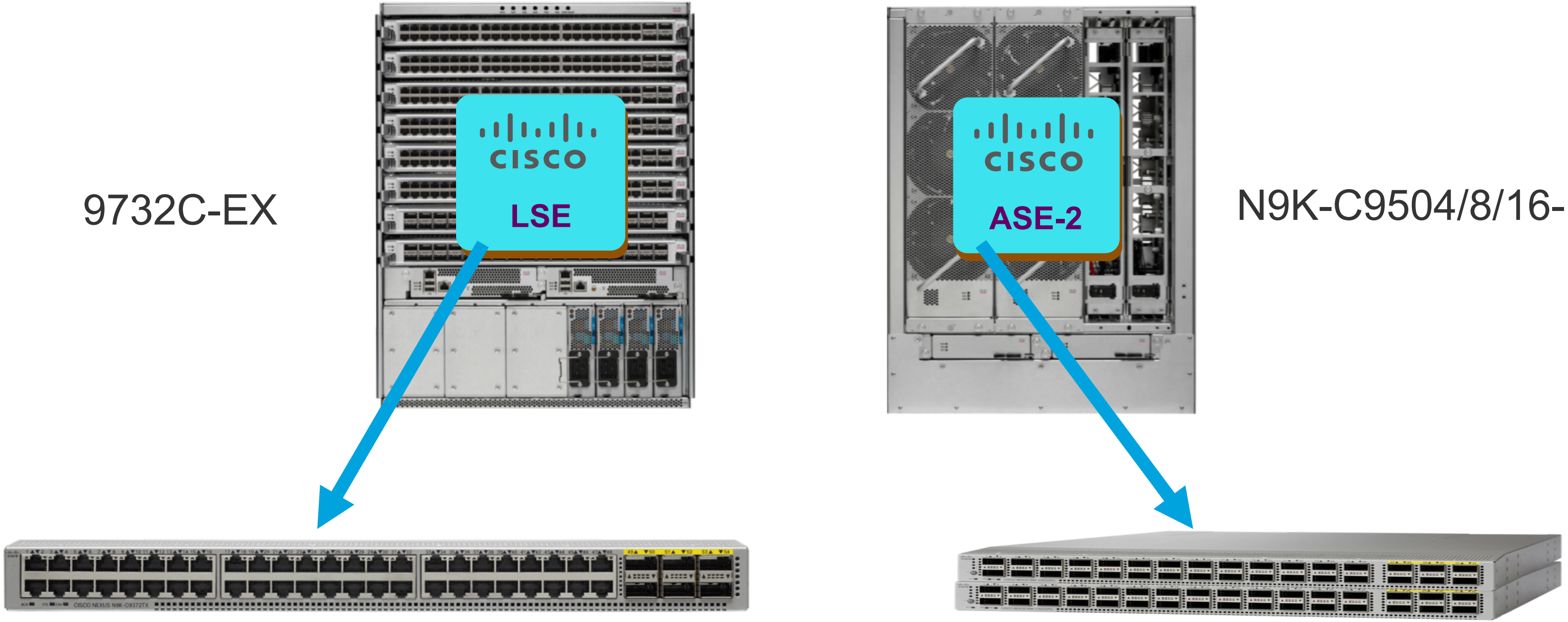


- Sharing platform with UCS FI
  - 3rd Generation FI is based on first gen 9300
  - 4th Generation FI will be based on 2nd Generation 9300EX



Responding to Fast Market Changes

Sharing ASICs Among Platforms



9732C-EX

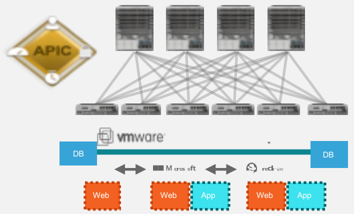
N9K-C9504/8/16-FM-E

N93180YC-EX

N9236C, 9272Q, …

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

Connection
Creation
Expansion
Reporting
Fault Mgmt
VTS

VxLAN-BGP EVPN standard-based

3rd party controller support

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

Programmable Network

rpm
CHEF
ANSIBLE
python
puppet labs

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

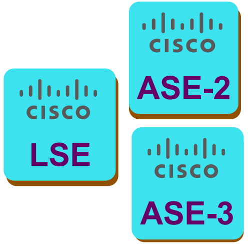
No Changes to EOS and EOL

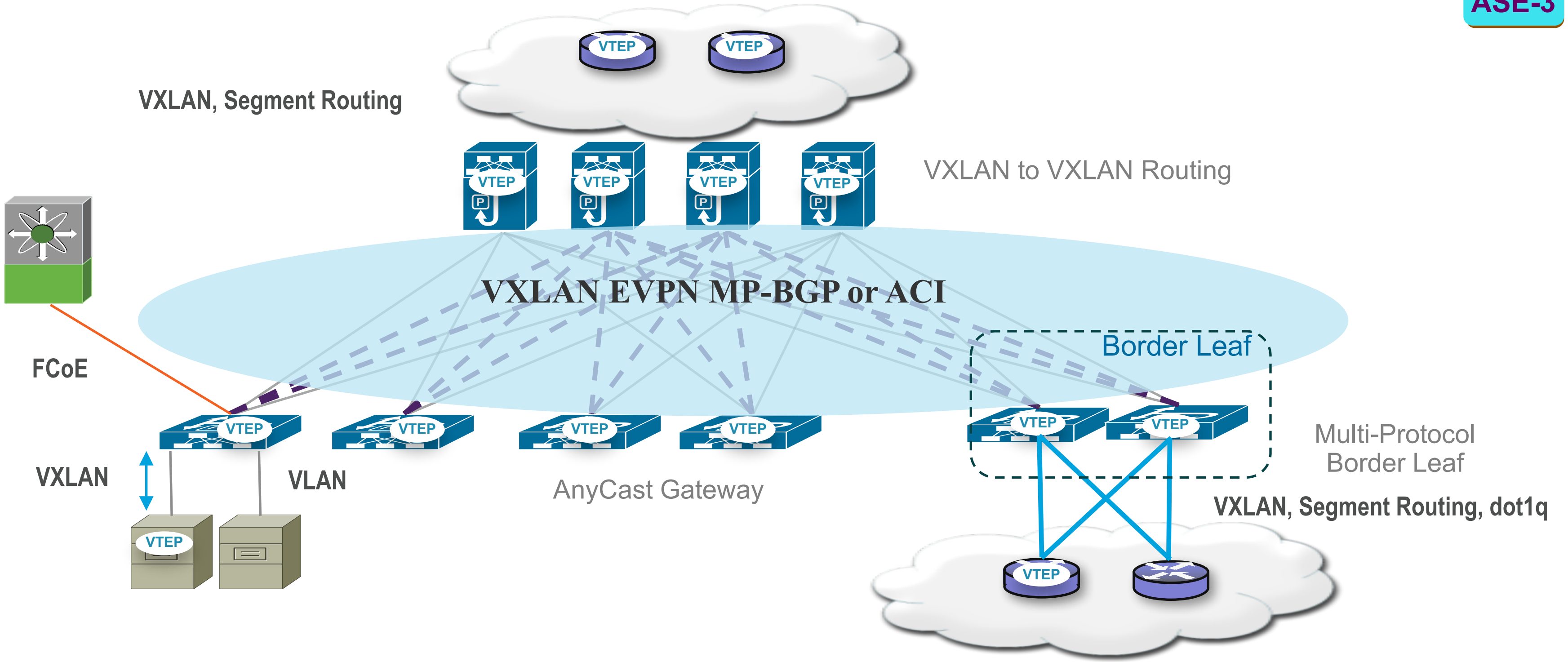
- Will you see more rapid changes in the Networking Space from the Industry?
  - **YES**
- Does this mean you will be forced to upgrade faster?
  - **NO**
- EoS and EoL policies will still be the same
- The choice is still yours

Agenda

- What's New
  - 2nd Generation Nexus 9000
  - Moore's Law and 25G SerDes
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- Design Impacts of 25G, 50G and 100G
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)

Nexus 9000 Forwarding





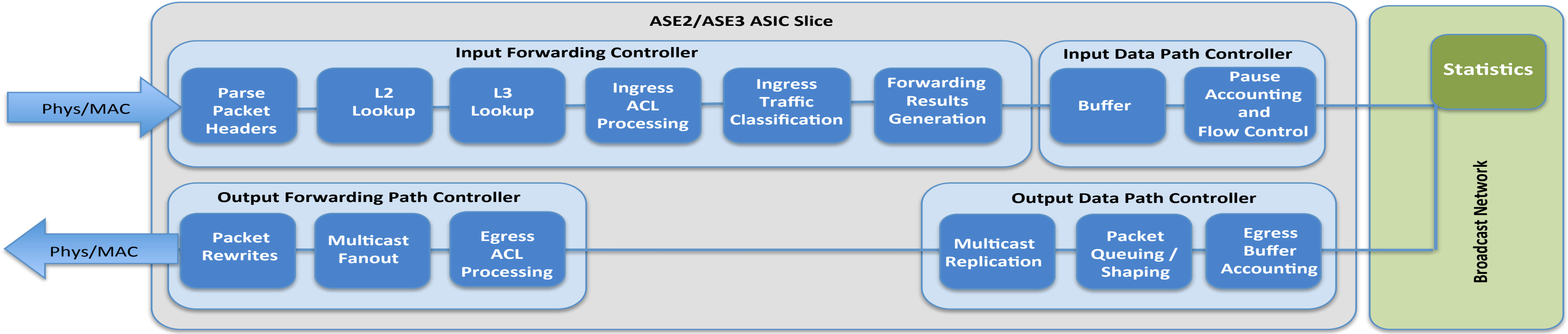
Nexus 9000 Life of a Packet

ASE2 / ASE3 / LSE

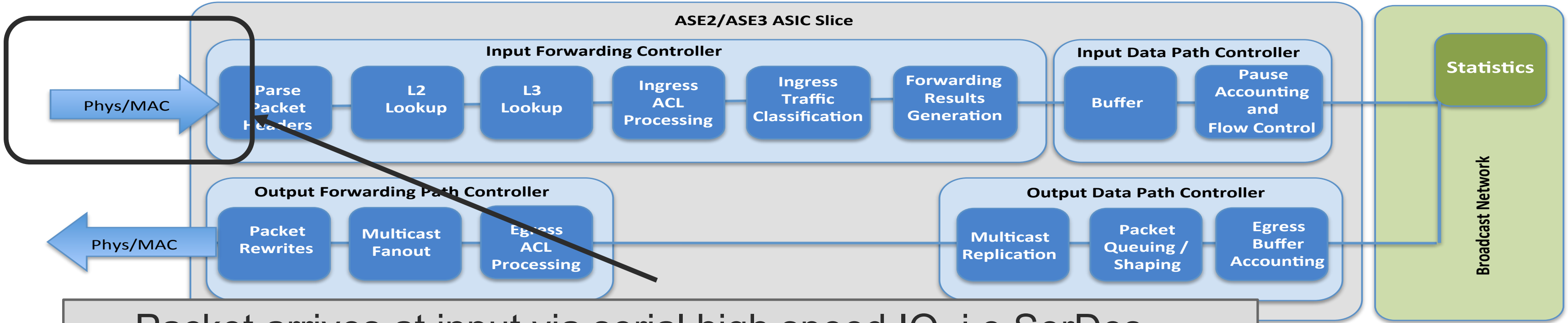
LSE

ASE-2

ASE-3

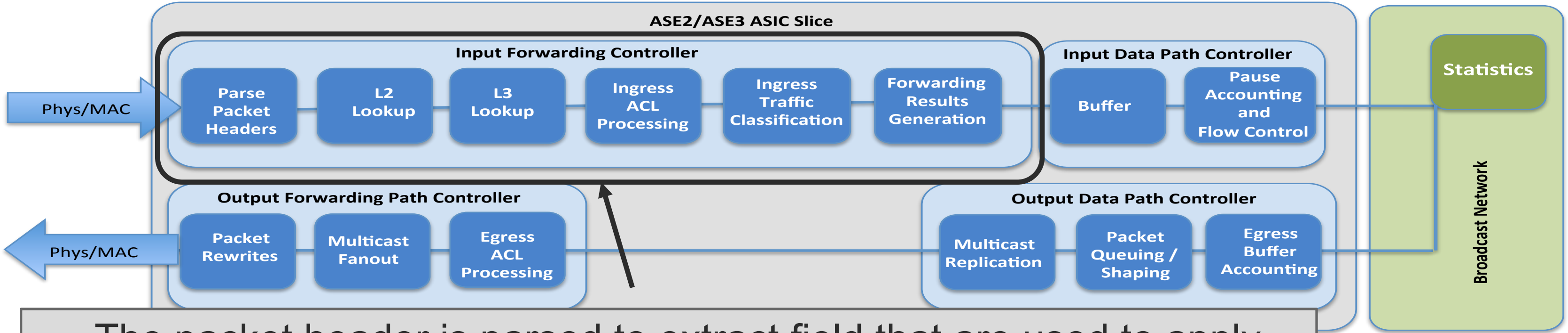


Life of a Packet in ASE2 / ASE3 / LSE ASIC



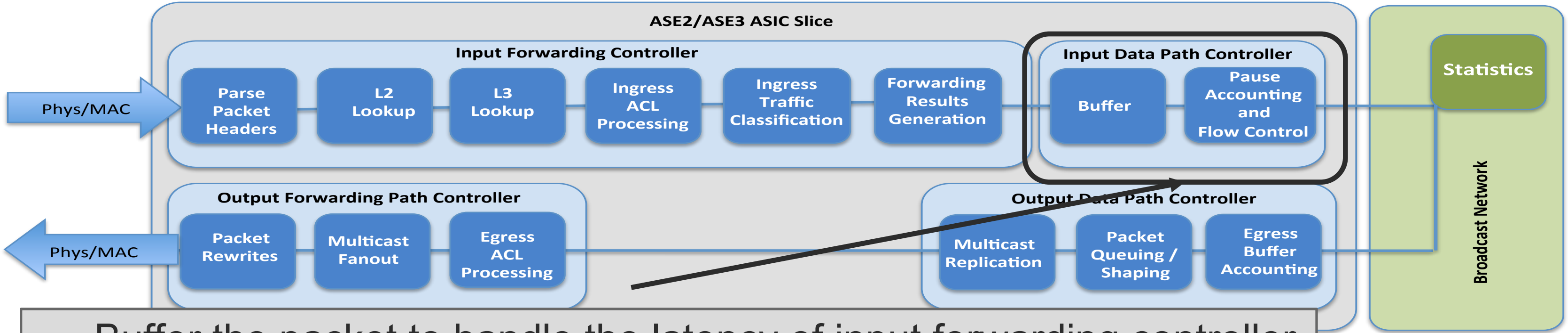
- Packet arrives at input via serial high speed IO, i.e SerDes
- The serial data is converted to parallel stream and MAC is responsible to validate framing protocol
- The MAC operates in cut through and pass the packet to client interface

Life of a Packet in ASE2 / ASE3 / LSE ASIC



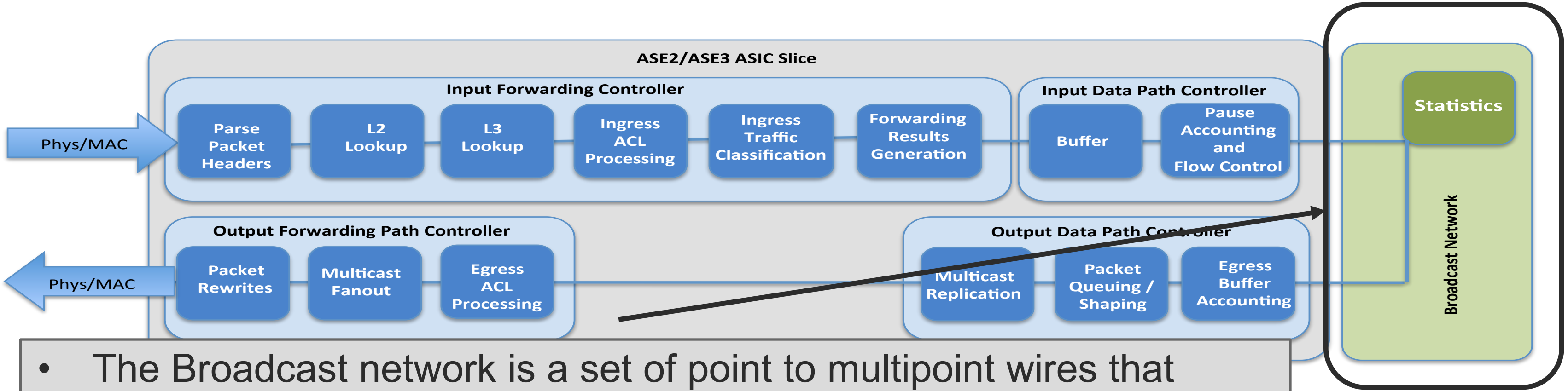
- The packet header is parsed to extract field that are used to apply policy and making forwarding decision and load-balancing
- The parsed field are used in a series of forwarding table and access control list lookup
- Flow Table Analytics

Life of a Packet in ASE2 / ASE3 / LSE ASIC



- Buffer the packet to handle the latency of input forwarding controller pipeline
- Perform pause accounting and flow control generation
- Implements headroom buffers for PAUSE absorption

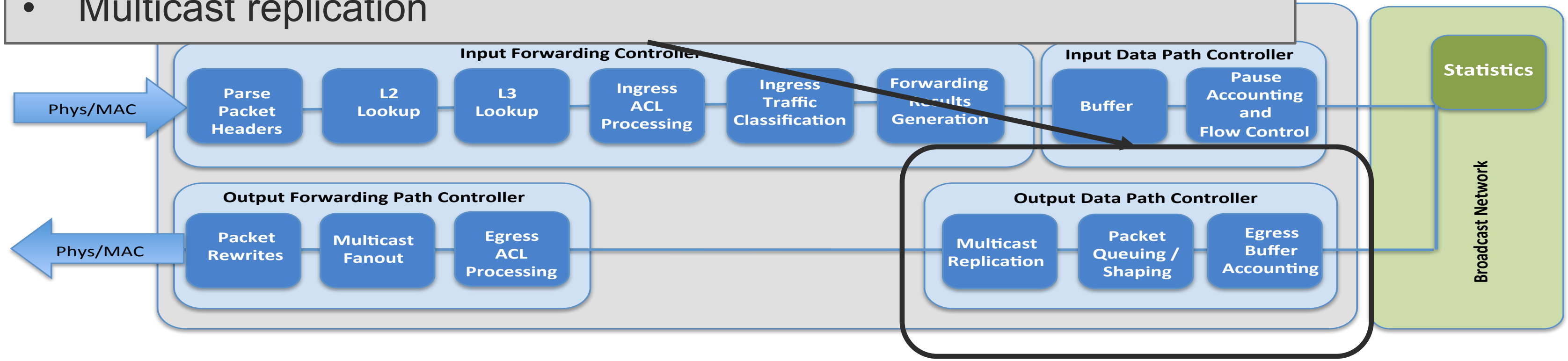
Life of a Packet in ASE2 / ASE3 / LSE ASIC



- The Broadcast network is a set of point to multipoint wires that allows any to any connectivity between the slices.
- Each input slice drives wires that is connected to all output slices
- This is ***not*** a scheduled network, each output slice has bandwidth to accept data from all input slices *simultaneously*

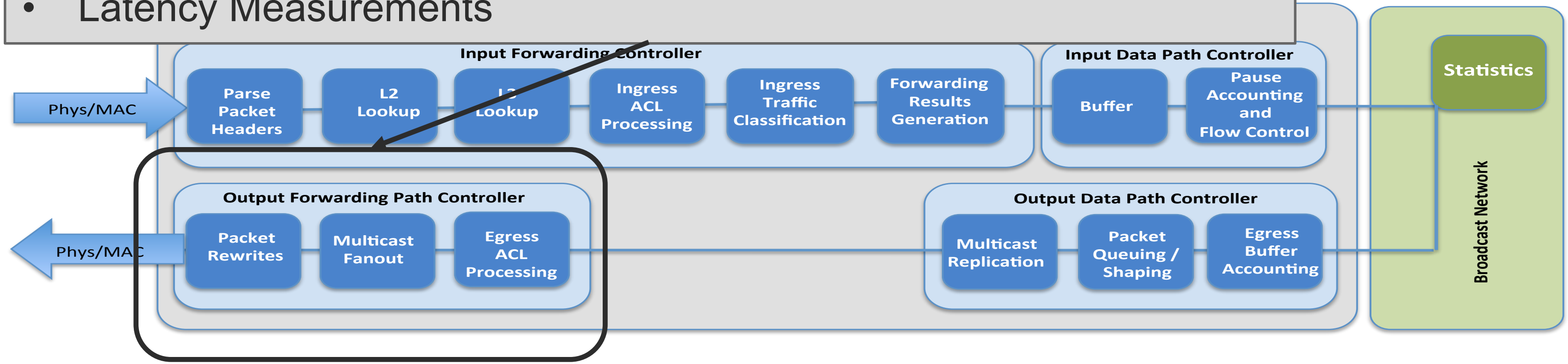
Life of a Packet in ASE2 / ASE3 / LSE ASIC

- Output packet buffering
- Packet buffer accounting
- Output queuing and scheduling
- Multicast replication



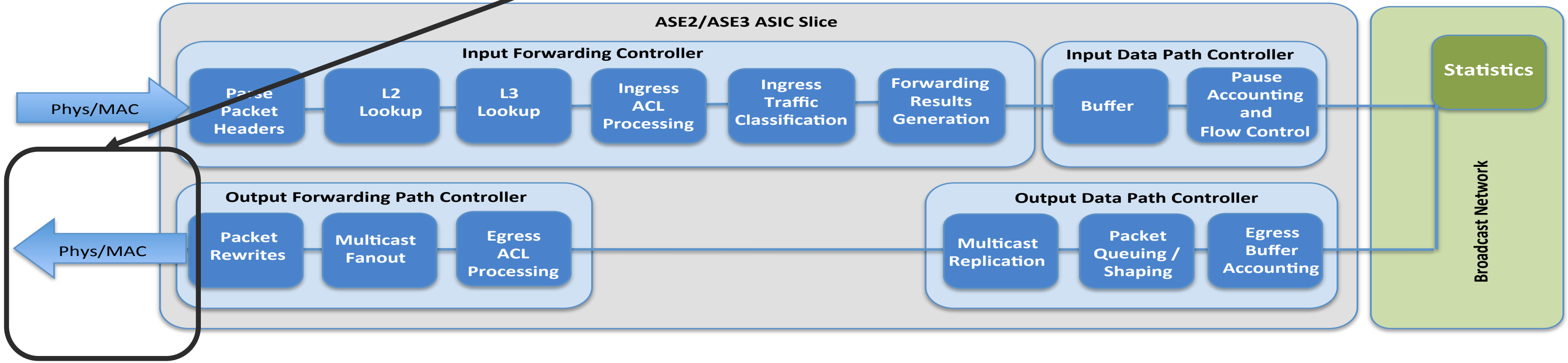
Life of a Packet in ASE2 / ASE3 / LSE ASIC

- Output forwarding controller performs egress ACLs
- It performs packet rewrite and encapsulation
- It performs multicast expansion
- Latency Measurements

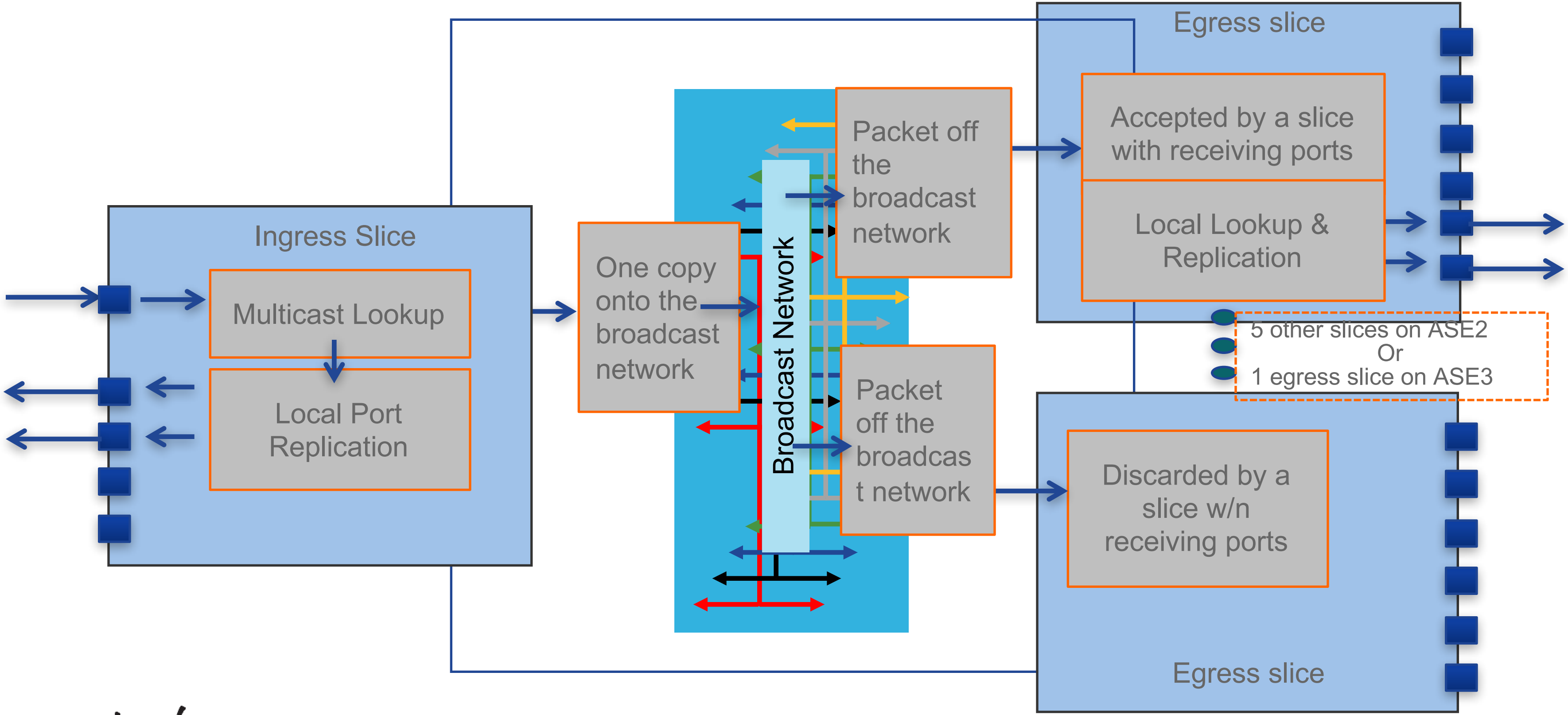


Life of a Packet in ASE2 / ASE3 / LSE ASIC

- Packet leaves the output via serial high speed IO, i.e SerDes



Multicast Packet Forwarding


Ingress Slice
Multicast Lookup
Local Port Replication
One copy onto the broadcast network
Broadcast Network
Packet off the broadcast network
Packet off the broadcas t network
Egress slice
Accepted by a slice with receiving ports
Local Lookup & Replication
5 other slices on ASE2
Or
1 egress slice on ASE3
Discarded by a slice w/n receiving ports
Egress slice

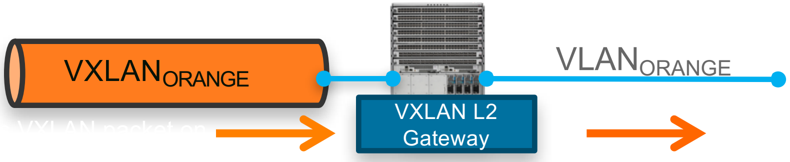
Agenda

- What's New
- Next Generation Capabilities
  - Forwarding – Packet Walks
  - Forwarding – Protocol Support
  - Forwarding - Table Templates
  - Telemetry
  - Encryption (MACSEC and CloudSEC)
  - QoS & Buffering
- Design Impacts of 25G, 50G and 100G
- Next Gen Nexus 9000 Switch Platforms

VXLAN Support
Gateway, Bridging, Routing*

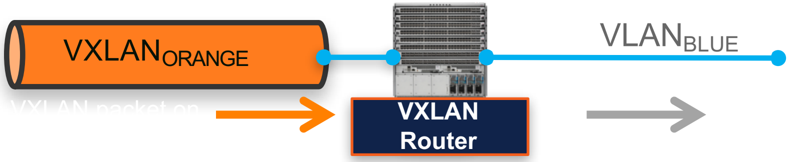
VXLAN to VLAN Bridging

(L2 Gateway)



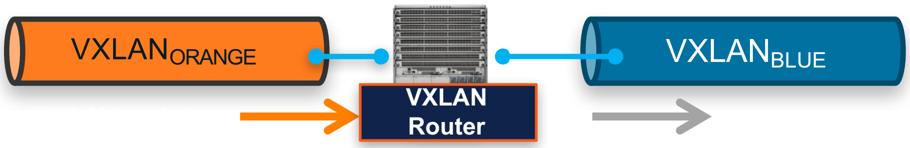
VXLAN to VLAN Routing

(L3 Gateway)



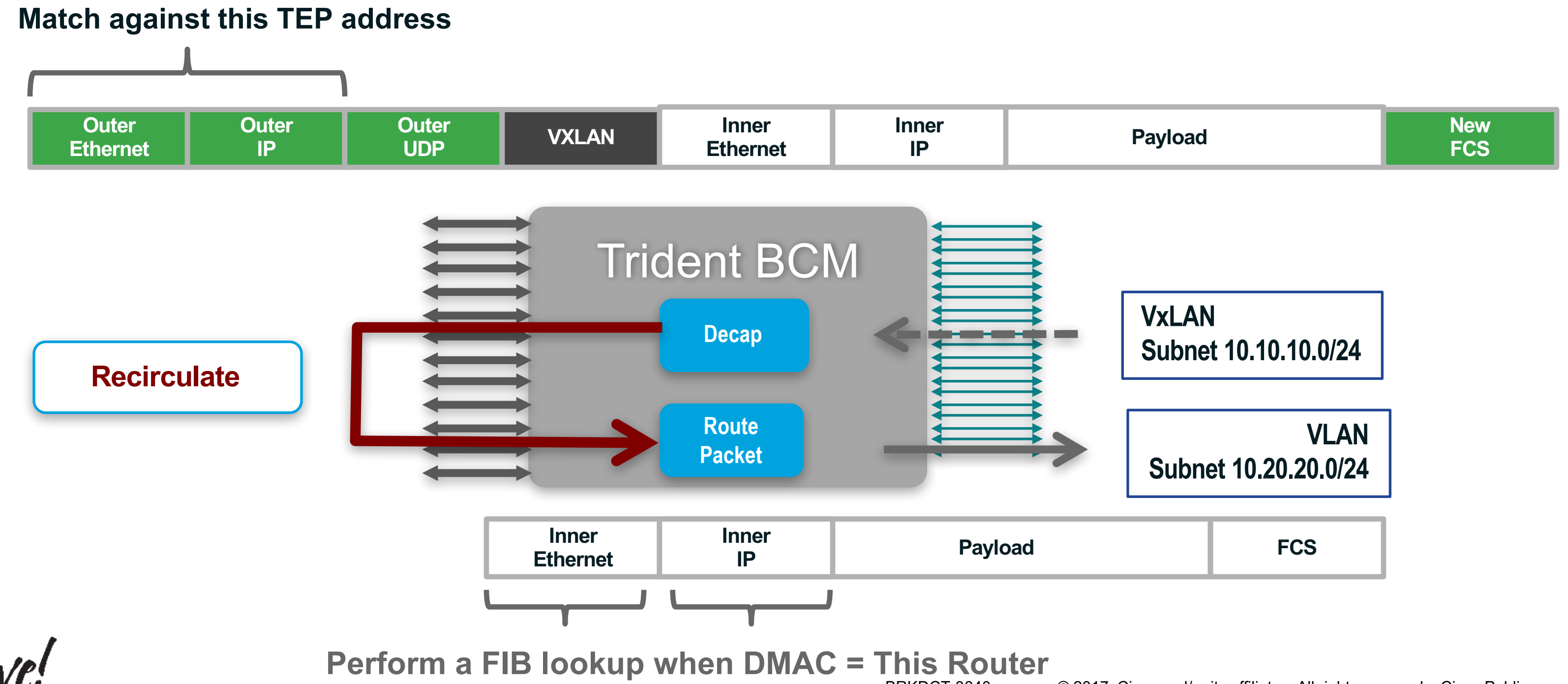
VXLAN to VXLAN Routing

(L3 Gateway)



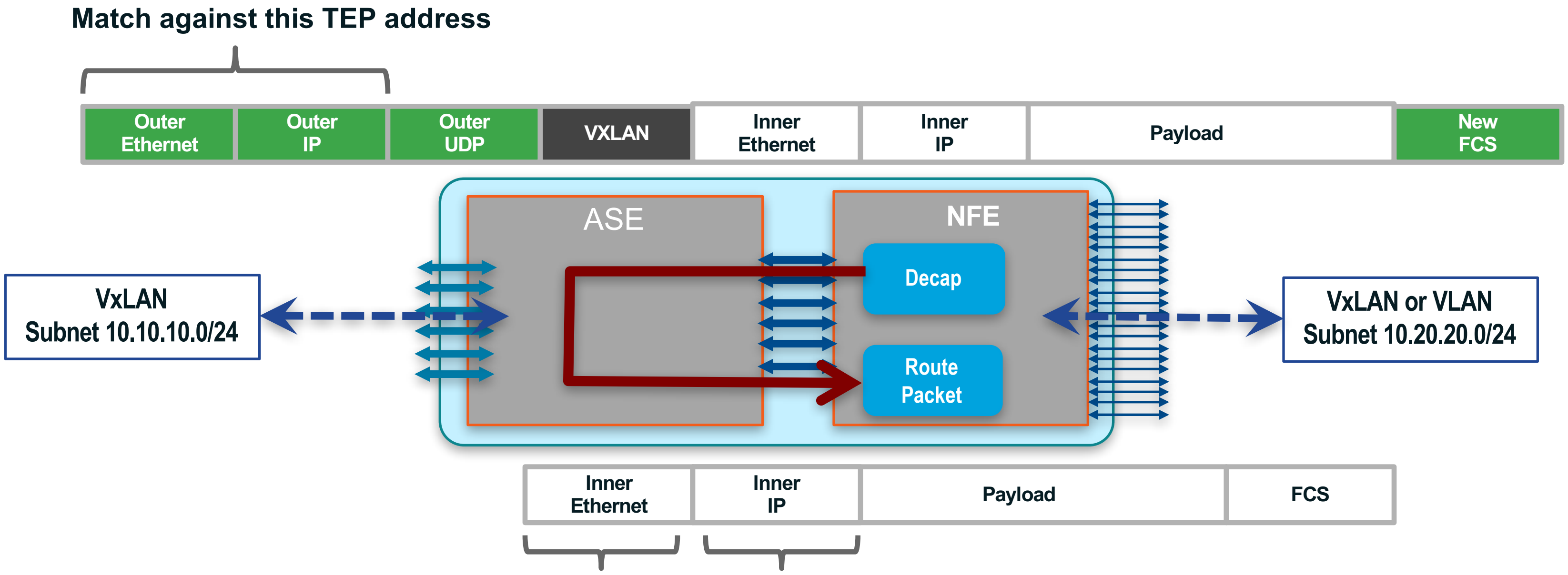
VxLAN to VLAN Routing – Trident 2

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



VLAN/VxLAN to VxLAN Routing
First Gen Nexus 9300 NX-OS Mode

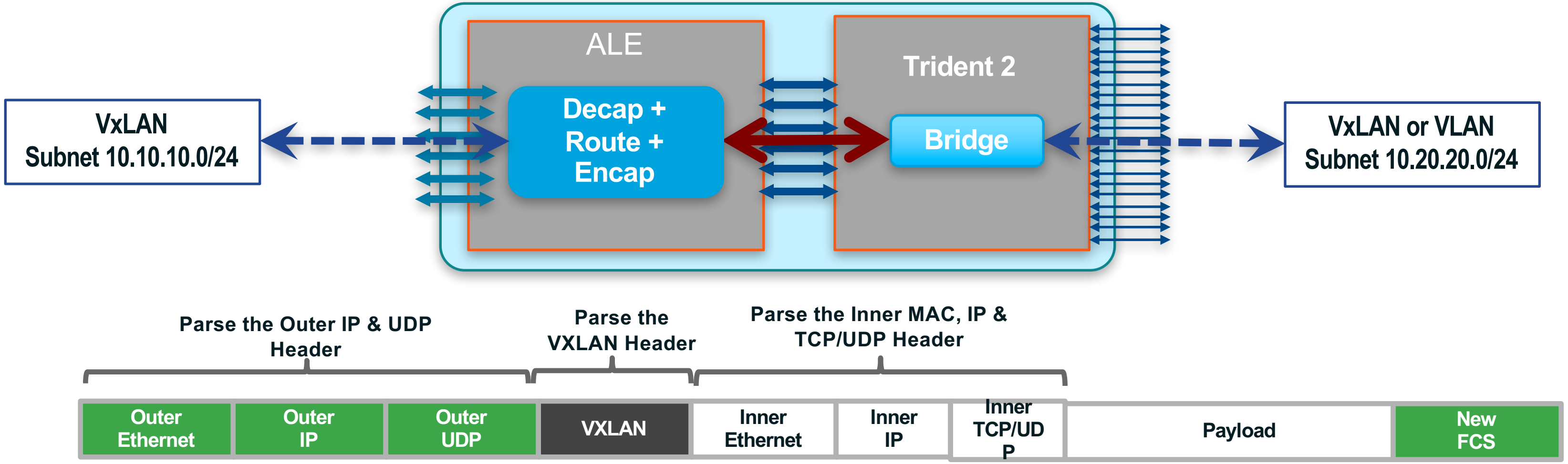
- In NX-OS mode forwarding is performed by the NFE (Trident-2) ASIC
- ALE provides extended buffer, some SPAN and ERSPAN functions
- Re-circulation is performed for VXLAN Routing



Perform a FIB lookup when DMAC = This Router

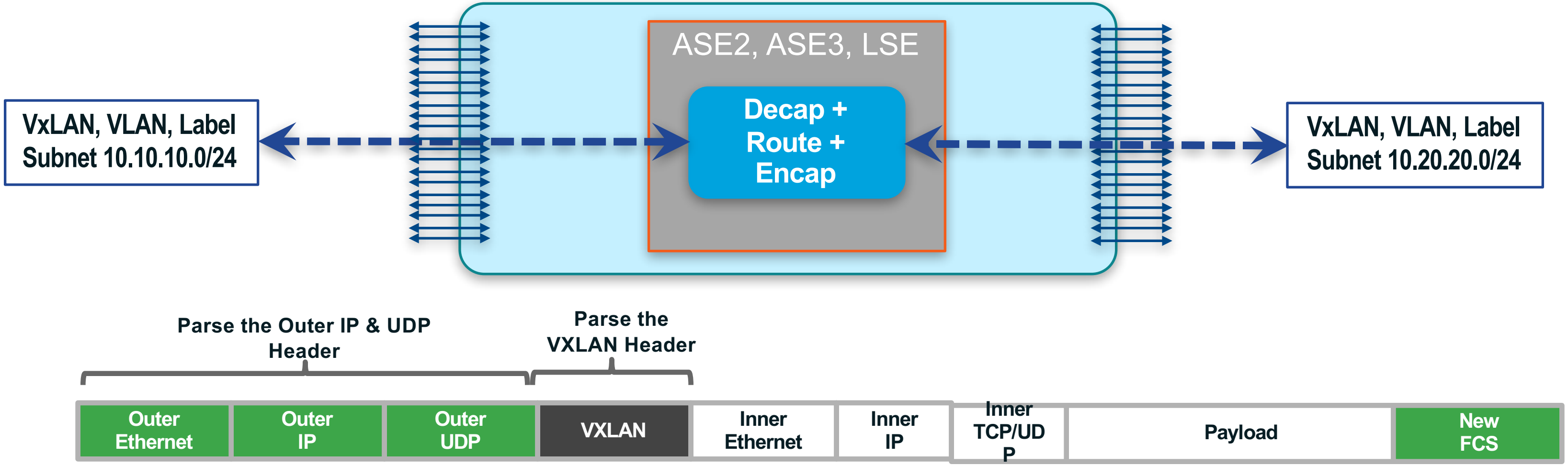
VLAN/VxLAN to VxLAN Routing
First Gen Nexus 9300 ACI Mode

- 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

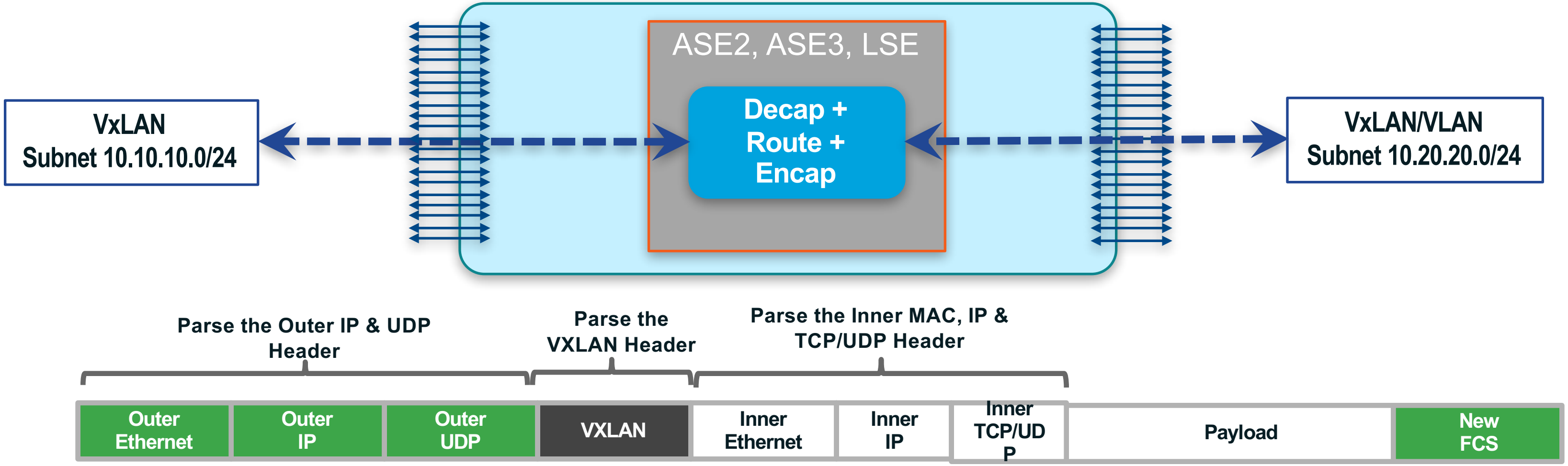
- ASE2, ASE3 & LSE ASIC parse the full outer MAC, IP/UDP header, VXLAN header in one pipeline pass
- VLAN to VXLAN 'and' VXLAN to VXLAN routing is performed in a single pass
- Line rate performance for all encapsulations with all packet sizes



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



VxLAN
Subnet 10.10.10.0/24

ASE2, ASE3, LSE

Decap +
Route +
Encap

VxLAN/VLAN
Subnet 10.20.20.0/24

Parse the Outer IP & UDP Header

Parse the VXLAN Header

Parse the Inner MAC, IP & TCP/UDP Header

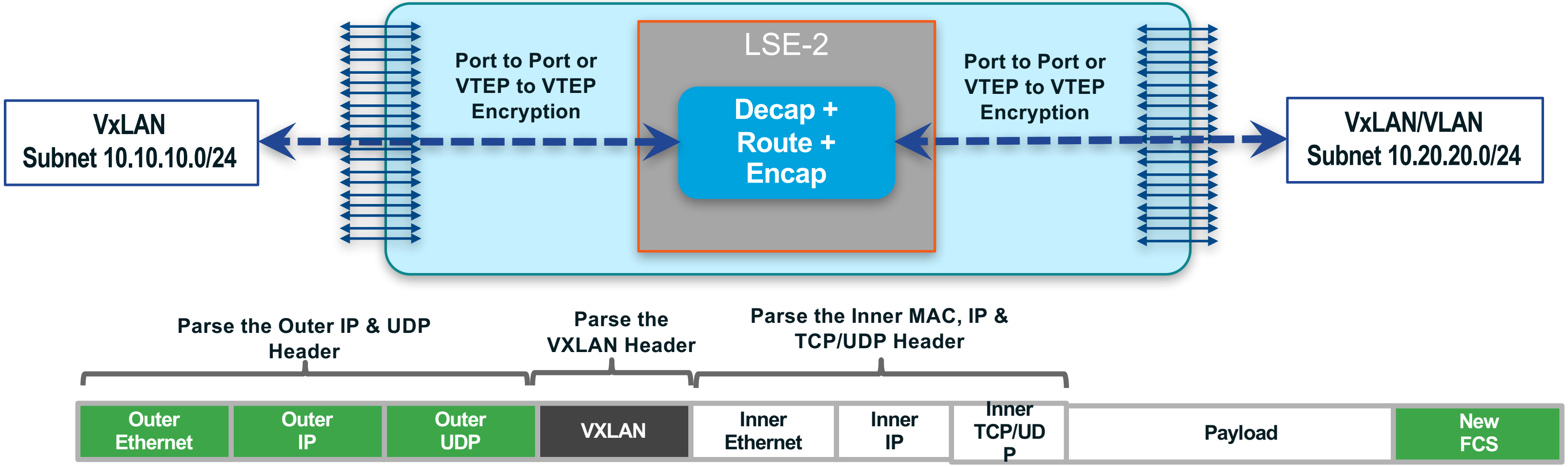
Outer Ethernet	Outer IP	Outer UDP	VXLAN	Inner Ethernet	Inner IP	Inner TCP/UDP	Payload	New FCS

Cisco live!

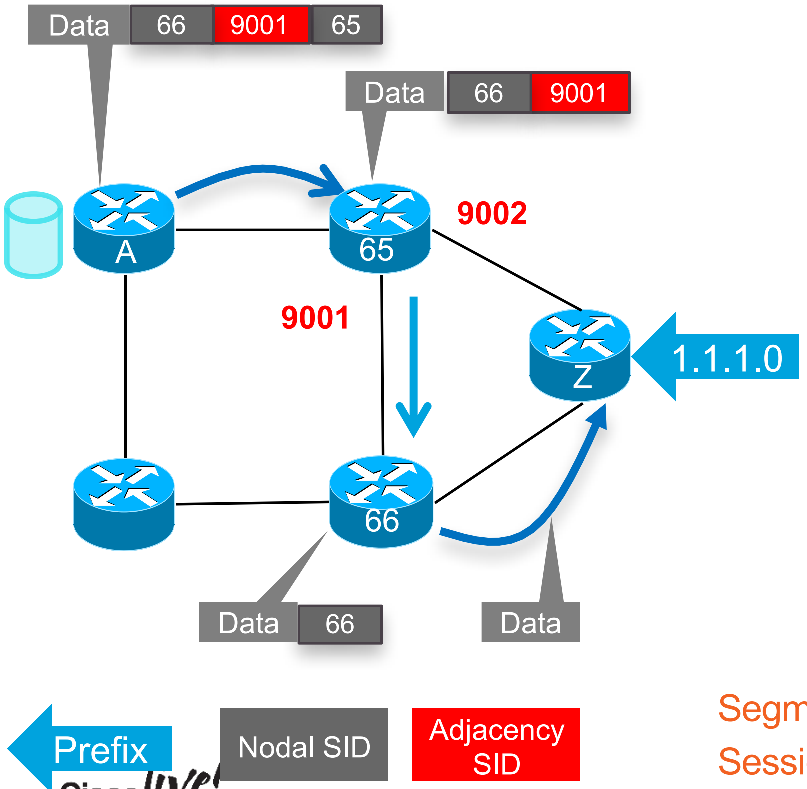
VLAN/VxLAN to VxLAN Routing and Encryption
Nexus 9300**FX** ACI and Standlaone Mode

FYI

- All of the ‘EX’ Capability ‘plus’
- MACSEC – Encryption of traffic port to port (100Gbps per port)
- CloudSec – Encryption of traffic over L3 backbone at line rate (100Gbps per port)
  - GCM-AES-128 (32-bit PN), GCM--AES-256 (32-bit PN), GCM-AES-128-XPN (64-bit PN), GCM-AES-256-XPN (64-bit PN)



Segment Routing – MPLS w/ Explicit Path Control 9200 and 9300EX



Data-Plane: Uses MPLS label stack to perform Source Routing

Control-Plane: BGP-LU, BGP endpoints and IP Prefixes are learned through hop by hop LU underlay

A stack of Segments can be used by the source to steer any flow along any desired path by encoding it in packet header as an ordered list of segments

Shipping – N3k/N9K
- Node-SID/Prefix-SID
- BGP-LU for control plane

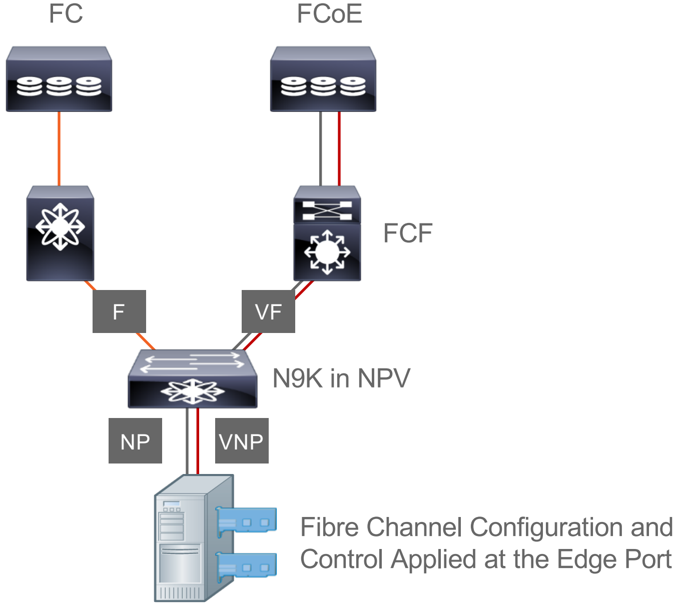
Q3CY16 – N3K/N9K
- Adjacency-SID; Binding SID
- Egress Peer Engineering with BGP-LS
- L3VPN/EVPN support over SR (Q4CY16)

Segment Routing in Data Centre using Nexus 9000 and 3000
Session ID: BRKDCN-2050 & Session ID: LABRST-2020

FCoE NPV – Unified Fabric Switching
Nexus 9300 & 9300EX

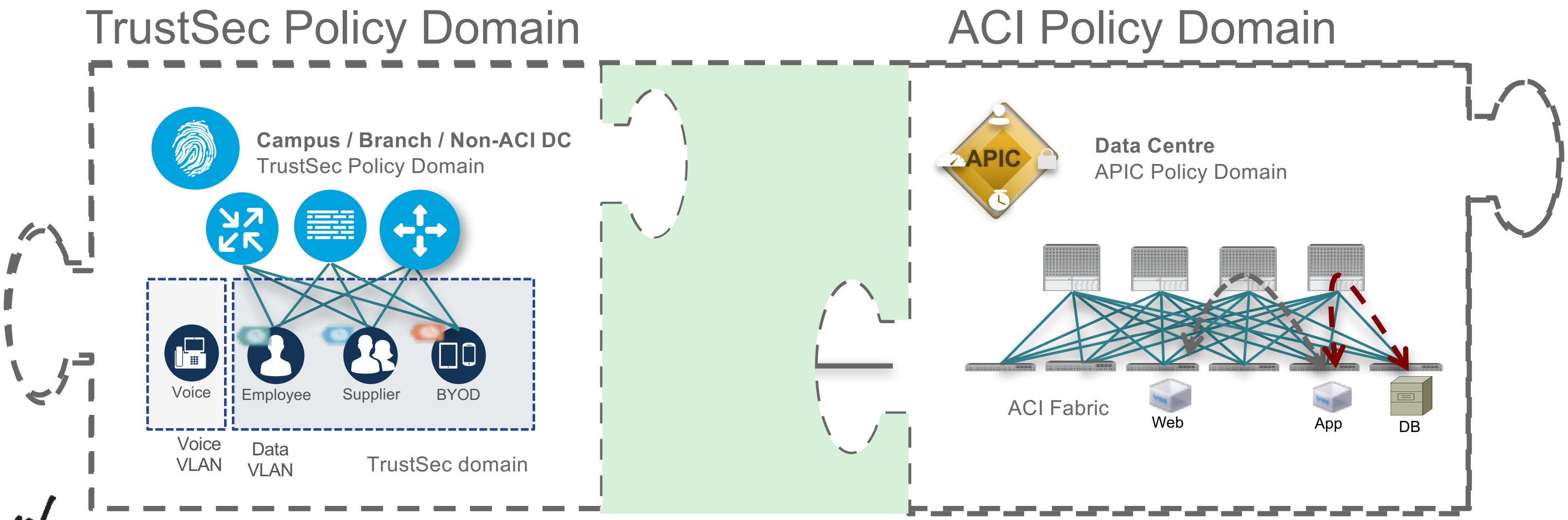
Connect FCoE-capable Hosts to a FCoE-Capable FCoE Forwarder (FCF) Device

- Standalone NX-OS support
  - FCoE NPV on N92xx and N93xx
  - FCoE on FEX N2348UPQ
- ACI support
  - 9300-EX
  - FEX including B22


FC
FCoE
FCF
F
VF
N9K in NPV
NP
VNP
Fibre Channel Configuration and Control Applied at the Edge Port

Enabling Group-Based Policies Across the Enterprise
VXLAN-GPE (ACI EPG) and TrustSec SGT

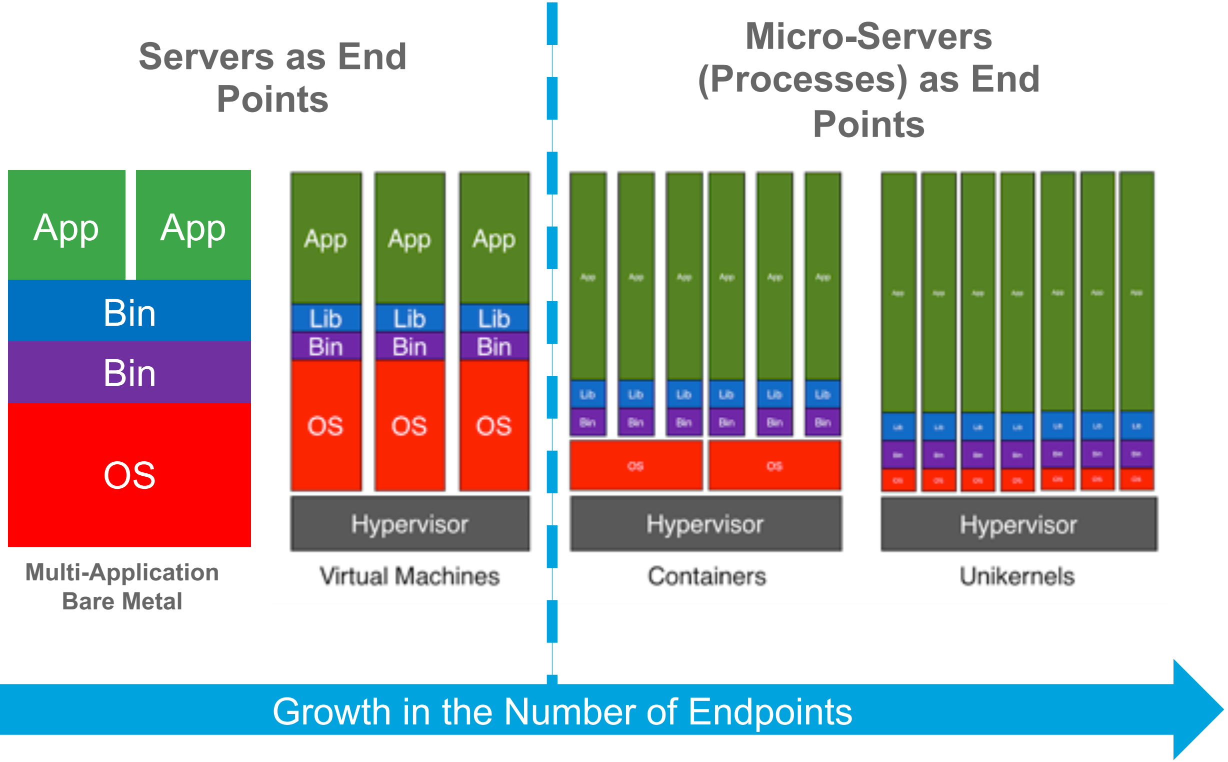
- Goal: **Consistent Security Policy Groups** and **Identity** shared between TrustSec and ACI domains
- Allow TrustSec security groups to be used in ACI policies
- Allow ACI EndPoint Groups to be used in policies across the Enterprise



Agenda

- What’s New
- Next Generation Capabilities
  - Forwarding – Packet Walks
  - Forwarding – Protocol Support
  - Forwarding - Table Templates
  - Telemetry
  - Encryption (MACSEC and CloudSEC)
  - QoS & Buffering
- Design Impacts of 25G, 50G and 100G
- Next Gen Nexus 9000 Switch Platforms

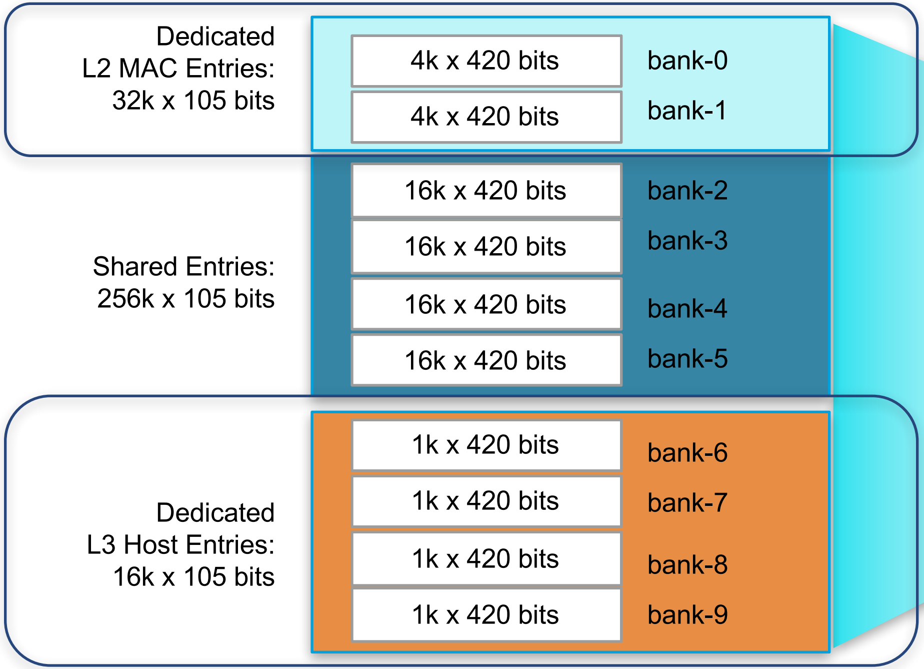
Nexus Forwarding Table Templates
Responding to changes in End Point Density



Unikernels, also know as "virtual library operating system"

NFE (Trident 2) Unified Forwarding Table Modes

- 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)



Dedicated
L2 MAC Entries:
32k x 105 bits

4k x 420 bits	bank-0
4k x 420 bits	bank-1

Shared Entries:
256k x 105 bits

16k x 420 bits	bank-2
16k x 420 bits	bank-3
16k x 420 bits	bank-4
16k x 420 bits	bank-5

Dedicated
L3 Host Entries:
16k x 105 bits

1k x 420 bits	bank-6
1k x 420 bits	bank-7
1k x 420 bits	bank-8
1k x 420 bits	bank-9

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

First Gen Nexus 9300 Forwarding Templates

```
N9k-1(config)# system routing max-mode l3
Warning: The command will take effect after next reload.
Note: This requires copy running-config to startup-config before switch reload.
N9k-1#
```

	Nexus 9300	
	Default	Maximum Layer-3 Mode
LPM Routes	16K	128K
IP Host Entries	120K (208K protocol learned IPv4 host routes)	16K
MAC Address Entries	96K	32K
Multicast Routes	32K* (hardware capable of 72K)	8K*
Multicast Fan Outs	8K (no vPC)	8K (no vPC)
IGMP Snooping Groups	32K* (hardware capable of 72K)	8K*

http://www.cisco.com/c/dam/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-736548.pdf

* Shared with IP hosts

First Gen Nexus 9300 Forwarding Templates

	Switch CLI	T2 BCM-shell
MAC Table	show mac address-table count	l2 show
IP Host Table	RIB: show ip route sum Show ip route FIB: sh forwarding route summary mod <#> sh forwarding route	l3 l3table show [on LC] n9k# bcm-shell mod 1 "l3 l3table show" | count
IP LPM Table	RIB: show ip route sum show ip route FIB: show forwarding route sum mod <#> show forwarding route	l3 defip show [on FM] n9k# bcm-shell mod 22 "l3 defip show" | count
egress next-hop table		l3 egress show [on both LC and FM] n9k# bcm-shell mod 1 "l3 egress show" | count

BRKDCT-3101 - Nexus 9000 (Standalone) Architecture Brief and Troubleshooting

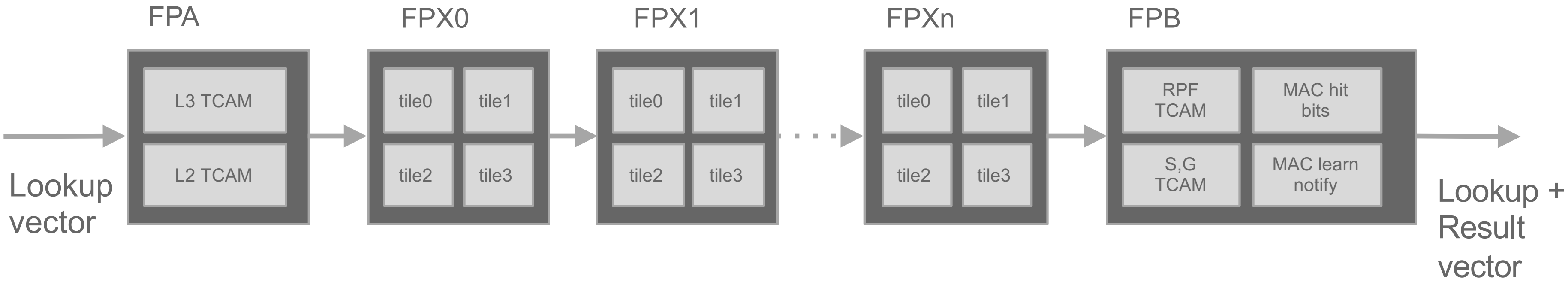
BRKCLD-2601 - Layer 3 Forwarding and Troubleshooting Deep Dive on Nexus 9000

Nexus 9000 2nd Generation Templates Tile Based Forwarding Tables

ASE-2

LSE

ASE-3



- 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

Forwarding Table Compression

- Eliminating repetitive information from forwarding table. Increased table scale with same amount of SRA. Effectively compress forwarding table entries.
- Applicable for IPv4 host, IPv4 LPM routes and IPv6 /64 LPM routes

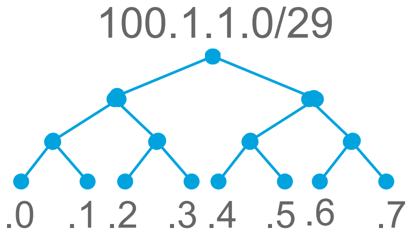
Destination IP	Next_hop
100.1.1.1/32	2.2.2.2
100.1.1.2/32	2.2.2.2
100.1.1.3/32	2.2.2.2
100.1.1.4/32	2.2.2.2
100.1.1.5/32	2.2.2.2

Common Information that can be eliminated

Pivot Entry
100.1.1.0/29

TRIE Entry	Next_Hop
.1	2.2.2.2
.2	2.2.2.2
.3	2.2.2.2
.4	2.2.2.2
.5	2.2.2.2

3 bits required per entry. Able to pack more entries with same amount of memory



N9300-EX Forwarding Table Templates

Examples

- Initial template supporting for standalone
- ACI Support for Templates with 3.0 release (Q3CY17)

Sample template 1

Table Type	IPv4 Hosts	IPv4 LPM	IPv6 Hosts	IPv6 LPM	MAC	Multicast	Next_Hop	IPv4 MPLS
Scale	700K*	700K*	2K	2K	96K	32K	32K	16K

* shared entry. IPV6 entries in TCAM and are shared

Sample template 2: High IPv4 Host route and IPv4 LPM Scale with IPv6 entries

Table Type	IPv4 Hosts	IPv4 LPM	IPv6 Hosts	IPv6 LPM	MAC	Multicast	Next_hop	IPv4 MPLS
Scale	640K*	640K*	16K	2K	96K	32K	32K	16K

* shared entry. IPv6 LPM entries in TCAM

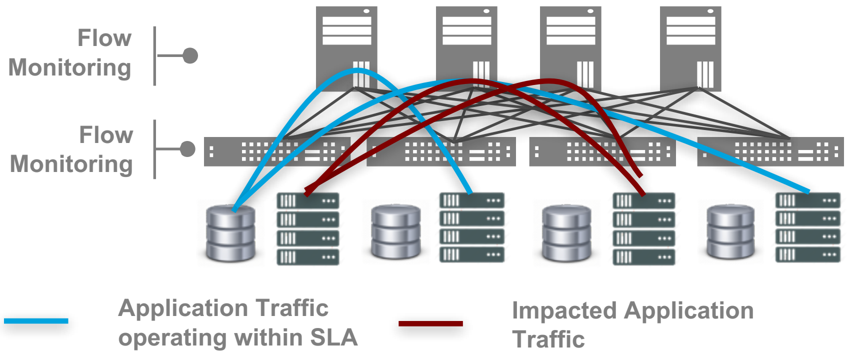
Agenda

- What's New
- Next Generation Capabilities
  - Forwarding – Packet Walks
  - Forwarding – Protocol Support
  - Forwarding - Table Templates
  - Telemetry
  - Encryption (MACSEC and CloudSEC)
  - QoS & Buffering
- Design Impacts of 25G, 50G and 100G
- Next Gen Nexus 9000 Switch Platforms

Fabric Wide Troubleshooting
Real Time Monitoring, Debugging and Analysis

Granular Fabric Wide Flow Monitoring Delivering Diagnostic Correlation

“Tetration Analytics”


Flow Monitoring
Flow Monitoring
Application Traffic operating within SLA
Impacted Application Traffic

Debug

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

Monitor

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

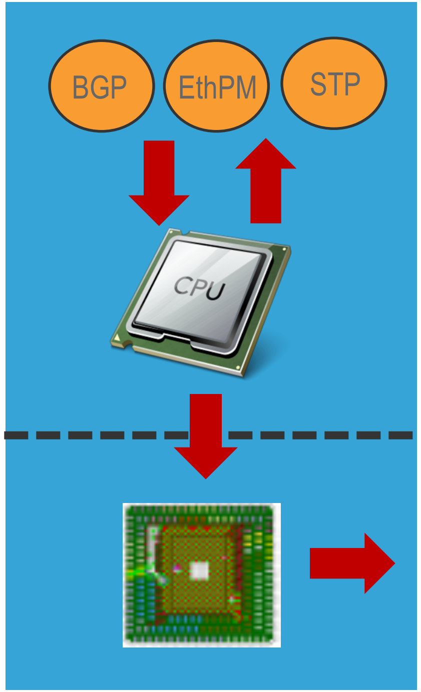
Analyse

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

Improving the Efficiency of Accessing HW state

Direct Export of the Hardware State

Monitor SW State (polled, timer driven, on demand, …)



Configure Desired Triggers (Events, Flows, …)

CPU sources the SW Telemetry Data (everything not in the HW export)

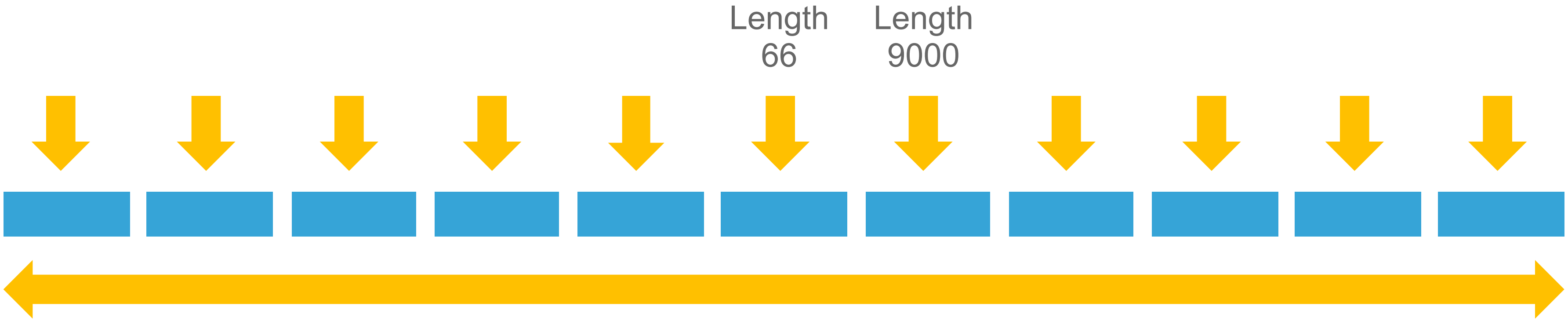
Configure Required Telemetry (Process State, Flow Cache, Events, SSX)

ASIC Directly Transmits HW Telemetry Data (Timer and Event Triggers)

Real-time Flow Sensors
ASE-3 & LSE (the ‘X’ in the 9200-X and 9300-X)

- Granular flow information
  - Per flow statistics
  - Per packet visibility

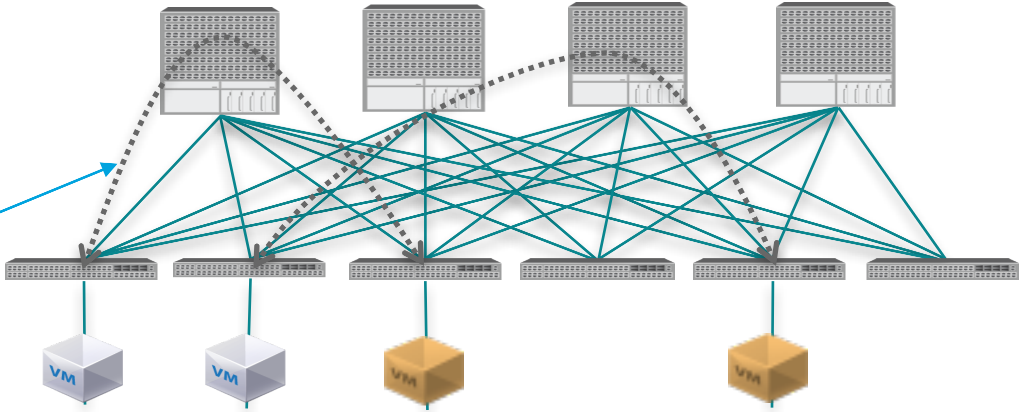
Per Packet Variations



Accumulated Flow Information (Volume…)

Latency Measurements
LSE

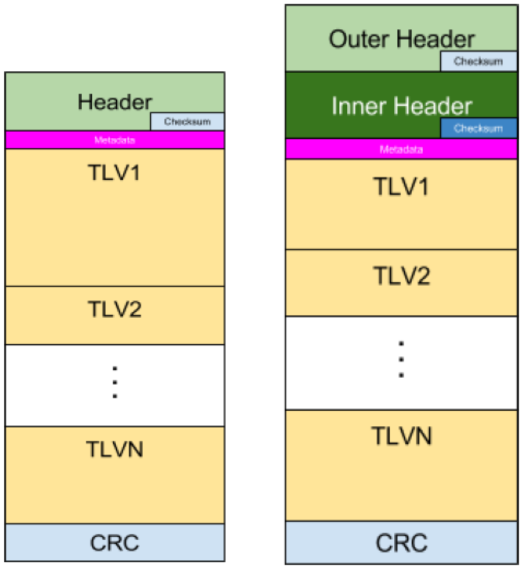
- PTP based network latency metrics will be supported with ACI 2.3 release (Q2CY17)
    - Per Port Average Latency & Variance
    - Per Port 99% Latency (99% of all packets have recorded latency less than this value)
- Supported with Multi-Pod
- Two modes with different degrees of granularity
    - 10 msec end to end latency
    - 100 usec end to end latency



Leverages 8 bytes of T-Tag (same time stamp format leveraged by Nexus 3500 used in HFT envrionments)

Real-time Hardware Telemetry ASE-4 & LSE-2

- Streams ASIC-level statistics to one or more collectors
- User defines which statistics, how often, and to which collector(s) using which encapsulation(s), should be streamed
    - Could provide predefined frequency and 'statistics sets' (interface stats, ACL stats, etc.) – user just specifies collector
    - Or, all configuration options can be exposed to end user (JSON-type definition file) – assumes we publish full list of supported statistics

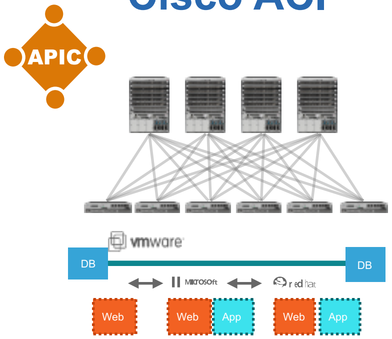


Agenda

- What’s New
- Next Generation Capabilities
  - Forwarding – Packet Walks
  - Forwarding – Protocol Support
  - Forwarding - Table Templates
  - Telemetry
  - Encryption (MACSEC and CloudSEC)
  - QoS & Buffering
- Design Impacts of 25G, 50G and 100G
- Next Gen Nexus 9000 Switch Platforms

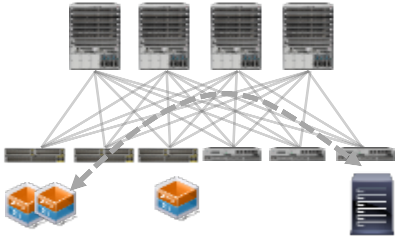
ACI and Nexus 9k Standalone
MACSEC PHY HW Encryption Capability

Cisco ACI



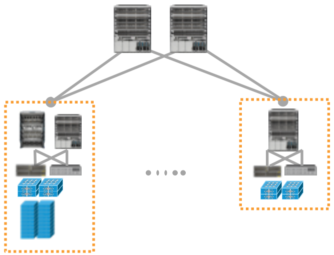
1 Encryption for ACI

Programmable Fabric



2 Encryption for Programmable Fabric

Programmable Network



3 Encryption for Programmable Network

Goal: Solve Encryption for All 3 Usecases

ACI and Nexus 9k Standalone
MACSEC PHY HW Encryption Capability

Q2/Q3 –CY17

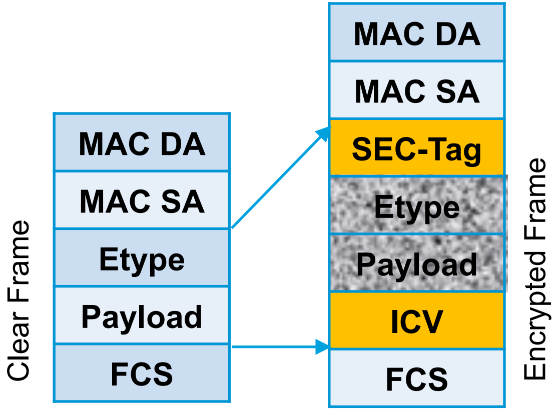
- Breakout for 100GE/40GE/10GE
- IEEE Compliant 802.1AE,bn,bw,cg
- Security Ciphers Suites:
  - GCM-AES-128 (32-bit PN)
  - GCM--AES-256 (32-bit PN)
  - GCM-AES-128-XPN (64-bit PN)
  - GCM-AES-256-XPN (64-bit PN)
- Key Exchange Protocol
  - MKA (IEEE Standard)
  - PSKs
- FIPS 140-2 Certified (Planning)

- MACSEC HW Capability:
  - Link MACSEC (Underlay)
  - INS-SEC (ie. VXLAN Overlay Encryption)
  - ClearTag (MPLS, Segment Routing, VPN, EVPN support, L2 encap)

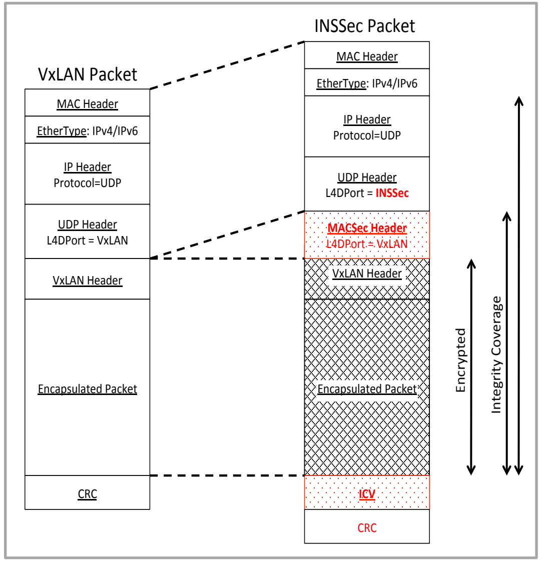
- 128 Security Associations per 100G
- Man in Middle Attack protection per SA replay window checking over the full range
- ECC protection on all memories

MACSEC Frame Format for:
Link, VXLAN and ClearTag Encryption

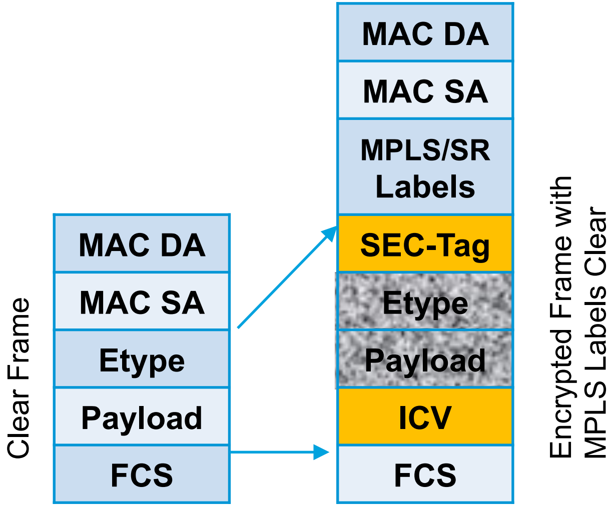
MACSEC Link Encryption



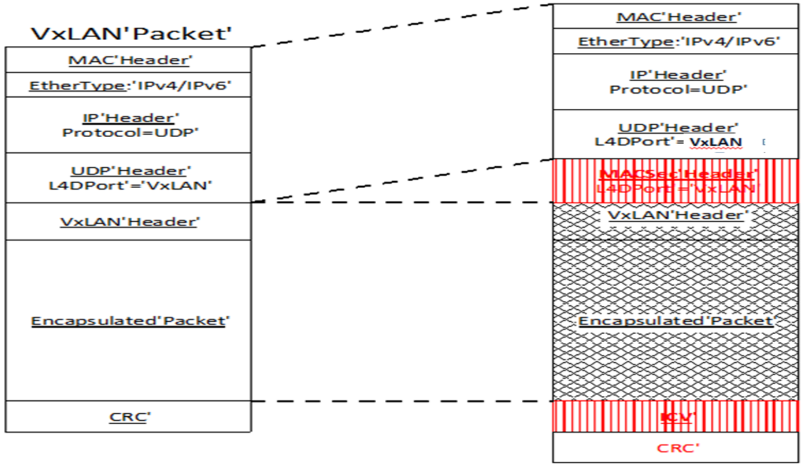
Overlay VXLAN Encryption

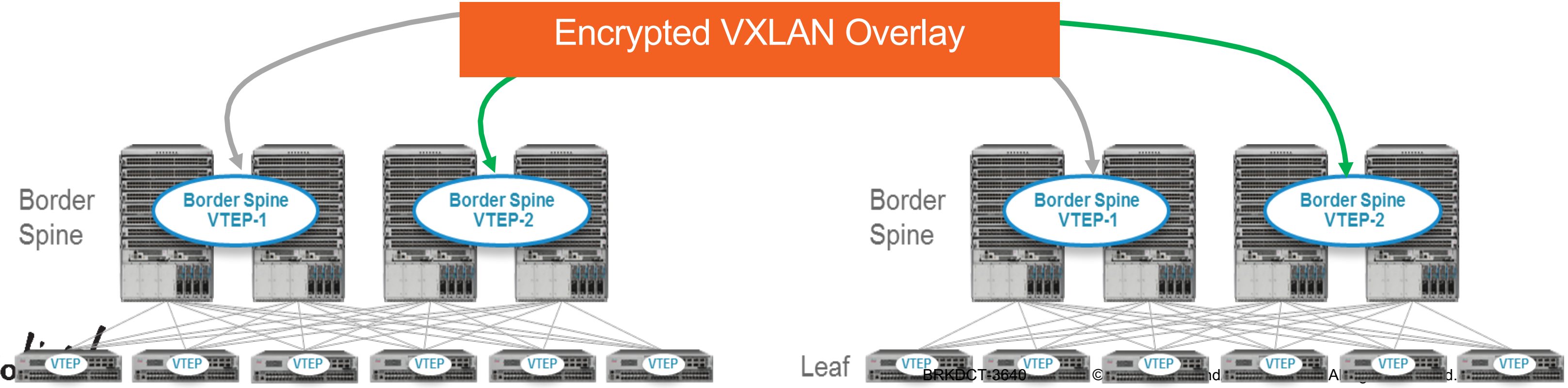


Transport ClearTag Encryption

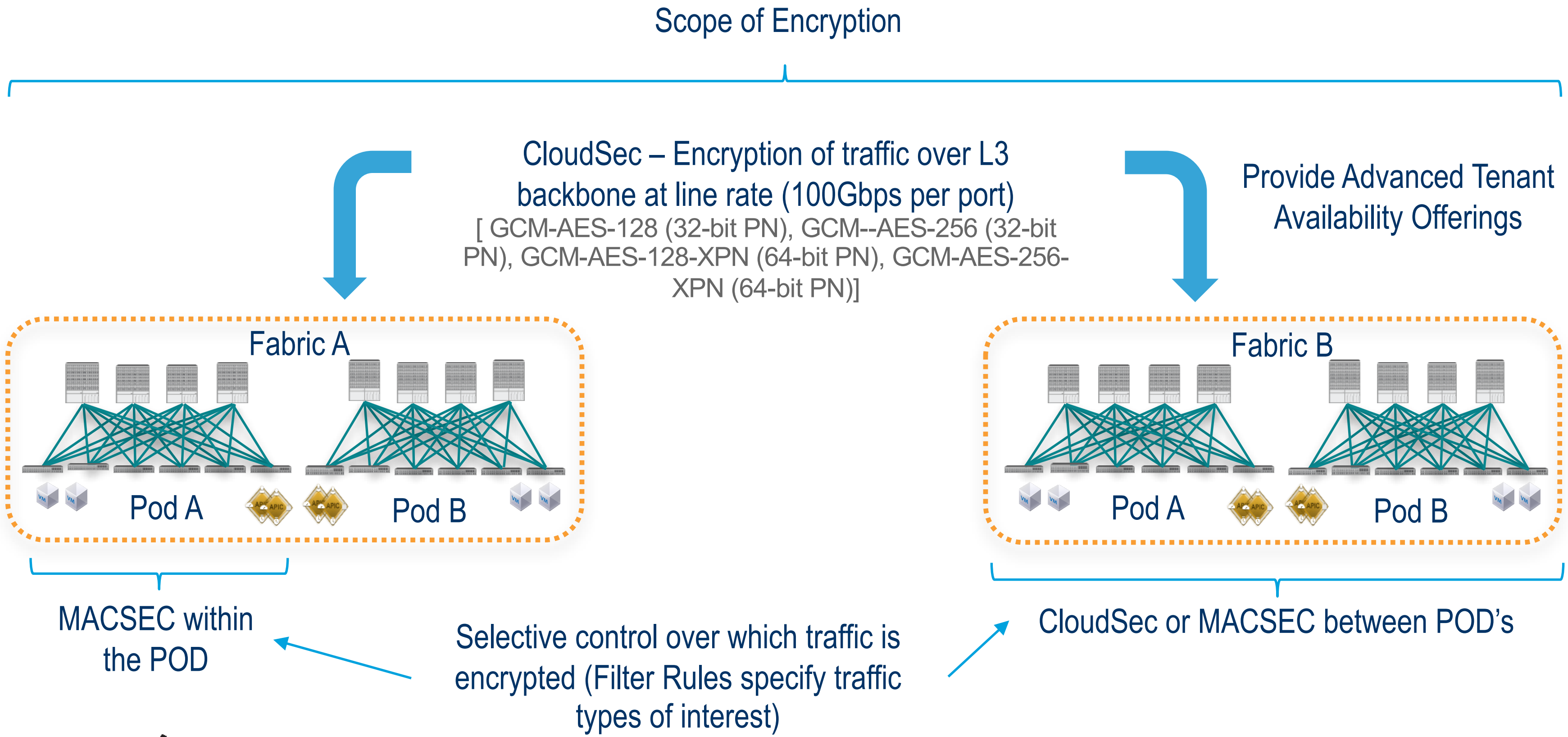


VXLAN Encryption
(Nexus 9k Standalone)





ACI End to End Encryption
Secure Communications for all traffic



Scope of Encryption

CloudSec – Encryption of traffic over L3 backbone at line rate (100Gbps per port)
[GCM-AES-128 (32-bit PN), GCM--AES-256 (32-bit PN), GCM-AES-128-XPN (64-bit PN), GCM-AES-256-XPN (64-bit PN)]

Provide Advanced Tenant Availability Offerings

MACSEC within the POD

CloudSec or MACSEC between POD's

Selective control over which traffic is encrypted (Filter Rules specify traffic types of interest)

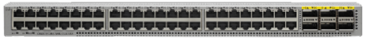
Agenda

- What’s New
- Next Generation Capabilities
  - Forwarding – Packet Walks
  - Forwarding – Protocol Support
  - Forwarding - Table Templates
  - Telemetry
  - Encryption (MACSEC and CloudSEC)
  - QoS & Buffering
- Design Impacts of 25G, 50G and 100G
- Next Gen Nexus 9000 Switch Platforms

Nexus 9000 QoS and Buffering
Shared Memory & Egress Queuing

Cat4900 – Shared Memory Egress buffering

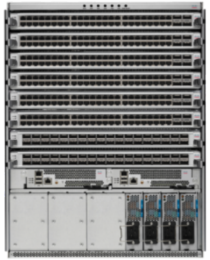
Nexus 5x00 – VoQ Ingress Buffering



Nexus 9200/9300 Shared Memory Egress buffering

Cat6500 – Egress Buffering

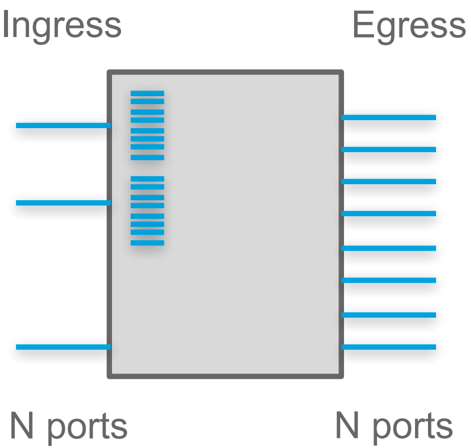
Nexus 7x00 – VoQ Ingress Buffering



Nexus 9000 Egress Buffering

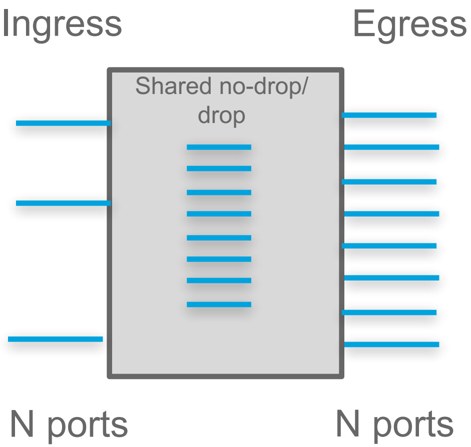
Nexus 9000 QoS and Buffering VoQ vs. Output Queue Design

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

Nexus 9000 QoS and Buffering

NX-OS 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/ no drop class

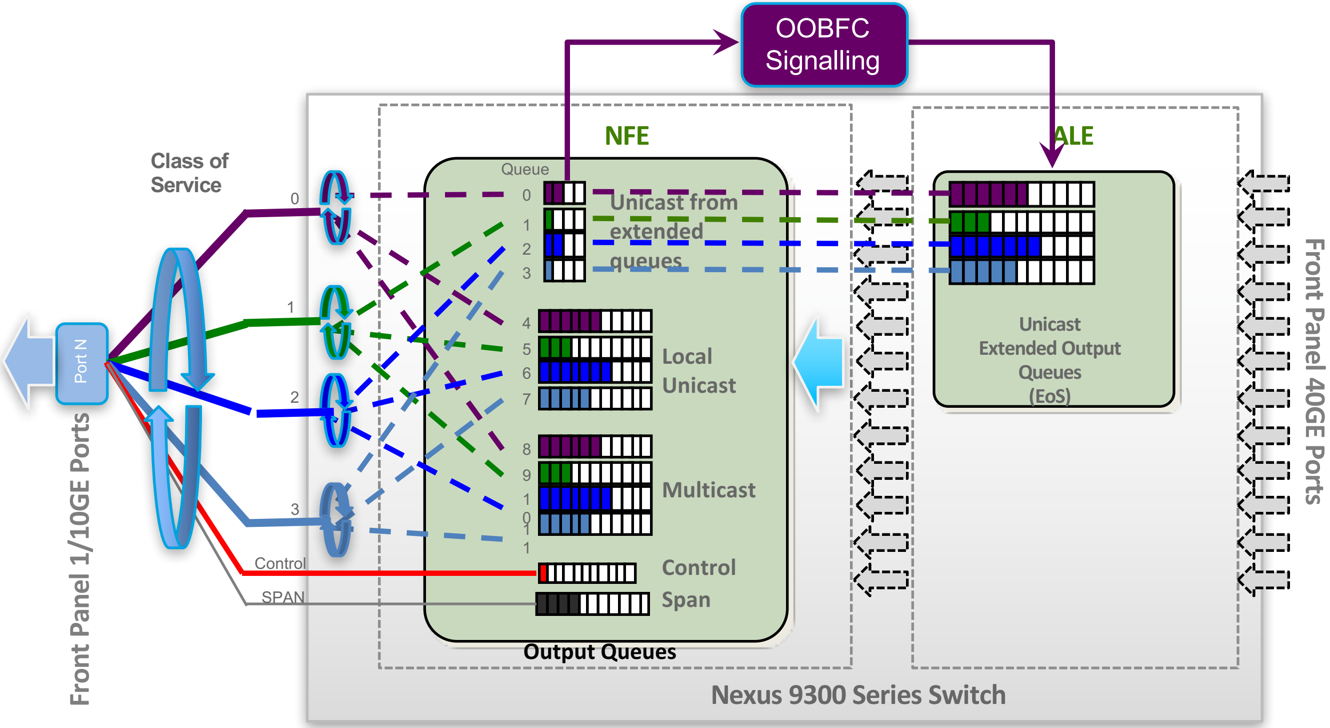
- **Egress Queuing and Shaping**
  - Policy-map type queueing
  - 8 user-defined classes based on qos-group (8 unicast + 8 multicast)
  - 1 control class for CPU and 1 class for SPAN traffic
  - 7 no-drop classes

End-to-End QoS Implementation and Operation with Cisco Nexus Switches

Session ID: BRKDCT-3346

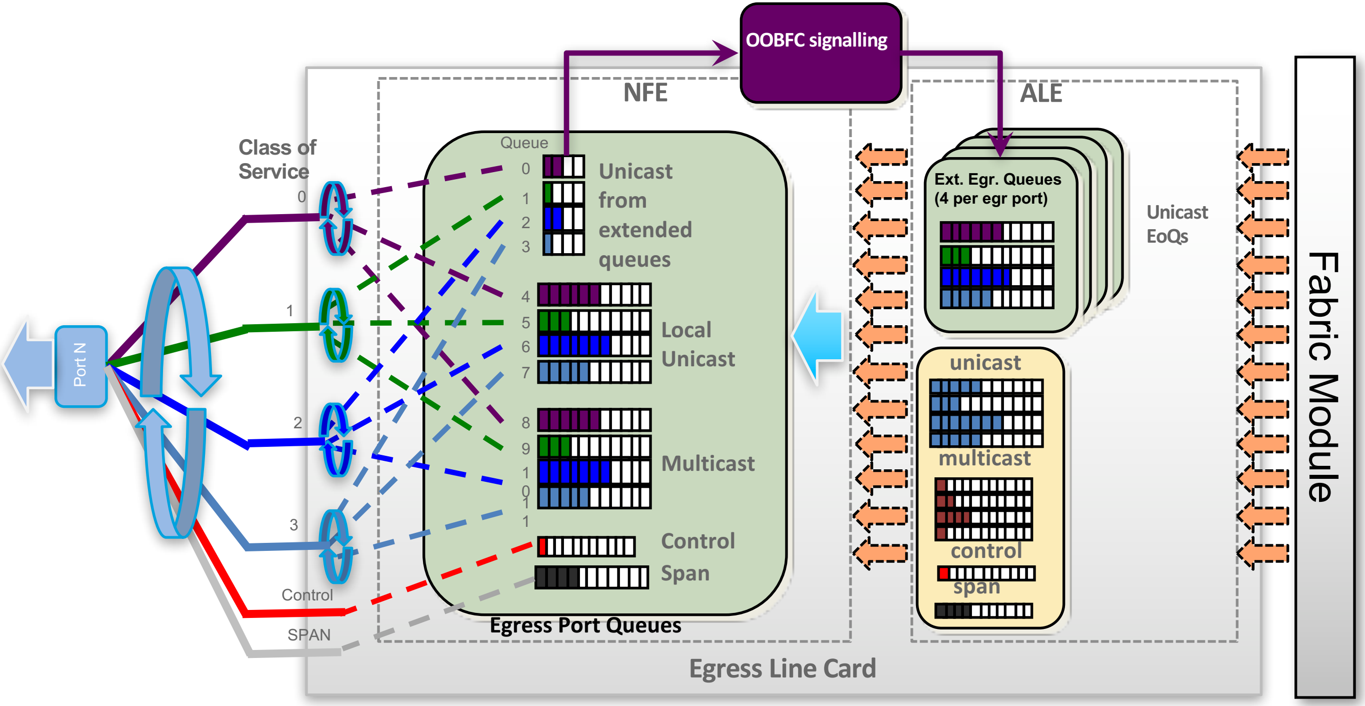
Queuing & Scheduling on First Gen Nexus 9300 Switches

4 Unicast + 4 Multicast + 2 Services Queues per Port

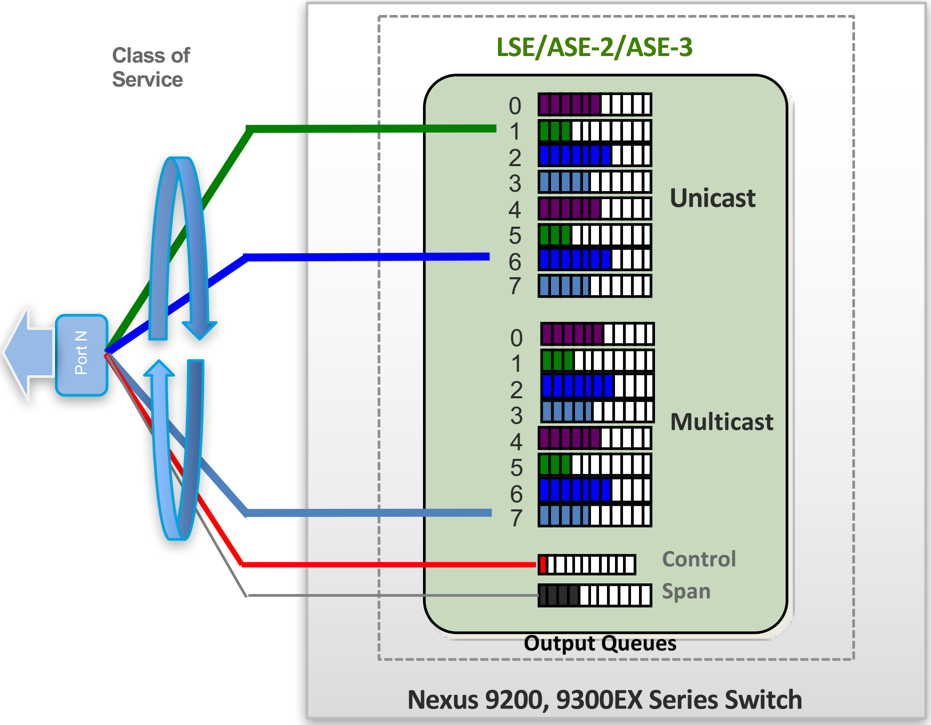


Queuing & Scheduling on First Gen Nexus 9000 Switches

4 Unicast + 4 Multicast + 2 Services Queues per Port



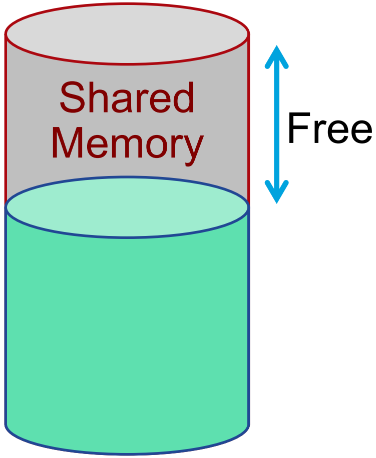
Queuing & Scheduling on 2nd Gen Nexus 9000 Switches
8 Unicast + 8 Multicast + 2 Services Queues per Port



- For each port up to 18 distinct queues could be scheduled
  - CPU queue
  - 8 unicast queue
  - 8 multicast queue
  - SPAN queue
- The CPU queue has strict priority
- The SPAN queue is best effort and lowest priority
- The scheduling between the 16 user queues is configurable
- By default the selection between unicast and multicast is 50-50 DWRR in each group and then among the groups based on DWRR with each group receiving 12.5 %
- Any number of queues or groups could be strict priority (SP), among SP groups the lowest queue number wins

Shared Memory Buffering
Dynamic Buffer Protection (DBP)

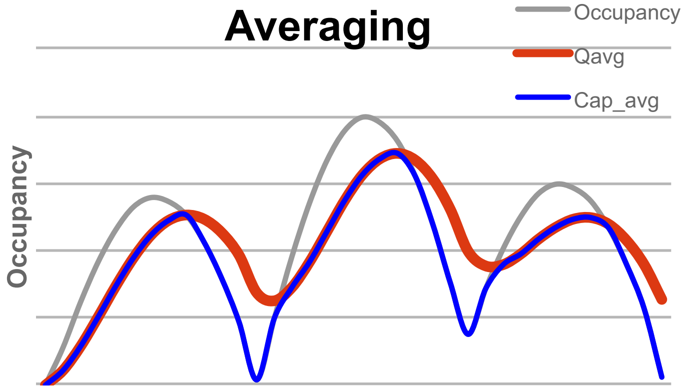
- **Requirement**
  - In a shared memory switch it is necessary to prevent any output queue from taking more than its faire share of the buffer when its output is oversubscribed
  - It can take more than its fair share to handle burst if the output is not oversubscribed.
- **Basic Algorithm (Deployed on Merchant and First Gen Nexus 9000)**
  - The algorithm defines a dynamic max threshold for each queues sharing the same buffer, if the queue length is less than threshold packets are accepted otherwise packet are discarded
  - The dynamic threshold is calculated by multiplying the amount of free memory by a parameter Alpha
- **Enhanced Algorithm (Deployed on 2nd Generation Nexus 9000)**
  - The algorithm is expanded to include the concept of pool (class of service) and it is also adapted to multicast traffic.
  - The dynamic buffer algorithm is extended to allocate memory among buffer pools then to allocate among queues within each pool

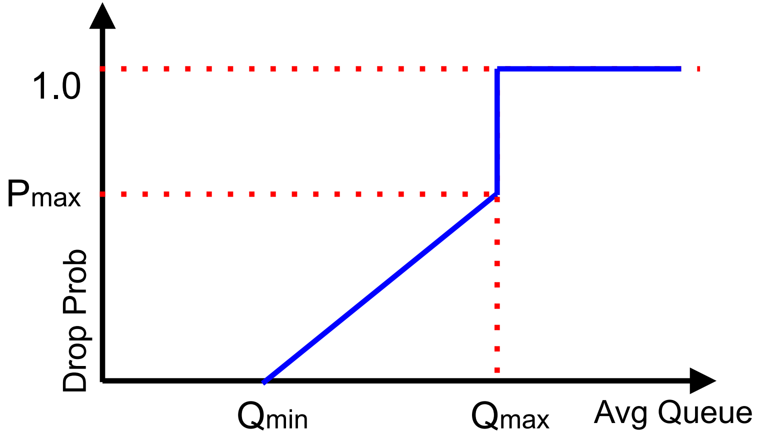


Nexus 9000 QoS and Buffering

Active Queue Management (AQM)

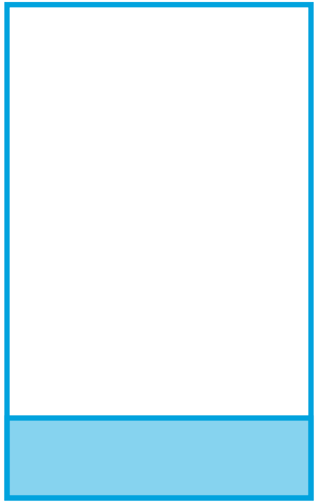
- AQM
  - Mode and parameters defined by profiles mapped to queues
  - Averaging timer per profile
  - Drop/ECN-mark per profile
- WRED
  - Each queue mapped to a profile
  - Averaging with Cap_Avg
- AFD
  - Drop/mark only elephant flows
  - Arrival rate measured by ETRAP
  - “Fair” rate computed using a continuous feedback loop
- ECN
  - Mark/drop ECN Capable flows
  - Ignore/drop non-ECN capable flows





Buffering Data Centre
Two Requirements for Buffers

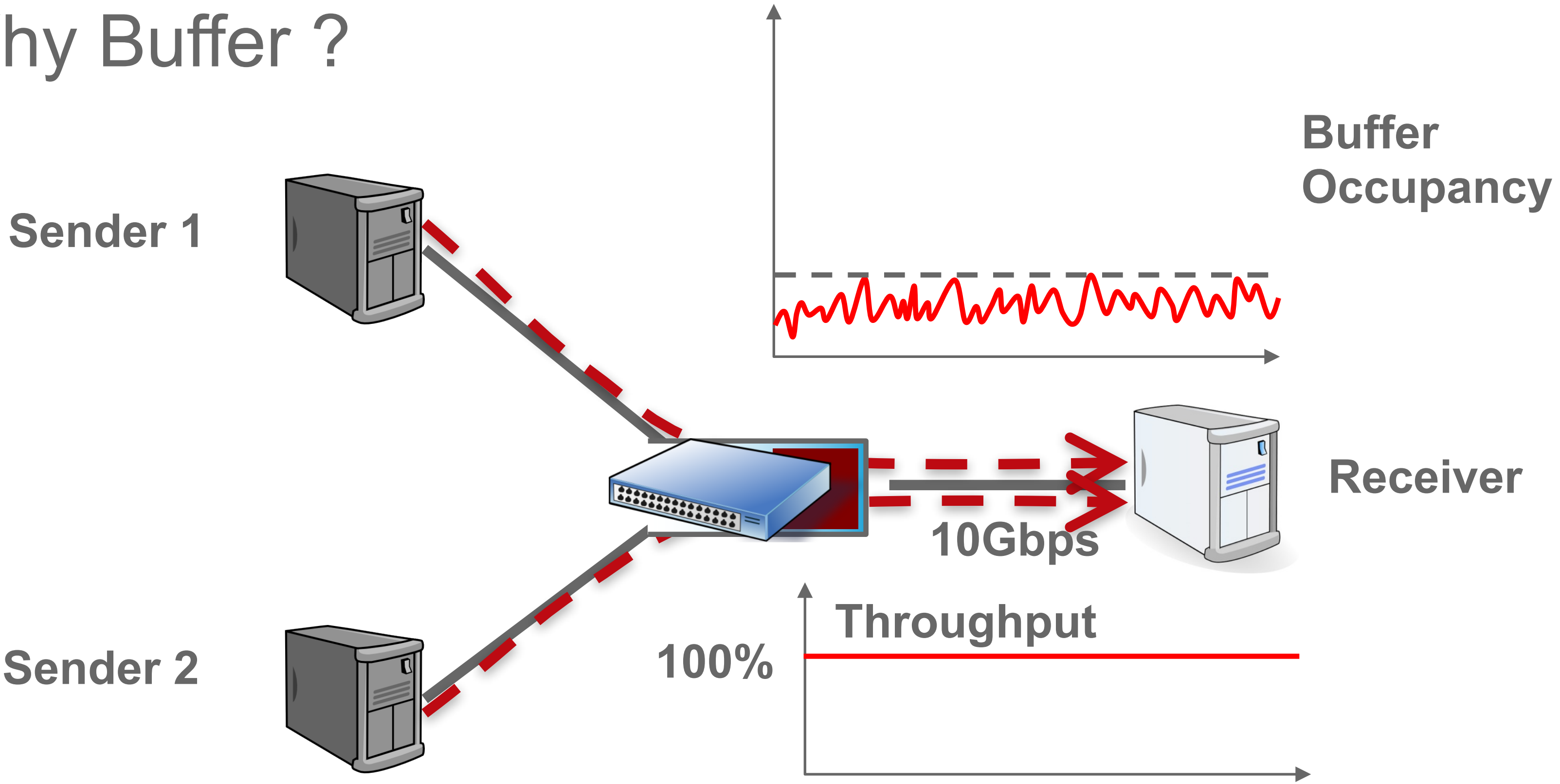
- Long Lived TCP flows
  - Maximise the utilisation of the available network capacity (ensure links are able to run at line rate)
  - Window Size Increases to probe the capacity of the network
  - Delay x Bandwidth Product (C x RTT)*
    - e.g if your network had 100 msec of latency with 10G interface, 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 Incast Burst

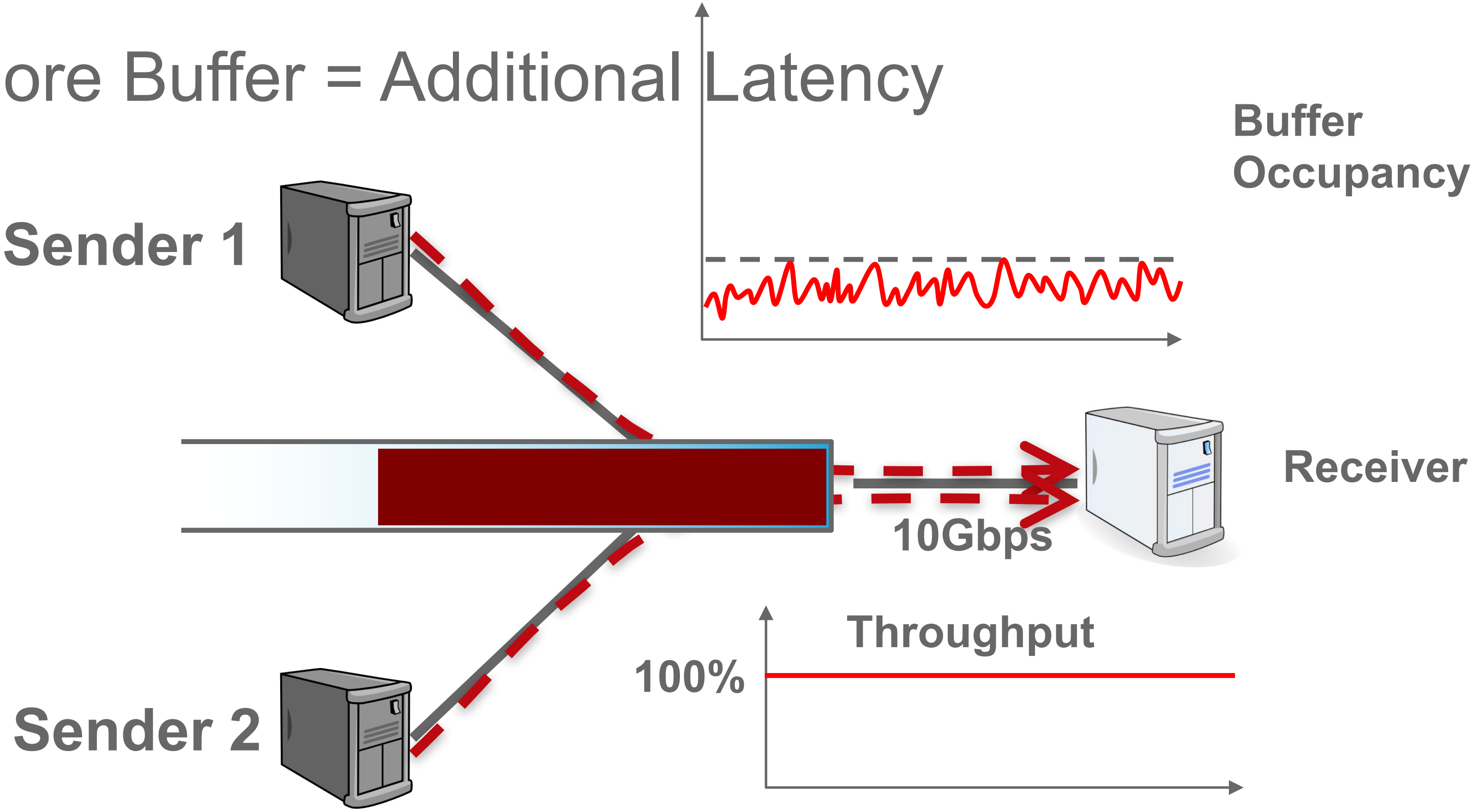
Buffer Required for Maximising Network Utilisation

Why Buffer ?



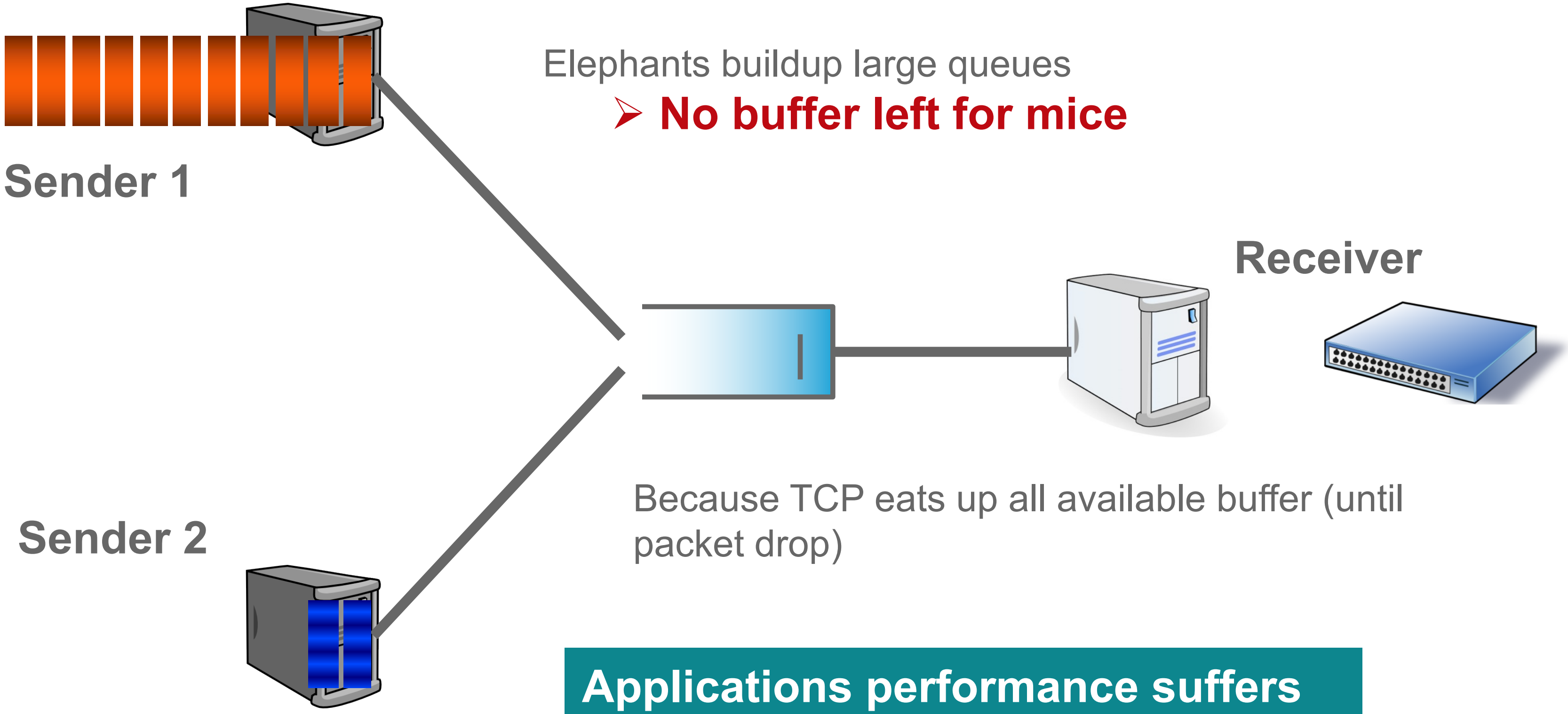
Buffer Occ > 0 ➔ Throughput = 100%

More Buffer = Additional Latency



Application does not go faster

Elephants Waste Buffer



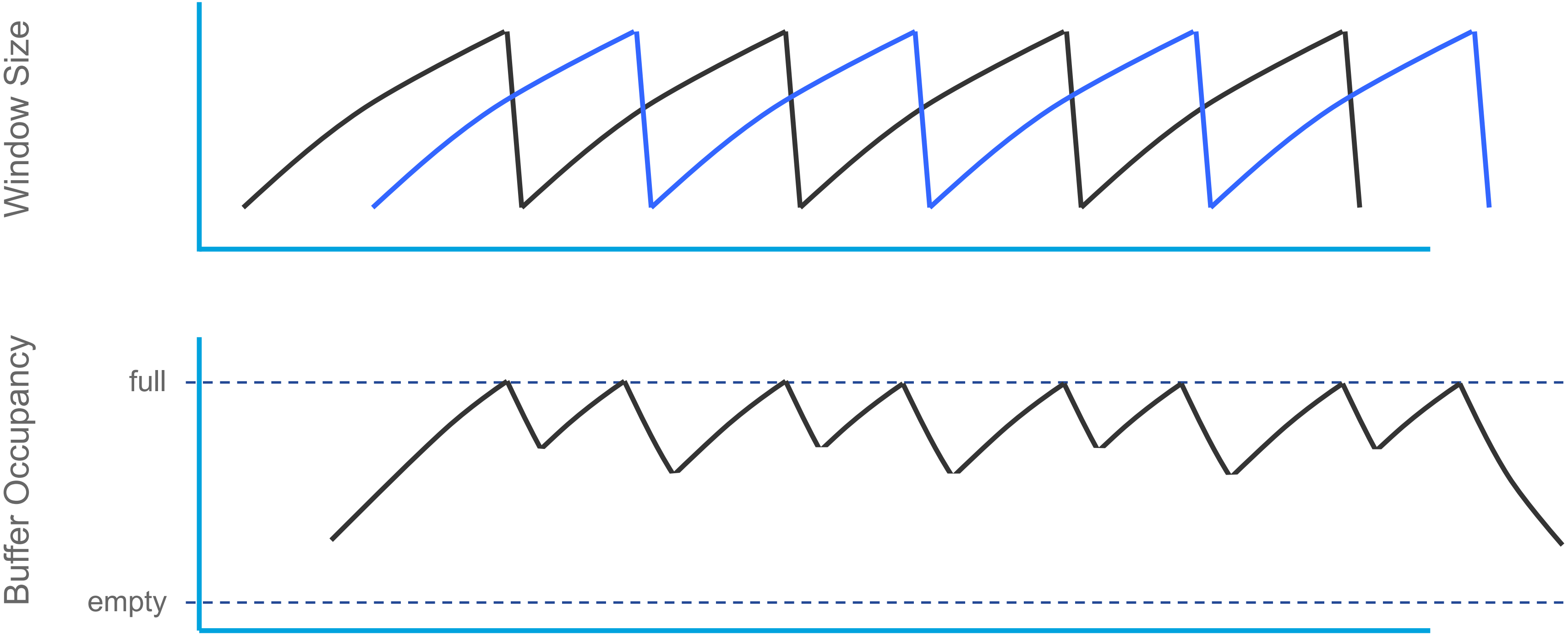
Elephants buildup large queues

- ➤ **No buffer left for mice**

Because TCP eats up all available buffer (until packet drop)

Applications performance suffers

Multiple TCP Flows in Reality



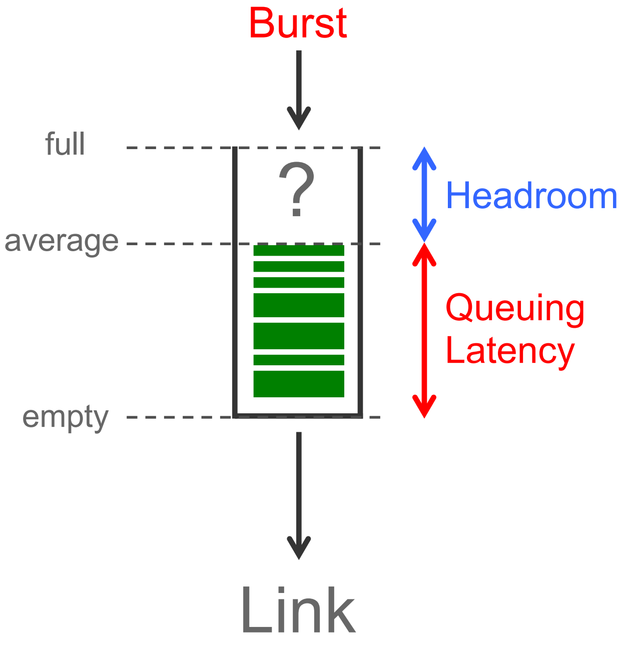
Long Lived TCP Flows
TCP Congestion Control and Buffer Requirements

- 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 $\dfrac{C \times RTT}{\sqrt{n}}$ instead of $C \times RTT$

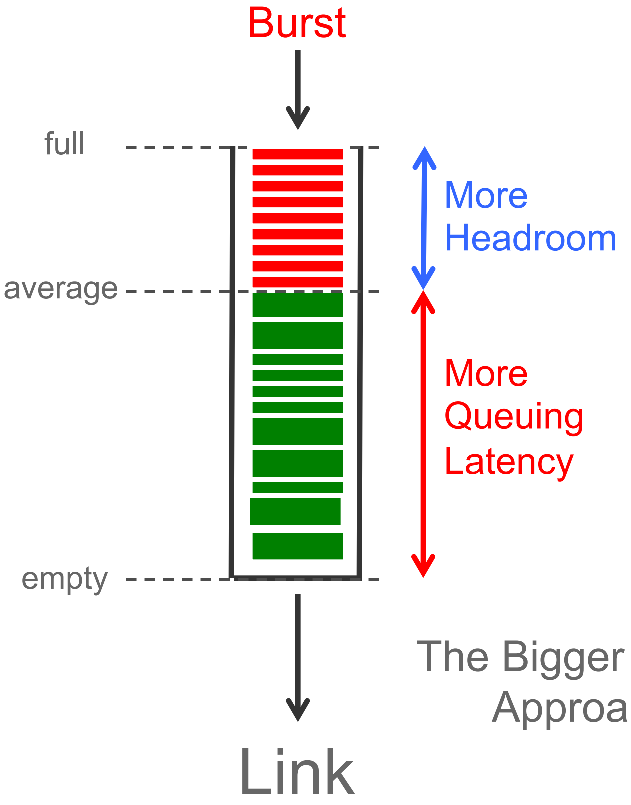
- Proven by theory and experiments in real operational networks
- For example, see Beheshti et al. 2008: "Experimental Study of Router Buffer Sizing"

Micro-bursts Need Headroom
Where Does it Come From?



Micro-bursts Need Headroom
Where Does it Come From?

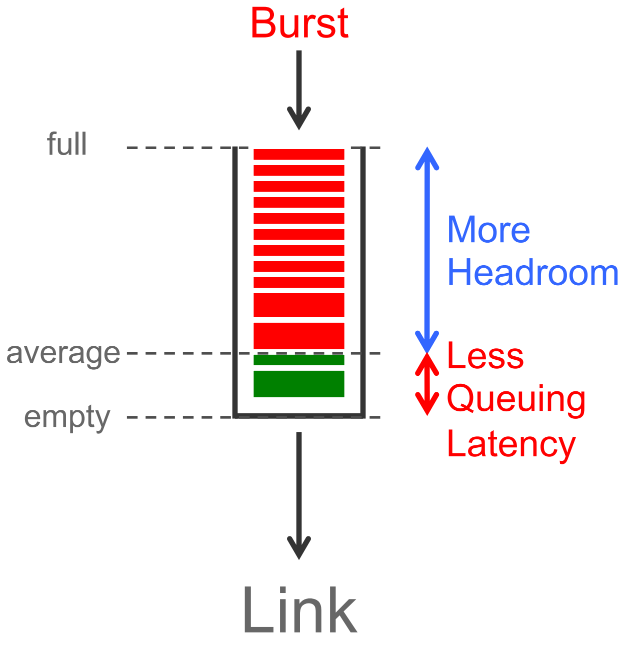
- Larger Buffer can increase the burst headroom but
  - Increases queuing latency which decreases application performance
- You can still have large flows fill up the entire buffer resulting in no increase in burst headroom
  - Impacts application performance



The Bigger Buffer Approach

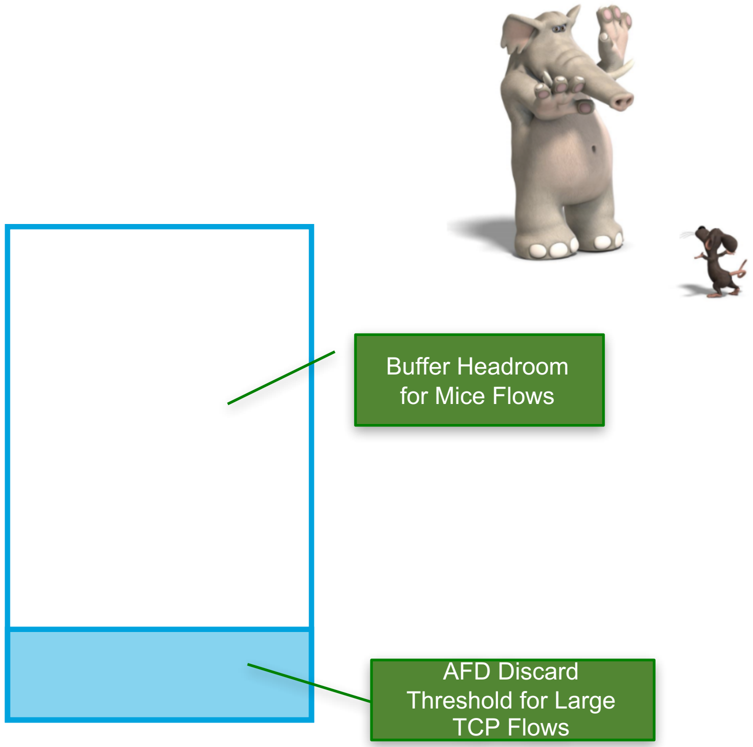
We Want the Best of Both Worlds

- Maximise the amount of buffer always available for bursts
- Minimise the latency for high throughput flows
- Better application performance for both types of traffic

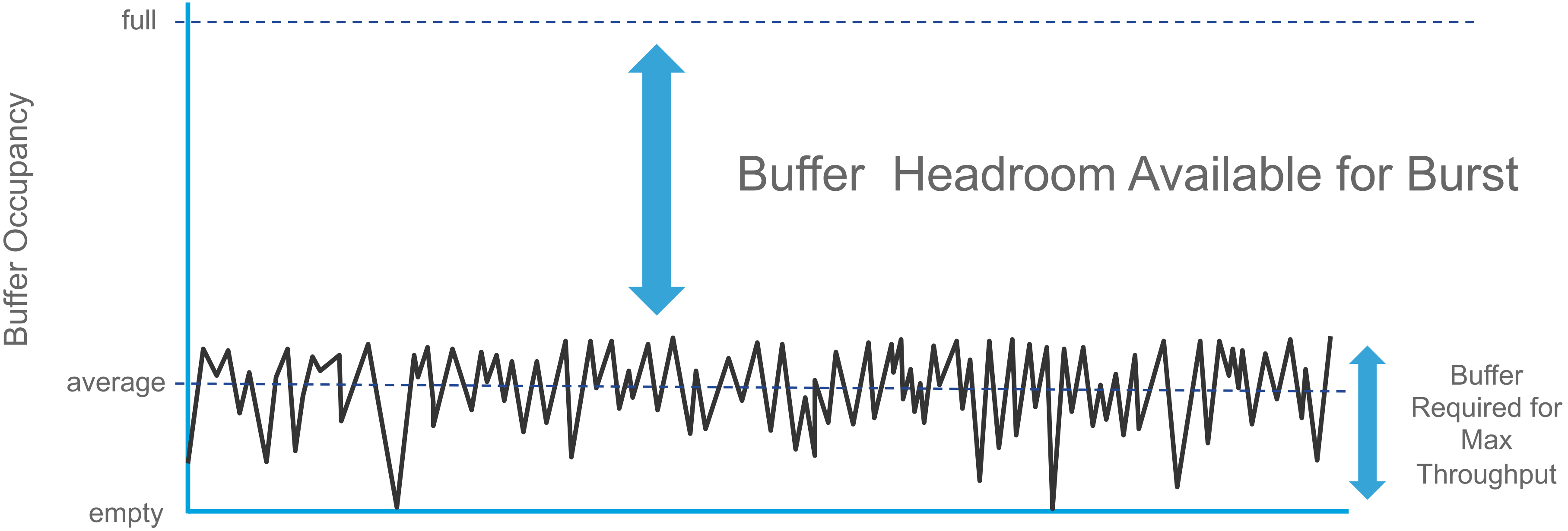


Innovation Gives us the Best of Both Worlds
AFD & DPP

- 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 Headroom for Mice Flows
AFD Discard Threshold for Large TCP Flows

AFD Increases Headroom, Reduces Latency

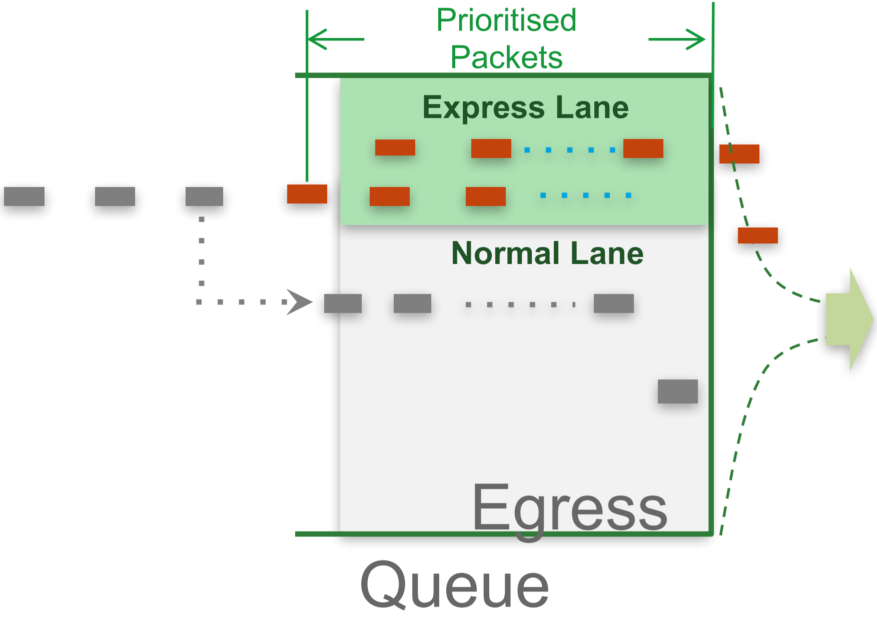


Buffer Headroom Available for Burst

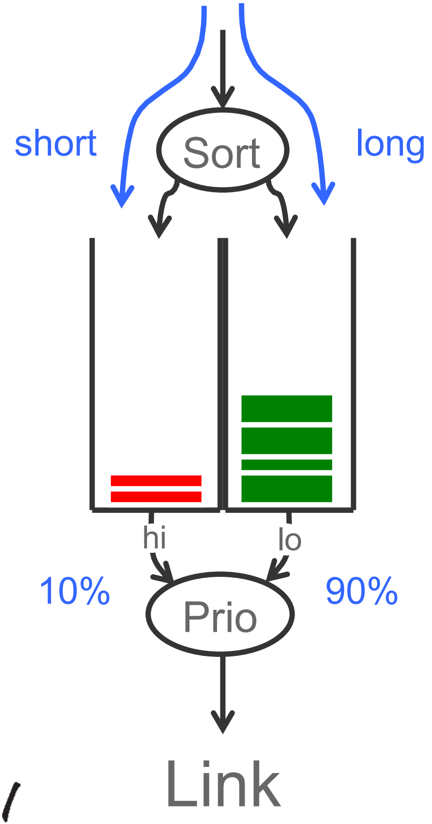
Buffer Required for Max Throughput

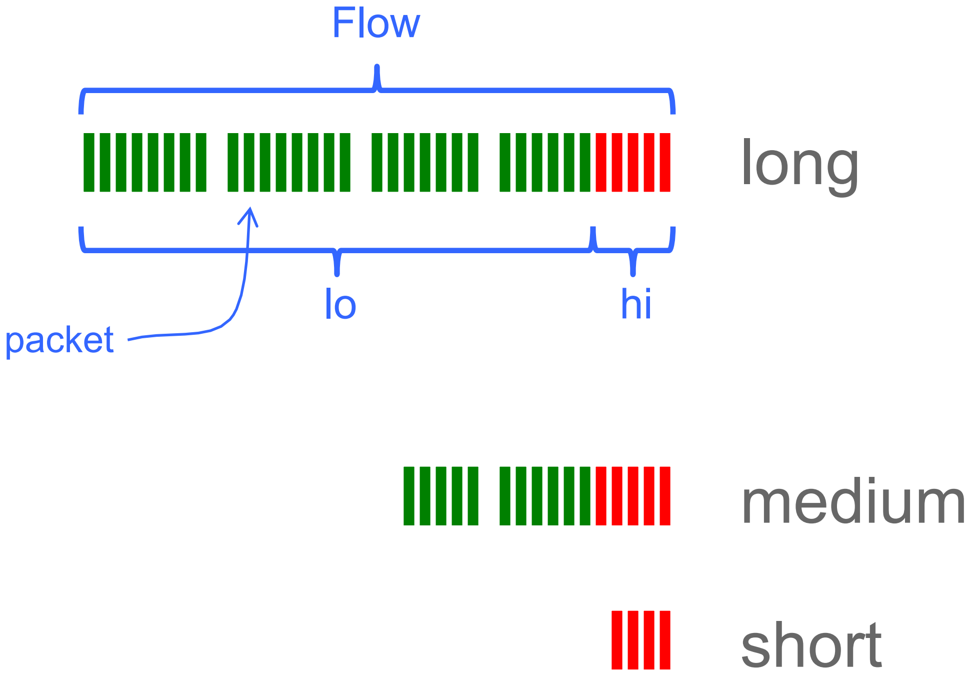
DPP (Dynamic Packet Prioritisation)

- Separate flows into short and long
- Put short flows in high priority queue
- Put long flows in low priority queue
- The 10% of bytes that are in short flows means high priority queue will be empty
- Prioritisation guarantees packet order
- We want to prevent the drops of the mice, the incast and burst traffic



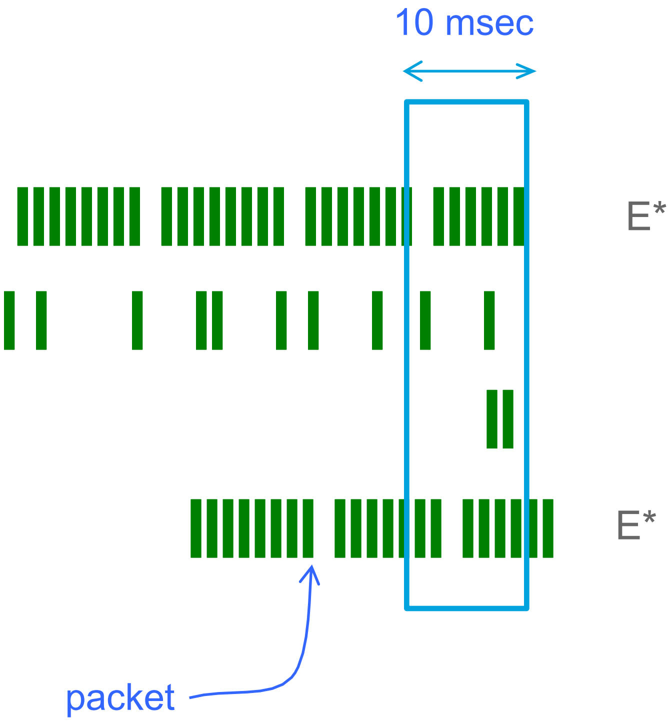
All Flows are Short Until They Become Long





Elephant Trap

- Mechanism to identify large volume flows
  - Identified based on 5-tuple
- Elephant trap threshold is byte-count-based.
  - When received packets in a flow exceeds the number of bytes specified by the threshold, the flow is considered an elephant flow
  - Only elephant flows are submitted to AFD dropping algorithm. Mice flows are protected and not subject to AFD dropping
  - Arriving data rate is measured on the ingress, and compared against a calculated fair rate on the egress port to decide dropping capability

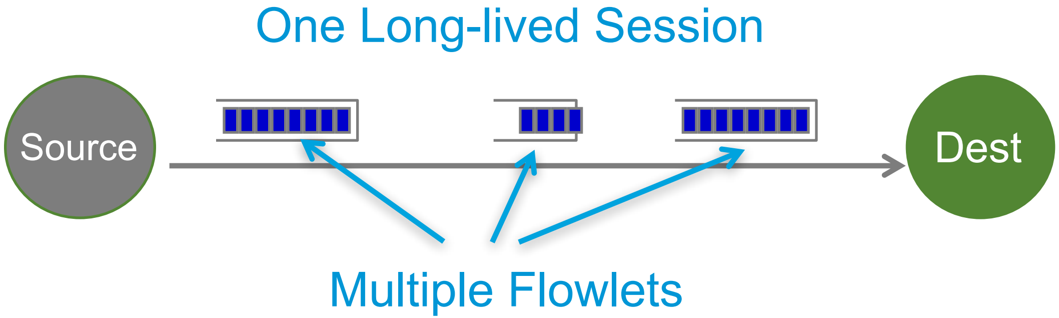


E* == Elephant Flow

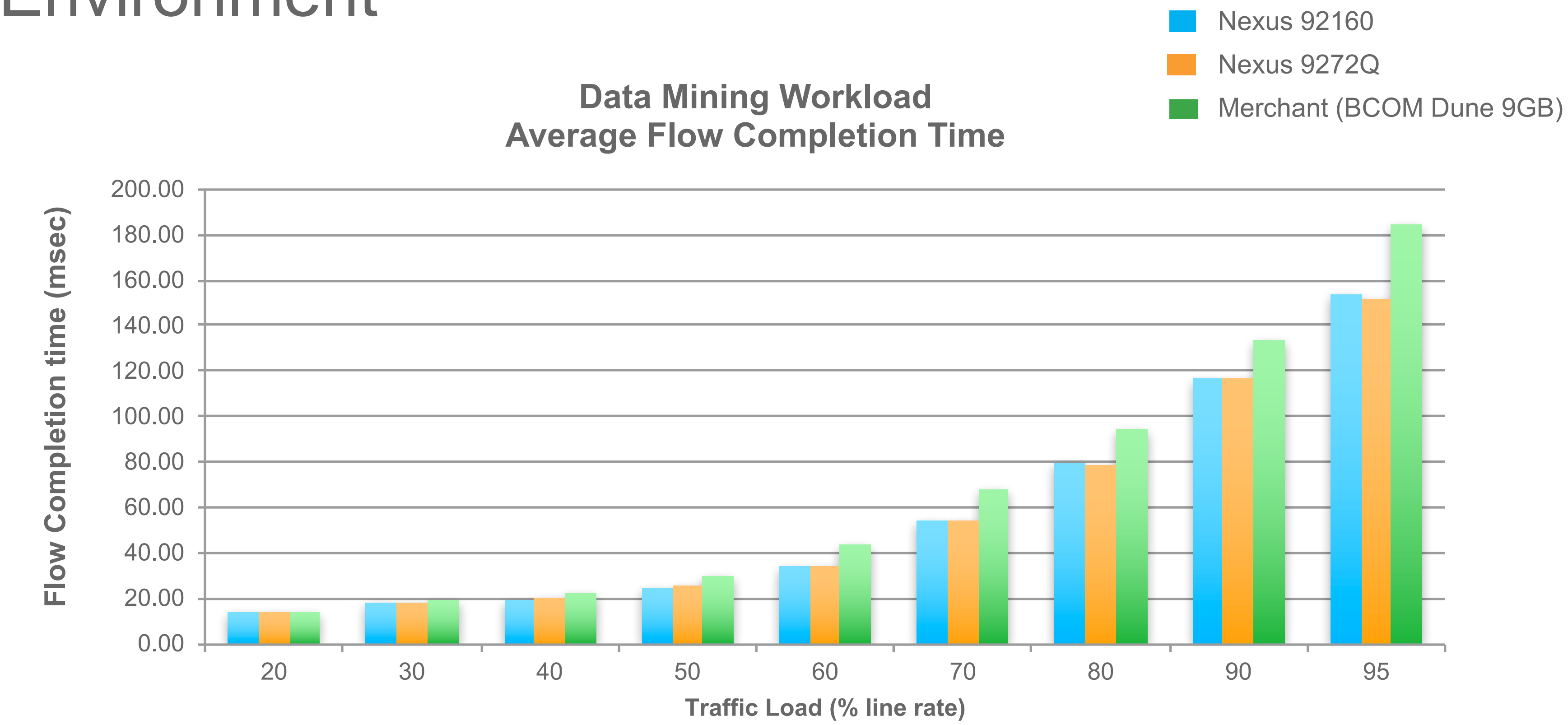
DPP Looks for Any Burst TCP, UDP, Multicast, ..

A Long-lived TCP Session ≠ An Elephant Flow

- The elephant trap and DPP algorithm are **not** tracking only TCP sessions
- The algorithm is 5-tuple based which means it can find TCP, UDP, Unicast and Multicast bursts
  - A very long lived session that is quiet and then bursts will be prioritised for that burst
  - Traffic arriving due to a link failure will be prioritised, etc …

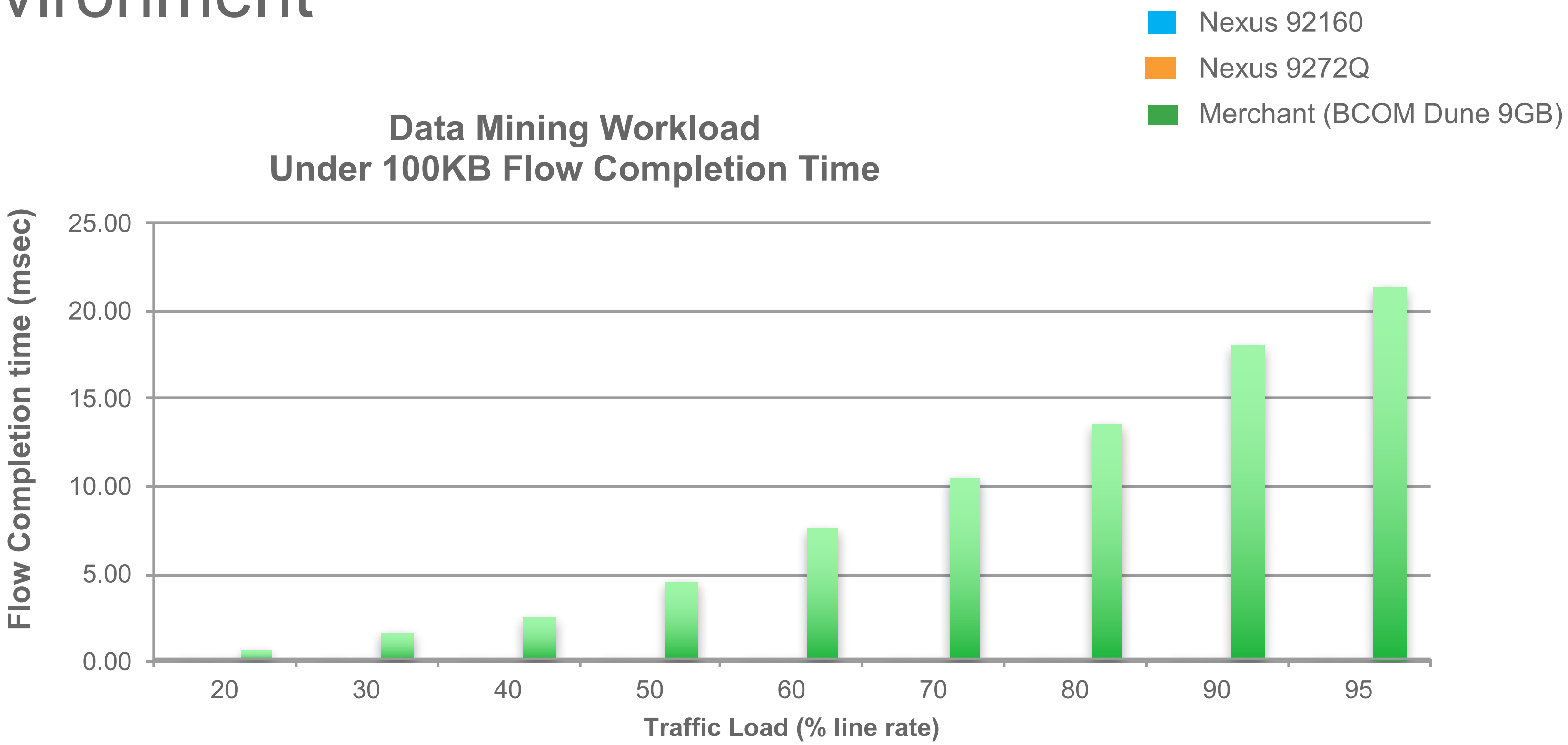


Better Application Performance in an Incast Environment



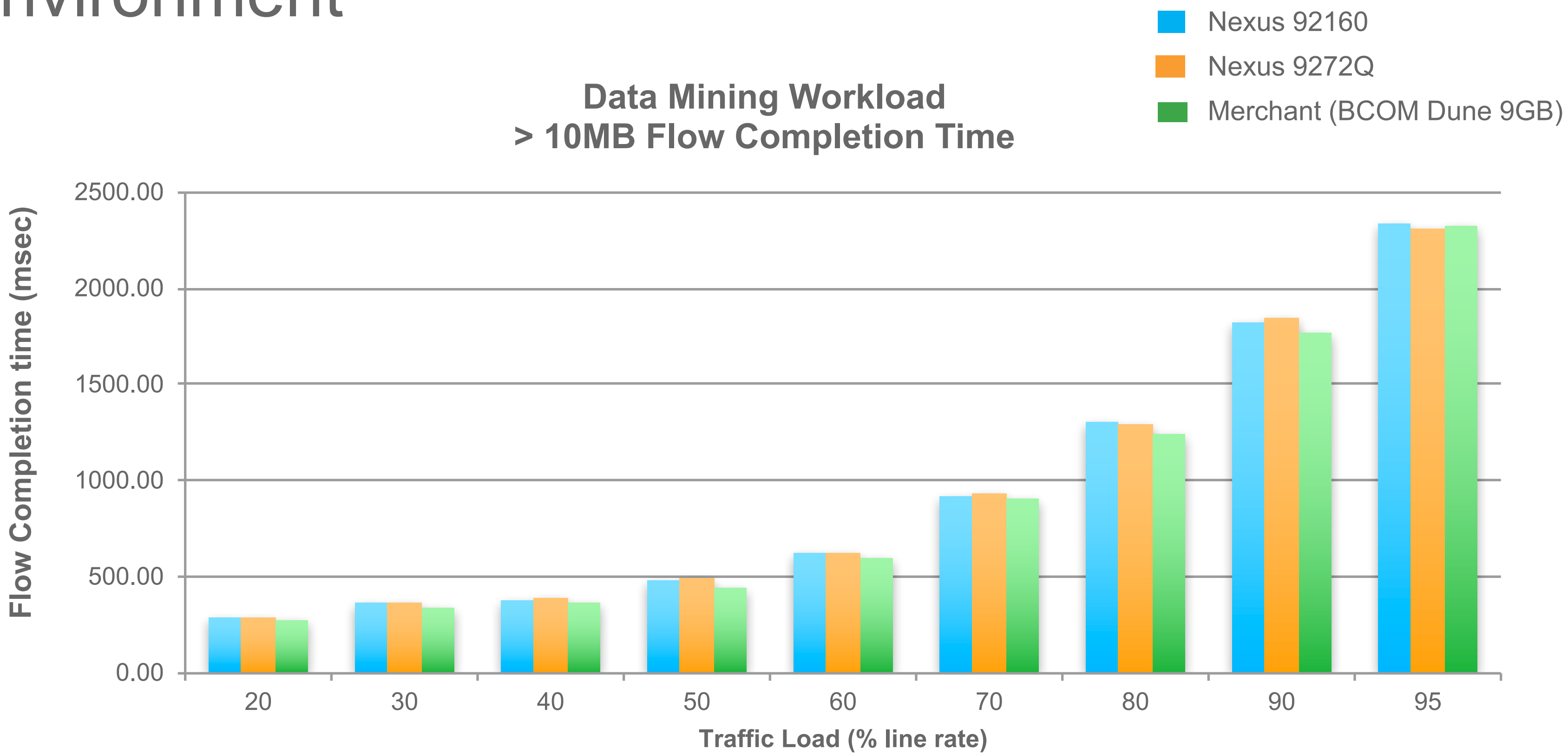
http://miercom.com/cisco-systems-speeding-applications-in-data-Centre-networks/

Better Application Performance in an Incast Environment



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

Better Application Performance in an Incast Environment



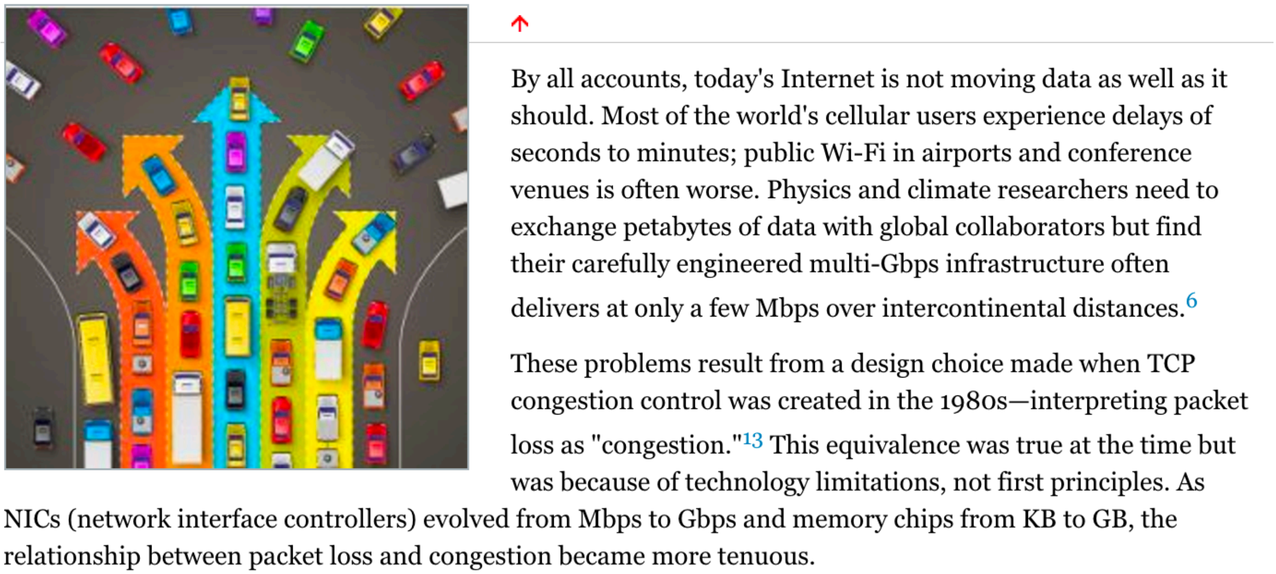
http://miercom.com/cisco-systems-speeding-applications-in-data-Centre-networks/

So Why Are You Now Shipping a Big Buffer Switch?

- There are a few specific environments, synchronised UDP bursts, where latency and throughput are much less important than loss, e.g. need to ensure that every trading customer gets the pricing update
- Some customers just want it
- The majority of the Internet and Cloud environments are moving beyond the older requirements

BBR: Congestion-Based Congestion Control

Latest Google Research into TCP Congestion Control

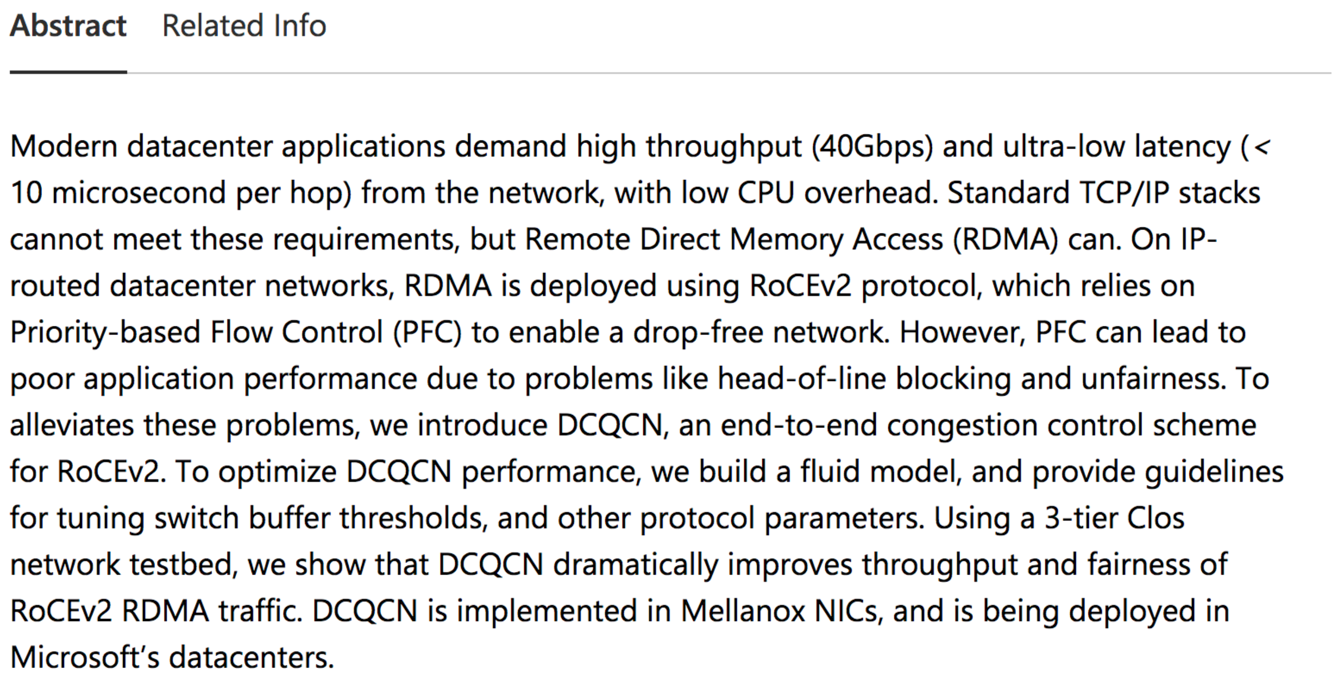


By all accounts, today's Internet is not moving data as well as it should. Most of the world's cellular users experience delays of seconds to minutes; public Wi-Fi in airports and conference venues is often worse. Physics and climate researchers need to exchange petabytes of data with global collaborators but find their carefully engineered multi-Gbps infrastructure often delivers at only a few Mbps over intercontinental distances.[6]

These problems result from a design choice made when TCP congestion control was created in the 1980s—interpreting packet loss as "congestion."[13] This equivalence was true at the time but was because of technology limitations, not first principles. As NICs (network interface controllers) evolved from Mbps to Gbps and memory chips from KB to GB, the relationship between packet loss and congestion became more tenuous.

http://cacm.acm.org/magazines/2017/2/212428-bbr-congestion-based-congestion-control/fulltext

Congestion Control for Large-Scale RDMA Deployments



Abstract Related Info

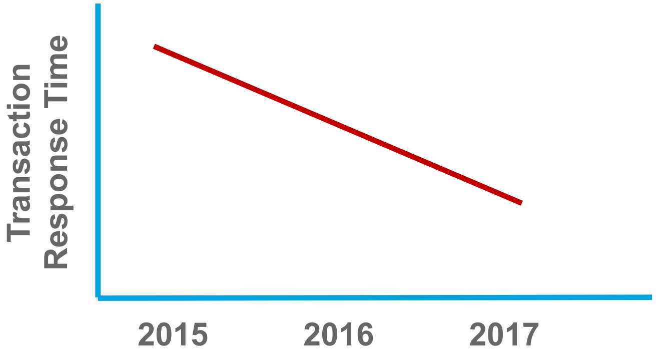
Modern datacenter applications demand high throughput (40Gbps) and ultra-low latency (< 10 microsecond per hop) from the network, with low CPU overhead. Standard TCP/IP stacks cannot meet these requirements, but Remote Direct Memory Access (RDMA) can. On IP-routed datacenter networks, RDMA is deployed using RoCEv2 protocol, which relies on Priority-based Flow Control (PFC) to enable a drop-free network. However, PFC can lead to poor application performance due to problems like head-of-line blocking and unfairness. To alleviates these problems, we introduce DCQCN, an end-to-end congestion control scheme for RoCEv2. To optimize DCQCN performance, we build a fluid model, and provide guidelines for tuning switch buffer thresholds, and other protocol parameters. Using a 3-tier Clos network testbed, we show that DCQCN dramatically improves throughput and fairness of RoCEv2 RDMA traffic. DCQCN is implemented in Mellanox NICs, and is being deployed in Microsoft's datacenters.

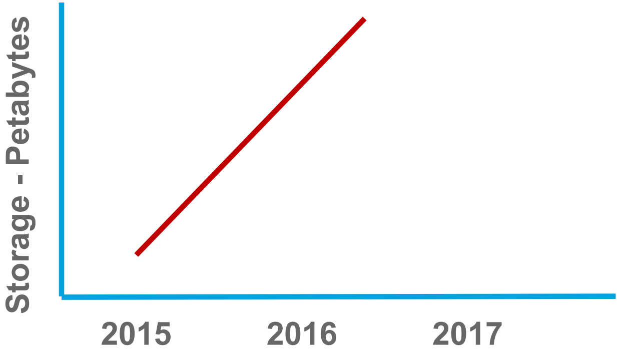
https://www.microsoft.com/en-us/research/publication/congestion-control-for-large-scale-rdma-deployments/

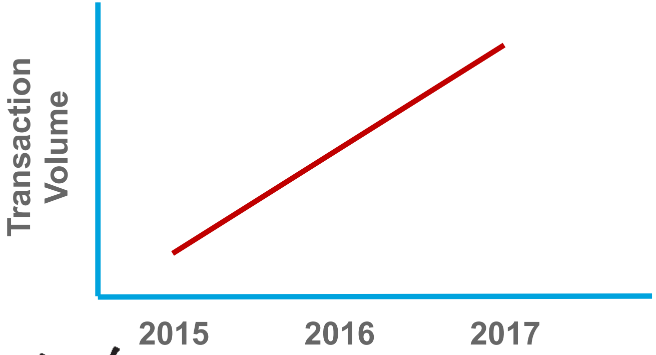
Agenda

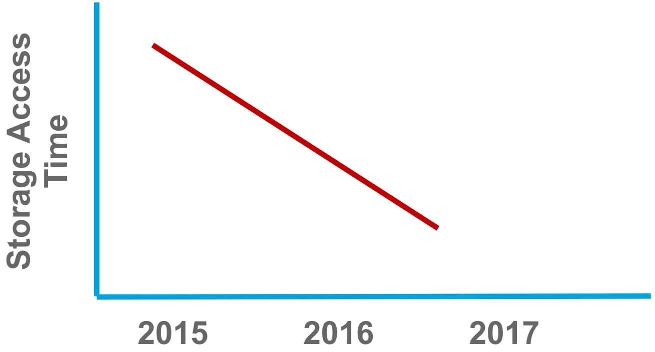
- What's New
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- Design Impacts of 25G, 50G and 100G
  - 100G Design Thoughts
  - 40/100G Optics
  - 25/50G
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)

Design for Optimal Capacity
What does 25/50/100G mean



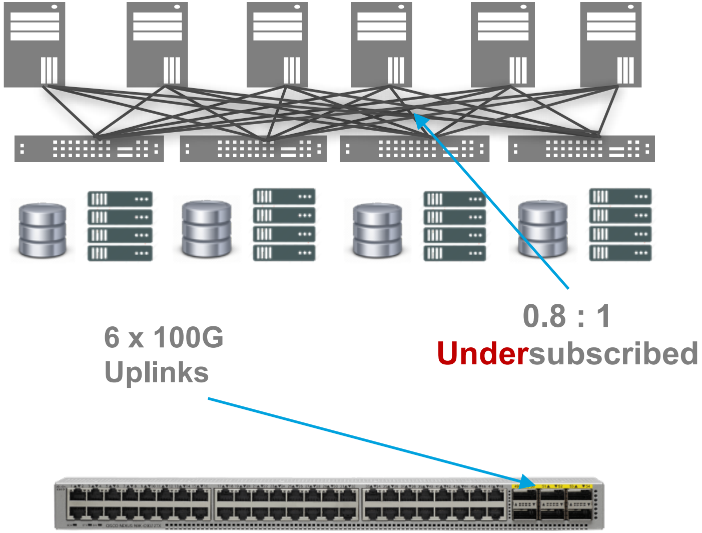






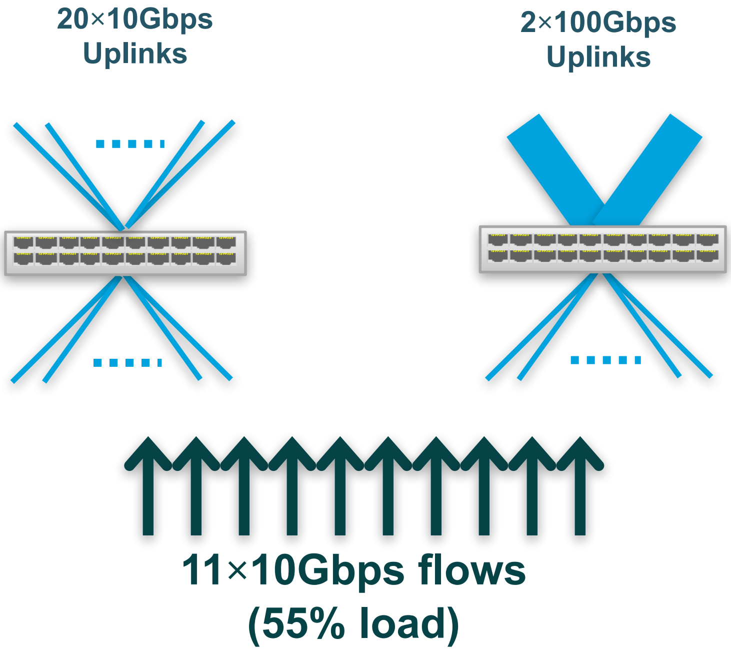
Design for Optimal Capacity
What does 100G mean

- You do not need to ‘and’ should not be designing a network that is capacity constrained
  - Capacity has a ‘very’ different cost point than it did even as recently as last year
- Design for the Optimal Capacity Requirements
  - Bandwidth solves problems, buffering at best masks them
- Consider “undersubscription”
  - It is now possible



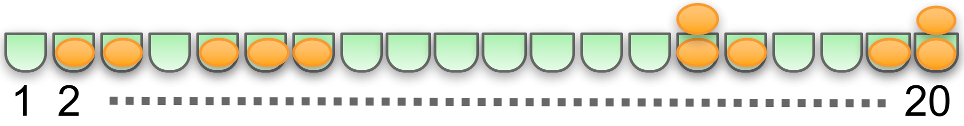
N9K-C93180YC-EX

Design for Optimal Capacity
What does 100G mean

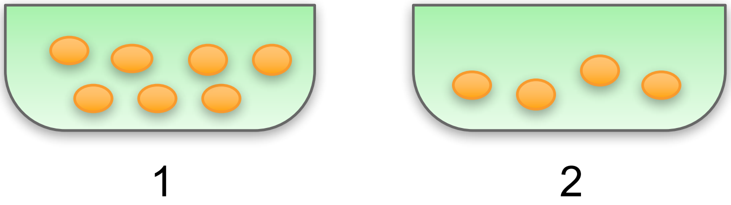


11×10Gbps flows
(55% load)

Prob of 100% throughput = 3.27%



Prob of 100% throughput = 99.95%



Agenda

- What's New
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- Design Impacts of 25G, 50G and 100G
  - 100G Design Thoughts
  - 40/100G Optics
  - 25/50G
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)

Optics Pluggable Multispeed Interfaces SFP & QSFP

SFP

Pluggable Options

- 100M SFP
- 1G SFP
- 10G SFP+, Twinax, AOC
- 25G SFP+, Twinax, AOC

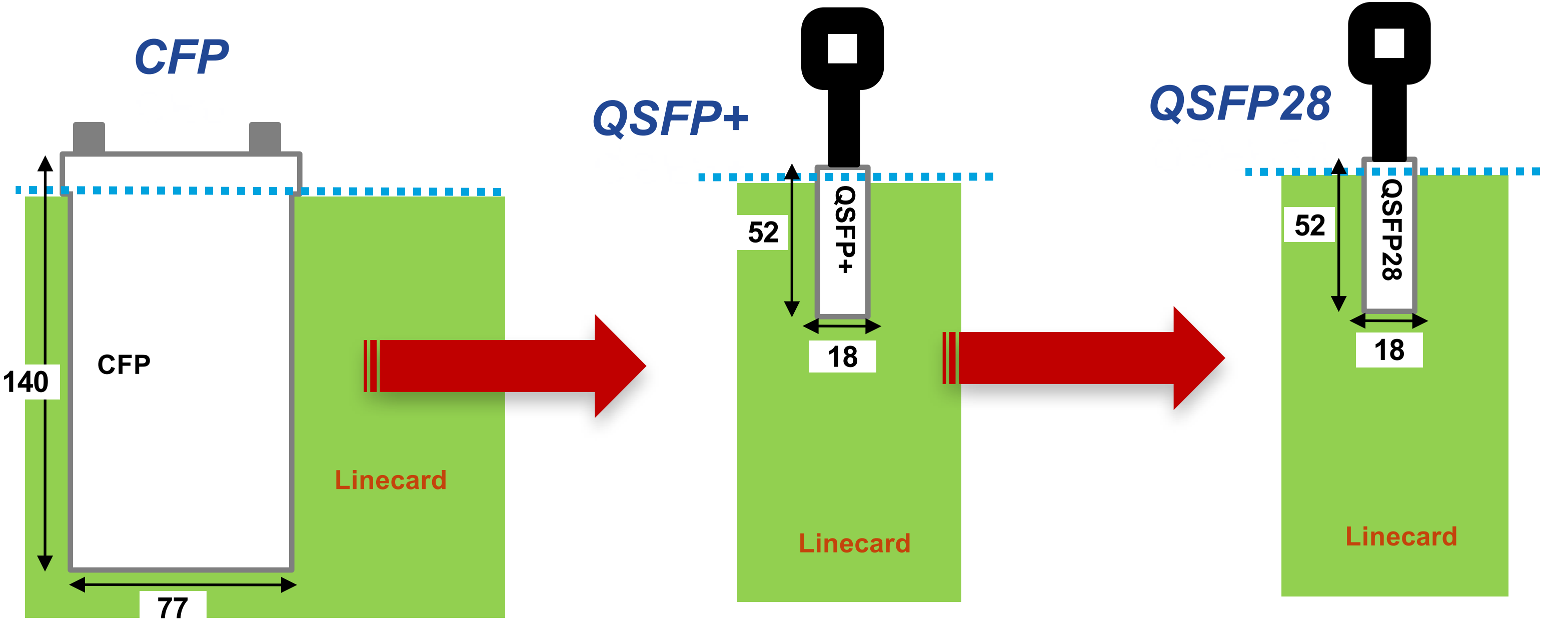
QSFP

Pluggable Options

- 100M SFP (via QSA)
- 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

Next Generation Packages for 40/100G

QSFP+ & QSFP28



QSFP28

1/2 the power & 1/5 the size of CPAK

44% the size of CFP4

	QSFP+	QSFP28
Power (W)	3.5	~3.5
Electrical	4x10G	4x25G

Support for 40G Optics QSFP+

100m, MMF

SR4 QSFP+

10km, SMF

LR4 QSFP+

40km, SMF

ER4 QSFP+

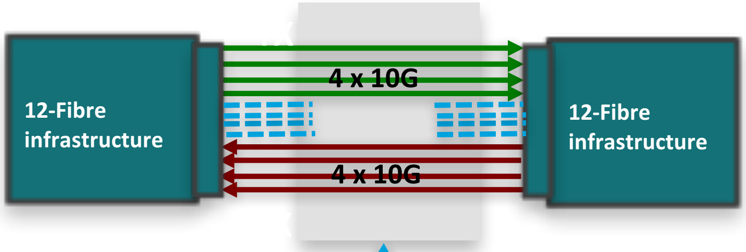
2km, SMF

WSP-Q40GLR4L QSFP+

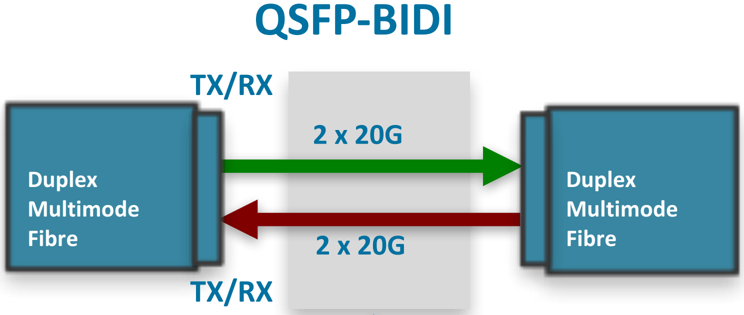
10m, Copper

4x10G AOC QSF+s

QSFP-BIDI vs. QSFP-40G-SR4



12-Fibre ribbon cable with MPO connectors at both ends



Duplex multimode fibre with Duplex LC connectors at both ends

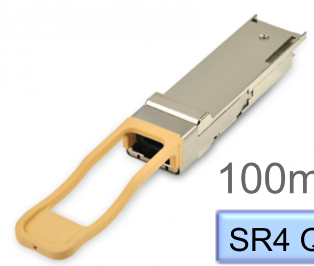
Higher cost to upgrade from 10G to 40G due to 12-Fibre infrastructure

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

Support for 40G Optics

QSFP+	Fibre	Connectors	Distance
QSFP-40G-SR4	MMF	MPO	100m
QSFP-40G-SR4-S	MMF	MPO	150m
QSFP-40G-CSR4	MMF	MPO	400m
QSFP-40GE-LR4	SMF	LC pair	10km
QSFP-40G-LR4	SMF	LC pair	10km
QSFP-40G-ER4	SMF	LC pair	40km
WSP-Q40GLR4L	SMF	LC pair	2km
QSFP-40G-LR4-S	SMF	LC pair	10km

Support for 100G Optics QSFP28



100m, MMF

SR4 QSFP28

10km, SMF

LR4 QSFP28

500m-2km, SMF

SM-SR QSFP28
CWDM QSFP28

1/2/3/5m, Copper

CU QSFP28

Built-in Cable/Optics

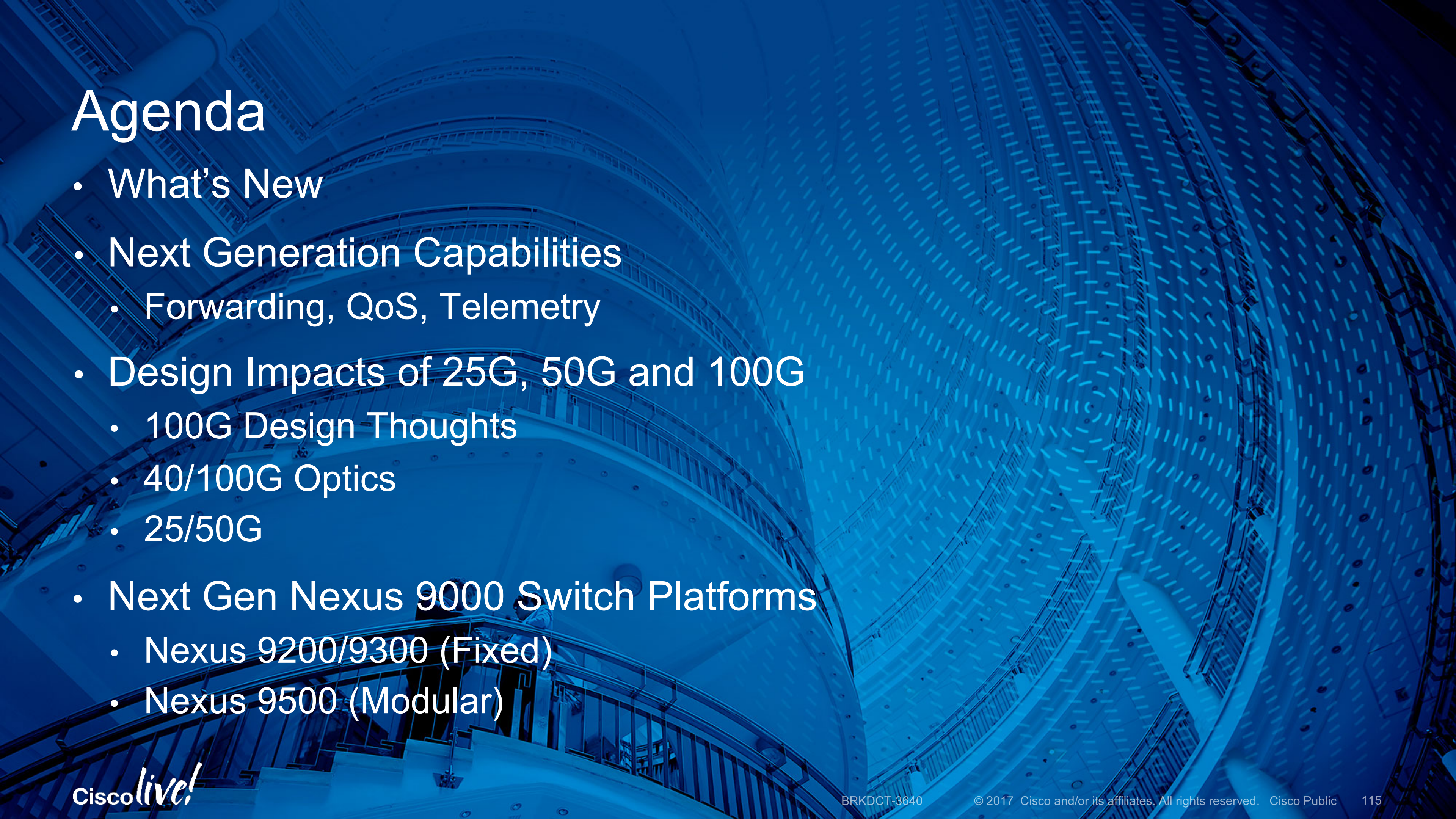
1/2/3/5/7/10/15/20 m, Copper

AOC QSFP28

Built-in Cable/Optics

Support for 100G Optics

QSFP28	Fibre	Connectors	Distance
SR4	MMF	MPO-MTP12	Up to 100m
LR4	SMF	LC pair	Up to 10km
SM SR	SMF	LC pair	Up to 500m
CWDM4	SMF	LC pair	Up to 2km
CU 1/2/3/5 m	Copper	Build-in QSFP28	Up to 5m
AOC 1/2/3/5/10/15/20m	Copper	Build-in QSFP28	Up to 20m



Agenda

- What’s New
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- Design Impacts of 25G, 50G and 100G
  - 100G Design Thoughts
  - 40/100G Optics
  - 25/50G
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)

25/50G Ethernet Standards

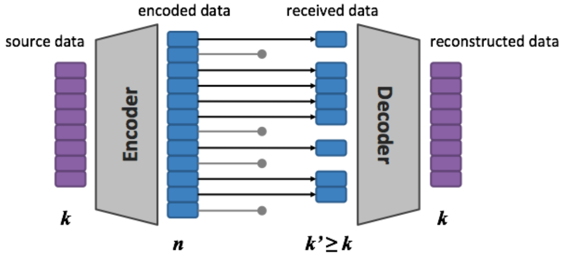
Successful plug fest UNH IOL 25/50G
Aug 1-4 2016

	Consortium	IEEE	Cisco TMG Cables*
Distance	Passive: 1,2,3 meter	Passive: 1,2,3,5 meter Optics: SR	AOC cables: 1,2,3,5,7,10M (Shipping Jan CY17')
Deployment	Within Rack	Across Rack	Within/Across Rack
Supporting Platform	N9200, N9300-EX N3200, X9700-EX	Roadmap N9300-FX X9700-FX, X97160YC-EX	N9200, N9300-EX, N3200, X9700-EX, N9300-FX
Forward Error Correction	3m needs FC FEC	3m needs FC FEC >3m need RS FEC	Can work with either FC FEC or RS FEC
NIC (Verified)	Mellanox		NIC needs to support the same FEC mode as the switch
NIC (Ongoing Testing)	Qlogic, BRCM, Intel		

*IEEE or Consortium does not spec for AOC 1,2,3,5, 7,10 meter.

What About 25G?
FEC (Forward Error Correction)

- FEC greatly reduce uncorrected errors across the media and help to extend the usable reach of those media
- FEC introduces latency penalty and depending on the distance FEC could be disabled to optimise the latency (~250 nsec)
- 25G standard support 3 modes of FEC to support different twinax cable reach
  - Clause 74 Fire code FEC: FC FEC
  - Clause 108 Reed-Solomon FEC: RS FEC
- Passive cable 1 and 2 meter does not require FEC
- Passive cable 3 meter requires FC FEC
- Passive cable more than 3 meter or 100m MMF SR optics requires RS FEC
- RS FEC introduce more latency than FC FEC



Raw BER*	BER after FEC*
5.7E-7	1.97E-29

* Example of FEC improvement of realised BER with 56G PAM4 encoding

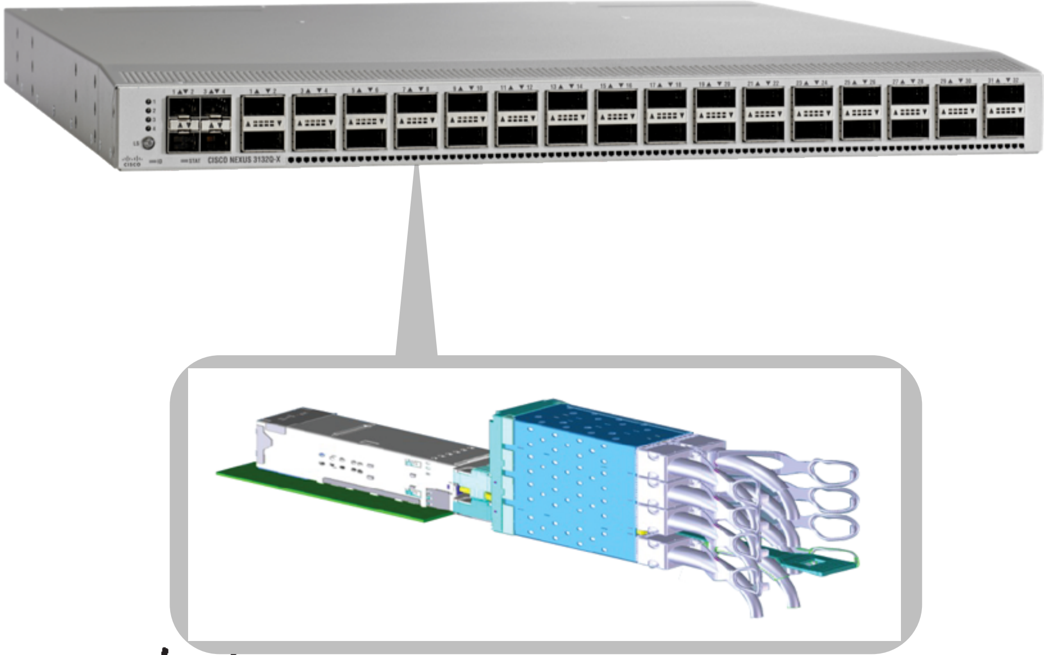
25G / 10G Backward Compatibility

- 25G Ethernet passive cable support both 10G and 25G speed
- 10G and 40G Ethernet passive cable are not designed to run at 25G Ethernet single lane

Optics		Platform
Passive Cables	1/2/3/5 meter	Nexus 92160YC-X
Active Cables	1/2/3 meter *	Nexus 92160YC-X
Breakout Cables	1/2/3 meter	Nexus 9232C Nexus 9236C Nexus 92160YCX

* Active cable greater than 3 meter requires FEC RS which is not supported on Nexus 92160YCX

Cisco QSFP-to-SFP Converters



Q1CY16

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

What is Next?
50/400G

Pulse **A**mplitude **M**odulation
4 Indicates the number of valid signal levels

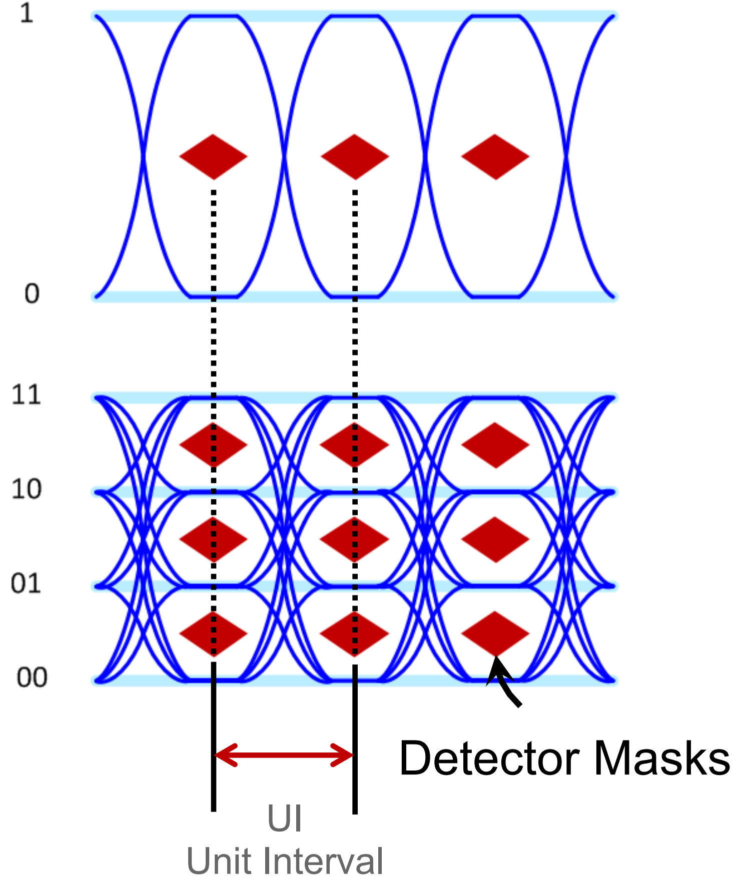
- NRZ is the same as PAM2
- PAM3 is used in 100Base-T
- PAM5 is used in 1000Base-T
- PAM16 is used in 10GBase-T

Higher order modulation with PAM has been used for decades to achieve higher bit rates

NRZ:
2 Levels
1 Bit per UI

PAM4:
4 Levels
2 Bits per UI

Ideal Differential Eye Diagrams



What is Next?
50/400G

IO Signalling	2006-2008	2009-2012	2013-2015	2016-2017	2018-2019
Backplane	3.125G NRZ	7.5G NRZ	15G NRZ	25G NRZ	56G PAM4
Chip to Chip	156MHz DDR 3.125G NRZ	7.5G NRZ	15G NRZ	25G NRZ	56G PAM4
Chip to Module	10G NRZ	10G NRZ	25G NRZ	25G NRZ	53G PAM4
LC bandwidth	320G (oversubscribed)	480G	1.44T	3.6T	14.4T
LC feature	32x10G	48x10G	14x100G 36x40G	36x100G	36x400G

Agenda

- What's New
  - 2nd Generation Nexus 9000
  - Moore's Law and 25G SerDes
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- Design Impacts of 25G, 50G and 100G
- Nexus 9000 Switch Platforms
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)

Nexus 9300 Series Switches Portfolio
First Generation

N9K-C93120TX

N9K-C9332PQ

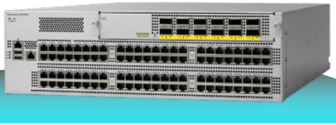
N9K-C9372PX

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+

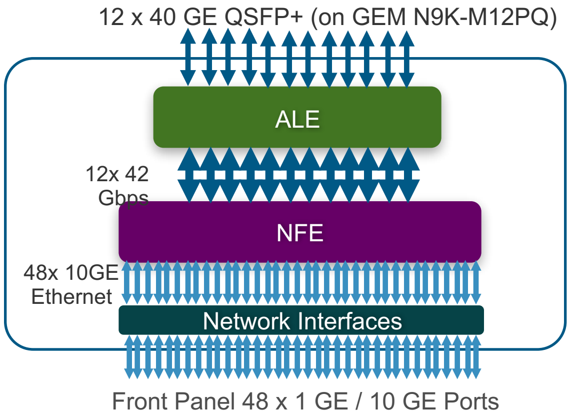
Nexus 93120TX
- 2 RU w/n GEM module slot
- 1200Gbps
- 6-port 40 Gb QSFP+
- 96-port 1/10 G-T

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

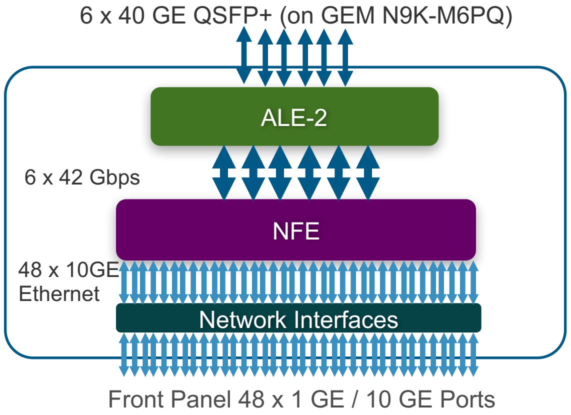
Nexus 93128TX/ 93128PX
- 3 RU with 1 GEM module slot
- 1,280Gbps
- 96-port 1/10 G-T on Nexus 93128TX
- 96-port 1/10 SFP+ on Nexus 93128P
- 6 ports 40 Gb QSFP+ on N9K-M6PQ GEM module
- 8 ports 40 Gb QSFP+ on N9K-M12PQ GEM module
- 2 ports 100 Gb CFP2 on N9K-M4PC-CFP2 GEM module

First Gen Nexus 9300 Series Switch Architecture

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



12 x 40 GE QSFP+ (on GEM N9K-M12PQ)
ALE
12x 42 Gbps
NFE
48x 10GE Ethernet
Network Interfaces
Front Panel 48 x 1 GE / 10 GE Ports



6 x 40 GE QSFP+ (on GEM N9K-M6PQ)
ALE-2
6 x 42 Gbps
NFE
48 x 10GE Ethernet
Network Interfaces
Front Panel 48 x 1 GE / 10 GE Ports

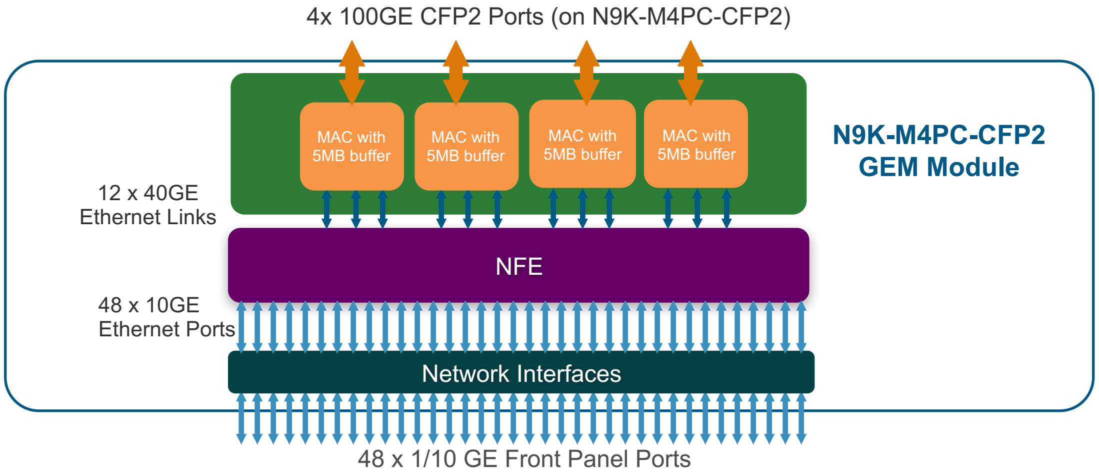
Nexus® 9396PX/TX with N9K-M12PQ GEM Module

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

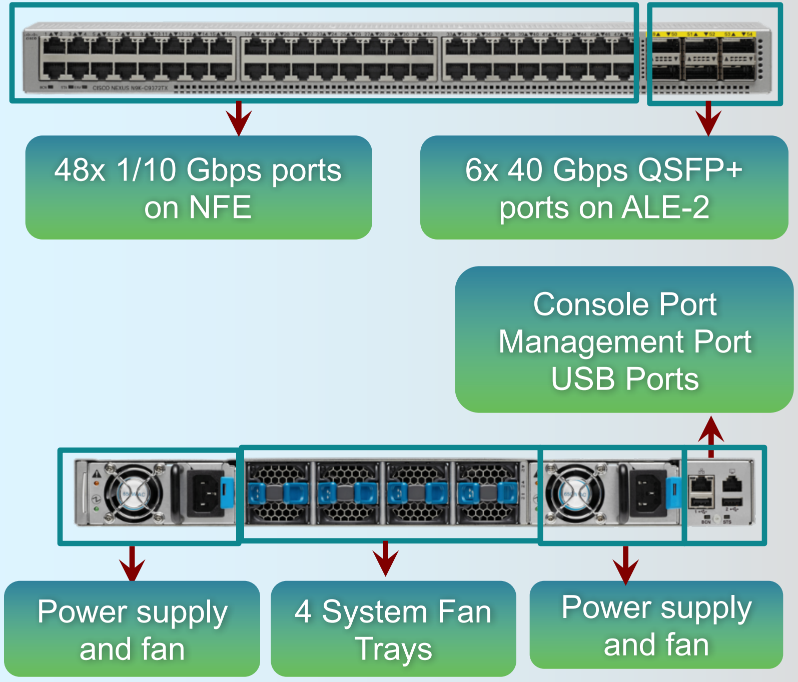
- Hardware is capable of VXLAN bridging only
- Hardware is capable of supporting NX-OS only
- Line rate performance for packet sizes > 200-Bytes

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

N9K-C9372PX / N9K-C9372TX



48x 1/10 Gbps ports on NFE

6x 40 Gbps QSFP+ ports on ALE-2

Console Port
Management Port
USB Ports

Power supply and fan

4 System Fan Trays

Power supply and fan

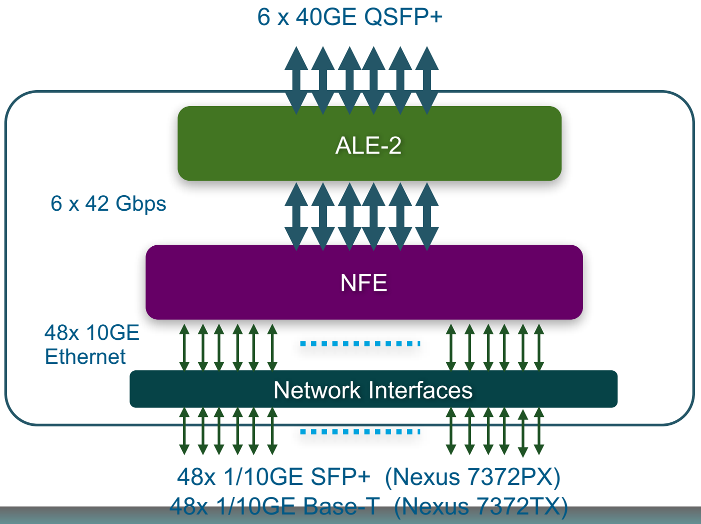
Nexus 9300 Series Switch Architecture

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



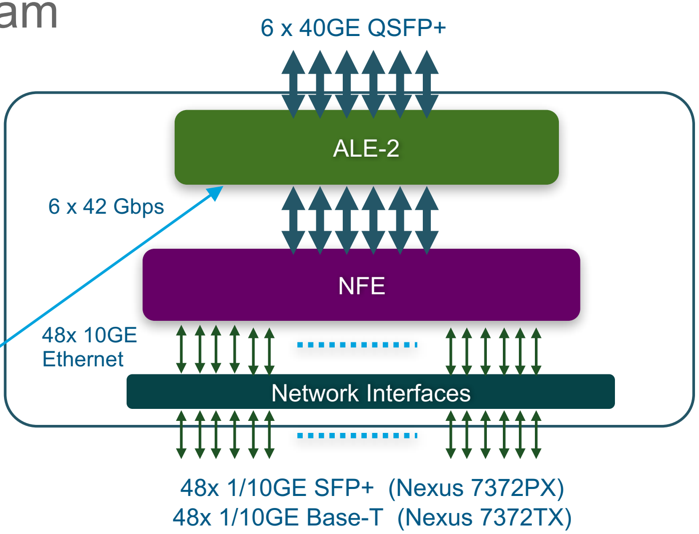
Nexus® 9372PX, Nexus 9372TX

- 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

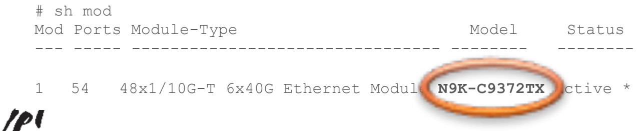
Nexus 9300 ’E’ Series

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



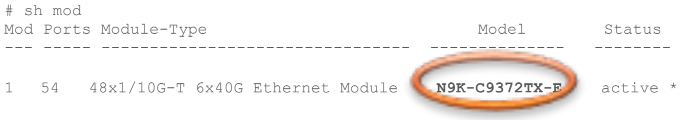
N9K-C9372TX
Show module information:



```
# 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:

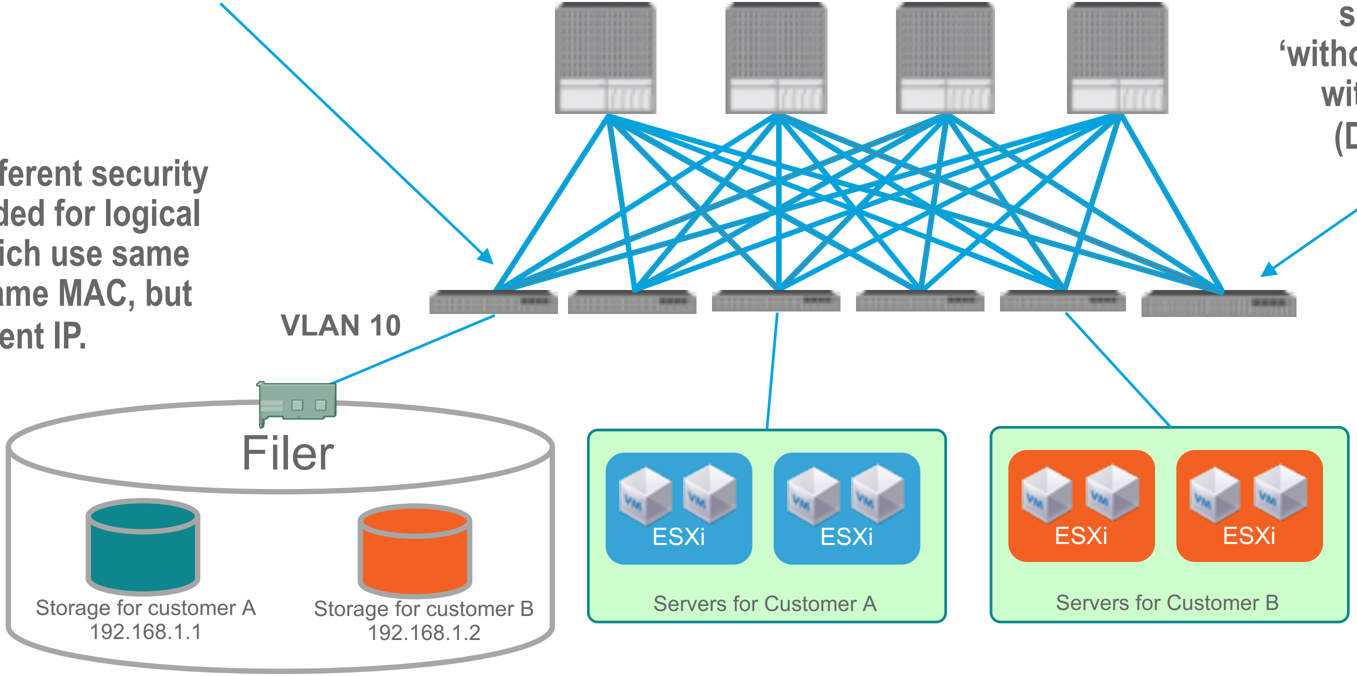


```
# sh mod
Mod Ports Module-Type                         Model           Status
--- ----- ----------------------------------- --------------  --------

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

Nexus 9300 ’E’ Series - IP Based EPG

- 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



Use Case: Different security policy is needed for logical storages which use same VLAN and same MAC, but different IP.

IP Based EPG supported ‘without’ E version with vSwitch (DVS, AVS)

Next Gen – 9200 & 9300EX
2nd Generation

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

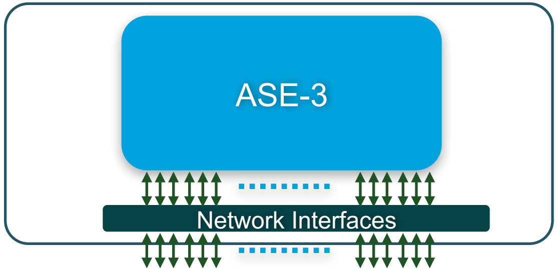
Industry’s first 36p 100G 1RU switch

Industry’s first native 25G VXLAN capable switch

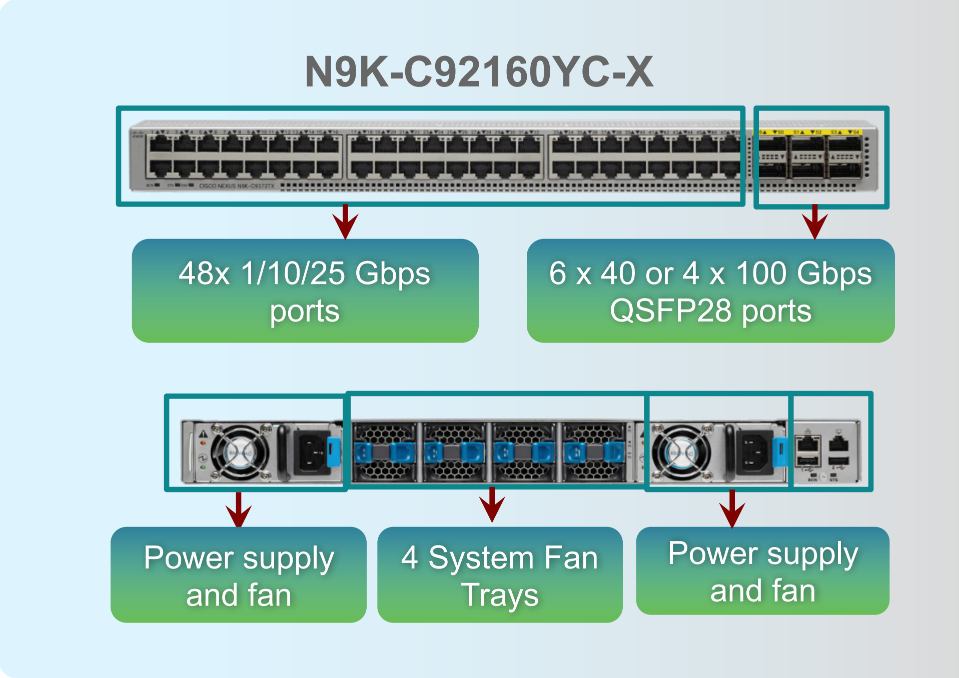
Up to 30 MB shared buffer

High density compact 40/100G aggregation

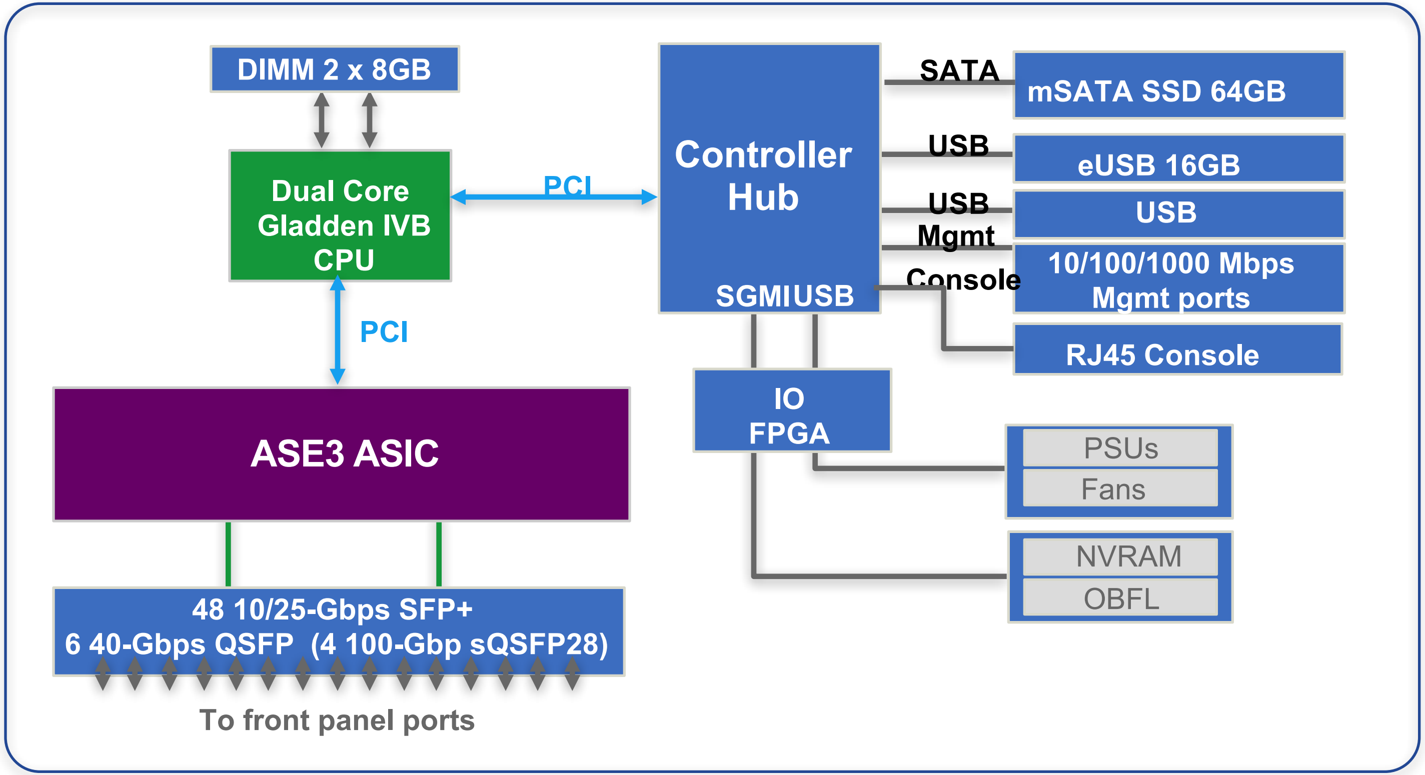
Nexus 92160YC-X
ASE3 Based



- ASIC: ASE3
- 1 USB + 1 RS232 Serial
- 2-core CPU (Intel Ivy Bridge Gladden 1.8Ghz)
- 2MB NVRAM
- 16GB DRAM + 64GB SSD
- Two Power supply (650W) 1 + 1 redundant
- Typical Power Usage
  - 10G mode : 150 W
  - 25G mode : 170 W
- Maximum Power Usage 430 W
- Four Fans 3 + 1 redundant

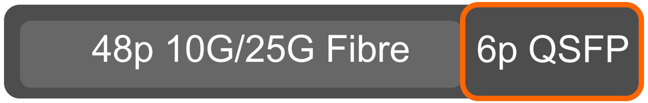


Nexus 92160YC-X
ASE3 Based



Nexus 92160 Port Configuration

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



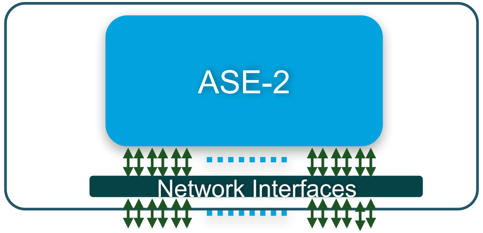
CLI to find the operation mode:

```
    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**

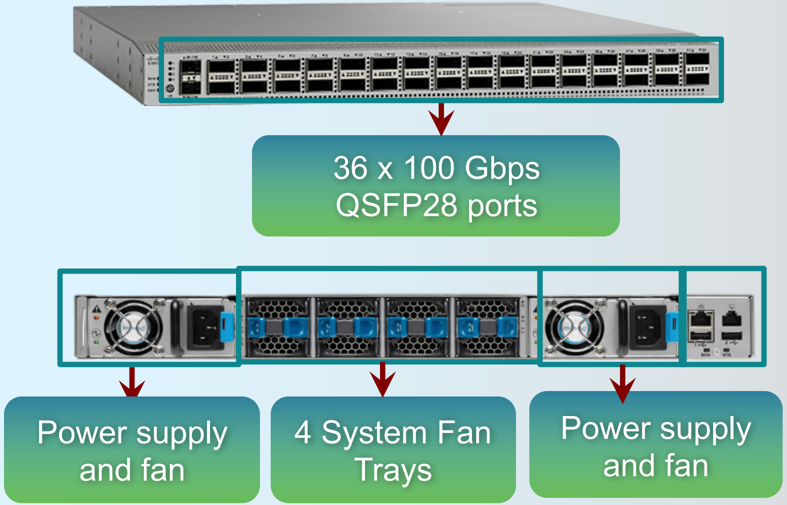
Nexus 9236C
ASE2 Based



36 x 100G QSFP28

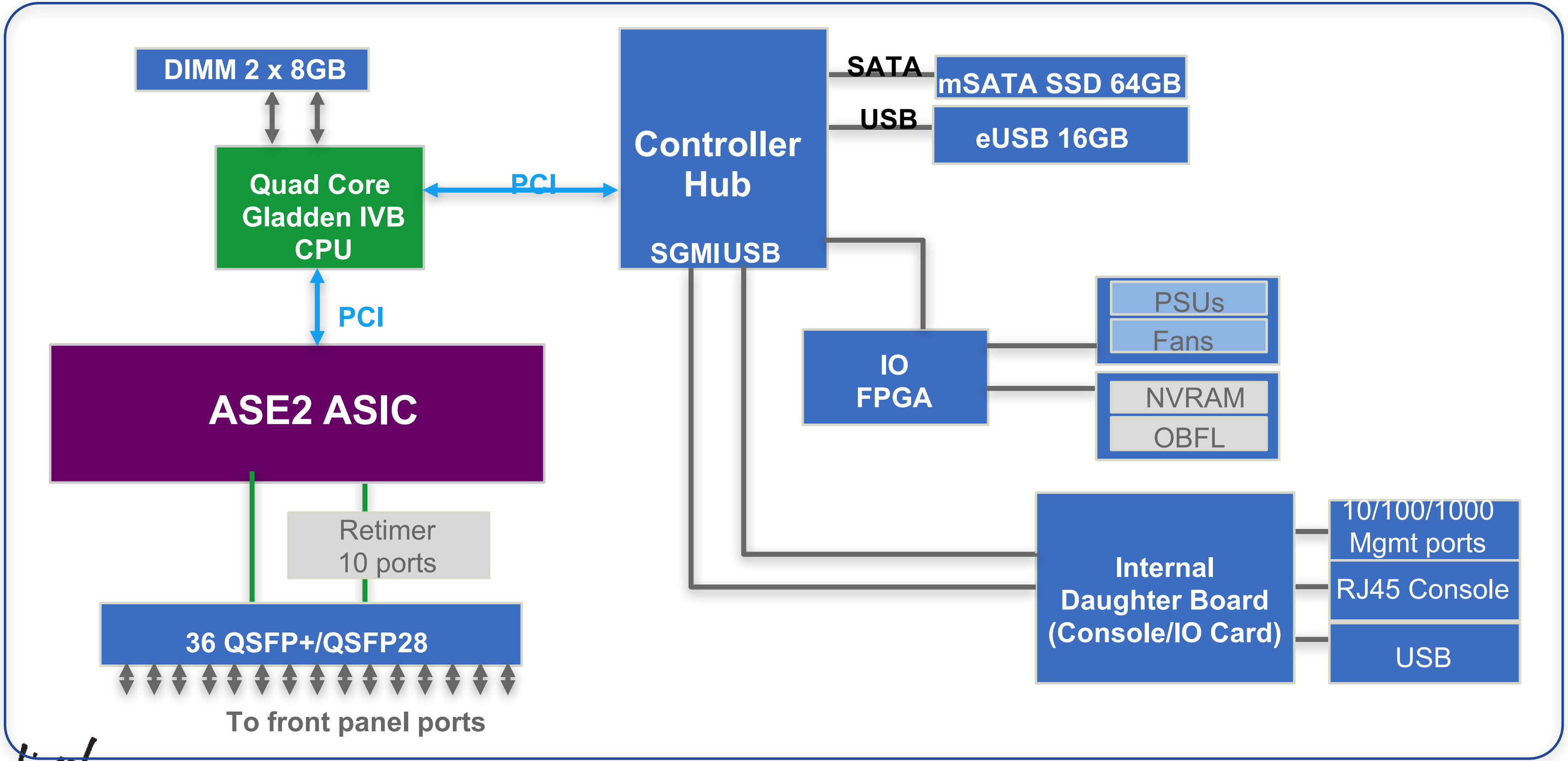
- ASIC: ASE2
- 4-core CPU (Intel Ivy Bridge Gladden 4 core at 1.8 GHz)
- 16GB DRAM + 64GB SSD
- 2MB NVRAM
- Two Power supply (1200W) 1 + 1 redundant
  - Typical Power Usage 375 W
  - Maximum Power Usage 640 W
- Two Fans 3 + 1 redundant
- 36 x 40/100G ports
- 144 10/25G ports (when all ports in breakout mode

N9K-C9236C



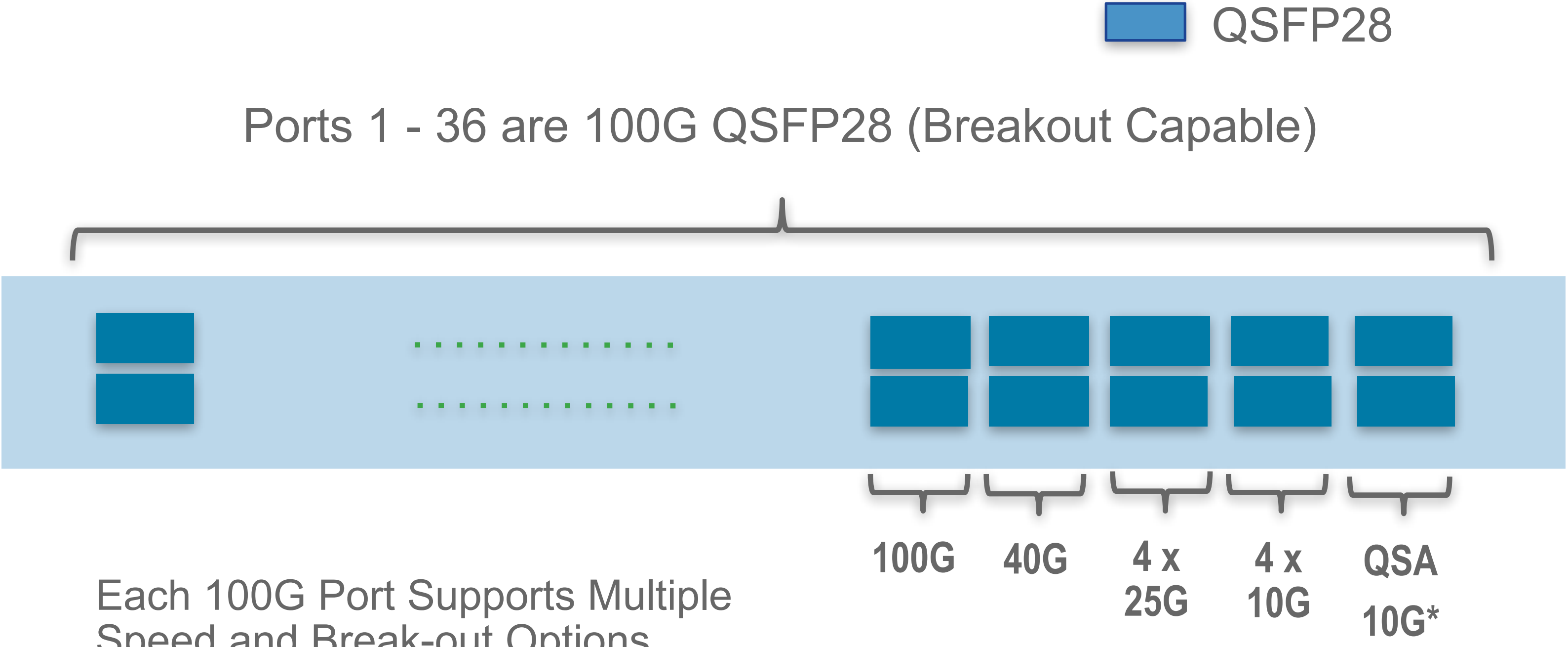
Nexus 9236C

ASE2 Based



Nexus 9236C Port Configuration

1 RU 36 Port 100G Fibre

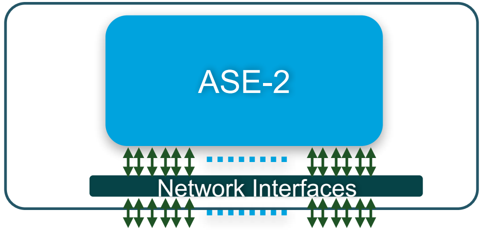


Each 100G Port Supports Multiple Speed and Break-out Options

* (QSA in a future SW release)

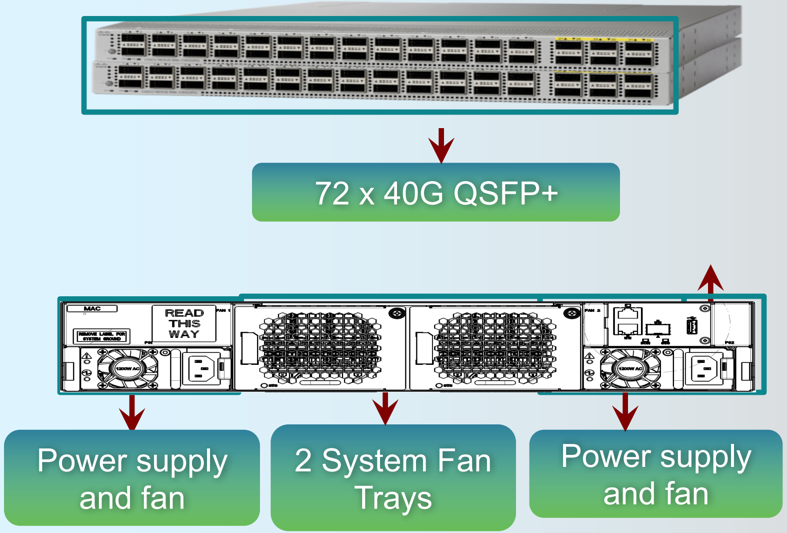
Nexus 9272Q
ASE2 Based

- ASIC: ASE2
- 4-core CPU (Intel Ivy Bridge Gladden 4 core at 1.8 GHz)
- 16GB DRAM + 64GB SSD
- 2MB NVRAM
- Two Power supply (1200W) 1 + 1 redundant
  - Typical Power Usage 310 W
  - Maximum Power Usage 1050 W
- Two Fans 3 + 1 redundant
- 36 x 40/100G ports
- 144 10/25G ports (when all ports in breakout mode



72 x 40G QSFP+

N9K-C9272Q



72 x 40G QSFP+

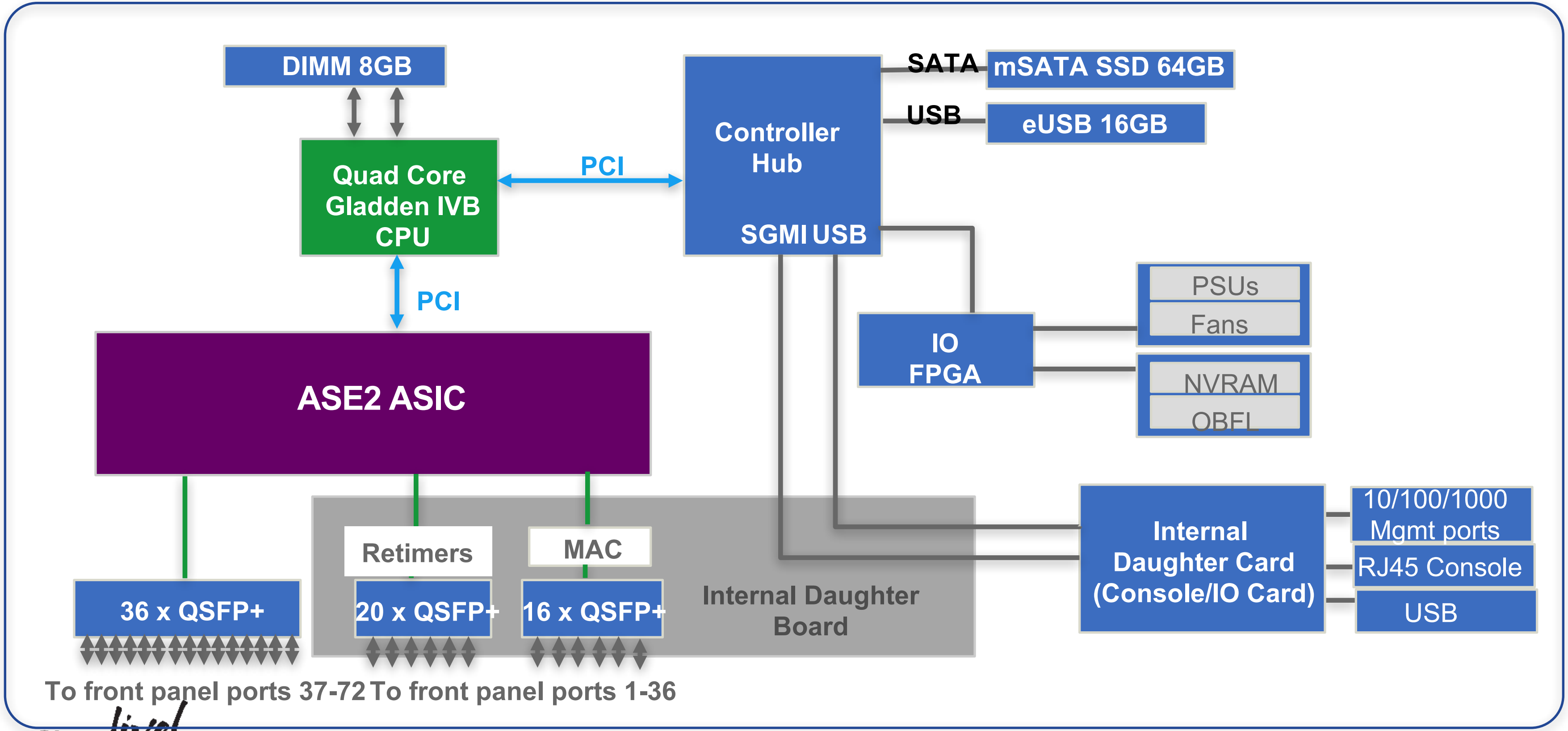
Power supply and fan

2 System Fan Trays

Power supply and fan

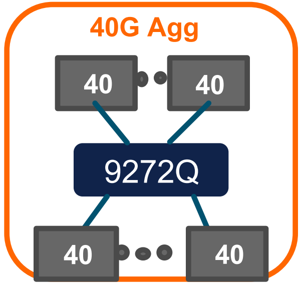
Cisco live!

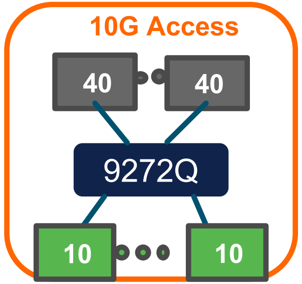
Nexus 9272Q Architecture



Nexus 9272Q Port Configuration

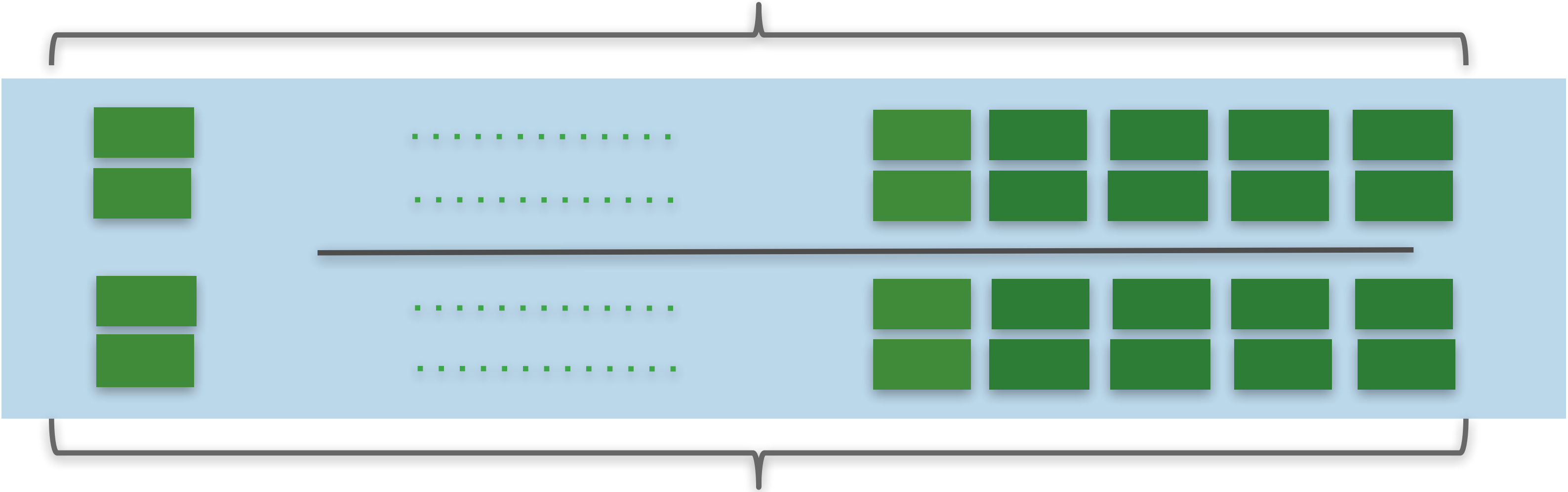
2RU 72 Port 40G Fibre





🟩 QSFP+

Ports 1 - 36 are 40G QSFP+

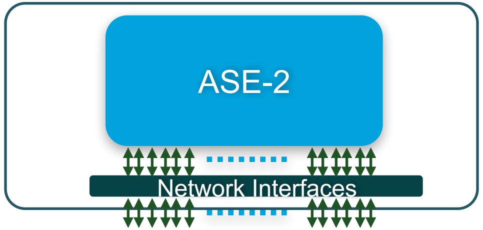


Ports 37 - 72 are 40G QSFP+ (Breakout Capable 144 x 10G)

Nexus 92304QC
ASE2 Based

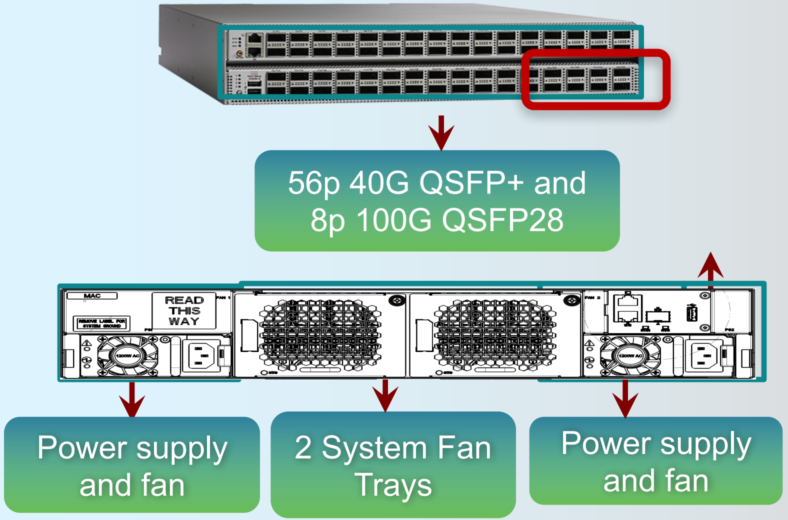
- ASIC: ASE2
- 4-core CPU (Intel Ivy Bridge Gladden 4 core at 1.8 GHz)
- 16GB DRAM + 64GB SSD
- 2MB NVRAM
- Two Power supply (1200W) 1 + 1 redundant
    - Typical Power Usage 305 W
    - Maximum Power Usage 720 W
- Two Fans 3 + 1 redundant

- 56 x 40 Gbps + 8 x 100 Gbps



56 x 40G QSFP+ & 8 x 100G QSFP28

N9K-C92304QC



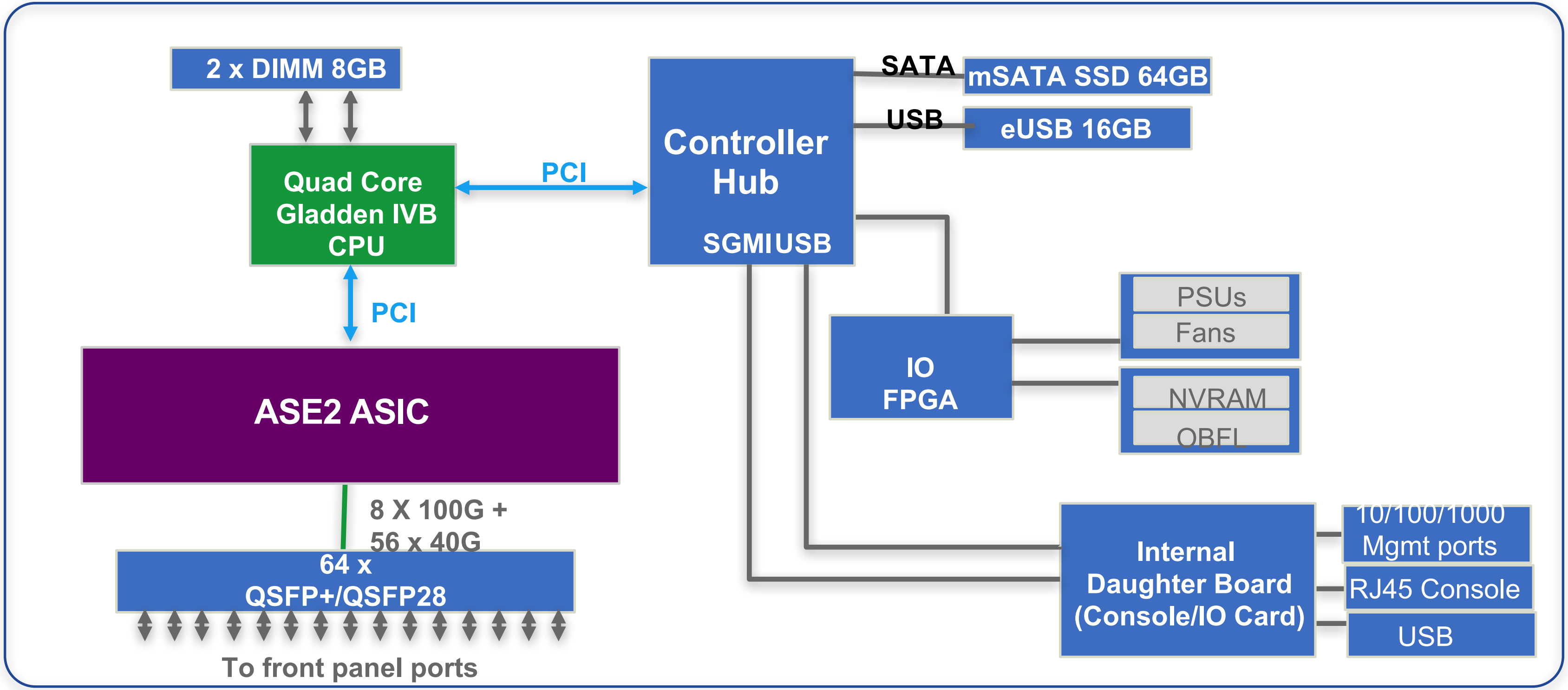
56p 40G QSFP+ and 8p 100G QSFP28

Power supply and fan

2 System Fan Trays

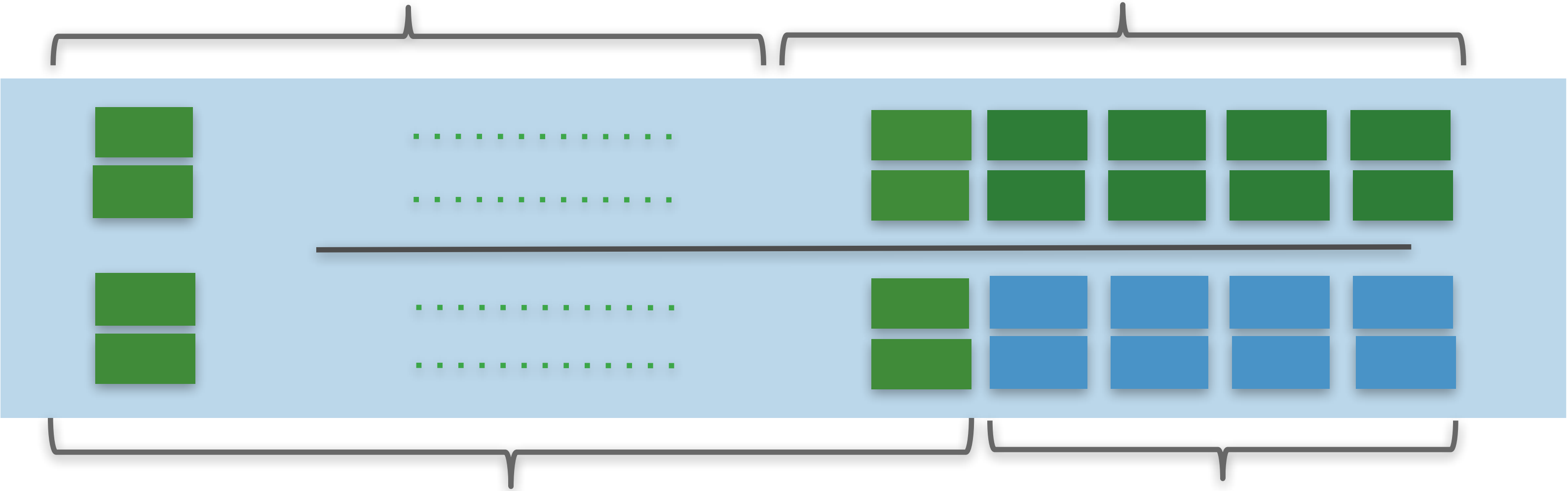
Power supply and fan

Nexus 92304 Architecture



Nexus 92304QC Port Configuration

2RU 56p 40G Fibre + 8p 40G/00G



🟦 QSFP28 🟩 QSFP+

Ports 1-16 are 40G QSFP+
(Breakout Capable 4 x10G)

Ports 17-32 are 40G QSFP+

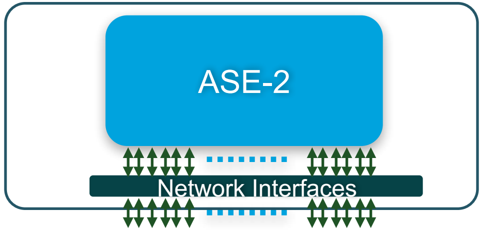
Ports 33-56 are 40G QSFP+

Ports 57-64 100G
QSFP28

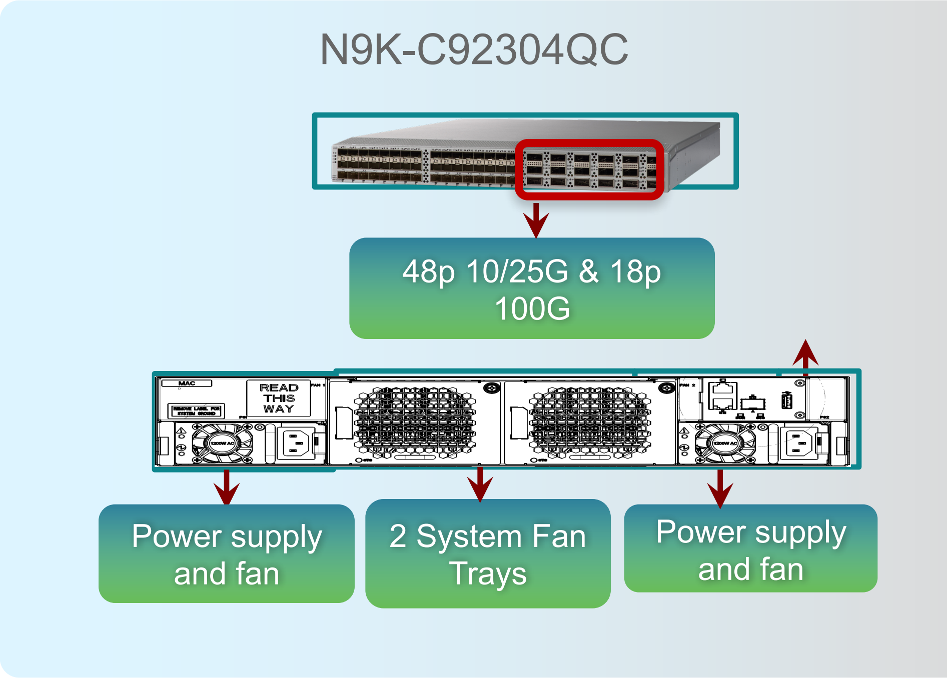
N9K-C92300YC
ASE2 Based

- ASIC: ASE2
- 2-core CPU (*Uses Intel Broadwell CPU*)
- 16GB DRAM + 64GB SSD
- 2MB NVRAM
- Two Power supply (1200W) 1 + 1 redundant
- Two Fans 3 + 1 redundant

- 10/25G access and 100G uplink in a compact form factor
- 12p 100G for 1:1 subscription and additional 6p 100G for peer links

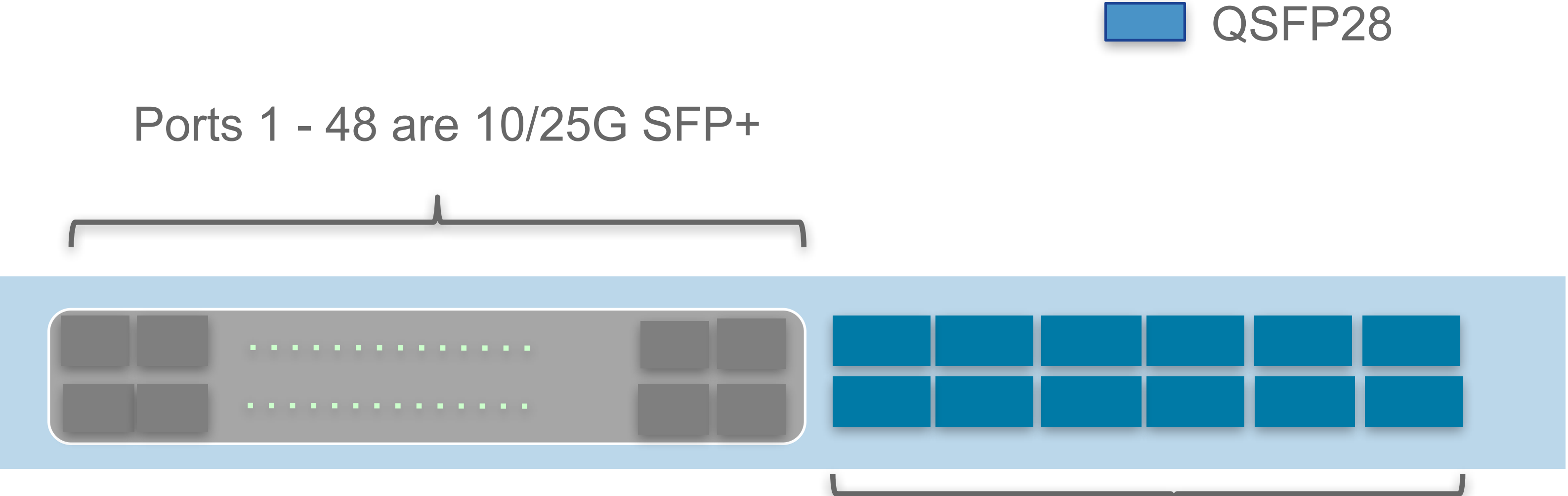


48p 10/25G & 18p 100G



N9K-C92300YC

48p 10/25G & 18p 100G



🟦 QSFP28

Ports 1 - 48 are 10/25G SFP+

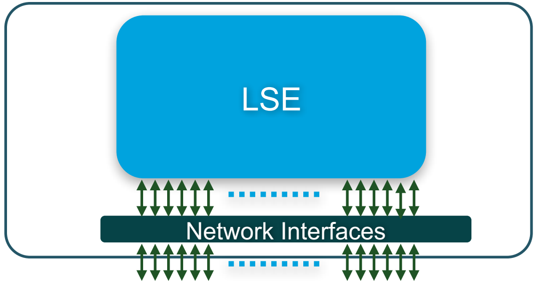
40/100G

No Breakout

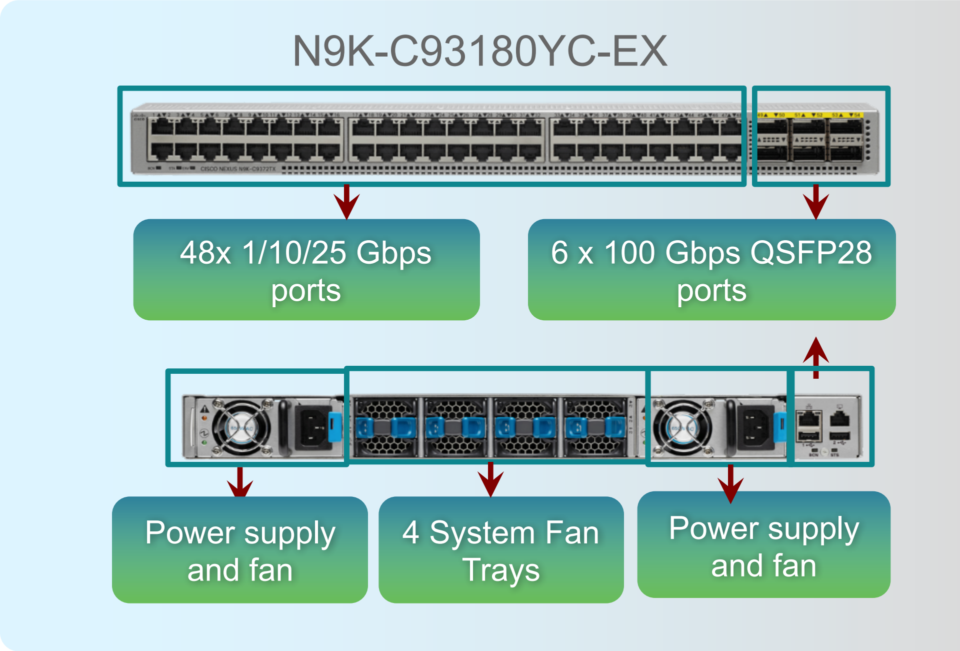
Nexus 93180YC-EX Series
LSE Based

- ASIC: LSE
- 2-core CPU (Intel Ivy Bridge Gladden)
- 16GB DRAM + 64GB SSD
- 2MB NVRAM
- Two Power supply (650W) 1 + 1 redundant
- Power consumption 248 W
- Four Fans 3 + 1 redundant

- Support both NX-OS mode and ACI mode (ACI leaf)
- Flow Cache


LSE

Network Interfaces


N9K-C93180YC-EX

48x 1/10/25 Gbps ports

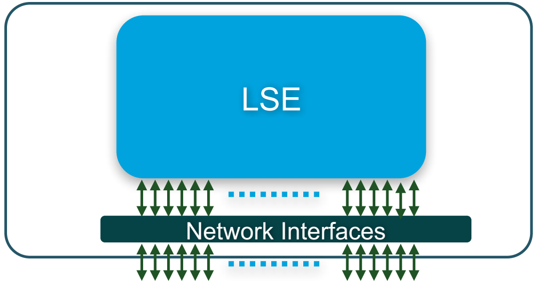
6 x 100 Gbps QSFP28 ports

Power supply and fan

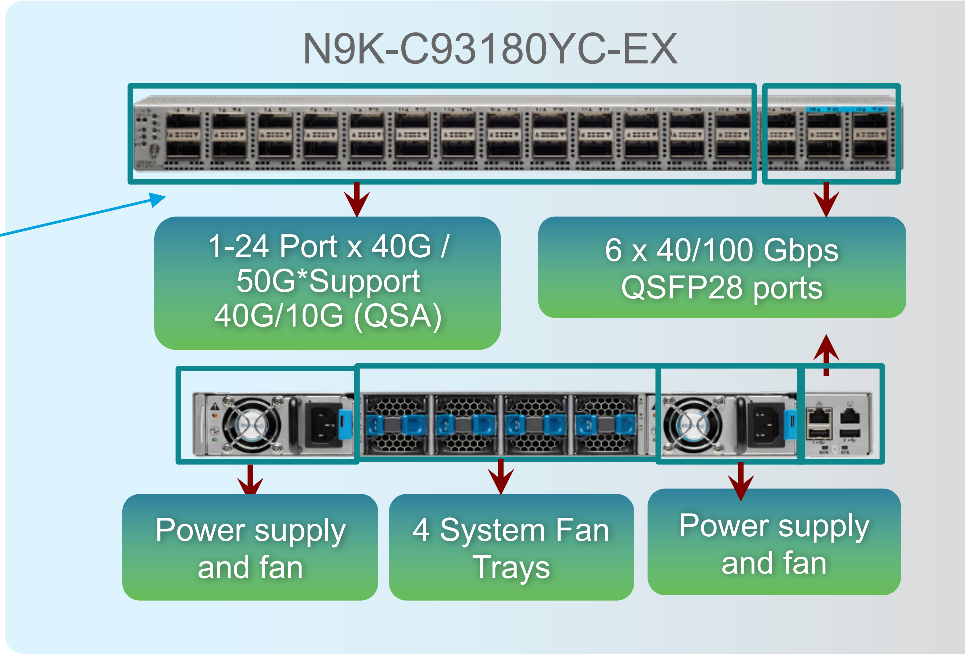
4 System Fan Trays

Power supply and fan

Cisco Nexus 93180LC-EX Switch
LSE Based



- ASIC: LSE
- 2-core CPU (*Uses Intel Broadwell CPU*)
- 1RU 32-Port (24p 40/50G, 6p 100G) QSFP Switch
- *Hardware is capable of 50G / 100G on the 'server ports' but software support will come later
- Support both NX-OS mode and ACI mode (ACI leaf)
- Flow Cache


N9K-C93180YC-EX

1-24 Port x 40G / 50G*Support 40G/10G (QSA)

6 x 40/100 Gbps QSFP28 ports

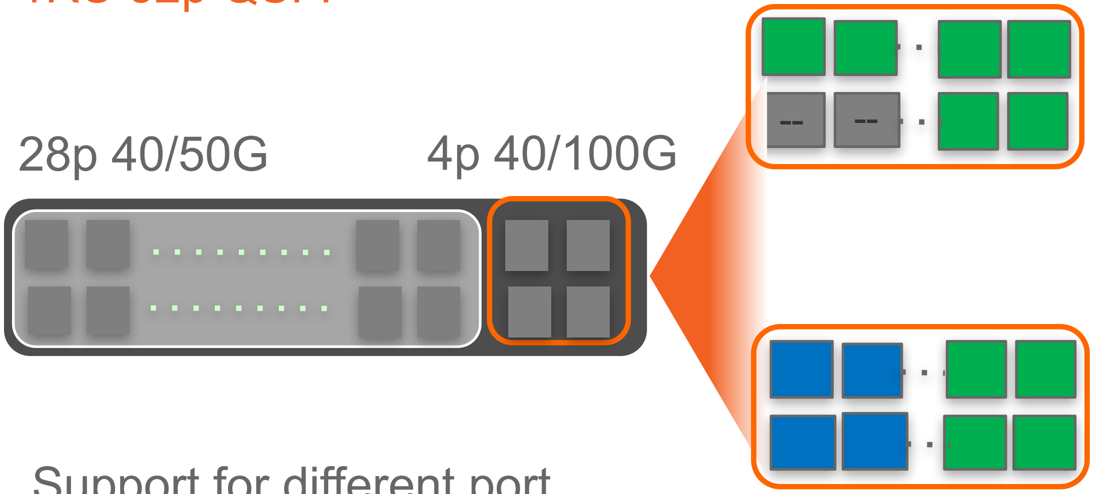
Power supply and fan

4 System Fan Trays

Power supply and fan

Nexus 93180LC-EX Port Configuration

1RU 32p QSFP



Port configuration supported:

- Ports 1,3,5…27, 29, 30, 31, 32 are 100G capable (ports 2, 4, … 28 are shut down)
- All active ports are breakout capable: 4x 10G, 4x25G

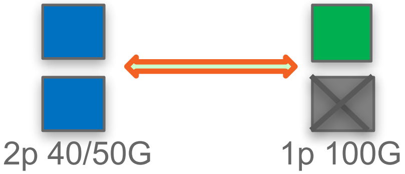
28p 40/50G & 4p 40/100G (40G Leaf)

Port configuration supported:

- Ports 1 – 28 support 40/50G, ports 29 – 32 support 100G
- Ports 1, 3, 5…27, 29, 30, 31, 32 are breakout capable. (If breakout is enabled the port below is shut down – except for ports 29 – 32)

Support for different port templates (reboot required)

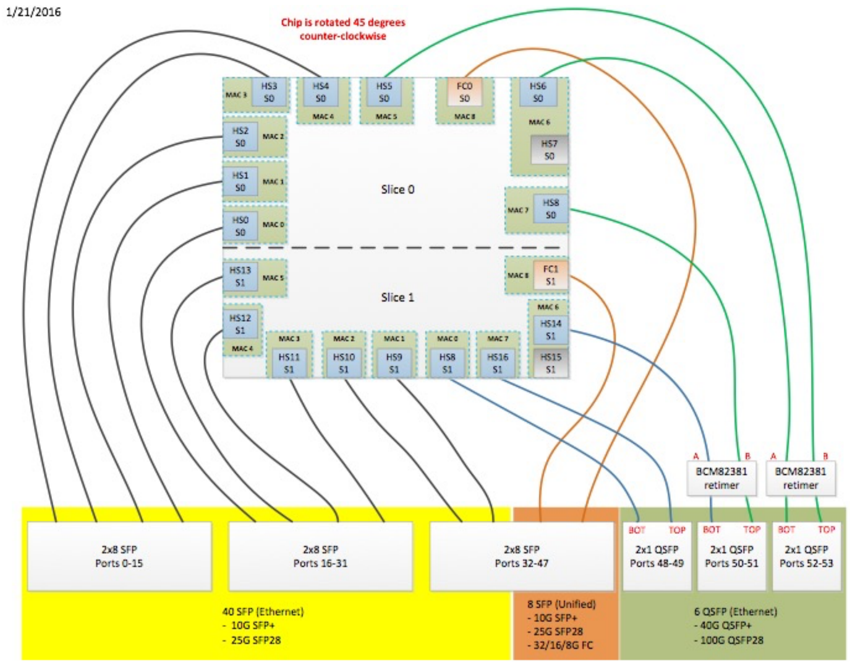
- 28p 40/50G + 4p 40/100G
- 24p 40/50G + 6p 40/100G
- …
- 18p 40/100G



Note: Please check software roadmap for supported configs at FCS.

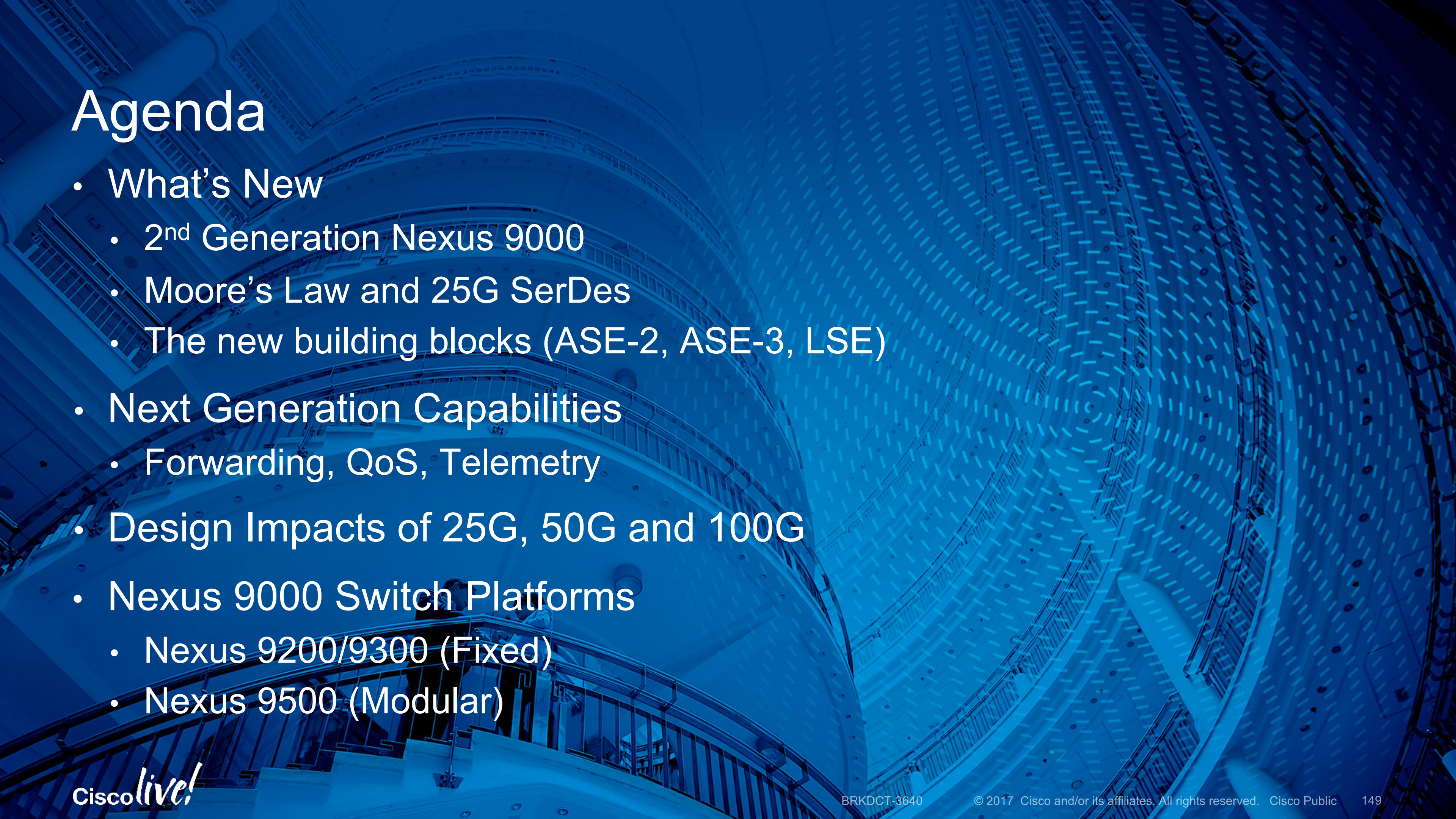
- In ACI mode 24p 40G and 6p 100G is supported at FCS
- In NX-OS mode 28p 40G and 4p 100G is supported at FCS. (Also supports 32p 40G)

Nexus 93180YC-EX Series
LSE Based



- Does it help to wire different ports to different ‘slices’
  - NO
  - Unlike a line card that has power supply, connectors, …, a slice is more like a CPU core
  - Plan for device redunandcy
- That said you like to know so ...

```
N9K-2nd-Gen#show hardware internal tah ?
```



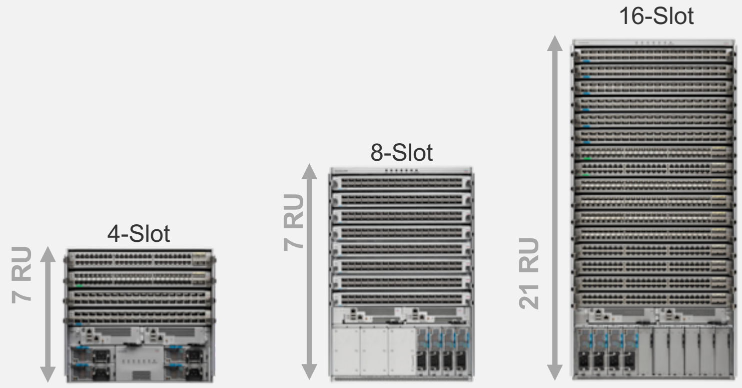
Agenda

- What’s New
  - 2nd Generation Nexus 9000
  - Moore’s Law and 25G SerDes
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- Design Impacts of 25G, 50G and 100G
- Nexus 9000 Switch Platforms
  - Nexus 9200/9300 (Fixed)
  - Nexus 9500 (Modular)

Cisco Nexus 9500 Platform Switches

Density in DC Optimised Footprint

Cisco Nexus® 9500



	Nexus 9504	Nexus 9508	Nexus 9516
Payload Slots	4	8	16
Cloud Scale	Shipping	Shipping	Mid CY17
BRCM TH	Shipping	Shipping	No Plans
BRCM T2	Shipping	Shipping	Shipping
BRCM Jericho	**Q2CY17**	**Shipping**	**Future**

Common Components

Chassis, Supervisor, System Controller, Power Supply, Fan Tray

Deployment Options

Choice of ACI and NX-OS, Choice of BRCM and Cisco ASIC

Multi-Generation Investment Protection:

No Mid-plane, Power Supply Headroom for 100/400G and Line rate encryption

Nexus 9500 – Modular
1/10/25/40/50/100G Capable

9500 Series

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)

- Analytics Ready
- Smart Buffer
- FX Support for MACSEC & Cloud-SEC

+

Cisco ASIC

16nm Technology

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

X9400-S (NX-OS)

- BCOM Trident and Tomahawk

+

Merchant ASIC

28 and 40nm Technology

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

X9600-R (NX-OS)

- Off Chip Buffer
- BCOM Jericho

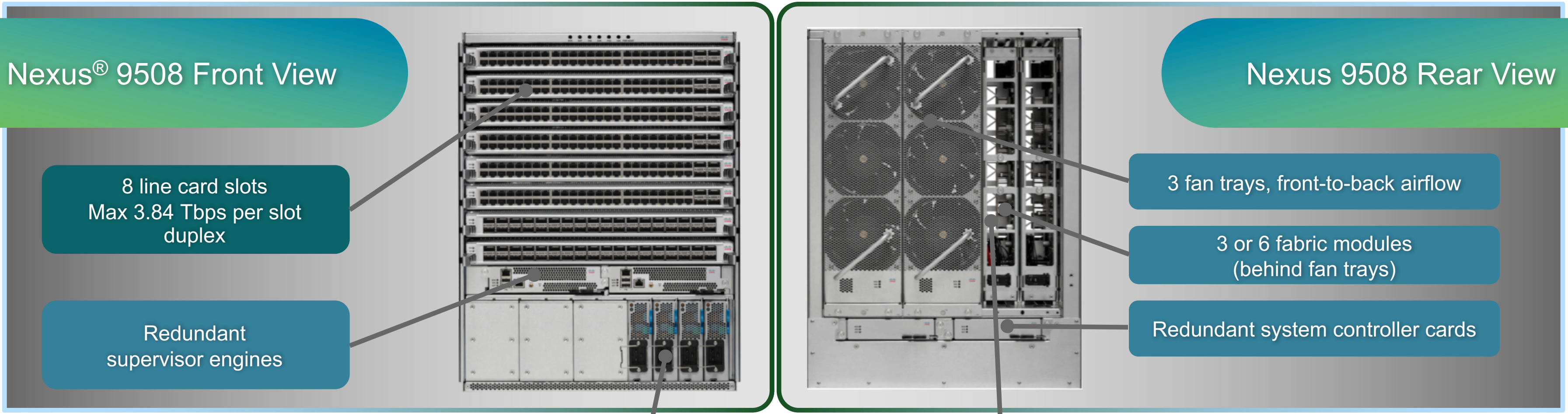
+

Merchant ASIC

28nm Technology

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

Nexus 9500 Platform Architecture



8 line card slots
Max 3.84 Tbps per slot duplex

Redundant supervisor engines

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 – Supervisors

SUP-A
4-core/4-Thread
1.8-GHz x86 Sandy Bridge
16GB of RAM
64GB SSD

SUP-B
6-core12-Thread
2.2-GHz x86 IVY Bridge
24GB of RAM
256GB SSD

SUP-A+ **Q3CY17**
4-core/**8-Thread**
1.8-GHz x86 **Broadwell**
16GB of RAM
64GB SSD

SUP-B+ **Q3CY17**
6-core/12-Thread
1.9-GHz x86 **Broadwell**
32GB of RAM
256GB SSD

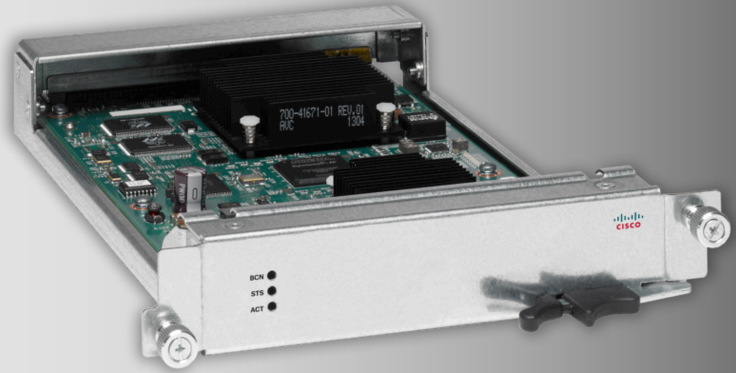
Intel CPU Generations

Westmere 32nm → **New Architecture** → Sandy Bridge 32nm → **New Mfr Process** → Ivy Bridge 22nm → **New Architecture** → Haswell 22nm → **New Mfr Process** → Broadwell 14nm

Nexus 9500 Platform Architecture

System Controller Module

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



Nexus 9500 Platform Architecture
Energy Efficient Power Supply Options

N9K-PUV-3000W-B

N9K-PDC-3000W-B

N9K-PAC-3000W-B

Q4CY17

80 PLUS

Platinum Energy efficient

One SKU option for AC or DC input

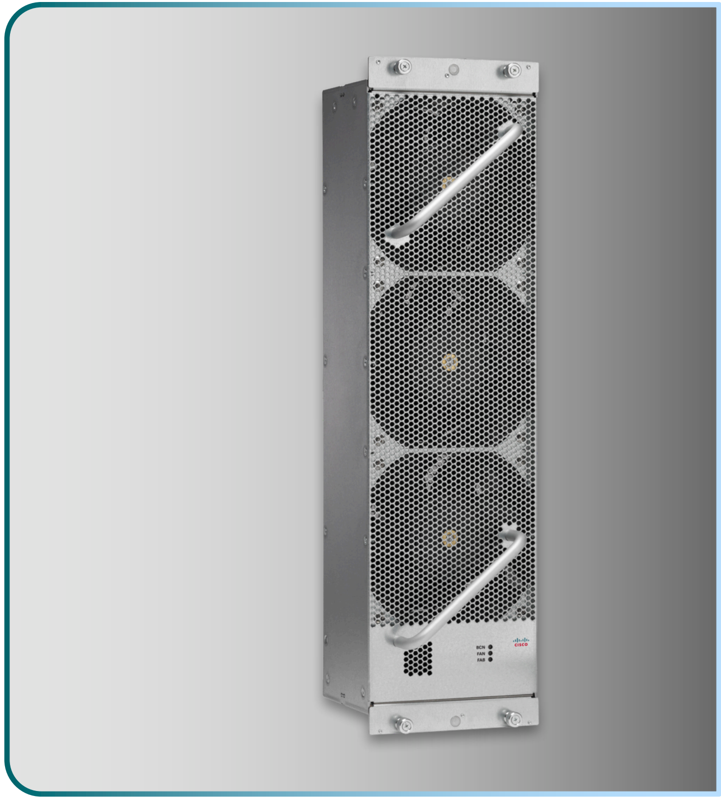
Input Voltage Range	
AC	200V – 240V
DC	-40V – -72V
High Voltage AC/DC	200 to 277V AC 240 to 380V DC
*Grid Redundant HV AC/DC	

For Nexus 9504 , 9508 and 9516
Online Insertion & Removal Capable

*Picture not shown

Nexus 9500 Platform Architecture
Fan Tray

- 3 fan trays
  - 3 dual fans per tray
  - Dynamic speed control driven by temperature sensors
  - Straight airflow across line cards and fabric modules
  - If one fan tray is removed, the other two fan trays will speed up 100% to compensate for the loss of cooling power
- N+1 Redundancy per tray



Nexus 9500 – Modular
1/10/25/40/50/100G Capable

9500 Series

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)

- Analytics Ready
- Smart Buffer
- FX Support for MACSEC & Cloud-SEC

+

Cisco ASIC

16nm Technology

Fabric Module

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

X9400-S (NX-OS)

- BCOM Trident and Tomahawk

+

Merchant ASIC

28 and 40nm Technology

Fabric Module

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

X9600-R (NX-OS)

- Off Chip Buffer
- BCOM Jericho

+

Merchant ASIC

28nm Technology

Fabric Module

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

Nexus 9400, 9500, 9600 Series Line Cards
Merchant & Merchant +

94xx Series

100G
X9432C-S: 32p 40/100G QSFP
X9408PC-CFP2: 8p 100G CFP2

40G
X9432PQ: 32p 40G QSFP+

10G
X9464PX: 48p 10G SFP+ & 4p 40G
X9464TX: 48p 10GT & 4p 40G

BROADCOM Trident T2 ‘or’ BROADCOM Tomahawk

95xx Series

40G
X9536PQ: 36p 40G QSFP+ (24p linerate)

10G
X9564PX: 48p 10G SFP+ & 4p 40G
X9564TX: 48p 10GT & 4p 40G

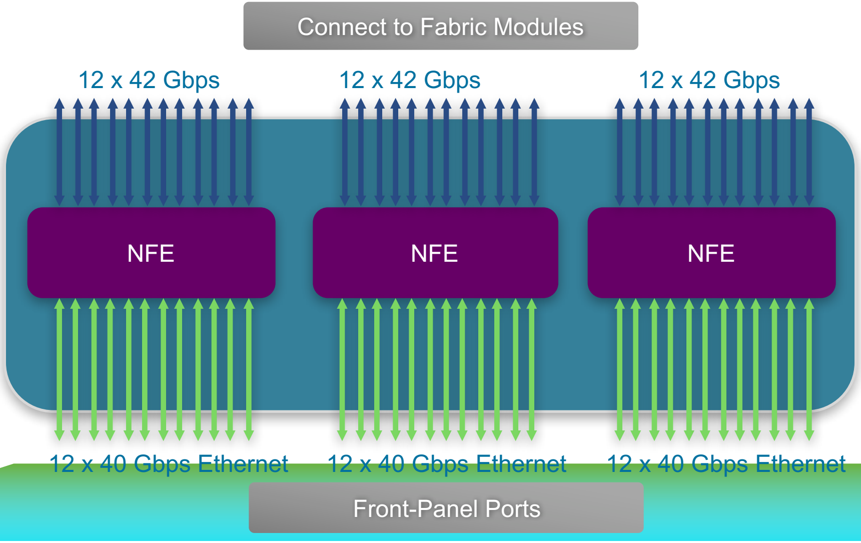
CISCO ASE + BROADCOM Trident T2

9600 Series

40G
X9636PQ: 36p 40G QSFP+

BROADCOM Trident T2

Nexus 9500 N9K-X9600 Series Line Cards
N9K-X9636PQ



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

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

Nexus 9500 N9K-X9500 Series Line Cards

N9K-X9564PX & N9K-X9564TX

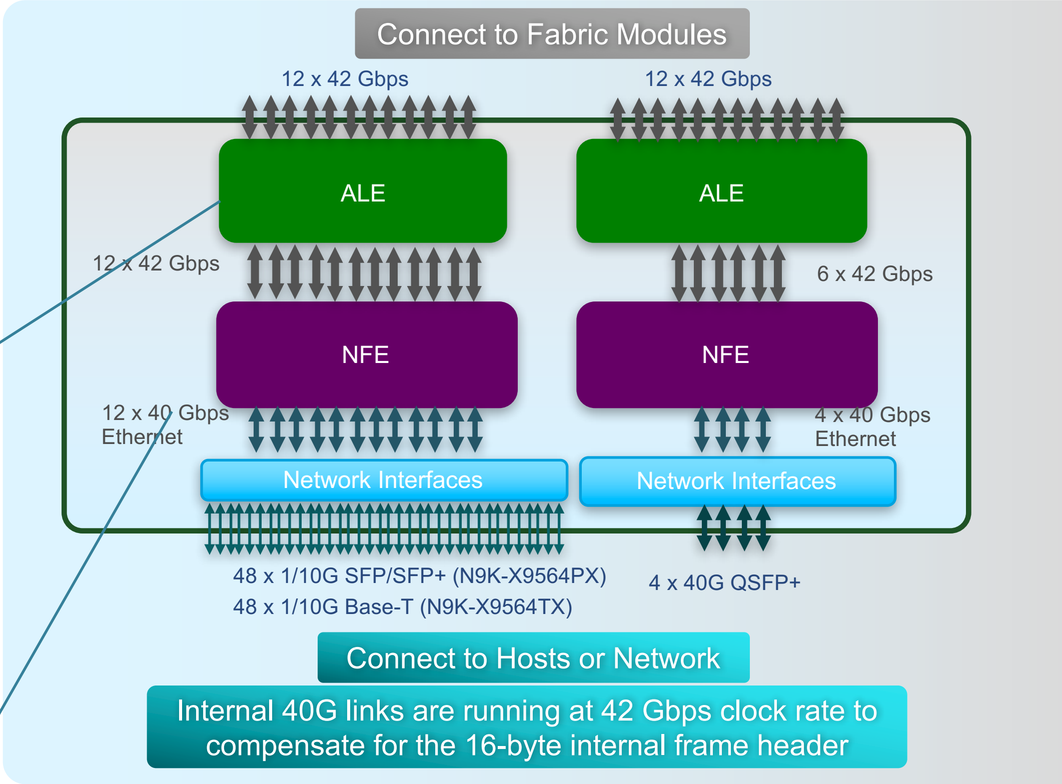
2 network forwarding engines (NFEs)

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

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

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

NFE ASICs act as main forwarding engines for standalone mode.



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

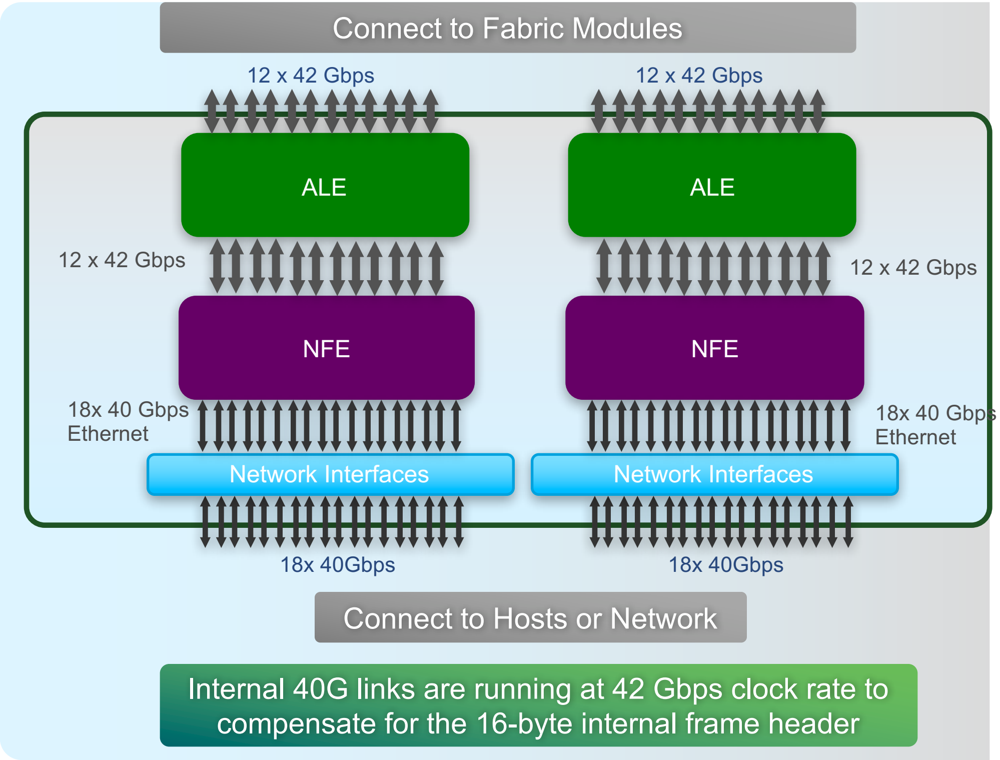
Nexus 9500 N9K-X9500 Series Line Cards

N9K-X9536PQ Line Card

2 network forwarding engines (NFEs)

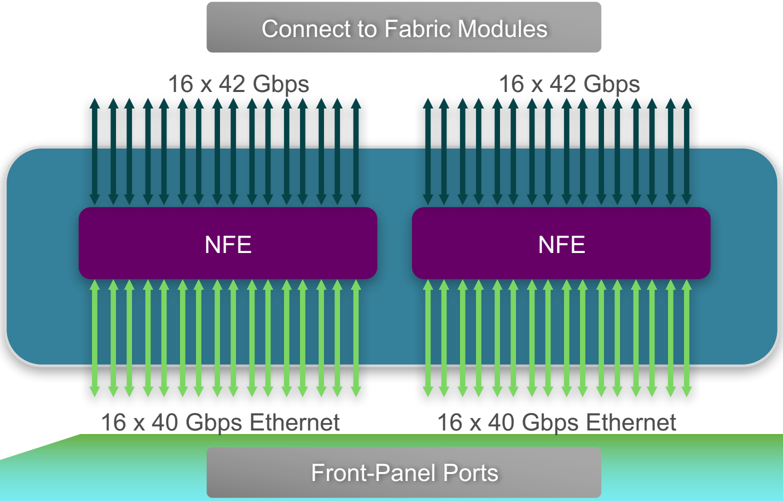
2 application leaf engines (ALEs) for additional buffering and packet handling

Need 3 fabric modules, can work with 6



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

Nexus 9500 N9K-X9400 Series Line Cards
N9K-X9432PQ



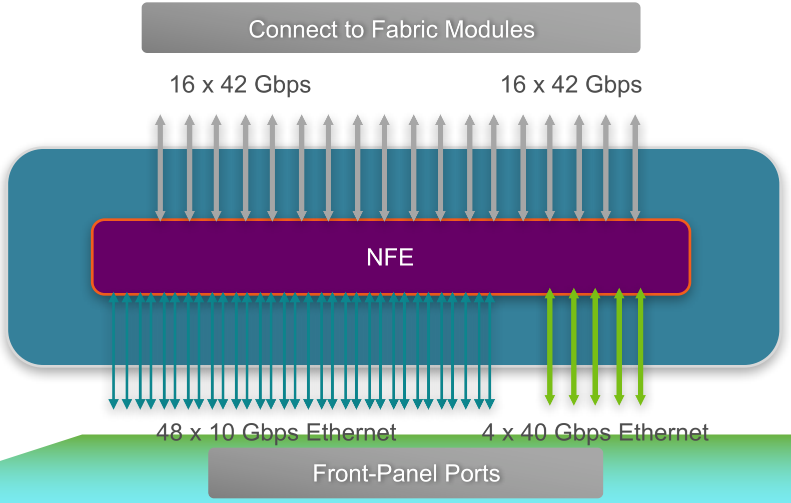
Internal 40G links are running at 42 Gbps clock rate to compensate for the 16-byte internal frame header

N9K-X9432PQ is supported in all Nexus 9500 chassis types.

- Two network forwarding engines (NFE)
- Each NFE supports 16x 40 Gbps front panel ports
- Oversubscribed for small packets (<193 Bytes)
- Line rate performance for larger packet sizes (> 193 Bytes)

Nexus 9500 N9K-X9400 Series Line Cards

N9K-X9464PX and N9K-X9464TX



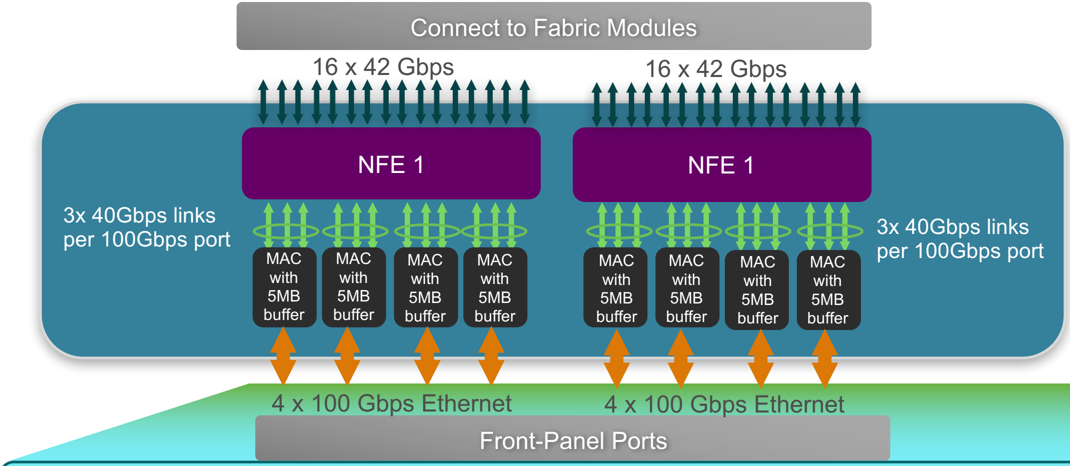
Internal 40G links are running at 42 Gbps clock rate to compensate for the 16-byte internal frame header

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

- One NFE supports all 48x 1/10 Gbps and 4x 40 Gbps front panel ports
- Oversubscribed for smaller packet sizes (<193 Bytes)
- Line rate performance for larger packet sizes (> 193 Bytes)

Nexus 9500 N9K-X9400 Series Line Cards

N9K-X9408PC-CFP2



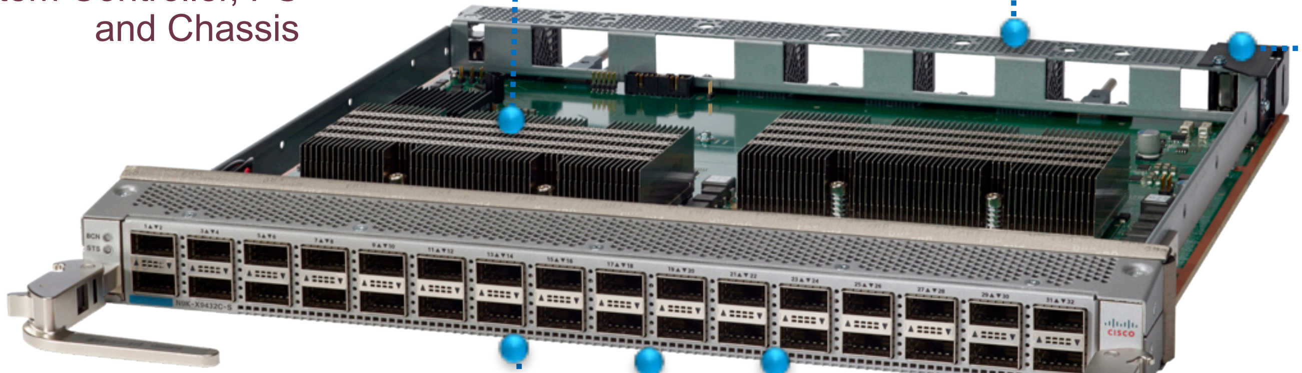
Internal links to fabric modules are running at 42 Gbps clock rate to compensate for the 16-byte internal frame header

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

40/100G - Merchant

N9K-X9432C-S



Investment Protection with Supervisors, System Controller, PS and Chassis

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

Supported in NX-OS mode

Supports Mix and Match Current Linecards*

QSFP28 Connector, Pin compatible with 40G QSFP+

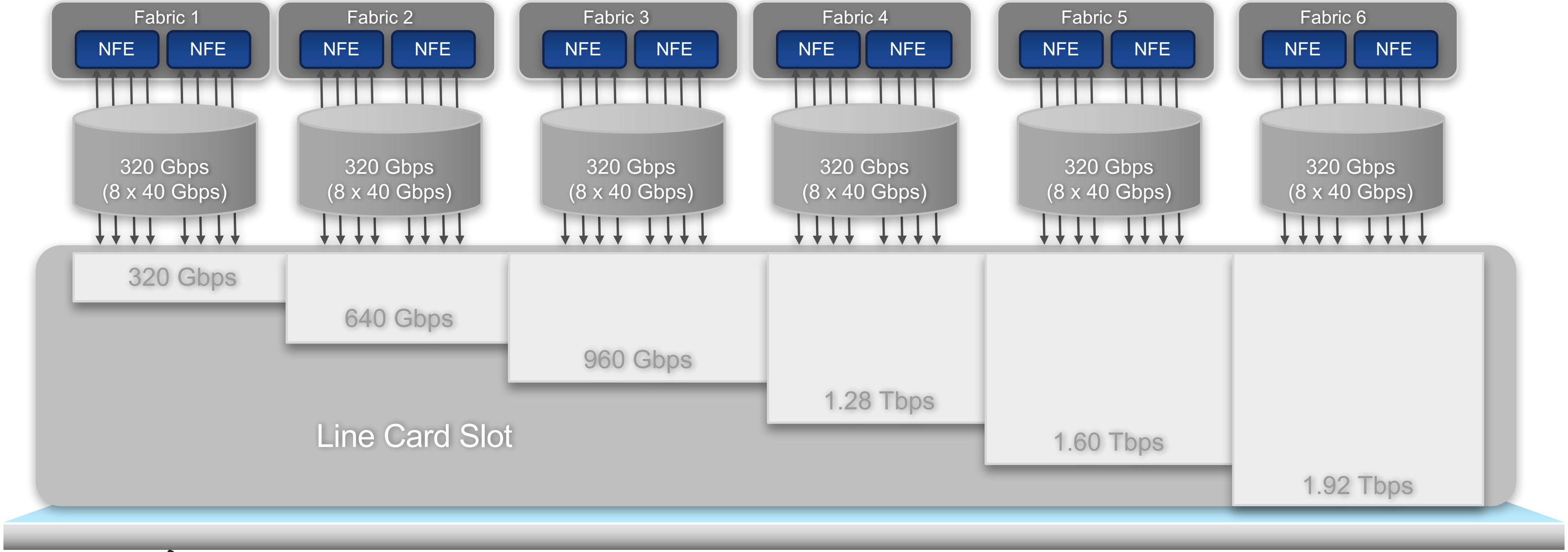
4, 8 and 16* Chassis

* future

First Gen Nexus 9500 Series Switch Fabric Module

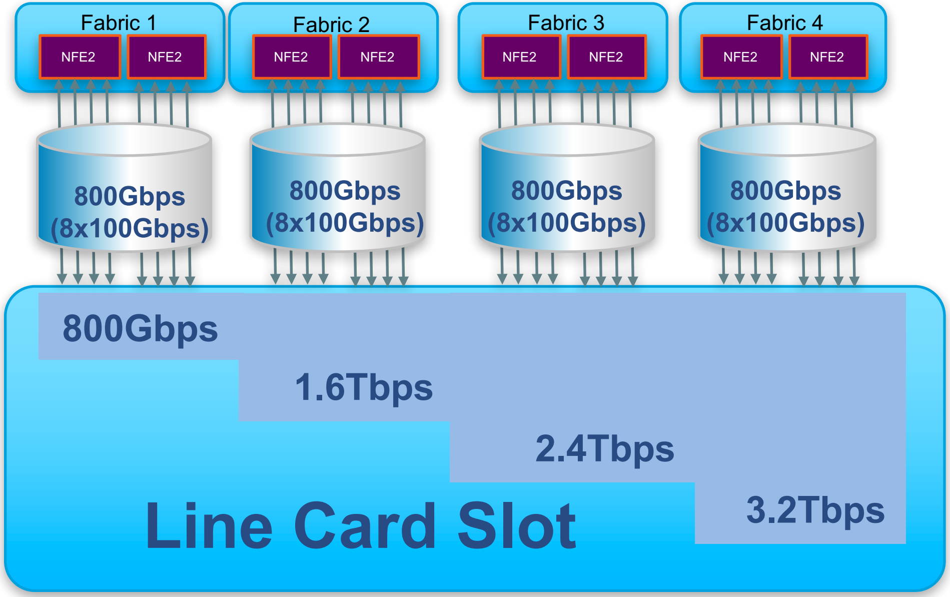
Data Plane Scaling (Using Nexus 9508 as an example)

- 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.



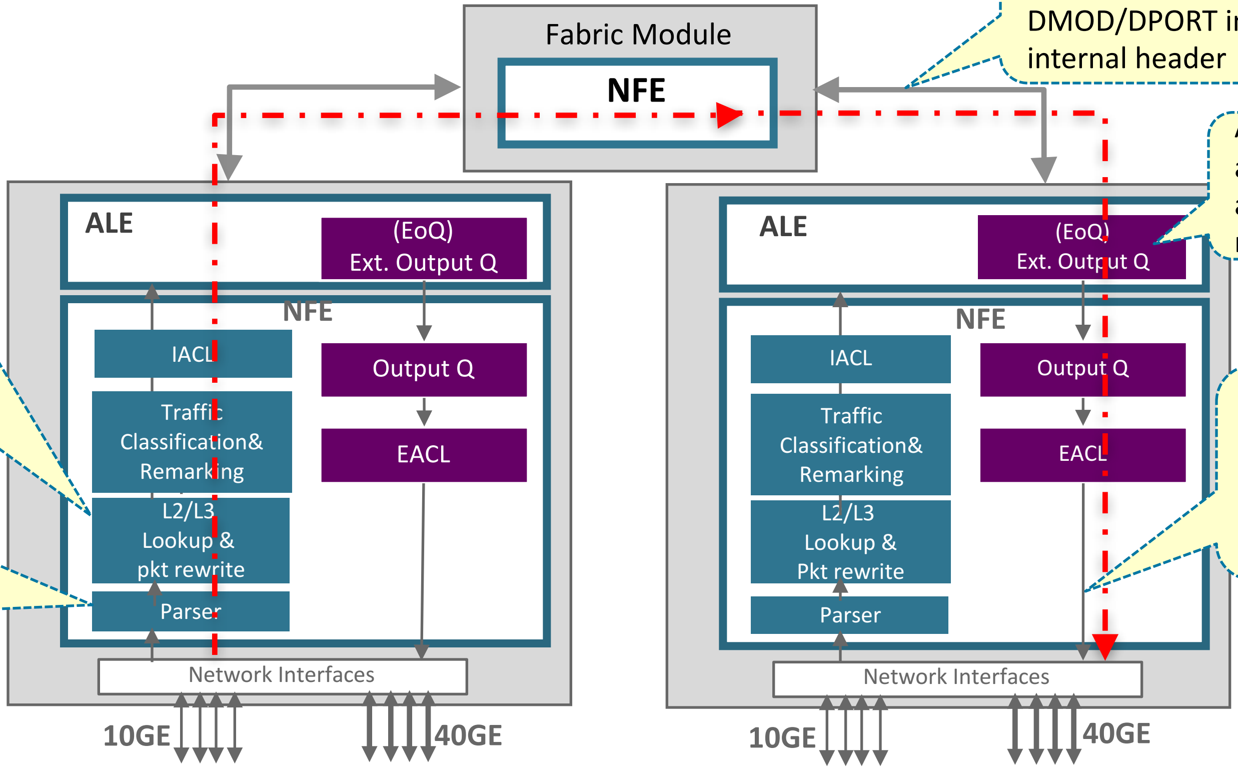
Second Gen Nexus 9500 Series Switch Fabric Module
Data Plane Scaling (Using Nexus 9508 as an example)

- With 4 Fabric Modules, each I/O module slot can have up to 3.2 Tbps forwarding bandwidth.



- N9K-C9504-FM-E
  - One NFE2 ASIC per FM
  - 32x100G ports per FM
- N9K-C9508-FM-E
  - Two NFE2 ASICs per FM
  - 64x100G ports per FM
- N9K-C9516-FM-E
  - Four NFE2 ASICs per FM
  - 128x100G ports per FM

Nexus 9500 Layer-2 Unicast Packet Walk



L2/L3 Lookup finds the dst MAC in Layer-2 MAC table and resolves DMOD/DPORT. Packet is sent to FM with DMOD/DPORT in the internal header

Examines ingress packet. Get packet headers for processing.

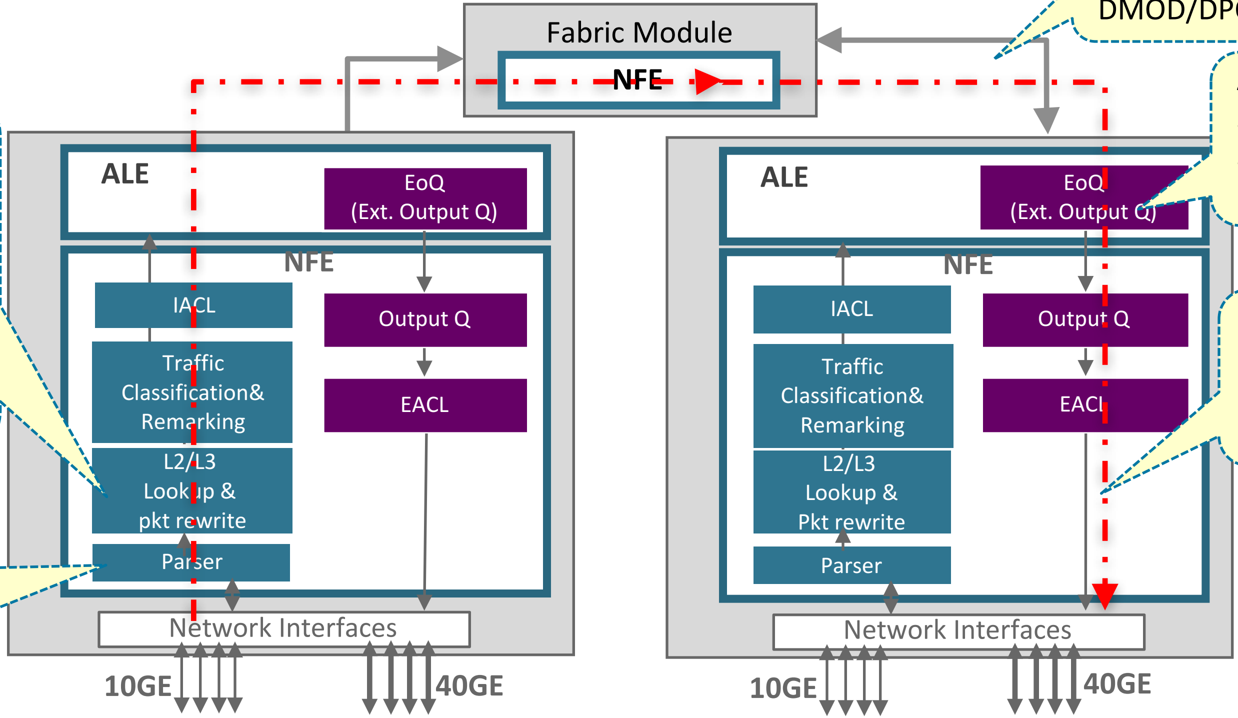
FM sends packet to egress LC based on DMOD/DPORT in the internal header

Additional buffer to absorb Microburst and Port Speed mismatch

Egress LC sends packet to the egress port based on DMOD/DPORT in the internal header

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

Nexus 9500 Layer-3 Host Unicast Packet Walk



L2/L3 Lookup finds the dst IP address in Layer-3 Host table, and resolves DMOD/DPORT. Packet is sent to Fabric Module with DMOD/DPORT in the internal header

Examines ingress packet. Get packet headers for processing.

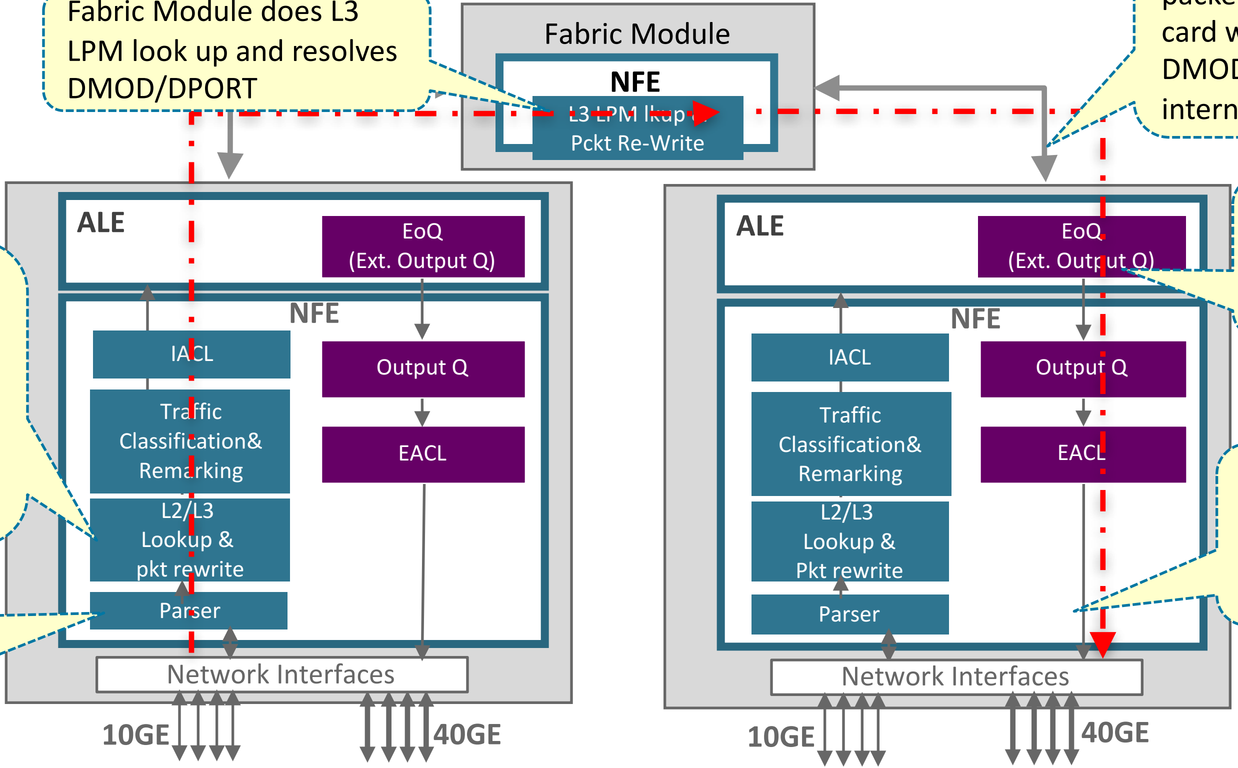
FM sends packet to egress LC based on DMOD/DPORT

Additional buffer to absorb Microburst and Port Speed mismatch

Egress LC sends packet to the egress port based on DMOD/DPORT

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

Nexus 9500 Layer-3 LPM Unicast Packet Walk



Fabric Module does L3 LPM look up and resolves DMOD/DPORT

Fabric Module sends packet to egress line card with DMOD/DPORT in the internal header

L2/L3 Lookup hits the default route. Packet is sent to Fabric Module using its virtual MOD ID as DMOD in the internal header.

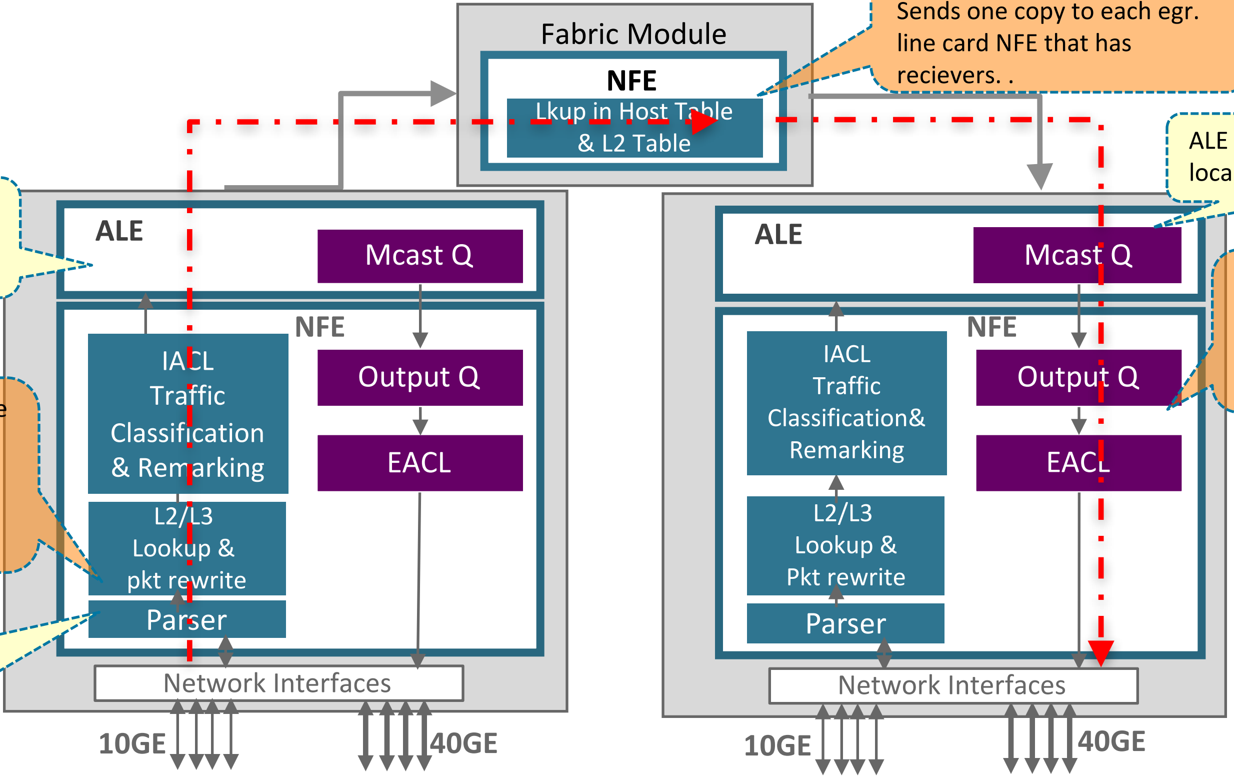
Additional buffer to absorb Microburst and Port Speed mismatch

Egress line card sends packet to egress port based on DMOD/DPORT

Examines ingress packet. Get packet headers for processing.

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

Nexus 9500 L2/L3 Multicast Packet Walk


Fabric Module
NFE
Lkup in Host Table & L2 Table
ALE
Mcast Q
NFE
IACL Traffic Classification & Remarking
Output Q
EACL
L2/L3 Lookup & pkt rewrite
Parser
Network Interfaces
10GE
40GE
ALE
Mcast Q
NFE
IACL Traffic Classification& Remarking
Output Q
EACL
L2/L3 Lookup & Pkt rewrite
Parser
Network Interfaces
10GE
40GE

ALE forward pckts to Fabric Module

L2/L3 mcast lkup; Replicate pckts to local receiving ports;
Send 1 copy to fabric module;

Examines ingress packet. Get packet headers for processing.

Lookup to resolve egr. Lin card NFE;
Sends one copy to each egr. line card NFE that has recievers. .

ALE forward pckts to local NFE

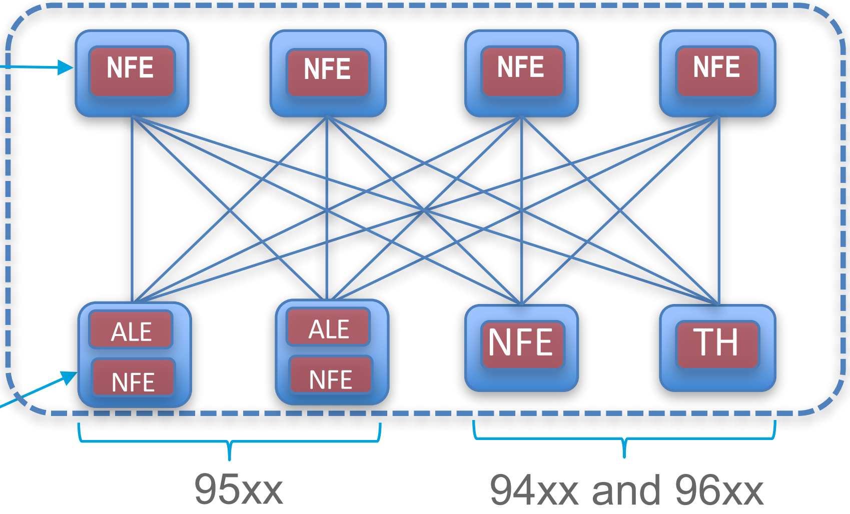
Lookup for local receiving ports; replicate pkts onto those ports.

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

Modular Nexus 9500
CLOS Based Hierarchical Forwarding

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

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

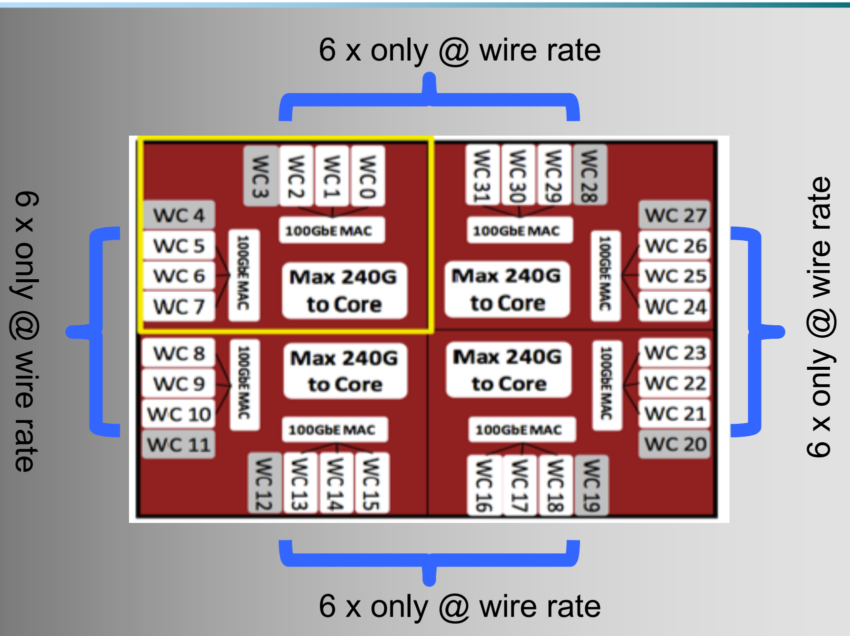


Network (LPM) Routes

Host Routes

In standalone NX-OS Mode Line Card Forwarding is performed in the NFE (Trident 2)

NFE Forwarding Capacity



- 24 of 32 ports are full line rate for all packet sizes

Two forwarding Modes on NFE

- Full Late-Rate Mode (FLM)
- Over-subscribed Mode (OSM)

Full Late-Rate Mode (FLM):

- Only 24 40GE ports are used
- Every port is full line-rate for all packet sizes

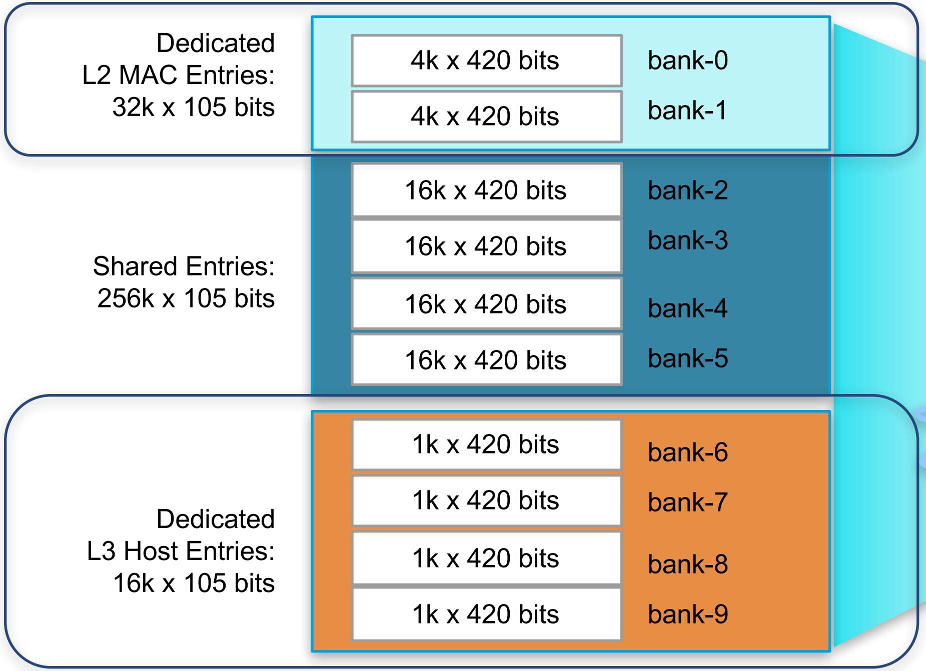
Over-subscribed Mode (OSM)

- All 32 40GE ports are used
- Every ports is line-rate for packet sizes > 193 Bytes

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



Dedicated
L2 MAC Entries:
32k x 105 bits

4k x 420 bits	bank-0
4k x 420 bits	bank-1

Shared Entries:
256k x 105 bits

16k x 420 bits	bank-2
16k x 420 bits	bank-3
16k x 420 bits	bank-4
16k x 420 bits	bank-5

Dedicated
L3 Host Entries:
16k x 105 bits

1k x 420 bits	bank-6
1k x 420 bits	bank-7
1k x 420 bits	bank-8
1k x 420 bits	bank-9

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

	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#
```

Nexus 9500 Merchant/Merchant+

BRCM T2

Fabric Module

- N9K-C9504-FM
- N9K-C9508-FM
- N9K-C9516-FM

Line Cards

- N9K-X9736PQ
- N9K-X9636PQ
- N9K-X9536PQ
- N9K-X9564PX
- N9K-X9564TX
- N9K-X9408PC-CPF2
- N9K-X9432PQ
- N9K-X9464PX
- N9K-X9464TX

BRCM Tomahawk

Fabric Module

- N9K-C9504-FM-S
- N9K-C9508-FM-S

Line Card

- N9K-X9432C-S

BRCM Jericho

Fabric Module

- N9K-C9508-FM-R

Line Cards

- N9K-X9636C-R
- N9K-X9636Q-R

Cisco CloudScale

Fabric Module

- N9K-C9504-FM-E
- N9K-C9508-FM-E

Line Cards

- N9K-X9732C-EX
- N9K-X97160YC-EX

Note: No mix match of different types of fabric modules in same chassis

Nexus 9500 – Modular
1/10/25/40/50/100G Capable

9500 Series

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)

- Analytics Ready
- Smart Buffer
- FX Support for MACSEC & Cloud-SEC

+

Cisco ASIC

16nm Technology

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

X9400-S (NX-OS)

- BCOM Trident and Tomahawk

+

Merchant ASIC

28 and 40nm Technology

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

X9600-R (NX-OS)

- Off Chip Buffer
- BCOM Jericho

+

Merchant ASIC

28nm Technology

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

Cisco Nexus 9636x-R Line Cards

Line Cards Comparison

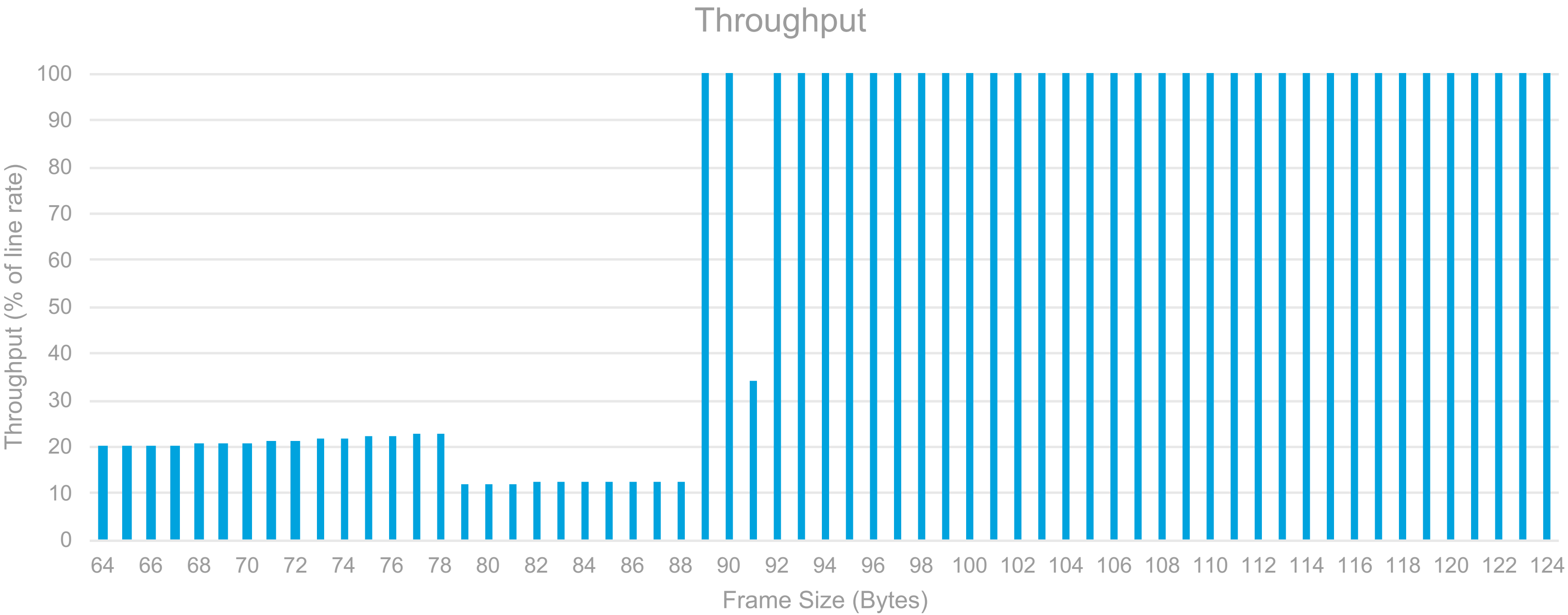
	N9K-X9636C-R	N9K-X9636Q-R
Maximum Number of 100 Gb Ports QSFP28	36	--
Maximum Number of 40 Gb Ports QSFP+	36	36
Line rate ports @ 64 bytes packets	24 ports @100 Gbps or 36 ports @ 40 Gbps	36 ports @ 40 Gbps
Line rate ports @ > 64 bytes packets	36 ports @100 Gbps or 36 ports @ 40 Gbps	36 ports @ 40 Gbps
Total Module Capacity	2.4 Tbps @ 64 bytes 3.6 Tbps @ > 92 bytes	1.4 Tbps @ 64 bytes 1.4 Tbps @ > 64 bytes
Ports per Jericho	6	12
Minimum Number of Fabric Modules for Full Line-Rate Performance	5	4
N9K-SUP-B Required	Yes	Yes

Cisco Nexus 9636C-R Line Cards

N9K-X9636C-R

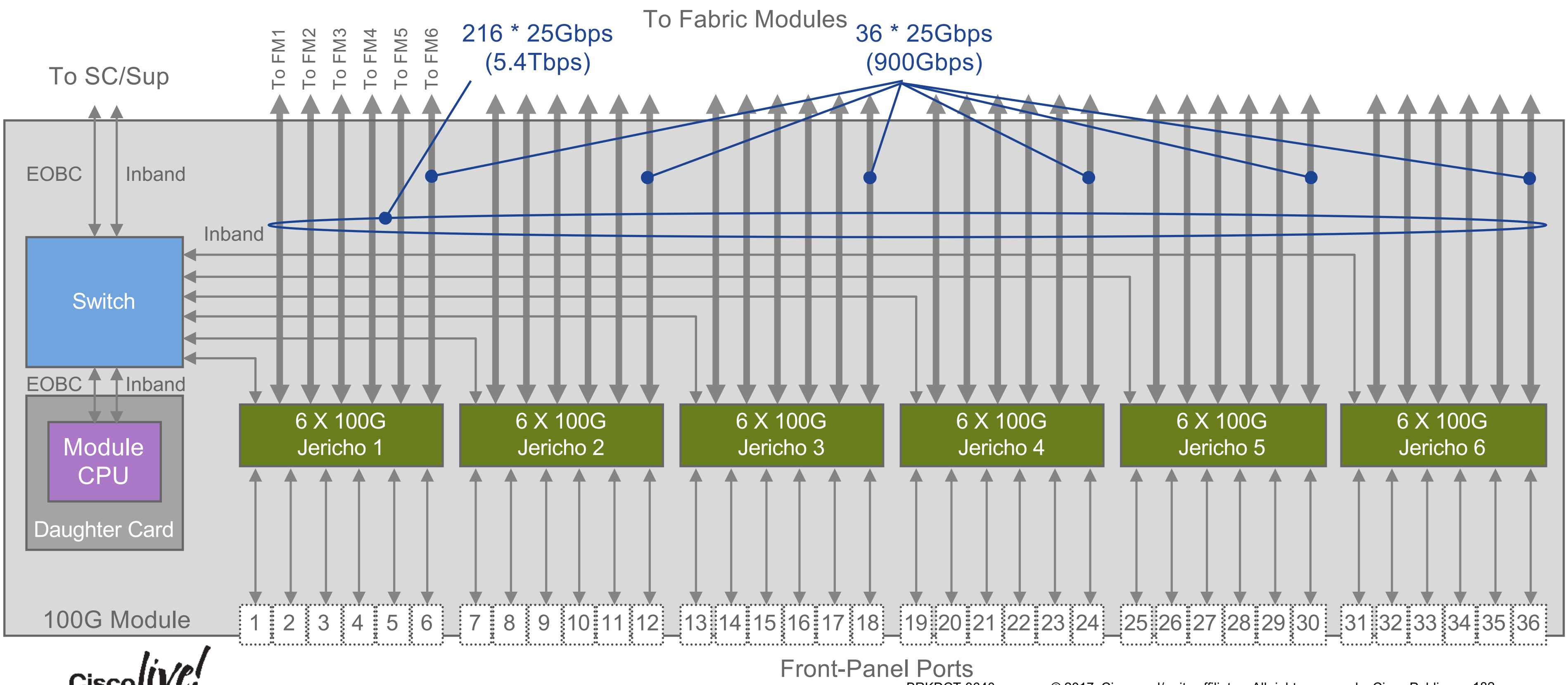
- 36 x 100G QSFP28 ports
  - Can also operate as 40G ports with 40G QSFP
- Six forwarding ASICs, one per 6 front-panel 100G ports (4GB GDDR5 DRAM-based buffer per ASIC)
- Up to 36 line rate ports at larger packet sizes (higher than 92B)
  - 3.6Tbps total module capacity
- 24 line rate 100G ports at 64 bytes
  - 4 ports per Jericho (total bandwidth 480Gbps per Jericho)
- 8-core 2.4GHz module x64 module CPU with 16GB DDR3 DRAM
- Requires N9K-SUP-B and N9K-9508-FM-R

Cisco Nexus 9636C-R Line Cards - Throughput



Cisco Nexus 9636C-R Line Card

N9K-X9636C-R Module Architecture

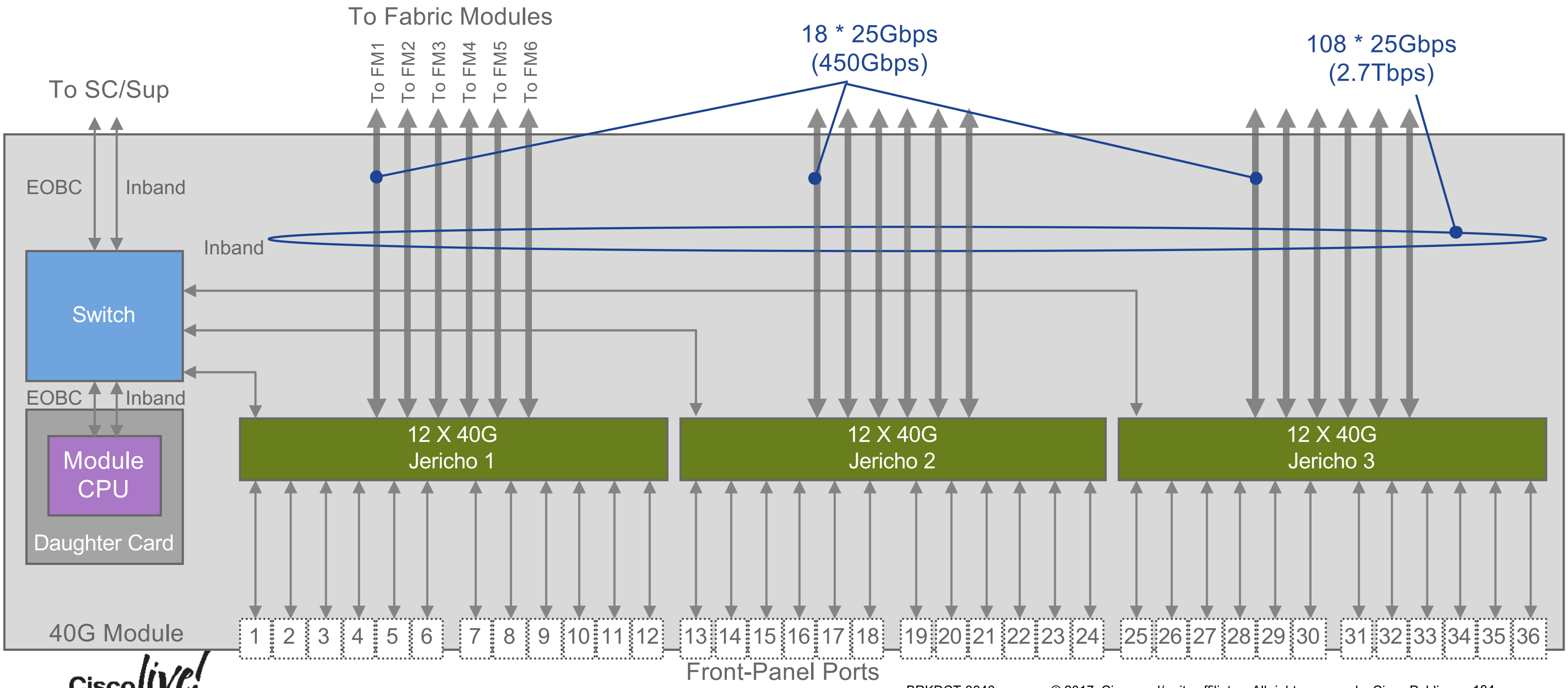


Cisco Nexus 9636Q-R Line Cards
N9K-X9636Q-R

- 36 x 40G QSFP+ ports
- Three forwarding ASICs, one per 12 front-panel 40G ports
- 4GB GDDR5 DRAM-based buffer per ASIC
- 36 line rate 40G ports at all packet sizes
    - 1.4Tbps total module capacity
- 8-core 2.4GHz module x64 module CPU with 16GB DDR3 DRAM
- Requires N9K-SUP-B and N9K-9508-FM-R

Cisco Nexus 9636Q-R Line Cards

N9K-X9636Q-R(Potenza-40) Module Archicture

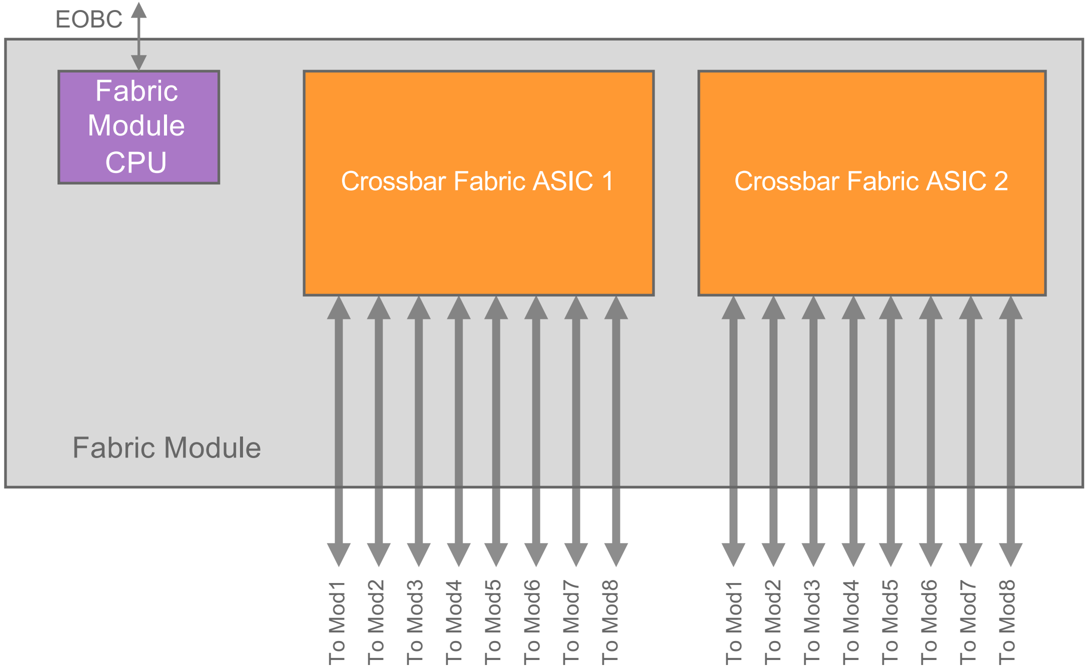


Nexus 9508-FM-R Fabric Module

- Up to 6 fabric modules per system
- Two crossbar ASICs per fabric module
- Provides cell-based fabric interconnect between I/O modules
  - Variable length cells (64..256 bytes)
- Each fabric module provides 900G bandwidth per I/O module slot
- 900Gbps * 6 fabrics = 5.4Tbps/slot
- 8-slot chassis: 5.4Tbps/slot * 8 I/O modules = 43.2Tbps total system bandwidth
- 4-core ARM CPU with 4GB DDR3 DRAM

Nexus 9508-FM-R 8-Slot Fabric Module

Architecture

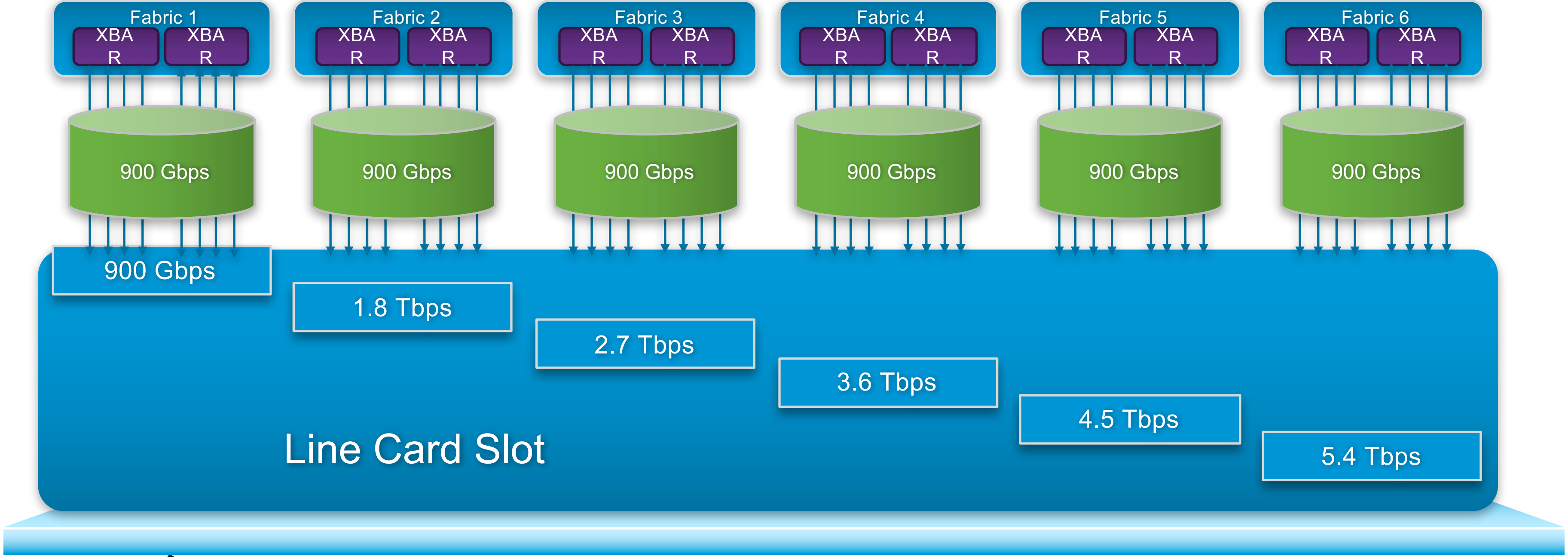


To I/O Modules

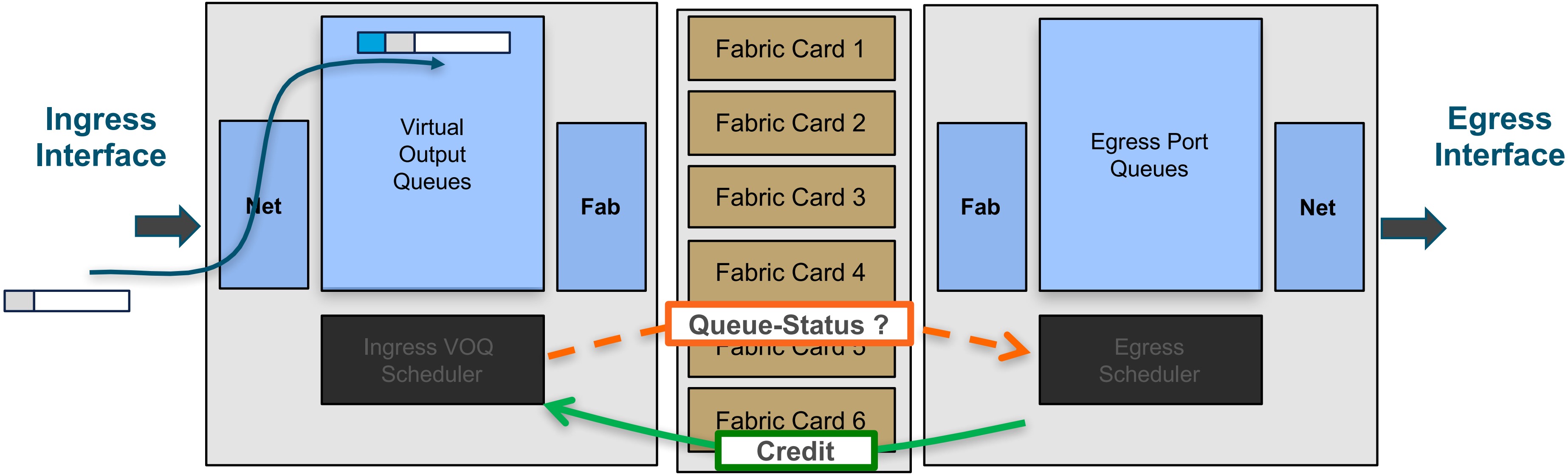
Nexus 9508-FM-R Fabric Module

Data Plane Scaling For 8-Slot Chassis

- Each fabric module for the 8-slot chassis can provide up to 900 Gbps to each I/O module slot
- With 6 fabric modules, each I/O module slot can have up to 5.4Tbps forwarding bandwidth in each direction

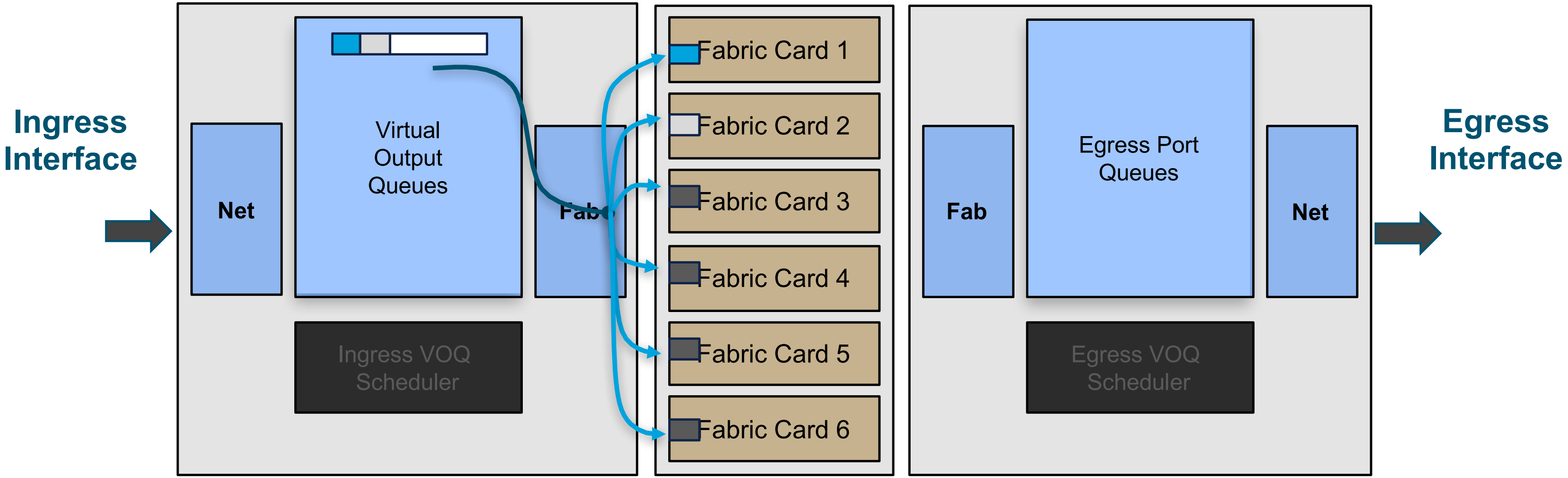


VOQ Architecture



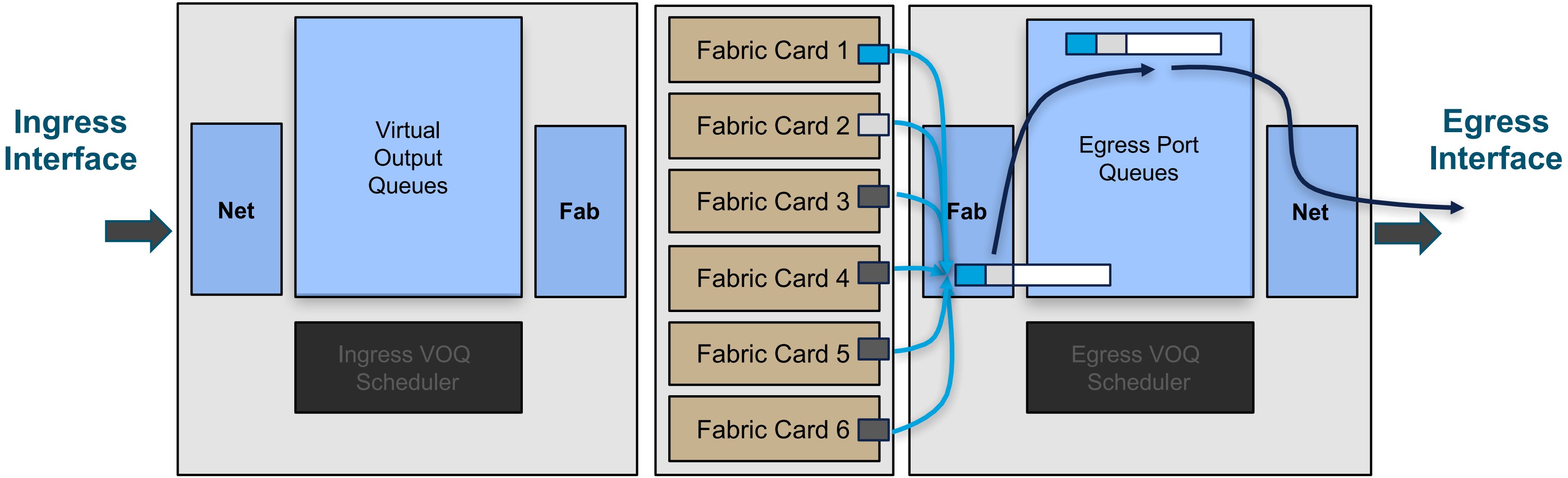
- Packet is received on ingress interface, classified, and stored in internal buffer
- Ingress VOQ scheduler polls Egress scheduler (maintaining a local VOQ DB)
- Egress answers with a credit-message

VOQ Architecture



- Packet is split in cells and load balanced among the fabric cards
- Cells are transported to the egress line card

VOQ Architecture



- Cells are collected and packet re-assembled
- Packet is stored in the port queue
- Finally packet is transmitted through the egress interface

Cisco*live!*

Forwarding Tables

Q4CY16 Supported

Forwarding Table	System Scale
IPv4 prefixes	192K (shared with v6 prefixes/hosts)
IPv4 ARP/host routes (/32)	750K (shared with MAC)
IPv6 prefixes/host routes	62K (shared with v4 prefixes)
IPv4 multicast routes	32K (shared with ACL)
Adjacency table (rewrite table)	80K directly connected / system
IPv4/IPv6 ACL entries	48K / system (spread over multiple instances) (max IPv4/IPv6 per instance: 48K/12K)
MAC table	64K at FCS (shared with IPv4 host routes)

Nexus 9500 Merchant Off Chip Buffer

BRCM T2

Fabric Module

- N9K-C9504-FM
- N9K-C9508-FM
- N9K-C9516-FM

Line Cards

- N9K-X9736PQ
- N9K-X9636PQ
- N9K-X9536PQ
- N9K-X9564PX
- N9K-X9564TX
- N9K-X9408PC-CPF2
- N9K-X9432PQ
- N9K-X9464PX
- N9K-X9464TX

BRCM Tomahawk

Fabric Module

- N9K-C9504-FM-S
- N9K-C9508-FM-S

Line Card

- N9K-X9432C-S

BRCM Jericho

Fabric Module

- N9K-C9508-FM-R

Line Cards

- N9K-X9636C-R
- N9K-X9636Q-R

Cisco CloudScale

Fabric Module

- N9K-C9504-FM-E
- N9K-C9508-FM-E

Line Cards

- N9K-X9732C-EX
- N9K-X97160YC-EX

Note: No mix match of different types of fabric modules in same chassis

Nexus 9500 – Modular
1/10/25/40/50/100G Capable

9500 Series

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)

- Analytics Ready
- Smart Buffer
- FX Support for MACSEC & Cloud-SEC

+

Cisco ASIC

16nm Technology

Fabric Module

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

X9400-S (NX-OS)

- BCOM Trident and Tomahawk

+

Merchant ASIC

28 and 40nm Technology

Fabric Module

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

X9600-R (NX-OS)

- Off Chip Buffer
- BCOM Jericho

+

Merchant ASIC

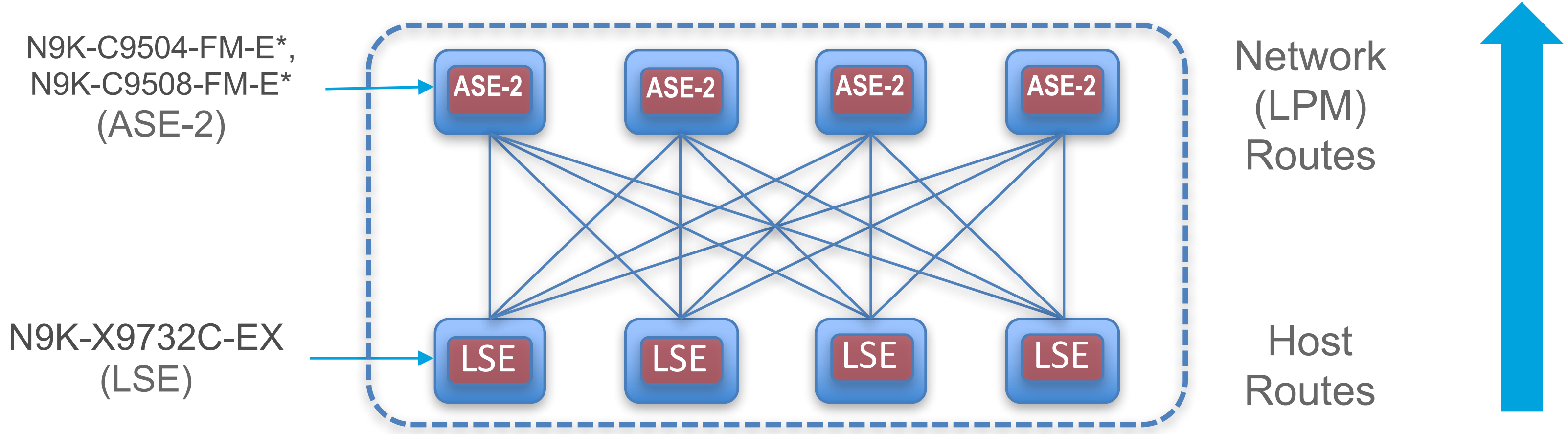
28nm Technology

Fabric Module

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

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

Nexus 9500 Forwarding Programming Mode

Generation 2 Line Cards and FM’s

Default template

Table Type	IPv4 Hosts	IPv4 LPM	IPv6 Hosts	IPv6 /64 LPM	MAC	Multicast	Next_Hop	IPv4 MPLS
Scale	**768K***	**768K***	16K	256K	96K	32K	64K	16K
Location	LC	LC	FM	FM	LC	LC and FM	LC + FM	LC

* shared entry. Non-/64 IPv6 routes in separate TCAM

High Host route and LPM Scale with Multicast

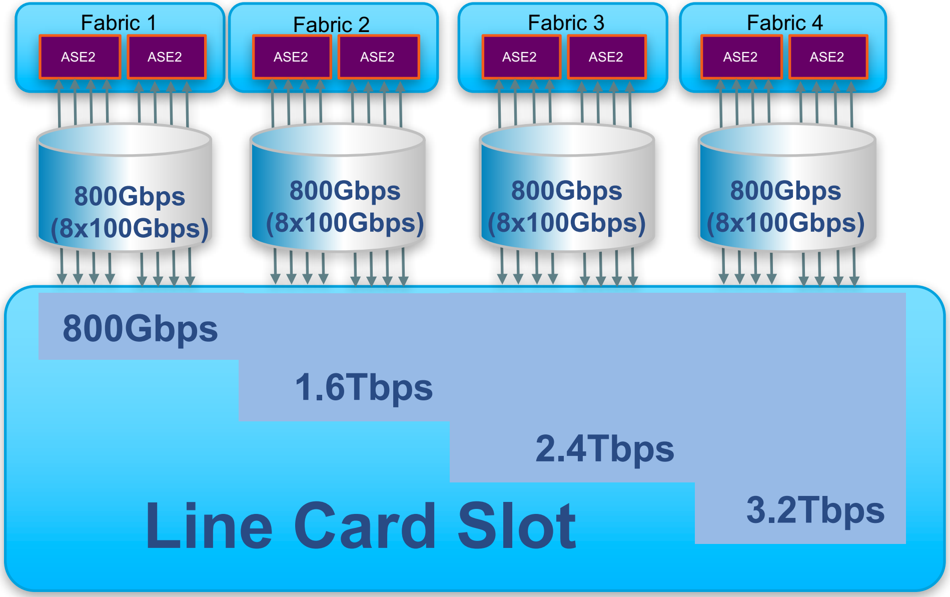
Table Type	IPv4 Hosts	IPv4 LPM	IPv6 Hosts	IPv6 /64 LPM	MAC	Multicast	Next_hop	IPv4 MPLS
Scale	**1M***	**1M***	16K	256K	16K	32K	64K	16K
Location	LC	LC	FM	FM	LC	LC + FM	LC + FM	LC

* shared entry. Non-/64 IPv6 routes in separate TCAM table

Second Gen Nexus 9500 Series Switch Fabric Module

Data Plane Scaling (Using Nexus 9508 as an example)

- With 4 Fabric Modules, each I/O module slot can have up to 3.2 Tbps forwarding bandwidth.



- N9K-C9504-FM-E
  - One ASE2 ASIC per FM
  - 32x100G ports per FM
- N9K-C9508-FM-E
  - Two ASE2 ASICs per FM
  - 64x100G ports per FM
- N9K-C9516-FM-E
  - Four ASE2 ASICs per FM
  - 128x100G ports per FM

Nexus 9500 Series Line Cards – Cisco ASICs

Deployment Options: Aggregation, Spine

ACI

ACI Spine

40G

X9736PQ: 36p 40G QSFP+

ALE

97xx Series Line Cards

NX-OS & ACI

Access/Aggregation/Spine

10/25/40/50/100G

X9732C-EX: 36p 40/100G QSFP

+ ACI-DCI, Buffer & Analytics

LSE

97xx-EX Series Line Cards

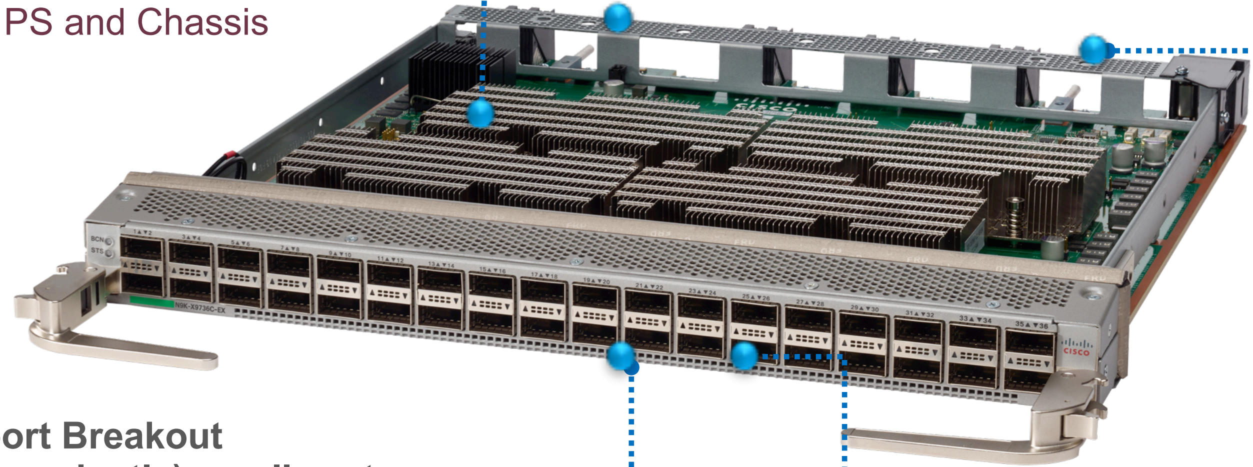
Nexus 9500 N9K-X9732C-EX

LSE Based

Supported in ACI and NX-OS mode

N9K-X9732C-EX line card needs 4 fabric modules to operate at full line rate on all 32 ports. Line Rate for all packet size.

Investment Protection with Supervisors, System Controller, PS and Chassis



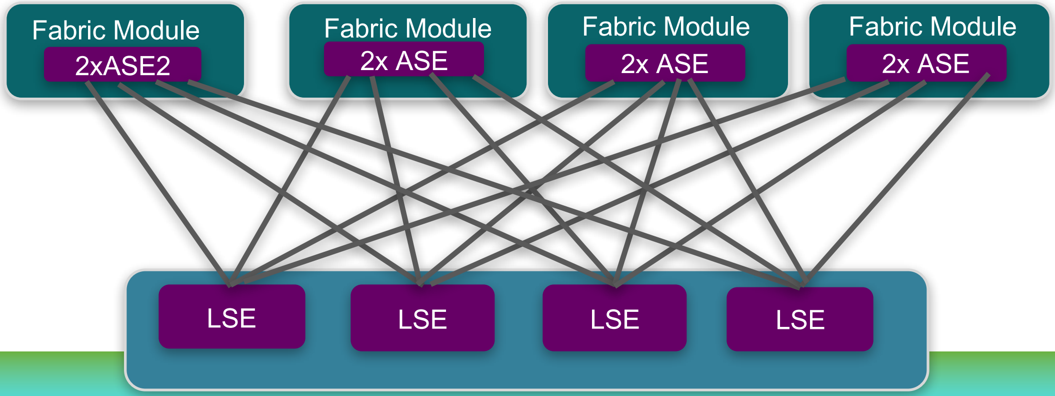
Support Breakout (independently) on all ports

QSFP28 Connector, Pin compatible with 40G QSFP+

Ports Modes:
4x10G,4x25G,40G,2x50G,100G

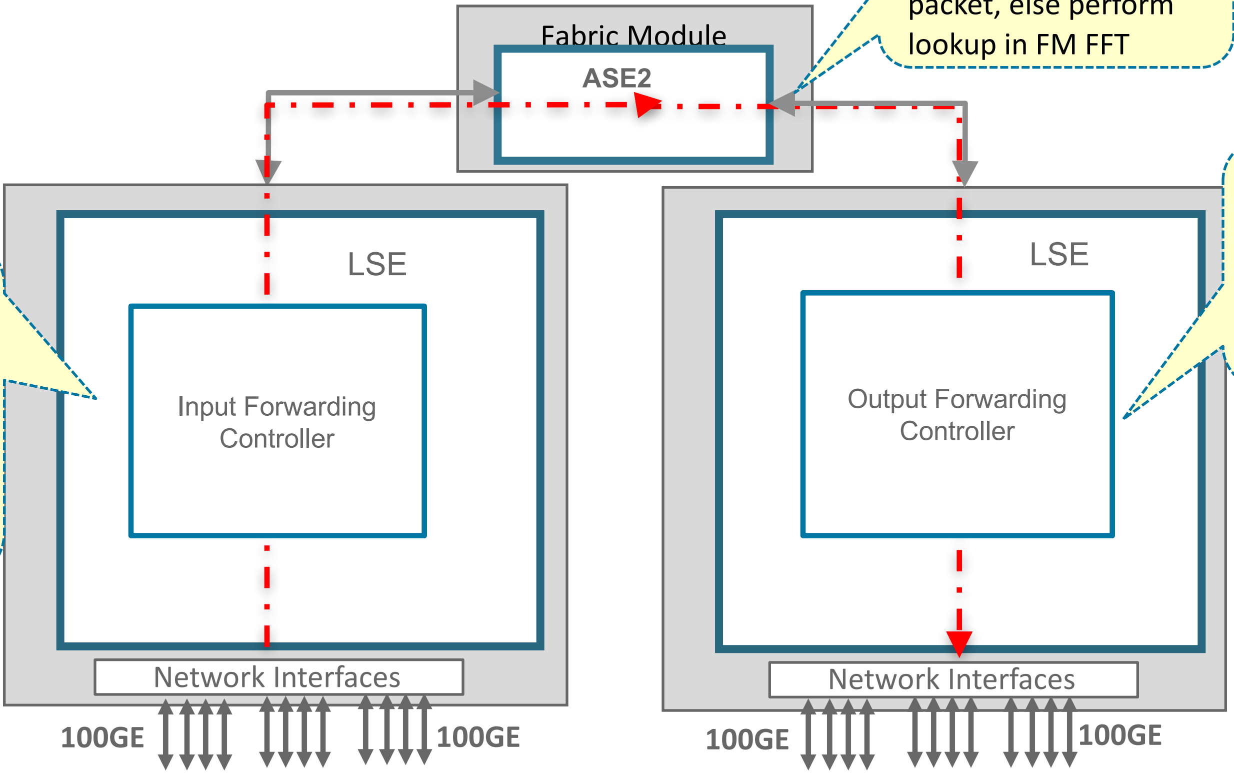
Nexus 9500 N9K-X9732C-EX Line Card

- N9K-X9732C-EX Fabric Connectivity with N9K-C9508-FM-E Fabric Module



- Needs 4 fabric modules (fabric module slot 2, 3, 4 and 6)
- Each LSE provides 8 x 100 Gbps front-panel ports and 8 x 100 Gbps internal links to the fabric modules
- Line rate for packet sizes

L2/L3 Unicast Packet Walk

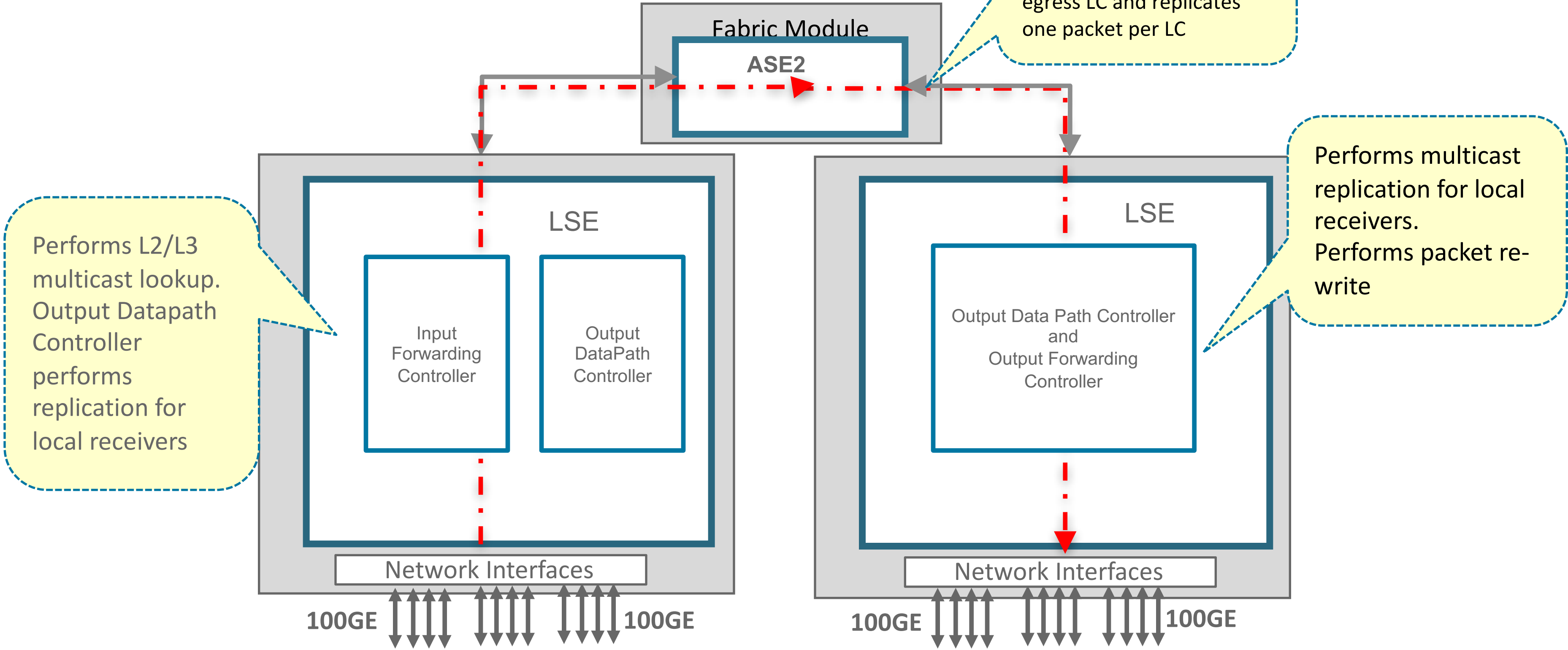


If ingress LC has lookup results, FM simply transfers packet, else perform lookup in FM FFT

Header Parser, DMAC lookup for L2/L3 forwarding. With FFT LC can have host or/and LPM route. 64B Internal header added to pass metadata

Gets input packet and associated metadata. Extracts internal header and performs packet re-write

Multicast Packet Walk


Fabric Module
ASE2
LSE
Input Forwarding Controller
Output DataPath Controller
Network Interfaces
100GE
100GE
LSE
Output Data Path Controller and Output Forwarding Controller
Network Interfaces
100GE
100GE

Fabric Module performs one more lookup to determine receivers on egress LC and replicates one packet per LC

Performs L2/L3 multicast lookup. Output Datapath Controller performs replication for local receivers

Performs multicast replication for local receivers.
Performs packet re-write

Nexus 9500 Fabric Modules

BRCM T2

Fabric Module

- N9K-C9504-FM
- N9K-C9508-FM
- N9K-C9516-FM

Line Cards

- N9K-X9736PQ
- N9K-X9636PQ
- N9K-X9536PQ
- N9K-X9564PX
- N9K-X9564TX
- N9K-X9408PC-CPF2
- N9K-X9432PQ
- N9K-X9464PX
- N9K-X9464TX

BRCM Tomahawk

Fabric Module

- N9K-C9504-FM-S
- N9K-C9508-FM-S

Line Card

- N9K-X9432C-S

BRCM Jericho

Fabric Module

- N9K-C9508-FM-R

Line Cards

- N9K-X9636C-R
- N9K-X9636Q-R

Cisco CloudScale

Fabric Module

- N9K-C9504-FM-E
- N9K-C9508-FM-E

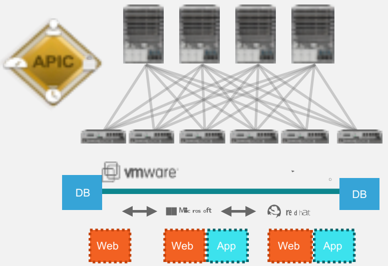
Line Cards

- N9K-X9732C-EX
- N9K-X97160YC-EX

Note: No mix match of different types of fabric modules in same chassis

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

Connection
Creation
Expansion
Reporting
Fault Mgmt
VTS

VxLAN-BGP EVPN standard-based

3rd party controller support

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

Programmable Network

rpm
CHEF
ANSIBLE
python
puppet labs

Modern NX-OS with enhanced NX-APIs

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

Nexus 9400 & 9600 (line cards), 9200, 3100, 3200

Nexus 9700EX + 9300EX

Q & A

Complete Your Online Session Evaluation

Give us your feedback and receive a **Cisco Live 2017 Cap** by completing the overall event evaluation and 5 session evaluations.

All evaluations can be completed via the Cisco Live Mobile App.

Caps can be collected Friday 10 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

Cisco live!

Thank you

Cisco live!

7–10 March 2017 • Melbourne, Australia