
cisco
Cisco live!
July 10–14, 2016 · Las Vegas, NV
Your Time Is Now

Cisco Nexus 9000 Architecture

Eddie Tan - Distinguished Engineer

BRKARC-2222

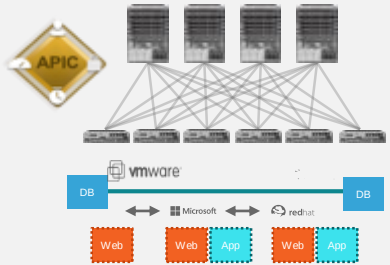
Cisco live!

Agenda

- What's New
  - 2nd Generation Nexus 9000
  - Moore's Law
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9500 (Modular)
  - Nexus 9200/9300 (Fixed)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- 40G/100G Transceiver
  - 25G technology

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

Application Centric Infrastructure

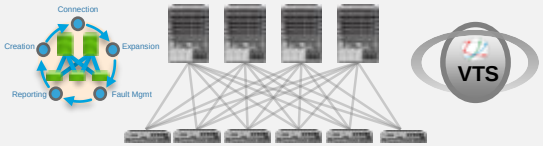


Turnkey integrated solution with security, centralized management, compliance and scale

Automated application centric-policy model with embedded security

Broad and deep ecosystem

Programmable Fabric

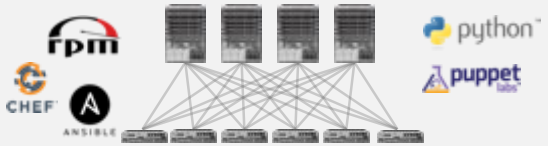


VxLAN-BGP EVPN standard-based

3rd party controller support

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

Programmable Network



Modern NX-OS with enhanced NX-APIs

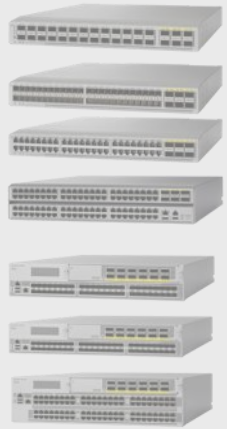
DevOps toolset used for Network Management
(Puppet, Chef, Ansible etc.)

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

Nexus 9700EX + 9300EX

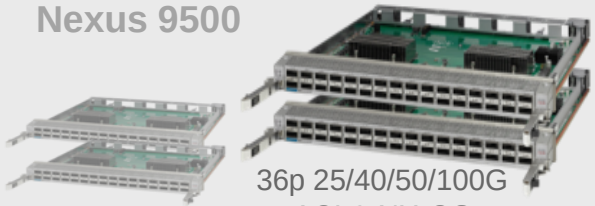
Nexus 9000 Portfolio
10/25/40/50/100G on Merchant or Cisco Silicon

Nexus 9300



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

Nexus 9500



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

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

Existing Chassis Delivering on Investment Protection Promise

Continued Support for All Existing Line Cards Merchant and Cisco Silicon

Nexus 9504 Nexus 9508 Nexus 9516

Nexus 9300EX

Industry Only 25G Native VXLAN

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

Nexus 9200

Industry Only 25G Native VXLAN

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

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

Nexus 3100

64p 40G

32p 40G

48p 10G & 6p 40G

48p 1G & 4p 10G

Nexus 3100V

32p 40G

48p 10G & 6p 100G

VXLAN routing, 100G uplinks, No 25G

T2+

Nexus 3200

32p 25/50/100G

64p 40G Single Chip

VXLAN bridging, **25/100G**

Tomahawk

Shipping for 3+ months

Single NX-OS Image for Nexus 3000 & Nexus 9000

Agenda

- What's New
  - 2nd Generation Nexus 9000
  - Moore's Law
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9500 (Modular)
  - Nexus 9200/9300 (Fixed)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- 40G/100G Transceiver
  - 25G technology

Gordon Moore

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

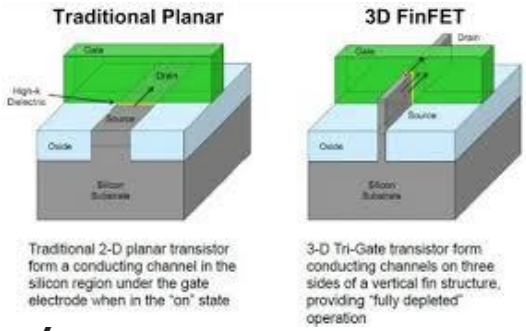
"Moore's Law" - 1975

Moore's Law

It's all about the Economics

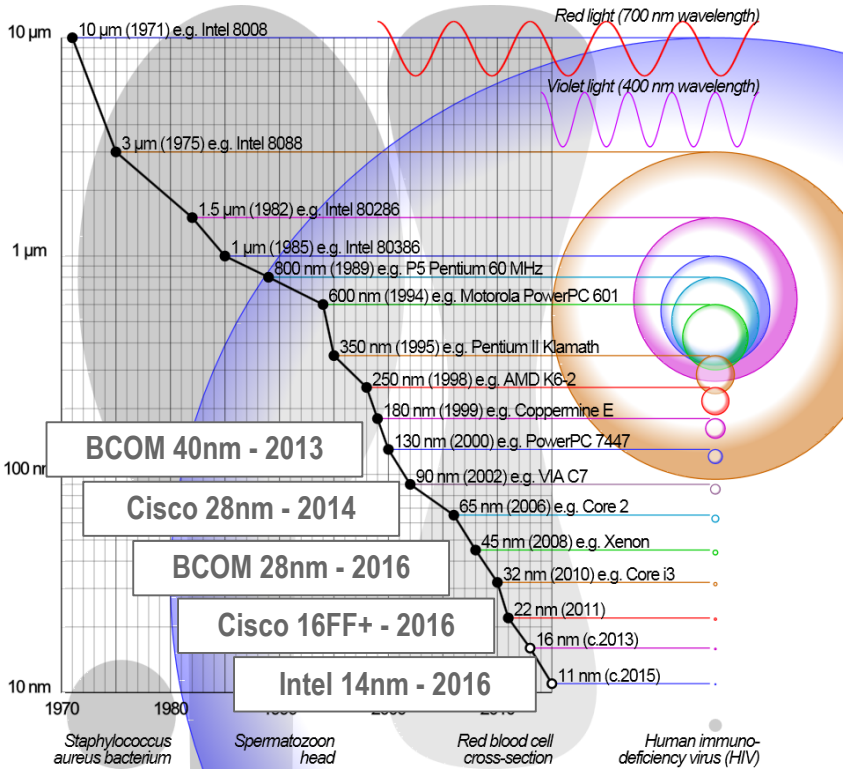
- Increased function, efficiency
- Reduced costs, power
- ~ 1.6 x increase in gates between process nodes

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



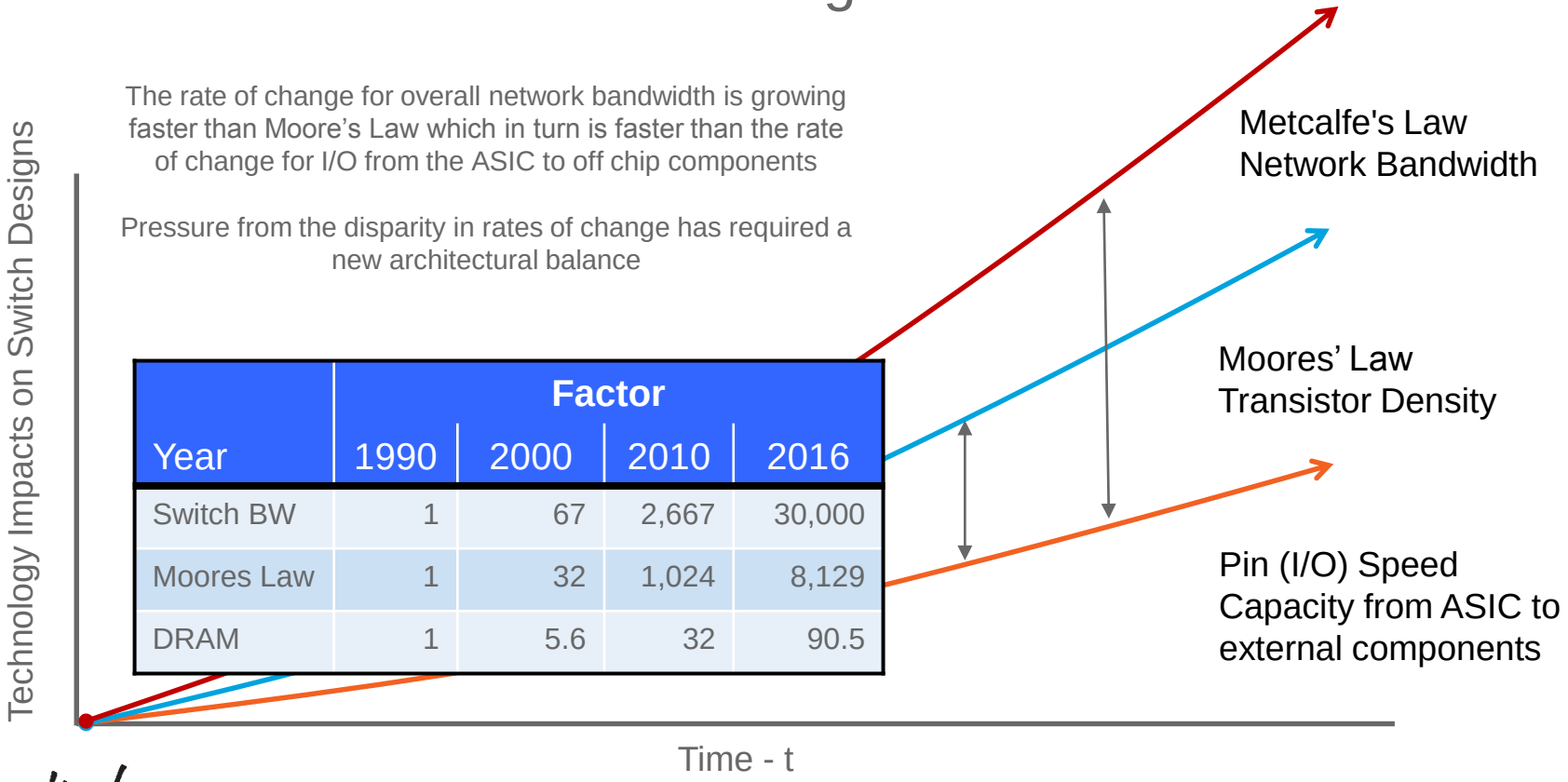
http://en.wikipedia.org/wiki/Semiconductor_device_fabrication

Metcalfe, Moore and ASIC Pin I/O Rates
The Switch Architectural Challenge

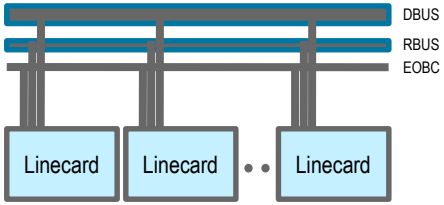
The rate of change for overall network bandwidth is growing faster than Moore's Law which in turn is faster than the rate of change for I/O from the ASIC to off chip components

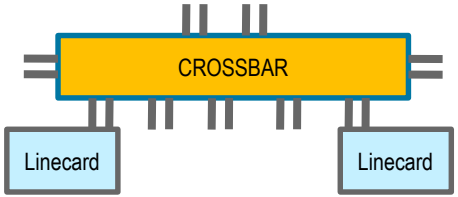
Pressure from the disparity in rates of change has required a new architectural balance

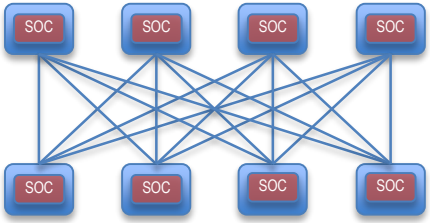
	Factor			
Year	1990	2000	2010	2016
Switch BW	1	67	2,667	30,000
Moores Law	1	32	1,024	8,129
DRAM	1	5.6	32	90.5



Switching Architecture Changes
Shifting of Internal Architecture





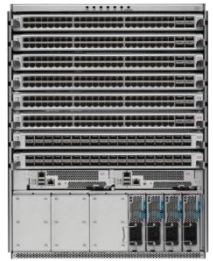


Design Shifts Resulting from Increasing Gate Density and Bandwidth

10/100M

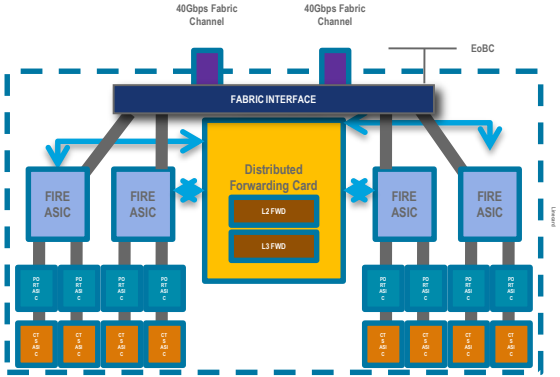
100M/1G

1G/10G

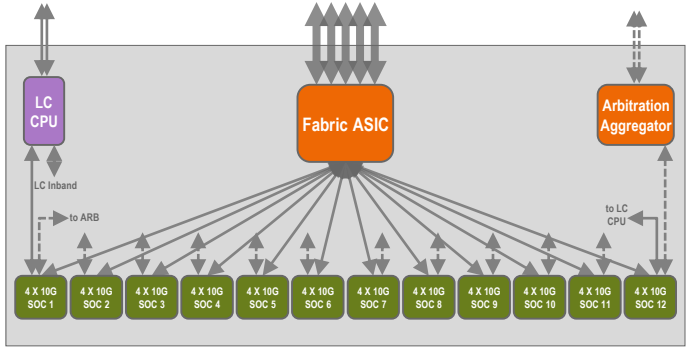


10G/100G

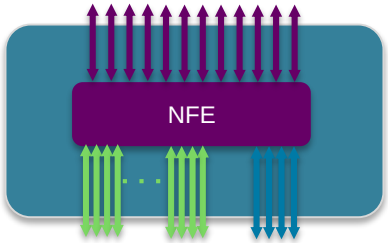
Switching Architecture Changes
Consolidation of Functions onto fewer components



32 x 10G Ports

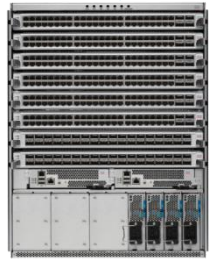


48 x 10G Ports



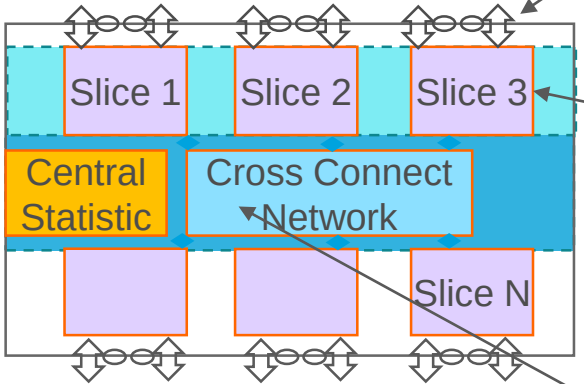
64 x 10G Ports

Design Shifts Resulting from Increasing Gate Density and Bandwidth



Switch On Chip (SOC)

It is a full multi-stage switch on an ASIC



- 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

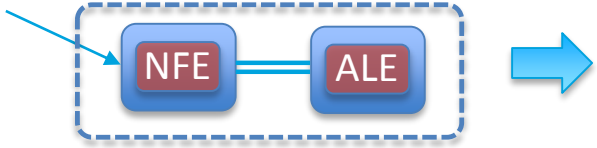
- Multi-mode MAC
- Packet parser
- Forwarding controller
- Input packet buffering for pause
- Output packet buffering
- Buffer accounting
- Output queuing and scheduling
- Output Rewrite

- Gen2 PCIe controller for register and eDMA access
- Cross connect 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

Nexus 9372E

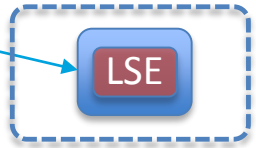
Switch on Chip (SOC)



Leverages Merchant (BCOM) + Cisco

Fixed Second Generation Nexus 9200 & 9300EX
A Single ASIC based Switch

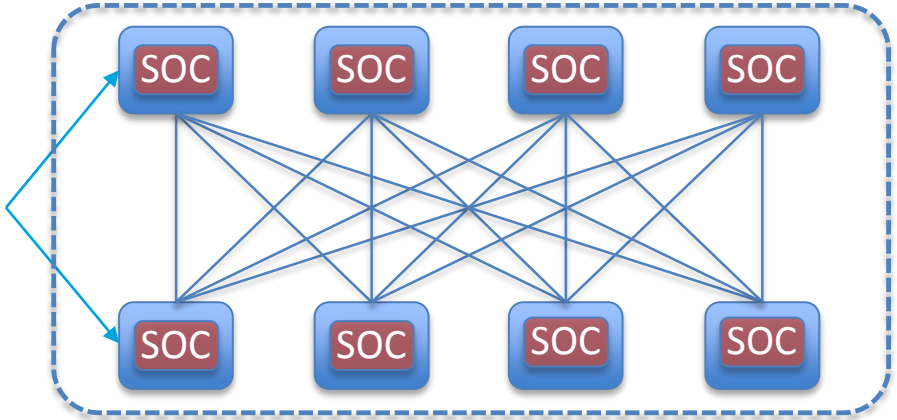
Switch on Chip
(SOC)

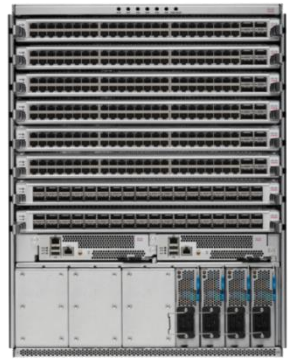


The Switch ‘is’ the ASIC

Modular Nexus 9500
A CLOS Based SOC Architecture

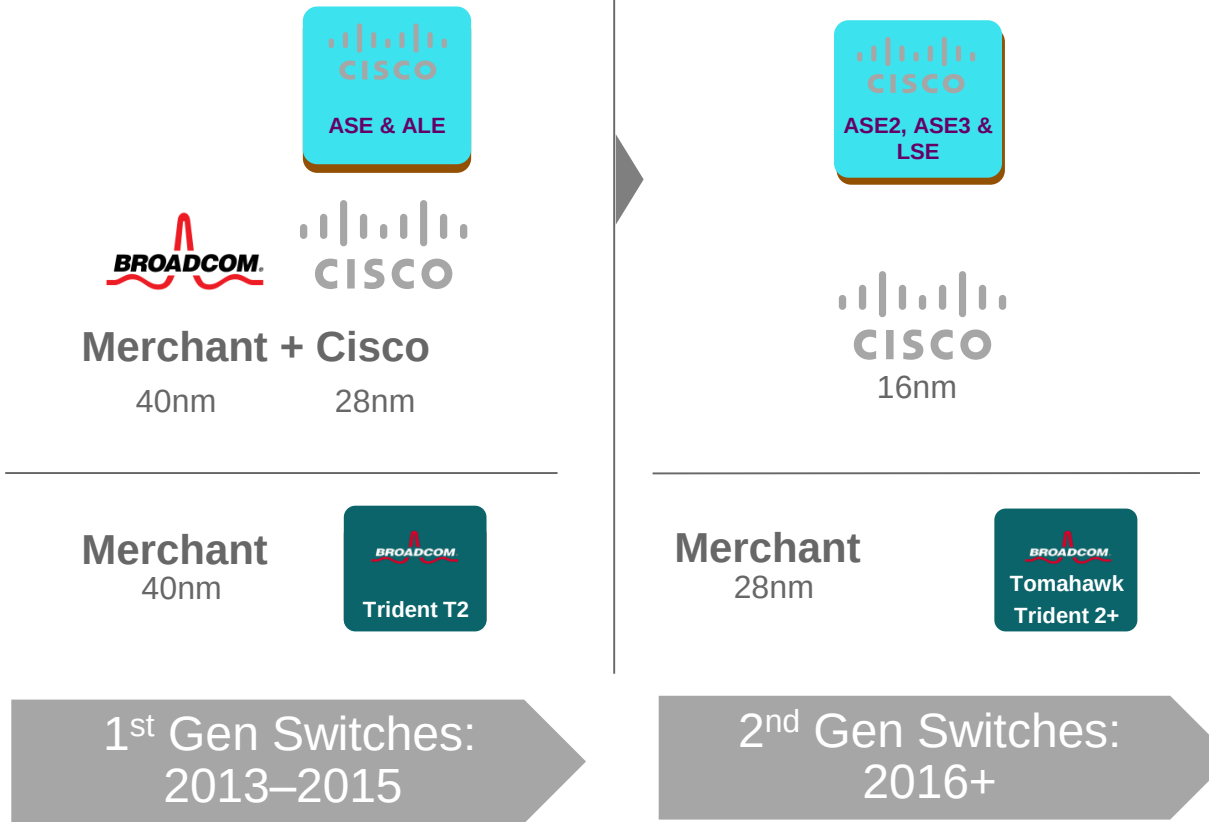
Leverage Switch on Chip (SOC) Based components





Non Blocking Leaf and Spine based CLOS Network inside the Switch

ASIC Used by Nexus 3000/9000



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

Telemetry
- Analytics
- Netflow
- Atomic Counters

Optimization
- Smart Buffers
- DLB/ Flow Prioritization

ASIC Used by 2nd Gen. 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 Prioritization & Elephant-Trap for trapping 5 tuple of large flows

ASIC Used by 2nd Gen N9000

16nm

- LSE – Leaf Spine Engine
- Standalone leaf & spine, ACI leaf and spine
- Flow Table (Netflow, …)
- ACI feature and service and security enhancement
- FabricPath
- 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 Prioritization, Elephant-Trap for trapping 5 tuple of large flows

2nd Gen. N9K ASIC Summary

	CISCO LSE	CISCO ASE-2	CISCO ASE-3
Capacity/Performance	1.8Tbps	3.6Tbps	1.6Tbps
Use Case	ACI/NX-OS TOR/Leaf Line card for N9500	Fabric Module for N9500 High density 40/100G NX-OS TOR	Cost effective TOR
Platforms built with	N93180C-EX N93108TC-EX N93180YC-EX X9732C-EX	N9K-C9504-FM-E N9K-C9508-FM-E N9236C N9272Q N92304QC	N92160YC-X
Netflow/Tetration HW Sensor	Yes	No	Yes
VXLAN Routing	Line rate	Line rate	Line rate
Intelligent buffering	Yes	Yes	Yes
Flexible Forwarding Table	Yes	Yes	Yes

ASIC Used by Nexus 3000/9000

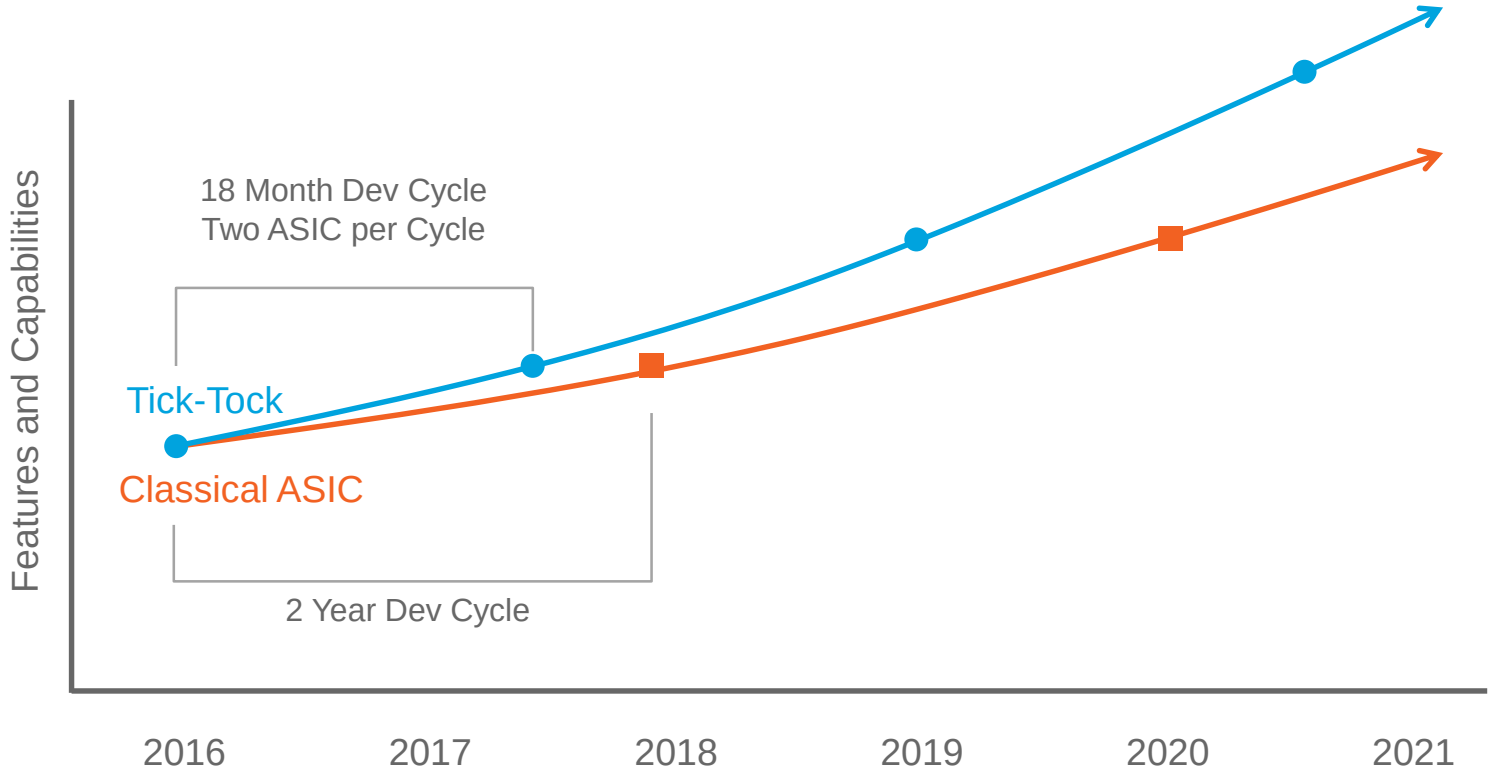
Merchant
28nm

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

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

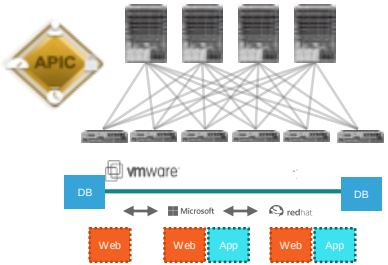
- Broadcom Trident 2+
- 1.28Tbps I/O & 0.96T Core (< 192B pkt)
  - 32 x 40GE (line rate for 24 x 40G)
- Standalone leaf and spine
- VXLAN Bridging & Routing (with-out recirculation)

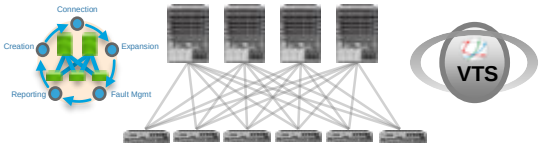
Development Cycle Decreasing
Time to Leverage Moore's Law is Reducing



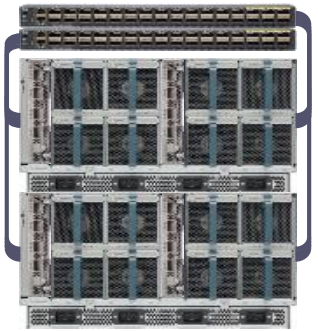
Responding to Fast Market Changes
Sharing Platforms Among Different Architectures

- Common hardware platforms for ACI and NX-OS fabric



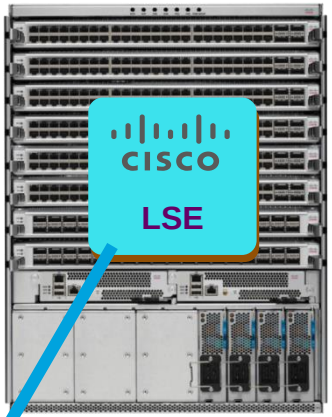


- Sharing platform with UCS FI
  - 3rd Generation FI is based on first gen 9300
  - 4th Generation FI will be based on 2nd Generation 9300EX

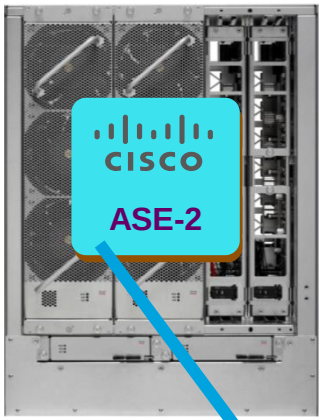


Responding to Fast Market Changes
Sharing ASICs Among Platforms

9732C-EX



N93180YC-EX

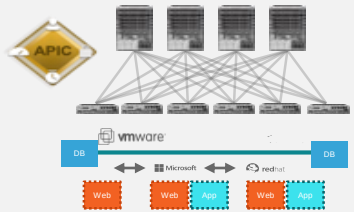


N9K-C9504/8/16-FM-E

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

VxLAN-BGP EVPN standard-based

3rd party controller support

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

Programmable Network

Modern NX-OS with enhanced NX-APIs

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

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

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

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
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9500 (Modular)
  - Nexus 9200/9300 (Fixed)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- 40G/100G Transceiver
  - 25G technology

Nexus 9500 – Modular

9500 Series

9504 9508 9516

Existing 4-, 8-, 16- slot chassis

No mid-plane to update

Power and cooling within existing shipping system profile

Existing shipping Power Supply, Supervisor and System Controllers

X9700-EX (NX-OS and ACI)

32p 100G QSFP Line card
- 10/25/40/50/100G
- Analytics Readiness

\+

Cisco ASIC

16nm Technology

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

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

X9400-S (NX-OS)

32p 100G QSFP Line card
- 10/25/40/50/100G

\+

Merchant ASIC

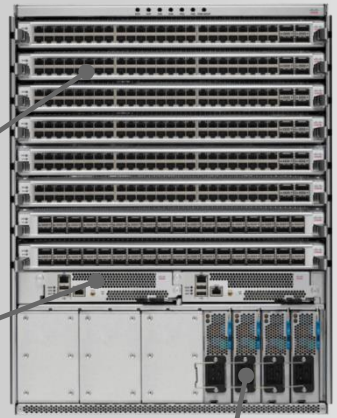
28 and 40nm Technology

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

1/10/25/40/50/100G Capable

Nexus 9500 Platform Architecture

Nexus® 9508 Front View

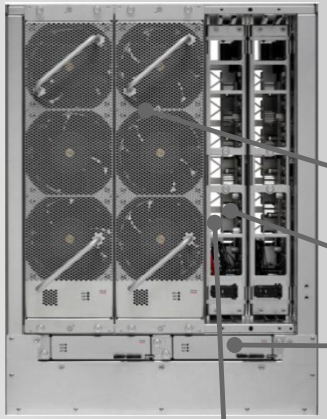


8 line card slots
Max 3.84 Tbps per slot
duplex

Redundant
supervisor engines

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

Nexus 9508 Rear View



3 fan trays, front-to-back airflow

fabric modules
(behind fan trays)

Redundant system controller cards

No mid-plane for
LC-to-FM connectivity

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

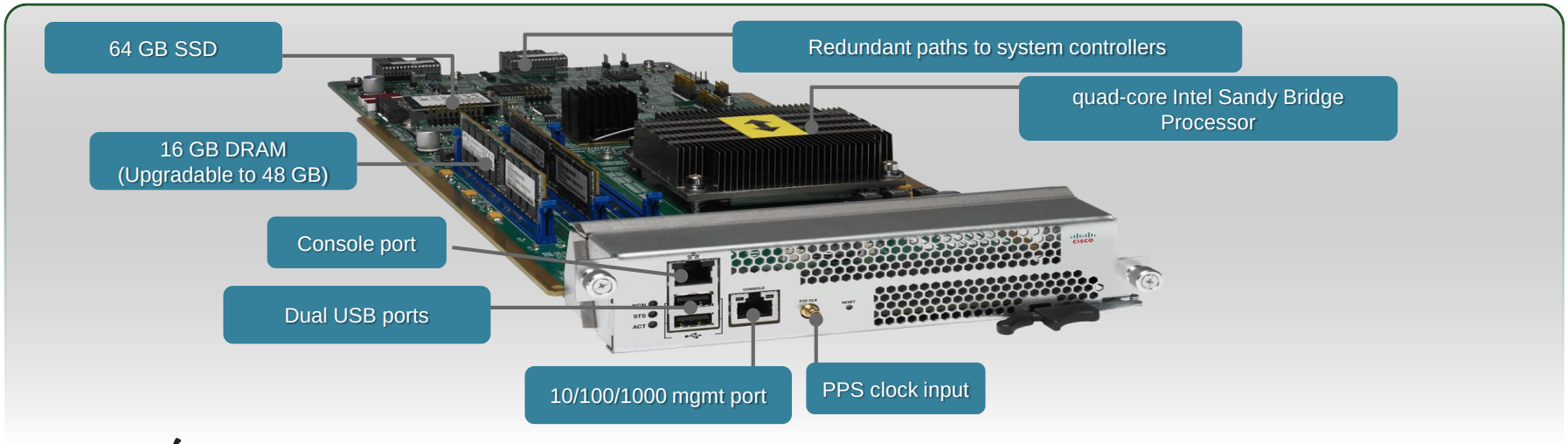
Designed for Power and Cooling Efficiency
Designed for Reliability
Designed for Future Scale

Nexus 9500 Platform Architecture

Supervisor Module Sup-A

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

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

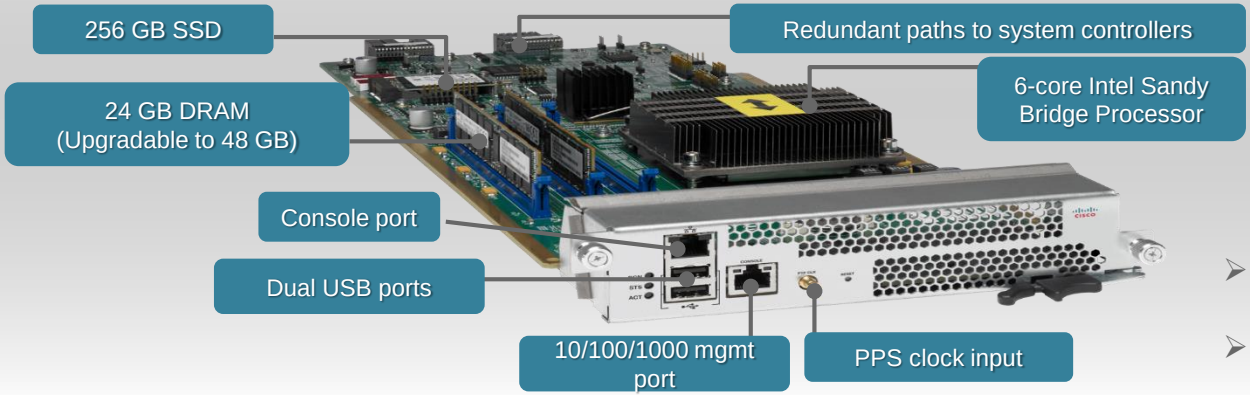


Nexus 9500 Platform Architecture

Supervisor Module Sup-B

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

Supervisor Module	
Processor	2.1 GHz, 6 cores 2.2GHz IVY Bridge
System Memory	24 GB, upgradable to 48 GB
RS-232 Serial Ports	One (RJ-45)
10/100/1000 Management Ports	One (RJ-45)
USB 2.0 Interface	Two
SSD Storage	256 GB



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

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

Nexus 9500 Platform Architecture

System Controller 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

Universal 3000W Power Supply

High Voltage AC and DC
180-305 VAC, 192-400 VDC

Universal AC and DC Supply

NX-OS/ACI*

New S-Grid safety cables for high voltage deployments

Resiliency with N+N, N+1

Chassis Port side Air Intake

Nexus 9500 Platform Architecture
Fan Tray

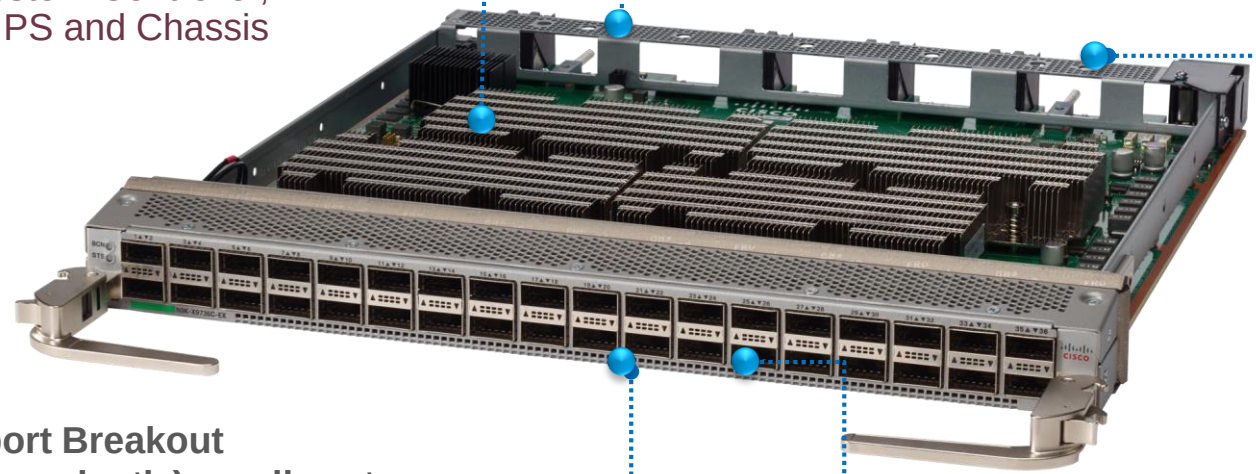
- 3 fan trays
  - 3 dual fans per tray
  - Dynamic speed control driven by temperature sensors
  - Straight airflow across line cards and fabric modules
  - If one fan tray is removed, the other two fan trays will speed up 100% to compensate for the loss of cooling power
- N+1 Redundancy per tray

Nexus 9500 N9K-X9732C-EX
LSE Based

Investment Protection with Supervisors, System Controller, PS and Chassis

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.



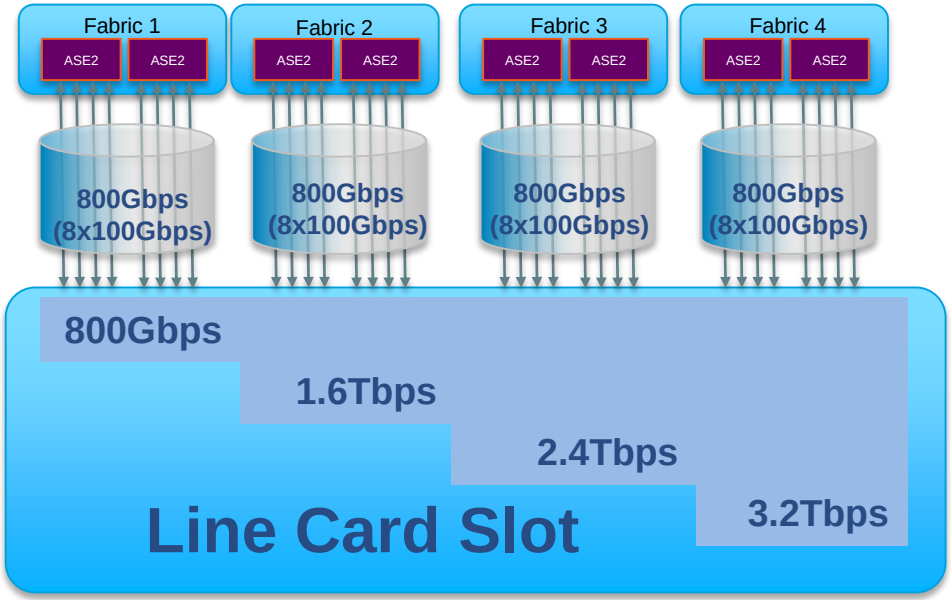
Support Breakout (independently) on all ports

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

QSFP28 Connector, Pin compatible with 40G QSFP+

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

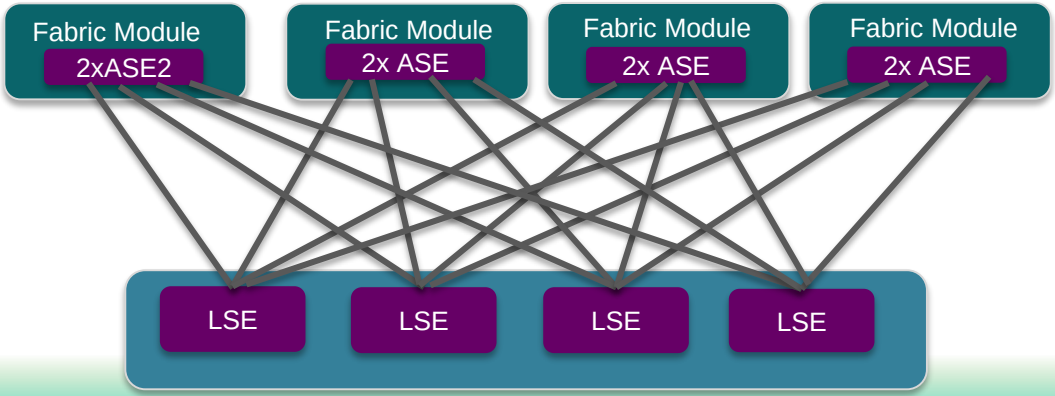
- With 4 Fabric Modules, each I/O module slot can have up to 3.2 Tbps forwarding bandwidth.



- 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

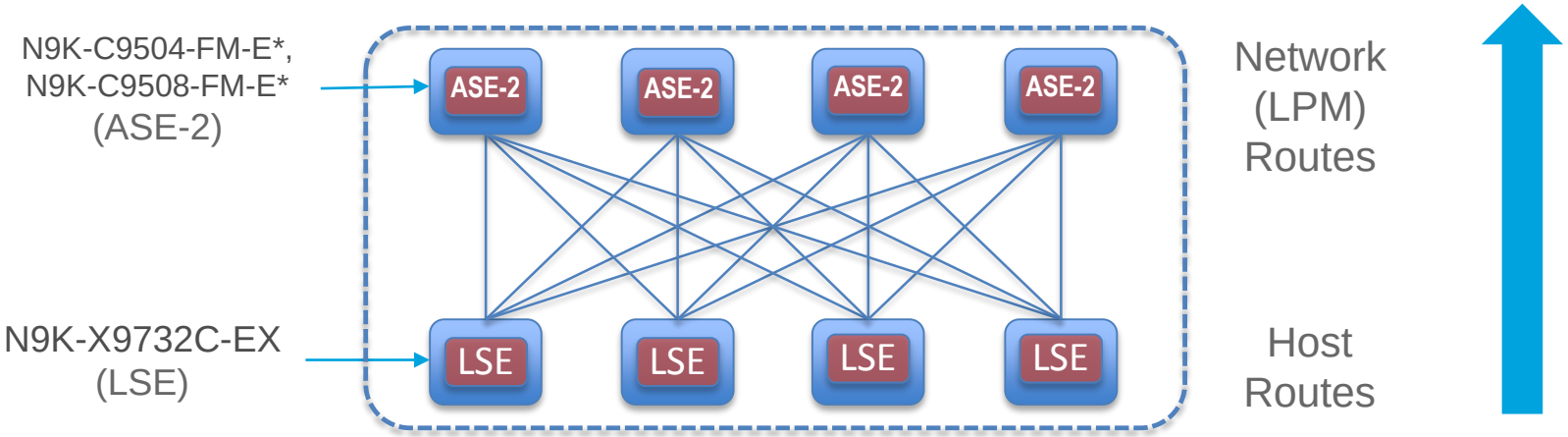
Nexus 9500 N9K-X9732C-EX Line Card

- 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

Modular Nexus 9500
Generation 2 Line Cards and Fabric Modules



Summarization and Balance of Host and Network Routes Shift

Nexus 9500 Series Line Cards – Cisco ASICs
Deployment Options: Aggregation, Spine

ACI

ACI Spine

40G
X9736PQ: 36p 40G QSFP+

CISCO **ASE**

97xx Series Line Cards

NX-OS & ACI

Enhanced Agg/ Spine

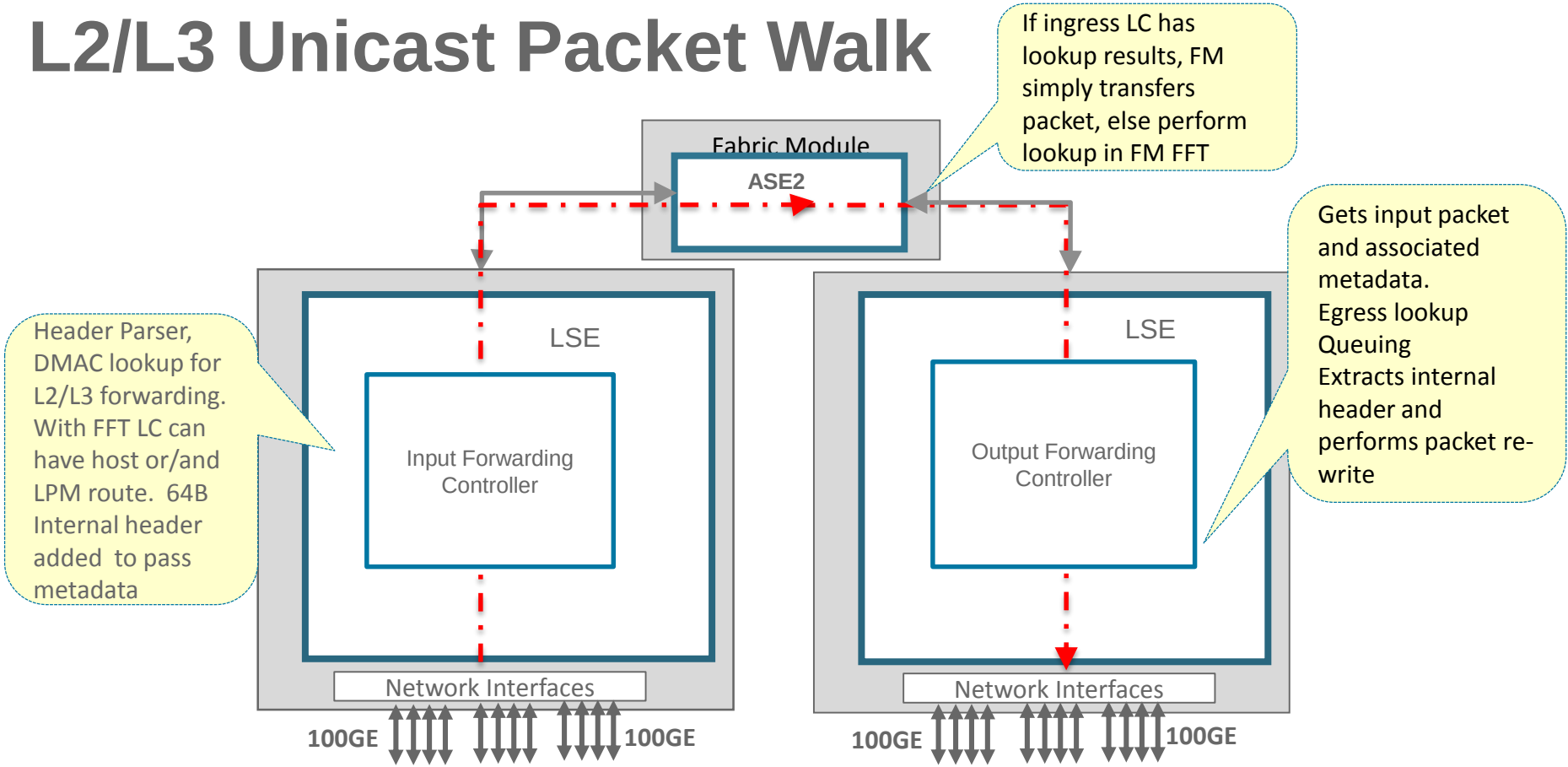
100G
X9732C-EX: 32p 40/100G QSFP

CISCO **LSE**

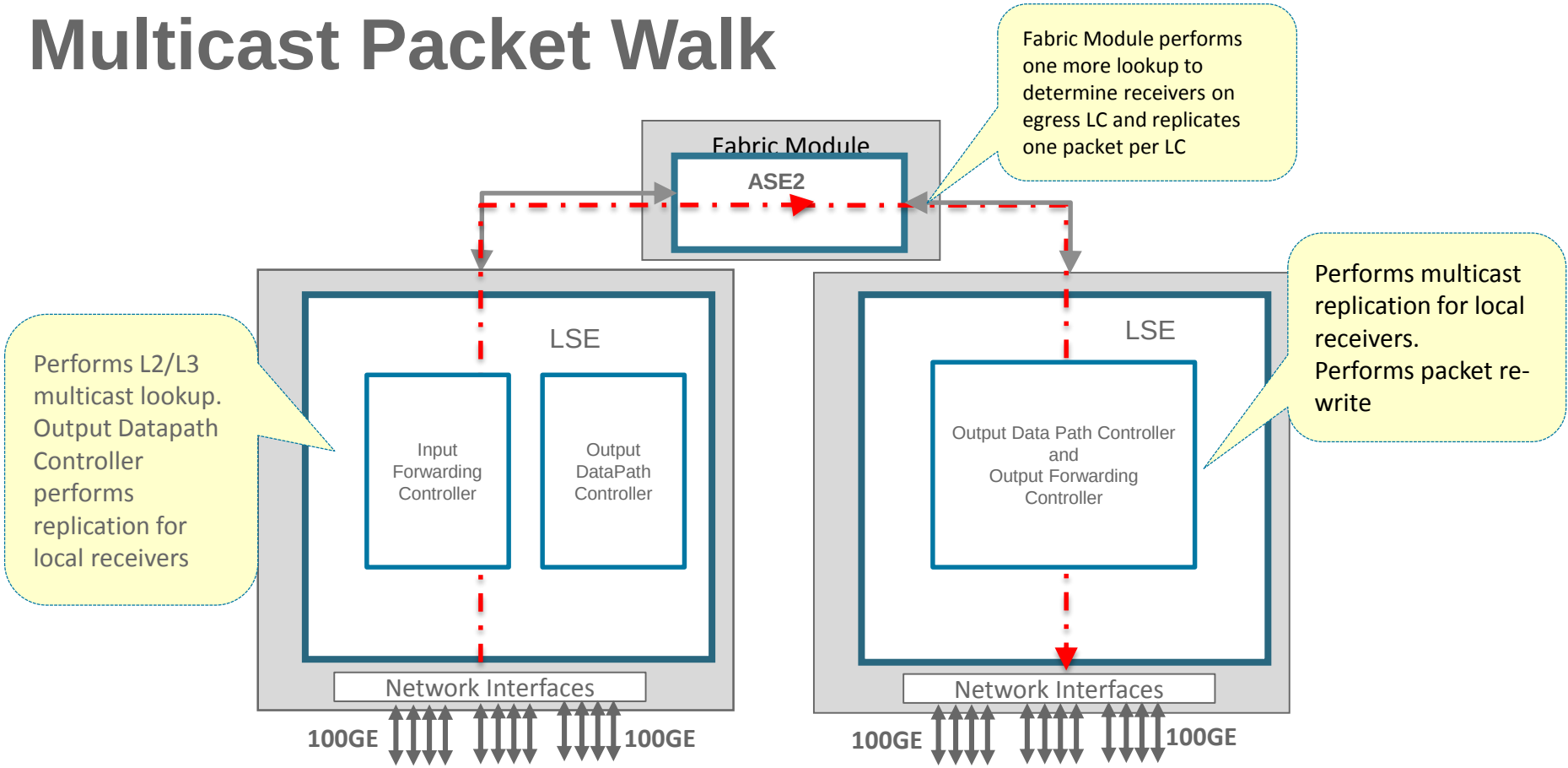
+ ACI-DCI, Buffer & Analytics

97xx-EX Series Line Cards

L2/L3 Unicast Packet Walk

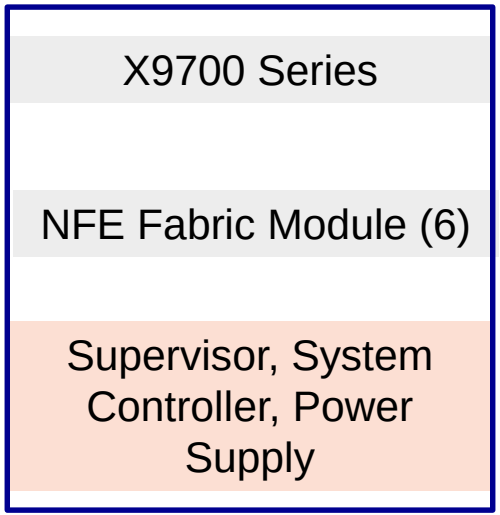


Multicast Packet Walk



Nexus 9500 – LC and Fabric Compatibility
Cisco ASIC

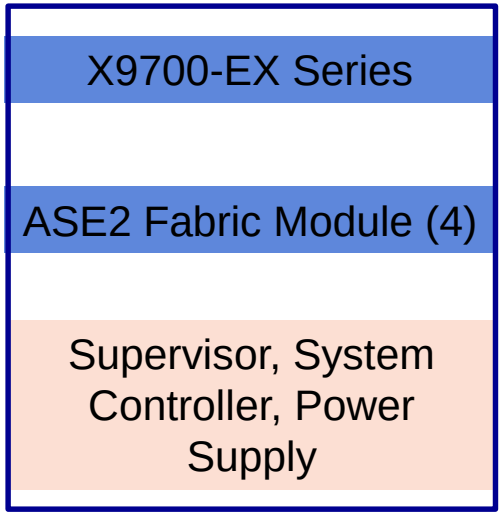
ACI Only



4, 8 and 16 Slot

Shipping

ACI, NX-OS



4, 8 and 16* Slot

Shipping

* Future item

Nexus 9500 – Modular

9500 Series

9504 9508 9516

Existing 4-, 8-, 16- slot chassis

No mid-plane to update

Power and cooling within existing shipping system profile

Existing shipping Power Supply, Supervisor and System Controllers

X9700-EX (NX-OS and ACI)

Cisco ASIC

16nm Technology

32p 100G QSFP28 Line card

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

+

Fabric Module

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

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

X9400-S (NX-OS)

Merchant ASIC

28 and 40nm Technology

32p 100G QSFP28 Line card

- 10/25/40/50/100G

+

Fabric Module

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

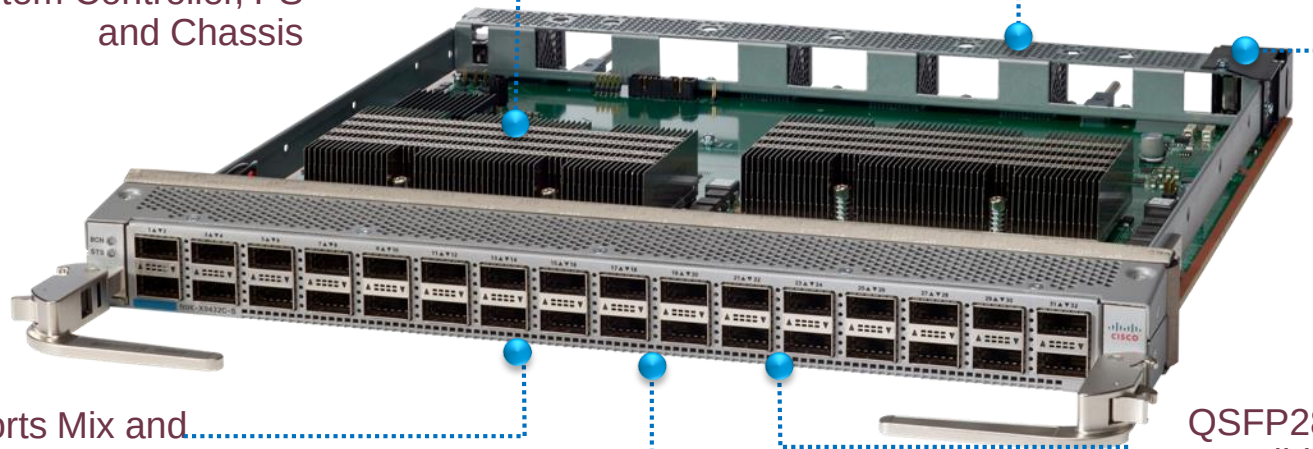
1/10/25/40/50/100G Capable

40/100G - Merchant
N9K-X9432C-S

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

Investment Protection with Supervisors, System Controller, PS and Chassis

Supported in NX-OS mode



Supports Mix and Match Current Linecards*

QSFP28 Connector, Pin compatible with 40G QSFP+

4, 8 and 16* Chassis

* future

Agenda

- What's New
  - 2nd Generation Nexus 9000
  - Moore's Law
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9500 (Modular)
  - Nexus 9200/9300 (Fixed)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- 40G/100G Transceiver
  - 25G technology

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

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

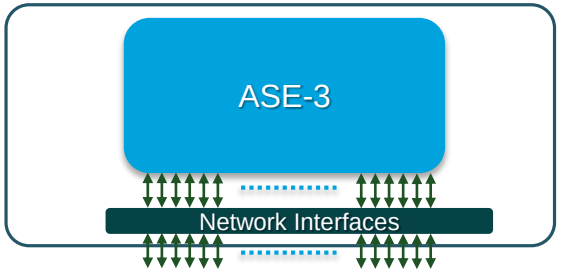
48p 10/25G SFP + 4p 100G/ 6p 40G QSFP
Nexus 92160YC-X

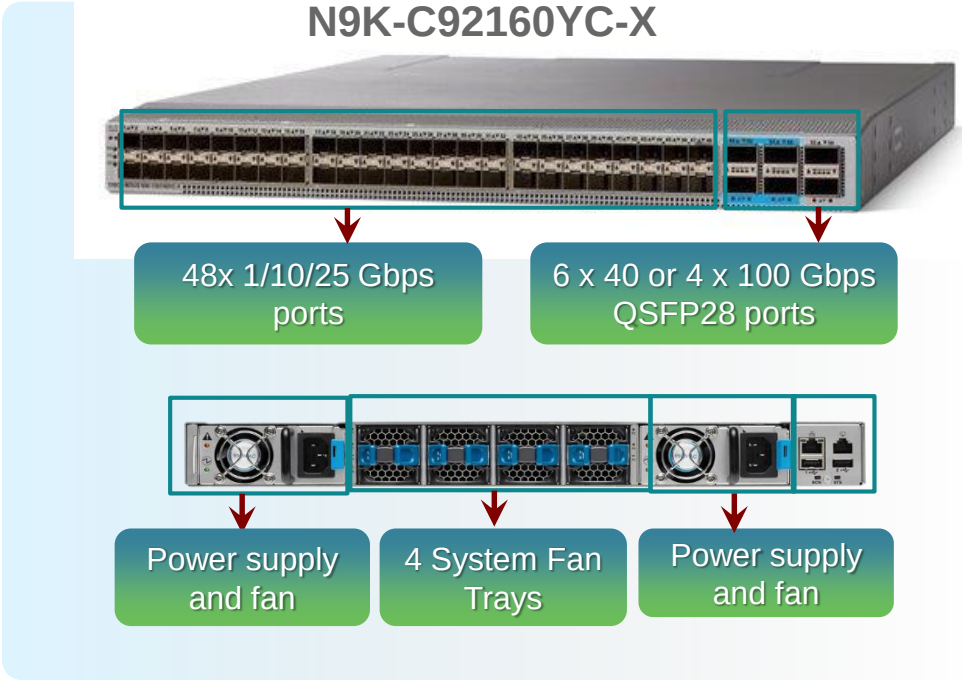
NX-OS switches
Industry's first 36p 100G 1RU switch
Industry's first native 25G VXLAN capable switch
Up to 30 MB shared buffer
High density compact 40/100G aggregation

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 92160 Port Configuration

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

CLI to find the operation mode:

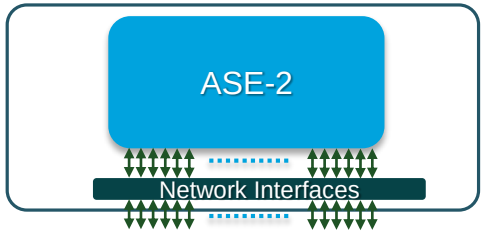
```
drvly15(config-if-range)# sh running-config | grep portmode
hardware profile portmode 48x25G+2x100G+4x40G
```

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

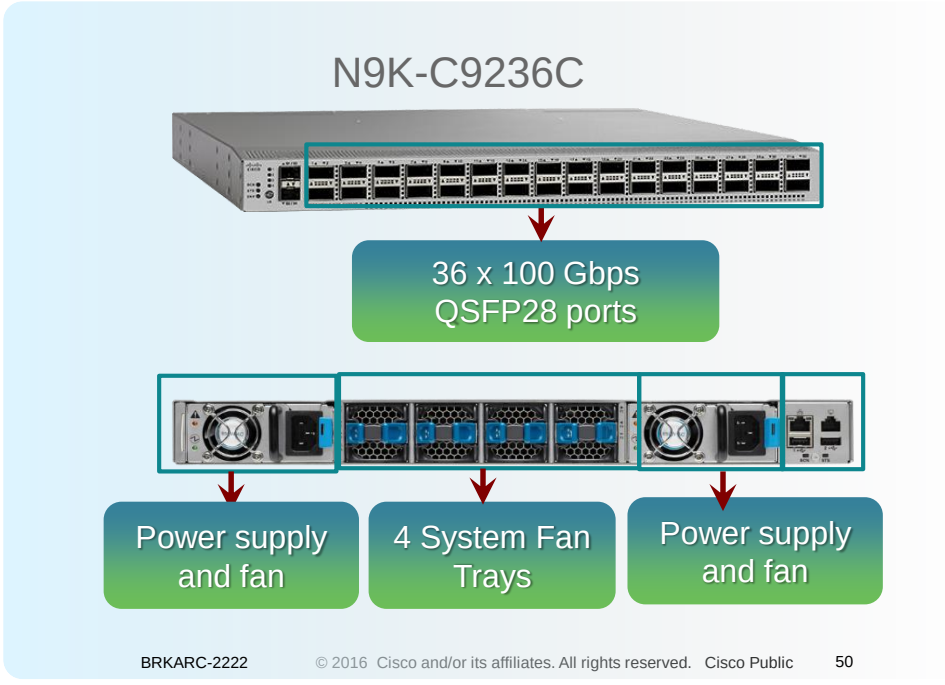
- **Breakout modes**
- **There are two breakout modes**
  - **40G to 4x10G breakout.**
    - **This breaks out 40G ports into 4 X 10G ports**
    - **Cli command**
- **interface breakout module 1 port <x> map 10g-4x**
  - **100G to 4x25G breakout.**
    - **This breaks out 100G ports into 4 X 25G ports**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

Nexus 9236C
ASE2 Based

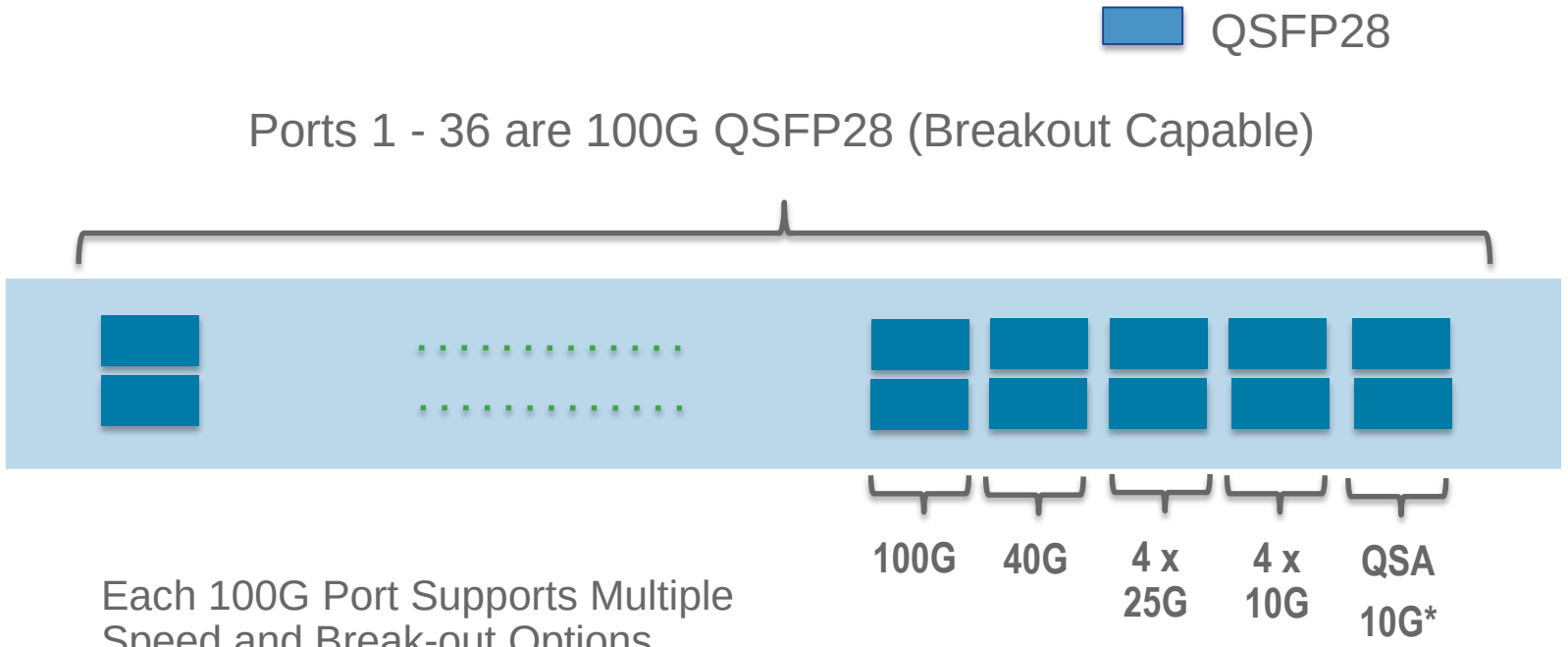
- ASIC: ASE2
- 4-core CPU (Intel Ivy Bridge Gladden 4 core at 1.8 GHz)
- 16GB DRAM + 64GB SSD
- 2MB NVRAM
- Two Power supply (1200W) 1 + 1 redundant
  - Typical Power Usage 375 W
  - Maximum Power Usage 640 W
- Four Fans 3 + 1 redundant
- 36 x 40/100G ports
- 144 10/25G ports (when all ports in breakout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



36 x 100G
QSFP28



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

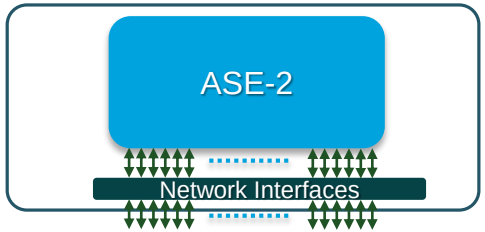


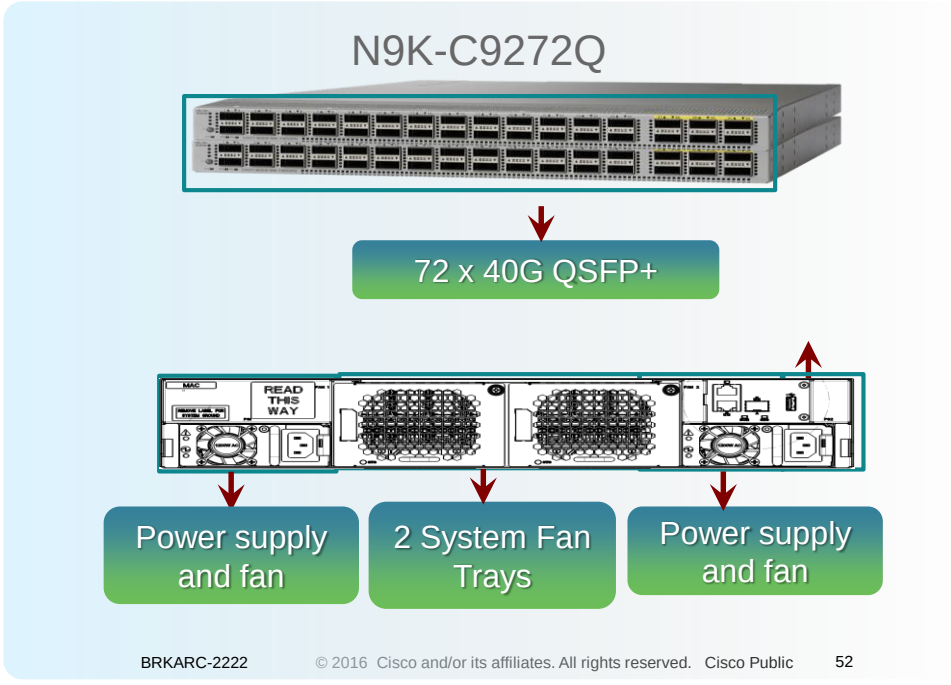
Each 100G Port Supports Multiple Speed and Break-out Options

* (QSA in a future SW release)

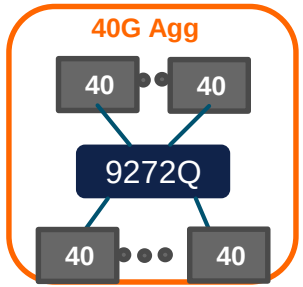
Nexus 9272Q
ASE2 Based

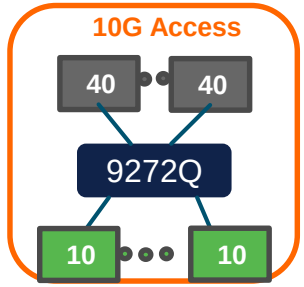
- ASIC: ASE2
- 4-core CPU (Intel Ivy Bridge Gladden 4 core at 1.8 GHz)
- 16GB DRAM + 64GB SSD
- 2MB NVRAM
- Two Power supply (1200W) 1 + 1 redundant
  - Typical Power Usage 310 W
  - Maximum Power Usage 1050 W
- Two Fans 1 + 1 redundant
- 36 x 40/100G ports
- 144 10/25G ports (when all ports in breakout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





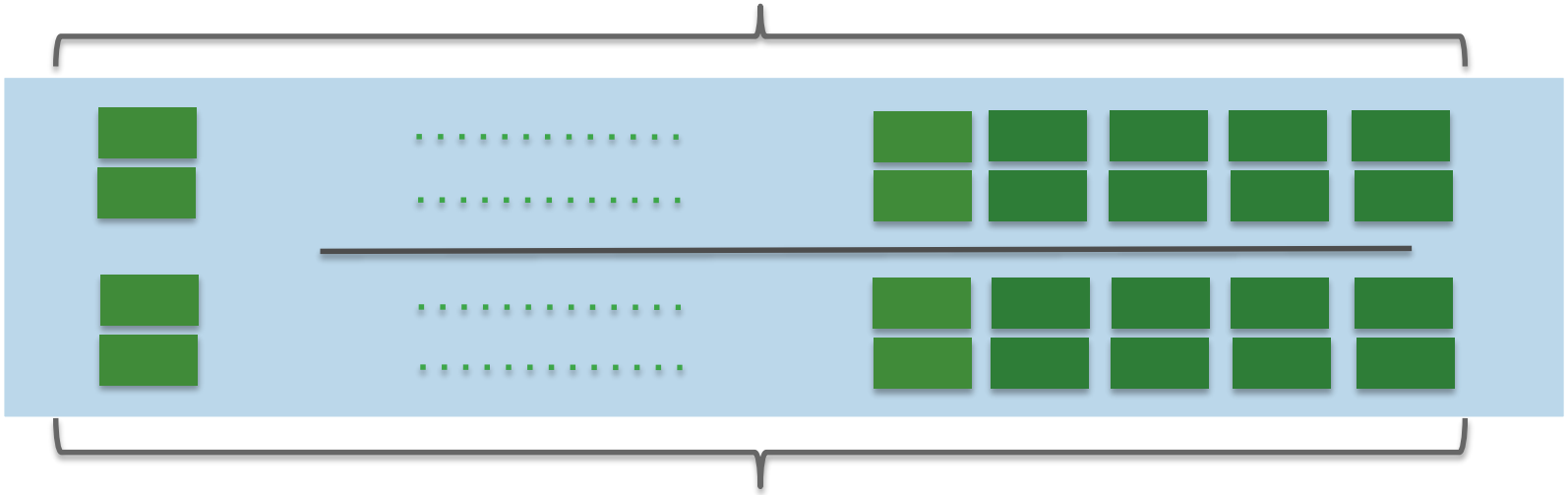
Nexus 9272Q Port Configuration
2RU 72 Port 40G Fiber





QSFP+

Ports 1 - 36 are 40G QSFP+

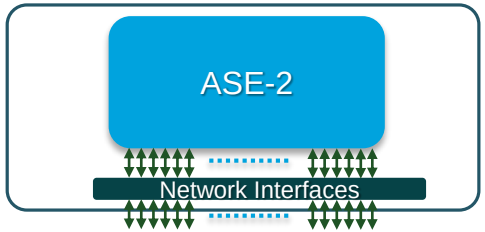


Ports 37 - 72 are 40G QSFP+ (Breakout Capable 144 x 10G)

Nexus 92304QC
ASE2 Based

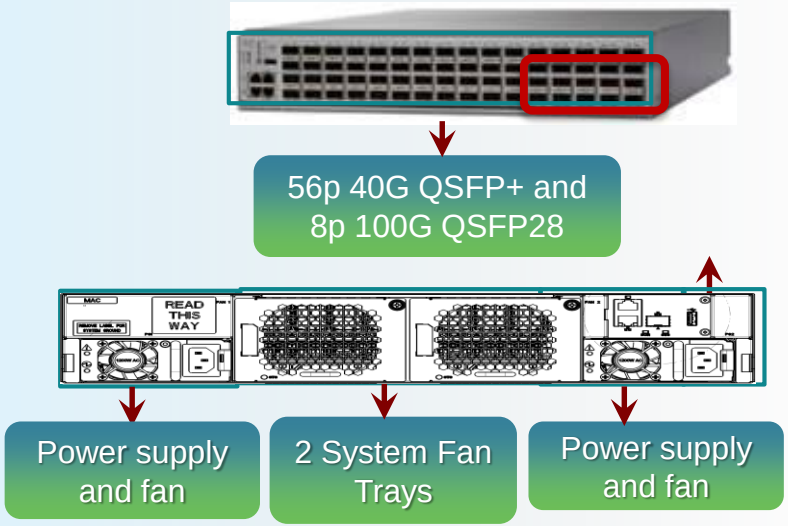
- ASIC: ASE2
- 4-core CPU (Intel Ivy Bridge Gladden 4 core at 1.8 GHz)
- 16GB DRAM + 64GB SSD
- 2MB NVRAM
- Two Power supply (1200W) 1 + 1 redundant
  - Typical Power Usage 305 W
  - Maximum Power Usage 720 W
- Two Fans 1 + 1 redundant

- 56 x 40 Gbps + 8 x 100 Gbps

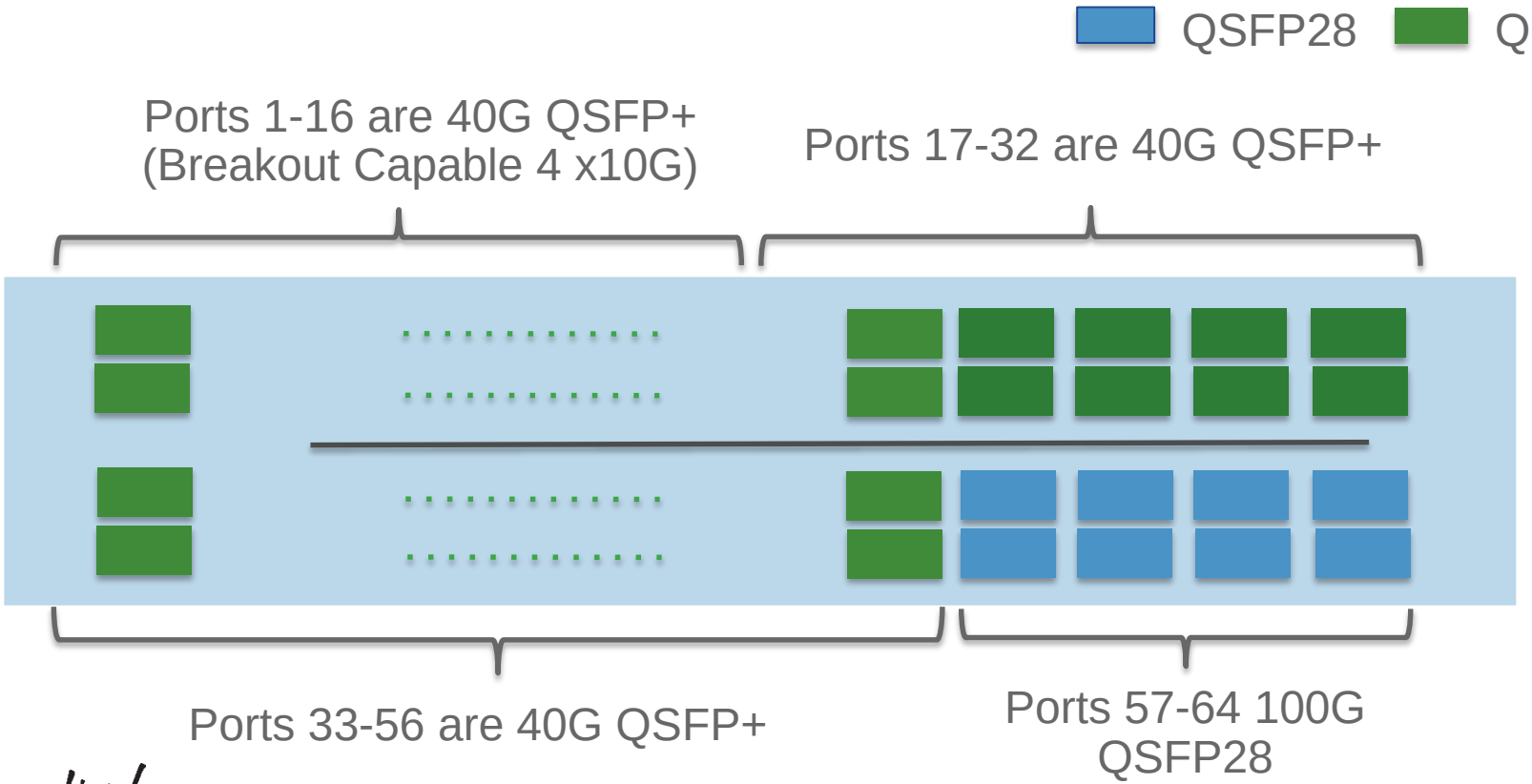


56 x 40G QSFP+ & 8 x 100G QSFP28

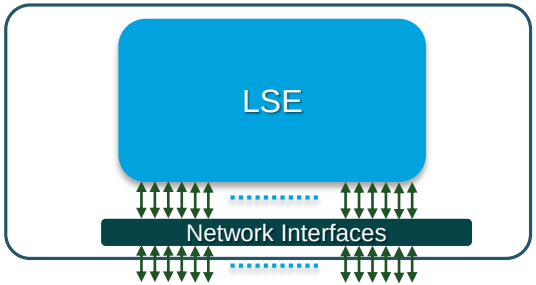
N9K-C92304QC



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

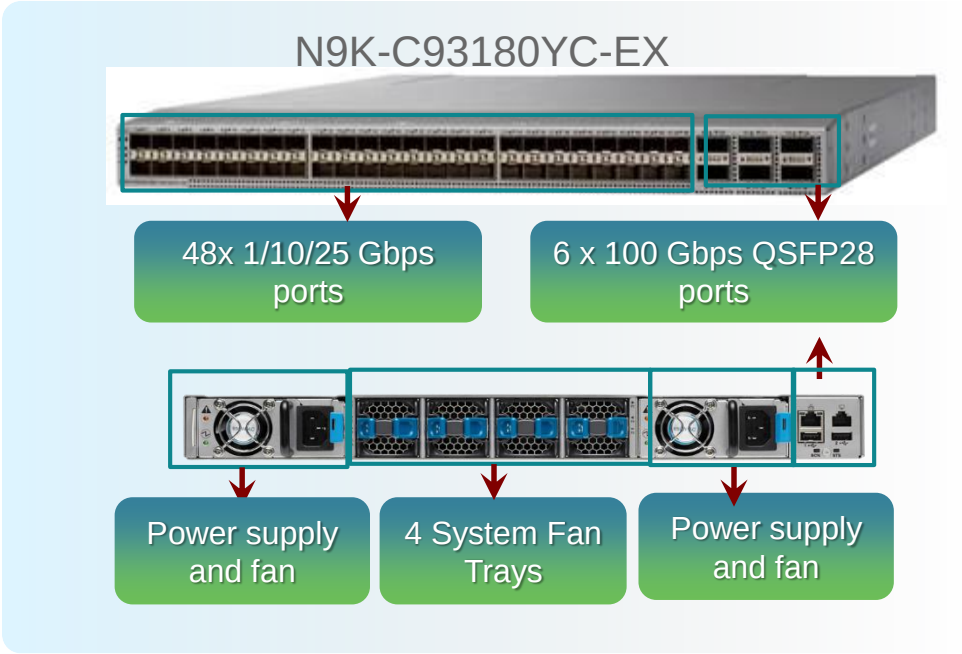


Nexus 93180YC-EX Series
LSE Based



- 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

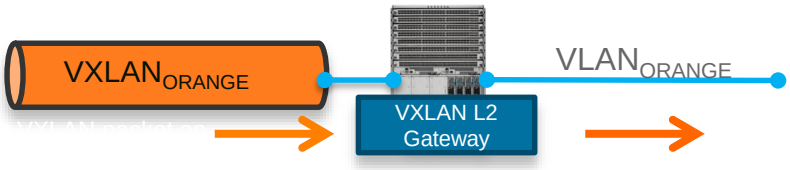


Agenda

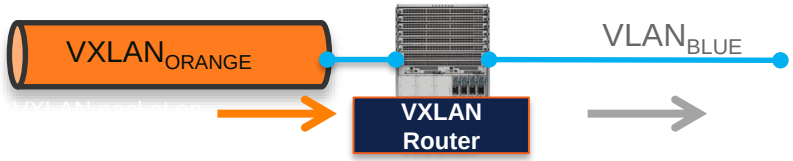
- What's New
  - 2nd Generation Nexus 9000
  - Moore's Law
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9500 (Modular)
  - Nexus 9200/9300 (Fixed)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- 40G/100G Transceiver
  - 25G technology

VXLAN Support
Gateway, Bridging, Routing*

VXLAN to VLAN Bridging
(L2 Gateway)



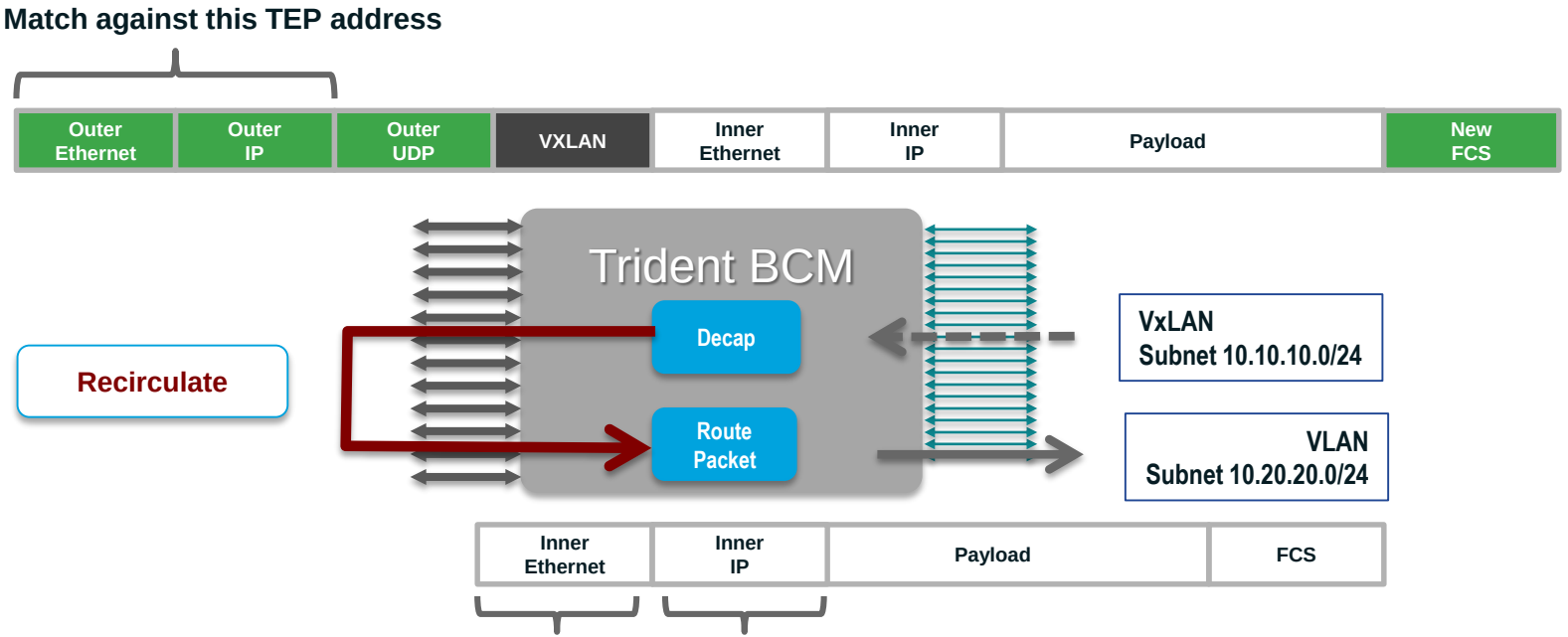
VXLAN to VLAN Routing
(L3 Gateway)



VXLAN to VXLAN Routing
(L3 Gateway)

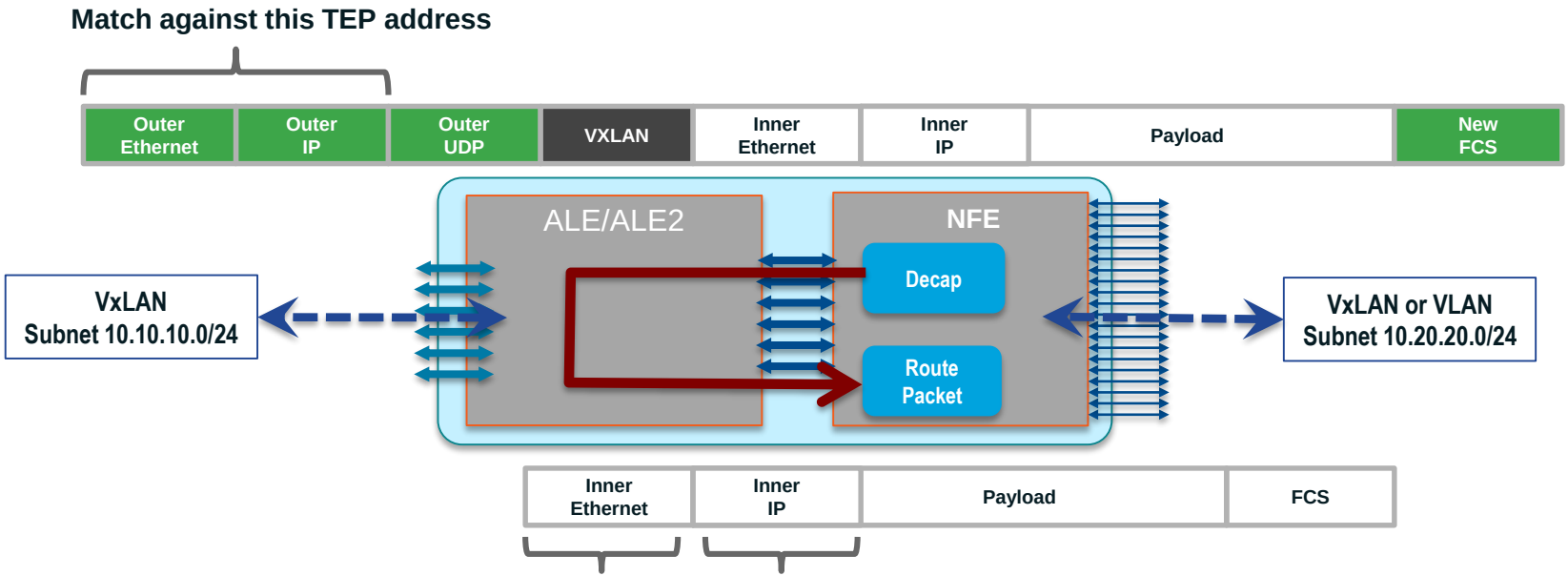
VxLAN to VLAN Routing – Trident 2

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



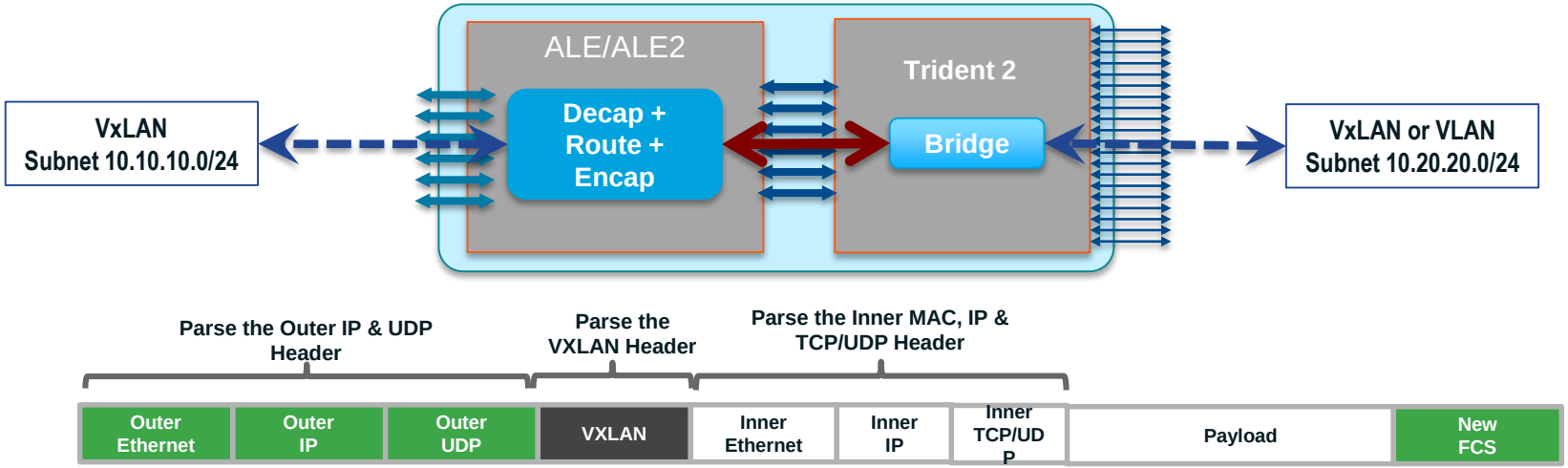
VLAN/VxLAN to VxLAN Routing
First Gen Nexus 9300 NX-OS Mode

- 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



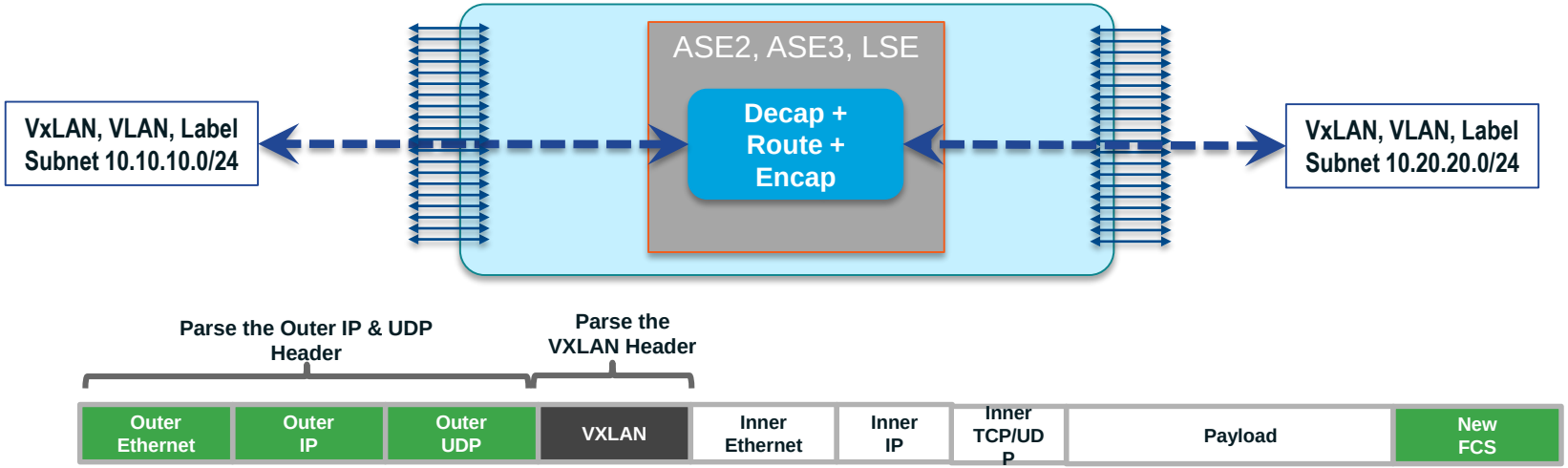
VLAN/VxLAN to VxLAN Routing
First Gen Nexus 9300 ACI Mode

- 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



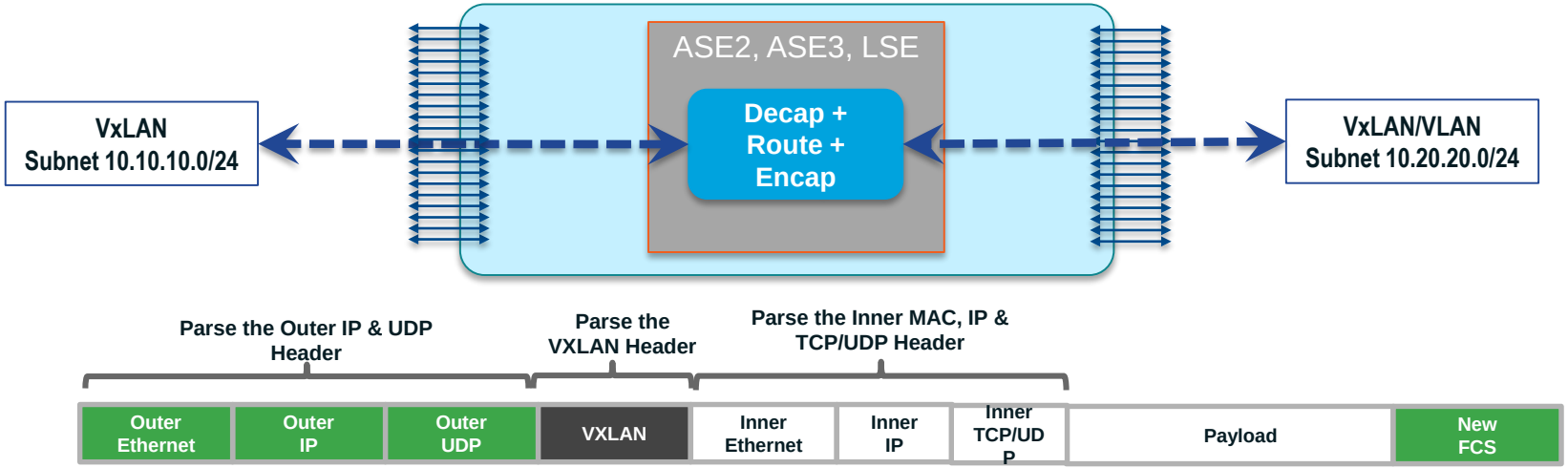
VLAN/VxLAN to VxLAN Routing
Nexus 9300EX, 9200 Standalone Mode

- 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



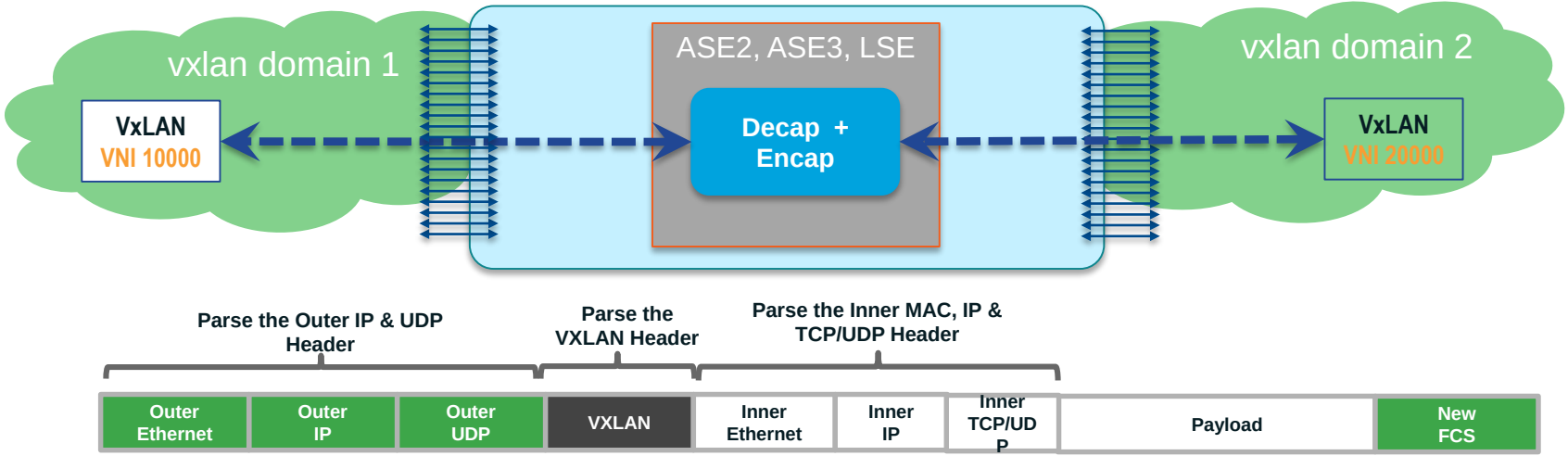
VLAN/VxLAN to VxLAN Routing
Nexus 9300EX ACI Mode

- 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



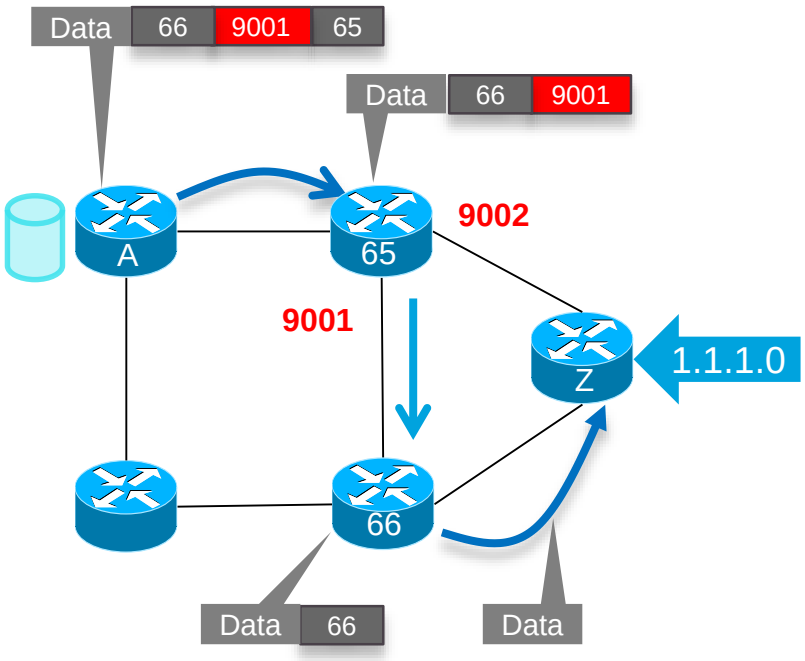
VXLAN to VXLAN Bridging*
Nexus 9200 and Nexus 9300-EX

- 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
- Decapsulate packets, terminate vxlan tunnel, encapsulate with new VNI and new outer header
- Solution to connect two vxlan domain2(multi-pod or DCI)



* Future software feature

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)

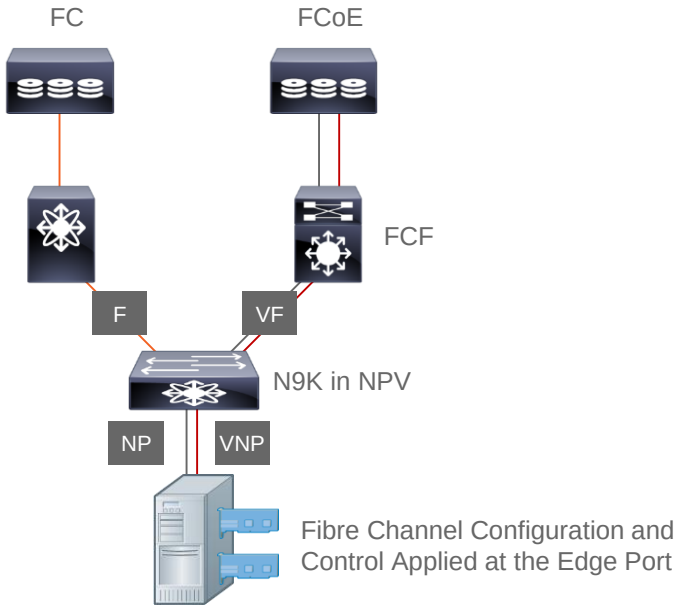
Prefix | Nodal SID | Adjacency SID

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

FCoE NPV – Unified Fabric Switching Nexus 9300 & 9300EX

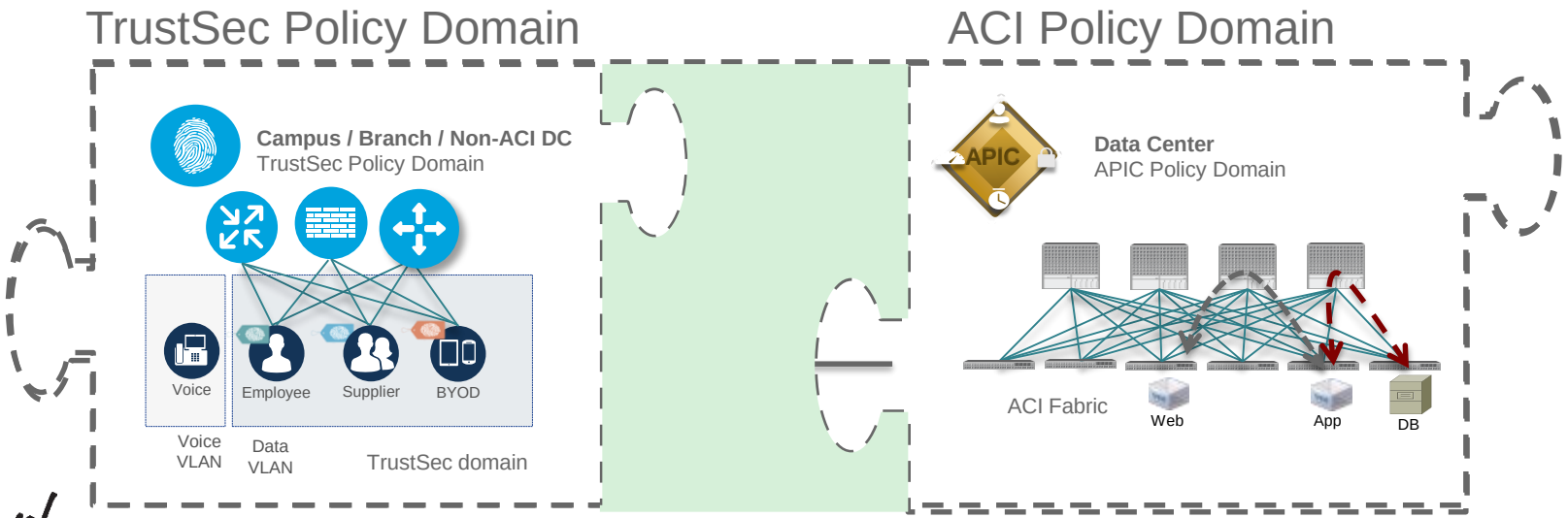
Connect FCoE-capable Hosts to a FCoE-Capable FCoE Forwarder (FCF) Device

- Standalone NX-OS support
  - 9300 with FCoE + vPC - 7.0(3)I4(1)
  - FCoE NPV on N92xx and N93xx (Q3CY16)
  - FCoE on FEX N2348UPQ
- ACI support
  - 9300-EX (Q3CY16)
  - FEX including B22 (Q4CY16)

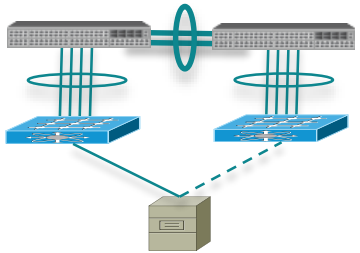
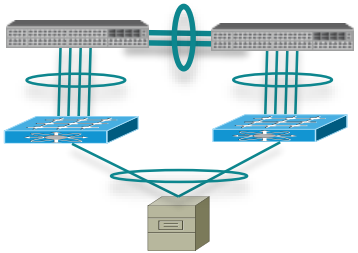
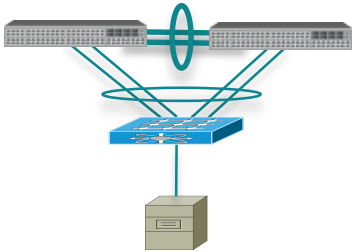
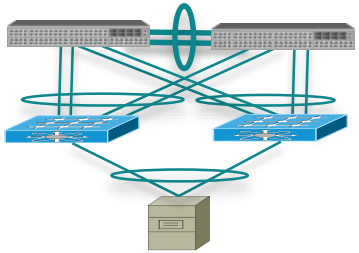
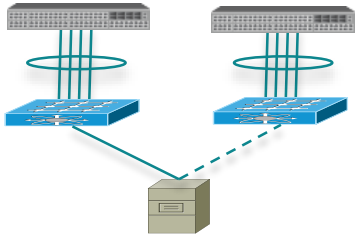
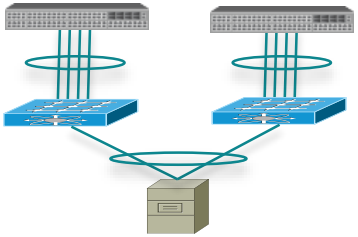
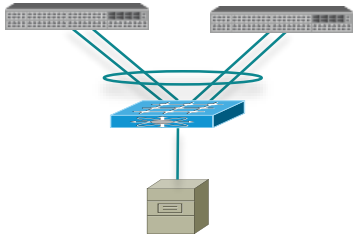
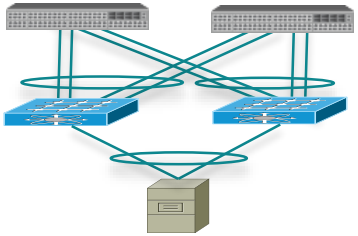


Enabling Group-Based Policies Across the Enterprise
VXLAN-GPE (ACI EPG) and TrustSec SGT

- 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



VNTAG - FEX Topology Support Roadmap

Active/Standby Teaming	Straight Through (Single Homed)	vPC (Dual Homed)	EvPC	
				**Nexus 9300 Standalone**
6.1(2)I2(3)	6.1(2)I2(3)	Target 2HCY16	Future	
				**Nexus 9300 ACI Leaf**
11.0(1d) - Shipping	11.1(x) - Shipping	Future	Future	

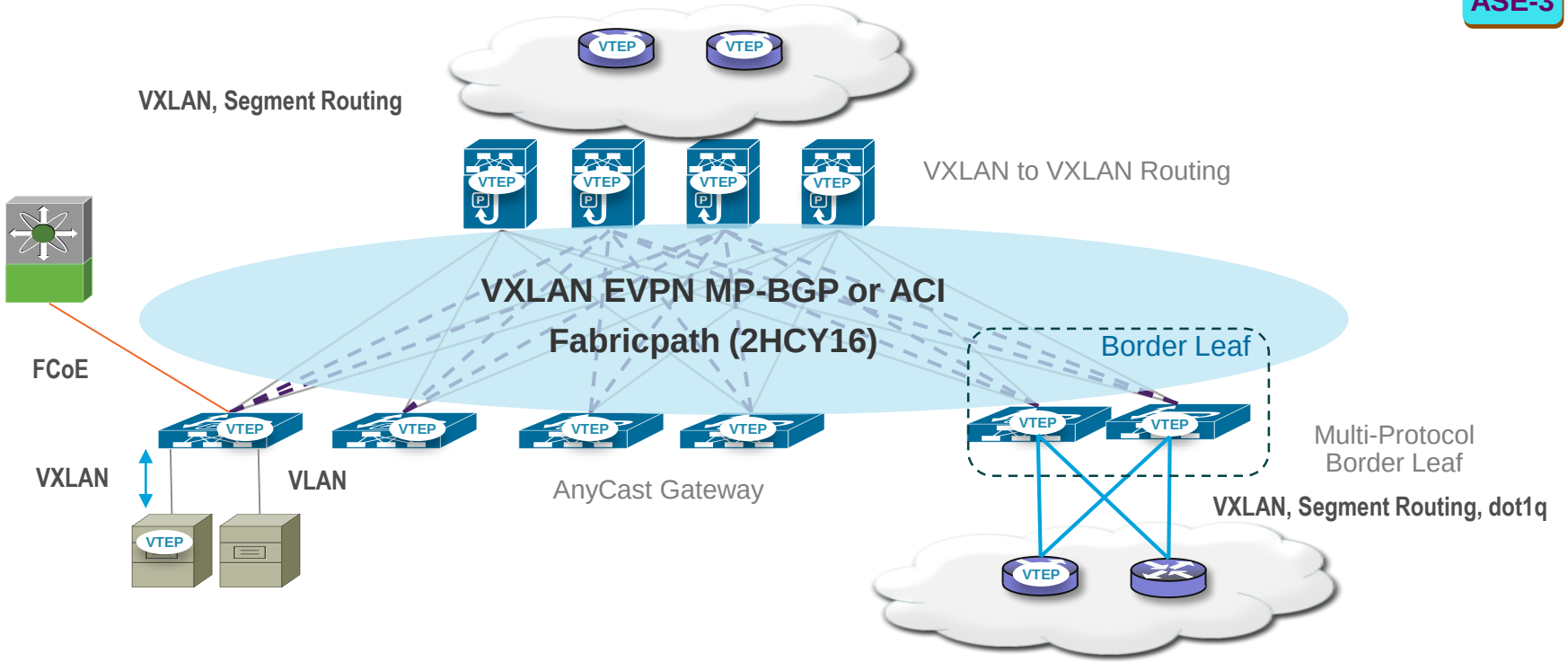
FEX Support Matrix (As of June 2016)

	N9K-C9332PQ	N9K-C9372PX	N9K-C9372TX	N9K-C9396PX	N9K-C93120TX	N9K-C93128TX	N9K-C93180YC	N9K-C93108TC-EX
N2K-C2224TP	√ (Supported with 4x10G breakout on switch)	√	x	√	x	x	2HCY16	x
N2K-C2232PP	√ (Supported with 4x10G breakout on switch)	√	x	√	x	x	2HCY16	x
N2K-C2232TM	√ (Supported with 4x10G breakout on switch)	√	x	√	x	x	2HCY16	x
N2K-C2232TM-E	√ (Supported with 4x10G breakout on switch)	√	x	√	x	x	2HCY16	x
N2K-C2248TP	√ (Supported with 4x10G breakout on switch)	√	x	√	x	x	2HCY16	x
N2K-C2248TP-E	√ (Supported with 4x10G breakout on switch)	√	x	√	x	x	2HCY16	x
N2K-C2248PQ	√ QSFP From FEX uplink to QSFP on Switch (w/ internal breakouts on FEX uplink)	√(Supported with 4x10G breakout on FEX uplink)	x	√(Supported with 4x10G breakout on FEX uplink)	x	x	2HCY16	x
N2K-C2332TQ	Future	Roadmap	x	Future	Future	x	Future	x
N2K-C2348UPQ	√ Operates in native 40G on uplink (QSFP to QSFP)	√ Supported with 4x10G breakout on FEX uplink	x	√ Supported with 4x10G breakout on FEX uplink	x	x	2HCY16	x
N2K-C2348TQ	√ Operates in native 40G on uplink (QSFP to QSFP)	√ Supported with 4x10G breakout on FEX uplink	x	√ Supported with 4x10G breakout on FEX uplink	x	x	2HCY16	x
N2K-C2348TQ-E	Future	Future	x	Future	x	x	Future	x
B22-HP	x	√	x	√	x	x	2HCY16	x
B22-DELL	x	√	x	√	x	x	2HCY16	x
B22-IBM	x	√	x	√	x	x	2HCY16	x
B22-Fujitsu	x	√	x	√	x	x	2HCY16	x

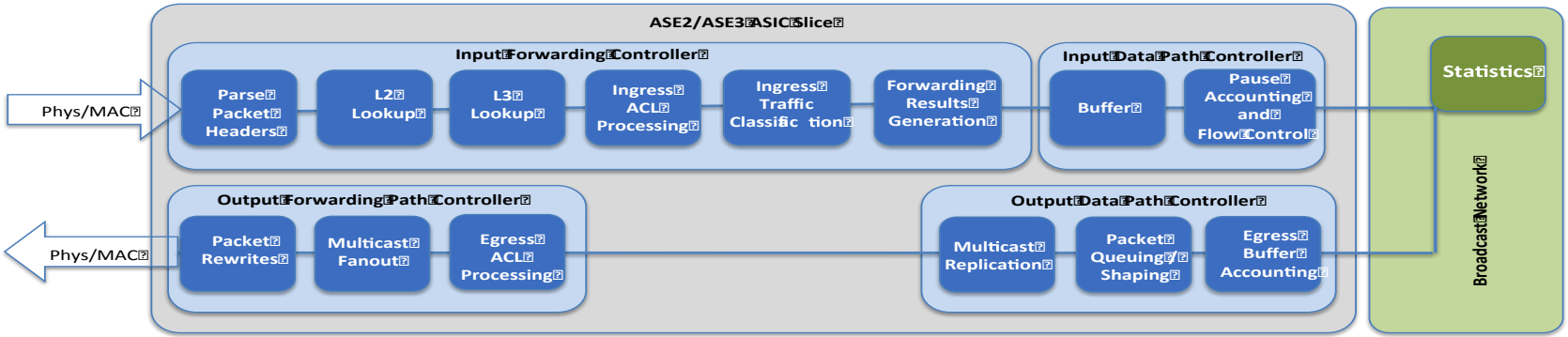
Agenda

- 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)

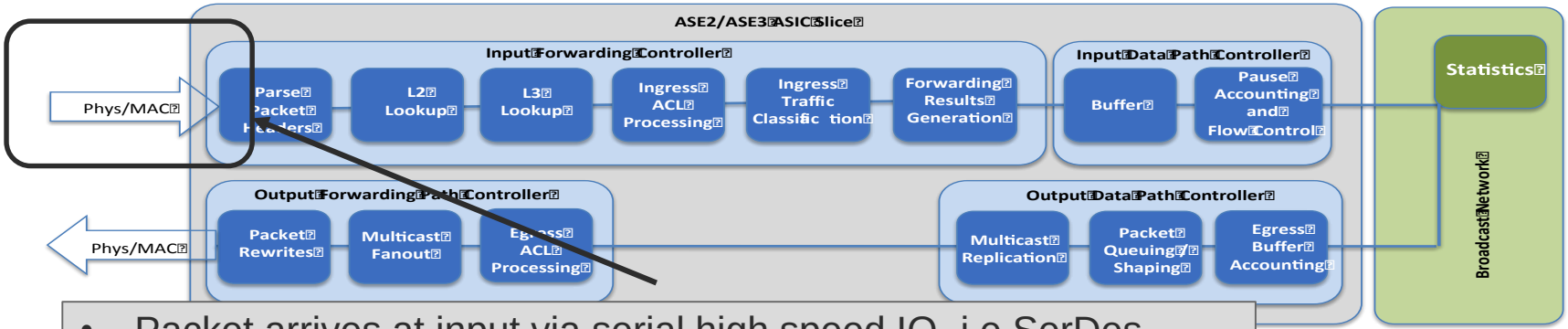
Nexus 9000 Forwarding



Nexus 9000 Life of a Packet
ASE2 / ASE3 / LSE

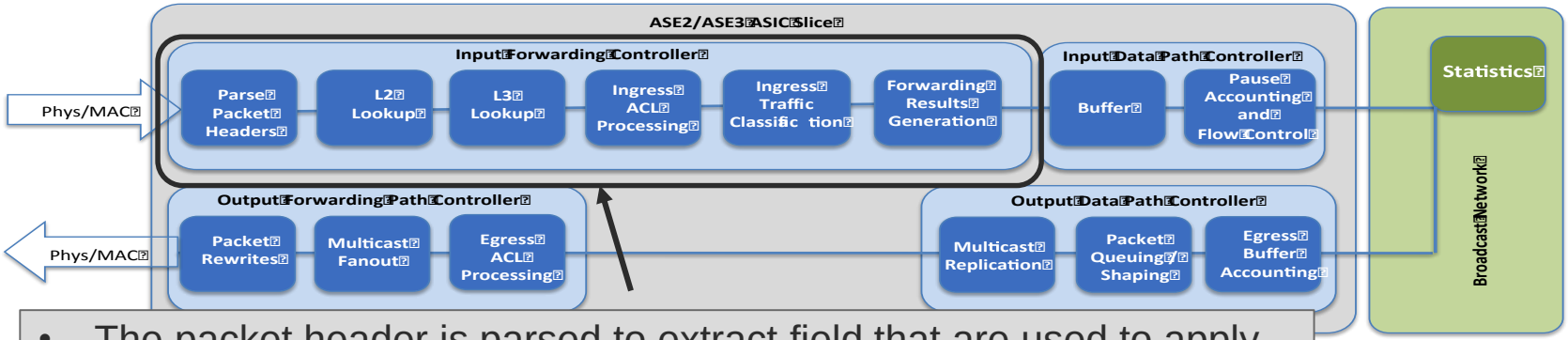


Life of a Packet in ASE2 / ASE3 / LSE ASIC



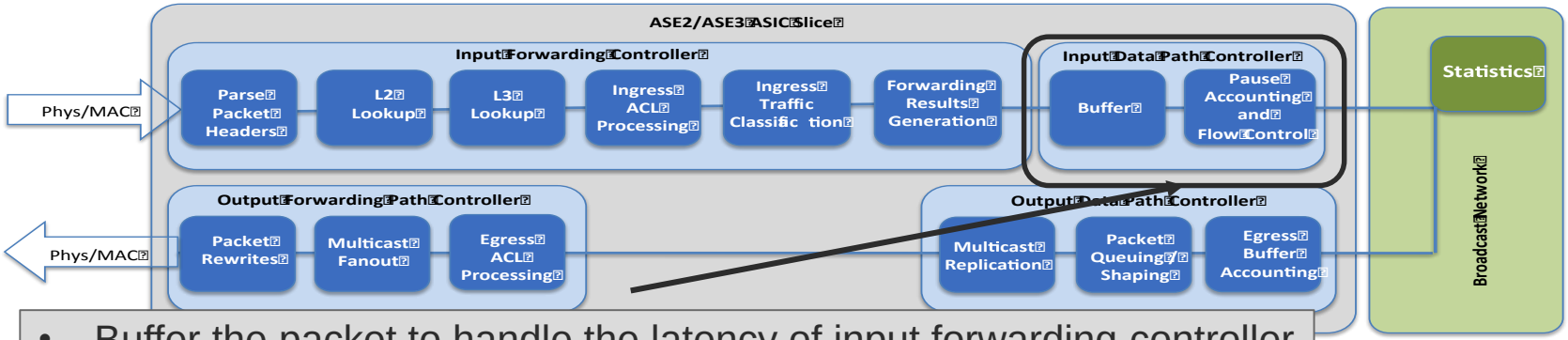
- Packet arrives at input via serial high speed IO, i.e SerDes
- The serial data is converted to parallel stream and MAC is responsible to validate framing protocol
- The MAC operates in cut through and pass the packet to client interface

Life of a Packet in ASE2 / ASE3 / LSE ASIC



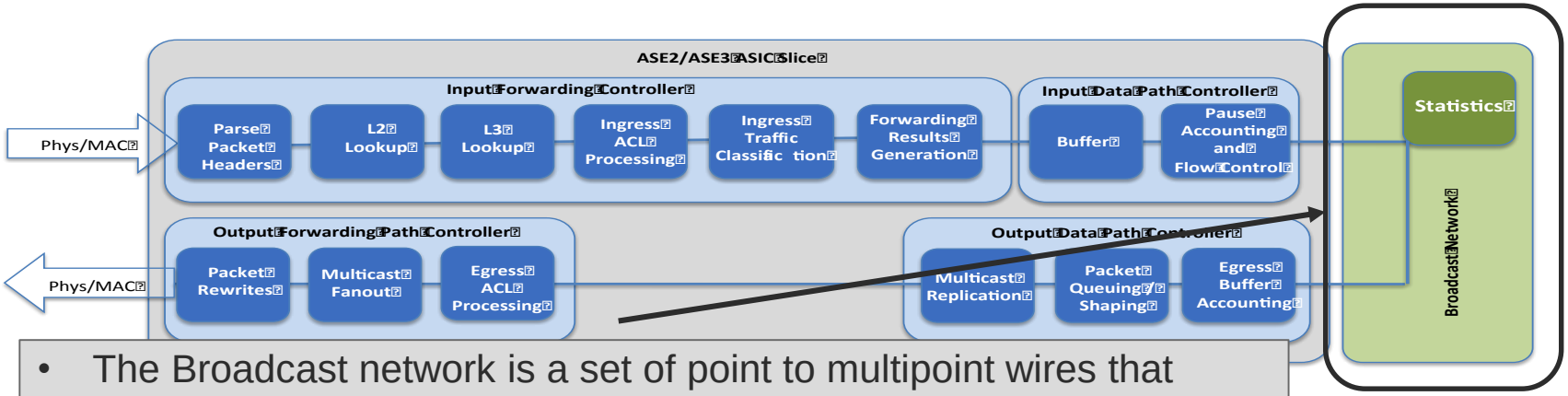
- The packet header is parsed to extract field that are used to apply policy and making forwarding decision and load-balancing
- The parsed field are used in a series of forwarding table and access control list lookup
- Flow Table Analytics

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

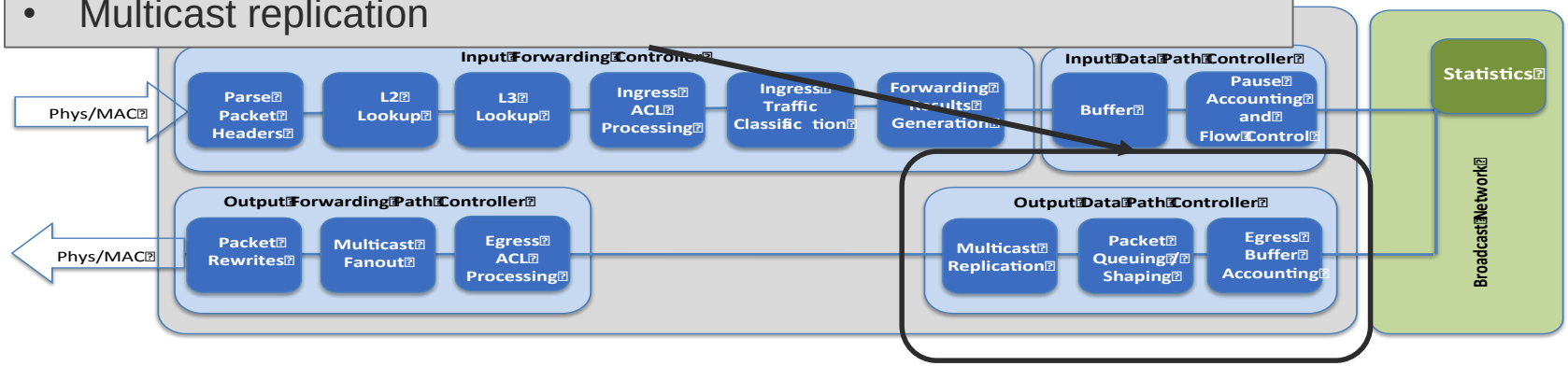
Life of a Packet in ASE2 / ASE3 / LSE ASIC



- 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*

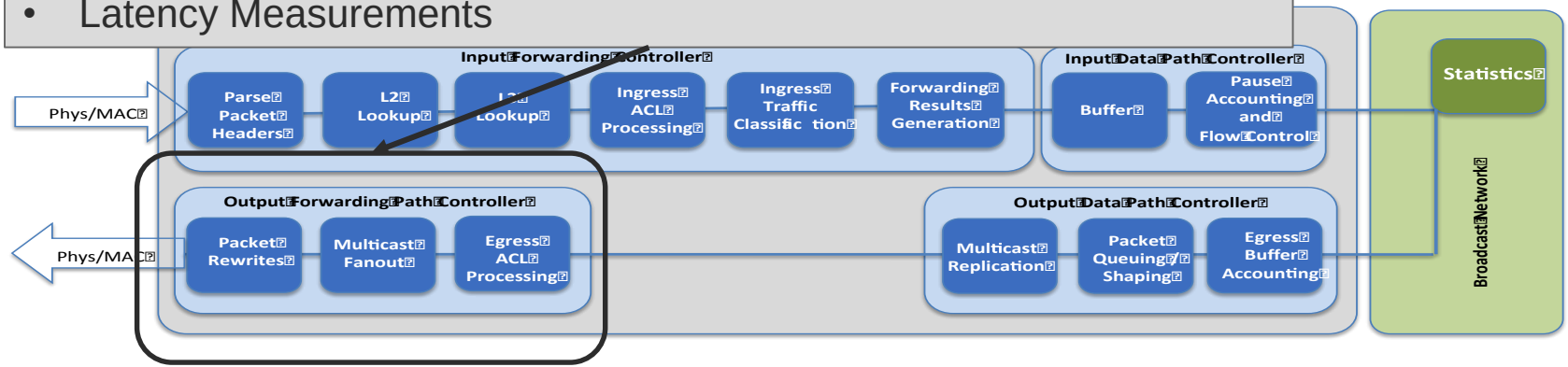
Life of a Packet in ASE2 / ASE3 / LSE ASIC

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



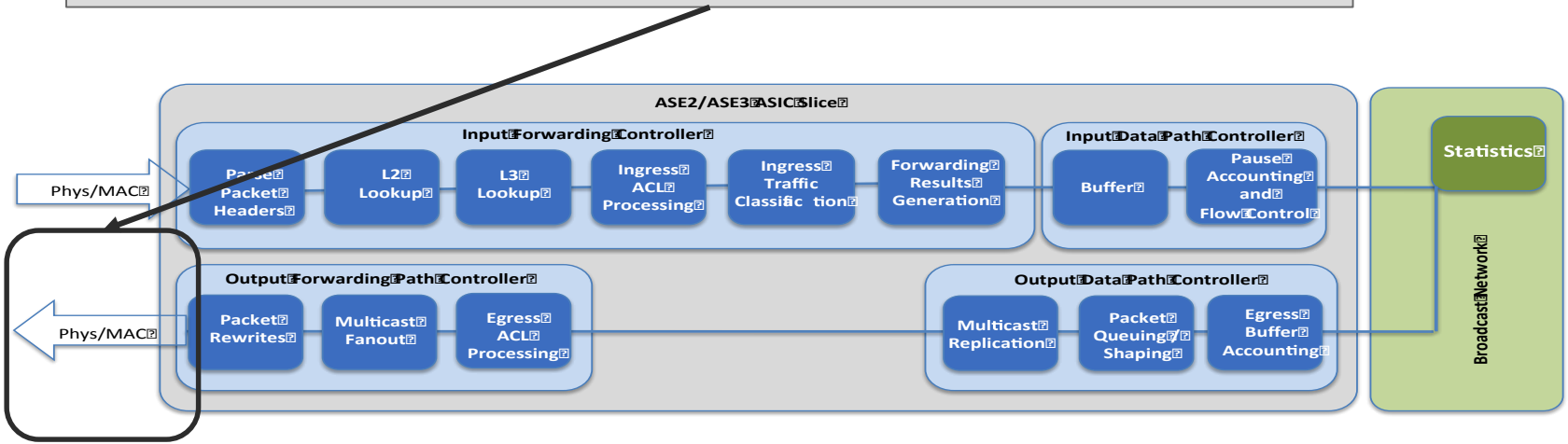
Life of a Packet in ASE2 / ASE3 / LSE ASIC

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

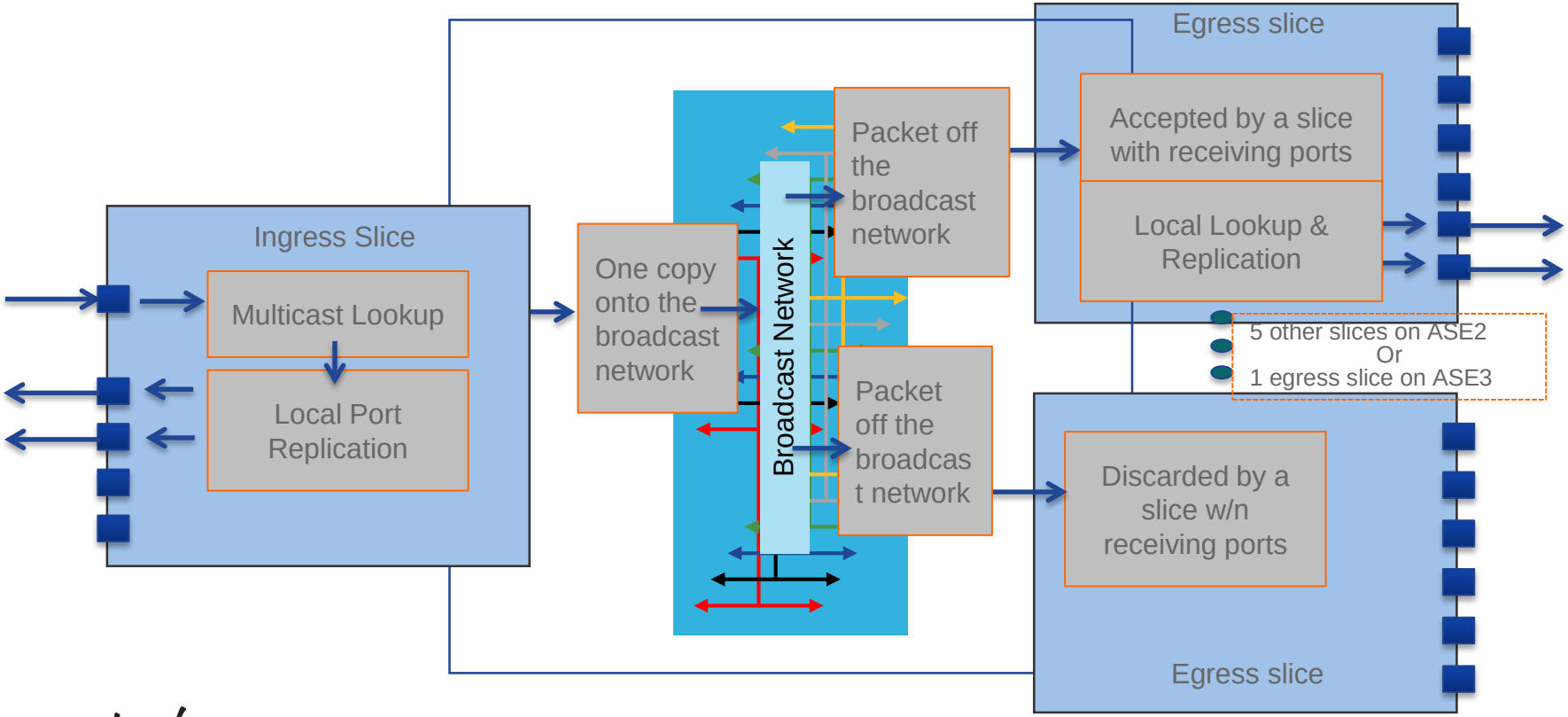


Life of a Packet in ASE2 / ASE3 / LSE ASIC

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



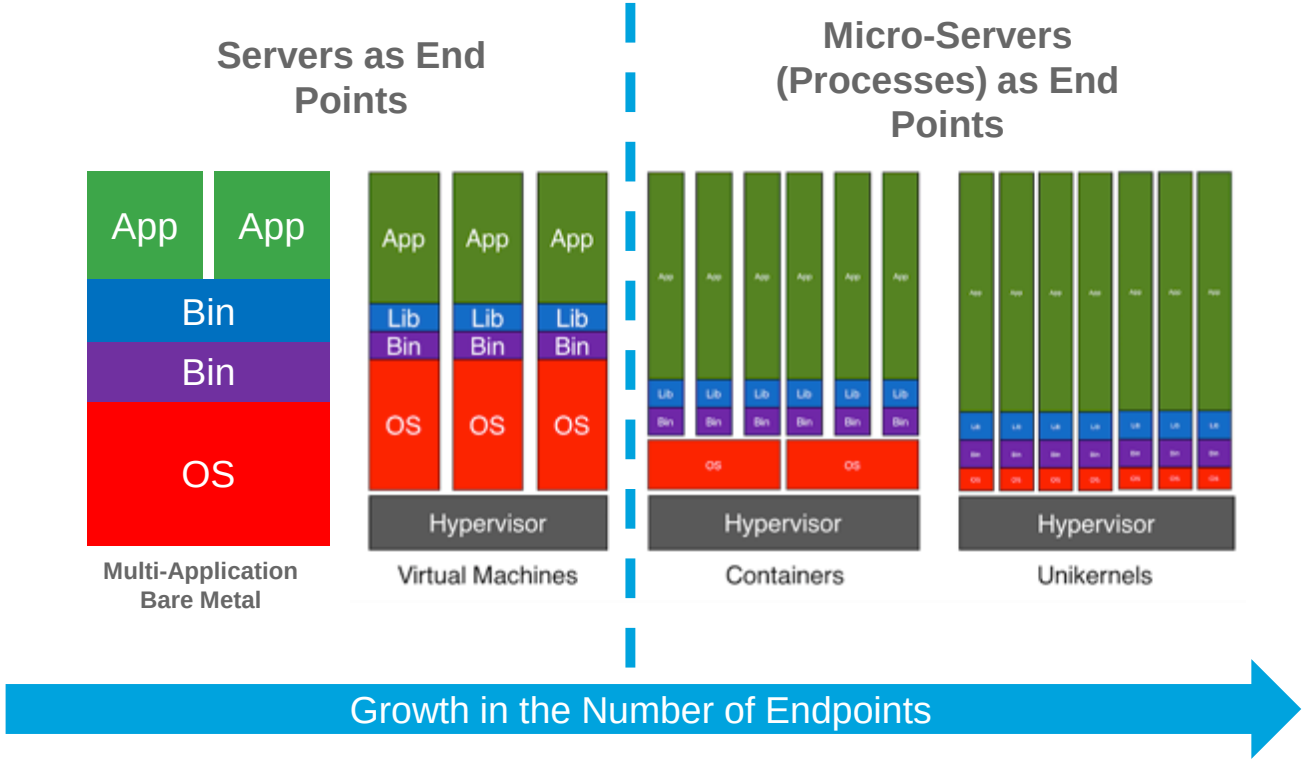
Multicast Packet Forwarding



Agenda

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

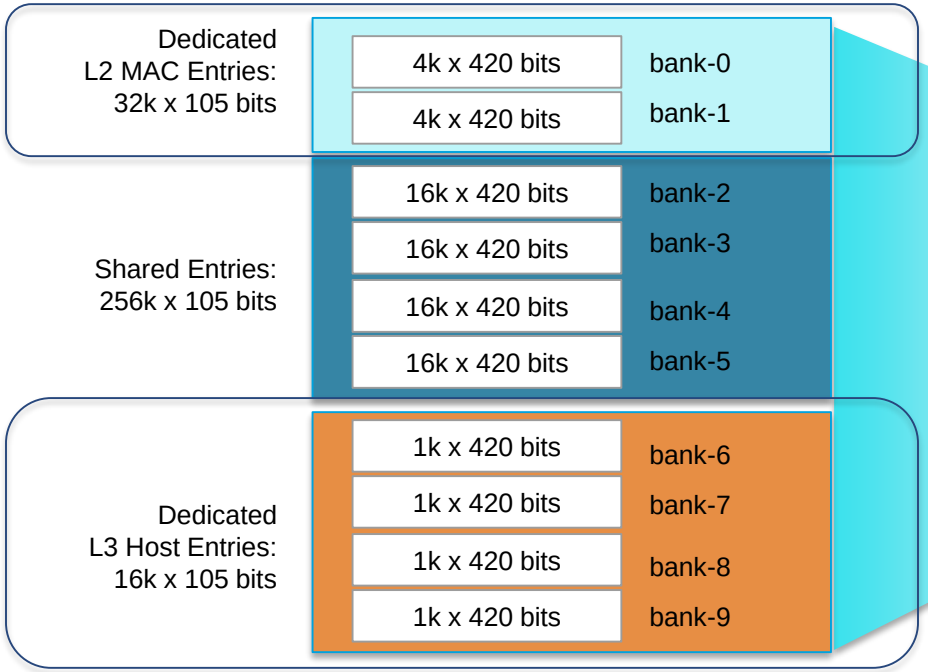
Nexus Forwarding Table Templates
Responding to changes in End Point Density



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)



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

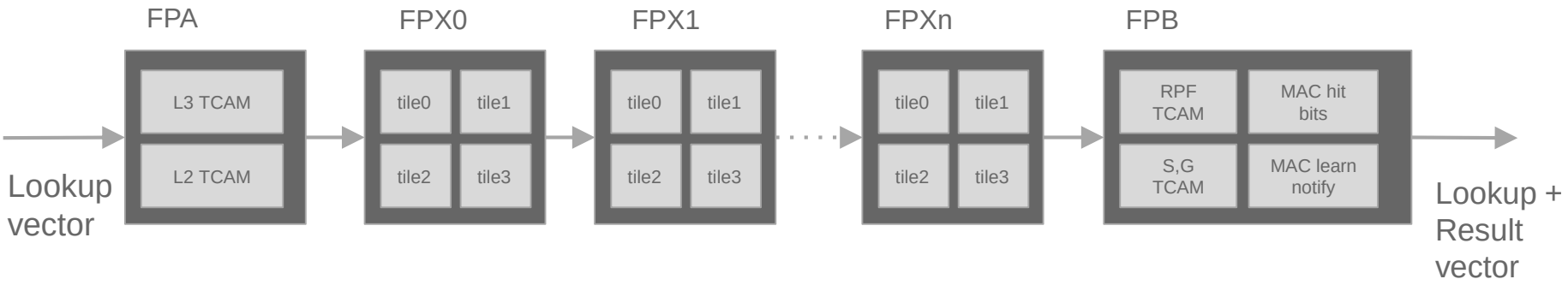
First Gen Nexus 9300 Forwarding Templates

	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



- 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 maximized utilization for each node in the network
  - Templates will be supported initially

Features Sharing Forwarding Tables

Sharing Memory Among Features with 2nd Gen. N9K

Shared Forwarding Table

Standalone(NX-OS)

- Prefix routes
- Host routes
- MAC
- Adjacency, ECMP
- Multicast
- MPLS

ACI

- LST/GST-Host routes
- Prefix routes
- MAC
- Adjacency, ECMP
- Policy
- External EPG subnets
- Multicast

Dedicated Table

- 16K VRF+BD(L2VNI +L3 VNI)
- 64K logical port to VNI mapping(port local VLAN to VNI Mapping)
- Hardware scale. Check verified scale document for software support

Forwarding Table Compression

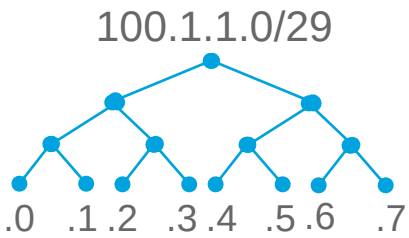
- Eliminating repetitive information from forwarding table. Increased table scale with same amount of SRA. Effectively compress forwarding table entries.
- Applicable for IPv4 host, IPv4 LPM routes and IPv6 /64 LPM routes

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 templates will be pre-defined.
- Customizable templates will be supported in a future SW release
- Raw table size. Please check software release for actual supported scale

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

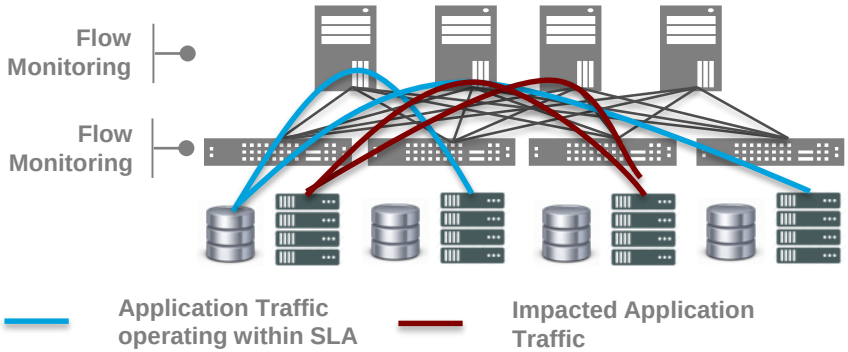
Agenda

- What's New
  - 2nd Generation Nexus 9000
  - Moore's Law
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9500 (Modular)
  - Nexus 9200/9300 (Fixed)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- 40G/100G Transceiver
  - 25G technology

Fabric Wide Troubleshooting
Real Time Monitoring, Debugging and Analysis

Granular Fabric Wide Flow Monitoring Delivering Diagnostic Correlation

“Tetration Analytics”



Debug

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

Monitor

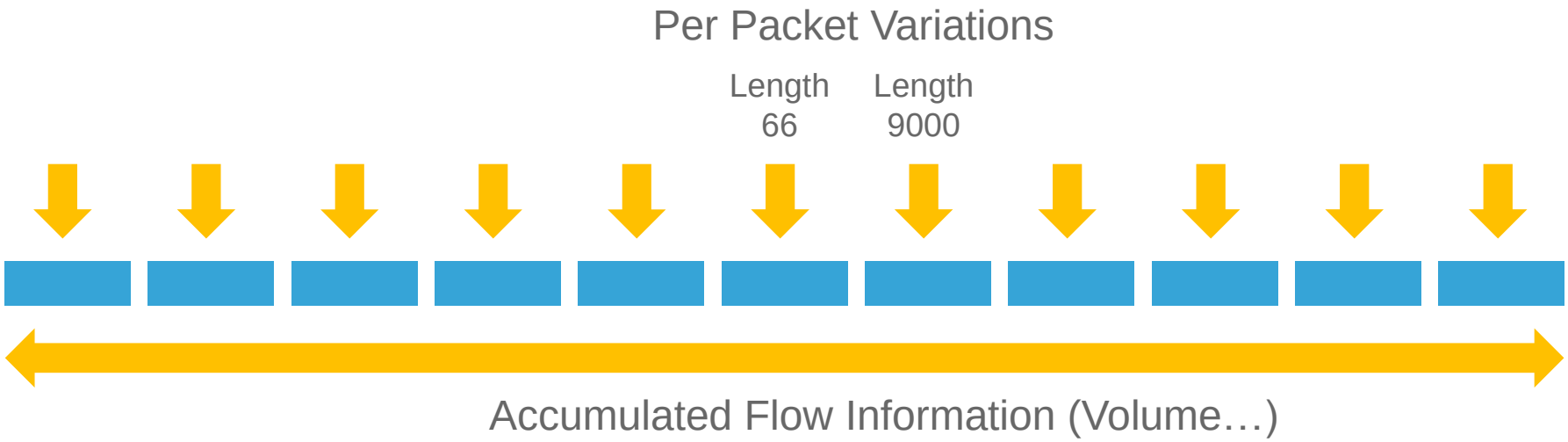
Track Latency (avg/min/max), buffer utilization, network events

Analyze

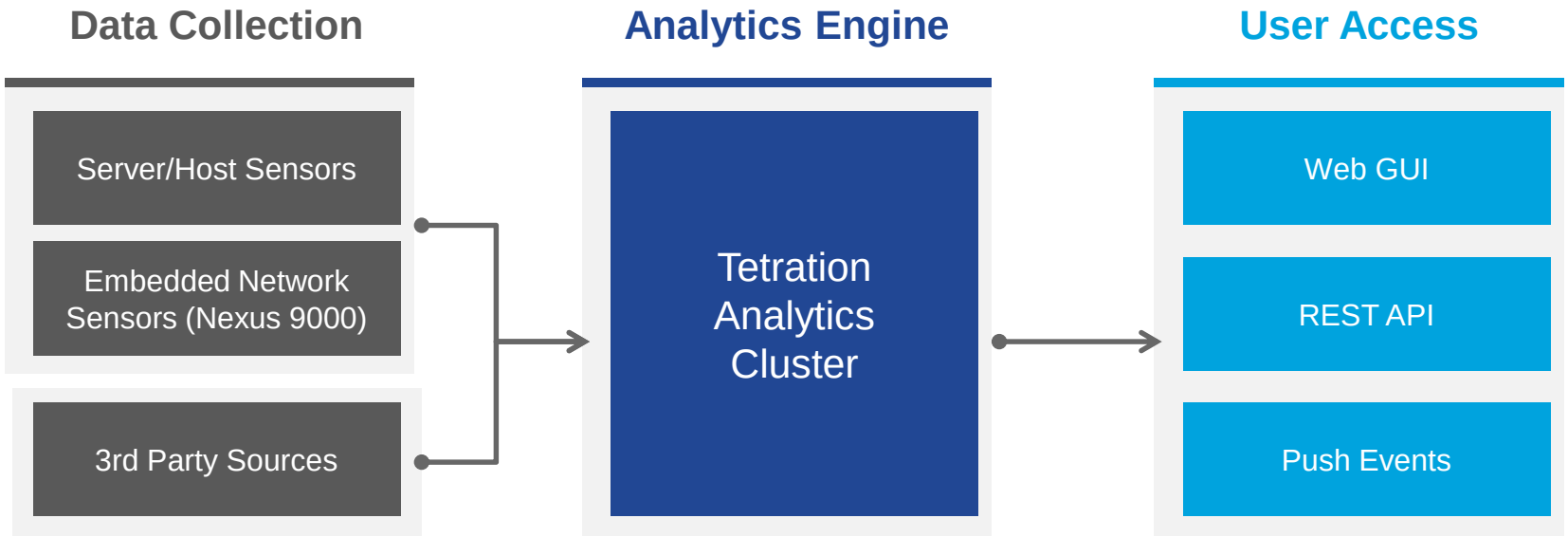
Specific events and suggest potential solution (e.g. trigger automatic rollback)

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



Cisco Tetration Analytics
Nexus 9000 Hardware Sensors



Tetration Analytics - Data Center Analytics Deployment and Use Cases

Session ID: BRKACI-2060

Tetration Analytics - Network Analytics & Machine Learning Enhancing Data Center Security and Operations

Session ID: BRKDCN-2040

Agenda

- What's New
  - 2nd Generation Nexus 9000
  - Moore's Law
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9500 (Modular)
  - Nexus 9200/9300 (Fixed)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- 40G/100G Transceiver
  - 25G technology

Nexus 9000 QoS and Buffering
Shared Memory & Egress Queuing

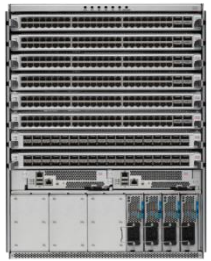
Cat4900 – Shared Memory Egress buffering

Nexus 5x00 – VoQ Ingress Buffering

Nexus 9200/9300 Shared Memory Egress buffering

Cat6500 – Egress Buffering

Nexus 7x00 – VoQ Ingress Buffering

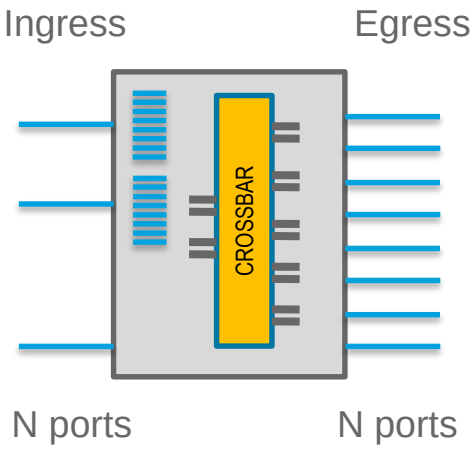


Nexus 9000 Egress Buffering

Nexus 9000 QoS and Buffering VoQ vs. Output Queue Design

VOQ

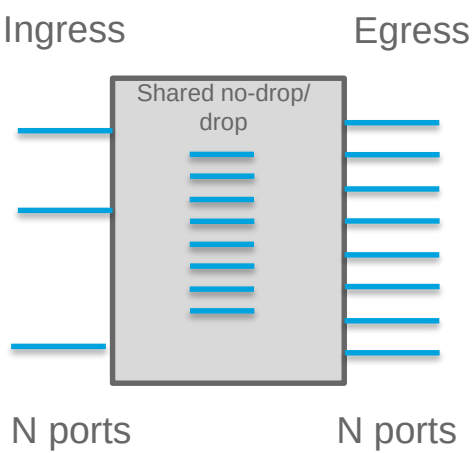
Virtual Output Queue



Input buffer for every egress port

NxN buffer size

Output Queue Buffer



Shared buffer for N egress ports

N buffer size

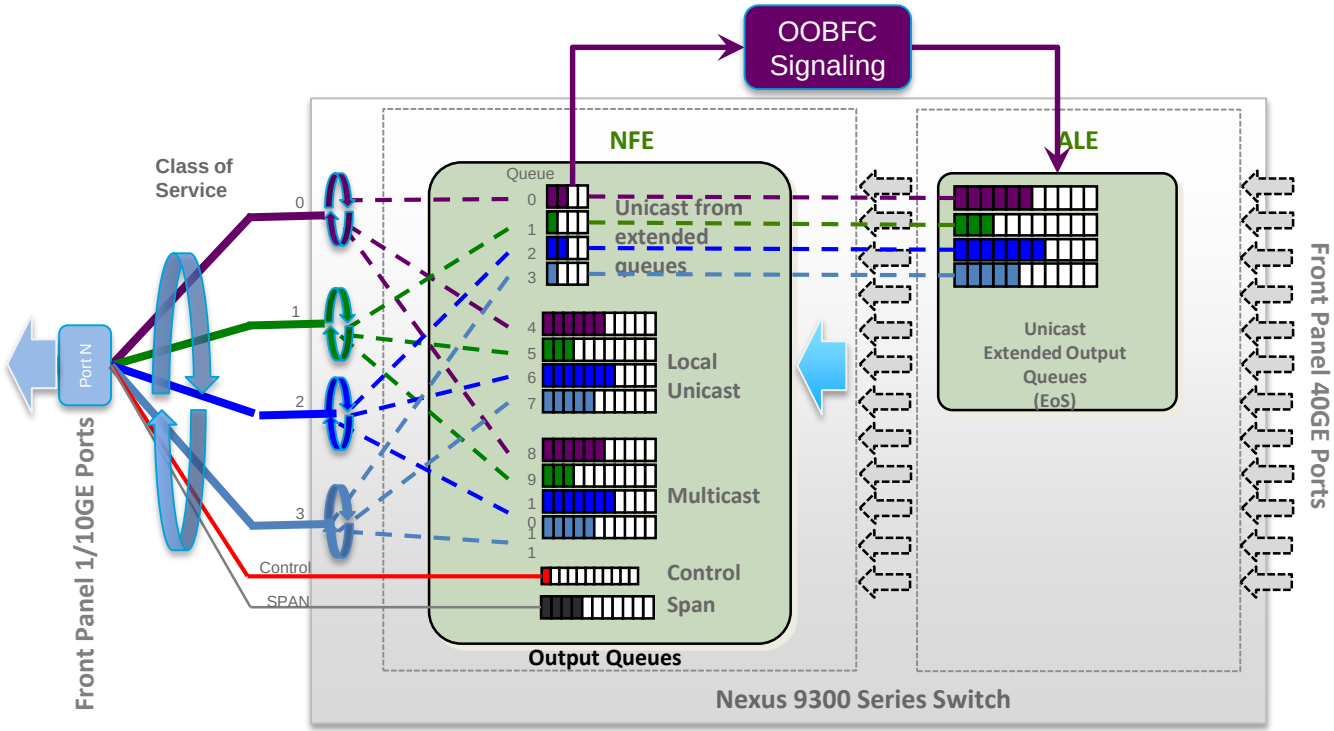
Nexus 9000 QoS and Buffering
NX-OS QoS

- **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

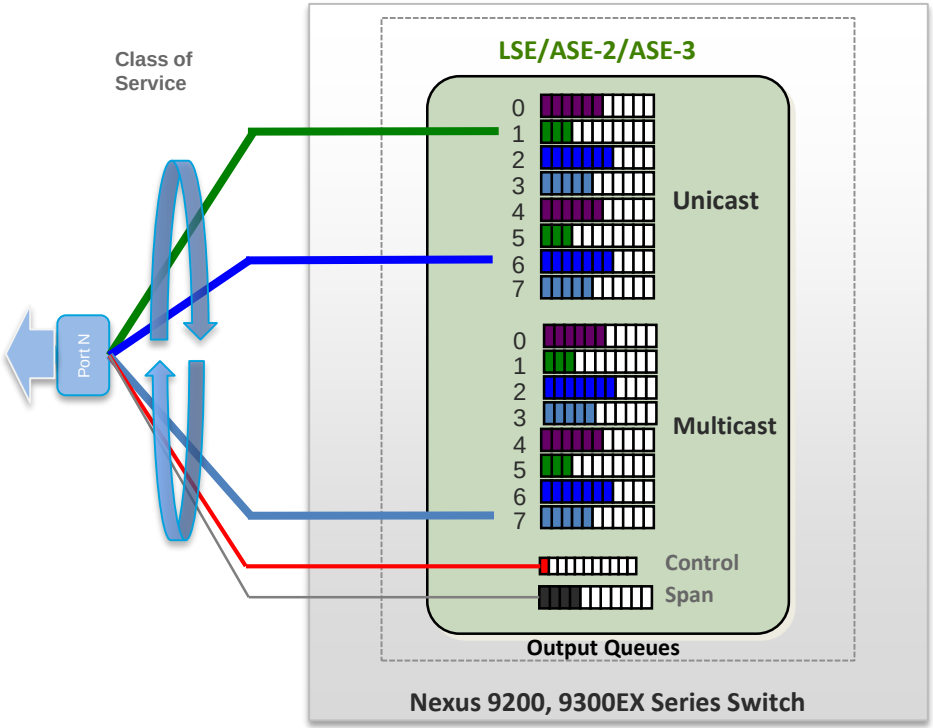
End-to-End QoS Implementation and Operation with Cisco Nexus Switches

Session ID: BRKDCT-3346

Queuing & Scheduling on First Gen Nexus 9300 Switches
4 Unicast + 4 Multicast + 2 Services Queues per Port

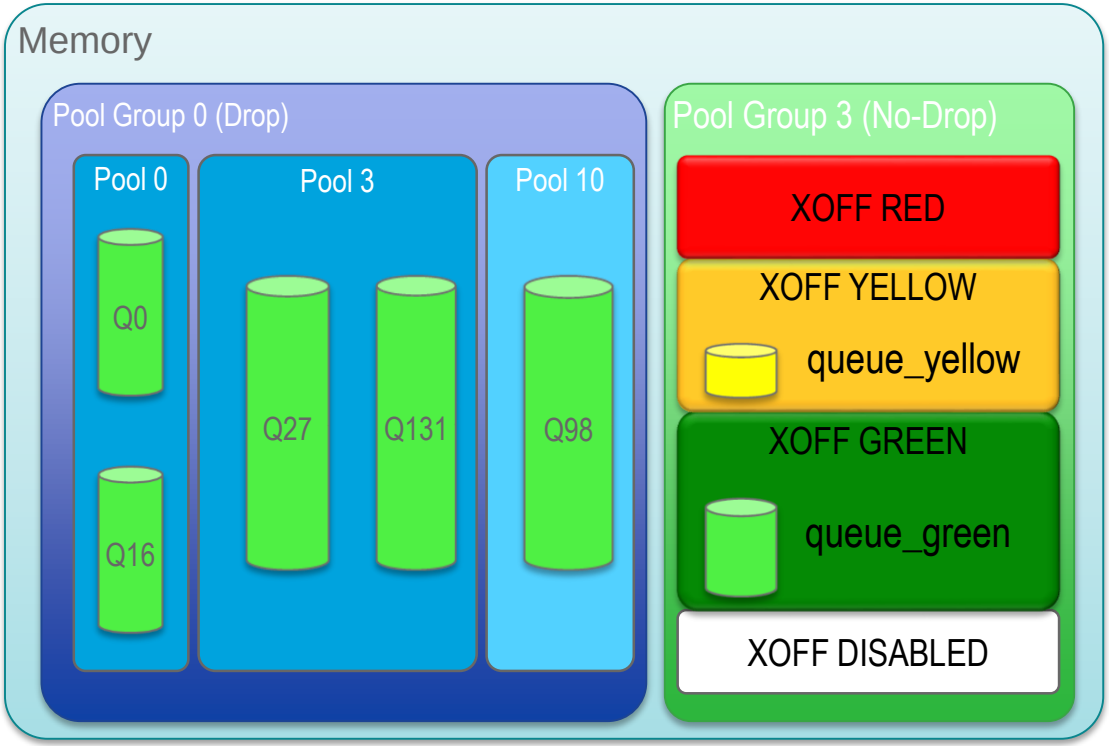


Queuing & Scheduling on 2nd Gen Nexus 9000 Switches
8 Unicast + 8 Multicast + 2 Services Queues per Port



- 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

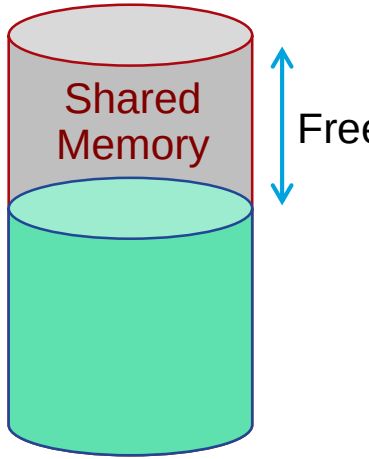
Shared Memory Buffering
Output Buffer Architecture



- Memory
- Pool Groups (PG)
  - Up to 4 Pool Groups
  - Can be Drop or No-drop
  - Static allocation
- Pools
  - 8 UC and 8 MC
  - Many:1 mapping of Pools:PG
- Queues
  - N queues per pool, where N = number of ports
  - Parameters defined by queue profile

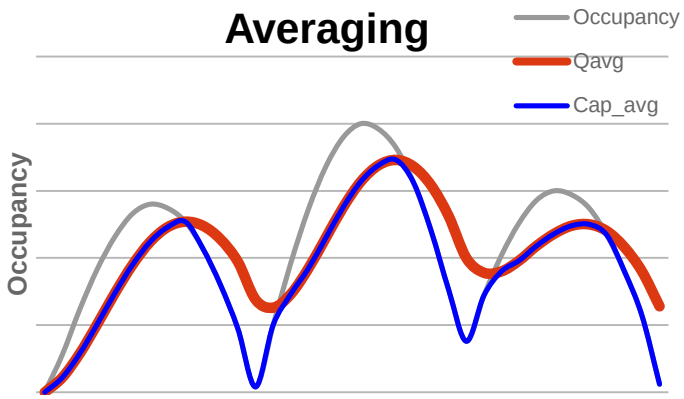
Shared Memory Buffering
Dynamic Buffer Protection (DBP)

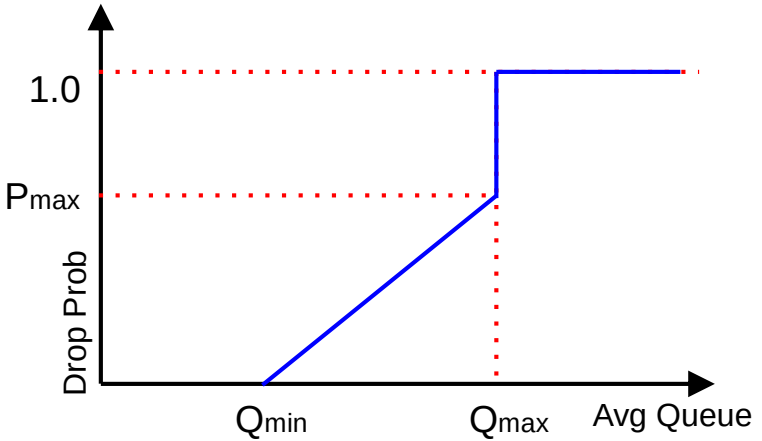
- **Requirement**
  - In a shared memory switch it is necessary to prevent any output queue from taking more than its faire share of the buffer when its output is oversubscribed
  - It can take more than its fair share to handle burst if the output is not oversubscribed.
- **Basic Algorithm (Deployed 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)

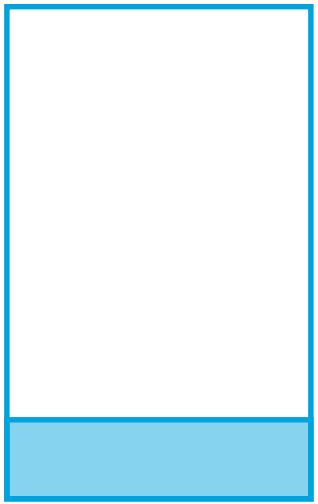
- AQM
  - Mode and parameters defined by profiles mapped to queues
  - Averaging timer per profile
  - Drop/ECN-mark per profile
- WRED
  - Each queue mapped to a profile
  - Averaging with Cap_Avg
- AFD
  - Drop/mark only elephant flows
  - Arrival rate measured by ETRAP
  - “Fair” rate computed using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





Buffering Data Centre
Two Requirements for Buffers

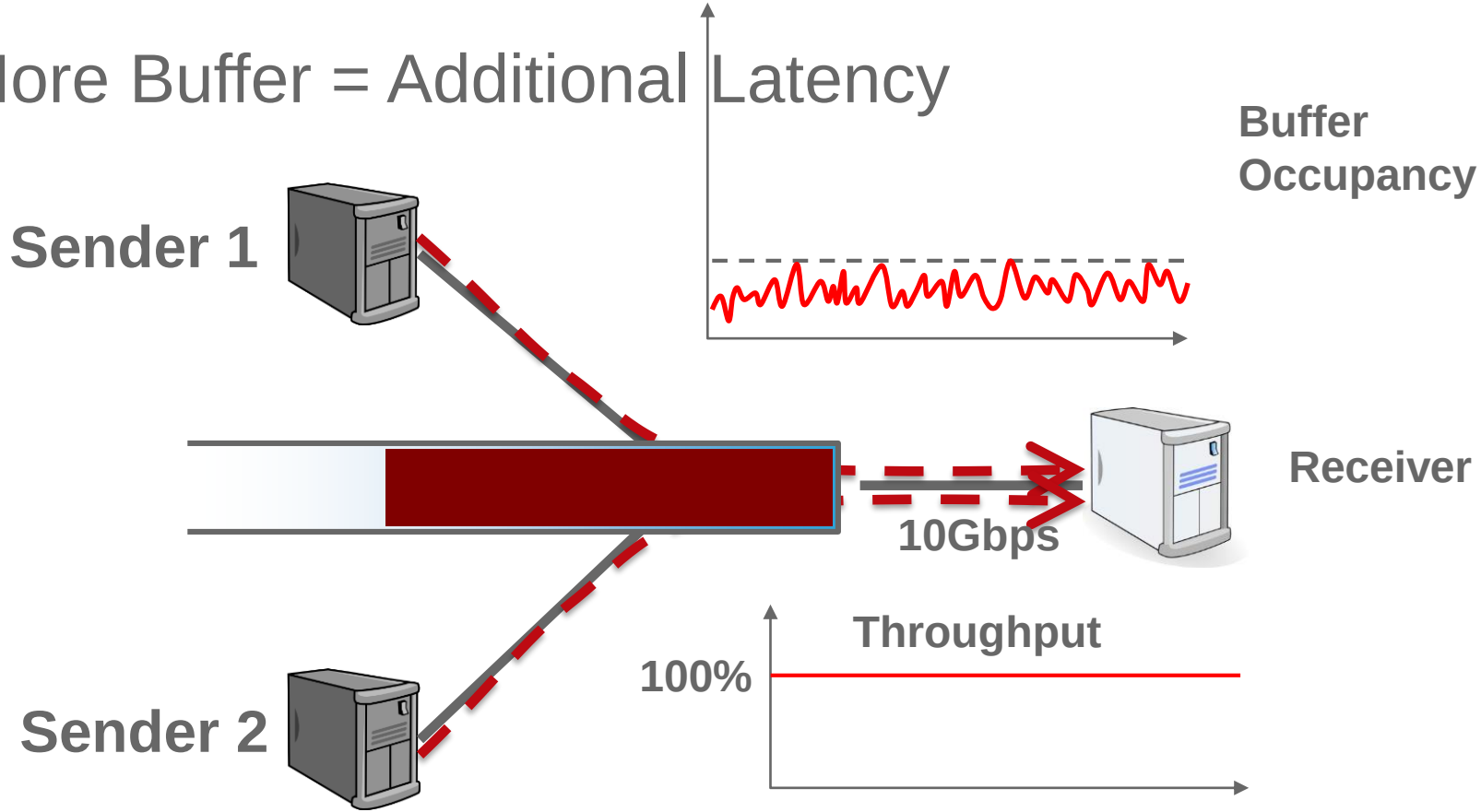
- Long Lived TCP flows
  - Maximize the utilization of the available network capacity (ensure links are able to run at line rate)
  - Window Size Increases to probe the capacity of the network
  - Delay x Bandwidth Product (C x RTT)*
    - e.g if your network had 100 micro-sec of latency with 10G interface,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



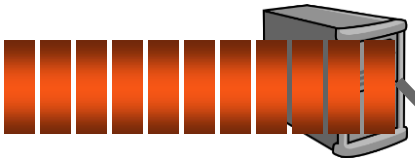
Buffer Available for Incast Burst

Buffer Required for Maximizing Network Utilization

More Buffer = Additional Latency



Elephants Waste Buffer



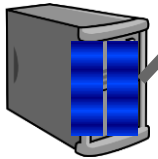
Sender 1

Elephants buildup large queues

- **No buffer left for mice**

Receiver

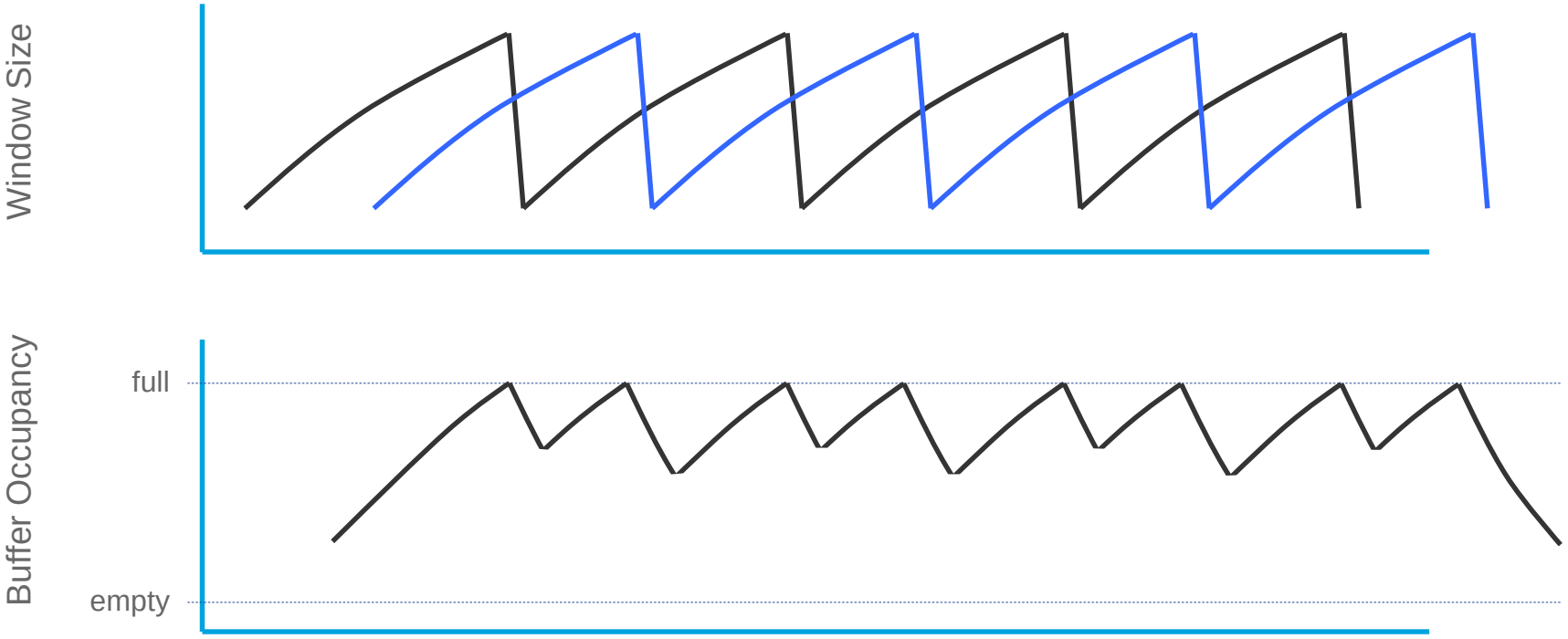
Sender 2



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

Applications performance suffers

Multiple TCP flows in reality



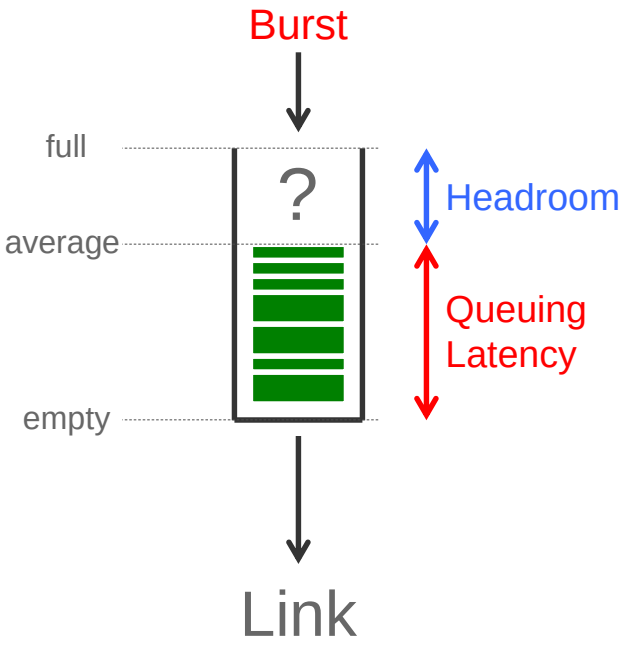
Long Lived TCP Flows
TCP Congestion Control and Buffer Requirements

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

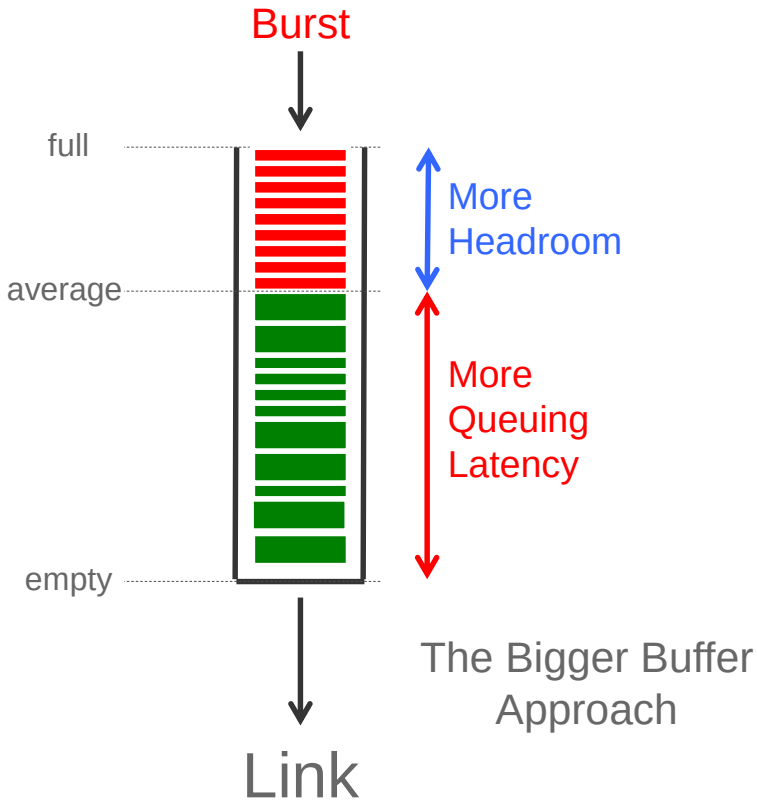
- Proven by theory and experiments in real operational networks
- For example, see Beheshti et al. 2008: "Experimental Study of Router Buffer Sizing"

Micro-bursts Need Headroom
Where does it come from?



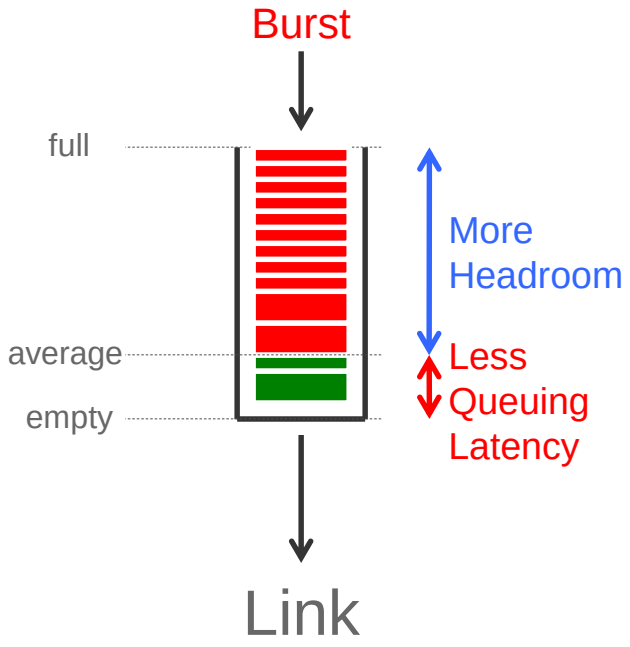
Micro-bursts Need Headroom Where does it come from?

- 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



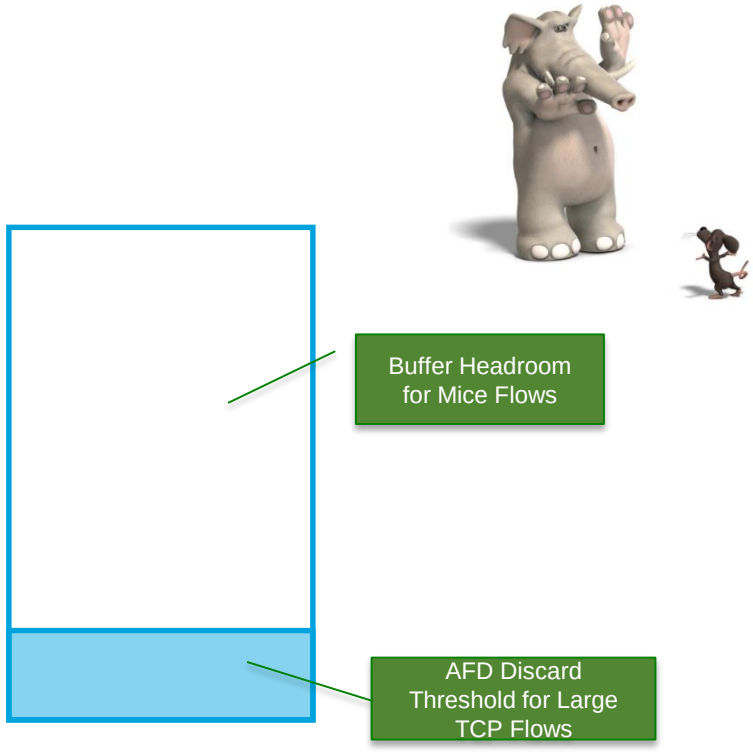
We want the best of both worlds

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

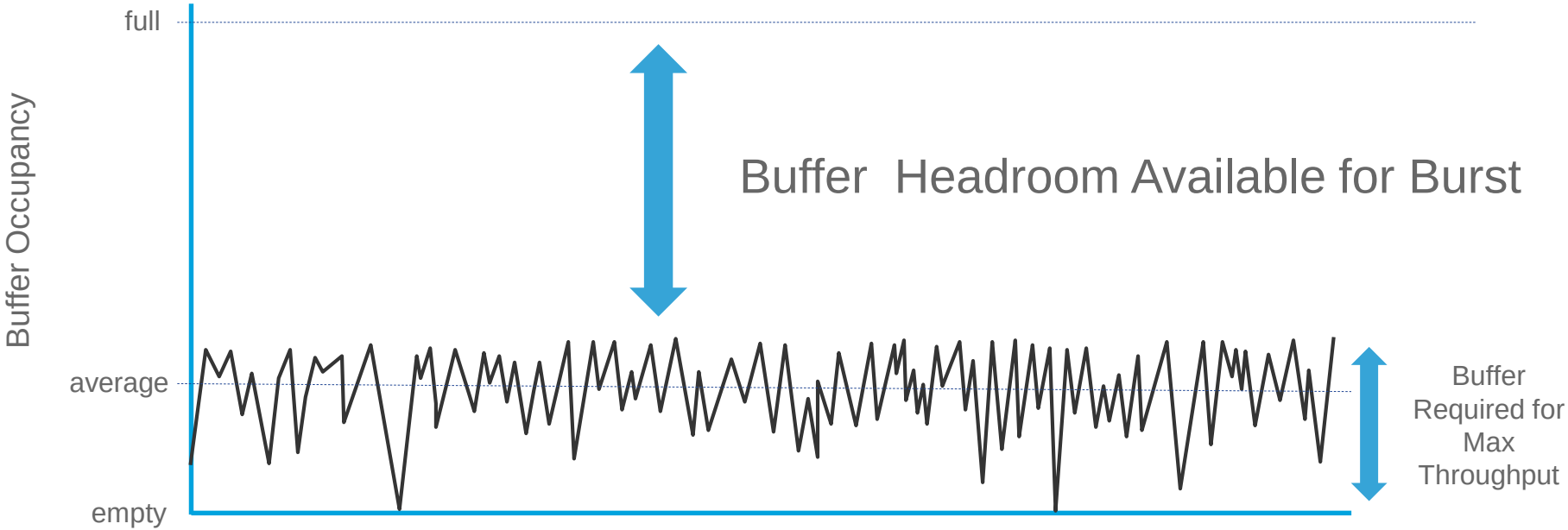


Innovation gives us the best of both worlds
AFD & DPP

- How to minimize 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) Prioritization (DPP)

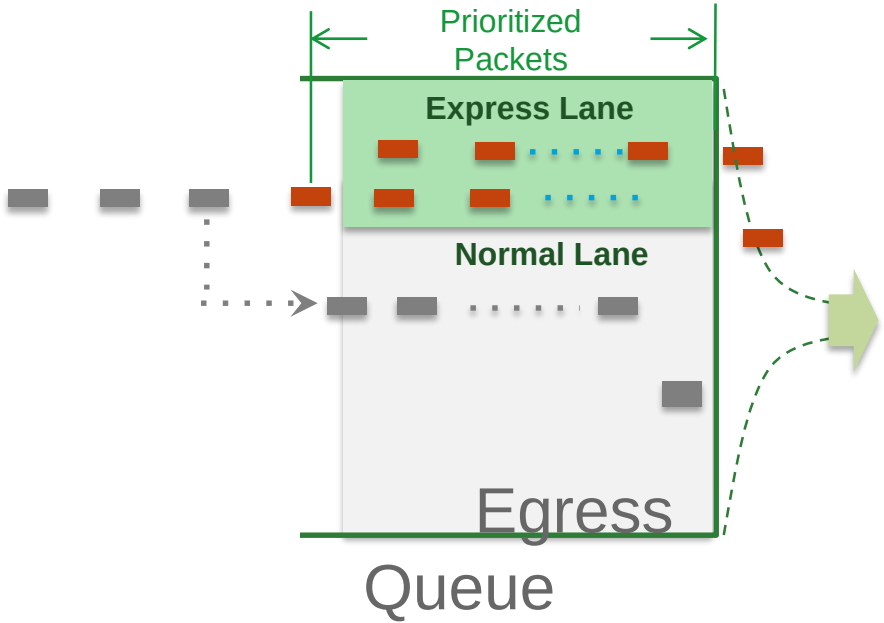


AFD Increases Headroom, Reduces Latency

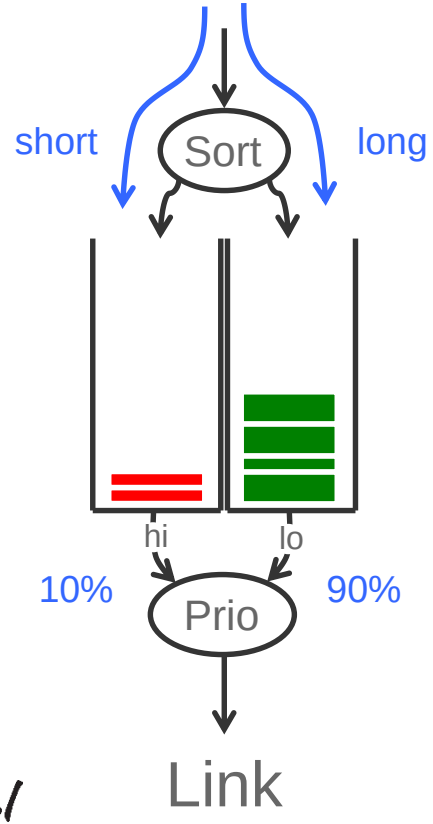


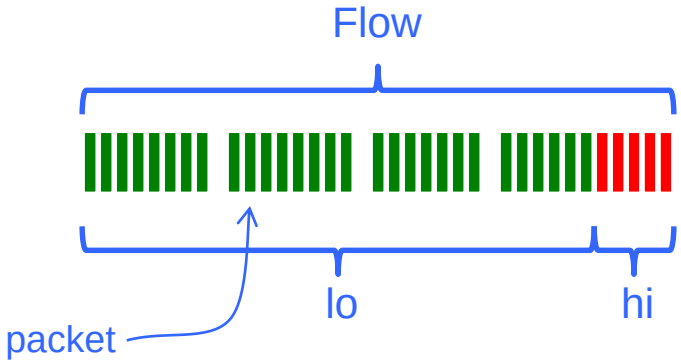
DPP (Dynamic Packet Prioritization)

- 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
- Prioritization 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





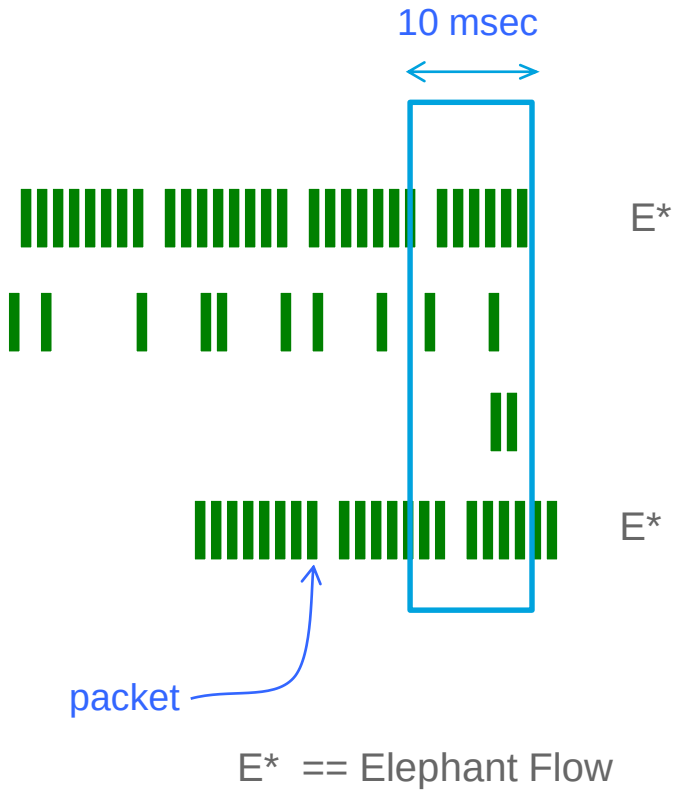
long

medium

short

Elephant Trap

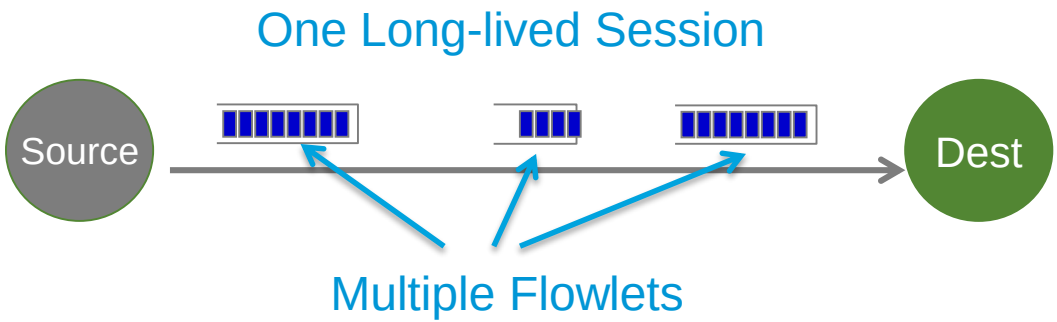
- Mechanism to identify large volume flows
  - Identified based on 5-tuple
- Elephant trap threshold is byte-count-based.
  - When received packets in a flow exceeds the number of bytes specified by the threshold, the flow is considered an elephant flow
  - Only elephant flows are submitted to AFD dropping 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



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

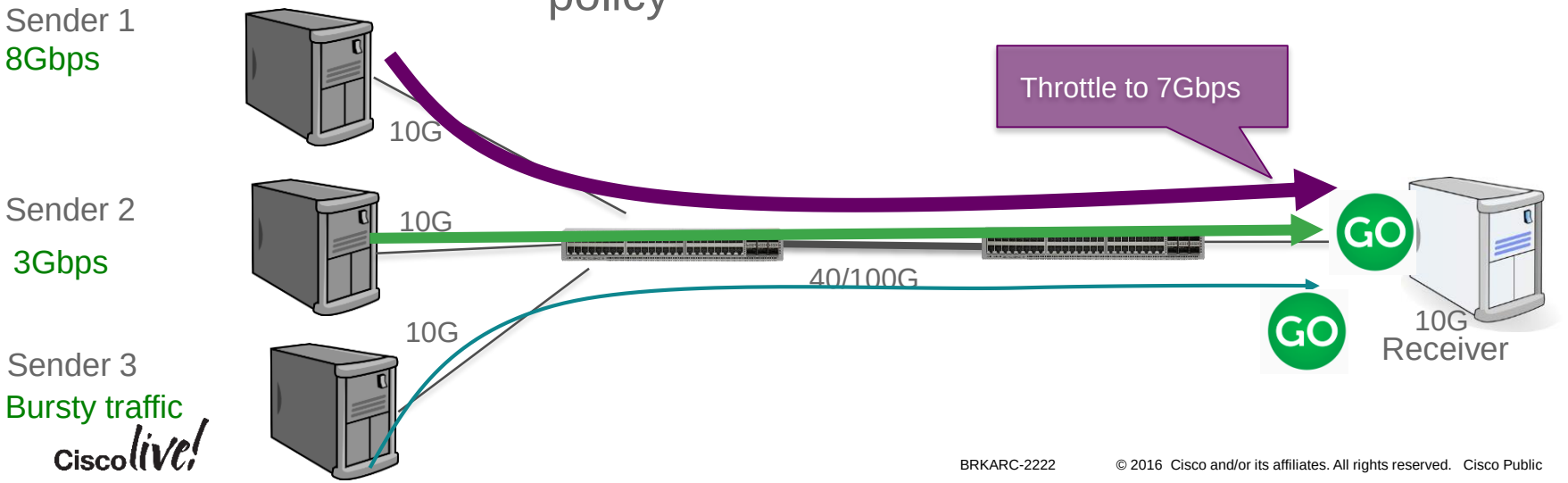
A Long-lived TCP Session ≠ An Elephant Flow

- 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 prioritized for that burst
  - Traffic arriving due to a link failure will be prioritized, etc …

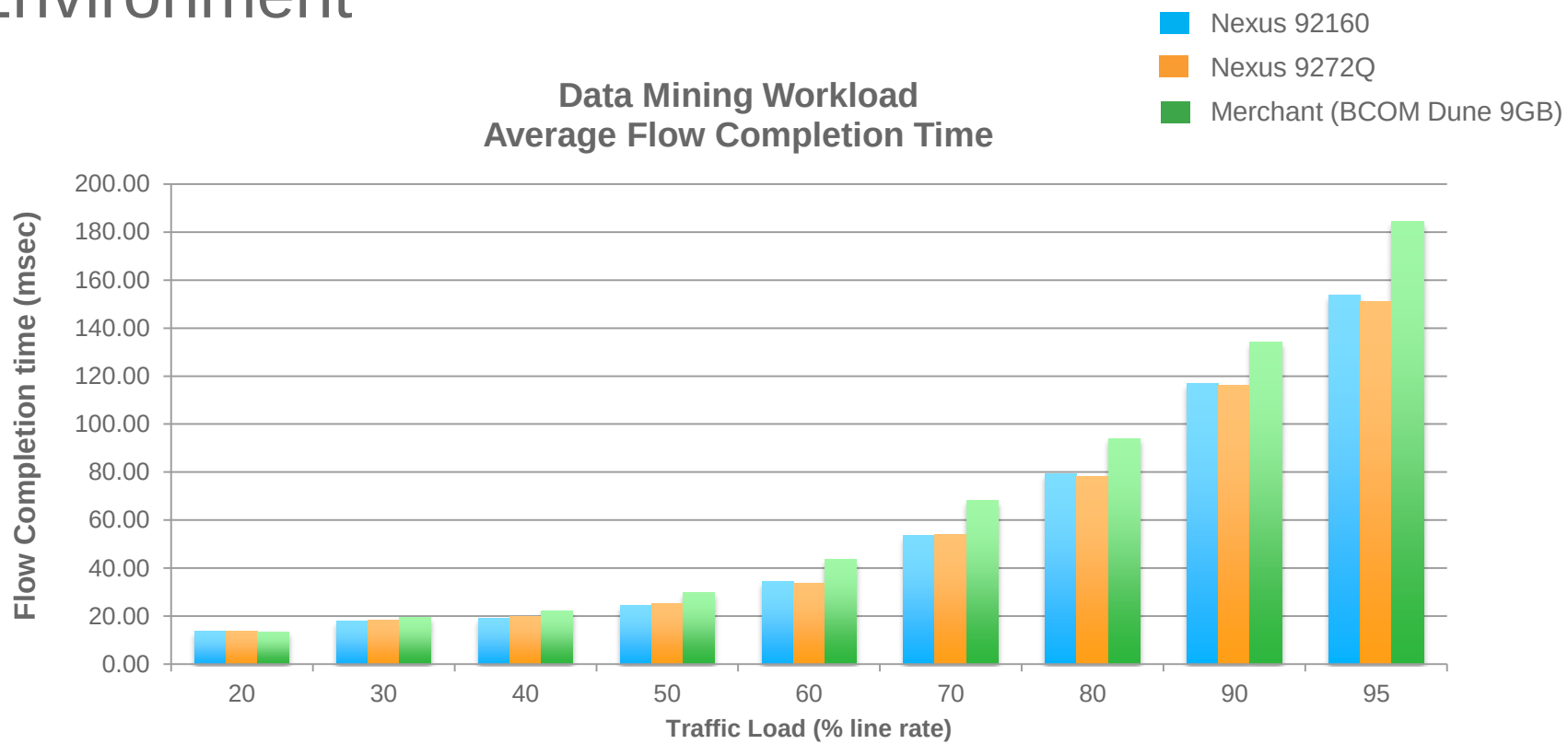


Protecting Bursty Traffic and Ensure Fairness

- Protecting bursty traffic with DPP
- Ensure fairness among big flows that belongs to same class of service.
- Traditional QoS is static. Hard to come up with QoS policy

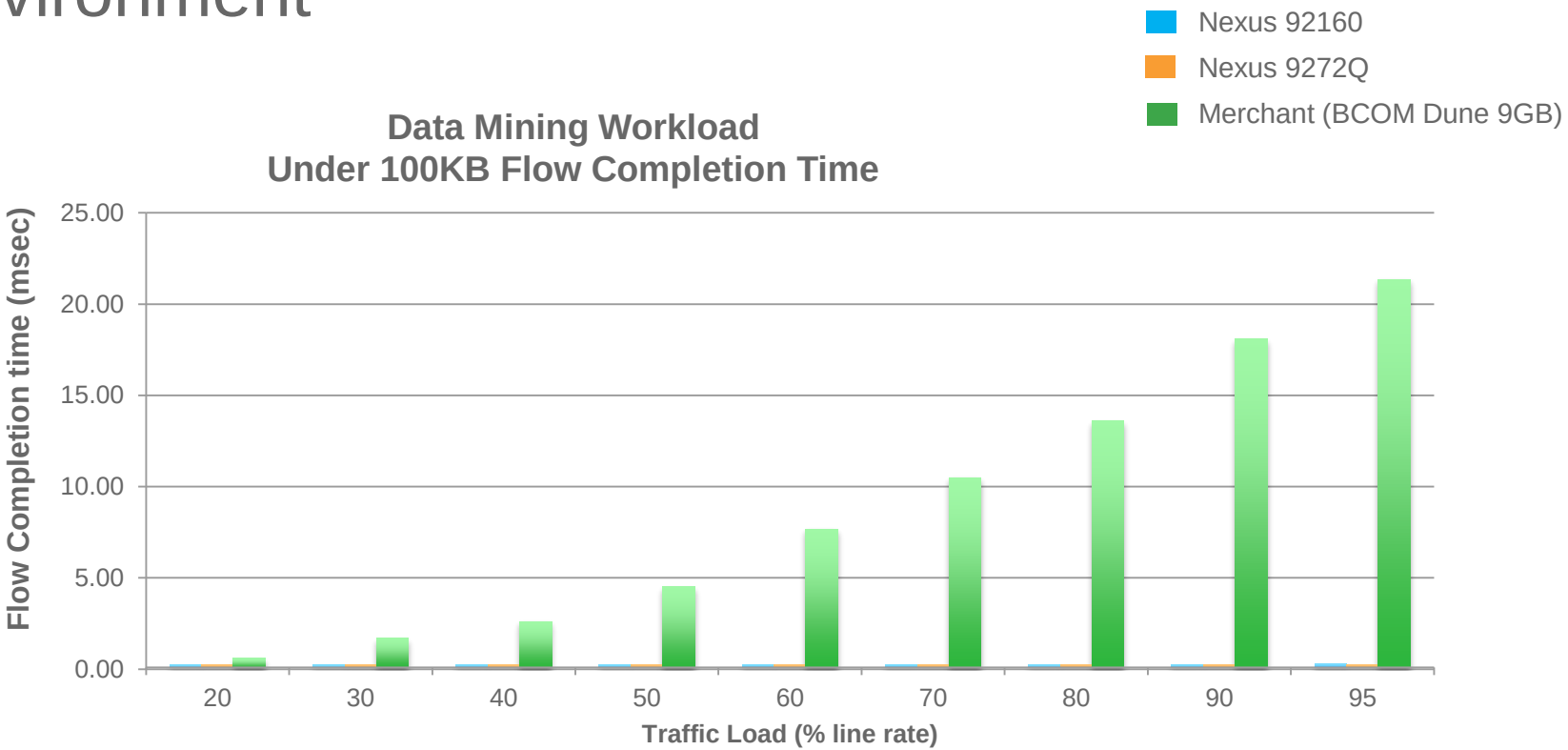


Better Application Performance in an Incast Environment



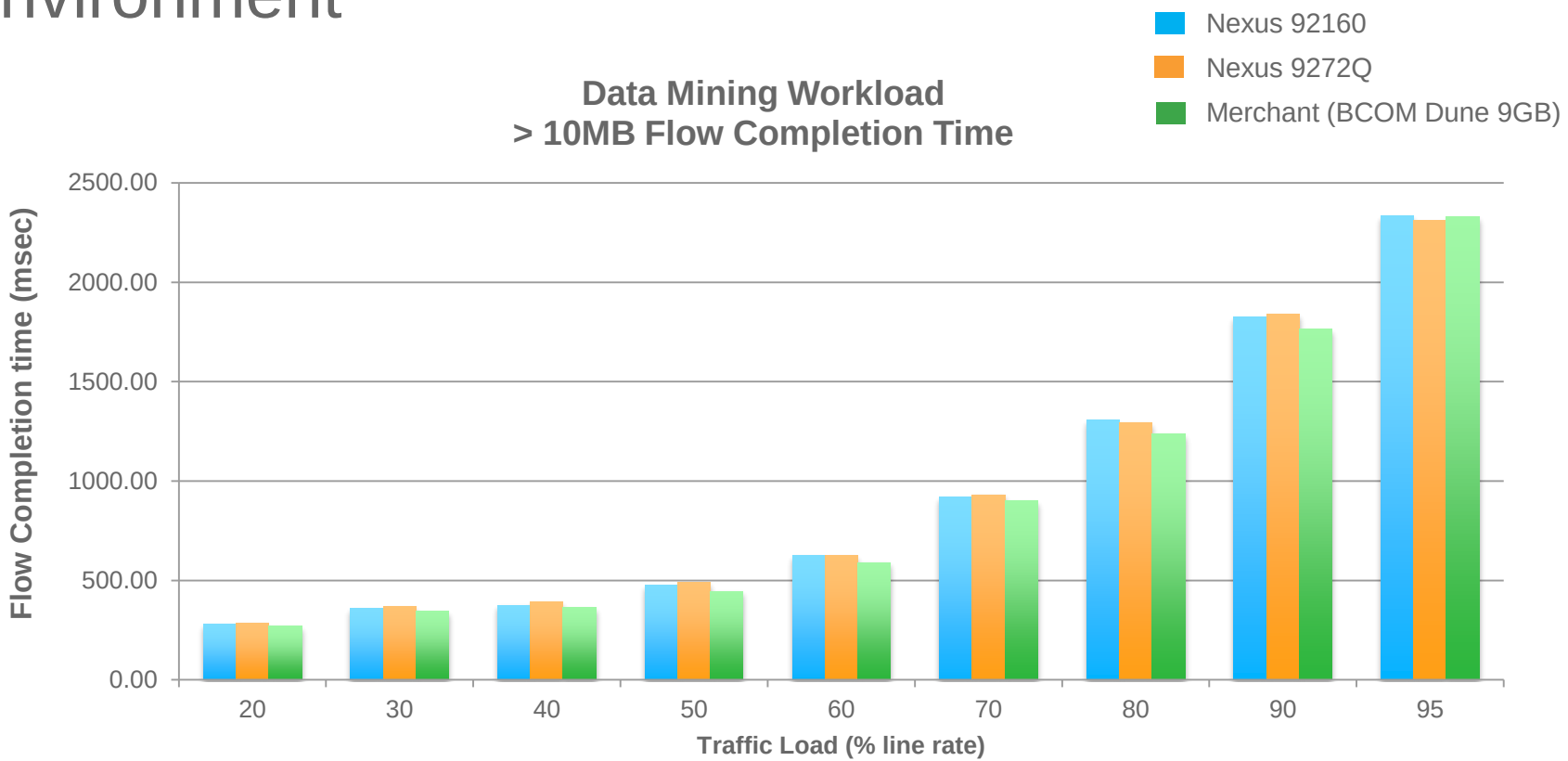
http://miercom.com/cisco-systems-speeding-applications-in-data-Centre-networks/

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



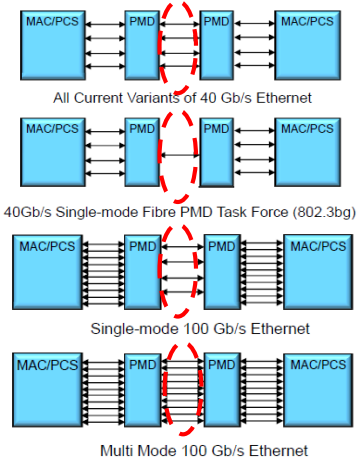
http://miercom.com/cisco-systems-speeding-applications-in-data-Centre-networks/

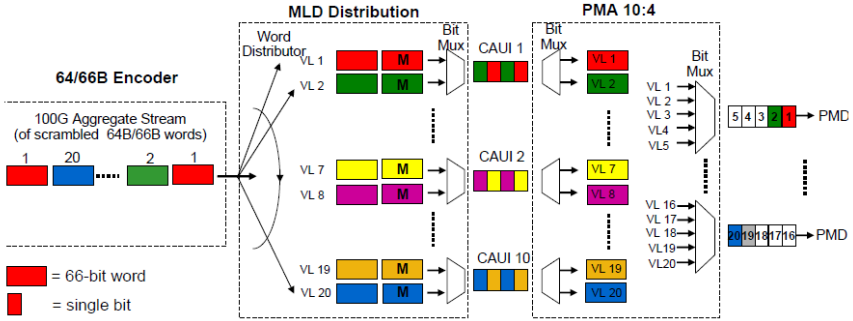
Agenda

- What's New
  - 2nd Generation Nexus 9000
  - Moore's Law
  - The new building blocks (ASE-2, ASE-3, LSE)
- Next Gen Nexus 9000 Switch Platforms
  - Nexus 9500 (Modular)
  - Nexus 9200/9300 (Fixed)
- Next Generation Capabilities
  - Forwarding, QoS, Telemetry
- 40G/100G Transceiver
  - 25G technology

Multi Lane Distribution (MLD)

MLD (Multi Lane Distribution)



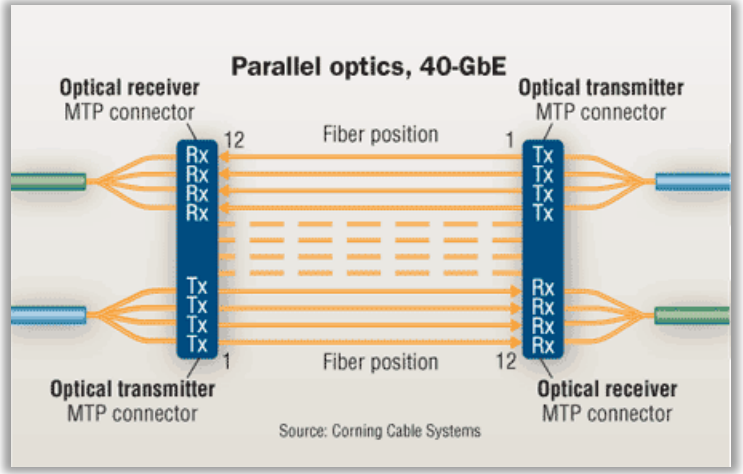


- 40GE/100GE interfaces have multiple lanes (coax cables, fibers, wavelengths)
- MLD provides a simple (common) way to map 40G/100G to physical interfaces of different lane widths

QSFP and QSFP28 Parallel Lanes

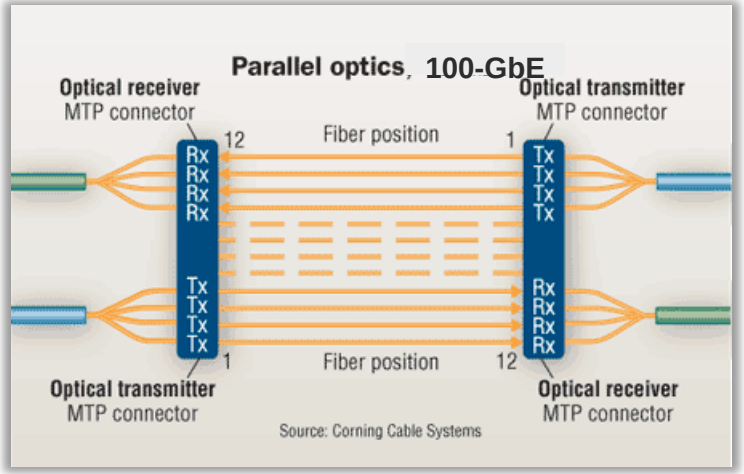
4 x10 = 40G shifts to 4 x 25 = 100G

- Same form factor for QSFP and QSFP28
- Same cable plant for QSFP and QSFP28



QSFP

Backed by 10G SerDes



QSFP28

Backed by 25G SerDes

Optics Pluggable Multispeed Interfaces SFP & QSFP

SFP

Pluggable Options

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

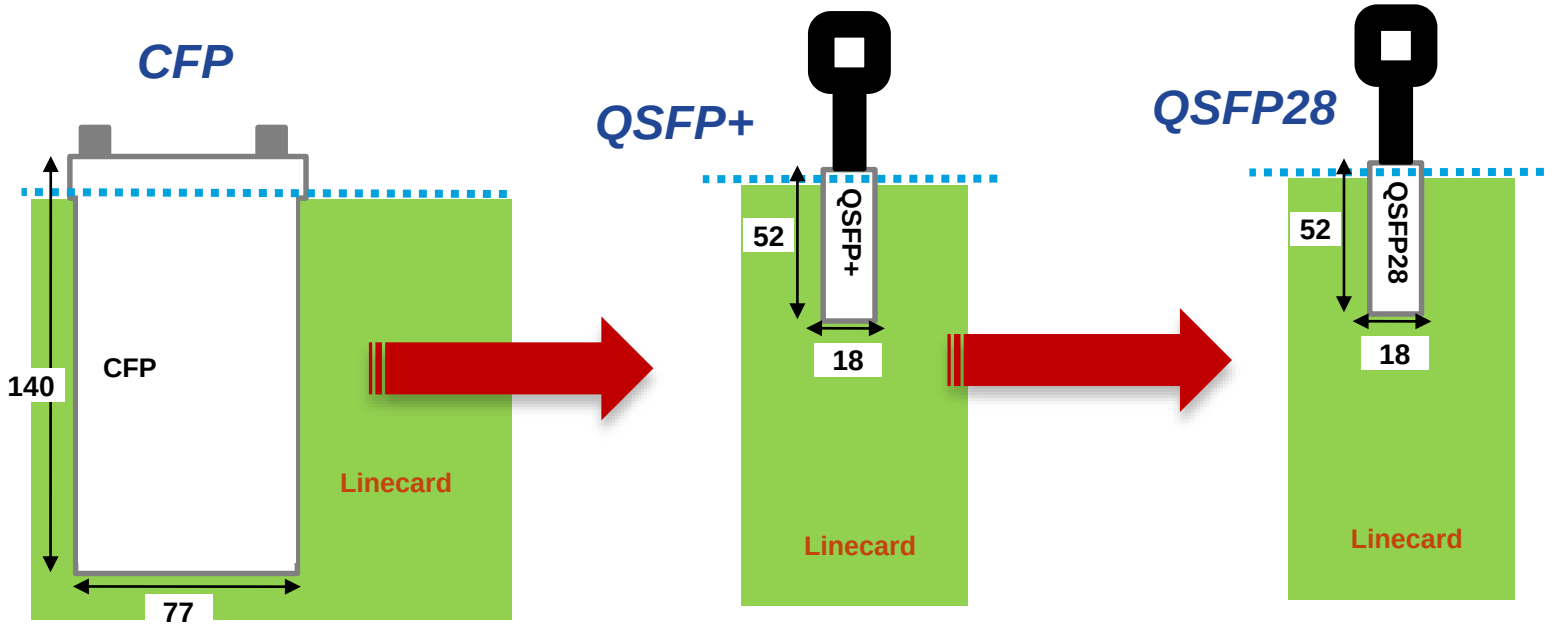
QSFP

QSFP28

Pluggable Options

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

Next Generation Packages for 40/100G
QSFP+ & QSFP28



QSFP28

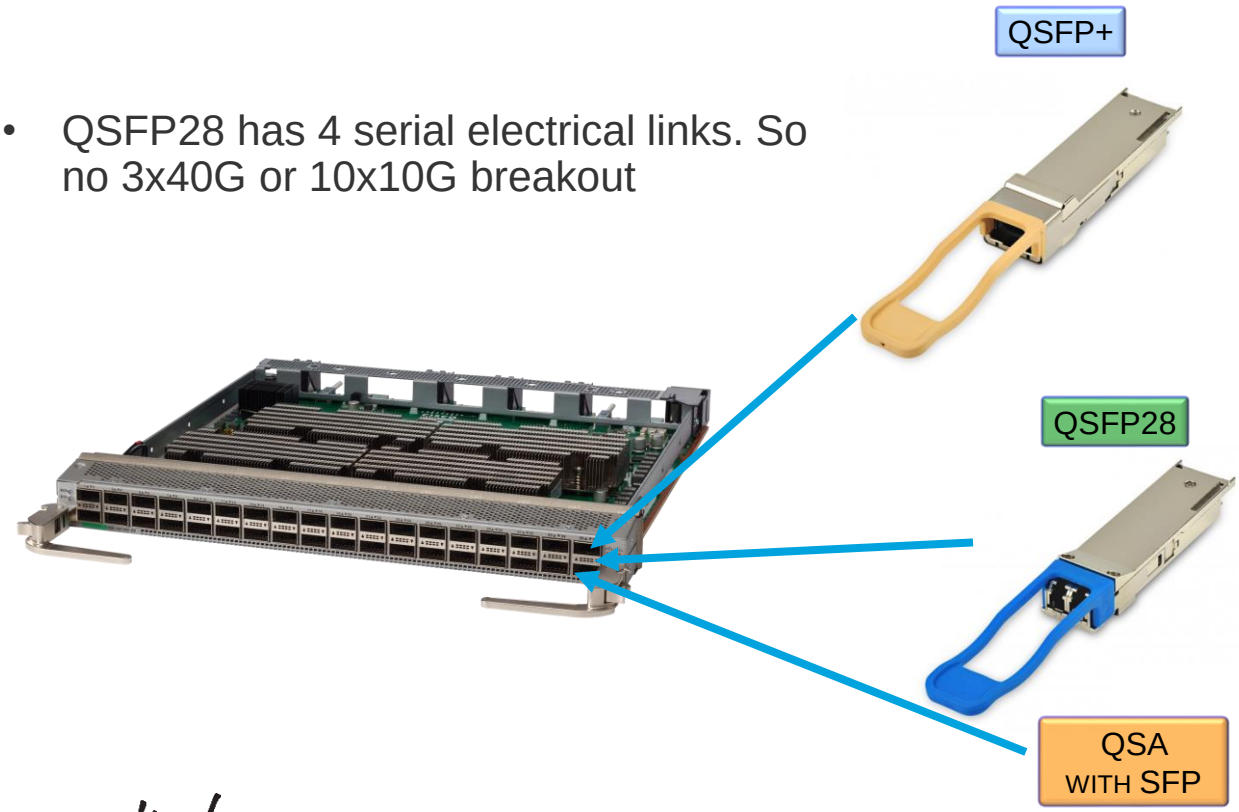
1/2 the power & 1/5 the size of CPAK

44% the size of CFP4

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

Multiple Speeds with QSFP28 interface

- QSFP28 has 4 serial electrical links. So no 3x40G or 10x10G breakout



✓ 1x40G
✓ 4x10G

✓ 1x100G
✓ 4x25G
✓ 2x50G

✓ 1x10G
✓ 1x1G

Support for 40G Optics QSFP+

100m, MMF

SR4 QSFP+

10km, SMF

LR4 QSFP+

40km, SMF

ER4 QSFP+

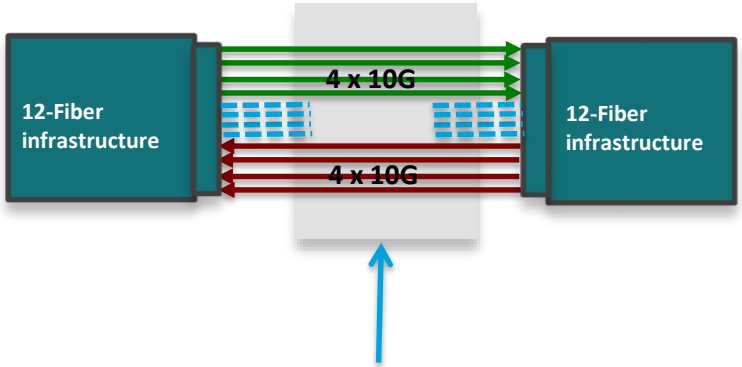
2km, SMF

WSP-Q40GLR4L QSFP+

10m, Copper

4x10G AOC QSF+s

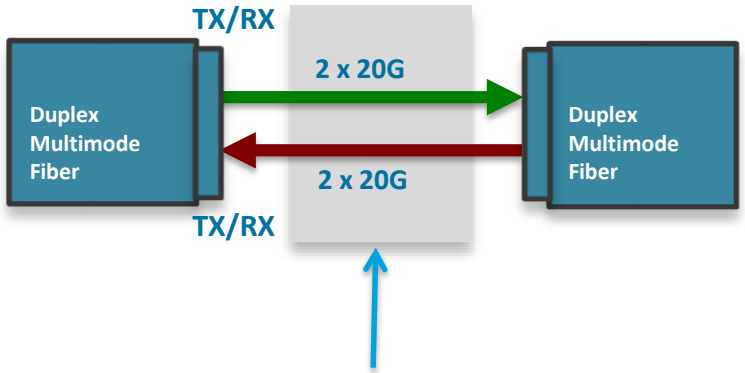
QSFP-BIDI vs. QSFP-40G-SR4



12-Fiber ribbon cable with MPO connectors at both ends

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

QSFP-BIDI



Duplex multimode fiber with Duplex LC connectors at both ends

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

Support for 40G Optics

QSFP+	Fiber	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, SMF

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	Fiber	Connectors	Distance
SR4	MMF	MPO-MTP12	Up to 100m
LR4	SMF	LC pair	Up to 10km
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

Cisco 100G QSFP28 Optics Portfolio

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

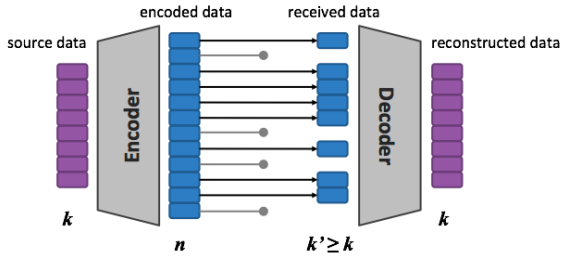
Optics Type	Description	Connector	Availability
QSFP-100G-SR-BD	**40/100G, 100m**	**MMF LC**	2HCY16
QSFP-100G-SR4-S	100GBASE-SR4, 100m	MMF MPO	Q4CY15
QSFP-100G-LR4-S	100GBASE-LR4, 10km	SMF LC	Q4CY15
QSFP-100G-CWDM4-S	100GE CWDM4, 2km	SMF LC	Q4CY15
QSFP-100G-PSM4-S	100GBASE-PSM4, 2km	SMF MPO	Q4CY15
QSFP-100G-CU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

What about 25G?
25 Gbps Ethernet & Standards

	Consortium	IEEE
Distance	Passive: 1,2,3 meter	Passive: 1,2,3,5 meter Optics: SR
Deployment	Within Rack	Across Rack
Supporting Platform	N9200, N9300-EX N3200	Roadmap N9300-EXU
25G NIC (Verified)	Mellanox	None available yet
25G NIC (Ongoing Testing)	Qlogic, BRCM	

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 optimize 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 realized BER with 56G PAM4 encoding

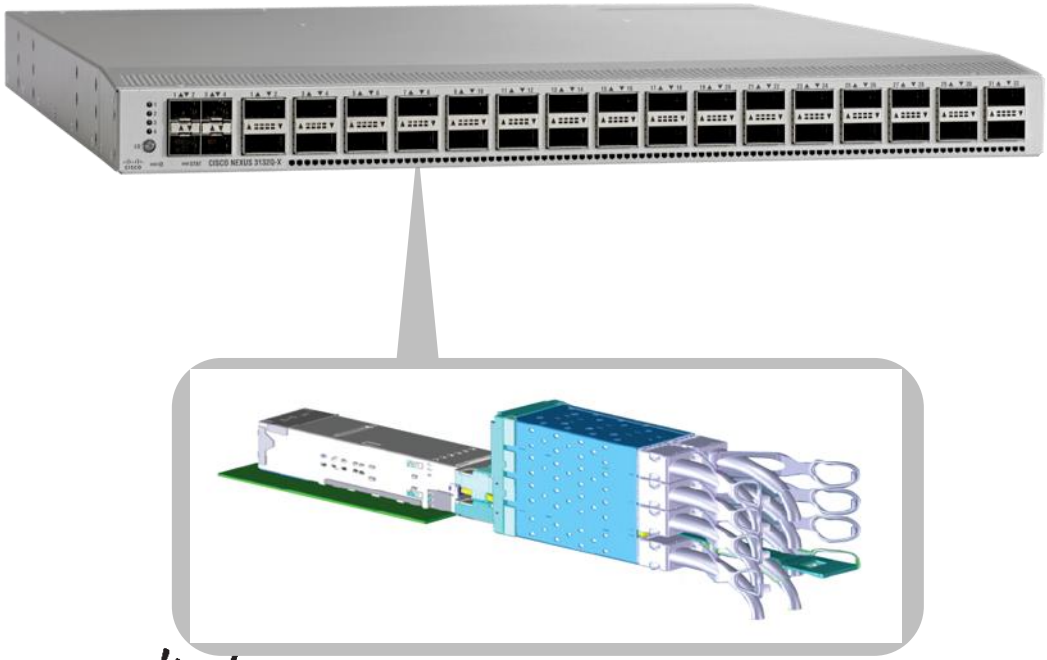
25G / 10G backward compatibility

- 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

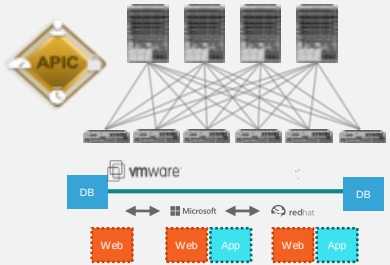
32p 100G -> 48p 50G & 8p 100G

No break-out cable

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

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

Application Centric Infrastructure

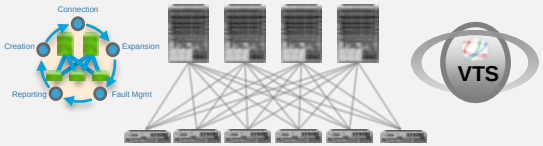


Turnkey integrated solution with security, centralized management, compliance and scale

Automated application centric-policy model with embedded security

Broad and deep ecosystem

Programmable Fabric

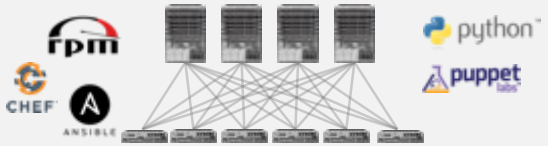


VxLAN-BGP EVPN standard-based

3rd party controller support

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

Programmable Network



Modern NX-OS with enhanced NX-APIs

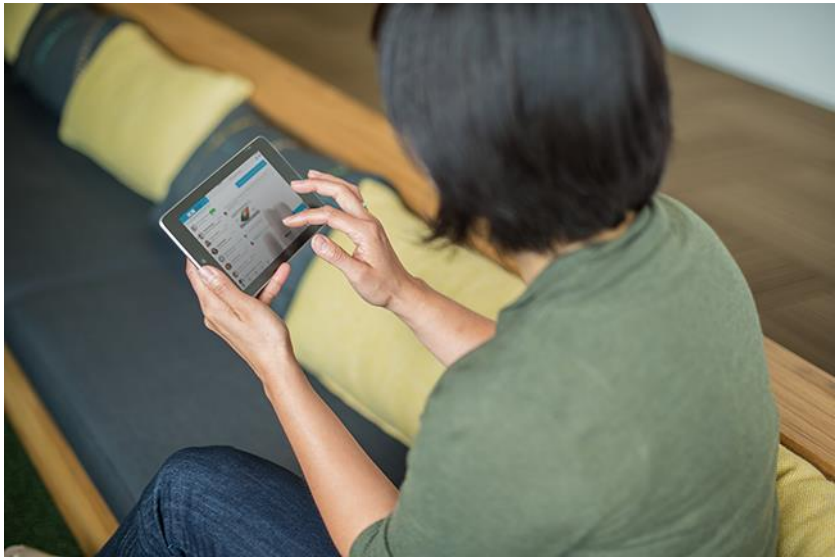
DevOps toolset used for Network Management
(Puppet, Chef, Ansible etc.)

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

Nexus 9700EX + 9300EX

Complete Your Online Session Evaluation

- Give us your feedback to be entered into a Daily Survey Drawing. A daily winner will receive a $750 Amazon gift card.
- Complete your session surveys through the Cisco Live mobile app or from the Session Catalog on CiscoLive.com/us.



Don’t forget: Cisco Live sessions will be available for viewing on-demand after the event at CiscoLive.com/Online

Continue Your Education

- Demos in the Cisco campus
- Walk-in Self-Paced Labs
- Lunch & Learn
- Meet the Engineer 1:1 meetings
- Related sessions

Thank you


cisco
Cisco live!
July 10-14, 2016 • Las Vegas, NV