
You make possible

Nexus 9000 Architecture

Tim Stevenson
Distinguished Engineer, Technical Marketing
BRKARC-3222

Session Abstract

This session presents an in-depth study of the architecture of the latest generation of Nexus 9000 modular and top-of-rack data center switches. Topics include forwarding hardware, switching fabrics, and other physical design elements, as well as a discussion of key hardware-enabled features and capabilities that combine to provide high-performance data center network services.

What This Session Covers

- Latest generation of Nexus 9000 switches with Cloud Scale ASICs
- Nexus 9500 modular switches with Cloud Scale linecards
- Nexus 9300 Cloud Scale top-of-rack (TOR) switches
- System and hardware architecture, key forwarding functions, packet walks

Not covered:

- First generation Nexus 9000 ASIC/platform architectures
- Nexus 9500 merchant-silicon based architectures
- Other Nexus platforms
- *Catalyst* 9000 platform

Agenda

- Data Center and Silicon Strategy
- Cloud Scale ASIC Architecture
- Cloud Scale Switching Platforms
- Packet Walks
- Key Takeaways

Cisco Data Center Strategy

Application Centric

Capture Intent

Multicloud Ready

Platform Innovation

Nexus 9000 Cloud Scale Switching Portfolio

Key Elements of the Data Center Strategy

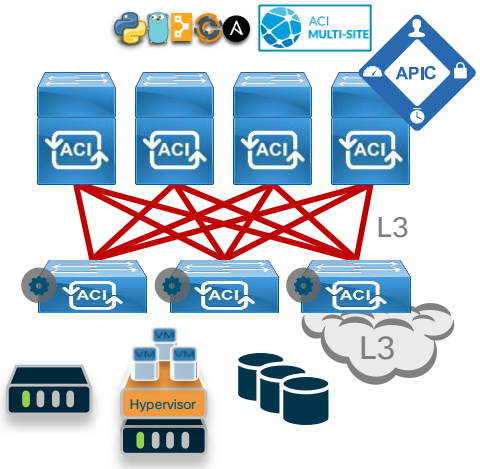
Nexus 9300-EX and 9300-FX/FX2

Premier TOR platforms

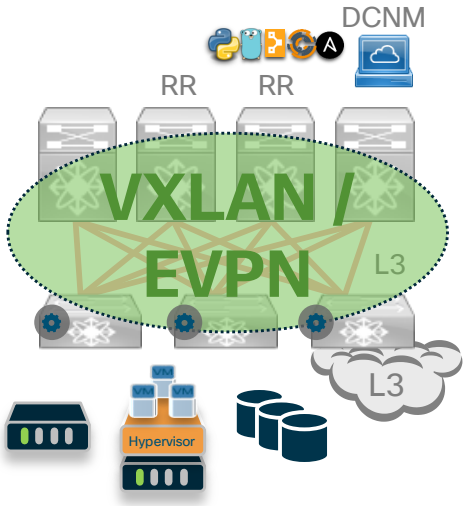
Nexus 9500 with X9700-EX and X9700-FX Modules

Flagship switching modules for Nexus 9500 modular chassis

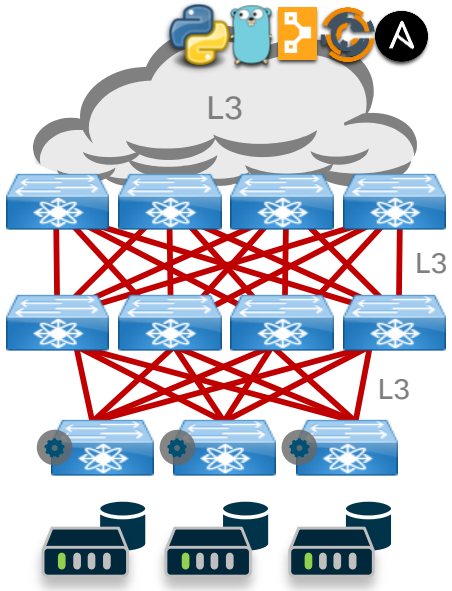
Building Data Center Fabrics with Nexus 9000



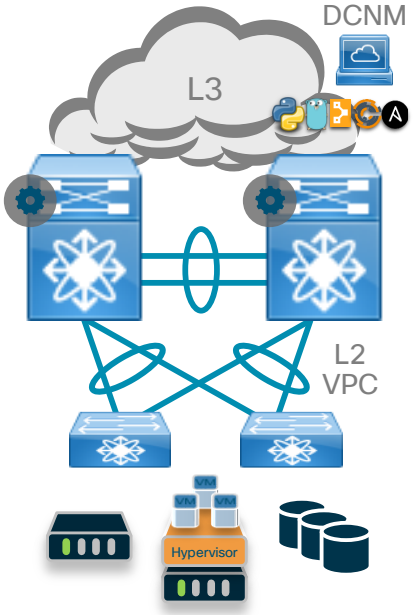
ACI – Turnkey Fabric



Standalone – Programmable Fabric with VXLAN+EVPN

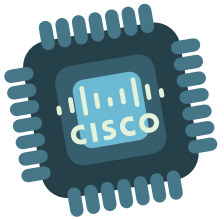


Standalone – Programmable IP Network



Standalone – Traditional Data Center Network

Why Custom Silicon?



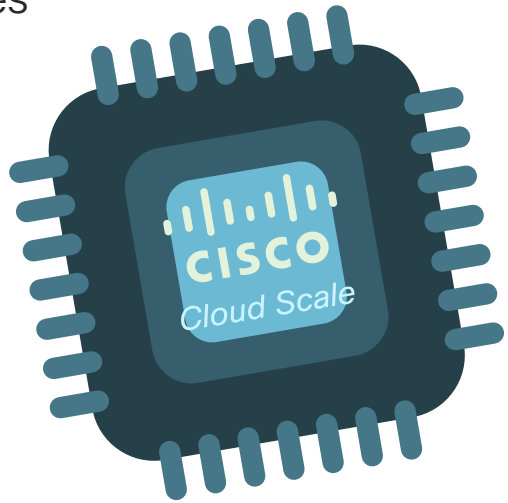
- Cisco competitive advantage – vehicle for differentiating innovations
  - ACI policy model + congestion-aware flowlet switching
  - Flexible forwarding tiles
  - Single-pass tunnel encapsulations
  - In-built encryption technologies
  - Intelligent buffers
  - Streaming hardware telemetry
- Tight integration between hardware / software / marketing / sales / support
- Closely aligns hardware designs with software innovations, strategic product direction, competitive differentiators, serviceability

Agenda

- Data Center and Silicon Strategy
- Cloud Scale ASIC Architecture
- Cloud Scale Switching Platforms
- Packet Walks
- Key Takeaways

Cisco Cloud Scale ASIC Family

- **Ultra-high port densities** → Reduces equipment footprint, enables device consolidation, denser fabric designs
- **Multi-speed 100M/1/10/25/40/50/100G** → Flexibility and future proofing
- **Rich forwarding feature-set** → ACI, Segment Routing, single-pass L2/L3 VXLAN routing
- **Flexible forwarding scale** → Single platform, multiple scaling alternatives
- **Intelligent buffering** → Shared egress buffer with dynamic, advanced traffic optimization
- **In-built analytics and telemetry** → Real-time network visibility for capacity planning, security, and debugging



Cloud Scale Family Members

LSE

- 1.8T chip – 2 slices of 9 x 100G each
- X9700-EX modular linecards; 9300-EX TORs

LS1800FX

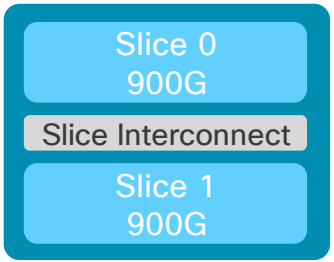
- 1.8T chip – 1 slice of 18 x 100G with MACSEC
- X9700-FX modular linecards; 9300-FX TORs

S6400

- 6.4T chip – 4 slices of 16 x 100G each
- 9364C/9332C TORs; E2 fabric modules

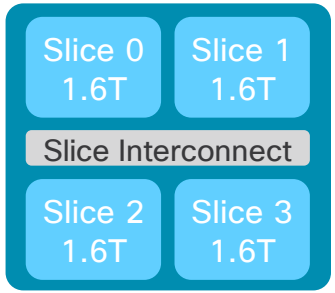
LS3600FX2

- 3.6T chip – 2 slices of 18 x 100G with MACSEC + CloudSec
- 9300-FX2 TORs

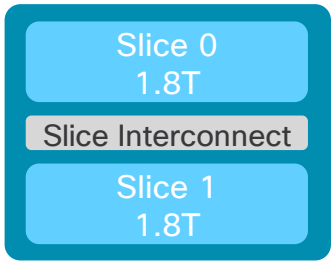


LSE – 18 x 100G

LS1800FX – 18 x 100G



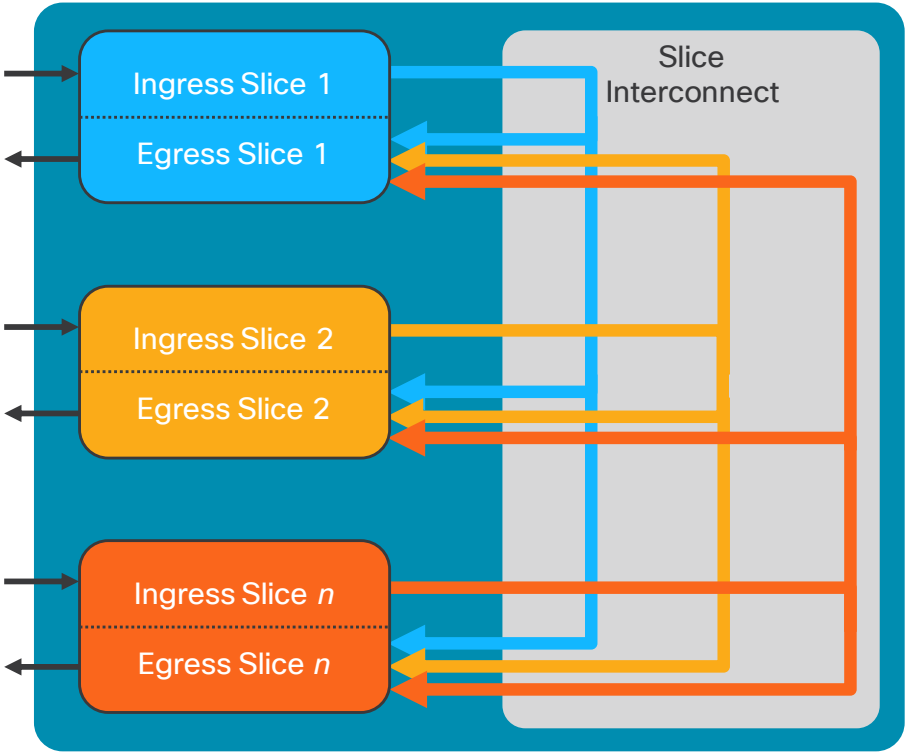
S6400 – 64 x 100G



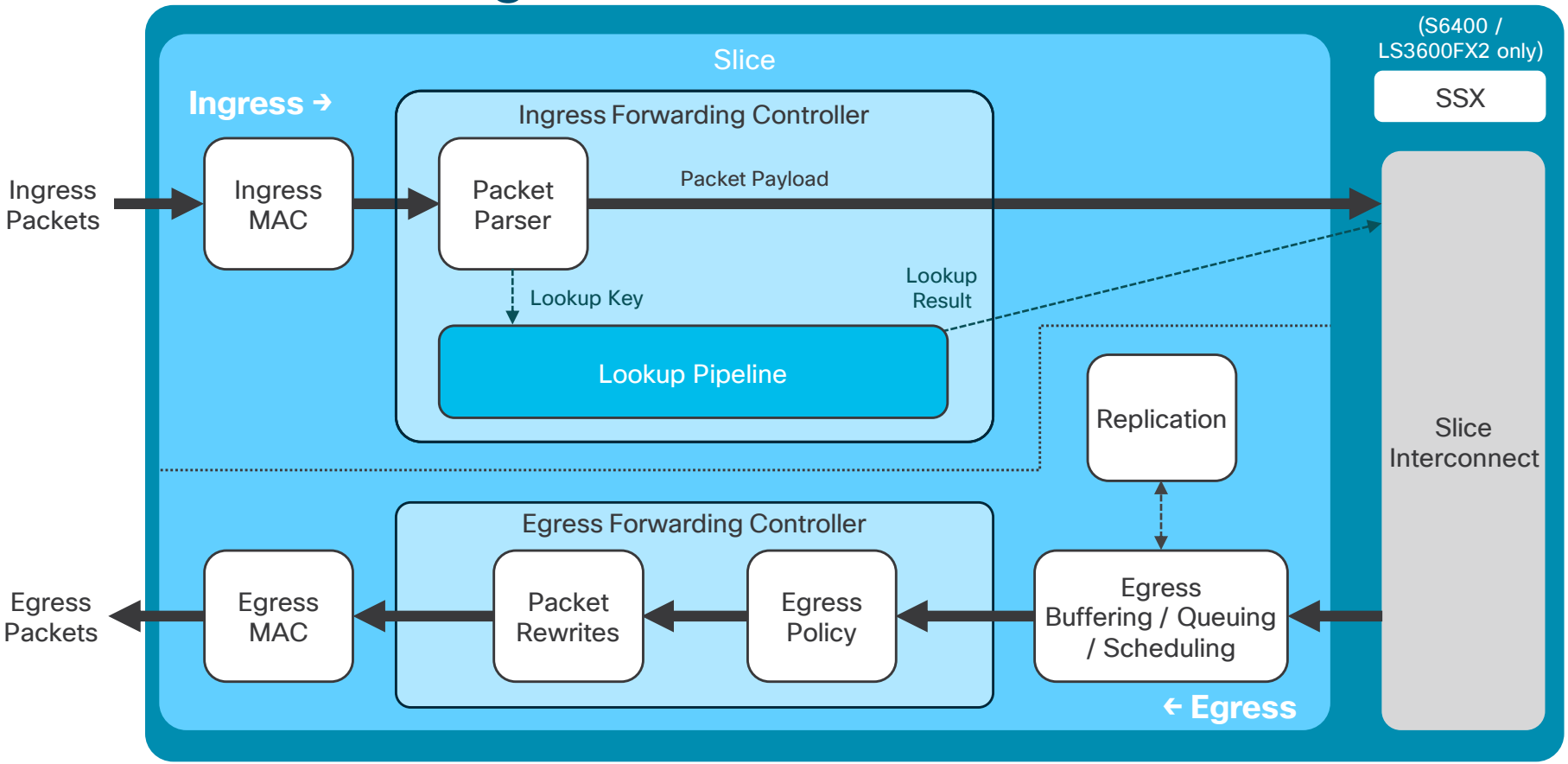
LS3600FX2 – 36 x 100G

What Is a “Slice”?

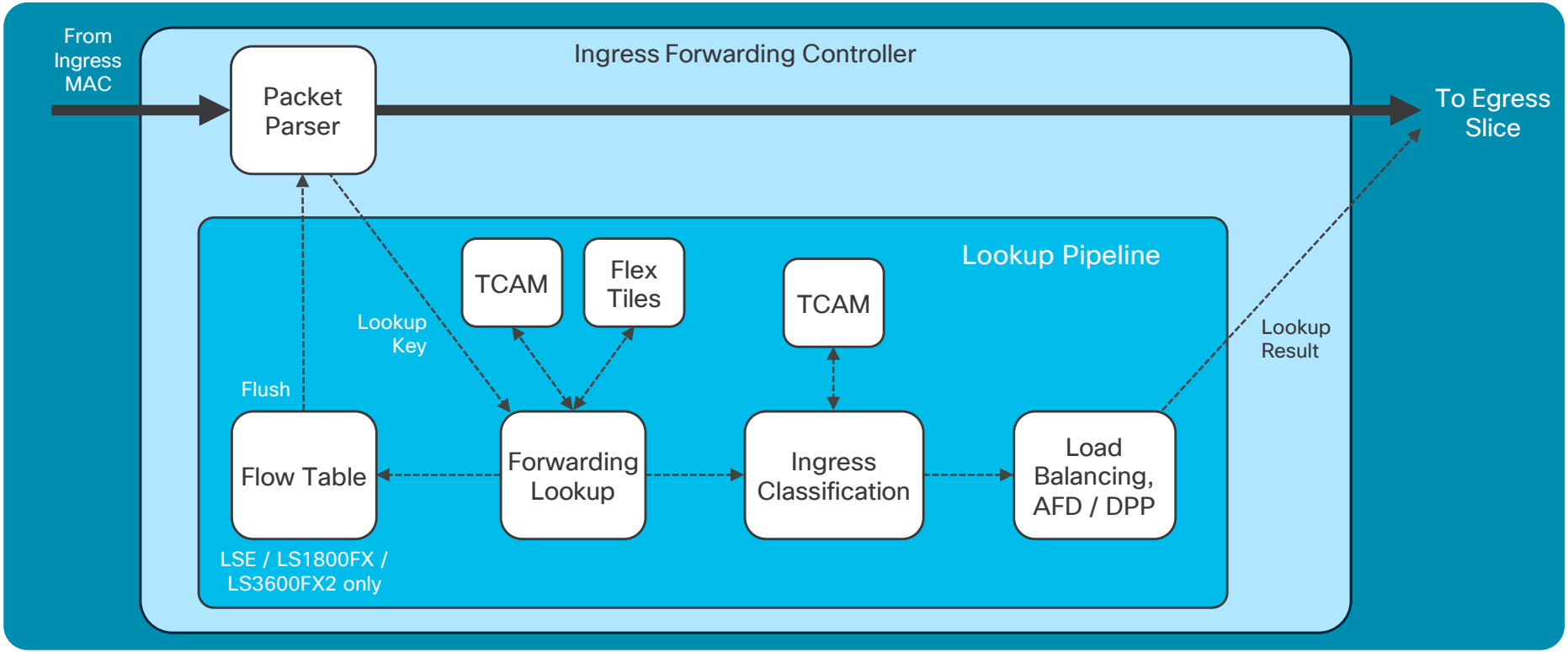
- Self-contained forwarding complex controlling subset of ports on single ASIC
- Separated into Ingress and Egress functions
- Ingress of each slice connected to egress of all slices
- Slice interconnect provides non-blocking any-to-any interconnection between slices



Slice Forwarding Path

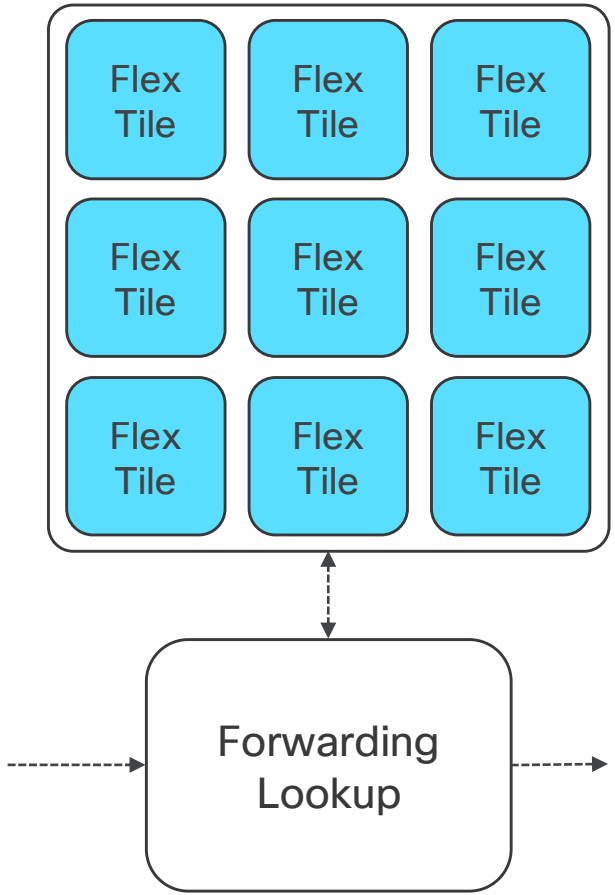


Ingress Lookup Pipeline



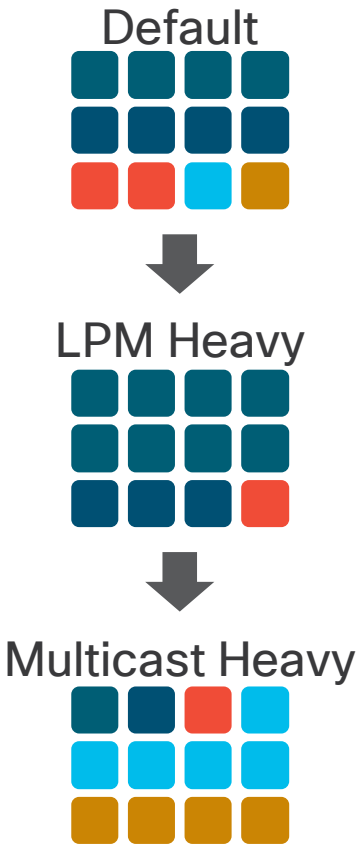
Flexible Forwarding Tiles

- Provide fungible pool of table entries for lookups
- Number of tiles and number of entries in each tile varies between ASICs
- Variety of functions, including:
  - IPv4/IPv6 unicast longest-prefix match (LPM)
  - IPv4/IPv6 unicast host-route table (HRT)
  - IPv4/IPv6 multicast (*,G) and (S,G)
  - MAC address/adjacency tables
  - ECMP tables
  - ACI policy



Flex Tile Routing Templates

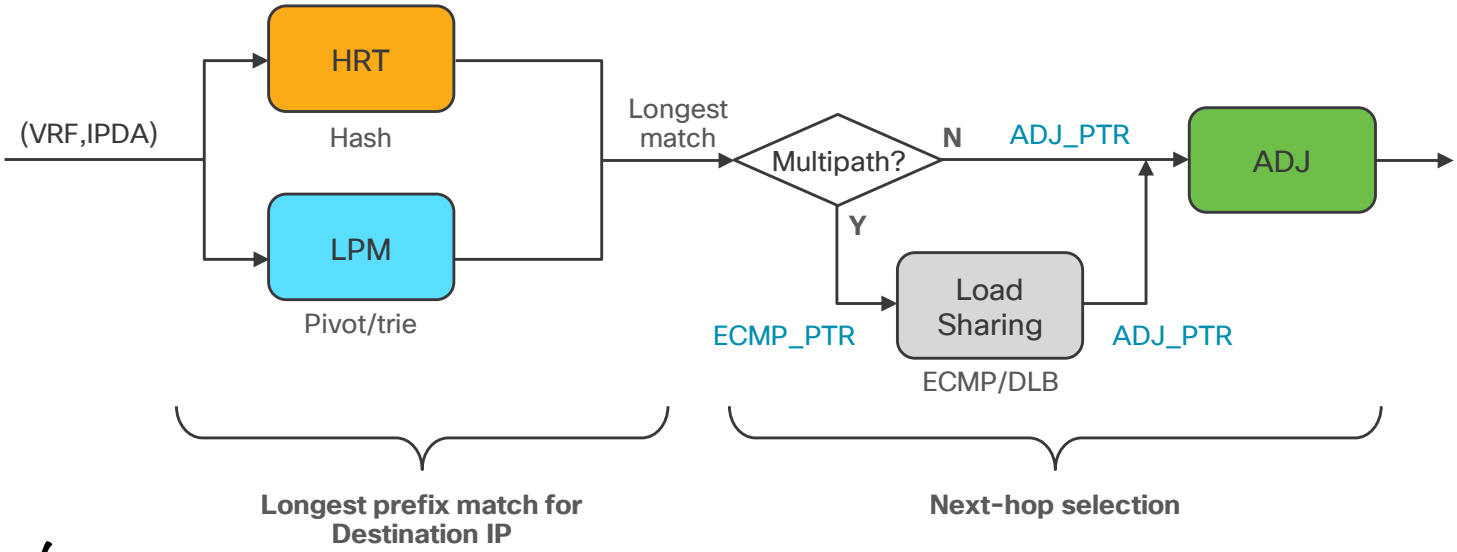
- Configurable forwarding templates determine flex tile functions
  - “system routing template” syntax
- Templates as of NX-OS 9.2(2):
  - Default
  - Dual-stack host scale*†
  - Internet peering*
  - LPM heavy
  - MPLS heavy*
  - Multicast heavy
  - Multicast extra-heavy
  - Service provider
- Defined at system initialization – reboot required to change profile



* Template does not support IP multicast
† Template not supported on modular Nexus 9500
** Template not supported on TORs

IP Unicast Forwarding

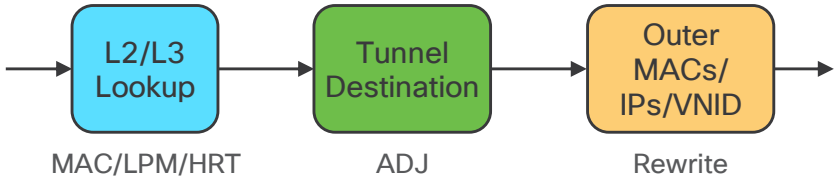
- Hardware lookup in flex tiles based on (VRF, IPDA)
- Longest-match from hash-based exact match (**HRT**) + pivot/trie match (**LPM**)
- Lookup result returns **adjacency** – directly or via **load-sharing** decision (ECMP/DLB)



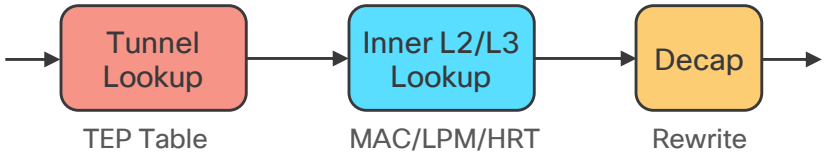
VXLAN Forwarding

- VXLAN and other tunnel encapsulation/ decapsulation performed in single pass
- Encapsulation
  - L2/L3 lookup drives tunnel destination
  - Rewrite block drives outer header fields (tunnel MACs/IPs/VNID, etc.)
- Decapsulation
  - Outer lookup determines if tunnel is transit or terminated on local TEP
  - Inner lookup determines final output port and rewrites

Encapsulation



Decapsulation



Load Sharing

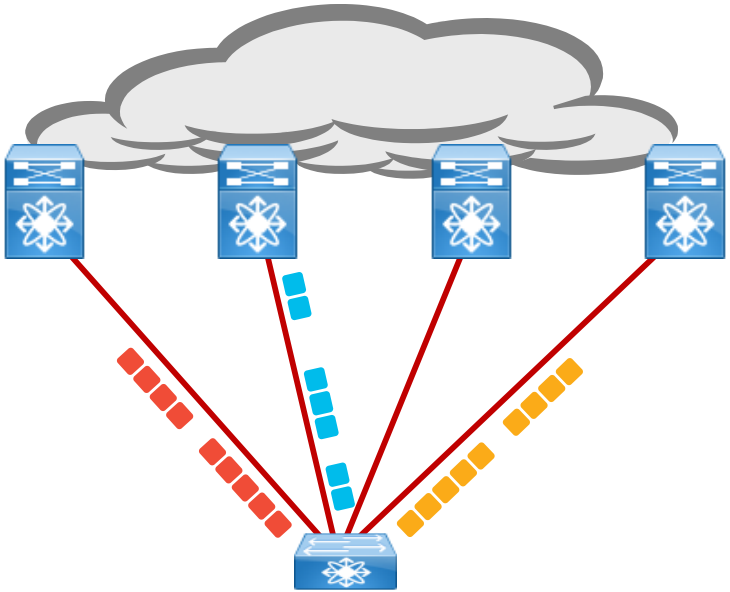
Equal-Cost Multipath (ECMP)

- Static flow-based load-sharing
- Picks ECMP next-hop based on hash of packet fields and universal ID
  - Source / destination IPv4 / IPv6 address (L3)
  - Source / destination TCP / UDP ports (L4)
  - L3 + L4 (default)
  - GRE key field

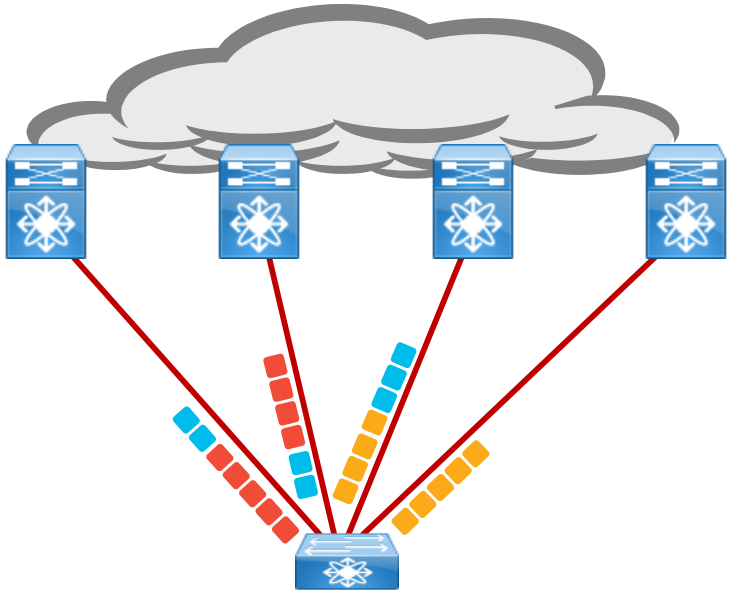
Dynamic Load-Balancing (DLB)

- Supported on leaf switches in ACI fabric
- Congestion aware, flow-based or flowlet-based – rebalances flows/flowlets based on path congestion

ECMP versus DLB Load-Sharing



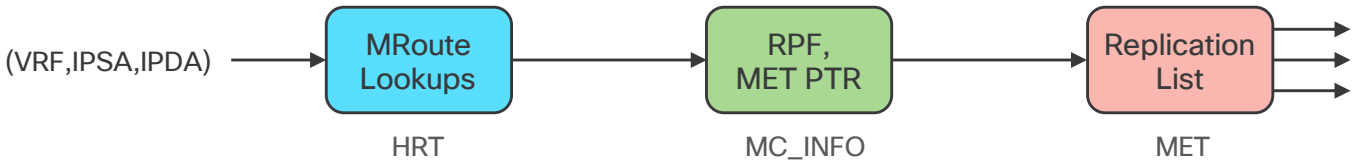
ECMP



DLB

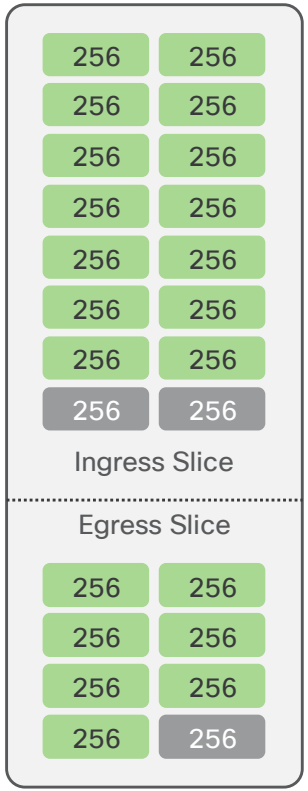
Multicast Forwarding

- Hardware performs multicast lookups in **HRT**
- Additional, secondary table for multicast also provisioned (“MC_INFO”) from flex tiles – **RPF check** and **MET pointer**
- **MET** in egress slice holds local output interface list (OIL)
- Replication is single copy, multiple reads

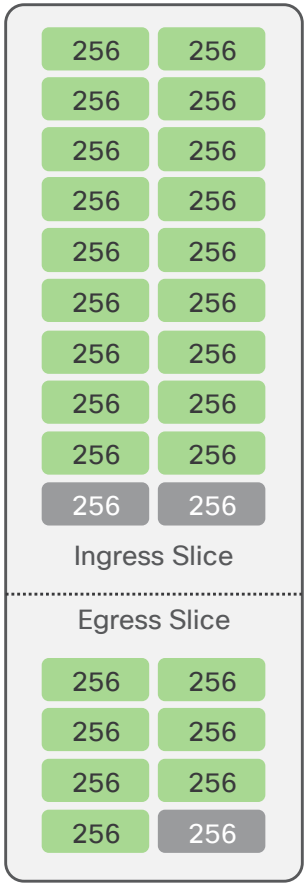


Classification TCAM

- Dedicated TCAM for packet classification
- Capacity varies depending on platform
- Leveraged by variety of features:
  - RACL / VACL / PACL
  - L2/L3 QOS
  - SPAN / SPAN ACL
  - NAT
  - COPP
  - Flow table filter (LS1800FX / LS3600FX2)



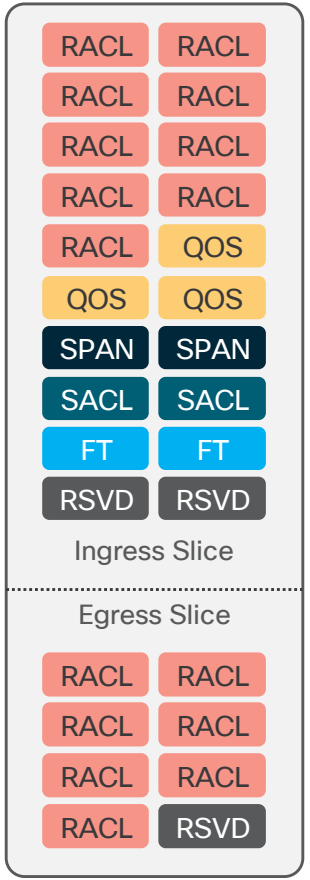
LSE / S6400
4K ingress ACEs /
2K egress ACEs per slice



LS1800FX / LS3600FX2
5K ingress ACEs /
2K egress ACEs per slice

TCAM Region Resizing

- Default carving allocates 100% of TCAM and enables:
  - Ingress / Egress RACL
  - Ingress QOS
  - SPAN
  - SPAN ACLs
  - Flow table filter (LS1800FX / LS3600FX2 only)
  - Reserved regions
- Based on features required, user can resize TCAM regions to adjust scale
  - To increase size of a region, some other region must be sized smaller
- Region sizes defined at initialization – changing allocation requires system reboot
  - Configure all regions to desired size ("hardware access-list tcam region"), save configuration, and reload



Cloud Scale Hardware Telemetry

Flow Table (FT)

- Captures full data-plane packet flow information, plus metadata

Flow Table Events (FTE)

- Triggers notifications based on thresholds / criteria met by data-plane packet flows

Streaming Statistics Export (SSX)

- Streams ASIC statistics based on user configuration

Data-Plane Flow Data

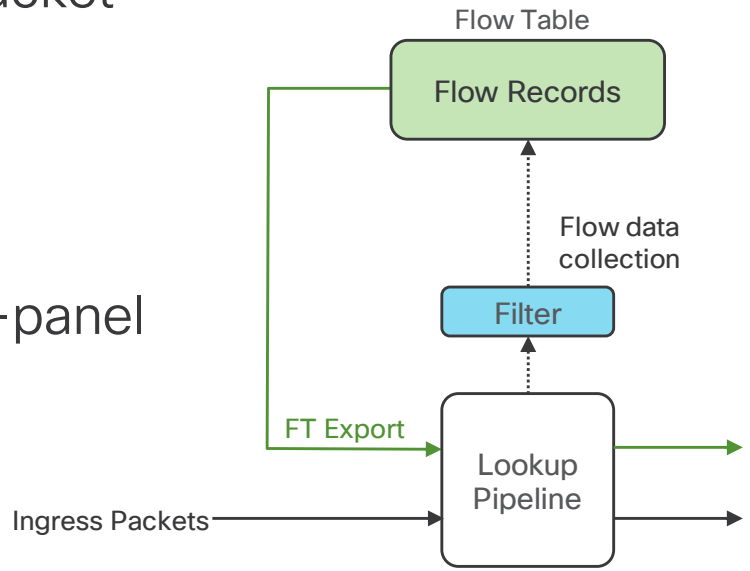
ASIC State

Flow Table

- Collects full flow information plus metadata
  - 5-tuple flow info
  - Interface/queue info
  - Flow start/stop time
  - Flow latency
- 32K flow table entries per slice
- Direct hardware export
- Leveraged by Network Insights, Netflow, Tetration
- LSE / LS1800FX / LS3600FX2 platforms support hardware flow table

Flow Table Operation – Network Insights

- Determine if collection enabled for packet (filter TCAM)
- If so, install FT record
- Flush records, encapsulate in IP/UDP
- Perform lookup and forward on front-panel port

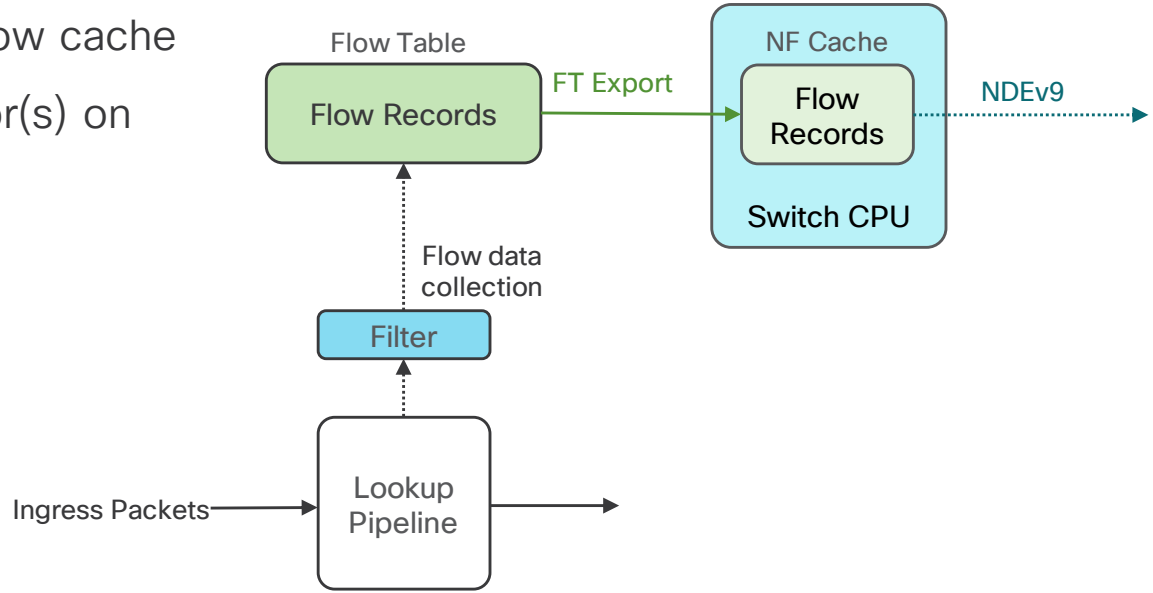


Flow Table Operation – Netflow

1. Install FT records as usual
2. Flush records to switch CPU
3. CPU builds traditional Netflow cache
4. NDEv9 exported to collector(s) on front-panel port or mgmt0

Netflow v9 support:

- 9300-FX TORs: 7.0(3)I7(1)
- 9300-EX TORs: 7.0(3)I7(2)
- 9300-FX2 TORs: 7.0(3)I7(3)
- 9500-EX Modular: 9.2(2)

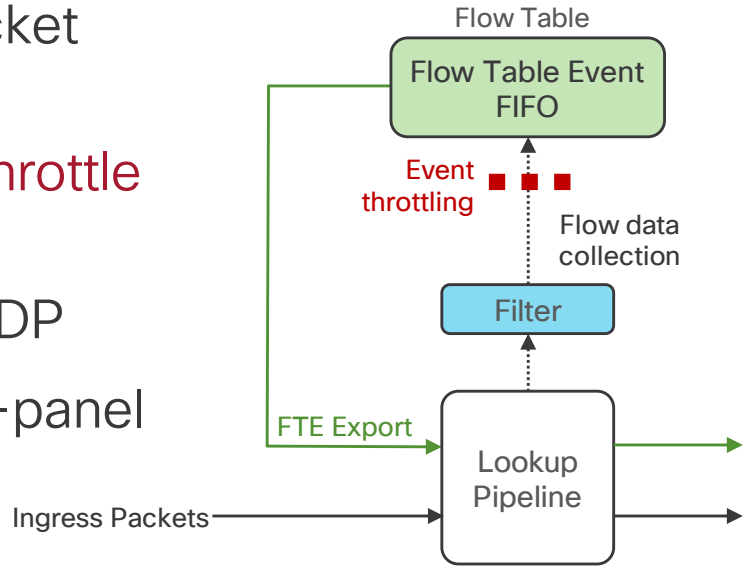


Flow Table Events

- Triggers notifications based on criteria / thresholds met by data-plane packet flows
- Collects full flow information plus metadata
  - 5-tuple flow info with timestamp
  - Interface/queue info
  - Buffer drop indication
  - Forwarding drop, ACL drop, policer drop indication
  - Latency/burst threshold exceeded indication
- Direct hardware export, with flow-level and global throttling
- LS1800FX / LS3600FX2 platforms support triggered flow table events

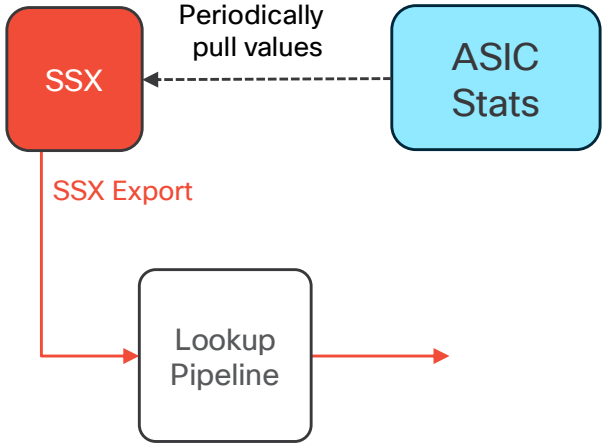
Flow Table Events Operation

- Determine if event(s) enabled for packet (filter TCAM)
- If so, collect flow data in FTE FIFO; Throttle excess events
- Flush and encapsulate record in IP/UDP
- Perform lookup and forward on front-panel port



Streaming Statistics Export (SSX)

- Streams ASIC statistics at rapid cadence based on user config
  - Interface counters (RMON counters)
  - Ingress/egress queue depth and queue drops
  - Egress total buffer depth
  - Egress queue microbursts
- User defines streaming parameters – which statistics, how often, and to which collector
- Direct export from ASIC to front-panel port – no switch CPU involvement
- Hardware support in S6400 / LS3600FX2

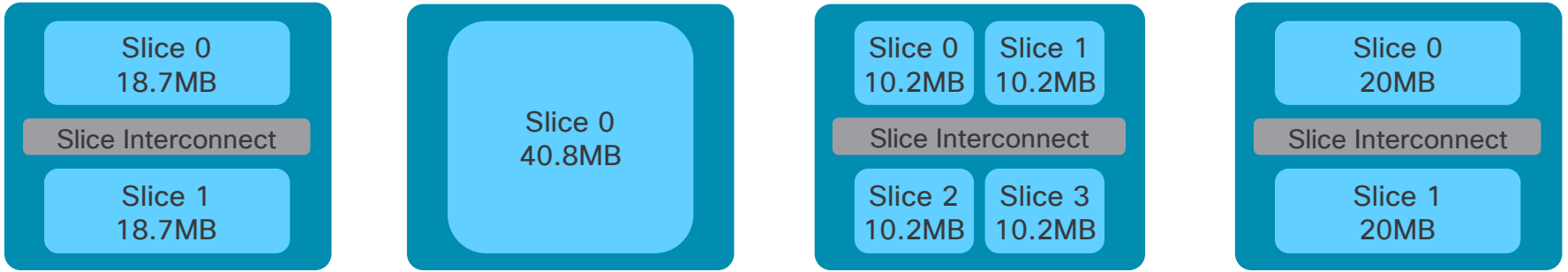


Hardware Telemetry Platform Support

Platform	ASIC	FT	FTE	SSX
9300-EX	LSE	✓	X	X
9300-FX	LS1800FX	✓	✓	X
9364C	S6400	X	X	✓
9300-FX2	LS3600FX2	✓	✓	✓

Buffering

- Cloud Scale platforms implement shared-memory egress buffered architecture
- Each ASIC slice has dedicated buffer – only ports on that slice can use that buffer
- Dynamic Buffer Protection adjusts max thresholds based on class and buffer occupancy
- Intelligent buffer options maximize buffer efficiency



LSE
18.7MB/slice
(37.4MB total)

LS1800FX
40.8MB/slice
(40.8MB total)

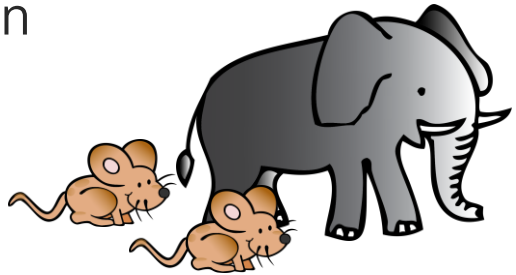
S6400
10.2MB/slice
(40.8MB total)

LS3600FX2
20MB/slice
(40MB total)
Shared option (10MB/slice + 20MB shared)

Intelligent Buffering

Innovative Buffer Management for Cloud Scale switches

- **Dynamic Buffer Protection** (DBP) – Controls buffer allocation for congested queues in shared-memory architecture
- **Approximate Fair Drop** (AFD) – Maintains buffer headroom per queue to maximize burst absorption
- **Dynamic Packet Prioritization** (DPP) – Prioritizes short-lived flows to expedite flow setup and completion



Images courtesy of:
https://www.clker.com/clipart-206333.html
https://www.clker.com/clipart-catroon-mouse.html

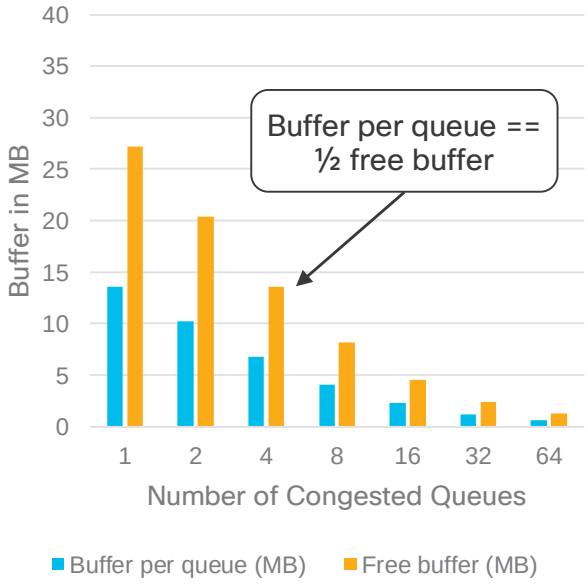
Miercom Report: Speeding Applications in Data Center Networks
http://miercom.com/cisco-systems-speeding-applications-in-data-center-networks/

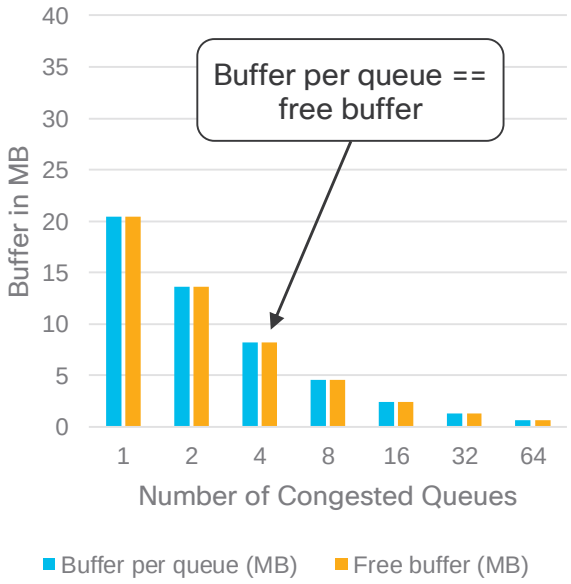
Dynamic Buffer Protection (DBP)

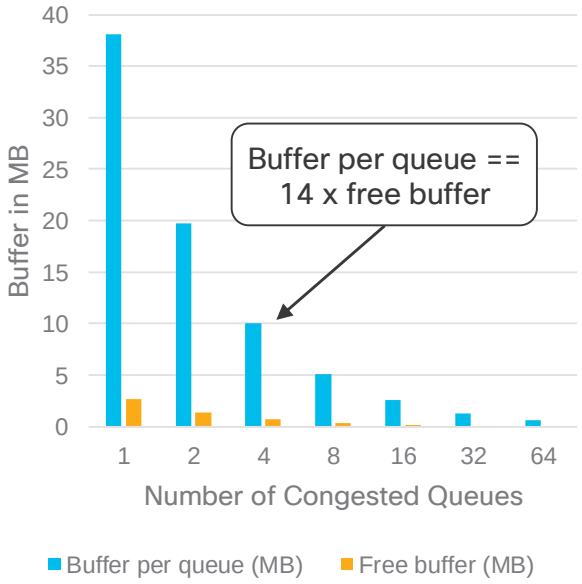
- Prevents any output queue from consuming more than its fair share of buffer in shared-memory architecture
- Defines dynamic max threshold for each queue
  - If queue length exceeds threshold, packet is discarded
  - Otherwise packet is admitted to queue and scheduled for transmission
- Threshold calculated by multiplying free memory by configurable, per-queue **Alpha** (α) value (weight)
  - Alpha controls how aggressively DBP maintains free buffer pages during congestion events

Alpha Parameter Examples

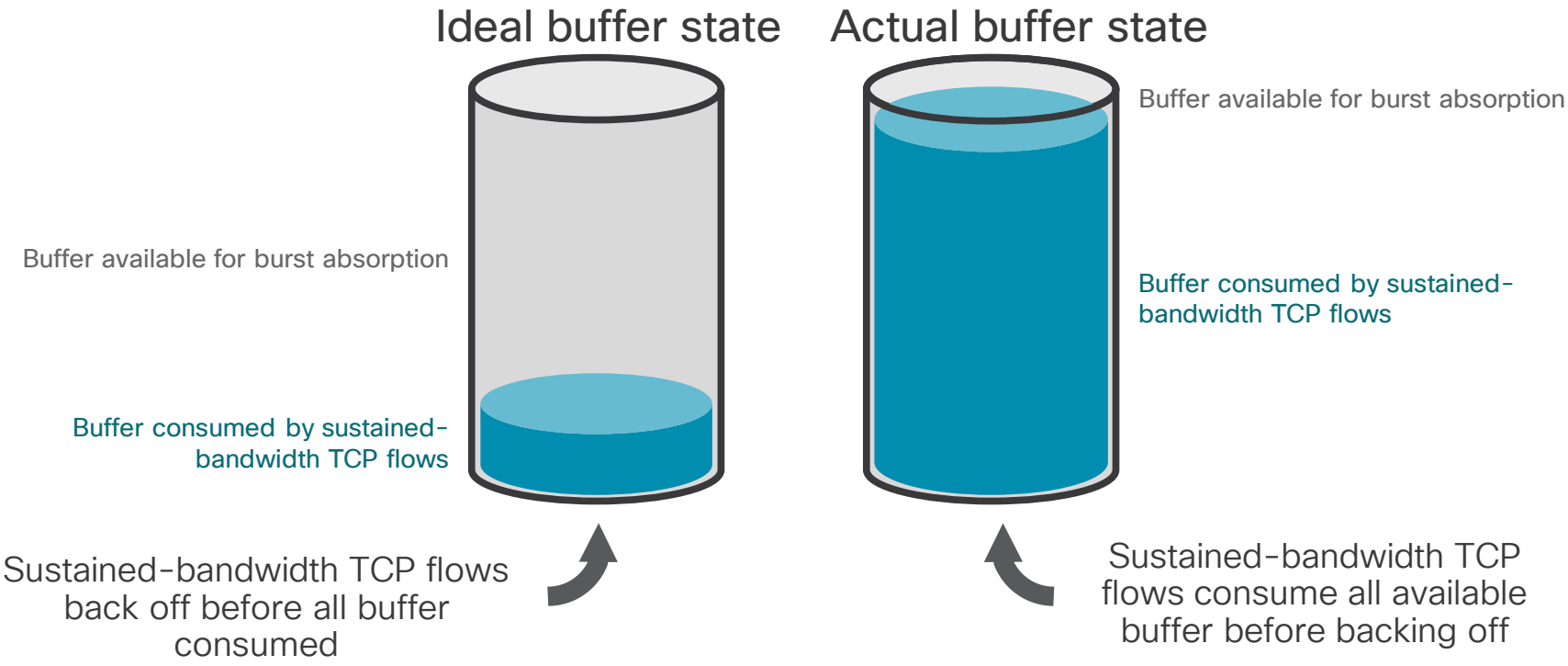
Default Alpha on Cloud Scale switches





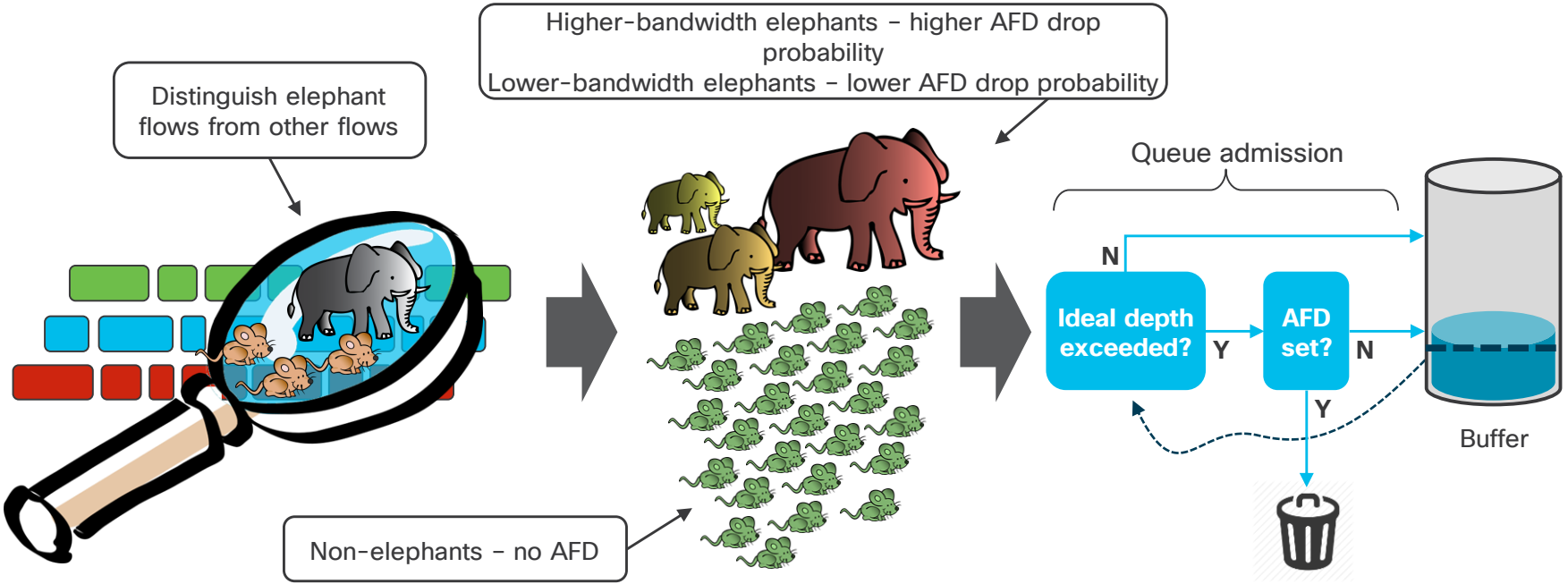


Buffering – Ideal versus Reality



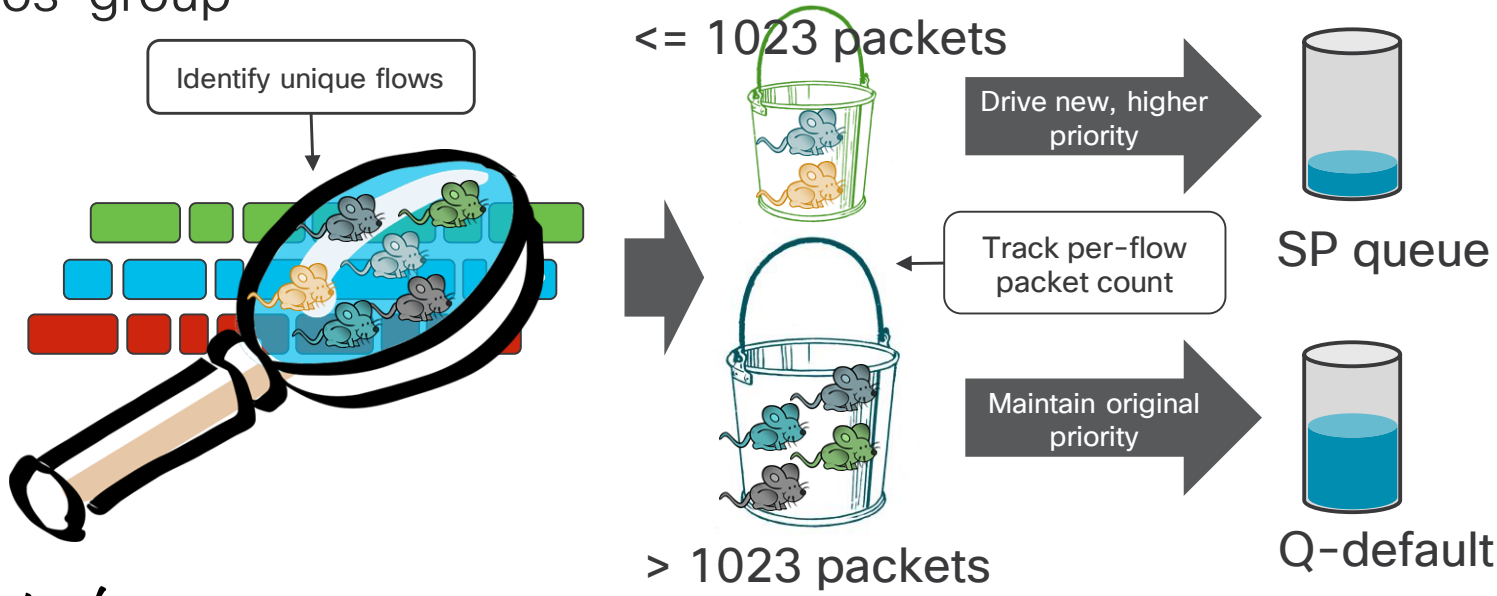
Approximate Fair Drop (AFD)

Maintain throughput while minimizing buffer consumption by elephant flows – **keep buffer state as close to the ideal as possible**

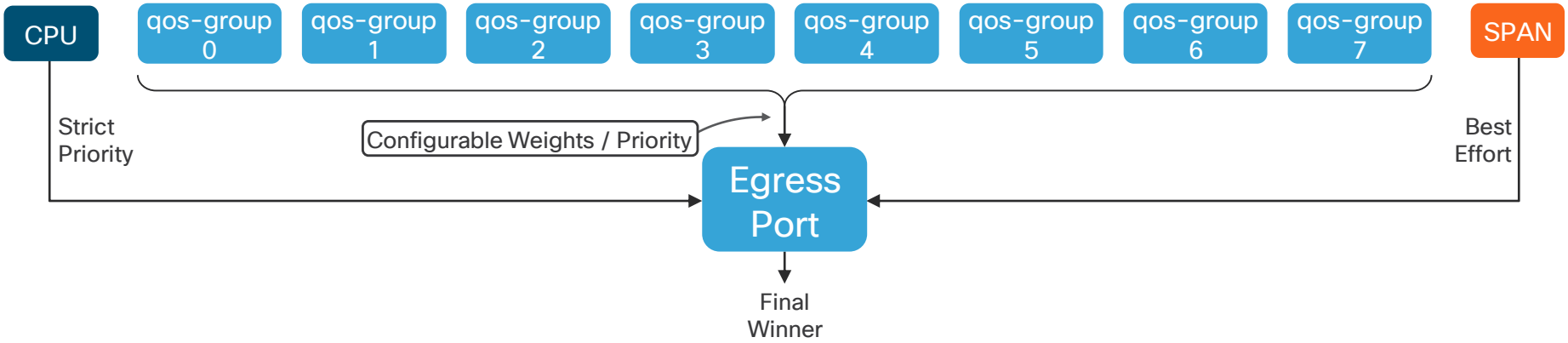


Dynamic Packet Prioritization (DPP)

- Prioritize initial packets of new / short-lived flows
- Up to first 1023 packets of each flow assigned to higher-priority qos-group



Queuing and Scheduling

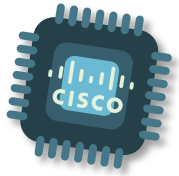


- 8 qos-groups per output port
- Egress queuing policy defines priority and weights
- Dedicated classes for CPU traffic and SPAN traffic

Agenda

- Data Center and Silicon Strategy
- Cloud Scale ASIC Architecture
- Cloud Scale Switching Platforms
- Packet Walks
- Key Takeaways

Cloud Scale Platforms



Nexus 9300-EX and 9300-FX/FX2

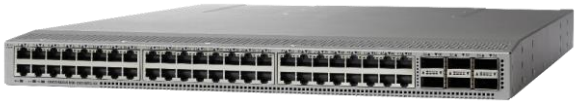
- Premier TOR platforms
- Full Cloud Scale functionality
- ACI leaf / standalone leaf or spine
- FX option with MACSEC using LS1800FX silicon
- FX2 option with key enhancements using LS3600FX2 silicon

Nexus 9500 with X9700-EX and X9700-FX Modules

- Switching modules for Nexus 9500 modular chassis
- Full Cloud Scale functionality
- ACI spine / standalone aggregation or spine
- FX option with MACSEC using LS1800FX silicon

Nexus 9300-EX Cloud Scale TOR Switches

48-port 10/25G SFP28 + 6-port 100G QSFP28
N9K-C93180YC-EX – LSE-based
ACI: 1.3(1)
NX-OS: 7.0(3)I4(2)



48-port 1/10GBASE-T + 6-port 100G QSFP28
N9K-C93108TC-EX – LSE-based
ACI: 2.0(1)
NX-OS: 7.0(3)I4(2)

32-port 40G/100G QSFP28
N9K-C93180LC-EX – LSE-based
ACI: 2.2(1)
NX-OS: 7.0(3)I6(1)

Key Features

Dual capability – ACI and NX-OS mode

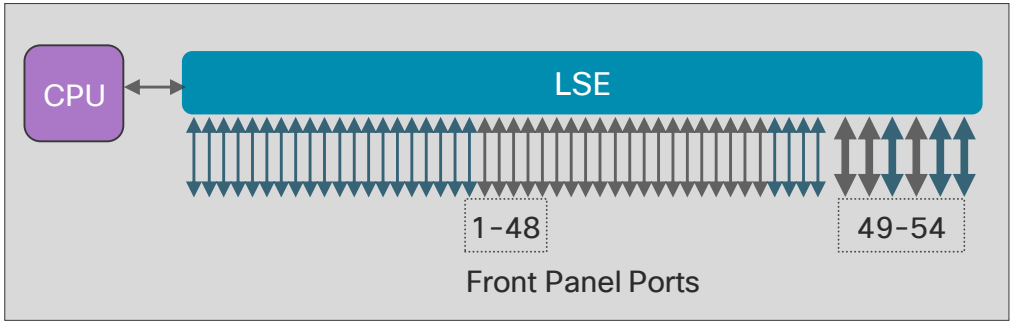
Flexible port configurations – 1/10/25/40/50/100G

Native 25G server access ports

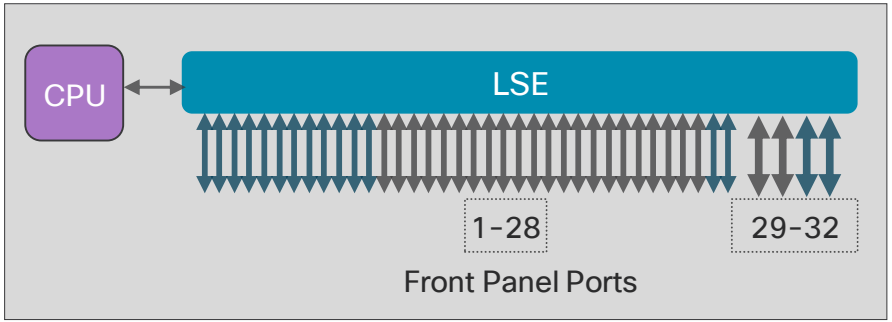
Flow Table for Tetration Analytics, Netflow

Smart buffer capability (AFD / DPP)

Nexus 9300-EX Switch Architectures



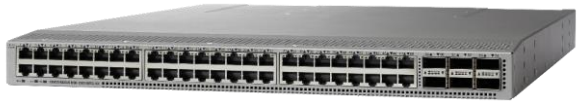
C93180YC-EX (10/25G + 100G) /
C93108TC-EX (10G + 100G)



C93180LC-EX (40G + 100G)

Nexus 9300-FX Cloud Scale TOR Switches – Pervasive MACSEC

48-port 10/25G SFP28 + 6-port 100G QSFP28
N9K-C93180YC-FX – LS1800FX-based
ACI: 2.2(2e)
NX-OS: 7.0(3)I7(1)



48-port 1/10GBASE-T + 6-port 100G QSFP28
N9K-C93108TC-FX – LS1800FX-based
ACI: 2.2(2e)
NX-OS: 7.0(3)I7(1)

48-port 100M/1GBASE-T + 4-port 10G/25G + 2-port 100G QSFP28
N9K-C9348GC-FXP – LS1800FX-based
ACI: 3.0(1)
NX-OS: 7.0(3)I7(1)

Key Features

Dual capability – ACI and NX-OS mode

Flexible port configurations – 100M/1/10/25/40/50/100G

Line-rate 256-bit encryption on all ports

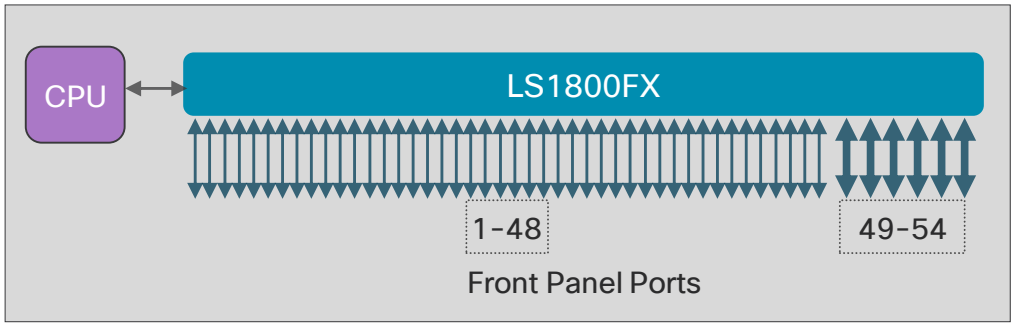
32G FC support on all SFP ports

25G distances beyond 3m (RS-FEC)

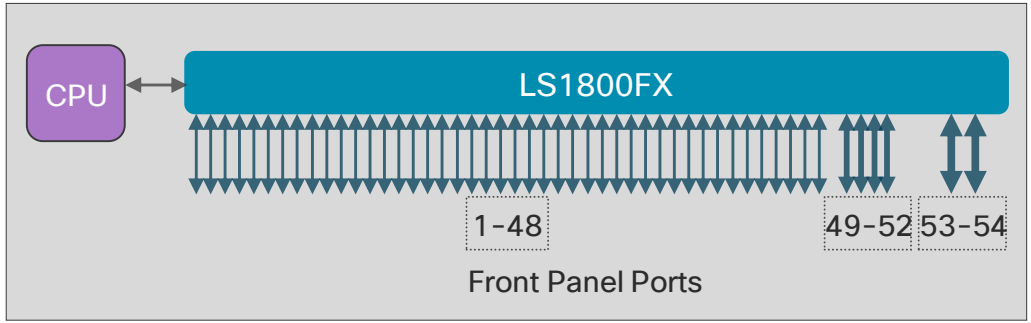
Flow Table for Tetration Analytics, **Network Insights**, Netflow

Smart buffer capability (AFD / DPP)

Nexus 9300-FX Switch Architectures



C93180YC-FX (10/25G + 100G) /
C93108TC-FX (10G + 100G)



C9348GC-FXP (100M/1G + 10/25G + 100G)

Slice 0

Nexus 9364C / 9332C 100G Cloud Scale Switches

64-port 100G QSFP28
+ 2-port 10G SFP+
N9K-C9364C – S6400-based
ACI: 3.0(1)
NX-OS: 7.0(3)I7(2)

32-port 100G QSFP28
+ 2-port 10G SFP+
N9K-C9332C – S6400-based
ACI: 4.0(1)
NX-OS: 9.2(3)

Key Features

Dual capability – ACI and NX-OS mode

Compact, high-performance fixed ACI spine

100G/50G/40G/10G (single port mode – no breakout)

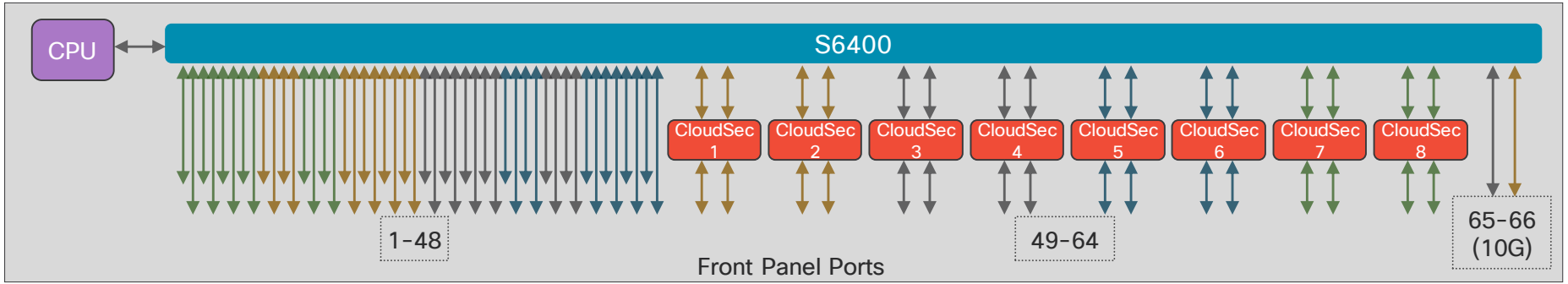
2 x 100M/1G/10G SFP+ ports

MACSEC/CloudSec on 16 / 8 ports

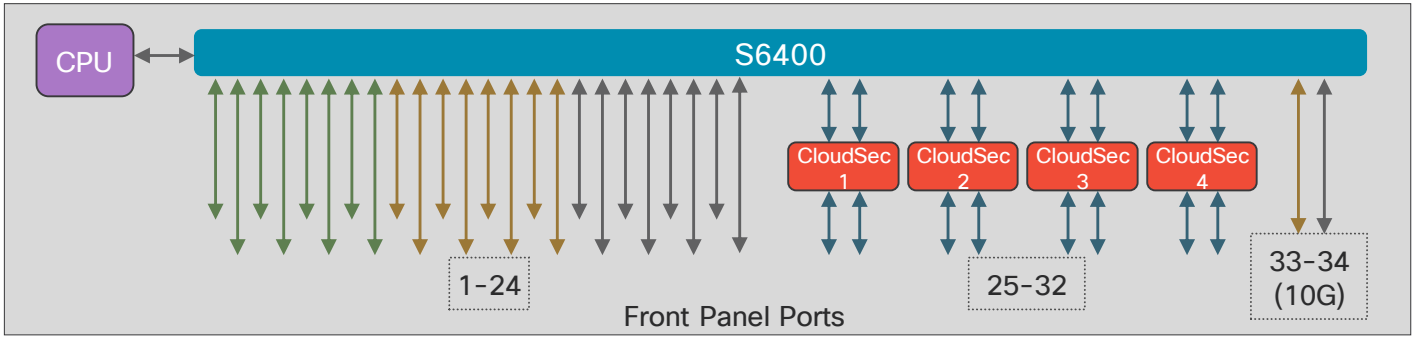
Streaming Statistics Export (SSX)

Smart buffer capability (AFD / DPP)

Nexus 9364C / 9332C Switch Architecture



C9364C (100G + 10G)



C9332C (100G + 10G)

Nexus 9300-FX2 Cloud Scale TOR Switches

36-port 100G QSFP28
N9K-C9336C-FX2 – LS3600FX2-based
ACI: 3.1(2)
NX-OS: 7.0(3)I7(3)

48-port 10/25G SFP28 +
12-port 100G QSFP28
N9K-C93240YC-FX2 – LS3600FX2-based
ACI: 4.0(1)
NX-OS: 7.0(3)I7(3)

Key Features

Dual capability – ACI and NX-OS mode

Versatile standalone 100G switch

High-performance 100G ACI leaf switch (9336C)

100G/50G/40G/10G with breakout capability

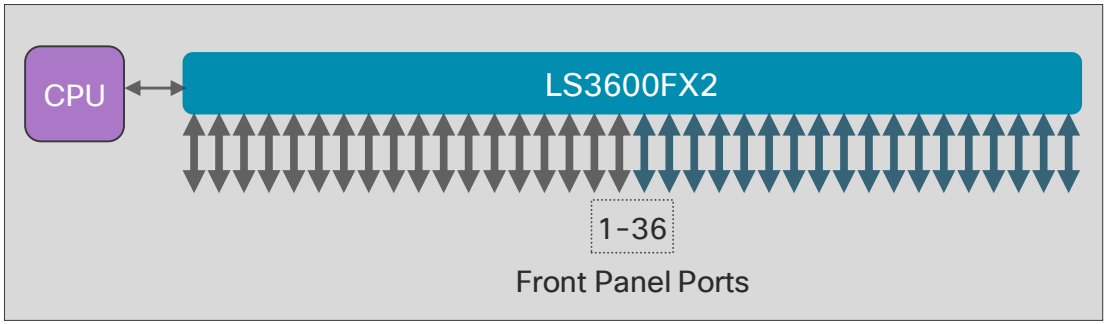
Flow Table for Tetration Analytics, Network Insights, Netflow

Streaming Statistics Export (SSX)

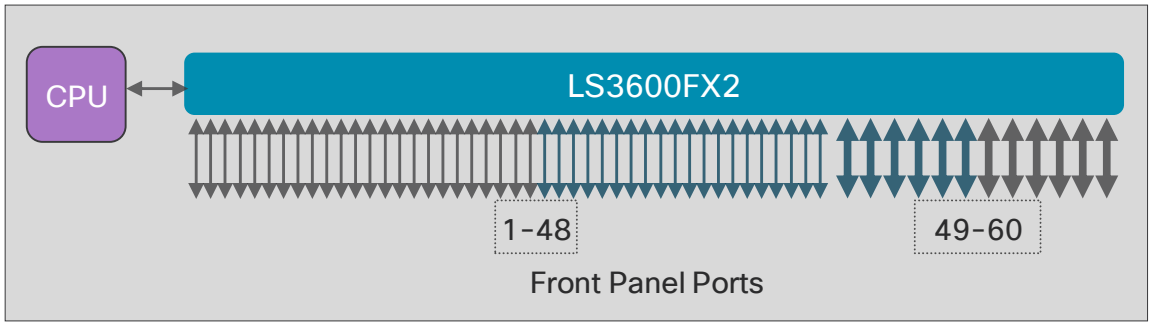
MACSEC/CloudSec on all ports

Smart buffer capability (AFD / DPP)

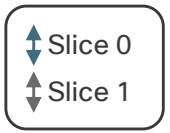
Nexus 9300-FX2 Switch Architecture



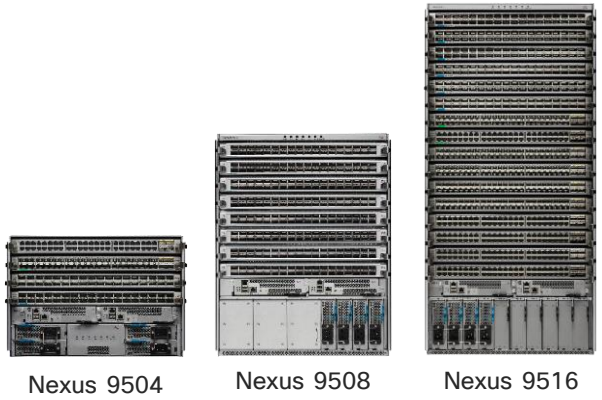
C9336C-FX2 (100G)



C93240YC-FX2 (10/25G + 100G)

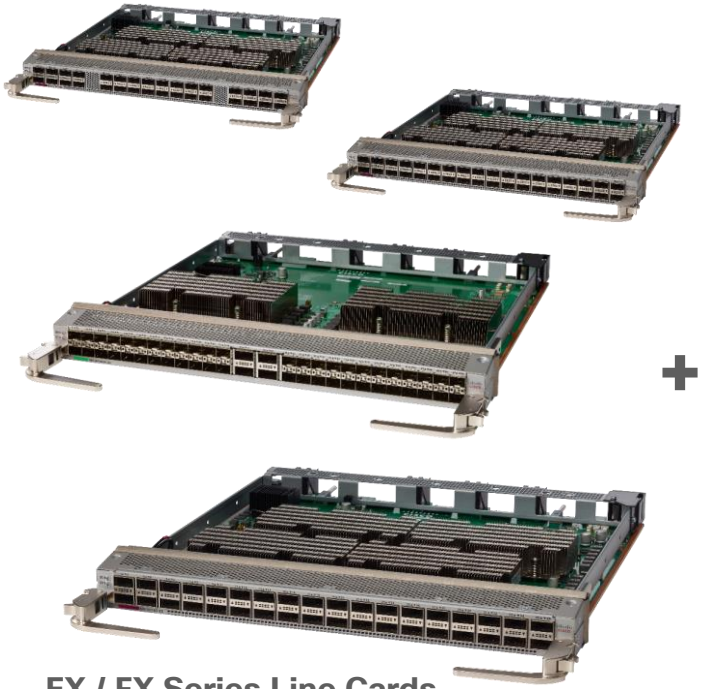


Nexus 9500 Modular Cloud Scale Switches

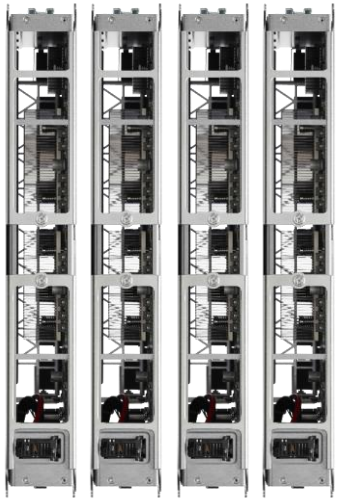


Nexus 9504 Nexus 9508 Nexus 9516

Common Equipment



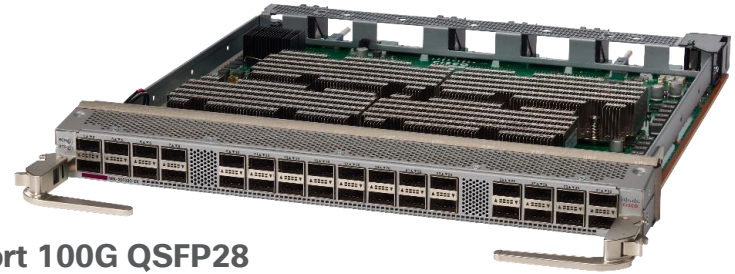
EX / FX Series Line Cards



E/E2-Series Fabric Modules

X9700-EX 100G Cloud Scale Modules

N9K-X9732C-EX / N9K-X9736C-EX



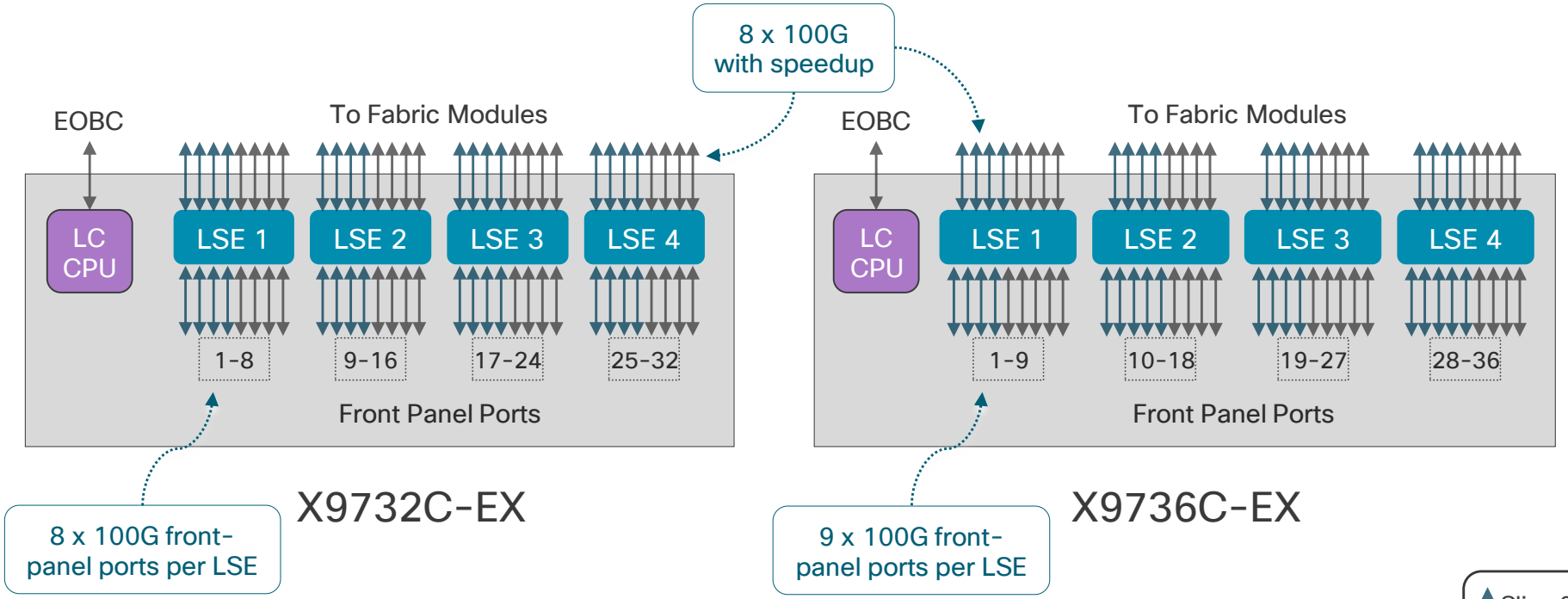
32-port 100G QSFP28
X9732C-EX – LSE-based
ACI: 1.3(1)
NX-OS: 7.0(3)I4(2)

36-port 100G QSFP28
X9736C-EX – LSE-based
ACI: Not supported
NX-OS: 7.0(3)I6(1)

Key Features

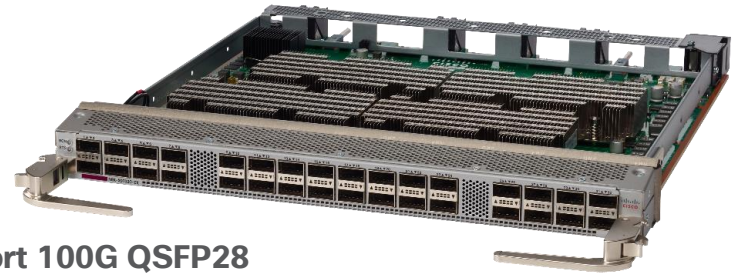
- 9732C-EX – Dual capability ACI and NX-OS
- 9736C-EX – NX-OS only
- Line-rate performance up to 3.2Tbps capacity
- Flexible port configurations – 10/25/40/50/100G with breakout
- Flow Table for Netflow, Tetration
- Smart buffer capability (AFD / DPP)

N9K-X9732C-EX / N9K-X9736C-EX Architecture



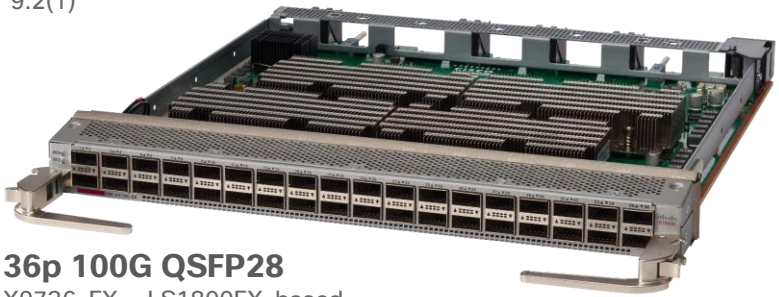
X9700-FX 100G Cloud Scale Modules

N9K-X9732C-FX / N9K-X9736C-FX



32-port 100G QSFP28

X9732C-FX – LS1800FX-based
ACI: Not supported
NX-OS: 9.2(1)



36p 100G QSFP28

X9736-FX – LS1800FX-based
ACI: 13.0(1)
NX-OS: NX-OS: 7.0(3)I7(3)

Key Features

9732C-FX – NX-OS only

9736C-FX – Dual capability ACI and NX-OS

3.2Tbps capacity line-rate performance at 170-byte frames

N+1 fabric redundancy on 9732C-FX

3.6Tbps capacity with optional 5th fabric module on 9736C-FX

Flexible port configurations – 10/25/40/50/100G with breakout

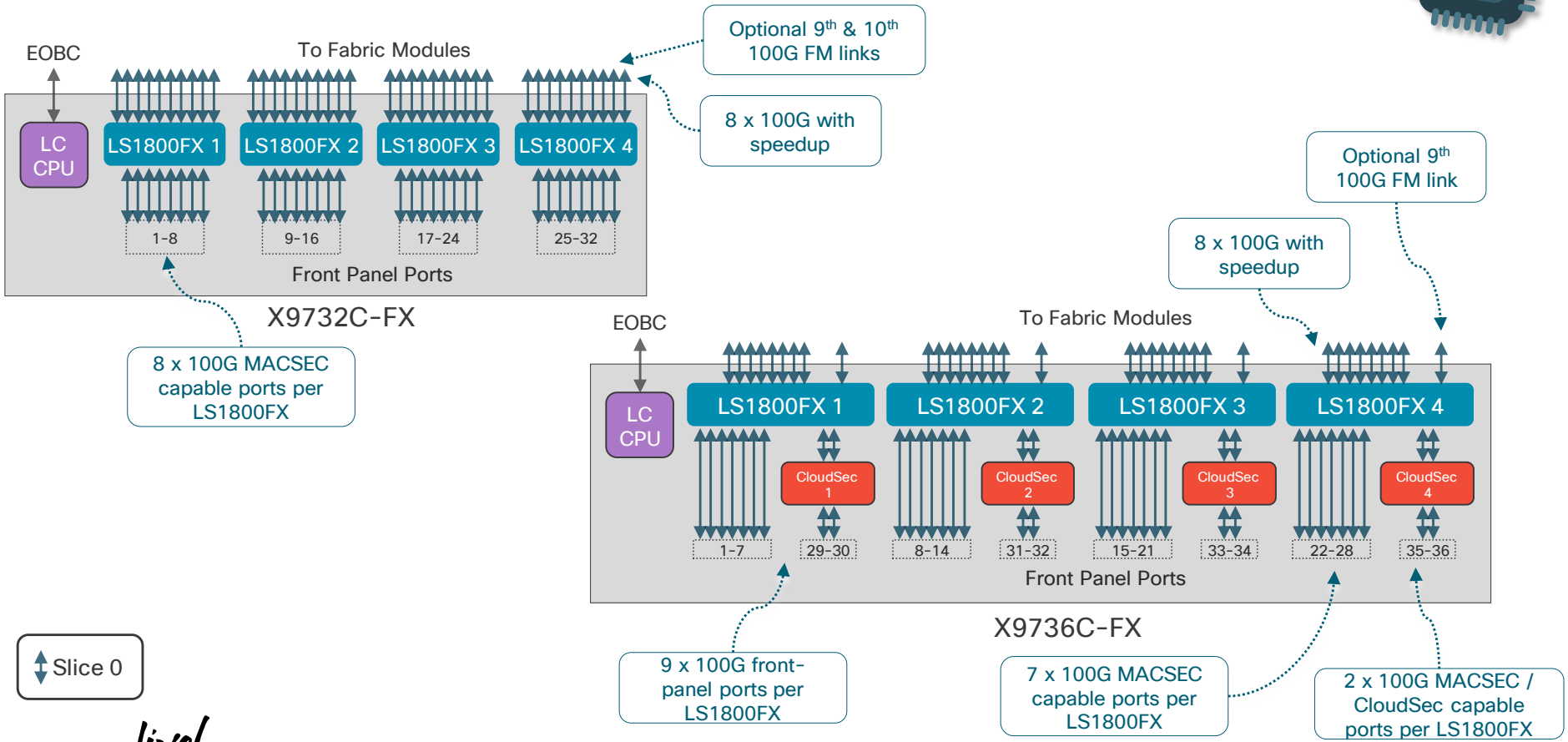
Line-rate MACSEC on all ports

CloudSec encryption (8 ports) on 9736C-FX

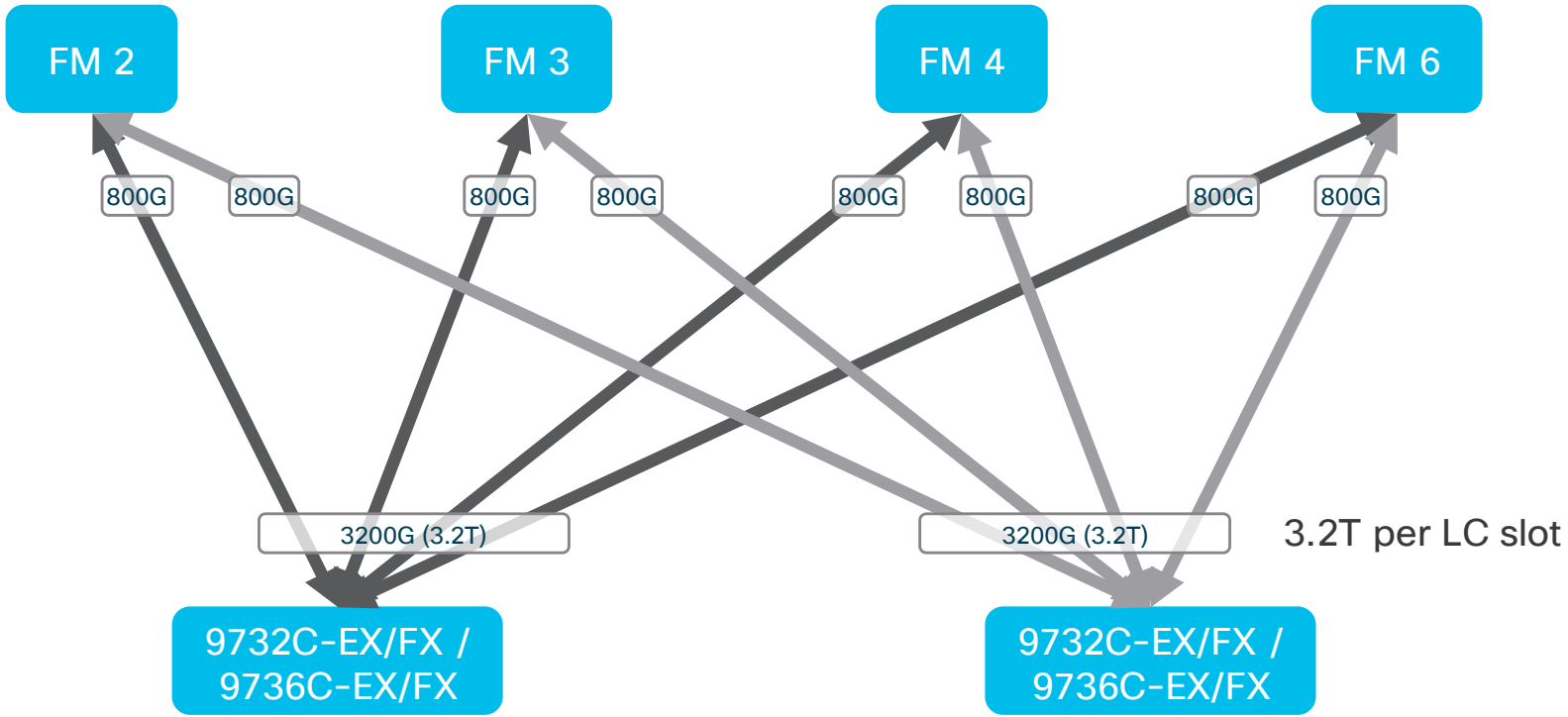
Flow Table for Network Insights, NetFlow (roadmap), Tetration

Smart buffer capability (AFD / DPP)

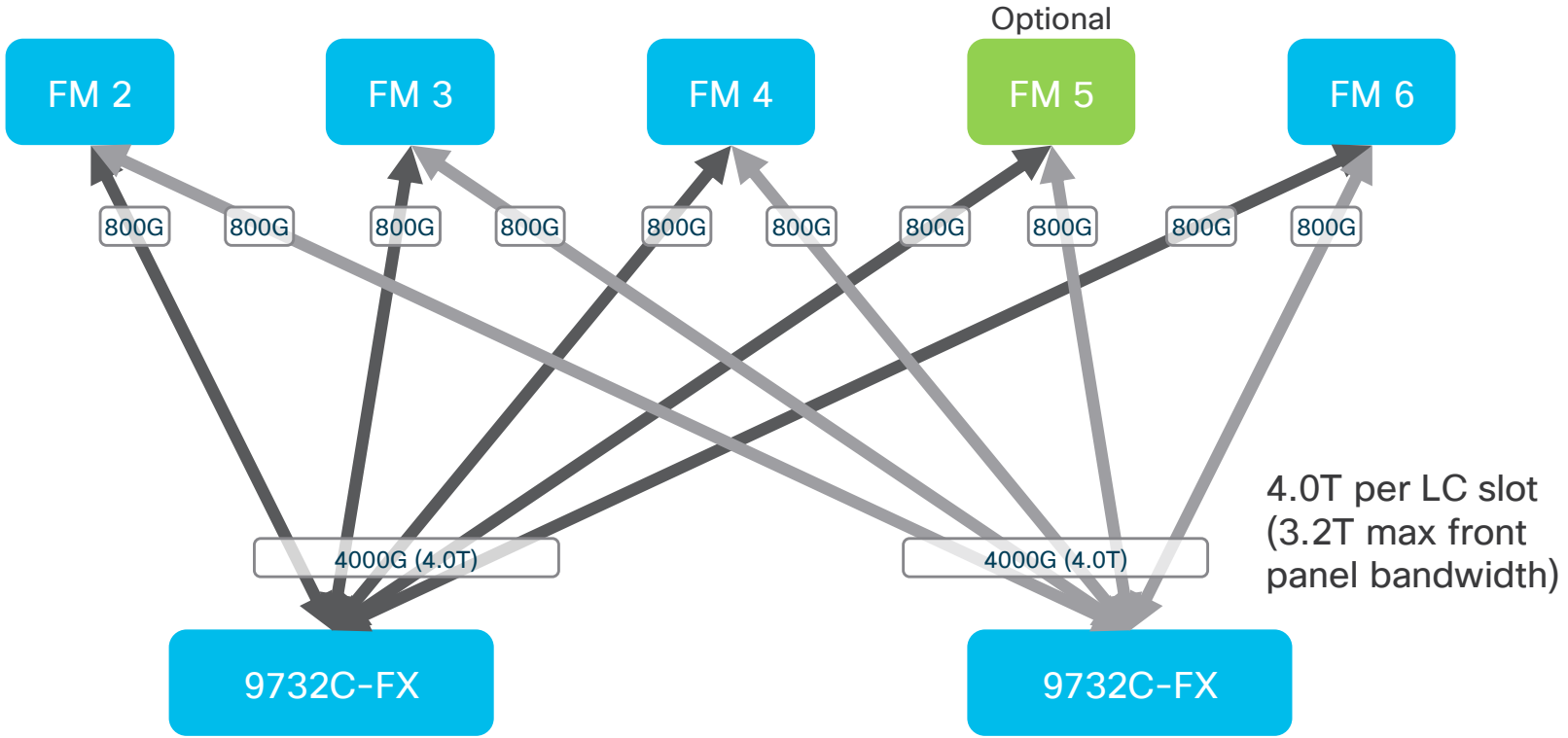
N9K-X9732C-FX / N9K-X9736C-FX Architecture



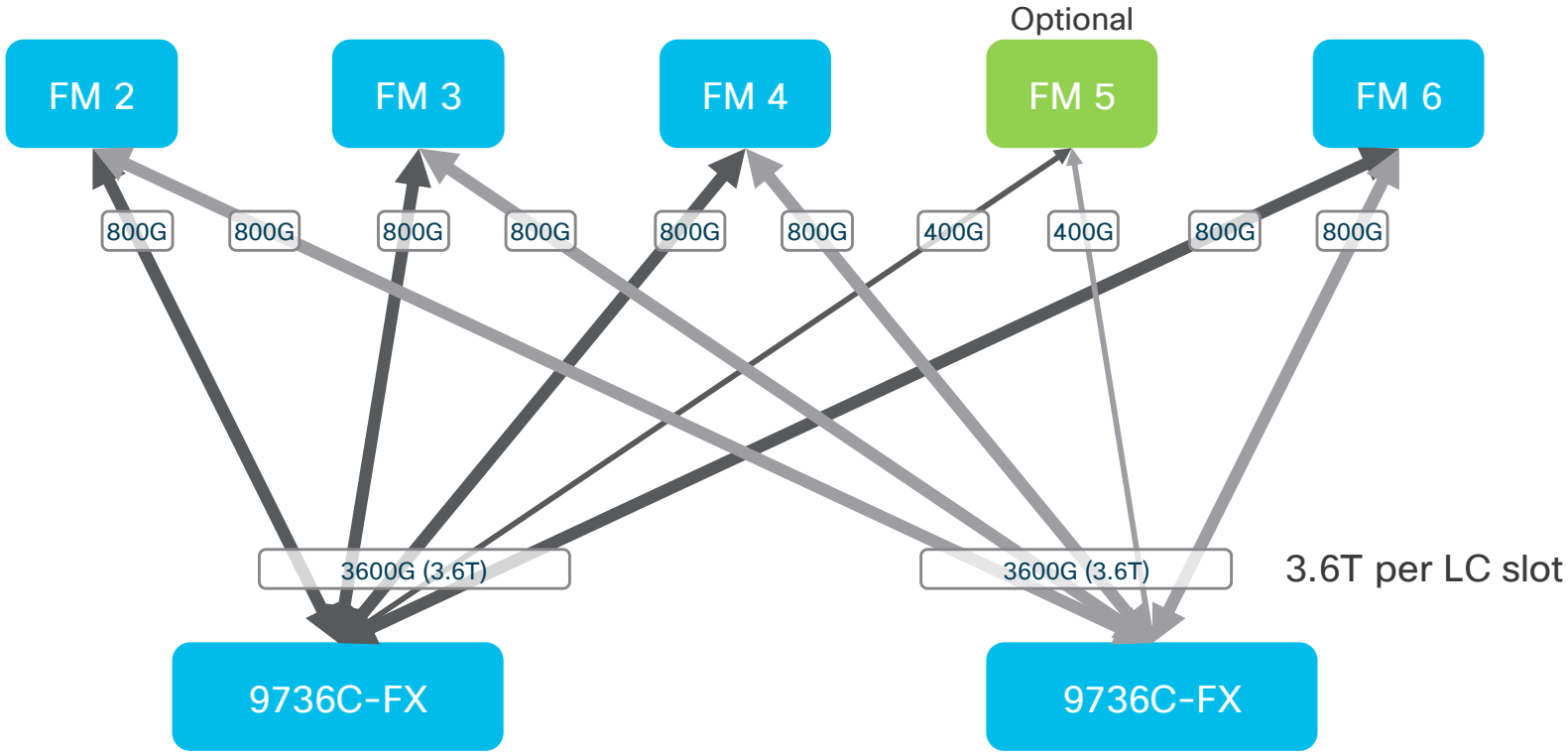
9732C-EX / 9736C-EX / 9732C-FX / 9736C-FX Fabric Connectivity



9732C-FX Fabric Connectivity – 5 FMs



9736C-FX Fabric Connectivity – 5 FMs



Fabric Redundancy – 5 FMs

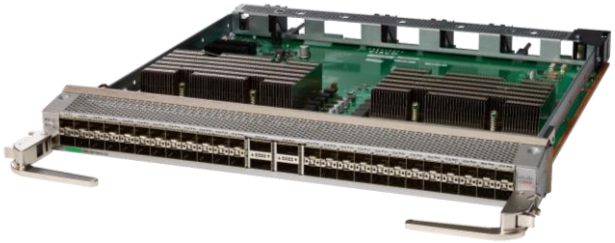
Limitations and Notes

- **All** modules installed in chassis must be either 9732C-FX or 9736C-FX to use 5 FMs
  - If other module type installed, 5th FM powered off automatically
- **9732C-FX:**
  - 5 FMs required for N+1 fabric module redundancy
- **9736C-FX:**
  - 5 FMs required for full bandwidth
  - Bandwidth reduction on FM failure varies depending on which FM failed

Note: 5 x FMs supported on all chassis types in standalone from 7.0(3)I7(2). 5 x FMs with 9736C-FX supported from in ACI 13.2(2).

X9700-EX/FX EOR/MOR Cloud Scale Modules

N9K-X97160YC-EX / N9K-X9788TC-FX



48p 10/25G SFP+ and 4p 100G QSFP28

X97160YC-EX – LSE-based
ACI: Not supported
NX-OS: 7.0(3)I5(2)

48p 1/10GBASE-T and 4p 100G QSFP28

X9788TC-FX – LS1800FX-based
ACI: Not supported
NX-OS: 7.0(3)I7(3)

Key Features

NX-OS mode only

Flow Table for Tetration Analytics, NetFlow

Smart buffer capability (AFD / DPP)

97160-EX:

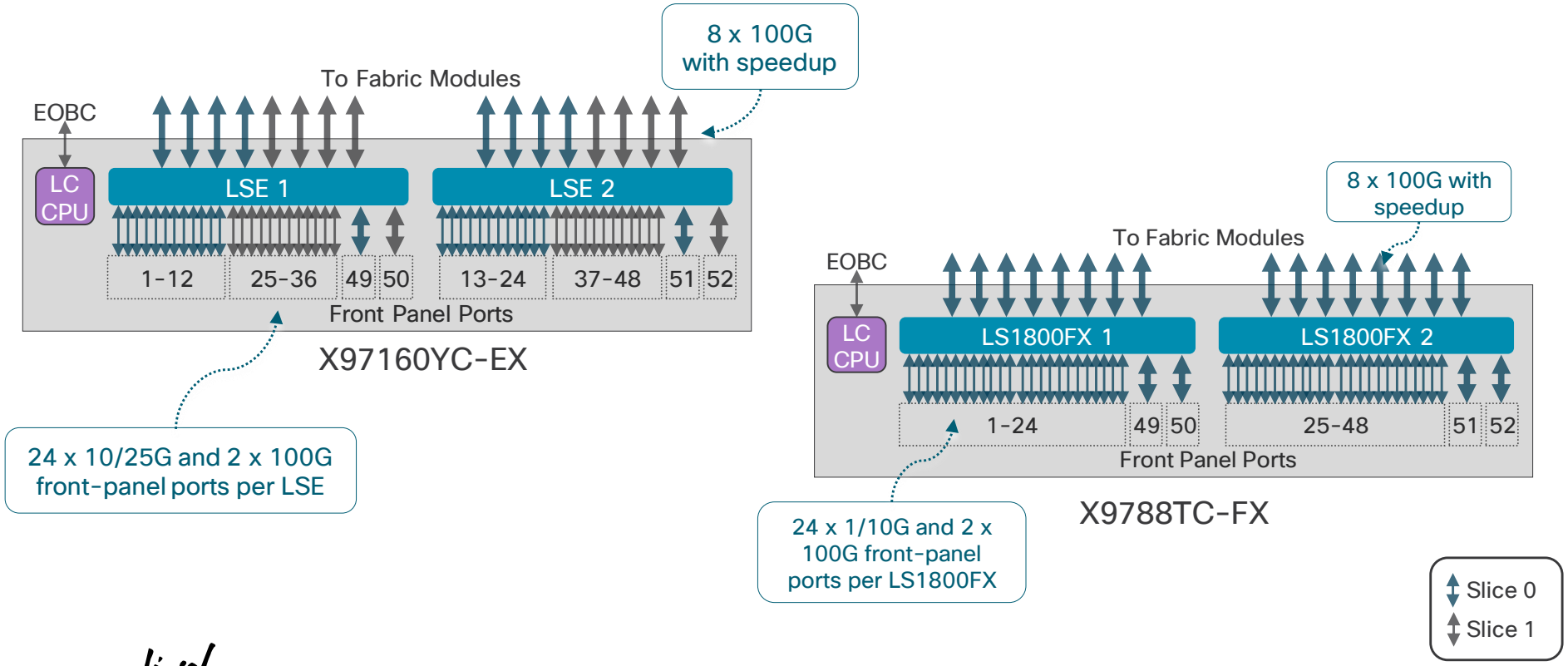
- 1.6Tbps capacity with line-rate performance
- Flexible port configurations –1/10/25G SFP28 ports, 1/10/25/40/50/100G QSFP28 ports

9788-FX:

- 880Gbps capacity with line-rate performance
- Flexible port configurations –1/10GBASE-T ports, 1/10/25/40/50/100G QSFP28 ports

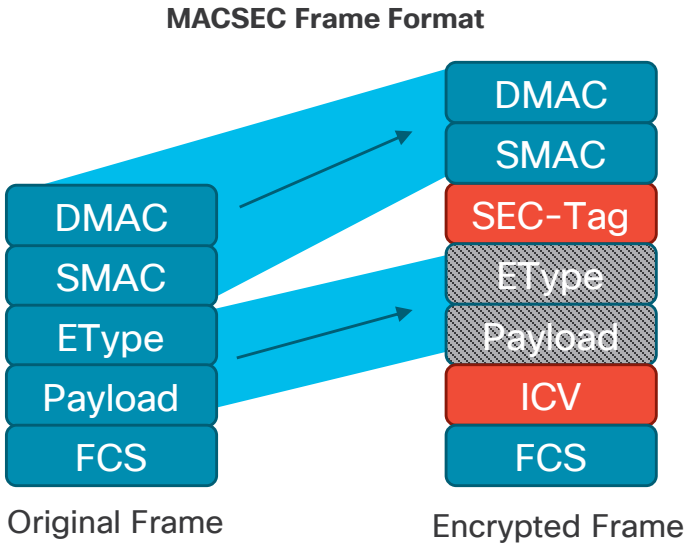
Line-rate MACSEC on all ports

N9K-X97160YC-EX / N9K-X9788TC-FX Architecture

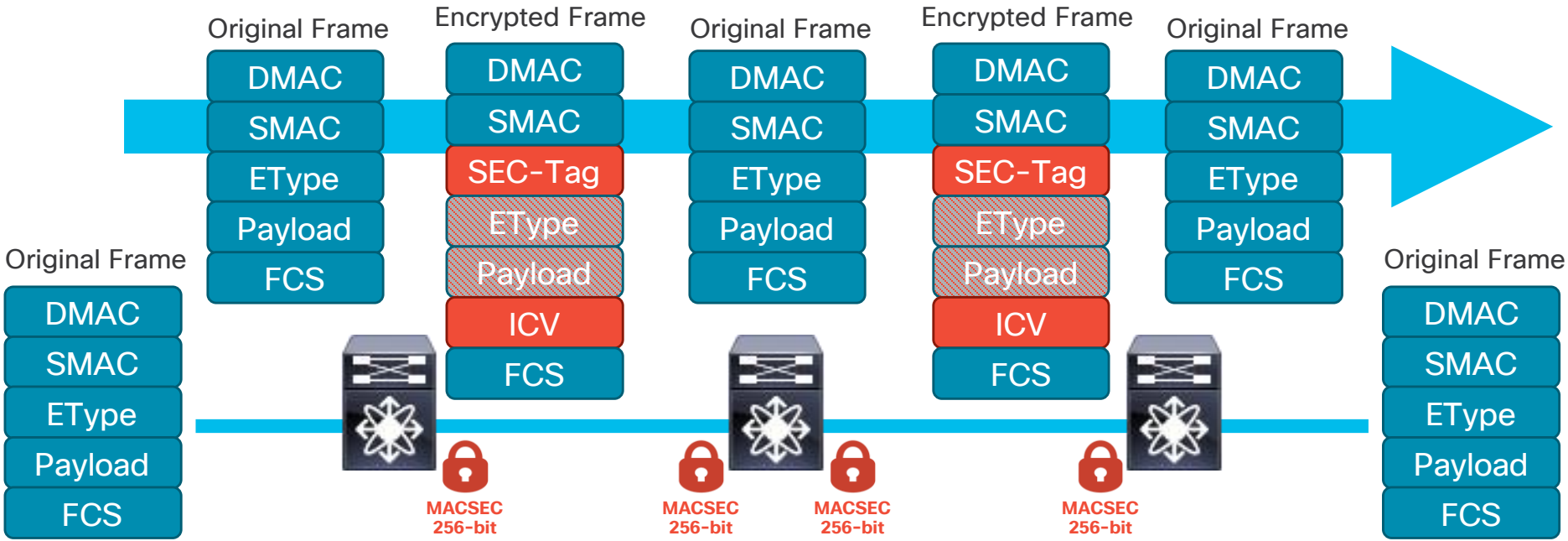


MACSEC Hardware Encryption

- Provides link-level hop-by-hop encryption
- IEEE 802.1AE 128-bit and 256-bit AES encryption with MKA Key Exchange
- Native hardware support available on:
  - All ports on X9736C-FX linecard
  - All ports on Nexus 93180YC-FX / 93108TC-FX switches
  - 16 x 100G ports on Nexus 9364C switch
  - All ports on Nexus 9336C-FX2 / N9K-C93240YC-FX2 switches

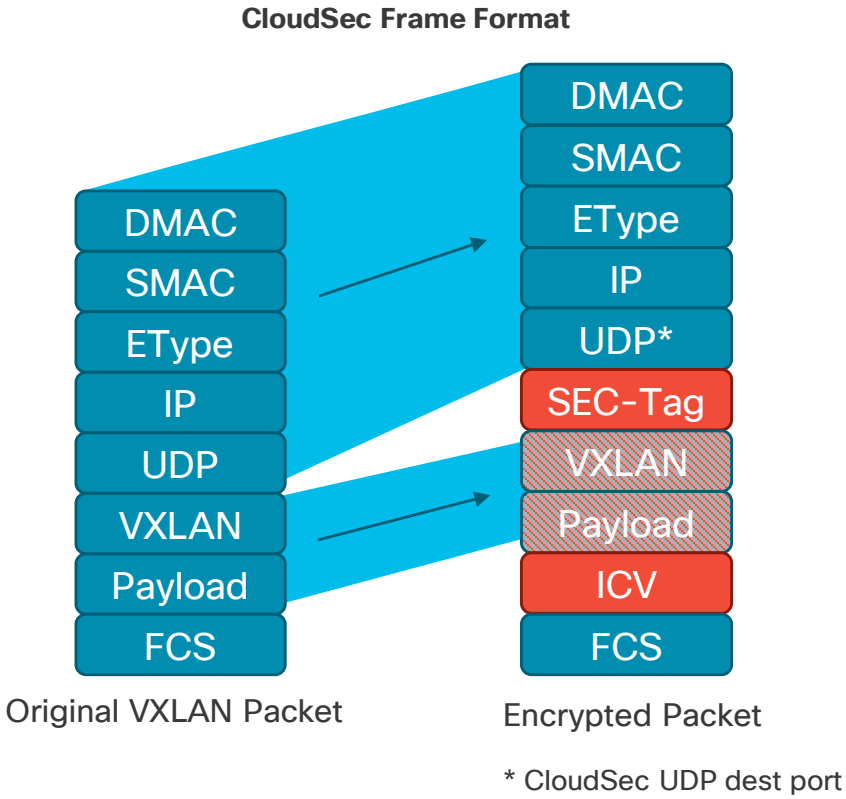


MACSEC

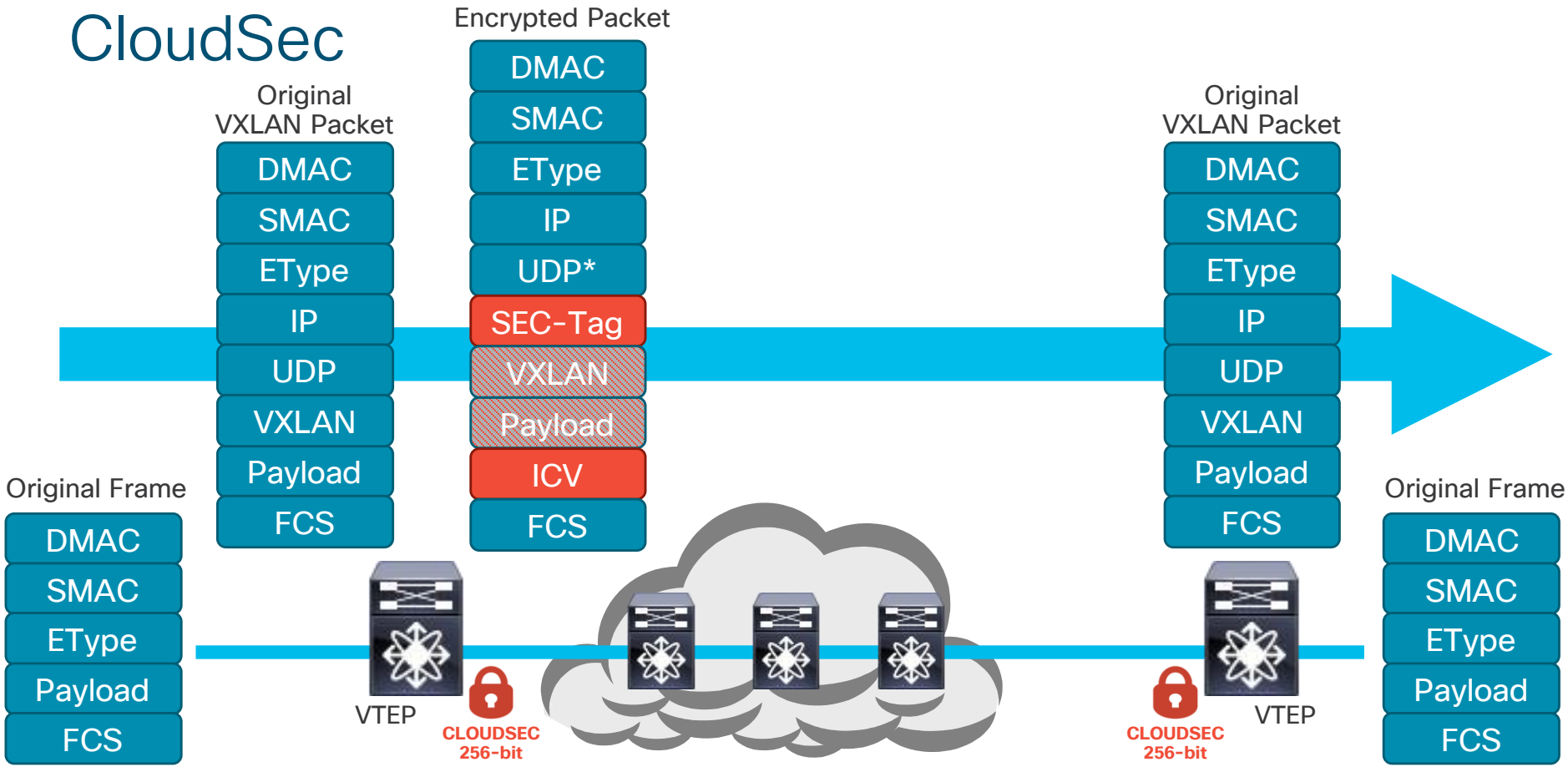


CloudSec Hardware Encryption

- Provides VTEP-to-VTEP encryption
- Encrypts VXLAN header and payload for transport over arbitrary IP network
- Hardware support available on:
  - 8 x 100G ports on X9736C-FX linecard
  - 16 x 100G ports on Nexus 9364C
  - All ports on 9300-FX2 TORs
- No support on other TOR switches



CloudSec

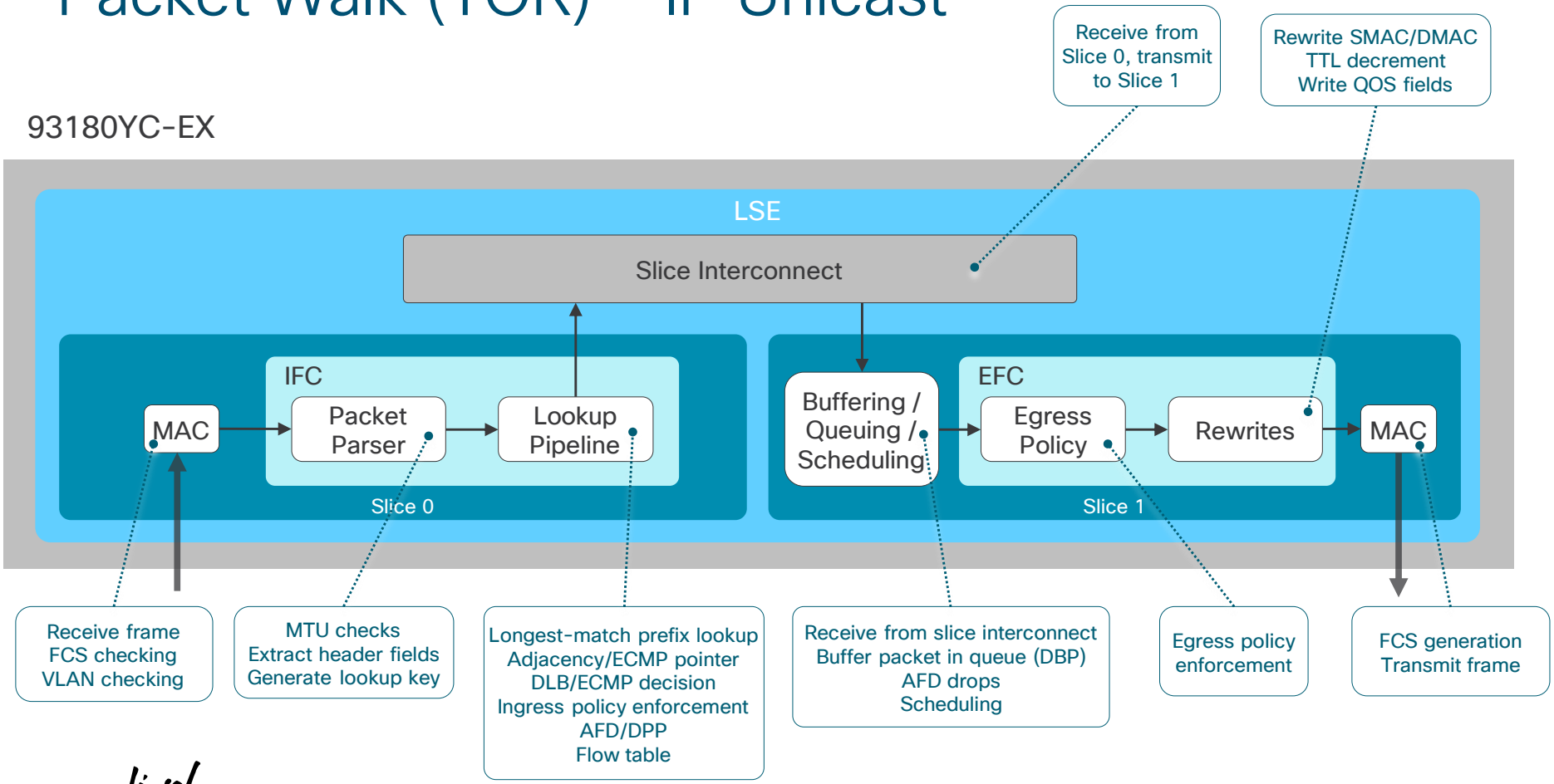


* CloudSec UDP dest port

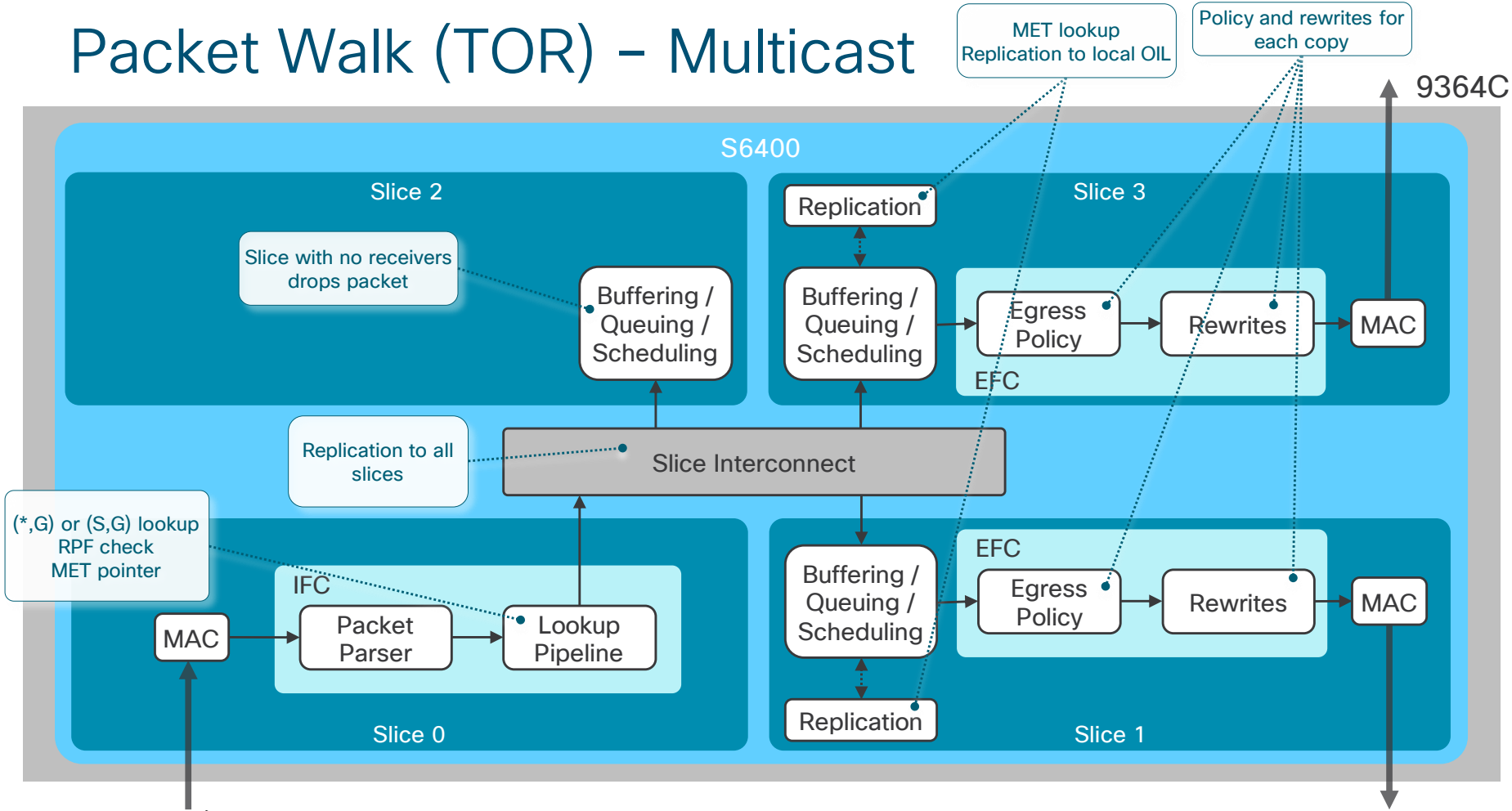
Agenda

- Data Center and Silicon Strategy
- Cloud Scale ASIC Architecture
- Cloud Scale Switching Platforms
- Packet Walks
- Key Takeaways

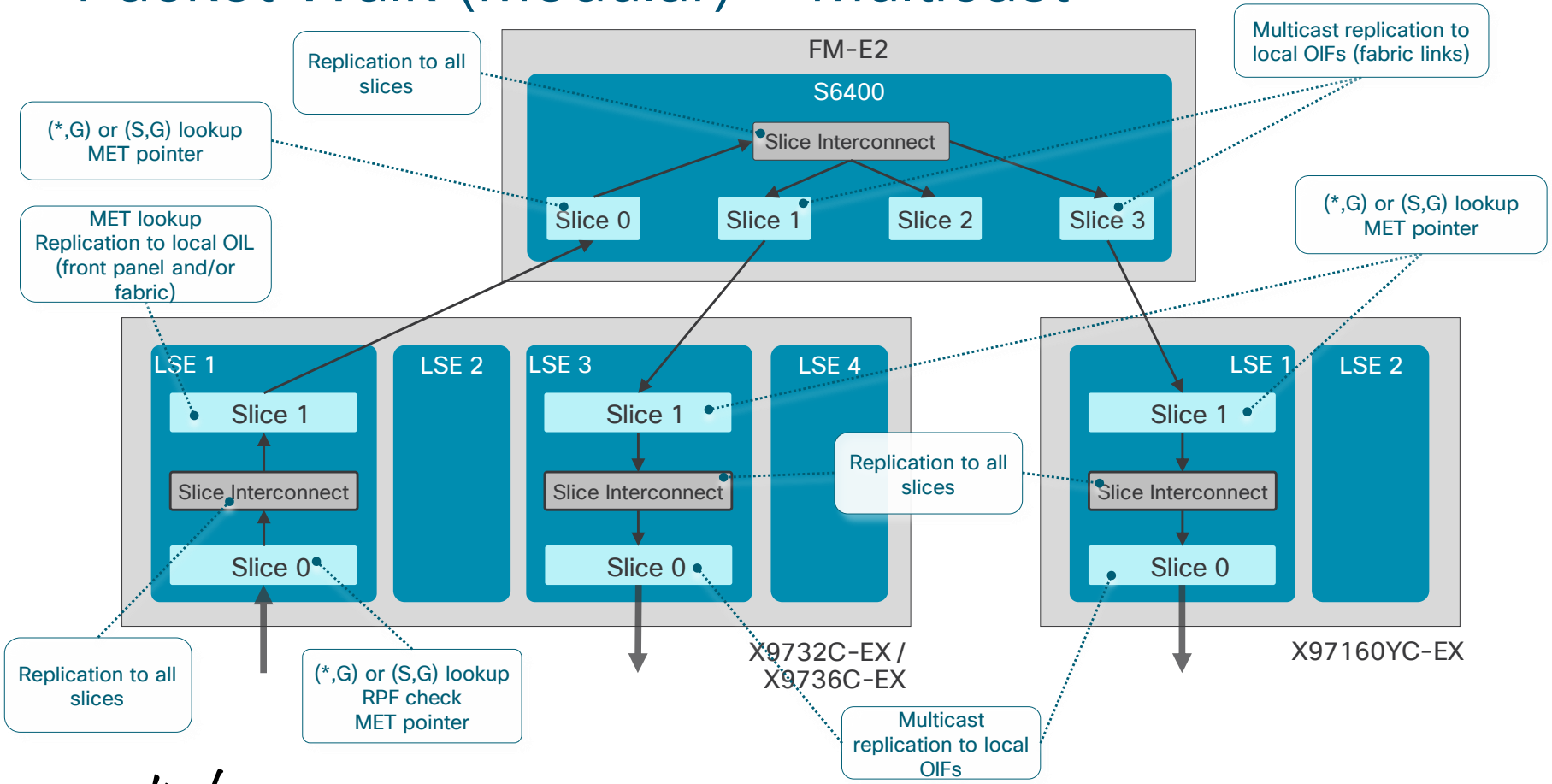
Packet Walk (TOR) – IP Unicast



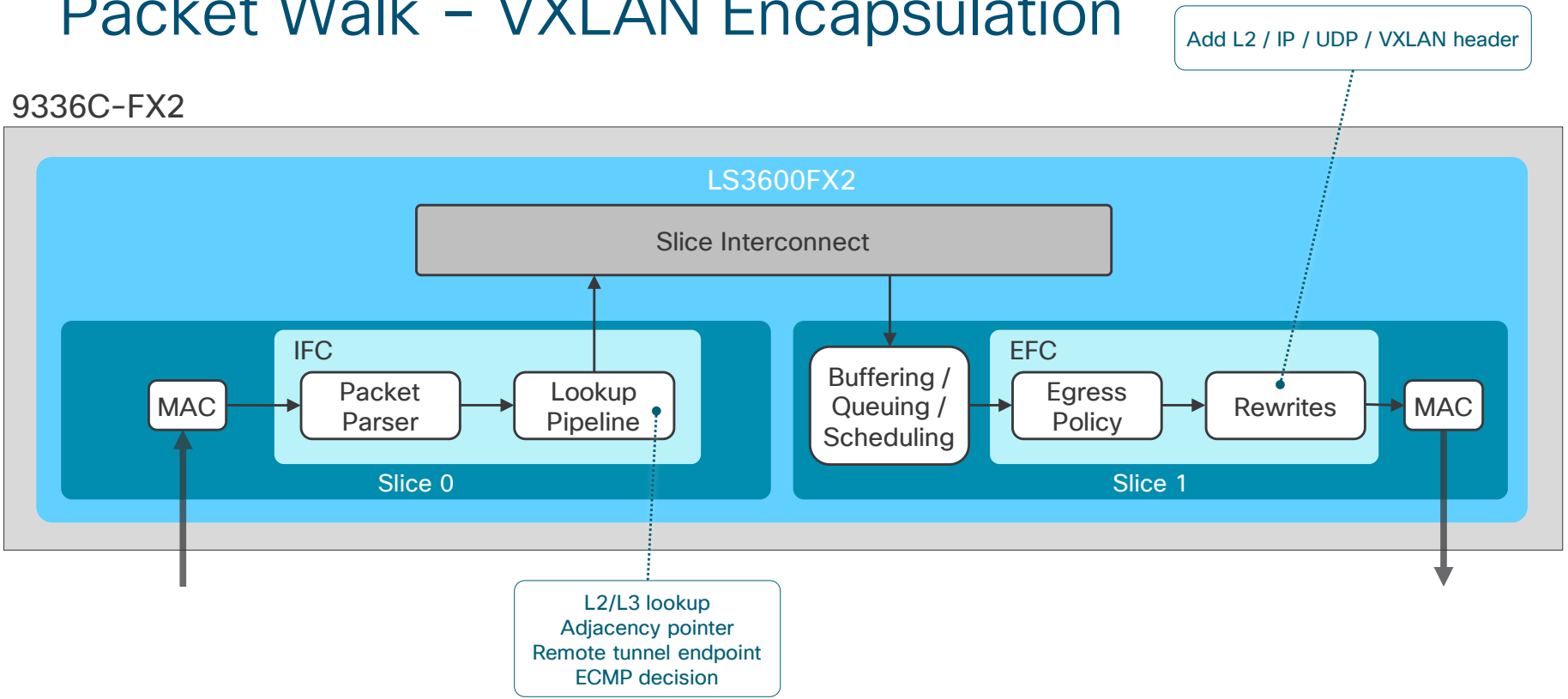
Packet Walk (TOR) – Multicast



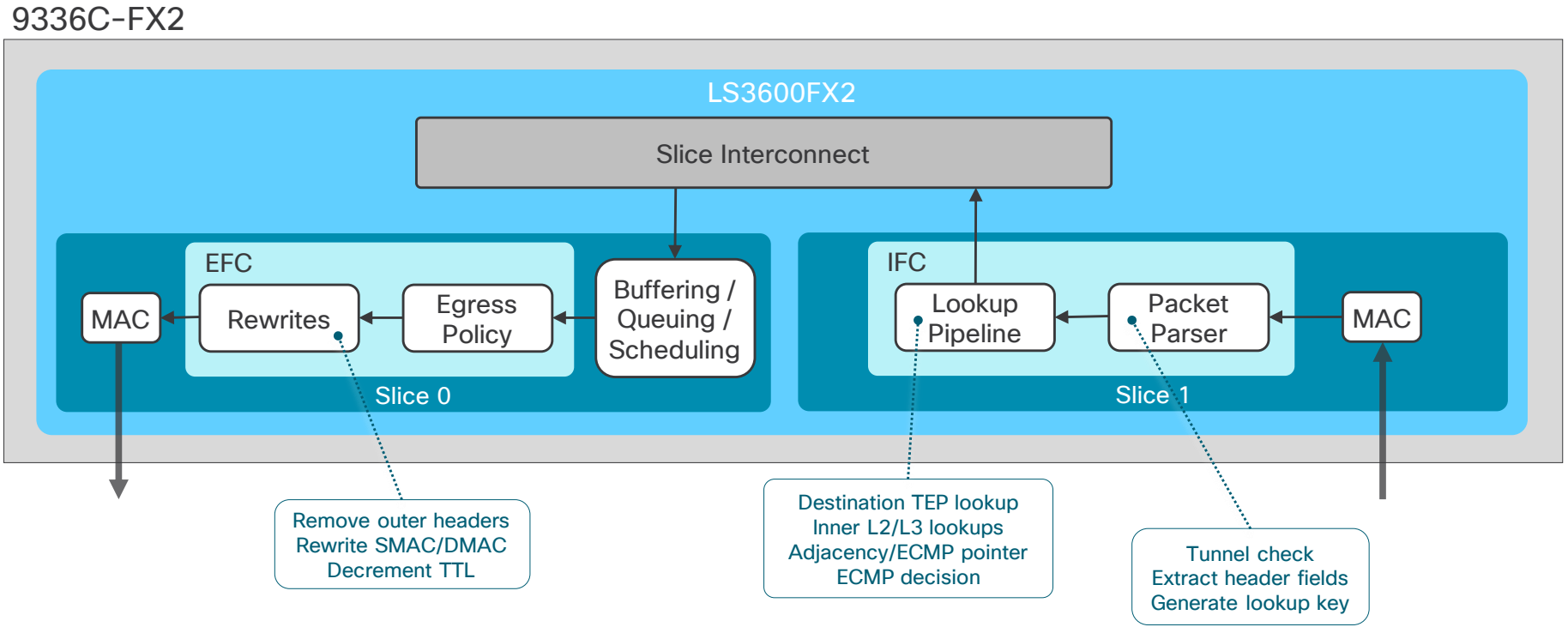
Packet Walk (Modular) – Multicast



Packet Walk – VXLAN Encapsulation



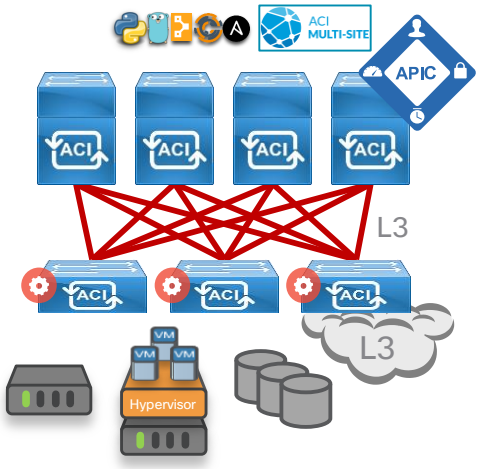
Packet Walk – VXLAN Decapsulation



Agenda

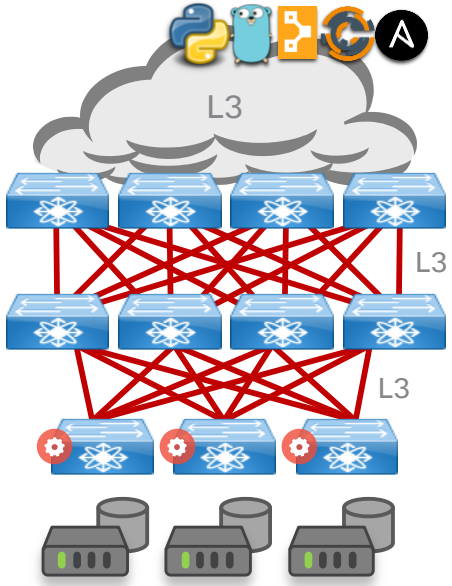
- Data Center and Silicon Strategy
- Cloud Scale ASIC Architecture
- Cloud Scale Switching Platforms
- Packet Walks
- Key Takeaways

Building Data Center Fabrics with Nexus 9000

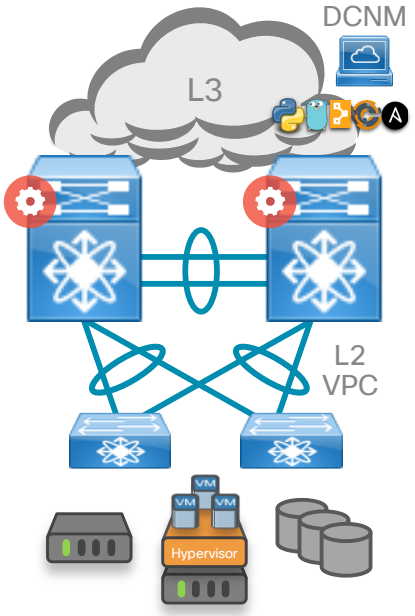


ACI – Turnkey Fabric

Standalone – Programmable Fabric with VXLAN+EVPN



Standalone – Programmable IP Network



Standalone – Traditional Data Center Network

Key Takeaways

- You should now have a thorough understanding of the Nexus 9000 Cloud Scale switching platform architecture
- Feature-rich, innovative switching platform addresses virtually every deployment scenario
- Nexus 9000 Cloud Scale platform forms foundation of Cisco Data Center strategy

Complete your online session evaluation

- Please complete your session survey after each session. Your feedback is very important.
- Complete a minimum of 4 session surveys and the Overall Conference survey (starting on Thursday) to receive your Cisco Live water bottle.
- All surveys can be taken in the Cisco Live Mobile App or by logging in to the Session Catalog on ciscolive.cisco.com/us.

Cisco Live sessions will be available for viewing on demand after the event at ciscolive.cisco.com.

Cisco Webex Teams

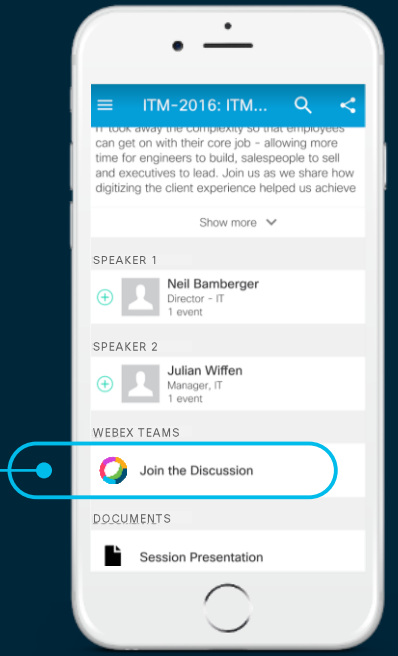
Questions?

Use Cisco Webex Teams to chat with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install Webex Teams or go directly to the team space
4. Enter messages/questions in the team space

Webex Teams will be moderated by the speaker until June 16, 2019.



cs.co/ciscolivebot#BRKARC-3222

Continue your education

Demos in the Cisco campus

Walk-in labs

Meet the engineer 1:1 meetings

Related sessions

Thank you

You make **possible**