Cisco live!

What You Make Possible

TOMORROW
starts here.

CISCO

Cisco live!

Cisco Nexus 7000 Switch Architecture

BRKARC-3470

TOMORROW starts here.

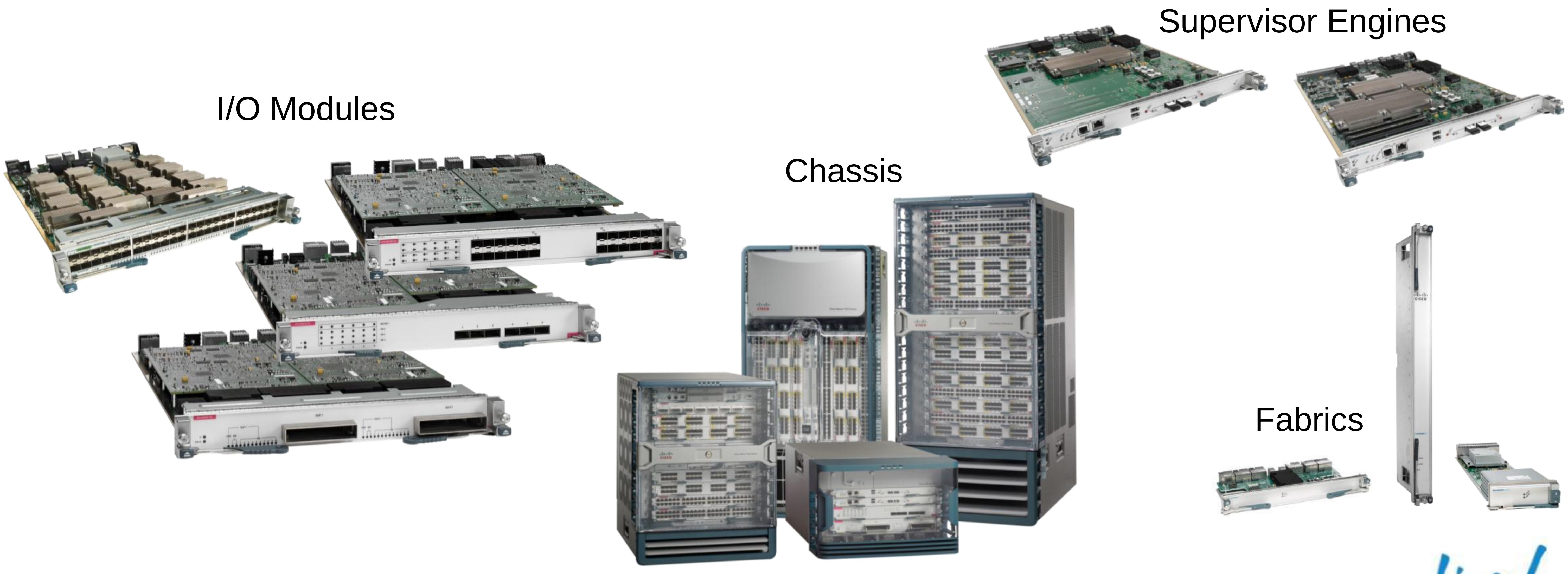
Session Goal

- To provide a thorough understanding of the Cisco Nexus™ 7000 switching architecture, supervisor, fabric, and I/O module design, packet flows, and key forwarding engine functions
- This session will examine only the latest additions to the Nexus 7000 platform
- This session will not examine NX-OS software architecture or other Nexus platform architectures

© 2013 Cisco and/or its affiliates. All rights reserved.

What is Nexus 7000?

Data-center class Ethernet switch designed to deliver high-availability, system scale, usability, investment protection



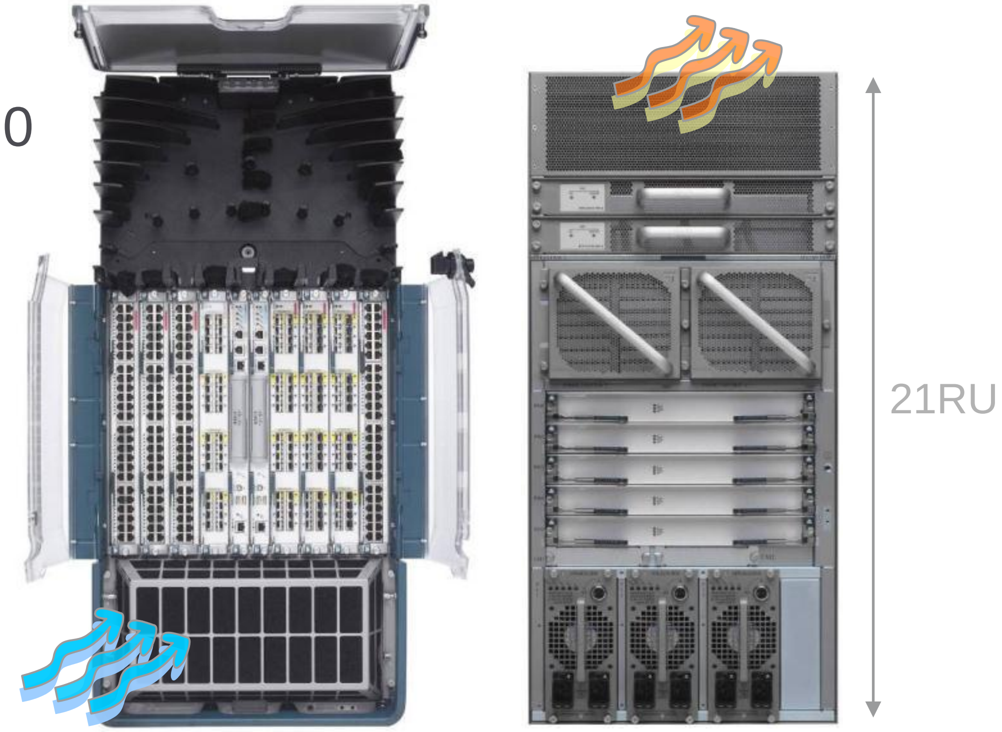
Agenda

- Chassis Architecture
- Supervisor Engine and I/O Module Architecture
- Forwarding Engine Architecture
- Fabric Architecture
- I/O Module Queuing
- Layer 2 Forwarding
- IP Forwarding
- Classification
- NetFlow
- Conclusion

© 2013 Cisco and/or its affiliates. All rights reserved.

Nexus 7000 Chassis Family

Nexus 7010



21RU

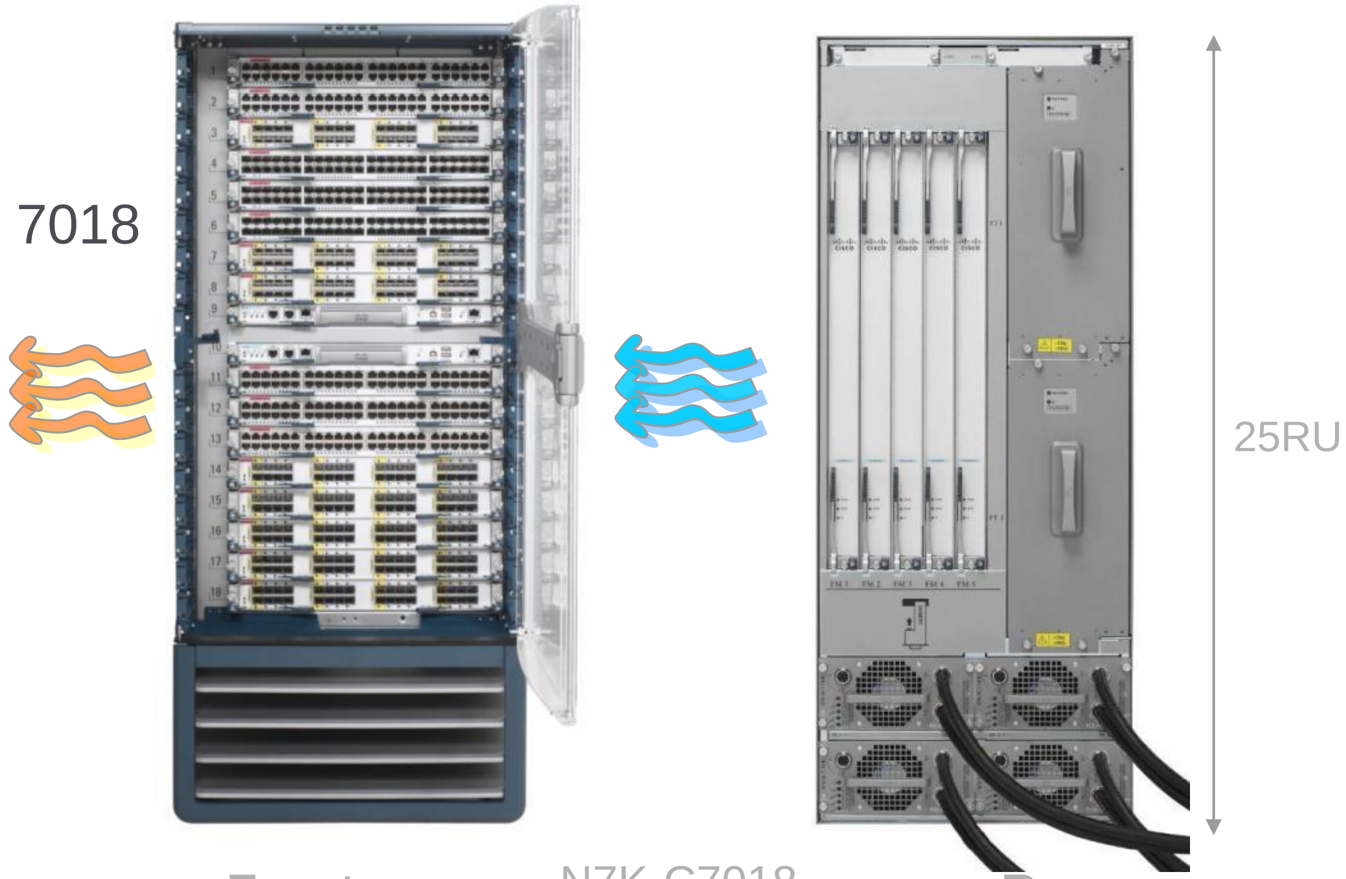
Front

N7K-C7010

Rear

NX-OS 4.1(2) and later

Nexus 7018



25RU

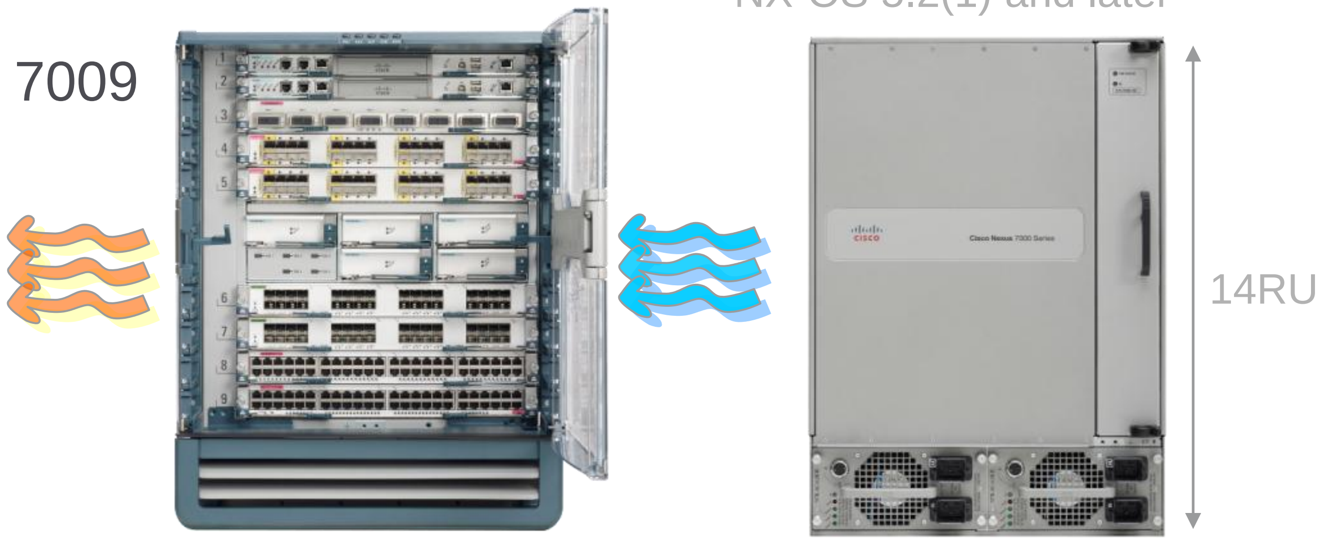
Front

N7K-C7018

Rear

NX-OS 5.2(1) and later

Nexus 7009



14RU

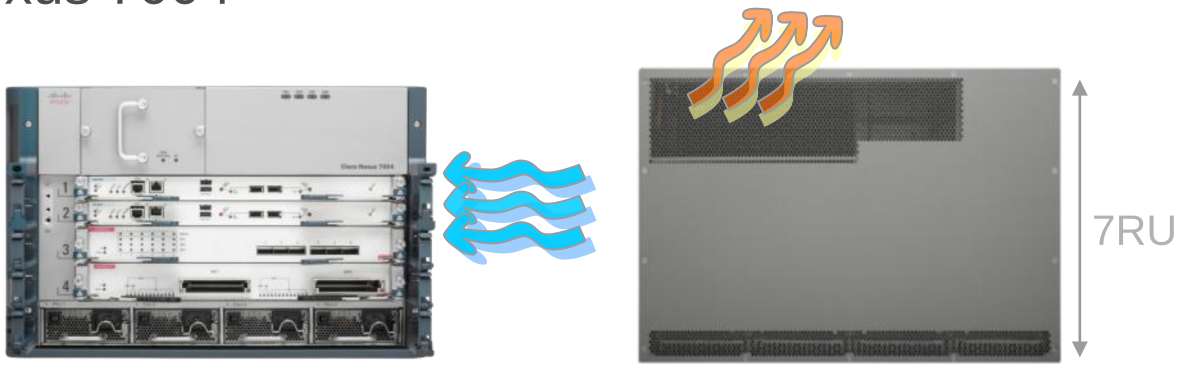
Front

N7K-C7009

Rear

NX-OS 6.1(2) and later

Nexus 7004



7RU

Front

N7K-C7004

Rear

© 2013 Cisco and/or its affiliates. All rights reserved.

Supported in NX-OS release 6.1(2) and later

Nexus 7004 Chassis

- 4 slot chassis – 2 payload slots, 2 supervisor slots
- No fabric modules – I/O modules connect back-to-back
- Side-to-back airflow
- 3 X 3000W power supplies (AC or DC)
- All FRUs accessed from chassis front
- Supports Sup2 / 2E only
- Supports M1L, M2, F2, F2E modules
  - No support for M1 non-L, F1 modules

Key Chassis Components

- Common components:
  - – Supervisor Engines
  - – I/O Modules
  - – Power Supplies (except 7004)
- Chassis-specific components:
  - – Fabric Modules
  - – Fan Trays

Agenda

- Chassis Architecture
- Supervisor Engine and I/O Module Architecture
- Forwarding Engine Architecture
- Fabric Architecture
- I/O Module Queuing
- Layer 2 Forwarding
- IP Forwarding
- Classification
- NetFlow
- Conclusion

© 2013 Cisco and/or its affiliates. All rights reserved.

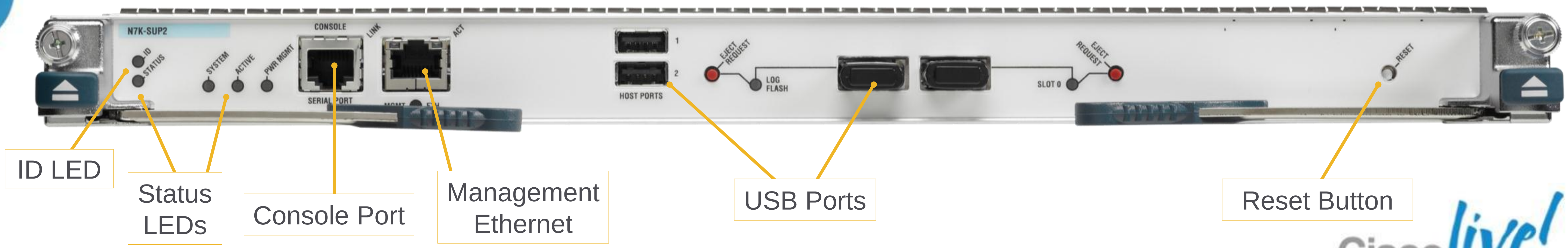
Supervisor Engine 2 / 2E

- Next generation supervisors providing control plane and management functions

Supervisor Engine 2	Supervisor Engine 2E
Base performance	High performance
One quad-core 2.1GHz CPU with 12GB DRAM	Two quad-core 2.1GHz CPU with 32GB DRAM

- Second-generation dedicated central arbiter ASIC
  - Controls access to fabric bandwidth via dedicated arbitration path to I/O modules
- Interfaces with I/O modules via 1G switched EOBC

N7K-SUP2/N7K-SUP2E



© 2013 Cisco and/or its affiliates. All rights reserved.

Nexus 7000 I/O Module Families

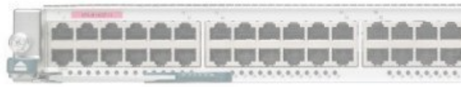
M Series and F Series

- M Series – L2/L3/L4 with large forwarding tables and rich feature set

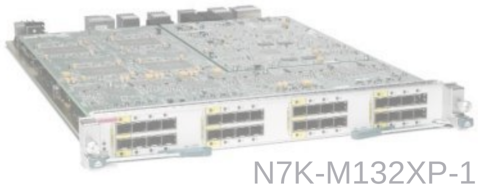
N7K-M108X2-12L

N7K-M224XP-23L

N7K-M206FQ-23L



N7K-M148GT-11L

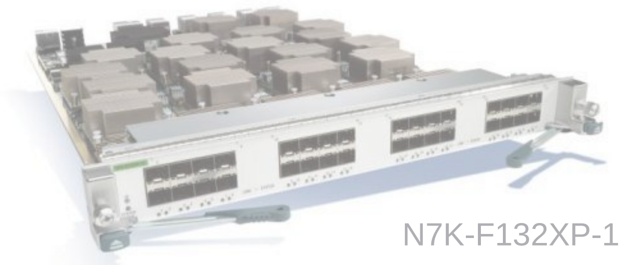


N7K-M132XP-12L

N7K-M202CF-22L

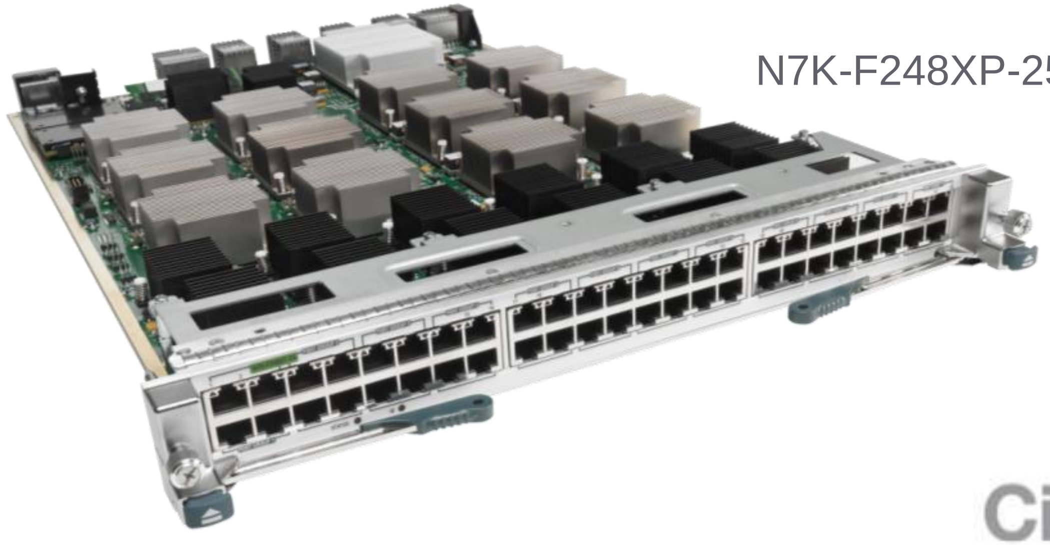
N7K-M148GS-11L

- F Series – High performance, low latency, low power with streamlined feature set



N7K-F132XP-15

N7K-F248XP-25



N7K-F248XP-25E

© 2013 Cisco and/or its affiliates. All rights reserved.

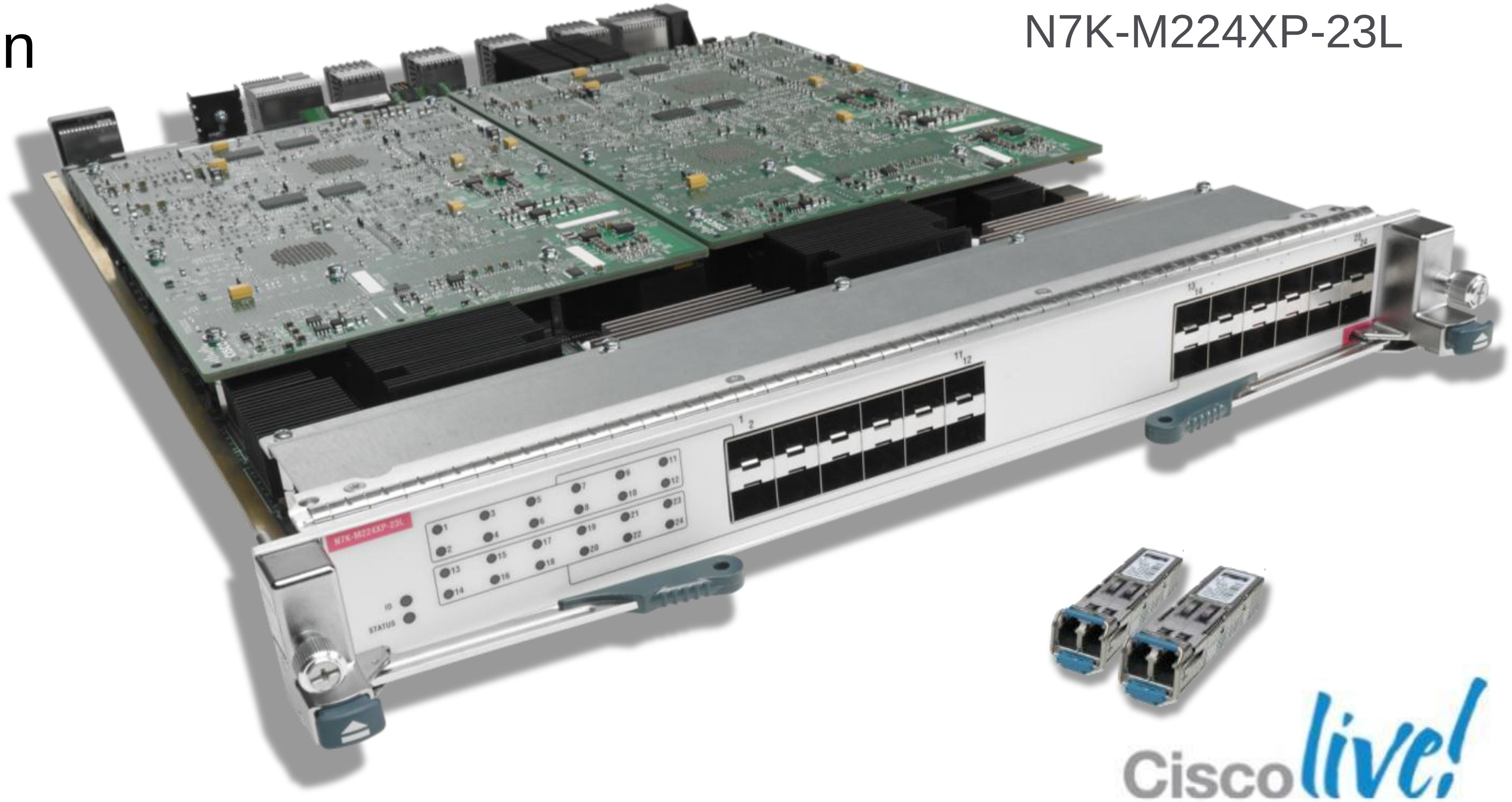
Cisco Public

Supported in NX-OS release 6.1(1) and later

24-Port 10G M2 I/O Module

N7K-M224XP-23L

- 24-port 10G with SFP+ transceivers
- 240G full-duplex fabric connectivity
- Two integrated forwarding engines (120Mpps)
  - Support for “XL” forwarding tables (licensed feature)
- Distributed L3 multicast replication
- 802.1AE LinkSec on all ports

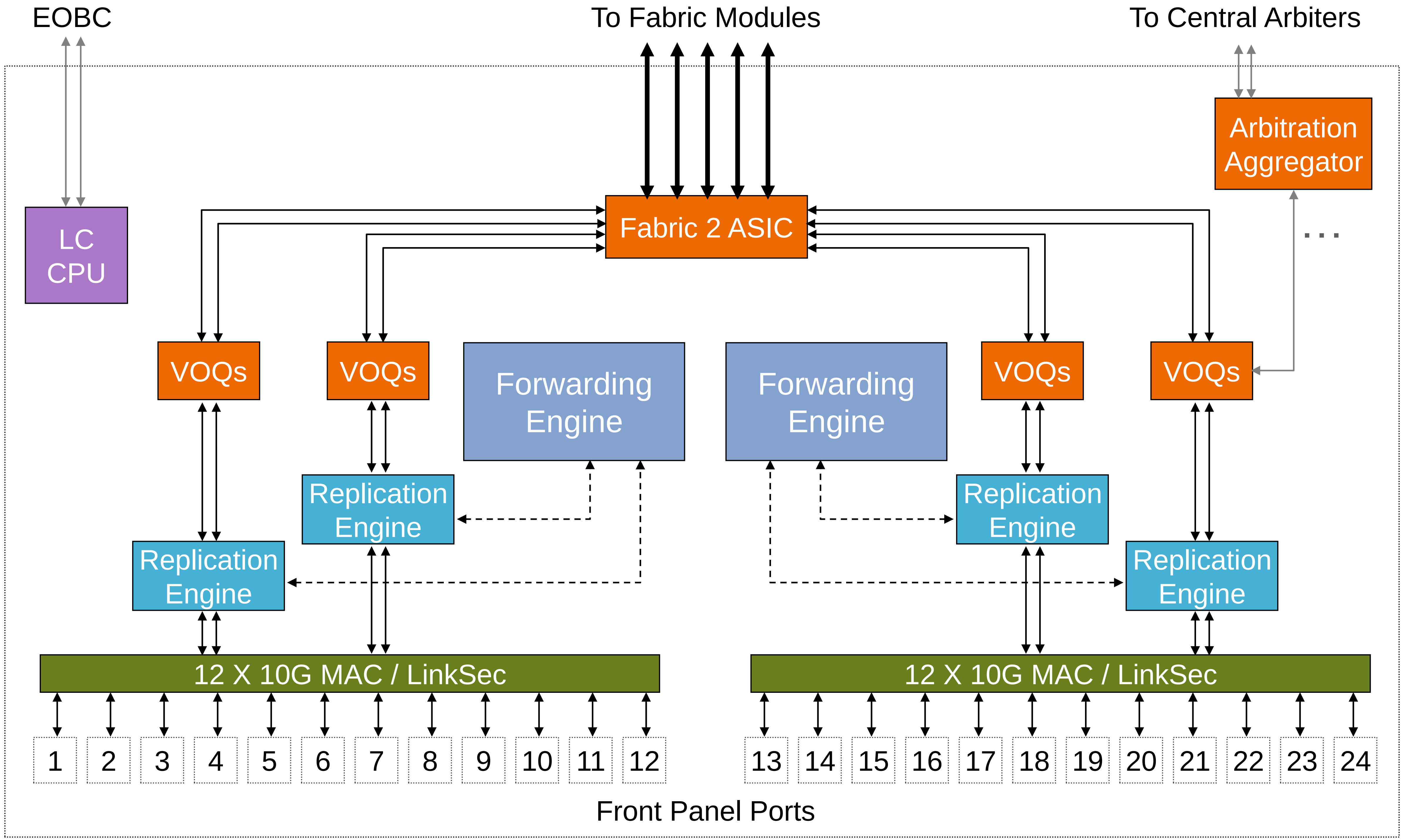


N7K-M224XP-23L

© 2013 Cisco and/or its affiliates. All rights reserved.

24-Port 10G M2 I/O Module Architecture

N7K-M224XP-23L



© 2013 Cisco and/or its affiliates. All rights reserved.

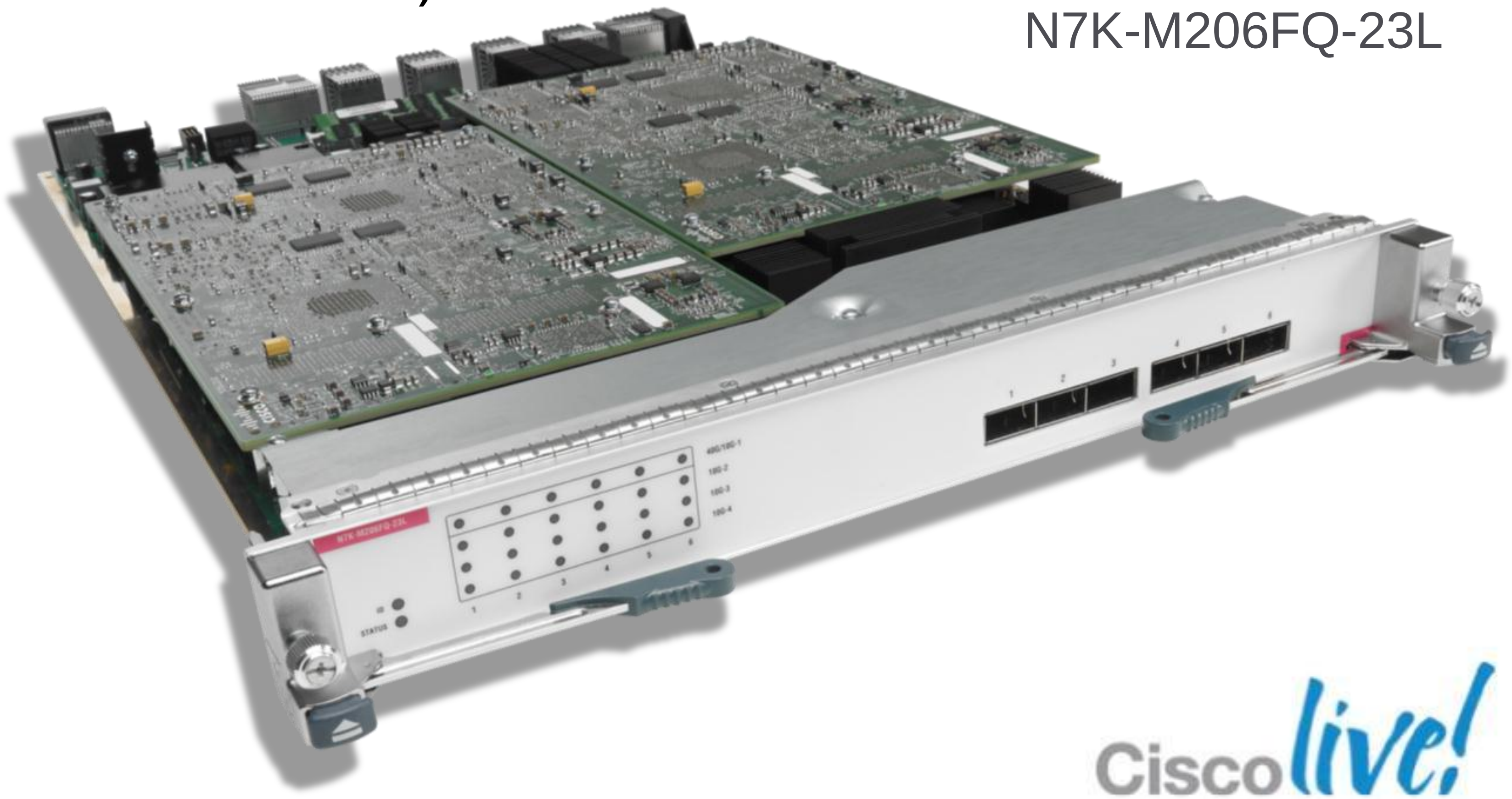
Supported in NX-OS release 6.1(1) and later

M Series

6-Port 40G M2 I/O Module

N7K-M206FQ-23L

- 6-port 40G with QSFP+ transceivers
  - Option to breakout to 4X10G interfaces per 40G port*
- 240G full-duplex fabric connectivity
- Two integrated forwarding engines (120Mpps)
- Support for “XL” forwarding tables (licensed feature)
- Distributed L3 multicast replication
- 802.1AE LinkSec on all ports

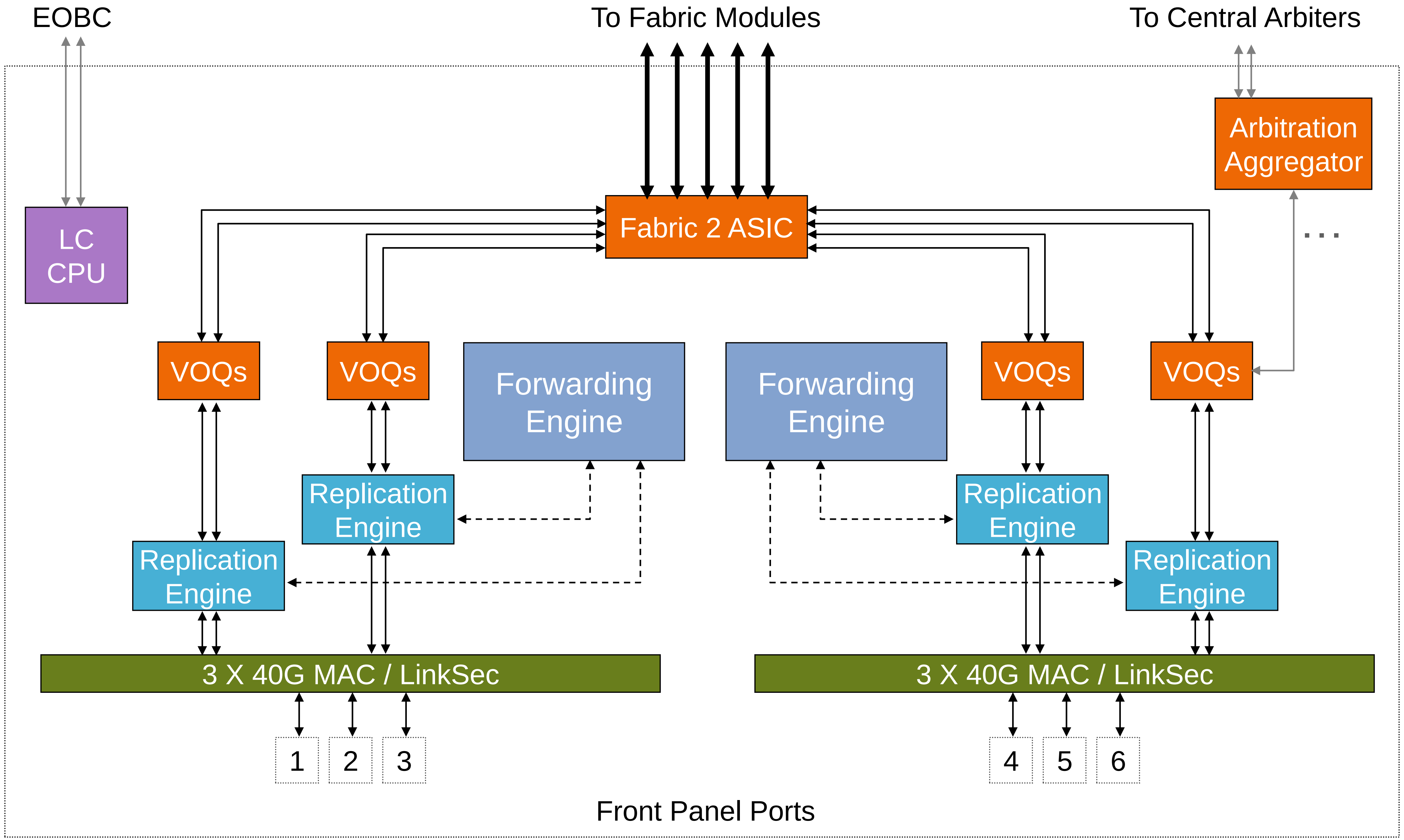


N7K-M206FQ-23L

* Roadmap feature

6-Port 40G M2 I/O Module Architecture

N7K-M206FQ-23L



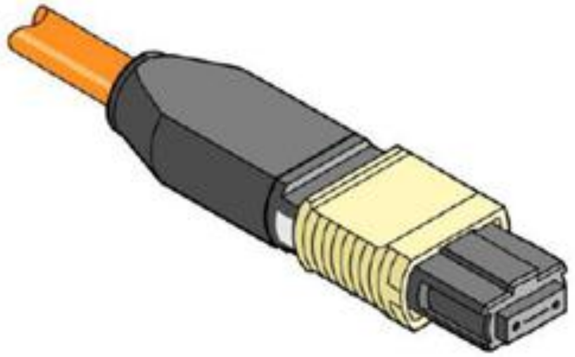
© 2013 Cisco and/or its affiliates. All rights reserved.

M Series

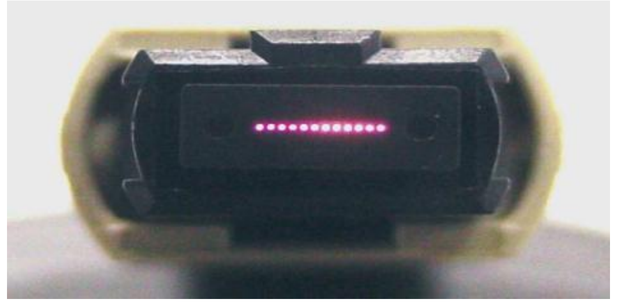
40G Transceivers – QSFP+

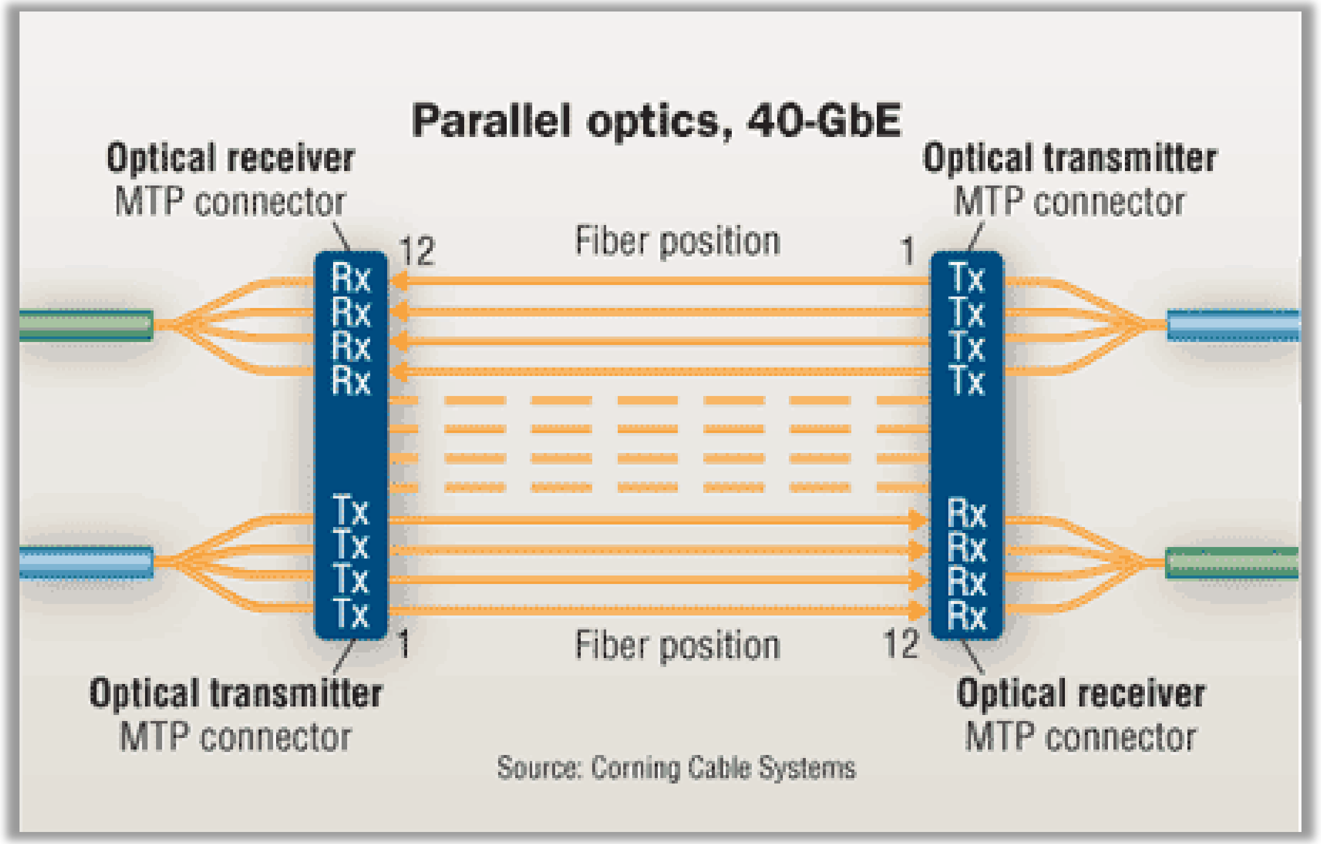
QSFP-40G-SR4

- 40GBASE-SR4 supported in 6.1(1)
  - 12-fibre MPO/MTP connector
  - 100m over OM3 MMF, 150m over OM4 MMF
- Other form-factors TBA


MPO Optical Connector

Interior of ribbon fibre cable


40G MPO interface
(one row of 12 fibres)



Source: Corning Cable Systems

40G 12-strand ribbon fibre
(4 middle fibres unused)

© 2013 Cisco and/or its affiliates. All rights reserved.

M Series

Supported in NX-OS release 6.1(1) and later

2-Port 100G M2 I/O Module

N7K-M202CF-22L

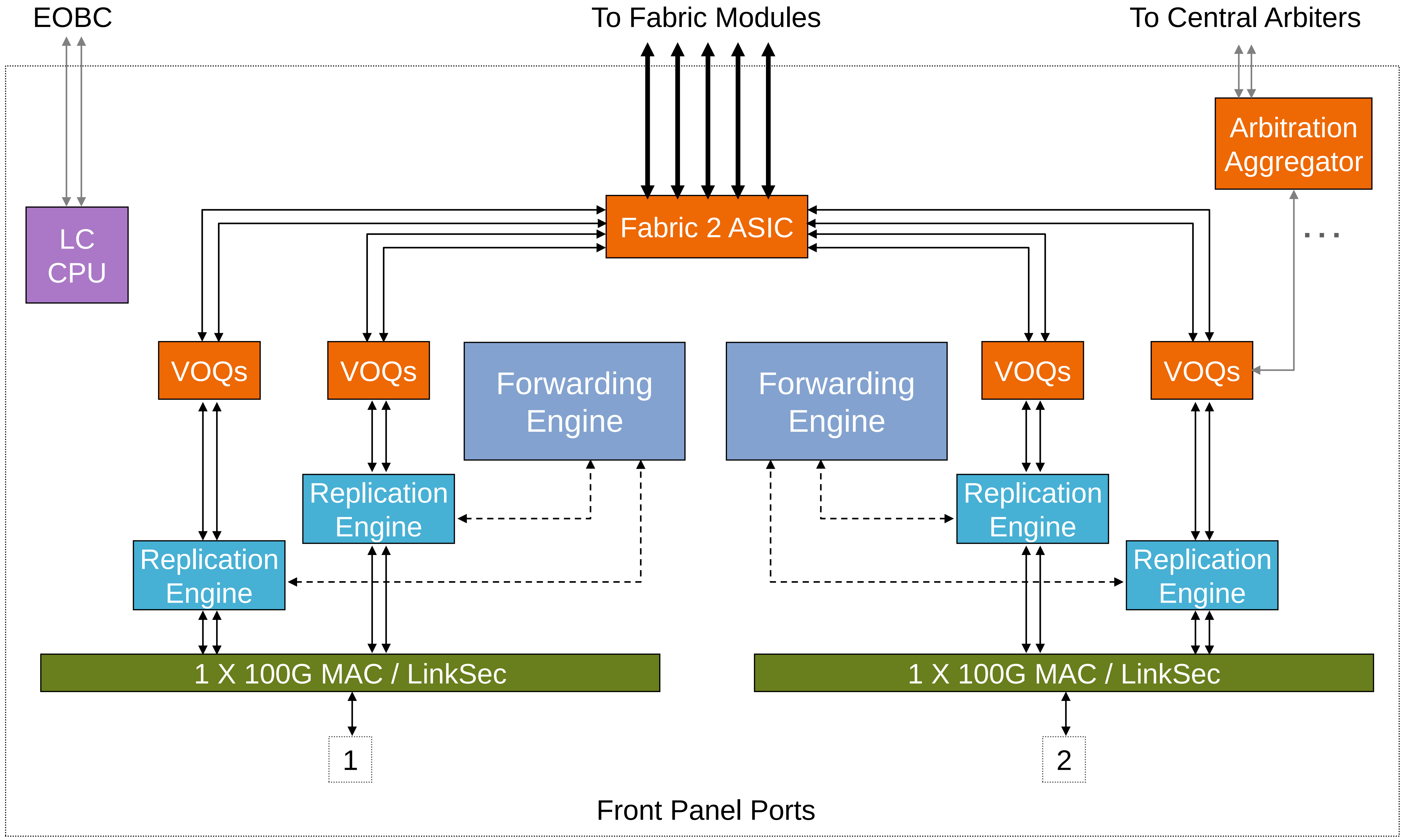
- 2-port 100G with CFP transceivers
  - Option to breakout to 2X40G or 10X10G interfaces per 100G port*
- 200G full-duplex fabric connectivity
- Two integrated forwarding engines (120Mpps)
- Support for "XL" forwarding tables (licensed feature)
- Distributed L3 multicast replication
- 802.1AE LinkSec on all ports

N7K-M202CF-22L

* Roadmap feature

2-Port 100G M2 I/O Module Architecture

N7K-M202CF-22L



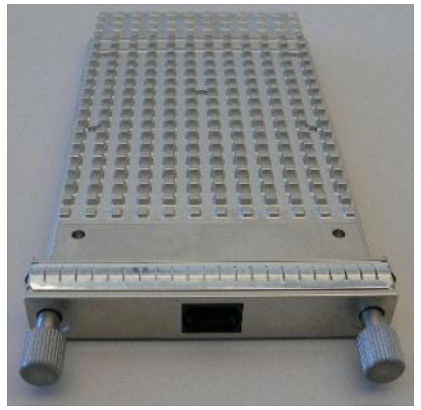
© 2013 Cisco and/or its affiliates. All rights reserved.

100G Module Transceivers – 40G and 100G CFP

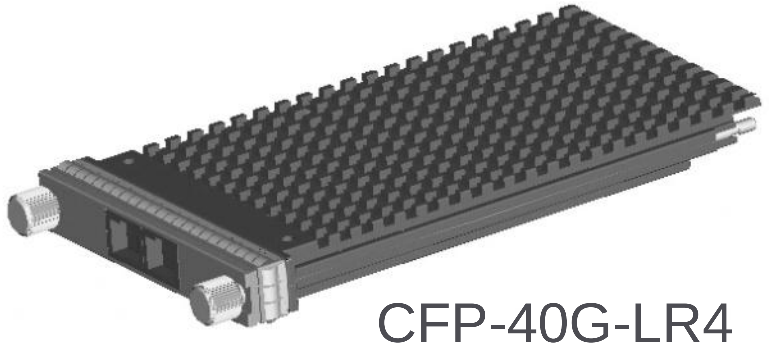
- 100GBASE-LR4 supported from 6.1(1)
  - SC connector
  - 10km over SMF
- Other form-factors on roadmap

CFP-100G-LR4

- 40GBASE-SR4 supported from 6.1(2)
  - 12-fibre MPO/MTP connector
  - 100m over MMF
- 40GBASE-LR4 supported from 6.1(2)
  - SC connector
  - 10km over SMF



CFP-40G-SR4



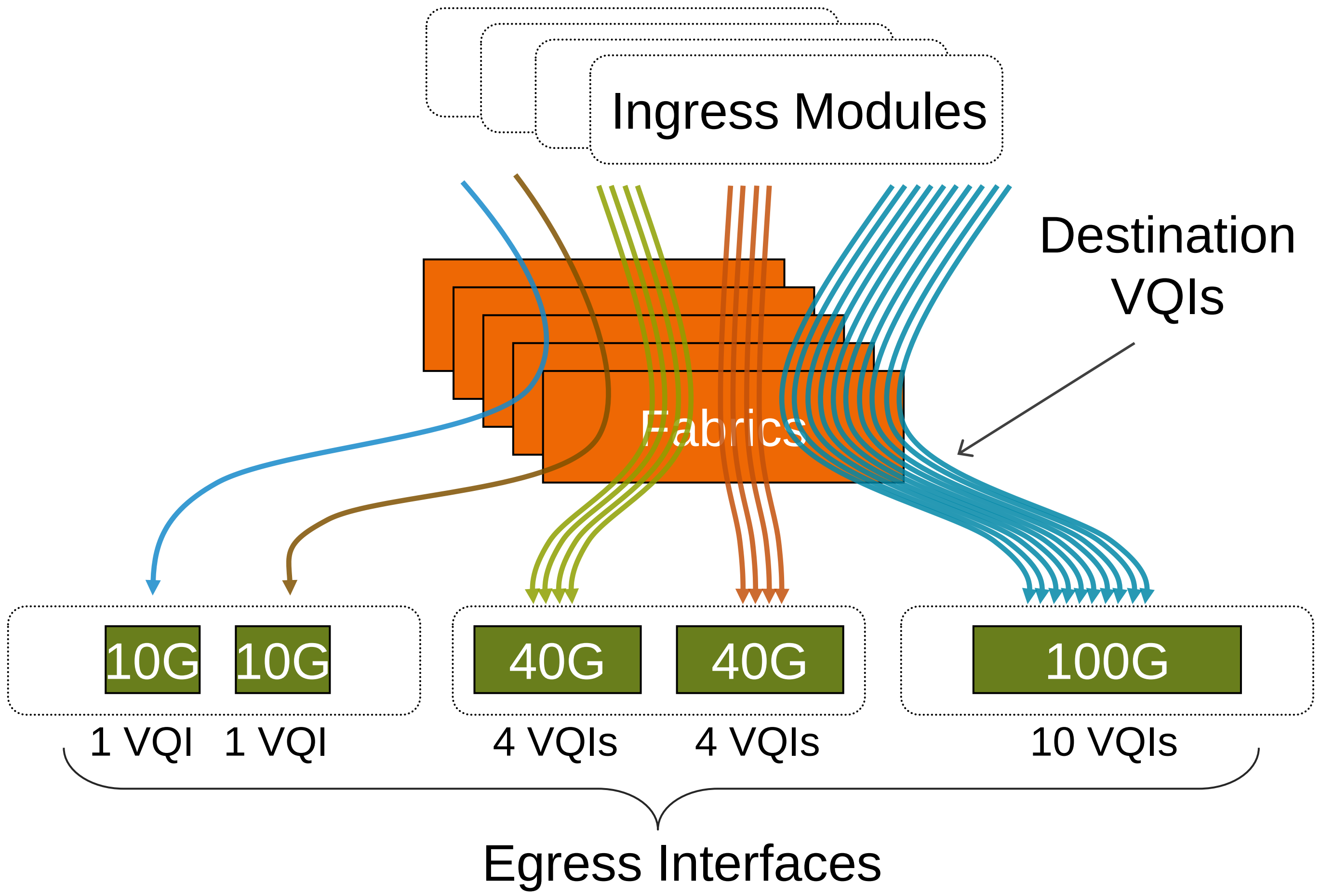
CFP-40G-LR4

© 2013 Cisco and/or its affiliates. All rights reserved.

M Series

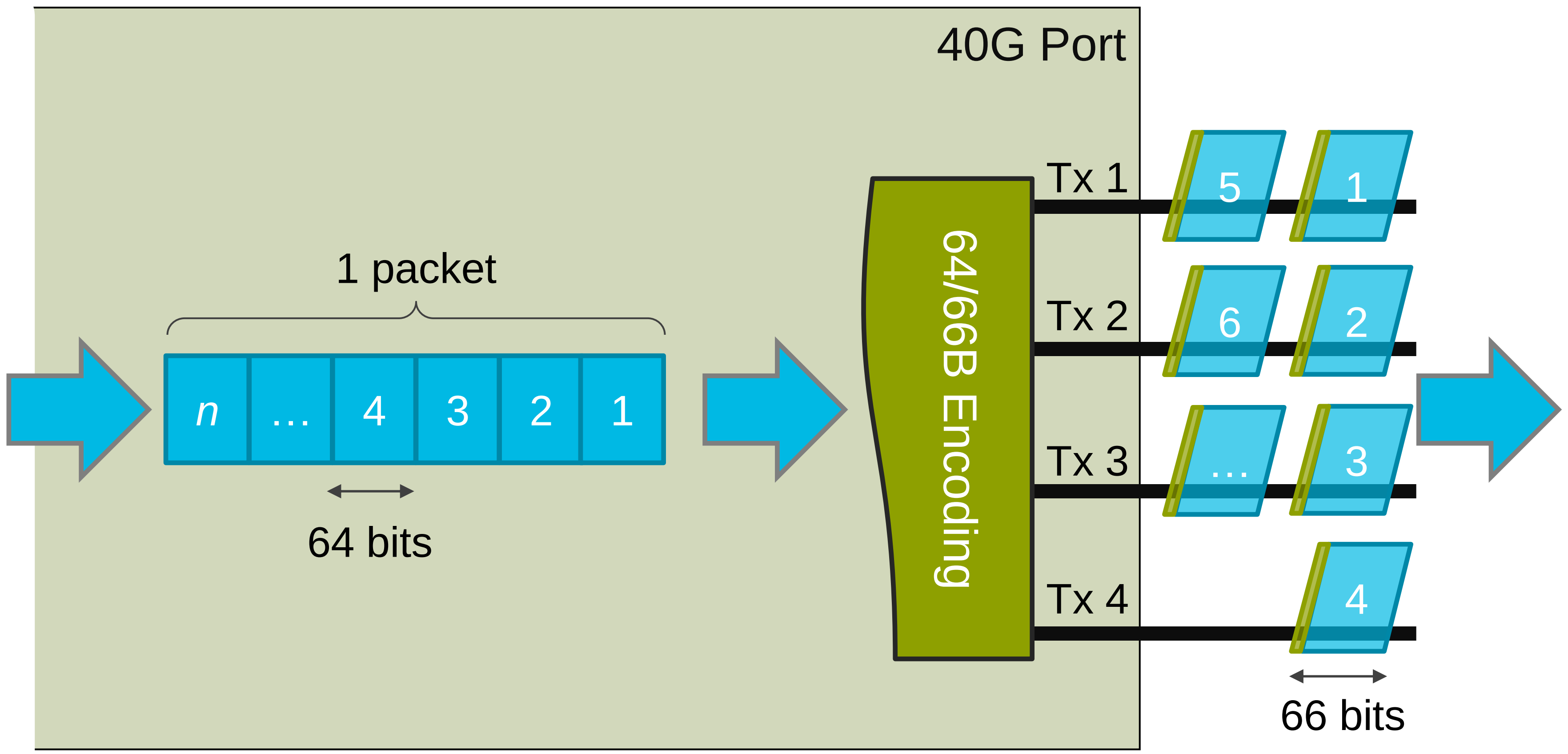
40G and 100G Flow Limits – Internal versus “On the Wire”

Internal to Nexus 7000 System



- Each VQI sustains 10-12G traffic flow
- Single-flow limit is ~10G

On the Wire (40G)



- Packets split into 66-bit “code words”
- Four code words transmitted in parallel, one on each physical Tx fibre
- No per-flow limit imposed – splitting occurs at physical layer

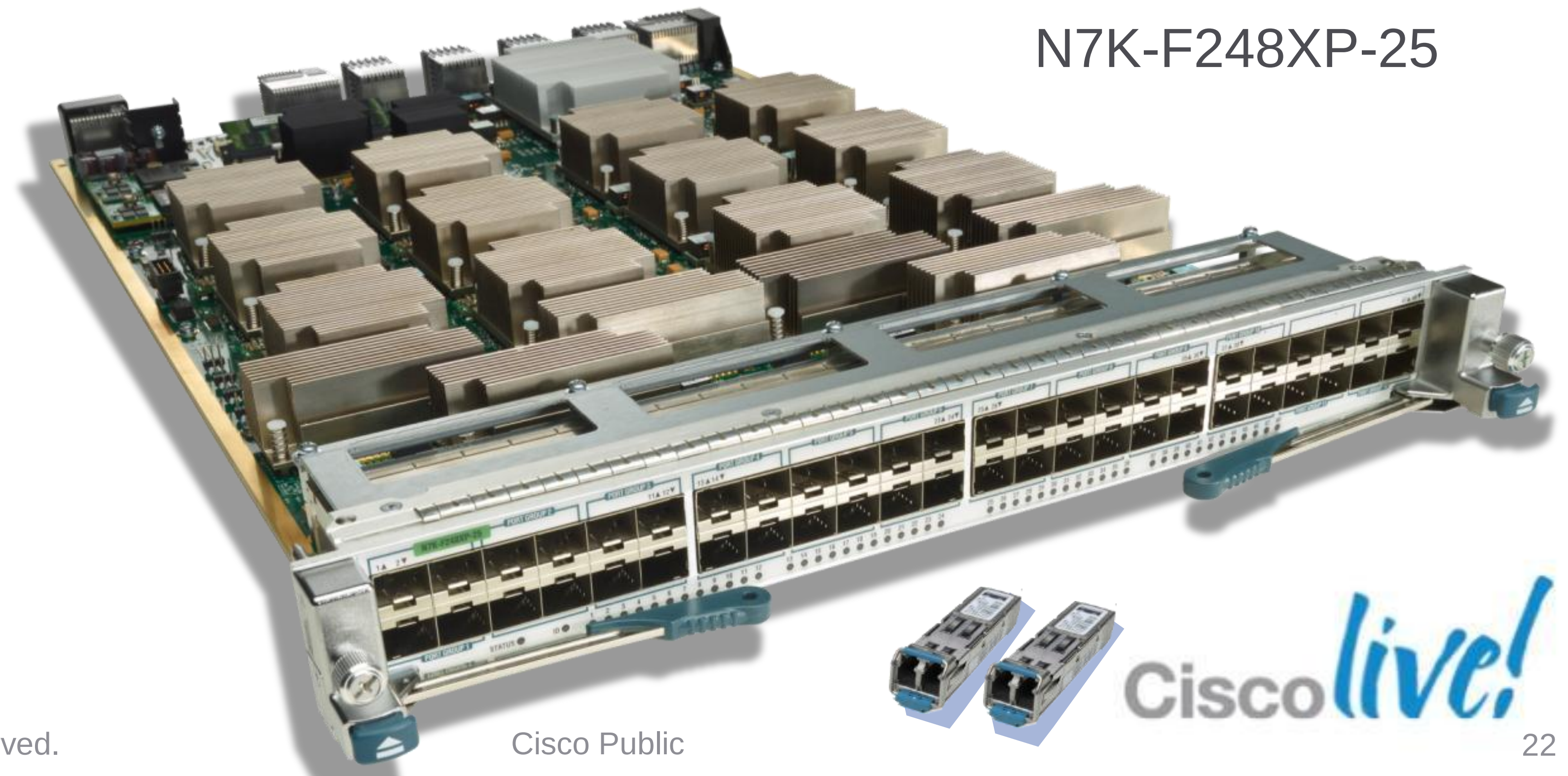
© 2013 Cisco and/or its affiliates. All rights reserved.

Supported in NX-OS release 6.0(1) and later

48-Port 1G/10G F2 I/O Module

N7K-F248XP-25

- 48-port 1G/10G with SFP/SFP+ transceivers
- 480G full-duplex fabric connectivity
- System-on-chip (SoC)* forwarding engine design
  - 12 independent SoC ASICs
- Layer 2/Layer 3 forwarding with L3/L4 services (ACL/QoS)
- Supports Nexus 2000 (FEX) connections
- FabricPath-capable
- FCoE-capable



N7K-F248XP-25

* sometimes called “switch-on-chip”

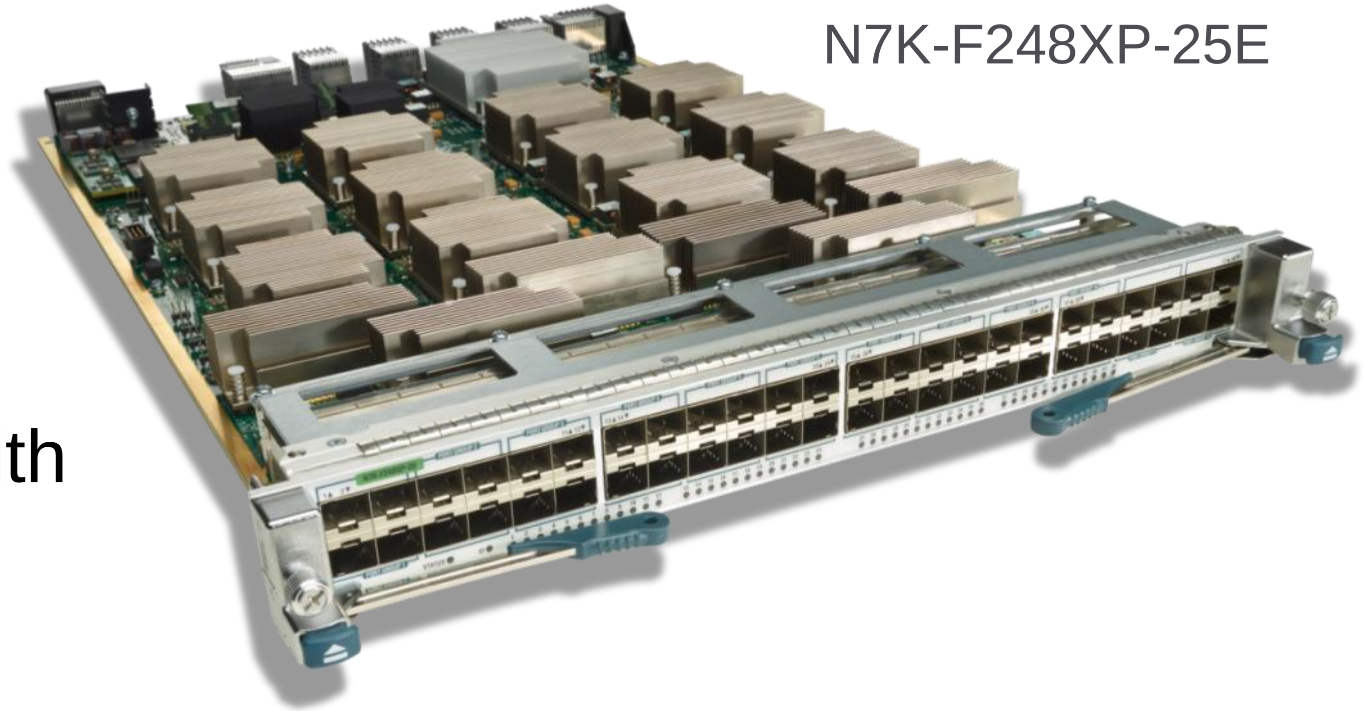
Supported in NX-OS release 6.1(2) and later

F Series

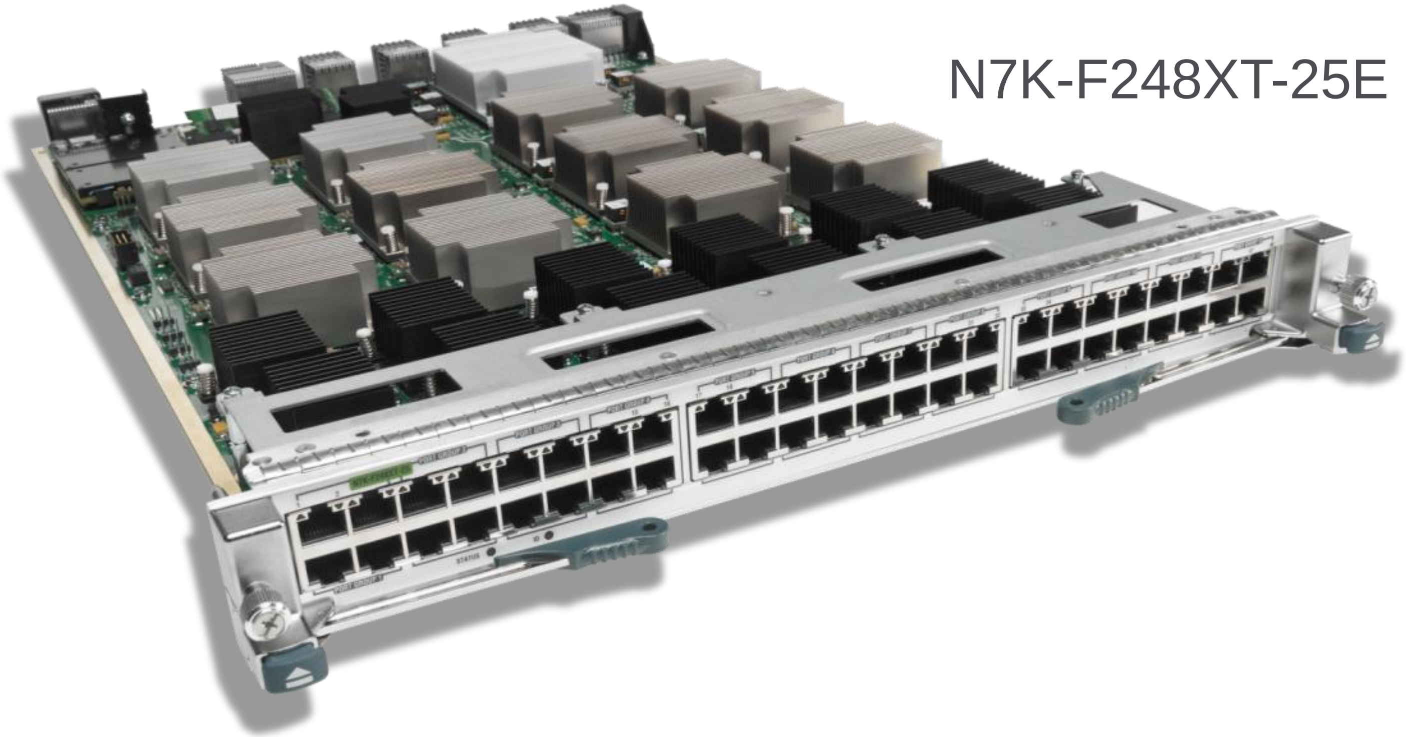
48-Port 1G/10G F2E I/O Modules (Fibre and Copper)

N7K-F248XP-25E / N7K-F248XT-25E

- Enhanced version of original F2 I/O module
- Fibre and copper version
- 480G full-duplex fabric connectivity
- Same basic SoC architecture as original F2 with some additional functionality



N7K-F248XP-25E



N7K-F248XT-25E

What’s Different in F2E?

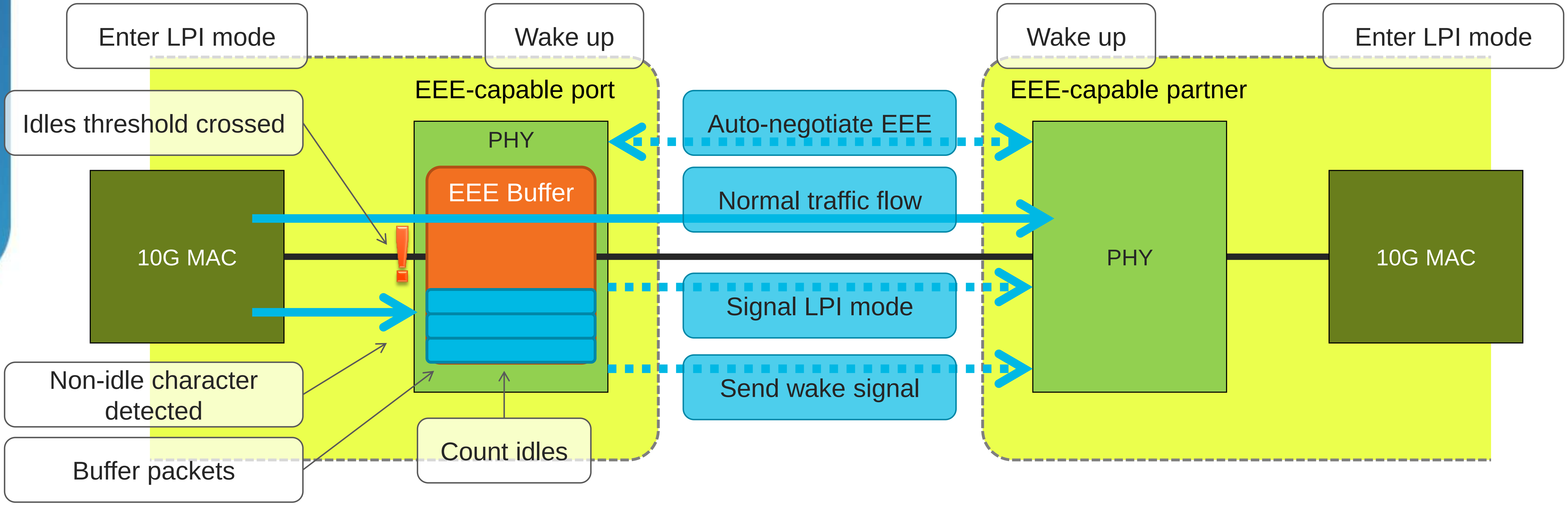
- Interoperability with M1/M2, in Layer 2 mode*
  - Proxy routing for inter-VLAN/L3 traffic
- LinkSec support*
  - Fibre version: 8 ports
  - Copper version: 48 ports
- Energy Efficient Ethernet (EEE) capability on F2E copper version
- FabricPath learning enhancements
  - No learning on broadcast frames

* Roadmap feature

© 2013 Cisco and/or its affiliates. All rights reserved.

Energy Efficient Ethernet (IEEE 802.3az)

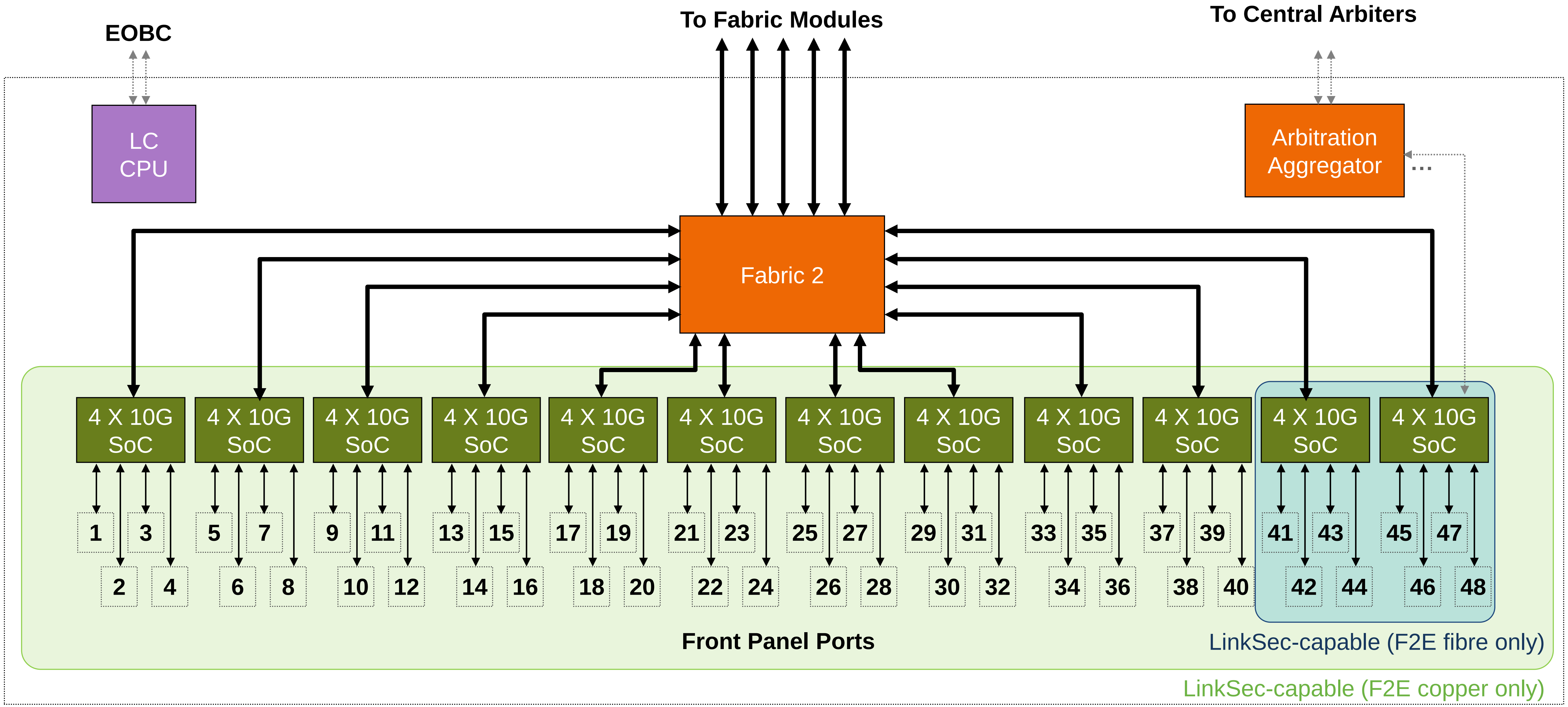
- IEEE standard for reducing power consumption during idle periods
- Auto-negotiated at Layer 1, like speed and duplex
- Introduces Low Power Idle (LPI) mode for Ethernet ports
  - Systems on both ends of link save power in LPI mode
  - Transparent to upper layer protocols



© 2013 Cisco and/or its affiliates. All rights reserved.

48-Port 1G/10G F2 / F2E I/O Module Architecture

N7K-F248XP-25 / N7K-F248XP-25E / N7K-F248XT-25

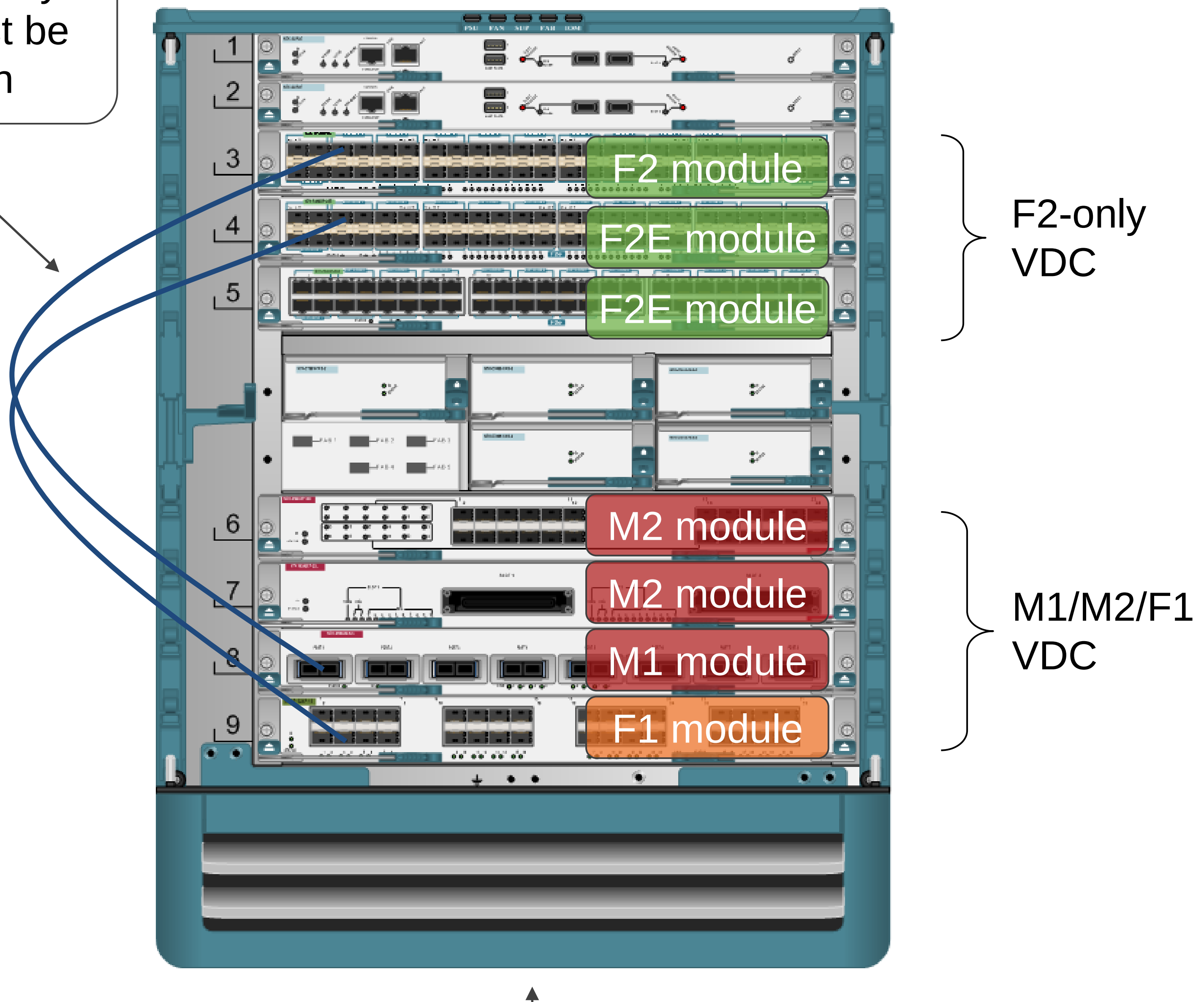


© 2013 Cisco and/or its affiliates. All rights reserved.

F2-Only VDC

- F2/F2E modules do **not** interoperate with other Nexus 7000 modules*
- Must deploy in an “F2 only” VDC
- Can be default VDC, or any other VDC
  - Use the **limit-resource module-type f2** VDC configuration command
- System with only F2 modules and empty configuration boots with F2-only default VDC automatically

Communication between F2-only VDC and M1/M2/F1 VDC must be through external connection



M1/M2/F1 modules can exist in same **chassis** as F2/F2E modules, but **not** in the same VDC

* F2E will interoperate in Layer 2 mode with M1/M2 in a future software release

Agenda

- Chassis Architecture
- Supervisor Engine and I/O Module Architecture
- Forwarding Engine Architecture
- Fabric Architecture
- I/O Module Queuing
- Layer 2 Forwarding
- IP Forwarding
- Classification
- NetFlow
- Conclusion

© 2013 Cisco and/or its affiliates. All rights reserved.

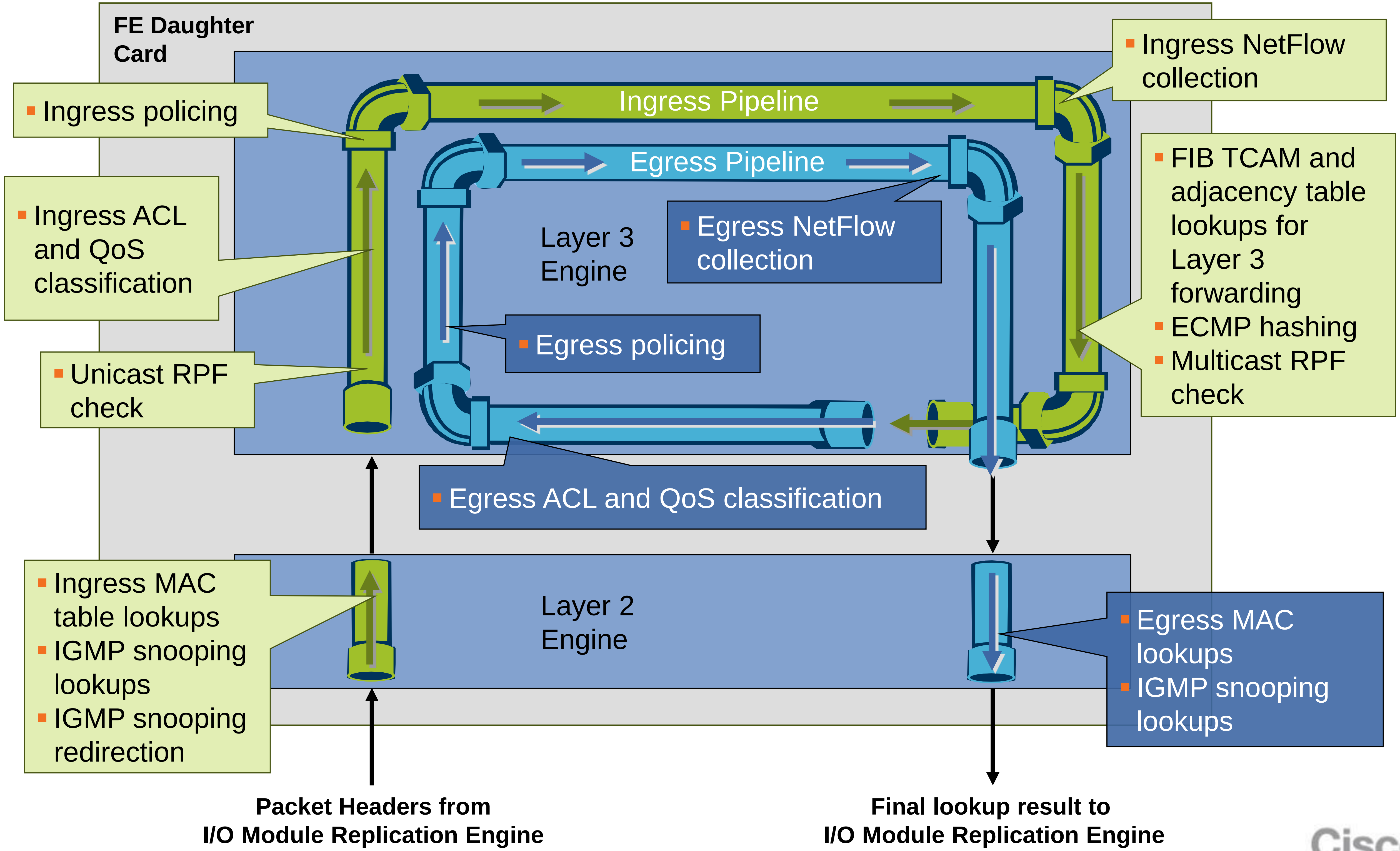
M-Series Forwarding Engine Hardware

- Hardware forwarding engine(s) integrated on every I/O module
- 60Mpps per forwarding engine Layer 2 bridging with hardware MAC learning
- 60Mpps per forwarding engine Layer 3 IPv4 and 30Mpps Layer 3 IPv6 unicast
- Layer 3 IPv4 and IPv6 multicast support (SM, SSM, bidir)
- MPLS
- OTV
- IGMP snooping
- RACL/VACL/PACL
- QoS remarking and policing policies
- Policy-based routing (PBR)
- Unicast RPF check and IP source guard
- Ingress and egress NetFlow (full and sampled)

Hardware Table	M-Series Modules without Scale License	M-Series Modules with Scale License
FIB TCAM	128K	**900K**
Classification TCAM (ACL/QoS)	64K	**128K**
MAC Address Table	128K	128K
NetFlow Table	512K	512K

© 2013 Cisco and/or its affiliates. All rights reserved.

M-Series Forwarding Engine Architecture



- Ingress policing
- Ingress ACL and QoS classification
- Unicast RPF check
- Ingress NetFlow collection
- FIB TCAM and adjacency table lookups for Layer 3 forwarding
- ECMP hashing
- Multicast RPF check
- Egress NetFlow collection
- Egress policing
- Egress ACL and QoS classification
- Ingress MAC table lookups
- IGMP snooping lookups
- IGMP snooping redirection
- Egress MAC lookups
- IGMP snooping lookups

Packet Headers from I/O Module Replication Engine

Final lookup result to I/O Module Replication Engine

F Series

F2/F2E Forwarding Engine Hardware

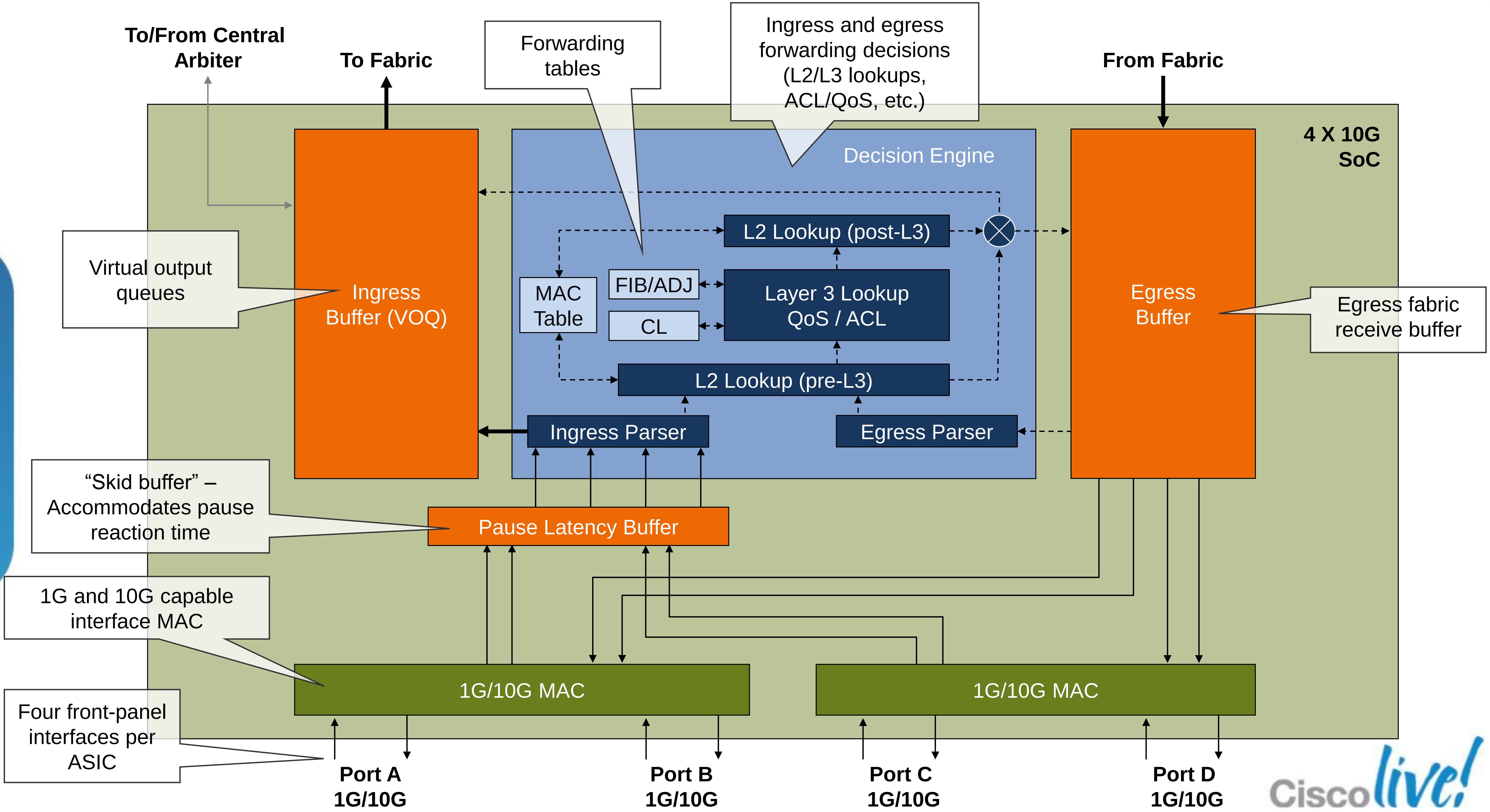
- Each SoC forwarding engine services 4 front-panel 10G ports (12 SoCs per module)
- 60Mpps per SoC Layer 2 bridging with hardware MAC learning
- 60Mpps per forwarding engine Layer 3 IPv4/ IPv6 unicast
- Layer 3 IPv4 and IPv6 multicast support (SM, SSM)
- IGMP snooping

- RACL/VACL/PACL
- QoS remarking and policing policies
- Policy-based routing (PBR)
- Unicast RPF check and IP source guard
- FabricPath forwarding
- Ingress sampled NetFlow
- FCoE

Hardware Table	Per F2 SoC	Per F2 Module
MAC Address Table	16K	256K*
FIB TCAM	32K IPv4/16K IPv6	32K IPv4/16K IPv6
Classification TCAM (ACL/QoS)	16K	192K*

* Assumes specific configuration to scale SoC resources

F2/F2E Forwarding Engine



© 2013 Cisco and/or its affiliates. All rights reserved.

Agenda

- Chassis Architecture
- Supervisor Engine and I/O Module Architecture
- Forwarding Engine Architecture
- Fabric Architecture
- I/O Module Queuing
- Layer 2 Forwarding
- IP Forwarding
- Classification
- NetFlow
- Conclusion

Crossbar Switch Fabric Modules

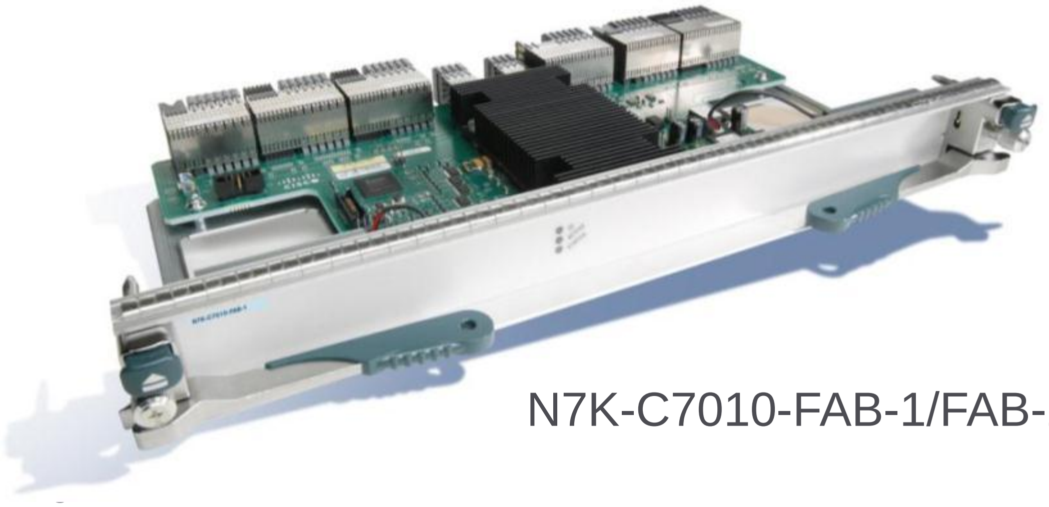
N7K-C7018-FAB-1/FAB-2

- Provide interconnection of I/O modules in Nexus 7009 / 7010 / 7018 chassis
- Each installed fabric increases available per-payload slot bandwidth
- Two fabric generations available – Fabric 1 and Fabric 2

Fabric Module	Supported Chassis	Supported I/O Modules	Per-fabric module bandwidth	Total bandwidth with 5 fabric modules
Fabric 1	7010 / 7018	All	46Gbps per slot	230Gbps per slot
Fabric 2	7009 / 7010 / 7018	All	110Gbps per slot	550Gbps per slot

- Different I/O modules leverage different amount of fabric bandwidth
- Access to fabric bandwidth controlled using QoS-aware central arbitration with VOQ

N7K-C7009-FAB-2



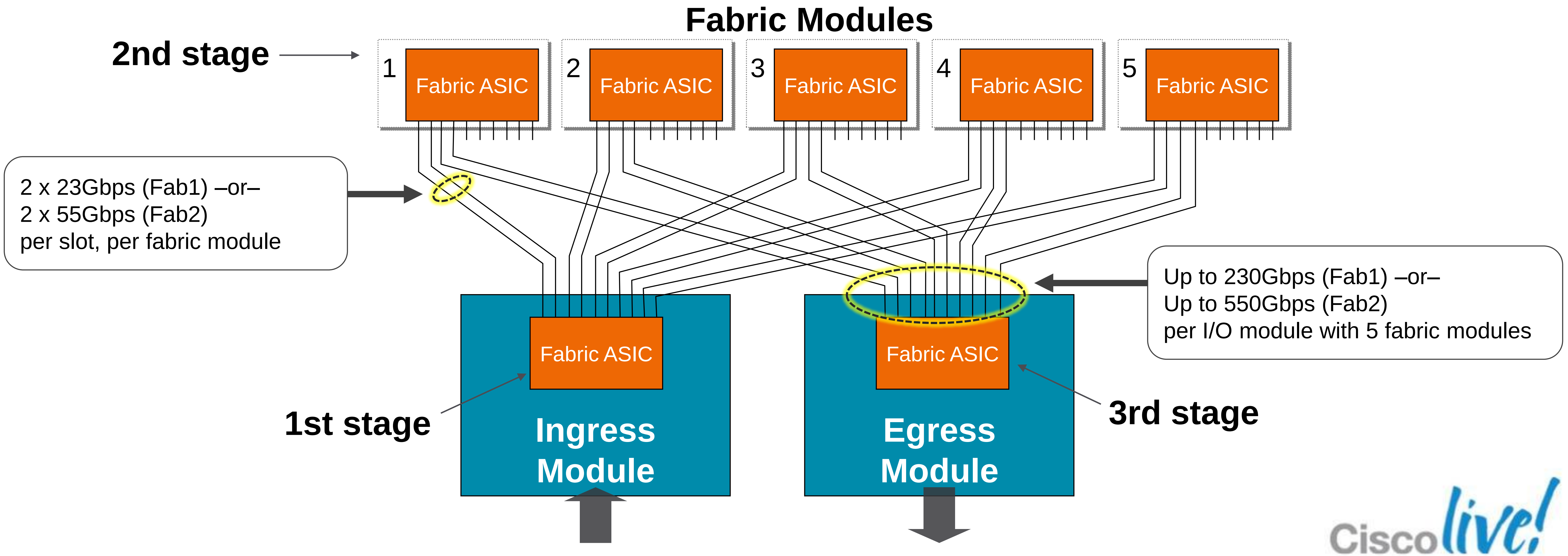
N7K-C7010-FAB-1/FAB-2

© 2013 Cisco and/or its affiliates

Multistage Crossbar

Nexus 7000 implements 3-stage crossbar switch fabric

- Stages 1 and 3 on I/O modules
- Stage 2 on fabric modules



2 x 23Gbps (Fab1) –or–
2 x 55Gbps (Fab2)
per slot, per fabric module

Up to 230Gbps (Fab1) –or–
Up to 550Gbps (Fab2)
per I/O module with 5 fabric modules

© 2013 Cisco and/or its affiliates. All rights reserved.

I/O Module Capacity – Fabric 1

230Gbps

per slot bandwidth

One fabric

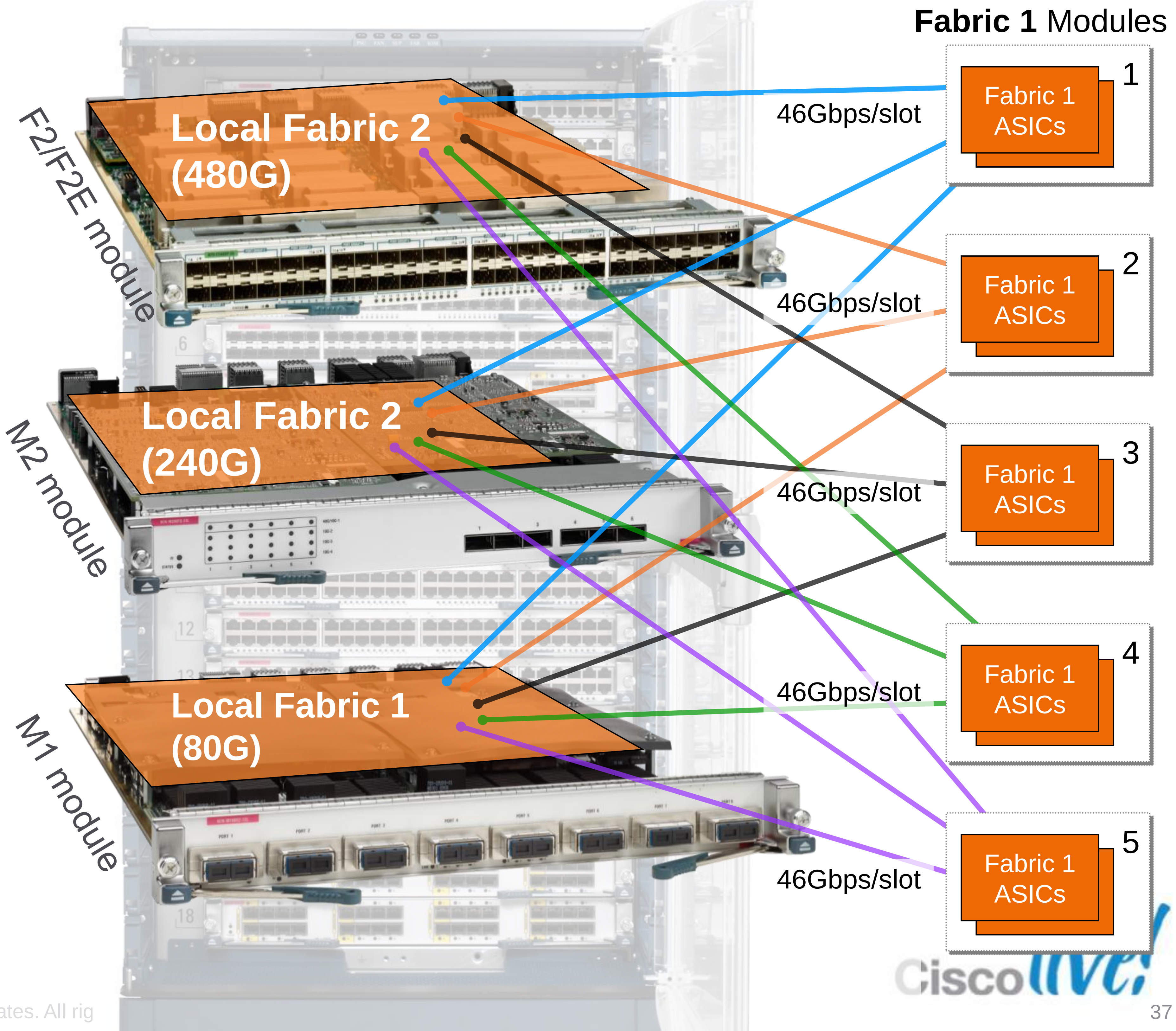
- **Any port** can pass traffic to **any other port** in system

Two fabrics

- 80G M1 module has full bandwidth

Five fabrics

- 240G M2 module limited to 230G per slot
- 480G F2/F2E module limited to 230G per slot



© 2013 Cisco and/or its affiliates. All rig

I/O Module Capacity – Fabric 2

Fab2 does **NOT** make Fab1-based modules faster!!

550Gbps

per slot bandwidth

One fabric

- **Any port** can pass traffic to **any other port** in system

Two fabrics

- 80G M1 module has full bandwidth

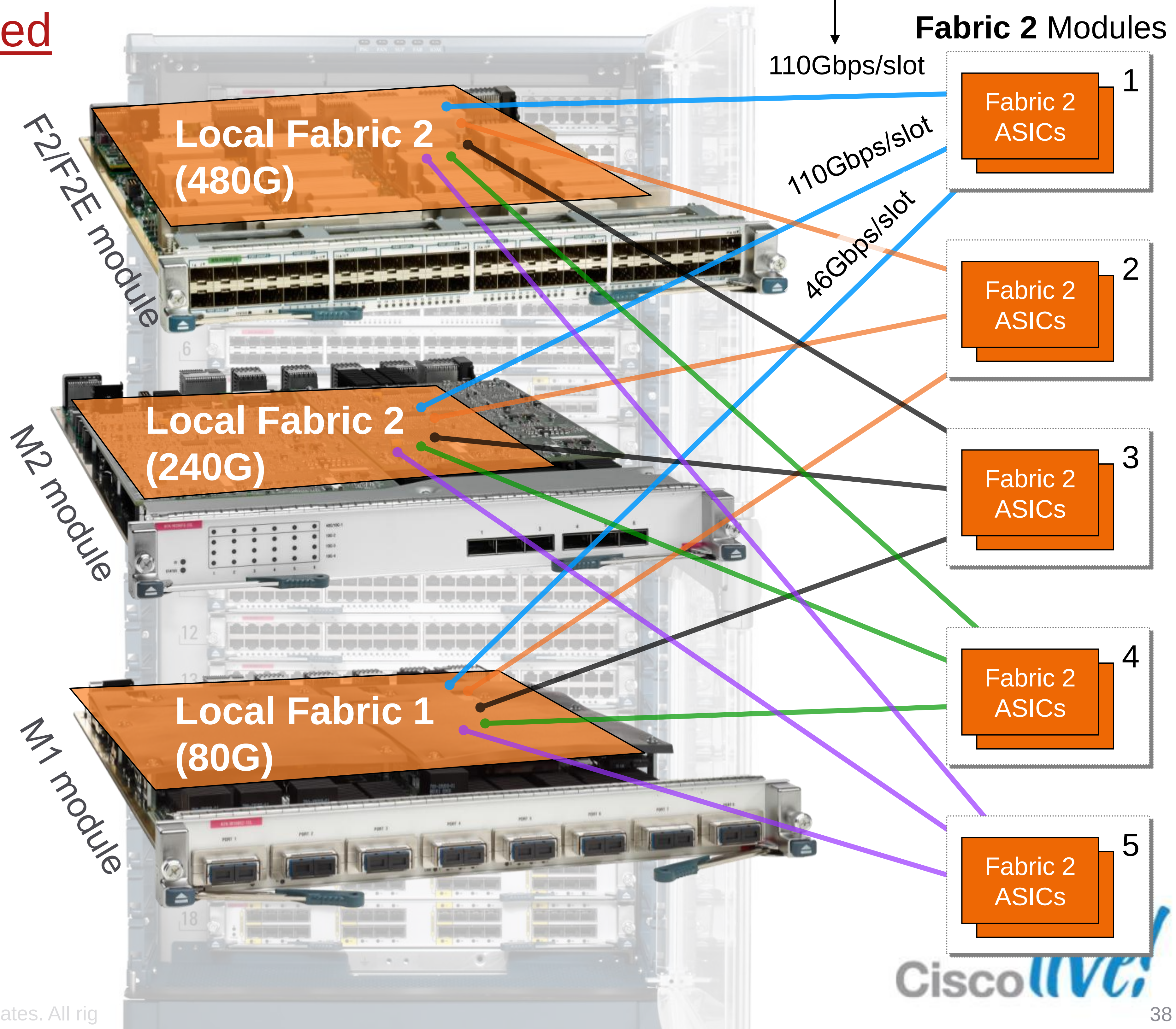
Three fabrics

- 240G M2 module has maximum bandwidth

Five fabrics

- 480G F2 module has maximum bandwidth

Fabric channels run at lowest common speed

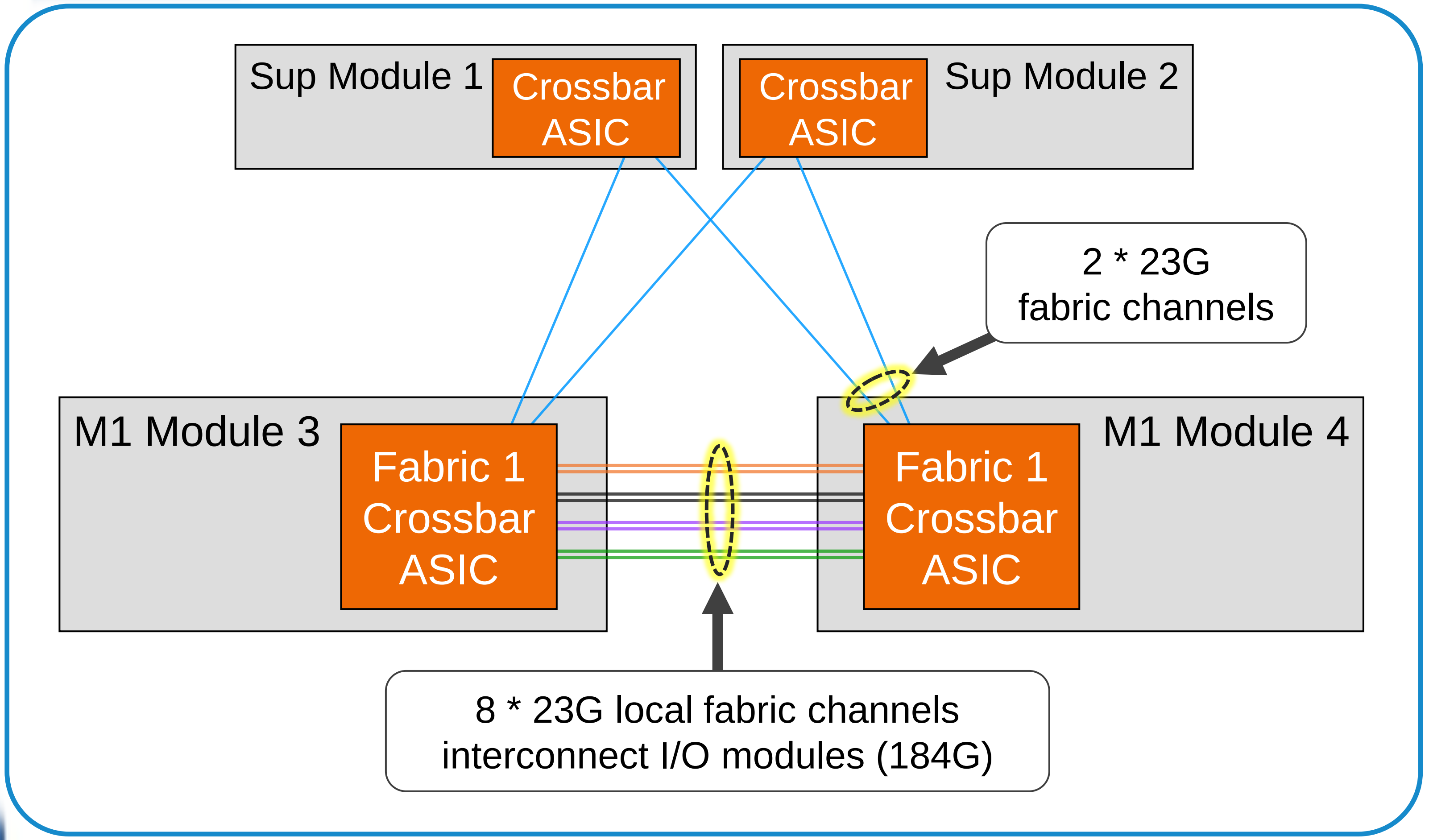


© 2013 Cisco and/or its affiliates. All rig

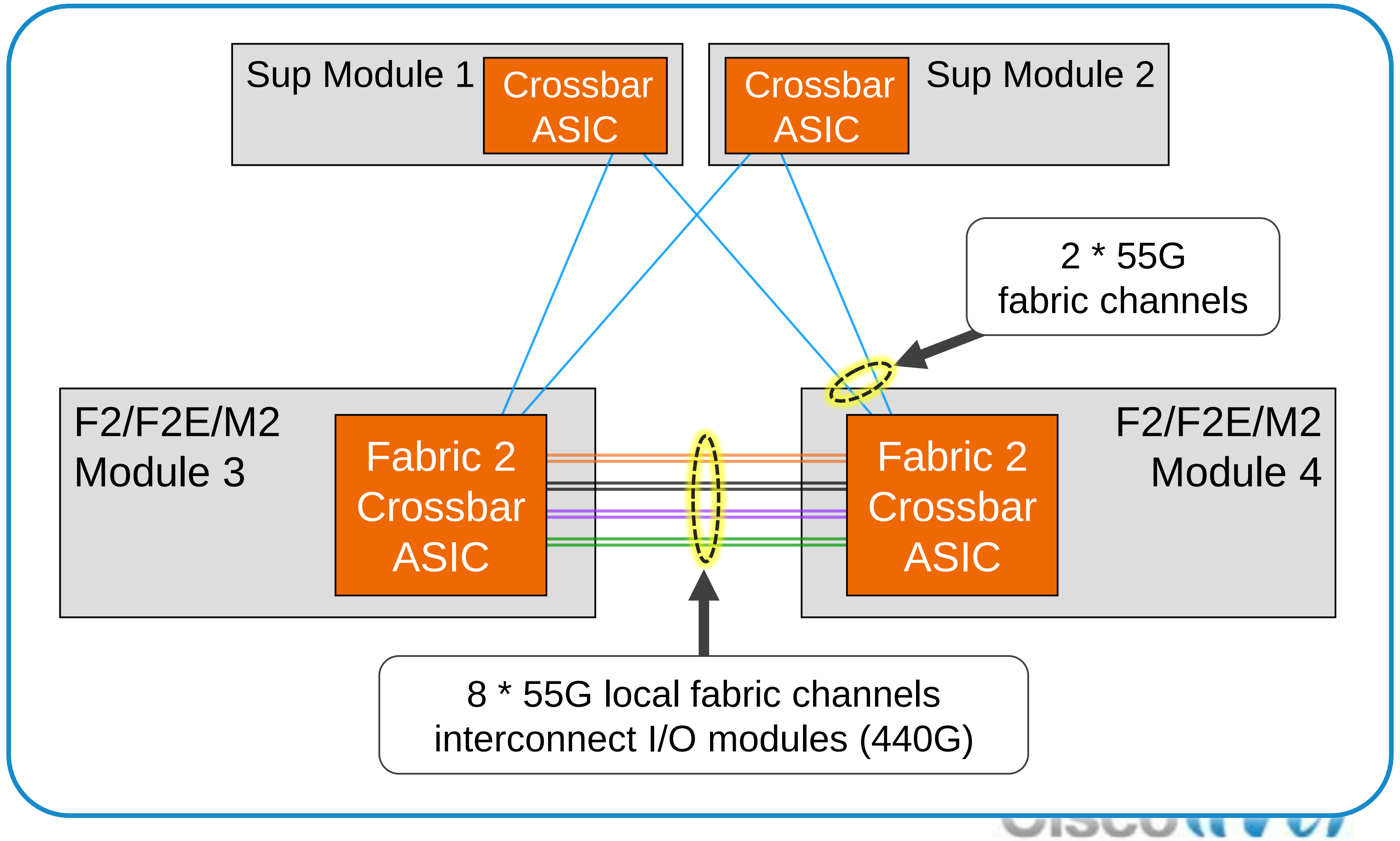
What About Nexus 7004?

- Nexus 7004 has no fabric modules
- I/O modules have local fabric with 10 available fabric channels
  - I/O modules connect "back-to-back" via 8 fabric channels
  - Two fabric channels "borrowed" to connect supervisor engines
- Available inter-module bandwidth dependent on installed module types

M1 Modules in Nexus 7004



F2/F2E/M2 Modules in Nexus 7004



© 2013 Cisco and/or its affiliates. All rights reserved.

Arbitration, VOQ and Crossbar Fabric

- Arbitration, VOQ, and fabric combine to provide all necessary infrastructure for packet transport inside switch

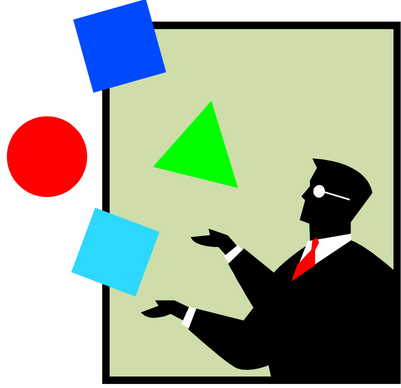
- **Central arbitration** – Controls scheduling of traffic into fabric based on fairness, priority, and bandwidth availability at egress ports
- **Virtual Output Queues (VOQs)** – Provide buffering and queuing for ingress-buffered switch architecture
- **Crossbar fabric** – Provides dedicated, high-bandwidth interconnects between ingress and egress I/O modules

Central Arbitration

- Access to fabric for unicast traffic controlled using central arbitration
  - Ensures fair access to available bandwidth on each egress port
  - Can provide no-drop service for some traffic classes
- Arbiter ASIC on Supervisor Engine provides central arbitration via dedicated arbitration path to every module
- Arbitration performed on per-destination, per-priority basis
  - Ensures high priority traffic takes precedence over low priority traffic
- For multidestination traffic, no central arbitration
  - Ingress broadcast, multicast, unknown unicast frames sent unarbitrated

© 2013 Cisco and/or its affiliates. All rights reserved.

Virtual Output Queues (VOQs)



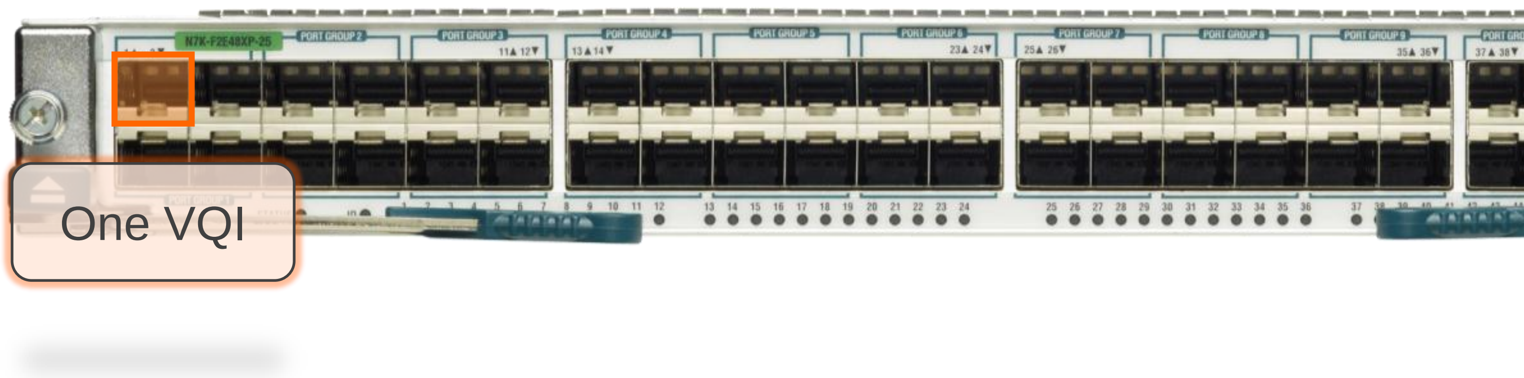
- VOQs at ingress to fabric provide buffering and queuing for egress destinations reached through the fabric
- Queuing of traffic entering fabric based on destination port (VQI) and packet priority
  - Four levels of priority per destination
- VOQs prevent congested egress ports from blocking ingress traffic destined to other ports
  - Provide independent scheduling for individual egress destinations

© 2013 Cisco and/or its affiliates. All rights reserved.

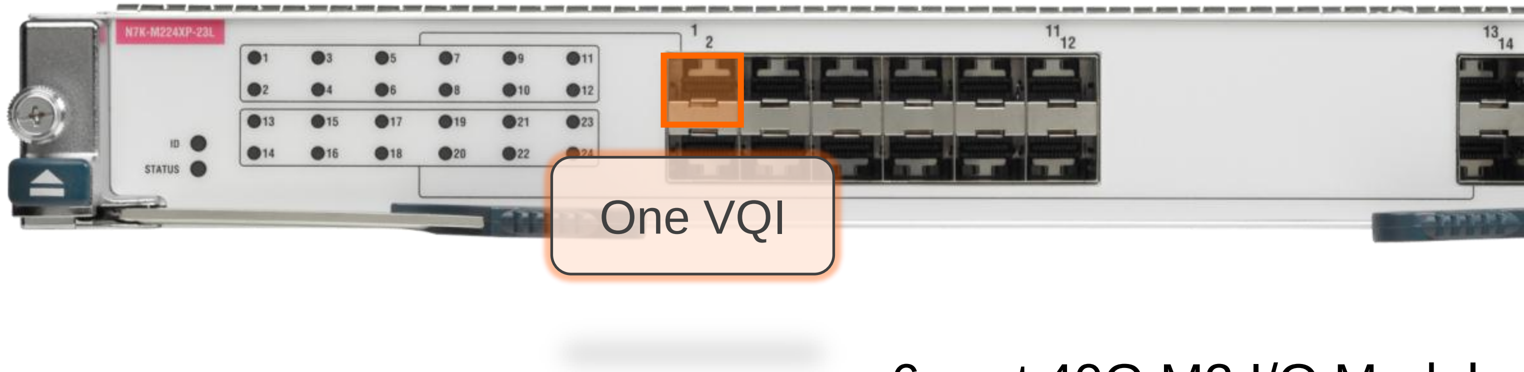
VOQ Destinations (VQIs)

- Each egress interface has one or more associated “Virtual Queuing Indexes” (VQIs) or “VOQ Destinations”
- Each VQI has four priority levels / classes
- For 1G / 10G interfaces:
  - One VQI for each 1G or 10G port
- For 40G interfaces:
  - Four VQIs for each 40G port –or–
  - One VQI for each 10G breakout port
- For 100G interfaces:
  - Ten VQIs for each 100G port –or–
  - Four VQIs for each 40G breakout port –or–
  - One VQI for each 10G breakout port

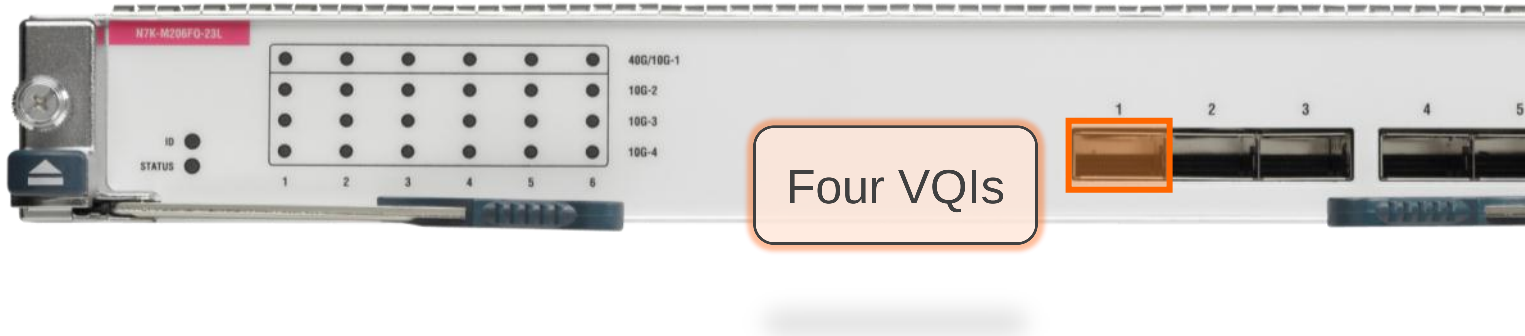
48-port 1G/10G F2/F2E I/O Module



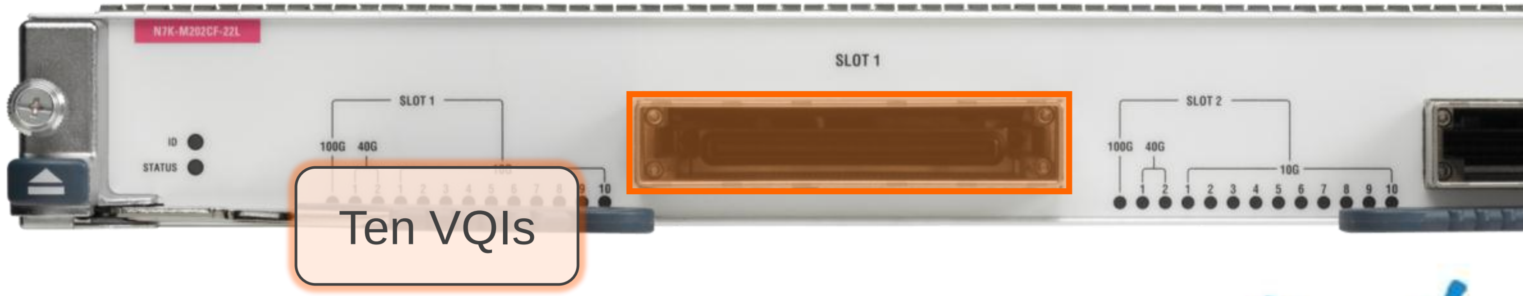
24-port 10G M2 I/O Module



6-port 40G M2 I/O Module



2-port 100G M2 I/O Module



© 2013 Cisco and/or its affiliates. All rights reserved.

Agenda

- Chassis Architecture
- Supervisor Engine and I/O Module Architecture
- Forwarding Engine Architecture
- Fabric Architecture
- I/O Module Queuing
- Layer 2 Forwarding
- IP Forwarding
- Classification
- NetFlow
- Conclusion

Buffering, Queuing, and Scheduling

- Buffering – storing packets in memory
  - Needed to absorb bursts, manage congestion
- Queuing – buffering packets according to traffic class
  - Provides dedicated buffer for packets of different priority
- Scheduling – controlling the order of transmission of buffered packets
  - Ensures preferential treatment for packets of higher priority and fair treatment for packets of equal priority

- Nexus 7000 uses queuing policies and network-QoS policies to define buffering, queuing, and scheduling behaviour
- Default queuing and network-QoS policies always in effect in absence of any user configuration

I/O Module Buffering Models

- Buffering model varies by I/O module family
  - **M-series modules**: hybrid model combining ingress VOQ-buffered architecture with egress port-buffered architecture
  - **F-series modules**: pure ingress VOQ-buffered architecture
- All configuration through Modular QoS CLI (MQC)
  - Queuing parameters applied using class-maps/policy-maps/service-policies

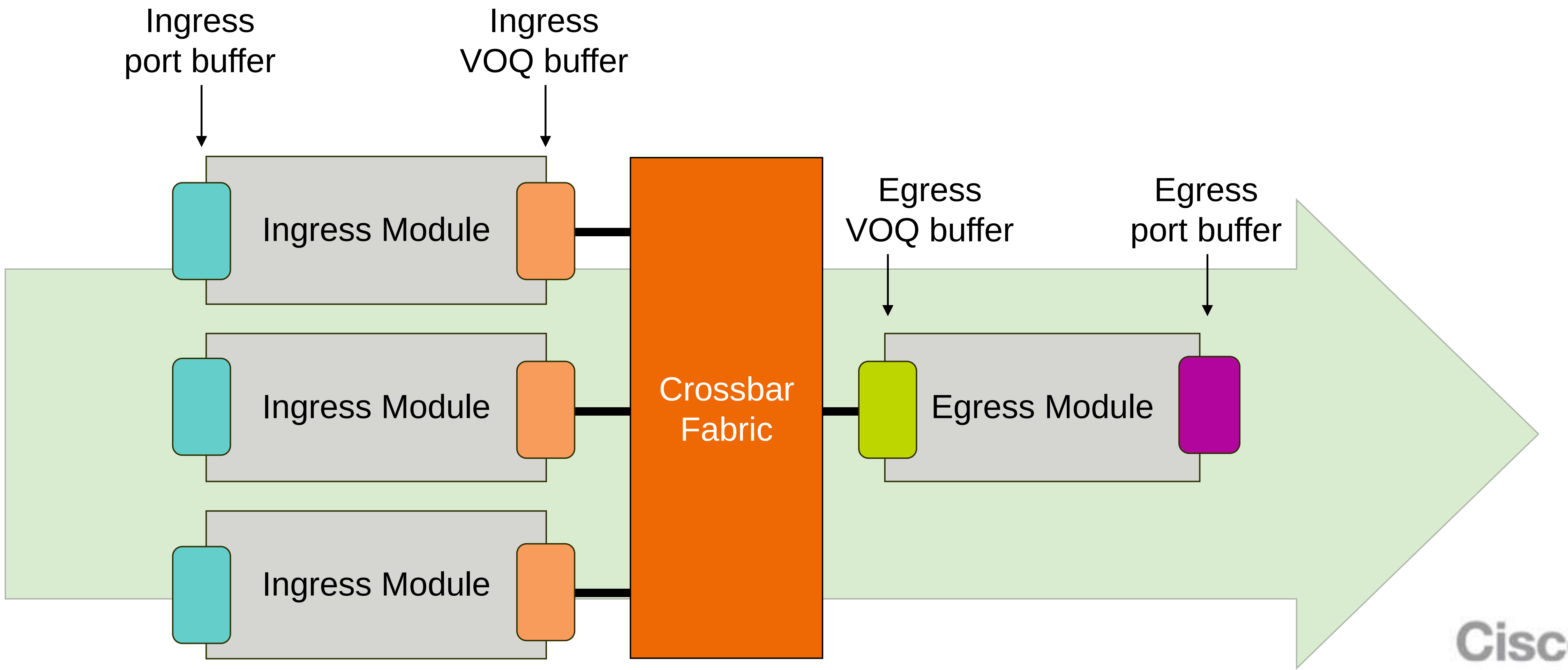
© 2013 Cisco and/or its affiliates. All rights reserved.

M Series

Hybrid Ingress/Egress Buffered Model

M-Series I/O Modules

- 🟦 Ingress port buffer – Manages congestion in ingress forwarding/replication engines only
- 🟧 Ingress VOQ buffer – Manages congestion toward egress destinations (VQIs) over fabric
- 🟩 Egress VOQ buffer – Receives frames from fabric; also buffers multidestination frames
- 🟪 Egress port buffer – Manages congestion at egress interface



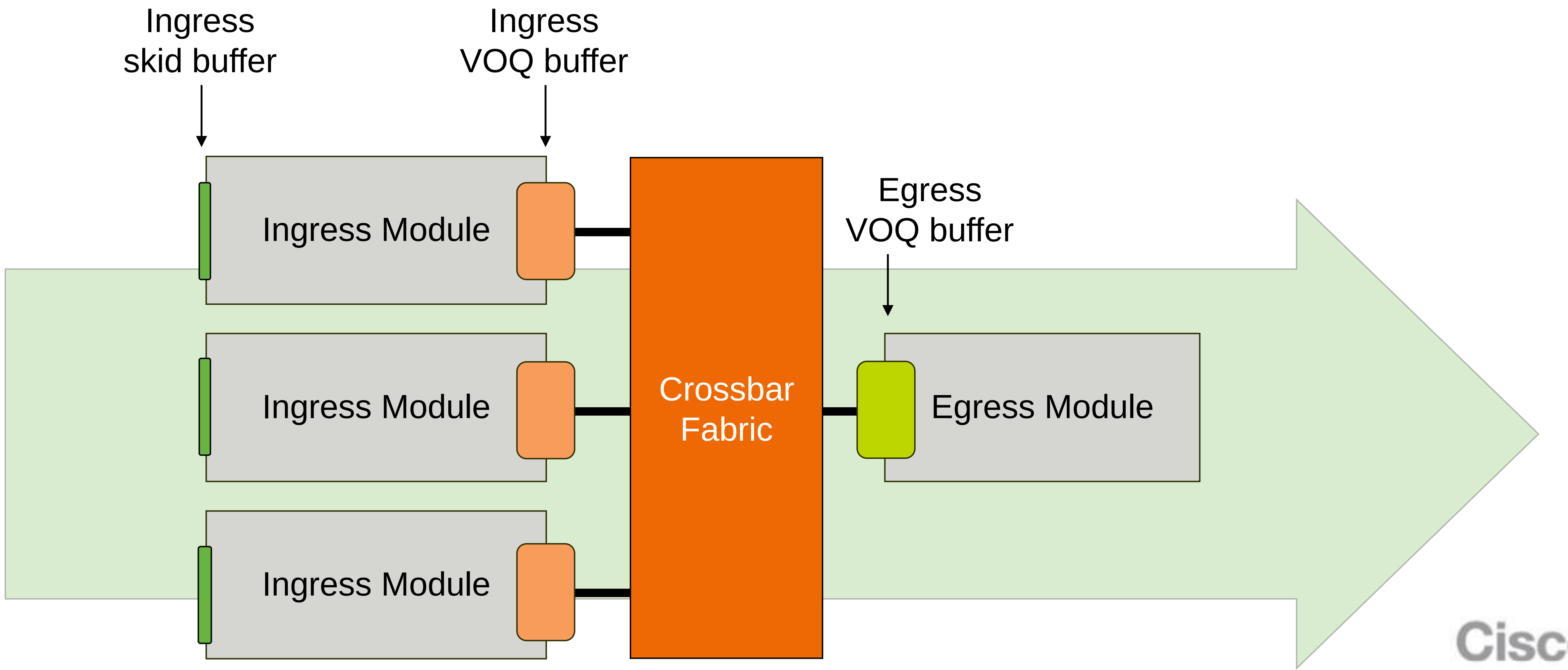
© 2013 Cisco and/or its affiliates. All rights reserved.

F Series

Ingress Buffered Model

F-Series I/O Modules

- 🟩 Ingress “skid” buffer – Absorbs packets in flight after external flow control asserted
- 🟧 Ingress VOQ buffer – Manages congestion toward egress destinations (VQIs) over fabric
- 🟨 Egress VOQ buffer – Receives frames from fabric; also buffers multidestination frames

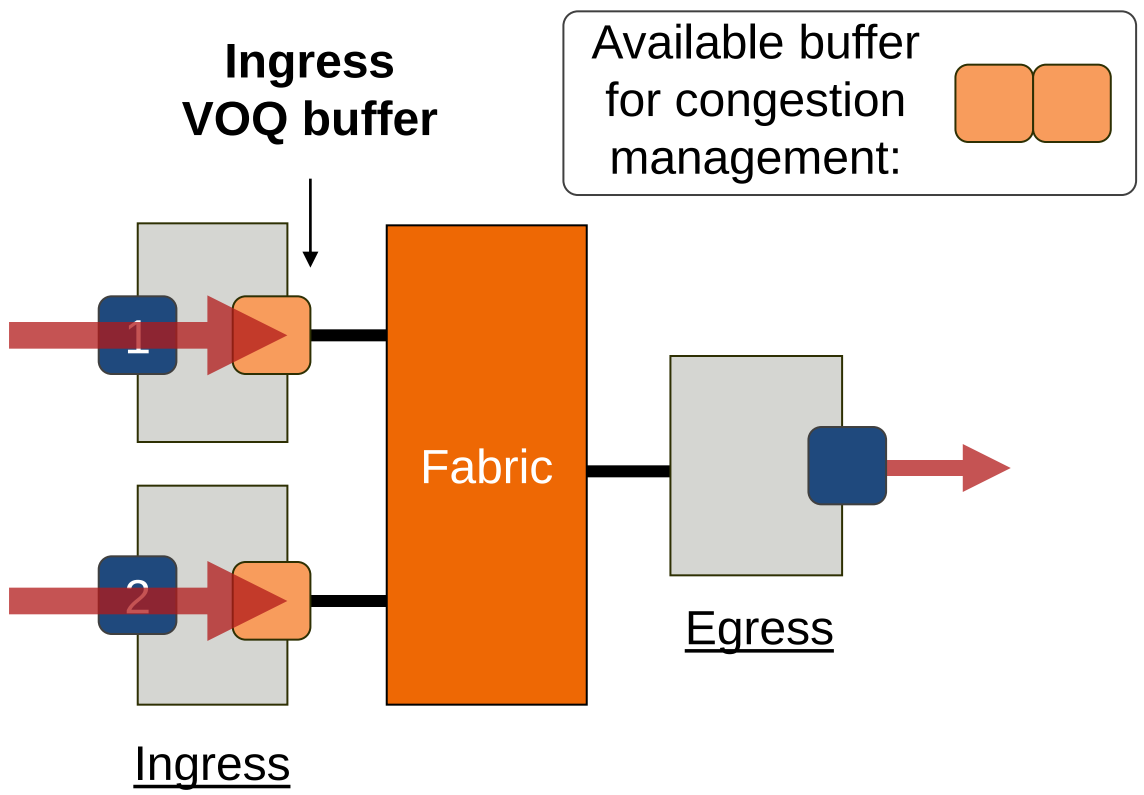


© 2013 Cisco and/or its affiliates. All rights reserved.

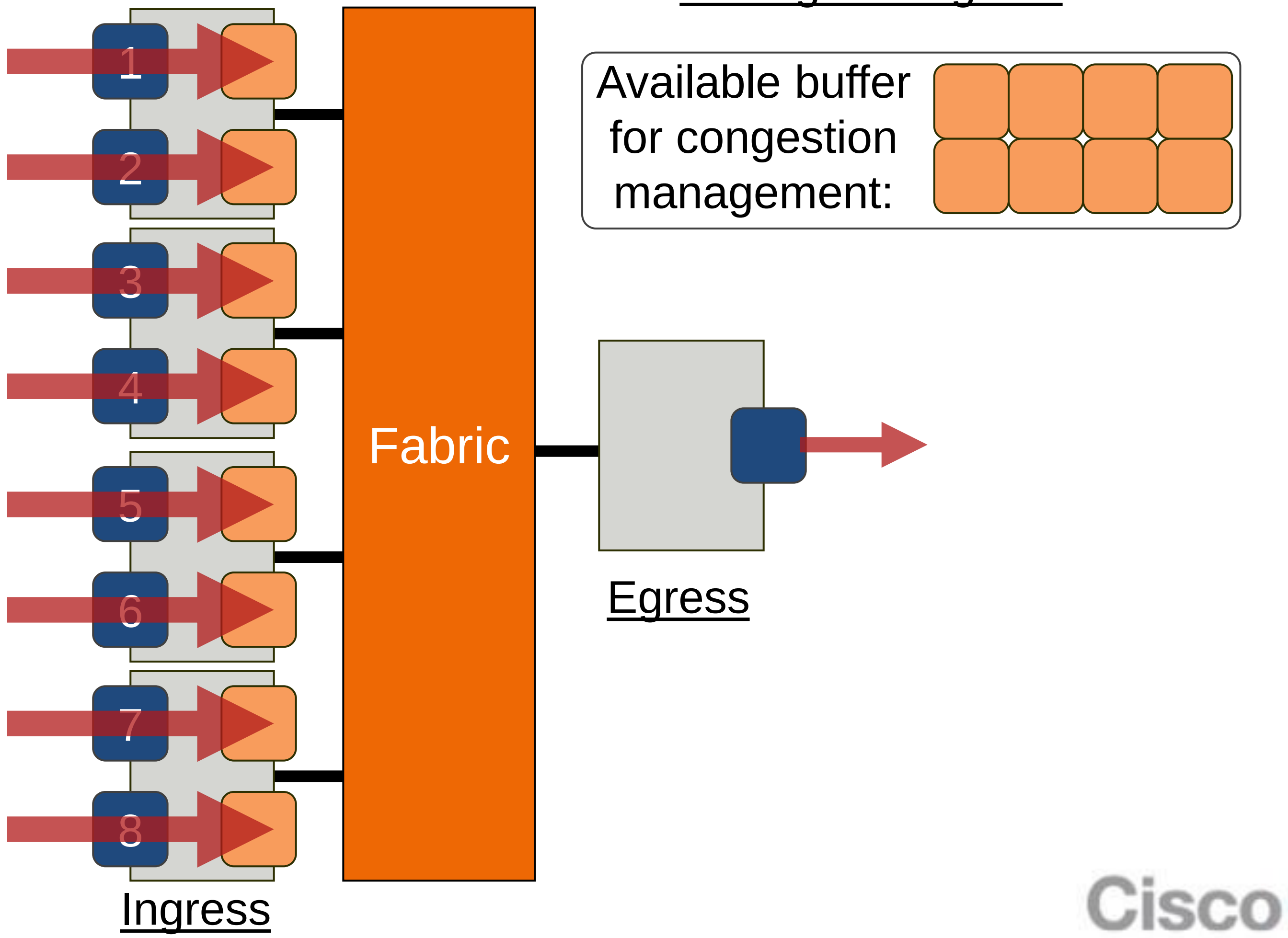
Distributed Buffer Pool

- Ingress-buffered architecture implements large, distributed buffer pool to absorb congestion
- Absorbs congestion at all ingress ports contributing to congestion, leveraging all per-port ingress buffer
- Excess traffic does not consume fabric bandwidth, only to be dropped at egress port

2:1 Ingress:Egress



8:1 Ingress:Egress



© 2013 Cisco and/or its affiliates. All rights reserved.

Agenda

- Chassis Architecture
- Supervisor Engine and I/O Module Architecture
- Forwarding Engine Architecture
- Fabric Architecture
- I/O Module Queuing
- Layer 2 Forwarding
- IP Forwarding
- Classification
- NetFlow
- Conclusion

© 2013 Cisco and/or its affiliates. All rights reserved.

Layer 2 Forwarding

- Layer 2 forwarding – traffic steering based on destination MAC address
- Hardware MAC learning
  - CPU not directly involved in learning
- Forwarding engine(s) on each module have copy of MAC table
  - New learns communicated to other forwarding engines via hardware “flood to fabric” mechanism
  - Software process ensures continuous MAC table sync
- Spanning tree (PVRST or MST), Virtual Port Channel (VPC), or FabricPath ensures loop-free Layer 2 topology

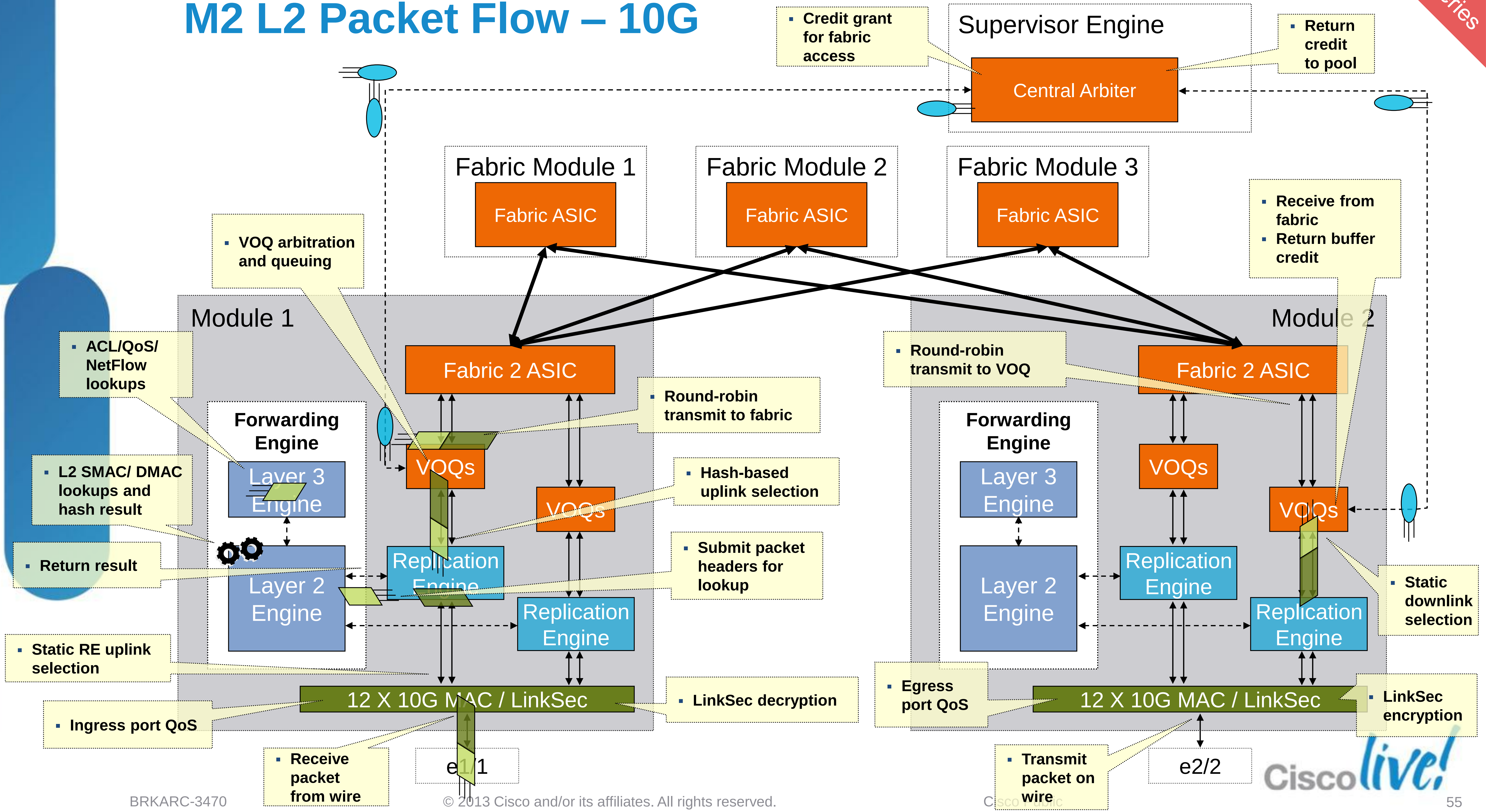
Hardware Layer 2 Forwarding Process

MAC table lookup drives Layer 2 forwarding

- Source MAC and destination MAC lookups performed for each frame, based on {VLAN,MAC} pairs
- Source MAC lookup drives new learns and refreshes aging timers
- Destination MAC lookup dictates outgoing switchport

HDR = Packet Headers DATA = Packet Data CTRL = Internal Signalling

M2 L2 Packet Flow – 10G



- Credit grant for fabric access

- Return credit to pool

- Receive from fabric
- Return buffer credit

- VOQ arbitration and queuing

- ACL/QoS/ NetFlow lookups

- L2 SMAC/ DMAC lookups and hash result

- Return result

- Static RE uplink selection

- Ingress port QoS

- Receive packet from wire

- Round-robin transmit to fabric

- Hash-based uplink selection

- Submit packet headers for lookup

- LinkSec decryption

- Round-robin transmit to VOQ

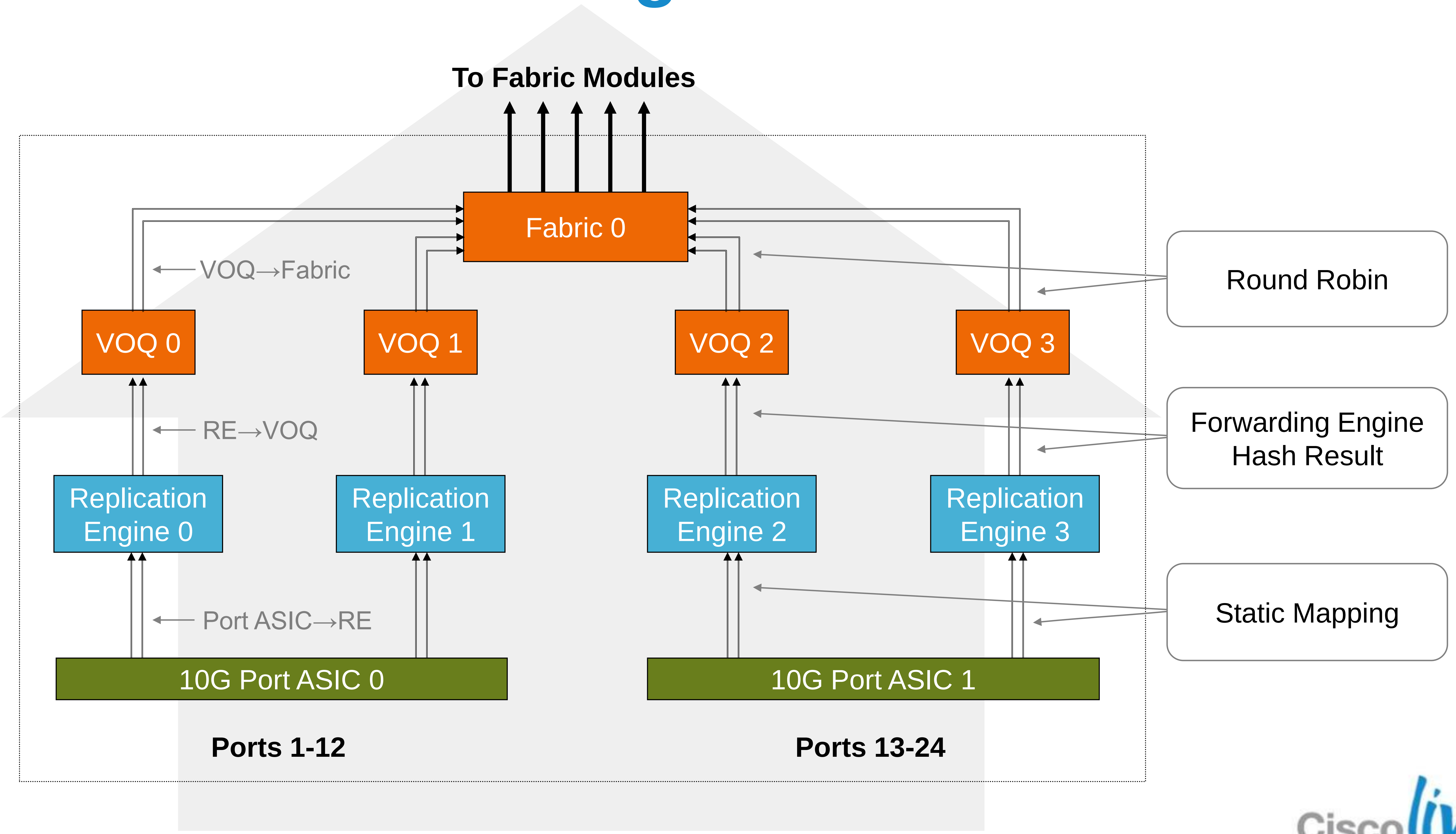
- Static downlink selection

- Egress port QoS

- LinkSec encryption

- Transmit packet on wire

10G M2 Module Ingress Path

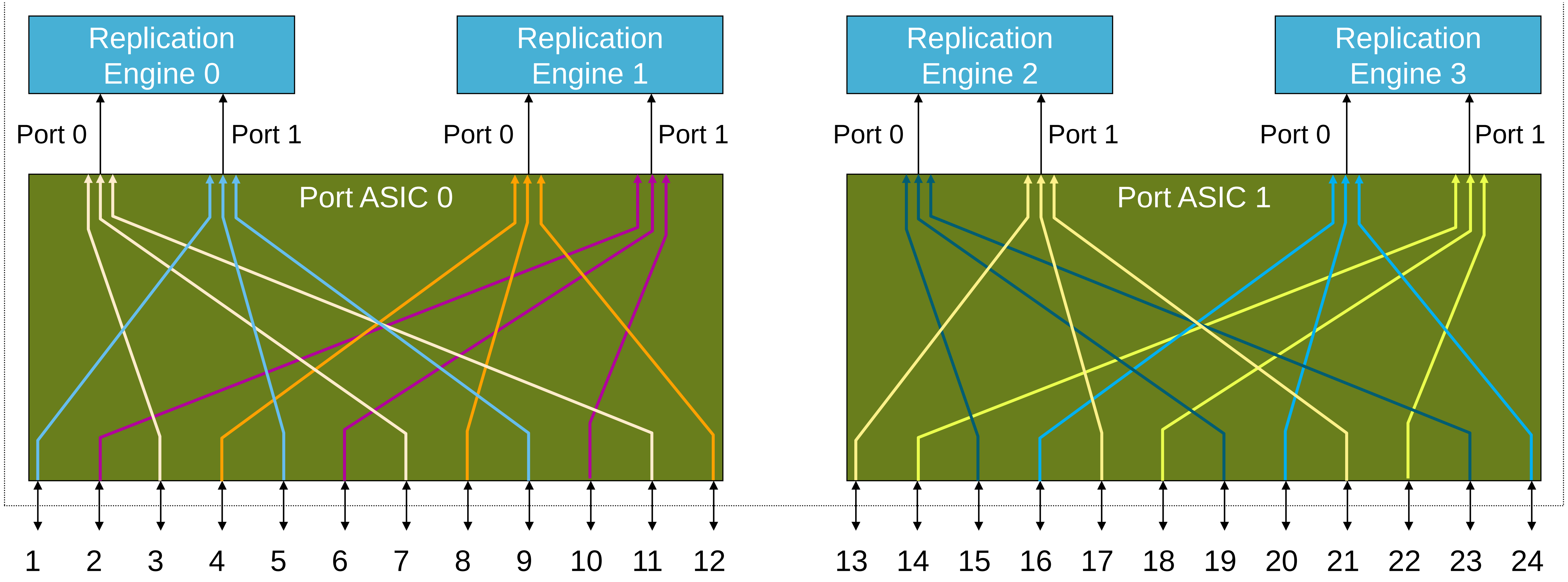


© 2013 Cisco and/or its affiliates. All rights reserved.

M Series

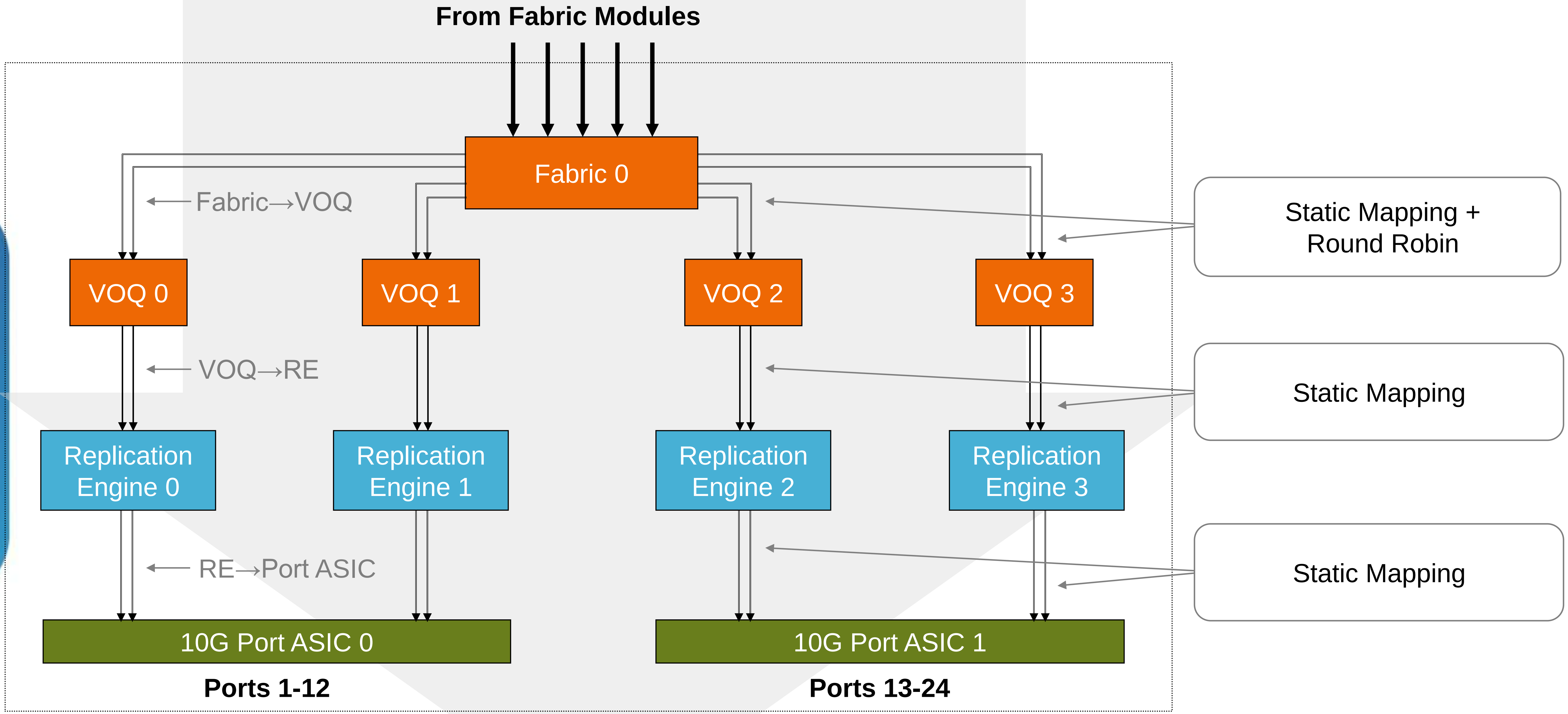
Replication Engine Selection on Ingress – 10G M2 Module

- Front-panel ports statically mapped to replication engine uplinks



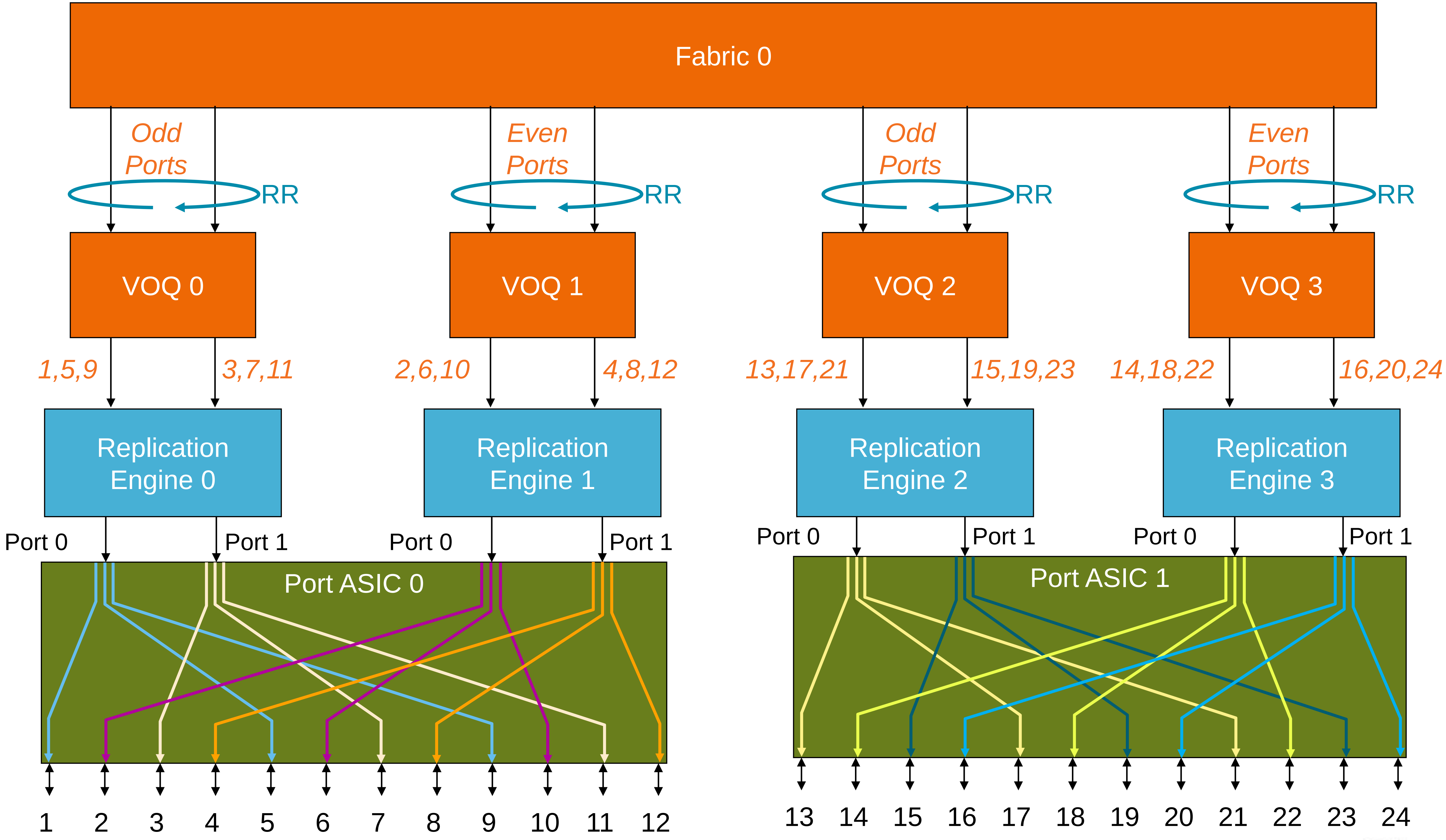
© 2013 Cisco and/or its affiliates. All rights reserved.

10G M2 Module Egress Path



© 2013 Cisco and/or its affiliates. All rights reserved.

10G M2 Module Egress VQI Mapping

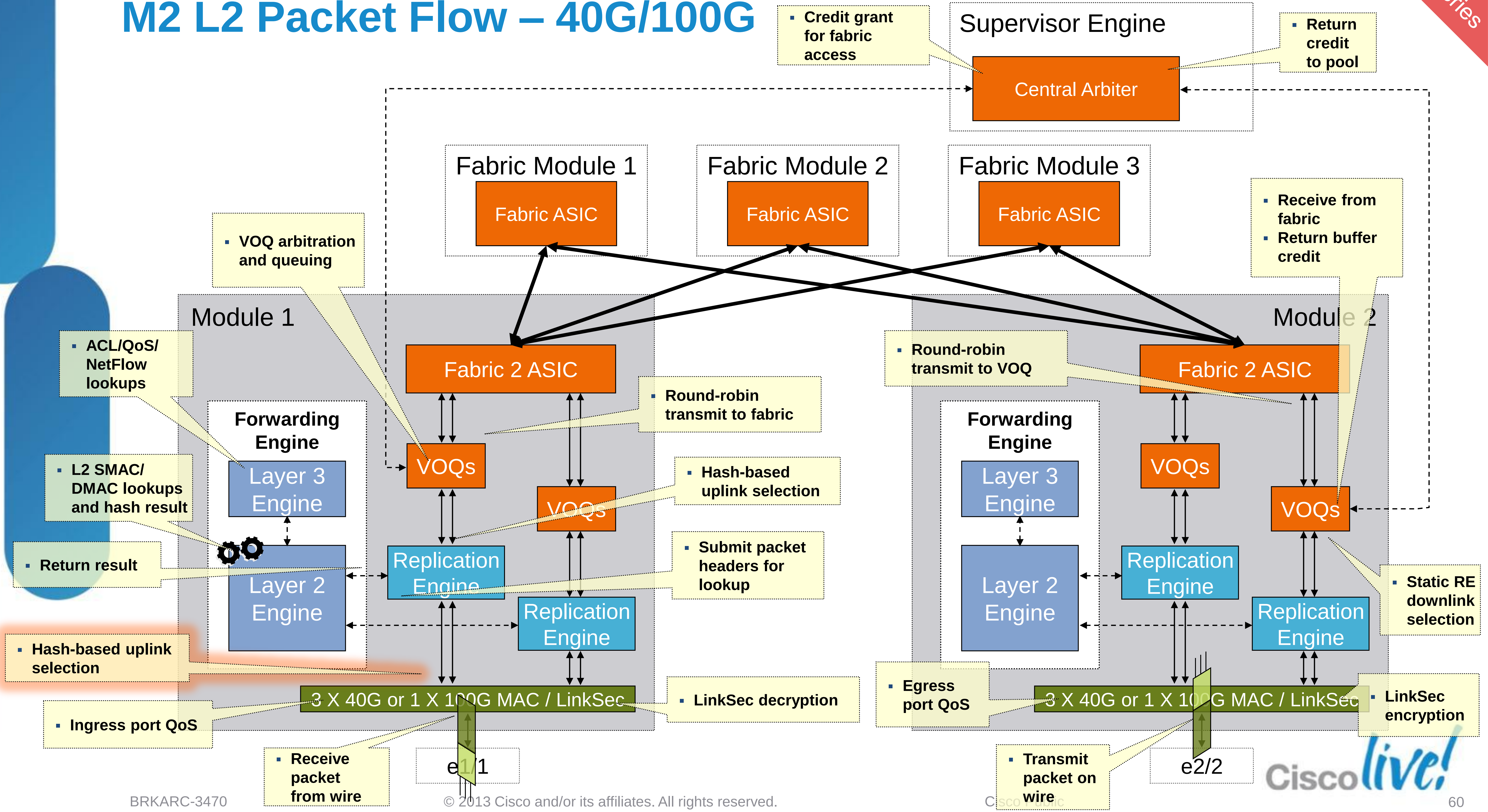


© 2013 Cisco and/or its affiliates. All rights reserved.

HDR = Packet Headers DATA = Packet Data CTRL = Internal Signalling

M Series

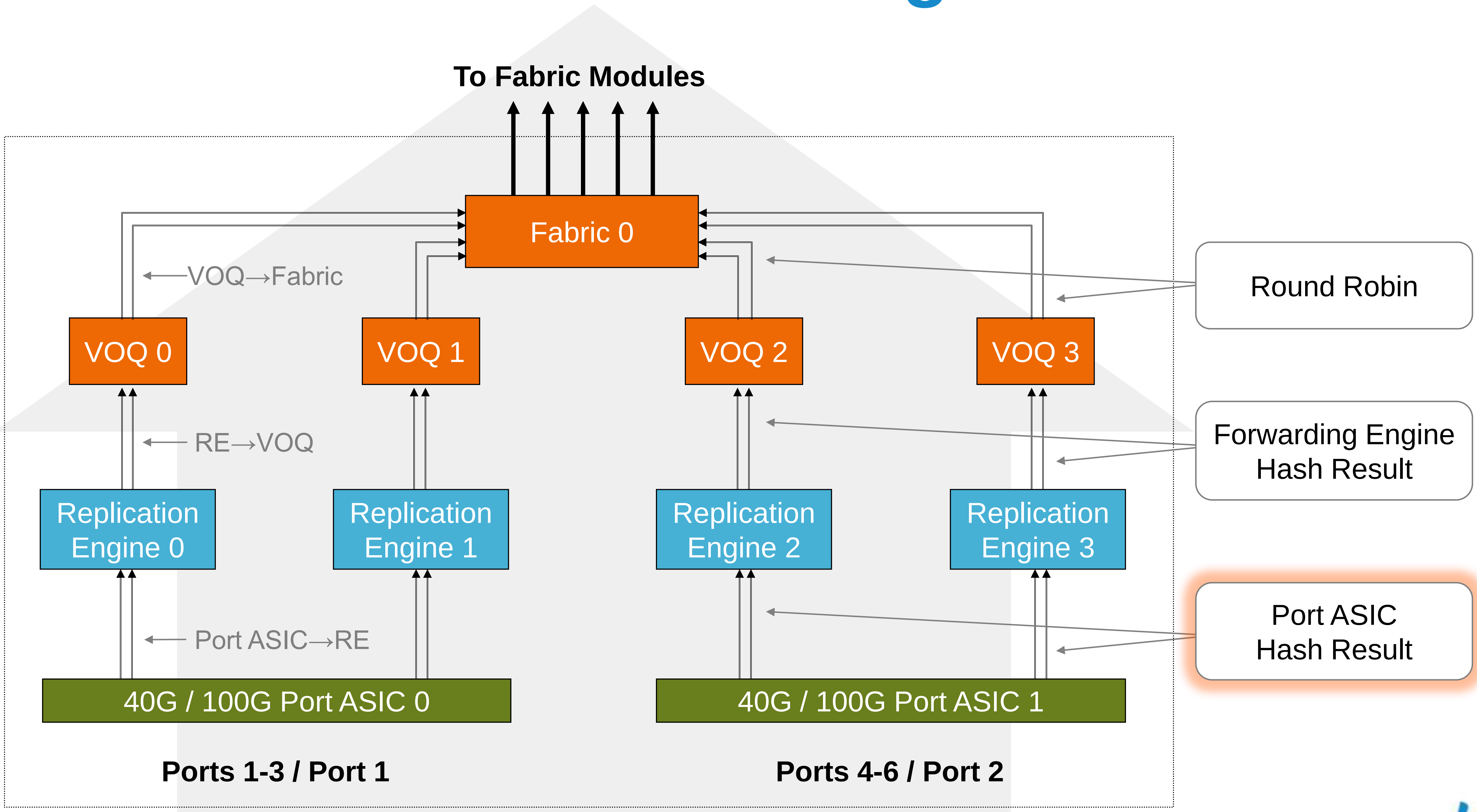
M2 L2 Packet Flow – 40G/100G



- **Credit grant for fabric access**
- **Return credit to pool**
- **Receive from fabric**
- **Return buffer credit**
- **VOQ arbitration and queuing**
- **ACL/QoS/ NetFlow lookups**
- **L2 SMAC/ DMAC lookups and hash result**
- **Return result**
- **Hash-based uplink selection**
- **Ingress port QoS**
- **Receive packet from wire**
- **Round-robin transmit to fabric**
- **Hash-based uplink selection**
- **Submit packet headers for lookup**
- **LinkSec decryption**
- **Round-robin transmit to VOQ**
- **Static RE downlink selection**
- **Egress port QoS**
- **LinkSec encryption**
- **Transmit packet on wire**

© 2013 Cisco and/or its affiliates. All rights reserved.

40G / 100G M2 Module Ingress Path

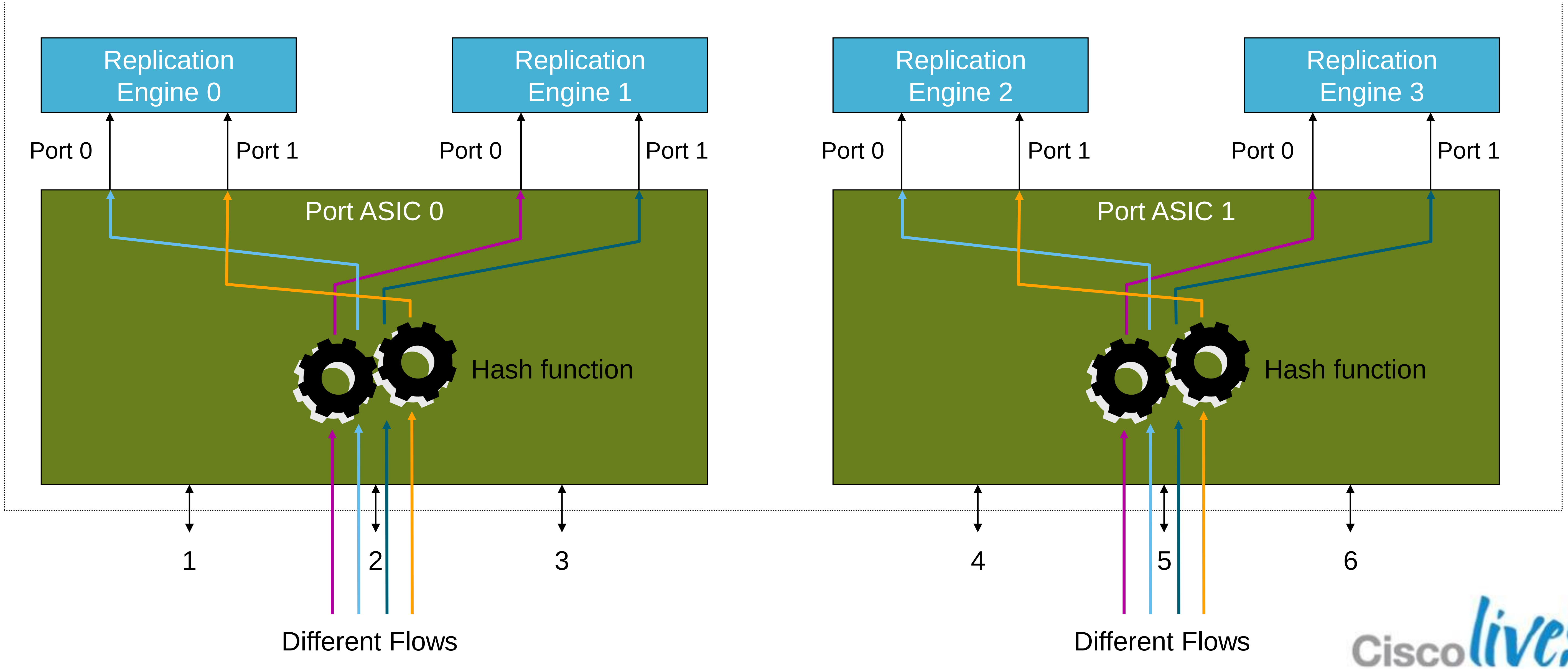


© 2013 Cisco and/or its affiliates. All rights reserved.

Replication Engine Selection on Ingress

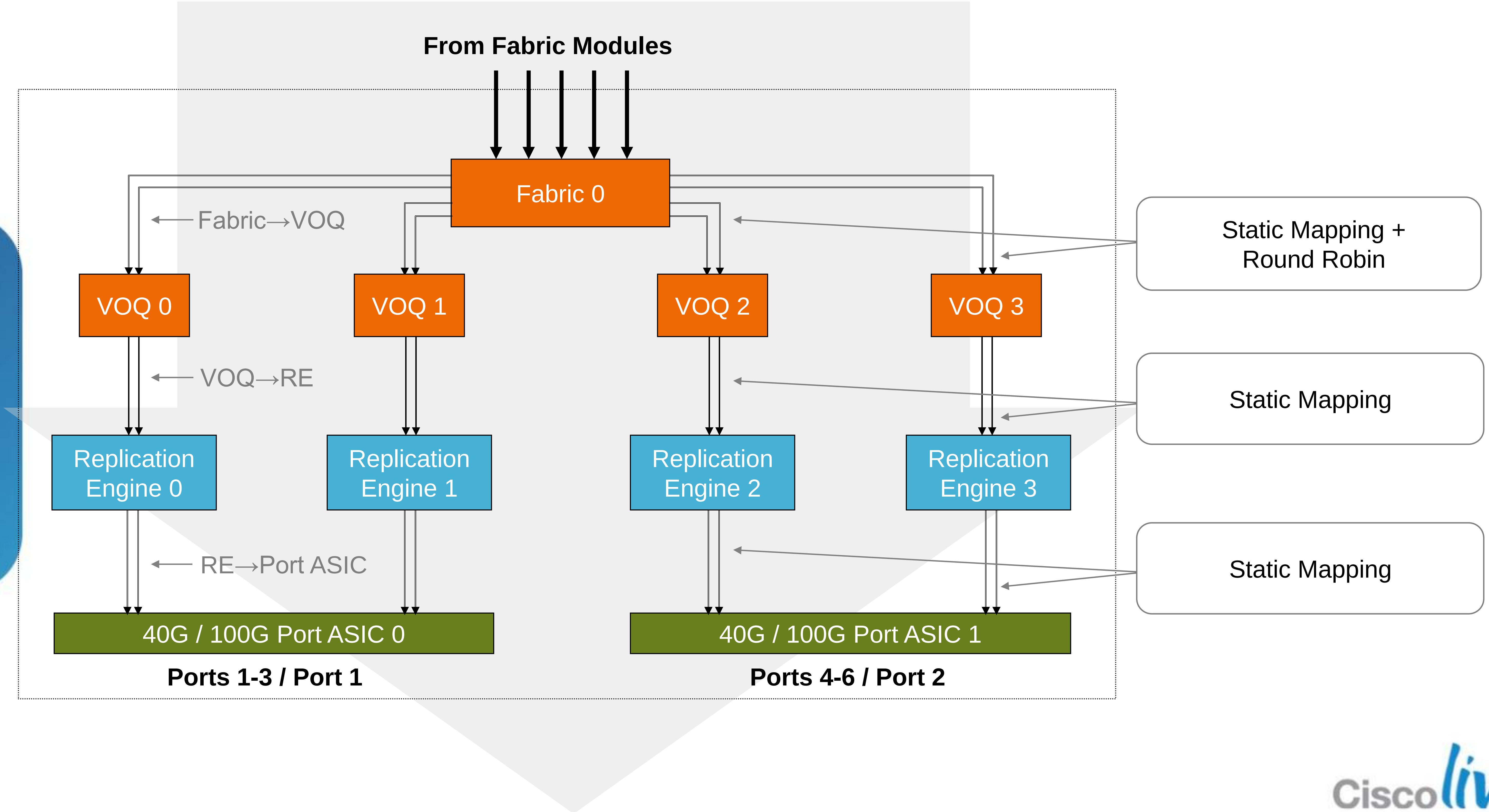
– 40G / 100G M2 Module

- Hash Result generated by Port ASIC selects replication engine uplink
- Hash input uses Layer 3 + Layer 4 information



© 2013 Cisco and/or its affiliates. All rights reserved.

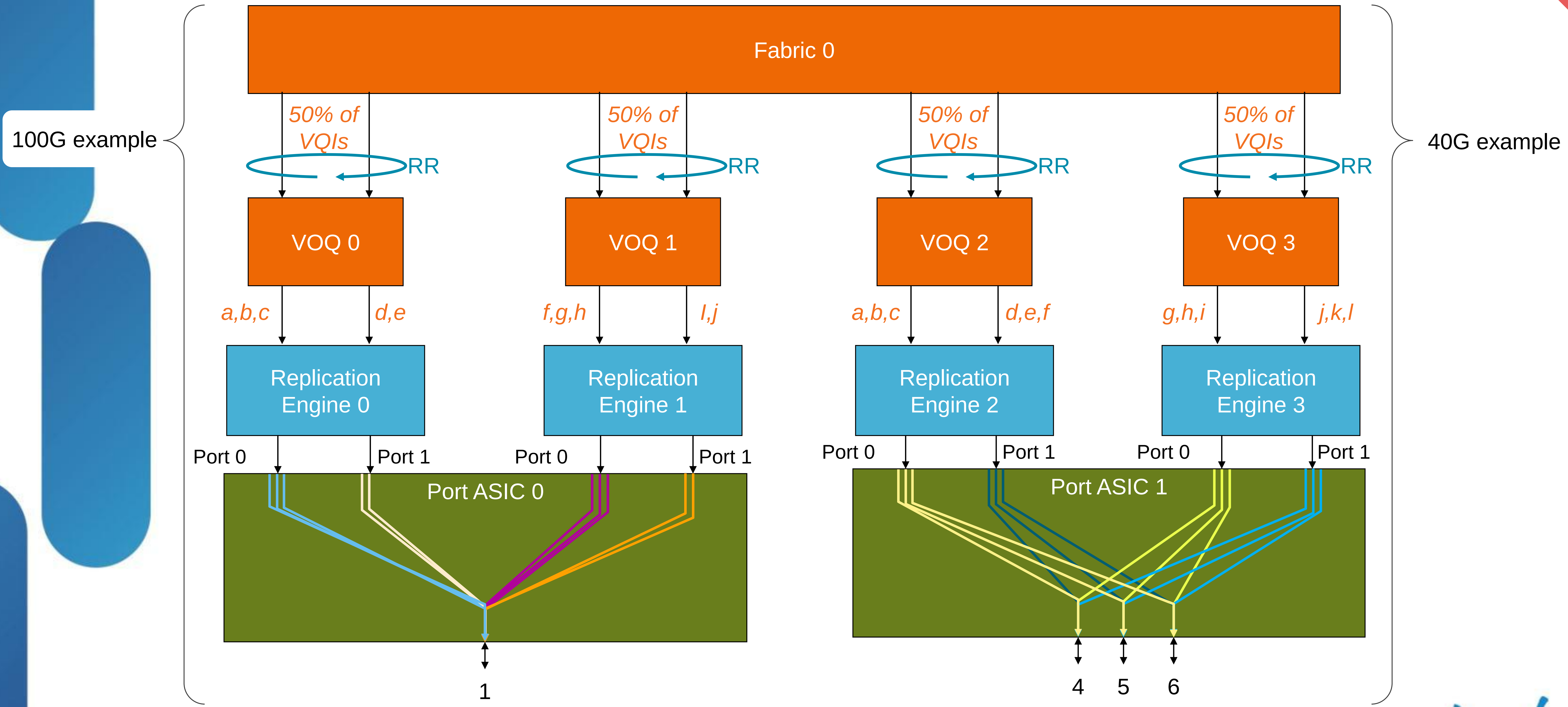
40G / 100G M2 Module Egress Path



M Series

© 2013 Cisco and/or its affiliates. All rights reserved.

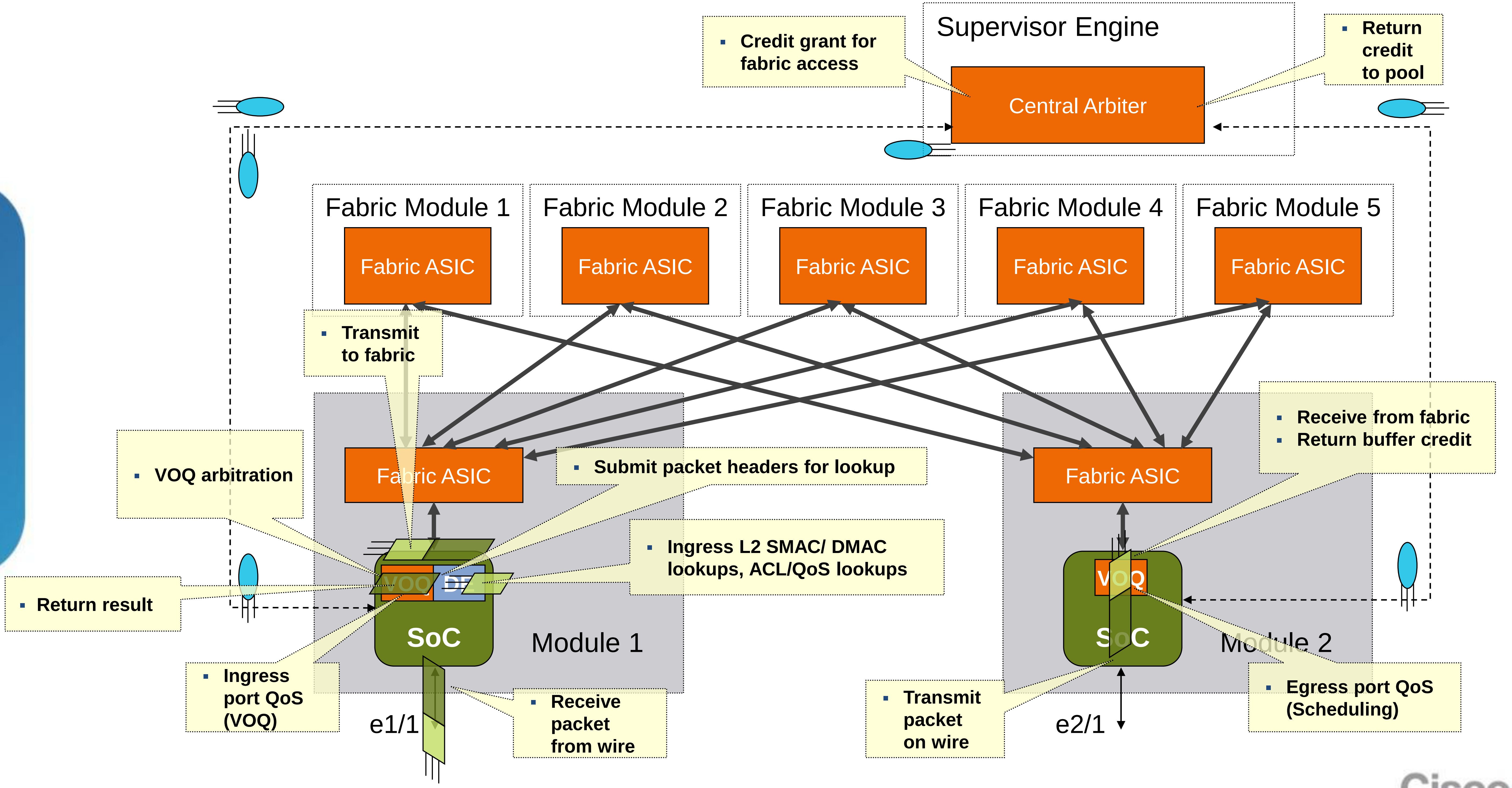
40G / 100G M2 Module Egress VQI Mapping



HDR = Packet Headers DATA = Packet Data CTRL = Internal Signalling

F Series

F2 / F2E L2 Packet Flow



- Credit grant for fabric access
- Return credit to pool
- Transmit to fabric
- VOQ arbitration
- Submit packet headers for lookup
- Receive from fabric
- Return buffer credit
- Ingress L2 SMAC/ DMAC lookups, ACL/QoS lookups
- Return result
- Ingress port QoS (VOQ)
- Receive packet from wire
- Transmit packet on wire
- Egress port QoS (Scheduling)

© 2013 Cisco and/or its affiliates. All rights reserved.

Agenda

- Chassis Architecture
- Supervisor Engine and I/O Module Architecture
- Forwarding Engine Architecture
- Fabric Architecture
- I/O Module Queuing
- Layer 2 Forwarding
- IP Forwarding
- Classification
- NetFlow
- Conclusion

© 2013 Cisco and/or its affiliates. All rights reserved.

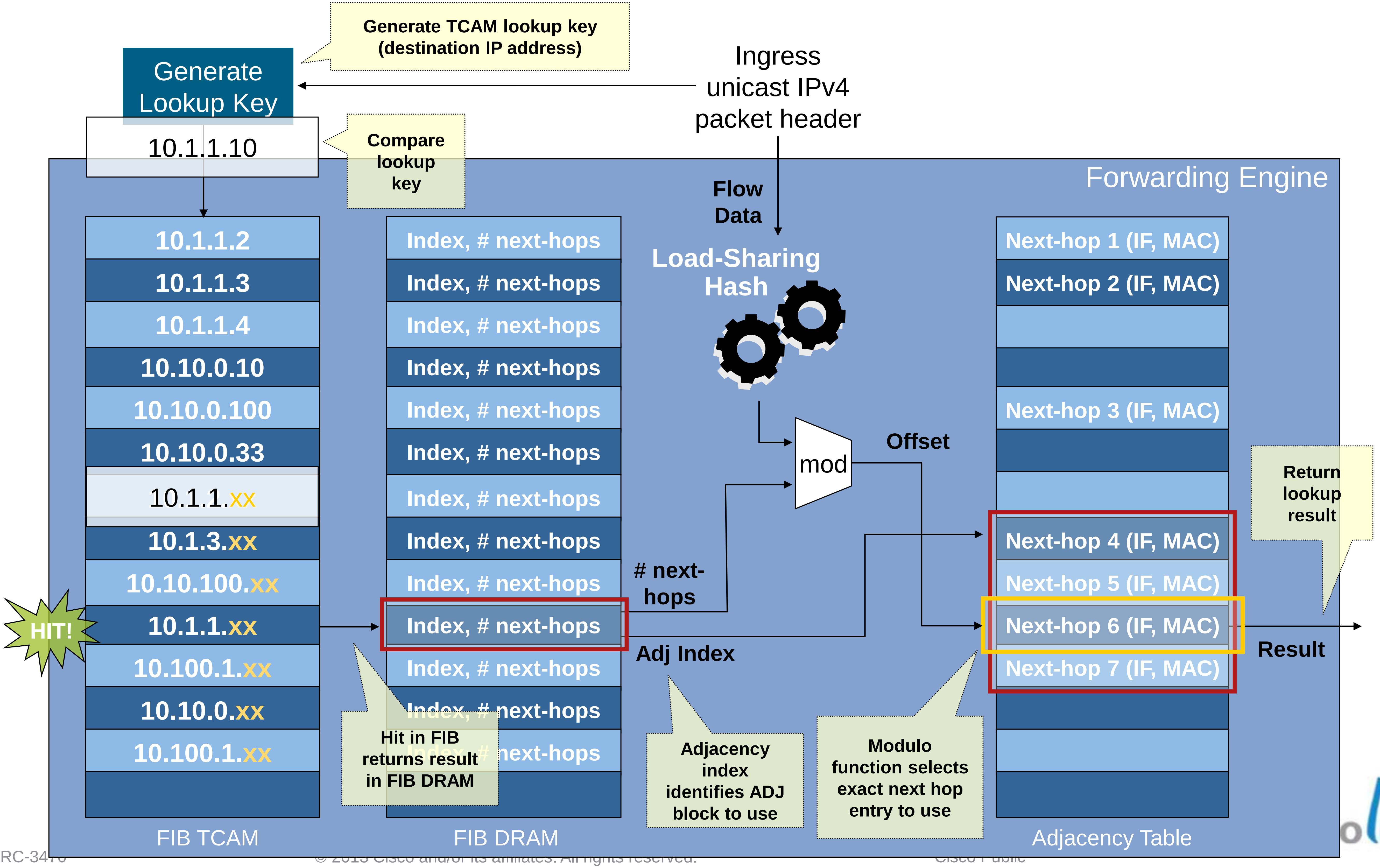
IP Forwarding

- Nexus 7000 decouples control plane and data plane
- Forwarding tables built on control plane using routing protocols or static configuration
  - –OSPF, EIGRP, IS-IS, RIP, BGP for dynamic routing
- Tables downloaded to forwarding engine hardware for data plane forwarding
  - –FIB TCAM contains IP prefixes
  - –Adjacency table contains next-hop information

Hardware IP Forwarding Process

- FIB TCAM lookup based on destination prefix (longest-match)
- FIB "hit" returns adjacency, adjacency contains rewrite information (next-hop)
- Pipelined forwarding engine architecture also performs ACL, QoS, and NetFlow lookups, affecting final forwarding result

IPv4 FIB TCAM Lookup (M1/M2)



Generate TCAM lookup key (destination IP address)

Generate Lookup Key

Ingress unicast IPv4 packet header

10.1.1.10

Compare lookup key

Forwarding Engine

Flow Data

Load-Sharing Hash

FIB TCAM:

- 10.1.1.2
- 10.1.1.3
- 10.1.1.4
- 10.10.0.10
- 10.10.0.100
- 10.10.0.33
- 10.1.1.xx
- 10.1.3.xx
- 10.10.100.xx
- 10.1.1.xx
- 10.100.1.xx
- 10.10.0.xx
- 10.100.1.xx

HIT!

FIB DRAM:

- Index, # next-hops
- Index, # next-hops
- Index, # next-hops
- Index, # next-hops
- Index, # next-hops
- Index, # next-hops
- Index, # next-hops
- Index, # next-hops
- Index, # next-hops
- Index, # next-hops
- Index, # next-hops
- Index, # next-hops
- Index, # next-hops

Hit in FIB returns result in FIB DRAM

mod

Offset

# next-hops

Adj Index

Adjacency index identifies ADJ block to use

Modulo function selects exact next hop entry to use

Adjacency Table:

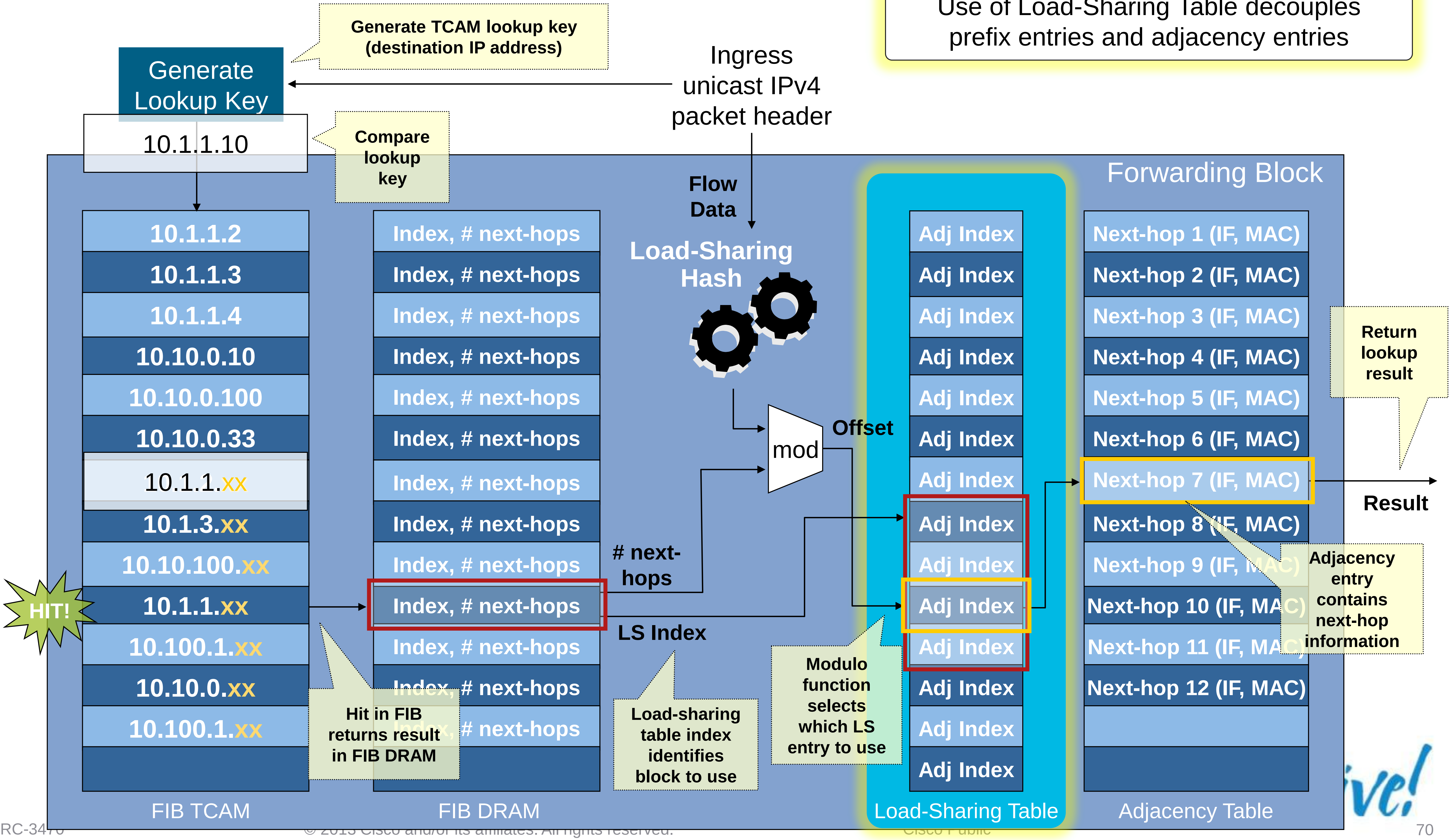
- Next-hop 1 (IF, MAC)
- Next-hop 2 (IF, MAC)
- Next-hop 3 (IF, MAC)
- Next-hop 4 (IF, MAC)
- Next-hop 5 (IF, MAC)
- Next-hop 6 (IF, MAC)
- Next-hop 7 (IF, MAC)

Return lookup result

Result

IPv4 FIB TCAM Lookup (F2 / F2E)

Use of Load-Sharing Table decouples prefix entries and adjacency entries



Generate TCAM lookup key (destination IP address)

Generate Lookup Key

10.1.1.10

Ingress unicast IPv4 packet header

Compare lookup key

Forwarding Block

Flow Data

Load-Sharing Hash

mod

Offset

# next-hops

LS Index

FIB TCAM	FIB DRAM	Load-Sharing Table	Adjacency Table
10.1.1.2	Index, # next-hops	Adj Index	Next-hop 1 (IF, MAC)
10.1.1.3	Index, # next-hops	Adj Index	Next-hop 2 (IF, MAC)
10.1.1.4	Index, # next-hops	Adj Index	Next-hop 3 (IF, MAC)
10.10.0.10	Index, # next-hops	Adj Index	Next-hop 4 (IF, MAC)
10.10.0.100	Index, # next-hops	Adj Index	Next-hop 5 (IF, MAC)
10.10.0.33	Index, # next-hops	Adj Index	Next-hop 6 (IF, MAC)
10.1.1.xx	Index, # next-hops	Adj Index	Next-hop 7 (IF, MAC)
10.1.3.xx	Index, # next-hops	Adj Index	Next-hop 8 (IF, MAC)
10.10.100.xx	Index, # next-hops	Adj Index	Next-hop 9 (IF, MAC)
10.1.1.xx	Index, # next-hops	Adj Index	Next-hop 10 (IF, MAC)
10.100.1.xx	Index, # next-hops	Adj Index	Next-hop 11 (IF, MAC)
10.10.0.xx	Index, # next-hops	Adj Index	Next-hop 12 (IF, MAC)
10.100.1.xx	Index, # next-hops	Adj Index	
		Adj Index	

Hit in FIB returns result in FIB DRAM

Load-sharing table index identifies block to use

Modulo function selects which LS entry to use

Adjacency entry contains next-hop information

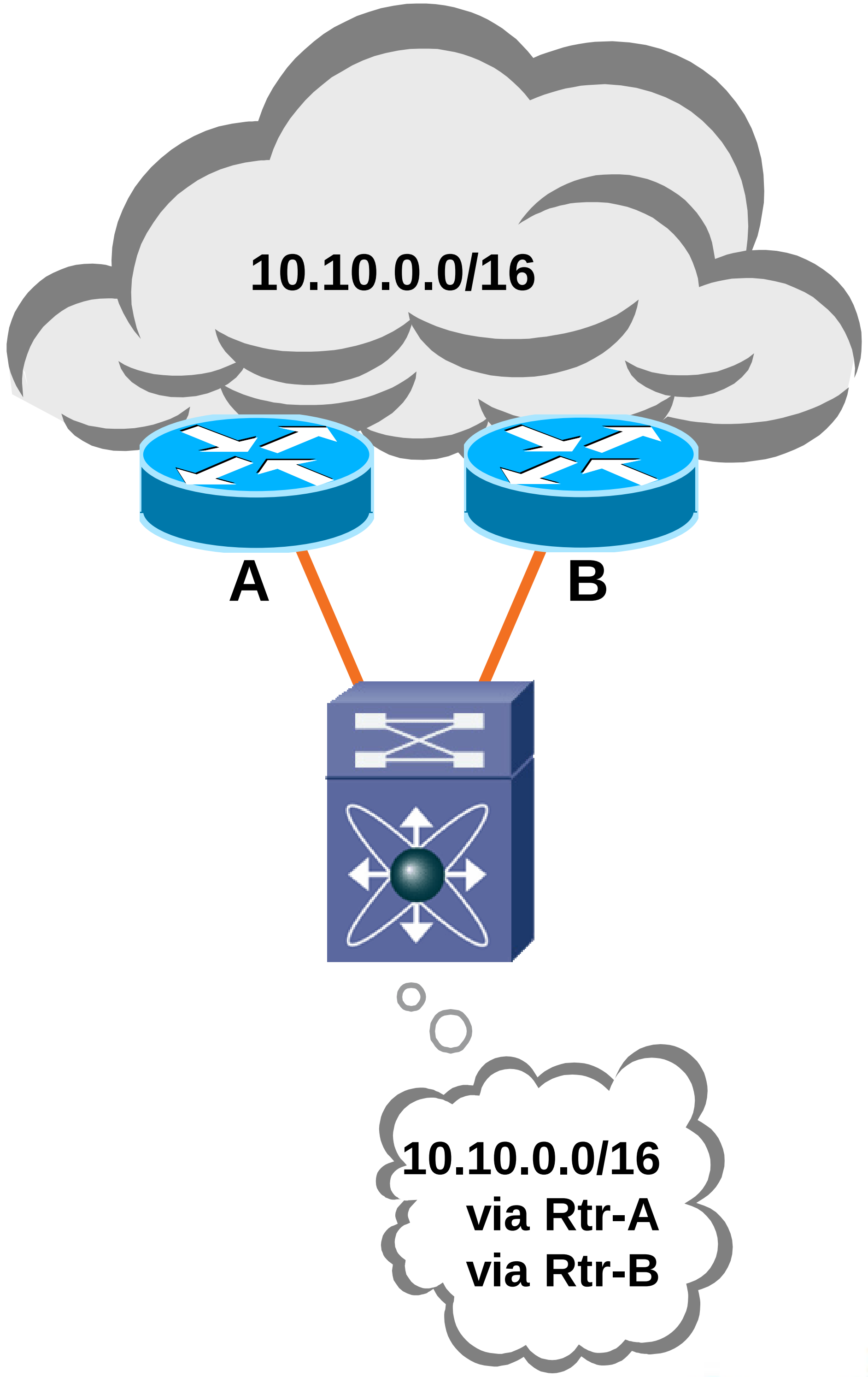
Return lookup result

Result

© 2013 Cisco and/or its affiliates. All rights reserved.

ECMP Load Sharing

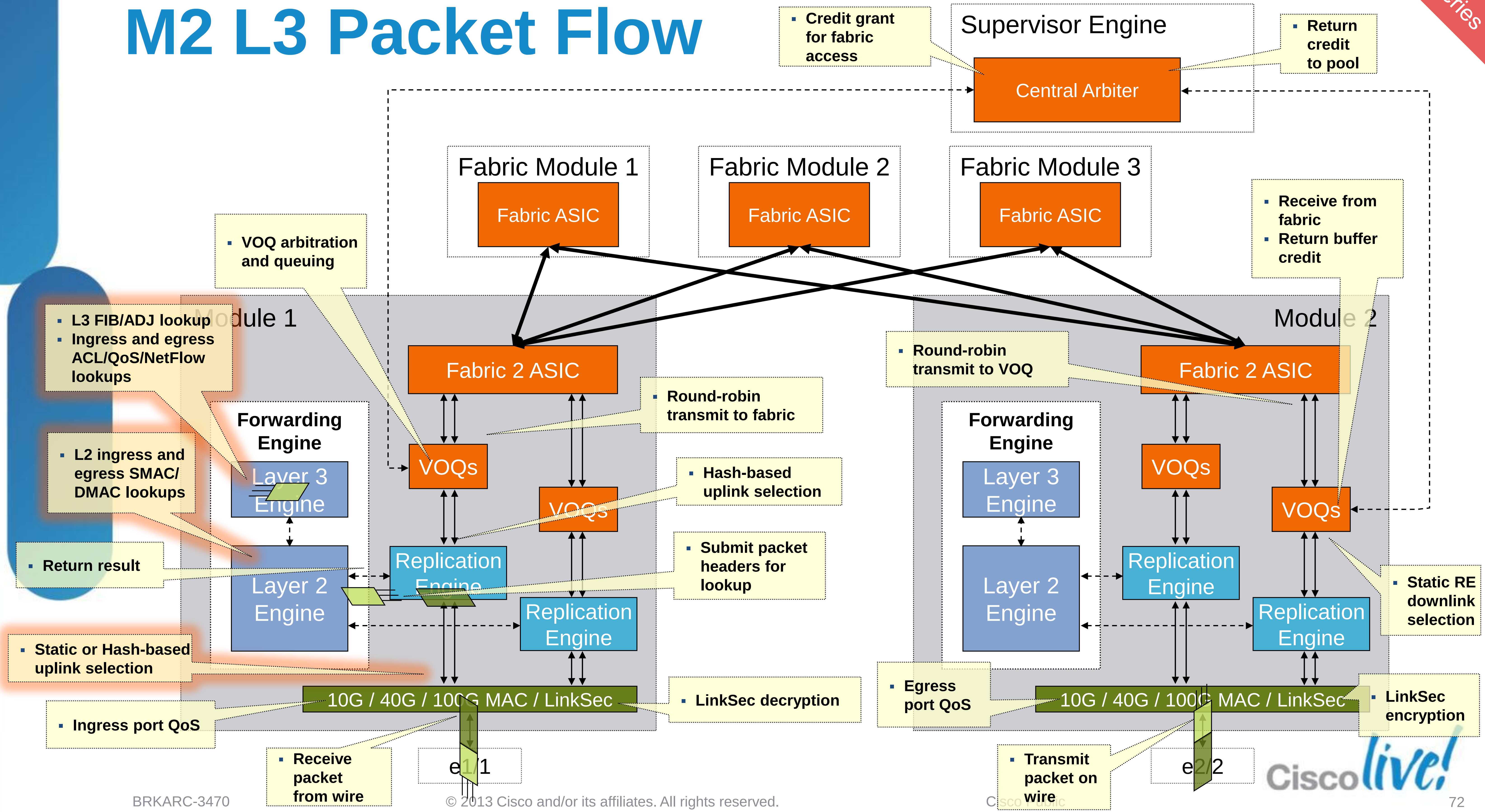
- Up to 16 hardware load-sharing paths per prefix
- Use maximum-paths command in routing protocols to control number of load-sharing paths
- Load-sharing is per-IP flow
- Configure load-sharing hash options with global ip load-sharing command:
  - Source and Destination IP addresses
  - Source and Destination IP addresses plus L4 ports (default)
  - Destination IP address and L4 port
- Additional randomised number added to hash prevents polarisation
  - Automatically generated or user configurable value



HDR = Packet Headers DATA = Packet Data CTRL = Internal Signalling

M Series

M2 L3 Packet Flow



Supervisor Engine

Central Arbiter

- Credit grant for fabric access

- Return credit to pool

Fabric Module 1 — Fabric ASIC

Fabric Module 2 — Fabric ASIC

Fabric Module 3 — Fabric ASIC

- Receive from fabric
- Return buffer credit

- VOQ arbitration and queuing

Module 1

- L3 FIB/ADJ lookup
- Ingress and egress ACL/QoS/NetFlow lookups

- L2 ingress and egress SMAC/ DMAC lookups

- Return result

- Static or Hash-based uplink selection

- Ingress port QoS

- Receive packet from wire

Fabric 2 ASIC

Forwarding Engine

Layer 3 Engine

Layer 2 Engine

VOQs

VOQs

Replication Engine

Replication Engine

10G / 40G / 100G MAC / LinkSec

e1/1

- Round-robin transmit to fabric

- Hash-based uplink selection

- Submit packet headers for lookup

- LinkSec decryption

Module 2

- Round-robin transmit to VOQ

Fabric 2 ASIC

Forwarding Engine

Layer 3 Engine

Layer 2 Engine

VOQs

VOQs

Replication Engine

Replication Engine

10G / 40G / 100G MAC / LinkSec

e2/2

- Static RE downlink selection

- Egress port QoS

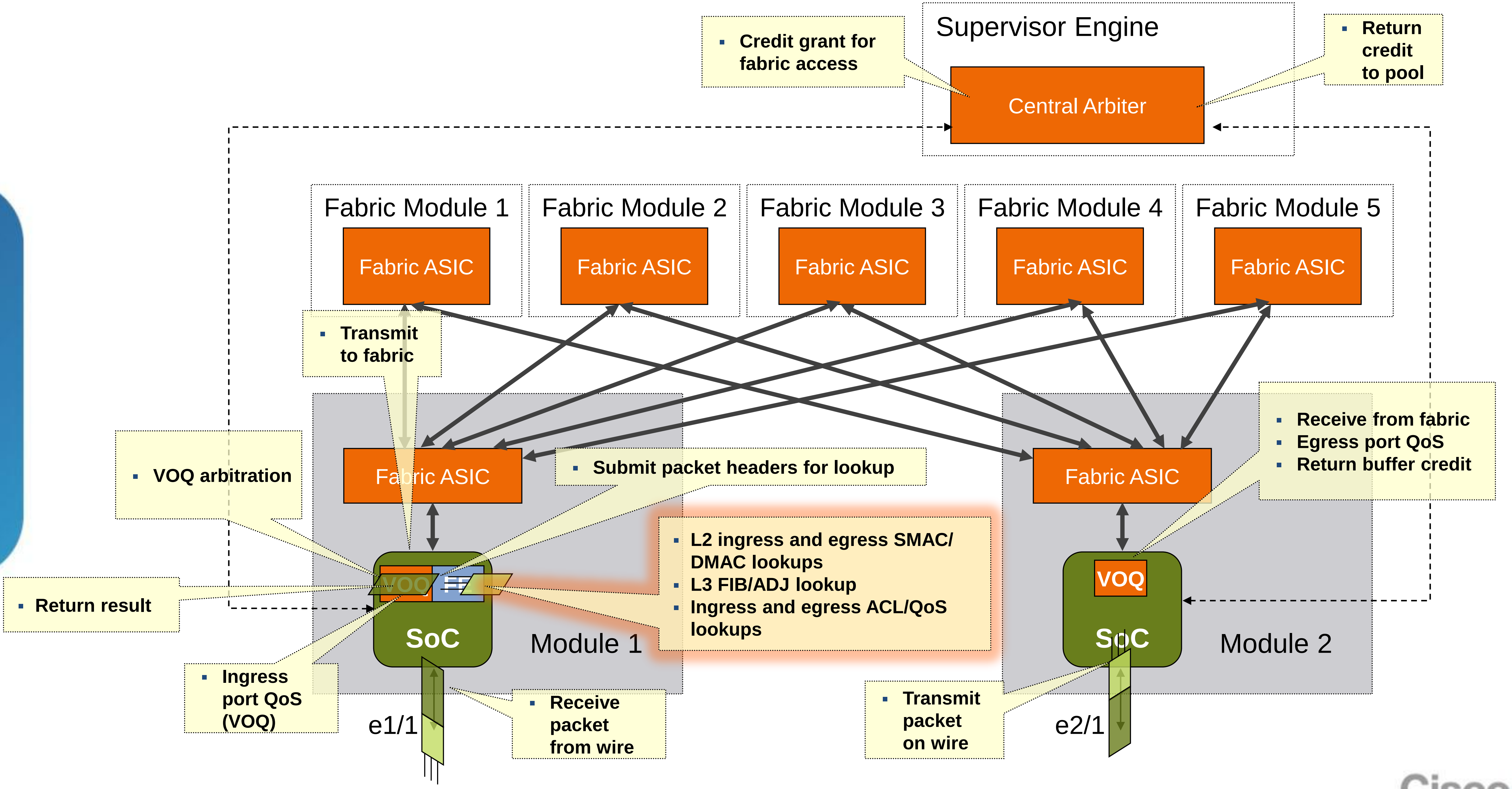
- LinkSec encryption

- Transmit packet on wire

Cisco live!

HDR = Packet Headers DATA = Packet Data CTRL = Internal Signalling

F2 / F2E L3 Packet Flow



- Credit grant for fabric access

- Return credit to pool

- Transmit to fabric

- VOQ arbitration

- Submit packet headers for lookup

- Receive from fabric
- Egress port QoS
- Return buffer credit

- L2 ingress and egress SMAC/ DMAC lookups
- L3 FIB/ADJ lookup
- Ingress and egress ACL/QoS lookups

- Return result

- Ingress port QoS (VOQ)

- Receive packet from wire

- Transmit packet on wire

© 2013 Cisco and/or its affiliates. All rights reserved.

Agenda

- Chassis Architecture
- Supervisor Engine and I/O Module Architecture
- Forwarding Engine Architecture
- Fabric Architecture
- I/O Module Queuing
- Layer 2 Forwarding
- IP Forwarding
- Classification
- NetFlow
- Conclusion

© 2013 Cisco and/or its affiliates. All rights reserved.

What is Classification?

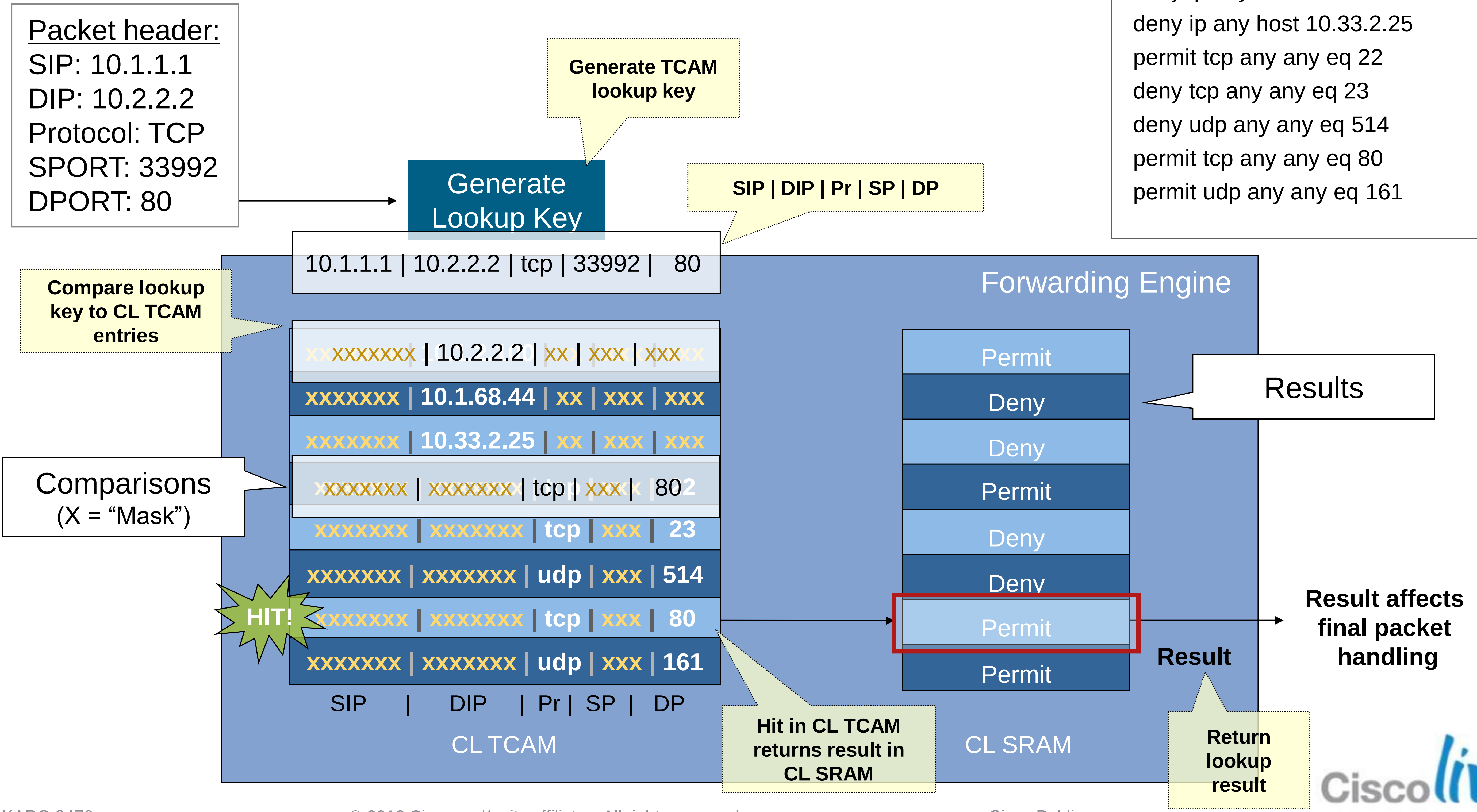
- Matching packets
  - – Layer 2, Layer 3, and/or Layer 4 information
- Used to decide whether to apply a particular policy to a packet
  - – Enforce security, QoS, or other policies
- Some examples:
  - – Match TCP/UDP source/destination port numbers to enforce security policy
  - – Match destination IP addresses to apply policy-based routing (PBR)
  - – Match 5-tuple to apply marking policy
  - – Match protocol-type to apply Control Plane Policing (CoPP)
  - – etc.

© 2013 Cisco and/or its affiliates. All rights reserved.

CL TCAM Lookup – ACL

Security ACL

```
ip access-list example
 permit ip any host 10.1.2.100
 deny ip any host 10.1.68.44
 deny ip any host 10.33.2.25
 permit tcp any any eq 22
 deny tcp any any eq 23
 deny udp any any eq 514
 permit tcp any any eq 80
 permit udp any any eq 161
```



Packet header:
SIP: 10.1.1.1
DIP: 10.2.2.2
Protocol: TCP
SPORT: 33992
DPORT: 80

Generate Lookup Key

Generate TCAM lookup key

SIP | DIP | Pr | SP | DP

10.1.1.1 | 10.2.2.2 | tcp | 33992 | 80

Compare lookup key to CL TCAM entries

Comparisons (X = “Mask”)

SIP	DIP	Pr	SP	DP	CL SRAM
xxxxxxxx	10.2.2.2	xx	xxx	xxx	Permit
xxxxxxx	10.1.68.44	xx	xxx	xxx	Deny
xxxxxxx	10.33.2.25	xx	xxx	xxx	Deny
xxxxxxx	xxxxxxx	tcp	xxx	80	Permit
xxxxxxx	xxxxxxx	tcp	xxx	23	Deny
xxxxxxx	xxxxxxx	udp	xxx	514	Deny
xxxxxxx	xxxxxxx	tcp	xxx	80	Permit
xxxxxxx	xxxxxxx	udp	xxx	161	Permit

Hit in CL TCAM returns result in CL SRAM

Return lookup result

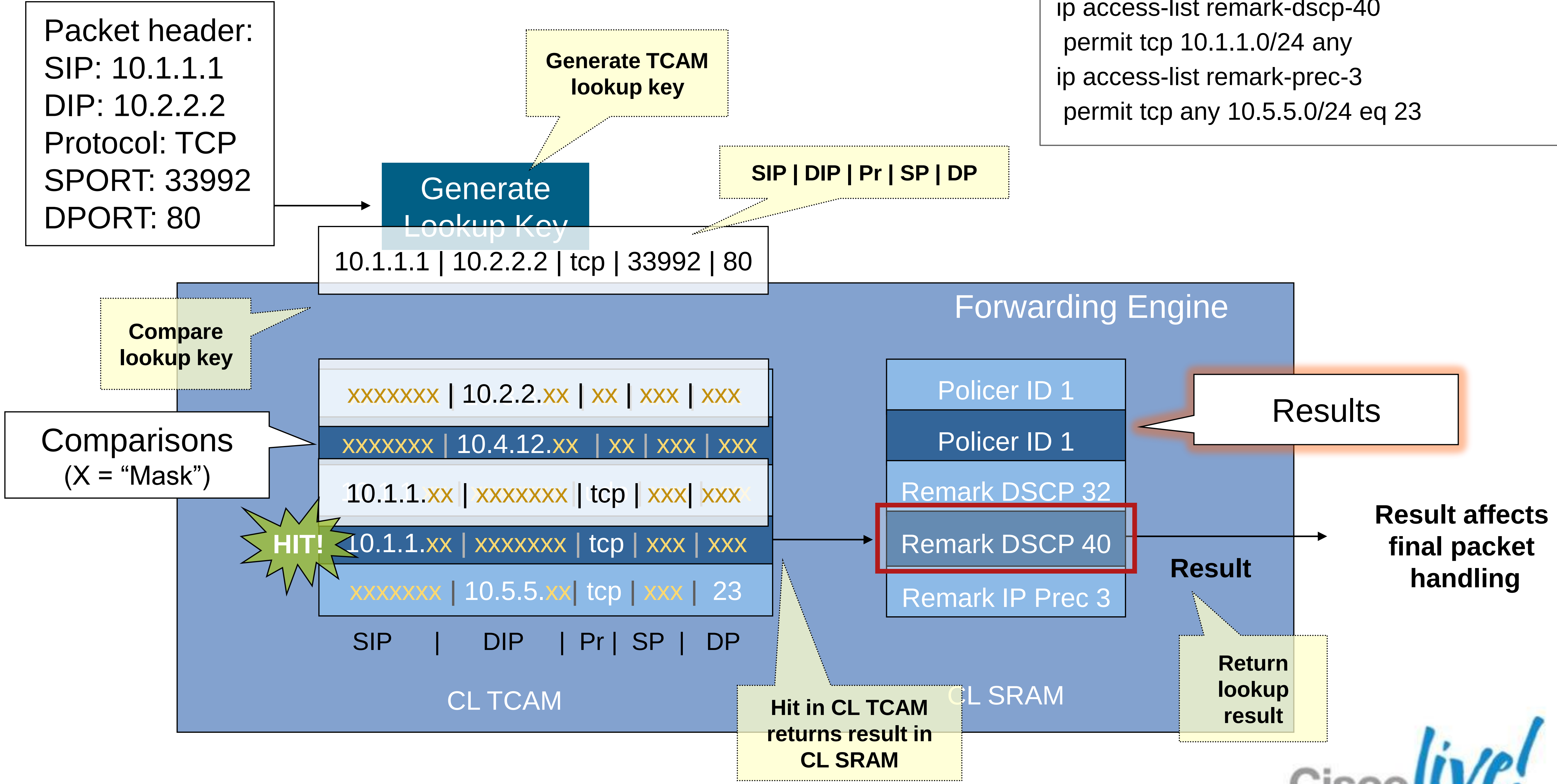
Result affects final packet handling

© 2013 Cisco and/or its affiliates. All rights reserved.

CL TCAM Lookup – QoS

QoS Classification ACLs

```
ip access-list police
 permit ip any 10.3.3.0/24
 permit ip any 10.4.12.0/24
ip access-list remark-dscp-32
 permit udp 10.1.1.0/24 any
ip access-list remark-dscp-40
 permit tcp 10.1.1.0/24 any
ip access-list remark-prec-3
 permit tcp any 10.5.5.0/24 eq 23
```

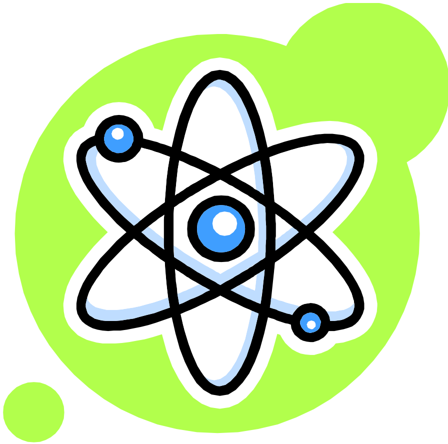


Packet header:
SIP: 10.1.1.1
DIP: 10.2.2.2
Protocol: TCP
SPORT: 33992
DPORT: 80

SIP	DIP	Pr	SP	DP	
xxxxxxx	10.2.2.xx	xx	xxx	xxx	Policer ID 1
xxxxxxx	10.4.12.xx	xx	xxx	xxx	Policer ID 1
10.1.1.xx	xxxxxxx	tcp	xxx	xxx	Remark DSCP 32
10.1.1.xx	xxxxxxx	tcp	xxx	xxx	Remark DSCP 40
xxxxxxx	10.5.5.xx	tcp	xxx	23	Remark IP Prec 3

Atomic Policy Programming

- Avoids packet loss during policy updates
- Enabled by default
- Atomic programming process:
  - – Program new policy in free/available CL TCAM entries
  - – Enable new policy by swapping the ACL label on interface
  - – Free CL TCAM resources used by previous policy



© 2013 Cisco and/or its affiliates. All rights reserved.

Atomic Policy Programming Cont.

- To support atomic programming, **software reserves 50% of available TCAM**
- If insufficient resources available, system returns an error and no modifications made in hardware
  - `Failed to complete Verification: Tcam will be over used, please turn off atomic update`
- Disable with **no platform access-list update atomic**
  - Disabling may be necessary for very large ACL configurations
  - Atomic programming attempted but not mandatory
- User can disable atomic programming and perform update non-atomically (assuming ACL fits in CL TCAM)
  - “Default” ACL result (deny by default) returned for duration of reprogramming
  - Use **[no] hardware access-list update default-result permit** to control default resul

Classification Configuration Sessions

Two ways to configure ACL/QoS policies:

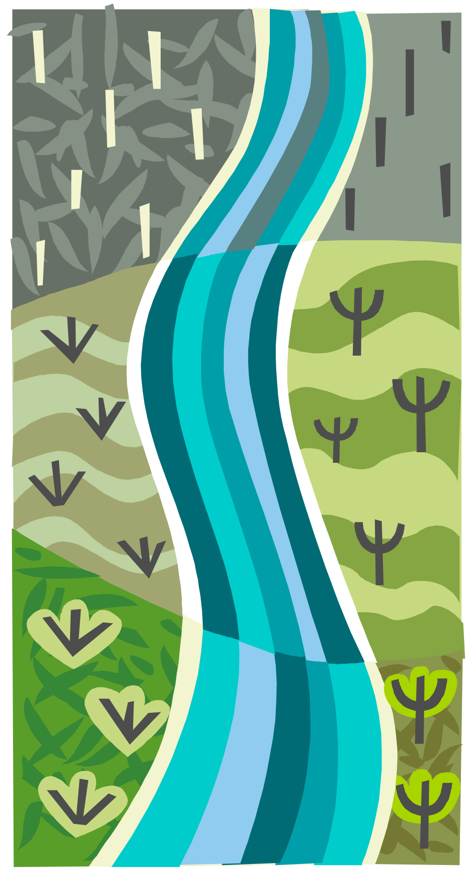
- Normal configuration mode (**config terminal**)
  - Configuration applied immediately line by line
  - Recommended only for small ACL/QoS configurations, or non-data-plane ACL configuration
- Session config mode (**config session**)
  - Configuration only applied after **commit** command issued
  - Recommended for large ACL/QoS configurations
- Config session mode also provides **verify** facility to "dry-run" the configuration against available system resources
  - No change to existing hardware configuration after verification (regardless of verification result)

© 2013 Cisco and/or its affiliates. All rights reserved.

Agenda

- Chassis Architecture
- Supervisor Engine and I/O Module Architecture
- Forwarding Engine Architecture
- Fabric Architecture
- I/O Module Queuing
- Layer 2 Forwarding
- IP Forwarding
- Classification
- NetFlow
- Conclusion

NetFlow on Nexus 7000



- NetFlow collects flow data for packets traversing the switch
- Each module maintains independent NetFlow table

	M1 / M2	F2 / F2E
Per-interface NetFlow	Yes	Yes
NetFlow direction	Ingress/Egress	Ingress only
Full NetFlow	Yes	No
Sampled NetFlow	Yes	Yes
Bridged NetFlow	Yes	Yes
Hardware Cache	Yes	No
Software Cache	No	Yes
Hardware Cache Size	512K entries per forwarding engine	N/A
NDE (v5/v9)	Yes	Yes

© 2013 Cisco and/or its affiliates. All rights reserved.

Full vs. Sampled NetFlow

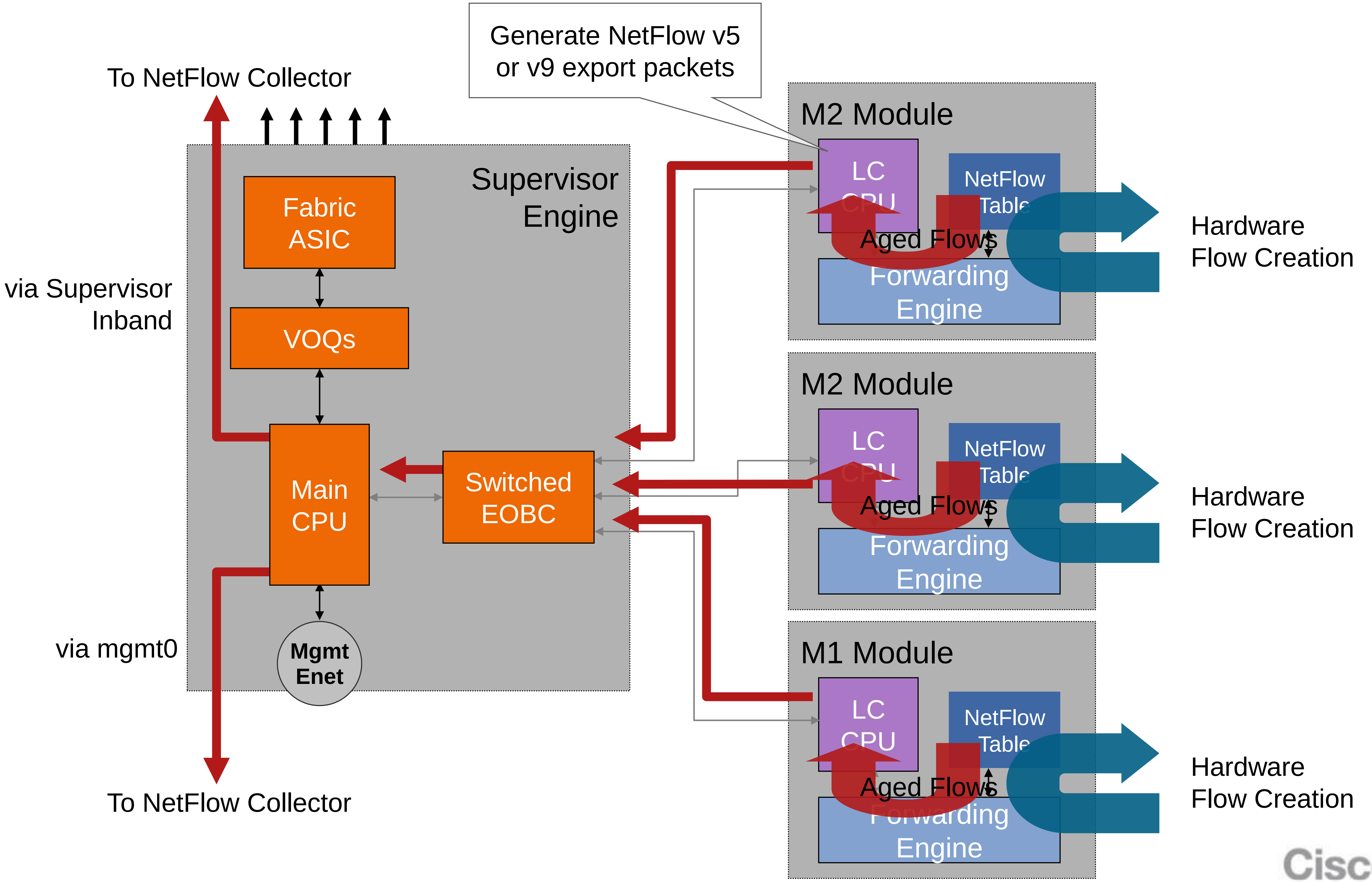
- NetFlow collects full or sampled flow data
- Full NetFlow: Accounts for every packet of every flow on interface
  - Available on M-Series modules only
  - Flow data collection up to capacity of hardware NetFlow table
- Sampled NetFlow: Accounts for M in N packets on interface
  - Available on both M-Series (ingress/egress) and F2/F2E (ingress only)
  - M-Series: Flow data collection up to capacity of hardware NetFlow table
  - F2/F2E: Flow data collection for up to ~1000pps per module

© 2013 Cisco and/or its affiliates. All rights reserved.

Sampled NetFlow Details

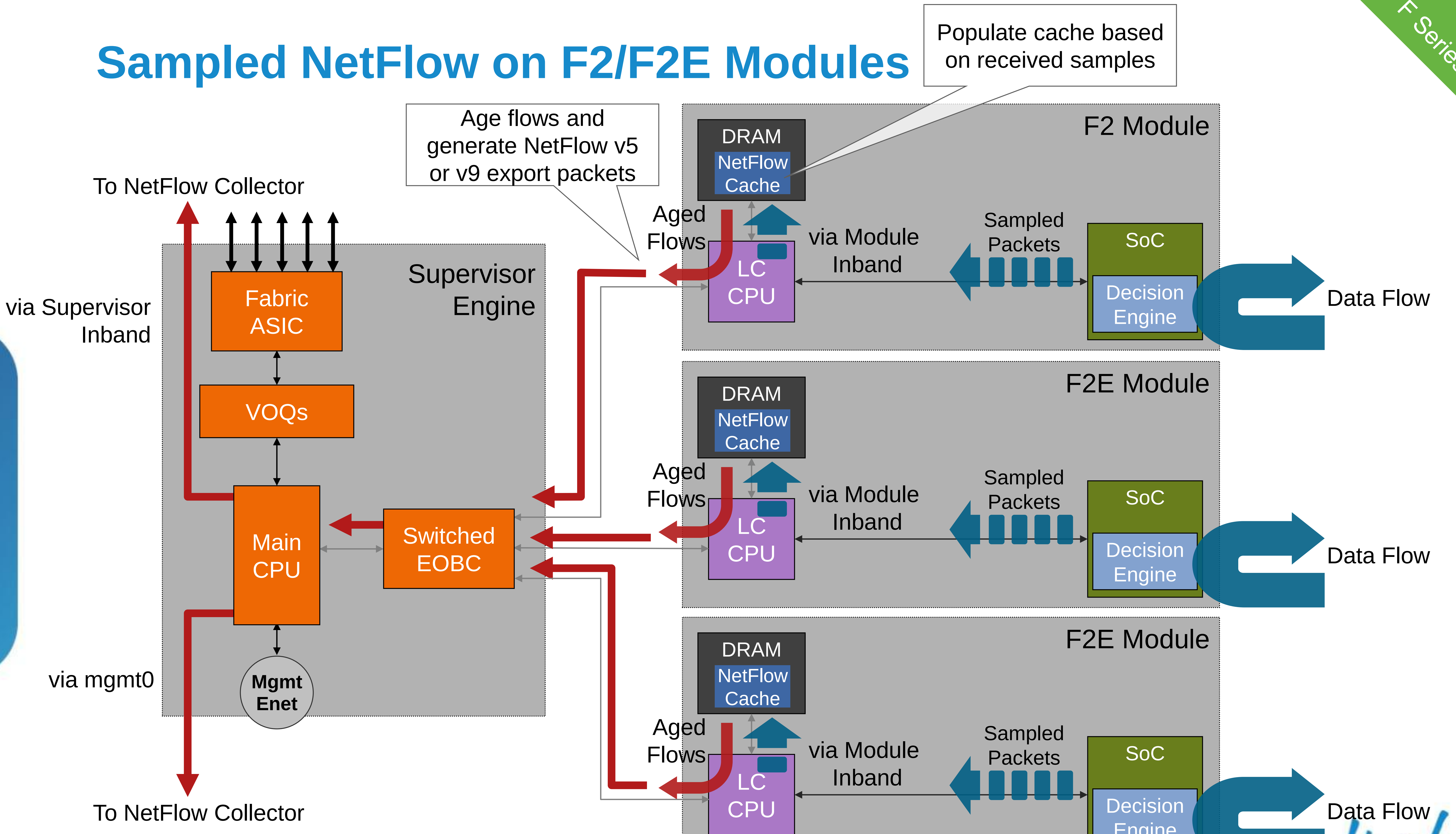
- Random packet-based sampling
- M:N sampling: Out of N consecutive packets, select M consecutive packets and account only for those flows
- On M-Series, sampled packets create hardware NetFlow table entry
- On F2/F2E, sampled packets sent to LC CPU via module inband
  - Rate limited to ~1000pps per module
- Software multiplies configured sampler rate by 100 on F2/F2E modules
  - Example: when using 1 out-of 100 sampler on F2/F2E interface, sampled rate becomes 1:10000

NetFlow on M1/M2 Modules



© 2013 Cisco and/or its affiliates. All rights reserved.

Sampled NetFlow on F2/F2E Modules



© 2013 Cisco and/or its affiliates. All rights reserved.

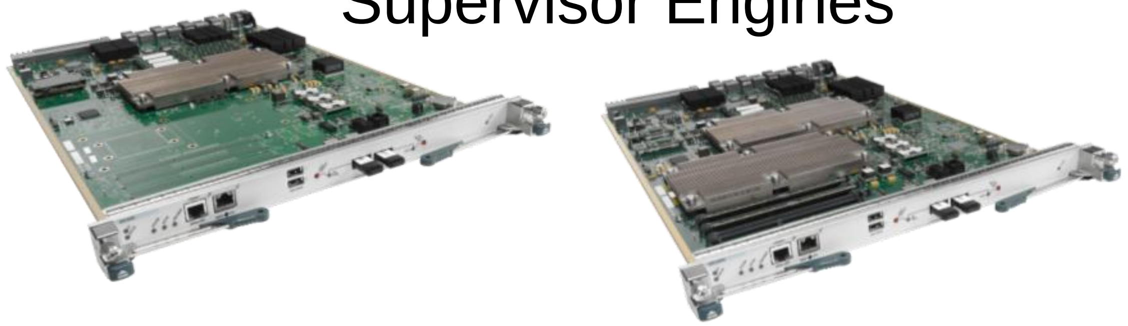
Agenda

- Chassis Architecture
- Supervisor Engine and I/O Module Architecture
- Forwarding Engine Architecture
- Fabric Architecture
- I/O Module Queuing
- Layer 2 Forwarding
- IP Forwarding
- Classification
- NetFlow
- Conclusion

© 2013 Cisco and/or its affiliates. All rights reserved.

Nexus 7000 Architecture Summary

Supervisor Engines

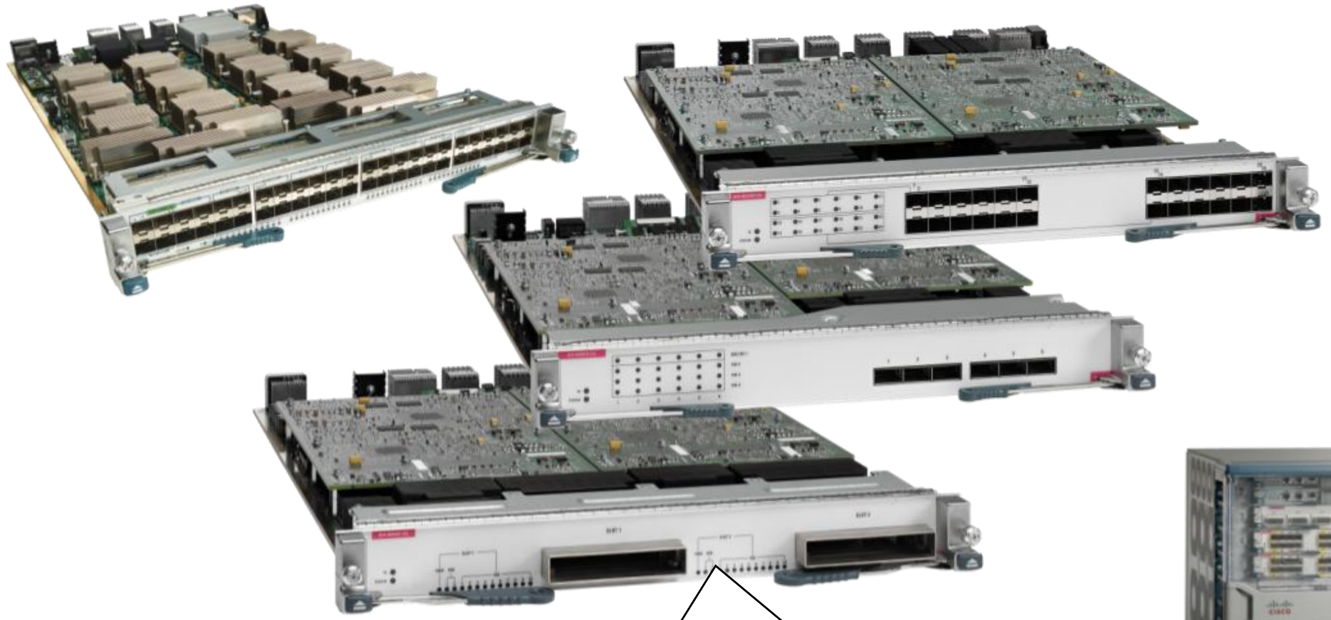


Control plane protocols, system and network management

Chassis

Multiple chassis designs with density and airflow options

I/O Modules



Variety of front-panel interface and transceiver types with hardware-based forwarding and services, including unicast/multicast, bridging/routing, ACL/QoS classification, and NetFlow statistics

Fabrics

High-bandwidth fabric to interconnect I/O modules and provide investment protection

© 2013 Cisco and/or its affiliates. All rights reserved.

Conclusion

- You should now have a thorough understanding of the Nexus 7000 switching architecture, I/O module design, packet flows, and key forwarding engine functions…
- **Any questions?**

© 2013 Cisco and/or its affiliates. All rights reserved.

Q & A

Complete Your Online Session Evaluation

Give us your feedback and receive a Cisco Live 2013 Polo Shirt!

Complete your Overall Event Survey and 5 Session Evaluations.

- Directly from your mobile device on the Cisco Live Mobile App
- By visiting the Cisco Live Mobile Site www.ciscoliveaustralia.com/mobile
- Visit any Cisco Live Internet Station located throughout the venue

Polo Shirts can be collected in the World of Solutions on Friday 8 March 12:00pm-2:00pm

Cisco live! 365

Don't forget to activate your Cisco Live 365 account for access to all session material, communities, and on-demand and live activities throughout the year. Log into your Cisco Live portal and click the "Enter Cisco Live 365" button. www.ciscoliveaustralia.com/portal/login.ww

© 2013 Cisco and/or its affiliates. All rights reserved.