cisco Live!

Let's go

#CiscoLiveAPJC

The bridge to possible

Cisco 8000 Forwarding and QoS Deep Dive

Shinya Horiuchi, Technical Leader

BRKSPG-2208

cisco Live!

#CiscoLiveAPJC

Cisco Webex App

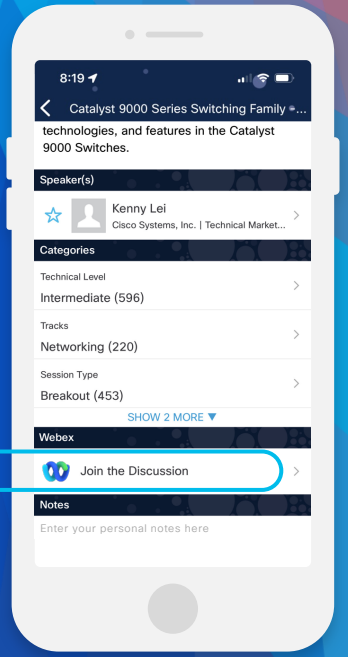
Questions?

Use Cisco Webex App 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 the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until December 22, 2023.



https://ciscolive.ciscoevents.com/ciscolivebot/#BRKSPG-2208

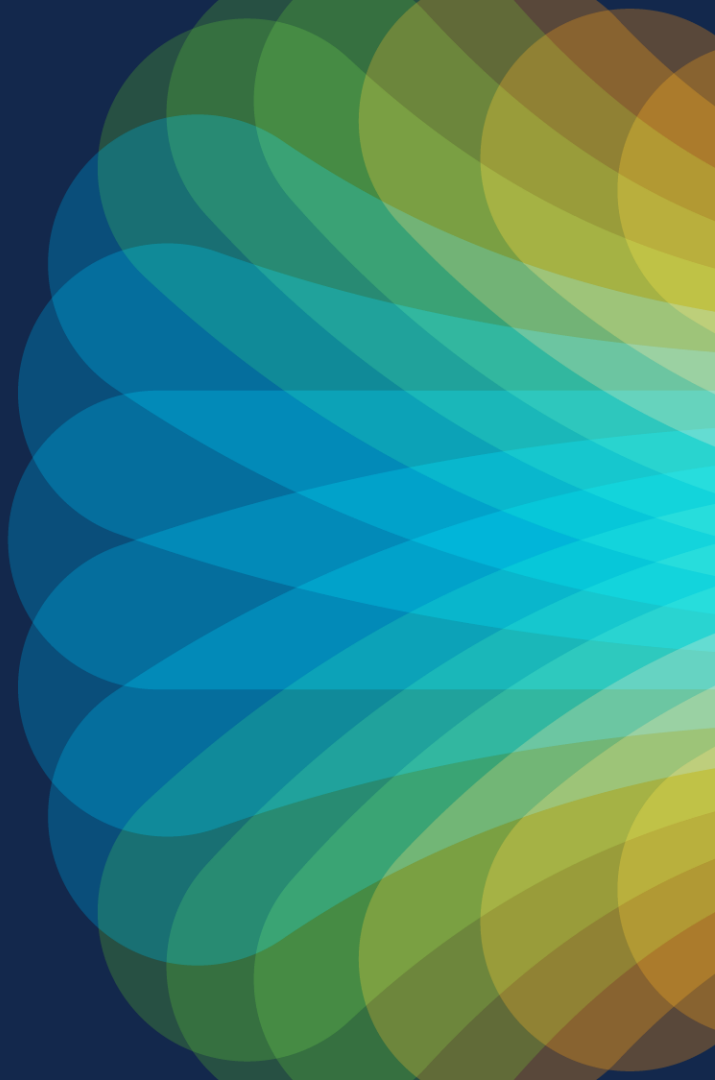
© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Agenda

- System Architecture
  - VOQ
  - Packet Memory & Flow
  - NPU Building Block
- Forwarding Resources
- ECMP/Label optimization
- VOQ Scale & Counter Bank
- QOS Config model
- Scheduling & Queueing
  - RED & Dual Queue Limit
- Policer
- Bundle/MPLS QOS
- QOS Trouble shooting CLI

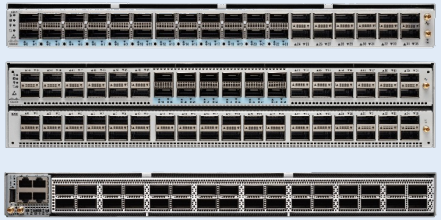
© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

System Architecture

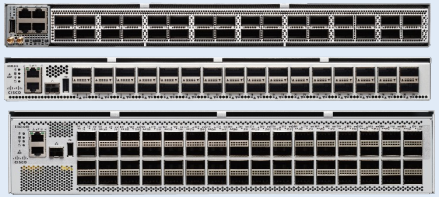


Cisco 8000 Series Product Line

8200

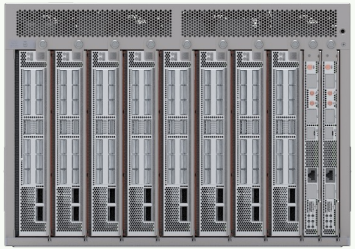


8100



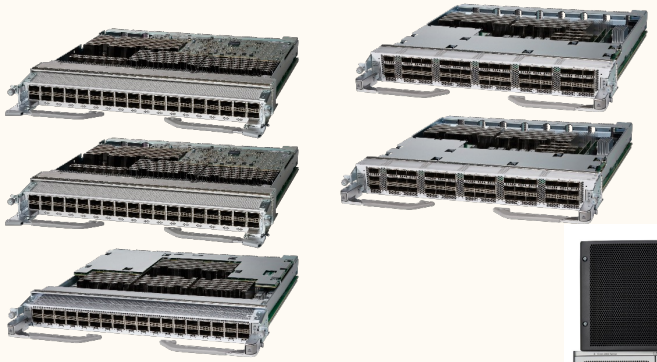
Fixed

8600

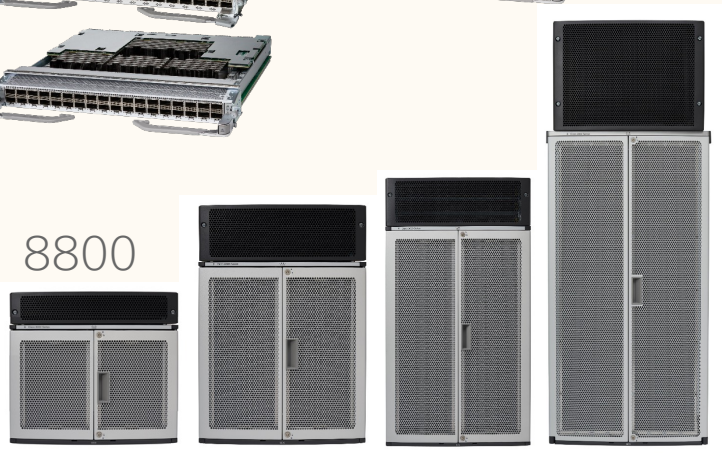


Centralized Modular

Linecards



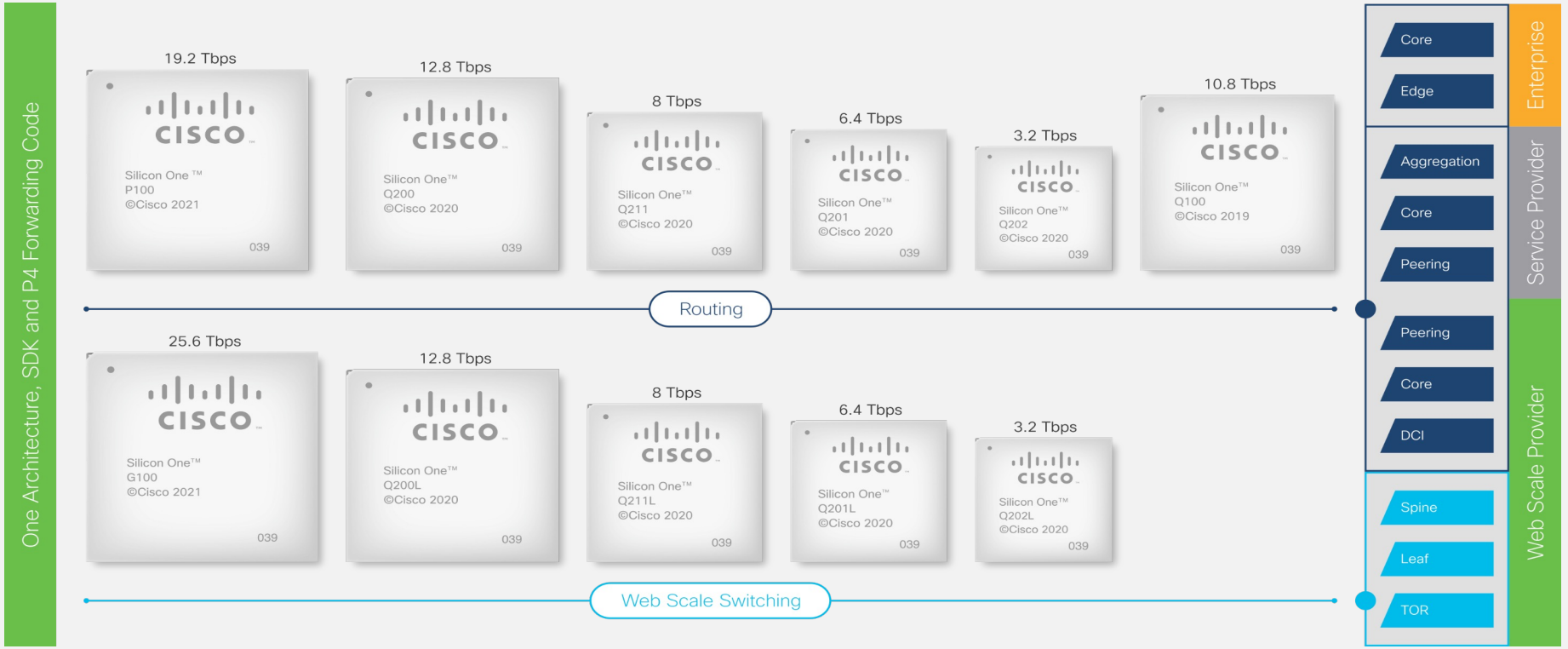
8800



Distributed Modular

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Cisco Silicon One



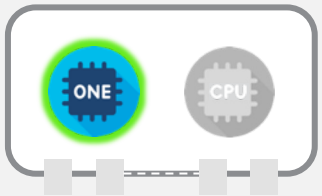
© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

3 System Designs with One ASIC Architecture

Router-on-Chip mode
Linecard mode
Fabric mode

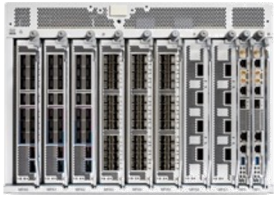
Fixed System

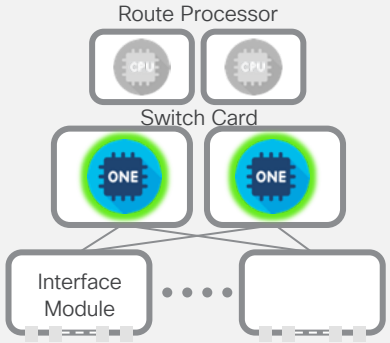
Cisco 8100/8200



Centralized System

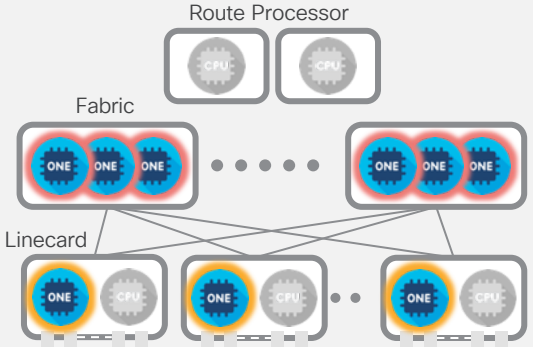
Cisco 8600





Modular System

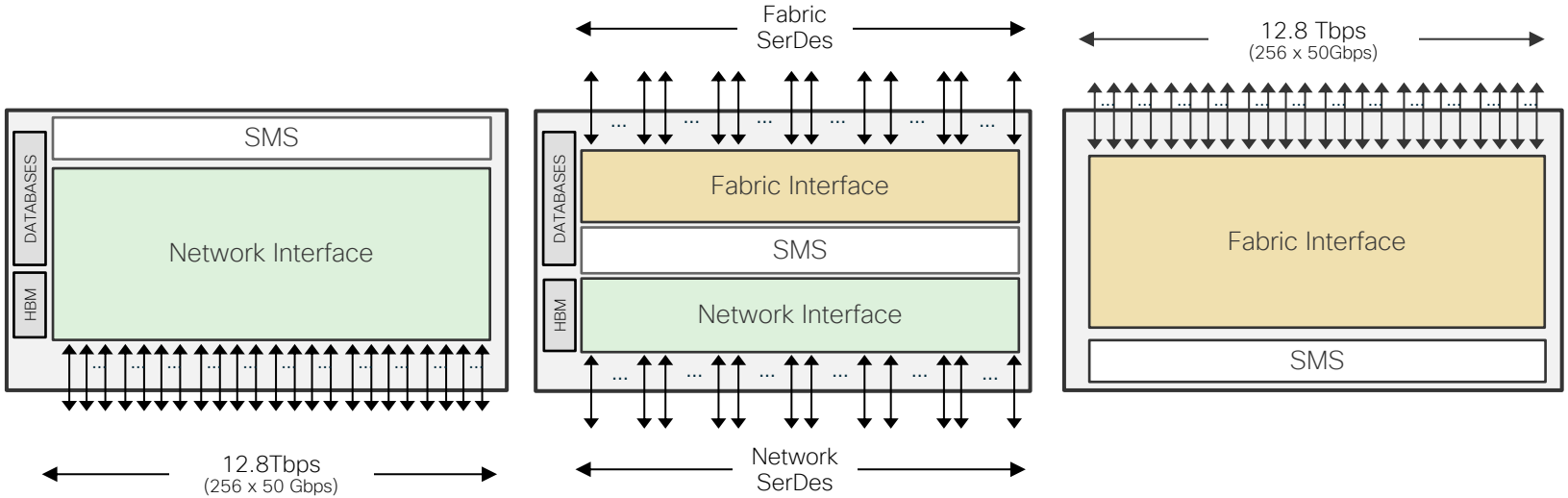
Cisco 8800



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Cisco Silicon One Q200
Modes of Operation

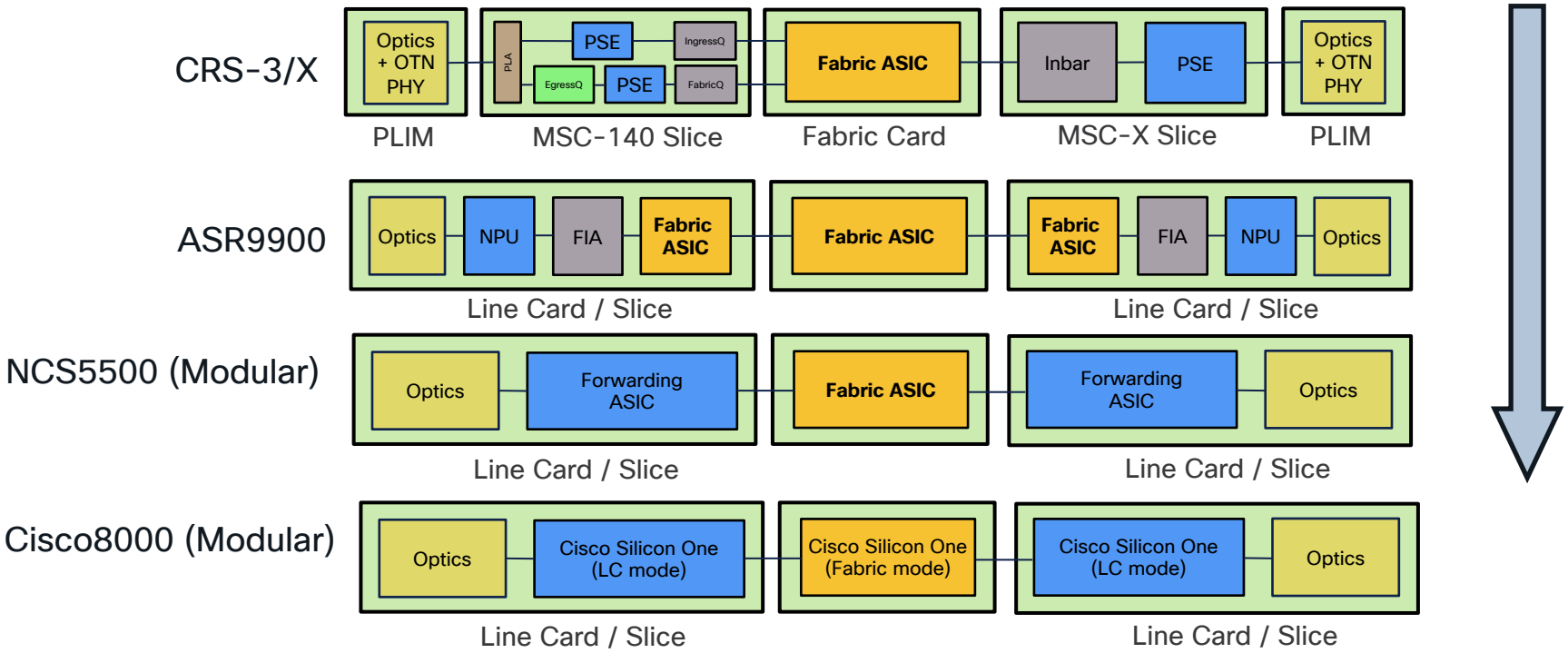
- Cisco Silicon One ASIC works with 3 modes
  - As NPU on fixed chassis
  - As NPU + fabric interface on line card module
  - As fabric ASIC on fabric card



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Simplification is Key

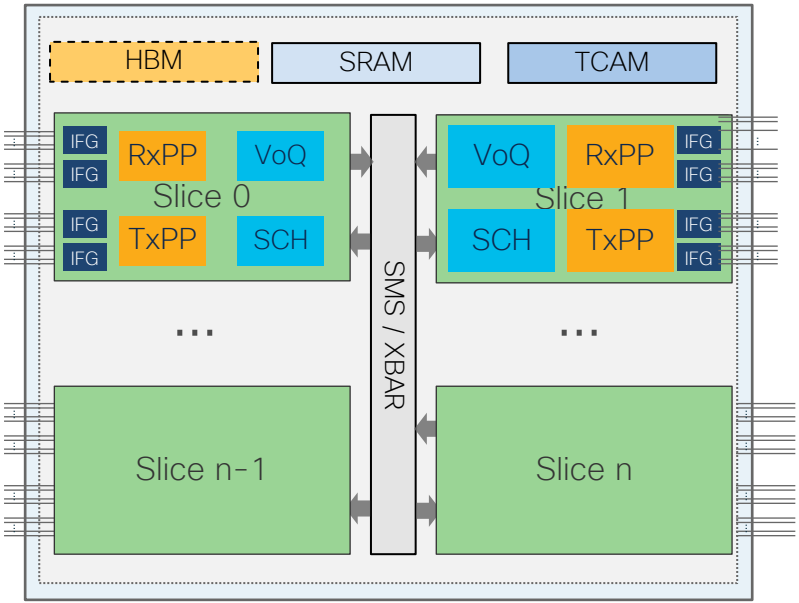
- Fewer Components: Cost Optimization and Lower Power Consumption



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

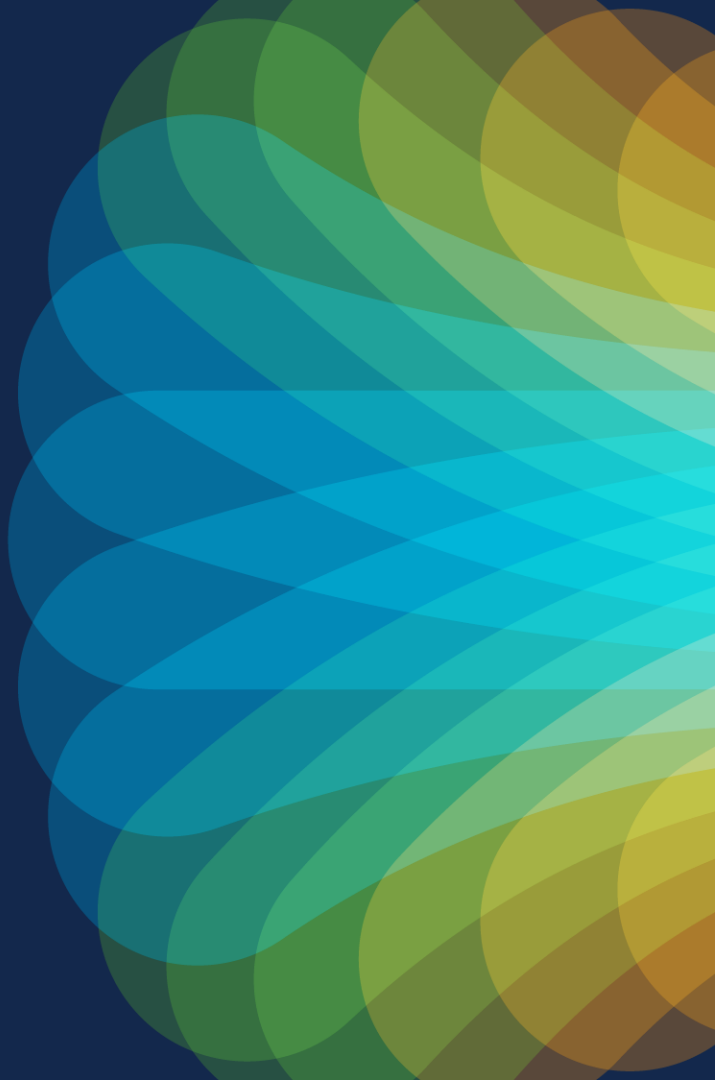
Common ASIC Elements

- Variable number of slices — Slice n
  - 2 Interface Groups (IFGs) per slice
- Separate Rx and Tx components
  - Packet Processors, Schedulers — RxPP, TxPP, SCH
- Shared Memory Switch (SMS)
  - Connects slices via common buffers — SMS / XBAR
- Internal SRAMs and TCAM — SRAM, TCAM
- Multi-rate SerDes down to 10G
- Different P4 flavors based on usage
  - Line card, fabric, or fixed
- HBM memory is optional — HBM
  - Buffer expansion
  - FIB expansion



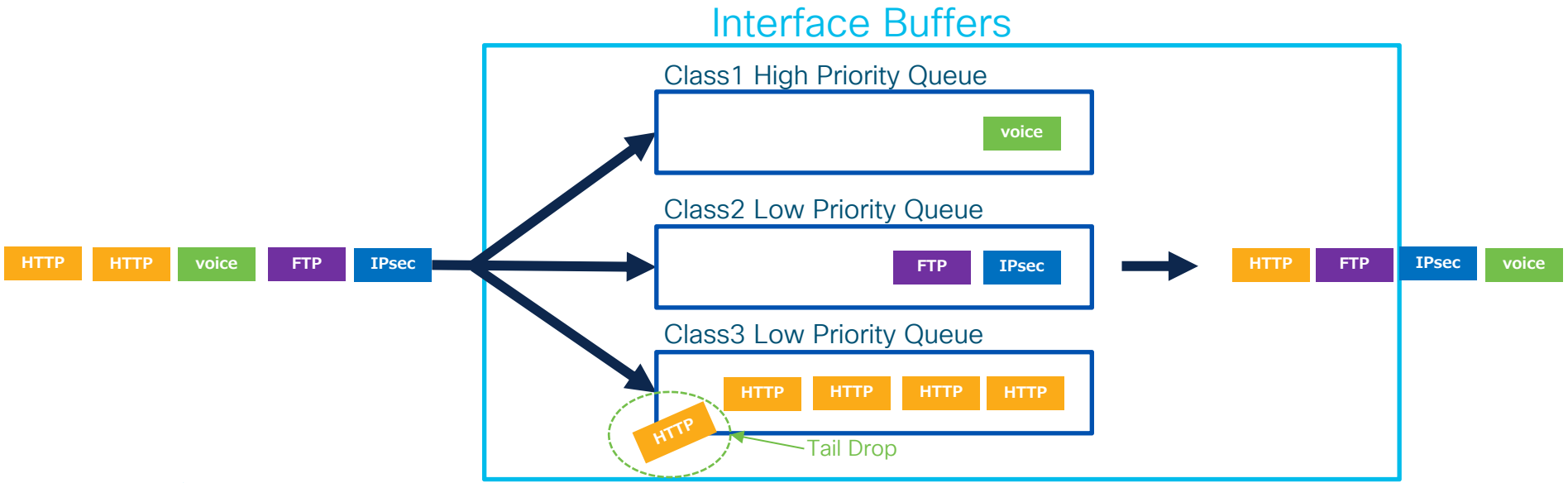
© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

VOQ



What is queueing?

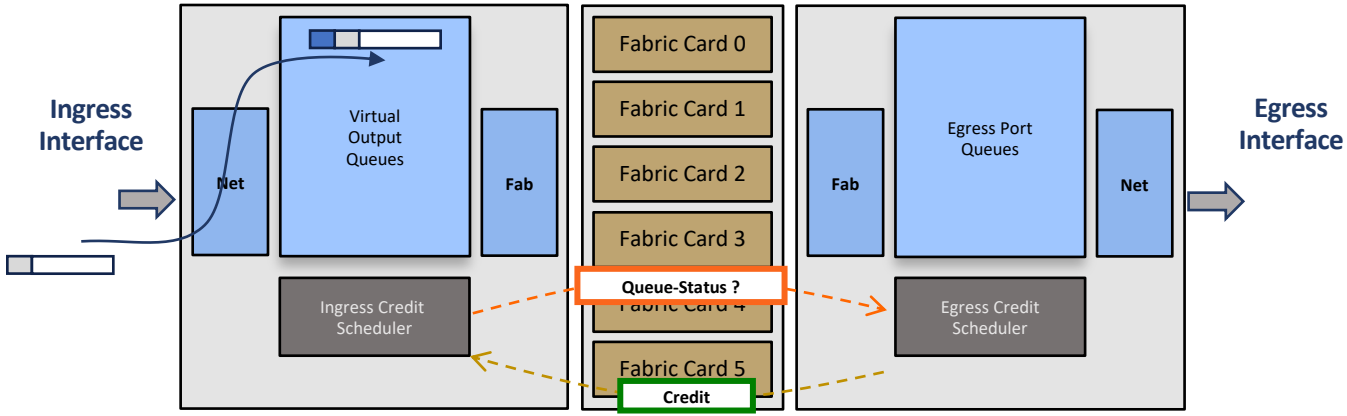
- Queueing is designed to accommodate temporary congestion.
- The packets in the queue are stored in the buffer (memory) until bandwidth becomes available.



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

What is VOQs(Virtual Output Queue)?

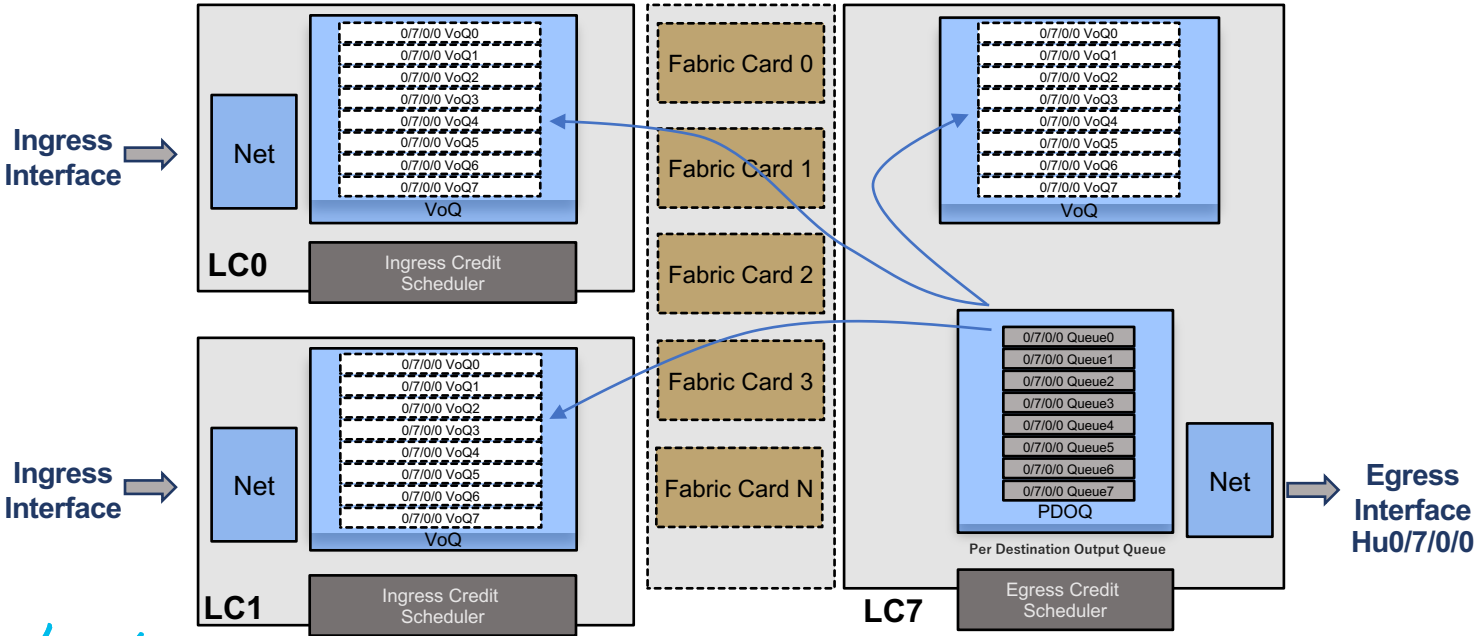
- Cisco 8000 uses an ingress queueing/VOQ architecture, where queues exist on each ingress line card that represent an output destination in the system
- Packets are queued to each VOQ at ingress, and arbitrated across the fabric using a credit grant/request system
- If credit granted, the packets are placed into very shallow output queues that live on the egress line cards.



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

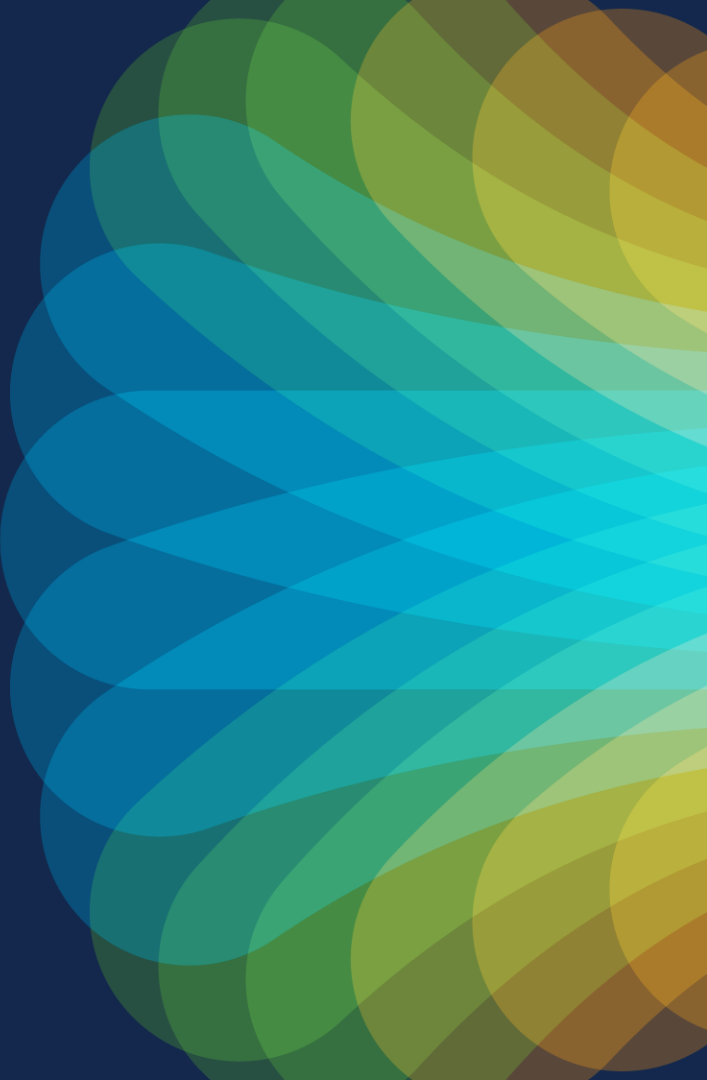
What is VOQs? – contd

- System has 8 output queues(OQs) per physical port
- System replicates set of 8 VOQs for all physical ports in bootup on all Rx slices
  - VOQs represents the OQs at ingress. There's 1:1 mapping between VOQs and OQs.
  - Typically, the traffic-class to VoQ mapping is 1:1.

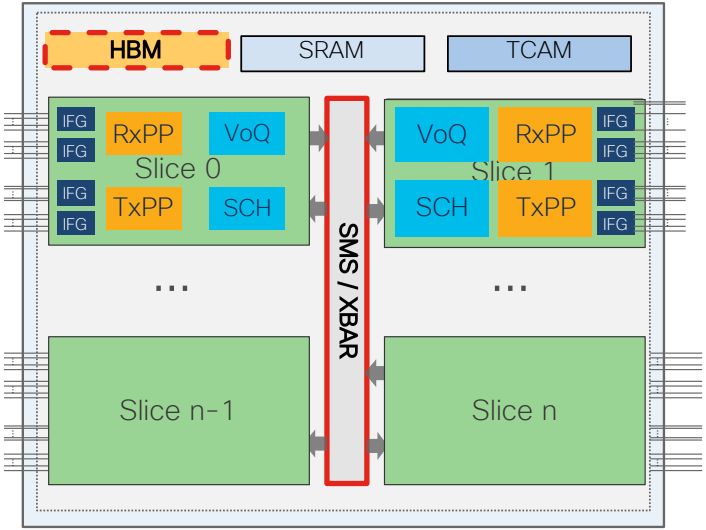


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Packet Memory & Packet Flow



Packet Memory



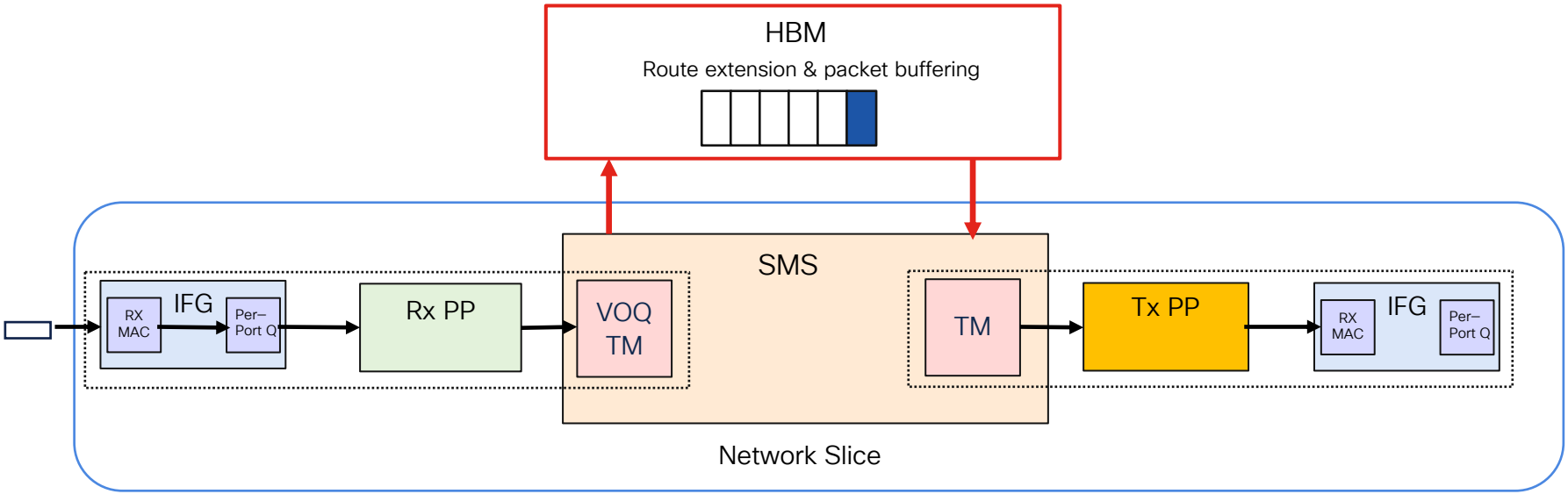
- SMS (Shared Memory System): On-die memory
  - Primary packet memory
  - Fully shared memory
  - Each buffer block can hold 384byte

- HBM (High Bandwidth Memory): Off-die memory
  - 8GB per NPU
  - Secondary packet memory
  - capacity: 1 M buffers [1 buffer=8kbyte]
  - HBM storage is shared between LPM (prefixes storage) and QOS deep buffering
  - HBM buffer block can store a maximum of 16 packets

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

VoQ Congestion Operation

Congestion

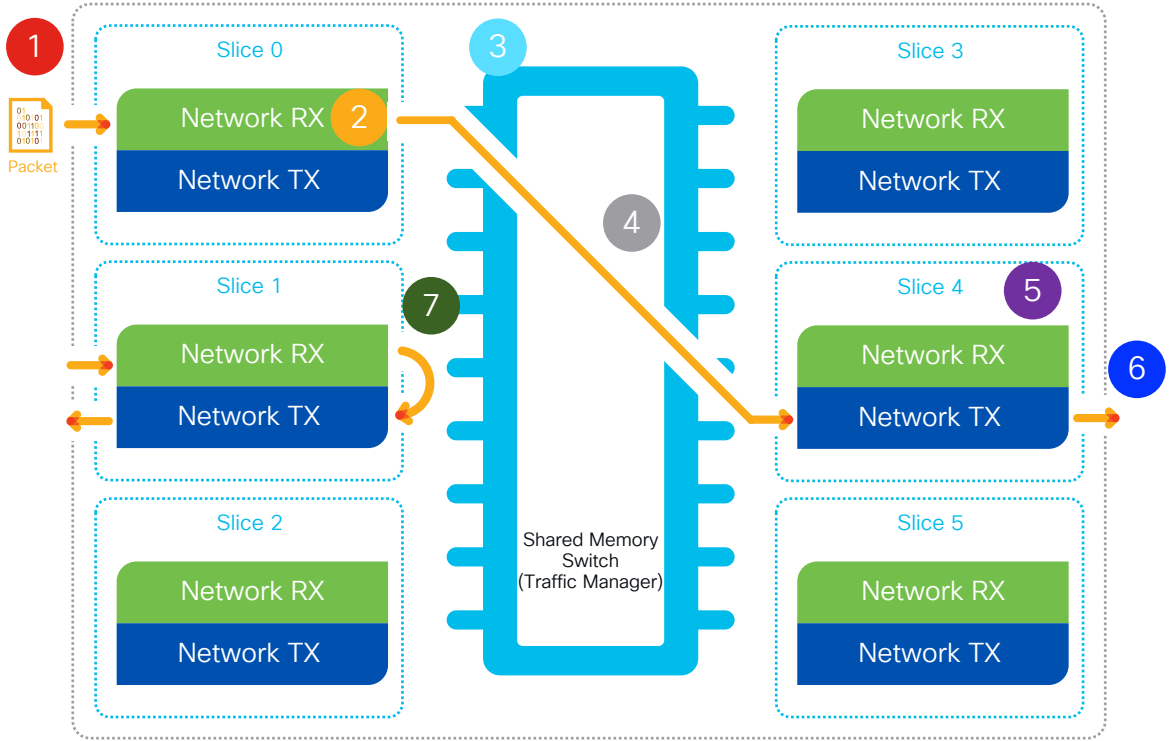


- When the egress port is experiencing congestion, VOQ starts building up in SMS
- VOQs get evicted once the congestion/latency in VOQ increases beyond a certain limit
- VOQs get moved back to the SMS once congestion eases

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Standalone
All slices are network facing

1. Packets arrive on the input slice of the receive device.
2. The NPU performs the packet lookup, feature checks, and identifies the destination for the packet.
3. The ingress NPU enqueues the packet on the corresponding VoQ for the {destination, traffic class}
4. When credits are available for the destination VoQ, the packet is switched across the SMS to the destination (egress) slice.
5. On the egress slice, the transmit direction lookups and feature checks are performed.
6. The final encapsulations are added to the packet and it's transmitted from the TX NPU out the physical interface
7. All packets are switched through the same SMS regardless of whether they are bound for the same NPU slice or a different NPU slice.



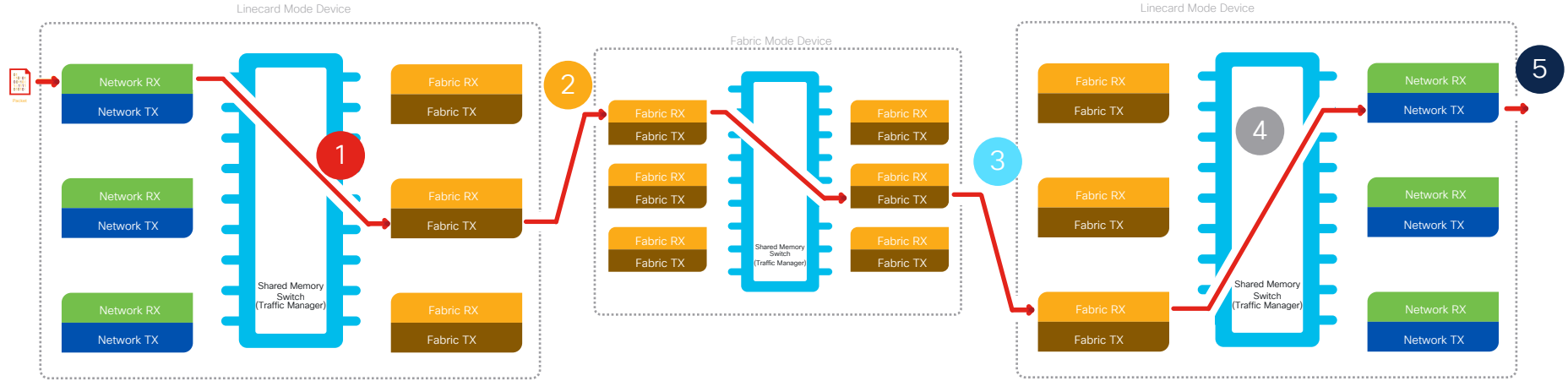
© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Linecard Mode

Some slices network facing and some fabric facing

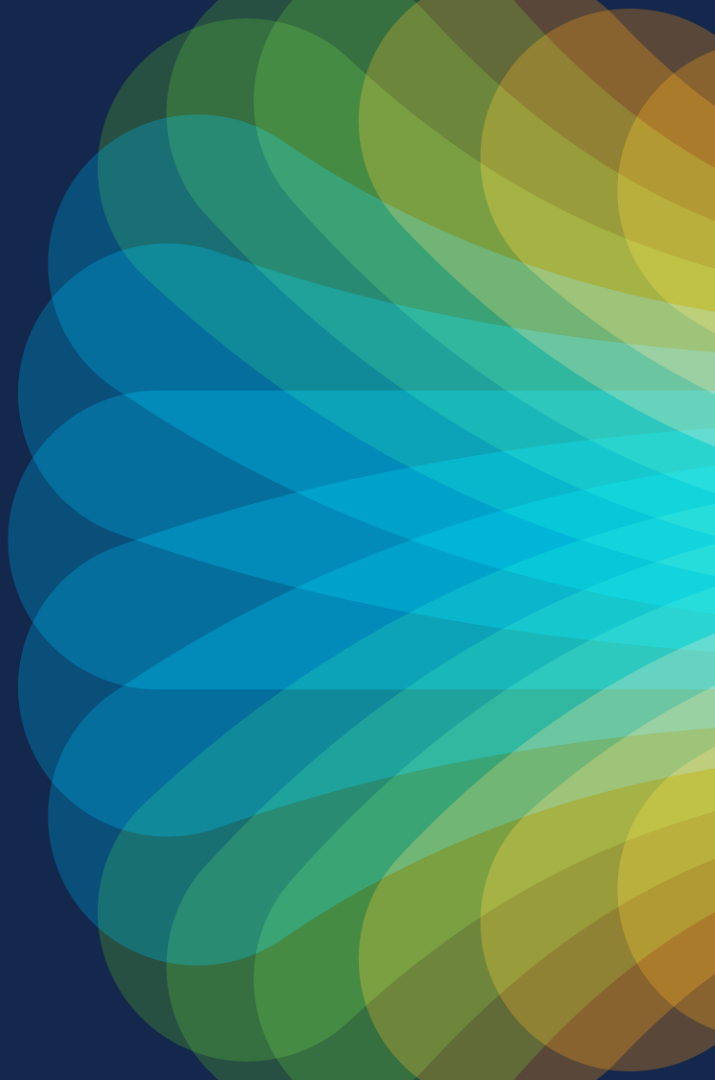
In the system with multiple LCs, the packet processing on RX/TX NPU is basically same. Packets are assigned on the ingress NPU to a VoQ that represents the destination {interface, traffic class}, and credits are still allocated by the TM scheduler in the same manner.

1. Packets are dequeued from the VOQ on the ingress NPU, switched across the local SMS onto a slice that operates in fabric mode.
2. Packets are switched to fabric devices (exist physically on the dedicated fabric cards).
3. Packets are switched to the correct destination NPU based on the address that is encoded in the internal fabric header.
4. Packets are received by the fabric facing slice of the transmit side NPU and switched to the slice of the output interface.
5. Output lookups/features/encapsulation are done in the transmit NPU slice, and the packet is transmitted to the output interface.



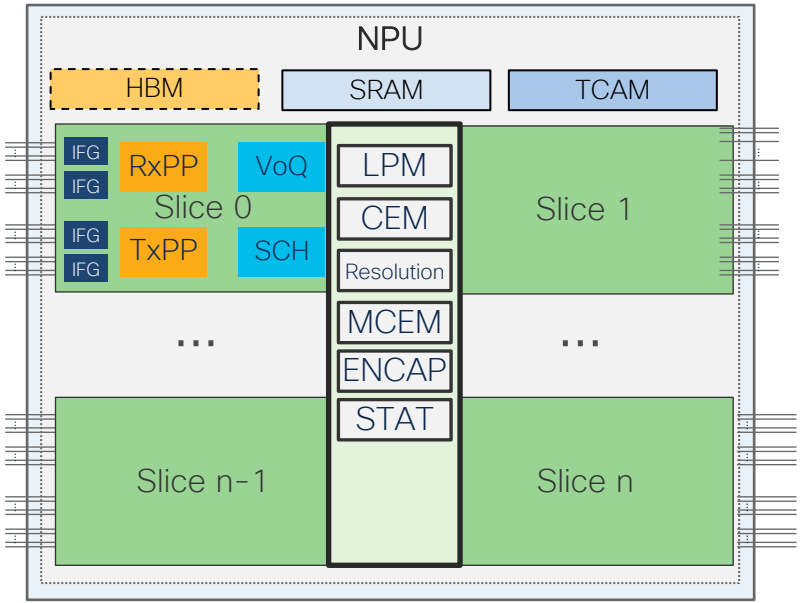
© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Forwarding Resources



Silicon One Memories

- SRAM, TCAM can be used as information memory
- On-die databases used for
  - Route tables: prefixes/ next-hop/ load-balancing
  - Classification / filters
  - Encapsulation DBs
  - Statistics
  - ACLs
- HW resources must be stored in the memory

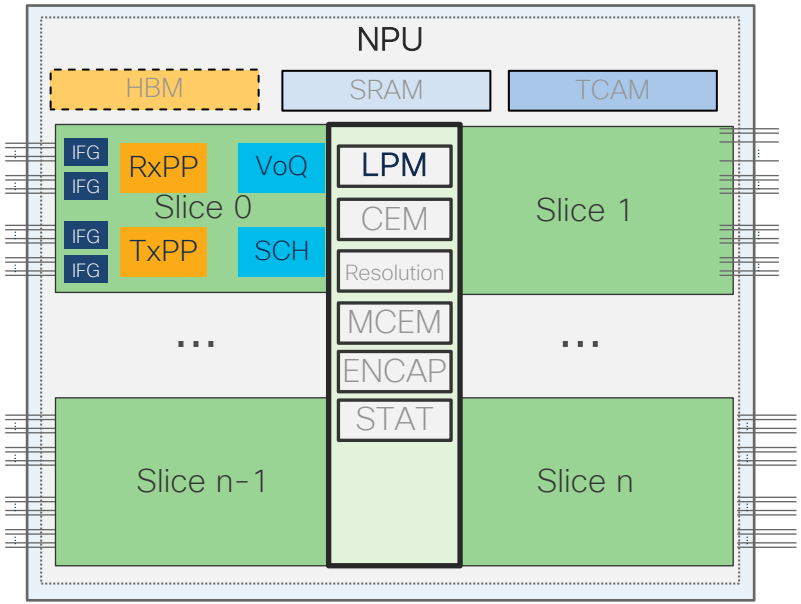


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

LPM/CEM/TCAM/HBM
Mapping applications to memories

- LPM (Longest Prefix Match):
  - IPv4: All prefixes (local, connected, /32)
  - IPv6: All prefixes (except /128)
  - Hybrid ACL 1st look-up

- LPM is built from a combination of TCAM, SRAM, and optional HBM

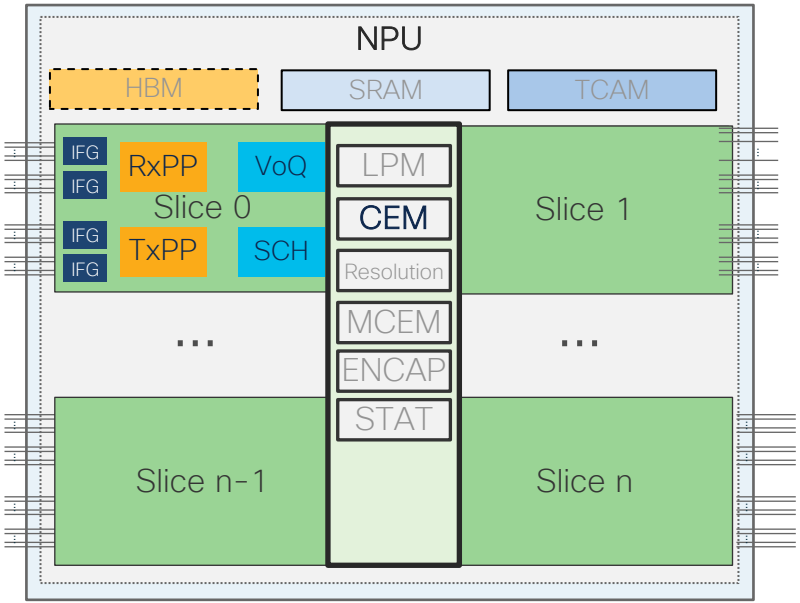


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

LPM/CEM/TCAM/HBM
Mapping applications to memories

- Central EM (Exact Match):

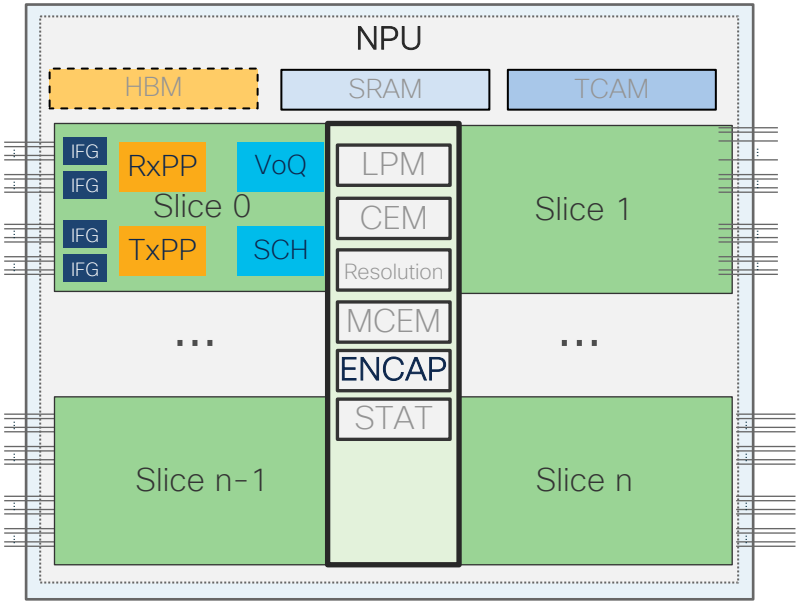
Forwarding information	Details
IPv6	All /128 (/127 & /126 based on configuration)
MPLS local labels	SR/LDP, TE Headend, TE Mid-point
Multicast(v4, v6)	(*,G)/32, (S,G)/64, (*,G)/128, (S,G)/256,
IPv4 Multicast	(*,G)/m (m<32)
IPv6 Multicast	(*,G)/m (m<128)
Hybrid ACL 1st look-up	
LPTS 2nd lookup	
Bridge MAC addresses	



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

LPM/CEM/TCAM/HBM
Mapping applications to memories

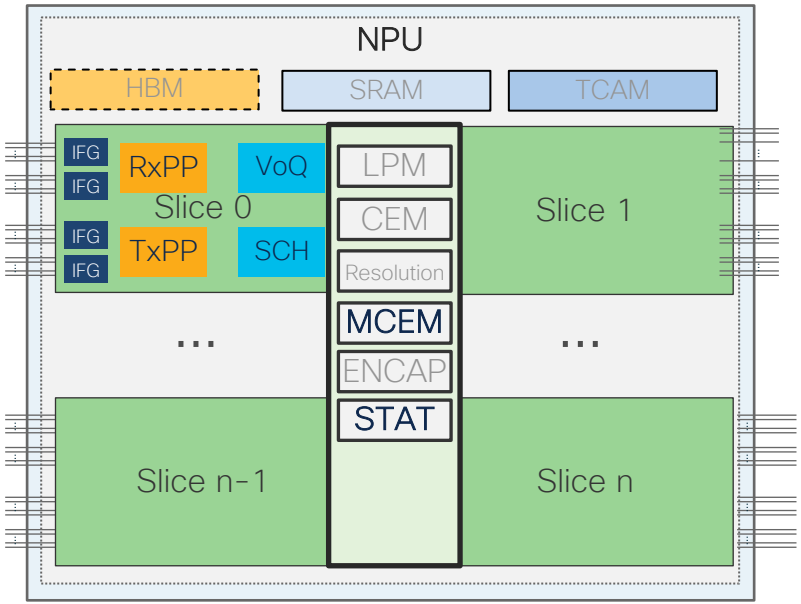
- Egress Large Encap EM (Exact Match):
  - Remote label: MPLS LDP/SR, L3VPN, L2VPN, 6PE VPN, MPLS TE Tunnel, BGP-LU, and SR-TE
- Egress Small Encap EM (Exact Match):
  - Transport label for BGP LU, LDP over TE Tunnel, SR over SR-TE , SR-TE(High scale: If gt 2 stacks), and SRv6 encap



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

LPM/CEM/TCAM/HBM
Mapping applications to memories

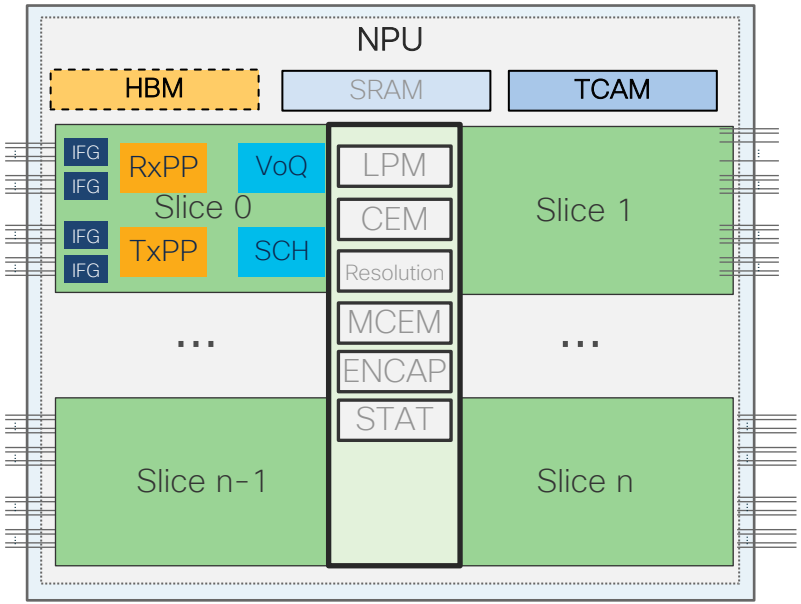
- Multicast EM (MC-EMDB):
  List of OLEs, Multicast route stats, and L2VPN Bridge L2 interface
- Stats:
  Used to store all counters(Forwarding, VoQ, Multicast stats, QoS, Meter)



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

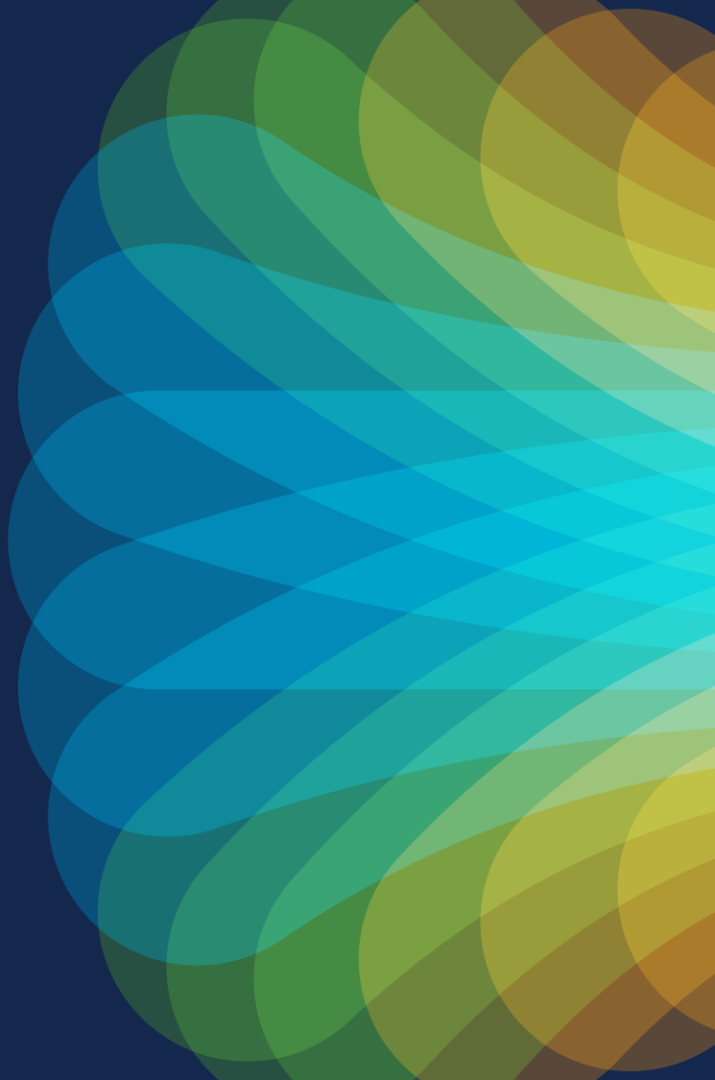
LPM/CEM/TCAM/HBM
Mapping applications to memories

- HBM(High Bandwidth Memory):
  - Deep packet buffer,
  - L2 LPM Prefix as FIB Expansion
- TCAM:
  - Packet classification(Security ACL, QoS ACL, BGP-Flowspec, LPTS),
  - Hybrid ACL



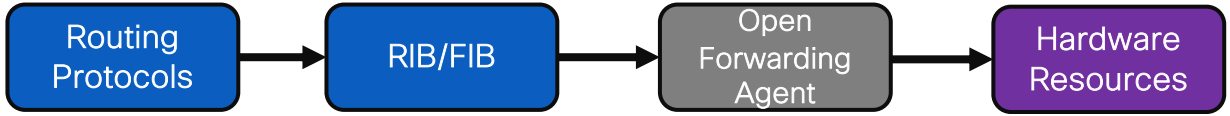
© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Resources



Monitoring Memory Resources
Thresholds Yellow / Red

- Hardware programming is done through an abstraction layer: OFA



- Each database is using two thresholds: Yellow at 80% and Red at 95%(default)

```
RP/0/RP0/CPU0:C8818#show controllers npu resources lpmtcam location 0/0/CPU0
HW Resource Information
    Name                            : lpm_tcam
    Asic Type                       : Q200

NPU-0
OOR Summary
        Estimated Max Entries       : 100
        Red Threshold               : 95 %
        Yellow Threshold            : 80 %
        OOR State                   : Green

### once each thresholds are crossed, syslog warnings are generated

RP/0/RP0/CPU0:Mar 11 14:50:41 : npu_drvr[388]: %PLATFORM-OFA-4-OOR_YELLOW : NPU 0, Table npu, Resource lpm_tcam
RP/0/RP0/CPU0:Mar 11 15:02:12 : npu_drvr[388]: %PLATFORM-OFA-5-OOR_GREEN : NPU 0, Table npu, Resource lpm_tcam
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources

Forwarding Table Resources accounting supported

```
RP/0/RP0/CPU0:C8818#show controllers npu resources ?
  all                       all hardware resources
  centralem                 central_em
  counterbank               counter_bank
  dipindex                  dip_index
  egressacltcam             egress_acl_tcam
  egressl3dlp               egress_l3_dlp
  egresslargeencap          egress_large_encap
  egresssmallencap          egress_small_encap
  ingressacltcam            ingress_acl_tcam
  ipv6compressedsips        Multicast IPv6 Source Addresses
  l2serviceport             L2 Forwarding interface
  l3acport                  L3 Forwarding interface
  lpmtcam                   lpm_tcam
  lptsmeters                lpts_meters
  mcemdb                    Multicast Replication and Route Statistics
  myipv4tbl                 myipv4tbl
  nativefecentry            native_fec_entry
  protectiongroup           protection_group
  sipidxtbl                 sipidxtbl
  stage1lbgroup             stage1_lb_group
  stage1lbmember            stage1_lb_member
  stage2lbgroup             stage2_lb_group
  stage2lbmember            stage2_lb_member
  stage2protectionmonitor   stage2_protection_monitor
  tunneltermination         tunnel_termination
  tunneltermination1        tunnel_termination_1
  v4lpts                    v4_lpts
  v6lpts                    v6_lpts
```

For Exact Match Resource
For Counter Resource

Egress Large Encapsulation
Egress Small Encapsulation

L2 interface related Resource
L3 interface related Resource

Longest Prefix Match Resource
LPTS meters Resource
Multicast OLE Resource
GRE decap resource

GRE encap/decap resource
Recursive ECMP group resource
Recursive ECMP members resource
Non-recursive ECMP group resource
Non-recursive ECMP members resource

GRE Tunnels resource
V4 LPTS TCAM Resource
V6 LPTS TCAM Resource

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources

OOR: show controller npu stats trap counters

- HW resource issue often result in lookup miss for each resource
- Errors like label lookup miss, route lookup miss, adjacency lookup miss etc. are counted by trap counters.
  - Trap counter is stats for punted packets
- show controllers np descriptions to check the meaning of each stats

```
RP/0/RP0/CPU0:C8818#show controllers npu stats traps-all instance all location 0/0/CPU0 | in MISS
ETHERNET_SMAC_MISS(D*)                0    44   LC_CPU      232   1586  0    67      145        IFG     64      0                  0
ETHERNET_MC_SNOOP_LOOKUP_MISS(D*)     0    46   LC_CPU      232   1586  0    67      145        IFG     64      0                  0
MPLS_TE_MIDPOINT_LDP_LABELS_MISS(D*)  0    88   LC_CPU      232   1586  0    67      145        IFG     64      0                  0
MPLS_ILM_VRF_LABEL_MISS(D*)           0    89   LC_CPU      232   1586  0    67      145        IFG     64      0                  0
MPLS_MISSING_FWD_LABEL_AFTER_POP(D*)  0    92   LC_CPU      232   1586  0    67      145        IFG     64      0                  0

RP/0/RP0/CPU0:C8818#show controllers np descriptions location 0/0/CPU0  | inc MISS
44   | ETHERNET_SMAC_MISS                              | Yes  | Ethernet Ingress         |  SMAC of bridged packet is not found in
MAC table.
46   | ETHERNET_MC_SNOOP_LOOKUP_MISS                   | Yes  | Ethernet Ingress         |  IGMP snooping lookup miss.
88   | MPLS_TE_MIDPOINT_LDP_LABELS_MISS                | Yes  | MPLS Egress              |  on egress failed to resolve the midpoint
Traffic-Engineering or LDP (Label Distribution Protocol) label(s) in encapsulation data-base.
89   | MPLS_ILM_VRF_LABEL_MISS                         | Yes  | MPLS Ingress             |  Label lookup in VRF fails on a CsC
(Carrier Supporting Carrier) interface., miss in EM database..
92   | MPLS_MISSING_FWD_LABEL_AFTER_POP                | Yes  | MPLS                     |  Cannot pop and forward when forward label
is missing, terminated label is BOS
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
OOR: show controller npu stats trap counters

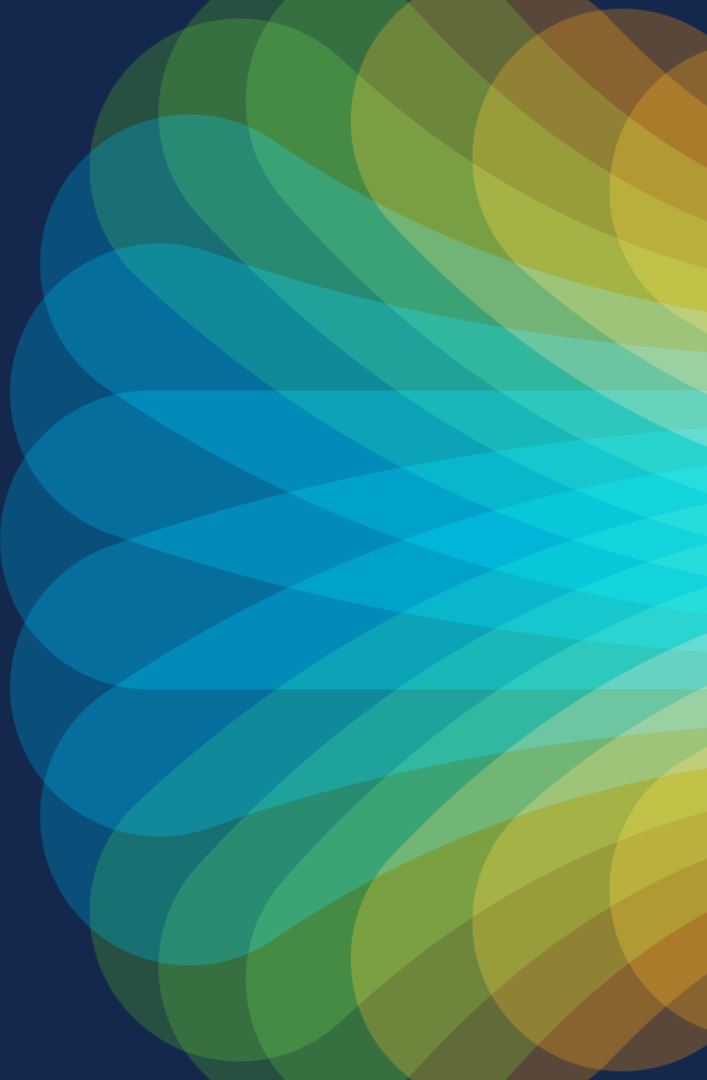
- show captured packet to check the content of dropped packet in trap counter

```
RP/0/RP0/CPU0:C8818#show controllers npu stats traps-all instance all location 0/1/CPU0
Trap Type                       NPU  Trap  Punt       Punt  Punt  Punt Configured Hardware   Policer Avg-Pkt Packets    Packets
                                ID   ID    Dest       VoQ   VLAN  TC   Rate(pps)  Rate(pps)  Level   Size    Accepted   Dropped
=============================================================================================================================
L3_ROUTE_LOOKUP_FAILED          0    129   LC_CPU     233   1538  1    67         145        IFG     64      9293       14198635134

RP/0/RP0/CPU0:C8818#show captured packets traps 129 location 0/1/CPU0
Captured packets are also saved in /harddisk:/nputrap_pcap_traps_129_loc_LC1.txt
***** Packet 1 (len 124) *****
Captured at: 2023-09-12 15:53:35.900

DMAC: 00:00:5E:00:53:01
SMAC: 00:00:5E:00:53:FA

IPv4 header
Internet Header length:20
TOS:0x20
Total length:68
Fragment [ID:0 Is-frag:0 Do-not-frag:0 More-Frags:0 Offset:0]
TTL:64
Protocol:6 (TCP)
Checksum:0x5482
Source: 192.0.2.17
Destination: 198.51.100.1
TCP header
- snip -
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Resources:

LPM/CEM/Adjacency



Monitoring Memory Resources

Forwarding Table resources: LPM(Longest Prefix Match)

```
RP/0/RP0/CPU0:C8818#show controllers npu resources lpmtcam location 0/0/CPU0
HW Resource Information
    Name                            : lpm_tcam
    Asic Type                       : Q200
NPU-0
OOR Summary
        Estimated Max Entries       : 100
        Red Threshold               : 95 %
        Yellow Threshold            : 80 %
        OOR State                   : Green
OFA Table Information
(May not match HW usage)
        iprte                       : 1220484
        ip6rte                      : 7
        ip6mcrte                    : 0
        ipmcrte                     : 0
Current Hardware Usage
    Name: lpm_tcam
        Estimated Max Entries       : 100
        Total In-Use                : 34        (34 %)
        OOR State                   : Green
       Name: v4_lpm
           Total In-Use               : 1220515
       Name: v6_lpm
           Total In-Use               : 12
```

Current IPv4 routes

Current IPv6(Non /128) routes

Total actual In-Use usage % (Directly from HW)

Actual In-Use entries for v4

Actual In-Use entries for v6

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources

Forwarding Table resources: CEM(Central Exact Match)

```
RP/0/RP0/CPU0:C8818#show controllers npu resources centralem location 0/0/CPU0
HW Resource Information
    Name                            : central_em
    Asic Type                       : Q200
NPU-0
OOR Summary
        Estimated Max Entries       : 100
        Red Threshold               : 95 %
        Yellow Threshold            : 80 %
        OOR State                   : Green
OFA Table Information
(May not match HW usage)
        brmac                       : 0
        iprte                       : 14
        ip6rte                      : 102
        mplsrte                     : 20034
        puntlptspolicer             : 96
        ipmcrte                     : 0
        ip6mcrte                    : 0
        aclpolicer                  : 0
Current Hardware Usage
    Name: central_em
        Estimated Max Entries       : 100
        Total In-Use                : 3         (3 %)
        OOR State                   : Green
```

Bridge L2 MAC

IPv4 LPTS, Hybrid ACL

Current IPv6(/128)

MPLS label number

IPv4 multicast routes

IPv6 multicast routes

Total actual In-Use usage %
(Directly from HW)

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
Adjacency(ARP/NDP)

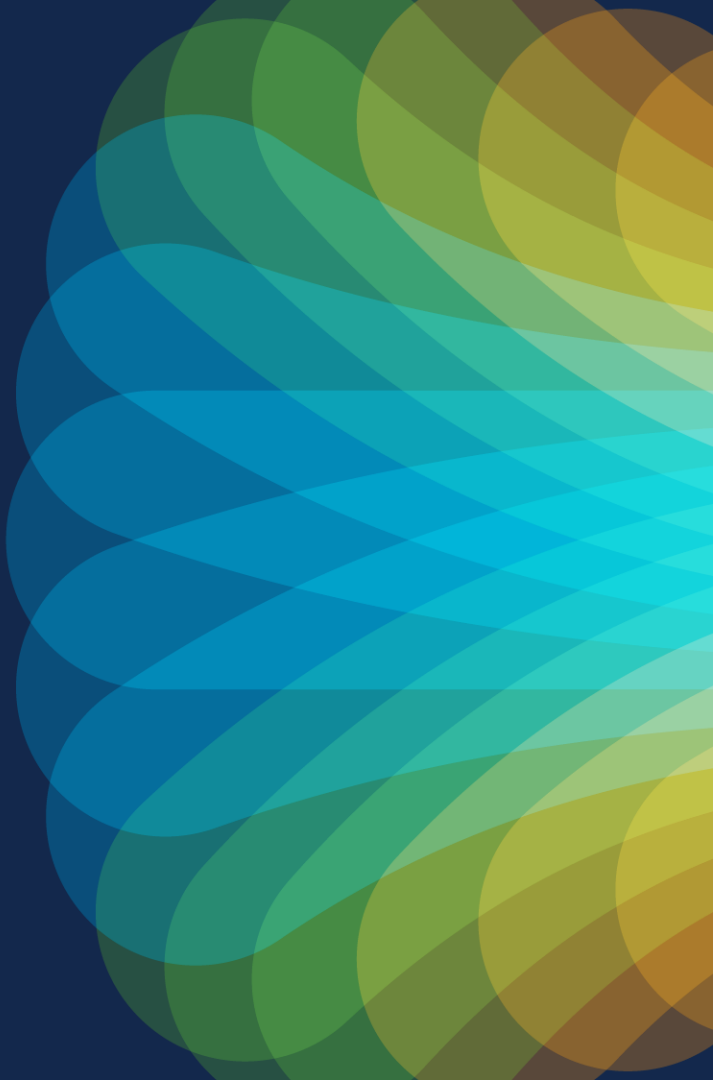
- The encap resource allocation controlled by Global Resource IDs (GRID)
- Adjacency resource can be monitored from show grid command.
- Below is OOR example and GRID Bank Full message will be seen

```
RP/0/RP0/CPU0:8201#show grid pool 2 bank 0
Bank Ptr                        : 0x308d075ae0
Bank ID                         : 0
Pool                            : GLIF (id 2)
Bank Start                      : 2
Bank End                        : 4095
Max Bank Size                   : 4094
Max Resource Pages              : 128
Available resource IDs          : 0 (0.000% free)
Bank statistics:                          Success      Error         (since last clear)
  Resource IDs reserved                     12057    2128757           12057    2128757
  Resource IDs returned                      7963          0            7963          0
Client                          : fib
  Resource IDs reserved                     12057    2128757           12057    2128757
  Resource IDs returned                      7963          0            7963          0
current usage                   : 4094

RP/0/RP0/CPU0:Nov  6 07:21:34.087 UTC: grid_svr[127]: %L2-GRID-4-BANK_FULL : GRID POOL:GLIF(2), BANK 0 FULL. Max size 4091, Curr RIDs 4091.
RP/0/RP0/CPU0:Nov  6 07:21:34.087 UTC: grid_svr[127]: %L2-GRID-4-BANK_MAX_CLIENT : GRID POOL:2,BANK 0 FULL. Max Clt dpa_fib_cl, RIDs 4091.
LC/0/0/CPU0:Nov  6 07:21:34.087 UTC: fib_mgr[206]: %ROUTING-FIB-3-PLATF_UPD_FAIL : FIB platform update failed:
Obj=FIB_DATA_TYPE_NHINFO[ptr=0x309b5128c0,refc=0,flags=0x8000008] Action=CREATE Proto=ipv4. Cerr='RID' detected the 'warning' condition
'GRID': Invalid argument  : fib_mgr : (PID=5713) :  -Traceback= 560471085b90 560471091bd8 5604710893a4 560471089668 5604711efe6f
7f59a2e4708b 7f59a3142fe9 7f59a314cf38 5604710048ab
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

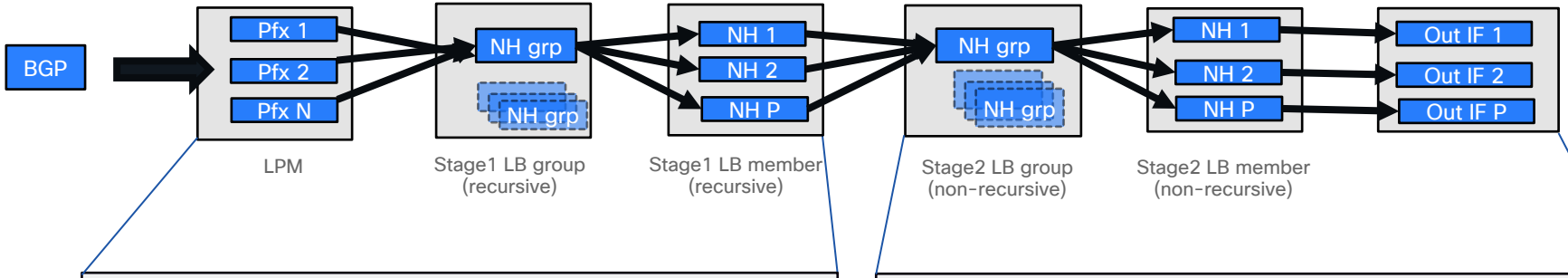
Monitoring Resources:

ECMP Group and Member



ECMP Group and Member

- ECMP is equal cost multi path. Formed by 2 or more next-hops for same destination with same cost.
- With ECMP, additional information needed to be managed on what interfaces are in same ECMP group




```
RP/0/RP0/CPU0:C8818#show route 10.0.0.0/24

Routing entry for 10.0.0.0/24
  Known via "bgp 65000", distance 200, metric 0, type internal
  Installed Oct 11 00:00:32.753 for 1d05h
  Routing Descriptor Blocks
    20.0.0.1, from 20.0.0.1, BGP multi path
      Route metric is 0
    20.0.0.2, from 20.0.0.2, BGP multi path
      Route metric is 0
  No advertising protos.
```

```
RP/0/RP0/CPU0:C8818#show route 20.0.0.1/24

Routing entry for 20.0.0.0/24
  Known via "ospf 1", distance 110, metric 51, type inter area
  Installed Oct 11 00:00:10.767 for 1d05h
  Routing Descriptor Blocks
    30.1.0.2, from 30.1.0.2, via Bundle-Ether100
      Route metric is 51
    30.2.0.2, from 30.2.0.2, via Bundle-Ether100.1
      Route metric is 51
    30.3.0.2, from 30.3.0.2, via Bundle-Ether100.2
      Route metric is 51
    30.4.0.2, from 30.4.0.2, via Bundle-Ether100.3
      Route metric is 51
  No advertising protos.
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources

Forwarding Table resources: stage1_lb_group

```
RP/0/RP0/CPU0:C8818#show controllers npu resources stage1lbgroup
location 0/0/CPU0
HW Resource Information
    Name                            : stage1_lb_group
    Asic Type                       : Q200
NPU-0
OOR Summary
        Estimated Max Entries       : 16384
        Red Threshold               : 95 %
        Yellow Threshold            : 80 %
        OOR State                   : Green
OFA Table Information
(May not match HW usage)
        nhgroup                     : 205
Current Hardware Usage
    Name: stage1_lb_group
        Estimated Max Entries       : 16384
        Total In-Use                : 7         (0 %)
        OOR State                   : Green
```

Total Max entries of Recursive ECMP Group

Actual In-Use of Recursive ECMP Group

- stage1_lb_group is recursive ecmp group to next hop resolution mainly related to BGP path
- ECMP with unique set of nexthops will consume 1 stage1_group entries.

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
Forwarding Table resources: stage1_lb_member

```
RP/0/RP0/CPU0:C8818#show controllers npu resourc stage1lbmember
location 0/0/CPU0
HW Resource Information
    Name                            : stage1_lb_member
    Asic Type                       : Q200
NPU-0
OOR Summary
        Estimated Max Entries       : 196640
        Red Threshold               : 95 %
        Yellow Threshold            : 80 %
        OOR State                   : Green
OFA Table Information
(May not match HW usage)
        nhgroup                     : 213
Current Hardware Usage
    Name: stage1_lb_member
        Estimated Max Entries       : 196640
        Total In-Use                : 10239     (5 %)
        OOR State                   : Green
```

Total Max entries of Recursive ECMP Members

Actual In-Use of Recursive ECMP Members

- stage1_lb_member is recursive ecmp member to next hop resolution mainly related to BGP path
- In stage1_lb_group entries, if 1 prefix has 4 equal cost path/next-hop, 4 stage1_lb_member entries is consumed
- Q200: stage1_lb_member includes stage0. It means both lb members & prefix objects is included as well.
    - Hence the big difference on scale number between Q100(16k) and Q200(192k)

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources

Forwarding Table resources: stage2_lb_group

```
RP/0/RP0/CPU0:C8818#show controllers npu resourc stage2lbgroup
location 0/0/CPU0
HW Resource Information
    Name                            : stage2_lb_group
    Asic Type                       : Q200
NPU-0
OOR Summary
        Estimated Max Entries       : 16384
        Red Threshold               : 95 %
        Yellow Threshold            : 80 %
        OOR State                   : Green
OFA Table Information
(May not match HW usage)
        nhgroup                     : 470
Current Hardware Usage
    Name: stage2_lb_group
        Estimated Max Entries       : 16384
        Total In-Use                : 24        (0 %)
        OOR State                   : Green
```

Total Max entries of ECMP Group

Actual In-Use of Non-recursive ECMP Group

- stage2_lb_group is non-recursive ecmp group to next hop resolution mainly related to IGP path

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
Forwarding Table resources: stage2_lb_member

```
RP/0/RP0/CPU0:C8818#show controllers npu resources stage2lbmember
location 0/0/CPU0
HW Resource Information
    Name                            : stage2_lb_member
    Asic Type                       : Q200
NPU-0
OOR Summary
        Estimated Max Entries       : 32800
        Red Threshold               : 95 %
        Yellow Threshold            : 80 %
        OOR State                   : Green
OFA Table Information
(May not match HW usage)
        nhgroup                     : 540
Current Hardware Usage
    Name: stage2_lb_member
        Estimated Max Entries       : 32800
        Total In-Use                : 52        (0 %)
        OOR State                   : Green
```

Total Max entries of ECMP Members

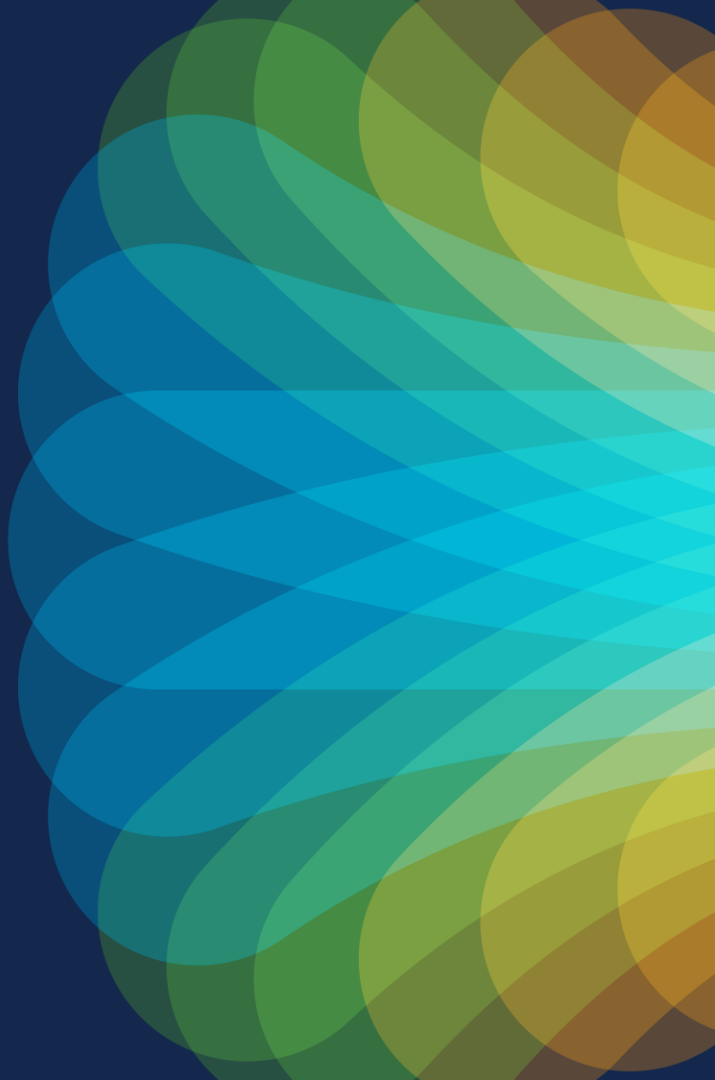
Actual In-Use of Non-recursive ECMP Members

- stage2_lb_member is non-recursive ecmp member to next hop resolution mainly related to IGP path
- In stage2_lb_**group** entries, if 1 prefix has 4 equal cost path/next-hop, 4 stage2_lb_member entries is consumed

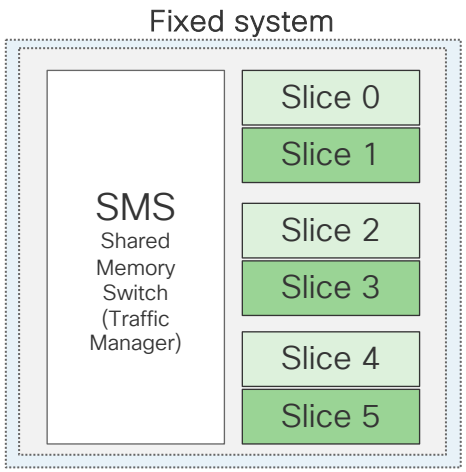
© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

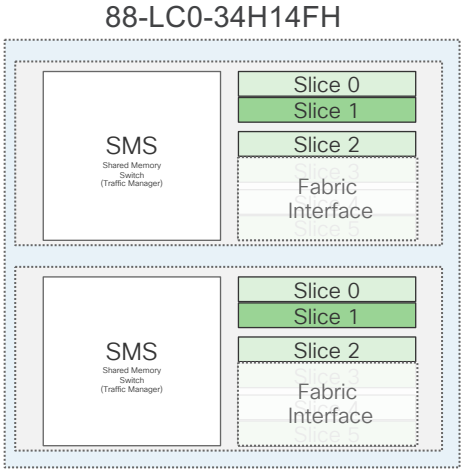
Monitoring Resources:

MPLS labels



Slices & Slice-Pair





	Fixed	Modular
# of NPU	1/system	2/Line Card
# of Slices/NPU	6(Network I/F)/NPU	3(Network I/F) + 3(Fabric I/F)/NPU
# of Slice Pair/NPU	3/NPU	1.5/NPU
# of Slice Pair/LC	3/System	1.5 x 2/LC

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Verify NPU/VOQ/Slice/IFG for a given interface

```
RP/0/RP0/CPU0:C8818#show controllers npu voq-usage interface all instance all location 0/0/CPU0

--------------------------------------------------------------------
Node ID: 0/0/CPU0
Intf         Intf     NPU Slice IFG  Sys   VOQ   Flow   VOQ    Port
name         handle    #    #    #   Port  base  base   port   speed
             (hex)                                    type
--------------------------------------------------------------------
- snip -
FH0/0/0/11   168       0   0    0    48   1496      0 local   400G
FH0/0/0/13   170       1   2    1    56   1560      0 local   400G
FH0/0/0/14   178       1   2    0    60   1568      0 local   400G
FH0/0/0/15   180       1   2    0    64   1576      0 local   400G
FH0/0/0/16   188       1   1    1    68   1584      0 local   400G
```

- If interface belongs to slice # 0,1 then its slice-pair 0
- If interface belongs to slice # 2,3 then its slice-pair 1
- If interface belongs to slice # 4,5 then its slice-pair 2

- Fixed systems has 3 slice-pairs with 6 n/w slices & Distributed systems will have 2 slice-pairs with 3 n/w slices (2 n/w slices belong to 1 slice-pair + 1 n/w slice belongs to another 1 slice-pair.)

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
Forwarding Table resources: Egress Large Encap

```
RP/0/RP0/CPU0:C8818#show controllers npu resources egresslargeencap location 0/1/CPU0
HW Resource Information
    Name                            : egress_large_encap
    Asic Type                       : Q200
NPU-0
OOR Summary
        Estimated Max Entries       : 131136
        Red Threshold               : 95 %
        Yellow Threshold            : 80 %
        OOR State                   : Green


OFA Table Information
(May not match HW usage)
        l2if                        : 0
        pweport                     : 0
        iptnlencap                  : 0
        lspnh                       : 98954
        spansession                 : 0
        puntpolicer                 : 149
```

32k per slice-pair(Q100)
64K per slice-pair(Q200)

MPLS labels

ERSPAN session

PUNT policer

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
Forwarding Table resources: Egress Large Encap

```
Continued

Current Hardware Usage
    Name: egress_large_encap
        Estimated Max Entries       : 131136
        Total In-Use                : 10371    (7 %)
        OOR State                   : Green


       Name: slice_pair_0
           Total In-Use               : 10191
           OOR State                  : Green


       Name: slice_pair_1
           Total In-Use               : 180
           OOR State                  : Green
```

Total-in-use for all slice-pairs combined

Overall OOR state logically combined from all slice-pairs

Slice-pair0 OOR state

- Large encapsulation table is mainly used to lookup for MPLS out label(MPLS LDP/SR, L3VPN, 6VPE, MPLS TE, BGP-LU)
- In this example, Hu0/1/0/10 is Out IF for LDP forwarding. Hu 0/1/0/10 belongs to NPU0 and slice-pair0.

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
Forwarding Table resources: Egress Small Encap

```
RP/0/RP0/CPU0:Cisco8800#sh controllers npu resources egresssmallencap location 0/2/CPU0

HW Resource Information
    Name                            : egress_small_encap
    Asic Type                       : Q200

NPU-0
OOR Summary
        Estimated Max Entries       : 65600
        Red Threshold               : 95 %
        Yellow Threshold            : 80 %
        OOR State                   : Green


OFA Table Information
(May not match HW usage)
        l2macstatic                 : 0
        lspnh                       : 1
        redirectvrf                 : 0
Continued
```

32K per Slice-pair
64K per NPU

Transport Label for BGP LU

- Small encapsulation table is used for Transport Label for BGP LU, LDP over TE Tunnel, SR over SR-TE, SR-TE and SRv6
- Shared with L2 Mac Static table and Redirect VRF for 6PE VPNs

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
Forwarding Table resources: Egress Small Encap

```
Continued

Current Hardware Usage
    Name: egress_small_encap
        Estimated Max Entries       : 65600
        Total In-Use                : 3         (0 %)
        OOR State                   : Green

       Name: slice_pair_0
           Total In-Use               : 2
           OOR State                  : Green


       Name: slice_pair_1
           Total In-Use               : 1
           OOR State                  : Green
```

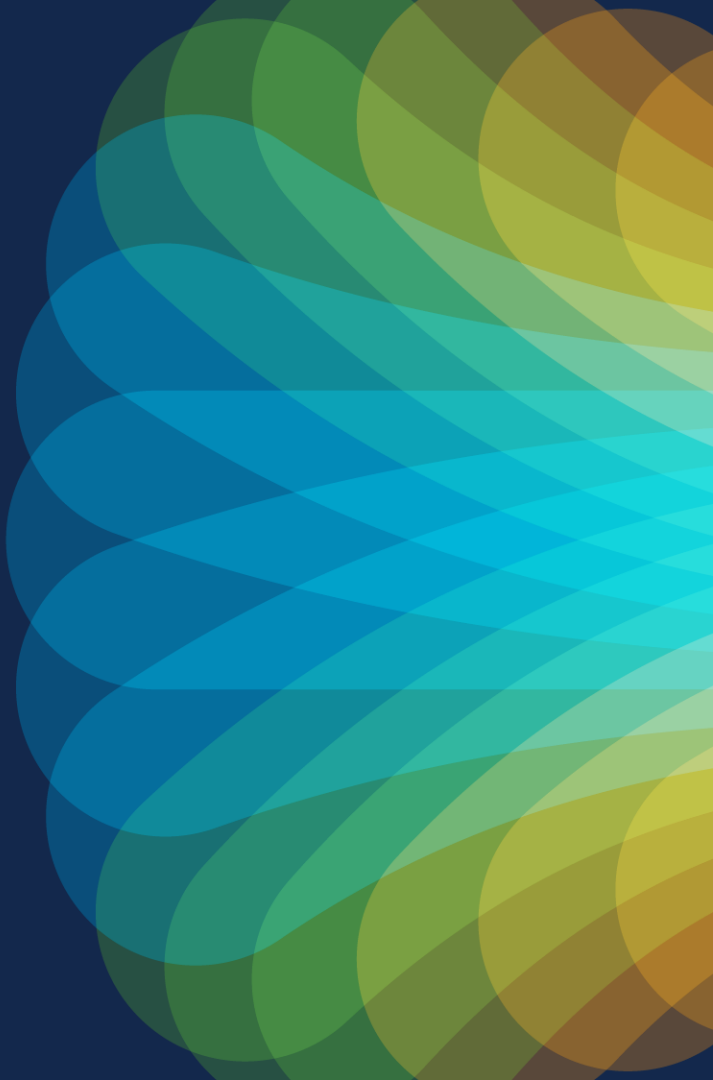
Total-in-use for all slice-pairs combined

Overall OOR state logically combined from all slice-pairs

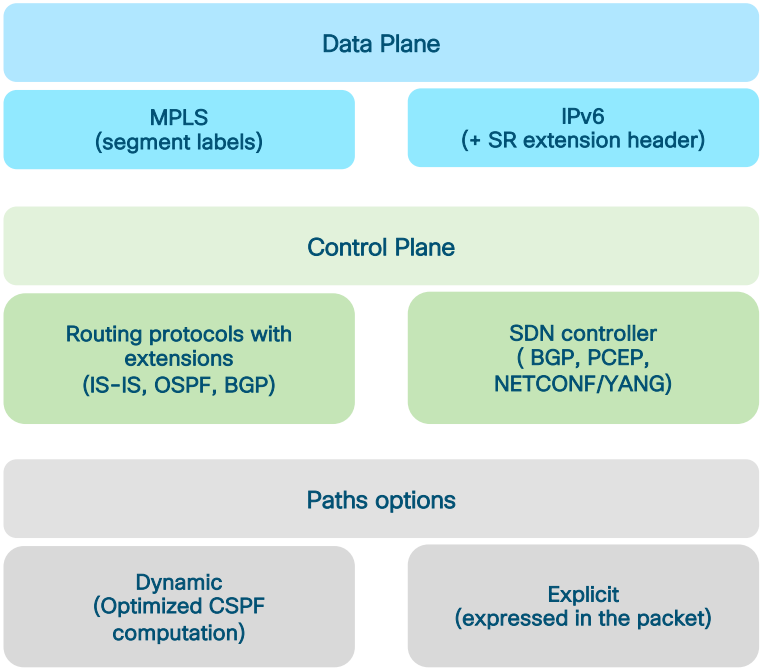
© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Resources:

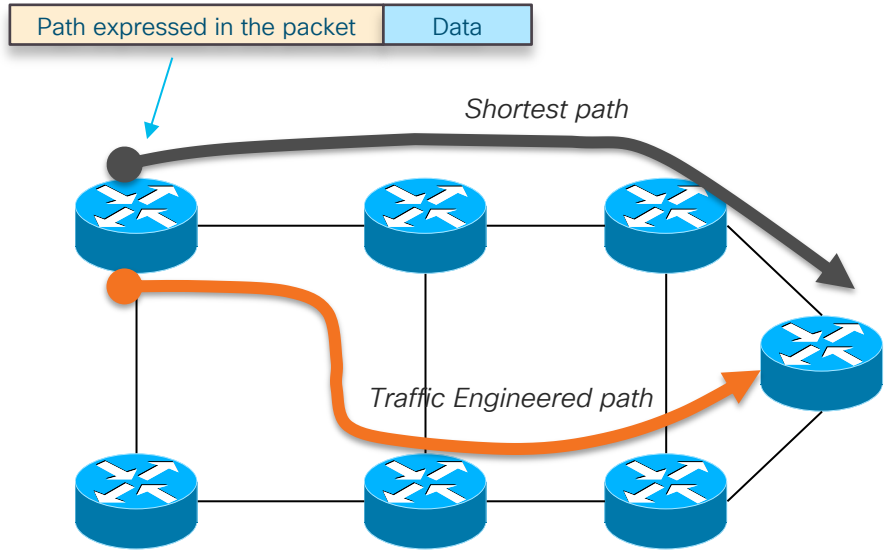
SRv6



Segment Routing

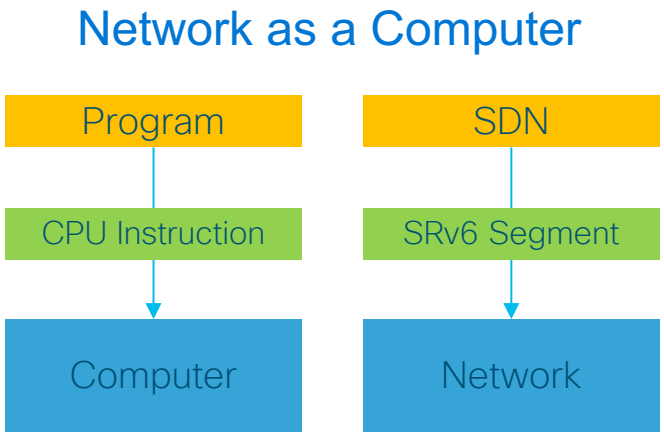


- Source Routing paradigm
  - Stateless IP fabric



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

SRv6 on Cisco8000



Feature	Release introduced
SRv6 Traffic Engineering	IOS-XR 7.10.1
SRv6 Traffic Accounting	IOS-XR 7.10.1
SRv6 Services: L3 EVPN	IOS-XR 7.9.1
SRv6 Services: Services with Remote SIDs from W-LIB	IOS-XR 7.9.1
L3 EVPN/SRv6 and L3VPN/MPLS Interworking Gateway	IOS-XR 7.9.1
L3 EVPN/SRv6 and L3 EVPN/MPLS Interworking Gateway	IOS-XR 7.9.1
SRv6 Services: IPv4 L3VPN	IOS-XR 7.8.1
SRv6 Services: IPv6 L3VPN	IOS-XR 7.8.1

- SRv6 is supported on Q200-based line cards and fixed-port routers but not on Q100
- Related scale numbers:
  - Max-SIDs : 24K
  - Max Locators : 16
  - Number of uSID blocks : 256 [00-ff]

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
Forwarding Table resources: tunnel_termination

```
RP/0/RP0/CPU0:C8818#show controllers npu resources tunneltermination location 0/0/CPU0
HW Resource Information
    Name                            : tunnel_termination
    Asic Type                       : Q200
NPU-0
OOR Summary
        Estimated Max Entries       : 4128
        Red Threshold               : 95 %
        Yellow Threshold            : 80 %
        OOR State                   : Green
Current Hardware Usage
    Name: tunnel_termination
        Estimated Max Entries       : 4128
        Total In-Use                : 271       (6 %)
        OOR State                   : Green
       Name: slice_0
           Total In-Use               : 271
           OOR State                  : Green
```

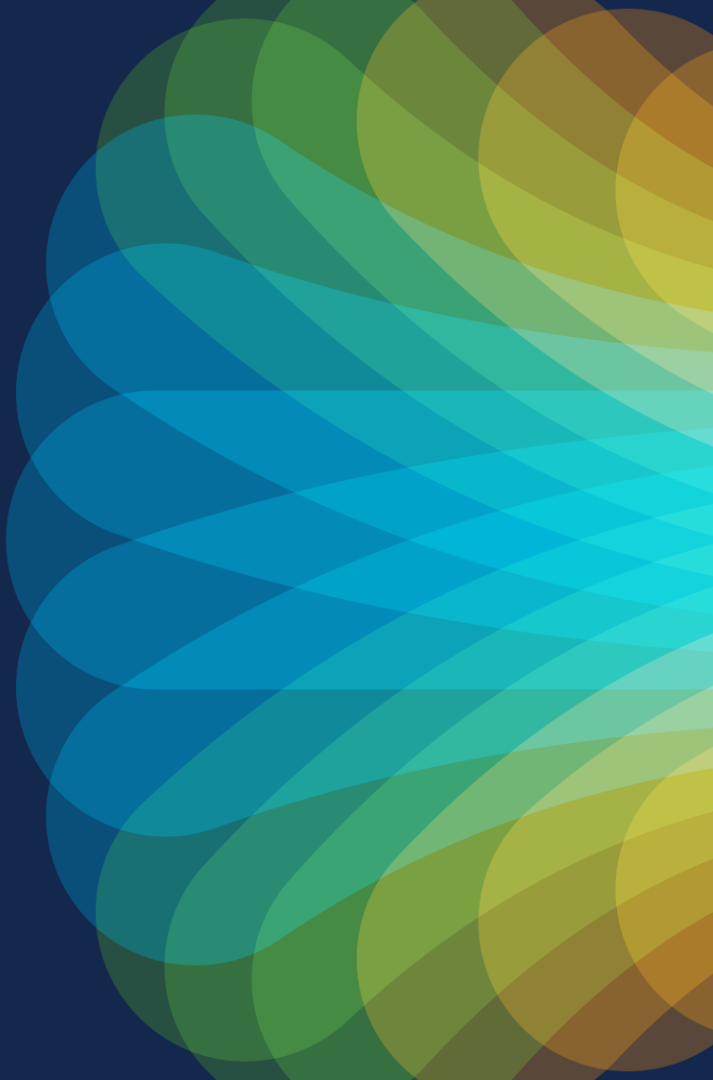
4k per NPU

Shared between GRE/IPinIP/SRv6

- SRv6 forwarding is closely related to pure IPv6 forwarding, and HW resource is basically same
  - LPM for IPv6 prefix, Small Encap for SRv6 encap
- Tunnel_termination is also unique HW resource to manage SRv6 function
  - End.DT4/DT6 SID (F1 & uSID /64) consume 1 tunnel_termination entry

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

ECMP/Encap Optimization



ECMP resource: Cisco8000 vs NCS5500

- Forwarding resource for ECMP need careful consideration and testing as it's usually sparse resource
- NCS5500 with Jericho/Jericho+ NPU has 4K resource(32k for Jericho2) as encap id for ECMP (ECMP FEC)
  - It's per-path resource for each unique prefix for MPLS
  - E.g. 1 prefix with 2 paths consumes 2 resource
  - SR FIB optimizations in-place but some caveats
- Cisco8000 has more resource. Q100 has 8K, Q200 has 16k
  - Furthermore, it's per-prefix resource. 1 prefix with 2 paths consumes 1 resource

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Optimization for ECMP resource

- IGP (ISIS) level, we can restrict the number of ECMP with knob “maximum-paths 1”
  - maximum-path 1 can pick different next-hops randomly for every prefix for LB perspective
  - Think about if the ECMP is really needed
- IP labeled / MPLS LSP ECMP / BGP LU, L3 VPN multipath is common abusers for ECMP resource
  - Use LDP filters to reduces the local label allocation only to the remote PE loopbacks
  - Use allocate-label to restrict label allocation only to important prefixes for BGP LU
  - Use per-vrf label allocation on all the remote PE nodes.
- Prefer Bundles to ECMP when it’s possible

```
RP/0/RP0/CPU0:8201(config)#mpls ldp address-family
ipv4 label local allocate for ?
  WORD         IP access-list
  host-routes  Allocate label for host routes only
```

```
RP/0/RP0/CPU0:8201(config-bgp-af)#allocate-label ?
  all           Allocate labels for all prefixes
  route-policy  Use a route policy to select prefixes for label allocation

RP/0/RP0/CPU0:8201(config-bgp-af)#label mode ?
  per-ce        Set per CE label mode
  per-vrf       Set per VRF label mode
  route-policy  Use a route policy to select prefixes for label allocation
mode
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Optimization for Encap resource
LDP / SR (Segment Routing)

- LDP programs 2 types of labels
  - Local label which every LDP enabled node allocates for all prefixes in routing table
    - These labels are stored in CEM table
  - Remote label which is received from remote peer node
    - These labels are stored in large encap table
- A prefix can have multiple out labels (remote labels) based on multiple paths/nexthop
  - So Large encap is constructed by prefix & associated path/nexthop
    - Ex:- A prefix with 5 ecmp paths will have 5 remote labels and HW program 5 encap entries
- Encap get programmed in the corresponding slice-pair based on the exit interface which is pointing the nexthop
- If the path is over bundle interface with members on all slice-pairs then same encap entry get programmed on all slice-pairs
- For SR, egress encap programming is same as LDP

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Optimization for Encap resource
Best Practice: Prefixes Filtering

- One way to limit number of MPLS labels is… to limit the number of prefixes
- For IGP
  - ISIS: advertise passive-only
  - OSPF: prefix-suppression
- For BGP
  - Should a given PE have full view of tens of thousands PEs spread across multiple continents or states?
  - Sometimes yes, sometimes no.
  - Allocate BGP communities, filter BGP routes, to install only what's relevant
- Prefer Bundles to ECMP
  - When possible (some operators prefer independent L3 links to facilitate isolation or for TE)
  - Bundles are more efficient - single encap entry is used for multiple members of the same bundle

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Optimization for Encap resource
Best Practice: LDP/BGP MPLS Label Allocation

- By default, LDP will allocate a label for every single prefixes
  Most of time, only loopacks are required as typically LDP is for MPLS forwarding within local NW.
  Best practice is to control label allocation scheme
- LDP Label advertisement can be controlled by label local advertise config.

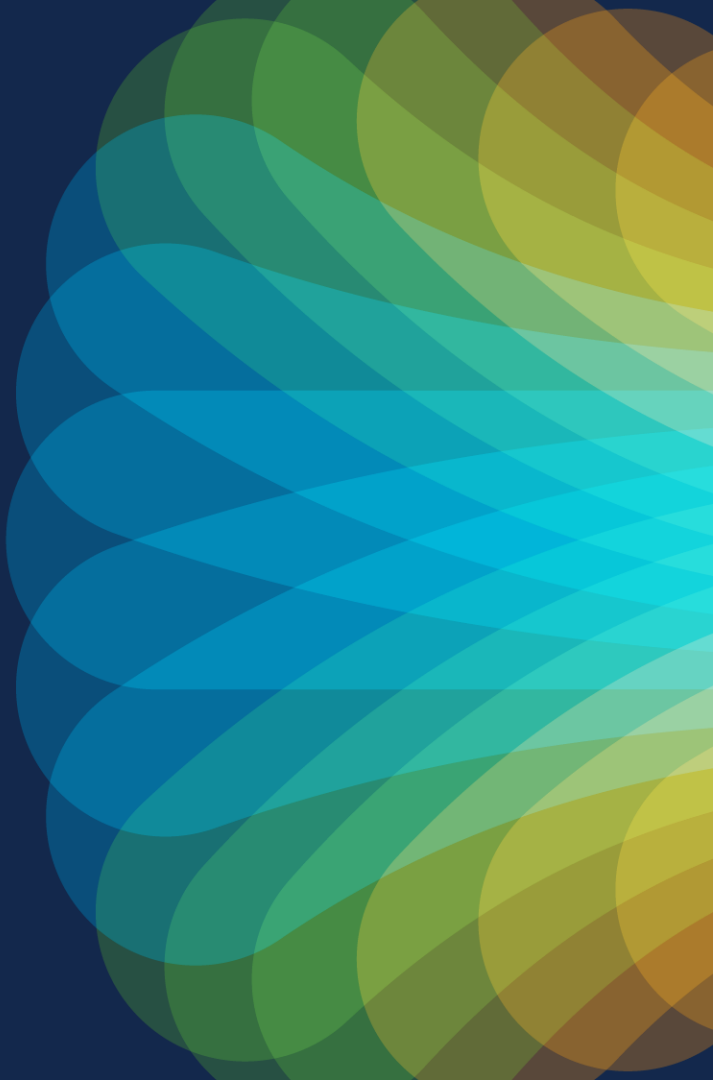
```
RP/0/RP0/CPU0:8818(config-ldp-af)#label local advertise ?
  disable        Disable label advertisement to all peers for all prefixes
  explicit-null  Configure explicit-null advertisement
  for            Prefix centric outbound label filtering
  interface      Advertise interface host address
  to             Peer centric outbound label filtering
```

- BGP label allocation can be controlled by allocate-label config.

```
RP/0/RP0/CPU0:8818(config-bgp-af)#allocate-label ?
  all           Allocate labels for all prefixes
  route-policy  Use a route policy to select prefixes for label allocation
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

VOQ Scale & Counter Bank



VOQ scale

- VOQ scale can be checked with “show controllers npu voq-usage resource-manager” command.
- 8 VOQ get consumed per interface with 8 VOQ mode (default)

```
RP/0/RP0/CPU0:C8818#show controllers npu voq-usage resource-manager instance all location 0/0/CPU0

--------------------------------
Node ID: 0/0/CPU0
IFG         VOQS         VOQS
            USED         FREE
--------------------------------
NPU 0:
IFG00         16           48

### adding egress queueing policy to Hu0/0/0/10.1 (Hug0/0/0/10 belongs to NPU0/IFG0)

RP/0/RP0/CPU0:C8818#show controllers npu voq-usage resource-manager instance all location 0/0/CPU0

--------------------------------
Node ID: 0/0/CPU0
IFG         VOQS         VOQS
            USED         FREE
--------------------------------
NPU 0:
IFG00         24           40
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

4 VOQ mode

- 4xVOQ mode mainly introduced to meet the requirements of higher scale with less queues
- Configure “hw-module profile qos voq-mode 4/8” is used to change between 8xvoq & 4xvoq modes.
  - system reload is required to get the mode effective
- Up to 4 priority level can be configured
- VOQ scaling will also depend on free counters available in counter bank. 4 VOQ mode also reduce the counter usage

VOQ usage in 8 VOQ mode

```
--------------------------------
Node ID: 0/0/CPU0
IFG          VOQS          VOQS
             USED          FREE
--------------------------------
NPU 0:
IFG00          16            48
IFG01          16            48
IFG02          16            48
IFG03          16            48
IFG04          16            48
IFG05          16            48
NPU 1:
IFG00          16            48
IFG01          16            48
IFG02          16            48
IFG03          16            48
IFG04          16            48
IFG05          16            48
- snip -
```

VOQ usage in 4 VOQ mode

```
--------------------------------
Node ID: 0/0/CPU0
IFG          VOQS          VOQS
             USED          FREE
--------------------------------
NPU 0:
IFG00           8            56
IFG01           8            56
IFG02           8            56
IFG03           8            56
IFG04           8            56
IFG05           8            56
NPU 1:
IFG00           8            56
IFG01           8            56
IFG02           8            56
IFG03           8            56
IFG04           8            56
IFG05           8            56
- snip -
```

TC/VOQ mapping in 4xVOQ mode

Traffic-Class	VoQ offset	OQ	Priority
7	7	7	P1
6	6	6	P2
5	5	5	Normal
4	0	0	Normal
3			
2			
1			
0			

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
Forwarding Table resources: counter_bank

```
RP/0/RP0/CPU0:C8818#show controllers npu resources counterbank location 0/0/CPU0
HW Resource Information
    Name                            : counter_bank
    Asic Type                       : Q200

NPU-0
OOR Summary
        Estimated Max Entries       : 108
        Red Threshold               : 95 %
        Yellow Threshold            : 80 %
        OOR State                   : Green

Current Hardware Usage
    Name: counter_bank
        Estimated Max Entries       : 108
        Total In-Use                : 54
        OOR State                   : Green
```

Max 96 Bank for Forwarding(8K) , 12 Banks for Meter/QoS(6K)

Actual In-Use counter banks

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
Forwarding Table resources: counter_bank

- Used counters per user and bank

```
RP/0/RP0/CPU0:C8818#show controllers npu debugshell 0 "script /pkg/bin/print_internal_state.py counter_banks"
location 0/0/CPU0

Logical bank 5, Slice: 1, Num slices: 1, Allowed Users: ['VOQ'], Direction: INGRESS
        Physical bank Id     Slice/Ifg    Total
        50                   1 /0          8192

        Allocations:                      Counter User                      Used
                                          VOQ                                912
- snip -
Logical bank 8, Slice: 4, Num slices: 2, Allowed Users: ['QOS', 'METER'], Direction: INGRESS
        Physical bank Id     Slice/Ifg    Total
        104                  4 /0          6144

        Allocations:                      Counter User                      Used
                                          QOS                                0
                                          METER                              1536
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Monitoring Memory Resources
Forwarding Table resources: counter_bank

- 1 class in policy-map has 3 counters (matched, transmitted & dropped)

```
Class prec0
  Classification statistics          (packets/bytes)      (rate - kbps)
    Matched             :                  0/0                   0
    Transmitted         :                  0/0                   0
    Total Dropped       :                  0/0                   0
```

- 1 policy with 9 classes = 30 counters increased
  (qos counters can only be allocated in class groups of 2, so we round up 9 to 10, hence 30 counters)

```
RP/0/RP0/CPU0:C8818#show controllers npu debugshell 0 "script /pkg/bin/print_internal_state.py counter_banks"
location 0/0/CPU0 | inc QOS
Logical bank 7, Slice: 0, Num slices: 2, Allowed Users: ['QOS', 'METER'], Direction: INGRESS
                                         QOS                                  0

### configure 1 policy-map with 9 classes to Hu0/0/0/10.1

RP/0/RP0/CPU0:C8818#show controllers npu debugshell 0 "script /pkg/bin/print_internal_state.py counter_banks"
location 0/0/CPU0 | inc QOS
Logical bank 7, Slice: 0, Num slices: 2, Allowed Users: ['QOS', 'METER'], Direction: INGRESS
                                         QOS                                  30
```

- 100 policies(attached to sub-interface on the same IFG/Slice) takes 100 * 30 = 3000 counters.

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

VOQ stats sharing example

- VOQ stats sharing mode can be enabled to allocate more counters for the VOQs
- Configure “hw-module profile stats voqs-sharing-counters 1/2/4” to change system to different stats mode

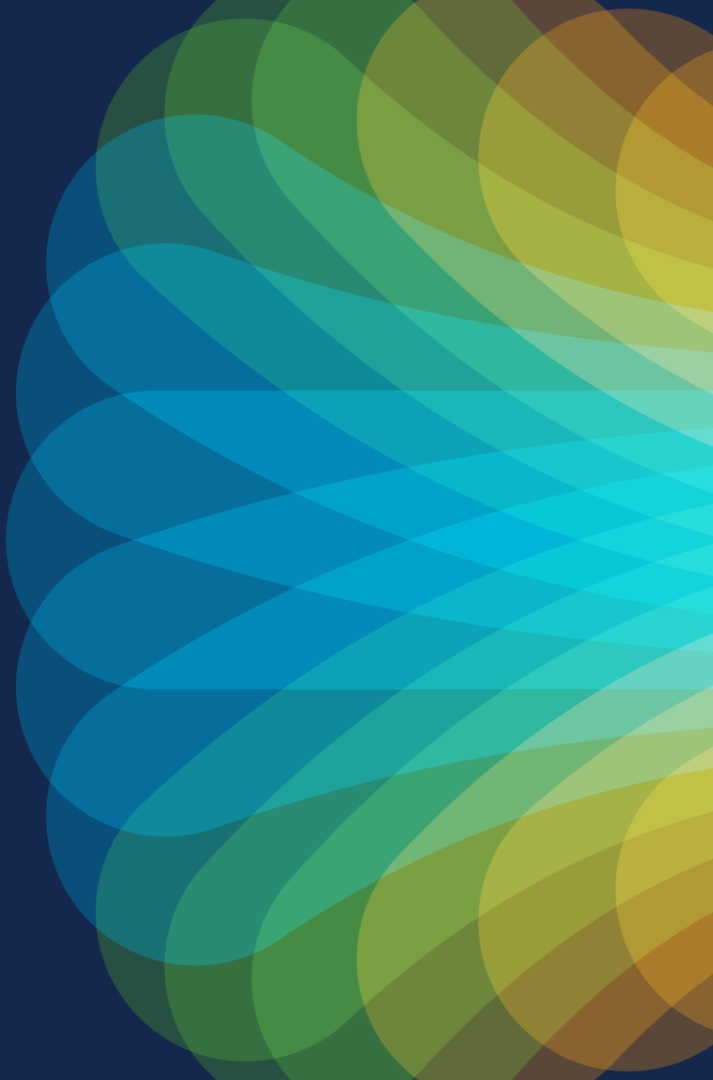
  1 is default mode

  System reload is needed to get the new profile applied
- In the below 2 sharing mode example, one consolidated stats entry for 2 VOQs

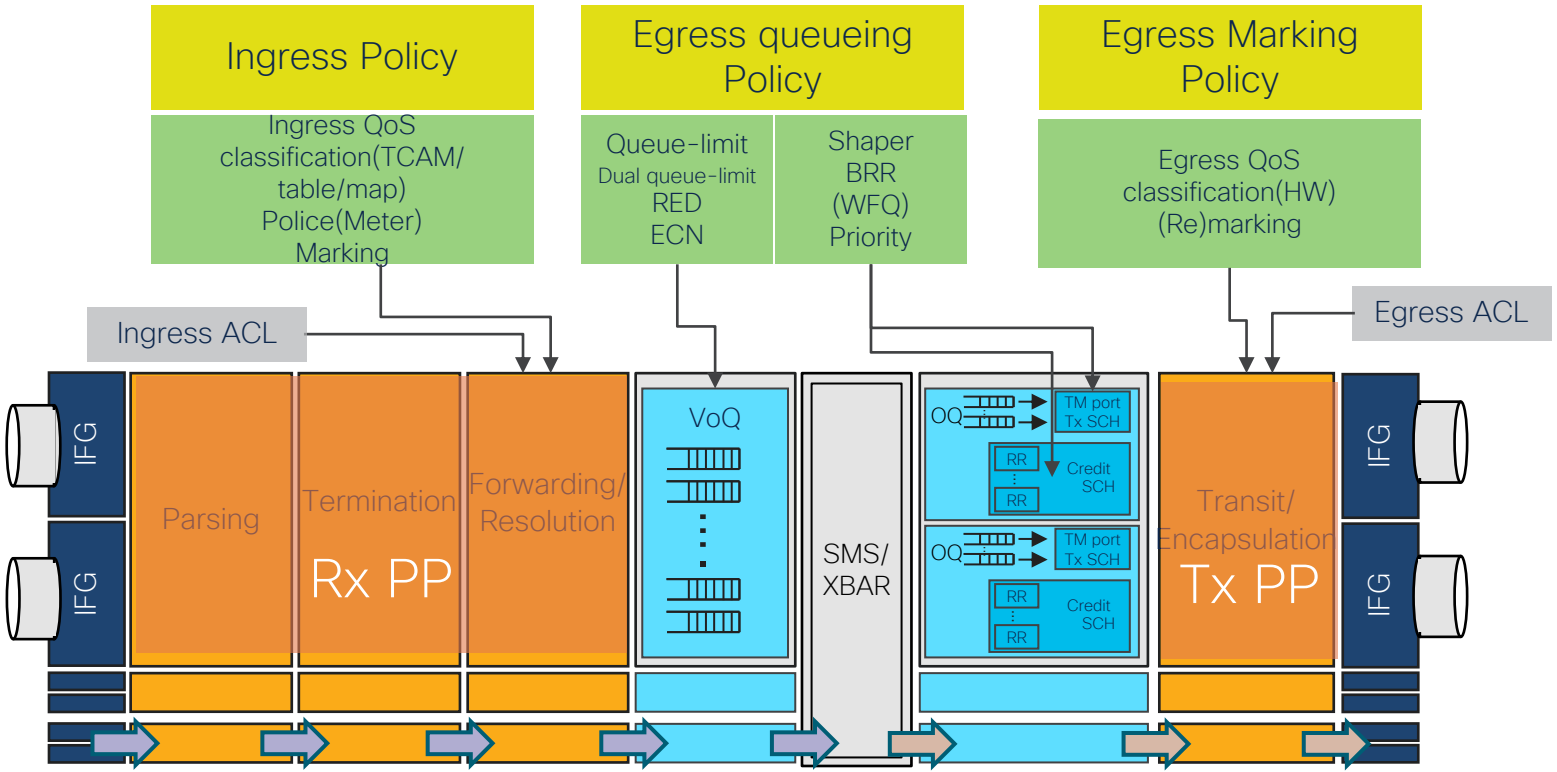
```
RP/0/RP0/CPU0:C8818#show controllers npu stats voq ingress interface all instance all location 0/1/CPU0

Interface Name          =   Te0/0/0/12
Interface Handle        =          160
Location                =     0/1/CPU0
Asic Instance           =            0
VOQ Base                =         1552
Port Speed(kbps)        =     10000000
Local Port              =       remote
VOQ Mode                =            8
Shared Counter Mode     =            2
       ReceivedPkts    ReceivedBytes   DroppedPkts     DroppedBytes
-------------------------------------------------------------------
TC_{0,1} = 37600624        3233653664      2213717150      190379674900
TC_{2,3} = 0               0               0               0
TC_{4,5} = 0               0               0               0
TC_{6,7} = 303902899       26135649314     854991851       73529299186
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Config Model of QOS

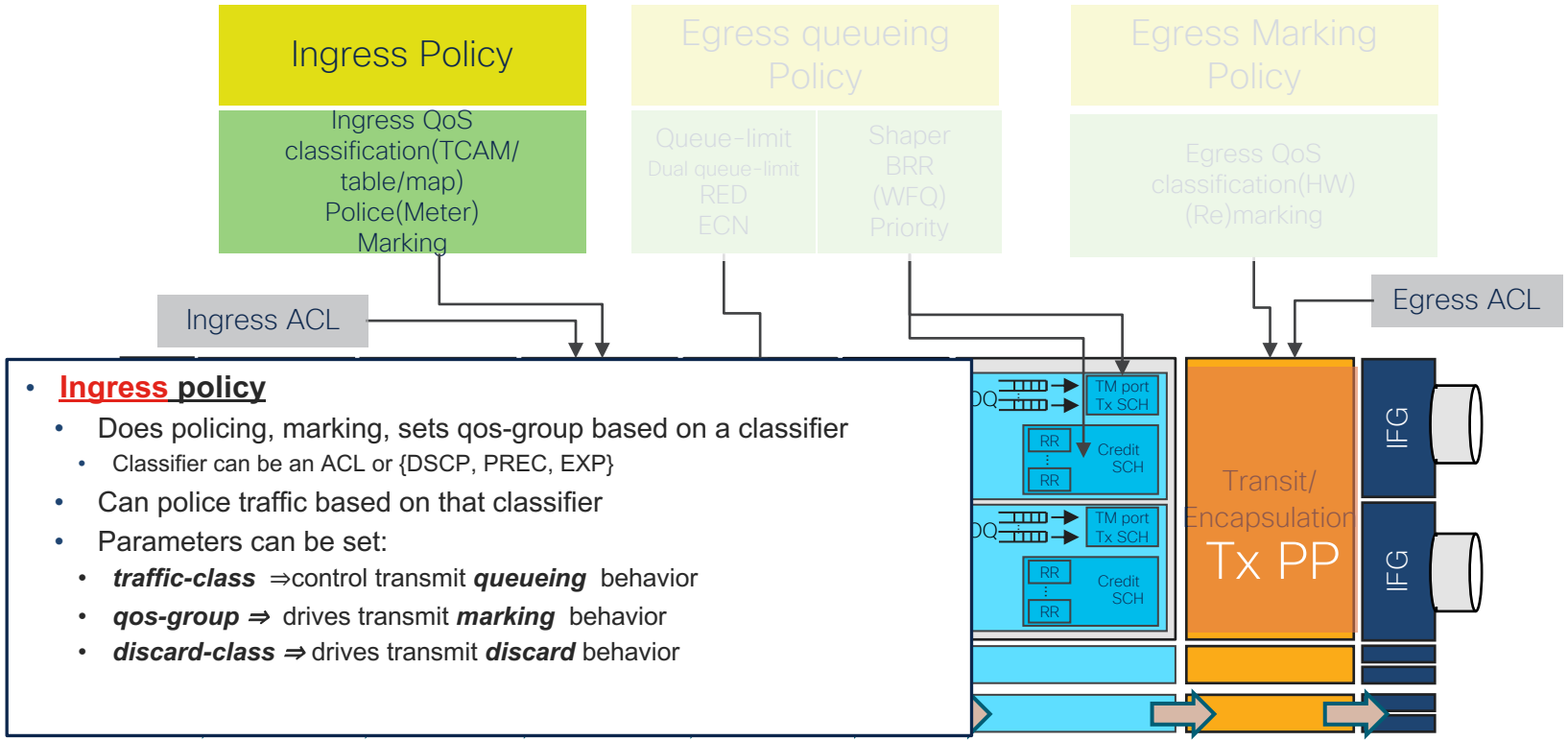


8000 QoS: 3 policy config model



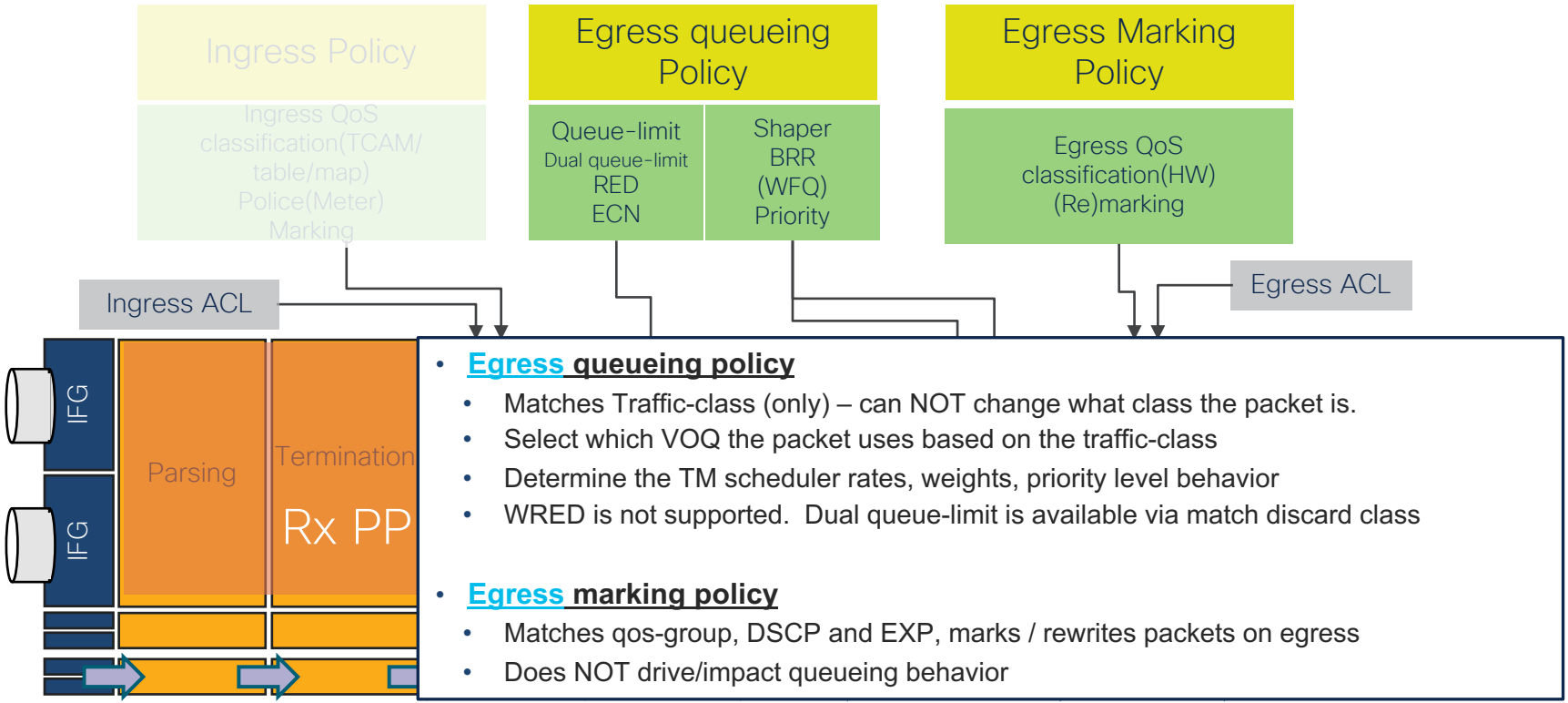
© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

8000 QoS: 3 policy config model



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

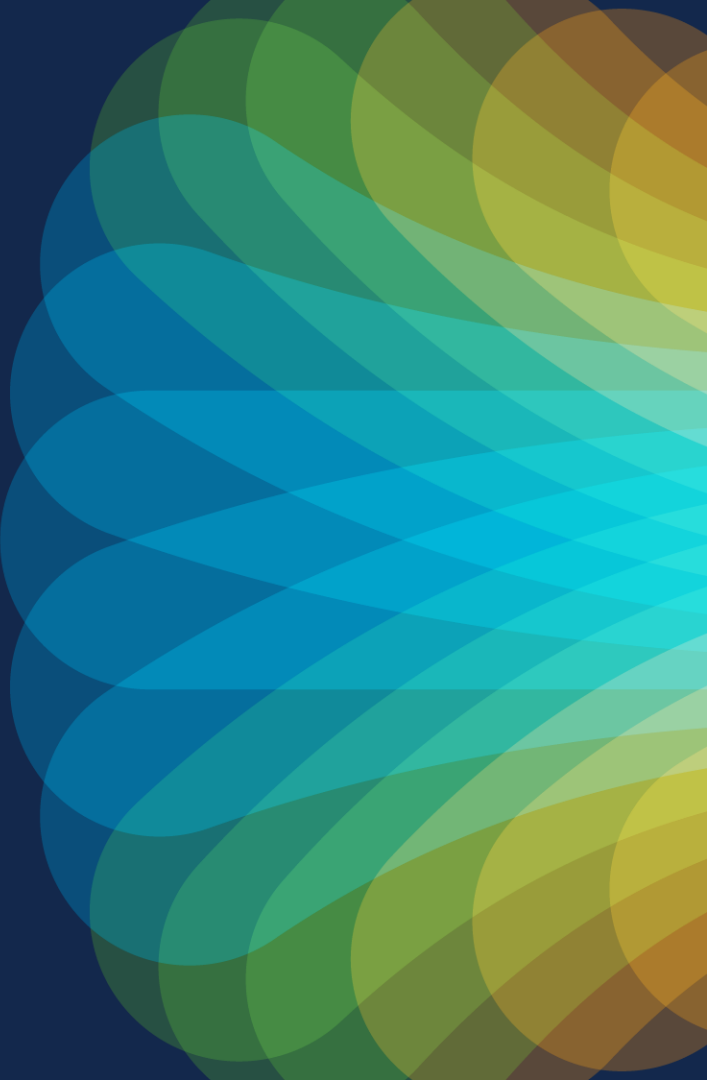
8000 QoS: 3 policy config model



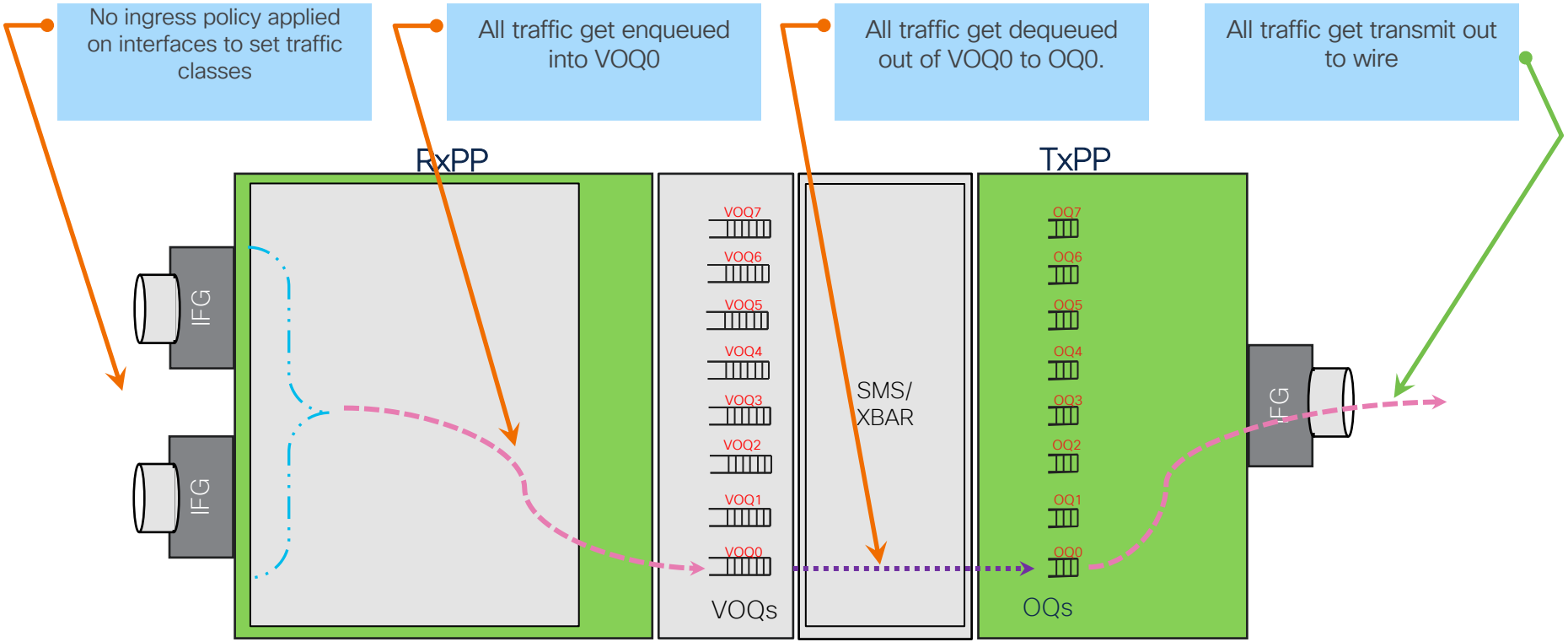
- **Egress queueing policy**
  - Matches Traffic-class (only) – can NOT change what class the packet is.
  - Select which VOQ the packet uses based on the traffic-class
  - Determine the TM scheduler rates, weights, priority level behavior
  - WRED is not supported. Dual queue-limit is available via match discard class
- **Egress marking policy**
  - Matches qos-group, DSCP and EXP, marks / rewrites packets on egress
  - Does NOT drive/impact queueing behavior

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Scheduling & Queueing



Queueing behavior
Default



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Queueing behavior

Traffic classes marked at ingress

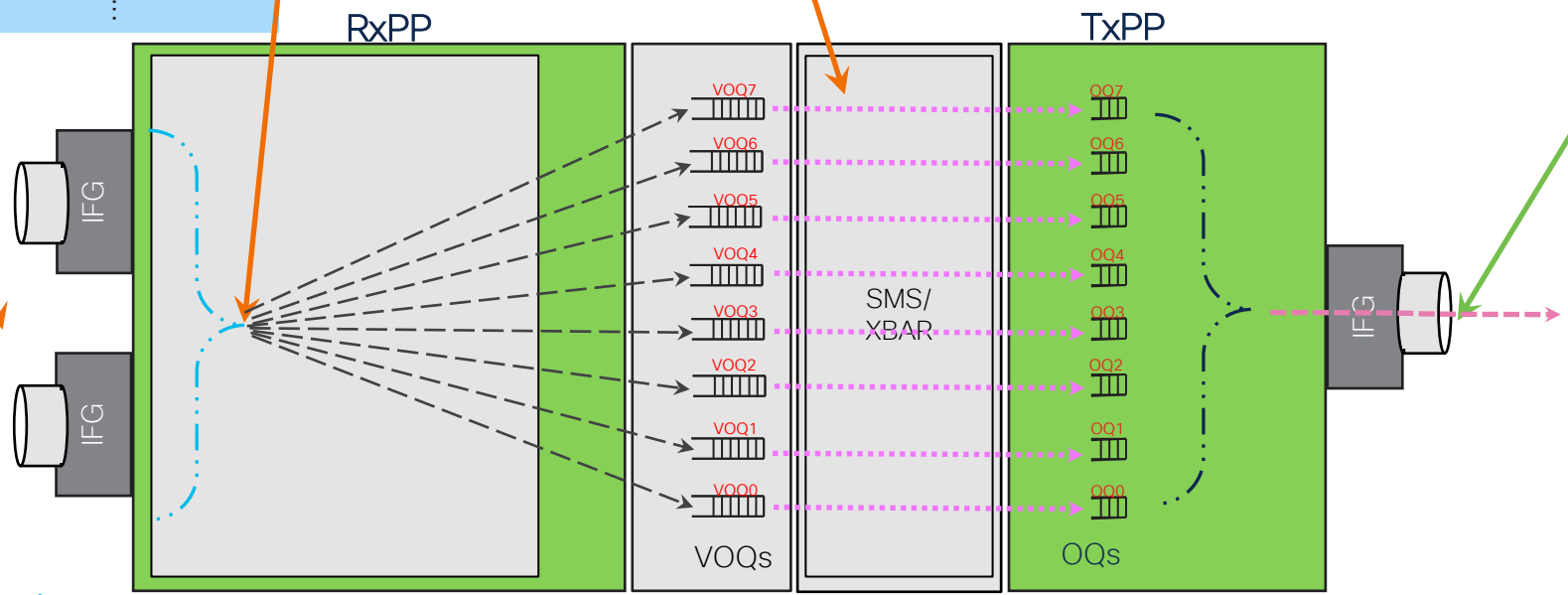
Ingress policy applied on interfaces:
Ex:

```
class tc1
   set traffic-class 1
class tc2
   set traffic-class 2
```

TC0 – TC7 traffic get enqueued into VOQ0 – VOQ7 correspondingly because TCs are marked at ingress

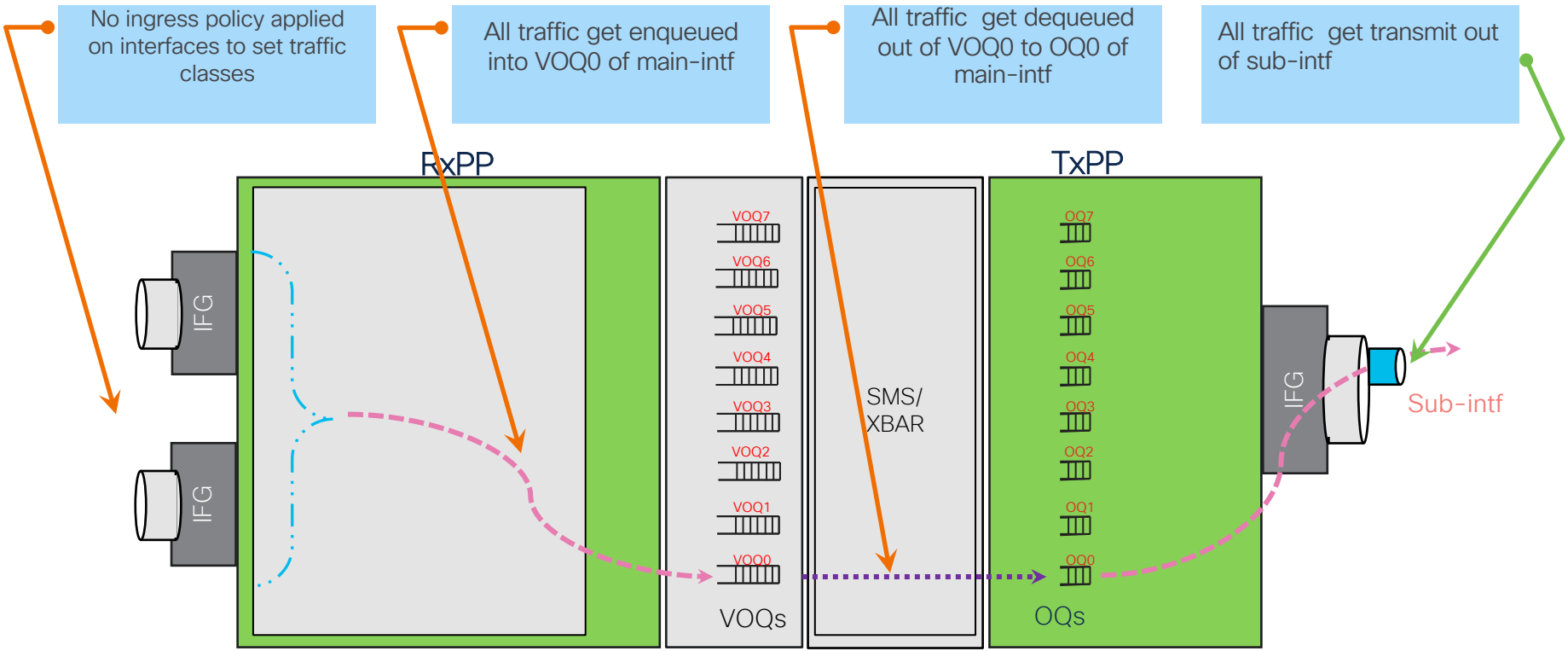
All traffic get dequeued out of VOQ0 – VOQ7 to OQ0 – OQ7 for transmitting out

All traffic get transmit out to wire



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Sub Interface Queueing behavior
Default



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Sub Interface Qing behavior

Traffic classes marked at ingress, no Qing policy applied on sub-intf

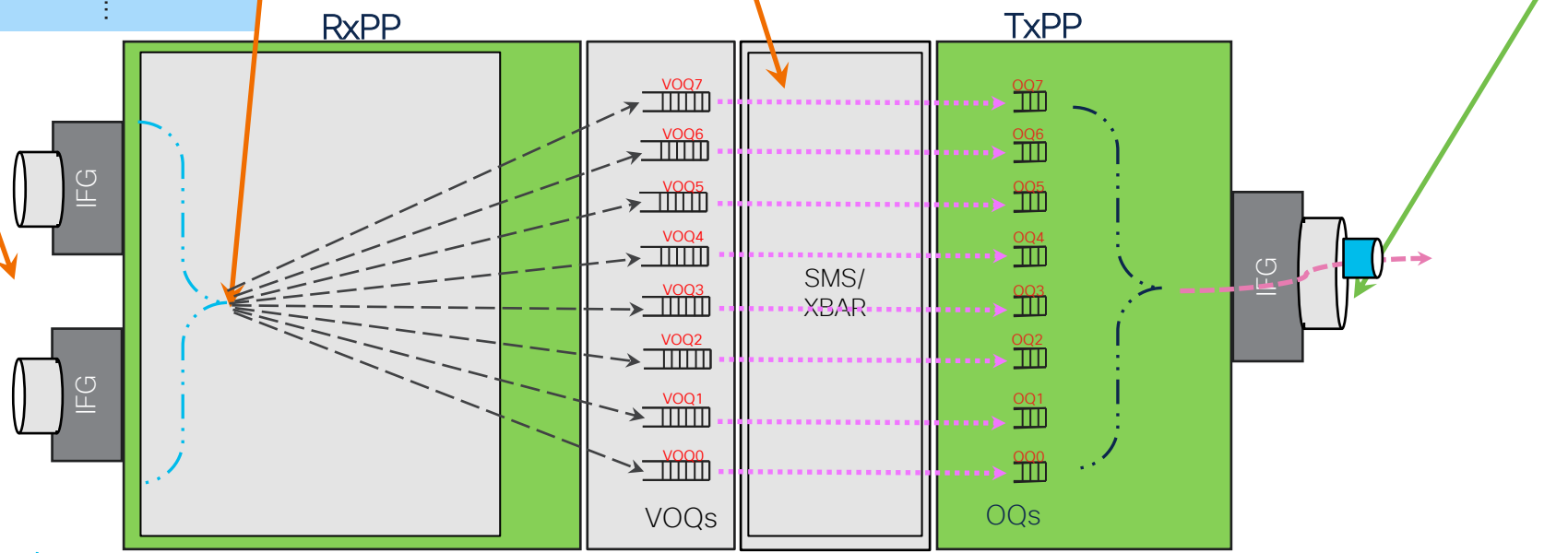
Ingress policy applied on interfaces:
Ex:

```
Class-map tc1
   set traffic-class 1
Class-map tc2
   set traffic-class 2
```

TC0 – TC7 traffic get enqueued into VOQ0 – VOQ7 of main-intf

All traffic get dequeued out of VOQ0 – VOQ7 to OQ0 – OQ7 of main-intf

All traffic get transmit out of sub-intf



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Sub Interface Qing behavior:

Traffic classes marked at ingress and Qing policy applied on sub-intf

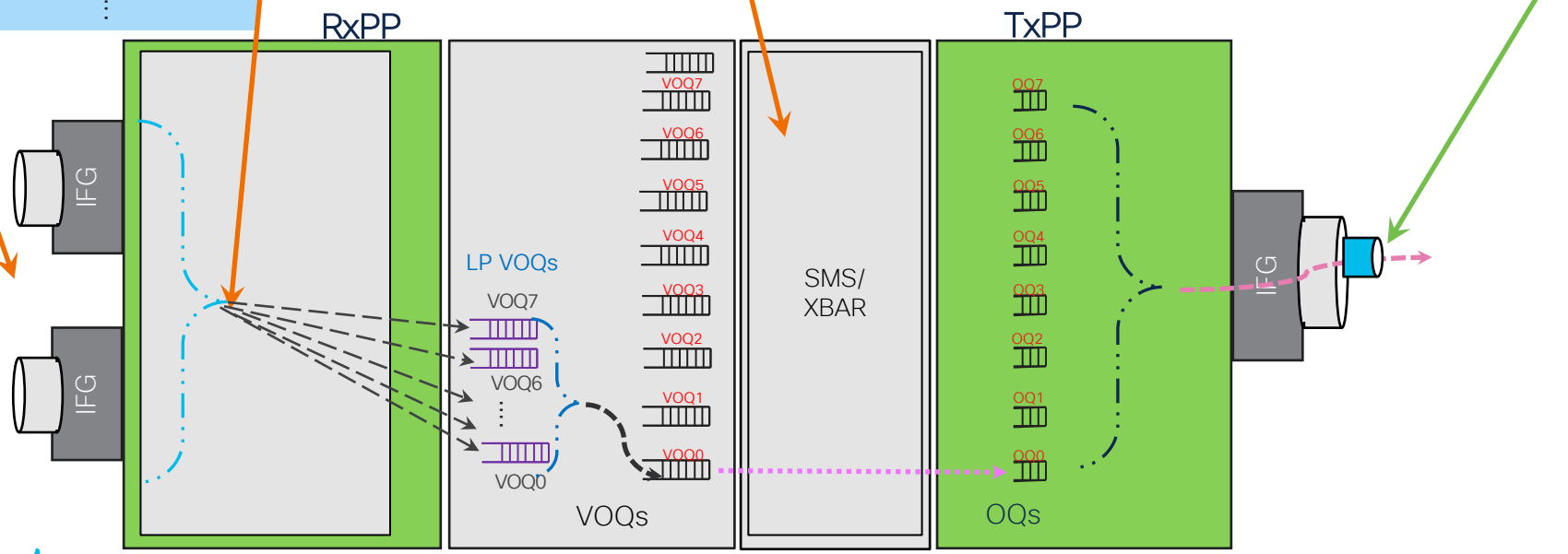
Ingress policy applied on interfaces:
Ex:

```
Class-map tc1
   set traffic-class 1
Class-map tc1
   set traffic-class 1
```

TC0 – TC7 traffic get enqueued into VOQ0 – VOQ7 of sub-intf

All traffic get dequeued out of VOQ0 – VOQ7 of sub-intf to OQ0 of main-intf

All traffic get transmit out of sub-intf



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Priority

- Traffic-class 7 must be set as “priority level 1” in the policy
- Priority can’t be skipped in policy config
- Default priority profile (w/o egress policy) is P1(TC7), P2(TC6), PNs(TC0-TC5)
- System originated control plane traffic are marked as TC7 which are marked by system internally and can’t be changed manually

Default mapping without policy

Traffic-Class	VoQ offset	OQ	Priority
7	7	7	P1
6	6	6	P2
5	5	5	Normal
4	4	4	Normal
3	3	3	Normal
2	2	2	Normal
1	1	1	Normal
0	0	0	Normal

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Bandwidth Remaining Ratio (BRR)

- BRR can't be configured in priority classes
- BRR give minimum guaranteed service rate for the classes
  - If we don't limit HP BW then LPs won't get any guaranteed service rate
- Each class can get higher than guaranteed service rate based on the congestion and over all BW consumption at that instant of time
- In right sample config and output,
  - TC7 has configured shaper as 10% of port BW (100G) -> 10Gbps
    - remaining BW of 90G is for other same priority classes and it is distributed as per the BRR value configured for each class
    - Config is omitted but total configured BRR value in the policy is 76
  - TC6 has shaper 10%, BRR 60%. Service rate = 10% of 100G-> 10Gbps
  - TC5 has BRR 4%: Service rate = (4/76) x 90G = 4.736842 Gbps

```
# sample config
policy-map Egress-Policy
 class traffic_out7
  priority level 1
  shape average percent 10
 !
 class traffic_out6
 queue-limit 10 mbytes
  bandwidth remaining ratio 60
  shape average percent 10
 !
 class traffic_out5
  queue-limit 600 ms
  bandwidth remaining ratio 4
 (total configured BRR value in this policy is 76)
```

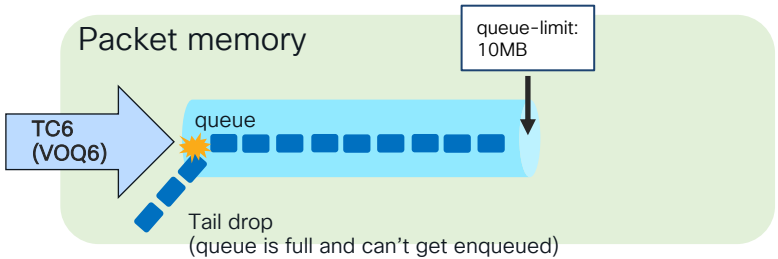
```
# output of show qos interface
Level1 Class (HP1)                    =   traffic_out7
Egressq Queue ID                      =   3095 (HP1 queue)
Queue Max. BW.                        =   10000000 kbps (10 %)
Guaranteed service rate               =   10000000 kbps
TailDrop Threshold                    =   7495680 bytes / 6 ms
(default)
WRED not configured for this class

Level1 Class                          =   traffic_out6
Egressq Queue ID                      =   3094 (LP queue)
Queue Max. BW.                        =   10000000 kbps (10 %)
Inverse Weight / Weight               =   1 / (60)
Guaranteed service rate               =   10000000 kbps
TailDrop Threshold                    =   9996288 bytes / 8 ms
(10 megabytes)
WRED not configured for this class

Level1 Class                          =   traffic_out5
Egressq Queue ID                      =   3093 (LP queue)
Queue Max. BW.                        =   no max (default)
Inverse Weight / Weight               =   15 / (4)
Guaranteed service rate               =   4736842 kbps
TailDrop Threshold                    =   355258368 bytes / 600
ms (600 ms)
WRED not configured for this class
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Queue limit

- Default Queue limit is 6ms on C8K systems
  - For 100G ports the default queue limit will be 75MB
  - For 400G ports the default queue limit will be 300MB
- Supported queue limit configuration units are
  - ***Bytes***: b, kb, mb
  - ***Time:*** ms, us
  - ***Percentage***: % of maximum supported queue limit (390mb)
  - ***# of Packets***: 256bytes/packet is considered to convert it to total bytes
- Supported queue limit range is 614400 byte (100 HBM block) – 390070272 byte (62k HBM block)

```
# output of show qos interface for 100G port with 10Gbps shaper
Level1 Class (HP1)                =   traffic_out7
Egressq Queue ID                  =   3095 (HP1 queue)
Queue Max. BW.                    =   10000000 kbps (10 %)
Guaranteed service rate           =   10000000 kbps
TailDrop Threshold                =   7495680 bytes / 6 ms (default)
```




```
class TC6
 queue-limit 10 mbytes
 bandwidth remaining ratio 60
 shape average percent 10
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Queue limit - contd

- Calculation example how queue limit is derived for each default, byte, packets, percent cases.
- Effective q-limit will be multiple of 6144 (block size).
- TC7 / Q-limit is default(6ms)
  - Q-limit in bytes = (6ms/10G)/8= 7.5 Mbytes
  - Effective Q-limit = (7.5MB/6144) x 6144 = 7495680 bytes
- TC6 / Q-limit is 10MB
  - Q-limit in bytes = 10 Mbytes
  - Effective Q-limit = (10MB/6144) x 6144 = 9996288 bytes
- TC5 / Q-limit is 2500 packets
  - Q-limit in bytes = 256 byte x 2500 = 640 kbyte
  - Effective Q-limit = (640KB/6144) x 6144 = 638976 bytes
- TC4 / Q-limit is 5%
  - Q-limit in bytes = 5% of 390070272 (max supported) = 19503513 byte
  - Effective Q-limit = (7.5MB/6144) x 6144 = 7495680 bytes

```
# sample config

policy-map Egress-Policy-
queuelimit
 class traffic_out7
  priority level 1
  shape average percent 10
 !
 class traffic_out6
  queue-limit 10 mbytes
  bandwidth remaining ratio 60
  shape average percent 10
 !
 class traffic_out5
  queue-limit 2500 packets
  bandwidth remaining ratio 4
 !
 class traffic_out4
  queue-limit percent 5
  bandwidth remaining ratio 3
```

```
# output of show qos interface for 100G port

Level1 Class (HP1)          =   traffic_out7
Egressq Queue ID            =   3095 (HP1 queue)
Queue Max. BW.              =   10000000 kbps (10 %)
Guaranteed service rate     =   10000000 kbps
TailDrop Threshold          =   7495680 bytes / 6 ms (default)

Level1 Class                =   traffic_out6
Egressq Queue ID            =   3094 (LP queue)
Queue Max. BW.              =   10000000 kbps (10 %)
Inverse Weight / Weight     =   1 / (60)
Guaranteed service rate     =   10000000 kbps
TailDrop Threshold          =   9996288 bytes / 8 ms (10
megabytes)

Level1 Class                =   traffic_out5
Egressq Queue ID            =   3093 (LP queue)
Queue Max. BW.              =   no max (default)
Inverse Weight / Weight     =   15 / (4)
Guaranteed service rate     =   4736842 kbps
TailDrop Threshold          =   638976 bytes / 1 ms (2500
packets)

Level1 Class                =   traffic_out4
Egressq Queue ID            =   3092 (LP queue)
Queue Max. BW.              =   no max (default)
Inverse Weight / Weight     =   20 / (3)
Guaranteed service rate     =   3552631 kbps
TailDrop Threshold          =   19501056 bytes / 44 ms (5 %)
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

QOS Resource
CGM profile

- VOQ CGM profile is one of the QOS resource for congestion management and definition of internal parameters at enqueueing in VOQ block
- Maximum of 32 VOQ CGM profiles are available **per system**
- VOQ CGM profile must be allocated/reserved at system bootup:
  - 4 profiles for internal use
  - 2 profiles for each unique interface speed (if system has 100G & 400G port, 4 profiles allocated)
- Remaining CGM profile are used for unique queue-limit and pair of min/max value of random-detect on a first come, first served basis.

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

QOS Resource
CGM profile

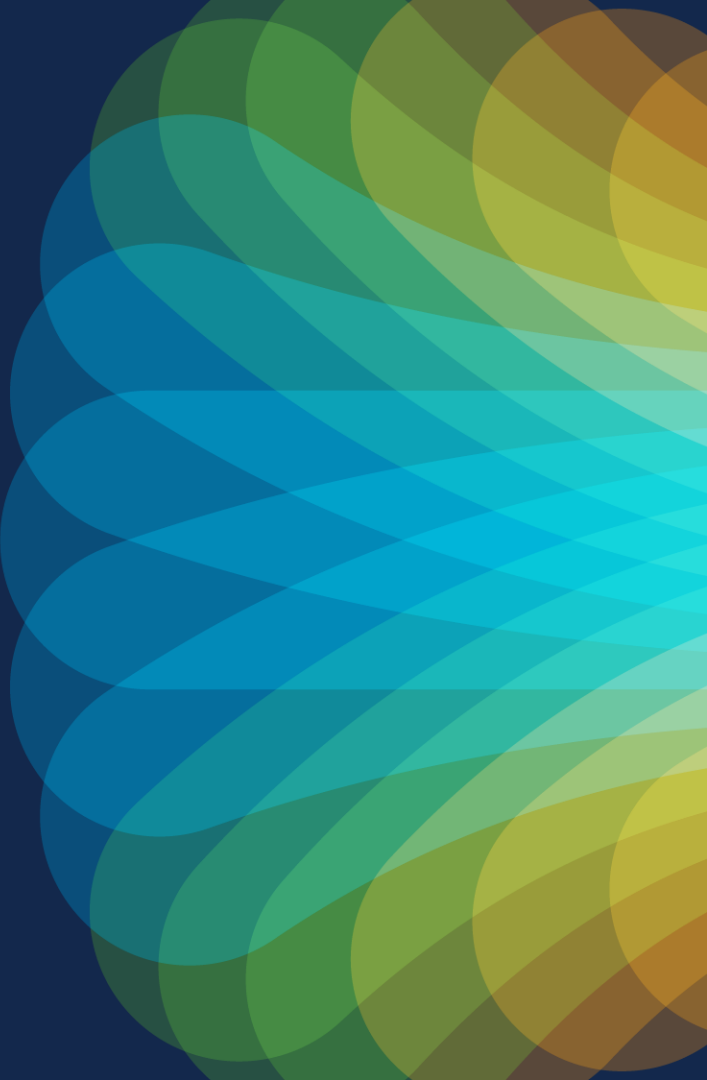
- This is an example of how to check available CGM profile and how many is allocated
- Internally allocated 4 profile is not shown in the ”show dpa objects” output
  - This system has 100G and 400G port, so default allocation is 4 entries
  - By adding a policy with 4 unique queue-limit, 4 entries are allocated
- Commit is failed if the config exceed the profile limit

```
### clean system with 100G and 400G ports

RP/0/RP0/CPU0:8201#show dpa objects tmrateprofile object-count location 0/RP0/CPU0
Wed Sep 13 23:46:20.454 JST
Table [TMRATEPROFILE] has 4 entries in DB
Table [TMRATEPROFILE] had 4 as highest count @ Sun Aug 28 20:34:09 2022


### attach Egress-Policy-subif to Hu0/0/0/3

RP/0/RP0/CPU0:8201#show dpa objects tmrateprofile object-count location 0/RP0/CPU0
Thu Sep 14 00:03:16.653 JST
Table [TMRATEPROFILE] has 8 entries in DB
Table [TMRATEPROFILE] had 11 as highest count @ Wed Sep 13 23:57:32 2023
```

```
### policy with 4 unique queue-limit
policy-map Egress-Policy-subif
 class traffic_out7
  priority level 1
  queue-limit 3500 kbyte ①
 !
 class traffic_out6
  queue-limit 3200 kbyte ②
  bandwidth remaining ratio 60
 !
 class traffic_out5
  queue-limit 3100 kbyte ③
  bandwidth remaining ratio 4
 !
 class class-default
  queue-limit 3000 kbyte ④
  bandwidth remaining ratio 2
 !
 end-policy-map
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

RED & Dual Queue Limit



Random Early Detection (RED)

- Cisco 8000 platform only support one curve of Random Early Detection(RED) for flows in a single class-map.
- WRED is not supported, and Dual queue-limit is the alternate solution.
- RED is a dropping mechanism that randomly drops packets before a VoQ is full.
  - every packet would have the same probability of being dropped inside a single class no matter if the packet has differentiated marking.
  - RED will be more likely to drop an incoming packet than when the average queue length is longer.
- RED can't be configured in priority classes
- RED can be enabled on a queueing policy indicating the value of minimum and maximum thresholds defining the curve.
  - Thresholds can be defined in bytes (up to gbytes) or in time (us to ms range).

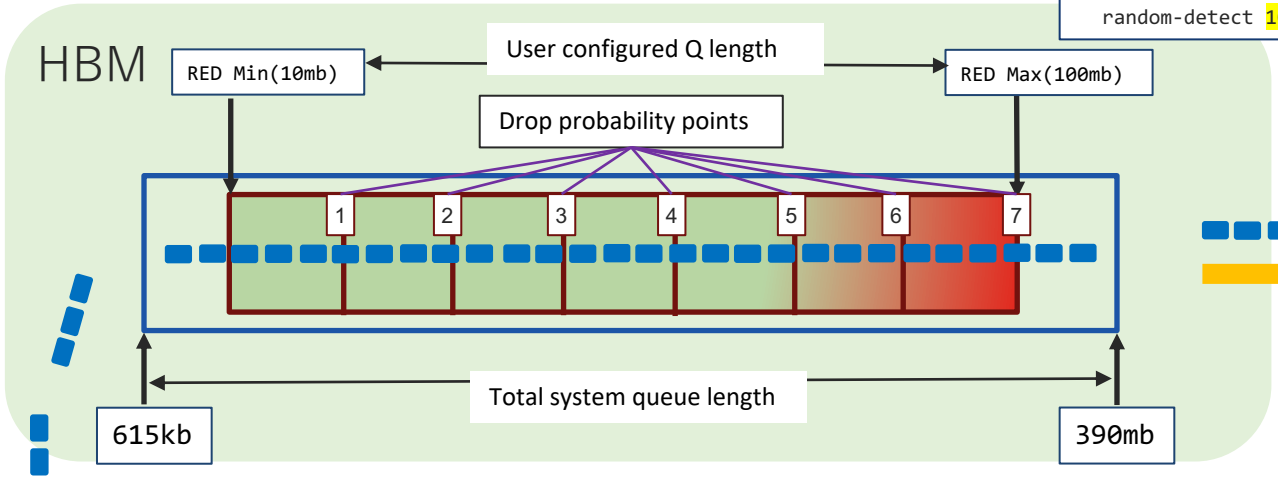
```
# sample config

 class traffic_out1
  bandwidth remaining ratio 2
  random-detect 10 mbytes 100 mbytes
```

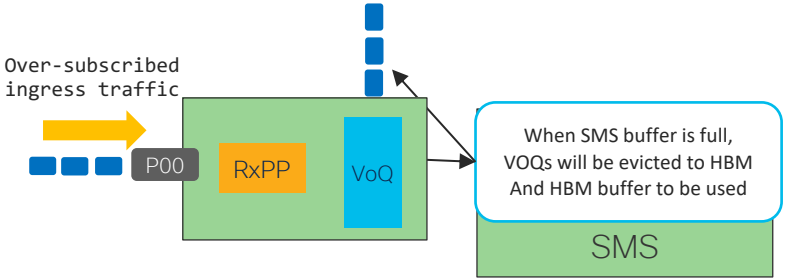
```
# sample output of show qos interface

Level1 Class                              =   traffic_out1
Egressq Queue ID                          =   3089 (LP queue)
Queue Max. BW.                            =   no max (default)
Inverse Weight / Weight                   =   30 / (2)
Guaranteed service rate                   =   2368421 kbps
TailDrop Threshold                        =   390070272 bytes / 1318 ms (default)

Default RED profile
WRED Min. Threshold                       =   9996288 bytes (10 megabytes)
WRED Max. Threshold                       =   99999744 bytes (100 megabytes)
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Random Early Detection (RED)

```
class traffic_out1
 bandwidth remaining ratio 2
 random-detect 10 mbytes 100 mbytes
```



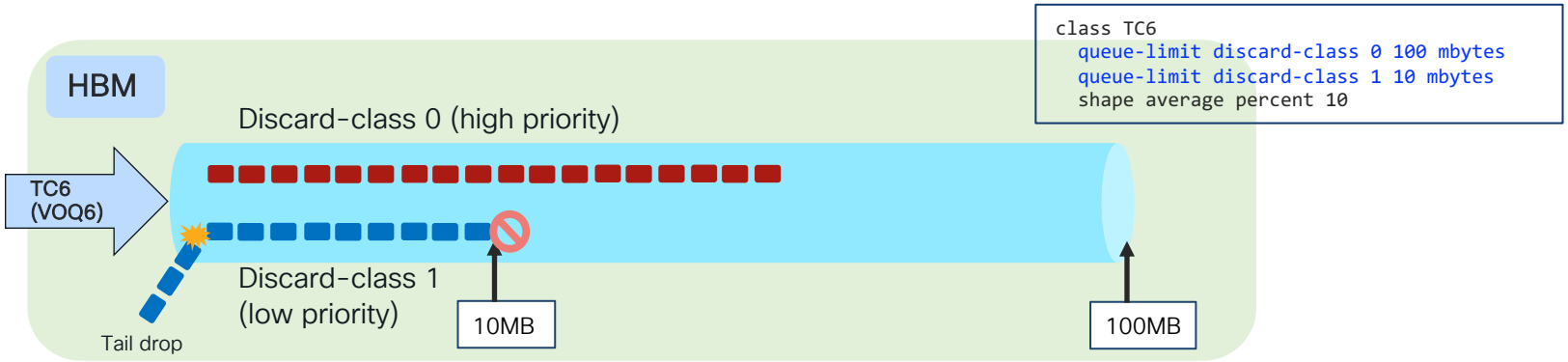


- Packets drop at different thresholds by different probablity
- example
  - drop probability - 100% over #7 where Avg. Q-length is over Max threshold
  - drop probability - 86% if Avg. Q-length reaches the point between #6 and #7
  - ...
  - drop probability - 14% if Avg. Q-length reaches the point between Min and #1
  - drop probability - 0% if Avg. Q-length is under Min threshold

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Dual Queue Limit (DQL)

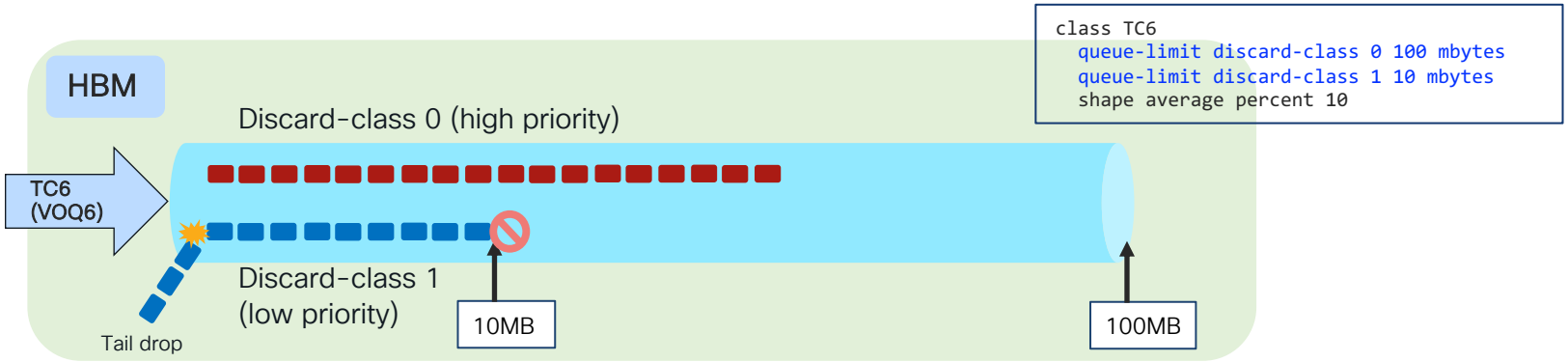
- As an alternative to two-curve WRED (not supported in 8000 platform)
- Protects High Priority traffic from Low Priority traffic
- To configure DQL, 2 queue limits to be configured in one Traffic Class; One queue-limit per discard-class(DC)
- Do not have any separate Drop counters for Discard class 0 and Discard class 1 packets



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

DQL – How it works

- Ingress policy classifies traffic and mark discard-class0(high priority) or discard-class1(low priority). Traffic class is same for both discard-class(DC).
- With congestion:
  - When queue length hits queue limit for DC1, then packets marked DC1 are no longer queued, and dropped
  - All bandwidth given to DC0 stream
- When queue length falls below DC1 queue-limit, DC1 packets start to enqueue and get serviced



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

DQL configuration example

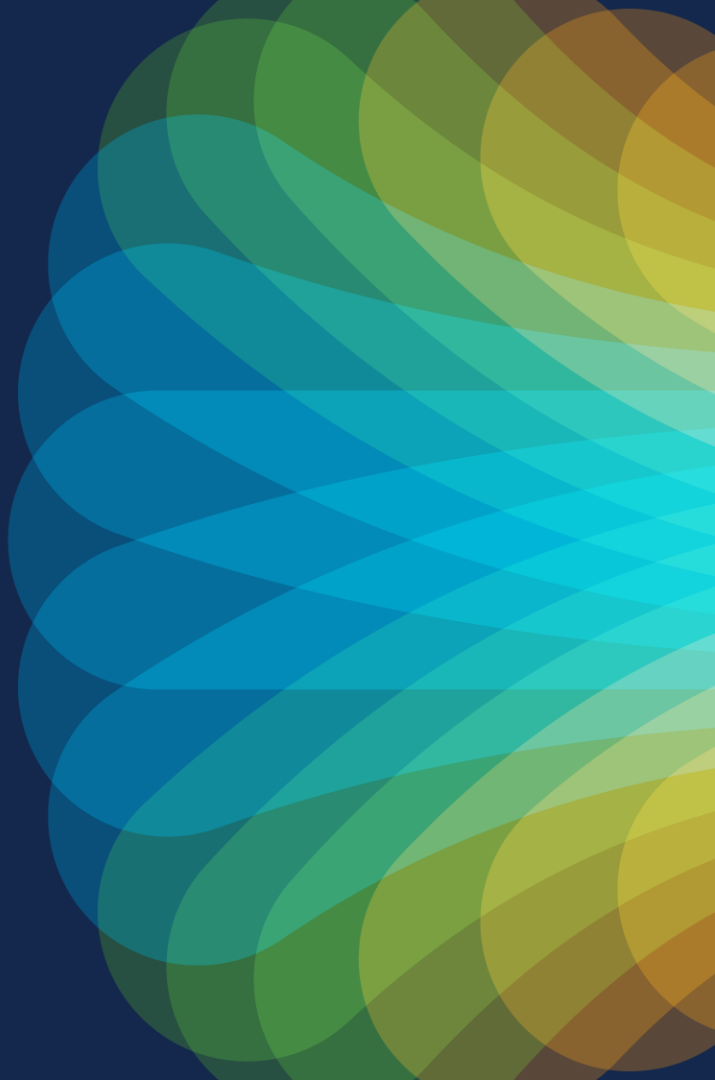
- Similar syntax, allowed units as 'queue-limit'
- Both queue-limit values for discard-class in class-map require same unit
  - mbytes and ms can't be mixed in same class
- Value of queue-limit for DC1 has to be less than DC0

```
policy-map DQL_egress
 class traffic_out7
  priority level 1
  shape average percent 20
 !
 class traffic_out6
  queue-limit discard-class 0 100 mbytes
  queue-limit discard-class 1 10 mbytes
  shape average percent 10
 !
 class traffic_out5
  queue-limit discard-class 0 50 ms
  queue-limit discard-class 1 30 ms
  shape average percent 5
 !
 class class-default
 !
 end-policy-map
!
end
```

```
Level1 Class (HP1)                      =   traffic_out7
Egressq Queue ID                        =   3095 (HP1 queue)
Queue Max. BW.                          =   20000000 kbps (20 %)
Guaranteed service rate                 =   20000000 kbps
TailDrop Threshold                      =   14997504 bytes / 6 ms (default)
WRED not configured for this class

Level1 Class                            =   traffic_out6
Egressq Queue ID                        =   3094 (LP queue)
Queue Max. BW.                          =   10000000 kbps (10 %)
Inverse Weight / Weight                 =   2 / (BWR not configured)
Guaranteed service rate                 =   10000000 kbps
Discard Class 1 Threshold               =   1474560 bytes / 1 ms (10 megabytes)
Discard Class 0 Threshold               =   75497472 bytes / 60 ms (100 megabytes)
WRED not configured for this class

Level1 Class                            =   traffic_out5
Egressq Queue ID                        =   3093 (LP queue)
Queue Max. BW.                          =   5000000 kbps (5 %)
Inverse Weight / Weight                 =   2 / (BWR not configured)
Guaranteed service rate                 =   5000000 kbps
Discard Class 1 Threshold               =   12582912 bytes / 20 ms (30 ms)
Discard Class 0 Threshold               =   25165824 bytes / 40 ms (50 ms)
WRED not configured for this class
```

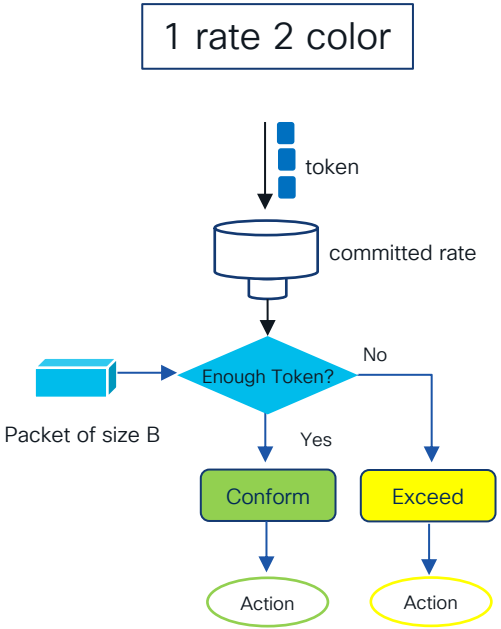

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

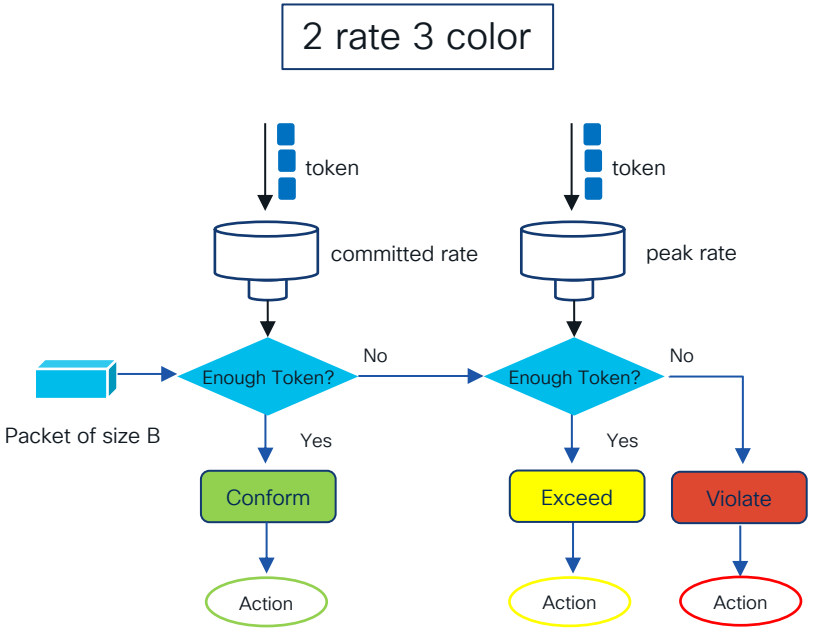
Policer



Policer – limit the bandwidth

- Single Rate Two Color Policer (1R2C) & Two Rate Three Color Policer (2R3C) is supported
  - 1R2C policer is enabled when user configures just one rate associated to a policer in the policy-map.
  - 2R3C policer is enabled when user specifies two rates (committed rate & peak-rate) to a policer in the policy-map.





© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Policer Config sample

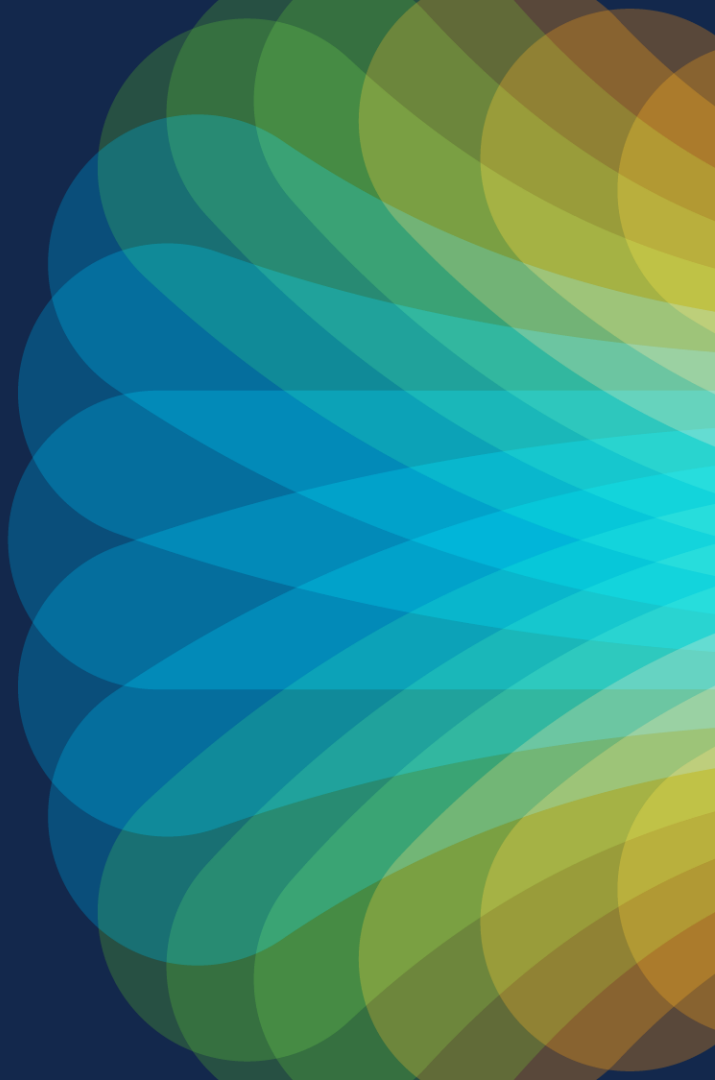
- By default, if user does not specify any action, the default conform action is transmit and the default exceed & violate action(for 2R3C) is drop.
- Policer actions supported (Transmit/Drop/Set discard-class)

```
# sample config
class-map match-any 1R2C
 match mpls experimental topmost 5
 match precedence critical
 end-class-map

class-map match-any 2R3C
 match mpls experimental topmost 4
 match precedence flash-override
 end-class-map

policy-map policer
 class 1R2C
  police rate 500 mbps
 class 2R3C
  police rate 10 gbps peak-rate 20 gbps
 class class-default
```

```
# sample output of show qos interface
Level1 Class                          =   1R2C
Policer committed rate                =   500000 kbps (500 mbits/sec)
Policer conform burst                 =   10240 bytes (default)
Policer conform action                =   Just TX
Policer exceed action                 =   DROP PKT

Level1 Class                          =   2R3C
Policer committed rate                =   10000000 kbps (10 gbits/sec)
Policer peak rate                     =   20000000 kbps (20 gbits/sec)
Policer conform burst                 =   1024000 bytes (default)
Policer exceed burst                  =   2048000 bytes (default)
Policer conform action                =   Just TX
Policer exceed action                 =   DROP PKT
Policer violate action                =   DROP PKT

Level1 Class                          =   class-default
Policer not configured for this class
```

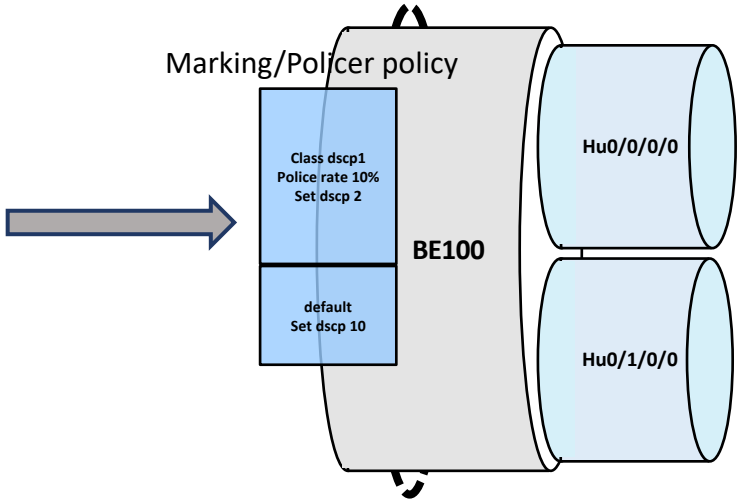

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Bundle QOS



Classify/Marking on Bundle interface

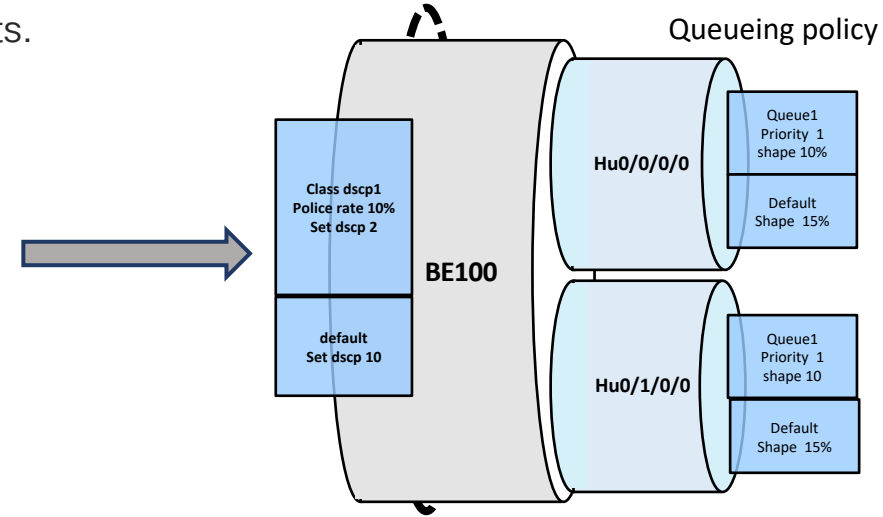
- A classification, marking & policing policy can be applied on bundle main interface or bundle sub interface but not on both.
- Classification and Policer resources (policers) are allocated per bundle interface but not per bundle member interface.
  - Classification and policing policy stats are per bundle interface but not per member interface.
- Policer rates can be configured only in percent unit.



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

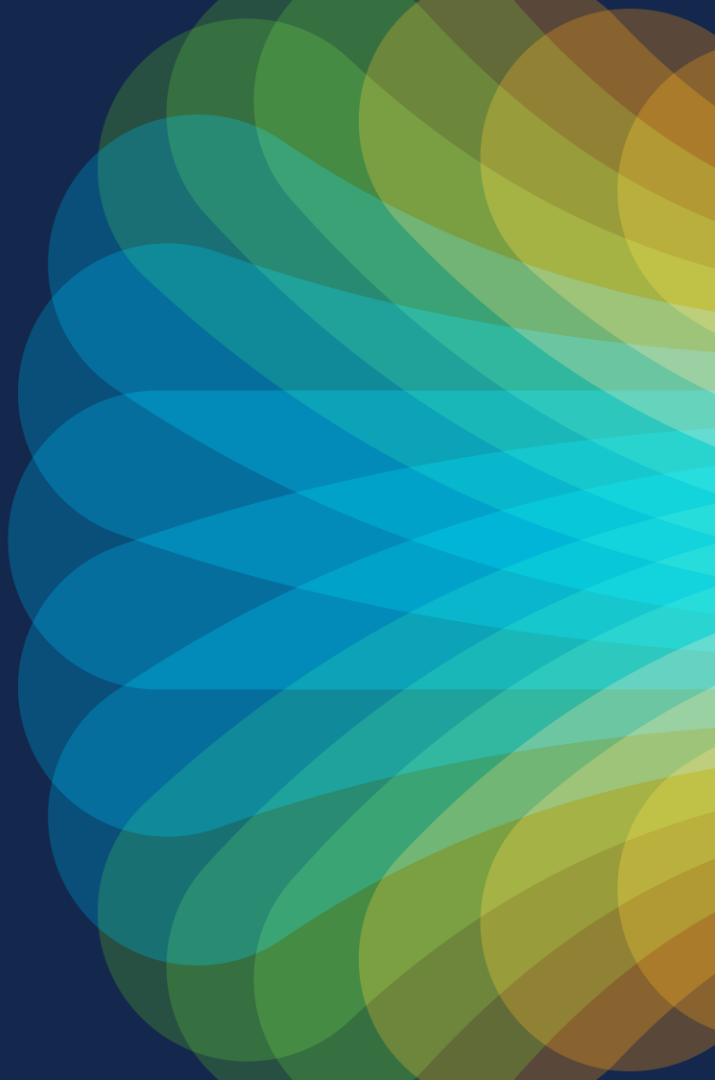
Queueing on Bundle interface

- A Queueing policy can be applied on bundle main interface or bundle sub interface but not on both.
- Queueing policy resources (voq's) are allocated per bundle member interface.
  - Queueing policy stats are per bundle interface and per member interface.
- Shaper rates can be configured only in percent units.



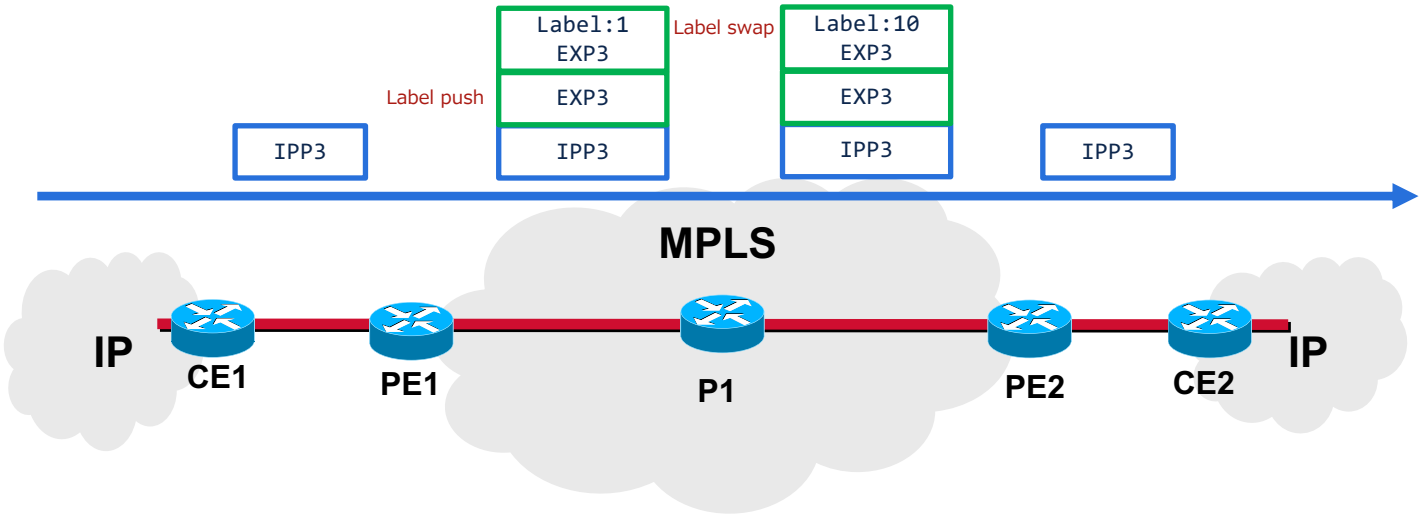
© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

MPLS QOS



Default Diffserv behavior
without any marking policy

- MPLS-EXP is derived from IP DSCP value while imposing label at the ingress PE node
  - It's same for pushing both single/multiple labels case
  - Outmost label's EXP is used for swap case on P-node

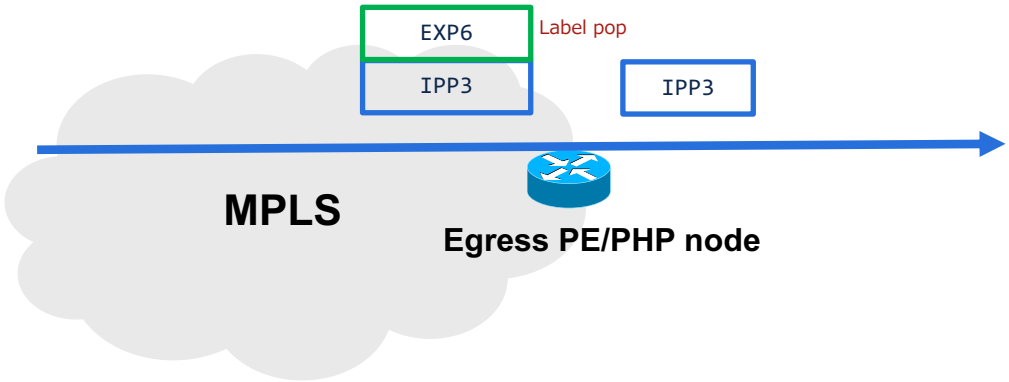


NOTE: if TCAM based (ACL) classification is applied at ingress on the ingress-PE then MPLS-EXP does not get derived from IP DSCP and it always get zero. So in such deployment its must to have MPLS-EXP marking present in either ingress or egress policy on that node.

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

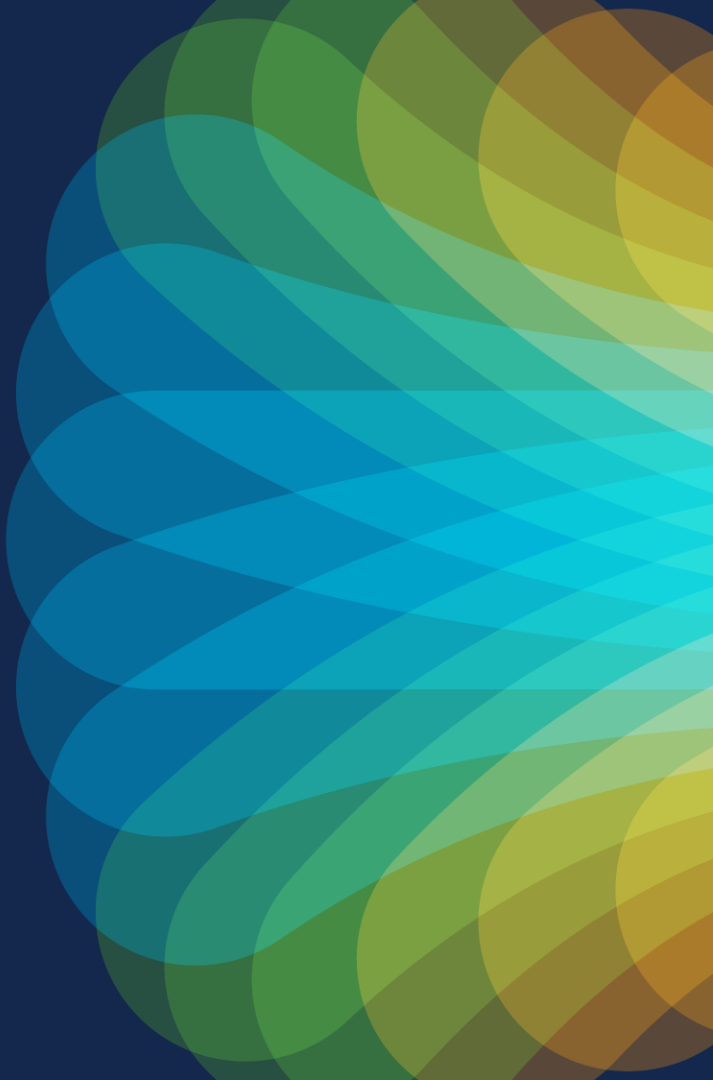
Default Diffserv behavior
without any marking policy

- Egress PE/PHP node does not copy EXP bits correspondingly to IP DSCP bits while popping labels



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

QOS Trouble Shooting CLI



Get a snapshot of QOS configuration

- “show policy-map targets” and its options can be used to get a quick snapshot of all policy-maps.
- You can find usage of a particular policy-map across interfaces without show run interface

```
RP/0/RP0/CPU0:C8818#show policy-map targets
1) Policymap: Egress-Policy-subif    Type: qos
     Targets (applied as main policy):
       HundredGigE0/0/0/5.1 output
       HundredGigE0/0/0/5.2 output
       HundredGigE0/0/0/5.3 output
       HundredGigE0/0/0/5.4 output
       HundredGigE0/0/0/5.5 output
       HundredGigE0/0/0/5.6 output
     Total targets: 6

     Targets (applied as child policy):
     Total targets: 0

2) Policymap: marking1_out    Type: qos
     Targets (applied as main policy):
       Bundle-Ether100 output
     Total targets: 1

     Targets (applied as child policy):
     Total targets: 0
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Check mapping between IF and VOQ

- “show controllers npu voq-usage interface” to check the mappings
- You can find corresponding NPU/Slice/IFG/VOQ base of specific interface.
  - Same entry is created on other LC as “remote” for VOQ port type

```
RP/0/RP0/CPU0:C8818#show controllers npu voq-usage interface tenGigE 0/0/0/12 instance
all location 0/0/CPU0
------------------------------------------------------------------
Node ID: 0/0/CPU0
Intf         Intf     NPU Slice IFG  Sys   VOQ   Flow   VOQ    Port
name         handle    #    #    #   Port  base  base   port   speed
             (hex)                                  type
------------------------------------------------------------------
Te0/0/0/12   160       1    2    1    52   1552      0 local    10G

RP/0/RP0/CPU0:C8818#show controllers npu voq-usage interface tenGigE 0/0/0/12 instance
all location 0/1/CPU0
------------------------------------------------------------------
Node ID: 0/1/CPU0
Intf         Intf     NPU Slice IFG  Sys   VOQ   Flow   VOQ    Port
name         handle    #    #    #   Port  base  base   port   speed
             (hex)                                  type
------------------------------------------------------------------
Te0/0/0/12   160       1    2    1    52   1552      0 remote   10G
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Typical QOS monitoring commands

- “show qos interfaces” to check service rate, VOQ base, and other QOS parameters

```
RP/0/RP0/CPU0:C8818#show qos interface tenGigE 0/0/0/12 out
NOTE:- Configured values are displayed within parentheses
Interface TenGigE0/0/0/12 ifh 0x160  -- output policy
NPU Id:                        1
Total number of classes:       3
Interface Bandwidth:           10000000 kbps
Policy Name:                   TS_congestion
VOQ Base:                      1552
Accounting Type:               Layer1 (Include Layer 1 encapsulation and above)
VOQ Mode:                      8
Shared Counter Mode:           1
------------------------------------------------------------------------------
Level1 Class (HP1)                       =   traffic_out7
Egressq Queue ID                         =   1559 (HP1 queue)
Queue Max. BW.                           =   9000000 kbps (90 %)
Guaranteed service rate                  =   9000000 kbps
TailDrop Threshold                       =   49999872 bytes / 44 ms (50 megabytes)
WRED not configured for this class

Level1 Class                             =   traffic_out1
Egressq Queue ID                         =   1553 (LP queue)
Queue Max. BW.                           =   800000 kbps (8 %)
Inverse Weight / Weight                  =   63 / (BWR not configured)
Guaranteed service rate                  =   500000 kbps
TailDrop Threshold                       =   620544 bytes / 10 ms (10 ms)
WRED not configured for this class
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Typical QOS monitoring commands

- “show policy-map interface” to check QOS stats

```
RP/0/RP0/CPU0:C8818#show policy-map interface tenGigE 0/0/0/12 output

TenGigE0/0/0/12 output: TS_congestion

Class traffic_out7
  Classification statistics          (packets/bytes)     (rate - kbps)
    Matched             :         10617695830/4431612637967        10022122
    Transmitted         :          9245145010/3859802379379        8524105
    Total Dropped       :          1372550820/571810258588         1498017
  Queueing statistics
    Queue ID                              : 1559
    Taildropped(packets/bytes)            : 1372550820/571810258588
Class traffic_out1
  Classification statistics          (packets/bytes)     (rate - kbps)
    Matched             :          8990589798/3752352492778        10022297
    Transmitted         :           475255557/198487540270         520055
    Total Dropped       :          8515334241/3553864952508        9502242
  Queueing statistics
    Queue ID                              : 1553
    Taildropped(packets/bytes)            : 8515334241/3553864952508
- snip -
```

Matched: stats for packet match this class
Transmitted: stats for packet transmitted
Total Dropped: stats for dropped packet by QOS

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

QOS stats monitoring
Egress congestion

- “show controllers npu stats voq” to check VOQ counters
- Egress congestion can happen
  - Flows from several ingress interfaces add up to more than the destination interface rate
  - Egress interface is shaped to a rate lower than the ingress traffic bound to it

```
RP/0/RP0/CPU0:C8818#show controllers npu stats voq base 1552 instance all location 0/1/CPU0

Location            =     0/1/CPU0
Asic Instance       =            0
VOQ Base            =         1552
VOQ Mode            =            8
Shared Counter Mode =            1
       ReceivedPkts    ReceivedBytes   DroppedPkts     DroppedBytes
-------------------------------------------------------------------
TC_0 = 27979845        11675149996     511100178       213322506836
TC_1 = 27980511        11675141804     511099467       213330425738
TC_2 = 0               0               0               0
TC_3 = 0               0               0               0
TC_4 = 0               0               0               0
TC_5 = 0               0               0               0
TC_6 = 0               0               0               0
TC_7 = 458651386       191355104168    80428636        33635443206
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

QOS stats monitoring
Egress congestion

- “show drops all” to check overall drop counters including VOQ drops
  - ongoing option is useful to check the drops since last read

```
RP/0/RP0/CPU0:C8818#show drops all ongoing location 0/1/CPU0

-------------------
MODULE voq_drops
-------------------

VOQ Base               =         1552
Interface Name         =   Te0/0/0/12
Te0_0_0_12
Asic Instance          =            0
Local Port             =       remote
Location               =     0/1/CPU0
Interface Handle       =          160
----------------------------------
TC_0 : 85182275
TC_1 : 85183007
TC_7 : 13412802
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

QOS stats monitoring
Egress congestion

- “show controller npu stats counters-all ” for overall NPU stats
  - detail option is available
  - Packets that cannot be enqueued will be tail dropped. And VOQ drops are expected.

```
RP/0/RP0/CPU0:C8818#show controller npu stats counters-all instance all location 0/1/CPU0
Tue Sep 12 08:50:16.778 UTC

Statistics Rack: 0, Slot: 1, Asic instance: 0

Counters update: Tue Sep 12 08:50:08 2023

Total number of blocks: 30
Per Block Statistics:

Slice 0 RX Summary counters:
  IFG_RX 0 packets                    :     405451435203
  IFG_RX 0 bytes                      :  123696684538339
  IFGB_RX 0 port packets              :       1724304718 (ovf)
  IFG_RX 1 packets                    :                0
  IFG_RX 1 bytes                      :                0
  PDVOQ drop packets                  :      20529009665
  PDVOQ drop bytes                    :    1233540489358
```


© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

QOS Troubleshooting
show tech-support

- show tech-support ofa
- show tech-support qos pi
- show tech-support qos platform
- show tech-support fabric
- show tech-support platform-fwd

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

cisco Live!

Did you know?

You can have a
one-on-one session with
a technical expert!

Visit Meet the Expert in The HUB
to meet, greet, whiteboard & gain
insights about your unique questions
with the best of the best.

Meet the Expert Opening Hours:

Tuesday	**3:00pm – 7:00pm**
Wednesday	**11:15am – 7:00pm**
Thursday	**9:30am – 4:00pm**
Friday	**10:30am – 1:30pm**

Session Surveys

We would love to know your feedback on this session!

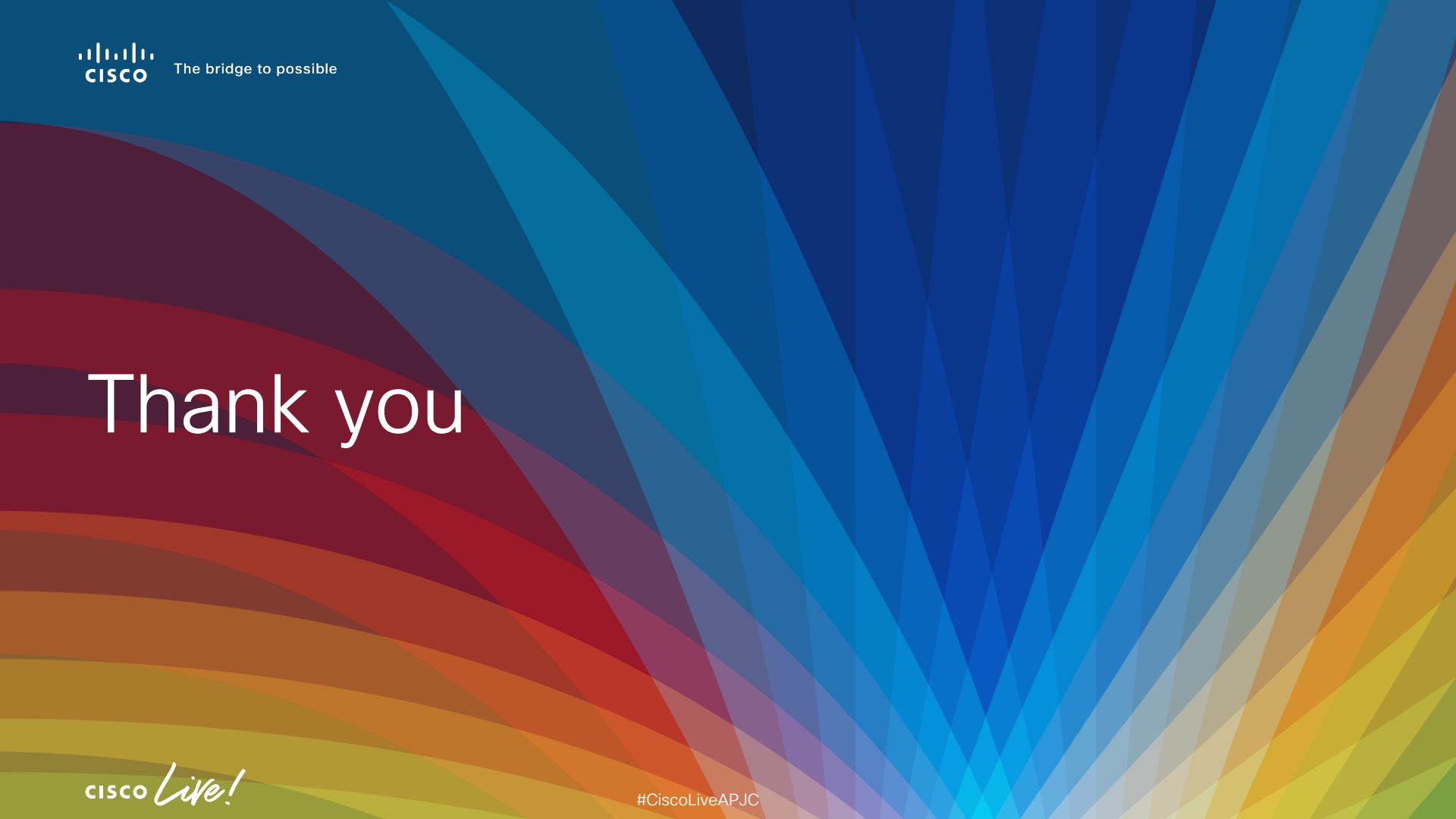
- Complete a minimum of four session surveys and the overall event surveys to claim a Cisco Live T-Shirt

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Expert meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public


cisco
The bridge to possible
Thank you
cisco Live!
#CiscoLiveAPJC

cisco Live!

Let's go

#CiscoLiveAPJC