White Paper
Public

Cisco N9300-H Series Switches White Paper

© 2022 Cisco and/or its affiliates. All rights reserved.

Contents

Introduction 4

Target Audience 4

CloudScale ASIC Family 4

Cisco CloudScale LS12800H2R, LS12800H2 and LS6400H1 ASIC Architecture 6

CloudScale LS12800H2R, LS12800H2 and LS6400H1 ASIC Overview 6

CloudScale H2R/H2/H1 ASIC Slice 6

CloudScale H2R/H2/H1 ASIC Slice Forwarding Path 7

CloudScale H2R/H2/H1 ASIC Ingress Forwarding Controller 8

CloudScale H2R/H2/H1 ASIC Egress Forwarding Controller 8

Cisco CloudScale H2R, H2 and H1 ASIC Forwarding lookup 9

CloudScale H2R/H2/H1 ASIC Forwarding TCAM 9

CloudScale H2R/H2/H1 ASIC Flexible forwarding Tiles (flex tiles) 10

CloudScale H2R/H2/H1 ASIC Unicast Forwarding 12

CloudScale H2R/H2/H1 ASIC Multicast Forwarding 13

CloudScale H2R/H2/H1 ASIC VXLAN Forwarding 14

Cisco CloudScale H2R, H2 and H1 ASIC Clasification TCAM 14

CloudScale H2R/H2/H1 ASIC TCAM Scalability 15

CloudScale H2R/H2/H1 ASIC Default ACL TCAM Allocation 15

CloudScale H2R/H2/H1 ASIC Egress RACL Example 16

Cisco CloudScale H2R, H2 and H1 ASIC Hardware Telemetry 17

CloudScale H2R/H2/H1 ASIC Flow Table (FT), Flow Table Event (FTE) and Netflow 18

CloudScale H2R/H2/H1 ASIC Inband Network Telemetry (INT) 19

CloudScale H2R/H2/H1 Streaming Statistics Export (SSX) 21

Cisco CloudScale H2R, H2 and H1 Span on Drop (SOD) 21

CloudScale H2R/H2/H1 ASIC Ingress Forwarding Drop 23

CloudScale H2R/H2/H1 ASIC Egress Buffer Drop 23

CloudScale H2R/H2/H1 ASIC Egress Rewrite Engine Drop 24

CloudScale H2R/H2/H1 ASIC Packet Recirculation 24

Cisco CloudScale H2R, H2 and H1 ASIC Packet Buffer 24

CloudScale H2R/H2/H1 ASIC On-Die packet Buffer 25

CloudScale H2R/H2/H1 ASIC Intelligent Buffering 27

CloudScale H2R HBM Packet Buffer 29

Cisco CloudScale H2R, H2 and H1 ASIC AI/ML 30

Explicit Congestion Notification (ECN) 31

Priority Flow Control (PFC) 31

© 2022 Cisco and/or its affiliates. All rights reserved.

Cisco N9300-H Switches 32
Cisco N9332D-H2R 32
Cisco N93400LD-H1 34
Cisco N9364C-H1 35
Cisco N9348C8D-H2 37
Cisco N9300-H MACsec Hardware Encryption 39
Cisco N9300-H SyncE and PTP 40
Conclusion 42
Additional Resources 42

© 2022 Cisco and/or its affiliates. All rights reserved.

Introduction

In 2016, Cisco made its debut of the Cisco Nexus® 9000 Series Switch line with products built on the revolutionary Cisco® Cloud Scale intelligent Application-Specific Integrated Circuits (ASICs). These initial Cloud Scale platforms enabled customers to establish high-performance, cost-effective data center networks, offering a transition from 10G and 40G Ethernet to more robust 25G and 100G connectivity options. Over the following years, Cisco continued to expand and enhance the product family with additional platforms, introducing numerous innovations to address the evolving demands of Cloud Scale data centers, converged and hyperconverged infrastructure, and virtualized and containerized applications.

Large-cloud and data-center networking teams require a flexible, reliable solution that efficiently manages, troubleshoots, and analyzes their IT infrastructure. In addition, they need security, automation, visibility, analytics, and assurance. Coupled with tools such as Cisco Nexus Dashboard (ND) for visibility and automation, Cisco N9300 Series switches are ideal platforms to build a high-performance AI/ML network fabric.

The inclusion of 400G Ethernet technology further enriches the Cloud Scale portfolio by providing platforms that deliver high-density, top-of-rack aggregation, spine aggregation, and backbone connectivity options. These platforms utilize the GX, GX2A, GX2B, and the latest additions, H2R, H2 and H1 Cloud Scale ASICs. With 400G options available for both top-of-rack and modular systems, the Cisco N9000 series chassis stand at the forefront of the industry, incorporating innovations such as 400G QSFP-DD (double density) transceivers that offer full backward compatibility with existing QSFP28 (100G) and QSFP+ (40G) transceivers.

This comprehensive document provides an in-depth explanation of Cisco CloudScale LS12800H2R, LS12800H2 and LS6400H1 ASICs utilized in Cisco N9300-H2R, N9300-H2 and N9300-H1 platfroms. The primary aim of this document is to highlight the latest feture enhancments with the CloudScale ASICs.

Target Audience

This document is for planning, implementation, and maintenance in DevOps teams.

CloudScale ASIC Family

The Cisco N9000 platform switches are equipped with cutting-edge Cisco Custom CloudScale technology ASICs. This unique approach of designing and owning both the silicon, Software SDK, and operating system confers significant competitive advantages to the Cisco N9000 series. It serves as a powerful vehicle for delivering an array of advanced features and functions.

Leveraging the advantages of the latest semiconductor device fabrication, these ASICs boast higher transistor density and lower power consumption. These features are crucial in enabling the creation of ASICs with increased bandwidth, a higher number of ports, larger forwarding tables, generous buffers, and the opportunity to implement novel, advanced capabilities.

Cisco's cloud-scale ASICs introduce a wide range of Ethernet speeds, including 25, 50, 100, 200, and 400 Gigabit Ethernet (GE), to data center networks at a cost point optimized for optimal performance. See Table 1 for the specific Cloud Scale ASICs used in each Cisco N9000 Switch Family. Table 2 highlights the major feature sets associated with each Cloud Scale ASIC.

Table 1. Cisco CouldScale ASIC family with their corresponding Cisco N9000 switch

© 2022 Cisco and/or its affiliates. All rights reserved.

CloudScale ASIC	N9000 Family	Platforms and Line Cards
LS 1800 EX	9300-EX, X9700-EX (LCs)	C93180YC-EX, C93108TC-EX, C93180LC-EX, X9732C-EX, X9736C-EX, X97160YC-EX
LS 1800 FX	9300-FX, X9700-FX (LCs)	C93180YC-FX, C93108TC-FX, C9348GC-FXP, X9732C-FX, X9736C-FX, X9788TC-FX
LS 3600 FX2	9300-FX2	C9336C-FX2, C9336C-FX2-E, C93240YC-FX2, C93360YC-FX2, C93216TC-FX2
LS 1800 FX3	9300-FX3	C93180YC-FX3, 93180YC-FX3S, C93108TC-FX3P, C93108TC-FX3, C9348GC-FX3, C9348GC-FX3PH
S 6400	9300C, FM-E2 (FMs)	9364C, 9332C,9508-FM-E2, 9516-FM-E2
LS 6400 GX	9300-GX, X9700-GX (LCs)	C9316D-GX, C93600CD-GX, C9364C-GX, X9716D-GX, 9504-FM-G, 9508-FM-G
LS 25600 GX2A LS 12800 GX2B	9300-GX2A, 9300-GX2B, 9408	9364D-GX2A, 9348D-GX2A, 9332D-GX2B, C9400-SW-GX2A
LS 12800 H2R, LS 12800 H2 LS 6400 H1	9300-H2R, 9300-H2, 9300-H1	9332D-H2R, 9364C-H1, C93400LD-H1, 9348C8D-H2

Table 2. Cisco CouldScale ASIC family features

ASIC Feature	LS 1800 EX	LS 1800 FX	LS 3600 FX2	LS 1800 FX3	S 6400	LS 6400 GX	LS25600 GX2A	LS12800 GX2B	LS12800 H2/R	LS6400 H1
Bandwidth (Tbps)	1.8T	1.8T	3.6T	1.8T	6.4T	6.4T	25.6T	12.8T	12.8T	6.4T
No. of Slices	2 Slices	1 Slice	2 Slices	1 Slice	4 Slices	4 Slices	8 Slices	4 Slices	4 Slices	2 Slices
Fabrication node	16nm	16nm	16nm	14nm	16nm	14nm	7nm	7nm	7nm	7nm
Max Ports (400/200/ 100/50 GE)	0/0/ 18/72	0/0/ 18/72	0/0/ 36/72	0/0/ 18/72	0/0/ 64/0	16/32/ 64/128	64/128/ 256/256	32/64/ 128/128	32/64/ 128/128	16/32/ 64/128
Buffer/Max-per-port (MB)	37.4/18.7	40/40	40/30	40/40	40/10	80/20	120/30	120/60	80/40	40/40
HBM Packet Buffer	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	8G HBM (H2R)	N/A

© 2022 Cisco and/or its affiliates. All rights reserved.

ASIC Feature	LS 1800 EX	LS 1800 FX	LS 3600 FX2	LS 1800 FX3	S 6400	LS 6400 GX	LS25600 GX2A	LS12800 GX2B	LS12800 H2/R	LS6400 H1
Telemetry	FT	FT, FTE	FT, FTE, SSX	FT, FTE, SSX	SSX	FT, FTE, SSX, INT-XD, INT-Transparent	FT, FTE, SSX, INT-XD	FT, FTE, SSX, INT-XD	FT, FTE, SSX, INT-MX, INT-MD, INT-XD	FT, FTE, SSX, INT-MX, INT-MD, INT-XD
Hardware Entries/Tiles	544K/17	1088K/34	544K/17	1088K/34	352K/11	1088K/34	640K/20	1280K/40	1280K/40	1280K/40
Classification TCAM (Ingress/ Egress) Per Slice	4K/2K	5K/2K	5K/2K	5K/2K	4K/2K	5K/2K	6K/3K	6K/3K	14K Shared	14K Shared

Cisco CloudScale LS12800H2R, LS12800H2 and LS6400H1 ASIC Architecture

CloudScale LS12800H2R, LS12800H2 and LS6400H1 ASIC Overview

LS12800H2/R and LS6400H1 are the latest addtions to the Cisco cloudScale ASIC family with multiple enhacmnets to the privoues ClouldScale ASICs. LS12800H2R comes with dual die, 4 network slices and High Bandwidth Memory (HBM) and LS6400H1has a single die with 2 slices. We will refer to them as H2R, H2 and H1 ASICs for the rest of this document. Figure 1 illustrate the H2R and H1 ASICs.

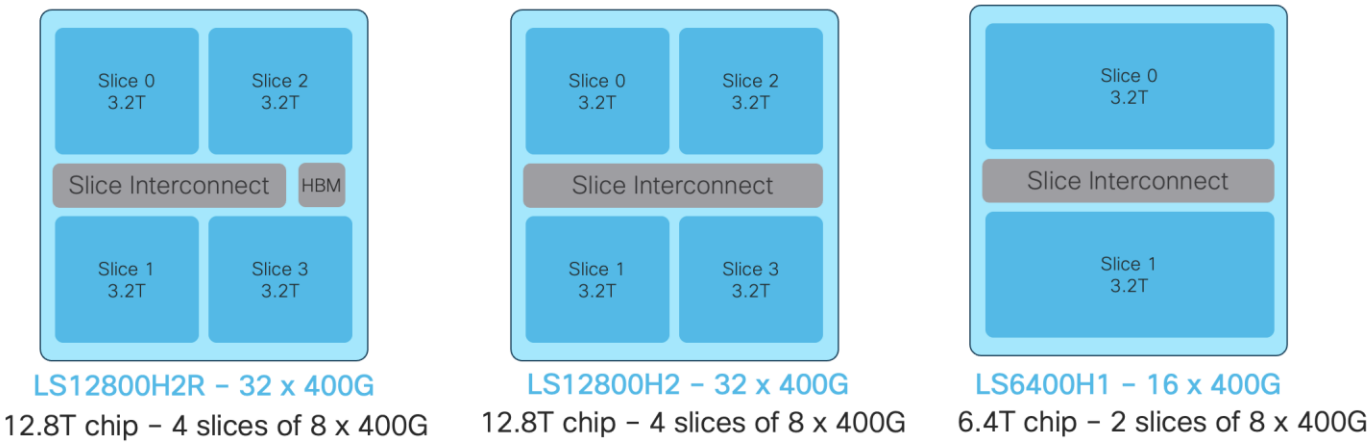


Figure 1. Cisco CouldScale LS12800H2R, LS12800H2 and LS6400H1 ASICs

CloudScale H2R/H2/H1 ASIC Slice

A CloudScale ASIC slice refers to a self-contained forwarding complex responsible for handling both ingress and egress functions for a specific subset of ports. Each slice is further divided into separate sections for ingress and egress functions. The ingress portion of each slice is interconnected to all egress portions of other slices via a slice interconnect, which enables non-blocking any-to-any interconnection between slices. For a visual representation of the Cisco CloudScale ASIC slice architecture, see Figure 2.

© 2022 Cisco and/or its affiliates. All rights reserved.

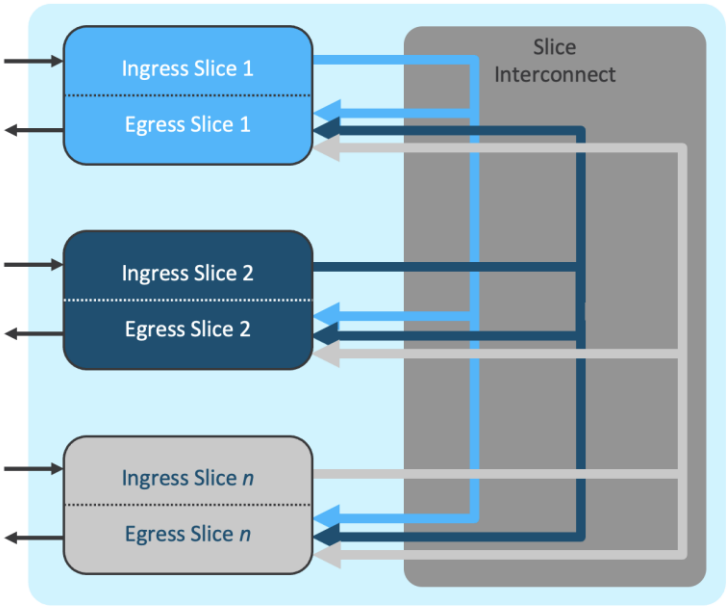


Figure 2. Cisco CloudScale H2R/H1 ASIC slice architecture

CloudScale H2R/H2/H1 ASIC Slice Forwarding Path

Each of the pipeline elements shown in Figure 2 can consist of multiple forwarding processing steps. Figure 3 illustrates the detailed steps in the Cloud Scale ASIC pipelines.

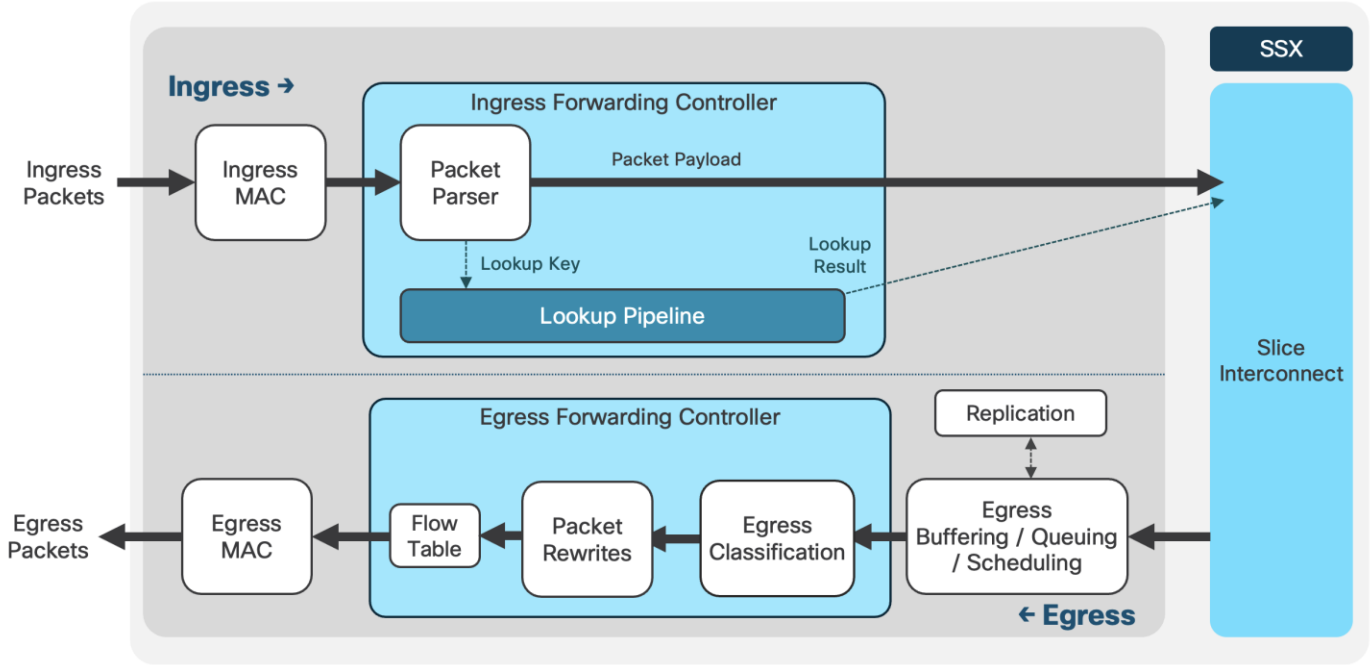


Figure 3. Cisco CouldScale H2R/H2/H1 ASIC Slice Forwarding Path

© 2022 Cisco and/or its affiliates. All rights reserved.

CloudScale H2R/H2/H1 ASIC Ingress Forwarding Controller

The ingress forwarding controller operates as follows: when a packet is received from the MAC layer, it parses the packet headers and conducts several lookups to determine whether the packet should be accepted and how it should be forwarded to its intended destination. Additionally, the controller generates instructions for the data path to handle the storage and queuing of the packet. For a visual representation of this process, see Figure 4, showcasing the CloudScale ASIC ingress forwarding controller.

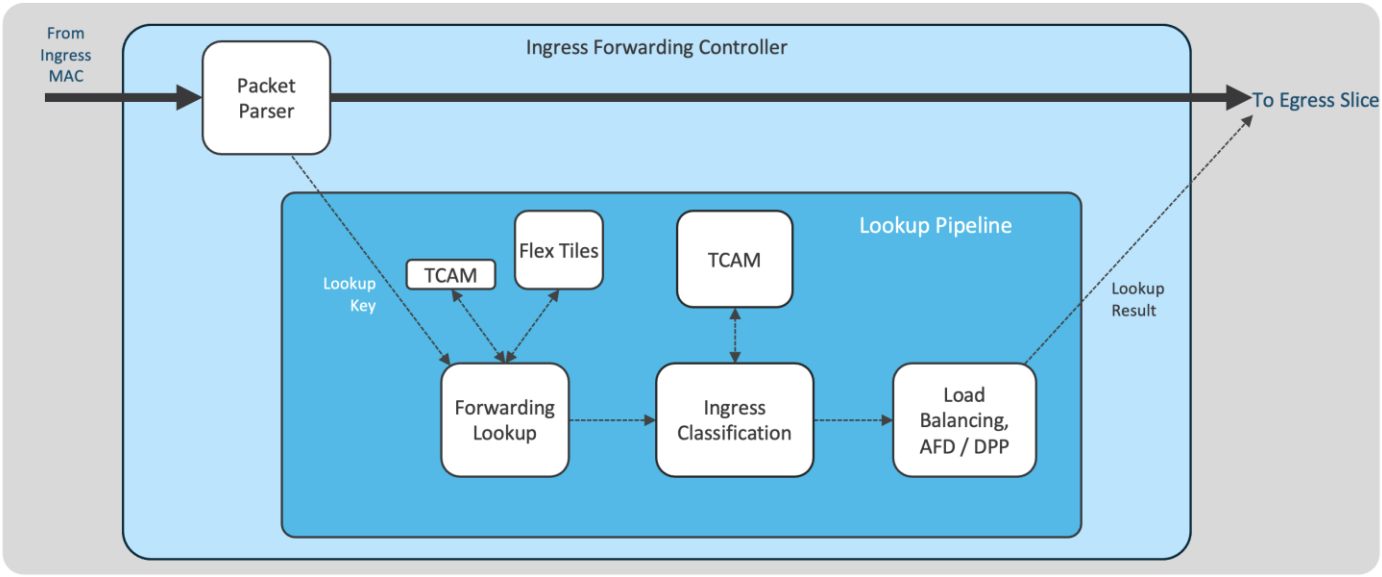


Figure 4. Cisco CouldScale H2R/H2/H1 ASIC ingress forwarding controller

CloudScale H2R/H2/H1 ASIC Egress Forwarding Controller

The egress forwarding controller is tasked with receiving packets from the buffer manager, along with their corresponding metadata, for transmission purposes. Its primary responsibilities include egress classification and managing all packet rewrites. You can observe the representation of the " Egress Forwarding Controller" in Figure 5, where the egress classification takes place. Egress RACLs and VACLs are the most commonly allocated TCAM resources for egress classification. Additionally, Egress PACLs are also supported with H2R, H2 and H1 ASICs.

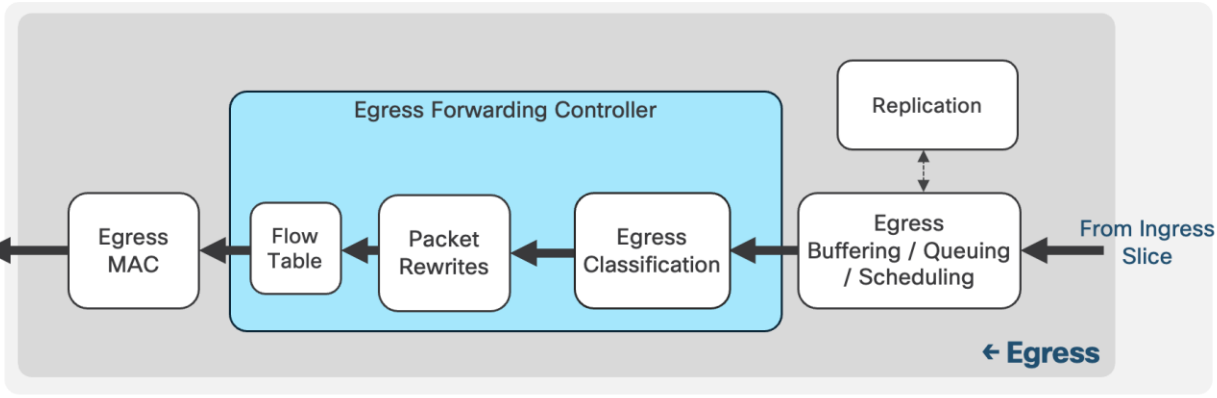


Figure 5. Cisco CouldScale H2R/H2/H1 ASIC egress forwarding controller

© 2022 Cisco and/or its affiliates. All rights reserved.

Cisco CloudScale H2R, H2 and H1 ASIC Forwarding lookup

The forwarding lookup is the initial block in the forwarding pipeline. It comprises two components: the "Forwarding TCAM" and "Flexible tiles". It's important to note that the Forwarding TCAM is not utilized for classification ACLs; instead, a separate resource, the ACL/QoS TCAM, is used for this purpose. The H2R/H2/H1 forwarding lookup block is depicted in Figure 6.

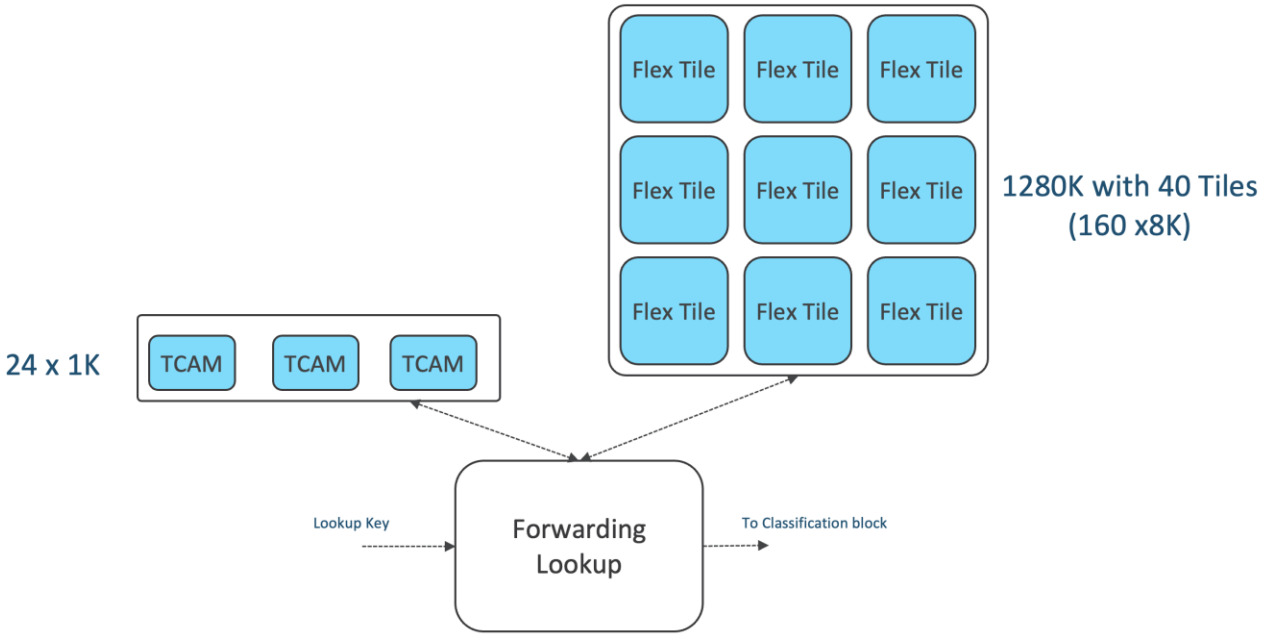


Figure 6. Cisco CouldScale H2R/H2/H1 ASIC Forwding Lookup blocks

CloudScale H2R/H2/H1 ASIC Forwarding TCAM

The Forwarding TCAM, a conventional TCAM memory, manages overflow and hash collisions, serving as the front-end for flexible forwarding lookups. The H2R, H2 and H1 ASICs feature a 24K Forwarding TCAM divided into smaller 1K banks. The default allocation of the Forwarding TCAM is illustrated in Figure 7. By default, ten 1K banks are assigned to IPv4, and twelve 1K banks are allocated to IPv6. When Unicast Reverse Path Forwarding (uRPF) is activated on the Cisco N9300-H, these entries are shared with uRPF.

© 2022 Cisco and/or its affiliates. All rights reserved.

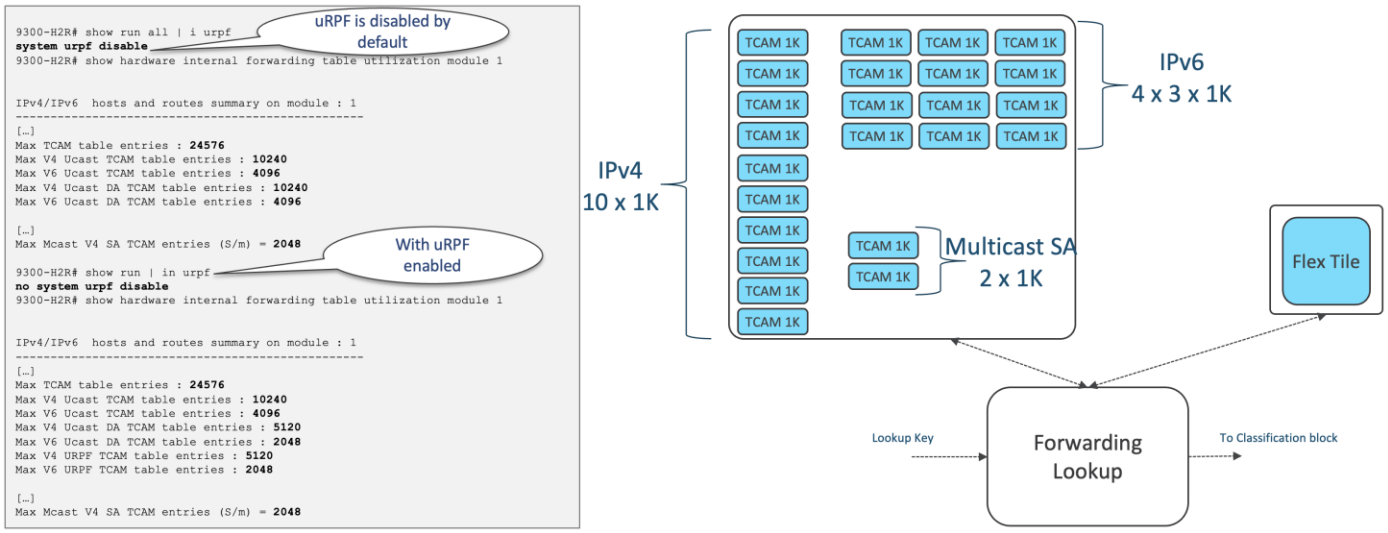


Figure 7. Cisco CouldScale H2R/H2/H1 ASIC Forwding TCAM banks

CloudScale H2R/H2/H1 ASIC Flexible forwarding Tiles (flex tiles)

The "Flexible tiles" serve as the second component of the forwarding lookup. They offer a versatile pool of table entries for lookups, capable of performing a variety of functions, including:

- IPv4/IPv6 unicast longest-prefix match (LPM)
- IPv4/IPv6 unicast host-route table (HRT)
- IPv4/IPv6 multicast (*,G) and (S,G)
- MAC address table / adjacency table
- ECMP tables
- ACI policy

Figure 8 illustrates the "Flexible tiles" component within the forwarding lookup block.

© 2022 Cisco and/or its affiliates. All rights reserved.

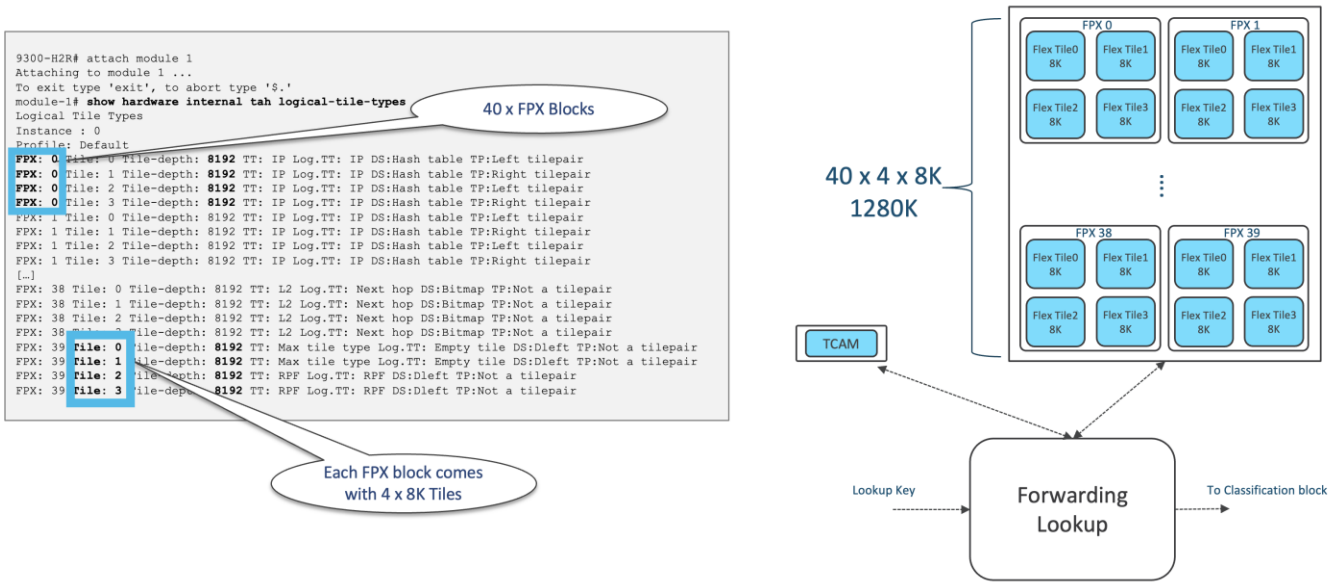


Figure 8. Cisco CouldScale H2RH2//H1 ASIC “Flexible tiles”

The H2R, H2 and H1 ASICs feature a total of 160 x 8K tiles, which can be restructured using the " system routing template" CLI. Figure 9 shows how users can alter the default template, while Table 3 presents the supported forwarding templates for the H2R, H2 and H1 ASICs, including the scalability of each profile. Please refer to the N9000 Verified Scalability Guide (VSG) for additional information on scalability.

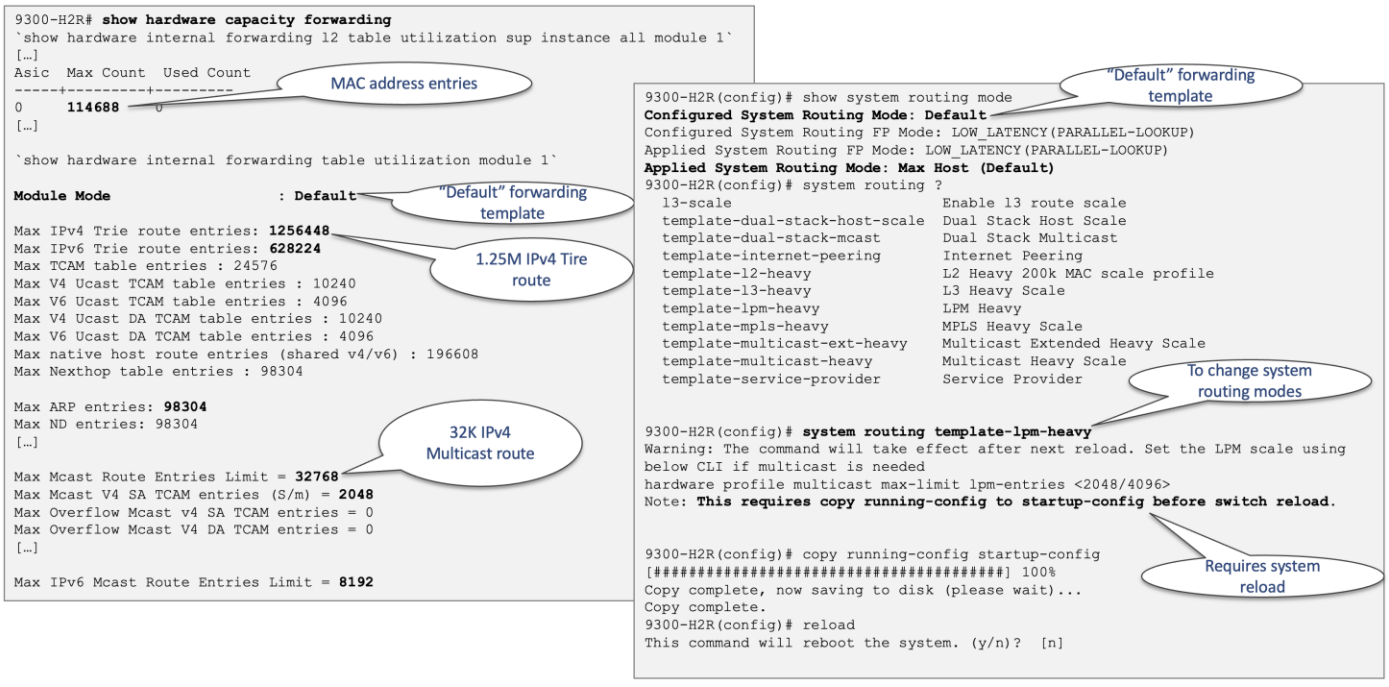


Figure 9. Cisco CouldScale H2R/H2/H1 ASIC Hardware forwarding templates

© 2022 Cisco and/or its affiliates. All rights reserved.

Table 3. Cisco CouldScale H2R/H2/H1 ASIC forwarding templates

Template	MAC	IPv4 Trie	IPv6 Trie	IPv4 TCAM	IPv6 TCAM	IPv4/IPv6 Host	Nexthop Table entries	ARP/ ND	IPv4 Multicast	IPv4 Multicast SA	IPv6 Multicast
Default	114,688	1,256,448	628,224	10,240	4,096	196,608	98,304	98,304	32,768	2,048	8,192
L2 Heavy	262,144	1,256,448	628,224	10,240	4,096	196,608	253,952	98,304	0	2,048	0
Dual Stack Host Scale	114,688	419,430	235,929	10,240	4,096	393,216	131,072	98,304	32,768	2,048	8,192
Internet Peering	32,768	2,000,000	628,224	10,240	4,096	114,688	49,152	49,152	0	2,048	0
LPM Heavy	114,688	1,256,448	628,224	10,240	4,096	196,608	98,304	98,304	8,192	2,048	2,048
MPLS Heavy	98,304	681,574	383,385	10,240	4,096	196,608	196,608	98,304	0	2,048	0
Multicast Heavy	49,152	1,256,448	628,224	9,216	4,096	327,680	65,536	65,536	32,768	1,024	8,192
Multicast-Ext-Heavy	57,344	1,256,448	628,224	9,216	4,096	327,680	65,536	65,536	131,072	1,024	8,192

CloudScale H2R/H2/H1 ASIC Unicast Forwarding

Once the forwarding template configures the flexible tiles, the H2R/H2/H1 ASIC conducts hardware lookups based on the packet type. Specifically, for IP unicast packets, the lookup key comprises the Virtual Routing and Forwarding (VRF) and the IP destination address derived from the packet.

At the same time, the ASIC accesses two distinct tables to find the longest prefix match for the packet's IP address. The first table, the Host Route Table (HRT), includes exact match entries for locally learned /32 or /128 host addresses. Through a hash-based lookup, the ASIC verifies if there's an existing entry for the specific IP address. If such an entry is present, it signifies the longest prefix match.

In parallel, the ASIC conducts a lookup in the Longest Prefix Match (LPM) table. This table permits variable-length prefix matching, enabling the matching of IP prefixes such as /24 or /16. The ASIC identifies which of the two tables has the longest match.

Once the longest prefix match is established, the ASIC moves on to verify if there are multiple paths to the destination address. The destination address could be linked with a single interface or multiple next hops.

If the destination is directly connected and not a multi-path entry, the ASIC fetches an adjacency pointer from the table. This pointer holds information about reaching the next hop, including the output interface, the new destination MAC address (if crossing an L3 boundary), and the MTU.

For multi-path entries, the ASIC consults an ECMP (Equal-Cost Multipath) or load balancing table to choose the suitable next hop from the available options. The ASIC determines which path to select, and the chosen path is assigned an adjacency pointer. Both IPv4 and IPv6 unicast packets undergo these processes. The IP Unicast lookup process is depicted in Figure 10.

© 2022 Cisco and/or its affiliates. All rights reserved.

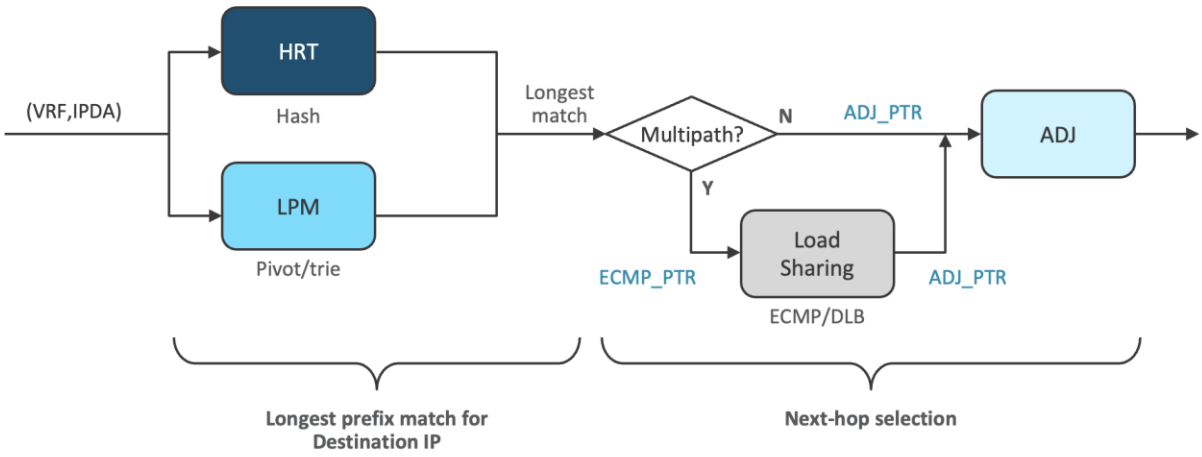


Figure 10. Cisco CouldScale H2R/H2/H1 ASIC unicast forwarding

CloudScale H2R/H2/H1 ASIC Multicast Forwarding

The process varies slightly when handling multicast packets. An exact match lookup is conducted in the Host Route Table (HRT), involving the VRF, IP source address, and IP destination address. These lookups aim to identify mroute entries by matching the source and destination group IP addresses in the packet. The objective is to discern whether the mroute is a specific (S,G) or shared tree (star,G) entry.

If a match is detected in the HRT, the system refers to a separate table known as MC info. This additional table, implemented using flex tiles, holds information like the RFP (Reverse Path Forwarding) interface and the MET (Multicast Expansion Table) pointer.In multicast, an RPF check is necessary for each packet to ensure that we are not creating a loop. If a packet arrives from a non-RPF interface, it will be dropped to prevent looping.

Assuming the RPF check passes, a MET pointer is acquired. The MET table is found at the egress side of the slice. When the packet reaches the egress side (as seen in the packet walk), the MET table decides the number of packet copies required based on the OIL (Outgoing Interface List). The OIL designates the interfaces where the multicast packet should be duplicated.

Overall, this process guarantees efficient multicast forwarding, averting loops and determining suitable replication of multicast packets based on the mroute entries and OIL. The Multicast lookup process is illustrated in Figure 11.

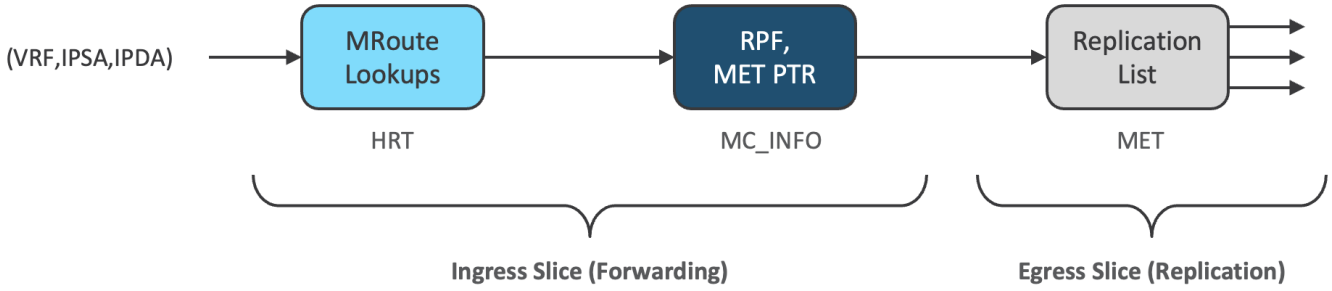


Figure 11. Cisco CouldScale H2R/H2/H1 ASIC multicast forwarding

© 2022 Cisco and/or its affiliates. All rights reserved.

CloudScale H2R/H2/H1 ASIC VXLAN Forwarding

VXLAN and other tunnel encapsulation/ decapsulation performed in single pass with Cisco CloudScale H2R/H2/H1 ASIC.

In the ingress direction, the adjacency signifies the existence of a tunnel. This tunnel initiates a rewrite function for the outer header, including MAC and IP addresses. The actual rewrite operation takes place during the data processing's egress phase. When a packet enters the H2R/H2/H1 ASIC, it scrutinizes the inner packet to determine if it's destined for a tunnel. The ASIC then chooses the appropriate tunnel and appends the necessary information to the packet, such as the outer source MAC address and VXLAN VNID.

On the other hand, in the reverse direction, after the packets have been subjected to lookup processes, the parser recognizes them as tunnel packets. It captures both the inner and outer packet header fields. The ASIC's Tunnel End Point (TEP) table is consulted to ascertain whether the outer header's destination address corresponds to a local tunnel within the switch. If it does, additional lookup operations are performed on the inner header. If it doesn't match a local tunnel, the packet is forwarded based solely on the outer header. For packets destined for local tunnels, packet termination is required. This process involves conducting inner Layer 2/Layer 3 lookups and, if necessary, the rewrite block removes the VXLAN encapsulation and makes any necessary adjustments to the original inner frame before forwarding it. Figure 12 depicts both the encapsulation and decapsulation steps.

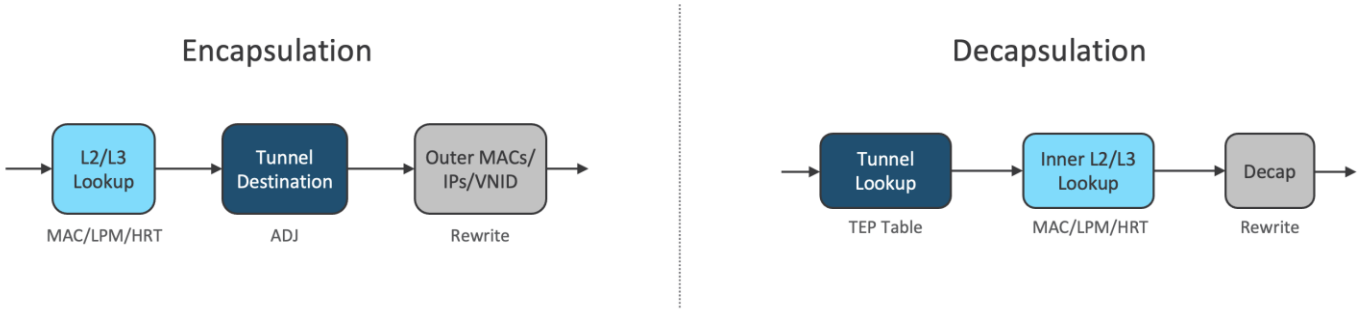


Figure 12. Cisco CouldScale H2R/H2/H1 ASIC VXLAN forwarding

Cisco CloudScale H2R, H2 and H1 ASIC Clasification TCAM

Following forwarding lookups, the packet proceeds to ingress classification processing. The ingress matches are checked against the classification TCAM. These ACLs include various types such as Routed ACLs (RACLs), VLAN ACLs (VACLs), Port ACLs (PACLs), Quality of Service (QoS), Network Address Translation (NAT) ACLs, and more.

The classification entries are restricted to individual ASIC slices and are programmed only where needed. This method enhances the utilization of the classification TCAM in the Cisco N9300-H platform switches. The H2R, H2 and H1 ASICs share a 14K TCAM space, which caters to both ingress and egress classifications. The Cisco CloudScale H2R/H1 classification TCAM architecture is demonstrated in Figure 13.

© 2022 Cisco and/or its affiliates. All rights reserved.

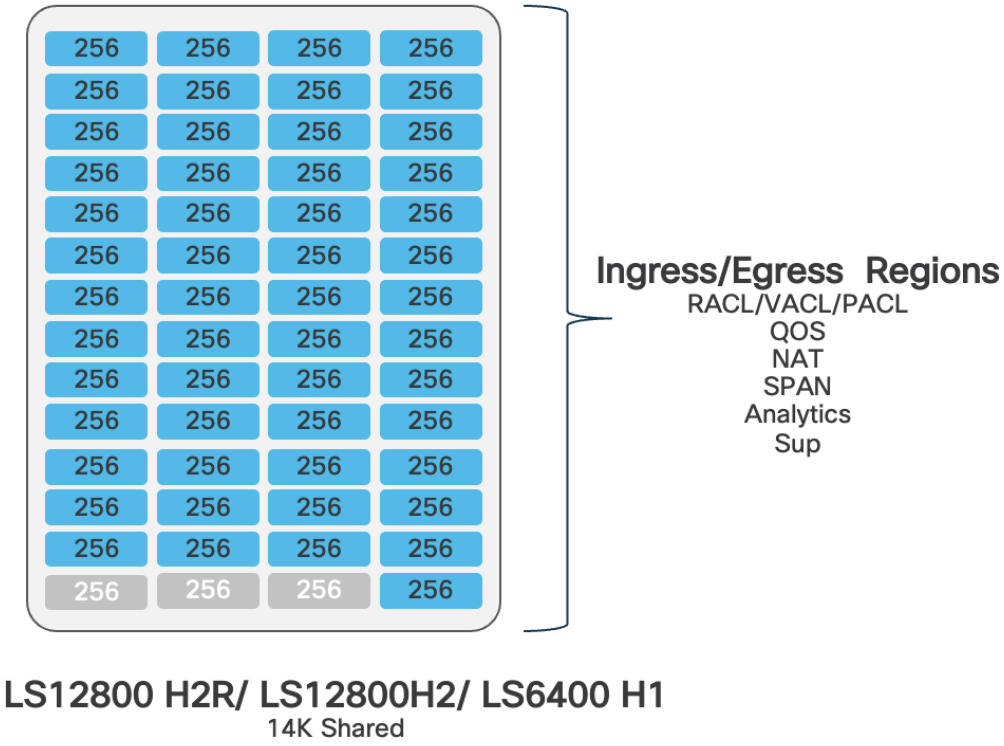


Figure 13. Cisco CouldScale H2R/H2/H1 ASIC ACL TCAM architecture

CloudScale H2R/H2/H1 ASIC TCAM Scalability

The CloudScale H2R/H2/H1 ASIC TCAM is optimized for scalability through slice-aware policy programming. PACL and RACLs are programmed only on slices where ports with these ACLs are present, and port channel ACLs are programmed solely on slices where the respective members exist. Similarly, VLAN ACLs are programmed exclusively on slices where the corresponding VLAN is present. To enable policy sharing, labels are used with label space localized to each slice. For Cisco CloudScale H2R/H2/H1 ASIC ACL TCAM scalability numbers, see Table 4. It's important to note that the TCAM scalability numbers were verified with all TCAM regions freed up except for the 'ing-sup' and 'egr-sup' regions.

Table 4. Cisco Cloud Scale H2R/H2/H1 ASIC ACL TCAM scalability numbers

ASIC	Ingress PACL		Egress PACL		Ingress RACL		Egress RACL		Ingress VACL		Egress VACL	
Name/Slices	Per Slice	Total	Per Slice	Total	Per Slice	Total	Per Slice	Total	Per Slice	Total	Per Slice	Total
LS12800 H2/R (4 slices)	13568	54272	13568	54272	13568	54272	13568	54272	13568	54272	13568	54272
LS6400 H1 (2 slices)	13568	27136	13568	27136	13568	27136	13568	27136	13568	27136	13568	27136

CloudScale H2R/H2/H1 ASIC Default ACL TCAM Allocation

The allocation of both ingress and egress ACL TCAM to different ACL types is user configurable. Each ACL type requires its dedicated bank/banks, and ACL programming is localized on a per ASIC basis. ACL

© 2022 Cisco and/or its affiliates. All rights reserved.

entries are programmed into the TCAM only where they are required. By default, all available TCAM regions are allocated.

For the Cisco N9300-H2R, N9300-H2 and N9300-H1, they are the only N9000 platforms featuring shared TCAM space for both Ingress and Egress slices. By default, 10K is assigned for ingress and 4K for egress, but any bank can be reconfigured to serve either as ingress or egress.

To visualize the default classification TCAM regions for the N9300-H2R/H2/H1 series switches, refer to Figure 14.

```
9300-H2R# show hardware access-list tcam region
                                  NAT ACL[nat] size =  512
                          Ingress PACL [ing-ifacl] size = 1536
                                  VACL [vacl] size = 1024
                            Ingress RACL [ing-racl] size = 4096
                       Ingress L2 QOS [ing-l2-qos] size =  512
             Ingress L3/VLAN QOS [ing-l3-vlan-qos] size =  512
                            Ingress SUP [ing-sup] size =  512
     Ingress L2 SPAN filter [ing-l2-span-filter] size =  256
     Ingress L3 SPAN filter [ing-l3-span-filter] size =  256
                          Ingress FSTAT [ing-fstat] size =    0
                                  span [span] size =  512
                            Egress RACL [egr-racl] size = 2816
                              Egress SUP [egr-sup] size =  256
                  Ingress Redirect [ing-redirect] size =    0
                       Egress L2 QOS [egr-l2-qos] size =    0
             Egress L3/VLAN QOS [egr-l3-vlan-qos] size =    0
        Ingress Netflow/Analytics [ing-netflow] size =  512
                          Ingress NBM [ing-nbm] size =    0
                           TCP NAT ACL[tcp-nat] size =    0
               Egress sup control plane[egr-copp] size =    0
        Ingress Flow Redirect [ing-flow-redirect] size =    0
     Ingress PACL IPv4 Lite [ing-ifacl-ipv4-lite] size =    0
     Ingress PACL IPv6 Lite [ing-ifacl-ipv6-lite] size =    0
                    Ingress CNTACL [ing-cntacl] size =    0
                    Egress CNTACL [egr-cntacl] size =    0
                     MCAST NAT ACL[mcast-nat] size =    0
                      Ingress DACL [ing-dacl] size =    0
       Ingress PACL Super Bridge [ing-pacl-sb] size =    0
      Ingress Storm Control [ing-storm-control] size =    0
           Ingress VACL redirect [ing-vacl-nh] size =    0
                        Egress PACL [egr-ifacl] size =    0
                  Egress Netflow [egr-netflow] size =    0
                        Egress SPAN [egr-span] size =    0
```

```
9300-H2R# show system internal access-list globals
-------------------------------------------------------------------------------
                INSTANCE 0 TCAM Region Information:
-------------------------------------------------------------------------------
Ingress:
--------
                     Region          TID     Base     Size     Width
-------------------------------------------------------------------------------
            Ingress PACL ALL           4        0     1536         1
            Ingress RACL ALL           7        0     4096         1
            Ingress VACL ALL          11        0     1024         1
                Ingress PQOS          15        0      512         1
           Ingress L3QOS ALL          22        0      512         1
                 Ingress SUP          23        0      512         1
            Ingress MAC SPAN          25        0      256         1
                Ingress SPAN          28        0      512         1
           Ingress Vlan SPAN          29        0      256         1
   Ingress Netflow/Analytics          31        0      512         1
                         NAT          66        0      512         1
-------------------------------------------------------------------------------
Total configured tcam size: 10240
Egress:
--------
                     Region          TID     Base     Size     Width
-------------------------------------------------------------------------------

            Egress RACL ALL          42        0     2816         1
            Egress VACL ALL          46        0     1024         1
             Egress SUP ALL          62        0      256         1
-------------------------------------------------------------------------------
Total configured tcam size: 4096

-------------------------------------------------------------------------------
Remaining free tcam size (can be used for ingress and egress): 0
Unusable tcam size due to fragmentation - ingress: 0, egress: 0

-------------------------------------------------------------------------------
```

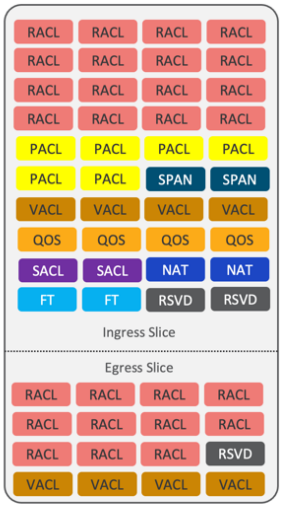


Figure 14. Default classification TCAM regions for Cisco Nexus 9300-H2R/H2/H1 chassis

CloudScale H2R/H2/H1 ASIC Egress RACL Example

The H2R, H2 and H1 ASIC offers the greatest number of TCAM entries for both Ingress and Egress regions. The maximum Egress RACL allocation with H2R/H2/H1 ASICs is illustrated in the example shown in Figure 15. To achieve the maximum Egress RACL, the user must first deallocate the default regions.

© 2022 Cisco and/or its affiliates. All rights reserved.

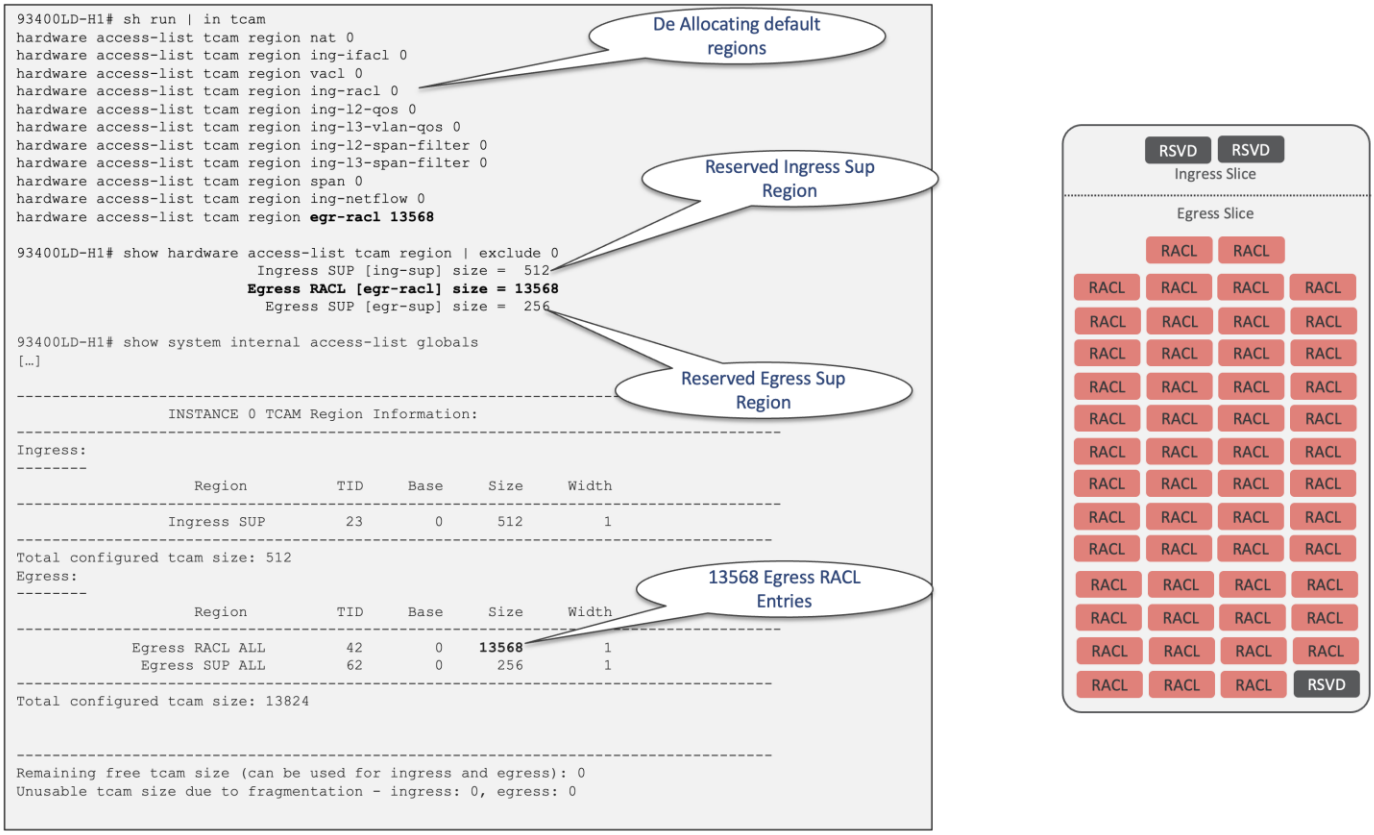


Figure 15. Maximum Egress RACL allocation with Nexus 9300-H2R/H2/H1 chassis

Cisco CloudScale H2R, H2 and H1 ASIC Hardware Telemetry

Besides Software Telemetry, Cisco H2R, H2 and H1 ASICs also incorporate Hardware Telemetry, further divided into two primary components. These components include 'data plane flow data,' obtained from the actual traffic moving through the forwarding pipeline, and 'basic ASIC data.' The latter signifies statistics that can be extracted from the ASIC at remarkably high rates. These telemetry features are visually represented in Figure 16. It should be noted that Nexus Dashboard (ND) currently uses Flow table (FT), Flow Table Events (FTE), and Streaming Statistics (SSX). Furthermore, at FCS, there is no software support available for In-Band Network Telemetry (INT).

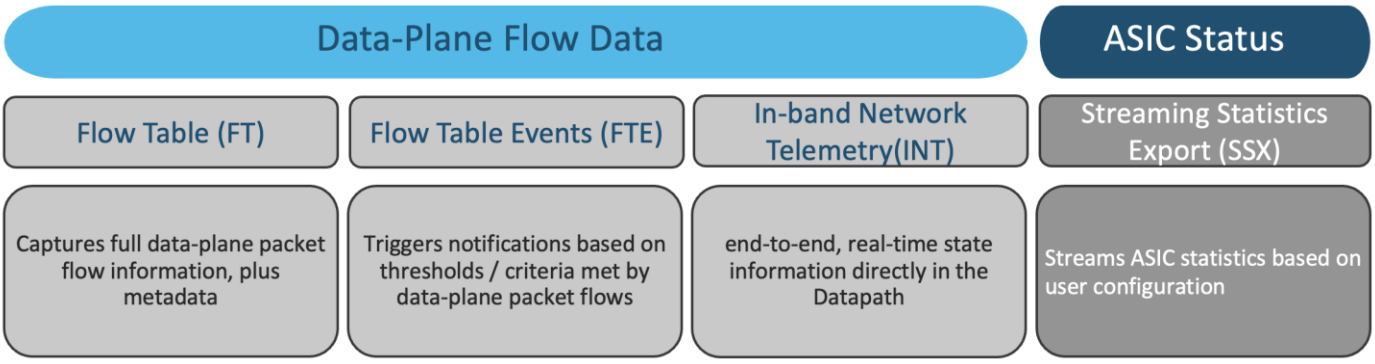


Figure 16. Cisco CloudScale H2R/H2/H1 ASIC Hardware Telemery

© 2022 Cisco and/or its affiliates. All rights reserved.

CloudScale H2R/H2/H1 ASIC Flow Table (FT), Flow Table Event (FTE) and Netflow

The flow table, in conjunction with H2R, H2 and H1 ASICs, serves the purpose of gathering comprehensive information pertaining to 5-tuple flows and additional metadata related to packet forwarding. This includes:

- 5-tuple flow information
- Interface/queue details
- Flow start/stop timestamps
- Packet disposition indicators (for drop events)
- Burst measurements

A notable feature of the flow table in the context of H2R, H2 and H1 ASICs is that each Egress slice possesses a dedicated flow table, and the size of each table has been expanded to accommodate 128,000 entries per slice. What sets the Flow table apart is its capability to export flow data directly from the ASIC without relying on CPU intervention. The ASIC can encapsulate the flow records into packets and transmit them directly to the front panel port, all without involving the CPU or software. This approach ensures minimal impact on switch performance.

In addition, flow table event (FTE) triggers notifications based on criteria or thresholds met by data-plane packet flows.

These recorded flow table entries find application in Nexus Dashboard (ND) and netflow analysis. In Figure 17, we observe the utilization of the Flow table in conjunction with ND, while Figure 18 illustrates the utilization of the Flow Table in the context of Netflow.

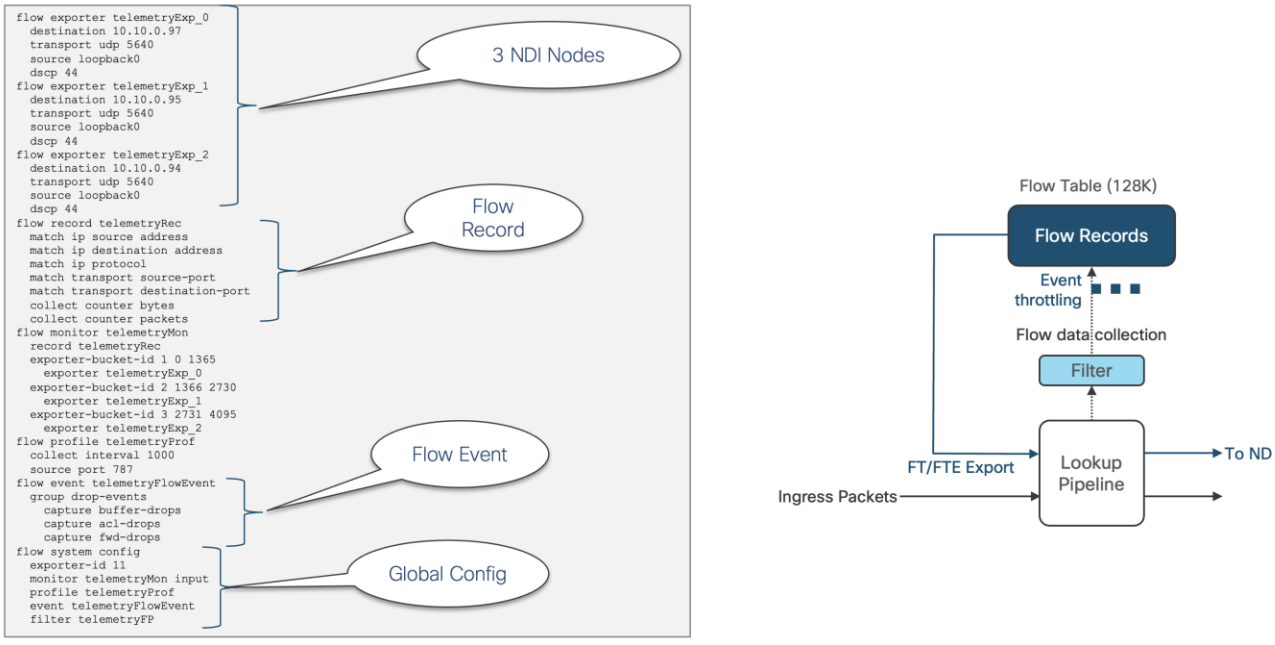


Figure 17. Flow Table useage with Nexus Dashboard (ND)

© 2022 Cisco and/or its affiliates. All rights reserved.

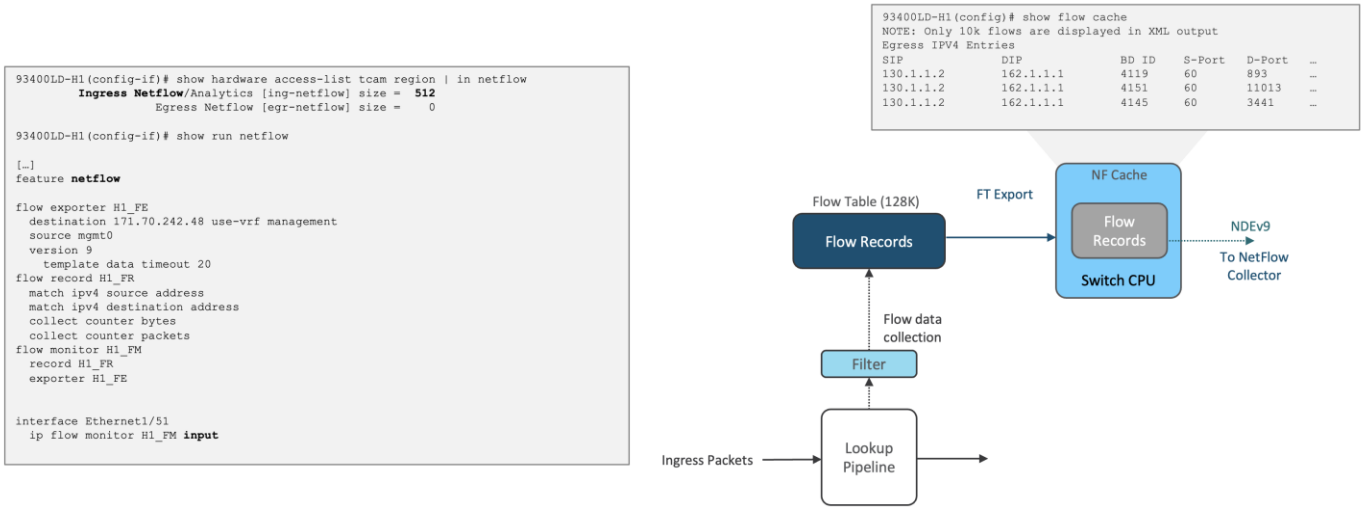


Figure 18. Flow Table useage with Netflow

CloudScale H2R/H2/H1 ASIC Inband Network Telemetry (INT)

INT Is a framework that is designed to monitor, collect, and report flows and network states, by the data plane. INT sources embed the monitored INT flow information in normal data packets, and all the downstream devices add the same information with their details that are based on the source information received. Is used to achieve per-packet network visibility with low overheads. First record (INT instruction +metadata) will be inserted in data packet at INT Source node where as the second record (INT metadata) will be appended to same data packet at INT Transit node. The third Record (INT metadata) will be appended to INT stack at INT Sink. INT Sink will remove INT record and forward to INT Collector while original packet will be forwarded to server facing port. Figure 19 illustrate the INT architecture while figure 20 extands on the INT modes.

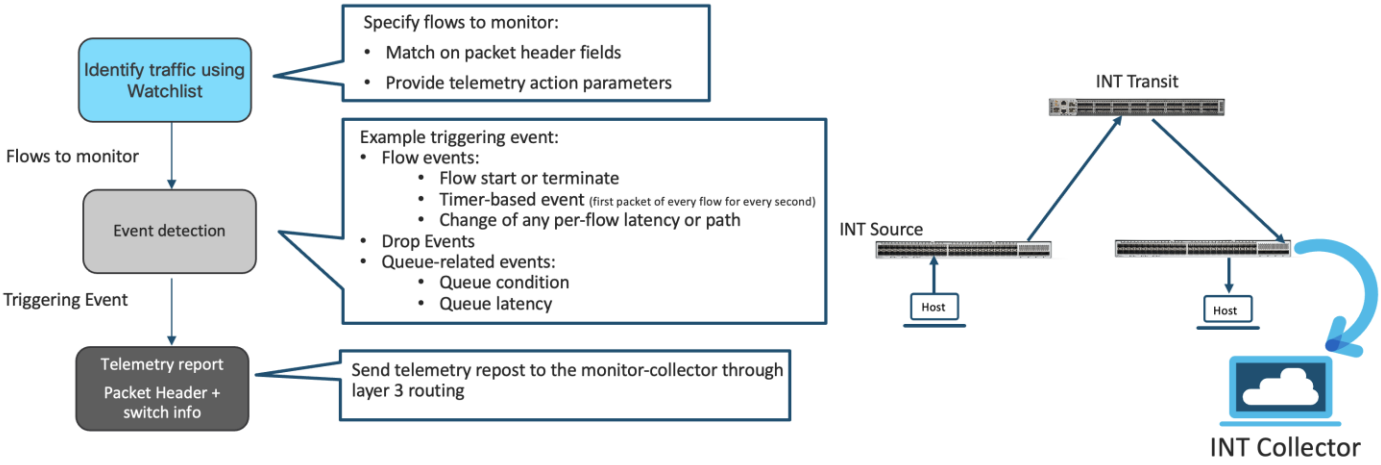


Figure 19. INT architecture

© 2022 Cisco and/or its affiliates. All rights reserved.

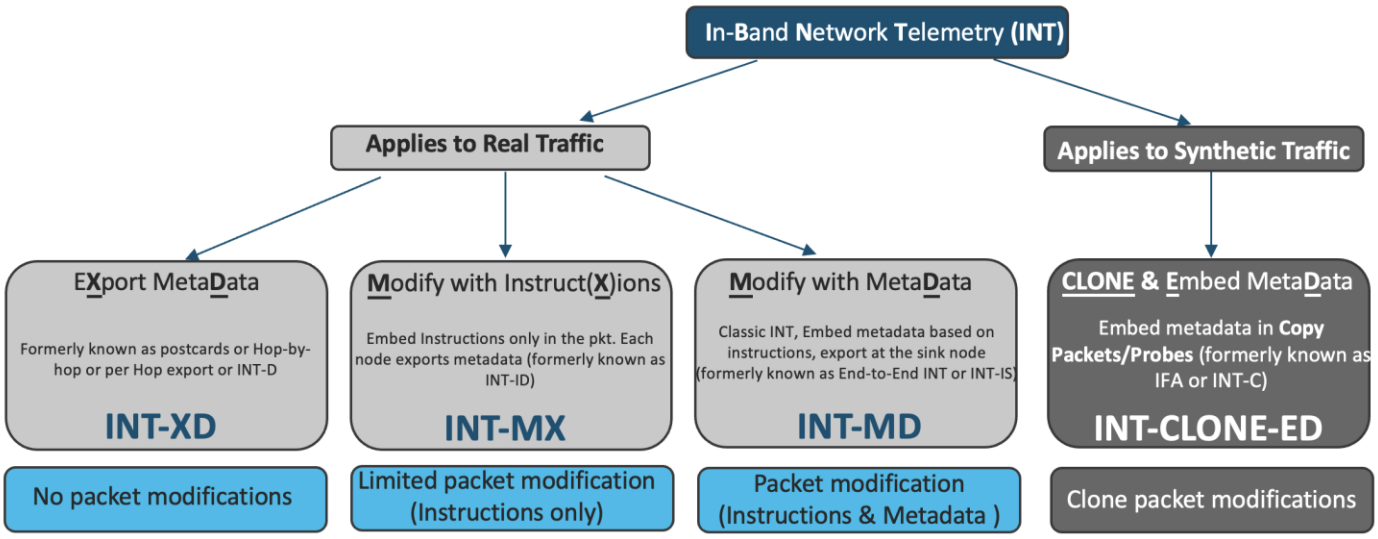


Figure 20. INT supported modes with H2R/H2/H1 ASIC

Cisco H2R, H2 and H1 ASICs support INT version 1, version 2 and overload-INT for backward compatibility. ASIC supports all the three INT v2 modes:

- XD: Also called as postcard mode. Each node in the INT domain exports metadata based on the watchlist configuration.
- MX: Each node exports metadata based on the instructions embedded in the packet. Each intermediate node embeds instruction set in the packet.
- MD: Each transit node, starting from source, embed instructions and metadata in the packet, at the end the sink node, strips the complete metadata stack and exports it to the collector.

Figure 21 presents the INT export fields supported by the hardware in the Cisco CloudScale H2R/H2/H1 ASICs.

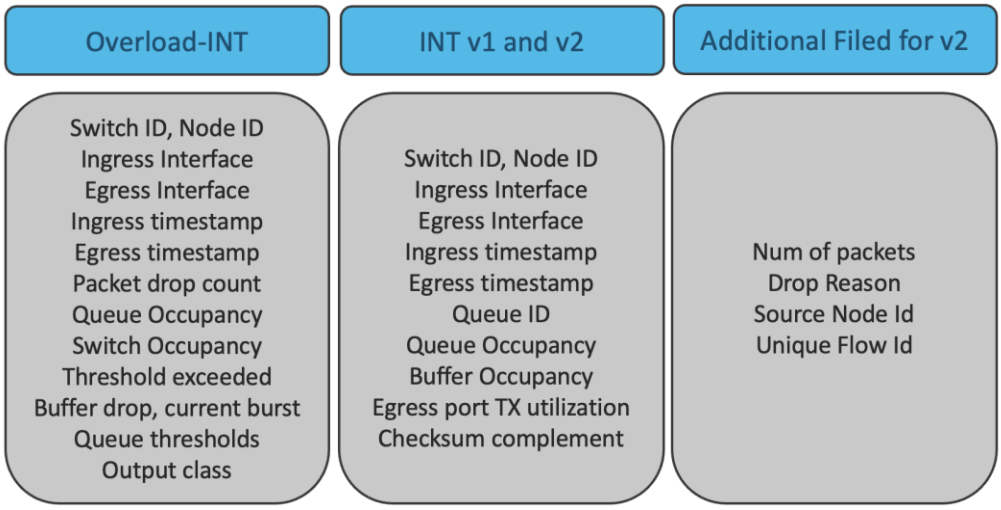


Figure 21. Cisco CloudScale H2R/H2/H1 ASIC HW supported Exported Fileds

© 2022 Cisco and/or its affiliates. All rights reserved.

CloudScale H2R/H2/H1 Streaming Statistics Export (SSX)

SSX swiftly streams ASIC statistics based on user configurations. Here are some of the statistics supported by SSX:

- Interface counters (RMON counters)
- Ingress/egress queue depth and queue drops
- Egress total buffer depth
- Egress queue microbursts

The user has the ability to define streaming parameters which include deciding which statistics to use, determining their frequency, and selecting the collector to which they are sent. Another key feature is the direct export from the ASIC to the front-panel port. This process eliminates the need for switch CPU involvement, making it more efficient.

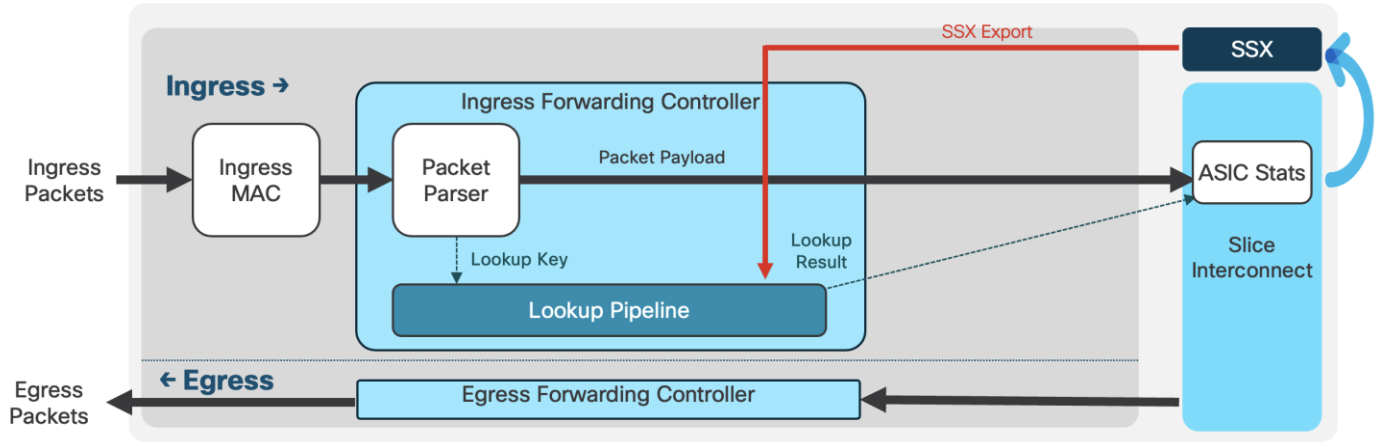


Figure 22. Streaming Statistics Export (SSX) with Cisco H2R/H2/H1 ASIC

Cisco CloudScale H2R, H2 and H1 Span on Drop (SOD)

Cloudscale H2R/H2/H1 ASICs offer a crucial feature of mirroring dropped packets within the switch, which could be a result of configuration issues or a lack of buffer/queue space. Packet drops, though generally unwanted, are sometimes inevitable. The ability to mirror these dropped packets, also known as SPAN on Drop or SOD, allows network administrators to analyze the situation. This analysis can aid in fine-tuning system configurations or adjusting traffic flow to reduce packet drops. As displayed with figure 22, the process of SPAN on Drop involves three steps: detecting the dropped packet, generating a copy of it, and then transmitting this dropped packet outside the switch.

© 2022 Cisco and/or its affiliates. All rights reserved.

```
93400LD-H1(config)# show run monitor
<snip>

version 10.4(2) Bios:version 02.07
monitor session 1 type erspan-source
  erspan-id 3
  vrf default
  destination ip 10.20.1.2
  source forward-drops rx
  no shut

monitor erspan origin ip-address 10.20.1.1 global

93400LD-H1(config)# sh monitor session 1
   session 1
---------------
type              : erspan-source
state             : up
erspan-id         : 3
vrf-name          : default
acl-name          : acl-name not specified
ip-ttl            : 255
ip-dscp           : 0
destination-ip    : 10.20.1.2
origin-ip         : 10.20.1.1 (global)
source intf       :
    rx            :
    tx            :
    both          :
source VLANs      :
    rx            :
    tx            :
    both          :
filter VLANs      : filter not specified
source fwd drops  : high priority

marker-packet     : disabled
packet interval   : 100
packet sent       : 0
packet failed     : 0
egress-intf       : Eth1/51
source VSANs      :
    rx            :
```

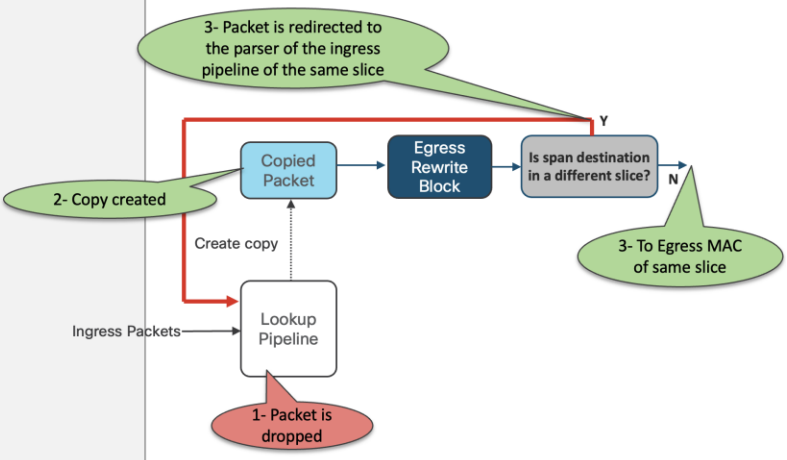


Figure 22. The process of SPAN on Drop with Cisco CloudScale H2R/H2/H1 ASIC

A packet might be dropped for numerous reasons at different stages. From a mirroring standpoint, H2R/H2/H1 ASIC support the following types of drops: those occurring in the ingress lookup pipeline or as a result of ingress forwarding, those happening in the egress buffer manager due to buffer occupancy, and those in the egress rewrite engine. Figure 23 ilusstrate the potential packet drop points in the H2R/H2/H1 forwarding path.

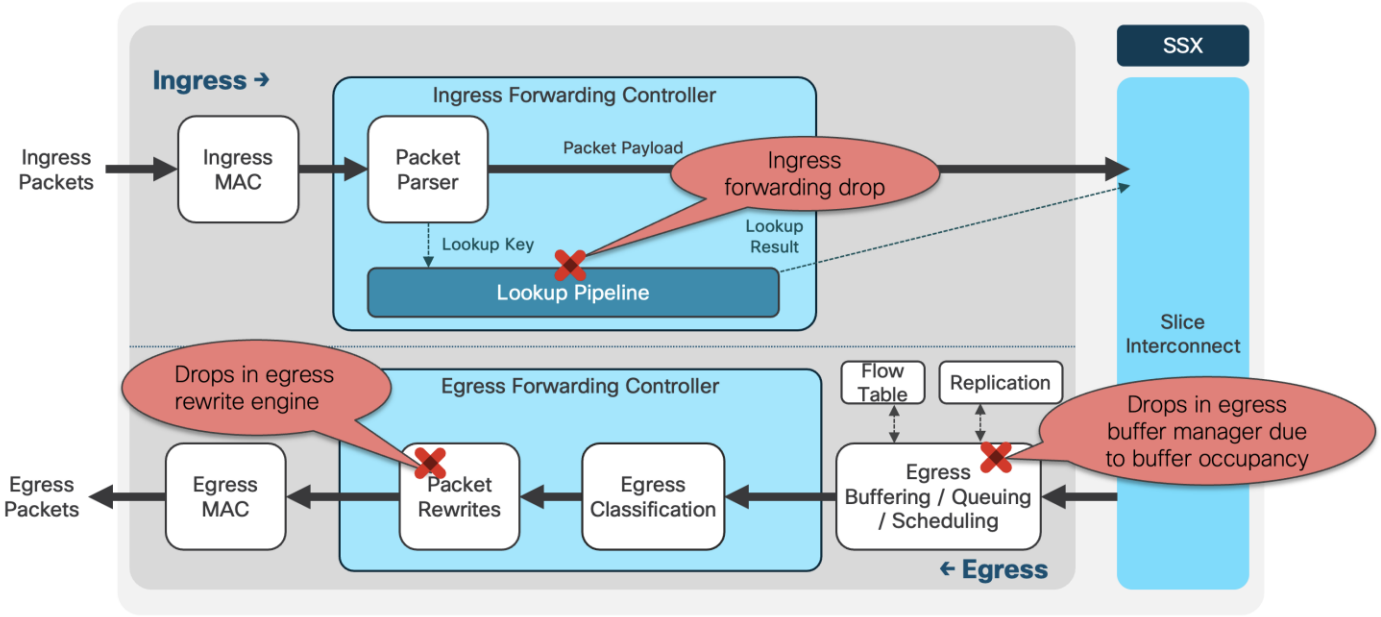


Figure 23. The potential packet drop points in the H2R/H2/H1 forwarding path

© 2022 Cisco and/or its affiliates. All rights reserved.

CloudScale H2R/H2/H1 ASIC Ingress Forwarding Drop

A packet may get dropped in ingress pipeline due to various reasons including:

- VLAN mismatch
- MTU exception
- Time To Live (TTL) expiration
- Missing Forwarding Information Base (Fib Miss)

If a packet gets discarded, the egress buffer manager won't accept it for any forwarding action. Span supports all the ingress forwarding drops except ACL drops.

Cloudscale H2R/H2/H1 ASICs use following steps to mirror a dropped packet:

1. SOD cap bit is set in internal header if a dropped packet is to be mirrored. NX-OS implementation is to have a catch all entry and mark all packets with SOD cap bit.
2. Egress buffer manager will accept a dropped packet if SOD cap bit is set. It will set mirror action in the opcode and assigns a span index to be this packet.
3. The packet is set to be placed in the replication block, where it locates the span destination port using span table lookups guided by the span-index. After this, the replication block duplicates the packet. The replicated packet is then lined up to be transmitted out of the destination.
4. Packet be passed to packet rewrite block which will add ERSPAM-GRE header to the packet and send packet to egress MAC. Egress MAC will put packet on the wire.

Figure 24 ilusstrate the CloudScale H2R/H2/H1 SOD process.

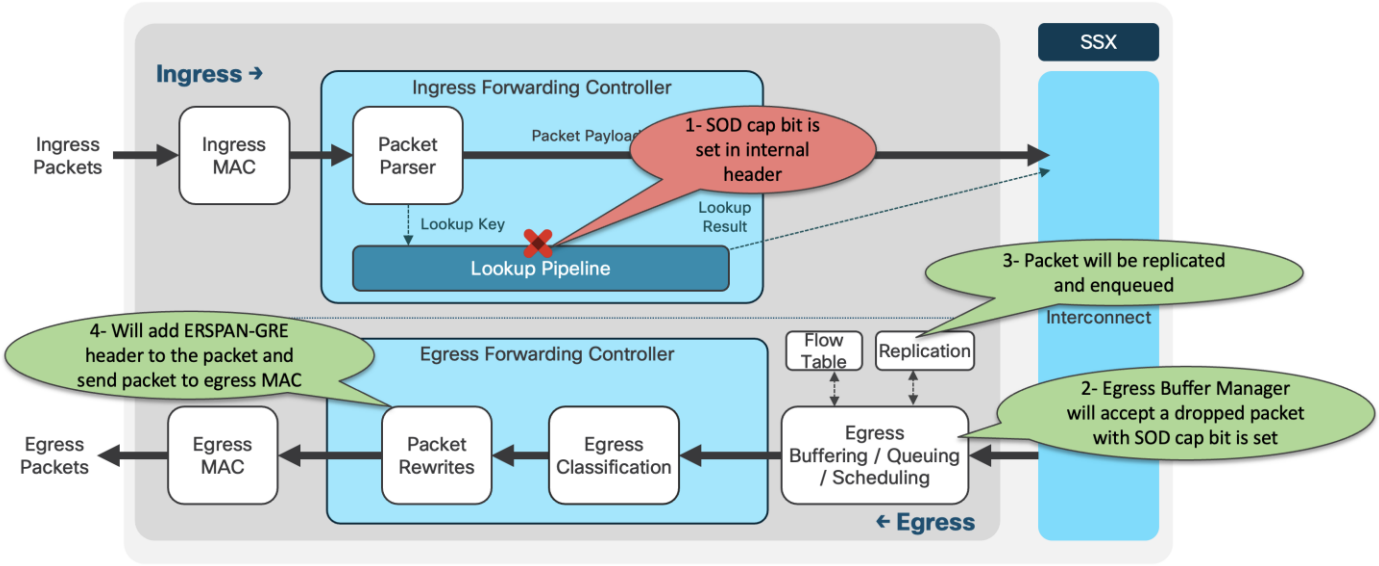


Figure 24. Cisco CloudScale H2R/H2/H1 SOD process

CloudScale H2R/H2/H1 ASIC Egress Buffer Drop

A buffer drop may occur if the egress buffer manager discards the packet due to memory limitations such as oversubscription drop and tail drop. This is often the case when the ingress rate exceeds the egress rate, leading to buffer congestion and the subsequent dropping of the unicast packet.

© 2022 Cisco and/or its affiliates. All rights reserved.

If the SOD cap bit is set, the egress buffer manager will convert the opcode to span for these packets and assign a span-index. The rest of the procedures will continue as per the ingress forwarding drop.

CloudScale H2R/H2/H1 ASIC Egress Rewrite Engine Drop

With the H2R/H2/H1 ASIC, any unicast packet that is dropped during the rewrite process can be mirrored. The Egress Packet Rewrite block conducts an Egress TCAM lookup and during the ACL lookup, a packet can be tagged with an Egress SOD cap. If a packet is dropped in the Egress Rewrite block, it will assign a span-index to the packet and notify the buffer manager. The buffer manager then transfers it to the replication block. Following this, the process continues as it does with an ingress forwarding drop.

CloudScale H2R/H2/H1 ASIC Packet Recirculation

If the span destination is in the same slice where SOD processing is taking place, it's forwarded from the egress pipeline. However, if it's in a different slice, the packet, after ERSPAN encapsulation in the Egress Rewrite block, is redirected to the parser of the ingress pipeline of the same slice using the recirculation port. Afterward, processing occurs as if the packet was received from a normal interface. Therefore, forwarding is based on the GRE-ERSPAN header. These packets are constrained by the recirculation bandwidth. The H2R/H2/H1 ASIC provides two recirculation interfaces for each slice of 200 Gig. Span-on-drop uses one recirculation port, so the rate of span traffic is capped at 200 Gig. Figure 25 ilusstrate the CloudScale H2R/H2/H1 ASIC Packet Recirculation.

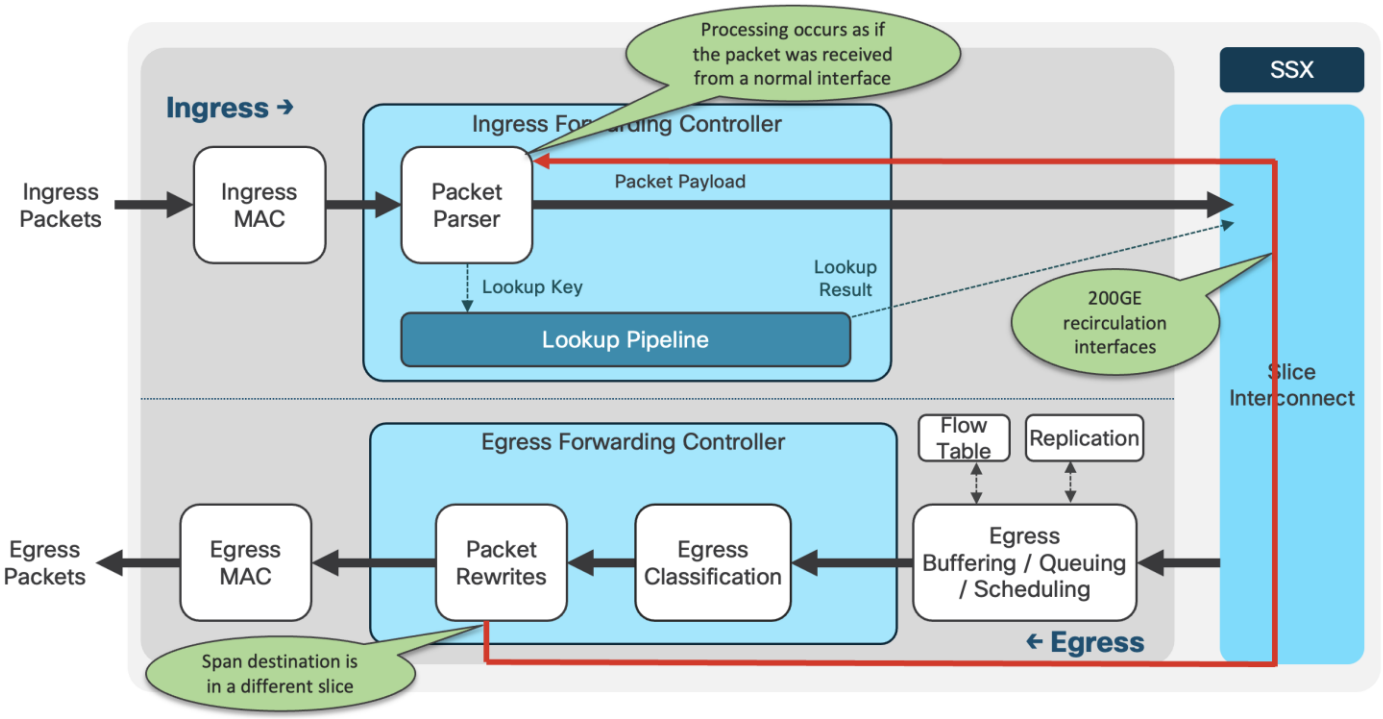


Figure 25. Cisco CloudScale H2R/H2//H1 ASIC Packet Recirculation

Cisco CloudScale H2R, H2 and H1 ASIC Packet Buffer

© 2022 Cisco and/or its affiliates. All rights reserved.

CloudScale H2R/H2/H1 ASIC On-Die packet Buffer

The shared-memory egress buffered architecture is implemented in the Cisco CloudScale H2R/H2/H1 ASICs. The H2/R ASIC is equipped with an 80MB total, where each slide shares 40MB of the buffer. On the other hand, the H1 ASIC features a completely shared on-die packet buffer of 40MB. The structure of the on-die packet buffer for H2R, H2 and H1 ASICs is displayed in Figure 26. The 40MB is distributed into 32 banks, each comprising of 2K cells, with every cell containing 624Bytes.

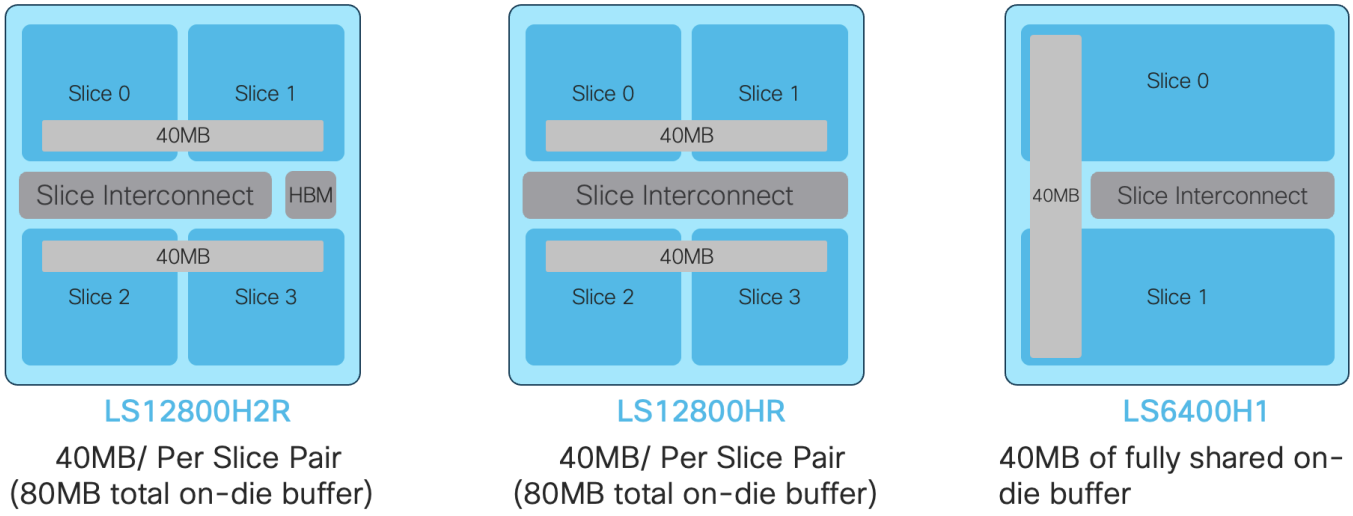


Figure 26. the on-dipe packet buffer atricturef for H2R, H2 and H1 ASICs

The CloudScale H2R/H2/H1 ASICs offer two cell allocation modes. In the default allocation, most of the buffer cells are dedicated to the user-defined space. When set to the 'no drop' mode, a separate space is created for the 'no drop' traffic classes. The cell allocations for the H2R in both 'drop' and 'no-drop' modes are illustrated in Figures 27 and 28, respectively. Figures 29 and 30 show these allocations for the H2 ASIC, while Figures 31 and 32 display these allocations for the H1 ASIC across two slices.

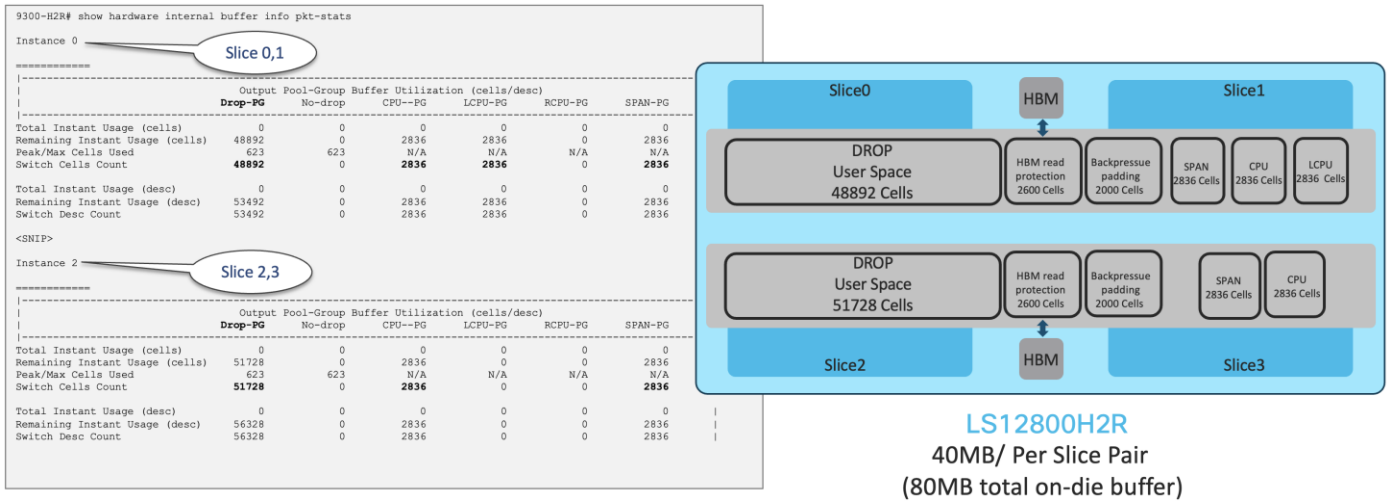


Figure 27. The default mode one-die packet buffer allocation for H2R

© 2022 Cisco and/or its affiliates. All rights reserved.

```
9300-H2R# show hardware internal buffer info pkt-stats

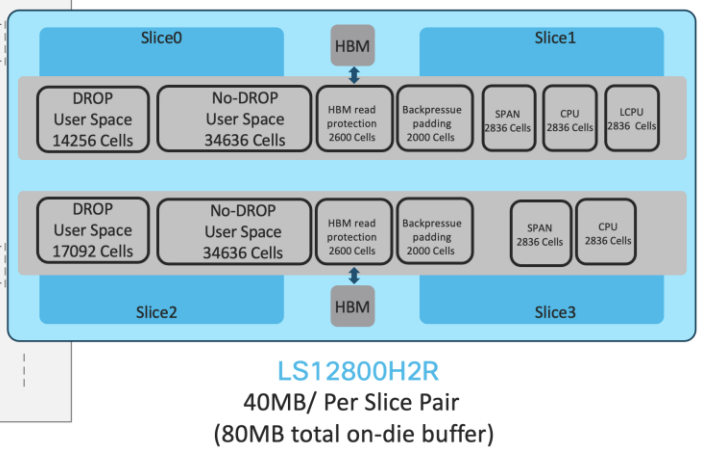
Instance 0                                          Slice 0,1
============
|-------------------------------------------------------------------------------------------------------|
|                                Output Pool-Group Buffer Utilization (cells/desc)                      |
|                          Drop-PG      No-drop      CPU--PG      LCPU-PG      RCPU-PG      SPAN-PG     |
|-------------------------------------------------------------------------------------------------------|
Total Instant Usage (cells)          0            0            0            0            0            0
Remaining Instant Usage (cells)  14256        34636         2836         2836            0         2836
Peak/Max Cells Used                623          623          N/A          N/A          N/A          N/A
Switch Cells Count              14256        34636         2836         2836            0         2836

Total Instant Usage (desc)           0            0            0            0            0            0
Remaining Instant Usage (desc)   53492            0         2836         2836            0         2836
Switch Desc Count                53492            0         2836         2836            0         2836

<SNIP>

Instance 2                                          Slice 2,3
============
|-------------------------------------------------------------------------------------------------------|
|                                Output Pool-Group Buffer Utilization (cells/desc)                      |
|                          Drop-PG      No-drop      CPU--PG      LCPU-PG      RCPU-PG      SPAN-PG     |
|-------------------------------------------------------------------------------------------------------|
Total Instant Usage (cells)          0            0            0            0            0            0
Remaining Instant Usage (cells)  17092        34636         2836            0            0         2836
Peak/Max Cells Used                623          623          N/A          N/A          N/A          N/A
Switch Cells Count              17092        34636         2836            0            0         2836

Total Instant Usage (desc)           0            0            0            0            0            0
Remaining Instant Usage (desc)   56328            0         2836            0            0         2836
Switch Desc Count                56328            0         2836            0            0         2836
```



LS12800H2R
40MB/ Per Slice Pair
(80MB total on-die buffer)

Figure 28. The no-drop mode one-die packet buffer allocation for H2R

```
9300-H2# show hardware internal buffer info pkt-stats

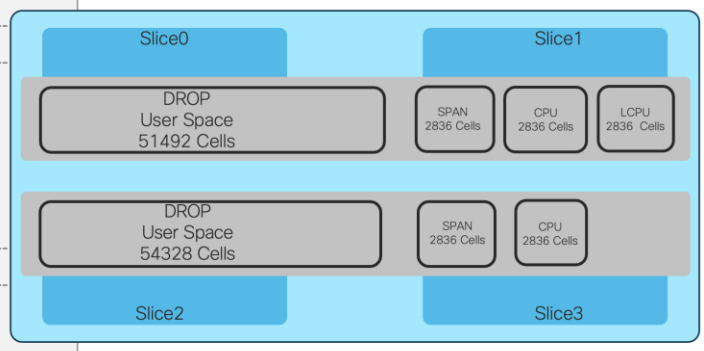
Instance 0                                          Slice 0,1
============
|-------------------------------------------------------------------------------------------------------|
|                                Output Pool-Group Buffer Utilization (cells/desc)                      |
|                          Drop-PG      No-drop      CPU--PG      LCPU-PG      RCPU-PG      SPAN-PG     |
|-------------------------------------------------------------------------------------------------------|
Total Instant Usage (cells)          0            0            0            0            0            0
Remaining Instant Usage (cells)  51492            0         2836         2836            0         2836
Peak/Max Cells Used                319          319          N/A          N/A          N/A          N/A
Switch Cells Count              51492            0         2836         2836            0         2836

Total Instant Usage (desc)           0            0            0            0            0            0
Remaining Instant Usage (desc)   53492            0         2836         2836            0         2836
Switch Desc Count                53492            0         2836         2836            0         2836

<SNIP>

Instance 2                                          Slice 2,3
============
|-------------------------------------------------------------------------------------------------------|
|                                Output Pool-Group Buffer Utilization (cells/desc)                      |
|                          Drop-PG      No-drop      CPU--PG      LCPU-PG      RCPU-PG      SPAN-PG     |
|-------------------------------------------------------------------------------------------------------|
Total Instant Usage (cells)          0            0            0            0            0            0
Remaining Instant Usage (cells)  54328            0         2836            0            0         2836
Peak/Max Cells Used                319          319          N/A          N/A          N/A          N/A
Switch Cells Count              54328            0         2836            0            0         2836


Total Instant Usage (desc)           0            0            0            0            0            0
Remaining Instant Usage (desc)   56328            0         2836            0            0         2836
Switch Desc Count                56328            0         2836            0            0         2836
```



LS12800H2
40MB/ Per Slice Pair
(80MB total on-die buffer)

Figure 29. The default mode one-die packet buffer allocation for H2

```
9300-H2# show hardware internal buffer info pkt-stats

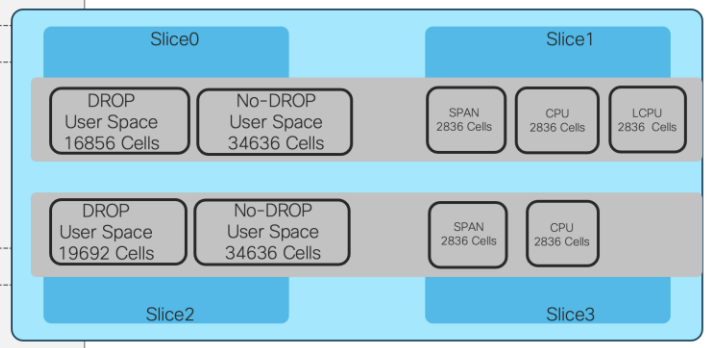
Instance 0                                          Slice 0,1
============
|-------------------------------------------------------------------------------------------------------|
|                                Output Pool-Group Buffer Utilization (cells/desc)                      |
|                          Drop-PG      No-drop      CPU--PG      LCPU-PG      RCPU-PG      SPAN-PG     |
|-------------------------------------------------------------------------------------------------------|
Total Instant Usage (cells)          0            0            0            0            0            0
Remaining Instant Usage (cells)  16856        34636         2836         2836            0         2836
Peak/Max Cells Used                319          319          N/A          N/A          N/A          N/A
Switch Cells Count              16856        34636         2836         2836            0         2836

Total Instant Usage (desc)           0            0            0            0            0            0
Remaining Instant Usage (desc)   53492            0         2836         2836            0         2836
Switch Desc Count                53492            0         2836         2836            0         2836

<SNIP>

Instance 2                                          Slice 2,3
============
|-------------------------------------------------------------------------------------------------------|
|                                Output Pool-Group Buffer Utilization (cells/desc)                      |
|                          Drop-PG      No-drop      CPU--PG      LCPU-PG      RCPU-PG      SPAN-PG     |
|-------------------------------------------------------------------------------------------------------|
Total Instant Usage (cells)          0            0            0            0            0            0
Remaining Instant Usage (cells)  19692        34636         2836            0            0         2836
Peak/Max Cells Used                319          319          N/A          N/A          N/A          N/A
Switch Cells Count              19692        34636         2836            0            0         2836


Total Instant Usage (desc)           0            0            0            0            0            0
Remaining Instant Usage (desc)   56328            0         2836            0            0         2836
Switch Desc Count                56328            0         2836            0            0         2836
```



LS12800H2
40MB/ Per Slice Pair
(80MB total on-die buffer)

© 2022 Cisco and/or its affiliates. All rights reserved.

Figure 30. The no-drop mode one-die packet buffer allocation for H2

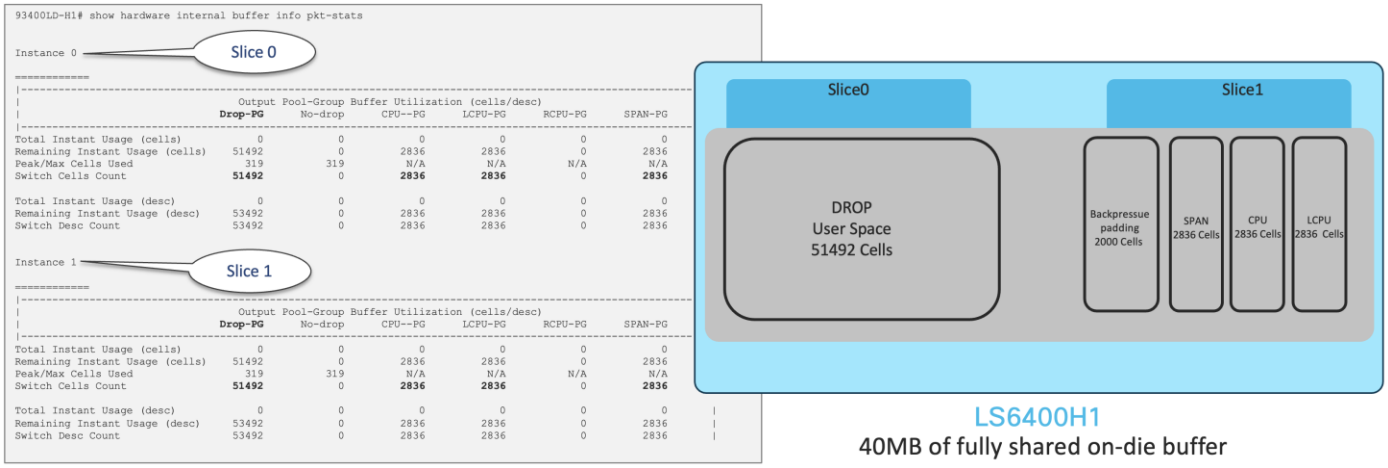


Figure 31. The default mode one-die packet buffer allocation for H1

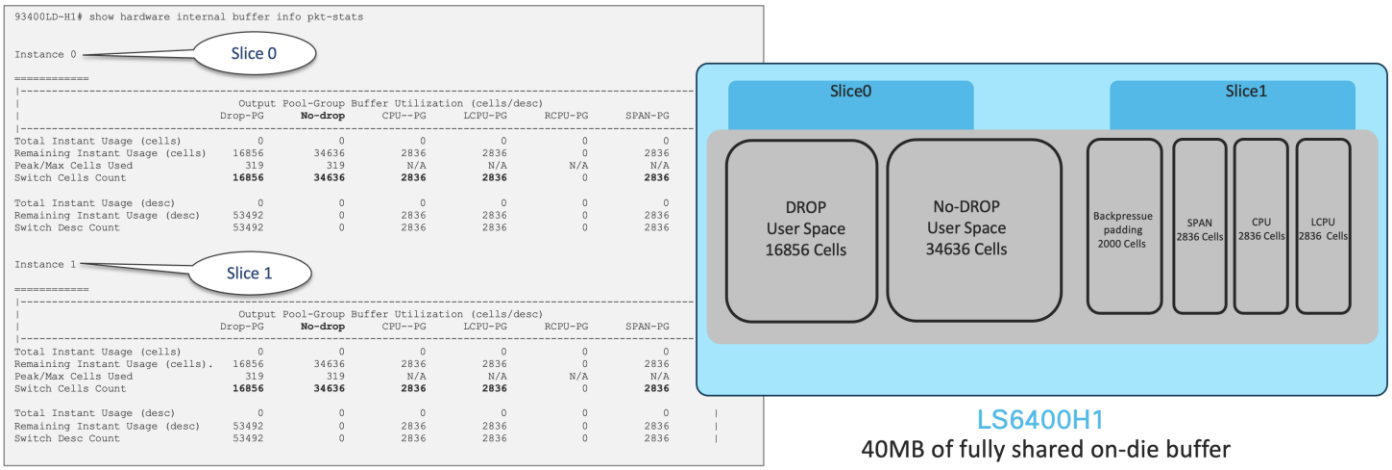


Figure 32. The no-drop mode one-die packet buffer allocation for H1.

CloudScale H2R/H2/H1 ASIC Intelligent Buffering

The on-die packet buffer, also referred to as smart buffers, of the CloudScale H2R/H2/H1 ASICs are composed of three primary components:

1. Dynamic Buffer Protection (DBP)
2. Approximate Fair Drop (AFD)
3. Dynamic Packet Prioritization (DPP)

Dynamic Buffer Protection (DBP)

DBP is activated by default and manages buffer allocation for congested queues within the shared-memory architecture. It is the method through which the ASIC regulates buffer admission, ensuring equitable access to the available shared memory buffer across all ports and queues. There is a dynamically adjusted maximum threshold - if this threshold is exceeded, buffer admission is denied. If it is below the threshold,

© 2022 Cisco and/or its affiliates. All rights reserved.

buffer admission is granted. This threshold is computed by multiplying the free memory by a configurable, per-queue Alpha (α) value (or weight). Alpha determines how assertively DBP preserves free buffer pages during congestion events. Figure 33 displays the buffer allocation for a specific port with varying alpha values, while Figure 34 provides an illustrative configuration example.

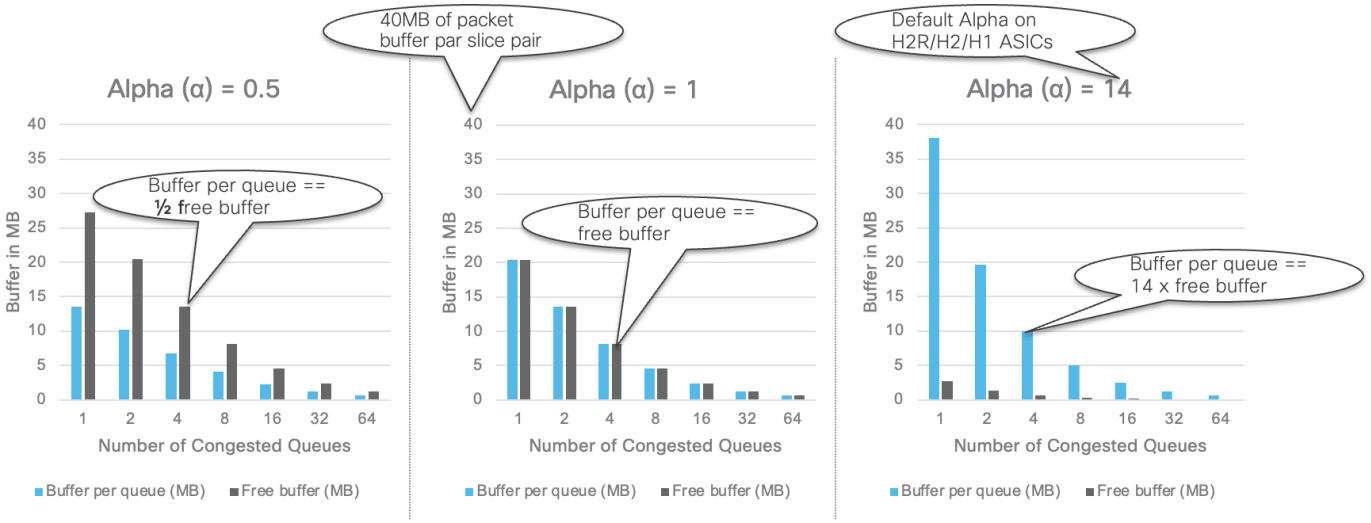


Figure 33. Cisco ClouldScale H2R/H2/H1 different alpha values

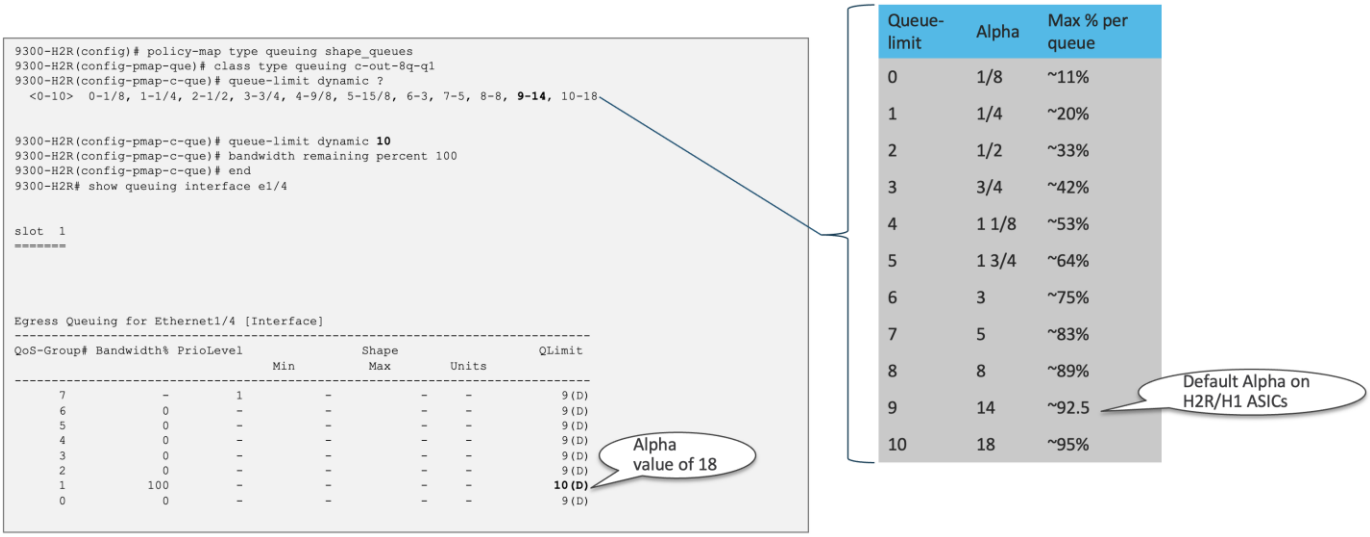


Figure 34. Cisco CloudScale H2R/H2/H1 Queue limit (Alpha) configurartion

Approximate Fair Drop (AFD)

AFD is an active queue-management scheme whose fundamental goal is to provide fair bandwidth allocation among flows that share a common egress queue. The fairness has two aspects. First, AFD distinguishes long-lived elephant flows from short-lived mice flows and exempts mice flows from the dropping algorithm so that mice flows will get their fair share of bandwidth without being starved by bandwidth-hungry elephant flows. Second, AFD tracks elephant flows and subjects them to the AFD algorithm in the egress queue to grant them their fair share of bandwidth. Figure 35 provides an illustrative configuration example.

© 2022 Cisco and/or its affiliates. All rights reserved.

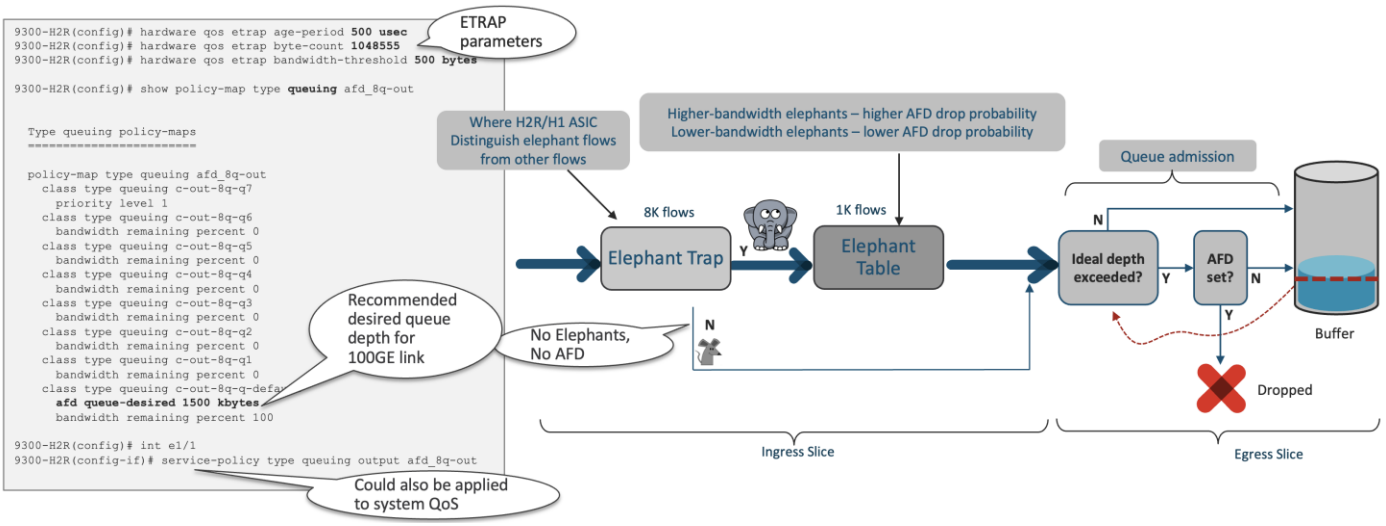


Figure 35. Cisco CloudScale H2R/H2/H1 Approximate Fair Drop (AFD)

Dynamic Packet Prioritization (DPP)

Dynamic packet prioritization, or DPP, provides the capability to manage mice flows and elephant flows in the same traffic class separately for queuing. It effectively isolates them in two separate queues despite the fact that they belong to the same traffic class. This separation is not possible with the traditional simple queue management technologies because they lack flow awareness within a class. Figure 36 provides an illustrative configuration example.

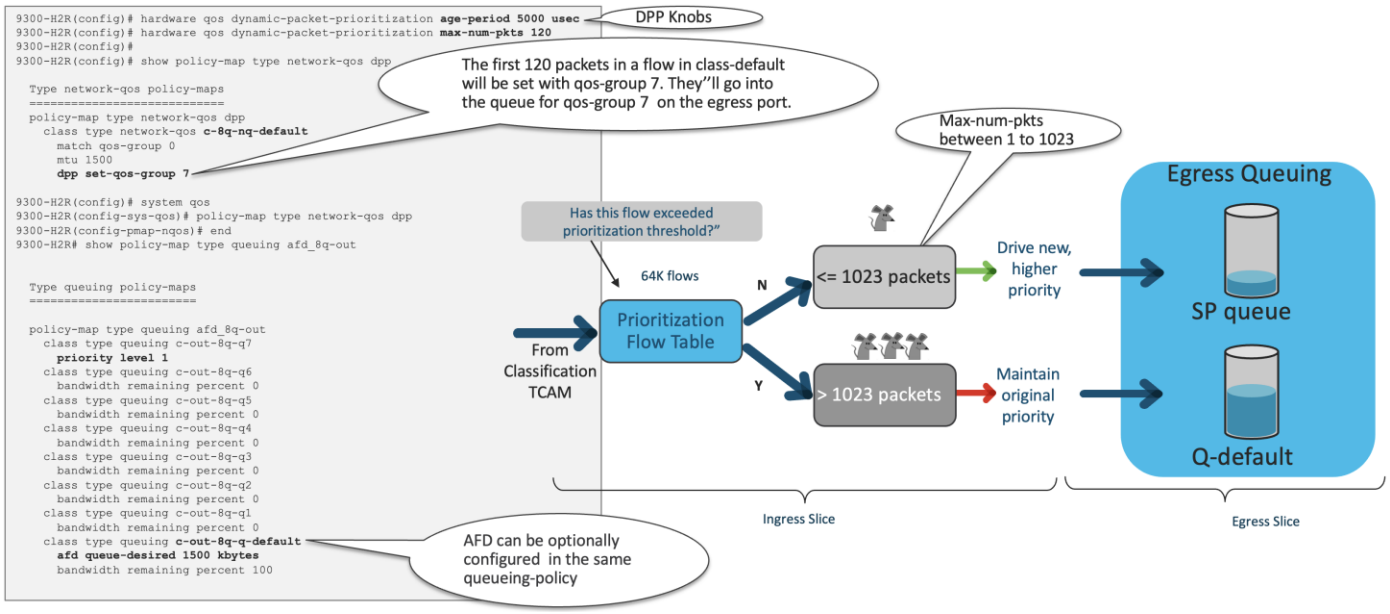


Figure 36. Cisco CloudScale H2R/H2/H1 dynamic packet prioritization (DPP)

CloudScale H2R HBM Packet Buffer

Along with Smart buffers, the Cisco LS 12800 H2R ASIC also features an 8GB Deep buffer. The Deep Buffer, also known as High Bandwidth Memory (HBM), is crucial for handling large traffic surges in the switch by providing an extra 8GB buffer on top of the existing buffers in the switch. The Deep Buffer

© 2022 Cisco and/or its affiliates. All rights reserved.

system is distributed across four slices of the LS12800 H2R, with each slice receiving up to 2GB of buffer space. If this limit is exceeded, noticeable drops will occur.

The system includes the Deep Buffer feature by default, enabling any queue to tap into the buffers during periods of congestion. Notably, the Deep Buffer is situated outside the traditional switch pipeline, thereby offering independent accounting and buffering functions. Packets in the Deep Buffer are stored as 'novels', with one novel equating to 256KB. The LS12800 H2R handles the conversion each time a packet is added or removed from the High Bandwidth Memory (HBM). To constrain the queues, Deep Buffer implements static thresholding, with any packets that surpass this threshold being tail dropped.

The on-chip packet buffer is solely designed to support dynamic alpha configuration, while the High Bandwidth Memory (HBM) is limited to static thresholding. As a result, the alpha programming is performed only in the on-chip system, while the off-chip system continues to validate the queue based on maximum cell/novels thresholds. This architecture ensures that each queue functions optimally based on its particular supporting configuration.

It's important to mention that the Quality of Service (QoS) Command Line Interface (CLI) does not provide any specific sub-options for deep buffer. Rather, all existing QoS configurations are reused, and the internal code base guides the deep-buffer programming using a predefined logic. The architecture of the H2R ASIC HBM is depicted in Figure 37. Each slice is assigned with 6,881,280 Cells, or 16384 novels. Here, one cell equals 624 Bytes while one novel amounts to 256K Bytes.

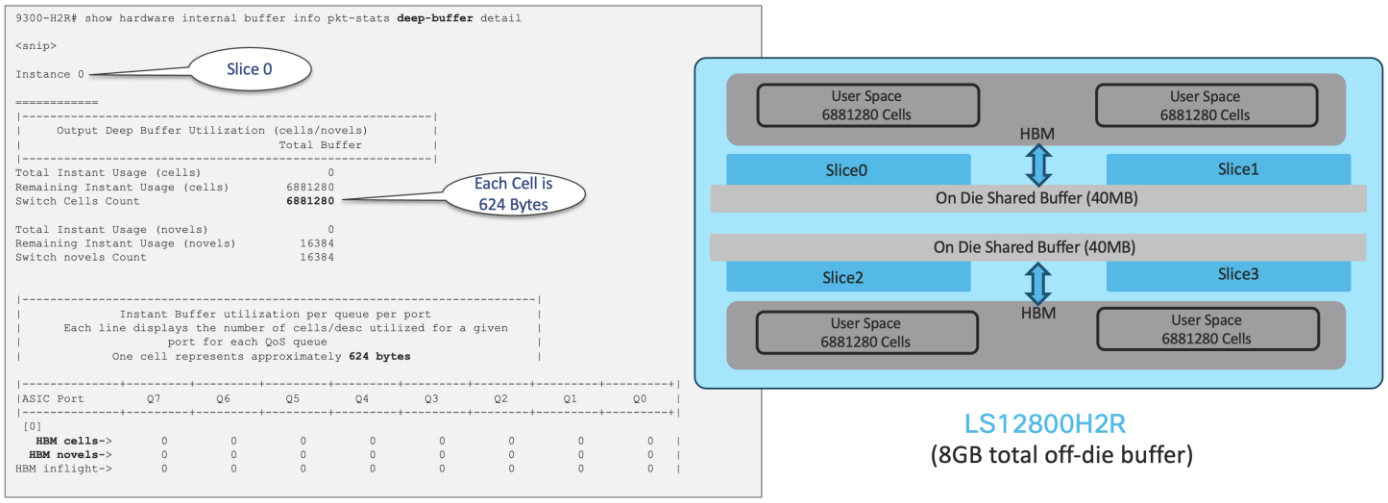


Figure 37. Cisco CouldScale H2R ASIC deepbuffer architecture

Cisco CloudScale H2R, H2 and H1 ASIC AI/ML

Artificial Intelligence and Machine Learning (AI/ML) applications are being used increasingly in today's data centers, and the Cisco CloudScale H2R, H2 and H1 ASIC have the hardware and software capabilities to provide the right latency, congestion-management mechanisms, and telemetry to meet the requirements of those applications.

Remote direct memory access (RDMA) is a well-known technology used for high performance computing (HPC) and storage networking environments. The advantages of RDMA are the high throughput and low latency transfer of information between compute nodes at the memory-to-memory level, without burdening the CPU. This transfer function is offloaded to the network adapter hardware to bypass the operating system software network stack. This technology also has the advantage of reduced power requirements. Figure 38 illustrates the direct GPU memory communication using RDMA..

© 2022 Cisco and/or its affiliates. All rights reserved.

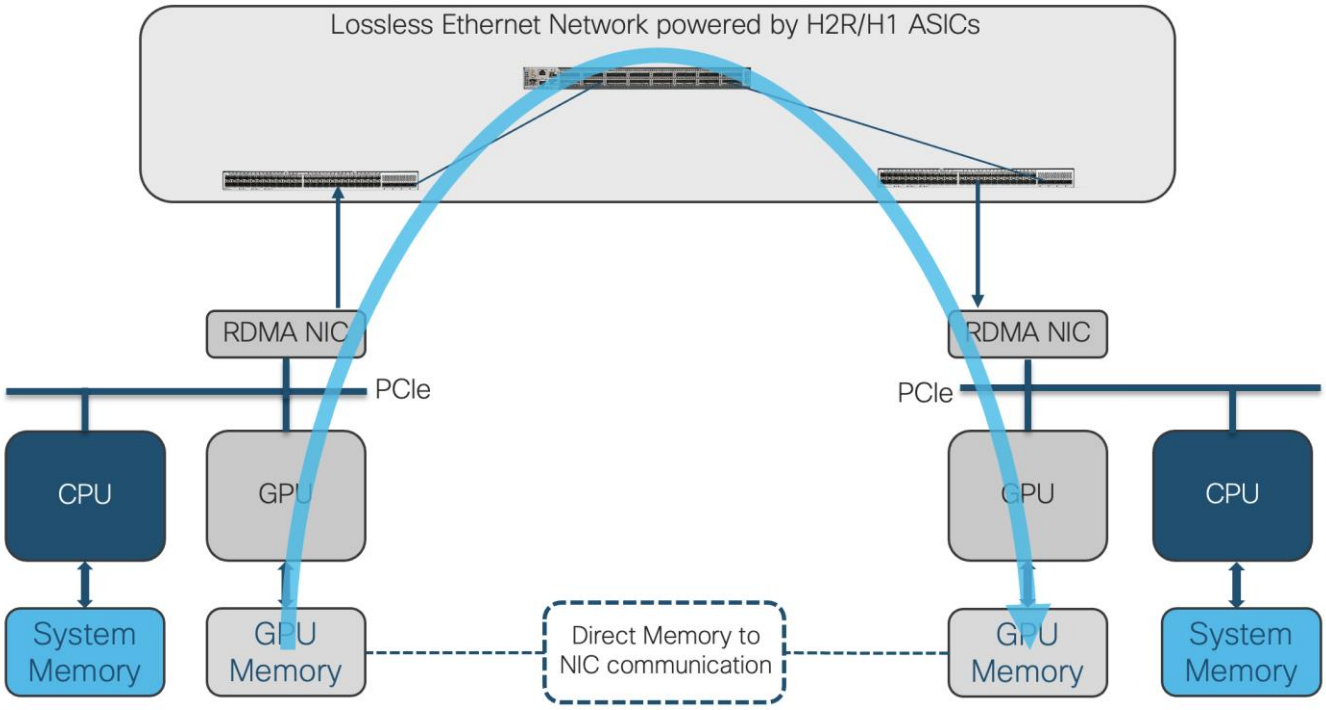


Figure 38. Direct GPU memory communication using RDMA

In AI/ML clusters, RDMA is used to communicate memory-to-memory between GPUs over the network. This implementation is called GPUDirect RDMA. RoCEv2 is an excellent transport for GPUDirect RDMA.

For RoCEv2 transport, the network must provide high throughput and low latency while avoiding traffic drops in situations where congestion occurs. The Cisco CloudScale H2R, H2 and H1 are built for data center networks and provide the required low latency. With up to 12.8 Tbps of bandwidth per ASIC, these switches provide the very high throughput required to satisfy AI/ML clusters running on top of RoCEv2 transport. The CloudScale H2R, H2 and H1 also provide support and visibility in a lossless network through software and hardware telemetry in both ECN and PFC.

Explicit Congestion Notification (ECN)

In situations where congestion information needs to be propagated end-to-end, ECN can be used for congestion management. ECN is marked in the network node where congestion is experienced inside the IP header type of service (TOS) field in the 2 least significant bits. When a receiver gets a packet with the ECN congestion experience bits set to 0b11, it generates and sends a congestion notification packet (CNP) back to the sender. When the sender receives the congestion notification, it slows down the flow that matches the notification. This end-to-end process is built in the data path, and as such is an efficient way to manage congestion.

Priority Flow Control (PFC)

Priority Flow Control was introduced in Layer 2 networks as the primary mechanism to enable lossless Ethernet. Flow control was driven by the class of service (COS) value in the Layer 2 frame, and congestion is signaled and managed using pause frames, and a pause mechanism. However, building scalable Layer 2

© 2022 Cisco and/or its affiliates. All rights reserved.

networks can be a challenging task for network administrators. Because of this, network designs have mostly evolved into Layer 3 routed fabrics.

As RoCEv2 can be routed, PFC was adjusted to work with differentiated services code point (DSCP) priorities to signal congestion between routed hops in the network. DSCP is a mechanism used for classifying network traffic on IP networks. It uses the 6-bit differentiated services field in the IP header for packet classification purposes. Using Layer 3 marking enables traffic to maintain classification semantics across routers. Since PFC frames use link local addressing, the network devices can receive and perform pause signaling for both routed and switched traffic. PFC is transmitted per-hop, from the place of congestion to the source of the traffic. This step-by-step behavior could take time to be propagated to the source. PFC is used as the primary tool to manage congestion for RoCEv2 transport.

Cisco CloudScale H2R, H2 and H1 ASICs support both PFC congestion management and ECN marking with either weighted random early detection (WRED) or approximate fair drop (AFD) to indicate congestion in the network node. Figure 39 illustrates the RoCEv2 configuration example with the Cisco N9300-H1.

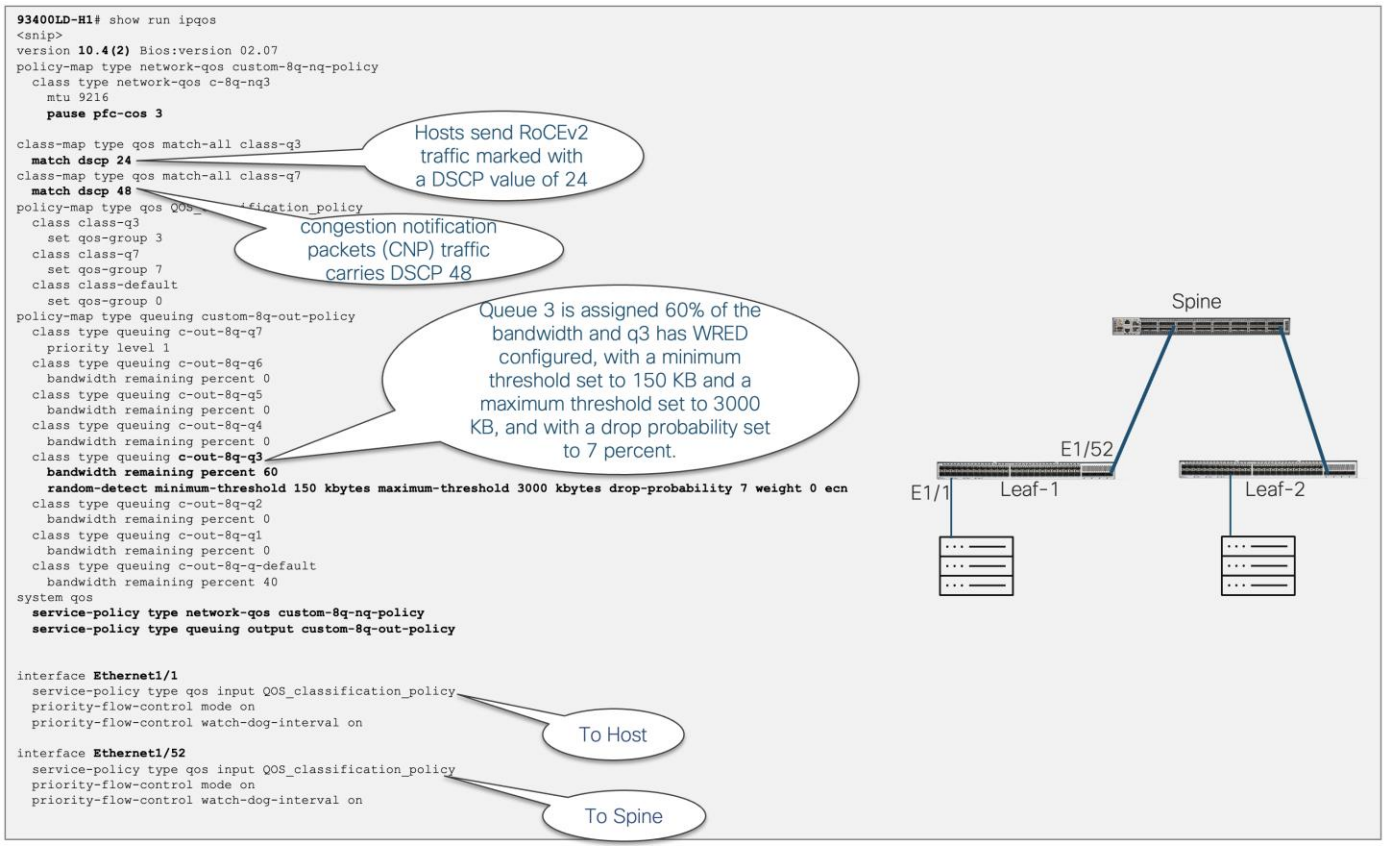


Figure 39. RoCEv2 configuration example with cisco N9300-H1

Cisco N9300-H Switches

Cisco N9332D-H2R

The Cisco N9332D-H2R Switch introduces the first switch with deep-buffer capability to the Cisco N9000 fixed-switch portfolio. The switch is built with backward-compatible 400G optical interface Quad Small Form-Factor Pluggable-Double Density (QSFP-DD) ports to migrate existing data center fabrics from 40-Gbps and 100-Gbps speeds to 400 Gbps; it also offers various lower port speeds and densities, including

© 2022 Cisco and/or its affiliates. All rights reserved.

10, 25, 40, 50, 100, and 200 Gbps, using breakouts. The Cisco N9332D-H2R Switch is based on Cisco® CloudScale technology equipped to support next-generation cloud architecture.

The Cisco N9332D-H2R is a 1-Rack-Unit (1RU) 32-port 400 Gigabit Ethernet switch that supports 12.8 Tbps of bandwidth. The switch provides 80MB of on-die packet buffer and 8GB of high bandwidth memory (or HBM) with MACsec capability on all 32 ports.

The Cisco N9332D-H2R is equipped with support for the Precision Time Protocol (PTP) providing Class C timing accuracy suitable for telecom and media profiles. Leveraging PTP and Synchronous Ethernet, it ensures highly accurate frequency and time distribution. The front and rear views of the N9332D-H2R are depicted in Figures 40 and 41, respectively, while Figure 42 provides a visual representation of the Switch on Chip (SOC) Architecture.



Figure 40. Cisco N9332D-H2R Switch, frontside



Figure 41. Cisco N9332D-H2R Switch, rearside

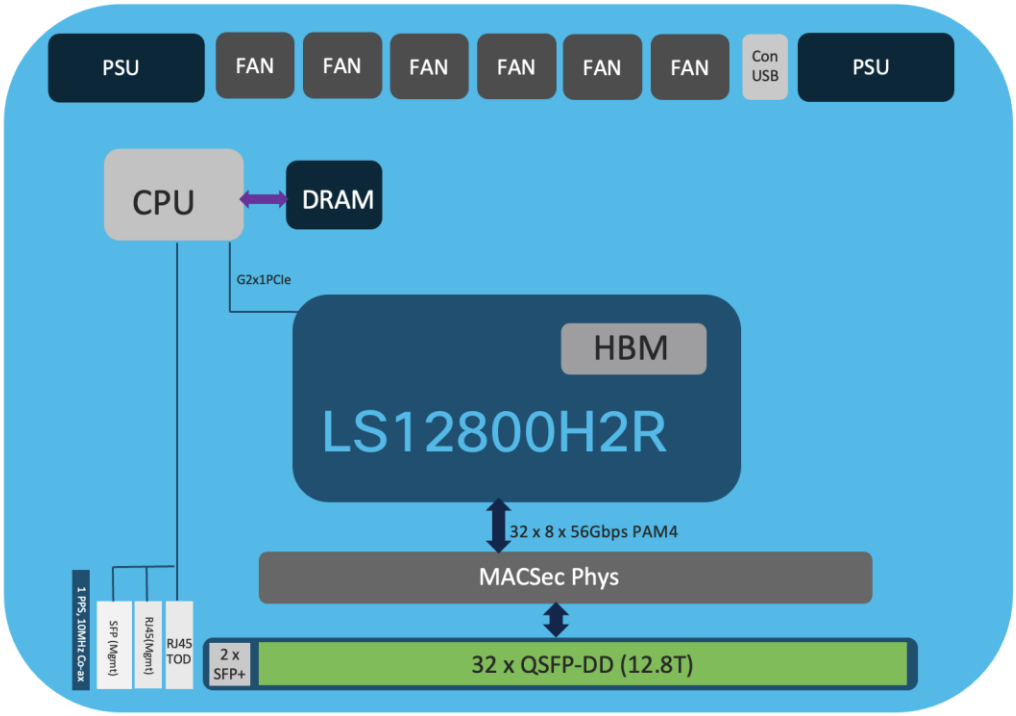


Figure 42. Cisco N9332D-H2R Switch on Chip (SOC) Architecture

Leveraging the multi-slice architecture of the LS12800H2R ASIC, it's critical to comprehend how the front panel ports correspond to the CloudScale ASIC slices. The mapping of the front panel ports of the Cisco N9332D-H2R is depicted in Figure 43.

© 2022 Cisco and/or its affiliates. All rights reserved.

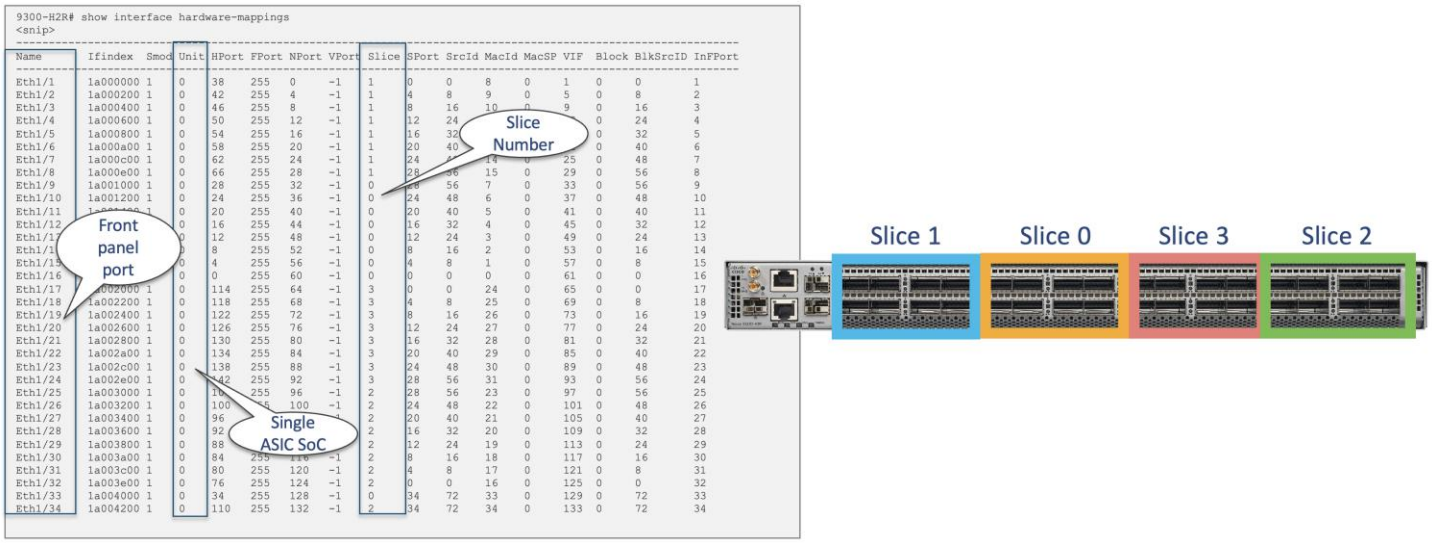


Figure 43. Cisco N9332D-H2R Switch Port-Mapping

Cisco N93400LD-H1

The Cisco N93400LD-H1 Switch introduces Precision Time Protocol (PTP) with Class C timing accuracy to the Cisco N9000 fixed-switch portfolio. Through PTP and Synchronous Ethernet, it delivers frequency and time distribution with high accuracy The switch is built with backward-compatible 400G optical interface Quad Small Form-Factor Pluggable-Double Density (QSFP-DD) ports for high density aggregation at 400 Gbps speed; it also offers various lower port speeds and densities, including 10, 25, and 50 Gbps. The Cisco N93400LD-H1 Switch is based on Cisco® CloudScale technology equipped to support next-generation cloud architecture.

The Cisco N93400LD-H1 is a 1-Rack-Unit (1RU) switch featuring 48 ports of 50 Gigabit Ethernet and 4 ports of 400 Gigabit Ethernet, supporting a bandwidth of 4 Tbps. The switch includes a fully shared on-die packet buffer of 40MB, with MACsec capability available on all ports. Figures 44 and 45 display the front and rear views of the N93400LD-H1, and Figure 46 shows the Switch on Chip (SOC) Architecture.



Figure 44. Cisco N93400LD-H1 Switch, frontside



Figure 45. Cisco N93400LD-H1 Switch, rearside

© 2022 Cisco and/or its affiliates. All rights reserved.

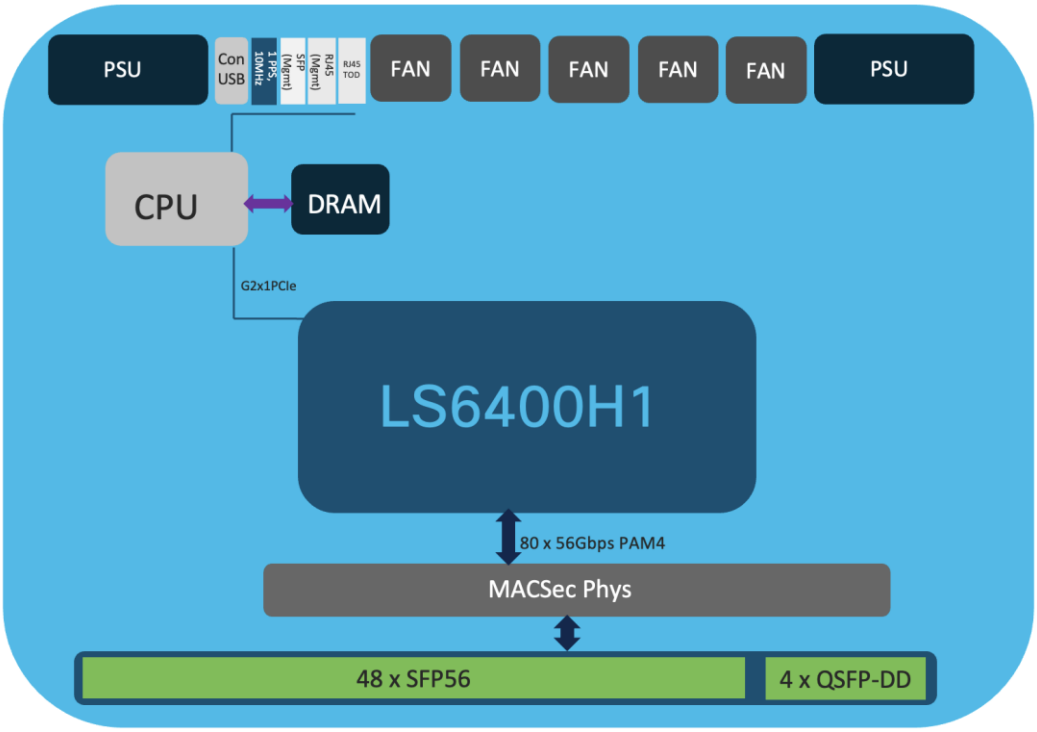


Figure 46. Cisco N93400LD-H1 Switch on Chip (SOC) Architecture

Figure 47 provides a visual representation of the front panel port arrangement of the Cisco N93400LD-H1.

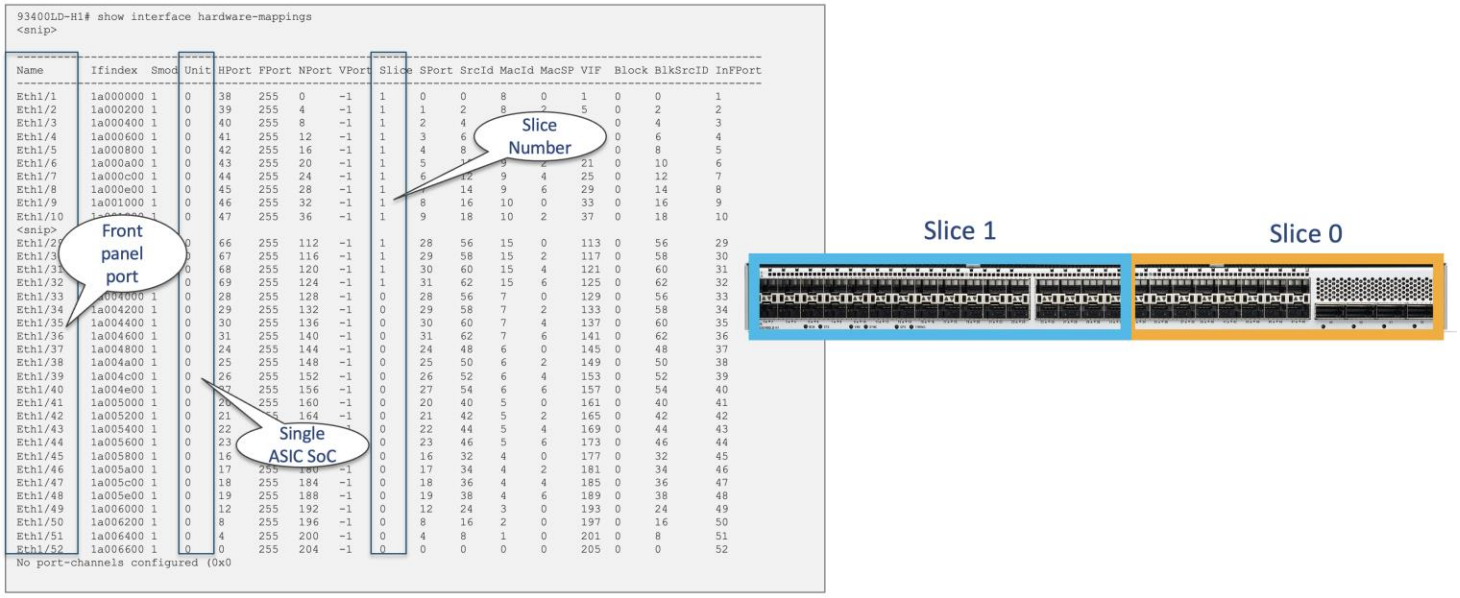


Figure 47. Cisco N93400LD-H1 Switch Port-Mapping

Cisco N9364C-H1

The Cisco N9364C-H1 switch introduces high-density 100G aggregation for the data center fabric. It also offers various lower port speeds and densities, including 10, 25, and 40 Gbps. The Cisco N9364C-H1

© 2022 Cisco and/or its affiliates. All rights reserved.

switch is based on Cisco® Cloud Scale technology equipped to support next-generation cloud architecture.

The Cisco N9364C-H1 is a 2-Rack-Unit (2RU) 64-port 100 Gigabit Ethernet switch that supports 12.8 Tbps of bandwidth and 2.5 bpps across all ports. The switch provides 40MB of on-die packet buffer with MACsec capability on 16 ports. Figures 48 and 49 display the front and rear views of the N9364C-H1, and Figure 50 shows the Switch on Chip (SOC) Architecture.



Figure 48. Cisco N9364C-H1 Switch, frontside



Figure 49. Cisco N9364C-H1 Switch, rearside

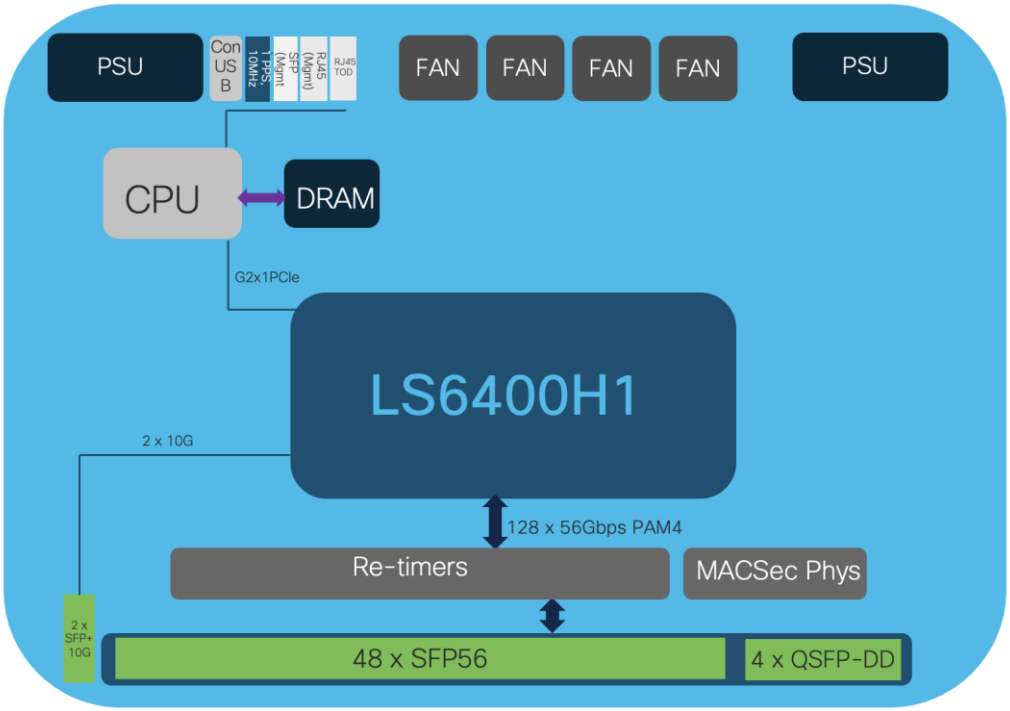


Figure 50. Cisco N9364C-H1 Switch on Chip (SOC) Architecture

© 2022 Cisco and/or its affiliates. All rights reserved.

Figure 51 provides a visual representation of the front panel port arrangement of the Cisco N9364C-H1.

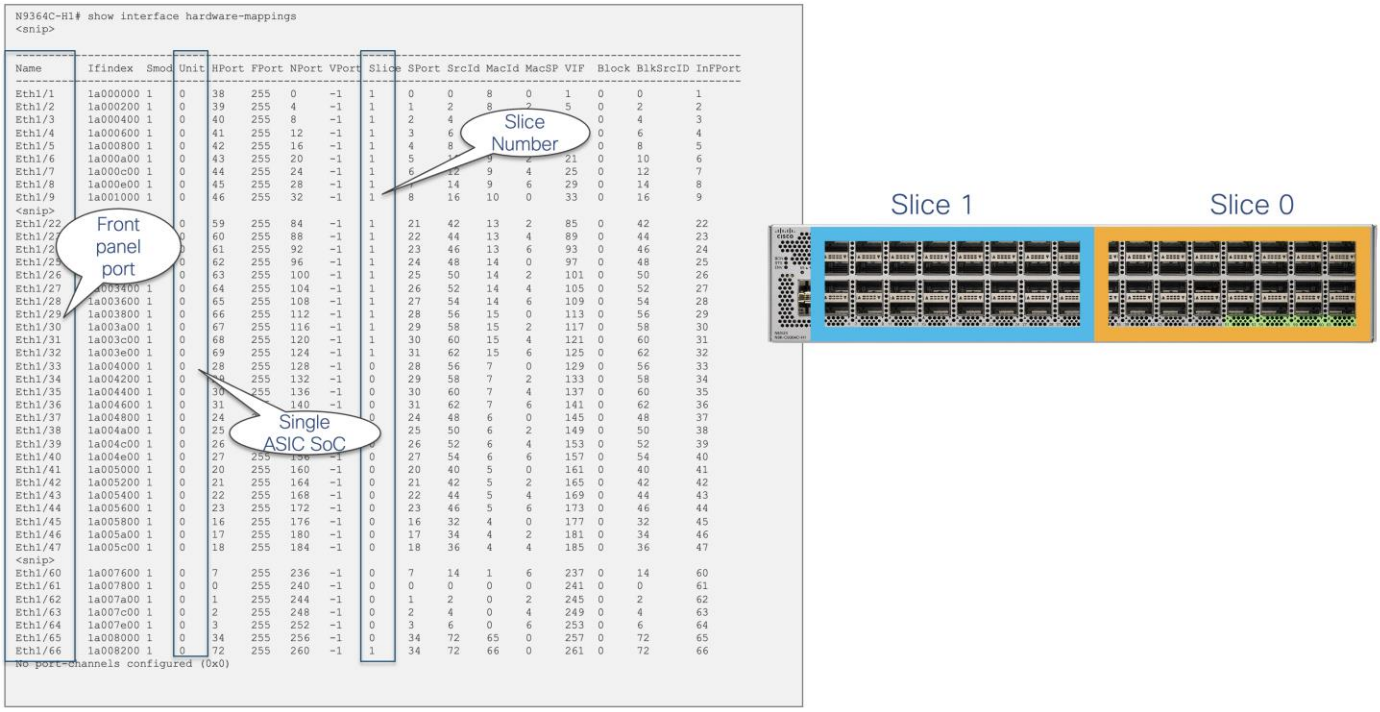


Figure 51. Cisco N9364C-H1 Switch Port-Mapping

Cisco N9348C8D-H2

Data center networks are evolving in more than one direction. Spine and core links may move to 400G or 800G, while many server-facing connections remain at 10G, 25G, 50G, or 100G. That leaves a clear design challenge at the leaf: add more 100G capacity while making efficient use of rack space.

AI, high-performance computing (HPC), and hyperscale cloud environments are increasing demand for high-throughput connections. For many deployments, 100G remains a cost-efficient workhorse, and the move to faster spine links can increase the need for dense 100G leaf capacity.

The Cisco N9348C8D-H2 addresses this need by combining dense, flexible 100G access with higher-speed uplinks in a compact 1RU design.

The Cisco N9348C8D-H2 is a fixed 1RU switch that brings together 48 100G SFP-DD ports and eight 400G QSFP-DD uplinks. SFP-DD ports support 10G, 25G, 50G, and 100G; QSFP-DD uplinks support 40G, 100G, 200G, and 400G. Together, these options let organizations connect varied endpoints at the leaf and link to faster spine layers as the fabric grows. Figures 52 and 53 display the front and rear views of the N9348C8D-H2, and Figure 54 shows the Switch on Chip (SOC) Architecture.

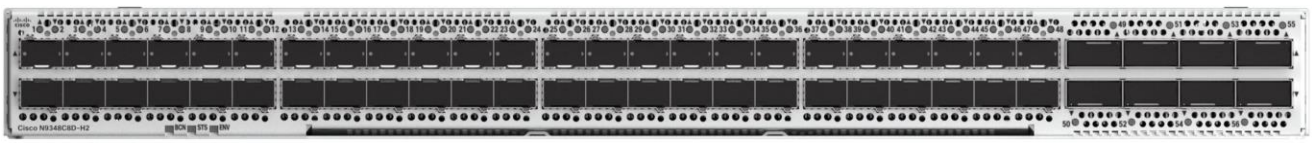


Figure 52. Cisco N9348C8D-H2 Switch, frontside

© 2022 Cisco and/or its affiliates. All rights reserved.



Figure 53. Cisco N9348C8D-H2 Switch, rearside

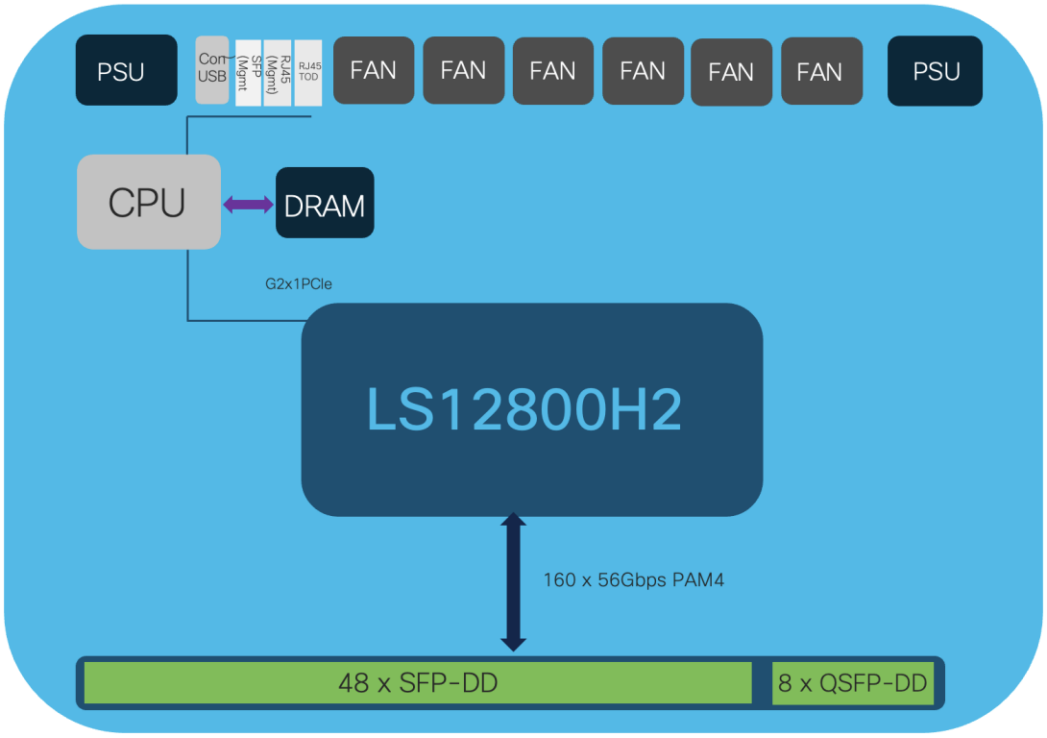


Figure 54. Cisco N9348C8D-H2 Switch on Chip (SOC) Architecture

Figure 55 provides a visual representation of the front panel port arrangement of the Cisco N9348C8D-H2.

© 2022 Cisco and/or its affiliates. All rights reserved.

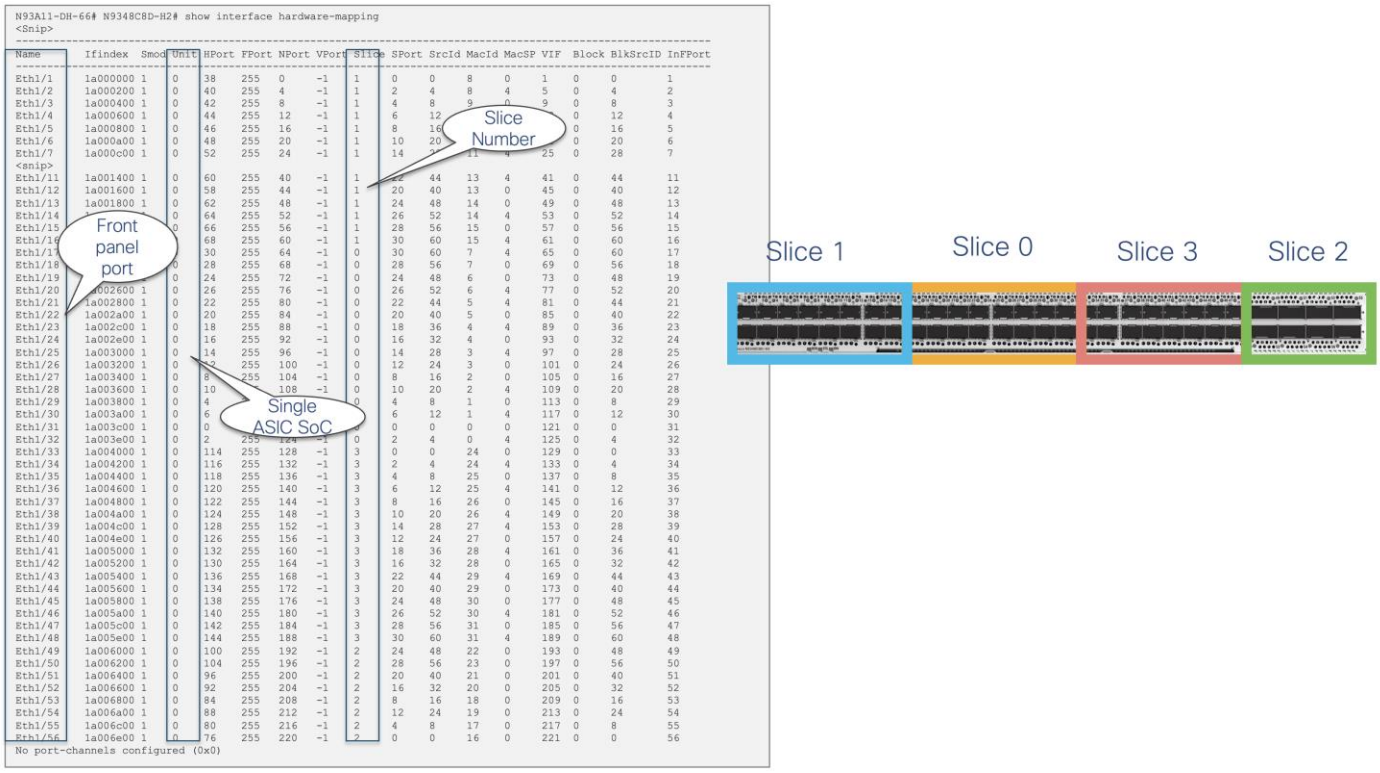


Figure 55. Cisco N9348C8D-H2 Switch Port-Mapping

Cisco N9300-H MACsec Hardware Encryption

MACsec offers encryption at the line rate on a link-level, hop-by-hop basis. Both Cisco N9332D-H2R and N93400LD-H1 models support MACsec across all ports. Cisco N9364C-H1 provides MACsec on the last 16 ports. They also support IEEE 802.1AE 128-bit and 256-bit AES encryption with MKA Key Exchange. An example of a MACsec configuration with the Cisco N9300-H is demonstrated in Figure 56.

© 2022 Cisco and/or its affiliates. All rights reserved.

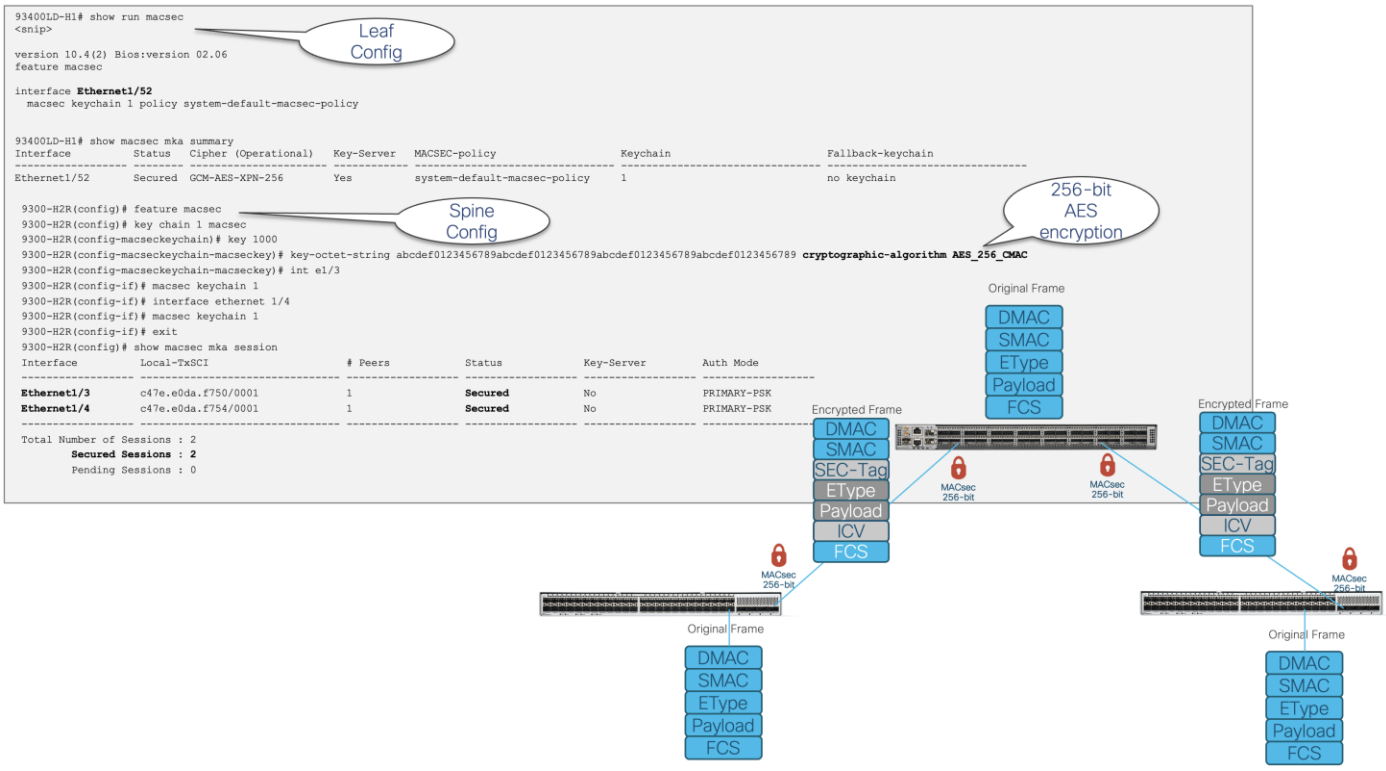
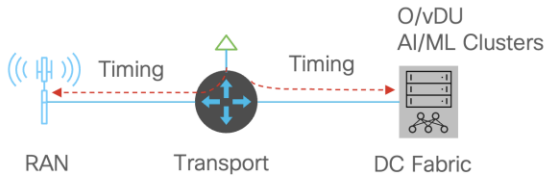


Figure 56. Example of MACsec encryption on Cisco N9300-H

Cisco N9300-H SyncE and PTP

The advent of 5G networks creates unprecedented opportunities and challenges for telcos and enterprises to deliver secure services and new user experiences. Each new generation of mobile networks has driven the need for increased precision and accuracy in synchronization standards and solutions. 5G technology is needed to support Ultra-Reliable Low-Latency Communications (URLLC) services, and new revenue generating services like autonomous vehicle connectivity and eHealth services are simply not possible without highly accurate and reliable time synchronization. Even services like massive Industrial Internet of Things (IIoT) and other industrial automation services require reliable time synchronization. Figure 57 showcases the time synchronization within DC Fabric.



- On Fronthaul, Radios should be within 400ns budget from UTC for efficient performance
- O/vDU may have control to decide if RU is out of sync and needs reboot / reset to align with time
- O/vDU may or may not support SyncE, but it is must for RU

5G RAN functions in Data Center fabric require support for accurate timing with telecom timing profiles

- To accurately sync RAN functions within DC Fabric
- Ensures redundancy and resiliency
- Provides better performance and holdover for 5G operations
- Accurate time-stamping and recovery to improve efficiency for 5G AI/ML clusters and enhances Security network operations

Figure 57. The Time Synchronization within DC Fabric

© 2022 Cisco and/or its affiliates. All rights reserved.

The Open Radio Access Network Architecture (O-RAN) or virtual RAN (vRAN) present a novel approach for service providers to meet the requirements of 5G networks. The Cisco N9300-H holds a significant role in the O-RAN/vRAN transition, as the currently deployed solution for far edge Data Centers (DC) necessitates the following features:

- Telecom profile G.8275.1(Full timing)
- Telecom profile G.8275.2 (Partial timing support) whenever G.8275.1 is not possible
- Synchronous Ethernet (SyncE)
- Line rate 10G/25G/50G, 100G/400GE interfaces

The PTP telecom profile G.8275.1 necessitates both PTP and SyncE across the entire path from T-GM to T-TSC. Furthermore, each network node requires boundary clock support, with PTP frames distributed using L2 multicast. This makes the Cisco N9300-H an ideal candidate due to its Class-C accuracy. Figure 58 depicts the deployment of O/vRAN aided by Far Edge DC, while Figure 59 showcases the deployment of the Far Edge DC using the Cisco N9300-H that operates PTP 8275-1 telecom profile.

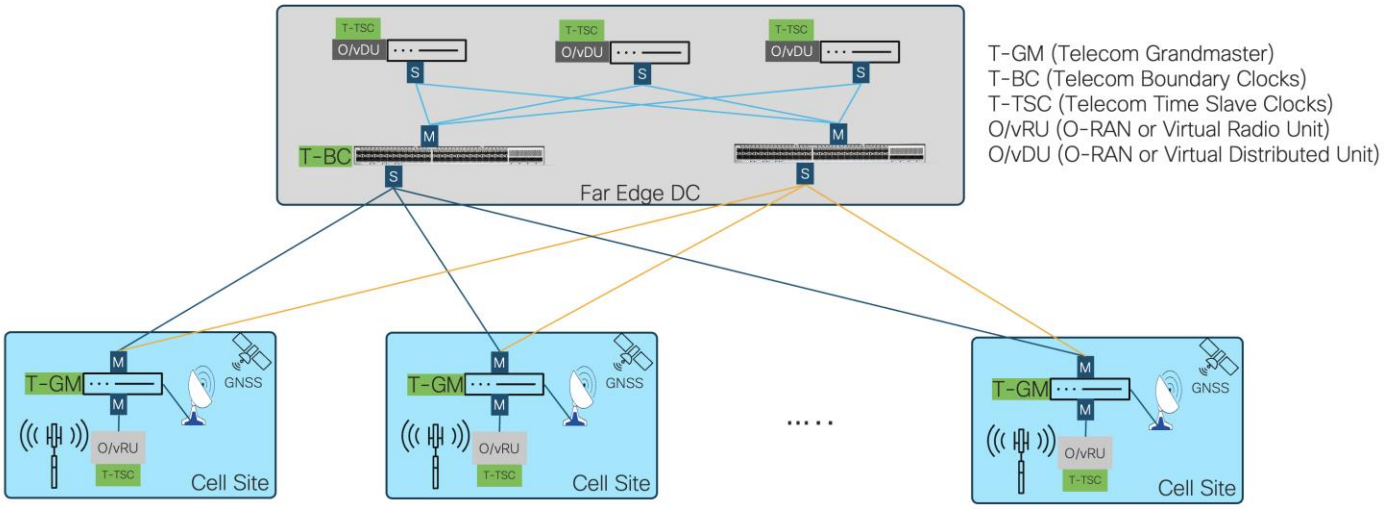


Figure 58. O/vRAN Architecture with Far Edge DC deployment

© 2022 Cisco and/or its affiliates. All rights reserved.

```
93400LD-H1(config)# feature ptp
93400LD-H1(config)# ptp profile ?
  8275-1   PTP telecom profile matching 8275.1
  8275-2   PTP telecom profile matching 8275.2
  default  Default profile


93400LD-H1(config-ptp-profile)# ptp profile 8275-1

Please enable SyncE ('feature frequency synchronization') on at least one port
to operate correctly in hybrid mode (default for 8275.1 profile)

NOTE: All PTP configurations have been reset to default values corresponding
to the configured profile except on interfaces with profile override. Some
parameters (transport, cost, anno timeout, anno intv, dreq intv, sync intv)
remain unchanged on such interfaces
```

Both 8275-1 and 8275-2 are supported with 9300-H2R/H1

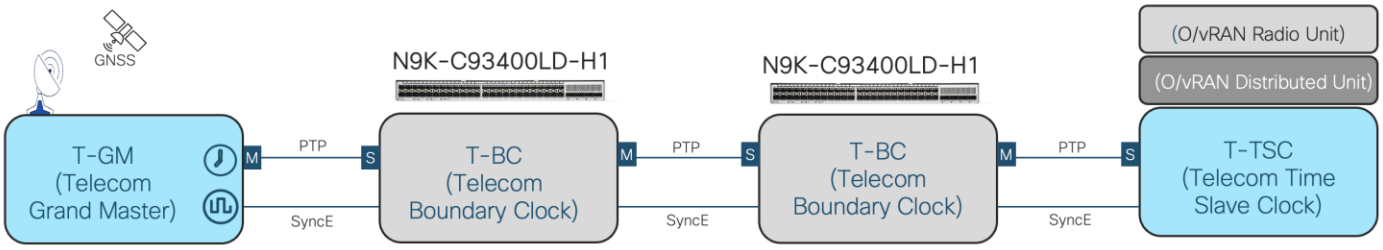


Figure 57. Cisco N9300-H deployments in Far Edge DC with 8275-1 PTP telecom profile

In addition to 8275-1 telecom profile, Cisco N9300-H also supports the G.8275.2 teleome profile which is based on the partial timing support from the network, hence nodes are not required to be directly connected. This profile is aimed at operation over existing networks, where Boundary clocks are placed at strategic locations to reduce noise and it uses PTP over IPv4 and IPv6 in unicast mode instead of L2 multicast frames used in G.8275-1. Figure 58 showcases the deployment of the Far Edge DC using the Cisco N9300-H that operates PTP 8275-2 telecom profile.

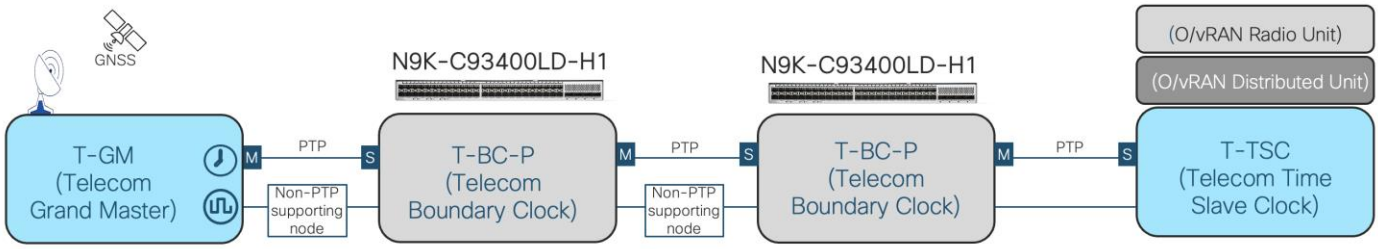


Figure 59. Cisco N9300-H deployments in Far Edge DC with 8275-2 PTP telecom profile

Conclusion

The Cisco N9300-H platform switches, incorporating Cisco's cutting-edge CloudScale H2R, H2 and H1 ASICs, establish a benchmark for contemporary data center switching solutions. These switches offer a variety of features including adaptable forwarding tiles, enhanced ACL TCAM regions, smart buffer management, deep buffer, advanced hardware telemetry and visibility, MACsec, and PTP with Class C timing accuracy. These features equip customers with the necessary tools to optimize performance and efficiency in their data center operations.

Additional Resources

- Cisco Nexus 9500 Cloud Scale Line Cards and Fabric Modules White Paper
- Flexible Forwarding Table on Nexus 9000

© 2022 Cisco and/or its affiliates. All rights reserved.

- Cisco Data Center Networking Blueprint for AI/ML Applications
- Cisco Validated Design for Data Center Networking Blueprint for AI/ML Applications
- Classification TCAM with Cisco CloudScale ASICs for Nexus 9000 Series Switches White Paper

© 2022 Cisco and/or its affiliates. All rights reserved.