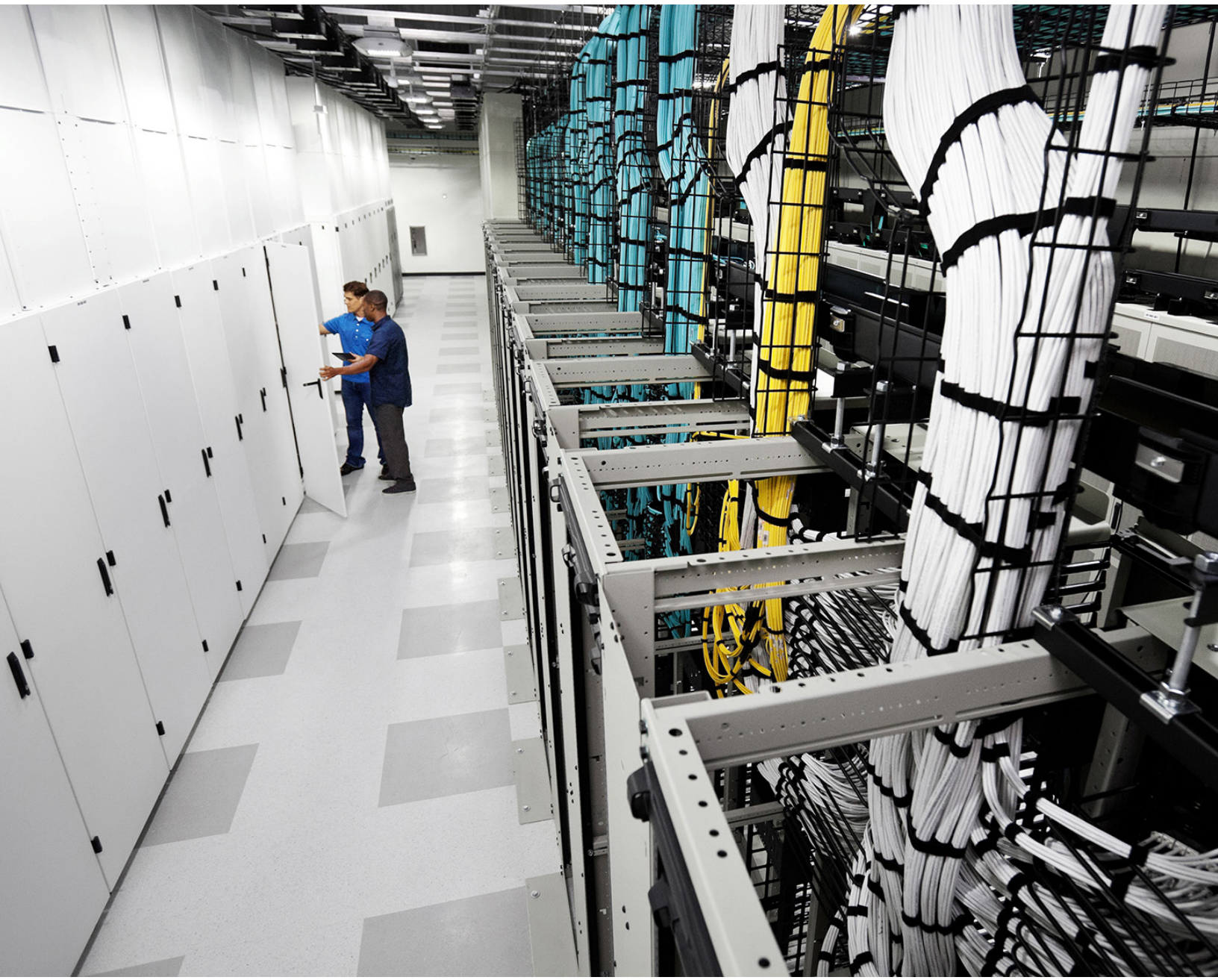
Cisco Nexus 9200 Platform Switches Architecture



Contents

Introduction

Starting in 2016, the data center switching industry will begin the shift to new capacity and capabilities with the introduction of 25, 50, and 100 Gigabit Ethernet connectivity. This new Ethernet connectivity supplements the previous 10 and 40 Gigabit Ethernet standards, with the similar cost points and power efficiency, and represents a roughly 250 percent increase in capacity.

Cisco is releasing a number of new data center switching products to help our customers build higher-performance and more cost-effective data center networks to accommodate greater application workloads and different types of connectivity. To support both existing and next-generation data center network infrastructure, the new Cisco® switches support both existing and new standard Ethernet speeds, including 1, 10, and 40 Gbps and 25, 50, and 100 Gbps.

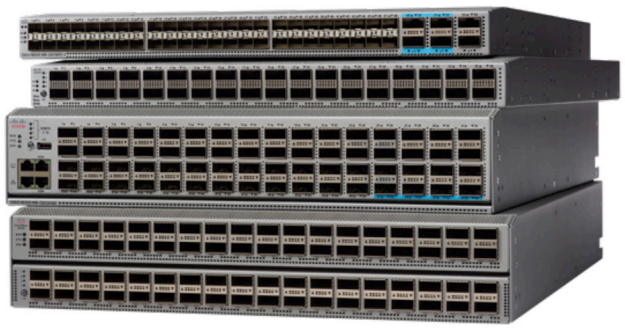
This document discusses the hardware architecture of the new switch platforms in the Cisco Nexus® 9000 Series Switches product family: in particular, the Cisco Nexus 9200 platform switches. The Cisco Nexus 9200 platform is a new extension of the Cisco Nexus 9000 Series. The platform uses a compact fixed-switch form factor and operates in the NX-OS mode, based on Cisco NX-OS Software. Built with next-generation Cisco application-specific circuits (ASICs), the Cisco Nexus 9200 platform brings high-performance and high-density 25, 50, and 100 Gigabit Ethernet connectivity to data center networks, providing industry-leading performance, power efficiency, and capabilities in compact fixed-form-factor switches.

Cisco Nexus 9200 Platform Architecture

The Cisco Nexus 9200 platform consists of fixed-configuration switches built with the new Cisco Application Centric Infrastructure (Cisco ACI™) Spine Engine 2 (ASE2) ASIC or ASE3 ASIC.

The initial introduction of the Cisco Nexus 9200 platform offers four models: Cisco Nexus 9272Q, 92304QC, 9236C, and 92160YC-X Switches (Figure 1). The 9272Q, 92304QC, and 9236C models are built using the ASE2 ASIC. The 92160YC-X uses the ASE3 ASIC. Table 1 summarizes the Cisco Nexus 9200 platform models.

Figure 1. Cisco Nexus 9200 Platform Switch Models



Following the naming conventions for the Cisco Nexus 9000 Series, the characters in the Cisco Nexus 9200 platform product names indicate supported port speeds or additional hardware capabilities:

- Q: Native 40-Gbps front-panel ports
- Y: Native 25-Gbps front-panel ports
- C: Native 100-Gbps front-panel ports
- X (after the hyphen): Cisco NetFlow and data analytics capabilities

Table 1. Cisco Nexus 9200 Platform Switch Models

Model	Description	Cisco ASIC
Cisco Nexus 9272Q Switch	72 x 40-Gbps Enhanced Quad Small Form-Factor Pluggable (QSFP+) ports	ASE2
Cisco Nexus 92304QC Switch	56 x 40-Gbps QSFP+ ports and 8 x 100-Gbps ports	ASE2
Cisco Nexus 9236C Switch	36x 100-Gbps AWFP28 ports	ASE2
Cisco Nexus 92160YC-X	40 x 1/10/25-Gbps SFP and 6 x 40-Gbps QSFP or 4 x 100-Gbps QSFP28 ports	ASE3

The Cisco Nexus 9272Q, 92304QC, and 9236C Switches all use the same CPU and memory configuration, which is a four-core Intel Ivy Bridge Gladden processor, and 16 GB of system memory. The Cisco Nexus 92160YC-X is built with a dual-core Intel Ivy Bridge Gladden processor and 16 GB of system memory. All the Cisco Nexus 9200 platform switches are built with redundant fans and support reversible airflow. Table 2 summarizes the hardware characteristics of the Cisco Nexus 9200 platform.

Table 2. Cisco Nexus 9200 Platform Hardware Characteristics

	Cisco Nexus 9272Q	Cisco Nexus 92304QC	Cisco Nexus 9236C	Cisco Nexus 92160YC-X
CPU	4 cores	4 cores	4 cores	2 cores
System memory	16 GB	16 GB	16 GB	16 GB
Solid-state disk (SSD) drive	64 GB	64 GB	64 GB	64 GB
Shared System buffer	30 MB	30 MB	30 MB	20MB
Management ports	1 RJ-45 and 1 SFP+	3 RJ-45 ports	1 RJ-45 and 1 SFP+	1 RJ-45 and 1 SFP+
USB ports	1	1	1	1
RS-232 serial ports	1	1	1	1
Power supplies (up to 2)	930W DC, 1200W AC, or 1200W HVAC/DC	1200W AC, 930W DC, or 1200W HVAC/HVDC	930W DC, 1200W AC, or 1200W HVAC/DC	650W AC, 930W DC, or 1200W AC/HVDC
Input voltage (AC)	100 to 240V	100 to 240V	100 to 240V	100 to 240V
Input voltage (HVAC)	200 to 277V	200 to 277V	200 to 277V	200 to 277V
Input voltage (DC)	–48 to –60V	–48 to –60V	–48 to –60V	–48 to –60V
Input voltage (HVDC)	–240 to –380V	–240 to –380V	–240 to –380V	–240 to –380V

	Cisco Nexus 9272Q	Cisco Nexus 92304QC	Cisco Nexus 9236C	Cisco Nexus 92160YC-X
Frequency (AC)	50 to 60 Hz	50 to 60 Hz	50 to 60 Hz	50 to 60 Hz
Fans	2	2	4	4
Airflow	Port-side intake and exhaust	Port-side intake and exhaust	Port-side intake and exhaust	Port-side intake and exhaust
Physical (H x W x D)	3.5 x 17.4x 24.5 in. (8.9 x 44.2 x 62.3 cm)	3.5 x 17.5 x 22.5 in. (8.9 x 44.5 x 57.1 cm	1.72 x 17.3 x 22.5 in. (4.4 x 43.9 x 57.1 cm)	1.72 x 17.3 x 22.5 in. (4.4 x 43.9 x 57.1 cm)
RoHS compliance	Yes	Yes	Yes	Yes

Cisco Nexus 9272Q Switch Architecture

The Cisco Nexus 9272Q (Figure 2) is an ultra-high-density 2-rack-unit (2RU) switch. It supports 5.76 terabits per second (Tbps) of duplex bandwidth and a packet forwarding rate of more than 4.5 billion packets per second (bpps) across 72 x 40-Gbps QSFP+ front-panel ports.

As shown in Figure 2, of the 72 ports, the top 36 (ports 1 through 36) operate in 40-Gbps mode only. Of the bottom 36 ports (ports 37 through 71), 35 support breakout to 4 x 10-Gbps connectivity (port 72 does not support breakout), providing a total port density of 140 x 10 Gbps.

The Cisco Nexus 9272Q supports per-port dynamic breakout on ports 37 through 71. Any of these ports can individually break out to 4 x 10 Gbps without restriction to port-group assignments. They also can break out with the need to reload the switch to change the port speed.

Figure 2. Cisco Nexus 9272Q Switch

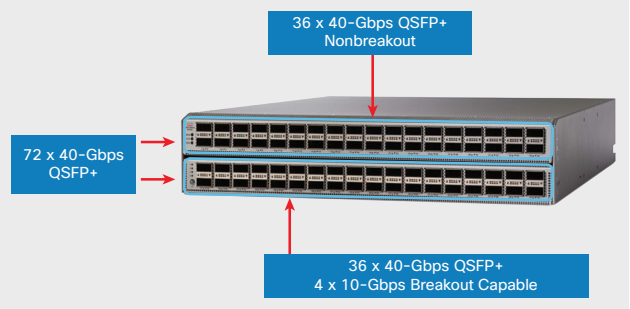


Figure 3 shows the Cisco Nexus 9272Q hardware architecture.

The Cisco Nexus 9272Q is equipped with a quad-core Gladden Ivy Bridge 1.8-GHz CPU and 16 GB of system memory.

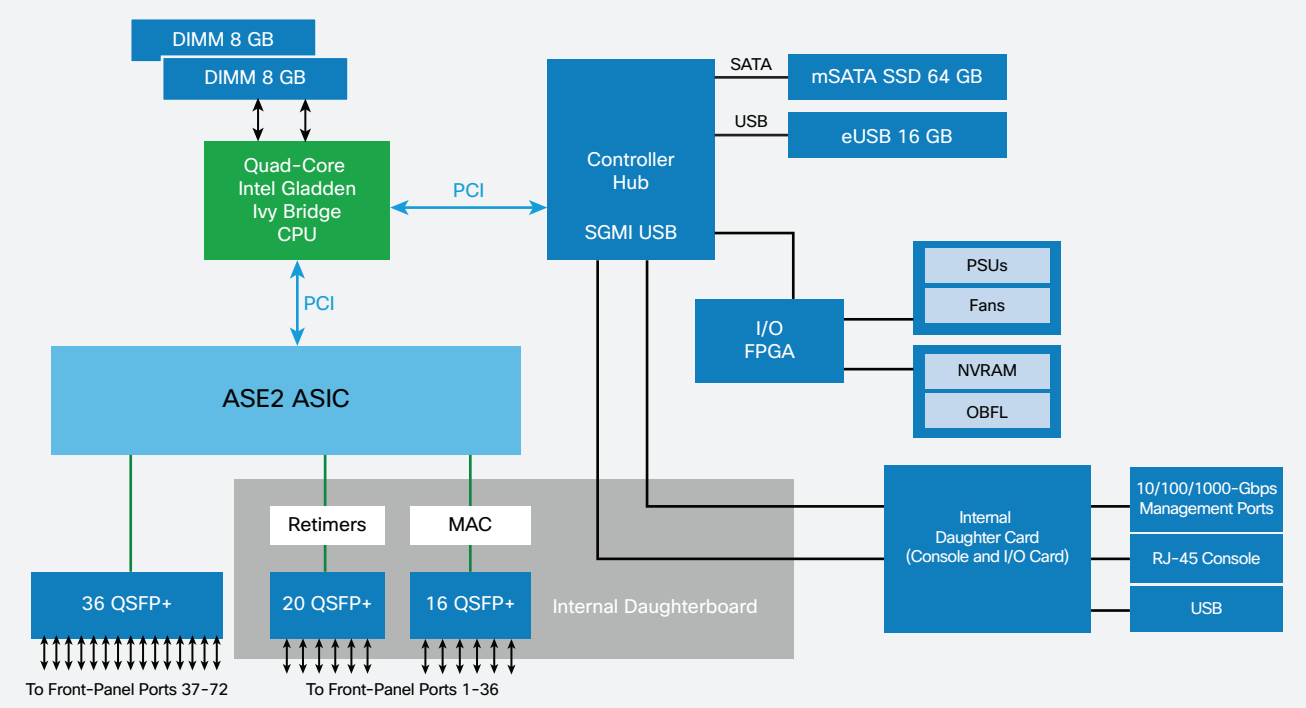
Two 8-GB DDR 3 DIMMs are populated using both memory channels of the CPU. This design provides a high-performance control plane that is critical for a data center switch, especially in a large-scale and dynamic network. It also provide the CPU power, memory, and storage capacity needed for the extensive programmability functions introduced in the NX-OS operating system for the Cisco Nexus 9200 platform to support the modern operating mode for data center networks.

The CPU is connected to the controller hub through PCI Express (PCIe) connections. The controller hub provides standard interfaces (SATA, USB, Ethernet, etc.) to the storage, power, fan, and management I/O components. The Cisco Nexus 9272C is equipped with a 64-GB SSD drive.

The console and I/O daughterboard includes an RG-45 serial console port connection and dual-media Ethernet management ports supporting either 10/100/1000BASE-T or 1-Gbps SFP for fiber connections. Only one of the two management ports can be active at any given time. The switch will automatically select the port with an active link status. If both links are connected, the copper interface will have priority. The console and I/O card includes a USB 2.0 port.

The data-plane forwarding components on the Cisco Nexus 9272Q include a single multiple-slice ASE2 ASIC. The ASE2 ASIC has direct connections to 36 of the 72 front-panel ports: ports 36 through 72. It supports 40-Gbps mode and 4 x 10-Gbps breakout mode on ports 36-71. Front-panel ports 1 through 36 are on an internal daughterboard that connects them to the ASE2 ASIC. The internal daughterboard provides retimers or additional MACs for ports 1 through 36. These ports do not support the 4 x 10-Gbps breakout mode.

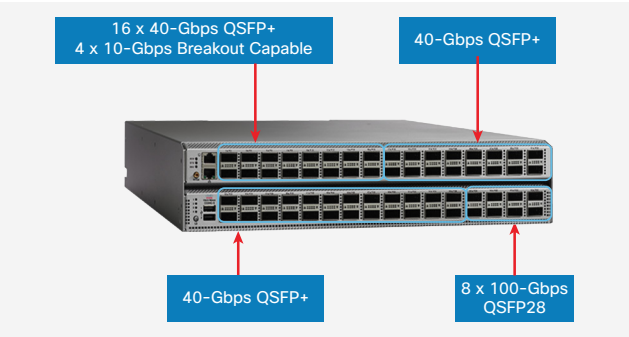
Figure 3. Cisco Nexus 9272Q Switch Hardware Architecture



Cisco Nexus 92304QC Switch Architecture

The Cisco Nexus 92304QC (Figure 4) is an ultra-high-density 2RU switch that supports 6.1 Tbps of duplex bandwidth and a packet forwarding rate of more than 9.5 bpps across 56 x 40-Gbps QSFP+ ports and 8 x 100-Gbps QSFP28 ports.

Figure 4 Cisco Nexus 92304QC Switch



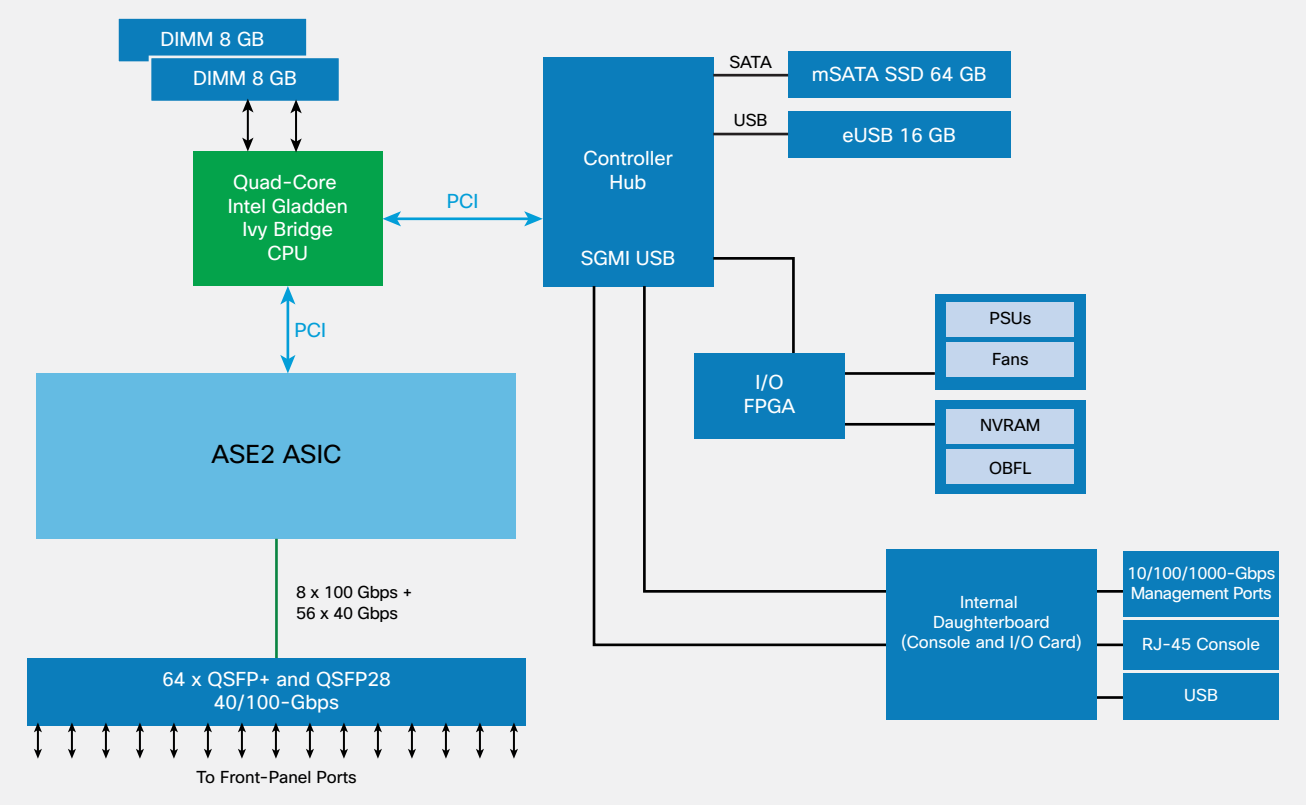
As shown in Figure 4, the front-panel ports on the Cisco Nexus 92304QC provide 56 x 40-Gbps QSFP+ ports (ports 1 through 56) and 8 x 100-Gbps QSFP28 ports (ports 57 through 64). The first 16 ports (ports 1 through 16) support breakout to 4 x 10-Gbps speed, providing a total of 64 x 10-Gbps ports on a Cisco Nexus 92304QC.

The Cisco Nexus 92304QC supports per-port dynamic breakout on the breakout-capable ports (ports 1 through 16). Any of these ports can break out to 4 x 10-Gbps port. No reload process is required when you convert 40-Gbps ports to 4 x 10-Gbps breakout ports, and there are no port-group assignment restrictions.

The last 8 ports are 100-Gbps ports using 100-Gbps QSFP28 optics.

Figure 5 shows the Cisco Nexus 92304QC hardware architecture. This switch has the same CPU, memory, and system architecture as the Cisco Nexus 9272Q. Its data plane is also built on a single ASE2 ASIC, which is directly connected to all 64 front-panel ports. No retimer or additional MACs are needed.

Figure 5. Cisco Nexus 92304QC Switch Hardware Architecture



Cisco Nexus 9236C Switch Architecture

The Cisco Nexus 9236C (Figure 6) is a 1RU switch that supports 7.2 Tbps of duplex bandwidth and a packet forwarding rate of more than 11.25 bpps across the 36 front-panel ports. Each of the 36 front-panel ports on a Cisco Nexus 9236C can be individually configured into 1 x 100-Gbps, 1 x 40-Gbps, 4 x 25-Gbps, or 4 x 10-Gbps ports. No switch reload is required to change the port speed configuration, and there are no port-group assignment restrictions.

The Cisco Nexus 9236C provides the industry's best 100-Gbps port density and performance in the compact 1RU form factor. It is best suited as a lean data center spine or aggregation switch. With the support of up to 144 x 1, 10, and 25-Gbps breakout ports, it can also provide ultra-dense 1, 10, and 25-Gbps host connectivity when deployed in the data center network access layer. In addition to the port density, the switch also benefits the network through its flexibility to support different port speeds, enabling a data center network design that connects a mixture of hosts with different port speeds.

Figure 6. Cisco Nexus 9236 Switch

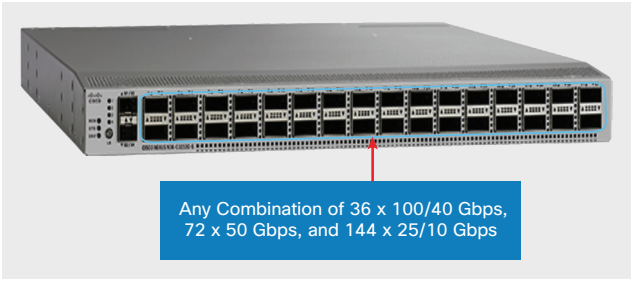
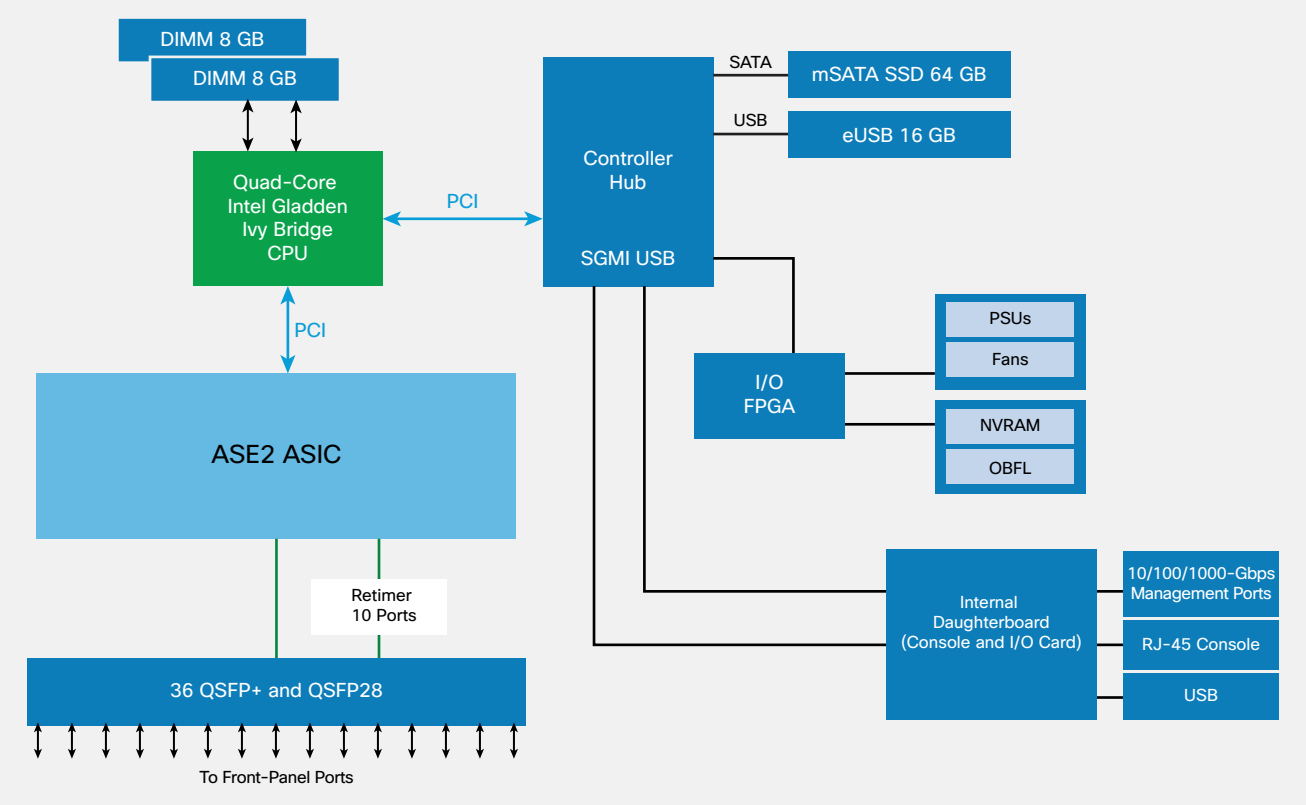


Figure 7 shows the Cisco Nexus 9236C hardware architecture. This switch has the same internal architecture as the Cisco Nexus 92304QC, except for the retimers for some of the front-panel ports. Its data-plane is also built on a single ASE2 ASIC, which connects to 36 front-panel ports either directly or through a retimer.

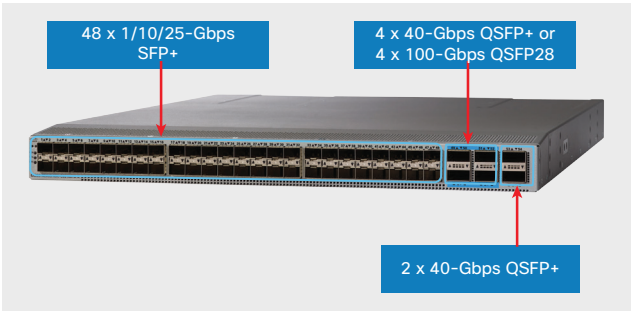
Figure 7. Cisco Nexus 9236C Switch Hardware Architecture



Cisco Nexus 92160YC-X Switch Architecture

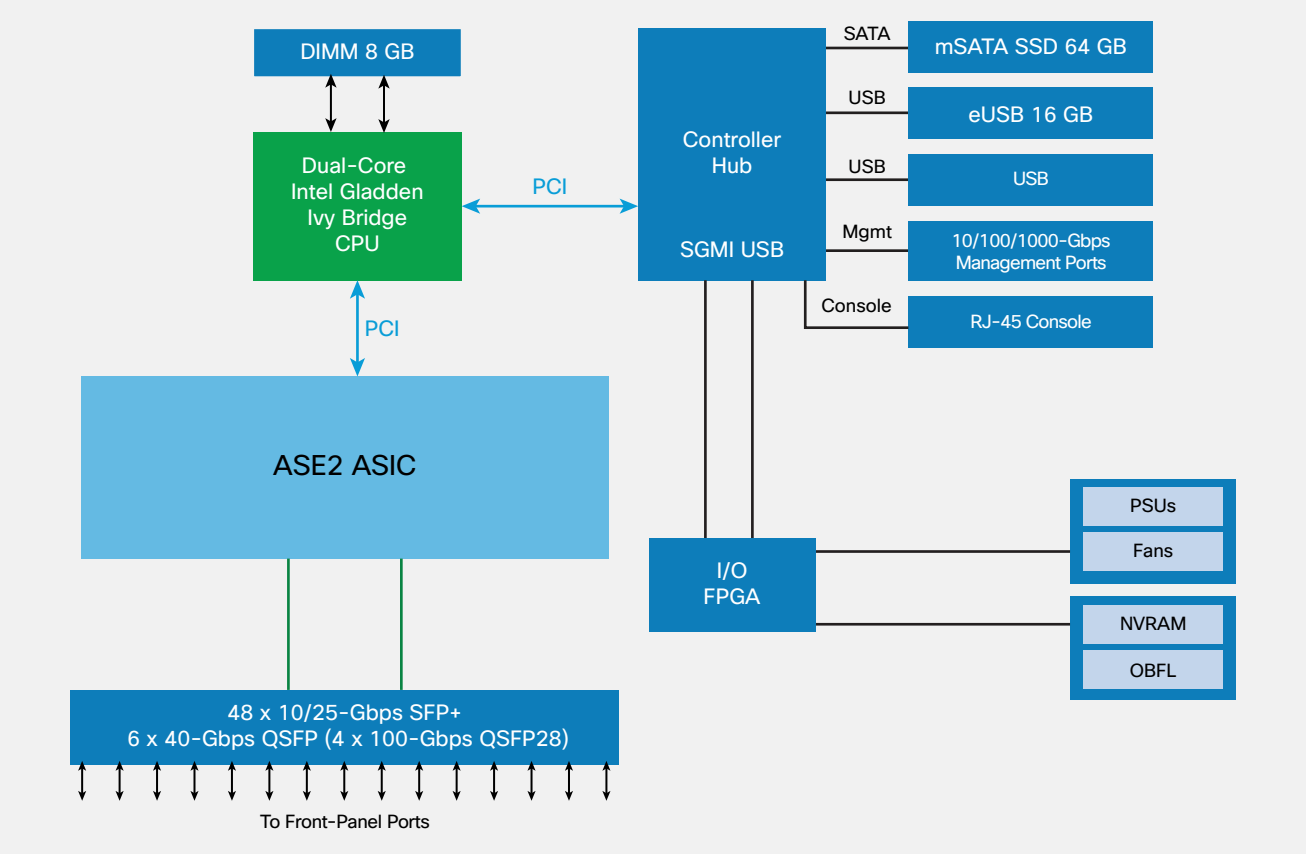
The Cisco Nexus 92160YC-X (Figure 8) is a 1RU switch that supports 3.2 Tbps of duplex bandwidth and a packet forwarding rate of 2.5 bpps. It provides 48 SFP+ ports, which can individually run at 1, 10, and 25-Gbps speeds, and 6 QSFP+ ports at 40 Gbps or 4 QSFP28 ports at 100 Gbps. The Cisco Nexus 92160YC-X provides dense 1, 10, and 25-Gbps connectivity with flexible 40 or 100-Gbps uplink capability. The QSFP ports on the switch can operate as 4 x 100 Gbps, 2 x 100 Gbps + 4 x 40 Gbps, or 6 x 40 Gbps. Two of the QSFP ports (eth1/50 and eth1/52) also support breakout to 4 x 10 Gbps, 2 x 25 Gbps, or 2 x 50 Gbps.

Figure 8. Cisco Nexus 92160YC-X Switch



As shown in Figure 9, the Cisco Nexus 92160YC-X has a dual-core Intel Ivy Bridge Gladden processor with 16 GB of system memory. Its data-forwarding plane is based on a single ASE3 ASIC.

Figure 9. Cisco Nexus 92160YC-X Switch Hardware Architecture



New-Generation ASICs in Cisco Nexus 9200 Platform

The Cisco Nexus 9200 platform switches are built with either the new Cisco ASE2 ASIC or ASE3 ASIC. As members of the Cisco next-generation data center ASIC family, the ASE2 and ASE3 ASICs are purposely built to provide a single-chip solution for data center switches with the following goals:

- Increase density. The platform supplements 40 Gigabit Ethernet density with 100 Gigabit Ethernet using QSFP28. It supplements 10 Gigabit Ethernet density with 25 Gigabit Ethernet SFP+ or 40 or 50 Gigabit Ethernet QSFP.
- Decrease the number of physical chips needed to build a given system configuration to achieve greater performance and power efficiency.
- Continue innovations in the networking space. Innovations include increased forwarding scalability, standalone IP and Layer 2 address scaling, large-scale Port Address Translation (PAT), increased packet-buffer memory, telemetry, analytics, and diagnostics.

ASE2 and ASE3 ASIC Architecture

ASE2 and ASE3 both have a multiple-slice design. Their architecture is similar, but they differ in port density, buffering capability, forwarding scalability, and some features.

Each ASIC has three main components:

- Slice components: The slices make up the switching subsystems. They include multiprotocol MAC, packet parser, forwarding lookup controller, I/O packet buffering, buffer accounting, output queuing, and scheduling components.
- I/O components: The I/O components consist of high-speed serializer/deserializer (SerDes) blocks.
- Global components: The global components make up the broadcast network. They include a set of point-to-multipoint wires to connect all the slices together. They also include the central statistics counter modules.

Both ASE2 and ASE3 support different port speeds, including 1, 10, 25, 40, 50, and 100 Gbps. Table 3 lists the port density of each ASIC. The 10-Gbps ports also support 1 Gbps.

Table 3. ASIC Port Characteristics

ASIC	1/10 Gigabit Ethernet Ports	25 Gigabit Ethernet Ports	40 Gigabit Ethernet Ports	100 Gigabit Ethernet Ports
ASE2	144	144	64	36
ASE3	72	64	18	16

ASE2 and ASE3 Forwarding Table

ASE2 and ASE3 ASICs both use a shared hash table known as the unified forwarding table (UFT) to store Layer 2 and Layer 3 forwarding information. The UFT size is 352,000 entries on both ASE2 and ASE3 ASICs. It is partitioned into different regions to support MAC address, IP host address, IP address longest-prefix match (LPM) and multicast lookups. The UFT is also used for next-hop and adjacency information and reverse-path forwarding (RPF) check entries for multicast traffic.

The UFT is internally composed of multiple tiles. Each tile can be independently programmed for a particular forwarding table function. This programmable memory sharing provides flexibility to address a variety of deployment scenarios and increases the efficiency of memory resource utilization.

In addition to the UFT, the ASICs have a 16,000-entry ternary content-addressable memory (TCAM) that can be used for forwarding lookup information.

With the programmable shared hash table memory, forwarding table carving for different forwarding functions on the Cisco Nexus 9200 platform can be configured in hardware to address different deployment scenarios in the data center network. The switch operating system, NX-OS, can place a software control on top of the flexible hardware to support validated common forwarding table profiles.

Table 4 lists the maximum capacity in the ASIC hardware for each table type and for the default forwarding table profile set by NX-OS. For additional profiles, refer to the validated scalability white paper for the particular NX-OS release of interest.

Table 4. ASIC Table Capacity

Table	Maximum in ASIC Hardware	Default in Cisco NX-OS Software
LPM IPv4 routes	256,000	16,000
LPM IPv6 (/64) routes	256,000	4000
LPM IPv6 (/65-/127) routes	128,000	4000
IPv4 host routes	256,000	192,000 or 96,000 (Equal-Cost Multipath [ECMP])
IPv6 host routes	128,000	96,000 or 48,000(ECMP)
Multicast	32,000	10,000
MAC addresses	256,000	96,000

ASE2 and ASE3 Buffer Architecture

The slices in the ASE2 and ASE3 ASICs function as switching subsystems. Each slice has its own buffer memory, which is shared among all the ports on this slice.

To efficiently use the buffer memory resources, the raw memory is organized into 208-byte cells, and multiple cells are linked together to store the entire packet. Each cell can contain either an entire packet or part of a packet. Table 5 summarizes the amount of buffer space on each ASIC.

Table 5. ASIC Buffer Capacity

ASIC	Number of 100 Gigabit Ethernet Ports	Number of Slices	Number of Buffer Cells per Slice	Buffer Size per Slice	Total Buffer Size
ASE2	36	6	24,000	5.1 MB	30.6 MB
ASE3	16	2	48,000	10.2 MB	20.4 MB

Both ASE2 and ASE3 ASICs support 10 classes of service: 8 user-defined classes of service, 1 Cisco Switched Port Analyzer (SPAN) class of service, and 1 CPU class of service. The software can partition the buffer into a maximum of four pool groups. For example, drop and no-drop (enabled with Priority Flow Control [PFC]) classes have different pool groups; and CPU and SPAN classes have different pool groups than user-defined classes. A certain number of cells are allocated to each pool group, and they are not shared among pool groups. This approach helps guarantee buffer resources for each pool group for the traffic types that the group serves.

Buffer Allocation

The bulk memory of the packet buffer can be statically partitioned by software between input and output through the switch configuration. By default, the Cisco Nexus 9200 platform uses class-based egress queuing, so most buffer cells are allocated to the egress queue. However, if PFC is enabled, the switch will use ingress queues for the no-drop classes to handle Pause operations. With this configuration, more buffer cells are dedicated to the ingress queue. This configuration-based buffer partitioning between ingress and egress queues increases the effective buffer resources for the queuing strategy deployed on the switch.

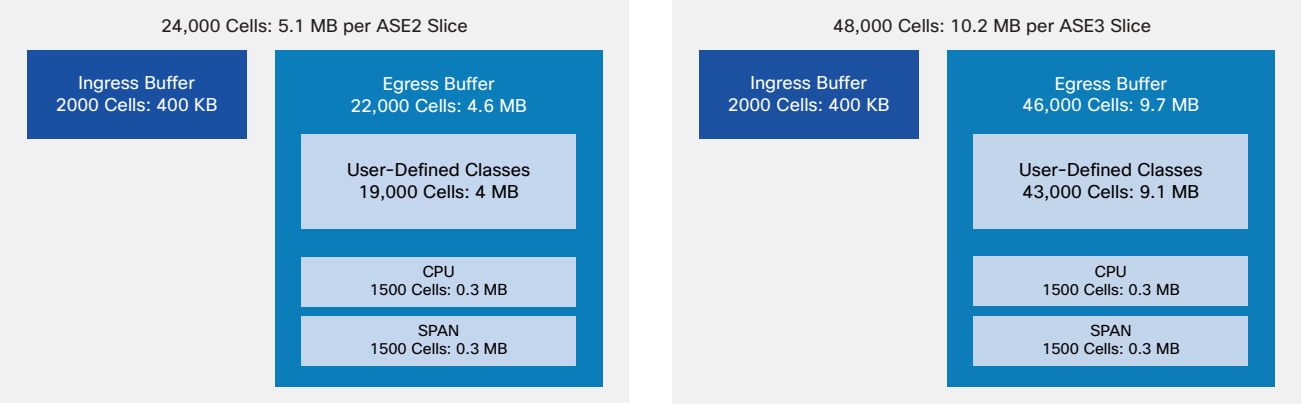
Figure 10 shows the default per-slice buffer allocation on ASE2 and ASE3. It shows that most buffer cells are allocated to the egress pool groups, except for a minimum buffer allocation for the ingress buffer.

Three egress buffer pool groups are used:

- User-defined classes
- CPU
- SPAN

Within the pool group for the user-defined classes, up to 16 pools can be created and maintained: two for each class of service (one for unicast traffic and one for multicast traffic in each class).

Figure 10. Default Buffer Allocations on ASE2 and ASE3 ASICs

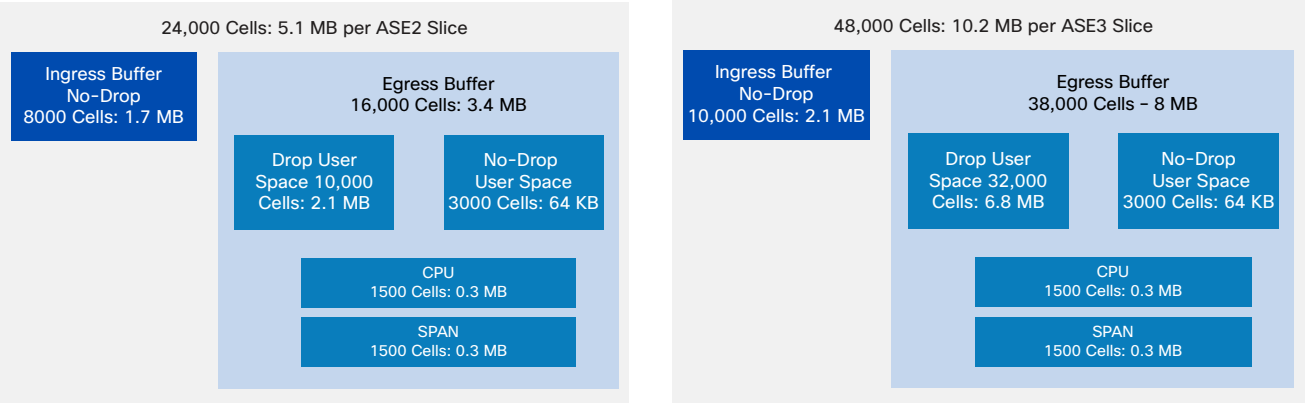


Both ASE2 and ASE3 support PFC. PFC provides lossless semantics for traffic in the no-drop classes by using the per-class and per-port Pause mechanism for the upstream devices. ASE2 and ASE3 ASICs handle Pause using an ingress buffer and can support up to three no-drop classes. In a design with a large port count, use of an ingress buffer to handle Pause is more efficient because the buffer size needs to accommodate the Pause latencies only for the input port. If the Pause buffer is implemented at output, then the shared memory needs to handle the worst case for the sum of all the ports on the switch.

When PFC is enabled on the Cisco Nexus 9200 platform, the switch allocates a certain amount of buffer to the ingress queue on each ASIC slice. This ingress buffer is shared across all the ports in the slice and is partitioned per pool and per port. A pool is an internal construct, and software configuration defines the mapping of classes to pools.

Figure 11 shows the buffer allocation on the ASICs when PFC is enabled. A large number of buffer cells are reserved for the ingress no-drop queues.

Figure 11. Buffer Allocation with PFC on ASE2 and ASE3 ASICs



Intelligent Buffer Management

Both ASE2 and ASE3 ASICs have built-in intelligent buffer management functions, primarily Approximate Fair Drop (AFD) and Dynamic Packet Prioritization (DPP), for active queue management. The intelligent buffer functions add per-flow control to the existing congestion avoidance and congestion management mechanisms to provide better application performance.

Approximate Fair Discard

AFD is a flow-aware early-discard mechanism to signal network congestion to TCP. Prior to AFD, Weighted Random Early Discard (WRED) was the primary technology for congestion signaling, also known as Active Queue Management (AQM). WRED applies an early-discard buffer threshold to each class-based weighted queue, but it doesn't have flow awareness within a class. Hence, it has to treat all traffic flows equally and drops packets randomly for all flows. This random discard process can yield detrimental packet drops to short-lived small (mice) flows, which are more sensitive to packet loss, while potentially still leaving long-lived large (elephant) flows occupying most of the buffer. As a result, the flow completion time for the mice flows can suffer drastically, and the elephant flows cannot achieve fairness among themselves either.

AFD, to the contrary, takes into account information about flow sizes and data arrival rates before making a drop decision. Therefore, the algorithm can protect packet-loss-sensitive mice flows and provide fairness to competing elephant flows.

Using an elephant trap (ETRAP), AFD can differentiate short-lived mice flows from long-lived elephant flows within a given traffic class and submit only the elephant flows to the AFD early-discard function. A flow can be defined using multiple parameters, but typically the 5-tuple is used. AFD uses a hash table to track all the active flows and measure their byte counts on ingress. A user-configurable byte-count-based ETRAP threshold is deployed to decide whether a flow is a mice flow or an elephant flow. A flow is a mice flow if it transmits fewer bytes than the ETRAP threshold during its lifespan. After the byte count of a flow exceeds the ETRAP threshold, the flow is considered an elephant flow, and it is moved to the elephant flow table for further tracking and is subject to AFD drop decisions.

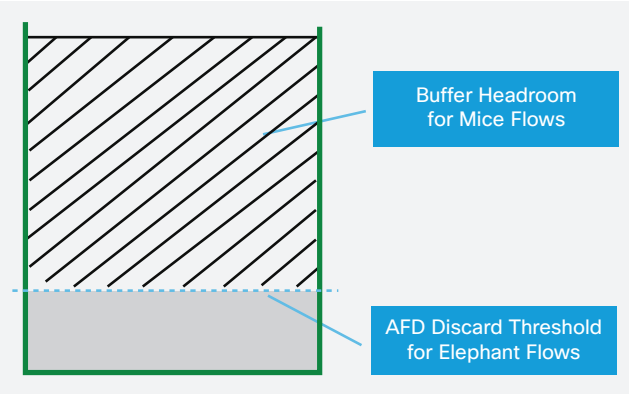
In addition, AFD has the intelligence to apply fair discards among elephant flows based on their data arrival rate when the AFD early-discard buffer threshold is crossed. The algorithm has two main elements.

One element is rate measurement: ETRAP measures the arrival rate of each flow in the elephant flow table on the ingress port, and the measured arrival rate is carried in the packet header when packets are internally forwarded to the egress port.

The other main element of AFD is the fair-rate calculation: the AFD algorithm dynamically computes a per-flow fair rate on an egress port using a feedback mechanism based on the egress port queue occupancy. When a packet of an elephant flow enters the egress port queue, the AFD algorithm compares the measured arrival rate of the flow with the computed fair share. If the arrival rate of an elephant flow is less than the per-flow fair rate, the packet is not dropped. However, if the arrival rate exceeds the computed per-flow fair rate on the egress port, packets will be dropped from that flow in proportion to the amount that the flow exceeds the fair rate. The drop probability is computed using the fair rate and the measured flow rate. As a result, all elephant flows achieve the fair rate. The AFD parameters for the output queues are configured using profiles. The profile, as with WRED, can be configured to mark a packet with explicit congestion notification (ECN) instead of dropping it.

Figure 12 shows the overall effect of AFD. By submitting only elephant flows to the early-discard algorithm, AFD can prevent unwanted packet drops in mice flows and preserve enough buffer headroom to accommodate bursts caused by a large number of simultaneous mice flows (incast and microburst traffic). Among long-lived elephant flows, the AFD algorithm applies fair early discard based on the data arrival rate.

Figure 12. AFD Flow-Based Early Discard



Dynamic Packet Prioritization

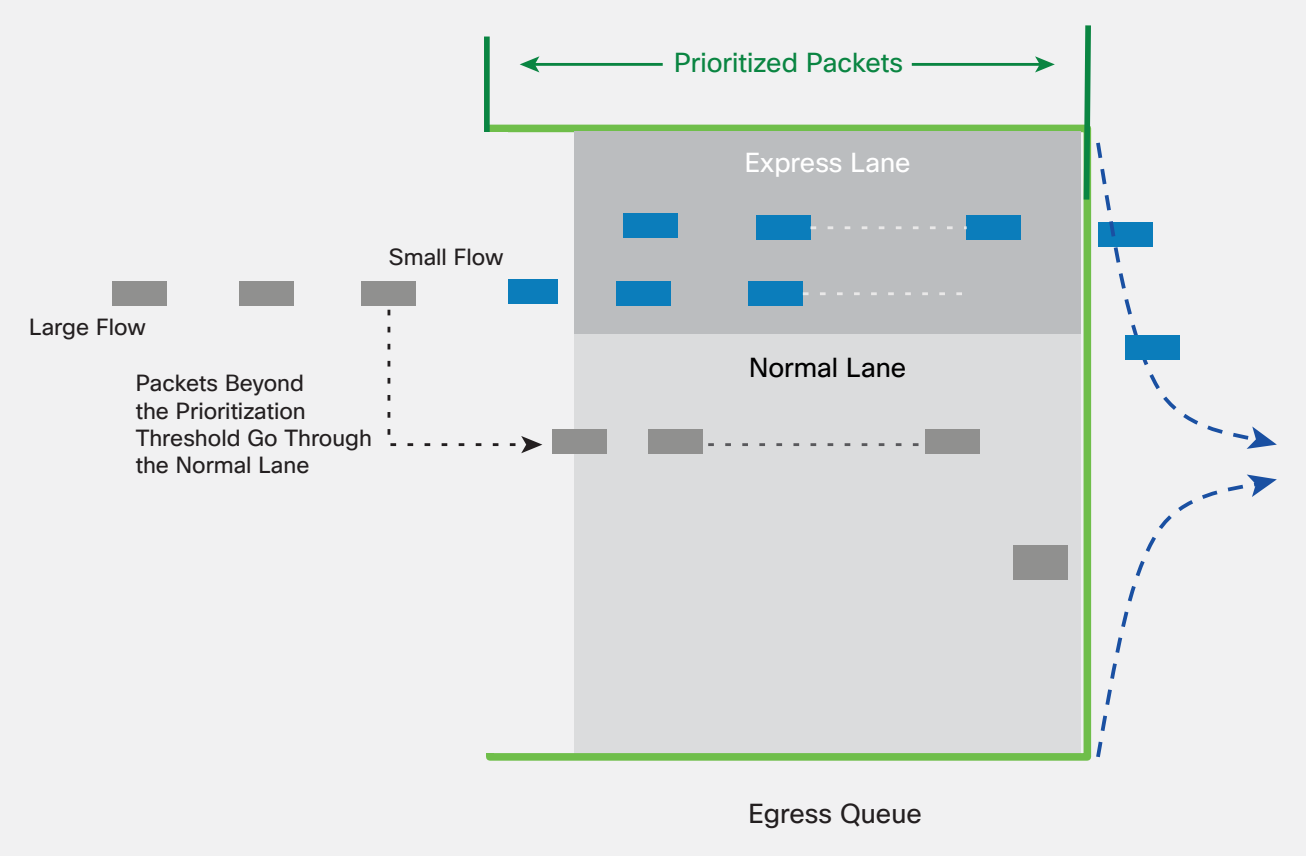
DPP can provide significant latency benefits for short-lived small flows under network congestion by automatically giving priority to the first few packets from each flow.

As a traffic flow traverses an egress queue, its packet count is measured and checked against a user-configurable packet-count-based prioritization threshold. If the number of packets received in a flow is below the prioritization threshold, the packets are prioritized to bypass the rest of the queue. If the packet count of the flow exceeds the threshold, the excessive packets in the flow will not be prioritized any more. Because short-lived small flows, such as microburst flows, consist of very few packets per flow, they will not cross the threshold, and hence the entire small flow is prioritized. For long-lived large flows, after the initial few packets allowed by the threshold, the rest of the flow will go through the normal queuing process.

As shown in Figure 13, DPP essentially creates an express lane for short-lived small flows, and leaves long-lived large flows in the normal lane. This approach allows small flows to have priority both in the switch and the network to reduce the number of drops and decrease latency. Because small flows in most data center applications are more sensitive to packet loss and long latency than are long-lived large flows, prioritizing small flows improves overall application performance.

Flow prioritization can be combined with the AFD algorithm to drop fairly among the long-lived large flows and prioritize the small flows with sufficient buffer space to accommodate a large number of simultaneous small flows (incast and microburst traffic). This approach reduces the mean queue length without increasing the number of timeouts for small flows, providing significant performance improvement.

Figure 13. Dynamic Packet Prioritization



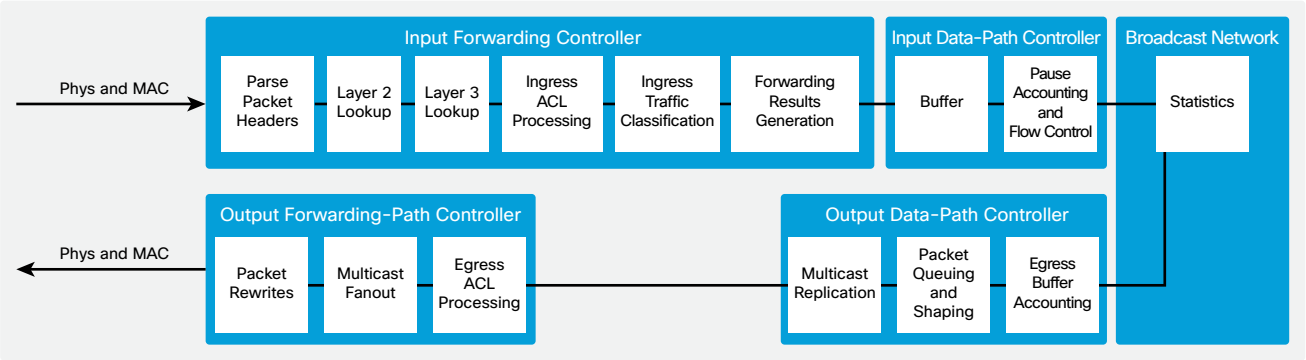
Cisco Nexus 9200 Platform Unicast Packet Forwarding

Forwarding Pipelines on ASE2 and ASE3 ASICs

Unicast packet forwarding on the Cisco Nexus 9200 platform is performed by the network forwarding engine: either the ASE2 or ASE3 ASIC. The ASE2 and ASE3 ASICs have multiple-slice designs, using six or two slices, respectively. Each slice represents a switching subsystem with both an ingress forwarding pipeline and an egress forwarding pipeline. The ingress forwarding pipeline on each slice consists of an I/O component, input forwarding controller, and input data-path controller. The egress forwarding pipeline consists of the output data-path controller, output forwarding-path controller, and another I/O component. All slices are connected to a broadcast network that provides point-to-multipoint connections from each slice, allowing all-to-all connectivity between slices. The broadcast network provides enough bandwidth to support full-line-rate forwarding between all slices concurrently.

Figure 14 shows the forwarding pipelines on a Cisco Nexus 9200 platform switch. When a packet enters a Cisco Nexus 9200 platform switch, it goes through the ingress pipeline of the slice on which the ingress port resides, traverses the ASIC internal broadcast network to get onto the egress slice, and then goes through the egress pipeline of the egress slice.

Figure 14. Forwarding Pipelines on the Cisco Nexus 9200 Platform



Ingress Pipeline: Input Forwarding Controller

The input forwarding controller receives the packet from the ingress port MAC, parses the packet headers, and performs a series of lookups to decide whether to accept the packet and how to forward it to its intended destination. It also generates instructions to the data path to store and queue the packet. Because the Cisco next-generation ASIC switches are cut-through switches, input forwarding lookup is performed while the packet is being stored in the pause buffer block. The input forwarding controller performs multiple tasks in the sequence shown earlier in Figure 14:

- Packet-header parsing
- Layer 2 lookup
- Layer 3 lookup
- Ingress access control list (ACL) processing
- Ingress traffic classification
- Forwarding results generation

Packet-Header Parsing

When a packet enters through a front-panel port, it goes through the ingress pipeline, and the first step is packet-header parsing. The flexible packet parser parses the first 128 bytes of the packet to extract and save information such as the Layer 2 header, EtherType, Layer 3 header, and TCP IP protocols. This information is used for subsequent packet lookup and processing logic.

Layer 2 and Layer 3 Forwarding Lookup

As the packet goes through the ingress pipeline, it is subject to Layer 2 switching and Layer 3 routing lookups. First, the forwarding process examines the destination MAC address (DMAC) of the packet to determine whether the packet needs to be switched (Layer 2) or routed (Layer 3). If the DMAC matches the switch's own router MAC address, the packet is passed to the Layer 3 routing lookup logic. If the DMAC doesn't belong to the switch, a Layer 2 switching lookup based on the DMAC and VLAN ID is performed. If a match is found in the MAC address table, the packet is sent to the egress port. If there is no match for DMAC and VLAN combination, the packet is forwarded to all ports in the same VLAN.

Inside the Layer 3 lookup logic, the destination IP address (DIP) is used for searches in the Layer 3 host table. This table stores forwarding entries for directly attached hosts and learned /32 host routes. If the DIP matches an entry in the host table, the entry indicates the destination port, next-hop MAC address, and egress VLAN. If no match for the DIP is found in the host table, an LPM lookup is performed in the LPM routing table.

Ingress ACL Processing

In addition to forwarding lookup processing, the packet undergoes ingress ACL processing. The ACL TCAM is checked for ingress ACL matches. Each ASIC has an ingress ACL TCAM table of 4000 entries per slice to support system internal ACLs and user-defined ingress ACLs. These ACLs include port ACLs (PACLs), routed ACLs (RACLs), and VLAN ACLs (VACLs). ACL entries are localized to the slice and are programmed only where needed. This approach makes the best use of the ACL TCAM in the Cisco Nexus 9200 platform switch.

Ingress Traffic Classification

Cisco Nexus 9200 platform switches support ingress traffic classification. On an ingress interface, traffic can be classified based on the address field, IEEE 802.1q class of service (CoS), and IP precedence or differentiated services code point (DSCP) in the packet header. The classified traffic can be assigned to one of the eight quality-of-service (QoS) groups. The QoS groups internally identify the traffic classes that are used for subsequent QoS processes as packets traverse the system.

Ingress Forwarding Result Generation

The final step in the ingress forwarding pipeline is to collect all the forwarding metadata generated earlier in the pipeline and pass it to the downstream blocks through the data path. A 64-byte internal header is stored along with the incoming packet in the packet buffer. This internal header includes 16 bytes of iETH (internal communication protocol) header information, which is added on top of the packet when the packet is transferred to another the output data-path controller through the broadcast network. This 16-byte iETH header is stripped off when the packet exits the front-panel port. The other 48 bytes of internal header space is used only to pass metadata from the input forwarding queue to the output forwarding queue and is consumed by the output forwarding engine.

Ingress Pipeline: Input Data-Path Controller

The input data-path controller performs ingress accounting functions, admission functions, and flow control for the no-drop class of service. The ingress admission control mechanism determines whether a packet should be admitted into memory. This decision is based on the amount of buffer memory available and the amount of buffer space already used by the ingress port and traffic class. The input data-path controller forwards the packet to the output data-path controller through the broadcast network.

Broadcast Network and Central Statistics Module

The broadcast network is a set of point-to-multipoint wires that allows connectivity between all slices on the ASIC. The input data-path controller has a point-to-multipoint connection to the output data-path controllers on all slices, including its own slice. The central statistics module is connected to the broadcast network. The central statistics module provides packet, byte, and atomic counter statistics.

Egress Pipeline: Output Data-Path Controller

The output data-path controller performs egress buffer accounting, packet queuing, scheduling and multicast replication. All ports dynamically share the egress buffer resource. The details of dynamic buffer allocation are described earlier in this document.

The output data-path controller also performs packet shaping. Following the design principle of simplicity and efficiency, the Cisco Nexus 9200 platform uses a simple egress queuing architecture. In the event of egress port congestion, packets are directly queued in the buffer of the egress slice. There are no virtual output queues (VoQs) on the ingress slice. This approach greatly simplifies system buffer management and queuing implementation.

A Cisco Nexus 9200 switch can support up to 10 traffic classes on egress, 8 user-defined classes identified by QoS group IDs, a CPU control traffic class, and a SPAN traffic class. Each user-defined class can have a unicast queue and a multicast queue per egress port. This approach helps ensure that no single port can consume more than its fair share of the buffer memory and cause buffer starvation for other ports.

Egress Pipeline: Output Forwarding Controller

The output forwarding controller receives the input packet and associated metadata from the buffer manager and is responsible for all packet rewrites and egress policy application. It extracts internal header information and various packet-header fields from the packet, performs a series of lookups, and generates the rewrite instructions.

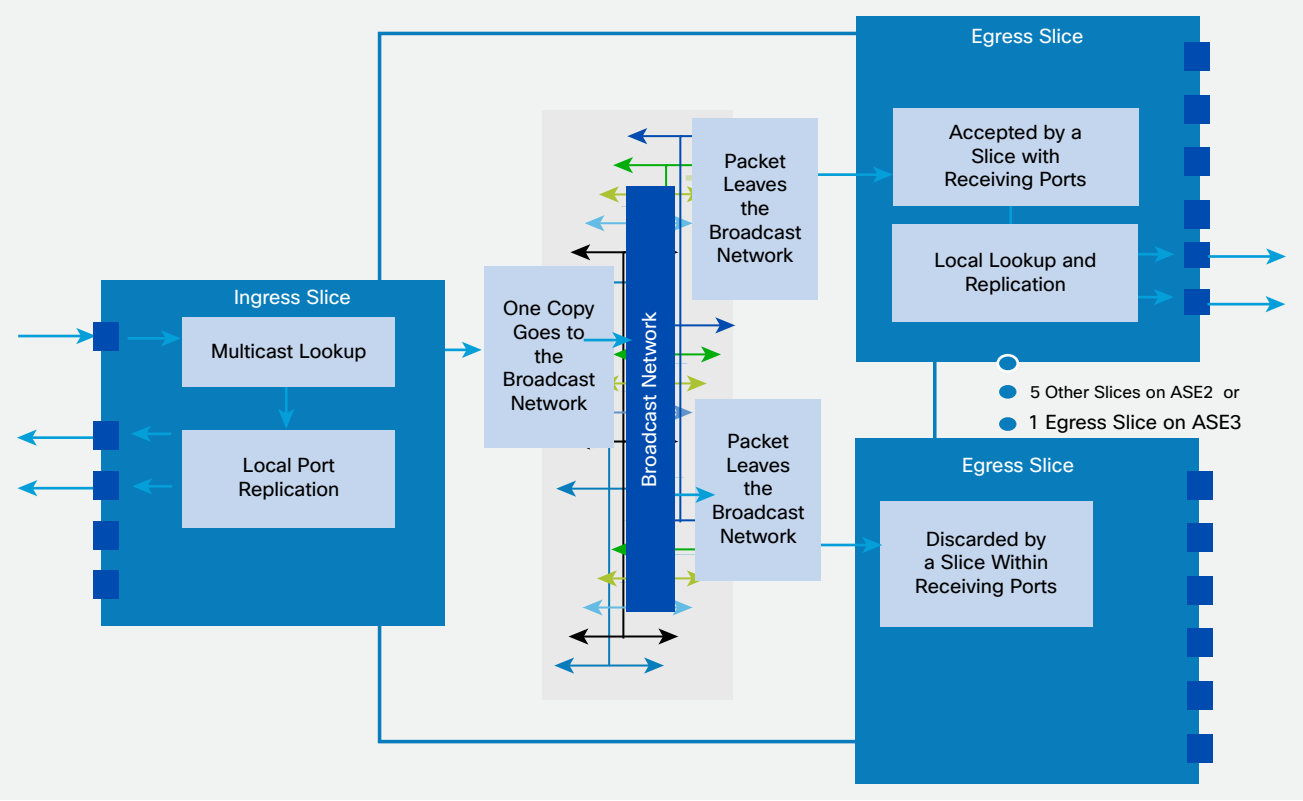
Cisco Nexus 9200 Platform Multicast Packet Forwarding

Multicast packets go through the same ingress and egress forwarding pipelines as the unicast packets on a Cisco Nexus 9200 platform switch, except that multicast forwarding lookup uses multicast tables, and multicast packets go through a multistage replication process to be forwarded to multiple destination ports.

ASE2 and ASE3 ASICs both consist of multiple slices that are interconnected by a nonblocking internal broadcast network. When a multicast packet arrives at a front-panel port, the ASIC performs a forwarding lookup. This lookup resolves local receiving ports on the same slice as the ingress port and provides a list of intended receiving slices that have receiving ports in the destination multicast group. The packet is replicated on the local ports, and one copy of the packet is sent to the internal broadcast network, with the bit vector in the internal header set to indicate the intended receiving slices. Only the intended receiving slices will accept the packet off the wire of the broadcast network. The slices without receiving ports for this group will simply discard the packet. The receiving slice then performs local Layer 3 replication or Layer 2 fanout lookup and replication to forward a copy of the packet to each of its local receiving ports.

Figure 15 shows the multicast forwarding process.

Figure 15. Multicast Forwarding Process



Conclusion

Cisco Nexus 9200 platform switches are industry-leading, ultra-high-density, fixed-configuration data center switches. They offer line-rate Layer 2 and 3 features that support enterprise applications, service provider hosting, and cloud-scale environments. Built with the new generation of Cisco ASICs for data center switching, the Cisco Nexus 9200 platform delivers 25, 50, and 100Gbps speeds at the cost of 10 and 40 Gbps, which represents 250 percent increase in capacity.

The Cisco Nexus 9200 platform also supports the current 1, 10, and 40Gbps speeds with flexible combinations of 1, 10, 25, 40, 50, and 100Gbps connectivity. This flexibility allows organizations to deploy the Cisco Nexus 9200 platform in their current networks for 1, 10, and 40-Gbps connectivity and link speeds, while getting ready to transform to 25, 50, and 100-Gbps speeds without additional costs for the switch platforms. This approach provides a platform that can support your future needs and let you smoothly transition your data center networks to sustain increasing requirements from applications, allowing you to offer new and better services.

For More Information

For more information about the Cisco Nexus 9000 Series Switches, see the detailed product information at the product homepage at http://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/index.html.

C11-737204-02 09/16