Cisco Nexus 3100 Platform Switch Architecture

White Paper

October 2013

What You Will Learn

The Cisco Nexus® 3100 platform switches are second-generation Cisco Nexus 3000 Series Switches and offer improved port density, scalability, and features compared to the first-generation switches. This platform includes the Cisco Nexus 3132Q and 3172PQ Switches. This document provides an overview of the features of the Cisco Nexus 3100 platform and a detailed description of the internal architecture.

Cisco Nexus 3132Q Switch Overview

The Cisco Nexus 3132Q Switch is a dense, high-performance Layer 2 and 3 10 and 40 Gigabit Ethernet switch. It has a compact one-rack-unit (1RU) form factor and runs the industry-leading Cisco® NX OS Software operating system, providing customers with comprehensive features and functions that are widely deployed.

Similar to the other Nexus devices, the ports on the Nexus 3100 switches are rear facing. The Cisco Nexus 3132Q (Figure 1) has 32 40-Gbps Quad Small Form-Factor Pluggable (QSFP+) ports. As an added benefit, the first QSFP+ ports are physically shared with four SFP+ ports allowing customers the ability to use SFP+ form factor transceivers natively on the device. All of the QSFP+ ports on the device can operate as a native 40-Gbps port or a four independent 10-Gbps ports. Rear side of the switch also has a serial console port, USB port, PPS connector and an out-of-band 10/100/1000-Mbps Ethernet management ports.

The front side of the switch has two redundant hot-pluggable power supplies and four individual fan modules. Since datacenters cannot afford downtime, the built-in redundancy of the switch allows full operations even after the failure of a power supply and a fan module. To further add operational flexibility for customer's, the device supports both in server rack and in network rack deployments. This is made possible with a choice of airflow power supplies and fan. For the datacenters that have deployed VDC power to devices, there is also the option of VAC and VDC supplies for the devices.

The Cisco Nexus 3132Q is well suited for data centers that require a cost-effective, power-efficient, line-rate Layer 2 and 3 40 Gigabit Ethernet aggregation-layer or spine switch.

Figure 1. Cisco Nexus 3132Q Rear Panel

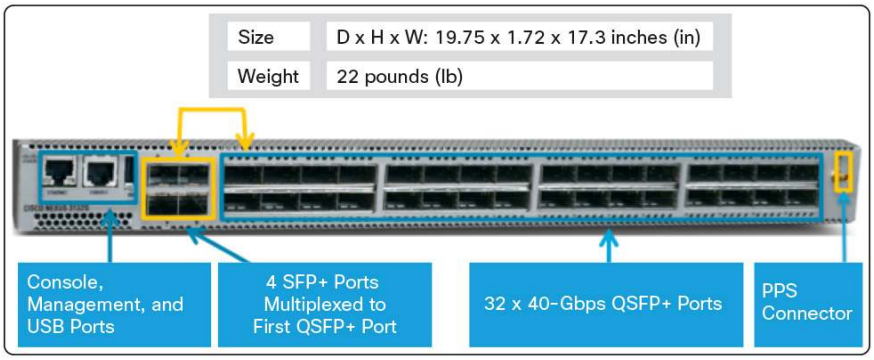


Figure 2. Cisco Nexus 3132Q Front Panel

Cisco Nexus 3172PQ Switch Overview

The Cisco Nexus 3172PQ Switch is a dense, high-performance Layer 2 and 3 10 and 40 Gigabit Ethernet switch. It has a compact 1RU form factor and runs the industry-leading Cisco NX‑OS Software operating system, providing customers with comprehensive features and functions that are widely deployed.

The rear side of the Cisco Nexus 3172PQ has 48 x 10-Gbps SFP+ ports and 6 additional QSFP+ ports. The QSFP+ ports can operate in native 40-Gbps or 4 x 10-Gbps mode.

The front side of the switch has two N+N redundant hot-pluggable power supplies and four individual fan modules. The switch can run with a single power supply and with one failed fan module. It supports both forward- and reverse-airflow schemes with AC and DC power inputs. The switch also has a serial console port, a USB port, and an out-of-band 10/100/1000-Mbps Ethernet management port.

The Cisco Nexus 3172PQ is well suited for data centers that require a cost-effective, power-efficient, line-rate Layer 2 and 3 access or leaf switch.

Cisco Nexus 3100 Platform Features

The Cisco Nexus 3100 platform switches offer the following features:

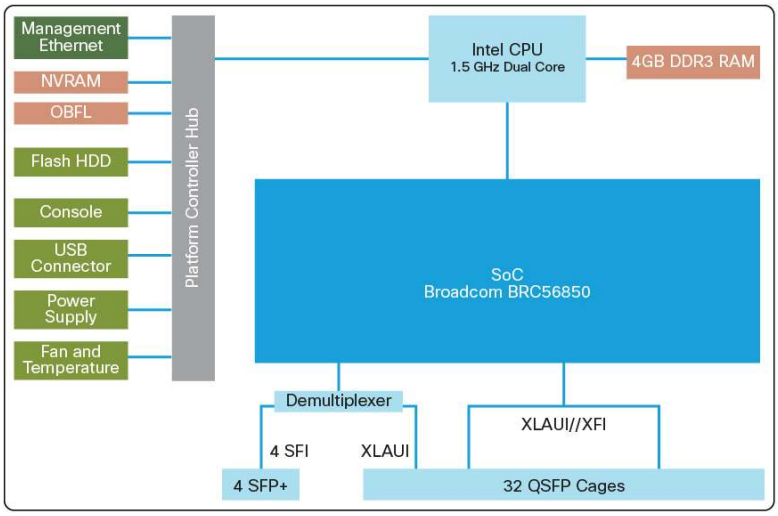
- **High density and high availability:** The Cisco Nexus 3132Q provides 32 x 40-Gbps ports or up to 104 x 10-Gbps ports in 1RU. The Cisco Nexus 3100 platform is designed with redundant and hot-swappable power supplies and individual fan modules that can be accessed from the front panel, where status lights offer an at-a-glance view of switch operation. The switch can function properly with one failed fan and one failed power supply. For high reliability, any fan or power supply can be hot swapped during operations. To support efficient data center hot- and cold-aisle designs and deployments into server racks or network racks, both front-to-back (port side air exhaust) and back-to-front (port side air intake) cooling is supported. The switch supports graceful restart helper capability for Open Shortest Path First (OSPF), Enhanced Interior Gateway Routing Protocol (EIGRP), and Border Gateway Protocol (BGP).
- **Nonblocking line-rate performance:** The Cisco Nexus 3100 platform can handle packet flows at wire speed on all ports to provide throughput of up to 2.5 terabits per second (Tbps) of bidirectional bandwidth, with forwarding performance of up to 1.4 billion packet per second (bpps).
- **Ultra-low latency:** The cut-through switching technology along with the use of a single ASIC switch-on-a-chip (SoC) architecture of the Cisco Nexus 3100 platform enables the switches to offer ultra-low latency.
- **Shared-buffer architecture:** The Cisco Nexus 3100 platform switches have 12.2 MB of buffer space, including per-port and dynamically allocated shared buffer space.

- **Separate egress queues for unicast and multicast traffic:** The Cisco Nexus 3100 platform supports 16 egress data queues: 8 each for unicast and multicast traffic configurable in 8 quality-of-service (QoS) groups. In addition, it supports dedicated queues for control-plane traffic: 2 each for unicast and multicast traffic.
- **Robust Layer 3 mode:** The Cisco Nexus 3100 platform has a comprehensive Layer 3 feature set that includes full BGP support along with many other features. For more information, see the Cisco Nexus 3100 platform data sheet.
- **Multicast:** Protocol-Independent Multicast sparse mode (PIM-SM), PIM source-specific multicast (PIM-SSM), and Multicast Source Discovery Protocol (MSDP) multicast protocols are supported. The Cisco Nexus 3100 platform also supports Bidirectional PIM (PIM-Bidir) in the hardware; software support will be added after the initial release.
- **QoS:** The Cisco Nexus 3100 platform supports extensive QoS features, including traffic classification and marking, traffic shaping, policing, congestion management, weighted random early detection (WRED), and explicit congestion notification (ECN).
- **Virtual Extensible LAN (VXLAN):** The Cisco Nexus 3100 platform can support VXLAN in the hardware. VXLAN is a Layer 2 network isolation technology that uses a 24-bit segment identifier to scale beyond the 4000-instance limitation of VLANs. VXLAN technology creates LAN segments by using an overlay approach with MAC address-in-IP encapsulation.

Cisco Nexus 3100 Platform Architecture

The Cisco Nexus 3100 platform is implemented with an SoC design. Figure 3 shows the Nexus 3132Q architecture with the switch's onboard components and connectivity.

Figure 3. Cisco Nexus 3132Q Data-Plane and Switch-on-a-Chip Architecture



The Cisco Nexus 3100 platform control plane runs Cisco NX-OS Software on a dual-core 1.5-GHz Intel processor with 4 GB of DDR 3 RAM. The supervisor complex is connected to the data plane in band through two internal PCI Express (PCIe) links, and the system is managed with in band or through the out-of-band 10/100/1000-Mbps management port.

Table 1 summarizes the control-plane specifications for the Cisco Nexus 3100 platform.

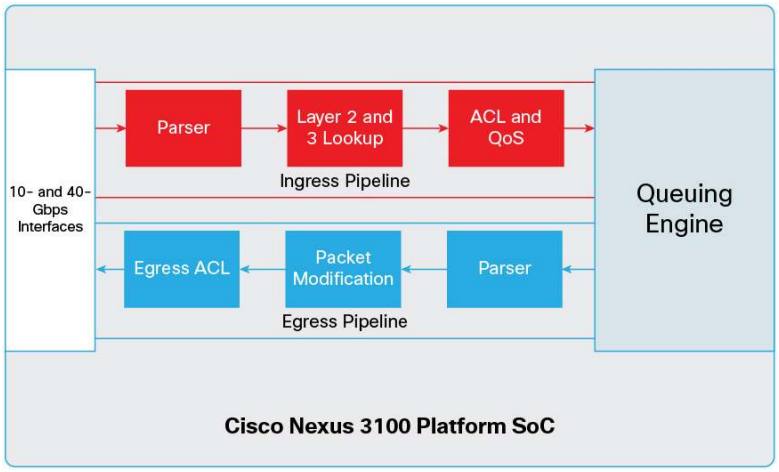
Table 1. Control-Plane Specifications

Component	Specification
CPU	1.5-GHz Intel Processor (dual core)
DRAM	4 GB of DD
Persistent disk	2 GB of embedded flash memory for base system storage
NVRAM	16 MB to store syslog, licensing information, and reset reason
On-board fault log	512 MB of flash memory to store hardware-related fault and reset reasons
Boot and BIOS flash memory	64 MB to store images
Management interface	RS-232 console port and 10/100/1000BASE-T management ports: mgmt0; 1 external USB flash port can be used for updating the system configuration

Cisco Nexus 3100 Platform Switch-on-a-Chip Forwarding

Figure 4 shows the internal blocks in the forwarding ASIC that every packet traverses before it is forwarded out of the egress interface. The ASIC has three blocks: ingress pipeline, queuing engine, and egress pipeline.

Figure 4. Cisco Nexus 3100 Platform Switch-on-a-Chip Packet Flow



- The ingress pipeline is responsible for most of the switch features, such as MAC address learning, forwarding lookup, access control list (ACL) policy lookup, and QoS classification and policing. In essence, the ingress pipeline block makes the forwarding decision. When a packet arrives at the ingress port, the parser engine parses the incoming packets and extracts the fields required for lookup. Then the Layer 2 and 3 forwarding lookup is performed in the MAC address and Layer 3 forwarding tables. The forwarding lookup would take additional hardware table resources in addition to the MAC address and Layer 3 forwarding table based on the type of traffic. Which could be Layer 2 unicast switched, L2 multicast switched, Layer 3 unicast routing or layer 3 multicast traffic.

  For Layer 2 unicast switching traffic, VLAN assignment is performed for untagged packets using a port-based table. A tagged packet has a VLAN ID from the packet. The next step is the learning phase: the source MAC address is learned in the hardware for the given VLAN. Depending on the destination MAC address lookup result, the packet can be forwarded to a destination, to the Layer 3 processing engine, or to the CPU, or it can be flooded to all members of a VLAN.

For Layer 3 unicast routing traffic, which would have destination MAC address of the switch interface, MAC address table lookup would forward internally to Layer 3 forwarding lookup. Layer 3 forwarding lookup is done in the Layer 3 host and longest-prefix match (LPM) tables. This lookup provides a result that is indexed to the next-hop table, which lists the outgoing interface and the destination MAC address. The outgoing interface provides an index to the Layer 3 interface table that supplies the source MAC address and the VLAN.

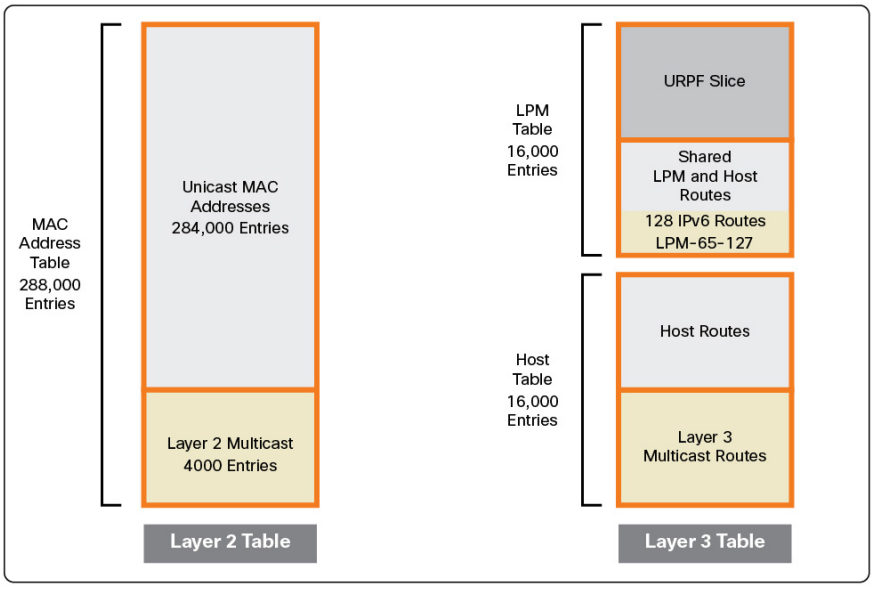
For multicast packets, when a Layer 2 lookup returns a hit, the packet is forwarded based on the destination MAC address in the Layer 2 multicast table. If the lookup returns a hit in the Layer 3 IP multicast table, the packet is replicated on the egress ports and VLANs using the multicast group and VLAN table.

- The queuing engine is responsible for flow control, congestion management, replication, buffering. and scheduling.
- The egress pipeline is responsible for packet rewrite based on the lookup result in the ingress pipeline. When the packet leaves the queuing engine, the lookup result is used to rewrite the packet after it has been parsed; then the packet is forwarded out of the egress interface.

Cisco Nexus 3100 Platform Forwarding Table

The Cisco Nexus 3100 platform has three main forwarding tables: the MAC address table, the host table, and the LPM table. The MAC address table is used for Layer 2 unicast and Layer 2 multicast forwarding. The host table is used for Layer 3 host routes, IPv4 routes with prefix 32 and IPv6 routes with prefix 128, and Layer 3 multicast routes. The LPM table is used for IPv4 and IPv6 unicast summary routes. The LPM table is also used for Unicast Reverse-Path Forwarding (URPF); when this feature is enabled, half the LPM table is reserved for URPF. IPv6 summary routes with prefixes 65 through 127 are placed in a separate region in the LPM table. Cisco NX-OS provides a command-line interface (CLI) option to change the table carving: to change the Layer 2 and Layer 3 multicast table sizes, disable URPF, and change the IPv6 LPM-65-127 region. Figure 5 shows the default table carving.

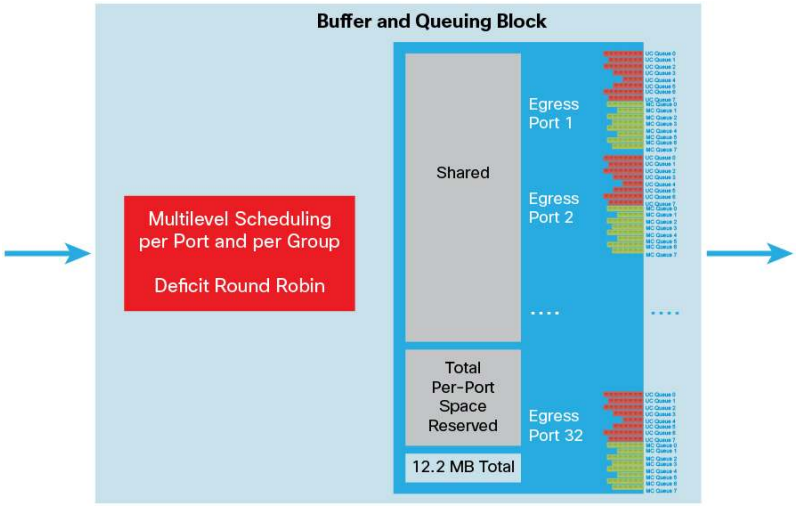
Figure 5. Cisco Nexus 3100 Platform Forwarding Table



Cisco Nexus 3100 Platform Queuing Engine

The Cisco Nexus 3100 platform has shared buffer-pool memory, which is part of the queuing engine (Figure 6). The buffer block has a total of 12.2 MB of buffer space for the platform. This buffer block provides some dedicated buffer space assigned per port and per queue, with the remaining portion of the buffer space used as a dynamic shared buffer pool for all the ports. There are eight unicast queues and eight multicast queues per port for data traffic, and two control-plane queues per port for unicast and multicast control traffic.

Figure 6. Cisco Nexus 3100 Platform Buffer Block



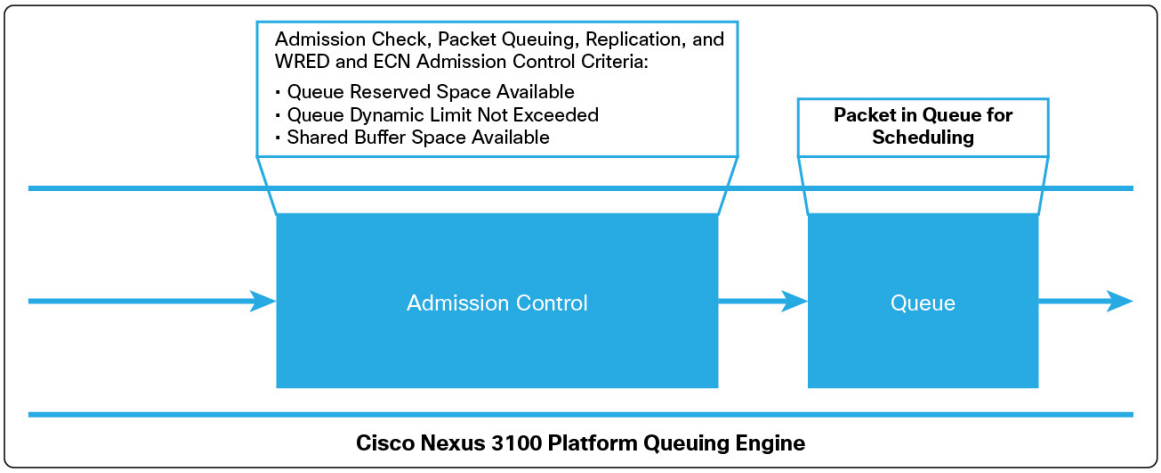
Flow Control

The Cisco Nexus 3100 platform implements both ingress and egress accounting for the packets being switched. Ingress accounting is used mainly for congestion management and flow control. As part of flow control, the queuing engine performs the admission control check (Figure 7) for all the packets entering the system. Depending on the amount of buffer space available in the queuing engine, the packet will be assigned to the reserved per-port and per-queue buffer or to the dynamic shared buffer space. If the packet buffer is not available to store the packet for the assigned priority group, a flow control message is sent on the ingress port.

Congestion Management

The Cisco Nexus 3100 platform allows congestion management mechanisms using WRED and ECN marking. Admission control monitors the egress queue size for the configured threshold. When the queue depth reaches the WRED threshold according to the drop-curve profile configured, the packet may be marked for dropping, or the packet may be randomly marked with ECN bit to notify the end host about the network congestion. ECN marking is performed so that the receiver of the packet echoes the congestion indication to the sender, which must respond as though congestion had been indicated by packet drops.

Figure 7. Packet Flow to the Queuing Engine



Packet Replication

The incoming packet needs to be replicated in the event of unknown unicast, multicast, broadcast, or Switched Port Analyzer (SPAN) traffic. The packet replication decision occurs on the egress admission control module based on the lookup result from ingress pipeline. The replication occurs in the queuing engine as the multiple descriptors referencing the packet are being placed in the egress queues for scheduling.

Buffering

Buffering occurs based on the egress port and the internal priority (QoS group) assigned to the packet and the traffic type (unicast or multicast). On egress, there are 16 queues for each port: 8 egress queues each for unicast and multicast. Buffering uses allocation from the reserved per-port and per-queue buffer space first; it then uses the dynamic shared buffer space if needed. The buffer block supports dynamic limitation of the shared buffer use per port based on the available unused buffer space cell count through dynamic thresholds. If no buffer space can be assigned to store a packet, the packet is controlled at the ingress port or dropped as a result of ingress admission control or tail dropped at the egress queue depending on the configuration.

Link Bandwidth Management

The Cisco Nexus 3100 platform implements 16 egress queues for each port, with eight system classes (QoS groups). Each system class represents one unicast and one multicast egress queue. The eight-system class (QoS group) shares the same link bandwidth, and the user can set the desired bandwidth for each egress class through the Cisco NX-OS CLI. The user can also set the scheduling weight for the unicast and multicast egress queues in the same system class. By default, 100 percent of the link bandwidth is assigned to the default class, with equal scheduling weight between the unicast and multicast queues.

Scheduling

The Cisco Nexus 3100 platform supports strict priority and deficit weighed round-robin (DWRR) scheduling for the egress queues of each port. The packet scheduler uses multilevel scheduling for the egress queues of every port. With multilevel scheduling, scheduling is performed based on the unicast and multicast queues in same class, then based on the traffic classes of the same scheduling scheme for the port (strict priority or DWRR), and then based on the traffic classes across scheduling schemes for the port. By default, only the default DWRR data-plane class is enabled along with the two strict-priority control-plane classes for each port.

Conclusion

Cisco designed the Cisco Nexus 3100 platform to extend the industry-leading versatility of the Cisco Nexus Family to provide ultra-low-latency 10 and 40 Gigabit Ethernet data center-class switches with full and robust Layer 3 support built into the switches. The Cisco Nexus 3100 platform, with high port density, forwarding-table scalability, and the Cisco NX-OS feature set that is demanded by data centers that require a cost-effective and power-efficient line-rate Layer 2 and 3 access or leaf switch.

For More Information

- Cisco Nexus 3000 Series Switches: http://www.cisco.com/go/nexus3000.
- Cisco NX-OS Software: http://www.cisco.com/go/nxos.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Printed in USA C11-729693-00 10/13