White Paper

Cisco Nexus 3064PQ Switch Architecture

What You Will Learn

The Cisco Nexus® 3064PQ Switch (Figure 1) is a high-performance, high-density, ultra-low-latency Ethernet switch that is part of the new Cisco Nexus 3000 Series Switches. This compact one-rack-unit (1RU) form factor 1 and 10 Gigabit Ethernet switch provides line-rate Layer 2 and 3 switching. The switch runs the industry-leading Cisco® NX-OS Software operating system, providing customers with robust features and capabilities that are widely deployed globally. The Cisco Nexus 3064PQ is well suited for financial co-location deployments that require support for robust unicast and multicast routing protocol features at ultra-low latencies. This document describes the architecture of the Cisco Nexus 3064PQ Switch; it provides an overview of the switch features and benefits and a detailed description of the internal architecture.

Figure 1. Cisco Nexus 3064 PQ Layer 3 Switch

Cisco Nexus 3064PQ Overview

The Cisco Nexus 3064PQ is a 1RU 1 and 10 Gigabit Ethernet Layer 3 switch that is built to provide throughput with ultra-low latency. It has 48 fixed 1 and 10 Gigabit Ethernet ports that accept modules and cables meeting the Enhanced Small Form-Factor Pluggable (SFP+) form factor. It has 4 fixed Quad SFP+ (QSFP+) ports (each QSFP+ port can handle 4 x 10 Gigabit Ethernet). The switch has a single serial console port and dual out-of-band 10/100/1000-Mbps Ethernet management ports. Two N+N redundant hot-pluggable power supplies and a unified four-fan cooling tray capable of cooling with one failed fan provide highly reliable front-to-back cooling.

All ports are at the rear of the switches, simplifying cabling and reducing cable length (Figure 2). Cooling is front to back, supporting hot- and cold-aisle configurations that help increase cooling efficiency. The front panel (Figure 3) includes status indicators and hot-swappable, N+N redundant power supplies and their power entry connections and cooling modules. All serviceable components are accessible from the front panel, allowing the switch to be serviced while in operation and without disturbing network cabling.

Figure 2. Cisco Nexus 3064PQ Rear Panel

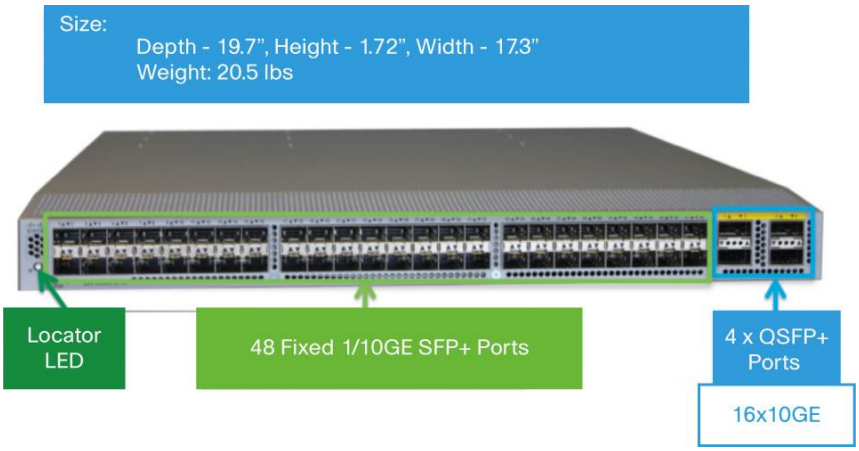


Figure 3. Cisco Nexus 3064PQ Front Panel

The Cisco Nexus 3064PQ is equipped with 4 QSFP+ ports and 4 fixed QSFP+ ports (Figure 4). Refer to the Cisco Nexus 3000 Series data sheet for information about supported optics.

Figure 4. Cisco Nexus 3064PQ Rear Right

Cisco Nexus 3064PQ Features

The Cisco Nexus 3064PQ is the first switch in the Cisco 3000 Series of high-performance, high-density, ultra-low-latency Ethernet switches providing 64 ports of 10 Gigabit Ethernet line-rate connectivity. The switch offers these features:

- High density and high availability: The Cisco Nexus 3064PQ provides 48 1/10-Gbps ports and 16 x 10-Gbps ports in 1RU. The Cisco Nexus 3000 Series is designed with redundant and hot-swappable power supply and fan modules that can be accessed from the front panel, where status lights offer an at-a-glance view of switch operation. The fan tray includes four fans and can function properly with one failed fan and one failed power supply and is hot swappable. To support efficient data center hot- and cold-aisle designs, front-to-back cooling is used for consistency with server designs. The switch supports graceful restart helper within Open Shortest Path First (OSPF), Enhanced Interior Gateway Routing Protocol (EIGRP), and Border Gateway Protocol (BGP).
- Nonblocking line-rate performance: All the 10 Gigabit Ethernet ports in the Cisco Nexus 3000 Series can handle packet flows at wire speed. The Cisco Nexus 3064PQ can have 48 Ethernet ports at 10 Gbps and 16 ports at 10 Gbps sending packets simultaneously without any effect on performance, offering true 1.28 terabits per second (Tbps) of bidirectional bandwidth, with more than 950 million packet per second (mpps).
- Ultra-low latency: The cut-through switching technology used within a switch-on-a-chip architecture of the Cisco Nexus 3000 Series enables the product to offer an ultra-low latency. The ultra-low latency on the Cisco Nexus 3000 Series together with a dedicated buffer per port and dynamic shared buffer make the Cisco Nexus 3000 Series platform the best choice for ultra-latency-sensitive environments.
- Shared buffer architecture: The Cisco Nexus 3064PQ has 9 MB of buffer space, including a per-port and dynamically allocated shared buffer.
- Separate egress queues for unicast and multicast traffic: The Cisco Nexus 3064PQ increases the number of egress groups by supporting 12 queues: 8 egress unicast and 4 egress multicast queues configurable in 8 quality-of-service (QoS) groups.
- Weighted Random Early Detection (WRED): Random Early Detection is a congestion avoidance mechanism that takes advantage of TCP's congestion control mechanism. By randomly dropping packets prior to periods of high congestion, Random Early Detection tells the packet source to decrease its transmission rate. Assuming that the packet source is using TCP, it will decrease its transmission rate until all the packets reach their destinations, indicating that the congestion is cleared. WRED is useful on any output interface where you expect to have congestion. Congestion avoidance is configurable with WRED on egress network-QoS policy maps on the Cisco Nexus 3064PQ. By default, tail-drop is the congestion control mechanism.
- Explicit congestion notification (ECN) marking: ECN is an extension to TCP/IP defined in RFC 3168. ECN allows end-to-end notification of network congestion without dropping packets and is used on WRED thresholds. Traditionally, TCP detects network congestion by observing dropped packets. When congestion is detected, the TCP sender takes action by controlling the flow of traffic. However, dropped packets can sometimes lead to long TCP timeouts and consequent loss of throughput. The Cisco Nexus 3000 Series can set a mark in the IP header, instead of dropping a packet, to signal impending congestion. The receiver of the packet echoes the congestion indicator to the sender, which must respond as though congestion had been indicated by packet drops.
- Robust Layer 3 mode: The Cisco Nexus 3000 Series can operate in Layer 3 mode without any hardware add-on. It has a comprehensive Layer 3 feature set that includes full BGP support along with many other features. For more information, see the Cisco Nexus 3000 Series 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 switch can forward multicast traffic at line rate on all 64 10 Gigabit Ethernet ports. Multicast packets are replicated in

hardware. When a Layer 2 lookup returns a hit, the packet is forwarded based on the destination MAC address in the Layer 2 multicast table. Next, if the lookup result points to an entry in the Layer 3 IP multicast table, the packet is replicated at the egress ports and VLANs referred to by the table. When the switch receives an IP multicast packet with a group address not yet learned by the switch, the link-local unknown multicast packets are flooded within the source VLAN. By default, the Cisco Nexus 3000 Series Switches will distribute (S, G) PIM joins among ECMP paths. Non-link-local unknown multicast packets will be dropped in hardware without any effect on performance.

Cisco Nexus 3064PQ Architecture

The Cisco Nexus 3000 Series control plane runs Cisco NX-OS Software on a dual-core 1.86-GHz Intel Arrandale processor with 4 GB of DDR 3 RAM. The supervisor complex is connected to the data plane in-band through two internal ports running 1-Gbps Ethernet (1 Gbps for each direction), and the system is managed in-band, or through the out-of-band 10/100/1000-Mbps management ports.

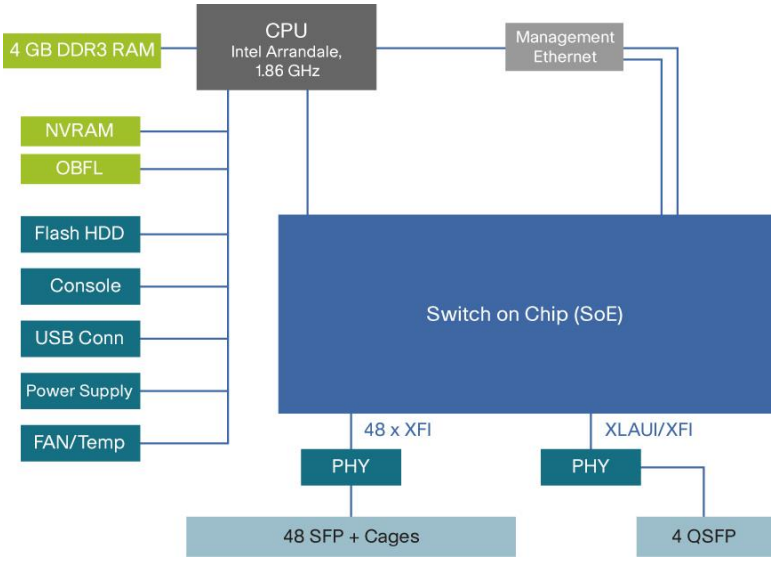
Table 1 summarizes the control-plane specifications.

Table 1. Control Plane Specifications

Component	Specification
CPU	1.86 GHz Intel Arrandale processor (dual core)
DRAM	4 GB of DD3 in 2 DIMM slots
Persistent disk	2 GB of embedded USB (eUSB) 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 upgradable and golden images
Management interface	RS-232 console port and two 10/100/1000BASE-T management ports: mgmt0 and mgmt1; 1 external USB flash port can be used for updating the system configuration

The Cisco Nexus 3000 Series is implemented with a switch-on-a-chip design. Figure 5 shows the connectivity between the components of the Cisco Nexus 3064PQ.

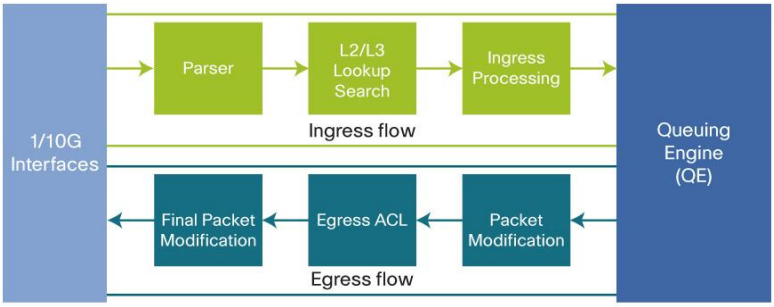
Figure 5. Cisco Nexus 3064PQ Data Plane and Switch-on-a-Chip Architecture



Cisco Nexus 3064PQ Switch-on-a-Chip Forwarding

A Queuing Engine (QE) allocates the buffer dynamically as needed (Figure 6). The buffering details are discussed in the next section.

Figure 6. Cisco Nexus 3000 Switch-on-a-Chip Packet Flow



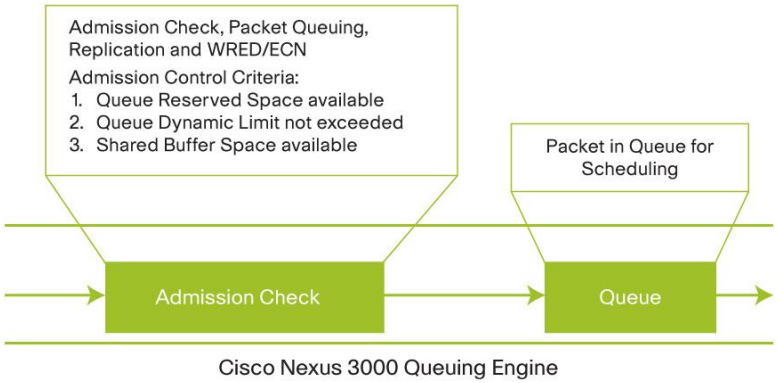
The switch-on-chip design allows a low-latency bypass mode. The ingress flow is responsible for most of the switch features such as VLAN assignment and Layer 2 and 3 table lookups. In essence, the ingress flow component makes the forwarding decision.

The parser engine parses the incoming packets and extracts the fields required. Then the Layer 2 and 3 lookup search module looks up the fields extracted by the parser. For Layer 2 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 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, the packet arrives at the Layer 3 processing engine, and the source IP address is looked up in the Layer 3 table. The destination IP address is looked up and indexed in the next-hop table, which lists the outgoing interface and the destination MAC address. The outgoing interface provides an index in the Layer 3 interface table that supplies the source MAC address and the VLAN. When the packet leaves the queuing engine, this information is used to rewrite the packet after it has been parsed; then it is forwarded out of the egress interface. The learning and aging rates and takes full advantage of the switch-on-a-chip architecture. Also, with this implementation, the 128,000-entry MAC address table will not cause any significant CPU overhead.

The packet flows receives admission checks (Figure 7). Depending on the amount of buffer space available in the Queuing Engine, the packet will be stored in the reserved per-port per-queue buffer or in the dynamic shared buffer space. These components are part of the shared buffer. The queuing, replication, or WRED and ECN process takes place in the admission control component. Then the packets are sent to the queue for scheduling. For packet replication, the decision occurs on the egress admission control module; 64 bytes of overhead will be added if replication is needed. The replication will occur in the Queuing Engine, as the packets are being placed in queues for scheduling.

Figure 7. Packet Flow to the Queuing Engine

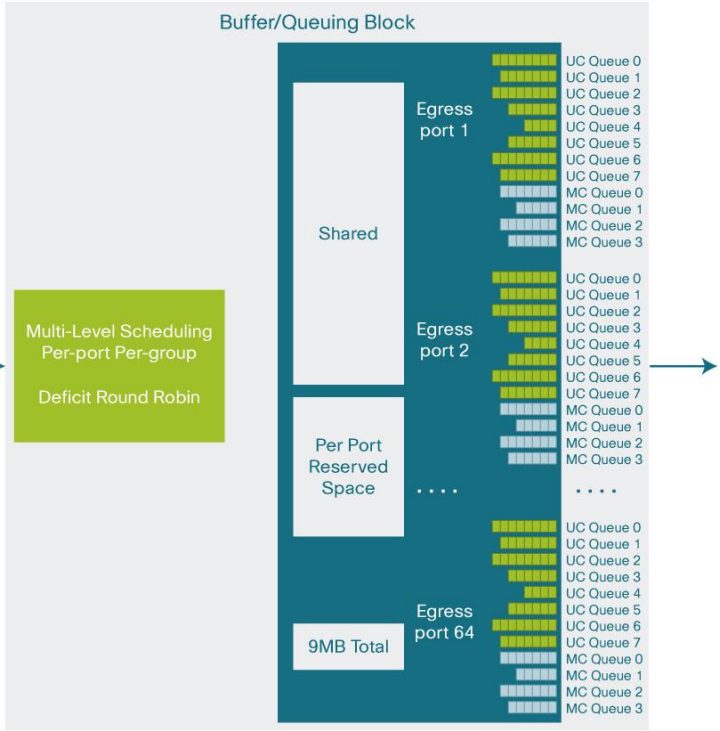


Cisco Nexus 3000 Queuing Engine

Cisco Nexus 3000 Series Queuing Engine: Buffering and QoS

Each Cisco Nexus 3000 Series interface is supplied with a buffer block that resides in the queuing engine (Figure 8). The buffer block has a per-port dedicated and a dynamic shared buffer pool, with a total of 9 MB of buffer space for the platform. There are eight unicast queues and four multicast queues per port, with multilevel scheduling per port and per-group, using a deficit round-robin mechanism. The buffer repartition is shared, and buffer limits can be set per port and per queue. The queuing engine provides ingress and egress admission control for packets. At ingress flow control and admission check, the packets are stored based on the space reserved per port, the reserved headroom, or dynamic need. At egress, packet queuing, replication, and WRED and ECN are stored based on the space reserved per port, the reserved headroom, or dynamic need.

Figure 8. Cisco Nexus 3000 Series Buffer Block



The main QoS functions consist of the following:

- Traffic classification (ingress)
- Marking: class of service (CoS) or Differentiated Services Code Point (DCSP) (ingress)
- Maximum transmission unit (MTU) checking (ingress)
- Congestion management (ingress and egress)
- Queuing and scheduling (ingress and egress)
- Link bandwidth management (egress)
- WRED (egress)
- ECN (egress)

In the Cisco Nexus 3000 Series, all data path and QoS system resources can be configured on a system-class basis. The QoS configuration uses the Cisco NX-OS command-line interface (CLI) with three QoS types: QoS, network-QoS, and queuing.

Traffic Classification

The first step in QoS processing consists of classifying the incoming packet so that it can be associated with a system class. This classification information will be carried from ingress to egress and will be used for all the QoS processing. The classification can be based on the CoS or DSCP bits of the incoming packet or on user-defined QoS access control lists (ACLs) that match Layer 2, 3, and 4 information. Alternatively the Cisco Nexus 3000 Series allows the user to set a default CoS value for an interface, and classification then is performed based on the marked IEEE 802.1p value.

CoS and DSCP Marking

The Cisco Nexus 3064PQ can mark IEEE 802.1p (CoS) bits or DSCP bits in the IP header. This function can be applied either on ingress or egress and is performed by the forwarding controller block of the unified port controller (UPC).

MTU

The Cisco Nexus 3000 Series Switches are next-generation high-performance Ethernet switches that enable multiple types of traffic: unicast and multicast at high performance with ultra-low latency, transmitted through the same interfaces. Each traffic class can have different MTU requirements in the range of 1518 to 9216 bytes. Consequently, the MTU setting needs to be per system class, and not per interface, because multiple traffic classes share the same physical interface. When operating as Layer 2 switches, the Cisco Nexus 3000 Series supports per-system-class MTU for Layer 2, and per-interface MTU for Layer 3.

Congestion Management

Each interface of the Cisco Nexus 3000 Series has 8 QoS groups, with 12 queues: 8 egress unicast queues and 4 egress multicast queues. Deficit Weighted Round Robin (DWRR) decrements the proxy queue at a programmable rate. When the proxy queue reaches a threshold that indicates 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.

Queuing and Scheduling

The Cisco Nexus 3064PQ uses an input queuing system architecture to meet high-performance and high-density requirements. The architecture implements both ingress and egress queues. The Cisco Nexus 3064PQ does not provide the same virtual output queuing (VOQ) as performed on Cisco Nexus 5000 Series Switches. Buffering occurs at the end of the ingress admission process. By the time the packet reaches the end of the ingress admission process, the egress port information is already known and resolved. The buffering occurs based on the egress port and the internal priority (QoS group) and the traffic type (unicast or multicast). This process can be compared with VOQ. On egress, there are 12 queues for each interface: 8 egress queues for unicast, and 4 egress queues for multicast.

Link Bandwidth Management

The Cisco Nexus 3064PQ implements 12 egress queues for each interface, with an egress queue corresponding to a system class for unicast or multicast. The 8 QoS groups share the same link bandwidth, and the user can set the desired bandwidth for each egress queue through the Cisco NX-OS CLI. DWRR is scheduled between the egress queues. In DWRR, each queue is assigned a scheduling weight, and the bandwidth of each port is shared according to the weight.

WRED

Random Early Detection is a congestion avoidance mechanism that takes advantage of TCP's congestion control mechanism. By randomly dropping packets prior to periods of high congestion, Random Early Detection tells the packet source to decrease its transmission rate. Assuming that the packet source is using TCP, it will decrease its transmission rate until all the packets reach their destinations, indicating that the congestion is cleared. WRED is useful on any output interface where you expect to have congestion. Congestion avoidance is configurable with WRED on egress network-QoS policy maps on the Cisco Nexus 3064PQ. By default, tail-drop is the congestion control mechanism.

ECN Marking

ECN is an extension to TCP/IP defined in RFC 3168. ECN allows end-to-end notification of network congestion without dropping packets. Traditionally, TCP detects network congestion by observing dropped packets. When congestion is detected, the TCP sender takes action by controlling the flow of traffic. However, dropped packets can sometimes lead to long TCP timeouts and consequent loss of throughput. The Cisco Nexus 3000 Series can set a mark in the IP header, instead of dropping a packet, to signal impending congestion. The receiver of the packet echoes the congestion indicator to the sender, which must respond as though congestion had been indicated by packet drops.

Conclusion

Cisco designed the Cisco Nexus 3000 Series 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 a full and robust Layer 3 built into the switches. The ASICs allow an ultra-low-latency bypass mode, while retaining a comprehensive Cisco NX-OS feature set.

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 C11-661242-01 04/11