

Cisco Nexus 5548P Switch Architecture

What You Will Learn

Cisco introduced Nexus® 5000 Series Switches both as high-bandwidth, low-latency, access-layer switches for rack deployment and as the basis for a unified network fabric that can help simplify data center infrastructure while reducing capital and operational costs.

The switch family, using cut-through architecture, supports line-rate 10 Gigabit Ethernet on all ports while maintaining consistently low latency independent of packet size and services enabled. It supports a set of network technologies known collectively as IEEE Data Center Bridging (DCB) that increases the reliability, efficiency, and scalability of Ethernet networks. These features allow the switches to support multiple traffic classes over a lossless Ethernet fabric, thus enabling consolidation of LAN, SAN, and cluster environments. Its ability to connect Fibre Channel over Ethernet (FCoE) to native Fibre Channel protects existing storage system investments while dramatically simplifying in-rack cabling.

Today Cisco announces the next-generation platforms of the Cisco Nexus 5000 Series: the Cisco Nexus 5500 platform. The Cisco Nexus 5500 platform extends the industry-leading versatility of the Cisco Nexus 5000 Series of purpose-built 10 Gigabit Ethernet data center-class switches and provides innovative advances toward higher-density, lower-latency, multilayer services. The Cisco Nexus 5500 platform is well suited for enterprise-class data center server access-layer deployments and smaller-scale, midmarket data center aggregation deployments across a diverse set of physical, virtual, storage access, and unified data center environments. The Cisco Nexus 5500 platform has the hardware capability to support Cisco® FabricPath and IETF Transparent Interconnection of Lots of Links (TRILL) to build scalable and highly available Layer 2 networks. The Cisco Nexus 5500 platform can be used as a Layer 3 switch through the addition of a routing module, enabling customers to deploy Layer 3 at the access layer.

This document describes the architecture of the Cisco Nexus 5548P Switch (Figure 1), the first platform in the Cisco Nexus 5500 Switches. It provides a brief overview of the switch features and benefits, followed by a detailed description of the internal architecture.

Figure 1. Cisco Nexus 5548P Switch



Cisco Nexus 5548P Overview

The Cisco Nexus 5548P is a one-rack-unit (1RU), 1 and 10 Gigabit Ethernet and FCoE access-layer switch built to provide 960 Gbps of throughput with very low latency. It has 32 fixed 1 and 10 Gigabit Ethernet ports that accept modules and cables meeting the Small Form-Factor Pluggable Plus (SFP+) form factor. One expansion module slot

can be configured to support up to 16 additional 1 and 10 Gigabit Ethernet ports or 8 Fibre Channel ports plus 8 1 and 10 Gigabit Ethernet ports. The switch has a single serial console port and a single out-of-band 10/100/1000-Mbps Ethernet management port. Two N+N redundant, hot-pluggable power supplies and two N+N redundant, hot-pluggable fan modules provide highly reliable front-to-back cooling.

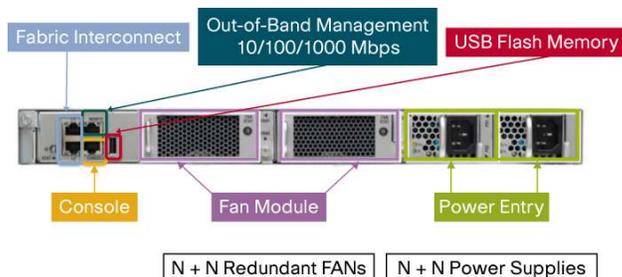
All ports are at the rear of the switches, simplifying cabling and reducing cable length (Figure 2). Cooling is front-to-back, supporting hot- and cold-aisle configurations that help increase cooling efficiency. The front panel (Figure 3) includes status indicators and hot swappable, N+N redundant power 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.

1G Support on the Nexus 5548 switch will be available in a future software release targeted for Q1CY11

Figure 2. Cisco Nexus 5548P Rear Panel



Figure 3. Cisco Nexus 5548P Front Panel



The Cisco Nexus 5500 platform is equipped to support expansion modules that can be used to increase the number of 10 Gigabit Ethernet and FCoE ports or to connect to Fibre Channel SANs with 1/2/4/8-Gbps Fibre Channel switch ports, or both. The Cisco Nexus 5548P supports one expansion module from the following offerings (Figures 4 and 5):

- Ethernet module that provides sixteen 1 and 10 Gigabit Ethernet and FCoE ports using the SFP+ interface

Figure 4. Ethernet Expansion Module



- Fibre Channel plus Ethernet module that provides eight 1 and 10 Gigabit Ethernet and FCoE ports using the SFP+ interface, and 8 ports of 1/2/4/8-Gbps native Fibre Channel connectivity using the SFP interface

Figure 5. Combination Expansion Module



Cisco Nexus 5500 Platform Features

The Cisco Nexus 5500 Series is the second generation of a family of outstanding access switches for 10 Gigabit Ethernet connectivity. The Cisco Nexus 5500 platform provides a rich feature set that makes it well suited for top-of-rack (ToR), middle-of-row (MoR), or end-of-row (EoR) access-layer applications. It protects investments in data center racks with standards-based 1 and 10 Gigabit Ethernet and FCoE features, and virtual machine awareness features that allow IT departments to consolidate networks based on their own requirements and timing. The combination of high port density, lossless Ethernet, wire-speed performance, and extremely low latency makes the switch family well suited to meet the growing demand for 10 Gigabit Ethernet that can support unified fabric in enterprise and service provider data centers, protecting enterprises' investments. The switch family has sufficient port density to support single and multiple racks fully populated with blade and rack-mount servers.

- High density and high availability: The Cisco Nexus 5548P provides 48 1/10-Gbps ports in 1RU, and the upcoming Cisco Nexus 5596 Switch provides a density of 96 1/10-Gbps ports in 2RUs. The Cisco Nexus 5500 Series is designed with redundant and hot-swappable power and fan modules that can be accessed from the front panel, where status lights offer an at-a-glance view of switch operation. To support efficient data center hot- and cold-aisle designs, front-to-back cooling is used for consistency with server designs.
- Nonblocking line-rate performance: All the 10 Gigabit Ethernet ports on the Cisco Nexus 5500 platform can handle packet flows at wire speed. The absence of resource sharing helps ensure the best performance of each port regardless of the traffic patterns on other ports. The Cisco Nexus 5548P can have 48 Ethernet ports at 10 Gbps sending packets simultaneously without any effect on performance, offering true 960-Gbps bidirectional bandwidth. The upcoming Cisco Nexus 5596 can have 96 Ethernet ports at 10 Gbps, offering true 1.92-terabits per second (Tbps) bidirectional bandwidth.
- Low latency: The cut-through switching technology used in the application-specific integrated circuits (ASICs) of the Cisco Nexus 5500 Series enables the product to offer a low latency of 2 microseconds, which remains constant regardless of the size of the packet being switched. This latency was measured on fully configured interfaces, with access control lists (ACLs), quality of service (QoS), and all other data path features turned on. The low latency on the Cisco Nexus 5500 Series together with a dedicated buffer per port and the congestion management features described next make the Cisco Nexus 5500 platform an excellent choice for latency-sensitive environments.
- Single-stage fabric: The crossbar fabric on the Cisco Nexus 5500 Series is implemented as a single-stage fabric, thus eliminating any bottleneck within the switches. Single-stage fabric means that a single crossbar fabric scheduler has full visibility into the entire system and can therefore make optimal scheduling decisions without building congestion within the switch. With a single-stage fabric, the congestion becomes exclusively a function of your network design; the switch does not contribute to it.
- Congestion management: Keeping latency low is not the only critical element for a high-performance network solution. Servers tend to generate traffic in bursts, and when too many bursts occur at the same time, a short period of congestion occurs. Depending on how the burst of congestion is smoothed out, the overall network performance can be affected. The Cisco Nexus 5500 platform offers a full portfolio of congestion management features to reduce congestion. These features, described next, address congestion at different stages and offer granular control over the performance of the network.
- Virtual output queues: The Cisco Nexus 5500 platform implements virtual output queues (VOQs) on all ingress interfaces, so that a congested egress port does not affect traffic directed to other egress ports. Every IEEE 802.1p class of service (CoS) uses a separate VOQ in the Cisco Nexus 5500 platform architecture, resulting in a total of 8 VOQs per egress on each ingress interface, or a total of 384 VOQs per ingress interface on the Cisco Nexus 5548P, and a total of 768 VOQs per ingress interface on the Cisco Nexus 5596. The extensive use of VOQs in the system helps ensure high throughput on a per-egress, per-CoS basis.

Congestion on one egress port in one CoS does not affect traffic destined for other CoSs or other egress interfaces, thus avoiding head-of-line (HOL) blocking, which would otherwise cause congestion to spread.

- Separate egress queues for unicast and multicast: Traditionally, switches support 8 egress queues per output port, each servicing one IEEE 802.1p CoS. The Cisco Nexus 5500 platform increases the number of egress queues by supporting 8 egress queues for unicast and 8 egress queues for multicast. This support allows separation of unicast and multicast that are contending for system resources within the same CoS and provides more fairness between unicast and multicast. Through configuration, the user can control the amount of egress port bandwidth for each of the 16 egress queues.
- Lossless Ethernet with priority flow control (PFC): By default, Ethernet is designed to drop packets when a switching node cannot sustain the pace of the incoming traffic. Packet drops make Ethernet very flexible in managing random traffic patterns injected into the network, but they effectively make Ethernet unreliable and push the burden of flow control and congestion management up to a higher level in the network stack.
- PFC offers point-to-point flow control of Ethernet traffic based on IEEE 802.1p CoS. With a flow-control mechanism in place, congestion does not result in drops, transforming Ethernet into a reliable medium. The CoS granularity then allows some CoSs to gain a no-drop, reliable, behavior while allowing other classes to retain traditional best-effort Ethernet behavior. The no-drop benefits are significant for any protocol that assumes reliability at the media level, such as FCoE.
- 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. 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 5500 platform 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.
- FCoE: FCoE is a standards-based encapsulation of Fibre Channel frames into Ethernet packets. By implementing FCoE and enabling a broad range of partners to terminate FCoE on the host side, the Cisco Nexus 5500 platform enables storage I/O consolidation on top of Ethernet.
- Network interface virtualization (NIV) architecture: The introduction of blade servers and server virtualization has increased the number of access-layer switches that need to be managed. In both cases, an embedded switch or soft switch requires separate management. NIV enables a central switch to create an association with the intermediate switch, whereby the intermediate switch will become the data path to the central forwarding and policy enforcement under the central switch's control. This scheme enables both a single point of management and a uniform set of features and capabilities across all access-layer switches.
- One critical implementation of NIV in the Cisco Nexus 5000 Series is the Cisco Nexus 2000 Series Fabric Extenders and their successful deployment in data centers. A Cisco Nexus 2000 Series Fabric Extender behaves as virtualized remote I/O module, enabling the Cisco Nexus 5500 platform to operate as a virtual modular chassis.
- IEEE 1588 Precision Time Protocol (PTP): In financial environments, particularly high-frequency trading environments, transactions occur in less than a millisecond. For accurate application performance monitoring and measurement, the systems supporting electronic trading applications must be synchronized with extremely high accuracy (to less than a microsecond). IEEE 1588 is designed for local systems requiring very high accuracy beyond that attainable using Network Time Protocol (NTP). The Cisco Nexus 5500 platform supports IEEE 1588 boundary clock synchronization: that is, the Cisco Nexus 5500 platform will run PTP and synchronize to an attached master clock, and the boundary clock will then act as a master clock for all attached slaves. The Cisco Nexus 5500 platform also supports packet time stamping by including the IEEE 1588 time stamp in the Encapsulated Remote Switched Port Analyzer (ERSPAN) header.

- Cisco FabricPath and TRILL: Existing Layer 2 networks based on Spanning Tree Protocol have a number of challenges to overcome: suboptimal path selection, underutilized network bandwidth, control-plane scalability, and slow convergence. Although enhancements to Spanning Tree Protocol and features such as Cisco virtual PortChannel (vPC) technology help mitigate some of these limitations, these Layer 2 networks lack fundamentals that limit their scalability.
- Cisco FabricPath and the TRILL are two emerging revolutionary solutions for creating scalable and highly available Layer 2 networks. Cisco Nexus 5500 Series' hardware is capable of switching packets based on Cisco FabricPath headers or TRILL headers. This capability enables customers to deploy scalable Layer 2 networks with native Layer 2 multipathing.
- Layer 3: The design of the access layer varies depending on whether Layer 2 or Layer 3 is used at the access layer. The access layer in the data center is typically built at Layer 2, which allows better sharing of service devices across multiple servers and allows the use of Layer 2 clustering, which requires the servers to be adjacent to Layer 2. In some design, such as two-tier designs, the access layer may be Layer 3, although this may not imply that every port on these switches is a Layer 3 port. The Cisco Nexus 5500 platform can operate in Layer 3 mode with the addition of a routing module.
- Hardware-level I/O consolidation: The Cisco Nexus 5500 platform ASICs can transparently forward Ethernet, Fibre Channel, FCoE, Cisco FabricPath, and TRILL, providing true I/O consolidation at the hardware level. The solution adopted by the Cisco Nexus 5500 platform reduces the costs of consolidation through a high level of integration in the ASICs. The result is a full-featured Ethernet switch and a full-featured Fibre Channel switch combined in one product.

Cisco Nexus 5548P Architecture

The Cisco Nexus 5548P control plane runs Cisco NX-OS Software on a dual-core 1.7-GHz Intel Xeon Processor C5500/C3500 Series with 8 GB of DRAM. The supervisor complex is connected to the data plane in-band through two internal ports running 1-Gbps Ethernet, and the system is managed in-band, or through the out-of-band 10/100/1000-Mbps management port. Table 1 summarizes the control-plane specifications.

Table 1. Cisco Nexus 5548P Control Plane Components

Component	Specification
CPU	1.7 GHz Intel Xeon Processor C5500/C3500 Series (dual core)
DRAM	8 GB of DDR3 in two DIMM slots
Program storage	2 GB of eUSB flash memory for base system storage
Boot and BIOS flash memory	8 MB to store upgradable and golden image
On-board fault log	64 MB of flash memory to store hardware-related fault and reset reasons
NVRAM	6 MB of SRAM to store syslog and licensing information
Management interface	RS-232 console port and 10/100/1000BASE-T mgmt0

The Cisco Nexus 5500 platform data plane is primarily implemented with two custom-built ASICs developed by Cisco: a set of unified port controllers (UPCs) that provides data-plane processing, and a unified crossbar fabric (UCF) that cross-connects the UPCs.

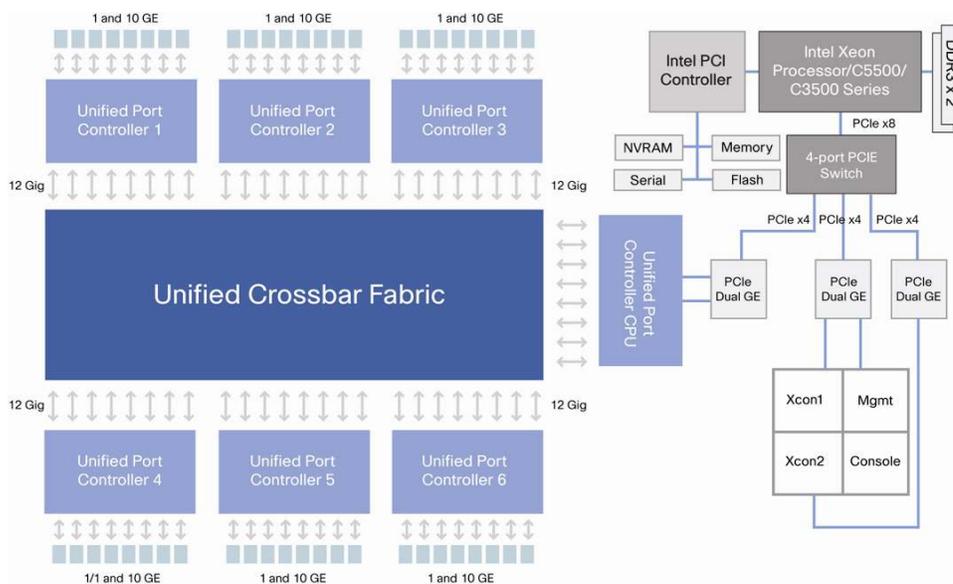
The UPC manages eight ports of 1 and 10 Gigabit Ethernet or eight ports of 1/2/4/8-Gbps Fibre Channel. It is responsible for all packet processing and forwarding on ingress and egress ports. Each port in the UPC has a dedicated data path. Each data path connects to UCF through a dedicated fabric interface at 12 Gbps. This 20 percent over-speed rate helps ensure line-rate throughput regardless of the internal packet headers imposed by the ASICs. Packets are always switched between ports of UPCs by the UCF.

The UCF is a single-stage high-performance 100-by-100 crossbar with an integrated scheduler. The scheduler

coordinates the use of the crossbar between inputs and outputs, allowing a contention-free match between I/O pairs. The scheduling algorithm is based on an enhanced iSLIP algorithm. The algorithm helps ensure high throughput, low latency, and weighted fairness across inputs, and starvation- and deadlock-free best-match policies across variable-sized packets.

The Cisco Nexus 5548P is equipped with seven UPCs: six to provide 48 ports at 10 Gbps, and one used for connectivity to the control plane. Figure 6 shows the connectivity between the control plane and the data plane.

Figure 6. Cisco Nexus 5548P Data Plane and Control Plane Architecture



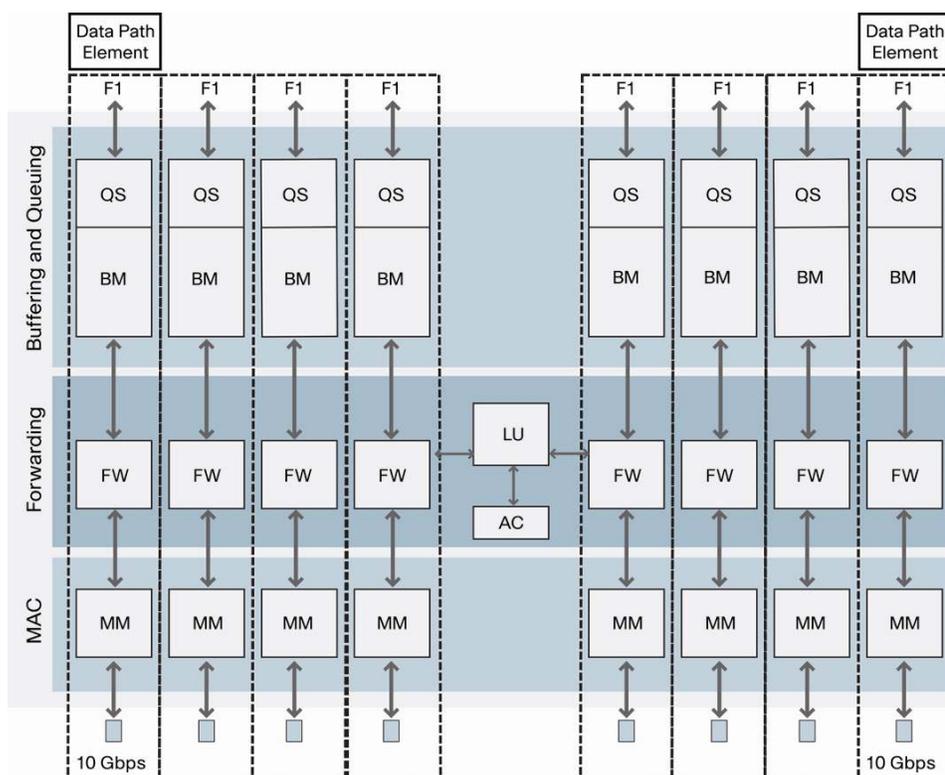
Unified Port Controller Details

The UPC has three major elements: media access control (MAC), forwarding control, and the buffering and queuing subsystem.

The multimode MAC is responsible for the network interface packet protocol and flow-control functions. It consists of encoding-decoding and synchronization functions for the physical medium, frame cyclic redundancy check (CRC), and length check. The flow-control functions are IEEE 802.3x Pause, IEEE 802.1Qbb PFC, and Fibre Channel buffer-to-buffer credit. The multimode MAC supports 1 and 10 Gigabit Ethernet and 1/2/4/8-Gbps Fibre Channel.

The forwarding controller is responsible for the parsing and rewrite function (FW), lookup (LU), and access control list (ACL). Depending on the port mode, the parsing and editing element parses packets to extract fields that pertain to forwarding and policy decisions; it buffers the packet while waiting for forwarding and policy results and then inserts, removes, and rewrites headers based on a combination of static and per-packet configuration results from the forwarding and policy decision. The lookup and ACL receive the extracted packet fields, synthesize the lookup keys, and search a series of data structures that implement Fibre Channel, Ethernet, FCoE, Cisco FabricPath, TRILL forwarding modes, QoS, and security policies.

The buffering and queuing components consists of bulk memory (BM) and the queue subsystem (QS). The bulk memory is responsible for data buffering, congestion management, flow control, policing, ECN marking, and Deficit Weighted Round-Robin (DWRR) link scheduling. Packets are sent from bulk memory to the crossbar fabric through the fabric interface (FI). The queue subsystem is responsible for managing all queues in the system. At ingress, it manages the VOQ and multicast queues. At egress, it manages the egress queues. Figure 7 shows the UPC block. Each dedicated data path element has its own components except the lookup and ACL, which are shared among data path elements within the UPC.

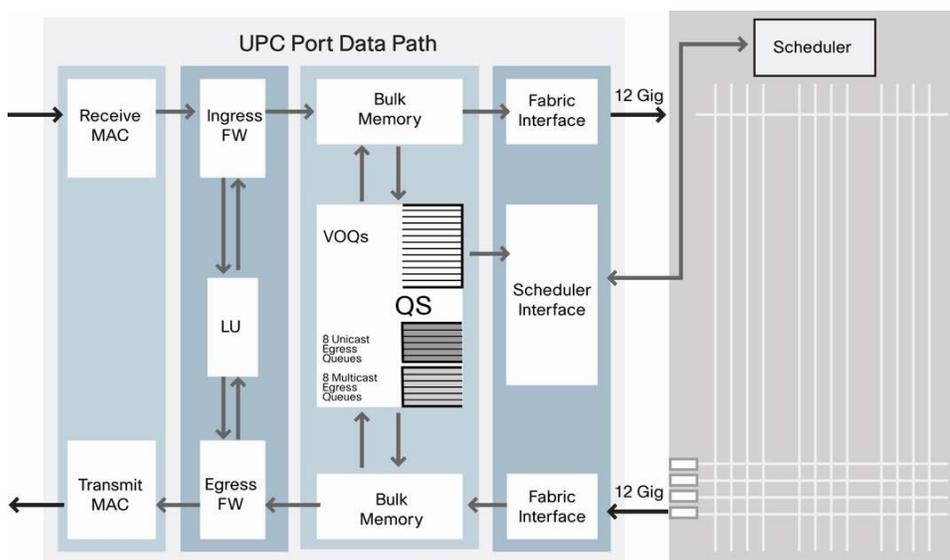
Figure 7. Unified Port Controller

On ingress, a packet received through the MAC (Figure 8) goes through the parsing and editing element that is responsible for parsing and editing fields out of the incoming packets. The parsed fields are then fed to the lookup engine in the UPC for a forwarding decision. After the forwarding decision is received, the frame is edited based on the forwarding decision result and sent to bulk memory. The parsing and editing logic understands Ethernet, IPv4 and IPv6, IP Layer 4 transports (TCP and User Datagram Protocol [UDP]), Fibre Channel, FCoE, Cisco FabricPath, and TRILL. The parsing and editing block feeds inputs to the forwarding engine as soon as the relevant frame header fields have been extracted, enabling true cut-through switching.

When a frame is present in bulk memory, the frame is queued in a unicast VOQ or multicast queue, and a request is sent to scheduler to gain access the crossbar fabric. For unicast, each VOQ represents a specific CoS for a specific egress interface, giving high flexibility to the unicast scheduler in selecting the best egress port to serve an ingress at each scheduling cycle and completely eliminating head-of-line blocking. For multicast, there are 128 queues on every ingress port; each multicast queue can be used by one or more multicast fanout. When a grant is received from the scheduler, the packet is sent through the fabric interface to the crossbar fabric.

On egress, a packet received from the crossbar fabric is sent to bulk memory through the fabric interface. The packet is queued in one of the 16 egress queues, allowing complete separation between unicast and multicast traffic even within the same CoS. The packet then goes through the same forwarding and lookup logic before it is transmitted out of the port.

Figure 8. Unified Port Controller Data Path

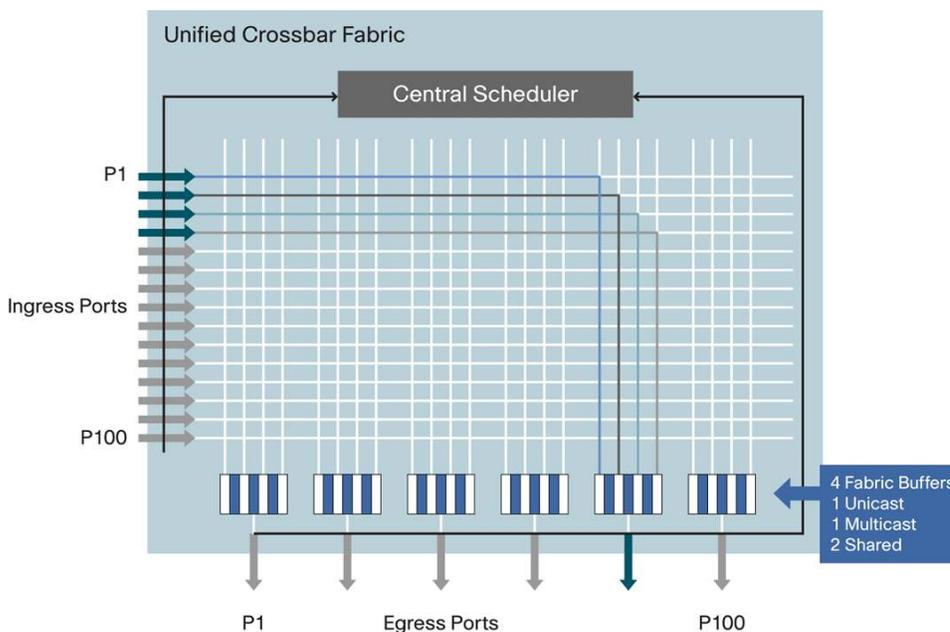


Unified Crossbar Fabric Details

The UCF is a single-stage, high-performance 100-by-100 nonblocking crossbar with an integrated scheduler (Figure 9). The single-stage fabric allows a single crossbar fabric scheduler to have full visibility into the entire system and therefore make optimal scheduling decisions without building congestion within the switch.

The crossbar provides the interconnectivity between input ports and output ports. Each row in the crossbar is associated with an input port, and each group of four columns is associated with an egress port; thus, there are four cross-points per egress port. In addition, there are four fabric buffers with 10,240 bytes of memory buffer per egress port. The four fabric buffers and four cross-points per egress interface allow four different ingress ports to simultaneously send packets to an egress port, allowing up to a 300 percent speed-up rate for unicast or multicast traffic. The four fabric buffers are shared between unicast and multicast traffic, with one reserved fabric buffer for unicast, one reserved fabric buffer for multicast, and two fabric buffers shared.

Figure 9. Unified Crossbar Fabric



The scheduler coordinates the use of the crossbar between input and output ports. The original iSLIP algorithm, which is based on iterative round-robin scheduling, has been enhanced to accommodate cut-through switching of different packet sizes. There is a separate scheduler for unicast and a separate scheduler for multicast. The scheduler uses a credit system when allocating bandwidth to each egress port. The credit system monitors fabric buffer and egress buffer use per egress port before a grant is sent to an ingress port to give access to the fabric. This approach helps ensure that the crossbar fabric is lossless, and it enables flow control to ingress ports when congestion occurs.

Cisco Nexus 5500 platform Unified Forwarding

The unified forwarding engine is a single forwarding engine implementation within each UPC capable of making forwarding decisions for Ethernet, Fibre Channel, Cisco FabricPath, and TRILL. To reduce the possibility of bottlenecks in making forwarding decisions, the unified forwarding engine is designed to use a local coherent copy of the forwarding and policy enforcement data structure in each UPC. Figure 10 summarizes the forwarding pipeline.

Figure 10. Unified Forwarding Pipeline



Interface Resolution

The first step is to map the incoming packet to an interface so that the configuration applied to the interface can take effect as the packet traverses the switch. The Cisco Nexus 5500 Series implements the concept of the logical interface (LIF). LIF is a hardware data structure that is associated with an incoming packet so that its properties, such as the forwarding setting and policy enforcement, are used in the rest of the packet processing. Each physical and virtual interface will have an associated LIF. A virtual interface is a logical entity that is used for FCoE and Cisco Nexus 2000 Series Fabric Extenders ports.

When a packet is received on a Cisco Nexus 5500 Series physical interface, the physical interface alone does not provide enough information to look up the appropriate virtual interface configuration, and therefore the physical interface information must be augmented with data parsed from the incoming packet header. Interface resolution is therefore the process of associating the incoming packet with a LIF.

Forwarding Lookup

The next step is to determine the destination of the packet. For each protocol, the appropriate forwarding table is looked up for the forwarding decision. For Ethernet, the 32,000-entry station table is used for both unicast and multicast. For unicast, the result of the lookup returns the corresponding egress interface, or a flooding index when the destination of the packet is unknown. For multicast, a flooding index is returned. The flooding index is used as input for an 8000-entry flooding index table that returns a list of outgoing interfaces. For Fibre Channel, the station table is used when the Fibre Channel ID (FC-ID) is local; otherwise, the FC-IDs are looked up in the 8000-entry Fibre Channel table for a forwarding decision. The result of the lookup can be up to 16 equal-cost forwarding paths. For Cisco FabricPath or TRILL, the switch ID or egress RBridge is looked up in the 8000-entry Layer 2 Multipathing (L2MP) table for a forwarding decision. The result of the lookup can be up to 16 equal-cost forwarding paths.

Policy Enforcement

Policy enforcement is performed with a multistage policy engine that is responsible for manipulating the forwarding results with a combination of parallel searches in memory arrays, hash tables, and ternary content-addressable memory (TCAM). The parallel search results are then evaluated and prioritized in a pipeline to create a final policy decision of ACL permit, ACL deny, QoS policing, redirect, or SPAN replication. Each TCAM located on each UPC offers 4096 matched access control entries. QoS, role-based, and SPAN ACLs have local scope and are allocated

independently on each UPC. VLAN ACLs and control-plane ACLs have global scope and are kept synchronized on all the UPCs. The policy engine evaluates the following elements:

- VLAN membership
- Interface, VLAN, and MAC binding
- MAC and Layer 3 binding (for IP and Fibre Channel)
- Fibre Channel zone membership
- Port ACLs
- VLAN ACLs
- Role-based ACLs
- QoS ACLs
- SPAN and diagnostic ACLs
- Control-plane ACLs (supervisor redirect and snooping)

Multipath Expansion

Multipath expansion is the process of choosing a physical interface when multiple paths or aggregated interfaces are present. For Fibre Channel, Cisco FabricPath, and TRILL, multiple equal forwarding paths can exist, and a path needs to be chosen for a given flow. For Ethernet, the destination interface that is returned as a result of a forwarding lookup could be an aggregated interface such as a PortChannel, and a physical port needs to be chosen per flow.

The definition of a flow changes depending on the protocol being forwarded. In Ethernet, a flow is a software-configurable selection of source and destination MAC addresses, source and destination IP addresses, and source and destination TCP and UDP ports. In FCoE and Fibre Channel, a flow is a software-configurable selection of source and destination MAC addresses, source and destination FC-IDs, and origin exchange identifiers (OX-IDs).

The Cisco Nexus 5500 platform uses multiple polynomial functions to hash a flow to obtain a numerical value that can be used to choose among up to 16 physical interfaces.

Ethernet Learning in Hardware

The last step in the forwarding process consists of MAC address learning. When an unknown source MAC address is seen for the first time by the ingress UPC's forwarding engine, the local UPC learns the MAC address in hardware. For any traffic flow involving unknown source MAC addresses, both the ingress and the egress UPC learn the MAC address in hardware, which reduces flooding for the return path. The ingress UPC subsequently generates an interrupt to the supervisor, which updates all the other UPCs.

Cisco Nexus 5500 platform Buffering and QoS

Each Cisco Nexus 5500 platform interface is supplied with a dedicated pool of 640 KB distributed by the QoS subsystem among eight CoSs (called system classes in the QoS command-line interface [CLI]). Defined in the IEEE 802.1Q tag by the IEEE 802.1p bits, each CoS can have an independent QoS policy configured through Cisco NX-OS, and the QoS subsystem's goal is to help ensure the best possible throughput to each class within the constraints defined by each policy.

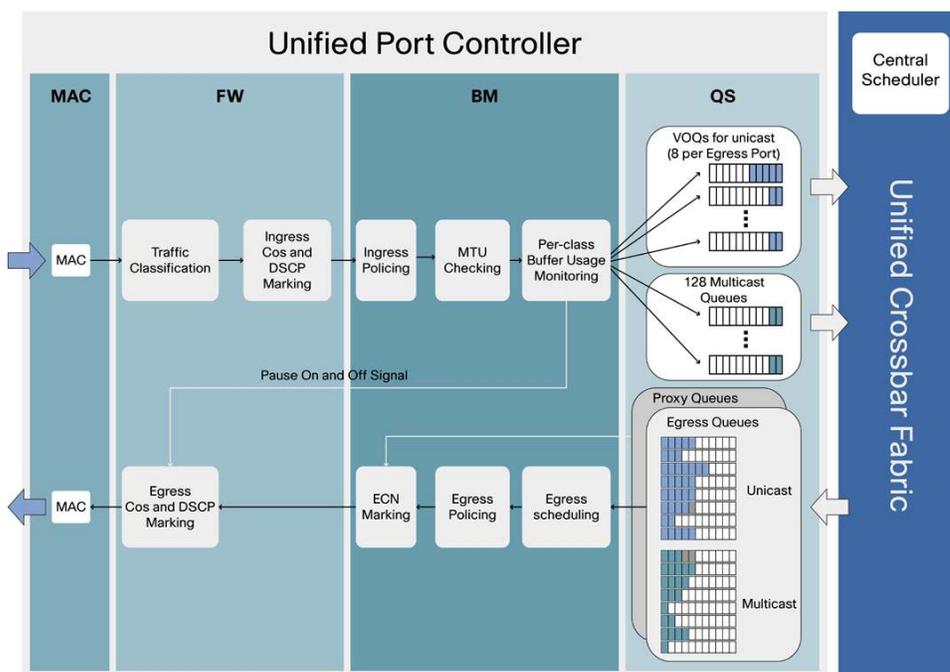
The buffering strategy on the UPC includes ingress and egress buffers from the pool of 640 KB of memory. Ingress buffering constitutes the majority of the buffering needs, and therefore most buffers are assigned to the ingress side; egress buffering is used mainly to sustain flow control for both Ethernet and Fibre Channel and to create an egress pipeline to increase throughput.

In the Cisco Nexus 5500 platform, all data path and QoS system resources can be configured on a system-class basis. Each system class has an independent buffer allocation, independent maximum transmission unit (MTU) setting, independent drop behavior, independent ingress VOQs, independent egress queues, and independent congestion management. Figure 11 summarizes the QoS processing flow.

The main QoS functions consist of the following:

- Traffic classification (ingress)
- Marking (ingress and egress)
- Policing (ingress and egress)
- MTU checking (ingress)
- Congestion management (ingress and egress)
- Queuing (ingress and egress)
- Link bandwidth management (egress)

Figure 11. Cisco Nexus 5500 platform's QoS Flow



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 CoS or Differentiated Services Code Point (DSCP) bits of the incoming packet or on user-defined QoS ACLs that match Layer 2, 3, and 4 information. Alternatively the Cisco Nexus 5500 platform 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

Cisco Nexus 5500 platform can mark IEEE 802.1p (CoS) bits or DSCP bits in the IP header. This function can be applied either on ingress or on egress and is performed by the forwarding controller block of the UPC.

Policing

Traffic policing is the process of monitoring network traffic for compliance with a service-level agreement (SLA) and taking steps to enforce that contract. Traffic exceeding a traffic contract can be discarded, marked as noncompliant, or left as is, depending on administrative policy and the characteristics of the excess traffic. The Cisco Nexus 5500 platform implements policing using the token-bucket mechanism, which uses a burst size, mean rate, and time interval.

Each token bucket has a counter (Tc) that keeps track of available tokens for the bucket. Tc is initially set to the burst size (Bc), and tokens are added at a configurable committed information rate (CIR) per time interval. When a packet is received, the packet is allowed to proceed if the token bucket is not empty ($Tc > 0$). The token counter is then decremented by the packet length. When the token bucket is empty ($Tc \leq 0$), the packet will be dropped or marked until the token becomes available again. The Cisco Nexus 5500 platform provides the policing function for QoS, control-plane redirection, storm control, and SPAN.

MTU Checking

Cisco Nexus 5500 platform are next-generation, high-performance unified fabric switches that enable I/O consolidation when multiple types of traffic, such as Ethernet, FCoE, and high-performance computing traffic, are transmitted through the same physical interface. Each traffic class can have different MTU requirements: for example, the MTU for FCoE can be 2112 bytes, and the MTU for Ethernet can be in the range of 1518 to 9216 bytes. Consequently, the MTU setting needs to be per system class and not per interface since multiple traffic classes share the same physical interface. When operating as Layer 2 switches, Cisco Nexus 5500 platform support per-system class MTU, but not per-interface MTU.

After the packet is associated with a system class, the UPC checks the MTU of the received packet against the MTU of the corresponding configured system class. In the cut-through switching mode, packets that exceed the MTU value are truncated. In the store and forwarding mode, MTU violations lead to packet drop.

Congestion Management

On ingress, the congestion management module in the UPC monitors the buffer use for each system class. For the drop system class, the buffer management module starts to drop the packets when the maximum buffer size allocated for the system class is reached. For the no-drop system class, a pause signal is sent to the MAC when buffer use reaches a threshold (that is lower than the total allocated buffer size); the MAC will send a PFC frame with the priority-enable vector set for this class. When the buffer use is below the threshold, the congestion management module removes the pause signal, and the MAC then generates PFC frame with the priority-enable vector cleared, indicates to the sender that transmission can be resumed for this system class.

On egress, the congestion management module uses a proxy queue to measure the average egress queue. Each interface of the Cisco Nexus 5500 platform has 16 egress queues: 8 egress unicast queues and 8 egress multicast queues. A proxy queue is associated to each of 16 egress queues. The proxy queue is a packet-length counter, and it is incremented by the frame length every time a packet is sent to bulk memory from the fabric interface. 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

The Cisco Nexus 5500 platform uses input queuing system architecture to meet high-performance and high-density requirements. The architecture implements both ingress and egress queues. On ingress, VOQs are used to eliminate head-of-line blocking so that a congested egress port or congested CoS does not affect any other CoS or egress port. For each ingress interface of the Cisco Nexus 5500 platform, there are 8 VOQs per egress port, for a total of 384 VOQs on each ingress interface of the Cisco Nexus 5548P. Multicast uses separate queues, and there

are 128 ingress queues for multicast, so each ingress multicast queue can be used by one or more multicast fanouts.

On egress, there are 16 queues for each interface: 8 egress queues for unicast, and 8 egress queues for multicast. The Cisco Nexus 5500 platform takes the traditional per-class egress queues one step further and allows separation for unicast and multicast within the same system class. This approach allows more fairness and granularity when both unicast and multicast contend for system resources.

Link Bandwidth Management

The Cisco Nexus 5500 platform implements 16 egress queues for each interface, with an egress queue corresponding to a system class for unicast or multicast. The 16 queues share the same link bandwidth, and the user can set the desired bandwidth for each egress queue through the Cisco NX-OS CLI. One of the 16 queues can be configured for strict-priority queuing; otherwise DWRR is scheduled between the egress queues. The strict-priority queue, if configured, is always scheduled and serviced first; when the strict-priority queue is empty, the rest of the queues are scheduled using DWRR. In DWRR, each queue is assigned a scheduling weight, and the bandwidth of each port is shared according to the weight.

Conclusion

Cisco designed the Cisco Nexus 5500 Series to extend the industry-leading versatility of the Cisco Nexus 5000 Series high-performance, low-latency, 10 Gigabit Ethernet data center-class switch. The Cisco Nexus 5500 platform provides innovative advances toward higher density, lower latency, and multilayer services. Its ASICs are built to transparently forward Ethernet, Fibre Channel, FCoE, Cisco FabricPath, and TRILL, providing true I/O consolidation at the hardware level. The Cisco Nexus 5500 platform enables customers to deploy a next-generation data center network with unified network fabric that can help simplify data center infrastructure, reducing capital and operational costs.

For More Information

- Cisco Nexus 5000 Series Switches: <http://www.cisco.com/go/nexus5000>
- Cisco Nexus 5500 Series Switches: <http://www.cisco.com/go/nexus5000>
- Cisco Nexus 2000 Series Fabric Extenders: <http://www.cisco.com/go/nexus2000>
- 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.

Cisco and the Cisco Logo are trademarks of Cisco Systems, Inc. and/or its affiliates in the U.S. and other countries. A listing of Cisco's trademarks can be found at www.cisco.com/go/trademarks. Third party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1005R)