Cisco Nexus 5000 Series Architecture: The Building Blocks of the Unified Fabric

What You Will Learn

Multicore computing and virtualization are rapidly changing the data center landscape, furthering the need for high-bandwidth, low-latency switching. These technologies increase efficiency by increasing server utilization, but they also promote an ever-increasing demand for bandwidth. Most data centers grappling with the bandwidth challenge are migrating to 10 Gigabit Ethernet to alleviate their IP network bottlenecks. In addition, most data centers support dual Fibre Channel links per server to access their storage networks, and some data centers supporting high-performance computing (HPC) environments also support multiple interprocess communication (IPC) networks per server.

Cisco® offers a better solution to these challenges in the form of the Cisco® Nexus 5000 Series Switches. Designed as access-layer switches for in-rack deployment, the Cisco Nexus 5000 Series helps simplify data center infrastructure and reduce total cost of ownership (TCO). It supports I/O consolidation at the rack level, reducing the number of adapters, cables, switches, and transceivers that each server must support, all while protecting investment in existing storage assets.

The Cisco Nexus 5000 Series delivers these benefits to data centers through the following product features:

- **High performance 10 Gigabit Ethernet:** The Cisco Nexus 5000 Series is a family of line-rate, low-latency, cost-effective 10 Gigabit switches designed for access-layer applications.
- **Fibre Channel over Ethernet (FCoE):** The Cisco Nexus 5000 Series is the first open-standards-based access-layer switch to support I/O consolidation at the rack level through FCoE.
- **IEEE Data Center Bridging (DCB):** The switch family incorporates a series of Ethernet enhancements designed for the data center, including flow control and network congestion management.
- **VM Optimized Services:** The switch family supports end-port virtualization and virtual machine optimized services, helping increase the scalability of virtual Layer 2 networks and enhancing application performance and security.

This document describes how Cisco has designed the Cisco Nexus 5000 Series Switches as both 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. This document provides a brief overview of the switch features and benefits and then details the series' 10 Gigabit Ethernet, I/O consolidation, and virtualization capabilities. Internally, the switches are based on only two custom application-specific integrated circuits (ASICs): a unified port controller that handles all packet-processing operations on ingress and egress, and a unified crossbar fabric that schedules and switches packets. Every design decision made in these two devices is precisely targeted to support I/O consolidation and virtualization features with the most efficient use of transistor logic, helping minimize power consumption and maximize performance.

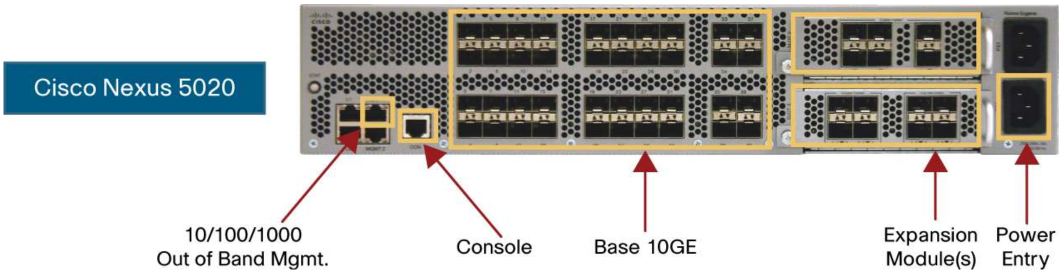
Introducing the Cisco Nexus 5000 Series

The Cisco Nexus 5000 Series is designed to be deployed in server racks, and the series is designed much like the servers it supports. All ports and power entry connections are at the rear of the switches, simplifying cabling and minimizing cable length (Figure 1). Cooling is front-to-back, supporting hot- and cold-aisle configurations that help increase cooling efficiency. The front panel includes status indicators and hot-swappable, N+1 redundant power supplies 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. The switch family's port density is such that, depending on the switch model and server rack configuration, switches can support top-of-rack, adjacent-rack, and end-of-row configurations.

Cisco Nexus 5020 56-Port Switch

The Cisco® Nexus 5020 is a two-rack-unit (2RU), 10 Gigabit Ethernet and FCoE access-layer switch built to provide 1.04 Tbps of throughput with very low latency. It has 40 fixed 10 Gigabit Ethernet/FCoE ports that accept modules and cables meeting the Small Form-Factor Pluggable Plus (SFP+) form factor. Two expansion module slots can be configured to support up to 12 additional 10 Gigabit Ethernet/FCoE ports, up to 16 Fibre Channel ports, or a combination of both. The switch has a single serial console port and a single out-of-band 10/100/1000-Mbps Ethernet management port. Two N+1 redundant, hot-pluggable power supplies and five N+1 redundant, hot-pluggable fan modules provide highly reliable front-to-back cooling.

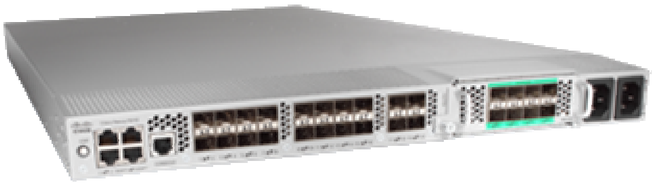
Figure 1. Cisco Nexus 5020 Rear Port Configuration



Cisco Nexus 5010 28-Port Switch

The Cisco Nexus 5010 Switch is a 1RU, 10 Gigabit Ethernet/FCoE access-layer switch built to provide more than 500 Gigabits per second (Gbps) throughput with very low latency (Figure 2). It has 20 fixed 10 Gigabit Ethernet/FCoE 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 6 additional 10 Gigabit Ethernet/FCoE ports, up to 8 Fibre Channel ports, or a combination of both. The switch has a single serial console port and a single out-of-band 10/100/1000-Mbps Ethernet management port. Two N+1 redundant, hot-pluggable power supplies and five N+1 redundant, hot-pluggable fan modules provide highly reliable front-to-back cooling.

Figure 2. Cisco Nexus 5010, Supporting 20 Fixed 10 Gigabit Ethernet/FCoE Ports and One Expansion Module Slot

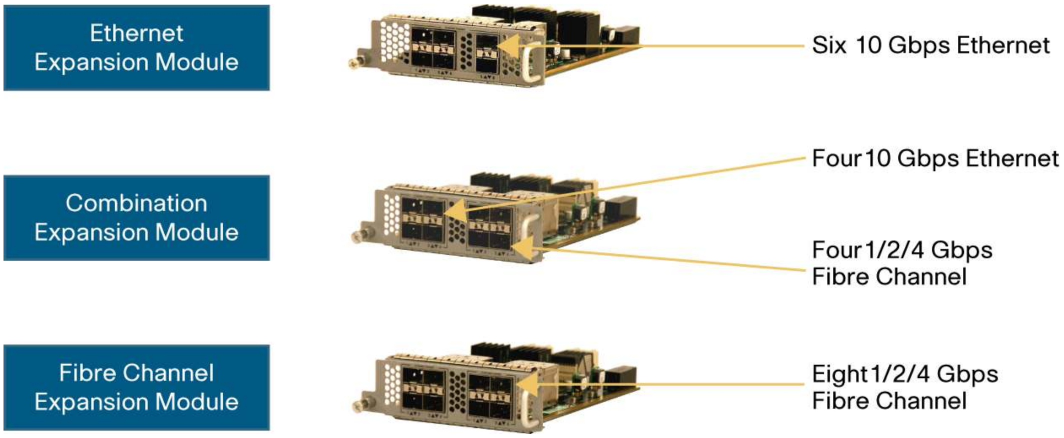


Expansion Module Options

The Cisco Nexus 5000 Series is equipped to support three expansion module options that can be used to increase the number of 10 Gigabit Ethernet/FCoE ports, connect to Fibre Channel SANs, or both. The Cisco Nexus 5010 supports a single module, with the Cisco Nexus 5020 supporting any combination of two modules that allow the switches to be optimized for specific data center applications (Figure 3):

- A 10 Gigabit Ethernet module provides an additional six 10 Gigabit Ethernet/FCoE SFP+ ports per module, helping the switch support even denser server configurations.
- A Fibre Channel module provides 8 ports of 1-, 2-, or 4-Gbps Fibre Channel through SFP ports for transparent connectivity with existing Fibre Channel networks, ideal in environments where storage I/O consolidation is the main focus.
- A combination Fibre Channel and Ethernet module provides 4 10 Gigabit Ethernet/ FCoE ports through SFP+ ports and 4 ports of 1-, 2-, or 4-Gbps native Fibre Channel connectivity through SFP ports.

Figure 3. Three Expansion Module Options Allow the Cisco Nexus 5000 Series to be Optimized for Specific Data Center Applications


Ethernet Expansion Module
Six 10 Gbps Ethernet
Combination Expansion Module
Four10 Gbps Ethernet
Four1/2/4 Gbps Fibre Channel
Fibre Channel Expansion Module
Eight1/2/4 Gbps Fibre Channel

Cisco Nexus 5000 Series Feature Highlights

Features and Benefits

The switch family's rich feature set makes the series ideal for rack-level, access-layer applications. It protects investments in data center racks with standards based Ethernet and FCoE features that allow IT departments to consolidate networks based on their own requirements and timing.

- The combination of high port density, wire-speed performance, and extremely low latency makes the switch an ideal product to meet the growing demand for 10 Gigabit Ethernet at the rack level. The switch family has sufficient port density to support single or multiple racks fully populated with blade and rack-mount servers.
- Built for today’s data centers, the switches are designed just like the servers they support. Ports and power connections are at the rear, closer to server ports, helping keep cable lengths as short and efficient as possible. Hot-swappable power and cooling modules can be accessed from the front panel, where status lights offer an at-a-glance view of switch operation. Front-to-back cooling is consistent with server designs, supporting efficient data center hot- and cold-aisle designs. Serviceability is enhanced with all customer-replaceable units accessible from the front panel. The use of SFP+ ports offers increased flexibility to use a range of interconnect solutions, including copper for short runs and fiber for long runs.
- Fibre Channel over Ethernet and IEEE Data Center Bridging features supports I/O consolidation, eases management of multiple traffic flows, and optimizes performance. Although implementing SAN consolidation requires only the lossless fabric provided by the Ethernet pause mechanism, the Cisco Nexus 5000 Series provides additional features that create an even more easily managed, high-performance, unified network fabric.

10 Gigabit Ethernet and Unified Fabric Features

The Cisco Nexus 5000 Series is first and foremost a family of outstanding access switches for 10 Gigabit Ethernet connectivity. Most of the features on the switches are designed for high performance with 10 Gigabit Ethernet. The Cisco Nexus 5000 Series also supports FCoE on each 10 Gigabit Ethernet port that can be used to implement a unified data center fabric, consolidating LAN, SAN, and server clustering traffic.

Nonblocking Line-Rate Performance

All the 10 Gigabit Ethernet ports on the Cisco Nexus 5000 Series Switches 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 5020 can have 52 Ethernet ports at 10 Gbps sending packets simultaneously without any effect on performance, offering true 1.04-Tbps bidirectional bandwidth.

Single-Stage Fabric

The crossbar fabric on the Cisco Nexus 5000 Series Switches is implemented as a single-stage fabric, thus eliminating any bottleneck within the switch. Single-stage fabric means that a single crossbar fabric scheduler has full visibility of the entire system and can therefore make optimal scheduling decisions without building congestion within the switch. With a single-stage fabric, the bandwidth you see is the bandwidth you get, and congestion becomes exclusively a function of your network design; the switch does not contribute to it.

Low Latency

The cut-through switching technology used in the Cisco Nexus 5000 Series ASICs enables the product to offer a low latency of 3.2 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 5000 Series enables application-to-application latency on the order of 10 microseconds (depending on the network interface card [NIC]). These numbers, together with the congestion management features described next, make the Cisco Nexus 5000 Series a great choice for latency-sensitive environments.

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 5000 Series offers a full portfolio of congestion management features to minimize congestion. These features, described next, address congestion at different stages and offer maximum granularity of control over the performance of the network.

Virtual Output Queues

The Cisco Nexus 5000 Series implements virtual output queues (VOQs) on all ingress interfaces, so that a congested egress port does not affect traffic directed to other egress ports. But virtual output queuing does not stop there: every IEEE 802.1p class of service (CoS) uses a separate VOQ in the Cisco Nexus 5000 Series architecture, resulting in a total of 8 VOQs per egress on each ingress interface, or a total of 416 VOQs on each ingress interface. The extensive use of VOQs in the system helps ensure maximum 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.

Lossless Ethernet (Priority Flow Control)

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 at a higher level in the network stack.

IEEE 802.1Qbb Priority Flow Control (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. A networking device implementing PFC makes an implicit agreement with the other end of the wire: any accepted packet will be delivered to the next hop and never be locally dropped. To keep this promise, the device must signal the peer when no more packets can reliably be accepted, and that, essentially, is the flow control function performed by PFC. The benefits are significant for any protocol that assumes reliability at the media level, such as FCoE.

Delayed Drop

Traditional Ethernet is unreliable, and the only way to postpone packet drops in case of congestion is to increase the buffering capabilities of the interfaces. With more buffers, short-lived congestion can be handled without causing any packet drops, and the regular drop behavior takes over if the congestion lasts longer. Tuning the amount of buffer space available effectively means tuning the definition of “short-lived congestion.”

PFC changes the equation by pushing back the buffering requirements to the source. PFC works very well for protocols like FCoE that require a reliable medium, but it makes short-lived congestion and persistent congestion undistinguishable.

Delayed drop mediates between traditional Ethernet behavior and PFC behavior. With delayed drop, a CoS can be flow controlled and the duration of the congestion monitored, so that the traditional drop behavior follows if the congestion is not resolved. Delayed drop offers the capability to tune the definition of “short-lived congestion” with PCF, hence removing the need to increase physical buffers on the interfaces.

Fibre Channel over Ethernet

FCoE is a standard-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 5000 Series introduces storage I/O consolidation on top of Ethernet.

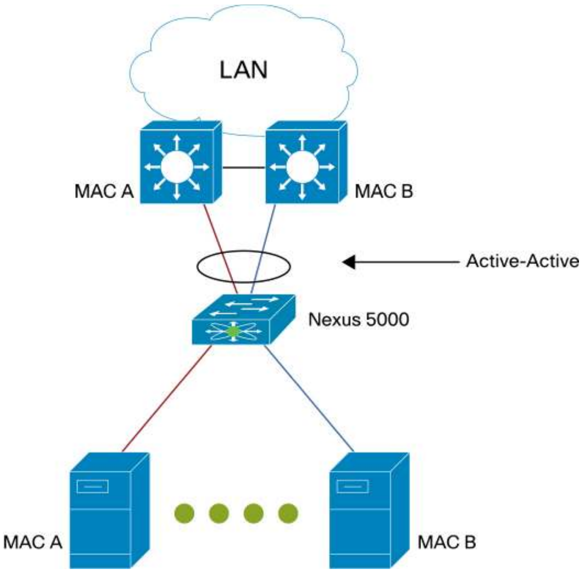
Hardware-Level I/O Consolidation

The Cisco Nexus 5000 Series ASICs can transparently forward Ethernet, Fibre Channel, and FCoE, providing true I/O consolidation at the hardware level. The solution adopted by the Cisco Nexus 5000 Series minimizes 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.

End-Port Virtualization

- **Ethernet:** Ethernet host virtualizer (EHV): In most network designs, access switches are attached to multiple distribution switches for high-availability purposes. The physically redundant paths are not all active in the loop-free logical topology created by the Spanning Tree Protocol, and that affects the amount of active bandwidth available to the LAN core. Using EHV, the default switching behavior can be changed in the Cisco Nexus 5000 Series and replaced by a different loop-prevention scheme at the access layer. EHV allows the switch to behave like a giant end-host for the network, representing all the hosts (servers) directly attached to it (Figure 4). Because of this behavior, EHV is completely transparent to the rest of the network and shrinks the Spanning Tree domain one level up to the distribution layer, giving full access to all the bandwidth physically available between the access and distribution layers.

Figure 4. Ethernet Host Virtualizer (EHV)



- **Fibre Channel:** N_port virtualization (NPV): Because of the use of hierarchically structured addresses (Fibre Channel IDs [FC_IDs]), Fibre Channel switches can offer L2MP, thus resolving the forwarding limitations of the Spanning Tree Protocol in Ethernet. However, the fixed address structure limits the scalability of a Fibre Channel fabric to a maximum of 239 switches, constraining the network design choices available to SAN architects. The Cisco Nexus 5000 Series frees the SAN of these constraints by offering the option to run the switch in NPV mode. When NPV mode is enabled on the Cisco Nexus 5000 Series, the switch becomes a transparent proxy that does not participate in the SAN fabric services, and it can aggregate all the directly attached initiators and targets directed toward the SAN fabric core as if it were a simple multipoint link. Used in conjunction with NPIV on the perimeter of the SAN fabric, NPV is a powerful tool for scaling the SAN beyond the port density of traditional Fibre Channel switches.

Cisco Nexus 5000 Series Internal Architecture

Supervisor Architecture

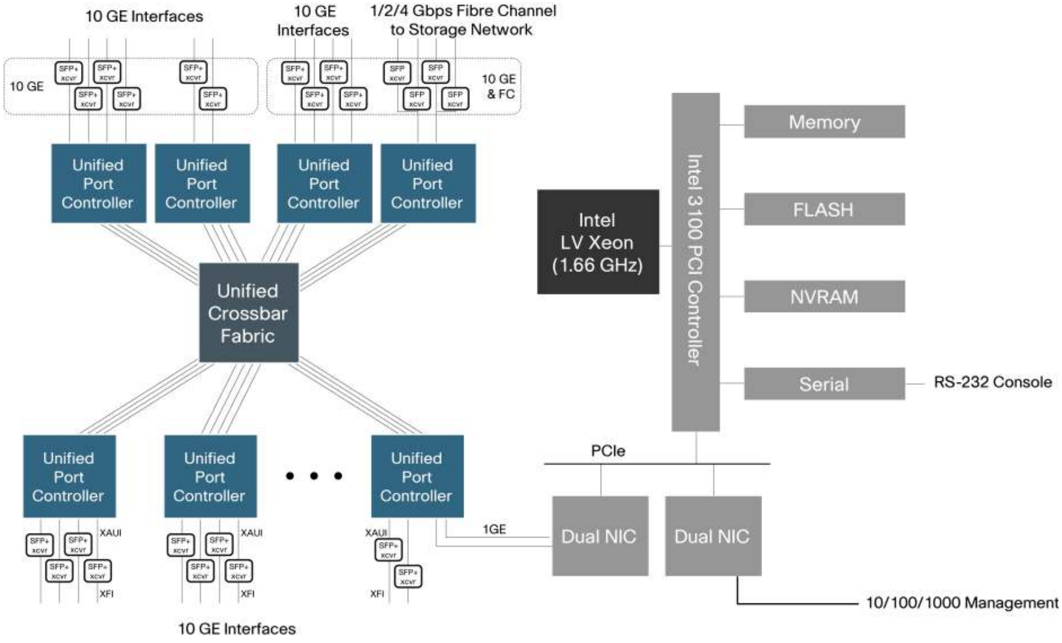
On the control plane side, the Cisco Nexus 5000 Series runs Cisco® NX-OS on a single-core 1.66-GHz Intel LV Xeon CPU with 2 GB of DRAM. The supervisor complex is connected to the data plane in-band through 2 internal ports running 1-Gbps Ethernet, and the system is managed in-band, or through the out-of-band 10/100/1000-Mbps management port. Table 1 summarizes the architecture specifications.

Table 1. Cisco Nexus 5000 Series Architecture

Item	Specification
CPU	1.66-GHz Intel LV Xeon: LF80538KF0281M
Memory	2 GB of DDR2 400 (PC2 3200) in 2 DIMM slots
Boot flash memory	1 GB of USB-based (NAND) flash memory
BIOS	2 MB of EEPROM with locked recovery image
Onboard fault log (OBFL)	64 MB of flash memory for failure analyses, kernel stack traces, boot records, and fault logs
NVRAM	2 MB of SRAM: Syslog and licensing information

Data Plane

Figure 5. Supervisor and Data Plane Architecture



The Cisco Nexus 5000 Series uses a scalable cut-through input queuing switching architecture. The architecture is implemented primarily by two ASICs developed by Cisco:

- A set of unified port controllers (UPCs) that perform data plane processing
- A unified crossbar fabric (UCF) that cross-connects the UPCs

Each UPC manages 4 10 Gigabit Ethernet/FCoE ports and makes forwarding decisions for the packets received on those ports. After a forwarding decision is made, the packets are queued in VOQs, waiting to be granted access to the UCF. (Because of the cut-through characteristics of the architecture, packets are queued and dequeued before the full packet contents have been received and buffered on the ingress port.) The UCF is responsible for coupling ingress UPCs to available egress UPCs, and it internally connects each 10 Gigabit Ethernet/FCoE interface through fabric interfaces running at 12 Gbps. This 20 percent over speed helps ensure line-rate throughput regardless of the packet manipulation performed in the ASICs.

The Cisco Nexus 5020 is equipped with 14 UPCs, giving it a total of 56 available interfaces at 10 Gbps; 52 of these interfaces are wired to actual ports on the chassis back panel, 2 are used for supervisor CPUs in-band connectivity, and the remaining 2 are currently unused. A single UCF is a 58-by-58 single-stage crossbar switch, and it is therefore sufficient to support all 56 internal fabric interfaces from the 14 UPCs (Figure 5).

Unified Port Controller

The UPC handles all packet processing operations within the Cisco Nexus 5000 Series server switch. It is an L2MP device with the capability to operate simultaneously and at wire speed with the following protocols:

- Classical Ethernet
- Fibre Channel
- FCoE

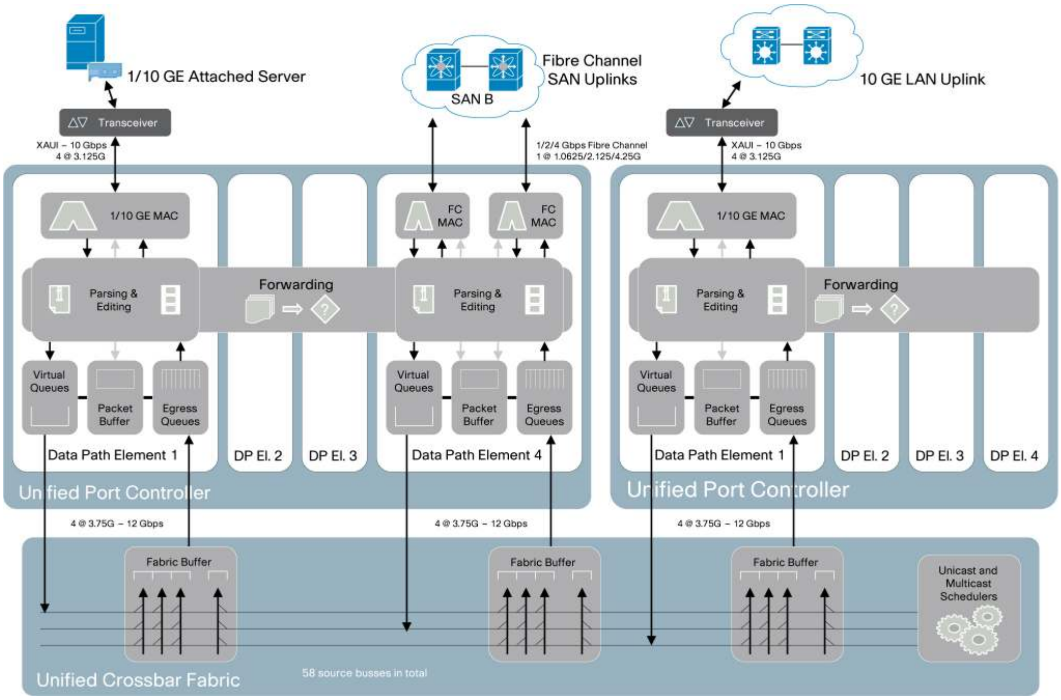
On the ingress side, it handles the physical details of different media, maps the received packets to a unified internal packet format, and makes forwarding decisions based on protocol-specific forwarding tables stored locally in the

ASIC. On the egress side, it remaps the unified internal format to the format supported by the egress medium and Layer 2 protocol and transmits the packet.

Each external-facing 10-Gbps interface on a UPC can be wired to serve as two Fibre Channel interfaces at 1/2/4 Gbps for an expansion module, and therefore a single UPC can connect up to 8 Fibre Channel interfaces through expansion modules.

As Figure 6 shows, the UPC ASIC is partitioned into four data path elements, one for each 10 Gigabit Ethernet interface. Most of the resources on the UPC are physically assigned on a per-data-path-element basis, with the exception of the forwarding logic, which is shared by the four elements.

Figure 6. UPC ASIC Architecture



At the front of each data path element are the four Media Access Controllers (MACs) needed to support Ethernet and Fibre Channel convergence in the ASIC, each with integrated flow control handling:

- One Gigabit Ethernet MAC (with flow control based on IEEE 802.3X pause and PFC)
- One 10 Gigabit Ethernet MAC (with flow control based on 802.3X pause and PFC)
- Two 1/2/4-Gbps Fibre Channel MAC (with flow control based on buffer-to-buffer credit).

The parsing and editing block is responsible for parsing fields out of the incoming packets. The parsed fields are then fed to the forwarding engine in the UPC for a forwarding decision. They are also used for the encapsulation and decapsulation of packets, both by adding or removing internal headers and by FCoE–to–Fibre Channel translation. The parsing and editing logic understands Ethernet, IPv4 and IPv6, IP Layer 4 transports (TCP and UDP), Fibre Channel and FCoE. 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.

The cut-through technology implemented in the UPC enables packets destined for free egress ports to flow immediately through the UCF and out, without the need to be fully buffered anywhere in between. Under these circumstances, the switch can produce the first bit of a packet on the egress interfaces just 3.2 microseconds after the first bit of the packet was received by the ingress interface (tested with SFP+ copper transceivers), and the 3.2-microsecond latency does not change regardless of the overall packet size.

Each interface is supplied with a dedicated pool of 480 KB of ECC-protected SRAM, distributed by the QoS subsystem among eight CoS (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 maximum 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 480 KB 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.

On the ingress side, each data path element is equipped with a VOQ for each port and system class, as well as a multicast queue for each system class. Each unicast VOQ represents a specific CoS for a specific egress interface, giving maximum flexibility to the UCF unicast scheduler in selecting the best egress port to serve an ingress at each scheduling cycle and completely eliminating head-of-line blocking.

On the egress side, each interface uses a queue for each system class to prevent flow control in one CoS from affecting the performance of the other CoSs.

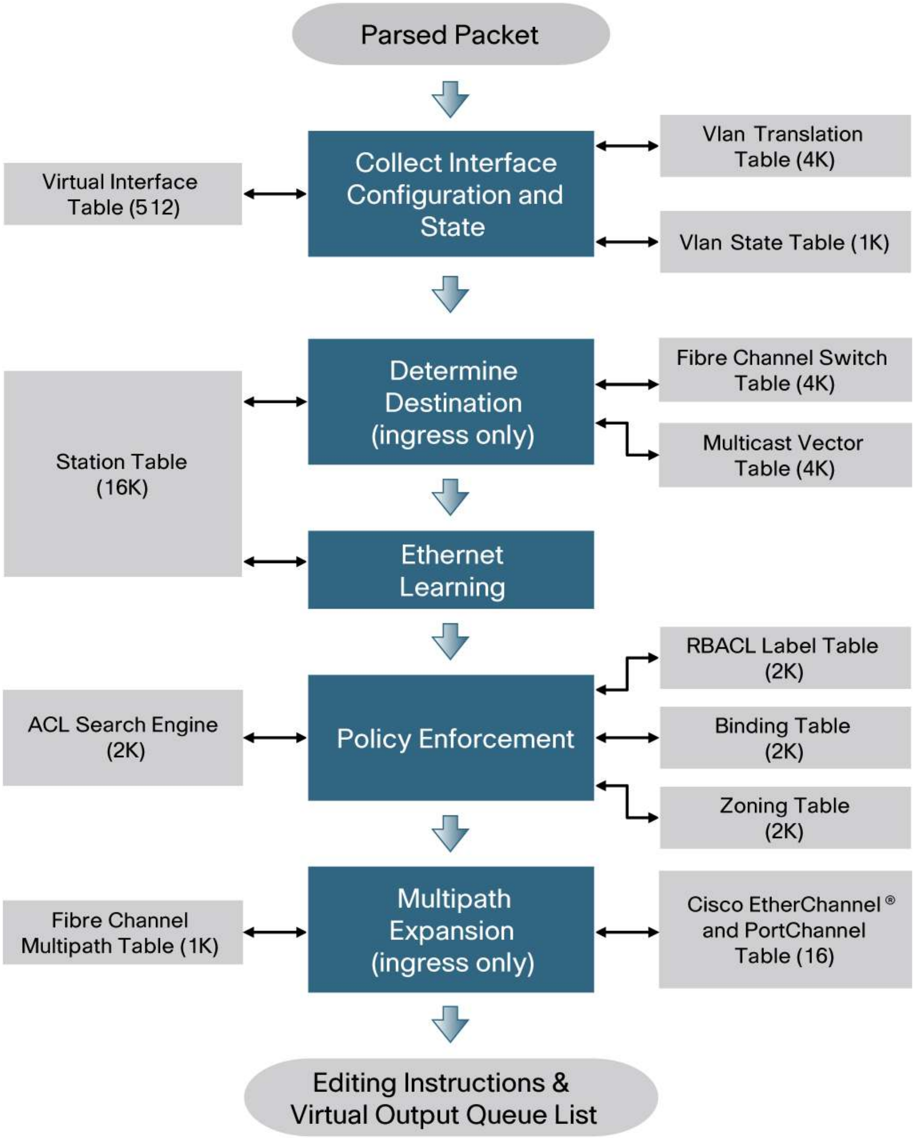
Unified Forwarding Engine

The most significant component of the Cisco Nexus 5000 Series unified fabric is the unified forwarding engine implemented in the UPC. The unified forwarding engine is a single forwarding engine implementation capable of making forwarding decisions for Ethernet and Fibre Channel. The unified forwarding engine design takes into account the similarities and differences among these two forwarding styles, to maximize the common blocks and therefore minimize the amount of logic and the number of transistors needed in the UPC. In the ASICs, the savings amount to reduced die size, power consumption, and heat dissipation, overall allowing the UPC to reach the desired density goal of 4 line-rate ports at 10 Gbps on a single chip.

To minimize bottlenecks in making forwarding decisions, the unified forwarding engine is designed to use a local coherent copy of the forwarding station table that is on the UPC silicon. The station table on the UPC is implemented in hardware with a modern dLeft hash table of 32,000 entries.

Figure 7 shows the steps in making a forwarding decision.

Figure 7. Unified Forwarding Engine Decision Making


Parsed Packet
Collect Interface Configuration and State
Virtual Interface Table (512)
Vlan Translation Table (4K)
Vlan State Table (1K)
Determine Destination (ingress only)
Station Table (16K)
Fibre Channel Switch Table (4K)
Multicast Vector Table (4K)
Ethernet Learning
Policy Enforcement
ACL Search Engine (2K)
RBACL Label Table (2K)
Binding Table (2K)
Zoning Table (2K)
Multipath Expansion (ingress only)
Fibre Channel Multipath Table (1K)
Cisco EtherChannel® and PortChannel Table (16)
Editing Instructions & Virtual Output Queue List

The following sections summarize the steps.

Virtual Interface States

The first action taken in the forwarding pipeline is to build context for the received packet. This is done by mapping the packet to an interface configuration, so that the configuration applied to the interface can take effect as the packet traverses the switch. The Cisco Nexus 5000 Series implements the concept of virtual interfaces: logical entities with independent configuration mapped to a single physical interface. This concept is extremely powerful as it allows LAN and SAN administrators to apply independent configurations to virtual Ethernet ports carrying regular Ethernet traffic and virtual Fibre Channel N ports carrying FCoE traffic. As a result, although the actual data packets are multiplexed on the same physical wire, on the management plane LAN and SAN management are presented separately and in isolation, giving maximum flexibility and providing continuity with existing data center operation models.

When a packet is received on a 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. Typically, the parsing is very simple, and it involves only searching for an FCoE header to choose between the virtual Ethernet and the virtual Fibre Channel interface.

Destination Lookup

After the unified forwarding engine has determined the appropriate virtual interface to use for the rest of packet processing, the actual forwarding decision process can be started by looking up destination MAC addresses or FC IDs in the appropriate forwarding tables. For traditional Ethernet forwarding, only one station table needs to be consulted; for Fibre Channel forwarding, the selection of destinations involves both the station table (for locally attached stations) and a switch table that handles remote destinations though Layer 2 routing. The power of the switch table is its capability to enable hardware-based equal-cost multipathing on the Cisco Nexus 5000 Series, available for Fibre Channel forwarding. The Cisco Nexus 5000 Series UPC scales to maintain a link state database of up to 4000 Fibre Channel switches.

The 32,000 entry station table in each UPC is shared by all the forwarding styles, and each entry in the table is marked for the appropriate forwarding style with the VLAN and VSAN information stored together with the entry.

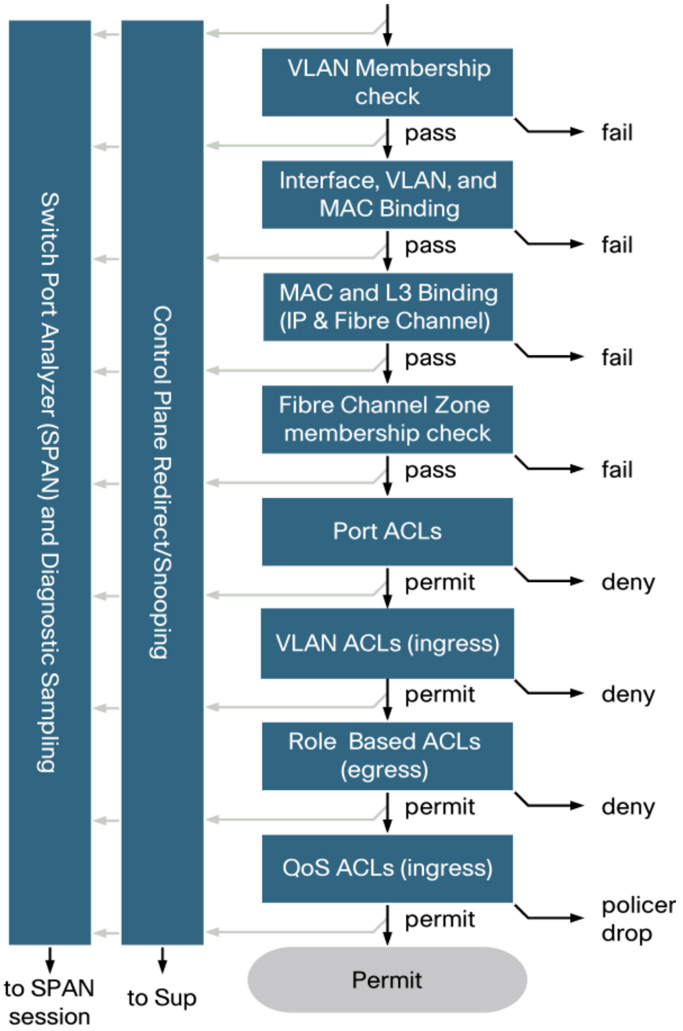
Hardware-Based Source Path Learning

When an unknown source MAC address is seen for the first time by one UPC’s unified 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, and the ingress UPC generates an interrupt to the supervisor, which updates all the other UPCs that are not touched by the flow. This technique minimizes the amount of unicast flooding needed, while still allowing a simple implementation of a distributed station table: the UPCs that are most likely to be involved in the reverse path for a flow learn the source MAC addresses in hardware.

Policy Enforcement

The Cisco Nexus 5000 Series UPC follows a strict set of comprehensive rules to help ensure that packets get forwarded or dropped based on the intended configuration. The multistage policy engine is responsible for this step and manipulates 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 Switched Port Analyzer (SPAN) replication. Specifically, ACLs are implemented in 1-Mb TCAM located on each UPC, offering 2048 match access control entries, each 432 bits wide (Figure 8).

Figure 8. Policy Enforcement



The policy engine evaluates the following elements:

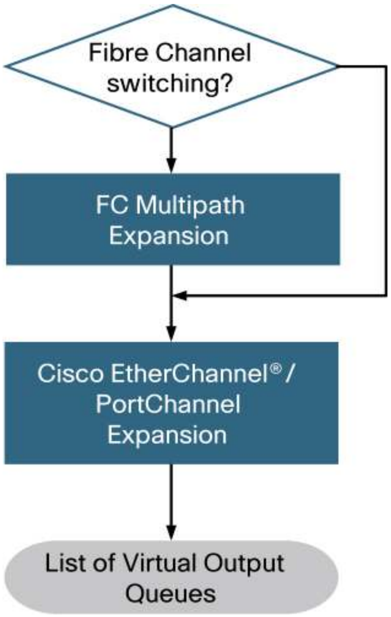
- VLAN membership
- Interface, VLAN, and MAC binding
- MAC and Layer 3 binding (for IP and Fibre Channel)
- Fibre Channel zone membership
- Port ACLs (768 access control entries)
- VLAN ACLs (1024 access control entries, only in ingress)
- Role-based ACLs (only in egress)
- QoS ACLs (64 access control entries, only in ingress)
- SPAN and diagnostic ACLs (64 access control entries)
- Control plane ACLs (supervisor redirect and snooping; 128 access control entries)

The UPC is very flexible in the allocation of access control entries, and therefore Cisco NX-OS partitions the ACLs into different functional regions. Cisco NX-OS distinguishes between the global scope of VLAN ACLs and control plane ACLs, which must be kept synchronized on all the UPCs, and the local scope of port, QoS, role-based, and SPAN ACLs, which are allocated independently on each UPC.

Multipath Expansion

When the forwarding logic looks up the station table and potentially the switch tables for a unicast packet, the egress interface that results from the lookup can be a physical or virtual interface, an aggregated interface (Cisco EtherChannel or SAN PortChannel), or an identifier describing a set of the above physical/virtual/aggregated interfaces that are equally good in reaching the specific destination. The final step of the forwarding engine is therefore to select a specific physical fabric path out of the list of logical paths that are available. This is the task of the multipath expansion logic (Figure 9).

Figure 9. Multipath Expansion Logic


Fibre Channel switching?
FC Multipath Expansion
Cisco EtherChannel® / PortChannel Expansion
List of Virtual Output Queues

The expansion logic takes into account packet flow semantics to help ensure in-order delivery of packets within a flow, while spreading different flows across different physical paths to maximize fair utilization of all the available egress interfaces.

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 5000 Series UPC hashes flows to obtain a numerical value that can be used to select among up to 16 physical interfaces. Up to 16 aggregated interfaces (Cisco EtherChannel or SAN PortChannel interfaces) can be created on the Cisco Nexus 5000 Series, each with up to 16 member physical interfaces.

VOQ Selection

When a packet is received, the UPC of the ingress interface is responsible for choosing a set of egress interfaces and UPCs that should be used to forward the packet to its final destination. Each external interface on a UPC reaches all the other external interfaces on all the UPCs through the UCF, with no exceptions: the UPC does not perform any local forwarding for the 4 ports that it manages. The goal of a forwarding decision is to select a set of internal egress fabric interfaces, put packet descriptors in the corresponding appropriate VOQs, and let the UCF drain the queues as the fabric schedulers see fit. Virtual output queuing is a practical solution to avoid head-of-line blocking, and the Cisco Nexus 5000 Series uses it extensively not only to avoid head-of-line blocking among egresses, but also to avoid head-of-line blocking among different priority classes destined for the same egress interface.

Unified Crossbar Fabric

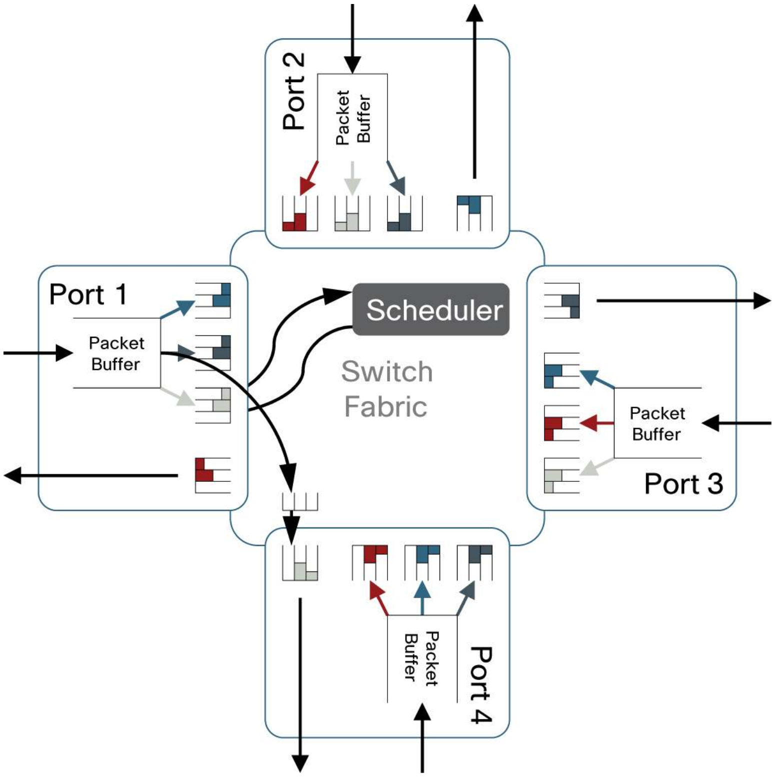
The UCF is a single-stage, high-performance 58-by-58 nonblocking crossbar with an integrated scheduler. The crossbar provides the interconnectivity between input ports and output ports with a total switching capacity of 1.04 Tbps. As packets traverse the crossbar, they are over speed by 20 percent to compensate for internal headers and to help ensure a 10-Gbps line rate for all packet sizes.

The integrated scheduler coordinates the use of the crossbar between inputs and outputs, allowing a contention-free match between input-output pairs (Figure 10). The scheduling algorithm is based on an enhanced algorithm. The original algorithm is not well suited for cut-through switching because the bounds on completing a packet in flight are not deterministic. The modified algorithm helps ensure high throughput, low latency, and weighted fairness across inputs, and starvation- and deadlock-free maximum-match policies across variable-sized packets.

All input buffering is performed by the UPC, so the UCF does not have any input buffer. For each packet, a request is sent to scheduler. There are, however, four fabric buffers and four crosspoints per egress interface, with 10,240 bytes of memory per buffer. Three fabric buffers are used for unicast packets, and one is reserved for a multicast packet. The four buffers allow use of the fabric to be granted to four ingress ports in parallel, resulting in a 300 percent speedup for unicast packets. The buffers are sent out in first-in, first-out (FIFO) order to the egress queues in the UPC, building a certain egress pipeline to fill up the egress bandwidth on the corresponding UPC and to increase throughput.

Another important characteristic of the scheduler is the credit management system, which helps ensure that space is available in the egress buffer before a VOQ a serviced. This feature implies that while the path from the UCF to an egress UPC is used to drain a fabric buffer, that fabric buffer is considered filled until the drain is complete. If either the fabric buffer on the UCF or the egress buffer pool on the UPC are unavailable for a specific (egress port or priority) pair, the scheduler will consider that egress busy.

Figure 10. UCF Integrated Scheduler



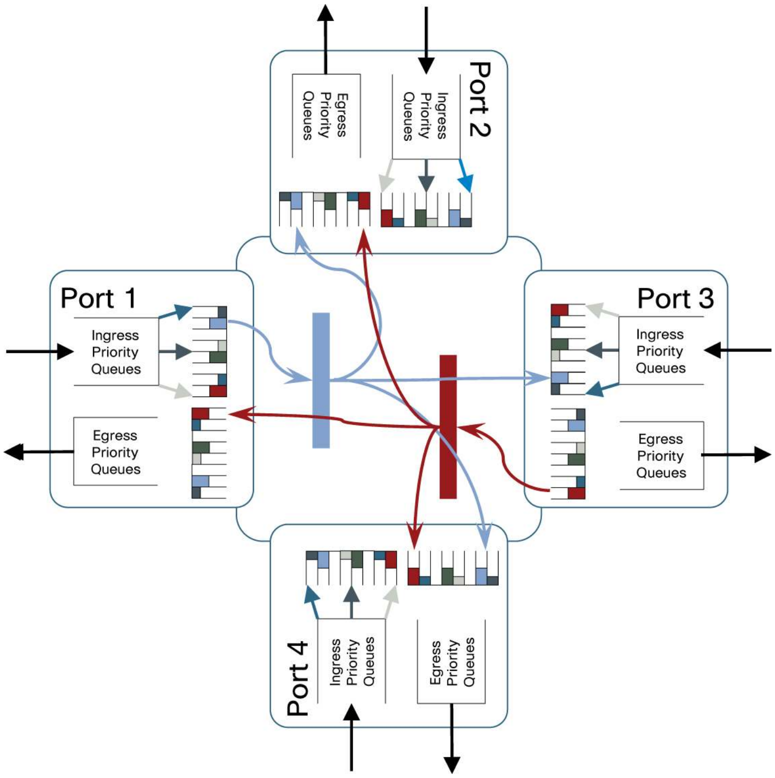
Multicast Fabric Replication

For Layer 2 multicast traffic (that is, traffic with destination MAC addresses that are MAC multicast addresses), packet replication is performed by the UCF. Fabric replication optimizes use of the internal fabric interface between the ingress UPC and the UCF, granting maximum throughput at the system level. To support fabric replication, each UPC maintains separate input queues for multicast traffic (a total of 8: one per priority class), and the UCF uses a separate multicast scheduling algorithm. The multicast scheduling algorithm can operate in two modes. In the default mode, the scheduler grants access to the internal fabric interface when the entire fan-out for the packet is available.

At that point, the ingress UPC sends one copy of the packet to the UCF, removes the packet from its internal buffer, and removes the packet descriptor from the input priority queue. Then the UCF internally replicates that single copy to all the egress fabric interfaces. The UCF helps ensure that under no condition will a multicast packet be subject to starvation in the input priority queue of a UPC; however, the ingress UPC cannot be granted access to the fabric interface for its multicast packet until all the fabric buffers in the UCF are free for all the egress ports in the fan-out of the multicast packet.

The UCF multicast scheduling algorithm can also operate in a different mode, in which it intelligently splits the packet fan-out into multiple subsets, thereby accelerating drainage for large fan-outs. As soon as a subset of the fan-out is available, the UCF grants access to the fabric interface, and the UPC sends the packet but keeps the descriptor at the head of the input priority queue. The UCF replicates the packet to the finalized subset of the fan-out and keeps track of which portion of the fan-out needs to be served later. After a minimal set of partial grants to nonoverlapping subsets of the fan-out, the entire fan-out will be served, and the UCF lets the UPC move ahead to the next packet in the input priority queue (Figure 11).

Figure 11. Multicast Fabric Replication



Conclusions

Cisco has designed the Cisco Nexus 5000 Series Switches to be the best solution for high-bandwidth, low-latency, access-layer switches for rack deployments. In the context of I/O consolidation, the Cisco Nexus 5000 Series Switches is also the basis for a Unified Fabric that can help simplify data center infrastructure, which translates into reduced capital and operational costs. This document has provided a brief overview of the switch features and benefits, followed by a detailed description of the internal implementation of the series' 10 Gigabit Ethernet, I/O consolidation, and virtualization capabilities. The document would not have been complete without introducing the two ASICs that make this possible: the unified port controller that handles all packet-processing operations on ingress and egress, and the unified crossbar fabric that schedules and switches packets. Cisco Nexus 5000 Series switches are the first members of the Cisco data center switch portfolio to deliver on the promise of a Unified Fabric and represent another step towards making the Cisco Data Center 3.0 strategy a practical reality.

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 C11-462176-03 06/09