Catalyst 6000 and 6500 Series Architecture

Executive Summary

The ever-increasing demands for bandwidth and performance, and even greater need for intelligent network services, challenge today's switches, routers, and associated devices. Customers require a wide range of performance and services, including high availability, quality of service (QoS), and security, to build scalable enterprise and service provider networks.

The Cisco Catalyst® 6000 Family, which includes the Catalyst 6000 and 6500 Series, delivers new high-performance, feature-rich, multilayer-switching solutions for enterprise and service provider networks. Designed to address the increased requirements for gigabit scalability, high availability, and multilayer switching in campus backbone and distribution points, Internet data centers, and Web hosting and e-commerce environments, the Catalyst 6000 Family complements the Catalyst 4000 and 3500 families for user aggregation, and the Cisco 7200, 7500, 7600, and Cisco 12000 Gigabit Switch Router (GSR), for high-end routing. Together, the Catalyst and Cisco router families deliver a wide range of intelligent solutions.

The Catalyst 6500 Series architecture supports scalable switching bandwidth up to 256 Gbps and scalable multilayer switching that exceeds 200 million packets per second (pps). For customers not requiring the performance of the Catalyst 6500 Series, the Catalyst 6000 Series provides a more cost-effective solution, with backplane bandwidth of 32 Gbps and multilayer switching that scales up to 15 million pps. The Catalyst 6000 Series architecture provides a superior schema for congestion management by using per-port buffering. In addition, the advanced switching engines of the Catalyst 6500 Series ensure that services such as QoS, security, and policing can be done with no performance degradation.

Catalyst 6509

Catalyst 6506

Catalyst 6513

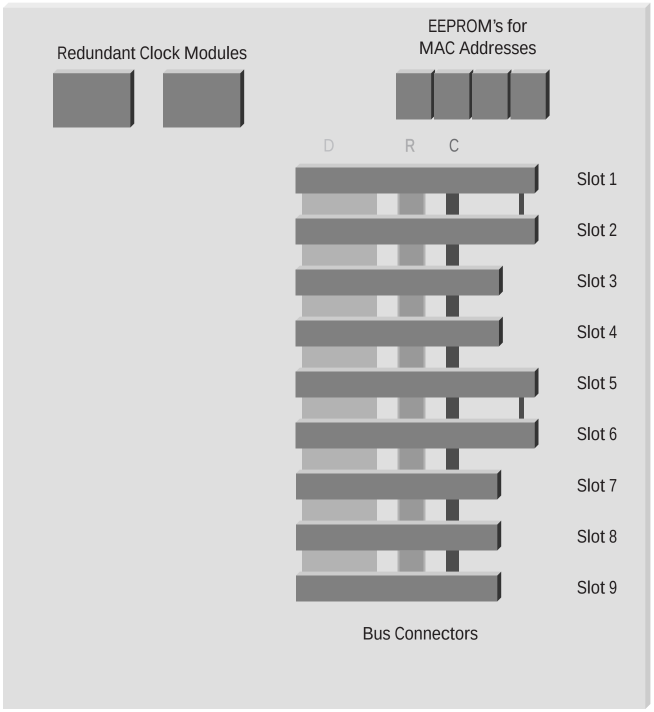
Catalyst 6000 Family Hardware Options

The Catalyst 6000 Family comprises the Catalyst 6000 and 6500 Series Switches. Cisco Systems offers the Catalyst 6000 Family in several different chassis options. Because customers look for modularity in offered slots, the Catalyst 6000 and 6500 offer six- and nine-slot chassis versions. The Catalyst 6500 also offers a 13-slot chassis and a “NEBS” chassis, in which slots are vertically arranged. Both Catalyst 6000 and 6500 Series Switches also support a wide range of interface types and densities, including 384 10/100 Ethernet ports, 192 100Base-FX Fast Ethernet ports, and up to 130 Gigabit Ethernet ports (in the nine-slot chassis). The 13-slot chassis offers up to 576 10/100 ports or 192 Gigabit Ethernet ports.

Backplane layout in the Catalyst 6000 and the Catalyst 6500 differs. The Catalyst 6000 uses a 32-Gbps switching bus. The Catalyst 6500 offers both the 32-Gbps switching bus and the option for the 256-Gbps Switch Fabric Module (SFM). The SFM uses slot 5 (and optionally slot 6 for redundancy) and provides the system with a high-speed switching path. The SFM cannot be used in the Catalyst 6000 Series. The following figures show the backplane layout of the Catalyst 6000 and 6500 Series.

Catalyst Backplanes

Figure 1 Catalyst 6009 Backplane



Catalyst 6000 Series

Slot 1—Supervisor Engine

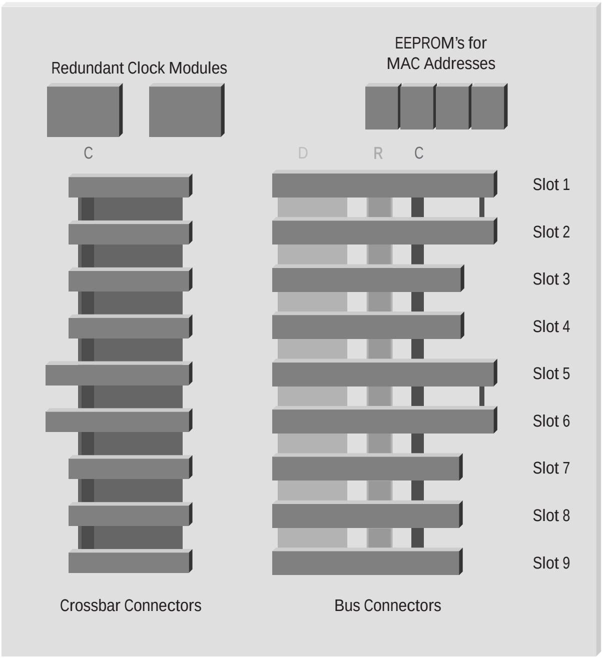
Slot 2—Line Card or Redundant Supervisor Engine

Connectors on Right for:
- EARL Data Bus (Dbus)
- EARL Results Bus (Rbus)
- Ethernet Out of Band Channel (EOBC)

Redundant Clock Modules

MAC Address EEPROMs

Figure 2 Catalyst 6509 Backplane



Catalyst 6500

Slot 1—Supervisor Engine

Slot 2—Line Card or Redundant Supervisor Engine

Connectors on Right for:
EARL Data Bus (Dbus)
EARL Results Bus (Rbus)
Ethernet Out of Band Channel (EOBC)

Redundant Clock Modules

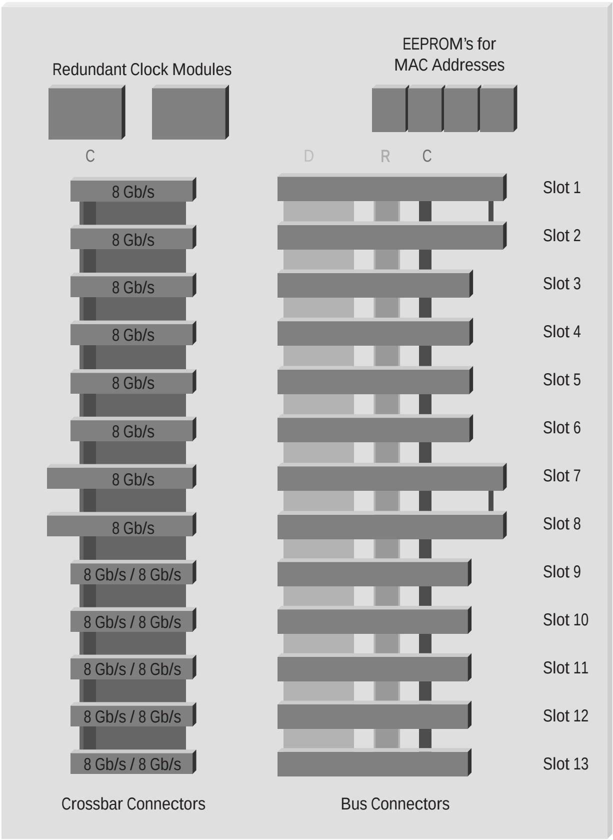
MAC Address EEPROMs

Slot 5—Switch Fabric Module

Slot 6—Line Card or Redundant Switch Fabric Module

Connectors on Left for 256 Gbps Crossbar Switching Fabric

Figure 3 Catalyst 6513 Backplane



Catalyst 6513

Slot 1—Supervisor Engine

Slot 2—Line Card or Redundant Supervisor Engine

Connectors on Right for:
EARL Data Bus (Dbus)
EARL Results Bus (Rbus)
Ethernet Out of Band Channel (EOBC)

Redundant Clock Modules

MAC Address EEPROMs

Slot 7/8—Optional Switch Fabric Module Support or Line Card

Supports 2500W/4000W Power Supply

Connectors on Left for 256 Gbps Crossbar Switching Fabric

Slots 2–8 Support 1 * 8 Gb/s Fabric Channel, Slots 9–13 Support 2 * 8 Gb/s Fabric Channels

Line Cards

Two line card versions are available for the Catalyst 6000 Family. Because customers have the option of using the 32-Gbps switching bus or the 256-Gbps SFM, the two line card versions provide connectivity into the different fabrics. The following table lists the line cards and their compatibility.

Table 1 Line Card Summary

Line Card Version	Chassis Supported	Explanation
Non Fabric-enabled	Catalyst 6000 and 6500	Available since the inception of the Catalyst 6000 Family. These line cards connect to the 32-Gbps bus and can be used in all chassis.
Fabric-enabled	Catalyst 6500	Connect into both the 32-Gbps bus and the SFM. The SFM is not supported in the Catalyst 6000 Series.
Fabric-only	Catalyst 6500	Connect only to the SFM and provide the highest levels of system performance and throughput. These line cards require the SFM because they do not connect to the 32-Gbps bus.

Because the non fabric-enabled line cards can be used in any chassis option they provide the greatest flexibility and investment protection for customers who have already deployed the Catalyst 6000 Family. For customers requiring very high bandwidth and throughput, Cisco recommends using the fabric-enabled or fabric-only line cards.

Note: All line cards interoperate with each other. For example, a non fabric-enabled line card in the Catalyst 6500 interoperates with a fabric-only line card in the same system. An explanation of how that works is provided in the section titled “Catalyst 6000 Family Packet Flow—A Day in the Life of a Packet.”

Supervisor Options

The Catalyst 6500 Series includes two versions of the Supervisor Engine. The Supervisor Engine is required for system operations; a chassis without a Supervisor will not operate. The Supervisor Engine uses slot 1 in the chassis. The second slot in the system, slot 2, can be used for a secondary redundant supervisor engine. Note that because of the switching implementation of the Catalyst 6000 and 6500, only one Supervisor Engine needs to be active at one time. However, with the High Availability feature enabled, both supervisors maintain the same state information, including Spanning-Tree topology, forwarding tables, and management information, so that if the primary supervisor fails, the redundant engine can take over within two seconds.

To address the needs of different customers who deploy the Catalyst 6500 in varying applications, Cisco provides two Supervisor Engines.

Supervisor Engine 1

Supervisor 1, the first switching engine for the Catalyst 6000 Family, provides performance levels of 15 million pps using a cache-based switching scheme. Supervisor 1 has three main components: the Network Management Processor (NMP), the Multilayer Switch Feature Card (MSFC) and the Policy Feature Card (PFC). Each component provides a critical function to the network.

Supervisor 1 is available in three options:

- Basic Layer 2 switching with no Layer 3-based QoS or security access control lists (ACLs), although port-based class of service (CoS) and destination MAC address-based CoS is supported. Basic switching based on the MAC address is supported.
- Supervisor 1 with the PFC, which provides Layer 2 switching with Layer 3 services (including QoS and security ACLs). QoS classification and queuing, as well as security filtering, is supported at data rates of 15 million pps. This functionality is supported at Layer 2 and 3 even though Layer 3 switching and routing is not performed.
- Supervisor 1 with PFC/MSFC1 (or 2), which provides full Layer 3 switching and routing. This combination enables the Catalyst 6500 to route IP and Internet Packet Exchange (IPX) traffic at 15 million pps.

Supervisor Engine 2

Supervisor Engine 2 is designed specifically for service provider and high-end enterprise core applications. This supervisor engine provides forwarding capability of up to 30 million pps when using fabric-enabled line cards and the SFM (both must be used). An important difference between the Supervisor 1 and Supervisor 2 is that Supervisor 2 supports Cisco Express Forwarding (CEF) in hardware. CEF is a switching implementation that is based on the topology of the network rather than the traffic flow. This causes the control plane of the Catalyst 6500 scale to converge faster in the event of a route flap and perform lookups for millions of flows (which occur in Internet service provider [ISP] deployments).

Supervisor 2 has two options:

- Supervisor 2 with PFC-2, which provides QoS, Private Virtual LAN (PVLAN), and ACL functionality at 30 million pps with no performance penalty.
- Supervisor 2 with PFC-2/MSFC-2, which enables full routing on the Catalyst 6500. This supervisor enables the Catalyst 6500 to provide Internet-class routing and high performance.

Supervisor Engine 2 can be used in the Catalyst 6000 and Catalyst 6500 chassis. The SFM requires use of Supervisor 2, but Supervisor 2 can operate independently of the SFM.

Catalyst 6000 Family Architecture

The Catalyst 6000 and Catalyst 6500 Series architecture is described in several sections of this paper. Each section details the system functionality and how the components operate. This section covers the following topics.

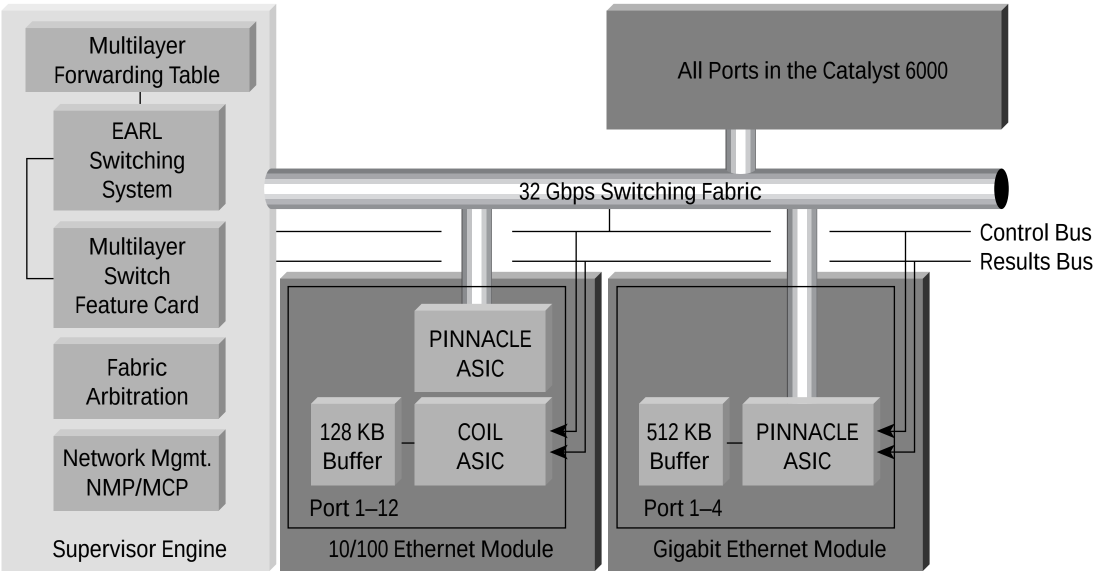
- Catalyst 6000 and 6500 switching bus architecture
- Catalyst 6500 SFM architecture
- The Multilayer Switch Feature Card (MSFC)
- Switching implementation on Supervisor 1
- Switching implementation on Supervisor 2
- Distributed Forwarding Card (DFC)
- QoS and ACL handling
- Line card buffering and ASIC overview

The Switching Fabric—Moving the Packet

Catalyst 6000 and 6500 Switching Bus Architecture

Figure 4 illustrates the Catalyst 6000 Family architecture.

Figure 4 Catalyst 6000 and 6500 Switching Bus Architecture



The Catalyst 6000 system is based on a 32-Gbps advanced pipelining switching bus. The switching bus is a shared medium bus; that is, all the ports attached to the bus see all the frames transmitting across it. Coupled with the pipelining mechanism, this switching is very efficient because after a decision is made the switching engine orders the nondestination ports to ignore the frame.

The Catalyst 6000 switching bus includes three distinct buses: the D-bus (or Data bus), the C-bus (or Control bus), and the R-bus (or results bus). All non fabric-enabled line cards connect to the switching bus through the connectors on the right side of the chassis (see Figure 1). The D-bus is the bus where data is forwarded from one port to another and realizes a bandwidth of 32 Gbps. The Results bus (R) takes information from the switching logic located on the Supervisor Engine back to all the ports on the switch. The control bus (C-bus) relays information between the port ASICs and the Network Management Processor (NMP).

Two notable features on the switching bus of the Catalyst 6000 are pipelining and burst mode. Pipelining enables the Catalyst 6000 Family systems to switch multiple frames onto the bus before obtaining the results of the first frame. Typically, on shared medium architectures, only a single frame or packet can reside on the bus at a time. The entire lookup process by the forwarding engine happens in parallel to the transfer of the frame across the switching bus.

If the frame is switched across the switching bus before the lookup is done, the switching bus is idle until the lookup is done. This is where pipelining comes into play. Ports are allowed to source frames on the switching bus before the results of the first frame lookup are done. The second frame (from any port) is switched across the bus and pipelined for a lookup operation at the forwarding engine. Thirty-one such frames can be switched across the switching bus before the result of the first frame is received. The 32nd frame must wait before it can be sourced on the switching bus.

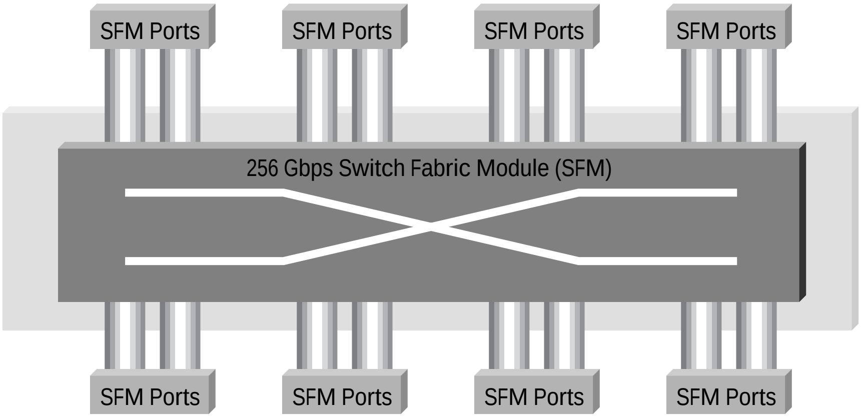
The Burst-Mode feature enables the port to source multiple frames on the switching bus. If the port sends just one frame each time it is granted access to the data bus, there is potentially an unfair allocation of bandwidth to the bus when it is heavily loaded. For example, if two ports are trying to send data and one has 100-byte frames while the other has 1000-byte frames, the port with the 1000-byte frames can switch 10 times as much data as the one with 100-byte frames. This is because they alternate in the arbitration and each sends one frame at a time.

In the Catalyst 6000 Family systems, a port can send multiple frames to the switching bus in a manner that controls the amount of bandwidth it consumes regardless of the frame size. To accomplish this, the port ASICs maintain a counter of the number of bytes it has transferred and compare it to a threshold. Provided that the count is below threshold value, the port continues to send frames as long as it has data to send. When the count exceeds the threshold, the port stops sending data after completing the current frame and stops transmitting because the arbitration logic at this point senses the condition and removes bus access. The threshold value is a function of the number of ports in the system, their thresholds, empirical data, simulation results, and so forth. The system automatically computes the threshold value to ensure fair distribution.

Catalyst 6500 Crossbar Switching Fabric

The Catalyst 6500 and the Switch Fabric Module (SFM) provide a 256-Gbps switching system with forwarding rates over 100 million pps. The SFM uses the connectors on the left side of the Catalyst 6500 chassis. Note that because these connectors are not in the Catalyst 6000, this chassis cannot use the SFM. The SFM uses a 256-Gbps crossbar switching fabric to interconnect the line cards on the switch. Figure 5 is a logical diagram of the SFM.

Figure 5 Catalyst 6500 Switch Fabric Module



The SFM can best be thought of as a 16-port “switch,” with the ports actually connecting to the line cards. In the Catalyst 6500, each slot in the chassis receives two crossbar ports, and each port is clocked at 8 Gbps (the actual bandwidth is 16 Gbps because there is one 8-Gbps path for transmitting into the crossbar and 8 Gbps for transmitting out of the crossbar). The fabric-enabled modules connect to one of the ports on the crossbar, providing 8-Gbps access into the switching fabric. The fabric-only line cards attach to both ports per slot into the crossbar, allowing them 16 Gbps of connectivity.

The Catalyst 6500 SFM uses overspeed to eliminate congestion and head-of-line blocking. Overspeed is a concept by which the internal “paths” *within* the crossbar fabric are clocked at a speed faster than the input rates *into* the crossbar. This allows packets to be switched out of the source module through the fabric to the output line card at high data rates. The SFM uses 3x overspeed, meaning that each internal trace is clocked at 24 Gbps relative to the input rate, which is clocked at 8 Gbps.

Local Switching on the Fabric-Enabled Line Cards

Each of the line cards connecting to the SFM uses a local switching fabric. The fabric-enabled cards, such as the WS-X6516, support the DFC to enable high-speed switching. These line cards have connectivity to one channel port on the SFM and also have a connection into the 32-Gbps centralized switching bus. The fabric-only line cards, such as the WS-X6816, connect only into the SFM via dual fabric channels. Figure 6 and Figure 7 show these line cards.

Figure 6 SFM Single-Attached Fabric-Enabled Card with Optional DFC

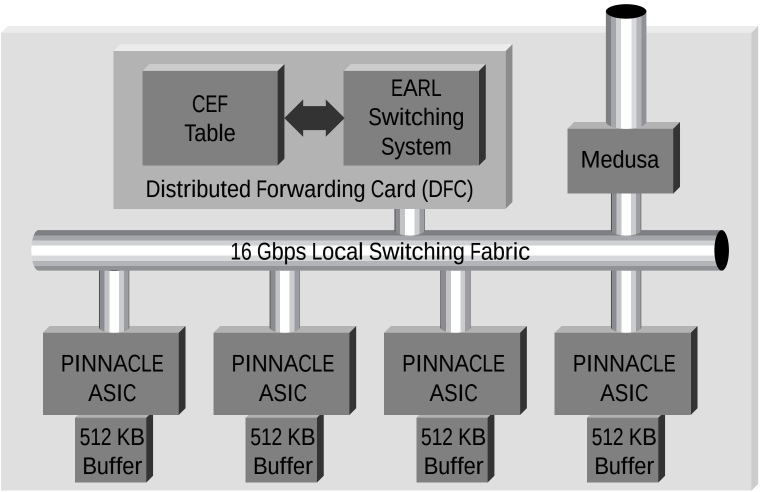
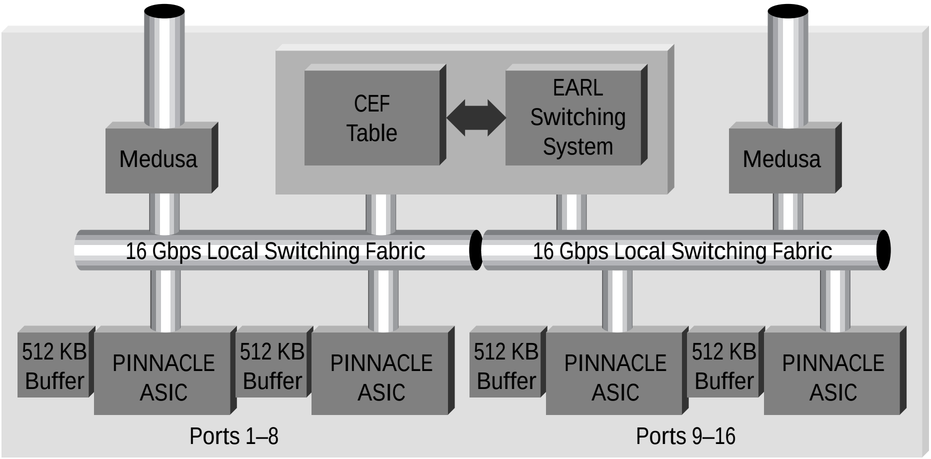


Figure 7 SFM Fabric-only Line Cards



The key difference between the two line cards is that the fabric-enabled cards use a single local switching bus with a bandwidth capacity of 16 Gbps. The fabric-only line cards use two local switching buses, each clocked at 16 Gbps. Both line cards can support distributed forwarding. The DFC daughter card is available as an add-on for the fabric-enabled cards. The fabric-only line cards have the DFC embedded in the system.

A critical component of the local-switch implementation is the connection point between the local system and the SFM. In the Catalyst 6500, this function is handled by an ASIC called Medusa. This ASIC is the interface between the local bus and the crossbar. On the fabric-enabled cards (*not* the fabric-only cards), Medusa also interfaces to the main 32-Gbps switching bus. How Medusa actually functions in packet switching is discussed in a later section titled “Catalyst 6000 Family Packet Flow—A Day in the Life of a Packet”.

The Switching Implementation—Making a Switching Decision

Switching implementations use two key functions: the control plane and the data-forwarding plane. The control plane maintains all of the overhead functions of the switch, including handling of the routing protocols, the routing table, flow initiation, and some access control. This function, often overlooked in the “race for speed,” is absolutely critical to the switching architecture and cannot be ignored. In the Catalyst 6000 Family, the MSFC handles the control plane function. The packet forwarding decision is done in hardware and can take place at data rates exceeding 100 million pps. This functionality is handled by the Supervisor Engines and, on some line cards, the Distributed Forwarding Cards (DFCs).

Multilayer Switch Feature Card (MSFC)

The Catalyst 6500 uses centralized routing control plane functionality. This capability is provided by a daughter card module called the MSFC on the Supervisor Engine. The MSFC handles all of the control plane functions within the switch architecture.

Note: There are two versions of the MSFC: MSFC-1 and MSFC-2. This document focuses on the MSFC-2.

The MSFC is based on an R7000 300-MHz processor. This gives the MSFC a forwarding performance rate in software of 650,000 pps. For a Catalyst 6500 using Supervisor Engine 1, this provides very high performance for the flow setup, which the architecture mandates.

The MSFC can handle and maintain large routing tables. There are three memory options available: the standard 128 Mbytes, an option for 256 MB, and an option for 512 MB. For networks that require the Catalyst 6500 to handle the entire Internet routing table, Cisco recommends the 512 MB version.

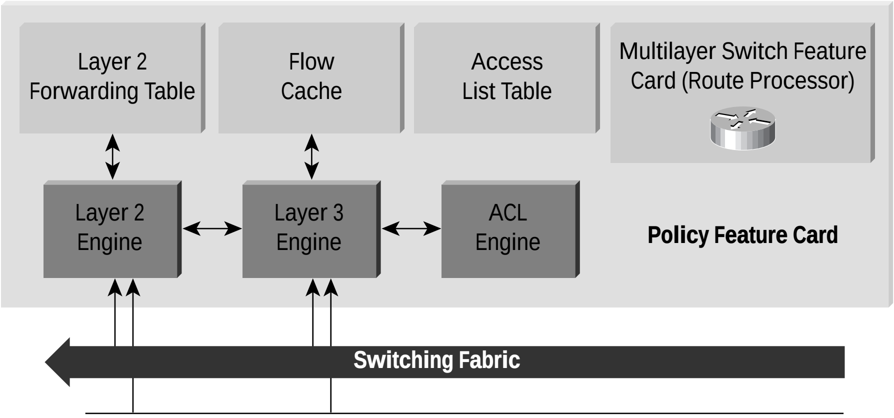
For Supervisor Engines 1 and 2, the MSFC maintains the routing table and communicates across an out-of-band bus to the hardware switching ASICs. On Supervisor Engine 1, the first packet in a flow that does not have an entry in the hardware-switching table is sent to the MSFC for software processing. The MSFC compares the destination IP address with the routing table and make a forwarding decision. After the MSFC has switched the first packet of a new flow in software, the hardware is automatically programmed to switch subsequent packets in the ASIC complex.

In Supervisor Engine 2, the MSFC does not forward IP frames. Instead, it builds a CEF table (also known as a Forwarding Information Base [FIB]) table, which is based on the contents of the routing table. The CEF table contains the same information as the routing table and uses a highly optimized search algorithm in order to “hit” on the destination network. The MSFC downloads the CEF table directly into the hardware so all packets are switched in hardware and not by the MSFC itself.

Supervisor Engine 1A and the Policy Feature Card—Traffic-based Switching

The second critical component of switching in the Catalyst 6000 Family also resides on the Supervisor engine and is called the Policy Feature Card (PFC). The PFC actually contains the switching ASICs that enable high-speed switch at data rates up to 15 million pps. Because there are substantial differences between Supervisor Engine 1 and Supervisor Engine 2, these differences are addressed separately. The following figure shows the functional components of the PFC.

Figure 8 Supervisor 1 and the Policy Feature Card



The PFC is the heart of the switching system. The forwarding decisions of the Catalyst 6500 are made by three ASICs: one for Layer 2 MAC-based forwarding, one for Layer 3, and the other for ACLs, whether for security or QoS. The Layer 2 ASIC handles two key items. First, and somewhat obviously, it looks up MAC addresses within a broadcast domain in order to switch at Layer 2. However, this ASIC also identifies a packet (or flow) that needs to be Layer 3 switched. The MSFC, shown in the figure above, primes an entry in the flow cache, which the Layer 3 engine uses to switch packets in hardware. The MSFC registers its MAC address with the Layer 2 Engine, so that, upon examination of a packet, the Layer 2 engine can decide to ignore the result of the lookup performed by the Layer 3 engine.

After the system determines that Layer 3 switching needs to take place, the result of the Layer 3 engine is used. The Catalyst 6500 with a Supervisor 1 uses a switching mechanism called traffic-based switching (also known as flow-caching). A flow is defined as a traffic stream from a source IP address to a destination IP address. Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) port information can also be stored as part of the flow cache entry. When the first packet in a flow enters the switch, a lookup is done in the hardware lookup table to see if an entry exists. If one does not, the packet is sent to the routing software running on the switch's CPU, which matches the destination IP address against the routing table, locates VLAN-of-exit, switches the packet, and automatically creates an entry in the hardware flow-cache. Subsequent packets can be switched in hardware. For this system to work effectively, a fast CPU, and more importantly, fast and efficient software, must be used.

The Catalyst 6500 flow cache can support a maximum of 128,000 entries. How the flow cache is used depends on how the Catalyst 6500 is "told" by the network manager to switch flows (the command line interface allows this configuration). There are three options in the Catalyst 6500.

- Destination-only—In this case, a flow is built and stored in the PFC based on the destination IP address. This allows for the best utilization of the PFC flow cache as multiple sources communicating with one destination IP address (a server, for example) only use a single flow entry.
- Source-destination—In this case, a flow is built based on both the source and the destination address. This takes up more entries in the cache if, for example, five sources are talking with one destination, that uses five entries in the cache.

- Full flow—This is a very expensive way of using the flow cache. In this mechanism, a flow is created on not only the source and destination IP, but also on the UDP or TCP port number. A single IP source and destination pair could potentially take up several entries, one for each TCP or UDP stream.

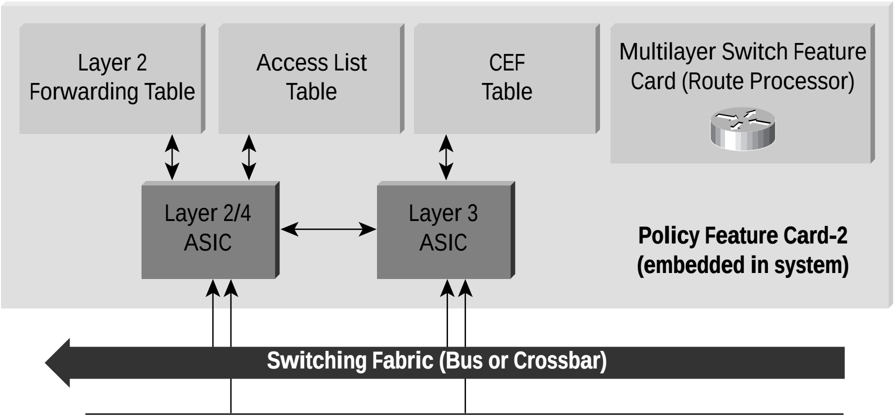
The flow table is broken up into 8 pages of memory; each page can store 16,000 entries. A hashing algorithm is used to perform lookups in the table. The hashing algorithm is critical both to learning packets and storing them at high speed as well as switching at data rates of 15 million pps. Because of the statistical nature of hashing algorithms, hash collisions can occur. A hash collision occurs when the lookup for two packets hashes to the same location in memory. To account for this, the Layer 3 engine turns to the next page in memory to see if that location is used. This will continue either until the address is learned or until the eighth page is reached. If the learning information still cannot be stored, then the packet is flooded (at Layer 2) or sent to the MSFC (for Layer 3). Note that, since network are very dynamic, with flows being learned and aged out, that this is a very rare occurrence.

The Supervisor Engine 1 with the PFC is designed to forward packets at 15 million pps. Cisco recommends it for deployment in most network scenarios, including the network access layer (such as a wiring closet or server farm) and enterprise network distribution points (such as the MDF of a large building).

Supervisor Engine 2—CEF-based Forwarding in Hardware

Supervisor Engine 2 for the Catalyst 6500 provides higher speed switching and resiliency relative to Supervisor Engine 1. The switching system on Supervisor 2, often referred to as PFC-2, can provide data rates up to 30 million pps. The major difference in the system is the switching implementation. Unlike Supervisor 1A, Supervisor 2 uses a CEF-based forwarding system in hardware. The following figure illustrates the PFC-2 on Supervisor Engine 2.

Figure 9 Policy Feature Card 2 on Supervisor Engine 2


Layer 2 Forwarding Table
Access List Table
CEF Table
Multilayer Switch Feature Card (Route Processor)
Layer 2/4 ASIC
Layer 3 ASIC
Policy Feature Card-2 (embedded in system)
Switching Fabric (Bus or Crossbar)

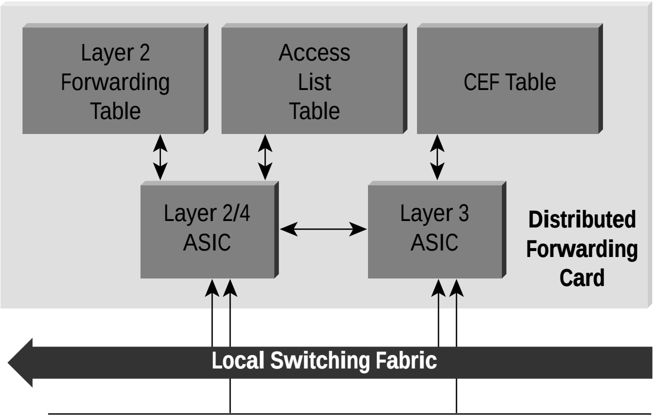
There are several differences in PFC-2. The most noticeable one is the combination of the Layer 2 and ACL engines into a single engine. Also important is the fact that a flow cache is no longer used in the system. The MSFC, located on Supervisor 2, downloads the CEF table to the Layer 3 engine, which in turn places the CEF table in hardware. Note that a flow cache table is also built in the PFC-2, although it is used for statistics gathering and not for packet switching.

Unlike a flow cache, which is based on traffic flow, the CEF table is based on the network topology. When a packet enters the switch, the switch performs a longest match lookup based on the destination network and the most specific netmask. Instead, for example, of switching based on a destination address of 172.34.10.3, the PFC-2 looks for the network 172.34.10.0/24 and switches to the interface connecting to that network. This scheme is highly efficient and does not involve the software for anything other than the routing table and prepopulation of the FIB table. In addition, cache invalidation because of a route flap does not occur; as soon as a change is made in the routing table, the CEF is updated immediately. This makes the CEF table more resilient to changes in the network topology.

Supervisor 2 can also enable distributed switching through the DFCs, which are daughter cards for the fabric-enabled line cards (such as the WS-X6516). To achieve 30 million pps, the Supervisor Engine 2 uses the 32-Gbps switching bus for control traffic from the source line card to the PFC-2. By compressing the header required for the lookup and sending it across the switching bus, the PFC-2 can look up packets much faster. The data forwarding actually takes place over the SFM. By using the DFCs, the forwarding decision is localized to the line card and, instead of sending the headers to the Supervisor Engine, the packet can be switched directly over the SFM.

Conceptually, the DFC looks the same as the PFC-2 on the Supervisor Engine. It contains virtually the same components with one exception: the routing engine (which maintains the routing table) is not present. Therefore, although the forwarding decision is localized, the control plane is still centralized. This provides the best of both worlds: centralized device control and high-speed packet forwarding. The following figure illustrates the DFC.

Figure 10 Distributed Forwarding Card Architecture



The key difference in the DFC-enabled system is that a switching decision is made on the input module instead of centrally at the Supervisor Engine. The CEF table used for switching, however, is still calculated centrally at the Supervisor Engine and then downloaded to the CEF table local to the line card. Therefore, the Supervisor CEF table and the local CEF tables are always synchronized.

The distributed nature of the Catalyst 6500 with Supervisor Engine 2 and the DFC daughter cards enables forwarding rates exceeding 100 million pps.

The Supervisor Engine 2 and the DFCs are intended for very large enterprise and service provider backbone networks. These networks, more often than not, can take advantage of the high-speed switching capability of more than 100 million pps and the resilience and scalability of CEF-based forwarding.

Quality of Service and Security Access Control Lists

The Catalyst 6500 system uses a unique mechanism for handling QoS and security ACLs in hardware. This mechanism, called Ternary CAMs (TCAMs), is used on Supervisor 1, Supervisor 2, and the DFCs. Following is a brief explanation of TCAMs and how they operate.

The PFC switching system can support up to 512 ACL labels. This switching system can identify 512 unique ACLs (by names) where each ACL can have multiple access control entry (or ACE) for a total of 32,000 maximum ACEs across 512 ACLs. The TCAM has the capability of storing 64,000 entries. This is broken up into four main blocks: input checking and output checking for QoS ACLs and input and output checking for security ACLs. Within these sections, there is a further distinction made between QoS and security ACLs.

The operation of a TCAM is similar to a hashing table, but in the TCAM, the “hashing value” is the mask. Out of the 32K possible locations, each set of eight locations has a mask associated with it. When an access control entry (or ACE) is configured, a mask in the command syntax tells the ASIC which bits to check for in the 134-bit long flow label. Suppose the ACE has a mask, that instructs it to check for the first 20 bits, and the rest of the bits are designated “don’t care” (meaning they can be any value). This results in a specific checking pattern that falls into, for example, Mask # 1.

In this example, the next time an ACE is entered that has the same mask, it will be populated as ACE 2 (if that location is available) and the process continues. If an ACE is entered that says to check for some other combination of 0, 1 or Don’t Care bits, this will become Mask # 2. All ACEs matching this mask will now go in this bank, if you will. So, if every entry (flow label) in the TCAM has a different mask then a single TCAM will be able to hold 1K entries only, because there is a single mask for every eight entries.

On the other hand, if flow labels have maximum sharing of masks then we will be able to use the entire 32K within the TCAMs. This sharing of mask dictates the total number of entries that can be stored in the TCAM. Worst case is 4K entries (shared between Security ACLs and QoS) and best case is 32K.

Table 2 Example Showing TCAM Usage with Mask Sharing

ACE 1	ACE 9	ACE 17	ACE 22	ACE 30
ACE 2	ACE 10	ACE 18	ACE 23	ACE 31
ACE 3	ACE 11	ACE 19	ACE 24	ACE 32
ACE 4	ACE 12	ACE 20	ACE 25	ACE 33
ACE 5	ACE 13	ACE 51	ACE 26	Blank
ACE 6	ACE 14	Empty	ACE 27	Empty
ACE 7	ACE 15	Empty	ACE 28	Empty
ACE 8	ACE 16	Empty	ACE 29	Empty
MASK # 1	MASK # 2	MASK # 3	MASK # 4	MASK # 5

The TCAM implementation and behavior is similar for PFC-1 and PFC-2. This functionality allows for ACL lookups to take place at data rates up to 15 million pps on Supervisor Engine 1 and 30 million pps on the DFC and Supervisor Engine 2. Because ACL lookups happen in parallel to the Layer 2 and 3 lookups, there is no performance loss when using security or QoS ACLs.

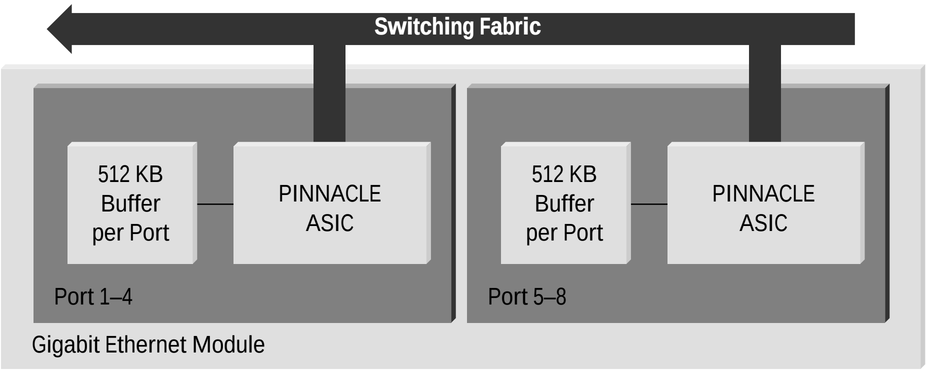
Buffering and Congestion Management for the 10/100 and Gigabit Ethernet Line Cards

Buffering and congestion management are a critical component of a switching architecture. In almost every network design and architecture, there is a congestion point somewhere in the switch. Unlike many vendor’s claims, that congestion point is almost never the switching fabric in a real network. Instead, the congestion point is typically an outgoing port or interface. This is seen in network designs such as in the access layer, where multiple user ports are accessing the uplink, multiple distribution switches are accessing the port to the core, and the core network aggregating traffic to the server farm, WAN, or the Internet.

The Catalyst 6500 has built the congestion management techniques into custom ASICs. These ASICs also serve to provide access from the ports onto the 32-Gbps main switching bus or the 16 Gbps local bus on the fabric-enabled and fabric-only line cards. These ASICs are also informed by the PFC (or the DFC) how to queue a packet for quality of service. There are two types of port ASICs, one for Gigabit Ethernet, the other for 10/100.

The subsystem used for the Gigabit Ethernet ports is referred to as Pinnacle. A diagram of Pinnacle is shown in the figure below. Each Pinnacle ASIC controls four Gigabit Ethernet ports. On the 16-port Gigabit Ethernet line card, for example, there are four Pinnacle ASICs. Each Pinnacle ASIC maintains a 512-K buffer per port. To avoid head-of-line blocking to the switching fabric, a minimal amount of buffer is provided for the receive (or RX) queue, which is taking frames coming in from the network. The bulk of the buffering is allocated for the transmit (or TX) queue. The ratio between transmit and receive queues is 7-to-1; this means that, on each port, there is 448 Kbytes of TX buffer compared to 64 Kbytes of RX buffer. This makes the Catalyst 6500 an outbound queuing model switch.

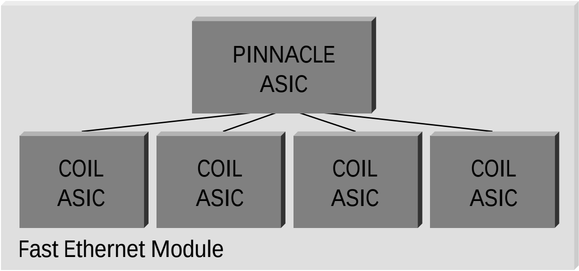
Figure 11 PINNACLE Port ASIC and Buffering



The Pinnacle ASIC takes the division of buffering a step further in regards to its handling of QoS. Each port has two receive queues and three transmit queues. Of the three TX queues, one is handled in a strict priority fashion, meaning that it always has a fixed amount of bandwidth guaranteed by the outbound scheduling logic. The other two queues are handled via the weighted round robin (WRR) scheduler. In the WRR scheduling mechanism, the two queues are weighted relative to each other and given proportional access to the outgoing port. The TX queues themselves are, by default, allocated 15 percent for strict priority, 15 percent for high priority, and 70 percent for low priority. A common example used to explain the proportion is the airline model: most of the seats are economy class, analogous to a high amount of low priority traffic; many fewer seats for business, or high priority traffic, and still fewer for first or premium class.

The 10/100 ports have a slightly different mechanism for handling congestion management. These line cards use a combination of Coil and Pinnacle to achieve congestion management. As shown in the figure below, each Coil ASIC supports 12 10/100 ports. In turn, each Pinnacle ASIC supports four COILs. This allows support for all 48 ports on the line card. Like the Gigabit Ethernet line card, COIL provides support for buffering on a per-port basis. Each 10/100 port in the system has 128 Kbytes of buffer. This buffer is broken down into a 7-to-1 ratio between TX and RX. This eliminates any head-of-line-blocking issues because of outbound interface congestion.

Figure 12 COIL ASIC Configuration on 10/100 Line Cards



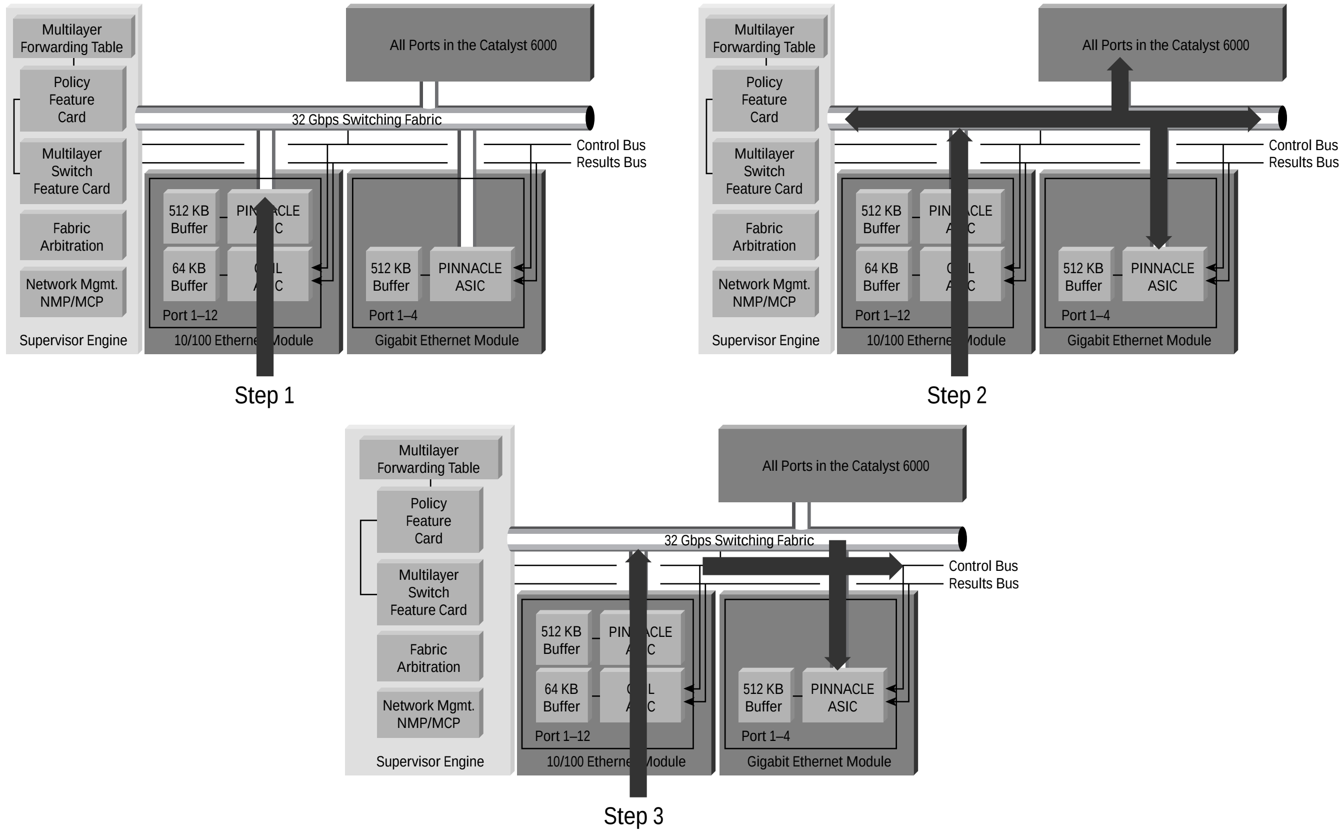
Congestion management is an important system consideration within the architecture. The ability of the switch to handle situations of many-to-one traffic flows is key to ensuring that the architecture will perform in a real network.

Catalyst 6000 Family Packet Flow—"A Day in the Life of a Packet"

Catalyst 6000 and the 32-Gbps Switching Bus

This section details how packets actually flow through the switch. Refer to the diagram below for a step-by-step process.

Figure 13 Catalyst 6000 and the 32-Gbps Switching Bus Packet Switching



Step 1—Packet enters the switch

When a frame enters the Catalyst 6000 switch, the input port takes the frame in and places it in the input buffer. The input buffer is design to handle store-and-forward checking and hold the frame while PINNACLE arbitrates for access onto the switching bus. There is a local arbiter on each line card that is responsible for allowing each port on each PINNACLE access to the switching bus. This local arbiter signals the central arbiter (on the Supervisor Engine), which is responsible for allowing each local arbiter to allow frames onto the switching bus.

Step 2—Packet sent across switching fabric and lookup takes place

The switching bus on the Catalyst 6000 is a shared medium, meaning that all the ports on the switch see the packet as it is on the bus. Once the central arbiter has granted access to the switching fabric, the packet is sent across the bus. All ports begin downloading that packet into their transmit buffers. The PFC is also looking at the switching bus, sees the frame and initiates a lookup. First, the Layer 2 table is consulted. If the packet is local to the VLAN, a switching decision is finished. If the Layer 2 table is the router's MAC address, then the Layer 3 Engine examines the packet and determines if a forwarding entry exists in hardware. If not, the packet is sent to the MSFC. If an entry does exist, the destination VLAN is identified and the Layer 2 lookup table is consulted again, this time to determine the MAC address within the VLAN and is associated outbound port.

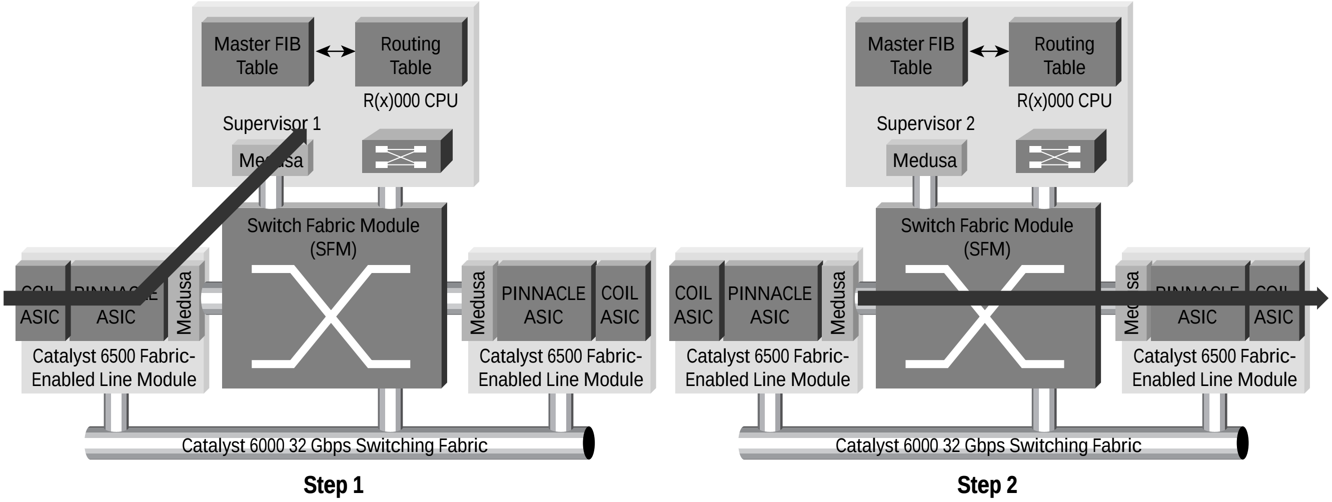
Step 3—Forwarding the packet

Now that the outgoing interface has been identified, all the ports on the switch that are not the destination port are told, over the Results Bus, to flush their buffers of that packet. The Results Bus also carries to the destination port the MAC re-write information and the appropriate QoS parameters to use (so the packet can be queued correctly on the outgoing port). Once the destination port has received the packet, PINNACLE queues the packet in the correct queue and then uses the SP/WRR scheduler to switch the frame out of memory to the destination external to the switch.

Catalyst 6500 and the SFM with Centralized Switching

This section details how packets actually flow through the Catalyst 6500 equipped with the Switch Fabric Module. Refer to the diagram below for a step-by-step process.

Figure 14 Catalyst 6000 and the SFM with Centralized Switching Packet Flow


Master FIB Table
Routing Table
R(x)000 CPU
Supervisor 1
Medusa
Switch Fabric Module (SFM)
COIL ASIC
PINNACLE ASIC
Medusa
Catalyst 6500 Fabric-Enabled Line Module
Medusa
PINNACLE ASIC
COIL ASIC
Catalyst 6500 Fabric-Enabled Line Module
Catalyst 6000 32 Gbps Switching Fabric
Step 1
Master FIB Table
Routing Table
R(x)000 CPU
Supervisor 2
Medusa
Switch Fabric Module (SFM)
COIL ASIC
PINNACLE ASIC
Medusa
Catalyst 6500 Fabric-Enabled Line Module
Medusa
PINNACLE ASIC
COIL ASIC
Catalyst 6500 Fabric-Enabled Line Module
Catalyst 6000 32 Gbps Switching Fabric
Step 2

Step 1a—Handling the frame on the local line card.

A fabric-enabled system is different from the Catalyst 6000 bus-based system, however, there are remarkable similarities. Each SFM-enabled line card can be thought of conceptually as a Catalyst 6000 on a line card. The switching functionality, therefore, is fairly similar. The packet first enters the switch and its handled by PINNACLE the same way as in the Catalyst 6000 system. Arbitration is required on the local 16 Gbps bus and, while the packet is on the local bus, the header information required for look-up is parsed by the Medusa ASIC, compressed, and sent across the 32-Gbps bus to the Supervisor Engine. All of the Medusa ASICs on all the line cards see that frame and download the information.

Step 1b—Packet Lookup.

The packet is received by Supervisor Engine 2 and is presented to the PFC-2, which includes both the Layer 2 and 3 forwarding tables. Like PFC-1, a Layer 2 lookup is performed to determine whether a Layer 3 switching decision needs to be made. If a Layer 3 decision is needed, the header information is looked up by the Layer 3 engine, which is utilizing the CEF table. Once a destination VLAN is identified, a second Layer 2 lookup is performed to determine the correct MAC address. This information is then sent across the results bus, which, again, all line cards Medusas see. All Medusas except the destination drop the frame from its buffers.

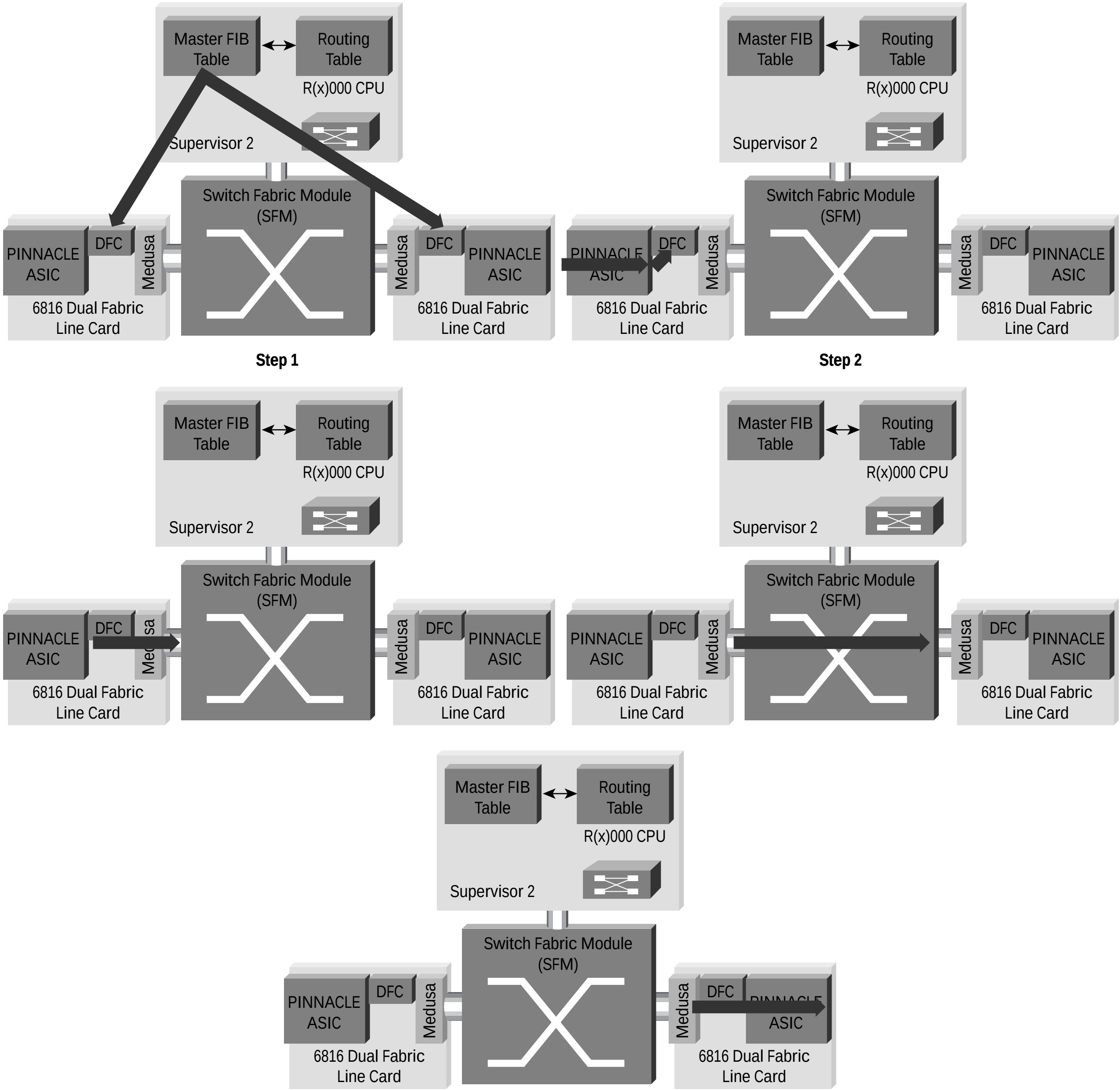
Step 2—Forwarding the packet.

The source line card now knows where the destination is. The line card, via the crossbar interface, adds a tag, which identifies the destination line card and sends the packet into the SFM. The SFM switches the frame to the appropriate destination line card. The information on the Results bus informs the destination line card what the destination port is and what the port of exit is. The packet is queued, for QoS, the same way it is in the Catalyst 6000 system. The WRR scheduler then sends the frame out to the network.

Catalyst 6500 and the SFM with Distributed Switching

This section details how packets flow through the Catalyst 6500 equipped with the SFM and distributed switching. Refer to the following diagram for a step-by-step process.

Figure 15 Distributed Forwarding Packet Switching



Step 1—Downloading the CEF table

The first step in the distributed switching model of the Catalyst 6500 is to calculate the CEF table and download that to the line cards. As stated earlier in this paper, the CEF table is calculated based on the entries in the routing table. This table is computed centrally at the MSFC-2 on the Supervisor Engine and downloaded to the PFC-2 and DFC (or integrated CEF table). Therefore, the local and central CEF tables contain the same information.

Step 2—Packet lookups

When a packet enters the switch, it is handled by PINNACLE and arbitration is requested for the local switching bus. All ports on the local bus see that frame, including the DFC. The DFC performs a lookup in the local table and identifies whether the destination is local to the line card or across the switching fabric.

Step 3—Switching the packet to the SFM

If the destination is across the SFM, the DFC tells the SFM interface controller (called Medusa) to prepend a tag on the packet identifying the exit "port" on the SFM.

Step 4—Packet switching in the SFM

Once the packet is received by the SFM, the SFM examines the tag prepended to the packet and makes its own switching decision. Remember that the fabric uses 3x overspeed, so, although the inputs to the SFM is 8 Gbps, internal switching takes place at 24 Gbps. The SFM identifies the outgoing port and switches the frame to the Medusa ASIC on the outgoing line card.

Step 5—Switching the frame to the outbound port

The Medusa on the outgoing port takes the frame out of the SFM and places it on the switching bus. Since a switching decision has already been made, the local data bus and results bus are driven with data at the same time. The data bus broadcasts the frame and the results bus indicates what the destination port is. The information on the Results bus informs the destination line card what the destination port is and what the port of exit is. The packet is queued, for QoS, the same way it is in the Catalyst 6000 system. The WRR scheduler then sends the frame out to the network.

Conclusion

The Catalyst 6000 Family, including the Catalyst 6000 and 6500, is an advanced multilayer switching system capable of switching at speeds of more than 100 million pps. Subsequent documents will review the actual performance capabilities tested and verified in the lab. By incorporating Layer 2–4 capabilities in hardware, the system delivers high performance, superior functionality, and high availability to enterprise and service provider networks.

CISCO SYSTEMS

Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters
Cisco Systems Europe
11, Rue Camille Desmoulins
92782 Issy-les-Moulineaux
Cedex 9
France
www-europe.cisco.com
Tel: 33 1 58 04 60 00
Fax: 33 1 58 04 61 00

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

Asia Pacific Headquarters
Cisco Systems Australia, Pty., Ltd
Level 9, 80 Pacific Highway
P.O. Box 469
North Sydney
NSW 2060 Australia
www.cisco.com
Tel: +61 2 8448 7100
Fax: +61 2 9957 4350

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on the

Cisco Web site at www.cisco.com/go/offices

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia
Czech Republic • Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia
Ireland • Israel • Italy • Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru
Philippines • Poland • Portugal • Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa
Spain • Sweden • Switzerland • Taiwan • Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

07/01 LW2338