

Catalyst 8500 CSR Architecture

Overview

The Catalyst® 8500 family of Campus Switch Routers provide the next-generation switching paradigm for both switching performance and critical network services. Including support for all IP and IPX routing standards, the Catalyst 8500 family of backbone switches offer a more complete suite of sophisticated features than any other Layer 3 switches in their class. The Catalyst 8500 family contains several key components of the new enterprise campus network architecture, providing for high-speed performance, resiliency, and quality of service (QoS) within the network backbone. By incorporating Cisco IOS® technology, the Catalyst 8500 family provides seamless integration with the Catalyst 5000 family (including the Route Switch Module [RSM] and NetFlow feature card), as well as the Cisco 7500 routers for WAN access.

This paper will focus on the Catalyst 8500 Campus Switch Router. Unlike many other switches, which are only now stabilizing their software, the Catalyst 8500 builds on years of expertise in developing complex campus LAN and WAN routers. The Catalyst 8500 family offers a complete set of campus-required features such as Hot Standby Router Protocol (HSRP), Protocol Independent Multicast (PIM) for multicast routing, and a full set of routing protocols, including Routing Information Protocol (RIP), RIP version 2, Open Shortest Path First (OSPF), Interior Gateway Routing Protocol (IGRP), Enhanced IGRP, and Border Gateway Protocol (BGP) -4 (future). As QoS mechanisms become more important in the campus, the

Catalyst 8500 series provides, in its first phase, four queues per port based on IP precedence. In the following phase, the Catalyst 8500 family will provide support for per-flow queuing, a much more granular method of determining QoS and ensuring minimal latency and packet loss for mission-critical applications.

The Catalyst 8500 delivers the performance, scalability, and robustness required for network production backbones. The Catalyst 8510 family delivers wire-speed IP and IPX performance for all ports, scaling to 6 million pps. The Catalyst 8540 delivers the same wire speed functionality for IP and IPX, scaling to 24 million pps. With the Catalyst 8500 family, Cisco ensures investment protection and migration to new technologies. The Catalyst 8510 offers redundant power supply modules and uses the same power supplies used for the Catalyst 5000 and LightStream 1010 chassis. In addition, switching modules from the Catalyst 8510 can be used in the bottom five slots of the Catalyst 5500 switch, ensuring customers who have purchased Catalyst 5500 switches a clear migration to high-speed Layer 3 switching and QoS, if required. The Catalyst 8540 provides a high level of system redundancy, providing for redundant switch fabrics, route processors and power supply modules.

Flexible Chassis Options

Catalyst 8510

The Catalyst 8510 is a five-slot chassis that supports up to 32 ports of 10/100 Fast Ethernet connectivity or four ports of Gigabit Ethernet capacity for uplink or server connections. The Catalyst 8510 provides 10 Gbps of nonblocking switching capacity for

both Layer 2 and Layer 3 switching. Because the Catalyst 8500 family of switches is ATM-capable, future releases of the Catalyst 8510 software and line cards will support ATM uplink capacity for OC-3, OC-12, and Packet over SONET (PoS) connectivity. These ATM modules will offer the capability for LAN Emulation (LANE) and RFC 1483. The Catalyst 8510 switch route processor will populate one slot, allowing the remaining four slots to be used for connectivity modules. All modules for the Catalyst 8510 will be forward-compatible in the Catalyst 8540 switch.

Catalyst 5500

The Catalyst 5500 switch provides full integration of the Catalyst 8510 line cards by using the secondary passive backplane fabric in slots 9 through 12. Slot 13 in the Catalyst 5500 is used for either the LightStream 1010 ATM switch processor or the Catalyst 8510 Switch Route Processor (SRP) module. By using the additional switching fabric available in every shipping Catalyst 5500, Cisco is able to maintain investment protection for Catalyst 5500 customers by enabling them to use the high-speed Layer 3 forwarding and routing capability of the Catalyst 8510. (Currently, the Catalyst 8510 and LightStream 1010 line cards can not be used simultaneously.) Like the Catalyst 8510, the Catalyst 5500 fabric provides support for 10 Gbps of nonblocking switching capacity.

Catalyst 8540

The 13 slot Catalyst 8540 CSR provides a 40 Gbps non-blocking switching fabric with performance scaling to 24 million pps. Like the Catalyst 8510, the Catalyst 8540 provides support for Layer 2 bridging, Layer 3 routing as well as future capabilities for ATM uplink capacity. The Catalyst 8540 Switch Processor requires two slots, with a third slot for redundancy; should either of the Switch Processors fail, the third will take over. One slot is required for the Route Processor, which handles system management and control plane functions. A second Route Processor slot is reserved for redundancy. The remaining eight slots are used for connectivity modules.

Hardware Switching Features

Catalyst 8510 Switch Route Processor Module

The Switch Route Processor module provides the intelligence to the Catalyst 8510, interfacing to each port via the switch fabric. The SRP module runs the Cisco IOS software for high-speed, Layer 3 switching, including the Cisco Express Forwarding table, routing protocol control, and dynamic multicast. Also supported on the SRP module are the Simple Network Management Protocol (SNMP) management agent and the many Management

Information Bases (MIBs) used for the management of the device, as well as in the future, integrated management applications for advanced traffic management.

In order to support such capabilities, the SRP module uses a powerful, 100-MHz MIPS R4600 processing subsystem. Based on the RSP design used in the Cisco 7500 series of multiprotocol routers, together with its advanced memory management application-specific integrated circuits (ASICs), the SRP module can execute more than 100 million instructions per second. The use of such an advanced processor delivers not only the high performance required to operate such computationally intensive protocols as routing calculation and distribution, but also facilitates the operation of the Cisco IOS software. By taking advantage of the Cisco IOS software features, the Catalyst 8500 can exploit many of the sophisticated capabilities already found on Cisco's range of multiprotocol routers while delivering a familiar user interface to users of other Cisco products.

The SRP module supports a dual-height PCMCIA Type II slot, which can be used to support a variety of Flash EPROM modules adding from 8 MB to 20 MB of additional memory. These modules are necessary to support larger code images as the sophistication of the Cisco IOS software grows. Flash cards can also be used to program switches with standard configuration parameters. Note, however, that the Flash cards are not required for the operation of the Catalyst 8510.

Catalyst 8540 Switch Processor/Route Processor

The switch processing and route processing in the Catalyst 8540 are provided on two separate modules. The two switching fabric modules provide the shared memory architecture used for transporting frames from one interface to another. This fabric provides a full 40 Gbps of non-blocking switching connectivity. The Route Processor features a 200 MHz MIPS R5000 processor. This high performance CPU will, like the CPU on the Catalyst 8510, run the Cisco IOS software for high-speed, Layer 3 switching, including the Cisco Express Forwarding table, routing protocol control, and dynamic multicast. In addition, the Simple Network Management Protocol (SNMP) management agent and the many Management Information Bases (MIBs) used for the management of the device, as well as in the future, integrated management applications for advanced traffic management are run in the Route Processor. It is important to remember that two Switch Processors and one Route Processor are required for the Catalyst 8540 system.

Catalyst 8500 Series Switching Modules

Each of the interface modules connects into the Catalyst 8500 fabric. The modules from the Catalyst 8510 can be used in the future Catalyst 8540; however, the higher-density modules, developed specifically for the Catalyst 8540, cannot be deployed in the Catalyst 8510. There are no slot dependencies for the line cards. The fan tray is replaceable while the switch is operational, reducing the mean time to repair.

The Catalyst 8510 switch provides for two versions of line cards. One offers eight ports of 10/100 Fast Ethernet over Category 5 copper cable with RJ-45 connectors. The second offers eight ports of 100BaseFX over fiber-optic cable with SC fiber connectors. The Catalyst 8540 offers two versions of the line cards as well: a 16-port 10/100 Fast Ethernet over copper modules, and a 16-port 100Base-FX module. Each line card has the option for 16,000 or 64,000 table entries. This option translates into 16,000 or 64,000 routes and/or Layer 2 MAC addresses that can be stored locally within the line card. For most campus implementations, the 16,000 entry cards will provide sufficient address space. For Internet service providers (ISPs) who may deploy the Catalyst 8500 CSR with the future ATM or PoS modules, the higher number of table entries will usually be required.

The Catalyst 8510 will offer a one-port Gigabit Ethernet module for delivery in the Summer, 1998 timeframe. In the future, ATM uplink modules supporting RFC 1483, LANE, and Multiprotocol Label Switching (MPLS) will be available. The Catalyst 8540 will offer a two-port Gigabit Ethernet module towards the end of the fourth quarter, 1998. The Catalyst 8540 will also offer ATM and Packet over SONET uplinks for OC-3, OC-12 and OC-48 interfaces.

Cisco Express Forwarding

The Catalyst 8500 features Cisco Express Forwarding (CEF), the routing technology developed and implemented in the Cisco 12000 gigabit switch router (GSR). This technology offers a new paradigm for route distribution and forwarding by distributing routing information from the central processor to the individual line modules. This technology, used within the Internet, provides for scalability in large campus core networks. CEF provides Layer 3 forwarding based on a "shadow" of the routing table, resulting in very high-speed routing table lookups and forwarding. This feature provides for wire-speed IP and IPX forwarding for all ports in the Catalyst 8500 family.

This technology will be discussed in greater detail later in this paper.

Redundancy

The Catalyst 8500 features both hardware-level and network redundancy. The chassis features hot-swappable, redundant power supply modules as well as hot-swappable line cards. The redundancy capabilities of the Cisco IOS software allows for key network features such as Hot Standby Router Protocol (HSRP); routing protocol convergence via RIP, OSPF, or Enhanced IGRP; Fast EtherChannel; and load sharing across equal-cost Layer 3 paths and Spanning Tree (for Layer 2-based networks). These features allow network managers to build resilient networks not just from the hardware perspective, but from the software side as well.

Catalyst 8500 System Architecture

The Catalyst 8510 is based on a 10-Gbps, shared-memory fabric. The Catalyst 8540 is based on a 40-Gbps shared memory fabric. These fabrics provide for full non-blocking switching capacity for any port configuration, including Ethernet, Fast Ethernet, Gigabit Ethernet, or 155-Mbps/622-Mbps ATM. The Catalyst 850 CSR is designed as a “distributed switching system,” meaning that all line cards in the system work closely with the system processor to ensure current Layer 2 and Layer 3 address and routing information. The system processor is the “manager” of the system, providing for Layer 3 route calculation and Forwarding Information Base (FIB) distribution to the line cards, Layer 2 MAC address learning and distribution to the line cards, and system management. Figure 1 provides a high-level overview of the switching architecture. Keep in mind that, despite the bandwidth and memory differences between the Catalyst 8510 and Catalyst 8540, the functions of the architecture are identical.

The best way to understand the architecture of the Catalyst 8500 is to divide the switch into three distinct, functional segments: the switch fabric, switch line cards, and the processor engine.

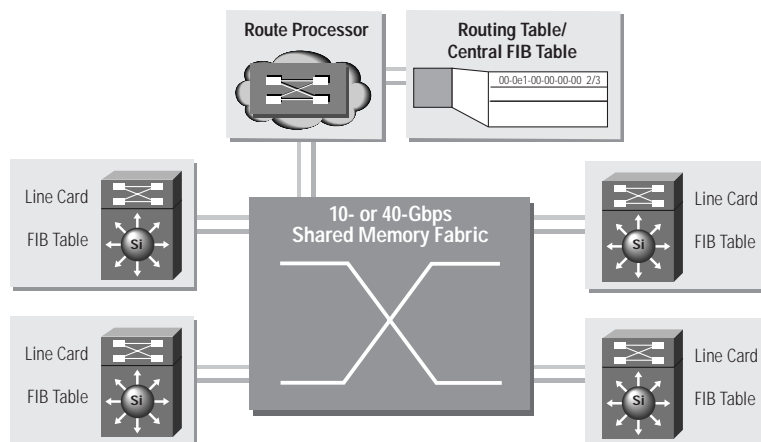
The Catalyst 8510 comprises 3 megabytes of shared-memory switching fabric. The Catalyst 8540 comprises a 12 megabyte shared memory fabric. This memory is dynamic, meaning that a packet stored in memory takes only as much memory as it needs. Access into and out of the shared memory is dynamically allocated by the direct memory access (DMA) ASIC. Because the switch fabric is non-blocking, per-port buffers are not

required; the fabric speed is faster than the combined speed of all the ports. Congestion will therefore only occur when an individual output port is congested.

The line cards are designed to carry considerable intelligence for the switching system. Each line card contains ASICs designed to provide input and output into the fabric as well as to maintain a Layer 3 FIB or a Layer 2 MAC address table. These tables allow the Catalyst 8500 system to make switching decisions very quickly prior to transmission across the switching fabric. The line cards, therefore, must work closely with the system processor to ensure that all address tables and routing information is current. The line cards are also responsible for presenting a uniform frame to the switching fabric for effective buffering, QoS policy enforcement, and packet switching.

The Processor Engine is responsible for all address and route learning and distribution. Because the Catalyst 8500 is designed as a distributed switching system, the system processor (CPU) needs to ensure that all Layer 3 routes and Layer 2 MAC addresses are maintained and update the line cards as needed. The system processor is also responsible for handling all system management, including SNMP and remote monitoring (RMON) statistics.

Figure 1 High-Level Catalyst 8500 CSR Architecture



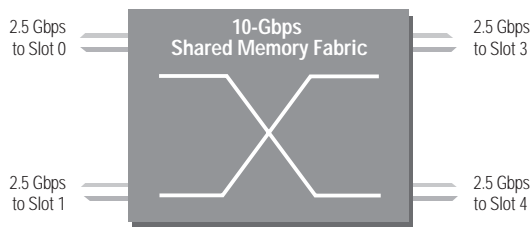
Switching Fabric and Arbitration

Shared Memory

The Catalyst 8510 is based on a 3-megabyte shared-memory architecture with a total system bandwidth of 10 Gbps. The Catalyst 8540 is based on a 12-megabyte shared memory architecture with a total system bandwidth of 40 Gbps. It is a completely non-blocking switch, meaning that all input ports have equal and full access into the shared memory for packet switching. The Catalyst 8500 also provides four queues per port, allowing the Frame Scheduler to make intelligent QoS decisions based on the priority of each queue.

Each of the line cards in the Catalyst 8510 is allotted 2.5 Gbps of capacity into the fabric (as shown in Figure 2). This allows for non-blocking switching capacity within the switching system by ensuring that each slot is given more bandwidth than all of the ports on the line card can generate. The 2.5-Gbps bandwidth is divided into transmit and receive paths, each of 1.25 Gbps, to ensure that both reads and writes to the shared memory can be accomplished simultaneously. In the Catalyst 8540, each slot has 5-Gbps into the shared memory fabric. This bandwidth is also divided into 2.5 Gbps transmit and 2.5 Gbps receive paths into the fabric.

Figure 2 Switching Bandwidth per Slot on Catalyst 8510



Because of the Catalyst 8500's non-blocking nature, every port in the switch has full access to every other port. Each packet entering the switch fabric is tagged with an internal routing tag. This routing tag provides the switching fabric with the appropriate port of exit information, the QoS priority queue the packet is to be stored in, and the drop priority. The Fabric-Switching ASIC (FSA) queues each packet into memory and creates a pointer, based on the internal routing tag, to the appropriate destination port. The Frame Scheduler is then responsible for scheduling the frame out of memory based on the queue where the packet is being stored.

Each port transmitting through the fabric is, by default, placed in the lowest-priority queue. This places all traffic at a "best-effort" QoS level. When a network manager configures a

policy, that traffic is transmitted in the queue corresponding to the specified IP precedence. That queue is granted more service, thereby reducing latency and the possibility that traffic on that queue will be dropped. All management and control plane traffic, such as BDPUs, routing protocol updates, and management frames are placed in the highest-priority queue for transmission to the CPU.

The Frame Scheduler

The Frame Scheduler has two main responsibilities within the Catalyst 8500: first, to schedule frames into the switching fabric based on the priority queue being requested, and second, to schedule frames out of the switching fabric based on the Weighted Round Robin (WRR) scheduling algorithm. Note that the QoS implementation will be discussed in detail later in this paper.

At the input to the switching fabric, the CEF ASIC posts a request to the Frame Scheduler for access to the fabric. The Frame Scheduler handles each request in a time-division multiplexing (TDM) fashion, meaning that each CEF ASIC will have the opportunity to clock an entire frame into the fabric when access has been granted. Because each CEF ASIC handles four ports, the Frame Scheduler allows the CEF ASIC to clock in a maximum of four packets into memory (the CEF ASIC will be discussed in the next section). Each packet in memory has been prepended with an internal routing tag which, as mentioned earlier, contains the port of exit, queuing priority, and drop priority. Based on the routing tag, the input Frame Scheduler places the packet in the correct queue (see Figure 3).

Figure 2b Switching Bandwidth per Slot on Catalyst 8540

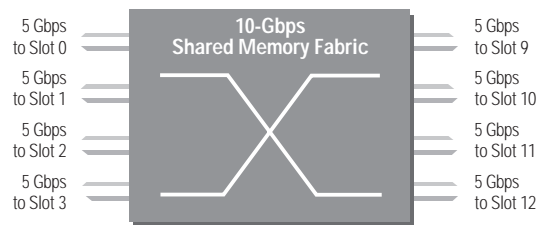
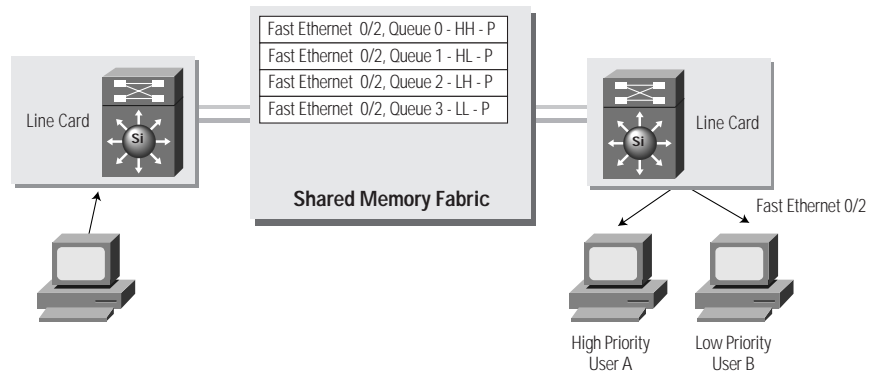


Figure 3 Input Scheduling and Queue Allocation



The “HH,” “HL,” “LH,” and “LL” designations refer to the IP precedence fields used by the Catalyst 8500 to determine the appropriate queue. Although not shown, a fifth, critical high-priority routing tag is prepended to all management and control plane packets for immediate delivery to the CPU.

On the output side, the Frame Scheduler is responsible for servicing each queue based on the WRR priority scheme. WRR allows the network manager to configure how much service each queue will receive. In a situation where there is no congestion, WRR and the weights provided do not play a real part on how packets are switched out of the fabric, because there is plenty of bandwidth available. However, if a link is congested, WRR services each queue per port based on the priority set by the network manager. Consider, for example, the weights assigned by a network manager in Table 1.

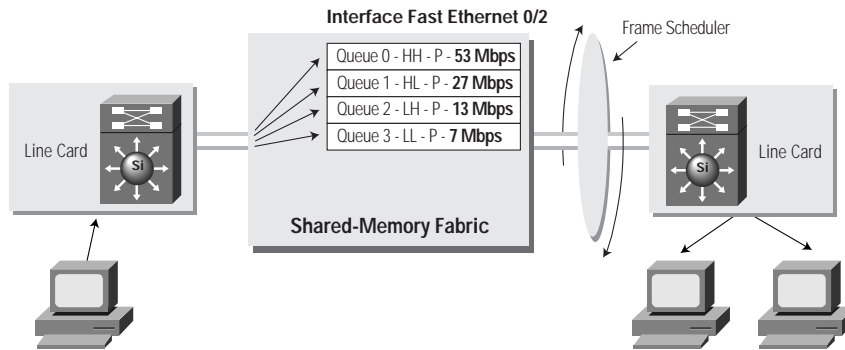
Based on these priorities and weights provided, the Frame Scheduler services QoS-0 more often, granting that queue 53 Mbps out of the 100 Mbps possible on the output link. The second queue, QoS-1, receives 27 Mbps of the bandwidth, and so forth. These commands are set globally on the Catalyst 8510 and function the same for all ports on the switch.

The Catalyst 8510 also allows network managers to override the global QoS settings by allowing port-to-port communications to have a different level of priority. Network managers have the option of configuring bandwidth based on a source-destination, destination, or source basis and provide weights based on certain IP addresses having more bandwidth than others. This feature will be available with the Hardware Access List Daughter Card, to be introduced in the future.

Table 1 Sample WRR Priority Weights

Quality of Service Priority	Weight Given by Network Manager	Bandwidth Assignment Calculation	Bandwidth Assigned
QoS-0	8	$= (8 / (8 + 4 + 2 + 1)) \times 100$	53 Mbps
QoS-1	4	$= (4 / (8 + 4 + 2 + 1)) \times 100$	27 Mbps
QoS-2	2	$= (2 / (8 + 4 + 2 + 1)) \times 100$	13 Mbps
QoS-3	1	$= (1 / (8 + 4 + 2 + 1)) \times 100$	7 Mbps

Figure 4 WRR Scheduling and Bandwidth Allocation

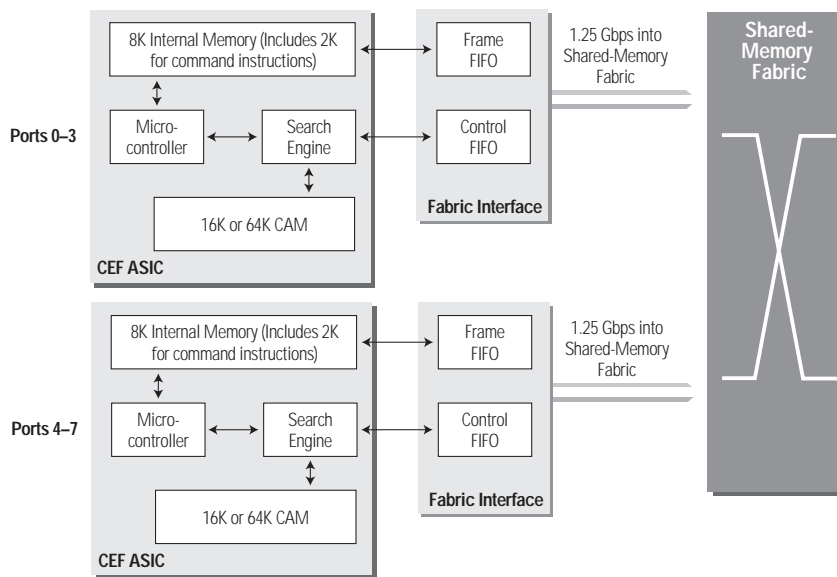


Line Card Architecture

The second major components of the Catalyst 8500 architecture are the line cards. Because the Catalyst 8500 uses a distributed architecture, the line cards must be intelligent enough to make both Layer 3 and Layer 2 forwarding decisions at wire speed for all media types, as well as enforce QoS policies. Figure 5 details the architecture of the Catalyst 8500 line cards. Note in Figure 5 that the Catalyst 8540 will utilize 4 CEFA per line card.

The Catalyst 8500 line cards are based on the Cisco Express Forwarding ASIC (CEFA). The CEF ASIC is based on the MMC Ethernet PIF (EPIF) ASIC. It is called the CEF ASIC since the Cisco Express Forwarding mechanism is programmed into the ASICs. This ASIC is responsible for the Ethernet MAC layer functions, address or network lookup in the CAM table, and forwarding of the packet with its correct rewrite information to the Fabric Interface. The Fabric Interface is also resident on the line card and is responsible for the packet rewrite, QoS classification, and signaling to the Frame Scheduler.

Figure 5 Catalyst 8500 Line Card Architecture



CEFA

The CEFA is at the heart of the line card architecture. This ASIC has several key components that will be discussed in detail. Each CEFA services four ports on the line card. In order to service eight ports, two CEFAs are used per line card. On the Catalyst 8540, four CEFAs are used in order to service 16 ports. Although not shown in Figure 5, the CEFA is responsible for all MAC layer functions. The MAC is 10/100 auto-sensing and auto-negotiating, if so configured. The MAC can also be run in either full or half duplex.

Packets entering the switch port and having passed through MAC functions are stored in an internal block of SRAM. This memory is 8 kilobytes in size, with 2K reserved for command instructions. This memory is used to hold the packet while the appropriate lookups take place.

The CEFA microcontroller is a mini-CPU that is local to four ports on the Catalyst 8500 line module. The microcontroller is designed to handle the traffic on each of the ports in a fair manner. This means the CEFA must ensure that all packets have equal access into internal memory and that lookups via the Search Engine are done fairly by arbitrating service between the four ports. This is handled in a round-robin manner, meaning that the microcontroller cycles between each port, processing requests as needed.

The microprocessor also has the critically important task of forwarding system messages such as Spanning Tree BPDUs, routing advertisements, Cisco Discovery Protocol (CDP) packets, Address Resolution Protocol (ARP) frames, and other control-type messages back to the central CPU. Those messages are forwarded by the CEFA to the CPU.

CEFA Search Engine

The Search Engine in the CEFA performs the address lookup or network output interface lookup. It performs its lookup in the content-addressable memory (CAM) table, which can hold either 16,000 or 64,000 entries. The Search Engine can make two types of switching decisions: Layer 2-based or Layer 3-based. With the hardware-based access list feature card (to be discussed later), the Search Engine can also perform lookups based on Layer 4 information. The Search Engine is therefore responsible for maintaining the Layer 2 MAC address table and the Layer 3 FIB.

An incoming packet is placed into the internal memory. As soon as the first 64 bytes of the frame are read into memory, the microcode signals the Search Engine with the relevant source or destination MAC address, destination network, or (in the future) Layer 4 port information. The Search Engine can then conduct a lookup in the CAM table for the corresponding entry. Using a

binary tree lookup method, the Search Engine can hit a MAC address or perform a longest match on the destination network address very quickly. The corresponding rewrite information, which is stored in the CAM table, is then delivered to the control FIFO of the Fabric Interface.

As stated earlier, the CAM table comes in two options: one supporting 16,000 entries and the second supporting 64,000 entries. In order to cycle through the number of entries quickly, a binary tree lookup is done in order to hit on the matching address and deliver relevant destination and rewrite information to the Fabric Interface. When the microcontroller signals the Search Engine with an entry to match, the address or network is presented in binary form. The Search Engine then cycles through the binary tree until the address is matched or the longest match to the destination network is reached.

Fabric Interface

The final stage in packet switching within the Catalyst 8500 CSR can now occur. The switching CEFA now knows the port-of-exit for the packet based either on its MAC address or on the Layer 3 IP or IPX network numbers. The packet must now be transferred across the switching fabric to the destination. The Fabric Interface is responsible for preparing the packet for its journey across the switching fabric.

The Fabric Interface consists of two main components: the frame FIFO and the control FIFO. As can be seen from Figure 5, the internal memory of the CEFA has a direct connection into the frame FIFO, and the Search Engine has a direct connection into the control FIFO. When the Search Engine completes the lookup, the packet moves from internal memory into the frame FIFO. In parallel, the Search Engine returns to the control FIFO all of the relevant rewrite and QoS information.

The Fabric Interface then rewrites the packet with the appropriate information and calculates the checksum. At the same time, the Fabric Interface prepends the internal routing tag containing port of exit, the QoS priority, and drop priority, onto the packet. Once completed, the Frame Scheduler is signaled to place the frame into the fabric.

At the output port, the Fabric Interface forwards the packet to its output MAC. Since all rewrite and error checking has been done at the ingress port, no additional work needs to be performed on that frame.

Switch Route Processor

The system processor is the final element of the Catalyst 8510 architecture and resides at the core of the switch. The system processor resides on the SRP module, along with the shared

memory fabric. The system processor CPU is a 64-bit 100MHZ R4600 RISC processor. Its architecture is very similar to that of the Cisco 7500 Route Switch Processor (RSP). The Route Processor for the Catalyst 8540 is a 200MHZ R5000 RISC processor, very similar to the RSP-4 engine. The software running on the Catalyst 8500 SRP is the Cisco IOS Version 12.0.

The system processor is responsible for all system-level switching and management functions, such as running the routing protocols, maintaining the routing table and Cisco Express Forwarding FIB table, and the Spanning Tree configuration. Each CEFA is responsible for identifying these frames and for sending them to the CPU. In designing a distributed switching system, it was critical to separate the control and data switching planes. This allows the CPU to handle all of the control plane activities and the switching fabric and line cards to handle the data forwarding.

Routing Protocols

The CPU is responsible for running all of the Catalyst 8500's routing protocols. The Catalyst 8500 provides support for IP and IPX forwarding and routing. IP routing support includes RIP versions 1 and 2, OSPF, IGRP and Enhanced IGRP. IPX routing support includes RIP and Enhanced IGRP. (NLSP will be included in a future release.) Other protocols, such as AppleTalk, DECNet, and VINES are bridged in the Catalyst 8510. Future routing protocol support includes BGP-4. The Catalyst 8540 is designed to support multiprotocol routing in its second software release.

The CPU is also responsible for maintaining state information regarding multicast routing. The Catalyst 8500 supports PIM (sparse mode and dense mode) as well as Distance Vector multicast Routing Protocol (DVMRP) interoperability. The CPU is responsible for responding to and forwarding joins and leaves as well as responding to pruning messages sent by PIM. Multicast forwarding takes place at the line card level.

Most importantly, the CPU is responsible for maintaining the routing table. By using Cisco Express Forwarding, the CPU creates a FIB, which contains a subset of the routing table. The FIB is based on a topology map of the network, allowing routing to take place via the network topology at high speed. The FIB is then downloaded to the line cards, allowing them to make Layer 3 routing decisions locally without having to interrupt the CPU. This capability allows the Catalyst 8500 to forward all frames at wire speed for all ports. The FIB and Cisco Express Forwarding will be discussed later in this paper.

Layer 2 VLAN and Switching

Although the switching decisions are made at the line cards, the CPU is still responsible for maintaining Layer 2 information. The CPU is responsible for bridge group configuration and Spanning

Tree calculation. Bridge groups are configured on the Catalyst 8500 in the same way they are in the other Cisco routers. Instead of routing traffic to an outgoing interface, the traffic is bridged via its Layer 2 address. Integrated Routing and Bridging (IRB) is also supported in the Catalyst 8500 in order to support both bridging and routing at the same time. The CPU is also responsible for maintaining all Spanning Tree information within the switch. This includes calculation of the root bridge, optimum path determination to the root, and determining the forwarding and blocking links.

Cisco Express Forwarding

CEF evolved to best accommodate the changing network dynamics and traffic characteristics resulting from increasing numbers of short-duration flows typically associated with Web-based applications and interactive multimedia sessions. Existing Layer 3 switching paradigms use a route-cache model to maintain a fast lookup table for destination network prefixes. The route-cache entries are traffic driven, in that the first packet to a new destination is routed via routing table information, and as part of that forwarding operation, a route-cache entry for that destination is added. This process allows subsequent packets flows to that same destination network to be switched based on an efficient route-cache match. These entries are periodically aged out to keep the route cache current and can be immediately invalidated if the network topology changes.

This "demand-caching" scheme—maintaining a very fast access subset of the routing topology information—is optimized for scenarios whereby the majority of traffic flows are associated with a subset of destinations. However, given that traffic profiles at the core of the Internet (and potentially within some large enterprise networks) no longer resemble this model, a new switching paradigm was required that would eliminate the increasing cache maintenance resulting from growing numbers of topologically dispersed destinations and dynamic network changes.

CEF avoids the potential overhead of continuous cache churn by instead using a FIB for the destination switching decision. The FIB mirrors the entire contents of the IP and IPX routing table. This means that there is a one-to-one correspondence between FIB table entries and routing table prefixes; therefore, there is no need to maintain a route cache. Note that although CEF has been specified for IP, it also applies to IPX as well, which will be described in subsequent sections.

CEF Operation

CEF provides features comparable to fast switching, including load sharing, recursive route resolution, and access lists. CEF uses two tables that are maintained in the SRP and downloaded to the

line cards: the FIB and adjacency table. The FIB database is a “shadow” of the routing table and is used for making forwarding decisions. The adjacency table maintains the adjacent nodes, and the link-layer information (such as packet rewrite information) necessary to reach that adjacent node. Every entry in the FIB table has a pointer to a corresponding entry in the adjacency table.

The FIB table is populated by callbacks (inputs) from the routing table. After a route is resolved, it points to a next hop, which should be an adjacency. This step is done at the SRP and then downloaded to the line cards, allowing the line cards to maintain a current topology of the network, which enables rapid switching decisions (within 10 microseconds) as well as fast convergence in the event of a routing topology change. The FIB is modified when a route is added, removed, or changed in the routing table. This information is immediately downloaded to the line cards.

The adjacency table is also populated by callbacks from the routing protocols, which include information such as next-hop information and (S,G) interfaces for multicast groups. Adjacencies are added when a protocol detects that there is an adjacent node via the routing protocol. When a packet arrives at the ingress port, the CEF ASIC performs a FIB lookup based on the destination IP address. The matching FIB entry points to an adjacency entry, which in turn provides the valid link layer rewrite and outgoing interface. The packet is forwarded based on this information. Figure 6 shows the relation of the FIB to the adjacency table.

CEF Performance

CEF implements a Cisco patent-pending expedited lookup and forwarding algorithm to deliver maximum Layer 3 switching performance. Additionally, Express Forwarding is less CPU intensive than route caching because the switching decisions are

outbound interface, and a cache entry is added for that destination. Because the cache information is derived from the routing table, routing changes cause existing cache entries to be invalidated and then reestablished to reflect any topology changes. In networking environments that frequently experience significant routing activity (such as the Internet backbone), this can cause traffic to be forwarded via the routing table (Process-Level Switching) as opposed to via the route cache (Fast Switching). During major network convergence or flux, performance can thus be suboptimal.

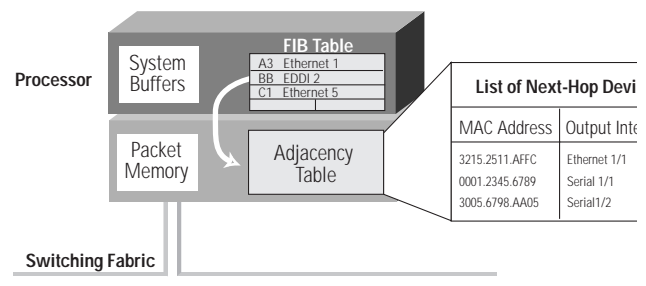
Express Forwarding obviates this Process Switching/Fast Switching scenario, and because the FIB is topology-driven rather than traffic-driven, CEF’s switching performance is largely independent of and unaffected by network size or dynamics. This makes the Catalyst 8500 ideal as a core Campus Switch Router placed at key distribution points in the network or in the network core itself.

made by each line card in the Catalyst 8500. CEF in the Catalyst 8500 allows for high-speed forwarding (wire speed on all ports) and low latency. Latency on the Catalyst 8510 has been measured at 38 microseconds (last-in, first-out measurement).

One of the key benefits of CEF in the Catalyst 8500 is its routing convergence. Because the FIB is distributed to all line cards, any time a route goes away, flaps, or is added, the FIB is able to update that information and provide it to the line cards. This means that CPU processor interrupts are minimized, because there is no route cache to invalidate and relearn. Line cards are able to receive the new topology quickly and reconverge around a failed link based on the routing protocol being used.

Layer 3 Forwarding performance on the Catalyst 8510 is over 5 million pps for IP. Performance for the 8540 is approximately 24 million pps for IP and IPX.

Figure 6 FIB and Adjacency Table



CEF Resilience

CEF in the Catalyst 8500 offers an unprecedented level of switching consistency and stability in large, dynamic networks. This high level results from the fact that the FIB lookup table contains all known routes, therefore eliminating the potential for “cache-misses” that occur with demand caching schemes. For example, if a route is not found in the forwarding cache, the first packet(s) then looks to the routing table to determine the

IPX Forwarding Information Base

Although not specified in the original CEF specification, support has been included in CEF for IPX in the Catalyst 8500. The functional elements of the IPX-based FIB are very similar in operation to the IP-based FIB described above. The structure of the FIB is the same, meaning that there is a FIB and an adjacency table for IPX (separate from the IP FIB and adjacency table). This information also exists on the line cards.

Each entry in the FIB table corresponds to a particular IPX network number. During IPX switching, the microcode on the CEFA will key a search off the destination network number, which will result in a “hit.” The hit will return two key pieces of information related to the destination network number: any flags that indicate Internal Network or Directly Connected network; and the next-hop pointer (NHP). It is the NHP that points to the entry in the adjacency table.

The adjacency table in IPX is known as the Node Address Table (NAT). Each entry in the IPX FIB has a pointer into the NAT. Located in this table are the node addresses for given destination IPX networks. Each address in the node table results into an entry containing encapsulation information (whether the address belongs to its own interface node address) and a pointer to the connected network prefix. All lookups are done at the ingress port and sent with the appropriate internal routing tag to the switching fabric.

IPX performance is over 5 million packets per second across 32 ports of 10/100.

Packet Switching in the Catalyst 8500

Now that the components of the Catalyst 8500 have been explained, it becomes much easier to understand how packets are switched across the fabric.

Layer 3 Forwarding

By using CEF, each of the line cards maintains a “shadow” of the routing table, known as the FIB. As stated earlier, the FIB consists of two parts: the FIB itself, which contains the network topology map, and the adjacency table, which contains the packet rewrite information. As networks are configured, the routing table is created and the FIB is populated. The FIB is then downloaded to the line cards. Any changes made to the routing table, caused by additions or deletions of routes or route flaps, are updated in the central FIB, which in turn updates the line cards. This means that at all times, all line cards have a correct map of the network topology.

Packet switching in the Catalyst 8500 takes place as follows:

1. A packet is received at the physical interface. The CEFA ASIC provides the MAC-layer functions, and the packet is stored in internal memory
2. As soon as the first 64 bytes of the frame are read, the microcode running on the microcontroller reads the source and destination IP addresses or IPX network information. If the destination address is the router’s MAC address, the packet is routed. If not, it is bridged
3. The information is used by the Search Engine to begin a lookup in the CAM table for the longest match entry
4. The destination network is matched within 64 clocks (or approximately 2.5 microseconds). The match is returned to the microcontroller, which in turn moves the frame from the internal memory to the Fabric Interface’s frame FIFO. At the same time, the Search Engine returns relevant information such as QoS classifications and MAC header rewrite information to the Control FIFO
5. Packet rewrite and QoS classifications take place here at the input Fabric Interface
6. The packet is prepended with the internal routing tag. The internal routing tag used corresponds to the particular quality of service being requested, the appropriate port of exit, and the drop priority. In other words, the level of QoS determines which of the four queues the packet will be placed in
7. As soon as the entire frame is received into the Frame FIFO, the scheduler is signaled, requesting arbitration based on the QoS level. When arbitration is granted, the frame moves into the shared fabric and is stored with a pointer to the output port
8. The destination port is signaled by the fabric ASIC to take the frame out of a known memory location. The destination port knows that it is receiving the correct frame because of the internal routing tag corresponding to a particular internal port-to-port circuit
9. The frame is sent out to the network

Layer 2 Bridging

When a port or groups of ports are running in bridging mode, the Search Engine initiates a lookup in the CAM table based on the Layer 2 MAC address. Because the Catalyst 8500 is a distributed switching system, each port (or in this case, CEFA) maintains a list of addresses and ports of exit that are of local significance. This means that if, for example, Address A is a destination learned on interface FastEthernet 0/1, the remaining interfaces on the Catalyst 8500 do not have to have that address stored in their CAM tables unless they have a packet to send to Address A.

The central CAM on the SRP maintains the master table for both Layer 2 addresses and Layer 3 routes. When a new address is learned by a CEFA, that address (not the packet) is sent to the CPU so that the CPU has an updated list of all MAC addresses being bridged. The CPU populates the central CAM with the new address (in the previously unknown address is a source address on the frame) When a new address is learned or updated, the CEFA is updated by the central CAM table. The central CAM (which will be discussed later in this document) contains all addresses that the switch has learned.

If the destination MAC address is a broadcast address (ffff.ffff.ffff), the packet is tagged with a destination as being all ports in that bridge group and is sent out to the switching fabric. The fabric ASIC creates a pointer from that point in memory to all ports in that bridge group. This means that if there were eight ports in a bridge group, all eight ports would receive that broadcast.

Assuming that both the source and the destination MAC address have been learned, the following procedure occurs during Layer 2 frame switching:

1. A packet is received at the physical interface. The CEFA ASIC provides the MAC-layer functions, and the packet is stored in internal memory
2. As soon as the first 64 bytes of the frame are read, the microcode running on the microcontroller reads the MAC source and destination addresses. If the destination MAC address is not that of the interface, Layer 2 switching is required. This information can now be used by the Search Engine.
3. Because the packet has been received on a particular VLAN, the Search Engine begins a search for the MAC address and its corresponding port of exit
4. The destination MAC address is found. The microcontroller moves the frame from the internal memory to the Fabric Interface's Frame FIFO. At the same time, the Search Engine returns relevant information such as QoS classifications or ISL information to the Control FIFO
5. The frame is prepended with the internal routing tag
6. As soon as the entire frame is received into the Frame FIFO, the scheduler is signaled, requesting arbitration. When arbitration is granted, the frame moves into the shared fabric and is stored sequentially
7. The destination port is signaled by the Fabric ASIC to take the frame out of memory. The destination port knows that it is receiving the correct frame because of the internal routing tag
8. The frame is re-encapsulated via ISL, if necessary, and sent out to the network

Catalyst 8500 Quality-of-Service Mechanisms

As network managers begin to deploy critical network applications, QoS becomes increasingly more important. The Catalyst 8510 provides extensive core QoS mechanisms that are built into the switch architecture. These functions, performed by the fabric ASIC and frame schedule, ensure policy enforcement via packet classification and queuing on the ingress port as well as scheduling via WRR at the egress port.

The Phase 1 implementation of QoS on the Catalyst 8500 is based on IP precedence. IP precedence information is gathered from the type-of service (ToS) field in the IP header. For an

incoming IP packet, the first three bits of the ToS (also called the Service Type field) are used to determine the delay priority and the drop priority. The least significant bit (bit 2) defines the drop priority. If this bit is turned on, the Catalyst 8510 drops that packet before it drops packets with the bit turned off when the destination queue becomes full. The higher 2 bits are used for the delay priority.

This means that there are eight different classes that the Catalyst 8500 can recognize. The classes are summarized in Table 2.

Table 2 IP Precedence Values

IP TOS Field Value	Delay Priority	Drop Priority	Queue Selected
0 0 0	0 0	0 (Drop packet first)	QoS-3 (lowest priority)
0 0 1	0 0	1 (Drop after last)	QoS-3
0 1 0	0 1	0	QoS-2
0 1 1	0 1	1	QoS-2
1 0 0	1 0	0	QoS-1
1 0 1	1 1	1	QoS-1
1 1 0	1 1	0	QoS-0
1 1 1	1 1	1	QoS-0 (highest priority)

Currently, the Catalyst 8500 has the ability to read the precedence field and switch accordingly. This means that it cannot reclassify traffic (in Phase I). Future phases of the Catalyst 8510 will allow more granular enforcement and reclassification of the IP precedence field.

Queuing

In the Catalyst 8500, packets are queued based on the delay priority and the target next-hop interface. The highest two bits of the IP precedence indicate to the Frame Scheduler which queue the packets must be placed in. The Fabric Switching ASIC then provides a pointer from the output port to that queue, indicating where in memory to retrieve the packet from. As a result of the ingress data-path processing on microcontroller on the CEF ASIC, the packet can be queued to one of 128 queues (for each of 32 possible ports there are four queues) based on next-hop interface and delay priority. The Fabric Switching ASIC is responsible for providing the output port with the correct pointer to the correct queue.

Each queue has an absolute queue limit and discard threshold limit (essentially they can be seen as the in-of-profile threshold and out-of-profile queue threshold, respectively). All packets are queued if the current queue depth is below the discard threshold. If the queue depth is between discard threshold and absolute queue limit, only packets with drop priority “1” (or in-profile packet) are queued, and packets with drop priority “0” are discarded. When queue depth is beyond absolute queue limit, all packets are discarded. The discard threshold is user configurable. This feature will be available in a future release.

Scheduling and Weighted Round Robin

As stated earlier, frame scheduling becomes increasingly important when a outgoing link is congested. To handle this problem network managers have the option of selecting weights to each of the different queues. This allows bandwidth to be granted to higher-priority applications (via IP precedence), yet still fairly grant access to lower-priority queues. In cases where there is no network congestion, all queues are granted the same weight and are serviced appropriately. However, when congestion occurs, the Frame Scheduler allows each queue the bandwidth set forth by the network manager.

Commands for QoS-Based Switching Features

The QoS commands in the Catalyst 8500 are new to the Cisco IOS software and are mentioned in order to demonstrate how to configure these new features. The goal behind QoS support is to sort traffic into a small number of classes and mark the packets

accordingly within the switching fabric. The QoS value is used to provide differential treatment to traffic in different classes, such that different QoS is provided to each class.

I. Enable/Disable QoS-Based Switching

```
Switch-Router> en
Switch-Router# config t
Switch-Router(config-t)# qos switching
```

The “no” version disables QoS-based switching on the entire system. By default, QoS-based switching is always enabled. All packets are sent across the fabric using the QoS-3 (or low-priority) queue.

II. Configure Precedence to WRR Scheduling Weight Mapping

A. System-Level Mapping

```
Switch-Router> enable
Switch-Router# configure terminal
Switch-Router(config)# qos mapping precedence
<value> wrr-weight <weight>
```

- <value>

The precedence value (0 to 3) derived from the IP precedence field in order to determine delay priority. The higher two bits of the precedence field are used for this purpose

- <weight>

The WRR scheduling weight (1 to 15). This parameter specifies the weight assigned to traffic with the given precedence.

To set the above precedence back to default for the Catalyst 8510, use the “no” version of the above command.

The following defaults are used to map the IP precedence to the WRR weights and are shown in Table 3.

Table 3 IP Precedence and WRR Weights

IP Precedence	WRR Weight
0	1
1	2
2	4
3	8

B. Interface-Level Mapping

This command allows the network manager to configure the mapping at the interface level, overriding the system-level mapping. Essentially, this allows the network manager to assign different WRR scheduling weight for a particular precedence traffic between a pair of interface(s). (Note that this is not per-flow queuing, which can key off of the source or destination address.)

```
Switch-Router> enable
Switch-Router# configure terminal
Switch-Router(config)# qos mapping [source fastether <x/y/z>] [destination fastether <a/b/c>] precedence <qos-val> wrr-weight <weight>
```

The keywords and parameters used are the same as for the system-level mapping

Both the source and destination interfaces are optional.

When both are not specified, the system-level QoS mapping is configured. Otherwise, the network manager can specify a source and/or destination interface to configure the WRR weight. The WRR weights can be based on (in order of priority):

- Traffic streams with a certain precedence from a particular source interface to a particular destination interface
- Traffic streams with a certain precedence to a particular destination interface
- Traffic streams with a certain precedence from a particular source interface

Using the “no” version of the above command sets the mapping for the specified precedence to the current system-level mapping.

III. Show Commands

A. QoS-Based Switching

```
Switch-Router> en
Switch-Router# show qos switching
```

This command indicates whether QoS-based switching is enabled or not. The output is as follows:

```
QoS Based IP Switching is [Enabled | Disabled].
```

B. QoS Mapping

```
Switch-Router> enable
Switch-Router# show qos mapping [source fastether <x/y/z>
destination fastether
<a/b/c>]
```

Catalyst 8500 Multicast Support

IP multicast allows IP traffic to be sent from one source or multiple sources and delivered to multiple destinations. Instead of sending individual packets to each destination, which is highly taxing to the switch fabric, a single packet is sent to a multicast group, which is identified by a single IP destination group address. That IP destination group consists of a number of IP destinations that require that frame. From a router perspective, an input multicast feed from a given source must be sent out through (possibly) multiple output interfaces based on the information received by the multicast routing protocols such as PIM.

The Catalyst 8500 CSR supports IP multicast at wire speed for all ports, allowing for high-speed switching of packets from input source ports to multiple destination ports. The Catalyst 8500 also supports IP multicast routing protocols such as PIM dense and sparse modes as well as DVMRP interoperability.

Internet Group Management Protocol

Internet Group Management Protocol (IGMP) provides a method for end stations to request multicast traffic as well as for routers to determine who on a locally attached segment is requesting traffic. IGMP uses IP datagrams to allow IP multicast applications to join a multicast group. IGMP relies on Class D IP addresses for the creation of multicast groups and is defined in RFC 1112.

Membership in a multicast group is dynamic, meaning that it changes over time as hosts join and leave the group. Multicast routers use IGMP host-query messages (sent to group address of 224.0.0.1 with TTL of 1) to keep track of the hosts that belong to multicast groups. When router receives a packet addressed to a multicast group, it forwards the packet to those interfaces that have hosts belonging to that group. Routers periodically send host-query messages to refresh their multicast group membership knowledge.

The Catalyst 8500 supports both IGMP version 1, which most end stations currently support, and IGMP version 2, which, unlike version 1, provides support for clients informing the network that they are leaving a multicast group.

Protocol Independent Multicast

As networks increase in size, multicast routing becomes critically important in order to determine, in a large routed network, which segments require multicast traffic and which do not. PIM is a routing protocol for multicast that uses existing unicast routing protocols such as RIP or OSPF for path forwarding determination and network location. PIM can be operated in two modes. PIM dense mode and PIM sparse mode. The mode selected determines how the Catalyst 8500 populates its multicast routing table and how the router forwards multicast packets it receives from its directly connected LANs. Note that enabling PIM on an interface also enables IGMP operation on that interface.

In dense mode, a router assumes that all other routers want to forward multicast packets for a group. Therefore, interfaces with PIM dense mode enabled receive the multicast feed as soon as a single user requests one. That segment will continue to receive the multicast until it times out. If a Catalyst 8500 receives a multicast packet and has no directly attached members or PIM neighbors present, a prune message is sent back to the source. Subsequent multicast packets are not flooded to this pruned branch. PIM builds source-based multicast distribution trees. PIM dense mode is most useful when:

- The senders and receivers are in close proximity to one another
- There are fewer senders than receivers
- Multicast traffic volume is high

- The stream of multicast traffic is constant

In sparse mode, a router assumes that other routers do not want to forward multicast packets for a group, unless there is an explicit request for the traffic. When hosts join a multicast group, the directly connected routers send PIM join messages to the rendezvous point (RP). The RP keeps track of multicast groups. Hosts that send multicast packets are registered with the RP by that host's first-hop router. The RP then sends joins toward the source. At this point, packets are forwarded on a shared distribution tree. When the data stream begins to flow from sender to RP to receiver, the routers in the path optimize the path automatically to remove any unnecessary hops. Sparse mode assumes that no hosts want the multicast traffic unless they specifically ask for it.

Sparse mode PIM is optimized for environments where there are many multipoint data streams and each multicast stream goes to a relatively small number of LANs in the internetwork. PIM sparse mode is most useful when:

- There are few receivers in a group
- Senders and receivers are separated by WAN links
- The type of traffic is intermittent

Distance Vector Multicast Routing Protocol

DVMRP is the first-generation multicast routing protocol most known for its use in the Multicast Backbone (MBONE). DVMRP uses a flood-and-prune approach to multicast packet delivery. This means that DVMRP assumes that all other routers in a network want to forward multicast packets for a group. This creates huge scalability problems as routers must now maintain state for multicast paths that may not require or want to handle multicast traffic. For that reason, the Catalyst 8500 does not support DVMRP, but does support DVMRP interoperability with PIM. This allows the Catalyst 8500 to interoperate with non-Cisco multicast routers that use DVMRP.

Cisco IOS software in the Catalyst 8500 supports dynamic discovery of DVMRP routers and can interoperate with them over traditional media or over DVMRP-specific tunnels. When a DVMRP neighbor has been discovered, the router periodically transmits DVMRP report messages advertising the unicast sources reachable in the PIM domain.

Cisco Group Membership Protocol

Cisco Group Membership Protocol (CGMP) addresses the issue of efficiently forwarding IP multicast packets across Layer 2 switches. CGMP allows Layer 2 switches to leverage IGMP information recorded on the Catalyst 8500 to make intelligent Layer 2 forwarding decisions based on the destinations requesting the multicast traffic. The net result is that with CGMP, IP multicast traffic is delivered only to those Layer 2 switch ports that are interested in multicast traffic. All other Layer 2 switch ports that have not requested the traffic do not receive it. When a router receives an IGMP join message, it records the source MAC address of the IGMP message and turns around and issues a CGMP Join message downstream to Layer 2 switch. The switch uses the CGMP message to dynamically build an entry in the switching table that maps the multicast traffic to the client's switch port.

The Catalyst 8500 uses PIM, not CGMP, for multicast forwarding determination. However, the Catalyst 8510 does function as a CGMP server, meaning that on a per-interface basis, it informs the connected LAN switch of multicast groups that it needs to be aware of. The Catalyst 8500 responds to IGMP version 1 and 2 multicast join and leave (for IGMP v2) requests and forwards them on the multicast tree via PIM.

The Multicast Routing Table

The Cisco IOS software running on the Catalyst 8500 uses PIM and DVMRP interoperability to exchange IP multicast network information. Each routing protocol runs as a separate IOS process in the SRP. The multicast routing table is a centralized routing information database that is resident on the SRP. The packet forwarding engine consults the routing table to route the packets to appropriate destinations.

A multicast routing table is different than a unicast routing table. A multicast routing table maps an ordered pair consisting of a source IP address and a multicast group to an ordered pair consisting of an input interface and a set of output interfaces. Packets from the given source to the given multicast group that arrives over an input interface are appropriate output interfaces. Packets that arrive on wrong input interface are discarded.

The Catalyst 8500 maintains the central multicast routing table at the SRP. By using CEF and the associated distribution of the forwarding information base, the line cards can forward multicast

traffic intelligently based on the multicast topology of the network. This feature allows the input port to decide which output interfaces require the multicast traffic and inform the switching fabric about which output ports to direct that packet to. Any change in the multicast routing table is instantly downloaded to the line cards, allowing the Catalyst 8500 to maintain a constant, up-to-date map of the network.

Catalyst 8500 Series Features

The Catalyst 8500 series provides a rich feature set designed to service both Layer 2 switched and Layer 3 routed networks. The following features are currently available in the Catalyst 8510.

Layer 3 Features

Fast EtherChannel

Link redundancy with the Fast EtherChannel® technology allows for physical link redundancy using Fast Ethernet connections. Up to four Fast Ethernet connections can be used as one Layer 3 forwarding path providing up to 800-Mbps aggregate capacity. If link detection determines a failure of any one link, packet switching is performed on the remaining active links in the Fast EtherChannel. Based on the source/destination address pair, Fast EtherChannel optimizes the available bandwidth by load-balancing traffic across the available links in the Fast EtherChannel.

Fast EtherChannel works via a source-destination IP or IPX address load-balancing scheme. In the Catalyst 8500, up to four ports can be configured in a channel group. Each channel group has its own IP address or IPX network number. When a packet is queued to exit out of the port channel interface, an Exclusive-OR (X-OR) operation takes place on the last two bits of the source and destination address. The result of that operation determines which link in the channel the packet will take. A given source/destination pair will consistently take the same link unless a failure occurs. Upon detection of a failure, all learned addresses switch over to one of the remaining links in the channel. Unlike the Catalyst 5000 Fast EtherChannel implementation, the Catalyst 8500 places no dependencies on which ports are configured in the channel. The ports can exist on the same or on different line modules in the chassis. Also, the Catalyst 5000, to accomplish link determination, uses the MAC address for the X-OR operation, while the Catalyst 8500 utilizes the Layer 3 address.

Hot Standby Router Protocol

Hot Standby Router Protocol (HSRP) is designed to provide high network availability by routing IP traffic from hosts on Ethernet networks without relying on the availability of any single router. This feature is particularly useful for hosts that do not support a

router discovery protocol (such as IRDP) and do not have the functionality to switch to a new router when their selected router reloads or loses power. Without this functionality, a router that loses its default gateway because of a router failure is unable to communicate with the network.

HSRP solves this problem by providing a virtual MAC address and IP address that are shared between the routers in an HSRP group. One of these devices is selected by the protocol to be the active router, either because its interface has a lower MAC address or because the network manager has pre-empted other interfaces. The active router receives and routes packets destined for the HSRP group's MAC address. When HSRP detects that the designated active router has failed, the selected backup router assumes control of the HSRP group's MAC and IP addresses. (A new standby router can also be selected at that time.) In order to detect a failure, devices that are running the Hot Standby Router Protocol send and receive multicast UDP-based hello packets. (Note that ICMP redirects are disabled for interfaces running HSRP.)

The chosen MAC address and IP address are unique and will not conflict with any others in the same network segment. The MAC address is chosen from a well-known pool of Cisco MAC addresses. The user configures the last byte of the MAC address by configuring the HSRP group number. The unique virtual IP address is also configured by the user. The IP address need to be specified only on a single router within a same group. Once HSRP starts running, it selects an active router and instructs its device layer to listen on an additional MAC address, which is a dummy MAC address.

Cisco IOS Routing Protocols

The Catalyst 8500 switch provides a comprehensive suite of routing protocols based on the Cisco IOS software. The key benefit of this implementation is that the Cisco IOS software has been in development and deployment for many years, ensuring a reliable implementation of all the routing protocols supported. This capability differs considerably from many competitive products that are attempting to implement complex routing protocols for the first time.

The Catalyst 8500 supports RIP and RIP version 2, OSPF, IGRP, and Enhanced IGRP routing for IP networks. Future support for routing protocols will include IS-IS and BGP-4. For IPX networks, the Catalyst 8500 supports RIP and Enhanced IGRP. Many of the Cisco IOS routing protocol features such as route redistribution and load balancing over equal cost paths (for RIP, OSPF, Enhanced IGRP and static routes) are currently supported.

Configuration of these routing protocols is identical to the configuration methods currently used on all Cisco router platforms.

Integrated Routing and Bridging

Integrated Routing and Bridging (IRB) enables a network manager to route a given protocol between routed interfaces and various bridge groups. This allows multiple ports in the Catalyst 8500 to reside in one bridge group with one IP address and be routed to other Catalyst 8500 interfaces with different IP addresses. Packet switching when running IRB takes place exactly as defined earlier in this paper, with the exception that a packet entering a port is allowed to have either the destination MAC address of the interface (meaning routing must take place) or a different MAC address, indicating that the address exists on the same subnet but on a different port in the bridge group.

IRB in the Catalyst 8500 works only for IP and IPX. Therefore, when configuring routing, it is NOT possible to create the following configuration:

```
Router# bridge irb
Router# bridge 1 route appletalk
```

Routing AppleTalk (or any other protocol other than IP and IPX) is not currently supported in the Catalyst 8500. Future multiprotocol support may add this functionality.

Layer 2 Features

Inter-Switch Link

In order to support VLANs between switches, the Catalyst 8500 identifies frames from end stations as belonging to a particular VLAN. It does this using a trunking protocol called Inter-Switch Link, or ISL, which runs over Fast Ethernet. The ISL technology uses a scheme known as packet tagging. This scheme allows the Catalyst 5000 series (along with the Catalyst 3000 series) to multiplex VLANs across a single physical link, maintaining strict adherence to the individual VLAN domains. The ISL frame is a standard Ethernet or IEEE 802.3 frame, tagged with a VLAN ID. This frame is sent as a multicast and is only meaningful to ISL devices. Because it is a standard frame, repeater hubs and transparent bridges forward it as they would any other frame. Any 100-Mbps Ethernet link can support this protocol. The link can run at either half-duplex or full-duplex.

Bridge Groups

A bridge group is a broadcast domain set up within a switch. Typically, a bridge group corresponds to a particular subnet, although not necessarily. Bridge groups in the Catalyst 8500 are designed to support non-IP/IPX traffic, which must be currently bridged.

The Catalyst 8500 supports up to 64 bridge groups per switch. It is important to differentiate a bridge group from a VLAN. A VLAN is assumed to terminate at the Catalyst 8500, because it is expected that routing will take place. An ISL link can trunk multiple VLANs, each one assuming to "end" at the router port. Should a VLAN need to go through the switch, a bridge group is configured. A VLAN configured on the Catalyst 8500 for an ISL link is not a bridge group since that traffic is not bridged through the switch. VLAN trunking is currently supported in the Catalyst 8510 via the ISL technology. The emerging IEEE 802.1Q VLAN trunking standard will be supported in the second software release.

Configuration of bridge groups within the Catalyst 8500 is similar to VLAN configuration in Cisco routers. A sub-interface is defined at the interface. A bridge group is defined and a VLAN is mapped to the sub-interface via the encapsulation isl <vlan number> command.

Cisco Discovery Protocol

CDP is a Layer 2 protocol available on all Cisco products. CDP allows Cisco products to exchange information with each other regarding their MAC addresses, IP addresses, and outgoing interfaces. This feature is a critical method for troubleshooting potential network problems.

Conclusion

The Catalyst 8500 family of Campus Switch Routers is designed to scale campus distribution and core networks by providing key performance for IP and IPX at 6 million packets per second (for the Catalyst 8510) and 24 million packets per second (for the Catalyst 8540). In addition, the Catalyst 8500 provides important network services not found in start-up vendors' switches, including HSRP, wire-speed multicast forwarding and routing, extensive QoS offerings, and most importantly, the key routing and routed protocol support of the tried-and-true Cisco IOS software.



Corporate Headquarters

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

European Headquarters

Cisco Systems Europe s.a.r.l.
Parc Evolic, Batiment L1/L2
16 Avenue du Quebec
Villebon, BP 706
91961 Courtaboeuf Cedex
France
<http://www-europe.cisco.com>
Tel: 33 1 6918 61 00
Fax: 33 1 6928 83 26

Americas

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

Asia Headquarters

Nihon Cisco Systems K.K.
Fuji Building, 9th Floor
3-2-3 Marunouchi
Chiyoda-ku, Tokyo 100
Japan
<http://www.cisco.com>
Tel: 81 3 5219 6250
Fax: 81 3 5219 6001

Cisco Systems has more than 200 offices in the following countries. Addresses, phone numbers, and fax numbers are listed on the Cisco Connection Online Web site at <http://www.cisco.com>.

Argentina • Australia • Austria • Belgium • Brazil • Canada • Chile • China (PRC) • Colombia • Costa Rica • Czech Republic • Denmark • England • France • Germany • Greece • Hungary • India • Indonesia • Ireland • Israel • Italy • Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland • Portugal • Russia • Saudi Arabia • Scotland • Singapore • South Africa • Spain • Sweden • Switzerland • Taiwan, ROC • Thailand • Turkey • United Arab Emirates • United States • Venezuela