

ARCHITECTURES WITH BUS-BASED SWITCH FABRICS: CASE STUDY— CISCO CATALYST 6000 SERIES SWITCHES

7.1 INTRODUCTION

The Cisco Catalyst 6000 family of switch/routers consists of the Catalyst 6000 and 6500 Series that are designed to deliver high-speed Layer 2 and 3 forwarding solutions for service provider and enterprise networks. This chapter describes the different Catalyst 6000 series switch/router architectures [CISCCAT6000, MENJUS2003]. The Catalyst 6000 series supports several architectural options in terms of Layer 3 routing and forwarding capabilities.

Layer 3 routing and forwarding capabilities in the Catalyst 6000 and 6500 switches are handled by two important components: the Multilayer Switching Feature Card (MSFC) and the Policy Feature Card (PFC). These two key components, along with other functions required for system configuration and operation, are implemented on a specialized module in the Catalyst 6000/6500 called the Supervisor Engine.

The PFC provides the hardware-based forwarding engine functions required to perform in the Catalyst 6000/6500, Layer 2 and 3 forwarding, quality of service (QoS) classification, policing, shaping, priority queuing, and QoS and security access control list (ACL) processing. For Layer 3 forwarding, the PFC requires the services of a routing engine (or route processor) in the chassis to generate the necessary routing information needed for the flow/route cache or forwarding tables used by the Layer 3 forwarding engine.

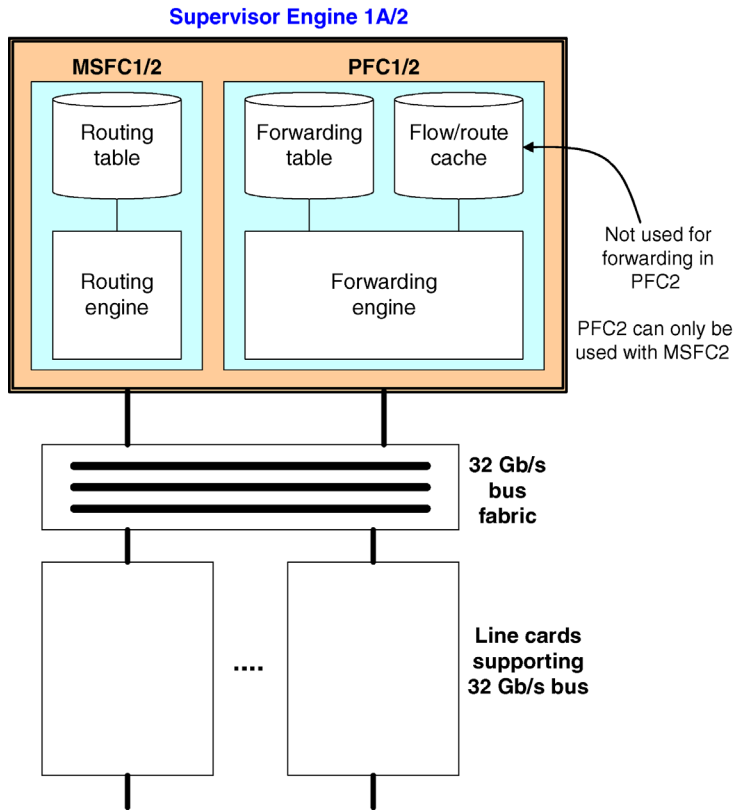


FIGURE 7.1 Bus-based architecture with forwarding engine in centralized processor.

The MSFC provides the control plane (routing engine) functions required by the PFC to perform Layer 3 forwarding. Based on the architecture categories defined in Chapter 3, the architectures discussed here fall under “Architectures with Bus-Based Switch Fabrics and Centralized Forwarding Engines” (see Figure 7.1)

7.2 MAIN ARCHITECTURAL FEATURES OF THE CATALYST 6000 SERIES

This section highlights the main architectural features of the Catalyst 6000 series of switch/routers. Compared to the Catalyst 6500 Series, the Catalyst 6000 Series offers lesser capabilities in terms of packet forwarding performance and scalability. The Catalyst 6000 Series offers a more cost-effective solution for enterprises and service providers not requiring the higher performing Catalyst 6500 Series.

The Catalyst 6000 Series has a backplane bandwidth of 32 Gb/s (using a shared switching bus architecture) and Layer 2/3 forwarding capacity of up to 15 million

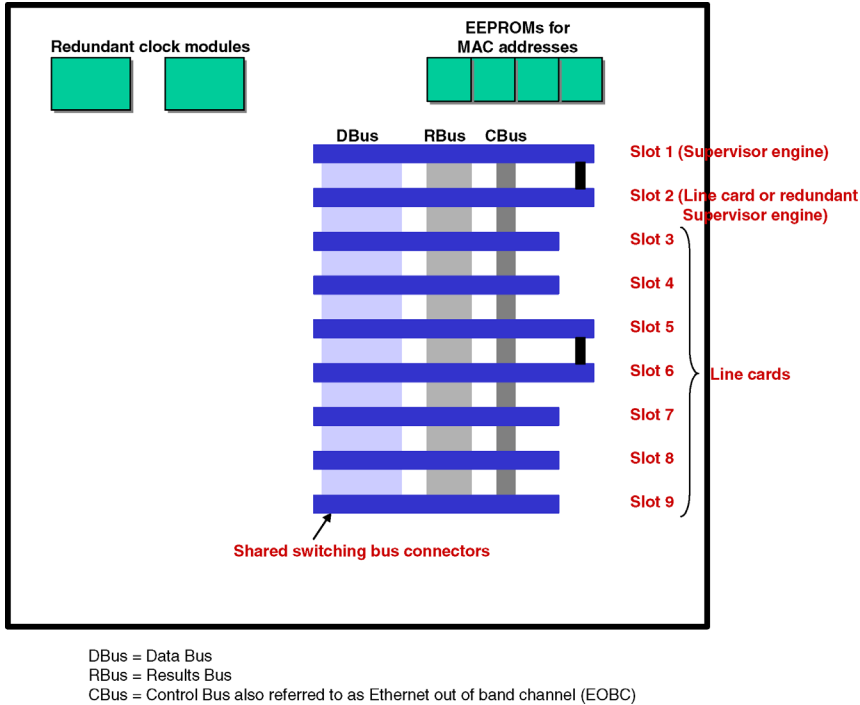


FIGURE 7.2 Catalyst 6009 backplane.

packets-per-second (Mpps). The Catalyst 6500 Series architecture, on the other hand, supports a backplane bandwidth of up to 256 Gb/s and Layer 2/3 forwarding in excess of 200 Mpps.

The Catalyst 6000 Series supports different chassis options, a six-slot chassis in the Catalyst 6006 and nine-slot chassis in the Catalyst 6009. These switch/routers support a wide range of network interface types and port densities: 384 10/100BASE-T/TX Ethernet ports, 192 100BASE-FX Ethernet ports, and up to 130GbE (Gigabit Ethernet) ports (in the nine-slot chassis). The main identifying components on the Catalyst 6006 and 6009 backplane (Figure 7.2) are the following:

- Slot 1 supports the Supervisor Engine.
- Slot 2 supports either a line card or a redundant Supervisor Engine.
- Slots 3–9 support line cards.
- Supports bus connectors for the 32 Gb/s shared switching bus: Data Bus (DBus); Results Bus (RBus); Control Bus (CBus) (or Ethernet out-of-band channel (EOBC))
- Clock module with redundancy (primary and secondary).
- Ethernet MAC address EEPROMs.

The line cards used in the Catalyst 6000 Series are referred to as the “nonfabric-enabled” or “Classic” line cards. These line cards can be used in both the Catalyst 6000 and Catalyst 6500 Series and connect to the 32 Gb/s shared switching bus. They can also be used in all Catalyst 6000 and Catalyst 6500 Series chassis types as long as they support the 32 Gb/s switching bus.

Other line card types referred to as “fabric-enabled” and “fabric-only” are available for the Catalyst 6500. These line card types support connectivity to a crossbar switch fabric.

7.3 HIGH-LEVEL ARCHITECTURE OF THE CATALYST 6000

Both Cisco Catalyst 6000 and 6500 Series support a common 32 Gb/s shared switching bus architecture as shown in Figure 7.3. The Catalyst 6000, in particular, supports only the 32 Gb/s shared switching bus as the backplane connectors in Figure 7.2 illustrate. The Catalyst 6500, on the other hand, supports both the 32 Gb/s shared switching bus (same architecture as in the Catalyst 6000 Series) and an option for a stand-alone 256 Gb/s Switch Fabric Module (SFM) or an integrated 720 Gb/s crossbar switch fabric on Supervisor Engine 720. The exception is the Catalyst 6500 with Supervisor Engine 32 that supports only a 32 Gb/s shared switching bus.

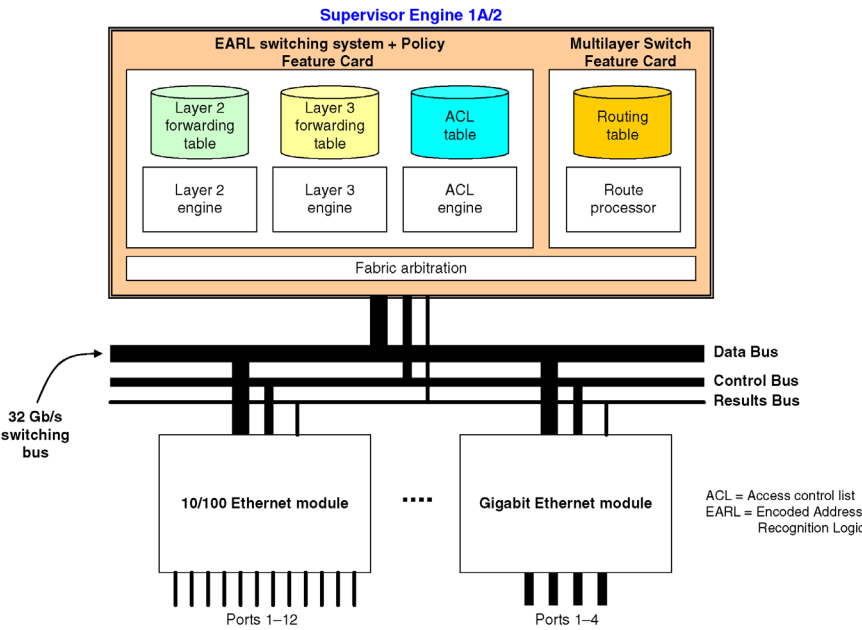


FIGURE 7.3 High-level view of the Cisco Catalyst 6000 and Catalyst 6500 switching bus architecture.

The 32 Gb/s backplane is designed as an advanced pipelining shared switching bus and consists of three separate sub-buses: the Data Bus, the Results Bus, and the Control Bus (which is also called the Ethernet out-of-band channel). The Classic (or nonfabric-enabled) line card types connect to the 32 Gb/s shared switching bus via the bus connectors on the backplane shown in Figure 7.2.

The DBus has a bandwidth of 32 Gb/s and is the main bus that carries data from one system module or line card to another. The RBus carries forwarding information (obtained after forwarding table lookup) from the forwarding engine located on the Supervisor Engine back to all the line cards. The CBus or EOBC carries control and management information between the line card port ASICs and the network management entity in the Supervisor Engine.

The 32 Gb/s pipelining switching bus is designed as a shared medium bus that allows all the frames transmitted on it to be visible to all the modules and ports attached to the bus. The pipelining mechanism working with the shared switching bus allows efficient reception of a transmitted frame after a forwarding decision is made by the forwarding engine. It also allows the flushing of a transmitted frame at the nondestination ports once the destination ports have received the frame (i.e., they are instructed to ignore the frame).

7.3.1 32 Gb/s Switching Bus Operating Modes

The 32 Gb/s switching bus has two operational modes: pipelining and burst modes. These two modes are described in the following sections.

7.3.1.1 Pipelining Mode In the typical or conventional shared medium bus architecture, only a single frame can be transmitted (i.e., propagated) on the bus at any given time. In such an architecture, if the frame is transferred across the shared switching bus from a port (to the forwarding engine) before the forwarding table lookup is completed, the shared switching bus stays idle until the address lookup is completed. Pipelining allows the ports to transmit multiple frames back-to-back on the switching bus while waiting for the results of the first frame sent to be received from the forwarding engine.

With pipelining, ports are allowed to transmit multiple frames on the shared switching bus before the result of the first frame address lookup is obtained. Immediately after transmitting the first frame, the second frame (which can be from any port) is transmitted across the shared switching bus and pipelined for forwarding table lookup at the forwarding engine. The system allows the address lookup process at the forwarding engine to occur in parallel to the transfer of the multiple frames across the shared switching bus.

In the pipelining mechanism, the 32 Gb/s switching bus allows 31 frames to be transferred back-to-back across the shared switching bus (and pipelined at the forwarding engine for address lookup operation) before the result of the first frame is received. The 32nd frame to be sent (after the sequence of transmitted 31 frames) must wait until the pipelining process once again allows it to be

transmitted on the shared switching bus as the first frame of a new sequence of 31 frames.

7.3.1.2 Burst Mode The burst mode feature together with a transmit threshold mechanism allows fair allocation of the 32 Gb/s shared switching bus bandwidth to the ports. To understand how this works, let us consider the following case. If a port transmits just a single frame each time it is permitted to transmit on the shared switching bus, depending on the lengths of the frames transmitted by all ports, there could potentially be an unfair bandwidth allocation to some ports during heavy traffic load conditions.

Let us assume, for example, that two ports (Ports A and B) have data to send and with Port A having 150 byte frames while Port B 1500 byte frames. With a simple allocation policy without thresholds, Port B with the 1500 byte frames can send 10 times more data than Port A with 150 byte frames. This unfairness arises because the ports alternate in the arbitration process for the shared switching bus and when granted access each port transmits just one frame at a time. The burst mode feature enables a port to transmit multiple frames on the shared switching bus but subject to a threshold.

With the burst mode feature, a port can transmit multiple frames on the shared switching bus but the amount of bandwidth it consumes is controlled independent of the frame size it transmits. A count of the number of bytes a port has transmitted is maintained by its port ASIC (in a counter) that is compared to a threshold. The threshold values are computed by the system to ensure fair distribution of the shared switching bus bandwidth.

As long as the port byte count is below the threshold value, the port is permitted to transmit more frames if it has any. When the byte count goes above the threshold, the port is not permitted to send additional frames after transmitting the current frame and stops further frame transmission. This is because the fabric arbitration logic (see Figure 7.3) upon sensing this threshold exceeding condition removes bus access for the port in question.

7.3.2 Control Plane and Data Plane Functions in the Catalyst 6000

Similar to routers, routing and switching in multilayer switches rely on two key functions commonly referred to as the control plane and data plane. The control plane (realized through a route processor, also called the control engine) is responsible for running the routing protocols, generating and maintaining the routing table, and maintaining all of the management functions of the switch including device security and access control.

The data plane (realized through the forwarding engine(s)) is responsible for forwarding a packet on its way to the destination using the routing information generated by the control plane.

- **The MSFC as the Routing Engine:** The Catalyst 6000 uses a centralized control plane functionality that resides in a daughter card module called the

MSFC, which is one of two key components on the Supervisor Engine. In Supervisor Engines 1A and 2, the MSFC runs the routing protocols and maintains the routing table. It communicates with the hardware forwarding engines in the system across an out-of-band bus called the CBus (or EOBC).

- **The EARL (Encoded Address Recognition Logic) as a Layer 2 Forwarding Engine:** The EARL is a centralized Layer 2 processing and forwarding engine (in the Catalyst 6000 and 6500 Supervisor Engines) for learning MAC address locations of connected stations and forwarding packets based upon the learned MAC addresses. The EARL maintains the VLAN, MAC address, and port relationships in a Layer 2 forwarding table as illustrated in Figure 7.3. These relationships are used to make Layer 2 forwarding decisions in hardware using Layer 2 forwarding ASICs. In some Supervisor Engine architectures as discussed below, the EARL functions (i.e., Layer 2 forwarding functions) are fully integrated into the PFC.
- **The PFC as the Forwarding Engine:** The PFC implements the packet forwarding engine in the Catalyst 6000. It supports the forwarding engine ASICs that enable packet forwarding at data rates of up to 15 Mpps in the Catalyst 6000. The PFC also provides Layer 3/4 level packet field inspection and processing, allowing some security and QoS features to be supported based upon the Layer 3 and Layer 4 parameters of user packet traffic. As discussed below, the PFC supports Layer 3 forwarding only with the addition of an MSFC (which supports the Layer 3 control plane functions) in the system. The PFC can be installed and used just by itself for Layer 2 forwarding and simple Layer 3/4 packet inspection/processing, without an MSFC installed in the system.

The MSFC is essentially an IP router on a daughter card installed in the system, providing full Layer 3 routing functionality and enabling the Catalyst 6000 to perform Layer 3 forwarding. In a Layer 3 forwarding configuration, the MSFC provides the control plane component of Layer 3 forwarding engine (i.e., populating and maintaining the routing table).

The PFC provides the data plane component of Layer 3 forwarding engine (i.e., forwarding table lookups, rewriting frame and packet headers, and forwarding packets undergoing routing to the appropriate egress port). This means an MSFC must be installed with a PFC for full Layer 3 control and data plane operations to take place.

7.4 CATALYST 6000 CONTROL PLANE IMPLEMENTATION AND FORWARDING ENGINES: SUPERVISOR ENGINES

The Catalyst 6000 Series supports two versions of the Supervisor Engine – Supervisor Engine 1A and Supervisor Engine 2. Supervisor Engine 1A is the first version of the integrated routing and forwarding engine designed for the Catalyst

6000 family. A Supervisor Engine must be installed in a Catalyst 6000 for it to function since all the required “intelligence” for system operation resides in this module. As shown in Figure 7.2, the Supervisor Engine sits in Slot 1 in the chassis. Slot 2 can accommodate a secondary redundant Supervisor Engine when required, or otherwise can be used for a line card.

For both Catalyst 6000 and 6500, one Supervisor Engine is sufficient for system operation, and in a redundant configuration, only one Supervisor Engine of the two engines needs to be active at one time. However, in the redundant configuration, both Supervisor Engines maintain the same state information, including Layer 3 routing information and tables, Layer 2 information including Spanning-Tree topology and Layer 2 forwarding tables, and system management information. In this configuration, if the primary Supervisor Engine fails, the redundant Supervisor Engine takes over without noticeable interruption in system operation.

7.4.1 Supervisor Engine 1A Architecture

Supervisor Engine 1A supports a flow/route cache-based forwarding scheme and provides forwarding performance of 15 Mpps. The Supervisor Engine 1A is targeted for deployment in network access layer scenarios such as in a wiring closet of an enterprise network, a server farm, and the main distribution frame (MDF) of an office building in an enterprise network.

As illustrated in Figures 7.3 and 7.4, Supervisor Engine 1A has the following main components: the EARL switching system, MSFC, and PFC. Each component supports a number of critical functions necessary for system operation. Supervisor Engine 1A is designed to support three different configurations as described in the following sections.

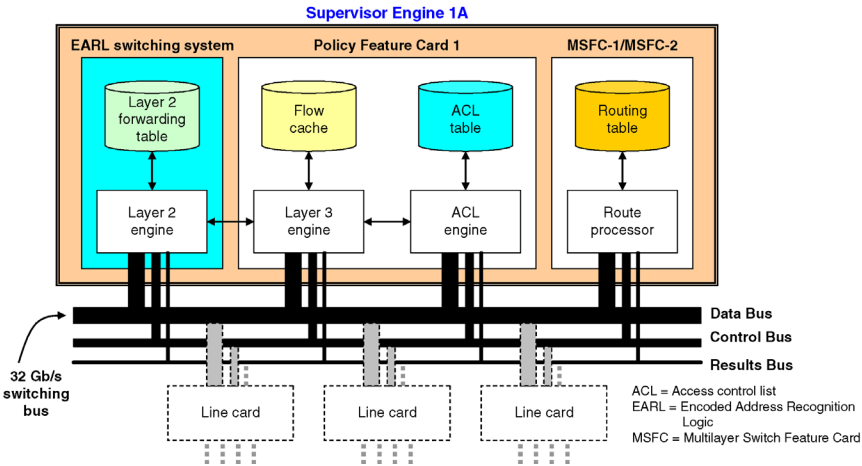


FIGURE 7.4 Catalyst 6000 with Supervisor Engine 1A with Policy Feature Card 1 (PFC1) and Multilayer Switch Feature Card (MSFC).

7.4.1.1 Supervisor Engine 1A with Only an EARL Switching System In this configuration, Supervisor Engine 1A has only an EARL switching system and with no MSFC and PFC present in the system. This is the most basic configuration allowing only basic Layer 2 forwarding based on MAC addresses.

The system does not support any Layer 3-based QoS or security ACLs, but only port-based and destination MAC address-based class of service (CoS). In this configuration, the Catalyst 6000 is reduced to a simple Layer 2 switch with no support for Layer 3 forwarding, QoS, and security classification capabilities. In this basic mode, the Supervisor Engine 1A can provide only Layer 2 forwarding of up to 15 Mpps.

7.4.1.2 Supervisor Engine 1A with an EARL Switching System Plus PFC1 In this configuration, Supervisor Engine 1A has an EARL switching system in addition to a PFC but no MSFC is present in the system. This is like removing the MSFC block from Figure 7.4 leaving only the EARL and PFC. This configuration provides Layer 2 forwarding with Layer 3 QoS and security ACLs services only. Layer 3 forwarding and routing is not supported. The system supports Layer 2 forwarding, Layer 3 QoS classification, queuing, and security filtering at data rates of 15 Mpps. These Layer 2 and 3 services are supported even though Layer 3 forwarding and route processing are not performed (unless an MSFC is added to provide route processor functions).

In this configuration, the Supervisor Engine 1A has the basic Layer 2 forwarding engine that inspects the local Layer 2 forwarding table to determine the egress port, and possibly VLAN, for Layer 2 forwarded packets. The PFC implements a Layer 3 forwarding engine, route/flow cache, and an ACL processing engine with a local ACL table. The PFC does not perform Layer 3 forwarding because no MSFC is present to provide the route processing functions required to generate the required Layer 3 routes and next hop information. The PFC, however, can perform Layer 3/4 QoS classification and ACL filtering using its local ACL engine and ACL table.

The local ACL table is maintained in ternary content addressable memory (TCAM). The TCAM stores and maintains the ACL information in a data format (i.e., data structures) that can be easily inspected by the ACL engine. A number of tasks are processed in parallel when a packet arrives at Supervisor Engine that requires ACL filtering. While the Layer 2 forwarding engine determines the egress port and VLAN for the packet by examining the forwarding information in the Layer 2 forwarding table, the ACL engine inspects its ACL table to determine if the packet is to be permitted or denied into the system. The lookups in Layer 2 forwarding table and the ACL table are performed in parallel, thus preventing QoS classification and ACL processing of traffic to not adversely affect the 15 Mpps forwarding rate of the switch.

7.4.1.3 Supervisor Engine 1A with an EARL Switching System Plus PFC1/MSFC1 (or 2) This configuration allows for full Layer 3 routing and forwarding in the Catalyst 6000 using the Supervisor Engine 1A. Here, the Supervisor Engine 1A module has a PFC1 and MSFC1 or MSFC2 in the system (Figure 7.4). This

configuration provides Layer 2 forwarding in addition to full Layer 3 routing and forwarding with the corresponding Layer 3 QoS and security services. This configuration enables the Catalyst 6000 to forward IP traffic at 15 Mpps.

The MSFC1 and MSFC2 share a similar architecture with the main difference being only in packet forwarding performance. The MSFC1 is designed with an R5000 200 MHz processor, up to 128 MB memory, and packet forwarding rate of up to 170 kpps in software. The MSFC2 supports an R7000 300 MHz processor, up to 512 MB memory, and packet forwarding rate of up to 650 kpps in software. They both can support Layer 3 packet forwarding in hardware at 15 Mpps in the Supervisor Engine 1A.

In this configuration, the Layer 3 forwarding engine on the PFC1 can perform Layer 3 forwarding, because route processing can now be done using the MSFC. The presence of the MSFC allows the Layer 3 forwarding engine to forward packets requiring routing such as is required in inter-VLAN communications. This Layer 3 forwarding can be done in addition to the other PFC features, such as QoS classification and ACL filtering. The Supervisor Engine 1A with PFC1 and MSFC1/MSFC2 employs a route/flow cache to forward Layer 3 traffic.

The flow cache (which is maintained on the PFC1 (Figure 7.4)) is used to forward Layer 3 packet flows through the Catalyst 6000. The first packet in a flow is always sent to the MSFC, which examines the local forwarding table it maintains to determine the next hop information for this first packet. The MSFC makes a Layer 3 forwarding decision and forwards the packet and the forwarding instructions back to the Layer 3 forwarding engine in the PFC1.

The Layer 3 forwarding engine then extracts the packet forwarding information and writes this information into its flow cache. When subsequent packets (belonging to the same flow as the first packet) are received and match the flow cache entries in the PFC1, they are Layer 3 forwarded directly by the PFC1 Layer 3 forwarding engine, rather than sent to the MSFC for processing and forwarding.

The main limitation of the flow cache-based method for Layer 3 forwarding is that the initial Layer 3 forwarding table lookup is performed by the MSFC software process. The first packet in a Layer 3 flow must be sent to the MSFC for Layer 3 forwarding table lookup and forwarding. This means that in a network environment that has many short-term Layer 3 flows being set up at the same time, the MSFC software process can easily be overwhelmed by the many flows it has to handle. This problem becomes more acute particularly in enterprise and service provider core network environments, where many short-term connections can be established at the same time.

7.4.1.4 Details of Packet Processing in the Supervisor Engine 1A The EARL module has its own local processor (for Layer 2 forwarding), which is referred to as the switch processor. The switch processor is responsible for running the Layer 2 protocols of the switch (e.g., Spanning Tree Protocol (STP), IEEE 802.1AB Link Layer Discovery Protocol (LLDP), and VLAN Trunking Protocol (VTP)), as well as implementing some QoS and security related services necessary for the PFC

Layer 3/4 data plane operations. The MSFC also has its own local processor, which simply can be referred to as the route processor and is responsible for implementing the Layer 3 control plane functions.

On Supervisor Engine 1A, the first packet in a flow that does not have an entry registered in the flow cache in the PFC1 is sent to the MSFC for software-based forwarding. The MSFC extracts the packet's destination IP address and performs a lookup in its local forwarding table to determine how to forward the packet. After the MSFC software process has forwarded the first packet of the new flow, the PFC1 receives and uses this forwarding information to program its flow cache so that it can forward subsequent packets in the same flow directly without MSFC intervention.

The forwarding decisions in the Supervisor Engine 1A are handled by three components (Figure 7.4): EARL switching system for Layer 2 MAC address-based forwarding, MSFC for Layer 3 forwarding of the first packet in a flow, and the PFC for Layer 3 forwarding of subsequent packets in a flow and ACLs processing (for implementing QoS and security services).

The EARL switching system (Layer 2 ASIC) learns MAC addresses within a broadcast domain (or VLAN) to create a Layer 2 forwarding table, which in turn is used to forward packets at Layer 2. The EARL module also identifies which packets (within a flow) need to be forwarded at Layer 3—Packets sent to destinations outside the broadcast domain have to be Layer 3 forwarded.

After forwarding the first packet, the MSFC generates an entry to be installed in the flow cache (in the PFC), which the Layer 3 forwarding engine (in the PFC) uses to forward subsequent packets in the flow in hardware. To facilitate the forwarding of packets requiring routing, the MSFC (considered the default gateway for routing traffic) registers its assigned MAC address with the Layer 2 forwarding engine so that upon examination of a packet, it can decide if the packet is to be sent to the MSFC or not.

The Layer 2 forwarding engine forwards packets requiring routing to the Layer 3 forwarding engine first, which in turn may forward them to the MSFC for further processing. Traffic to the MSFC (i.e., the default gateway for packets going to another VLAN) is sent to its known and registered MSFC MAC address at the Layer 2 forwarding engine.

After the Layer 2 forwarding engine determines that Layer 3 forwarding needs to take place (i.e., traffic to outside the broadcast domain), the services of the Layer 3 forwarding engine are engaged. The Supervisor Engine 1A uses flow-caching where a flow can be defined as a traffic stream from a specified source IP address to a specified destination IP address. The flow cache can also store in addition Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) port numbers as part of the flow cache entry.

When the first packet in a flow arrives at the PFC, a lookup is performed in the flow cache by the Layer 3 forwarding engine to determine if an entry for the packet exists. If the cache contains no related entry, the packet is sent to the MSFC, which uses the destination IP address to perform a lookup in the forwarding table it maintains (to determine the next hop IP address, the egress port, and destination

VLAN). The packet is then forwarded to the next hop and the forwarding information is used to create an entry in the PFC flow cache. Subsequent packets in the same flow can then be forwarded by the PFC hardware (by the Layer 3 forwarding engine using the newly created flow cache entry).

A maximum of 128,000 entries can be supported by the flow cache and there are three options for creating flow cache entries:

- **Destination-Only Flow Option:** In this configuration, a flow entry is created in the flow cache based on the destination IP address. This option utilizes less PFC flow cache space as multiple source IP addresses communicating with one destination IP address (e.g., a server) result in only a single-flow cache entry.
- **Source-Destination Flow Option:** In this configuration, a flow entry is created based on both the source and destination IP addresses. This option consumes more entries in the flow cache; for example, if four source IP addresses are communicating with one destination IP address, then four entries are created in the flow cache.
- **Full-Flow Option:** This is a more resource-consuming option of using the flow cache because a flow entry is created not only from the source and destination IP addresses but also from the UDP or TCP port numbers.

To facilitate searches, the flow cache is split into eight pages of memory with each storage page capable of handling 16,000 entries. The PFC then uses a hashing algorithm to carry out lookups in the flow cache. The hashing algorithm has to be efficiently implemented since it is critical to how entries can be stored and lookups can be performed at high speeds.

Hashing algorithms work in a statistical manner that can lead to hash collisions. A hash collision occurs when the lookups for two packets with different parameters hash to the same location in memory. To account for such collisions, the Layer 3 forwarding engine (in the PFC) moves to the next page in memory to check if that location is used. This next page search process continues until either the lookup information is stored or until the eighth page is reached. If the lookup information still cannot be stored after the eighth page, then the packet is flooded out the switch (at Layer 2) or sent to the MSFC (for further (and possibly, Layer 3) processing).

Supervisor Engines 1A and 2, and the Distributed Forwarding Cards (DFCs) in the Catalyst 6500 all use TCAMs for storing and processing QoS and security ACLs. A TCAM has the capacity of 64,000 entries and is split into four main blocks. Two blocks are used for QoS ACLs input checking and output checking, and two blocks for security ACLs input and output checking.

The PFC1 and PFC2 have similar TCAM implementation and behavior. The implementations allow ACL lookups to take place at packet rates of up to 15 Mpps on Supervisor Engine 1A and 30 Mpps on the DFC and Supervisor Engine 2. ACL lookups are performed in parallel to the Layer 2 and 3 lookups, which results in no performance degradation when processing QoS or security ACLs.

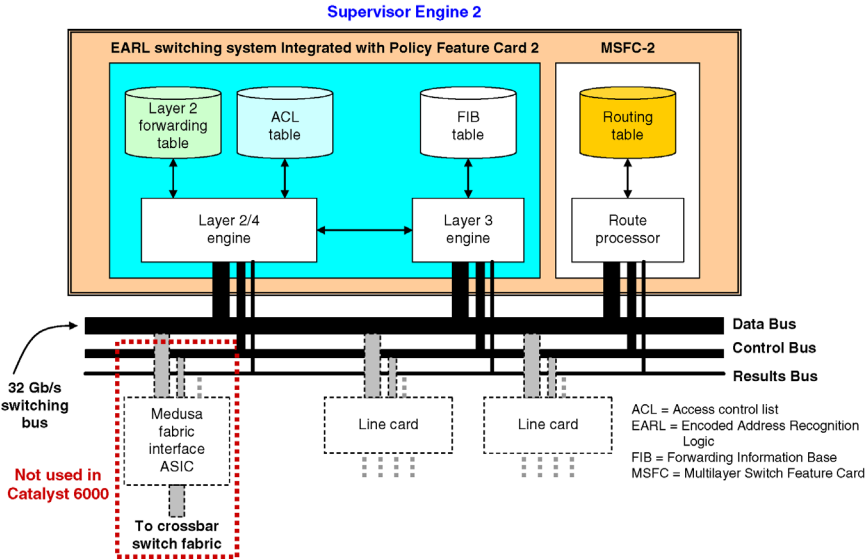


FIGURE 7.5 Catalyst 6000 with Supervisor Engine 2 with Policy Feature Card 2 (PFC2) and Multilayer Switch Feature Card 2 (MSFC2).

7.4.2 Supervisor Engine 2 Architecture

Supervisor Engine 2 provides higher packet forwarding performance and system resiliency compared to Supervisor Engine 1A. Supervisor Engine 2 (Figure 7.5) provides forwarding capacity of up to 30 Mpps when deployed in a Catalyst 6500 chassis using both fabric-enabled line cards and the Crossbar Switch Fabric Module (SFM).

The fabric-enabled line cards and SFM can only be used in the Catalyst 6500 Series and not in the Catalyst 6000. Furthermore, the SFM requires the use of Supervisor Engine 2 or higher, but Supervisor Engine 2 can still operate without a SFM in the system, that is, when used in the Catalyst 6000 that does not support the SFM. The backplane architecture in Figure 7.2 shows no support of a SFM.

Supervisor Engine 2 is more suitable for deployment in the core of service provider and large enterprise networks. The major difference between Supervisor Engine 1A and Supervisor Engine 2 is that Supervisor Engine 2 supports topology-based forwarding tables with highly optimized lookup algorithms (also referred to as Cisco Express Forwarding (CEF)) implemented in hardware. We have already seen above that Supervisor Engine 1A supports only flow cache-based forwarding. The forwarding tables in Supervisor Engine 2 can also be distributed to the line cards if they support local forwarding engines – distributed forwarding.

As explained in Chapter 1, distributed forwarding is a forwarding method that is based on distributing the forwarding tables (created from the topology of the network rather than from traffic flow caching) to the line cards so that forwarding

can be done locally there. Supervisor Engine 2 comes in two configurations: Supervisor Engine 2 with EARL switching system integrated with PFC2, and Supervisor Engine 2 with EARL switching system integrated with PFC2/MSFC2.

7.4.2.1 Supervisor Engine 2 with EARL Switching System Integrated with PFC2 This configuration provides only Layer 2 forwarding with Layer 3 QoS and security ACLs and services, in addition to Private Virtual LAN (PVLAN) services. The Supervisor Engine 2 in this configuration has only a PFC2 installed (Figure 7.5, without any MSFC block). The PFC2 and PFC1 have similar functions (including Layer 3/4 QoS classification and security ACL filtering), although the PFC2 is two times faster than the PFC1 and can store more QoS and security ACLs in hardware.

With switch fabric-enabled line cards and a SFM installed in a system, the Supervisor Engine 2 with PFC2 is capable of forwarding packets and performing Layer 3/4 QoS classification and ACL filtering at speeds of up to 30 Mpps. However, with no MSFC present in this configuration to provide routing information, Layer 3 forwarding (routing) cannot be done.

As shown in Figure 7.5, the EARL switching system is actually integrated into the PFC2. The Layer 2 and ACL engine are combined to obtain a single Layer 2/Layer 4 engine. The capabilities of the Layer 2 forwarding engine are enhanced to include Layer 3/4 QoS classification and ACL filtering. The Layer 3 forwarding engine is not used for Layer 3 forwarding, because an MSFC2 is not present to generate the routing information required to populate the forwarding table.

7.4.2.2 Supervisor Engine 2 with EARL Switching System Integrated with PFC2/MSFC2 In this configuration, an MSFC2 is added to enable Layer 3 forwarding on the Supervisor Engine 2 with PFC2. Supervisor Engine 2 does not support the MSFC1. The Layer 3 forwarding engine on the PFC2 can now perform Layer 3 forwarding because routing information is now provided by the MSFC2 (Figure 7.5). This configuration enables Layer 2 forwarding with full Layer 3 routing and forwarding on the Catalyst 6000 and Catalyst 6500.

The addition of the MSFC allows the Layer 3 forwarding engine on the PFC2 to Layer 3 forward packets (e.g., inter-VLAN traffic) while also supporting all other features of the PFC, such as QoS classification and ACL filtering. The PFC2 and MSFC2 both use topology-based forwarding tables with optimized lookup mechanisms (Cisco Express Forwarding (CEF)). The MSFC2 is responsible for running the routing protocols, building the routing tables, and generating the appropriate CEF tables (which include the Forwarding Information Base (FIB) table and adjacency table) to be used by the PFC.

In this configuration, as soon as packets need to be Layer 3 forwarded, the Layer 3 forwarding engine in the PFC already has the necessary information in its forwarding table to forward the packet to the next hop, without having to send the first in a flow to the MSFC. This forwarding architecture avoids the problems associated with flow cache-based forwarding when operating in an environment that has a high number of short flow connections being established in very short time intervals.

7.4.2.3 Details of Packet Processing in the Supervisor Engine 2 The MSFC2 in Supervisor Engine 2 does not forward IP packets (apart from exception packets directed to it by the PFC2). Instead, the MSFC2 constructs and maintains the main copy of a distributed forwarding table (also known as a Forwarding Information Base). The FIB contains the most important information required for packet forwarding and this information is distilled from the routing table created by the routing protocols running in the MSFC2. The MSFC2 copies the FIB it generates directly to the forwarding hardware in the PFC2 so that all packets are forwarded in the PFC2 hardware and not by the MSFC2 software process.

It is important to note that a flow cache is also generated in the PFC2, but this flow cache is used for statistics collection (e.g., as in NetFlow) and not for Layer 3 forwarding of packets.

7.4.3 Multilayer Switching Highlights in Cisco Catalyst Switches

This section summarizes the main features of multilayer switching in the Catalyst 6000/6500 switches and other Cisco Catalyst switches to be discussed in later chapters. The discussion here sets the context for the forwarding methods used here and in the other architectures. Cisco Catalyst switches support two methods of hardware-based Layer 3 forwarding, some aspects of which have already been described above for Supervisor Engines 1A and 2. The methods differ in how the data plane components of Layer 3 forwarding can get the necessary control plane information required to forward packets.

7.4.3.1 Front-End Processor Approach with Flow-Based Forwarding This method (called Multilayer Switching (MLS) by Cisco) represents the first method of hardware-based Layer 3 forwarding used by Cisco Catalyst switches. The method uses a flow-based model to populate a flow cache that includes the necessary control plane information required for the data plane to Layer 3 forward a packet. A flow simply represents a stream of IP packets, each sharing a number of identical parameters, such as the same source/destination IP address, or same source/destination TCP port or a combination of these.

An MLS Route Processor (MLS-RP) (i.e., the MSFC) provides control plane operations, while an MLS Switching Engine (MLS-SE) (i.e., the PFC) provides data plane operation. MLS requires that the first packet of a new flow (i.e., candidate packet) received by the MLS-SE be forwarded to the MLS-RP. The MLS-RP then makes a Layer 3 forwarding decision using a software process operating as part of its control plane and forwards the packet in software to its next hop but the packet exits the switch via the MLS-SE.

The MLS-SE receives the Layer 3 forwarding instructions in the returned processed (first) packet (i.e., enabler packet) from the MLS-RP that is on its way to the next hop. The MLS-SE then populates the flow cache with the forwarding information required to Layer 3 forward subsequent packets that belong

to the flow associated with the first packet. Subsequent packets received by the MLS-SE can then be Layer 3 forwarded in MLS-SE hardware without requiring the packets to be sent to the MLS-RP because the flow cache now has the required forwarding information.

7.4.3.2 Distributed Forwarding Approach (aka Cisco Express Forwarding)

This forwarding method, which uses optimized lookup algorithms and network topology-based distributed forwarding tables (method referred to by Cisco as CEF), is the newer generation of hardware-based Layer 3 forwarding. This is the preferred forwarding method used by modern-day Cisco and other high-performance switch/routers and routers.

In the CEF architecture, the forwarding (or CEF) table is prepopulated with all the necessary Layer 3 forwarding information (distilled from the routing table). This allows the Layer 3 forwarding engine ASIC to forward all IP packets in hardware, unlike in the MLS approach that requires the first packet of a flow to be forwarded in software by the MLS-RP.

The CEF architecture is more efficient and scalable and avoids the performance limitations of MLS method in environments where thousands of new short flows are established in very short time intervals. The CEF architecture is very scalable because the main CEF table information can be distributed to multiple Layer 3 forwarding engines. This means that a switch/router or router can perform multiple Layer 3 forwarding operations simultaneously, one per CEF table and forwarding engine.

The route processor (control plane) component of switch/router or router is responsible for generating the information in the CEF table and updating it as network routing topology changes occur. The CEF table can be viewed as consisting of two tables: the Layer 3 forwarding table and the adjacency table that hold the Layer 2 addresses of the next hops and directly attached hosts.

7.5 CATALYST 6000 LINE CARD ARCHITECTURES

The Catalyst 6000 and 6500 employ two types of port ASICs for network connectivity. The PINNACLE ASIC is designed for Gigabit Ethernet network ports (Figure 7.6), and the COIL ASIC for 10/100 Mb/s Ethernet ports (Figure 7.7). These port ASICs provide connectivity from a network to the 32 Gb/s main shared switching bus or the 16 Gb/s local bus supported on the fabric-enabled and fabric-only line cards.

The port ASICs also support the Catalyst 6000 and 6500 congestion management mechanisms. The PFC on the Supervisor Engine (or the DFC on fabric-enabled line cards) is responsible for instructing these port ASICs on how a packet should be classified and queued for QoS processing.

As illustrated in Figure 7.6, each PINNACLE ASIC supports four Gigabit Ethernet ports and provides congestion management with per port buffering. For

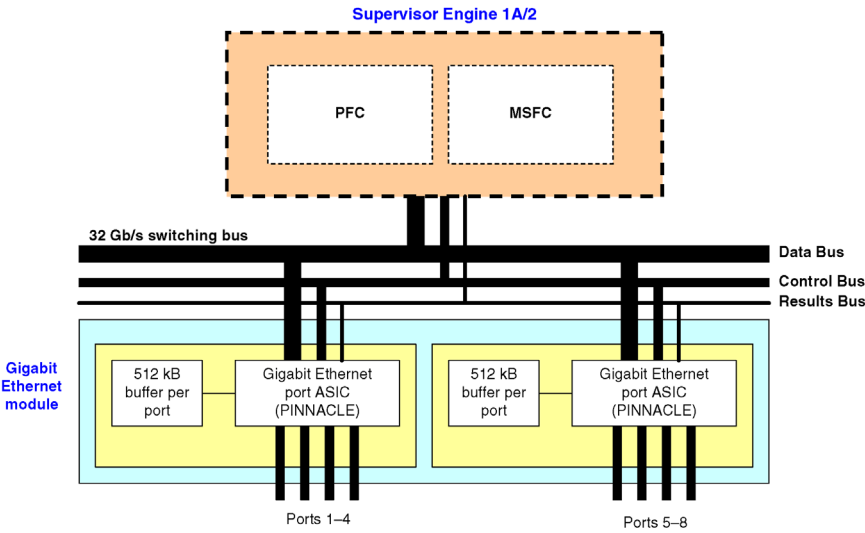


FIGURE 7.6 PINNACLE Gigabit Ethernet (GbE) Port ASIC and Buffering.

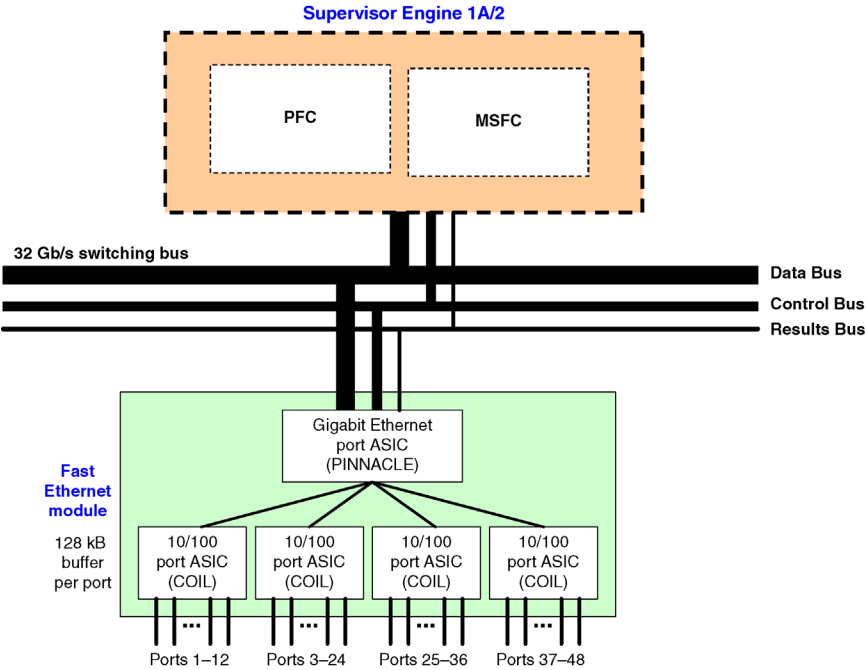


FIGURE 7.7 COIL Port ASIC Configuration on 10/100 line cards.

example, each 16-port Gigabit Ethernet line card holds four PINNACLE ASICs with each PINNACLE ASIC allocated a 512 kB buffer per port.

To prevent head-of-line (HOL) blocking to the shared switching bus fabric, a smaller amount of buffer is allocated to the receive (RX) queue (which is the queue that accepts frames coming from the network into the switch). A larger share of the buffering is allocated to the transmit (TX) queue (which is the queue that transmits frames from the switch to the network).

The allocation is done such that the ratio between the amount of buffers assigned to the transmit queue and receive queues is 7:1, which results in 448 kB of TX queue buffer and 64 kB of RX queue buffer. This buffer allocation strategy makes the Catalyst 6000 essentially an output queuing switch.

The PINNACLE ASIC handles QoS processing by assigning each port two RX queues and three TX queues. One queue out of the three TX queues is served in a strict priority fashion, while the other two queues are served using a weighted round-robin (WRR) scheduler. With the strict priority scheduler, the strict priority queue is allocated a fixed and guaranteed amount of bandwidth that is configured at the output port scheduling logic located in the port ASIC.

With the WRR scheduler, the two remaining queues are given scheduling weights that are relative to each other (i.e., sum of the normalized weights is equal to one) and where each queue is given bandwidth at the outgoing port proportional to its weight. If the three TX queues are configured as strict priority, high-priority, and low-priority queues, then the default port bandwidth allocation configuration will have 15% for strict priority queue, 15% for high-priority queue, and 70% for low-priority queue.

As Figure 7.7 illustrates, each COIL ASIC has 12 10/100 Mb/s Ethernet ports. Each PINNACLE ASIC, in turn, supports four COIL ASICs resulting in 48 ports on the line card. The 10/100 Mb/s Ethernet ports using the COIL ASIC work with their attached PINNACLE ASIC to implement congestion management. These dual ASIC line cards rely on a combination of control mechanisms in the COIL and PINNACLE ASICs to carry out congestion management.

Similar to the Gigabit Ethernet line card described above, the COIL ASIC supports buffering on a per port basis with each 10/100 Mb/s Ethernet port in the system allocated 128 kB of buffer. This 128 kB buffer is in turn divided between the TX and RX queues in a 7:1 ratio. The smaller RX buffer is used to prevent any HOL blocking problems as discussed above.

7.6 PACKET FLOW IN THE CATALYST 6000 WITH CENTRALIZED FLOW CACHE-BASED FORWARDING

This section describes the flow of packets through the Catalyst 6000 (with the 32 Gb/s shared switching bus) and a centralized flow cache maintained by the Supervisor Engine 1A. The processing steps are described in Figures 7.8–7.10.

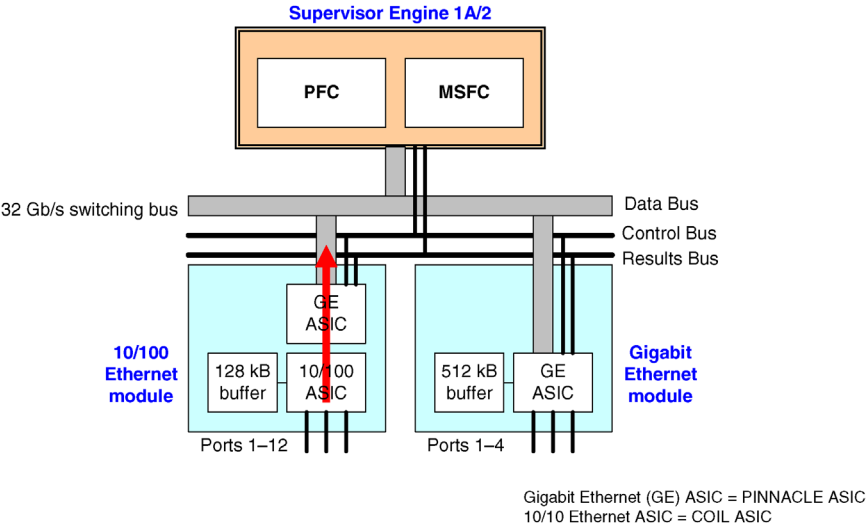


FIGURE 7.8 Step 1: Packet enters the switch from network.

Step 1 (Figure 7.8): Packet Enters an Input Port on the Switch from the Network

- A packet from the network enters an input port and is temporarily stored in the RX buffer. The packet is held in the RX buffer while the PINNACLE ASIC arbitrates for access to the 32 Gb/s shared switching bus.

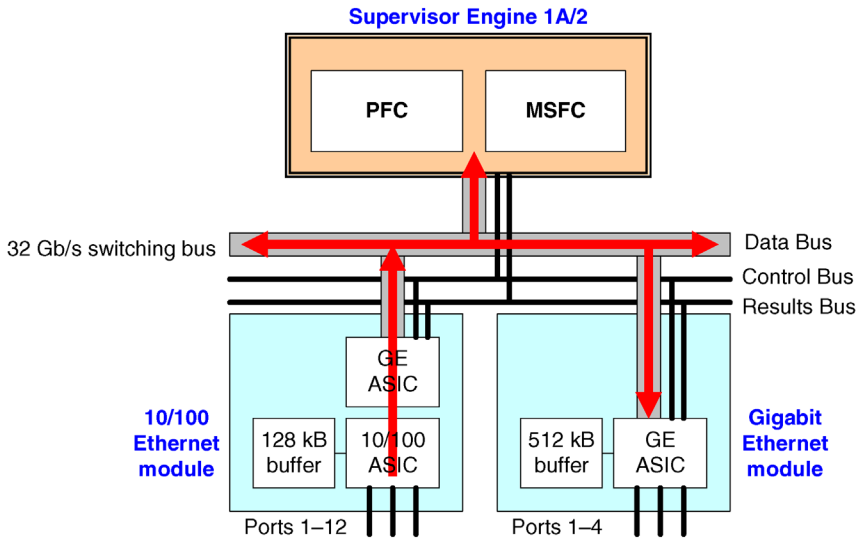


FIGURE 7.9 Step 2: Packet sent across 32 Gb/s switching bus and lookup takes place in Supervisor Engine.

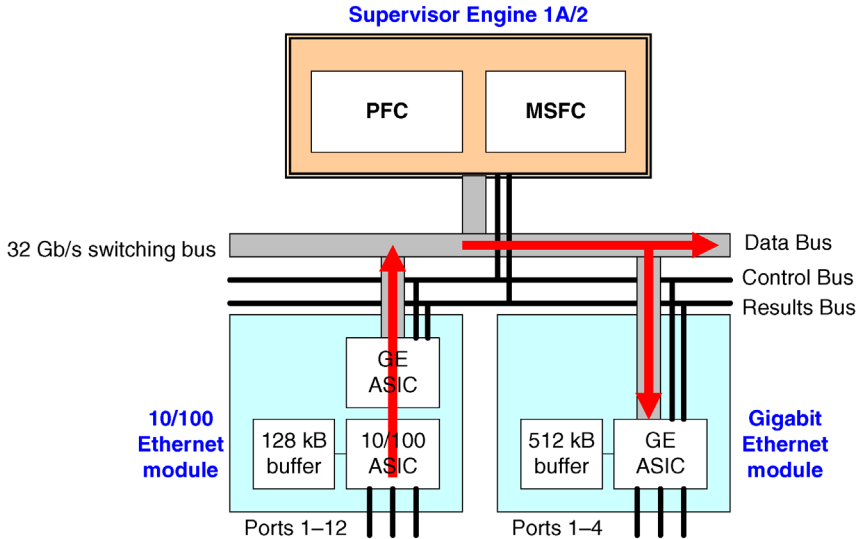


FIGURE 7.10 Step 3: Packet forwarded from switch to network.

- The 32 Gb/s shared switching bus is a shared medium allowing all the ports and modules connected to it to sense a transmitted packet as it propagates on the bus. Each line card has a local arbitration mechanism that allows each port on each PINNACLE ASIC to request for access to the shared switching bus.
- The local arbitration mechanism communicates with the central arbitration mechanism on the Supervisor Engine (see Figure 7.3), which then determines when each local arbitration mechanism is allowed to transmit packets on the shared switching bus.

Step 2 (Figure 7.9): Packet Sent Across the 32 Gb/s Switching Bus and Forwarding Table Lookup Takes Place in the Supervisor Engine

- Once the PINNACLE ASIC has been granted access to the shared switching bus by the central arbitration mechanism, the packet is transmitted across the bus.
- As the packet propagates along the shared switching bus, all the connected ports and modules start copying that packet into their TX buffers.
- The connected PFC, which also monitors the shared switching bus, senses the transmitted packet and initiates a forwarding table lookup process. First, the PFC references its Layer 2 forwarding table to determine if Layer 2 forwarding is required. If the packet is destined to a station located in the same VLAN as the source station served by the switch, then Layer 2 forwarding is carried out.

- However, if the Layer 2 destination address in the packet is the MSFC's registered MAC address, then the Layer 3 forwarding Engine in the PFC checks its flow cache to determine if a forwarding entry exists for the packet. If no flow entry exists, the packet is sent to the MSFC for further processing. But if a flow entry does exist in the flow cache, the PFC uses the packet's destination IP address to perform a lookup in the flow cache for the next hop MAC address, outbound port, and VLAN associated with the packet's destination.

Step 3 (Figure 7.10): Packet Forwarded from the Switch Through the Outbound Port to the Network

- After the lookup process above, the Supervisor Engine identifies the outgoing port for the packet. The Supervisor Engine also informs (over the Results Bus (RBus)) all the nondestination ports on the switch to flush the packet from their buffers.
- The RBus also conveys to the destination port or ports (in the case of multicast traffic) the outgoing Ethernet frame MAC address rewrite information (source MAC and next hop MAC addresses) and the relevant QoS instructions to be used to queue the packet correctly on the exit port.
- Upon receiving the packet, the PINNACLE ASIC on the destination port places the packet in the correct TX queue. The ASIC then uses its strict priority and WRR schedulers to transmit the framed packet out of its buffer on its way to the destination.