

Cisco 12000 Series Internet Router Architecture: Packet Switching

Document ID: 47320

Contents

Introduction

Prerequisites

- Requirements

- Components Used

- Conventions

Background Information

Packet Switching: Overview

Packet Switching: Engine 0 and Engine 1 Line Cards

Packet Switching: Engine 2 Line Cards

Packet Switching: Switching Cells across Fabric

Packet Switching: Transmitting Packets

Packet Flow Summary

Related Information

Introduction

This document examines the most important architectural elements of the Cisco 12000 Series Internet Router — switching packets. Switching packets are radically different from any of the shared memory or bus-based Cisco architectures. By using a crossbar fabric, the Cisco 12000 provides very large amounts of bandwidth and scalability. Furthermore, the 12000 uses virtual output queues to eliminate the Head of Line Blocking within the switch fabric.

Prerequisites

Requirements

There are no specific requirements for this document.

Components Used

The information in this document is based on the following hardware:

- Cisco 12000 Series Internet Router

The information in this document was created from the devices in a specific lab environment. All of the devices used in this document started with a cleared (default) configuration. If your network is live, make sure that you understand the potential impact of any command.

Conventions

For more information on document conventions, see the Cisco Technical Tips Conventions.

Background Information

(The switching decision on a Cisco 12000 is done by the line cards (LCs). For some LCs, a dedicated Application-Specific Integrated Circuit (ASIC) actually switches the packets. Distributed Cisco Express Forwarding (dCEF) is the only switching method available.

Remark: Engines 0, 1, and 2 are not the latest engines developed by Cisco. There are also Engine 3, 4, and 4+ line cards, with more to follow. Engine 3 line cards are capable of performing Edge features at line rate. The higher the Layer 3 engine, the more packets get switched in hardware. You can find some useful information about the different line cards available for the Cisco 12000 Series Router and the engine on which they are based at Cisco 12000 Series Internet Router: Frequently Asked Questions.

Packet Switching: Overview

Packets are always forwarded by the ingress line card (LC). The egress LC only performs outbound Quality of Service (QoS) that is queue-dependent (for example, Weighted Random Early Detection (WRED) or Committed Access Rate (CAR)). Most of the packets are switched by the LC using distributed Cisco Express Forwarding (dCEF). Only the control packets (such as routing updates) are sent to the Gigabit Route Processor (GRP) for processing. The packet switching path depends on the type of switching engines used on the LC.

This is what happens when a packet comes in:

1. A packet comes into the physical layer interface module (PLIM). Various things happen here:
 - ◆ A transceiver turns optical signals into electrical ones (most CSR line cards have fiber connectors)
 - ◆ L2 framing is removed (SANE, Asynchronous Transfer Mode (ATM), Ethernet, High-Level Data Link Control (HDLC)/Point-to-Point Protocol – PPP)
 - ◆ ATM cells are reassembled
 - ◆ Packets that fail the cyclic redundancy check (CRC) are discarded
2. As the packet is received and processed, it is Direct Memory Accessed into a small (approximately 2 x maximum transmission unit (MTU) buffer) memory called the "First In, First Out (FIFO) burst memory". The amount of this memory depends on the type of LC (from 128 KB to 1 MB).
3. Once the packet is completely in FIFO memory, an application-specific integrated circuit (ASIC) on the PLIM contacts the Buffer Management ASIC (BMA) and asks for a buffer to put the packet in. The BMA is told what size the packet is, and allocates a buffer accordingly. If the BMA cannot get a buffer of the right size, the packet is dropped and the "ignore" counter is incremented on the incoming interface. There is no fallback mechanism as with some other platforms. While this is going on, the PLIM could be receiving another packet in the FIFO burst memory, which is why it is 2xMTU in size.
4. If there is a free buffer available in the right queue, the packet is stored by the BMA in the free queue list of the appropriate size. This buffer is placed on the Raw Queue, which is examined by the Salsa ASIC or the R5K CPU. The R5K CPU determines the destination of the packet by consulting its local dCEF table in Dynamic RAM (DRAM), and then moves the buffer from the Raw Queue to a ToFabric queue corresponding to the destination slot.

If the destination is not in the CEF table, the packet is dropped. If the packet is a control packet (for example, routing updates), it is enqueued to the queue of the GRP and will be processed by the GRP. There are 17 ToFab queues (16 unicast, plus 1 Multicast). There is one toFab queue per line card (this includes the RP). These queues are known as "virtual output queues", and are important so that head-of-line blocking doesn't occur.

5. The ToFab BMA cuts the packet up into 44-byte pieces, which are the payload for what will eventually be known as "Cisco Cells". These cells are given an 8-byte header and 4-byte buffer

header by the frFab BMA (total data size so far = 56 bytes), and then enqueued into the proper ToFab queue (at which point, the #Qelem counter in the pool the buffer came from goes down by one, and the ToFab queue counter goes up by one).

The "decision maker" depends on the type of switching engines:

On Engine 2+ cards, a special ASIC is used to improve the way the packets are switched. Normal packets (IP/Tag, no options, checksum) are processed directly by the Packet Switching ASIC (PSA), then bypass the raw queue/CPU/Salsa combination and are enqueued directly onto the toFab queue. Only the first 64 bytes of the packet are passed through the Packet Switching ASIC. If the packet cannot be switched by the PSA, the packet is enqueued to the RawQ to be handled by the CPU of the LC as explained previously.

At this point, the switching decision has been made and the packet has been enqueued onto the proper ToFab output queue.

6. The toFab BMA DMA's (Direct Memory Access) the cells of the packet into small FIFO buffers in the fabric interface ASIC (FIA). There are 17 FIFO buffers (one per ToFab queue). When the FIA gets a cell from the toFab BMA, it adds an 8-byte CRC (total cell size = 64 bytes; 44 bytes payload, 8 bytes cell header, 4 bytes buffer header). The FIA has serial line interface (SLI) ASICs that then perform 8B/10B encoding on the cell (like the Fiber Distributed Data Interface (FDDI) 4B/5B), and prepares to transmit it over the fabric. This may seem like a lot of overhead (44 bytes of data gets turned into 80 bytes across the fabric!), but it is not an issue since fabric capacity has been provisioned accordingly.
7. Now that an FIA is ready to transmit, the FIA requests access to the fabric from the currently active card scheduler and clock (CSC). The CSC works on a rather complex fairness algorithm. The idea is that no LC is allowed to monopolize the outgoing bandwidth of any other card. Note that even if an LC wants to transmit data out of one of its own ports, it still has to go through the fabric. This is important because if this didn't happen, one port on an LC could monopolize all bandwidth for a given port on that same LC. It'd also make the switching design more complicated. The FIA sends cells across the switch fabric to their outgoing LC (specified by data in the Cisco Cell header put there by the switching engine).

The fairness algorithm is also designed for optimal matching; if card 1 wants to transmit to card 2, and card 3 wants to transmit to card 4 at the same time, this happens in parallel. That's the big difference between a switch fabric and a bus architecture. Think of it as analogous to an Ethernet switch versus a hub; on a switch, if port A wants to send to port B, and port C wants to talk to port D, those two flows happen independently of each other. On a hub, there are half-duplex issues such as collisions and backoff and retry algorithms.

8. The Cisco Cells that come out of the fabric go through SLI processing to remove the 8B/10B encoding. If there are any errors here, they'd appear in the show controller fia command output as "cell parity". See How To Read the Output of the show controller fia Command for additional information.
9. These Cisco Cells are DMA'd into FIFOs on the frFab FIAs, and then into a buffer on the frFab BMA. The frFab BMA is the one that actually does the reassembly of cells into a packet.

How does the frFab BMA know what buffer to put the cells in before it reassembles them? This is another decision made by the incoming line card switching engine; since all queues on the entire box are the same size and in the same order, the switching engine just has the Tx LC put the packet in the same number queue from which it entered the router.

The frFab BMA SDRAM queues can be viewed with the show controller frfab queue command on the LC. See How To Read the Output of the **show controller frfab | tofab queue** Commands on a Cisco 12000 Series Internet Router for details.

This is basically the same idea as the toFab BMA output. Packets come in and are placed in packets that are dequeued from their respective free queues. These packets are placed into the from-fabric queue, enqueued on either the interface queue (there is one queue per physical port) or the rawQ for output processing. Not much happens in the rawQ: per-port multicast replication, Modified Deficit Round Robin (MDRR) – same idea as Distributed Weighted Fair Queuing (DWFQ), and output CAR. If the transmit queue is full, the packet is dropped and the output drop counter is incremented.

10. The frFab BMA waits until the TX portion of the PLIM is ready to send a packet. The frFab BMA does the actual MAC rewrite (based, remember, on information contained in the Cisco Cell header), and DMA's the packet over to a small (again, 2xMTU) buffer in the PLIM circuitry. The PLIM does the ATM SAR and SONET encapsulates, where appropriate, and transmits the packet.
11. ATM traffic is reassembled (by the SAR), segmented (by the tofab BMA), reassembled (by the fromfab BMA) and segmented again (by the fromfab SAR). This happens very quickly.

That is the lifecycle of a packet, from beginning to end. If you want to know what a GSR feels like at the end of the day, read this entire paper 500,000 times!

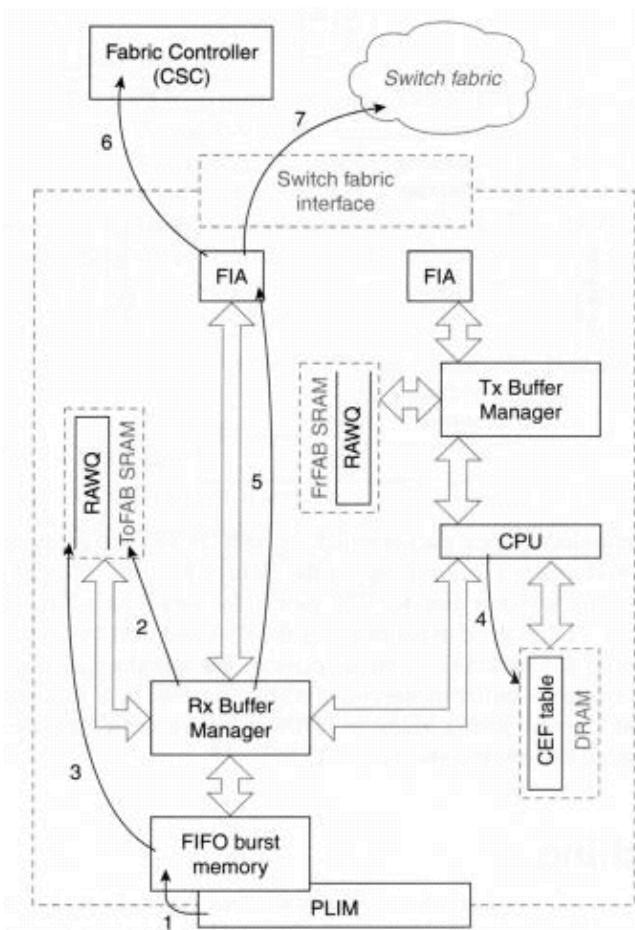
The packet switching path on the GSR depends on the type of forwarding engine on the LC. Now we'll go through all the steps for Engine 0, Engine 1, and the two LCs.

Packet Switching: Engine 0 and Engine 1 Line Cards

The sections below are based on the book *Inside Cisco IOS Software Architecture*, Cisco Press.

Figure 1 below illustrates the different steps during packet switching for an Engine 0 or Engine 1 LC.

Figure 1: Engine 0 and Engine 1 Switching Path



The switching path for the Engine 0 and Engine 1 LC is essentially the same, although the Engine 1 LC has an enhanced switching engine and buffer manager for increased performance. The switching path is as follows:

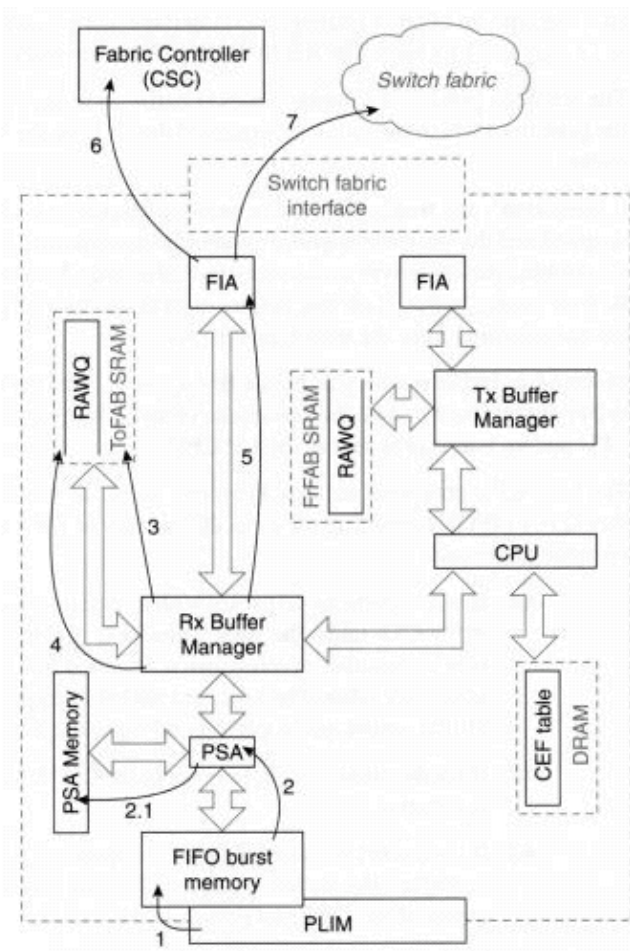
- **Step 1** – The interface processor (PLIM) detects a packet on the network media and begins copying it into a FIFO memory called **burst memory** on the LC. The amount of burst memory each interface has depends on the type of LC; typical LCs have 128 KB to 1 MB of burst memory.
- **Step 2** – The interface processor requests a packet buffer from the receive BMA; the pool from which the buffer is requested depends on the length of the packet. If there aren't any free buffers, the interface is dropped and the interface's "ignore" counter is incremented. For example, if a 64-byte packet arrives into an interface, the BMA tries to allocate an 80-byte packet buffer. If no free buffers exist in the 80-byte pool, buffers are not allocated from the next available pool.
- **Step 3** – When a free buffer is allocated by the BMA, the packet is copied into the buffer and is enqueued on the **raw queue** (RawQ) for processing by the CPU. An interrupt is sent to the LC CPU.
- **Step 4** – The LC's CPU processes each packet in the RawQ as it is received (the RawQ is a FIFO), consulting the local distributed Cisco Express Forwarding table in DRAM to make a switching decision.
 - ◆ **4.1** If this is a unicast IP packet with a valid destination address in the CEF table, the packet header is rewritten with the new encapsulation information obtained from the CEF adjacency table. The switched packet is enqueued on the virtual output queue corresponding to the destination slot.
 - ◆ **4.2** If the destination address is not in the CEF table, the packet is dropped.
 - ◆ **4.3** If the packet is a control packet (a routing update, for example), the packet is enqueued on the virtual output queue of the GRP and processed by the GRP.
- **Step 5** – The receive BMA fragments the packet into 64-bytes cells, and hands these off to the FIA for transmission to the outbound LC.

At the end of Step 5, the packet that arrived into an Engine 0/1 LC has been switched and is ready to be transported across the switch fabric as cells. Go to Step 6 in the section Packet Switching: Switching Cells across Fabric.

Packet Switching: Engine 2 Line Cards

Figure 2 below illustrates the packet switching path when the packets arrive into an Engine 2 LC, as described in the following list of steps.

Figure 2: Engine 2 Switching Path



- **Step 1** – The interface processor (PLIM) detects a packet on the network media and begins copying it into a FIFO memory called **burst memory** on the LC. The amount of burst memory each interface has depends on the type of LC; typical LCs have 128 KB to 1 MB of burst memory.
- **Step 2** – The first 64 bytes of the packet, called the header, are passed through the Packet Switching ASIC (PSA).
 - ◆ **2.1** The PSA switches the packet by consulting the local CEF table in the PSA memory. If the packet cannot be switched by the PSA, go to Step 4; otherwise, continue to Step 3.
- **Step 3** – The Receive Buffer Manager (RBM) accepts the header from the PSA and copies it into a free buffer header. If the packet is larger than 64 bytes, the tail of the packet is also copied into the same free buffer in packet memory and is queued on the outgoing LC virtual output queue. Go to Step 5.
- **Step 4** – The packet arrives at this step if it cannot be switched by the PSA. These packets are placed on the **raw queue** (RawQ) and the switching path is essentially the same as for the Engine 1 and Engine 0 LC from this point (Step 4 in the case of Engine 0). Note that the packets that are switched by the PSA are never placed in the RawQ and no interrupt is sent to the CPU.
- **Step 5** – The Fabric Interface Module (FIM) is responsible for segmenting the packets into Cisco Cells and sending the cells to the Fabric Interface ASIC (FIA) for transmission to the outbound LC.

Packet Switching: Switching Cells across Fabric

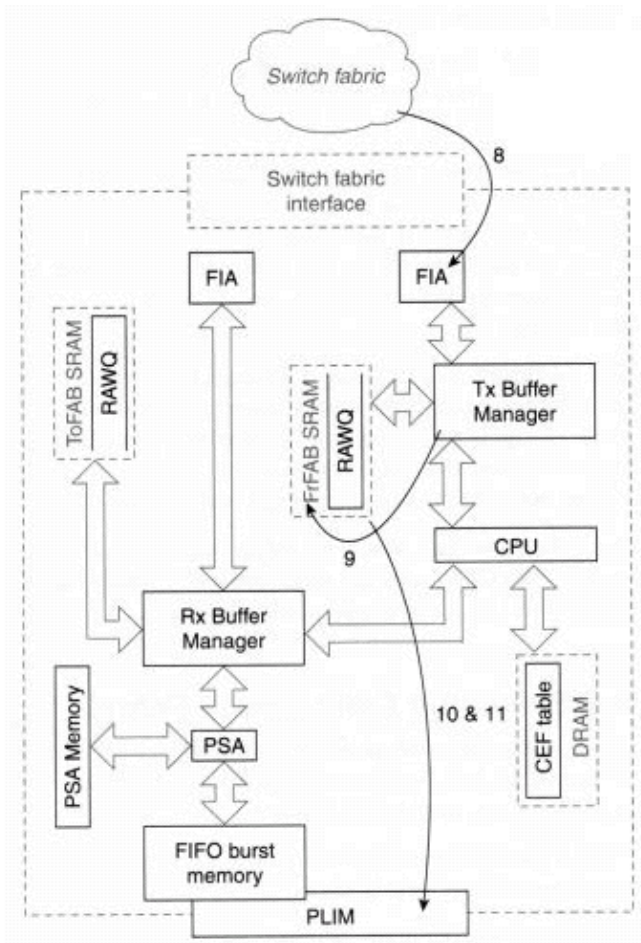
You arrive at this stage after the packet switching engine switches the packets. At this stage, the packets are segmented into Cisco Cells and are waiting to be transmitted across the switching fabric. The steps for this stage are as follows:

- **Step 6** – The FIA sends a grant request to the CSC, which schedules each cell's transfer across the switch fabric.
- **Step 7** – When the scheduler grants access to the switch fabric, the cells are transferred to the destination slot. Note that the cells might not be transmitted all at once; other cells within other packets might be interleaved.

Packet Switching: Transmitting Packets

Figure 3 below shows the last stage of packet switching. The cells are reassembled and the packet is transmitted onto the media. This takes place on the outbound line card.

Figure 3: Cisco 12000 Packet Switching: Transmit Stage



- **Step 8** – The cells switched across the fabric arrive into the destination line card through the FIA.
- **Step 9** – The transmit Buffer Manager allocates a buffer from the transmit packet memory and reassembles the packet in this buffer.
- **Step 10** – When the packet is rebuilt, the transmit BMA enqueues the packet onto the destination interface's transmit queue on the LC. If the interface transmit queue is full (the packet cannot be enqueued), the packet is dropped and the **output queue drop** counter is incremented.

Note: In the transmit direction, the only time packets are placed in the RawQ is when the LC CPU needs to do any processing before transmission. Examples include IP fragmentation, multicast, and output CAR.

- **Step 11** – The interface processor detects a packet waiting to be transmitted, dequeues the buffer from the transmit memory, copies it into internal FIFO memory, and transmits the packet on the media.

Packet Flow Summary

IP packets that traverse the 12000 are processed in three phases:

- Ingress Line Card in three sections:
 - ◆ Ingress PLIM (Physical Line Interface Module) – Optical to Electrical conversion, Synchronous Optical Network (SONET)/Synchronous Digital Hierarchy (SDH) un–framing, HDLC, and PPP processing.
 - ◆ IP Forwarding – Forwarding decision based on FIB lookup and queuing into one of the ingress unicast queues or multicast queues.
 - ◆ Ingress Queue management and Fabric Interface – Random Early Detection (RED)/Weighted Random Early Detection (WRED) processing on the ingress queues and de–queuing towards the fabric in order to maximize fabric utilization.
- Switching IP packets through the 12000 fabric from ingress card to egress card or egress cards (in case of multicast).
- Egress Line Card in three sections:
 - ◆ Egress Fabric Interface – Reassembling the IP packets to be sent and queuing into egress queues; processing multicast packets.
 - ◆ Egress queue management – RED/WRED processing on the ingress queues and de–queuing towards the egress PLIM to maximize the egress line utilization.
 - ◆ Egress PLIM – HDLC and PPP processing, SONET/SDH framing, Electrical to Optical conversion.

Related Information

- [Cisco 12000 Series Internet Router Architecture – Chassis](#)
- [Cisco 12000 Series Internet Router Architecture – Switch Fabric](#)
- [Cisco 12000 Series Internet Router Architecture – Route Processor](#)
- [Cisco 12000 Series Internet Router Architecture – Line Card Design](#)
- [Cisco 12000 Series Internet Router Architecture – Memory Details](#)
- [Cisco 12000 Series Internet Router Architecture – Maintenance Bus, Power Supplies and Blowers, and Alarm Cards](#)
- [Cisco 12000 Series Internet Router Architecture – Software Overview](#)
- [Understanding Cisco Express Forwarding](#)
- [Technical Support – Cisco Systems](#)

[Contacts & Feedback](#) | [Help](#) | [Site Map](#)

© 2009 – 2010 Cisco Systems, Inc. All rights reserved. [Terms & Conditions](#) | [Privacy Statement](#) | [Cookie Policy](#) | [Trademarks of Cisco Systems, Inc.](#)

Updated: Jul 07, 2005

Document ID: 47320
