Cisco 12000系列互联网路由器体系结构：分组交换

目录

简介

本文检查Cisco 12000SERIES互联网路由器的最重要的结构上元素--交换信息包。交换信息包是完全不同的与任何共享内存或基于总线的思科体系结构。通过使用Crossbar结构， Cisco 12000提供非常很多带宽和可扩展性。此外， 12000使用虚拟输出队列排除在交换矩阵内的Head of Line封闭。

先决条件

要求

本文档没有任何特定的要求。

使用的组件

本文档中的信息基于下列硬件：

- Cisco 12000 系列互联网路由器

本文档中的信息都是基于特定实验室环境中的设备编写的。本文档中使用的所有设备最初均采用原始（默认）配置。如果您使用的是真实网络，请确保您已经了解所有命令的潜在影响。

规则

有关文档规则的详细信息，请参阅 Cisco 技术提示规则。

背景信息

(在Cisco 12000的交换决定由线卡(LCs)完成。对于一些LCs，专用的Application-specific integrated circuit (ASIC)实际上转换数据包。Distributed Cisco Express Forwarding (DCEF)是唯一的交换方法联机。

注释：引擎0， 1和2不是思科开发的最新的引擎。有也引擎3， 4和4+线卡，与跟随的更多。引擎 3 板卡能够以线路速率执行边缘功能。第 3 层引擎越高，在硬件中交换的数据包就越多。您能找到关于不同线路卡的一些有用的信息可用为他们在Cisco 12000SERIES互联网路由器的Cisco 12000系列路由器和引擎：常见问题。

分组交换：概述

数据包由进入线路卡(LC)总是转发。出口LC只执行例如是依据队列的出站服务质量(QoS) (加权随机早期检测(WRED)或承诺接入速率(CAR))。使用Distributed Cisco Express Forwarding (DCEF)，大多数据包由LC交换。仅控制数据包(例如路由更新)被发送对处理的千兆路由处理器(GRP)。信息包交换的路径取决于在LC使用的交换引擎种类。

这是发生了什么，当数据包来在：

1. 数据包进入物理层接口模块(PLIM)。多种事发生此处：收发器把光信号变成电一个(多数CSR线卡有光纤连接器)L2帧删除(神志正常，异步传输模式(ATM)、以太网，高级数据链路控制(HDLC)/点对点协议- PPP)ATM信元被重新召集失败循环冗余冗余校验的数据包(CRC)丢弃
2. 因为数据包接收并且处理，它是直接存储器访问到呼叫“先入先出(FIFO)分段存储的”一个小(大约2个x最大传输单元(MTU)缓冲区)内存。相当数量此内存取决于LC种类(从128 KB到1 MB)。
3. 一旦数据包完全在FIFO内存，在PLIM的application-specific integrated circuit (ASIC)与缓冲区管理ASIC (BMA)联系并且请求缓冲区放置数据包。BMA告诉什么大小数据包是，并且相应地分配缓冲区。如果BMA不能获得适当大小的缓冲区，数据包丢弃，并且"ignore"计数器在流入接口被增加。没有回退机制如同一些其他平台。当这继续时， PLIM可能接收在FIFO突发内存的另一数据包，是它为什么在大小上是2xMTU。
4. 如果有在正确队列的一空闲缓存联机，数据包由在适当的大小的自由队列列表的BMA存储。此缓冲区在自然状态的队列被放置，由萨尔萨ASIC或R5K CPU检查。R5K CPU通过咨询其在动态RAM (DRAM)的本地dCEF表确定数据包的目的地，然后移动缓冲区从自然状态的队列向ToFabric队列与目的地地址槽相应。如果目的地不在CEF表里，数据包丢弃。如果数据包是控制数据包(例如，路由更新)，排队对GRP的队列，并且由GRP处理。有17个Tofab队列(16单播，加上1组播)。有每线卡一个tofab队列(这包括RP)。这些队列叫作“虚拟输出队列”，并且是重要，以便head-of-line封闭不发生。
5. Tofab BMA剪切数据包成44字节片段，是什么的有效负载最终叫作“思科信元”。这些信元由frfab BMA给8字节报头和4字节缓冲报头(到目前为止全部数据估量= 56个字节)，然后排队到适当的ToFab队列(到时，在缓冲区来自的池的#Qelem计数器由一个断开，并且Tofab队列计数器由一个上升)。“作决策者”取决于交换引擎种类：在引擎2+卡，特殊ASIC用于改进数据包交换的方式。正常数据包(IP/Tag、没有选项，校验和)直接地由Packet Switching ASIC (PSA)处理，然后绕过原始queue/CPU/Salsa组合和排队直接地在tofab队列上。数据包的仅前64个字节通过分组交换ASIC通过。如果数据包不可能由PSA交换，数据包被排列对LC的CPU将处理的RawQ如以前解释。这时，交换决定做了，并且数据包被排列了在适当的Tofab输出输出队列上。

6. tofab BMA DMA (直接存储器访问)数据包的信元到在矩阵接口ASIC (FIA)的小FIFO缓冲区里。有17个FIFO缓冲区(一每个Tofab队列)。当FIA从tofab BMA时获得信元，添加一个8字节CRC (总信元大小- 64个字节;44个字节有效载荷、8个字节信元头， 4个字节缓冲报头)。FIA有串行线路然后进行8B/10B在信元的编码的接口(SLI) ASIC (类似光纤分布式数据接口(FDDI) 4B/5B)，并且准备在结构传送它。这也许似乎类似很多开销(44字节的数据获得把变成在结构间的80个字节!)，但是它不是问题，因为结构产能相应地设置了。
7. 即然FIA准备传送， FIA请求对结构的访问从当前活跃的卡调度器和时钟(CSC)。CSC研究相当复杂公平算法。想法是LC没有允许垄断流出的带宽其他卡。注意，即使LC要从其自己的端口之一当中传送数据，必须仍然通过结构。这是重要，因为，如果这没有发生， LC的一个端口可能垄断一个给的端口的所有带宽该同样LC的。它也将做复杂化的交换设计。FIA发送在交换矩阵间的信元对他们的流出的LC (指定由在交换引擎放置的那里思科信元头的数据)。公平算法为最佳匹配也设计;如果card1要传送到卡2，并且卡3要同时传送到卡4，这平行发生。那是在交换矩阵和总线体系结构之间的大差值。认为它如类似于一台以太网交换机与集线器;在交换机，如果端口A要发送到端口B和端口C要与端口D谈，那两个流独立彼此发生。在集线器上，有半双工问题例如冲突和回退并且再试算法。
8. 从结构出来的思科信元通过`SLI处理删除8B/10B编码。如果那里此处任何错误，他们在show controller fia命令输出中将出现作为“信元奇偶校验”。请参阅如何阅读输出show controller fia命令其他信息的。
9. 这些思科信元是DMA'd到在frfab FIAs的FIFO，然后到在frfab BMA的一缓冲区。frfab BMA是实际上执行信元重组到数据包的那个。frfab BMA如何了解放置信元的什么缓冲区，在重新召集他们前？这是流入线路卡交换引擎做出的另一决定;因为在整个方框的所有队列是相同大小和按同一顺序，交换引擎在进入路由器的同一个编号队列安排Tx LC放置数据包。frfab BMA SDRAM队列可以用show controller frfab queue命令查看在LC。请参阅如何阅读**show controller frfab的**输出|在一个Cisco 12000SERIES互联网路由器的tofab队列命令关于详细信息。这基本上是想法和tofab BMA输出一样。数据包在从他们的各自自由队列离队的数据包进来和安置。这些数据包被放置到from-fabric队列，排队在接口队列(有每个物理端口一个队列)或输出处理的rawQ。并非在rawQ发生：每端口组播复制、改进的差额轮询(MDRR) -想法和分布式加权公平排队(DWFQ)一样和输出控制访问率。如果传输队列满，数据包丢弃，并且输出丢弃计数器被增加。
10. frfab BMA等待，直到PLIM的TX部分准备发送数据包。frfab BMA执行实际MAC重写(基于，请记住，在思科信元头包含的信息)和DMA数据包到在PLIM电路的一小(再， 2xMTU)缓冲区。PLIM执行ATM SAR，并且SONET封装，只要适合的话，并且传送数据包。
11. ATM流量被重新召集(由SAR)，被分段(由tofab BMA)，被重新召集(由fromfab BMA)并且再被分段(由fromfab SAR)。这非常迅速发生。

那是数据包的生命周期，自始至终。如果要了解什么GSR感觉类似当晚，读此整个文章500,000次!

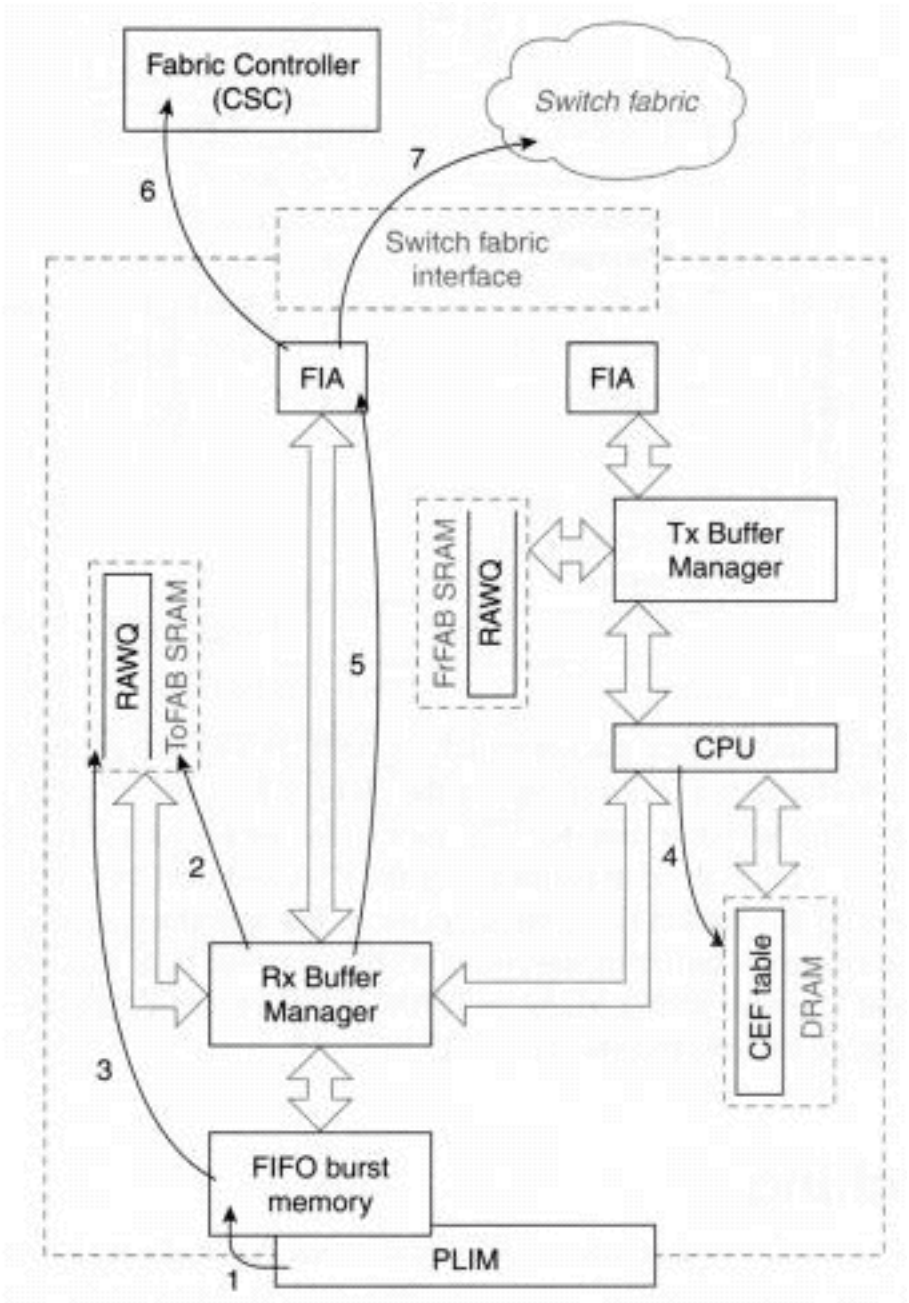
GSR的信息包交换的路径取决于转发引擎种类在LC的。现在我们将通过引擎0，引擎1和两个LCs的所有步骤。

分组交换：引擎0和引擎1线卡

下面的部分根据书*在Cisco IOS软件软件结构里面*， Cisco出版社。

在引擎0或引擎1 LC的，数据包交换期间下面的图1说明不同的步骤。

图 1：引擎0和引擎1交换路径



引擎0和引擎1 LC的交换路径根本是相同的，虽然引擎1 LC有一个增强版交换引擎和缓冲区管理程序更完善的性能的。交换路径如下：

- **Step1** -接口处理器(PLIM)检测在网络媒介的一数据包并且开始复制它到呼叫在LC的**分段存储的**FIFO内存。每个接口有的相当数量分段存储取决于LC种类;典型LCs有128 KB对分段存储1 MB。
- **步骤2** -接口处理器请求从接收BMA的一数据包缓冲;缓冲区请求的池取决于数据包的长度。如果没有任何空闲缓存，接口丢弃，并且接口的"ignore"计数器被增加。例如，如果64字节数据包到达到接口， BMA设法分配80字节数据包缓冲。如果空闲缓存在80字节池不存在，缓冲区从下个可用的池没有分配。
- **步骤3** -当BMA时分配空闲缓存，数据包在**自然状态的队列**(RawQ)复制到缓冲区和被排列处理的由CPU。中断发送对LC CPU。
- **步骤4** - LC的CPU进程在RawQ的每数据包，因为接收(RawQ是FIFO)，咨询在DRAM的本地分布式Cisco快速转发模式表做出交换决定。**4.1**如果这是与一有效目的地址的一单播IP数据包在CEF表里，信息包报头以从CEF邻接表得到的新的封装信息重写。交换数据包在虚拟输出输出队列被排列与目的地地址槽相应。**4.2**如果目的地址不在CEF表里，数据包丢弃。**4.3**如果数据包是控制数据包(例如路由更新)，数据包在GRP的虚拟输出输出队列被排列并且由GRP处理。

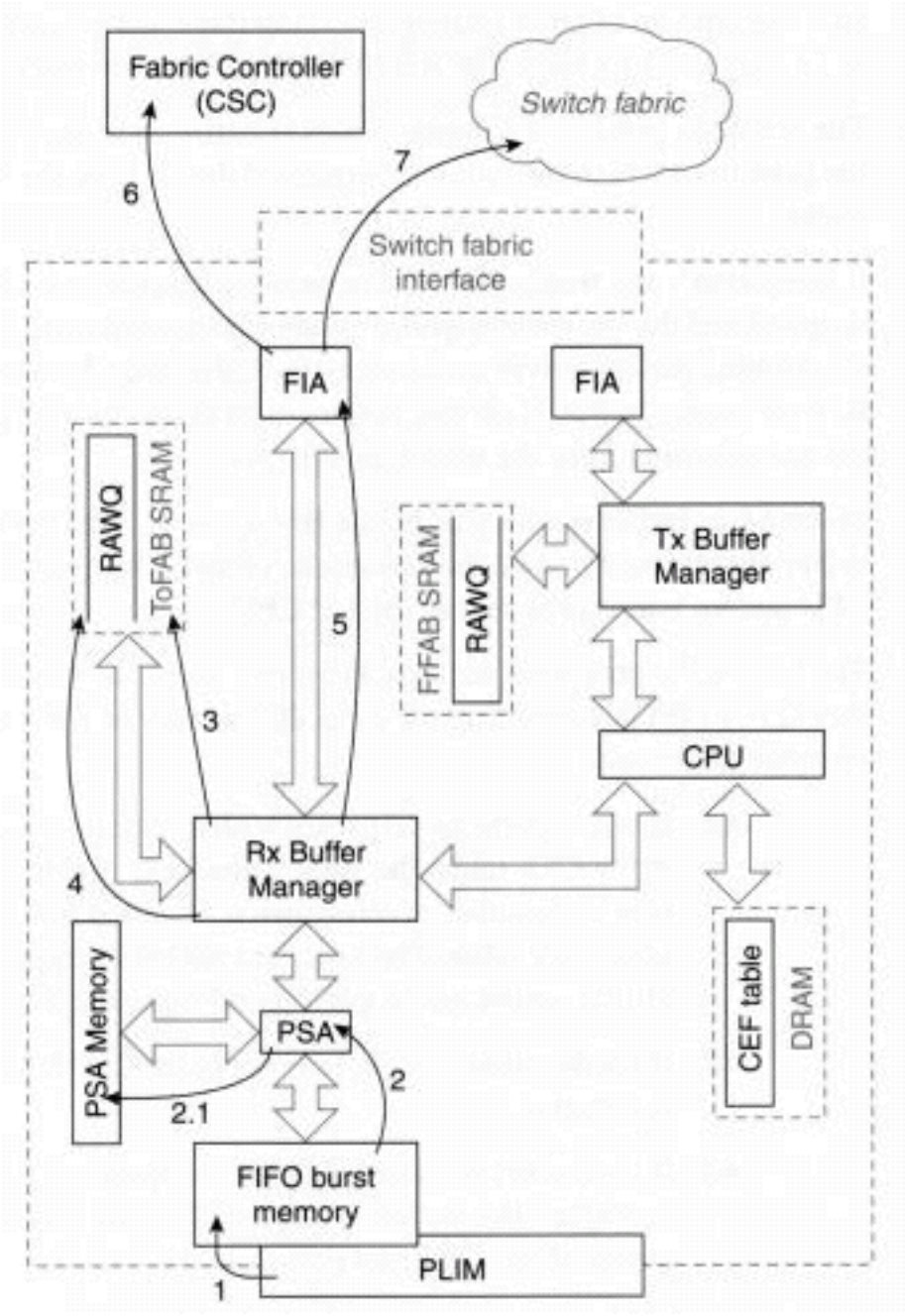
- **步骤5** -接收BMA分段数据包到64字节信元，并且递交这些对发射的FIA对出站LC。

在步骤5结束时，到达到引擎0/1 LC的数据包交换并且准备在交换矩阵间传输作为信元。进入在部分Packet Switching的步骤6 ：跨结构交换信元。

分组交换：引擎2线卡

下面的图2说明信息包交换的路径，当数据包到达到引擎2 LC时，正如步骤所描述以下列表。

图 2：引擎2交换路径



- **Step1** -接口处理器(PLIM)检测在网络媒介的一数据包并且开始复制它到呼叫在LC的**分段存储的**FIFO内存。每个接口有的相当数量分段存储取决于LC种类;典型LCs有128 KB对分段存储1 MB。
- **步骤2** -数据包的前64个字节，呼叫报头，通过Packet Switching ASIC (PSA)通过。**2.1** PSA通过咨询在PSA内存的本地CEF表转换数据包。如果数据包不可能由PSA交换，请进入步骤4;否则，请继续对步骤3。

- **步骤3** - Receive Buffer Manager (RBM)接受从PSA的报头并且复制它到空闲缓存报头。如果数据包大于64个字节，数据包的尾标在流出的LC虚拟输出输出队列也复制到在数据包内存的同一空闲缓存和排队。进入步骤5。
- **步骤4** -，如果不可能由PSA，交换数据包到达在此步骤。这些数据包在**自然状态的队列** (RawQ)被放置，并且交换路径根本是相同的象为从此点(一旦引擎0)的步骤4的引擎1和引擎0 LC。注意由PSA交换的数据包在RawQ和没有中断从未安置发送对CPU。
- **步骤5** - Fabric Interface Module (FIM)对分段数据包到思科信元和发送信元负责对发射的矩阵接口ASIC (FIA)对出站LC。

分组交换：跨结构交换信元

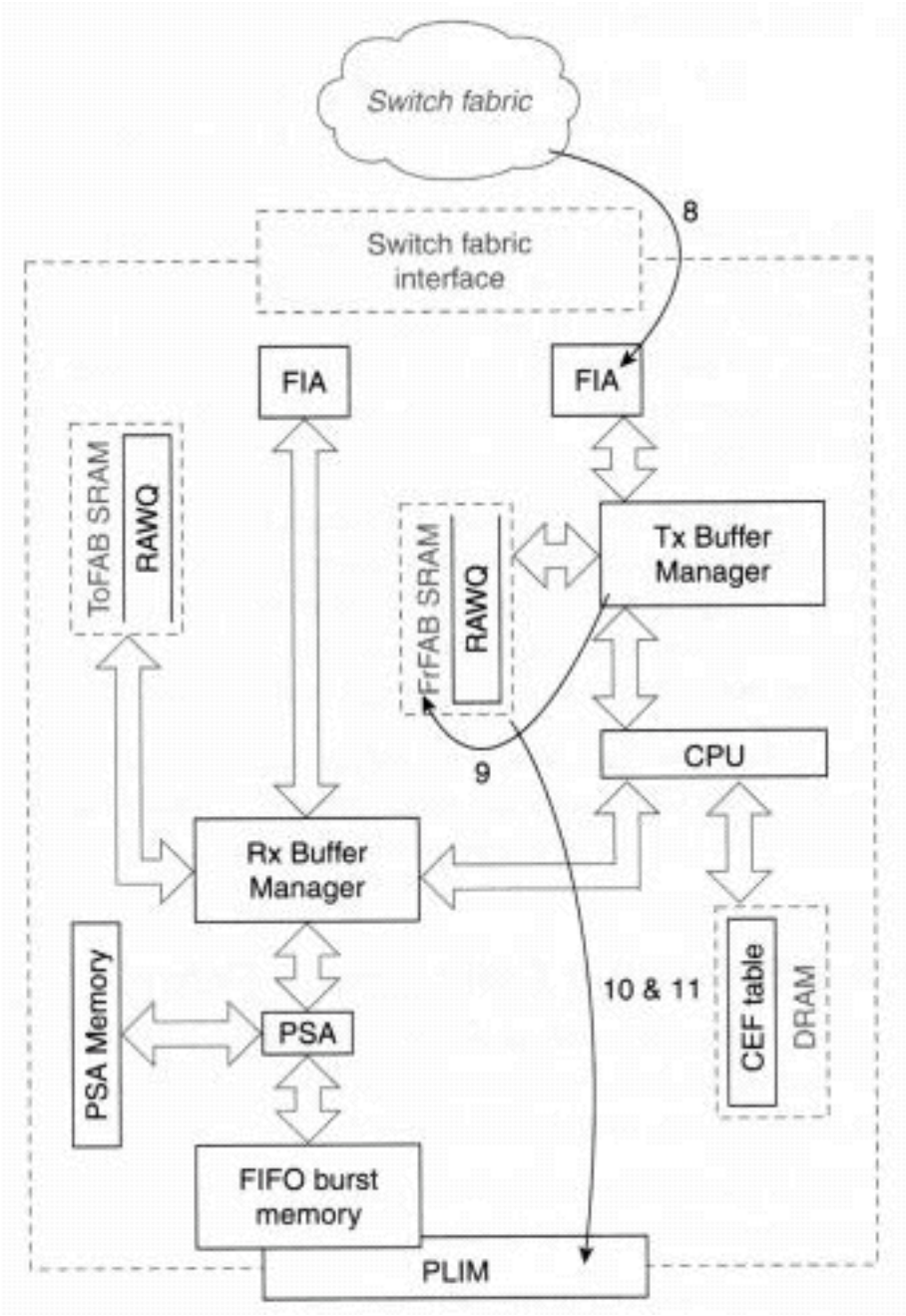
在信息包交换的引擎转换数据包后，您在此阶段到达。在此阶段，数据包被分段到思科信元和等待在交换结构间传送。此阶段的步骤如下：

- **步骤6** - FIA发送授予请求对CSC，安排在交换矩阵间的每个信元的转移。
- **步骤7** -当调度器准许对交换矩阵时的访问，信元转接对目的地地址槽。注意信元也许不同时传送;也许插入在其他数据包内的其他信元。

分组交换：传送信息包

下面的图3显示数据包交换最后阶段。信元被重新召集，并且数据包传送在媒体上。这在出局线路卡发生。

图 3：Cisco 12000数据包交换：传输阶段



- **步骤8** -在结构间交换的信元到达到目的地线路卡通过FIA。
- **步骤9** -传输缓冲区管理程序从传输数据包内存分配一缓冲区并且重新组装在此缓冲区的数据包。
- **步骤10** -当数据包重建时，传输BMA排列在目的地接口的传输队列上的数据包在LC。如果接口传输队列满(数据包不可能被排列)，数据包丢弃，并且**输出队列丢弃**计数器被增加。**注意：** 在传送方向，当数据包在RawQ之时安置是，当LC CPU需要在发射前执行其中任一处理。示例包括IP分段、组播和输出控制访问率。
- **步骤11** -等待的接口处理器检测数据包传送，离队从传输内存的缓冲区，复制它到内部FIFO内存，并且传送在媒体的数据包。

信息包流汇总

横断12000的IP信息包在三个相位内处理：

- 在三个部分的进入线路卡：入口PLIM (物理线路接口模块) -光学对电子转换，同步光网络(SONET) /Synchronous数字体系(SDH) un-framing、HDLC和Ppp处理。IP转发-根据FIB查找和

队列的转发决策到其中一个入口单播队列或组播队列。入口队列管理和矩阵接口-随机早期检测处理在入口队列和离队往结构的/Weighted随机早期检测(WRED)为了最大化结构利用率。
- 交换IP信息包通过12000结构从进入卡到输出卡或输出卡(在组播的情况下)。
- 在三个部分的出口线路卡：出口矩阵接口-重新组装IP信息包是发送和排队到出口队列;处理组播信息包。出口队列管理-在入口队列的RED/WRED处理和离队往出口PLIM最大化出口线路利用率。出口PLIM - HDLC和Ppp处理， SONET/SDH构建帧，对光转换的电。

相关信息

- 技术支持 - Cisco Systems