Cisco 12000系列互联网路由器体系结构：分组交换

目录

简介

本文档研究了Cisco 12000系列互联网路由器最重要的架构元素 — 交换数据包。交换数据包与任何共享内存或基于总线的思科架构都截然不同。通过使用纵横式交换矩阵，Cisco 12000可提供非常大的带宽和可扩展性。此外，12000使用虚拟输出队列来消除交换矩阵内的线路头阻塞。

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

(Cisco 12000上的交换决策由线卡(LC)完成。 对于某些LC，专用专用专用集成电路(ASIC)实际上会交换数据包。分布式思科快速转发(dCEF)是唯一可用的交换方法。

备注：引擎0、1和2不是思科开发的最新引擎。还有引擎3、4和4+线卡，后面还有更多。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。您可以找到一些有用信息，说明Cisco 12000系列路由器可用的不同线卡及其基于Cisco 12000系列互联网路由器的引擎，这些线卡包括：常见问题.

分组交换：概述

数据包始终由入口线卡(LC)转发。 出口LC仅执行与队列相关的出站服务质量(QoS)(例如，加权随机早期检测(WRED)或承诺接入速率(CAR))。 大多数数据包由LC使用分布式思科快速转发(dCEF)交换。 仅控制数据包（如路由更新）会发送到千兆路由处理器(GRP)进行处理。分组交换路径取决于LC上使用的交换引擎类型。

当数据包进入时，会发生以下情况：

1. 数据包进入物理层接口模块(PLIM)。 这里发生了各种情况：收发器将光信号转换为电信号（大多数CSR线卡具有光纤连接器）删除L2成帧(SANE、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信元重组不通过循环冗余校验(CRC)的数据包将被丢弃
2. 在接收和处理数据包时，它被直接存储器访问到称为“先进先出(FIFO)突发存储器”的小型(大约2 x maximum transmission unit(MTU) buffer)存储器。 此内存的大小取决于LC的类型（从128 KB到1 MB）。
3. 一旦数据包完全在FIFO内存中，PLIM上的专用集成电路(ASIC)就会与缓冲区管理ASIC(BMA)联系，并请求缓冲区将数据包放入。BMA会被告知数据包的大小，并相应地分配缓冲区。如果BMA无法获得大小适当的缓冲区，则丢弃数据包，并在传入接口上增加“ignore”计数器。没有其他平台的回退机制。在这种情况下，PLIM可能正在FIFO突发内存中接收另一个数据包，这就是它大小为2xMTU的原因。
4. 如果在正确的队列中有可用的空闲缓冲区，则BMA会将数据包存储在大小适当的空闲队列列表中。此缓冲区放在Salsa ASIC或R5K CPU检查的原始队列上。R5K CPU通过在动态RAM(DRAM)中查询其本地dCEF表来确定数据包的目的地，然后将缓冲区从原始队列移动到与目标插槽对应的ToFabric队列。如果目的地不在CEF表中，则丢弃数据包。如果数据包是控制数据包（例如，路由更新），则它将入队到GRP的队列，并由GRP处理。有17个ToFab队列（16个单播，加1个组播）。 每个线路卡有一个toFab队列（包括RP）。 这些队列称为“虚拟输出队列”，它们非常重要，因此不会发生行首阻塞。
5. ToFab BMA将数据包分割为44字节片段，这些片段是最终称为“思科信元”的负载。 这些信元由frFab BMA（到目前为止的总数据大小= 56字节）提供8字节报头和4字节缓冲区报头，然后入队到适当的ToFab队列(此时，缓冲区中的#Qelem计数器将关闭1,ToFab队列计数器将上行1)。“决策者”取决于交换引擎的类型：在引擎2+卡上，使用特殊ASIC来改进数据包的交换方式。普通数据包（IP/Tag、无选项、校验和）由分组交换ASIC(PSA)直接处理，然后绕过原始队列/CPU/Salsa组合，直接入队到toFab队列。只有数据包的前64个字节通过分组交换ASIC。如果数据包无法由PSA交换，则数据包将入队到RawQ，由LC的CPU处理，如前所述。此时，已做出交换决策，并且数据包已排入适当的ToFab输出队列。

6. toFab BMA DMA（直接内存访问）将数据包的单元在交换矩阵接口ASIC(FIA)中转换为小型FIFO缓冲区。 有17个FIFO缓冲区（每个ToFab队列一个）。 当FIA从toFab BMA获取信元时，它会添加一个8字节CRC(信元总大小 — 64字节；44字节负载、8字节信元报头、4字节缓冲区报头)。 FIA具有串行线路接口(SLI)ASIC，这些ASIC在信元上执行8B/10B编码(如光纤分布式数据接口(FDDI)4B/5B)，并准备通过交换矩阵传输。这看起来可能会产生大量开销（44字节的数据在整个交换矩阵中转换为80字节！），但这不是问题，因为交换矩阵容量已相应调配。
7. 现在FIA已准备好传输，FIA请求从当前活动的卡调度程序和时钟(CSC)访问交换矩阵。CSC使用一种相当复杂的公平算法。其思想是，不允许LC独占任何其他卡的传出带宽。请注意，即使LC希望从自己的一个端口传输数据，它仍必须通过交换矩阵。这很重要，因为如果不发生这种情况，LC上的一个端口可能会独占同一LC上给定端口的所有带宽。这也使交换设计变得更加复杂。FIA通过交换矩阵将信元发送到其传出LC（由交换引擎放置的思科信元报头中的数据指定）。公平性算法也设计为最优匹配；如果卡1想向卡2传输，而卡3想同时向卡4传输，则并行发生。这是交换矩阵和总线架构之间的巨大差异。将其视为类似于以太网交换机与集线器；在交换机上，如果端口A要发送到端口B，而端口C要与端口D通信，则这两个流会彼此独立地发生。在集线器上，存在半双工问题，例如冲突和回退和重试算法。
8. 从交换矩阵中取出的思科信元通过SLI处理来删除8B/10B编码。如果此处存在任何错误，它们会以“信元奇偶校验”的形式显示在show controller fia命令输出中。 有关其他信息，请参阅如何读取show controller fia命令的输出。
9. 这些Cisco单元是DMA'd在frFab FIA上进入FIFO，然后进入frFab BMA上的缓冲区。frFab BMA实际上是将信元重组到数据包中。frFab BMA如何知道在重组信元之前将信元放入哪个缓冲区？这是传入线卡交换引擎做出的另一个决定；由于整个框中的所有队列大小相同且顺序相同，因此交换引擎只让Tx LC将数据包放入其进入路由器的相同编号队列中。在LC上使用show controller frfab queue命令可以查看frFab BMA SDRAM队列。请参阅如何读取show controller frfab的输出 | Cisco 12000系列Internet路由器上的tofab queue命令以了解详细信息。这与toFab BMA输出基本相同。数据包进入并放入从各自空闲队列出列的数据包中。这些数据包被放入从交换矩阵队列，在接口队列（每个物理端口有一个队列）或rawQ上入队以进行输出处理。rawQ中不会发生太多情况：每端口组播复制、修改差额轮询(MDRR) — 与分布式加权公平队列(DWFQ)相同的理念，以及输出CAR。如果传输队列已满，则丢弃数据包并增加输出丢弃计数器。
10. frFab BMA会等待，直到PLIM的TX部分准备好发送数据包。frFab BMA执行实际MAC重写（请记住，根据思科信元报头中包含的信息），并将数据包DMA转换到PLIM电路中的小（同样是2xMTU）缓冲区。PLIM执行ATM SAR和SONET封装（如果适用）并传输数据包。
11. ATM流量重组（通过SAR）、分段（通过tofab BMA）、重组（通过fromfab BMA）和再次分段（通过fromfab SAR）。 这种情况发生得非常快。

这是数据包的生命周期，从头到尾。如果您想了解GSR在一天结束时的感觉，请阅读整篇论文50万次！

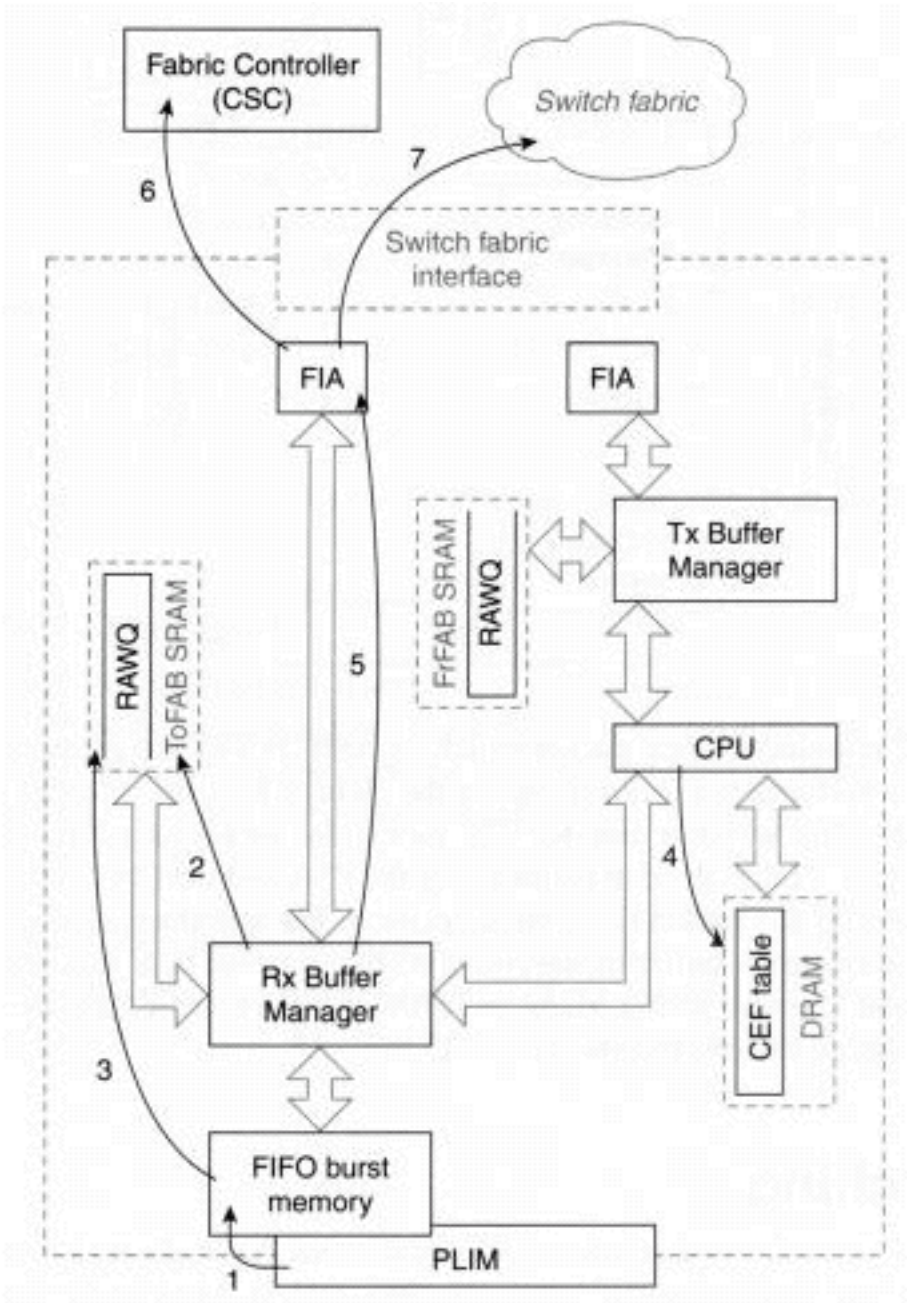
GSR上的分组交换路径取决于LC上转发引擎的类型。现在，我们将介绍引擎0、引擎1和两个LC的所有步骤。

分组交换：引擎0和引擎1线卡

以下各节基于Cisco Press的《*Inside Cisco IOS软件架构*》一书。

下图1说明了引擎0或引擎1 LC在分组交换过程中的不同步骤。

图 1：引擎0和引擎1交换路径



引擎0和引擎1 LC的交换路径基本相同，但引擎1 LC具有增强的交换引擎和缓冲区管理器以提高性能。交换路径如下：

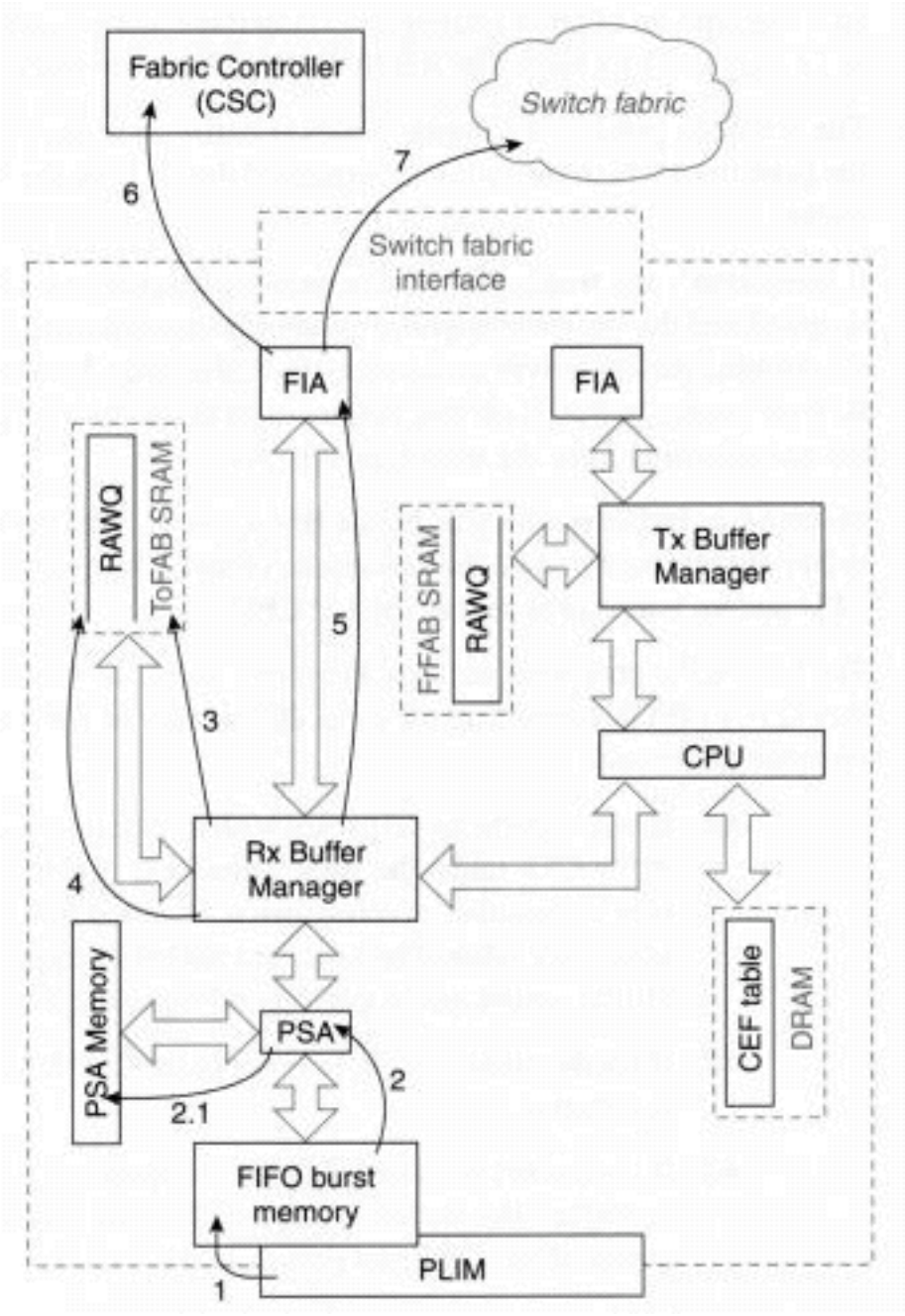
- **第1步** — 接口处理器(PLIM)在网络介质上检测数据包，并开始将其复制到LC上称为突**发内存**的FIFO内存。每个接口的突发内存量取决于LC的类型；典型的LC有128 KB到1 MB的突发内存。
- **第2步** — 接口处理器向接收BMA请求数据包缓冲区；从中请求缓冲区的池取决于数据包的长度。如果没有任何可用缓冲区，则会丢弃接口，并增加接口的“忽略”计数器。例如，如果64字节的数据包到达接口，BMA会尝试分配80字节的数据包缓冲区。如果80字节池中不存在空闲缓冲区，则不从下一个可用池分配缓冲区。
- **第3步** — 当BMA分配空闲缓冲区时，数据包被复制到缓冲区中，并在原始队列(RawQ)上排队，由CPU处理。中断被发送到LC CPU。
- **第4步** - LC的CPU在收到RawQ时处理该数据包（RawQ是FIFO），并咨询DRAM中的本地分布式Cisco快速转发表以做出交换决策。**4.1如**果这是单播IP数据包，其CEF表中具有有效目的地址，则使用从CEF邻接表获取的新封装信息重写数据包报头。交换分组在与目标插槽对应的虚拟输出队列上入队。**4.2如**果目的地址不在CEF表中，则丢弃数据包。**4.3如**果数据包是控制数据包（例如路由更新），则该数据包将入队到GRP的虚拟输出队列，并由GRP处理。
- **第5步** — 接收BMA将数据包分段为64字节的信元，并将这些信元交给FIA，以便传输到出站

LC。
在第5步结束时，到达引擎0/1 LC的数据包已进行交换，并准备作为信元通过交换矩阵传输。转至Packet Switching：部分的步骤6交换矩阵中的信元。

分组交换：引擎2线卡

下图2说明了当数据包到达引擎2 LC时的数据包交换路径，如以下步骤列表所述。

图 2：引擎2交换路径



- **第1步** — 接口处理器(PLIM)在网络介质上检测数据包，并开始将其复制到LC上称为突**发内存**的FIFO内存。每个接口的突发内存量取决于LC的类型；典型的LC有128 KB到1 MB的突发内存。
- **第2步** — 数据包的前64个字节（称为报头）通过分组交换ASIC(PSA)。**2.1 PSA**通过查看PSA内存中的本地CEF表来交换数据包。如果PSA无法交换数据包，请转至步骤4;否则，请继续步骤3。
- **第3步** — 接收缓冲区管理器(RBM)从PSA接受报头，并将其复制到空闲缓冲区报头。如果数据

包大于64字节，则数据包的尾部也复制到数据包内存中的同一空闲缓冲区中，并在传出LC虚拟输出队列中排队。转到第 5 步。
- **第4步** — 如果PSA无法交换数据包，则数据包到达此步骤。从此点开始，这些数据包被放置在原始队列**(RawQ)**上，交换路径基本上与引擎1和引擎0 LC的交换路径相同（在引擎0的情况下，步骤4）。 请注意，PSA交换的数据包从不放在RawQ中，也不会向CPU发送中断。
- **第5步** — 交换矩阵接口模块(FIM)负责将数据包分割为Cisco信元，并将信元发送到交换矩阵接口ASIC(FIA)，以便传输到出站LC。

分组交换：跨交换矩阵交换信元

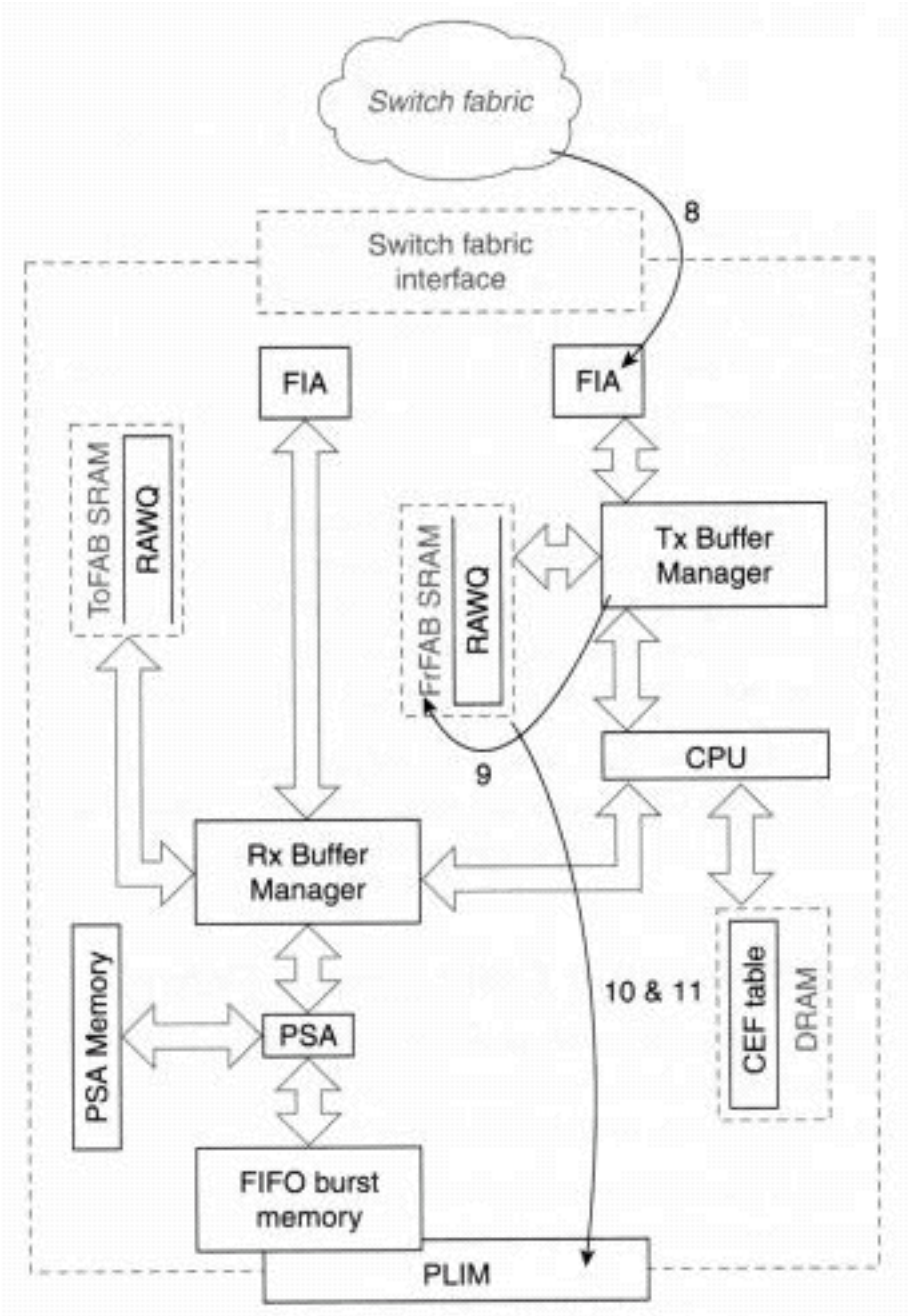
数据包交换引擎交换数据包后，您即到达此阶段。在此阶段，数据包被分段到思科信元，并等待通过交换矩阵传输。此阶段的步骤如下：

- **第6步**- FIA向CSC发送授权请求，CSC安排每个信元在交换矩阵中的传输。
- **第7步** — 当调度程序授予对交换矩阵的访问权限时，信元将被传输到目的插槽。请注意，信元可能不会一次全部传输；其他数据包中的其他信元可能是交错的。

分组交换：传输数据包

下图3显示了分组交换的最后阶段。信元重组后，数据包被传输到介质上。这发生在出站线卡上。

图 3：思科12000分组交换：传输阶段



- **第8步** — 通过交换矩阵交换的信元通过FIA到达目的线卡。
- **第9步** — 传输缓冲区管理器从传输数据包存储器中分配缓冲区，并在此缓冲区中重组数据包。
- **第10步** — 当数据包重建时，传输BMA将数据包排队到LC上目的接口的传输队列。如果接口传输队列已满（数据包无法入队），则丢弃数据包，并增加**输出队列**丢弃计数器。**注意**：在传输方向，在RawQ中放置数据包的唯一时间是LC CPU在传输前需要执行任何处理。示例包括IP分段、组播和输出CAR。
- **步骤11** — 接口处理器检测等待传输的数据包，将缓冲区从传输存储器中排队，将其复制到内部FIFO存储器中，并在介质上传输数据包。

信息包流汇总

通过12000的IP数据包分三个阶段处理：

- 入口线卡分为三个部分：入口PLIM（物理线路接口模块） — 光到电转换、同步光网络(SONET)/同步数字层次结构(SDH)取消成帧、HDLC和PPP处理。IP转发 — 根据FIB查找和队列到入口单播队列或组播队列之一的转发决策。入口队列管理和交换矩阵接口 — 入口队列上的

随机早期检测(RED)/加权随机早期检测(WRED)处理和向交换矩阵的去队列，以最大限度地提高交换矩阵利用率。
- 通过12000交换矩阵将IP数据包从入口卡交换到出口卡或出口卡（在组播情况下）。
- 出口线卡分为三部分：出口交换矩阵接口 — 重组要发送的IP数据包，并将其排队到出口队列；处理组播数据包。出口队列管理 — 在入口队列上进行RED/WRED处理，并向出口PLIM取消队列，以最大限度地提高出口线路利用率。出口PLIM - HDLC和PPP处理、SONET/SDH成帧、电气到光纤转换。

相关信息

- 技术支持 - Cisco Systems