

第一部分 概述

第1章 引言

产业界人士和观察家（甚至包括那些经过多年外层空间旅行刚刚返回这个世界的人）都已经很清楚，因特网（Internet）发展所达到的地位和其所产生的现象都不同于本世纪或上世紀所提出的任何一种技术。Internet的延伸和影响范围、有关Internet出版物、以及包括美国在线（AOL）、美国电报电话公司（AT&T）和微软公司等Internet产业界的大量风险投资者，这一切都会使我们有一种纷繁迷乱的感觉。所有这些都是通过这样或那样的方式与Internet连接起来。Internet也是Joe Sixpack和Fortune 1000这样的网站每天都关心、考虑和使用的唯一技术。或许Internet是世界上少有的几个能够以相同的平等程度来对待每一个用户的实体组织之一。一个企业的首席执行官（CEO）如果想给公司提供更好的网络服务保证，他必须建立一个专用网络。而在Internet中，每一个人对网络的访问都是平等的。

Internet的发展并没有损害到那些在过去150年中所发展起来的其他技术。的确，电话技术是相当重要的，它可以使我们能够在双方不见面的情况下通过声音与线路另一端的人通话。同样，汽车也改变了我们的生活，汽车的出现能够使我们在一天之内跨越更大的距离，而这个距离要比任何其他动物多出一个数量级。电灯、无线电和电视都曾经是改善我们日常生活的十分重要的技术，扩展了我们在非睡眠状态的时间，向我们传播各种信息，使我们享受更多的娱乐。我们已经在很大程度上解决了生存问题。大多数人的饭桌上有足够的食品、有温暖的住所，并且都有一个工作场所，可以每天早出晚归地工作。我们也可以不必被动地接收各种电视节目，而可以轻松地使用遥控器选择欣赏自己喜爱的频道。

Internet除了有把事情变得更好的能力外，也可能把事情搞得更糟。在好的一方面，Internet能够使我们在世界范围同人们进行对等通信；使我们能够访问那些存储在数以百万计的网络计算机上的几乎无限的大量信息。一些功能强大的搜索引擎能够使我们更加简单和迅速地实现对有用、有意义的信息资源的定位。不同阶段的商务活动，包括从最初的偶然兴趣直到成熟的采购定单等，都可以在Internet上完成。甚至于许多人已经开始幻想在将来的某天，Internet能使我们不再需要每天早起去上班了。人们可以靠在枕头上使用一台膝上型计算机（或许将来可能出现的任何先进的计算机）通过拨接Internet对所有的商务活动和某些消遣娱乐进行管理和维护。

在不利的一方面，Internet也可能使我们成为有电子怪癖的人，使我们缺乏与其他人进行直接交流的能力。人们仅有的非睡眠时间都将被耗费在计算机的荧光屏前，不停地键入Internet地址（URL）或指向其他的超级链接。最令人不安的是，由于“等待回应（WFRE, waiting for reply）”而浪费的时间是不可挽回的。WFRE现象的出现是由于Internet上太拥塞、太慢，以至于你的浏览器似乎进入了一个永久“等待回应”的状态。有时候它只是几秒钟的问题；另一些情况下可能是几分钟。你在WFRE状态下盯着计算机荧光屏等待所花费的时间

是相当大的，这些时间的总和可能会是一个令人吃惊的数字，其数量级或许是几个月甚至几年。我们所讨论的要点在于：

- 1) Internet已经经历了巨大的增长过程，并且这种增长将会继续。
- 2) 不论是居民用户或者是团体用户，Internet都受到了同等的欢迎。对于后者，Internet还意味着新的收入增长点。
- 3) 一些实力很强并且有创造力的产业巨头正在致力于Internet的应用，以便为其企业自身及其消费者提供有利条件。毋庸置疑，不论是偶尔对Internet的临时使用还是正式规范地应用Internet，都将导致对Internet更多的兴趣和广告宣传。与此同时，也将伴随着Internet应用和及其流量的成比例的增长。
- 4) 目前Internet的带宽和容量还是缺乏的，这导致了Internet上不稳定的响应时间和不可预知的性能。同时产生的问题是，Internet是否有能力支持未来的、高带宽需求的、时延敏感的应用？或者说Internet是否有能力支持居民对带宽容量的适度增长的需求？

我们是如何进入了这样一个不稳定的状态呢？这个问题有若干答案，但其中没有一个是真正有权威性的解释，或许还有一些是可以根本不考虑的。首先，Internet是其自身成功的一个受害者。每一天都有新的用户加入到Internet中，越来越多的人不停地使用浏览器通过一个Web站点搜寻他们所感兴趣的下一个Web站点。由于访问Internet的价格仅是电话的市话费用附加一个适度的费率，因此并没有一个价格上的保护手段来防止某些浏览者对Internet资源的长时间占用。另一种资源的缺乏不一定是由于网络资源的不足引起的，而更大程度上是由于服务器的资源不足造成的。对某些服务器或服务器阵列来说，突发性的连接请求所引起的负荷和突发的频度可能大大超过了这些服务器的处理能力。这种突发的大量的连接请求一般发生在大量的客户试图同时访问同一个Web服务器的时候。这个问题可以被认为是一个临时性的问题，因为服务器的供应商通常会不断地提供新型的内容服务器主机、负载均衡器、Web缓存器等来使该问题得到缓解^①。

另一个问题是某些链路可能正好没有足够的带宽来支持业务所提供的流量负荷。这个问题的部分解决方案当然是增加更多的带宽；一些新的技术，如波分复用（WDM）技术，似乎可以提供几乎无限的带宽。所有这些我们上述所讨论的问题都是造成Internet及Intranet（Intranet是Internet在企业范围内的一个著名的复制品）性能极其不稳定的重要因素。在这些问题上，有很多都已经被研究清楚了；虽然其中有些诸如价格等问题是不可能在一夜之间得到解决的，但是我们至少已经知道解决方案是存在的，并且可以在不久的将来得到应用。然而，有关Internet性能和基于IP协议进行网络互连的最基本问题，很大程度上还在于基本IP路由转发处理过程和该功能的实现平台。

1.1 早期的路由功能

路由功能在最初就已存在。这些机盒开始被称为网关（从一个网络到另一个网络的网关），后来被称为路由器，用来将不同的物理网络连接成一个更大的、统一的互连网络。每一个单个的信息分组都带有一个标识其目的网络的标签。路由器维护一个目的网络表，该表是一个存放已知目的网络地址信息的表结构。路由器接收到分组时，就在目的网络表中搜索与该分

^① 实际上，现在已经有了一些完美而巧妙的改善Web服务器性能的技术，例如采用代理服务器、Web缓存器、传输控制协议（TCP）连接路由器，以及可以使资源在整个网络中分布和共享的分布缓存器协议等。

组目的网络标签相匹配的表项。一旦找到匹配，路由器就可以将这个分组转发输出到一个适当的网络接口上，然后等待处理下一个到来的分组。对每一个到达路由器的单个分组都会执行这个相对简单的过程。即使大量具有相同目的地的分组连续到达路由器，路由器仍然以单个分组的方式来处理这些分组。

我们来考察图 1-1 所示的第一代路由器。它由单个的通用中央处理器和多个介质接口适配卡组成，处理器和适配卡都通过一条共享总线互相连接。处理器上运行路由协议，负责维护一个转发表，记录着到达分组所能转发到的下一跳路由器（next-hop router）。进入路由器的分组通过共享总线到达处理器中，在处理器中执行一个标准的路由表搜索过程来确定下一跳路由器。然后，分组再次经过共享总线传送到相应的输出接口上。这个系统的性能受到总线速度和中央处理器的处理能力的限制。此外，每一个分组需要两次经过总线。由于早期的 IP 网络都是基于软件实现的，并且与路由器的性能相比，早期的 IP 网络更注重于其连接性和可靠性，因此这种设计还是可以令人满意的。

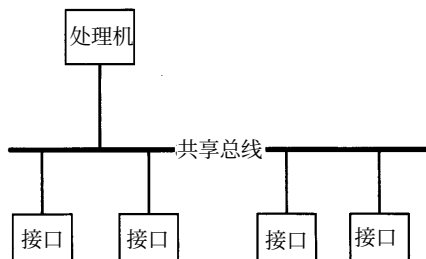


图1-1 第一代路由器

但是，随后 Internet 不断扩张。最初其增长速度还是比较平稳的，后来这种增长的速度加快了。越来越多的用户和越来越多的站点导致了越来越多网络的出现。网络数量的增加使得路由表不断增大，也导致了路由表查寻时间的增加。再加上用户流量的成比例增长，已经使高速路由器技术的引入和应用变得十分必要了。提高性能的一个通常的方法是将分组转发的计算过程分布到各个介质接口适配器中去。路由表的一部分或全部可以被缓存在输入适配器的存储器中。这使输入适配器能够进行转发路由的计算，并直接通过总线将分组传送到正确的输出适配器上，在这个过程中不需要中央处理器的干预。这种模型的性能仍然受到一定程度的限制。主要的制约因素是总线的速率，以及在路由表查寻过程中对大型表结构进行分类排序所耗费的时间。改善路由器技术的另一种方法是用一个交换机来替代路由器中的总线。这样，所有的输入和输出端口就可以通过一个无阻塞的交换机构互相连接起来。这个模型如

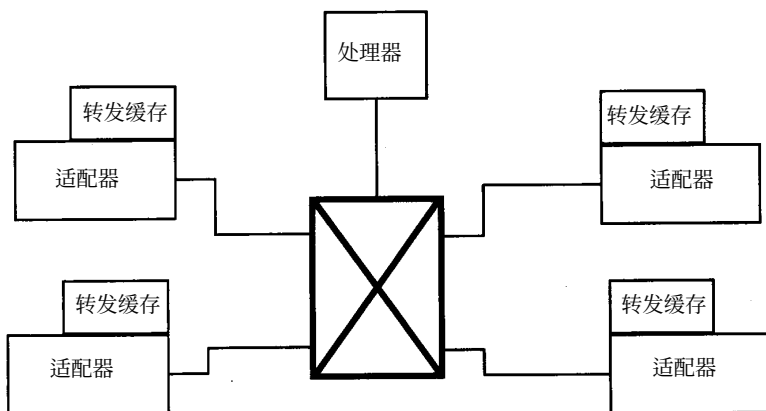


图1-2 具有交换机构的路由器

图1-2所示。在该模型中，尽管内部设计和效率有所提高，但是由于路由器自身所需的每个分组（per-packet）的处理过程以及在处理具有突发性和不可预测特征的 IP流过程中采用了先进先出（FIFO）的排队方式，会造成这些改进方案的价格很高，吞吐能力的增加有限，性能也不稳定。

1.2 ATM与IP

伴随着Internet的出现和增长，快速路由器技术在交换领域内引起了广泛的兴趣。宽带综合业务数据网（ISDN）曾经被视为是一种能支持通信端点之间进行实时通信的技术。它采用了一种简洁的传输技术，称为异步传输模式（ATM）。ATM可以承载任何信息流，包括数据、语音和图像等，ATM将这些信息分成很小的分组，称为信元（cell），然后将信元通过预先建立好的通路进行传输^①。这种预先建立好的通路称为虚连接（virtual connection）。由于ATM能够通过多种不同的宽带技术来支持数据、语音和图像等具有服务质量（QoS）特性的业务的传输，因此ATM被看作是最重要的交换技术，理所当然它也吸引了最多的注意力。

ATM在几个方面上与IP寻路存在不同。ATM是面向连接的，而不是无连接的。在进行任何数据传输之前，必须在两个通信实体之间建立一条端到端的连接。ATM需要通过人工配置手段或采用信令协议（十分类似于电话网中的电话信令）来建立连接。二者之间的另一个不同点是ATM不需要在每一个中间节点（交换机）为单个的分组（信元）确定路由^②。对于ATM来说，两个通信实体之间的网络通路在进行任何数据交换之前就已经计算好，并且在整个连接期间保持不变。在进行连接建立时，每一个中间节点的ATM交换机都为该连接分配一个标识符或标签，这个标识符或标签对于该交换机内的交换器、连接和分配及输入输出端口都是唯一的。这样可以实现两件事情：沿着一条固定的ATM交换通路上的资源将会为某个特定的连接所保留；单个的ATM交换机只需建立一个转发表，而这个转发表是局部的，并且只保存着有关穿越该交换机的活动连接的表项。这一点与路由器是不同的，路由器必须维护一个整个网络范围的路由表，路由表中包含所有可能的目的地址，其中或许还包含一些该路由器永远也不会用到的目的地址。

在ATM交换机中转发信元的过程十分类似于在路由器中转发IP分组。二者都在信头中包含了地址信息，被用来在一个列表中进行索引，该列表记录着可用于发送信元或分组的接口。ATM的速度很快，比传统的基于软件的路由器快得多。这是因为信元头中的标签长度是固定的，ATM在路由查找中所需要索引的连接表通常比路由器中所维护的IP转发表小得多，并且转发过程本身是由特殊的硬件实现。因此，ATM交换与基于软件的IP路由选择相比具有显而易见的性能优势，其中的原因就是硬件执行简单任务速度要比软件执行复杂任务的速度快得多^③。

我们列表1-1的目的只是为了说明IP和ATM之间原理上的几个差异，并非是为了重新挑起一场关于两种技术谁更优越的无谓的争论。从前这种令人厌烦的争论确实发生过，争论的焦点就是IP和ATM哪一个可以作为下一代网络技术的基础。这些争论已经结束了，很清楚，IP将会出现在所有的网络中。当然，IP的一个优势就是它几乎可以运行在任何一种数据链路

① 信元只有53个字节长，没有足够的空间来装载目的地址。因此，在ATM中不可能采用类似于IP和分组那样的基于目的地址进行信元转发。

② IP可以作到这一点，因为每一个IP分组头中都带有目的网络地址。

③ 在某些节点上，半导体技术的进步可以使信元交换和分组路由具有相同的性能价格比。

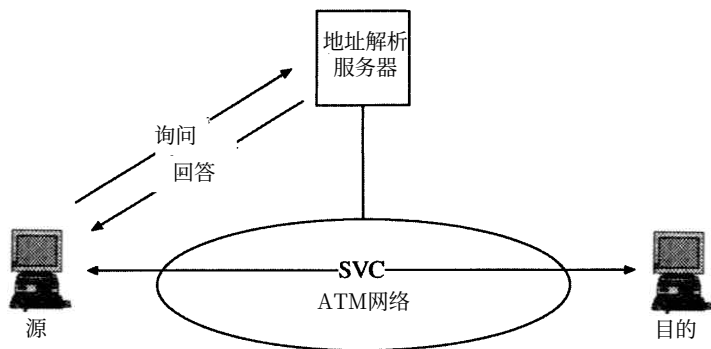
技术的高层，包括 ATM。ATM 应该被看作是一种 IP 能采用并且也应该采用的十分强有力的数据链路技术，而不是用来作为 IP 的替代技术。

表1-1 IP与ATM的比较

属 性	IP	ATM
连接性	无连接	面向连接
分组长度	可变	固定(53字节)
QoS	无	有
信息	数据	数据，语音和图像
通路确定	每分组	连接建立
转发状态	所有可能的网络	局部活动的传输连接
转发基础	最大匹配地址前缀	固定长度的标签
信令	无	有

最初，尝试将 IP 与 ATM 结合起来的研究并不很热烈。IP 所采用的无连接、逐级跳（hop-by-hop）地转发数据报的方式是不可改变的。事实上，IP 在一个具有带宽保留和时延限制能力的网络上运行的概念并不存在。而从 ATM 的观点来看，一系列的分离寻址、路由选择和信令协议都是专门为 ATM 开发的，其运行和操作也是不可改变的。那么什么方法能够将这两种完全不同的技术结合到一起呢？这个问题的唯一答案是地址解析方法。为了能与同一子网中的对等实体成功地实现通信，IP 的源主机需要将目的 IP 地址解析为相应的第二层地址。这使 IP 源主机能够分组寻址到目的 IP 地址，然后采用适当的第二层目的地址将 IP 分组封装到一个第二层的帧中。可是对 ATM 来说，目的 ATM 地址只是在连接建立之前被用于转发信令虚连接（SVC）的建立请求。不过，

IP 源主机仍然需要将目的 IP 地址解析为 ATM 地址，以便 IP 源主机能够建立和初始化一条通向目的地的 ATM SVC。在图 1-3 中举例说明了这个基本的解析方法。连接到 ATM 上的源主机向地址解析服务器发送解析请求，要求获得能正确到达目的主机的 ATM 地址。地址解析服务器将 ATM 地址返回给源主机，然后源主机就可以建立一条到目的主机的 SVC。



IETF（Internet Engineering Task Force）首先在 RFC1577 中建议了这种解决方案，也就是基于 ATM 的经典 IP 和 ARP（Classical IP and ARP over ATM）。在 RFC1577 中，定义了一个 ATM 地址解析协议（ATMARP）服务器来维护一个保存着 IP 地址和 ATM 地址的表。稍后 ATM 论坛将这种方法用于把介质访问控制（MAC）地址解析为 ATM 地址，同时增加了对广播的支持。提出这些方案的目的是为了开发一种系统，能够使局域网（LAN）中的应用可以在不做任何修改的情况下移植到 ATM 网络上。ATM 论坛所做的工作被命名为局域网仿真（LANE），现在已经十分流行，因为它可以使基于多种协议的应用程序能够在原有网络和仿真网络上透

明地运行。经典 IP 和局域网仿真都是将高层的 IP 和 LAN 的功能与 ATM 所支持的底层服务完全分离开，在这一点上二者是类似的。

这两种方法的另一个相似之处在于 ATM 连接的范围。经典 IP 把 ATM 通信连接限制在同一个逻辑 IP 子网（LIS）中的主机（或路由器）之间。即使是不同子网中的两个主机连接在同一个 ATM 网络上，二者之间仍然需要通过路由器来进行分组转发。对于逐级跳（hop-by-hop）路由选择方式，这将引入较大的例行开销和时延。LANE 在这一点上的限制是类似的，连接在同一个 ATM 网络上但属于不同逻辑 IP 网络的主机不能直接利用 ATM 进行通信，这样的主机之间也需要通过路由器交换信息。

这类技术演化的下一个步骤是使不同网络中的两个主机能够通过 ATM 网直接通信。现有的互连网络模型中所有子网间流量都要经过路由器，新的技术需要缓解这个压力。问题的解决方案是不仅要继续支持传统的逐级跳路由选择方式，并且还提供一种机制使 IP 源设备能够建立一条直接的 ATM 连接通往位于另一个网络上的 IP 目的设备。因此，这样的网络不仅可以为尽力而为传输（best-effort）模式和低流量的业务流提供通常的、逐级跳方式的路由通路，同时也可大量的、长持续时间的业务流提供第二条高性能高带宽的交换通路。

这项技术最早是在 IETF 的下一跳解析协议（NHRP）中实现的，后来被合并到了 ATM 论坛的“基于 ATM 的多协议（MPOA）”中。二者都扩展了经典 IP 和 LANE 所首先引入的地址解析服务器的功能，使其解析的地址能够跨越子网的边界。这样一种新的高性能的选路方式诞生了，称为直通选路或快捷连接选路（cut-through or shortcut routing）。它将传统的 IP 选路扩展为一个独立的“泛网络”的地址解析服务，再加上一套管理动态 SVC 所必需的底层 ATM 信令和路由协议。在图 1-4 中说明了直通连接选路的概念。

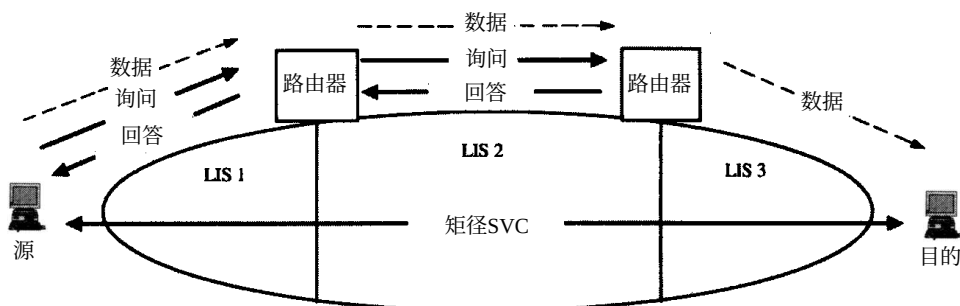


图1-4 直通选路

MPOA 使得分布转发的概念首次在独立的第二代路由器中的实现得到了具体化。中央处理器现在被称为路由服务器，而转发功能被集中到了各个独立的边缘设备上，边缘设备在 ATM 网络的边界上用于连接 ATM 网。在这个分布式的路由器模型中，共享总线实际上就是 ATM 网络本身。MPOA 将其称为虚拟路由器，它能够在子网之间以千兆比特每秒的速率传送分组。这个思想看起来是一个很好的方案，或许会很快流行起来。特别是它十分适合于具有折叠式骨干路由器（collapsed backbone router）的校园网环境，因为在这种环境中，该路由器常常是性能的瓶颈。

1.3 IP 交换

1996 年春天，IP 通信领域发生了一些非常特别的事情。美国加州的一个名为 Ipsilon 的刚

刚成立的小公司提出了一种革命性的思想，称为 IP 交换（IP switching）。这个概念的简单和巧妙性令人惊叹。它将一个 IP 路由处理器捆绑在一个 ATM 交换机上，去除了交换机中所有的 ATM 论坛信令和路由协议，这样的结构我们就可以称之为 IP 交换机了。ATM 交换机由与其相连的 IP 路由处理器控制；IP 交换机作为一个整体运转，执行通常的 IP 路由协议，并进行传统的逐级跳方式的 IP 分组转发。当检测到一个大数据量、长持续时间的业务流时，IP 路由处理器就将会与其邻接的上行节点协商，为该业务流分配一个新的虚通路和虚信道标识（VPI/VCI）来标记属于该业务流的信元，同时更新 ATM 交换机中连接表对应的内容。一旦这个独立的处理过程在路由通路上的每一对 IP 交换机之间都得到执行，那么每一个 IP 交换机就可以很简单地把交换连接表中的上行和下行节点的表项入口正确地连接起来。这样，最初的逐级跳选路方式的业务流最终被转变成了一个 ATM 交换的业务流。上述的基本概念由图 1-5 说明。

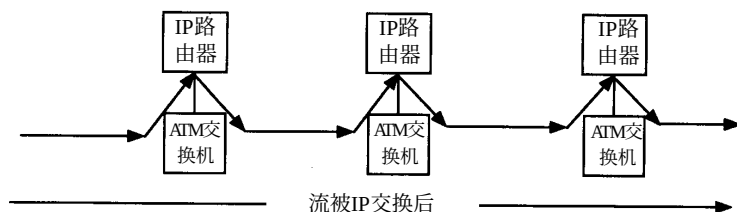


图1-5 Ipsilon的IP交换

抛开复杂的 ATM 论坛的信令、路由和 MPOA 等协议，而让业务流能够动态经过一个 ATM 交换机构进行交换，这个概念一提出就几乎把整个网络产业界闹翻了天，至少是在广告宣传方面。在商业展示会上，人群都聚集在 Ipsilon 的展台前，在各种会议上都有他们对这项技术大量的阐述和介绍。商业出版界开始表示对这种创新性的解决方案的赞同。其他的供应商也赶 Ipsilon 的时髦，纷纷宣布他们也将支持 IP 交换。

为什么不这样呢？毫无疑问，Ipsilon 有一批极其优秀的 IP 和 ATM 技术工程师和开发人员。Ipsilon 的创始人汤姆·里昂（Tom Lyon）曾经开发了 ATM 适配层（AAL5）。彼得·纽曼（Peter Newman）是一个在 ATM 交换技术领域著名的工程师和开发者，也加入了 Ipsilon。在 IPv6 的发展中具有影响力的人物罗伯特·海顿（Robert Hinden）也是其技术队伍中的一员，和其他许多人一样，他也来自像 Cisco 这样的产业界巨头。IP 交换具有如此魅力的另一个原因正如其名称一样，是可以进行 IP 的交换。除了通常正规的 IP 以外，它没有试图去支持任何其他的东西。实际上，IP 交换机就是一个运行通常路由协议的路由器。只是它采用了一个 ATM 交换器来作为分组转发的引擎。人们普遍相信，纯 ATM 信元的交换是进行大量数据转发的一种极其快速和有效的方法——它比当时市场上的任何一种路由器都要快得多。IP 交换也正是利用了这一点，因为在这个产业中，一旦谈及速度，总是会吸引人们注意的。

为了取悦于那些纯粹 IP 主义者，Ipsilon 又不厌其烦地将所有与 ATM 类似的东西从 IP 交换中分离出来。Ipsilon 首先强调，与他们的轻量（lightweight）业务流建立和交换管理协议（包括 Ipsilon 流量管理协议 IFMP 及通用交换管理协议 GSMP）相比，ATM 论坛的协议（包括 LANE 和 Q.2931 信令）是十分复杂的。ATM 中那些强有力的流量管理功能，包括 QoS（服务质量）和 ATM 的流量控制也被简化了。由于这些内容中的某些部分中存在假设和预测的成分，在许多有关下游 ATM 通路等内容的部分中显得有些牵强。在通信产业界中的多数人当然会认

为更大的带宽是必需的，而 ATM 正是一种能够满足这种需求的技术。但是，同样也有很多人十分关心 ATM 论坛所提出的那一系列协议的复杂度以及相关的实现成本，包括网络管理、培训等等。在当时，还没有真正使用了 ATM 中诸如 QoS 等高级特性的应用实例。可以说，当时 ATM 论坛本身并没有为 ATM 技术建立自身形象提供帮助，也没有为像 IP 交换这样的新思想敞开大门。回想一下就知道，当时 ATM 论坛的“基于 ATM 的 IP 选路”和 MPOA 的解决方案仍然处于研究过程之中。事实上，他们正处于为标准中的抽象概念建立名词缩写的阶段，还不能对其工作原理给出一个明确简洁的说明。从规范草案中只能找到一些诸如 Internet 地址子工作组（IASG）、IASG 协调功能组（ICFG）、边缘设备功能组（EDFG）和其他一些功能组的术语。而这些术语也不是为方便阅读而编写的^①。

1.4 路由器 IP 交换

如果我们把焦点聚集在 ATM 交换技术和 Internet 及 IP 开始所受到的关注上，就会很奇怪从前没有人曾经想到过 IP 交换^②。不过这并没有阻碍产业界的巨头开发相似的技术，把 IP 选路与交换合并到一个集成的解决方案中。在 1996 年秋天，Cisco System 发布了标记交换（tag switching）。标记交换提出了用 IP 交换来包含分组的概念。这种构思的产生动机是让路由器能采用一种标签交换的转发机制，而不是采用传统的最长地址匹配的查寻方法，以便提高以路由器为核心的网络的性能。Cisco 公司认为，在标记的基础上对分组或信元进行标签的交换比在路由表中提取每个分组的目的地址进行最长匹配查寻的方式更加高效。同时，将网络层的选路过程从实际的转发机制中分离出来，可使任何不同的 3 层服务都可以与一个简单而高速的转发机制即标签交换联系起来。它的工作原理是，首先根据路由器表项的内容向网络中的标记交换路由器（tag switch router）发布一组标记，然后根据每个分组所携带的标记进行标签交换，来完成业务流在网络中转发。标记交换的概念在图 1-6 中说明。

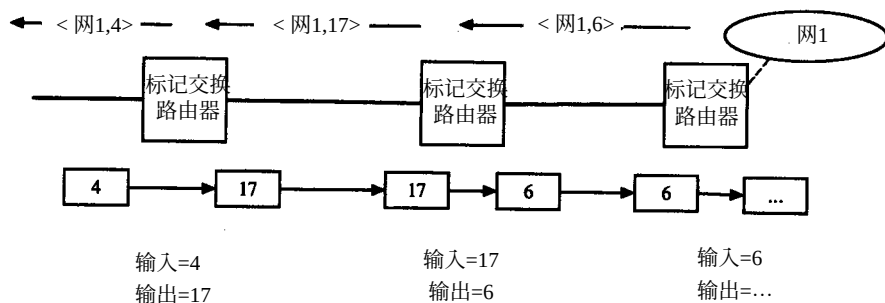


图1-6 Cisco的标记交换

标记交换与 IP 交换的不同点在于，标记交换在控制流（control traffic）出现的基础上，特

① 为了自身的声誉，MPOA 工作组在 1996 年末大大地简化了他们的规范。MPOA 的最终规范在 1997 年 7 月发表，这个版本十分易于理解，它清楚地说明了 MPOA 客户和 MPOA 服务器这两个组成部分的功能和行为。

② 实际上，日本的研究人员在若干年前就已经提出过信元交换路由器（CSR）的概念，它可以实现无连接的选路和面向连接的直通过路由交换。它被命名为基于 ATM 的常规 IP（Conventional IP over ATM），但是并没有引起更多的关注，或许是因为当时研究工作的主流集中在了经典 IP 和 LANE 上。有关 CSR 的概念将在第 10 章中详细阐述。

别在选路协议更新（routing protocol update）的基础上，生成和分配标记^①。标记交换为出现在标记交换路由器路由表中的每一个目的地都建立一条交换通路，去往某个特定目的网络的所有数据都被安排在相同的交换通路上。这种方式与IP交换是不同的，IP交换是为实际达到的数据流本身建立一条交换通路，该通路只能被该数据流使用。标记交换可以被归类为是拓扑驱动或者控制驱动的（topology- or control-driven），而IP交换是流驱动或者数据驱动（flow- or data-driven）的。

几乎就在Cisco宣布其标记交换的同时，IBM提出了一个名字听起来不同凡响的解决方案，即ARIS。ARIS的含义是“聚合的基于路由的IP交换”（Aggregate Route-Based IP Switching）。尽管这个概念与标记交换比较类似，但二者之间还是有一些明显的差别。总地来说，其中的一些内容明显地增强了拓扑驱动的IP交换功能，而另一些内容只能被视为是完成同一个目标的不同方法。ARIS首先引入了多点到点交换通路的概念。在一个采用选路方式的网络环境中，从多个输入（ingress）源到一个公共的输出（egress）路由器的转发通路构成了一个逐级跳（hop-by-hop）的树结构。ARIS将第3层的转发树映射到一个第2层的多点到点的树，这个树的根是输出路由器，由根分支出去，到达所有的输入路由器。所有目的地址位于在输出路由器之后的业务流都将沿着这个2层的交换通路的分支流向输出路由器。这个概念在图1-7中说明。

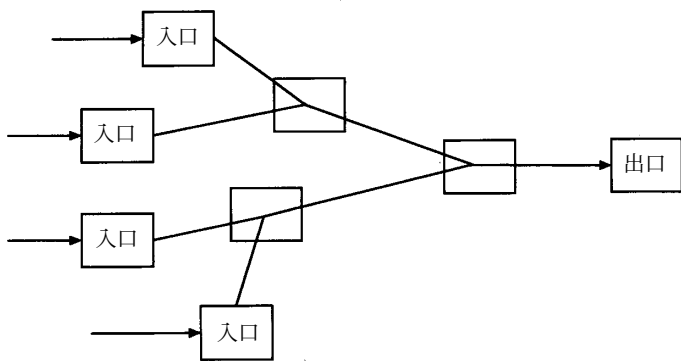


图1-7 ARIS多点到点的交换通路

ARIS允许每个分支节点执行虚连接（VC）的合并操作，因此ARIS能够建立多点到点的交换通路。VC合并的过程是一个将属于多个上游VC的信元合并到一个单一的下游VC上去的过程。在每一个合并点上还需要采用每个VC（per-VC）的缓冲和排队等技术，以便保证帧的完整性并防止新生成的下游VC上发生信元交错。一个能够支持多点到点交换通路的网络在保证所有数据都能实现交换的同时，只消耗很少的VC资源。这也给它提供较好的可扩展性；VC合并技术最终也被融合到了标记交换以及后来的标准中^②。

标记交换和ARIS之间的其他一些较小的差异还有环路预防、局部输出标签分配（local versus egress label distribution）、输出聚合的广度，以及在传输信令消息时，对TCP或用户数据报协议（UDP）的使用方式等。但是从根本上讲，这两种解决方案都在现有的IP网络拓扑基础上建立交换通路，并且可将所有的流量交换到特定的目的地上去。

1.5 一个IP交换的标准

一些利用2层交换来提高IP选路性能的其他技术和解决方案也被提出来。这些方案的出现一般都伴随着一个由其供应商的市场部门所创造出来的名称，例如3Com FastIP、Cascade IP

① 其他形式的控制流如资源预留流建立协议（RSVP）可以分配标记。这种技术在第11章中描述。

② ATM论坛也增强了它的用户网络接口（UNI）及PNNI的信令和选路以便支持多点到点的信令。

Navigator、Cabletron SecureFast等。其他一些方案的名称缩写也都很精巧而富有异国情调，或许能更好地被人们记住，如SITA、PLASMA、IPSOFACTO等，它们大都是研究开发的成果 and 思想。所有这些技术与支持IP的2层交换以及前面所涉及的技术的融合起来，就产生了一类新的解决方案，被统称为IP交换。不幸的是Ipsilon并没有将这个名字注册为其商标，因此这个名字不仅可以用来说明某个供应商的解决方案，而且可以用来代表所有在单一网络基础结构上实现IP选路与交换相结合的一整套技术。

1996年底，IETF成立了一个工作组，对集成路由与交换的解决方案进行标准化，这个标准称为多协议标签交换（MPLS）。这个工作组中的主要专家大都是来自各个设备厂商的研究机构中的工程师和开发人员，他们都曾经参与或正在参与IP交换技术的研究。但是Ipsolin并没有出现在这个工作组中。他们将许多最初在标记交换和ARIS建议中所描述的技术与其他的思想结合起来，以便形成一个最初的MPLS框架、体系结构和标签分配协议（LDP）规范^①。MPLS被看作是一种能够帮助Internet服务提供者（ISP）在帧中继或ATM网络基础上以更易于控制和扩展的方式提供IP路由服务的解决方案。不论有没有交换式的网络基础设施，在网络中利用标签来转发分组，都为引导业务流穿过特定的链路和节点创造了许多令人感兴趣的的可能性。与性能和可扩展性相比，它或许更接近于流量工程。

目前，存在许多不同的技术和协议，能使基于硬件的交换成为可能，这无疑是提高IP选路性能的一种手段。其中最具有影响力的硬件交换技术曾经是（也将继续是）ATM，因此各种目前已经提出的IP交换解决方案都采用了ATM交换。像经典IP和LANE这样的最初的研究成果解决了IP在ATM上连接性的问题。MPOA和MPLS等最新的努力成果试图采用ATM提供更好的性能并减少网络的传输时延。另外，由于同时提供了两条通路（一条默认的路由通道和一条分离的2层交换通路），它也为校园网或ISP的流量管理工程师们提供了一整套很有帮助的可选方案。现在，我们已经完全能够对网络进行合理配置，使其能对尽力而为（Best-effort）方式的数据流采取路由的方式，同时对高优先级的数据流采用交换方式。

使用IP交换系统需要付出什么代价呢？这个代价有赖于投资的风险。它有可能像我们给一个已安装的ATM交换机网络提供新软件那样简单；也可能我们所需要的是一个路由服务器和一套新的边缘设备；或许甚至需要更新整个骨干网络基础设施，这种情况很可能发生在校园网或ISP的核心中。在实践中，一般不会有有什么新东西需要添加到桌面，因为桌面系统一般都采用以太网，将来也不会有太大变化。IP交换实际上是一种骨干网中的解决方案，而不是一种端到端的解决方案。我们需要根据现有骨干网络的性能价格（或好或不好）以及可预见的未来的流量负荷及应用的需求来进行权衡。

1.6 结论

我们是否真的需要IP交换呢？从目前情况来看，我们完全有可能使用一种新设备以线路的最高速率在子网之间传送很大数量的分组，这种新的硬件设备家族称为3层交换机。3层交换机与其ISP级别的兄弟产品千兆位（Gigabit）路由器一样，是一种十分高速的路由器，与IP交换机相同，它们都在其转发通路上引入了特殊的硬件。在这种情况下，它已不是一个ATM交换机，而是一个专门设计用来进行3层选路的专用集成电路（ASIC）。它的选路过程与通常的选路是一样的，不需要任何ATM论坛或其他专门的控制协议来建立基于ATM VC上的转发

① MPLS将在第15章中详细介绍。

通路。基于硬件的选路可以将 ATM和特定的 IP交换解决方案引入到 Internet和企业 Intranet领域中的合适的位置上。

但是人们应该牢记几个过去历史和未来发展趋势的要点。首先，只考虑带宽不是一个解决方案。带宽是任何网络中最重要且最有价值的资源。但是历史的经验证明，用户带宽的需求总是要超过网络实际的能力。这种趋势在可预见的将来是不会改变的，可能永远也不会改变。如果的确是这样，网络供应商就会在他们网络中有选择性地确定允许哪些数据流通过他们的网络，如何处理这些数据流等。我们也能想象，网络的提供者将可能更加积极地为通过其网络的数据流提供更好的服务^①。假设一个网络基础设施有足够的带宽和有效的队列管理，网络提供者将会需要一种在选路时能够将允许的流量负荷、现有的网络性能和拓扑条件等因素分解出来的动态的路由协议，或者需要一种机制来为特定的数据流配置一条非缺省的路由。现有的动态 IP路由协议还不能在数据流或性能基础上选择路由。而配置非缺省的路由也需要每个路由器过滤器（per-router filters）和静态选路。然而，校园网中的 IP交换解决方案是能在骨干通路上采用 ATM的，这样专用网络—网络接口（PNNI）就可以被用来作为一种在带宽、QoS参数和可到达性基础上计算和选择路由的方法。而在 ISP的环境中，MPLS提供了一种简单的方法来选定的数据流配置和产生非缺省的路由。

这里提出的另一个要点是针对未来趋势的。下一代的 IP网络必须能够支持一些非缺省的服务才是具有实用价值的，这些非缺省服务可能是安全性、服务等级（COS）、QoS、组播（Multicast）、服务级别保证（SLA）等等。IP交换并没有包括更多额外的工作，但是它将无连接的 IP数据流放到了实质上是面向连接的一条 2层交换通路上。这种方式带来了一般面向连接的网络所具有的好处，例如高带宽、流量管理和整形、流分类、QoS、安全性、非缺省选路等。在这里要记住的一点是，新的协议和开发将使网络提供者能够提供新的 IP服务。网络提供者将不断提供这种新服务来吸引新的用户和消费者；或者在 Intranet的环境中，这些新服务能使终端用户群的生产效率更高。IP交换就是这样一项技术，它能够更高效和经济地帮助网络提供者支持多种新的服务。

本书的其他部分将包括一些有关 IP基础、LAN/WAN（广域网）交换、以及目前已开发或已提出的一组不同的 IP交换方案的细节。我们会惊奇地发现，那些与 IP交换有关的基本问题和解决方案，诸如快速转发查寻、提高带宽、降低时延、QoS、非缺省选路等等，仍然在产业界内继续被研究和开发。我们可能永远也不会满足，但是这是一个十分正确的态度，因为这可以使我们不断地去发现新的技术和途径来改善我们已经获得的一切。

^① 这需要假定网络提供者能保证向用户提供达到一个可预知的水平的网络性能，而用户或消费者需要为此付出一个额外的费用。

China-pub.com

下载

第2章 TCP/IP、寻址和选路

由于本书主要讨论的内容是 IP 和交换，因此回顾以下传输控制及网络互连协议系列 (TCP/IP) 应该是很有帮助的。TCP/IP 实际上是一组由不同协议构成的协议族。该协议族可使计算机及终端用户能够通过一个由不同介质构成的任意大小网络进行相互通信。本章将对 TCP/IP 体系结构、寻址及 IP 选路的操作进行概括地介绍。

2.1 TCP/IP 的历史

TCP/IP 的历史，包括更一般的互连网络的历史，可以追溯到 60 年代。Licklider、Clark^① 与 Kleinrock^② 分别在其研究报告中提出了一个“银河网” (Galactic Network) 的概念以及分组交换的理论。但是，实际上是冷战刺激了军事和科研机构去研究和开发一个能够抵挡核武器进攻的通信基础结构。在分组交换网络的分组数据 (称为数据报 datagram) 中采用目的寻址方式的概念是由 Paul Baran 于 1964 年在其发表论文中提出的^③。如果某些通信链路或节点在遭受核攻击时受到破坏，分组就可以重新选择一条可用的替代路由，最终被传送到目的地。

60 年代后期，美国国防部高级研究计划署 (ARPA) 在这个领域内投资进行了研究和开发。1969 年，被称为“接口消息处理器 (IMP)” 的第一个 ARPANET 的节点安装并投入了使用。到 1969 年底，ARPANET 的节点已达到了 4 个。

70 年代，最重要的网络技术进展除了 1973 年发明的以太网以外，就是互连网络协议的提出。ARPANET 上最初的主机到主机之间的协议实际上只是网络控制程序 (NCP)。这种 NCP 的方式后来被对一系列网络互连协议的进一步研究所取代，最终这一系列协议被定名为 TCP/IP。其中的第一部分称为 TCP，是由 Kahn 和 Cerf 在 1974 年开发的^④，该部分将 TCP 协议中端到端传输功能与分组转发、IP 的寻址功能结合了起来。后来这个模型又被重新进行组织和调整，我们今天所知道的 TCP 和 IP 的基本功能被分割成一组分离的协议。

一个有趣的值得注意的现象是多年以前所提出的 TCP/IP 和网络互连的基本原则中的大部分内容至今都没有改变。那些指导 TCP/IP 开发，并使其能够达到今天这样繁荣水平的基本原则包括：

- 每一个网络都是唯一的和可自维持的 (self-sustaining)。无论其是否与其他网络连接，都是可以自主运行的。由网关 (或路由器) 设备互相连接起来的一个或多个网络的集合被称为互连网络。
- 通信采用尽力而为的 (best-effort) 方式。如果源端没有收到目的端返回的应答，数据分组将会被重发。
- 路由器 (最初被称为网关) 负责连接各个网络。但路由器不维护任何与特定数据流有关

① J.C.R. Licklider and W. Clark, “On-Line Man-Computer Communications”, Aug. 1962.

② L. Kleinrock, “Information Flow in Large Communications Nets”, RLE Quarterly Progress Report, July 1961.

③ P. Baran, “On Distributed Communications Networks”, RM-3420-PR, Rand Corp., Aug. 1964.

④ R.E. Kahn and V.G. Cerf, “A Protocol for Packet Network Interconnection”, IEEE Transactions on Communications, Tech 5, May 1974.

的状态。

- 不存在任何集中控制。一个网络的运行不依赖于任何一个集中的管理实体。

由于TCP/IP的这些基本原则以及它能够在几乎任何一种介质上适应不同应用需要，TCP/IP在80年代和90年代快速发展，一直到今天这样的规模和水平。下面列出一些（但或许不是全部）TCP/IP发展的重要的时间和里程碑：

- 1983年，UNIX 4.2BSD发布，其中包含了TCP/IP，主机系统从NCP升级到TCP/IP。
- 1984年，域名服务系统（DNS）发明。
- 1986年，NSFNET形成，它与ARPANET并行成为了Internet的骨干。
- 1988年，在由于拥塞造成的一系列Internet失效事件之后，Van Jacobson出版了他的有关控制和避免TCP/IP拥塞的著名论文^①，该论文阐述了TCP慢速启动和拥塞控制的基本内容。
- 1989年，有关“开放最短路径优先（OSPF）的第一个请求注释（RFC）”的路由协议发表。
- 1991年，ATM论坛成立。第一批成员包括北方电信、Sun、数字设备公司（DEC）和Sprint。
- 1992年，由Tim Berners-Lee发明的万维网（WWW，World Wide Web）开始出现。下一代IP（IPng）工作组（WG）成立并开始研究下一代IP中的核心技术。
- 1993年到1994年，局域网中的以太网交换和广域网中的帧中继开始起步。日本的研究人员第一次提出了将IP选路与ATM交换相融合的概念。
- 1995年，作为下一代网络的传输技术，ATM流行程度达到了顶峰。
- 1996年，Ipsilon宣布了其新产品和支持的协议，网络技术的字典中出现了“IP交换”。稍后，IBM和Cisco也与之相呼应，推出了更易于扩展和升级的解决方案，在由路由器-交换机集成设备所构成的网络中交换IP分组。其他厂商后来也纷纷提出类似的方案。
- 1997年，IETF中MPLS工作组成立，它为设备厂商及其追随者们提供了一个进行技术发展的正式场所，使其能自由争论各自的方案并从事一项具有挑战性的工作——把网络层选路与2层交换结合起来的技术研究。一种被称为“3层交换机”的新的设备被引入，它能够采用硬件进行快速的选路。

对于系统网络体系结构（SNA）、Novell互连网协议交换（IPX）、Appletalk、Decnet和他们的信奉者来说，TCP/IP无疑是占据了支配地位的网络互连协议。关于TCP/IP后来又发生了很多事情，新技术不断出现，使其运行得更快、更可靠、更易于升级和扩展，而且这种发展趋势使TCP/IP将继续向更好的方向发展。

2.2 TCP/IP体系结构和组成部分

与SNA和开放系统互连（OSI）等其他的网络互连协议一样，我们可以很容易地用一个分层模型来说明TCP/IP的体系结构和功能。但是，TCP/IP协议的体系结构并没有采用七层模型，只实际定义了四个层次，如图2-1所示。其说明如下：

- 数据链路。数据链路层定义了与某特定介质的物理连接特

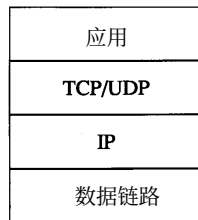


图2-1 TCP/IP体系结构

^① V. Jacobson, “Congestion Avoidance and Control”, ACM Sigcomm Proceedings, 1988.

性，以及用于在该介质上发送和接收的信息帧的格式。TCP/IP支持的数据链路技术很多，包括以太网（各种速率的以太网）、ATM、令牌环、光纤分布数据接口（FDDI）、帧中继等等。TCP/IP的完美之处就在于它可以在几乎任何一种物理网络上运行。

- IP。IP层负责将数据分组从源转发到目的地。在每一个分组中，都包含一个目的 IP 地址的字段，IP层利用这个字段信息来把分组转发到其目的地。

IP不仅可以运行在各种主机上（包括源主机和目的主机），也可以运行在分组交换和转发设备上，这些设备被称为路由器，我们将在下面进行说明。IP层是无连接的，这意味着任何数据开始传送之前，不需要首先建立一条穿过网络到目的地的通路或路由。从理论上说，每个分组的确都可以采用不同的路由转发至同一个目的地。此外，IP不能保证分组按正确的顺序到达目的地，甚至不能保证分组能够到达目的地。IP的设计者希望通过这种方式能让网络不必保存任何每个连接或每个数据流的附加状态。我们也应该注意到，尽管在IP层以上有多种传输和应用层功能，而在IP层以下也有多种数据链路技术，但只有其中的IP组件是TCP/IP和Internet的汇聚点。

目前的IP有两个版本：IPv4和IPv6。IPv4是我们今天正在使用的版本。IPv6被认为是下一代的IP（IPng），它是一个新的版本，IPv6内部进行了许多简化，许多功能得到了增强，其中最重要的改进就是它大大增加了地址空间。有关IPv6的内容将在下一章中讨论。

- TCP/UDP。传输层运行于IP层之上，它由两个协议组成。在源和目的之间，TCP提供一种面向连接的、可靠的传输服务；而UDP则提供一种无连接的、不可靠的传输服务。TCP和UDP都运行于主机上，能够分别为不同的应用程序提供特定的服务。

- 应用程序。各种应用程序都使用了底层TCP/IP的服务。例如，用于终端仿真的Telnet、用于文件传输的文件传输协议（FTP）、用于主页浏览的超文本传输协议（HTTP）及用于电子邮件的简单邮件传输协议（SMTP）等，都是TCP/IP许多著名的应用实例中的一部分。

除了协议栈以外，TCP/IP的另外两个基本组成部分是IP分组和路由器。分组（packet）的定义十分简单，如图2-2表示。IP分组由一个数据净荷字段及跟在其后面的传输头、IP头和数据链路头组成。当分组在某个特定的数据链路介质（如以太网）上传送时，数据链路头用于将一个分组从源引导到目的地。如果在分组从源到目的地的传输通路上存在不同的数据链路技术，那么数据链路头的内容将会发生改变。比如在某些实例中，一条源到目的地的通路由一个以太网LAN、一个帧中继永久虚连接（PVC）和一个令牌环LAN组成。传输头可能由一个TCP头和一个UDP头组成；但如果应用程序没有用到TCP或UDP所提供的传输层服务而是自己提供传输功能，那么TCP头或UDP头就可能被省略。IP头当然是必需的，因为其中包含着IP分组的最终目的地址。

净荷	传输头	IP头	数据链路头
----	-----	-----	-------

图2-2 简单IP分组

路由器是一种设备，它能将IP分组从一个网络转发到另一个网络。路由器上的每一个网络接口都分别连接到一个互不相同的网络上。一个路由器与其他路由器之间不断地交换信息，这些信息是关于整个互连网中各个网络的位置和可达性等方面的信息。路由器之间信息的交换必须通过采用适当的路由协议来进行。此外，路由器还采用路由协议来计算IP分组去往目的地的“最佳路径”（即实际上的最短或费用最低的通路）。

图2-3中给出了由路由器互相连接起来的3个网络所构成的一个互连网。在同一个网络内

的主机（如网络A中的主机）当然不需要路由器来进行通信。但是，如果源主机和目的主机位于不同的网络中，那么源主机必须通过路由器来发送IP分组去往目的主机。路由器将会查看各分组所携带的目的地址，并将这个目的地址与保存在一个网络目的表中的、路由器已知网络目的地址相比较，然后将分组转发到目的地。路由器是Internet以及其他更一般的、基于TCP/IP的网络的、最基本的构造单元。

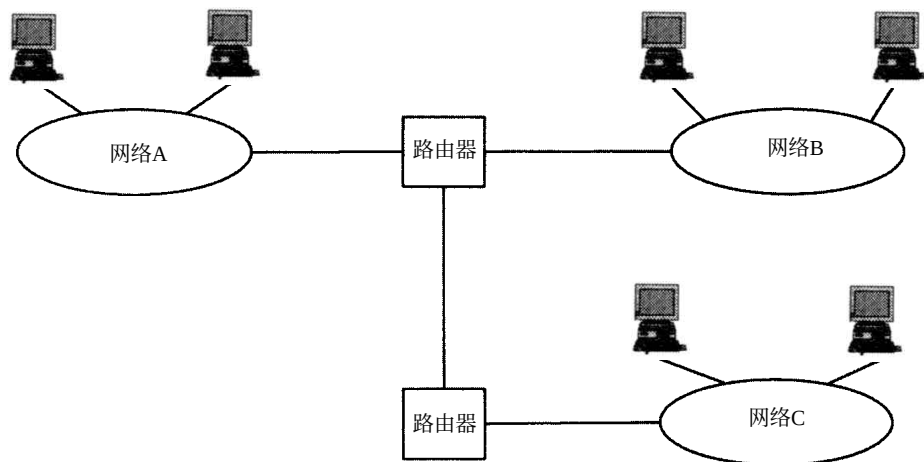


图2-3 简单的网络互连

2.3 IPv4头格式

在网络中，为将分组准确地引导到其最终目的地所必需的信息都保存在分组头中。IP层从它的上层接收到传输层或应用层的信息以后，就将IP分组头添加到接收的信息之后，然后将其向下提交给下层的数据链路层，通过某特定的介质发送。IPv4分组头的长度范围可以从20字节到最多60字节（假如使用了附加的选项）。IPv4分组头的格式如图2-4所示。其中每一个部分的含义如下：

- 版本（Version）。当前的版本是4。

- 分组头长度（HL）。以32位字为单位的IP分组头的长度。这样，一个典型的20字节的分组头应该包含5个32位字。

- 服务类型（TOS）。8比特的服务类型字段分为两个部分：优先级（precedence）和服务类型（TOS）。优先级由3个比特组成，定义优先级的最初目的是提供一种能够为特定分组流分配优先级的手段。其余的5个比特为穿越网络的IP分组定义了服务类型。实际上，TOS字段从未按其最初设想而被真正地使用过。但目前新的研究工作正在努力试图重新定义TOS字段，

版本	分组头长度	服务类型	总长度	
标识			标志	分段偏移
生存时间	协议		分组头检验和	
源IP地址				
目的IP地址				
选项				

图2-4 IPv4头格式

以使其能够为IP分组分配服务类型，这就是所谓的差分服务（Differentiated Services），有关的内容将在下一章讨论。

- 总长度（Total length）。是一个16比特长的数据报长度字段，用来定义以字节为单位的整个IP分组的总长。IP分组的最大长度为65536字节。

- 标识（Identification）。这个字段包含一个16比特的标识符，目的主机用它来识别和组合IP分组的片段。如果IP分组的入线数据链路的最大传输单元（MTU）长度大于其出线数据链路的MTU长度，IP路由器就会把该IP分组分段。MTU长度的定义是数据链路帧所能承载的最大的IP分组的长度。IP分组的组装只在目的主机上进行。

数据分段并不是一个好办法，因为它在路由器和目的主机上引入了额外的开销。一种被称为通路MTU发现[⊖]（Path MTU Discovery）的技术可以使发送分组的源主机发现从源到目的通路上的最长MTU长度，因而可以避免任何不必要的分段操作。

- 标志（Flag）。标志字段由3个比特构成，其含义如下：

- ☐ Bit 0：保留。

- ☐ Bit 1：0 = 可以对分组进行分段；1 = 不可分段。

- ☐ Bit 2：0 = 最后一个段；1 = 还有更多的段。

- 偏移（Offset）。偏移字段由13个比特构成。原始IP分组被分段后，每一个分段生成的数据块（Fragment block）都有一个从IP分组起始位置开始编号的序号，这个序号记录在偏移字段中。

- 生存时间（TTL，Time to live）。TTL字段用来防止在网络中出现IP分组的无限循环。一个IP分组每经过一个路由器，TTL值都会减小。当TTL的值被减到0时，路由器就会丢弃该分组。最初TTL值是一定数量的秒数；但目前它的含义是在路由器之间允许的最大跳（hop）数。

TTL值也被用来控制组播（multicast）分组在网络中传播的范围。组播分组的TTL值必须大于通往外部路由器接口中所配置的TTL阈值，否则该分组将被丢弃。

- 协议（Protocol）。协议字段用来鉴别利用了IP服务的邻接高层所采用的协议。例如，6 = TCP、17 = UDP、1 = ICMP（互连网控制消息协议）、89 = OSPF等。

- 分组头校验和（Header checksum）。这个16比特的字段是根据IPv4分组头中的其他字段计算出来的。分组每经过一个路由器，校验和的值都会被更新，因为TTL字段和IP选项字段中的值可能发生改变。

- 源IP地址（Source IP address）。发出IP分组的主机的32位IP地址。

- 目的IP地址（Destination IP address）。32位目的主机IP地址。

IP分组头中还可以有一系列的选项字段。这些选项由通路上的每一个路由器来处理。选项字段可以被添加到一个IP分组中，内容包括源选路和路由器警告等。实际中选项字段很少被使用，因为它所带来的任何功能的增强与它所引起的每个中间路由器的额外处理开销相比都是不值得的。

2.4 IPv4寻址

IPv4的地址长度是固定的，被限制在32位，通常用类似于uu.xx.yy.zz的一串用圆点隔开的

[⊖] RFC1191，通路MTU发现。

数字表示（例如121.17.13.1）。每一个主机或路由器的接口都被分配了一个IP地址，用来定义设备所连接的网络上的主机号和网络号。地址中代表网络的部分称为网络前缀（network prefix），代表主机的部分称为主机号（host number）。

分类的寻址（classful addressing）意味着网络前缀与主机号之间的边界落在地址字段中的一个字节边界上。这个边界落在32位地址中的什么位置取决于IP地址的类型和子网掩码（subnet mask）的值。子网掩码是一个按位的掩码，它用来确定在地址中有多少个比特位用于构成网络前缀。IP的单播（unicast）地址有3类，表2-1分别总结了这3类地址。子网掩码可以表示为数字格式（如255.0.0.0）或者是符号“/”后加掩码的比特数的形式。例如，一个掩码为255.255.0.0的B类地址189.23.0.0可以被表示为189.23.0.0/16。

表2-1 分类IP单播地址

类型	字节1	字节2	字节3	字节4	网络范围	网络总数	子网掩码
A类	网络	-----主机-----			1-126	126	255.0.0.0 或 /8
B类	----网络	----	----主机	----	128-191	16,384	255.255.0.0 or /16
C类	-----网络	-----	主机		192-223	2,097,152	255.255.255.0 or /24

2.4.1 子网划分

子网划分这项技术用来把一个单一的IP网络地址划分成多个更小的子网（subnet）。这种技术可使一个较大的分类IP地址能够被进一步划分。这样就可以使一个大的分类地址（classful address）能被一个企业中地理位置不同的许多站点所共享，而不需要为每个站点都分别申请一个分类网络地址。子网划分的过程是将分类IP类地址中的主机号部分进一步划分成一个子网部分和一个主机部分。与传统的分类地址一样，地址中的网络部分（网络前缀+子网）与主机部分之间的边界是由子网掩码来定义的。

图2-5给出了一个例子。B类地址187.15.0.0被分配给了ABC公司。ABC公司的网络规划者希望建立一个企业级的IP网络（Intranet），用于将数量超过200个的站点互相连接起来。由于在IP地址空间中“187.15”部分是固定的，网络的规划者就只有剩下的两个字节（或16位）用来定义子网和子网中的主机。他们决定利用第三个字节作为子网号，第四个字节作为给定子网上的主机号。这个互连的网络的子网掩码的值就是255.255.255.0。这意味着ABC公司的Intranet能够支持最多254个子网，每一个子网可以支持最多254个主机。至于从外部的Internet的角度来看，ABC公司的网络仍然看起来像一个地址为187.15.0.0的单一网络。

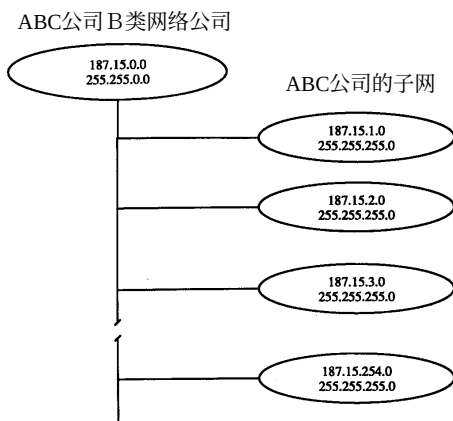


图2-5 ABC公司的网络

这个例子说明了为整个网络定义统一子网掩码（255.255.255.0）的情况。它意味着每个

子网中最大的主机数只能是 254。此外还存在一些情况，即在一个子网中支持的主机的数量需要比现在这种采用固定长度子网掩码的方式所能支持的主机数量更多（或更少）。在这种情况下，我们还可以使用可变长子网掩码（VLSM）。采用 VLSM 能够把一个分类地址网络划分成若干大小不同的子网。假设在上述 ABC 公司的例子中，我们需要支持一个 500 个主机组成的子网。网络规划者就不需要分配两个子网（每个支持 254 个主机），而分配一个子网掩码为 255.255.254.0 的子网就可以支持最多 512 个主机地址了。而另一个场合可能只需要 100 个地址，在这种情况下一个取值为 255.255.255.128（128 个主机）的子网掩码就足够了。另外还有一个特殊的例子，就是连接两个路由器的点到点的子网。由于在这种子网上只可能有两个“主机”，因此我们应该使用一个取值为 255.255.255.252 的掩码，就可以不浪费任何地址资源，以便其他子网能够使用。

VLSM 能更高效地管理地址空间，让分配给每个子网的主机地址的数量都符合实际需要。在动态选路的网络中，为了使 VLSM 工作得更有效，就必须配置能支持 VLSM 的路由协议。支持 VLSM 的路由协议在其发布的通告消息中把子网掩码包含在了其网络地址前缀中。目前的路由协议如路由信息协议-2（RIP-2）和 OSPF 等都支持 VLSM。

2.4.2 超网

子网划分将一个单一的 IP 地址划分成多个子网。超网（supernet）与子网相反，它的功能是将多个 C 类网络聚合起来，构成一个单一的、具有共同地址前缀的网络。采用这种地址汇聚操作的思想有两个出发点：

1) 通过减少路由表中 C 类网络的表项来减少路由器中所维护的路由表的大小。

2) 只为网络分配其实际所需的地址数，因而可以更高效地利用那些未使用的地址空间。例如，一个 B 类地址包含大约 65,000 个主机地址。而一个网络可能只用了其中的很小的一部分，这就造成了大量未使用的地址资源也不能被其他的网络所利用。同样，由于 C 类网络只有 254 个可用的地址，构成网络太小，有必要进行聚合。

图 2-6 中的例子可以最好地说明这个概念。16 个 C 类地址组成了一个地址空间块。在路由器中，并不是把所有的 16 个 C 类网络地址分别分配不同的表项，而是使用了一个单一的超网 192.18.0.0/20。从技术上讲，超网是一种子网掩码长度小于其特定分类子网掩码长度的一种网络。由符号“网络前缀 / ”所代表的连续 C 类地址块被称为无分类域间路由（CIDR）块。作为降低 IP 地址分配速度以及减少 Internet 路由器中表项数的一种方法，CIDR 技术在过去的几年内已经被广泛认同。现在希望获得一个 Internet 地址的组织都将被分配给一个 CIDR 块，而不是我们前面所描述的那种传统的分类地址。

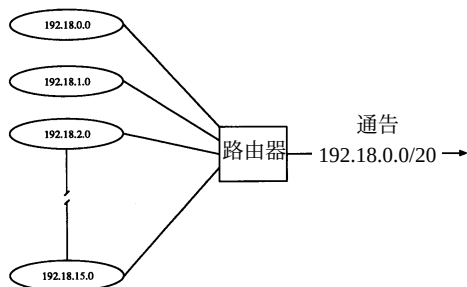


图2-6 CIDR 汇聚

2.5 IP 选路和转发

选路（routing）是一个通用的术语，用来描述某一个网络中的主机所发出的分组经过一

个或多个路由器传输到位于另一个网络中的主机的过程。更明确地说，选路由两个独立并且截然不同的操作构成：

- 路由数据库管理。所有路由器都维护着一个目的网络的数据库，数据库中还可能有一些拓扑数据，路由器以目的地址（目的地址包含在分组头中）为基础并根据这些数据计算分组应该转发到哪里去。对这些数据库的维护工作是通过在各路由器之间交换路由数据库更新消息的过程进行的。这些消息的格式和其中包含的信息是由各路由器所采用的某一个（或多个）动态路由协议（如 RIP、OSPF）来定义的。路由器用于维护数据库完整性、准确性以及建立转发表^①（以后用该术语来表示路由表）所使用的处理过程、算法和协议等也都被称为 IP 选路。

- 分组转发。把分组从一个网络转发到另一个网络的实际过程被称为分组转发或 IP 转发。当分组到达一个路由器时，目的 IP 地址信息从分组头中被提取出来，并与路由表中的表项进行比较。路由表是从路由数据库中计算得来的，由一个目的网络的列表和与之相应的通路中的下一跳（next-hop）地址构成；路由器认为这条通路可以通往目的地。分组头的校验和被重新计算，TTL 的值被递减，然后分组被转发到下一跳地址，最后一直被传递到其最终的目的地。这种方式被称为逐级跳（hop-by-hop）的选路，因为 IP 分组被转发到哪里是由每一个路由器根据其自身路由表 and 该分组目的地址来决定的。

图2-7从一个独立的路由器角度描述了 IP 选路过程的基本组成部分。路由数据库的更新消息到达路由器后，被传递给路由协议实体进行处理。一旦这些消息得到处理，一个新的 IP 路由表就生成了。当一个 IP 分组到达路由器后，其目的地址就会与路由表中的表项进行比较，然后被发送到其所需的通路上。此外，路由协议实体也能产生新的路由数据库更新消息，并传递给它相邻的其他路由器进行处理。

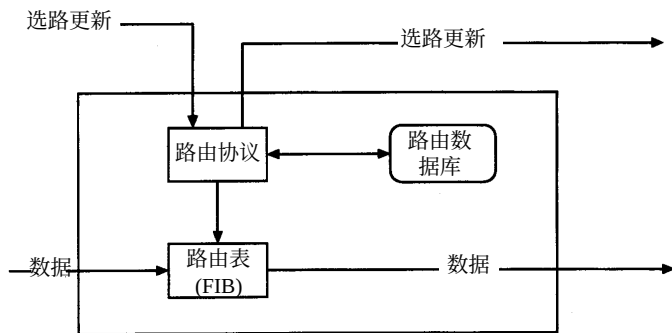


图2-7 IP选路的组件

尽管在大量的论文、标准和书籍中已经对 IP 选路处理和路由协议的问题进行了很多的讨论，但是这种在网络中转发分组的简单过程正重新开始受到更多的关注。由于多种不同的原因，IP 分组最终到达其目的地所需的时间很长。其原因包括：

- 通路上有太多的路由器。这实际上是一个网络设计问题；这个问题可以通过将路由器替换成交换机来解决，而目前许多网络供应商正在进行这种替换。

- 路由表的规模过大。路由表越大，在路由表中查找与 IP 分组的目的地址相匹配的表项所花费的时间就越长。目前，路由表的大小在一定程度上可以通过使用 CIDR 和其他地址汇聚技术得到控制，但是随着时间的推移，路由表仍然会不断增大。如果我们引入 IPv6，情况会进一步恶化，因为在 IPv6 中网络地址长度至少将会达到 64 位。

- 陈旧的路由器技术。我们经常会使用一些老式的设备。在老式路由器中，进入输入接口

① 转发表也被称为转发信息数据库（FIB）。

的IP分组需要经过总线传递到中央处理器（即保存路由表的地方），然后再次经过总线传递到输出接口，这种方式的效率不很高。在新的路由器中所采用的技术与此不同，新设计将把路由表的一部分或全部缓存在适配器接口上的特殊硬件中；此外，用于适配器之间数据传输的共享总线也已经被一个高性能的交换机所替代。

毫无疑问，所有这些新提出的或仍在研制的技术将大大提高路由器中IP转发过程的性能。另一个值得注意的技术是通过2层交换机（更明确地说就是ATM交换机）来对IP转发通路进行重定向和缓存，这个过程就称为IP交换，将在以后的章节中阐述。

2.6 路由协议

路由协议（routing protocol）使路由器能够与其他的路由器交换有关网络拓扑和可达性的信息。任何路由协议的首要目标都是保证网络中所有的路由器都具有一个完整准确的网络拓扑数据库。这一点是十分重要的，因为每一个路由器都要根据这个网络拓扑信息数据库来计算各自的转发表。正确的转发表能够提高IP分组正确到达目的地的概率；不正确或不完整的转发表意味着IP分组不能到达其目的地，更坏的情况是它可能在网络上循环一段较长时间，白白地消耗了带宽和路由器上的资源。

路由协议可以分为域内（Intradomain）和域间（Interdomain）两类。一个域通常又可以被称为一个自治系统（AS，Autonomous System）。AS是一个由单一实体进行控制和管理的路由器集合，采用一个唯一的AS号（如AS 3）来标识。域内协议（又称为内部网关协议或IGP）被用在同一个AS中的路由器之间，其作用是计算AS中的任意两个网络之间的最快或者费用最低的通路，以达到最佳的网络性能。域间协议（又称外部网关协议或EGP）被用在不同自治系统中的路由器之间，其作用是计算那些需要穿越不同自治系统的通路。由于这些自治系统是由不同的组织来管理的，因此在选择穿越AS的通路时，我们所依据的标准将不只局限于通常所说的性能，而是要依据多种特定的策略和标准，如费用、可用性、性能、AS之间的商业关系等等。边缘网关协议（BGP）是EGP的一个例子，而IGP的实例就包括OSPF和RIP等。图2-8给出了一个由3个不同自治系统构成的网络，在相邻AS之间运行EGP，在AS内部运行IGP。

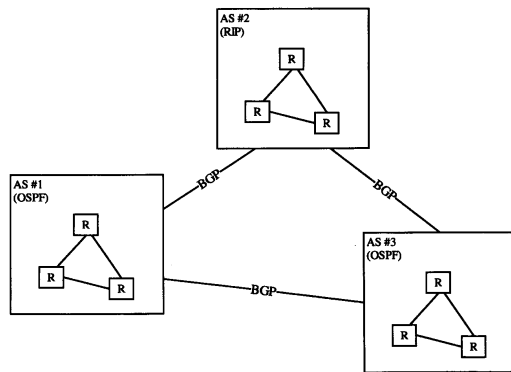


图2-8 自治系统

对路由协议进行划分的其他标准还包括：

- 可升级性（scalability）。可升级性是由路由协议自身的能力所决定的，也就是其在最大限度地减少路由器之间的控制流（路由数据库更新）及计算新路由表所消耗的路由器资源的同时，支持大量路由器和网络的能力。

- 避免环路（Loop Avoidance）。当一个路由协议计算一个新的路由表时，它应该能够防止和避免出现环路，即那些使IP分组多次经过同一个路由器或网络的通路。当路由器正在向网络中的各路由器发布网络拓扑改变消息时，它很难做到这一点。但无论如何，避免环路都

是一些协议，如BGP和EIGRP（增强内部网关路由协议）等所支持的一项重要的特征。

- 汇聚（Convergence）。当网络拓扑改变（如一条链路失效或一个新的网络加入）的时候，路由协议必须能将这个消息发布到整个网络中，以便于网络中的每一个路由器都根据这个新信息计算新的路由表，反应网络拓扑的变化。这个过程称为汇聚。路由器对网络拓扑正确汇聚得越快，IP分组就能越快地被成功传递到目的地。

- 标准。在IETF中开发的各種路由协议都记录在RFC中。这可以使不同的生产厂商能够在各自的平台上实现这些协议并提高互通性。

- 可扩展性（Extensibility）。可扩展性的定义是路由协议在不改变其基本操作和向后兼容的情况下加入新功能的能力。OSPF就是一个优秀的、可扩展路由协议的实例。很多新的功能已成功地加入到OSPF协议中，其中包括组播、QoS选路、链路层寻址支持等。

- 参数（Metrics）。除了目的网络以外，还存在一些参数和值是与之一起发布的，同时被用于路由表计算中。这类参数可能包括跳（hop）的级数、链路的费用、带宽及时延等等。

- 路由算法。路由协议一般会采用以下两种基本的路由算法中的一种，即距离向量（Distance vector）法或链路状态（Link state）法。

2.6.1 距离向量选路

距离向量选路（distance vector routing）的前提是基于这样一种思想，即每一个路由器都会把它所知道的所有网络以及到达每个网络的距离等方面的信息通知给与之相邻的路由器。运行距离向量路由协议的路由器会向与之直接相邻的路由器发布一个或多个距离向量。一个距离向量是由一个二元组 {network, cost} 构成的，其中network是目的网络，而费用（cost）是一个相对值，它反映了在发布向量的路由器与目的网络之间通路上的链路或者路由器的数量。这样，路由数据库就由许多距离向量构成，这些向量代表了从该路由器到达所有网络的距离或费用。

当一个路由器从与之相邻的路由器上接收到一个距离向量更新（update）消息时，它会把自身的cost值（通常为1）加到它所收到更新消息中的距离向量的cost值上。然后该路由器把这个新计算出来的到达目的网络的cost值与其自身记录的信息相比较（该信息是其在先前收到距离向量更新消息时计算出来的）。如果新的cost值小，那么路由器将用新的cost值来更新路由数据库，并计算生成一个新路由表，在新路由表中把发布距离向量的相邻路由器作为其到达目的网络的下一跳路由器。

图2-9说明了这个巧妙而简单的过程。路由器C发布了一条距离向量 {目的网络，费用}，指向网络1，网络1直接与路由器C相连。路由器B收到这条距离向量，并把自身的费用值加到其从路由器C上所收到的距离向量中的费用值上，并继续把这个向量发布给路由器A。这样，路由器A现在就知道它可以通过路由器B经过2步跳到达网络1。

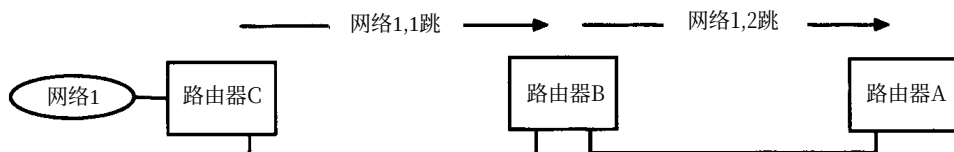


图2-9 距离向量选路

尽管距离向量选路十分简单，但它也会遇到一些众所周知的问题。例如，如果路由器 B 和 C 之间的链路中断，路由器 B 可能会试图对分组进行重新选路。重新选路的结果可能会选择到路由器 A，因为路由器 A 曾经向路由器 B 发布过一个距离向量 {网络 1, 4}。路由器 B 将接收这个向量，并且向路由器 A 返回一个距离向量 {网络 1, 5}，如此往复形成循环。这种情况被称为“无限计数 (count-to-infinity)”问题，它会导致比实际需要更长的汇聚时间。

解决这个问题的一种方法被称为“split horizon”法。其内容规定，禁止向某目的网络的下一跳路由器发布该目的地的可达性的信息。换句话说，就是路由器 A 不能向路由器 B 发布距离向量 {网络 1, 4}，因为路由器 B 是路由器 A 去往网络 1 的下一跳路由器。split horizon 方法还有一个变形，称为“poison reverse”方法，其中目的网络的可达性可以通过一个极高的费用值发布给其下一跳路由器。目的是保证原来的下一跳路由器 (B) 不会把发布路由器 (A) 当作是可到达网络 1 的。

距离向量路由协议还存在其他一些问题。由于整个的路由数据库需要在一个周期性的时间间隔内 (通常是每 60 秒一次) 发送到路由器的所有接口上，较大的路由数据库就会在低速链路上导致较大的链路开销。这个问题的一种可实现解决办法是只在网络拓扑发生了变化以后才发送路由数据库的内容，该方法称为触发更新 (triggered update)。

距离向量选路中采用的算法一般都基于十分流行的 Bellman Ford 算法，并且它已经在很多路由协议中得到实现，包括 RIP 和 IGRP (内部网关路由协议) 等。

2.6.2 链路状态选路

链路状态选路 (link state routing) 工作所基于的思想是，一个路由器能够把有关连接到该路由器的链路状态、链路费用以及任何连接到这些链路上的路由器的标识等信息通知给网络中的所有其他路由器。运行链路状态路由协议的路由器会向整个网络发布链路状态分组 (LSP)。一个 LSP 通常包含一个源标识符、一个相邻路由器的标识符以及二者之间链路的费用。LSP 被所有的路由器接收，用于建立一个网络整体的统一拓扑数据库，再根据这个拓扑数据库中的内容把路由表计算出来。事实上，网络中的所有路由器都保持了一张整个网络拓扑的地图。根据该地图，路由器能计算出从任何源到任何目的地的最短通路 (或费用最低通路)。

我们可以考察图 2-10 中的网络拓扑。附在路由器之间链路上的数值是链路的费用。路由器向网络中的所有其他路由器传播 LSP，用于建立链路状态数据库。然后网络中每一个路由器就采用最短或费用最低的通路建立一个以本路由器为根、分支到所有其他路由器的树。路由器 A 所生成的树如图 2-10 所示，路由器生成的树用于在下一步计算路由表。计算最短路径树的算法通常采用 Dijkstra 算法。

与距离向量路由协议相比，链路状态路由协议有若干优点：

- 更快速地汇聚 (convergence)。使链路状态路由协议的汇聚更快速的原因通常有几个。首先，LSP 能够迅速传遍整个网络，用来建立网络拓扑的一个准确视图。第二，在 LSP 中只反映网络拓扑的变化，而不是整个路由数据库。第三，无限计数 (count-to-infinity) 问题不会出现。

- 更小的网络开销。链路状态路由协议传送的 LSP 只反映网络拓扑的变化，而不是传送整个路由数据库。

- 可扩展性 (Extensibility)。链路状态路由协议具有扩展能力，用于支持和传播不同的网

络参数、地址和其他的拓扑信息。同时，由于路由器维护了一个拓扑数据库，当计算去往某一个特殊目的地的通路时，新的信息总是可以获得的。

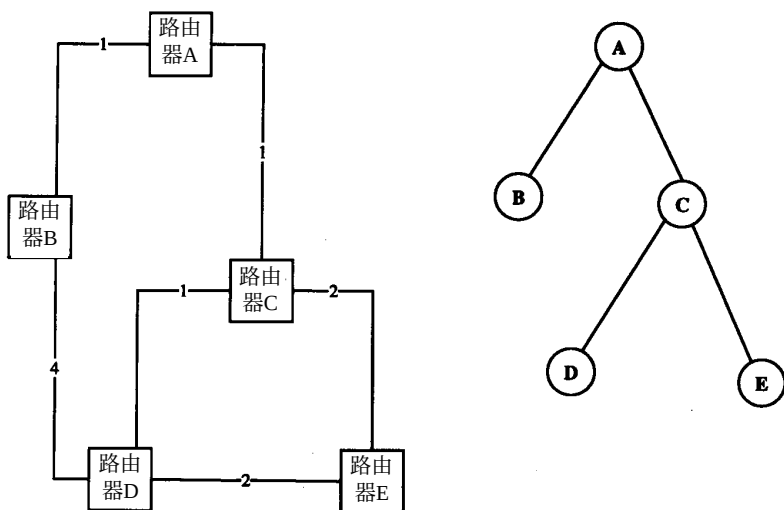


图2-10 链路状态选路

• 可升级性 (Scalability)。链路状态路由协议提供了更好的升级特性。这是因为在一个大型网络中，路由器可以被划分成多个组 (group)。在一个组中，路由器之间相互交换 LSP，并建立一个该组统一的拓扑数据库。为了在不同的组之间交换拓扑信息，一个特殊的路由器子集首先总结出该组的拓扑数据库，然后将这些总结出的数据在一个 LSP中发送给邻近的组中的特定路由器。由于在拓扑数据库中只需要本组中的路由器，并且只有拓扑结构发生变化的组中的路由器，而不是其他路由器组中的路由器，才需要重新计算最短通路树和路由表，因此这种协议能够减少路由器中所需存储器和处理过程的数量。这种分级的概念在图 2-11中说明。分级的概念是在链路状态路由协议 (如 OSPF及PNNI) 的实现过程中的一个十分重要的概念。

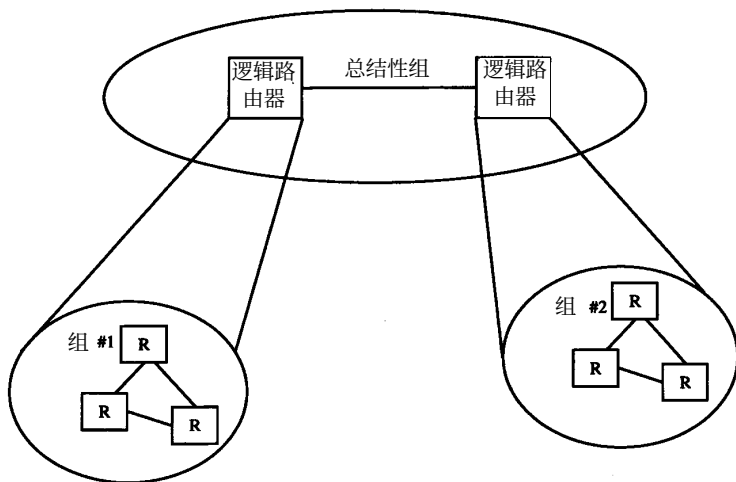


图2-11 链路状态选路的层次

2.6.3 RIP

RIP是一个几乎在任何一个TCP/IP主机或路由器中都实现的、最普通距离向量路由协议。事实上在80年代中期，随着一些UNIX版本的发行，RIP就已经被广泛传播开了。RIP在功能上的主要特征包括以下几个方面：

- RIP具有距离向量路由算法。
- RIP把转发跳（hop）的级数作为一个参数。
- 路由器每30秒广播一次整个路由数据库。
- 支持RIP的路由器网络的最大网络直径是15跳（hop）。
- RIP不支持VLSM。

目前在许多中小型企业网中，RIP的配置和运行十分简单。它属于内部或域内路由协议。为了弥补RIPv1的一些不足，RIPv2也被开发出来了。RIPv2的操作过程与RIPv1十分类似，但它增加了对VLSM的支持。这给那些在管理IPv4地址空间时需要更大灵活性的网络管理员提供了用一个OSPF来支持VLSM的替代方案。RIPv1和RIPv2分别在RFC1058和RFC1723文件中描述。

2.6.4 OSPF

今天，OSPF是一个众所周知的、采用链路状态路由算法的协议。OSPF也是一个内部/域内路由协议。市场上所有（如果不是全部也是绝大部分）的路由器都支持OSPF。OSPF在功能上的主要特点包括：

- 包含链路状态路由算法（Dijkstra），有时被称为最短路径优先（SPF）。
- 支持多条到达相同目的地的等价通路。
- 支持VLSM。
- 分为两个层次。
- 只有在网络拓扑结构发生变化时，才会产生链路状态的发布。
- 具有可扩展性。

图2-12中给出了一个由若干区域构成的OSPF网络的例子。在OSPF中，区域（area）的概念用来定义在一个自治系统（AS）中的路由器和网络的集合。在OSPF网络中，必须存在一个区域0用来定义网络骨干区域。如果配置了多个区域，那么所有的非0区域都必须通过一个区域边界路由器（ABR）连接到区域0。在一个区域中，路由器相互发布和交换链路状态通告（LSA - link state advertisement），并为该区域建立一个统一的映射图，称为链路状态数据库。区域之间通过ABR相互传递有关某一特定网络和拓扑的概括信息。因而路由器可以保存有关其所在区域中的所有网络及路由器的完整信息，以及有关在区域外网络及路由器的特殊信息。路由器中有足够的信息来引导分组通过合适的区域边界路由器到达另一个区域中的网络。

OSPF已经开始引起了网络管理和开发人员的许多关注，其原因有以下几点：

1) 由越来越多的路由器构成的大型网络正在被建造出来和投入使用，而OSPF比RIP及其他距离向量路由协议更易于升级和扩展。

2) 在这些网络中，现在需要或者将会需要采用更多的附加功能和服务。作为一个链路状态路由协议，OSPF能够通过简单地在其链路状态发布消息中增加和定义几个承载新信息的字段，来逐级扩展和增强功能。目前这种附加的功能包括组播地址和非透明LSA（一个用于承

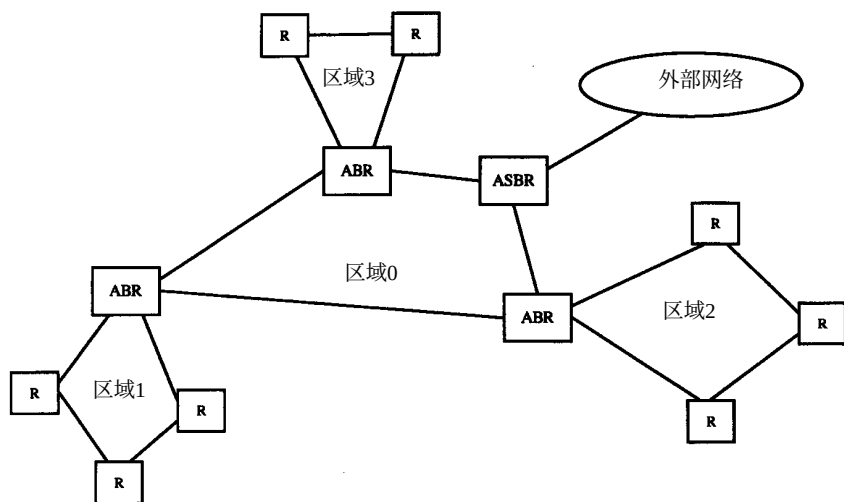


图2-12 OSPF网络

载未来发布信息的占位符)等。

OSPF在文献RFC2328中描述。

2.6.5 BGP

BGP属于外部或域间路由协议。BGP的主要目标是为处于不同AS中的路由器之间进行路由信息通信提供保证。BGP也常常被称为通路向量路由协议，因为BGP在发布到一个目的网络的可达性的同时，包含了IP分组到达目的网络过程中所必须经过的AS的列表。通路向量信息是十分有用的，因为只要简单地查找一下BGP路由更新中的AS编号就能有效地避免环路的出现。

BGP的主要功能特点包括：

- 是通路向量 (path vector) 路由协议。
- 通过对选择路由器的影响以及控制到其他BGP路由器的通路分布来实现对基于策略的选路 (policy-based routing) 的支持。
- 为了保证BGP路由器之间可靠地进行路由信息的交换，BGP中采用了TCP协议。
- 支持CIDR汇聚及VLSM。
- 对网络拓扑结构没有限制。

图2-13给出了一个由几个AS组成的、运行BGP的网络。不同AS中的路由器之间建立一个外部BGP (EBGP) 关系。同一个AS中的路由器之间建立一个内部BGP (IBGP) 关系。为了保证同一个AS中的全部路由器都保存一组一致的路由信息，在该AS中的每一个BGP路由器都必须与AS中任何其他BGP路由器建立一个IBGP关系。但是，在一个AS内进行IP分组选路时，

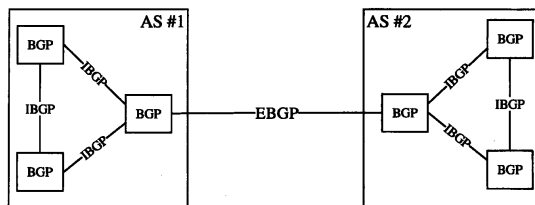


图2-13 BGP网络

只使用普通的域内路由协议如 OSPF 等，而不是使用 BGP。

在网络启动的时候，相邻的 BGP 路由器（在我们的例子中分别用路由器 1 和路由器 2 来代表）之间互相打开一个 TCP 连接，然后交换整个路由数据库。那以后，只有拓扑结构和策略的改变才会使用 BGP 更新消息发送。一个 BGP 更新消息可以声明或撤消到一个特定网络的可达性。在 BGP 更新消息中也可以包含通路的属性，属性信息可被 BGP 路由器用于在特定策略下建立和发布路由表。

目前 BGP 的版本号是版本 4（BGP4），在 RFC1771 中描述。

2.7 TCP 和 UDP

TCP/IP 协议体系结构支持两种基本的传输协议：TCP 和 UDP。TCP 代表传输控制协议，它在两个 TCP 端点之间支持面向连接的、可靠的传输服务。UDP 代表用户数据报协议，用于在两个 UDP 端点之间支持无连接的、不可靠的传输服务。

大部分传统的 TCP/IP 应用，如 telnet 和 FTP，以及类似 HTTP 这样的较新应用都使用了 TCP 作为传输功能服务。TCP 比较重要的功能特点包括：

- 面向连接。这意味着 TCP 端点之间必须在进行数据通信前相互建立一个连接。只有 TCP 端点才会关心连接的状态，而网络的中间路由器只关心 IP 分组的转发。

- 可靠的数据传递。TCP 中使用了顺序号，并采用了直接应答的方式，在必要时可通过重来保证发自源端的数据能成功地被传递到目的端。

- 流量控制。这一功能可以防止 TCP 发送者的发送流量超过接收者的接收处理能力。其实现方法是接收者向发送者发送一个接收窗口值。该窗口值告诉发送者接收者能够处理多少数据。在收到接收者发来的应答前（该应答通知发送者继续传送更多数据），TCP 发送者最多能发送的数据量只能等于窗口值。

- 拥塞控制。用于防止 TCP 发送者发送的信息量超过网络中链路或路由器的最大处理能力。在连接建立后，TCP 发送者首先通过网络发送少量的数据，然后等待接收者的应答。每收到一个应答，TCP 发送者都逐步地增加其发送到网络中的数据数，直到它检测出在某一点上出现拥塞。TCP 发送者能够根据多种因素（包括超时或接收到重复的应答等）来检测出网络中的拥塞是否到达了一个阈值。流量控制和拥塞控制结合起来，使 TCP 主机能够迅速而公平地调整其发送速率，以达到与网络及接收者的处理能力相匹配的目的。

- 只支持点到点的连接。

人们一直在不断提高 TCP 的性能和效率。在 RFC1323 中定义了许多在高速网络上提高 TCP 的性能和吞吐能力的扩展方案。RFC2018 定义了一种称为 TCP SACK（可选应答）的技术，它把窗口中所对应的数据量划分成许多数据块，接收者对每一数据块只需发送一个应答。如果某些数据丢失，只有丢失的（未应答的）数据块被发送者重传，而不是重传整个窗口所对应的全部数据。图 2-14 说明了两个主机之间 TCP 连接建立、数据交换及连接

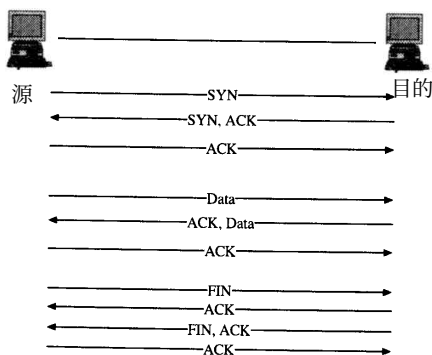


图2-14 TCP连接

撤消的过程。

在某些应用中，可能不希望使用或者根本不能使用 TCP所支持的那些面向连接和可靠的服务。这时，UDP提供了对这类应用的支持。在某些场合中，例如电子邮件等，只需传送少量的数据分组，这些应用只需要 UDP。而在另一些场合中，应用程序会自己提供可靠的传输和流量控制机制，这时UDP也是足够的。UDP协议具有以下特征：

- 无连接操作。
- 传输不可靠。
- 没有流量和拥塞控制。
- 在UDP分组头中的源及目的端口号提供了一种简单的复用 /解复用服务。
- 支持点到点和点到多点（组播）的传输。

更多的信息

Halabi, S., Internet Routing Architectures, Cisco Press, 1997.

<http://www.ietf.org>.

Huitema, C., Routing in the Internet, Englewood Cliffs, N.J.: Prentice Hall, 1995.

Moy, J. T., OSPF Anatomy of an Internet Routing Protocol, Addison-Wesley, 1998.

Stevens, W. Richard, TCP/IP Illustrated Vol. 1 The Protocols, Addison-Wesley, 1994.

China-pub.com

下载

第3章 新出现的TCP/IP协议

前一章阐述了 TCP/IP的基本体系结构、功能以及支持 IP选路的协议。本章的重点是 TCP/IP协议系列中的那些新近出现的和正在形成的协议和功能。下面我们所要讨论的这些新协议和功能在 TCP/IP的基本互连功能中引入了一些新的服务，例如 QoS、组播和安全性等。此外，有关 IPv6的功能和特点也将在本章作一个简单的讨论。

3.1 IP组播

到目前为止，我们所有的讨论都集中在 TCP/IP中的单播（即单点到单点或 unicast）通信上。IP组播（IP Multicast）是TCP/IP体系结构中的一个附加的部分，它可以在一个基于 IP的网络上进行高效率的成组通信。在那些需要进行一点对多点、多点对多点、少数点对多点以及少数点对少数点的通信应用中，IP组播是最理想的技术方案。应用的实例包括视频会议、共享公告板、“推进”技术（例如广告和信息订阅等）、资源/服务器发现（resource/server discovery）、财务数据发布以及服务器复制等等。

IP组播的基本思想是多个接收者可以接收同一个或一组源所发出的相同数据的一个拷贝。我们当然可以让这一个或这一组发送者将数据拷贝分别发送给每一个单独的接收者，但很明显由于多种原因，这种方式的效率不很高。发送者除了必须了解和管理每一个接收者的主机地址外，发送者还需要向所有接收者发送多个相同的数据拷贝，白白地消耗了网络的带宽和发送者自身的资源。

IP组播定义了一个更加高效的模型：

- 定义一个组地址（group address）。每个组地址代表一个或多个发送者与一个或多个接收者之间的一个会话（session）。
- 接收者主机可以采用组播地址告诉路由器该主机希望加入（或退出）哪一个组播组。
- 发送者使用组播组地址发送分组。发送者可以不知道任何有关接收者的信息，如接收者在什么位置等，实际上发送者也不需关心这些信息，而只需了解组地址。
- 路由器建立一棵从发送者分支出去的组播传递树，这棵树延伸到所有的、其中至少有一个组播组成员的网络中。利用这个传递树，路由器把那些寻址到组播组的 IP分组一直转发到具有组播组成员的网络中。

在图3-1中给出一个支持 IP组播的网络的构成。接收者采用某个组成员资格协议（group membership protocol）通知距它最近的路由器，表示它希望加入一个组播组。运行在路由器网络中的一个组播路由协议^①就会建立一棵组播传递树，用来将一个组播组中的分组传递给该组的成员。就发送者而言，它们只需要简单地发送一个以相应组播组地址为其目的地址的单个IP分组，剩下的工作都由路由器来完成。这个模型的规模可以从一个接收者扩展到数百或数千个接收者。任何时候，如果一个接收者希望离开一个组播组时，它也只需简单地通知距它最近的那个路由器，而与发送者无关，因为发送者根本不用关心组成员的变化。

^① 组播和单播路由协议均在同一个路由器上同时运行。

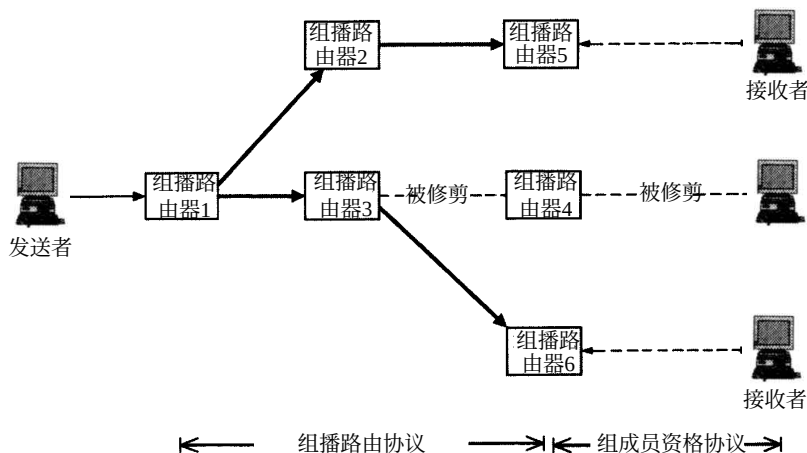


图3-1 IP组播的组件

3.1.1 D类寻址

D类的地址空间是专为IP组播组地址而定义的。每个组播组地址都落在从224.0.0.0到239.255.255.255的空间范围内。这个地址空间中的一部分地址被保留，被某些特殊的组功能、一些人们熟知的组播应用以及某些管理范畴的组播程序所使用。其余的地址部分可以在需要进行组播传送的时候动态分配。完整的保留组播组地址列表可以从Internet指定编号授权(IANA)^①那里获得。

IP组播组地址可以被映射到电气电子工程师协会(IEEE)所规定的802 MAC组播地址上。这种映射的实现过程是，取出IP组播组地址的低23位，将其添加到一个由IANA指定的特殊前缀01-00-5E之后。例如，组播组地址225.02.03.04可以映射成一个以太网组播地址01-00-5E-02-03-04。将IP组播组地址映射到IEEE 802 MAC层的组播地址可以使需进行组播传送的主机能够利用某些网络接口卡的硬件组播功能。D类地址的格式如图3-2。

	前4比特	剩下的28比特
D类	1110	组播组地区

图3-2 D类组播地址

3.1.2 IGMP

Internet组成员协议(IGMP)是一个IP组播模型中的组成员协议。IGMP被路由器用来向其直接连接的主机申请组成员资格，也被主机用来通知其直接连接且支持组播的路由器该主机希望接收地址为某个特定组播组地址的IP分组。

基本的IGMP消息交换在图3-3中进行了说明。路由器首先利用一个可寻址到所有主机的组地址(即244.0.0.1)发送一条IGMP主机成员资格查寻(IGMP Host Membership Query)消息。若一个主机希望加入某组播组，它就利用该组播组的组地址回应一条IGMP主机成员资格报告(IGMP Host Membership Report)消息。在我们给出的例子中，1号主机利用主机成员资格报告消息回应了查寻消息，并指明它属于“1号组”。2号主机用同样方式指明它属于“2

① <http://www.isi.edu/iana/>。

号组”。每个组播组路由器只需要接收一条主机成员资格报告。一种消息抑制算法可以防止相同子网上其他主机再次发送某特定组的主机成员资格报告（在已经有主机发送的情况下），它采用随机的延迟防止了来自同一个组播组的重复的成员资格报告把 LAN和路由器的接口淹没。

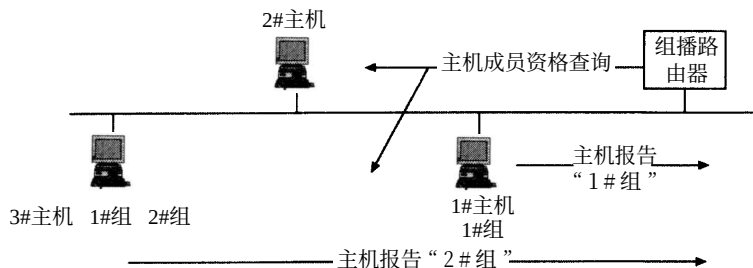


图3-3 IGMP流

在RFC1112中提出了IGMP的版本1；RFC2236定义了IGMP版本2，它引入了一些增强的新功能来提高IGMP的效率。其中包括一种在LAN上的组播查询者（multicast querier）的选举技术（假如一个LAN上有两个或多个询问者的情况下）；其中还引入了两条新消息：特定组询问（Group-Specific Query）和脱离组（Leave Group）。这两条新的消息可以使组播查询者能够查寻任何一个属于某特定组播组的主机，并且可使主机能够立即离开一个指定的组播组而不需要等待一个超时时间，也就降低了所谓的“脱离延迟（leave latency）”。

3.1.3 组播选路

组播路由协议通常采用两种基本算法来创建传递树（delivery tree），并通过这些传递树把组播分组转发到组播组的成员网络中。第一种算法称为反向通路转发（RPF），第二种通常被称为核心基础树（CBT）。

RPF算法十分简单：

- 1) 当一个组播分组到达路由器的一个接口时，如果这个接口是用于向发送者发送单播分组的接口，那么就把它这个组播分组转发到所有其他的接口上。
- 2) 否则，丢弃该分组。

图3-4说明了这个过程。路由器A在接口w1上接收到一个来自源S.1的组播分组。路由器A首先检查分组中的源地址（S.1），然后查寻路由表，检查哪一个接口可以用来向主机S.1发送单播分组。由于这个组播分组是在通往S.1的下一跳（next-hop）接口即w1上接收到的，因此这个分组被复制并被发送到所有的其他接口上。如果这个组播分组是从任何一个其他接口上接收到的，那么路由器A将丢弃该分组。

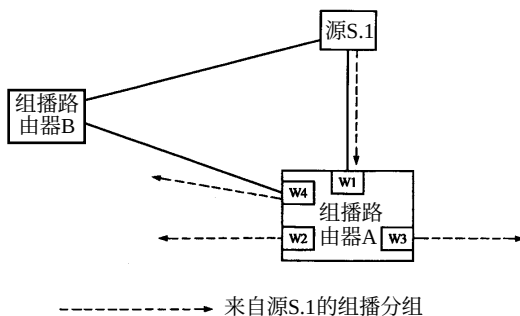


图3-4 反向通路转发

采用RPF算法的最终结果是每一个路由器都建立了一棵传递树，或者更准确地说，是一棵以源为根、分支到所有网络和路由器的最短路径树（SPT）。但是，单独使用RPF算法会导致

一些不必要的分组复制，也会使组播分组被转发到一条不存在活动组播组成员的链路上。其中导致后者的原因是在 RPF 处理过程中，由组播分组转发到一个路由器时并不关心组播组的成员资格。

为了克服这些不足，人们已经对原始的 RPF 算法进行了一些改进。如果一条链路所连接的两个路由器分别存在可到达发送者的不同路径，在这条链路上就可能会发生分组的重复复制（例如在图 3-4 中连接路由器 A 和 B 的链路）^①。如果每一个路由器都能够确定“在树上（on-tree）^②”谁是它的上游和下游邻接点，这种重复的复制就可以消除。存在多种技术可以用于实现这个判断过程，包括路由表检查、路由数据库消息交换等，或者还可以直接发送或接收一个显式的消息。通过这些努力，最终结果是分组不会被转发到不在传递树上的链路上去。

我们通常应该利用组成员资格信息来修剪传递树，把那些没有组成员的路由器接口和链路从传递树上剪去。截短的反向通路广播（TRPB）是对 RPF 的另一种改进，它的基本原理是，如果一个路由器接口所直接连接的子网上没有活动的组成员，就把该接口从传递树上删除。反向通路组播（RPM）是一个进一步优化的算法，它可使一个路由器在其直连接口及所有下游路由器的接口上都没有活动组成员时，能向其上游路由器发送一条传递树修剪消息。在图 3-1 中，SPT 的根位于发送者，分支到所有具有活动组成员的网络。可以注意到，我们可利用 RPM 算法把没有成员的链路和路由器修剪掉。

使用 RPF 算法并结合了上述的一项或多项改进的组播路由协议通常被归类为“广播和修剪”协议。其含义是，当从发送者收到一个组播分组时，路由器将把分组广播到其所有的接口上，由此形成一个以源为根、分支到所有路由器的最短路径树。一旦组成员发生变化，传递树将被动态修剪，只保留包含组成员的那些网络。经过一段时间后，修剪状态超时，广播过程将会重复执行。采用广播和修剪技术来支持组播的方式在实现上比较简单，但是会造成带宽的浪费；并且由于路由器必须为每个 {源主机，组地址} 对保存 $O(S \times G)$ 的信息，因此不容易进行扩展和升级。

CBT 是一些组播路由协议所采用的另一种算法。CBT 定义了一个核心路由器，包含有组成员的所有下行路由器都需要向核心路由器发送显式的加入消息。首先需要建立一个单一的特定组传递树（group-specific delivery tree），其根位于核心路由器，并且分支到（也只分支到）那些具有组成员的下行路由器上。加入和建立树过程与任何发送源或组播分组的存在无关。组播组的发送源将所有组播数据转发到核心路由器，核心路由器把组播流量注入到特定组传递树上。这种方法可以使相同组播组内的多个发送源共享相同的一个传递树；因此 CBT 组播路由协议也被称为共享树协议。

CBT 的扩展性比广播修剪技术好得多，因为对每一个组来说，路由器只需要将自己与一个统一的基于核心的传递树相关联。这意味着路由器只需要维护和存储 $O(*, G)$ ，其中 G 是组地址的数量。此外，CBT 也不会像广播修剪协议那样周期性地向网络中注入大量数据。核心路由器在收到来自下行路由器显式加入消息之前将阻塞组播分组的发送。而只有在路由器接收到来自主机的 IGMP 主机成员资格报告以后，这种显式的加入消息才可能产生。

① 这个链路也可以是一个共享介质的 LAN，使用相同的机制来防止分组的重复复制。

② 在树上（on-tree）的含义是指那些位于传递树上的路由器；该传递树以组播源为根，分支到具有活动组播组成员的所有路由器。传递树实际上是能够转发组播分组的一个路由器和链路的集合。上游是指向着根（发送源）的方向；下游是指远离发送源的方向。

3.1.4 组播路由协议

目前已经定义的多种组播路由协议一般可分为两大类：密集模式（dense-mode）和稀疏模式（sparse-mode）。密集模式的协议适合于工作在具有较高带宽并且其中的组成员较为集中、互相间十分接近的网络环境中。密集模式的协议通常采用广播和修剪方式，因为这种方式十分简单，而网络本身又有能力吸收其中的开销。能够良好运行密集模式协议的网络环境的典型实例是校园 LAN 和企业 Intranet。稀疏模式协议所适用的环境是带宽不够充裕，并且组成员分散在不同网络和地理区域中。在这种情况下，采用通过向单一核心路由器发送显式加入消息来实现的、由接收者驱动的方法是很有意义的。这种网络的应用实例可以是一个极大型的企业的 Intranet，甚至可以是 Internet。

距离向量组播路由协议（DVMRP）的出现和使用已经经历了相当一段时间，目前仍然被广泛地使用在 MBONE（覆盖在 Internet 上的 IP 组播骨干网）中。DVMRP 采用了 RPM 算法转发分组，并定义了一种“隧道”方法，使组播分组能够通过不具备组播能力的路由器来转发。MOSPF 协议提供了对 OSPF 协议的一系列组播能力的扩展，它使用 OSPF 的拓扑数据库来计算以源为根的最短路径树。PIM 是“与协议无关的组播（Protocol Independent Multicast）”的缩写。PIM 又分为两种模式：密集模式（DM）和稀疏模式（SM）。PIM-DM 与 DVMRP 十分类似，不同之处在于它不依赖于那些基于距离向量的路由功能，例如距离向量路由更新以及利用 split horizon 方式来通告树上游、下游路由器的存在。PIM-SM 能够建立一棵根为某聚合点（RP - rendezvous point）路由器的共享树，并且在需要更高性能的情况下，它能够动态地将共享树切换到一个以源为根的最短路径树中。

PIM 和 CBT 都是比较著名的算法，二者实现组播时都不依赖于任何现有的底层单播路由协议的功能。它们能够与任何单播协议如 RIP 及 OSPF 等一起运行在同一台路由器上。而 DVMRP 和 MOSPF 却依赖于某些单播路由功能：DVMRP 运行一个类似于 RIP 的距离向量选路算法；MOSPF 需要 OSPF 的功能。

表 3-1 总结了多种组播路由协议的功能和特点。

表 3-1 组播路由协议

	名 称	密集/稀疏模式	算 法	环 境	文 献
DVMRP	距离向量组播路由协议	密集模式	RPM	校园网, Intranet	RFC1075
MOSPF	OSPF 的组播扩展	密集模式	Link-State	校园网, Intranet	RFC1584
RIM-DM	PIM 密集模式	密集模式	RPM	校园网, Intranet	
PIM-SM	PIM 稀疏模式	稀疏模式	RPM/CBT	Intranet, Internet	RFC2117
CBTv2	核心基础树版本 2	稀疏模式	CBT	Intranet, Internet	RFC2189

3.2 集成服务

IP 网络所支持的基本服务模型只有一种服务类型：尽力而为（best effort）。其含义是不管提交给网络的业务流的类型是什么（声音、图像或数据），一旦进入了网络，就以先来先服务的方式对信息分组进行服务。在过去和现在，这种方式都能够较好地满足那些没有时延和带宽约束的纯数据应用的需要。但是 Internet 或者更普遍地说整个 IP 网络，正在迅速演化，它不仅需要支持传统的尽力而为的数据应用，还必须支持声音、图像和多媒体业务等新的数据类型。

IP集成服务体系结构（IntServ）就是对现有IP网络模型的一个增强结构，它可以同时支持和提供实时服务和尽力而为服务。开发这个模型的动机是很明显的：

- 尽力而为服务已不能继续满足需要。这是毫无疑问的，各种应用正在不断被引入到网络中，它们具有不同网络需求。终端用户也需要获得可靠和一致的性能，如果仅提供一种服务，网络提供者将不可能获得更大盈利。

- 多媒体应用组件的出现。IP网络不再限制于仅提供基于字符的应用，网络必须能够提供不同业务类型的综合，包括图像、声音和数据等，并提供QoS保证。

- 优化网络和资源的使用。采用不断提高网络容量的方式来支持各种实时和尽力而为的业务是不可能的，也不必要（这在LAN环境中是可能的，而在WAN环境中根本不可行）；而为实时业务的处理保留其所需的资源将是这一问题的一种更经济更理想的解决方案。在这种情况下，尽力而为的业务流将会使用剩下的网络容量来传送。

- 提供和传送额外服务。IntServ模型使网络提供者能够提供和支持一些特殊的额外服务，以体现其与竞争对手的区别。

在分组网络上支持QoS并不是一个新的概念，为了支持这项功能，在过去的几年中已经提出了大量的研究论文、实验和原型系统。IETF提出的RFC1633是第一个描述IntServ模型的需求和基本组成的正式文献。后来IETF最终成立了集成服务（IntServ）工作组^①，并在其宪章中详细说明了“这种增强的服务模型，并定义和标准化了实现这一新服务模型所必需的接口和需求。”目前，IntServ工作组已经完成了IntServ模型的定义、基本组成部分、新业务类型等方面的工作。该工作组的工作是与RSVP工作组协同进行的^②。

3.2.1 IntServ模型

图3-5给出了IntServ模型，该模型有以下几个组成部分：

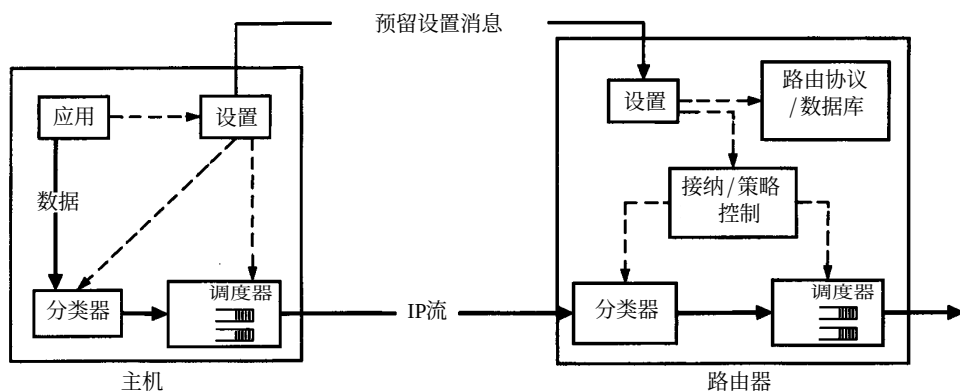


图3-5 集成服务模型

- 设置协议（setup protocol）。设置协议使主机或路由器能够动态地保留网络中的资源，以便满足特殊业务流的服务需求。RSVP就是设置协议的一个例子，其他的例子还包括ST2+和Q.2931。

- 流量说明（flow specification）。用于定义提供给一个特定业务流的QoS。在IntServ模型

① <http://www.ietf.org/html.charters/IntServ-charter.html>。

② <http://www.ietf.org/html.charters/rsvp-charter.html>。

的上下文中所定义的业务流（flow）是指具有相同QoS的、从某个源到目的端的一个分组流的序列。从概念上不太严格地说，流量说明可以被看作是一个保留的网络资源的集合（如最小带宽等），这些资源是网络支持该业务流QoS所必需保留的。

- 流量控制。在网络设备（如主机、路由器或交换机）中存在一些组件能够控制和管理为支持特定QoS所必需的网络资源。

流量控制组件的配置可以采用某种信令控制协议，如RSVP，也可以采用人工配置。流量控制组件由以下几个组成部分：

- 接纳控制（Admission Control）。确定一个设备（主机、路由器或交换机）是否能够支持某特定数据流所需的QoS。

- 分类器（Classifier）。以分组头中特定字段的内容为基础鉴别和选择某种业务类型。

- 调度器（Scheduler）。在设备的输出链路上提供所需的QoS服务类别。

从这个模型中可以看出，设置协议只是一个控制协议，它不承载任何数据。在IntServ模型中，转发通路的关键组成部分是分类器和调度器。

3.2.2 IntServ流类型

IntServ目前支持3种业务流类型：保证服务（GS）、控制负荷（CL）以及尽力而为的服务。GS^①能为符合该类型的分组提供专用的带宽和限定的时延，不会因为队列溢出而丢失分组。GS是为那些具有极严格时延和带宽特性需求的应用而设计的。能够采用和受益于GS的应用包括高质量会议电话、实时财政事物处理，以及那些需要进行电路仿真的网络业务。GS是唯一支持可计量QoS参数（时延）的IP流类型。

CL^②服务是一种在轻负载网络上的类似于尽力而为性质的网络服务。CL并不保证带宽或时延，但是也不像尽力而为的服务那样，CL的性能不会在网络负载增加时明显下降。CL是为能够容忍一定数量时延和丢失的应用而设计的。这类应用的例子包括中低质量的音频视频组播或者可能是为一个能够承载时延敏感应用的专用Intranet，如SNA或DEC局域传输（LAT）等提供隧道机制。

3.2.3 RSVP

RSVP^③是IntServ模型中的一个设置协议成员。从本质上说，RSVP是一个信令协议，它可使IPv4（或IPv6）的应用能够为某个固定网络元素（主机或路由器）通路上的分组流保留资源。尽管一个具有足够网络带宽和较低时延的传统尽力而为的IP网络可能会满足特定应用的QoS需求，甚至我们可以在路由器通路上人工配置一个IntServ流量控制组件来满足其QoS需求，但是RSVP给予了实际应用一种从网络中动态请求、接收和调整QoS需求的能力，以适应网络变化。

在描述RSVP的工作过程以前，让我们先来确定一下RSVP到底是什么和不是什么。首先，RSVP不为某个流定义或提供QoS。QoS的语义是由IntServ体系结构以对象的形式定义的。流说明（FlowSpec）就是一个IntServ对象的例子，它用于参数化地为一个特定的流定义分组调

① RFC2212, “Specification of Guaranteed Quality of Service”。

② RFC2211, “Specification of the Controlled-Load Network Element Service”。

③ RFC2205, “Resource ReSerVation Protocol (RSVP) — Version 1 Functional Specification”。

度器。一个流的 QoS 是由流量控制组件中的预留状态提供的，这些流量控制组件也就是分类器（用于识别流中的分组）和调度器（用于根据 QoS 级别在出口链路上调度分组）。RSVP 只简单地、不透明传输 IntServ 对象并将其交给通路上每一个主机或路由器的流量控制组件。第二，RSVP 不是一个路由协议。RSVP 可以与其他任何的任意路由协议共同工作，如 RIP、OSPF 或 BGP，并用这些协议为 RSVP 消息选择一条从源主机到目的主机的路由。RSVP 的设计者有意识地规定，每种选路过程都分别有各自的保留设置，以便在各种 IP 网络上都给 RSVP 以最大的灵活性，使其操作不必依赖于某个特定的路由协议。RSVP 作为 IntServ 模型的一部分，它只关心将 QoS 状态传递给某条通路上的主机和路由器，而该通路是由普通的路由协议计算出来的。

RSVP 的操作与其他的呼叫或预留设置协议类似，都是比较简单的。资源被预留给源与目的主机之间一条固定通路上的某个特定的呼叫（或流）^①。但是 RSVP 的工作方式与其他的呼叫或预留设置协议之间有一些明显的差异。造成这些差异的主要原因是：RSVP 本身的设计就是为了使其运行在一个网络动态（network dynamics）可能突然发生状态变化的无连接 IP 网络中；此外基于 IP 的终端用户和应用本身就有很大的差别。RSVP 所特有的功能如下：

- RSVP 是由接收者初始化的。负责初始化和向网络发送资源预留请求是流的接收者，而不是发送者。在大型组播组中，成员资格可能是动态的，不同的接收者可能占有不同的网络容量（例如有些可能使用 28.8KB 的模拟线路连接，而另一些可能采用 100MB 以太网）。在此情况下，不必让发送者负担所有这种信息，而让接收者自己决定自己的 QoS 能力，然后生成合适的资源预留请求。

- RSVP 是软状态（soft state）的。在一条固定通路上的预留状态如果不刷新，是会超时的。例如，大型组播组的组成员资格是动态的，我们可以采用简单地让预留状态过时的方式，来避免处理已释放的预留所造成的开销。另外，拓扑的变化也会造成通往某一目的地的最佳通路的变化。

RSVP 在组播环境中的工作过程在图 3-6 的概念框图中说明。它由以下几个基本步骤组成：

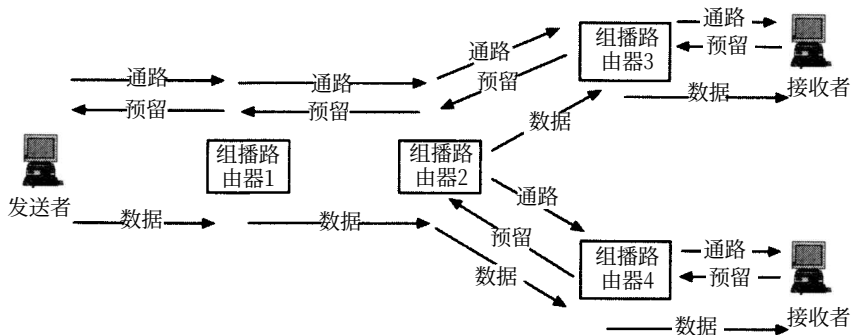


图3-6 RSVP 的操作

- 1) 发送者开始发送某一组播地址的分组。
- 2) 接收者利用 IGMP 主机报告消息向其直连的路由器申请加入这个组。借助于某种组播路由协议，路由器建立了一个发送者与接收者之间的传递树。

^① 在这种情况下，固定通路实际上是动态路由协议计算出来的路由通路，RSVP 在该通路上预留资源。如果动态路由协议计算了一条新的通路，RSVP 必须新的通路上保留资源。

3) 发送主机上的应用程序通过一个应用程序接口 (API) 向本地的 RSVP 实体注册并传递信息。这个信息标识出了发送主机、目的地 (单播地址或组播地址) 以及发送应用程序所生成的流的特性。

4) 发送主机产生一条 RSVP PATH 消息, 并将其发往接收主机。PATH 消息会在通路上每一个具有 RSVP 能力的路由器上都停下来, 存储某些路由器的通路状态 (通过留下前一跳或上游节点的 IP 地址), 并收集有关路由器的 QoS 特性的信息 (通过更新 ADSPEC 字段)。PATH 消息标明了发送者到接收者之间的通路, 并收集了通路的特性信息。一旦 PATH 消息到达目的端, 接收者就可以得到一个很完整的通路状况, 以便其根据发送者的发送特性及 PATH 消息所搜集到的路由上的 QoS 能力来确定它能申请怎样的 QoS 参数。

5) 接收主机上的应用程序通过一个 API 接口将预留请求传递给本地的 RSVP 实体。然后将会产生一个 RSVP RESV 消息, 并按 PATH 消息的相反方向上行发送给源主机。RSVP RESV 消息中包含一个流描述符, 它被传递给通路上每一个支持 RSVP 的路由器中的流量控制组件, 同时, 接纳控制和策略控制功能也被调用。如果预留请求被接受, 分类器和调度器将被更新。在返回发送者的分支点上 (如组播路由器 2), RSVP 将把多个下游预留汇聚成一个单一的预留请求, 这样可以减少路由器所需要维护的状态的数量。

6) 一旦 RESV 消息到达发送者, 一条从发送者到接收者方向的、具有预留资源的固定通路就被建立起来了。

7) 由于 RSVP 是软状态的, 因此必须周期性地产生 PATH 和 RESV 刷新消息, 否则状态就会超时。当然, 如果拓扑结构的变化导致了在发送者和接受者之间产生了一条新的通路, 那么整个的 PATH/RESV 过程将在新的路由通路上被重新执行。

在 RFC2205 到 RFC2216 的文献中对 IP IntServ 和 RSVP 作了详细描述。概括地说, RSVP 最重要的功能如下:

- 为单点流和组播流提供资源预留的支持。
- 预留是单向的, 也就是说只能建立从发送者到接收者方向的资源预留。
- 预留请求由接收者产生。
- 预留是软状态的, 并且需要定时刷新防止其超时。
- 不透明地传送流量控制和策略控制信息。

3.2.4 其他的 RSVP 和集成服务方案

IntServ 模型与 RSVP 相结合, 定义了一种使主机和路由器能支持 QoS 的机制。但是, 它还不够以在现有的 IP 模型上提供真正的、端到端的 QoS。通常在任何一个网络中, 主机和路由器之间都是通过不同的数据链路技术互相连接的。其中的一些链路技术如 ATM 可以支持 QoS。而其他的一些技术, 如共享的以太网等, 想为特定流提供 QoS 可能是十分困难的。IETF 成立了一个称为在特定链路层上的集成服务 (ISSLL)^① 的工作组, 来解决在不同的数据链路技术 (如 ATM、802 LAN 及低速链路等) 上支持 IntServ 模型和 RSVP 的问题。

另一个方案涉及到了 IP 路由协议中缺乏 QoS 选路的问题。这些路由协议所使用的参数是由非 QoS 变量组成的, 如转发的跳数、到目的网络的链路费用的总和等。类似于可用带宽容量、时延、抖动及其他的 QoS 属性都没有被采用。在 RSVP 中, PATH 消息是通过一条路由由协议

① <http://www.ietf.org/html.charters/issll-charter.html>。

采用非QoS参数计算出来的通路来转发的。这就可能导致我们试图去建立一条非最佳的通路，因此可能使预留请求被拒绝。即使另外一条通路上有足够可用的 QoS容量，路由协议也没有办法去检查和选择这条通路，甚至根本不可能使用替代通路。有关 QoS选路的问题定义和解决方案目前已经开始被IETF的服务质量选路^① (QoS SR) 工作组进行研究。

最后，正如RFC2208中所指出的，关于RSVP的操作还有以下一些方面是需要关注的：

- 可扩展性。对目前的路由器来说，要支持和维护许多小的 RSVP资源预留或对高带宽连接进行流分类可能是比较困难的。现在仍然没有一个明确的方法来为一组汇聚的流定义资源预留。
- 安全性。防止服务被窃取和破坏也是一个很重要的问题，它必须通过一个适当的分布式密钥管理系统来解决，但是目前还没有一种系统能够被实际应用。
- 策略。需要在某种策略基础上进行业务传递的网络也将会需要成熟的策略控制机制与RSVP实体共同工作。

3.3 差分服务

在基于IP的网络上提供差分的服务级别已经成为了一个十分活跃的技术研究及标准开发的领域。起初，人们认为ATM强大的QoS能力能够解决这个问题，但是我们一直（直到现在）没有实现ATM直接连接到桌面或家庭。另一个很有前途的解决方案是前一节所述的集成服务模式。然而，尽管理论研究、仿真和早期的测试^②结果都证明该模型是可行的，但是或许RSVP并不能成为每个人都期望的那种端到端的 QoS的解决方案。其原因很简单：维护每个流的状态和信令的开销使得这种模型极不利于扩展。因此，必须在目前 Internet的这种尽力而为的FIFO服务与RSVP及IntServ所提供的、针对每个业务流的复杂的机制之间寻找一种折衷方案。这个问题的答案被称为差分服务 (DiffServ)。DiffServ使用了分组标记和按类排队的方法，以便在基于IP的网络上支持具有优先级的服务。DiffServ^③目前是IETF的一个正式的工作组，该工作组的工作任务是开发在IP网络上的差分服务的标准。

3.3.1 DiffServ的模型和操作

DiffServ的基本思想包括以下几点：

- 定义一组数量较小的服务类型和优先级。例如简单地定义低、中和高三种优先级。或者简单地定义尽力而为和实时两种服务类型。每一种服务类型都可以有一个与之相关的业务流特性描述文件 (profile)，其中包含最大和最小的允许带宽以及突发长度和持续时间等信息，当然也可能只简单记录最小带宽数量。
- 在网络的边缘对所有分组进行分类，并标记每个分组所属的服务类型。该过程将检测每个分组中的IP头或传输头中的一个或多个字段。每个分组的头中都可能有一个或多个比特被标记，用来反映出该分组的服务类型以及某些每个跳 (per-hop) 的行为如丢弃优先选项 (drop preference) 等^④。

^① <http://www.ietf.org/html.charters/qosr-charter.html>。

^② Baugher, M., and Jarrer, S., “Test Results of the Commercial Internet Multimedia Trials,” Computer Communications Review, vol. 28, no. 1, Jan. 1998。

^③ <http://www.ietf.org/html.charters/diffserv-charter.html>。

^④ 丢弃优先选项指出该分组在遇到拥塞时是否自愿被丢弃。

- 核心网络交换机和路由器依据分组头中标记比特的内容来对分组进行处理。

这种模型与 IntServ 模型相比有很大的优越性。首先，在网络中不需要维护每个业务流（per-flow）的信令或状态。第二，可以为一个由许多单个流汇聚映射成的服务类型提供具有优先级的服务。这使得网络操作者能够轻松地所有可能使用网络的潜在终端用户及应用程序提供一种数量较少但又有所区别的服务级别。通过将多个单独的流汇聚成较少的一些特殊业务流，可以使这种模型能够比那些每个业务流（per-flow）的技术能够更高效地扩展和升级。这种技术的另一个通用的名称是 COS。

DiffServ 模型的另一个重要优越性是，每个单独的主机或应用程序都可以不需要修改就能接收具有优先级的服务。对进入网络的 IP 分组进行分类、标记以及可能的调整都是由网络边缘设备完成的。与此类似，核心路由和交换部件也只需要很少的改动或者根本不需要改动。今天这些设备中的大部分都支持某种分类的队列管理。与将其整体升级以支持 RSVP 和 IntServ 流量控制的方案相比，在其中加入一个能进行分组识别并把分组放进相应的队列的比特级分类器只是一个代价很小的升级。如果核心网络由帧中继或 ATM 构成，边缘设备只需要简单地将分组放入具有不同优先级的队列中就行了。

DiffServ 为网络服务提供者所支持的各种服务提供很大灵活性和自由度。分组头中的服务比特只被用于对分组进行分类的用途。定义和提供什么种类的服务完全由服务提供者来决定。一些服务提供者可能会为高优先级的业务流提供带宽和时延的保证；而其他提供者可能只能保证最小带宽。唯一的要求就是对这些比特语义的理解必须保持一致，以便使众多 ISP 的网络能以相同的方式理解这些优先级位。

DiffServ 的另一个重要的优点是将在下一节中讨论的虚拟专用网（VPN）和差分服务的概念。VPN 是通过在 Internet 中建立安全隧道形成的。流过一个安全隧道的 IP 分组的传输头和净荷的一部分或全部都会被加密。在此情况下，DiffServ 仍然能够为这些分组提供具有优先级的服务，因为它只检查分组头明文中的比特，这使差分服务能工作于 VPN 之上。

DiffServ 也存在一些问题需要解决。首先，与 IntServ 和 ATM 相比，DiffServ 网络所提供的服务在带宽和时延方面可能得不到 ATM 那样的保证，除非网络能够直接提供这种保证，而实际上这是由构成互连网络的各个独立网络所决定的。第二，边缘组件仍然需要高性能的每个分组（per-packet）分类器，这与 IntServ 模型中的同类部件在实际上并没有任何差别。此外，还有一个需要解决的问题是如何在潜在数量很大的边缘组件上进行分类器状态的管理。这个问题可能的解决方案包括人工配置、简单网络管理协议（SNMP）、轻量目录访问协议（LDAP）及 RSVP 等。最后的一个重要问题是如何给这些服务确定价格，以便只能有一部分的消费者或用户能够访问。如果我们提供的高优先级服务与传统的尽力而为的服务之间只有一个名义上的、无实际意义的价格差，我们就可以想像每个人都会去申请高优先级的服务，这将破坏和我们提出差分模型的初衷。

DiffServ 模型由下列组件构成：

- DS 字节。差分服务字节实际上是对 IPv4 TOS 字段的翻新，类似于新的 IPv6 中定义的业务流类型（Traffic Class）的字段。在该字段中的比特指明了分组希望获得的服务。

- 边缘组件。这些设备都位于 DiffServ 网络的入口和出口。它们负责对分组进行分类、标记和调整（如管制、整形等）。边缘组件可以是一个主机、防火墙、路由器或其他可能的边缘设备。

- 内部组件。这些设备可能是核心交换机或路由器，它们可以根据 DS字节的内容提供每个跳（per-hop）的服务。这些组件采用队列管理技术来控制队列的深度（如随机早期检测即RED），并且在每个服务（per-service）基础上采用一定策略对出口的发送过程进行调度（如加权公平排队[WFQ]、基于类的排队或优先权排队等）。

- 策略管理。我们需要对服务策略进行管理，或许还需要在网络和终端用户之间指定或强制某些服务级别的协定，这些工作都需要一定的工具和人员来完成。这是因为总会有某些人在其受到更好的服务的同时，需要花费更多金钱。

包含在DiffServ的边缘和核心路由器中的特殊要素由图 3-7说明。

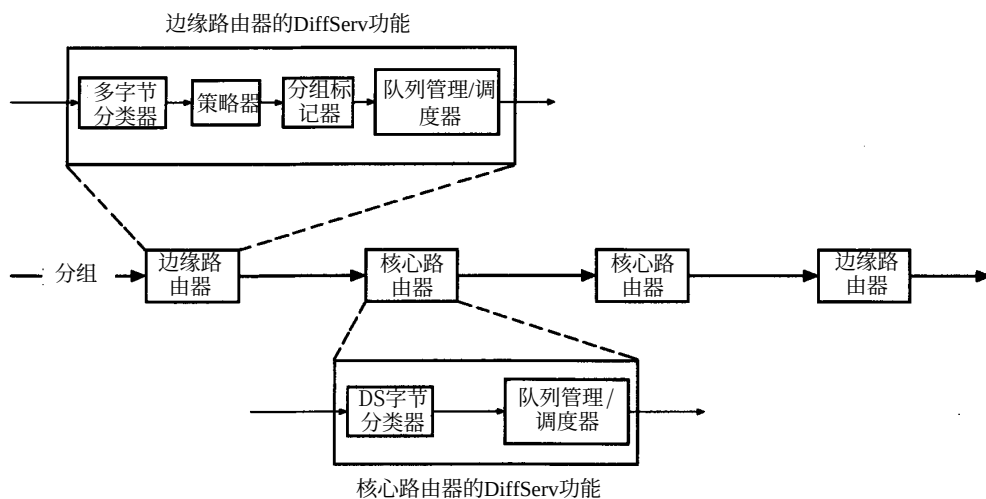


图3-7 DiffServ的边缘路由器和核心路由器的功能

边缘路由器的组成和功能如下：

- 多字节分类器（multibyte classifier）。可以检查分组中的一个或多个字段，包括 IP头和传输层的头等，甚至包括净荷。
- 策略器（policer）。可以对分组进行整形或丢弃，以便使之能够符合其所传输的服务的流量约定（profile）。
- 分组标记器（marker）。标记DS字节中的一个或多个比特，以标识这些分组是否与某个特定类或协定（profile）相关。
- 队列管理/调度器。管理队列的长度，调度分组的发送。

核心路由器的功能较少，也更简单：

- DS字节分类器。检查IP头内的DS字节。
- 队列管理/调度器。与边缘路由器中的相同。

正如图3-7中所示，DiffServ把每个分组（per-packet）分类等复杂的工作推到了网络的边缘，而让高性能的核心交换机和路由器去干它们最擅长的事情，即以极高的速度转发分组穿网络。

3.3.2 DS字节的格式

尽管在IPv4中最初设计TOS字节的目的是用来建立传送优先级和进行TOS路由选择，但是

在实际中该字段却从来没有被使用过；即使有人用过，也是在专用环境中相同设备之间的特殊实现。由于DiffServ的基本功能是提供每个分组的优先级服务，因此它决定重新使用 TOS字节的函数，并将其内容重新定义在 DS字节中，以便实现DiffServ模型的功能目标。

迄今为止已经出现了不少关于如何分配这些比特的建议^①。图3-8是一个综合了早期研究成果而形成的初步建议。开始的 6比特字段定义了每个跳的行为（per-hop behavior），它规定了在一个特定的交换机或路由器上该分组所应该获得的服务或转发行为。两个 CU比特在目前是未定义（currently undefined）的。

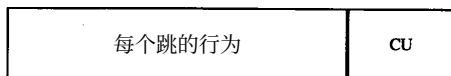


图3-8 DS字节的定义

在这个领域中的研究工作仍然在继续，DS字节的定义有可能会改变。一般地说，最终的定义应该包括一个每个跳行为，由它来给出一些建议的服务（如缺省的或高优先级的），但同时也为网络提供者的特定处理的实现留下足够的空间。当然其中也很可能有一个丢弃优先选项比特，它在一些活动队列管理技术^②（如RED）的应用中对控制拥塞是十分有用的。其他可能的字段还包括一个显式拥塞指示（ECN），它可以使源端的发送速率能自动地适应网络的可用容量。

3.4 IP安全性

网络的安全性一直是数据网络中的一个必需而重要的因素。在先前的网络中，只有计算机本身和应用程序是资源。保护这些资源是一个相对简单的工作。机密的数据被存储在一个安装在玻璃机房中的大型主机系统中。通过密码保护进行鉴权，只允许在企业网中封闭的一部分终端用户对其访问，而且企业网本身也是对外封闭的。对许多政府部门和学术机构来说，Internet只是一个用于转发电子邮件和进行文件传输的工具。

Internet和计算机技术的革命改变了所有这一切。现在，计算机十分高速、廉价并具有可移动性，一般用于储存机密信息。Internet在使人们能够全球范围内进行无障碍通信的同时，其不安全性也是可以想象的。这意味着终端用户传输数据所使用的网络元素可能会不在他们的控制和视野之内。如果数据本身是机密的，它就会暴露给可能的窃贼，被无权限用户查看或删改。

与此同时，专用的企业网也需要使用 Internet并与之交互。Internet在广告和电子商务方面蕴藏着巨大的商机，对用户来说 Internet必须是完全可见的。但是这样做将使企业网络的资源完全暴露在黑客、行骗者和那些爱恶作剧的人的攻击之下，其中可能还包括一些不道德的分子为了取乐或某些更险恶的动机所进行的破坏。不管原因如何，现在对网络安全要求比以往更严格，也更必要了。

有多种方法可用于保护网络中的数据流。可以通过在专用网络的边缘建立一个防火墙，过滤掉那些未经授权的数据流。应用程序和传输协议可以使用其自己的安全机制。另一种技术是在IP层或更明确地说是在每个 IP分组的基础上提供安全性保障。由于以下几个原因使这种方法具有一定的意义：

- Internet和企业的Intranet都是基于IP的。所有的数据流量都必须穿越 IP层。所有的数据

① <http://diffserv.lcs.mit.edu/>。

② <http://www-nrg.ee.lbl.gov/floyd/red.html>。

也都是由IP分组来承载的。

- 它可以保护和隔绝高层应用程序免遭安全性的攻击。
- 它支持现有的高层安全机制。
- 它可以支持在Internet上建立一个可扩展的、安全的VPN。

为了满足在IP层上实现安全性的需要，IETF成立了IP安全性（IPsec）工作组^①。经过努力，该工作组已经推出了一系列在IPv4和IPv6上提供网络层安全性的协议、机制和服务。

IPsec框架所支持的服务包括以下内容：

- 访问控制（access control）。
- 数据源鉴定（data origin authentication）。验证每一个IP分组的发送者的身份。
- 重播保护（replay protection）。防止攻击者偷听到某些分组并在以后某个时间回放。
- 数据完整性（data integrity）。检查以确保IP分组在传输过程中没有被篡改。
- 数据保密性和数据加密（data confidentiality & encryption）。通过加密将分组的一部分隐藏起来。

- 限定的业务流管理（limited traffic flow management）。隐藏原始发送者的IP地址。
- 密钥管理（key management）。

IPsec框架最初定义的协议包括鉴权头（AH）、封装安全净荷（ESP）及密钥管理。

3.4.1 IPsec安全关联

有许多信息是采用IPsec协议的通信实体所必须了解的。每一个通信实体都必须要知道另一个实体的网络地址。每个实体都必须对在网络中以安全方式进行通信时所采用的各种安全机制，如加密算法、共享密钥及数字签名等具有相同的解释和使用方法。

一个IPsec安全关联（SA）是两个或多个实体之间的一个单向的逻辑连接。这些通信实体定义了一个逻辑信道，并对应用实体在该逻辑信道上所传送的数据流上所采用的安全机制互相关达成了一致。IPsec SA为IPsec之间的数据交换提供了一条安全的通信信道。

在通信实体之间建立IPsec SA可以采用以下的两种模式之一：隧道（tunnel）和传输（transport）。隧道模式的SA可以在两个中间设备（如路由器和防火墙等）之间建立安全的信道。例如，在两个路由器之间可以建立一个隧道模式的SA。隧道模式的SA把原始的IP分组封装在一个新的IP分组中，其中的源和目的地址是隧道端点。安全服务只提供给流经这两个路由器之间的分组。

传输模式的SA在两个主机之间建立安全的信道，并可以在端到端的基础上提供安全服务。当然，也有可能在一个隧道模式的SA内嵌套一个传输模式的SA，以便在流经不同实体的分组上，增强整体的安全性。

一个特定的SA根据其目的IP地址、安全协议（AH或ESP）及安全参数索引（SPI）的值来标识。这些值装载在每一个IP包中，使得接收端的系统能够选择用于处理这些分组所使用的SA。

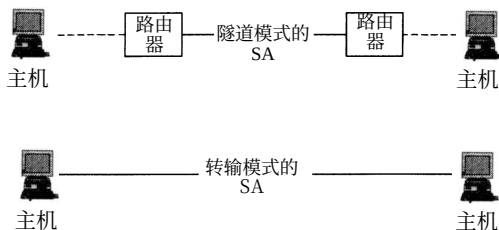


图3-9 IPsec安全机制

^① <http://www.ietf.org/html.charters/ipsec-charter.html>。

图3-9分别给出了隧道模式和传输模式的 SA的例子。

3.4.2 鉴权头和封装安全净荷

最初的IPsec协议族由AH和ESP协议构成。AH支持数据源鉴别、数据一致性，还有一个回放（replay）保护的选项。除了分组头中那些经常更改字段以外，整个IP分组的一致性都将得到验证。其中经常更改字段包括TTL和分组头的校验和，它们在每一跳都会被改变。IPsec AH不提供数据加密，数据净荷仍然以明文方式传送。此外，IPsec AH同时提供了对隧道模式和传输模式SA的支持。

IPsec ESP提供了更高级别的安全性，包括数据源鉴别、数据一致性、数据保密性（数据加密）及回放保护选项。ESP通过对数据加密实现对整个净荷的保护。如果建立了一个隧道模式的ESP AS，整个的原始IP分组包括全部分组头的信息都被加密，可以隐藏原始数据发送者和接收者的身份。

在隧道模式和传输模式下AH和ESP协议头在IP分组中的布置如图3-10所示。

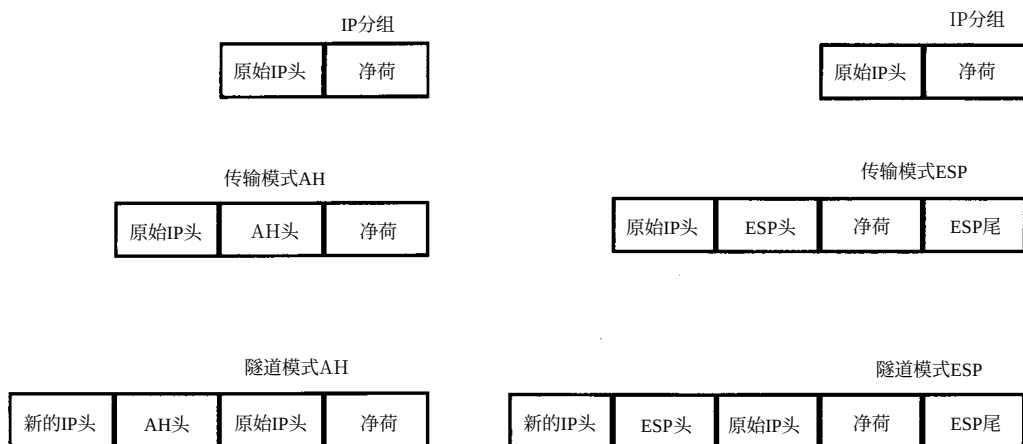


图3-10 IPsec AH和 ESP协议

3.4.3 密钥管理

要实现密码技术的基本操作，都要求相互通信的实体共享相同的密钥。发送者使用密钥进行数据的加密，接收者用它来进行解密。如果密钥丢失或被窃就会危及数据的安全，这时必须立即提供一个新的密钥。想要与其他实体建立ESP SA的IPsec实体必须被分配给一个共享的密钥。密钥可以是手工配置的，但是随着SA数量的增加，手工配置会对扩展性造成严重的制约。

IPsec工作组定义了一个Internet安全体系结构密钥管理协议（ISAKMP），以满足用户对密钥管理的需求。ISAKMP实际上并不是一个协议，而是一个对在动态网络环境中安全地建立SA过程进行管理的框架。ISAKMP（或ISAKMP/Oakley）定义了一个密钥生成、自动密钥刷新和密钥发布的框架。它独立于任何特定的加密算法，可以在非安全的网络链路上运行并建立SA。

3.5 RTP/RTCP

目前，在IP网络上的流式音频和视频（如 RealAudio，InternetPhone等）业务已经十分流

行。实时传输协议（RTP）是一个新的应用层的帧协议（不是一个真正的传输协议），它可以使实时应用程序在IP网络上以流的方式传送数据。多媒体流首先被编码（并压缩），然后被封装到RTP分组中，最后采用UDP/IP通过一个单点或组播会话发送出去。定时、排序及净荷识别等信息都从RTP发送者传递到接收者，这使RTP的接收者在其接收到最初几帧的时候就可以开始播放多媒体的数据流了。RTP并没有建立连接，也不能保证数据按时到达目的端。RTP是一个支持所有图像和声音编码的标准（RFC1890），并且能够支持H.323视频会议应用。实时传输控制协议（RTCP）是也一个受到好评的控制协议，它被用于向RTP会话的成员提供返回和统计信息。

3.6 IPv6

IPv6是下一代的IP版本，在它提出的众多改进中，最重要的是它将地址空间扩展到128位，即原先的4倍。扩展IPv6的地址空间主要是为了解决在90年代初出现的IPv4地址空间被迅速耗尽的问题。根据当时的估计，整个32位的地址将在10年或更短的时间内被耗尽。虽然这种情况还未发生，但是IETF利用这个机会提出了一个新的IP地址的版本来克服现有的、已察觉到的IPv4的缺点。其中的一些（不是全部）缺点和需求如下：

- 现有的IPv4的地址空间太小。更大的地址空间是必须的。
- 路由表大小。Internet中网络的数量和其对应的路由表正在迅速扩大。一个大的地址空间将能给层次化结构的建立以及采用CIDR思想把众多低级实体汇聚成一个高级实体提供更大的灵活性。
- 简单性。IPv4中的许多内容都很复杂，至少看起来很复杂，例如分段（fragmentation）、选项处理和配置等。
- 性能。大型的路由表、分段和选项处理等都会造成路由器中的延迟和时延。
- 配置的简易性。IPv4的设备必须配置一个合法的IP地址或者使用动态主机控制协议（DHCP）从服务器上获得一个合法的IP地址。
- 安全性。
- 可扩展性。即在不需要对现有协议作大修改的前提下，无缝增加新功能和特性的能力。
- 与IPv4共存并能转换到IPv4。一种理想的方案是能保证IPv4的设备和应用在较长时间内不被淘汰。任何IP的新版本都必须能够与IPv4共存和协同工作。

从这些需求出发，一些十分睿智且具有网络经验的工程师通过努力，推出了所谓的IPv6。IPv6是IP的一个新版本，其整体的操作与IPv4并没有根本上的不同。例如，所有的网络都需要被分配一个网络地址或地址前缀；IPv6仍然按照传统的无连接逐级跳的方式转发分组；IPv6路由器也需要运行路由协议；所有的网络设备也都需要配置一个合法的IPv6地址等等。或许，IPv6可以被看成是一个更简单、更易于扩展和更高效的IP版本。

IPv6的主要功能和特点如下：

- 扩展的地址空间（扩展到128位）。
- 增强的选路功能。IPv6中有许多特性和功能能够提高网络性能和改善选路控制。更大的地址空间使我们在建立地址层次的时候具有更大的灵活性。使用选路扩展头（routing extension header）可以使源选路更高效。在IPv6头中增加的一个流标签字段可以使一个流中的分组被区分出来，以便进行特殊的处理。组播地址空间也被扩展，并且包括了一个范围

(scope) 字段用来有效地限制一次组播发送的范围。分段 (fragmentation) 功能在路由器中被禁止；选项的处理也更直接，在许多情况下不会给中间路由器造成负担。IPv6中还定义了任意传播 (anycast) 地址，使用这种地址，一个分组可以被寻址到若干可能的接口中“最近”的一个接口上。

- 简化的头格式。IPv4的分组头有10个字段加上两个地址域及若干可能的选项。而IPv6的分组头只有6个字段加上两个地址域及若干可能的扩展头。

- 改善了对选项的支持。选项在IPv6中称为扩展分组头；其编码也有改变，这将会提供更好的性能并为今后引入新的扩展提供更大的灵活性。

- 流标签 (flow label)。源端可以对要求某个中间路由器进行特殊服务或处理的特定流的分组进行标记。这将可以简化在每个路由器上所进行的分组分类的处理过程^①。

- 转换机制 (transition mechanism)。在IPv6中定义了一系列机制，使其能够与IPv4的主机和路由器共存和交互。

版本	业务流类型	流标签	
净荷长度		下一分组头	跳数限制
源IP地址			
目的IP地址			
扩展头			

图3-11 IPv6头格式

IPv6的分组头有40字节长。其格式如图3-11，讨论如下：

- 版本。版本号是6。
- 业务流类型 (traffic class)。也就是IPv6分组中的DiffServ DS字节。
- 流标签 (flow label)。这是一个由源端分配的非零值，用来鉴别需要特殊服务或处理 (如QoS或非缺省特定服务等) 的特定流中的分组。流标签的长度为20比特。
- 净荷长度。是一个16比特的字段，它定义了 IPv6 分组头后面的扩展头加上净荷的字节长度。
- 下一分组头 (next header)。它定义了紧跟在后面的分组头的类型，它可能是一个扩展分组头、传输层分组头 (如TCP) 或者甚至是数据净荷。该字段的长度为1字节。
- 跳数限制 (hop limit)。一个8比特的值，其功能与IPv4中的TTL字节相同。
- 源地址。128比特的IPv6源地址。
- 目的地址。128比特的IPv6目的地址。

在40字节的基本IPv6分组头之后与实际的净荷以前，可能有一个或多个扩展分组头。就像IPv4中的选项，扩展分组头提供了一种方法来装载有关分组所需特殊处理的信息。大部分的IPv6扩展分组头只需要由目的端来处理，而不需中间路由器处理，因而提高了性能。扩展分组头中的内容由下一分组头 (next header) 字段来标识，该字段在基本IPv6分组头和任何扩展分组头中都存在。这个概念在图3-12中说明；而目前定义的IPv6扩展分组头的说明如表3-2。

^① 通常对IPv6的一个误解是认为它能够支持QoS。事实上IPv6所能支持的QoS是与IPv4相同的。IPv6仍然需要IntServ或DiffServ机制来进行服务区分 (包括QoS)。流标签和流类型 (traffic class) 字段也只是提供了一种为特殊处理而进行快速分组识别的机制。

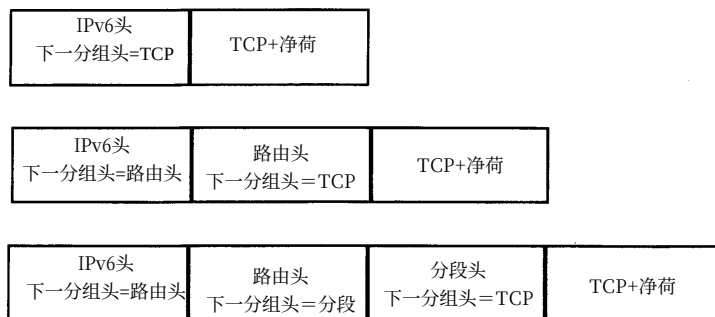


图3-12 IPv6的扩展头

表3-2 IPv6的扩展头

扩展头	描 述
逐级跳	必须被通路上的每一个路由器检查。例如，IPv6分组的巨型净荷 (Jumbo Payload, 长度超过 65535 字节) 以及路由器警告等，它们会指令路由器进一步地检查 IPv6 的净荷 (如 RSVP 消息等)
选路	定义源路由
分段	由源主机插入，并允许传递大于链路 MTU 的分组
鉴权	保证源地址是可信的，并保证分组没有被替换
封装安全性净荷	在源端和接收端之间保证机密性
目的端	仅由目的主机处理

更多的信息

Ferguson, P., and Huston, G., Quality of Service, Wiley Computer Publishing, 1998.

<http://playground.sun.com/pub/ipng/html/ipng-main.html>.

<http://www.ietf.org>.

<http://www.ipmulticast.com>.

Huitema, C., IPv6 The New Internet Protocol, 2d ed., Englewood Cliffs, NJ: Prentice Hall, 1998.

Huitema, C., Routing in the Internet, Englewood Cliffs, NJ: Prentice Hall, 1995.

Keshav, S., An Engineering Approach to Computer Networking, Addison-Wesley, 1997.

Maufer, Thomas A., Deploying IP Multicast in the Enterprise, Englewood Cliffs, NJ: Prentice-Hall, 1998.

Stevens, W. Richard, TCP/IP Illustrated Vol. 1 The Protocols, Addison-Wesley, 1994.

Thomas, Stephen A., IPng and the TCP/IP Protocols, John Wiley & Sons, 1996.

China-pub.com

下载

第4章 交换的概念和LAN交换技术

前面几个章节给出IP的一个总体框架，特别是详细介绍了构成TCP/IP的体系结构和协议。为了更深入地进行关于IP交换技术及其基础的讨论，我们首先讨论交换转发（switch forwarding）和交换通路控制（switch path control）。交换转发技术利用了保存在分组中及交换机中的信息来将分组通过交换机转发出去。交换通路控制用于在交换机构成的网络中建立和维护交换通路。

在数据网中，交换技术的应用可以有多种不同的形式，用交换技术可以定义一种在不同LAN网段之间快速传递数据包的技术，也可以用交换技术实现一个端到端的交换体系结构，例如ATM。本章将回顾局域网交换的基本要素。下一章将会讨论帧中继和ATM，我们希望这两章的内容能为本书中后面章节中所讨论的IP交换方案奠定基础。

4.1 交换的概念

在任何一个网络和应用技术中的术语后面加上“交换”这个词，好像就会得到一种极具吸引力的、可带来高性能、低费用、低复杂性的新技术。这种例子有很多，如以太网变成交换式以太网、令牌环变成交换式令牌环交换、IP变成IP交换等等。很多网络新技术已经能比我们现在所使用的网络设施速度快几个数量级了，把这些新技术与“交换”术语结合起来，当然会使我们听起来感觉更好（例如将千兆位以太网变成交换式千兆位以太网）。但是，为了赶时髦，也使一些与交换技术明显毫不相关的或至少是表面上无关的东西，也纷纷与“交换”拼凑在一起，形成了一些新的、千奇百怪的网络技术，例如应用层交换等。

什么是交换，为什么交换如此流行呢？第二个问题的答案很简单，就是价格和性能上的原因。交换技术，特别是与局域网结合的交换技术，可以用比传统共享的LAN技术低得多的价格来提供比传统技术高得多的性能。在现在和可预见的未来，半导体技术的进步可以把更多的网络处理集成到一个廉价的芯片中，这可以使价格不断下降，同时使性能比老式的、基于软件的处理方法提高几个数量级。因此交换广泛流行的根源就是其高速和廉价的事实。

那么交换到底是什么呢？在这个数据互连的世界上，交换可以有多种形式。LAN交换可用于在位于相同或不同网段上的工作站之间传递分组。广域网（WAN）交换通常采用的形式是在两个终端节点（如一对路由器）之间提供一条虚连接。不管其所处的环境如何，所有的交换形式都有如下共同的基本属性：

- 工作于任何协议栈的第二层和二层以下。这意味着LAN或WAN的交换对于运行在高层的IP或其他网络层协议及任何应用程序都是透明的。图4-1说明了两个IP终端节点通过一条2层交换

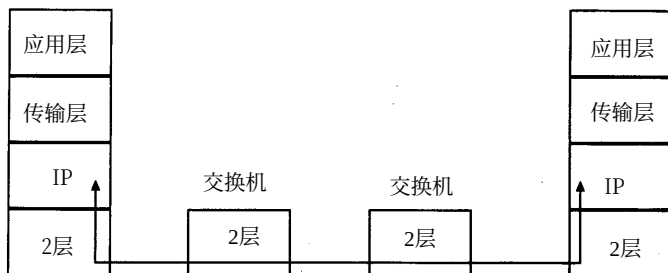


图4-1 协议栈中的2层交换

通路相互通信的情况。

- 交换由硬件实现。交换机把从一个输入端口上接收到的数据分组交换发送到一个输出端口，这个工作过程不受其他处理器的干预。将此功能用硬件（或固件）来实现通常可以带来更高的性能及更低的延迟、复杂度和费用。

交换的过程可以从两个不同的方面来观察：本地和端到端。从本地的出发点来看，我们需要研究的问题是，一个交换机是怎样根据接收到的数据分组来决定将其交换到哪一个输出端口上去。换句话说，就是数据分组及交换机内部有哪些信息可以用来使交换机能快速地将数据分组从输入端移动到输出端。这个问题称为交换转发问题。交换过程的第二个方面的问题是如何在交换网络中的两个端点之间建立和维护交换通路。这方面技术称为交换通路控制。

4.1.1 交换转发技术

从表面看，一个交换机很简单，只是一个具有一定数量端口的机箱，交换机通过这些端口可与多种不同的设备如工作站、路由器或其他交换机等连接。当某个分组到达一个输入端口时，交换机的任务就是决定要么丢弃这个分组，要么将其移动到正确的输出端口上进行发送。正确的输出端口是根据分组中所携带的信息来确定，在某些情况下可能根据交换机本身内部的信息决定。包含在分组中的信息由下列之一构成：

- 目的地址。
- 源路由向量（source-route vector）。
- 连接标识符。

在第一种情况下分组包含一个目的地址字段，该字段的意义很明确，就是分组所要到达的目的终端（主机）的地址。通过采用下面将讨论的各种技术中的一种，交换机建立一个地址和端口的对应表，对应表由目的地址入口和与该地址相关的输出端口构成。当分组到达输入端口时，交换机将检查分组中的目的地址，并用其在地址端口对应表的表项中寻找匹配，然后将分组发送到其正确的路由上。当然，如果寻找不到匹配的入口，交换机就可能将该分组丢弃。或者交换机会将其发送到所有的输出端口上，以希望分组能通过其中的一个端口到达目的端。如果在地址端口对应表中，分组中的目的地址与该分组的输入端口匹配，那么该分组将被丢弃。这种思想基于一个符合常理的假设，即交换机不应该把来自某端口的分组再次转发到该端口上。

另一个需要注意的是分组总是以其固有的格式到达交换机，这就隐含地说明了发送源可以只关心发送分组，剩下的工作都由交换机来完成。这种技术通常使用在以太网LAN交换中，如图4-2所示。具有目的地址“aaa”的分组到达交换机，交换机查找地址端口对应表后，将该分组转发到端口4。

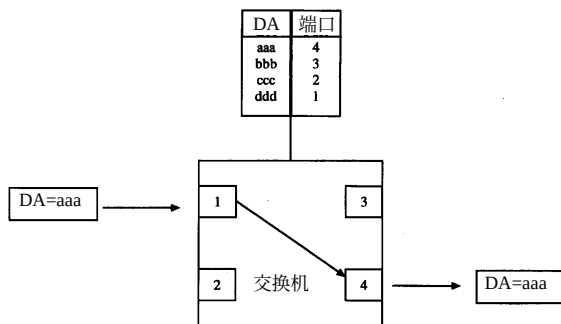


图4-2 目的地址交换

另一项技术利用了包含在分组中的源路由向量（source-route vector）。源路由向量是一个

网络中拓扑元素的序列，分组必须按照该序列给出的通路才能到达目的地。拓扑元素的范围包括通路上交换机的输出端口号，通路上每个交换机的地址或是在令牌环源选路的情况下的一个或多个环编号桥接器及环编号的组合。需要再次说明，采用这种技术时，一个源路由向量被插入到分组中，交换机就利用这个信息将分组转发到适当的输出端口上。应该注意，交换机在本质上不需要建立和维护像我们在前面所给的例子中那样的一个额外的交换表。在源路由向量中已经包含了通往目的地通路的所有信息，交换机只需要处理该向量就行了。由于其在网络方面做了一定工作，并且由于发送者向外发布了准确的构造源路由向量的必须信息，因此很显然，在通过网络交换分组时源选路方法是十分高效的。源选路常被用在令牌环网中，也被ATM PNNI阶段I用于为SVC的设置请求进行选路。图4-3给出了一个用输出端口号作为向量元素的源选路的例子。

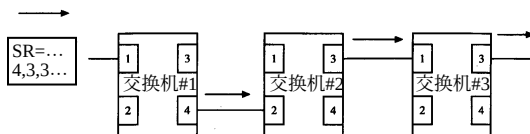


图4-3 源选路

第三种技术利用了装载在分组中的一个连接标识符来确定输出端口。连接标识符也可以称为标签（label）。这种技术与前两种技术有所不同，因为在这种情况下，分组头中的标签在经过交换机的时候是需要被交换的。使用我们下面所讨论的技术，一个标签被添加到了分组上，交换机需建立一个由输入端口号、输入标签、输出端口号和输出标签构成的连接表。当一个分组到达交换机时，其输入端口号和分组标签就被用于在连接表中搜索匹配的表项。然后，分组中的标签就被正确地交换，分组也被转发到正确的输出端口上。在大多数情况下这一过程能够用硬件快速完成，因为交换机的工作仅仅是从分组中提取一个固定长度的标签并将其与连接表中的表项相比较而已。标签交换现已更广泛地为人们所熟知，也被ATM、帧中继以及一些将在本书后面的章节中所讨论的IP交换协议用来转发数据包。

我们还应该注意到，从一个单个交换机的角度来看，一个特定的标签值是仅限于该交换机本地内部的，更明确地说，它被限定于交换机自身的端口（输入或输出）上。这一点我们可以从标签被交换（重写）为新值的事实中推断出来。但是，从完整的角度来说，新标签值也可能与旧标签值相等。但即使在这种情况下，标签仍然只在本地有意义^①。图4-4给出了一个标签交换的例子。

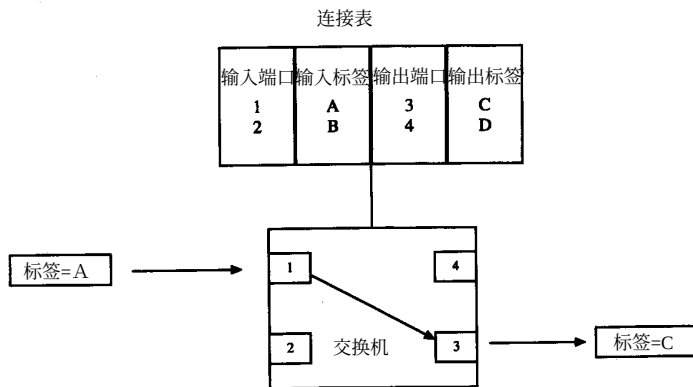


图4-4 标签交换

4.1.2 交换通路控制

前一节所讨论的交换转发技术全部都依赖于某种形式的交换通路控制来在交换机中建立

① 本地意义（Local significance）可能是一个容易引起误解的术语。实际上，一对交换机必须对特定标签值保持一致的用法。换句话说，上游交换机的输出端口和下游交换机的输入端口将会使用相同的标签值。

和维护信息，以便使分组能转发到目的地。交换通路控制可能会引起网络中交换机之间的控制消息的交换，从而使各交换机能建立一个一致的交换转发表。它也可能导致发送数据的源主机产生一条“通路设置”消息并将其发送到网络中。通路设置消息用来在网络中发现一条通路，以使后续的数据分组能按这个通路流动。

交换通路控制可以采用下列技术之一来实现：

- 地址学习 (address learning)。
- 生成树 (spanning tree)。
- 广播和发现 (broadcast and discover)。
- 链路状态选路 (link state routing)。
- 显式信令 (explicit signaling)。

地址学习可能是一种最简单的技术，交换机可用它产生和创建一个交换转发表。在某一个特殊的端口，交换机不加区别地监听所有的分组，并将分组的源地址与端口号一起保存在一个地址端口对应表高速缓存器中。同时，交换机也将检测分组的目的地址，并试图在地址 - 端口缓存中寻找其匹配的入口。如果成功地找到匹配，就将该分组交换到其对应的端口。如果没有找到匹配，就将该分组发送到除了其输入端口以外的所有输出端口上。如果目的地址对应的输出端口就是该分组的输入端口，交换机将丢弃该分组。经过一段时间，交换机就会使缓存器中的表项超时，以便使高速缓存中的信息不会因为过时而变得无用。

地址学习通常被用在以太网透明桥接中建立地址端口对应表缓存器（见图4-5）。可以看到，交换机1已经通过学习知道工作站aaa和bbb位于端口1上，而工作站ccc和ddd位于端口3上。

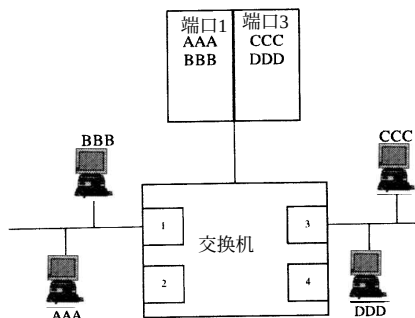


图4-5 地址学习

地址学习十分易于实现，并且在互相通信的主机上不需要作任何修改或处理过程。但是，它需要交换机能在其缓存器中存储一个可能是十分庞大的地址表。地址学习的另一个主要缺点是：如果源主机与目的主机之间有一条以上的通路，就可能产生环路（loop）。环路的形成是这样的，即一个分组从源网络 LAN 1 到目的网络 LAN 2 的转发过程中需要经过一个或多个桥接器，如果分组通过一条通路 1 被发出后又从另一条通路 2 被返回，就说明网络中形成了环路。在这种情况下，通路 1 上的桥接器认为分组的源在 LAN 1 上。而通路 2 相反方向上的桥接器认为分组的源在目的网络 LAN 2 上。由于没有办法来解决这个明显的冲突，分组将会在网络上无限循环下去。

为了避免环路问题，交换机可以建立一个生成树（spanning tree）。生成树是一个链路和交换机的子集，它能被用于产生连接到交换网络上的任何两个站点或网段之间的无环路的通路。交换机通过相互交换控制消息来建立和维护生成树，在控制消息中包含了根交换机的标识符和本交换机到根交换机的距离。只有那些回到根交换机具有最小代价（cost）的交换机和交换机上的端口才会出现在这棵树上。所有其他的交换机和端口都被修剪掉了，这使得交换机和交换机的端口可以以无环路的方式转发分组。换句话说，就是某交换机上的一些端口处

于“转发”状态，而其他的交换机或该交换机上的其他端口处于“阻塞”状态。处于转发状态的端口和链路就组成了生成树。图 4-6 给出了一个互连 LAN 集合的生成树。

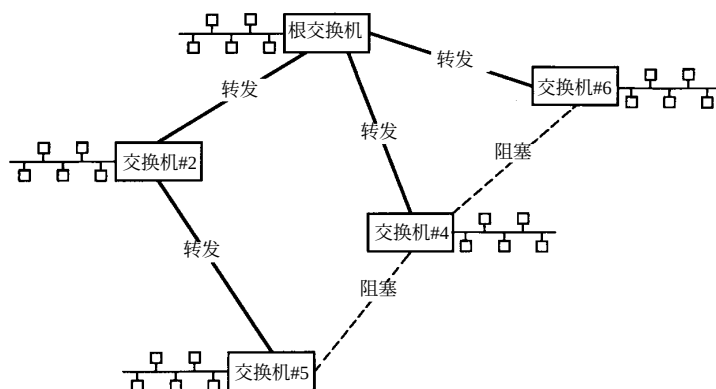


图4-6 生成树

生成树算法在IEEE 802.1d LAN桥接标准中定义。它被用于以太网透明桥接器和令牌环源选路桥接器中。在采用桥接器或交换机互连来对 LAN 环境进行扩展时，生成树是十分有用的。但是生成树在可扩展性和健壮性方面有一定的限制。首先，依据拓扑结构，数据流量可能会集中在网络中少数的几个链路和交换机上。即使是在活动链路和交换机上存在可用的替换通路，生成树也将阻塞数据流而不让其通过替换通路。这意味着生成树只利用了可用网络拓扑（链路和交换机）的一个子集来转发数据流。第二，在处理大量生成树控制消息时的开销很大，特别是在网络拓扑结构很大并且不稳定的时候。

广播和发现是另一种在 LAN 交换（或桥接）中常用的技术，用来在网络中确定交换通路。这里的想法是通过向目的端发送一条广播探测分组（explorer packet）来探测一条或多条可能的通路。目的端接收到探测分组后，会利用一条或多条可能的通路把探测分组返回给源端。在返回的分组中包含了一个源路由向量，发送者可以利用它来转发所有的后续分组到目的端。

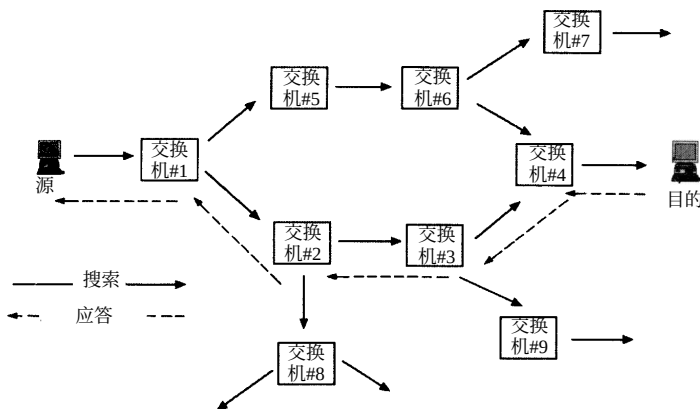


图4-7 广播和发现

利用广播和发现技术在网络中建立一条交换通路的概念如图 4-7 所示。可以注意到，探测分组将传播到整个网络中。这被称为全路由探测（ARE）。如果已存在生成树，那么探测分组将选择沿着生成树的方向到达目的地。这被称为生成树探测（STE）。目的端收到探测分组后，

可能会使用广播分组向发送端返回一个分组，它也可以像图中那样选择一条源路由。在这种情况下，目的端分别按顺序经过交换机 4、3、2、1 发送一个分组返回给发送源。发送者就将利用这个通路转发分组到目的端。如果使用了 ARE 帧，广播和发现技术就可以提供在源 LAN 到目的 LAN 的多条通路之间的统计负载平衡。由于源选路被用作了一种交换转发的机制，在交换部分的开销极小；但另一方面，探测分组也消耗了带宽资源，在网络初起阶段可能导致网络拥塞。

第四种技术假定交换机运行了一个链路状态（或拓扑状态）路由协议。交换机同与之相邻的交换机交换链路状态信息，并利用这些信息建立一个统一网络链路状态数据库，该数据库由所有的活动链路和节点构成。由于每个交换机都知道网络中哪些链路和节点是活动的，分组就能按照逐级跳（基于目的交换）的方式或使用源选路算法被转发到目的地。在后一种情况下，当入口交换机接收到一个源端希望与目的端通信的指示时，它可以参考其本地的链路状态数据库，组装一个源路由向量，将其加入某个分组中并发送出去。应注意，反映在本地链路状态数据库中的整个拓扑结构在每次组装源路由向量的时候（只会发生在交换链路的建立阶段）都会被检验。

链路状态选路被类似 OSPF 和中间系统到中间系统（IS-IS）这样的 IP 路由协议使用来建造和维护路由器中的转发表。PNNI 阶段 I 则利用链路状态选路在 ATM 交换机网络中转发 ATM SVC 的设置请求。

最后一种建立交换通路的方法是利用显式的信令。在任何数据分组开始流过一个专用的交换通路之前，源端必须首先通过网络向目的端发送一条连接建立请求消息。这条消息可以确保存在一条通往目的端的活动通路，并且可能会检验在该通路上是否有足够的容量来应付数据流。当源端收到来自目的端的一个应答并确认存在一条有足够容量的通路时，数据才能开始传送。显式信令的概念在图 4-8 中说明。

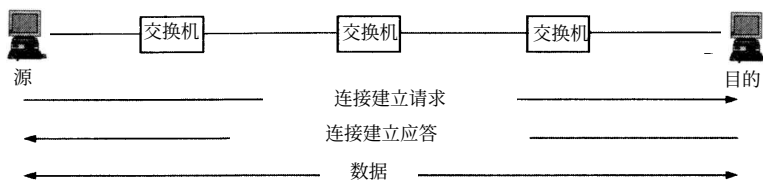


图4-8 显式信令

利用信令来建立交换通路有其优点也有其不足之处。在好的方面，信令可以在数据传输前确认实际上确实存在一条到达目的端的可用通路。这可以取消这样一个过程，即源端发送连续的分组，然后等待应答，当没有收到应答时，它才确定目的端是不可达的。采用信令的另一个优点是路由的计算只在连接建立的时候进行一次，后续的数据分组都将只需沿着相同的专用交换通路到达目的端。一旦数据传送开始，就不再需要路由的计算。从另一个角度看，就是为一系列分组而进行的通路计算开销被平均分摊到了通信实体的整个通信生存期中。

信令之所以吸引人的第三个原因是一些反映连接特性的参数如最小带宽和最大时延等可以传递到网络中，并被通路上的每一个交换机所处理。理想的结果是通往目的端的交换通路能够满足通信端点间数据流的特定带宽和时延特性。最后，显式的信令可以成为一种很有用的方法来在每个连接的基础上对网络的使用进行统计和计费。

信令的缺点主要是其增加了时延以及它给交换系统带来的复杂性。它所增加的额外时延

是连接建立延迟，也就是其进行初始化和完成连接设置过程的时间。信令增加了系统复杂性是因为，除了要处理数据以外，网络还必须处理信令请求。对信令请求的处理导致了其他控制组件的出现，如接纳控制等，这些组件必须以某种形式加入到连接设置的过程中。一个系统所能可靠应付的处理数量是有限的。在一个依赖显式信令的交换网络中，能够支持的最大交换通路的数量，以及能够同时处理的连接建立过程的数量都是有限的。

显式信令通常被用在那些所谓的硬状态的、面向连接的交换方案中，如帧中继和 ATM。它也可被用于一些软状态解决方案（如 RSVP）以及将在本书后面所要讨论的新 IP 交换协议家族中。硬状态和软状态的区别是前者在连接设置过程完成前不允许任何数据传送，而后者（软状态）则允许数据传输，但必须使用一条不同的通路。硬状态的连接一般不会终止，除非有终止请求或网络中出现某种不可预测的故障。软状态的连接将会在一定的周期以后超时，除非被新的连接建立消息所刷新。软状态的连接被认为是十分健壮的，因为即使是在交换通路建立之前数据也可以进行传送。

4.2 LAN 交换

LAN 交换是对传统 LAN 技术的一次本质的革命。自从人们第一次把同一地点的工作站联网的时候起，终端用户就已经很快发现互连共享文件和其他网络资源（如打印机）的方便性。在以太网环境中，终端通过挂在同一条电缆上相互连接起来。这条电缆作为各工作站的一个共同的传输介质被所有连接在上面的工作站所共享。但是由于它是共享的介质，在一个时刻只允许有一个工作站发送数据。一些协议被用来保证在任何时候只有一个工作站能够获得共享总线的控制权，如带有冲突检测的载波监听多路访问（CSMA/CD）及令牌传递等。图 4-9 描述了一个共享介质的 LAN。在连接域中的三个工作站共享一个共同的传输介质，但同时只能有一个工作站传输数据，其他工作站必须等待。

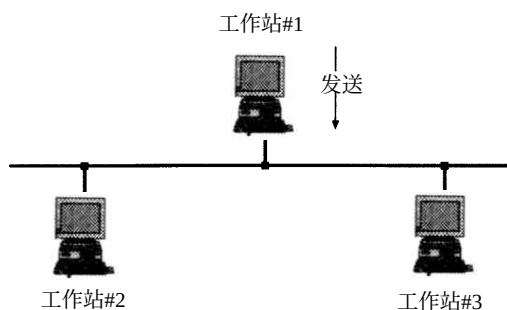


图4-9 共享LAN

尽管这些协议至今没有发生什么变化，但是在物理结构上，由工作站挂接到单一的固定长度的电缆上的 LAN 物理基础结构已经逐渐被星形连接的集线器（hub）所替代。共享的 LAN 集线器的优点在于其价格低、介质的灵活性和利于网络管理。这种简单的配置及结合高速工作站、更好的网络管理工具、客户服务器应用的增长以及向多协议网络环境发展的趋势等因素，共同使局域网互连技术得到了极大的发展。

但是在某些地方，同一个 LAN 上的多个工作站共享连接资源的形式越来越成为一个问题。当一个客户机正连接到一个文件服务器进行文件传输时，由于它消耗了大部分的可用带宽，其他的客户机不得不等待。这种等待会导致较低网络利用率和较低的应用程序吞吐率，特别是在业务量和终端用户大量增加的时候情况会更加恶化。因而在工作组 LAN 中必须要有更好的办法来解决这个问题。

最好的解决方法是将 LAN 分段，减少每个网段上的工作站的数量，最好是每个网段只有一个工作站。工作站因此可以访问全部的支持带宽（如在快速以太网中为 100Mbps）进行数

据传输，而不需要等待其他客户的作业完成。将通信能力由半双工扩展成全双工可以使工作站能使用全部的可用带宽同时发送和接收数据。专用全双工通信的概念对服务器是特别有帮助的，因为服务器必须试图同时为多个客户提供服务。将一个LAN分段成每网段只有一个工作站的概念被称为微分段（microsegmentation），在图4-10中对其进行了说明。它将每一个工作站连接到LAN交换机的一个端口上，并且给予每个端口全双工的带宽。

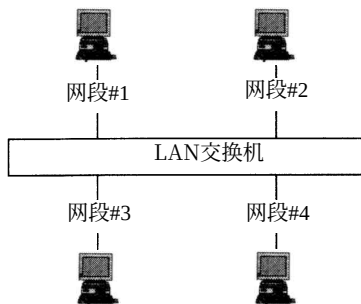


图4-10 LAN交换

微分段在工作组级别上提供了更大的带宽和更好的性能。从网络的角度来看，LAN交换在性能和容量上可以提供类似的益处。首先，LAN交换是对传统的双端口的桥接进行了多端口的扩展。因此一个交换机可以被认为是一个多端口的桥接器，只是它有更多的端口和无阻塞的结构并且它可以提供更高数量级的容量和更好的性能。其次，高速的中继可将交换机互连到一个高速的骨干网络上以实现工作组之间的通信。最后，一个由交换机构成的网络可以支持一个或多个虚拟LAN（VLAN）。每一个VLAN是一个广播域，其中的成员之间的通信就像它们共享一个网段一样，即使它们可能是通过了多个链路和交换机相互连接的。由于管理、安全和广播控制等方面的原因，VLAN是十分有用的。VLAN的内容将在4.2.3节中讨论。

4.2.1 短径转发

从一个输入端口将分组交换到一个输出端口的过程是依据分组内部自身包含的某些信息进行的。正如在4.1.1节中所讨论的，这些信息可能是目的地址、标签或者是一个源路由向量。LAN交换机通常会根据分组的目的地址来交换分组。此外，LAN交换机也可能使用两种内部的转发技术来处理分组：短径转发（Cut-through Forwarding）和存储转发（Store and Forward）。

在短径转发模式中，交换机检查和校验了目的地址后就立刻开始转发分组。对分组的第一部分的转发甚至可以在该分组的其它部分（或说分组尾）还在从输入端口的交换缓冲读出的时候进行。这种转发模式的优点是具有较低的交换延迟，这对于支持LAN上时延敏感的业务是很有好处。其缺点是短径交换把所有的分组，不管是合法的还是不合法的都转发到了输出端口上。

短径交换的概念如图4-11所示。

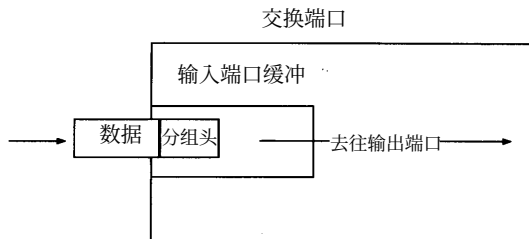


图4-11 短径转发

4.2.2 存储转发

存储转发是常规LAN交换机中所采用的另一种每个分组（per-packet）的转发机制。在这

种模式的操作中，LAN交换机在确定分组发送到哪里之前将把整个分组读入一个缓冲区中。这使交换机能检查帧长度的合法性，并进行循环冗余校验（CRC）。如果帧太长、太短或CRC出错，交换机就可以将其丢弃，而不必将其转发到下一个工作站或网段上。然而，这种交换机对整个分组进行的错误检测会增加延迟。在交换机中所花费的时间与帧长度直接成正比。

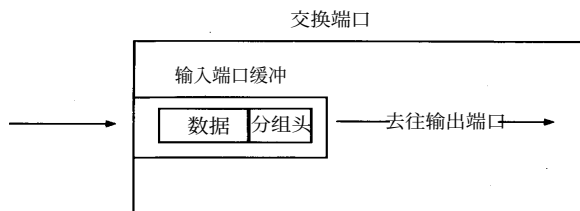


图4-12 存储和转发

存储转发技术被LAN交换机用于把分组从一种介质上转发到另一种介质上。例如，一个带有ATM上行链路的以太网LAN交换机在着手对分组进行分段前必须完成接收整个分组，然后将其分割为53字节长的信元，并将其交换到ATM上行链路上。

存储和转发的概念如图4-12所示。

4.2.3 虚拟LAN和IEEE 802.1Q

VLAN是与LAN交换技术相关的一项重要而有益的应用。从定义上说，VLAN是一个广播域，其中的成员利用LAN交换进行通信，仿佛共享同一物理网段一样。在VLAN中，划分在同一广播域中的成员并没有任何物理或地理上的限制，它们可以连接到一个交换网络中的不同交换机上。广播分组、未知分组及成员之间的数据分组都被限定在VLAN之内。对VLAN的另一个定义是，它能够使单一的交换结构被划分成多个小的广播域。

加入一个VLAN所基于的标准可以是任意数量的、不同的因素。通常这些因素包括：

- 物理端口。
- MAC地址。
- 3层单点地址。
- 组播地址。
- 若干数量的不同策略，如每天中的时间、应用及MAC地址等。

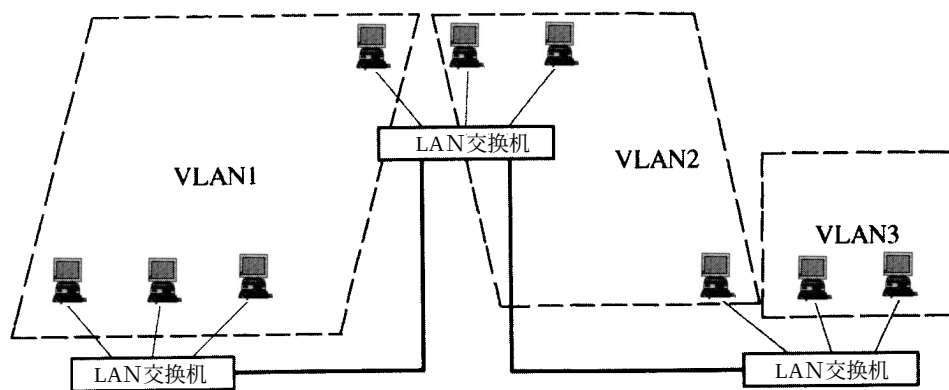


图4-13 虚拟LAN

VLAN的概念在图4-13中说明。可以看到，VLAN的成员在物理上可以连接到相同或不同的交换机上。把LAN交换用作传输的载体，使得同一个VLAN中的成员能够以实际线路速率

或以接近实际线路速率进行低延迟的通信。成员资格控制、拓扑独立性以及高速的 LAN 内部连接技术等，都为 VLAN 提供了许多优越性：

- 配置。成员的添加、移去和修改都可能对网络管理者造成操作和维护上的负担。在一个动态定义的 VLAN 中，成员资格是由一些客户参数（如地址等）来确定的，因此大大简化配置的任务。一个客户可以从一个交换机端口移动到另一个交换机端口上，但保持在相同的 VLAN 上。
- 安全性。成员在加入 VLAN 前必须提供授权，通过这种方式 VLAN 能够保护重要的资源。
- 网络效率。不同 VLAN 上的业务流在本质上是相互隔离的。这可以防止一个 VLAN 上的业务流到达另一个 VLAN，从而防止引起对工作站和网段资源可能的不必要占用。VLAN 之间的业务流必须通过一个桥接器或路由器。这也是一项很有用的技术。可以用来把一个高性能的服务器与若干具有亲缘关系的客户机组合在一起。
- 广播封闭（broadcast containment）。一个正确配置和操作的 VLAN 可以防止广播分组从一个 VLAN 泄漏到另一个 VLAN 中或把这种泄漏降低到最小程度。

除非 VLAN 的成员是物理上分离的（如划分交换机的端口），交换机必须有一些方法来识别分组并将其转发到正确的 VLAN 上。该方法的实现使用了一种标记（tag）技术，每一个分组都被加上了一个标记，用来确定该分组所属的 VLAN^①。VLAN 的标记使交换机能够将来自不同 VLAN 上的业务流复用到相同的交换网络上。在 VLAN 交换网络配置完成后，工作站发送的第一个分组将会到达一个入口交换机端口。除非成员资格的规则是基于物理端口的，否则分组的部分内容将会被检查并与一个 VLAN 配置数据库中的内容相比较。然后，一个标记或 VLAN 标识符就会被加到这个分组上。VLAN 标识符和目的地址被 VLAN 交换机用来将分组转发到相同 VLAN 上的适当目的地。同时，该信息也被 VLAN 交换机用来过滤非成员的业务流。

图4-14说明了一个 VLAN 交换实例的操作过程。交换机 1 和 2 被配置在两个 VLAN A 和 B 中，其中 VLAN A 的成员包括工作站 aaa 和 ccc，VLAN B 的成员包括工作站 kkk 和 jjj。应注意，配置信息必须发布到每一个可能具有 VLAN 成员的交换机上。现在可以假定交换机通过一个静态或动态的学习过程已经了解到哪些 VLAN 上的哪些成员分别位于哪些交换机端口上。当工作站 aaa 向工作站 ccc 发送一个分组时，它首先到达交换机 1 的端口 1。交换机在此处查询配置数据库，发现工作站 aaa 属于 VLAN A 后用一个 VLAN 标识符来标记该分组，然后将分组转发向工作站 ccc。被标记的分组沿着交换机 1 和 2 之间的链路到达交换机 2。交换机 2 注意到该分组属于 VLAN A，就将标记移除，并通过端口 3 将分组转发到工作站 ccc。

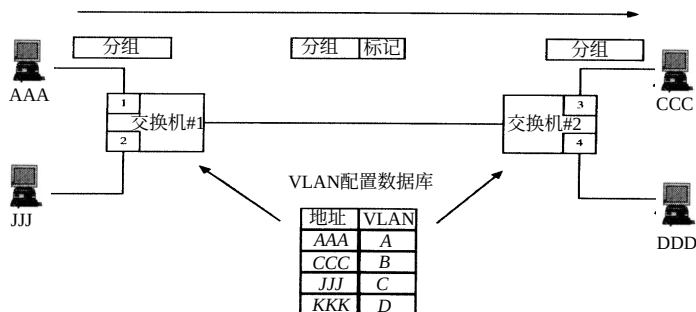


图4-14 VLAN交换

① 在VLAN中所定义的对标记（tag）的使用不应该与 Cisco 的标记交换（Tag Switching）中的标记相混淆。在 VLAN 环境中所说的标记标识了分组所属的某个 VLAN。而 Cisco 的标记交换中的标记用于提高 IP 转发的速度。

直到最近，交换机的生产厂商还在使用其自己的方法来标记分组。这导致了互操作性上的一些问题，因为不同厂商提供的交换机可能采用了不同的方法来进行分组标记。有一项技术已被一些厂商所采用，即利用 IEEE 802.1Q安全数据交换（Secure Data Exchange）作为 VLAN的标识符。IEEE 802.1Q是作为一种为MAC层设备提供授权和加密服务的方法而被开发的。尽管它可以解决这个问题，但是这毕竟不是这个标准的设计目标。人们对在这里使用该标准也有很多不赞成的意见。

因此IEEE就着手制订了一个通用 VLAN标准，经过努力，形成了虚拟桥接 LAN的IEEE 802.1Q规范。802.1Q规范中规定了下列内容的实现：

- 定义了一个体系结构，以便在现有的 IEEE 802 桥接LAN上提供VLAN服务。
- 定义了以太网/IEEE 802.3及令牌环/IEEE 802.5上承载VLAN标记的帧格式。
- 定义了VLAN识别（VLAN-aware）设备^①用来进行配置信息和成员资格信息通信的协议和机制。
- 定义了IEEE 802.1Q VLAN识别设备网络中转发帧的标准和过程。
- 保证与非VLAN识别设备（non-VLAN-aware devices）的互操作性和共存性。非VLAN识别设备就是一个工作站或路由器，它不能接收和发送带有VLAN标记或VLAN标识符（VID）的分组，也不能理解有关VLAN成员资格的信息。

标记头（tag header）的格式在以太网环境和令牌环网环境是不同的。标记头在以太网帧中的位置和格式如图4-15。标记头在原始的以太网帧上加入了4个字节，这样使以太网的最大帧长度变为1518字节。这个值大于IEEE 802.3标准中所规定的1514字节，但是目前正预期对其进行修改，以便使带有VLAN标记长度为1518字节以太网帧能够被支持。

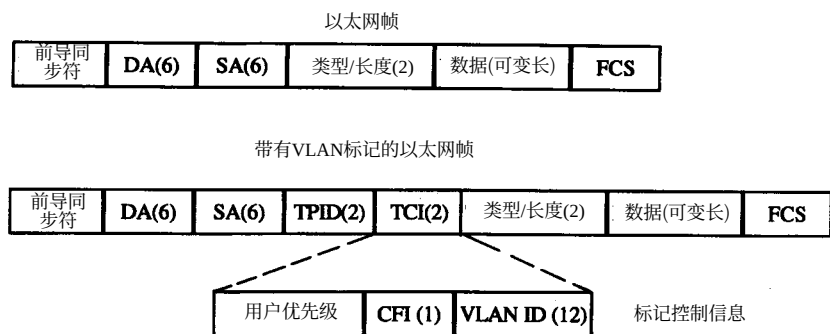


图4-15 以太网VLAN帧格式

4字节的标记头的组成如下：

- 标记协议标识符（TPID）。2字节的TPID字段的值为16进制的81-00，这个值必须区别于以太网类型字段中的任何值。
- 标记控制信息（TCI）。TCI中包含一个3比特的用户优先级字段，用来在支持 IEEE 802.1p规范的交换机进行帧转发的过程中标识帧的优先级，下面将作详细说明。TCI中还包含1比特的规范格式标识符（CFI），用于标识MAC地址信息是否是规范格式的。此外TCI中还有一个12比特长的VID，定义该帧所属的VLAN。

^① VLAN识别的设备（VLAN-aware devices）可以是一个连接到VLAN网络上的工作站、桥接器、交换机或路由器。它们可以接收、发送和解释具有VLAN标记的分组。

标记头在令牌环帧中的位置和格式如图 4-16。它与以太网帧的不同之处在于它使用了一个8字节的子网访问协议（SNAP）封装的TPID，而不是像以太网中的2字节的TPID。加上2字节的TCI，VLAN标记头将在令牌环帧的长度上加上 10 字节。

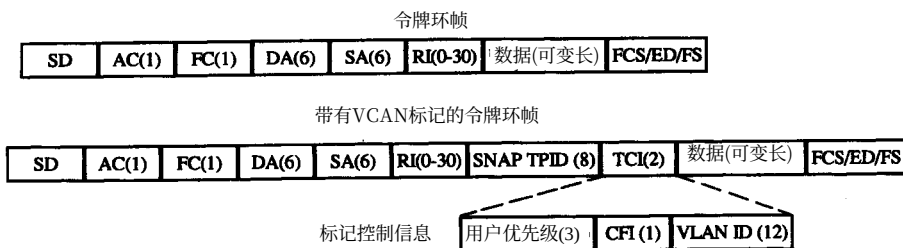


图4-16 令牌环VLAN帧格式

8字节的SNAP封装的TPID中包含的内容如下：

- 第1到第3字节装载基本的SNAP头，其16进制值为AA-AA-00。
- 第4到第6字节装载SNAP协议ID，其16进制值为00-00-00。
- 第7和第8字节装载上述的TPID，16进制值为81-00。

IEEE 802.1Q规范的操作过程在图 4-17中的网络中作了最好的说明。连接到VLAN交换机1、2和4的以太网网段上的工作站都在接入链路上（access link）。接入链路是用来将非VLAN识别（non-VLAN-aware）的工作站连接到一个VLAN交换机端口的一个LAN网段。接入链路可以是单独一个网段，也可以由非VLAN识别的桥接器和交换机连接起来的多个网段或工作站组成。接入链路不能承载标记分组。

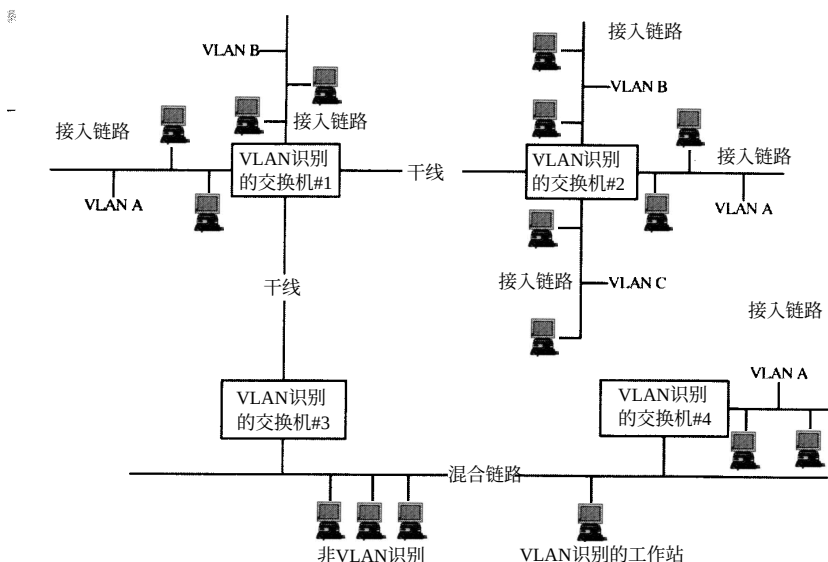


图4-17 IEEE 802.1Q VLAN网络

干线链路（trunk link）是只承载标记分组（即那些具有VID）的链路。干线链路只能支持那些能够理解VLAN帧格式和成员资格的设备。干线链路最通常的实现就是连接两个VLAN交换机的链路。但是干线链路也可以是一个由多个VLAN交换机和VLAN识别的（VLAN-aware）

工作站连接共享的LAN网段。

IEEE 802.1Q规范定义的第3种链路是混合链路（hybrid link）。混合链路可以同时传输标记的和未标记的分组并支持 VLAN 识别和非 VLAN 识别的设备。在图 4-17 中，混合链路同时连接了 VLAN 交换机 3 和 4、属于 VLAN C 成员的非 VLAN 识别的工作站以及属于 VLAN A 成员的 VLAN 识别的工作站。可以看到，混合链路既可以作为某些 VLAN 的接入链路，也可以作为另一些 VLAN 的干线链路。还应该注意的，对于连接到相同混合链路上的相同 VLAN 的成员来说，其所有分组必须被标记为一个相同的 VID 值。例如，如果 VLAN B 的非 VLAN 识别的成员连接到了如图 4-17 中的混合链路上，它们将不能与 VLAN B 中的其他 VLAN 识别的工作站通信。

VLAN 的可靠性和可扩展性操作要求在网络中的所有 VLAN 识别的设备上维护一个一致的成员关系。显然，小型的环境中在每个交换机（per-switch）基础上管理这些信息是可能的。然而，在一个具有大量 VLAN 识别的设备和大量 VLAN 以及具有动态 VLAN 成员资格（需要频繁地加入和离开）的网络环境中，某些形式的动态 VLAN 成员资格机制就是必须的了。IEEE 802.1Q 定义了通用属性注册协议（GARP）VLAN 注册协议（GVRP）。GVRP 被 VLAN 识别的设备（如工作站和交换机等）用于在 VLAN 识别的设备的网络中动态地注册和分配 VLAN 的成员资格信息。这可以使一批交换机能够建立和更新有关哪些 VLAN 是活动的、通过哪些端口可以到达等的信息。

GVRP 是利用了 GARP 的一个应用。GARP 是一个接收者驱动的协议，用于在具有 GARP 能力的设备构成的网络中注册属性。属性的组成包括组播地址，在 VLAN 环境下还包括 VLAN 的成员资格信息等。GARP 的概念（及其 GVRP 扩展）在图 4-18 中说明。在这个例子中，VLAN 交换机 4 检测到 VLAN A 的成员连接到了它的一个端口上。VLAN 交换机 4 因此将产生一个 GVRP 消息并传递给其相邻的 VLAN 交换机。接收到 GVRP 消息时，每个交换机将更新其 VLAN 转发表，把接收该消息的输出交换机端口与 VLAN A 的 VID 联系起来。

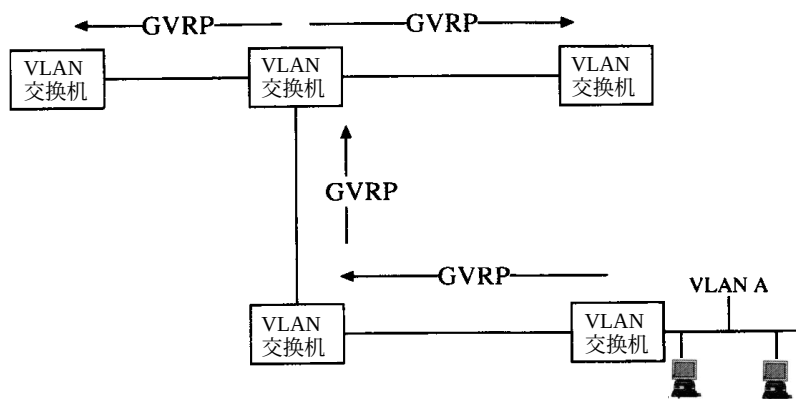


图4-18 使用GVRP的动态VLAN法册

4.2.4 IEEE 802.1p

与 IEEE 802.1Q 规范的应用紧密结合的是一项正在进行的工作，即在 IEEE 802 LAN 中为实时业务流和动态组播过滤定义一种规范。这些内容在 IEEE 802.1p 规范中得到了定义，并被

融合进了为MAC层桥接所制订的IEEE 802.1D规范中。

IEEE 802.1p扩展了基本LAN桥接器的功能，其中添加的内容包括：

- 对数据帧进行标识和划分优先级的机制，作为一种在桥接LAN环境中支持时延敏感业务流的手段。

- 提供了动态的过滤器，以便在桥接LAN环境中支持分组的组播通信。

在802LAN上支持不同的业务流类型并提供不同服务等级的能力是通过在帧中标记“优先级”比特然后让交换机根据帧的优先级比特来把帧放到一个传输队列中来实现的。IEEE 802.1p交换机利用了VLAN标记头TCI字段中的3个用户优先级比特。用户优先级字段中的3个比特意味着最多可以定义8个不同的业务流类型或优先级别。

在每一个数据帧中都具有一个用户优先级字段，可以控制并使帧经过网络总传输时延减小到最小。总的时延是排队时延和接入时延的总和。其中排队时延（queuing delay）是帧在出口链路上被选择发送前在队列中等待的时间量。接入时延（access delay）是帧在接入LAN且能够被发送前必须等待的时间量。某些LAN技术如以太网等，并没有控制这个过程。用户优先级是一种技术，它可以被用来为那些不支持这项功能的LAN技术提供访问的优先级。

用户优先级可以由数据源来设置，也可以由802.1p交换机来设置。交换机对优先级的设置可以根据某些策略或者根据配置信息。例如，可以把所有来自端口A的数据都标记成高优先级的。图4-19给出了802.1p交换机的组成和操作过程。

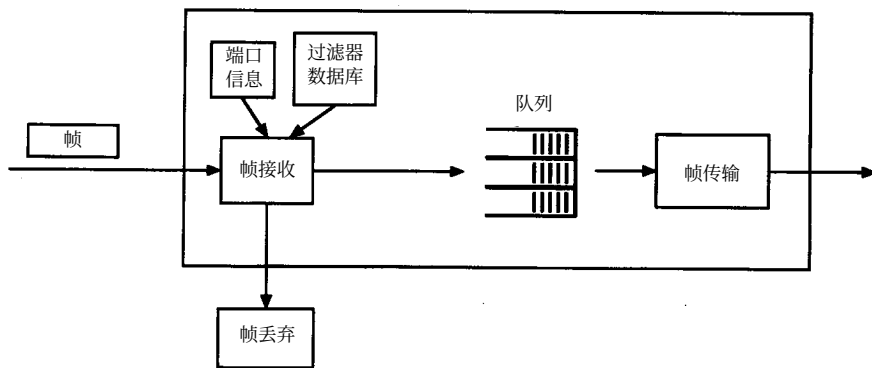


图4-19 IEEE 802.1p交换机

当帧到达一个交换机端口时，与MAC地址相关的不同准则、过滤器及正确的LAN操作都被应用到该帧上。如果帧在其中的检测过程中失败或不能通过一个过滤器，该帧将被丢弃；否则，它就会依据其用户优先级比特的内容被放在若干优先级队列中的一个。如果在帧中没有用户优先级字段（即该帧来自一个非802.1p的工作站），交换机可以选择一个缺省值，并将该值编码到帧中。

从历史上讲，LAN对待组播数据与对待广播分组是一样的，即将分组发送到所有的工作站及每一个桥接器或交换机的端口。但这并不是对网络资源的一种高效的使用方法，因为只有一个LAN成员的子集希望收到组播的数据帧。因此，制订IEEE 802.1p规范的第二个目标就是定义能在LAN环境中高效率地支持组播的技术。

这个目标的实现方法是，在每个交换机内建立一个桥接器组播过滤数据库，并采用动态组成员资格协议来更新数据库。这样交换机就能够跟踪哪些组播组是活动的，哪些端口上有

组播成员。对那些没有组播成员的端口，交换机将把组播流量过滤掉。

IEEE 802.1p中所规定的动态组成员资格协议是 GARP组播注册协议（GMRP）。GMRP是一种使终端系统和交换机能注册或注销组播组成员资格信息的应用，它能够让注册信息在整个网络中传播。GMRP采用了GARP来实现这一点，如图4-18。潜在的组成员注册其组地址及其邻近交换机的端口。组地址被发布到整个网络中的各交换机，最终产生一个无环路的生成树。这样，网络就能支持动态的组播组，同时最小程度地影响非组播组的资源。

4.3 LAN交换网络实例

LAN交换可以有各种的表现形式，其配置依赖于终端用户和应用的需求以及人们所希望的实际容量和性能的增加。下面我们将说明两种 LAN交换使用的场合，即工作组交换和骨干交换。

工作组交换（workgroup switching）是一种提高终端用户带宽和性能的简单而廉价的技术，这些终端用户通常使用邻近或本地的计算资源。工作组交换的思想如图4-20所示。以前，大部分工作都需要利用一个本地服务器的客户可以共享一个LAN网段（以太网或令牌环）。将共享介质的技术替换为一个LAN交换机消除了LAN访问的争用问题，每一个客户现在就可以拥有一条专用的网络访问通路而不被其他工作站、LAN适配器、服务器或应用程序所争用。此外，服务器可以被装备一个高速全双工的LAN适配器（如快速以太网），这将进一步改善它对客户提供的服务。再考虑到工作组交换的费用低廉，采用其他的方案几乎是没有什么意义的。

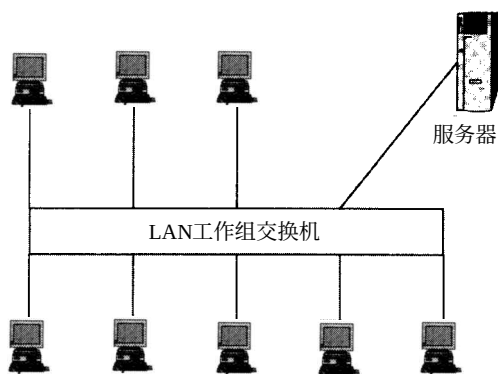


图4-20 LAN工作组交换

工作组只构成了不断扩展的LAN环境的一个很小的部分。使用LAN交换技术对整个网络全面地扩展及提高性能和容量需要进行组间业务流的交换。特别是在服务器资源与客户不在相同的地理位置的情况下。因而采用交换的骨干网具有十分重要的意义，因为它可以在工作组之间提供连接和高速的数据传输。

建立一个LAN骨干网有许多不同的选择方案。一种方案是把所有的工作组交换机和服务器按照一种集线器或放射状的拓扑结构折叠成一个单一的骨干交换机，这个骨干交换机也是一个集线器。从工作组客户的角度来看，它提供了一条与连接到骨干网上的服务器之间的直接高速链路，但是在工作组到骨干网链路失效的时候它不能提供任何冗余通路。

另一个方案是建立一个骨干交换机，并把所有的工作组都连接到骨干网上。骨干网在网络中任何两点之间提供多条交换通路。这为链路或交换机失效的情况提供了增强级别的冗余度和可靠性。图4-21给出了一个部分互连的交换骨干网的例子。为了便于说明，在同一位置中的工作组和骨干交换机被表示在2个不同的方框中，但在实际配置中可希望这两项功能配置在同一个机箱中。可以注意到，从工作组到服务器之间存在多条交换通路。如果单个工作组

或骨干交换机失效，也只会影响到该工作组。然而，存在多条可能的通路并不意味着业务流可以利用它们。可以回忆以下 4.1.2 节的内容，生成树被用来在交换机或桥接器网络中建立一条转发数据流的无环路的通路。在这个环境中，生成树将会把转发链路和交换机限制在一个网络拓扑的子集中。如果生成树的转发通路上出现拥塞就会造成一个问题——没有办法来使用其他可用的那些链路和交换机。为了消除这种局限，ATM 被作为骨干来对基于 LAN 的工作组进行互连。由于 ATM 采用面向连接的机制，它会首先定位所有能够适应带宽和业务流的其他需求的可用通路。因此，尽管对于工作组环境我们认为 LAN 交换是最好的选择，但是对于进行工作组互连的骨干来说，ATM 无疑是十分有用的。

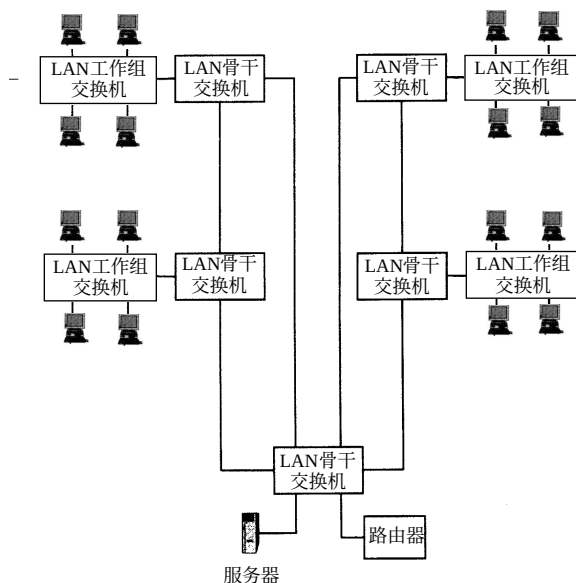


图4-21 LAN骨干

更多的信息

Draft Standard P802.1Q/D8, IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks, December 1997.

Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Common specifications—Part 3: Media Access Control (MAC) Bridges: Revision (Incorporating IEEE P802.1p: Traffic Class Expediting and Dynamic Multicast Filtering).

Sackett, G., IBM's Token-Ring Networking Handbook, New York: McGraw-Hill, 1993.

Saunders, S., The McGraw-Hill High-Speed LAN Handbook, New York: McGraw-Hill, 1996.

China-pub.com

下载

第5章 WAN交换技术

到目前为止我们已经讨论过了 TCP/IP、交换的概念和 LAN 交换技术，现在我们可以把注意力集中在 WAN 中的交换技术上。帧中继和 ATM 是营运商、ISP 以及企业 Intranet 用来通过 WAN 交换数据和图像声音所采用的两种最主要的技术^①。

5.1 帧中继

帧中继是一种面向连接的高速分组交换技术，在今天的许多公用和专用 WAN（包括 Internet）中已经十分普及了。通过帧中继互相连接的设备之间需要建立一条或多条 VC，并通过一个帧中继交换网络进行通信。帧中继作为一种提供 WAN 网络互连服务的技术十分流行，其原因不仅由于多个帧中继 VC 可以被复用到相同的网络上，而且它可以为单个用户提供满足其带宽和协议需要的 VC。

5.1.1 体系结构

帧中继技术是随着当今技术的发展和进步从早先的经验和技术中演化而来的。帧中继的前身是 X.25，它从 X.25 技术中继承了 VC 和每个 VC 连接标识符的概念。但是由于 X.25 是在 70～80 年代开发的，当时的线路速率不高（与现在相比），线路质量也时常出现问题。因此在每跳（per-hop）基础上提供差错检测和可靠性机制是十分必要的。虽然这些技术最终能够使数据安全地传递到目的端，但是在反应时间和网络延迟两个方面的代价却很大。

随着 LAN 的出现及 LAN 通过 WAN 进行互连需求的增加，我们越来越需要一种能够处理大流量、可变速率业务流的新技术的出现。可以想象，即使 X.25 通过改进其带宽和性能能够满足上述的要求，在实际上也是完全没有必要的了。因为现今的数字通信系统已经提供了比 X.25 高得多的网络容量，并且也不再需要在网络中提供两点之间的可靠性和差错检测机制了。此外，一个十分重要的功能是根据需要分配网络带宽，这样网络才能够支持数据流不断增加的变化性和随机特性。最后，新的网络技术还必须像 X.25 一样对网络层协议透明，而不是像 SNA 或 IP 那样。

帧中继是能满足这些要求的一项新技术，事实已经证明帧中继是一种极受欢迎的、能够在 WAN 上支持数据通信的通用技术。帧中继技术利用了最新的高质量、高速度和更高性能的通信链路技术（例如光纤通信）。差错检测仍然在每跳基础上实现，但是差错恢复过程被移到了网络的边缘。因此，许多高性能低费用的分组交换机制就能够被置于网络的中央部分。各种智能化的功能如流量控制和差错恢复等，都由连接到帧中继网络上的终端系统来实现。

帧中继功能上的主要特征如下：

- 可变的帧长度。
- 可变的带宽（DS0～OC3+）。

^① ATM 既可用于局域网又可用于广域网中。本章将介绍用于这两种环境中的 ATM 的基本要素。ATM 在 LAN 中的特殊应用将在以后的章节中讨论。

- 支持PVC和SVC。
- 端口共享。多个VC能够发起或终结于同一个网络设备端口。

• 带宽管理。定义帧中继 VC 的时候可以同时定义一些流量参数，用于在通常情况和重负载条件下控制注入到网络中的数据流量。约定信息率（CIR）定义了通常在某段时间内的可流过网络的数据量，单位是比特每秒（bit/s）。约定突发长度（Bc）定义了在一时间段内网络所能支持的最大数据流量，单位是也比特每秒。过度突发长度（Be）定义了网络能够试图传送的最大数据流量。

- 支持多种网络层协议，如 IP、IPX、SNA 等，同时支持桥接 LAN 业务流。
- 采用了基于数据链路控制标识（DLCI）连接标识符的标签交换转发。
- 拥塞控制。帧中继支持显式的拥塞指示。此外，如果信息的传输速率超过了一个定义的阈值，那么其中的一些分组会被标记成符合丢弃条件的分组，也可能被全部丢弃。

- 与 ATM 互通。
- 支持数据，也能支持语音。

• 帧中继已被路由器、交换机、帧中继接入设备及其他多种 WAN 网络设备所支持。尽管不能排除帧中继在主机系统上的应用，但是实际上帧中继的确不是一种基于主机的技术。

图5-1说明了帧中继网络的基本体系结构。帧中继交换机构成了帧中继网络的骨干，交换机之间通过帧中继网络-网络接口（NNI）进行互连。实现帧中继终端设备（FRTE）功能的设备通过帧中继用户-网络接口（UNI）与帧中继交换机上的端口相连。连接帧中继交换机的设备包括路由器、FRAD，也可能还有主机。FRTE负责流量控制、差错恢复和可靠性保证。帧中继交换机负责复用、帧定界、CRC检错和拥塞指示。要在一个帧中继交换机构成的网络中实现通信，需要在FRTE端点之间建立一个永久虚连接或交换虚连接。在图5-1中，FRTE 1与FRTE 2之间就建立了一个PVC。

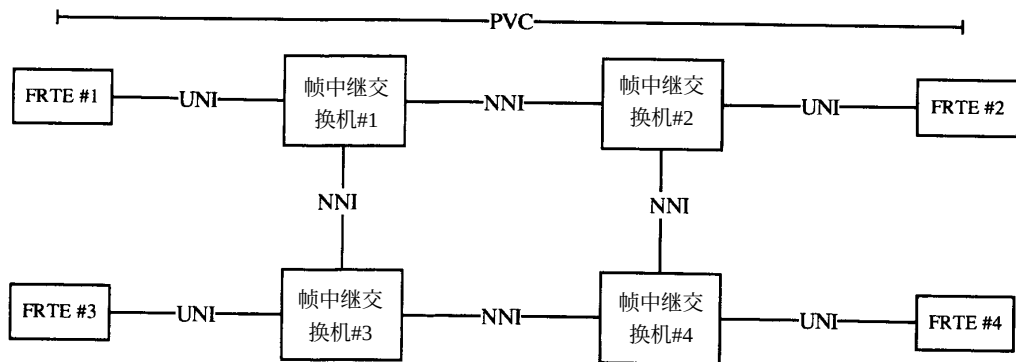


图5-1 帧中继体系结构

有关帧中继功能的规范在过去的几年中已经被多个标准化组织开发出来；这些组织包括美国国家标准委员会（ANSI）及国际电信联盟（ITU）。帧中继论坛^①（Frame Relay Forum）也以实现协议（IA）的形式提供了各种文献，这些文献是了解帧中继基本功能和规范最好的资料来源。帧中继论坛是一个由致力于帧中继技术发展和应用的供应商、开发者、营运者及终端用户组成的协会。表5-1列出了目前已有的帧中继的实现协议（IA）。

^① <http://www.frforum.com>。

表5-1 帧中继论坛实现的协议

标 准	描 述	标 准	描 述
FRF.1	帧中继UNI	FRF.7	帧中继PVC组播
FRF.2	帧中继NN2	FRF.8	帧中继/ATM业务互通
FRF.3	多协议封装	FRF.9	帧中继数据压缩
FRF.4	帧中继SVC	FRF.10	帧中继网络到网络SVC
FRF.5	帧中继/ATM网络互通	FRF.11	基于帧中继的语音
FRF.6	帧中继用户网络管理	FRF.12	帧中继分段

5.1.2 DLCI

帧中继VC是由DLCI标识的。一个DLCI是帧中继协议数据单元（PDU）中的一个多比特的字段，它标识了该分组所属的VC。更简单地说，DLCI就是穿过交换机时，被贴在属于某一VC的分组上的一个标签。VC只有在交换机对之间、端到端环境下有意义，而DLCI只在局部有意义，因为它在穿过网络中的每一个交换机时都会被交换和更改。因此我们可以说，帧中继交换采用了一种标签交换的转发机制。

DLCI最少有10比特长，包含在一个2字节长的地址字段中。在更长的地址字段中也已定义了更长的DLCI字段。图5-2给出了包含10比特DLCI字段的一个2字节基本地址字段的结构。帧中继PDU由下列字段构成：

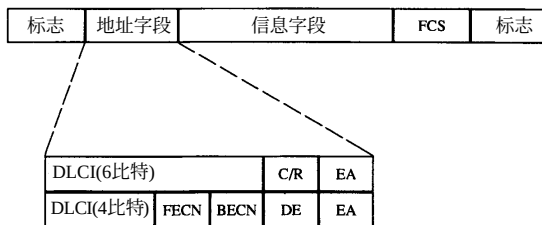


图5-2 帧中继PDU

- 标志（flag）。帧中继PDU定界符，值为十六进制7E。

- 地址字段。由2、3或4个字节构成，其中包含了DLCI。大部分应用实现都采用了基本的2字节格式。在一些文献中将其称为Q.922地址。

- DLCI。一个唯一的10比特值，标识分组所属的帧中继VC。
- 命令/应答（C/R）。1比特的标识，指明该帧是命令或是应答。
- 扩展地址位（EA）。
- 前向显式拥塞指示（FECN）。1比特字段，用来通知终端用户，在帧传送的方向上出现拥塞。
- 后向显式拥塞指示（BECN）。1比特的字段，用来通知终端用户，在帧传送的相反方向上出现拥塞。
- 丢弃认证（DE）。1比特的字段，指出拥塞发生时丢弃的优先级。
- 帧检测序列（FCN）。2字节的字段，用于对整个帧中的数据的CRC校验，校验范围不包括标志字节和FCN自身。

5.1.3 拥塞控制

拥塞控制通常被定义为一种机制，它可使网络成员和终端系统能够很好地适应网络中可用容量的变化。通常在注入网络的流量超过了链路和交换机的容量时，拥塞就会发生。因此拥塞控制的目标就是使业务源减少其发送到网络中的数据量，以便使输入到网络中的流量不超过网络的可用容量。数据源可以通过隐含或明确的方式检测到拥塞的存在。隐含的拥塞指

示可能是分组拥塞和延迟，数据源可以由此推断出网络中出现了拥塞，应该降低其数据发送的速率，直到拥塞消失，这一点与 TCP的工作过程相同。显式的拥塞指示是由网络向业务源发送显式的信号来指示网络拥塞的发生。

帧中继使用显式通知方式进行拥塞控制。显式拥塞指示信号装载在帧中继 PDU地址字段中的FECN或BECN位上。当分组从发送源传递到接收者的过程中出现拥塞，帧中继交换机就会设置FECN位；如果在相反的方向上出现拥塞，帧中继交换机就会设置 BECN位。接收到 FECN或BECN指示后，业务源可能会降低其 CIR或者可能将某些其他的分组标记为可以丢弃的。

5.1.4 RFC1490

帧中继的通用性表现在它既可以传输路由（3层）数据流又可以传输桥接（2层）数据流。RFC1490定义了帧中继数据字段的扩展方法，以支持对路由和桥接分组的封装。这使得帧中继成为在WAN上承载LAN互连数据流的最理想选择。RFC1490之所以重要的原因是它能够在同一个PVC上支持多种协议（如IP、IPX），而不需要为每一种协议申请不同的PVC。这种方式的一个十分明显的优点是它只需要提供很少量PVC，因此可以在网络及帧中继设备（如路由器）中降低费用和资源消耗。

在RFC1490中定义了一些不同的格式，分别用于指明业务流的类型是路由的还是桥接的以及在分组中承载的是什么协议。图 5-3中分别给出了路由及桥接格式。在路由格式中，RFC1490头紧跟在Q.922地址字段的后面。RFC1490包含一个控制字段、一个分组拆装（PAD）字段和一个网络层协议标识（NLPID），NLPID指明在信息字段中承载的内容所采用的协议，例如，IP协议所对应的NLPID值是0xCC。但是还有许多协议，包括桥接数据在内，并没有被分配一个NLPID值；在这种情况下，就需要定义附加的字段。在桥接格式中使用了一个5字节的SNAP封装头。它由一个3字节的OUI字段和一个2字节的协议标识符（PID）组成。把这两个字段的各种取值组合起来，就可以标识出信息字段中可承载的内容，例如是桥接的IEEE 802.3以太网帧等。这时还需要在SNAP头之后包含一个MAC地址。

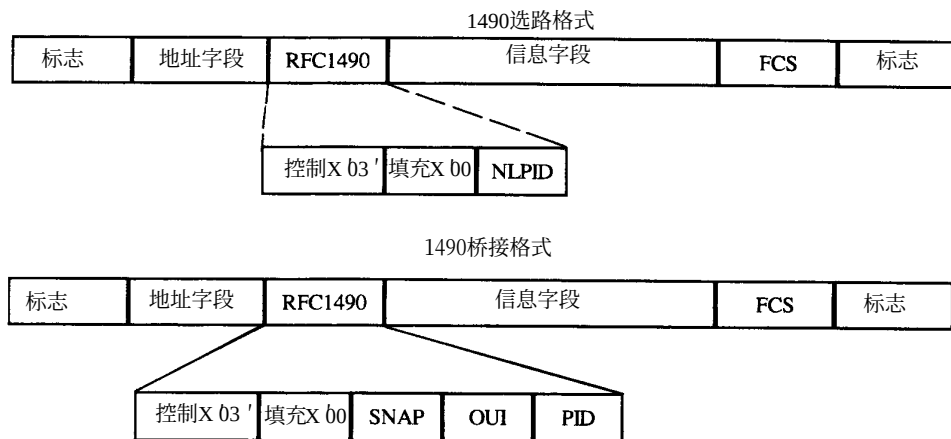


图5-3 RFC 1490

a)选路 b) 桥接格式

5.1.5 帧中继网络实例

帧中继是一种功能强大而通用的技术，它能够以较低费用通过 WAN 实现数据设备的互连。列举一些实例可以有助于我们了解帧中继实际使用方法以及我们前面所讨论的一些概念。

图5-4中的例子是由三个路由器通过一个帧中继网络实现的全互连。从中我们可以看到：

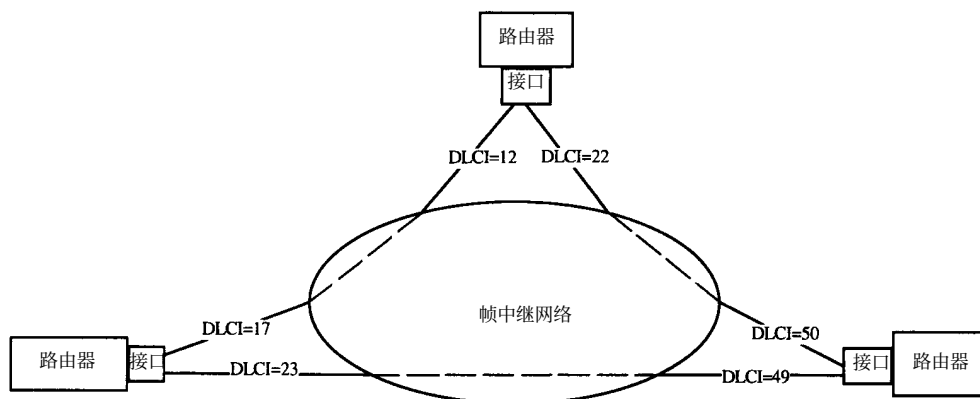
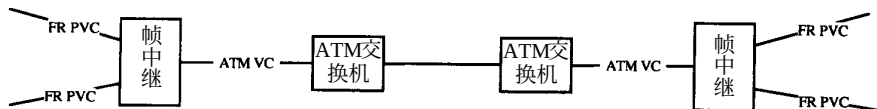


图5-4 连接到一个帧中继网上的路由器

- 发送源上的一个单独的帧中继接口可以支持去往不同目的地的多个 PVC。这种端口共享的能力可以降低网络中的路由器及交换机的端口价格。
- 每个路由器上所配置的 DLCI只是在本地才有意义。换句话说，在同一个 PVC 两端，其 DLCI 可以不同。
- 每一个 PVC 可以配置不同的业务流参数（如 CIR 等），因此可以适应任何两个路由器之间的、具有特殊流量特性的服务。



(a)

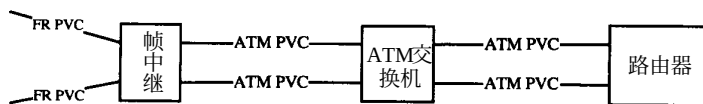


图5-5 帧中继与ATM

a)互连网络 b)服务互连

帧中继通用性的另一个证明是它具有与 ATM 互通的能力。图5-5给出了两个帧中继与 ATM 协同工作的例子。在帧中继与 ATM 网络互通的情况下，ATM 网络作为帧中继交换机之间的骨干网络。首先，帧中继 PVC 进入一个配备了 ATM 上行链路的入口帧中继交换机（ingress FR switch），然后帧中继的业务流就通过 ATM 网络传递到一个出口帧中继交换机（egress FR switch）。

switch), 在该出口交换机中业务流重新被还原成帧中继 PVC, 继续传递到其目的地。在 ATM 骨干网上汇聚和传输帧中继 PVC 业务流是 ISP 十分常用的一种利用 ATM 技术的方法。另一个例子是帧中继与 ATM 的业务互通。在这种情况下, 帧中继的 PVC 被映射到一个单独的、终结于路由器或主机的 ATM PVC 上。这种配置可能的应用场合是, 许多由帧中继网络互相连接起来的远程站点需要与高性能的 ATM 网中的某些资源实现通信。将帧中继 PVC 映射到 ATM PVC 的工作是由交换机中的互通功能实现的。

5.2 ATM

ATM 是一种面向连接的信元交换技术, 它可以在相同的物理基础设施上支持数据、图像和语音业务。应用程序和设备利用 ATM 技术在 ATM 交换网络中互相建立虚连接。与帧中继或 LAN 交换不同, ATM 支持包括 QoS 在内的多种业务, 支持 QoS 使得 ATM 能够像支持普通数据业务一样支持实时业务应用。正如我们下面将要讨论的, ATM 也是一种功能强大的 2 层交换技术, IP 和其他网络层协议都可以利用 ATM 的这一特点。

5.2.1 体系结构

ATM 的体系结构来源于 80 年代末为 ISDN 宽带化和增强 ISDN 性能所进行的工作。宽带 ISDN 技术开发的目标是实现高速、数字化和面向连接的服务, 它可以支持各种不同的应用, 例如数据、实时传送、交互式图像和声音等等。由于种种原因, ATM 被选作了宽带 ISDN 的传输模式。这些原因包括 ATM 是独立于传输的、可以按需要分配带宽、能十分简单地提供半永久连接、其硬件和相关的性能易于扩展和升级等等。由于 ATM 具有这些特点, 并且由于它能够在相同的基础设施上统计复用大量的数据、语音和图像的连接, ATM 在短短几年内就被公认为是下一代网络的技术。尽管 ATM 不一定能够完全发展到那个地步, 可是 ATM 能够也必将在未来的网络中扮演一个重要的角色。

图 5-6 简化了所有不必要的细节, 描述了一个 ATM 系统的体系结构。尽管所有其他的网络互连结构都提供了一些控制平面的功能, 但 ATM 所提供的控制平面功能是最出色的, 因为 ATM SVC 作为常用的功能, 它的建立和撤消都需要显式的信令过程。因此, ATM 终端系统——包括主机、路由器、边缘设备及其他系统——都有一个信令平面和一个控制平面。信令用于在 ATM 网络中建立和撤消连接。连接一旦被建立起来, 数据就从应用层向下传递到 ATM 适配层。适配层将高层的应用数据适配到底层的 ATM 服务上。目前已经定义了一些不同的 ATM 适配层的类型, 对适配层应用有赖于其所提供的 ATM 服务。适配层所提供的另一项功能是分段和组装。该功能在源端将高层净荷分成 48 字节长的净荷段, 而在目的端又将 48 字节的净荷段组装成高层的净荷。

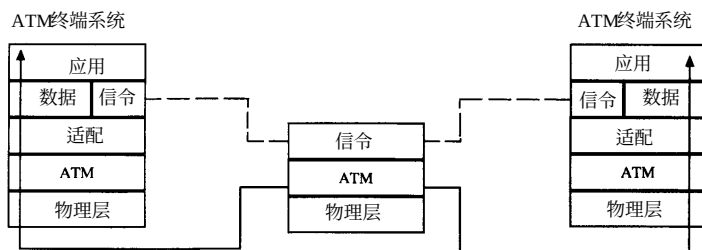


图 5-6 ATM 体系结构

当数据以48字节净荷段的形式传递到ATM层后，一个5字节的信元头被加到每个48字节的净荷上，构成一个信元。随后，信元向下传送到物理接口上。物理层接口采用多种不同的链路层技术构成，例如同步光纤网络 /同步数字系列（SONET/SDH）、非屏蔽双绞线（UTP）、T3及T1等。

在交换通路的每一个中间节点交换机上，单个的信元都是根据其信元头的内容而被交换的，交换过程中采用了标签交换的机制。信元在到达其目的端以前不会被组装成原来的分组。

终端系统与其相邻交换机之间的接口是UNI，相邻交换机之间的接口是NNI。这种区分是很重要的，因为流经UNI接口的数据流和流经NNI的数据有着不同的协议、消息和信元格式。

ATM主要的功能特点可以总结如下：

- 采用了面向连接的信元交换技术。
- 具有可变的带宽（DS1到OC48+）。
- 支持永久、半永久和动态的虚连接及永久和交换的虚通路。
- 提供了端口共享；多个VC可以起源或终止于相同的物理设备端口上。
- 采用53字节的固定长度分组（称为信元）。
- 为不同的业务流类型（实时、可变速率或尽力而为）提供每个连接的QoS支持。
- 支持点对点 and 点到多点的连接。
- 基于ATM信元头中VPI/VCI的标签交换。
- 为主机、路由器、交换机和边缘设备等所广泛支持。
- 可与帧中继及LAN等互通。
- 可支持LAN仿真，基于ATM的IP以及ATM固有的应用。
- 支持数据、语音和图像。

在很大程度上，ATM论坛^①推动了ATM的标准化进程。ATM论坛是一个由供应商、开发者、营运者和消费者组成的组织，成立于1991年。表5-2列出了一些ATM论坛所完成的、以多种ITU宽带ISDN的原始规范为基础的ATM标准。

表5-2 ATM论坛的标准

ATM论坛实现的协议	描 述
UNI 3.0/3.1/4.0	ATM用户网络接口信令的不同版本
业务流管理4.0	ATM流量管理机制，包括ABR基于速率的流量控制
PNNI Phase 1	动态交换机之间的信令及SVC设置请求的选路协议
IISP	临时交换机间信令协议，对SVC设置请求的静态选路
LAN仿真(LANE)1.0	在ATM网络上层仿真以太网或令牌环的服务和协议
基于ATM的多协议(MPOA)1.0	在ATM网络上用于分布式虚拟路由器操作的服务和协议

5.2.2 虚通路和虚连接

ATM支持两种类型的连接：VC和虚通路（VP）。VC定义为两个或多个终端系统之间传输信元的专用逻辑信道。VC的起源和终止都在ATM终端系统中，并且我们可以为VC配置特定的业务和QoS特性参数，如带宽和时延等。PVC是在网络初起时静态配置和初始化的VC。“软”PVC是在网络激活的状态下根据需要在多个终端系统之间配置和建立的一种半永久的PVC。

① <http://www.atmforum.com>。

另外，还可以通过信令动态建立 VC，称为 SVC。属于某个特定 VC 的信元由其信元头中的 VC 标识符（VCI）来标识。

VP 是一组 VC，其中每一个 VC 都与一个唯一的 VP 相关。属于一个特定 VP 的信元由其信元头中的虚通路标识符（VPI）来标识。VP 也可以是永久的（PVP）或交换的（SVP）。在网络中，VP 是进行大量 VC 交换的一种高效率的手段，因为这种交换（也同样是标签交换）仅基于信元中的 VPI，而忽略 VCI 值。在那些不支持 SVC 的广域 ATM 网上，PVP 对于支持基于 SVC 的通信是很有作用的，其实现的方法是利用 PVP 建立一个管道来传递 SVC 设置的请求。多个 VP 以及这些 VP 中的 VC 可以在同一个物理链路上传输。

物理链路、VP、VC 等都在图 5-7 中进行了描述。应该注意，在一个 VP 中，VCI 必须是唯一的。当然，在不同 VP 中可能有两个 VC 具有相同的 VCI 的值。

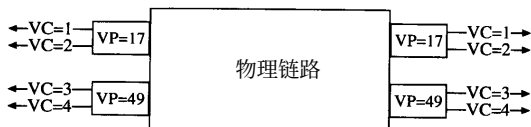


图5-7 ATM虚通路和虚连接

5.2.3 信元格式

ATM信元的长度为53字节，其中有48字节的净荷和5字节的信元头。UNI及NNI信元的格式分别在图5-8中给出，其各字段将在下面说明。需要再次说明的是，UNI格式的信元是在终端系统和交换机之间传送；NNI格式的信元是在两个交换机之间传送。

- 一般流量控制（GFC）。一个4比特的字段，定义GFC的最初目的是为了在终端系统和交换机之间传送流量控制指示，但目前并没有实现。GFC只在UNI格式中存在。

- VPI。这个字段在UNI中有8比特，在NNI格式中有12比特。NNI格式中多出来的VPI比特来自于GFC字段，因为GFC字段在NNI格式中不存在。

- VCI。该字段为16比特。

- 净荷类型指示（PTI）。一个3比特的字段，用来指出净荷里装载的是用户数据、网络管理数据还是流量管理信息。PTI中有一个比特是AAL指示位，当该信元是一个数据包的最后一个信元时为1，其他信元均为0。它被终端系统中用来确定一个数据包的结束和下一个数据包的开始。

- 信元丢失优先级（CLP）。一个1比特的值，用于区分那些符合流向约定和不符合流量约定的信元。不符合流量约定的信元在出现网络拥塞时可以被丢弃。

- 信头差错检测（HEC）。这个字段被ATM网络层用于检测和纠正信元头中的差错。

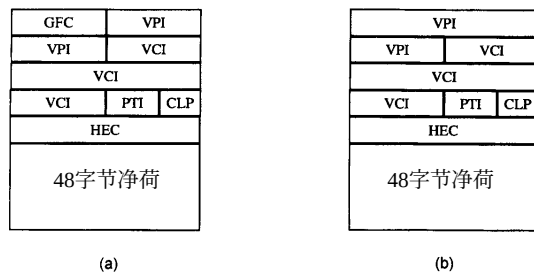


图5-8 ATM信元格式

a) UNI b) NNI

5.2.4 ATM寻址

ATM终端系统和交换机中模仿 OSI 网络业务接入点（NASP）的结构采用了一个 20 字节的寻址结构。根据 ATM 终端系统是在公用网还是在专用网中，存在几种可以采用的格式。不管格式如何，ATM 地址由一个最大 13 字节的地址前缀，一个 6 字节终端系统标识符（ESI）和一

个1字节的、只在本地接口有意义的选择符组成。20字节的ATM地址的格式如图5-9所示。

ATM地址前缀	ESI (6)	Sel (1)
---------	---------	---------

图5-9 ATM地址

5.2.5 AAL5格式

ATM支持不同的ATM适配层，这些适配层可以根据应用的需要提供特定的服务。AAL1中具有时钟恢复的机制，可以在源端和目的端之间进行时钟恢复，通常用于支持实时服务，如电路仿真和语音等。AAL3/4能够在同一个链路上支持来自不同数据源的信元实现复用和交错，通常被用于支持无连接的网络互连技术，如交换的多兆比特数据业务（SMDS）等。AAL1和AAL3/4都在48字节的净荷里附加额外的开销。由于已经存在了9%（5/53）的所谓信元税（cell tax），再使用净荷中的字节将进一步地降低ATM信元的使用效率。

为解决适配效率问题，AAL5出现了，它最初被称为简单和高效的适配层（SEAL），也几乎是所有实际的ATM系统中应用最多的AAL协议。AAL5不需要在48字节的净荷中附加任何开销，并且实现十分简单。这是因为它除了需要处理一个用于帧级别差错保护的32比特CRC值以外，不需要处理任何分组头和尾。

此外，由于在已实现的ATM系统中，很大比例的系统所承载的数据（特别是IP）都是无连接的和尽力而为的业务，因此，从利用ATM和支持IP的角度来看，除了要求尽量小的处理过程和开销及尽量小的实现费用外，实际上对AAL的功能并没有额外的要求，而AAL5正好可以符合这些要求。

用户信息(1-65535字节)	填充(0-47字节)	CPCS UUI 1字节	CPI 1字节	Length 2字节	CRC 4字节
-----------------	------------	-----------------	------------	---------------	------------

图5-10 AAL5格式

AAL5的帧结构如图5-10。用户数据的长度可以是1字节到最大65535字节。在用户数据后面有一个PAD字段，用于将用户数据填充成48的整数倍。在PAD字段之后，是一个8字节长的AAL尾（trailer），它主要由一个长度字段和一个CRC校验字段构成。

正如我们将在本书的剩余部分中所描述的那样，所有基于ATM的网络互连方案（除了LANE 1.0以外）都使用了AAL5作为其适配层。

5.2.6 UNI信令

能够根据需要在终端系统之间建立虚连接是ATM最引人注目的特性之一。网络资源一般以带宽、缓存器、标签等形式出现。系统应该能够按照需求分配网络资源，在整个连接期间资源被相互通信的终端系统所使用；一旦不再需要时这些资源将被释放，以便其他的用户能够继续使用。所有这一切都是动态完成的，而不需要像PVC那样进行预先的配置。SVC具有的另一个精巧而引人注目的特征是，它不会为了保证新建连接所需的资源而剥夺现有其他连接的资源。换句话说，当一个申请建立SVC的请求发送到网络时，如果网络没有足够的容量来支持该连接，这个SVC请求将会被拒绝。这有点类似于电话中遇到忙音。最后需要说明的是，每一个SVC申请都可以带有若干用于说明ATM终端系统间数据流的特定的流量和QoS参数。

UNI信令是一个ATM终端系统向其他一个或多个终端系统传递SVC建立请求的方法。更明确地说，UNI信令是由一些协议以及一些能在ATM终端系统和交换机之间的UNI接口上进行交换的消息构成的。UNI信令消息不承载任何数据，而是承载有关连接的特定的信息要素，

例如目的ATM地址、由平均带宽和峰值带宽等构成的业务流描述符以及QoS参数（如最大允许的时延抖动）等等。各种UNI规范中信息要素都在ATM论坛的UNI规范中作了描述。

目前的UNI信令能够支持建立点到点和点到多点的SVC。点到点的SVC可以是双向的，而点到多点的SVC是单向的，只能从发送者流向所有的接收者。图5-11中的例子说明了建立一个点对点的SVC所必需的UNI消息交互过程。

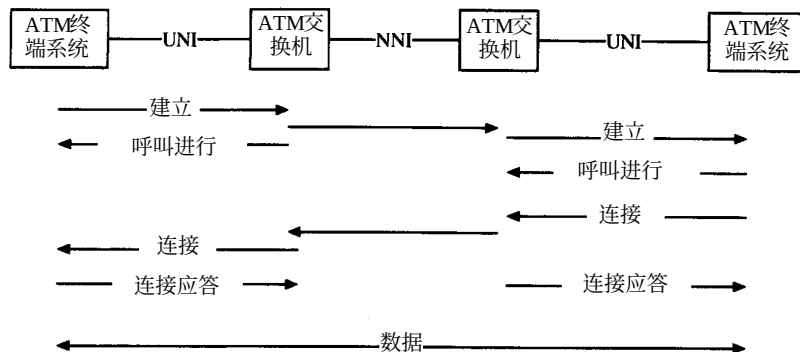


图5-11 ATM UNI信令

UNI 3.0和3.1规定，点到点和点到多点的SVC的请求者都必须事先知道目的地址。UNI 4.0中增加了叶节点启动加入（LIJ）功能，它可以使叶节点^①能动态地加入到一个已有的点到多点的SVC中。UNI信令的进一步完善的方向可能是，提出新的轻量的信令协议以及提供建立多点到点SVC的能力。

5.2.7 ATM的服务类型和流量管理

ATM支持多种不同的应用类型以便以最佳方式工作，这些应用类型对网络有不同的需求。例如实时应用需要有最小的延迟、低抖动、时延限制和最小带宽。而与之相对的尽力而为的应用会使用尽可能多的可用带宽，但能实现信息传输速率与网络可用容量的变化的自动适应，而不会对应用造成影响。除这两种应用以外还有一些应用，它们对业务流的需求在不同的时间段内是不断变化的，它们或许需要或许不需要网络在时延、抖动及带宽等方面的特殊保证。

为了在一个广泛的范围内实现对各种不同应用的支持，ATM体系结构定义了一系列不同的服务类型。ATM的服务分类在ATM网络中定义和支持了一组服务，任何一种应用都属于其中的某一类。ATM的分类可包含所有应用的服务类型，这种分类定义可以使规范更容易制订和遵循，并使ATM终端系统能够了解，它能向网络要求什么及不能向网络要求什么。

ATM的服务分类在表5-3中给出，在ATM论坛的TM 4.0规范中作了详细的定义^②。

根据终端系统向网络申请的服务，ATM可以采用多种流量管理技术。流量管理优化了网络资源的使用，避免和限制了网络拥塞的发生。流量管理也有助于保证和实现终端系统和网络所达成的相应连接QoS的协定。

TM 4.0规范详细定义了ATM流量管理的功能。一些很著名的流量管理机制包括：

- 连接接纳控制（CAC）。CAC首先检查是否有可用的资源能够分配给一个连接请求，然

^① 叶节点是指被连接到一个点到多点的虚连接的接收端上的一个ATM终端系统。

^② “ATM论坛流量管理版本4.0”，ATM论坛，af-tm-0056.000，1996年4月。

后决定允许连接建立过程继续进行或者拒绝该请求。

- 使用参数控制（UPC）。也被称为流量管制，UPC在网络的入口处检验业务流是否符合源与网络所达成的业务流协议。不符合的信元将会被标记或丢弃。

表5-3 ATM的服务分类

ATM服务分类	描 述
CBR	固定比特率，用于电路仿真或实时图像等应用，需要固定带宽及端点之间的时钟同步
Rt-VBR	实时可变比特率，用于以可变比特率传输实时数据的应用，如图像等
Nrt-VBR	非实时可变比特率，与rt-VBR服务相同，但没有时延的限制
ABR	可用比特率：为应用提供尽力而为的服务，可以根据网络的反馈调整传输速率
UBR	未规定的比特率：纯粹的尽力而为的服务，没有流量控制
GFR	保证的帧速率，类似于VBR，但要保证一个最小的信元速率

- 选择性的信元丢弃。CLP = 1的信元将会被拥塞的设备（如交换机）丢弃。
- 流量整形（traffic shaping）。可以改变信元流的特性，以便使之与流量协定保持更大的一致。

- 帧丢弃。当交换机中出现拥塞时，信元将被丢弃。如果丢弃了帧中的一个信元，帧丢弃机制将把该帧中剩余的其他信元也全部丢弃。该功能的变形也出现在早期分组丢弃（EPD）拥塞控制技术的文献中。帧丢弃是限制未指定比特率（UBR）业务拥塞的一种很有用的方法。

- ABR基于速率的流量控制。ABR源能够根据网络的反馈动态地调整其传输速率。ABR源周期性地产生一个特殊的资源管理（RM）信元穿过网络发往目的端。中间节点交换机将把拥塞指示标记在信元中或者可能会在信元中标记一个显式速率（ER）值。当RM信元到达目的端并返回到源端时，源端就可以根据RM信元中的拥塞指示调整其速率或把速率设置为ER字段所包含的值。

5.2.8 PNNI

许多人相信由于ATM是一个面向连接的交换技术，选路将是不必要的。对于在已建立的PVC或SVC上传输的信元来说确实是这样。VPI/VCI（或更一般地说即信元头中的标签）并不是一个可选路或可寻址的实体。它只简单地标识了该信元所属的连接，并且只在本地有意义，在信元通过一个交换机进行转发时，VPI/VCI的值很可能被交换。但是，在SVC连接建立之前，SVC连接请求必须从源端传到目的端。在该SVC连接请求中必然有一个可以选路或寻址的字段，即ATM目的地址。此外，SVC请求中还可以有一个业务流描述符，其中包括峰值比特率和持续比特率（PCR、SCR）及QoS参数如信元丢失率（CLR）、最大信元传输时延（CTD）等等。因此ATM的选路系统必须能识别和选择一条到达目的地的通路，该通路能够支撑和维持SVC请求中所规定的带宽和QoS需求。

这不是一项像我们想象的那样简单的任务。注意，通常你所熟悉的IP路由协议在计算通往目的端的路由时只考虑一维参数（如最少的跳数或费用）。为ATM SVC请求进行选路必须考虑的因素包括：到达目的端可靠性、通路上的可用容量、允许注入通路的流量负荷以及通路是否能达到连接的QoS要求等。考虑到在大型的ATM基础设施上需要管理的SVC建立请求的数量可能很大，任何ATM路由协议都必须是可扩展的和自主配置（self-configuring）的，并且必须尽量减少其对网络带来的负荷。同时，它还必须能在ATM网络的交换机中准确地跟踪和发布网络资源的状态、数量和质量等。

PNNI阶段I是由ATM论坛开发的信令和路由协议，它能在ATM专用交换机网络中动态地为SVC设置请求选择路由^①。PNNI阶段I中包含了一个路由组件，它能够向网络中的所有ATM交换机发布拓扑状态信息。PNNI阶段I中还包含了一个信令组件，它利用拓扑状态信息来在网络中计算一条从源到目的端的路由，用于转发SVC设置请求。

PNNI阶段I中借用了一些我们前面所讨论路由协议中的机制，以实现其高效和可扩展的目标。PNNI利用了下列的机制：

- 链路状态选路。ATM交换机向其他的交换机发布拓扑状态信息，并利用这些信息建立拓扑状态数据库。拓扑状态选路的含义是交换机将发布有关链路状态和交换机自身的信息。然后，拓扑状态数据库中的内容就可被用于计算出一条转发SVC设置消息的通路。

- 源选路。通过让源或入口ATM交换机根据拓扑状态数据库来计算整个穿越网络的通路，PNNI限制了SVC设置期间的资源消耗。源ATM交换机把一个源路由向量，称为指派传输列表（DTL），添加到SVC设置请求消息中并发送出去。每一个中间节点交换机只传送该DTL，而不关心通路的计算。还应注意在SVC建立过程中，在入口交换机中整个拓扑数据库将被用于在网络中为特定连接确定一条最优的通路。与生成树或IP路由协议不同，PNNI能够利用整个网络链路和交换机的拓扑结构来建立连接。

- 层次选路。PNNI必须能扩展，为了实现这一点就必须引入层次化的概念。一个共享相同的ATM地址前缀和拓扑状态数据库的ATM交换机集合被称为一个对等组（peer group）。其拓扑状态数据库中的内容是由一个唯一的对等组领导者（PGL）总结出来并以聚合的形式发布给其他对等组的PGL。各个对等组的PGL实际上还可以组成更高级别的对等组，并且作为这个高级别对等组的成员，这种成员可称为逻辑组节点（LGN）。LGN并不是实际的物理交换机本身，而是其低层对等组拓扑结构的一个抽象的代表。PNNI的层次概念如图5-12。

通过采用概括和汇聚的形式，层次方法限制了在网络中发布的拓扑状态信息的数量。在图5-12的例子中，对等组1中的交换机共享一个反映该对等组实际拓扑结构的单一拓扑数据库，但其中可能只有外部对等组（如对等组2）中交换机的概括性的信息。

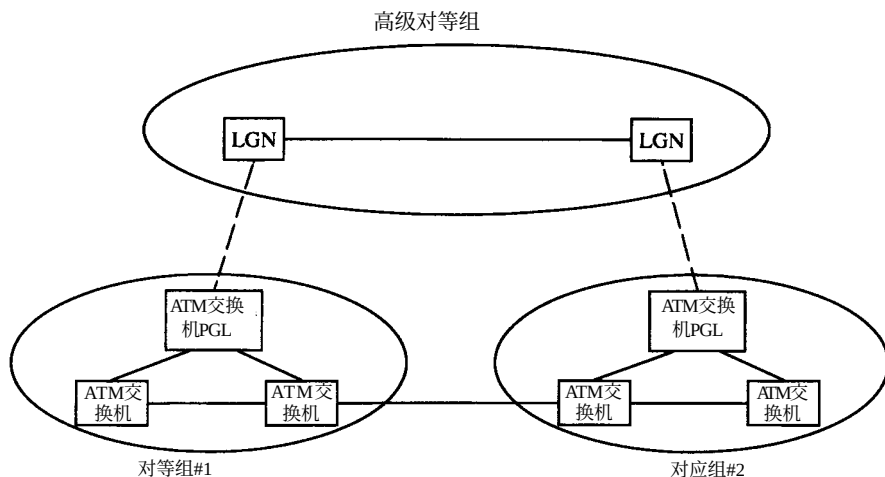


图5-12 PNNI的层次

PNNI阶段I是一个用于专用ATM网及某些公用ATM网的标准ATM路由协议。PNNI的一些

① “ATM专用网络-网络接口规范 版本1.0”，ATM论坛，af-pnni-0055.000，1996年3月。

扩展（包括I-PNNI、PNNI补充选路[PAR]、代理PAR等）都可以支持它与IP路由协议的集成和协同工作，有关内容将在第14章中讨论。

5.2.9 ATM网络实例

ATM能支持多种应用和网络拓扑，在实际应用中呈现出各种不同的应用形式。它可以作为一个完整的端到端的网络体系结构，拥有自己的一套应用、寻址、路由和数据传输功能。尽管在一些小型应用中，例如高质量MPEG2视频等，这种结构运转得十分理想，但是到目前为止，这种小型结构的ATM网络不是很普遍。ATM也可以作为一个高速高性能的2层基础设施，用来支持端到端的数据网络应用或者可以在校园网环境中作为基础骨干网。图5-13给出了这两种应用结合的例子。

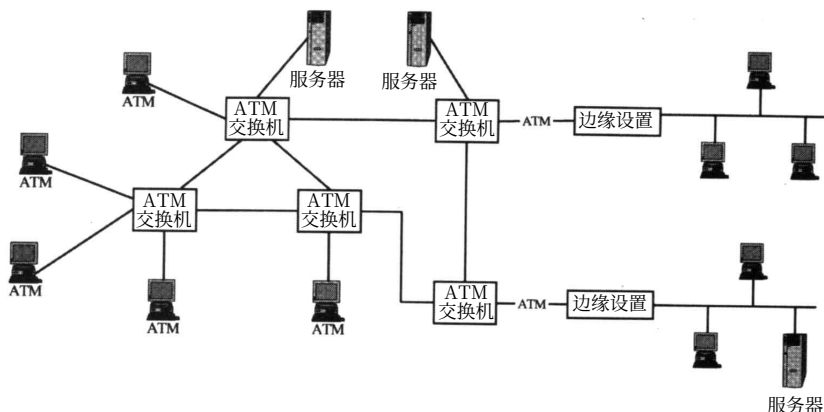


图5-13 ATM校园网

可以看到一些工作站和几个服务器都直接连接到了ATM交换机上。工作站通过使用现有的应用和网络协议如IP、IPX等，就能够把底层的ATM网络作为一个传输客户-服务器应用业务流的高速网络。同时，ATM网络可以作为骨干网络，连接两个以太网（图中的右侧）。可以假设连接在以太网上的工作站和服务器能够以一种透明的方式与那些连接在ATM网上的设备协同工作。这种方式的使用是很有可能，在实际应用中也很普遍。在这个环境中，底层一般采用基于ATM的IP协议，如LAN仿真和经典IP等，有关内容将在后面的章节中讨论。

ATM也作为一种骨干网技术而被ISP所采用。或许这个概念最好的证明在图5-4中，其中我们可以用ATM网络来替代帧中继网络。我们还可以像图5-5中那样，在一个帧中继ATM网络互通的环境中，将ATM作为网络骨干的核心。ATM对ISP和营运者都是十分有用的，因为ATM具有带宽管理功能，同时具备在显式通路上提供虚连接的能力，因而可以达到某种级别的负载平衡。

更多的信息

Black, U., ATM: Foundation for Broadband Networks, Englewood Cliffs, NJ: Prentice Hall, 1995.

Dutton, H., and Lenhard, P., Asynchronous Transfer Mode (ATM) Technical Overview, Englewood Cliffs, NJ: Prentice Hall, 1995.

Dutton, H., and Lenhard, P., High-Speed Networking Technology: An Introductory Survey, Englewood Cliffs, NJ: Prentice Hall, 1995.

Miller, M., Analyzing Broadband Networks, M&T Books, 1994.

Spohn, D., and McDysan, D., ATM Theory and Application, New York: McGraw-Hill, 1995.

China-pub.com

下载

第6章 IP交换的概念

在前面几章，我们从一些细节上回顾了 TCP/IP协议族和当前普遍用于局域网及广域网中的一些重要且流行的下层交换机制和协议。现在我们把这两个话题联系起来，讨论一下 IP交换的概念。IP交换可以定义为利用第二层交换来加速经过一个网络的 IP分组的转发的机制和协议。虽然其他第二层交换（如帧中继）技术也可行，但最普遍用于 IP交换的交换机制是 ATM。本章将介绍有关 IP交换的一些重要概念和定义，并探讨其发展背后的一些动力。

6.1 定义和术语

IP交换是利用第二层交换作为传送 IP分组通过一个网络的主要转发机制的一组协议和机制^①。IP交换利用交换的高带宽和低延迟优势尽可能快地传送一个分组通过网络。虽然如我们所知，在所有的 IP交换解决方案中，有一些的确支持 IP路由和转发，但它们都试图通过一个交换组件重定向所有或部分分组，从而避免执行任何第三层处理。

为更好地理解IP交换的功能和操作，我们先定义几个重要概念。

6.1.1 IP交换机

一般而言，一个 IP交换机是一个设备或系统，能够在第三层转发 IP分组并具有一个使分组也能在第二层被交换的交换组件。一个 IP交换机具有区分哪些分组将在第三层被转发及哪些分组将在第二层被交换的机制，然后通过一条第二层交换路径重定向一些或所有分组。许多 IP交换机采用 ATM交换机构，但其他的二层交换技术也可以使用^②。

IP交换系统的两个例子在图 6-1 中描述。两个例子中的 IP控制点（IPCP）运行一个典型的路由协议（如 RIP、OSPF、BGP），提供一条缺省的第三层逐级跳（hop-by-hop）路由路径，并且可以直接或间接地与交换机组件通信，从而经过交换组件重定向 IP分组。与传统的 ATM交换机一样，交换机组件维护一个包含输入端口、输入标签（VPI/VCI）、输出端口和输出标签的连接表。上图描述一个 IP交换机，下图描述一个虚拟 IP交换机。

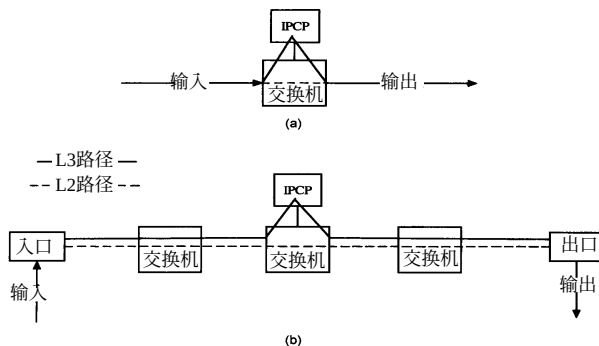


图6-1 IP交换系统

a) IP交换机设备 b) 虚拟IP交换机

① 虽然我们将讨论旁路中间路由跳的 IP交换技术，但应该注意，这个定义包括促进 IP连接和性能的 ATM交换的其他形式。ATM上的经典IP和LAN仿真是后面几章中将要讨论的两个例子。

② 有关 IP交换解决方案的一般和特殊的讨论都假设使用 ATM作为交换组件，这是因为众多的解决方案都采用它。其他的第二层交换技术也可采用。不要求 ATM的特殊情况（如标记交换、MLPS）将特别指出。

这两个模型有一些差别。其一在于第二层交换路径的范围。对于 IP交换机, 交换路径包含在直接受控于一个单一 IPCP的单一设备和交换机组件中。多个邻接 IP交换机必须以某种方式合作, 建立一条经过两个或两个以上 IP交换机的端到端交换路径。对于虚拟 IP交换机, 一条端到端交换路径可以跨越一个单一 IPCP间接控制下的多个交换机。另一个区别是输入和输出“端口”的位置。在 IP交换机设备配置中, IP交换系统的输入和输出端口位于一个机箱中。对于虚拟 IP交换机, 输入和输出组件可以位于同一个机箱中或位于物理上不同的机箱中。

这两个模型的另一个区别是 ATM论坛的 UNI/PNNI 信令和路由协议的使用和与其共存。一个 IP交换机依靠 IP网络拓扑结构和路由协议选择一条转发路径穿过网络, 然后使用一个在它自己和邻接交换机间的特殊控制协议交换, 将逐级跳转发路径映射到一条第二层交换路径上。一个虚拟 IP交换机也使用特殊控制协议来启动此过程, 但其后则依靠 ATM网络拓扑结构、信令和路由协议来选择和建立第二层交换路径。在这种情况下, 使用标准的 ATM论坛 UNI/PNNI 信令和路由协议来完成这一任务。

应该注意到, 一个 IP交换机能够从一个交换机也能从一个端系统上运行 ATM论坛协议。如果两个 IP交换机想通过一个中间 ATM交换机的网络通信, 则需要 IP交换机的 UNI信令和 VC管理。另一种方案是, 一个 IP交换机的交换机组件在一个“ships in the night”的配置上运行 ATM论坛协议(如 PNNI)的情况。这使得一个单一设备能够在一个机箱中提供所有的 IP路由和转发、IP交换及 ATM交换业务。可以假定在 IP交换业务的实际开发中, ATM交换机组件将真正支持 ATM论坛协议, 以给网络提供者提供一些不同的业务选择。

IP交换机这一术语是在 1996 年见诸于众时, 由 Ipsilon 首先使之普及的。这个公司的设备包括一个 IP交换机控制器(如路由器), 连接到一个运行 IFMP 和 GSMP 协议的独立的 ATM交换机构^①。东芝 CSR 和 IBM 集成交换机路由器 (ISR / Integrated Switch Router) 具有相似的配置且使用特殊控制协议, 但它们也包括对 ATM论坛协议的特定支持和与其共存。MPOA 是一个虚拟 IP交换机的例子, 事实上它使用虚拟路由器来描述它提供的功能。

6.1.2 IP交换机的入口和出口

一个 IP交换系统提供缺省的第三层 IP路由和转发及加速的第二层 IP交换的服务。这些业务的受益者可以是独立的应用, 或者是由 LAN 网段、网络或一个共享的目的地聚合成的一组端用户的集合。要制定标准来确定谁或什么能够接收 IP交换业务, 无论标准是什么, 必须有一个进入和离开 IP交换系统的方式。因此要构成一个 IP交换系统的必备组件, 需要一个入口和出口组件。

IP交换的入口和出口组件位于一个 IP交换系统的边缘。一个入口和出口组件可以包含运行于一个工作站的一组代码, 一个已有边缘设备或路由器的增加功能或者一个特殊的黑盒子。一个 IP交换的入口和出口包括下列重要功能:

- 为进入和离开网络的服务提供常规的缺省 IP转发。
- 为进入和离开 IP交换系统的分组提供介质转换(如以太网到 ATM)。
- 参与适当的入口和出口间的第二层交换路径的建立、保持和释放的控制过程。
- 在入口处, 划分出符合条件的分组并将它们转发到第二层交换路径上。一般包括检查分组头的一些字段来决定是否应该将分组放到一条交换路径上, 如果这样, 是哪一条, 查找一个标准路由表, 将标签插入分组或在 ATM交换系统的情况下, 分段并在一个虚连接上传送。

① IFMP 在 RFC1953 中描述。GSMP 在 RFC1987/2297 中描述。

- 在出口处，接收第二层交换路径上的分组，当分组离开 IP 交换系统时完成标准的 IP 转发过程。

一个 IP 交换的入口和出口的概念在图 6-2 中描述。首先假定，IP 交换系统已经确定一些分组可以在第二层交换路径上转发，且这样一条路径已经在网络入口和出口组件间建立。根据一些 IP 层标准，如源/目的 IP 地址或目的网络前缀，到达网络入口的分组被划分为缺省的第三层转发或第二层交换。划分为第二层交换的分组在交换路径上转发。所有其他的分组都使用常规 IP 转发过程在第三层转发。当分组到达网络出口时，或者在第二层或者在第三层，它们都将被传送到第三层进行常规 IP 处理，并转发至目的地。

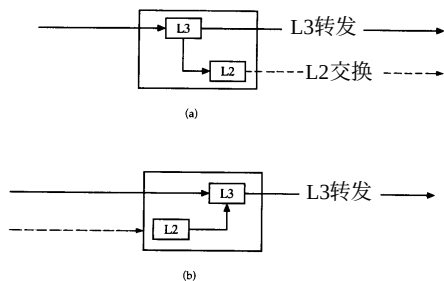


图6-2 IP交换

a) 入口 b) 出口

一个 IP 交换机并不是在真空中或者以独立的方式操作。为达到有效的目的，它必须在网络入口和出口处具有良好的功能，以使分组能够进入 IP 交换系统，在一条第三层路由路径或第二层交换路径上通过，然后离开。

6.1.3 直通路径

在常规的 IP 路由环境中，来自一个源网络的分组经过一系列路由器，被逐级跳转发到一个目的网络。与每一个路由器跳有关的是一些必备功能，如一个路由表的查询、分组头校验、TTL 减量、介质转换等等。这会经过路由路径的分组带来一定数量的累积时延和传输延迟。减少时延和传输延迟的办法是在可能的地方旁路中间路由跳。

在前面几节中提到的第二层交换路径称为一条直通路径。一条直通路径定义为一个旁路中间第三层路由跳，从源端到目的端的虚连接^①。一条直通路径可以在两个主机间，两个边缘设备（如路由器）间或在主机和路由器间建立。这全依赖于前面提到的入口和出口功能的放置，放在直通路径上的业务流的粒度及用于建立它的 IP 交换协议。

一条直通路径的基本特性是：

- 它旁路中间的第三层路由功能。
- 建立一条直通路径的触发系统可以基于数据业务流或控制业务流。数据驱动的直通路径可能导致在 IP 交换系统建立一条直通路径之前，在第三层路由一些数据业务流。控制驱动的直通路径基于控制业务流（如路由表更新）的出现而建立，从而所有的后续数据业务流将在直通路径上通过。

- 如果一个入口和出口间的直通路径不存在或突然消失，业务流仍然可以通过路由路径到达目的地。

- 它可以与路由路径（IP 交换机）遵循相同的物理通路（经过相同的端点和链路）或者它可以经过一个独立的第二层交换机拓扑（虚拟 IP 交换机）。对于虚拟 IP 交换机的情况，直通路径可以在拓扑结构上独立于 IP 路由路径（见图 6-1 的例子）。

① 一条直通路径由于它旁路或直连路由跳，可以区别于一条常规路径。常规路径是在同一个子网的两个设备间建立的第二层路径（如 ATM SVC）。

• 一个入口到出口的直通路可以在一条端到端通路上建立或者它可以通过连接多个更小的直通路来建立。

• 它可以起源于一个网络入口或出口或者根据在每个 IP 交换机上完成的当前本地直通路建立动作“无序地”开始。

• 它可以是点到点、点到多点或者多点到点。

• 它能够在路由路径上的可用功能之外提供一些加强的 QoS 或 COS。

非常清楚，有许多不同的标准和技术来建立通过一个网络的直通路。后面几节将讨论怎样和何时建立及管理直通路。

6.2 动力

IP 交换技术的发展受一些因素的推动，一些与商业有关，一些与技术有关。从技术上讲，由于 Internet 及许多内部网处于一定的或近似的拥塞状态，存在一个很大的动力发展更快和更便宜的 IP 转发设备。这并不奇怪，因为事实上许多 IP 网络设计上就有拥塞。提供一个能百分之百地处理处于突发速率的业务流的网络将过于昂贵，因此它们以相对少的容量被建立起来，期望最多的端用户在最多的时间里，通过网络获得其最多的业务量。如果发生拥塞，TCP 主机将发送速率适配到网络容量允许的大小，拥塞消除，人们可以继续他们的商务。当然，如果主机不使用 TCP（想想流动的音频/视频），并继续将分组注入一个已经阻塞的网络，而忽略或不理解隐含或显式的拥塞指示，问题可能不会消失。事实上问题可能更严重。

根据标准网络理论，性能瓶颈可以通过加快速度来减少或消除。提高数据分组的转发速率通常可以对网络性能起到一系列积极的作用。更快的传送速率意味着分组在队列中等待的时间会更少。队列将更快地清空，为瞬间的突发业务留出更多可用的队列容量。业务流的突发能得到更好的处理，不会由于队列容量不足而丢弃分组，造成重传或更坏的情况。整个网络的性能将得到改善。

现在事情并非真的这么简单。在任何网络总会有拥塞点，已经有建议提出，网络交换或路由结点在队列调度或管理中应起到更积极的作用。“拥塞忽略”主机将必须设法得到管制等等。但人们无疑将需要更多的带宽和更快的转发设备来处理业务量的增加并减少拥塞。

使用 ATM 技术被许多人看作是理想的解决方案，以回避与当前路由器（大约到 1997 年）技术相关的性能瓶颈。一个标准的 ATM 交换机提供多吉比特（multigigabit）转发速率，比现有的路由器快几个数量级。ATM 也提供极好的性能价格，且端口速率提高到 OC12 或更高，是十多年中工业研究和标准发展的一个主要焦点。

无论如何，在 IP 路由的 Internet 环境中使用 ATM 技术对那些开始研究 IP 交换解决方案的人来说是最可取的。IP 交换出现的技术原因包括下列内容：

• 路由器是瓶颈，不能支持 OC3 + 转发速率的业务量。

• 路由表太大，查询时间太长。如果许多路由被映射到一个更小的标签组（如 VPI/VCI），且一个查询包括在硬件中索引 VPI/VCI 表，则会有所改善。

• 下一代 IP 网络和应用需要 ATM 所能提供的带宽管理、性能和 QoS。

• 需要一个简单的方法在一个面向连接的网络中支持无连接 IP 业务流。

• 需要一个简单的方法为需要直通路的应用构造直通路。

• ATM 论坛的 UNI/PNNI 信令和路由被认为过于繁琐。需要一个更简单的东西以符合“IP

中间派”的口味。

- ISP核心网络已经或正计划开发 ATM。需要一个简单且可扩展的方法在 ATM面向VC的网络上支持大数量的IP业务流。

- 需要开发平台和网络以支持正在兴起且先进的 IP功能，如IP综合业务（RSVP）和流量工程（显式路由）。

毫无疑问还有其他原因。但 IP交换发展背后的主要技术推动力是这样—一个愿望，即通过使用交换特别是ATM交换作为一些或者所有业务量的转发机制从而有效地提高性能。

如今在网络工业做任何新事情的商业原因是模糊的，并且隐藏在市场和机构，或投机资本家公司发出的华丽的、夸大的宣言之后。根据一些后见之明的想法，不难认识到一些确立的交易者和启动者在IP交换技术上投资这么大并继续这样做的理由：

- 确认在ATM上的投资。
- 重定位ATM作为一个“IP友好”的技术。
- 采用提供IP路由和高容量ATM交换的综合平台，开发实际的及认识到的 Internet和内部网的性能问题。
- 提供新技术与已有的路由器基制竞争。
- 提供新技术以迷惑市场并通过模糊其缺点来保护路由器基制。

谁会确实知道呢。但这个生意的目的是赚钱，有些人赚到了，有的人还没有。很有可能，各种形式的IP交换将成为公共网和专用网的运营者因其增加的容量和特殊业务而开发的一个非常有用的技术。

6.3 IP交换的寻址模型

前面的章节讨论了IP和ATM寻址。一个采用ATM技术支持IP交换业务的网络必须至少支持IP地址空间。对于具有特定源点或目的地的IP分组，IP地址可以绑定到一个ATM地址，或者一个IP地址的组合可以映射到一个连接（借助于用一个VPI/VCI标志业务流）。因此，有两个基本的IP交换寻址模型：独立的和IP到VC。

6.3.1 独立寻址

独立寻址模型规定IP和ATM地址空间都应保持在网络中。因此，与ATM网络相连的一个路由器或主机必须装配一个IP地址和一个ATM地址。这与任何装配一个IP地址和一个MAC地址的连接到LAN的IP主机没什么不同。

除非一个PVC已经事先建立，一个源IP主机必须知道它要与之通信的目的IP主机的ATM地址。这个IP到ATM的地址解析是需要的，以使源IP主机可以向目的IP主机发起一个SVC连接请求。一个源IP主机可以通过配置或一些其他手段来了解这个信息，或者它可能必须查询一个保存IP和ATM地址映射表的地址解析服务器。虽然有可能一个单一的路由协议可以两者都支持（如I-PNNI[综合PNNI]），实际上需要独立的路由协议来通告IP和ATM目的前缀。

独立寻址模型有下列特性：

- 每一个连接到ATM的IP可寻址设备必须装配一个IP和ATM地址。
- 在这种环境中，典型地支持用于IP（如OSPF）和ATM（如PNNI）的独立的路由协议。
- 实际上，IP到ATM地址的动态解析需要一个额外的向服务器查询的机制，这可能给建立

一条常规路径（如同一子网的两个主机间）或一个源点和目的地间的一条直通路程增加时间延迟。

独立寻址用于下面描述的IP交换叠加模型。MPOA是一个使用独立寻址的IP交换系统的例子。

6.3.2 IP到VC

IP到VC(IP-to-VC)这一术语是一个尝试，用来描述一个仅支持IP地址的寻址模型，并且根据分组头中IP层信息的内容，分组被放在一条独立的直通路程，一个VC上。在这个模型中，不使用ATM地址来帮助IP交换的通信。后面几节描述的独立的控制协议用来将IP分组映射到直通路程上，而不使用ATM地址和ATM论坛协议。然而，映射过程使用ATM VPI/VCI值作为一个连接IP层信息与一条VC的途径。

IP-to-VC寻址模型的特性是：

- 保持一个单一的IP地址，不需要或不使用ATM地址。
- 使用一个单一的IP路由协议来通告IP地址前缀。
- 使用专门的IP-to-VC映射协议将IP分组与一条直通路程相关联。

IP-to-VC寻址模型被所有对等的IP交换实体使用且其中没有ATM论坛协议，或被用来帮助建立一条直通路程。使用IP-to-VC寻址模型的一个例子是运行IFMP和GSMP的IP交换机。

6.4 IP交换模型

IP交换有两个基本模型，它们的不同在于使用或不使用ATM论坛协议和组件包括IP交换机或虚拟IP交换机。这两种模型称为对等和叠加模型。

6.4.1 叠加模型

IP交换的叠加模型包括运行在一个独立的ATM层之上的一个IP层。换句话说，它由运行IP路由协议、具有IP地址的IP设备和运行ATM信令及路由协议、具有ATM地址的ATM设备（IP主机、IP路由器、ATM交换机等）组成。由于IP和ATM单元都可以利用且运行大部分基于标准的协议，叠加模型可能是最容易实现的。但在两个地址空间内有些叠加功能必须加以维护且必须支持两个路由协议。叠加模型的一个例子是MPOA。

叠加模型的特性是：

- 它使用独立寻址。
- 在IP（如OSPF）和ATM（如PNNI）上运行独立的路由协议，意味着要维护两个独立的拓扑结构且一个不了解另一个。例如，运行OSPF的IP路由器了解IP网的拓扑结构但看不到或了解ATM交换机。
- 如果SVC用于常规或直通路程的建立，则需要IP和ATM间的地址解析以及用户网络接口（UNI）/PNNI信令和路由。
- 通常使用虚拟IP交换机。另一个例子是连接到一个ATM骨干网边缘的一簇IP主机和路由器。
- 支持缺省的路由和直通路程。
- 对于在其他非广播多接入（NBMA）技术^①上的IP有同样的特性。

① NBMA网络的例子是ATM、帧中继和X.25。

图6-3描述了一个叠加模型的例子。IP与ATM连网设备同时存在意味着独立寻址。缺省路由路径是通过两个路由器跳，IP控制点1和2的一条路径。一条旁路路由器跳的直通路可以在所示的入口和出口间建立。一旦入口确定了出口的ATM地址，它就可以使用标准的ATM论坛信令和路由建立一条直通路经过ATM交换机1、3和4。注意置于直通路上的分组与遵循路由路径的分组没有经过相同的设备和链路。这是因为ATM路由协议（如PNNI阶段I）根据ATM交换机的拓扑状态而不是IP的拓扑结构^①为入口到出口的SVC建立请求选择一条路径。

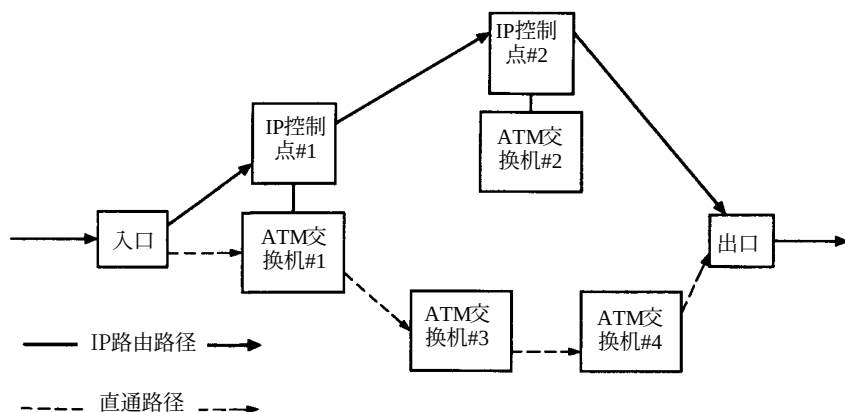


图6-3 叠加模型

6.4.2 对等模型

IP交换的对等模型规定，IP交换机组件维护一个单一的IP地址空间并支持一个单一的IP路由协议。对等模型也意味着用于将IP业务流映射到直通路上的一个独立控制协议的存在。对等模型的一个例子是运行IFMP和GSMP的IP交换机的一个网络。

对等模型的特性是：

- 维护一个单一的IP地址空间。
- 具有一个单一的IP路由协议。
- IP交换机使用特殊控制协议将IP业务流映射到直通路路上。
- 它支持缺省的路由路径和直通路。

图6-4描述了对等模型的一个例子。由于不存在运行ATM论坛协议的ATM交换机组件，意味着使用IP-to-VC寻址。注意路由路径和直通路经过相同的设备，分别是IP交换机1和2，且路由路径和直通路业务流经过相同的链路。这是因为一个单一的IP路由协议将基于一个单一的IP网络拓扑结构计算出到一个目的地的最好路径。IP交换对等

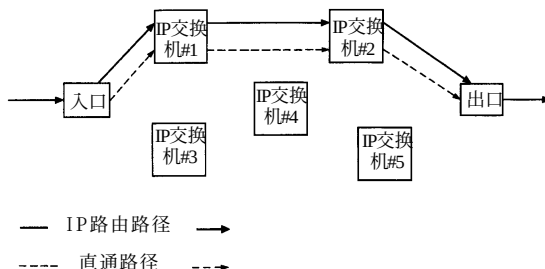


图6-4 对等模型

① 这对于需要支持一个能够利用拓扑结构中所有链路和结点的高容量骨干网的网络提供者来说可能是一个非常吸引人的性能。注意IP路由协议在选择一条路径通过网络时，不考虑可用网络容量、延迟和其他QoS参数。

模型与叠加模型之间的一个总结性比较在表 6-1 中描述。

后面几章将要详细论述的两个重要项目是路由器邻接和 VC 消耗。在叠加模型中，连接到一个大 ATM（或其他 NBMA 网络）网络的路由器可能需要相互对等地交换路由协议更新信息以得到一个对 IP 网络拓扑结构精确和及时的描述。支持大量对等连接的路由器将必须处理相当多的路由协议更新信息。这还不包括路由器已经相互建立的任何数据连接。在对等模型中，IP 交换机只需要与它们的邻接交换机对等交互，如同一个路由器网络中相互间的对等交互。

表6-1 叠加与对等之比较

属 性	叠 加	对 等
寻址	独立的(IP和ATM)	单一的(IP), 使用直接的 IP-to-VC 的映射
路由协议	IP和ATM都需要	只有IP
ATM论坛协议	有	无
专用的IP到直通路径协议	无	有
地址解析(IP-to-ATM)	有	无
直通路径与IP路由路径的拓扑结构相同	无	有
IP交换组件	虚拟IP交换机、路由器、主机	IP交换机、路由器、主机
例子	MPOA、连到ATM网络的路由器	运行IFMP/GSMP、ARIS等的IP交换机

另外，与一个大 ATM 网络相连的路由器自然希望相互间建立直接 VC，因为它提供了最好的性能。一个连接所有路由器的全连接网会消耗大量的（ $O(N^2)$ ）VC 资源。一些对等模型的实现将全连接网的连接所需要的 VC 资源消耗最小化。ARIS 引入的一个技术称为 VC 合并；它将多个上游 VC 合并为一个单一的下游 VC。这极大地减少了必须穿过 ATM 网络来建立并由每个路由器支持的 VC 数目。

6.5 IP交换类型

建立一条直通路径的标准和置于直通路径上的业务流的分类标准依赖于 IP 交换解决方案和协议的类型。一般而言，有两种 IP 交换解决方案：流驱动和拓扑驱动。以下内容讨论这两种类型的基本操作和结构。

6.5.1 流驱动的方案

IP 交换的流驱动模型在一个特定 IP 流上操作，此 IP 流定义为具有相同源 IP 地址，目的 IP 地址和端口号的一个分组序列。流驱动的 IP 交换的基本操作在图 6-5 中的概念图中描述，包括以下步骤：

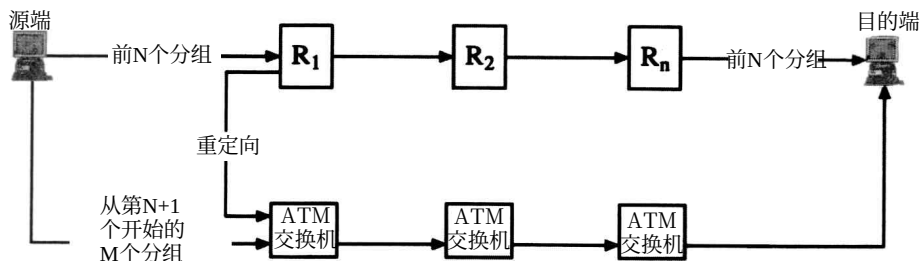


图6-5 流驱动的IP交换

• 一个流的前 N 个分组经过一个或多个IP路由实体 ($R_1, R_2, \dots, R_n$) 被逐级跳路由至目的地。IP路由实体是一个IP交换机或虚拟IP交换机的路由组件, 或者也可以是由ATM连接起来的实际路由器。

• 基于IP流的特性 (如业务流类别、端口号、源/目的IP地址、到达速率等等), IP路由实体 (在网络的边缘或中间) 触发一个重定向^①过程的启动。重定向过程包括建立一条直通路程并将IP流重定向到直通路程所需的动作, 从而旁路逐级跳路由器处理 ($R_1, R_2, \dots, R_n$)。这通过给属于IP流的信元分配一个新VPI/VCI标签并更新ATM交换机^②内的VPI/VCI连接表来完成。

• 一旦IP流被重定向到直通路程, 流中剩余的从第 ($N+1$) 个开始的 M 个分组被交换。

流驱动的IP交换有几个有趣的性质。首先, 它可以也应该不必交换所有的流。一些如ICMP “pings” 和尽力而为的email可能不会也不能保证从一个更快的交换路径获得任何好处。其次, 每个有资格在一条直通路程上通过的流的重定向过程单独完成。使一个特定的流有资格成为被交换的候选者的条件依赖于特定实现的策略和推断。没有一个标准说, “这些分组头字段所标志的这个流将被交换”。

最后注意, 流的前几个分组被路由通过网络, 只有重定向过程完成后, 流“尾”才在直通路程上通过。

流驱动的IP交换的基本功能特性是:

• 特定IP流的出现可以触发一条直通路程的建立, 流被定义为应用到应用或主机到主机的业务流。一个流由共享公共IP分组头信息, 如源和目的IP地址及TCP或UDP端口号的一个分组序列所标志。

• 分组可以被路由通过网络或置于在每个流基础上的一条直通路程上。

• 将放置于一条直通路程上的每个流的重定向过程独立完成。重定向过程的范围可以限于一对邻接IP交换机, 或者它可以是网络入口和出口间端到端的。

• 每个IP交换机可以为每个流单独完成重定向过程。换句话说, 重新标志一个流及由此更新一个IP交换机连接表的决定不依赖于另一个IP交换机对同一个流所进行的操作。

• 基于IP交换系统中实现的直通路程决定标准和策略, 一些流可能没有资格被交换。

• 一个基于流的入口组件可以将一些流置于一条直通路程上并路由其它的流。

• 如果一条直通路程不存在或突然消失, 分组可以被路由至目的地。

• 通常不是全部流都被交换。前面一些分组被路由以使IP交换机有一个机会分类出一个流并决定启动重定向过程。直通路程建立起来后, 流中剩余的分组被交换。

• 直通路程是典型的软状态, 这意味着如果不刷新, 它们将超时。

操作一个流驱动的IP交换解决方案会带来一些额外开销。包括在每个流的基础上, 由网中组件参与的交换和处理重定向或流建立 (flow-setup) 的消息。还有, 方案中的交换单元必须为每个将要被交换的流分配每个连接 (per-connection) 资源 (如标签、缓冲区)。大量有交换资格的流的到达可能会给路由常规业务量和建立直通路程带来一些时延, 也会将资源制约强加给交换机资源。

一个流驱动的IP交换解决方案在可能有大量单独IP流的一个大网的核心也许不会很好地

① 重定向过程的其他名字是流建立和流映射。

② 图中显示只有 R_1 与一个交换机组件通信。实际上, 与一个交换机组件相关的每一个IP路由实体可以独立操作, 通过交换机组件重定向分组并更新交换机的连接表。

运作。离开核心部分，源端和目的端共享相同位置的概率更高，使得这一方案更可配置和应用。特别是位于一个校园或内部网的单个主机对可以有一个或多个应用从被交换中受益。随着向端用户的靠近，IP流的数量没有像在核心处那样集中。这意味着有更多的交换机资源可利用，以处理位于核心同一侧的主机对之间的直通路径（即它们不必经过核心便可相互通信）。

流驱动的IP交换协议的例子包括IFMP和流属性通知协议（FANP / Flow Attribute Notification Protocol），将在后面章节详细论述。

6.5.2 拓扑驱动的方案

拓扑驱动的IP交换基于运行一个IP交换机里的IP路由实体中的路由协议（如OSPF、BGP）所维护的IP网络拓扑结构。与反映一个目的网络的目的IP前缀相关的新VPI/VCI标签被产生和分发至路由区域的其他IP交换机。所有指定到一个特定目的网络的业务流遵循一条基于此新VPI/VCI标签的交换路径。

拓扑驱动的IP交换的基本操作在图6-6中的概念图中描述，包括以下步骤：

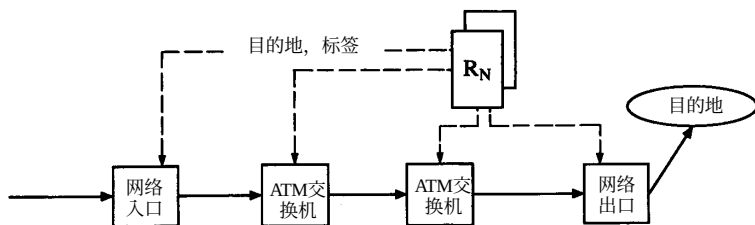


图6-6 拓扑驱动的IP交换

- 基于IP路由实体（ $R_1, R_2, \dots, R_n$ ）间路由协议消息的交换，IP交换机汇聚于一个网络拓扑结构。

- 与采用{目的前缀、标签}格式的目的IP前缀相关的新VPI/VCI标签被产生和分发至路由区域中IP交换机的交换组件。

- 入口IP交换机检查进入网络的分组的目的前缀。入口IP交换机将分组放在一条到网络出口的交换路径上而不是将分组转发至下一跳IP地址。所有到那一目的地的业务流在一条从网络入口到网络出口的交换路径上通过。

- 出口IP交换机接收交换路径上的分组并在第三层将其转发至目的地。

拓扑驱动的IP交换方法与流驱动方法相比，提供了一些在性能和可扩展性方面的增强。首先，含有一个或多个去往一个公共目的网络的流的所有业务流均被交换。换句话说，全部流均被交换而不是像流驱动的IP交换那样仅交换后面的流。其次，由于仅在拓扑结构有变化或控制业务流到达后才建立交换路径，拓扑驱动的IP交换的运行时间开销更低。如果拓扑结构是稳定的，交换路径已经建立起来，业务流将在其上通过。从可扩展性的角度看，拓扑驱动的方法更好。这是因为一个路由区域的交换路径的数目与路由器的数量或网络的大小成比例。另外，通过建立通向一个目的网络共享集合的多点到点直通路径（如使用VC合并），一个更高级别的汇聚是可能的。

当然，适用于所有流的更好的扩展性和性能特点会带来一定的花费。拓扑驱动的IP交换，不论业务流在其上通过或不通过，均建立直通路径。这意味着由于控制业务流（如路由协议更新）而不是数据业务流的结果，交换机资源被消耗。要考虑的另一种情况是，可能很难在

一条拓扑驱动的直通路经上通过的一个流提供 QoS。这是因为，可能有实时和尽力而为 IP 流的混合在一个拓扑驱动的尽力而为的直通路经上通过，且指定一个公共目的网络上的资源。当然，尽力而为的流不会在意，但实时业务流可能会获得不尽满意的性能。最后，直到一个网络汇聚，不准确的路由表的更新会引起瞬时循环的形成，导致形成一个交换循环的一组标签的产生。IP 路由器可以通过减量 TTL 字段至零并丢弃分组来限制瞬时循环引起的危害。第二层（如 ATM）交换机可能没有这样的机制，所以这种情况下的一个循环可能是成问题的。

拓扑驱动的 IP 交换的基本特性是：

- 直通路经基于目的网络的存在而建立，且仅在拓扑结构发生变化后才建立。
- 所有通向一个目的地的业务流均被交换。
- 入口 IP 交换机接收一个新的每一目的（per-destination）标签（VPI/VCI）。执行常规的 IP 转发，但不是将分组定向到下一跳 IP 地址，而是由入口 IP 交换机添加新标签^①并将分组置于一条直通路经上。

- 在每个中间交换组件上，分组（或信元）被交换标签。
- 出口 IP 交换机从交换路径上接收分组，去掉标签，接着使用常规第三层过程转发分组。
- 如果一条直通路经突然消失，分组可以被路由至目的地。
- 它可能在第二层引起瞬时循环。
- 拓扑驱动的直通路经保持其连接直到拓扑结构发生变化。

拓扑驱动的 IP 交换解决方案是标记交换和 ARIS。这些将在后面章节继续讨论。

6.6 IP 交换分类

本章介绍了 IP 交换的基本概念。定义了寻址模型，IP 交换模型及其两种类型，流驱动和拓扑驱动。更多的将在后几章介绍，包括那些使用 ATM 作为同一网络中两个主机间一个连接载体的解决方案。

到目前为止介绍的定义和概念为 IP 交换和相关技术的初步分类提供了依据，如图 6-7 所示。在对等模型分支的下面是两种可能的 IP 交换类型，流和拓扑。与每个类型相关的是一些特定实现，将在第二部分详细讨论。

在叠加模型分支的下面，有一个流驱动解决方案，MPOA，将在后面讨论。在这个分类中有一个独立的分支称为地址解析（AR）。基于 AR 的 IP 交换解决方案是那些将一个数据网络地址（如 IP、IP 组播、MAC）解析为一个 ATM 地址的解决方案，以达到使用 ATM 的 SVC 作为端点间一个通信手段的目的。尽管解决方案 NHRP 试图旁路中间路由跳，但基于 AR 的解决方案通常局限于一个局部子网。另外，基于 AR 的 IP 交换必须调用地址解析功能并建立 SVC，以在一个更实际的方式下工作。这与用于在一条已经存在的路由路径上提高转发速率的 IP 交换有根本的不同。

这个初步 IP 交换分类也可以作为本书其余部分所讨论话题的一个路线图。

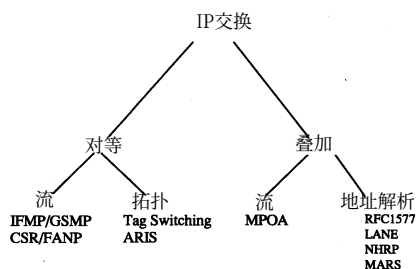


图6-7 初步IP交换分类

① 如果使用 ATM，入口 IP 交换机将分组分段成信元并应用适当的 VPI/VCI 标签。

更多的信息

Callon, R., “ Multilayer Switching including MPLS and Related Approaches, ” tutorial notes, Next Generation Networks 1997, November 1997, Washington, D.C.

[http://infonet.aist-nara.ac.jp / member/nori-d / m1r/](http://infonet.aist-nara.ac.jp/member/nori-d/m1r/).

Metz, C., “ MSS and IP Switching, ” IBM Whitepaper, May 1997.

Petrosky, M., “ Short-cut Routing, ” The Burton Group Network Strategy Report, February 1997.

White, P., “ ATM Switching and IP Routing Integration: The Next Stage in Internet Evolution, ” IEEE Communications, April 1998.

China-pub.com

下载

第7章 ATM上的经典IP解决方案

目前IP、交换和IP交换都已经涉及到，是回过头来重温 ATM上的经典IP(Classical IP over ATM)技术的体系结构、协议和操作的一个好时机。ATM上的经典IP是准确地为RFC1577中描述的功能赋予的一个名字，ATM上的经典IP和ARP在后来修改的RFC2225中也有描述。本章将把这个词应用到一个范围更广的定义，这样不仅涉及 RFC1577/2225，还包括IP over ATM协议的姐妹序列。除了 RFC1577，经典序列还包括 NHRP和组播地址解析服务器（MARS）。它们是经典的，意义在于它们支持传统的IP业务和行为，且对于运行其上的应用是透明的。

7.1 体系结构

经典IP/ATM (classical IP/ATM)[⊖] 技术的体系结构结合了前面所讨论的许多内容。共享一个公共地址前缀的IP主机和路由器被组合为与一个物理ATM基础结构相连的子网。IP是网络层协议，为穿过网络边界的分组提供一个无连接的逐级跳转发业务。路由器基于路由更新消息的交换来维护转发表并从一个网向另一个网转发IP分组。同一子网的IP设备使用ATM VC直接通信。此标准行为的一个例外是NHRP提供的使不同子网的IP设备能够在ATM网络上直接通信的功能。

在ATM层，ATM论坛协议维护一个一致的ATM网络拓扑并支持ATM端点间PVC和SVC的建立。IP主机和路由器被分配IP地址，连接到ATM的设备和交换机被分配ATM地址。由于IP和ATM的地址和拓扑是不同的，ATM上的经典IP协议使用独立寻址。事实上，除了NHRP，的确没有任何努力把ATM当作在同一网络通信的IP端点间的一个第二层传送以外的任何手段。

不过，IP必须进行调节以在一个ATM网上通信。首先，IP分组必须在AAL层进行适配，从而在源端分段成信元并在目的端组装成分组。因此，需要一个标准技术将IP分组封装成AAL帧。其次，ATM使用预先配置的PVC或者动态建立的SVC将数据从一个源端传送到目的端。在后一种情况，源端必须知道目的端的ATM地址。因此，一个想要在一条SVC上与一个目的IP设备通信的源IP设备必须在连接建立之前将目的IP地址解析为一个ATM地址。在一个LAN环境中，源IP主机发出ARP广播并被传送到目的IP主机，目的IP主机接着发送一个包含目的MAC地址的ARP响应。ATM不是一个广播介质，所以需要一些其他技术进行地址解析。

当通信模式是ATM SVC时，ATM上的经典IP协议都使用一个地址解析服务器（ARS）。ARS的功能是维护IP地址和相关ATM地址[⊗]的一个表或高速缓存，并响应来自源IP设备的想把一个IP地址解析为一个ATM地址的请求。ARS可以通过手工配置、登记或者通过被动地检查包含ATM地址绑定的数据或控制分组来设定。IP设备将与ARS一起保持一条控制VC并与其他IP设备建立一条或多条SVC来交换数据。当然，如果使用PVC，则不需要动态地址解析且

⊖ ATM上的经典IP、经典IP、CIP和CLIP都是用来表示最初在RFC1577定义并在RFC2225中修改的协议。为拓宽定义，本章将使用一个小“c”经典IP/ATM来描述包括RFC1577（和它后面的RFC2225）、NHRP、MARS及服务器高速缓存同步协议（SCSP / Server Cache Synchronization Protocol）的协议族，它们均源于IETF的NBMA（ION）上的互连工作组，且都支持IP并需要一个地址解析服务器或ARS（SCSP是例外）。

⊗ 这些称作IP-ATM地址绑定，记为{IP.1, ATM.1, IP.2, ATM.2,...IP.n, ATM.n}。

ARS的使用也不是必须的。

图7-1描述了ATM上的经典IP协议的通用体系结构。在源端和目的端，应用和TCP/IP分别位于一个AAL和ATM层之上。源和目的设备（如主机、路由器）装配有一个IP地址和一个ATM地址。一个源IP设备将向ARS询问一个IP/ATM地址绑定以便与目的设备建立一条SVC。如果使用PVC，ARS则不需要。

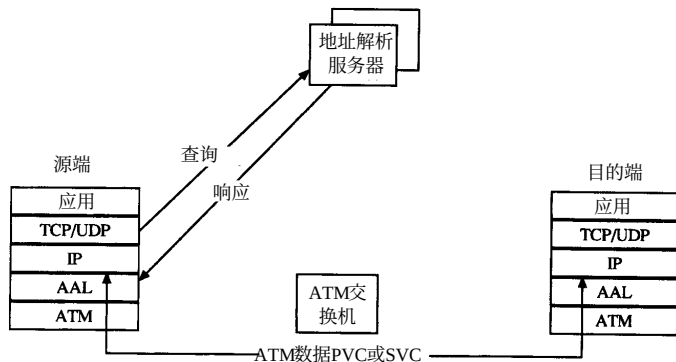


图7-1 ATM上的经典IP体系结构

7.2 RFC1483

IP必须以分组正在其上传送的当前数据链路的格式封装，对于ATM也是这样。RFC1483中描述的ATM上的多协议封装定义了一个封装IP分组的标准技术及更通用的，以AAL5帧的格式在一个ATM网络上传送的多协议数据分组。RFC1483定义了两种格式的封装：逻辑链路控制（LLC）/SNAP和VC复用。

7.2.1 LLC/SNAP封装

RFC1483中定义的第一种技术称为LLC/SNAP封装。它通过在每个包含一个IP分组或任何其他网络层协议分组的AAL5帧之前添加一个8字节的LLC/SNAP帧头来工作。LLC/SNAP帧头识别是否一个帧被路由或桥接及如果被路由，是哪一个网络层协议（如IP、IPX）被路由。包含一个IP分组的典型的LLC/SNAP封装的AAL5帧在图7-2中描述。所显示的OUI和以太类型（Ethernet type）构成SNAP的帧头。IP的Ethernet值是x'0800'。对于不同的协议可能有其他的值。

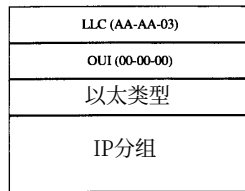


图7-2 RFC 1483 LLC/SNAP封装的帧头

由于协议实体包含在每一个AAL5帧中，LLC/SNAP封装使得多个协议能够分享同一个ATM VC。这把支持一个ATM网络中多协议业务流所需的VC数目减到最小。如果有成本、资金或其他的问题与网络中VC的数目有关，这也许是一个因素。另外，不依靠ATM信令在一个协议特定实体中识别协议和终结VC。否则，就得由ATM系统通过为每个协议建立一条VC来复用不同的网络层协议。相反地，由LLC/SNAP帧头及每个VC端点中的一个LLC实体来识别协议实体，并且随着数据被传递到应用，处理数据的复用和解复用。较低的VC消耗和信令开销之间的平衡是添加给每个AAL5帧的额外的8字节LLC/SNAP帧头。

图 7-3描述了 ATM上的 IP 数据的 LLC/SNAP封装的概念。当通信 ATM设备在 PVC或SVC上交换多协议数据业务流时, LLC/SNAP封装是有用的。这种情况的一个例子是由一条 PVC连接的两个路由器交换 IP 和IPX业务流。LLC/SNAP是RFC1577/2225中使用的缺省封装技术, 而且事实上在大多数ATM上的IP 解决方案中使用。

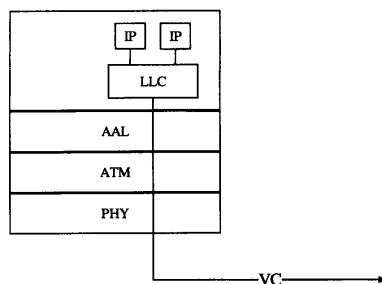


图7-3 RFC1483 LLC/SNAP封装

7.2.2 VC复用

RFC1483中定义的第二种技术称为 VC复用或空封装。在这种方式中不需要 LLC/SNAP帧头。网络层协议在终结于每个 ATM端点中的一个网络层实体的独立 VC上传送。网络层协议实体可以手工设置或由连接建立阶段的 ATM信令动态建立。网络层协议的复用和解复用通过每一协议 (per-protocol) VC的存在, 在ATM层隐含地完成。

VC复用在降低开销和提高性能方面提供了比 LLC/SNAP封装略大的优势。这是因为每个 AAL5帧不需要8字节的LLC/SNAP帧头, 且 VC直接终结于一个网络层协议实体。然而, 这种优势可以被较高的 VC消耗和支持一个多协议信令环境所需要的信令开销所抵消。如果VC资源丰富且建立起来价格低廉或者如果IP是唯一的网络层协议, 那么 VC复用是较好的选择。

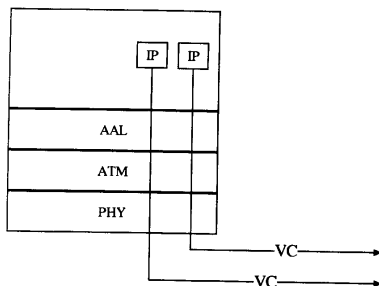


图7-4 RFC 1483VC复用封装

RFC1483中所描述的 VC复用的概念在图7-4中描述。

7.3 ATM上的经典IP和ARP

共同用ATM上的经典IP和ARP (Classical IP and ARP over ATM, CLIP) 这一术语表示的功能和过程在RFC1577和后来的RFC2225中描述^①。CLIP描述了为位于同一个子网 (称为一个 LIS) 的IP设备间建立点到点的 ATM虚连接的机制, 而且提供了一个在 LIS成员间跟踪和分发 IP/ATM地址绑定的业务。CLIP的宗旨不是在一个已经存在的网络拓扑如一个 LAN上仿真IP和ARP的操作, 而是使 ATM能够为在同一子网的两个 IP设备间的数据交换起到“线”的作用。子网间的通信保持不变且仍然要求分组经过一个路由器。

7.3.1 组件

一个支持CLIP连接的网络的组件包括下列内容：

- LIS。这是连接到一个物理 ATM网络的主机和路由器的一个集合, 它们共享同一 IP地址前缀或子网号。一个LIS中的任何两个设备应该能在一条点到点 ATM PVC 或SVC上直接通信。不同逻辑IP子网中的设备必须通过一个路由器通信。一个路由器可以连接到一个或多个 ATM

① 除非特别标注, 本章讨论的 CLIP功能和过程从属于RFC1577中所列的那些内容。

LIS以及包括其他介质（如以太网）的子网上。

- **ATMARP**。这定义了一个 ATMARP 客户机和服务器间使用的地址解析协议和消息格式。除了它已被修改用于承载 ATM 地址并在一个面向连接的 ATM 网络上工作以外，ATMARP 在概念上与传统 IP ARP 相似。另外由于 ATM 不支持纯广播，ARP 地址解析功能由一个“第三方”、ATMARP 服务器完成。

- **ATMARP 服务器**。ATMARP 服务器是一个 LIS 成员的 ARS。ATMARP 负责维护 LIS 成员的一个 IP/ATM 地址绑定 {IP.n, ATM.n} 表，并响应来自 LIS 客户机的 ATMARP 查询。每个 LIS 至少需要一个 ATMARP 服务器。ATMARP 服务器功能可以驻留于一台工作站、一个路由器或一个连接到 LIS 的独立设备中，且只有 SVC 被支持时才需要它。一个 ATMARP 服务器不转发任何数据，它仅仅处理控制数据。

- **ATMARP 客户机**。一个 ATMARP 客户机功能驻留于每个想要与同一 LIS 中的另一个设备建立一条 SVC 的 LIS 成员中。ATMARP 客户机与 ATMARP 服务器联系并交换 ATMARP 消息来实现 IP/ATM 地址映射。ATMARP 客户机的另一个名称是 RFC1577 客户机。

应当注意 ATMARP 的控制功能和在 VC 上的数据交换限于一个单一的 LIS。在 ATM 上的经典 IP 和 ARP 这部分，没有进行任何尝试来旁路或直连中间路由跳。即使在一个单一的物理 ATM 基础结构上叠加着多个 LIS，情况也是这样。ATM 上的经典 IP 和 ARP 没有给传统的端到端 IP 互连模型强加任何改变。

7.3.2 地址解析

一个 CLIP 网络的基本操作包括地址登记、地址解析、SVC 建立和数据交换^⑨。图 7-5 的上面部分显示了连接到一个 ATM 交换机的两个主机和一个 ATMARP 服务器共同构成一个 LIS。如 RFC1577 中所描述的，主机 1 到主机 2 建立一条 SVC 所需的消息交换包括如下内容：

- 支持 RFC1577 的 CLIP 客户机首先向 ATMARP 服务器登记它们的 IP 地址。这通过与 ATMARP 服务器进行 InvATMARP 消息的一个交换来完成。当 ATMARP 服务器寻找连接到 ATM 的 IP 主机（它也可以是 LIS 中的一个路由器）的 IP 地址时，它首先发出一个 InvATMARP 请求（InvATMARP request）。IP 主机用一个包含它的 IP 地址的 InvATMARP 应答（InvATMARP reply）来响应。接着 ATMARP 服务器将主机的 IP 和 ATM 地址存在一个表中达 20 分钟。到 20 分钟结束时，除非由主机或 ATMARP 服务器通过一条已存在的 VC 进行刷新，IP/ATM 地址绑定将过期。InvATMARP 过程在 IP 主机和 ATMARP 服务器间建立的一条控制 VC 上进行。控制 VC 由必须装配 ATMARP 服务器的 ATM 地址的 IP 主机启动。

- LIS 中的一个想要与同一 LIS 中的另一个设备建立一条 SVC 的源 IP 主机必须确定目的地的

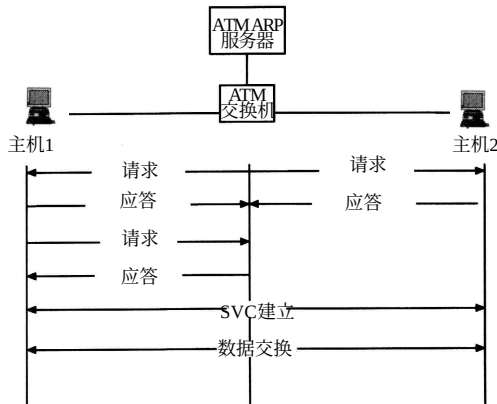


图7-5 RFC 1577地址解析

⑨ 大部分或全部 RFC1577 网络在 LIS 成员间使用 SVC。如果网络只使用 PVC，地址解析过程需要使用反向 ATMARP 机制来将一个 PVC 端点解析为一个 IP 地址。

IP和ATM地址。要这样做，源IP主机ATMARF客户机将在控制VC上与ATMARF服务器交换ATMARF消息。发向ATMARF客户机（图7-5中的主机1）的ATMARF应答包括目的地（主机2）的IP/ATM地址绑定。ATMARF消息的基本格式在图7-6中描述。

- 源IP主机现在使用标准ATM论坛协议建立一条到目的IP主机的SVC。

ATMARF消息被封装在一个LLC/SNAP帧头，且x'0806'这一以太类型定义了一个ARP过程。ATMARF消息中应注意的重要字段是：

- 操作码。定义了ATMARF消息的类型。其值为：

- ☐ ARP_Request = 1
- ☐ ARP_Reply = 2
- ☐ InARP_Request = 8
- ☐ InARP_Reply = 9
- ☐ ARP_NAK = 10

- 源ATM地址。源主机的ATM地址。
- 源协议地址。源主机的IP地址。
- 目的ATM地址。目的主机的ATM地址。
- 目的协议地址。目的主机的IP地址。

CLIP过程可以在一些方面得到改进以获得更高的效率和可靠性。这样的改进在ATM上的经典IP和ARP的新版本，RFC2225中描述。现在，由ATMARF消息的一个交换完成在ATMARF服务器上登记主机的ATM地址的任务。为了登记，IP主机发送一个在目的IP地址字段包含它自己的IP地址的ATMARF请求（ATMARF request）。ATMARF服务器在它的高速缓存中保存源IP和ATM地址并在ATMARF应答（ATMARF reply）中将那些值返回给IP主机。这克服了InvATMARF过程用于登记的不必要开销。在一些情况下，InvATMARF过程可能消耗不必要的资源，并且也不需要向ATMARF服务器登记一个IP主机的IP/ATM地址绑定。考虑用一个多协议路由器作为一个ATMARF服务器运行的情况，根据定义，它必须发送InvATMARF请求到LIS中的所有设备——即使他们不支持IP。一个成员的IP和ATM地址登记是能够简单地采用ATMARF消息交换来完成的。

另一个改进的方面是ATMARF服务器的可用性。如果LIS中单个的ATMARF服务器失效，则没有一个备份可用。服务器高速缓存同步协议（SCSP / Server Cache Synchronization Protocol）解决了多服务器问题，这将在本章的后面部分进行讨论，新CLIP客户机现在可以配置一个ATM地址表，代表服务于LIS的一个或多个ATMARF服务器。如果一个服务器失效，客户机可以与表中的下一个联系。

最后，CLIP只支持IP单播通信。虽然这不会变，CLIP能够通过在一个ATM网上提供IP组

硬件类型		协议类型	
T/L源ATM地址	T/L源ATM子地址	操作码	
L源协议地址	T/L目的ATM地址	T/L目的ATM子地址	L目的协议地址
源ATM地址			
源ATM子地址			
源协议地址			
目的ATM地址			
目的ATM子地址			
目的协议地址			

图7-6 ATMARF分组格式

播和广播支持的MARS功能来扩充。

7.3.3 ATM上的经典IP和ARP网络实例

CLIP是实现ATM上的IP的一个简单而有效的技术。它是一个很好的解决方案，主要是因为IP主机和设备能够从ATM的高带宽和低延迟上受益。在高容量高带宽网络上运行得最好的IP单播应用尤其是一个在CLIP网络上运行的极佳候选者。

图7-7描述了一个CLIP网络的例子并阐明了有关其操作的几个重要特性。首先，一个LIS的所有成员能够在ATM网上直接通信。不要求一个LIS的成员物理上连至同样的或邻近汇集的ATM交换机——只要它们共享同一IP地址前缀和同样的ATMARF服务器，且能够与ATM上的所有其他LIS成员通信。这一与物理和逻辑拓扑结构的分离赋予了CLIP一个类似VLAN的功能而不支持广播。第二，多个LIS可能在同一物理ATM网络上叠加。然而，虽然它们可能共享同一ATM网络，它们只能通过一个路由器通信。在我们的例子中，所有LIS间的通信必须经过路由器。

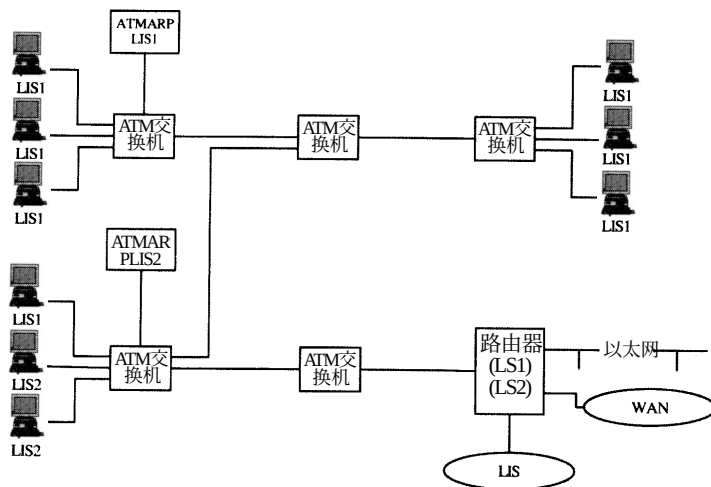


图7-7 ATM上的经典IP和ARP网络实例

第三，CLIP不改变IP互连的“连接网络”模型。换句话说，逻辑IP子网经路由器与其他LIS和其他网络相连。只要涉及到路由器，LIS就只是路由器将向其转发分组的另一个网络。最后，CLIP不支持广播或组播。这意味着它不会是传统LAN通信的一个直接替代。如果主机应用需要广播或组播支持，或者如果主机具有利用广播或组播传送（OSPF）的控制功能，这应该成为考虑的一个内容。然而，如果一个MARS客户机和服务器功能分别包括在（也许是共存）ATMARF客户机和服务器功能中，一个CLIP形式的网络便可以令人信服地提供一个传统LAN的所有业务。理论上这是可行的，但实际上不太可能，因为ATM论坛LANE协议提供单播和组播的ATM上的IP支持已经有一定时间了并且网络实现者希望开发已经运行的大多数厂商正建立和即将运作的功能。

7.4 下一跳解析协议

下面要讨论的经典IP/ATM解决方案序列当中的协议是NHRP，在RFC2332中定义。下一

跳解析协议(NHRP)将CLIP ATMARP协议的概念加以扩展,超过了LIS的范围。NHRP是一个子网间(LIS)地址解析协议,使得连接到同一NBMA网络上的主机和路由器能够穿过子网边界建立点到点VC(直通路程)^①。本质上它使得一个源端能够向网络查询一个ATM地址,用来同与实际目的端更接近或级跳更少的NBMA网络上的一个“下一跳”、“上一跳”或“出口”实体建立一条SVC。NHRP的目的是旁路连接到同一NBMA网络(如ATM)的主机和路由器间的中间路由跳,以获得更高的吞吐量、更低的延迟及减轻路由器和交换机的负荷。

NHRP的基本需求可以通过沿着分组经由与同一NBMA(或ATM)网络相连的路由器的若干跳转发的路径进行最好的描述。考虑图7-8表示的网络。主机A.1和C.1分别与路由器1和2连接到一个单一的物理ATM网络。注意路由器物理上经过一个单一的ATM接口连接到ATM网络,因此像一个“路由器棒棒糖”或“一根棍上的路由器”。到每一子网的逻辑连接由一条或多条特定LIS的VC完成,其VC终结于一侧路由器与另一侧同一LIS中的另一个主机或路由器的单个ATM接口。

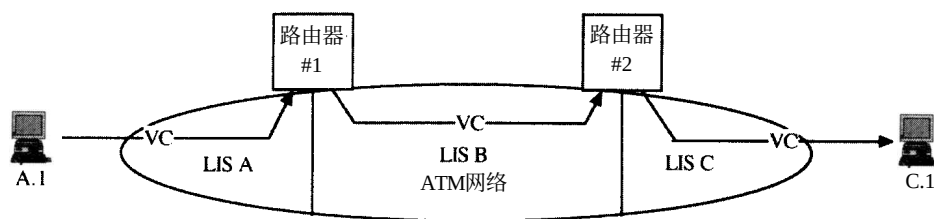


图7-8 经过一个NBMA网络的多跳路由

主机A.1想与主机C.1进行IP上的视频会话。双方当然能够受益于直接ATM通信提供的高性能、低延迟和QoS。但根据标准IP路由,所有子网间的通信必须经过路由器。因此,主机A.1必须首先与路由器1建立一条SVC,通过VC将去往主机C.1的分组转发到路由器1,路由器1接着将执行标准的IP转发。如果路由器1还没有这样做,它必须建立一条SVC到路由器2并接着将分组转发到路由器2,路由器2和主机C.1间也必须这样做。如果N是级跳数,合计需要建立N+1条每一LIS(per-LIS)的VC,且累积的时延和传输延迟与N个跳数有关^②。

也要注意在一根棍上的路由器配置这种情况下,分组(以信元的形式)两次通过ATM交换机和UNI物理装置。这对于急需带宽和时延敏感的IP业务流不是最有效的路径。NHRP通过使在分离的子网中的主机A.1与主机C.1能够在一条ATM VC上直接通信来解决这个问题。它提高了转发路径的性能并从一些分组转发中卸除中间路由器。因此,NHRP可以认为是对所有子网间通信必须通过路由器完成的传统需求的一个缓解。

7.4.1 组件

一个支持NHRP的网络的组件包括下列内容:

- NBMA网络。NHRP在与同一NBMA网络共享连接的主机和路由器上运行。在单个NBMA网络上可以叠加一个或多个IP子网。一个NBMA网络的例子是ATM。

① NHRP经常被称为“下一跳路由协议”。这是错误的。它根本就不是一个路由协议。事实上NHRP需要使用IP路由作为将NHRP消息正确地转发到目的地的一个途径。

② 在一个单一的物理NBMA网上叠加的一个多LIS网总是需要N+1条特定LIS的VC,作为任何两个主机间的一条缺省路由路径。真正性能上的节约通过旁路数据转发路径上的N个跳得到。

• 本地地址组 (LAG)^①。这是共享同一 IP 地址前缀、连接到一个物理 NBMA 网络的主机和路由器的集合。不像 LIS，与不同 IP 子网相关的 LAG 成员能够通过路由器转发数据业务流或者相互间建立直通过路。换一种方式说，“本地/远程”转发决定与目的地址是分离的，并与流的业务量或 QoS 特性有关^②。这对于希望为数量小、保持时间短、尽力而为的业务流进行路由，而为数量大、保持时间长、时延敏感的业务流预留 NHRP 直通过路来说，是有意义的。

• NHRP。这定义了一个 NHRP 客户机和服务器间使用的地址解析协议和消息格式。NHRP 与 ATMARP 相似，只是 ATM 地址与不同网络上的 IP 主机地址和地址前缀有关。

• 下一跳服务器 (NHS)。NHRP 服务器被称为 NHS，位于对 NHRP 客户机的一个集合提供 NHRP 服务的路由器或路由服务器上。一个 NHS 负责维护一个 IP/ATM 地址绑定表，描述在与不在 ATM 网络上的目的地址。这里的一个目的地址可以是一个单个主机的 IP 地址或代表一些主机的一个地址前缀。NHS 的基本功能是为它所服务的特定 NHRP 客户机和目的端提供地址解析。如果一个 NHS 不服务于一个特定主机或目的端，那么它必须试图将 NHRP 请求转发到那个 NHS 去。NHS 必须知道它作为一个服务器为之提供功能的目的实体（地址）。一个子网至少需要一个 NHS。

• 下一跳客户机 (NHC)、NHC 或 NHRP 客户机功能，位于一个主机或路由器上。它向一个 NHRP 服务器发起 NHRP 请求并能够穿过一个 ATM 网络创建到目的地的直通过路。一个 NHC 维护 IP/ATM 地址映射的一个高速缓存，这一高速缓存通过与 NHS 的 NHRP 协议交换获得或者可以通过其他途径构成。一个 NHC 也能够转发分组到缺省路由器。

7.4.2 地址解析

NHRP 地址解析过程的基本操作与 CLIP 的 ATMARP 地址解析过程类似，只是它穿越子网边界进行。首先，客户机必须向服务器登记它们的 IP 地址和 ATM 地址。其次，一个想要与一个目的实体建立直接 ATM 连接的 NHC 通过路由路径向目的端转发一个 NHRP 请求^③。服务于目的端的“最近的”NHS 将用一个 NHRP 应答来响应，应答中包含目的主机的或距目的端最近的 ATM 设备的 ATM 地址^④。NHRP 应答沿路由路径回到源 NHC。NHC 将信息进行高速缓存并且可以接着建立一条旁路中间路由器的直接 ATM VC 或直通过路。NHRP 必备消息头和客户机信息单元 (CIE/Client Information Element) 的通用格式在图

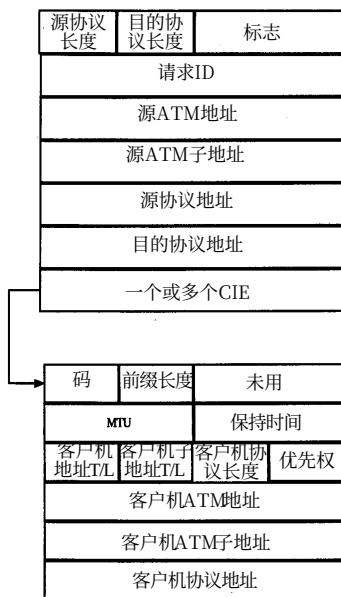


图7-9 NHRP必备消息头和CIE格式

① LAG是专用于NHRP的一个术语，且与大多数熟悉路由等的概念不同。为明确起见，在 NHRP 的讨论中，术语子网或 LIS 将用 LAG 代替，但应该明确主机能够发起和终结子网间的 VC。

② RFC1937, Y. Rekhter and D. Kandlur, “Local/Remote” Forwarding Decision in Switched Data Link Subnetworks.

③ 一个通过 NBMA 网络进行连接的愿望可能不会实际中启动 NHRP 过程的一个充分理由。典型地，如果业务流保持时间长并且可以从 ATM 网络所提供的高性能和低延迟中受益，这一理由便应该被剔除。做出那种决定所需的政策或推断不在 NHRP 规范的范围，但应适用于任何 NHRP 实现。

④ ATM 作为一个 NBMA 网络的例子用于解释中。

7-9中描述，值得注意的字段在下面解释。

一个NHRP消息的必备部分包含的重要字段由下列内容构成：

- 请求ID。用于同步一个NHRP请求与它相应的应答的唯一值。
- 源ATM地址。源NHRP客户机的ATM地址。
- 目的协议地址。NHRP试图用一个相应的ATM地址为其解析IP主机地址或IP地址前缀。

NHRP请求和响应消息可以包含一个或多个CIE子消息。CIE中的重要字段是：

• 前缀长度。这用来产生共享由前缀长度定义的地址前缀的所有IP地址的一个同等类别。这使得NHRP能够建立一条可以由同等类别中包含的多个目的端共享的直通路程。

• 保持时间。CIE中包含的时间量信息被认为是有效的。保持时间超过后，任何高速缓存信息将被丢弃。

• 客户机ATM地址。NHRP客户机的ATM地址。在一个NHRP应答中，这个字段将包含与目的协议地址相关的ATM地址。

图7-10显示了一个单一的ATM网络，包括在三个独立的IP子网上配置的一对路由器和主机。路由器1与LIS1和LIS2相连，路由器2与LIS2和LIS3相连。两个路由器都有一个NHS，每个主机有一个NHC。在两个主机间建立一条直通路程所需的数据流和NHRP消息包括下列内容：

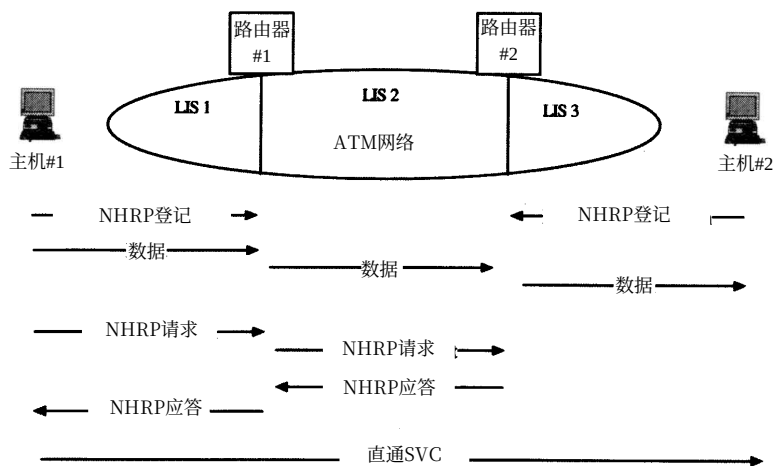


图7-10 NHRP地址解析和数据流

• NHS必须得到它们子网中的NHC的ATM地址。这可以通过若干方式实现，包括使用图中所示的显式NHRP登记消息。另一个技术是人工配置。第三个技术是当NHRP请求/应答分组来自或被路由到网络中的点时，利用NHS“监视”它们。前两种技术用来放到权威的NHS上。一个权威的NHS是子网的实际NHRP服务器且是唯一能够响应权威NHRP请求的NHS。而且它是NHRP请求将去往的ATM网络上的最后一个NHS。

• 源主机（主机1）通过转发分组到缺省路由器开始向一个远程目的主机发送分组。与一个“直通建立”策略相关，分组被转发出一个ATM接口这一事实能够触发主机发送一个NHRP请求。一个NHRP请求包含源IP地址、源ATM地址和目的IP地址并被路由到目的IP主机。

- NHRP请求分组被通过路由路径向目的的路由器转发。当它到达服务于目的的路由器（路由器2）的NHS时，一个NHRP应答被产生并通过路由路径发送到源NHC。NHRP应答分组包含源IP地址、目的IP地址、源ATM地址、客户机寻求的信息和目的ATM地址。

- 收到NHRP应答后，源NHC将这一信息进行高速缓存而且能够建立一条直达目的的路由器（路由器2）的SVC，这样便旁路了两个中间路由器跳。定义的保持时间超过后，高速缓存信息被丢弃。如果在某些点信息变成无效，NHRP便产生删除消息将客户机和NHS维护的高速缓存中的无效数据删除。

NHRP提供的基本功能是相当明确的，但有几点应该指出。首先，NHRP不是一个路由协议。NHRP功能（客户机和服务器）与支持动态IP路由的路由器共存，动态IP路由被要求用来转发NHRP和数据分组通过网络^①。其次，搜集到的对NHRP的一个共同的批评是潜在的高直通路径建立延迟。通过允许中间或转接NHS对它们不明确为之服务的目的端的NHRP请求提供非权威响应，可以将延迟减小到一定程度。将这个例子应用到图7-10中，路由器1已经事先知道了与主机2相关的IP/ATM地址绑定，只要非权威请求比特打开，它便能够响应主机1产生的NHRP请求。当比较NHRP与本书后面将要讨论的其它技术时，一个重要的问题是实际上它只支持点到点直通路径。NHRP目前没有一个明确的技术来支持点到多点直通路径。

有关NHRP的另一个考虑是这样一个事实，即由一个连接到ATM的NHS为之服务的目的端不必物理上连接到ATM网络。在这种特殊情况下，NHRP应答中返回的目的ATM地址是出口路由器或边缘设备的ATM地址。观察图7-11中的配置。已经为位于NBMA网络外的网络A.1的所有业务流建立了直通过程。位于路由器3的NHS的NHRP应答中返回的ATM地址是路由器3自己的ATM地址。



图7-11 通向一个外部NBMA网络的NHRP直通过程

随着协议中定义的可扩展TLV字段的增加，可以认为NHRP是一个通用的地址解析和控制协议，用于连接到和运行于一个NBMA网络的设备。例如，NHRP包含能够使几乎任何互连信息通过交换NHRP消息进行通信的扩展字段。连接到一个LAN的NHRP客户机可以使用NHRP来解析一个ATM骨干网另一边的连接到LAN的一个目的主机的MAC地址。正是这一灵活性和可扩展性提示了ATM论坛，将NHRP功能结合到将在下一章探讨的MPOA规范中。

另一个需要提到的与NHRP有关的重要内容是，在某种拓扑结构下可能会形成永久的路由回路。在一个运行NHRP的网络中形成回路必须具备的条件是：

- 业务流在一条路由器到路由器的直通过程上通过一个NBMA骨干网络。如果直通过程在两端都有一个主机，则不会形成环路，因为根据定义主机不转发业务流。
- 在目的网络和源网络之间存在一条非NBMA“后门”路径。
- 拓扑的一个突然变化对NHRP直通路由是未知的。

① 这意味着静态路由过时且不用，除非在极特别的情况。

• 直通路程上距离或路径向量信息的丢失，使得对于通过出口路由器离开 NBMA网络的分组来看，后门路径是到目的地的最短路径。

在一个使用NHRP的网络中形成一个永久回路的成熟条件在图 7-12中描述。RFC1932提供了一个在NHRP环境中形成一个回路的描述和例子以及避免这种情况的一些可能的补救方法。一个明显的解决方案是删除或阻塞任何后门路径。另一个解决方案是 NHRP直通路由器交换路由信息或可能对路由验证进行某种定期检查。由于前者路由拓扑和数据路径拓扑间内在的不一致，后者将额外的控制业务量和状态引入网络，这两种方案都不太具吸引力。不论采用什么方法，当直通路程穿越路由区域边界时，都将需要维护某种直通状态。因为在一个区域运行的一个协议的特定参数可能与另一个区域中的不一致。

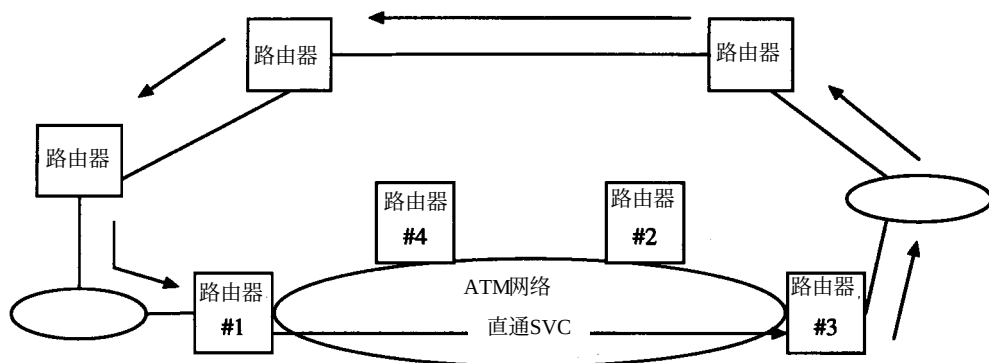


图7-12 概念上的NHRP回路拓扑结构

7.4.3 NHRP网络实例

NHRP用于穿过ATM网络建立直通路程。图 7-13描述了两个这样的情况。在上面的图中，

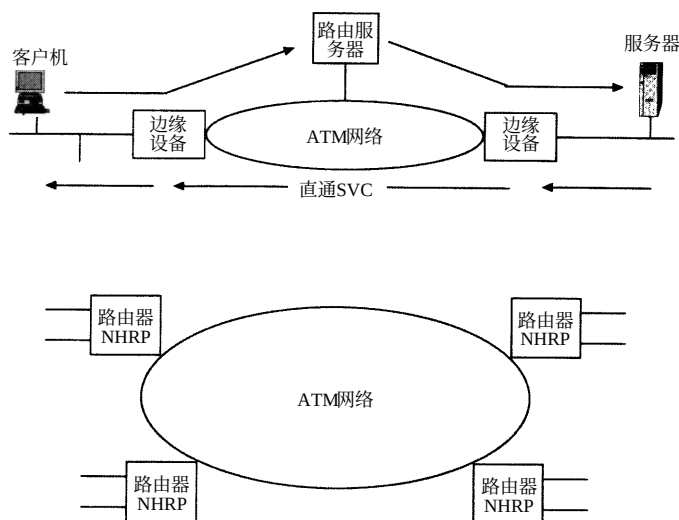


图7-13 NHRP网络实例

一个ATM校园网在与ATM相连的服务器和边缘设备上运行一个NHC功能。NHS功能在ATM路由服务器上运行。NHRP可以用于为经过ATM网络的业务流建立一条直通过程。另一个例子是，数量小的基于事物的客户机到服务器的查询被路由到服务器，从服务器到客户机的返回路径可以是一条NHRP启动的直通过程。这个例子描述了校园网中NHRP的功能。实际上NHRP的超集MPOA很可能将在客户机和服务器上运行，提供同样的直通过程服务。

在图7-13下面的图中，NHRP客户机和服务器功能并存于每个与ATM相连的路由器中，并且用来为单播业务流建立直通过程穿过交换WAN网络。在这个例子中，NHRP提供建立穿过WAN的直通过程的机制，但必须制定哪个业务流将启动和经过直通过程的策略。

7.5 组播地址解析服务器

组播地址解析服务器(Multicast Address Resolution Server, MARS)是经典IP/ATM技术序列中的第三个协议。MARS扩展了ARS概念，将组播地址解析包括进来。一般来讲，MARS为ATM上的IP组播操作提供了一个地址解析和登记服务。MARS在RFC2022中定义，并且也可以认为是一个ATM网络上的组登记的IGMP替代以及在LIS上仿真组播转发行为的一个途径。

7.5.1 ATM上的组播

在讨论MARS之前，对一个ATM网络上用于支持通用组播传输通信的两种技术进行概述也许是有用的。首先，ATM的确以点到多点VC的形式支持组播。这是一个单向的、单一QoS的树，树根位于源点（称为根）且树枝伸向称为叶的特定结点。在UNI3.1中，要求根结点了解它想加到树上的所有叶结点的ATM地址。在UNI4.0中，一个叶结点可以执行一个LIJ来加入一个已存在的组播树，但叶结点必须事先知道一个组ID。如果对通用组播或更明确些对IP组播的要求需要支持多点到多点流，那么这只有在ATM环境中，通过配置一个VC全连通网或通过使用一个组播服务器（MCS）才是可能的。

图7-14描述了这两种方法。一个VC全连通网包括全连通的点到多点的VC，根结点位于每个主机且树枝伸向ATM网上的所有其他主机。由于从根结点到树上的所有叶结点间存在一个直接ATM路径，并且“组播转发”在ATM连接上的信元级完成，所以VC全连通网提供最好的性能^①。然而，VC全连通网消耗的VC资源最多。并且如果组中有多个源点，那么每次加入或去掉一个成员，每个源点和ATM网络间都需要UNI信令。UNI信令处理随源点的数目和会员资格变化的频率成比例。

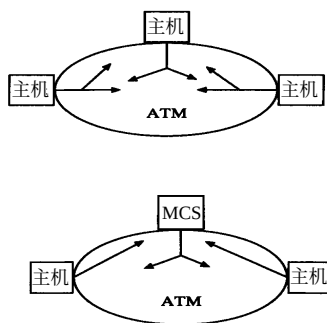


图7-14 ATM上的组播技术

支持ATM上的组播的另一个技术是借助于使用MCS。在这种情况下，源点与MCS间建立一条单一的点到点VC。一条单一的点到多点VC的根位于MCS，树枝伸向包括发送者在内的

^① ATM在ATM级没有任何组寻址的概念。在ATM背景下使用组播转发这一术语，意味着入口信元被复制并交换标记到一个或多个出口端口。

网上所有主机。起源于发送者的组播数据首先被送到 MCS，然后MCS将它转发到网上的所有主机。显然，这需要更少的VC和更小的信令开销。对于一个特定的组，只有一条单一的根位于MCS的点到多点VC需要管理。但是，由于所有业务流必须首先通过 MCS，与这种解决方案相伴随的是增加的传输延迟和时延。这不仅包括在信元级转发数据，也包括将信元组装成一个AAL5帧，并接着将它们分段成信元以在出口的点到多点 VC上传送。发送者也必须以某种方式处理从MCS发回的自己的分组。

MARS两种技术都支持。应由网络管理者决定使用哪种技术，在每组（per-group）的机制上，通过ATM网转发组播数据。

7.5.2 组件

一个支持MARS的网络组件包括下列内容：

- MARS。如 RFC2022中所定义，MARS是给予一组业务和协议的统称，这些业务和协议为支持UNI3.0/3.1 ATM网络上的IP组播所开发。包括用于向一个 MARS服务器登记组员资格和向簇中所有成员分发组员资格信息的协议和消息的格式，以及想要将 IP组播组地址解析为特定ATM地址的发送者所使用的协议和消息的格式。

- MARS簇。一个簇是一组与 ATM相连的端点，它们使用相同的 MARS服务器登记它们的组员资格信息及接收组员资格更新数据。

- MARS 服务器。MARS服务器或简单地称为 MARS，是支持 ATM上的IP组播的一簇与 ATM相连的设备（如主机、路由器）的地址解析服务器。它维护一个 IP组播D级组地址（记作IPmc.n）表和组员（IPmc.1, ATM.1, ATM.2, ...ATM.n）相应的ATM地址，称为主机映射表。MARS服务器也响应来自 MARS客户机的想要将组地址解析为一个或多个相应组员的ATM地址的查询。MARS可以与一个NHS和/或ATMARF服务器共存或者在一个独立的设备上运行。

- MARS客户机。MARS客户机驻留于支持 ATM上的IP组播的一个簇的所有成员上。MARS客户机负责与MARS服务器联系并登记它的 ATM地址及它想成为其中一员的IP组播组。一个MARS客户机发送者也可以查询一个MARS服务器，以将一个IP组地址解析为组成员相应的ATM地址。一个MARS客户机发送者可以根据 MARS所跟踪和分发的它们的成员资格，在一个点到多点VC上增加或删除组成员。一个MARS客户机也可以建立一条到一个MCS的VC，从而通过一个特定MCS将组播分组转发到一个组。

- 簇控制VC（ClusterControlVC）。ClusterControlVC是一个点到多点的VC，根结点位于MARS，树枝伸向所有簇成员。它由MARS服务器使用来分发组成员资格。

- MCS。MCS作为MARS发送者和接收者间的中间会合点来使用。尽管从一个 MARS协议的角度来看，它必须作为一个特定组的发送者的代理簇端点来操作并且当管理去向一个组中所有成员的点到多点VC时必须作为一个发送者运转。MCS负责登记它的ATM地址和它将担当其MCS的IP组播组地址。MARS维护一个独立的 IP组播组地址表和那一特定组 {Ipmc.1, MCS.1.}的MCS的ATM地址，称作服务器映射表。MARS还将MCS加到用于向MCS分发组成员资格的一条点到多点服务器控制 VC（ServerControlVC）上。希望向一个组发送组播分组的发送者需要建立一条VC到一个或多个MCS。接着MCS将分组转发出去到它的特定组的点到多点VC上。

7.5.3 地址解析

MARS协议的基本操作在概念上与迄今为止讨论的其他经典 IP/ATM技术非常相似。首先, 一个MARS客户机向MARS服务器传送它想加入哪个 IP组播组以及它自己的 ATM地址。这使得MARS服务器能够建立一个包括 IP组播组地址和相应的 ATM地址的一个主机映射表。同时, MARS将把客户机作为一个叶结点加到它的 ClusterControlVC上。当一个潜在的发送者想要将 IP组播分组发送到一个特定组时, 它向 MARS发送者查询组播组的主机映射表。当从 MARS服务器接收到信息后, 发送者将组员加到一个根位于源点的点到多点 VC上并开始发送。

MARS协议消息的基本操作和流在图 7-15中描述并在下面更详细地解释:

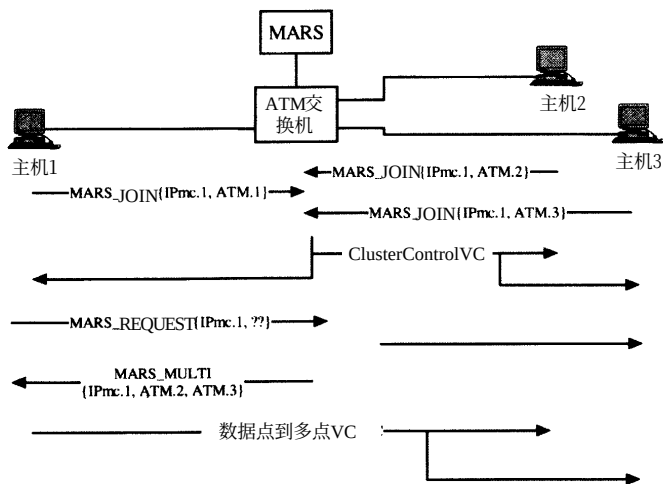


图7-15 MARS地址解析和数据流

- 簇成员主机2和主机3发送MARS_JOIN消息到MARS服务器。MARS_JOIN 消息中包含的信息有主机想加入的IP组播组地址 (Ipmc.x) 和它们自己的ATM地址 (ATM.x)。
- MARS保存MARS_JOIN消息中的地址, 并将IP主机加到它的ClusterControlVC上。
- 一个簇成员主机 1想发送编址为 Ipmc.1的组播分组, 它与 MARS联系并发送一个MARS_REQUEST消息。
- MARS在一个MARS_MULTI消息中返回与组播组地址Ipmc.1相关的组成员的一组ATM地址。
- 主机1建立一条点到多点VC到主机2和主机3, 并开始发送组播分组。
- 一个组成员可以通过向 MARS发送MARS_LEAVE消息而离开一个组播组。接着 MARS把消息传播到ClusterControlVC上, 在那里发送者接收消息并立即从它的点到多点 VC上删除此组员。同样的过程, 使用一个 MARS_JOIN消息, 能够使一个组员加入一个组并且接着由发送者将新组员加到一个已存在的特定组的点到多点 VC上。

发送者通过跟踪由 MARS传播并经过 ClusterControlVC送至发送者的 MARS_JOIN或MARS_LEAVE消息来管理组员资格。根据这些消息的内容, 发送者可以从点到多点 VC上增加或删除成员。

当一个MCS被引进到配置时, 端点(发送者和接收者)的基本行为保持不变, 但中间的机制有轻微修改。有一个MCS作为一个“反射器”将组播数据转发出去到所有的组员。MCS

必须通过向MARS发送一个MARS_MSERV消息告知MARS它将支持哪一个组播组。MARS更新服务器映射表中它的每组服务器映射表项并将 MCS加到一个 ServerControlVC上。ServerControlVC被建立起来,使得MARS能够将任何组成员资格变化 (MARS_JOIN/LEAVE)传送到MCS。MCS使用这一消息来管理根位于 MCS的特定组的点到多点VC。当MARS收到来自一个特定组播组的一个 MARS_REQUEST消息时, 它会将服务器映射表的内容返回。接着发送者将建立一条点到点VC到MCS, 服务于一个特定的组。

如前面所强调的那样, 发送者必须以某种方式处理已经从 MCS反射回自己的分组。这通过使用一个专用簇成员标识符 (CMI / Cluster Member Identifier) 来处理, CMI是当一个主机加入一个簇时由MARS分发的。CMI放在由发送者传送的每个分组的分组头中。接着一个发送者会检查从MCS到达的每个分组的CMI, 如果它与自己的CMI相符则分组被丢弃。这要求额外的AAL帧处理。但尽管如此, 当使用一个MCS时这也是必要的。

假设MARS为一个簇的成员提供一个组播业务, 那么使它能够支持通常情况下的广播是一个自然而然的扩展。RFC2226定义了一个使MARS能够向一个LIS的成员提供IP广播服务的技术。如果有一个支持客户机 /服务器应用的需求, 要求广播功能定位 LIS上的特定资源, 那么这一技术可能尤其有用。另一个例子是控制协议的开发, 如 RIP使用广播传输作为一个分发控制信息的手段。一个IP广播功能将完成已由RFC1577/2225定义的单播IP流和RFC2022定义的组播IP流的ATM上的IP的功能。

使用MARS提供的机制来合并IP广播支持是相当简捷的。想接收广播帧的 LIS成员使用IP广播地址255.255.255.255和它们的ATM地址发送一个MARS_JOIN消息。当一个发送者想发布一个广播时, 它使用一个标准的 MARS_REQUEST/MULTI消息与MARS交互, 获得所有其他LIS成员的ATM地址。多个ATM地址可以组装到一个单一的MARS_MULTI消息中, 降低了控制开销。另一个使用MARS支持广播的优势是使用组播服务器降低VC资源数量的能力。

虽然到目前为止的讨论集中在一个簇内的IP主机间的组播通信, 但有可能一个簇也将包括一个路由器。一个路由器用于转发去往或来自不同网络的单播业务流, 并且也可以用于转发组播数据。连到一个簇的一个路由器使用标准的MARS协议消息登记它的ATM地址和它想加入或离开的任何组播组。在仿真一个连接到一个广播LAN如以太网的具有组播功能的路由器时, 路由器可以选择加入一定范围的组播组, 可能包括全部D级地址空间。在这种方式下, 路由器在一个不加区别的状况工作。然而, 设有M个发送主机涉及到N个组播组, 则路由器必须终结 $M * N$ 个VC并涉及到相关的AAL处理开销。情况可能是, 一些组播业务流不需在LIS外传送, 因此将由网络管理者负责配置路由器, 以规划路由器适合加入的某些组播地址或地址范围。

与LAN相连的有组播功能的路由器的另一个功能是IGMP查询过程。目的是通过向全部主机224.0.0.1组播地址发送一个IGMP查询, 确定LAN上的组播组成员资格的存在。虽然在一个广播LAN上做起来很简单, 一个典型的MARS情况将需要额外的MARS_JOIN消息并消耗额外的VC资源。RFC2022中定义的一个替代过程是使用 MARS_GROUPLIST_REQUEST/REPLY消息。这些消息为路由器提供了簇中活跃组播组的一个目录。这个方法显然消耗了更少的资源且不需要一个显式的簇成员初始化。

7.5.4 MARS网络实例

MARS向一个簇中成员提供一个IP组播业务, 这里一个簇本质上是一个LIS。图7-16描述

了使用 MARS 进行 IP 组播的一个 LIS。如图所示，MARS 可以与 ATMARP 服务器共存。另外，一个路由器也连到 LIS 以提供去往和来自扩展互连网络中的其他网络的单播和组播帧的转发。

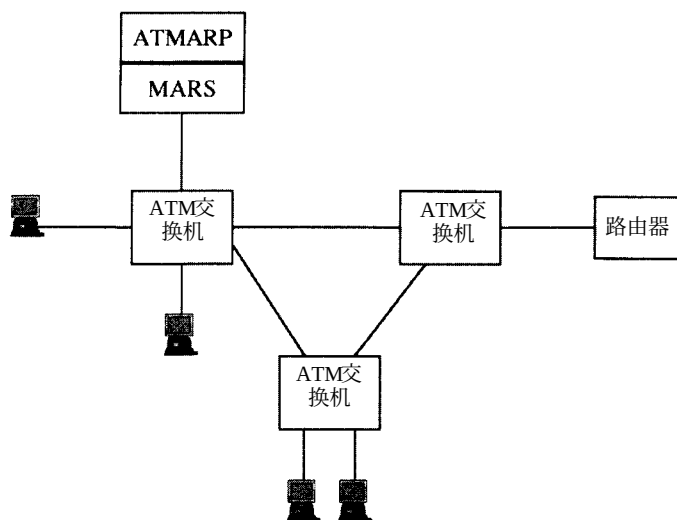


图7-16 MARS网络实例

7.6 服务器高速缓存同步协议

现有的经典 IP/ATM 协议序列依赖于一个单一服务器或多个服务器的可用性和可达性，以在其他设备间提供地址解析协议。如果那个单一的服务器消失，对经典 IP/ATM 客户机不可用，那么网络将停止工作。事实上，ATMARP、MARS 和 NHS 服务器都具有单点失效性。因此，要提高经典 IP/ATM 网络解决方案的强壮性和灵活性，必须想办法对一个单一 ARS 服务器的损失加以防备以免严重破坏网络运行。同时，任何失败保护机制必须努力减小加到客户机、服务器或网络本身的额外开销。

我们可以想象一个可能的解决方案，万一主服务器失效则使用一个“热备份”的备用服务器。在这种情况下，备用服务器连接到网络但配置成在热备份方式下运行。换句话说，虽然从备用服务器到网络可能有一个活动连接，但备用服务器不执行任何经典 IP/ATM ARS 功能。至多它可以建立一条 VC 到主服务器，作为暗中监视主服务器可用性的一个手段。因此，如果备用服务器和主服务器间的 VC 失效，那么备用服务器会知道要开始向客户机提供基本的 ARS 业务。

这个解决方案易于实现，而且热备份这一概念能够向客户机和网络运营者提供一个额外的安慰。然而，这个解决方案有一些缺点。首先是开销问题。很简单，将放在一边不做任何有用工作的备用 ARS 服务器置于网络中是不会节省开销的。其次，鉴于可扩展性的因素，将处理功能从一个单一的服务器卸下并将其分配在网络中的若干服务器上，这也许会有用。尽管这需要所有服务器同步它们的数据库^①，并且这是热备份方式不提供的功能。另外值得关注的是，一旦客户机切换到备用服务器，在网络返回到它的主服务器状态之前会有一些时延。

^① 在 SCSP 的情况下，数据库只是高速缓存或表项的另一个词。它是服务器维护的客户机信息，通过一个查询 / 应答协议对网络中的其他客户机有效。

再有一个问题是，一旦主服务器恢复正常，在什么情况下客户机将返回到主服务器。

为解决热备份方法中固有的局限性，并提供一个通用的多服务器同步和复制的解决方案，开发了SCSP。SCSP在RFC2334中定义。它使得多个服务器能够提供一个如 IP/ATM地址解析的服务，来同步它们的表或高速缓存信息以便所有服务器有同一信息的一个拷贝。如果一个服务器失效，剩余的服务器能够立即提供同样的 ARS业务。它也通过将处理负载分配到网络中的多个服务器上提高了可扩展性。另外，控制业务量的范围局限于客户机和最近的服务器间的距离。

SCSP没有定义一个新的高速缓存同步/复制技术，而是从OSPF借用了数据库同步的一套方法。事实上，SCSP的行为与OSPF的非常相似，所不同的是每次数据库内容更新时不需要进行一个Dijkstra计算。唯一的需求是包含客户机状态信息（如 IP/ATM地址绑定）的数据库在组中所有参与的服务器间进行同步，且数据库的更新被可靠地传送到组中所有的服务器。借用现有的OSPF机制是有意义的，因为它被证明可在一个可扩展的方式下工作，且数据库的更新能够以最小的网络开销迅速、可靠地进行传播。

为最好地描述它的操作，SCSP规范定义了三种类型的服务器：

- 本地服务器（LS / Local Server）。LS是应用操作和功能定义的服务器。仅称它为一个LS以提供一个参考结构。每个具有SCSP功能的服务器从自己的角度来说都是一个LS，并且在LS和其他SCSP服务器间没有功能上的差别。

- 直连服务器（DCS / Directly Connected Server）。一个DCS由一条VC直接连到一个LS。每个具有SCSP功能的服务器从自己的角度来说都是一个LS，并且在DCS和其他SCSP服务器间没有功能上的差别。

- 远程服务器（RS / Remote Server）。一个RS不直接连到一个LS上，但总是具有一个或多个SCSP服务器跳的距离。在RS和其他SCSP服务器间没有功能上的差别。

SCSP规范定义了三个子协议使得服务器能够相互联系，完成高速缓存同步并在需要的时候传播高速缓存更新信息。

- Hello协议。Hello协议由LS使用来决定是否邻近的DCS在运行，并且它们之间的连接是单向的，双向的还是无功能的。

- 高速缓存定位（CA / Cache Alignment）协议。CA协议由一个LS使用来与它的邻近DCS同步或定位其高速缓存。

- 高速缓存状态更新（CSU / Cache State Update）协议。CSU协议用于动态更新一组SCSP服务器中的高速缓存表项。更新的高速缓存表项的一个例子是一个新登记的 CLIP客户机与其相关的IP/ATM地址绑定。CSU公告被可靠地传送到组中的所有服务器。

图7-17描述了连接到一个单一的大LIS上的三个ATMARP服务器。从服务器1，LS的角度来看，服务器2是DCS，服务器3是RS。每次一个客户机向服务器1登记或者一个表项过期，服务器1会发送一个CSU更新消息到服务器2，服务器2接着将其转发到服务器3。因此SCSP使得所有三个ATMARP服务器

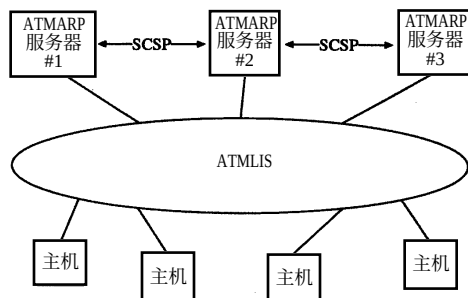


图7-17 SCSP网络实例

能够维护连接到LIS上的客户机的精确信息。

服务器支持的每个协议都必须运行 SCSP 的一个独立实例。以便使 SCSP 能够作为一个通用的服务器高速缓存的同步和复制协议来工作。因此有可能将它在支持经典 IP/ATM 协议的服务器上实现。另外，ATM 论坛将在 LAN 仿真网间接口（LNNI / LAN Emulation Network-to-Network Interface）2.0 规范中定义它的用途，作为提供分布的 LAN 仿真服务器（LES / LAN Emulation Server）/ 广播/未知服务器（BUS）/ LAN 仿真配置服务器（LECS / LAN Emulation Configuration Server）服务的一个途径。并且考虑用其他客户机 / 服务器状态协议如 DHCP 来实现。

7.7 ATM上的IPv6

迄今为止讨论的所有的经典 IP/ATM 协议都是有关 ATM 网络上 IPv4 操作的。假定 IPv6 是 IP 的下一代版本，且未来网络将包括高速的宽带传输，那么将不难想象有大量的 IPv6 业务量流过 ATM 网络^①。目前，使用 ATM 论坛的 LAN 仿真规范支持的仿真 LAN 方法在 ATM 上运行 IPv6 是可能的。许多 IP 交换方案在体系结构上能够将 IPv6 流或地址前缀映射到一条类似 ATM VC 的交换路径上。但没有一个特定的解决方案来将一个 ATM 网络上的 IPv6 操作近似地映射到 RFC1577/2225 中定义的 IPv4 和 ATM 上的经典 IP and ARP 功能上。

这有几个原因。IPv4 是普遍存在的，任何映射 ATM 上的 IP 的解决方案必须首先解析 IPv4。而且 ATM 及 IPv6 在任何有意义的范围内都有待开发，这将迫使研究人员、工程师和厂商开发和交付 ATM 上的 IPv6 功能。另一个原因是由于连续的调整和微调，IPv6 规范还不稳定。最后也许是最重要的，正确的 IPv6 主机的运行依赖于组播的使用，作为用同一网络的其他 IPv6 主机发现和解析 IPv6 到链路层寻址的一个途径。因此，ATM 上的 IP 组播操作问题必须首先解决。

IPv6 在它的邻接主机发现操作中使用组播传送。一个想与同一子网上的一个目的主机通信的源 IPv6 主机首先向一个特定目的地，请求结点（solicited-node）组播地址发送一个邻接主机请求消息。目的主机收到组播分组的一个拷贝，注意到它是指定到其特定组播地址的，便将带有其链路层地址的一个邻接主机公告消息返回给源主机。

定义 ATM 上的 IPv6 操作的初步工作是基于应该不触及邻接主机发现机制这一概念的。或者换句话说，它不应该从根本上变化来适应在一个如 ATM 的特定数据链路技术上的操作。尽管这意味着必须具有一个 ATM 网络上的组播操作所需要的方法，并且在 ATM 上的 IP 组播的情况下，这意味着 MARS。

对各种解决方案的初步研究指出了支持 ATM 上的 IPv6 的一些不同途径。一种技术特指共享同一链路的 IPv6 端点。在这种情况下，一条 IPv6 链路的概念已经推广到一条逻辑链路（LL），这类似于 IPv4 的逻辑 IP 子网。一个 ATM 网络中 IPv6 的内部 LL 连接通过一个 MARS 服务器向 LL 的其他成员分发邻接主机发现消息来完成。一个 ATM 网络中 IPv6 的 LL 间连接由 ICMP 重定向消息或者也许由路由器到路由器的 NHRP 来支持，以作为将一个远离链路的目标地址解析为相应的 ATM 地址的一个手段。

图7-18描述了 ATM 上 IPv6 的内部 LL 连接。主机1和主机2与一个 ATM 网络相连。MARS 作为一个提供 IP 组播支持的工具，更明确地作为一个向 LL 中的主机转发邻接主机发现消息的途

^① Internet2 将更有可能首先开发 ATM 上的 IPv6 和其他宽带传输。

径而存在。尽管在执行邻接主机发现之前，主机 1 和主机 2 必须作为簇成员向 MARS 服务器登记。如果主机 1 想与主机 2 通信，主机 1 将向 MARS 发送一个邻接主机请求消息。如果到目的簇成员存在一条点到点 VC，则 MARS 可以直接转发消息到此簇成员或者 MARS 可以将消息发回到它的 ClusterControlVC 上。注意在这种情况下，MARS 已经被改变用于转发非 MARS 控制信息。当目的主机收到邻接主机请求时，它会从消息中提取 IPv6 和 ATM 地址并建立一条 VC 回到源主机。现在同一 LL 上的两个 IPv6 主机间的 ATM 连接已经建立起来。

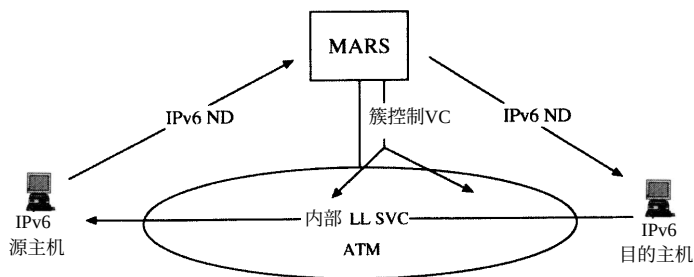


图7-18 ATM上的IPv6的内部LL联接

更多的信息

<http://www.ietf.org/html.charters/ion-charter.html>

Metz, C., and Sackett, G., ATM and Multiprotocol Networking, New York: McGraw-Hill, 1997.

Minoli, D., and Schmidt, A., Network Layer Switched Services, Wiley Computer Publishing, 1998.

第8章 LANE和MPOA

最初在RFC1577中定义的ATM上的经典 IP and ARP 是第一次尝试着在ATM提供的面向连接的服务之上映射一个无连接网络层协议。这一尝试由 IETP发起和执行，自然仅仅从IP的角度来解决问题。但数据联网的世界不是也永远不会是由单一协议构成的^①。除IP以外，还有大量基于其他网络协议如SNA/高级对等联网（APPN / Advanced Peer-to-Peer Networking）、Netbios、IPX、Appletalk等的老式LAN应用。如果ATM会成为数据网络应用的一个有用的传输手段，那么它必须能够适应包括IP在内的其他协议。

ATM论坛解决了这一问题并开发了ATM上的LAN仿真（LANE / LAN Emulation）规范^②。LANE仿真了一个物理ATM网络之上的一个典型广播LAN的功能和行为。不加修改的LAN应用能够在ATM网络上运行。老式LAN工作站和应用能够与它们连接到ATM上的对等端通信，如同它们通过一个LAN的网桥通信。正如同任何老式的LAN一样，LANE能够支持IP及其他网络协议。

LANE的工作完成后，ATM论坛并没有停在那里。为利用ATM网络的速度和容量来加速在网络间转发路由业务流，ATM论坛开发了MPOA规范。MPOA使用LANE规范和IETF的NHRP中定义的机制在不同网络的设备间建立直通路，从而避免了传统路由器中固有的相对较慢的逐个分组的处理。事实上MPOA是一个流驱动的IP交换技术。

本章提供了LANE和MPOA规范的组成和功能的概述。当然两者都充分利用下层的ATM传来加速IP端点间的数据流。

8.1 LAN仿真 V1.0

LAN仿真V1.0于1995年完成，是ATM论坛的LAN仿真分工作组的一个成果^③。它定义了有关的组件、服务、协议和功能，使得与ATM相连的设备能够如同它们连接到一个令牌环网或一个以太网的LAN网段上一样运行。LANE也允许老式LAN网段通过一个ATM LAN网桥连接到一个仿真LAN。对于应用来说，LANE制造了一种它们在连接到广播LAN的工作站上运行的假象。另外，LANE也使得老式LAN能够使用ATM的高带宽和低时延作为骨干传输手段。事实上，在互连老式LAN网段的ATM骨干网上使用LANE服务，是ATM在校园网中最普遍的应用。

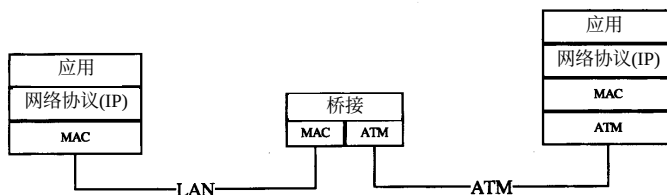


图8-1 ATM LAN仿真体系结构

① 大多数网络的方向当然是IP，但事实上总会有支持多种协议的网络。

② <http://www.atmforum.com>。

③ “LAN Emulation over ATM version 1.0,” ATM forum, af-lane-0021.000, Jan.1995。

图8-1描述了LAN仿真背后的概念。注意除非涉及应用和协议，一个与老式 LAN相连的设备和一个与LANE支持的仿真LAN相连的设备间没有差别。

LANE的操作与前面讨论的经典 IP/ATM技术相似，它定义了一个查询 ARS的方法用于地址解析。然而，对于 LANE，是MAC层地址而不是IP地址被映射到相应的 ATM地址，并根据要求分发给LANE客户机。LANE进一步补充了基本的ARS方法，支持广播、组播和未知帧的转发和自动配置的业务。毫无疑问，LANE比经典 IP复杂一些，但在一个ATM网络之上提供一个真正的仿真LAN服务需要这种复杂性。

8.1.1 需求

LANE的开发和引入背后的最终目的是提供一个无缝和透明的演变途径，使得 LAN应用当其从已存在的老式LAN基础结构向更新和更先进的 ATM网络演化时能够使用。在保证透明性的同时考虑向ATM的逐渐过渡，从而产生出LANE规范必须解决的一些特定需求：

- IEEE802.3/802.5 LAN的仿真。大多数 LAN的安装和应用在以太网或令牌环网介质上运行。LANE必须仿真这两种 LAN类型之一的功能，以便应用和协议能够不加修改地在一个ATM网络之上运行。为了应用的缘故，必须支持和仿真标准的 MAC驱动器接口，如网络驱动器接口规范（NDIS / Network Driver Interface Specification）、开放驱动器接口（ODI / Open Driver Interface）、数据链路协议接口（DLPI / Datalink Protocol Interface）和MAC层寻址结构，以便连接到 ATM的工作站能够通过标准桥接技术与他们的老式 LAN对等端通信。另外LANE必须仿真常规共享LAN技术提供的无连接传送服务。这使得一个源端不需要首先建立一个连接就能够开始向目的端发送数据。

- 对广播和组播的支持。许多 LAN应用和协议使用广播和 /或组播传送来分发控制信息，并发现同一 LAN网段或桥接段网络上的其他主机或服务。因此 LANE必须提供一个服务，其中一个源端产生的一个广播 /组播分组，被送到包括任何与 ATM相连的LAN网桥的仿真LAN的所有成员。

- 与老式LAN互连。LANE仿真一个标准以太网或令牌环网的功能。其中之一是将两个或多个网段桥接到一起的能力。必须提供一个功能使得一个老式 LAN网段能够被桥接到一个支持LANE的仿真LAN网段。

- 已存在的ATM论坛信令和路由协议的使用。根据应用，一个 LAN上的通信可能是动态的。例如，上午一个客户机和服务器可以在一段短时间内交换大量数据，接着在一天的其余时间内相互不再联系。在一个无连接的共享介质 LAN上，这个短暂的交互作用在事物处理期间除了带宽没有消耗任何东西。在一个 ATM网络上交换数据需要在通信端点间建立一条 VC。假定LAN通信具有动态特性并且需要通过一个面向连接的网络完成，那么使用已有的 ATM论坛信令和路由协议在LAN端点间动态建立SVC是有意义的。需要时动态分配带宽（SVC），从而在ATM交换机构和ATM端点内保存了资源。因此，需要 LANE在无连接LAN联网与ATM信令和路由间进行连接，以便 ATM SVC能够被建立起来，交换LAN业务流并且当交换完成后释放给其他业务流使用。

- 仿真LAN拓扑与物理ATM网络的分离。ATM网络的拓扑与LANE提供的仿真LAN的拓扑是分离的。事实上，LANE使得支持连接到同一物理ATM网络的多个仿真LAN成为可能。在一个LANE工作站和它成为其中一部分的仿真LAN之间没有任何关联，在LANE工作站和它物

理上连接到的 ATM 交换机间也没有任何关联。最后, 由于 LANE 支持一个广播功能, 它提供一个虚拟 LAN 的性能。

- 自动配置和连接。一个 LAN 工作站可以典型地插到一个集线器或交换机的一个端口并且变成激活状态。LANE 应该提供同样的服务。

开发和设计相对复杂的功能以融入一个公用标准要花费一些时间和代价, 并且需要进行折衷。不过, LANE 是一个工业标准, 在满足所有上述基本需求的同时, 给实现者留出一些空间来引入增值功能。

LANE 支持的基本功能是对在一个物理 ATM 网络上运行的一个令牌环网或以太网 LAN 进行仿真^①。具有 LANE 功能的设备——连接到同一仿真 LAN 的工作站、服务器、网桥或路由器——都“相信”它们使用一个标准 LAN 协议和 MAC 寻址方案相互寻址和通信, 并且它们能够在一个共享介质的 LAN 上发送和接收 MAC 层广播和组播帧。LANE 提供了产生和维护这一想法的服务。因此无连接传送以及广播和组播服务得以支持。发自仿真 LAN 的业务流能够被桥接或路由到老式 LAN, 反之亦然并且在仿真 LAN 之间也是如此。

除了一些极小的例外, 的确没有什么事情是一个老式 LAN 能做而一个仿真 LAN 不能做的^②。由于它与下层 ATM 机制的关联, 一个仿真 LAN 能够加强一个 LAN 的操作。能够减少广播的蔓延, 开发更多的带宽及可能的 QoS 并且动态简化一个 VLAN 的操作。

8.1.2 组件

LAN 仿真规范定义了一些概念和组件。一个是 LAN (ELAN) 本身。一个 ELAN 包括一组与 ATM 相连的设备, 其中包含一个或多个称为 LAN 仿真客户机 (LECs / LAN Emulation Clients) 或 LE 客户机的客户机组件^③。一个 ELAN 也必须提供在一组 LAN 仿真 (LE / LAN Emulation) 服务器上运行的一组服务和功能。LEC 是存在于每一个使用 LANE 服务的与 ATM 相连的设备中的逻辑实体。它本质上位于 ATM AAL 层之上且, MAC 驱动器之下。LEC 的基本功能是将上面的应用和 MAC 协议功能与下面的 ATM 服务相连。

LANE 规范中定义的 LE 服务包括 LES、BUS 和 LECS。LES 维护一个 MAC/ATM 地址绑定表并支持称为 LE_ARP 的特定地址解析协议。BUS 向 ELAN 中的其他成员转发广播、组播及未知或未解析帧。LECS 提供 LEC 在网络初始化时检索的配置和 ELAN 成员资格信息。每个 ELAN 必须有一对 LES/BUS, LECS 是可选的。

图8-2描绘了 ELAN 和它的组件。一个 LEC 使用一组共同称为 LAN 仿真用户到网络接口 (LUNI / LAN Emulation User-to-Network Interface) 的过程和协议与 LE 服务器接口。在

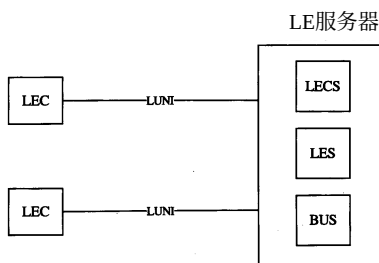


图8-2 仿真LAN的概念和组件

① 一个单一的LANE实例能够支持一个令牌环网或以太网网段——但不是两者都支持。如同它们的老式对等体, 仿真LAN可以由路由器或网桥互连。当然也可能在一个 ATM 网络的上面支持多个仿真 LAN。

② 一个仿真 LAN 并不仿真它的老式对等体的每一个功能上的微小部分。如, 一个仿真 LAN 不经历那些发生在以太网上的冲突。一个仿真 LAN 不会发送一个真正的令牌环网中存在的站管理 (SMT / Station Management) 或标志帧。

③ 术语 LEC 和 LE 客户机将互换使用。力争避免使用缩写 LEC 代替多个 LE 客户机, 原因是它会与 LAN 仿真配置服务器 (LECS / LAN Emulation Configuration Server) 的缩写混淆。

LUNI的LEC和LE服务器间执行的四个基本功能是网络初始化、配置、地址解析和数据传送。

LE服务器提供使LEC连接到一个ELAN的主要服务，发送和接收广播帧并且将MAC地址解析为ATM地址以便数据交换能够在ATM SVC上进行。每个LE服务器为与其通信的LEC提供一个特定服务。下面给出了对单个LE服务器所提供的那些服务的详细介绍以帮助解释LANE的操作。

- LES。LES为ELAN提供地址解析服务器功能。一个LEC向LES登记它的MAC（或在一个令牌环网网桥功能的情况下的路由描述符）和ATM地址。LES用适当的MAC-ATM地址绑定响应来自LE客户机的地址解析查询（LE_ARP）^①。如果LES不能用一个特定的MAC-ATM地址绑定来回答查询，那么它将地址解析请求转发到连接到ELAN的其他LE客户机。LES不转发任何数据帧。它的唯一目的是有利于为ELAN的LE客户机成员进行地址解析。

- BUS。BUS处理ELAN成员发向广播MAC地址（‘FFFFFFFFFFFF’）的广播数据、组播数据以及还没有解析出一个目的ATM地址的初始单播数据。BUS提供这样的业务，使得LE客户机能够以一个无连接方式开始发送单播、广播或组播帧。所有广播、组播和未解析的单播帧必须经过BUS。一个BUS通过以下方式操作：接收来自一个源LEC的广播帧的全部信元，将信元组装成一个帧，接着将它们分段并发回到作为ELAN成员的一些或全部LE客户机。

- LECS。LECS根据请求向一个LE客户机提供配置信息，以便它可以加入一个特定的ELAN。LECS提供的信息包括LAN类型（令牌环网或以太网）、最大帧尺寸以及最重要的，LES的ATM地址。LES ATM地址使得LE客户机能够与一个特定的LES建立联系。成功地连接到一个LES是向一个ELAN声明成员资格的隐含形式。如果一个LEC不能连接到LES，它就不能加入一个ELAN。当LECS决定允许或拒绝LEC得到一个特定ELAN的成员资格时，它也可以考虑请求的LE客户机所提供的一些信息（如LEC ATM地址、LEC MAC地址等）。因此LECS是ELAN成员资格策略能够借此进行集中配置，并根据需求动态分发给LE客户机的一个途径。LECS的使用是可选的但通常推荐使用它。如果不使用LECS，LES的ATM地址就必须在每个LEC中配置。

LEC是支持LANE并进行数据通信的与ATM相连的端点中包含的逻辑实体。一个LEC能够在一个主机或边缘设备上运行，如一个ATM LAN网桥或路由器。后面的实现称为一个代理LEC（proxy-LEC），因为它代表其后面的所有设备提供LANE转发和地址解析服务。一个LEC的基本功能包括与LE服务组件建立通信，登记配置的MAC和ATM地址，支持LE_ARP协议以及管理与ELAN中其他LE客户机的直接VC通信^②。

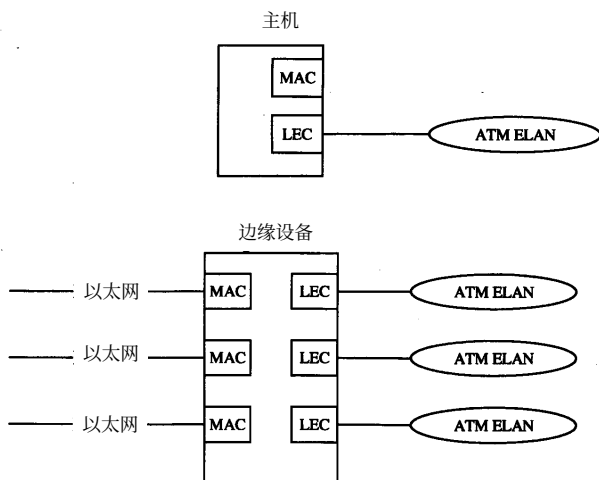


图8-3 LEC和代理LEC

^① 一个MAC-ATM地址绑定由{MAC.1, ATM.1}表示。在一个源路由向量路由描述符的情况下，它用{RD.1, ATM.1}表示。

^② 代理LEC不登记连接到一个老式LAN网段上的设备的MAC地址。但是它响应一个为连接到一个老式LAN网段的MAC地址之一进行的地址解析查询，提供它自己的ATM地址。

图8-3描述了一个主机中的单一 LEC实现和包括多个 LEC实体的一个边缘设备上的代理 LEC实现。

8.1.3 LANE连接

在一个仿真LAN中存在两种基本类型的连接。第一种类型包括在LE客户机和服务器组件间存在的一些控制连接。这些连接使得在LE客户机和服务器组件间能够进行控制信息的交换,以便一个LAN网段的操作能够在ATM网络上进行。第二种LANE连接类型严格地承载数据。广播、组播和未知帧通过BUS在LE客户机间传送,点到点SVC能够通过ATM网络在两个LE客户机间建立。图8-4描述了LANE的控制和数据连接与LE客户机和服务器组件间的关系。

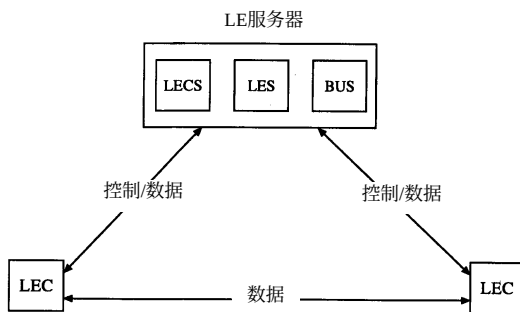


图8-4 LANE的控制和数据连接

LANE支持三种控制连接。它们是：

- 配置直接VCC。这是在LEC和LECS间建立的一条可选的点到点VC。包含LES的ATM地址的配置数据在此VC上转发。一旦信息交换完成,这个连接即可拆除。

- 控制直接VCC。这是在LEC和LES间建立的一条点到点VC。这个必备连接由LEC使用,以达到登记和地址解析的目的。这个连接必须在所有时间都处于激活状态。

- 控制分发VCC。这是一条单向的点到多点VC,根位于LES,树枝伸向连接到ELAN的所有LE客户机。这条可选VC由LES使用来分发地址解析查询并响应LE客户机。

这三个基本的控制连接在图8-5中描述。

LANE支持三种数据连接。它们是：

- 组播发送VCC。这是一条在LEC和BUS间建立的点到点VC,由一个源LEC用来发送并可能接收广播、组播和未知帧。

- 组播转发VCC。这是一条单向的点到多点VCC,根位于BUS,树枝伸向连接到ELAN的所有LE客户机。

- 数据直接VCC。这是在一个ELAN的两个LE间建立的一条点到点VC。

LANE支持的三个基本的数据连接在图8-6中描述。

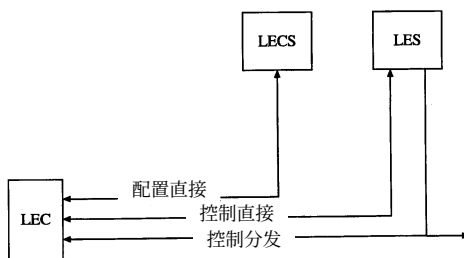


图8-5 LANE控制连接

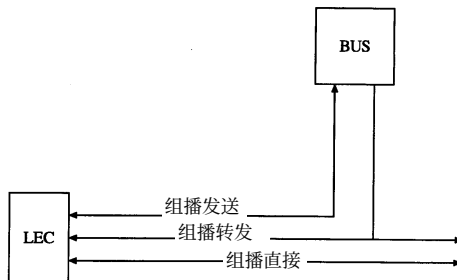


图8-6 LANE数据连接

8.1.4 LANE地址解析和数据流

在连接到ATM的端点间一条新建的 SVC上的地址解析和后续数据流交换是相对简明的，而且在概念上与上一章讨论的基于 ARS的解决方案相似。LE客户机向LES登记它们的MAC/ATM地址信息。LES用MAC/ATM地址绑定响应来自LE客户机的查询，以便LE客户机能够通过ATM连接相互交流。但是为完全仿真一个LAN的操作并在一个可扩展的和有效的方式下进行，需要一些附加的步骤。通过LECS进行自动配置和初始化使得LE客户机能够自引导其自己进入一个ELAN。BUS提供了一个途径，借此能够建立一个LAN广播信道，以便无连接广播帧能够传播到ELAN成员。并且如果LES不能解析一个MAC-ATM地址查询，它将把查询转发到能够响应的其他LE客户机。

图8-7描述了初始化两个客户机及接着执行地址解析和数据直接 VCC建立的基本流程。步骤如下：

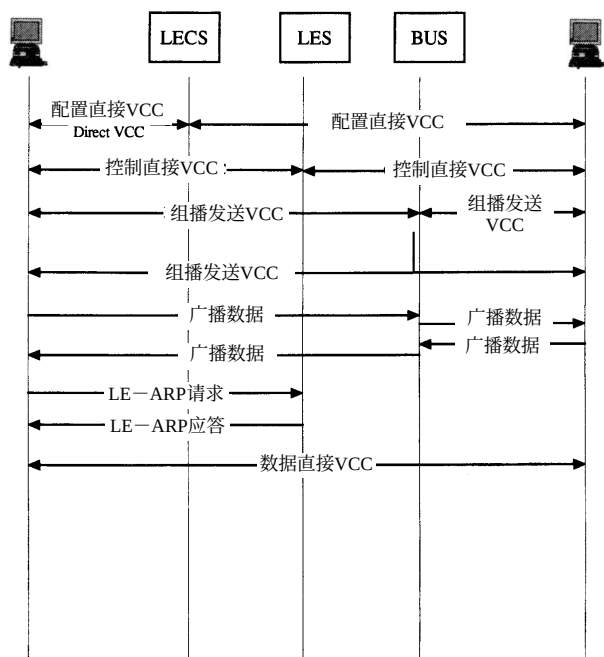


图8-7 LANE仿真的地址解析和数据流

- 第一步是网络初始化。这涉及手工配置或者也可以通过连接到 LECS并与其交互作用来自动完成。这个例子中的LEC1和LEC2已被配置为将与LECS建立一条配置直接VC。为达到这一目的，LE客户机必须以某种方式确定 LECS的ATM地址。可以通过几种途径之一完成：使用一个事先配置的LECS ATM地址，使用一个SNMP临时本地管理接口（ILMI / Interim Local Management Interface）获得操作，从一个ATM交换机检索一个LECS ATM地址或通过使用众所周知的LECS ATM地址进行^①。LECS使用LE_CONFIGURE消息的交换，为LEC1和LEC2提供LES的ATM地址、LAN类型、最大帧尺寸、ELAN名字等等。

- LE客户机的其余几个步骤涉及加入 ELAN和接着向LES登记相应的MAC及ATM寻址信

^① 众所周知的LECS ATM地址是X “47007900000000000000000000000000-00A03E000001-00”。

息。LEC1和LEC2首先与LES建立一条控制直接VCC。LE客户机接着与LES交换LE_JOIN消息，包含如客户机的ATM地址、LAN类型、代理LEC指示符等信息。一旦LEC1和LEC2成功地加入ELAN，它们便通过控制直接VCC与LES交换LE_REGISTER消息，登记它们所知道的所有MAC目的地址^①。另外，LE客户机向MAC广播地址（‘FFFFFFFFFFFF’）发送一个LE_ARP请求以获得BUS的ATM地址。LES接着提供BUS的ATM地址。

- 在加入和登记步骤之后，必须为LE客户机建立一条广播信道。组播发送和组播转发VCC分别为此目的服务。LEC1和LEC2与BUS间建立一条组播发送VCC，接着LEC1和LEC2被作为叶结点加到点到多点组播转发VCC上。ELAN现在可以运行，且能够进行真正数据的传送业务。

- 这个例子中首先发送的实际数据是一系列发自LEC1（IP ARP广播）的广播分组。广播帧沿组播发送VCC转发到BUS，接着沿组播转发VCC从BUS发回到所有LE客户机。一个源LEC通过检查每一帧中的LECID值，丢弃可能已从BUS发回的任何它自己的帧。如果帧中的LECID值与LEC本身的LECID值相匹配，则帧被丢弃。

- LEC2在IP ARP应答中将它自己的MAC地址返回给LEC1。接着帧又在组播发送和转发VCC所提供的LANE广播信道上分别传播。

- LEC1知道LEC2的MAC地址但不知道相应的ATM地址。为解析LEC2的ATM地址，LEC1在控制直接VCC上发送一个LE_ARP请求到LES。LES用包含目的LEC2的MAC-ATM地址绑定的一个LE_ARP应答来响应。如果LES不包含LEC2的一个登记的MAC/ATM地址绑定，它将LE_ARP请求转发到其他LE客户机和连接到ELAN的代理LE客户机。适当的LEC应当通过发回一个LE_ARP应答到LES来响应转发的LE_ARP请求，LES接着将它中继到源LEC。源LEC在一个内部LEC高速缓存中存储MAC-ATM地址绑定。

- LEC1使用常规ATM论坛信令和路由协议建立一条数据直接VCC到LEC2。

到达一个LEC的后续数据帧将首先检查LEC内部高速缓存以确认一个数据直接VCC。如果不存在一个数据直接VCC可用，LEC将查询LES。当等待一个应答时，LEC可以通过BUS发送数据帧。一旦LES响应了查询，LEC便建立一条数据直接VCC。空闲一段时间的连接被释放，并且资源也被释放用于新的连接。

整个过程表面看来显得复杂，并且可能就是复杂的。回过头可以体会到将附加的复杂性设计到LANE中有很好的理由。首先，如果每个具有多个ATM地址和参数的LEC必须手工配置，那么LANE将不是一个非常吸引人的解决方案。为使得配置和初始化一个LEC的过程尽可能简单，引入了LECS。其次，LAN协议充分利用了广播。一个LAN仿真的广播机制是必备的，于是BUS被包括进来。最后，将附加的控制机制和步骤包含进来，这样，如果LES不能直接回答一个LE_ARP请求，它将把这一请求传播给ELAN上的其他LE客户机，且接着将接收到的应答中继到源LEC。事实上，LES不仅仅是一个地址解析服务器，而且是一个地址解析代理者。因此，跟踪ELAN本身和那些它桥接到的老式LAN网段上存在的每个单一的MAC地址和相应的ATM地址，是一项不可能的任务，LES不必去完成。

LANE规范包括一些领域，使得有创造性的实现者能够开发出更好的性能并对网络资源进行控制。广播管理是一个这样的领域。在一个平面的桥接网络中，大量的广播会引起网桥和网段的拥塞。除了在网桥中安装MAC地址过滤器或完全改变应用和网络层协议这样费时且容

^① 一个代理LEC将不登记它所桥接到的老式LAN网段上的MAC地址。

易出错的措施外，真的没有办法来限制广播的蔓延。当然，可以转换到一个路由网络，但那会带来一系列新的可能的性能瓶颈和网络复杂性。在 LANE 中，所有的广播必须经过 BUS。BUS 可以被安排来检查特定广播帧并且将它们转发出组播转发 VCC 到所有的 LE 客户机，转发出一个组播发送 VCC 到一个特定 LE 客户机或者可能将其完全丢弃。

例如，当一个 IP ARP 广播到达一个 BUS 后，它通常会被传送到组播转发 VCC 到所有的 LEC 和 ELAN 的代理 LEC 组件上。包括连接到一个代理 LEC 的 LAN 网段上的工作站的每个 LEC 都会看到广播。一个智能的 BUS 能够搜索所有的 IP ARP 请求，并加以应答来认识 ELAN 上 IP 地址的位置且将那些值存储到一个广播高速缓存中。当一个后续的 IP ARP 请求到达时，如果其目的 IP 地址与广播高速缓存中的一个 IP 地址匹配，智能 BUS 就能够通过一条已有的组播发送 VCC 将 IP ARP 请求定向到特定的感兴趣的 LEC 上。这种情况下的智能 BUS 已经动态地将 N 个 LE 客户机的广播开销降低到一个单一的 LEC 开销。

实现者能够在 LANE 中开发的另一个领域是 ELAN 成员资格策略。应该记得在初始化阶段，一个 LE 客户机建立一个连接到 LECS 并交换 LE_CONFIGURE 消息。表面上，LEC 是从 LECS 接收 ELAN 特定的参数，包括 LES 的 ATM 地址。但是，有可能扩展这个配置过程来制定成员资格策略，以便一个 LEC 只能基于 LECS 中配置的一些策略被许可加入 ELAN。这一策略的一个例子可能是拒绝某个试图在非工作时间加入或其 MAC 地址超出某一特定地址范围的 ELAN 成员资格。的确应由实现者来决定怎样利用初始配置过程。这里关键在于 LECS 是一个用于管理 ELAN 成员资格的单一的控制点。对于 LAN 交换机中的标准 VLAN 成员资格技术，通常不是这种情况，VLAN 成员资格信息必须被配置和分发到网络中的所有交换机。

关于 LANE V1.0 有些问题需要指出。首先，整个 LES/BUS/LECS 的综合体具有单点失效性。如果 LE 业务消失，则 ELAN 变为不可运行。有特定厂商的技术用于提供一个备用 LES/BUS 服务器，当主 LES/BUS 失效时进行转换。这个解决方案可能的缺点是整个网络必须随备用 LES/BUS 重新初始化，这在网络崩溃后混乱的重启动阶段中会引入相当大的网络和信令开销。

LANE 另一个经常提到的局限是它从 LAN 中掩盖了下层的 ATM 机制。LAN 应用不能开发出一些 ATM 更强大的功能如 QoS。事实上，LANE V1.0 规定数据直接 VCC 应当使用 QoS 的级别 0（尽力而为）。坦诚地说，这个论点很空洞。应当记得 LANE 的主要目的是仿真一个令牌环网或以太网网段的功能和行为。自然这些 LAN 技术无论如何不提供任何 QoS 机制。实际上仅仅是最近 QoS（RSVP/IntServ）的概念才被引入到 IP 世界——这一尝试仍需要与下面的数据链路层（如以太网和 ATM）进行明确的交互作用。的确没有意义自底向上设计一个 QoS 机制，这确实是应用的任务，即自顶向下确定和请求网络的特定业务。还应当指出，如果使其这样做，一个 LEC 能够建立和接收具有 QoS 的一条数据直接 VCC，并且 LECS 或 LES 能够成为具有 QoS 能力的 LE 客户机存储和检索 QoS 参数的一个逻辑地点。这种情况下，由于 LANE 没有明确定义支持 QoS 的机制，它不承担责任。当 LAN 应用具有向网络申请 QoS 的能力时，它们将需要一个具有 QoS 能力的数据链路层如 ATM，以便适当地工作。只有那时才可以在一个 LANE 环境中开发 ATM QoS，但需要对参加的 LE 客户机进行非缺省的配置。

8.1.5 LANE 网络实例

在开发 LANE 时，曾设想 ATM 将会随时间的推移代替如以太网等老式 LAN 技术。当然，

增加的带宽以及将 ATM 送到桌面的功能具有诱人的可能性。特别是 25MB ATM 被视为一个将 ATM 引入到典型桌面的相对便宜的途径。应用服务器、具有 ATM 上行链路的边缘设备、还可能是路由器，都可以通过更高速的链路如 OC3 连接到 ELAN。将不同速度的设备连接到同一 ELAN 的能力是一个通常被忽略的 LANE 的性能。用一个或多个 OC3 或者甚至 OC12 链路互连 ATM 交换机将提供一个灵活和可扩展的骨干网。另外，特定的应用如需要纯 ATM 服务的 ATM 上的 MPEG2，可以在同一网络基础结构上开发和利用。这种情况在图 8-8 中描述。

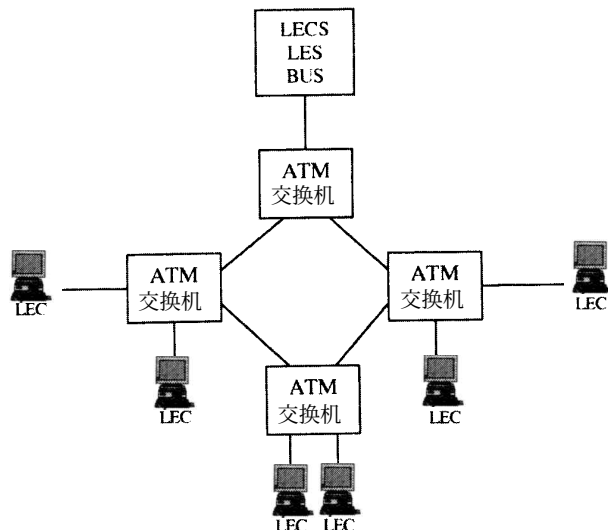


图8-8 标准LANE网络

在 ATM 作为下一代网络技术隆重进发的路上，一个有趣的事情发生了：

以太网。结果以太网能够以较低费用提供同等带宽，而不需要一系列新的和不同的协议和组件。并且如果没有应用需要或能够利用 QoS，安装 ATM 就没有意义。如果带宽是关键，则以太网实现起来更简单、更便宜、风险更小。

对于 ATM 在桌面没有以太网成功的原因大部分听起来是有道理的。但是，需要一个可扩展的和灵活的骨干网将所有那些以太工作组一起互连起来。令人惊讶地，ATM 作为一个优秀的骨干网解决方案悄悄地显现出来。如图 8-9 所示，以太网工作组可以使用 ATM 上行链路通过边缘交换机连接进 ATM 骨干网，或者可能借助直接终结于 ATM 交换机的快速以太网上行链路。两者都提供了高性能的通路进入 ATM 骨干网。骨干网运行 LANE，并且简单地用作工作组间业务流的一个高速传输通道。如果需要更多的带宽，可以升级或增加更多的 ATM 中继线或开发更多的边缘设备。

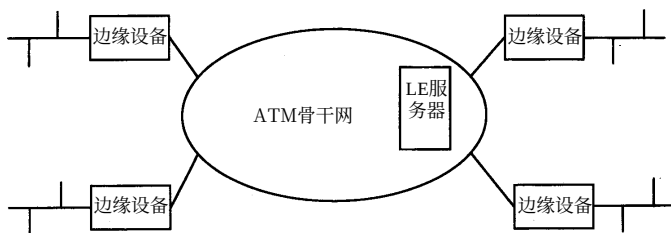


图8-9 使用LANE的以太网工作组骨干网

图例还显示，像传统 LAN 一样，单一 ELAN 能够支持一个或多个 IP 子网。为在子网间转发业务流，即使多个 ELAN 叠加到一个单一的 ATM 网络上，仍需要一个路由器。对于包括多个 ELAN 的一个子网，使用标准 LAN 桥接技术在不同 ELAN 的 LE 客户机间转发业务流。

8.1.6 LANE 与经典 IP 的比较

RFC1577 中定义的 ATM 上的经典 IP 和 LANE V1.0 是支持一个 ATM 网上的数据联网进行的

最初努力。表8-1比较了这两种解决方案。

表8-1 经典IP与LANE 的比较

属 性	经典IP(RFC 1577)	LAN仿真V1.0
支持的协议	仅IPv4	IP、IPX、Appletalk、APPN等
ARS	ATMARP	LES/BUS
自动配置	无	有
广播/组播	无	有
设备	主机/路由器	主机、路由器、网桥
可扩展	否	是
封装	RFC1483	LANE
子网间通信	路由器	路由器
环境	高性能IP子网	LAN、LAN骨干网

8.2 LANE V2.0

LANE V2.0是ATM论坛开发的LAN仿真规范中一个新的版本。它为在ATM网上仿真IEEE802.3和802.5提供了同样的基本服务。它提供了一些新的和增强的性能，允许LE客户机更好地利用下层的ATM机制。随着增加SCSP作为同步多个LES/BUS对间的LES/BUS高速缓存信息的一个手段，LE服务组件的强壮性和可扩展性得以极大地改进。这使得多个LES/BUS实体能够向ELAN成员提供地址解析和广播业务并消除LANE V1.0中固有的单点失效性的局限。

由于新的功能和增强的功能分别强调了LE客户机和服务器的行为和功能，规范分为两个独立的文件。LUNI 2.0强调了特定于LE客户机的行为和功能的那些功能^①。它支持类似LLC复用这类新功能以对VC更好地利用，还提供了对ABR流控和QoS的支持，对组播的增强的支持及对MPOA的支持。LNNI 2.0强调了LANE2.0中服务器间的部分。除了SCSP，定义了一个新的服务器组件称为特殊组播服务器（SMS / Special Multicast Server）。

8.2.1 体系结构

LANE 2.0规范的体系结构在图8-10中描述。LE客户机使用LUNI 2.0协议接入LE服务器组件。但是，在LANE 2.0的情况下，现在可能有多个LES/BUS实体可以供LE客户机利用。LE客户机也可以与一个SMS建立多个发送VCC，成为根位于SMS的一个组播转发VCC上的叶结点。LNNI 2.0协议的范围限于服务器间的通信和业务。

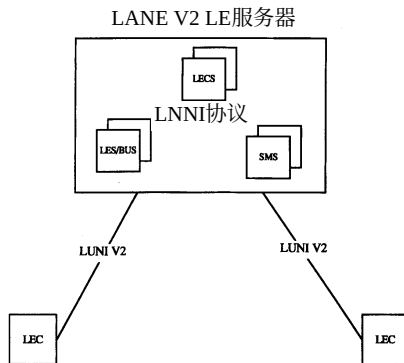


图8-10 LANE 2.0的体系结构

8.2.2 新功能

下面的新功能和增强功能在LANE V2.0规范中定义：

- LLC复用数据直接VCC。在LANE V1.0中，一个数据直接VCC专用于同一ELAN中的两个LE客户机之间的一个特定业务流。在LANE V2.0中，来自一个或多个ELAN的多个业务流，

① “LAN Emulation over ATM Version 2 — LUNI Specification,” ATM Forum, af-lane-0084.000, July 1997。

包括非LANE业务流，可以使用LLC复用VCC复用到一条单一的共享VCC上。图8-11中描述一个附加的12字节帧头，被添加到常规LE数据帧上。它包括一个6字节的LLC/OUI头、一个2字节的帧类型（指示一个LLC复用的IEEE802.3或802.5数据帧）及一个4字节的ELAN标识符。LLC复用VCC通过允许多个LE客户机共享同一数据直接VCC，使得消耗的资源更少。这对于减少在穿越一个ATM骨干网的多个LEC边缘设备间建立的VCC数目特别有用。

- ABR支持。LANE V2.0使得LE客户机能够使用UNI 4.0信令协议建立ABR连接。

- QoS支持。LANE V2.0的LE客户机能够建立具有指定的ATM服务质量的数据直接VCC。在登记过程中，LE客户机使用一个特殊TLV登记它们能够支持哪些QoS业务。接着在LE_ARP过程中，源LE客户机被告知一个目的LE客户机能够支持QoS。

- 选择性组播。在LANE V1.0中，所有组播业务流通过BUS转发。通过使LE客户机能够建立独立的组播发送VCC到其他LE客户机或甚至一个SMS上，LANE V2.0能够将组播处理从BUS（它也必须转发广播或未知帧）上卸下来。登记时，一个LE客户机必须指明它能够选择组播，而且它必须登记它想接收哪个组播组地址。于是BUS就不会将去往那些登记的组播地址的帧发到LE客户机。另一个原因是LES能够通知特定组的SMS：有一个LE客户机应当作为叶结点加到SMS组播转发VCC上。接着，要向组地址发送的LE客户机能够为其组播组地址使用LE_ARP，以便它们可以建立组播发送VCC到相应的SMS。一个典型的LANE V2.0组播情况在图8-12中描述。

- MPOA支持。LANE V2.0提供一组TLV，MPOA可以依靠其进行自动配置、设备发现和缺省选路。MPOA客户机可以从LECS检索配置参数。而且MPOA特定的TLV被添加到常规LANE登记和ARP控制过程，能够帮助MPOA客户机和业务发现彼此^①。

- SCSP支持。在LANE V1.0中，只允许一对有效的LES/BUS。LNNI 2.0吸收SCSP作为一种途径，使得多个LES/BUS对能够向一个ELAN的客户机提供一个单一的LANE服务。这加强了可扩展性和强壮性。

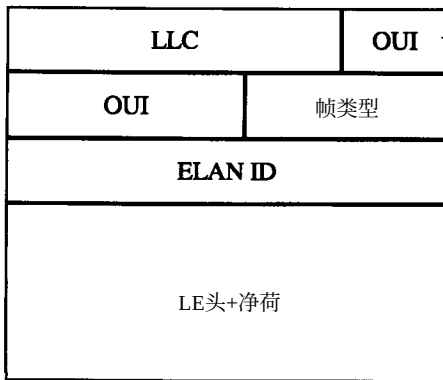


图8-11 LANE V2.0 LLC复用LE数据帧格式

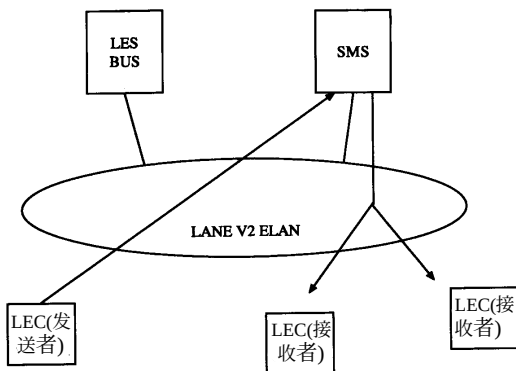


图8-12 LANE V2.0选择性组播

① MPOA仅需要LANE 2.0的一个最小版本来运行。LE客户机的一个最小版本支持LANE V2.0规范中标识的配置变量和TLV集，将LE_ARP应答中的TLV与LE_ARP高速缓存中的条目联系起来，并提供一个更高层的接口到与LE_ARP条目相关联的TLV。一个MPOA客户机不需要支持LLC复用VCC、QoS、选择性组播等。

8.3 ATM上的多协议

LAN仿真支持ATM网络上的一个基于IP（也可以是其他协议）的LAN的运行。但是仍然需要一个路由器穿越子网边界。LANE的下一个逻辑扩展将使得不同子网也可能是ELAN上的LE客户机能够直接通过ATM网络并旁路中间路由器跳进行通信。ATM论坛沿这一路线推进工作，开发了MPOA规范^①。

MPOA采用而且也的确需要LANE和NHRP的功能来支持一个ATM网络上的子网内和子网间的通信。一个MPOA客户机使用LANE与同一子网的其他客户机通信，使用“类似NHRP”的协议解析子网间的地址，并与其他子网的客户机建立直通路径。然而，LANE仿真ATM上的LAN功能，除了ATM带宽能够提供的以外没有任何性能上的提高，MPOA是第一次尝试着开发ATM的高带宽和低延迟，并将其应用到性能受常规路由处理约束的业务流上。事实上，MPOA的目标是将子网间转发路径脱离路由器重定向到ATM交换机构上。也许另一种阐述LANE和MPOA之间关系的方式是，LANE在ATM上提供了一个第二层服务，而MPOA在ATM上提供了一个第三层互连服务。

8.3.1 需求

MPOA的主要目标是能够在不同子网的客户机间建立直通路径。也就是说有一些需求和因素推动了MPOA的开发和导入，现将它们总结如下：

- 高性能子网间通信。许多校园网和内部网包括多个子网。大多数数据流必须经过一个或多个路由跳，从而带来传输延迟和时延。一个MPOA解决方案必须为穿过子网边界的业务流提供一个在性能上明显的改进。

- 影响已有标准。ATM论坛投入大量时间和努力开发LANE规范——它颇有成效。用这一成果对如IETF的NHRP这类受欢迎的工作加以影响，从而获得其机制更易于理解的解决方案，这是有意义的。它也缩短了进入市场的总的工作投入和时间，并为初始采用LANE者提供了一个可见的演进途径。

- 简化路由过程的开销和复杂度。在一个典型的多网环境中，多个路由器中的每个都运行一个特定的路由协议（如OSPF）来执行路径计算，这会带来额外的网络复杂度、操作涉及的范围、还可能有成本。一个MPOA解决方案应该努力将执行路径计算的实例减少或最小化。这样做可能会从总体上降低路由工作的全面复杂度，并导致更低成本、更高速度的转发设备的实现。

- 影响ATM技术。ATM除了其广为人知的技术强度外，也受益于整个网络工业的支持和连续投资。MPOA当然应该尽可能地努力开发ATM技术。同时MPOA解决方案必须允许组件连接到非ATM LAN基础结构，以从ATM交换提供的性能和容量增强中受益。

MPOA提供了在一个ATM网络上支持传统的第三层互连的机制和协议。它通过设立一个虚拟路由器的概念，将前面提到的需求和因素在提供路由服务中加以考虑。MPOA虚拟路由器在图8-13中描述。它包括运行传统路由协议并执行路径计算的一个路由服务器或“CPU”，如同在任何老式单独的路由器中的中央处理单元（CPU）。虚拟路由器的“适配器”是驻留于工作站或边缘设备的客户机转发组件，在输入端接收分组并基于适当的转发表项将它们转发到

^① “Multi-Protocol over ATM Specification V1.0,” ATM Forum, af-mpoa-0087.000, July 1997.

输出端。为加速转发过程，客户机组件可能仅为客户机位于转发路径的那些流高速缓存从路由服务器接收的转发表信息。最后，虚拟路由器的“背板”或交换机构是一个 ATM交换网络。

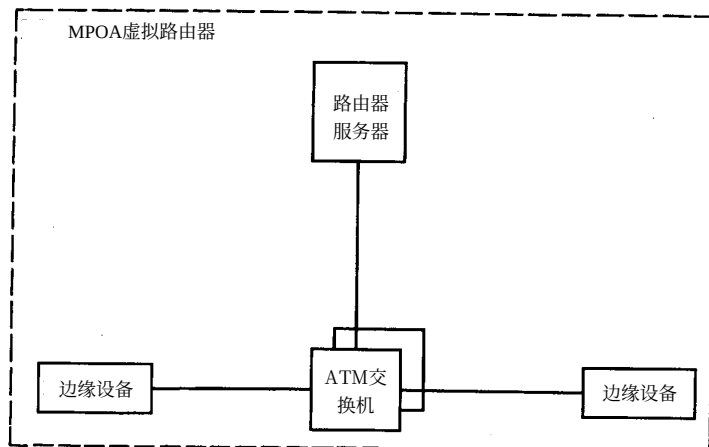


图8-13 MPOA 虚拟路由器模型

虚拟路由器模型的好处可以从内部和外部系统的角度看出。在内部，ATM交换的使用在客户机转发组件间提供了一个可扩展的和高速的传送。路径计算只在路由服务器中进行，并且根据请求只下载相关的转发信息并存储在客户机转发者中。这造成了路由计算与经过网络中多个组件的分组转发过程的分离和分布。简单地通过在边缘增加客户机转发组件或在核心增加ATM交换容量，从而向上扩展性能和容量也非常容易。在外部，虚拟路由器看起来表现得与任何其他路由器一样。它运行路由协议并与其他路由器对等地交换路由协议的更新信息。它能够连接到运行在ATM或非ATM网络上的工作站和设备上，并为其提供路由业务。由于它只运行路由协议的一个单一实例，整个系统的复杂度被降低。

8.3.2 组件

MPOA围绕虚拟路由器这一模型或结构而建立。特别是，MPOA规范定义了两个重要的组件，在图8-14中显示并描述如下：

- MPOA客户机（MPC / MPOA Client）。MPC能够驻留在一个工作站、边缘设备^①、路由

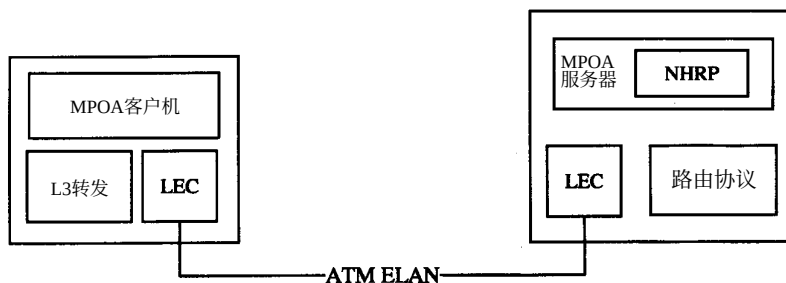


图8-14 MPOA客户机和服务器组件

① 一个边缘设备将LAN业务流转发到一个ATM网络上，反之亦然。在LANE中，边缘设备仅进行第二层桥接。在MPOA中，边缘设备能够进行第二层桥接和第三层转发。

器或路由服务器中。MPC的主要功能是作为直通路的一个源端和目的端工作。MPC包括一个LE客户机,以便它能够连接到一个ELAN上并使用LANE协议与同一子网的MPOA服务器和MPC通信。它也包含一个第三层转发组件,但它不运行一个路由协议。一个MPC将向MPOA服务器查询它到其他MPC建立直通路需要的信息。MPC功能的进一步差异可以通过分别检验入口处理和出口处理中的那些功能得到。一个入口MPC处理进入MPOA系统的分组,使用桥接将它们缺省地转发到使用LANE协议的MPOA服务器,也可能建立一条直通路到一个出口MPC。出口MPC终结直通路并管理一个特定连接的高速缓存,当帧离开MPOA系统时需要用它来处理帧。

- MPOA服务器 (MPS / MPOA Server)。MPS在一个路由服务器或一个路由器中操作。它运行标准路由协议 (如 OSPF), 包括与一个或多个 ELAN 接口的一个或多个 LE 客户机, 并运行如 NHRP 中定义的一个 NHS 版本, 但带有额外的 MPOA 扩展功能。与它的客户机对等体一样, 一个 MPS 能够在入口和出口模式下操作。入口 MPS 通过下载转发信息响应来自入口 MPC 的查询, 这样它就可以建立一条直通路。下载的信息包括网络层地址到 ATM 地址的一个绑定。入口 MPS 也可以作为一个缺省网关为连接到 ATM 的 MPC 主机或一个 MPC 边缘设备另一侧的老式 LAN 设备服务。入口 MPS 也能够将一个 MPOA 查询转换成一个 NHRP 格式以便它可以被转发到另一个 MPS。出口 MPS 能够接收 NHRP 查询并将它们转换到一个 MPOA 格式。同时, 它能够与出口 MPC 通信, 以强加给它特定于待直连连接的高速缓存信息。

8.3.3 MPOA 控制和数据流

与任何联网系统一样, MPOA 组件相互建立联系以交换控制信息和实际数据。MPOA 没有象 LANE 那样定义一组正式的控制和数据连接, 但它将其信息流划分为控制流或数据流。这些在图 8-15 中描述。

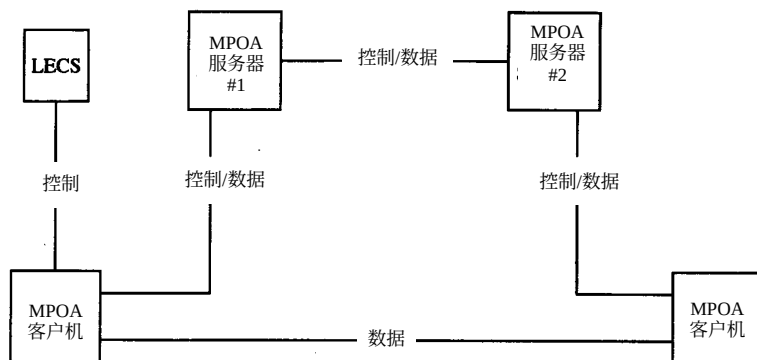


图8-15 MPOA 控制和数据流

MPC使用标准的LANE控制机制从LECS检索配置信息。入口MPC和MPS间的控制流用于地址解析, 出口MPS和MPC间的控制流用于高速缓存管理。除了标准路由和NHRP, 没有定义特定的MPS间协议。

MPS提供了一个缺省的逐级跳路径用于被路由的业务流, 并由时间短和/或数量小的数据流使用, 或者在直通路建立前由任何数据流使用。对于能够从ATM交换受益的数据流, 一条直通路可以在MPC间建立起来。

8.3.4 地址解析和数据流

MPOA的地址解析和后续的直通路建立过程与已经讨论的机制没什么不同。但有些差别与MPOA相关。MPOA利用标准的LANE配置机制从LECS检索MPOA参数。MPOA也利用LANE来发现同一子网上的其他MPOA设备。两者都通过把特殊MPOA TLV值加到LANE控制消息上来完成。一旦系统运行，入口MPC监视进入MPOA系统的各个分组，使用标准的桥接技术和LANE协议将它们转发到入口MPS。接着分组通过一条逐级跳路由路径到达目的端。如果检测到一个流，入口MPC转发一个MPOA解析请求去往实际第三层目的端^①。最后MPOA请求到达服务或出口MPS，MPS没有用一个IP/ATM地址绑定应答，而是首先与出口MPC交换高速缓存管理消息。高速缓存管理消息的交换为出口MPC和MPS提供必要的信息，以完成直通路建立并处理离开MPOA系统的分组。接着出口MPS形成一个应答并将它转发到入口MPC，入口MPC便建立一条直通路到出口MPC。

MPOA与和它关系密切的同辈LANE和NHRP相比有些独特，这是由于当且如果一个直通路请求被启动时，它在入口执行流检测。另外，MPOA在出口MPS和出口MPC间使用一个强制协议交换以在直通路建立之前更新出口高速缓存信息。

MPOA的基本操作在下面列出：

- 1) 配置。MPOA设备从LECS检索配置信息。
- 2) 发现。MPOA设备通过各种LANE控制消息中MPOA TLV的出现，在同一子网相互发现各自的存在。例如，一个MPC必须知道MPS的MAC和ATM地址以便它能够发送MPOA请求。
- 3) 入口MPC到入口MPS的处理。一旦检测到能够受益于一条直通路的一个流，入口MPC生成并接着转发一个MPOA解析请求到入口MPS。如果入口MPS是目的端的服务MPS，那么它能够启动出口处理。否则，它会把请求转换为一个NHRP格式，用它自己的地址代替源地址，并接着将消息沿路由路径转发到目的端。由于它使用自己的地址而不是入口MPC地址作为源地址，所以它能够保证当应答返回时它会看到^②。

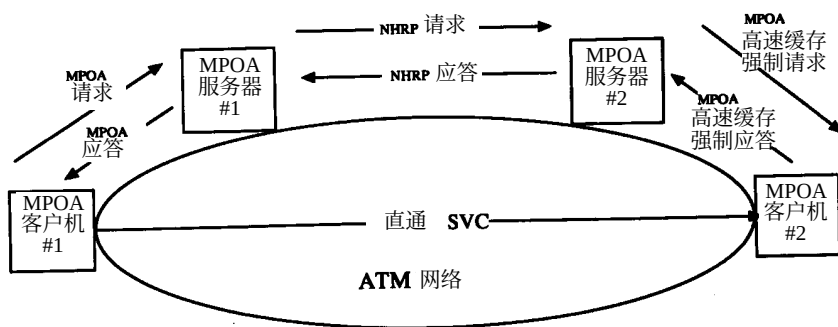


图8-16 MPOA流

- 4) 出口MPS到出口MPC的处理。当NHRP请求到达服务MPS时，它在出口MPS和出口

① 在MPOA的情况下，一个流定义为来自同一源点并去往同一目的地的一串分组，这些分组在一个特定时间区域内到达入口MPC或MPS。例如，如果分别具有同样第三层源和目的地址的10个帧在一个1秒的间隔内到达，入口MPC将其视为一个流并启动MPOA目的地址解析过程。在流检测和地址解析之前，所有的分组被使用LANE协议桥接到入口MPS。

② 一个入口MPC甚至可能没有一个IP地址。

MPC间触发一个高速缓存强制协议交换。高速缓存强制协议有几个目的。首先，它检验在出口MPC有足够的资源来接收直通路程。其次，它由出口MPS使用给出口MPC强加一个高速缓存状态。高速缓存状态包括入口MPC的ATM地址、目的网络层地址以及如果MPS是帧，离开MPOA系统所经过的实际设备，也包括出口MPS将使用的数据链路层（DLL）帧头。从图8-16可以看到，出口MPC（MPOA客户机#2）是分组离开MPOA系统的地方。需要对分组将被放置的介质进行适当的DLL封装，而且必须在出口MPC上安装这一功能。出口MPS使用前面讨论的LANE发现机制发现DLL信息。如果出口MPC有足够的资源支持一条直通路程，那么它用包含出口MPC的ATM地址的一个高速缓存强制应答消息进行响应，并产生一个如图8-17所示的高速缓存项。

出口MPC也可以在高速缓存强制应答中包括一个标记。这一标记用来区分共享同一网络层目的地址和同一源及ATM地址但具有一个不同DLL的业务流。在某些情况下，一个标记也可以加快在出口的高速缓存查询。标记值被包含在入口MPC发送的每个帧中。如果使用标记值，那么在出口MPC中将产生一个带有标记的出口高速缓存项，如图8-18所示。

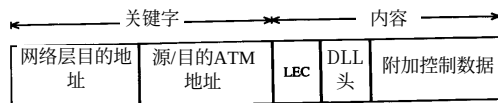


图8-17 出口MPC高速缓存项



图8-18 具有标记的出口MPC高速缓存项

出口MPS从出口MPC收到高速缓存强制应答后，它将其转换为一个NHRP应答并通过路由路径将其转发到入口MPS。

5) 入口MPS到入口MPC的处理。当入口MPS收到NHRP应答时，它将其转换为一个MPOA解析应答并发送到入口MPC。一个如图8-19描述的入口高速缓存项被产生，并且入口MPC能够接着建立一条直通路程到出口MPC。



图8-19 入口MPC高速缓存项

对于执行地址解析和直通路程建立所需流的一个更概括的描述要回过头来参阅图8-16。最后阶段当然是在不同子网的MPC间的一条单播直通路程的建立。有关通常情况下MPOA的操作有几个附加的项目需要指出。首先，所有MPOA业务流如RFC1483中定义的那样在LLC/SNAP帧头中封装。这使得MPOA流能够与其他LLC/SNAP封装业务流以及LANE V2.0 LLC复用业务流共享同一VC。其次，MPOA提供了一个途径，使得一个入口MPC能够建立一条具有QoS的直通路程。一个特殊的MPOA服务类型TLV能够在地址解析过程从出口MPC传达到入口MPC。这一TLV告知入口MPC和出口MPC能够支持什么QoS。

要提到的另一个方面是流检测。并非必须由入口MPC进行流检测。由于所有路由业务流最初通过入口MPS转发，所以在入口MPS安装一个特定的流建立策略是切实可行的。当入口MPS基于事先配置的策略检测到某一流的存在时，它能够向入口MPC发送一个MPOA触发消息，指示它启动MPOA地址解析过程。

8.3.5 MPOA网络实例

MPOA最适合校园网，特别是作为校园网的骨干网。从技术的角度看，MPOA是一个基于流的解决方案，不适合扩展到可能有成百上千个流需要管理的大型核心网络中。它在校

网中应用的另一个理由反映了大多数工作组和端站点位于共享和交换 LAN网络特别是以太网这一现实。将交换 LAN网络插入一个具有 MPOA功能的 ATM骨干网络，从而在校园网的工作组间提供可扩展的高性能的网络层路由会非常经济和简单。一个 MPC边缘设备是将交换 LAN工作组插入到 ATM骨干网的组件。为扩展路由系统的性能和容量，只需要添加更多的 MPC边缘设备或增加连接 ATM骨干交换机的带宽和中继线。

一个具有 MPOA功能的 ATM校园骨干网在图 8-20中描述。只要涉及到一个 WAN上的 MPOA操作，给予足够的信令和 VC容量当然是可能的。入口和出口 MPS能够分别发送和接收 NHRP请求和应答，所以就可能有两个 MPS通过一个基于 NHRP路由器的网络进行通信。只要流的数目不给系统带来过度的需求，MPOA就能够提供尽力而为的逐级跳路由服务和高性能的直通路交换服务。

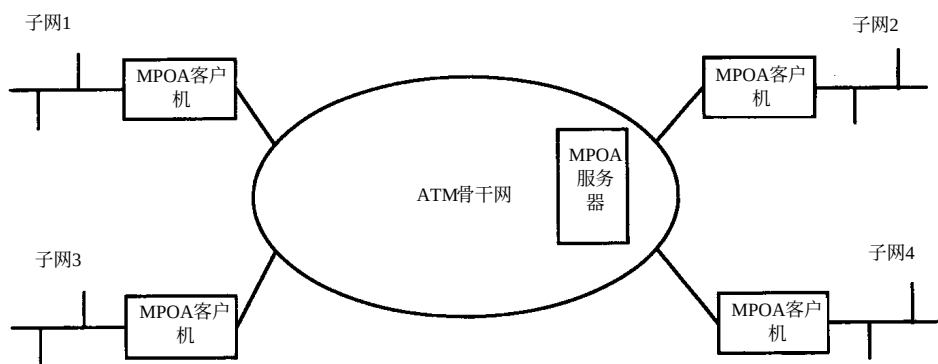


图8-20 MPOA校园骨干网

MPOA已经被建议进行轻微的修改，以便在一个 ATM基础结构上提供直通路由的同时，能够在 WAN上更好地扩展^①。很明显，通用MPOA的基于流的方法不太可能在一个大型网络扩展。如果机制是基于拓扑的，它就可能在一个更可扩展的方式下运行。要达到这一目的所需的基本修改是：

- 删除入口MPC的桥接功能。
- 下载全部转发表到入口MPC。将建立或完成一条直通路径的建立所需的拓扑信息在特定数据流到达之前放好。这导致这种方法有一个拓扑驱动的风格。
- 仍然由路由服务器运行路由协议并执行路径计算。它也必须监视边缘转发组件的状态。
- 使得入口MPC能够执行流分类、分组过滤和标准 IP转发。这会使得非缺省业务在网络边缘分发。例如，一个分类符合标识应该放在一条具有 ATM QoS的独立直通路径上的特殊流的所有分组。

更多的信息

Alexander, C., “MPOA Basics,” IBM Technical Report TR292320, Sept. 1997.

<http://www.mpoa.com>.

“MPOA Multiprotocol over ATM,” ATG’s Communications & Networking Technology Guide Series, 1997, available from <http://www.techguide.com>.

^① Siemens（西门子）和 Newbridge（新桥）已经建议了这一概念并称之为载波系统互连（CSI / Carrier Systems Interconnect）。

第二部分 IP交换协议

第9章 IFMP与GSMP

1996年Ipsilon引入（和定义）了新一类叫作IP交换机的互连设备并创造了IP交换这一词，同时开拓了以ATM技术来提高IP选路的性能的一个创新的方法。IP交换技术描述了动态转发独立的数据流的技术，它利用了由ATM交换机所构成的转发机制。为了说明如何分类独立的IP数据流，以及如何用ATM交换机构转发，Ipsilon首先定义了叫作IP交换机的新设备。本质上，IP交换机是连接到ATM交换机上的路由器，该ATM交换机去除了所有ATM论坛的协议。Ipsilon开发并颁布了两个RFC协议，即IFMP和GSMP。RFC1953所定义的Ipsilon流管理协议（Ipsilon Flow Management Protocol—IFMP）使得两个相邻的IP交换机能够对同一个数据流的信元进行分类并重新标记VPI/VCI值，而RFC1987所定义的通用交换机管理协议（General Switch Management Protocol—GSMP）使得IP交换机的路由器实体能够控制ATM交换机内部的资源和连接表。在统一运行标准IP选路协议的相邻IP交换机网络上，IFMP和GSMP提供了一种附加的功能，对数据流进行分类以判断是否可以交换而不是选路，进而利用由ATM信元交换机所构成的第二层交换式通路进行转发。

9.1 体系结构

Ipsilon IP交换模型的体系结构是基于若干假设，包括选路与交换的价格和性能的比较、Internet业务流的特点、所期望的IP转发设备的行为以及它们所维护的网络状态等。首先，Internet和IP业务量的容量的数量级在不断增长，为了适应这一增长，需要更多的带宽与容量，这是相当可靠的假设，不容置疑。其次，传统路由器以及共享式总线结构和基于软件的选路机制相配合速度太慢，无法满足IP业务量的增长。尽管引入基于硬件的千兆位路由器有所帮助，但由于其价格昂贵和在核心网络中的局限性，可能无法彻底解决问题。分析的依据是，传统路由器用作能够保障QoS的高性能转发机制，其能力不足。另一方面，ATM交换机用硬件所允许的最小代价，设计成以千兆位速率转发信元。作为转发机制，ATM交换机与能提供同等功能的路由器相比具有更低的价格和更高的性能^①。

Ipsilon模型所做的另一个判断是ATM论坛的信令和选路协议过于复杂。为了采用传统方法如ATM上的经典的IP、LANE以及MPOA支持ATM上的IP，则必须支持重叠模型，以管理双重地址空间、双重拓扑结构、双重选路协议。ATM信令和ATM交换机所管理的每个VC的状态都是基于端对端的，不难看出为什么有人认为ATM太复杂。图9-1比较了ATM论坛和IP交换模型把IP数据流在VC上进行映射的地址解析步骤^②。

① 当前这是一种看好的论断，但是基于芯片的选路的进展使得选路与交换在价格/性能上的差距快速靠拢。

② 正如第6章所讨论的，传统的LANE和MPOA的ATM上的IP的解决方案是重叠模型的例子，他们进行独立的寻址。Ipsilon的IP交换解决方案是对等模型的例子，它进行IP—VC的寻址。

最后两个假设与一般的 IP 原则是一致的。第一个假设是，所有的 IP 业务量可以用一系列的数据流来描述，数据流是从某个源端主机/应用到某个目的端主机/应用的分组的序列。这些数据流当中，有些是长持续期的，也就是说通信的双方在一定的时间间隔内交换了多个分组。另一些数据流是短持续期的，通信的主机交换了一两个分组以后就转而进行其他操作。对于长持续期的数据流，在给定的持续时间、分组数量、也许还包括带宽和时延的约束条件下，如果网络能够提供快速通道，自然将在性能和传输能力上得到改善。在同样的约束条件下，短持续期的数据流将用已有的逐级跳的缺省路径转发，因为对短持续期的数据流提供快速通道不会改善性能。实际上，为一两个分组建立快速通道将比利用缺省路径转发开销更大。表9-1列举了一些典型的长持续期和短持续期的 IP 数据流。

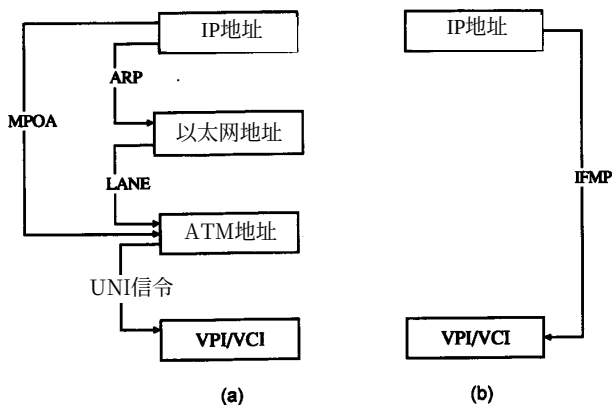


图9-1 地址解析步骤的比较

a) ATM论坛 b) IP交换解决方案

表9-1 典型的IP数据流

长持续期	短持续期	长持续期	短持续期
文件传送	DNS	远程登录	ICMP
NFS	NTP	多媒体下载	SNMP
HTTP 1.1持续连接	电子邮件		

最基本并且可能是最重要的假设是 IP 交换应该保持 IP 网络无连接和无状态的工作方式。标准的路由器不保留通过它的每个业务流的状态，并且不依赖于上游和下游邻机的工作方式。路由器根据本地的信息，对应向何处转发分组作出独立的判断，这种方式同样适用于 IP 交换。每个 IP 交换机根据本地的信息，对于哪些数据流应该被交换作出判断。每个 IP 交换机各自在本地管理关于每个数据流的连接状态，而不是用传统 ATM 中的端到端的工作方式。IP 转发设备中保存着每个数据流的状态，虽然这与 IP 无状态、无连接的原理不一致，但这是一种“软状态”，即这些状态如果不更新将超时。每个 IP 交换机中对于数据流的软状态进行本地的管理，这使得 IP 交换在动态性和健壮性方面与纯 IP 选路非常接近，同时能够在带宽和性能方面利用面向连接的信元交换的优点。实际上，IP 交换机可以看作是改造为路由器的 ATM 交换机。

Ipsilon IP 交换模型的基本体系结构可以用图 9-2 所示的两个相邻的 IP 交换机的功能进行最好的说明。可以看到数据流 1 的信元按照逐级跳的方式在两个 IP 交换机，即交换机 1 和 2 中进行选路。在 IP 交换机之间为该数据流建立的缺省通道是 $VPI/VCI=0/15^{\ominus}$ ，在这个通道上进行逐级跳的转发。在每一跳中进行第三层的处理，涉及标准的路由表查询、TTL 值的递减和分组头校验和的计算，这与在标准路由器中所进行的第三层的处理没有任何差异。每个 IP 交换机的路由器机构中都包含了数据流分类器，以判断后续的分组是否应该用快速的 ATM 交换机而不

^① 相邻的 IP 交换机在初始化阶段建立了缺省的 $VPI/VCI=0/15$ 的通道。所有缺省的业务流和控制消息都是在该 VC 上传送，这种 VC 存在于每一对 IP 交换机之间。

是低速的路由器机构进行转发。每个 IP 交换机完全是根据本地配置的规则进行该判断。当决定把数据流中剩余分组进行交换时，IP 交换机将向上游邻机发送消息，指示邻机为该数据流的信元重新标记新的 VPI/VCI 值。可以假定，数据流 1 经过的沿途的每个 IP 交换机都作出了类似的判断，并且指示上游的邻机重新标记该数据流的信元。一旦数据流在 IP 交换机的输入和输出端口重新标记了非缺省的 VPI/VCI 值，IP 交换机的路由器机构能够很容易把输入和输出的 VC “拼接”在一起，在设备内构成直通过径。从另一个角度，每个 IP 交换机进行本地的数据流分类和重新标记的过程，结果是建立了输入到输出的直通过径，旁路了中间的路由器跳。

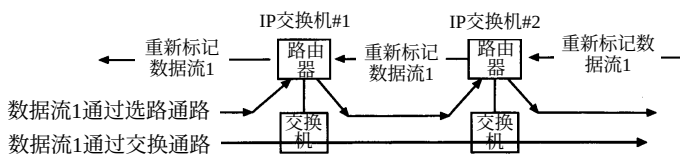


图9-2 IP交换功能模型

另一种看待IP交换体系结构的方法是分析组成IP交换机的基本模块。第一个模块是，IP交换机运行标准的IP选路协议并进行基本的IP转发。这种基本的“IP软件”模块普遍存在于所有IP路由器和任何能够在子网间进行分组转发的主机。第二个模块 ATM交换机只是基本的基于硬件的信元交换机构，而这种信元交换机构传统上是采用 ATM论坛定义的协议和规程控制的，但在IP交换机上去除了ATM论坛的软件。第三个模块是一组由 Ipsilon 独立开发的数据流分类、重新标记和交换机控制规程。于是 IP 交换机以及更普通的 IP 交换模型包含集成到 ATM 交换机的 IP 软件，该软件有附加的模块（IFMP 和 GSMP），能够动态地把数据流映射到 VC，并控制交换机资源。图 9-3 所示为 IP 交换机的基本模块。

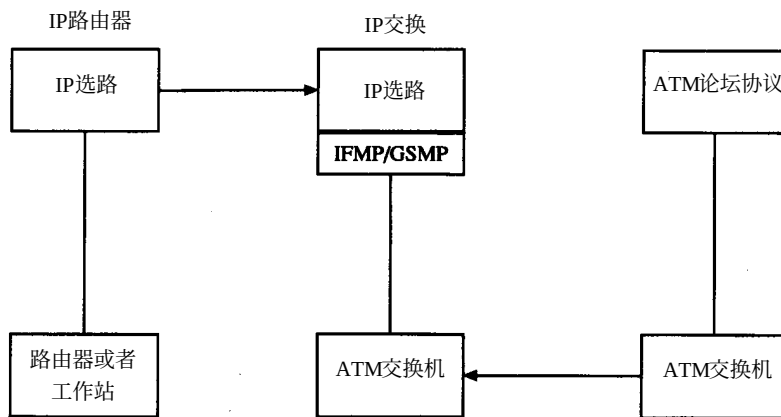


图9-3 构造IP交换模块

9.2 组件

Ipsilon 的 IP 交换解决方案包含以下基本功能和协议组件：

- 数据流。数据流是分组的序列，这些分组从某个源端到达某个目的端（单播或者组播），并与一定的路径以及该路径上每个转发设备给分组所提供的服务（缺省或者非缺省）相关。数据流中的分组通常具有相同的源端地址、相同的目的端地址，也可能有相同的 TCP/UDP 端口号。

• 数据流的类型。Ipsilon的IP交换是在主机之间或者应用之间的次上处理数据流^①。类型1的数据流定义为两个进程或者应用之间靠IP分组中的字段来区别的分组，这些必要字段包括源和目的端的IP地址以及各自的源和目的端的端口号。类型2的数据流定义为在两个主机间以源和目的端的IP地址字段来区分的分组。值得注意的是是一对主机能够支持一个类型2和多个类型1的数据流。

• 数据流标识符。一组IP和传输层的字段，这些字段能够区别某个数据流的类型。图9-4所示为用于类型1和类型2的数据流的标识符。

版本	IHL	TOS	TTL	协议
源端IP地址				
目的端IP地址				
源端口		目的端口		

(a)

版本	IHL	保留	TTL	保留
源端IP地址				
目的端IP地址				

(b)

图9-4 数据流标识符

a) 类型1 b) 类型2

• 数据流分类器。驻留在每一个IP交换机的选路实体中的组件，用以判断对某个数据流进行选路还是交换。网络性能优化的目标要求网络有统一的规则，但数据流分类是按照每个IP交换机自己的规则进行的。数据流分类器检查每个分组头中的特定字段，然后查询自己的规则库。例如，数据流分类器提取某一对源/目的端的IP地址，根据自己的规则库判断该流分组是否应该被交换。依据一定时间间隔内到达的分组的数目，数据流分类器触发重定向过程。

• IP交换机。IP交换机是由连接到ATM交换机上的IP交换控制器构成^②。IP交换控制器运行标准的IP选路协议并执行标准的IP转发，同时它还包括数据流分类器并与上游和下游的邻机交换IFMP协议的信息。IP交换机控制器包含GSMP主控制点，该点与ATM交换机中的GSMP辅控制点通信。图9-5所示为基本的IP交换机。

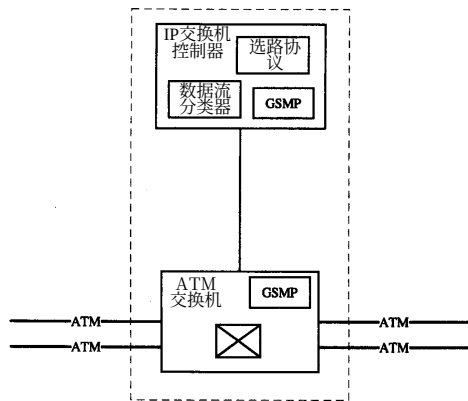


图9-5 IP交换机

• IP交换机的入口/出口。IP交换系统必须有入口和出口，以便业务流能进出网络。Ipsilon的IP交换机的入口和出口是数据流的端点，该数据流映射到第二层ATM VC。入口和出口的功能可以由路由器、边缘设备或IP交换机完成，入口和出口能发送和接收IFMP消息。

• IFMP。IFMP用于相邻的IP交换机对某个数据流的信元进行重新标记。IFMP协议消息从

① Ipsilon的IP交换实现了一种数据流驱动的IP交换，添加了其他的用地址前缀标识数据流的类型的方法，它也可能支持拓扑驱动的IP交换。

② IP交换机控制器是Ipsilon称呼路由器或者路由器模块的名字，它们连接到ATM交换机上，形成IP交换机。

IP交换机传递到其上游邻机。IFMP功能包括关联协议和重定向协议，前者发现链路上的对等IP交换机，后者管理指派给某个数据流的标志（VPI/VCI）。

- GSMP。GSMP被IP交换机控制器用于管理ATM交换机的资源，它能使IP交换机控制器建立和释放跨越ATM交换机的连接。

9.3 Ipsilon数据流管理协议

Ipsilon数据流管理协议（Ipsilon Flow Management Protocol—IFMP）作为IP交换机网络中分发数据流标记的协议，它在跨越外部数据链路的相邻的ATM交换机控制器（或者IP交换机入口和出口）之间工作。具体来说，下游IP交换机（接收端）利用它通知上游交换机（发送端）在某一段时间段内为某个数据流赋予某个VPI/VCI值。数据流用流标识符来标记。下游IP交换机必须在一定时间段内更新数据流的状态，否则数据流的状态将被删除。

IFMP仅在相邻的IP交换机之间操作，与其他IP交换机的工作方式和功能无关。换句话说，IP交换机从下游IP交换机接收到IFMP重定向消息后不会向其上游邻机发送类似的消息。IFMP的基本目的是分发与某个数据流相关的新的VPI/VCI值，以便加速转发功能并有可能基于每个数据流进行交换，从而提高总吞吐量。

9.3.1 IFMP重定向协议

IFMP定义了两个协议子集，在相邻IP交换机之间交换消息。当两个IP交换机利用外部ATM数据链路初次通信时，第一个协议子集，即IFMP关联协议被触发。它使具有IFMP功能的设备（IP交换机、主机、路由器等）发现在链路另一端的对等设备的标识，用于同步链路两端的状态、检测对端标识的变化、交换指派给链路的IP地址表。接收节点处理后续的IFMP消息之前，关联协议必须完成链路两端的同步状态（ESTAB状态）^①。

IFMP重定向协议处理数据流/标签的关联信息的分发和管理。下游（接收端）IP交换机负责发起重定向过程，此过程包括给接收端口分配新的标记（VPI/VCI），并在IFMP消息中把新的标记和数据流标识符以及高速缓存值的生存期一起发送到发端IP交换机。IFMP不要求明确的确认，带有缺省（VPI/VCI=0/15）或者非缺省标记的后续数据的到达将通知接收节点是否成功处理了以上的消息。

IFMP重定向协议是接收端发起的。原因之一是，可靠的网络协议中发送端不能使用非缺省方式发送信息，除非发送端能够确认（即使不是百分之百）接受端接受并处理了这些信息，否则接收端的IP交换机可能恰巧在输入端口已经分配了此非缺省的标记（VPI/VCI），而发送端的IP交换机却使用这些端口作为输出端口（尽管概率很低），这将导致信元丢弃，而后需要一定的错误恢复并进行标记同步的时间。发送端的IP交换机接收到包含所用标签的明确的消息，确知某个标记已经被接收端的IP交换机分配。因此正是接收端的IP交换机发出的这个IFMP消息向发送端的IP交换机通知已经被分配了非缺省的标记。原因之二是，只有已经在接收端口分配标记，基于信元的标记替换的非缺省通信方式才能成功。当发起重定向过程时，IFMP首先分配该资源。

IFMP消息的发送端用IPv4的分组头封装IFMP信息，目的端IP地址是外部链路另一端的对等IP交换机。对等IP交换机的IP地址通过关联协议获得。IFMP重定向协议支持五种类型的消

^① RFC1953的开头部分描述了IFMP关联协议的状态表。

息：

• 重定向消息。重定向消息通知相邻的上游 IP 交换机在某时间段内把某个标记赋予某个数据流。图 9-6 所示为重定向消息元素（封装在 IP 和 IFMP 分组头内）的格式，解释如下：

数据流类型	数据流标识符 长度	生存期
标记		
数据流标识符		

图9-6 IFMP重定向消息元素

□ 数据流类型字段。指明包含在数据流标识符字段内的数据流标识符的数据流类型（类型 1 或 2）。

□ 数据流标识符长度字段。以 32 位字为单位的数据流标识符字段的长度。

□ 生存期字段。以秒为单位的时间长度，指明在接收到重定向消息的 IP 交换机内部数据流标识符和标记的关联关系所维持的时间。

□ 标记字段。包含一个 32 位的标记值。对于 ATM 数据链路，这个值包含图 9-7 所示的 VPI 和 VCI 的值。

保留 (4)	VPI (12)	VCI (16)
-----------	----------	----------

图9-7 ATM数据链路的IFMP标记

□ 数据流标识符字段。指明将与标记关联在一起的数据流。

• 重声明消息。重声明消息通知相邻的上游 IP 交换机解除一个或多个流与标记的关联关系，并释放这些标记以便数据流标识符字段能够分配给其他数据流。当网络拓扑变化或者下游 IP 交换机的标记不够时，有必要进行这种操作。解除关联关系的数据流重新采用缺省的转发工作方式释放标记以便将来使用。每个重声明消息的元素包含数据流类型字段、数据流标识符长度字段、标记字段和数据流标识符字段。

• 重声明确认消息。在每个上述的重声明消息成功地释放了一个或多个标记后，相邻的 IP 交换机利用重声明确认消息给出确认。

• 标记范围消息。上游 IP 交换机向下游 IP 交换机发出标记范围消息，响应包含标记的重定向消息，指明上游 IP 交换机（发送端）不能分配该标记。标记范围消息包含了最小和最大的标记值，上游 IP 交换机只能支持此范围内的标记值。

• 出错消息。出错消息是对 IFMP 的重定向消息的响应，用来通知下游的 IP 交换机，发送端由于处于异常状态而无法成功处理重定向消息。异常状态的一个例子是数据流类型未知。

9.3.2 数据流标记封装

IP 交换支持在 ATM 数据链路上传输 IP 分组的几种不同类型的封装格式，在 RFC1954 进行了定义。封装格式依赖于下游 IP 交换机在 IFMP 重定向消息中传递的数据流类型信息。图 9-8 所示为在缺省通道上 IP 分组的缺省

LLC (AA-AA-03)	
SNAP (00-00-00-08-00)	
IP V4 分组	
填充 (0-47)	
AAL5 尾	

图9-8 缺省封装

封装格式。RFC1483定义了LLC/SNAP封装的技术。

IFMP重定向过程为IP分组打上新的标记，IP分组根据各自的数据流类型进行封装。图9-9所示为类型1和2的数据流的封装方式。两种方式的有效数据放在标准的AAL5 PDU中，使用RFC1483定义的VC复用封装或者无封装。有效数据前面没有LLC/SNAP的分组头。在类型1的数据流的封装中，IP分组头的下列字段不传送：版本、IP分组头长度（Internet Header Length—IHL）、服务类型（Type Of Service—TOS）、存活期（Time To Live—TTL）、协议、源和目的端的IP地址。另外，IP分组头后面的4个对应于源和目的端的TCP或者UDP的端口值的字节也不传送。在类型2的数据流的封装中，IP分组头的下列字段不传送：版本、IHL、TTL、协议、源和目的端的IP地址。两种方式中，不被传送的IP字段保存在下游发出重定向消息的IP交换机中，这些IP字段与非缺省的标记联系在一起。这使得发起重定向过程的IP交换机能够在必要的时候重新装配出完整的分组。它同时可提供安全机制，因为对于潜在的入侵者而言很难使用一个数据流与标记的映射关系建立交换的数据流连接，即使建立了连接，也很难交换到具有相同标记关联的另一个数据流上（不同的源/目的端的IP地址对），入侵者的目的是用IP交换过程逃过第三层的检查，并未经许可而访问计算资源或数据。但是下游或者终端的IP交换机能够由一些字段重建原来的分组，当最初的IFMP消息向上游发送时，这些字段被缓存起来。

总长度		标识符
标志	分段	校验和
数据		
填充		
AAL5尾		

(a)

保留	TOS	总长度	
标识符		标志	偏移量
保留	协议	校验和	
数据			
填充			
尾			

(b)

图 9-9

a) 数据流类型1 b) 数据流类型2的封装

9.4 通用交换机管理协议

通用交换机管理协议（General Switch Management Protocol—GSMP）是一个简单、通用的交换机管理和控制协议，用于对IP交换机内部交换机资源的管理^①。该控制协议是Ipsilon的IP交换解决方案的组成部分，而GSMP是两个该解决方案特定的控制协议之一。IFMP通过外部ATM数据链路运行在相邻的IP交换机控制器之间，而GSMP运行在IP交换机内部的IP交换机控制器和ATM交换机之间。GSMP使IP交换机控制器能够建立和释放穿越该交换机的连接、在点到多点连接中加入和删除叶节点、管理交换机端口、请求配置信息。ATM交换机也可以使用GSMP的IP交换机控制器主动地提供事件信息。

IP交换机的模块包括IP交换机控制器和ATM交换机，这两个模块交互GSMP协议消息以管理交换机资源。二者有主从关系，IP交换机是主，ATM交换机内的GSMP代理是从GSMP协议

① 该协议非常简单，许多不同的ATM交换机都实现了它，其他IP交换解决方案也采用了该协议来管理ATM交换机资源。

消息通过在IP交换机控制器和ATM交换机之间建立的标准控制VC（VPI/VCI=0/15）传递。RFC1483定义的LLC/SNAP封装用于封装GSMP消息的头和数据段，见图9-10所示。选用LLC/SNAP封装，其他协议，如SNMP就可以复用在同一VC。正如它的姊妹协议IFMP，GSMP定义了关联协议，用于同步链路两端的状况并识别链路两端的通信方^①。

消息类型区域表明在有效数据中GSMP消息的类型。结果字段用于指定对某个成功的请求的响应。由于GSMP是请求/响应协议，因而GSMP头含有处理标识符将特定请求与相对应的响应关联。

LLC (AA-AA-03)			
SNAP (00-00-00-88-0C)			
版本	消息类型	结果	代码
交互标识符			
GSMP 消息			
填充(0-47)			
AAL5尾			

图9-10 GSMP 消息格式

9.4.1 消息类型

GSMP支持五种消息：连接管理、端口管理、统计、配置和事件，对每一种消息类型的作用和功能说明如下：

- 连接管理。IP交换机控制器用连接管理建立、删除、修改和证明穿越ATM交换机模块的连接^②。这些是GSMP消息最常用的形式，因为在交换机中根据IP流和后续激发的IFMP消息的到达更新连接状态。表9-2描述了六种连接管理消息类型。

表9-2 连接管理消息类型

连接管理消息类型	描 述
增加分支	建立一个连接或者在点到多点连接中加入新的分支
删除分支	在一个点到多点连接中删除分支或者当该分支是唯一的分支/最后一个分支时，删除该连接
删除树	删除整个连接
检查树	在连接中检查分支的数量
全部删除	删除交换机输入端口的所有连接
转移分支	把单一分支从当前的输出端口 /VPI/VCI的关联关系转移到新的输出端口 /VPI/VCI

图9-11可以更好地说明GSMP为特定IP流建立连接所起的作用。在给上游发送IFMP重定向消息并接收到被分配的非缺省VC（VCI=27）值的流的信元后，IP交换机控制器检查流在下游是否被重新标记。在这个例子中，输出端或下游方向的流已被赋予非缺省VC（VCI=42）值。随后IP交换机控制器向ATM交换机发送GSMP加入分支命令，建立一个“内交换”连接，将输入端4和输入VCI=27与输出端2和输出VCI=42关联起来，结果属于流的信元将通过ATM交换模块被信元交换而旁路了IP交换机控制器。

GSMP连接管理有意思的一点是管理单播和组播连接没有分别。同样的加入分支命令也可用于建立一个单播连接或在已有的点到多点连接中加入新的分支。另外，大多数连接管理消息（除移动分支命令外）仅占一个ATM信元，所以开销是最小的而处理是快速的。

① 在RFC1987的后部定义了GSMP关联子协议和状态表。

② 在GSMP范畴内，一个连接是在同一个ATM交换机内部的输入端口、VPI、VCI和输出端口、VPI、VCI的关联关系。

• 端口管理。端口管理消息用于激发（引起）交换机端口、放下、运行后向循环测试并重新配置。当交换机端口被激发，所有到达特定输入端的连接被删除，只有 IP 交换式连接能获得由 GSMP 控制的在端口的 VPI/VCI 标记空间。新的端口段号产生，该号是在每一次端口被激发后由交换机产生的任意 32 位长的号。依于端口产生的所有 GSMP 请求都必须包含一个相匹配的端口段号，否则请求将被忽略。端口段号能使 IP 交换机控制器与当前端口的激活状态相同步，以保护端口免于处理过时的命令。交换机端口被激发后，端口状态被激活。

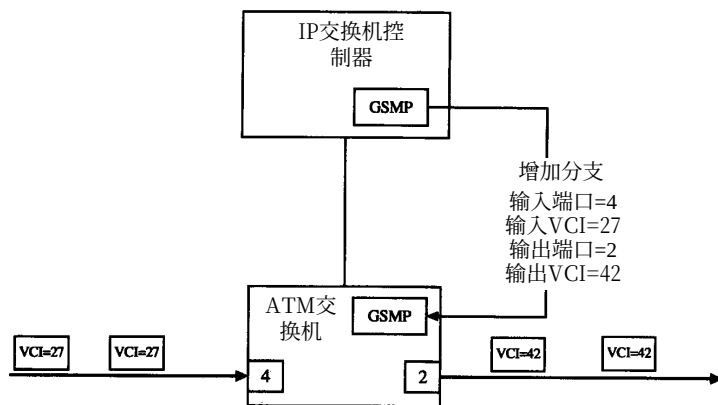


图9-11 GSMP增加分支的例子

• 统计。统计消息能使 IP 交换机控制器要求和接收与端口和连接有关的多种统计计数。IP 交换机控制器能获得的信息有：

- ☐ 输入和输出信元数。
- ☐ 输入和输出帧数。
- ☐ 输入和输出信元丢弃数。
- ☐ 输入和输出帧丢弃数。
- ☐ 输入HEC数。
- ☐ 输入无效的VPI/VCI数。

• 配置。配置消息使 IP 交换机控制器发现 ATM 交换机模块的容量和功能。除交换机名字、交换机类型和微程序语言的版次号外，配置信息可以使 IP 交换机控制器发现端口的特定信息，如支持的 VPI/VCI 标记范围、最大信元率、状态、界面类型及端口优先级^①。

• 事件消息。事件消息能使 ATM 交换机通知 IP 交换机控制器一些事件，如端口打开，端口关闭。

9.4.2 GSMP 版本2

RFC1987定义的GSMP版本1覆盖了标准ATM交换机管理的基本方面，它能使 IP 交换机控制器迅速容易地管理连接和端口状态、接收到有关端口、交换机及连接状况，但不表明也不支持ATM技术和交换机中常有的几个特定区域，如

- ☐ 虚通道交换。

^① RFC1987定义的GSMP版本1.1仅仅假定输出队列具有简单的严格优先级，也可以为一个端口定义多个优先级队列。

- 使用参数控制（Usage Parameter Control—UPC）业务流策略参数。
- 每条VC或每级定时。
- 使能/禁用VC功能，如早期分组丢弃。

GSMP版本2是RFC2297定义的GSMP新版本。除了几处提高的消息类型外，最重要的改进是对QoS的支持。新的QoS消息允许IP交换机控制器为虚通道连接和虚信道连接划分等级，并在每级或每个VC基础上定位资源。这不是一个简单的交错连接交换结构，图9-12可抽象地描述出ATM交换机所提供的QoS功能。

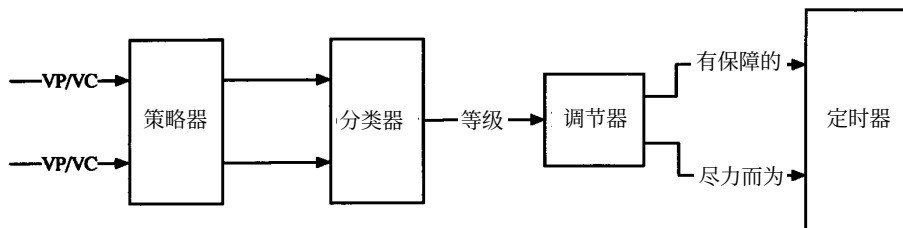


图9-12 GSMP—QoS抽象交换机模型

它由四个主要组件构成：分类器、策略器、调节器、定时器。分类器将多连接（VP或VC）分成单一等级从而QoS可应用于聚集组中。策略器通过可能的标记或丢弃信元来加强在单个级或连接上的业务量。调节器可以运行策略或定形功能。信元由调节器的策略作用评价来决定它们是否与调节参数所规定的速率一致。不一致的信元或被标记、丢弃、或分类定时，而一致的信元接收QoS，过量的信元被处理。调节器的定形功能在每级或每个连接的基础上按调节参数规定的速率定信元步速。定时器在输出端口上，按照QoS对每级或每个连接的要求定时要传送的信元。

除QoS支持外，GSMP版本2还在以下方面有所提高：

- QoS消息。
- 虚通道交换。
- 从交换机端口提取全部连接状态的消息。
- 在交换机端口对VPI和VCI重配置的消息。
- 删除一系列连接的消息。
- 更具体的出错消息定义。
- 避免溢流到接收缓冲器的流控制窗口。
- 对关联损失的定义。
- 对关联协议的提高。
- 命令消息组的规定。
- 多通道链路、以太封装的扩展。
- 版本协商。

9.5 IP交换实例

在以上描述和解释IP交换组件和协议的基础上，下面的例子有助于说明单个IP交换机如何将流映射为ATM连接。

• 缺省转发。图9-13a给出了一个IP交换机及其上游和下游的节点。在相邻的IP交换机之间所有的业务量都是在缺省通道 $VPI/VCI=0/15$ 中转发。当信元到达IP交换机控制器后，将被重新装配为分组、进行第三层处理、再拆成信元、在缺省通道上向下游节点传送。

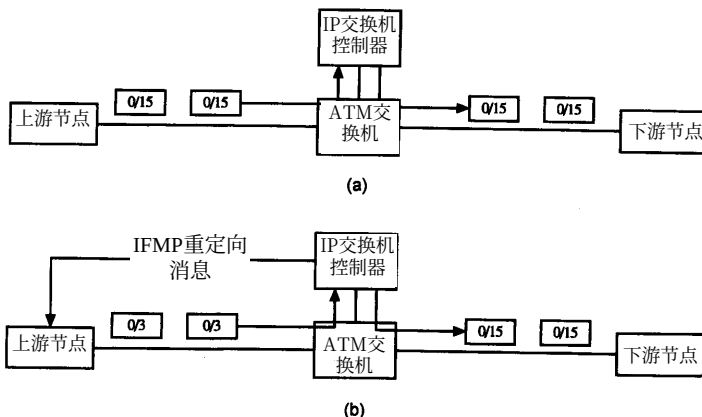


图 9-13

a) 缺省转发 b) 数据流分类

• 流分类。在图9-13b中，根据配置策略，IP交换机控制器检测到一个适于交换的IP流。这个IP交换机向上游节点发送IFMP重定向消息，这个消息中包含了流标识符和新的标签 $VPI/VCI=0/3$ 。上游节点接到重定向消息后开始把流的信元在通道 $VPI/VCI=0/3$ 上发送。

• 下游节点流分类。在图9-14a中，下游节点同样决定对流进行交换。下游IP交换机向本IP交换机控制器发送IFMP重定向消息，其中包含新的标签 $VPI/VCI=0/7$ 。本IP交换机开始在 VPI/VCI 值为0/7的通道上传送流的信元。这时流仍然经过第三层的IP交换机控制器转发但是流的信元在IP交换机的输入和输出端口都获得了新的标记。

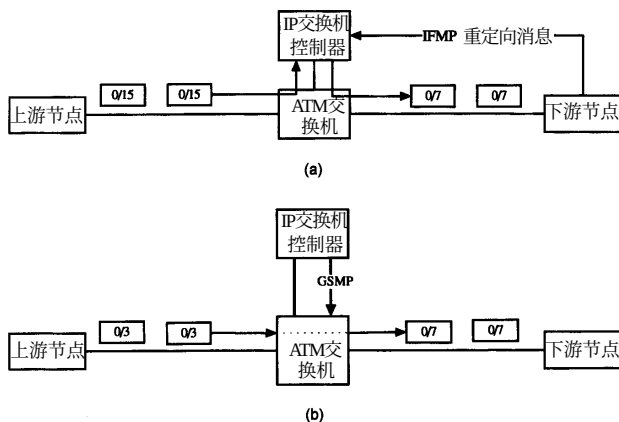


图 9-14

a) 下游节点数据流分类 b) VC连通

• VC连通。一旦IP交换机控制器检测到流在输入和输出端口都重新标记，便发出GSMP ADD BRACH命令把两个VC通道连通起来。连接便完成了，流的信元以ATM交换机构的速度

在IP交换机中转发。

9.6 IP交换和组播

IP组播业务量属于需要大带宽和低时延的应用。一个例子是多方电视会议或者一方对多方的可视报告，这时视频流沿着传送树向下游传送，要求极小的信元丢失率、足够的带宽和低延迟。IP交换机可以用相对直接的方式支持IP组播，与支持单播业务量相似。IFMP重定向消息仍然是从接收端逆流而上，把流映射为新的VC。实际上区别仅在于在网络不同的分叉点。流标识符中的目的地址是组播地址，GSMP根据传递树中下游分叉的数目，执行多个ADD BRANCH操作。图9-15说明了IP交换支持组播的概念。

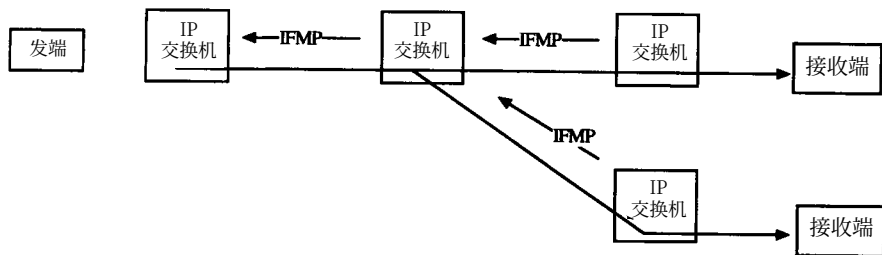


图9-15 IP组播的支持

9.7 IP交换网络实例

从互连的角度，IP交换可以在多种不同的网络环境中运行。它用逐级跳的处理转发IP业务流，从这个角度上它与标准的IP路由器是兼容的。然而，为了加速IP业务流，这个方案中的精心设计的细节要求足够的每个流的交换机的资源，还要求入口和出口的边缘模块兼容IFMP。容易想到的是图9-16所示的校园或小企业的IP网络。这种环境中，基于以太网的工作组通过兼容IFMP的边缘设备接入IP交换骨干网。高速服务器用IFMP适配卡直接连到由IP交换机构成的骨干网。由于通信方的局部性，可能的和实际被交换的IP数据流的总容量不会超过网络中交换机资源的最大容量。总的端到端吞吐量依赖于入口/出口边缘设备的转发能力。这种情形提供的支持与MPOA相似，不同的是IFMP是一种控制协议。

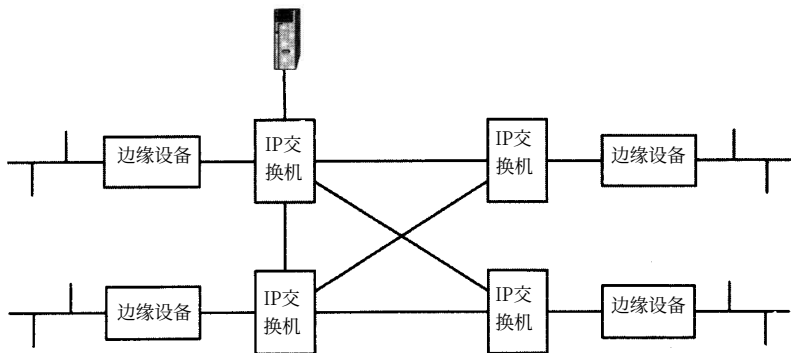


图9-16 企业内部的IP交换

另一种IP交换可能发挥作用的舞台是网络的边缘。IP交换机制可用作流鉴别器或业务量整形器，对某个数据流进行识别、标记和服务，速率较高的接入ISP骨干网或由ISP骨干网进入速率较低的用户环路。考察图9-17所示的两个例子。第一个例子中，IP交换机位于Web服务器组的前端，分类器设计为对要求高优先级的某个数据流进行识别和标记。这使得高优先级的流能够通过高性能的交换通路进入ISP骨干网，而对其他的流仅提供基本的尽力而为的逐级跳的服务。第二个例子非常有趣，使用一个IFMP兼容的本地的宽带调制解调器（电缆或xDSL）向上游的前端传送IFMP重定向消息，使某个IP数据流的业务高速进入目的端。

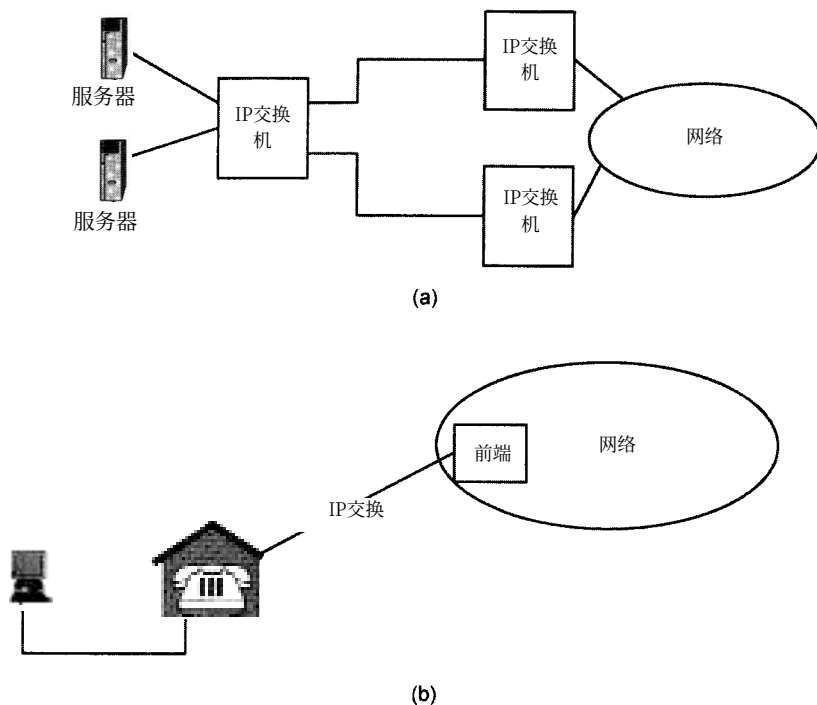


图9-17 IP交换边缘设备

更多的信息

<http://www.ipsilon.com/~pn/>.

McKeown, N., and Lin, S., "A Simulation Study of IP Switching," ACM/Sigcomm Proceedings, 1997.

Newman et al., "IP Switching and Gigabit Routers," IEEE Communications, vol. 35, no. 1, Jan. 1997.

Newman et al., "IP Switching: ATM Under IP," IEEE/ACM Transactions on Networking, April 1998.

第10章 CSR与FANP

IP与ATM技术的结合不是一件容易的事情。IP具有无连接、动态和健壮的特点，ATM具有硬态和面向连接的特性，二者要以某种方式结合起来。ATM提供的大带宽、低时延和 QoS 对IP是无法拒绝的诱惑。早期定义的 ATM上的IP 的解决方案为了保持可扩展性，把 IP对ATM带宽的使用限制在一个子网或者仿真 LAN(LANE)内部，而基于路由器的网络互连限制了吞吐量。NHRP和MPOA在某种程度上放松了这种要求，允许 ATM建立的直通路旁路中间的路由器，因此使IP业务量有更高的带宽。虽然 ATM直通路能够提供大带宽和低时延，但代价是地址解析过程的开销和 IP选路和ATM转发之间可能不连续。Ipsilon的IP交换是革命性的方法，它解决了以上提到的问题，但是需要所有的端系统和中间路由器（IP交换机）用新的 IFMP和 GSMP协议替换基于标准的 ATM论坛协议。当 IETF、ATM论坛和其他的组织正在进行早期开发的努力时，一组优秀的日本研究人员正在开发不同的 IP与ATM结合的解决方案。这种解决方案在1994年末和1995年初以一系列 Internet草案和白皮书的形式出现，它不再把 ATM网络看作一大片云，而是看作由路由器连接在一起的一系列 ATM数据链路（或子网）^①。然而，与传统路由器对每一个分组进行第三层处理所不同，这种体系结构中的路由器能够在第三层按分组转发或者在第二层按信元转发。

这种体系结构的核心是信元交换机路由器（Cell Switch Router, CSR），它是一个通过标准ATM UNI接口与多个ATM子网相连的设备，它运行标准的 IP选路协议（如 OSPF、PIM）并在子网间执行传统的 IP分组转发。另外，它还能够两个不同的 ATM子网间利用第二层的信元交换建立内部CSR直通路，这是通过把来自一个 ATM子网的输入 VC与另一个子网的输出 VC连接起来实现的。

这个解决方案中的另一个部分是数据流属性通告协议（Flow Attribute Notification Protocol, FANP），用来在一对CSR之间把不同颗粒度的IP业务流映射为缺省VC。一旦CSR检测到在上游和下游（相对于数据流的方向）子网中某个 IP数据流已经映射为非缺省的 VC，FANP能够把它们连接起来，产生一个高效的内部 CSR直通路。沿着路由通路，相邻的 CSR完成了把输入和输出非缺省VC连接起来的过程，IP数据流的入口到出口的直通路就建立起来了。

CSR/FANP解决方案在许多方面与 Ipsilon的IP交换相似。两者都是数据流驱动、都提供缺省和直通路选路、都利用了专门的基于对等实体的数据流驱动的控制协议，该控制协议用新的 VPI/VCI重新标识数据流的信元。然而与 IP交换所不同的是 CSR/FANP模型在ATM子网内部继续使用标准的 ATM论坛的信令和选路，并且它包含了经过中间 ATM交换机进行数据流与标记映射的技术，这使得 CSR/FANP比IP交换更容易与 ATM论坛相容。

10.1 体系结构

CSR/FANP体系结构模型开发的目標是利用 ATM信元交换的大带宽和低时延，同时保持 IP

① 在这种体系结构中，每个 ATM数据链路表示一个IP子网，RFC1577/2225所定义的逻辑IP子网就是一个例子。这一章剩余部分将使用 ATM子网这个词表示叠加在 ATM数据链路上的单个的IP子网。

网络的可扩展性和健壮性。为了达到这个目标，CSR/FANP没有像MPOA一样试图在跨越整个ATM网络的传统的端到端的ATM VC上传送IP业务流，而是保持了传统Internet的体系结构，也就是使用IP选路设备进行子网互连。但是在这种模型中，与每个子网相联系的底层的数据链路技术是ATM，路由器（更确切的说是CSR）不仅能够进行传统的IP分组转发，而且能够进行ATM信元交换。利用CSR在子网间转发信元的功能，就可能构建直通过程，旁路中间逐个分组的处理，从而利用了ATM提供的大带宽和低时延。

图10-1所示为CSR/FANP模型。包括主机和边缘设备的输入和输出模块提供了进出CSR网络的手段。CSR把多个ATM子网互连起来，执行基本的IP转发功能，运行标准的单播和组播选路协议。分组采用逐级跳的方式转发，使用起始和终止于CSR设备的一系列缺省通道。或者分组可以在端到端的直通过程上传输，这种直通过程同样起始和终止于CSR设备，并且由一系列连接在一起的VC构成。CSR把某个分别由输入和输出VC传送的数据流联系起来，进行连接过程。把分组或者更确切的说分组组成的数据流映射为专用VC的过程是FANP的功能。

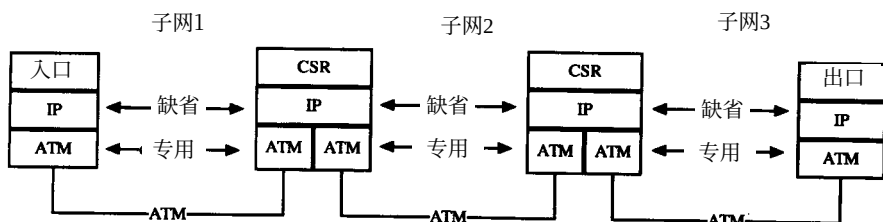


图10-1 FANP模型

CSR/FANP的操作包含几个独立的层面。首先，ATM论坛的信令和选路都限制在子网内部。CSR和其它ATM子网的设备利用这些机制建立缺省VC，或者根据需要建立专用VC。一个合理的假设是，一个ATM子网中包含多个运行标准ATM论坛协议的ATM交换机。所以尽管整个大范围的CSR网络可能利用ATM数据链路技术提供连通性以及可选的大带宽服务，但是ATM VC的管理操作仍然限制在IP子网边界以内。其次，只在子网内部相邻的或者VC相连的模块之间，FANP进行数据流到VC的映射操作。与NHRP不同的是FANP不能跨越子网边界。最后，IP路由由通路上每个CSR自己作出决定，是否对数据流的信元进行交换而不是对每个分组进行第三层的处理。

与其他方法相比，在支持ATM上的IP的同时保持标准Internet模型的操作有几个重要的优点。首先是可扩展性，Internet的存在证明了可扩展性的重要性。逻辑上说CSR/FANP与传统的基于选路的网络没有区别，只是把ATM作为子网中的数据链路技术。其次，CSR/FANP十分可行，并且能够比较直接地把RSVP的数据流映射到提供一定QoS等级的直通过程上。值得注意的是RSVP的PATH和RESV消息按照标准的方式沿着交换式通路传播。CSR能够解释RSVP消息并在子网内部把数据流映射到具有QoS的专用VC上，进而提供输入和输出专用VC上具有期望的QoS的信元交换服务。

另一个不经修改就能够支持的标准IP服务是IP组播。IP路由器负责根据组内成员的信息建立和维护传递树，这在CSR/FANP模型中并没有改变，只是由CSR完成组播路由器的功能。如果在不含路由器的ATM网络上支持大规模组播并且仅使用点到多点的VC，那么要求根节点维护每个叶节点的明确的信息。这将带来扩充性的问题并对资源造成很大负担。

CSR/FANP网络中的缺省通路和直通路都是动态 IP 选路计算的，这是它的另一个重要的优点。实际上缺省通路和直通路含有相同的子网、链路和 CSR。对于大规模 ATM 网络，这种方法减少了 PNNI（或者网络运营者通过配置静态路由表，如 IISP）在一个或多个层次上维护拓扑一致性和准确性的负担。把大型 ATM 网络划分为较小的 ATM 子网使得运行标准和专用的选路协议的不同的 ATM 交换机能够协调工作。例如，一个 ATM 子网可能包含支持 PNNI 的交换机，而另一个子网中的 ATM 交换机可能只能运行 IISP。请注意动态 IP 选路的动态性，一旦一个 CSR 或者链路出现故障，可以为业务流重新选路并且倒换到替代通路上。当然在新通路上的中间子网和 CSR 中执行 FANP 并进行连接过程都有一些时延，然而不需要像 PNNI 中 VC 意外中断时那样重新对端到端的连接进行初始化。

10.2 组件

RFC2098 描述了与 CSR/FANP 有关的定义和组件：

- 数据流。数据流是从某个源端到某个目的端的分组的（单播或组播）序列，与一定的通路和沿途每个路由器为分组提供的（缺省或非缺省）服务相关。数据流的分组通常具有相同的源端地址、目的端地址、可能还有相同的 TCP/UDP 端口号。

- 数据流标识符（数据流 ID）。数据流的分组具有共同的一组 IP 层和传输层的字段。一个数据流 ID 的例子是图 10-2 所示的源/目的端的 IP 地址对。另一个例子是源/目的端的 IP 地址/端口对。

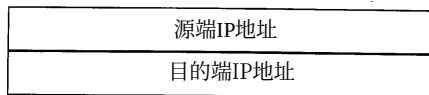


图10-2 数据流ID

- CSR。CSR 是能够进行传统的第三层 IP 转发和第二层信元交换的设备。它与一个或多个 ATM 子网相连，使用任意数量的标准 ATM 接口并在这些接口上运行标准的 ATM 论坛的 UNI 信令。图 10-3 所示为一个通用的 CSR，它包含 CSR 控制点和 ATM 交换机构。设备中 CSR 控制点的作用类似 IP 路由器，运行标准的 IP 单播和组播选路协议。它同时管理 ATM 交换机组件中连接的状态。另外 CSR 控制点必须能够识别 IP 数据流，并在同一 ATM 子网中与对等 CSR 交换 FANP 消息，以便把数据流映射为 VC。对等 CSR 设备通过交换控制信息来管理缺省 VC 并以缺省选路的方式转发分组。对等 CSR 设备还可以管理一个或多个专用的 VC，这些专用的 VC 传输某个 IP 数据流并形成端到端直通路的路网段。

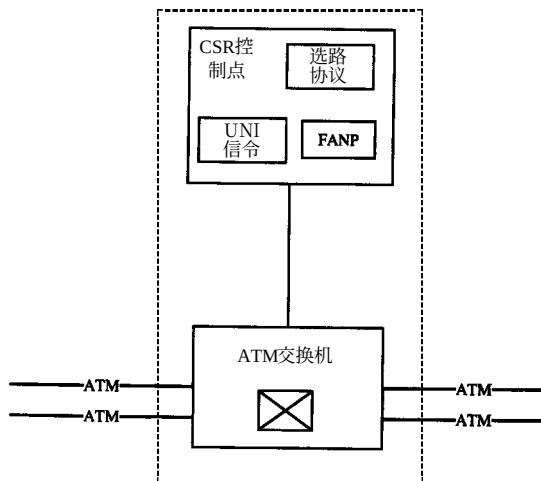


图10-3 信元交换机路由器

- CSR 入口/出口。为了与外部网络互操作并使业务流能够从具有 CSR 的网络中获益，必须在 CSR 网络的边缘放置入口和出口模块。CSR 入口/出口可以是主机、边缘设备、路由器或者 CSR 本身。它作为通过一个和多个 CSR 设备的直通路路径的源和终点。CSR 入口/出口能够交换 FANP 消息，以便把数据流映射到专用的 VC。

• 缺省VC。是两个CSR之间的通用VC，缺省VC提供一种连接IP分组在其上第三层进行逐级跳的转发，通过CSR的网络。从缺省VC上输入的信元装配成分组并进行第三层的处理。CSR也可以利用缺省VC交换控制信息，例如FANP消息和选路协议的更新消息。

• 专用VC。一个专用VC建立在ATM子网内的两个（或者在组播的情形下多个）CSR设备之间，以传输特定IP数据流的信元。本地CSR的规则决定哪一个IP数据流映射到专用VC上。CSR可以把两个专用VC连接起来，以形成直通路径，从而旁路第三层的处理。

• 旁路通路。这是CSR入口到出口的ATM信元交换的连接，由沿着路由通路被一个或多个CSR连接到一起的一系列专用VC构成。旁路通路是CSR描述直通路径的术语^①。旁路通路遵循动态IP选路所计算出的物理通路，区别是在每个中间CSR中进行信元交换而不是在第三层处理。它可以建立尽力而为和基于QoS的旁路通路。旁路通路可以按点到点方式承载单播IP数据流，也可以沿着拓扑树承载组播数据流。图10-4所示为不同CSR连接和旁路通路的概念。

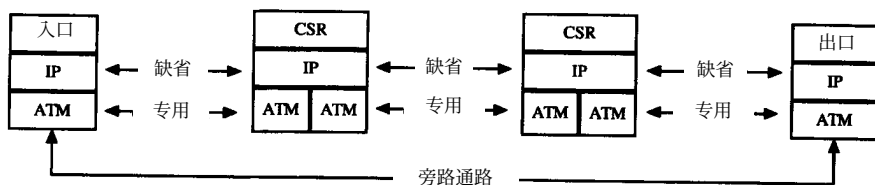


图10-4 CSR连接

• VCID。VCID是唯一的标识符或“句柄”，专用VC的CSR端点使用（共享）该VCID，把某个数据流和专用VC联系在一起。使用VCID而不是VPI/VCI的原因是在通过一个或多个ATM交换机时每个信元的VPI和/或VCI标记可能被替换或者改写。使用端到端的方式没有意义，VPI/VCI可能变化，所以试图把某个IP数据流映射为VPI/VCI将无法工作^②。VCID的格式如图10-5所示，包含上游节点的6字节的ESI和上游节点选定的唯一的6字节本地标识符。

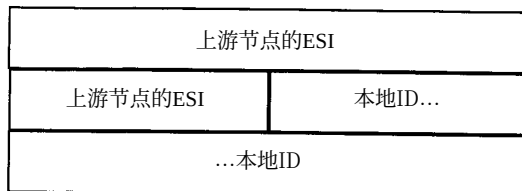


图10-5 VCID格式

10.3 CSR与ATM交换机的互操作

CSR/FANP模型设计的基本出发点是与ATM交换机共存和互操作。与Ipsilon的IP交换不同，CSR/FANP模型能够与ATM交换机协同工作。它是在假定大范围地理区域的网络已经覆盖了ATM数据链路的条件下提出的。它扩展了旁路通路可达的范围，从而提高了整个网络的容量和性能。与Ipsilon的IP交换不同的是，它不必要替换ATM交换机中符合ATM论坛标准的软件，改用专用技术。实际上，CSR引入网络以后，ATM交换、信令和选路保持不变。为了适应ATM交换，CSR/FANP模型引入下面几个特点：

- CSR之间在标准的ATM接口上使用ATM论坛UNI信令通信^③。

① CSR文献中把直通路径称作旁路通路或者直通路径。

② Ipsilon的IP交换把数据流映射为VPI/VCI标记，但是该体系结构不建议采用中间的ATM交换机进行标记替换。在Ipsilon的IP交换模型中，利用点到点的ATM数据链路，IP交换机彼此直接相连。

③ 当以CSR/FANP的方式运行时，CSR中的ATM交换机组件不必支持PNNI，尽管它提供信元交换功能。

• 对等CSR组件（例如入口/出口CSR设备、CSR）交换VCID值，把IP数据流与专用VC关联起来。在对等CSR组件之间与共享的VCID通信的过程称作VCID通知过程。图10-6所示为在两个CSR设备之间使用VCID把数据流N与专用的VC联系在一起的概念。

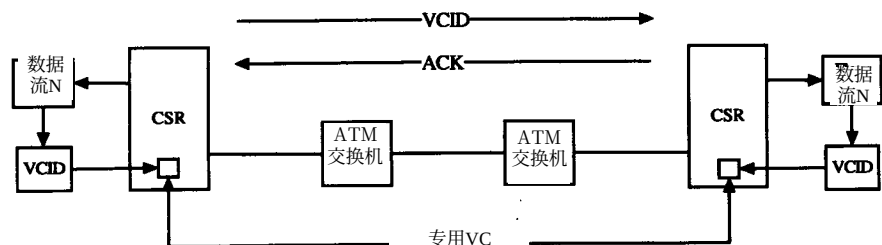


图10-6 在CSR网络中使用VCID值

CSR/FANP操作随着当前在CSR设备之间采用的ATM网络的类型不同而不同。区别在于为传输特定IP数据流提供专用VC的方法，以及对等CSR之间VCID通知的必要性和手段。表10-1归纳了在对等CSR设备之间可能存在的四种不同的ATM配置，并评价了VCID通知所需技术。

表10-1 CSR与ATM交换的互操作

ATM配置	VPI/VCI值是否改变	是否需要VCID	VCID通知方式
点到点透明链路	否	否	不适用
VP隧道	只替换VPI值；VCI保持不变	因为VPI标记被替换，所以建议采用VCID；但是因为VCI值保持不变，所以VCID不是必需的	带内
PVC网络	是	是	带内
SVC网络	是	是	带内或带外

在点到点透明配置中，在专用VC的两端，VPI/VCI标记相同。这种配置的一个例子是两个CSR用点到点链路直联起来，中间没有ATM交换机。这时没有必要使用VCID把IP数据流与专用VC联系起来，实际的VPI/VCI值本身就足够了，因为它在两端相同。第二个例子是VP通道，在信元穿过ATM交换机网络时VPI值可能被替换。由于VPI/VCI的一部分发生了变化，推荐使用VCID通知过程，尽管这一过程不是必须的。因为在VP交换机中每个ATM信元的VCI值保持不变。所以如果在数据流与标记的关联的实现中标记是VCI值，就可以在VP通道上支持它^①。最后两个例子中，由于中间ATM交换机进行标记替换的过程，VCI值在两端不同。这时需要VCID值，并且必须执行VCID通知过程以便专用VC的两端能够了解特定VCID值。

VCID通知过程能够采用多种方式中的一种。带内通知是VCID在数据流所使用的VC上传输。换句话说，首先在两个CSR设备之间建立专用VC，然后VCID通知过程在这条VC上进行。带外VCID通知过程涉及使用单独的“通道”或另一条通路，VCID在对等CSR模块之间通知。

通过在ATM SVC建立消息中包含VCID（或一些临时句柄）可执行带外VCID通知，具体来说，VCID值包含在SETUP消息的信息元素中，SETUP消息在CSR源和终点之间传递。VCID放在宽带底层信息（Broadband Lower Layer Information, BLLI）IE的“用户定义的第二层协议信息字段”。然而，为BLLI IE分配的7比特对于一对CSR设备之间专用VC的编号可

① 这是包括Ipsilon、标记交换和ARIS等在内的所有IP交换解决方案所支持的标准技术。

能是不够的。一种可能的解决方案是在 BLLI IE 中定义一个临时句柄，作为“占位者”直到随之进行的实际 VCID 值的明确交换，实际 VCID 值替代临时句柄。

10.4 数据流属性通知协议

CSR/FANP 模型使得映射为专用 VC 的 IP 数据流获得高吞吐量和低时延。除了提供选路和信元交换转发能力的 CSR 设备以外，RFC2129 定义的 FANP 版本 1 是 CSR/FANP 模型的另一个重要组成部分。FANP 是控制协议，CSR 设备用它交换建立专用 VC 必要的信息，把 IP 数据流与专用 VC 关联起来，然后把 IP 数据流的信元放在专用 VC 上。FANP 消息交换的目的是为每个数据流在 ATM 子网内部 CSR 设备之间建立专用的 VC，以便最终能够形成入口到出口的旁路通路的一段。

本质上 FANP 是数据流驱动的控制协议。这意味着 FANP 过程是由 IP 数据流的到达所触发的，这与 IFMP 在许多方面是相似的。数据流由什么组成或者更一般地说什么类型的数据能够触发 FANP 过程则是根据每个 CSR 内部配置的规则，但是这些规则在整个网络内部应该一致。“触发”分组的例子可能是 TCP SYN 报文（指明建立 TCP 连接）、包含特定端口号的分组（例如 HTTP 的端口为 80）、包含相同源和目的 IP 地址的分组或者在一定时间间隔内到达的具有相同头信息的一组分。另外像 RSVP RESV 消息等特定控制消息的出现也可能触发 FANP 过程。图 10-7 所示为 3 个中间 CSR 设备之间的典型的 FANP 消息流。

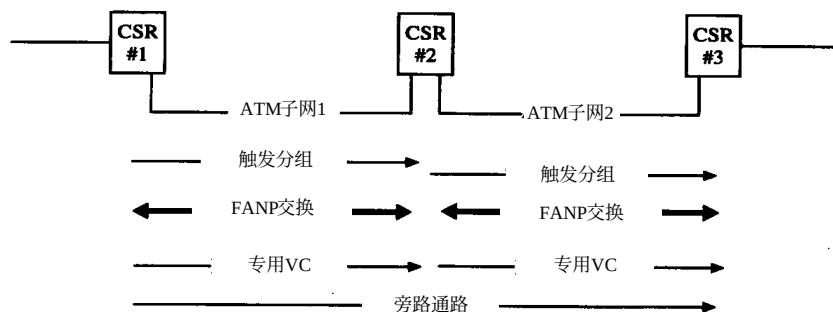


图10-7 FANP过程

触发分组到达 CSR1 引发了与 CSR2 交换 FANP 消息。同一数据流在子网 2 中 CSR2 和 CSR3 之间触发了 FANP 消息交换，这与子网 1 内部的 CSR1 和 CSR2 之间的 FANP 动作无关。与 Ipsilon 方法不同，这里是上游 CSR 引发 FANP 过程。基本的 FANP 过程包含专用 VC 的分配、VCID 协商和数据流 ID 通知。数据流 ID 通知是把数据流 ID 与 VCID 联系起来进而与专用 VC 联系起来的。一旦 CSR 检测到输入和输出端口上某个数据流都被放置在专用 VC 上，CSR 控制点命令 ATM 交换机组件对连接表中的两个表项进行“交叉连接”，进而在设备内部形成信元交换通路。FANP 还包含检测是否专用 VC 上的数据流已经停止了一段时间（死亡时间）的机制。如果数据流的确停止了，那么 FANP 将删除相应的数据流 ID 和 CSR 中的 VCID 表项并释放专用 VC。

FANP 定义了 CSR 之间管理专用 VC 的五个过程：

- 专用 VC 选择。
- VCID 协商。
- 数据流 ID 通知。

- 专用VC更新。
- 专用VC删除。

10.4.1 专用VC选择

一旦接受到触发分组，CSR必须执行的第一步是选择专用VC。有几种实现方法。第一种技术叫做VC池，这种方法中对等CSR设备之间预先建立VC池。可以通过人工设置（PVC）或者通过信令（SVC）。CSR在未用VC池中选择一个VC，作为特定数据流的专用VC。第二种技术是通过SVC信令实现的，这种方法中对等CSR设备之间使用标准的ATM论坛信令和选路来建立SVC。值得注意的是这种方法中源CSR在SVC建立之前必须对目的CSR的IP和ATM地址进行解析。这可以通过前面提到的传统式的ATM上的IP或者局域网仿真的方法实现。第三种上游CSR设备提供专用VC的方法是在预先配置的已经在对CSR设备之间建立起来的PVP中选择一个未用的VCI。

VC池方法的优点是VC已经建立起来，在FANP过程进行下一步骤之前没有时延。于是减小了把数据流放置在专用VC上以及建立旁路通路的时延。另一个优点是CSR ATM子网（CSR UNI端点和ATM交换机）可以比较容易的吸收每个数据流的FANP消息风暴。因为CSR设备和ATM交换机没有必要使用UNI/PNNI信令和选路^①。在确保VC池中有足够的未用VC的前提下，可以设置一个未用VC的低位，以便当未用VC的数量低于某个阈值时，CSR将用额外的SVC来填充VC池。

VC池方法的缺点是它消耗了CSR设备和ATM网络的VC资源。在VC资源紧张或者配置VC昂贵（例如广域网）的网络中动态的按需的SVC方法更有吸引力。另一个应考虑的问题是VC池中VC的QoS可能与特定数据流期望的QoS不一致。如果某个数据流要求特定的QoS等级，最简单和有效的方法是建立具有QoS的SVC。图10-8所示为RSVP RESV消息触发建立具有特定QoS的专用VC的例子。

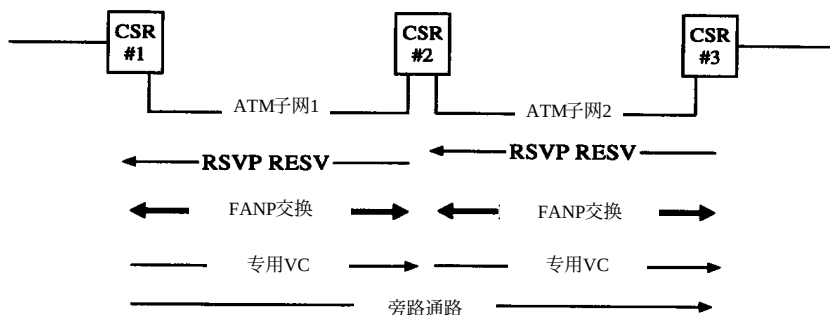


图10-8 RSVP驱动的FANP专用VC的建立

10.4.2 VCID协商

选择专用VC之后进行VCID协商过程。如果存在中间ATM交换机，专用VC端点的VPI/VCI标记不同，则需要这一过程。在上游CSR向下游CSR在专用VC上发出PROPOSE消息

^① 随着快速交换机处理器的出现，SVC容量及其建立/拆除的性能都提高了，因此这不再是一个重要的问题。

后, 开始 VCID协商过程。PROPOSE消息包含 VCID和发端和收端 CSR的IP地址。如果下游 CSR接受这个值, 它在缺省 VC上向上游 CSR发出PROPOSE ACK消息。上游和下游的 CSR设备这时具有相同的与专用 VC联系在一起的VCID。

10.4.3 数据流ID通知

FANP过程的下一个步骤是数据流 ID通知。上游 CSR在缺省 VC上向下游 CSR发出 OFFER消息。图10-9所示为OFFER消息, 包含与专用 VC关联的VCID和将在专用 VC上传输的IP流的流ID。另外还包含一个刷新闻隔字段, 说明下游 CSR在多长的时间间隔内需要向上游 CSR发出READY消息, 指示专用 VC上仍有信元在传输。

版本	操作代码	校验和
VCID类型	数据流ID 类型	刷新闻隔
VCID		
数据流ID		

图10-9 OFFER 消息的格式

10.4.4 专用VC刷新

下游CSR在专用VC上接受信元, 定期向上游 CSR发送READY消息, 通知上游 CSR可以继续专用 VC上发送IP数据流的信元。当专用 VC上的信元流停止后, 下游 CSR将不再向上游 CSR发送READY消息, 暗示流不再是激活状态。如果上游 CSR在一定的时间间隔 (死亡期) 内没有收到READY消息, 将认为数据流不再是激活状态, 专用 VC可以释放。

图10-10所示为FANP的VCID协商、数据流ID通知和专用VC更新的流程图。

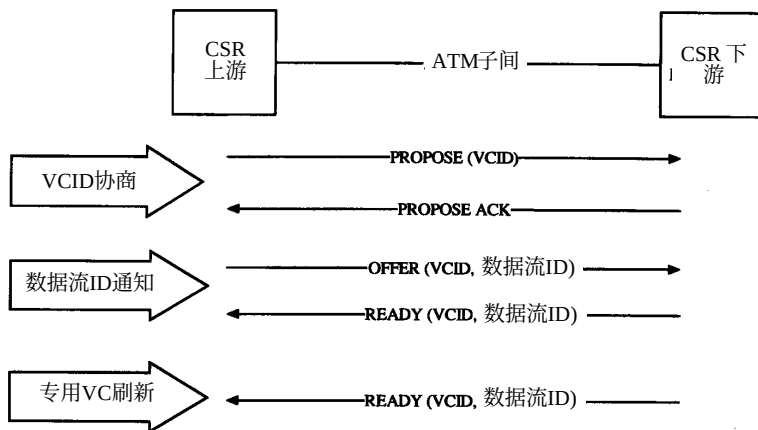


图10-10 FANP数据流

10.4.5 专用VC删除

当上游CSR检测到专用VC不再使用, 将根据ATM配置采取几种操作中的一种删除或者释放专用VC。如果专用VC是PVC, 上游CSR向下游CSR发出REMOVE消息, 其中包含VCID和流ID。结果是删除了与VC相关的数据流ID与VCID的映射关系, 释放后该VC可以被其他数据流使用。如果专用VC是SVC, ATM信令可以方便地拆除连接。

图10-11所示为FANP删除专用VC过程的流程图。

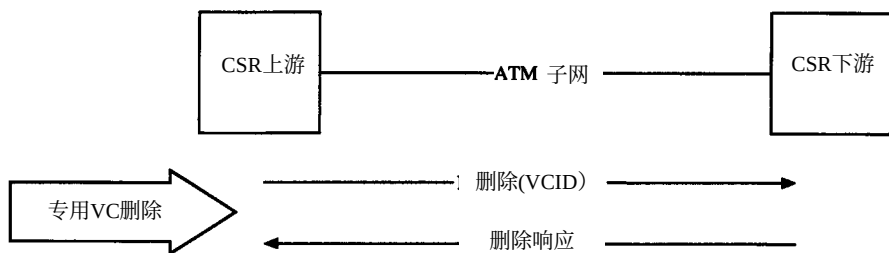


图10-11 FANP专用VC的删除

10.5 版本2的数据流属性通知协议

可能受其他一些互相竞争的方法（例如标记分布协议和 ARIS）启发，CSR/FANP的开发者开发了FANP的新的一个版本。除了进一步定义了以上所讨论的特定 ATM配置下的工作方式，版本2的FANP还引入了邻机发现（Neighbor Discovery—ND）协议，并说明了通过网络中的CSR设备建立旁路通路的两种操作模式。CSR使用FANP邻机发现协议来发现 ATM子网中的任何兼容版本2的FANP的设备，在CSR邻机之间交换特定的涉及 FANP的协议属性，并监测和验证CSR邻机所维护的状态的一致性。版本 2的FANP的两个互相区别的操作模式是分布式控制（Distributed Control—DC）和入口控制（Ingress Control—IC）。在分布式控制模式中，在路由通路沿途的CSR设备对之间FANP操作的发起和运行是独立和分布式的。在入口控制模式中，FANP操作由入口CSR设备发起，并沿着路由通路顺流而下的传播，直到出口 CSR。另外，版本2的FANP还定义了共同的分组头格式、几个新的消息以及一组新的数据流 ID，这些数据流 ID使得CSR/FANP能够支持数据流驱动、拓扑驱动和请求驱动的建立直通过径的方法。采用这些改进后CSR/FANP模型被认为既适用于数据流丰富的环境（校园网、企业内部网），也适用于数据流挑战的环境（大型 IP网络）。

10.5.1 邻机发现

在采用邻机发现之前，除非使用人工配置，否则共享相同 ATM子网的CSR设备无法验证其他CSR设备的存在或者对方的功能。例如，上游 CSR设备怎样判断下游 CSR能够支持的VCID的范围？版本2的FANP中邻机发现协议解决了这个问题，使得共享相同网络的 CSR设备能够按照下列方式工作：

- 自动监测网络中其他支持 FANP的设备的存在。
- 交换FANP协议属性，例如对入口控制或者分布式控制操作模式的支持。
- 验证CSR邻机所维护的状态信息的一致性。

在同一 IP子网内部邻机发现消息在 CSR设备之间交换。邻机发现协议能够在点到点和支持组播的网络上工作。在点到点的链路中，起始端的 CSR向邻机发送INIT消息。INIT消息中包含发端的本地标识符（本地ID）、描述FANP操作所支持模式的能力对象以及 VCID范围对象。收端用RECOGNIZE消息给予确认，其中包含收端的本地 ID、发端的本地 ID、能力对象以及 VCID范围对象。发送端接收到 RECOGNIZE消息并确认自己的本地 ID包含在该消息中，如果这样则向接收端发送 KEEP_ALIVE消息，其中包含接收端的本地 ID。当收发端的 CSR设备确

认彼此的本地ID已经交换之后，两者之间建立了一种操作状态。只有当收发端都处于操作状态，专用VC才能建立或者释放。CSR设备定期与对等体交换KEEP_ALIVE消息以判断状态的变化。

在支持组播的网络中，起始端的CSR发送HELLO消息，其中包含了发送端的本地ID。如果发送端是骨干或者内部路由器，HELLO消息发送到所有节点地址（224.0.0.1），否则发送到所有路由器地址（224.0.0.2）。当邻机CSR设备接收到HELLO消息，它向发送端发出INIT消息，这个过程与上面描述的过程基本相同。

图10-12所示为点到点和组播网络中典型的邻机发现的消息流。

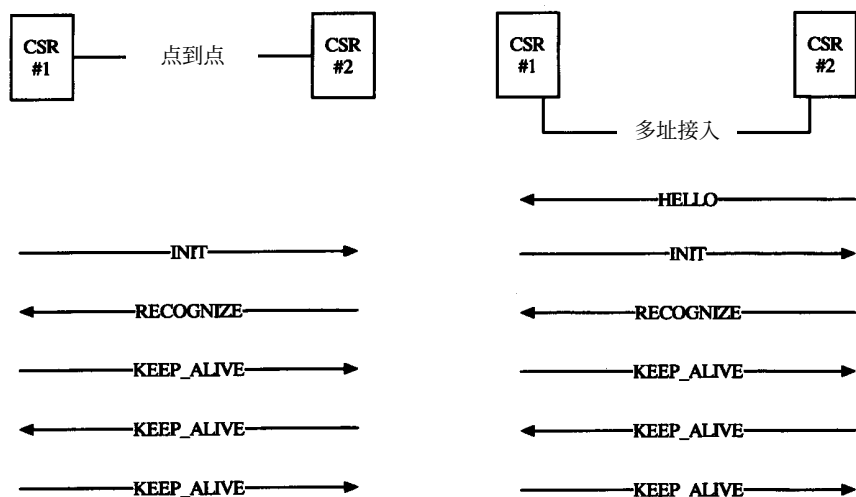


图10-12 FANP 版本2的邻机发现数据流

10.5.2 分布控制模式

在分布式控制模式的操作中，沿着路由通路每个CSR设备用独立和分布式的方式各自发起FANP操作。一旦收到触发分组或者根据其他预先定义的机制，ATM子网中的上游CSR将与下游CSR交换FANP消息，把某个数据流映射到专用VC上。执行这个操作时，并不强制下游

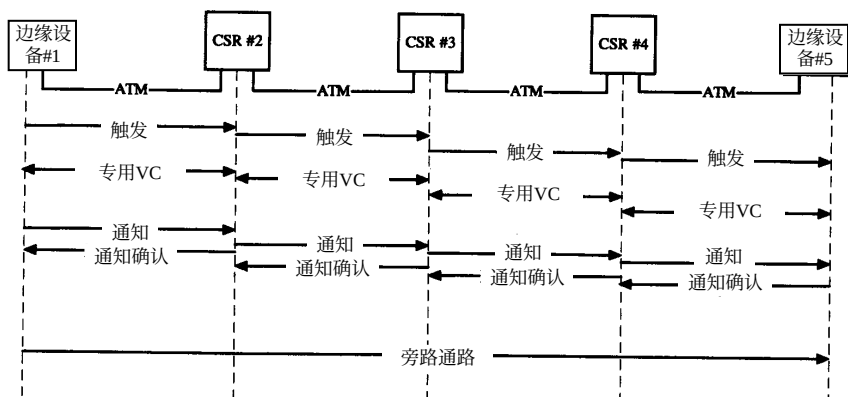


图10-13 FANP 版本2的DC模式操作

CSR引发与其各自下游的CSR交换FANP消息。然而在整个CSR网络中要保持触发机制的一致性，否则始终无法形成入口到出口的旁路通路。尽管与上游 FANP事件没有依赖关系，在相邻的上游ATM子网中触发FANP过程的事件很可能将在下游CSR引发相似的过程。

图10-13所示为分布式控制模式中通过网络中五个支持 FANP的设备建立旁路通路的过程。开始阶段，某个IP数据流或者触发分组（例如 TCP SYN）的到达在四个单独的 ATM子网中的每一个都独立得触发了FANP。FANP过程包含下列步骤：

- 选择专用VC。如前面讨论的，专用VC可以用ATM的SVC信令过程建立，也可以从预先设置的未用专用VC池中选择，或者在VP情况下指派未用的VCI。如果相邻的CSR设备之间的链路是点到点链路并且不经过中间ATM交换机，则没必要进行这个步骤。

- VCID通知。这个过程通知专用VC两端之间共用的唯一的VCID值，这可以用交换带内的PROPOSE消息进行，或者用NOTIFY消息传送VCID及SVC建立请求的BLLI IE中的临时VCID句柄。如果相邻的CSR设备之间的链路是点到点链路并且不经过中间ATM交换机，则没必要进行这个步骤。

- 数据流ID通知。NOTIFY消息中包含数据流ID与VCID的映射关系，在专用VC上从上游CSR向下游邻机发送，可能在同一NOTIFY消息中把VCID与数据流ID编组。

一旦每个CSR发现输入和输出专用VC是与同一数据流ID联系，两者可以连接起来，如果路由通路沿途的其他CSR设备完成了这个操作，那么就可以形成入口到出口的旁路通路。如果拓扑变化或者需要从旁路通路中删除数据流，释放专用VC后将执行一系列数据流ID和VCID的删除过程。由于旁路通路是在每个子网中被CSR分别建立，分布式控制模式有利于对网络拓扑变化、组播组成员变化以及RSVP预留状态的快速响应。

10.5.3 入口控制模式

在入口控制模式的操作中，发生在入口CSR设备的触发事件将引发其下游邻机的FANP过程，然后下游邻机进一步引发自己的下游邻机的相似的FANP过程，如此重复。这个过程在CSR设备和ATM子网构成的路由通路上重复进行直到出口CSR设备。具体来说，入口CSR产生FANP的NOTIFY消息并向下游邻机发送，通知数据流与VC的映射。与分布式控制模式不同，下游邻机不是对这个消息进行应答而是记录数据流ID和专用VC信息，然后产生NOTIFY消息

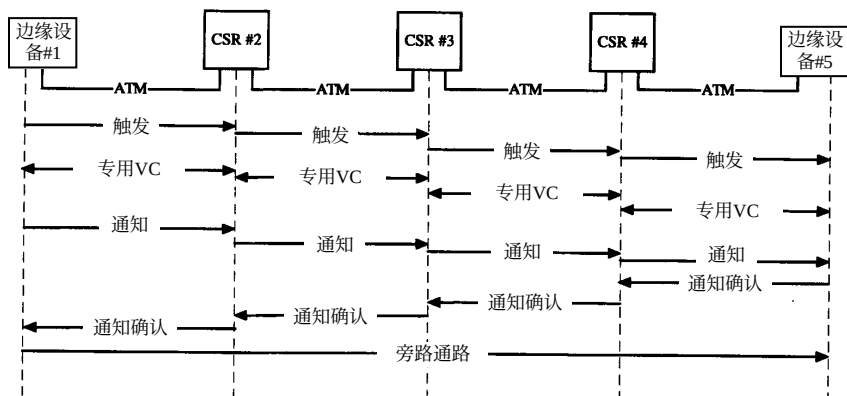


图10-14 FANP版本2的IC模式操作

并向自己下游邻机发送。继续这个过程直到 NOTIFY消息到达出口CSR，这时出口CSR将发送 NOTIFY ACK消息，经过逐级跳的逆流而上的转发直到入口 CSR。沿返回通路的每个 CSR，把对应相同数据流ID的输入和输出专用 VC连接起来，形成CSR内部的直通路，直通路共同组成入口到出口的旁路通路。图 10-14所示为入口控制模式中在五个支持 FANP的设备所构成的网络中建立旁路通路的过程。可见这是一个入口到出口的过程。

入口控制模式的重要特点是能够创建无环路的旁路通路，显然这可以通过消除环路的一个点很容易的实现。工作在入口控制模式 FANP的CSR设备可使用两个方法：

- 跳数检查。如果FANP消息中的跳数值超过了预先设置的阈值，CSR将拒绝这个消息。
- 数据流ID检查。如果NOTIFY消息包含已经注册过的数据流ID的信息，CSR将拒绝这个消息。

10.6 CSR和组播

由于CSR/FANP采用Internet模型，因此可支持 IP组播。在ATM子网的内部，从上游 CSR到一个或者多个属于某个组播组的下游 CSR设备（和主机）建立点到点或者点到多点的 VC，进而在组播传递树上分叉。像单播业务流一样，当 CSR监测到一个数据流ID（包含组播组地址）把入口专用 VC与一个或者多个出口专用 VC关联起来，便进行 VC连接过程，从而形成CSR内部的直通路。向点到多点 VC中增加叶节点可以由组播数据的到达、PIM-JOIN消息、IGMP报告或者当CSR包含MARS时的MARS JOIN消息而触发。图10-15所示为组播旁路通路。

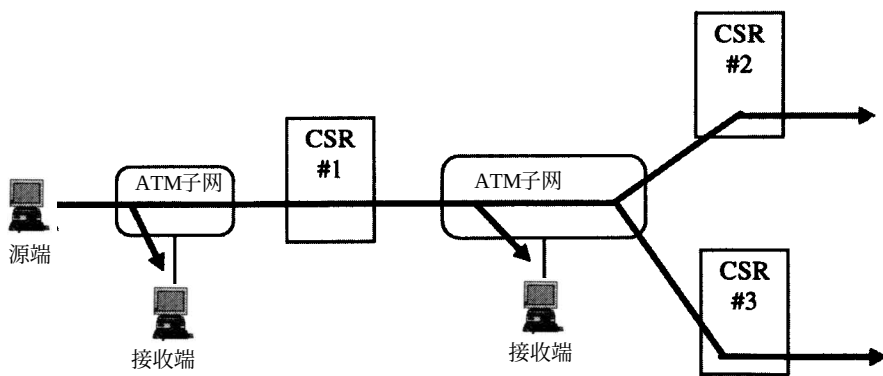


图10-15 CSR组播旁路通路

10.7 CSR网络实例

CSR设备可以在多种不同的网络环境中应用，第一种是连接到同一 ATM网络上的大量边缘路由器。通常需要在任意的入口和出口路由器对之间配置全互联的 PVC以优化性能，然而这需要 $O(N^2)$ 的VC以及路由器、交换机和所涉及的操作成员内部与之成正比的资源消耗。把CSR设备放置在骨干网络可以减小对全互联 VC的要求，减小了整个网络中传播的和路由器处理的控制信息的流量，同时仍然提供在任意入口和出口路由器对之间建立直接 VC（旁路通路）的能力。当然，与所有专用方法一样，直通路的最大范围是控制机制所作用的范围。因此，这种环境下理想的情形是边缘路由器使用 FANP，见图10-16。

另一种适于使用CSR的环境是扩展的校园骨干网或者城域网（Metropolitan Area Network

—MAN)，其中既包含高层的IP转发设备也包含直接使用ATM的设备。例如，若干由交换式10/100/1000兆以太网组成的不同的校园工作组通过上行ATM链路连接到包含CSR的四个集线器之一，以OC12或者OC48运行的标准ATM链路又把CSR设备组成了骨干网。值得注意的是四个校园区域中的每一个都包含直接使用ATM的设备。这种情况下，IP业务流转发到骨干网，并按照业务类型和网络配置沿着标准的逐级跳的路由通路或者信元交换旁路通路转发。

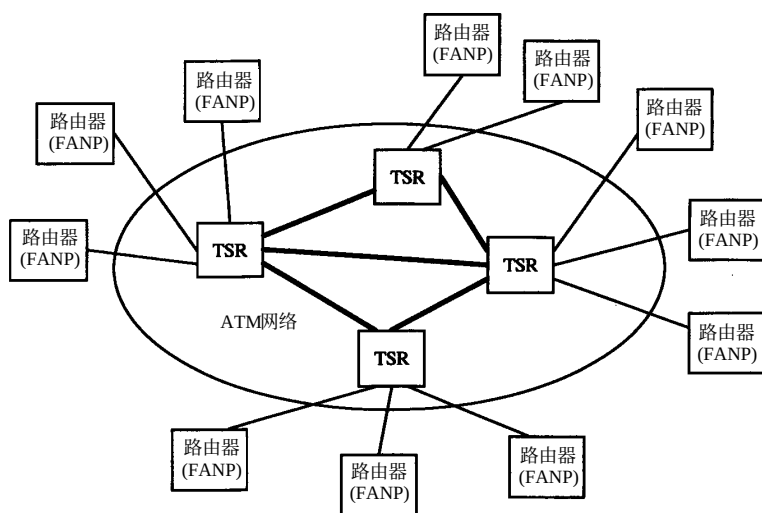


图10-16 大型ATM网络中的CSR

更多的信息

Esaki et al., "Internetworking Based on Cell Switch Router—Architecture and Protocol Overview," IEEE Proceedings, vol. 85, no. 12, Dec. 1997.

<ftp://ftp.wide.toshiba.co.jp/pub/csr/>

第11章 标记交换

Ipsilon公司的IP交换和东芝公司的CSR/FANP都是数据流驱动的IP交换的应用，也就是说，它们按照独立的业务流的到达来安排交换机资源。具体来说，分配标签（VPI/VCI）的动作以及把数据流映射到VC上的信令消息（IFMP和FANP）的交换过程都是独立于单个IP数据流而进行的。两种方法都保持了Internet模型的扩展性以及第三层按照逐级跳方式对所有业务流进行转发，两种方法都引入了特定的控制协议，把IP数据流转移到端到端的直通路径上。有意或无意中这些解决方案的设计者都选择了数据流驱动的方法，其原因如下。首先这是一种适应Internet模型的方法，在Internet模型中缺省操作是逐级跳选路。另外是对独立的IP业务流进行分类、从路由通路把它们分离出来并与面向连接的ATM联系起来从概念上更直接。无论什么原因，Ipsilon的IP交换以及CSR/FANP为IP业务流提供了所期望的大带宽和低时延。在Internet成指数增长背景下，CSR/FANP模型的出现以及随后工业界对于Ipsilon的IP交换的狂热至少说明了对于提高IP选路性能的密切关注。

1996年秋天，Cisco公司提出了一套通用和本质上不同的解决方案，它集成了选路和交换。正如它的名字一样，标记交换使用控制协议信息把IP业务流映射到交换通路上，而不是对每个数据流进行分类并映射到动态建立的直通路径上。具体来说，标记交换依据选路协议信息（如目的端地址前缀）并沿着路由通路把相应的标记分配给有关的设备^①。去往特定目的端的分组被添加了适当的标记并根据标记的内容在标记交换网络中转发。实质上，标记交换用简单的第二层的标记表查询和替换操作代替了标准的第三层的路由表查询，并且在某个路由通路上分配标记以构建交换通路。

标记交换在叫做标记交换路由器（Tag Switch Router—TSR）的设备内进行。TSR支持标准的单播和组播选路协议，如OSPF和PIM，并且可能具有沿着缺省路由通路进行IP业务流转发的功能。TSR设备使用附加的控制协议TDP来分发标记关联，即地址前缀与标记的映射关系。在TSR设备的网络中分发与目的地址前缀相联系的标记将创建通往目的端的交换通路（直通路径）。与标准的路由表查询相比，按照所携带的标记进行标记替换并转发分组改善了性能。标记交换使用与ATM交换和帧中继交换相同的标签替换和转发机制。对于参与标记分发的网络层业务，标记交换是透明的，这使得任意数量的不同的网络层功能能够映射到简单和快速的转发机制中，而且这种转发机制的性能和扩展性已经在大型的帧中继和ATM网络中得到证明。

与前面讨论的方法相区别的是：标记交换是拓扑驱动的IP交换协议，换句话说，通过路由表的更新通告了目的网络前缀的存在，从而创建了一条通往该目的端的交换通路。标签交换转发方法不仅包含在ATM中，而且包含在帧中继、PPP和实际上任何的介质封装中，标记交换使标签交换转发方法更一般化。标记交换的基本目标是提高骨干路由器的转发性能，它使用了简单的按照标记的标签替换转发功能，并把不同的网络层选路服务（例如，单播、组播、COS等等）与这种标签替换转发的机制联系起来，同时保持与介质无关。

^① 标记是标签的别名，在这一章中标记和标签互换使用。

11.1 体系结构

标记交换的体系结构可以从几个不同的层面来看。在概念层面上，它是为提高性能以及大规模选路系统所支持的业务而设计的。考虑到大型 ISP 网络中骨干路由器在任意时刻可能有成百上千的数据流通过，所有这些数据流的共同的要素是它们都是去往选路系统知晓的某个目的地址前缀。一种简单的提高整个网络性能的技术是把标记分配给所有的路由器，以便加速所有通过选路系统的数据流的转发过程。另外，标记的创建和分发与特定数据业务流的到达无关，主要是控制信息驱动的，或者换句话说标记的创建和分发是依靠反映网络拓扑变化的选路协议的更新消息。

在动态 IP 选路协议根据目的端地址选择的通路上转发数据流未必是最佳的，例如，应该采用在整个网络中分配业务量负荷的方法，从而使业务流均摊在几个费用相同或者相近的通路上。通过一系列选定的节点和链路的通路可以提供非缺省服务，例如安全监测、业务量计费或者一些其他特殊业务。

另一个看待标记交换的层面是 TSR 的转发功能。典型的路由器要执行基于目的地址的路由表查询，以及 TTL 递减、分组头校验和介质转换。有效的转发业务流不仅依赖于路由器自身的能力，而且依赖于选路协议计算的通往目的地的最佳的下一跳。标记交换简化了转发功能，它按照每个分组中固定长度的标记的内容进行简单的查表和标签替换操作，从而替代了标准的路由表查询。图 11-1 表示了这个基本概念。标记值为 4 的输入分组查询连接表，确定相应的输出端口和输出标记，然后替换标记并在端口 6 发送。值得注意的是，分组的输出链路是依据标记本身而不是目的地前缀。

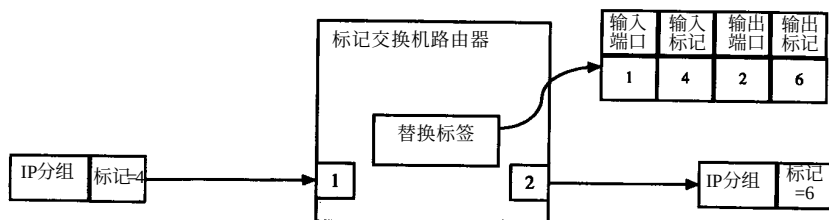


图11-1 基于标记进行标签替换

标记交换的一个重要的结构特征是把标记交换转发操作从网络层的控制功能中分离出来。二者的分离是一个深思熟虑的设计，使得网络运营者能够把若干当前和未来的业务与简单和可扩展的转发机制联系起来，例如目的地选路、组播选路以及显式选路等特定业务与一组标记联系起来，当这些标记在网络中分配时，将形成针对每一种业务的端到端的交换通路。尽管业务可能不同，但是基本的转发机制仍然保持不变。这样，如果引入新的网络层控制功能，就不必重新优化或者升级转发通路上的组件和设备。当发生不可预见的必要的网络层的变化时，已有的投资可以得到保护。例如，突然需要引入 IPv6 以获得更大的地址空间时，不需要对现有的转发通路进行任何修改^①。

标记交换由两个基本组件组成：控制和转发。控制组件负责在标记交换机设备中创建和

^① 这可能过分简化了。现实中，设置在标记交换网络边缘的设备仍然需要在第三层转发分组并且需要支持网络层增加的改进。

维护一组正确的标记，标记的分发是通过独立的控制协议（如 TDP）或者经过修改的能在 TSR 设备之间传送标记的现有的控制协议（如 RSVP、PIM）来完成。转发组件使用分组内部的标记信息和标记交换机设备中存储的标记信息执行分组转发过程。尽管标记的形式可能随着物理介质的不同而不同（例如，VPI/VCI、DLCI等），但是基本的转发机制保持不变。这意味着标记信息和标记交换能在任何数据链路技术上操作，而不是限于某种数据链路实现。

图11-2所示为把标记交换操作应用于基本的按照目的地选路的实例。网络由三个内部 TSR 设备、输入 TSR 设备和输出 TSR 设备构成。每个 TSR 都包含目的网络 184.16/16 的路由表表项。通往目的网络沿途的下游 TSR 在输入端口分配一个标记，然后向上游 TSR 设备发送标记关联 <184.16/16, tag>。上游 TSR 设备接到该消息并在输出端口上设置相应的标记。标记关联交换是各自执行的，可以创建入口到出口的直通过程，传送所有去往网络 184.16/16 的业务流。

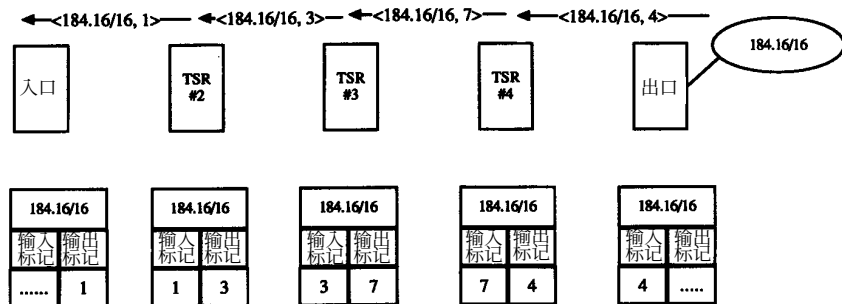


图11-2 基于目的地址的标记交换实例

从这个例子中可以看到几个重要的标记交换的功能。首先，直通过程是根据控制信息而不是数据流来建立的，换句话说，虽然在标记交换式通路上可能没有去往特定目的网络的数据流，但是仍然需要分配交换机资源。一旦确实有分组去往该目的网络时，就可以在已经存在的直通过程上立刻转发，而无需进行数据流分类以判断是否建立直通过程。因此没有建立交换通路的时延，因为交换通路已经存在。其次，入口和出口的 TSR 设备需要执行标准的第三层处理。入口 TSR 设备在发往网络中的分组上添加标记时查询转发表，而出口 TSR 设备删除标记并按照选路协议计算的下一跳转发分组。再次，对每个分组，内部 TSR 设备不再进行第三层处理。可以推测，这将提高转发性能并能够把标记与若干不同的网络层业务联系在一起。

11.2 组件

标记交换模型包含若干不同的功能组件和设备定义，解释如下：

- 控制组件。控制组件负责在一组 TSR 设备之间产生和维护标记的一致性。创建标记涉及分配标记然后把它与特定目的地关联。目的地可以是主机地址、地址前缀、组播组地址或者只是任何关于网络层的信息。标记关联的分配可以使用 TDP 或者在现有控制协议如 RSVP 或者 PIM 上通过“捎带”来实现。

- 转发组件。转发组件利用分组中的标记以及每个 TSR 设备中的标记信息库（TIB）来转发分组。具体来说，当 TSR 接收到包含标记的分组，它按照标记来检索 TIB 的表项。TIB 的表项包含输入标记、输出标记、输出接口以及一些输出数据链路的信息或者封装信息。如果分组中的标记与 TIB 中的一个输入标记匹配，则对每个相应的输出表项，进行标记替换，并应用

数据链路层的信息，在输出链路上发送分组。图 11-3所示为转发操作的例子。

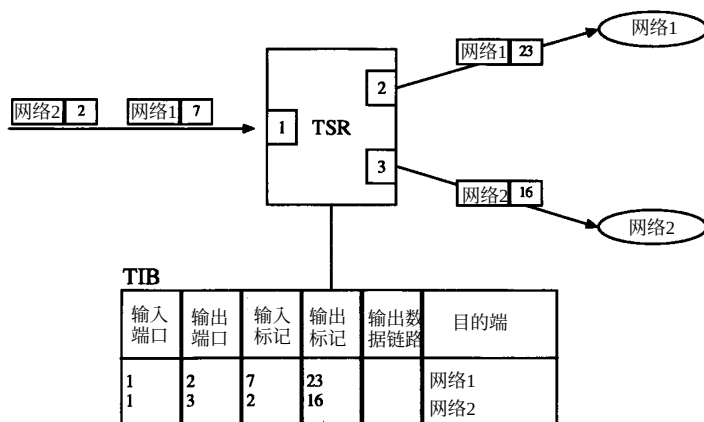


图11-3 标记交换机的转发

- 标记交换。把网络层的信息与标记关联在一起并使用标签（标记）替换机制进行分组转发的体系结构、协议和过程。

- TSR。TSR是转发设备，它运行标准的单播和组播选路协议，可能具有第三层转发分组的能力和硬件或者软件驱动的标签替换转发机制。TSR能够在相邻或者互联的TSR设备之间分发标记关联。TSR实现的实例包括传统路由器、ATM交换机、以及混合式交换机/路由器。

- 标记边缘路由器（Tag Edge Routers—TER）。TER设置在基于TSR的网络的入口和出口。它能够在输入或者输出端口上分配标记、执行第三层的转发、运行标准的单播和组播选路协议。TER（或者TER功能）在入口把未打标记的分组打上标记，并在网络出口删除分组的标记。

- 标记。标记是分组包含的一个短的固定长度的分组头字段。标记的例子可以是ATM信元的VPI和/或VCI、帧中继PDU的DLCI头或者分组中第二层和第三层寻址信息之间插入的“薄垫片”标志。图 11-4所示为薄垫片标志的例子，它的总长度为32比特，其中前20比特分配给标记，3比特分配给COS，1比特作为堆栈标志，8比特作为TTL。图11-4所示为在通用PPP或者以太帧中设置薄垫片标记头。标记交换要求每个分组包含标记，因此如果介质封装信息不能提供一个类似ATM和帧中继的“专门的标记”，薄垫片标记能作为TSR设备网络中进行分组标签交换的依据^①。

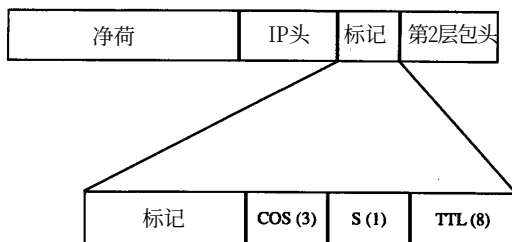


图11-4 标记头

TIB。TIB是连接或者标签替换表，由标记交换设备创建和维护。指示网络中分组的转发是依靠TIB而不是路由表（FIB）。图11-3所示为TIB的例子。一旦接收到TDP或者其他控制协

^① 在分组中插入附加的字段将增加分组的长度。由于不希望分段，TER设备应该执行路径MTU发现，以便判断在不分段的前提下路径上允许的最大MTU的大小。

议中的标记，TIB的表项将被更新。

- TDP。TDP是基于对等实体的控制协议，TSR设备用它分发标记关联。在标记交换网络中TDP消息的交换只发生在对等实体或者相邻节点之间。

- 标记堆栈。这是标记交换引入的有趣的功能之一，大致相当于对IP分组又一次应用IP封装，使得一个分组能够携带多个标记。通过在已标记的分组（层次1）之上“压入”一个新标记（层次2），网络中分组的转发依据层次2的标记，然后“弹出”层次2的标记，继续进行依据层次1标记的标记交换操作。图11-5所示为一个例子。分组到达TSR1，TSR1在已有的层次1的标记（X）之上压入一个层次2的标记（A）。分组在网络中转发到TSR4，在这里层次2的标记（C）被弹出，分组按照层次1的标记（X）转发。标记堆栈使得标记交换适用于分层选路。例如，层次1的标记可以与域间BGP选路相联系，而层次2的标记可以支持域内OSPF选路。

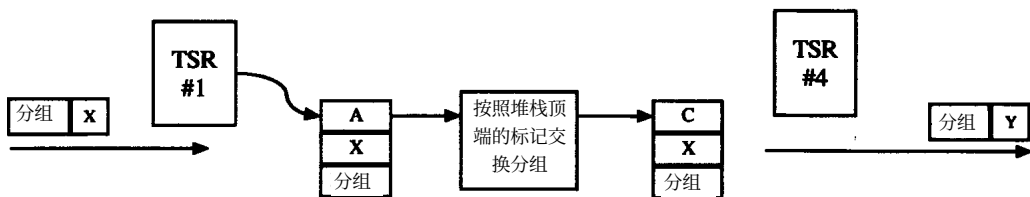


图11-5 标记堆栈

- 转发等价类（Forwarding Equivalence Class, FEC）。FEC是具有一组共同特性的分组的流，他们以相似的方式在网络中转发。为了标记交换的目的，一个FEC映射到沿路由通路的一组标签。例如，所有具有相同目的地址前缀的分组可看作一个FEC，并且能关联到标记关联中的一个标签。FEC也可以细化为主机到主机的数据流。

- 标记交换通路。它是通过把FEC与一组标记联系起来形成的，跨越了一系列TSR设备，是入口到出口的交换通路。

- 捎带。标记交换允许用现有的类似RSVP和PIM的控制协议进行标记分发。这种分发标记的方式称为捎带。

11.3 标记分配方法

标记交换操作的一个重要组成部分是标记和标记关联的分配和分发。利用TDP或者在现有的控制协议上捎带进行标记的分发。目前定义了三种如何分配标记以及由哪个TSR设备分配标记的方法：下游分配、下游按需分配、上游分配。上下游分配的基本区别在于上游分配方法是由将要应用标签（标记分组）的TSR设备进行标记关联，而下游分配方法是由要接收带标记的分组并进行TIB查找的TSR设备进行标记关联。值得注意的是下面描述的标记分配机制与分发标记关联采用TDP还是捎带无关。

11.3.1 下游分配

在下游标记分配机制中，标记由下游TSR设备分配，并传送到上游TSR邻机。下游标记分配机制（依据路由表项）包含下列基本步骤，如图11-6所示。

对路由表中的每个表项，下游TSR设备分配一个标记并更新TIB的输入表项，标记关联〈地址前缀，标记〉发送到上游TSR设备。

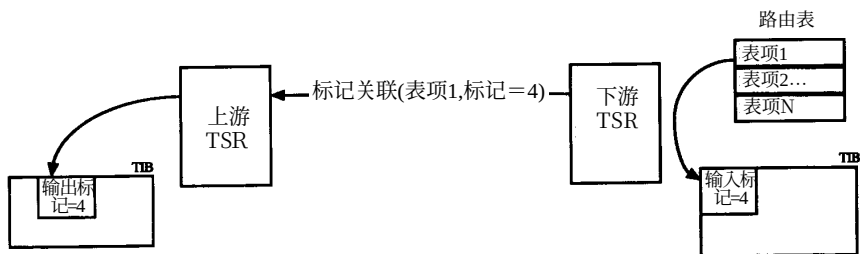


图11-6 下游标记分配

当上游 TSR 设备接收到标记关联并判断出该信息来自通往目的地的下一跳，它把该标记以及输出链路层的信息放入自己的 TIB 的输出标记表项。

11.3.2 下游按需分配

在下游按需分配机制中，用与下游分配机制相同的方法将标记由下游 TSR 设备分配并传送到上游。但是下游 TSR 设备仅当上游 TSR 设备有某个请求时才分配标记。这种方法非常适合包含 ATM 交换机组件的 TSR 设备，因为 ATM 交换机通常只支持有限的标记集合（VPI/VCI 标签）。在 ATM 环境中标记交换的实际实现很可能同时支持标记交换和直接的 ATM 业务。这意味着整个 VPI/VCI 标签的范围分为两个部分，进一步减少了标记交换能够使用的标记的数量。因此仅当需要时才分配标记比较合理。另外下游标记分配机制与发起选路协议的更新消息（实际目的网络）的方向一致，与类似 RSVP 和 PIM 等接收端驱动的控制协议的控制流一致。

下游按需标记分配机制包括下面的基本步骤，如图 11-7 所示。

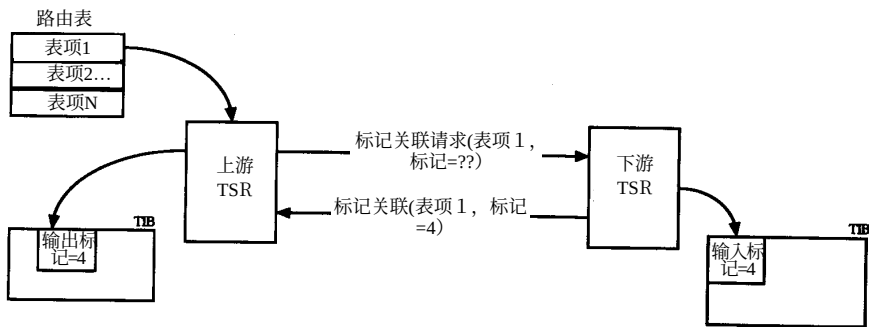


图11-7 下游按需标记分配

- 对路由表中的每一个表项，上游 TSR 设备产生一个标记关联的请求并发送到去往目的地的下一跳。
- 当去往目的地的下一跳的下游 TSR 设备收到请求后，分配一个标记并更新自己的 TIB 的输入表项，产生标记关联〈地址前缀，标记〉并向上游 TSR 设备发送。
- 当上游 TSR 设备收到标记关联，它把标记以及输出链路层信息设置到自己的 TIB 的输出标记表项。

11.3.3 上游分配

- 在上游标记分配机制中，上游 TSR 设备分配标记并通过点到点链路传送到下游 TSR 设备。

上游分配仅限于点到点配置，因为对于某个标记关联，上游 TSR设备所维护的输出 TIB的表项必须在给定输出端口上保持唯一。

上游标记分配机制包含以下基本步骤：

- 对于路由表中包含通过点到点链路可达的下一跳地址的每个表项，上游 TSR首先分配一个标记，接着把分配的标记以及某个每个接口的链路层信息设置在输出标记表项字段，以便更新输出 TIB表项，然后产生标记关联〈地址前缀，标记〉。
- 上游TSR把标记关联发送到下游TSR设备，它位于通往目的地的下一跳。
- 下游TSR设备收到标记关联，替换 TIB输入表项对应于目的网络的标记。

11.4 标记分布协议

标记分布协议（Tag Distribution Protocol，TDP）是标记交换体系结构采用的几种控制协议之一，用来在相关的 TSR设备之间传送标记关联信息，它是唯一的专门用于这个目的的控制协议。TDP与标准的单播和组播选路协议一起运行在 TSR和TER设备上。它的运行与要求创建和分发标记关联的消息无关。根据请求，TDP能够有效和可靠地在 TSR设备之间分发标记关联。

11.4.1 TDP功能

TDP的基本功能是在相关的 TSR设备之间支持标记关联的分发、请求和释放。TDP运行在对等TSR设备之间建立的TCP连接上。用TCP传输有几个原因。首先，TDP采用增量更新的方式，即只传送新的或者修改的状态。这与 OSPF和BGP所采用的概念和运行方式相似。为了达到这个目的，信息必须按照正确的顺序可靠地传送到目的地。TCP具有这个能力，因此所有的TDP消息都是在TCP连接上传送的。其次，TCP提供的可靠传送机制不再需要在 TDP中完成这个功能，使得TDP的设计和扩展更容易。

准备交换标记关联的两个 TSR设备首先建立彼此之间的 TCP连接。TCP连接是双向的，因此 TDP消息能够在两个方向上传送。每个TDP消息包含一个固定的分组头以及一个或者多个可变长度的协议信息单元（Protocol Information Elements，PIE）。PIE 由一个或者多个 TLV字段构成。图 11-8所示为TDP分组的通用结构和两个PIE的结构。

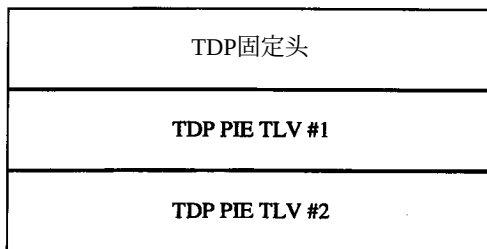


图11-8 TDP分组格式

在TCP连接建立后，通过交换一系列的初始化信息，使得对等 TSR设备处于操作状态。此后，可以开始用TDP交换标记关联。具体来说，标记关联放入 TDP的PIE结构中，与TDP固定的分组头组成帧，在 TCP连接上传送。按照前一节中描述的不同的标记分配机制传送 TDP消息。如果TCP连接中断，则删除标记关联并且回收标记。

假如把ATM当作一种自然的标签交换技术，那么在 ATM网络上进行标记交换操作是可行的。TDP以及更一般的标记交换对 ATM网络提供了以下支持：

- TDP消息在VPI/VCI=0/32的虚通路上传送。
- TDP消息的缺省封装是按照RFC1483定义的LLC/SNAP封装。

- 标记可以包含VPI和/或VCI值。
- TDP允许两个TSR设备决定ATM封装的类型（LLC/SNAP或者VC复用）以及所支持的VCI值的范围^①。

11.4.2 TDP协议信息单元的类型

TDP最初的版本支持以下的PIE类型：

- TDP_PIE_OPEN。这是与对等体建立 TCP连接后TSR发出的第一个PIE。收到该PIE的TSR需要立刻用TDP_PIE_KEEPALIVE或者TDP_PIE_NOTIFICATION进行响应。这个PIE支持的TLV包括是否需要下游按需分配、所支持的VCI范围以及VC封装类型。
- TDP_PIE_BIND。这个PIE从一个TSR发往另一个TSR，分发标记关联。它可能是由事件触发（例如路由表的更新）或者是对TDP_REQUEST_BIND的响应。标记关联包含在TLV结构中，由标记值、地址前缀和长度组成。图11-9所示为TDP_PIE_BIND的格式。

类型(0×200)	长度
请求ID	
AFAM	BLIST_TYPE
BLIST_LENGTH	关联表
关联表	
可选参数	

图11-9 TDP_PIE_BIND格式

- 请求ID。把响应与TDP_BIND_REQUEST对应起来。
- AFAM。指定标记关联中的网络层地址族。有关地址族的细节参见 RFC1700。
- BLIST_TYPE。指定BINDING_LIST中的BLIST表项的格式和语法。表 11-1所示为BLIST_TYPE的值。

表11-1 BLIST_TYPE表

BLIST_TYPE	BLIST表项格式	BLIST_TYPE	BLIST表项格式
0	空	3	22-位组播上游指定(*.G)
1	32-位上游指定	4	32-位组播上游指定(S, G)
2	32-位下游指定	5	32-位上游指定的VCI标记

- BLIST_LENGTH。关联表的长度。
- BINDING_LIST。由一个或者多个BLIST表项构成的可变长度字段，BLIST_TYPE指明了表项的类型。BLIST表项由32位的标记值和相关的单播地址前缀或者组播组地址组成。在单播地址前缀的情况下，要包含地址前缀长度。BLIST类型5和6中定义的跳数（Hop Count－HC）值指示了带有标记的分组在逐级跳的转发中可能通过的路由器的个数^②。图11-10所示为不同BLIST表项的格式。

- TDP_PIE_REQUEST_BIND。查询特定地址前缀的标记关联或者所查询的TSR设备中全部的标记关联。
- TDP_PIE_WITHDRAW_BIND。由分配标记关联的TSR设备发出，通知其他TSR设备可

① 标记交换假定TSR设备的邻机彼此直接相连，或者如果存在中间ATM交换机，二者通过预先配置的ATM虚通道连接。

② 值得注意的是在诸如ATM的一些情形下，不存在用来递减的TTL字段。这种操作将在网络入口和出口进行。

以使用已分配的标记关联。

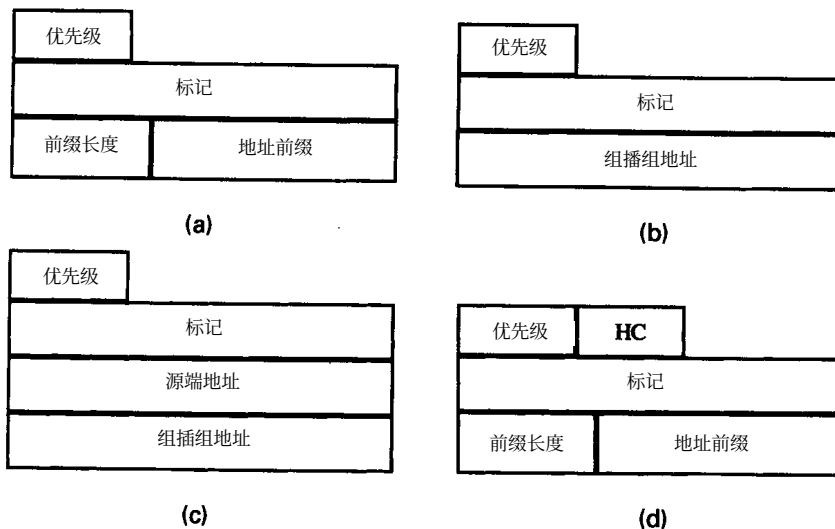


图11-10 BLIST_TYPE格式

a) 类型1和2 b) 类型3 c) 类型4 d) 类型5和6

• TDP_PIE_KEEP_ALIVE。替代实际TDP消息来复位定时器。如果定时器超时，对等体TSR设备将关闭TCP连接。

• TDP_PIE_NOTIFICATION。通知对等体TSR设备“重要”事件。例子包括TSR功能或者操作状态的错误和变化。

• TDP_PIE_RELEASE_BIND。由不再需要某个标记关联的TSR发出，该标记关联来自先前的上游请求/下游指派操作。

11.5 标记交换服务

每个分组中的标记用来检索每个端口或者交换机中的TIB。就整个网络而言，标记只在局部有意义，标记并不说明标记交换的业务流的类型。因此，标记交换最强大的特性之一是：把某个网络层的业务流或者业务与一组标记联系起来，从而利用相同的标签替换转发机制。

11.5.1 基于目的地的选路

对单播业务流的基于目的地的选路可能是标记交换最直接的应用。TSR和TER设备运行标准的单播选路协议，用标准的方式与网络中的对等体交换选路协议的更新消息。在网络拓扑变化以后，路由器搜集并建立反映当前网络拓扑的新的转发信息库（Forwarding Information Base—FIB）。用TDP和前面讨论过的一种标记分配机制，地址前缀映射为标记并且在网络中相关的TSR和TER设备中分发。作为标记交换过程的参与者，入口TER设备将把标记与FIB中的每个表项联系起来。入口TER设备对进入网络的不含标记的分组进行第三层处理，并查询FIB，为分组添加适当的标记。然后分组经过标记交换穿过网络，并且在适当的出口TER设备离开网络。

图11-2所示为基于目的地的标记交换的例子。值得注意的是TDP独立的运行在任意的两个

TSR设备之间。这使得标记分配和分发过程能够以分布式的方式进行。与基于目的地的选路的其他方面相比,这是TDP更普遍的特点,在这里它显得重要是因为标记交换通路建立得越快,分组就能够越早的在这个通路上去往各自的目的地。另外,所有去往特定网络的业务量都被标记并以交换的方式通过网络,而IFMP或者FANP中只有业务流的尾部通过网络交换。

11.5.2 层次结构

Internet中采用的IP选路体系结构本质上是分层次的。自治系统(Autonomous Systems—AS)之间的选路由诸如BGP等外部网关协议(Exterior Gateway Protocol, EGP)处理,而自治系统内部的选路由诸如OSPF等内部网关协议(Interior Gateway Protocol, IGP)支持。然而二者之间没有严格的界线。中间AS的路由器要维护内部(AS内)和外部(所有其它中间AS)的信息。正是这样,中间网络的路由器才能够把分组转发到适当的边界路由器并去往下一个AS。这时将不能访问内部中间路由器所维护的选路信息,并且导致路由表查询缓慢和拓扑变化以后汇聚缓慢。

标记交换提供了一种把内部和外部路由信息分离的技术,使得中间AS的路由器只维护其自身AS的路由信息。运行诸如BGP等EGP的边缘路由器仍然维护与外部拓扑有关的路由信息。标记交换使用标记堆栈来提供内部和外部拓扑之间明确的分界。根据BGP边缘路由器所维护的路由表项,创建并分配一组标记。根据运行诸如OSPF的IGP的内部路由器所维护的路由表项,创建并分配另一组标记。中间AS中的分组携带由EGP和IGP相关标记组成的标记堆栈。

图11-11所示为把分层次选路应用于标记交换(具体来说是标记堆栈)的示意图。值得注意的是在进入中间AS之前,分组包含了单一的与BGP管理的通路相联系的标记。为了穿越中间AS到达下一跳,BGP边缘路由器要使用IGP,具体来说是IGP计算出的一组标记。因此,入口TSR将在BGP标记之上放置一个IGP特定的标记并且在网络中用标准的标记转发机制转发分组。BGP标记不被替换或者操作,因为它不是处于堆栈的顶端,而IGP标记处于堆栈的顶端。当分组到达出口TSR时,IGP标记被去掉,使用BGP标记查询TIB并进行后续的分组转发。

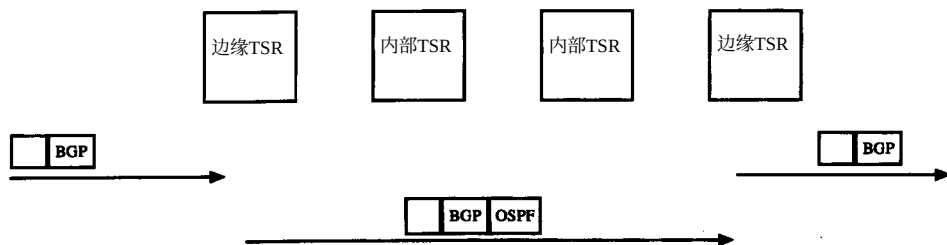


图11-11 选路层次和标记交换

11.5.3 ATM

标记交换和ATM本质上是兼容的,有这样几个原因。首先,ATM网络使用标签替换转发机制在网络中转发信元。其次,每个信元中包含的VPI和/或VCI值可以作为标记(或者标记堆栈)来提供一个或者两个层次的标记,网络层信息与这些标记联系在一起。与标记交换设计的目标相一致,ATM提供了高带宽的传输,独立于网络层业务流(FEC),FEC与标记并且进而与标记交换式通路相联系。

为了支持标记交换，ATM交换机必须首先运行一个诸如 OSPF 的选路协议的实例并且与对等体 TSR/TER 设备和路由器交换路由表更新信息。它还必须运行一个控制组件，能够分配和分发基于网络层信息的标记，这意味着它必须支持 TDP 或者能够携带标记的控制协议。如果 ATM-TSR 设备提供入口和出口功能，它就应该能够转发 IP 分组。实际上，用 TDP 控制组件实现了这些附加的路由器特有的功能以后，普通的 ATM 交换机就转化成 TSR 设备了。同时，转化后的 ATM 交换机就不必使用甚至不再需要 ATM 论坛的协议了。以 ATM 交换作为标记交换机制是相对直接的。标记关联由地址前缀和新的 VPI/VCI 值构成，由 TDP 创建并分发。为了避免不必要的 VPI 和/或 VCI 标签的消耗，建议使用的标记分配方法是下游按需分配。正因为这样，除非该 ATM-TSR 设备是上游 ATM-TSR 设备通往目的网络的下一跳，否则在 ATM-TSR 设备的入口端口不分配 VPI/VCI 标签。标准的第三层处理和必要的拆装功能分别在入口和出口 ATM-TSR 设备中执行。

一般来说，标记交换的规模随着路由表中表项数量的增加而增长。换句话说，标记交换机在 TIB 中所维护的标记的数量等于 TSR 路由表中表项的数量^①。然而，由于 ATM 使用标记转发机制，可能存在一种情形，TSR 需要为某个路由表项维护多个标记，这种情形出现在多个上游 ATM-TSR 设备使用相同的下一跳 ATM-TSR 设备去往某个目的网络时。如图 11-12 所示，两个入口 ATM-TSR 设备共享相同的下一跳 ATM-TSR 设备去往目的网络（网络 X）。为了避免不同源端（ATM-TSR 1 和 ATM-TSR 2）的信元在同一个下游 VC 上穿插起来，必须如图所示在 ATM-TSR TIB 中建立独立的 VCI 值。这使得不同源端的信元在到达出口 ATM-TSR 时具有不同的 VCI 值。出口 ATM-TSR 就能够很容易的把信元装配成分组并进行标准的第三层处理。

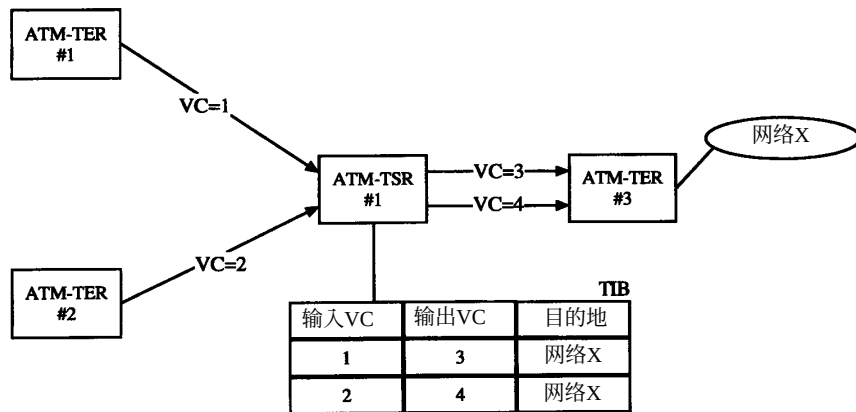


图11-12 标记交换和ATM TIB实例

信元穿插问题有几个解决方法^②。首先可以使用 VPI 值在网络中转发信元，在出口 ATM-TSR 用 VCI 值区别源端。然而，由于在出口 ATM-TSR 所有的 VCI 标签（代表多个不同的源端）可能具有同一 VPI，需要使用一些方法避免可能出现的 VCI 冲突。一种避免冲突的方法是为每个入口 ATM-TSR 配置不同的 VCI 值的范围，但是这种避免冲突方法的扩展性受到 VCI 空间大

① 如果一个标记表示一组路由表项，TIB 中将只需要较少的标记。

② 第12章将讨论称作 VC 汇聚的另外一种技术。

小的限制。

在经典的 ATM 上的 IP 方法中,大量的路由器连接到 ATM 网络的边界,而标记交换控制和转发平面是在 ATM 交换机构上实现的,与前者相比后者具有几个可能的优点。首先,与多个对等体路由器建立联系的负担减小了,因为就像标准路由器一样,ATM-TSR 只与邻机交换选路协议的更新消息。其次,不再需要标准的 ATM 论坛的协议,直通路径(标记交换通路)可以在入口和出口设备之间利用标记交换定义的控制过程建立。当在实际网络拓扑下选择通路时,还可以采用动态 IP 选路。另外,不必对数据流进行分类。当所有的数据流进入网络时,去往特定目的地的直通路径已经建立起来。最后,由于标记交换提供了控制和转发组件的分离,使得多种不同的网络层服务(如基于目的端的选路、层次结构、显式选路等等)联系起来成为可能,因此能够最大限度地使用 ATM 提供的大容量和高性能。

关于在 ATM 网络上进行标记交换操作,还有几点值得指出。首先,拓扑的变化可能造成错误的选路信息的分发,从而可能形成环路的标记交换通路。由于 ATM 信元头中不包含 TTL 字段,这类环路将浪费链路和 TSR 资源。可以从另一个角度来看待这种情况,也就是转发的环路是在第二层形成的,而消除环路(例如 TTL 递减)是在第三层。为了解决这个问题,或者在形成环路的第一个点进行消除,或者把环路中的业务流的破坏作用通过一些机制进行限制,例如按照 VC 排队和 EPD,使得拓扑的汇聚性或者非环路业务流不会受影响。其次,与 ATM 上的 IP 解决方案所支持的叠加模型不同,标记交换仍然依赖于动态 IP 选路来在网络中选择通路。即使转发通路由 ATM 传输构成,通路的选择仍然是基于现有 IP 选路协议的度量值的集合,例如跳数或者最低链路成本。在计算通路时,不考虑可用容量、给定的负荷或者通路时延。然而为了改善性能或者平衡负载,通过把一组标记与网络中的非缺省通路联系起来,可以把业务流切换到指定通路。

11.5.4 组播

除了单播 IP 业务流,标记交换还可以支持组播 IP 业务流。具体来说,标记可以与组播传递树联系。当上游 TSR 设备收到一个分组时,根据标记查找到一个或者多个 TIB 表项,这些表项对应传递树的下游分支。进行标记替换后,分组继续沿传递树传送。如果 TIB 中没有匹配的表项,则丢弃分组。利用标记交换进行组播转发的优点是速度和简单性。传统的组播转发涉及两次查表:上游发送端方向和下游接收端方向。而标记交换的组播转发与单播业务流相同,只涉及在 TIB 中进行直接的匹配查找和标记值的简单替换。反向通路转发(RPFF)检查大大简化。如果 TIB 中没有匹配的标记值,则丢弃分组。

有些情况下,组播传递树的一个或者多个分支可能具有诸如 LAN 的固有的组播能力。例如,一个以太网段连接着在传递树中一个上游和多个下游路由器。如果在这种环境下采用标记交换,上游 TSR 设备可以采用相同的标记,在支持组播的 LAN 网段向多个下游 TSR 设备转发组播分组。

为了支持这种环境,连接到同一个 LAN 的 TSR 设备必须首先决定标记分配方法。换句话说,所有的 TSR 设备都可能是传递树的一部分,但是只有部分 TSR 设备属于传递树的特定组,表示传递树中特定组的标记由谁分配?解决方法是为 LAN 上的所有 TSR 设备分配一个互不重叠的完整的标记空间。从而每个 TSR 设备只负责分配它所管理的特定范围的标记。一旦把标记空间分块,每个 TSR 设备就可以在自己的标记空间内分配标记、把标记与组播组联系起来,

然后把标记关联分发给连接到 LAN 的所有其他 TSR 设备。

可以采用上游或者下游分配的方法分发表示组播组的标记。上游 TSR 设备选择标记并在 LAN 上组播到下游邻机可能是最简单的解决方案，但是必须考虑几个问题。一则上游 TSR 可能有多个自己的上游组播源端，从而需要分配更多的标记。二则拓扑的变化可能产生了新的上游 TSR 并需要标记的重新关联。最后，组播组成员加入组播组的方向是逆流而上，而这种方法却是顺流而下，二者不一致。因此更倾向的方法是下游 TSR 设备完成标记关联和分发功能。这与组播组的成员的信息传送的方向一致，并且使标记的分配和分发与单播选路一致，还使得 PIM 组播选路消息能够捎带标记信息，因为这些标记是逆流而上传送的。如果拓扑变化产生了新的上游 TSR 设备，也不再需要重新设定标记关联。

标记空间的划分可以通过 PIM 的 Hello 消息的扩展项实现。连接到 LAN 上的 TSR 设备在 LAN 上组播 PIM 的 Hello 消息，其中包含声明要求分配的标记的范围。如果与另一个 TSR 设备冲突，具有较高 IP 地址者获胜，而失败者必须重新分配。PIM 的 Hello 消息还能够向 TSR 设备通知在相同的 LAN 上连接着不能支持标记的路由器。

通过在下游 TSR 设备产生的 PIM_JOIN 消息中设置标记关联完成标记分发。如图 11-13，TSR 1 向 LAN 上的所有其他 TSR 设备组播 PIM_JOIN 消息，其中包含关于组播组的标记关联。上游 TSR 设备（TSR 3）收到该标记关联，在 TIB 中为某个组播组设置输出标记项。LAN 上的其他 TSR 设备（TSR 2）可能有属于传递树的下游链路，所有它在 TIB 中为特定组播组设置输入标记项。值得注意的是，这种情况下唯一的下游 TSR 设备负责分配标记，把标记与组播组关联，然后用现有控制协议（PIM）分发标记。

应该指出，支持标记交换的组播的实现需要使用 PIM 稀疏模式（Sparse Mode—SM）的组播选路协议。之所以不使用 TDP 分发标记关联是因为使用了 TCP 面向连接的传送。如果标记交换域因为各种原因没有运行 PIM—SM，可以使用更简单的 PIM 密集模式（Dense Mode—DM）支持这个能力。这将要求下游 PIM—DM 路由器产生 PIM_JOIN 消息，偏离（修改）其利用泛洪和删除的方法进行组播的本质。

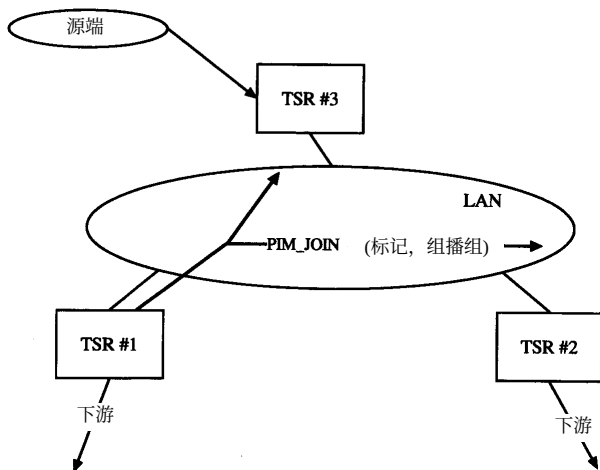


图11-13 PIM_JOIN标记分发

11.5.5 服务等级

除了提高对单播和组播业务流的 IP 选路的性能，标记交换能够支持差分服务。可以采用多种技术达到这个目的。一种方法是在标记中包含服务等级（COS）标识符，TSR 设备以此调度发送出的分组。第二种方法是建立多个标记并扩展为多个标记交换通路，分别对应不同的服务等级。两种方法都保持了标准的尽力而为服务的简单的标记转发机制不变。支持 COS 的标记交换与 DiffServ（在第 3 章讨论过）的概念是一致的，把多个数据流映射到几个服务等

级中，网络按照数据流的显式的或者隐含的等级处理其分组。

为了在标记交换体系结构中支持 COS，首先需要对分组进行分类，以确定其应该接受的服务，这是依据各自的决策规则在源端主机到入口 TER 之间的网络的各点进行（包括二者在内）。分类以后，以某种方式对分组进行标识，指示网络需要为它们提供的服务级别。同样这是发生在源端主机到入口 TER 之间的网络的各点（包括二者在内）。然后分组将进入标记交换域，在标记交换通路上传送，提供预期的基于等级的差分服务。

DiffServ 模型提供了一个非常粗糙的服务，其中仅仅把所有数据流映射到几个等级。如果需要更精细的业务划分，甚至对每个数据流分别进行服务，标记交换可以采用 RSVP 功能在标记交换机内部建立和预留资源。为了实现这一点，在每个 TSR 设备中为每个会话分配一个标记，并当 RSVP 消息到达时为其设置 QoS 状态。分发 RSVP 流的标记信息需要在 RSVP 的 PATH 和 RESV 消息中引入标记对象。RSVP 作为控制协议，负责在 TSR 设备之间传递标记信息，这是用捎带方法进行标记分发的又一个例子。图 11-14 所示为一个简单例子。值得注意的是 RSVP 的 PATH 和 RESV 消息以标准的方式经过每一跳（TSR 设备）的处理。这个例子中，某个数据流 N 的 RSVP 的 RESV 消息向上游发送端（图中未画）传送标记对象。这是标记分发中下游分配技术的一个例子。当 RESV 消息到达每个 TSR，其内容传递给 TSR 设备的流量控制组件。TSR 设备检测能否支持数据流的带宽和 QoS 特性，然后假定检测通过，设置适当的资源以支持下游方向的数据流。同时，在 TIB 的输出项为这个数据流设定标记对象。简单的标记查表机制代替了每个 TSR 中对每个分组分类的多个字段的处理开销，这是 RSVP 和标记交换的一个重要优点。另一个应用标记交换的优点是无论粗糙或者细致划分服务，都可以采用相同的转发机制和相同的基础设施。

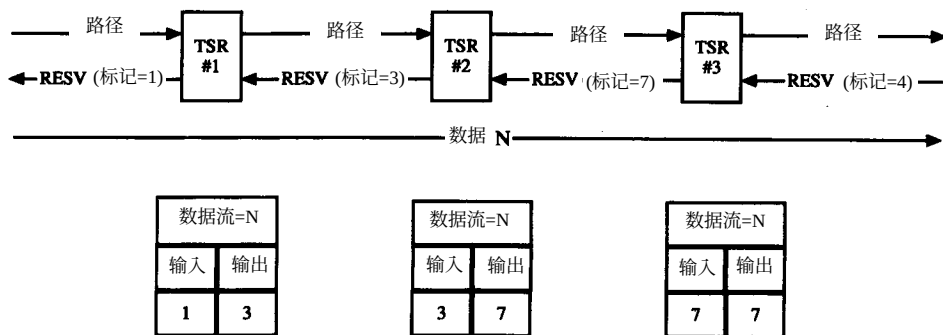


图11-14 RSVP和标记交换

11.5.6 显式选路

标记交换另一个潜在的有用的功能是支持显式选路。显式选路可以认为是预先配置的跳的顺序，用来转发某个子网的业务流。实际上它是非缺省的通路，替代了动态 IP 选路计算的基于目的地的通路。有几个原因使得显式选路非常有用。网络运营者可能希望通过把业务流在多个通路上转发来更好的平衡骨干网的负荷。特殊的服务（例如 RSVP）可能要求区别于标准缺省路由通路的专用通路。用户定义的规则（例如安全性、计费、VPN 连接）可能要求某些业务流通过某个链路和节点。加上其他未讨论到的原因，对于显式选路的概念以及更一般

在网络中主动和有效的配置业务量的功能，需求十分迫切。

为了支持显式选路，必须在某个通路沿途的 TSR 设备中设置不同于标准缺省选路的标记关联。还必须配置分类器，并将其设置在显式通路的入口，识别将在显式通路上传送的子网的业务流。这有几种实现方法。第一种方法，在每个 TSR 设备中人工配置标记关联。第二种方法，“引导” TDP 消息通过一系列 TSR 设备，形成显式选路。尽管在技术上是可行的，但是这种方法在当前 TDP 处理的范围内涉及某些局部的修改，因此只限于相邻的 TSR 设备^①。第三种方法是，修改 RSVP 使其基本上沿着跳的顺序形成入口到出口的显式选路。

图11-15所示为使用 RSVP 在标记交换网络建立显式选路。这个例子中入口 TER 设备 A 发起 RSVP 的 PATH 消息，其中包含显式通路（Explicit Route—ER）对象。ER 对象指示 PATH 消息通过一系列 TSR 设备。RESV 消息在相同的通路上沿着相反的方向传送，设置标记，从而为去往网络 X 的业务流创建显式选路。值得注意的是这种情形下源端和目的端分别是入口和出口的 TER 设备，而不是 RSVP 通常情形下的源端和目的端的主机。实际上尽管 RSVP 仍然可能设置 QoS 状态，这时其基本作用是沿着显式通路传送标记关联。把通过其他机制人工配置或者安装的几种现有的 RSVP 对象（例如发端模板）组合起来实现这一点。还应该注意没有明确定义的动态选择形成显式通路的跳的顺序的机制。这些信息仍然必须根据网络规程人工配置或者计算。RSVP 只提供了建立交换通路的机制，而没有定义交换通路由哪些链路和节点组成。

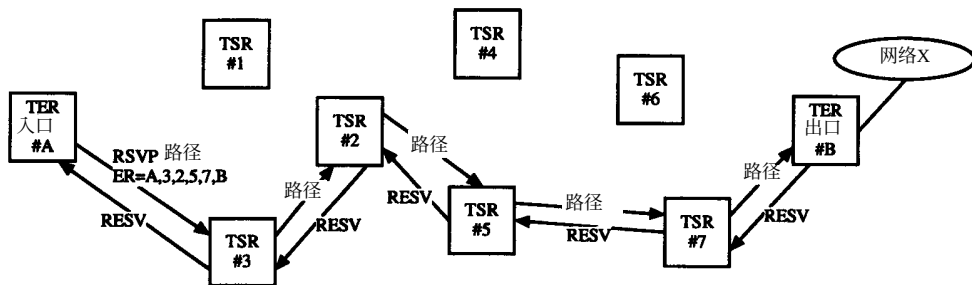


图11-15 利用RSVP建立显示通路

11.6 标记交换网络实例

对于给定的选路和拓扑环境，标记交换非常适用于大规模 IP 网络，例如企业内部网或者 ISP 网络。在现有路由器中安装软件，或者在多业务 ATM 网络中建立 VC 交换通路时，作为控制组件选项，可以实现标记交换。图 11-16 为一个基本的标记交换网络的例子。骨干网由几个骨干 TSR 设备构成，周围有多个 TER 设备，这与大型交换网络的概念（帧中继或者 ATM）没有什么不同，都是作为大量边缘路由器的骨干网。与标记交换所不同的是图中的所有组件都参与计算和分发路由表信息。实际上，标记交换是对等模型的一个例子，所以只需要单一的地址和选路协议。附加组件 TDP 把通路映射为标记，然后在 TSR 设备之间分发标记。在性能和显式选路控制方面它具有优势，无需帧中继或者 ATM 就可以提供交换虚连接。

^① 因为显式通路是从入口到出口，所以从网络边缘开始建立显式通路可能更简单。而 TDP 不是从网络边缘开始运行的，它是在 TSR 设备对之间分布式和自治的运行的，因此它不十分适合这种特定的解决方案。

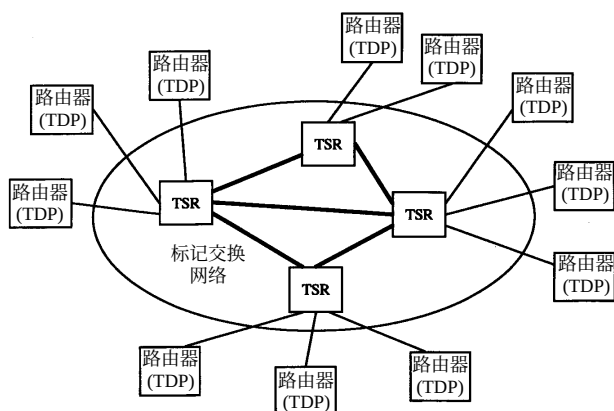


图11-16 标记交换网络实例

更多的信息

Doolan et al., Switching in IP Networks, Morgan Kaufmann, 1998. <http://www.cisco.com>.

Rekhter et al., "Tag Switching Architecture Overview," IEEE Proceedings, vol. 85, no. 12, Dec. 1997.

第12章 ARIS

在Ipsilon的IP交换以及东芝公司的CSR/FANP模型中，IP数据流与虚连接存在紧密的联系，而Cisco引入的标记交换将这种紧密的联系分离开。标记交换不是将由主机到主机或者应用到应用通信的单个IP数据流映射到实质上的ATM虚连接，而是极大地拓宽了交换通路上的IP业务流范围。实际上所有去往指定目的端的业务流都放在交换通路上，标记关联来源于路由表项，而交换通路的建立是通过创建和分发标记关联来完成。这一拓扑驱动的IP交换提高了选路系统的扩展性，使得所有的业务流都能够得到标签替换转发机制所提供的快速服务。另外，其他网络服务，例如与基于目的端选路没有特定关系的非缺省的显式选路，也可以被映射为标记，从而某个业务流就可以不降低性能而经过不同的路径。

1996年秋出现了标记交换以后，IBM公司很快宣布了一个类似的拓扑驱动的IP交换解决方案，它有一个容易记的名字ARIS。ARIS表示汇聚的基于通路的IP交换（Aggregate Route-Based IP Switching），而其基本功能是按照汇聚的IP数据流的目的端建立并且维护交换通路。ARIS利用了选路协议的信息，并且把这些信息与网络入口到出口的交换通路联系起来，从而实现上述功能。交换通路可以采用多点到点的树的形式，以出口路由器为根并且向所有的入口路由器分叉。ARIS采用的拓扑驱动的方法实际上与标记交换提供的功能非常相似。然而二者有一些重要的区别，当前二者在IP交换领域产生了有趣的（而且是激烈的）讨论，但是从长远来看，这种讨论将提高二者的性能并且有利于对拓扑驱动的IP交换的全面理解。

具体来说，ARIS是运行在集成的交换机/路由器（Integrated Switch/Router，ISR）设备网络上的控制协议。ISR是支持IP选路的第二层交换机（例如ATM）。ARIS利用选路协议中的信息（例如OSPF、BGP）把IP分组映射到第二层交换通路，通过ISR设备的网络。ARIS是第一种引入汇聚的交换通路的IP交换机制，在多个不同源端和同一目的端之间可以建立多点到点的交换通路（树），因此降低了对交换机资源的消耗。出口ISR表示目的端，它与一个或者多个目的端地址前缀相联系。共享同一出口ISR的业务流在多点到点树的分叉上传送并在出口ISR离去。因此ARIS网络能够把所有业务流交换到任意数量的出口ISR，而仅仅使用 $O(N)$ 个交换通路（多点到点树）。这种独特的扩展性使得ARIS网络能够包含更多的路由器，并且解放出交换机资源以支持附加服务，例如基于QoS的连接。

ARIS操作和理解的核心是出口标识符（出口ID^①）。出口ID是一个可配置的值，用来标识在某个多点到点树上传送到出口ISR的IP业务流。出口ID配置在出口ISR，可以用现有选路协议的信息动态创建或者人工配置。出口ID的值从单个IP主机地址一直分布到组的无类型的域间选路（Classless Inter-Domain Routing，CIDR）前缀，甚至BGP的NEXT_HOP。在出口ISR设置出口ID将引发从出口ISR建立交换通路。ARIS从网络的边缘（出口和入口ISR）开始建立交换通路，而不是像标记交换和IFMP那样局部的建立。另外ARIS协议提供了避免环路的机制，可以只建立无环路的交换通路。这一点对于在ATM网络上运行是特别重要的，因为信元中不

① 本质上出口ID就是标记交换中的FEC，但是标记交换并没有明确如何以及在何处定义或者使用FEC。出口ID明确了这个问题，指出在出口ISR配置，并且为去往出口ISR指定值的所有业务流建立唯一的交换通路。

包含环路控制的TTL域^①。

12.1 体系结构

ARIS体系结构建立在对于IP选路系统的行为的简单而有时微妙的观察上。首先，可以预见的是所有穿越中间选路域的业务流都是去往某个目的端的，其经过某个入口路由器进入域并且经过某个出口路由器离开。业务流由多个不同源端创建，并通过数量相似的入口路由器进入域。如果它们都是去往同一出口路由器之后的多个目的端，它们甚至仍然必须通过该单一的出口路由器离开域。因此转发通路中的共同的网络元素未必是中间路由器的路由表所包含的一个或者多个目的端地址前缀，而只是出口路由器本身。如果采用IP交换机制在中间域中加速分组的转发，出口路由器将是配置和发起这个过程的便捷和合理的位置。

另外，穿过中间域的数据流沿着组播树相反的方向传送，该组播树源于出口路由器并且分叉到所有入口路由器。考察图12-1的拓扑。网络A、B以及C可以通过出口路由器4到达。所有去往这些网络的业务流都在不同的入口点进入中间域，而选路协议将为分组选择去往出口路由器的最小跳数的通路。往往一些来自不同上游链路的分组被路由器放置在同一下游链路上传送，这种N比1的上游到下游的复用能力是路由器的基本功能并且使基于路由器的网络易于扩展。如果采用IP交换机制在穿过中间域的时候提供更快的性能，并且要求使它像标准IP选路一样易于扩展，那么在第二层提供这种“汇聚”功能是很吸引人的。

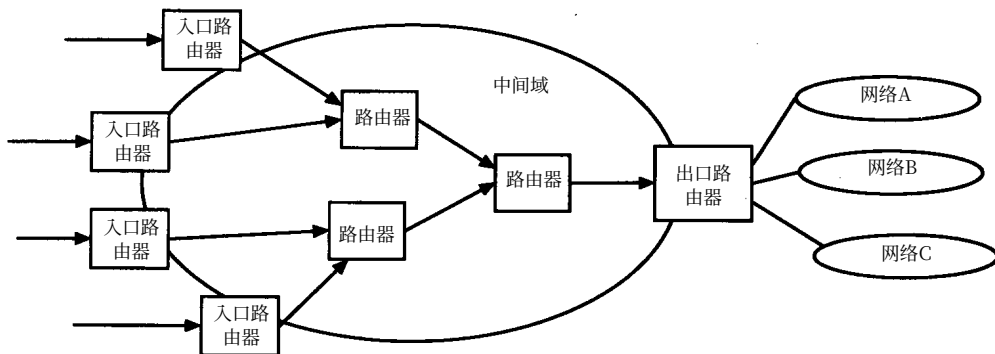


图12-1 中间域出口路由器

从实用的角度看，大规模网络已经并且必将包括交换技术，例如帧中继或者ATM，当然还包括IP路由器。交换具有价格和性能优势，总是尽量采用交换。无可质疑的入口和出口路由器之间最快速的通路是交换通路，而不是经过中间路由器。尽管如此，在大规模网络中，随着源端、接入设备和边缘路由器数量的增长，为了进行任意两点之间的交换，所要求的交换机资源（标签、缓存、存储器、VC端点等）将呈指数增长。每个新连接到具有N个其他路由器的交换式网络上的路由器都必须与其对等路由器建立（N-1）个交换式连接。另外，可能在两个路由器之间还需要有多个按照业务划分的交换通路，这进一步消耗了有限的交换机资源。IP交换解决方案目的是解决规模的扩展性，避免消耗网络的资源，在交换全部或者大部分业务流的同时，尽量减小对交换机资源的需求。节省下来的交换机资源将留给特殊的非缺省服务，例如基于QoS的连接。

最后，在大型ISP骨干网以及载体网中ATM正发挥越来越大的作用。原因很清楚：该传送

^① 回忆纯ATM的选路协议，PNNI阶段I利用源端选路机制来避免产生环路。

技术以适当的价格提供了可管理的带宽。IP交换解决方案的引入省去了传统ATM信令和选路，而根据路由表进行工作。它特别注意减小或者避免IP选路过渡状态的影响。IP选路是非常动态的，并且能够对变化进行快速响应，但是并不是网络中的每个路由器都能够同时接收到有关网络拓扑变化的信息。这可能导致形成暂时的选路环路。每个路由器靠递减TTL来控制循环业务流的破坏作用。如果把基于拓扑的IP交换解决方案引入到ATM交换机机构中，必须避免形成环路。

我们应该始终注意前面提到的这一点，ARSI解决方案的开发者所倾向的IP交换机制中，配置以及交换通路建立过程都是由出口或者边缘设备的发起的。他们还在协议中设计了支持交换通路汇聚的能力，以便节省交换机资源。ARIS还可以看作为在ATM网络上运行而特别设计的IP交换解决方案。ARIS假定所有离开中间域的业务流必然通过一个出口路由器（ISR）。因此引入了出口ID的概念，并使之配置到出口ISR上，作为选路协议信息和交换式通路之间的“纽带”。其思想是为一个出口ISR配置单一的出口ID，用出口ID就可以表示出口ISR背后的一个或者多个目的端（或者甚至只是出口ISR本身），这样更简单也更直接。出口ID可以用来为所有经过出口ISR去往这些目的端的业务流建立交换通路。

图12-2所示为出口ID的概念以及它在ARIS网络中所起作用的示意图。在这个例子中，所有去往目的端1到N的IP业务流都与一个出口ID相联系，出口ISR可以从选路协议中获取出口ID或者人工配置。ARIS使用出口ID在ISR设备的网络中建立和识别交换式通路，其终点是出口ISR。这个交换式通路将传送所有去往目的端1到N的业务流，因此出口ID可以看作业务流的标识符，这些业务流具有相同的一个或者多个目的端（例如路由表项）并且共享第二层交换式通路。ARIS创建交换式通路的最终结果是具有相同的一组目的端的会聚业务流在同一个第二层交换式通路上传送。ARIS为每个出口ID创建一个交换式通路，当然一个ISR可以支持多个出口ID。从实用角度，出口ID是可配置的值，网络运营者可以为交换式通路上传送的业务流选择颗粒度。

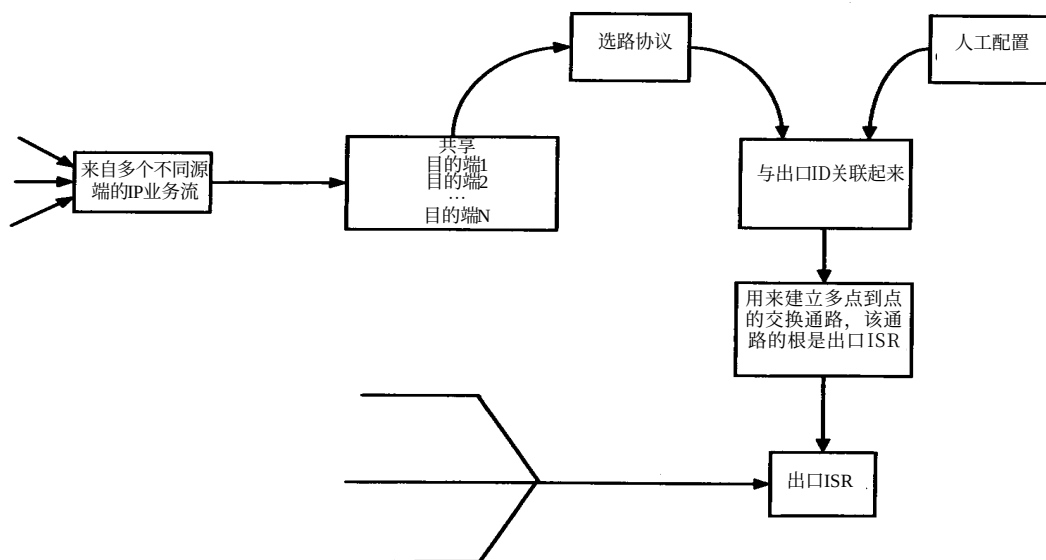


图12-2 ARIS 出口ID的概念

运行在ISR上的ARIS是一个轻负荷的控制协议。ISR设备基本上是ATM交换机，其中运行IP选路协议并且可以在第三层转发分组。ISR类似于IP交换机、CSR或者TSR设备，只是它仍然能够在透明的配置中支持纯ATM交换（通过ATM论坛的UNI/PNNI信令和选路）。实际上ARIS设计能够在多业务基础设施上运行，其中所有尽力而为的IP业务流都进行交换并且能够利用任意数量的不同机制（包括纯ATM和/或ATM上的RSVP）来提供特殊服务。

由出口发起并且具体来说在出口ISR上配置出口ID，使得从出口ISR发起建立交换通路变得很自然^①。换句话说，当网络拓扑变化或者配置了新的出口ID以后需要建立新的交换通路时，ARIS通路建立消息是由出口ISR发出的。它们按照逐级跳的方式转发，从出口ISR最终到达所有可能的入口ISR设备，入口ISR把业务流转发到与出口ID联系在一起的目的地。只有当通路建立消息到达入口ISR设备的时候，完整的入口到出口的交换通路才建立起来。ARIS通路建立消息包含出口ID以及在ISR交换机组件内部设置的标志。基于边缘的方法能够建立具有某种入口到出口特性的交换通路，例如，基于边缘的建立机制容易提供无环路的通路或者显式通路。另外，这种方法与其他的可扩展的IP控制协议保持一致，例如PIM-SM和RSVP。然而，缺点是建立通路存在一定的延迟并且在传递通路建立消息时对“其他”ISR存在依赖。

通过支持交换通路的汇聚，ARIS解决了节省交换机资源的问题。利用交换通路的汇聚，来自多个上游链路的信元汇聚到同一个下游链路。利用对交换式通路汇聚的支持，多点到点树能够以出口ISR为根并且分叉到所有入口ISR设备。因此可能提供任意交换机之间的通路而只使用 $O(N)$ 个交换通路，其中N是网络中ISR设备的数量^②。值得注意的是交换通路的汇聚未必支持基于QoS的业务流，汇聚功能本身将引入额外的时延和时延抖动。更确切的说，交换通路的汇聚建立了多点到点树，以较少的交换机资源的消耗来传输所有尽力而为服务。如果需要支持基于QoS的连接，ARIS允许使用非ARIS的连接建立方法，例如RSVP或者纯ATM信令和选路。

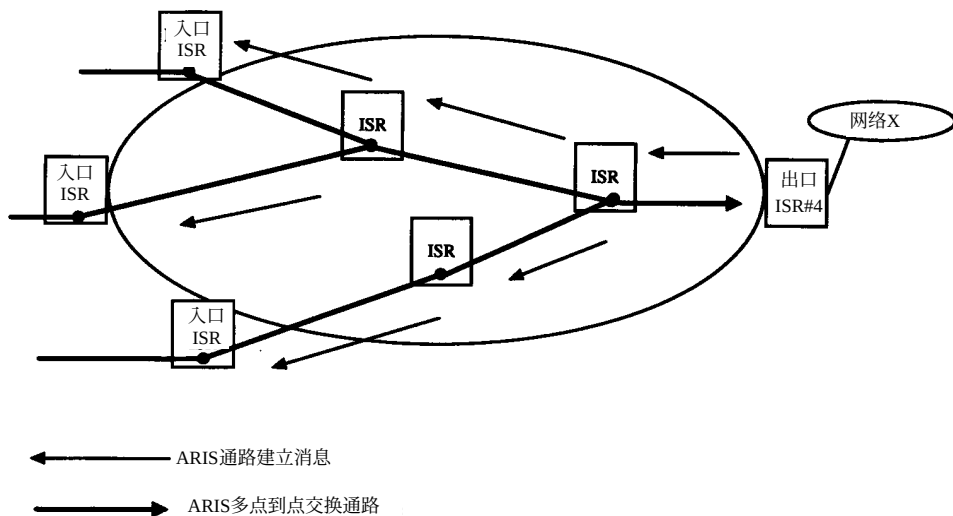


图12-3 ARIS体系结构

① 边缘这个词用来说明ARIS中建立组播通路的消息是由出口ISR发起的，而不是入口ISR。但是在这一章中边缘和出口这两个词将互换使用。

② 这时使用N个多点到点树。

图12-3所示为ARIS提供的基本体系结构和功能，图中为一个运行 ARIS的ISR设备域。某个ISR（ISR 4）被配置为“网络X”的出口ID，因为它是去往网络X的所有业务流的出口ISR。ARIS的通路建立消息以反向通路组播（Reverse Path Multicast-RPM）的方式传送，由出口ISR到达每个转发去往网络X的分组的入口ISR。沿着路由通路到达出口ISR的中间ISR设备将处理通路建立消息，在每个交换机组件中分配标记（与出口ID相联系）。在入口ISR，标准FIB配置了出口ID的参数，以便当分组进入网络时，标准路由表查询能够体现出出口ID的存在。该出口ID与到达出口ISR的交换通路相联系。然后去往网络X的分组在整个交换式通路上进行标签替换，直到出口ISR，无需中间的第三层处理。因为所有业务流都是交换，选路系统将具有更好的整体性能。

12.2 组件

ARIS解决方案由下列组件构成：

- ISR。是运行IP选路协议、ARIS控制协议以及能够在第三层转发分组同时能够在第二层交换分组的第二层交换机（例如ATM）。ISR可以同时在透明配置下支持纯ATM信令和选路（Q.2931/PNNI）并且可以作为MPOA通路服务器。如果ISR支持PNNI，必须在PNNI和ARIS之间划分VPI/VCI标签的范围。图12-4所示为一个ISR的例子。

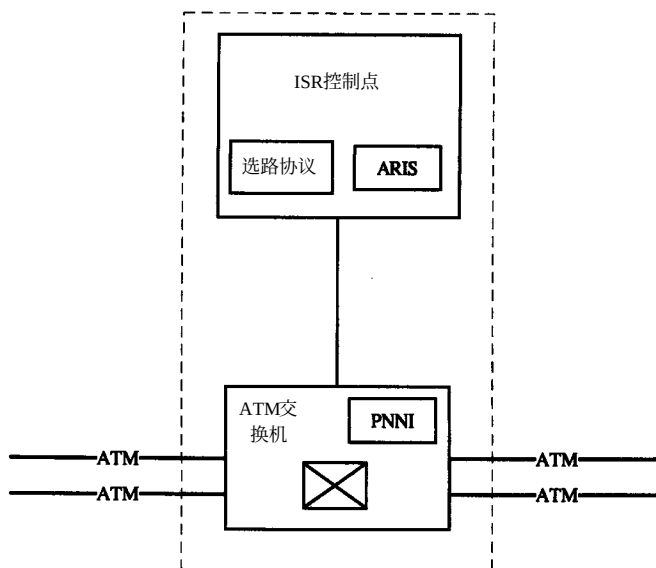


图12-4 ARIS ISR

- ARIS。是能够根据IP选路协议的信息建立第二层交换式通路

的控制协议。在ISR中除了IP选路协议，还要运行ARIS。另一种解释ARIS的方法是，它把许多IP目的端映射为一个出口ID，该出口ID与一组标签联系在一起，形成穿越ISR设备网络的交换通路。

- 出口ISR。是一个把穿越ARIS域的IP业务流输出到直接相连IP子网（例如IP主机的以太网域网）、非ARIS域或者另一个ARIS域的ISR。值得注意的是一个ISR可以同时配置为输入、输出和中间节点。出口ISR发出ARIS通路建立消息。

- 出口标识符。是可配置参数，标识交换通路并把它与一个或者多个“背后的”源端（目的端）联系在一起，这些目的端与一个出口ISR相联系并通过它可达。

- VC汇聚。是第二层交换功能，能够把多个上游VC会聚到同一个下游VC，而不会把不同帧的信元间隔开。支持VC汇聚的交换机必须收集输入VC上一个帧的所有信元，然后在输出链路用新的VCI值以连续的方式发送帧的信元。在新的VC上发送的信元必须按照帧来分组，以便目的端的设备能够区别不同源端的信元。VC汇聚可能需要特殊的硬件，或者在交换机中的每条VC缓存的功能。图12-5所示为运行VC汇聚功能的ISR（ISR 3）的例子。值得注意的是

因为VCI=7的信元是按照帧分组的，出口ISR可以把它们装配成分组并且转发到目的端。VC汇聚的优点是减少VC资源（例如标志）消耗而同时提供第二层交换转发功能。与没有VC汇聚的情形相比，N个上游VC只需要单一的而不是N个下游标签，从而节省VC资源。

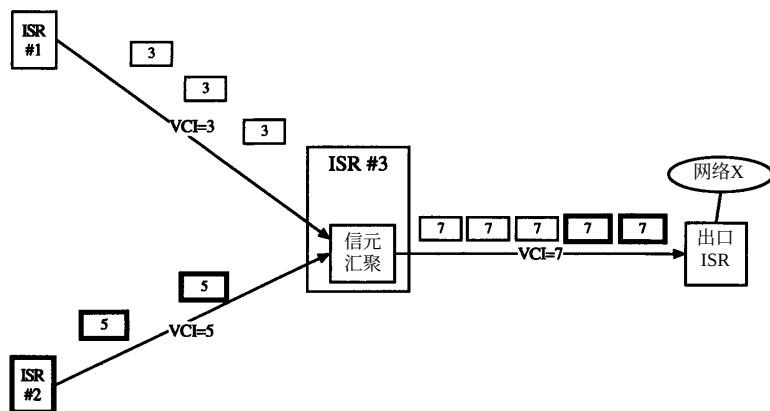


图12-5 VC 汇聚

- VP汇聚。是标准ATM交换功能，使多个上游VP汇聚到同一个下游VP。输入和输出VCI值保持唯一并且不变，因此在交换机中不需要特殊的VC汇聚功能。唯一的VCI值保证了源端（例如入口ISR）的唯一性。然而因为VPI值可能具有相同的VCI值，必须把唯一的VCI值的范围发送到入口ISR设备以避免VCI冲突。

- 多点到点的交换通路。是ARIS建立的交换通路的别名，它以出口ISR为根并且分叉到所有入口ISR设备。业务流沿相反的方向从树的叶子传送到出口ISR的根。一个多点到点树与一个出口ID联系在一起。多点到点树利用VP或者VC汇聚来减少交换机资源的消耗。

- VC信息库（VC Information Base, VCIB）。它是出口ID的表，并且与ISR交换机组件维护的第二层标签信息相联系。图12-6所示为图12-5中ISR 3的VCIB的例子。

出口ID	输入端口	输入VC	输出端口	输出VC
网络X	1	3	3	7
网络X	2	5	3	7

图12-6 VCIB实例

12.3 出口标识符

为了提供最大的灵活性并且对所选择的将通过交换通路的业务流进行控制，ARIS规范支持许多不同的出口标识符。出口ID可以从选路协议的信息中获得，或者人工配置。在交换通路建立阶段，ARIS协议从出口ISR向所有路由通路沿途的ISR设备发送出口ID值。它安装在路由通路沿途的ISR设备的VCIB内（与第二层交换信息一起）。另外，在FIB中为每个下一跳ISR包含其出口ID。

ARIS规范定义了下列的出口标识符：

- IPv4地址前缀。这一标识符用以建立到某个主机或者CIDR前缀的交换通路。这类出口ID建议用在选路协议（例如RIP）不提供汇聚功能或者网络中目的端地址前缀的数量少的环境。如果需要建立单独的到某个子网或者已知主机的交换通路时也会用到。

• BGP Next_Hop。它包括在BGP Next_Hop参数中的值。NEXT_HOP参数是一个众所周知的命令参数，用来标识 BGP路由器的IP地址以通告通路。概念上它表示下一跳地址，可能代表大量的目的端。这个出口标识符可以是连接到 BGP边缘路由器的IP地址或者外部对等体的IP地址^①。两种方法都可以达到相当程度的汇聚，因为所有到 BGP下一跳路由器的业务流将在单一的交换通路上传送。

• OSPF路由器ID。这个标识符包含 ISR的OSPF路由器ID，该ISR发起链路状态的通告。ARIS与OSPF链路状态通告（Link-State Advertisements, LSA）的信息相互作用，建立到出口ISR的交换通路，该出口ISR连接到OSPF域外部的资源^②。第一种情况是，利用静态配置或者其它选路协议（例如RIP、BGP）重新分配，远端网络可以由某个ISR插入OSPF选路域，这些远端网络在路由器中可能由多个无法合并的目的端地址前缀来表示。ARIS把OSPF路由器ID（出口ISR的）作为出口ID，以建立到出口ISR的多点到点树，而不是采用多个出口ID（以及同样数量的交换通路）来覆盖这些远端网络。换句话说，ARIS能够把所有插入OSPF选路域的远端网络抽象为单一的值，即通告ISR的OSPF路由器ID。图12-7所示为这种功能。这使得汇聚的形式简单、消耗交换机资源少，同时为所有通过出口ISR去往远端网络的业务流提供交换通路。

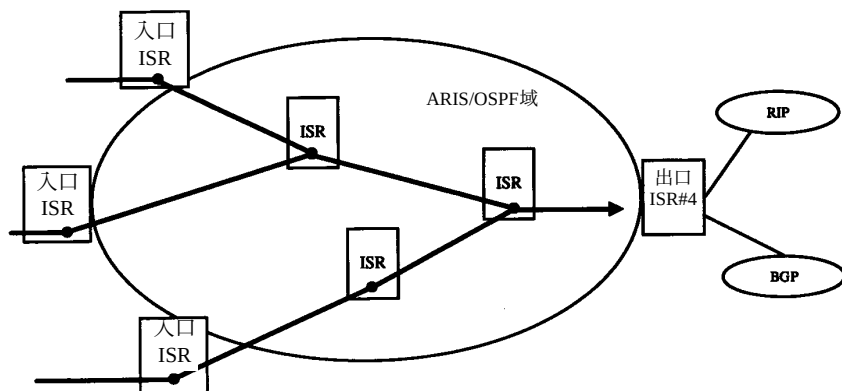


图12-7 把OSPF路由器ID作为出口ID的ARIS。出口ID=出口ISR的OSPF路由器

• OSPF区域边缘路由器ID。使用OSPF路由器ID的例外是当ISR收到一个包含非零转发地址的AS外部链路通告。转发地址指路由器的IP地址，该路由器可以达到OSPF选路域外部的网络。OSPF利用转发地址把业务流从发起通告消息的路由器中旁路。因此当把转发地址作为出口ID时，可以更好的利用网络资源。

• 组播〈源端，组〉。标识了组播传递树中的唯一的源端地址/组播组地址对，该传递树利用泛洪以及诸如DVMRP或者PIM-DM的裁剪协议。这种情况下，交换通路的建立由入口ISR发起，或者换句话说，由最靠近源端的ISR发起。它包括一个点到多点的传送树，以入口ISR为根并且分叉到所有出口ISR设备。为每个〈源端，组〉建立一个交换通路。图12-8所示为一个例子。

• 组播〈*，组〉。这个出口ID用来建立交换通路，根据共享树的诸如PIM-SM的组播选路协议。这种情况下，交换通路的根是ISR汇聚点（Rendezvous Point, RP）。

① 在后一种情况下，ARIS开发者建议采用被比喻为把某人的叶子扔到邻居的院子的方式。

② 更清楚的说是运行ARIS和OSPF的ISR设备的域。

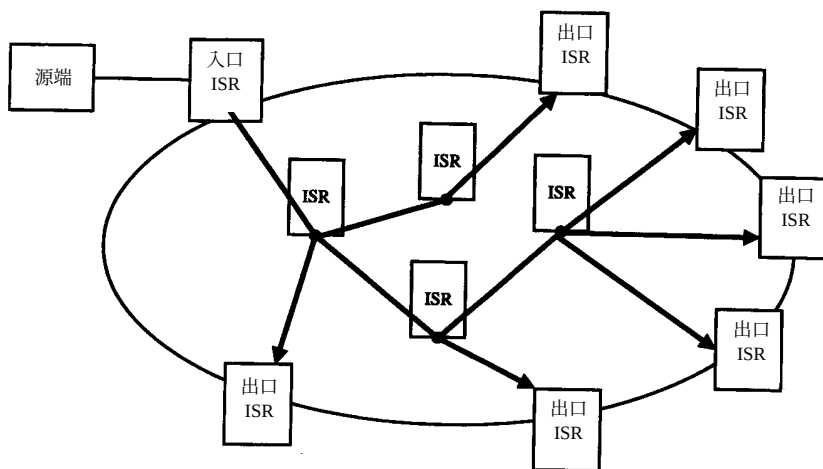


图12-8 ARIS对组播的支持

- 数据流。标识某个源端和目的端的IP地址对，并且用来为某个数据流建立交换通路。
- CIDR组。该出口ID包含CIDR前缀的列表，它们恰巧具有共同的出口ISR。如果需要把多个不同的CIDR前缀汇聚起来，而无法由现有的选路协议完成汇聚，该出口ID十分有用。

12.4 ARIS协议

ARIS是运行在ISR设备网络上的控制协议。正如已经讨论过的，它把IP路由通路映射为交换通路。具体来说，ARIS设置路由通路上的交换通路的状态，通往与某个出口ISR相联系的目的端或者一组目的端的业务流在该通路上传送。ARIS的通路建立消息（ESTABLISH）由出口ISR发起并且以反向通路组播的方式在ISR网络中转发，直到入口ISR。ESTABLISH消息包含若干对象，即出口ID和ESTABLISH消息所通过的ISR设备的列表。后者用来防止形成将带来问题的第二层的交换通路的环路。与TDP和IFMP所不同的是，ARIS通路建立消息必须在交换通路建立之前从出口一直传送到入口。例外ARIS消息需要显式确认以保证可靠传送^①。

12.4.1 消息类型

图12-9所示为ARIS协议消息的结构。通用头中的字段包括版本、消息类型、长度、报头校验、发端路由器ID、发端序列号以及发端和收端的任务号。通用头对ARIS消息进行鉴权并且指示了消息类型。通用头后面是通用对象头，定义了对象类型、对象子类型以及长度。接下来是对象。ARIS消息通常包含多个通用对象头和对象。

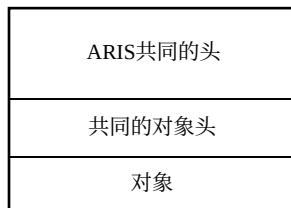


图12-9 ARIS消息结构

ARIS规范定义的五条消息类型是

- INIT。它在ARIS邻机之间交换并用来建立关联关系。一旦成功的收到了INIT消息，

^① ARIS使用自己的确认机制，TDP使用TCP的确认机制，这是二者之间竞争的主要方面之一。TCP去除了TDP的可靠性功能但是却需要一些附加的配置，而协议自定义的确认机制可能更灵活一点。

ARIS邻机将转换为激活状态。只有在激活状态下才能够建立交换通路。

- **KEEPALIVE**。由ISR发出以通知其相邻的ISR设备自己处于激活状态。INIT消息设置了定时器对象，在定时器对象超时期限之内必须传送 KEEPALIVE消息，否则关联将超期。

- **ESTABLISH**。它用来创建交换通路。它由出口 ISR发出或者发送到响应 TRIGGER消息的ISR。它用RPM方式传送，直到所有 ISR设备接收到该消息并建立交换通路状态，这些 ISR设备都属于该指定的出口 ISR所在的转发通路。接收到 ESTABLISH消息的ISR首先检查并确认该消息来自去往目的端的下一跳 ISR。它还通过查找路由器通路对象中自己的 ISR ID，验证该消息以前未曾处理。如果任何一项检查未通过，向发端都发送一个否定的确认。如果两项检查都通过了，接收端用一个肯定的确认对发端响应。然后它更新 VCIB、增加一跳、在路由器通路对象中添加自己的 ISR ID、在通往相邻ISR设备上游端口分配一个新的标志、并向所有相邻的上游ISR设备发送 ESTABLISH消息。ESTABLISH消息中包含的重要的对象有：

- 出口ID。该对象包含出口标识符。

- 路由器通路。这是 ESTABLISH消息所通过的ISR设备的列表，并且只有当需要避免回路时才包含它。

- 标志。包含ISR的输入端口分配的标志，并在 ESTABLISH消息中顺流而上发送到相邻的ISR。

- **TRIGGER**。它由上游ISR发送到下游ISR，请求一个ESTABLISH消息。通常当拓扑变化创建了新的到出口 ISR的转发通路（进而创建新的交换通路），从而需要重新计算转发通路，会出现这种情况。这时过期的通路将被拆除，并向新的下游 ISR（去往出口ISR的新的下一跳）发送TRIGGER消息。

- **TEARDOWN**。该消息删除已建立的交换通路。例如，如果选路协议或者人工配置取消了一个出口ID的服务，该ISR设备应该与删除出口ID以后剩余的ISR设备交换TEARDOWN消息。

- **ACKNOWLEDGE**。该消息提供对 ARIS消息的肯定或者否定的应答。

12.4.2 对象

ARIS规范所定义的 ARIS协议消息的对象是：

- **标签**。这是由下游ISR指派给某个出口ID的标签。图12-10所示为标签对象的格式。

E	保留	V	VPI	VCI
---	----	---	-----	-----

图12-10 标签对象

- **E**。E位是一个端到端的确认位，如果开始于入口 ISR的上游ISR设备分配了标志，该位置位。

- **V**。只有当VPI字段重要时，才置位V位。

- **VPI**。12位的VPI字段。

- **VCI**。16位的VCI字段。

- **出口ID**。这个对象包含12.3节所描述的出口标识符之一。图12-11为几种常用值的格式。字段的内容一目了然。

- **多路**。这个对象是一个标识符，定义对于同一出口ID的几个可能的成本相同的通路之一。

- **路由器通路**。这个对象包含 ISR标识符的列表，这些 ISR标识符表示去往一个给定出口ID的ESTABLISH消息所经过的通路。

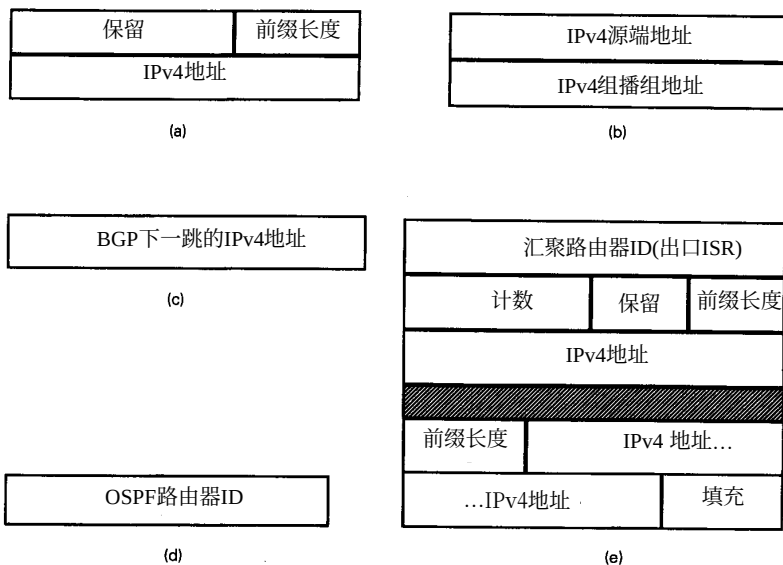


图12-11 出口ID格式

a) IRv4 地址 b) 组播 (源端, 组播组) c) BGP下一跳 d) OSPF 路由器ID e) CIDR

• 显式通路。显式通路对象包含 ISR 设备的列表，ESTABLISH 消息必须经过这些 ISR 设备以建立非缺省的交换通路。

- 隧道。这个对象包含链路层的标签以及与这个标签相联系的 IPv4 地址前缀。
- 定时器。根据定时器对象所在的消息类型，它可以有不同的含义。在 INIT 消息中，它是邻机超时值。在 ESTABLISH 消息中，它是建立刷新的间隔。
- ACK。这个对象包含被应答的消息的序号。
- Init。该对象包含 ISR 能够支持的 VPI 和 VCI 值的范围。

12.4.3 交换通路的建立

图12-12所示为利用 ARIS 协议建立交换通路的例子。ISR 设备的网络包含在中间域内部，中间域连接到非 ARIS 域。网络 1 是包含在非 ARIS 域中的几个目的端网络之一，ISR 设备可能需要向这些目的网络转发业务流。

首先 ISR 设备与邻机交换 INIT 消息，以形成关联并进入激活状态。在处理任何通路建立消息之前这个过程是必须的。

网 1 的一个出口 ID 配置到出口 ISR^①。这可以动态的通过“下载”路由表或者通过人工网络配置来完成。网络 1 是 ARIS 域以外的网络并且可以通过出口 ISR 4 到达。

出口 ISR 顺流而上向 ISR 3 发送 ESTABLISH 消息。建立消息包含出口 ID、新标签 (VCI=9)、以及 ESTABLISH 消息所通过的 ISR 设备的列表 (路由器通路对象)。这种情况下列表只包含出口 ISR 的 ID (4)。

ISR 3 收到 ESTABLISH 消息并且 (1) 通过查看自己的 FIB 表，检查以确认它来自正确的通往出口的下一跳地址，ISR 2 通过查看 ISR 列表，确保通路是无环路的。然后 ISR 3 用新的出

① 在这个例子中，出口 ID 的类型是 IPv4 地址前缀。

口ID更新自己的FIB，并且在自己的VCIB创建一项，其中包含出口ID（网络1）和新的标签（VCI=9）。ISR 3向出口ISR发回ACKNOWLEDGE消息。

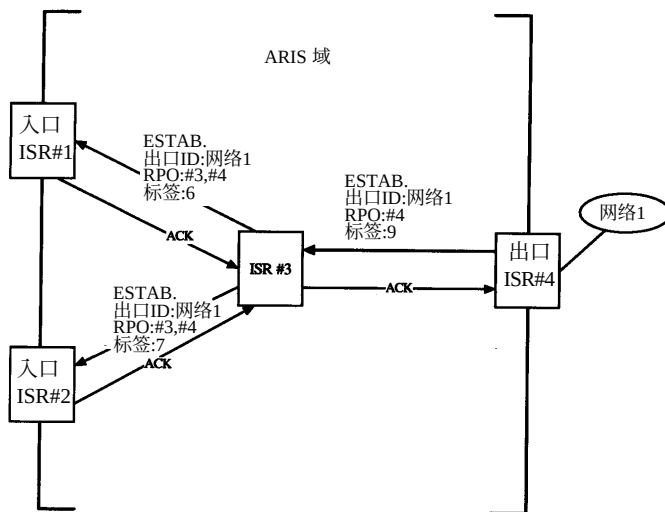


图12-12 ARIS 通路建立实例

接下来ISR 3为它的上游邻机ISR 1分配一个新的标签（VCI=6），ISR 2分配一个新的标签（VCI=7）。它还把自己的ISR ID（3）添加到路由器通路对象中，然后向ISR 1和2分别发送ESTABLISH消息。当ISR 3收到所有上游邻机的ACKNOWLEDGE消息，它更新VCIB，把输入端口的上游标签与输出端口的下游标签连接起来，该下游标签与网络1的出口ID相联系。这个过程递归式的重复进行，直到通往网络1的路由通路上的所有ISR设备已经相应的更新了自己的FIB和VCIB。最终结果是形成一个从所有入口ISR设备到出口ISR的多点到点交换通路。

12.5 环路避免

当实际网络拓扑与路由表的内容不一致时，可能发生选路环路，将导致IP分组在路由器和网络中超过一次的“循环”。利用IP头的TTL字段可以控制这种情况，当分组通过每个路由器的时候，递减该值。一旦分组的TTL值减为零，路由器就丢弃该分组。而ATM中没有这种机制。由于通常ATM硬件运行的速度远远超过IP路由器，第二层的环路将消耗大量带宽和交换机资源。这将对其它无环路的业务流产生有害的影响，并且可能影响网络基础设施支持新的或者已有的业务流的能力。

在ATM网络上运行是问题的焦点，通过在建立交换通路的同时就避免环路，ARIS提供环路问题的解决方案。由于它是由出口发起并且从网络边缘建立通路，ARIS通路建立过程容易跟踪直到出口ISR的通路。具体来说，在ESTABLISH消息中包含路由器通路对象就可以实现，每个处理ESTABLISH消息的ISR的ID都更新该对象。如果路由器通路对象中已经包含了ISR自己的ID，该ISR将不处理ESTABLISH消息。

从出口来避免环路是相当直接的。但是拓扑的变化可能发生在IP网络的任何位置，如果网络内部发生变化，如何防止环路？ARIS向通往出口ISR的新的下一跳ISR发送TRIGGER来解决这个问题。然后下一跳ISR发出ESTABLISH消息，以替代从入口ISR到发出TRIGGER消

息的ISR的上游交换通路。但这个过程有一定的代价，因为可能拆除一个最优的上游交换通路，而在新ESTABLISH消息到达后重新建立交换通路。换句话说，可能没有必要拆除原有的上游交换式通路。

利用分布式计算可以解决这个问题，当拓扑变化时计算无环路的新的通路^①。分布式计算的思想是当下游ISR切换到新的交换通路时，首先判断然后裁剪可能形成环路的上游链路和节点。在分布式计算完成之前，仍然使用原有的交换通路。具体来说，当ISR（A）从新的下一跳收到ESTABLISH消息，它检查自己是否属于新的通路。如果是，则形成了环路。否则，发起分布式计算过程。这时向上游ISR邻机发出查询消息，以判断当上游邻机和ISR A在新的交换通路连接起来时是否形成环路。如果的确形成了环路，把那些上游邻机从新的交换通路中裁剪掉（在这一过程中删除掉而不再继续考虑）。如果没有形成环路，上游邻机用允许消息响应，并且添加到上游树中。当ISR A从所有自己的邻机收到允许消息，它丢弃原通路并把新的上游树连接到新的交换通路上。这种解决方案的优点是当下游拓扑变化时上游交换通路未必受到影响。

12.6 ARIS服务

出口ID的概念和边缘的发起过程（入口或者出口）提供了具有单控制点的ARIS网络，同时在支持网络层业务时具有灵活性。通过调整单一出口ISR的出口ID，可以把可变的业务量颗粒度的等级映射到不同的多点到点交换通路。另外，ARIS不需要修改任何现有的选路协议就能够支持单播和组播选路。同时VC汇聚节省资源的机制以及多点到点的交换通路解放了更多的交换机资源，以便利用ARIS或者非ARIS方法建立特殊的非缺省的连接。实际上ARIS是仅有的几个运行在多业务ATM基础设施上的控制协议之一。正如标记交换一样，第二层的标签替换转发功能独立于所采用的网络层业务的类型。

12.6.1 出口汇聚

基于IP的大型骨干网同时支持成百上千的IP数据流。与把单个IP数据流或者甚至单个路由表项映射到单个交换通路相比，把共享同一出口的IP数据流汇聚到单个交换通路将减少网络资源的消耗。建立交换通路的控制消息以及网络中交换通路的状态都比较少。配置出口ID把一组目的端与一个出口ISR联系在一起，使得ARIS利用最少的交换机资源交换所有出口方向的业务流。

12.6.2 组播

ARIS可以为IP组播业务流建立交换通路。交换式点到多点传送树的建立由根或者入口发起。利用中间ISR设备的基于硬件的交换，交换通路树把所有组播业务流从入口ISR传送到出口ISR^②。与路由器基于软件的组播选路相比，这是显著的改进。

ARIS独立于所应用的任何组播选路协议。ARIS支持每个〈源端，组地址〉对建立数据驱动的以源端为根的传送树，DVMRP和PIM密集模式（PIM-DM）需要这种传递树。同时它

① 参见J.J. Garcia-Luna-Aceves, “Loop-Free Routing Using Diffusing Computations,” IEEE/ACM Transactions on Networking, vol. 1, no. 1, Feb. 1993.

② 在入口和所有出口ISR仍然需要标准的第三层组播转发。

支持为每个〈*, 组地址〉建立接收端驱动的共享式传递树, PIM稀疏模式 (PIM-SM) 需要这种传送树。

12.6.3 显式选路

在网络中设置非缺省选路是重要和所期望的功能, 其目的是提供特殊服务 (例如 QoS、性能、安全性等等) 或者仅仅为平衡负载。利用出口 ID 和边缘发起使得 ARIS 能够很自然的提供这种功能。ARIS 通过上游显式通路对象支持这种功能。显式通路对象的内容由 ISR 设备矢量构成, 对于给定出口 ID, ARIS ESTABLISH 消息必须通过这些 ISR 设备。从而形成入口和出口 ISR 设备之间的无环路的非缺省的交换通路。显式选路对象的内容可以人工定义、由选路协议 (例如 OSPF) 计算、或者通过其他一些方法建立。ARIS 支持出口或者入口节点发起的点到点、点到多点以及多点到点的交换通路。图 12-13 所示为通往网络 B 的业务流的显式通路的例子。

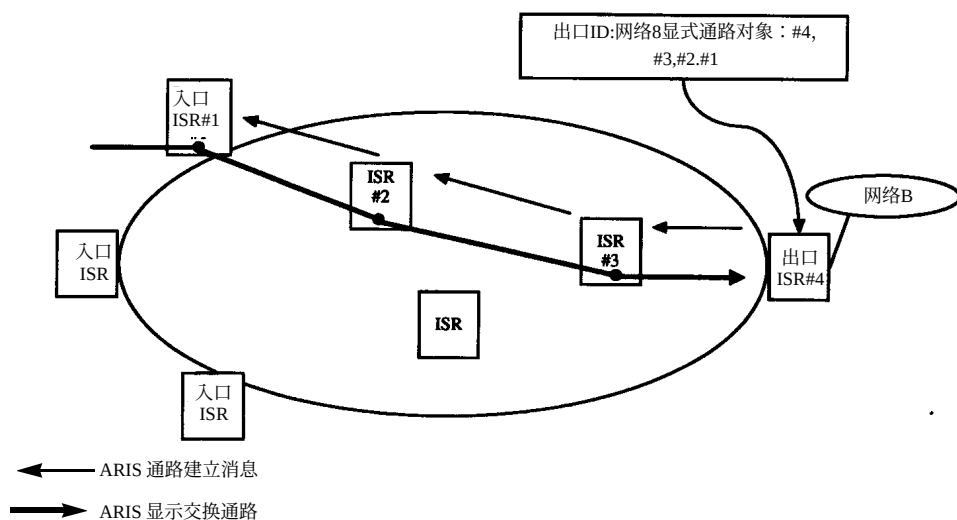


图12-13 ARIS显式选路

12.6.4 ATM

ARIS 设计为工作于任何第二层的交换技术。例子包括帧中继、ATM 以及 IEEE 802 局域网交换。然而出于几种原因, ATM 值得特别关注, 包括支持 QoS、在 OC12 以及更高的速率下的可扩展性、大量供货商所提供的 ATM 交换机。而且许多大型 IP 网络都采用了 ATM 以提供附加的网络能力和性能。

为了利用 ATM, ARIS 所做的调整包括:

- 环路避免。ARIS 具有内在的避免形成第二层环路的机制。由于 ATM 没有环路避免机制, 这个能力十分重要。
- 与透明配置共存。具有 ATM 交换机组件的 ISR 支持诸如 ATM 论坛的信令和选路以及 PVC 和 PVP 的 ATM 协议。
- 互操作性。采用图 12-14 所示的两种可能的技术, ARIS ISR 设备可以通过纯 ATM 交换机

的网络进行彼此的通信^⑨。第一种方法是在相邻的 ISR 设备之间通过 ATM 交换机网络建立 VP(见图12-14a)。所有的数据和控制业务流利用这个 VP 传送, 并从 VP 未用的 VCI 值中选择与出口标识符相联系的标签。第二种方法是在相邻的 ISR 设备之间建立 PVC 或者 SVC(见图12-14b)。它是缺省 VC, 传送选路业务流以及包括 ESTABLISH 消息在内的控制业务流。当 ARIS ESTABLISH 消息向上游传送, 下游 ISR 通过 ATM 交换机网络建立与出口 ID 相联系的 SVC。ESTABLISH 消息以及 ATM SVC 建立消息都包含一些信息(称作句柄), 以便使收端 ISR 设备(收到 ESTABLISH 消息者)能够通过新建立的 ATM SVC 与出口 IP 联系起来。

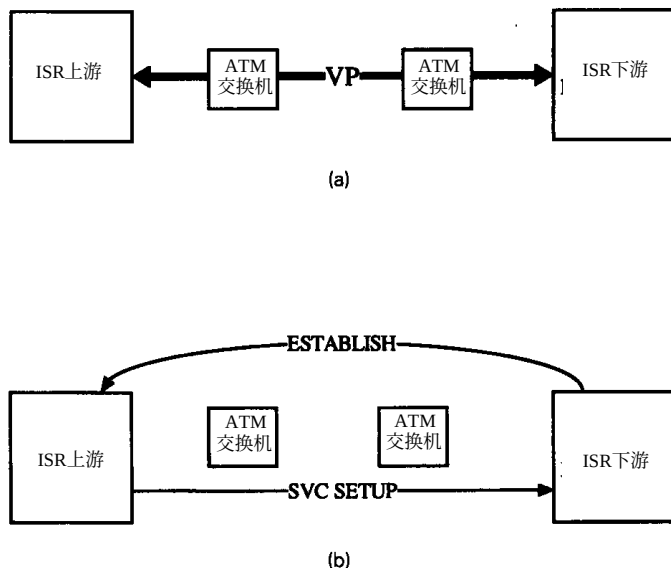


图12-14 ARIS 对ATM支持的两种方式

12.6.5 第二层隧道

ARIS支持建立交换通路的隧道。交换通路的隧道在第二层把数据交换到出口 ISR或者隧道端点。隧道中传送的数据可以携带标签堆栈, 其中只有最顶端的标签被处理(进行标签替换)。在隧道的端点, 顶端标志(例如 VPI)被“移去”, 然后根据堆栈中的下一个标签(例如 VCI)交换数据。因此在出口 ISR没有第三层处理。值得注意的是隧道中的数据仅仅根据堆栈顶端的标签进行交换, 即使堆栈中存在其他的标签, 也不会涉及。ARIS建立的第二层隧道的可能的应用包括层次结构(域内选路使用一种标签、域间选路使用另一种标签)、DVMRP通道以及第二层的数据流的汇聚/拆散。

图12-15所示为 ARIS 网络支持的入口和出口 ISR 之间建立的第二层隧道。根据 VPI 值(在信元头中)在隧道中转发数据。当数据到达出口 ISR(或者隧道的端点), VPI 值被移去, 然后使用堆栈中的下一个标签, 即 VCI 值, 在第二层转发数据。利用 ESTABLISH 消息中的通道对象, ARIS 能够通知入口 ISR 堆栈包含哪些标签(例如 VCI 值)。用这个值可以在已有的通道上把业务流交换到某个目的端。这个例子中, 用 VCI=1 来交换通往网络 1 的数据,

^⑨ 这种机制与第 10 章中讨论的 CSR/FANP 模型相似。

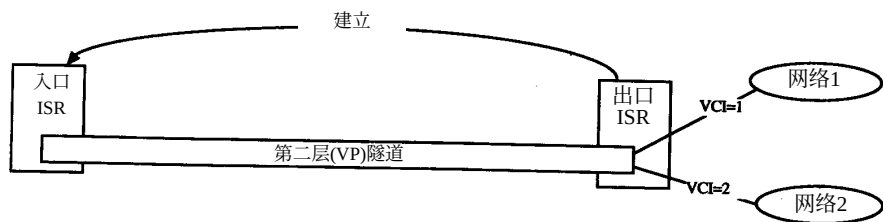


图12-15 ARIS对第二层隧道的支持

而用 VCI=2 来交换通往网络 2 的数据。由于避免了在第三层拆散并且处理分组，性能得到提高。

12.6.6 服务分类

有几种不同的技术使基于 ARIS 的网络可以支持基于 COS/QoS 的连接。首先，可以使用 RSVP 在 ISR 设备的通路沿途设置 QoS 和交换通路状态。这种方法与上一章介绍的标记交换所使用的方法非常相似。沿着由 RSVP PATH 消息建立的通路，包含在 RSVP RESV 消息中的标志对象顺流而上直到发端。RESV 消息包含设置到 ISR 交换机组件的标志以及 QoS 参数。一旦为某个数据流在输入和输出端口分配了标志，ISR 可以把二者连接起来形成交换通路。当入口到出口通路上的每个 ISR 完成了该操作，IP 数据流就可以利用下面的第二层（例如 ATM）的基础实施的高带宽和 QoS 能力。

另一种虽然与 ARIS 无关但是也能够节省交换机资源的技术是采用纯 ATM 的信令和选路。这要求 ISR 在交换机组件中支持 ATM 论坛的信令和选路（PNNI 的第 I 阶段）。网络运营者因此利用 ARIS 交换所有的尽力而为的 IP 业务流，同时为那些要求 ATM QoS 的应用和最终用户提供 ATM 支持。

还有一种方法是由 ARIS 支持通往某个目的端的多点到点交换通路。例如，去往某个目的端的任何业务流都利用缺省的交换通路 1。利用独立的出口 ID，可以为源端发出的去往相同目的端的业务流建立第二个交换通路，即交换通路 2。这种情况下，ARIS 允许为每个源端分别建立到达相同目的端的交换通路，从而提供一些服务差分等级。

12.7 ARIS 网络实例

ARIS 是拓扑驱动 IP 交换解决方案，它最适合大型基于 IP 的网络。通过精心设计，交换任意两个节点之间的所有 IP 业务流所需的网络资源达到了最小。这意味着更多的网络（交换机和路由器）资源可以分配给非缺省的特殊业务。这些业务可能包括虚拟专用网（Virtual Private Networks, VPN）、QoS 业务流、显式选路、隧道、纯 ATM 连接等等。总之，ARIS 允许大型网络交换大量的汇聚的 IP 业务流，从而提高整个网络的性能。

ARIS 非常适合于运行在 ATM 基础实施上的大型多业务 ISP 应用。图 12-16 为这种网络的一个例子。请注意，这种 ISR 设备构成的网络支持 IP 选路、利用 ARIS 的 IP 交换业务以及纯 ATM 交换。路由器作为网络的外设，在第三层把分组转发到骨干 ISR 设备，或者如果有 ARIS 协议，可以建立直接的入口到出口的交换通路。

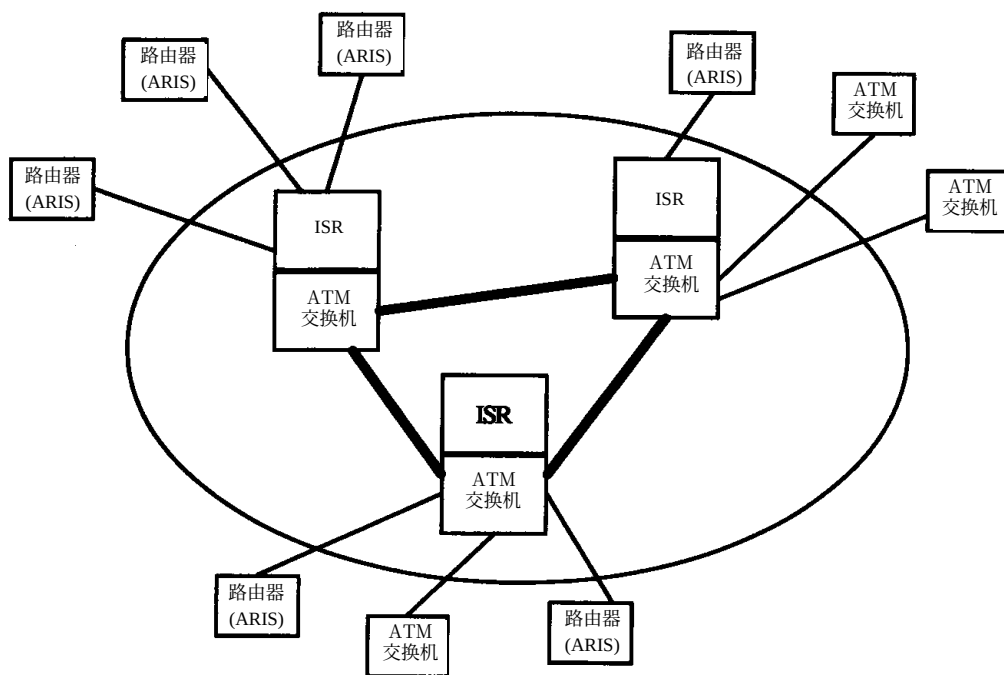


图12-16 ARIS网络实例

更多的信息

Feldman et al., "Aggregate Route-Based IP Switching (ARIS)," IBM Technical Report, TR 29.2353.

Feldman et al., "ARIS Protocol Specification," IBM Technical Report, TR29.2368.

Metz, C., "Aggregate Route-Based IP Switching (ARIS)," IBM White Paper, Aug. 1997.

China-pub.com

下载

第13章 其他解决方案

到此为止，本书归纳了IP协议集、局域网和广域网交换技术、ATM上的经典IP的解决方案、ATM论坛的局域网仿真和MPOA解决方案以及Ipsilon、东芝、Cisco和IBM公司的四种著名的IP交换解决方案。然而，采用基于硬件的交换来加快IP转发过程的解决方案远不止这些。当然仅仅考察ATM方面的研究和开发，用ATM交换机来转发或者映射IP数据流的解决方案也远不止这些。供货商和研究者提出了大量的其他解决方案，通过把选路和交换结合起来，能够快速地为IP网络和应用提供更好的服务。这一章将对这些已经提出的支持IP交换功能的其他解决方案作一归纳。

13.1 FastIP

3Com公司开发和引入了叫做FastIP (FIP)的在第二层交换网络上进行IP交换的解决方案^①。FastIP在端点之间采用了NHRP，这些端点连接到相同的物理第二层的网络，但是逻辑上它们属于不同的子网。实际上FastIP可以看作NHRP的扩展，它使得以太网局域网上的不同子网的端点可以建立直接的二层交换通路，而不必通过路由器转发分组。FastIP采用NHRP把目的端的IP地址解析为MAC地址，而不是NBMA地址（例如ATM）。一旦发端把目的端的IP地址解析为目的端的MAC地址，两个端点就可以建立基于MAC地址的第二层交换式通信。然而这种环境中仍然出现路由器，以便必要时在子网之间转发分组。路由器还可以作为直通路径的管理者，换句话说，可以把路由器配置为用来选择适合交换式通路服务的某些业务流。

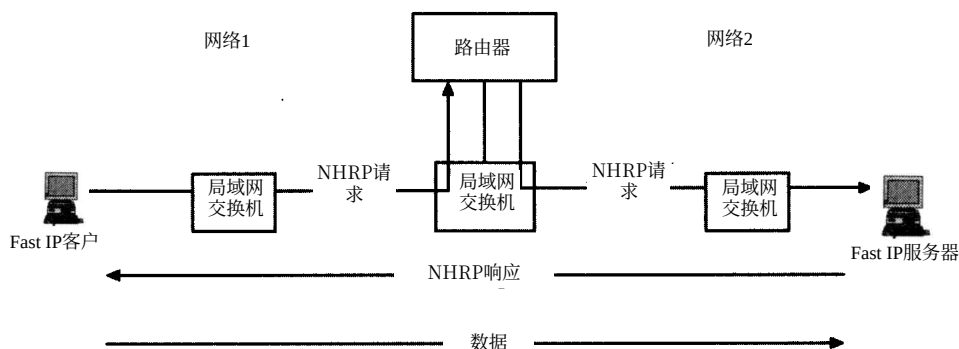


图13-1 FastIP操作

如图13-1所示为FastIP的基本操作。在相同的第二层交换网络（例如以太网）上连接了一个客户机和服务器，但它们属于不同的子网。首先客户机和服务器都要配置FastIP的驱动程序。初始状态下数据流通过缺省的网关路由器。在一定的时间间隔内发送了一定数量的分组以后，客户机发出NHRP请求并通过路由器转发这个请求。NHRP请求包含源端的MAC地址，并通过传统的IP路由转发到目的端，在这个例子中也就是服务器。而转发分组的路由器首先判断源

^① <http://www.3com.com>。

端（客户机）和目的端（服务器）是否适合用直通路通信。根据路由器中配置的规则决定是否允许NHRP请求通过。如果其规则指明客户机和服务器通过直通路通信，则把NHRP请求直接转发到目的端，即服务器。否则，路由器将丢弃NHRP请求，分组将继续通过路由器转发。

运行FastIP驱动程序的服务器识别出NHRP请求，利用其中的源端MAC地址，直接向发起请求的客户机发出NHRP响应。NHRP响应的返回数据通路上的第二层的交换机根据客户机的MAC地址交换分组。

在客户机接收到NHRP响应以后，它利用服务器的MAC地址而不是通过路由器为分组寻址，转发去往非本地目的端（服务器）的分组。像其他IP交换的解决方案一样，FastIP的目的是旁路中间路由器的处理，使得不同子网的端点能够利用彼此之间的第二层交换式通路直接通信。

FastIP的另一个特点是能够支持交换式虚拟局域网之间通信。在NHRP请求中包含IEEE 802.1Q虚拟局域网标记可以实现。目的端可以安装MAC地址和虚拟局域网ID为NHRP响应寻址。源端和目的端之间的所有后续通信都使用标准的802.1Q局域网标记。这使得本质上是一个封闭广播域的虚拟局域网向非虚拟局域网成员“开放”并支持直接的交换式连接，当然允许进行一定的管制。

FastIP可以归类为数据流驱动的IP交换解决方案，采用与MPOA相似的“一次选路——多次交换”的方法。FastIP扩展到以太网介质，所以它不仅仅是ATM解决方案。为了组建FastIP网络，服务器和所有客户机都必须用FastIP驱动程序升级。在交换式的基础结构上存在NHRP响应泛滥的问题。不考虑这一点，把数据流从基于路由器的处理中解放出来，为交换式的IP业务流提供了更好的性能。

13.2 交换式多协议服务

除了ARIS的贡献，在IP交换的范畴内，IBM公司还开发和引入了几种其他技术^①。交换式多协议服务（Multiprotocol Switched Services, MSS）是一种增强在交换网络（特别是ATM网络）中IP互联的性能和可扩展性的服务。它是对基于标准的局域网仿真和NHRP协议的功能扩展。具体来说，服务器组件（LES/BUS、NHRP）采用智能的方式维护信息，因此优化了数据转发通路，减小了网络开销。以另一种方式，服务器组件协助客户机建立彼此之间的交换式通路，同时减少了额外的广播业务量。

第一个MSS功能称作通路交换（Route Switching, RS）。像FastIP一样，它使得连接到局域网上的不同子网的端点能够建立第二层交换式连接，以避免中间选路。通路交换扩展了NHRP的能力，支持连接到ATM骨干网上的以太网或者令牌环网的端点进行IP到MAC的地址解析。因此通路交换的目的是在ATM骨干网上支持零跳选路。

图13-2所示为一个典型的通路交换网络。客户机和服务器连接到不同的以太网段，所有以太网段都连接到由一个或者多个仿真局域网组成的ATM骨干网。客户机安装路由交换客户（Route Switching Client, RSC）软件，它可以请求一条通往目的端的直通路。路由交换服务器（Route Switching Server, RSS）作为子网之间的缺省IP路由器、连接到仿真局域网的主机、边缘设备的LES/BUS以及下一跳服务器（NHRP服务器）。在这个例子中，首先客户机发

^① <http://www.networking.ibm.com>。

出NHRP请求，查询IP地址127.0.0.1，发现RSS的MAC地址。RSS回应以NHRP响应，RSC保存响应中的源端MAC地址。初始状态下，客户机发往服务器的分组寻址到缺省IP路由器（这个例子中是RSS）的MAC地址。当RSC检测到寻址到缺省IP路由器的MAC地址的分组，RSC发出NHRP请求，试图把目的端的IP地址解析为MAC地址。RSS用服务器的MAC地址响应NHRP请求，该MAC地址是通过以前的注册或者ARP过程知道的。另外，NHRP响应可能包含附加的信息（例如入口ATM地址、令牌环网通路的描述器），在客户机和/或源端边缘设备建立从客户机到服务器的第二层的交换式通信通路时，这些信息是必要的。例如，入口边缘设备需要出口边缘设备的ATM地址，以便建立穿越ATM骨干网的SVC。一旦这个信息传送到源端和入口组件，就能够建立穿越ATM骨干网的直通过程。客户机用NHRP过程学习到MAC地址，把所有分组寻址到服务器。入口边缘设备沿着直通过程，把分组放置在通往出口边缘设备的适当的SVC上传送。沿着直通过程不会遇到选路处理，因此称为零跳选路。

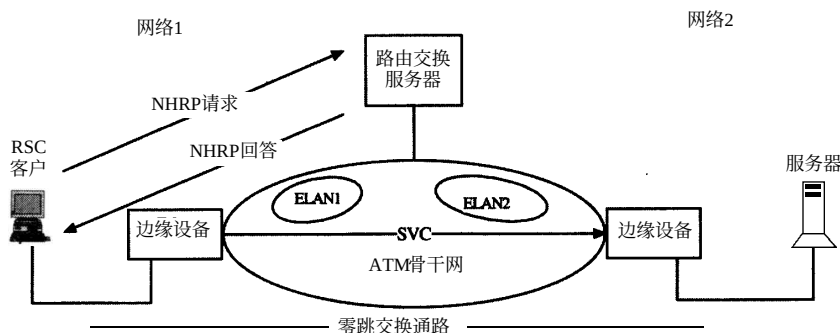


图13-2 路由交换网络

通路交换支持连接到以太网和令牌环网上的主机，而源端和目的端必须连接到相同类型的局域网。与MPOA或者NHRP客户的功能不同，通路交换不需要增加边缘设备的功能。另外它还支持非对称直通过程的概念，使得直通过程可以只在一个方向建立。例如，在图13-2所示的服务器配置RSC程序，使它能够建立通往任何客户机的直通过程，尽管客户机并没有安装RSC程序，也没有参与RS过程。如果客户机没有RSC程序，所有客户机到服务器方向的业务量都通过IP路由器在缺省选路转发。这种特殊的功能使得大量的服务器到客户机的业务流享受第二层交换式通路的服务，而客户机到服务器业务量仍然采用选路。唯一的变化是需要服务器安装RSC程序。客户机没有任何变化。

另一个MSS支持的IP交换功能是超仿真局域网。超仿真局域网是两个或者多个仿真局域网的集合，它们用超仿真局域网的桥接功能连接到一起。每个仿真局域网支持诸如LES/BUS/LECS的标准局域网仿真1.0版本的服务。超仿真局域网支持不同仿真局域网中的LE之间的直接VCC数据通路，这消除了仿真局域网之间通信所需要的中间桥接处理。这个过程也叫做直通路桥接（Short-Cut Bridging, SCB）。

如图13-3所示，SCB使得LE客户在穿越仿真局域网边界时旁路了标准的桥接处理。超仿真局域网的桥接功能可以驻留在各个仿真局域网的LES/BUS实体中，或者如图所示运行在独立的设备上。它配置为传统的网桥，其桥接端口作为LEC代理模块并提供特殊的SCB处理。SCB处理是在仿真局域网之间转发某个局域网仿真控制帧，包括LE_ARP请求/响应。初始化的点到点的帧也是通过SCB端口在仿真局域网之间转发，其结果是一个仿真局域网上的源端

LEC A.1与另一个仿真局域网上的目的端 LEC B.2建立数据直通路VCC。

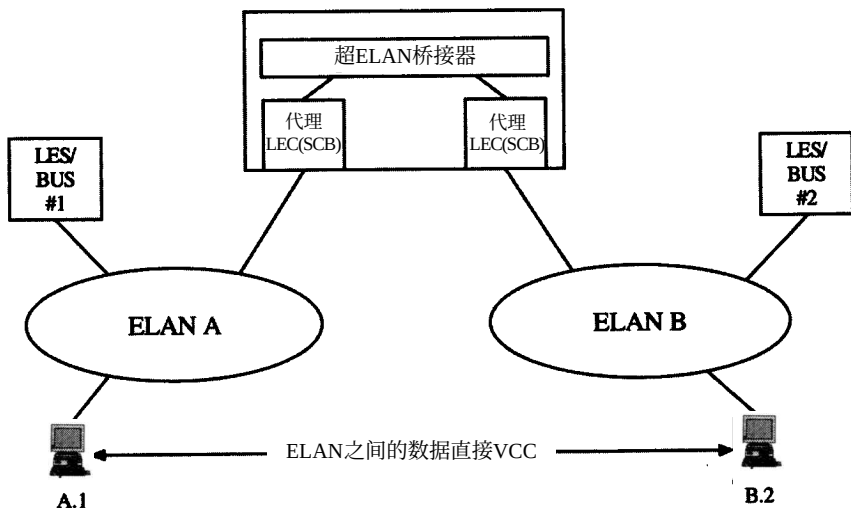


图13-3 超ELAN直通路桥接

为了控制广播的扩散并且使基于 IP 的仿真局域网更易于扩展，MSS支持称作广播管理（Broadcast Manager，BCM）的附加功能。对局域网仿真 1.0版本的IP子网，广播帧通过点到点的组播传送VCC发送到BUS，然后BUS利用点到多点的组播转发VCC向所有LE客户转发广播帧。把某个ARP请求转移到已有的组播发送VCC上，IP子网广播（ARP请求）的规模可以大大缩减。BUS之所以能够这样工作是因为BUS探测所有ARP请求并予以响应，以及在高速缓冲存储器中保存了IP/MAC/LEC的关系。事实上，广播帧被转化为单播帧，这使得大规模点到多点的广播树被“裁剪”成单一的分支，因此降低了网络资源的开销，并且能够产生大型基于IP的仿真局域网。值得注意的是广播的缩减是在IP子网内部。

两种附加的MSS功能控制了超仿真局域网中广播的扩散：桥接广播管理器（Bridging Broadcast Manager，BBCM）和动态协议过滤器（Dynamic Protocol Filtering，DPF）。BBCM与上面描述的广播管理器功能相似，只是它是用超仿真局域网的桥接功能实现的。它减少了不必要的IP ARP广播业务量和仿真局域网网段的处理。通过了解每个桥接器端口所支持的协议，DPF能产生某种协议的虚拟局域网。某种协议或者网络的广播分组的范围限定在已知能够支持这种协议的桥接器的端口上。这进一步缩减了大规模桥接网络中可能遇到的广播业务量。

总之，MSS支持增强基于IP的局域网和ATM交换环境的扩展性和性能的几种功能。对于跨越ATM网络的以太网或者令牌环网的端点，选路交换构成零跳选路。超仿真局域网功能在现有ATM骨干网上实现大型IP子网并且提供与第三层选路所提供的广播控制功能相同的功能。实际上，由于BCM功能减少了子网内部的广播业务量，超仿真局域网性能更好。

13.3 IP导航器

Ascend公司的IP导航器（IP Navigator，IPnav）是在其广域网的帧中继和ATM交换机网络上提供IP交换业务的解决方案^①。传统的连接到交换机网络上的路由器彼此之间需要建立虚连

^① <http://www.ascend.com>。IP导航器首先在Cascade公司的广域网交换机技术中开发出来。Ascend公司1997购买了Cascade公司。

接, 导致 $O(N^2)$ 个VC, 其中 N 是路由器的数量。IPnav解决了这种扩展性问题, 在广域网交换机的网络上用多种方法提供IP交换业务。首先, 广域网交换机运行诸如OSPF的IP选路协议的实例, 使广域网交换机能够发现网络的IP拓扑, 并与外部路由器交换更新路由表的消息。其次, 广域网交换机建立多点到点的树, 以出口广域网交换机作为根并分叉到所有入口广域网交换机, 这使得消耗的VC资源最小, 并且提供了超越网络的交换式通路。最后, IPnav采用了Ascend的虚拟网络导航器(Virtual Network Navigator, VNN)所支持的VC信令和选路协议作为选择符合数据流特性的最佳通路的方法^①。

图13-4所示为一个运行IPnav的Ascend的广域网交换机网络。对于某个出口广域网交换机, 建立了一个以此为根的多点到点的树并分叉到所有入口广域网交换机。这是基于出口节点标识符, 与ARIS提供的功能相似。对于基于ATM的网络, 多点到点的交换式通路实际上由虚通路构成, 入口广域网交换机选择唯一的VCI值, 以保障信元到达出口广域网交换机时源端的唯一性。当一个分组到达入口广域网交换机, 进行标准的路由表查询。然而, 与标准IP选路情形下把分组转发到下一跳路由器不同, 入口广域网交换机把分组放到预先建立的通往出口广域网交换机的VC上。IPnav网络内部没有中间的第三层处理。

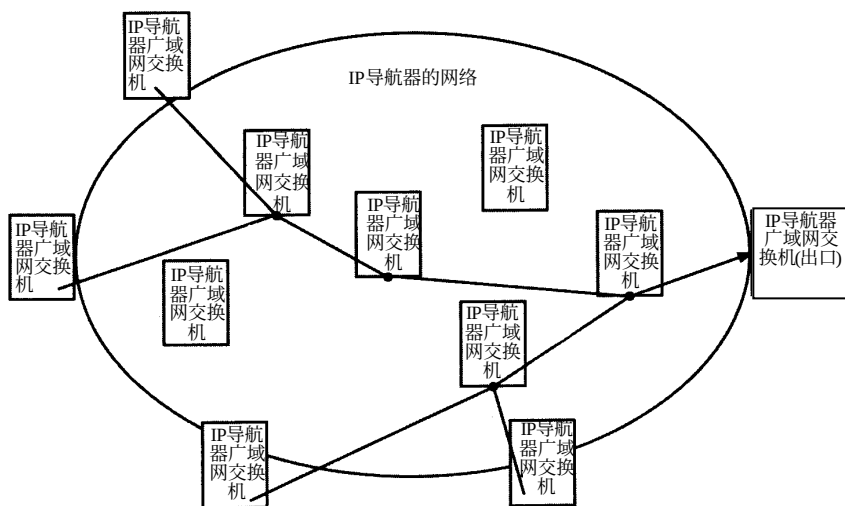


图13-4 IP导航器的网络

IPnav可以归类为拓扑驱动的IP交换解决方案, 它是为运行在Ascend的广域网交换机上的大型IP网络而设计的。尽管它可以与外部路由器互操作, 并且支持所有的单播和组播选路协议, 但是它的应用仅局限在Ascend的广域网交换机的范畴。还应该指出的是广域网交换机除了IPnav提供的IP交换功能, 还支持帧中继和ATM交换业务。对于具有Ascend的广域网交换机的网络运营者, 当希望把自己的网络升级到支持IP交换, IPnav是一个很好的解决方案。

13.4 IPSOFACTO

IPSOFACTO这个巧妙的别名表示快速ATM信元交换传送上的IP交换(IP Switching over

^① VNN是对运行在Ascend公司的广域网交换机上的OSPF的扩展。诸如带宽、时延以及寻址信息的网络资源将进行通告, 这使得VNN交换机能够根据可用资源和业务量特性选择通路。

Fast ATM Cell Transport), 指的是通过 ATM 交换机组件对 IP 数据流的信元进行交换的技术。IPSOFACTO 是数据流驱动的 IP 交换解决方案, 每个 IP 数据流的分组由支持 IPSOFACTO 的 ATM 交换机进行交换。与 IFMP 和 FANP 不同, 它没有定义把数据流映射到 VC 的明确的控制协议, 而是由 ATM 交换机内部的 IPSOFACTO 实体来管理特定 IP 数据流的 VCI 值的分配和指派。另外, IPSOFACTO 实体管理 ATM 交换机的连接表项, 以便把特定 IP 数据流的输入和输出 VC 连接到一起, 形成内部直通路径。假如沿着路由通路的一组邻近的 IPSOFACTO 交换机都建立内部直通路径, 就可以在入口和出口之间形成交换通路。

为了支持 IPSOFACTO, ATM 交换机必须支持标准 IP 选路和转发并在交换机控制器中运行 IPSOFACTO 组件的一个实例。这种环境中不使用 ATM 论坛的信令和选路。图 13-5 所示为 IPSOFACTO IP 交换机以及把 IP 数据流映射到 ATM 信元交换式连接上的操作。初始情况下, 输入端口所有未用的 VCI 值都映射到 IPSOFACTO 实体。上游发送端在输出链路分配未用的 VCI, 并且用新的 VCI 标识这个数据流的信元。下游 IPSOFACTO IP 交换机在输入端口接收到信元, 把它们转发到交换机控制器。在 IPSOFACTO 组件用一个未用的 VCI 值标识了数据流的信元之前, IP 选路组件把分组转发到适当的下游节点, 其后输入和输出端口的数据流的信元都重新进行了标识, 因此交换机控制器可以简单地在连接表中增加一项〈输入端口、输入 VCI、输出端口、输出 VCI〉。

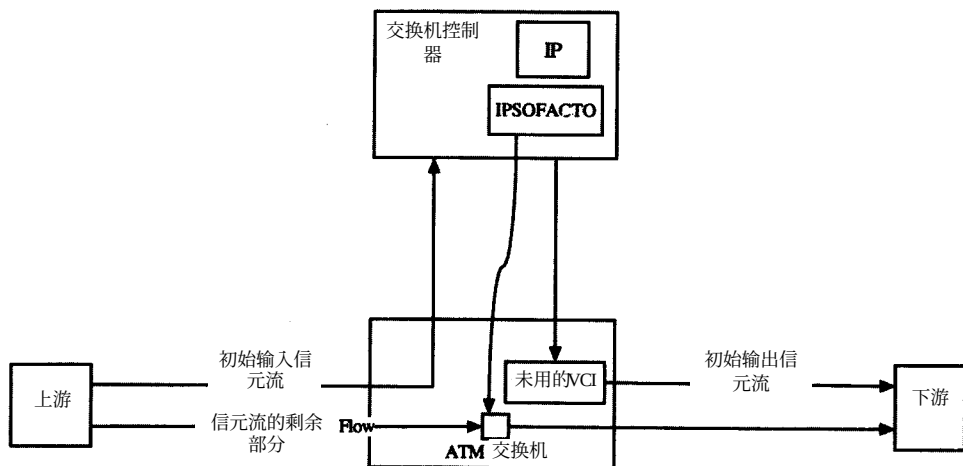


图13-5 IPSOFACTO

IPSOFACTO 支持单播和组播直通路径。在输出链路上分配未用的 VCI 可以以某些数据类型 (例如 TCP SYN)、控制流 (例如 PIM Join) 或者某些数据流到达的速率为基础来确定。

13.5 虚拟网络交换

虚拟网络交换 (Virtual Network Switching, VNS) 是由 Nortel 公司开发和引入的基于广域网的多协议交换技术^①。它运行在 Nortel 的广域网交换机上, 并且与所有 IP 交换解决方案一样, 把第三层处理限制在入口和出口设备。在 VNS 网络中根据目的端地址对分组进行第二层交换。另外, VNS 支持基于 COS 的选路、流量工程、虚拟专用网以及组播。VNS 网络可以与

① <http://www.nortel.com>。

大量外部逻辑网络和接口互通，例如局域网、PPP链路、局域网仿真 /MPOA网络以及RFC1483封装的ATM链路。RFC2340包含了VNS的概要。

VNS网络的外围是能够识别IP的入口和出口节点，并且由多个由点到点链路互联起来的支持VNS的交换机构成，看起来像单一的具有广播能力的局域网，其中所有节点都彼此直连起来。逻辑上交换机形成的这种结构称作逻辑网络（Logical Network，LN）。当分组进入VNS网络，入口节点判断分组转发到的出口节点。如图13-6所示入口节点在分组上附加特定的相同的VNS头。在网络中采用VNS选路而不是IP选路选择通往出口节点的通路。这一相同的头指示了分组是VNS封装的分组。VNS头包含了TTL字段、分组从属的逻辑网络号（Logical Network Number，LNN）、丢弃优先级、COS、链路选择标志、源端ID以及出口节点ID头。IP组播分组包含附加的组播组地址的4字节字段，它是LNN和出口节点ID值的组合，形成在VNS网络中转发分组的标签。该标签用OSPF的扩展项分发到LN的特定节点。



图13-6 VNS头

VNS解决方案的一个吸引人的特点是能够把网络配置成多个LN，从而把业务量分解到多个VPN。一个入口节点可能包含多个路由表的实例，其中每一个反映了一个特定用户VPN的拓扑。

由于VNS同时使用IP和交换机的拓扑在网络中选择通路，可以认为它是拓扑驱动的IP交换解决方案。

13.6 响应者发起的直通路程协议

响应者发起的直通路程协议（RISP）由富士通公司的研究者在1998 Engineering Interop会议的论文中提出^①。RISP的进一步的细节可以参阅在NBMA互通的邮件组中的提案^②。RISP是对等实体之间的协议，它使得主机或者路由器能够通过第二层交换式网络建立直通路程。它设计为运行在IP/ATM叠加网络上，其中同时保持IP和ATM技术以及选路协议。NHRP作为子网间基于客户机/服务器模型的地址解析协议，与之不同，RISP不是基于客户机/服务器的协议，也不需要网络中类似NHS的特定服务器建立直通路程。实际上RISP是对等实体之间的协议，发起和建立直通路程的实体用它通信。针对一个IP-ATM地址映射关系，不需要在建立直通路程之前首先查询NHRP服务器的功能。甚至可能根本不用任何地址解析服务而运行基于RISP的网络。

与所有IP交换的解决方案相似，RISP的目标是在不同子网的设备之间建立直通路程，以便旁路中间选路处理。RISP首先使用一个称作请求者的主机或者路由器，它向目的端（远端主机或者路由器）发出回呼请求的消息，企图建立交换式通路（VC）。回呼请求消息沿着路由通路向目的端转发，其中包含请求者的IP和ATM地址。如果实际目的端不是在ATM网络中，出口路由器将解释和处理回呼请求消息。无论目的端还是出口路由器（响应者）然后将建立到请求者的直接SVC。从请求者到响应者的数据将沿着直通路程传送。图13-7所示为这个基

① Ogawa et al., "Shortcut Routing for Networks with Heterogeneous Routing Mechanisms," Interop Engineer Conference, 1998.

② <http://netlab.ucs.indiana.edu/hypermail/ion/9708/0027.html>.

本过程。

与HNRP相比, RISP在子网间建立直通路路径的优越性具有两面性。首先, 中间路由器不需要运行NHRP并且维护IP-ATM地址对应关系表。其次, 建立直通路路径的延迟减小了, 因为发端不需要等待远端或者中间 NHRP服务器的响应。目的端对回呼请求的响应是建立直通路路径。

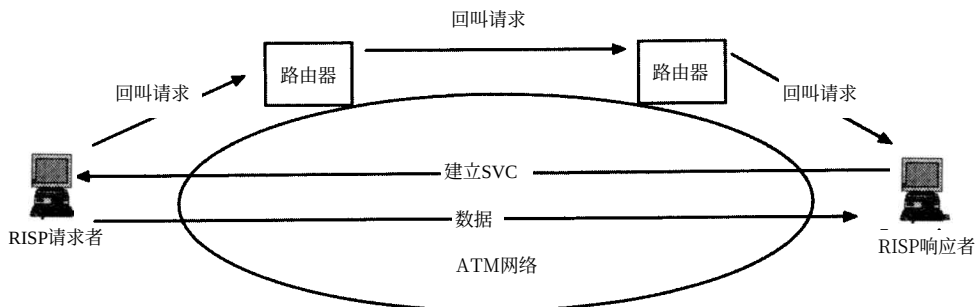


图13-7 RISP

RISP协议不要求中间路由器支持 RISP; 然而中间路由器必须能够转发 RISP消息。换句话说, 传统的连接到 ATM的路由器可以采用直通路路径的建立, 而无需运行特定的控制协议 (NHRP)。唯一需要支持 RISP协议的实体是请求者和响应者。RISP主机和路由器可以配置为经典IP的逻辑IP子网 (Logical IP Subnet, LIS), 仅需要一个ATM ARP服务器。ATM ARP服务器用来解析 IP/ATM映射关系, 以便在 LIS内部进行SVC通信。因此在子网之间建立直通路路径不需要中间路由器运行 NHRP。还可能通过独立的点到点链路, 把一个主机配置为连接到一个路由器上, 并使路由器为每个主机维护 (通告) 主机通路。这种情况下, 不需要诸如 ATM ARP或者NHRP的地址解析协议, 仅需要 IP选路协议 (例如 OSPF) 和RISP两个控制协议。值得注意的是RISP只支持单播连接。

13.7 DirectIP

Nbase通信公司的DirectIP是通过Nbase的局域网交换机网络在不同子网的连接到局域网的设备之间建立交换式连接的技术^①。在DirectIP解决方案中, 交换机监听并存储连接到交换机上的IP设备的IP和MAC地址。交换机使用这个信息来“欺骗”一个子网上的 IP源端, 使它认为在与同一子网上的目的端通信, 而实际上目的端在另一个子网中。尽管是明显的强词夺理, 但是不需要对端点进行任何修改, 子网中间的数据流可以享受线速第二层交换。DirectIP与DHCP配合可以简化交换机和主机配置的要求, 并支持虚拟局域网的概念。

图13-8中的两个图是基于 DirectIP的网络实现交换式连接的方式。首先, 发端发出对目的端的ARP请求。中间交换机检查目的端 IP地址并判断是否可以根据规则库建立交换式连接。如果不能建立交换式连接, 交换机无需目的端的 MAC地址, 只需将源端的后续分组进行第三层转发。如果可以建立交换式连接, 交换机将在对源端的 ARP响应中返回目的端的MAC地址。源端把目的端的IP/MAC地址映射关系保存到自已的ARP高速缓存区。源端现在使用目的端的

^① <http://www.nbase.com>。

MAC地址向目的端发送分组。交换机在第二层转发分组，而无需第三层处理。

DirectIP可以归类为IP学习的解决方案。它学习连接到交换机区段的设备的地址信息，然后使用这些信息在不同子网的设备之间建立第二层转发表。实际上IP自学习交换机作为目的端MAC的代理蒙蔽源端，使源端认为在与相同子网的另一个主机交换分组。由于在自学习过程中依赖广播，以及需要存储IP和MAC地址，IP自学习交换机非常适合校园局域网环境。根据实现的不同，它们可能需要也可能不需要运行动态IP选路。如果是后者，需要在交换机中人工配置IP信息。一般来说，尽管如此，仍然可以把IP自学习交换机看作保持局域网环境的简单快速IP转发器。它们把广播限制在配置的子网内部，并在子网之间以线速转发分组。但是它们可能没有与传统路由器相联系的完整的功能，例如动态IP选路或者组播。实际上一种DirectIP所推荐的应用情形是校园网路由器的前端，并且主要是为路由器分摊转发负荷。

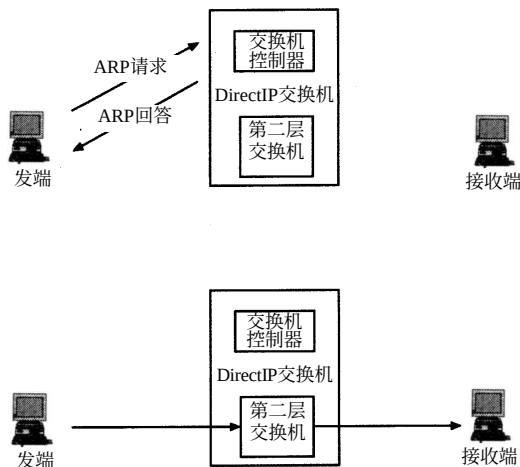


图13-8 DirectIP

13.8 AREQUIPA

一种已知的IP交换的局限是缺乏QoS支持，特别是当它运行在ATM上的时候。回想IP交换的目的是加速IP转发的过程，它没有隐含或者显式地要求IP交换在快速IP转发服务之外做任何事情。在IP交换网络中支持QoS的技术之一是RSVP。正如在标记交换和ARIS章节中所归纳的，可以在RSVP RESV消息中简单的增加标签或者标记对象，以便QoS和交换式通路状态能够同步建立。然而，这种解决方案局限在标记和ARIS的范畴之内，无法支持端到端的业务。

采用RSVP并没有解决经典IP解决方案集缺乏QoS支持的问题^①。正如ATM上的经典IP（RFC1577/2225）和NHRP（RFC2332）的解决方案都是建立尽力而为UBR服务的ATM SVC直通路。为了消除这个限制，瑞士的研究者开发了一种解决方案，其中对等IP应用能够在彼此之间建立特定QoS的虚连接^②。该解决方案在RFC2170中定义，称作AREQUIPA，其含义为ATM上应用请求的IP（Application Requested IP over ATM）。

开发AREQUIPA的动机很简单：开发在ATM上能够支持IP QoS的解决方案。AREQUIPA可以看作经典IP解决方案集的扩展。如图13-9所示，它使得两个主机和具体来说两个应用彼此之间建立具有QoS的ATM SVC。这是靠在套接字层次的应用去请求具有QoS的基于套接字

① 在第3章描述的集成服务模型提出之前，IP对QoS无能为力。因此把QoS工程化到经典的IP解决方案集并没有实际的基础。但是在RFC1755（UNI3.1）和RFC（UNI4.0）所定义的ATM上的IP的信令规范支持建立具有QoS的SVC。

② <http://icawww.epfl.ch/>。

的连接或者期望这样的连接将会建立来完成的。这涉及定义几个新的应用所调用的 API 功能 (RFC2170 描述)。一旦进行了 API 调用并且两个主机之间的 SVC (具有 QoS) 已经建立, 应用的数据将在这个特定的 SVC 上转发, 其他尽力而为的业务流将在标准的尽力而为的连接上传送。

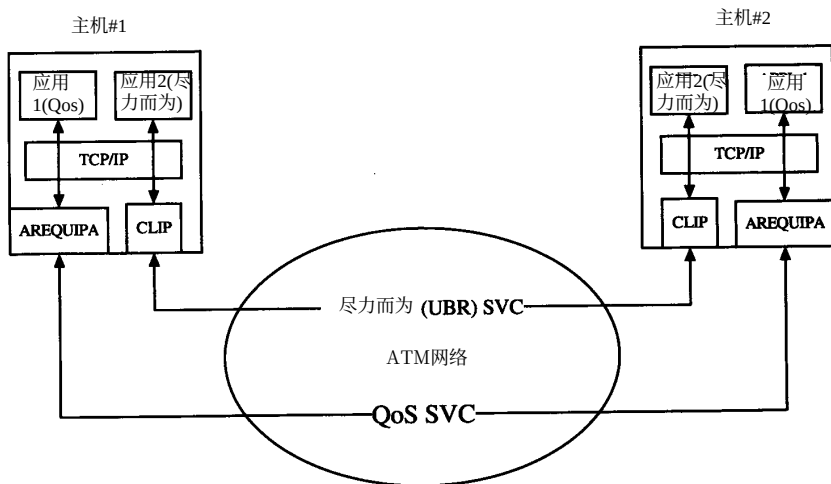


图13-9 AREQUIPA

其开发者指出, AREQUIPA 采用 ATM QoS 作为传送实时 Web 内容的技术是有用的。这种情形下, 用元信息扩展了 Web 的文档, 元信息描述了 ATM 业务以及相应地在 ATM 网络上传送的 QoS 要求。希望访问这些文档的 Web 浏览者在 HTTP 请求中包含其 ATM 地址以及可能的套接字号。服务器然后可以使用 AREQUIPA 传送文档。当然这种解决方案要求所有客户机和服务器都连接到同种 ATM 网络上。尽管这种情况可能不实际, 但是不远的未来将可能把 QoS (特殊服务) 与特定文档 (内容) 联系起来, 然后根据客户机的请求在网络上利用特殊服务 (QoS) 传送文档。

China-pub.com

下载

第三部分 新的相关工作

第14章 大型交换式网络上的路由

在上一章中，我们讨论了一组不同的 IP 交换解决方案。这些方案都结合了 IP 选路控制功能和高性能的 ATM 交换能力。它们要求在一个单一的设备上，即一个混合式的交换机 / 路由器上实现 IP 控制平面和 ATM 用户平面的综合，并需要引入一个专用的控制协议。这个协议将 IP 层的结构映射到一条交换路径中，而不管这些结构究竟是一段数据流的源端和目的端地址对还是简单的目的端前缀。此外，这些 IP 解决方案都基于对等模型，都保留了单一的 IP 地址空间和单一的 IP 选路协议。而且通过一些理论研究和物理实现证明了它们具有较好的性能。但是，鉴于这些技术尚未成熟，在任何现有大规模网络中实施这些技术还不是时候。

到目前为止，通过一个网络交换 IP 分组将会采用传统的叠加（overlay）模型来完成。运行 IP 选路协议（如 OSPF）的路由器将连接到一个帧中继或 ATM 交换网络上，而路由器之间则通过虚连接进行通信。不过，在获得交换性能的同时，也必须考虑到在网络规模和流量增加的过程中，出现的诸如 VC 容量和交换协议开销一些问题。因此，网络供应商们，特别是那些运行帧中继和 ATM 基础设施的运营商们，就必须处理好这样一个问题：即如何在一个大规模交换网络上最好地支持 IP 网络互通，从而使得网络性能达到最优并获得最小的控制处理开销，同时避免不必要的人工干预。

一种可能的解决方案是采用 IP 交换对等模型，不过在目前这个阶段实现它还为时过早，因为在实施过程中还有若干个未知因素需要考虑[⊖]。另一方面，仍有不少网络供应商出于对可运营性和运行费用的考虑，不想采用 IP 交换，他们对运营传统的交换网络非常在行。另外一个重要原因是他们目前仍依赖一些特定的交换技术来提供一组网络业务，比如帧中继和 ATM PVC 以及有可能提供的 SVC 业务就是这样的例子。为了支持 IP 业务，这些传统交换网络的供应商只是简单地寻找可以更好地利用他们交换基础设施的解决方案。为了支持现有和未来业务并最佳地利用网络资源，他们也许会采用一种更加动态地按照需求（on-demand）的资源分配优化方案。

前面我们已经讨论了颇有希望的一些解决方案。其中，第一种比较可行的方案是在路由器之间建立部分或全连的 PVC 网络。尽管这一方案尚不能进行动态管理，它还是实现了这样一个目标——为互连路由器建立一个类似租用线路的网络。另外一个方案，我们称之为子网模型，它将大型交换网络分解成单独的逻辑 IP 子网并在它们之间建立起路由。这一方案易于理解和操作，但是在子网间转发数据时却可能引入一次或多次的转接。第三种解决方案是第二种模型的扩展。它采用 NHRP 协议在不同子网的主机或路由器之间建立子网间的直通过路，从而避免了中间路由器对分组的处理。不过这一方案需要引入一个新的控制协议，而且所有路由器都必须支持 NHRP 并需要保留 IP/ATM 地址映射。

⊖ 如下一章所述，许多供应商专有的解决方案将最终汇集成一个标准。

有两种最新方案不需要采用额外的控制协议就可以把 IP 层业务流映射到第二层交换通路。它们扩展了现有的链路状态选路协议作为传播 IP/ATM 地址映射的途径。第一种方案是 OSPF 的扩展, 称作地址解析通告 (ARA)。一条 ARA 本质上是一条 IP/ATM 地址映射, 它由连接到 ATM 上的路由器采用标准扩散机制进行发布。在 OSPF 网络中连接到 ATM 上的所有其他路由器就可以将该信息存储在拓扑数据库中并据此计算路径。当一个源端路由器想要与一个目的端路由器通信时, 如果它预先知道目的端的 IP 地址和 ATM 地址, 就可以立即向目的端发起 SVC 建立请求。网络基础设施所要做的唯一变动就是更新路由器上的 OSPF。

第二种解决方案则扩展了另一个链路状态选路协议, PNNI。众所周知, PNNI 是一个通过 ATM 交换网络转发 SVC 建立请求的信令和选路协议。PNNI 采用传统链路状态扩散算法分发 ATM 寻址信息。扩展后的 PNNI 允许向网络发布 IP 层的信息 (如 IP 地址前缀、路由器 ID 等)。第一种扩展方案称作 PNNI 扩充选路 (PAR), 在连接到 ATM 上的路由器上运行 PNNI 以及任何其他动态 IP 选路协议 (如 OSPF)。PAR 允许路由器参与 PNNI 拓扑信息的分发过程。实际上, 连接到 ATM 上的路由器在一个或多个对等组的 PNNI 拓扑中成为节点。PAR 路由器采用 PNNI 扩散机制分发拓扑信息并发现连接到 ATM 网络上的非 ATM 设备和服务。

第二种 PNNI 的扩展方案称作代理 PAR (proxy-PAR), 它只对 PAR 稍做修改。连接到 ATM 上的路由器不运行 PNNI 而是启动一个客户程序, 通过一个运行在 ATM 交换机上的代理 PAR 服务器对路由器的 IP 和 ATM 信息进行注册, 然后采用 PNNI 在网络中发布这些信息。其他连接到 ATM 上的路由器如果需要的话, 可以从距离最近的代理 PAR ATM 交换机上重新获得这些信息。通过利用 PAR 和代理 PAR, IP 设备可以发现其他连接到 ATM 网络上的 IP 设备的位置和服务。代理 PAR 的一个应用是为连接到 ATM 上的路由器提供信息, 以建立与 ATM 上的经典 IP 叠加模型相符的 SVC 集合。

第三种 PNNI 的扩展方案是综合 IP 和 ATM 选路协议形成一个基于 PNNI 的选路协议。综合 PNNI (I-PNNI) 可以同时运行在 ATM 交换机、连接到 ATM 上的 IP 路由器和老式 IP 路由器上。在一个对等组中路由器和交换机都变成了节点, 彼此交换包含 ATM 和 IP 地址前缀在内的拓扑信息。该方法的目的是支持包含拓扑信息的单一选路协议, 为 SVC 选路和 IP 转发提供路径选择, 同时也可以减小叠加模型的复杂性。

14.1 问题

网络供应商如果想要实施叠加模型, 即在一个交换网络上运行 IP 选路, 那么就会遇到很多问题^①。首先, 需要对两个单独的选路协议、地址空间和拓扑结构进行配置、操作和维护。然而 IP 路由器只与临近的路由器交换路由表更新信息、维护 IP 拓扑结构数据库以及转发分组。除了逻辑链路 (点到点、广播 LAN) 之外它对底层的交换基础网络一无所知。类似地, 运行 ATM 协议的 ATM 交换机只与临近的 ATM 交换机交换路由表更新信息, 维护 ATM 拓扑结构数据库以及转发 SVC 请求。即使网络供应商不支持 SVC, 它们也必须处理 PVC 管理的诸多问题 (如配置、业务量监控和错误恢复)。

叠加模型的另外一个问题是连接到路由器上所需 VC 的数量。图 14-1 是一个由三个 ATM 交换机和八个路由器组成的网络。图 14-2 表示出 VC 全互连拓扑结构。由于在任意两个路由器之间都建立有虚连接, 所以这一特殊的逻辑配置方式为路由器之间的通信提供了最优性能。然

^① 在下面的讨论中, 交换网络将以 ATM 为例, 但同样适用于帧中继或其他 NBMA 类型的网络。

而，随着连接到 ATM 网络上路由器数目的增加，为保持全互连配置和优化性能所需要建立的 VC 的数目也会迅速增加。事实上，如果设 N 是路由器的数目，那么所需 VC 的数目是按 $O(N^2)$ 增加。在这个例子中，对 8 个路由器进行全互连需要建立 28 个点对点双向 VC 连接^①。这个数量的 VC 看上去还是容易管理的，对于中小型网络也很适合。但是如果 N 继续增大，那么所需 VC 的数目就会急剧增大。如果 $N=100$ ，那么所需 VC 的数目为 4950。如果 $N=1000$ ，那么所需 VC 的数目为 499,500。而且，如果这些网络需要支持多种服务，那么在任意两个路由器之间可能还会需要不止一条 VC 连接。

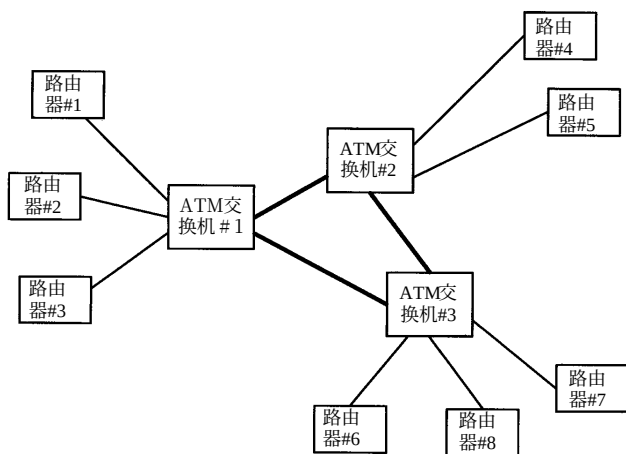


图14-1 物理拓扑

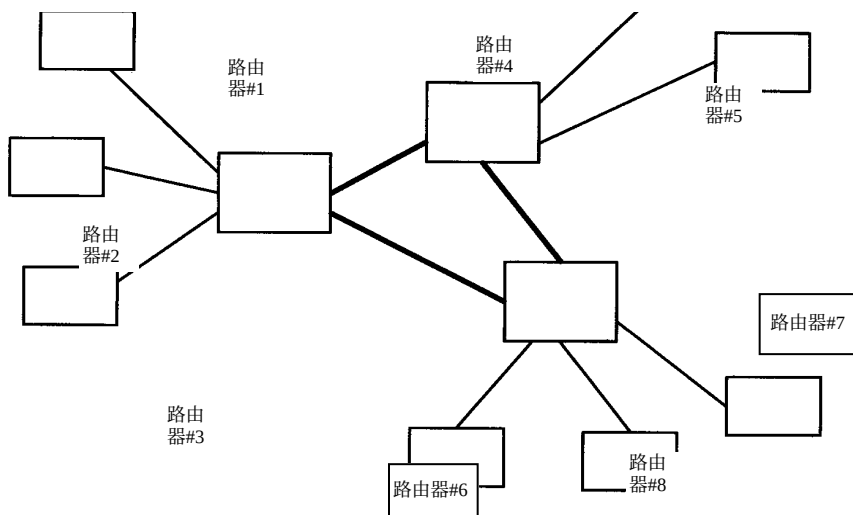


图14-2 全互连逻辑拓扑

随着路由器数目的增加，相应地 VC 数目的增加会受到 VC 预定（provisioning）和容量的限制。VC 预定需要管理多个边缘路由器之间很大数量的 VC。大多数情况下，VC 是通过手工配置的。当然，那些需要承载大量 VC 的基础交换网络也存在这个问题。因此，需要消耗一

^① 如果 $N=8$ ，那么 $(N*(N-1))/2=28$ 。

些人力、物力以保证一个稳定工作的 VC 网络环境。此外，网络需要为每一个连接分配标签、内存、控制块、缓存器等网络资源，而网络资源的有限性限制了 ATM 交换机只能支持有限数目的交换式连接。

在 ATM 上实施 IP 选路叠加模型时需要考虑的另外一个重要问题是 IP 选路协议的可扩展性。尤其是当路由器还必须提供和处理大量表示每条 VC 链路的信息的时候。如图 14-2 是一个全互连逻辑网络拓扑结构，每一个路由器需要与其他连接到网络中的路由器保持邻接关系。选路邻接可以使路由器进行路由表交换和更新。转发邻接可以作为去往目的端网络的下一跳。如果路由器运行一个距离向量选路协议，例如 RIP 或 BGP，那么一个路由器需要处理的信息与选路上相邻接的路由器的数目以及选路表的项目成正比。换句话说，是线性 $O(\text{路由的数目} * \text{选路邻接的数目})$ 。如果路由器运行的是链路状态选路协议，就会得到链路状态信息（LSA）并在网络中扩散。如果 VC 的数目很大，就会产生很大数目的链路以及由每个与 ATM 相连路由器提供的 LSA 信息。依照 OSPF 协议，链路状态数据库的大小以及计算路由的开销与链路（即 VC）的数目成正比。此外，大部分 LSA 可能包含复制信息，因此会给每个路由器增加额外的处理负担。邻接路由器可能会重复传播从其他路由器接收到的 LSA 信息。

要想解决这些问题可以采取不同的方法。首先，选路协议和 VC 的可扩展性问题可以通过减小路由器间 VC 的数量来解决。小数目的 VC 明显地降低了网络中交换资源的消耗并减少邻接选路的数目。要实现这一目的，可以建立一个部分互连的 VC 网络或者将一个交换网络分解成若干个独立子网并利用路由器连接这些子网。但是尽管这些方案增强了可扩展性，它们几乎都在数据转发路径上引入额外的路由转接，从而引入额外的延迟和欠优化的性能。子网模型的性能可以通过利用 NHRP 建立子网间的直通路程得到提高。不过，这一方法需要额外的地址解析控制协议并要求在选路路径上存在 NHRP 服务器。

另外一个方案是采用 MPT-PT 连接，我们在第 12 章中介绍 ARIS 协议时引入了这个概念。应用 MPT-PT 连接并不只限于 ARIS 协议，包含独立 IP 和 ATM 控制和数据平面的叠加模型也可以应用这一方案。支持 MPT-PT 连接的网络只需要 $O(N)$ 个连接，而非 $O(N^2)$ 个点点对点连接，其中 N 是路由器的数目。这一方案的优点在于它只用少量的 VC 连接就提供了路由器间的完全互连业务。图 14-3 示例出随着路由器数目的增长节省下来的 VC 数目。

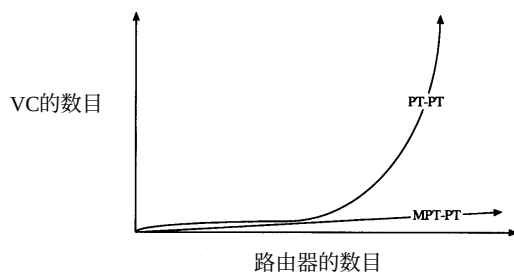


图14-3 MPT-PT和PT-PT环境下的VC消耗

ARIS是一个基于对等的IP交换控制协议，它支持建立 MPT-PT 连接。叠加模型需要连接建立控制协议支持这一功能。在 ATM 网络环境中，UNI 和 PNNI 信令和选路协议联合完成连接建立过程。尽管目前它们还只能支持点到点和点到多点的交换式连接，通过对 UNI/PNNI 信令进行扩展就有可能支持 MPT-PT 连接的建立。如图 14-4 所示，一种可能的技术是利用标准根启动方式的点到多点信令来建立 MPT-PT 连接，不同的是在连接建立阶段会指示出数据流的方向是从叶子到根，而不是通常情况下的从根到叶子。于是，ADD PARTY 消息将依旧从根发起，并为一个点到多点连接添加叶子。但是，真实数据是从叶子节点上行流向根。这就要求 ATM 交换机具有支持 VC 合并的能力并且 MPT-PT 只支持 UBR 连接。鉴于可扩展性带来的好处，ATM

论坛可能会增强UNI/PNNI信令和选路功能以支持MPT-PT连接。

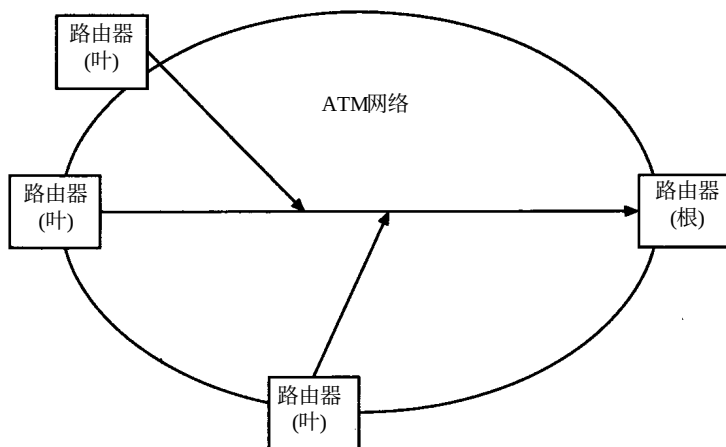


图14-4 ATM MPT-PT连接

为了建立一个可扩展和高效的IP/ATM叠加网络，只依靠优化网络配置或增强VC管理功能还不是一个完全的解决方案。增值业务如VPN以及带宽的按需分配就需要快速可靠地配置网络资源。因此需要采用ATM基础设施的SVC能力来动态建立承载IP分组的网络。然而，在路由器之间建立SVC时需要在呼叫建立之前绑定目的端IP地址和相关联的ATM地址。目前，基于服务器的技术例如NHRP提供这种绑定。但是如果网络环境只包含路由器和交换机，也许该考虑在已有的选路协议中添加IP与ATM的映射。

14.2 OSPF扩展

OSPF是一个链路状态选路协议，用于建立和维护网络拓扑结构数据库。根据数据库的内容，路由器计算转发分组到目的地址的路由表。OSPF是可扩展的（scalable），它能够迅速传播拓扑结构的变化，并减少业务量的控制开销。同时，OSPF也是可拓展的（extensible），在引入新功能的同时，保持OSPF协议的向后兼容性。在ATM环境下扩展OSPF的一种可能方案是通过网络同时传播链路层和IP层信息。换句话说，对OSPF进行扩展，使IP和相关ATM地址信息通过网络进行分发。因此，OSPF可以看作是传递IP/ATM地址映射的工具。得到这些信息，在路由器之间就可以根据需要建立SVC。

14.2.1 Opaque LSA

在RFC 2370中^①，描述了一种链路状态新类型，即Opaque LSA。它是一个通用化机制，可以包容任何有关OSPF的扩展。Opaque LSA采用标准扩散算法，允许新信息通过一个OSPF域进行传播并存储到路由器的拓扑数据库中。OSPF可以直接利用包含在opaque LSA中的信息，而且一些想要用OSPF分发信息的应用也可以利用这些信息。

在OSPF Hello协议、数据库描述分组和LSA的选项字段中，OSPF路由器采用一个比特位

^① 有关OSPF opaque LSA和ARA可以在位于<http://www.ietf.org/html.charters/ospf-charter.html>的相关Internet草案或RFC中找到。

表示支持 opaque LSA 的能力。这使得 OSPF 路由器能够确认它们的邻接路由器是否支持处理 opaque LSA。同时，它还包含一个扩散范围 (flooding scope) 的功能，这一功能定义了一个特定 opaque LSA 的拓扑分发范围。这用于限制由 opaque LSA 承载去往 OSPF 一个特定端口的信息。

如图14-5所示，一个 opaque LSA 包含一个标准的 20 个字节的 LSA 头和可变长度的数据信息。它支持三种不同类型的链路状态，它们的基本扩散范围分别为：

链路状态		选项	链路状态类型
Opaque 类型	Opaque ID		
通告路由器			
链路状态序列号			
链路状态校验和		长度	
Opaque 数据			

图14-5 OSPF opaque LSA 格式

- 链路状态类型 9。定义了本地链路范围。这一类型可以只在同一子网的路由器间传播信息。
- 链路状态类型 10。定义了本地区域范围。这一类型可以只在同一区域的路由器间传播信息。
- 链路状态类型 11。定义了自动系统区域范围。这一类型可以通过整个 OSPF AS 进行扩散。

Opaque LSA 头的链路状态 ID 包含一个 8 比特的 opaque 类型和一个 24 比特的 opaque ID。这意味着最多可以定义 256 种不同的 opaque 类型。根据定义，IETF/OSPF 预留了数值 0 到 127、数值 128 到 255 用于实验用途。

14.2.2 地址解析通告

OSPF 分发 ATM 地址信息需要两个条件。第一，opaque LSA 必须作为一个在 OSPF 网络中插入和扩散新信息的端口。第二，真实的 IP/ATM 地址解析信息/映射必须包含路由器的 IP 和 ATM 地址或者一个或多个 IP 地址前缀以及一个相关联的 ATM 地址。ARA 定义了地址映射类型和格式，它们装载在 opaque LSA 中并被分发到各路由器中。尤其，ARA 允许路由器分发 IP/ATM 地址映射（称作链路层关联）并允许路由器决定在 OSPF 选路域内与一个特定目的网络最近的链路层关联。ARA 可以用来分发路由器的链路层关联（称作 R-ARA）或者是可以通过路由器到达网络的链路层关联（称作 N-ARA）。

ARA 机制的基本目标是利用 OSPF 为路由器迅速可靠地提供 IP/ATM 地址映射，从而建立起通过 ATM 网络内部及地区间的直通 SVC。

OSPF ARA 选项包含下列特性：

- ARA 信息按链路状态类型 10 opaque LSA 传送。ARA 选项支持四种不同的 opaque 类型 (在 opaque LSA 头中指定 opaque 类型)：内部区域路由器、内部区域网络、区域间网络和区域间 ASBR 链路层关联。
- 只有连接到 ATM 的 OSPF 路由器才能参与 ARA 过程。
- 因为 ARA 在一个 OSPF 区域内采用可靠的扩散机制进行分发，所以它不受分组丢失的影响。而分组丢失会对查询服务器 (query-to-server) 机制如 NHRP 产生很大的影响。
- 支持建立点到点、点到多点和多点到点的直通路径 (SVC)。

- 可以与现有地址解析机制如 NHRP/MPOA 共同工作。

- 提供将一组路由器关联到一个逻辑网络的能力。通过在 ARA 中包含一个逻辑网络 ID，多个不相交的 OSPF 网络（如 VPN）可以共存于一个单一的 ATM 网络中。OSPF 将从路由表中去除任何逻辑网络 ID 与路由器上预设值不符的链路层关联。

图14-6给出在单一区域中 OSPF ARA 过程的一个例子。路由器 1 到 4 支持 ARA 选项，并通过如图所示的一个部分互连前置的 PVC 网络连接到 ATM 网络上。路由器 5 也连接到 ATM 网络上，但只支持 opaque LSA 功能^①。路由器 6 通过一个非 ATM 串行链路连接到路由器 4 上，但不支持 opaque LSA 功能。在这个例子中，具有 ARA 功能的路由器产生并分发包含在 ARA 中的链路层关联信息。由具有 ARA 功能的路由器分发的链路层信息允许这些路由器初始化 SVC。例如，如果路由器 1 需要转发分组到网络 N2，它可以建立一个直接连到路由器 2 的 SVC。或者如果路由器需要转发分组到 N6，它可以建立一个直接连到路由器 4 的 SVC。在这两种情况下，路由器 1 接收从路由器 2 和路由器 4 发出的包含链路层信息，完成 ARAs 过程。

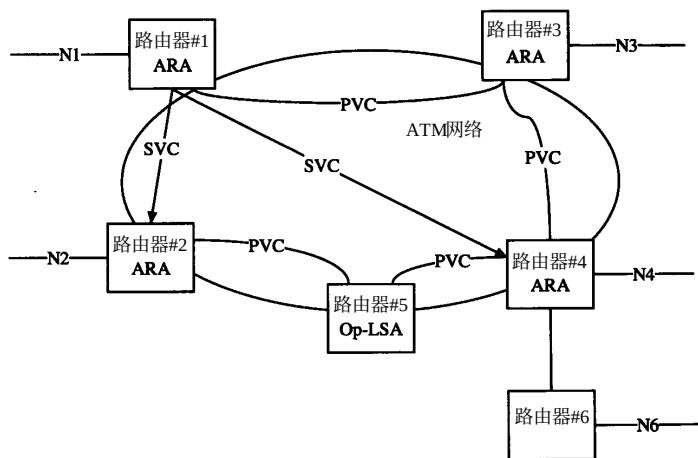


图14-6 OSPF ARA功能

图14-7为OSPF ARA的分组结构。它包含一个 opaque LSA 头和 ARA 选项。四种 ARA 选项类型均包含顶点（vertex）关联。顶点关联包含一个链路服务类型（ATM 点对点、点对多点、多点对点）、IntServ 服务名称（用于映射 IntServ 到 ATM）、管理加权、长度、逻辑网络 ID（用于 VPNs）和解析信息。解析信息为加入已有点对多点的 VC 又定义了 ATM 地址或一个 LIJ 呼叫标识符。

设计 OSPF ARA 的目的是为了在 ATM 网络上建立 OSPF 选路域，并在路由器之间建立直通路径。OSPF ARA 采用现有可靠的 OSPF 拓扑分发机制为所有具有 ARA 能力的路由器提供 IP/ATM 地址映射。因此并不要求一个入口路由器查询地址解析服务器的恢复信息。因为路由器可以在 OSPF 网络中迅速地与其他路由器建立 SVC，因此它不负责建立一个完全互连的 VC 网络。如果需要更多的 VC 增强连接性或支持更大的业务量负载，就可能要建立 SVC。支持更少的选路邻接也可以减少选路协议业务量的开销。

值得注意的是，只有路由器才有 OSPF ARA 功能。ATM 交换机没有该功能。真实 ATM 网

^① 路由器 5 传播 ARA 信息需要这个功能。

络拓扑结构对路由器透明。事实上，底层 ATM基础设施仍作为一个虚拟点对点链路（如 RFC1483）、广播LAN（如LANE）或根据网络配置的其他结构。除了扩展了 OSPF以支持更好的VC管理功能之外，这个解决方案仍保持了 ATM上的经典IP 叠加模型。

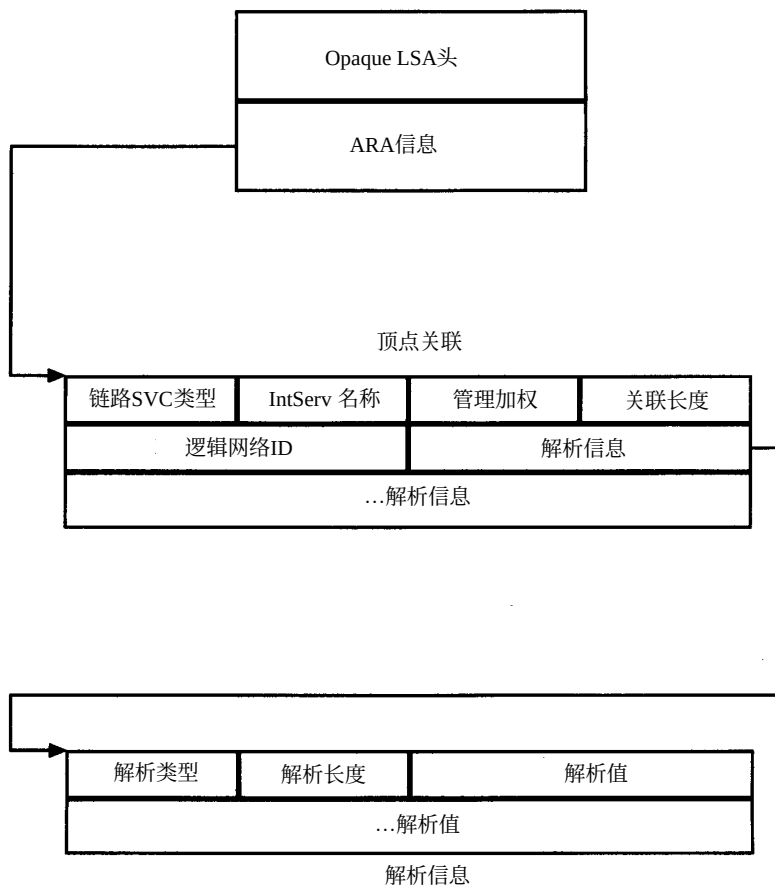


图14-7 OSPF ARA格式

14.3 PNNI扩展

与扩展OSPF以承载ATM信息类似，扩展 PNNI以承载IP层信息也是有可能的。其实，PNNI第一个阶段就是一个分等级的链路状态选路协议，适用于专用的 ATM交换机网络。PNNI的基本功能是向其他 ATM交换机分发拓扑信息，计算源端选路的路径，并通过网络向目的端传送SVC建立请求。同时它能够基于业务量负载的 QoS特性和网络的可用容量来选择路径，并在连接建立过程中提供一个动态选路功能。因为需要传输大量 IP业务量的ATM网络（企业网、ISP等）正在逐渐配置PNNI，因此有必要考察将IP综合进PNNI选路协议的各种可能方案。

14.3.1 PNNI扩充选路

扩展PNNI选路功能支持IP的第一种方案是PNNI扩充选路（PAR）^①。在一个PAR网络中，

① PAR是ATM论坛的PNNI子工作组的一个工作项目。

所有路由器都运行传统 IP 选路协议如 OSPF 以建立并管理 IP 转发表。所有 ATM 交换机都运行 PNNI 选路协议。PAR 使得连接到 ATM 的路由器也参与 PNNI 拓扑分发过程。特别的，如果一个 PAR 路由器要运行一个 PNNI 选路实例，它先广播它的存在和能力（如选路协议、路由器 ID、IP 地址等），然后了解其他与 ATM 相连的路由器所支持的地址范围和服务。通过过滤从 PNNI 拓扑分发过程中接收的信息，一个 PAR 路由器提取建立 SVC 所需的 IP 和 ATM 信息。例如，一个 OSPF 路由器可以通过 PAR 了解其他属于同一 OSPF 区域的连接到 ATM 网络上的路由器。PAR 允许具有普通 IP 配置参数的路由器建立并保持与其他路由器的 SVC，这样就简化了网络的配置和管理。实际上，PAR 的主要贡献是它极大地减少了配置 PVC 的需求。

PAR 也提供在一个单一 ATM 网络上支持多个 VPN 的能力。它采用一个 VPN id 值标记所有与 PAR 相关的信息。如果 PAR 更新的数值与路由器的 VPN id 值不符，就滤掉这个值。PAR 路由器也可以过滤 IP 层信息如 IP 地址和协议标志。这保证了 PAR 路由器只处理与操作有关的信息。

如图 14-8 是一个 PAR 网络。路由器 1 到 10 支持 OSPF。路由器 4 到 7，连接到 ATM 网络上，与 ATM 交换机都支持 PNNI。PAR 路由器参与 PNNI 拓扑管理过程。它们交换 PNNI Hello 消息然后加入对等组。它们与临近的 PNNI 节点交换 PNNI 拓扑状态元素（PTSE）并建立和维护一个 PNNI 拓扑数据库，也在 PTSE 中插入非 ATM 的路由器专用信息，并传送给对等组。运行标准 PNNI 的 ATM 交换机会忽略任何路由器的专用信息并透明地把它传送给其他节点。这使得与 ATM 相连的交换机可以采用 PNNI 向其他路由器分发 IP 专用信息。因为 PAR 路由器维护一个 PNNI 数据库，它们完全掌握一个 ATM 网络拓扑的真实情况。PAR 更新信息也可以在一个分等级的 PNNI 网络中不同级别进行分发。

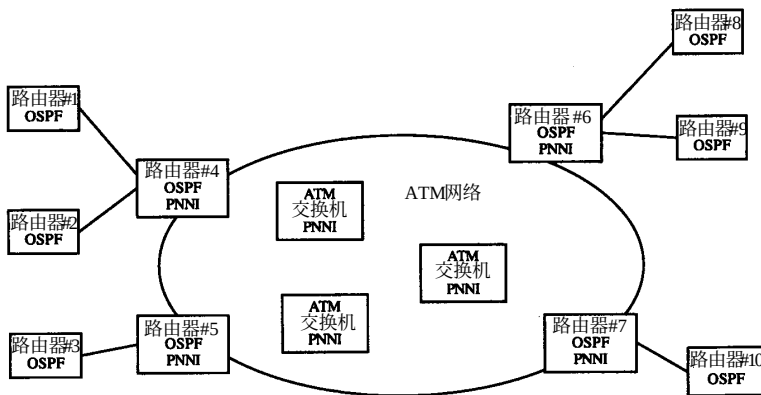


图14-8 PAR网络

由PAR路由器维护的PNNI拓扑数据库也包含其他 PAR 路由器支持的地址和服务。这些共享普通 IP 配置信息的路由器能够建立或在网络初始化时自引导一个部分互连的 SVC 网络。这避免了预置一组 PVC 的要求（如 OSPF ARA 的情况）。例如，属于同一个 OSPF 区域的路由器可以应用 PAR 发现其他路由器然后建立一个完全互连的 SVC 网络。

除了透明地传播 PNNI PTSE 之外，PAR 不要求对运行 PNNI 的 ATM 交换机作任何变动。支持 PAR 的一个设备必须能够产生和解释与 PAR 相关的信息。需要注意的是，运行在路由器上的 PNNI 以及更通用的 PAR 解决方案本身并不指定一旦路由器收到 PAR 更新信息该做什么。这是一个可以由网络供应商解决的配置问题。最后，它仍是一个路由器上的 IP 选路协议，决定

去往一个特定目的地的最佳下一跳。PAR只为连接到ATM网络上的路由器提供了一个在路由器上支持有关功能和服务的一个较好的方案。

14.3.2 代理PAR

PAR的一个缺点是它要求在每一个路由器上都运行 PNNI。为了降低这一要求,在传播相似的自动检测和配置功能时,定义了一个简单的 PAR版本,称作代理PAR[⊖]。代理PAR是PAR的最小版本,允许与ATM相连的路由器利用PNNI拓扑分发来发现其他路由器支持的服务[⊕]。代理PAR包含一个运行在路由器上的客户组件、一个运行在直接连到ATM上的服务器组件以及一系列协议消息格式和交换。客户与服务器注册IP层信息。注册信息通过PNNI域分发。一个客户为获得有关的IP层信息可能查询一个服务器。类似PAR,代理PAR的目的是允许非ATM设备为服务的注册和发现采用PNNI扩散机制。

图14-9示例出一个运行代理PAR的网络的例子。路由器1到5连接到运行PNNI的ATM网络上。每一个路由器被单独地配置成同一OSPF区域的部分。每一个路由器也运行代理PAR客户。此外,这些直接连接到路由器的交换机运行代理PAR服务器。在网络初始化时,服务器上的代理PAR客户会联系它们的服务器方并注册IP层信息如OSPF区域ID、路由器ID等等。接下来这些信息通过采用PNNI扩散机制的ATM网络进行分发。然后路由器上的每一个代理PAR客户查询交换机上的代理PAR服务器,获得任何路由器进行操作所需的IP层信息。在这个例子中,查询提供给每个路由器属于路由器上同一OSPF区域的其他路由器的信息(包括ATM地址)。于是建立SVC对路由器来说就是个简单的过程。结果是一个自动配置的SVC叠加模型,它反映了OSPF区域的逻辑拓扑结构。在这个例子中不需要预置PVC,同一区域OSPF路由器的发现和SVC建立都是自动的。此外,每一个路由器不得不向服务器单向查询相关PAR信息,不像PAR那样自动传播。

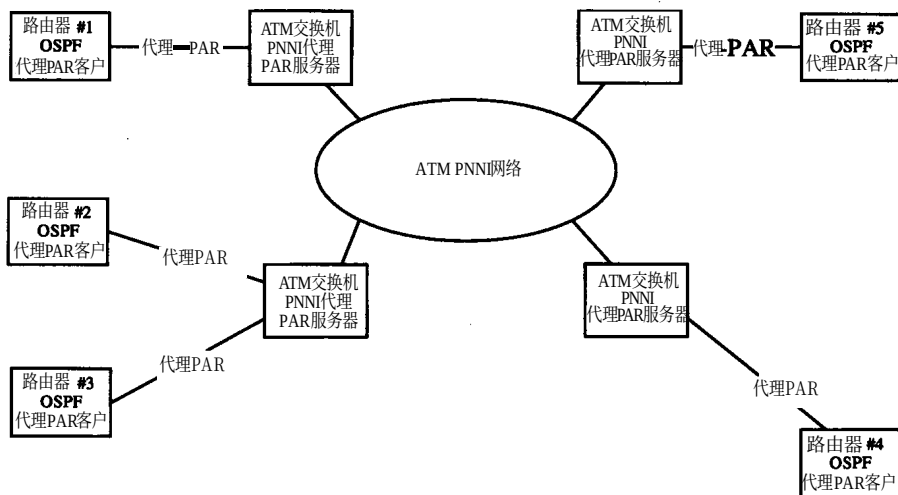


图14-9 代理PAR网络

⊖ 代理PAR是ATM论坛从事的PAR规范的一部分。

⊕ Droz, P., and Przygienda, T., "Proxy PNNI Augmented Routing (Proxy PAR)," Proceedings from the 1998 First International Conference on ATM ICATM'98, June 1998, Colmar, France.

代理PAR规范定义了三种特定的功能协议。第一种是 Hello协议，它是 PNNI阶段I规范中定义的 PNNI Hello协议的基本版本。Hello协议首先建立代理 PAR 客户和服务器的邻接关系，然后就可以交换接下来的注册和查询信息。服务器也通过 Hello协议通知客户服务器它们分配给数据的生命周期。

代理PAR支持的第二种协议是注册协议。一个代理 PAR 客户采用这一协议来注册它支持的特定IP协议和服务。与这个信息相关的是在 PNNI系列内的一个 ATM地址和一个成员资格范围。后者允许与 PAR 相关的特定信息分发到 PNNI 系列内的一定级别。例如，一个路由器可能注册OSPF服务的目的是为了在相同 PNNI 对等组中的其他 OSPF路由器分发信息，或者注册BGP服务的目的是为了在高级别对等组中的其他BGP路由器分发信息。

图14-10表示一个典型代理 PAR 服务注册分组的格式。它包含一个 PNNI头、ATM地址、序列号和成员资格范围。接下来的可能是一个或多个服务定义信息组（IG）。在这个例子中，IPv4业务定义的IG专用于IPv4业务。它包含一个 IPv4地址、地址掩码和业务掩码。业务掩码指示哪个特殊业务（OSPF、BGP、PIM等）被注册。然后每个注册信息通过基于成员资格范围的 PNNI网络传播。它不能直接传递到其他客户而必须精确地查询。如果一个客户想要注册新的信息，任何先前注册的信息会被无条件地覆盖。因此在客户和服务器的间不保持涉及到先前注册信息的状态。

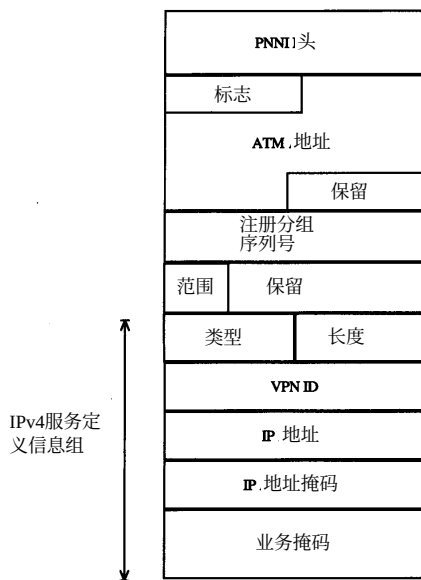


图14-10 代理PAR 服务注册分组

代理PAR支持的第三种协议是查询协议。一个代理 PAR 客户采用查询协议从服务器重新获得与 PAR 相关的信息。客户并不更新所有信息，而是基于成员资格范围查询 IP 地址、地址掩码、VPN ID 和服务。这使得代理 PAR 客户只查询和接收在 ATM 网络上与正确配置和操作有关的信息。客户定期地查询服务器以更新任何可能过期的信息。图 14-11 表示出代理 PAR 的客户和服务器的间的一般消息流。

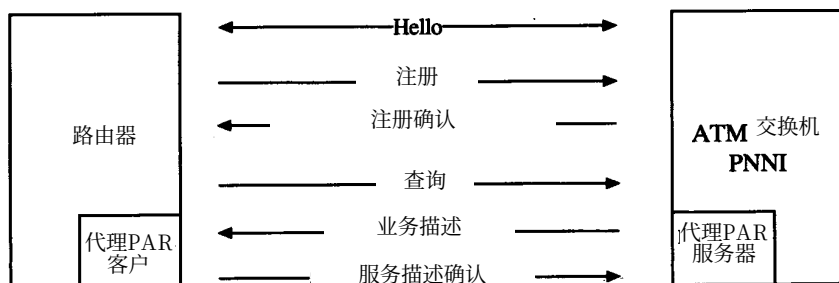


图14-11 代理PAR消息流

PNNI的可扩展性使得 PAR和代理 PAR路由器只注册和访问相关信息，这些信息只与在 ATM网络上应用的 IP层服务（或协议）有关。表 14-1列出 PAR和代理 PAR支持的一些最常用的服务。在某些情况下，服务信息可能会附带一些附加的特定协议信息，用于完成网络配置。例如，支持 OSPF服务的路由器包含一个单独的 IG，它描述了区域 ID、指定路由器（DR）选举优先级和 OSPF接口类型。另一个例子是 BGP，它还需要一个额外的 IG传送 AS编号、BGP id和路由器的反馈信息。此外，一个具有系统功能的 IG用于注册专用和实验信息。

表14-1 PAR/代理PAR服务

服 务	附 加 信 息	服 务	附 加 信 息
OSPF	有	PIM-DM	无
RIP	无	MARS	无
BGP4	有	NHRP	无
MOSPF	有	ATMARP	无
DVMRP	无	DHCP	无
PIM-SM	有	DNS	有

代理 PAR在客户端对内存的要求和处理相对不多，这是开发人员进行代理 PAR设计的初衷。但是，这暗示客户不能存储和要求大量的数据。再下一步是采用 PNNI扩散算法为一群小数量的稳定客户分发服务信息，而不考虑环境频繁变化的情况。不过，它提供了一个有效、无服务器和自由广播的机制，用于广播和发现与 ATM相连，共享 IP配置信息的设备的存在。总体上讲，代理 PAR和 PAR使得这些组件引导了一个 SVC叠加配置。它完全兼容现有的运行在叠加模型中的 IP/ATM地址解析机制（如 ATMARP、NHRP等）。甚至可以同时使用代理 PAR和 OSPF ARA功能。前者自动发现和连接属于同一区域的路由器，后者用于分发 IP/ATM地址映射。

14.3.3 综合的PNNI

扩展 PNNI的第三种方案是综合 IP和 ATM选路，称作综合的 PNNI（I-PNNI）。I-PNNI与 OSPF ARA方案相反，不是采用 IP选路协议分发 ATM地址，而是应用 ATM选路协议（PNNI）分发 IP地址。特别地，I-PNNI巩固了 IP和 ATM选路协议形成基于 PNNI的一个单一协议。实际上它只是在 PNNI上加入承载 IP选路信息的 IG。I-PNNI可以运行在 ATM交换机、与 ATM相连的路由器、非 ATM路由器和主机上。I-PNNI完全兼容使用 PNNI的 ATM交换机。对于连接到传统介质如以太网或帧中继的路由器来说，I-PNNI取代了动态 IP选路协议如 OSPF或 BGP。

从一个路由器的角度，I-PNNI只是另一种链路状态选路协议，它用于在网络中分发 IP地址前缀并计算跳转路径。就 ATM来说，I-PNNI只是一个为传播非 ATM数据而有所扩展的基本 PNNI，交换机会简单地忽略掉这些数据。运行 I-PNNI的路由器和交换机维护包含路由器和交换机的整个网络的一个单一的拓扑数据库。与叠加模型的传统双协议方案相比，在一个混合的路由器和交换机网络中，运行一个单一的选路协议的主要优点是它消除了选路协议的复制并极大地降低了选路邻接的数目。这是由于网络中的任何节点，一个路由器或是交换机，只与它直接相邻的节点形成一个选路邻接关系。另一个重要的优点是可以优化 IP选路，这是由于路由器对真实网络（包含路由器、ATM交换机和连接链路）的拓扑已经有了全面的了解。

图14-12给出一个二层网络，它包含了运行 I-PNNI的路由器和交换机。在对等组 C中有一个路由服务器和边缘设备。一个典型的路由服务器（如 MPOA服务器）运行一个 IP选路协议，

在这个例子中是I-PNNI。边缘设备没有运行I-PNNI而是运行一个标准的IP/ATM客户协议（如LANE或MPOA）。

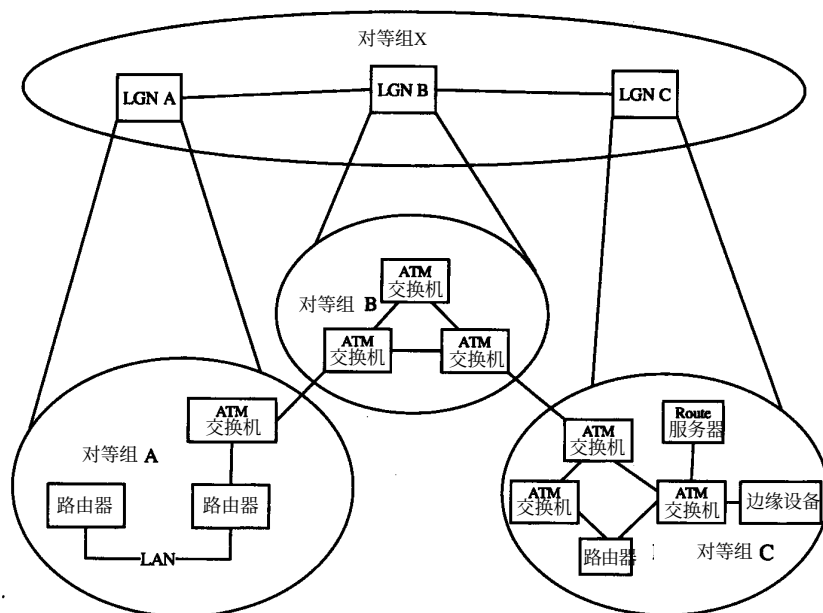


图14-12 I-PNNI网络

I-PNNI的功能特点包括：

- 所有路由器、ATM交换机和路由服务器都运行一个单一的选路协议，即 I-PNNI。
- 路由器和交换机加入进对等组。一个对等组能够包含所有 ATM交换机、所有老式路由器或二者的混合。在一个混合的环境中，对等组领导者（PGL）必须支持I-PNNI，目的是在对等组之外通告IP的可达性。
- 在路由器之间、交换机之间、路由器与交换机之间交换 PNNI拓扑通告（PTSE）。
- IP和ATM地址在单独的PNNI PTSE中通告。因此，尽管I-PNNI综合了IP和ATM选路，它仍然支持和维护二个分开的地址空间。
- 可以通过以下三种方式通告IP地址的可达性：
 - 直接（Direct）通过通告服务器直接到达这个IP地址。
 - 查询（Query）通过通告服务器可以到达这个IP地址，但如果发出请求的IP实体查询了这个地址就可能会找到更好的途径。
 - 关心（In-care-of）IP地址属于第三方。它被路由器用来通告一个知名主机或子网的IP地址和关联的ATM地址。不需要用NHRP进行查询。
- 一个对等组的成员（路由器和交换机）建立和维护一个单一的、通用拓扑数据库。
- 由于路由器不负责转发SVC请求，路由器只作为“受限制的转运”节点。然而，路由器能够利用标准PNNI选路技术（如建立DTL）来启动和终结SVC请求。
- 对于连接到具有广播能力的LAN（如以太网）且运行I-PNNI的路由器来说，它们借用了

OSPF设计路由器和伪节点的概念来表示在拓扑数据库中的 LAN。

- 可以产生一个多级别的选路等级。
- 当有不完全信息时，I-PNNI节点需要用NHRP解决IP/ATM地址映射。例如，一个选路等级将只在对等组间提供汇总的 ATM寻址信息或者一个运行 I-PNNI的路由服务器不向个别主机通告它支持的所有 ATM主机的路由。

更多的信息

R.Callon et al., “ ATM Routing Options, ” Connexions, March 1996.
[http:// www.zurich.ibm. com /%7Edro/proxy/](http://www.zurich.ibm.com/%7Edro/proxy/).

第15章 多协议标签交换

很明显，工业界如 Ipsilon、Cisco、IBM 等对 IP 交换解决方案给予了很大的关注，这表明 IP 交换是一个非常重要的技术。网络供应商们希望开发出能够显著提高基于 IP 连网性能和灵活性的新平台。用户对这一技术也充满兴趣，因为它承诺可以减轻由于路由器技术的过负荷和业务量的指数增长而造成网络容量上的瓶颈。许多网络供应商们把它看作是更有效利用帧中继和 ATM 网络的一个途径。工业界权威人士在考察了 IP 交换技术后对它的前途持有谨慎的乐观态度^①。玩世不恭地讲，IP 中心主义者实际上已经采用了 ATM 但是没有叫 ATM 而已。

到 1996 年秋季出现了很多不同 IP 交换方案，它们分别由新兴公司（如 Ipsilon）和计算机业界公司（如 Cisco、IBM）提出。其中一些是流驱动方案，它们采用基于硬件的 ATM 交换技术为端到端的主机和业务流提供改善的性能和更多的带宽。其他则是拓扑驱动方案，它们基于路由器和选路协议来提高业务流的吞吐量。总的来说，这些不同的方案都采用一个简单的标签交换方式。所有这些方案都意识到将选路和交换综合起来具有极大的优点。选路当然是 Internet 的要求所在，而交换则提供便宜的、高容量的分组/信元转发硬件。但是，这些方案虽然引起了极大的轰动，但它们面临很多实际的问题而举步为艰。

1996 年 12 月份在 MIT 举办了一个关于标记交换的 BOF 会议。之后，IETF 成立了一个从事综合选路和交换问题研究的工作组^②。事实上，它的诞生是有其技术、政治以及其他方面的必然性的，并以多协议标签交换（MPLS）来命名^③。这个名称是基于这样一个事实，即简单标签交换是作为底层转发机制。名称中的多协议（multiprotocol）意味着除了 IP 之外还支持多种网络层协议。在体系结构上这是对的，但实际上网络层协议会将只限于 IPv4 和 IPv6。

MPLS 工作组的主要目标是开发一个综合选路和交换的标准。特别地，MPLS 将合并网络层选路和标签交换而形成一个单一的解决方案，它有如下的优点：

- 改善选路的性能和成本。
- 提高传统叠加模型选路的扩展性。
- 引进和实施新业务时更具灵活性。

上述最后一点对网络供应商具有特殊的重要性。为了吸引更多的客户并产生更大的利润，网络供应商常常需要发布新业务和经过改善的业务。考虑到 IP 交换和 ATM 的密切关系，MPLS 是非常有希望的一种解决方案，ATM 的应答式业务量管理和 QoS 能力给它带来了好处。

工业界在从事 MPLS 及其推荐方案的工作时，表现出了广泛的兴趣和参与意识。由此，MPLS 包含了许多熟悉的 IP 交换概念。一个 MPLS 网络采用标准分组处理方式对第三层的分组进行转发，采用标签交换对第二层分组进行交换。它基于对等模型，即每一个 MPLS 设备运行一个单一的 IP 选路协议，交换选路表更新信息并维护一个拓扑结构和一个地址空间。它把转发等价类（FEC）映射成标签。为了在被称作标签交换服务器（LSR）的 MPLS 设备间分配和

① 对任何新鲜或挑战性的事物做谨慎不承担责任但清楚的表述是标准工业界权威人士的行为。

② 这不只是把这些单独的各方团结起来的利他主义。或许更令人害怕的是某个解决方案可能会主宰其他方案。

③ <http://www.ietf.org/html.charters/mpls-charter.html>。

分发FEC/标签绑定信息，制定了一个标签分发协议（LDP）。一旦LDP完成它的任务，从入口到出口就建立了一条标签交换式路径。当分组进入网络，入口LSR检查分组头的多个字段来确定分组属于哪个FEC。如果存在一个FEC/标签关联，入口LSR就给分组贴上一个标签然后引导到合适的出口。分组经标签交换通过网络到达出口LSR，而后去除标签并处理第三层分组。由于第三层处理被推移到边缘（入口）并只被处理一次，因此提高了性能。而在中间节点的处理包括：将分组中固定长度的标签与LSR连接表中的一个相应项目进行匹配，然后交换标签，这通常是由基于硬件的高性能交换机来完成。

尽管性能和效率是非常重要的，但MPLS带来的好处还不止这些。在那些运行大型基于路由器网络的人们看来，他们最感兴趣的是MPLS执行了先进的业务量管理而没有带来性能的损伤。回想一下，利用标记交换和ARIS建立一个非缺省的显式路由也是可能和比较简单的，它首先鉴别FEC然后在沿着非缺省路径的设备上分配相关联的标签^①。对MPLS而言，转发缺省和非缺省业务量的方式是完全一样的。显式选路的一个可能应用是在不同路径上分发不同FEC的业务量以平衡穿过网络的负荷。另一个应用可能是利用一些配置策略（比如网络入口的位置）在不同的LSP上疏导流量。而且，正像人们所想象的，任何一类FEC可以与一组标签相连并通过一个MPLS网络进行转发。还有一个非常有趣的应用可能是将VPN内部业务量映射成标签，于是可以在同一个VPN的地点之间形成高性能的LSP。

从ATM的观点来看，MPLS应该被看作是另一个控制平面，是IP选路的一部分。它是另外一种建立ATM VC的方法。当运行在ATM硬件上时，MPLS和ATM论坛协议都采用了相同的分组格式（53字节的信元）、相同的标签（VPI/VCI）、相同的标签交换和转发机制以及相同的入口和出口功能^②。他们都需要连接建立协议（例如，MPLS的LDP和ATM的UNI/PNNI）。

根本的差别在于MPLS没有采用ATM寻址、ATM选路和ATM协议。MPLS采用IP寻址、动态IP选路和另一个控制协议（LDP），LDP把FEC映射成标签而后形成LSP。一般情况下，MPLS只涉及创建和分发FEC/标签映射，这样就可以通过一个网络的缺省或非缺省路径更好和更有效地转发IP业务量。当运行在ATM硬件上时，MPLS只是另一种ATM控制平面，虽然它主要是通过IP选路协议驱动并唯一地为选路的IP业务量而设计。当然实际情况是MPLS大多数情况下会与ATM共存，它们以双方各自的方式运行或许MPLS还通过ATM交换机进行通信。最后应该指出的是，MPLS体系结构具备运行在任何数据链路上而不仅仅是ATM的能力。因此，一个网络供应商可能在一个包含点对点协议（PPP）、帧中继、ATM和广播LAN数据链路的域上配置并运行MPLS。

对于所有运行和应用基于IP网络的人们来说，MPLS工作组的形成及其标准化成果是一个积极的发展。它成为一个重要而热烈的论坛，讨论与网络扩展性、路由器设计及性能、协议行为和更经济地引入新服务的能力等相关的重要问题。为了开发出可以更好地利用基础网络传输大量IP业务量的解决方案，它为许多已经装备了ATM设备的网络供应商提出明确的需求。在工业界中许多最优秀的工程师都参加了MPLS的开发。实际上，由专业公司构思的一些聪明想法首先出现在MPLS的讨论中。例如，倒数第二跳弹出和标签栈（Tag）、关联FEC和VC的VCID（Toshiba）以及环路预防和VC合并（IBM）等。可以通过折衷或多数人的意见来解决

① 处理路径上所有节点序列的开销只在LSP建立阶段，与此相反的是，在普通IP选路中，每个分组要处理这些开销。

② 在一定程度上，入口和出口功能是一样的，MPLS和ATM上的经典IP叠加模型都处理分组的第三层协议。

有争议的问题，从而保证了我们所有人都不至于走极端^①。

15.1 体系结构

MPLS的出现是源于早期的IP交换解决方案，因此它的体系结构是基于已经提出的IP交换的想法、概念和组件。它的基本目标之一，正如先前的解决方案一样，是简化通过网络转发IP分组。这个目标是完全有希望实现的。在传统IP转发机制中，每个路由器分析包含在每个分组头中的信息，然后解析分组头、提取目的地址、查询路由表、决定下一跳地址、计算头校验、减值TTL、完成合适的出口链路层封装，最后发送分组。或者简单地说，每个路由器处理每个分组的过程是：分析分组的网络层头字段，根据目的地址前缀为分组分配一个FEC，然后将FEC映射到下一跳路由器。

MPLS旨在简化在路由器入口处处理网络层头分析和FEC分配功能的过程。入口路由器不是将FEC映射到下一跳路由器，而是在分组上添加表示分组归属FEC的一个标签。在下一跳路由器上，因为分组已经与FEC关联，所以没有必要再检查网络层头。标签用于索引一个包含出口端口和一个新标签的连接表。旧标签被新标签取代，然后分组从出口端口转发到下一跳路由器。与传统IP转发相比，MPLS简化的转发机制示例于图15-1。

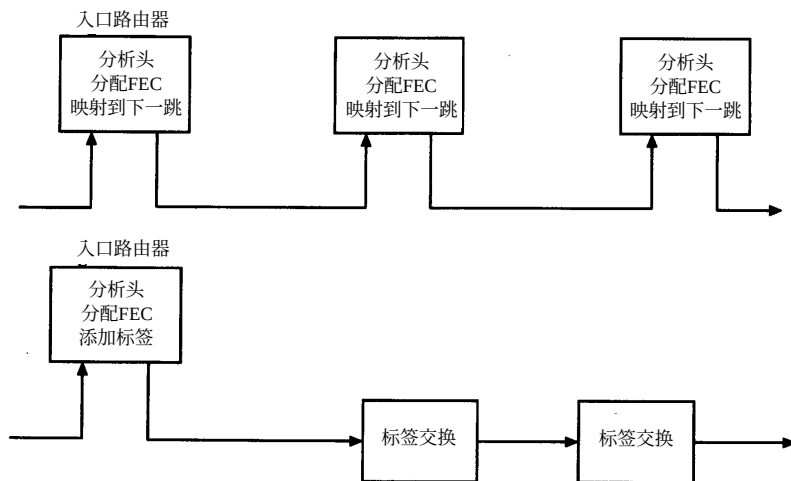


图15-1 MPLS和传统的IP转发

① Bala R.'s "Guide to MPLS Terminology," Aug. 3, 1997, submission to MPLS Mailing List: MPLS: My Product Line Needs a New Story.

标签 (Label)：在盒子上贴标签作产品标识之用。

标签交换 (Label Swapping)：交换路由器标签。标签交换通常需要一个或两个工程师，但是可以由整个市场部实行。

标签信息库 (Label Information Base)：一个包含所有目前工业界中用到的标签数据库。已占用的标签包括“IP Switching”、“Cell Switch Router”、“Tag Switching”。

流 (Flow)：离开一个大公司加入一个IP交换新兴公司的雇员流。

径流 (Stream)：许多流的汇聚；许多离开一个大公司加入一个IP交换新兴公司的雇员。

层 (Layers)：OSI定义三个MPLS层。第三层是天花乱坠的市场广告宣传。第二层是支持市场宣传的策略。第一层是支持第二层的性能方面的数字。

非常混淆 (VC：Very Confounding)：支持标准化MPLS的争论。

直通VC (Shortcut VC)：一个非常混淆的争论，跳过不必要的细节，如性能结果并直接得出MPLS是解决方案的结论。

环路(Loop)：一个间接策略，一个卖主利用对传统路由器性能的怀疑来销售IP交换产品并利用对IP交换的怀疑来销售传统路由器。

环路避免 (Loop Prevention)：一些客户坚持用帧中继上的SNA以避免涉及所有业务。

栈 (Stack)：平稳地累积有关IP交换的草案和白皮书。

LDP：让我们处理分组，是ATM厂商的新口号。

图中MPLS部分显示入口路由器可能将分组映射到不同 FEC的任何编号上。例如，一个FEC可能基于目的端网络地址、一组目的端地址、一个源端 /目的端地址对、一个源端地址或者甚至是网络入口的物理点。一个 FEC也可以表示所有经过一个显式非缺省路径的分组。无论为分组分配FEC的机制多么复杂，通过网络转发分组仍然基于标签交换。于是，与应用传统IP转发机制相比，MPLS使得基于策略的选路以一种更简单和更直接的方式进行工作。

在开发MPLS规范的过程中，产生了一批特殊的功能性要求。不过，这些要求使得MPLS将以一种可扩展的方式解决综合选路和交换的问题。MPLS必须完成的基本要求是：

- 简化分组转发以降低成本并提高性能。
- 独立运行在特殊数据链路上；也就是说，它必须工作在帧和信元介质上。
- 兼容但也独立于现有和将来的网络层选路协议。
- 支持环路避免和检测。前者防止形成环路，后者发现已出现的环路。
- 允许汇聚多个业务流在一个LSP上转发。
- 兼容现有的IP网络管理工具。
- 支持分级操作。
- 支持单播和组播业务量。
- 支持O(N)个交换式路径，其中N是网络中MPLS节点的数目。
- 支持与非MPLS交换机互操作。
- 用非MPLS交换技术支持SIN操作。
- 同时支持拓扑驱动和流驱动的IP交换模型。

在MPLS方案中有一些核心技术和组件。第一，MPLS是一个标签交换路由器（LSR）。类似一个通用IP交换机，它具有第三层转发分组和第二层交换分组的功能。它也能运行传统IP选路协议并可能执行一个特殊控制协议以与邻接LSR协调FEC/标签的绑定信息。第二个核心技术或组件是标签。一个标签是一个包含在每个分组中的短固定长度的数值，用于通过网络转发分组。一对LSR必须在标签的数值和意义上达成一致。例如，下游LSR会告知上游LSR一个特殊标签X代表一个特定称作A的FEC。于是，一个标签只在一对正通信的LSR之间起作用，并用来表示属于一个从上游LSR流向下游LSR特殊FEC的分组。MPLS可以支持添加到现有帧或分组结构（如以太网，PPP）的标签或者它也可以利用包含在数据链路层（如帧中继，ATM）中的标签结构。MPLS采用的第三个核心技术是转发机制，即标签交换。从典型的帧中继和ATM的性能和容量来看，实现标签交换是一个快速和简单的转发过程。没有必要像传统IP选路那样分析分组头中的变长部分。标签作为一个整体（也可能是标签区域中附加的字段如TTL和COS）由交换机组件处理。即使一个分组包含一个标签栈，MPLS设备处理栈中的顶部标签。

图15-2表示出沿着一个LSP转发分组的过程。在入口LSP上，在分组上贴上一个标签，产生一个深度为M的标签栈。沿着LSP的中间MPLS节点接收和处理这个分组。只有栈中顶部标签被处理，并与对应于下一跳下游LSR的新标签进行交换。LSP出口的LSR根据栈中下一个标签的内容作出转发的决定^①。这意味着出口LSR只需要弹出栈就可以得到栈中的下一个标签。如果出口节点告诉仅倒数第二个LSR弹出栈，就可以获得一个更好的优化方案，称作倒数第二跳弹出栈^②。然后分组到达出口MPLS设备并携带用于转发分组的已经在栈中顶部的标签。

① 如果只有一个标签，将基于网络层头的内容作出转发的决定。

② 一个特殊的标签，隐含的空标签，用于告知上行MPLS节点弹出栈。

这样出口设备就不必处理未必有用的标签表查询。

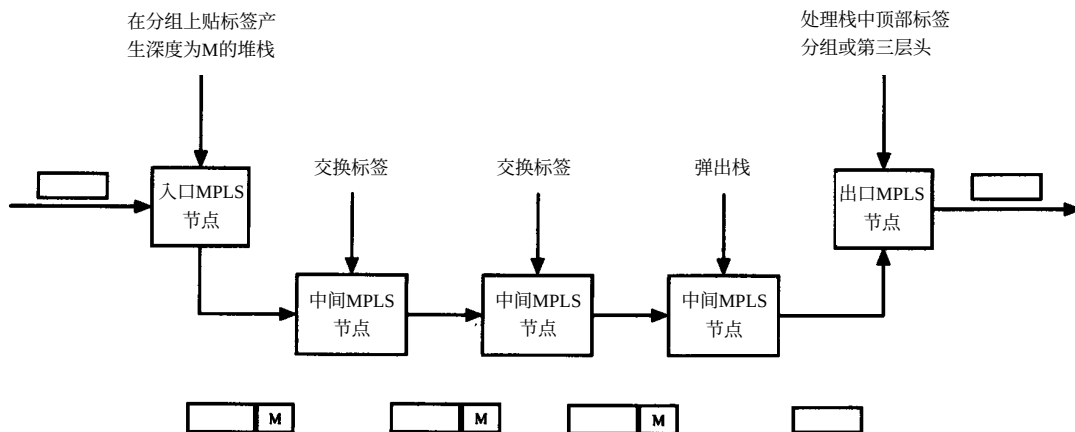


图15-2 沿着LSP的MPLS转发

MPLS的第四个核心技术是标签分发。标签分发是分发 FEC/标签绑定信息的过程，目的是为了形成一个 LSP 并且标签交换属于特定 FEC 的分组。这可以通过一个单独的标签分发协议 (LDP) 来完成或者在现有控制协议中传输 FEC/标签绑定信息^①。对下行 LSR 而言，基本的操作包括分配标签和分发 FEC/标签绑定信息给相邻的上行 LSR。在这种情况下，其中 FEC 对应一个由一个动态选路协议分发的地址前缀，它能够以一种独立或有序的方式完成一个 LSP 的形成（采用 LDP 或另一个控制协议）。其中，独立的 LSP 控制是这样一种情况：一个 LSR 绑定一个标签到 FEC 并告知它的 LSR 邻居这一绑定信息；排序的 LSP 控制只发生在 LSR 绑定一个标签到一个 FEC 的时候，条件是 LSR 是 LSP 的出口或者已经从下一跳 LSR 收到了特定 FEC 的绑定信息。排序的 LSP 控制用于建立一个非缺省特性的路径。有两个非缺省特性路径的例子，一个是经过一个特定顺序节点的路径，另一个是没有环路的路径。MPLS 体系结构具有支持独立和排序 LSP 控制的能力。

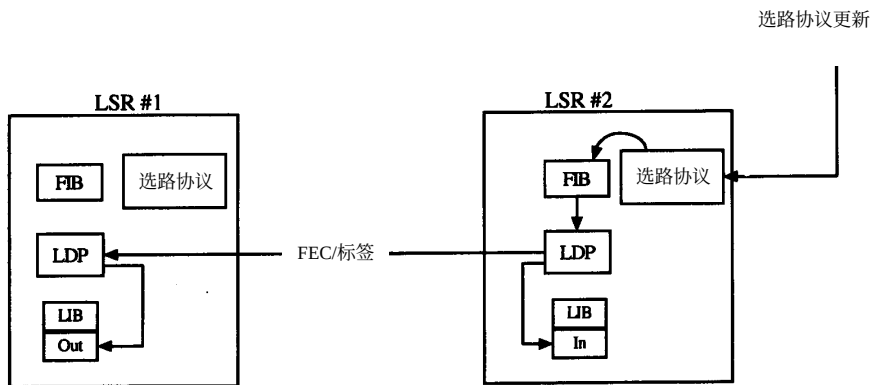


图15-3 拓扑驱动标签分配

要建立 MPLS 的 LSP，可以基于三种方式：特定数据流的到达（流驱动）、保留建立消息

① 如果在现有协议如 BGP 和 RSVP 中尾部附加 (piggy-backing) 标签，就意味着需要改动这些协议。

(如RSVP) 或者选路表更新消息 (拓扑驱动)。考虑到这样的事实, 即可扩展性是最主要的要求之一以及 MPLS 是为特大型 IP 网络而设计, 拓扑驱动的方案是最有可能被普遍采用的。图 15-3 概念性地表示出在一对 LSR 之间路由表更新信息如何启动 FEC/标签绑定信息的交换过程。当 LSR #2 收到选路表更新信息时, 触发计算新路径的过程, 由此在路由表 (FIB) 中进行添加或修改。LSR #2 识别这个过程, 并且对于每个 FEC (这种情况下为地址前缀) 在入口端口上分配一个标签并放置在标签信息库 (LIB) 中。这个标签和相关联的 FEC 就与上行的 LSR #1 通过 LDP 进行通信, 而后 LSR #1 将这一标签放置在它的 LIB 中相应的出口端口上。

15.2 组件

针对于 MPLS 的组件和功能, 有下列一些专业术语:

- 多协议标签交换 (MPLS)。IETF 为综合选路和交换提出的标准。
- 标签交换路由器 (LSR)。支持 MPLS 并负责第三层转发分组和第二层标签交换分组的设备。更具体地说, 一个 LSR 可以是一个传统的交换机 (如 ATM) 扩充 IP 选路, 或者升级为支持 MPLS 的一个传统路由器。在后者的情况下, 路由器根据包含在每个分组中一个显式标签的内容转发分组。
- 标签边缘路由器 (LER)。到并且从一个 MPLS 域转发分组的传统路由器。它也可以与内部 MPLS LSR 通信以交换 FEC/标签关联信息。
- 标签分发协议 (LDP)。它是 MPLS 的控制协议, 用于在 LSR 之间交换 FEC/标签关联信息。
- 标签交换式路径。一个从入口到出口的交换式路径。它由 MPLS 节点建立, 目的是采用一个标签交换转发机制转发一个特定 FEC 的分组。
- 标签信息库 (LIB)。保存在一个 LSR (LER) 中的连接表, 在 LSR 中包含有 FEC/标签关联信息和关联端口以及媒质的封装信息。表 15-1 给出一个 LIB 的例子。LIB 中的项目包含以下内容:

表15-1 样例标签信息库

输入端口	输入标签	下一跳	标识符	标签操作	出口标签	出口端口	链路层封装
1	1	LSRA	NetA	Replace	2	2	ATM
1	4	LSRD	NetD	Replace	5	3	ATM
2	7	LSRA	NetA	Replace	2	2	ATM
3	3	LSRC	NetC	Push	3	2	ATM
4	2	LSRA	LSRA	Pop			ATM

- ☐ 入口端口。
- ☐ 入口标签。
- ☐ 下一跳 LSR。
- ☐ FEC 标识符 (如地址前缀)。
- ☐ 在标签上执行的操作。用一个出口标签替代入口标签, 弹出标签栈或者用出口标签替代入口标签并在标签栈顶部压入一个新标签。
- ☐ 出口标签。
- ☐ 出口端口。
- ☐ 出口链路层封装。

• 径流(stream)。在相同路径上转发并以相同方式处理的分组流。一个径流包含一个或多个流(flow)。在MPLS体系结构中一个径流由一个径流成员描述符(SMD)标识。

• 转发等价类(FEC)。在相同路径上转发,以相同方式处理并因此被一个LSR映射到一个单一标签的一组IP分组^①。一个FEC也可以被定义为将分组映射到一个特定径流的一个操作符。

• 标签。一个固定长度物理上连续的数值,用于标识一个径流。一个标签的格式依赖于分组封装所在的介质。例如,ATM封装的分组(信元)采用VPI和/或VCI数值作为标签,而帧中继PDU采用DLCI作为标签。对于那些没有内在标签结构的介质封装,则采用一个特殊的数值填充。图15-4给出4字节填充标签的格式,它包含一个20比特的标签数值、一个3比特的COS数值、一个1比特的堆栈指示符和一个8比特的TTL数值。此外,如果填充数值被插入到一个PPP或以太网帧中,包含在各自帧头中的一个协议ID(或以太网类型)表示一个帧或者一个MPLS单播或组播帧。

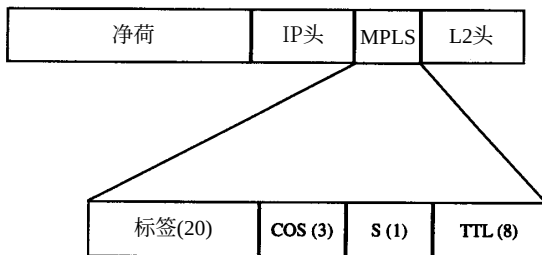


图15-4 MPLS 标签格式

• 标签栈。一个排序的标签集。在一个分组中添加,可以隐含地承载多于一个FEC的信息,即分组的归属以及分组可能经过的LSP。一个标签栈使得MPLS支持分级选路(一个标签用于EGP而另一个标签用于IGP)并且汇聚多个LSP到一个单一的中继(trunk)LSP上。

• 径流合并。一些小径流合并进一个单一的大径流,例如ATM的VP合并和VC合并。一个具有径流合并能力的MPLS网络可以支持O(N)个交换式路径用于传输所有尽力而为业务量。

15.3 标签分发协议

MPLS LDP是一个单独的控制协议,LSR应用它交换和协调FEC/标签绑定信息^②。具体地说,LDP是消息交换和消息格式的序列,它们使得对等LSR就一个特定标签的数值达成一致,这个标签指示出分组所属的一个特定FEC。在对等LSR之间需要建立一个TCP连接,以确保LDP消息能够按照正确的顺序可靠地传送。LDP映射消息可以从任何本地LSR(独立的LSP控制)发起,或者从出口LSR(排序的LSP控制)发起,并从下行LSR流向上行LSR。一个特定数据流的到达、一个保留建立消息(RSVP)或选路更新消息都可以触发交换LDP消息。一旦一对LSR交换用于一个特定FEC的LDP消息,每个LSR关联它们LIB中的入口标签与一个对应的出口标签,之后就形成了一个从入口到出口的LSP。

15.3.1 消息流

LDP消息被分成三类:发现(discovery)、邻接(adjacency)和映射(mapping)。发现类消息发布并维护网络中的LSR。它广播一个LDP Link Hello消息给所有路由器组地址,通知同

① 如果LSR不支持一个径流合并功能FEC到一个标签映射的一个例外就会发生。

② 为了与MPLS术语保持一致,在本章剩余部分将采用径流/标签映射(stream/label mapping)代替FEC/标签的说法。

一链路上其他LSR它的存在，或者引导一个LDP Hello消息给一个特定IP地址，这个地址被没有直接与其相连的LSR接收^①。邻接类消息用于建立、维护和终结LSR对等对之间的邻接关系。它包含了建立一个TCP连接然后交换对话协商消息的过程。协商参数包括LDP协议版本、时间值、VPI/VCI范围等。通告类消息用于建立、修改和删除LSR对等对之间的径流/标签映射信息。一个典型的通告类消息是一个LDP映射消息，它被一个LSR用于与相邻LSR交换一个径流/标签映射信息。这个消息将包含一个径流标识符和一个相关联的标签，可能还有一些补充对象包括一个COS值、LSR ID向量（用于环路预防）、跳计数（在LSP中LSR跳的次数）和MTU尺寸。

图15-5给出三个相邻LSR之间一般的LDP消息流。每个LSR通过发送和接收Hello消息发现在相同链路上一个相邻LSR的存在。然后建立起一个TCP连接并交换初始化（Initialization）消息。之后，由下行LSR为径流=A产生径流/标签映射并传送给上行相邻LSR。

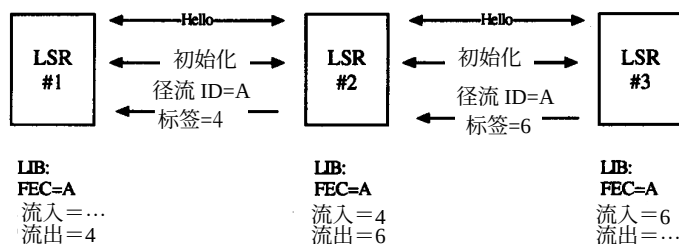


图15-5 LDP消息流

15.3.2 径流成员描述符

MPLS LDP指定了一些径流成员描述符，它们定义映射到一个LSP中的业务量类型和粒度。LDP支持的SMD包括：

- 通配符。它指定任何与给定标签相关联的SMD。
- 网络地址。变长的地址前缀。
- BGP下一跳。
- OSPF路由器ID。
- OSPF ABR路由器ID。
- 汇聚列表（aggregation list）。包含汇聚到一个LSP上的地址前缀的列表。
- 不透明通道（opaque tunnel）。用于识别通过一个显式路由LSP（ERLSP）的分组。
- 流。源端/目的端地址对。

今后还可能补充一些有关组播流的SMD。

15.4 MPLS服务

一个MPLS网络具有支持不同网络业务的能力。它提供的最基本服务是通过网络快速转发分组。尽管这是一个重要功能，而且是任何网络供应商都努力追求的，它还不是选择实现

^① 如果对等BGP被多个中间LSR分开，可能需要在它们之间这样做以交换径流/标签映射信息。

MPLS的唯一原因。实际上，MPLS提供了一个比ATM上的经典IP 叠加模型容易扩展同时有效地支持了ATM网络上的IP业务量的模型。当然，任何网络供应商绝对都会首先要求性能和扩展能力。但是实施MPLS的主要原因是因为它有一个简单的、固定长度的标签，它不是网络层信息，却用于通过网络转发分组。它可以灵活地运送任何用户业务量，申请一个缺省或非缺省的业务，把它与一组标签相关联，然后采用相同的、高性能、高容量的标签交换机制运送业务量。它使得网络供应商可以开发并提供网络层新服务（这些服务对用户设施是可见的），并维持一个单一的高性能传输网络。因为数据转发路径是脱离网络层服务而完全独立的，于是MPLS使得网络供应商可以开发和提供有意思、具吸引力、客户和用户也愿意付钱的网络层服务。

15.4.1 ATM

尽管MPLS可能运行在任何数据链路上，但是出于多种原因，ATM往往成为关注的对象和焦点。首先，在前面讨论IP交换技术中，MPLS是基于这样一个出发点，它们的大多数都试图利用基于硬件的ATM交换技术。这样做具有绝对重要的意义，因为这样可以更好地利用第N代ATM交换技术。其次，许多目前采用传统叠加模型的ISP/承载者类型网络在其核心都应用了ATM交换机。当这些网络的规模和路由器数量增长的时候，扩展VC数目就成了一个主要问题。而采用现有的ATM基础设施可以提供更加有效和可扩展的IP选路能力。同时假设MPLS会立即取代ATM是不现实的。事实上更新现有ATM交换机以支持MPLS有时也是很困难的。此外，传统网络可能会满足并依赖于ATM论坛的协议来支持已有的一个特殊业务。因此，在一个ATM网络上采用MPLS机制必须与现有ATM协议共存并相互合作。最后，ATM是目前唯一一种数据链路技术，能够支持具有QoS要求的数据、语音和视频。实际上，多业务网络供应商看中了ATM的特性，并利用ATM来支持先进的、高性能的IP交换。

在ATM网络上可以按照不同的方式支持MPLS。第一种方案是启用对等模型。在这种情况下，MPLS设备转化成一个LSR，或者讲是一个ATM-LSR的ATM交换机。ATM-LSR支持普通IP选路协议，可能具有转发第三层分组的能力或者运行LDP第三层分组并标签交换第二层信元。ATM-LSR通过串行链路直接相连，彼此交换选路表和LDP消息。在这个例子中，ATM-LSR没有应用任何ATM协议，而是采用LDP建立通过网络的LSP。因为每个ATM-LSR只与它周围的ATM-LSR而不是每个VC另一端的路由器形成邻接关系，所以对等MPLS/ATM实现方案比IP/ATM叠加模型减轻了系统选路协议的负担。由于网络拓扑是基于实际LSR间物理链路的数目，于是它不是多VC（VC-rich）叠加模型中的链路密集（link-dense）型。不过，这里可能忽略了ATM一些很有吸引力的特点，比如基于QoS无环路路径选择和一些更广泛的QoS支持。此外，如果没有利用ATM固有的QoS特性，就很难判断在这一解决方案中是否有必要引入信元头开销了。

另一个方案是综合方案，其中ATM-LSR通过ATM交换机与其他每个ATM-LSR通信。相临ATM-LSR可能建立PVC或利用SVC信令建立VC。VC承载的信息包括：控制业务流、缺省IP业务流、LDP消息和每条径流（per-stream）的LSP业务量。当一个径流映射到一个VC，就需要一个特殊的标识符VCID关联径流和VC。究其原因，是可以像信元穿过ATM交换机一样进行标签交换。也可以事先建立一个VP连接。沿着一个路径的ATM-LSR连接入口VC和对应的出口VC时，也就建立了一个入口到出口的LSP。

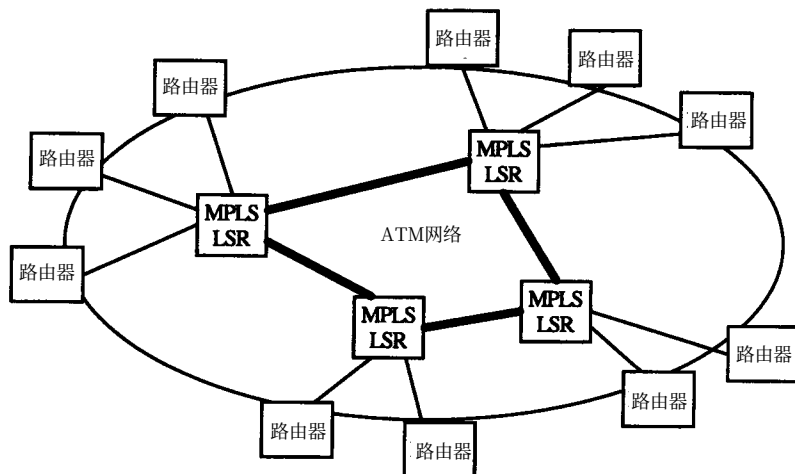


图15-6 ATM上的MPLS综合方案

综合方案的一个突出特点是它将 MPLS引入到现有 ATM网络中。例如，可以在一个大型 ATM网络中间安装一组核心 MPLS LSR作为“VC汇聚”（VC-aggregation）集线器。集线器承载从边缘路由器 VC发送的汇聚业务量。从选路邻接的角度来看，边缘路由器只是简单地与最近的核心 LSR保持对等关系。尽管这可能在核心网络中引入额外的路由器转接，这一方案仍被看作是摆脱“N平方”困境的一个暂时解决方案^①。将来，MPLS功能可以分散到接近网络的边缘，于是允许建立边缘到边缘（edge-to-edge）的 LSP。另一个重要特点是现有 ATM VC管理技术依然适用。这对实际运营很有好处（网络规程和运营可以围绕 ATM VC管理解决），同时考虑一些 ATM交换机可能是无法升级到 MPLS这个实际情况也是有好处的。综合方案可能是从 IP/ATM 叠加模型到 MPLS 最容易的过渡途径。需要说明的是 ATM VC 操作仅提供了 LSR 间的连通性，它依赖于 ATM-LSR 的配置和实现。图 15-6 给出一个 ATM 上的 MPLS 综合方案的例子。

第三种支持 ATM 上的 MPLS 的方案称作 SIN (ship-in-the-night) 配置。在这一模型中，物理的 ATM 模型被分成两个不连接的网络：一个由 PNNI 支持和维护，另一个由 IP 和 MPLS 支持和维护。在图 15-7 中，一个 MPLS LSR 负责为 MPLS LSP、ATM VC 和可能的 IP 分组提供中间交换业务。尽管这是一个与交换有关的工程问题，它一般要求将 VPI/VCI 标签空间划分为 MPLS 和 ATM 部分。此外，一些物理端口可能配置给 MPLS 及 ATM。这样，MPLS 和 ATM 交换机共存但相互独立，共同控制网络交换资源。交换机也可能运行 VC 选路协议如 PNNI。于是，一个采用综合方案的 ATM-LSR 网络会支持现有 MPLS 和

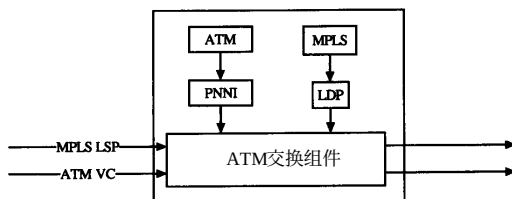
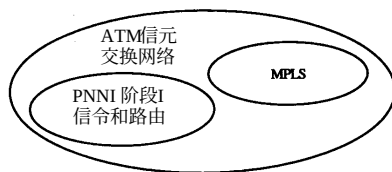


图15-7 ATM上的MPLS的SIN方案

^① 通过在入口 LSR 上把入口 VC 汇聚到一个 VP 中继上然后在出口 LSR 上做相反处理，可以避免核心路由器的转接。

ATM的拓扑、选路和VC管理。

尽管这个方案看起来颇为复杂，它仍具有一定的优势。首先，它支持 ATM。由此带来的是单一的操作过程、管理任务、管理工具和硬件设备。其次，就复杂性而言，ATM实际上只负责为数据、语音和视频提供虚连接上的 QoS。网络供应商可以利用这一技术来支持住户宽带服务（xDSL）、高质量视频服务（MPEG2 over ATM）、电路仿真服务等业务。它不像叠加模型那样可能会耗尽 VC资源，有控制的实施具有 SIN能力的 ATM-LSR是个不错的方案。这样做可以使以 ATM为中心的网络供应商可以继续提供 ATM业务，同时引入一个优化的控制平面来扩展 IP选路功能。

因为 ATM在转发路径上不提供任何环路检测的功能，因此必须考虑防止环路的方法。在实际过程当中，解决可能发生的 ATM环路依赖于若干个因素，比如交换机的实现、预防环路的要求和配置选项等。由于 MPLS采用了 ARIS和标记交换方案中的一些技术，它可以采用多种途径解决环路问题：

- 不要担心环路问题也不要期望环路不会发生。那是一个极端乐天派的想法。
- 依靠 ATM-LSR的每条 VC(per-VC)排队和队列管理方法（如 EPD、RED）来减少环路造成的破坏。这个方案具有一定的合理性，不过它可能要求在交换机中添加新硬件设备。
- 采用面向出口排序的 LSP控制来建立一个无环路的路径。这一方法防止一个 LDP路径建立消息多次通过一个交换机。这是最可靠的方案，只是在 LSP路径建立过程中会耗费大量的时间。
- 采用扩散过程系统地除去会造成环路的上游链路。
- 采用更快、更简单的环路检测算法。

径流合并是对在 ARIS规范中提到的 ATM特有的 VP合并和 VC合并功能的推广。它使得一个 MPLS网络仅以 $O(N)$ 个交换式路径的方式进行扩展。就 ATM网络上的操作而言，必须要考虑与 MPLS功能相关的一些因素：

- 交换机上的 VC合并功能需要一个每条 VC(per-VC)缓存和/或排队机制来防止来自不同源的信元在合并的下行 VC上产生交叉。
- 如果仅支持 VP合并，就必须要防止在到出口 ATM-LSR的 VP上发生 VC冲突，这可以通过对每个入口 ATM-LSR上分配一个无重叠的 VCI范围来实现。当然用其他动态方法（如单节点分配 VCI范围）实现这一点也是可能的，不过它是通过配置过程最容易实现的方法。
- ATM-LSR广播他们支持 VP或 VC合并的能力。
- 如果一个 ATM-LSR必须把一个径流的信元合并到一个单一的下行路径上但不具有完成 VC-合并的能力，它必须要求来自下行 ATM-LSR的一个特殊 FEC的多个标签。标签的数目由 ATM-LSR自己决定（它可能是一个特殊 FEC的入口节点）。一般来说，一个 VP合并节点要求每个 FEC对应一个 VPI，而一个 VC节点要求每个 FEC对应一个 VPI/VCI。一个非合并节点可能要求多个上游节点的 VPI/VCI标签以及一个它自身要求的 per-FEC的标签。

15.4.2 业务量工程

MPLS最有用的应用之一是完成业务量工程。业务量工程，不严格地讲，是指在独立和缺省的路径上引导业务流量的能力。MPLS可以为共享一个入口和出口节点的径流建立独立的 LSP，而不是依赖动态 IP选路协议为所有共享相同出口节点的径流计算路径，这些路径往往基

于单一的尺度公式，如转接数目或最低链路成本等。此外，MPLS对缺省和非缺省LSP的转发路径是相同的，因此也就不会产生性能损伤。

有很多原因可以解释为什么网络供应商们想要实行业务量工程。其中一个原因是网络供应商们想要在整个网络中平均地分配网络业务量负荷，目的是为了提供更大容量的传输能力或者是要避免网络的瓶颈。另外一个原因是网络供应商们试图利用所有可用的网络资源支持用户的业务量。如果存在可用资源（如链路、交换机）未被利用而其他网络资源又没有可用容量或者超负荷运行，这是不合理的。还有一个原因可能是，为了提供一个高带宽和低时延的增值业务，它们需要通过一个单独的路径承载某一客户的业务量。

根据网络的情况不同，可以采用不同的技术实现在一个特定路径上引导业务量的要求。例如，在一个纯路由器环境中，可以采用IPv4的松散或严格源端选路来实现在一个特定路径上转发分组。不过，这不是一个特别有吸引力的解决方案，因为每个分组都需要包含整个IPv4选项头。而且，处理IPv4选项时还会引入一定的开销。在IPv6中选路头是一个更有效的实现方案，不过这当然需要在网络中有IPv6兼容的路由器。

在一个IP叠加环境网络中，网络供应商可能会提供一些额外的VC（如帧中继、ATM），这些VC经过一个特定的链路集和节点集，从而提供更好的网络利用率或为客户提供更大的容量和更好的性能。就业务量工程而言，VC具有一些有用的功能：

- 可以基于每条VC应用QoS特性如带宽和时延等。
- 一条路径上的每个节点都执行接纳控制从而保证存在足够的资源来支持业务量。
- 除非路径上的一个链路或节点受到了影响，拓扑条件的变化不会影响路径本身。于是VC方案存在一定级别的稳定性。

- 可以手工预定或采用明确信令动态建立VC。
- 业务量管理功能（如管制、整形）可以基于per-VC实施。

在一个特定路径上预定VC的吸引力在于其价格。除非采用SVC信令和选路，在预定一个VC时总会有一些管理延迟。这个延迟可能导致当VC预定完成时，却不再需要这条VC了。此外，为了按照需要预定VC，叠加模型在网络中得到了实施。不过，由于叠加模型引入了开销，因此还需要一个能够提供叠加模型的业务量工程功能但不产生开销的解决方案。

MPLS被认为是有效解决业务量工程的一个理想方案。它可以在任何数据链路上轻松地预定一个特定径流的非缺省显式路径。特别地，MPLS具有下列适合于业务量工程的特性：

- 具有预定每条径流LSP的能力。
- 拓扑的变化可能引起动态IP选路重新计算新缺省路径，在此期间MPLS维护非缺省LSP。
- 可以在一个显式路径上应用QoS特性。
- 分组穿过显式路径但不产生性能损伤。
- 可以人工预定或利用选路协议维护的拓扑信息动态计算显式路径。
- 包含显式路径的LSP实际序号只在路径建立期间传输，与之相对的是IP的每个分组都包含整个路径序号。
- 比叠加模型更有效和更具扩展性地支持ATM网络上的显式路径。
- 在传统面向连接的网络中（如帧中继、ATM）以及基于分组无连接网络中都支持显式路径的功能和特性。

在一个MPLS环境中预定显式路径可能采用两个技术。第一个技术是应用LDP建立一个通

过LSR网络的显式选路式LSP（ER-LSP）。LDP的ER-LSP由出口LSR发起并转发到上行LSR，它访问一个特定LSR序列上的每个LSR，直到到达入口LSR。该消息包含ER-LSP消息必须经过的LSR列表、一个标签、一个SMD以及一个任选的用于标签的COS值。另一个技术用到了RSVP和一个显式路由对象（ERO）及一个标签请求对象。在这个方案中入口LSR启动一个包含ERO和标签请求对象的RSVP PATH消息。标签请求对象指示出所要求的一个径流/标签映射和网络层协议类型。ERO包含有PATH消息必须访问的一组LSR。出口LSR在收到PATH消息后向上行发送方（这里为LSP的入口LSR）发送一个RESV消息。RESV消息包含由下行LSR为输入业务量分配的一个标签并应用相同标签将输出业务量传送到上行LSR。此外，RESV消息可能包含QoS参数（如TSPEC、RSPEC），他们源自入口LSR发送的信息，因此允许在ER-LSP上保持QoS状态。这两个方案的概念性的描述参见图15-8。

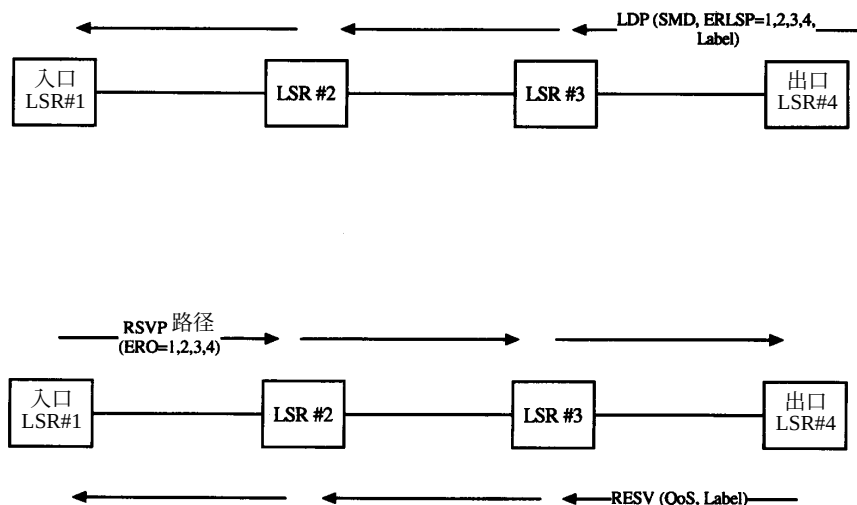


图15-8 MPLS 采用LDP和RSVP建立ER-LSP

15.4.3 QoS/COS

MPLS网络必须提供的另一个重要服务是COS。它之所以重要是因为它允许个别用户流从特殊的COS属性中得到好处，而且网络供应商提供这一服务可以产生额外的收入。MPLS首先在网络边界识别客户流（通过分组头中的多个字段）然后将这些流放置在一个可能具有一些COS或QoS属性的特定LSP上，MPLS实际上使这一过程变得容易了。采用一些可能的技术MPLS就可以支持LSP上的一个非缺省服务：

- 附加在每个分组上的标签可以明确传输一个COS指示符。除了在每个LSR交换标签外，在出口链路上的分组可以根据其COS属性得到服务。MPLS shim头包含一个COS字段。入口LSR会认为所有源端地址为A的分组都应按照高优先级处理并执行差分服务策略，由此可以得到COS指示符。

- 一个COS值可以隐含地与一个特殊LSP关联。这要求LDP或RSVP为LSP分配一个非缺省的COS值，从而可以正确处理该径流的分组。倘若标签不包含一个显式COS指示符，则要求该功能。

- 采用ATM信令和选路（综合方案）建立的一个特殊 LSP可以利用 ATM固有的QoS机制。

15.4.4 虚拟专用网络

VPN是在公用网络上叠加的一个逻辑网络。通常采用公用 Internet作为传输网络提供VPN。在外部选路的IP分组上封装内部VPN分组从而在VPN成员站点之间的链路上建立虚拟通道。这意味着尽管内部VPN业务量穿过一个公用网络，它的性能和容量受现有IP选路的限制。

MPLS可以关联任何抽象网络层的实体和一组独立于转发业务量的标签。因此，关联VPN的成员（及它们各自的网络地址前缀）和一组标签也是很有可能的，这样就可以建立VPN的内部LSP并按照一个安全和可预测的方式交换业务量。由于网络供应商可能支持多个VPN（或者提供VPN成员间的传输业务），因此需要有区分网络路由表目录的方法。通过为每个VPN分配一个VPN标识符（VPN ID）的方法可以解决这个问题。同时它可以用来关联一个特定网络的选路表目录。这样，它就变成了一个通告径流/标签映射（这里SMD地址前缀对于特定VPN是唯一的）到一个VPN相应成员的问题。采用一些技术如在BGP更新消息中携带标签或采用LDP可以解决这个问题。图15-9表示一个利用MPLS形成VPN的实例。

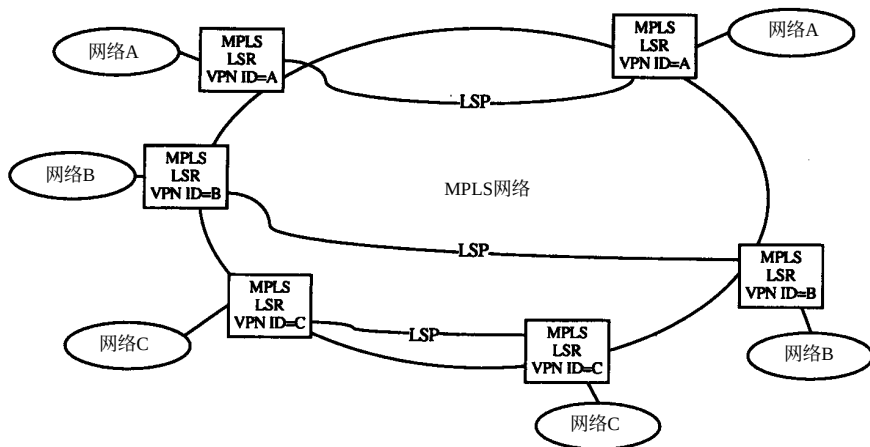


图15-9 MPLS VPN支持

15.4.5 组播

IP组播业务量通常持续一段时间并要求合适的带宽和低时延。典型的应用有多方会议电视和共享白板工具。MPLS通过把一个三层传递树映射成一个LSP树，使得组播分组的有效传输变得容易。此外，组播分组在各自的LSP中流动并与其他LSP上的其他业务量类型分开，这样做的好处是提高了性能并降低了时延。对于组播业务量这一点很重要，因为它完全不关心可用的网络容量。换句话说，它不像TCP或TCP应用那样会对网络拥塞有所反应，这意味着一个组播应用可能会向网络中发送大量的分组并引起拥塞，从而导致TCP作出反应并减少其信息传输速率，影响性能。

MPLS为一个组播传递树的分支分配一个标签，这个组播传递树由一个源端寻根树的（源，组地址）或仅仅由一个共享传递树的（*，组地址）来标识。标签的分配可以基于一个组播转发高速缓存中的新登录条目（由组播数据分组的到达驱动）或者基于显式组播选路协议的控

制消息（如PIM-join）。支持MPLS网络上的组播业务还需要考虑以下几点：

- 因为ATM没有环路检测，必须要避免在ATM网络上层形成一个环路LSP。
- 如果一个特殊组播LSP的物理链路具有广播能力（如以太网），有必要分割整个可用标签空间并为LSR分配不重叠区间。因此传送到不同组播组的LSR不会采用相同的标签。

15.5 MPLS网络实例

MPLS的主要内容是一个拓扑驱动的IP交换解决方案，这一方案在服务提供商的网络中以及其他采用ATM和基于分组基础设施的大型IP网络中得到了实施。前者提供了一个比现有叠加模型更有效地支持ATM上的IP交换式路径的方法。实际上，尽管后者要求 $O(N^2)$ 个链路而且在有大量路由器（N）的网络上会引起路由协议的处理过负荷，采用径流合并的MPLS也只要求 $O(N)$ 个链路。对于一个基于分组设备上的操作，MPLS提供了一个有效和直接的技术建立和维护非缺省的显式路径。对于一个任意规模、复杂度和容量且基于分组路由器网络的业务量工程而言，MPLS技术是一个有用的机制。

图15-10给出两种可能情况的实例。MPLS在相同LER入口到出口对之间建立了单独的LSP。在第一个例子中，一个LSP承载所有去往网络A、B或C的分组而另一个LSP承载所有去往网络D的分组。即使具有相同的入口和出口点，也可以建立单独的LSP来承载粒度变化的业务量。如果网络供应商愿意的话，LSP上业务量的粒度甚至可以下降到主机到主机级别或应用到应用级别。在第二个例子中，为传输所有去往网络D的分组建立了一个单独的ER-LSP。这是出于性能或安全的方面的原因考虑或仅仅为了平衡多个路径上的骨干业务量。

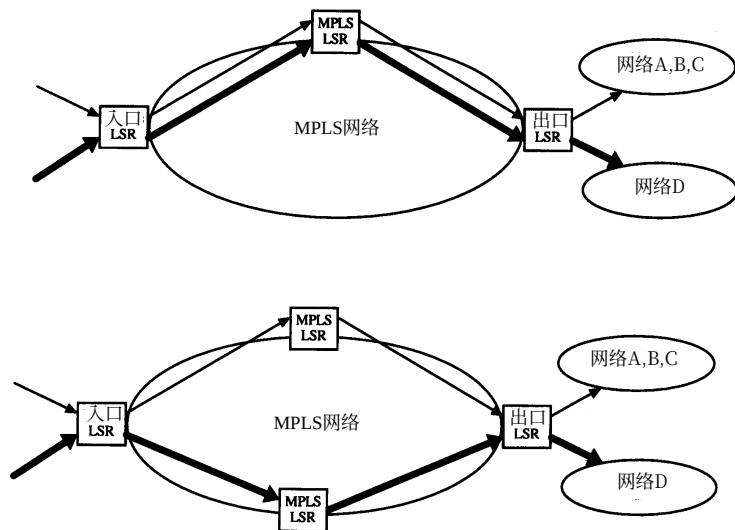


图15-10 MPLS网络实例

15.6 MPLS与MPOA的比较

MPLS规范的主要贡献包括一些有关各种IP交换方案的工作，还有从这些方案中汲取的许多MPLS的有用的特征。虽然任何标准都需要折衷地选择一些可能的功能，对大多数人而言MPLS是个不错和稳定的方案。

与其把MPLS与先前的方案做比较，不如比较它与另一个著名的ATM上的IP交换方案，MPOA。它们之间最基本的差别在于方案所设计的环境不同。MPOA基本上是一个基于校园网的解决方案，它允许在主机和边缘设备间建立通过ATM网络的直通VC。提供了一个可以叠加在ATM网络上的虚拟路由器服务。它的作用依赖于校园网中ATM交换机的密度。另一方面，MPLS可以应用在一个不止包含ATM交换机而且还有其它数据链路技术的网络中。它为业务量工程提供了一个简单机制，在某些情况下增强了基础设施的扩展特性。这两个方案都比传统IP选路提供了更好的性能而且可以动态地应用骨干网络资源或者至少简化了它们的应用。MPOA利用PNNI选路而MPLS采用业务量工程。

表15-2比较了MPOA和MPLS的一些功能特性。

表15-2 MPOA与MPLS的比较

属 性	MPOA	MPLS
环境	校园、WAN	WAN
路由器实体	虚拟路由器	标签交换路由器
模型	叠加	对等
寻址	分开的IP和ATM	仅IP
选路协议	IP单播和组播及ATM选路(PNNI)	IP单播和组播
交换式路径的建立	流驱动	拓扑驱动(也可以支持流和预留驱动)
控制协议	IP、MPOA、NHRP、ATM论坛协议	IP和LDP
设备	主机、边缘设备路由器	路由器和交换机
对ATM支持	是	不支持但可以共存
数据路径选择	PNNI阶段I	动态IP选路或显式路由
标准	ATM论坛	IETF
非ATM数据链路	通过边缘设备	是
故障单点	有，MPOA路由服务器	否

更多的信息

Feldman et al., “Evolution of Multiprotocol Label Switching,” IEEE Communications, vol. 36, no. 5, May 1998.

Hagard, G., et al., “Multiprotocol Label Switching in ATM Networks,” Ericsson Review, no. 1, 1998.

<http://www.employees.org/~mpls/>

<http://dcn.soongsil.ac.kr/~jinsuh/home-mpls.html>

第16章 结 束 篇

IP交换技术提出的最初思想是想要为IP业务流提供更大的容量和更好的性能，它希望将路由器中分组的处理过程改进为一种提高网络性能的方法。当时的设计目标并不是提供更快的路由器，而是想利用已有的千兆比（multigigabit）转发设备，如ATM交换机作为大容量高性能的转发器。然而，ATM交换机本质上不是分组转发装置，它们只作为信元交换机将输入端口的信元交换到相应的输出端口，而且ATM信头中也没有可用于IP转发的字段。ATM信头中最重要的一个字段是连接标识——也相当于一种标签，用于将信元绑定到一个端到端的虚连接上。因此，为了将ATM交换机转换为路由器（或作为路由系统的一部分），需要增加控制功能，以便在IP业务流的转发路由和ATM虚连接间建立映射关系。一旦这种控制功能发挥作用，IP业务流将不需要经过路由器中基于分组的处理过程，而可以无中断地由交换设备通过网络转发。

事实上，所有的IP交换解决方案，包括MPLS技术，都是简单的将IP业务流映射到ATM虚连接上。正如我们在前面提到标记交换和MPLS技术时所说，这种将分组流映射到短固定长度标签的思想可以推广到基于各种传输媒体的IP网络中，用于映射IP业务流和标签。与ATM虚连接类似，标签交换的路由可以通过网络中的分组路由器来建立，如果网络不提供传统的第二层交换，标签交换至少可以通过非缺省路径传送分组，而且不会降低传输性能。当然，建立IP业务流与标签交换间的映射关系也是有代价的。它需要增加额外的负载和复杂度，用于将无连接的分组转发结构映射到一个静态的面向连接的下层体系上。即使是对等模型中包含的软状态IP中心方案，也要求在网络出入节点间的每一段路径上进行绑定操作，这种操作并不要求在连接路径上独立或有序的进行，而只需要在端到端的意义上实现。同样，ATM论坛的信令协议系列中提出的所谓重量级（heavyweight）信令机制也要付出这样的代价。不过，通过这些机制，网络可以获得更高的吞吐量，也可以更高效的利用网络资源。总的说来，这些代价是值得的。

这一时期也出现了其他简单的解决方案，它们可以或至少许诺可以解决IP交换所能解决的一些问题。如下一代的以太网——千兆以太网，可以支持1000兆比的传输速率；在校园网中采用第三层交换技术，可以在以太网间通过费用较低的方式，以每秒上百万比特的速率转发分组；还有为下一代IP网设计的WAN吉比特和特比特路由器，它们可以支持每秒千亿比特的转发容量。这些技术将会运行在光纤网上，并通过高速的传输和转发机制对不同的服务提供支持。通过对以上技术的支持，IP网络不仅可以传输数据，还可以承载几乎所有的业务，包括语音和视频业务。

16.1 变化的网络需求

不论是从终端用户，还是从提供服务的网络运营商的角度来看，信息网络所扮演的角色和应该提供的服务都发生了快速而巨大的变化。那些公开提出并已经作书立传的解决方案既包括基本的实验研究，也有相应的产品出现。它们逐渐引进了一系列新的设想，这些设想有时甚至超出了当时的实际情况。回想一下，不久以前人们还在冥思苦想如何解决连接另一台

主机下载文件的问题。而且，当你匆忙给国内另一所大学发出电子邮件后，你甚至对收信人是否能够收到邮件都没有完全的把握，这被认为是计算机文化中一种奇怪的现象。各公司拥有并且掌握他们自己的网络，包括自己的设备和专用的通信线路，用于本公司职员间的业务交流。而除了这种通信方式外，还有面对面会议，电话联系或硬拷贝交换方式。

随着全球 Internet 和 URL 逐步取代了电话号码、商业卡以及电视广告，一切与之相关的事件都希望有所变化。仅仅与计算机相连已经不能满足需要，人们希望获得更快、更透明、更安全的连接。现在，终端用户以及代表着终端用户的应用（不单指下载文件的应用），将执行网络程序并且表示网络内容。这些内容可能存在于全世界不同角落的不同计算机中，也就是说对于终端用户而言，谁将立即获得这些内容并不重要，而且在这种时候，等待消息发送到目的地也是可以容忍的^①。

另外，电子邮件的通信方式也需要改进。尽管过去它采用一种原始但有效的基于文本的形式表示，现在也出现了一系列新的表示方法，可以支持各种各样不同的附件，包括语音和视频等。语音和视频附件提供了对人类交流至关重要的语言和视觉信息的通信方式。虽然基于 IP 网络的实时语音和视频通信现在仍处于发展初期，过不了多久，它们将像今天的电子邮件一样被普遍使用。

各公司也取消了将自己的专用信息网与世界其他网络隔离的硬件物理屏障，而且在 Internet 上很快建立了自己的主页，向经常上网的成千上万的人们宣传他们的产品。许多人认为，电子商务是赢得新客户、获得更多商机和最大限度提高市场占有率不可错过的机会。同时，各公司也把 Internet 看成与商业伙伴和投资商建立“电子联系”简单而便宜的方式。为了在全球范围内管理商务，各公司迫切需要与远离总部的雇员建立高效安全的通信。

网络运营商们共同组建了包括全球 Internet 在内的互连网络，他们需要改进基础网络来支持新兴的各种应用和业务。他们现在面临的问题是如何处理空前增长的业务流量以及如何满足客户们希望获得更好更可靠服务的需求。简而言之，网络运营商必须提供具有竞争性的新业务，才能让客户乐意支付额外的费用，否则，他们将面临被淘汰的危险。

即使一项新技术与当前的网络状况非常兼容，还必须考虑未来网络需求的变化，这些变化主要包括：

- 提供差分服务。客户希望有各种质量等级的服务可供选择，而且只需支付相应的费用。这种需求不能通过简单的连接或采用宽带网来解决，还必须要有相应的技术支持。

- 简单性和可扩展性。网络运营商掌握的核心网络所采用的新技术必须简单可行，具有可扩展性并易于管理。这些技术也必须易于实现，这样网络运营商才能提供给客户乐意付费的特殊业务。而且，客户还希望有多种等级的服务可供选择。由此引起一个矛盾：简化核心网络 and 对外提供复杂服务。

基于 IP 网。毫无疑问，今天的世界是 IP 网的世界，所有的信息通信都将增加 20 字节的 IP 头进行传送。

16.2 流行的解决方案

IP 交换的一个最主要优势是提高性能。特别是在采用了 ATM 交换硬件设备后，将进一步

^① 当时人们认为这种等待是游戏的一部分，但它逐渐不能被容忍了，而且这种忍耐会随着人们对网络要求的进一步提高而更加难以接受。

提高性能。一般来说，ATM交换技术具有非常强大的交换能力，与同为分组处理的路由器相比，ATM交换设备具有很高的性能价格比。然而，在校园网和广域网范围内，也出现了许多新的数据链路技术以及可以将传统路由功能封装到硬件设备中的更快更低费用的方法。这些更便宜更快的交换技术（包括IP和其他技术）可能更具有竞争性，事实上也的确如此。试想一下，如果采用一种IP转发设备提供更快的传输能力，可以代替IP交换设备并且与交换技术保持同等的价格，那么有什么理由选择IP交换技术呢？

千兆以太网实验就证明它的能力，可以处理每秒1千兆比特速率的数据^①。就带宽而言，它比快速以太网增加了十倍，比传统的10兆以太网快了100倍。对于超过5公里的长距离通信，千兆以太网采用单模光纤（SMF）；而对近距离则采用便宜的多模光纤。随着工作站和服务器中CPU性能的提高以及企业成批生产大规模集成电路带来的价格下降，千兆以太网将解决下一代校园网的问题。速度提高了，费用降低了，而且实现也简单了，除此之外，也就没有什么问题需要解决了。

不过，在校园网环境中加入千兆以太网需要进行仔细的配置。千兆以太网只是一个更快的局域网段而已，它为每个工作站或服务器提供快速发送和接收数据的能力。校园网不能直接将一个单独的局域网段扩展到终端用户桌面，事实上，千兆以太网、快速以太网、令牌环网等多个单独的局域网段间，必须通过网桥或路由器连接。而且，如果校园网希望提供一定程度上的健壮性，它也需要具有容纳链路和节点差错的容错性能，也就是说，校园网应该提供冗余路径，而且各个互连的局域网段间也应该提供动态的重选路由能力，保证以太网数据可以选择激活的冗余路径进行传送。

图16-1是千兆以太网局域网段与采用第二层交换的骨干网相连的示意图。图中第二层交换设备除了实现普通LAN交换机的功能外，还可以在网段间按承载媒体的速率提供无阻塞的转发能力。从图中可以看到，生成树算法在任意两个网段之间形成了一个无环路径，而且这条特殊路径穿过了网络拓扑节点的某个特定子集，换句话说，骨干网的路径集中在特定的链路和交换设备上。事实上，当网络拓扑发生改变时往往会重新计算新的生成树，但是在某一确定的生成树情况下，即使其他链路和节点能够使用，也会处于阻塞状态而无法利用（除非有广播方式的业务流通过）。相反地，组成生成树的链路和节点处于激活或转发状态，以太网数据只能通过它们转发。因此，即使存在那些千兆以太网段和无阻塞线速的第二层交换设备，特定网络拓扑结构也只能使用整个网络的一部分元素，剩余的元素不能使用而且也无需考虑。如果生成树包含的链路或节点发生拥塞，整个骨干网的性能也会受到影响。不过这并不是千兆以太网本身的问题，毕竟在同一网段中的设备仍可以用千兆速率来转发数据，而且这也不是第二层交换设备的问题。相对而言，这应该是生成树路由选择算法的问题，因为为了提供无环路径，生成树算法不能最优地使用网络资源。当然，可以设计一种网络拓扑结构，使任意两个网段间只有一条可能的路由，这样的网络就能避免以上问题。

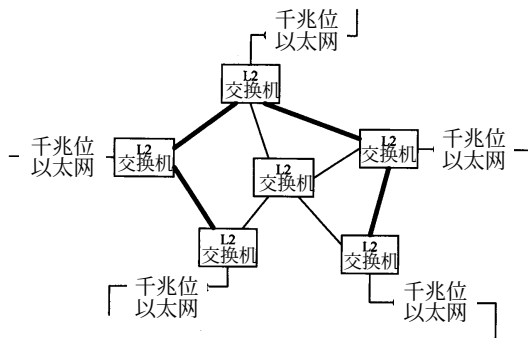


图16-1 二层交换网络中的千兆以太网

① <http://www.gigabit-ethernet.org/>。

另一个解决校园网问题的方案被称为第三层交换，这是一种具有许多不同特点的新型互连设备。第三层交换设备基本上可以算作便宜的、基于硬件线速的路由器，它们具有典型的以太网多端口特性，可以支持 10/100兆以及千兆以太网的交换，在多数情况下使用传统的单播和组播协议，并且拥有每秒 1兆比分组的转发速率。目前的校园网刚好需要配备便宜而且性能高的路由器，看上去采用第三层交换技术可以解决所有问题。

经过以上讨论，我们关心的仍然是骨干网的性能，因为它是具有千兆速率的终端用户竞争的主要资源。图 16-2演示了与图 16-1相同网络拓扑的骨干网，但是它采用的是第三层交换设备。在这个网络中，动态 IP选路协议可以基于最少的转发节点数（如 RIP）或最低的链路费用（如 OSPF）选择路由，将分组按其特定的目的地址转发。图中分组需要从网络 A 传送到网络 B，它将经过第三层交换设备 1、2和3。如果交换设备2或者连接交换设备2和3的链路发生拥塞并导致分组丢失，会出现什么情况呢？除了网络 A到网络B的业务流性能会受到影响外，其他业务性能不会降低。动态 IP选路协议的缺点是即使存在另一条可选路径穿过网络，它也不能将分组从拥塞链路引导到可选路径中；同样，即使存在另一条非拥塞路由到达目的端，动态IP选路协议也不能将新的或已有的业务流引导到非拥塞路由中，有时甚至不能发现这样的路由。

图16-2演示的网络再一次表明，尽管存在着连接到桌面的千兆以太网以及连接这些网段的线速转发设备，也不能完全利用网络资源（如带宽，骨干网转发容量等）。这是因为动态 IP选路协议只使用一个静态参数计算路由，既不考虑路由的剩余带宽，也不考虑所选路由是否拥塞，甚至不知道也不关心传送的数据内容是什么^①。这样即使网络中的一部分链路和节点可能过度使用，网络的另一部分却处于空闲状态。因此，在采用第三层交换设备连接千兆以太网，并通过动态 IP选路协议选路的骨干网中，业务流很可能流过拥塞路径。另一种类似的情况是，当网络拓扑发生变化时，动态 IP选路协议可能将所有业务流转移到转发次数少且费用低的路由中，而这个新路由也可能是拥塞路由。

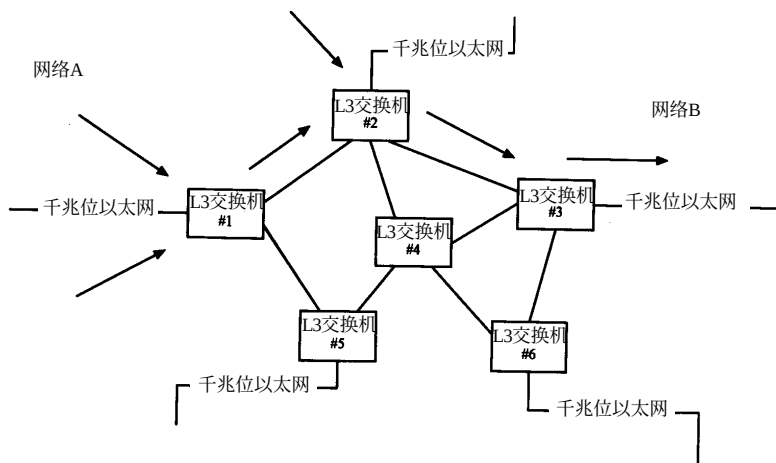


图16-2 三层交换网络中的千兆位以太网

^① RFC2386, A Framework for QoS-based Routing in the Internet, 这篇文章讨论了以上问题并提出了可能的解决方案。

现在的问题是，IP交换方案能否解决校园骨干网中的问题？关键在于采用什么样的IP交换机制。IFMP或FANP仍然依赖动态IP选路协议并且需要一个额外的控制协议；标记交换和ARIS适用于大型的ISP类型网络。对于校园网，需要采用便宜且有多家厂商可供选择的解决方案。假设以太网可以连接到桌面，骨干网络需要一个可以利用整个网络、拓扑结构健壮且动态的选路协议。

在众多IP交换方案中能够较好满足以上要求的只有MPOA虚拟路由器。首先，它通过在网络边缘增加MPOA客户机或在核心网络中增加ATM交换机以提供千兆比转发容量，这种公开的标准已经得到许多生产厂商的支持。其次，它具有多端口特性。它采用标准的IP单播和组播路由协议，连接以太网路由器与ATM上行线路（或其他传输媒体），同时兼容传统的路由器。MPOA虚拟路由器的最重要优势在于它可以采用PNNI阶段1的规则选路。本书前面提到过，PNNI计算路由时会考虑众多因素，包括目的端点的可达性，可选路由的剩余容量和服务质量（QoS）特性以及流过所选路由的业务流及其QoS需求。一旦路由确定，对路由有效性的隐含验证将在连接建立请求时，由沿路交换机的连接接纳控制（CAC）实现。在骨干网的路由建立之后，业务流就可以通过建好的路径到达目的端，这样可以避免上面提到的业务量转移（flip）问题。

当然，更多高效的骨干网路由选择策略也应该考虑费用问题。首先，在骨干网中增加ATM交换机，需要在网络边缘入口点将帧结构转变成信元，然后在网络边缘出口点将信元转换为帧结构。另外，不论是IP选路协议还是ATM论坛协议都需要在控制平面上增加一比特的额外负载和一个单独的地址解析协议（MPOA/NHRP），用于解决IP与ATM的地址映射问题。所幸的是，MPOA是一个稳定的公开标准，可以进行改进，而且已经存在许多查询地址解析服务器并建立相应SVC的方法，如经典IP，局域网仿真等。

表16-1比较了用第三层交换技术和IP交换（MPOA）技术分别组建校园骨干网的一些特性。表中考虑的IP交换是第二层交换，如果网络拓扑比较简单，采用第三层交换就够了。

表16-1 三层交换与IP交换

特 性	三 层 交 换	IP交换(MPOA)
完全IP路由	是	是
介质	Ethernet(10/100/1000)	Ethernet.ATM.和其他
环境	校园	校园网, WAN
QoS	没有	有
应用	IP	IP,ATM
数据路径计算	基于IP选路和拓扑	基于ATM选路和拓扑
额外控制协议	没有	有, ATM论坛协议及MPOA/NHRP
故障单点	没有	有, MPOA服务器

广域网中也开始出现新的千兆比或特比特路由器，它们具有比高速路由器甚至交换机更大的转发容量。同样，这些设备也采用传统路由器的动态IP选路协议，除非它们需要与其他高速网络（如ATM、千兆以太网、光纤信道、分组SONET等）的接口相连。随着光纤的引进和密集型波分复用（DWDM）设备的应用，带宽不足的问题不再像过去那么严重了。而且，千兆比路由器还可以在这些高速数据链路间转发分组。但是，IP网络仍然有两大问题需要解决：一、IP网络和主机都采用IP地址标识，而不是光谱中的波长标识；二、分组只能通过路由器转发到目的地址。

IP交换能够在具有千兆比特或特比特路由器的光纤网络中发挥作用吗？也许它不能提高网络性能，但是可以作为非缺省路由时的转发设备。图 16-3演示的是分组 SONET链路连接千兆路由器形成的网络，它具有千兆比特的转发速率，90%以上的有效负载效率，在叠加模型中无额外负载等特点。看上去它可以解决所有的问题了。网络运营商可以在网络边缘增加差分服务策略，而在核心网络中利用 IP 分组头的 DS 域提供有区别的服务。这正是网络运营商们追求的内部简化并可以分别处理客户业务的极佳网络。

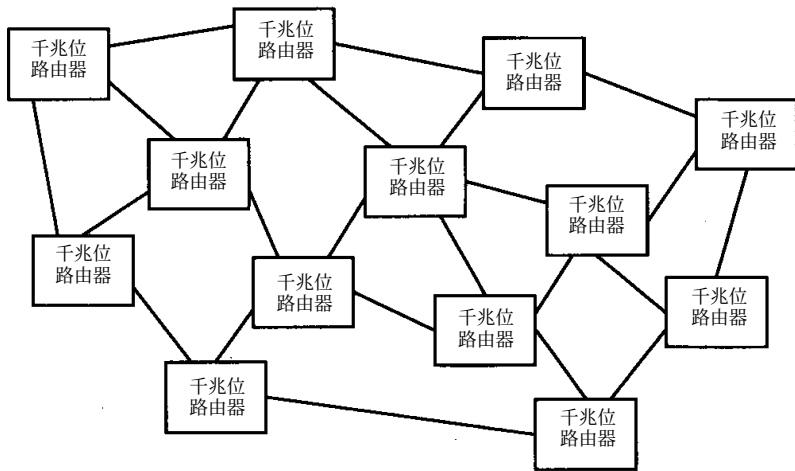


图16-3 分组Sonet互联网的千兆位路由器

然而，校园网中的第三层交换设备仍然采用动态 IP 选路协议。如果一个突发业务量到达，分组很可能由于路径拥塞而丢失。从安全或其他因素考虑，一些特定的业务流可能不需要动态 IP 选路，而是直接通过一条非缺省路由转发分组。另一种可能是，业务流特性突然改变，导致动态 IP 选路选择的链路和节点被突然增加的流量淹没。因此，有必要采用非缺省的显式选路规则。上一章曾经讨论过，如果只用本地 IP 设备（如处理分组的源路由设备、基于路由的过滤器和静态路由器等），通过非缺省路径转发分组将是很困难的。图中的 MPLS 机制没有底层的 ATM 或帧中继支持，却采用了非缺省的显式选路规则。总而言之，在图 16-3 所示的网络中，MPLS 机制应该支持快速高效的非缺省选路规则。

16.3 IP交换能够为我们做什么？

网络产业已经经历或正在经历许多令人惊奇的，富有创造性并且成效颇丰的革新思想。有些思想可以说像大海中的鲸鱼，它们巨大和美丽的景象使我们目瞪口呆，然而，经过最初的欣赏时期，看着它们的喷水孔喷射出水雾，伴随着震耳的泼溅声响后，只剩下了平静的海面，就好像什么都没有发生过。而另一些思想，又像异域的美食，最初不被我们接收，但慢慢的我们习惯了它的味道甚至喜欢上它。还有一部分思想，则像有争议的立法，最后因为暗地的阴谋，妥协，武力扭曲或隐含的威胁等原因而只以一种淡化的形式出现。

IP交换技术属于以上哪种思想呢？是鲸鱼，是异域美食还是有争议的立法？这很难说，但可以肯定的是IP交换中包括了以上三种思想。1996年当IP交换首次出现时，它像鲸鱼的水

柱一样令人们瞠目结舌；随后，由于 Ipsilon 和东芝提出的流驱动（flow-driven）方案在扩展性方面的考虑不够而失去吸引力。IP 交换中的一些解决方案，如 MPOA、标签交换和 ARIS 中的一部分，可以看作是异域美食，只不过还需要一些改进来处理它的味道而使人们喜欢。另外，IETF 对 MPLS 的标准化处理可以算作是有争议的立法。

总的说来，IP 交换在互连网进化过程中起到了积极的推动作用。它为我们做到了以下几点：

- 证实 IP 是当前网络的传输方式，并预测未来的发展趋势。
- 明确指出过去和现在的 IP 选路机制速度太慢。
- 论证了 ATM 是传送 IP 分组的极有效技术。
- 可以明确的为客户分组标记“特定服务”标识。
- 提供建立和管理 ATM 虚连接的轻权（lightweight）控制平面。
- 提供更多高效的可扩展的支持 ATM 网络中 IP 业务流的选路机制。
- 提供更好的显式选路机制。
- 间接鼓励投资商生产更快更便宜的路由器。

China-pub.com

下载

第四部分 附录

术 语 表

802.1.P. IEEE标准 定义了802局域网中支持动态组播过滤和业务量分优先级。通过包含在802.1Q VLAN标记中的3比特用户优先级字段来表示业务量优先级。

802.1Q. IEEE标准 定义了802局域网内支持虚拟局域网(VLAN)的方法。当每个分组经过一个或多个VLAN兼容的桥接器或交换机时,插入一个4字节或10字节的VLAN标记。

AAL(ATM ADAPTATION LAYER, ATM适配层) 一组标准化协议,将用户的业务量适配成信元格式为高层提供服务。

AAL1 (AAL TYPE 1, 第一类AAL) 一种AAL,用于传送固定比特率(CBR)业务量(如音频和视频)并仿真基于TDM的电路(如DS1, E1)。

AAL2 (AAL TYPE 2, 第二类AAL) 一种AAL,用于支持依赖时间并面向连接的可变比特率(VBR-RT)业务量(如分组化的视频和音频)。

AAL3/4 (AAL TYPE 3/4, 第3/4类AAL) 用于支持无连接及面向连接的可变比特率业务,同时支持SMDS。

AAL5 (AAL TYPE 5, 第5类AAL) 最常用的一类AAL,用于支持数据分组的传输。

ABR (AVAILABLE BIT RATE, 可用比特率) 是尽力而为业务类型的一种(另一种是UBR),网络不对信元传递进行绝对保证;但是保证用户传输的最小比特率,并尽可能地降低信元丢失率。

ACCESS RATE (接入速率) 表示一个用户可以发送到网络线路上的速率,单位是比特每秒(bps)。

ACK (ACKNOWLEDGMENT, 应答) 接收到一个发送分组后的应答消息。ACK可以附加在反向业务量分组上或是独立的分组。

ACR (ALLOWED, OR AVAILABLE, CELL RATE, 允许或可用比特率) 一个给定QoS类型的可用带宽,以每秒信元数为单位,并受网络动态控制。

AH (AUTHENTICATION HEADER, 鉴权头) 插入到每个提供数据初始鉴权和完整性的IP分组中的IPsec对象。

ANSI (AMERICAN NATIONAL STANDARDS INSTITUTE, 美国国家标准化协会) 一个美国标准化组织。

ARIS (AGGREGATE ROUTE-BASED IP SWITCHING, 汇聚基于路由的IP交换) IBM提出的一种拓扑驱动IP交换解决方案。

ARP (ADDRESS RESOLUTION PROTOCOL, 地址解析协议) TCP/IP协议的一种,通过将一个物理地址(即MAC地址)映射到一个IP地址来解决本地网络寻址问题。

ARPANET Internet的原始基于网络。

ARS (ADDRESS RESOLUTION SERVER, 地址解析服务器) 描述服务器功能的通用词汇, 它包含的功能有: (1) 维护一个局域网、网络层地址及其相关联的 ATM地址缓存表; (2) 响应来自相关客户的查询信息。

ASYNCHRONOUS (ASYNCHRONOUS TRANSMISSION, 异步传输) 一种基于字节方式获取同步的方法。采用起始和停止比特位进行定界。

ASYNCHRONOUS TRANSFER (异步传输) 一种传输信息的有效方式, 时间槽的概念是基于需求(ATDM, ATM)而不是基于周期(TDM, STM)。

ASYNCHRONOUS TRANSFER MODE (异步传输模式) 一种面向连接的宽带交换和复用技术。由于没有时钟控制, 所以是异步的。信息被封装在固定尺寸格式的分组 (即信元)中, 以很高的速率(最高达每秒上千兆比特)传送。

ATM FORUM (ATM论坛) 一个ATM标准化组织。最初由一些供应商和电信公司资助, 包含多个委员会, 负责制定建议和提出实施规范。

ATM LAYER ATM协议栈模型的第二层, 完成构造和处理 ATM信元、使用参数控制(UPC)和业务质量支持等功能。

ATMARP 客户向ATMARP服务器请求或接收对目的端 ATM地址的IP地址解析所使用的协议消息格式。它只解决连接到同一 LIS上设备的寻址。在RFC1577/2225中定义了ATMARP, 即所谓的 ATM上的经典IP和ARP。

ATMARP CLIENT (ATMARP 客户) 查询ATMARP服务器地址映射和缓存的主机或路由器, 与同一LIS上的设备建立SVC连接。

ATMARP SERVER(ATMARP 服务器) 保存IP/ATM地址映射的服务器, 响应ATMARP客户查询请求, 可以运行在一个独立设备或路由器上。

BEC (BACKWARD ERROR CORRECTION, 反向错误校正) 一种基于接收端反馈的错误校正方法, 发送端重传任何被发现含有错误的数据。

BEST EFFORT (尽力而为) 一类没有具体业务量参数和绝对保证的 QoS。

BT (BURST TOLERANCE, 突发容限) 突发容限与MBS成比例, 作为漏桶机制的参数用于检查SCR的一致性。

BURSTINESS (突发) 一种业务源特性, 定义为峰值信元速率(PCR)与平均信元速率之比, 用于测量信元间间隔。

BUS (BROADCAST AND UNKNOWN SERVER, 广播和未知服务器) 转发组播、广播和未知目的地地址业务流到所附 LEC 的一种服务器。

CAC (CONNECTION ADMISSION CONTROL, 连接接纳控制) 决定是否接受一个虚连接(VC)建立请求。

CAT-3 (CATEGORY-3 UNSHIELDED TWISTED PAIR 第3类非屏蔽双绞线) 一类非屏蔽双绞线, 通常用于低速率信元传输 (25M~50Mbps)的ATM接口, 距离最多100m。

CAT-5 (CATEGORY-5 UNSHIELDED TWISTED PAIR 第5类非屏蔽双绞线) 一类非屏蔽双绞线, 通常用于高速率信元传输 (超过50Mbps)的ATM接口。

CBR (CONSTANT、OR CONTINUOUS、BIT RATE, 恒定或连续比特速率) 一种ATM服务类型, 支持需要满足特定 QoS要求的连续信息比特流的传输, 如语音和视频 (参见QoS分类)。

CBT (CORE BASED TREES, 基于核心的树) 在一个核心路由器上确定一个特定组组播传递树的组播选路算法, 所有该组的发送者都要通过核心路由器转发分组。例如, PIM-SM和CBR是支持CBT算法的组播选路协议。

CDV (CELL DELAY VARIATION, 信元延迟变化) QoS参数, 测量单个信元传输延迟(CTD)和期望传输延迟之差, 表示在一个VC中信元间隔的程度。在ATM复用器(MUXs)或交换机中会引入CDV。

CDVT (CELL DELAY VARIATION TOLERANCE, 信元延迟变化容限) 指定CBR业务流的CDV可接受容限。

CELL (信元) 一个53字节分组, 包含5字节的信元头和48字节的净荷。在源端, 用户业务量被分段成信元, 在目的端被组装成信元。

CELL HEADER (信元头) 5字节的信元头包含目的端路径和流量控制信息, 具有以下字段: GFC、VPI、VCI、PT、CLP和HEC。

CER (CELL ERROR RATE, 信元错误率) 一个QoS参数, 测量发送信元错误的比例(在目的端发现错误)。

CES (CIRCUIT EMULATION SERVICE, 电路仿真业务) 一种由ATM提供的服务类型, 通过AAL1仿真TDM类型的CBR电路业务。

CI (CONGESTION INDICATION, 拥塞指示) RM信元中指示拥塞的一个比特位, 当目的端最后收到的信元被标记时置位。

CIR (COMMITTED INFORMATION RATE, 承诺的信息速率) 帧中继中的术语, 定义在任何网络条件下许诺提供给用户的信息速率。

CIRCUIT EMULATION (电路仿真) 提供给终端用户的一种虚连接业务, 用于仿真一条实际数字比特流(如视频流)线路(如2-或45-Mbps信号)。

CLASSICAL IP 由IETF指定, 在ATM环境下为支持各种通用应用, 如FTP、Telnet、SMTP、SNMP而开发的ATM上的IP的网络协议。包括IP对QoS类的支持以及SVC和PVC网络上的ARP。ATM上的IP传输的主要问题是分组封装和地址解析。

CLP (CELL LOSS PRIORITY, 信元丢失优先级) ATM信元头中对应于信元丢失优先级的一个比特位。在拥塞情况下可以丢弃低优先级 (CLP=1)的信元。

CLR (CELL LOSS RATIO, 信元丢失率) 一个QoS参数, 定义为丢失信元与发送信元总数之比。

CONGESTION CONTROL (拥塞控制) 资源和业务量管理机制, 用于避免和防止网络瘫痪的情况发生(如缓冲器溢出, 带宽不足)。

CONNECTIONLESS NETWORK (无连接网络) 从源端发送分组到目的端不需要事先建立连接的通信业务。如IP和SMDS。

CONNECTION-ORIENTED NETWORK (面向连接的网络) 端点之间建立连接的通信业务。如ATM和帧中继。

COS (CLASS OF SERVICE, 服务类型) 网络对分组进行分类, 然后根据每一类分组的优先级处理分组。

CPCS (COMMON PART CONVERGENCE SUBLAYER, 公共部分汇聚子层) AAL汇聚子层(CS)的一部分。其作用是传递原语到其他AAL子层(SAR, SSCS)。它支持标准公共部分

AAL : AAL1, AAL3/4, AAL5的功能。

CPE (CUSTOMER PREMISES EQUIPMENT, 用户地基设备) 由用户使用并位于用户区域的计算机和通信设备(包括硬件和软件)。(见DTE)

CPI (COMMON PART INDICATOR, 公共部分指示符) 在AAL3/4 CPCS-PDU头中的1-字节字段, 用于指示BASize字段包含的比特数。

CRC (CYCLIC REDUNDANCY CHECK, 循环冗余校验) 一种比特错误检测技术, 采用一种基于已发送比特的数学算法, 该算法对同一分组中的信息比特进行计算, 将计算出的数值一同送出。接收端按照相同的算法, 重新计算该值并与接收到的值进行比较。如果二个值不一致, 则认为发送的分组有错。

CS (CONVERGENCE SUBLAYER, 汇聚子层) AAL的上半层。分为二个子层, 公共部分汇聚子层(CPCS)和特定服务汇聚子层(SSCS)。它与服务有关, 功能包括处理信元延迟抖动(CDV)、源端时钟频率恢复和前向错误校正(FEC)。尽管每个AAL都有自己的功能, 一般来讲, CS描述了在ATM和非ATM协议之间所需的服务和功能。

CS-PDU (CONVERGENCE SUBLAYER PROTOCOL DATA UNIT, 汇聚子层协议数据单元) 在CS层中, 用于在高层和SAR间传送信息的PDU, 信息在此被转换成信元。

CSR (CELL SWITCH ROUTER, 信元交换路由器) 在CSR/FANP体系结构中定义的一种设备, 负责第三层分组转发和第二层ATM信元交换。

CSU (CHANNEL SERVICE UNIT, 信道服务单元) 在用户和通信网络间提供接口的用户端设备。CSU可以和DSU合并在一个设备中(参见DCE)。

CTD (CELL TRANSFER DELAY, 信元传输延迟) 一个QoS参数, 用于测量一个信元从源端到目的端传输的平均耗时。它是编码、解码、分段、组装、处理和排队时延的总和。

DATAGRAM (数据报) 一种分组传输模式。分组独立地选路并可能经过不同的路径, 因此它对分组的传送顺序没有保证。

DCE (DATA CIRCUIT-TERMINATING EQUIPMENT OR DATA COMMUNICATIONS EQUIPMENT, 数据电路终端设备或数据通信设备) 用户端设备, 作为发送媒体的接入点, 通常是一个调制解调器或其他通信设备。

DIFFSERV (DIFFERENTIATED SERVICE, 差分业务) IETF为支持IP网络上的COS而制定的技术。当分组通过网络转发时, 它在网络边缘对分组进行标记, 在网络设备上对分组按照优先级进行排队。

DS-0 (DIGITAL SIGNAL 0, 数字信号0) 速率为64Kbps的数字传输物理接口。

DS-1 (DIGITAL SIGNAL 1, 数字信号1) 速率为1.544Mbps的数字传输物理接口。作为T-1标准, 它可以支持24个DS-0电路。

DS-3 (DIGITAL SIGNAL 3, 数字信号3) 速率为44.736Mbps的数字传输物理接口。

DSU (DATA SERVICE UNIT, 数据服务单元) 终端用户设备, 作为低速(如56kbps)服务和高速线路间的接口。

DTE (DATA TERMINAL EQUIPMENT, 数据终端设备) 为终端用户提供接入通信网络服务的主机(PC或工作站)。DTE与进行信令处理的DCE相连。

DVMRP (DISTANCE VECTOR MULTICAST ROUTING PROTOCOL, 距离向量组播选路协议) 基于RPM的组播选路协议, 采用距离向量机制(如横向分割、距离向量选路表)。同

时也支持无组播能力的单向通道。

DWDM (DENSE WAVE DIVISION MULTIPLEXING, 密集波分复用) 通过不同光波长上的复用, 实现在一根光纤上同时承载大量 (如超过32)多吉比特业务流。

DXI (DATA EXCHANGE INTERFACE, 数据交换接口) 在DTE(如路由器或本地交换机)和DCE之间的一种基于帧的接口。ATM论坛采用DXI接口作为在WAN上提供ATM能力的一种可行性方案。

E.164 由ITU-T定义的8字节地址格式。它一般应用在ATM公用网络, 并由电信商提供。在专用网内部, 采用20-字节的NSAP地址格式。

E-1 (EUROPEAN DIGITAL SIGNAL 1, 欧洲数字信号1) 数字物理接口的速率为2.048 Mbps的欧洲标准。

E-3 (EUROPEAN DIGITAL SIGNAL 3, 欧洲数字信号3) 数字物理接口的速率为34.368Mbps的欧洲标准, 它可以同时支持16个E-1电路。

EFCI (EXPLICIT FORWARD CONGESTION INDICATION, 显式前向拥塞指示) 在PTI中包含的1比特信息字段, 它表示在中间节点中是否发生拥塞。例如, 当EFCI被置1时表示有缓冲器发生溢出。

EGP (EXTERIOR GATEWAY PROTOCOL, 外部网关协议) 域内或AS间选路协议的别称, 如EGP和BGP。

ELAN (EMULATED LAN, 仿真LAN) ATM论坛定义的一种网络, 一组LE客户可以通过它连接到同一个LES/BUS。

ER (EXPLICIT RATE, 显式速率) RM信元头中的一个字段, 指定用户在虚连接上应该传输的信元速率, 参见RM。

ESP (ENCAPSULATING SECURITY PAYLOAD, 封装安全净荷) 在每一个提供数据机密性、鉴权和完整性的IP分组中插入的一个IPsec对象。

FANP (FLOW ATTRIBUTE NOTIFICATION PROTOCOL, 流属性通知协议) 在东芝CSR交换机间采用的一种控制协议。上游CSR设备为下游CSR设备提供流/标签的映射信息。如果上游和下游CSR需要通过ATM交换机通信, 也可能包含对VCID值的转换。

FDDI (FIBER DISTRIBUTED DATA INTERFACE, 光纤分布数据接口) ANSI定义的标准, 用于实现光纤上的高速(100Mbps)LAN。

FEC (FORWARD ERROR CORRECTION, 前向错误校正) 一种不用重传的错误校正技术, 接收端负责纠正分组中的任何错误。

FEC (FORWARDING EQUIVALENCE CLASS, 转发等价类) 在同一路径上以相同的方式处理并被转发的一组IP分组, 可以被映射到一个单一的标签。

FIB (FORWARDING INFORMATION BASE, 转发信息库) 包含目的端、下一跳地址和接口的路由表的别名。

FLOW CONTROL (流控制) 用于网络拥塞避免和业务量控制的方法。它有三种技术: 基于窗口的控制、基于速率的控制和基于信用的控制。基于窗口的控制是采用一个滑动窗口来确定在一个预设的时间内允许发送的信元数。基于速率的控制负责监视和控制源端发送速率。基于信用的控制是当源端有信用(credit)时才可以发送信元。CAC也是流控制的一部分。

FLOW LABEL (流标签) 在IPv6头中用于识别属于一个特定流分组的一个20比特字段。

FLOW TYPE (流类型) 定义在Ipsilon IP交换机中映射到VC中流的类型(如主机到主机、应用到应用)。

FLOW-DRIVEN IP SWITCHING (流驱动的IP交换) 基于单独IP流到达建立直通路路径的一种IP交换解决方案。例如, IFMP、FANP和MPOA。

FRAME RELAY (帧中继) 提供一个非常可靠的虚电路分组传递技术。帧中继中的一些概念被用到了ATM网络中。

FRTE (FRAME RELAY TERMINATING EQUIPMENT, 帧中继终端设备) 终结帧中继VC的网络互通设备(如路由器、FRAD)。

FTP (FILE TRANSFER PROTOCOL, 文件传输协议) 用于在网络中不同机器间传输文件的协议。

GARP (GROUP ADDRESS REGISTRATION PROTOCOL, 组地址注册协议) 一个通过网络注册属性的接收端驱动协议。这些属性包含组播地址和 VLAN成员资格信息。在 IEEE 802.1Q规范中定义了GARP。

GBPS (GIGABITS PER SECOND, 千兆位每秒) 千兆位每秒的传输速率。

GCRA (GENERIC CELL RATE ALGORITHM, 通用信元速率算法) 由ATM论坛提出, 并根据一些特定业务量参数定义信元速率一致性的一个参考模型。通常采用漏桶算法。

GFC (GENERIC FLOW CONTROL, 通用流控制) 在ATM信元头中支持复用功能的一个4比特字段。当GFC协议尚未启动时默认值为‘0000’。GFC支持ATM连接中的简单流控制。

GMRP (GARP MULTICAST REGISTRATION PROTOCOL, GARP组播注册协议) GMRP允许终端系统和交换机注册, 删除组播成员资格的注册信息, 并通过网络发布信息。在IEEE 802.1P 规范中定义了GMRP。

GVRP (GARP VLAN REGISTRATION PROTOCOL, GVRP虚拟局域网注册协议) VLAN设备(如工作站和交换机)采用GVRP进行动态注册并通过一个 VLAN设备向网络分发VLAN成员资格信息。允许一组 VLAN交换机建立和更新处于激活的 VLAN。在IEEE 802.1Q 规范中定义了GVRP。

HEC (HEADER ERROR CHECK OR HEADER ERROR CONTROL, 头错误检验或头错误控制) 信元头中用于头错误校正和检测的1字节字段。由于头包含一些重要信息, 所以 HEC也非常重要。

HOL (HEAD-OF-LINE, 线路前部) 交换机内部缓冲器的头位置。与 HOL有关的是阻塞现象, 即队列中的信元必须要等待 HOL信元头先离开队列。

IEEE (INSTITUTE OF ELECTRICAL AND ELECTRONIC ENGINEERS, 电气和电子工程师协会) 在计算机和电子领域有广泛影响的一个标准化组织。

IETF (INTERNET ENGINEERING TASK FORCE, INTERNET工程任务组) 一个为实现IP互操作而开发规范的组织。

IFMP (IPSILON FLOW MANAGEMENT PROTOCOL, IPSILON流管理协议) 一个在Ipsilon IP交换机间运行的控制协议, 其中下行流 IP交换机采用一个新VPI/VCI标签重新标记一个特定流信元的方法指导上行流 IP交换机。在RFC 1953中定义了IFMP并在Ipsilon IP交换机和边缘设备上得到了实现。

IGMP (INTERNET GROUP MANAGEMENT PROTOCOL, INTERNET组管理协议) 组

成员资格协议，运行在支持IP组播的主机和路由器上。

IGP (INTERIOR GATEWAY PROTOCOL, 内部网关协议) 域内或AS间选路协议的别名。如RIP、OSPF和IGRP。

IISP (INTERIM INTERSWITCH SIGNALING PROTOCOL, 临时交换机间信令协议) 采用基于UNI信令和静态地址前缀表转发SVC请求的一个协议。

ILMI (INTERIM LOCAL MANAGEMENT INTERFACE, 临时本地管理接口) ATM论坛在SNMP基础上制定的网络管理系统(NMS), 可以提供在UNI接口上与虚连接配置、性能和故障管理有关的信息。它在AAL3/4和AAL5上运行。它一旦用ITU-T进行标准化就最终会被取代。

INTSERV (INTEGRATED SERVICES, 集成服务) 支持尽力而为和实时业务流的IP模型。包含建立协议(如RSVP)、业务量控制、业务量分类和流规范。从RFCs 2205到2216定义了INTSERV。

IP (INTERNET PROTOCOL, INTERNET协议) 为高层传输协议提供无连接业务的一个连网协议。它负责发现和维护拓扑信息并穿过同质或异质网络转发分组。与TCP一起, 通称为TCP/IP协议。

IP MULTICAST (IP组播) 一种IP网络传输业务, 其中分组寻址到一个组地址, 路由器传递这些分组给组成员。主机应用一个组成员资格协议(IGMP), 通知路由器它们想加入或离开哪个组播组, 然后路由器允许一个组播选路协议建立一个从源端网络到所有包含组成员网络的传递树。

IP SWITCH (IP交换机) 可以在第三层转发IP分组并能在第二层交换分组的设备或系统。一个IP交换机区分哪些分组需要在第三层转发, 哪些分组需要在第二层交换, 然后在第二层交换式路径上重定向一些或所有分组。大部分IP交换机利用ATM交换结构, 不过也可以采用其他第二层交换技术。

IP SWITCHING (IP交换) 采用第二层交换加速IP分组转发的一组技术的总称。

I-PNNI (INTEGRATED PNNI, 集成PNNI) PNNI的扩展, 在PNNI的基础上集成了IP和ATM选路形成单一的选路协议。

IPNG (IP NEXT GENERATION, 下一代IP) IPv6的别称。

IPSEC (IP SECURITY, IP安全) 在IPv4和IPv6中为支持网络层安全性而定义的协议和机制, 保障每个IP分组的安全性。它支持数据的原始鉴定和加密, 并基于在非安全网络上启动安全通道的技术, 从而可以在公用Internet上建立虚拟专用网(VPN)。

IPX 由Novell开发并用于Novell客户/服务器通信的网络层协议。

ISAKMP/OAKLEY 在通信IPsec实体间支持共享式密钥动态生成、分发和交换的Internet密钥交换(IKE)。

ISDN (INTEGRATED SERVICES DIGITAL NETWORK, 综合业务数字网) CCITT早期采用的一种协议参考模型, 目的是为数据、音频和视频提供一种普遍适用、端到端、交互式的数字业务。

ISOCRONOUS (同步的) 参考一个时间槽可以分解成等尺寸的微小时间槽情况, 微小时间槽分配给了同步传输信息的不同信道。

ISSLL (INTEGRATED SERVICES OVER SPECIFIC LINK LAYERS, 特殊链路层上的集

成服务) 一个IETF工作组, 负责开发在特殊链路层如 ATM、802 LAN和低速率链路上支持 IntServ的规范。

ITU-T (INTERNATIONAL TELECOMMUNICATIONS UNION-TELECOMMUNICATIONS STANDARDS SECTOR, 国际电信联盟-电信标准部) 一个前身为CCITT的正式国际标准、规范和建议组织。ITU-T是1948年成立的国际电信联盟 (ITU)的一部分, 由联合国 (UN)主办, 目的是促进电话和电报等业务的发展。

JITTER (抖动) 信元或分组延迟变化。

LABEL (标签) 一个固定长度物理连续的数值, 用于标识一条径流。标签的格式依赖于封装分组的介质。例如, ATM封装的分组(即信元)采用VPI和/或VCI作为标签而帧中继的PDU采用DLCI作为标签。对于那些没有内在标签结构的介质封装, 通常采用一个特殊的 4字节字段。

LDP (LABEL DISTRIBUTION PROTOCOL, 标签分布协议) 用于在LSR间交换FEC/标签绑定信息的MPLS控制协议。

LER (LABEL EDGE ROUTER, 标签边缘路由器) 负责从MPLS域转发分组的一个传统路由器。也可以与内部MPLS LSR通信以交换FEC/标签绑定信息。

LABEL INFORMATION BASE (LIB, 标签信息库) 保存在LSR(LER)内的连接表, LSR内包含有FEC/标签绑定信息及相关联端口和媒体封装信息。

LABEL STACK (标签栈) 一组追加在一个分组上已排序的标签, 允许隐式地传送有关FEC的信息, 这些信息包括分组的归属和分组经过的 LSP。一个标签栈允许MPLS支持一个选路系列(EGP的标签和IGP的标签)并将多个LSP汇集到一个单独的“中继”LSP上。

LSR (LABEL SWITCH ROUTER, 标签交换路由器) 支持MPLS并负责第三层分组转发和第二层分组标签交换的设备。

LSP (LABEL SWITCHED PATH标签交换路径) 通过一系列LSR建立一个从入口到出口的交换路径, 并利用一个标签取代转发机制转发一个单独的 FEC分组。

LAG (LOGICAL ADDRESS GROUP, 逻辑地址组) 连接到一个物理NBMA网络上的主机和路由器的集合, 负责在不同子网间建立主机和路由器间的直连路径。

LAN (LOCAL AREA NETWORK, 局域网) 在近距离(通常在同一层或大楼内部)高速互连PC、终端、工作站、服务器、打印机和其他外设的一个网络。在各种各样的 LAN标准中, 以太网应用得最为广泛。

LAN EMULATION (局域网仿真) 为在ATM环境中提供支持LAN功能性和连通性所需的接口和协议, 可以使ATM协议、接口和设备与传统协议互操作。

LANE 见LAN Emulation。

LEAKY BUCKET (漏桶) 一种流量控制算法, 网络监控信元是否遵守连接参数, 对不一致的信元进行标记或丢弃。类似一个底部有洞的桶 (存储器), 允许流体(信元)以一定的速率流出(参见GCRA, Traffic contract, UPC)。

LE-ARP (LAN EMULATION ARP, 局域网仿真ARP) 在LANE中为绑定一个请求ATM地址到MAC地址所使用的ARP。

LEC (LAN EMULATION CLIENT, 局域网仿真客户) 一般位于一个终端系统上, 如ATM主机或LAN交换机上。其任务是维护地址解析表并转发数据业务量。它与一个 ATM地址

唯一关联。

LECS (LAN EMULATION CONFIGURATION SERVER, 局域网仿真配置服务器) 主要功能是为LEC提供配置信息的服务器。

LES (LAN EMULATION SERVER, 局域网仿真服务器) 为LANE地址解析协议 (LE-ARP)提供支持的服务器。LEC在LES上注册各自的ATM和MAC地址。一个LES被一个ATM地址唯一标识。

LIS (LOGICAL IP SUBNET, 逻辑IP子网) 一个ATM子网, 包含共享相同地址前缀的设备并且可以采用ATM PVC或SVC相互通信。

LLC/SNAP ENCAPSULATION (LLC/SNAP封装) RFC1483支持的封装技术, 将AAL帧封装在一个通常包含一个网络层标识符的 8字节LLC/SNAP头中。它允许多协议分组通过单一的VC进行传送。

LNNI (LAN EMULATION NETWORK NODE INTERFACE, 局域网仿真网络节点接口) 一个为在LANE服务器成员 (LES、LECS、BUS)间交换控制信息而定义的协议。

LUNI (LAN EMULATION USER NETWORK INTERFACE, 局域网仿真用户网络接口) 在LEC和网络提供的LANE之间定义的UNI。

MARS (MULTICAST ADDRESS RESOLUTION SERVER, 组播地址解析服务器) 解析IP组播组地址和ATM地址的一种地址解析协议, 从而实现在ATM网络上运行IP组播业务。

MBPS (MEGABITS PER SECOND, 兆比特每秒) 百万比特每秒的传输速率。

MBS (MAXIMUM BURST SIZE, 最大突发尺寸) 一个业务量参数, 指定以峰值速率允许发送的最大信元数。

MCR (MINIMUM CELL RATE, 最小信元速率) 给出业务源在一个虚连接上可以发送信元的最小速率。

MCS (MULTICAST SERVER, 组播服务器) 接收由源端启动的组播业务量并转发 MCS寻根组播树的服务器。

MIB (MANAGEMENT INFORMATION BASE, 管理信息库) 定义了参考变量(如整数和串)对象的一种数据结构。一般包含有关网络管理和性能的信息。

MOSPF (MULTICAST EXTENSIONS FOR OSPF, OSPF的组播扩展) 对OSPF的功能扩展, 支持建立内部和网间组播传递树。

MPEG (MOTION PICTURE EXPERTS GROUP, 运动图像专家组) 一种数字编码、传输和解码的视频技术标准, 兼容现有VCR质量的运动图像。

MPLS (MULTIPROTOCOL LABEL SWITCHING, 多协议标签交换) IETF为综合选路和交换功能而定义的标准化技术。它将一个 FEC的分组与一个标签集相关联。在网络的入口识别FEC和标记分组, 然后传送LSP到网络的出口。

MPOA (MULTIPROTOCOL OVER ATM, ATM上的多协议) 除了IP之外的一组支持分布式选路协议的标准。它在LANE和NHRP上开发, 支持将交换机、路由服务器和主机连接到ATM网络上。

MULTIMEDIA (多媒体) 组合不同形式的信息如文本、数据、图像、视频、音频、图片提供给用户, 如电视会议。

MULTIPOINT-TO-POINT (MPT-PT) SWITCHED PATH (多点到点交换路径) ARIS建立

的交换路径，它以出口 ISR 为根延伸到所有入口 ISR 设备。在相反方向上业务流从树叶返回树根，即出口 ISR。MPT-PT 树利用 VP 或 VC 合并来限制交换资源的消耗。

MUX (MULTIPLEXER, 复用器) 一种本地网络设备，组合多个信息流以共享一条公共的物理信道。

NBMA NETWORK (NONBROADCAST MULTIACCESS NETWORK, 非广播多路访问网络) 连接到一个通用基础结构上的网络，但不具备任何广播能力。例如 X.25、帧中继和 ATM。

NDIS (NETWORK DRIVER INTERFACE SPECIFICATION, 网络驱动程序接口规范) NIC 设备驱动的通称，独立于任何硬件或软件的实现。

NHC (NEXT HOP CLIENT, 下一跳客户) 在需要 NHRP 服务器的主机或路由器上的客户功能。

NHRP (NEXT HOP RESOLUTION PROTOCOL, 下一跳解析协议) 内部子网地址解析协议。

NHS (NEXT HOP SERVER, 下一跳服务器) 运行在路由服务器或路由器上的 NHRP 服务器。

NIC (NETWORK INTERFACE CARD OR CONTROLLER, 网络接口卡或控制器) 工作站和 PC 等 DTE 设备接入网络时所需的硬件通信接口电路板，也称作适配卡。

N-ISDN (NARROWBAND INTEGRATED SERVICES DIGITAL NETWORK, 窄带综合业务数字网络) B-ISDN 的前身，它包含早期为 ISDN 设计的标准。

NMS (NETWORK MANAGEMENT SYSTEM, 网络管理系统) 设置所需硬件和软件参数的一组 OAM&P 功能，用于管理一个网络。

NNI (NETWORK NODE INTERFACE OR NETWORK-TO-NETWORK INTERFACE, 网络节点接口或网络到网络接口) ITU-T 制定的标准接口，位于同一网络内部的节点间。ATM 论坛区分二种标准，一是专用网络下的 PNNI，另一个是公用网络下的公用 NNI。

OAM (OPERATIONS AND MAINTENANCE, 操作维护) 有关网络性能监测、故障检测和系统保护的一组管理和监督措施。

OC-N (OPTICAL CARRIER-N, 光载体-N) ITU-T 为在光纤上以 51.84Mbps 的 n 倍速率进行传输而定义的物理接口，即 OC-3 为 155.52Mbps，OC-12 为 622.08Mbps，OC-48 为 2.488Gbps。

OCTET (8 位的字节) 8 比特或一个字节。

OVERLAY MODEL (叠加模型) 当 IP 路由拓扑独立地运行在如 ATM 的第二层交换拓扑上时使用。它需要定义和维护单独的地址空间、单独的选路协议和单独的拓扑结构。

PAR (PNNI AUGMENTED ROUTING, PNNI 扩充选路) PNNI 协议的扩展，在绑有 ATM 的路由器上运行具有任何其他动态 IP 选路协议的 PNNI 实例。

PAYLOAD (净荷) ATM 信元的组成部分，包含需要承载的实际信息，占用 48 字节。(参见 PTI)

PCR (PEAK CELL RATE, 峰值信元速率) 刻划业务源的一个业务量参数，规定信元可以发送的最高速率。它是在一个给定虚连接上最小信元间隔 (二信元间时间) 的倒数。

PDU (PROTOCOL DATA UNIT, 协议数据单元) 最初用于 OSI 模型中描述穿过不同层的

原语。它包含头、数据和尾信息，是一条消息。

PEAK DURATION (峰值持续时间) 一个源端业务量特性，给出以峰值信元速率发送的持续时间，它等于以信元为单位的突发长度。

PEER MODEL (对等模型) 当网络转发节点运行在单一的IP拓扑结构上采用的模型。该模型支持单一的IP拓扑和单一的IP地址空间并运行单一的IP选路协议。

PHY (PHYSICAL LAYER, 物理层) ATM协议参考模型的底层，分为传输汇聚（TC）和物理介质（PM）二个子层。在连接ATM设备的物理接口上发送ATM信元。

PIM (PROTOCOL INDEPENDENT MULTICAST, 协议独立式组播) 二种截然不同的组播选路协议：PIM密集模式和PIM疏松模式的合称。PIM-DM基于RPM，PIM-SM采用会合点(RP)的概念接收所有源端寻根方式的组播业务量并转发每组RP寻根方式的组播传递树。

PNNI (PRIVATE NETWORK NODE INTERFACE, 专用网络节点接口) 用于通过ATM交换网络转发SVC请求的动态信令和选路协议。

PRI (PRIMARY RATE INTERFACE, 基群速率接口) 一个ISDN规范，在一个DS1或E-1线路上提供23个64Kbps的B信道和一个64Kbps的D信道。

PROXY PAR 为在ATM网络上运行网络互通业务设备而服务的PNNI扩展部分，包括注册、查询和发现机制。

PTI (PAYLOAD TYPE IDENTIFIER, 净荷类型标识符) 考虑AAL和EFCI编码信息用途的一个3比特信元头字段。

PVC (PERMANENT OR PROVISIONED VIRTUAL CONNECTION, 永久或预定虚连接) 在ATM网络中为不确定应用而预定的虚连接（VPC/VCC），由网络管理系统（NMS）建立。

QOS (QUALITY OF SERVICE, 服务质量) 刻划一个给定虚连接上业务量特性的一组ATM性能参数。这些参数包括CLR、CER、CMR、CDV、CTD和平均信元传输延迟。

RFC (REQUEST FOR COMMENT, 请求评注) 包含建议、标准和规范的草案。RFC只作为历史建议或存档备案。

ROUTING (选路) 一种网络管理功能，负责从源端到目的端转发分组。存在多种算法满足各种网络拓扑和要求。

RPF (REVERSE PATH FORWARDING, 反向路径转发) 组播转发协议，在所有接口上转发分组（除了在反向网络源端最短路径的接口上）。

RPM (REVERSE PATH MULTICASTING, 反向路径组播) 通过从传递树上除掉没有组成员的线路和接口来优化网络资源利用率的RPF算法的扩展功能。

RSVP (RESERVATION PROTOCOL, 资源预留协议) 用于保留通过网络路由器和主机资源（如带宽）的一种协议。RSVP是IP集成服务模型的一个组件。

RTCP (RTP CONTROL PROTOCOL, RTP控制协议) 在一个实时会话中为参与者提供会话管理信息的附加控制协议。

RTP (REAL-TIME TRANSPORT PROTOCOL, 实时传输协议) 一个应用层协议，提供序列编号、净荷类型标识符、时间戳和源标识符。源端可以发送实时数据（如语音和视频）到目的端，目的端可以按照正确顺序和正常时间进行重放。

RTT (ROUND-TRIP TIME, 往返时间) 在源端和设备(如交换机)之间的往返时间。通常根据设备的缓冲能力以信元的数目测量，用于流量控制中窗口的计算。

SAAL (SIGNALING AAL SERVICE, 信令AAL业务) AAL协议中的信令部分。ITU-T借鉴N-ISDN开发了其规范。

SAR (SEGMENTATION AND REASSEMBLY, 分段和组装) AAL的下半层。负责将数据从信息帧插入到信元中, 添加一些必要的头或尾开销比特, 形成 48字节送到ATM层。每一种AAL类型都有自己的SAR格式。在目的端, 信元的净荷被提取出来然后转换成合适的协议数据单元。

SAR-PDU (SEGMENTATION AND REASSEMBLY PROTOCOL DATA UNIT, 分段和组装协议数据单元) SAR子层与ATM层交换的48字节协议数据单元。包含 SAR-PDU净荷和一些SAR子层可能添加的控制信息。

SCR (SUSTAINABLE CELL RATE, 可维持信元速率) 刻画一个突发源的业务量参数, 指定在一个给定VC上信元可以发送的最大平均速率。定义为 MBS和最小突发间隔时间之比。

SCSP (SERVER CACHE SYNCHRONIZATION PROTOCOL, 服务器缓存同步协议) 一种服务器间缓存同步和复制的协议。SCSP允许基于服务器的协议如 LANE、ATMARP、NHRP、MARS和MPOA通过多台服务器分布网络互联服务。

SDH (SYNCHRONOUS DIGITAL HIERARCHY, 同步数字系列) 为光纤线路上的高速数字传输而制定的信号接口系列。

SEAL (SIMPLE EFFICIENT ADAPTATION LAYER, 简单有效的适配层) AAL5的初始名称和建议。

SERVICE TYPES (服务类型) 有四种服务类型: CBR、VBR、UBR和ABR。CBR和VBR是有保证服务, 而UBR和ABR被认为是尽力而为服务。

SHORTCUT PATH (直通路程) 在不同网络实体间建立的旁路掉任何中间路由跳的第二层交换路径。直通路程的建立和维护是所有 IP交换协议的主要功能目标。

SMDS (SWITCHED MULTIMEGABIT DIGITAL SERVICE, 交换多兆比特数字业务) 一种基于53字节分组的无连接网络互联业务, 目的是将不同的局域网接入交换公用网络。

SMTP (SIMPLE MAIL TRANSFER PROTOCOL, 简单邮件传输协议) 为支持电子邮件(email)而开发的协议标准。

SNA (SYSTEMS NETWORK ARCHITECTURE, 系统网络体系结构) IBM引入的一种基于主机的网络体系结构, 它在端点间建立逻辑信道。

SNMP (SIMPLE NETWORK MANAGEMENT PROTOCOL, 简单网络管理协议) IETF为处理管理信息而制定的标准。通常作为用户数据报协议(UDP)的一个上层应用。

SONET (SYNCHRONOUS OPTICAL NETWORK, 同步光网络) ANSI为高速、高质量数字光纤传输制定的标准。北美的标准为SDH。

SS7 (SIGNALING SYSTEM NUMBER 7, 七号信令系统) CCITT制定的公共信道信令标准, 目的是为ISDN提供内部控制和网络智能化功能。

SSCOP (SERVICE SPECIFIC CONNECTION-ORIENTED PROTOCOL, 面向连接特定服务协议) SAAL中SSCS的一部分。SSCOP是一个端到端协议, 通过在发送方和接收方之间重传和状态报告来提供错误发现和校正, 从而保证传递的完整性。

STM (SYNCHRONOUS TRANSFER MODE, 同步传输模式) 一种分组交换方式, 时间轴被分割成时间槽再分配给单个信道, 用户可以定期地发送数据。时间槽基本上表示全部可

用带宽的给定部分(参见TDM)。

STM-1 (SYNCHRONOUS TRANSPORT MODULE-1, 同步传输模块-1) ITU-T为在ATM中以155.52Mbps速率进行数字传输而定义的SDH物理接口。

STM-N (SYNCHRONOUS TRANSPORT MODULE-N, 同步传输模块-N) ITU-T为在ATM中以STM-1的n倍速率进行数字传输而定义的SDH物理接口。

STP (SHIELDED TWISTED PAIR, 屏蔽双绞线) 二条相互隔离的铜线绞和并包上保护屏蔽层。

STREAM (径流) 通过同一路径转发并以相同方式处理的分组径流。一条径流可能包含一条或多条流(flow)。在MPLS体系结构中一条径流由一个径流成员描述符(SMD)确定。

STS-1 (SYNCHRONOUS TRANSPORT SIGNAL-1, 同步传输信号-1) 以51.84Mbps速率光纤传输的SONET信号标准。

STS-N (SYNCHRONOUS TRANSPORT SIGNAL-N, 同步传输信号-N) 以基本STS-1信号的n倍速率(即STS-3为155Mbps)发送的SONET信号格式。

SVC (SWITCHED VIRTUAL CONNECTION, 交换虚连接) 一种通过信令动态建立和拆除的连接。

T1 一种速率为1.544Mbps的TDM数字通道。

T3 一种速率为44.736Mbps的TDM数字通道, 可以复用28个T1信号, 通常表示为DS-3。

TAG SWITCHING (标签交换) 由Cisco建议的拓扑驱动的IP交换解决方案。

TAXI (TRANSPARENT ASYNCHRONOUS TRANSMITTER/RECEIVER INTERFACE, 透明异步发送器/接收器接口) 在速率为100Mbps的多模光纤上提供连通性的接口。

TCP (TRANSMISSION CONTROL PROTOCOL, 传输控制协议) 基于IP设备的一种可靠的、面向连接的传输协议。

TCP/IP 互联网协议集, 用于互联Internet和企业网上的主机和网络。

TDM (TIME-DIVISION MULTIPLEXING, 时分复用) 将全部带宽(链路容量)分割成若干个允许比特流相联接(复用)的若干个信道的技术。将时间轴划分为固定长度的时间槽, 一个特定的信道只能在一个特定的时间槽期间传送。

TDP (TAG DISTRIBUTION PROTOCOL, 标签分发协议) 用于标签交换设备之间分发路由/标签绑定的控制协议。

TM (TRAFFIC MANAGEMENT, 业务量管理) 提供连接接纳控制(CAC)和流量拥塞控制的方法, 如UPC和业务量整形。

TOPOLOGY-DRIVEN IP SWITCHING (拓扑驱动的IP交换) 基于路由表内容建立直通路经的一种IP交换解决方案, 如ARIS和Tag Switching。

TOS (TYPE OF SERVICE, 服务类型) 包含在IPv4头中的8比特字段, 由IETF DiffServ工作组重新定义以指定一个单独的分组服务优先级, 已被重新命名为DS字节(或字段)。

TRAFFIC CLASS (业务量分类) 保证业务、可控负载和尽力而为业务量分类的IntServ术语。

TRAFFIC CONTRACT (业务量合同) 考虑到期望网络提供的QoS和用户对预设业务量参数(即PCR、MBS、突发和平均信元速率)的认可, 而在用户和网络之间达成的协定。

TRAFFIC DESCRIPTORS (业务量描述符) 描述源端业务量的一组参数, 在ATM中有

PCR、MBS、CDV和SCR，在IP IntServ中有Tspec/Rspec。

TRAFFIC SHAPING (业务量整形) 对非一致业务量(即违反业务量合同中指定的 PCR、CDV和MBS业务量参数)进行管制的方法。

UBR (UNSPECIFIED BIT RATE, 未规定比特率) 主机可以按峰值信元速率发送的尽力而为服务。对带宽、时延或其他 QoS参数没有任何保证。

UDP (USER DATAGRAM PROTOCOL, 用户数据报协议) 没有任何分组顺序或传递保证的无连接传输协议，直接运行在 IP层上。

UNI (USER-NETWORK INTERFACE, 用户网络接口) 在CPE(用户)和ATM网络(ATM交换机)间定义的一组协议和业务量特性的接口。ATM论坛制定了二个标准，一个是在用户和公用ATM网络之间，称作公共UNI；一个是在用户和专用ATM网络之间，称作P-UNI。

UPC (USAGE PARAMETER CONTROL, 使用参数控制) 一种业务量控制方式，用于检查和加强业务量合同和QoS参数的用户一致性。通常称作业务量管制，工作在 UNI级别。

UTP (UNSHIELDED TWISTED PAIR, 非屏蔽双绞线) 没有任何包装外套的双绞线，用于短距离连接。

VBR-NRT (VARIABLE BIT RATE-NONREAL TIME, 非实时可变比特率) 一类支持不要求定时信息业务流的服务类型，适于长数据分组传输。

VBR-RT (VARIABLE BIT RATE-REAL TIME, 实时可变比特率) 提供给依赖定时信息业务流的一种服务类型，业务流由平均和峰值速率刻画，适于承载分组化的(压缩)视频和音频。

VC (VIRTUAL CHANNEL, 虚信道) 描述在连接(交换或终端用户) 点间共享一个公共标识符数值的单向ATM信元流。

VC MERGE (VC合并) 属第二层交换机功能，允许多个上游 VC流合并进一个下行VC流而不用交互来自不同帧的信元。一个具有 VC合并功能的交换机必须搜集入站 VC上一个帧的所有信元并在出站链路上以一个新的 VCI值按照连续方式发送信元。在新 VC上发送的信元必须连续按帧成组以使目的端设备区分来自不同源的信元。VC合并可能要求交换机上具有专用硬件或每个VC的缓存能力。

VC MULTIPLEXING ENCAPSULATION (VC复用封装) RFC1483支持的封装技术，使得AAL帧在VC端点间传输。它需要为每一个网络层协议提供一个单独的 VC。

VCI (VIRTUAL CHANNEL IDENTIFIER, 虚信道标识符) ATM信元头中的一个16比特字段，为VC提供唯一标识符，可作为标签用于IP交换技术。

VCID (VC IDENTIFIER, VC标识符) 与一个特定VC流唯一联系的数值。

VIRTUAL CIRCUIT (虚电路) 在一个源端和一个目的端间穿网络建立的一个连接，其中，一个固定路由被选择用于整个通话过程，带宽动态分配。

VIRTUAL CONNECTION (虚连接) 在端用户(源端和目的端)间建立的一个连接，分组沿同一路径转发，并且带宽不是永久分配直到其被使用。

VIRTUAL IP SWITCH (虚拟IP交换机) 增加了一个IP路由业务的交换机的一个集合。它的组件包括入口和出口边缘交换机，一个运行路由协议并提供缺省 IP转发的路由服务器以及交换机，典型的如ATM。虚拟IP交换机的一个例子是MPOA虚拟路由器。

VLAN(VIRTUAL LAN, 虚拟LAN) 一个联网环境。其中，物理上相互独立的 LAN中的

用户以这样一种方式互连，使得它们看起来如同在同一 LAN 工作组中。

VLSM (VARIABLE LENGTH SUBNET MASK, 可变长子网掩码) 一个单一的路由系统支持可变长掩码的能力，以将主机地址空间容量与一个子网的实际主机数目进行最好的匹配。RIPv2、OSPF和E-IGRP是支持VLSM的路由协议。

VP (VIRTUAL PATH, 虚通道) 在交换机间共同组合的一组虚信道 (VC)。

VP (VP合并) 使得多个上游虚通道 (VP)合并为一个单一的下游 VP的标准ATM交换机功能。入口和出口 VCI值保持唯一不变，从而不需要交换机具有特殊的 VC合并功能。唯一的 VCI值用于保持源端(如入口ISR)的身份。但是，由于VCI值可以共享一个公共VPI,所以必须将VCI值的一个特定范围告知入口ISR设备以防止VCI冲突。

VPCI/VCI (VIRTUAL PATH CONNECTION IDENTIFIER/VIRTUAL CHANNEL IDENTIFIER, 虚通道连接标识符/虚信道标识符) 两个数字的结合，一个标识 VP，一个标识 VCI。它是在ATM数据链路上运行的IP交换解决方案的一个标记。

VPI(VIRTUAL PATH IDENTIFIER,虚通道标识符) 在信元头的一个8比特值，标识VP及相应的信元所属的VC。VC可以作为IP交换技术中的一个标记使用。

VPN(VIRTUAL PRIVATE NETWORK,虚拟专用网) 叠加在公众Internet上的一个逻辑网络。使用IPsec 协议形成的内部VPN是安全的信道。

WAN(WIDE AREA NETWORK,广域网) 覆盖长距离范围并且通常使用公众电话电路的网络。