第 3 章　Internet 组播

3.1　引　　言

Internet 网络传输和处理能力的大幅提高使得基于网络的新应用越来越多，特别是音频和视频压缩技术的发展和成熟，使得音频和视频应用成为 Internet 上最重要的应用之一，出现了如视频点播、视频会议、远程学习和计算机协同工作等以多媒体为特征的新应用。与一般的网络应用相比，Internet 的视频点播、可视电话和视频会议等多媒体应用有着数据量大、时延要求高、持续时间长等特点，因此要解决这些应用所要求的传输带宽大、实时性强等问题，需要采用不同于传统单播和广播机制的转发技术来实现。而组播技术正是解决这一问题的理想方案。组播是一种点到多点和多点到多点的通信方式，即多个接收者同时接收一个组播源发送的相同信息，它能够有效地利用网络带宽，提高网络资源的利用率。

3.1.1　计算机网络中的通信方式

计算机网络中主要包括单播、组播、汇播、任意播和广播 5 种通信方式。

1. 单播(Unicast)

单播是目前计算机网络中使用最普遍的通信方式，所谓单播就是点到点的通信方式。这种通信方式只涉及一个发送端和一个接收端。请注意，在网络中提到单播，大家一般默认数据是单向传输的，但是控制数据的传输是双向的。在这种通信方式下，从一台主机发出的每个数据包只能发送给一台主机。Web 浏览和 FTP 文件传输都是常见的单播应用。如果有多台主机想从一个发送端得到同样的信息，使用单播通信方式，发送主机必须向每个主机单独发送多份同样数据的副本。从此可以看出使用单播实现群组通信，可扩展性不好。举例来说，有 n 台主机要进行群组通信，如果采用单播的方式，那么就要建立 $\frac{n(n-1)}{2}$ 个单播连接。这种发送端发送冗余信息，同一信道上传送多份同样信息的副本的方式，不仅给发送端主机带来沉重的负担，同时又占用了大量的网络带宽。

2. 组播(Multicast)

组播(也称多播)技术是一种允许一台主机(称为组播源或发送端)一次同时发送单一数据分组到多台主机(称为接收端)的技术。组播作为一点对多点的通信方式，是节省网络带宽的有效方法之一。在网络音频/视频广播等多媒体应用中，当需要将一个节点的数据分组传送到多个节点时，无论是采用重复单播通信方式还是广播方式，都会严重浪费网络带宽，而只有组播才是最好的选择。这是因为组播能使一个或多个组播源只把数据分组发送给特定的组播组，只有加入该组播组的主机才能接收到数据分组。目前，组播技术被广泛应用在网络音频/视频广播、视频点播、网络视频会议、多媒体远程教育、推送(push)技术(如股票行情等)和虚拟现实游戏等方面。组播和单播的比较见图 3.1，图中箭头表示信息流。从图 3.1 中可以看出，组播和单播相比节省了网络带宽，减轻了服务器的负载。

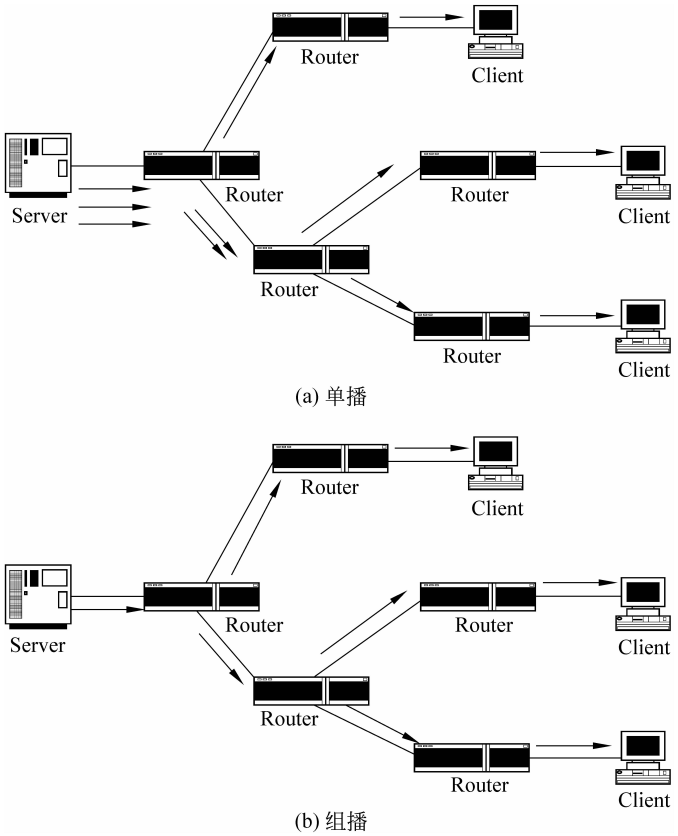

图 3.1　组播和单播的比较

目前文献中提到组播一般是指单点到多点的通信方式。多点到多点的通信方式是组播的扩展方式，这种方式一般被称为**多点通信**或**群播**(Multipoint Communication)。

3. 汇播(Concast)

汇播这种通信方式是指多个发送端(主机)将信息发送给一个接收端(主机)，这是一种 $n:1$ 的通信方式[1]。Concast 的核心思想是能够让接收端通过尽可能少的分组掌握尽可能多的信息，因此其关键就是如何在网络的中间节点合并多个发送端发送来的分组。目前的 Internet 还不支持汇播，因此研究人员建议采用主动网络技术实现汇播。使用 Concast 一个例子是在仿真系统中，多台主机将仿真结果发送给一台主机进行仿真结果的分析。另外一个 Concast 的例子是在网管系统中，多台主机将管理数据发送给主管理站点。这种通信方式也经常用在远程教育中，所有学生可以远程提交作业给指导老师。Concast 还可以用在可靠组播中解决反馈爆炸问题，多个接收端发回的确认分组可以在中间节点进行合并。

4. 任意播(Anycast)

任意播是指源主机把数据传输给由任意播地址标识的一组目的主机中最近的一个。使用任意播的应用只需要把任意播分组发送给一组主机中的任意一个就可以，而并不关心到底是哪台主机收到了分组。如图 3.2 所示，两台服务器以任意播地址 A 标识，它们分别位于不同的网络区域提供相同的服务。网络中的路由器将自动地将客户的服务请求发送到最近的服务器。

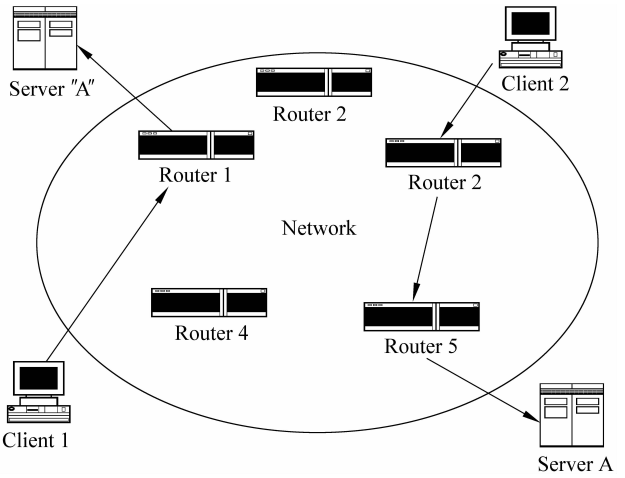

图 3.2　任意播功能示意图

5. 广播(Broadcast)

广播是指某台主机将信息发送给网络中所有其他的主机。广播是局域网络中经常使用的通信方式。单播和广播都可以看做是组播的特殊形式。广播的主要缺点是每次广播都要发送数据到所有主机,不管这些主机是否需要此数据,这样不仅浪费了网络带宽,同时也消耗了所有主机的资源。当广播出现频繁时,往往会严重影响网络的性能,该问题通常被称为"广播风暴"。

3.1.2　组播发展的历史、现状及挑战

1988 年,当时在斯坦福大学攻读博士学位的 S. E. Deering 发表了组播的经典论文"Multicast Routing in Internetworks and Extended LANs[3]",奠定了 IP 组播的技术基础。S. E. Deering 也因为在组播和 IPv6 方面的重要贡献获得 2010 年 IEEE Internet Award。

1988 年,D. Waltzman、C. Portridge 和 S. E. Deering 完成了 RFC 1075,它是组播路由协议的首次实践。

1991 年 12 月,S. E. Deering 完成了他的博士论文"Multicast Routing in a Datagram Network",它奠定了组播网络体系结构和路由协议的基础。该文也成为 Internet 组管理协议(IGMP)的原型。

1994 年 3 月,形成了对 OSPF 协议的扩展协议 MOSPF(RFC 1584)。

1996 年 11 月,出现了对于基于 UNI3.0/3.1 的 ATM 组播网络支持协议(RFC 2022)。

1997 年 9 月,有核树(CBTv2)组播路由体系结构形成(RFC 2189)。

1997 年 11 月,组管理协议 IGMPv2 得到 IETF 的批准,成为标准(RFC 2336)。

1998 年 6 月,评估可靠组播传输协议 RMTP 的 IETF 标准出台(RFC 2357)。

1998 年 7 月,在制定 IPv6 地址体系标准时,确定 IPv6 组播地址分配方案(RFC 2373),这为组播技术在下一代 Internet 上的应用做出了必要的准备。

1999 年 10 月,Cisco、AT&T 和 Microsoft 制定组播地址动态客户分配协议 MADCAP(RFC 2730)。

2000 年底 2001 年初，人们着手制定各种组播 MIB 库，这标志组播技术正向可管理、可控制方向发展。

3.1.3 IP 组播技术的优缺点

与单播和广播通信方式相比，组播具有只发送单个数据包就可以将信息传送到所有需要它的接收者手中的特点，因此可以有效提高网络的工作效率，缓解网络瓶颈，可广泛应用于多媒体应用、网络游戏和任意点到多点的信息推送等场景。

归纳起来，组播主要有以下优点。

（1）**节约带宽**。运用组播技术发送数据常常能从根本上减少整个网络的带宽需求。当多个用户要求同一服务器提供同样信息时，如果使用单播技术，带宽消耗将随用户的增多不断增加；而对于组播，由于在共用链路上只传递信息的一份副本，因此带宽的需求并不会随用户数量的增加而增加。

（2）**减轻服务器负载**。对于网络上的许多应用，常常有一定数量的用户在接收完全相同的数据流。如果采用 IP 单播技术来为这些用户服务，需要发送者为每个用户单独建立一个数据流，由于这些数据流重复地发送完全相同的数据，所以将大大加重发送主机和通信网络的负载，同时也难以保证对不同接收者的服务公平性。举例来说，利用音频服务器传送一个无线节目给互联网上的实时连接用户。如果使用单播传送机制，由于服务器必须为每一个收听节目的用户发出各自的数据分组，随着用户数量的增加，需要不断增加实时音频服务器的能力和数量。如果使用组播来发布节目，服务器只需要发布单个实时数据流。用这种方式，不需要购买越来越多的高性能实时音频服务器以适应用户数目的增长。

（3）**减轻网络负载**。当将相同的内容传送给多个用户时，组播能明显地减少带宽要求，带宽消耗的降低等同于路由器上的负载降低。但在某些情况下，在特定点工作的路由器的负载可能会增大。请再参考图 3.1 的例子，我们知道第一跳路由器（直接与服务器相连的路由器）从服务器接收一个数据流。然而要注意的是，第一跳路由器把单个数据流复制成两个输出数据流以便将该数据送交给下游用户。这个复制过程增加了路由器的工作负载，在网络设计中需要考虑这个因素。如果路由器没有有效的复制机制，则当输出接口数很大时该路由器负载将明显增加。

组播在具备了以上的优点的同时，也存在一些由于自身特点所带来的缺点。

（1）**组播缺乏可靠性保证和拥塞控制机制**。由于组播是一对多的传输方式，无法直接使用面向单播的可靠传输协议 TCP 来保证数据的可靠传输和流量控制，而且由于组播应用往往传输的是视频流，因此现有的组播应用传输组播数据时通常采用 UDP 协议。UDP 协议是一种尽力而为的协议，这意味着数据的传输可能发生丢失、乱序以及重复到达等。因此如果要实现组播的可靠传输，就需要在应用层设计方案或通过一种在 UDP 之上的可靠组播协议来实现。但是和单播相比，可靠组播实现相对困难。更为严重的问题是，组播传输目前缺乏有效的拥塞控制机制。组播数据是基于 UDP 这种没有拥塞控制机制的协议进行传输的，如果组播本身不采用拥塞控制机制，那么组播数据流就很可能占满网络带宽，使网络中的 TCP 流量难以获得足够的带宽，造成对 TCP 流的不公平。组播拥塞控制机制是目前组播研究的一个难点问题，组播拥塞控制有两个重要的目标：可扩展性和 TCP-Friendly。可扩展性是指随着组规模的增大，拥塞控制协议不会造成组播性能下降。TCP-Friendly 则

要求组播和 TCP 流量公平地竞争网络带宽。

(2) **组播缺乏足够的安全性**。安全组播指的是只有注册过的发送者才可以向组发送数据;并且只有注册的接收者才可以接收组播数据。然而目前的 IP 组播很难保证这一点。因为 IP 组播使用无连接的协议 UDP。UDP 协议不使用肯定确认或否定确认机制来确保可靠传送,并且组播也不能被防火墙检测到,因此无法对组播进行安全认证。其次,Internet 缺少对于网络层的访问控制。除此以外,组成员可以随时加入/退出组播组的动态性使得对组成员的安全联盟的建立非常复杂,它必须能够根据组成员的变化进行动态的更新。以上这几点使组播安全问题同组播的可靠性问题一样难以解决。

(3) **组播缺乏有效的用户管理功能**。具体表现为:①认证难:组播协议不提供用户认证功能,用户可随意地加入或离开;②计费难:组播协议不涉及计费,加上组播源无法得知用户何时加入或离开,也无法统计某时间段到底有多少用户在接收组播数据,因此无法进行准确的计费;③管理难:组播源缺乏有效的管理手段去控制组播信息在网上传递的范围和方向。

(4) **组播实现复杂**。由于组播组成员分布在网络的不同地方,通过不同的链路和互连设备相连,并且接收者自身的处理能力不同。当所有接收者都要与同一组播源交互时,就必须采取某些方法使得每一个接收者接收到与其接收能力和从组播源到接收者之间带宽相适应的数据流。网络的异构性导致了组播应用的实现复杂性。所以在设计和实现组播时,必须充分考虑到网络的异构性特点。

按照传输数据的性质,组播应用可以分为图 3.3 所示的 4 类。

图 3.3　组播业务分类

(1) **实时多媒体应用**:这类应用主要包括视频会议。一种视频会议属于多点对多点的应用,在这种视频会议中,每个参会者既可以发言,也可以当听众。另一种视频会议属于一点对多点的应用,例如 IETF 使用 MBone 进行大会的网络直播。无法亲自到会的人可以作为听众通过网络听到参会者的报告。实时多媒体业务不需要很高的可靠性,但是对延时抖动的要求较高,因为在这种应用中,要求数据能够以一种平稳的速率进行发送,以便使图像看上去更自然,并且口形和声音要保持一致。

(2) **实时数据应用**:实时数据应用种类很多,其中一种典型的应用是电子白板,参会者可以在此白板上写字、绘图以及通过其他的书面形式和其他参会者共享和交流信息。在这种应用中要保证数据的无差错传输以保证白板上的数据的正确性,因此这种应用要求延时低。第二类实时数据应用是网络游戏。基于单播的网络游戏已经存在于 Internet,但是组播非常适合网络游戏或者仿真应用的使用,参与的计算机只需进入 IP 组播组就开始发送和接收游戏及仿真数据。

(3) **非实时多媒体应用**:非实时多媒体应用的例子如远程教育,此外最近涌现出的如

Web 页面的 Cache 技术和镜像技术。利用镜像和 Cache 技术能够使高品质的多媒体传输成为可能。这类应用对时延的要求不高。

(4) **非实时数据应用**：许多非实时数据业务需要高可靠性。一个典型应用是通过软件中心向各个网点进行软件发布和软件更新。此外推送业务也是一类新兴的使用组播技术的非实时数据应用，如新闻标题、天气变化等信息的发布，它们要求的带宽较低，对延时的要求不高。

3.2 组播地址

IP 组播地址指定了一个 IP 主机的集合，集合内的主机属于同一个组播组，拥有同样的组播地址，并接收发向该组播组的数据。需要注意的是，组播地址具有其特殊性，即组播地址对应于一个逻辑组播组，它并不代表每个组成员的实际网络位置，不能像单播地址一样进行聚合。以下介绍组播地址结构，以及它们和第二层组播地址的对应关系。

3.2.1 IPv4 组播地址

互联网名称与数字地址分配机构(ICANN，The Internet Corporation for Assigned Names and Numbers)控制着 IP 地址的分配。ICANN 将 D 类地址分配给组播，所有的组播地址都位于 224.0.0.0～239.255.255.255 的地址范围内[4]。IPv4 组播地址又分为保留的局部链路地址、全局范围地址、限制范围地址和 GLOP 地址。

1. 保留的局部链路地址

ICANN 保留的 224.0.0.0～224.0.0.255 范围内的 IP 地址给局部网段上的网络协议使用。目的地址为局部链路地址的数据包的生存时间(TTL，Time-To-Live)通常都设为 1，这样就不会被路由器转发。网络协议使用这些地址进行重要路由选择信息的传递。例如 OSPF 使用 IP 地址 224.0.0.5 和 224.0.0.6 来交换链路状态信息。表 3.1 列出了一些专用的局部链路地址。

表 3.1 专用组播局部链路地址

IP 地址	用　途	IP 地址	用　途
224.0.0.1	子网上所有主机	224.0.0.6	OSPF 代表路由器
224.0.0.2	子网上所有路由器	224.0.0.12	动态主机配置协议服务器
224.0.0.5	OSPF 路由器		

2. 全局范围地址

224.0.1.0～238.255.255.255 范围内的地址称为全局范围地址，这些地址用来在组织之间以及域间进行组播数据传递。在这个范围内的一些地址已经被预留给 ICANN 的一些组播应用。例如，IP 地址 224.0.1.1 预留给**网络时间协议**(Network Time Protocol，NTP)。

3. 受限范围组播地址[5]

239.0.0.0～239.255.255.255 范围内的地址被称为受限范围地址(Limited Scope

Address)或者受限管理范围地址(Administratively Scope Address)。这些地址被限制在一个本地组或者组织范围内使用。公司、大学或其他组织可以使用这类地址来开展局域网内的组播应用。这类应用的数据不会被转发到其所属的范围以外。在路由器上,通常配置过滤器来防止该组播地址范围内的组播数据包传递到自治系统之外。在一个自治系统或者自治域范围内,受限范围地址可以进一步进行划分,以便定义局部组播的边界,这种子划分称为地址范围限定(Address Scoping),它使得地址可以在多个小区域间重用。

4. 分配给 SSM 协议的地址[6]

232.0.0.0/8 范围内的地址预留给 SSM(Source-Specific Protocol Independent Multicast)协议。

5. GLOP 地址[7]

RFC 3180 提出将 233.0.0.0/8 地址范围预留给那些已经获得预留自治系统(AS)编号的组织使用,这些地址是静态分配的。GLOP 地址的构成方式如下:将自治域的 AS 编号嵌入 233.0.0.0/8 的第二个和第三个字节中。举例来说,对于自治域 AS 5662,将其转换为二进制是 0001011000011110,然后将此二进制值的高 8 位放入 233.0.0.0/8 的第二个字节,将低 8 位放入 233.0.0.0/8 的第三个字节,就得到了可以在这个自治域使用的组播地址 233.22.30/24。

3.2.2 组播 MAC 地址

现在简单回顾 S. Deering 为何在 1988 年提出 IP 组播。当时 Deering 在一个被称为 Vsystem 的分布式操作系统项目组从事研究工作,该项目要求系统内计算机能够利用以太网组播把消息分发给由不同计算机构成的组播组。随着项目的进展,校园另一侧的一组主机也需要加入系统,这时 S. Deering 就面临如何把以太网组播扩展到 IP 网络层次的问题。

首先面临的问题是如何实现该组主机的 IP 地址和 MAC 地址的映射。可能是由于经费的原因(IEEE MAC 地址空间必须付费使用),可以用于 IP 组播的 MAC 地址前缀固定为 01-00-5e (16 进制),也就是说 48 位的 MAC 地址中仅剩余后 24 位可用,而且这 24 位的第一位被固定预置为 0,因此实际只有 23 位的 MAC 地址空间。前面介绍过,IPv4 组播地址空间是 28 位的,因此以太网卡就必须处理 28 位 IP 组播地址到 23 位 MAC 组播地址的映射,其结果如图 3.4 所示,32 个 IPv4 组播地址映射到同一 MAC 地址上。

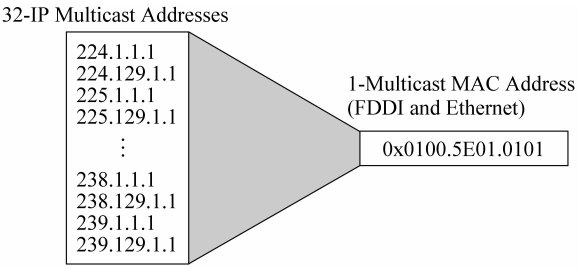

图 3.4 IP 组播 MAC 地址匹配

3.3 Internet 组管理协议 IGMP

在组播中，组的概念是十分重要的。由组播的定义可知，组播数据是从一个源发送到一组主机。在 IP 组播中，这一组主机用一个 IP 组播地址标识，它指定了发送报文的目的组。若一个主机想要接收发到一个特定组的组播报文，它需要监听发往那个特定组的所有报文。为解决 Internet 上组播报文的选路，主机需通过通知其子网上的组播路由器来加入一个组，组播中采用 **Internet 组管理协议**(Internet Group Management Protocol，IGMP)[8][9][10]来实现组播组的管理。

IGMP 协议目前有 3 个版本：IGMPv1[8]、IGMPv2[9]和 IGMPv3[10]。主机使用 IGMP 消息通知本地的组播路由器它想接收组播数据的主机组地址。如果主机支持 IGMPv2，它还可以通知组播路由器它退出某主机组。组播路由器通过 IGMP 协议为其每个端口都维护一张主机组成员表，并定期探询表中的主机组的成员，以确定该主机组是否仍然存在。

IGMP 消息用 IP 分组进行传送。IGMPv1 中定义了两种消息类型：主机成员询问和主机成员报告。当某主机想要接收某个组播组的数据时，它向本地的组播路由器发送“主机成员报告”消息，告知欲接收的组播地址。组播路由器收到“主机成员报告”消息后把该主机加入指定的主机组，并在设定的周期内向组播地址 224.0.0.1(代表所有支持组播的主机)发送“主机成员询问”消息。主机如果还想继续接收组播数据，必须发送“主机成员报告”消息。

IGMPv2 与 IGMPv1 的不同是它将版本字段和消息类型字段融合，把未使用字段作为“最大响应时间”字段。IGMPv2 报文的消息类型字段定义了 4 种消息类型：成员询问、IGMPv1 成员报告、IGMPv2 成员报告和退出主机组。IGMPv2 向前兼容 IGMPv1 协议，IGMPv1 的设备可以接收处理 IGMPv2 的消息报文。IGMPv2 中允许路由器对指定的主机组地址作“成员询问”，非该组的主机不必响应。如果某主机想退出，它可以主动向路由器发送“退出主机组”消息，而不必像 IGMPv1 中那样只能被动退出。

IGMPv3 和 IGMPv2 相比，最重要的是增加了源过滤功能，即不仅允许主机指定其需要接收的特定组的数据，还可以指定接收这个组中特定源的数据。主机可以通过 INCLUDE 和 EXCLUDE 两种模式来指定接收范围。使用 INCLUDE 模式时，主机给出其希望接收的源 IP 地址的列表；使用 EXCLUDE 模式时，则给出其不希望接收的源 IP 地址的列表。路由器根据主机指定的范围转发组播数据。在图 3.5 所示的实例中，组播组 224.1.1.1 的成

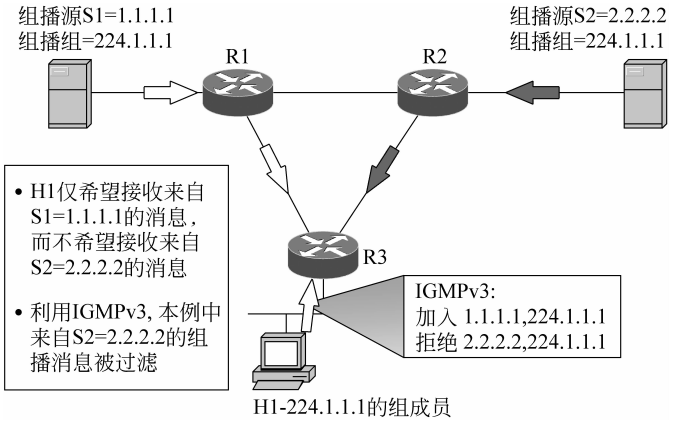

图 3.5 IGMPv3 的源过滤功能

员 H1 仅接收了来自组播源 S1＝1.1.1.1 的数据，而拒绝接收来自组播源 S2＝2.2.2.2 的数据。

3.4 组播转发

在单播转发中，从源节点到目的节点的转发路径一般是一条最短路径。与单播不同的是，组播在进行数据转发时，要将同一信息发送到不同的接收端，为了节约网络带宽和减少复制信息的次数，应该尽量使得在同一链路上信息只复制一次。因此组播的转发结构通常采用树型结构，在这种结构下，组播包的复制只在树的分支处进行，这样可以使全网范围的分组复制数量达到最少。组播组的成员可以随时加入或离开组播组，当位于某个分支上的所有接收者都不再接收发往某个组播组的数据时，路由器便将该分支从转发树上剪去，并停止沿该分支转发数据包。如果这个分支上的接收者又被激活并要求接收组播组数据，路由器就会动态地修改转发树，重新开始向该分支转发组播数据。本节介绍组播技术通常采用的两种转发树结构：源树和共享树，以及组播的逆向路径转发原理。

3.4.1 源树

源树是最简单的组播转发树结构。它的根是组播组的源节点，各个枝干形成一棵覆盖网络中所有组播组成员的生成树。这种树使用网络中的最短路径，即转发树上从源节点到各个组成员的路径都是最短路径，在数学上，这种树也被称为**最短路径树**(Shortest Path Tree，SPT)。在最短路径树上的每个路由器要保存的路由状态是(S，G)，这里 S 代表组播源，G 代表组播组地址，也就是对每个组播组的每个组播源都要建立一棵源树。图 3.6 表示的是组播源 S1 建立的到接收端 1 和 2 的源树。图 3.7 表示的是组播源 S2 建立的到接收端 1 和 2 的源树。图中箭头表示沿已建立的 SPT 传输的组播信息流。

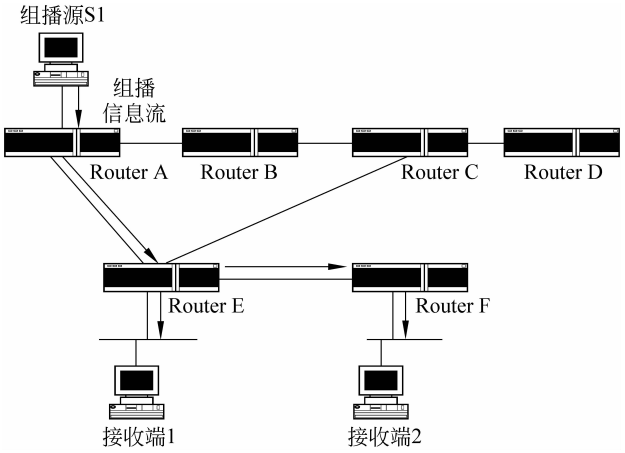

图 3.6　组播源 S1 建立的源树

基于源树的转发结构的好处在于不同组播源发出的数据包被分散到各自分离的组播树上，因此有利于网络中数据流量的均衡。同时，因为从源节点到各个组播组成员之间的路径是最优的，所以端到端(End-to-End)的时延性能较好，有利于流量大、时延性能要求较高的

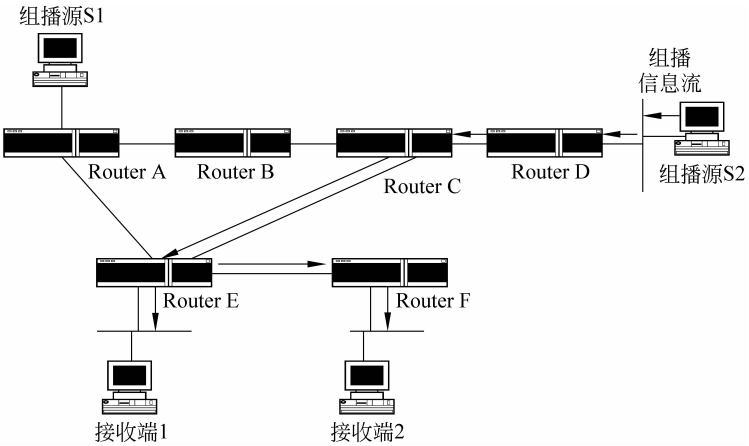

图 3.7　组播源 S2 建立的源树

实时媒体应用。但是这种源树的转发结构的可扩展性不是很好，如果一个组播组有 N 个发送源，那么就要为每个源节点建立一棵源树，同时，每台路由器都要保存 $N \times M$（M 是组播组的数量）个路由转发状态，引入了较大的状态维护开销。

3.4.2　共享树

与源树以发送端作为根不同，共享树使用一个共用的根，这个根位于网络中的某个点，这个共享的根称为**汇集点**（Rendezvous Point，RP）。当组播源发送信息时，它先将信息发送到这个共享根，然后由共享根将其发送到各个组成员。这个组播组的所有发送源都使用这棵组播树。为此每个路由器中保存的路由状态是（*，G），其中 * 是一个通配符，代表所有该组播组的所有组播源，G 代表该组播组。如图 3.8 所示，路由器 C 是这个共享树的根，即 RP，实线箭头表示共享树。当组播源要发送信息时，它们首先沿着到 RP 的最短路将信息发送至 RP，然后 RP 负责将组播信息沿共享树发送到各个接收端。图 3.8 中虚线表示各个组播源到共享根的最短路径。目前网络上广泛使用的 PIM-SM 协议[11]使用共享树方案。

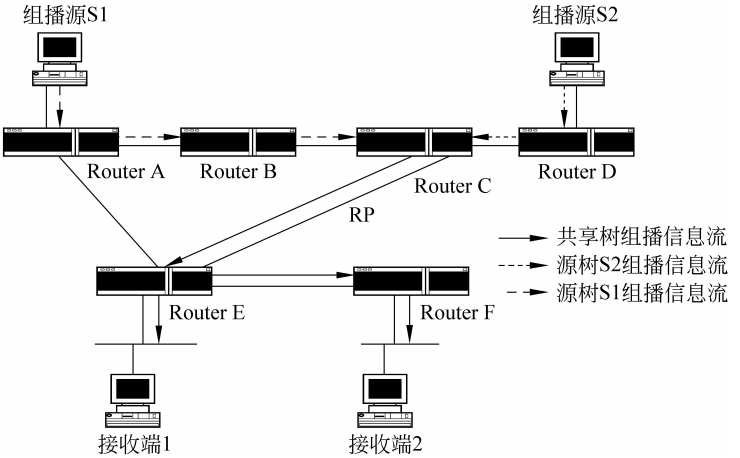

图 3.8　共享树结构