学位论文精选

IP 组播路由协议的研究与实现

李 炜
（信息产业部电信科学技术研究院，北京 100083）

[摘要] 概述了组播路由协议。分析了协议独立的组播路由—稀疏模式（PIM－SM）的缺陷，提出多个会聚点（RPs）的 PIM－SM 的改进机制。多个 RPs 机制有效地提高了原有单个 RP 的 PIM－SM 协议的健壮性，有利于实现负载均衡、分类业务及提高系统的容错性能，并讨论多个 RPs 机制的开销问题。分析了 PIM-SM 实现细节和组播技术的前景。

[关键词] 协议独立的组播路由—稀疏模式（PIM-SM）；多会聚点（RPs）；组播
[中图分类号] TN911 [文献标识码] A [文章编号] 1009－11742（2002）01－0082－07

1 引言

组播是将数据报送到一个组播组的所有成员的过程。节点可以动态加入和退出某个组播组。IP 组播介于 IP 单播和 IP 广播之间，能使主机将 IP 信息包发送到 IP 网络中的任何一组主机上。为此，把 IP 组播信息包中的目标地址安排成特殊形式的 IP 地址，称为 IP 组播地址。通过组播可以在一次传输中将数据报发送到多个接收者。子网（sub-network）中的非组播组成员可以在硬件上将组播数据报过滤掉，减少不必要的处理开销。组播提供了一种有效的通信、传输手段，并且充分利用网络资源。它可以使网络性能优化，使真正的分布式应用成为可能。在经济上，组播可节省网络和服务器的资源，使那些用单播或广播不可行的新型增值应用（如会议电视）成为可能。

组播路由协议用于生成组播源到组播组所有成员间的分布树。根据组播组成员在网络中的分布情况，IP 组播路由协议可分为两类：密集模式和疏松模式。密集模式组播路由协议有 Distance vector multicast routing protocol（DVMRP），Multicast open shortest path first（MOSPF），Protocol-independent multicast-dense mode（PIM－DM）。疏松模式组播路由协议有 Core-based trees（CBT）和 Protocol-independent multicast-sparse mode（PIM－SM）。由于 PIM－SM 使用显式的加入模型，组播信息被更好地约束在确实需要它的网络部分。因此，它不像 DVMRP 和 PIM－DM 那样低效，更适合在广域网末端有潜在成员的组播网络。PIM－SM 是组播网络域间组播路由的最好选择。

2 协议独立组播路由的稀疏模式（PIM－SM）

在 PIM－SM 中，组播组的成员要加入会聚点（RP），周围或下游所有组成员的路由器必须向 RP 发送加入信息，以加入到疏松分布树上。子网中 IP 地址最高的 PIM－SM 路由器被选为指定路由器（DR），负责向 RP 发送删除和加入信息。当一个接收者加入某个组播组，它通过因特网组管理协议（IGMP）通知指定路由器。指定路由器通过 Hash 函数计算这个组的 RP，并向它发送一个单播 PIM 加入报文，如果必要的话，还为该组播组创建一个转发表。

RP 的信息首先是通过收集自举信息得到的。一个路由器被选为它所在域的自举路由器，由它产生自举信息，并在必要时动态选择自举路由器，发

[收稿日期] 2001－10－11；修回日期 2001－11－09
[作者简介] 李 炜（1976－），男，湖北大悟县人，硕士，兴唐通信科技股份有限公司助理工程师

布状态稳定的 RP 的信息。当主机要加入某个组播组时，指定路由器用自举信息选择这个组的 RP。然后，它将组播数据报封装在 PIM 登记报文中，转发给该组的 RP。反过来，RP 送回 PIM 加入信息给源的指定路由器，允许后续的组播数据报直接发给它，而不用装在单播数据报中。

当组成员离开组时，指定路由器向 RP 发送删除消息，分布树相应的部分就被删掉。这棵树用于转发所有该组的业务，但这棵树不一定是最短路径。

共享树加入的第一步如图 1①所示。在这一步中，接收站点 1 通过 IGMP 成员关系报告加入到组播组 G。

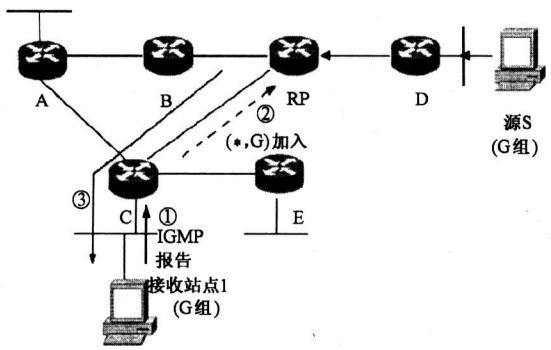

图 1　PIM－SM 共享树加入

Fig.1　The process of joining a PIM－SM shared tree

由于接收站点 1 是第一个加入组播组 G 的主机，因此，路由器 C 在它的组播路由表中为这个组播组建立一个（*，G）状态条目。然后，路由器 C 把到接收站点 1 的接口放在（*，G）状态条目的输出接口列表中。由于路由器建立了一个新的（*，G）状态条目，所以它必须向 RP（路由器 C 使用它的单播路由表确定通向 RP 的接口）发送一个 PIM（*，G）加入消息以便能够加入共享树（图 1 中②所示）。

RP 收到（*，G）加入消息，由于它原来没有组播组 G 的状态，所以在它的组播路由表中建立一个（*，G）状态条目，而且把到路由器 C 的链路添加到其输出接口列表。此时，组播组 G 的共享树已经在 RP 和路由器 C 及接收站点 1 之间建立起来，如图 1③实线所示。现在，从源 S 发送到 RP 的组播组 G 的任何信息都能沿着共享树下行发送到接收站点 1。

3　PIM－SM 的隐患及其解决

3.1　PIM－SM 的隐患

从上面可以看到，在 PIM－SM 协议中，RP 扮演着相当重要的角色。几乎所有的 PIM－SM 的操作都要通过 RP 来进行。RP 在 PIM－SM 中的地位至关重要。这样，由于 PIM－SM 对它的依赖太多，很容易使它成为整个系统的单个失效点。现有的 PIM－SM 机制每个组播组只用一个 RP 来建立共享树转发组播包。RP 失效会引起很严重的问题，而恢复过程很复杂，要用新的 RP 重新构造组播发送树。在新的组播发送树建立之前，肯定会有丢包的情况发生。同时，随着源和接收者的增多，RP 上会引起阻塞，通过这个 RP 的组播包会经历更多的延时。

因此，现有的 PIM－SM 机制不够健全。RP 的失效会引起严重的包丢失和性能异常，这对像流式媒体这样的应用来说是不可接受的。

3.2　多个 RP——隐患的解决

每个组使用多个 RP 机制来解决 PIM－SM 的上述隐患。在该机制中，每个源向多个 RP 注册并向它们发送组播包，但接收者只加入一个 RP。如果一个 RP 失效，加入该 RP 的接收者将不能到达这个 RP，然后，它就加入另一个替补的 RP。而源则无须采取特别的行动。为了支持多个 RP，源为每个 RP 设置出接口表项，而接收者则为每个 RP 设置入接口表项，建立了从源到多个 RP 及从多个 RP 到组播组的接收者的组播发送树。这些组播发送树可以预先建立以便在主 RP 失效时迅速恢复。

为支持多个 RP 机制，PIM－SM 需要做一些修改。

3.2.1　新的控制消息和定时器

1）*加入/删除消息*　为了支持多个 RP，在加入/删除消息时不仅是一个 RP 地址，而是对应组播组的多个 RP 地址及组播组地址。同时为每个 RP 设置一个优先级。这样，每个 RP 收到加入消息时，知道是哪个组播组的加入及自己在该组 RP 列表中的地位。

2）*存活消息*　组播组的各个 RP 定期向组内的其他 RP 发送存活消息，以宣告自己是活动的，这样，优先级比自己低的 RP 就不会转发组播包，如果自己收到优先级更高的存活消息，也不转发组播包。在存活消息中有组播组地址、RP 自己的 IP

地址及其优先级。

3）*定时器* 为了支持多个RP的机制，要增加两种新的定时器。**a**. 组播组的各个RP为优先级高于自己的RP设置一个定时器，以监测它们的活动情况。如果，所有这些定时器都到期，该RP就担负起转发组播组数据包的任务。**b**. 组播组的各个RP为自己设置一个发送存活消息的定时器。一旦该定时器到期，就发送自己的存活消息。如果在该定时器未到期之前RP收到优先级比自己高的RP的存活消息，就重置自己的存活消息定时器，这样就抑制了自己的存活消息，减少了网络上不必要的开销。

3.2.2 *RP的行为*

1）*多个RP的选择* 在PIM－SM协议中，组播组到RP的影射是通过Hash函数

$$\text{Value}(G,M,C(i)) = (1\,103\,515\,245\times((1\,103\,515\,245\times(G\&M)+12\,345)\ \text{XOR}\ C(i))+12\,345)\bmod 2^{31} \quad (1)$$

来实现，其中$C(i)$是侯选RP地址，G为组播组地址，M是自举消息中的Hash掩码。Hash掩码允许几个连续的组播组（如4个）影射到同一个RP。将结果排在前N位的侯选RP作为组播组G的RP，而RP的个数由网络管理者依具体情况确定。

2）*为RP设置优先级* 由式（1）中的算法得出的N个RP之间要有优先级的区别，以便确定谁负责转发组播组的数据报，谁只接收数据包而并不转发。RP的优先级由以下规则确定：

a. N个RP中那些路由更精确的RP优先级高；

b. 在规则a中比较结果相同的RP，按式（1）计算，结果大的RP优先级高；

c. 经过规则a和规则b的比较，有些RP的优先级仍然相同，则IP地址小的优先级高。

3）*RP决定是否转发组播包* 每个RP都接收来自源的组播包，但是RP要根据自己在当前活动的RP中的优先级地位来决定是否转发组播包，只有优先级最高的RP，才开始转发组播包。

每个RP收到来自源的PIM注册消息后，要向源发出（S，G）加入消息，在自己的组播转发表入接口列表上加上到源的接口。在收到来自其他路由器（如某个接收者的DR）的（S，G）加入消息后，也在自己的组播转发表出接口列表上加上相应的接口。

RP收到带有RP列表及RP优先级的加入消息后，知道自己在该组播组RP中的地位。每个RP定期向其他的RP发送存活消息。同时它也接收其他RP发来的存活消息。每个RP为优先级大于自己的RP设置一个存活定时器，每次收到来自优先级大于自己的RP的存活消息，相应的存活定时器复位。如果该RP所有的存活定时器都到期，则它就开始转发该组播组的数据包。如果在一段时间以后，它又收到优先级比自己高的存活消息，就立即停止转发组播包，由优先级高的RP转发。

4）*防止RP之间的频繁切换* 在实际的应用中，可能有这样的情况：优先级高的RP状态很不稳定，一会儿可以收到它的存活消息，一会儿又收不到，这样，优先级低的RP就在收不到存活消息时转发组播数据报，收到存活消息后又切换到高优先级的RP。在这种情况下，RP之间的切换过于频繁，也不利于组播接收者的接收。为了避免这种情况，设置优先级低的RP，只有它连续收到高优先级的RP的多个存活消息后，才认为该RP状态已经稳定，切换到该RP。

3.2.3 *DR及中间路由器的行为* DR收到组播组G的一个加入请求后，用式（1）算法找出N个RP。然后分别向这N个RP发送式（1）中的加入消息，建立到这N个RP的组播路径。当然，通往各RP的组播路径是不同的。这样中间路由器要为不同的RP建立（S，G，RPi）表项，以便相应的RP在负责转发时提供组播路径。

3.2.4 *一个例子* 下面用图2所示的网络来说明多个RP机制。

如图2所示，组播组G的源S向它的第一跳路由器D发送组播包。D根据式（1）找出2个RP——RP1和RP2，其中RP1优先级高于RP2。然后D分别向RP1和RP2发送注册消息，随后，RP1和RP2分别加入到以S为根的发送树。

接收站点在接收组播组G的数据包时向DR发送加入请求，DR根据式（1）找出2个RP：RP1和RP2，并分别建立到它们的路径。而中间路由器B建立（S，G，RP2）的转发表项，而C则建立（S，G，RP1）的转发表项。

开始时，由于RP1优先级高，由它负责转发组播包，RP2只是接收来自S的组播包，并不转发它。同时，RP1和RP2定期向对方发送存活消

息以示自己处于活跃状态。如果由于某种原因 RP1 失效，它就不会发出存活消息。经过一段时间后，RP2 为 RP1 设置的定时器到期，RP2 就认为 RP1 失效，开始转发来自 S 的组播包。

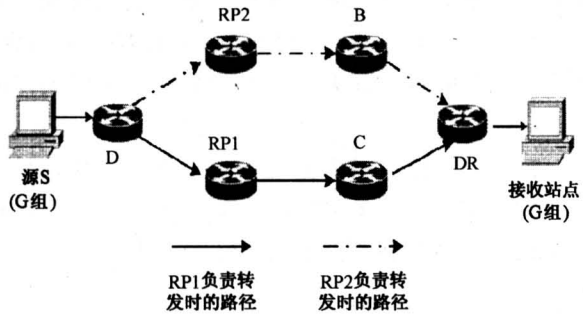

图 2　多 RP 的例子

Fig.2　An example of the multiple-RP scheme

为了防止从 RP1 向 RP2 切换过程中的数据丢失，在 RP2 将数据进行缓存。并用类似滑动窗口的机制来管理缓存。从源发来的数据被缓存在 RP2（大小为 N 个包）上的滑动窗口中。滑动窗口的机制：在窗口没有满时，所有进来的包都被存储；当窗口满时，如果再进来一个包，这时窗口前移。

检测到 RP1 失效后，RP2 取代 RP1 的位置。此时它还是要将进来的数据包缓存在滑动窗口中。为了避免缓存溢出，可以设置一个门限（比如窗口的一半），一旦超过此门限，RP2 向源发送消息，让源减小发送速率。状态稳定后，RP2 停止向源发送消息，正常运行。

3.3　多个 RP 的优点

多个 RP 机制可以带来如下的好处：负载均衡、可提供分类业务及容错性能。

在 PIM－SM 共享树机制中，RP 是源和接收者之间的中间点。随着接收者的增多，RP 的压力太大。如果经过 RP 的负载过重，那么延时和丢包就不可避免。在多个 RP 的机制下，每个 RP 负责转发一定数目的接收者的数据包，就可以在 RP 之间实现负载均衡。

用多个 RP 的机制可以提供较小程度的 QoS 支持。可以有选择的放置 RP，使每个 RP 负责不同类型的业务。接收者可以根据业务的要求选择不同的 RP。如图 3 所示，接收者对带宽和时延要求高时，可选择 RP1，而在接收者对带宽和时延要求不高时，可选择 RP2。当业务类型发生变化，导致业务要求变化时，可在 RP 之间切换。

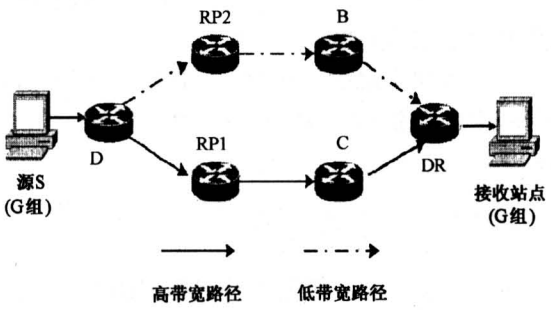

图 3　用多个 RP 实现分类业务

Fig.3　Realizing diff-serv with multiple-RP

在需要高度有保证的应用中，只有一个 RP 是不行的。这时多个 RP 机制的优点就显现出来了。一旦负责转发的 RP 失效，经过很短的一段时间后，其他的 RP 就会检测到这种情况，于是有新的 RP 担负起转发组播包的责任。这就为系统提供了很好的容错性能。

3.4　多个 RP 的折衷考虑和开销

3.4.1　*多个 RP 机制的折衷考虑*　为了提高 PIM－SM 机制的可靠性，正在转发的 RP 失效后，经过很短的时间，由另外 RP 担负起转发组播包的责任，在上述的 RP 机制中作了如下的折衷：

1）为了让替补的 RP 迅速的开始转发，源向所有的 RP 发送组播包，而所有的 RP 也接收组播包。并没有一个机制来通知源只向正在转发的 RP 发送组播包，这是为了减少 RP 切换的延时。

2）在中间路由器也为各个 RP 建立了（S，G，RPi）的转发表项，而不是只为正在转发的 RP 建立转发表项。这是为了在替补的 RP 开始转发组播包时，中间路由器无须临时再建立转发表项，减少组播包的发送延时。

3.4.2　*多 RP 机制的开销*　采用多个 RP 机制需要增加一些开销：

1）在加入/删除消息不仅有一个 RP 的信息，而还有组播组中其他的 RP 的信息，还要有它们的优先级。

2）在 RP 之间有定期的存活消息要传送和处理。每个 RP 都要接收组播包，要监测其他 RP 的活动情况，并作出自己是否要转发的决定。

3）源要向每个 RP 发送组播包。为了把组播包传给不同的 RP，在中间路由器上要建立相应的

转发表项。

4）为了防止数据的丢失，在不负责转发的RP上要对组播包进行缓存。

在实际的应用中，用两个RP已经足够，因此它所带来的开销也不大，是一 种有很好应用前景的机制。

4 PIM－SM的实现

为了在Linux环境下支持组播路由协议，系统内核和组播程序要协调工作，以完成处理和转发组播数据报的功能。

内核的基本功能就是根据组播的转发缓存转发组播数据包。而组播程序则负责处理从其他路由器和内核来的控制消息，并通过陷阱或系统调用来维护组播转发缓存。组播程序还根据PIM－SM协议的要求维护组播路由表中的定时器。

4.1 数据处理流程

当内核收到一个IP报，用ip-intr（）进程处理。进程ip-intr（）基于目的地址和报文的IP协议号将包发送到合适的处理状态机。在此，只关注以下几种情况：

1）如果报文是组播报，它先通过内核的组播转发状态机ip-mforward（），如果该报文的入接口与内核的转发表中的一项匹配，报文就沿着相应表项的出接口发送出去。否则，如果源和组不匹配，那么，缓存的报文丢失，产生一个内部错误的控制信息。

2）如果报文是PIM－SM报（如它的协议号为IPPROTO-PIM），它被传给内核的PIM状态机pim-input，然后按顺序由raw-input传给socket队列。

3）如果报文是IGMP报（如它的协议号为IP-PROTO-IGMP），它传给内核的PIM状态机igmp-input，然后按顺序由raw-input传给socket队列。

4.1.1 *组播数据报的处理流程* 设有一个来自源S的组播组G的数据报到达PIM组播路由器的X接口，PIM组播路由器上的PIM－SM程序对该数据报的处理流程如图4所示。

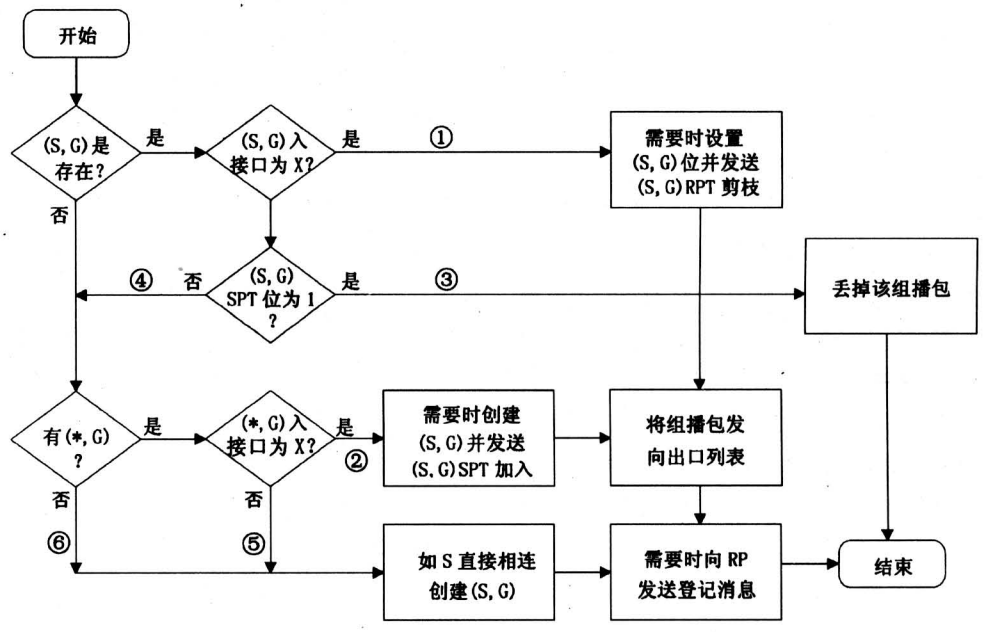

图4 组播数据报的处理流程

Fig.4 The processing of multicast datagram

1）流程到①时，组播路由器上有（S，G）表项，而且该表项的入口为接口X，说明组播包是由SPT树接收到，然后将组播包发往该表项的出口列表，需要时发送（S，G）RPT剪枝或向RP发送登记消息。

2）流程到②时，组播路由器上有（*，G）表项，而且该表项的入口为接口X，说明组播包是由RPT树接收到，然后将组播包发往该表项的出口列表，需要时向RP发送登记消息。

3）流程到③时，组播路由器上有（S，G）表

项，而且该表项的入口不是接口 X，而此时 SPT 树是活动的，应该由 SPT 树转发组播包，这时该包丢掉。

4）流程到④时，组播路由器上有（S，G）表项，而且该表项的入口不是接口 X，而此时 SPT 树不是活动的，这时要看 RPT 树是否活动。

5）流程到⑤时，组播路由器上有（*，G）表项，而且该表项的入口不是接口 X。这时，如果组播路由器和 S 直接相连，就在组播路由表中建立（S，G）表项并向 RP 发送登记消息，否则，由于 RPF 检查不通过，对该包不作任何处理。

6）流程到⑥时，组播路由器上没有组 G 的转发表项。这时，如果组播路由器和 S 直接相连，就在组播路由表中建立（S，G）表项并向 RP 发送登记消息，否则，对该包不作任何处理。

4.2.2　PIM－SM 加入消息的处理流程　设 PIM 组播路由器从它的 X 接口上接收到 PIM－SM 加入消息，该 PIM－SM 加入消息中的加入列表含有源 S。PIM 组播路由器上的 PIM－SM 程序对它的处理流程如图 5 所示。

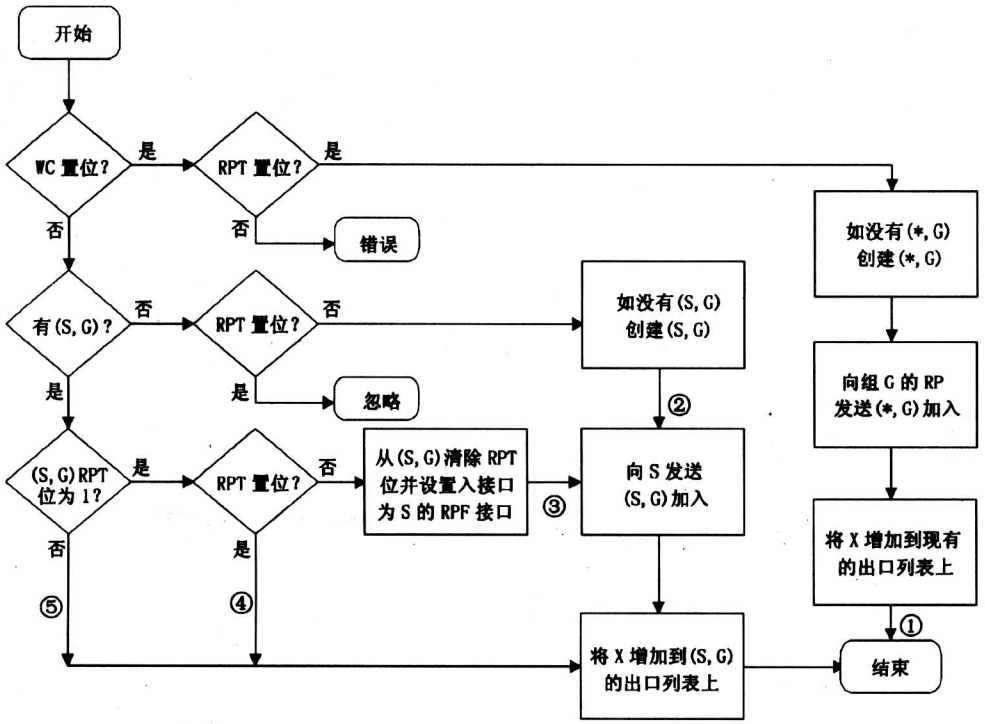

图 5　PIM－SM 加入消息的处理流程

Fig.5　The processing of PIM－SM Join message

1）流程到①时，WC（通配符）位和 RPT 位均为 1，因此这是一个加入（*，G）的加入消息。如果没有（*，G）表项就创建它，并向组 G 的 RP 发送（*，G）加入，将接口 X 添加到该表项的出口列表上。

2）流程到②时，WC（通配符）位和 RPT 位均未置位，且没有（S，G）转发表，因此这是一个新的源加入（S，G）。创建（S，G），并向 S 发送（S，G）加入消息，将接口 X 添加到该表项的出口列表上。

3）流程到③时，WC（通配符）位和 RPT 位均未置位，但是有（S，G）转发表，且（S，G）转发表的 RPT 位为 1，因此这是源 S 重新加入（S，G）的 RPT。向 S 发送（S，G）加入消息，将接口 X 添加到该表项的出口列表上。

4）流程到④时，WC（通配符）未置位，RPT 置位，但是有（S，G）转发表，且（S，G）转发表的 RPT 位为 1，这是（S，G）RPT 的加入更新，将接口 X 添加到该表项的出口列表上。

5）流程到⑤时，WC（通配符）未置位，但是有（S，G）转发表，且（S，G）转发表的 RPT 位为 0，这是（*，G）RPT 的加入更新，将接口 X

添加到该表项的出口列表上。

4.2 对多个 RP 的支持

为了支持多个 RP，在以上的 PIM－SM 实现的基础上还要做如下修改：

1）源的第一跳路由器和接收者的 DR 要建立到多个 RP 的路径，源为每个 RP 设置出接口表项，而接收者则为每个 RP 设置入接口表项。

2）每个 RP 上有进程处理收到的带有 RP 列表和优先级的加入消息，还要有进程设置定时器，定期发送存活消息，并根据其他 RP 的活动情况决定自己是否转发组播包。

3）中间路由器要有进程为不同的 RP 设置不同的转发表项。

4）在不负责转发的 RP 上要有进程对组播包进行缓存。

这些修改可支持多个 RP 机制，实现更好的可靠性。

5 结语

IP 组播将在 Internet 上扮演重要的角色，是未来 Internet 的必选项，而不是可选项。它允许应用开发者给网络增加很多功能而无须对网络做显著的改动。

开发 IP 组播应用并不难，为了发送组播数据报，可将数据发往该组播地址，如果要将它发往本地网以外，可将其 TTL 值增加。为了接收组播数据报，可加入相应的组播组。但是，让 IP 组播在组成复杂的网络上很好的运行并不简单，现在很多网络上采用了 MPLS 技术，在这样的网络上实现组播也是一个新的课题。提供可靠的、安全的数据传输并解决域间组播路由的问题，还有很多工作要做。

参考文献

[1] Estrin D, Farinacci D, Helmy A, et al. Protocol independent multicast-sparse mode (PIM－SM): protocol specification [S]. RFC 2362, June 1998.

[2] Deering S, Estrin D, Farinacci D, et al. The pim architecture for wide-area multicast routing [M]. ACM Transactions on Networks, April 1996

[3] Fenner W. Internet group management protocol, version 2 [S]. RFC 2236, November 1997

[4] Semeria C, Maufer T. Introduction to multicast routing [M], Internet Draft, Internet Engineering Task Force (IETF), March 1996.

[5] Handley M, Crowcroft J. Multicast today [J]. The Internet Protocol Journal, 1999, 2(4): 2～19

[6] Comer D E, Stevens D L. 用 TCP/IP 进行网际互连(第二卷):设计、实现和内部构成 [M]. 第二版. 张娟, 王 海. 北京:电子工业出版社 PRENTICE HALL 出版公司, 1998

[7] Parkhurst W R (CCIE). Cisco 组播路由与交换技术 [M]. 京京工作室. 北京:机械工业出版社 McGrraw-Hill 出版公司,1999

Research and Realization of IP Multicast Routing Protocol

Li Wei

(*China Academy of Telecommunication Technology, Ministry of Information Industry, Beijing* 100083, *China*)

[**Abstract**] This paper first explains why IP multicasting must be used. Then it summaries the fundamentals of IP multicasting and multicast routing protocols. After that it specially discusses protocol independent multicast-sparse mode (PIM-SM), points out its defect and proposes a scheme to solve it. In this scheme each source registers and sends data packets toward multiple rendezvous points (RPs) but receivers only join to a single RP. If one of the RPs fails, receivers will join to one of the alternative RPs quickly. On implementation, the paper centers on the details of implementing PIM-SM. Finally it discusses the problems and prospect of IP multicasting.

[**Key words**] protocol independent multicast-sparse mode (PIM-SM); multiple rendezvous point (RPs); multicast