全面认知 PIM – SM

Fully understand the PIM – SM

朱洪涛 （中南大学湘雅二医院 信息中心 湖南长沙 410011）

摘 要： PIM – SM（独立于协议的组播稀疏模式）具有自身的特性和优点被认为是大多数通用组播网络组播路由协议的最好选择。它更适合应用于广域网链接末端有潜在成员的组播网络中，在更多实际应用中它常常与 PIM – DM、IGMP 一起使用。

关键词： 汇合点 共享树 最短路径树 引导路由器

1 引言

随着宽带多媒体网络 的不断发展，各种宽带网络应用层出不穷。IP TV、视频会议、数据和资料分发、网络音频应用、网络视频应用、多媒体远程教育等宽带应用都对现有宽带多媒体网络的承载能力提出了挑战。采用单播技术构建 的传统网络已经无法满足新兴宽带网络应用在带宽和网络服务质量方面的要求，随之而来的是网络延时、数据丢失等等问题。此时通过引入 IP 组播技术，有助于解决以上问题。组播网络中，即使组播用户数量成倍增长，骨干网络中网络带宽也增加很少，从而最大限度的解决目前宽带应用对带宽和网络服务质量的要求。与 IP 单播类似，IP 组播路由协议建立了组播数据流在整个网络中的转发路径，但是与单播中数据流经过网络中单一路径从源主机到目的主机传递机制不同，IP 组播数据流的接收方通常是一组在网络中物理位置分散分布，但是以相同组播地址标识的主机组。由于从组播源到各组播目的的的转发路径拓扑均为树状，在组播中有一更形象的术语——组播分布树，我们用它来形容组播流在网络中经过的路径。组播路由协议又分为域内组播路由协议及域间组播路由协议两类，域内组播路由协议包括 PIM – SM、PIM – DM、DVMRP 等协议，域间组播路由协议包括 MBGP、MSDP 等协议。本文将重点介绍 PIM – SM。

2 PIM – SM 的特点和优势

PIM – SM 是一种稀疏模式的组播路由协议，比较适合应用于接收站点分布稀疏的网络（如广域网）。它通过设置汇合点路由器 RP 和引导路由器 BSR 来向所有 PIM – SM 路由器通告 RP – Set 信息、以及路由器的显式发送加入 – 剪枝（Join/Prune）信息建立起基于 RP 的共享树 RPT，组播数据沿着共享树流到加入到该组播组的网段。当数据流量达到一定程度，组播数据流可以切换到基于源的最短路径树 SPT，以减少网络延迟及负担。它都通过周期性地发送 Hello 报文与相邻的其他运行 PIM – SM 的路由器建立并且保持相互邻接关系。保持对邻接 PIM – SM 路由器的追踪对于建造和维护 PIM – SM 共享树是十分重要的。

在稀疏模式中，组播数据流只被发送到需要它的网络接收节点上。在 PIM – SM 中，这种发送是通过一跳一跳主动地向组播分布树的汇合点（RP）发送加入（Join）报文来完成的。当加入报文沿着树上行发送时，沿途的路由器建立组播转发状态以便需要的组播信息可以沿着树下行被传送回去。同样，当不再需要组播数据流时，路由器为了剪枝不需要的分支，沿着树上行朝汇合点发送剪枝（Prune）报文。当此剪枝报文沿着树向上一跳一跳地传送时，沿途每个路由器适当地更新它的转发状态。其更新经常导致与组播组或源有关的转发状态被删除。在这里，PIM – SM 与 PIM – DM 的最本质的区别在于：在 PIM – DM 下，路由器转发状态是通过组播信息的到达而建立。而在 PIM – SM 下，路由器转发状态是通过显式发送加入报文而建立的。

PIM – SM 通过建立共享树来实现对组播数据流的转发。共享树的建立是由需要收到来自指定组播组数据的最后一跳路由器（至少有一个直接相连的组播组

接收站点的路由器)来启动。当最后一跳路由器不再需要指定组播组数据时(即不再有任何组播组的接收站点时)路由器把自己从共享树上剪枝掉。

与其他的稀疏模式协议(如 CBT 即有核树)不同,PIM－SM 的主要优势之一就是它不限制只能通过共享树接收组播数据。具体来说就是:显式连接机制能被用来连接到根是某个特定源的最短路径树(SPT)。这是一个明显的优势——通过加入 SPT,组播数据不必通过 RP 即可直接路由到接收站点,因此减少了网络延迟以及在 RP 上可能出现的阻塞。通过加入 PIM－SM 中的 SPT,获得了最优分布树的益处,并且不会有和其他密集模式协议相关的开销和低效率。

PIM－SM 使用一个单向共享树,组播数据流从共享树的汇合点(RP)沿着树向下游流动。因此,组播数据源首先的任务就是设法使它的数据流到达 RP 以便数据流能够从共享树流出,PIM－SM 通过让 RP 加入 SPT 再返回源来完成这一任务,以便它能收到这个源的数据流。但是,首先 RP 必须设法知道源组是存在的,为了完成这个任务,PIM－SM 利用 PIM 注册和注册抑制报文来完成源注册过程。PIM 注册报文由第一跳路由器(DR:直接与组播源相连的路由器)发送到 RP。PIM 注册报文有两个用处:第一是通知 RP 源 S1 正在有效地向组 G 发送数据;第二个是为了沿着共享树向下发送信息,向 RP 转发源 S1(每一个都封装在一个单独的 PIM 注册报文中)发送最初的组播数据包。当 RP 接收到 PIM 注册报文时,它首先解封此报文以便它能全面检查组播数据包。如果数据包是来自一个存在的组播组(也就是组的共享树加入已经收到)则 RP 沿着共享树向下转发数据包,然后 RP 加入基于源的 SPT 以便它能够收到(S,G)原始的数据包而不是接收封装在 PIM 注册报文中的组播数据包。另一方面,如果没有对应组的共享树,RP 只是简单地丢弃组播数据包,并不向源发送 SPT 加入消息。当源发送的组播数据包经过 SPT 流向 RP 或者 RP 并不需要组播数据包,RP 就不再需要继续收到封装在注册报文的(S,G)数据包。于是 RP 向 DR 单播注册抑制报文,指示 DR 停止向 RP 发送注册报文。

对于一个特定的源,PIM－SM 能够把最后一跳 DR(也就是直接与已加入一个组播组的主机相连的 DR)从共享树切换到 SPT。这一步通常是通过指定一个利用带宽的 SPT 阈值来实现。如果超出了该阈值,最后一跳 DR 加入 SPT。请注意,在这里是最后一跳路由器而不是接收站点启动 SPT 切换。这种情况下,可能存在两种路径使得组播数据流流到接收站点:共享树和 SPT 树,这将导致数据包的副本被发送到接收站点,造成一种网络带宽的浪费。PIM－SM 使用一种特定的剪枝报文,将该报文沿着共享树发送给 RP,从而剪枝来自共享树的源组播数据。

为了使 PIM－SM 正常的工作,在 PIM－SM 域内的所有路由器必须知道 RP 地址。在所有组播组使用一个单独的 RP 的小网络中,在每个路由器的配置中人工指定 RP 的 IP 地址是可能的。但是如果网络规模比较庞大,或者如果 RP 经常变化,每个路由器上的配置就造成非常繁重的工作量。PIM－SM 使用自举(Bootstrap)机制允许在同一个域内的所有 PIM－SM 路由器动态地了解所有组到 RP 的映射,从而避免了人工配置 RP 的问题。

由于以上 PIM－SM 的诸多优点使得 PIM－SM 被认为是大多数通用组播网络域间组播路由协议的最好选择。它更适合应用于广域网链接末端有潜在成员的组播网络中。

3　PIM－SM 的工作过程(共享树的加入、剪枝)

PIM－SM 的操作是围绕着一个单向的共享树来展开的,这里的单向是指:从源到接收者方向。在共享树上,有一个根节点——RP,共享树上的组播数据流要依赖于 RP 来向下转发,因此共享树也叫 RP 树,通常称作 RPT。

那么源的数据流是如何到达接收者的呢?见下图:

ReceiverB 是个接收者,想接收 HostA 的数据流,则它向路由器 C 发送一个 IGMP 加入报文(该报文中包含一个组播组,即 B 想接收的那个组播流的多播地址),RouterC 收到这个加入报文后,它要检查看是否存在有关于该多播地址的路由条目,如没有则创建一个(*,G)路由条目(这里的 G 就是目标多播组的组地址),并将收到加入报文的接口添加在这个路由条目的出接口中。

同时这也引发了 RouterC 向 RP(图中的 RouterD)

发送一个 PIM(*,G)加入消息,以便能够加入共享树。至于 routerC 是如何知道 RP 的,我们将在以后讨论。

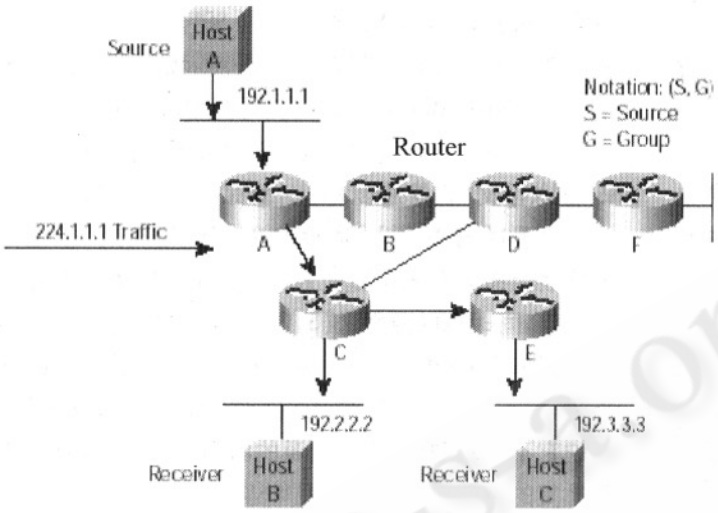

图 1

RP 收到这个(*,G)加入消息,也检查看是否存在有关于该多播地址的路由条目,有则将收到消息的接口加入到相应条目的出接口表中(即自己与 RouterC 相连的接口),如果没有相应的路由条目则创建,并也在其出接口表中添加收到消息的接口。

其实组播路由器在转发组播数据流的时候,并不关心其下面有多少个接收者,他们分别位于何处,它只关心组播数据流是否有相应的出接口,有就将它们从出接口转发出去,没有就丢掉。

假设,此时另外的一个接收者 ReceiverC 也想加入组播组 G 的共享树,则它也向它的直连 routerE 发送一个 IGMP 加入报文,E 一看没有有关 G 组的多播路由条目,它也建立一个(*,G)路由,并将相应的接口添加到该路由的出接口表中。由此又引发了 E 向 RP 发 PIM(*,G)加入消息。

当这个加入消息到达 RouterC 后,C 检查看到已经有了有关组播组 G 的多播路由条目(*,G)(也就是说它已经在该组的共享树上了),则 C 简单的将收到该加入消息的接口添加到这个(*,G)的出接口表中。并不再向 RP 发加入消息了。

至此,这个单向的共享树的下半段(即 RP 到接收者)路径就建立好了。

那么,组播源的信息是如何到达 RP,并向下流到接收者的呢? PIM - SM 是通过源注册来完成这一步的。

PIM 注册消息由第一跳路由器 DR(也就是直接与一个组播源相连的路由器)发送到 RP,注册消息的目的是:

(1) 通知 RP 源(图中的 HostA)正在有效地向组 G 发送信息。

(2) 为了沿着共享树向下发送信息,向 RP 转发源最初的组播信息包。

当组播源开始传输组播数据时,DR 收到组播数据包,它查看数据是从其直连的网络中收到的,则 routerA 知道自己是第一跳路由器——DR,并在它的组播路由表中建立一个(S,G)状态条目。由于是 DR,所以 routerA 将组播信息包封装在一个独立的 PIM 注册消息中,并把它单播给 RP。

当 RP 接收到 PIM 注册消息时,它先解封装此消息,检查组播包,看是否存在相应的组播组。如果不存在就简单的丢弃,并不向源发送加入消息;如果存在相应的组播组,RP 就沿着共享树向下转发信息,然后把 SPT 加入到源,以便它能接收到源的原始数据包,而不是封装在 PIM 注册消息中的。以图一为例说明这一过程:

组播数据源 HostA,开始转发数据,数据先是到达 RouterA,RouterA 发现数据来自与自己直连的网络,则它知道自己是 DR。然后,它创建(S,G)路由,并将组播数据包封装在 PIM 注册消息中并将它们单播到 RP(RouterD)。

当 RP,即 RouterD 收到 PIM 注册消息后,它解封装,查看数据包是否来在一个存在的组播组。因为在此之前 RP 建立了(*,G)路由项,所以 RP,将解封装了的数据从(*,G)路由的出接口转发出去。RP 还向源的方向发送一条(S,G)加入消息以把源加入到 SPT,即把(S,G)消息拉到 RP。

(S,G)加入信息一跳一跳的被传送到 DR,当 DR 收到加入消息时,从路由器 A 到 RP 的(S,G)SPT 就建立成了。为了终止 DR 的注册以及接收不封装的数据流,RP 还要向 DR 单播一个 PIM 保留消息。收到 PIM 保留消息后 DR 就不再发注册消息,并且不封装组播数据而是直接将其发送到 RP。

RP 接收到不封装的数据流就从相应的出接口转

发出去,到了 RouterC,RouterC 收到后检查出接口,一个是直接转发给了 HostB,一个是转发给了 RouterE。RouterE 收到后再从其多播路由的相应出接口转发出去,转发给了 HostC。

我们来做个总结,看看 PIM – SM 的工作过程:整个过程分为两部分,一部分是接收者到 RP,一部分是源到 RP。

接收者到 RP 的共享树建立是通过接收者向第一跳路由器发加入报文,再由第一跳路由器朝着 RP 方向发加入报文并在途中的所有路由器上建立相应的(*,G)路由条目,添加相应的出接口。

源到 RP 的共享树的建立,是通过 DR 向 RP 单播 PIM 注册消息,RP 解封装消息,查看相应的路由,转发数据,并向 DR 发送 PIM 保留消息,以终止 DR 的注册和接收不封装的数据流。

到此,一个组播数据流就从源通过 RP 流到了接收者 HostB 和 HostC。

以上是共享树的加入过程。那么,如果接收者不想再接收数据了么办呢,这就是我们要讲到的共享树的剪枝。

假设 HostC 不想再接收组播数据了,则它发送一个 IGMP 剪枝报文给路由器 E,E 收到剪枝后查看多播路由条目,然后这个接口从(*,G)的出接口列表中被删除。当接口被删除后,作为(*,G)条目的输出接口列表为空了,这表示 RouterE 不再需要这个组的信息了,则 RouterE 通过向 RP 发送(*,G)剪枝来把自己从共享树上剪枝掉。

RouterC 收到 E 的剪枝后,从它的(*,G)条目的出接口列表中将相应的接口删掉,因为 RouterC 的(*,G)的出接口列表中还有一个出口(到达 HostB 的)存在,所以它并不需要向 RP 发什么剪枝报文,它只是简单的删除掉到 E 的接口。

4 PIM – SM 的最短路径树

4.1 加入 SPT

PIM – SM 的一个主要好处就是它还可以使用 SPT 来接收组播信息。通过加入 SPT,组播信息不必通过 RP 即可直接路由到接收站点,因此减少了网络延迟以及 RP 上可能出现的堵塞。

在启了 PIM – SM 的 router 上,有个阀值叫 SPT – Switchover,当组播的数据流量大于设定的 SPT – Threshold 值时,router 就会向源发送一个(S,G)加入消息(它是通过 RPF 计算来确定把此报文从那个接口发送出去),以便加入到这个源的 SPT。消息被一跳一跳的流向源,并建立 SPT。说明一点:我们的设备缺省的 SPT – Threshold 值为 0,即实时要加入源的 SPT。

以图一为例讲解这个过程:routerE 向源发送(S,G)加入报文(它是通过 RPF 计算来确定把此报文从那个接口发送出去),当 RouterC 收到该报文后,就在它的组播组播转发表中建立(S,G)条目,并添加相应的出接口,同时也向源发送(S,G)加入报文。

因为 RouterC 检测到,共享树和 SPT 的路径在此分离,所以它沿着共享树向 RP 发一个 RP 位剪枝消息,就是说它不想从 RP 那里得到源的组播数据了,以避免收到重复的信息。

最后,当 RouterA 收到加入报文时,它向现有的(S,G)条目的输出接口表中添加到 RouterC 的接口,这样从源来的数据流就可以通过 A 直接到达 C。

4.2 剪枝 SPT(与剪枝共享树基本相同)

HostC 发送 IGMP 离开报文,routerE 收到后剪枝相应(*,G)、(S,G)的出口列表,因为只有 HostC 一个直连的成员,所以出口列表为空了,routerE 要向 RP 发送(*,G)剪枝,并停止发送定期的(S,G)加入消息。RouterC 收到剪枝消息也从(*,G)出口表中剪枝相应的出口,如果出口表为空则继续向 RP 发(*,G)剪枝,而 routerE 和 routerC 的(S,G)路由项就等着超时而被删掉。

5 RP

以上的讲解中,我们遗留了一个问题,就是 RP。RP:Rendezvous Point,集合点、汇合点的意思,它是共享树的根。

对于每一个多播组来说,都必须有且只有一个 RP,组播数据流要想下流到接收者就必须经过 RP,或至少最初要经过 RP(对于源的 SPT 来说),因为源不知道接收者在那里,它只能把数据交给 RP,RP 知道接收者与其相连的接口,并将数据从这些接口转发出去。与接收者相连的第一跳路由器最初甚至有的最终也不知道源在那里,它们只是将加入组的信息向 RP 发送,并在沿途建立起相应的(*,G)路由,最终到达 RP。

那么，源和接收者是如何知道 RP 在何处的呢？

一种方式，administrator 可以手工在网络上的每台 pim 路由器上指定静态的 RP，用命令：ip pim rp – address X. X. X. X。这样所有的 pim 路由器就都知道 RP 的位置了，但很明显这样有个弊端：它不适合用在大型以及经常变化的网络上。

另一种方式，动态的 RP，也就是 PIMv2 中定义的“将组到 RP 的映射信息发布到网络的所有 PIM 路由器”，这种方法常被简单的叫做“自举路由器机制”。

自举路由器机制，允许网络中有多个候选的 RP，这样可以容错，避免单一的静态 RP 失败而造成组播网络的瘫痪。那么多个候选的 RP，究竟哪个会被如何选为网络中的真正有效的 RP 呢？这就依靠 BSR（bootstrap router）。

BSR（bootstrap router）称作自举路由器。

网络中的某台路由器被配置成 BSR，则它要先向所有的 PIM 路由器发送 bootstrap 报文，通知大家“我是 BSR”。当候选的 RP 得知这个消息后，它们就把自己的信息单播给 BSR，这样 BSR 会收到网络中所有候选 RP 的信息，然后 BSR 就定期发送带有所有候选 RP 信息的 bootstrap 报文，给所有的 PIM 路由器，之后所有的 PIM 路由器按照相同的 hash 算法在本地算出一个 RP。因为大家收到的候选 RP 信息都相同，算法也相同，所以算出的 RP 也一定相同，这样就保证了网络同一个组播组可以映射到同一个 RP。（BSR 也可以配置多个，以提供容余，它们之间的竞争还是通过 bootstrap 报文。）

6 总结

结合以上例子，我们可以清楚认识到 PIM – SM 独立于协议，主要是指 PIM 不依赖于某种特定的单播路由协议，它只是利用单播路由协议建立起来的单播路由表来完成 RPF 校验，而非维护一个组播路由表来实现组播的转发。因为 PIM 不需要保持自己的路由表，所以它不需要象其它协议那样发送或接收组播路由更新，这样 PIM 的开销也就低了许多，可适用于很多应用。

参考文献

1 孙宝林、李腊元，可靠多播协议的分类研究，《武汉理工大学学报（交通科学与工程版）》，2005 年 03 期.

2 张士文，组播技术和网络技术，《计算机工程与设计》，2003 年第 1 期.

3 Winmag 社区，《IP 组播技术简介》2004 – 2 – 9.

4 思科全方位服务分支机构解决方案，《IP 多播技术及其应用》.

5 蒋华平、鲁东明，可靠多播技术研究，《计算机工程与应用》，2004 年 30 期.