基于多点隧道的组播 VPN 研究①

Study of Multicast VPN Based on Multipoint Tunnel

杨 波　廖建新　武家春　(北京邮电大学网络与交换技术国家重点实验室 100876)

摘要:介绍了组播 VPN (Virtual Private Networks)的三种候选方案,在比较分析各自优缺点的基础上,本文重点研究了基于多点组播隧道的组播 VPN 方案。该方案利用公网上的 PIM – SM(Protocol Independent Multicast – Sparse Mode)协议建立组播转发树作为多点组播隧道,在公网上透明传输私网的组播报文。

关键词:第三层 VPN　PIM – SM　多点隧道　组播 VPN

1　概述

目前广泛使用的各种 VPN 技术,在支持地点分散的 VPN IP 组播业务方面存在扩展性和可管理性方面的问题,不适用于电信级应用。本文深入研究了基于多点组播隧道的组播 VPN (Multicast VPN,MVPN)的解决方案,详细分析了该方案的关键技术细节,包括:多点隧道的建立过程,反向路径转发机制的扩充,私网数据在公网上的转发流程,组播业务流量优化等。上述问题的明确解决,有助于组播 VPN 技术的深入研究和产品化。

2　组播 VPN 的方案选择

2.1　BGP/MPLS VPN 体系结构

BGP/MPLS(Border Gateway Protocol/Multi – protocol Label Switching)技术为运营商提供了一种新的 IP VPN 解决方案[2]。在 BGP/MPLS VPN 网络结构中,每个用户接入点称作一个 Site;VPN 网络中包括如下实体:骨干网边缘路由器(Provider Edge Router,PE),用于存储虚拟路由转发实例(Virtual Routing Forwarding Instance, VRF),处理 VPN – IPv4 路由,接入位于各个 Site 的 VPN 子网,是 VPN 的主要实现者;用户网边缘路由器(Customer Edge Router,CE),汇聚本 Site 的私网路由,发布和接收用户网络路由;骨干网核心路由器(Provider Router,P),基于 MPLS 标签转发私网数据报文。

每个 Site 的 VPN 用户经 CE 连接本地 PE,并在 PE 上对应特定的 VRF。VRF 上配置一些策略,规定该 Site 的路由器可以接受哪些 Site 的路由信息,可以向外发布哪些 Site 的路由信息。每个 PE 根据 BGP 的扩展信息进行路由计算,生成每个相关 VPN 的路由表。PE 设备维护多个路由表,支持动态路由协议的多实例。

2.2　组播 VPN 方案及其比较

在基于 BGP/MPLS 的 IP VPN 上提供组播 VPN 解决方案,推动了 VPN 用户多媒体业务的开展,为网络提供商创造新的业务增长点。

针对 IP 组播的特点,目前有三种组播 VPN 解决方案[3]:

(1) 多点组播隧道方案。在公网上利用组播转发树的思路建立的连接该 VPN 所有 Site 的通道,称为组播隧道(Multicast Tunnel,MT),通道的端点位于 PE 路由器上,称作组播隧道接口(Multicast Tunnel Interface,MTI)。为了在公网上建立 MT,要在提供商的网络上启动一个公网组播路由协议实例,考虑到协议的成熟性与公网的伸缩性及带宽的优化,通常选用 PIM – SM (Protocol Independent Multicast – Sparse Mode)。PE 路由器要接入多个 VPN 的 CE,对应每个 CE 启动一个私网的组播路由协议实例。如图 1 所示,一个 VPN 在公网上对应一个 MT,MTI 负责将私网的组播协议和数据封装在公网的组播数据报文中,发送给远端的 PE;

① 基金项目:高等学校博士学科点专项科研基金资助课题(No. 20030013006);电子信息产业发展基金重点项目(下一代网络核心业务平台)。

远端的 PE 对报文解封装还原出私网数据，根据 VPN 标识转发给相连的 CE；这样私网的数据和协议报文就透明的穿越了公网，顺利到达了对端 VPN Site。

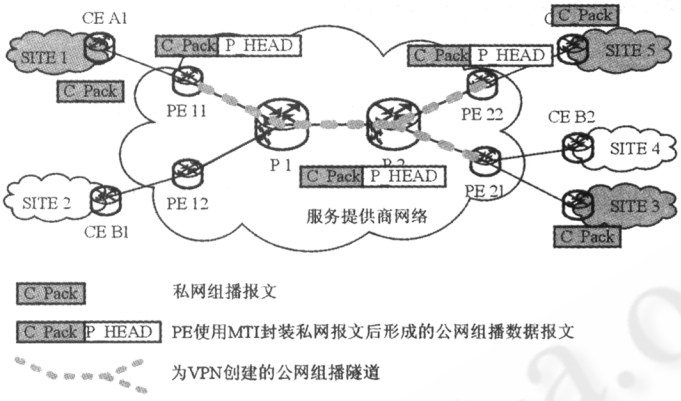

图 1　公网组播隧道方案原理

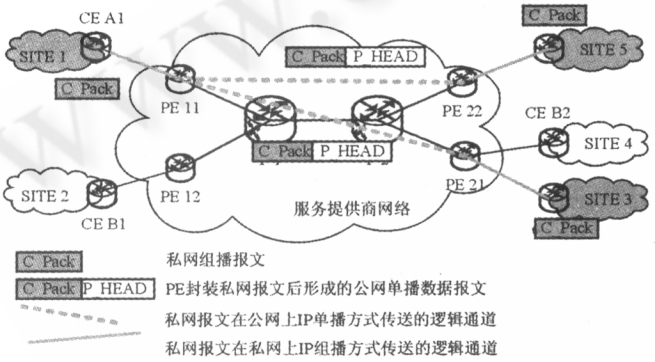

图 2　公网上的单播复制方案原理

（2）单播复制方案。本地 PE 对来自 CE 的组播报文，为 VPN 中每个需要该流量的对端 PE 复制一份，以对端 PE 的 IP 地址作为报文的目的地址，用 IPinIP[2] 方式封装成公网单播报文，使用公网单播方式发送给对端 PE。对端 PE 上拆封装以私网组播方式转发给相应的 CE，如图 2 所示。

（3）路由区分符方案。该方案借鉴 BGP/MPLS 单播 VPN 的思想，在公网上使用唯一的路由区分符 RD（Route Distinguisher）来标识特定的 VPN，为各私网的组播路由和 PIM 协议报文增加不同的 RD 作为区分，结合组播树的源和 MVRF（Multicast VPN Routing and Forwarding Instance）信息，构成（RD：S，G）和（RD：＊，G）形式的路由项，使得私网组播路由项全局唯一；对私网的数据报文，匹配对应 RD 的私网路由项在公网上转发。

表 1 对这三种方案的优缺点进行了对比，依据三种方案各有的特点，考虑到骨干网升级的难度和代价、网络可运营、可管理性问题，我们首先排除方案 3；方案 2 对于 PE 路由器的要求高，对网络带宽消耗大；本文选择多点组播隧道方案，下一节将对该方案中的关键技术细节进行深入研究。

3　基于多点组播隧道的组播 VPN 关键技术

在方案 1 中，PE 路由器要支持组播协议的多个实例：在公网上启动公网 PIM－SM 协议，建立多点组播隧道，封装和转发私网 VPN 组播数据，以下称为公网实例；还需要与各个 CE 交互，因此还需要为每个 CE 启动一个对应的私网组播路由实例，以下称为私网实例。但并不要求公网和私网都必须运行 PIM－SM 协议，VPN 用户可以根据以往网络部署或者应用需要选择适合的组播路由协议，只要 PE 支持该协议。考虑到 PIM－SM 目前是广泛使用、扩展性和性能较好的组播路由协议，本节以公网/私网均运行 PIM－SM 协议为例说明。

3.1　MVPN 接口和多点组播隧道

根据位置和功能将 PE 上的接口划分为 4 类：公网接口（Provider Network Interface，PNI）：用于连接 P 设备，来自该接口的数据依据公网 MVRF 路由；私网接口（Customer Network Interface，CNI）：用于连接 CE 设备，来自该接口的数据报文使用对应的私网 MVRF 路由；组播隧道接口（Multicast Tunnel Interface，MTI）：是 MVRF 为私网组播动态创建的虚接口，是多点组播隧道的终点，位于同一个 VPN 中的 PE 路由器间通过 MTI 交换私网的 PIM 协议控制报文并建立邻居关系；组播 VPN 标识接口（简称 MVI）：是一类逻辑接口，仅出现在 PE 公网组播路由表的接口列表中，用于标识该报文来自或者将转交对应组播 VPN 的 MVRF 处理。

多点组播隧道 MT 建立在同一个 VPN 的 PE 之间，用户的协议控制报文和业务数据报文在公网上都是通过该隧道透明传送的。MT 实际上是通过公网 PIM－SM 实例创建的组播转发树，与普通的 PIM 组播转发树不同的是：多点组播隧道的终点都是 MTI 接口，MT 用于创建组播树的组地址是在 PE 上通过配置命令指定

的 D 类组播地址，而不是从组播数据报文中获得的；作为特殊的组播转发树，MT 与常规的 PIM 转发树的创建和维护机制不同，必须对 PE 路由器上的 PIM 公网路由协议进行扩充。

表 1　三种 MVPN 方案的比较

方案名称	优点	缺点
方案 1：多点组播隧道	• 组播 VPN 对骨干网路由器透明，无需对 P 路由器现有的 BGP/MPLS 路由协议升级 • 只用为每个 VPN 在公网骨干路由设备上建立一组组播路由项以维护 MT，而不用记录私网组播业务的路由项，公网 P 路由器没有的额外协议处理开销 • 私网可以运行任何 IP 组播协议，私网协议透明于公网 • 无需显式配置隧道和隧道接口，管理维护费用小	• 由于私网报文在多点组播隧道中全转发，每个 MTI 都要接收该 VPN 的组播流量，对特定组播组不感兴趣的 PE 也会收到该组播流，公网上的网络业务流量没有最优化 • 公网路由器上必须启动组播路由协议 PIM－SM • PE 要支持组播路由协议的多实例
方案 2：单播复制	• 公网无需运行组播路由协议，也不用维护私网的组播路由表	• PE 设备上的数据复制和封装量大，对 PE 的性能提出了严峻的考验 • 公网上采用单播方式，增加了公网的数据流量负担，扩展性差
方案 3：路由区分符	• 充分利用了组播的优势，在公网上建立了依据 RD＋路由项形成的优化组播通道，按需转发私网报文，优化了公网通信流量 • 只有对 VPN 的特定组播组感兴趣的 PE 才会接收私网组播流，降低了 VPN 业务对 PE 的负担	• 骨干网设备上的组播路由项数量由各私网组播业务数量决定，网络运营者对此不可控 • 需要对公网路由器进行软硬件升级，以支持私网数据基于 RD 的路由转发 • 公网和私网必须运行相同的协议：PIM－SM

对 PE 上运行的 PIM 公网实例的扩充体现在如下方面：由于在组播路由表中引入了新接口类型，组播转发流程发生了变化；组播域地址是通过配置命令指定的，并不会有对应的 IGMP 组地址报告，因此必须在配置完成后在公网 MVRF 上生成静态 IGMP 成员，并显式向公网汇聚点 RP（Rendezvous Point）发送 PIM 加入报文；MTI 上并不会象常规的 PIM－SM 协议那样收到该组的组播数据报文，无法触发向公网 RP 发送 PIM 注册报文和创建（S，G）路由，因此 PE 使用 BGP 的“Update－Source”接口地址来创建（S，G）路由。“Update－Source”接口一般选择本路由器的 LoopBack 接口，以避免因物理端口状态改变引起的路由振荡。

3.2　PIM 邻居和 RPF 机制

由于多实例和隧道接口的引入，PIM 邻居的概念大大扩充，通常有三种邻居关系：CE 路由器与对应于该 Site 的 PE MVRF 建立私网的 PIM 邻居关系；PE 路由器之间通过 MT 建立 PE 私网 MVRF 邻居关系；另外 PE 路由器还与 P 路由器之间建立公网邻居关系。

反向路径转发（Reverse Path Forwarding，RPF）是 IP 组播中避免数据环路和数据重复而采取的策略，通过 RPF 机制决定上游邻居和相应的 RPF 接口。RPF 接口的作用是检查数据报文是否从正确的接口进入路由器；RPF 邻居用于检查 PIM 的加入/剪枝报文是否来自正确的上游邻居。RPF 的基本思想是：依据单播路由表，判断组播数据进入的接口是否为到达该组播源路由代价最小的接口和邻居。

引入了组播 VPN 后，CE 和 P 设备上仍然采用常规的 RPF，但在 PE 上由于存在多实例，其 RPF 机制要进行调整。PE 的公网 MVRF RPF 检查使用常规方式；对私网 MVRF，若 RPF 查找到的接口位于同一 MVRF 中，则 RPF 接口和上游邻居使用普通组播 RPF 规则获得；若 RPF 发现到组播源的路由是经 BGP 协议从远端 PE 获得的，则以该 MVRF 的 MTI 作为 RPF 接口。MTI 接口是 MVRF 的 PIM 虚接口，但其上并没有运行单播路由，因此也没有单播的路由邻接关系，RPF 邻居的判断则必须依赖 BGP。若到组播源的 BGP 路由“Next－Hop”属性是该 PE 经 MT 建立的 MVRF 邻居，则认为该邻居就是合法的 RPF 邻居。

3.3　MVPN 的数据转发流程

为了方便说明，对图 2 进行了简化，用图 3 描述了在路由稳定的情况下各骨干网设备的组播路由项，以下说明 MVPN 如何将位于 Site 1 的 VPN 组播报文传送到位于 Site 3 的接收者的过程。假设私网的组播源地址为 Sc，私网组播组地址为 Gc；VPN_A 在公网上的组

播隧道组地址为 Gp，PE A1 的 BGP Update – Source 接口地址为 Sp。

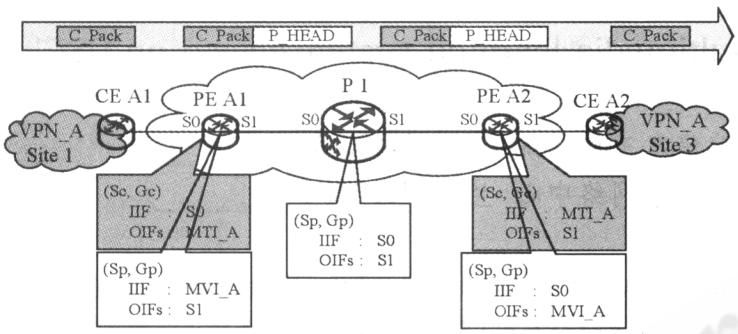

图 3　MVPN 数据转发流程

在入端边界路由器 PE A1 的 S0 接口上收到来自 CE A1 发送的组播数据报文[C_Pack]，S0 接口是 VPN_A 的 MVRF 接口，RPF 检查正确后，第一次查找 VPN_A 的 MVRF 组播路由表，以获得出接口列表；判断出接口的类型，若出接口为 MTI 类型，则根据该 MTI 的源地址 Sp 和目的地址 Gp 将私网封装成{[C_Pack]P_HEAD}形式的公网报文，再查找公网的 MVRF 组播路由表，获得出接口 S1，将该报文向 P1 转发出去。

出端边界路由器 PE A2 的 S0 接口上收到封装后的私网报文{[C_Pack]P_HEAD}后，由于 S0 绑定到了公网 MVRF 上，RPF 检查成功后，查找公网路由表获得了类型接口 MVI_A，这表明{[C_Pack]P_HEAD}报文到达了对端终点，此时需要拆封报文头，对内层的数据报文[C_Pack]再次查询 VPN_A 的 MVRF，从出接口 S1 将[C_Pack]转发到 CE2。

CE A1、CE A2 和 P1 上的转发行为，等同于常规的组播转发，此处不再赘述。

3.4　多点数据隧道的对 MVPN 流量的优化

第 2.2 节提到，方案 1 基于公网的组播转发树传送私网组播报文，由于 MT 是连接所有属于同一个 VPN 所有 PE 的多点组播隧道，一个 MTI 发送的组播数据，将传送到该 VPN 内的所有 PE 上，没有组播数据接收者的 PE 只是简单丢弃该报文。随着组播业务流量的增加，将对骨干网和 PE 路由器造成较大的额外负担，因此方案 1 没有做到业务流量最优化。

参考 PIM – SM 的基于流量门限的组播树切换机制[5]和方案 3 的思想，可以对方案 1 优化。在骨干网上引入一种动态树，一旦某个 VPN 的特定私网组播流量超过系统设定的切换门限，我们就在公网上建立一个新的组播树，并通过原有的多点组播隧道通知所有需要该组流量的 PE 动态加入该新的组播树，此后该组的私网流量将切换到新创建的组播树上。由于该新创建的隧道只用于传送大流量数据，因此我们称之为多点数据隧道；VPN 内的所有组播控制报文仍然通过原有的多点组播隧道传送。多点数据隧道是动态创建和维护的组播树，VPN 内的 PE 根据其所连接的 Site 中是否有对该业务感兴趣的接收者来决定加入或退出该多点数据隧道以及是否接收组播业务，这就保证了私网业务流量在核心网上的按需传输，优化了核心网的性能。

4　总结

本文介绍的多点隧道的组播 VPN 方案，不需对骨干网核心路由器进行升级和修改，不限制 VPN 用户使用的组播协议的类型，实施代价小，对用户透明，配合 BGP/MPLS 可提供一定的安全性保障。

参考文献

1 Gleeson B., J. Heinanen, G. Armitage, et al, A Framework for IP Based Virtual Private Networks[S], IETF RFC 2764, 2000.2.

2 E. Rosen, Y. Rekhter, BGP/MPLS VPNs[S]; IETF RFC2547bis, 2002.1.

3 Eric C. Rosen, Yiqun Cai, et al, Multicast in MPLS/BGP IP VPNs[S], Internet Draft, 2004.12.

4 D. Estrin, D. Farinacci, A. Helmy, Protocol Independent Multicast – Sparse Mode (PIM – SM)[S], IETF RFC 2362, 1998.

5 Beau Williamson, Developing IP Multicast Networks (Volume 1)[M], CISCO Press, 2000.

6 Simpson, IP in IP Tunneling[S], IETF RFC1853, 1995.10.

7 Cisco, Multicast VPN Data Sheet[R], http://www.cisco.com/Juniper, Multicast over Layer 3 VPNs Overview[R], http://www.juniper.net/.