目 录

MSDP

MSDP 简介

MSDP 概述

MSDP 是 Multicast Source Discovery Protocol（组播源发现协议）的简称，是为了解决多个 PIM-SM（Protocol Independent Multicast Sparse Mode，协议无关组播—稀疏模式）域之间的互连而开发的一种域间组播解决方案，用来发现其它 PIM-SM 域内的组播源信息。

在基本的 PIM-SM 模式下，组播源只向本 PIM-SM 域内的 RP 注册，且各域的组播源信息是相互隔离的，因此 RP 仅知道本域内的组播源信息，只能在本域内建立组播分发树，将本域内组播源发出的组播数据分发给本地用户。如果能够有一种机制，将其它域内的组播源信息传递给本域内的 RP，则本域内的 RP 就可以向其它域内的组播源发起加入过程并建立组播分发树，从而实现组播数据的跨域传输。

基于这一设想，MSDP 通过在网络中选取适当的路由器建立 MSDP 对等体关系，以连通各 PIM-SM 域的 RP。通过在各 MSDP 对等体之间交互 SA（Source Active，信源有效）消息来共享组播源信息。

⚠ 注意：

- MSDP 的适用前提：域内组播路由协议必须是 PIM-SM。
- MSDP 仅对 ASM（Any-Source Multicast，任意信源组播）模型有意义。

MSDP 原理

1. MSDP 对等体

通过在网络中配置一对或多对MSDP对等体，形成彼此相连的一张“MSDP连通图”，以连通各个 PIM-SM 域的 RP。通过这些 MSDP 对等体之间的接力，可以把某 RP 发出的 SA 消息传递给其它所有的 RP。

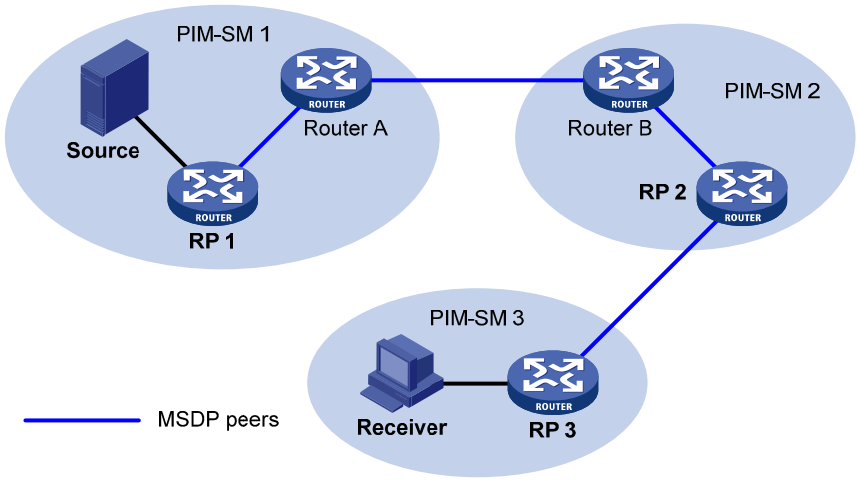

图1 MSDP 对等体的位置

如 图 1所示，MSDP对等体可以创建在任意的PIM-SM路由器上，在不同角色的PIM-SM路由器上所创建的MSDP对等体的功能有所不同：

(1) 在 RP 上创建的 MSDP 对等体

- 源端 MSDP 对等体：即离组播源（Source）最近的 MSDP 对等体（通常也就是源端 RP，如 RP 1）。源端 RP 创建 SA 消息并发送给远端 MSDP 对等体，通告在本 RP 上注册的组播源信息。源端 MSDP 对等体必须配置在 RP 上，否则将无法向外发布组播源信息。
- 接收者端 MSDP 对等体：即离接收者（Receiver）最近的 MSDP 对等体（如 RP 3）。接收者端 MSDP 对等体在收到 SA 消息后，根据该消息中所包含的组播源信息，跨域加入以该组播源为根的 SPT；当来自该组播源的组播数据到达后，再沿 RPT 向本地接收者转发。
- 中间 MSDP 对等体：即拥有多个远端 MSDP 对等体的 MSDP 对等体（如 RP 2）。中间 MSDP 对等体把从一个远端 MSDP 对等体收到的 SA 消息转发给其它远端 MSDP 对等体，其作用相当于传输组播源信息的中转站。

(2) 在普通的 PIM-SM 路由器（非 RP）上创建的 MSDP 对等体

如 Router A 和 Router B，其作用仅限于将收到的 SA 消息转发出去。

📖 **说明：**

对于通过 BSR 机制动态选举 RP 的 PIM-SM 网络来说，RP 是由 C-RP 选举产生的。为了增强其网络的健壮性，一个 PIM-SM 域内往往存在不止一个 C-RP。由于无法预计 RP 选举的结果，为了保证选举获胜的 C-RP 能始终位于“MSDP 连通图”上，需要在所有的 C-RP 之间建立 MSDP 对等体关系。而选举落败的 C-RP 在“MSDP 连通图”上所担当的角色相当于普通的 PIM-SM 路由器。

---

2. 借助 MSDP 对等体实现域间组播

如 图 2所示，PIM-SM 1 域内存在激活的组播源（Source），RP 1 通过组播源注册过程了解到了该组播源的存在。如果PIM-SM 2 和PIM-SM 3 域也希望知道该组播源的具体位置，进而能够从该组播源获取组播数据，则需要在RP 1 与RP 3、RP 2 与RP 3 之间分别建立MSDP对等体关系。

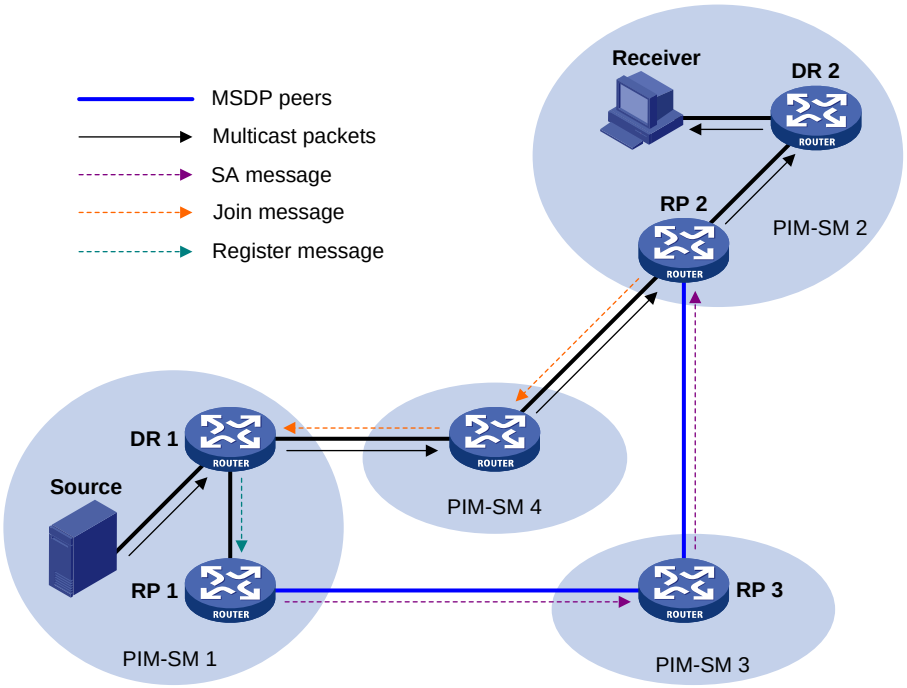

图2 MSDP 对等体示意图

借助 MSDP 对等体进行域间组播的工作过程如下：

(1) 当 PIM-SM 1 域内的组播源向组播组 G 发送第一个组播数据包时，DR 1 将该组播数据封装在注册消息（Register Message）中，并发给 RP 1。RP 1 因此获知了该组播源的相关信息。

(2) RP 1 作为源端 RP，创建 SA 消息，并周期性地向其它 MSDP 对等体发送。SA 消息中包含组播源的地址 S、组播组的地址 G 以及创建该 SA 消息的源端 RP（即 RP 1）的地址。

(3) MSDP 对等体对收到的 SA 消息进行 RPF（Reverse Path Forwarding，逆向路径转发）检查，以及各种转发策略的过滤，从而只接受和转发来自正确路径并通过过滤的 SA 消息，以避免 SA 消息传递环路；另外，可以在 MSDP 对等体之间配置 MSDP 全连接组（Mesh Group），以避免 SA 消息在 MSDP 对等体之间的泛滥。

(4) SA 消息在 MSDP 对等体之间转发，最终该组播源的相关信息将传遍所有建立了 MSDP 对等体关系的 PIM-SM 域（即 PIM-SM 2 和 PIM-SM 3）。

(5) PIM-SM 2 中的 RP 2 在收到该 SA 消息后，检查本域内是否有组播组 G 的接收者（Receiver）存在：

- 如果有接收者，RP 2 与接收者之间维护组播组 G 的 RPT。RP 2 创建（S，G）表项，向源端的 DR 1 逐跳发送（S，G）加入消息（Join Message），从而跨越各 PIM-SM 域直接加入以该组播源为根的 SPT。组播数据沿 SPT 到达 RP 2 后，再沿 RPT 向接收者转发。当接收者端的 DR 2 收到来自组播源的组播数据后，可以自行决定是否发起从 RPT 向 SPT 的切换；
- 如果没有接收者，RP 2 不会创建（S，G）表项，也不加入以该组播源为根的 SPT。

📖 **说明：**

- MSDP 全连接组：要求所有组成员之间两两建立 MSDP 对等体关系，且所有组成员均使用相同的组名称。
- 在使用 MSDP 进行域间组播时，RP 在收到组播源的信息后就不再需要依赖其它 PIM-SM 域内的 RP，此时接收者可以跨越各 PIM-SM 域内的 RP，而直接加入基于组播源的 SPT。

---

3. SA 消息的 RPF 检查规则

如 图 3所示，网络中有五个自治系统AS 1～AS 5，AS内部使用IGP互联，AS之间使用BGP或MBGP互联。每个AS中包含至少一个PIM-SM域，且每个PIM-SM域中包含至少一个RP。各RP之间建立起MSDP对等体关系，其中RP 3、RP 4 和RP 5 之间建立MSDP全连接组，并在RP 7 上将RP 6 配置为其静态RPF对等体。

📖 **说明：**

当PIM-SM域内只存在一个MSDP对等体时，该域又称为STUB域（如 图 3中的AS 4）。STUB域内的MSDP对等体可以同时拥有多个远端MSDP对等体，用户可以从中选取其中一个或多个配置为静态RPF对等体。对于来自静态RPF对等体的SA消息不进行RPF检查，直接接受并向其它对等体转发。

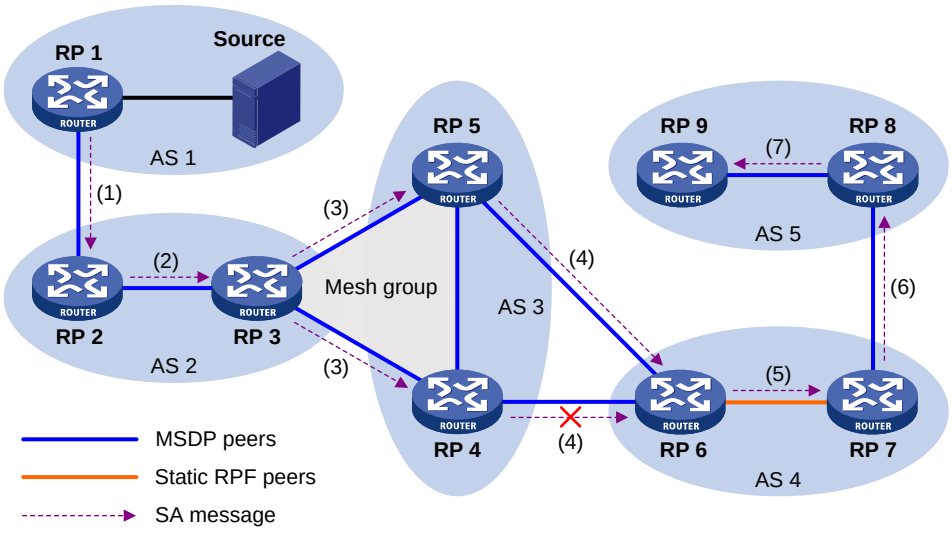

图3 SA 消息的 RPF 检查规则

对照 图 3，这些MSDP对等体将按照如下RPF检查规则处理收到的SA消息：

(1) 当 RP 2 收到 RP 1 发来的 SA 消息时

由于 SA 消息中所携带的源端 RP 的地址与 MSDP 对等体的地址相同，说明发出 SA 消息的 MSDP 对等体就是创建该 SA 消息的 RP，于是 RP 2 接受该 SA 消息并向其它对等体（RP 3）转发。

(2) 当 RP 3 收到 RP 2 发来的 SA 消息时

由于 SA 消息来自同一个 AS 的 MSDP 对等体（RP 2），且该对等体是到源端 RP 最佳路径上的下一跳，于是 RP 3 接受该 SA 消息并向其它对等体（RP 4 和 RP 5）转发。

(3) 当 RP 4 和 RP 5 分别收到 RP 3 发来的 SA 消息时

由于 SA 消息来自同一个全连接组的 MSDP 对等体（RP 3），于是 RP 4 和 RP 5 均接受该 SA 消息并不再向本组其它成员转发，而只向本组之外的其它 MSDP 对等体（RP 6）转发。

(4) 当 RP 6 收到 RP 4 和 RP 5（假设 RP 5 的 IP 地址较大）发来的 SA 消息时

尽管同处 AS 3 的 RP 4 和 RP 5 都与 RP 6 建立了 MSDP 对等体关系，但由于 RP 5 的 IP 地址较大，于是 RP 6 只接受 IP 地址较高的 MSDP 对等体（RP 5）发来的 SA 消息。

(5) 当 RP 7 收到 RP 6 发来的 SA 消息时

由于 SA 消息来自其静态 RPF 对等体（RP 6），于是 RP 7 接受该 SA 消息并向其它对等体（RP 8）转发。

(6) 当 RP 8 收到 RP 7 发来的 SA 消息时

属于不同 AS 的 MSDP 对等体之间存在 BGP 或 MBGP 路由。由于 SA 消息来自不同 AS 的 MSDP 对等体（RP 7），且该对等体是到源端 RP 的 BGP 或 MBGP 路由的下一跳，于是 RP 8 接受该 SA 消息并向其它对等体（RP 9）转发。

(7) 当 RP 9 收到 RP 8 发来的 SA 消息时

由于只有一个 MSDP 对等体（RP 8），于是 RP 9 接受该 SA 消息。

对于由其它路径到来的 SA 消息，MSDP 对等体将不接受也不转发。

4. 借助 MSDP 对等体实现域内 Anycast RP

Anycast RP（任播 RP）是指在同一个 PIM-SM 域内设置两个或多个具有相同地址的 RP，并在这些 RP 之间建立 MSDP 对等体关系，以实现域内各 RP 之间的负载分担和冗余备份。

如 图 4所示，在一个PIM-SM域内，组播源（Source）向组播组G发送组播数据，接收者（Receiver）是组播组G的成员。分别在Router A和Router B上配置相同的IP地址（称为Anycast RP地址，通常使用私有地址），同时将这些接口配置为C-RP，并在Router A和Router B之间建立MSDP对等体关系。

📖 **说明：**

通常在设备的逻辑接口（如 Loopback 接口）上配置 Anycast RP 地址。

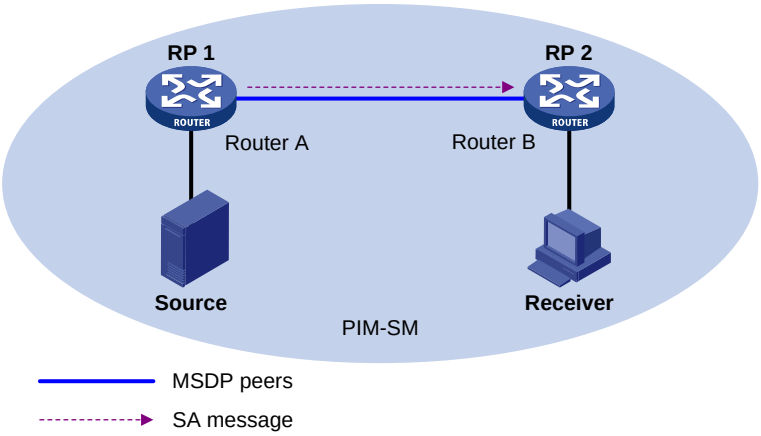

图4 Anycast RP 典型组网图

Anycast RP 的工作过程如下：

(1) 组播源选择距离最近的 RP 进行注册。如：Source 向 RP 1 注册，注册消息中封装有 Source 发出的组播数据。当该注册消息到达 RP 1 后，进行解封装。

(2) 接收者向距离最近的 RP 发送加入消息，加入以该 RP 为根的 RPT。如：Receiver 加入以 RP 2 为根的 RPT。

(3) RP 之间通过发送 SA 消息，共享注册的组播源信息。如：RP 1 创建一个 SA 消息，发送给 RP 2，该 SA 消息中封装有 Source 发出的组播数据。当该 SA 消息到达 RP 2 后，进行解封装。

(4) 接收者沿 RPT 收到组播数据后，直接加入以该组播源为根的 SPT。如：RP 2 沿 RPT 将组播数据向下转发。当 Receiver 收到来自 Source 的组播数据后，直接加入以 Source 为根的 SPT。

Anycast RP 的意义如下：

- RP 路径最优：组播源向距离最近的 RP 进行注册，建立路径最优的 SPT；接收者向距离最近的 RP 发起加入，建立路径最优的 RPT。
- RP 间的负载分担：每个 RP 上只需维护 PIM-SM 域内的部分源/组信息、转发部分的组播数据，从而实现了 RP 间的负载分担。
- RP 间的冗余备份：当某 RP 失效后，原先在该 RP 上注册或加入的组播源或接收者会自动选择就近的 RP 进行注册或加入操作，从而实现了 RP 间的冗余备份。

⚠ 注意：

- 必须为 Anycast RP 地址配置 32 位的子网掩码（即 255.255.255.255），也即将其配置为一个主机地址。
- MSDP 对等体的地址不能与 Anycast RP 地址相同。

多实例的 MSDP

属于同一实例的组播路由器各接口之间可以建立 MSDP 对等体。通过在 MSDP 对等体之间交互 SA 消息，可以实现跨域的 VPN 组播。

应用多实例的组播路由器，为其所支持的每一个实例都独立维护了一套 MSDP 机制，包括：SA 缓存、对等体连接、定时器、发送缓存和 PIM 交互的缓冲区。同时，保证不同实例之间信息隔离。所以，只有属于同一实例的 MSDP 和 PIM-SM 信息才可以交互。