第 34 卷　第 23 期　　计　算　机　工　程　　2008 年 12 月
Vol.34　No.23　　Computer Engineering　　December 2008
• 网络与通信 •　　文章编号：1000—3428(2008)23—0113—03　　文献标识码：A　　中图分类号：TP393

PIM-SM 注册报文发送的分布式实现

刘媛妮[1]，庄　雷[1]，李　丹[2]，张建辉[2]
(1. 郑州大学信息工程学院，郑州 450052；2. 国家数字交换系统工程技术研究中心，450002)

摘　要：针对指定路由器向汇聚点路由器注册过程中，主控生成报文数量过多对路由器造成负担太大的问题，提出分布式注册报文发送的实现方案，其主要包括理论分析与仿真测试。结果证明该方法减轻了主控的处理负担，缩短了数据包在路由器里的平均处理时间，减少了主控与底层信息交互的次数。

关键词：协议无关独立组播；注册报文；指定路由器；汇聚点路由器

Distributed Implementation of PIM-SM Register Message Sending

LIU Yuan-ni[1], ZHUANG Lei[1], LI Dan[2], ZHANG Jian-hui[2]
(1. School of Information Engineering, Zhengzhou University, Zhengzhou 450052);
2. National Digital Switching System Engineering & Technology Center, Zhengzhou 450002)

【Abstract】This paper proposes a distributed register message forwarding way to cope with the problem when Designated Router(DR) send register message to Rendezvous Point(PR), which may result in too many register packets on the controller plane of the router. Theoretical analysis and simulated data results show that the method lightens the controller plane's execution load dramatically, shortens the equal execution time of multicast packets in the router, and reduces the interactive time between controller plane and data plane.

【Key words】 Protocol Independent Multicast-SparseMode(PIM-SM); register message; Designated Router(DR); Rendezvous Point(RP) router

1　概述

文献[1]利用随机Petri网(Stochastic Petri Net, SPN)模型对整个PIM-SM[2-3]复杂的协议行为进行了建模，并在其SPN模型的基础上，结合路由器的实现，对协议中每种消息消耗的路由器处理负载进行了分析和实验，发现Register消息消耗的路由器处理负载比较多。

从协议本身来看，组播源的 DR 需要在组播会话初期把组播包封装在 Register 消息中发送到 RP，直到收到 RP 的 Register-Stop 消息为止。对于该过程，传统的发送 Register 报文的方法是将整个组播数包上报到主控，主控运行 PIM-SM 协议，根据组播数据包的组地址，选举出对应的 RP，将组播数据包封装成 Register 消息，再下发并通过交换网络发送(即上报－封包－下发)，如图 1 所示。由于 Register 消息数量大，如果将收到的数据包全部上报主控，由主控来完成 RP 地址的查找以及数据包的封装，会增大主控的处理负担。

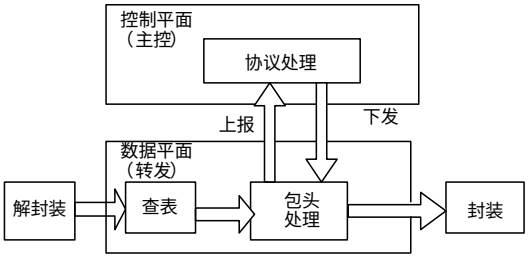


图 1　数据包处理流程

为减轻主控的处理负担，通常是在收到注册停止报文之前定期将部分数据包上报。这样在一定程度上减轻主控的处理负担，但是会将注册过程中发送的大部分数据包丢弃。

本文提出一种在路由器上改进的处理方法。理论与仿真分析表明，从DR收到源发送的组播数据包到收到RP的注册停止消息的时间段$T_{register}$中，使用该方法能大大降低组播数据包在路由器里的平均处理时间，并且可以减少数据平面与控制平面之间的交互次数，降低主控的负担。

2　PIM-SM 注册报文发送的分布式实现

2.1　问题分析

当前路由器的处理方式是，当收到直连源发送的组播数据包时，如果转发表没有相应的转发表项，就将该组播数据包上报主控。这种解决方法存在的一个严重问题，就是在注册过程中，组播数据包不停地上报主控，由主控查表、封装，再下发到路由器的数据平面，增加了主控的处理负担。

本文提出一种分布式实现方法，使数据包尽可能都在数据平面转发，即将组播数据包的组地址以及对应的 RP 地址存放在数据平面，发送注册报文时，就可以根据数据平面提供的 RP 信息，将数据包封装后转发。

根据 PIM-SMv2，注册报文发送的分布式实现在数据平面需要解决的问题如下：

(1)判断收到组播包的路由器接口是不是 DR；

(2)如何只将组播包的组地址以及对应RP存入数据平面，并且该信息只在整个$T_{register}$时间段内有效，注册过程结束后删除，以释放其占用的空间；

(3)怎样获得组播组地址对应的 RP 地址；

基金项目：国家“863”计划基金资助重点项目“大规模接入汇聚路由器系统性能与关键技术研究”
作者简介：刘媛妮(1982－)，女，硕士，主研方向：IP 组播路由协议，P2P 技术；庄　雷，教授、博士生导师；李　丹，博士；张建辉，讲师
收稿日期：2008-03-05　**E-mail：**lynzkxd_82@163.com

(4)在控制平面将 RP 信息下发到数据平面之前，如何限制组地址的上报。

2.2 总体设计

根据上文中提出的问题，给出分布式实现的设计思路，具体流程如下：

(1)在路由器启动后，通过运行协议 PIM-SM 将 DR 接口、转发表等信息形成表项下发给数据平面。

(2)路由器的 DR 接口只将收到的直连源发送的具有相同组地址的第 1 个组播数据包上报主控，用于触发上层 PIM-SM 协议进行 RP 选举，并将 RP 与组地址形成一个表项(即 RP 表项，具体定义见下文)下发到数据平面。

(3)第 2 个数据包到 RP 表项下发到数据平面之前的数据包需要先存入缓冲区，以等待 RP 表项下发后根据 RP 地址封装转发。

(4)RP 表项建立后再收到的组播数据包只需在数据平面查 RP 表，根据 RP 表里的 RP 信息，将组播数据包封装后转发。

在数据平面定义以下表项：

定义 1 DR 表：用于判断收到组播数据包的路由器接口是否是 DR 接口。通过查询该表可以判断收到组播包的接口是不是 DR 接口。

定义 2 RP 表：维护多条 Group 与 RP 的对应关系，以及此 RP 的状态信息。路由器通过查询该表获得组播组地址对应的 RP 地址，以及当前的 RP 状态，即是否允许注册。

定义 3 Group 表：记录组播数据包的组地址，通过查询该表来判断是否需要将组地址信息上报主控以达到限制上报报文数量的目的。

2.3 具体实现

(1)DR 表

DR 表是一个 120 bit 的列表(见图 2)，分别对应路由器的 120 个接口。每一位的值为 1/0，表示其对应路由器接口是否为 DR 接口。

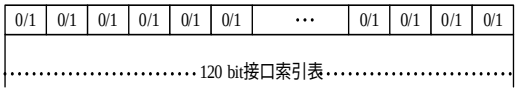


图 2 DR 列表格式

单板启动后，会请求主控下发 DR 接口信息。协议收到请求后，进行 DR 接口的选举，最终形成一个 120 bit 的列表。主控会定期将 DR 更新消息下发给单板，以保证平面一致性。

(2)RP 表

RP 表格式见表 1。

表 1 RP 表格式

Group address	RP address	Re-s (1/0)	B (1/0)
Group address$_1$	RP address$_1$	Re-s$_1$	B$_1$
Group address$_2$	RP address$_2$	Re-s$_2$	B$_2$
…	…	…	…
Group address$_n$	RP address$_n$	Re-s$_n$	B$_n$

在表 1 中，Re-s 表明是否收到注册停止消息；B 位是边界位，为了方便转发平面封装数据包时计算校验和，此处不再赘述。

RP 表初始化为空，主控收到单板上报的组地址后，运行 PIM-SM 协议，将 RP、Re-s 位、B 位信息形成如图 2 中的一个 RP 表项下发给单板。这样，再收到同一个组 G 的组播数据包时，就可以直接在数据平面转发。如果某一个 RP 失效，需要及时将该信息下发给单板，以删除对应的 RP 表项；如果协议收到注册停止消息，要将 Re-s 位置 1，不允许注册。

(3)Group 表

该表初始化为空，当查 RP 表没有对应的 RP 表项时，就需要查组地址列表，看是否需要将该组地址上报。

2.4 实现流程

当路由器的接口收到直连源发送的组播数据包并且查转发表没有对应的表项时，其实现流程如下：

(1)以收到组播包的路由器的端口号+接口号，查 DR 列表，查表结果为 1 转入第(2)步；为 0，丢弃。

(2)查 RP 表，有对应的 RP 表项，转入第(3)步；没有，转入第(6)步。

(3)Re-s 位为 1，不允许注册，丢弃；为 0，允许注册，转入第(5)步。

(4)RP 地址是与 DR 地址相等，不需要注册，丢弃；不相等，转入第(5)步。

(5)将数据包按照注册报文的格式封装并转发，数据包的目的 IP 地址为 RP 地址。

(6)组地址列表里有对应的组地址，将数据包放入缓冲区等待 RP 表项下发后处理；没有，转入第(7)步。

(7)将组地址填入组地址列表里，并将该地址上报主控。

注册过程实现流程示意图如图 3 所示。

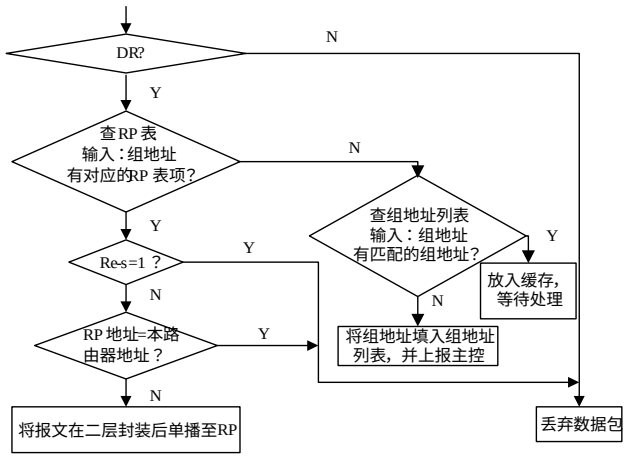


图 3 注册过程实现流程

3 理论分析与仿真

3.1 理论分析

在时间段T_{register}内，根据数据包在路由器里的不同处理方法，可以将处于同一个组的组播数据包分为以下 3 类：

(1)C_1：收到源发送的第 1 个组播数据包。该数据包被上报主控，触发协议针对该组地址进行RP选举。

(2)C_2：第 2 个数据包到RP表项下发之前收到的数据包。路由器收到这些数据包后先将其缓存，待RP表项下发后再处理。

(3)C_3：RP表项下发后收到的数据包，这些数据包直接被路由器转发。

表 2 中列出了使用改进方法后，3 类数据包在路由器需要进行的操作以及对应的时间。带*表示需要该步处理。

表 2 不同类型数据包在路由器中的处理时间

包类型	解封装t_p	查表				包头处理t_a	上报、协议处理、下发t_e	封装t_p
		转发表t_f	DR 表t_d	RP 表t_r	Group 表 t_g			
C_1	*	*	*	*	*	*	*	*
C_2	*	*	*	*	*	*		*
C_3	*	*	*	*		*		*

设源发送组播数据包的速率为每单位时间n个，则在时间段T_{register}内，3 类数据包的数量及平均处理时间如下：

(1)$T_{c1}=2t_p+t_f+t_d+t_r+t_g+t_a+t_e$, $N_{c1}=1$

(2)从第 2 个数据包到RP表项建立之前$t_p+t_f+t_d+t_r$时间段内的数据包，其处理时间为

$$T_{c2}=2t_p+t_f+t_d+t_r+t_g+t_a+t_w$$

其中，t_w为数据包放入缓冲区到RP表项下发后的等待时间。

$$N_{c2}=n\times((t_p+t_f+t_d+t_r+t_g+t_a+t_e)-(t_p+t_f+t_d+t_r))-1=n\times(t_g+t_e)-1$$

数据包查Group列表没有后，先放入缓冲区，因此，每一个数据包还有一个在缓冲区里的等待时间 tw_i。其中，第 i 个数据包的等待时间：$tw_i=T_3^{'}-\dfrac{i\times T_{c2}^{'}}{N_{c2}}$，$T_{c2}^{'}=t_e$(即上报-处　理-下发的时间)，则$N_3$个数据包总的等待时间为

$$T_w=\sum_{i=0}^{N_{c2}-1}tw_i=\frac{(N_{c2}+1)\times T_{c2}^{'}}{2}$$

因此，对于C_2，总的处理和等待时间为

$$T_{c2}\times N_{c2}+T_w$$

(3)C_3为从RP表项建立之前$(t_p+t_f+t_d+t_r)$时间到t时间段内收到的数据包为

$$T_{c3}=2t_p+t_f+t_d+t_r+t_a$$

$$Nc3=n\times(t-((t_p+t_f+t_a+t_e)-(t_p+t_f+t_d+t_r)))=n\times(t-t_e-t_a+t_d+t_r)$$

由(1)~(3)可得，使用本方法后数据包在时间段T_{register}，数据包在路由器里的平均处理时间为

$$T_{\text{new}}=\frac{T_{c1}+T_{c2}\times N_{c2}+\dfrac{(N_{c2}+1)\times T_{c2}^{'}}{2}+T_{c3}\times N_{c3}}{n\times t}$$

而使用传统的方法：$T_{\text{traditional}}=2t_p+t_f+t_a+t_e$。

可以看出，使用改进的方法后数据包的平均处理时间会降低，同时随着n, t值的增大，这种变化会越来越明显。另外，在T_{register}时间内，传统的处理是将所有的数据包都上交给主控来处理，而改进后的方法中上报主控的报文只有该组播组的第 1 个数据包，这样做大大降低了主控的处理负担。

3.2　仿真

发送组播注册报文的测试环境如图 4 所示，测试环境包括测试系统和待测系统。通过测试控制台和被测系统控制台配合完成测试配置，测试结果在测试控制台或者被测系统控制台查看。在测试系统控制台可以配置使测试仪以一定的速率发送特定的数据包，在待测系统控制台 1, 2 可以配置转发单元中的各种表项，模拟数据包在不同情况下的处理过程。

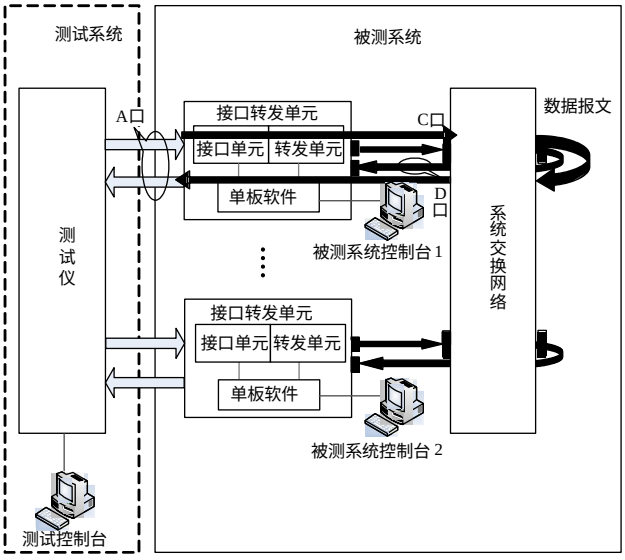
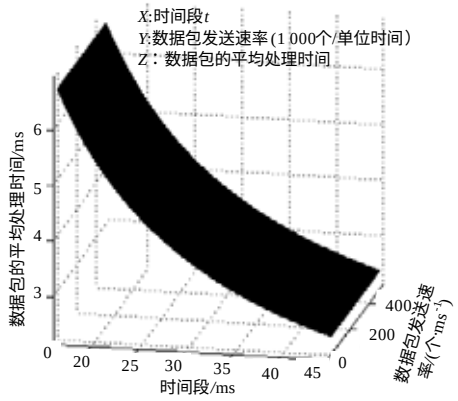
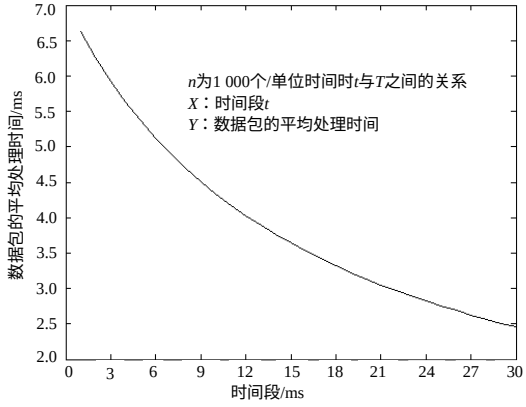


图 4　测试场景

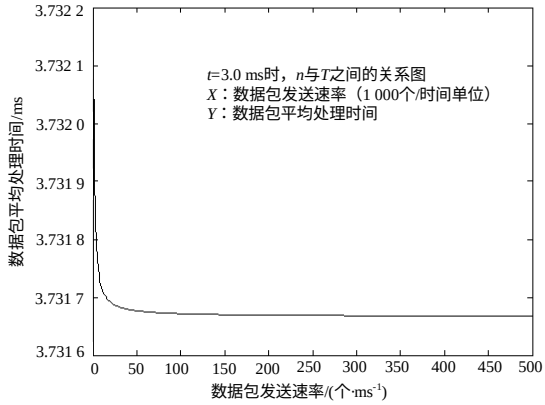
通过测试，得出表 2 中各参数的值分别为：t_p=40 ns, t_f=200 ns, t_r=200 ns, t_d=1 ns, t_a=300 ns, t_g=300 ns, t_e=1.5 ms。用Matlab对数据包在路由器里的平均处理时间T与n, t之间的关系进行分析得到图 5。其中，a表示n, t, T之间的关系；b, c分别为t=3.0 ms和n=1 000 个/单位时间时T分别与n和t之间的关系。



(a)n, t, T 之间的关系



(b) n 为 1 000 时 t, T 之间的关系



(c) t 为 3.0 ms 时 n, T 之间的关系

图 5　仿真数据曲线

由图 5(a)可得影响 T 的主要参数为 t。从图 5(b)中可以看出，在 t 值很小时，使用分布式处理方案的 T 要大于传统处理方案的值，但是随着 t 值的增大，前者的值就会越来越小，其优越性也就体现出来了。通常情况下，时间段 t 与 DR 收到注册停止消息的时间长短相关，从理论上讲，该时间是很长的，因此，使用分布式方法来完成注册报文的发送将大大降低数据包在路由器里的平均处理时间。 （下转第 118 页）