ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.4, April 2010, pp.827−837
doi: 10.3724/SP.J.1001.2010.03516

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

汇聚组播:新型 MPLS 服务质量组播体系结构*

江 勇[1+], 胡松华[1,2]

[1](清华大学 深圳研究生院,广东 深圳 518055)

[2](清华大学 计算机科学与技术系,北京 100084)

Rendezvous Multicast: A Novel Architecture for MPLS QoS Multicast

JIANG Yong[1+], HU Song-Hua[1,2]

[1](Graduate School of Shenzhen, Tsinghua University, Shenzhen 518055, China)

[2](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: E-mail: jiangy@mail.sz.tsinghua.edu.cn

Jiang Y, Hu SH. Rendezvous multicast: A novel architecture for MPLS QoS multicast. *Journal of Software*, 2010,21(4):827−837. http://www.jos.org.cn/1000-9825/3516.htm

Abstract: IP multicast and MPLS (multi-protocol label switching) are proposed to support emerging network applications effectively, extending current IP routing and switching mode in different ways. It is still in progress to combine IP multicast with MPLS properly. In this paper, a novel architecture, called Rendezvous MPLS Multicast, is proposed to support IP multicast in MPLS networks and to have the scalability of multicast routing, to achieve the label space reduction, and to solve some practical problems in implementation. A two-layered structure of control plane without multicast routing protocols is implemented by overlaying novel service control plane over the existing routing control plane to support multicast state aggregate. The interface between the two planes is formulated to support the interaction and cooperation. Moreover, the label distribution process for MPLS P2MP (point to multi-point) connection, with RSVP-TE (resource reservation protocol-traffic engineering) protocol, is extended to support aggregate of multiple label switching paths with traffic engineering and guarantee of quality of service. An algorithm of selecting Rendezvous Routers for Rendezvous MPLS Multicast has been presented. A test-bed with Linux-based implementations of MPLS multicast router and IP multicast service control system has also been constructed. The experimental results show its efficiency in terms of label space reduction and multicast traffic balancing, with the whole system adapting to the dynamic change of group members and network topology.

Key words: MPLS (multi-protocol label switching); IP multicast; quality of service; control plane; P2MP (point to multi-point); RSVP-TE (resource reservation protocol-traffic engineering); Linux

摘 要: 为支持新兴网络应用,IP 组播(multicast)和 MPLS(multi-protocol label switching)技术分别从不同方向扩展了当前的 IP 路由和交换模式.MPLS 和 IP 组播的结合是当前研究的一个热点,MPLS 网络中的服务质量组播面临着标签资源匮乏、组播路由状态的可扩展性以及具体实现上的困难.针对这些问题,提出了基于汇聚方法的新型 MPLS

* Supported by the National Natural Science Foundation of China under Grant No.60503053 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z209 (国家高技术研究发展计划(863))

Received 2008-04-24; Revised 2008-08-07; Accepted 2008-10-27

服务质量组播体系结构,提出在现有的路由控制平面上叠加一层面向IP组播服务的控制平面,取代组播路由协议并支持组播聚集,形成2层控制平面结构.定义了两平面之间的协作和交互方式,并通过扩展RSVP-TE(resource reservation protocol-traffic engineering) P2MP(point to multi-point)协议,在新的体系结构中融合了服务质量控制能力.另外,还探讨了汇聚组播中基于距离约束选择汇聚路由器的算法,实现了基于Linux的MPLS组播路由器和IP组播服务控制系统,并组建了实验平台.实验和模拟结果表明,基于汇聚组播的双平面网络控制结构能够适应组播用户和网络拓扑的动态变化,能够有效节省MPLS标签资源,平衡网络中组播流量的分布.

关键词: MPLS(multi-protocol label switching);IP组播;服务质量;控制平面;P2MP(point to multi-point);RSVP-TE(resource reservation protocol-traffic engineering);Linux

中图法分类号: TP393 **文献标识码:** A

如何优化网络资源利用,如何在IP网络中支持服务质量保证一直是研究的热点.以此为目标,IP组播(multicast)和MPLS(multi-protocol label switching)从不同方向上扩展了当前的IP路由和交换模式.

IP组播[1,2]与IP广播、单播相对应,是一种点到多点的分组传递模式.IP组播与单播相比,其主要优势是:

1) 当源主机同时向多个目的主机分发数据时,能够有效缓解源主机的带宽瓶颈;
2) 由路由器复制分组,能够减少链路重复占用,提高网络资源的利用率.IP组播被认为是支持新兴流媒体应用(Internet protocol television,简称IPTV等)的重要手段[3,4].

MPLS[5,6]对现有基于IP的分组路由和交换模式做出了重大改进,融合了面向连接和IP路由这两个特性,支持流量工程和服务质量控制.然而在当前,MPLS除了在承载IP单播方面发展得相对成熟之外,在支持其他服务上仍然也在不断地完善之中.其中,MPLS对IP组播的支持和MPLS网络中服务质量组播体系结构是相关领域的研究热点.

本文针对MPLS网络中的服务质量组播在体系结构上面临的标签资源匮乏、组播路由状态的可扩展性以及具体实现上的困难,提出了新的MPLS服务质量组播体系结构,实现了基于Linux的MPLS服务质量组播系统.新的组播体系结构通过单条路径共享和汇聚的方法节省MPLS标签资源,克服了基于分发树的聚集组播[7,8]中存在的带宽浪费问题,适用于大量组存在且组成员不断变化的情况.新的组播体系结构在现有的路由控制平面之上叠加了面向组播服务的控制平面,用以取代组播路由协议,避免了IP组播路由状态的可扩展性问题.我们扩展了RSVP-TE(resource reservation protocol-traffic engineering) P2MP(point to multi-point)[9]标签分发过程,用于建立基于汇聚组播的标签交换路径(label switched path,简称LSP),并为新的组播体系结构融合了服务质量保证能力.

我们实现了Linux环境下的MPLS组播路由器和MPLS组播服务控制系统,组建了实验环境.通过大量的实验和模拟,结果表明,这种双控制平面的组播体系结构虽然在节省MPLS标签资源这一单项指标上落后于聚集组播,但在标签资源和带宽资源整体性能上却优于聚集组播,并且能够适应组播组和网络拓扑的动态变化.

本文第1节综述与本文相关的研究工作.第2节描述新型MPLS组播体系结构的设计和实现、RSVP-TE协议扩展模块的实现.第3节通过实验和模拟对比我们提出的汇聚组播、基于RSVP-TE P2MP非聚集组播、聚集组播三者的性能.第4节总结全文并指出下一步的研究方向.

1 相关工作

IP组播虽经多年研究,依旧面临很多困难[10].文献[11]从ISP的角度分析了现有的组播协议,认为主要困难是:1) 组的管理、接纳控制等尚未进行规范;2) 组播地址分配与位置无关,导致聚集困难,而且组播地址数量有限,很难解决地址重复的问题;3) 组播对网络安全的影响远大于单播;4) 组播应用只对网络运营商和内容发布者有利,而对接收者没有吸引力;5) 目前组播路由协议与现有的Internet核心无状态模型不符.

尽管如此,IP组播的研究仍从两个方面作了有益的探索:1) 多点通信模式对于支持新兴应用是必要的;2) 使用网络层组播方法时需要对路由状态、多点之间的连接加以管理,避免对现有网络造成不利影响.近年来,

应用层组播[12]有了较快发展,在缓解源主机的带宽瓶颈上部分地满足了一些流媒体应用的需求,但在防止重复消耗带宽上却没有发挥出组播的优势.因此,网络层的组播仍有重要意义,但却需要引入新的思路.

在 MPLS 网络中,承载 IP 组播的基本方法是建立 P2MP 和 MP2MP(multi-point to multi-point)的标签交换路径,主要有两个实现难点:一是需要一种映射机制把 IP 组播分发树映射到 MPLS 多点连接上;二是需要一种信令机制建立多点连接.在设计和实现这两种机制时,都要考虑如何避免基于 IP 网络的组播所面临的问题.而且 MPLS 网络中的标签也是一种有限资源,使用的标签数量会影响路由器对 MPLS 帧转发的速度[13].

在 MPLS 信令机制方面,文献[9]使用不依赖组播路由协议的信令机制建立 P2MP 连接,该机制通过扩展 RSVP-TE[14]来实现,可建立满足服务质量要求的 P2MP 标签交换路径.然而,在存在大量组播组的情况下,只依赖该机制会消耗大量的标签资源.文献[9]没有给出组播分发树的映射机制.在映射机制方面,文献[7]提出了聚集组播方法,通过把组和分发树从概念上分开,提出用一个分发树为多个组传递数据.这种方法能够显著减少核心网络中的组播状态,减少标签资源的消耗,但是存在着带宽浪费的问题,当组和分发树之间不完全匹配时,一些组只是分发树的真子集,分发树会把组的流量传递到没有组成员存在的主机上.虽然文中提出为这种漏匹配建立多个分发树并把漏匹配约束在一定范围内,但当组中的成员发生动态变化[15]时,仍然会产生带宽浪费,且组在多个分发树之间迁移会带来显著的协议开销.

文献[8]模拟了聚集组播在 MPLS 网络中的性能,文中引入了一个管理实体 Tree Manager,用集中控制的方式在标签交换路由器 LSR(label switching router)上建立组播转发状态,在边界路由器 LER(label edge router)上建立组播流到 LSP 的映射.文献[16]使用 RSVP-TE 协议为 IP 组播建立 MPLS P2MP 标签交换路径,使用 NIMS(network information management system)控制实体在 LSR 和 LER 上建立 MPLS 状态.

文献[17–19]提出了用合并(merge)和非对称隧道(asymmetric tunneling)的方法,减少标签资源节约.其基本思想包括:1) 对于多条到达同一个目的地的 LSP 可以用合并的方法,共用尾部的一段 LSP;2) 对于到达不同目的地的 LSP 可以用标签栈的方法,构造多个层次的 LSP,逐段共用底层的 LSP.这种方式与 ATM 网络中的 VCC (virtual channel connection)和 VPC(virtual path connection)两层虚拟拓扑之间的映射有相似之处.采用这种方式可以最大程度地聚集 LSP,但却被认为是 NP 难问题[20],目前只有启发式算法,而且其中协议的设计和实现比较困难,有待进一步探索.

2 设计和实现

2.1 基本思想

当前解决 MPLS 服务质量组播的方法存在如下一些问题:

1) 只有针对个别目标的算法和模拟结果,缺乏整个体系结构上的实际支持.没有说明如何收集需要的信息、如何建立标签交换路径、如何与主机交互;
2) 引入控制实体[7,21–24]后,控制平面的功能设计不够合理.集中式的控制实体通常较难准确、及时、连续地获取网络实时流量信息;
3) 聚集组播中基于分发树的聚集粒度较大,虽然能够有效减少标签资源消耗,但却存在带宽浪费现象,而且当组的成员发生变化时,原来聚集在一起的一致的组会变得不一致.

我们认为,这些困难可以通过重新设计组播的体系结构来克服.我们推广了控制实体的思想,引入了服务控制平面,组成一个两层的 MPLS 控制平面,如图 1 所示.数据转发平面和路由控制平面与现有路由器的体系结构一致,数据转发平面负责 MPLS 帧的转发,路由控制平面由分布在各路由器上的控制实体组成,

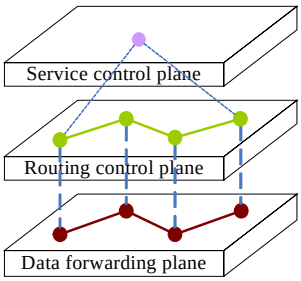

Fig.1 Logical structure of MPLS network
图 1 MPLS 网络的逻辑结构

运行着现有的路由协议[25,26]、标签分发协议[14,27].叠加的服务控制平面是面向具体服务的控制实体,负责收集和处理具体服务的信息,实现面向特定服务的优化,如组播聚集,再利用路由控制平面里信令协议、路由协议构造出需要的数据传递路径.

服务控制平面的思想源自现有的 Tree Manager,NIMS,不同之处在于,我们重新规划了控制平面的功能和控制平面间的交互.上述控制实体直接在路由器建立分发树的状态,没有考虑到与现有标签分发过程和服务质量控制过程的融合,因此依赖于网络的实时状态信息,实现也较为困难.我们应用服务控制平面的思想,控制实体只保留与 IP 组播服务直接相关的信息,把有突变特性的网络实时状态放在路由控制平面中,因此控制实体只记录组的分布、组聚集策略,分发树的实际建立由路由控制平面完成.

双层控制平面的组播体系结构具备一些优点:

1) 避免了在路由控制平面引入过多的特定服务的信息.我们认为,IP 组播路由协议试图在路由控制平面引入组播组的分布等组播状态信息,造成组播路由不仅依赖于单播路由协议而且受到组播组动态变化的困扰[28],支持大量组时协议开销急剧增加;
2) 我们设想在路由控制平面只反映出网络的本质特性,放入基本的、面向网络的控制元素.对 MPLS 网络而言,IP 路由协议和面向连接的信令协议体现了 MPLS 的主要特点,其他如组播、任播等都可以通过在服务控制平面增加控制实体实现,新的服务只需要扩展服务控制平面;
3) 服务控制平面能够完成针对特定服务的优化,本文中提出的汇聚组播方法对组播的聚集能够有效减少标签的消耗,能够适应网络状态和组的动态变化,并融合了服务质量控制能力.

服务控制平面的实现方式可以是集中式的,由存储资源丰富和较强数据处理能力的工作站来实现,这样能够有效地缓解路由控制平面上内存、计算能力等资源不足的问题.集中式服务控制平面的一个缺点是单点故障,但是我们认为可以用集群、分布式处理、冗余备份等方法增强其可靠性,而且服务控制系统不负责分组转发,对网络的实际影响是很有限的.

2.2 新型MPLS服务质量组播的体系结构

2.2.1 IP 组播服务

支持 IP 组播的 MPLS 网络中,物理实体主要是路由器(LER,LSR)和 IP 组播服务控制系统,如图 2 所示.为了完成组播服务,入口和出口 LER 通过 IGMP(Internet group management protocol)[29]获取组在本地的分布,然后通过驻留在 LER 上的服务控制系统接口,把这些信息传递到 IP 组播服务控制系统.IP 组播服务控制系统从所有 LER 上收集网络中所有组在网络中的分布,计算出基于汇聚方法的 LSP 聚集策略和服务质量控制策略,然后把这些策略返回给 LER.LER 上的 RSVP-TE 信令协议模块基于这些服务策略,通过本地路由模块(最短路径路由或约束路由[25,26])构造出需要的标签交换路径.同时,RSVP-TE 信令过程把服务质量保证技术和标签分发融合在一起,既可以提供基于每次连接的集成服务[30],也可以作为区分服务的接纳控制过程.

体系结构的层次模型如图 3 所示,分为:

1) 应用层的交互.位于用户主机上的组播应用进程与组播内容提供者设置的控制服务器之间进行单播会话,协商组播地址、认证、授权等信息;

2) 网络层的用户网络信令交互.组播应用进程建立到组播地址的 Socket 连接,操作系统会与本地组播路由器进行 IGMP 会话.本地组播路由器通过 IGMP 协议获取组在本地的分布,并通过服务控制协议把这些信息传递到 IP 组播服务控制系统.IP 组播服务控制系统为 MPLS 网络生成组播聚集策略和服务质量控制策略;

3) 网络层的网络信令交互.本地组播路由器基于服务策略,通过标签分发协议和路由协议构造穿越 MPLS 网络的标签交换路径.

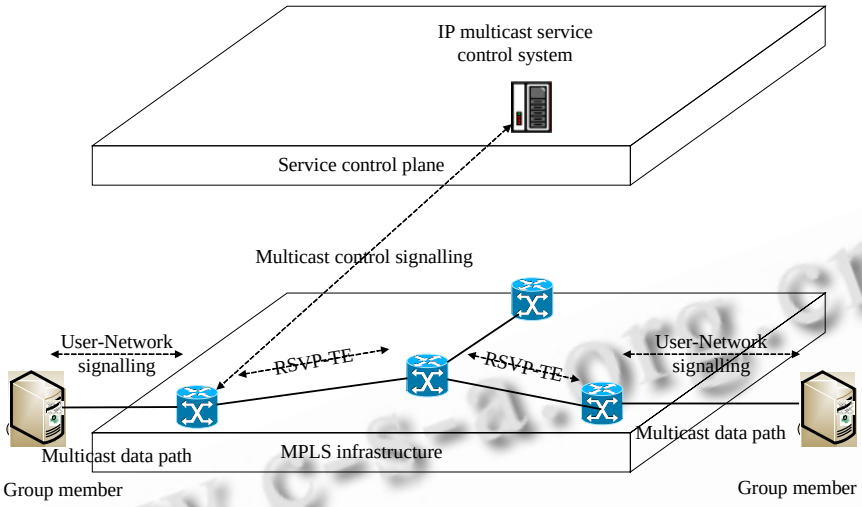

Fig.2 IP multicast in MPLS networks

图 2 支持 IP 组播的 MPLS 网络

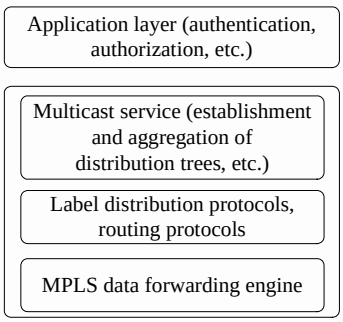

Fig.3 Layers of MPLS multicast architecture

图 3 MPLS 组播体系结构分层

2.2.2 汇聚组播

IP 组播服务控制系统提供的一个核心功能就是组播的聚集,减少标签资源的消耗.我们提出的聚集方法称为汇聚组播 RMM(rendezvous MPLS multicast),如图 4 所示.

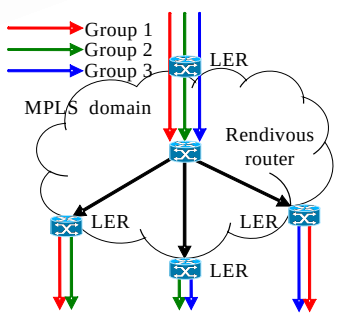

Fig.4 Rendevous MPLS multicast

图 4 汇聚组播 RMM

基本思想是:在网络中设置一个或多个汇聚路由器,从汇聚路由器到出口 LER 之间的标签交换路径 LSP 为所有组共享,因此对每个 LER 只需建立一条 LSP 就可以为所有组传递.这种方法比当前的基于分发树和分发子树的聚集有两个优势:1) 没有带宽浪费;2) 适应组的动态变化.付出的代价是 LER 与 IP 组播服务控制系统之间的交互增加了用户加入组播组的延迟.我们的考虑是:LER 与 IP 组播服务控制系统之间的交互采用的是端到端的通信机制,而不是传统的组播路由协议(PIM-SM,CBT).RSVP 这样的逐跳处理,在 IP 组播服务控制系统有足够处理能力的情况下,处理延迟可以忽略,所以这段延迟只相当于为组播路由协议增加了一跳的距离,是可以接受的.

令网络中边界路由器(LER)的数量为 N_r,组播组(multicast group)的数量为 N_g,那么 RMM 中标签交换路径 LSP 的数量级为 $O(N_r)$,不聚集时路径数量级为 $O(N_g)$,当 $N_g>>N_r$ 时,RMM 相比于不聚集的优势是显著的.聚集组播[7]中路径的数量级在理想状态下为 $O(1)$,因为只有一棵覆盖整个网络的树.但是,因聚集组播造成的带宽浪费

也是显著的.而对于一些类似的子树聚集的方法,也是用带宽资源置换标签资源.同时,因为子树的可能数量是LER集合子集的数量,达到$O(2^{N_r})$.聚集组播的改进算法中提出为组播树的聚集施加约束,如果带宽浪费超出约束,则新建分发路径.最坏情况下,这种策略产生的分发树的数量级也达到$O(2^{N_r})$.而且,组的分布通常是变化的[15],组在不同分发路径上迁移的开销也是显著的.在我们提出的RMM方法中,由于是单条路径的共享和聚集,组的动态变化只需在汇聚点切换即可.

RMM 算法的一个关键之处就是选择合适的汇聚路由器,这与组播路由协议(PIM-SM,CBT)有类似之处.但是,由于我们采用了IP组播服务控制系统,掌握了组的分布信息和网络拓扑信息,所以能够优化汇聚路由器选择算法.优化的目标可以是网络中的流量分布、组播代价等.本文中,我们提出了一种基于距离约束的汇聚路由器选择算法,选择到与 LER 集合平均距离最近的汇聚路由器.本文中,我们使用跳数作为距离的度量值,因为跳数越小,LSP占用的网络资源越少.

2.2.3 汇聚路由器选择算法

我们利用IP组播服务控制系统选择汇聚路由器,克服了困扰IP组播路由协议的汇聚点选择问题.

汇聚路由器的选择取决于网络的拓扑、参与组播的边界路由器集合、网络优化的目标.网络优化目标是设计和评价选择算法的依据,可以是网络流量平衡或与平均距离最短.本节中,我们提出基于距离约束的最短距离选择算法,其中距离的度量是跳数(hop).一条路径的跳数越少,占用的网络资源越少.图5给出了距离约束选择算法的伪码描述,图6是该算法的基本思想.

Assumptions: N_r nodes, N_e edge nodes denoted as 0,1,…,N_e−1
Input: Set of edge nodes *LERSet*,
Network topology repersented by an adjacency matrix $M_{adj}=(a_{i,j})_{N_r\times N_r}$ where $a_{i,j}$ denotes the adjacency relationship between node *i* and node *j*,
Max distance (hops) D_{hop};
Output: Set of rendezvous nodes, *CandidateSet*.
Algorithm:
CandidateSet=*LERSet*; //Initialization
for (each node *i* in *CandidateSet*)
for (*j*=0 to N_r)
$M[i][j]=M_{adj}[i][j]$; /**ith* row denotes the candiate node *i**/
While (1) {
for (each node *i* in *CandidateSet*) {
find adjacent nodes *j*, subject to *Distance*(*i*,*j*)⇐D_{hop};
$M[i][j]=1$;
} //broad search the *M* matrix from *i*th node, to find the adjacent node *j* to which the distance is less then D_{hop}
for (*j*=0 to N_r){
$\sum_{i=1to N_r} M[i][j]$; //sum M by column and obtain a vector with N_r elements
}
if (max($\sum_{i=1to N_r} M[i][j]$)==1) break; //exit of the algorithm
get *j*, subject to max($\sum_{i=1to N_r} M[i][j]$); /*the node *j* cover the maximum number of nodes */
put *j* into *CandidateSet*;
remove any node *i* from *CandidateSet*, with $M[i][j]=1$;
}

Fig.5 Selection algorithms of redevous router under constraints of distance

图5 距离约束的汇聚路由器选择算法

显然,在连通图中,如果 $D_{hop}\geq\lceil GloableDiameter/2\rceil$,其中,$D_{hop}$ 是汇聚路由器距离边界路由器最大跳数且$0<D_{hop}<255$,*GloableDiameter* 是以跳数度量的网络直径,那么算法一定能够收敛到一个汇聚路由器.

直观地看,聚集组播会把组播流量集中在该汇聚路由器周围,容易超出局部容量,而且由路由器失效带来的影响较大.为避免这种情况的发生,一个解决办法是收敛到多个局部的中心节点上.因此给定 $D_{hop}\geq\lceil LocalDiameter/2\rceil$,*LocalDiameter* 是局部网络的直径.实验中,D_{hop} 采用经验值30.

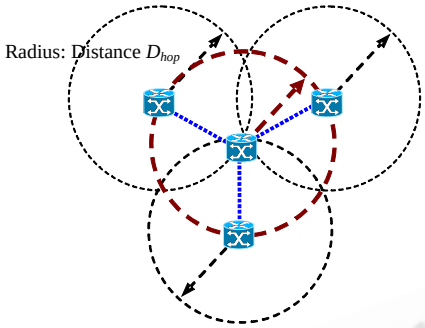

Fig.6 Distance constraints of redevous router

图 6 汇聚路由器的距离约束

2.3 实 现

我们实现了一个运行在 Linux 操作系统上的原型系统,Linux 网络内核中的 MPLS 代码基于开源项目 MPLS-Linux[31],增加了组播模块.因为本文提出的体系结构和算法不依赖于硬件结构,所以我们的设计同样适用于分布式交换路由器.

路由器实现上分为数据平面和控制平面两部分.数据平面以 MPLS-Linux 代码为基础增加了组播转发模块,并提供了配置接口.控制平面中,路由协议处理和 RSVP-TE 标签分发协议处理属于路由控制平面,IGMP 协议和服务接口模块属于我们提出的服务控制平面.

IGMP 协议模块处理 IGMP 消息,维护本地组播接口(virtual interface,简称 VIF)、组播路由块(multicast forwarding cache,简称 MFC),并把组标识(IP 组播地址){*GroupID*}传递到服务接口模块.服务接口模块收到{*GroupID*}后,连同本地 LER 的标识{*LERID*}组成消息{*GroupID*,*LERID*,*qosmetric*}传递给 IP 组播服务控制系统.实验中,由于 IGMP 还不能传递服务质量参数,而且主流操作系统的 Socket 接口都不支持服务质量参数,所以我们的 *qosmetric* 暂时由本地路由器按物理接口指定.

IP 组播服务控制系统收集组的分布和服务质量参数后,产生组播聚集策略和服务质量控制策略.服务策略形如:

$$RRID_1\{(ERID_{11},qosmetric_{11}),(ERID_{12},qosmetric_{12}),\ldots,(ERID_{1n},qosmetric_{1n})\};$$
$$RRID_2\{(ERID_{21},qosmetric_{21}),(ERID_{22},qosmetric_{22}),\ldots,(ERID_{2n},qosmetric_{2n})\};$$
$$RRID_n\{(ERID_{n1},qosmetric_{n1}),(ERID_{n2},qosmetric_{n2}),\ldots,(ERID_{nn},qosmetric_{nn})\}.$$

其中:*RRID* 为汇聚路由器的标识(环回地址);*ERID* 为出口路由器的标识,标识了 RSVP P2MP 消息中一条 Sub-LSP 的出口路由器;*qosmetric* 是 Sub-LSP 需要的服务质量保证策略.实验中,我们使用区分服务 DSCP (differentiated services code point)作为服务质量参数.

RSVP-TE 处理模块基于服务策略构造 RSVP-Path 消息,启动点到多点的标签分发过程[9].在没有服务质量需求的情况下,Path 消息的传递只需依赖路由表;在有服务质量需求和流量工程时,需要依赖路由模块提供路由.

RSVP-TE 由网络拓扑驱动,跟踪路由表的变化.采用的是端到端的消息传递、逐跳处理的机制,有序分发标签,并融合了服务质量控制过程[32].其 P2MP 扩展增加了点到多点 LSP 的标签分发过程.而且,为了更快地响应网络拓扑的变化,RSVP-TE 提供快速重路由机制.我们的 IP 组播服务控制系统中服务质量保证的实现主要依赖于 RSVP-TE.

我们以 RSVP-TE P2MP 协议为基础积极性了小扩展,便于 LSP 在汇聚路由器上汇聚.在维持 RSVP-TE P2MP 基本消息机制[9]不变的前提下,我们为 RSVP 协议增加了一类对象,以满足汇聚组播的需要.汇聚组播中,信令过程分为两段:从源到汇聚路由器的 P2MP LSP 和从汇聚路由器到宿的 P2MP LSP 或 P2P LSP.协议的扩展是为了支持这两段路径在汇聚路由器上连接,形成完整的 P2MP LSP.

为此,对于从源到汇聚点的 P2MP LSP,其 PATH 消息中 S2L_SUB_LSP 对象中的 Sub-LSP 目的地址域(sub-LSP destination address)填入的是多个汇聚路由器的地址.PATH 消息新增了一类对象:出口 LER 集.形如

$$RRID\{(ERID,qosmetric),(ERID,qosmetric),\ldots,(ERID,qosmetric)\},$$

其中,*RRID* 为汇聚路由器的标识(路由器环回地址),*ERID* 为出口路由器的标识.

出口 LER 集从属于某个汇聚路由器,由 IP 组播服务控制系统通过汇聚路由器选择算法,把所有参与组播边界路由器划分到一个或多个汇聚路由器的附属集合中.汇聚路由器接收到包含该类对象的 PATH 消息,判断本地是否已经存在于集合中 LER 的 LSP,且相应 LSP 的服务质量参数是否需要修改.若不存在,则新建一条 LSP.

服务控制系统包括网络拓扑测量模块和组播服务模块.实验中,为简单起见,拓扑测量通过扩展 OSPF(open shortest path first)路由协议实现,增加了拓扑输出模块.当网络不使用基于链路状态的路由协议时,我们设想使用基于探测机制的路由模块获取网络拓扑信息.组播服务模块收集组的分布{*GroupID*,*LERID*},为 *LERID* 选择汇聚路由器.

3 性能评价

3.1 评价标准

为了比较基于 RSVP-TE P2MP 协议的非聚集组播、AQoSM[7,8]和本文提出的汇聚组播 RMM 的性能,我们定义了以下 3 类度量.

定义 1(标签使用量和标签使用率). 标签使用量定义为 IP 组播服务占用标签的数量,反映了组播方案使用标签资源的效率.标签使用量等于 LSP 的数量 N_{path}.因为标签使用量与网络拓扑、存在的组播组数量 N_g 以及参与组播的 LER 数量 N_e 有关,因此,为了更一般性地反映标签资源的使用效率,我们定义了相对量标签使用率 $Ratio_{label}$.

$$Ratio_{label}=\frac{N_{path}}{N_g\times N_e}.$$

定义 2(带宽损失率). 在 AQoSM 中存在着带宽浪费,原因是多个组在边界路由器上分布不可能完全一致,因此,分发树覆盖的是它们的并集.这样,有些组的流量就被传递到没有组成员存在的边界路由器上.这些流量虽然被丢弃,但是已经占用了部分链路的带宽和路由器的处理时间,这里定义了带宽损失率 $Ratio_{band}$.

$$Ratio_{band}=\frac{B_{loss}}{B_{total}},$$

其中,B_{loss} 是各段链路上带宽浪费的总和,B_{total} 为各链路上占用的带宽的总和.

定义 3(综合资源使用率). 综合资源反映了标签和带宽的两种资源的整体使用情况.

令 $Ratio_{total}=\alpha\cdot Ratio_{label}+\beta\cdot Ratio_{band}$,其中,$0\le\alpha\le1$,$0\le\beta\le1$,$\alpha+\beta=1$.$\alpha/\beta$反映了两种资源的相对稀缺程度,在本文的实验中,我们取$\alpha=\beta=0.5$,表示 2 者同等重要.

3.2 实验结果

我们在实验系统上进行了功能测试和小规模的组播通信实验,对于大网络大量组播组存在的情况则通过模拟来验证.模拟中,网络拓扑采用的 nem 软件[33]随机生成,并随机生成了组播组在边界路由器上的分布.

图 7 刻画的是当参与组播的边界节点数量为 100 时,非聚集组播 Native、基于树的组播聚集 AQoSM、本文提出的基于汇聚的组播实现 RMM 三者在标签资源使用率上的对比.从图中可以看出,RMM 和 AQoSM 显著优于非聚集组播 Native.虽然 AQoSM 要优于 RMM 方法,但是二者差距不大.

图 8 刻画的是当边界节点数量为 100 时,三者在带宽资源使用率上的对比.在设计上,非聚集组播 Native 和 RMM 都没有带宽损失.从图中可以看出,AQoSM 随着组的数量增加,带宽损失率有显著增长.

图 9 刻画的是当边界节点数量为 100 时,在综合资源使用率上 Native,AQoSM 和 RMM 之间的比较.从图中可以看出,RMM 在综合代价上显著优于非聚集组播和基于树聚集的 AQoSM.

图 10 刻画的是当边界节点数量和组播组数量发生变化时,RMM 组播的综合资源使用率.可以看出,随着组播组数量的增加,曲面趋于平缓.其物理含义是,当组播组数量远大于边界节点数量时,几乎所有的边界节点都参与到组播中来,汇聚节点到所有边界节点都建立了 LSP,因此 LSP 的数量趋于恒定.即使组的数量继续增加,都可以共享已有的 LSP,体现了组播聚集的优势.

图 11 刻画的是当边界节点数量和组播组数量发生变化时,在综合资源使用率上非聚集组播 Native,AQoSM 和本文提出的 RMM 之间的比较.从图中可以看出,即使网络的规模不同,在总的资源消耗上,RMM 均比非聚集组播 Native,AqoSM 要更加节省,而且其优势在网络规模大(边界节点数量较大)、存在的组较多时更为显著.

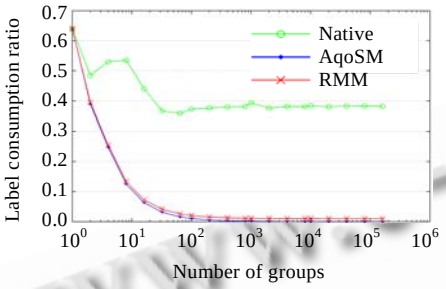

Fig.7 Consumption of label resources with multicast on 100 LERs

图 7 边界节点数量为 100 时的标签资源使用率

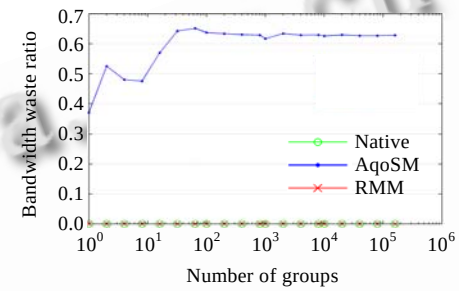

Fig.8 Waste of band resources with multicast on 100 LERs

图 8 边界节点数量为 100 时的带宽损失率

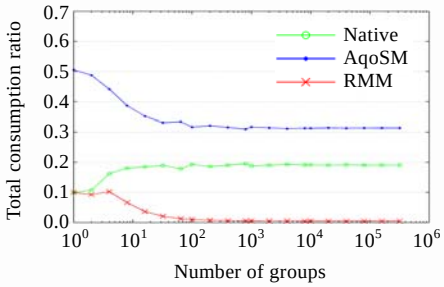

Fig.9 Total comsumption of network resources with multicast on 100 LERs

图 9 边界节点数量为 100 时的综合资源使用率

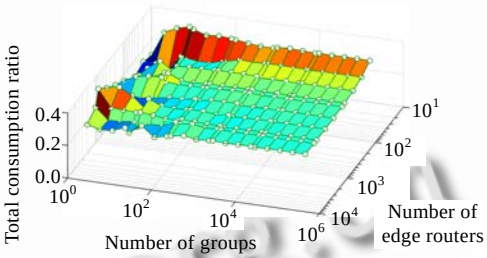

Fig.10 Total comsumption of network resources in RMM

图 10 RMM 的综合资源使用率

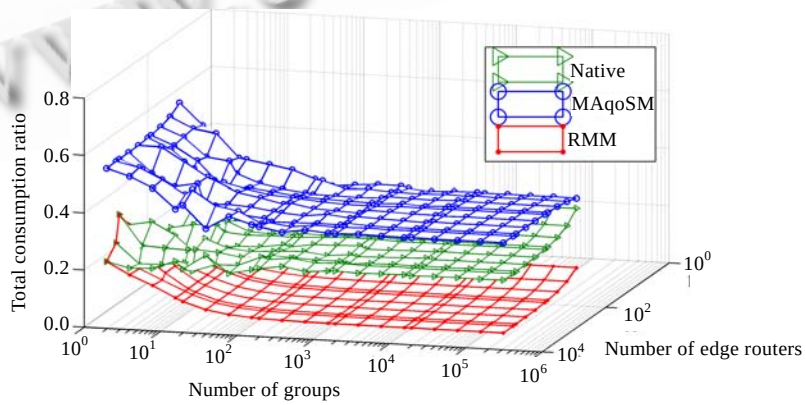

Fig.11 Comparison of total comsumption of network resources

图 11 综合资源使用率对比

4 结束语

本文中,我们针对 MPLS 网络承载 IP 组播时遇到的标签资源消耗、带宽浪费等问题,设计和实现了基于汇聚方法的新型的 MPLS 网络服务质量组播的体系结构.提出在现有的路由控制平面上叠加一层 IP 组播服务控制平面的方法,实现 2 层控制平面的网络架构;并通过扩展 RSVP-TE P2MP 协议,在新的体系结构中融合了服务质量控制能力.本文还探讨了汇聚组播中的汇聚路由器的选择算法.通过实验环境下的模拟和测试,汇聚组播的体系结构提供无带宽浪费的组播聚集能力,有效地减少了 MPLS 标签资源的消耗,并且能够适应组的动态性和网络拓扑的动态性.

下一步的研究目标主要包括扩展 RSVP 机制建立双向 LSP,提高 M2MP 时 RSVP 协议的效率,扩展 RSVP 机制建立层次性 LSP.同时,设计基于本文体系结构的、更高效的约束路由模块也是我们努力的方向.

References:

[1] Deering S, Cheriton D. Host groups: A multicast extension for datagram Internetworks. In: Proc. of the 9th Data Communications Symp. ACM/IEEE, 1985. 172−179. http://portal.acm.org/citation.cfm?id=892355

[2] Deering S, David C. Multicast routing in datagram Internetworks and extended LANs. ACM Trans. on Computer Systems, 1990, 8(2):85−110. [doi: 10.1145/78952.78953]

[3] Quinn B, Almeroth K. IP multicast applications: Challenges and solutions. IETF RFC 3170, 2001.

[4] Ratnasamy S, Ermolinskiy A, Shenker S. Revisiting IP multicast. In: Proc. of the ACM SIGCOMM 2006 Conf. ACM, 2006. 15−26. http://www.sigcomm.org/sigcomm2006/

[5] Rosen E, Viswanathan A, Callon R. Multiprotocol label switching architecture. IETF RFC 3031, 2001.

[6] Armitage G. MPLS: The magic behind the myths. IEEE Communications Magazine, 2000,38(1):124−131. [doi: 10.1109/35.815462]

[7] Fei A, Cui JH, Gerla M, Faloutsos M. Aggregated multicast: An approach to reduce multicast state. In: Proc. of the 3rd Int'l COST264 Workshop (NGC 2001) UCL. London: Springer-Verlag, 2001. 172−188.

[8] Cui JH, Lao L, Faloutsos M, Gerla M. AQoSM: Scalable QoS multicast provisioning in Diff-Serv networks. Computer Networks, 2006, 50(1):80−105. [doi: 10.1016/j.comnet.2005.03.003]

[9] Aggarwal R, Papadimitriou D, Yasukawa S. Extensions to resource reservation protocol—Traffic engineering (RSVP-TE) for point-to-multipoint TE label switched paths (LSPs). IETF RFC 4875, 2007.

[10] Wang B, Hou JC. Multicast routing and its QoS extension: Problems, algorithms, and protocols. IEEE Network, 2000,14(1):22−36. [doi: 10.1109/65.819168]

[11] Diot C, Levine BN, Lyles B, Kassem H, Balensiefen D. Deployment issues for IP multicast service and architecture. IEEE Network Magazine, Special Issue on Multicasting, 2000,14(1):78−88. [doi: 10.1109/65.819174]

[12] Chu YH, Rao S, Zhang H. A case for end system multicast. In: Proc. of the SIGMETRICS 2000. ACM, 2000. 1−12. http://portal.acm.org/citation.cfm?id=345063

[13] Gupta A, Kumar A, Rastogi R. Exploring the trade-off between label size and stack depth in MPLS routing. In: Proc. of the IEEE INFOCOM 2003. IEEE, 2003. 544−554. http://www.ieee-infocom.org/2003/

[14] Awduche D, Berger L, Gan D, Li T, Srinivasan V, Swallow G. RSVP-TE: Extensions to RSVP for LSP tunnels. IETF RFC 3209, 2001.

[15] Mieghem P, Hooghiemstra G, Hofstad R. On the efficiency of multicast. IEEE/ACM Trans. on Networking, 2001,9(5):719−732. [doi: 10.1109/90.974526]

[16] Ogashiwa N, Uda S, Uo Y, Shinoda Y. Implementation and evaluation of MPLS multicast mechanism considering deployment in established MPLS networks. Electronics and Communications in Japan, Part 3, 2007,90(11):40−49. [doi: 10.1002/ecjc.20354]

[17] Apostolopoulos G, Ciurea I. Reducing the forwarding state requirements of point-to-multipoint trees using MPLS multicast. In: Proc. of the ISCC 2005. IEEE, 2005. 713−718. http://www.ieee-iscc.org/2005/

[18] Solano F, Fabregat R, Donoso Y, Marzo JL. Asymmetric tunnels in P2MP LSPs as a label space reduction method. In: Proc. of the IEEE ICC 2005. IEEE, 2005. 43−47. http://www.ieee-icc.org/2005/

[19] Solano F, Fabregat R, Marzo JL. Full label space reduction in MPLS networks: Asymmetric merged tunneling. IEEE Communications Letters, 2005,9(11):1021−1023. [doi: 10.1109/LCOMM.2005.11016]

[20] Gerstel O, Cidon I, Zaks S. The layout of virtual paths in ATM networks. IEEE/ACM Trans. on Networking, 1996,4(6):873−884. [doi: 10.1109/90.556344]

[21] Gunther M, Braun T. Evaluation of bandwidth broker signaling. In: Proc. of the Int'l Conf. on Network Protocols ICNP'99. Switzerland: IEEE, 1999. 145−152. http://www.ieee-icnp.org/1999/

[22] Trimintzios P, Andrikopoulos I, Pavlou G, Flegkas P. A management and control architecture for providing IP differentiated services in MPLS-based networks. IEEE Communications Magazine, 2001,39(5):80−88. [doi: 10.1109/35.920861]

[23] Aukia P, Kodialam M, Koppol PW, Lakshman TV. RATES: A server for MPLS traffic engineering. IEEE Network Magazine, 2000,14(2): 34−41. [doi: 10.1109/65.826370]

[24] Stoica I, Adkins D, Zhuang S, Shenker S, Surana S. Internet indirection infrastructure. In: Proc. of the SIGCOMM 2002. Pittsburgh: ACM, 2002. 73−86. http://www.sigcomm.org/sigcomm2002/

[25] Katz D, Kompella K, Yeung D. Traffic engineering (TE) extensions to OSPF version 2. IETF RFC 3630, 2003.

[26] Apostolopoulos G, Williams D, Kama S, Guerin R, Orda A, Przygienda T. QoS routing mechanisms and OSPF extensions. IETF RFC 2676, 1999.

[27] Andersson L, Doolan P, Feldman N, Fredette A, Thomas B. Label distribution protocol specification. IETF RFC 3036, 2001.

[28] Moy J. Multicast routing extensions to OSPF. IETF RFC 1584, 1994.

[29] Cain B, Deering S, Kouvelas I, Fenner B, Thyagarajan A. Internet group management protocol, version 3. IETF RFC 3376, 2002.

[30] Braden R, Zhang L, Berson S, Herzog S, Jamin S. Resource ReSerVation protocol (RSVP), version 1: Functional specification. IETF RFC 2205, 1997.

[31] Leu JR. MPLS for Linux. 2007. http://sourceforge.net/projects/mpls-linux

[32] Andersson L, Swallow G. The multiprotocol label switching (MPLS) working group decision on MPLS signaling protocols. IETF RFC 3468, 2003.

[33] Magoni D. nem: A software for network topology analysis and modeling. In: Proc. of the 10th IEEE Int'l Symp. on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems (MASCOTS 2002). Mascots: IEEE, 2002. 364−371. http://www.computer.org/portal/web/csdl/abs/proceedings/mascots/2002/1840/00/18400364abs.htm

江勇(1975－),男,重庆人,博士,副教授,主要研究领域为计算机网络体系结构,服务质量,组播.

胡松华(1980－),男,助理工程师,主要研究领域为计算机网络体系结构,MPLS,组播.