文章编号:1001 - 9081(2009)09 - 2428 - 04

基于组播网关的可靠组播体系结构

姜腊林[1,2],徐蔚鸿[1],于 枫[2],樊俊青[2,3]
(1. 长沙理工大学 计算机与通信工程学院,长沙 410004; 2. 东南大学 计算机科学与工程学院,南京 210096;
3. 中国地质大学 计算机学院,武汉 430074)
(lnlnjiang@163.com)

摘 要:保证组播通信的可靠性是许多 Internet 上的组播应用的前提。针对 IP 组播在 Internet 中难以规模化部署的现状,提出了一种使用组播网关将 IP 组播岛与应用层组播(ALM)区域连接起来的可靠组播(RM)体系结构,对组标识、组播网关、组管理、差错控制和拥塞控制等关键问题给出了解决方案,并设计了组播网关竞争算法。该结构能够屏蔽底层组播技术差异,从而支持 Internet 上统一化的可靠组播服务部署。

关键词:可靠组播;IP 组播;应用层组播;组播网关; 竞争算法
中图分类号: TP393 **文献标志码**:A

Reliable multicast architecture based on multicast gateway

JIANG La-lin[1,2], XU Wei-hong[1], YU Feng[2], FAN Jun-qing[2,3]
(1. *School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha Hunan* 410004, *China*;
2. *School of Computer Science and Engineering, Southeast University, Nanjing Jiangsu* 210096, *China*;
3. *School of Computer, China University of Geosciences, Wuhan Hubei* 430074, *China*)

Abstract: Guaranteeing the reliability of multicast communication is the premise of many applications based on multicast in Internet. Since it was difficult to deploy IP multicast on a large scale in Internet, the authors presented a Reliable Multicast (RM) architecture, which utilized multicast gateways to interconnect IP multicast islands and Application Layer Multicast (ALM) domain. Solutions to key points in the architecture, such as group identification, multicast gateway, group management, error control, and congestion control were also proposed. At the same time, a competitive algorithm for multicast gateway was designed. This architecture could shield the difference existing in underlying multicast techniques and thereby support uniform deployment of reliable multicast services in Internet.

Key words: Reliable Multicast (RM); IP multicast; Application Layer Multicast (ALM); Multicast Gateway (MG); competitive algorithm

0 引言

随着视频会议、分布式交互仿真、交互式协同和数据分发等网络组播应用的发展,在 Internet 上进行大规模可靠组播(Reliable Multicast, RM)的研究一直是工业界和学术界热点问题之一。

1988 年,Deering 提出的 IP 组播[1]有效降低了网络带宽需求和服务器负载,得到了广泛应用。然而,IP 组播提供的是尽力而为的服务,为了保证组播数据的可靠传输,一系列运行在 IP 组播上的可靠组播传输协议被提出来。尽管 IP 组播作为最有效的数据群发方法研究已历经二十余年,但由于其在组状态维护、可扩展性、可靠性、拥塞控制、安全以及商用化等方面遇到的困难推迟了 IP 组播的广泛部署[2]111,至今仍未成为一种广泛使用的开放因特网服务。因此研究者们开始研究不依赖于 IP 组播的应用层组播(Application Layer Multicast, ALM)[3],并在此基础上实现组播的可靠性。ALM 不依赖 IP 组播,不需要改变网络底层结构,具有易于部署、灵活和易于实现可靠性的优点,但传输效率不如 IP 组播。

目前,IP 组播仅限于局域网或单一管理控制域内,这种局部范围的 IP 组播网络被称为 IP 组播岛[4]1366 - 1367。如何利用 IP 组播的高效性,如何将这些 IP 组播岛连接起来提供大范围的可靠组播应用,如何将 IP 组播与 ALM 结合起来,是值得研究的问题。研究者们提出了一些将 IP 组播岛连接起来的方案。在 MBone[5]中,利用 IP 单播隧道连接这些 IP 组播岛,这种方法需要对隧道两端的多播路由器进行手工配置,使得 MBone 的建立和维护费用太高,因而限制了其应用。Yoid[6]是一系列协议的集合,允许基于主机的内容通过单播隧道或 IP 组播(如果可用)分发。在可靠性方面,Yoid 可以选择 TCP 协议作为覆盖网络链路协议来保证逐跳间可靠性,但不能保证端到端的可靠性,且只能连接小范围 IP 组播岛(一跳或两跳)。Host Multicast[4]1375 在应用层组建一颗共享树连接 IP 组播岛中的一个指定成员(Designated Member, DM)以及其他非 IP 岛中的组成员,利用 UDP 隧道技术连接 IP 组播岛,使其能兼容 ALM 与 IP 组播,但没有提供可靠组播服务,也没有给出 DM 选定算法。

针对 Internet 中存在 IP 组播与应用层组播两种技术,以及存在 IP 组播岛的现状,本文提出了一种基于组播网关的可靠组播体系结构(Reliable Multicast Architecture Based on

收稿日期:2009 - 03 - 23;**修回日期**:2009 - 05 - 18。
基金项目:教育部重点项目(208098);湖南省教育厅资助科研项目(06C111);湖南省教育厅重点项目(07A056)。
作者简介:姜腊林(1964 -),女,湖南岳阳人,副教授,硕士,主要研究方向:可靠组播、网络体系结构; 徐蔚鸿(1963 -),男,湖南湘潭人,教授,博士,主要研究方向:人工智能在网络中的应用、模式识别; 于枫(1974 -),女,山东济南人,讲师,博士研究生,主要研究方向:计算机网络、Petri 网理论及应用; 樊俊青(1973 -),男,讲师,湖北汉川人,硕士,主要研究方向:并行与分布式计算、无线传感器网络、地学信息。

Multicast Gateway,RMABMG)。

1　基于组播网关的可靠组播体系结构

RMABMG 的基本思想是:在局部、小规模并且支持 IP 组播的网络中使用 IP 组播,在不支持 IP 组播的环境中使用 ALM;通过引入 MG,实现 ALM 与 IP 组播岛之间的互联,以解决 ALM 与 IP 组播的兼容性,如图 1 所示。每个 IP 组播岛中选定一个成员作为 MG,MG 的主要功能有两个:一是完成 ALM 与 IP 组播之间的相互转换;二是同时作为 IP 组播岛的成员和 ALM 的成员参与组播数据分发树的构造和组播路由。

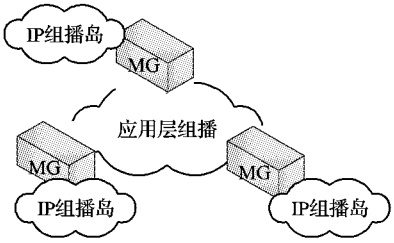

图 1　ALM 与 IP 组播通过 MG 互联

RMABMG 协议栈如图 2 所示,以粗粒度分层可以分四层,即组播网络层(图 2 中虚线框)、组播接口层、可靠组播服务层和可靠组播应用层。组播网络层包括底层的 TCP/IP 网络以及覆盖网络,主要提供组成员加入/退出、组播树的构建/维护以及组播数据的转发等基本组播功能。在 IP 组播岛环境中,由 IP 组播提供上述功能,而在 ALM 环境中,则由 ALM 实现上述功能。ALM 可以根据组播应用程序的特点以及所使用的网络环境而使用一些成熟的 ALM 协议,例如,在大规模单源组播应用环境中,可以使用 Zigzag[7] 协议,而在小规模多源组播应用环境中可以使用 ALMI[8] 协议。组播接口层屏蔽 IP 组播环境与 ALM 环境的差异,向 RM 服务层提供统一的接口。组播接口层根据所处的组播环境,将上层组播服务请求转化为相应组播环境中的功能调用。RM 服务层通过组播接口层提供的统一接口,利用底层 IP 组播或者覆盖网络层提供的组播数据路由转发功能,提供差错控制与拥塞控制等 RM 功能,并为组播应用程序提供接口;应用系统调用接口创建/加入/退出组播组,实现组播数据的可靠传输。RM 应用层即 RM 应用程序。

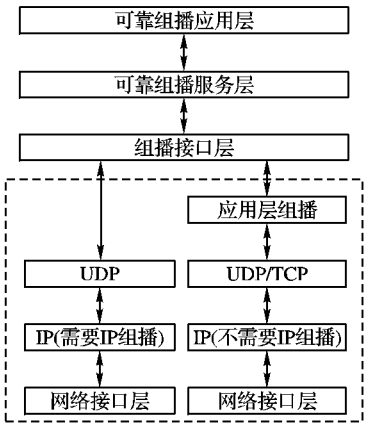

图 2　RMABMG 协议栈

MG 同时运行两套协议栈。当 MG 收到 ALM 覆盖网中的组播数据后,MG 进行 ALM 协议栈到 IP 组播协议栈的转换,然后在 IP 组播岛进行 IP 组播,将数据组播到 IP 岛中的各成员。反之,当 MG 收到 IP 组播岛中某个成员发送的数据后,进行 IP 组播协议栈到 ALM 协议栈的转换,将数据沿着 ALM 分发树送达各组播成员(包括其他的 MG)。

2　组标识与网关竞争算法

由于存在两种不同的组播环境,RMABMG 除了要解决传统可靠组播协议要解决的问题,如差错检测、确认方式、重传机制、流量控制、拥塞控制、可扩展性与网络的异构性,还要解决一些与本结构相关的特定问题,如组标识、组播网关的确定。

2.1　组标识

ALM 与 IP 组播使用不同的组标识方法,如何完成 ALM 与 IP 组标识间的转换是 RMABMG 要解决的关键问题之一。在 RMABMG 中,主机的 RM 服务层使用全局唯一的应用层组标识,而在 IP 组播岛中,IP 组播层使用 IP 组地址。这样,在 IP 组播岛中,就需要一种 ALM 组标识到 IP 组地址的转换,有两种方法完成这种转换。一种是使用会话目录服务[9],当某台主机利用全局组标识加入对应 IP 组时,首先检查会话目录,如果会话目录中存在与该全局组标识相关联的 IP 组地址,则返回该 IP 组地址,主机加入该 IP 组地址标识的组中;否则,分配新的 IP 组地址,并在会话目录中将新的 IP 组地址与全局组标识相关联。另一种方法是使用 Hash 函数,将全局组标识映射成 IP 组标识。

2.2　MG 竞争算法

如何将 IP 组播岛中的一个成员指定为 MG 是 RMABMG 要解决的另一个关键问题。可以使用静态的方法手工配置 MG,也可以运行某种算法动态确定 MG。由于组成员具有动态加入与退出组播组的特点,使用静态的方法不太适合组播成员动态变化的特点,因此,在 RMABMG 中,设计了一种竞争算法动态地确定 MG。当 IP 组播岛中某个成员加入组时,运行网关竞争程序,网关竞争算法参见伪代码表示的算法 1。如果竞争成为 MG,则运行网关应用程序,组建或参与应用层转发树的构造;否则,作为普通成员加入。算法中,使用哈希算法完成应用层组地址到 IP 组地址的转换,使用定时器和随机等待时间防止产生多个 MG,在有多个成员竞争时,可以使具有最高优先级的成员成为 MG。优先级可以代表主机的处理能力、带宽、存储容量等指标。考虑到协议的健壮性,MG 应该定期向 IP 岛中组内成员组播自己存在的消息。如果组成员在预先给定的时间间隔内收不到该消息,则可以认为 MG 失效,因此组播成员可以竞争生成一个新的 MG。由于 IP 组播岛中成员都有可能竞争成为 MG,因此必须支持两个协议栈。但 IP 组播岛中成员只有在竞争成为 MG 后才运行网关应用程序,因此,只有 MG 成员才需要运行两套协议栈,而岛中普通成员只需要提供 IP 组播服务即可。

算法 1　网关竞争算法

```
//将全局组标识符影射成 IP 组播地址
IPMulticastAdress = hash( GroupID) ;
//加入 IP 组
Join( IPMulticastAdress) ;
//以优先级 ThisPriority 请求成为网关
Lable: Send( IPMulticastAdress, MGRequest( ThisPriority) ) ;
//设置定时器
Set( MGRequetTimer) ;
//如果定时器时间未到
While( ! timeout( MGRequetTimer) )
//收到其他网关的应答或者其他即将成为网关的消息,则退出竞争
{ if( receive( MGRequestReply( ) ) ‖ receive( ToBeMG( ) ) ) exit( ) ;
  //收到其他成员请求成为网关消息,比较其优先级,优先级比
  //本成员高退出竞争,相同等待一随机时间
  if ( receive( MGRequest( Priority) ) )
    if( Priority > ThisPriorty) exit( ) ;
```

```
    else if( Priority == ThisPriorty)
    { set( MGRequetTimer); t = wait( randomize( MGRequetTimer));
      //等待期间收到其他网关或希望成为网关的消息则退出竞争
      while (! timeout( t))
        if( receive( MGRequestReply()) || receive( ToBeMG()))
        exit();
      //随机等待时间到转标号处执行
      goto lable;
    }
}
//定时器超时,发送希望成为网关消息,设置另一定时器
Send( IPMulticastAdress, ToBeMG()); set( ToBeMGTimer);
while (! timeout( ToBeMGTimer))
  if( receive( MGRequestReply()) || receive( ToBeMG()))
  exit();
//定时器超时,成为网关,运行网关程序
MG();
```

3 RM服务层协议要解决的关键问题

RM服务层协议主要解决差错和拥塞控制等可靠性问题。不同的组播应用需求需要不同的差错控制和拥塞控制机制。本章针对单源、延迟不敏感型RM应用,提出相应的RM服务层关键问题的解决方案。

3.1 组成员控制拓扑结构

为了保证可靠性,发送方需要知道组成员是否正确收到发送的数据,并及时重传修复,因此,组成员之间通常需要交换一些控制信息。为了方便控制,通常将组成员按某种逻辑图的形式组织起来,这种逻辑图被称为控制拓扑。控制拓扑主要有环状[10]、树状[11]409和超立方体结构[12]。有些可靠多播协议没有提供控制拓扑来分发控制信息,因此,这些协议必须限制控制信息的量,以满足可扩展性的需要。例如,可扩展可靠组播协议采用否定应答抑制机制[13]787。

综合考虑协议的可扩展性和对异构网络的适应性,认为树状结构的组成员控制结构比较适合单源组播应用需求。组成员控制拓扑结构如图3所示,该结构将组成员组成一颗以数据源为根的树,其中圆圈表示数据源,方框表示接收成员,虚线表示成员间是一种逻辑结构,箭头表示由父节点指向子节点。除源没有父节点,其他各组成员只有一个父节点。每个组成员只需知道其父节点和孩子节点。目前,有许多关于树拓扑结构的构造算法[4]1369-1371可以借鉴。

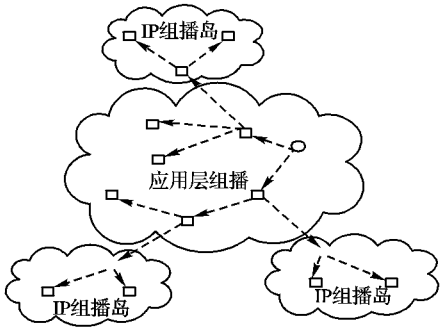

图3 组成员控制拓扑结构

3.2 状态控制、重传机制和差错控制机制

为了保证传输可靠性,对接收方未成功接收的数据进行重传是必要的。根据由哪一方进行传输状态跟踪和控制数据重传,可靠数据传输一般分为基于发送方的可靠传输和基于接收方的可靠传输[13]785。重传机制包括谁负责重传、什么时候重传以及重传方式。可以由发送源、路由器或从接收节点选出代表来负责重传[11]407。大规模通信的可靠播协议一般采用后者。在重传时间上,一般的协议都是在接收到重传请求后立即或等待一个随机时间后发送重传数据包[13]787,但在组播文件传输协议中,一般等文件传输完后再重传出错数据包[14]。重传方式分为单播方式、全局组播方式[13]798和局部组播方式[11]410。可靠组播协议主要处理数据包的丢失和传输乱序问题,其差错控制方法主要有自动重发请求和前向纠错[15]。

考虑到前向纠错主要用于延迟敏感型应用,协议采用自动重发请求差错控制机制。同时,协议采用基于发送方的可靠传输以及局部重传机制。发送节点通过底层数据分发树向组内所有成员组播数据,子节点收到数据后,在规定时间内向其父节点发送应答。父节点检测到数据丢失,向子节点重传丢失的数据。考虑到子节点数量有限,重传采用单播方式。这样,发送节点只需掌握其直接相连的孩子节点的状态,只负责向孩子节点重传差错数据,而其他节点的状态由各自父节点掌握,各父节点完成对子节点的局部重传。由于发送方不需要掌握所有组成员情况以及使用局部重传,有效抑制了反馈爆炸,缓解了发送方瓶颈,使得协议具有较好的可扩展性。重传采用单播方式,避免了全局组播方式浪费网络带宽以及局部组播方式实现复杂的缺点。

3.3 拥塞控制

组播传输的参与者比较多,可靠组播发送者与接收者常需交换大量控制信息,很容易引起网络拥塞,因此拥塞控制对可靠组播协议来说尤其显得重要。组播拥塞控制的主要方法可以分为基于窗口、基于速率、基于层的以及基于局部恢复的拥塞控制方法[2]116。拥塞控制主要解决可扩展性、异构性以及公平性[16]三类基本问题。

本协议使用局部重传机制能够有效减少可靠性相关的控制信息流量,有利于防止网络拥塞,并使协议具有较好的扩展性,因此,在拥塞控制方面着重考虑TCP友好性[17]。本协议采用发送端基于速率的拥塞控制方法。发送端根据收集到的反映系统拥塞状况的参数,根据Padhye等人提出的TCP连接平均分享带宽的解析模型[18]确定一个最大发送速率,并保证其瞬时发送速率不超过预先给定的速率。

4 结语

IP组播是一种高效的组播方式,但其对路由器的依赖、计费困难、应用程序不多等原因限制了它在Internet上的广泛部署。目前,一些网络厂商(如思科,华为)所开发的高端路由器都支持IP组播,一些研究者将主动网络技术结合到IP组播中,相信IP组播在下一代Internet中一定会有广阔的应用前景。ALM虽然在组播效率方面不如IP组播,但其不依赖路由器、灵活的特点是目前许多ISP部署组播应用的首选,其应用前景也是不可估量的。因此,如何结合二者的优势,使其能提供高效、可靠、安全、可扩展性、有服务质量(Quality of Service, QoS)保障、利于商业化的组播服务,是值得深入研究的问题。本文针对Internet中存在IP组播与ALM两种组播技术以及IP组播岛的现状,提出利用组播网关将IP组播和ALM相结合的RM体系结构,创造性地设计了组播网关竞争算法,为Internet上的RM部署提供了一种新思路。

本文虽然针对单源、延迟不敏感型的应用给出了一个RM协议的相关机制,但还很粗糙。我们下一步的主要工作是根据本文所提出的RMABMG,针对某类组播应用设计一个RM协议原型系统,通过仿真实验分析该系统性能,并在协议可扩展性、网络异构性、QoS方面做进一步的研究工作。

参考文献:

[1] DEERING S. Host extensions for IP multicasting [EB/OL]. [2008 -11 -04]. http://www.ietf.rfc/rfc.1112.txt.

[2] POPESCU A, CONSTANTINESCU D, ERMAN D, *et al.* A survey of reliable multicast communication [C]// Proceedings of the 3rd EURO-NGI Conference on Next Generation Internet Network. Trondheim, Norway: IEEE Computer Society, 2007: 111 -118.

[3] 章淼,徐明伟,吴建平.应用层组播研究综述[J].电子学报, 2004,32(12A): 22 -25.

[4] ZHANG BEI-CHUAN, JAMIN S, ZHANG LI-XIA. Host multicast: a framework for delivering multicast to end users [C]// Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies. New York, USA: IEEE Computer Society, 2002: 1366 -1375.

[5] ERIKSSON H. MBONE: The multicast backbone [J]. Communications of the ACM, 1994, 37(8): 54 -60.

[6] FRANCIS P. Yoid: Extending the multicast Internet architecture [EB/OL]. [2008 -12 -02]. http://www.ciri.org/yoid/docs/ycHtml/htmlRoot.html.

[7] TRAN D A, HUA K A, DO T T. Zigzag: An efficient peer-to-peer scheme for media streaming [C]// Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies. San Francisco, CA, USA: IEEE Computer Society, 2003: 1283 -1292.

[8] PENDAKARIS D, SHI S. ALMI: An application level multicast infrastructure [C]// Proceedings of the 3rd USENIX Symposium on Internet Technologies and Systems. San Francisco, CA, USA: USENIX Association, 2001: 49 -60.

[9] HANDLEY M. Session directories and scalable Internet multicast address allocation [J]. ACM Computer Communication Review, 1998, 28(4): 105 -116.

[10] WHETTEN B, KAPLAN S, MONTGOMERY T. A high performance totally ordered multicast protocol [C]// Proceedings of International Workshop on Theory and Practice in Distributed Systems, LNCS 938. London: Springer-Verlag, 1994: 33 -57.

[11] LIN J C, PAUL S. RMTP: A reliable multicast transport protocol [C]// Proceedings of the 15th Annual Joint Conference of the IEEE Computer and Communications Societies. San Francisco, CA, USA: IEEE Computer Society, 1996: 1414 -1424.

[12] ZHANG ZU-PING, CHEN HAO, CHEN JIAN-ER. A new control topology of information reliable spread [C]// Proceedings of 2003 IEEE International Conference on Robotics, Intelligent Systems and Signal Processing. Washington, DC: IEEE Computer Society, 2003: 789 -793.

[13] FLOYD S, JACOBSON V, LIU C, *et al.* A reliable multicast framework for light-weight sessions and application level framing [J]. IEEE/ACM Transactions on Networking, 1997, 5(6): 784 -803.

[14] BLACKMORE, F A. Multicast file transfer in a military satellite broadcast system [C]// Proceedings of the 21st Century Military Communications Conference Proceedings. Los Angeles, California: IEEE Computer Society, 2000: 1094 -1098.

[15] RIZZO L, VICISANO L. Effective eraseure codes for reliable computer communication protocols [J]. ACM Computer Communications Review, 1997, 27(2): 24 -36.

[16] LI V O K, ZHANG ZAI-CHEN. Internet multicast routing and transport control protocols [J]. Proceedings of the IEEE, 2002, 90(3): 360 -391.

[17] FLOYD S, FALL K. Promoting the use of end-to-end congestion control in the Internet [J]. IEEE/ACM Transactions on Networking, 1999, 7(4): 458 -472.

[18] PADHYE J, FIROIU V, TOWSLEY D F, *et al.* Modeling TCP Reno performance: A simple model and its empirical validation [J]. IEEE/ACM Transactions on Networking, 2000, 8(2): 133 -145.

(上接第 2427 页)

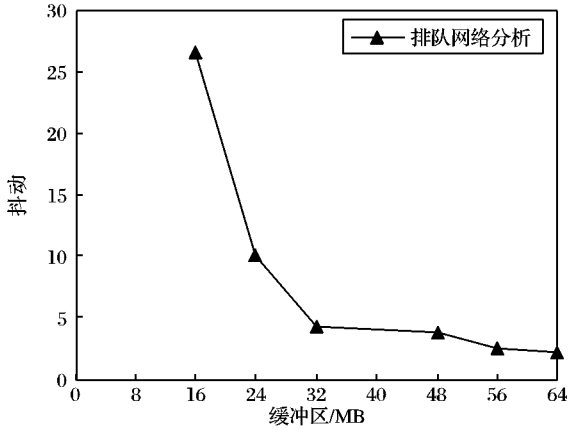

图 7　缓冲区与抖动关系

5　结语

本文针对资源配置优化问题开展研究,基于排队网络理论,提出一种多业务网络的资源分析方法,对网络特征及业务特性进行了定义及描述,建立了资源分析模型,包括多业务网络的排队服务模型、流量分配方法、模型的求解与计算,给出了多种资源与网络服务质量之间的配置关系,通过仿真实验,验证了该方法的正确性和有效性。由于该方法相对于测试等其他方法具有简单、省时、费用低等特点,因此更有利于应用于新建网络设计和已建网络的资源优化中。

参考文献:

[1] 蔡康,李洪,朱英军. IP 宽带业务与运营[M]. 北京: 人民邮电出版社, 2003.

[2] VEGESNA S. IP 服务质量[M]. 信达工作室,译. 北京: 人民邮电出版社, 2001.

[3] ANDREW L L H, HANLY S V, MUKHTAR R G. Active queue management for fair resource allocation in wireless networks [J]. IEEE Transactions on Mobile Computing, 2008, 7(2): 220 -247.

[4] DOVROLIS C, STILIADIS D, RAMANATHAN P. Proportional differentiated services delay differentiation and packet scheduling [J]. IEEE/ACM Transactions on Networking, 2002, 10(1): 12 -26.

[5] BOUDEC J L, THRAN P. Network calculus a theory of deterministic queuing system for Internet [M]. Heidelberg: Springer-Verlag, 2004.

[6] IMENZ D H, ONDA A. Optimal partition of QoS requirement on unicast paths and multicast trees [J]. IEEE/ACM Transactions on Networking, 2002, 10(1): 102 -114.

[7] BOLCH G, GREINER S. Queuing networks and Markov chains modeling and performance evaluation with computer science applications [M]. 2nd ed. New Jersey: John Wiley & Sons, 2006.

[8] GROSS D, SHORTLE J F, THOMPSON J M, *et al.* Fundamentals of queuing theory [M]. 3rd ed. New Jersey: John Wiley & Sons, 1998.

[9] 盛友招. 排队论及其在计算机通信中的应用[M]. 北京: 北京邮电大学出版社, 1998.

[10] 林元烈. 应用随机过程[M]. 北京: 清华大学出版社, 2002.