

# Clos网络中的组播路由算法

## The Analysis of Multicast Routing Algorithm in Clos Networks

石增增/ SHI Zeng-zeng, 顾华玺/ GU Hua-xi,  
王长山/ WANG Chang-shan  
(西安电子科技大学, 陕西 西安 710071)  
(Xidian University, Xi'an 710071, China)

中图分类号: TN915 文献标识码: A 文章编号: 1009-6868 (2008) 03-0042-05

**摘要:** 对于三级Clos网络, 扇出机制会影响Clos网络的阻塞率、算法的时间复杂度及网络成本, 因此选择好的扇出方式能充分发挥网络的组播能力。根据输出级扇出、中间级扇出、输入级扇出等不同的扇出机制分类, 可将组播算法分为输入级扇出算法(IFMA)、最迟扇出算法(LFMA)、切割扇出算法(SFMA)、中间级优先扇出算法(CMFFMA)。在对4种算法仿真比较的基础上, 文章提出针对不同的业务采用不同的处理方法的路由方案, 对于固定扇出业务可采用CMFFMA算法进行路由, 针对递增业务采用先输出级、再中间级、最后输入级扇出的策略, 可有效地降低阻塞率。

**关键词:** Clos网络; 组播; 路由算法; 扇出

**Abstract:** Fan-out mechanism would affect the blocking probability of the three-stage Clos network, the time complexity of the algorithm and network costs. Good fan-out approach can give full play to the multicast capacity of the network. According to the output-stage fan-out, middle-stage fan-out, and input-stage fan-out mechanisms, multicast algorithms include Input Fan-out Multicast Algorithm (IFMA), Lazy Fan-out Multicast Algorithm (LFMA), Split Fan-out Multicast Algorithm (SFMA), and Central Module First Fan-out Multicast Algorithm (CMFFMA). Comparing the analyses of these four algorithms, this article proposes a routing scheme in which different businesses use different algorithms. Bundled traffic can adopt CMFFMA, and incremental traffic can route from output-stage through middle-stage to input-stage fan-out. Therefore, the blocking probability can be effectively reduced.

**Key words:** Clos network; multicast; routing algorithm; fan out

随着宽带技术的不断发展, 视频点播、远程教学、新闻发布、网络电视等业务成为新一轮运营竞争的焦点, 它们的特点是, 信息由一个服务器向大量的客户端发布。组播技术非常适合这类业务, 并具有如下优点: 服务器不必知道某个客户端是否存在, 它只负责按多播地址将媒体流

发送出去, 即使有成千上万个客户端, 也仅发送一份即可; 客户端如果希望接收某媒体流服务器的数据, 只需加入该媒体流服务器播放数据使用的组播组即可<sup>[1]</sup>。

目前智能光网络的发展要求节点设备的交叉矩阵具有容量高、快速的端口配置和组播支持能力, 组播业

务根据目的节点数的不同, 可以分为单播、组播和广播3种类型<sup>[2]</sup>。单播是指待转发的消息在传送网中要求实现点对点的传输, 广播业务是指在传送网中把待转发的一个消息从源节点转发到传送网的全部输出端口上, 而组播业务是则把消息转发到传送网中的一组输出端口上。从广义上来讲, 单播和广播是组播的一个特例。

根据组播请求的多个目的输出端口的产生时间, 可以把组播分为两类<sup>[3]</sup>: 第一类是固定扇出业务, 所有的目的输出端口是在请求一开始就确定; 第二类是递增业务, 它的目的端口递增, 是不确定的。

### 1 Clos网络的组播业务

为了支持网络中的组播业务, 网络中的核心设备交换设备也应当具有组播功能。Clos网络自提出以来<sup>[4]</sup>由于其低成本、易大规模实现, 在交换设备中得到了广泛的应用。

图1为一个对称的三级Clos网络, 用 $n$ 表示输入输出模块的端口数量,  $N$ 表示总的输入端口数,  $f$ 表示扇出值,  $m$ 表示中间模块的数量,  $r$ 表示输入和输出模块的数量, 则一个三级Clos网络可以表示为 $C(n, m, r)$ 。如果用 $I_p$ 表示输入端口,  $P_p$ 表示输出端口, 那么一个组播请求可表示为 $(I_p: P_1, P_2, \dots, P_k)$ 。对称的三级Clos网络在任意级有扇出功能的组播严格无阻塞的条件为 $m \geq \min\{(n-1)f+n, (N-1)f, N\}$ <sup>[5]</sup>, 而且对于任意一个组播严格无阻塞网络, 需要的开关数最少为 $O(N^2)$ <sup>[6]</sup>, 但是在实际应用中并不需要达到严格无阻塞就可以有很好的性能。

#### 1.1 Clos网络扇出机制

对于三级Clos网络, 不同的扇出机制不但影响Clos网络的阻塞率, 而且影响算法的时间复杂度及网络成本, 因此选择好的扇出方式才能充分发挥网络的组播能力。以下将对Clos网络各级扇出的性能特点进行分析。

(1) 输出级扇出

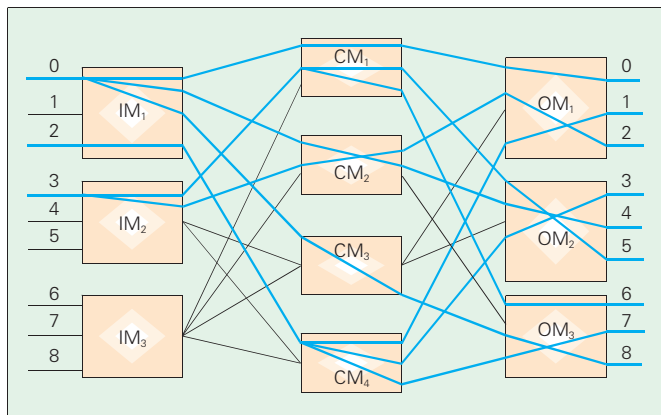


图1  
C(3, 4, 3)三级Clos  
网络示意图

输出级扇出指输出级模块具有扇出功能,如果输出级具有扇出功能,那么对于同一个业务源到一个输出模块中的多个输出端口只需要经过一个中间模块,否则有多少输出端口就需要经过多少个中间模块,在三级Clos网络中路由分配主要就是中间级模块的分配,因此必须降低对中间级模块的需求,而第三级扇出可以降低对中间级模块的要求,所以采用第三级扇出可以有效的降低阻塞率,这样我们就可以将一个组播请求由原来的端口表示( $I_p: \{P_1, P_2, \dots, P_k\}$ )转化成模块表示( $I: \{O_1, O_2, \dots, O_k\}$ ),其中 $I$ 表示输入模块, $O$ 表示输出模块。

### (2) 中间级扇出

中间级扇出指中间级的模块具有扇出功能。假如第一级没有扇出功能,那么所有组播分支只能在一个中间级模块进行扇出,因此只有那些满足所有扇出要求的中间交换单元才可以成功建立连接。所以在组播请求的扇出值很大的情况下,网络的阻塞概率将会急剧上升,但是由于只使用一个中间模块,可以避免外部阻塞的发生。

### (3) 输入级扇出

输入级扇出指输入级模块具有扇出功能,可以从一个输入端口到达不同的中间级模块。如果第三级有扇出的话,那么组播请求要到达几个输出级模块,就需要占用几个中间级模块。对于输入级扇出可以将组播分解成不同的单播请求进行处理,这样可

以利用单播中成熟的算法来进行处理,实现简单,而且可以降低内部阻塞率。但是由于每个组播请求只在第一级扇出,因此需要大量的中间模块,容易出现外部阻塞问题。

## 1.2 Clos网络组播算法介绍

Clos网络中的组播算法性能主要受扇出机制的影响,这样我们就根据扇出策略的不同将组播算法分为以下几种。

输入级扇出算法(IFMA)<sup>[7]</sup>是基于输入级扇出的算法,其主要思想是通过将一个扇出值为 $f$ 的组播请求转化成 $f$ 个单播请求,然后按照单播请求的路由算法进行路由,这样在Clos网络中每个组播请求只在输入级进行扇出,这样可以将组播业务理解为多个相互独立的单播业务,这样就可以利用单播算法中的成熟算法。如图1中的输入端口0到输出端口0、输出端口4和输出端口8的组播业务采用输入级扇出方式,在输入模块 $IM_1$ 中完成所有的扇出,分别经过中间模块 $CM_1$ 、 $CM_2$ 和 $CM_3$ 到不同的目的模块。

最迟扇出算法(LFMA)<sup>[6]</sup>是基于中间级扇出的算法,该算法的思想是只有在必须进行扇出时才进行扇出,即先在输出级扇出再在中间级扇出。因此对于每一个组播请求只使用一个中间模块,如图1中输入端口2中的请求( $2: \{1, 3, 7\}$ ),只使用了一个中间模块 $CM_4$ 。

这两种扇出机制都存在着自身

的局限性,但是又有很强的互补性,因此将两种扇出相结合的思想就应运而出。在三级Clos网络中,内部阻塞的产生主要是由于级间链路的竞争,如果没有第三级扇出,那么每个组播请求在一个输出模块的每个输出端口都要占用一个从中间级到输出级的链路,否则只需要一个链路。同样,如果中间级没有扇出,那么每一个子请求都要占用一条输入模块到中间模块之间的链路,这样就会出现外部阻塞。各种扇出机制各有优缺点,可以结合使用。在在输入级和输出级同时扇出的机制中又可以根据不同的分配方式分为切分扇出算法及先中间级后输入级算法两种。

切割扇出算法(SFMA)<sup>[8]</sup>是把目的输出模块进行分组,分组数 $g$ 为扇出值 $F$ 和切割值 $s$ 的比值向上取整,然后在进行路由时在第一级就进行扇出,即需要在第二级选择 $g$ 个可用的中间交换单元,然后再在第二级和中间级扇出机制一样进行同样的处理。如图1中输入端口3的请求,如果按照切割算法理解的话其扇出值 $F$ 为3,切割值 $s$ 为2,分为两组,一组通过中间模块 $CM_1$ 路由,另一组通过中间模块 $CM_2$ 路由。

最后一种算法是中间级优先扇出算法(CMFFMA)<sup>[9-10]</sup>,利用尽量少的中间模块完成扇出,即首先选择一个可以建立尽量多扇出的中间单元,建立其到输出模块的连接。如果到全部输出模块的连接均建立完成则路由成功,否则将余下的尚未完成的连接继续按照上一步的方法处理,利用其他中间级单元的扇出能力完成扇出。例如在图1中,由于没有一个中间模块能够满足输入端口3的所有扇出请求,因此通过 $CM_1$ 建立其中的两条,然后再通过 $CM_2$ 建立剩余的连接。

通过以上对扇出的分析,我们可以得到采用先中间级后输入级算法的扇出机制是最优的。与切割扇出机制相比,它少了盲目性,多了预先检测性,可以在第一级进行有目的扇

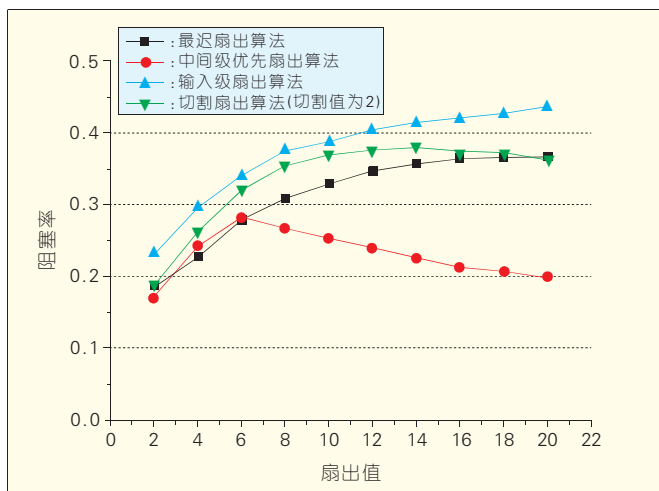


图2  
阻塞率随扇出值变化  
曲线图

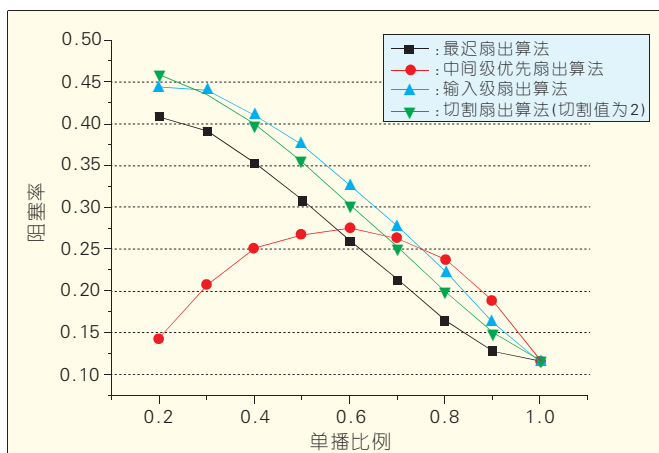


图3  
阻塞率随单播比例变  
化曲线图

出;与最近扇出机制相比,它又有很强的灵活性。

### 1.3 Clos网络组播算法仿真

#### (1) 仿真条件

采用OPNET软件对不同的组播算法进行仿真,仿真中的请求是按照占用-空闲源模式产生,即每个输入端口有占用和空闲两种状态,占用状态表示该输入端口当前存在一个链接,每种状态的持续时间均服从指数分布,如果 $1/\mu$ 表示占用的平均时间,  $1/\lambda$ 表示空闲的持续时间,那么在以输入端的状态判断,网络中的负载

$$\rho_1 = \frac{\frac{1}{\mu}}{\frac{1}{\mu} + \frac{1}{\lambda}} = \frac{\lambda}{\mu + \lambda}, \text{ 如果用 } f \text{ 表示组播}$$

的平均扇出,  $P_{rp}$  表示业务中单播的比率,那么网络中的实际负载  $\rho = (P_{rp} +$

$(1 - P_{rp}) \times f) \times \rho_1$ 。每个组播的扇出值按指数分布产生。

#### (2) 仿真结果

在具有组播业务的Clos网络中网络的阻塞率主要受组播业务的扇出值、组播比例和中间模块数的影响,下面就分别进行仿真分析。

图2是4种不同的算法在C(16, 16, 16)网络规模、0.8负载以及单播比例为0.5时的阻塞率随扇出值变化的曲线图。

从图2中可以看出随着扇出值的增加阻塞率会有所增加,但是当扇出值达到一定值时,阻塞率将趋于稳定,这是因为在负载固定、输出级有扇出的情况下,随着扇出值的增加请求数量会减少。同时由于输出级具有扇出功能,而输出级的模块数固定,所以当扇出值超出一定值后扇出的目的模块数不会有太多的变化,因此在扇出值大于一定范围后,阻塞则趋于稳定。在这几种算法里CMFFMA的阻塞率最低,因为他的扇出顺序是先输出级、再中间级、最后输入级,这样可以最低限度地节约网络中的链路资源,避免阻塞发生。

图3为C(16, 16, 16)的Clos网络在负载为0.8时的阻塞率随单播比例变化的仿真结果。

从图3中可以看出随着单播比例的增加IFMA算法、SFMA和LFMA算法的阻塞率单调下降,而CMFFMA算法的阻塞率随着单播比例的变化成抛物线状,这是因为这两种算法适宜于组播请求的建立,能够最大程度的利用已有的空闲资源,因此在单播比例较低时网络的阻塞率比较低,但是随着单播比例的增加阻塞率会逐渐增加,当到达一定的比例时阻塞率又随着单播比例的增加而下降,直到单播比例为1时,以上几种算法的阻塞率

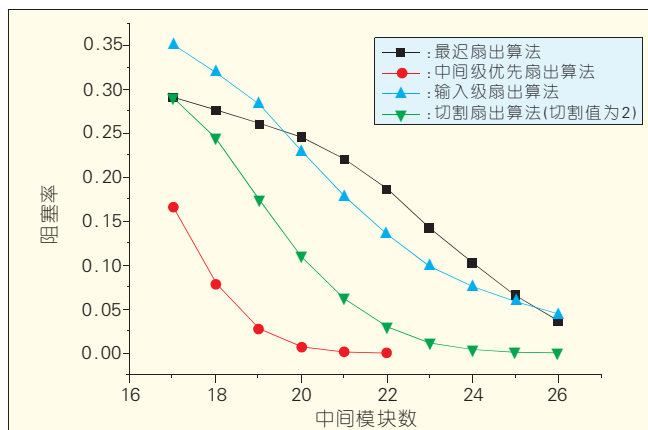


图4  
阻塞率随中间模块数  
的变化曲线图

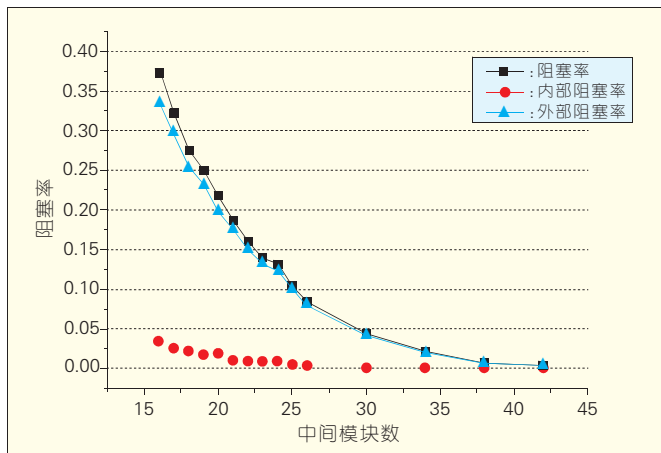


图5  
IFMA算法阻塞率与中间模块数的关系

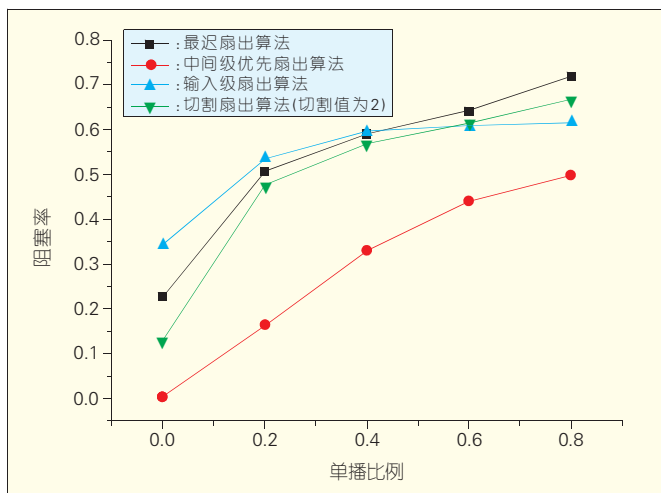


图6  
混合业务中的组播业务阻塞率

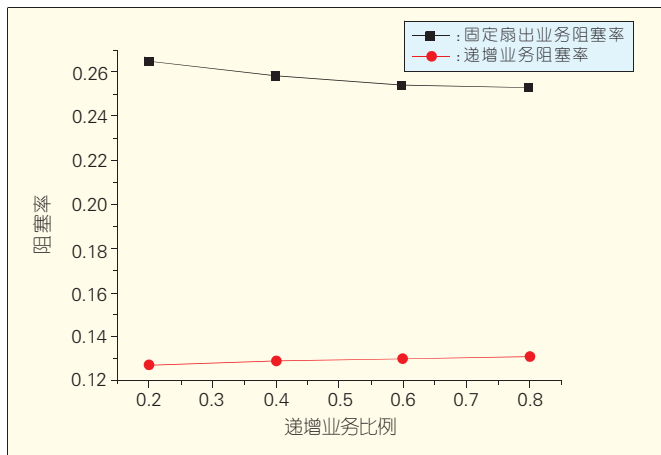


图7  
递增业务的阻塞率

均达到一个固定值。

图4为C(16,16,16)规模的Clos网络在0.8的负载,平均扇出值为8时及单播比例为0.5时各种算法的阻塞率随中间模块数的变化曲线。

从图4中可以看出随着中间模块

数的增加,不同算法的阻塞率下降的速度不同,其中LFMA算法和IFMA算法的下降最缓慢,其他两种算法的下降速度很快;而且在中间模块数远小于严格无阻塞所需要的中间模块数的情况下,Clos网络的阻塞率可以下

降到很低。

从以上分析可知IFMA算法的阻塞率在所有算法中是最高的,这是因为该算法采用输入级扇出,组播业务的扇出均要在输入级实现,这样会造成很高的外部阻塞,而且占用的第一级链路数与第二级链路数相等。

图5为IFMA算法的阻塞率随中间模块数的变化趋势,网络规模为C(16, m, 16),平均扇出值为8,负载为0.8,全组播业务。

图5中可以看出内部阻塞率较小,故网络的整体阻塞率主要由外部阻塞率决定。

上面分析了在单、组播业务混合的情况下网络的整体阻塞率,但是由于单播和组播业务的不同,其阻塞率不尽相同,图6为不同算法随单播比例变化对组播业务阻塞率的影响。从图6中可以看出随着单播比例的增加,组播业务的阻塞率单调递增。其中,CMFFMA算法的阻塞率最低,这是由于其更好的利用了网络中的空闲链路资源;IFMA算法采用输入级扇出,所以单播比例的增加并没有影响其可用的链路资源的减少,因此阻塞率的增长最慢。

## 2 组播实现方案

在对组播业务及常见算法的比较分析的基础上,本文设计出一种路由方案,针对不同的业务采用不同的处理方法。

由于三级均有扇出的CMFFMA算法的阻塞率最低,因此对于固定扇出业务可以采用该算法进行路由。

针对递增业务的特点,同时为了降低对链路资源的占用,采用先输出级、再中间级、最后输入级扇出的策略。由于递增业务是在固定扇出业务的基础上增加的业务,因此首先判断是否可以在固定业务已占用的输出模块内完成扇出,如果路由成功则退出;否则再判断是否可以通过固定业务已经占有的中间级模块完成路由,如果成功则退出;否则采用输入模块



进行扇出,如果成功则退出;否则返回路由失败。

图7为采用本方案后的C(16, 16, 16)规模的Clos网络,在单播比例为0.5、负载为0.8、平均扇出为8的时的阻塞率变化图,其中递增业务比例为递增业务占组播业务的比例。由于递增业务均是以单播的形式处理,而且对于递增业务处理思想与固定组播业务类似,首先从输出模块进行扇出、再中间模块、最后输入级,因此递增业务的阻塞率接近于单播业务的阻塞率,而且随着递增业务量的增加,网络的阻塞率无太大变化。

### 3 结束语

随着单播比例的增加,网络中的组播业务的阻塞率会随之增加。其中,中间级优先扇出算法要求输入级和输出级都要有扇出功能,充分利用了交叉矩阵中的链路资源,因此阻塞率最低。虽然组播严格无阻塞所需要的中间模块数很多,但是在实际的应用中并不需要很多就可以达到很低的阻塞率。而且在相同的条件下,随着中间级模块数量的增加,输入级和输出级同时扇出的算法的阻塞率下降更快。对于递增业务处理时可以按照组播扇出的思想进行处理,这样对整体网络中的阻塞率无明显影响。

下一步的工作是将重排算法引入Clos网络中的组播业务,通过对已建立的业务进行重排来降低阻塞率。

### 4 参考文献

- [1] SUN Shutao, HE Simin, ZHENG Yanfeng, et al. Multicast scheduling in buffered crossbar switches with multiple input queues[C]//Proceedings of 2005 Workshop on High Performance Switching and Routing(HPSR'05), May 12-14, 2005, Hong Kong, China. Piscataway, NJ, USA: IEEE, 2005: 73-77.
- [2] FU Hunglin, HWANG F K. On 3-stage Clos networks with different nonblocking requirements on two types of calls[J]. Journal of Combinatorial Optimization, 2005, 9(3): 263-266.
- [3] HWANG F K, SHENG-CHYANG L. On nonblocking multicast three-stage Clos networks[J]. IEEE/ACM Transactions on Networking, 2000, 8(4): 535-539.
- [4] CLOS C. A study of non-blocking switching network[J]. Bell System Technical Journal, 1953, 32(2): 406-424.
- [5] HWANG F K. A survey of nonblocking multicast three-stage Clos networks[J]. IEEE Communications Magazine, 2003, 41(10): 34-37.
- [6] FRIEDMAN J. A lower bound on strictly non-blocking network[J]. Combinatorica, 1988, 8(2): 185-188.
- [7] PARK Won-Bae, HENRY L. Owenand ellen wine zegura, SONET/SDH multicast routing algorithms in symmetrical three stage networks[C]//Proceedings of International Conference on Communications (ICC'95): Vol 3, Jun 18-22, 1995, Seattle, WA, USA. Piscataway, NJ, USA: IEEE, 1995: 1912-1917.
- [8] Kim D S, DU Dingzhu. Performance of split routing algorithm for three-stage multicast networks[J]. IEEE/ACM Transactions on Networking, 2000, 8(4): 526-534.
- [9] YANG Yuanyuan, MASSOG G M. Fast path

routing techniques for nonblocking broadcast networks[C]//Proceedings of IEEE 13th Annual International Phoenix Conference on Computers and Communications, Apr 12-15, 1994, Tempe, AZ, USA. Piscataway, NJ, USA: IEEE, 1994: 358-364.

- [10] YANG Yuanyuan, WANG Jianchao. A more accurate analytical model on blocking probability of multicast networks[J]. IEEE Transactions on Communications, 2000, 48(11): 1930-1935.

收稿日期: 2007-09-27

### 作者简介



石增增, 西安电子科技大学计算机学院在读硕士研究生。主要研究方向为Clos交换网络。



顾华玺, 西安电子科技大学ISN国家重点实验室副教授。博士毕业于西安电子科技大学。主要研究方向为互连网络、片上网络以及无线传感器网络中的关键技术等,已发表论文30余篇。



王长山, 西安电子科技大学计算机学院副教授。毕业于吉林大学, 主要研究方向为计算机软件与网络技术。已发表论文40余篇。

### 中兴通讯承建巴基斯坦最大波分骨干网

2008年3月,中兴通讯对外公布获得巴基斯坦电信公司(PTCL)Quetta波分环和Rawalpindi至Mensehra波分链两个项目,这是其年初独家中标PTCL 400G波分干线项目的延伸。网络建成后,将会成为巴境内网络规模最大,业务承载量最大的波分干线网。全网覆盖巴基斯坦境内绝大部分重要城市和地区,长近6 000公里,承载了巴基斯坦全国60%以上的长途数据、语音、互联网等业务。

中兴通讯通过深入了解PTCL的需求和发展规划,最终赢得了400G波分干线项目的承建权。该工程采用业界唯一能提供6种波分层保护方式的大容量长途密集波分

设备ZXWM M900以及新一代MSTP设备ZXMP S385混合组网,并采用先进的IP Over DWDM解决方案,直接通过DWDM设备承载PTCL的核心IP骨干网。

中兴通讯建设了多个国家的大型骨干传输网络,如印度BSNL国家骨干传输网、巴基斯坦PAKTEL国家波分干线、保加利亚CableTel骨干传输网、欧洲跨国运营商GTS DWDM国干传输网,以及最近的葡萄牙AR Telecom骨干传输网、马来西亚国家骨干传输网、卢旺达MTN国家干线传输网、突尼斯全国网、哥伦比亚Orbitel国家干线波分网络等。目前,中兴通讯光网络产品已广泛应用于全球90个国家的250个运营商。