支持多时隙业务的 MTS-Clos 网络结构及其建模分析

孙 倩* 许 都
(电子科技大学宽带光纤传感与通信教育部重点实验室 成都 611731)

摘 要：多级 Clos 网络是一种典型的可扩展交叉互连结构，在数据通信和计算机并行网络中有着广泛应用。基于传统三级 Clos 网络 C(m, n, r)的理论分析表明，当满足 $m \geq 2n-1$ 时该网络是严格无阻塞的。该文针对多时隙业务，利用可快速实现时隙交叉的单级交换模块，构建了一种新型的 MTS-Clos 交换网络结构 C(m, n, r, t)(Multiple Time Slot Clos network)。利用时隙交叉能力，该结构在保留原有网络特性的同时可提供更为理想的交换性能。采用随机分析模型，对该结构的阻塞率进行了理论分析，结果表明当中间级规模 $m = n + k$，k 是一个很小的非负整数，网络即达到无阻塞。对该结构的数值仿真有同样的结论。因此，该结构及分析结果对下一代大容量交换设备的设计，具有良好的参考价值。

关键词：多级互联网络；阻塞率；多时隙；Clos 网络
中图分类号：TN915 **文献标识码**： A **文章编号**：1009-5896(2012)05-1226-05
DOI: 10.3724/SP.J.1146.2011.00744

Modeling and Analysis on MTS-Clos Networks Serving Multi-slot Traffic

Sun Qian Xu Du
(*Key Laboratory of Optical Fiber Sensing & Communication Ministry of Education, University of Electronic Science & Technology of China, Chengdu* 611731, *China*)

Abstract: The multistage Clos network or C(m, n, r) is widely deployed in data communications and parallel computing systems because of its scalability. It is well known the network is strictly non-blocking if $m \geq 2n-1$. A new architecture of Multiple Time Slot Clos network (MTS-C (m, n, r, t)) is proposed in this paper, which is based on fast time slot interchangeable unit and serving for multi-slot traffic. The MTS-C (m, n, r, t) network keeps the original features while providing better switching performance. A new analytical model assuming a random routing strategy is established, and under this model the blocking probability of the C (m, n, r, t) network is analyzed. The analytical and simulation results show that a C (m, n, r, t) network with a small number of middle stage switches m, such as $m = n + k$, where k is small constant, is almost non-blocking for unicast connections.

Key words: Multistage interconnection networks; Blocking probability; Multi-slot; Clos networks

1 引言

随着网络业务融合、移动应用、物联网、云计算等新兴战略产业的发展，信息网络基础设施加速向宽带融合、泛在智能方向演进。IP技术作为承载网的必然选择，需要满足高带宽、多样化的数据、语音、视频等大量融合业务的传输与交换压力[1]。特别是在核心网部分，IP网络的流量和规模的膨胀发展，必然使得系统容量、端口速率等方面性能的提升成为急需解决的问题。

当交叉连接容量较大时，Clos矩阵[2,3]需要控制的交叉结点数量比平方矩阵(即Crossbar)大为减少[4]。同时，若Clos矩阵的中间级设为固定容量，则当需要扩容时仅需扩大输入级和输出级的容量即可。这在很大程度上满足了容量平滑增长的要求，因此Clos矩阵是目前交叉连接设备的主流应用矩阵。

目前对Clos网络结构的研究，主要根据构成交换网络的交换模块是否具有缓存能力而分为两类：无缓存和有缓存Clos交换结构。对于前者，典型的SSS (Space-Space-Space)Clos网络已大量应用于包交换[5,6]，同时有学者提出可以将无缓存的Clos网络应用于数据中心网络[7]；对于后者，在传统共享缓存MMM(Memory-Memory-Memory)Clos结构的基础上，文献[8]在近期提出了MM^{e}M结构，以克服MMM

2011-07-20 收到，2012-02-29 改回
国家自然科学基金(60872031)资助课题
*通信作者：孙倩 xuelang185@hotmail.com

结构下的队头阻塞[8]。此外，文献[9]从理论和仿真方面研究了SMM(Space-Memory-Memory)Clos网络在高速突发业务下的网络性能。

对于一个交换结构，其在一定网络资源或规模下的阻塞率是其交换性能的关键表征参数。针对Clos交换结构，经典的概率分析模型有Lee模型[10]，Jacobaeus模型[11]和Yang模型[12]，这3个模型都是用来计算在随机选路策略下三级Clos网络的阻塞概率。其中Yang模型由于考虑了级间链路复用，故此其分析结果更为精确。 此外，文献[13]利用了类似的分析方法，对Clos网络的多播阻塞率进行了深入的研究。

不同于传统的Clos网络，本文提出一种应用于多时隙业务的三级Clos网络，并对其阻塞性能进行理论和仿真研究。多时隙业务是指输入模块接收的主要是SONET/SDH或OTN的数据流，此类数据流的特点是：高等级的数字信号系列可通过将低速率等级的模块通过字节间插复用而成。同时，近期VLSI设计技术与半导体工艺水平的发展，也使得MTS-Clos(Multiple Time Slot Clos network)结构的可实现性问题得以解决，如Velio公司的VC2002芯片即可高密度地实现快速时隙交叉。针对这种多时隙多级Clos网络结构，目前尚没有较为严格的理论分析模型可供参考，这使得基于传统分析结论所进行的工程应用设计具有一定的盲目性。

本文第2节将给出基于多时隙交换的三级Clos网络结构MTS-Clos；在第3节对多时隙环境下的三级Clos网络阻塞率分析模型进行论述；第4节是数值仿真结果及分析计论；最后是结论。

2　基于 TST 的三级 Clos 结构

传统的三级对称 Clos 交换结构如图 1 所示，其中一、三级是 r 个 $n\times m$ 的交换模块，中间级是 m 个 $r\times r$ 的交换单元。根据网络的阻塞特性可将其分为 3 类[4]，当满足 $m\geq 2n-1$ 时，是严格无阻塞网络；当满足 $m\geq n$ 时，是可重排无阻塞网络；第三是广义无阻塞网络。明显地，严格无阻塞条件下的网络硬件实现代价很高，这使得实际应用中可重排无阻塞网络成为首选。这类网络中业务寻径的性能主要决定于路由算法，而目前应用中的路由算法普遍存在重排次数多、重排路径长等缺点。

鉴于此，我们提出一种新的结构：即每一级不再是简单的空分结构或是共享缓存交换结构，而是由具有时分和空分功能的 TST(Time-Space-Time)构成，我们称具有 n 个输入、m 输出端口，链路容量是 t 的 TST 的交叉规模是 $nt\times mt$ 。图 2 是 4 个输入端口，链路容量是 4 时隙的 TST。图中每个字母代表一个颗粒，即 SONET/SDH 中的低速信号。由于低速信号是以字节间插方式复用进高速信号的帧结构中的，这样低速信号在高速信号里的位置是固定的、有规律性，也就是有可预见性。这样就能从高速信号中直接插/分出低速信号。

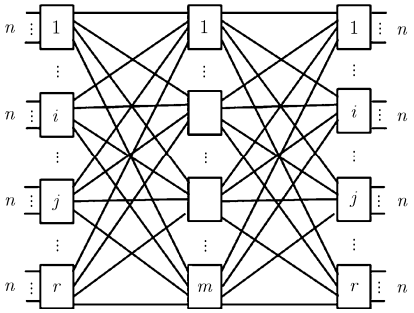

图 1 三级 Clos 网络 C(n, m, r)

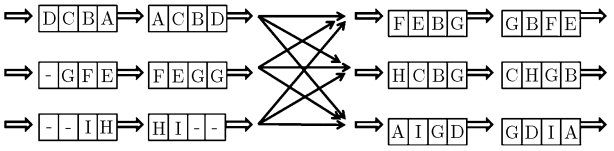

图 2 TST 交换结构

MTS-Clos 交换网络结构 C(m, n, r, t)是如下结构：输入、输出级是由 r 个交叉规模 $nt\times mt$ 的 TST；中间级是 m 个交叉规模 $rt\times rt$ 的 TST，如图 3 所示。输入级模块可将大颗粒业务拆分成小颗粒(基本颗粒)，如将任意 STM-N 的业务请求拆分成 STM-1 请求，在网络中独立寻路，然后在输出级模块进行时隙整合，恢复为原始大颗粒业务输出。

3　多时隙环境下的阻塞率分析模型

现有的分析模型主要针对传统的不具有多时隙交叉能力的 Clos 网络进行阻塞率分析，在此基础上，本节中对分析多时隙环境下 MTS-Clos 网络的阻塞率进行研究。

3.1 假设与符号定义

定义 1　如图 3 所示网络，所有模块(输入级、

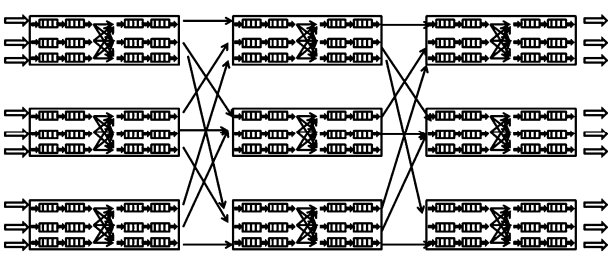

图 3 由 TST 构成三级 Clos 交换网络

中间级、输出级)都具有时隙交叉能力。设链路容量为 t 时隙，定义这样的网络为 MTS-C(m, n, r, t)。

定义 2 输入级第 g 个模块的输入端口定义为 $I_g = \{i_{g,i} \mid i \in \{1,\cdots,n\}\}$，其中 $g \in \{1,\cdots,r\}$。输出级第 h 个模块的输出端口定义为 $O_h = \{o_{h,j} \mid j \in \{1,\cdots,n\}\}$，其中 $h \in \{1,\cdots,r\}$。

定义 3 输入级第 g 个模块的第 i 个端口到输出级第 h 个模块的第 j 个端口的请求定义为 $C(i_{g,i}, o_{h,j})$。

定义 4 请求 $C(i_{g,i}, o_{h,j})$ 到达之前，I_g 模块 n_1 条链路忙，O_h 模块 n_2 条链路忙。设 n_1，n_2 中有 k 对链路共用中间级，则称为 k-链路复用(k- overlapped)。

假设 1 每条链路繁忙的事件是相互独立的。当业务负荷不是很高、网络规模较大时，该假设近似成立。

假设 2 每个时隙可以被随机地分配到有空闲的中间级链路上，即负载均匀分布在所有链路上。设 $p_{\text{slot}} \in [0,1]$ 是级间链路忙的概率，$q_{\text{slot}} = 1 - p_{\text{slot}}$ 是级间链路空闲的概率。

假设 3 输入级-中间级忙的链路与中间级-输出级忙的链路，其概率分布函数服从二项式分布。

假设 4 外部输入/输出链路和内部交叉链路的容量都是 t 个时隙，且交叉业务的带宽需求是基本带宽单元的倍数。带宽需求 $\beta(1 \le \beta \le t)$ 个时隙的业务用 β 表示，定义 β 为交叉业务的倍数因子。假设当前共有 s 个业务请求，第 i 请求的倍数因子是 $\beta_i\,(1 \le i \le s)$。定义 α 为外部链路利用率。则 $\alpha = \sum_{i=1}^{s} \beta_i \big/ (nt)$。

3.2 级间链路繁忙概率分析

首先求解级间链路忙的概率。由假设 4 可知，$B = \sum_{i=1}^{s} \beta_i$ 个基本时隙需求，被随机地分配在 m 条容量是 t 的链路上。这里有两种求解方法：

方法 1 求取方程整数解向量的个数

方程 1：$x_1 + x_2 + \cdots + x_m = B$, $(x_1, x_2, \cdots, x_m) \in [1,t]$ 的整数解向量的个数是 C_1。

方程 2：$x_1 + x_2 + \cdots + x_{m-1} = B - t$, $(x_1, x_2, \cdots, x_{m-1}) \in [1,t]$ 的整数解向量的个数是 C_2。

这种方法物理意义明确：方程 1 表示 B 个基本时隙需求，被随机地分配在 m 条容量是 t 的链路上，其整数解向量的个数 C_1 是 B 个基本时隙的选路方式。方程 2 表示 $B-t$ 个基本时隙需求，被随机地分配在 $m-1$ 条容量是 t 的链路上。整数解向量的个数 C_2 是 $B-t$ 个基本时隙的选路方式。综合上述两式，则某条链路忙的概率是

$$p_{\text{slot}} = C_2 / C_1 \tag{1}$$

上述两方程可以用贪婪算法求解，但其运算量与 m 成指数倍关系，当 m 较大时，运算量大，求解困难。

方法 2 母函数法

假设存在 m 种物品，每个物品的个数都是 t，这 m 种物品分别为 $x_1, x_2, \cdots, x_m$；重集 $Q_1 = \{t \cdot x_1, t \cdot x_2, \cdots, t \cdot x_m\}$。则上述方程 1 解向量的个数是求重集 Q_1 的 B 排列数。由于每种物品最多可以选 t 种，故母函数是：$f(x) = (1 + x + x^2 + \cdots + x^t)^m$。

$$\begin{aligned} f(x) &= \left(1 + x + x^2 + \cdots + x^t\right)^m \\ &= \sum_{a=0}^{m} C_m^a\,(-1)^a\, x^{a(t+1)} \cdot \sum_{b=0}^{\infty} C_{b+m-1}^b (x)^b \end{aligned} \tag{2}$$

令

$$f_1(x) = \sum_{a=0}^{m} C_m^a\,(-1)^a\, x^{a(t+1)} \cdot \sum_{b=0}^{\infty} C_{b+m-1}^b x^b \tag{3}$$

$$f_2(x) = \sum_{a=0}^{m} C_{m-1}^a\,(-1)^a\, x^{a(t+1)} \cdot \sum_{b=0}^{\infty} C_{b+m-2}^b x^b \tag{4}$$

其中 a，b 是正整数，t 表示链路容量，m 表示中间级数目。求出 $f_1(x)$ 中 x^B 的系数为 C_1，$f_2(x)$ 中 x^{B-t} 的系数为 C_2，则所要求的某个链路忙的概率是 $p_{\text{slot}} = C_2 / C_1$，空闲的概率：$q_{\text{slot}} = 1 - p_{\text{slot}}$。

3.3 网络 MTS-Clos(m, n, r, t)的阻塞率

引理 1 满足假设 1, 2, 3, 4，给定事件 n_1，n_2，在 MTS-Clos(m, n, r, t)中 k 对链路复用的概率是

$$\Pr\{k \mid n_1, n_2\} = \frac{\binom{n_1}{k}\binom{m-n_1}{n_2-k}}{\binom{m}{n_2}} = \frac{\binom{n_2}{k}\binom{m-n_2}{n_1-k}}{\binom{m}{n_1}} \tag{5}$$

其中 k 是链路复用的对数。

证明 在传统 Clos 网络中，此引理在文献[12]已给出证明方法，实际上，在 MTS-Clos(m, n, r, t)中的证明方法相同。

给定事件 n_1，n_2，k 对链路复用，则请求不被阻塞的条件是：$n_1 + n_2 - k < m$。注意到 k 的值应小于输入模块中忙的链路(n_1)和输出模块中忙的链路(n_2)，因此：$k \le \min\{n_1, n_2\}$。因此，给定事件 n_1，n_2，请求 $C(i_{g,i}, o_{h,j})$ 不被阻塞的概率是

$$\Pr\left\{C(i_{g,i}, o_{h,j}) \mid n_1, n_2\right\} = \frac{1}{\binom{m}{n_2}} \sum_{\delta_1}^{\delta_2} \binom{n_1}{k}\binom{m-n_1}{n_2-k} \tag{6}$$

其中 $\delta_1 = \max(0, n_1 + n_2 - m + 1)$，$\delta_2 = \min(n_1, n_2)$。根据假设 1，

$$\Pr\{n_1, n_2\} = \Pr\{n_1\} \cdot \Pr\{n_2\} \tag{7}$$

下面计算 $\Pr\{n_1\}$ 和 $\Pr\{n_2\}$，由假设 2，假设 3 在网络 MTS-C(m, n, r, t)中，　n_1 条输入级-中间级链路忙的概率是：$\binom{m}{n_1}p_{\text{slot}}^{n1}q_{\text{slot}}^{n1}$，但由于最多由 $n-1$ 条输入级-中间级链路忙，我们可得出更准确的概率：

$$\Pr\{n_1\}=\frac{\binom{m}{n_1}p_{\text{slot}}^{n_1}q_{\text{slot}}^{m-n_1}}{\sum_{j=0}^{n-1}\binom{m}{j}p_{\text{slot}}^{j}q_{\text{slot}}^{m-j}} \tag{8}$$

n_2 条中间级-输出级链路忙的概率：

$$\Pr\{n_2\}=\frac{\binom{m}{n_2}p_{\text{slot}}^{n_2}q_{\text{slot}}^{m-n_2}}{\sum_{j=0}^{n-1}\binom{m}{j}p_{\text{slot}}^{j}q_{\text{slot}}^{m-j}} \tag{9}$$

由式(6)–式(9)中，我们可以得出请求 $C(i_{g,i}, o_{h,j})$ 不被阻塞的概率为

$$\Pr\left\{C\left(i_{g,i},o_{h,j}\right)\right\}=\frac{\sum_{n_1=0}^{n-1}\sum_{n_2=0}^{n-1}\sum_{k=\delta_1}^{\delta_2}\binom{m}{n_1}\binom{n_1}{k}\binom{m-n_1}{n_2-k}p_{\text{slot}}^{(n_1+n_2)}q_{\text{slot}}^{(2m-n_1-n_2)}}{\left[\sum_{j=0}^{n-1}\binom{m}{j}p_{\text{slot}}^{j}q_{\text{slot}}^{m-j}\right]^2} \tag{10}$$

其中 $\delta_1=\max(0,n_1+n_2-m+1)$，$\delta_2=\min(n_1,n_2)$。

$$P_{B\text{slot}}=1-\Pr\left\{C\left(i_{g,i},o_{h,j}\right)\right\} \tag{11}$$

4　仿真及分析

仿真平台采用 VC++编写，网络 C(m, n, r, t) 的所有参数均可配置，主要由业务生成模块和业务处理模块构成。业务生成模块用于生成合法业务请求，然后由业务处理模块按照随机策略分配中间级，如若不能配通，即计为一次阻塞。

Yang 模型适用于基于 Crossbar 的空分 Clos 网络的分析，这类网络每一级都不具备时隙调整能力。图 4 是两种模型下同等配置下的 Clos 网络阻塞率分析结果。从图中的分析结果可以看出，基于 Crossbar 的空分 Clos 网络的阻塞率明显高于基于 TST 的 Clos 网络，同时三级时隙可调的 Clos 网络的阻塞率比传统的线路交换网络的阻塞率下降的更快。当然这是以牺牲硬件复杂度为代价的，Yang 模型所适用的是空分的 Clos 网络，这类网络硬件复杂度比文中所提的基于 TST 的 Clos 网络要小。

图 5 是网络 C(m,20,20,4)的仿真结果和分析结果的对照图，链路利用率 $\alpha=0.9$；从图中可以看出阻塞率随着中间级规模的增大迅速下降。随着中间级规模的增大，阻塞率变小就意味着配通率的增大，重构次数的降低。分析结果稍大于仿真结果，也可以说分析模型保守地估计了网络的阻塞率；但两结果相差不大且保持着相同的趋势说明分析模块有着良好的精度。

下面考查链路容量 t 对阻塞率的影响。图 6 所示是网络 C(20,20,20,t)，$\alpha=0.9$，阻塞率随链路容量 t 的变化图。图 7 所示是网络 C(20,21,20,t)，$\alpha=0.9$，阻塞率随链路容量 t 的变化图。从两图都可以看出，随着链路容量的增加，阻塞率呈下降趋势。是因为随着链路容量 t 的增加，p_{slot} 呈下降趋势，并且当 $m>n$ 时，p_{slot} 呈下降的更快，比较图 6，图 7，可以得出：当 $m>n$ 时阻塞率随链路容量下降比 $m=n$ 时更快，在我们所配置的网络中，阻塞率下降为 0；即当链路利用率 $\alpha=0.9$，网络 C(20,21,20,t)在足够大的链路容量下可以将业务全部配通。所以在构建交换网络时，如果能提供冗余的交换模块将会极大地降低重排次数。

5　结论

本文所提出的基于 TST 的 Clos 网络，与基于 Crossbar 的空分 Clos 网络相比虽然在构成上略为复杂，但在相同的网络规模下，基于 TST 的 Clos 网络的阻塞率更低；尤其特别的是当中间级规模 $m>n$ 时，阻塞率下降得更快，当中间级规模 $m=n+d$，d 是一个很小的非负整数；配通率就可以达到 100%。$m=n$ 时 Clos 网络是可重排无阻塞

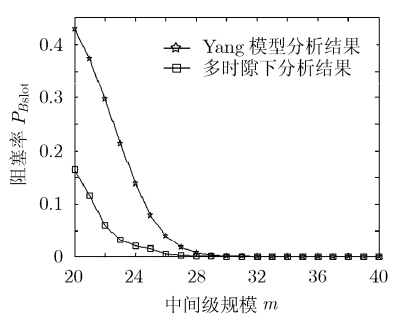

图 4 两种模型下的 Clos 网络阻塞率分析结果

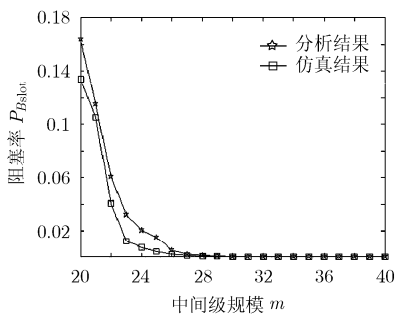

图 5 网络 C(m,20,20,4)的仿真与分析结果

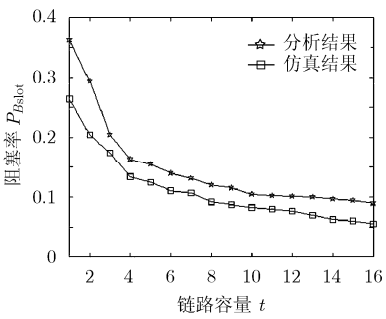

图 6 网络 C(20,20,20,t)阻塞率随 t 变化图

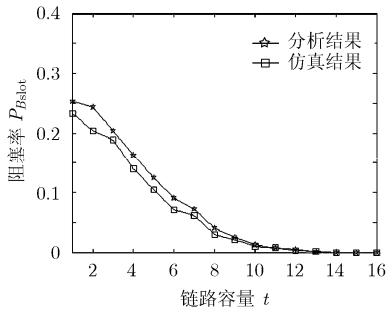

图 7 网络 C(20,21,20,t)阻塞率随 t 变化图

网络，但重排会引入噪声和误码；在可靠性要求很高的场合，可采用 MTS-Clos 结构，只需增加少量中间级就使网络无阻塞。

本文通过理论分析和仿真实验证明了 MTS-Clos 网络的阻塞率与链路容量成负相关：输入链路容量和级间链路容量同时增加时，网络的阻塞率随之降低。该性质是基于 Crossbar 的 Clos 网络所不具备。

参 考 文 献

[1] 代晓慧. 对十二五期间信息通信业发展的思考[OL]. www.c114. net/topic/2561/a570678.html, 2010.12.

[2] Clos C. A study of non-blocking switching networks [J]. *The Bell System Technical Journal*, 1953, 32(3): 406–424.

[3] Chao H J and Liu B. High Performance Switches and Routers [M]. New Jersey: Wiley-IEEE Press, 2007: 382–408.

[4] Benes V E. Mathematical Theory of Connecting Networks and Telephone Traffic [M]. New York: Academic Press, 1965: 53–65.

[5] Dorren H J, Calabretta N, and Raz O. A 3-stage CLOS architecture for high-throughput optical packet switching[C]. Communications and Photonics Conference and Exhibition (ACP), Shanghai, China, Nov. 2–6, 2009: 1–6.

[6] Oki Eiji, Kitsuwan Nattapong, and Rojas-Cessa R. Analysis of space-space-space Clos-network packet switch[C]. 2009 Proceedings of 18th International Conference on Computer Communications and Networks, San Francisco, CA, USA, Aug. 3–6, 2009: 1–6.

[7] Chao H J and Kang Xi. Bufferless optical Clos switches for data centers[C]. Optical Fiber Communication Conference and Exposition (OFC/NFOEC), Los Angeles, CA, USA, March 6–10, 2011: 1–3.

[8] Dong Zi-qian and Rojas-Cessa R. Non-blocking memory-memory-memory Clos-network packet switch[C]. Sarnoff Symposium, Princeton, NJ, USA, 2011: 1–5.

[9] Ruepp S, Rytlig A, and Manolova A V. Performance evaluation of 100 Gigabit Ethernet switches under bursty traffic[C]. Optical Network Design and Modeling (ONDM), Bologna, Italy, Feb. 8–10, 2011: 1–6.

[10] Lee C Y. Analysis of switching networks [J]. *The Bell System Technical Journal*, 1955, 34(6): 1287–1315.

[11] Jacobaeus C. A study of congestion in link system [J]. *Ericsson Techniques*, 1950, 51(3): 1–68.

[12] Yang Y. An analytical model on network blocking probability [J]. *IEEE Communications Letters*, 1997, 1(5): 143–145.

[13] Pattavina A and Tesei G L. Modeling the blocking behavior of multicast Clos networks[C]. INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications, IEEE Societies, San Francisco, CA, USA, 2003, Vol.1: 756–763.

孙 倩： 男，1984 年生，硕士，从事通信网络技术研究.

许 都： 男，1968 年生，教授，从事路由器体系结构和高速网络设备研究.