软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(3):593−603 [doi: 10.3724/SP.J.1001.2013.04251]

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

Multi-log$_2N$ 交换网络的性能分析模型及控制算法*

刘晓锋[1,3], 赵有健[2], 吴亚娟[3]

[1](电子科技大学 计算机科学与工程学院,四川 成都 611731)

[2](清华大学 计算机科学与技术系,北京 100084)

[3](西华师范大学 计算机学院,四川 南充 637002)

通讯作者: 刘晓锋, E-mail: xhxfliu@163.com

摘 要: 高速多平面交换网络解决了其内部冲突问题,但需要相应的路由控制算法的辅助,否则,内部冲突不能彻底解决.这是因为包在输入级路由平面的选择不够恰当,容易导致路由冲突的产生.因此,根据冲突链路集的思想,给出一种 Multi-log$_2N$ 交换网络的控制算法.该算法控制分组在路由平面间的选择,不仅能够适用于 RNB 和 SNB,还能实现单播和多播的控制,保障 Multi-log$_2N$ 完全实现无阻塞.另一方面,Multi-log$_2N$ 消除了内部的链路冲突,提高了交换速率,但对其交换性能缺乏系统的理论分析.给出一种基于嵌入式马尔可夫链的分析模型,对 Multi-log$_2N$ 网络中队列的使用及分组在队列中的平均等待时间、平均队长等相关性能指标进行了系统的分析,为基于 Multi-log$_2N$ 的光交换节点的设计提供了良好的理论依据.

关键词: Multi-log$_2N$;交换网络;多级网络;控制算法;自选路由;阻塞

中图法分类号: TP393 **文献标识码**: A

中文引用格式: 刘晓锋,赵有健,吴亚娟.Multi-log$_2N$ 交换网络的性能分析模型及控制算法.软件学报,2013,24(3):593−603. http://www.jos.org.cn/1000-9825/4251.htm

英文引用格式: Liu XF, Zhao YJ, Wu YJ. Control algorithm and performance analysis model of multi-log$_2N$ switching networks. Ruanjian Xuebao/Journal of Software, 2013,24(3):593−603 (in Chinese). http://www.jos.org.cn/1000-9825/4251.htm

Control Algorithm and Performance Analysis Model of Multi-log$_2N$ Switching Networks

LIU Xiao-Feng[1,3], ZHAO You-Jian[2], WU Ya-Juan[3]

[1](School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

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

[3](School of Computer, China West Normal University, Nanchong 637002, China)

Corresponding author: LIU Xiao-Feng, E-mail: xhxfliu@163.com

Abstract: Although high-speed multi-plane switching networks have removed their internal conflict problem, a routing control algorithm is necessary for realizing conflict-free routing. Otherwise, the conflict phenomenon cannot be totally avoided. This is because the routing plane may be chosen inappropriately by the incoming packet at the input stage. Therefore, a control algorithm based on the idea of conflict links set is presented in this paper. This algorithm controls the allocation of packets among routing planes in the multi-log$_2N$ switching networks, and hence, the conflict-free routing is totally guaranteed. Moreover, it is not only applicable for the RNB and SNB, but also suitable for unicast and multicast. On the other hand, inner link conflicts are removed in multi-log$_2N$ networks. The switching efficiency is improved, but no performance analysis models can be used to analyze the switching performance of Multi-log$_2N$ switching networks. So an analysis model based on embedded Markov chain is proposed in this paper, and is adopted to analyze the queue

* 基金项目: 国家自然科学基金(60903184, 61073167); 国家高技术研究发展计划(863)(2011AA010704); 西华师范大学重大培育项目(09A003); 西华师范大学科研启动项目(07B015)

收稿时间: 2011-10-18; 定稿时间: 2012-04-01; jos 在线出版时间: 2012-07-27

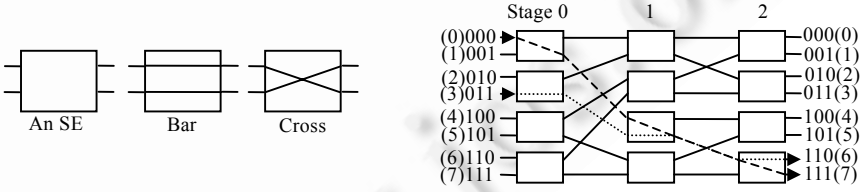
CNKI 网络优先出版: 2012-07-27 10:55, http://www.cnki.net/kcms/detail/11.2560.TP.20120727.1055.001.html

management and the relevant performance measures in detail, such as the mean waiting time, queue length and the probability of packets loss. All these conclusions are capable of providing well theoretical support for the design of the optical switching architecture based on multi- $\log_2 N$ switching networks.

Key words: multi-$\log_2 N$; switching network; MIN; control algorithm; self-routing; blocking

在通信系统与计算机网络技术的发展中,交换网络起了至关重要的推动作用.如果没有这项技术,所有用户通过电缆或光纤直接相连,整个地球表面除了电缆(或光缆)可能什么都没有.交换网络的发展经历了几个重要阶段[1],其总体发展趋势表现为由通用器件向专用器件,由串行处理向并行处理,从集中式到分布式.

在高性能交换网络中,多级互联网络(multistage interconnection network,简称 MIN)因具有良好的结构属性[2],在交换网络中一直倍受关注.特别在网络业务多样化、网络用户急剧增长的形势下,交换能力与成本产生冲突,此时,单级交换结构已有些力不从心了,而 MIN 在解决这对矛盾时却显得游刃有余.MIN 是由多个交换单元(switching element,简称 SE)通过链路互连成多级,每个 SE 可以是一个 $m\times n$ 的定向耦合器(directional coupler,简称 DC).不失一般性,本文假设 $m=n=2$,即每个 SE 具有两个输入和两个输出,如图 1(a)所示.一个 $N\times N$ 的 MIN 网络具有 $n=\log_2 N$ 级,通常称为 $\log_2 N$ 网络.$\log_2 N$ 网络具有良好的结构属性:首先是自选路由(self routing)或路由唯一性,这种属性确保了整个路由过程不需要额外的控制,只需知道源-目的端口号就能够顺利完成整个路由;其次,$\log_2 N$ 具有较低的网络直径 $O(\log_2 N)$;第三,$\log_2 N$ 的任何输入/输出请求具有相同的路由距离.后两点属性使 $\log_2 N$ 网络非常适合光交换节点,因为具有较低的衰减而且衰减均匀.遗憾的是,自选路由决定了 $\log_2 N$ 是一种带阻塞的体系结构,这对 $\log_2 N$ 的性能发挥有极大影响.阻塞(blocking)是指一个 SE 的两个输入争用同一输出,如图 1(b)中请求(0→7)和(3→6),在第 1 级的输出链路 5 发生了冲突.面临冲突,通常是在每个 SE 中设立缓存,在产生冲突的两个分组中任选一个通过,另一个存于缓存等待下一时隙再传送.这会增加传输延迟,降低交换速度.因此在交换系统中,无冲突的交换网络可在任意的输入/输出之间无冲突地建立请求,具有很强的吸引力.解决冲突的一个办法是构建 Multi-$\log_2 N$[3]网络.Multi-$\log_2 N$ 网络是将多个 $\log_2 N$ 通过水平级连(horizontal cascading,简称 HC)、垂直堆叠(vertical stacking,简称 VS)以及这两种的组合来实现内部链路无阻塞.在 Multi- $\log_2 N$ 作为光交换节点时,这 3 种构造方法会有不同的影响.HC 方法导致 Multi-$\log_2 N$ 的网络直径增大,光的衰减增加,因此,HC 型 Multi-$\log_2 N$ 不宜做光交换节点;VS 不会增加 Multi-$\log_2 N$ 的网络直径,非常适合做光交换节点,但在 Multi-$\log_2 N$ 的输入级需要相应的控制算法来控制路由平面的选择,否则无法真正实现无阻塞.



(a) An SE and its two states bar and cross (b) A 8×8 的 $\log_2 8$ 网络(baseline)

Fig.1 An example of MIN, SE and its two states

图 1 多级互联网络、交换单元及其两状态

随着光波技术的日渐成熟,网络业务走向多元化,光通信网络逐渐成为互联网络的主题.其中,光交换结构为光信号的交换提供了物理支持,应该具备低串音(crosstalk)、均匀损耗、内部无阻塞及灵活扩散等特点.VS 型 Multi-$\log_2 N$ 能够满足这些基本要求,因此,VS 型 Multi-$\log_2 N$ 作光交换结构又引起了人们的关注.Multi-$\log_2 N$ 通过破坏其路径唯一性来实现无阻塞,与此同时也破坏了自选路由的特性,以至于在实现无阻塞时需要相应的路由算法的支持,否则无法保障无阻塞.但在 Multi-$\log_2 N$ 的相关研究中,主要集中在实现无阻塞所需要 $\log_2 N$ 网络(routing plane,路由平面)的个数,很少涉及 Multi-$\log_2 N$ 网络的路由算法.Lea[3]给出了 Multi-$\log_2 N$ 在单播环境中实现无阻塞所需要的路由平面数,Tscha[5]给出了 Multi-$\log_2 N$ 实现严格非阻塞所需的路由平面数,但缺乏相应的

切实可行的路由算法来控制路由平面的选择问题.本文的一部分工作就是解决VS型Multi-$\log_2N$网络中路由平面的选择问题,当Multi-$\log_2N$网络作为光交换节点,这是相当重要的理论基础.

1 Multi-$\log_2N$相关研究

多级互连交换网络作为高速、高性能交换网络得到了广泛研究,其中,关于MINs的性能分析是其中非常重要的研究课题.由于$\log_2N$网络的阻塞特性影响其交换性能,人们为了提高其交换性能,不仅在输入/输出端设立缓冲队列,而且在其内部的每个SE也设立了相应的缓存队列(internal buffers,内部缓存).大量关于$\log_2N$性能分析的文献都是以某种网络(如Banyan网络[6]、Omega网络和Delta网络[7]等)为背景分析分组在内部缓存中的行为.Patel[7]对无缓冲的Delta网络做了性能分析的研究,得到了其吞吐量是关于网络中每一级输出强度的递归表达式,且随网络规模的增大而增大.Kumar[8]在Patel的研究基础上对结论进行了扩充,得到了无缓冲Banyan的吞吐量的上下界,且通过网络复制(networks replication)和链路扩张(links dilation)对其进行优化. Kruskal[9]给出关于Patel结论的渐近解形式.这是对无缓冲$\log_2N$的早期性能研究.为了降低阻塞的可能性,在$\log_2N$中添加内部缓冲,使阻塞只可能发生在一个SE(2×2)的两个缓存中,都有一个分组去争用同一个输出链路.根据缓存容量、缓存部署位置和流量模型,进行分类研究.Theimer等人[10]、Turner[11]和Jenq[12]对有限前置缓存(缓存设置在SE的输入端)$\log_2N$进行了研究.Lin[13]考虑了任意的流量模式.这些研究几乎都采用了同一种研究方法——近似马尔可夫链(approximated Markov chains).用马尔可夫链模拟$\log_2N$,所需的状态数随$\log_2N$的级数呈指数增长,当$N\to\infty$时,几乎不可能对其进行精确分析,因此只能用近似的马尔可夫链作为分析工具.当然,得到的结果存在一定的误差.另外,所有这些分析没有涉及Multi-$\log_2N$网络的性能研究,而是立足于$\log_2N$网络,通过在SE里设立缓存来解决冲突问题.本文的另一部分研究工作就是对Multi-$\log_2N$的交换过程、队列的使用及相关性能指标进行分析研究.

交换网络通过控制器执行路由算法在输入/输出之间建立链路.$\log_2N$网络是通过目的地址自动控制链路的建立,而Multi-$\log_2N$网络是由多个$\log_2N$网络构成,破坏了$\log_2N$网络的自选路由的特性.如果没有相应的控制算法控制到达的分组在多个$\log_2N$网络之间的选择,是难以实现无冲突的.Lea[3]根据$\log_2N$网络的阻塞情况构建路径相交图(path-intersection graph,简称PIG),然后利用着色原理对PIG着色,最后用不同颜色代表Multi-$\log_2N$网络中不同的路由平面,到达的分组根据自己目的端口对应的颜色分别路由到相应的路由平面. Kabacinski[14]提出一种实现WNB的路由算法,该算法具有较高的时间复杂度,即使使用并行计算技术亦如此. Lu等人[15]根据图论提出了实现RNB的路由算法,利用并行计算技术可使时间复杂度达到$O(\lg^2N)$.本文在Wu等人[16]研究的基础上,根据请求在一个路由平面上产生的冲突链路集提出一种路由算法,虽然时间复杂为$O(N^2)$,但形式简单,易于实现,既可控制RNB的单播路由,也可控制SNB的多播路由,具有较好的适应性.

2 Multi-$\log_2N$网络的路由算法

Lea[3]提出的Multi-$\log_2N$网络具有容错性好、网络直径小($O(\log_2N)$)且所有请求的路径长度相等的优点,特别适合光交换结构.在Multi-$\log_2N$的相关研究中,很多研究是关于使Multi-$\log_2N$成为非阻塞所需要的路由平面数,较少有关于分组在这些路由平面间的选择问题.在构建Multi-$\log_2N$的3种方法[3,4]中,HC型Multi-$\log_2N$不太适合光交换结构,因为这种结构不仅破坏了路径的唯一性且增加了网络直径,从而复杂了系统容错、查错能力以及增加光传输过程中损耗;而VS型Multi-$\log_2N$更适合光交换结构,这也是在交换节点成为光通信系统的速度瓶颈时,VS版Multi-$\log_2N$引起众多学者研究的一个原因.在本文的研究中,如无特别声明,Multi-$\log_2N$均指VS型.

Lea[3]给出Multi-$\log_2N$在单播环境中实现可重排非阻塞所需的平面数,$m\geqslant 2^{\lfloor n/2\rfloor}$,$n(=\log_2N)$为$\log_2N$网络的级数.不足的是,这个结论只适用于单播(unicast)系统,能否推广到多播(multicast)系统中还需进一步研究.

Tscha[5]给出了Multi-$\log_2N$实现严格非阻塞所需的平面数,当n为偶数时,有$m\geqslant\left(\frac{\delta n}{4}\right)+1$;$n$为奇数时,有

$m \geqslant \left(\frac{\delta}{2}\right)(n-1)+1$.其中,$n=\log_2 N$,$\delta=2^{\lfloor n/2 \rfloor}$.Tscha 的结论虽然是针对严格非阻塞的,但它既适用于单播,也适用于多播.

VS 型的 Multi-$\log_2 N$ 虽然破坏了自选路由的特性,但每个路由平面仍具有此特性,只要到达的分组确定了相应的路由平面,该分组就可以顺利完成路由.本文在 Wu 等人[16]研究的基础上,给出分组在 Multi-$\log_2 N$ 网络中一种形式化的路由算法.在描述路由算法之前,先给出如图 2 所示模型及相应假设.

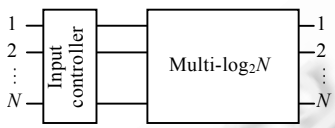


Fig.2 Multi-$\log_2 N$ model

图 2 Multi-$\log_2 N$ 模型

在图 2 所示的模型中,Multi-$\log_2 N$ 是由 m 个路由平面通过 VS 模式构建的非阻塞网络,隐去了路由平面之间的连接链路,在输入端与 Mulit-$\log_2 N$ 之间添设一个输入控制器(input controller).输入控制器主要完成以下 3 个方面的功能:首先是检测当前到达的分组是否存在冲突;其次,对存在冲突的分组执行相应路由算法,在 Multi-$\log_2 N$ 内实现无冲突传输.一个时钟分成两个阶段,即 $\tau=\tau 1+\tau 2$,在 $\tau 1$ 内计算路由表,在 $\tau 2$ 内按所得路由表路由各请求对;第三,处理多播问题.模型中也隐去了输入/输出的缓冲队列,这将在稍后的第 3 节讨论.

Multi-$\log_2 N$ 网络的路由算法的基本思想是,输入控制器根据到达分组源-目的端口计算它们在 $\log_2 N$ 中的冲突情况,存在冲突的两个分组分别由两个不同路由平面来完成相应的路由.计算冲突的方法[16]是:先根据分组的请求计算出其需要经过的链路集,如 16×16 网络中的请求对 1100→0101,该请求对所经过的链路集是 $P=\begin{pmatrix}1100\\0101\end{pmatrix}=\begin{pmatrix}12\\5\end{pmatrix}=\{12,12,7,6,5\}$ (为便于描述,2 进制形式化成 10 进制);再计算这些链路集的交集,根据交集的结果找出产生冲突的请求对.算法具体描述为:

Step 1:设有路由平面集数组(路由表)*Plane*[*m*],其中,元素 *Plane*[*k*]是一个请求集合,存放经过平面 *k* 到达目的地的请求对,其初值为 *Plane*[*k*]=∅,0≤*k*<*m*.

Step 2:计算请求对的链路集,记第 *i* 个请求对产生的链路集为 *l*(*i*),0≤*i*<*N*;

因为多级网络是位置换网络(bit permutation network)[17],一个请求从第 *i* 级传输到第 *i*+1 级,经过一个轮换函数的作用,就可以准确地定位到第 *i*+1 级的链路上,所以整个过程经过 *N* 次轮换,就得到相应的链路集.

Step 3:计算链路集的交集,记为 $L(i,j)=l(i)\cap l(j)$,$0\leqslant i,j\leqslant N, i\neq j$.

For (*i*=0; *i*<*N*; *i*++)

For (*j*=*i*+1; *j*<*N*; *j*++)

[$L(i,j)=l(i)\cap l(j)$;

If $L(i,j)\neq\varnothing$ then

把请求 *i*,*j* 分别加入两个不同的路由平面集合 *Plane*[*k*]和 *Plane*[*t*],要求不与 *Plane*[*k*]或 *Plane*[*t*]中已存在的请求发生冲突,否则,需更换新的路由平面集合;

]

$L(i,j)=\varnothing$,请求 *i*,*j* 可以进入任意一个平面 *Plane*[*k*].

根据平面集 *Plane* 将请求对成功地路由到相应的目的地.

下面以图 3 所示的 8×8 Baseline 交换网及表 1 所示的请求对来描述该算法.根据前述有 $m\geqslant 2$,所以取 $m=2$.路由平面集中各元素的初值为∅.因为 $L(0,1)=\{1\}$,所以请求对 *r*0 和 *r*1 要进入不同的路由平面,即 *Plane*[0]={*r*0},*Plane*[1]={*r*1};又 $L(1,2)=\{4,5\}$,请求对 *r*1 和 *r*2 进入不同路由平面,但要不和路由平面里已有的请求对产生

冲突,因此 *Plane*[0]={*r*2};*L*(4,6)={2},*r*4 和 *r*6 进入不同的路由平面,有 *Plane*[0]={*r*0,*r*4},*Plane*[1]={*r*1,*r*6};*L*(6,7)={6},*Plane*[0]={*r*2,*r*7};其他的请求对 *r*3 和 *r*5 没有发生任何冲突,可以在 2 个路由平面内任意路由,因此可得最后的路由为 *Plane*[0]={*r*0,*r*2,*r*3,*r*4,*r*7},*Plane*[1]={*r*1,*r*5,*r*6},如图 4 所示.该算法的效率为 $O(N^2)$.

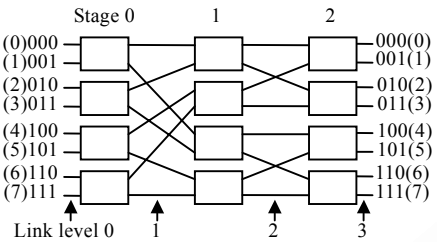


Fig.3 8×8 log$_2N$ (*N*=8)

图 3 8×8 log$_2N$ 网络(*N*=8)

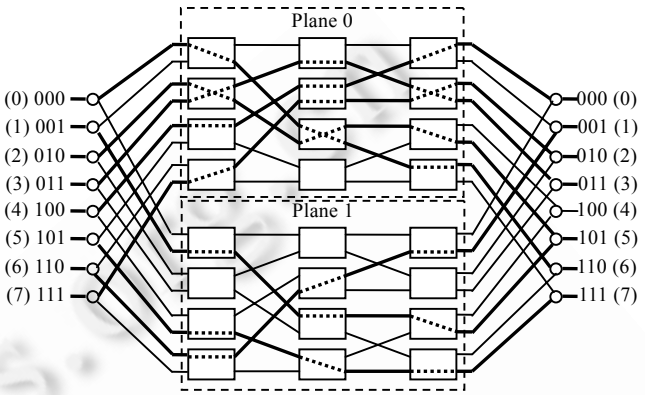


Fig.4 Routing illustration of requests *r*0 to *r*7

图 4 请求 *r*0~*r*7 的路由示意图

Table 1 A request table

表 1 请求表

请求对(10 进制)	经过的请求链路集(10 进制)
*r*0:0→6	*l*(0)={0,1,5,6}
*r*1:1→5	*l*(1)={1,1,4,5}
*r*2:2→5	*l*(2)={2,3,4,5}
*r*3:3→3	*l*(3)={3,2,1,3}
*r*4:4→0	*l*(4)={4,4,2,0}
*r*5:5→7	*l*(5)={5,5,7,7}
*r*6:6→1	*l*(6)={6,6,2,1}
*r*7:7→2	*l*(7)={7,6,3,2}

该路由算法不仅适用于单播,也适用于多播情形.对多播的情况,在计算链路集和路由表时将多播当成多个单播来处理.但需要注意的是,一个多播在传输过程中可以重叠,不存在冲突,应放在同一个路由平面,只是在路由过程中要对分组做另外的处理.其处理思想[18]如下:

输入控制器根据多播要到达的目的端口地址,按位串编码的方案重新构造一个新目的地址串 $d=d_0d_1\ldots d_{N-1}$,$d_i\in\{0,1\}$,$0\leqslant i\leqslant N-1$.如果目的端口地址为 *i*(10 进制),则 $d_i=1$;否则,$d_i=0$.源端口按新目的地址 *d* 进行路由.交换单元将 *d* 平分成两部分,如果前后两部分都含有 1,则该分组要经过此交换单元的上、下链路传送到下一级;如果只是前半部分有 1,则只通过交换单元的下链路传送到下一级;如果只是后半部分含有 1,则只通过交换单元的上链路传送到下一级;到了下一级重复此过程,直至分组到达相应目的地,如图 5 所示.

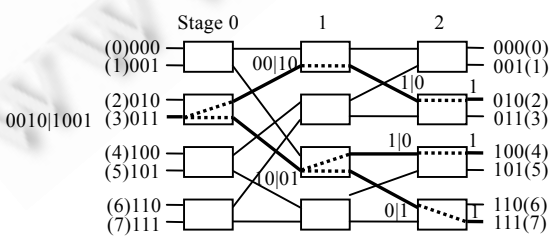


Fig.5 Multicast requests (Req:3→2,4,7; d=00101001)

图 5 请求多播路由(Req:3→2,4,7;d=00101001)

3 Multi-log$_2N$ 网络的性能分析模型

Multi-log$_2N$ 解决了其内部的链路阻塞问题,到达输入端的分组可以被无阻塞地传送到输出端,但 Multi-log$_2N$ 在分组交换时存在以下两个问题:(1) 要实现完全无阻塞,只知道所需要的路由平面数是不够的.如果没有路由算法来控制分组在多个平面间的分配,是无法保障无阻塞.这是本文第 2 节所解决的问题;(2) Multi-log$_2N$ 虽然解决了内部链路阻塞问题,但没有解决 I/O 争用问题,即在 Multi-log$_2N$ 网络的输入/输出可能出现争用输入端口/输出端口的现象.分组的到达过程并非按照某种计划有条不紊地进行,而是具有一定的突发性和随机性.所以,在 Multi-log$_2N$ 中设立缓冲队列是不可避免的,如图 6 所示.

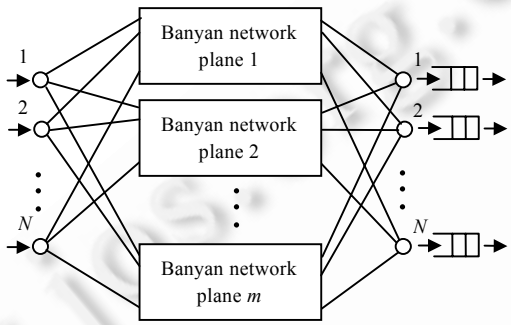


Fig.6 Multi-log$_2N$ without blocking

图 6 无阻塞的 Multi-log$_2N$ 网络

实际上,Multi-log$_2N$ 是一种带输出缓存且无阻塞的多级交换结构.Mulit-log$_2N$ 和 Crossbar(带输入缓存)都属内部无阻塞交换结构,但这两种结构有本质的不同.Crossbar 可能导致队首(head of the line,简称 HOL)阻塞,而 Multi-log$_2N$ 是通过多个路由平面来实现内部无阻塞,不可能导致 HOL 现象.如果 Multi-log$_2N$ 的每个输入都处于饱和状态(在任何时候都有分组),其交换速度不可能等于输入端的线速,而且一定会是线速的 N 倍.所以分组不可能在输入端排队等候,只是在输出端排队.表 1 中,请求 $r1$ 和 $r2$,如果在 Crossbar 结构中就会造成 HOL 阻塞,而在 Multi-log$_2N$ 结构中却不会.因此,Multi-log$_2N$ 是一种带输出缓存且无阻塞的多级交换结构.虽然这种体系结构在全光交换环境中也许不能充分发挥其性能,因为目前缺乏可适用的光缓存,还只能使用光延迟线(FDL),但随着技术的发展,这些问题都能得到很好地解决.

性能问题是任何系统都不能回避的.对网络系统而言,性能分析的精准性和服务质量(QoS)保证等都直接受控于某个较准确的流量模型.但是,随着网络应用的日趋复杂,网络流量特性亦不断发生变化,需要一种比较准确的流量模型来捕获不同网络应用的性能是比较困难的,因此,流量模型跟随网络应用的多样化而不断发展完善[19].下面首先对与本文工作相关的流量模型进行简单讨论,其次根据队列的容量是否有限详细分析此性能模型.

3.1 相关流量分析模型

在进行网络相关性能分析时,可通过流量模型来捕获相关统计特征.因此,流量模型的准确性直接决定性能分析的准确性.目前,对交换结构性能的研究有很多理论成果,大多集中在单平面交换网络,而其重要的基础是均匀的流量模式,即分组等概率到达交换结构的输入端和输出端等.有研究表明,很多时候,流量模式对交换结构并非是均匀的.特别是在网络节点呈指数增长以及网络应用纷繁复杂的情形下,网络流量呈现出自相似特性(或突发性)[20],传统的流量模型已不能很好地刻画网络流量的这种特性.在表征网络流量的自相似特性时,通常采用间歇泊松过程(interrupted Poisson process,简称 IPP),通常称为 ON-OFF 模型.ON-OFF 模型是一个 2 状态的 Markov 过程,分组只在 ON 状态下产生,而且可以连续产生多个分组(称为一个突发).即,系统可停留在 ON 状态一段时间(连续多个时隙);在 OFF 状态下不产生分组,当然,在 OFF 状态也可以停留连续多个时隙.在 ON 状态或

OFF 状态停留的时间是一个随机变量,通常认为,该随机变量服从几何分布或重尾分布(heavy tailed distribution)[20].在图 7 中,p 和 q 分别表示从 ON 状态到 OFF 状态和从 OFF 状态到 ON 状态的转移概率.如果假定在 ON 状态或 OFF 状态的停留时间服从几何分布的话,则有 $\Pr(\text{ON}=i \text{ slots})=p(1-p)^{i-1}, i\geq 1$ 和 $\Pr(\text{OFF}=j\text{slots})=q(1-q)^j, j\geq 0$.通过停留时间的概率分布,可以得到一个突发的平均长度及平均负载量等性能指标.

尽管目前网络流量相对以前有很大不同,呈现自相似特性,但在分析网络性能时,通常还是在均匀和非均匀两种流量模式下进行讨论.我们在建立交换结构性能分析模型时,假定分组的传输与时隙同步,即传输任务随一个时隙开始而开始、随一个时隙的结束而结束.而分组是定长的,交换结构传输一个分组就是一个时隙.考虑到在一定条件下(如 $p+q=1$),贝努利(Bernoulli)到达过程实际是上述突发到达过程的一个特例,以及篇幅的原因,我们假定分组是按独立同分布的贝努利过程到达 Multi-log$_2N$ 的输入端.在一个给定时隙,分组以概率 p 到达输入端,同时,p 也表示该端口的利用率(utilization)和分组的到达强度.假设 Multi-log$_2N$ 网络是由 m 个路由平面构成,到达输入端口的分组以概率 p_j 选择平面 j,最后,等概率且独立地路由到相应的输出端口.由于 p_j 的计算比较困难,与分组的阻塞情况有关,在确定阻塞概率时需要 $O((N!)^2)$[4]的复杂度,这很难在大型网络中实现,因此,本文假设到达的分组以独立且随机路由的方式选择路由平面.于是,所有的输出端口就具有相同的统计特性.因此,我们的分析以某个输出队列为讨论对象,记为标记性输出队列(tagged output queue).依据此构建如图 8 所示的分析模型.如果输出队列容量有限,则分组可能会丢失;否则,是不会出现丢包现象的.

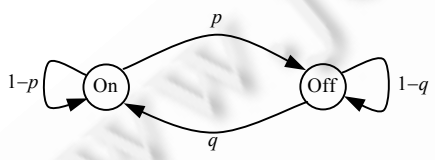


Fig.7 On-Off model

图 7 On-Off 模型

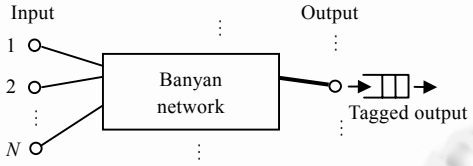


Fig.8 Analysis model of packet switching

图 8 分组交换的分析模型

3.2 无限容量的输出队列

设随机变量 A 表示在一个时隙内通过路由平面 j 到达标记输出队列的分组数量,根据前述,A 服从两项分布:

$$a_i=\Pr(A=i)=\binom{N}{i}\left(\frac{p}{mN}\right)^i\left(1-\frac{p}{mN}\right)^{N-i},\ i=0,1,...,N \tag{1}$$

当 $N\to\infty$时,公式(1)收敛于泊松分布:

$$a_i=\Pr(A=i)=\frac{(p/m)^i\mathrm{e}^{-p/m}}{i!},\ i=0,1,... \tag{2}$$

其概率生成函数是

$$A(z)=\sum_{i=0}^{N}z^i\Pr(A=i)=\left(1-\frac{p}{mN}+z\frac{p}{mN}\right)^N \tag{3}$$

由于分组的到达与离开并非完全与时钟同步,存在一定的传输时差,因此服务时间可以具有任意的分布,就可能失去无后效性的特性,致使不能从任意一个时间点来考察队列的变化.因此,设 Q_k^j 表示在第 k 个时隙结束的瞬间通过路由平面 j 到达标记输出队列的分组数量,A_k 在第 k 个时隙内到达分组数,则可得到关于 Q_k^j 的 Linkley 方程[21]:

$$Q_{k+1}^j=\max(0,Q_k^j+A_{k+1}-1) \tag{4}$$

其中,$Q_k^j>0, A_{k+1}>0$.如果当 $Q_k^j=0, A_{k+1}>0$ 时,有 $Q_{k+1}^j=A_{k+1}-1$.因为当 $Q_k^j=0$ 时,服务台在第 k 个时隙结束时成为闲置状态,当第 k+1 个分组到来时立即接受服务,服务台才被占用;到了第 k+1 个时隙结束时,到达的 A_{k+1} 个分组中有一个分组被送走了.由文献[21]可知,$[Q_k^j]$ 构成一个(隐)马尔可夫链,设此链在相邻时刻的转移概率为

$p_{ij}=\Pr[Q_{k+1}^j=j\mid Q_k^j=i]$.在确定 p_{00},即当 $Q_k^j=0$ 时,$Q_{k+1}^j=0$ 的概率有两种情况可能发生:一是在第 k+1 个时隙根本就没有包到达,即 A_{k+1}=0;二是在第 k+1 个时隙只有一个包到达,但被送走了,即 A_{k+1}=1.发生这两种情况的概率分别是 a_0 和 a_1,而且发生这两种情况是互斥的,因此有 $p_{00}=a_0+a_1$.其他情况与标准的 $M/G/1$ 一致,因此转移概率矩阵 $P=[p_{ij}](i,j=0,2,3,\ldots)$为

$$P=\begin{bmatrix} a_0+a_1 & a_2 & a_3 & a_4 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{5}$$

其中,$a_i=\Pr[A_{k+1}=i]$.设 $q_i=\lim_{n\to\infty}\Pr[Q_n^j=i]$,则可得相应的母函数为 $Q^j(z)=\sum_{i=0}^{\infty}q^iz^i$. 当输入强度$\lambda(=p/m)<1$ 时,队列存在极限分布[21].因此,在队列达到稳态时有方程$[q_0,q_1,q_2,\ldots]=[q_0,q_1,q_2,\ldots]P$,解这个方程,得到:

$$Q^j(z)=\frac{q_0a_0(z-1)}{z-A(z)} \tag{6}$$

对公式(6)的分子、分母分别微分,再取 z=1,得到 $q_0a_0=1-p/m$,代回到公式(6),得到稳态时队列长度分布的概率生成函数:

$$Q^j(z)=\frac{(1-p/m)(1-z)}{A(z)-z} \tag{7}$$

对公式(7)微分,并令 z=1,得到队列在稳态时长度的期望:

$$E(Q^j)=\frac{N-1}{N}\cdot\frac{(p/m)^2}{2(1-p/m)}=\frac{N-1}{N}\frac{p^2}{2m(m-p)} \tag{8}$$

根据 Little 公式[21],当系统达到稳定状态时,标记队列中分组的平均等待时间为

$$E(W^j)=\frac{(N-1)}{N}\frac{p}{2(m-p)} \tag{9}$$

上面分析的是在一个时隙内,通过平面 j 成功到达标记队列的情况.而在 Multi-$\log_2N$ 网络中有 m 个这样的平面,即在整个系统达到稳定状态时,有:

$$E(Q)=mE(Q^j)=\frac{(N-1)}{N}\frac{p^2}{2(m-p)} \tag{10}$$

$$E(W)=mE(W^j)=\frac{(N-1)}{N}\frac{mp}{2(m-p)} \tag{11}$$

在网络中有很多因素导致分组的丢失,如果这里只考虑队列因素,那么因为队列容量无限,请求服务的分组不会被丢失,迟早会被接受服务,所以分组丢失率 Pr[*packet loss*]=0.

3.3 有限容量的输出队列

假设输出队列的容量为 b(包括正被输出端口传送的分组,即队列中实际等待传送的分组有 b−1 个),则当队列达到饱和状态时,后来的分组会被丢失.其他条件同第 3.2 节,现在的 Linkley 方程[22]为

$$Q_{k+1}^j=\min\{\max(0,Q_k^j+A_{k+1}-1),b\} \tag{12}$$

相邻两状态的概率转移矩阵 $P=[p_{ij}](i,j=0,2,3,\ldots,b)$为

$$P=\begin{bmatrix} a_0+a_1 & a_2 & a_3 & \cdots & a_b & \sum_{k=b+1}^{N} a_k \\ a_0 & a_1 & a_2 & \cdots & a_{b-1} & \sum_{k=b}^{N} a_k \\ 0 & a_0 & a_1 & \cdots & a_{b-2} & \sum_{k=b-1}^{N} a_k \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_0 & \sum_{k=1}^{N} a_k \end{bmatrix} \tag{13}$$

在概率转移矩阵中,最后一列 $P[i,b]=\sum_{k=b-i+1}^{N} a_k, 0\leqslant i\leqslant b$ 表示 $\Pr[Q_{k+1}^{j}=b \mid Q_k^j=i]$.即标记队列在第 k 个时隙结束时有 i 个分组,在第 k+1 个时隙结束时有 b 个分组(饱和了).考虑到输出端口在第 k+1 个时隙内还要传送一个分组,则在第 k+1 个时隙内至少要到达 b–i+1 个分组.当然,最多只能到达 N 个分组.在每个时隙结束时,队列里最多能有 b 个分组,所以转移矩阵是一个(b+1)×(b+1)型矩阵.现在,队列的状态数是有限的,故存在极限分布.当队列达到稳态时,有方程:

$$[q_0,q_1,q_2,\ldots,q_b]=[q_0,q_1,q_2,\ldots,q_b]P \tag{14}$$

解方程组(14),可得:

$$\begin{cases} q_1=\dfrac{1-a_0-a_1}{a_0}q_0 \\ q_k=\dfrac{1-a_1}{a_0}q_{k-1}-\dfrac{1}{a_0}\sum_{i=2}^{k}a_i q_{k-i}, 2\leqslant k\leqslant b \end{cases} \tag{15}$$

根据归一化原则 $\sum_{i=0}^{b} q_i=1$,可求得 $q_0=1-\sum_{i=1}^{b} q_i$,可求得平均队长为 $E(Q^j)=\sum_{i=0}^{b} iq_i$.根据 Little 公式[21],可进一步求得分组在队列中的平均等待时间为 $E(W^j)=\dfrac{m}{p}E(Q^j)$.最后,根据公式(10)和公式(11),得到 Multi-log$_2N$ 的平均队长和平均等待时间.

当队列容纳的分组数已经达到了队列的最大容量,后来的分组只有被丢弃,在一个时隙内最多只可能到达 N 个分组.如果假设 A 是表示到达标记队列被丢失分组的随机变量,则分组的平均丢失率为

$$\begin{aligned} \Pr[packet\ loss]&=\frac{m}{p}E[A]=\frac{m}{p}[1\cdot a_{b+1}+2\cdot a_{b+2}+\ldots+(N-b)\cdot a_N] \\ &=\frac{m}{p}\sum_{k=b+1}^{N}(k-b)a_k=\frac{m}{p}\sum_{k=b+1}^{N}(k-b)\binom{N}{k}\left(\frac{p}{mN}\right)^k\left(1-\frac{p}{mN}\right)^{N-k} \end{aligned} \tag{16}$$

4 结论及未来的研究

VS 型 Multi-log$_2N$ 是一种无阻塞高速交换网络,在通信系统中,交换能力成为通信瓶颈时,这种结构作为光交换节点被应用到光网络中.因为 Mulit-log$_2N$ 的结构特点,决定了在设计 Multi-log$_2N$ 时需要考虑以下 3 个方面的问题:

(1) 路由问题.设计 Multi-log$_2N$ 的目的是消去内部阻塞,单个 log$_2N$ 网络具有自选路由,但由多个 log$_2N$ 组成的 Multi-log$_2N$ 不一定仍具有自选路由特性.实际上,如果没有相应的路由算法来控制分组在 Multi-log$_2N$ 的多个平面间的选择,无阻塞是难以保障的;只有分组进入某个平面内,自选路由才能发挥作用.因此,本文根据单平面内的冲突链路集的思想,提出了相应的路由算法来控制分组在多个平面间的路由选择,完全保障了 Multi-log$_2N$ 网络的无阻塞性;

(2) 内部缓存问题.消去了内部阻塞,那么就应该消去内部缓存,降低硬件成本;

(3) 性能问题.由于是多个平面参与分组的调度及交换,因此交换速度极大地提高,而且是线速的 N 倍.这决定了在 Multi-$\log_2 N$ 的输入端无需缓存,到达的分组可以立即得到服务;而分组在输出端口处却必须等待,因此在其输出端应有相应的缓存队列.

目前,有关交换网络的性能分析还只涉及单个平面(即 $\log_2 N$)网络,几乎没有对多平面(Multi-$\log_2 N$)网络的性能分析.因此,本文以嵌入式马尔可夫链为分析工具,提出相应的分析模型,并以此模型系统地分析了相应的性能指标,如平均队长、平均等待时间等.这些分析结论对用 Multi-$\log_2 N$ 作为光交换节点的体系结构是一个有益补充,或提供了相应的理论支持.

在本文研究中,我们假设分组是等概率地选择每一个路由平面,即文中 $p_j=1/m(1\leqslant j\leqslant m)$.如果除去等概率的假设,分组在路由平面选择时,尽可能地在同一平面路由多的分组,留下更多的空平面实现后来更多分组的无阻塞路由.即路由平面对分组存在一定的喜好程度.如第 2 节例子所示,在第 1 个平面路由了 5 个请求,在第 2 平面路由了 3 个请求,据此认为,第 1 个平面被选中的概率要大于第 2 个平面.如何更为合理地设置这个喜好程度是未来的研究工作.其次,Lea 等人[3]给出了 Multi-$\log_2 N$ 实现可重排无阻塞时所需的路由平面数,但这个结果只适用于单播.Tscha[5]的结果虽然适用多播,但是是实现严格非阻塞,其实现成本高于可重排.因此,Multi-$\log_2 N$ 网络实现可重排且适用于多播时所需的路由平面数还有待进一步研究.第三,本文提出的路由算法的时间复杂度为 $O(N^2)$,在未来的研究中去寻找时间复杂度更低的路由算法来支持 Multi-$\log_2 N$ 网络分组交换.最后,为了更加准确地评估多平面交换网络的交换性能,必须对网络流量有一个较准确的建模,因此,这也是以后研究工作中的一项重要内容.

致谢 感谢同行评审专家给出的修改意见,使文章的内容更加充实完善,同时也让作者受益匪浅,在此表示衷心的感谢.

References:

[1] Zhang XP, Liu ZH, Zhao YJ, Guan HT. Scalable router. Ruanjian Xuebao/Journal of Software, 2008,19(6):1452−1464 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/1452.htm [doi: 10.3724/SP.J.1001.2008.01452]

[2] Zheng SQ, Gumaste A, Shen H. A parallel self-routing rearrangeable nonblocking multi-$\log_2 N$ photonic switching network. IEEE/ACM Trans. on Networking, 2010,18(2):529−539. [doi: 10.1109/TNET.2009.2036173]

[3] Lea CT. Multi-$\log_2 N$ networks and their applications in high-speed electronic and photonic switching systems. IEEE Trans. on Communications, 1990,38(10):1740−1749. [doi: 10.1109/26.61445]

[4] Lea CT, Shyy DJ. Tradeoff of horizontal decomposition versus vertical stacking in rearrangeable nonblocking networks. IEEE Trans. on Communications, 1991,39(6):899−904. [doi: 10.1109/26.87179]

[5] Tscha Y, Lee KH. Yet another result on multi-$\log_2 N$ networks. IEEE Trans. on Communications, 1999,47(9):1425−1431. [doi: 10.1109/26.789678]

[6] Goke LR, Lipovski GJ. Banyan networks for partitioning multiprocessor systems. In: Proc. of the 1st Annual Symp. on Computer Architecture. 1973. 21−28. [doi: 10.1145/800123.803967]

[7] Patel JH. Performance of processor-memory interconnections for multiprocessors. IEEE Trans. on Computers, 1981,C-30(10):771−780. [doi: 10.1109/TC.1981.1675695]

[8] Kumar M, Jump JR. Performance of unbuffered shuffle-exchange networks. IEEE Trans. on Computers, 1986,C-35(6):573−577. [doi: 10.1109/TC.1986.5009435]

[9] Kruskal CP, Snir M. The performance of multistage interconnection networks for multiprocessors. IEEE Trans. on Computers, 1983,C-32(12):1091−1098. [doi: 10.1109/TC.1983.1676169]

[10] Theimer TH, Rathgeb EP, Huber MN. Performance analysis of buffered Banyan networks. IEEE Trans. on Communications, 1991, 39(2):269−277. [doi: 10.1109/26.76464]

[11] Turner JS. Queueing analysis of buffered switching networks. IEEE Trans. on Communications, 1993,41(2):412−420. [doi: 10.1109/26.216516]

[12] Jenq YC. Performance analysis of a packet switch based on single-buffered Banyan network. IEEE Journal on Selected Areas in Communications, 1983,SAC-1(6):1014−1021. [doi: 10.1109/JSAC.1983.1146023]

[13] Lin T, Kleinrock L. Performance analysis of finite-buffered multistage interconnection networks with a general traffic pattern. SIGMETRICS Performance Evaluation Review, 1991,19(1):68−78. [doi: 10.1145/107972.107973]

[14] Kabacinski W, Michalski M. The routing algorithm and wide-sense nonblocking conditions for multiplane baseline switching networks. IEEE Journal on Selected Areas in Communications, 2006,24(12):35−44. [doi: 10.1109/JSAC.2006.258221]

[15] Lu E, Zheng SQ. Parallel routing algorithms for nonblocking electronic and photonic switching networks. IEEE Trans. on Parallel and Distributed Systems, 2005,16(8):1−12. [doi: 10.1109/TPDS.2005.89]

[16] Wu CL, Feng TY. On a class of multistage interconnection networks. IEEE Trans. on Computers, 1980,C-29(8):694−702. [doi: 10.1109/TC.1980.1675651]

[17] Chang GJ, Huang FK, Tong LD. Characterizing bit permutation networks. Networks, 1999,33(4):261−267. [doi: 10.1002/(SICI)1097-0037(199907)33:4<261::AID-NET3>3.0.CO;2-Q]

[18] Zhao YJ, Luan GX, Guo J, Fu LZ. Design and throughput analyses of a new Banyan ATM switch with bypass queues. Journal of China Institute of Communications, 1999,20(1):42−47 (in Chinese with English abstract).

[19] Zhang B, Yang JH, Wu JP. Survey and analysis on the Internet traffic model. Ruanjian Xuebao/Journal of Software, 2011,22(1):115−131 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3950.htm [doi: 10.3724/SP.J.1001.2011. 03950]

[20] Willinger W, Taqqu MS, Sherman R, Wilson DV. Self-Similarity through high-variability: Statistical analysis of Ehernet LAN traffic at the source level. ACM SIGCOMM Computer Communication Review, 1995,25(4):100−113. [doi: 10.1145/217391.217418]

[21] Kleinrock L. Queueing Systems, Vol. I: Theory. New York: John Wiley & Sons, 1975. 275−277.

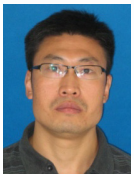
[22] Hluchyj MG, Karol MJ. Queueing in high-performance packet switching. IEEE Journal on Selected Areas in Communications, 1988,6(9):1587−1597. [doi: 10.1109/49.12886]

附中文参考文献:

[1] 张小平,刘振华,赵有健,关洪涛.可扩展路由器.软件学报,2008,19(6):1452−1464. http://www.jos.org.cn/1000-9825/19/1452.htm [doi: 10.3724/SP.J.1001.2008.01452]

[18] 赵有健,栾贵兴,郭景,付立政.一种新的基于旁路队列 Banyan 交换的 ATM 交换机结构的设计与性能分析.通信学报,1999,20(1):42−47.

[19] 张宾,杨家海,吴建平.Internet 流量模型分析与评述.软件学报,2011,22(1):115−131. http://www.jos.org.cn/1000-9825/3950.htm [doi: 10.3724/SP.J.1001.2011.03950]



刘晓锋(1972－),男,重庆人,博士生,副教授,CCF 会员,主要研究领域为计算机网络体系结构,路由与交换.
E-mail: xhxfliu@163.com

赵有健(1969－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机网络体系结构,路由与交换,计算机网络安全.
E-mail: zhaoyoujian@tsinghua.edu.cn

吴亚娟(1974－),女,博士,副教授,主要研究领域为基于网络的图像处理,数值计算.
E-mail: scwuyajuan@yahoo.com.cn