排序集线器多级互连交换结构的多路径自路由模型

李 辉[1], 何 伟[1], 伊 鹏[2], 王秉睿[1], 雷 凯[1], 安辉耀[1], 汪斌强[2]
(1. 北京大学深圳研究生院集成微系统重点实验室, 广东深圳 518055;
2. 信息工程大学国家数字交换系统工程技术研究中心, 河南郑州 450002)

摘 要: 目前已提出多种能提供 100% 吞吐率的分组交换结构, 如共享总线、共享内存、交叉矩阵及输入输出排队等. 它们的结构性缺陷是存在某个瓶颈限制了其规模的有效扩展, 如带宽瓶颈、调度算法运算处理瓶颈等. 本研究提出了一类新的结合群组排序集线器和多级互连网络的多路径自路由交换结构, 并证明了该类结构构建于代数群论的自路由数学模型. 该结构具有: 完全分布式自路由、无需端口匹配调度、无内部缓存、无缓存时延及无抖动、按位置换群建模及可递归扩展和模块化属性. 理论分析及仿真结果表明该结构适合作为提供 QoS 保证的超大规模宽带交换结构.

关键词: 双调; 集线器; 多级互连网络; 自路由; 交换结构
中图分类号: TP393, TN256 文献标识码: A 文章编号: 0372-2112 (2008) 01-0001-08

Modeling Multi-path Self-routing Switching Structure from Multistage Interconnection of Sorting Concentrators

LI Hui[1], HE Wei[1], YI Peng[2], WANG Bing-rui[1], LEI Kai[1], AN Hui-yao[1], WANG Bin-qiang[2]
(1. *Key Laboratory of Integrated Microsystems, Shenzhen Graduate School, Peking University, Shenzhen, Guangdong* 518055, *China*;
2. *NDSC, Information Engineering University, Zhengzhou, Henan* 450002, *China*)

Abstract: Various 100% throughput packet switching structures have been proposed for broadband network, such as Shared Bus, Shared Memory, Crossbar Matrix with Combined Input and Output Queuing, etc. Topologically speaking, their major demerit, such as bandwidth bottleneck and insufficient processing ability to schedule I/O matching, greatly limits their scalability for large scale switching routers. This paper proposes and models a novel multi-path self-routing switching fabric by integrating bitonic sorters and the multistage interconnection networks. This kind of structure possesses the properties of complete distributing and self-routing, free of I/O matching scheduling algorithm, no internal buffer, no buffered delay and jitter, modeled with algebraic permuting group, as well as high modularity and recursive scalability. Mathematical analysis and simulations show this structure is suitable for building super large scale switching fabric with QoS guaranteed application.

Key words: bitonic; concentrator; multistage interconnection network; self-routing; switching fabric

1 引言

网络系统由传输线及交换路由器组成, 网络的传输能力由传输线带宽和交换路由器带宽共同决定. 目前光纤传输技术使得单根光纤可以长距离传输每秒若干太位(1Tera= 1000Gega)以上的信息, 光的传输能力未来仍将按超过半导体行业中摩尔定理的速度增长. 相比之下, 目前规模最大的交换路由系统是 Cisco 公司的 CRS-1 系列, 其单机架的交换容量就是每秒若干太位, 多机架结构才能组成每秒几十太位的交换容量. 可见, 交换路由器的带宽及其发展速度远落后于传输带宽的需要.

从组成结构上看, 网络的路由交换设备可以分成两级: 即组成长途骨干网的高端交换路由器及城域网本地环路的接入交换路由器. $N \times N$ (N 输入, N 输出) 高端交换路由器的特点是大城市之间的互连交换, 端口数通常较少 ($N \leqslant 100$), 但每个端口的带宽很大, 如 OC768 (40Giga bps) 以上. 城域网接入交换路由器的功能是完成终端用户的数据交换, 其特点是用户数量巨大, 要求能提供万门 ($N \geqslant 10000$) 以上端口的交换能力, 而每个端口的带宽通常较小, 如 OC1(50Mbps); 从网络配置和使用数量上看, 城域网交换路由器的需求量远超过高端交换路由器.

当前网络的主要瓶颈是最后英里接入的带宽, 故 ITU-T 制定的 5 公里 ADSL 传输带宽从原来的 1.5Mbps

收稿日期: 2007-01-23 修回日期: 2007-12-03
基金项目: 国家自然科学基金(No. 60572042); 国家 863 高技术研究发展计划(No. 2007AA01Z218); 国家支撑计划(No. 2006BAH02A10); 广东省自然科学基金(NSFGD2007No. 295)

增长到 2003 年 ADSL2 + 标准的 25Mbps. 目前长途骨干网光纤的传输带宽利用率是较低的,但是今天终端用户仍经常遇到达不到 ADSL2 + 标称下行带宽 10% 的下载速度. 为什么? 从整个通信系统看,现在只剩下两个因素:一是提供数据的服务器处理容量及带宽不够,另外就是宽带接入交换机的交换能力不足. 而从网络系统来看,城域网交换路由器的交换容量不足正是问题所在. 当万门以上的城域网交换路由器进入网络时,当今网络本身的瓶颈可以消除. 本文的研究就是为该类交换路由器提供底层交换结构及模型.

宽带交换技术在物理层存在电信号和光信号之分,光交换在近 10 多年取得快速发展,光器件与多级互连网络结构的结合是近年的热点之一,但光的存储及逻辑处理能力仍远未达到实用,这一点在技术上制约了全光交换在大规模 $N \times N$(N 输入,N 输出,$N \geqslant 10000$)交换结构上的应用,故电信号交换在最近若干年内才是大规模交换结构的可行技术.

在分组交换路由技术的发展上,交叉矩阵(crossbar)在 N 较小时是一类实现无阻塞及自路由的理想交换结构. 它每个分组的交换过程是:先由调度器对活跃输入端口进行避免外部输出端口争用的配对,决定活跃输入端口下一个时隙输出分组的输出端口,然后再分组传输.

该结构本身的主要问题是:

(1)其需要的交换单元的数量是 N^2,两阶增长 $O(N^2)$的硬件实现复杂性,当 N 较大时,其成本增加变得不能接受. 因为从信息论熵的角度看,设由 M 个 2×2 基本交换元件构成的网络就可以提供所有 $N!$ 种输入输出一一对应的交换连接映射,M 个 2×2 基本交换元件构成的所有可能连接模式的数目 2^M 必须大于等于 $N!$,即 $2^M \geqslant N!$ 当 N 为较大的数时,其渐近线主导项为 $N\log_2 N$. 故由 2×2 交换单元构造 $N \times N$ 交换网络复杂度的信息论下界是 $O(N\log_2 N)$. 因此如何得到复杂度在 $O(N\log_2 N)$和 $O(N^2)$之间的最优结构也是交换理论研究的基本问题之一.

(2)其 N 行 N 列的拓扑结构决定其可能的最长路径是$(2N-1)$个交叉节点及$(2N-1)$段连线,而最短路径只是 1 个交叉节点及其连线,最大差 $2(N-1)$. 分组长度通常在 500 到 1000 比特,按 200Mbps 数据率计算,1000 比特分组的时间长度就是 5μs,即有效载荷时隙 $Ts = 5\mu s$;信号经每个交叉点的时延按最少一级门的延时 $\tau_d = 1ns$ 算(在开关速度为 1GHz),分组到达的最大时间差 $\delta = 2(N-1)\tau_d \approx 2N\tau_d = 2 \times 10000 \times 1ns = 20us = 4Ts$. 故为了同步传送一个时隙的 N 个分组,必须给予 5 个时隙的时间,开销占了 80%.

(3)仅靠交叉矩阵结构还不能实现无阻塞交换,它必须与一定的排队调度结构结合. 输出排队[19](Output Queuing)交换结构能够为业务流提供 100% 吞吐量、速率以及时延等多方面的服务质量 QoS(Quality of Service)保障,然而它需要核心交换结构的加速比达到 N,故 N 较大时是很难实现的. 比较而言,输入队列结构[20](IQ: Input Queuing)的交换单元和存储单元均只需工作于线路速率,因而对于构建大容量交换结构是一种十分经济的解决方案. 但 IQ 结构由于队头阻塞(HoL: Head of Line Blocking)问题,吞吐率只能达到 58.6%;另外 IQ 结构需要采用集中控制的调度算法,这使得在 IQ 结构实现服务质量保障十分复杂,很难具有现实意义. 因此提出了联合输入输出排队结构(CIOQ[7,21,22]),并证明了当加速比为 2 时,能使吞吐率达到 100%. 但 CIOQ 仍然需要使用基于稳定婚姻或极大匹配等调度算法,其时间复杂度都是 $O(N^2)$或以上,虽然有若干改进[13~17]但瓶颈仍然存在. 事实上,因为调度是每个分组时隙都必须执行一次的,且必须在一个或若干时隙内完成. 为了构造可扩展的万门以上的交换结构,任何 $O(N)$及以上的调度时间复杂性都不可行,因为分组的时隙长度不可能随 N 线性增长. 理论和实际上很早就成熟的可重排无阻塞 Benes[11]和严格无阻塞 Clos[12]结构由于相同原因不能用于大规模分组交换结构.

上述不足使得带调度的交叉矩阵结构也不具备线性扩展的可能性,因此如何避免每时隙的调度是关键所在. 不同类型的数据对于带宽,丢失率,时延及其抖动的 QoS 要求是不同的[23]. 用户数据如电子邮件等对实时性要求很低,对丢失率的要求高,丢失的分组必须重传;而最大量的流量是音视频,实时流媒体对实时性要求高,但在一定丢失率下可以不影响服务质量不必也不可能重送,如不压缩的语音在 10^{-3}丢失率,视频在 10^{-4}丢失率都是可接受的. 另外,传输线路本身也有丢失率,如双绞线约为 10^{-5}到 10^{-6},电缆约为 10^{-6}到 10^{-7},光纤约为 10^{-8}到 10^{-9}. 换句话说,交换结构不必追求 100% 吞吐率或无阻塞率,事实上,只要保证丢失率在 QoS 规定的范围内即可,而丢失或传错的数据必要时可由上层协议进行重发处理,或者忽略.

根据上述思路,本研究提出的是满足一定丢失率,不需要每时隙调度的大规模 $N \times N$ 交换结构. 交换结构实际上是一个两级交换结构,设 $M = 2^m$, $G = 2^g$, $N = M \times G = 2^m \times 2^g = 2^n$; $n = m + g$;把 G 个输出端口一起构成一个输出群组,故大规模的 $N \times N$ 交换结构只有 M 个输出群组;分组不必进行调度而是先交换到各自所属的输出群组. 交换到同一个输出群组的分组再经小规模的 $G \times G$(如 $G \leqslant 128$)进行第二次交换,送到各自的目的地址,该小规模结构可采用各种成熟技术. 因此该结构中,一个时隙中送到某个输出群组中属于同

一个目标端口的分组最多可以有 G 个，G 的大小表示了该结构多路径的统计复用能力，及吸收突发性流量负载的能力. 当然，当一个时隙中有超过 G 个分组输出到同一个输出群组时将会引起丢失. 该结构的中央控制部分只要完成很简单的功能，即在每个连接建立阶段，保证 $\sum_{i=1}^{N}\lambda_{ij}\leqslant 1$ 就可以了，λ_{ij} 为第 i 个输入端口到第 j 个输出端口的归一化负载.

本文第 2 节主要介绍多级互连网络及其自路由机制. 第 3 节介绍了排序集线器的递归构造方法. 第 4 节，将排序集线器与多级互连网络相结合构造了“排序集线器多级互连交换结构的多路径自路由模型”，并给出了阻塞率的计算方法. 第 5 节对该结构进行了仿真，分析其阻塞率与负载及交换结构参数的关系；最后，对全文进行了简单总结及对进一步研究的展望.

2　多级互连网络模型

$N\times N$ 多级互连网络(MIN: multistage interconnection network，)[1,4,5,18,24]，当 $N=2^n$ 时，由 n 级每级 $N/2$ 个 2×2 的基本自路由交换单元构成. MIN 由于其并行处理的特性及结构上可以大规模扩展，使得它在并行处理和分组交换[2,3]等领域一直吸引众多研究者. 下面我们定义其基本单元，级间交换及自路由模型.

2.1　基本自路由交换单元

定义 1　2×2 基本自路由交换单元是一个具有 2 输入和 2 输出端口(分别称为 0/1 端)的逻辑电路.

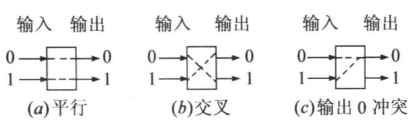


图 1　基本路由元件

参见图 1，当输入端口数据分组未争用某输出端口时，它有两种基本连接状态，即平行/交叉(Bar/Cross). 当有争用时，路由元件将随机选择其中一个输入分组作为胜者，并送到其希望的输出端口，败者的数据将被丢弃或误路由.

基本自路由元件可用带内信令来实现交换(见表 1)，如在数据分组的前加上两位带内信令. 第一位 A_1 活跃比特表示当前时隙是否有数据分组，1 表示有活跃的有效数据分组，0 表示当前没有有效数据的空分组. 可以推广到后面介绍的排序集线器或其它自路由元件.

表 1　基本路由单元的路由控制机制

连接状态		输入 1 带内信令		
		‘10’	‘00’	‘11’
输入 0 带内信令	‘10’	任意	平行	平行
	‘00’	交叉	任意	平行
	‘11’	交叉	交叉	任意

2.2　级间位置换交换模型

考虑作用于整数 1 至 n 上的置换群. 类似于群论中常用的循环标记，例如用(321)表示置换 σ，则有 $\sigma(1)=3,\sigma(2)=1,\sigma(3)=2$ 且 $\sigma(k)=k$，对所有 $k>3$. 该置换群中的单位元素即使整数 1 至 n 均映射到自身的映射，记为 id. 群论中的乘法是自左向右的复合函数. 因此 $\sigma_1\sigma_2(k)$ 定义为 $\sigma_2(\sigma_1(k))$，例如乘积(23)(123)等于(21). 故置换 $\sigma=(321)$ 导出如下 3 位二进制串中一一对应的映射：

$$X_\sigma: b_3b_1b_2\rightarrow b_1b_2b_3 \tag{1}$$

注意，这里源端口地址 $S_1S_2S_3=b_3b_1b_2$，目的地址 $d_1d_2d_3=b_1b_2b_3$；一般的，对于 n 位二进制串，定义于整数 1 至 n 上的置换 σ 可导出如下的映射[10,25]：

$$X_\sigma: b_{\sigma(1)}b_{\sigma(2)}\cdots b_{\sigma(n)}\rightarrow b_1b_2\cdots b_n \tag{2}$$

$$X_\sigma^{\wedge}: b_1b_2\cdots b_n\rightarrow b_{\sigma^{-1}(1)}b_{\sigma^{-1}(2)}\cdots b_{\sigma^{-1}(n)} \tag{3}$$

然而并非所有的位置换 σ 都可用于级间交换连接. 因此下面我们将介绍“级间位置换交换”[6]这一概念.

定义 2　级间位置换交换是满足如下条件的 n 位二进制串 $s_1s_2\cdots s_n$ 到 $d_1d_2\cdots d_n$ 的映射，X_σ. 若两个串在映射前仅最右一位地址(最低位)不同，则映射后在左边的高 $(n-1)$ 位中至少有一位不同，其集合记为 X.

简言之，同一路由元件的两个输出端口必须通过级间位置换交换送到下一级的不同路由元件的输入端，即 $(\sigma(n)\neq n)$ 否则该级间位置换连线没有实际交换意义.

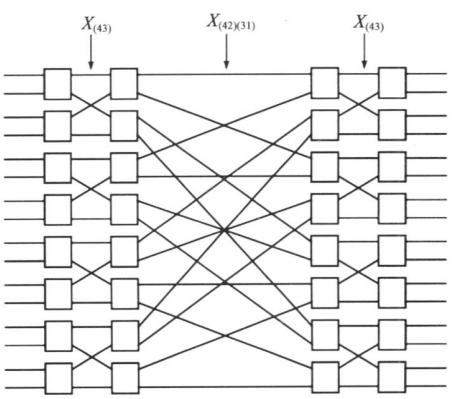


图 2　榕树型网络例子

2.3　多级互连网络以级间位置换交换描述的模型

定义 3　一个 $N\times N(N=2^n)$ 多级互连网络是由 n 级每级 $N/2$ 个 2×2 交换路由单元以输入级位置换 σ_0，$(n-1)$ 级级间位置换交换 $\sigma_1,\sigma_2,\cdots\sigma_{(n-1)}$ 及输出级位置换 σ_n 构成，可以表示为 $[\sigma_0:\sigma_1:\sigma_2:\cdots:\sigma_{(n-1)}:\sigma_n]$，其中冒号“:”表示一级 2×2 单元. 如其满足任何一对输入输出端口之间均有一条路由通路，则称其为榕树型网

络.

定义 3 描述了一类 MIN 网络,有很多是不常见的.常见的 $2^n \times 2^n$ 榕树网 *Banyan* = [*id*:(*n* 1):(*n* 2):⋯:(*n n* − 1):*id*],混洗网络 *Shuffle* = [*id*:(*n n* − 1 ⋯ 1):(*n n* − 1 ⋯ 1):⋯:(*n n*-1 ⋯ 1):*id*].图 2,16 × 16 分治网络 *N*2 = [*id*:(43):(42):(31):(43):*id*]是一类同等规模中模块化最好和版图复杂度最小的网络,我们称其为分治网络,其它规模的分治网络如:

64 × 64:[*id*:(65):(654):(63)(52)(41):(65):(654):*id*];

256 × 256:[*id* :(87) :(86)(75) :(87) :(85)(73)(62)(51) :(87) :(86)(75) :(87) :*id*];

1024 × 1024:[*id* :(10 9):(10 9 8) :(6 9 7 10 8):(10 9) :(10 5)(9 4)(8 3)(7 2) (6 1):(10 9):(10 9 8) :(6 9 7 10 8):(10 9) :*id*];

定义 4 定义 3 的 $N \times N(N = 2^n)$ *n* 级网络[σ_0:σ_1:σ_2:⋯:$\sigma_{(n-1)}$:σ_n],其级间位置换决定了如下两个 1 到 *n* 的 *n* 个数的某种序列,*T* 序列和 *R* 序列.

$$T_k = (\sigma_0\sigma_1\cdots\sigma_{k-1})^{-1}(n) \quad 1 \leqslant k \leqslant n; 1 \leqslant T_k \leqslant n;$$

$$R_k = (\sigma_k\sigma_{k+1}\cdots\sigma_n)(n) \quad 1 \leqslant k \leqslant n; 1 \leqslant R_k \leqslant n;$$

并不是所有描述成[σ_0:σ_1:σ_2:⋯:$\sigma_{(n-1)}$:σ_n]的网络都是榕树型网络,下面定理给出了其是否属于榕树型可路由网络的判定标准,及描述了其自路由的过程.

定理 1 定义 3 的 $N \times N(N = 2^n)$ *n* 级网络[σ_0:σ_1:σ_2:⋯:$\sigma_{(n-1)}$:σ_n]是榕树型网络的充分必要条件是其按定义 4 的路由序列 $R_1, R_2, \cdots R_n$, *n* 个数各不相同,即 1,2,⋯ *n* 均出现一且只有一次.

$$R_k = (\sigma_k\sigma_{k+1}\cdots\sigma_n)(n) \quad 1 \leqslant k \leqslant n; 1 \leqslant R_k \leqslant n;$$

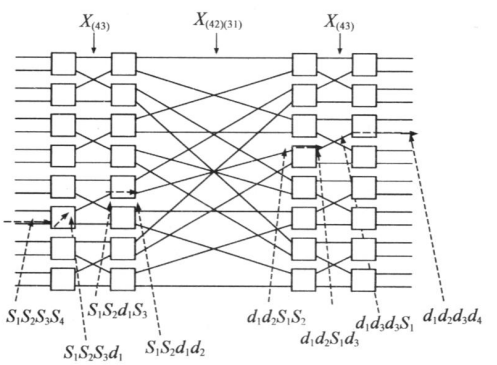


图 3 自路由过程例子

定理 2 当定义 3 的 $N \times N(N = 2^n)$ *n* 级网络[σ_0:σ_1:σ_2:⋯:$\sigma_{(n-1)}$:σ_n]是榕树型网络时,从任何 *n* 位二进制地址 $S_1S_2\cdots S_n$ 输入端口到任何 *n* 位二进制地址 $d_1d_2\cdots d_n$ 输出目标端口,其经 *n* 级 2 × 2 单元的逐级自路由地址序列是 $d_{R1}, d_{R2}, \cdots d_{Rn}$,即目标地址的第 R_k 位 d_{R_k} 用于第 *k* 级单元的自路由.

证明 由定理 1 及定义 4 关于 R_k 序列的定义可得.

表 2 例子列出了一些常见及不常见的网络的 T_k, R_k 序列及其自路由的过程.

表 2 网络的 T_k, R_k 序列及其自路由的过程

网络	T_k	R_k	自路由地址序列
*N*10 = [*id*:(432) :(42)(31) :(43) :(234)]	4,2,3,1	1,3,4,2	$d_1d_3d_4d_2$
[(321):(31):(32):*id*]	1,2,3	1,2,3	$d_1d_2d_3$
*N*11 = [*id*:(43):(42)(31):(43):*id*]	4,3,2,1	1,2,3,4	$d_1d_2d_3d_4$
[*id*:(43):(432):(4321):*id*]	4,3,2,1	1,2,3,4	$d_1d_2d_3d_4$

图 3 表示出对表中网络 *N*11 从源地址 $S_1S_2S_3S_4$ = 1011 到目标地址 $d_1d_2d_3d_4$ = 0101 按定理 2 定义的自路由过程.

本节自路由模型的讨论是为了与下面的二元排序集线器结合构造超大规模的群组多路径自路由模型.在本构造中,$N \times N$ 自排序器本身是一种自路由结构,下面介绍相应定义.

3 自路由排序器和集线器构造

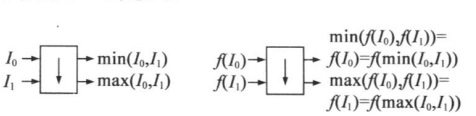
集线器在通信系统中大量使用,本研究中关注的是具有自路由属性的集线器.自路由集线器以排序器[9]按一定结构递归构造.当定义好一定的方向或顺序后,能比较两个数大小的比较器就是排序器.

3.1 比较器和排序器网络

定义 5 一个 2 × 2 比较器单元把两个输入端各自的输入数值进行比较,把其中小的输出到定义为地址 0 的上端口,而大的数输出到定义为地址 1 的下端口.定义了从低地址到高地址的顺序后,该比较器单元就成为排序器单元.如图 4(*a*)示.(信号按大小自路由排序后送输出端口,数据按箭头方向从小到大排列)

一个 $K \times K$ 比较器网络是由若干级 2 × 2 比较器单元构成,它具有顺序继承的属性如下.

比较器网络保序性质:假设一个比较器网络把输入序列 $I = \langle I_0, I_1, \cdots I_{k-1}\rangle$ 转换为输出序列 $O = \langle O_0, O_1, \cdots O_{k-1}\rangle$;那么对任何一个单调增的函数 f,网络把输入序列 $f(I) = \langle f(I_0), f(I_1), \cdots f(I_{k-1})\rangle$ 变成输出序列 $f(O) = \langle f(O_0), f(O_1), \cdots f(O_{k-1})\rangle$(图 4(*b*)进行了简单的证明).介绍上述比较器网络的保序性质,是为了证明下面排序网络的 0 − 1 定理.



(*a*) (*b*)

图 4

定理 3(0-1 定理) 若一个 *N* 输入的排序网络能对 0,1 组成的任意序列(共 2^N 种情况)准确排序,则该

网络也能对任意数字组成的输入序列准确排序. 证明见文献[8].

0-1 定理的存在使得在构造和证明任意排序网络时,只需要考虑由数值 0 和 1 组合的二元输入序列就可以了,便于用枚举法. 有多种方法构造排序器,本研究采用便于递归构造的双调排序器 Batcher 网络结构.

3.2　基于双调排序器的排序网络及其构造

定义 6　一个 2×2 双调排序单元(bitonic sorter)是能根据一位的地址信息,把两个输入端的信号按大小自路由排序后送输出端口的组合逻辑电路.

如图 5 所示,输出端数据按箭头方向从小到大排列,I_0, I_1 是一位二进制地址,上(下)输出端口为地址 0(1). $O_0 = \min(I_0, I_1)$, $O_1 = \max(I_0, I_1)$.

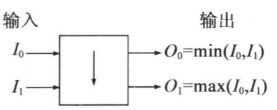


图 5　2×2双调排序单元

定义 7　一个双调(Bitonic)序列是一个仅由若干个 0 和 1 组成的序列. 它们的排列顺序可能是先递增后递减,或者先递减后递增,也可以是单调递增或者单调递减. 例如:形如 0…01…10…0 或者 1…10…0111, 0…01…1, 1…10…0 的序列都是双调序列.

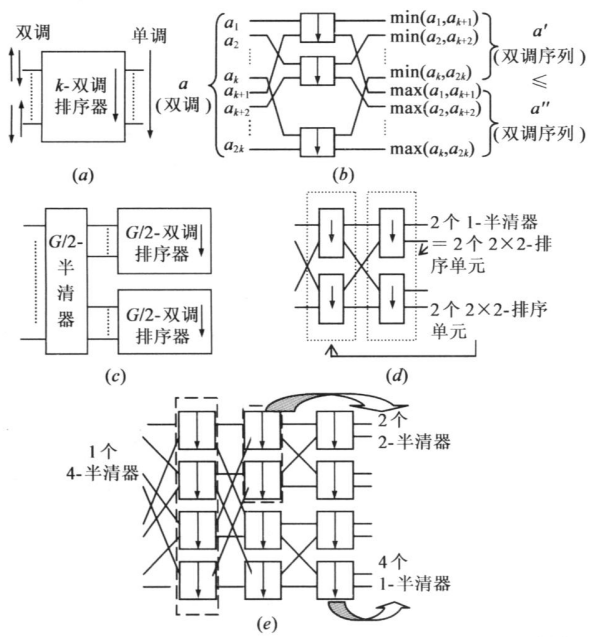


图 6　(a)k-双调排序器;(b) k-半清器结构;(c) G×G 双调排序器递归构造;(d) 4×4-双调排序器的构造;(e) 8-双调排序器的构造

定义 8　一个 k-双调排序器是能将一个长度为 k 的双调序列排序为单调序列的网络. 如图 6(a)所示,输出端数据按箭头方向从小到大排列. 即是能把输入为单调增接单调减,或单调减接单调增,总长度为 k 的双调序列排序成长度为 k 单调增线性 0…01…1 序列.

定理 4　一个 k-半清器是这样一级网络,它能将一个长度为 2k 的双调序列 a,拆分为 2 个双调序列 a′, a″,并且能保证 a′ ≤ a″. k-半清器具体的构造方法如图 6(b)所示.

运算符"≤"定义如下:对于两个等长双调序列,a_1 和 a_2,如果 a_1 序列中的每个元素均小于等于 a_2 中的元素,则有 $a_1 \leqslant a_2$. 注意并非所有的双调序列之间都存在这种关系. 只有两个待比较序列中有一个是全 0 或全 1 时,才存在此关系. 如 000000≤001100, 111100≤111111.

推论 1　任意两个 0,1 组成的序列,本身均构成一个长度为 2 的双调序列,故 $K = 1$ 的半清器就是一个定义 5 中规定的 2×2 双调排序单元(bitonic sorter),其输出已经是线性排序.

$G \times G(G = 2^g)$双调序列排序器递归构造

根据上面定义及半清器结构,可用如下方法递归构造输入规模为 $G = 2^g$ 双调序列排序器.

第 1 级是 1 个 $k = G/2$ 的半清器,其输出是 2 个 $G/2 = 2^{(g-1)}$的双调序列,其中有一个已经是完全相同的 0 或 1 序列;

第 2 级是 2 个 $(G/2) \times (G/2)$ 双调序列排序器. 而每个 $(G/2) \times (G/2)$双调排序器的第 1 级是一个 $k = G/4$ 的半清器,其输出是 2 个 $G/4 = 2^{(g-2)}$的双调序列,其中有一个已经是完全相同的 0 或 1 序列,故第 2 级共输出 4 个 $G/4$ 双调序列;

第 3 级 4 个 $(G/4) \times (G/4)$双调序列排序器;…如此递归.

第 g 级是 $2^{(g-1)}$个 2×2 双调序列排序器,即是 $k = 1$ 的半清器.

故所以共用了 g 级规模逐级减半的半清器,每级的半清器的数量是按倍递增.

如图 6(c)是一般递归构造结构,(d)具体显示了一个 $G = 4$ 的双调排序器可以通过一个 2-半清器和两个 2×2 排序单元构造出来.(e)表示了 $G = 8$ 双调排序器由 1 个 4-半清器和 2 个 $G = 4$ 双调排序器构成.

3.3　自路由集线器构造方法

定义 9　一个 2G-to-G 集线器是指一个 $2G \times 2G$ 的排序交换模块,它将 2G 个输入信号中最大的 G 个路由到具有最大输出地址的 G 个输出端口,并将其余的 G 个路由到具有最小输出地址的 G 个输出端口.

定理 5　一个 2G-to-G 集线器的结构,就是用 2 个 $G \times G$ 任意 0-1 二元序列排序器网络后接一级 $k = G$ 半清器.

自然地,可以把 G 个最大排序输出端口看作"1-群组输出",其余 n 个小的排序输出端口作为"0-群组输出". 由于在自路由交换结构中,每个时隙每个输入端口可能空闲没数据,也可能有数据发到"1-群组输出"或"0-群组输出",故至少有三种情况,如表 1 示.

尽管前面一直讨论的是对 0-1 二元序列的排序器.但根据‘0-1’定理可知,任何能对 0-1 二元序列准确排序的排序器也能对由任意数字组成的序列排序.因此,上面构造的排序器可以对上述三种输入分组状态的两位信息排序,排序大小按 1-输出群组,空闲,0-输出群组.图 7(a)(b)就是 $G=4$ 的 2G-to-G 自路由群组集线器实例.

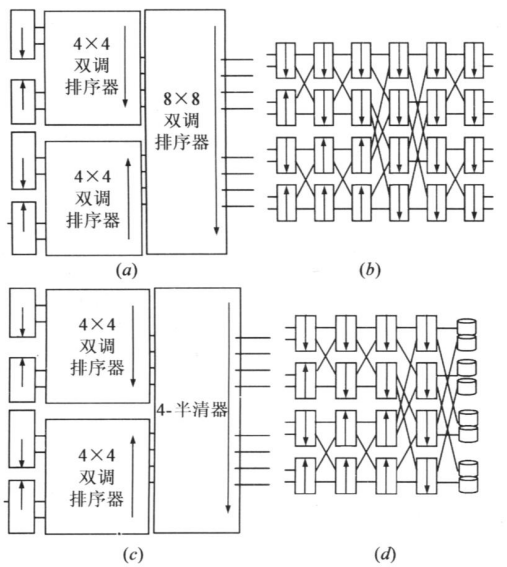


图 7 (a)建立 G=8 任意排序网络;(b)(a)的内部结构;(c)8-to-4 集线器;(d) (c)的内部结构

定义 10 一个 2G-to-G 群组集线器是指一个 $2G\times 2G$ 的自路由排序网络,其输出分成两个规模为 G 的群组,即上面 G 个端口构成的 0-输出群组 $G0$,和下面 G 个端口构成的 1-输出群组 $G1$.它能将 $2G$ 个输入信号中最多 G 个分组地址为 1(或 0)的信号自路由到 $G1$(或 $G0$)输出群组端口,而不管具体信号在 $G1$(或 $G0$)群组端口中的具体位置.

因此,要构造一个 2G-to-G 的集线器,就可以用图 7(c)(d)所示的方法,递归构造任意规模 G 群组集线器.

4 排序集线器多级互连交换结构的群组多路径自路由模型

图 8(a)显示了在图 2 基础上构建的 $N=128, M=16, G=8$,将排序集线器与多级互连网络相结合构造了“排序集线器多级互连交换结构的群组多路径自路由模型”.

一般的,设 $N=2^n, N=M\times G, M=2^m, G=2^g$,先构造一个 $M\times M$ 的可路由榕树类网络.然后将网络中各级 2×2 路由单元替换为 2G-to-G 自路由群组集线器.这样就建立了一个拥有 M 个输出群组,每群组包含 G 个输出端口的 $N\times N$ 网络.

下面分析该结构在不同负载下的阻塞率,为了便于分析,假定到达同一输出群组内的分组相互独立.设随机变量 X_j(j 从 1 到 m)代表在某一时隙,到达第 j 级群组集线器某输出群组的分组数目.显然 X_j 的上界是 G.

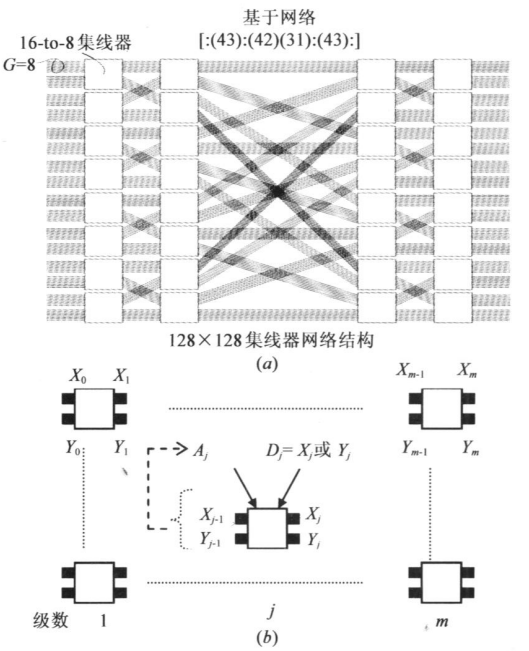


图 8 (a)$N=128, M=16, G=8$,多路径自路由交换结构;(b)阻塞率分析

定理 6 X_m 的分布可以通过如下的递推方法得到:设 X_0 是一个服从二项分布 $B(G,p)$的随机变量.实际上,X_0 反映了交换结构第 1 级集线器输入的负载情况.对 $j\geqslant 1$,设随机变量 Y_j 和 X_j 独立同分布(i.i.d),并分别对应于集线器的 1-输出群组和 0-输出群组.而随机变量 D_j 服从二项分布 $B(X_{j-1}+Y_{j-1},1/2)$,它表示在第 j 级集线器,当输入端有 $X_{j-1}+Y_{j-1}$ 个分组的情况下,分组数据到达集线器指定输出群组的数目.如图 8(b)所示.那么,对于所有 $z\leqslant G$,有 $\Pr(X_j=z)=\Pr(D_j=z)$.即:X_j 服从二项分布 $B(X_{j-1}+Y_{j-1},1/2)$ $(0\leqslant X_j\leqslant G)$.显然,分布 $B(X_{j-1}+Y_{j-1},1/2)$ 可根据 $X_{j-1}+Y_{j-1}$ 的值加以计算.

假定随机变量 A_j 代表进入第 j 级某 2G-to-G 群组集线器的分组数目($1\leqslant j\leqslant m$),则:

$$\Pr(A_j=t)=A_j(t)=\sum_{x+y=t}X_{j-1}(x)Y_{j-1}(y),\quad 0\leqslant x,y\leqslant G, 1\leqslant j\leqslant m \tag{4}$$

即共有 t 个分组到达第 j 级某群组集线器的输入端,如图 8(b)所示.从而这些分组中有 z 个到达集线器某输出群组(可能是 X_j 所在输出组,也可为 Y_j 所在输出组,两者等价)的概率可以计算如下:

$$\Pr(D_j=z)=\begin{cases}\sum_{t=z}^{2G}A_j(t)2^{-t}\binom{t}{z}, & \text{若 } z<G\\ \sum_{t=z}^{2G}\left(A_j(t)2^{-t}\sum_{k=G}^{t}\binom{t}{k}\right), & \text{若 } z=G\end{cases} \tag{5}$$

从交换结构的第 1 级直到最后一级重复上述的计算过程. 由 $\Pr(D_j = z)$的定义可知:若要求有 z 个分组顺利到达集线器某输出组,本级集线器输入端至少应有 z 个分组到达. 对于 $z < G$,若当前有 $t(z \leqslant t \leqslant 2G)$个分组到达,可以用二项分布 $B(t, 0.5)$并综合所有可能的输入情况来计算 $\Pr(D_j = z)$. 对于 $z = G$,这时 $G \leqslant t \leqslant 2G$,至少有 G 个分组竞争同一输出组. 此时,必须同时考虑可能有 G 个,或 $G+1$ 个,…一直到 t 个分组争用该输出群组的情况. 这样,总体的分组阻塞率可由定理 7 给出:

定理 7　设 $M = 2^m$, $N = M \times G$,则该交换结构 $N \times N$ 的分组阻塞率为 $1 - (E[X_m]/(G \times p))$即:

$$\Pr(blocking) = 1 - \left(\sum_{Z=0}^{G} z \cdot X_m(z)\right)/(G \cdot p) \quad (6)$$

其中 $X_0(z) = B(G, p)$,通过式(4),(5)计算 $X_1(k)$

$$A_1(t) = \sum_{x+y=t} X_0(x)\, Y_0(y), 0 \leqslant x, y \leqslant G$$

$$X_1(z) = \Pr(D_1 = z) = \begin{cases} \sum_{t=z}^{2G} A_1(t) 2^{-t} \binom{t}{z}, & \text{若 } z < G \\ \sum_{t=G}^{2G} \left(A_1(t) 2^{-t} \sum_{k=G}^{t} \binom{t}{k}\right), & \text{若 } z = G \end{cases}$$

继续迭代,就可计算 $X_2(z), \cdots, X_m(z)$.

5　性能仿真与分析

本节对上面介绍的交换结构在不同负载下的阻塞率用 Matlab 软件进行了相关仿真,仿真主要根据(4)(5)(6)的递推公式,逐级计算分组在各级的阻塞率并将当前级的阻塞率记录下来作为下一级输入流量的信息,以进一步计算. 结论如下:由图 9(a)可以看出,当网络的级数 $m = 16$ 时(此时对应的 $M = 2^{16}$), G 越大,对于指定的阻塞率来说,网络每个输入口可承受的负载 p 也越大. 从目前的芯片制造和封装技术看, G 为 64, 128, 256 都是可行的. 图 9(b)可以看出,如果群组大小 G 固定,扩大网络的级数 m,不会使网络丢失率发生大的变化. 图 9(c)反映了在相同网络规模下($N = 2^{16} = 65536$)时,使用不同规模的群组集线器,对网络阻塞率的影响. 显然, G 越大,在负载一定的情况下,阻塞率有显著降低. 因此要在输入端负荷很高时,若希望网络吞吐率能接近 100%,即能容许更高的输入负荷,可以类似构造并行 Banyan 网络的方法,同时使用若干个上述交换结构.

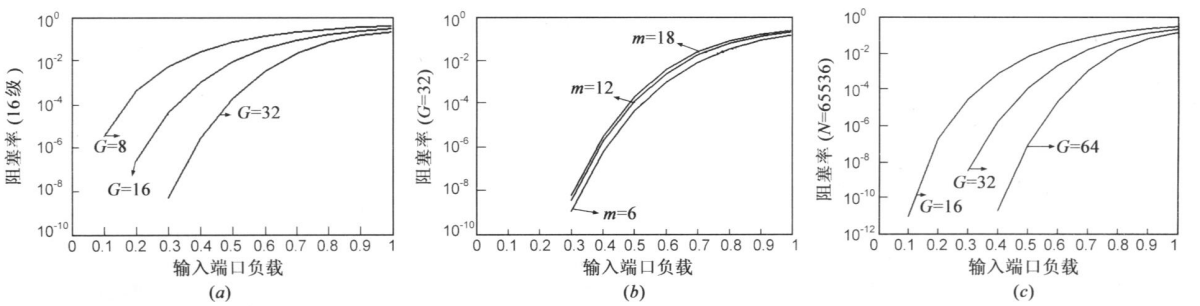


图 9　负载及阻塞率仿真结果

基于以上分析不难得出结论. 如果要求网络的阻塞率在指定范围之内,对于相同规模的网络,应尽量使用大规模的 2G-to-G 集线器. 但随着 G 的增长,单个群组集线器的构造复杂度也会增加,使网络时延增加. 因此,在实际设计中应权衡两者对网络性能的影响. 与 CIOQ 等结构对比来看, CIOQ 结构还需要一定附加存储器的支持,从一定程度上增加了网络实现代价. 而群组集线器结构具有无缓存时延和无抖动的优势.

6　结论

目前已提出多种能提供 100% 吞吐率的分组交换结构,如共享总线,共享内存,交叉矩阵及输入输出排队等. 它们的结构性缺陷是存在某个瓶颈限制了其规模的有效扩展,如带宽瓶颈,调度算法运算处理瓶颈等. 由于实际应用并不要求对所有种类的流量都是 100% 无阻塞,故本研究构造的是满足一定丢失率 QoS 保证的不需要每时隙调度的大规模 $N \times N$ 交换结构. 交换结构实际上是一个两级交换结构构成. 设 $M = 2^m$, $G = 2^g$, $N = M \times G = 2^m \times 2^g = 2^n$; $n = m + g$;大规模的 $N \times N$ 交换结构只有 M 个输出群组;分组不必进行调度而是先交换到各自所属的输出群组. 交换到同一个输出群组的分组再经小规模的 $G \times G$(如 $G \leqslant 128$)进行小规模的第二次交换,它的结构可采用各种成熟技术.

本研究对大规模 M 个群组的路由交换提出了结合群组排序集线器和多级互连网络的多路径自路由交换结构,并证明了该类结构构建于代数群论的自路由数学模型. 该结构具有完全分布式自路由,无需端口匹配调度、无内部缓存、无缓存时延及无抖动、按位置换群建模、及可递归扩展和模块化属性. 理论分析及仿真结果表明该结构适合于作为 QoS 保证的超大规模宽带交换结构.

参考文献:

[1] 任开新, 顾乃杰, 潘伟, 刘刚. 一种递归构造的合成 BANYAN 网络[J]. 电子学报, 2003, 32(2), 228- 231.

[2] 贺飞云, 闻懋生. 一种自选路由 ATM 容错交换网络[J]. 电子学报, 1997, 26(1): 28- 32.

[3] X H Jiang, P H Ho, H Shen. Fault tolerance analysis of optical switching systems built on the vertical stacking of banyan network[A]. Workshop on High Performance Switching and Routing[C]. 2004. 360- 364.

[4] C L Wu, T Y Feng. On a class of multistage interconnection networks[J]. IEEE Trans Comp, 1980, 29: 694- 702.

[5] Y M Yeh, T Y Feng. On a class of rearrangeable networks[J]. IEEE Trans Comput, 1992, 41(11): 1361- 1379.

[6] 李挥. Ω 等价类网络自路由研究[J]. 深圳大学学报(理工版), 1998, 15(4): 28- 36.

[7] S T Chuang, A Goel, N McKeown, B Prabhakar. Matching output queueing with a combined input output queued switch[J]. IEEE J Select. Areas Commun, 1999, 17: 1030- 1039.

[8] S C Liew, T T Lee. Principles of Broadband Switching and Networking[M]. The Chinese University of Hong Kong, 1995.

[9] K E Batcher. Sorting networks and their applications[A] Proc of the AFIP Spring Joint Computer Conference[C]. 1968, , 32: 307- 314.

[10] S Y R Li. Formalization of self-route networks and the rotary switch[A]. Proceedings of Infocom 1994[C]. Toronto, 1994. 438- 446.

[11] V E Benes. Mathematical theory of connecting network and telephone traffic[M]. Academic press, 1965.

[12] C Clos, A study of nonblocking switching networks[J]. Bell Syst Tech J, 1953, 14: 406- 424.

[13] 伊鹏, 汪斌强, 郭云飞, 李挥. 一种可提供 QoS 保障的新型交换结构[J]. 电子学报, 2007, 35(7): 1257- 63.

[14] P Yi, Y F Li, Hui Li, B Q Wang. Obtaining high performance switching with port distributed memories[A]. IEEE 4th International Conference on Communications, Circuits &Systems (ICCCAS' 06)[C]. Guilin, China, 2006. 1686- 1691.

[15] P Yi, B Q Wang, Hui Li, Y F Guo. Matching output queuing with a multiple input-crosspoint-output queued switches[A]. IEEE CHINACOM2006[C]. 2006.

[16] P Yi, H C Hu, Hui Li, B Q Wang. A distributed diffServ supporting scheduling scheme[A]. ICWMMN 2006[C]. The IET International Conference on Wireless, Mobile and Multimedia Networks, Hangzhou, China, 2006.

[17] P Yi, Hui Li, Q Yu, B Q Wang, Scheduling multicast and unicast traffic in buffered crossbar switches[A]. Wi0592, ICWMMN 2006[C]. The IET International Conference on Wireless, Mobile and Multimedia Networks 2006, Hangzhou, China, 2006.

[18] E Lu, M Yang, B Yang, S Q Zheng. A class of self-routing strictly nonblocking photonic switching networks[A]. GLOBECOM 2004[C]. IEEE, 2004, 29(2): 1011- 1015.

[19] G Kesidis, N McKeown. Output-buffer ATM packet switching for integrated-services communication networks[A]. Proc IEEE ICC' 97[C]. Montreal, Canada, 1997.

[20] N McKeown. Scheduling algorithms for input-queued cell switches[D]. Ph. D. dissertation, Univ. California, Dept Elect Eng Comput Sci Berkeley, CA, 1995.

[21] C S Chang, W J Chen, H Y Huang. On service guarantees for input buffered crossbar switches: a capacity decomposition approach by birkhoff and von Neumann[A]. IEEE IWQoS' 99[C]. London, U. K, 1999. 79- 86.

[22] S Iyer, S T Chuang, N McKeown. Practical algorithms for performance guarantees in buffered crossbars[A]. Proc of IEEE INFOCOM 2005[C]. Miami, FL, USA, 2005.

[23] A S Tanenbaum. Computer Networks[M]. 4thEdition, Prentice Hall, 2003.

[24] Y R Tsai, C W Lo. Banyarr based architecture for quasi-circuit switching[A]. IEEE ICNS 2006[C]. 23- 28.

[25] S Y R Li. Algebraic Switching Theory and Broadband Applications[M]. Academic Press, 2001.

作者简介:

李 挥 男, 1964 年出生于广东, 1986 年获清华大学工学学士, 2000 年 5 月获香港中文大学信息工程学哲学博士学位, 现为北京大学信息学院副教授, 深圳研究生院集成微系统学与工程重点实验室副主任, 流媒体研究中心主任, 国家集成电路设计产业化基地专家组成员, IEEE 通信学会会员. 主要研究方向为通信理论与系统, 流媒体理论与系统及其 VLSI 设计. 在国内外期刊和会议发表论文超过 30 篇, 申请发明专利 15 个.

E-mail: huilihuge@ yahoo. com. cn

何 伟 男, 1982 年出生于上海, 2005 年获复旦大学计算机专业理学学士学位, 现为北京大学深圳研究生院软件工程专业硕士研究生. 目前主要从事提供 QoS 保障的新型交换结构与相关调度算法的研究工作.

E-mail: heweist@ gmail. com

伊 鹏 男, 1977 年出生于湖北黄冈, 2003 年获通信与信息系统专业工学硕士学位, 现为信息工程大学通信与信息系统专业博士研究生, 已发表学术论文 16 篇, 获国家发明专利 2 项, 省部级科技进步一等奖 1 项. 目前的主要研究方向为高性能路由器交换结构与调度算法研究.

王秉睿 男, 1981 年出生于新疆乌鲁木齐, 2004 年获石河子大学计算机专业工学学士学位, 现于北京大学深圳研究生院攻读软件工程硕士学位. 目前从事计算机网络交换结构缓存容量分析方面的研究与仿真工作. E-mail: wbr912@ 163. com