文章编号：0258-7025(2004)Supplement-0245-04

一种 Tb/s 量级带宽的光电混合路由交换系统的结构

周新军，曹明翠，罗志祥，罗风光，徐 军
(华中科技大学激光技术国家重点实验室，湖北 武汉 430074)

摘要 提出一种采用光互连网络作为大容量路由器中的信息传输载体，以 VCSEL/PIN 作为光电、电光转换收发接口，用高速 IC 芯片来处理信息的光电混合太比特量级的大容量路由交换系统方案。通过采用高速分组交换芯片、可扩展的交换结构与共享储存排队方法，实现了交换机构规模可扩展到 1024×1024，端口速率为 2.5 Gb/s，信元丢失率小于 10^{-9} 的新型的光电混合路由交换核心模块。

关键词 计算机网络；路由器；光学互连；交换；VCSEL
中图分类号 TP393.4 **文献标识码** A

A Tbit/s Hybrid Optical Router Switching System Architecture

ZHOU Xin-jun, CAO Ming-cui, LUO Zhi-xiang, LUO Feng-guang, XU Jun
(*State Key Laboratory of Laser Technology, Huazhong University of Science and Technology, Wuhan, Hubei* 430074, *China*)

Abstract A new Tbit/s large capacity hybrid optical router switching system architecture was discribed, which employs high speed optical interconnection network to transmit information, VCSEL/PIN as optoelectronic and electrooptical device, and high speed large scale integrate circuit to process the information. The proposed architecture uses high speed packet switching circuit chip, scalable switching fabric and shared buffer output queue scheme. The result shows that it is feasible to construct a router switching core with cell loss rate below 10^{-9}, port rate at 2.5 Gb/s and scalable 1024×1024 switching fabric.

Key words computer network; router; optical interconnection; switching; VCSEL

近几年，IETF 提出了多协议标签交换 MPLS 技术，把路由技术和 ATM 交换技术相结合，既保留了传统路由器的可扩展性和灵活性，也集成了交换机的快速交换能力以及服务质量保证。

不管是当今的高速路由器、还是 ATM 交换机和 MPLS 交换机，交换结构都是它们的核心组成部分，其吞吐量、扩展性、可靠性等因素都决定于交换结构的性能。随着网络业务流的急剧增长，需要交换机的吞吐量也相应增加，这就导致一方面需要研究大容量的交换机构，另一方面，也需要研制的交换机构具有可扩展性，以满足不断增长的业务流的需求。

1 总体设计方案

1.1 高速路由器的技术方案

研制一个 Tb/s 量级的路由交换系统需要解决以下几个关键问题：

1) 要有足够的内部带宽和内部交换能力，能实现快速的路由查找。

2) 如何为大量的具有较小容量的路由器模块提供一个高速的互连模块来实现大容量的交换机结构。

3) 由于 IP 分组的长度是可变的，如何设计宽带的交换体系机构来快速转发 IP 分组。

4)由于太比特路由器主要应用在比较关键的网络节点，路由器内部要具有完整的协议栈，从而可以与大多现存的网络设备互连，能够将路由与交换功能有机地集成在一起。

我们的方案是采用光互连网络作为信息传输载体，VCSEL/PIN 作为光/电、电/光转换收发接口，以高速 IC 芯片处理信息的光电混合路由交换系统结构来实现，其原理如图 1 所示。

为了以较低的代价实现高速的分组转发，我们提出了一种可扩展的 Tb/s 量级的交换式路由器，其内部采用可扩展的 ATM 交换机构来支持高速的交换，如图 2 所示。它主要由输入接口、交换机控制器、输出接口、转发控制模块以及分组交换机构组成。通过分组交换机构互连的路由模块以一种分布的方

基金项目：国家 863 高技术计划(No.2002AA103064)资助课题。

作者简介：周新军(1973-)，男，华中科技大学激光技术国家重点实验室讲师，博士研究生，主要从事光互连、光交换与光电器件研究。E-mail: zhouxin@mail.hust.edu.cn

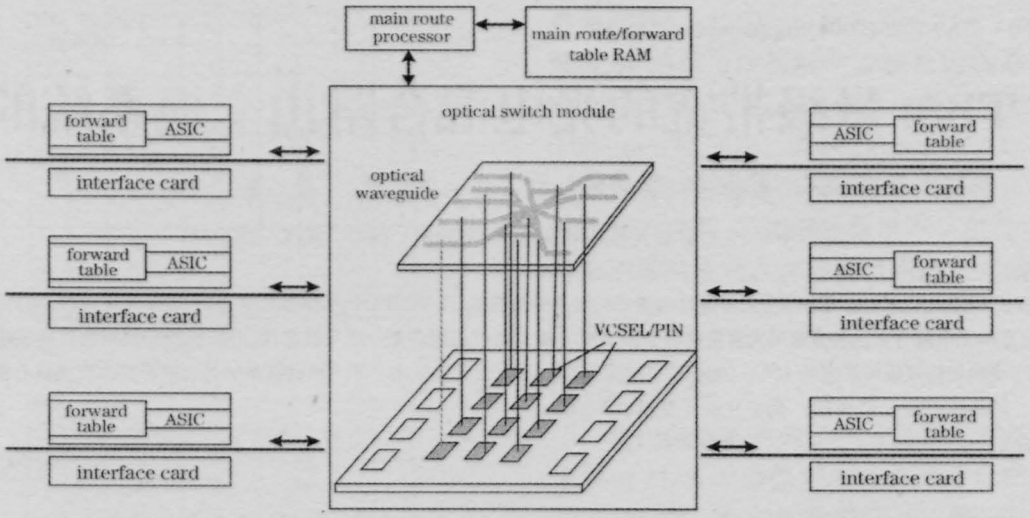

图 1 光电混合路由交换系统结构图

Fig.1 Block diagram of hybrid optical router switching system architectures

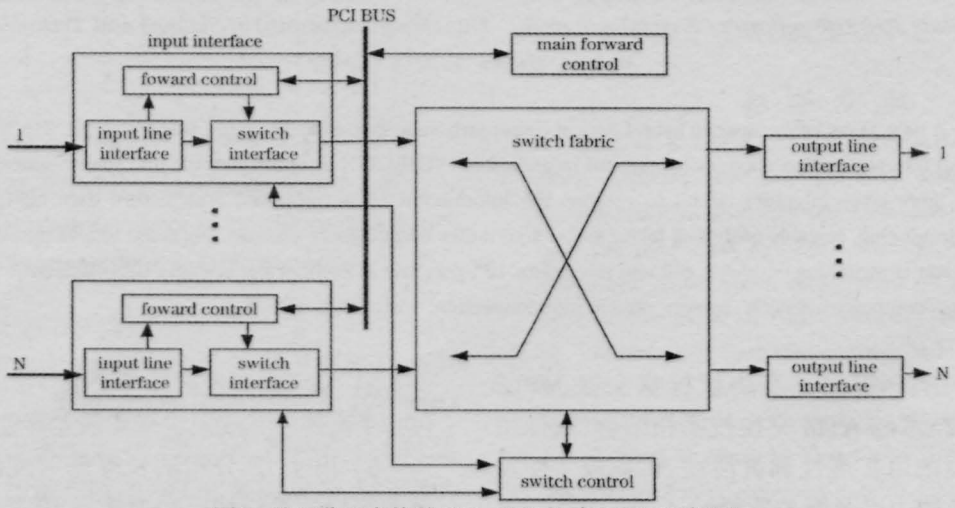

图 2 基于分组交换的 Tb/s 量级的路由器原理框图

Fig.2 Block diagram of Tb/s router based on packet switch technology

式执行分组转发。每一个接口都包含一个转发控制模块,执行标签的查找、分组分割以及分组转发。每个接口中的转发控制模块通过 PCI 总线或者分组交换机构相互连接。运行路由协议、标签分布协议以及其它控制协议的任务主要由主转发控制器承担,以加速在每个接口上的数据传送。

每个输入接口主要由一个输入线接口、输入交换机接口和转发控制模块组成。由于 IP 分组是可变长度的分组,而分组交换机构是固定长度分组交换机。为了使内部交换机对外部是透明的,因此必须在输入接口上把可变长度的分组打包成固定长度的分组,经过分组交换机构后在输出接口上再重新组装成分组。

当一个输入接口接收到一个输入分组时,它将首先验证其 IP 头。如果该 IP 头无效,那么此分组被丢弃。否则,转发控制模块将确定此分组是属于控制消息还是数据,然后分别处理。如果是控制消息,转发控制模块就通过一条控制路径 PCI 总线到主转发控制来转发分组。如果是数据,转发控制将根据此分组的标签查询,找出它的目的输出接口,并将此信息传递给输入交换机接口。然后输入交换机接口将通过分组交换机构转发该分组到输出接口。

这种结构的 MPLS 交换机具有以下特点:第一,它不需要额外的网络协议来支持流分类以及虚通道连接。其次,由于采用分组交换芯片以及可扩展的交换结构, 交换机构规模可以扩展到 1024×1024,端口速率为 2.5Gb/s,10Gb/s 甚至更高,所以该 Tb/s 的路由器具有良好的可扩展性和灵活性。

1.2 可扩展的分组交换体系结构

采用基于分组交换的互连模块,在每个接口上

将可变长度的 IP 分组转变成固定长度的分组，便于分组交换模块进行快速的分组交换。同时，采用基于 MPLS 的标签交换方式，把路由与交换有机地结合在一起。但实现太比特量级高速路由器的另一个关键问题，是如何为大量的较小规模的路由交换模块提供一个高速的互连模块来实现大容量的交换机构，即采用何种互连机构能够有效地将各个小的路由模块互连起来以实现低阻塞率的路由交换？

我们知道，互连机构是一个扇出比为 $F=m/n$ 的扇出网络，可以用 F 个 $N\times N$ 的 crossbar 网络平行叠加而成，但这会使得网络的节点数很大。为了减少互连网络的节点数，Bell 实验室采用了分组 crossbar 网络的思想，互连机构完全由 16×16 crossbar 构成。

引入分组 crossbar 一方面带来了扩展性的提高，但另一方面也增加了阻塞率。在输入业务流为均匀业务流的情况下，对于一般的广义 Knockout 交换机构而言，基本交换单元为 16×16，扇出比 $F=4$，在负载 $\rho=0.9$ 的情况下，阻塞率为 10^{-3} 量级，这远远不能满足现代通信系统的要求。为了得到更低的阻塞率，需要对交换网络结构进行改进或引入新的寻径算法。

贝尔实验室提出了一种“滚动”(rolling)算法[1]，其核心思想是将一个分组的寻径过程扩展到两个分组时隙完成，这样减少了分组冲突的机会。但阻塞率仍较高。在均匀业务流输入且负载为 $\rho=0.9$ 的情况下，其阻塞率为 10^{-7} 量级，这仍然不能满足需求。

我们提出了一种新的方案，把多路输入端的链路共享，即一组共享链路的输入端口中的任意一个输入端口都可以利用共享的多个输入端口的链路，如图 3 所示。对于多路输入端口共享链路的情况，由于每个输入端通过互连机构不同 pipe 中相同位置的 G 个 crossbar，共有 $G\times F$ 条路径可连往某一输出模块。这时在每个 pipe 中相邻的 G 个小 crossbar 的各输入端来的分组之间可能会发生路径争用的情况，即产生内部阻塞。该结构在性能上可简化为如图 4 所示的等价结构。

可看出，这一等价图与 $N=G\times n_c$，$m=G\times F$，$n=G$ 的广义 knockout 交换网络的结构完全相同。在均匀业务下，G 路输入端口共享链路的广义 knockout 交换网络输入端分组被阻塞的概率为：

$$P_{\text{loss}}=\frac{1}{n\rho}\sum_{k=m+1}^{N}(k-m)\binom{N}{k}\left(\frac{n\rho}{N}\right)^{k}\left(1-\frac{n\rho}{N}\right)^{N-k}=\frac{1}{G\rho}\sum_{k=G\times F+1}^{G\times n_c}(k-G\times F)\binom{G\times n_c}{k}\left(\frac{G\rho}{G\times n_c}\right)^{k}\left(1-\frac{G\rho}{G\times n_c}\right)^{G\times n_c-k}$$

当扇出比 $F=4$ 时，共享链路的输入端口数增加到 4 时，系统的内部阻塞率可到达接近 10^{-9}。当共享链路的输入端口数增加到 8 时，系统的内部阻塞率可低于 10^{-12}。

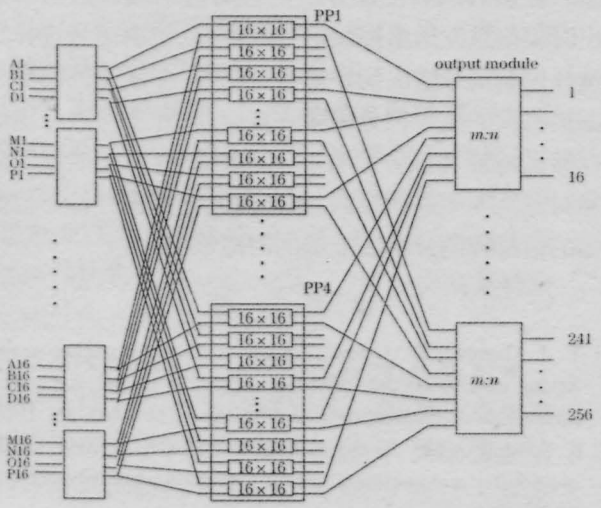

图 3　4 个输入端口共享链路的交换机构框图

Fig.3 Block diagram of switch fabric with 4 input port shared link

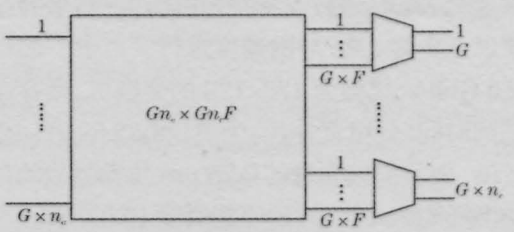

图 4　多输入端口共享链路的简化性能等价图

Fig.4 Simplified equivalence diagram with shared multi-input-ports

1.3 分布控制的多缓存共享输出排队结构

由于分组到达的随机性，在同一个分组时隙可能有多个分组到达同一个目的输出端口。但是，由于输入输出速率是相同的，在一个分组时隙只能有一个分组输出，因此将有分组不能输出而被丢弃。为了有效控制因竞争失败而造成的分组丢失，交换机必须提供缓存功能，对竞争失败的分组进行缓存排队。

在共享存储器类型的交换机中，所有的输入和输出端口都有权使用一个共享的存储器模块。在一个时

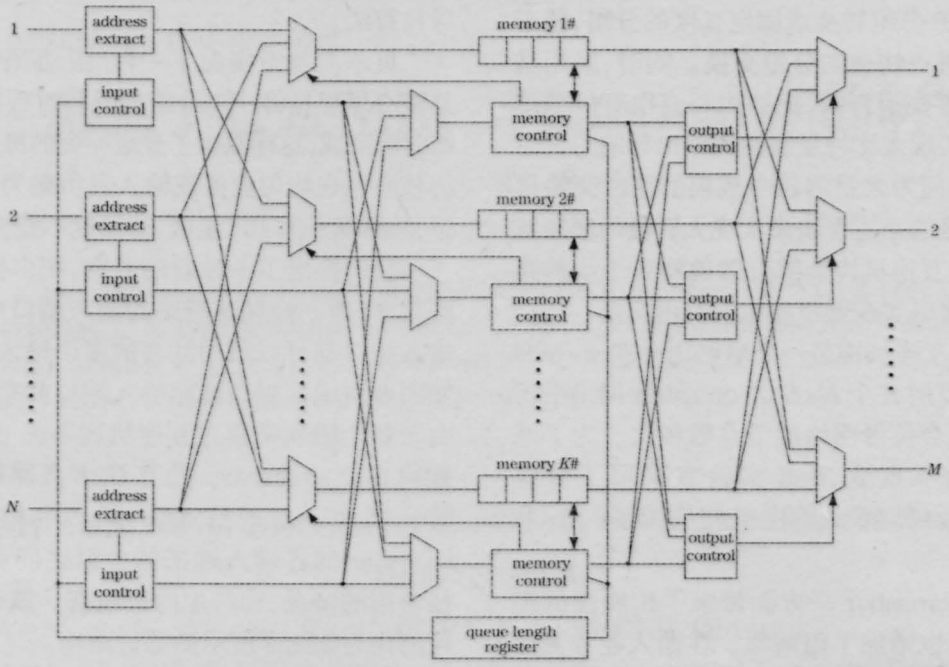

图 5 分布控制的多缓存共享输出模块结构框图

Fig.5 Block diagram of output module with multipile buffer memory shared

间隙内，最多有 N 个来自输入端口的分组可以被写入共享的存储器，并且最多有 N 个分组可以从共享的存储器中被读出。在一个 $N\times N$ 的共享存储器类型的交换机中，存储器模块必须能够在一个时间隙内执行 N 个写入和 N 个读出。因此，每个时间隙由 $2N$ 个子隙组成以执行 N 个写入和 N 个读出。对于输入/输出线速为 L bit/s 的 $N\times N$ 交换机而言，共享的存储器需要以至少为 $2NL$ /s 的速度运行。对于一端口速度为 2.5 Gbit/s，规模为 128×128 的交换机，则容许的最大访问时间(分组宽并行)为$(53\times 8)/(2.5\times 10^9\times 128\times 2)=0.66$ ns。而目前较先进的 0.25 μm 技术制造的存储器的访问周期为几纳秒，远不能满足上述要求。

H. Kondoh 等提出了一种多缓存共享的交换机构[2,3]，在逻辑上将所有的存储器模块作为一个大的联合的缓冲存储器来考虑。利用交换矩阵并行传输替代传统共享存储器交换结构中的复用和解复用，因而消除了存储器访问时间的限制。但这种集中控制的方式存在的一个缺点是，如果交换机的规模扩大，那么集中控制就将成为整个交换机的瓶颈。交换机规模的增加，将要求集中控制器在一个固定长度的时间间隙内执行更多的任务(例如，对分组的读写操作，存储和管理地址队列)。同样地，随着交换机规模的增加，在一个固定长度的时隙内必须向输入输出空分交换结构提供更多的寻径信息。

我们采用一种新的分布控制的多缓存共享的交换结构(如图 5 所示)，很好地解决了这个问题。每一个缓存储存器都有一个缓存控制器与之相对应，每一个缓存储存器在同一个时隙内读出目的地各不相同的分组，这样，既减少了分组读出时的冲突，又降低了对缓存控制器速率的要求。分布控制的多缓存共享输出模块主要由 N 个地址提取电路，N 个输入控制模块，$N+M-1$ 个存储器，$N+M-1$ 个存储器控制器，M 个输出控制模块，以及一些选择器组成。

地址提取电路从到达的分组中提取该分组的目的输出端口，并将该信息传递给输入控制器。输入控制器根据此目的端口信息，计算该分组的输出时隙。输入控制器然后把该分组的目的输出端口信息和输出时隙信息发送给控制信号选择器。控制信号选择器选择控制信号输出到存储器控制器，存储器控制器根据接收到的信号把分组写入存储器，并把写入分组的存储器地址和该分组将输出的时隙分别在地址队列和时隙队列中进行排队。输出控制器根据存储器控制器的信号控制分组在准确的时隙输出。

参 考 文 献

1 T. J. Cloonan, D. Grove. Terabit per second packet switch having distributed out-of-band control of circuit and packet switching communications[P]. USA Patent 5537403, Jul, 1996

2 H. Kondoh. A 622-Mb/s 8×8 ATM switch chip set with shared mutlibuffer architecture[J]. *IEEE J. Solid State Circuits*, 1993, **28**:808~815

3 S. H. Kang, C. Oh, D. K. Sung. A high speed ATM switch with common parallel buffers[C] *IEEE GLOBECOM* 95, 1995, 2087~2091