# 一种新的光突发交换节点结构研究

王 晟,罗蕴翰,王 雄,王 靖,姚 锐,彭 亮
(电子科技大学通信与信息工程学院,成都 610054)

**摘 要**: 与波长路由技术和光分组交换相比较,光突发交换(Optical Burst Switching,OBS)更适合于构建未来的光因特网. OBS 核心交换节点的设计必须在保证一定的交换性能的同时,尽量降低光器件成本. 本文提出了一种新的光突发交换结构,称为 PM ×1 ×P 结构#. 该结构有效地克服了使用光纤延迟线或光缓存等器件所导致的内部阻塞问题,提高了交换结构的吞吐量;同时,与具有相似性能的交换结构相比,该结构所需要的光器件较少,成本较低. 通过对比仿真实验,结果验证了该结构的优越性.

**关键词**: 光交换节点; 内部阻塞; 光纤延迟线; 光突发交换
**中图分类号**: TN929. 11 **文献标识码**: A **文章编号**: 0372-2112 (2004) 12A-086-07

# A Novel Design of Photonic Switches in Optical Burst Switching Networks

WANG Sheng,LUO Yun-han,WANG Xiong,WANG Jing,YAO Rui,PENG Liang
(*School of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan*,610054, *China*)

Abstract: Compared to the Wavelength Routing (WR) and Optical Packet Switching (OPS) networks,Optical Burst Switching (OBS) is believed to be a more attractive technical solution for future Internet infrastructure. One of the issues of design of OBS switches is how to keep the implementation cost low while maintain high performance. A novel design of photonic switch for OBS networks is proposed,which is called PM ×1 ×P structure. The Internal Blocking can be effectively avoided in this structure,hence much higher throughput of data burst can be achieved. On the other hand,the proposed design uses less optical elements in terms of optical switching elements,multiplexers and couplers. It turns out to be a more cost-saving design than other photonic switches with similar throughput performance. Results of extensive simulations are presented,and the performance of proposed design is proved.

Key words: photonic switch;internal blocking;fiber delay line;optical burst switching

## 1 引言

随着密集波分复用(Dense Wavelength Division Multiplexing,DWDM)光传输技术和设备的成熟和广泛应用,直接在光层实现业务交换的需求越来越大. 因此,全光实现的"电路"交换(即所谓波长路由)和分组交换受到广泛关注,被认为是克服电交换速率瓶颈的有效途径. 但是,这两种交换方式都存在各自缺陷. 一方面,随着分组业务(尤其是 IP 业务)的爆炸式增长,使得人们普遍认为未来的通信网络设计应以优化支持分组业务为主要目标,而波长路由交换方式本身的特点决定了它不适于支持突发性强的 IP 业务;另一方面,全光的分组交换在实现上存在若干难点,许多关键技术仍有待于光器件技术的成熟(如灵活有效的光逻辑器件、光存储器件,以及光域的同步技术等). 因此,近年来提出的光突发交换(Optical Burst Switching,OBS)技术日益受到人们的关注[1,2]. OBS 技术在支持分组业务的性能上高于波长路由方式,而实现难度(尤其是对光器件技术的要求)低于光分组交换[3].

一个 OBS 网络主要由边缘节点、核心节点和 DWDM 链路构成. 在 OBS 网络中,突发包(Burst)可以看作是由一些 IP 分组组成的超长分组,而这个超长分组的分组头就是突发包的控制分组(Burst Header Packet,BHP). 与传统分组交换不同的是,BHP 与突发包在物理通道上是分离的,在 DWDM 传输系统中,可以采用一个(或多个)专门的波长作为控制通道,用于传送 BHP,而把其他的波长作为数据通道. 边缘节点负责对数据分组进行缓存和封装,组装成突发包,然后发送给与之最邻近的 OBS 核心节点. 封装时边缘节点生成 BHP,先于突发包在特定的控制通道上发送. BHP 与突发包一一对应,在边缘节点设置 BHP 与突发包之间的偏移时间(offset time). BHP 中包含突发包的有关特征信息,如偏移时间、突发长度、所占用的波长等. BHP 在核心节点转换为电信号进行处理,包括路由的确定、资源的预约以及交换矩阵的配置等,保证当突发包到达时相应的数据通道已经配置好,从而实现数据在光域的透明传输. BHP 不必等待目的端的反馈确认,资源预约是单向的. 这种"单向预约"机制减小了连接建立延迟,提高了信道利用率.

在 OBS 核心节点,当多个突发包同时要交换到同一输出端口的某个特定波长通道上去的时候,会产生"冲突". 此时为防止数据丢失,在交换结构设计上主要采用波长变换器(Tunable Wavelength Converter,TWC)、光缓存(Optical Buffer,OB)或光纤延迟线(Fiber Delay Line,FDL). FDL 是一种最简单的 OB,而通常 OB 是由多个 FDL 通过光开关组合而成的. 本文第 2 节将进一步讨论 OB 的结构.

OBS 核心节点中需要配置各种高性能的光器件或组件. 例如切换速度低于微秒量级的光开关、高速可调谐激光器、稳

收稿日期:2004-10-18;修回日期:2004-11-18
基金项目:国家 863 计划重大项目资助(No. 2002AA122021)

定时间低的全光波长变换器等.因此光器件成本是设计中必须重点考虑的问题.

目前 OBS 核心节点光交换结构主要包括三类[4]:(1)基于空分交换阵列的输入缓存结构;(2)基于空分交换阵列的环回缓存结构;(3)基于广播 - 选择机制的交换结构.其中第 3 种交换结构大量采用了分光器件,光信号能量损失严重,难以满足实际需要.第 2 种结构中,空分交换阵列必须具有冗余的输出端口,才能提供环回支路以支持缓存.因此为了支持相同的端口数目,第 2 种结构必须采用比第 1 种结构更大规模的空分交换阵列.因此,目前最受关注的就是第 1 种结构(本文第 2.3 节讨论的 $PM \times PM$ 结构就属于这种类型).但是,光域的存储器件远没有达到成熟的地步,目前应用的光缓存大多以 FDL 和高速光开关为基本组件构造而成.这种光缓存的延迟粒度和最大延迟时间都是有限值,突发包因冲突而必须等待的时间如果超过了最大延迟时间,则只能丢弃.更重要的是,光缓存本质上是 FIFO(First In First Out,先入先出)系统,存在队头(Head Of Line,HOL)阻塞.根据分组交换机的经典理论,输入缓存且缓存为 FIFO 系统时,交换机的吞吐量约为 0.586[5].

综上所述,在设计 OBS 核心交换结构时,需要综合考虑两方面的因素.一方面,由于高速高性能的光器件价格昂贵,必须尽量降低光器件成本;另一方面,受制于现有光存储技术,OBS 交换节点解决冲突的能力无法与电分组交换机相比拟.而现有的 OBS 交换结构都没有很好地解决这些问题.

基于以上考虑,本文提出一种新的交换结构,称为 $PM \times 1 \times P$ 结构.该结构可以有效解决包括队头阻塞在内的各种内部阻塞问题,有效提高交换机的吞吐能力,降低丢弃率.并且,与具有相似交换性能的其它交换结构相比,该结构所需要使用的光器件数目较少,成本较低.

本文后续部分是这样安排的.第 2 节将首先讨论 OBS 交换节点的功能结构,然后给出 $PM \times 1 \times P$ 结构的具体描述.为了便于比较,第 2 节还简要讨论了一类具有典型意义的交换结构.针对第 2 节讨论的几种交换结构,第 3 节详细分析了各种冲突/阻塞类型及其解决机制,并给出相应的调度算法.随后的第 4 节对比研究了各种结构的交换成本.第 5 节通过计算机仿真实验进一步研究了影响交换性能的几个关键因素,并验证了 $PM \times 1 \times P$ 结构的优越性.最后的第 6 节总结全文.

## 2　OBS 核心交换节点

本文中,OBS 网络采用 LOBS(Labeled Optical Burst Switching)方案组网[6].这种方案中,每条 LSP(Label Switching Path)的路径是固定的,这样既便于实施业务量工程以优化配置资源,同时也有利于 OBS 边缘节点设定偏移时间.核心交换节点根据 BHP 携带的标签信息以及核心节点内的标签交换表可以唯一地确定输出端口,即该 BHP 对应的突发包的出端口是唯一的(在某些 OBS 组网方案中允许核心节点采用 Deflection Routing[7]解决冲突,此时一个突发包的输出端口是不一定的).

### 2.1　OBS 核心节点功能结构

OBS 核心交换节点通常分成两大功能部分:电控制部分和光交换部分.电控制部分主要包括交换控制单元,光交换部分主要包括空分交叉矩阵、TWC 和 OB 等器件或组件.如图 1 所示.

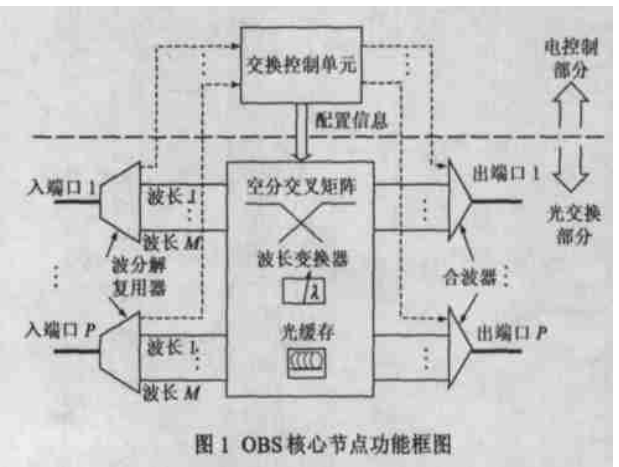

图 1 OBS 核心节点功能框图

$P$ 个输入输出端口采用 DWDM 光纤链路.每根光纤上复用 $M+1$ 个波长,其中 $M$ 个波长用于突发包的传输,剩余的 1 个波长承载 BHP 信息.输入光纤经波分解复用器(Wavelength Demultiplexer)后,BHP 被送到交换控制单元.OBS 核心节点通常采用集中控制方式,所有输入端口中的 BHP 信息都送到交换控制单元,集中进行调度、处理和交换.如果调度失败,则 BHP 及其所对应的突发包都被丢弃;如果调度成功,则交换控制单元负责产生对光交换部分各相关部件的配置命令,更新 BHP 中的信息,然后经 E/O 变换后将 BHP 送到相应的输出端口.

本文的研究重点在核心节点的光交换部分,对电控制部分不做深入讨论.本文后续内容中出现的交换结构示意图都只针对光交换部分,略去电控制部分.

### 2.2　PM ×1 ×P 结构

本文提出的 $PM \times 1 \times P$ 结构如图 2 所示.其交换功能主要靠 $PM$ 个 $1 \times P$ 光开关实现,因此命名.每个入端口的光纤链路上复用的 $M$ 个波长信号经波分解复用器后,通过 $M$ 路光纤分别与 TWC 相连.每个 TWC 输出的信号与一个 $1 \times P$ 光开关相连.光开关的 $P$ 个出口与整个交换结构的 $P$ 个出端口一一对应.通过配置光开关,去往不同出端口的突发包被送到光开关相应的输出端.来自不同光开关,去往相同出端口的信号经合波器复用到一根光纤链路上输出.在合波器与光开关之间的每条通路上都串连了一个 FDL.

FDL 的构造方式如图 3 所示.FDL 由 $K$ 级 $1 \times 2$ 光开关、一个合波器、以及延迟光纤构成.每级 $1 \times 2$ 光开关都有 2 路输出,其中 1 路经延迟光纤后与下一级光开关的入口相连;另 1 路直接与 FDL 出端的合波器相连.每级延迟光纤长度都是相同的.其延迟时间称为单位延迟,记为 $d$.光开关个数 $K$ 称为最大延迟级数,相应地,FDL 的最大延迟时间为 $D = Kd$.通过配置这 $K$ 个 $1 \times 2$ 光开关,经过 FDL 的突发包所经历的延迟可以取集合 $\{0, d, 2d, \cdots, Kd\}$ 中的任何一个值.

定性地看,单位延迟越小,对突发包的延迟调整能力越精确,交换性能越好.但是,单位延迟越小,为达到相同的最大延迟,所需要的延迟级数越大,所需的光开关越多,成本越高.

给定 $K, d$ 越小,则 $D$ 越小.单位延迟越小意味着时间调

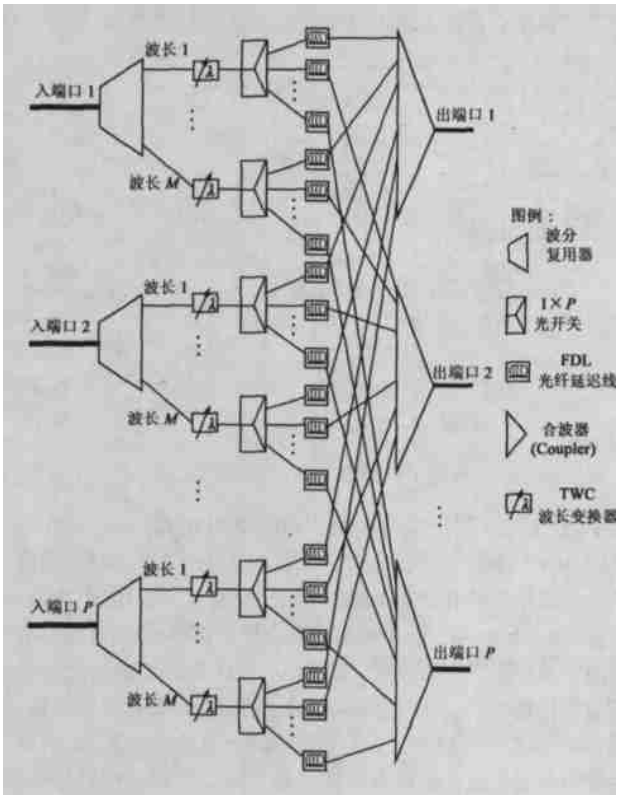


图 2 $PM\times1\times P$ 结构

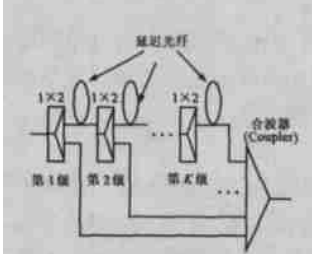


图 3 光纤延迟线 (FDL)示例

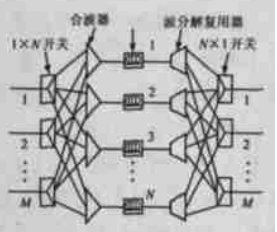


图 4 光缓存结构示意

整可以越精细;而另一方面,最大延迟越小意味着缓存能力越差.在这两个因素的共同作用下,对于给定的 $K$ 值,一定存在一个最佳的 $d$ 值.

### 2.3 $PM\times PM$ 结构

为便于进行对比研究,本节简要介绍一类具有代表性的交换结构,称为 $PM\times PM$ 结构.图 5 给出了 4 种 $PM\times PM$ 结构的变型.其中图 5($d$)给出的 $PM\times PM$ - Ⅳ结构在文献[8]中讨论过.其它 3 种结构可以看作是该结构的简化版本.

$PM\times PM$ 结构中完成交换功能的主要是 $PM\times PM$ 无阻塞的空分交叉矩阵,因此命名.4 种变型结构还具有 2 个共同特点:(1)每个入端口分解出的 $M$ 个波长通道上都分别配置了 TWC,称为"前置 TWC";(2)出端口采用合波器,将空分交换阵相应的 $M$ 个出口上的波长信号复用到一个光纤链路上输出.这 4 种变型在结构上的区别主要在从 TWC 出口到空分阵入口之间的配置上.图 5(a)所示的 $PM\times PM$ - Ⅰ结构中,TWC 出口与空分阵入口直接相连.$PM\times PM$ - Ⅱ结构在每个通道上分别配置了一个 FDL,如图 5($b$)所示.图 5($c$)所示的 $PM\times PM$ - Ⅲ结构中,每个入端口都配置了一个共享的 OB.从外部连接方式上看,共享式的 OB 是一个 $M$ 入 $M$ 出的结构.其内部构造如图 4 所示.$PM\times PM$ - Ⅳ结构在 $PM\times PM$ - Ⅲ结构的基础上,为每个通道各增加了一级 TWC,称为"后置 TWC",如图 5($d$)所示.

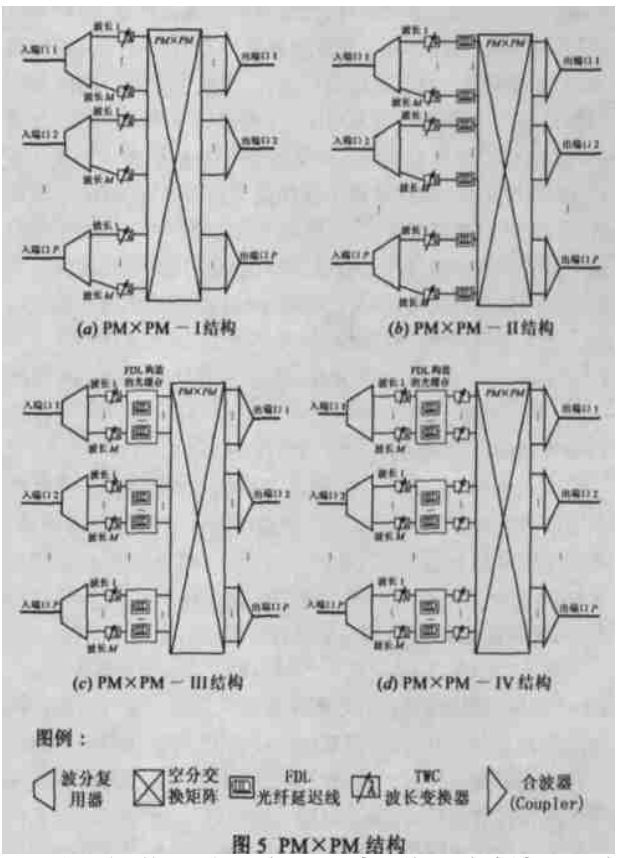


图 5 PM×PM 结构

图 4 所示的 OB 由 $M$ 个 $1\times N$ 光开关、$N$ 个合波器、$N$ 个 FDL、$N$ 个波分解复用器,和 $M$ 个 $N\times1$ 光开关构成.每个 OB 入口进入的突发包依次经过 $1\times N$ 光开关、合波器进入 FDL,通过配置光开关,任一入口进入的突发包都可以使用 $N$ 个 FDL 中的任意一个进行延迟.延迟后的突发包依次通过波分解复用器和 $N\times1$ 开关到达整个 OB 的出口.具体去往哪个出口,取决于进入 OB 时该突发包所占用的波长.经波分解复用器后,$i\,(1\le i\le M)$ 波长上的突发包被固定地送往 OB 的第 $i$ 个出口.

## 3 冲突解决及调度算法

### 3.1 冲突分析

突发包从交换结构入端口到出端口之间的完整通路(称为突发包的交换通路)上,所有需要竞争使用的资源,称为"冲突资源".在不同的交换结构中,冲突资源的种类和数量也不同.从冲突产生的机理来说,OBS 交换结构中的冲突主要有以下四种类型:

**(1)出波长冲突.** 当两个突发包来自不同的入端口,并要求去往相同出端口的某个特定波长通道时,就会出现冲突.这种因竞争出端口波长通道导致的冲突称为"出波长冲突".出波长冲突是输入业务量的具体特性导致的,本文研究的 5 种

交换结构中都存在这种冲突.

**(2) FDL 出口冲突.** 图 3 所示的 FDL 出口配置了一个合波器,来自各级光开关的光纤在此汇聚成一个通路.假定先后有两个突发包 $B_1$ 和 $B_2$ 进入 FDL,它们使用相同的波长 $\lambda$,但去往不同的出端口.若 $B_1$ 需要延迟,那么为了避免在 FDL 出口发生重叠,即使 $B_2$ 不需要延迟,它也可能不得不等到 $B_1$ 完全离开 FDL 后才能通过.这种因延迟时间不同而导致突发包竞争 FDL 出口的情况就称为"FDL 出口冲突".仅就 FDL 而言,如果进入 FDL 的突发包使用了不同的波长,即使存在时间上的交叠,在 FDL 出口处也不会冲突.由于 $PM \times PM$ - Ⅰ结构中没有 FDL;而 $PM \times 1 \times P$ 结构中虽然使用了 FDL,但是进入 FDL 的突发包一定去往同一出端口,因此这两种结构中都不存在 FDL 出口冲突.

**(3) 空分阵入口冲突.** 以 $PM \times PM$ - Ⅲ结构为例.该结构中,来自同一入端口不同波长通道上的多个突发包,经过 OB 后可以到达 $PM \times PM$ 空分阵的同一入口.如果这些突发包通过该空分阵入口的时间相互交叠,就会导致冲突.由于 $PM \times 1 \times P$ 结构中没有空分阵;而 $PM \times PM$ - Ⅰ和 $PM \times PM$ - Ⅱ结构中空分阵入口只与一个输入波长通道相连,故这种冲突只存在于 $PM \times PM$ - Ⅲ和 $PM \times PM$ - Ⅳ结构中.

**(4) 空分阵出口冲突.** 从不同的空分阵入口进入的突发包,如果希望同时到达同一个空分阵出口,就会导致这种冲突.除了 $PM \times 1 \times P$ 结构外,4 种 $PM \times PM$ 交换结构都存在这种冲突.

在特定情况下,上述四种冲突是相互关联的.例如,$PM \times PM$ - Ⅱ结构中,FDL 出口与空分阵入口直接相连,如果调度时避免了 FDL 出口冲突,空分阵入口冲突也就不会出现了.而在 $PM \times PM$ - Ⅲ和 $PM \times PM$ - Ⅳ结构中,虽然 FDL 出口与空分阵入口并不直接相连,但是即使 FDL 出口复用了多个波长,也会被波分解复用器送到不同的空分阵入口.因此只要空分阵入口有空闲,就说明 OB 中的 FDL 出口没有冲突.考虑到这种关联性,交换结构中需要避免的冲突比可能出现的冲突要少.

综合以上讨论,可以确定各种交换结构中的冲突资源,如表 1 所示.

沿用分组交换机的术语,因出波长冲突导致的突发包延迟或者丢弃称为"外部阻塞";反之,其它冲突引起的阻塞称为"内部阻塞".其中因 FDL 出口冲突导致的突发包丢弃或延迟称为"队头阻塞".

$PM \times PM$ - Ⅰ结构虽然只有外部阻塞,但解决阻塞问题的手段只有波长变换,一旦出端的所有波长都不空闲,则突发包只能被丢弃.与之相比,$PM \times PM$ - Ⅱ结构可以利用 FDL 对突发包进行延迟,以等待出端口出现空闲波长.但引入 FDL 后又不可避免地引入了队头阻塞,因此总的交换性能提升有限.$PM \times PM$ - Ⅲ结构中配置了共享式的 OB,通过配置 OB 中的 $1 \times N$ 开关,突发包可以选择进入 $N$ 个 FDL 中的任一个,因此与 $PM \times PM$ - Ⅱ结构相比,解决队头阻塞的能力有所提高.但是,该结构只配置了前置 $TWC$,假定为了解决出波长冲突,突发包被变换为 $\lambda_i$,则该突发包只能使用空分阵的第 $i$ 号入口,一旦空分阵入口 $i$ 出现冲突,突发包只能被阻塞.因此,该结构在解决 FDL 出口冲突的同时也引入了空分阵入口冲突.$PM \times PM$ - Ⅳ结构中配置了后置 TWC,可以利用前置 TWC 来任选一个空分阵入口,再利用后置 TWC 任选出波长.从而在解决出波长冲突的同时,部分地(而不是完全地)解决了空分阵入口冲突问题.

**表 1　5 种交换结构中涉及到的冲突资源**

| | $PM \times 1 \times P$ | $PM \times PM$ - Ⅰ | $PM \times PM$ - Ⅱ | $PM \times PM$ - Ⅲ | $PM \times PM$ - Ⅳ |
|---|---|---|---|---|---|
| 出波长 | Yes | Yes | Yes | Yes | Yes |
| FDL 出口 | No | No | Yes | No | No |
| 空分阵入口 | No | No | No | Yes | Yes |
| 空分阵出口 | No | No | Yes | No | No |

通过上述讨论可以看出,$PM \times PM$ - Ⅰ结构解决外部阻塞问题的手段有限,而 $PM \times PM$ - Ⅱ/Ⅲ/Ⅳ虽然改善了这种状况,但都不同程度地引入了新的内部阻塞问题.而在本文提出的 $PM \times 1 \times P$ 结构中,进入交换结构的每个波长通道都按照出端口进行分流,从而有效避免了内部阻塞;同时,该结构可以通过波长变换和延迟两种手段来解决外部阻塞问题.因此在本文所研究的 5 种交换结构中,$PM \times 1 \times P$ 结构具有最佳交换性能.

### 3.2　调度算法

为了便于比较各种交换结构本身的性能,本文针对不同的交换结构采用相似的调度算法,并且算法设计的思路是一致的.其主要特点包括:

(1) 采用"顺序调度"方案.即核心节点按照 BHP 到达的先后顺序逐个进行调度,如果调度成功,调度的结果将携带在 BHP(通过修改偏移时间字段的值)中发送到下一节点.先完成调度的 BHP 所预约的资源不能更改,也不能抢占.如果调度失败,则丢弃该 BHP 及其对应的突发包.

(2) 对于每种冲突资源的占用状态,只记录最后可用时间(Last Available Time,LAT).

(3) 相应于占用状态的记录方式,在调度突发包时不进行插空(Void Filling)[9].

(4) 选择可用资源时,优先采用"最小延迟"策略,然后考虑"最小空隙"策略.例如,突发包 $B$ 的到达时间为 $t_a$,希望到达的出端口有 $M$ 个波长 $\lambda_1$,$\lambda_2$,…,$\lambda_M$,它们的 LAT 分别为 $t_1$,$t_2$,…,$t_M$,FDL 的单位延迟为 $d$,则按照最小延迟策略,$B$ 应使用所需延迟级数最小的波长,假定该波长为 $\lambda_{opt}$,则有 $\lambda_{opt} = \min_{\lambda_i \in \{\lambda_1, \lambda_2, \cdots, \lambda_M\}} \{(t_i - t_a)/d\}$.如果存在多个波长满足上述条件,它们构成一个集合 $A$,最小延迟级数为 $k$,则按照最小空隙策略,应选择 $B$ 延迟 $k$ 级后造成时间空隙最小的那个波长,即有 $\lambda_{opt} = \min_{\lambda_i \in A} \{t_a + kd - t_i\}$.

给定一个 BHP,以及相关的冲突资源占用状态,算法按以下步骤调度:

第 1 步,根据 BHP 中携带的标签信息确定其出端口.

第 2 步,根据 BHP 中的信息估计其对应的突发包的到达时间以及突发包长度.

第 3 步,查询相关冲突资源的占用状态,确定最佳交换通路. 若交换结构中存在 $C$ 种类型的冲突资源,每种冲突资源包括 $R_i$ 个$(1 \le i \le C)$,则存在 $\prod_{1 \le i \le C} R_i$ 种可能的交换通路. 对于每条可能的交换通路,都可以计算出最早通过该通路所需要的延迟级数,取所有这些延迟级数的最小值 $k$. 若所需要的最小延迟级数大于 FDL 的最大延迟级数 $K$,即 $k > K$,则丢弃该 BHP 及其突发包. 否则选择需要最小延迟级数的交换通路作为最佳通路. 如果存在多条最小延迟通路,则选择造成空闲时间间隔最小的作为最佳通路.

第 4 步,如果存在最佳交换通路,更新占用状态信息. 即把最佳交换通路上的所有冲突资源的 LAT 都更新为突发包占用该资源的结束时间.

## 4 成本比较

第 2 节给出的 5 种交换结构主要包括了以下 4 种光器件:(1) 波分解复用器;(2) 合波器;(3) 波长变换器和 (4) 高速光开关. 在本文涉及的交换结构中,前 3 种作为单独的器件使用,规格相同或相似,为比较成本只需列出各种结构所需的器件个数就可以了. 但光开关的成本与光开关的规模直接相关,不能直接比较. 为此,本节给出一种估计光开关成本的统一标准. 基本思路是计算构造各种规模的光开关所需的 1 ×2 或 2 ×1 开关的数量,并称 1 ×2 (或 2 ×1) 开关为一个基本开关单元. 由于 1 ×2 开关的成本一致,故可以按基本开关单元的数量来比较各种交换结构所使用的光开关成本.

假定 $N = 2^n$,则 1 ×$N$ 空分开关可以用 $n$ 级 1 ×2 空分开关构造,其中第 $i$ 级开关共需 $2^{i-1}$ 个 1 ×2 开关,如图 6($a$) 所示. 故构造完整的 1 ×$N$ 开关共需要 $2^n - 1 = N - 1$ 个基本开关单元. 同理,$N$ ×1 开关共需要 $N - 1$ 个基本开关单元. 利用 1 ×$N$ 开关和 $N$ ×1 开关可以按照图 6($b$) 构造 $N$ ×$N$ 开关. 故 $N$ ×$N$ 开关阵所需使用的基本开关单元数量为 $2N(N - 1)$. 注意到图 2 所示的 FDL 结构中,使用的开关数量同样可用基本开关单元的数量来计算. 因此按照这种方式,可以计算出图 2 和图 5 所示的各种交换结构中使用的基本开关单元数目. 结合其它光器件的数目,可以得到表 2 所示的成本比较结果. 注意表中给出的波分解复用器和合波器的数目包括了 FDL 或者 OB 中使用的器件.

## 5 仿真实验

### 5.1 仿真模型及参数

本文仿真实验基于 OPNET 平台. 网络模型如图 7($a$) 所示,包括 4 个边缘节点和 1 个核心交换节点;节点间链路复用了 9 个波长;其中 8 个数据通道(即 $M = 8$),通道速率 1Gbps,1 个控制通道,速率 155Mbps. 核心节点的端口数 $P = 4$,光缓存中 FDL

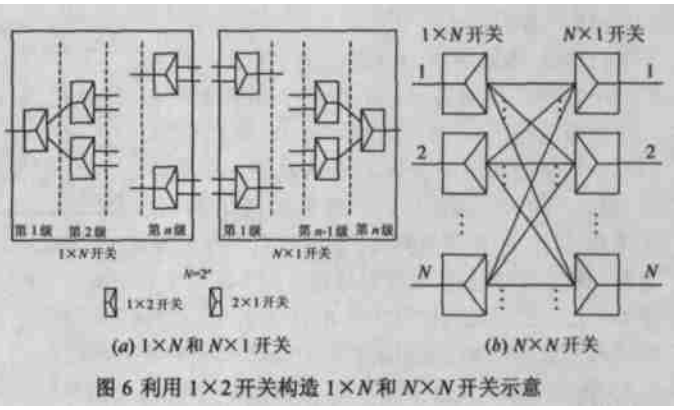

图 6 利用 1×2 开关构造 1×$N$ 和 $N$×$N$ 开关示意

的数目 $N = 8$. 边缘节点模型如图 7($b$) 所示,包括 8 个 IP 源,1 个汇聚器(Assembler),1 个调度器(Scheduler),输出为 8 个数据通道,1 个 BHP 通道. IP 源与汇聚器之间的通道速率为 1Gbps. IP 包长度固定为 1500 字节,IP 包间隔时间服从负指数分布. 所有源发出的 IP 包在汇聚器按目的边缘节点分成不同的队列. 图 7($a$) 所示的网络中,每个边缘节点发出的流最多有 3 个目的边缘节点,故汇聚器中有 3 个汇聚队列. 汇聚算法采用最大突发长度/最大汇聚时间,即累计汇聚队列中 IP 包的长度和汇聚时间,一旦长度达到最大突发长度,或者汇聚时间达到最大汇聚时间,则立刻将队列中已有的 IP 包组装成突发包发送给调度器. 实验中最大突发长度设置为 16500 字节,最大汇聚时间为 120μs. 调度器负责发送 BHP,并确定突发包的发送时间和输出波长. 调度器将收到突发包的时间作为 BHP 的发送时间,然后等待一段时间(偏移时间)后发送突发包,偏移时间在 30～100μs 之间均匀分布. 选择波长时采用负载均衡原则.

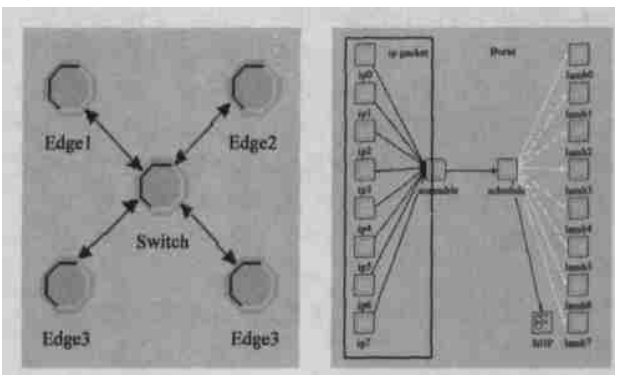

图 7 仿真模型

经过测试,上述模型可以保证业务量模型满足以下要求:

(1) 每个边缘节点的 8 个输出波长上业务量一致,因此核心节点输入端的 32 个波长通道上业务量是一样的.

(2) 每个边缘节点发出的,到其它 3 个边缘节点的业务量是一样的. 同样,每个边缘节点收到的其它 3 个边缘节点的业务量是一致的.

表 2 5 种交换结构所需光器件数目的比较

| | $PM \times 1 \times P$ | $PM \times PM$ - I | $PM \times PM$ - II | $PM \times PM$ - III | $PM \times PM$ - IV |
|---|---|---|---|---|---|
| 基本开关单元 | $PM(P - 1) + P^2MK$ | $2PM(PM - 1)$ | $2PM(PM - 1) + PMK$ | $2PM(PM - 1) + 2PM(N - 1) + PNK$ | $2PM(PM - 1) + 2PM(N - 1) + PNK$ |
| 波分解复用器 | $P$ | $P$ | $P$ | $P + PN$ | $P + PN$ |
| 合波器 | $P + P^2M$ | $P$ | $P + PM$ | $P + 2PN$ | $P + 2PN$ |
| 波长变换器 | $PM$ | $PM$ | $PM$ | $PM$ | $2PM$ |

表中,$P$ 为交换结构端口数目,$M$ 为波长数目,$K$ 为 FDL 级数,$N$ 为 OB 中 FDL 的个数.

在这种业务量模型下,核心交换节点的业务量模型是均匀的,即业务量在所有输入通道和输出通道之间均匀分布.以下各小节中给出的业务量值都是指单个波长通道上的业务量,实际上等价于链路利用率.例如,所谓业务量为 0.8,指的是每个波长通道上的业务量为 0.8,每个端口的业务量为 8×0.8 = 6.4.由于每个波长通道业务量一致,故整个 DWDM 链路的链路利用率仍为 0.8.

本节中,每个仿真数据都是 5 次仿真实验(使用 5 个不同的随机数种子)的平均值,每次仿真实验统计约 1000000 个突发包.本文考察的性能指标主要包括以下 4 个:

(1)突发包丢弃率.定义为调度失败的突发包与总的突发包数目的比值.

(2)平均延迟.定义为所有进入 OB/FDL 的突发包的延迟时间累积值与总的突发包数目的比值.

(3)内部阻塞率.定义为遭遇内部阻塞的突发包数目占突发包总数的比值.遭遇内部阻塞的突发包是指它的出端口和出波长都是空闲的(即没有外部阻塞),但由于其它原因它不得不被丢弃或者延迟.

(4)归一化阻塞率.定义为 1 减去入端与出端链路利用率的比值.一个达到 100 %吞吐的理想交换结构,其入端和出端链路利用率的比值应该为 1,无论是丢弃还是延迟都可能导致出端利用率降低,因此,与丢弃率和平均延迟相比,归一化阻塞率可以更好地反映出交换结构的综合性能.

### 5.2　仿真结果及讨论

**5.2.1　FDL 单位延迟**　首先研究了 FDL 级数一定的情况下,单位延迟时间对交换性能的影响.图 8 给出了 4 种交换结构中,最大延迟级数分别为 1、2、4、8、16 和 32 时,归一化阻塞率随单位延迟变化的曲线.从中不难看出以下几点规律:

(1)无论哪种交换结构,在延迟级数 $K$ 一定的情况下,都存在最佳的单位延迟时间.正如 2.2 小节所分析的那样:一方面,给定 $K$,单位延迟越大,则总的延迟时间越大,交换结构对突发包延迟时间的调整范围越大,避免冲突的可能性越大;另一方面,单位延迟小,调整的精细程度越高,出端的链路利用率也就越高.这两个因素同时起作用,就必然存在一个最佳单位延迟时间.

(2)延迟级数不同,最佳单位延迟不同.总的趋势是:$K$ 越大,最佳单位延迟时间越小.这是因为 $K$ 值较大时,总的延迟时间相应得到了保证,因而决定最佳性能的因素主要是调整的精细程度.

(3)相同入端链路利用率下,交换结构不同,最佳单位延迟不同.如图 8(a)和(b)所示,$K$ = 4 时,$PM \times 1 \times P$ 结构和 $PM \times PM$ - Ⅳ结构的最佳单位延迟分别为 40μs 和 30μs.这说明在业务量相同,且延迟级数相同的情况下,为了达到最佳交换性能,$PM \times 1 \times P$ 结构比 $PM \times PM$ - Ⅳ结构所要求的时间调整精细程度更低.反之,这正好证明了 $PM \times 1 \times P$ 结构对输出端链路的利用率要高于后者.同样的对比可以在 $PM \times PM$ - Ⅲ和 $PM \times PM$ - Ⅱ之间进行(如图 8($c$)和($d$)所示),结论相似.

为进一步比较各种交换结构的成本及其交换性能,后续实验中均假定 $K$ = 4.单位延迟时间为 40μs.根据第 4 节中的成本对比,可以得到各种交换结构所配置的光器件数目,如表 3 所示.

表 3　仿真实验中各种交换结构所配置的光器件数目

| | $PM \times 1 \times P$ | $PM \times PM$ - Ⅰ | $PM \times PM$ - Ⅱ | $PM \times PM$ - Ⅲ | $PM \times PM$ - Ⅳ |
|---|---|---|---|---|---|
| 基本开关单元 | 608 | 1984 | 2112 | 2560 | 2560 |
| 波分解复用器 | 4 | 4 | 4 | 36 | 36 |
| 合波器 | 132 | 4 | 36 | 68 | 68 |
| 波长变换器 | 32 | 32 | 32 | 32 | 64 |

交换结构端口数目 $P$ = 4,波长数目 $M$ = 8,FDL 级数 $K$ = 4,OB 中 FDL 的个数 $N$ = 8

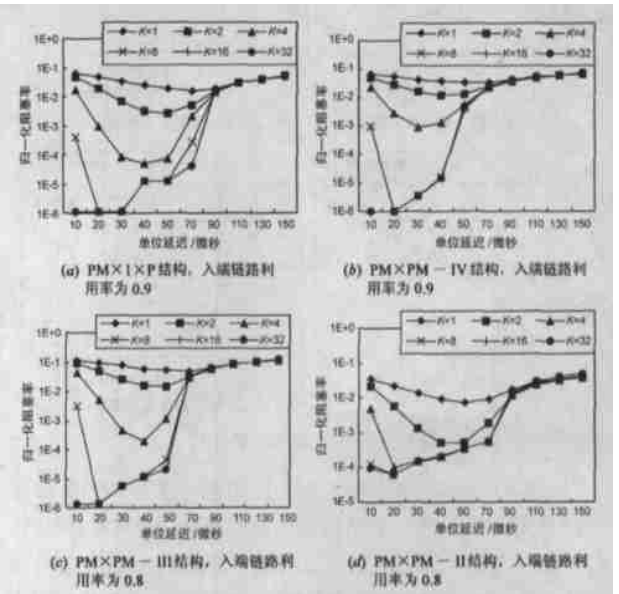


图 8　单位延迟与归一化阻塞率的关系

**5.2.2　交换结构性能比较**　图 9 给出了本文所涉及的各种交换结构的性能比较.其中图 9($a$)比较的是 5 种交换结构的归一化阻塞率;图 9($b$)比较的是突发包丢弃率;图 9($c$)比较了平均延迟,由于 $PM \times PM$ - Ⅰ结构没有配置 FDL/OB,故只比较了其它 4 种结构;图 9($d$)主要比较的是内部阻塞率,根据 3.1 小节的分析,$PM \times 1 \times P$ 和 $PM \times PM$ - Ⅰ结构没有内部阻塞,因此只比较了其它 3 种结构.

从图 9($a$)、($b$)和($c$)可见,$PM \times 1 \times P$ 结构在归一化阻塞率、丢弃率和平均延迟这 3 项指标上都优于其它交换结构,但是优越程度随输入端链路利用率的不同而有所差异.总的趋势是业务量较小时,$PM \times 1 \times P$ 结构在交换性能上的优势并不明显,甚至没有优势.例如图 9($a$)中,输入端链路利用率低于 0.7 以后,$PM \times 1 \times P$ 与 $PM \times PM$ - Ⅲ/Ⅳ结构的归一化阻塞率都低于 $10^{-6}$,已难以区分.而当业务量较大时,$PM \times 1 \times P$ 结构的性能优势很明显.同样以图 9($a$)为例,当业务量为 0.8 和 0.9 时,与性能最接近的 $PM \times PM$ - Ⅳ结构相比,$PM \times 1 \times P$ 结构的归一化阻塞率也要低 1～2 个数量级.丢弃率和平均延迟也存在类似趋势.考虑到这几种结构的成本比较(参见表 3),本文提出的 $PM \times 1 \times P$ 结构在得到相似交换性能的同时,大大降低了成本.例如,从表 3 可见,在光突发交换结构中最为昂贵的高速光开关这一项上,$PM \times 1 \times P$ 结构比性能最接近的 $PM \times PM$ - Ⅳ结构所需要的开关单元数目

少了近 75 %,而在同样昂贵的 TWC 这一项上, *PM* ×1 ×*P* 结构节省了 50 %.

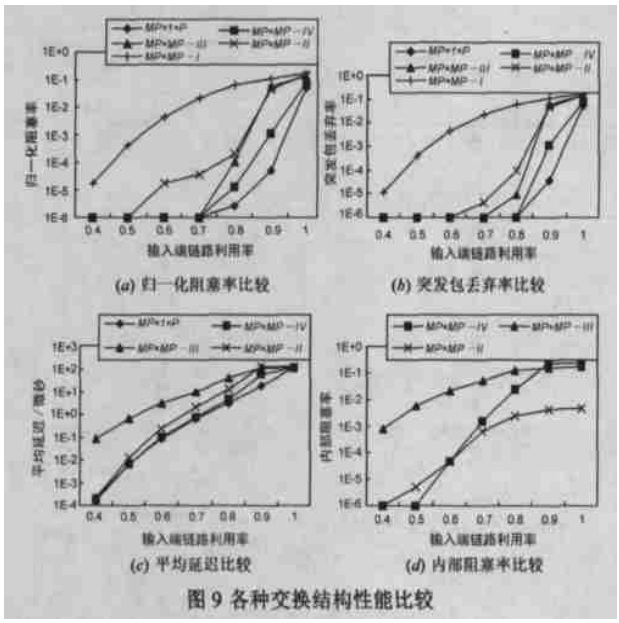


图 9 各种交换结构性能比较

图 9(*d*)给出的内部阻塞比较中,有两个明显的趋势值得研究.首先, *PM* ×*PM* - II 结构的内部阻塞小于 *PM* ×*PM* - III 结构. *PM* ×*PM* - II 结构中内部阻塞主要来源于 FDL 出口冲突,而发生 FDL 出口冲突的突发包都来自相同入端口的相同入波长;与之对应地, *PM* ×*PM* - III 结构中来自相同入端口的所有入波长上的突发包都可能竞争 OB 中某个 FDL 的出口或者某个空分阵入口.因此, *PM* ×*PM* - III 发生内部冲突的可能性要大于 *PM* ×*PM* - II 结构,需要 FDL 以解决冲突的可能性也越大,这一点从图 9(*c*)所示的平均延迟比较中也可以看出.其次, *PM* ×*PM* - IV 结构的内部阻塞在业务量较小时优于 *PM* × *PM* - II(业务量小于 0.6)和 *PM* ×*PM* - III(业务量小于 0.8)结构,但在业务量较大时内部阻塞反而较大.业务量较小时,无论哪种类型的内部冲突都较少发生,因而具有解决 FDL 出口冲突和空分阵入口冲突能力的 *PM* ×*PM* - IV 结构可以有效避免内部阻塞.但是随着业务量的加大,这种解决内部冲突的机制将导致更多的突发包进入 FDL,在效果上反而加剧了内部冲突的可能性.与之相比, *PM* ×*PM* - II 结构由于更倾向于将冲突的突发包丢弃,而不是引入 FDL,因此虽然具有更高的突发包丢弃率,反而具有较小的内部阻塞率.从上述分析不难看出,无论采用什么机制来解决内部冲突,其效果都是有限的,而且在业务量较大时反而可能加剧内部阻塞,这是由 *PM* ×*PM* 结构中输入端缓存机制决定的.因此,真正解决冲突的方式应该是尽量避免输入端缓存.而本文提出的 *PM* ×1 ×*P* 结构通过为每个输入流提供独立的、互不干扰的交换通路,有效地避免了输入端缓存的弊端,提高了交换性能.

## 6 小结

在光突发交换网络中,核心交换节点的设计必须在保证一定交换性能的同时尽量降低光器件的成本.与现有的典型设计方案( *PM* ×*PM* 交换结构及其变型)相比,本文提出的 *PM* ×1 ×*P* 结构更好地达到了上述要求.一方面, *PM* ×1 ×*P* 结构有效地解决了包括 FDL 队头阻塞在内的各种内部阻塞,从而大大提高了交换吞吐量,降低了丢弃率;另一方面,与具有相似交换性能的交换结构相比, *PM* ×1 ×*P* 结构需要的高速光开关数量、TWC 数量、波分解复用器和合波器数量更少,有效降低了成本.总之,本文所提出的交换结构具备结构简单,高性能和低成本的特点,具有广泛的应用前景.

## 参考文献:

[1] M Yoo, M Jeong, C Qiao. A high speed protocol for bursty traffic in optical networks[A]. Proc. of Conf. All-Optical Networking: Architecture, Control, Management Issues[C]. Boston, USA: SPIE, 1997, 1: 79 - 90.

[2] C Qiao, M Yoo. Optical burst switching - a new paradigm for an optical Internet[J]. Journal of High Speed Networks, 1999, 8(1): 69 - 84.

[3] D Blumenthal, P Prucnal, J Sauer. Photonic packet switches: Architectures and experimental implementation[J]. Proceedings of IEEE, 1994, 82: 1650 - 1667.

[4] L Xu, H Perros, G Rouskas. Techniques for optical packet switching and optical burst switching[J]. IEEE Communications Magazine, 2001, 39(1): 136 - 142.

[5] M J Karol, M G Hluchyj, S P Morgan. Input versus output queueing on a space-division packet switch[J]. IEEE Transactions on Communications, 1987, 35(12): 1347 - 1356.

[6] C Qiao. Labeled optical burst switching for IP and WDM integration [J]. IEEE Communications Magazine, 2000, 38(9): 104 - 114.

[7] X Wang, H Morikawa, T Aoyama. Deflection routing protocol for burst switching WDM mesh networks[A]. Proc. of Terabit Optical Networking: Architecture, Control, and Management Issues[C]. Boston, USA: IEEE/SPIE, 2000, 1: 242 - 252.

[8] M Yoo, C Qiao, S Dixit. QoS performance of optical burst switching in IP-over-WDM networks[J]. IEEE Journal on Selected Areas in Communications, 2000, 18(10): 2062 - 2071.

[9] Y Xiong, M Vandenhoute, H Cankaya. Control architecture in optical burst-switched WDM networks[J]. IEEE Journal on Selected Areas in Communications, 2000, 18(10): 1838 - 1851.

## 作者简介:

**王 晟** 男,1971 年生于成都,博士,副教授,目前主要研究方向为通信网与宽带通信技术. E-mail: key. lab @uestc. edu. cn.

**罗蕴翰** 女,1981 年生于成都,硕士研究生,就读于电子科技大学通信与信息工程学院.