OBS边缘节点接收调度模块的硬件实现

戴 睿，胡 钢，李 扬

(电子科技大学 宽带光纤传输与通信网技术教育部重点实验室 成都 610054)

【摘要】给出了一种用于光突发交换网络中边缘节点接收调度模块的电路实现方案。该方案以基于虚拟输出队列机制的公平、高效的交换开关仲裁算法—输入串行为核心，利用两片高速现场可编程门阵列芯片，同时进行6路千兆光突发交换网络数据的接收、交换以及以太网封装。六路数据完全独立，并且两片现场可编程门阵列芯片之间可以相互通信。

关 键 词 调度； 仲裁算法； 交换； 变长输入串行轮询； 光突发交换； 虚拟输出队列

中图分类号 TN929.11 文献标识码 A

Hardware Implementation of the Scheduling Module in OBS Edge Node Receiver

Dai Rui，Hu Gang，Li Yang

(Key Laboratory of Brodband Optical Fiber Transmission and Communication Networks UEST of China, Ministry of Education Chengdu 610054)

Abstract In this paper, a hardware implementation scheme of the scheduling module in Optic burst switching(OBS) edge node receiver is presented. Input serial polling(ISP), which is based on virtual output queuing(VOQ) mechanism, is a kind of fair and high-performance algorithm for crossbar arbitrating. Focused on ISP, the design allows receiving, switching and Ethernet-assembling for 6 routs of 1 000M-OBS data with two high-speed FPGA chips. The 6 routs of data mentioned above are totally independent, and there exists communication between two FPGA chips.

Key words scheduling; arbitrating algorithm; switching; variable length ISP; OBS; VOQ

由于密集波分复用(Dense Wavelength Division Multiplexing, DWDM)技术的日趋成熟，Tb/s 量级甚至更高速的传输网络已初步形成。但网络交换的发展却远远滞后于网络传输，成为满足全球数据业务量几何级数增长需求的最大瓶颈。光突发交换(Optical Burst Switching, OBS)技术克服了基于MPëS(Multiprotocol ë Switching)的光电路交换(Optical Circuit Switching, OCS)技术虚网络拓扑控制复杂的困难，弥补了光分组交换(Optical Package Switching, OPS)技术中光缓存方面(主要是还不存在光存储器)的缺陷，同时融合了上述两种技术的优点，成为当今光网络发展的宠儿[1]。

输入串行轮询(Input Serial Polling, ISP)算法是一种基于虚输出队列(Virtual Output Queuing, VOQ)的调度仲裁算法。与其他crossbar调度算法相比，ISP不但在重负载时带宽利用率高，而且实现简单。文章提出的调度方案实际上是ISP算法在变长交换中的应用。

1 基于变长交换的ISP算法

传统的crossbar采用简单的先进先出(First In First Out, FIFO)输入队列技术，从而不可避免地产生了对头

收稿日期：2004 – 07 – 09
基金项目：国家863计划资助项目(2002AA122021)
作者简介：戴 睿(1980 –)，男，硕士生，主要从事光突发网络方面的研究.

阻塞(Head Of Line, HOL)现象。HOL极大地降低了crossbar的吞吐量。例如当分组到达服从独立同分布的贝努里过程，N趋近于无穷时，crossbar的吞吐量只有58.6%，对于burst分组到达，利用率更低[2]。

与传统的ISP算法不同，本文设计方案提出的ISP算法是面向变长交换的，该算法具体过程如下：

1) 初始化：将RR_i置全0，S置全1，其中RR_i 表示第i个输入端口中优先级最高的VOQ；S表示输出端口状态向量，为0表示输出端口i已经分配给某个输入端口，为1表示该端口还没有被分配。

2) $Ar[k,i]$表示第k次调度中的第i次仲裁，其中一次调度迭代N次，每次迭代称为一次仲裁，一次仲裁只为一个输入端口选择输出端口。每次$Ar[k,i]$包括如下步骤：

(1) 轮询输入端口$P_{\text{in}}[i]=(k+i) \bmod N$；

(2) 判断$S \cap \boldsymbol{req}_i$(第$i$个输入端口的请求向量 $\boldsymbol{req}_i$)，若为空集则将i加1，并返回(1)进行下一次迭代，否则进入步骤(3)；

(3) 根据RR_i，$\boldsymbol{req}_i$以及S为当前输入端口$P_{\text{in}}[i]$选择输出端口$P_{\text{out}}[j]$，并输出In／Out 端口集合$W=\{<i,j>, 0<i, j<N-1\}$，然后将$S[j]$置0并返回(1)操作(若$i=0$，则$RR_i=(j+1) \bmod N$)。

对于S，若输出端口$P_{\text{out}}[t]$($t=0, 1, \cdots, N-1$)中的数据已传送完毕，则释放$P_{\text{out}}[t]$，置$S[t]$为1。

2　系统电路设计

OBS接收调度模块主要完成以下功能：

1) 接收从OBS核心节点传来的突发(burst)包，并将每个burst包拆卸成一个或多个IP包；

2) 对所拆卸的IP包进行交换，根据其目的地址将其分配到不同的端口上；

3) 把已经完成交换的IP包封装成以太网帧，并将其传送到用户端(以太网卡接口)。

本文所使用的现场可编程门阵列芯片(Field Programmable Gate Array, FPGA)内嵌于现场可编程系统芯片(Field Programmable System Chips, FPSC)中。由于每片FPSC仅提供4路数据通道，因此6路独立的数据必须由两片FPGA共同处理。又因为本文所采用设计方案中每片FPGA所完成的逻辑功能完全一致，所以仅给出一片FPGA的系统实现框图。

如图1所示，每片FPGA处理4路数据，其中3路为突发包数据，另一路数据是另1片FPGA传来的IP分组。因此两片FPGA可以相互通信，并协同处理六路突发包数据。下面介绍各子模块实现情况。

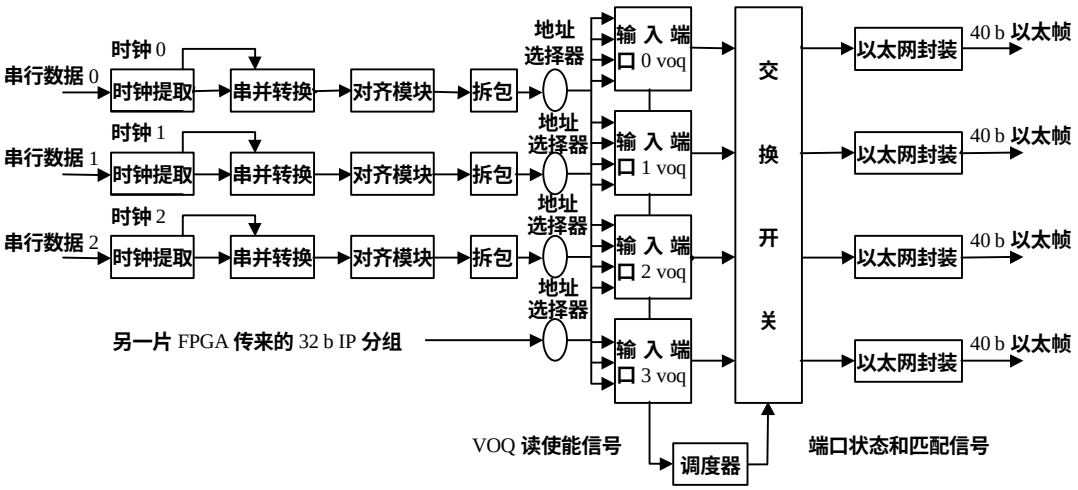

图1　OBS接收调度系统硬件框图

2.1　时钟提取与串并转换模块

OBS网络中的数据流是突发的，而不是连续传送的。这种突发性要求接收端必须快速地从输入数据流中恢复时钟信号。具体来讲，本文设计方案中需要进行时钟提取的锁相环的锁定时间保持在纳秒数量级。因此，为了保证突发接收的高效率，在突发包进入后续操作以前必须经过一个高速锁相环来获得发送端时钟信号，从而做到收发同步。另外，目前的FPGA无法处理1.25 Gb/s的串行高速信号，所以必须通过串并转换将高速串行信号变成低速并行信号才能实现FPGA内部操作。时钟提取与串并转换模块的功能便是从OBS

数据流中恢复出频率为1 250 MHz的发端时钟，然后进行1:40串并转换产生40 b的低速数据流并送往对齐模块。

3.2 突发包对齐(alignment)模块

鉴于突发包到达接收端口时刻的随机性，burst分组在进入拆包模块之前必须进行对齐，否则会极大地增加后续设计的复杂性，降低工作频率。突发包对齐(alignment)模块实际上是一个序列检测器，一旦检测到前导码(见图2突发包格式)便将其放置于40 b数据单元的最高8 b上，即对齐突发包。

保护带 (10 kb)	同步码 (8 b)	前导码 (8 b)	包 长 (24 b)	净荷类型 (8 b)	净荷长度 (16 b)	分组个数 (8 b)	标签 1 (32 b)	IP 包 1	…	保护带 (32 b)	IP 包 *n*

图2 突发包格式

3.3 拆包与封装模块

拆包模块通过判断前导码(突发包起始位置的标志)来确认突发包是否到达接收模块。若突发数据到达，则从包中提取IP分组个数和长度信息，然后将突发包分解成一个个不带标签的原始IP分组。由于芯片内部的数据交换宽度是32 b，因此要把这些分组数据填充成长度为32 b整数倍的IP包以完成后续处理。另外，拆包模块会发送一个标志IP分组起始与结束的信号start_end来避免交换和封装模块对数据包的变模计数，从而简化了逻辑，提高了工作频率。这是本文设计方案的一个特色。

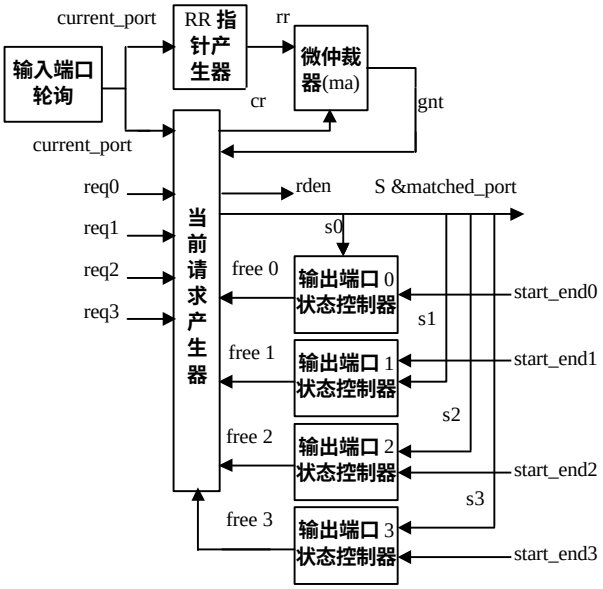

图 3 调度器硬件实现框图

封装模块按照以太网帧的格式将IP包封装成以太网帧。由于以太帧头为22字节，仅为16 b而不是芯片内部数据交换宽度32 b的整数倍，所以存在一个高16位是以太帧头，而低16 b是IP分组的数据单元，因此用宽度围32 b的FIFO是无法正确产生这一数据单元的。本设计方案把完成调度的32 b数据分成高16位和低16位，并分别将它们存入两个宽度为16位的FIFO中，通过控制两个FIFO的读信号来产生这个数据单元，从而解决了这一问题。

3.4 VOQ与crossbar

VOQ用于存放不同输出端口的IP包，所以VOQ的长度与丢包率有直接关系(长度越长丢包率越小)。因为突发包的最小长度为12.5 k字节，每一个输入端口对应3个或4个VOQ，每3个或4个VOQ竞争一个输出端口，那么在带宽利用率为90%的情况下，平均每个VOQ的最小长度为8.44 k字节。crossbar由4个带使能的多路选择器(Multiplexer, MUX)组成。其中有3个4×1多路选择器，1个3×1多路选择器。

3.5 基于变长交换的ISP调度器

图3是变长交换ISP调度器的实现框图。它主要由输入端口轮循 、RR指针产生器 、当前请求产生器 、微

仲裁器以及输出端口状态控制器组成。具体实现过程如下。

3.5.1　输入端口轮询

输入端口轮循模块相当于一个时序发生器，产生本次仲裁所轮循的输入端口。它由两个模4计数器cnt1和cnt2组成。cnt2表示每次调度中迭代的次数，每4个时钟周期加1。cnt1表示当前迭代应首先轮询的端口。当cnt2等于3时, cnt1加1。cnt1+ cnt2表示当前迭代应轮询的端口号。

3.5.2　RR指针产生器

RR指针产生器由4个两位的RR指针寄存器以及一个4×1多路选择器构成。4×1MUX根据当前迭代端口号选择相应的RR指针送微仲裁器。当迭代次数为0时，RR指针产生器根据输入／输出匹配结果，按照ISP算法修改相应的RR指针。

3.5.3　当前请求产生器

当前请求产生器根据从VOQ传来的请求信号(re) 、当前轮询端口信(rr)及输出端口的状态信息(s)产生该次仲裁的请求信号(cr)。同时通过微仲裁器发来的仲裁信号(gnt)和输出端口状态控制器给出的释放信号(free)，修改In／Out端口状态和匹配信号(s & matched port)，并控制相应VOQ的读使能信号(rden)。信号(cr)有效的条件是req与s均有效。s有效(输出端口空闲)的条件是free有效而gnt无效。而当gnt有效时输出端口忙，s无效。

3.5.4　微仲裁器

微仲裁器主要根据rr与cr信号实现函数ma，从而产生gnt。本文设计方案使用了一种叫温度计编码的编码器[3]，这种编码器对最高优先级作特殊的编码，并以此预处理输入请求，从而巧妙地消除了可编程优先级编码器的优先级可编程性，并且硬件实现简单，节约资源。

3.5.5　输出端口状态控制器

前面已经提到，本文设计方案使用了一种类似于帧定界的方法，它利用拆包模块发送的一个名为start_end的信号来标志IP分组的起始与结束。当start_end为“10”时状态控制器开始工作；当start_end为“01”时表明IP包传送完毕，此时发送free信号释放相应的输出端口。这种做法实际上替代了变模计数器的作用，同时简化了逻辑，提高了工作频率。

3.5.6　性能分析及仿真结果

文献[2]指出，对于keep full到达过程，ISP算法的交换带宽利用率为100%；在独立同分布或burst到达条件下，ISP算法的延时和信元延迟的均方差小于其他算法。这说明ISP在带宽利用率 、延时和公平性方面都是优秀的，特别是在轻负荷和端口数较小的情况下，ISP性能最优。图4为本文设计方案中调度器模块布局的布线后仿真波形图。

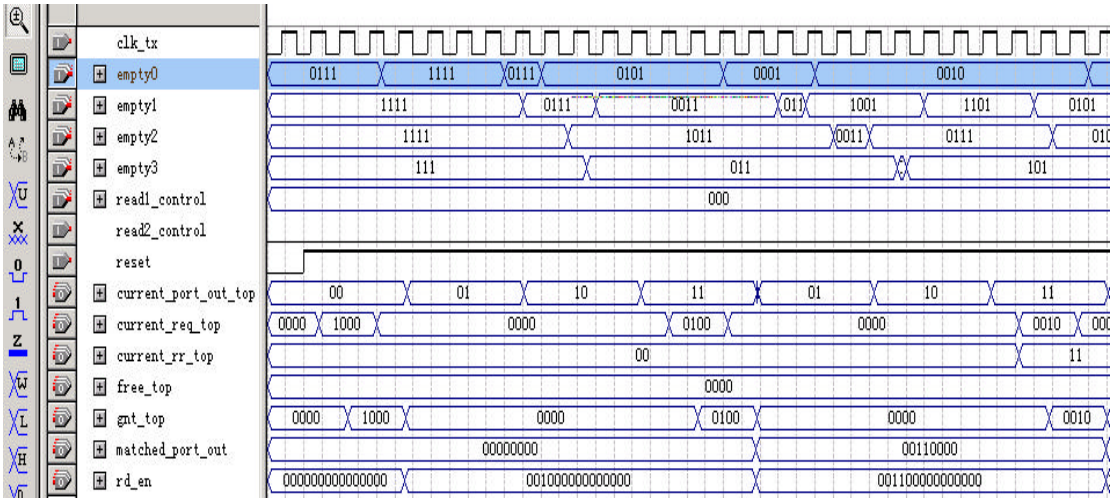

图4　调度器布局布线后仿真时序图

(下转第705页)

参　考　文　献

[1] 于金辉, 范　戈. 光突发交换技术[J]. 光通信技术, 2002, 26(5)：38-40

[2] Berger L. Generalized multi-protocol label switching (GMPLS) signaling functional description.rfc3471[DB/OL]. www.ietf.org, 2003-01-15

[3] Berger L. Generalized Multi-Protocol Label Switching (GMPLS) Signaling Constraint-based Routed Label Distribution Protocol (CR-LDP) Extensions. rfc3472[DB/OL]. www.ietf.org, 2003-01-15

[4] 吴　伟. 多协议标签交换[M]. 北京：清华大学出版社, 2002

[5] 陈俊峰, 李新琬, 吴龟灵,等. 光突发交换网络的突发包组装和调度[J]. 光通信技术, 2003, 27(2)：11-14

[6] 张　骞, 庞湘绮, 文爱军. 基于GMPLS的新一代网络流量工程[J]. 计算机网络世界, 2003, 12(3)：71-73

[7] 司　昕, 施杜平, 罗忠生. 基于GMPLS的光网络保护和恢复机制[J]. 高技术通讯, 2003, 13(9)：10-15

编　辑　熊思亮

(上接第693页)

4　总　结

ISP算法是一种高效、公平、简单的交换仲裁算法。通过理论分析和电路综合仿真结果可以发现ISP算法在轻负荷和端口数较小的情况下是目前所有同类算法中性能最优的算法。本文设计方案调度模块中每片FPGA分别处理并输出4路数据，即有4个输入／输出端口，符合端口数小的条件，适合使用ISP算法。

参　考　文　献

[1] 罗洪斌, 胡　钢, 李乐民. 光突发网络边缘节点突发排队方案[J]. 电子科技大学学报, 2003, 32(3): 289-291

[2] 孙志刚, 苏金树, 卢锡城. 高效的crossbar仲裁算法—ISP[J]. 计算机学报, 2000, 23(10): 1 078-1 082

[3] 彭来献, 郑少仁. 基于iSL IP算法的高速Crossbar调度器的FPGA设计与实现[J]. 解放军理工大学学报, 2001, 2[6]: 32-36

编　辑　熊思亮

(上接第696页)

参　考　文　献

[1] AS1773，Fiber optics Mechanization of a Digital Time Division Command Response/Multiplex Data Bus[S]. Washington，SAE，1995

[2] Su Chao , Chen Liankuan, Cheung Kwokwai. Theory of burst-mode receiver and its applications in optical multi-access networks[J]. Journal of Lightwave Technology ,1997,15 (4) :589-606

[3] Yusuke O, Robert G S. High-speed burst-mode packet-capable optical receiver and instant aneous clock recivery for optical bus opreation[J]. Journal of Lightwave Technology ,1994 ,12 (2): 325-330

[4] Topliss S, Beler D, Altwegg L. Synchronization for passive optical networks[J]. Journal of Lihgtwave Technology ,1995 ,13 (5): 947-953

编　辑　熊思亮