流水线技术在OBS调度模块中的应用

张 平, 胡 钢, 胡 明

(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】根据波长资源预约的原理，在光突发交换边缘节点的设计中讨论了波长状态表在其中的应用。重点阐明了流水线技术在状态表的筛选和改写中的应用及其FPGA实现。分析结果表明，流水线技术能够适应高速实时光突发交换边缘节点的调度任务。

关 键 词 光突发交换; 边缘节点; 流水线; 状态表

中图分类号 TN929.11 文献标识码 A

Application of Pipelining Technology in Scheduler Module of OBS

Zhang Ping, Hu Gang, Hu Ming

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

Abstract According to the idea of wavelength reserving,the application of the state table of wavelength is discussed in this paper. In this way,fair,on-chip resource and frequnece is considered in design of optical burst switch edge node.Through comparation,pipeling technology can be applied in this design successsfully.Finally the paper illustrate implementation of pipeling technology in FPGA device.

Key words optical burst switch; pipelining; edge node; station table

近年来，光网络的发展中出现了两个非常明显的趋势：动态波长预留(伴随着波长预留时间的减少)和将多协议标签交换(Multiprotocol Label Switch, MPLS)引入光网络，即多协议波长交换(Multiprotocol λ Switch, MPλS)。随着这两大趋势的不断发展，仅仅利用基于MPλS的光电路交换技术来传输突发数据(比如IP数据)遇到了诸如缺乏光RAM，光分组交换技术不成熟的一些困难。因此融合了光分组交换技术和光电路交换技术的一种新的技术就应运而生，这就是光突发交换(Optical Burst Switch, OBS)。在配备光缓存的交换机内，对于处理突发竞争问题的交换控制策略大体分为预约方式(Tell And Go, TAG)、固定延迟方式(Reserve a Fixed Delay, RFD)和充裕时间方式(Just Enough Time, JET)，研究表明交换模式和光缓存的配置方式对系统的性能有重要的影响[1, 2]。实际设计中采用预约线路资源并且设置偏置时间的JET交换控制策略得到了广泛的应用。本设计同样采用了JET策略，着重讨论了与之相适应的“流水线”技术在“波长状态表”的筛选和改写中的应用以及硬件实现。

1 设计梗概

OBS边缘节点调度模块的结构如图1所示。图中各个模块的功能为：

1) R-R模块：该模块采用(Round-Robin, R-R)(时间片轮转)算法按照优先级的分类来对输入端口的请求信息进行轮询，选出当前时刻最需响应请求队列。

收稿日期：2004 – 07 – 09
基金项目：国家863计划基金资助项目(2002AA122021)
作者简介：张 平(1977 –)，男，硕士生，主要从事光突发交换网络方面的研究.

2) 调度模块：其主要功能是对各个波长状态表进行实时比较，筛选出最短波长缓存，并根据发送波长缓存的状态进行改写。该部分的运行效率和资源占用情况直接影响到整个调度模块的工作效率。

3) Switcher模块：完成数据从入端口到波长发送缓存的交换任务。该模块包含两个主要部分：状态机(用来实现对CrossBar的打通过程)，FIFO(用来存储各个端口的输入请求)。

4) BHP模块：完成对相应数据包的突发头控制包(Burst Head Package, BHP)数据的组装。

5) CrossBar模块：主要完成数据包的交叉选通工作。

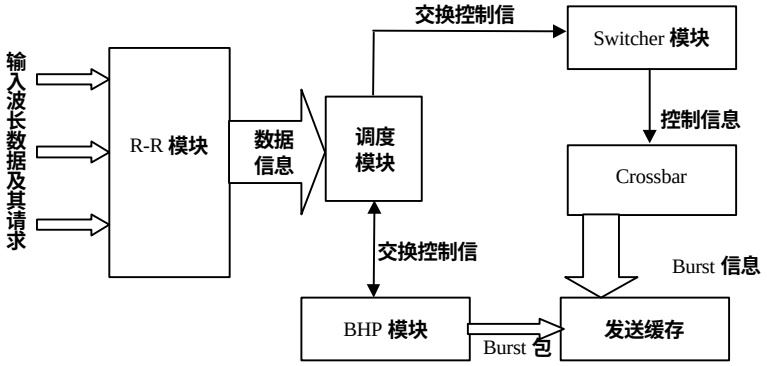

图1　调度模块内部组成

整个设计中最核心的内容就是调度模块对各个光波长发送缓存的存储状态的判别。存储状态包括当前包的长度信息，发送等待时间，发送状态等一系列重要数据。通常定义第i个波长的状态表的数据存储格式如图2所示。从图中看出，状态字说明该波长发送缓存是否为禁用；发送缓存长度为整个发送缓存当前的存储数据量；第i个突发包的长度(i=1, 2, …, N, N的取值取决于发送缓存可以容纳的最大突发包数目)。

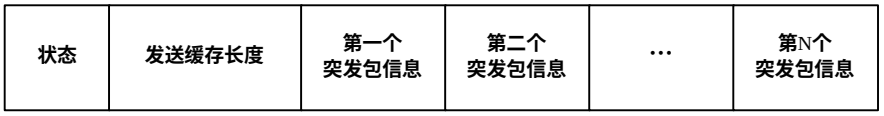

图2　状态表数据存储格式图

为了体现输入突发包的优先级，并且防止发送缓存的“溢出”，调度模块在系统时钟的控制下实时对状态表的“发送缓存长度”信息进行侦测，筛选出最短发送缓存，将当前最高优先级突发包交换入该发送缓存并及时将这一操作记录入该波长状态表，操作流程如图3所示。图中各项分别为：

1) 轮询状态表各表项：仲裁模块轮询查看各个状态表的“发送缓存长度”信息并予以比较，筛选出当前最短的波长发送缓存。

2) 根据筛选信息，将当前数据包存入该发送缓存：该部分的主要功能是就是调用调度模块中的交换单元完成突发数据包从端口到波长的交换任务。

3) 改写相应状态表的值：刷新状态表信息，以便于调度模块能够在下个时钟周期进行同样的数据交换操作。

4) 根据状态表相应信息组装预约资源的BHP包：通过对状态表中对应包的对应位置，测算出准确的偏置时间T_{offset}，并将此信息存入BHP包来预约光路资源。

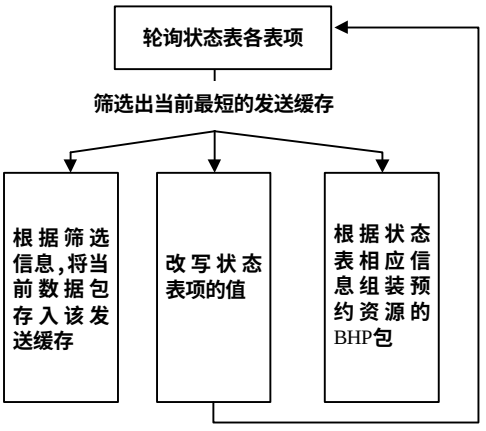

图3　状态表仲裁流程

由于该表单的规模随着输出缓存的容量，输出波长的数目的增多而增大。因此筛选的效率就是制约整个仲裁模块运行效率的关键。

2 常规的设计方式

通常的设计会对N个波长发送缓存的状态表的对应表项直接进行轮询和筛选。在Xilinx系列FPGA的结构中，最基本逻辑单元是Slices，每个Slices包括LUT(Look-Up-Table)，D触发器和相关逻辑。类似表项比较的这种组合逻辑就是通过LUT来实现的。

假设有N个波长，对应于这些波长的状态表项设为$T_1, T_2, \cdots, T_N$，则采用如下的算法进行筛选：

if $T_1<T_2$ and $T_1<T_3$ and $\cdots$ and $T_1<T_N$ then

if $T_2<T_1$ and $T_2<T_3$ and $\cdots$ and $T_2<T_N$ then

$\vdots$

if $T_N<T_1$ and $T_N<T_2$ and $\cdots$ and $T_N<T_{N-1}$ then

如果$N=8$，每个突发包的最大长度为2^{14} b。ISE6.1i进行后仿真的实际工作频率只能达到68 Mbps。远远不能满足设计要求的80 Mbps的频率指标的。

千兆以太网实际应用中一个非常重要的系统指标就是交换速率。对于波长状态表这种庞大的数据直接以上面所述的方式进行比较或者改写等操作，将会消耗很多的片上资源，且时序也会出现不稳定态。考虑到FPGA内部触发器延迟是由芯片本身属性所决定的，因此可以通过减少组合逻辑延迟来提升系统最终的交换速率，使之符合时序要求。

经过细致的比较和对硬件特性的分析，最终选用了“流水线”技术来完成表项的比较和改写动作。

3 “流水线”技术的机理

流水线技术能有效提高系统工作频率。通常的做法是将较大的组合逻辑进行分解，并且插入级间寄存器(用以暂存上级数据和合理控制时序)，这也是所谓“流水线”(PipeLining)技术的基本原理。如果某个设计的工作流程可分为若干步骤，而且整个数据处理是“单流向”的，即没有反馈或者迭代运算，前一个步骤的输出是下一个步骤的输入则可以考虑采用流水线设计来提高系统的工作频率。

“流水线”操作分为“异步流水操作”和“同步流水操作”两种模式，流水结构如图4所示(S_1,S_2,S_3分别代表状态1，状态2，状态3)：

1) 异步流水操作模式：通过相邻两级间的“握手”信息对数据的流向进行控制。“握手”信息包括上级对下级的准备信号和下级正确读取数据后的反馈信号。其操作流程如图4a所示；

2) 同步流水操作模式：需要在每一级插入时钟控制的寄存器，所有的存储器都同步的将上一级的数据加以锁存，每一级通常都是组合逻辑电路。系统的工作时钟频率是由最大级延迟决定的。其操作流程如图4b所示[4]。

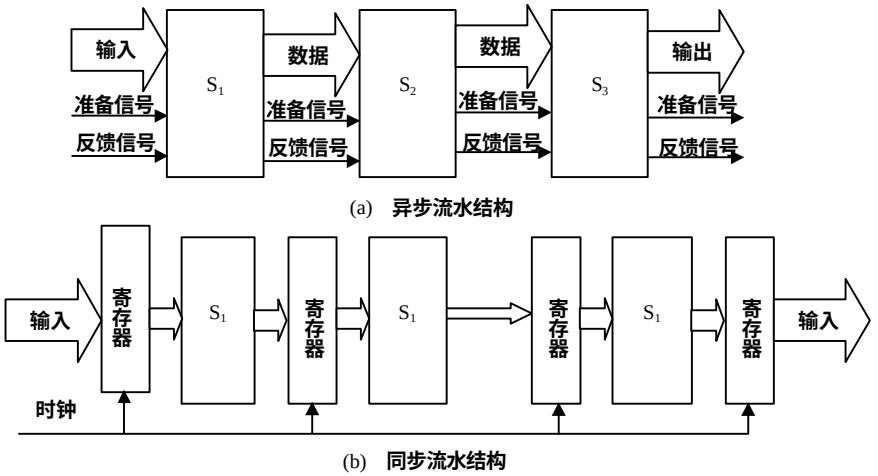

图4 异步流水与同步流水结构

本设计主要是通过“ PipeLining ”来进行逻辑分割和做到精确的时序配合，因此 “ 同步流水 ”操作。以下是同步流水技术的性能指标：

假设t_i是组合逻辑级Si的延迟，$t_m=\max(t_i)$是最大级间延迟。t_r是寄存器延迟。那么流水线的时钟周期t由该式决定：$t=t_r+t_m$；流水线的最大工作频率为：$f=1/t=1/(t_r+t_m)=1/(\max(t_i)+t_r)$。

从全局来看，“ 流水线 ”是一种通过对简单组合逻辑进行复制来提升整个系统工作频率，并且做到各个模块进行相同操作时无等待延迟的一项技术。流水线操作的关键就是整个设计时序的严格匹配。

4　硬件实现

由于必须在系统时钟上升沿对存储于状态表中的发送缓存长度信息进行比较或者改写，这就要求以上的操作和系统时钟严格匹配。但在FPGA中过大的组合逻辑延迟是不确定的，如果在本设计的仲裁模块中对所有的状态表信息作比较，则造成的不确定延迟将会严重降低整个系统的工作频率，以致造成时序紊乱而不能正常的工作。在这样的前提下，对整个的比较判断逻辑进行了如图5所示的分解操作。

为了做到每一级运算的延时在系统可预知的范围内，通常要在级与级之间插入寄存器单元，整体运算时间是一个可预知的固定时间$N\times T_{clk}$(N：分割后的运算级数；T_{clk}：系统时钟周期)。通过实际操作的后仿真，使系统工作频率达到了170 Mbps，同时也达到了既定的设计目标。

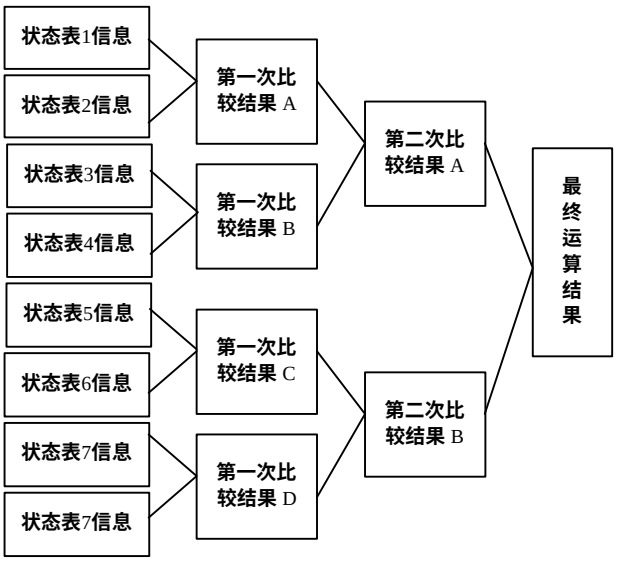

图 5　“ 流水线 ”操作示意图

5　结 束 语

本文主要讨论了OBS设计中对应于JET交换控制策略的的硬件实现方法——流水线方式。重点讨论了“ 流水线 ”方式实现的机理和FPGA的硬件实现方法。该设计在Xilinx Virtex-II Pro系列FPGA上实现。通过实际的仿真分析，采用“ 流水线 ”技术能够有效的提高系统工作频率，避免因级间延时不确定而造成的系统时序紊乱现象。

参 考 文 献

[1] 罗洪斌，胡　钢，李乐民. 光突发交换边缘节点突发排队方案[J]. 电子科技大学学报, 2003, 32(3): 289-292

[2] Qiao C, Yoo M. Choices, features and issues in optical burst switching[J]. Optical Network Magazine, 2000, 1(2): 36-44

[3] Manolis D, Georgios P, Dimitrios S I, *et al*. Nikos chrysos iCS-Forth[C]. 2003: 1-8

[4] Muhamed M. Principles of pipelined design[EB/OC]. http://www.cs.aucegypt.edu/~mudawwar/csci530, 2003-1-05

编　辑　刘文珍