OBS 网络中的多 BHP 延迟调度算法

姚 锐 王 晟 徐世中 虞红芳
(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 光突发交换 (Optical Burst Switching，OBS) 是下一代光网络中的有效核心交换技术之一。在OBS网络中，核心节点的调度方案是影响网络交换性能的重要因素。如何有效地对突发分组 (Burst) 进行合理调度，以减少丢包率正是关键问题所在。该文提出了一种多控制分组 (Burst Header Packet，BHP) 延迟调度算法，对于一个收集周期内到达的多个BHP所对应的突发数据，按照一定的方式对它们进行集中调度，实现对已预约资源的更改，使调度结果得到了优化。该算法根据不同的网络应用，可细化为 4 种具体的算法实现，以适用于不同的应用场景。仿真结果表明，该文提出的调度算法和已有算法相比，能够优化对突发分组的调度，提高对信道资源的有效使用，从而降低交换网络的丢包率。

关键词 光突发交换，多 BHP 延迟调度，可行实例，偏移时间
中图分类号：TN915.63 **文献标识码**：A **文章编号**：1009-5896(2006)01-0163-05

Multiple BHPs Delayed Scheduling Algorithm in Optical Burst Switching Network

Yao Rui Wang Sheng Xu Shi-zhong Yu Hong-fang
(*Key Lab of Optical Fiber Transmission and Communication Networks, University of Electronic Science and Technology of China, Chengdu* 610054, *China*)

Abstract Optical Burst Switching (OBS) is considered as one of core technologies for next generation optical switching networks. In OBS network, the scheduling algorithm for core router is important for the performance of switching network. And how to schedule bursts efficientlly and reduce the lost probability of bursts are the key issue. This paper proposes a new multiple BHPs (Burst Header Packets) delayed scheduling algorithm. With this algorithm, data bursts of all BHPs collected in one collection period are schedued collectively to a policy. And resources has been arranged before for the former BHP can be altered and rearranged by the latter one, and the scheduling result is optimized. In the light of different network application, this algorithm is divided into four realities. Through simulation and comparing the results with now existing algorithms, it is showed that the algorithm can optimize the scheduling of bursts, improve the channel utilization and decrease the loss possibility of bursts.

Key words Optical Burst Switching (OBS) , Multiple BHPs delayed scheduling, Feasible instance, Offset

1 引言

近年来，网络中的业务数据量呈现出急剧增长的态势，因此对下一代骨干网提出了具备极高容量的要求。随着网络带宽需求的增加，波分复用 (WDM) 技术也被越来越广泛地采用。WDM 光网络中实现交换与传输有 3 种方案：光路交换 (Optical Circuit Switching，OCS) ，光分组交换 (Optical Packet Switching，OPS) 和光突发交换 (Optical Burst Switching，OBS) 。由于传统的光路由交换不能有效地利用网络带宽，而光分组交换缺乏高速的光逻辑器件和光缓存等，仍处于研究阶段，因此光突发交换作为一种居中方案被提了出来。它融合两者的优点，被认为是下一代光网络中的核心交换技术之一。

在 OBS 网络中，突发数据 (Burst) 和控制分组 (BHP) 分别在不同的物理信道上进行传输，控制分组先于对应的突发数据一定偏移时间 (Offset) 发送。由于业务是随机到达的，突发数据之间可能存在时间上的重叠，如果不通过合理的方式处理则会引起丢包。当业务量较大的时候，冲突情况会变得很严重。核心路由器上的调度算法正是要在这段偏移时间内通过对 BHP 的调度，为其对应的突发数据提前预约好资源，从而实现突发数据的全光交换。而控制分组需要经

2005-05-26 收到, 2005-09-02 改回
国家“863”计划重大项目(2002AA12202)和电子科技大学科研启动费(Y02002010101016)资助课题

过光电转换到电域上，经过预约资源和处理后再变换到光域进行传输。OBS 网络中，处理冲突最常用的手段包括光纤延迟线 (Fiber Delay Line，FDL) ，可调谐波长变换器 (Tunable Wavelength Converter，TWC) 。FDL 通过对突发数据进行一定的延时来避免冲突，TWC 则通过将其变换到其它波长从而避免在本波长上出现的冲突。

不少文献针对OBS网络提出了一些调度算法，例如 Tell-And-Go protocol (TAG)[1], Just Enough Time protocol (JET)[2], Horizon scheduling [3], and Just In Time protocol (JIT) [4]。现有文献提出的OBS调度方案多采用“顺序调度”的策略，即核心节点按照BHP到达的先后顺序逐个进行调度，调度的结果被携带在BHP中发送至下一节点。先完成调度的BHP所预约的资源不能更改，也不能抢占。这种方式下的调度效率较低。本文提出一种新的多BHP“延迟调度”算法，可以有效解决这一问题。采用延迟调度的核心节点控制模块定期收集所有端口上到达的BHP，其重复周期称为收集周期。收集周期可以固定配置，也可以根据业务量情况动态调整。一个收集周期内收到的所有BHP集中统一调度，因此这些BHP所需要的资源可以被更加有效地安排，使有限的交换资源得到更最充分的利用。

本文后面的章节是这样安排的，第 2 节介绍提出的调度算法及 4 种具体的算法实现，第 3 节给出仿真和结果分析，最后一节为结论。

2 多 BHP 延迟调度算法

多 BHP 延迟调度算法的主要思想是，在每个输出端口上，对一段时间内 (BHP 收集周期)到达的所有 BHP 均延迟到该周期结束的时候，按照其所对应突发数据的各种可能的延迟安排进行优化调度。对突发数据的调度有时间段约束。每个突发数据都有一个到达时间，但是其需要去往的出端口波长通道在该时刻可能并不空闲，因此突发数据的起始调度时间 (下限) 取决于数据的到达时间和相应波长通道的最后可用时间二者中较大的那个值。另一方面，由于光缓存中 FDL 的最大延迟有限，因此每个突发数据的发送都有一个最后期限 (上限) 。我们称在这样一个时间区间内，对一个突发数据的一个可能的安排为一个可行实例 (Feasible instance，见图 1) 。一个突发数据对应于多个可能的安排，亦即多个可行实例，其个数取决于 FDL 的最大延迟级数。而一个 FDL 单元的延迟级数是有限的，因此突发数据对应的可行实例都是有限的。可见，是否配置 FDL，配置多少，只影响可行实例的个数，对具体算法流程并无影响。如果突发数据不能在这个时间段之间调度成功，则会被丢弃。

本算法思想按照是否配置 TWC 和突发数据是否有权重分为 4 种具体调度算法。在没有配置 TWC 的情况下，从一

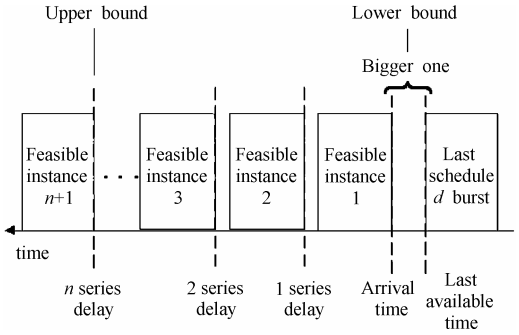

图 1 可行实例的调度

Fig.1 Scheduling of feasible instance

个波长通道到达的数据只能交换到出端口的相同波长上。在一个出端口收集到的 BHP 首先按照出波长进行分类 (图 2 中的分类缓存器模块) ；而对于配置了 TWC 的情况，由于可以进行波长变换，BHP 则只按照出端口收集和集中调度，不需要按照波长分类了。另外，在有的网络应用中需要考虑业务等级，此时的业务分组是带有权重信息的。对于突发数据有权重的情况，前面已经预约的资源可以按照一定的规则 (比如，使成功调度的数据总权重最大) 进行更改；当然，没有权重的情况下是不允许修改的。

对于数据的调度和对资源的安排和更改都在图 2 中的调度器模块完成。对所有突发数据对应的可行实例，按照结束时间早晚的顺序调度。由于每一个突发数据只能有一个调度结果，因此一旦一个可行实例被采纳，就表示该突发数据被成功调度，该数据的其他可行实例也就不再被考虑了。另外，收集周期是和偏移时间紧密相关的，而偏移时间的确定又与受到业务流所能忍受的最大延迟，边缘节点的缓存深度，以及突发数据必须穿过的核心节点的数目等因素的制约。我们假设能够接受的最大偏移时间为L_{OFFSET}，数据必须穿过的最多节点个数为H_{MAX}，那么，在一个核心节点中，BHP的最大延迟为 $L_{\mathrm{BHP}} = L_{\mathrm{OFFSET}}/H_{\mathrm{MAX}}$ 。我们用T_{BHP}来表示BHP在一个核心节点停留的总时间，如果T_{BHP}大于L_{BHP},那

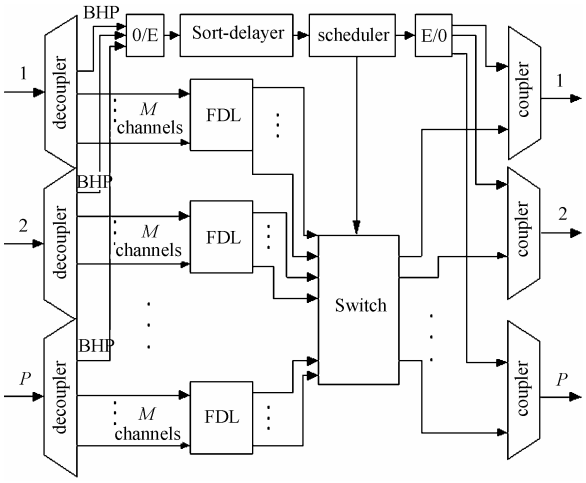

图 2 OBS 交换网络结构

Fig.2 Structure of OBS network

么很有可能当数据包到达核心节点时，对它的调度还没有完成。收集周期的选择不能超过该停留时间。

在本文中，假定核心节点有P个输入/输出端口，每端口复用M个数据波长。光信号在输入端口经波分复用器解复用后分为M路光纤接入空分交换阵；经空分阵交换后在输出端口通过合波器复用到一根光纤上 (见图 2) 。核心节点采用类似LAUC的方式[6]记录资源预约状态，即每个输出端口的每个波长通道只记录一个最后可用时间。FDL采用文献[7]中给出的FDL池，且具有足够多的FDL单元。令某个输出端口上待调度的一批突发数据构成一个集合$B=\{B_1,B_2,\cdots,B_n\}$，可供使用的波长通道集合为$X=\{C_1,C_2,\cdots,C_M\}$，相应于每个输出波长通道的最后可用时间集合为$T=\{T_1,T_2,\cdots,T_M\}$。以下分别描述 4 种具体算法实现。

2.1 无 TWC，不考虑权重时的具体调度算法 (WI_UW)

由于没有配置 TWC，不能进行波长转换，每个突发数据只有一个固定输出波长通道，在每个波长上的调度是独立的，对每个输出波长的算法：

(1) 每个收集周期内，到达一个 BHP 仅提取相关信息，将其可行实例集合插入一个按出端口、出波长分类的队列$B(i)$中等待, 直到一个收集周期结束；

(2) 在第 i 个波长上，将 $B(i)$按照 FDL 级数建立可行实例集合 $I(i)$；

(3) 对$I(i)$中开始于当前波长最后可用时间T_i之前的可行实例进行删除和更新；

(4) 对结束时间最早的可行实例进行调度，突发数据放入调度成功集合，并从$B(i)$中删除，其对应的所有可行实例也都从$I(i)$中删除，更新T_i；

(5) 重复步骤(3)，(4)，直到所有突发数据都调度完成($B(i)$为空) 或者可行实例集合 $I(i)$为空。

2.2 配置 TWC，不考虑权重时的具体调度算法 (WC_UW)

由于配置了 TWC，一个突发数据可以在 M 个输出波长通道上选择，去往相同出端口的各个波长通道上的所有突发数据进行统一调度。对每一个输出端口的算法：

(1) 每个收集周期内，到达一个 BHP 仅提取相关信息，将其可行实例集合插入一个按出端口分类的队列 B 中等待，直到一个收集周期结束；

(2) 将 B 按照 FDL 级数建立可行实例集合 I；

(3) 对I中开始于当前波长i (波长 1 开始) 最后可用时间T_i之前的可行实例进行删除和更新；

(4) 对结束时间最早的可行实例在当前波长进行调度，突发数据放入调度成功集合，并从B中删除，其对应的所有可行实例也都从I中删除，更新T_i；

(5) 重复步骤(3)，(4)，直到该波长调度完成 (B 为空或者可行实例集合 I 为空) ；

(6) 当 $i<M$ 时，转到步骤(2)对 i+1 波长继续进行调度，直到所有波长都已经安排完毕 ($i>M$)或者 B 为空。

2.3 无 TWC，考虑权重时的具体调度算法 (WI_W)

由于每个突发数据都有一个权重，调度算法的优化目标是使权重函数F_w最大。F_w定义为突发数据权重之和。对每一个输出波长的算法：

(1) 每个收集周期内，到达一个 BHP 仅提取相关信息，将其可行实例集合插入一个按出端口、出波长分类的队列$B(i)$中等待, 直到一个收集周期结束；

(2) 在第 i 个波长上，将 $B(i)$按照 FDL 级数建立可行实例集合 $I(i)$；

(3) 对结束时间最早的可行实例进行调度，如果该可行实例f_i与相同收集周期内，之前已经安排的一个或者多个可行实例集合S发生冲突，则比较权重函数。如果$F_w(f_i)>F_w(S)$，则取消S集合的安排，f_i调度成功，从$B(i)$中删除对应的突发数据，其对应的所有可行实例也都从$I(i)$中删除；否则，直接从$I(i)$中删除f_i；

(4) 重复步骤(3)，直到所有突发数据都调度完成($B(i)$为空) 或者可行实例集合 $I(i)$为空。

2.4 配置 TWC，考虑权重时的具体调度算法 (WC_W)

由于配置了 TWC，需要对每个出端口进行集中调度，且每个突发数据都有一个权重。对每一个输出端口的算法：

(1) 每个收集周期内，到达一个 BHP 仅提取相关信息，将其可行实例集合插入一个按出端口分类的队列 B 中等待，直到一个收集周期结束；

(2) 将 B 按照 FDL 级数建立可行实例集合 I；

(3) 在i波长上 (波长 1 开始) ，对结束时间最早的可行实例f_i进行调度，如果与相同收集周期内，之前已经安排的一个或者多个可行实例集合S发生冲突，则比较权重函数。如果$F_w(f_i)>F_w(S)$，则取消S集合的安排，f_i调度成功，从B中删除对应的突发数据，其对应的所有可行实例也都从I中删除；否则，直接从I中删除f_i；

(4) 重复步骤(3)，直到该波长调度完成 (B 为空或者可行实例集合 I 为空) ；

(5) 当 $i<M$ 时，转到步骤(2)对 i+1 波长继续进行调度，直到所有波长都已经安排完毕 ($i>M$)或者 B 为空。

3　仿真和结果分析

仿真实验在 OPNET 网络仿真平台上完成。网络采用内部无阻塞的核心交换阵，每端口复用 8 个数据波长，1 个控制通道。数据通道速率为 1Gbps，FDL 级数分别为 0，2，8，16，每级延迟为 50μs，边缘节点为 3 个 (见图 2，其中 P=3，M=8) 。我们分别仿真了上面提出的 4 种具体算法实现，研究了丢包率和收集周期的关系。

(1) WI_UW 算法 仿真参数为突发数据长度为 130kb (相应的持续时间为 130μs) ，偏移时间满足 (900，2000) μs 均匀分布，到达间隔服从负指数分布，业务强度为 0.8，BHP 收集周期 100~700μs。

仿真结果见图 3。横坐标为收集周期，纵坐标为丢包率。其中采用 BHP 顺序调度策略时 (见横坐标最左边一个点) ，BHP 是即时处理的，并不存在收集周期。从仿真结果可以看出，多 BHP 延迟调度算法中的 WI_UW 算法，其丢包率性能优于顺序调度算法，随着收集周期的增加，丢包率呈逐渐下降趋势；FDL 级数越大，丢包率越低。当收集周期为 700μs 的时候，在不同的 FDL 级数下，丢包率比顺序调度平均下降了 33%。

(2) WC_UW算法 仿真参数同前。仿真结果见图 4。其中，当FDL级数为 16，收集周期 300μs以上时，丢包率已经小于 10^{-6}。从仿真结果同样可以看出，多BHP延迟调度算法中的WU_UW算法，其丢包率性能优于顺序调度算法；由于配置了TWC，与前一个算法相比较，网络丢包率大幅的降低。当收集周期为 700μs的时候，丢包率比顺序调度平均下降了 56%。

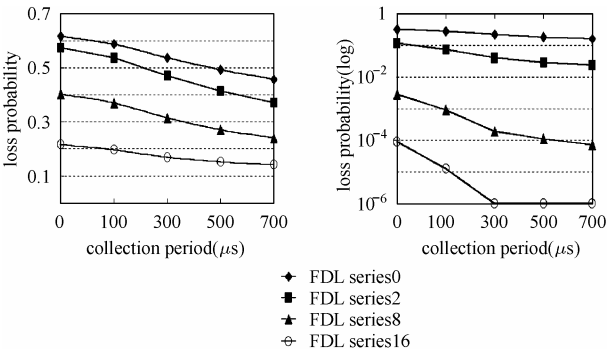

图 3 WI_UW 算法 图 4 WC_UW 算法

Fig.3 WI_UW algorithm Fig.4 WC_UW algorithm

(3) WI_W 算法 仿真参数为突发数据长度满足均值 1000kb (相应的突发数据持续时间为 1000μs) 的 Pareto 分布，偏移时间满足 (1100，2000) μs 均匀分布，以突发数据长度作为权重，到达间隔服从 Pareto 分布，业务强度为 0.8，收集周期 100~700μs。

与前面两个算法的仿真不同，为了仿真系统有各种不同权重的突发数据的情况， 突发数据长度不再取定长，而是呈 Pareto 分布，取突发数据长度作为权重。仿真结果见图 5。横坐标是收集周期，纵坐标为数据丢失率。从仿真结果可以看出，多 BHP 延迟调度算法中的 WI_W 算法，其数据丢失率性能优于顺序调度算法。当收集周期为 700μs 的时候，在不同的 FDL 级数下，数据丢失率比顺序调度平均下降了 29%。

(4) WC_W 算法 仿真参数同前。仿真结果见图 6。从仿真结果可以看出，多 BHP 延迟调度算法中的 WI_UW 算法，其数据丢失率性能优于顺序调度算法，随着收集周期的增加，数据丢失率呈下降趋势；由于配置了 TWC，与前一个算法比较，数据丢失率大幅降低。当收集周期为 700μs 的时候，数据丢失率比顺序调度平均下降了 21%。

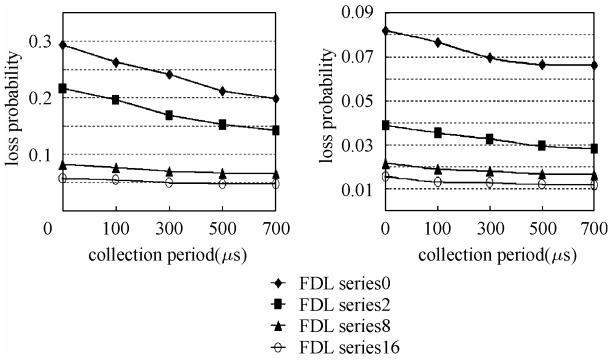

图 5 WI_W 算法 图 6 WC_W 算法

Fig.5 WI_W algorithm Fig.6 WC_W algorithm

以上 4 种具体算法均使网络丢包率大幅下降，收集周期越长，丢包率越低。引入的代价是，随着收集周期的增加，在一个核心节点需要的处理时间增大，因而较大的缩短了偏移时间，使得进入下一节点时，所剩余的偏移时间相应变短。同时，收集周期还受到偏移时间最小值的制约，所以并不是越小越好，必须根据具体网络配置和应用场景选取相应的参数。

以上 4 种具体算法均使网络丢包率大幅下降，收集周期越长，丢包率越低。引入的代价是，随着收集周期的增加，在一个核心节点需要的处理时间增大，因而较大的缩短了偏移时间，使得进入下一节点时，所剩余的偏移时间相应变短。同时，收集周期还受到偏移时间最小值的制约，所以并不是越小越好，必须根据具体网络配置和应用场景选取相应的参数。

4 结束语

本文从 OBS 网络现有的顺序调度策略的不足出发，进行考虑，提出了一种新的多 BHP 延迟调度算法。并且在此算法思想的基础上，针对各波长是否配置波长变换器，以及是否考虑突发数据的权重问题进一步提出了 4 种具体的算法实现。较顺序调度方式而言，本算法一次对多个 BHP 对应的突发数据进行集中调度，实现了对已预约资源的优化，从而使网络性能得到了改善。通过仿真，我们可以看出，由于多 BHP 延迟调度算法优化了信道资源的利用率，从整体上降低了网络丢包率，使网络性能得到提升。

参 考 文 献

[1] Hudek G C, Muder D J. Signaling analysis for a multi-switch all-optical network. Proceedings of IEEE International Conference on Communications ICC 95, Seattle, 1995: 1206 – 1210.

[2] Yoo M, Qiao C. Just-Enough-Time (JET): A high speed protocol for bursty traffic in optical networks. IEEE/LEOS Technologies for a Global Information Infrastructure, Montreal Quebec Canada, Aug. 1997: 26 – 27.

[3] Turner J. Terabit burst switching. *Journal of High Speed Networks*, 1999, 8: 3 – 16.

[4] Wei J, Pastor J, Ramamurthy R, Tsai Y. Just-in-time optical burst switching for multiwavelength networks. IFIP Broadband Commun., Hong Kong, Nov. 1999: 339 – 352.

[5] Vokkarane V M, Thodime G P V, Challagulla V U B, Jue J P. Channel scheduling algorithms using burst segmentation and FDLs for optical burst-switched networks. IEEE International Conference on Communications ICC '03, Alaska, May 2003, 2: 1443 – 1447.

[6] Xiong Yijun, Vandenhoute M, Cankaya H C. Control architecture in optical burst-switched WDM networks. *IEEE Journal on Communications*, 2000, 18(10):1838 – 1851.

[7] Yoo Myungsik, Qiao Chunming, Dixit Sudhir. QoS performance of optical burst switching in IP-over-WDM networks. *IEEE Journal on Communications*, 2000, 18(10) : 2062 – 2071.

姚　锐：　男，1978 年生，硕士生，研究方向为光通信网络. Email: yoyorui@uestc.edu.cn

王　晟：　男，1971 年生，教授，研究方向为光通信网络.Email: wsh_keylab@uestc.edu.cn

徐世中：　男，1972 年生，副教授，研究方向为网络技术和信号处理. Email: xsz@uestc.edu.cn

虞红芳：　女，1975 年生，副教授，研究方向为网络核心技术.Email: yuhf@uestc.edu.cn