一种简单的VOQ交换机时延确保分组调度算法

张福阳 熊庆旭

(南昌大学信息工程学院，南昌 330031) ((北京航空航天大学 电子信息工程学院，北京 100191)

摘 要：采用EDF(Earliest Deadline First)与轮询结合的方法，提出了一种简单的VOQ(Virtual Output Queueing)分组调度算法提供基于流的时延确保. VOQ队列采用EDF的策略裁决分组流的竞争，输入输出端口采用轮询方式匹配. 此时VOQ中分组到达至成为头分组的时间以及分组成为头分组至传输到相应输出端口的时间，分别对应于OQ中的分组排队等待时间及服务时间. 通过对所得算法详细的理论分析，给出了流时延界及流分组到达的显性关系. 更为重要的是，本文的理论结果不仅为设计更为有效的算法奠定了基础，同时为判别不同流的时延要求是否冲突提供了一种新的直接分析的手段.

关 键 词：分组调度；交换机；时延确保；虚拟输出排队

中图分类号：TP 393

文献标识码：A **文 章 编 号**：1001-5965(2008)11-1323-04

Simple scheduling algorithm for delay guarantee in VOQ architecture switches

Zhang Fuyang

(School of Information Engineering, Nanchang University, Nanchang 330031, China)

Xiong Qingxu

(School of Electronics and Information Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: Combining the earliest deadline first (EDF) policy and round robin manner, a simple scheduling algorithm was proposed to provide flow-based deterministic delay guarantee in virtual output queueing (VOQ) architecture switches. The EDF policy was employed to arbitrate the flow competitions in VOQ queues, and the round robin strategy was used to build input and output matching. In this case, the interval time between a packet arrival and that when it becomes the head cell in the VOQ, and the time from the packet becomes the head cell to it has been conveyed to the correspond output can be regard as the queuing time and service time in OQ architecture, respectively. The relation between the delay bounds and the packet inter-arrival times were derived. More important, the obtained results not only pave the way for design of more efficient algorithms, but also provide a novel approach to estimating that if there exist conflicts among the distinct delay bounds or not.

Key words: packet scheduling; switches; delay guarantee; virtual output queueing

服务质量(QoS, Quality of Service)保证是目前Internet发展中急需解决的问题之一. 而分组交换机/路由器中的分组调度提供QoS保证是分组交换网络提供QoS保证的基础和前提. 目前输出排队(OQ, Output Queued)结构交换机/路由器中提供QoS保证的分组调度取得了较为成熟的研究成果. 但由于OQ要求交换机内部的工作速度是线速的N倍，N为交换机输入端口数目，因此难以在高速网络中应用.

输入排队(IQ, Input Queued)结构的交换机内部工作速度与线速相同，具有良好的扩展性(scalability). 由于能够彻底解决输入缓存中头分

收稿日期：2007-11-15

作者简介：张福阳(1966—)，男，江西抚州人，副教授，zfuyang@126.com.

组(HOL, Head-of-Line)堵塞问题,虚拟输出排队VOQ是目前研究最多的一种IQ结构.在VOQ中一个分组要通过交换机必须参与三次竞争:输入缓存排队竞争、输入竞争、输出竞争,因此其分组调度比OQ相对复杂得多.

目前已有大量VOQ分组调度的研究结果,文献[1]对此进行了较全面的分析和对比.已有研究中绝大多数以100%的吞吐率为目标,研究较为成熟.但在提供QoS保证方面,研究还不成熟,存在的问题较多.

VOQ结构中提供QoS保证的调度算法大致可分为三类:固定帧长,矩阵分解以及业务均衡.固定帧长的算法[2-3]存在帧长选取的"粒度"(granularity)问题,为此文献[4]提出了UWRR(Uniform Weight Round Robin)算法,帧长随接入业务的改变而改变,但只能提供带宽保证不能提供时延保证.矩阵分解[5-6]方法的突出问题是计算复杂度高达$O(N^{4.5})$,限制了其在高速网络中的应用.业务均衡的方法[7]采用两级交换结构,但存在分组"失序"(out of sequence)的问题,需在输出端口增加缓存器再次采用OQ调度整序;或者在第二级交换机的输入缓存中采用固定帧长组帧,采用服务曲线的调度策略.业务均衡方法的另一个问题是存在时延界与端口数目和业务流数目的耦合问题,限制了交换机的扩展性.

文献[8]提出了业务均衡两级交换分组调度中避免分组失序的新算法,但不能提供QoS保证.文献[9]则进一步讨论了多种QoS要求并存时如何满足尽量多要求的问题,采用了基于帧的方式提供时延确保,但该方法计算复杂度高达$O(N^3)$.

输入排队交换机的应用场合是高速网络,为保证实际物理实现,分组调度算法必须简单有效.为此本文提出了一种简单的VOQ调度算法,将EDF[10]策略与轮询算法相结合,提供基于流的确定性时延保证.

1 系统模型

本文讨论端口为$N\times N$的VOQ排队结构交换机.为简化讨论,假设输入缓存容量足够大,不存在分组溢出丢失的现象.接入交换机的业务流的分组到达为可接入的,即任一输入端口一个时隙最多到达一个分组.所有分组的长度是固定的,任意输入输出端口对之间一个时隙最多只传输一个分组,即交换机的加速比(speedup)为1.

假设进入交换机业务流的总数目为M,进入$VOQ_{i,j}$缓存排队的流的数目为$M_{i,j}$.相应地

$$M=\sum_{i=1}^{N}\sum_{j=1}^{N}M_{i,j} \tag{1}$$

为讨论方便,本文所用符号说明如下:

$VOQ_{i,j}$:进入输入端口i,前往输出端口j的虚拟输出队列,$i,j\in(1,2,\cdots,N)$;

$f_{i,j}^{k}$:进入$VOQ_{i,j}$队列的第k个流,$k\in(1,2,\cdots,M_{i,j})$;

$a_{i,j}^{k}$:流$f_{i,j}^{k}$当前分组的到达时间;

$d_{i,j}^{k}$:流$f_{i,j}^{k}$的时延上界;

$x_{i,j}^{k}$:流$f_{i,j}^{k}$的最小到达时间间隔,$k\in(1,2,\cdots,M_{i,j})$;

$t_{i,j}^{k}$:流$f_{i,j}^{k}$的分组到达$VOQ_{i,j}$队列头部至传输到相应输出端口的最长时间.

2 调度算法

在VOQ中提供确定性QoS保障,必须保证任意端口对在任意时间段中的最小服务量.具体体现在两个方面,一是输出端口必须提供传输带宽的保证,即在适当的时间与相应的输入端口匹配;二是当与输出端口匹配时,相应输入端口中的分组要赢得输入竞争,使用传输机会.基于流的调度,还要裁决VOQ队列中的分组竞争.

对于输入、输出竞争,为了算法易于物理实现,本文采用轮询的方式,保证每个$VOQ_{i,j}$队列都能在每个轮询周期内获得一次服务机会.裁决VOQ队列的竞争采用EDF策略,即分组在到达时被赋予一个时限(deadline),该时限用于竞争裁决.本文假设所有分组都是在时隙开始后到达,时隙结束前从输入端口传输到输出端口.调度算法具体如下:

1) 在0时刻,对端口匹配矩阵$\boldsymbol{\Pi}$初始化

$$\pi_{i,j}=\begin{cases}1 & \text{若 } j=i\\ 0 & \text{其它}\end{cases} \tag{2}$$

其中$\pi_{i,j}=1$表示输入端口i与输出端口j匹配;否则,未建立匹配.

2) 对于时隙l开始时到达的分组(不失一般性地,假设属于流$f_{i,j}^{k}$),赋予其一个时限值$D_{i,j}^{k}$

$$D_{i,j}^{k}=a_{i,j}^{k}+d_{i,j}^{k} \tag{3}$$

根据$D_{i,j}^{k}$的大小顺序排入$VOQ_{i,j}$队列,$D_{i,j}^{k}$越小的分组排列越靠近队列头部.

3) 按如下规则更新端口匹配矩阵$\boldsymbol{\Pi}$:

$$\pi_{i,j}=\begin{cases}1 & \text{若 } j=i+l(\mathrm{mol})N\\ 0 & \text{其它}\end{cases} \tag{4}$$

4) 依次将 $\pi_{i,j}=1$ 所对应的输入端口 i 中 $VOQ_{i,j}$ 的头分组传输到输出端口 j.

3　理论分析

本文分析的核心思想是将 OQ 中的分组排队等待时间及服务时间,分别对应于 VOQ 中分组到达至成为头分组的时间以及分组成为头分组至传输到相应输出端口的时间,因此 OQ 中不同流分组最大服务时间对应于本文中的 $t_{i,j}^k$.

将同一 $VOQ_{i,j}$ 队列中的流分成 U 和 V 两个集合.U 集合中流的时延上界均小于该 $VOQ_{i,j}$ 队列中不同流的分组最大服务时间之和;V 集合中流的时延上界均不小于同一 $VOQ_{i,j}$ 队列中不同流的分组最大服务时间之和.

$$U=\left\{k\,\middle|\,k=1,2,\cdots,u;d_{i,j}^k<\sum_{m=1}^{M_{i,j}}t_{i,j}^m\right\} \tag{5}$$

$$V=\left\{k\,\middle|\,k=u+1,u+2,\cdots,M_{i,j};\ d_{i,j}^k\geqslant\sum_{m=1}^{M_{i,j}}t_{i,j}^m\right\} \tag{6}$$

不失一般性,把 U 集合中的流按时延上界的大小依次编号(1, 2,…,u),当 U 集合中的流竞争传输机会时则 $f_{i,j}^1$ 最先得到调度,$f_{i,j}^u$ 最后得到调度.

引理　假设流 $f_{i,j}^k$ 的分组到达时间间隔满足

$$x_{i,j}^k\geqslant\sum_{m=1}^{M_{i,j}}t_{i,j}^m\quad k\in(1,2,\cdots,M_{i,j}) \tag{7}$$

则如果流 $f_{i,j}^k$ 的时延上界满足式(8)的条件,则本文调度算法可确保每个流的时延上界.

$$d_{i,j}^k\geqslant\sum_{m=1}^{k}t_{i,j}^m+T_{i,j}\quad k\in(1,2,\cdots,u) \tag{8}$$

其中 $T_{i,j}$ 为所有 $t_{i,j}^k(k\in V)$ 中的最大值.

证明　假设在 0 时刻,$VOQ_{i,j}$ 队列为空,这时到达了第一个分组.$VOQ_{i,j}$ 队列中每个不同流的分组到达时刻是任意的,且相互独立.但同一个流 $f_{i,j}^k$ 的第一个分组到达之后,第二个分组到达时刻与第一个分组的到达时刻是相关的,即到达的时间间隔不小于 $x_{i,j}^k$.

必要性:即如果式(8)不满足,则至少有一个分组不能保证时延上界.假定在 0 时刻,具有服务时间为 $T_{i,j}$ 的分组到达了 $VOQ_{i,j}$ 队列,并且 U 集合中所有的流同时在 0^+ 时刻到达,该分组最早只能在 $\sum_{m=1}^{k}t_{i,j}^m+T_{i,j}$ 得到服务,这就超出了该流的时延上界.

充分性:即如果式(8)满足的话,则所有流的时延不会超过其上界.对任一流 $f_{i,j}^k$ 的分组到达时刻分以下几种情况:①若在 0 时刻到达,其最大离去时刻为 $\sum_{m=1}^{k}t_{i,j}^m<d_{i,j}^k$;② 若在 0^+ 时刻到达,其最大离去时刻为 $\sum_{m=1}^{k}t_{i,j}^m+T_{i,j}\leqslant d_{i,j}^k$;③ 若在 $(0^+,\ T_{i,j}]$ 时刻到达,其最大离去时刻小于 $d_{i,j}^k$;④ 若在 $T_{i,j}^+$ 时刻到达,其最大离去时刻为 $\sum_{m=1}^{k}t_{i,j}^m<d_{i,j}^k$;⑤ 在这之后到达的分组,则最大离去时刻将会更小.　　**证毕**

当输出竞争采用轮询匹配策略,可得定理 1.

定理 1　轮询匹配算法保证每个 $VOQ_{i,j}$ 队列的最大等待时间为 N 个时隙.

证明　考虑一个轮询周期 N 时隙之内一个 $VOQ_{i,j}$ 队列的情况,其它 VOQ 队列的情况类似.一个分组到达 $VOQ_{i,j}$ 队列头部的时刻有以下几种情况:①若在 0 时刻到达,正好赶上传输机会,等待时间为 0;②若在 0^+ 时刻到达,至少也至多等待 N 个时隙后获得传输机会;③若在$(0^+,\ N]$ 区间到达,最大等待时间小于 N 个时隙;④若在 N 个时隙之后到达,就是前一个轮询周期的循环,同样不会超过 N 个时隙.　　**证毕**

结合定理 1 和引理可以得到关于本文所得算法的定理 2.

定理 2　假定流的分组的到达时间间隔满足

$$x_{i,j}^k\geqslant M_{i,j}N\quad k\in(1,2,\cdots,M_{i,j}) \tag{9}$$

且流的时延上界要求满足

$$d_{i,j}^k\geqslant(k+1)N\quad k\in(1,2,\cdots,u) \tag{10}$$

则本文所得算法可以保证任何一个流的时延上界.

证明　由定理 1 和引理可得 $t_{i,j}^k=T_{i,j}=N$,将该值带入式(7)和式(8),即可得到定理 2.

证毕

4　讨　论

本文算法复杂度为 $O(N^{2.5})$ 远低于矩阵分解+PGPS 算法[5] 的 $O(N^{4.5})$.复杂度与业务均衡的算法[7] 相同,但只需一级而不是两级交换结构,且无需输出缓存器处理分组失序的问题.同时这两种典型算法均是基于端口的(port-based),而本文算法是基于流的(flow-based),这是目前研究相对较少的问题.

本文算法的时延上界如式(8)所示.文献[5]中算法在 $VOQ_{i,j}$ 的分组到达符合$(\sigma_{i,j},\rho_{i,j})$约束时,时延上界为$(s_{i,j}+\sigma_{i,j})/\rho_{i,j}$,其中 $s_{i,j}$ 是由在线调度策略决定的常数.文献[7]中算法时延上界为

$(N-1)L_{\max}+NM_{\max}+D_{OQ}$，其中 $L_{\max}$，$M_{\max}$ 分别为输入、输出端口业务流的最大数目，D_{OQ} 为分组通过 OQ 交换机的时延. 显然本文算法更利于显性地调控分组时延上界.

本文算法更重要的意义是为 VOQ 基于流时延确保调度提供了一种一般性的理论分析的手段. 本文虽然讨论的是简单轮询的方式，对于不同去向分布的业务可以通过 $x_{i,j}^k$ 的不同来体现，并调整一个轮询周期中相应端口的匹配次数.

首先是作为一种算法分析的手段，通过对不同调度算法的 $t_{i,j}^k$ 的分析以及分组到达时间间隔的确定，利用引理即可初步判断能否提供时延确保. 其次当存在三种以上的不同时延要求时，是否存在相应 VOQ 调度算法是一个 NP-Complete 问题，利用本文理论分析的结果可以直接计算出相应的最小的 $t_{i,j}^k$，进而通过不同 $t_{i,j}^k$ 的关系加以判断. 最后本文所得理论结果可以直接作为设计新算法的理论依据，在计算得到 $T_{i,j}^k$ 的基础上，设计相应的端口匹配算法以满足相应的时延上界的要求. 同时时延与最小分组到达间隔的关系是进行业务整形的基本依据，可用于设计相应的业务整形方案.

5 结束语

本文讨论了一种简单的 VOQ 结构交换机提供基于流的时延确保的分组调度算法. 采用了 EDF 与轮询相结合的方式，将 VOQ 调度中输出竞争过程中的最大等待时间映射为 EDF 调度所需要的最大服务时间. 所得算法不存在目前较为普遍的高复杂度或分组失序等问题，易于实际的物理实现.

在本文引理的证明过程中，要求分组的到达是非强占性的，因此本文理论分析得到的可调度域相对是保守的. 本文研究结果更为重要的意义是为 VOQ 基于流时延确保调度提供了一种一般性的理论分析的手段，这也是未来进一步具体研究的工作.

参考文献（References）

[1] 熊庆旭. 输入排队结构交换机分组调度研究 [J]. 通信学报，2005，26(6)：118－129
Xiong Qingxu. Research on packet scheduling in input-queued switches [J]. Journal of China Institute of Communications, 2005, 26(6):118－129(in Chinese)

[2] Li S, Ansari N. Input-queued switching with QoS guarantees [C]// IEEE INFOCOM '99. New York: IEEE Inc, 1999:1152－1159

[3] Hung A, Kesidis G, Mckeown N, et al. ATM input-buffered switches with guaranteed-rate property [C]// IEEE ISCC '98. Athens, Greece:IEEE Inc, 1998:331－335

[4] Rai I A, Alanyali M. Uniform weighted round robin scheduling algorithms for input queued switches [C]// IEEE ICC'01. Helsinki, Finland:IEEE Inc, 2001:2028－2032

[5] Chang C S, Chen W J, Huang H Y, et al. On service guarantees for input buffered crossbar switches: a capacity decomposition approach by Birkhoff and von Neumann [C]// IEEE IWQoS'99. London, England:IEEE Inc, 1999:79－86

[6] Li J, Ansari N. QoS guaranteed input queued scheduling algorithms with low delay[C]// IEEE HPSR'01. Dallas, TX, USA:IEEE Inc, 2001:412－414

[7] Chang C S, Lee D S, Yue C Y, et al. Providing guaranteed rate service in the load balanced Birkhoff-von Neumann switches[C]// IEEE INFOCOM'03. San Francisco, CA, USA: IEEE Inc, 2003:1662－1632

[8] Lee Hyoung-Il. A two-stage switch with load balancing scheme maintaining packet sequence [J]. IEEE Communications Letters, 2006, 10(4):290－292

[9] Lee Yong, Lou Jianyu, Luo Junzhou, et al. An efficient packet scheduling algorithm with deadline guarantees for input-queued switches [J]. IEEE/ACM Trans on Networking, 2007, 15(1):212－225

[10] Ferrari D, Verma D C. A scheme for real-time channel establishment in wide-area networks [J]. IEEE J SAC, 1990, 8(3): 368－379

[11] Bonuccelli M A, Clò M C. Scheduling of real-time messages in optical broadcast-and-select networks [J]. IEEE/ ACM Trans Networking, 2001, 9(5): 541－552

（责任编辑：娄 嘉）