文章编号:1001 - 2486(2005)05 - 0035 - 04

高性能路由器中的分组非精确调度技术*

陈一骄,孙志刚
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:随着网络带宽的不断提高,分组到达路由器的时间间隔不断缩短,对路由器处理分组的速度提出了新的要求。传统的分组调度算法,如 WFQ,由于性能和可扩展性等问题,难以在高性能核心路由器中实现。为此,提出了分组非精确调度技术,在不影响应用 QoS 的前提下对经典的调度算法进行修改,通过简化硬件设计提高调度器的服务能力。模拟分析表明,采用非精确调度技术的 SLQF 算法的延时特性与传统算法 LQF 基本一致。

关键词:高性能路由器;分组;非精确调度
中图分类号:TN393 文献标识码:A

Un-precise Packet Scheduling in High Performance Routers

CHEN Yi-jiao, SUN Zhi-gang
(College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Along with the continuing increase of network bandwidth, the time gaps between the arrived packets are decreasing. The changed situations demand the core router to improve the ability of packet processing. The traditional packet scheduling algorithms, such as WFQ, are difficult to implement in high performance core router because of the problems of performance and extensibility. The paper proposes a new un-precise packet scheduling scheme, which improves the performance of the scheduler by modifying the classical packet scheduling algorithms to simplify the design of hardware. The scheme doesn't affect the QoS quality of the algorithms. The delay performance of this scheme is analyzed and computer simulation results show that the delay performance of SLQF is consistent with that of LFQ on the whole.

Key words: high performance core router; packet; un-precise scheduling

由于 WEB 应用的出现,互联网从 20 世纪 90 年代初开始经历了 10 多年的飞速发展。在这期间,关于互联网的新技术、新应用不断出现。路由器作为组成互联网的核心设备,也在不断更新换代,从最初的多网卡微型计算机发展成为规模可与 MPP 计算机系统相媲美的复杂系统。与路由器相关的高性能网络交换、网络协议和 QoS(Quality of Service)技术的研究也取得了丰硕的成果。

例如,针对路由器的服务质量控制,Stiliadis 和 Varma 提出了 WFQ(Weighted Fair Queue)类调度算法[1,2],这类算法可确保应用分组获得的延时和带宽,对对象的公平性也有一定的改善。针对路由查找的最长前缀匹配问题,Nilsson 和 Karlsson 通过多层次压缩增强 Radix Tries 算法[3]的路径压缩率,减少一次路由查表的访存次数,提高路由查找的性能。针对 CrossBar 交换网络的调度问题,Nick McKeow 提出了 iSLIP 算法[4],该算法仅用 $\log_2 N$ 次迭代就可高效地实现具有 N 个端口的输入缓冲 CrossBar 的调度问题。针对缓冲区管理问题,Floyd 结合 TCP 协议拥塞控制机制,提出 RED[5](Random Early Detection)缓冲区管理算法,避免了 TCP 端系统出现窗口变化同步的现象,提高了网络的性能。上述算法对互联网的发展做出了巨大贡献,在 20 世纪 90 年代中后期,对这些算法的研究和改进一直是互联网学术界讨论的热点。

* 收稿日期:2005 - 05 - 20
基金项目:国家自然科学基金重点项目(90104001);国家重点基础研究发展计划项目(2003CB314802);国家 863 高技术研究发展计划基金项目(2003AA115130)
作者简介:陈一骄(1972—),男,助理研究员,博士生。

互联网的不断发展,要求路由器的性能和可扩展性不断提高。尤其在网络核心,路由器端口速率已经由10年前的2Mbps发展到现在的10Gbps,端口数目也由原来的几个上升为几十个甚至上百个,许多经典算法已经不适合在高性能路由器的设计中使用。

1 传统分组调度算法面临的挑战

分组调度是路由器中最常见的处理事件,其主要任务是从多个等待队列中按照一定策略选取一个分组从输出链路发出。输出队列的WFQ算法、输入缓冲CrossBar交换网络的iSLIP算法都是路由器中广泛使用的分组调度算法。然而随着网络规模的扩大和网络带宽的增加,这些调度算法很难在核心路由器中使用。

1.1 算法执行窗口减小的挑战

为了提高调度性能,当一个调度分组输出时,调度器进行下一个调度决策,确定下一个被调度的分组。调度器可用来进行调度决策的时间称为调度算法的执行窗口。随着网络带宽的不断提高,分组到达路由器的时间间隔不断缩短,调度算法的执行窗口也逐渐减小。以40字节的分组为例,当输出链路带宽为2Mbps和10Gbps时,算法执行窗口分别为:2Mbps链路:$(40\times 8b)/2Mb/s = 160\mu s$;10Gbps链路:$(40\times 8b)/10Gb/s = 32ns$。

逐渐变小的算法执行窗口使得分组调度算法的设计与实现越来越困难。

1.2 执行的复杂性提高对传统分组调度算法的挑战

随着网络规模的不断扩大,当前互联网核心并发存在几十万条流[6],即使使用排序链表组织流状态信息(虚时钟),在32ns左右时间内完成每个流状态的比较和修改也几乎是不可能的。因此在网络核心一般采用区分服务QoS模型,不针对每一条流进行调度,以削弱对QoS的支持换取路由器的高速处理。

由此可见,由于无法满足网络核心对处理性能和可扩展能力的要求,许多类似WFQ这样的算法无法在高性能路由器中使用。

2 非精确调度

分组调度算法的调度模型如图1所示。其中$Q(t)$表示t时刻等待处理的分组队列状态,f和f'为调度算法,$S(t)$表示t时刻处理算法保存的控制状态,A和B表示待调度对象。

设算法调度分组的延时为P_d,则P_d是$Q(t)$、$S(t)$、f和A的函数,即

$$P_d = f(Q(t), S(t), A) \tag{1}$$

通过对算法f进行部分修改,可以加速调度的快速实现,从而适应不断缩小的算法执行窗口。由于修改后算法f'确定的调度结果可能与算法f的调度结果不同,因此称算法f'为非精确调度算法,相应地,称算法f为精确调度算法。

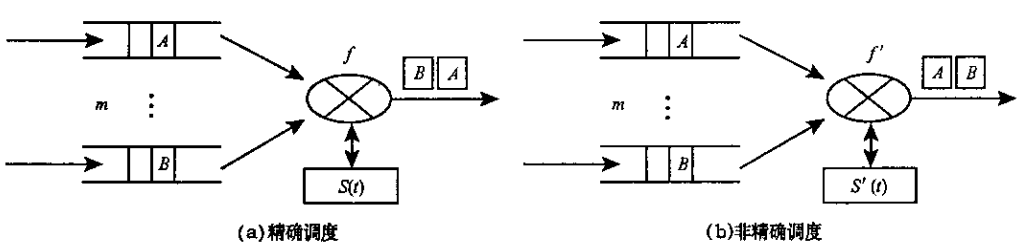

图1 非精确调度可能带来分组延时的变化

Fig.1 The packet delay changed because of un-precise scheduling

非精确调度可能会改变单个分组的处理延时,以图1所示的分组调度为例,精确调度f使队列m中分组A和队列n中分组B先后从输出链路发出;但对于非精确调度f',可能分组A和分组B输出的顺序正好相反,即B先于A从输出链路发出,导致分组A经历的延时增大,分组B经历的延时减小。因此,非精确调度在简化算法执行复杂度的同时,不能降低调度对QoS的保证能力。

对分组延时的控制是路由器 QoS 控制的重要手段,因此对算法非精确度的衡量以非精确调度带来的分组延时变化为基准。不同类型的调度算法可使用不同的衡量标准。设分组 p 进入队列的时刻为 t_0,到达队列头部的时刻为 t_1,被调度离开队列的时刻为 t_2,那么分组在队列中等待时间 $DQ = t_2 - t_0$,分组在队列头部等待时间为 $DH = t_2 - t_1$,对于精确算法和非精确算法,分组在队列中的等待时间记为 pDQ 和 $upDQ$,分组在队列头部等待时间记为 pDH 和 $upDH$。

2.1 非精确 WFQ 调度

WFQ 调度对单个分组的延时比较敏感,因此可使用流队列头部分组等待延时的变化衡量非精确性,即对于任意分组 p 有:

$$|pDH(p) - upDH(p)| < \delta_1 \tag{2}$$

其中,δ_1 是 QoS 控制可以接受的分组延时偏差,考虑网络应用对 QoS 的要求以及端系统对延时处理的精确度,δ_1 可设为 ms 级。设网络中最大分组为 1500 字节,对于 10Gbps 链路,图 1 所示的非精确调度造成的非精确度为:

$$(1500 \times 8)/10G = 1.2(\mu s) \ll \delta_1 \tag{3}$$

因此,图 1(b)中的调度虽然造成分组输出顺序改变,但变化的延时对应用并不敏感,是可接受的。而当链路速率为 2Mbps 时,非精确度达到 6ms,显然是不可接受的。

2.2 非精确交换网络调度

交换网络调度算法衡量的标准是带宽利用率或 VOQ(Virtual Output Queue)中信元的平均延时,因此非精确的交换网络调度的非精确度可使用以下方法计算:

对于任意 t 时刻开始的任意时间段 T,对于任意 $S(t)$、$Q(t)$,对于精确调度算法有 pM 个分组被调度到输出链路输出,对于非精确调度算法有 upM 个分组被调度到输出链路输出,那么:

$$\left| (\sum pDQ)/pM - (\sum upDQ)/upM \right| < \delta_2 \tag{4}$$

如果 δ_2 足够小,(4)式表示在较长时间内,调度的性能不会因非精确调度而显著下降,队列内分组的平均延时与精确调度最多相差 δ_2,且这个值与时间长度 T 无关。

2.3 非精确调度的一个实例

Manolis Katevenis, Georgios Passas 等提出了基于缓冲区的可变长度报文交换体系结构[7,8],将传统的以定长信元为基础的交换体系结构,扩展为以可变长度的报文为单位的基于报文的交换结构,是一种非精确调度方法。由于改定长交换为变长交换,加大了一次交换的长度,从而延长了调度算法的执行窗口,简化了调度算法的实现。在实际应用中,该交换体系结构不需要分片与重组电路以及由此带来的加速度需求,简化了输入输出端口间的时钟域同步,消除了输出队列,获得了很好的性能。

3 一种非精确交换网络调度算法

输入缓冲 CrossBar 交换开关是目前高性能核心路由器中广泛采用的交换结构[9]。为提高效率,交换开关在第 n 个时间槽对信元交换的同时,调度器根据各 VOQ 的状态确定开关在第 $n+1$ 个时间槽内的拓扑,并将配置信息写入开关。第 n 个时间槽结束后,开关根据预先的配置迅速改变内部拓扑,并启动第 $n+1$ 个时间槽的交换。

由于链路带宽不断增加,交换开关对调度器的速度要求不断提高。例如,对于端口速率为 10Gbps、端口数目为 8、每个时间槽交换 64 字节的交换开关,调度器要在 50ns 左右根据 64 个 VOQ 队列的状态完成调度决策,并将调度结果写入交换开关。由于时间短,目前,调度器中实现的都是简单的非加权调度算法[4],而许多高性能的调度算法,如 LQF(Longest Queue First)[10],因实现困难,一般在高性能路由器中都得不到应用。

大步调度[9]是一种典型的非精确调度,这种调度方法将每个时间槽调度一次改为 K 个时间槽调度一次。由于调度算法每 K 个时间槽执行一次,因此可以有足够的时间(K 个时间槽)执行较复杂的加权调度计算,从而使算法设计实现的压力大大减轻。图 2 所示为 LQF 与在 $K=2$,4 和 8 的情况下 SLQF 算

法延时特性的比较。通过比较可以看出,非精确 SLQF 算法与精确的 LQF 算法的延时特性基本一致,在高负载情况下性能更佳。同时,在理论上可以证明 SLQF 与 LQF 一样,都可以 100%利用交换网络带宽,关于 SLQF 算法的详细分析见文献[9]。

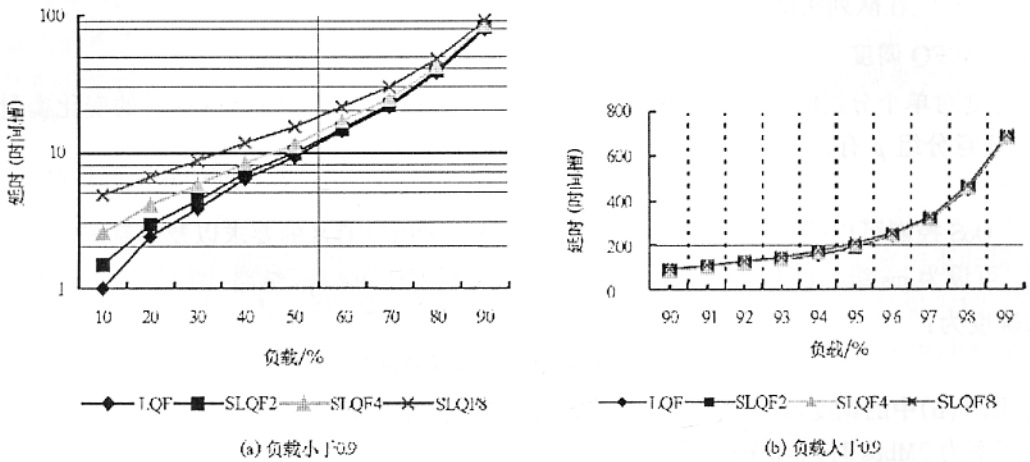

图 2 LQF 与 SLQF 的延时特性比较

Fig.2 The comparison of delay characteristic between LQF and SLQF

4 结束语

非精确调度技术是实现高性能核心路由器中调度算法的有效途径。其主要思想是在不影响应用服务质量的前提下,对网络核心中的调度算法进行改造,通过改造简化调度算法处理的复杂性,降低实现难度。

非精确调度技术还可以进一步发展成为非精确处理技术,即除队列调度外,核心路由器中的其他处理,如转发查表,也可以使用非精确的思想。即在使用多级交换网络的核心路由器中,传统的处理过程为转发查表然后进行多级交换,当多级交换网络有多条从输入端口到达目的输出端口的交换路径时,转发查表只需得到内部交换网络的下一跳,并不一定必须查出精确的输出端口号,而输出端口号可在交换的过程中逐次查表得到。

非精确处理技术及其在高性能核心路由器中的应用将是我们下一步研究的重点。

参 考 文 献:

[1] Stiliadis D, Varma A. Efficient Fair Queueing Algorithms for Packet-switched Networks[J]. IEEE/ACM Trans. on Networking, 1998,6(2):175 – 185.

[2] Stiliadis D, Varma A. Rate-Proportional Servers: A Design Methodology for Fair Queueing Algorithms[J]. IEEE/ACM Trans. on Networking, 1998,6(2):164 – 174.

[3] Nilsson S, Karlsson G. Fast Address Lookup for Internet Routers[A]. In P. K¨uhn and R. Ulrich, editors, Broadband Communications: The Future of Telecommunications[C], 1998:11 – 22.

[4] McKeown N. iSLIP: A Scheduling Algorithm for Input-queued Switches[J]. IEEE Transactions on Networking, 1999, 7(3):188 – 201.

[5] Floyd S. Random Early Detection Gataways or Congetion Avoidance[J]. IEEE/ACM Transaction, on Networking, August 1993,1(4):397 – 413.

[6] 程光,等.基于统计分析的高速网络分布式抽样测量模型[J].计算机学报,2003,26(10).

[7] Katevenis M, Passas G, et al. Variable Packet Size Buffered Crossbar (CICQ) Switches[R]. Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH) ICS – FORTH, Vassilika Vouton, Heraklion, Crete, GR – 711 – 10 Greece, http://archvlsi.ics.forth.gr/bufxbar/.

[8] Passas G. Performance Evaluation of Variable Packet Size Buffered, Crossbar Switches, Technical Report FORTH – ICS/TR – 328[R]. Inst. of Computer Science, FORTH, Heraklion, Crete, Greece; B.Sc. Thesis, Univ. of Crete; November 2003; http://archvlsi.ics.forth.gr/bufxbar.

[9] 孙志刚,卢锡城.输入缓冲交换开关的多步调度策略[J].软件学报,2002,(8).

[10] McKeown N, et al. A Starvation-free Algorithm for Achieving 100% Throughput in an Input-queued Switch[A]. Proceedings of ICCCN '96[C], 1996.