对数比例公平队列：一种新的调度算法 [1]

郑友泉 冯振明

(清华大学电子工程系 北京 100084)

摘 要 Internet 路由器中的队列调度算法关系到用户数据流能够获得的服务质量以及整个网络的性能. 该文提出了一种新的队列调度算法—对数比例公平队列调度算法, 并通过理论分析和仿真说明了该队列调度算法的一些特性. 这种算法不仅实现了网络总效用、网络传输性能与资源分配公平性的折衷, 还具有对恶意用户的惩罚作用, 鼓励用户使用合理的端到端流量控制算法.

关键词 公平队列, 调度算法, 服务质量, 对数比例公平队列

中图号 TN919.3

1 引 言

随着科学技术和社会需求的发展, 通信网络正在向综合业务网络演进. 由于不同的业务有着不同的服务质量 (QoS) 需求, 网络必须进行合理的资源分配, 以满足或尽量满足这些业务的服务质量需求. 资源共享式的通信网络 (如 Internet) 应该对那些行为友好 (有合理端到端流量控制机制) 的用户进行奖励, 对恶意 (没有合理端到端流量控制机制, 或在短时间内向网络中大量注入数据) 的用户进行惩罚 [1].

现行的 Internet 体系中, 广泛使用 TCP(Transmission Control Protocol) 算法对网络中的数据流进行端到端的流量控制, 目的是避免或缓解网络的拥塞状况, 提高网络的可用性. 这种流量控制算法只在用户终端处进行, 不在网络内部节点处增加任何复杂的功能, 因而使网络具有很强的可扩展性. 然而, 正是由于端到端流量控制机制过于简单, 导致目前 Internet 中经常出现网络拥塞、资源分配不公平等现象, 并且缺乏业务分级和服务质量保证的能力 [2,3]. 目前主要的解决办法是在网络内部节点中使用公平队列对用户的数据流进行控制, 用一定的复杂度换取网络性能的提高. 已经提出的公平队列调度算法有 WFQ(Weighted Fair queueing) 、 CSFQ(Core Stateless Fair Queueing) 、 Virtual Clock 、 Delay-Earliest-Due-Data 等等. 这些算法基本上实现了网络资源在数据流之间的公平分配, 并且对服务质量能够作出一定的保证. 但是, 它们都过于强调公平性问题而忽略了网络总效用、网络传输性能以及资源分配公平性之间的折衷, 也没有对恶意用户进行惩罚的机制.

在 Internet 中, 网络节点将主要通过队列管理和调度算法进行缓冲区和输出带宽的分配. 本文提出了一种新的公平队列调度算法: 对数比例公平队列 (LPFQ) 调度算法. 这种算法为每个数据流分配的带宽和延时性能均有上界和下界; 通过这种算法, 我们可以实现网络利用率、延时性能、资源分配公平性之间的折衷; 另外, 这种算法对恶意数据流具有抑制和惩罚作用, 鼓励用户采用一定的端到端流量控制机制以获得好的服务质量.

本文第 2 节提出了 LPFQ 的定义, 并对其性能进行了理论分析; 第 3 节通过数值仿真来说明此 LPFQ 队列调度算法的性能; 第 4 节是对全文的总结.

2 对数比例公平队列定义及其性能分析

现在的 Internet 中, 绝大多数网络内部节点 (路由器) 中使用的仍然是先来先服务的单个 FIFO(First-In First-Out) 队列, 这种排队机制使得进入该节点的数据流之间没有隔离, 单个数

[1] 2000-06-27 收到, 2000-10-30 定稿

据流所得到的性能受其他数据流的负面影响很大，资源分配的公平性得不到保证．为了隔离不同数据流，提高公平性，有些路由器中已经实现了 WFQ 机制．这种机制为进入路由器的每个流提供一个单独的 FIFO 队列，并为每个队列公平地分配输出带宽．前面已经提到，WFQ 虽然公平性很好，但不考虑网络的传输性能以及网络利用率，也不能对恶意用户进行惩罚．下面，我们提出 LPFQ 调度算法，并说明该算法能够得到一定的折衷效果，而且能够对恶意用户进行相应的惩罚．

2.1 LPFQ 定义

设某路由器的输出链路容量为 $C>0$，有 N 个数据流经过该路由器，设为 $l_i, 1\le i\le N$；该路由器为每个数据流分配一个单独的 FIFO 队列，这些 FIFO 队列的容量相同，设为 B，当队列中有数据在等待发送时，路由器就不停止发送工作；数据流 l_i 的队列中等待发送的队列长度设为 L_i，并通过丢包保证 $L_i<B$，设其队列占用率为 B_i，则 $B_i=L_i/B, 0<B_i<1$；队列长度和数据量均以包为单位．比例公平队列调度算法定义为：路由器为任意两个流分配的输出带宽 $\lambda_i, \lambda_j, 1\le i\le N, 1\le j\le N$ 满足下式

$$\frac{\lambda_i}{\lambda_j}=\frac{\log B_i}{\log B_j},\quad \sum_{i=1}^{N}\lambda_r=C \tag{1}$$

由 (1) 式可以看出：对于发送大量数据，导致队列占用率较大的恶意用户，不仅仅是丢包率大，其获得的发送速率以及延时性能也均低于队列占用率小的用户，这就使突发性强的流、没有端到端流量控制机制的流 (恶意的流) 受到了惩罚；队列占用率较小的流中的包获得的发送速率较大，等待延时较小；另外，我们还可以设置一个公平的占用门限，当队列占用率超过这个门限的时候再对数据流实行比例公平队列调度算法．这样，为了获得高速率和低延时，用户数据流最好的策略就是在路由器队列中仅保持少量的数据等待发送．

2.2 LPFQ 的发送速率

通过下面的证明可以看到，在采用 LPFQ 的路由器中，我们可以求出发送速率的上下界，或是按发送速率的上下界求出队列占用率的值，可以把这些值通过适当的反馈机制传送给用户，作为用户调整其端到端流量控制算法的参考．

定理 1　采用 LPFQ 的路由器中，任何一个流 l_i 的发送速率 λ_i 都有下界 $\lambda_{i\min}$ 和上界 $\lambda_{i\max}$：

$$\lambda_{i\min}=\frac{\log[(B-1)/B]}{\log[(B-1)/B]+(N-1)\log(1/B)}\cdot C \tag{2}$$

此下界在该流的队列占用率最大，其他流的队列占用率最小时发生；

$$\lambda_{i\max}=\frac{\log(1/B)}{\log(1/B)+(N-1)\log[(B-1)/B]}\cdot C \tag{3}$$

此上界在该流的队列占用率最小，其他流队列占用率最大时发生．

证明　由于 $0<B_i<1$，所以有 $1/B\le B_i\le(B-1)/B$, $\log(1/B)\le\log B_i\le\log[(B-1)/B]<0$，同样，$\log(1/B)\le\log B_j\le\log[(B-1)/B]<0$．所以，$i\ne j$ 时，

$$\frac{\log[(B-1)/B]}{\log(1/B)}\le\frac{\log B_i}{\log B_j}\le\frac{\log(1/B)}{\log[(B-1)/B]} \tag{4}$$

由 (1) 式可以求得流 l_i 的发送速率为

$$\lambda_i=\frac{\log B_i}{\log B_i+\sum_{i=1,i\ne j}^{N}\log B_j}\cdot C=\frac{1}{1+\sum_{i=1,i\ne j}^{N}\frac{\log B_j}{\log B_i}}\cdot C \tag{5}$$

根据 (4)，(5) 式可得：$\lambda_{i\min}=\dfrac{\log[(B-1)/B]}{\log[(B-1)/B]+(N-1)\log(1/B)}\cdot C$，当且仅当该流的队列占用率最大 $B_i=(B-1)/B$，而其他流的队列占用率最小 $B_j=1/B, i\neq j$ 时，等号成立.

$\lambda_{i\max}=\dfrac{\log(1/B)}{\log(1/B)+(N-1)\log[(B-1)/B]}\cdot C$，当且仅当该流的队列占用率最大 $B_i=1/B$，而其他流的队列占用率最小 $B_j=(B-1)/B, i\neq j$ 时，等号成立. 当路由器中只有一个流时，该流可以得到的最大发送速率为 C. 证毕

2.3 LPFQ 的延时性能分析

包在路由器上的延时就是包的排队等待延时加上包的发送时间. 等待延时显然是队列长度与发送速率的商，发送时间是包长度与发送速率的商，设平均包长度固定为 1，包到达路由器相应队列后，队列中等待发送的数据长度为 L_i，可得

$$\mathrm{Delay}_i=(L_i-1)/\lambda_i+1/\lambda_i=L_i/\lambda_i \tag{6}$$

由 (6) 式及定理 1 可以简单地证明下面的定理 (证明略).

定理 2 在采用 LPFQ 的路由器中，任何一个流 l_i 的数据包在该路由器处的延时 Delay_i 有下界 $\mathrm{Delay}_{i\min}$ 和上界 $\mathrm{Delay}_{i\max}$：

$$\mathrm{Delay}_{i\min}=1/\lambda_{i\max} \tag{7}$$

$$\mathrm{Delay}_{i\max}=(B-1)/\lambda_{i\min} \tag{8}$$

下界在该流队列占用率最小，其他流队列占用率最大时发生；上界在该流队列占用率最大，其他流队列占用率最小时发生.

与发送速率一样，延时的最差情况也是在其他用户均只有一个包等待发送时发生的. 我们可以通过队列占用率求得延时的最大值，也可以通过延时界限求得应该在路由器处保持多大的队列占用率.

可见，采用 LPFQ 可以保证任何一个流都得到一定的输出带宽和一定的延时性能. 每个流输出速率上下界以及延时上下界与其他的流无关；但实际输出速率、延时性能与其他流的表现是有关系的，任何一个流都将因为其它流的恶意行为而获得好处. 另外，考虑各个流的随机性时，平均发送速率和平均延时会在上下界之间随机变动.

2.4 公平性与网络性能

上面只讨论了单个路由器的情况，下面证明，在由多个使用 LPFQ 的路由器组成的网络中，可以实现网络总效用、网络总延时与公平性之间的折衷.

设一个网络由一个路由器集合 M 组成，其中每个路由器 m 对应的输出链路带宽为 $C_m>0$. 多个数据流竞争这些路由器的输出带宽和缓冲区资源，每个数据流的路径由 M 的一个子集 r 组成. 当路径 r 经过路由器 m 时，记为 $m\in r$. R 表示数据流路径的集合. 设网络中数据流的个数为 N. λ_r 表示分配给路径 r 的带宽. 显然，带宽分配必须满足带宽限制条件：

$$\sum_{r\ni m}\lambda_r\leq C_m,\quad m\in M \tag{9}$$

由于采用 LPFQ 算法，路由器 m 为每个数据流提供一个队列，每个队列容量均为 B_m；与数据流 r 对应的队列中排队等待发送的数据量为 $L_{r,m}$，设数据流 r 在路由器 m 处的队列占用率为 $B_{r,m}$，通过丢包保证 $L_{r,m}<B_m$，则 $B_{r,m}=L_{r,m}/B_m$，且 $0<B_{r,m}<1$；再假设数据流 r 的发送窗口为 W_r；设数据流 r 的传输速率为 λ_r，往返时间为 T_r. 下面的分析中使用流体流模型 [4].

定理 3　在上面假设的条件下，网络中数据流的稳态速率分布为最优化问题：

$$\max \sum_{R}(\log\lambda_r - a/\lambda_r), \quad \text{其中 } \lambda_r \geq 0, \quad \sum_{r\ni m}\lambda_r \leq C_m$$

且

$$a = \left(W_r - \sum_{m\in r} B_m\right) \Big/ \left[T_r - \sum_{m\in r}\left(B_m \log\left(\prod_{r\ni m} B_{r,m}\right)\right)\right], \quad 0 < B_{r,m} < 1 \tag{11}$$

的解.

证明　由于采用 LPFQ 算法，路由器 m 为经过它的任意两个队列占用率满足 $0 < B_{r,m} < 1$ 的数据流 $r1$, $r2$ 按如下方式分配输出速率：

$$\lambda_{r1}/\lambda_{r2} = \log B_{r1,m}/\log B_{r2,m} \tag{12}$$

设路由器 m 不空闲，即存在数据流，$B_{r,m}$ 不为 0，则

$$\lambda_r T_r + \sum_{m\in M} B_m B_{r,m} = W_r \tag{13}$$

$$\sum_{r\ni m}\lambda_r = C_m \tag{14}$$

由 (12)，(13) 式，可得

$$B_{r,m} = \left(\prod_{r'\ni m} B_{r',m}\right)^{\frac{\lambda_r}{C_m}} \approx 1 + \log\left(\prod_{r'\ni m} B_{r',m}\right)\frac{\lambda_r}{C_m} + \frac{1}{2!}\left(\log\left(\prod_{r'\ni m} B_{r',m}\right)\right)^2\left(\frac{\lambda_r}{C_m}\right)^2 \tag{15}$$

再代入 (13) 式，可得

$$\begin{aligned} W_r &= \lambda_r T_r + \sum_{l\in r} B_m B_{r,m} \\ &= \lambda_r T_r + \sum_{m\in r} B_m + \sum_{m\in r} B_m \log\left(\prod_{r'\ni m} B_{r',m}\right)\frac{\lambda_r}{C_m} \\ &\quad + \sum_{m\in r}\frac{1}{2!}B_m\left(\log\left(\prod_{r'\ni m} B_{r',m}\right)\right)^2\left(\frac{\lambda_r}{C_m}\right)^2 \end{aligned} \tag{16}$$

现在考虑 (15)，(16) 式的最优化问题，设 μ_m 为待定常数，对应的拉格朗日函数为

$$F = \sum_{R}\left(\frac{W_r - \sum_{m\in r} B_m}{\lambda_r} + \left(T_r + \frac{B_m \log\left(\prod_{r'\ni m} B_{r',m}\right)}{C_m}\right)\log\lambda_r + \sum_{r\ni m}\mu_m\left(\sum_{r\ni m}\lambda_r - C_m\right)\right)$$

根据 Kuhn-Tucker 定理 [5]，这个最优化问题的解满足：$\partial F/\partial\lambda_r = 0$, $r \in R$ 和 $\mu_m \geq 0$；且 $\sum_{r\ni m}\lambda_r < C_m \Rightarrow \mu_m = 0$. 经过计算可得

$$W_r = \lambda_r T_r + \sum_{m\in r} B_m + \sum_{m\in r} B_m \log\left(\prod_{r'\ni m} B_{r',m}\right)\frac{\lambda_r}{C_m} + \mu_m\lambda_r^2 \tag{17}$$

若其中 $\mu_m = \sum_{m\in r} \frac{1}{2!}\frac{B_m}{C_m^2}\left(\log\left(\prod_{r'\ni m} B_{r',m}\right)\right)^2$，则 (16) 式与 (17) 式相同．又由于目标函数是严格凸函数，最优解是唯一的，所以这个满足 (11) 式的最优化解就是稳态的速率分配．

证毕

Kelly 指出，对于数据业务，带宽分配结果 $\{\lambda_r\}$ 应该使总效用函数 $\sum_R \log\lambda_r$ 达到最大，实现比例公平 [6]；而 Massoulie 认为，带宽分配方案应该使总延时 $\sum_R 1/\lambda_r$ 达到最小，实现延时最小化公平 [4]．实际上，当往返时间很小、可以忽略不计时，通过网络设计可以保证 (16) 式中的参数 a 为正数．这时，(15) 式中定义的最优化问题实际上就是在最大化总效用函数与最小化总延时这两个目标之间的折衷，同时也是比例公平准则与延时最小化公平准则之间的一种折衷．

3 仿真结果

考虑多个流经过一个路由器的情况．设该路由器的输出链路带宽为 $C = 2000$ packet/s．有 N 个数据流 $\{l_i, 1 \geq i \leq N\}$ 竞争这个路由器的输出带宽和缓冲区资源，设 λ_i 为分配给数据流 l_i 的带宽．该路由器采用 LPFQ，为每个数据流提供一个队列，每个队列容量为 $B = 100$packet；与数据流 i 对应的队列中排队等待发送的数据量为 L_i，则数据流 l_i 的队列占用率 $B_i = L_i/B$，且 $0 < B_i < 1$．

第 1 种情况 设 $N = 4$，数据流 l_2, l_3, l_4 的队列中等待发送的数据量 L_2, L_3, L_4 较小，都为 1 到 10packet 之间的离散均匀分布．在这种情况下，数据流 l_1 所能得到的发送速率与其队列占用率之间的关系示于图 1．我们可以看到，对于一个流来说，当其他多数数据流行为友好、队列占用率较小时，它也必须保持较小的队列占用率才能得到与其他流相当的发送速率和延时性能，否则，它的发送速率和延时性能会收到 LPFQ 算法的惩罚．

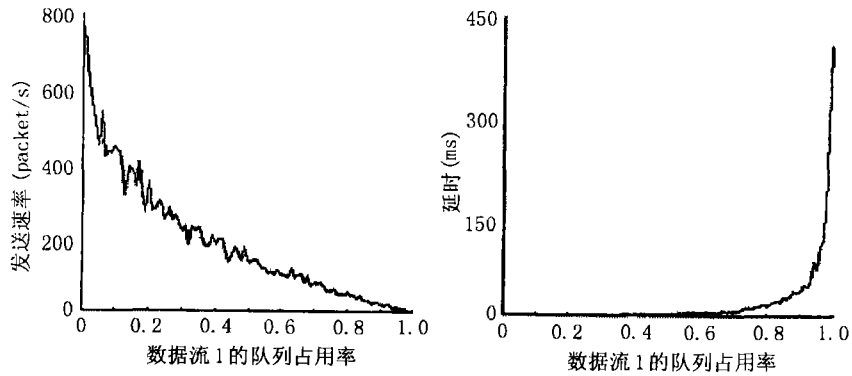

图 1 其他流行为友好，队列占用率较小时，发送速率、延时性能与队列占用率的关系

第 2 种情况 设 N =4，数据流 l_2, l_3, l_4 的队列中等待发送的数据量 L_2, L_3, L_4 较大，都为 90 到 100 packets 之间的离散均匀分布．在这种情况下，数据流所能得到的发送速率与其队列占用率之间的关系示于图 2．我们可以看到，当多数数据流不友好，占用大量队列容量时，那些友好的数据流将获得较大的发送速率和较好的延时性能．

如果我们假设数据流的到达过程为互相独立的泊松过程，通过仿真，可以得到与上面类似的结果．由于篇幅所限，这里不再列出图表，基本结果是：到达率高的数据流得到的发送速率和延时性能的稳态平均值均低于到达率较低的数据流．这是很容易理解的，到达率高，平均队列

占用率就较大，根据 LPFQ 的服务规则，它所得到的平均发送速率就较小，平均延时较大．另外，得到的网络利用率、网络总延时以及公平性均介于比例公平准则和延时最小化准则之间，达到了折衷的效果．

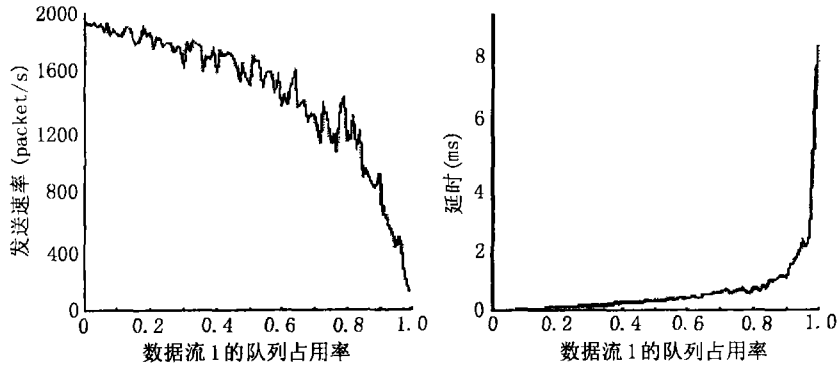

图 2　其他流不友好，队列占用率较大时，发送速率、延时性能与队列占用率的关系

4 结　论

通过前面的分析、仿真和比较，我们可以看到 LPFQ 的合理性和可行性．与其它的公平队列算法相比，LPFQ 算法的最大特点是：通过 LPFQ 的使用，可以使友好的用户数据流获得较高的传输性能，使恶意的数据流获得较差的传输性能，鼓励用户积极自觉的使用适当的端到端流量控制算法来调整注入网络的数据量；在使用 LPFQ 算法的网络中，我们可以实现网络总效用、网络性能以及资源分配公平性之间的一种较好的折衷．

虽然在网络内部节点处进行每个流的单独控制会引入一定的的复杂度，但随着简化算法的研究以及软硬件技术的发展，做到这些是不困难的 [7,8]．

本文中，除了公平性问题是在多节点的网络下证明以外，在对 LPFQ 性能研究时，只对数据流在单个节点的速率、延时等性能进行了分析和仿真．以后将进行多节点网络中 LPFQ 性能的仿真工作．另外，对前面提到的有门限的 LPFQ 以及带反馈信息或分级业务支持的 LPFQ 机制还在研究之中．

参 考 文 献

[1] S. Floyd, Promoting the use of end-to-end congestion control in the Internet, IEEE/ACM Trans. on Networking, 1999, 7(4), 458–472.

[2] T. R. Henderson, R. H. Katz, On improving the fairness of TCP congestion avoidance, In Proc. IEEE Globecom'98, Sydney, Australia, November 1998, 539–544.

[3] C. Albuquerque, T. Suda, B. J. Vickers, Fair queuing with feedback-based policing: Promoting fairness and preventing congestion collapse in the Internet, UCI-ICS Technical Report 99-26, University of California, Irvine, USA, September 1998.

[4] L. Massoulie, J. Roberts, Bandwidth sharing: objectives and algorithms, In Proc. IEEE Infocom'99, New York, USA, March 1999, 1395–1403.

[5] 郑宝玉，糜正琨，王良元，编著，通信工程中的最优化方法，北京，北京邮电学院出版社，1996, 13–19.

[6] F. P. Kelly, A. K. Maulloo, D. K. H. Tan, Rate control for communication networks: Shadow prices, proportional fairness and stability, Journal of the Operational Research Society, 1998, 49(1), 237–252.

[7] D. C. Stephens, J. C. R. Bennett, H. Zhang, Implementing scheduling algorithms in high-speed networks, IEEE J. on SAC, 1999, 17(6), 1145–1158.

[8] B. Suter, T. V. Lakshman, D. Stiliadis, A. K. Choudhury, Buffer management schemes for supporting TCP in gigabit routers with per-flow queueing, IEEE J. on SAC, 1999, 17(6), 1159–1169.

LOGARITHMIC PROPORTIONAL FAIR QUEUEING: A NEW SCHEDUCLING ALGORITHM

Zheng Youquan Feng Zhenming

(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract The scheduling algorithms of routers in the Internet are very important to the QoS (Quality of Service) which can be derived by users' flows. A new scheduling algorithm—Logarithmic Proportional Fair Queueing (LPFQ) scheduling algorithm is presented, and its characteristics are discussed using theoretical analysis and simulations. Theoretical analysis and simulation results demonstrate that LPFQ can not only achieve the compromise among fairness, network utility and total delay, but also punish malicious flows and encourage users to use proper end-to-end flow control algorithms.

Key words Fair queueing, Scheduling algorithm, Quality of service, Logarithmic proportional fair queueing

郑友泉： 男， 1974 年生，博士生，主要研究现代通信与计算机网络的建模与仿真、流量管理与拥塞控制等.

冯振明： 男， 1946 年生，教授，长期从事信号处理、 HFC 网络以及多媒体通信等多方面教学与科研工作.