Revisiting WFQ: Minimum Packet Lengths Tighten Delay and Fairness Bounds

Anirudha Sahoo and D. Manjunath

***Abstract*— In this paper we consider the bounds on the sample path discrepancy between the 'idealized' generalized processor sharing (GPS) and the 'practical' weighted fair queueing (WFQ) scheduling disciplines. We show that when both the minimum packet lengths and the weights are non zero, the discrepancy bounds can possibly be tighter than that in [1] and [2]. This new upper bound on the delay discrepancy is then used to provide an upper bound on the discrepancy in the instantaneous throughput, which can also be significantly tighter than those in [1] and [2].**

***Index Terms*— Scheduling, Weighted Fair Queueing (WFQ), Generalized Processor Sharing (GPS), fairness bound.**

I. Introduction

GENERALIZED processor sharing (GPS) generalizes the ideal of processor sharing (PS) and is a max-min fair way of sharing link capacity. Weighted fair queueing (WFQ) was first proposed as an emulation of the PS in [3] and its fairness issues were first analyzed in [1]. By assuming a maximum packet length, they obtained a bound on the *discrepancy*, the difference in the instantaneous throughput of a flow between the ideal GPS and the WFQ systems. It can be shown that this is the best discrepancy possible by a non preemptive scheduler. Since the instantaneous throughput difference between GPS and WFQ is bounded, it follows that the instantaneous queue lengths differences are also bounded. They also bound the difference in the delays in the WFQ and the GPS systems. In a widely cited paper, Parekh and Gallager [2] refer to WFQ as packetized-GPS or PGPS, and provide fairly simplified proofs for the delay and the fairness bounds between a GPS and the PGPS or WFQ systems. This is now considered the standard proof, e.g., see [4]. Since the above seminal papers, many variations of WFQ have been introduced, e.g., [5]–[9] that address implementation complexity and provide other useful properties.

To the best of our knowledge, all analyses of PGPS, WFQ and related schedulers consider only the maximum packet length in describing the discrepancy and none of them consider the case of a link also prescribing a minimum packet length. This is important because most networks do prescribe such a minimum corresponding to, for example, the header and the trailer lengths. In this paper, we show that for two important performance measures—delay and instantaneous throughput, the discrepancy between the GPS and WFQ (or PGPS) is much tighter when the link also prescribes a minimum packet length. Recall that the latter is also a relative fairness measure.

In the next section, we first derive an upper bound on the delay discrepancy when the minimum packet length is non zero. We will see that the correction to the well known bound of [1] and [2] can be significant. The new upper bound on the delay discrepancy is then used to provide an upper bound on the discrepancy in the instantaneous throughput, which is also significantly tighter than those in [1] and [2]. We conclude in Section III with examples and a brief discussion.

Manuscript received October 13, 2006. The associate editor coordinating the review of this letter and approving it for publication was Dr. Alex Sprintson.

A. Sahoo is with the Kanwal Rekhi School of Information Technology, Indian Institute of Technology Bombay, Powai, Mumbai - 400076, India (e-mail: sahoo@it.iitb.ac.in).

D. Manjunath is with the Department of Electrical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai - 400076, India (e-mail: dmanju@ee.iitb.ac.in).

Digital Object Identifier 10.1109/LCOMM.2007.061677.

II. Discrepancy Analysis with Minimum Packet Lengths

Recall that in the WFQ system, scheduling instants occur at the departure times (end of transmission). The new packet that is scheduled for transmission is the one that would have departed earliest in the corresponding GPS system from among those that are present in the system at the scheduling instant.

We consider a link of capacity C serving N flows indexed $i = 1, \ldots, N$. The packets from each flow are served according to FIFO. Let $L_{\min}$ and $L_{\max}$ be the minimum and maximum packet lengths that the link supports. Let ϕ_i be the weight of flow i. Recall that GPS and WFQ guarantee a minimum rate of $\frac{\phi_i}{\sum_i \phi_i}$ to flow i. Let

$$\phi_{\min} := \min_i \phi_i, \qquad \phi_{\max} := \max_i \phi_i, \qquad \phi = \sum_i \phi_i.$$

The arrival instants of the packets also remain the same in both systems. Note that these packets can arrive from any flow.

Our sample path analysis is along the lines of [2] and [4] and the notations are similar to that in [4]—we consider identical input sequences to both the GPS and the WFQ systems and analyze the discrepancies. Packets are numbered according to their departure times in the WFQ system.

Let $\hat{d}_k$ be the time at which the k-th packet, p_k, departs from the WFQ system, a_k its arrival time to both the systems and d_k its departure time in the GPS system and L_k the length of the packet. Let $\hat{s}_k$ be the time at which the transmission of packet p_k starts in the WFQ system.

The following lemma will be useful in the proof of the theorem and is a fairly straightforward consequence of the definition of WFQ. See [1], [2], [4] for a proof.

Lemma 1: If packets p_{k_1} and p_{k_2} are present in the system at a scheduling instant and if packet p_{k_1} is scheduled to transmit before packet p_{k_2} in WFQ, then packet p_{k_1} departs before packet p_{k_2} in the GPS. This, of course, implies that packet p_{k_1} departs before packet p_{k_2} in the WFQ system.

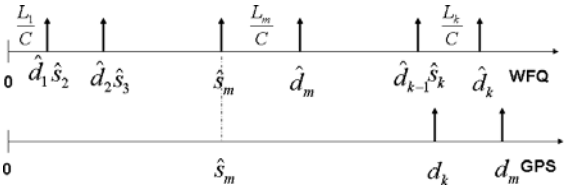

Fig. 1. Events of interest in the WFQ and GPS systems. (Adopted from [4])

The following theorem presents the tighter bound on the departure time of a packet in WFQ and GPS system.

Theorem 1: For all packets in the WFQ scheduler,

$$\hat{d}_k \leq d_k + \frac{L_{\max}}{C} - \frac{\left(\frac{\phi_{\min}}{\phi_{\max}}\right) L_{\min}}{C}. \tag{1}$$

Proof: Since both the GPS and WFQ are work conserving queueing disciplines, their busy periods will be identical and it suffices to consider a busy period that starts, without loss of generality, at time 0. Further, during a busy period, WFQ and GPS schedulers will transmit the same set of packets. Since the packets are indexed according to their departure in the WFQ, $\hat{d}_1 < \hat{d}_2 < \cdots$.

Since the busy period starts at 0, clearly, $\hat{d}_k = \frac{\sum_{i=1}^{k} L_i}{C}$. We need to consider two cases for the ordering of the departure times of packets in GPS relative to that in WFQ.

Case 1: None of the packets $p_1, p_2, \ldots, p_{k-1}$ departs after d_k in GPS. Hence, in GPS, at least $\sum_{i=1}^{k} L_i$ bits have been transmitted up to d_k and we have

$$d_k \geq \frac{\sum_{i=1}^{k} L_i}{C}.$$

Since $\hat{d}_k = \frac{\sum_{i=1}^{k} L_i}{C}$, the theorem is trivially true.

Case 2: There are some packets that depart earlier in WFQ than in GPS. These packets consume bandwidth that was meant for other packets and hence delay them. Consider such a packet p_m, $1 \leq m \leq k-1$ that has the largest index, such that $d_m > d_k$. Of the packets that should have departed after packet p_k, p_m is the last packet that departs earlier than p_k in WFQ and packets $p_{m+1}, \ldots, p_{k-1}$ depart earlier than packet p_k in both WFQ and GPS. This means, $d_{m+1} \leq d_k, d_{m+2} \leq d_k, \ldots, d_{k-1} \leq d_k$. By Lemma 1, packets $p_{m+1}, p_{m+2}, \ldots, p_k$ were not present in the WFQ system at time $\hat{s}_m$, because if they were, one of them would have been chosen for service instead of p_m. Figure 1 shows an example situation.

In the interval $[\hat{s}_m, d_k]$, GPS has transmitted complete packets $p_{m+1}, \ldots, p_k$. In the same interval, it would have also served some bits of packet p_m. Let L'_m be the minimum number of bits of packet p_m that was transmitted by GPS in this interval $[\hat{s}_m, d_k]$. Hence,

$$d_k \geq \hat{s}_m + \frac{\sum_{i=m+1}^{i=k} L_i}{C} + \frac{L'_m}{C}. \tag{2}$$

Let f_k and f_m be the flow to which packet p_k and p_m belong respectively. Since packet p_k was completely transmitted by GPS in the interval $[\hat{s}_m, d_k]$, (it arrived after $\hat{s}_m$ and departed at d_k) the number of bits of packet p_m that were transmitted in this interval is at least $\frac{\phi_{f_m}}{\phi_{f_k}} L_k$, i.e.,

$$L'_m \geq \frac{\phi_{f_m}}{\phi_{f_k}} L_k. \tag{3}$$

From Figure 1, it is clear that

$$\hat{d}_k = \hat{s}_m + \frac{L_m}{C} + \frac{\sum_{i=m+1}^{k} L_i}{C}. \tag{4}$$

Using (3) and (4) in (2), we get

$$d_k \geq \hat{d}_k - \frac{L_m}{C} + \frac{\phi_{f_m}}{\phi_{f_k}} \frac{L_k}{C}. \tag{5}$$

But $L_m \leq L_{\max}$ and $\left(\frac{\phi_{f_m}}{\phi_{f_k}}\right) L_k \geq \left(\frac{\phi_{\min}}{\phi_{\max}}\right) L_{\min}$. Hence, from (5), we get

$$d_k \quad \geq \quad \hat{d}_k - \frac{L_{\max}}{C} + \frac{\left(\frac{\phi_{\min}}{\phi_{\max}}\right) L_{\min}}{C} \tag{6}$$

and the theorem follows. □

This delay bound also impacts the fairness discrepancy of WFQ. Let $U^j(t)$ and $\hat{U}^j(t)$ be the number of bits of flow j transmitted by GPS and WFQ respectively up to time t. [1], [2] have shown that for all t and for all j, $U^j(t) - \hat{U}^j(t) \leq L_{\max}$. Theorem 2 below shows that this discrepancy is tighter if the link prescribes a minimum packet length.

Theorem 2: (i) Let L_k^j denote the number of bits in k-th packet of flow j, p_k^j. The fairness discrepancy between the WFQ and GPS systems satisfies the following inequality. For $\hat{d}_{k-1}^j \leq t \leq \hat{d}_k^j$,

$$U^j(t) - \hat{U}^j(t) \leq \begin{cases} L_{\max} - \frac{\phi_{\min}}{\phi_{\max}} L_{\min} \\ \qquad \text{if } L_k^j \leq L_{\max} - \frac{\phi_{\min}}{\phi_{\max}} L_{\min} \\ L_k^j - \left(\frac{\phi_j}{\phi} \left| L_{\max} - L_k^j - \frac{\phi_{\min}}{\phi_{\max}} L_{\min} \right| \right) \\ \qquad \text{if } L_k^j \geq L_{\max} - \frac{\phi_{\min}}{\phi_{\max}} L_{\min} \end{cases} \tag{7}$$

(ii) Further, for all t

$$U^j(t) - \hat{U}^j(t) \;\leq\; L_{\max} - \frac{\phi_{\min}}{\phi_{\max}} \frac{\phi_j}{\phi} L_{\min} \tag{8}$$

for all $t \geq 0$ and for $j = 1, \ldots, N$.

Proof: Consider packet p_k^j. Since packets within a flow are served in FIFO order, the total number of bits of flow j until the departure of packet p_k^j under the two systems should be equal, i.e.,

$$U^j(d_k^j) = \hat{U}^j(\hat{d}_k^j). \tag{9}$$

Since WFQ is a packet based scheduler, the entire link bandwidth is dedicated to a packet. Hence,

$$\hat{U}^j(\hat{d}_k^j) = \hat{U}^j(\hat{s}_k^j) + L_k^j, \tag{10}$$

From Theorem 1 we have

$$d_k^j \geq \hat{d}_k^j - \left(\frac{L_{\max}}{C} - \frac{\frac{\phi_{\min}}{\phi_{\max}} L_{\min}}{C} \right).$$

Changing signs, and hence the direction of the above inequality, and adding $\hat{s}_k^j$ to both sides, we get

$$\hat{s}_k^j - d_k^j \leq \hat{s}_k^j - \hat{d}_k^j + \frac{L_{\max}}{C} - \frac{\frac{\phi_{\min}}{\phi_{\max}} L_{\min}}{C}.$$

Since $\hat{d}_k^j - \hat{s}_k^j = \frac{L_k^j}{C}$, the above inequality can be rewritten as

$$\hat{s}_k^j \leq d_k^j + \frac{L_{\max} - L_k^j - \frac{\phi_{\min}}{\phi_{\max}} L_{\min}}{C}. \quad (11)$$

Denote $Z_k^j := L_{\max} - L_k^j - \frac{\phi_{\min}}{\phi_{\max}} L_{\min}$. Since both $U^j(t)$ and $\hat{U}^j(t)$ (and also the departure time of packets in both the systems) are non-decreasing, we can write the following inequality.

$$U^j(\hat{s}_k^j) \leq U^j\left(d_k^j + \frac{Z_k^j}{C}\right) \quad (12)$$

Using (10) in (12) we get

$$U^j(\hat{s}_k^j) - \hat{U}^j(\hat{s}_k^j) - L_k^j \leq U^j\left(d_k^j + \frac{Z_k^j}{C}\right) - \hat{U}^j(\hat{d}_k^j). \quad (13)$$

Z_k^j could be positive or negative and we consider these cases separately.

Case 1: $Z_k^j \geq 0$. In this case $U^j\left(d_k^j + \frac{Z_k^j}{C}\right)$ will be upper bounded by $U^j(d_k^j) + Z_k^j$. This is because the maximum number of bits that can be transmitted in time Z_k^j/C is Z_k^j. Using this bound and (9) in (13), we get

$$U^j(\hat{s}_k^j) - \hat{U}^j(\hat{s}_k^j) \leq L_k^j + Z_k^j.$$

Substituting for Z_k^j we get

$$U^j(\hat{s}_k^j) - \hat{U}^j(\hat{s}_k^j) \leq L_{\max} - \frac{\phi_{\min}}{\phi_{\max}} L_{\min}. \quad (14)$$

Case 2: $Z_k^j < 0$. In this case a simple upper bound on $U^j\left(d_k^j + \frac{Z_k^j}{C}\right)$ would be $U^j(d_k^j)$ and

$$U^j(\hat{s}_k^j) - \hat{U}^j(\hat{s}_k^j) \quad \leq \quad L_k^j$$

This can be further tightened by upper bounding $U^j\left(d_k^j + \frac{Z_k^j}{C}\right)$ by $U^j(d_k^j) - W_{\min}\left(\frac{Z_k^j}{C}\right)$ where $W_{\min}\left(\frac{Z_k^j}{C}\right)$ is the minimum work that will be done in the interval $\frac{|Z_k^j|}{C}$ in the GPS system. From (13), we get

$$U^j(\hat{s}_k^j) - \hat{U}^j(\hat{s}_k^j) \leq L_k^j + U^j\left(d_k^j\right) - \hat{U}^j(\hat{d}_k^j) - W_{\min}\left(\frac{Z_k^j}{C}\right)$$

$$\leq \quad L_k^j - \left(\frac{\phi_j}{\phi}|Z_k^j|\right) \quad (15)$$

In obtaining the last equality above, we have used (9) and the fact that minimum number bits of flow j that is transmitted in the GPS system in a time interval T is $\left(\frac{\phi_j}{\phi}\right)TC$.

The slope of $\hat{U}^j$ alternates between C when a packet from flow j is served and 0 when flow j is not being served. Since the slope of U^j also follows the same limits, for $\hat{d}_{k-1} \leq t < \hat{d}_k$, the difference $U^j(t) - \hat{U}^j(t)$ has its maximum value at $\hat{s}_k$, when the packet of flow j starts service in WFQ. Thus part (i) of the theorem holds.

To prove part (ii) consider the case of $Z_k^j < 0$, when the fairness discrepancy is given by (15). Note that $Z_k^j < 0$ occurs while $Z_k^j \geq 0$ need not occur. Now observe that $|Z_k^j|$ increases linearly with L_k^j when $Z_k^j < 0$ and that the slope of $\left(\frac{\phi_j}{\phi}|Z_k^j|\right)$ is $\frac{\phi_j}{\phi} \leq 1$. In the RHS of (15), we are subtracting $\left(\frac{\phi_j}{\phi}|Z_k^j|\right)$ from L_k^j. Hence the maximum value of the RHS in (15) occurs when $L_k^j = L_{\max}$ and we have

$$U^j(\hat{s}_k^j) - \hat{U}^j(\hat{s}_k^j) \quad \leq \quad L_{\max} - \frac{\phi_{\min}}{\phi_{\max}} \frac{\phi_j}{\phi} L_{\min}$$

Since $\frac{\phi_j}{\phi} \leq 1$, the bound obtained above for $Z_k^j < 0$ is looser than that obtained for $Z_k^j \geq 0$ in (14) and part (ii) of the theorem follows. □

III. Discussion

It is instructive to consider some special cases to see the impact of the tighter bounds on the discrepancy. First consider the case when all the flows have equal weight, i.e., $\phi_{\max} = \phi_{\min}$. Then (1) says that the delay discrepancy between the WFQ and GPS schedulers is upper bounded by $\frac{L_{\max} - L_{\min}}{C}$.

A second case of interest is that of fixed size packets, i.e., $L_{\max} = L_{\min}$ for which the delay discrepancy is upper bounded by $(1 - \frac{\phi_{\min}}{\phi_{\max}})L_{\max}$. This can be substantially smaller than the $L_{\max}$ bound. This case is of special interest because most streaming traffic (e.g., VoIP) use fixed size packets. Note that, if in this system $\phi_{\min} = \phi_{\max}$ then the delays in the WFQ system is never greater than the delay in the GPS system.

References

[1] A. Greenberg and N. Madras, "How fair is fair queuing?," *Journal of the ACM*, vol. 39, pp. 568–598, July 1992.

[2] A. Parekh and R. G. Gallager, "A generalised processor sharing approach to flow control in integrated services networks: the single–node case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 344–357, June 1993.

[3] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," *Internetworking Res. and Exp.*, vol. 1, pp. 3–26, 1990.

[4] A. Kumar, D. Manjunath, and J. Kuri, *Communication Networking - An Analytical Approach*. Morgan Kaufmann, 2004.

[5] S. J. Golestani, "A self clocked fair queueing scheme for broadband applications," in *Proc. IEEE INFOCOM 1994*, pp. 636–646.

[6] J. C. Bennett and H. Zhang, "Hierarchical fair queuing algorithms," *IEEE/ACM Trans. Networking*, vol. 5, pp. 675–689, Oct. 1997.

[7] P. Goyal, H. M. Vin, and H. Cheng, "Start-time fair queueing: a scheduling algorithm for integrated services packet switching networks," *IEEE/ACM Trans. Networking*, vol. 5, pp. 690–704, Oct. 1997.

[8] A. Varma and D. Stiliadis, "Hardware implementation of fair queueing algorithms for asynchronous transfer networks," *IEEE Commun. Mag.*, vol. 35, no. 12, pp. 54–68, Dec. 1997.

[9] D. Stiliadis and A. Varma, "Efficient fair queueing algorithms for packet-switched networks," *IEEE/ACM Trans. Networking*, vol. 6, pp. 175–185, April 1998.