Probability in the Engineering and Informational Sciences, **24**, 2010, 485–490.
doi:10.1017/S0269964810000124

RATIO BETWEEN AVERAGE SOJOURN TIMES UNDER PROCESSOR SHARING AND FAIR SOJOURN PROTOCOL

YINGDONG LU
IBM T.J.Watson Research Center
Yorktown Heights, NY 10598
E-mail: yingdong@us.ibm.com

Under very general assumptions, we prove that the ratio between the average sojourn time in a queue under the processor sharing (PS) and the same queue under the fair sojourn protocol (FSP) can be bounded above, and we derive such a bound in terms of the long-run average system size of the PS queue seen by arrivals.

1. INTRODUCTION

The fair sojourn protocol (FSP) is a scheduling policy proposed by Friedman and Henderson [1] for the purpose of balancing fairness and efficiency in a queuing system. More specifically, whenever an arrival or departure event happens, the server will evaluate the jobs that still need to be served and then serve the one that will depart first if a processor-sharing (PS) rule, which allocates the service capacity evenly to each job in the system, had been followed. Therefore, while maintaining the departure sequences of the PS queue, the FSP allows the server to be fully devoted to a job and thus reduce the overall sojourn time. In their work, Friedman and Henderson also showed that a queue under the FSP can outperform a queue with the same arrival sequence under PS in the following sense: The jobs in the FSP queue have shorter sojourn times than in the PS queue. It will then be of considerable theoretic and practical interests to quantitatively characterize this advantage. In Wierman and Harcol-Balter [2], it is shown that under the assumption of Poisson arrivals, the ratio between the long-run average sojourn time of PS queue and FSP queue cannot be more than $1 + \rho$, where ρ is the traffic intensity.

In this article, we intend to establish a similar bound for a very general class of queuing systems. To be more specific, we show that if the first two moments of the

© Cambridge University Press 2010 0269-9648/10 $25.00

https://doi.org/10.1017/S0269964810000124 Published online by Cambridge University Press

sojourn time and the system sizes seen by arrival can have ergodic limit for queues under both FSP and PS policies, then the ratio between the average sojourn time of the PS queue and the FSP queue can be bounded by $(q^{\mathrm{PS}} \vee 1) + 1$, where q^{PS} denotes the average system size under the PS seen by arrivals. Evidently, there is still plenty of room for improvement between this general result and the special case of the Poisson arrivals, where the bound is less than 2 for a stable queue; much more detailed analysis for subclasses of queuing system should be conducted in the future.

2. THE RESULT AND THE PROOF

Consider the following queuing system: The arrivals follow a general point process, with interarrival time A_n between the nth and $(n+1)$st arrival, and the workload of the nth arrival is S_n. We assume the following.

ASSUMPTION A: *For the queue under each policy,*

$$\frac{1}{N}\sum_{n=1}^{N}\left(D_n^{\mathrm{FSP}}\right)^i, \qquad \frac{1}{N}\sum_{n=1}^{N}\left(D_n^{\mathrm{PS}}\right)^i,$$

$$\frac{1}{N}\sum_{n=1}^{N}\left(Q_n^{\mathrm{PS}}\right)^i, \qquad \frac{1}{N}\sum_{n=1}^{N}\left(Q_n^{\mathrm{PS}}\right)^i, \qquad i = 1, 2,$$

converge almost surely to real numbers, where D_n^{PS} and D_n^{FSP} denote the sojourn time for the nth arrival under the PS and FSP, respectively, and Q_n^{PS} and Q_n^{FSP} denote the number of jobs in the system that the jth arrival sees.

The long-run average sojourn time under the FSP and PS respectively will be denoted by $\mathsf{E}[D^{\mathrm{PS}}]$ and $\mathsf{E}[D^{\mathrm{FSP}}]$. It is apparent that many practical systems such as a $GI/GI/1$ queue with traffic intensity less than one easily satisfy this condition.

Now define the following:

- $D_{k,j}^{\mathrm{PS}}$: sojourn time for the k-arrival under the PS assuming that there are only the first j arrivals;
- $D_{k,j}^{\mathrm{FSP}}$: sojourn time for the k-arrival under the FSP assuming that there are only the first j arrivals.

The following lemma ensure that the long-run average of these two quantities are the same as the long-run average of the sojourn time.

LEMMA 1: *Under Assumption A, we have, almost surely,*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} D_{n,N}^{\mathrm{PS}} = \mathsf{E}[D^{\mathrm{PS}}], \tag{1}$$

https://doi.org/10.1017/S0269964810000124 Published online by Cambridge University Press

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}D_{n,N}^{\mathrm{FSP}}=\mathsf{E}[D^{\mathrm{FSP}}]. \tag{2}$$

PROOF: Let us first consider the case of the PS queue. We know that $D_{n,N}^{\mathrm{PS}}\le D_n^{\mathrm{PS}}$ for any $N\ge n$; furthermore, if job n finishes before the $(N+1)$st arrival, then $D_{n,N}^{\mathrm{PS}}=D_n^{\mathrm{PS}}$. Therefore,

$$\sum_{k=1}^{n}[D_k^{\mathrm{PS}}-D_{k,n}^{\mathrm{PS}}]=\sum_{k=n-Q_n^{\mathrm{PS}}}^{n}[D_k^{\mathrm{PS}}-D_{k,n}^{\mathrm{PS}}]\le\sum_{k=n-Q_n^{\mathrm{PS}}}^{n}S_k. \tag{3}$$

Meanwhile,

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{j=n-Q_n^{\mathrm{PS}}}^{n}S_j\le\frac{1}{N}\sum_{j=1}^{N}S_jQ_j^{\mathrm{PS}}\le\frac{1}{2N}\sum_{j=1}^{N}[S_j^2+(Q_j^{\mathrm{PS}})^2],$$

where the first inequality is the result of exchange of the summation. From Assumption A, we know that $1/N\sum_{j=1}^{N}[(Q_j^{\mathrm{PS}})^2]$ converges as $N\to\infty$. Therefore,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}[D_n^{\mathrm{PS}}-D_{n,N}^{\mathrm{PS}}]=0,$$

from basic properties of infinite series.

For the case of the FSP, we have a similar bound to (3):

$$\sum_{k=1}^{n}[D_k^{\mathrm{FSP}}-D_{k,n}^{\mathrm{FSP}}]=\sum_{k=n-Q_n^{\mathrm{FSP}}}^{n}[D_k^{\mathrm{FSP}}-D_{k,n}^{\mathrm{FSP}}]\le\sum_{k=n-Q_n^{\mathrm{FSP}}}^{n}S_k. \tag{4}$$

By the same arguments, Eq. (4) will lead to the conclusion. ■

Suppose that we have the same Q_0 number of initial jobs in both systems. Define, for $n=1,2,\ldots,$

$$\Delta_n^{\mathrm{PS}}=\sum_{k=1}^{Q_0+n}D_{k,n}^{\mathrm{PS}}-\sum_{k=1}^{Q_0+n-1}D_{k,n-1}^{\mathrm{PS}},$$

$$\Delta_n^{\mathrm{FSP}}=\sum_{k=1}^{Q_0+n}D_{k,n}^{\mathrm{FSP}}-\sum_{k=1}^{Q_0+n-1}D_{k,n-1}^{\mathrm{FSP}}.$$

We have the following relationship between these two increments.

LEMMA 2: *Under Assumption A, the following inequality holds almost surely for every sample path, for* $n=1,2,\ldots,$

$$\frac{\Delta_n^{\mathrm{PS}}}{\Delta_n^{\mathrm{FSP}}}\le(Q_n^{\mathrm{PS}}\vee 1)+1. \tag{5}$$

https://doi.org/10.1017/S0269964810000124 Published online by Cambridge University Press

PROOF: In the proof of the this lemma, we will use a sample path argument, so all of the statements are true almost surely on every sample path. For the nth arrival, let us first assume that both queues are not empty and denote the remaining workload for jobs in the PS queue respectively as follows,

$$\tilde{R}_1, \tilde{R}_2, \ldots, \tilde{R}_{Q_n^{\mathrm{PS}}},$$

with the assumption of $\tilde{R}_k \le \tilde{R}_{k+1}$. Then for those jobs that are also present in the FSP queue, denote their remaining workload as $\hat{R}_j$ for the same index. Furthermore, assume that the nth arrival has workload S_n, which satisfies $S_n \in [\tilde{R}_i, \tilde{R}_{i+1}]$ for some $i \in \{0, 1, \ldots, Q_n^{\mathrm{PS}}\}$ where $\tilde{R}_0$ and $\hat{R}_0$ are defined to be zero.

It is apparent that the sojourn time of any job that has been finished before the arrival of the nth job will not be affected. For those remaining job with index $k \le i$ in the PS queue, assuming that there are only the first $n-1$ arrivals, the remaining sojourn time will be $\sum_{j=0}^{k-1}(Q_n^{\mathrm{PS}} - j)(\tilde{R}_{j+1} - \tilde{R}_j)$. Considering the new arrival, the sojourn time becomes $\sum_{j=0}^{k-1}(Q_n^{\mathrm{PS}} - j + 1)(\tilde{R}_{j+1} - \tilde{R}_j)$. Taking their differences, we get, for $k \le i$,

$$D^{\mathrm{PS}}_{\sigma(k),n} - D^{\mathrm{PS}}_{\sigma(k),n-1} = \tilde{R}_k, \tag{6}$$

where $\sigma(k)$ denotes the arrival order of the job whose remaining service job is $\tilde{R}_k$. For the nth arrival,

$$D^{\mathrm{PS}}_{n,n} = \sum_{j=0}^{i-1}(Q_n^{\mathrm{PS}} - j + 1)(\tilde{R}_{j+1} - \tilde{R}_j) + (Q_n^{\mathrm{PS}} - i + 1)(S_n - \tilde{R}_i). \tag{7}$$

Similarly, for $k > i$, the sojourn time assuming only the first $n-1$ arrivals, $\sum_{j=0}^{k-1}(Q_n^{\mathrm{PS}} - j)(\tilde{R}_{j+1} - \tilde{R}_j)$; the sojourn time after considering the nth arrival is

$$\sum_{j=0}^{i-1}(Q_n^{\mathrm{PS}} - j + 1)(\tilde{R}_{j+1} - \tilde{R}_j) + (Q_n^{\mathrm{PS}} - i + 1)(S_n - \tilde{R}_i)$$
$$+ (Q_n^{\mathrm{PS}} - i)(\tilde{R}_{i+1} - S_n) + \sum_{j=i+1}^{k-1}(Q_n^{\mathrm{PS}} - j)(\tilde{R}_{j+1} - \tilde{R}_j).$$

Again, their differences will be

$$D^{\mathrm{PS}}_{\sigma(k),n} - D^{\mathrm{PS}}_{\sigma(k),n-1} = S_n. \tag{8}$$

Combining expressions (6)–(8), we obtain

$$\Delta_n^{\mathrm{PS}} = 2\left(\sum_{k=1}^{i}\tilde{R}_k + (Q_n^{\mathrm{PS}} - i)S_n\right) + S_n \le (2Q_n^{\mathrm{PS}} + 1)S_n. \tag{9}$$

https://doi.org/10.1017/S0269964810000124 Published online by Cambridge University Press

Now, let us consider the FSP system; note that the scheduling is conducted according to status of the PS queue. Therefore, if for all $i = 1, 2, \ldots,$ is Q_n^{PS}, $\tilde{R}_i \le S_n$, then the nth job will be scheduled after all the existing jobs; hence, $\Delta_n^{\mathrm{FSP}} = \sum_{i=1}^{Q^{\mathrm{FSP}}} \hat{R}_i + S_n$. However, $\sum_{i=1}^{Q^{\mathrm{FSP}}} \hat{R}_i$ is just the total work load before the nth arrival, which is the same as $\sum_{i=1}^{Q^{\mathrm{PS}}} \tilde{R}_i$; thus in this case,

$$\frac{\Delta_n^{\mathrm{PS}}}{\Delta_n^{\mathrm{FSP}}} = \frac{2\sum_{i=1}^{Q^{\mathrm{PS}}} \tilde{R}_i + S_n}{\sum_{i=1}^{Q^{\mathrm{FSP}}} \hat{R}_i + S_n} \le 2. \tag{10}$$

If, otherwise, there is at least one $\tilde{R}_i > S_i$, then there is at least one job in the FSP queue that will be processed after job n; its sojourn time will increase at least by S_n. Therefore, the difference will include this difference and sojourn time of the nth job, which is lower bounded by S_n:

$$\Delta_n^{\mathrm{FSP}} \ge 2S_n. \tag{11}$$

The case when either Q_n^{PS} or Q_n^{FSP} is zero, from the fact that the workloads of the two queues are always the same, we can conclude that both queues are empty; hence, $\Delta_n^{\mathrm{FSP}} = \Delta_n^{\mathrm{PS}}$. In conjunction with Inequalities (9)–(11), we can obtain inequality (5). ■

To obtain our main result, we need the following technical lemma,

LEMMA 3: *Let* $\{a_i\}$ *and* $\{b_i\}$ *be two nonnegative series, satisfying* $(1/N)\sum_{i=1}^{N} a_i \to a$ *and* $(1/N)\sum_{i=1}^{N} b_i \to b$ *as* $N \to \infty$ *for* $a \ge 0$ *and* $b \ge 0$, *then* $(1/N)\sum_{i=1}^{N} a_i b_i \to ab$ *as* $N \to \infty$.

PROOF: The Abel summation formula gives

$$\sum_{i=1}^{N} a_i b_i = \sum_{i=1}^{N-1} A_i(b_i - b_{i+1}) + A_N b_N,$$

where $A_i = \sum_{j=1}^{i} a_i$. So, we can write

$$\begin{aligned}\frac{1}{N}\sum_{i}^{N} a_i b_i &= \frac{1}{N}\sum_{i=1}^{N-1}(A_i - ai)(b_i - b_{i+1}) + a\frac{1}{N}\sum_{i=1}^{N-1} i(b_i - b_{i+1}) + \frac{1}{N}A_N b_N \\ &= \frac{1}{N}\sum_{i=1}^{N-1}(A_i - ai)(b_i - b_{i+1}) + \frac{1}{N}(A_N - aN)b_N + a\frac{1}{N}\sum_{i}^{N} b_i.\end{aligned}$$

The second equality is, again, because of the Abel summation, which implies $\sum_{i=1}^{N-1} i(b_i - b_{i+1}) = \sum_{i=1}^{N} b_i - Nb_N$. The fact that $\lim_{i\to\infty} A_i/i = a$ implies that

$$\lim_{N\to\infty} \frac{\sum_{i=1}^{N-1} A_i(b_i - b_{i+1})}{\sum_{i=1}^{N-1} i(b_i - b_{i+1})} = a.$$

Therefore, the result follows. ■

https://doi.org/10.1017/S0269964810000124 Published online by Cambridge University Press

THEOREM 4: *For queue systems satisfing Assumption A,*

$$\frac{\mathsf{E}[D^{\mathrm{PS}}]}{\mathsf{E}[D^{\mathrm{FSP}}]} \le (q^{\mathrm{PS}} \vee 1) + 1. \tag{12}$$

PROOF: From Lemma 1, we know that it is enough to show that

$$\lim_{j\to\infty} \frac{\sum_{k=1}^{j+Q_0} D_{k,j}^{\mathrm{PS}}}{\sum_{k=1}^{j+Q_0} D_{k,j}^{\mathrm{FSP}}} \le (q^{\mathrm{PS}} \vee 1) + 1.$$

Meanwhile,

$$\begin{aligned}\frac{\sum_{k=1}^{j+Q_0} D_{k,j}^{\mathrm{PS}}}{\sum_{k=1}^{j+Q_0} D_{k,j}^{\mathrm{FSP}}} &= \frac{\sum_{k=1}^{Q_0} D_{k,0}^{\mathrm{PS}} + \sum_{k=1}^{j} \Delta_k^{\mathrm{PS}}}{\sum_{k=1}^{Q_0} D_{k,0}^{\mathrm{FSP}} + \sum_{k=1}^{j} \Delta_k^{\mathrm{FSP}}} \\ &\le \frac{\sum_{k=1}^{Q_0} D_{k,0}^{\mathrm{PS}} + \sum_{k=1}^{j} ((2Q_k^{\mathrm{PS}}) \vee 1 + 1)\Delta_k^{\mathrm{FSP}}}{\sum_{k=1}^{Q_0} D_{k,0}^{\mathrm{FSP}} + \sum_{k=1}^{j} \Delta_k^{\mathrm{FSP}}}, \quad \forall j,\end{aligned}$$

where the inequality is due to Inequality (5). With the condition that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} Q_n^{\mathrm{PS}} = q^{\mathrm{PS}},$$

due to Assumption A, Lemma 3 implies that, almost surely,

$$\lim_{j\to\infty} \frac{\sum_{k=1}^{j} Q_k^{\mathrm{PS}} \Delta_k^{\mathrm{FSP}}}{\sum_{k=1}^{j} \Delta_k^{\mathrm{FSP}}}$$

exists and it is equal to q^{PS}. Furthermore, because both $\sum_{j=1}^{\infty} \Delta_j^{\mathrm{PS}}$ and $\sum_{j=1}^{\infty} \Delta_j^{\mathrm{FSP}}$ diverge, the finite terms $\sum_{k=1}^{Q_0} D_{k,0}^{\mathrm{PS}}$ and $\sum_{k=1}^{Q_0} D_{k,0}^{\mathrm{FSP}}$ will not affect the limit. The result follows. ■

References

1. Friedman, E.J. & Henderson, S.G. (2003). Fairness and efficieny in web server protocols. *Proceedings of ACM/SIGMETRIC'03*.
2. Wierman, A. & Harchol-Balter, M. (2003). Bounds on a fair policy with near optimal performance. Technical report, Pittsburg: Cornegie Mellon University.

https://doi.org/10.1017/S0269964810000124 Published online by Cambridge University Press