J. Appl. Prob. **43,** 523–534 (2006)
Printed in Israel

ON THE NONOPTIMALITY OF THE FOREGROUND–BACKGROUND DISCIPLINE FOR IMRL SERVICE TIMES

S. AALTO,* *Helsinki University of Technology*
U. AYESTA,** *CWI*

Abstract

It is known that for decreasing hazard rate (DHR) service times the foreground–background discipline (FB) minimizes the mean delay in the M/G/1 queue among all work-conserving and nonanticipating service disciplines. It is believed that a similar result is valid for increasing mean residual lifetime (IMRL) service times. However, on the one hand, we point out a flaw in an earlier proof of the latter result and construct a counter-example that demonstrates that FB is not necessarily optimal within class IMRL. On the other hand, we prove that the mean delay for FB is smaller than that of the processor-sharing discipline within class IMRL, giving a weaker version of an earlier hypothesis.

Keywords: Queueing theory; scheduling; M/G/1; IMRL; mean delay; FB; PS; MLPS

2000 Mathematics Subject Classification: Primary 60K25
Secondary 68M20; 90B22; 90B36

1. Introduction

Consider an M/G/1 queue with arrival rate λ, mean service time $\mathrm{E}[S]$, and load

$$\rho = \lambda\,\mathrm{E}[S] < 1.$$

Jobs are served according to a work-conserving and nonanticipating service (scheduling) discipline π. A discipline is work conserving if it does not idle when there are jobs waiting, and nonanticipating if the remaining service times of jobs are not known by the server. Let Π denote the family of such service disciplines. For example, the well-known disciplines FCFS (first-come–first-served) and PS (processor-sharing) belong to this family, while SRPT (shortest remaining processing time) does not. Let $F(x) = \mathrm{P}\{S \le x\}$, $x \ge 0$, denote the cumulative service time distribution function of any job. Define $\bar{F}(x) = 1 - F(x)$, and assume that $\bar{F}(x) > 0$ for all x.

If the service time distribution has density $f(x)$, the hazard rate, $h(x)$, is defined by

$$h(x) = \frac{f(x)}{\bar{F}(x)} = \frac{f(x)}{\int_x^\infty f(y)\,\mathrm{d}y}.$$

A service time distribution belongs to class DHR (decreasing hazard rate) if $h(x)$ is decreasing for all x, i.e. $h(x) \ge h(y)$ whenever $x \le y$.

Received 23 September 2005; revision received 15 February 2006.
* Postal address: Networking Laboratory, Helsinki University of Technology, PO Box 3000, FIN-02015 HUT, Finland.
Email address: samuli.aalto@tkk.fi
** Current address: LAAS-CNRS, 7 Avenue de Colonel Roche, 31 077 Toulouse Cedex 4, France.
Email address: urtzi.ayesta@laas.fr

Yashkov [7] has shown that within the class DHR the mean delay is minimized by the foreground–background discipline (FB), which gives priority to the job with the least attained service. In fact, Righter and Shanthikumar [4] have proved that FB minimizes not only the mean delay and the mean queue length, but even the queue length in the stochastic sense. FB is also known as FBPS (feedback processor-sharing), LAST (least attained service time), and LAS (least attained service).

For all x, define

$$H(x) = \frac{\bar{F}(x)}{\int_x^\infty \bar{F}(y)\,\mathrm{d}y}. \tag{1.1}$$

A service time distribution belongs to class IMRL (increasing mean residual lifetime) if $H(x)$ is decreasing for all x, i.e. $H(x) \geq H(y)$ whenever $x \leq y$. This is due to the fact that

$$\mathrm{E}[S - x \mid S > x] = \frac{\int_x^\infty \bar{F}(y)\,\mathrm{d}y}{\bar{F}(x)} = \frac{1}{H(x)}. \tag{1.2}$$

It is known that IMRL is a weaker condition than DHR. In other words, DHR $\subset$ IMRL. Righter *et al.* [5, Theorem 3.14] stated that FB minimizes the mean delay even within class IMRL. (Unfortunately, there is a misprint in the abstract of [5] stating just the opposite. The correct form is given in [5, Theorem 3.14].)

A still more general class consists of those service time distributions for which $C^2[S] \geq 1$, where $C^2[S]$ denotes the squared coefficient of variation of the service time distribution, written in terms of the variance of the service time distribution, $D^2[S]$, as $C^2[S] = D^2[S]/\mathrm{E}[S]^2$. Wierman *et al.* [6, Example 1] demonstrated, by constructing a counter-example, that FB is not optimal within this class. In particular, they disproved the hypothesis of Coffman and Denning [2, pp. 188–189] that the mean delay for FB would be smaller than that of PS whenever $C^2[S] > 1$. The distribution given in their counter-example, while having a greater squared coefficient of variation than 1, does not belong to class IMRL. (Unfortunately, there is a misprint in [6, Example 1]. The corrected version reads as follows: $\mathrm{P}\{S = 1\} = \frac{4}{5} + \varepsilon$ and $\mathrm{P}\{S = 6\} = \frac{1}{5} - \varepsilon$. Then $C^2[S] > 1$ for any ε, $0 < \varepsilon < \frac{1}{10}$.)

In this paper we prove that, in contradiction with [5, Theorem 3.14], FB does *not* minimize the mean delay within class IMRL. More specifically, we first identify a flaw in the proof of [5, Theorem 3.14] that cannot be overcome. Then we choose a service time distribution that belongs to IMRL but not to DHR, and construct a discipline for which the mean delay is smaller than that of FB. However, we prove that the mean delay for FB is smaller than that of PS within class IMRL, giving a weaker version of the hypothesis of Coffman and Denning [2, pp. 188–189].

The rest of the paper is organized as follows. First, in Section 2, we recall an essential point from [5], which is a relationship between the mean delay and the so-called level-x workload. Then, in Section 3, we consider the sample paths of the level-x workload process and prove that FB is not pathwise optimal. In Section 4, we prove that FB is also not optimal with respect to the mean level-x workload, but still outperforms PS. In Section 5, we finally construct the counter-example demonstrating that FB does not minimize the mean delay within class IMRL.

2. Relationship between mean delay and level-x workload

Consider a single-server queueing system starting empty at time $t = 0$ and obeying a service discipline $\pi \in \Pi$. We assume that jobs arrive one at a time. They are indexed by $i = 1, 2, \ldots$, according to their arrival order.

https://doi.org/10.1239/jap/1152413739 Published online by Cambridge University Press

Let A_i denote the arrival epoch of job i and S_i its service time. In addition, let $X_i^\pi(t)$ denote the attained service of job i at time t. Then let $\mathcal{A}(t)$ and $\mathcal{N}^\pi(t)$ respectively denote the set of jobs that have arrived by time t and those still in the system at time t:

$$\mathcal{A}(t) = \{i\colon A_i \le t\}, \qquad \mathcal{N}^\pi(t) = \{i \in \mathcal{A}(t)\colon X_i^\pi(t) < S_i\}.$$

For any $x > 0$, let $\mathcal{N}_x^\pi(t)$ denote the set of those jobs in the system whose attained service is less than x:

$$\mathcal{N}_x^\pi(t) = \{i \in \mathcal{A}(t)\colon X_i^\pi(t) < \min\{x, S_i\}\}.$$

Furthermore, let $A(t) = |\mathcal{A}(t)|$, $N^\pi(t) = |\mathcal{N}^\pi(t)|$, and $N_x^\pi(t) = |\mathcal{N}_x^\pi(t)|$, where the modulus of a set denotes its cardinality.

For any $x \ge 0$, let $V_x^\pi(t)$ denote the workload in the system at time t contributed by those jobs with attained service less than x, which for brevity we call the *level-x workload*:

$$V_x^\pi(t) = \sum_{i \in \mathcal{N}_x^\pi(t)} (S_i - X_i^\pi(t)). \tag{2.1}$$

Note that in the limit $x \to \infty$ the level-x workload equals the ordinary workload of this system, i.e. the sum of remaining service times of all jobs, which is the same for all work-conserving disciplines.

To obtain another expression for the level-x workload, let $R_x^\pi(t)$ denote the total rate at which service is provided to the jobs with attained service less than x at time t. Whenever there are such jobs, the level-x workload decreases continuously with this rate. However, it may also decrease discontinuously: whenever the attained service of a job reaches the truncation threshold x, such that the job is no longer a member of $\mathcal{N}_x^\pi(t)$, the level-x workload decreases by a step that equals the remaining service time of that job. Thus, we have

$$V_x^\pi(t) = \sum_{i \in \mathcal{A}(t)} S_i - \int_0^t R_x^\pi(u)\,\mathrm{d}u - \sum_{i \in \mathcal{A}(t)\setminus\mathcal{N}_x^\pi(t)} (S_i - \min\{x, S_i\}). \tag{2.2}$$

Consider the M/G/1 queue with $\rho < 1$. Let $\bar{N}^\pi$ denote the steady-state mean number of jobs in the system and $\bar{V}_x^\pi$ the steady-state mean level-x workload. Righter *et al.* [5, Lemma 3.12] showed that, for any $\pi \in \Pi$,

$$\bar{N}^\pi = \int_{0-}^\infty H(x)\,\mathrm{d}\bar{V}_x^\pi, \tag{2.3}$$

where $H(x)$ is as given in (1.1). In fact, Righter *et al.* [5] defined the level-x workload as the sum of the remaining service times of those jobs in the system whose attained service is less than *or equal to* a given truncation threshold x. However, since $H(x)$ is continuous from the right, (2.3) is valid also with our definition.

Assume, then, that the function $H(x)$ is monotone, meaning that the service time distribution belongs to either IMRL or DMRL (decreasing mean residual lifetime). In this case the mean number of jobs in two systems with disciplines $\pi, \pi' \in \Pi$, respectively, may be compared as follows:

$$\bar{N}^\pi - \bar{N}^{\pi'} = -\int_0^\infty (\bar{V}_x^\pi - \bar{V}_x^{\pi'})\,\mathrm{d}H(x). \tag{2.4}$$

This equation follows from (2.3) after partial integration, and can be found from the proof of [5, Theorem 3.14]. Therefore, if $\bar{V}_x^\pi \le \bar{V}_x^{\pi'}$ for all x and the service time distribution belongs to class IMRL, and $1/H(x)$ and $-H(x)$ are thus increasing, then $\bar{N}^\pi \le \bar{N}^{\pi'}$.

https://doi.org/10.1239/jap/1152413739 Published online by Cambridge University Press

Let $\bar{T}^{\pi}$ denote the mean delay of a job. By applying Little's result, i.e. $\bar{N}^{\pi} = \lambda \bar{T}^{\pi}$, we finally obtain the following relationship between the mean delay and level-x workload.

Proposition 2.1. *Let $\pi, \pi' \in \Pi$. Assume that the service time distribution belongs to class IMRL. If $\bar{V}_x^{\pi} \leq \bar{V}_x^{\pi'}$ for all x, then $\bar{T}^{\pi} \leq \bar{T}^{\pi'}$.*

Righter *et al.* [5] used this result, which is valid as stated, to show that FB minimizes the mean delay within class IMRL. The problem lies in their Lemma 3.5, where they stated that FB is optimal with respect to the level-x workload even for each *sample path*. However, in Proposition 3.2, below, we show that this is not the case.

3. Sample path results for the level-x workload

Consider disciplines $\pi \in \Pi$ such that priority is given to jobs with attained service time less than some threshold $x > 0$, and assume that FCFS is used for all such jobs. Let FCFS_x denote the family of such disciplines.

First we show that these disciplines are at least as good as FB with respect to the level-x workload.

Proposition 3.1. *Let $x > 0$ and $\pi^* \in \mathrm{FCFS}_x$. Then, for any $t \geq 0$, $V_x^{\pi^*}(t) \leq V_x^{\mathrm{FB}}(t)$.*

Proof. Disciplines π^* and FB both give full priority to the jobs with attained service less than x. Thus, $R_x^{\pi^*}(t) = R_x^{\mathrm{FB}}(t)$ for all $t \geq 0$ and, consequently,

$$\int_0^t R_x^{\pi^*}(u)\,\mathrm{d}u = \int_0^t R_x^{\mathrm{FB}}(u)\,\mathrm{d}u.$$

It follows that the second term in (2.2), as well as the first, is the same for both disciplines.

Let us consider the third term in (2.2). Note first that, for any $\pi \in \Pi$,

$$\sum_{i \in \mathcal{A}(t) \setminus \mathcal{N}_x^{\pi}(t)} (S_i - \min\{x, S_i\}) = \sum_{i \in \mathcal{A}_x(t) \setminus \mathcal{N}_x^{\pi}(t)} (S_i - \min\{x, S_i\}),$$

where

$$\mathcal{A}_x(t) = \{i \in \mathcal{A}(t) \colon S_i > x\}.$$

In the FB system the jobs with $S_i > x$ reach the attained service level x in batches. In the corresponding π^* system, where the discipline is FCFS below this level, the same jobs reach level x one by one, in order, in such a way that the job that arrived last among those with service time greater than x reaches this level not later than the whole batch in the FB system. This is due to the work conservation principle. Thus,

$$\mathcal{A}_x(t) \setminus \mathcal{N}_x^{\mathrm{FB}}(t) \subset \mathcal{A}_x(t) \setminus \mathcal{N}_x^{\pi^*}(t),$$

which implies that

$$\sum_{i \in \mathcal{A}_x(t) \setminus \mathcal{N}_x^{\mathrm{FB}}(t)} (S_i - \min\{x, S_i\}) \leq \sum_{i \in \mathcal{A}_x(t) \setminus \mathcal{N}_x^{\pi^*}(t)} (S_i - \min\{x, S_i\}).$$

Thus, from (2.2), $V_x^{\pi^*}(t) \leq V_x^{\mathrm{FB}}(t)$.

Now we show that FCFS_x disciplines are strictly better than FB with respect to the level-x workload.

https://doi.org/10.1239/jap/1152413739 Published online by Cambridge University Press

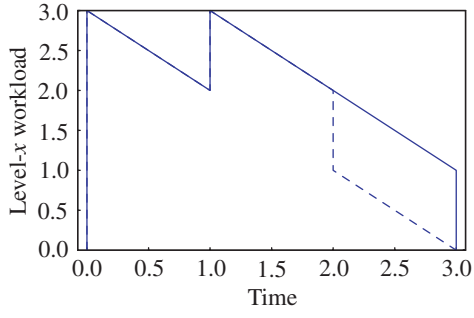

FIGURE 1: The level-x workload, $V_x^{\pi}(t)$, as a function of time for FB (*solid*) and π^* (*dashed*), with $x = 2$.

Proposition 3.2. *Let $x > 0$ and $\pi^* \in \mathrm{FCFS}_x$. There exist a sample path and a $t \geq 0$ such that $V_x^{\pi^*}(t) < V_x^{\mathrm{FB}}(t)$.*

Proof. Assume that $A_1 = 0$, $S_1 = 3x/2$, $A_2 = x/2$, $S_2 = x/2$, and $A_3 = 3x$. FB serves job one in the intervals $[0, x/2)$ and $[x, 2x)$ and job two in the interval $[x/2, x)$, whereas π^* serves job one in the intervals $[0, x)$ and $[3x/2, 2x)$ and job two in the interval $[x, 3x/2)$. As a result,

$$V_x^{\pi^*}(5x/4) = x/4 < 3x/4 = V_x^{\mathrm{FB}}(5x/4).$$

In fact, $V_x^{\pi^*}(t) < V_x^{\mathrm{FB}}(t)$ for all $t \in (x, 3x/2)$, as can be seen from Figure 1, where we have chosen $x = 2$.

Remark 3.1. We note that Proposition 3.2 contradicts [5, Lemma 3.5], which states that FB is pathwise optimal with respect to the level-x workload. This is essentially due to the fact that Righter *et al*. [5] confused the level-x workload, $V_x^{\pi}(t)$, with the variable

$$U_x^{\pi}(t) = \sum_{i \in \mathcal{N}_x^{\pi}(t)} (\min\{S_i, x\} - X_i^{\pi}(t)),$$

which we call *truncated level-x workload*, thus causing the remaining *truncated* service times to be summed, instead of the ordinary remaining service times as in (2.1). It is true that FB is pathwise optimal with respect to the truncated level-x workload [1, Proposition 5], but not, as Proposition 3.2 reveals, with respect to the level-x workload. The confusion in [5] can be explained as follows. As given in [1, Equation (16)], we have

$$U_x^{\pi}(t) = \sum_{i \in \mathcal{A}(t)} S_i - \int_0^t R_x^{\pi}(u)\,\mathrm{d}u.$$

This corresponds to [5, Equation (3.1)]. Thus, according to (2.2), Righter *et al*. [5] omitted the downward steps in the sample paths of the level-x workload process, $V_x^{\pi}(t)$.

Remark 3.2. Strictly speaking, if the level-x workload were defined as in [5], i.e. as the sum of the remaining service times of those jobs in the system whose attained service is less than *or equal to* the given truncation threshold x, then the disciplines π^* and FB considered in the proof of Proposition 3.2 would have the same level-x workload for all t. However, in that case, a slightly modified discipline, $\pi^* \in \mathrm{FCFS}_{x+\varepsilon}$, would serve as a counter-example for sufficiently small $\varepsilon > 0$.

https://doi.org/10.1239/jap/1152413739 Published online by Cambridge University Press

4. Mean value results for the level-x workload

Righter *et al.* [5] applied their Lemma 3.5 to prove that, for all $\pi \in \Pi$ and $x > 0$,

$$\bar{V}_x^{\mathrm{FB}} \leq \bar{V}_x^{\pi}, \tag{4.1}$$

which would be a condition sufficient (but not necessary) for the optimality of FB with respect to the mean delay, according to Proposition 2.1. However, as shown above, their Lemma 3.5 is not valid.

In this section we show that (4.1) is not valid either. We start, in Section 4.1, with some preliminary results related to a modified queue wherein the service times are replaced by their truncated versions. In Section 4.2, we derive some fundamental formulae for the mean level-x workload and give the mean level-x workload formulae for FB, PS, and FCFS$_x$. Finally, in Section 4.3, we show that FB is not optimal with respect to the mean level-x workload, although it still outperforms PS.

4.1. Truncated service times

Let $x \geq 0$ and consider a modified M/G/1 queue in which the original service times, S, are replaced by their truncated versions, $S \wedge x = \min\{S, x\}$. It is easy to see that

$$\mathrm{E}[S \wedge x] = \int_0^x \bar{F}(y)\,\mathrm{d}y, \qquad \mathrm{E}[(S \wedge x)^2] = 2\int_0^x y\bar{F}(y)\,\mathrm{d}y. \tag{4.2}$$

Furthermore, let

$$\rho_x = \lambda\,\mathrm{E}[(S \wedge x)]$$

denote the truncated load. The mean workload for a work-conserving M/G/1 queue with truncated service times is, by the Pollaczek–Khinchin formula (cf. [3, Equation (4.26)]),

$$\bar{W}_x = \frac{\lambda\,\mathrm{E}[(S \wedge x)^2]}{2(1-\rho_x)}. \tag{4.3}$$

By letting $x \to \infty$, we recover the ordinary Pollaczek–Khinchin formula,

$$\bar{W}_\infty = \frac{\lambda\,\mathrm{E}[S^2]}{2(1-\rho)}.$$

Regarding the derivative, it is easy to verify that

$$\frac{\mathrm{d}}{\mathrm{d}x}\bar{W}_x = \lambda\frac{\bar{W}_x + x}{1-\rho_x}\bar{F}(x). \tag{4.4}$$

4.2. Mean level-x workload

Once more consider the original M/G/1 queue with service times S and load

$$\rho = \lambda\,\mathrm{E}[S] = \lambda\int_0^\infty \bar{F}(x)\,\mathrm{d}x.$$

Note that $\rho > \rho_x$ for all $x > 0$, since we have assumed that $\bar{F}(x) > 0$ for all x.

https://doi.org/10.1239/jap/1152413739 Published online by Cambridge University Press

Recall that $\bar{V}_x^\pi$ is the steady-state mean level-x workload. In addition, let $\bar{N}_x^\pi$ denote the steady-state mean number of those jobs in the system whose attained service is less than x. Since π is work conserving and nonanticipating, we have (cf. [5, Equation (3.13)])

$$\bar{V}_x^\pi = \int_{0^-}^{x^-} \mathrm{E}[S - y \mid S > y] \, \mathrm{d}\bar{N}_y^\pi. \tag{4.5}$$

As was noted earlier, in Remark 3.1, the level-x workload, V_x^π, differs from the truncated level-x workload, U_x^π. For a work-conserving and nonanticipating discipline $\pi \in \Pi$, the steady-state mean truncated level-x workload, $\bar{U}_x^\pi$, reads as follows:

$$\bar{U}_x^\pi = \int_{0^-}^{x^-} \mathrm{E}[(S \wedge x) - y \mid S > y] \, \mathrm{d}\bar{N}_y^\pi.$$

Note further that the limit $\bar{V}_\infty^\pi = \bar{U}_\infty^\pi = \bar{W}_\infty$ is the same for all $\pi \in \Pi$.

Let $\bar{T}^\pi(x)$ denote the conditional mean delay of a job with service time x. We note that $\bar{T}^\pi(x)$ is increasing and continuous from the left, i.e. $\bar{T}^\pi(x^-) = \bar{T}^\pi(x) \le \bar{T}^\pi(x^+)$. It is also known [3, Equation (4.11)] that

$$\mathrm{d}\bar{N}_y^\pi = \lambda \bar{F}(y) \, \mathrm{d}\bar{T}^\pi(y).$$

Thus, by (4.5) and (1.2),

$$\begin{aligned}
\bar{V}_x^\pi &\overset{(4.5)}{=} \lambda \int_{0^-}^{x^-} \mathrm{E}[S - y \mid S > y] \bar{F}(y) \, \mathrm{d}\bar{T}^\pi(y) \\
&\overset{(1.2)}{=} \lambda \int_{0^-}^{x^-} \int_y^\infty \bar{F}(t) \, \mathrm{d}t \, \mathrm{d}\bar{T}^\pi(y) \\
&= \lambda \int_0^x \bar{T}^\pi(y) \bar{F}(y) \, \mathrm{d}y + \lambda \bar{T}^\pi(x) \int_x^\infty \bar{F}(y) \, \mathrm{d}y \\
&= \lambda \int_0^x \bar{T}^\pi(y) \bar{F}(y) \, \mathrm{d}y + \bar{T}^\pi(x)(\rho - \rho_x).
\end{aligned} \tag{4.6}$$

Note that $\bar{V}_x^\pi$ is increasing and continuous from the left, i.e. $\bar{V}_{x^-}^\pi = \bar{V}_x^\pi \le \bar{V}_{x^+}^\pi$.

Since, by [3, Equation (4.60)],

$$\bar{U}_x^\pi = \lambda \int_0^x \bar{T}^\pi(y) \bar{F}(y) \, \mathrm{d}y, \tag{4.7}$$

the mean level-x workload can also be given as follows:

$$\bar{V}_x^\pi = \bar{U}_x^\pi + \bar{T}^\pi(x)(\rho - \rho_x).$$

Furthermore, by (1.2),

$$\rho - \rho_x = \lambda \bar{F}(x) \, \mathrm{E}[S - x \mid S > x],$$

implying that

$$\bar{V}_x^\pi - \bar{U}_x^\pi = \lambda \bar{F}(x) \bar{T}^\pi(x) \, \mathrm{E}[S - x \mid S > x],$$

the factors of which have the following intuitive interpretations: by Little's result, $\lambda \bar{F}(x) \bar{T}^\pi(x)$ refers to the mean number of customers with service time longer than x and attained service at most x, and $\mathrm{E}[S - x \mid S > x]$ is the expectation of the truncated part of their service times.

https://doi.org/10.1239/jap/1152413739 Published online by Cambridge University Press

For FB, by [3, Equation (4.27)] we have

$$\bar{T}^{\mathrm{FB}}(x) = \frac{\bar{W}_x + x}{1 - \rho_x}.$$

Thus, by (4.4),

$$\lambda \bar{T}^{\mathrm{FB}}(x)\bar{F}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\bar{W}_x,$$

implying, by (4.6), that

$$\bar{V}_x^{\mathrm{FB}} = \bar{W}_x + \frac{\bar{W}_x + x}{1 - \rho_x}(\rho - \rho_x) = \bar{W}_x\left(1 + \frac{\rho - \rho_x}{1 - \rho_x}\right) + x\frac{\rho - \rho_x}{1 - \rho_x}. \tag{4.8}$$

Now consider PS. By [3, Equation (4.17)],

$$\bar{T}^{\mathrm{PS}}(x) = \frac{x}{1 - \rho}.$$

Thus, by (4.6), (4.2), and (4.3),

$$\bar{V}_x^{\mathrm{PS}} = \bar{W}_x\frac{1 - \rho_x}{1 - \rho} + x\frac{\rho - \rho_x}{1 - \rho}. \tag{4.9}$$

Finally, consider any $\pi \in \mathrm{FCFS}_x$. By [3, Equation (4.35)], for all $y \le x$ we have

$$\bar{T}^{\pi}(y) = \bar{W}_x + y.$$

Thus, by (4.6),

$$\bar{V}_y^{\pi} = \bar{W}_x\rho + \bar{W}_y(1 - \rho_y) + y(\rho - \rho_y)$$

for all $y \le x$. In particular,

$$\bar{V}_x^{\pi} = \bar{W}_x(1 + \rho - \rho_x) + x(\rho - \rho_x). \tag{4.10}$$

In the sequel we will also need the following equality, which is valid for any $\pi \in \mathrm{FCFS}_x$:

$$\bar{U}_x^{\pi} = \bar{U}_x^{\mathrm{FB}} = \bar{W}_x. \tag{4.11}$$

This can easily be verified using (4.7) and the conditional mean delay formulae given above. In fact, this property follows from the pathwise local optimality of these policies with respect to the truncated level-x workload; see [1, Section 3.1].

4.3. Comparison with FB

Now we are ready to show that FB is not optimal with respect to the mean level-x workload. Proposition 3.1 implies that any $\pi^* \in \mathrm{FCFS}_x$ is at least as good as FB. Below we show that such policies are strictly better than FB with respect to the mean level-x workload.

Proposition 4.1. *Consider any service time distribution and let $x > 0$ and $\pi^* \in \mathrm{FCFS}_x$. Then $\bar{V}_x^{\pi^*} < \bar{V}_x^{\mathrm{FB}}$.*

Proof. By (4.8) and (4.10),

$$\bar{V}_x^{\mathrm{FB}} - \bar{V}_x^{\pi^*} = \frac{(\bar{W}_x + x)(\rho - \rho_x)\rho_x}{1 - \rho_x} > 0,$$

since $0 < \rho_x < \rho < 1$.

https://doi.org/10.1239/jap/1152413739 Published online by Cambridge University Press

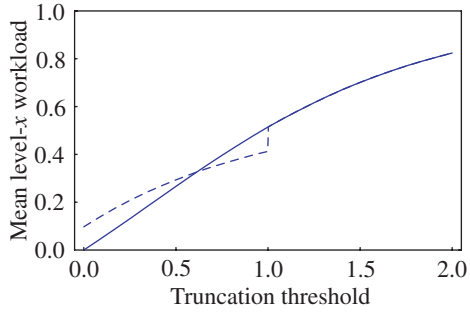

FIGURE 2: The mean level-x workload, $\bar{V}_x^{\pi}$, as a function of the truncation threshold, x, for FB (*solid*) and FCFS + FB(1) (*dashed*).

As a numerical example, consider the exponential service time distribution with rate parameter equal to 1, $\bar{F}(x) = e^{-x}$, and $E[S] = 1$. Let $\lambda = \frac{1}{2}$, implying that we have a stable system with load $\rho = \lambda E[S] = \frac{1}{2} < 1$. Let $\text{FCFS}+\text{FB}(x) \in \bar{\text{FCFS}}_x$ refer to the discipline that applies FB to all the jobs with attained service time greater than or equal to x. Then

$$\bar{V}_1^{\text{FB}} = 0.514, \qquad \bar{V}_1^{\text{FCFS}+\text{FB}(1)} = 0.413.$$

This result is illustrated in Figure 2, where we have depicted the mean level-x workload $\bar{V}_x^{\pi}$ as a function of the truncation threshold x for disciplines FB and FCFS + FB(1).

Finally we show that, while not being optimal, FB still outperforms PS with respect to the mean level-x workload.

Proposition 4.2. *Consider any service time distribution and let $x > 0$. Then $\bar{V}_x^{\text{FB}} < \bar{V}_x^{\text{PS}}$.*

Proof. By (4.8) and (4.9),

$$\bar{V}_x^{\text{PS}} - \bar{V}_x^{\text{FB}} = \frac{(\bar{W}_x + x)(\rho - \rho_x)^2}{(1-\rho)(1-\rho_x)} > 0,$$

since $\rho_x < \rho < 1$.

As an immediate consequence of Propositions 4.1 and 4.2, for any service time distribution, for $x > 0$, and for $\pi^* \in \text{FCFS}_x$, we have $\bar{V}_x^{\pi^*} < \bar{V}_x^{\text{PS}}$.

5. Mean delay results

In this section we justify our principal claim that FB does not necessarily minimize the mean delay within class IMRL, in contradiction with [5, Theorem 3.14]. However, we show that FB still outperforms PS with respect to the mean delay within class IMRL, which can be considered a generalization of [6, Theorem 1].

Theorem 5.1. *There exist a service time distribution belonging to class IMRL and a $\pi \in \Pi$ such that*

$$\bar{T}^{\pi} < \bar{T}^{\text{FB}}.$$

Proof: Step 1. First we must find a distribution that belongs to IMRL but not to DHR. A candidate (for any $c > 1$) is

$$\bar{F}(x) = \begin{cases} c^{-x}, & 0 \le x \le c, \\ x^{-c}, & x > c. \end{cases} \tag{5.1}$$

https://doi.org/10.1239/jap/1152413739 Published online by Cambridge University Press

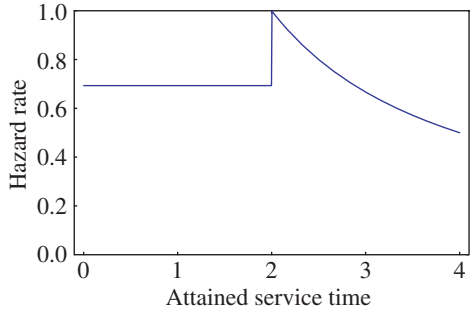

FIGURE 3: Hazard rate for the service time distribution defined in (5.1), with $c = 2$.

We thus first have an exponential section and then a Pareto-type tail. The corresponding density function is

$$f(x) = \begin{cases} c^{-x} \ln c, & 0 \le x \le c, \\ c x^{-c-1}, & x > c, \end{cases} \tag{5.2}$$

and the hazard rate is

$$h(x) = \begin{cases} \ln c, & 0 \le x \le c, \\ c x^{-1}, & x > c. \end{cases}$$

It is easy to see that $F(x)$ belongs to DHR if and only if $h(c-) \ge h(c+)$, which is equivalent to the requirement that $c \ge \mathrm{e}$. In Figure 3 we have depicted the function $h(x)$ for $c = 2$.

The mean residual lifetime function is as follows:

$$\frac{1}{H(x)} = \begin{cases} \dfrac{1}{\ln c} + \left(\dfrac{c}{c-1} - \dfrac{1}{\ln c} \right) c^{x-c}, & 0 \le x \le c, \\ \dfrac{x}{c-1}, & x > c. \end{cases} \tag{5.3}$$

Since, for all $c > 1$,

$$\frac{c}{c-1} - \frac{1}{\ln c} > 0,$$

we deduce from (5.3) that $F(x)$ belongs to IMRL for any $c > 1$. In Figure 4 we have depicted the function $1/H(x)$ for $c = 2$. To summarize, $F(x)$ belongs to IMRL but not to DHR if and only if

$$1 < c < \mathrm{e} \approx 2.718\,28.$$

Step 2. Recall from the previous section that FCFS + FB(x) gives priority to jobs with attained service time less than x, and that FCFS is used for all such jobs, while FB is applied to the remaining jobs with attained service time greater than or equal to x. It is known that FCFS is optimal within class IHR (increasing hazard rate), while FB is optimal within class DHR. Thus, for the distribution defined in (5.1) with $1 < c < \mathrm{e}$, the hazard rate of which is first increasing and then decreasing, it is reasonable to consider disciplines of type FCFS + FB(x).

Let us compare FB and FCFS + FB($c + \varepsilon$) with $\varepsilon \ge 0$. The conditional mean delays read as follows:

$$\bar{T}^{\mathrm{FB}}(x) = \frac{\bar{W}_x + x}{1 - \rho_x}, \qquad x \ge 0, \tag{5.4}$$

https://doi.org/10.1239/jap/1152413739 Published online by Cambridge University Press

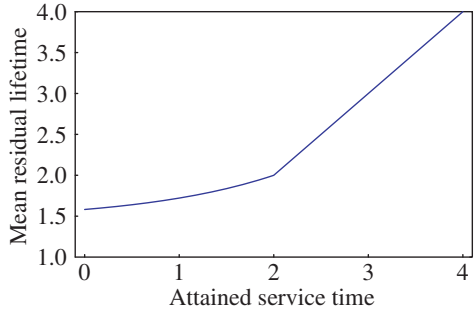

FIGURE 4: Mean residual lifetime for the service time distribution defined in (5.1), with $c = 2$.

and

$$\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c+\varepsilon)}(x) = \begin{cases} \bar{W}_{c+\varepsilon} + x, & 0 \le x \le c+\varepsilon, \\ \bar{T}^{\mathrm{FB}}(x), & x > c+\varepsilon. \end{cases} \tag{5.5}$$

We first consider the FCFS + FB(c) discipline with $\varepsilon = 0$ and show that

$$\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c)} = \bar{T}^{\mathrm{FB}}. \tag{5.6}$$

By (5.4) and (5.5),

$$\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c)} - \bar{T}^{\mathrm{FB}} = \int_0^c (\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c)}(x) - \bar{T}^{\mathrm{FB}}(x)) f(x)\,\mathrm{d}x.$$

Since, by (5.2), $f(x) = \bar{F}(x) \ln c$ for any $x \le c$, we obtain

$$\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c)} - \bar{T}^{\mathrm{FB}} = \ln c \int_0^c (\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c)}(x) - \bar{T}^{\mathrm{FB}}(x)) \bar{F}(x)\,\mathrm{d}x,$$

implying, by (4.7) and (4.11), that

$$\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c)} - \bar{T}^{\mathrm{FB}} = \frac{\ln c}{\lambda} (\bar{U}_c^{\mathrm{FCFS}+\mathrm{FB}(c)} - \bar{U}_c^{\mathrm{FB}}) = 0.$$

Thus, if we can now prove that

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c+\varepsilon)}\Big|_{\varepsilon=0^+} < 0,$$

then it follows from (5.6) that there exists a $\delta > 0$ such that, for any ε, $0 < \varepsilon < \delta$,

$$\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c+\varepsilon)} < \bar{T}^{\mathrm{FB}},$$

revealing the nonoptimality of FB for this distribution.

By (5.4) and (5.5),

$$\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c+\varepsilon)} = \int_0^{c+\varepsilon} (\bar{W}_{c+\varepsilon} + x) f(x)\,\mathrm{d}x + \int_{c+\varepsilon}^{\infty} \frac{\bar{W}_x + x}{1-\rho_x} f(x)\,\mathrm{d}x.$$

https://doi.org/10.1239/jap/1152413739 Published online by Cambridge University Press

Thus, by (4.4), we obtain the desired result as follows:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c+\varepsilon)} &= \left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bar{W}_{c+\varepsilon}\right)F(c+\varepsilon) + (\bar{W}_{c+\varepsilon}+c+\varepsilon)f(c+\varepsilon) \\ &\quad - \frac{\bar{W}_{c+\varepsilon}+c+\varepsilon}{1-\rho_{c+\varepsilon}}f(c+\varepsilon) \\ &\overset{(4.4)}{=} \lambda\frac{\bar{W}_{c+\varepsilon}+c+\varepsilon}{1-\rho_{c+\varepsilon}}\bar{F}(c+\varepsilon)F(c+\varepsilon) - \frac{\bar{W}_{c+\varepsilon}+c+\varepsilon}{1-\rho_{c+\varepsilon}}f(c+\varepsilon)\rho_{c+\varepsilon} \\ &= \lambda\frac{\bar{W}_{c+\varepsilon}+c+\varepsilon}{1-\rho_{c+\varepsilon}}\left((c+\varepsilon)^{-c}(1-(c+\varepsilon)^{-c}) - c(c+\varepsilon)^{-c-1}\frac{\rho_{c+\varepsilon}}{\lambda}\right) \\ &\to \lambda\frac{\bar{W}_c+c}{1-\rho_c}\left(c^{-c}(1-c^{-c}) - c^{-c}\frac{\rho_c}{\lambda}\right) \quad \text{as } \varepsilon\to 0^+.\end{aligned}$$

Since

$$\frac{\rho_c}{\lambda} = \mathrm{E}[S\wedge c] \overset{(4.2)}{=} \int_0^c \bar{F}(x)\,\mathrm{d}x \overset{(5.1)}{=} \int_0^c c^{-x}\,\mathrm{d}x = \frac{1}{\ln c}(1-c^{-c}),$$

we finally obtain

$$\left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c+\varepsilon)}\right|_{\varepsilon=0^+} = \lambda\frac{\bar{W}_c+c}{1-\rho_c}c^{-c}(1-c^{-c})\left(1-\frac{1}{\ln c}\right) < 0,$$

where the inequality follows from the fact that $1/\ln c > 1$ for all c, $1 < c < \mathrm{e}$.

As numerical examples, we have computed the following results:

- $c = 2.0, \lambda = 0.5, \rho = 0.791, \varepsilon = 1.0$: $\bar{T}^{\mathrm{FB}} = 5.901$, $\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c+\varepsilon)} = 5.811$,
- $c = 2.1, \lambda = 0.5, \rho = 0.733, \varepsilon = 0.7$: $\bar{T}^{\mathrm{FB}} = 4.640$, $\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c+\varepsilon)} = 4.584$,
- $c = 2.5, \lambda = 0.5, \rho = 0.575, \varepsilon = 0.2$: $\bar{T}^{\mathrm{FB}} = 2.561$, $\bar{T}^{\mathrm{FCFS}+\mathrm{FB}(c+\varepsilon)} = 2.558$.

Thus, in all these cases FCFS + FB($c+\varepsilon$) is found to be better than FB with respect to the mean delay.

Theorem 5.2. *Assume that the service time distribution belongs to class IMRL. Then* $\bar{T}^{\mathrm{FB}} \le \bar{T}^{\mathrm{PS}}$.

Proof. This follows immediately from Propositions 2.1 and 4.2.

References

[1] Aalto, S., Ayesta, U. and Nyberg-Oksanen, E. (2004). Two-level processor-sharing scheduling disciplines: mean delay analysis. In *ACM SIGMETRICS Performance Evaluation Review* (Proc. SIGMETRICS 2004/PERFORMANCE 2004), Association for Computing Machinery, New York, pp. 97–105.

[2] Coffman, E. G., Jr. and Denning, P. J. (1973). *Operating Systems Theory*. Prentice-Hall, Englewood Cliffs, NJ.

[3] Kleinrock, L. (1976). *Queueing Systems*, Vol. II, *Computer Applications*. John Wiley, New York.

[4] Righter, R. and Shanthikumar, J. G. (1989). Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Prob. Eng. Inf. Sci*. **3,** 323–333.

[5] Righter, R., Shanthikumar, J. G. and Yamazaki, G. (1990). On extremal service disciplines in single-stage queueing systems. *J. Appl. Prob*. **27,** 409–416.

[6] Wierman, A., Bansal, N. and Harchol-Balter, M. (2004). A note on comparing response times in the M/GI/1/FB and M/GI/1/PS queues. *Operat. Res. Lett*. **32,** 73–76.

[7] Yashkov, S. F. (1987). Processor-sharing queues: some progress in analysis. *Queueing Systems* **2,** 1–17.

https://doi.org/10.1239/jap/1152413739 Published online by Cambridge University Press