Probability in the Engineering and Informational Sciences, **19**, 2005, 351–361. Printed in the U.S.A.

SOJOURN TIMES IN THE *M*/*G*/1 FB QUEUE WITH LIGHT-TAILED SERVICE TIMES

M. MANDJES
CWI
Amsterdam, The Netherlands
and
KdV Institute for Mathematics, University of Amsterdam,
Amsterdam, The Netherlands
E-mail: michel.mandjes@cwi.nl

M. NUYENS
Department of Mathematics
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
E-mail: mnuyens@few.vu.nl

The asymptotic decay rate of the sojourn time of a customer in the stationary *M*/*G*/1 queue under the foreground–background (FB) service discipline is studied. The FB discipline gives service to those customers that have received the least service so far. We prove that for light-tailed service times, the decay rate of the sojourn time is equal to the decay rate of the busy period. It is shown that FB minimizes the decay rate in the class of work-conserving disciplines.

1. INTRODUCTION

The sojourn time of a customer (i.e., the time between his arrival and departure) is an often used performance measure for queues. In this article, we compute the asymptotic decay rate of the tail of the sojourn-time distribution of the stationary *M*/*G*/1 queue with the foreground–background (FB) discipline. This decay rate is then used to compare the performance of FB with other service disciplines like processor-sharing (PS) and first-in first-out (FIFO).

 0269-9648/05 $16.00

The FB discipline gives service to those customers who have received the least amount of service so far. If there are n such customers, each of them is served at rate $1/n$. Thus, when the *age* of a customer is the amount of service a customer has received, the FB discipline gives priority to the *youngest* customers. In the literature, this discipline has been called LAS or LAST (least-attained service time first) as well.

Let V denote the sojourn time of a customer in the stationary $M/G/1$ FB queue. Núñez Queija [6] showed that for service-time distributions with regularly varying tails of index $\eta \in (1,2)$, the distribution of V satisfies

$$P(V > x) \sim P(B > (1-\rho)x) \qquad x \to \infty, \tag{1}$$

where ρ is the load of the system, B is the generic service time, and $\sim$ means that the quotient converges to 1. Using Núñez Queija's method, Nuyens [7] obtained (1) under weaker assumptions. In the case of regularly varying service times, the tail of V under the disciplines last-in first-out (LIFO), PS, and shortest remaining processor time (SRPT) satisfies relations similar to (1). For nonpreemptive disciplines however, like FIFO, it is heavier than the tail of B; see Borst, Boxma, Núñez Queija, and Zwart [1].

Additional support for the effective performance of FB under heavy tails is given by Righter [8], Righter and Shanthikumar [9], and Righter, Shanthikumar, and Yamazaki [10]. They show that for certain classes of service times (including, e.g., the Pareto distribution), the FB discipline minimizes the queue length, measured in number of customers, in the class of all disciplines that do not know the exact value of the service times.

For light-tailed service times, the FB discipline does not perform as well, although for gamma densities $\lambda^\alpha x^{\alpha-1}\exp(-\lambda x)/\Gamma(\alpha)$ with $0 < \alpha \leq 1$, FB still minimizes the queue length, and for exponential service times, the queue length is independent of the service discipline. However, for many other light-tailed service times (e.g., those with a decreasing failure rate), the queue shows opposite behavior and the queue length is maximized by FB (see Righter and colleagues [8–10]). This undesirable behavior of the FB discipline is very pronounced for deterministic service times. In this extreme case in the FB queue, all customers stay until the end of the busy period, and the sojourn time under the FB discipline is *maximal* in the class of all work-conserving disciplines.

In this article, we consider the (asymptotic) decay rate of the sojourn time, where the *(asymptotic) decay rate* $\mathrm{dr}(X)$ of a random variable X is defined as

$$\mathrm{dr}(X) = \left| \lim_{x\to\infty} x^{-1} \log P(X > x) \right|,$$

given that the limit exists. Hence, a larger decay rate means a smaller probability that the random variable takes on very large values. In this sense, sojourn times are better when they have larger decay rates.

Assume that the service times for an $M/G/1$ queue have a finite exponential moment or, equivalently, the Laplace transform is analytic in a neighborhood of

https://doi.org/10.1017/S0269964805050205 Published online by Cambridge University Press

zero. In other words, the tail of the service-time distribution decreases exponentially. Our main result is that for the $M/G/1$ FB queue under the light-tailed assumption, large sojourn times are relatively likely.

THEOREM 1: *Let V be the sojourn time of a customer in the stationary $M/G/1$ FB queue and let L be the length of a busy period. If the service-time distribution has a finite exponential moment, then the decay rate of V exists and satisfies*

$$\mathrm{dr}(V) = \mathrm{dr}(L). \tag{2}$$

It is shown later that the decay rate of the sojourn time in an $M/G/1$ queue with any work-conserving discipline is bounded from below by the decay rate of the residual life of a busy period. For service times with an exponential moment, the decay rates for normal and residual busy periods are equal. Hence, (2) is the lowest possible decay rate for the sojourn time under a work-conserving discipline. Using the decay rate of V as a criterion to measure the performance of a service discipline then leads to the following conclusion: For service times with an exponential moment, the FB discipline is the worst discipline in the class of work-conserving disciplines.

The article is organized as follows. In Section 2, we present the notation, some preliminaries, and prove the lower bound for the decay rate of the sojourn time under any work-conserving discipline. In Section 3, Theorem 1 is proved. Section 4 discusses the result and the decay rate of the sojourn time in queues operating under several other service disciplines.

2. PRELIMINARIES

Throughout this article, we assume that the generic service time B with distribution function F in the $M/G/1$ queue satisfies the following assumption.

ASSUMPTION 1: *The generic service time B has an exponential moment (i.e., $E\exp(\gamma B) < \infty$ for some $\gamma > 0$).*

In addition, let the stability condition $\rho = \lambda EB < 1$ hold, where λ is the rate of the Poisson arrival process. The proofs in this article rely on some properties of the busy-period length L and related random variables, which we derive in this section.

Under Assumption 1, Cox and Smith [3] have shown that $P(L > x) \sim bx^{-3/2}e^{-cx}$ for certain constants $b, c > 0$. In particular, L has decay rate c. In fact, by expression (46) in Cox and Smith [3, p. 154], $c = \lambda - \zeta - \lambda g(\zeta)$, where g is the Laplace transform of the service-time distribution and $\zeta < 0$ is such that $g'(\zeta) = -\lambda^{-1}$. Hence, ζ is the root of the derivative of the function $m(x) = \lambda - x - \lambda g(x)$. Since $m(x)$ attains its maximum at the point ζ, we can write c in terms of the Legendre transform of B:

$$c = \mathrm{dr}(L) = \sup_{\theta}\{\theta - \lambda(Ee^{\theta B} - 1)\}. \tag{3}$$

Remark: This expression shows up as well in the following context. Consider a Poisson stream, with intensity λ, of independent and identically distributed (i.i.d.)

https://doi.org/10.1017/S0269964805050205 Published online by Cambridge University Press

jobs, where every job is distributed according to the random variable B. Let $A(x)$ denote the amount of work generated in an arbitrary time window of length x. It is an easy corollary of Cramér's theorem that

$$\lim_{x\to\infty} \frac{1}{x} \log P(A(x) > x) = -\sup_{\theta}\{\theta - \log Ee^{\theta A(1)}\}. \tag{4}$$

Noting that

$$Ee^{\theta A(1)} = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} (Ee^{\theta B})^k = \exp(\lambda(Ee^{\theta B} - 1)),$$

we observe that $P(L > x)$ and $P(A(x) > x)$ have the same decay rate. This is somewhat surprising, as $\{A(x) > x\}$ obviously depends just on $A(x)$ (i.e., the amount of traffic in a window of length x), whereas $\{L > x\}$ depends on $A(y)$ for *all* $y \in [0, x]$, due to

$$\{L > x\} \stackrel{\mathrm{d}}{=} \{B_1 + A(y) > y,\ \forall y \in [0, x]\}.$$

Here, B_1 is the first service time in the busy period L.

In renewal theory, the notion of *residual life*, also known as excess or forward-recurrence time, is standard. Let $\tilde{L}$ be the residual life of a busy period. Then $P(\tilde{L} > x) = (EL)^{-1}\int_x^\infty P(L > y)\, dy$; see, for instance, Cox [2]. Using standard calculus, we find

$$\mathrm{dr}(\tilde{L}) = \left| \lim_{x\to\infty} \frac{1}{x} \log \int_x^\infty y^{-3/2} e^{-cy}\, dy \right| = c = \mathrm{dr}(L). \tag{5}$$

Hence, $\tilde{L}$ has the same decay rate as L.

Another ingredient used in the proofs below is the $M/G/1$ queue with truncated generic service time $B \wedge \tau$, $\tau > 0$. Call this the *τ-queue* and let $L(\tau)$ denote the length of a busy period (a *τ-busy period*) in this queue. Let $\tilde{L}(\tau)$ be its residual life and define $L^*(\tau)$ to be the length of a τ-busy period in which the *first* service time B_1 is at least τ; that is,

$$P(L^*(\tau) > x) = P(L(\tau) \mid B_1 \geq \tau).$$

We now show that the random variables $L(\tau)$, $\tilde{L}(\tau)$, and $L^*(\tau)$ have the same decay rate.

Lemma 2: *Let $\tau > 0$ be such that $P(B \geq \tau) > 0$. Then*

$$\mathrm{dr}(L(\tau)) = \mathrm{dr}(L^*(\tau)) = \mathrm{dr}(\tilde{L}(\tau)) > 0.$$

Proof: We show that L and L^* have the same decay rate. The proof is then completed by using (5). Assume that $\tau > 0$ is such that $P(B \geq \tau) > 0$. If $B_1 \geq \tau$, then the first service time is maximal in the τ-queue, as all service times are bounded by τ. Hence,

https://doi.org/10.1017/S0269964805050205 Published online by Cambridge University Press

$$P(L(\tau) > x) \le P(L(\tau) > x \mid B_1 \ge \tau) = P(L^*(\tau) > x), \qquad x \ge 0.$$

Further,

$$\begin{aligned} P(L(\tau) > x) &\ge P(L(\tau) > x, B_1 \ge \tau) \\ &= P(L(\tau) > x \mid B_1 \ge \tau) P(B_1 \ge \tau) \\ &= P(L^*(\tau) > x) P(B_1 \ge \tau). \end{aligned} \tag{6}$$

From (6), it follows that $P(L^*(\tau) > x)$ and $P(L(\tau) \ge x)$ differ only by a factor independent of x. Hence, $\mathrm{dr}(L) = \mathrm{dr}(\tilde{L})$. ■

In this article, we need the following lemma about the decay rate of the sum of two independent random variables.

LEMMA 3: *Let X and Y be nonnegative, independent random variables such that $\mathrm{dr}(X) = \alpha_1$ and $\mathrm{dr}(Y) = \alpha_2$ for some $\alpha_1, \alpha_2 > 0$. Then $\mathrm{dr}(X + Y) = \min\{\alpha_1, \alpha_2\}$.*

PROOF: Since both X and Y are positive, $-\min\{\alpha_1, \alpha_2\}$ is clearly a lower bound for $\liminf_{x\to\infty} x^{-1} \log P(X + Y > x)$. For the upper bound, let $\varepsilon > 0$ and $n \in \mathbb{N}$ be fixed. Then,

$$P(X + Y > x) \le \sum_{i=0}^{n-1} P\left(X \ge \frac{ix}{n}\right) P\left(Y \ge \frac{(n-i-1)x}{n}\right).$$

For x sufficiently large, for all $i \in \{0, \ldots, n-1\}$,

$$\begin{aligned} P\left(X \ge \frac{ix}{n}\right) P\left(Y \ge \frac{(n-i-1)x}{n}\right) &\le \exp\left(-(\alpha_1 - \varepsilon)\frac{ix}{n} - (\alpha_2 - \varepsilon)x\frac{n-i-1}{n}\right) \\ &\le \exp\left(-(\min\{\alpha_1, \alpha_2\} - \varepsilon)\frac{(n-1)x}{n}\right). \end{aligned}$$

Hence,

$$\limsup_{x\to\infty} \frac{1}{x} \log P(X + Y > x) \le -(\min\{\alpha_1, \alpha_2\} - \varepsilon)\left(1 - \frac{1}{n}\right). \tag{7}$$

Since (7) holds for every $n \in \mathbb{N}$ and $\varepsilon > 0$, we can take the limits $n \to \infty$ and $\varepsilon \downarrow 0$, and the result follows. ■

Let D be the time from the arrival of a customer until the first moment that the system is empty. The following proposition is valid even when Assumption 1 does not hold.

PROPOSITION 4: *Consider a stationary queue with an arbitrary service-time distribution, Poisson arrivals, and a work-conserving discipline. Then $D \stackrel{\mathrm{d}}{=} A\tilde{L} + L$, where $P(A = 1) = \rho = 1 - P(A = 0)$ and $A, \tilde{L}$, and L are independent.*

https://doi.org/10.1017/S0269964805050205 Published online by Cambridge University Press

PROOF: The value of the random variable D does not depend on the service discipline. There are two possibilities. With probability $1-\rho$, the customer finds the system empty. In this case, D is just the length L of the busy period started by the customer. Second, if our customer enters a busy system, then the server can first finish all the work in the system apart from the work of our tagged customer. The moment the remainder of the original busy period, which has length $\tilde{L}$, is finished, our customer starts a subbusy period. This length of this subbusy period, which is independent of $\tilde{L}$, is distributed like L. ■

For the stationary τ-queue with Poisson arrivals and a work-conserving discipline, we have the following corollary.

COROLLARY 5: *In the stationary τ-queue, the random variable D satisfies $D \stackrel{\mathrm{d}}{=} A(\tau)\tilde{L}(\tau)+L(\tau)$, where $P(A(\tau)=1)=\lambda E(B\wedge\tau)$. If the customer has service time τ in the τ-queue, then $D \stackrel{\mathrm{d}}{=} A(\tau)\tilde{L}(\tau)+L^*(\tau)$.*

Since the system is work-conserving, the sojourn time of a customer is not longer than D. Hence, $V \leq_{\mathrm{st}} D$ for every service discipline. Since $A\tilde{L}$ and L satisfy the conditions of Lemma 3, the following corollary holds.

COROLLARY 6: *For every work-conserving service discipline, the sojourn time V of a customer in the stationary queue satisfies*

$$\limsup_{x\to\infty}\frac{1}{x}\log P(V>x)\leq\lim_{x\to\infty}\frac{1}{x}\log P(A\tilde{L}+L>x)=-\mathrm{dr}(L).$$

An immediate consequence of this corollary and Theorem 1, which will be proved in the next section, is the following.

COROLLARY 7: *The FB discipline minimizes the decay rate of the sojourn time in the class of work-conserving disciplines.*

In Section 4, we indicate that for service times with a finite exponential moment, there are disciplines with a strictly larger decay rate than FB (e.g., FIFO).

Interestingly, for service times with certain Gamma distributions, the FB discipline stochastically minimizes the queue length, as was mentioned in Section 1, but the sojourn time has the smallest decay rate. This shows that optimizing one characteristic in a queue could have an ill effect on other characteristics.

The existence of a finite exponential moment in the corollary is crucial: For heavy-tailed service times, the tail of V cannot be bounded by that of L. For example, in the $M/G/1$ FIFO queue with service times satisfying $P(B>x)=x^{-\nu}\mathcal{L}(x)$, where $\mathcal{L}(x)$ is a slowly varying function at ∞ and $\nu>1$, De Meyer and Teugels [4] showed that

$$P(L>x)\sim(1-\rho)^{-\nu-1}x^{-\nu}\mathcal{L}(x).$$

It may be seen that in this case, the tail of $\tilde{B}$, the residual life of the generic service time B, is one degree heavier than that of B. Now, note that for the FIFO discipline,

https://doi.org/10.1017/S0269964805050205 Published online by Cambridge University Press

we have $V_{FIFO} \geq A\tilde{B}$. Hence, the tail of V is at least one degree heavier than that of L; see also Borst et al. [1] for further references. In the light-tailed case, this phenomenon is absent because the tails of L and $\tilde{L}$ have the same decay rate.

3. PROOF OF THE THEOREM

In this section, Theorem 1 is proved. The results in this section rely on the following decomposition of V. Let $V(\tau)$ be the sojourn time in the stationary $M/G/1$ queue of a customer with service time τ. The sojourn time V of an arbitrary customer in the stationary queue satisfies

$$P(V > x) = \int P(V(\tau) > x)\, dF(\tau). \tag{8}$$

Here, F is the service-time distribution. Hence, we may write $P(V > x) = E_B P(V(B) > x)$, where B is a generic service time independent of $V(\tau)$, and E_B denotes the expectation w.r.t. B. Theorem 1 is proved using this representation of V. In Proposition 8, we compute the decay rate of $V(\tau)$.

PROPOSITION 8: *Let* $\tau > 0$ *be such that* $P(B \geq \tau) > 0$. *If the service-time distribution satisfies Assumption 1, then* $\mathrm{dr}(V(\tau)) = \mathrm{dr}(L(\tau))$.

PROOF: By the nature of the FB discipline, the sojourn time $V(\tau)$ of a customer with service time τ who enters a stationary queue is the time until the first epoch that no customers younger than τ are present. This is the time until the end of the τ-busy period that he either finds in the τ-queue, or starts. By Corollary 5, $V(\tau)$ then satisfies

$$V(\tau) \stackrel{\mathrm{d}}{=} A(\tau)\tilde{L}(\tau) + L^*(\tau), \tag{9}$$

where $\tilde{L}(\tau)$ is the residual life of a τ-busy period, $L^*(\tau)$ is a τ-busy period that starts with a customer with service time τ,

$$P(A(\tau) = 1) = 1 - P(A(\tau) = 0) = \lambda E(B \wedge \tau),$$

and $A(\tau)$, $\tilde{L}(\tau)$, and $L^*(\tau)$ are independent. By Lemma 2, the random variables $A(\tau)\tilde{L}(\tau)$ and $L^*(\tau)$ satisfy the condition of Lemma 3. From (9) and Lemma 2, it follows that

$$\mathrm{dr}(V(\tau)) = \mathrm{dr}(A(\tau)\tilde{L}(\tau) + L^*(\tau)) = \mathrm{dr}(L(\tau)). \tag{10}$$

This completes the proof. ■

Having found the lower bound for the decay rate in Corollary 6, the following lemma provides the basis for finding the upper bound. The *endpoint* $x_F \in [0,\infty]$ of the service-time distribution F is defined as $x_F = \inf\{u \geq 0 : F(u) = 1\}$.

https://doi.org/10.1017/S0269964805050205 Published online by Cambridge University Press

LEMMA 9: *Let V be the sojourn time of a customer in the stationary $M/G/1$ FB queue. Suppose the service-time distribution satisfies Assumption 1. If $\tau_0 > 0$ and $P(B \geq \tau_0) > 0$, then*

$$\liminf_{x\to\infty} \frac{1}{x} \log P(V > x) \geq -P(B \geq \tau_0)^{-1} \int_{[\tau_0, x_F]} \mathrm{dr}(L(\tau))\, dF(\tau). \tag{11}$$

Here, F is the distribution function of the generic service time B.

PROOF: We have

$$P(V > x) \geq P(V > x, B \geq \tau_0) = P(V > x \mid B \geq \tau_0) P(B \geq \tau_0). \tag{12}$$

Using the representation (8), we find

$$\log P(V > x \mid B \geq \tau_0) = \log E_B[P(V(B) > x) | B \geq \tau_0]. \tag{13}$$

Since $\log x$ is a concave function, applying Jensen's inequality to the conditional expectation in (13) yields

$$\log E_B[P(V(B) > x) | B \geq \tau_0] \geq E_B[\log P(V(B) > x) | B \geq \tau_0]. \tag{14}$$

From (12)–(14), it follows that $\Theta := \liminf_{x\to\infty}(1/x)\log P(V > x)$ satisfies

$$\Theta \geq \liminf_{x\to\infty} \frac{1}{x} \int_{[\tau_0, x_F]} \log P(V(\tau) > x)\, dF(\tau)/P(B \geq \tau_0). \tag{15}$$

Applying Fatou's lemma to (15) yields

$$\Theta \geq P(B \geq \tau_0)^{-1} \int_{[\tau_0, x_F]} \lim_{x\to\infty} \frac{1}{x} \log P(V(\tau) > x)\, dF(\tau).$$

The result now follows from Proposition 8. ■

The following lemma is used to develop the upper bound for the decay rate of V from Lemma 9. We introduce the notation $c(\tau) = \mathrm{dr}(L(\tau))$, so that $c = \mathrm{dr}(L) = c(x_F)$.

LEMMA 10: *The function $c(\tau)$ is decreasing in τ. Furthermore, $c(\tau) \to c(x_F)$ as $\tau \to x_F$.*

PROOF: For all τ, the function $h_\tau(\theta) = \theta - \lambda(Ee^{\theta(B\wedge\tau)} - 1)$ is concave in θ, since any moment generating function is convex. Furthermore, $\lim_{\theta\to-\infty} h_\tau(\theta) = \lim_{\theta\to\infty} h_\tau(\theta) = -\infty$. By definition of $L(\tau)$ and (3), we can write $c(\tau) = \sup_\theta\{h_\tau(\theta)\}$. Then $c(\tau)$ is decreasing in τ, since $h_\tau(\theta)$ is decreasing in τ. Since $c(\tau) \geq h_\tau(0) = 0$ for all τ and $c(\tau)$ is decreasing, $c(\tau)$ converges for $\tau \to x_F$. Now, note that $h_\tau(\theta)$ is continuous in τ for all $\theta \in [0, \sup\{\eta : Ee^{\eta B} < \infty\})$, even if B has a discrete distribution. Since the supremum of $\theta - \lambda(Ee^{\theta B} - 1)$ is attained in this interval, we have $\lim_{\tau\to x_F} c(\tau) = c(x_F)$. ■

https://doi.org/10.1017/S0269964805050205 Published online by Cambridge University Press

PROPOSITION 11: *Let V be the sojourn time of a customer in the stationary $M/G/1$ FB queue. If the service-time distribution satisfies Assumption 1, then*

$$\liminf_{x\to\infty} \frac{1}{x} \log P(V > x) \geq -\mathrm{dr}(L). \tag{16}$$

PROOF: If $P(B = x_F) > 0$, then choosing $\tau_0 = x_F$ in (11) yields

$$\liminf_{x\to\infty} \frac{1}{x} \log P(V > x) \geq -c(x_F) = -\mathrm{dr}(L),$$

and (16) holds. Assume $P(B = x_F) = 0$ and let $\varepsilon > 0$. By Lemma 10, there exists an $x_\varepsilon < x_F$ such that $c(\tau) \leq c + \varepsilon$ for all $\tau \geq x_\varepsilon$. Choosing $\tau_0 = x_\varepsilon$ in (11) then yields

$$\begin{aligned}\liminf_{x\to\infty} \frac{1}{x} \log P(V > x) &\geq -P(B \geq x_\varepsilon)^{-1} \int_{[x_\varepsilon, x_F]} c(\tau)\, dF(\tau) \\ &\geq -P(B \geq x_\varepsilon)^{-1} \int_{[x_\varepsilon, x_F]} (c + \varepsilon)\, dF(\tau) = -c - \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, the lower bound (16) follows. ■

PROOF OF THEOREM 1: The lower bound for $\mathrm{dr}(V)$ is established in Corollary 6, and the upper bound in Proposition 11. ■

4. DISCUSSION

The decay rate of the sojourn time V in the $M/G/1$ FB queue is the same as for the preemptive LIFO queue. Indeed, the sojourn time of a customer in the stationary $M/G/1$ queue under the preemptive LIFO discipline is just the length of the sub-busy period started by that customer. From Theorem 1, it follows that the decay rates of the sojourn times for LIFO and FB are equal.

The sojourn time of a customer in the stationary queue under FIFO satisfies $V_{\mathrm{FIFO}} = B + W$, where W is the stationary workload. From the Pollaczek–Khinchin formula,

$$Ee^{-sW} = \frac{s(1-\rho)}{s - \lambda + \lambda E \exp(-sB)}, \tag{17}$$

it follows that the decay rate of W is the value of s for which the denominator in (17) vanishes. Hence, $\mathrm{dr}(W)$ is the positive root θ_0 of $h(\theta) = \theta - \lambda(Ee^{\theta B} - 1)$. Furthermore, since $\mathrm{dr}(B) = \inf\{\theta : h(\theta) = -\infty\}$, we have $\theta_0 < \mathrm{dr}(B) \leq \infty$. An analog of Lemma 3 then yields that $c_{\mathrm{FIFO}} := \mathrm{dr}(V_{\mathrm{FIFO}}) = \theta_0$.

Since h is concave, $h(0) = 0$, and $h'(0) = 1 - \lambda EB < 1$, we have by Theorem 1 and (3) that

$$c_{\mathrm{FB}} := \mathrm{dr}(V_{\mathrm{FB}}) = \mathrm{dr}(L) = \sup_\theta h(\theta) < \theta_0 = c_{\mathrm{FIFO}} < \mathrm{dr}(B); \tag{18}$$

https://doi.org/10.1017/S0269964805050205 Published online by Cambridge University Press

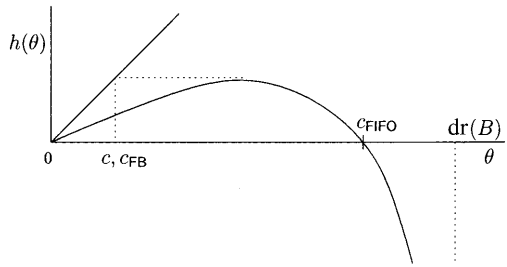

FIGURE 1. The decay rates of the sojourn time under FB and FIFO.

see also Figure 1. Hence, in the FIFO system, the decay rate of the sojourn time is strictly larger than that in the FB queue. As an illustration, consider the $M/M/1$ queue in which the service times have expectation $1/\mu$. For stability, we assume $\lambda < \mu$. Straightforward computations then yield that $c_{\mathrm{FB}} = (\sqrt{\mu} - \sqrt{\lambda})^2$, $c_{\mathrm{FIFO}} = \mu - \lambda$, and $\mathrm{dr}(B) = \mu$. Since $\lambda < \mu$, we conclude that for the $M/M/1$ queue, inequality (18) is satisfied.

Finally, Mandjes and Zwart [5] consider the PS queue with light-tailed service requests. They show that the decay rate of V_{PS} is equal to $\mathrm{dr}(L)$ as well, under the additional requirement that, for any positive constant k,

$$\lim_{x\to\infty} \frac{1}{x} \log P(B > k \log x) = 0.$$

For deterministic requests, clearly this criterion is not met. Indeed, in [5], it is shown that the decay rate of V in the $M/D/1$ queue with the PS discipline is larger than $\mathrm{dr}(L)$.

Acknowledgments

The authors thank A. P. Zwart for kindly commenting on an earlier version of the article and the referee for his clear suggestions and comments, which have seriously improved the presentation of the article. M. Nuyens research was done while he was at the KdV Institute for Mathematics, University of Amsterdam.

References

1. Borst, S., Boxma, O., Núñez Queija, R., & Zwart, A. (2003). The impact of the service discipline on delay asymptotics. *Performance Evaluation* 54: 175–206.
2. Cox, D. (1962). *Renewal theory*. London: Methuen.
3. Cox, D. & Smith, W. (1961). *Queues*. London: Methuen.
4. De Meyer, A. & Teugels, J. (1980). On the asymptotic behaviour of the distributions of the busy period and service time in M/G/1. *Journal of Applied Probability* 17(3): 802–813.
5. Mandjes, M. & Zwart, A. Large deviations for waiting times in processor sharing queues (submitted).
6. Núñez Queija, R. (2000). Processor-sharing models for integrated-services networks. Ph.D. thesis, Eindhoven University, Eindhoven, The Netherlands.

7. Nuyens, M. (2004). The foreground–background queue. Ph.D. thesis, University of Amsterdam, Amsterdam.
8. Righter, R. (1994). Scheduling. In M. Shaked and J. Shanthikumar (eds.), *Stochastic orders and their applications*. San Diego, CA: Academic Press.
9. Righter, R. & Shanthikumar, J. (1989). Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probability in the Engineering and Informational Sciences* 3: 323–333.
10. Righter, R., Shanthikumar, J., & Yamazaki, G. (1990). On extremal service disciplines in single-stage queueing systems. *Journal of Applied Probability* 27: 409–416.