J. Appl. Prob. **42,** 883–891 (2005)
Printed in Israel

THE EFFECT OF SERVICE TIME VARIABILITY ON MAXIMUM QUEUE LENGTHS IN M^X/G/1 QUEUES

GER KOOLE * ** AND
MISJA NUYENS,* *** *Vrije Universiteit Amsterdam*
RHONDA RIGHTER,**** *University of California, Berkeley*

Abstract

We study the impact of service time distributions on the distribution of the maximum queue length during a busy period for the M^X/G/1 queue. The maximum queue length is an important random variable to understand when designing the buffer size for finite-buffer (M/G/1/n) systems. We show the somewhat surprising result that, for three variations of the preemptive last-come–first-served discipline, the maximum queue length during a busy period is smaller when service times are more variable (in the convex sense).

Keywords: Maximum queue length; busy period; service discipline; LCFS; variability; stochastic ordering; buffer overflow

2000 Mathematics Subject Classification: Primary 60K25
Secondary 90B22

1. Introduction

An important design issue for telecommunications systems and other applications is determining the buffer size when buffers are finite. We can better understand the effect of a particular buffer size by understanding the distribution of the maximum queue length during a busy period in an infinite buffer system. We give a characterization of the busy-period maximum queue length, M, for the M^X/G/1 queue for three types of preemptive last-come–first-served (LCFS) discipline:

(i) LCFS-p-resume disciplines, in which preempted services are resumed when service recommences;

(ii) LCFS-p-repeat disciplines with resampling, in which preempted services must be restarted from scratch when service recommences and a new service time is chosen from the service time distribution;

(iii) LCFS-p-repeat disciplines without resampling, in which preempted services must be restarted from scratch when service recommences, but the total service requirement for a given customer is the same each time it restarts its service.

Received 29 March 2004; revision received 3 May 2005.
* Postal address: Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands.
** Email address: koole@few.vu.nl
*** Email address: mnuyens@few.vu.nl
**** Postal address: Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA. Email address: rrighter@ieor.berkeley.edu

These characterizations of M for each of the queueing disciplines allow us to show the effects of service time distribution on M, as stated in (a)–(c) below. For a fixed service discipline, let M and M' be the maximum number of customers during a busy period in two M^X/G/1 queues with respective generic service times S and S', and with the same arrival rate λ and the same batch size distribution. We assume that the distributions of S and S' are such that the queues are stable. In this paper, we show that the following relations hold; the definitions of the various stochastic orders can be found in the next section.

(a) Under the LCFS-p-resume discipline, if $S' \leq_{\mathrm{LT}} S$ then $M' \leq_{\mathrm{st}} M$.

(b) Under the LCFS-p-repeat discipline with resampling, if $\mathrm{E}(\mathrm{e}^{-\lambda S'}) \geq \mathrm{E}(\mathrm{e}^{-\lambda S})$ then $M' \leq_{\mathrm{st}} M$.

(c) Under the LCFS-p-repeat discipline without resampling, if $S' \leq_{\mathrm{icv}} S$ then $M' \leq_{\mathrm{st}} M$.

A consequence of our results is the somewhat surprising conclusion that M will be stochastically smaller when service times are more variable (in the convex sense) under the preemptive LCFS disciplines. Miyazawa (1990) and Miyazawa and Shanthikumar (1991) showed that for the finite-buffer M^X/G/1/n queue under a nonpreemptive discipline, the loss rate, i.e. the probability that a random customer is lost, will be larger when service times are more variable in the convex sense. Our result relates to the loss rate, but the effect goes in the other direction. That is, we find that for preemptive LCFS disciplines, $\mathrm{P}(M > n)$ is smaller when service times are larger in the convex sense, where $\mathrm{P}(M > n)$ can be interpreted as the probability of at least one loss occurring during a busy period in the M^X/G/1/n queue; see also Chang *et al.* (1991).

For other results on the impact of the service time and batch size distributions on various performance measures of queueing systems, see, for example, Hordijk (2001), Makowski (1994), Shanthikumar and Yao (1994), and the references therein. For other applications of the preempt–repeat service discipline, see, for example, Adiri *et al.* (1991), Birge *et al.* (1990), and Cai *et al.* (2004), (2005).

The paper is organized as follows. First, in the next section, we recall some definitions of stochastic ordering. We then study M for each of the preemptive LCFS disciplines, providing some numerical illustrations of our results.

2. Preliminaries

Recall the following stochastic ordering relations for random variables X and Y.

Definition 2.1. X is larger than Y in the stochastic sense, i.e. $X \geq_{\mathrm{st}} Y$, if $\mathrm{E}(\phi(X)) \geq \mathrm{E}(\phi(Y))$ for all increasing functions ϕ for which the expectations exist.

Equivalently, $X \geq_{\mathrm{st}} Y$ if and only if $\mathrm{P}(X > t) \geq \mathrm{P}(Y > t)$ for all t.

Definition 2.2. X is larger than Y in the convex sense, i.e. $X \geq_{\mathrm{cx}} Y$, if $\mathrm{E}(\phi(X)) \geq \mathrm{E}(\phi(Y))$ for all convex functions ϕ for which the expectations exist.

Note that $X \geq_{\mathrm{cx}} Y$ implies $\mathrm{E}(X) = \mathrm{E}(Y)$ and $\mathrm{var}(X) \geq \mathrm{var}(Y)$. In this sense, the convex ordering is an ordering of variability in random variables.

Definition 2.3. X is larger than Y in the increasing-concave sense, i.e. $X \geq_{\mathrm{icv}} Y$, if $\mathrm{E}(\phi(X)) \geq \mathrm{E}(\phi(Y))$ for all increasing concave functions ϕ for which the expectations exist.

Definition 2.4. X is larger than Y in the Laplace transform sense, i.e. $X \geq_{\mathrm{LT}} Y$, if $\mathrm{E}(\mathrm{e}^{-\theta X}) \leq \mathrm{E}(\mathrm{e}^{-\theta Y})$ for all $\theta > 0$ for which the expectations exist.

Note that $X \geq_{\mathrm{cx}} Y$ implies $X \leq_{\mathrm{icv}} Y$, which in turn implies $X \leq_{\mathrm{LT}} Y$.

Finally, for reasons of brevity we use the following notation. When we say that $X = (Y \mid Z = z)$, we mean that $\mathrm{P}(X = x) = \mathrm{P}(Y = x \mid Z = z)$ for all x.

3. Preemptive LCFS disciplines

3.1. LCFS preempt–resume discipline

We first consider the M^X/G/1 queue with the LCFS preempt–resume (LCFS-p-resume) discipline. That is, the customer who has been in the system the least amount of time is always served, and newly arriving customers preempt earlier arrivals already in service. Within a batch, customers are arbitrarily labeled, so that we may think of them as arriving sequentially, though immediately after each other. Thus, one customer in a newly arriving batch will be considered the most recent arrival and will immediately enter service, and the rest of the batch cannot be served until that customer, as well as all customers arriving in later batches that preempt that customer, have been served.

Customers who resume service after being preempted restart their service where they left off. Hence, a random service with service time S that is preempted when t units of service have already been received has remaining service time $(S - t \mid S > t)$. We also assume that service is nonidling. Let T be a generic interarrival time, where T has an exponential distribution with rate λ, and let X be a generic batch size with arbitrary distribution and mean μ. We assume that the queue is stable, i.e. $\lambda\mu\,\mathrm{E}(S) < 1$.

Let customer 0 be the last customer in the first batch in the busy period, i.e. the first customer to enter service, and let S_0 be the service time of customer 0. Let $N \equiv N(S_0)$ be the number of Poisson batch arrival times that occur during the service of customer 0, and let $N(s) = (N(S_0) \mid S_0 = s)$. Note that the service will be interrupted if $N(S_0) > 0$. Let X_0 be the number of customers in the first batch of the busy period and define $M(k, n) = (M \mid X_0 = k,\ N = n)$ and $M(k) = (M \mid X_0 = k)$, meaning that $M(X_0, N) = M = M(X_0)$. Let M_i, $i = 1, 2, \ldots$, be independent, identically distributed copies of M, and define $\max_{i=1,\ldots,n} M_i$ to be 0 if $n = 0$. For the LCFS-p-resume discipline, we then have the following characterization of $M(k, n)$. The symbol '$\stackrel{\mathrm{D}}{=}$' stands for equality in distribution.

Lemma 3.1. *The maximum queue length $M(k, n)$ for the M^X/G/1 queue under the LCFS-p-resume discipline satisfies*

$$M(k, n) \stackrel{\mathrm{D}}{=} \max\left\{ \max_{i=1,\ldots,n} M_i + k,\, M(k-1) \right\}, \qquad k \geq 1,\ n \geq 0,$$

where $M(0) = 0$ and M_i, $i = 1, 2, \ldots$, and $M(k-1)$ are independent.

Proof. We can think of constructing the busy period, conditional on having $X_0 = k$, $S_0 = s$, and $N(s) = n$, as follows. Denote the arrival epochs, on a clock that only ticks when customer 0 is being served, by $0 < t_1 < \cdots < t_n < s$. A batch of customers arrives at time t_1 and starts a new independent busy period (and stops our clock temporarily), except that there are k other customers in the queue (the original customers) throughout that busy period. When this first sub-busy period is over, at time $t_1 + \tau$ say, customer 0 returns to service and our clock resumes ticking. Another batch arrives at time $t_2 + \tau$, starting a new independent busy period, and so on, until the n sub-busy periods have been completed and the original service time s has elapsed. A new busy period then starts, serving one of the other $k - 1$ customers that arrived in the first batch. The maximum queue length during this busy period has the same distribution

https://doi.org/10.1239/jap/1127322037 Published online by Cambridge University Press

as $M(k-1)$. Because the arrival process is memoryless, this construction is stochastically equivalent to the dynamics of a generic $M^X/G/1$ busy period starting with k customers.

Let $P(k,b) = \mathrm{P}(M(k) \le b)$ and $P(b) = \mathrm{P}(M \le b) = \mathrm{E}(\mathrm{P}(X_0, b))$. Then $P(0,b) = 1$ and $P(0) = 0$. Using the fact that $\mathrm{E}(\mathrm{P}(b-k)^N)$ is the z-transform, or probability-generating function of N evaluated at $z = P(b-k)$, we have, from Lemma 3.1,

$$\begin{aligned} P(k,b) &= \mathrm{E}(\mathrm{P}(M + k \le b)^N) P(k-1,b) \\ &= \mathrm{E}(P(b-k)^N) P(k-1,b) \\ &= \mathrm{E}(\mathrm{e}^{-\lambda(1-P(b-k))S}) P(k-1,b), \qquad 1 \le k \le b. \end{aligned}$$

Corollary 3.1. *For b and k such that $b \ge k \ge 1$, we have*

$$P(k,b) = \prod_{i=1}^{k} \mathrm{E}(\mathrm{e}^{-\lambda(1-P(b-i))S}).$$

If we restrict ourselves to unit batch sizes only, meaning that $X \equiv 1$ and $P(b) = P(1,b)$, we have the following corollary.

Corollary 3.2. *If $X \equiv 1$ then, for $b \ge 1$, we have $P(b) = \mathrm{E}(\mathrm{e}^{-\lambda(1-P(b-1))S})$.*

We can now see how the distribution of S affects M.

Theorem 3.1. *For $M^X/G/1$ queues operating under the LCFS-p-resume discipline, if $S' \le_{\mathrm{LT}} S$ then $M' \le_{\mathrm{st}} M$. In particular, if $S' \ge_{\mathrm{cx}} S$ then $M' \le_{\mathrm{st}} M$.*

Proof. To show that $M' \le_{\mathrm{st}} M$, we show that $P(k,b) \le P'(k,b)$ (with the obvious definition for P') for all k and b, by induction on k and b. We have $P(0,b) = 1 = P'(0,b)$ for each b, and $P(k,1) = 0 = P'(k,1)$ for $k > 1$. Furthermore, since $S' \le_{\mathrm{LT}} S$, we have

$$P(1,1) = \mathrm{P}(S < T) = \mathrm{E}(\mathrm{e}^{-\lambda S}) \le \mathrm{E}(\mathrm{e}^{-\lambda S'}) = P'(1,1).$$

Suppose that $P(i,a) \le P'(i,a)$ for $a < b$ and all $i \ge 0$, meaning that $P(a) \le P'(a)$ for all $a < b$, and suppose that $P(i,b) \le P'(i,b)$ for all $0 \le i < k$. Now consider $P(k,b)$. From Corollary 3.1, the induction hypothesis, and the assumption that $S' \le_{\mathrm{LT}} S$, it follows that

$$\begin{aligned} P(k,b) &= \prod_{i=1}^{k} \mathrm{E}(\mathrm{e}^{-\lambda(1-P(b-i))S}) \\ &\le \prod_{i=1}^{k} \mathrm{E}(\mathrm{e}^{-\lambda(1-P'(b-i))S}) \\ &\le \prod_{i=1}^{k} \mathrm{E}(\mathrm{e}^{-\lambda(1-P'(b-i))S'}) \\ &= P'(k,b). \end{aligned}$$

This completes the proof.

This theorem is illustrated in Figures 1, 2, and 3 for three convex-ordered families of distribution (namely uniform, Pareto, and hyperexponential distributions). In each figure, we plot $\mathrm{P}(M \le n)$, calculated using Corollary 3.2, against n.

https://doi.org/10.1239/jap/1127322037 Published online by Cambridge University Press

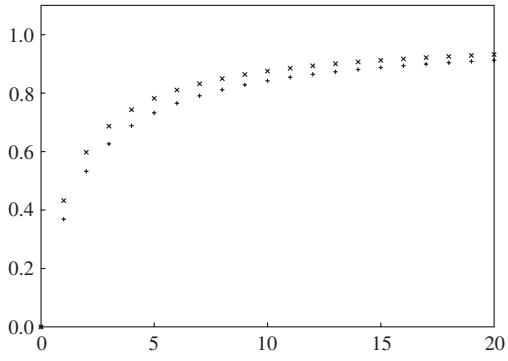

FIGURE 1: A plot of $P(M \leq n)$ against n for uniform service time distributions uniform(0.9, 1.1) (+) and uniform(0, 2) (×), with $E(S) = 1$. We have set $\lambda = 0.9$.

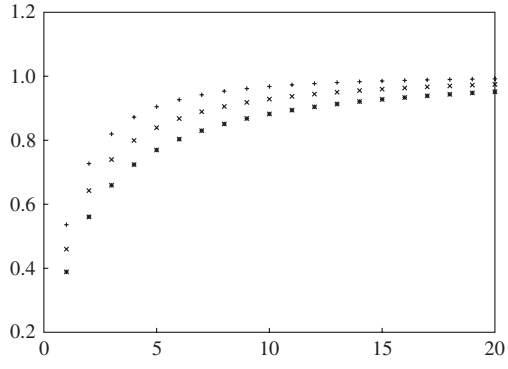

FIGURE 2: A plot of $P(M \leq n)$ against n for Pareto(α) service time distributions with distribution function $F_\alpha(x) = 1 - ((\alpha - 1)/(\alpha x))^\alpha$, $x \geq (\alpha - 1)/\alpha$ (meaning that $E(S) = 1$). We have used $\alpha = 1.5$ (+), $\alpha = 2$ (×), and $\alpha = 10$ (∗), and have set $\lambda = 0.95$.

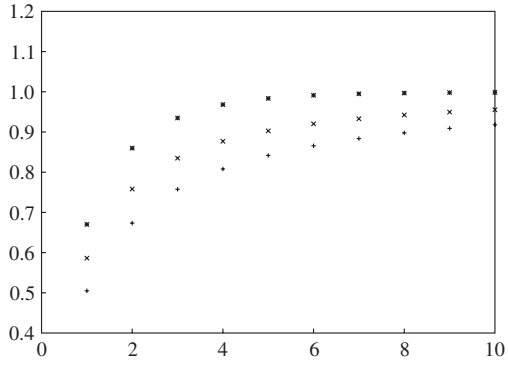

FIGURE 3: A plot of $P(M \leq n)$ against n for hyperexponential(k) service time distributions with $P(S = X_1) = 1 - 2^{-k} = 1 - P(S = X_2)$, where X_1 and X_2 are exponentially distributed with means $1/2(1 - 2^{-k})$ and $2^k/2$, respectively (meaning that $E(S) = 1$). We have used $k = 1$ (+), $k = 3$ (×), and $k = 10$ (∗), and have set $\lambda = 0.95$.

https://doi.org/10.1239/jap/1127322037 Published online by Cambridge University Press

Remark 3.1. It is well known (Kelly (1979)) that the M/G/1 queue under the LCFS-p-resume discipline exhibits service time insensitivity in the sense that the marginal distribution of the number in the stationary system, L, depends on the service time distribution only through its mean. At first this seems at odds with our results, but we must bear in mind that the maximum number in the system during a busy period depends on the sample path evolution of the queue length over a busy period. Hence, the behavior of M and L may be very different. This idea is further illustrated in the following heuristic example.

Example 3.1. Let M be the maximum number in an M/G/1 LCFS-p-resume queue with $S \equiv 1$ (call this system I) and let M' be the corresponding maximum when the first service time in a busy period, S', is equally likely to be ε or $2 - \varepsilon$, meaning that $S \leq_{\text{cx}} S'$, and the other service times in the busy period are identically equal to 1 (call this system II). Then, for ε very small, the first busy period in system II is equally likely to be very short and have a maximum of 1, or to consist essentially of two busy periods, each evolving as a busy period of system I. The second of these busy periods starts when the initial customer has received $1 - \frac{1}{2}\varepsilon$ units of service. That is, roughly speaking, M' is equally likely to be 1 or to have the same distribution as $\max\{M_1, M_2\}$, meaning that $M' \neq_{\text{st}} M$. Note, however, that L and L' have roughly the same distribution. Indeed, $\mathrm{P}(L = 0) = \mathrm{P}(L' = 0)$, since the workload is the same in both systems. Furthermore, a random arrival during a busy period in system II will either see a customer with $S' = \varepsilon$ in service, with very small probability, or will arrive during one of the two busy periods that each evolve as in system I. Finally, note that the distribution of the length of a busy period does depend on the distribution of S.

The M^X/G/1/b LCFS-p-resume queue also exhibits insensitivity, i.e. the distribution of the number in system, L_b, depends on the distribution of S only through its mean. Hence, the loss rate in the M^X/G/1/b queue, $\mathrm{P}(L_b = b)$, is insensitive to the distribution of S. In contrast, our result shows that the probability of at least one loss occurring during a busy period, $\mathrm{P}(M > b)$, does depend on the distribution of S, and is greater when S is larger in the Laplace transform sense.

3.2. LCFS preempt–repeat discipline with resampling

Now we suppose that when services are preempted they must be restarted from scratch. The new service time is assumed to be an independent random variable with the same distribution. We call this the LCFS-p-repeat discipline with resampling. Of course, the behavior of the queue under the LCFS-p-resume and LCFS-p-repeat disciplines is the same when the service times are exponential.

We use the same notation as in the previous subsection. Now, for stability, we need $\lambda\mu\,\mathrm{E}(S_{\mathrm{e}}(S)) < 1$ and $\lambda\mu\,\mathrm{E}(S_{\mathrm{e}}(S')) < 1$, where $S_{\mathrm{e}}(S)$ is the effective service time, i.e. the total time a random customer must spend in service, including restarts due to interruptions. Thus, $\mathrm{E}(S_{\mathrm{e}}(S)) = \mathrm{E}(S \wedge T) + \mathrm{P}(S > T)\,\mathrm{E}(S_{\mathrm{e}}(S))$, where $a \wedge b = \min\{a, b\}$, and, hence,

$$\mathrm{E}(S_{\mathrm{e}}(S)) = \frac{\mathrm{E}(S \wedge T)}{\mathrm{P}(S \leq T)}. \tag{3.1}$$

For T exponentially distributed with rate λ, it is not hard to show that

$$\mathrm{E}(S_{\mathrm{e}}) = \frac{1 - \mathrm{E}(\mathrm{e}^{-\lambda S})}{\lambda\,\mathrm{E}(\mathrm{e}^{-\lambda S})}$$

and, hence, that we need $\mathrm{E}(\mathrm{e}^{-\lambda S}) > \mu/(\mu + 1)$ for stability.

https://doi.org/10.1239/jap/1127322037 Published online by Cambridge University Press

For the M^X/G/1 LCFS-p-repeat queue, we can identify the following embedded random walk. The number in the system at arrival and departure epochs during a busy period is equivalent to a random walk on the nonnegative integers with absorbing state 0. The random walk starts at the random point X_0, decreases by 1 if $T > S$ (a departure), and increases if $T < S$ (an arrival). When it increases, it increases by X, where X is independent of S and T. Thus, we have the following characterization of M, where $I = 1$ if $T < S$, $I = 0$ otherwise, and $M(k) = (M \mid X_0 = k)$.

Lemma 3.2. *The maximum queue length, $M(k)$, for the M^X/G/1 queue under the LCFS-p-repeat discipline satisfies*

$$M(k) \stackrel{\mathrm{D}}{=} I\,M(k+X) + (1-I)\max\{k, M(k-1)\},$$

where $M(0) = 0$; I, X, and $M(k-1)$ are mutually independent; and $M(k+X)$ is independent of I and $M(k-1)$.

Let $I' = 1$ if $T > S'$ and let $I' = 0$ otherwise. If $\mathrm{P}(T > S') \geq \mathrm{P}(T > S)$ then $I' \geq_{\mathrm{st}} I$. From Lemma 3.2 and a coupling argument, it then follows that $M' \leq_{\mathrm{st}} M$. Therefore, we have the following theorem.

Theorem 3.2. *For M^X/G/1 queues operating under the LCFS-p-repeat discipline, if*

$$\mathrm{E}(\mathrm{e}^{-\lambda S'}) \geq \mathrm{E}(\mathrm{e}^{-\lambda S})$$

then $M' \leq_{\mathrm{st}} M$.

Note that for the LCFS-p-repeat discipline, we only need the Laplace transform of the service time evaluated at (the arrival rate) λ to be ordered for two service time distributions, rather than a complete Laplace transform ordering. Thus, all possible distributions of service times can be completely ordered and, hence, we have a complete stochastic ordering of the corresponding maximum queue lengths. Of course, it is also true that $S' \geq_{\mathrm{cx}} S$ implies $S' \leq_{\mathrm{LT}} S$, which in turn implies that $\mathrm{E}(\mathrm{e}^{-\lambda S'}) \geq \mathrm{E}(\mathrm{e}^{-\lambda S})$.

3.3. LCFS preempt–repeat discipline without resampling

For our final model, we suppose again that when services are preempted they must be restarted from scratch, but that the service time is now drawn from the service time distribution only once. We call this the LCFS-p-repeat discipline without resampling. Note that the LCFS-p-repeat disciplines with and without resampling are the same for deterministic service times. For stability, we again need $\lambda\mu\,\mathrm{E}(S_{\mathrm{e}}(S)) < 1$ and $\lambda\mu\,\mathrm{E}(S_{\mathrm{e}}(S')) < 1$, where $S_{\mathrm{e}}(S)$ is the effective service time. Given that $S = s$, the service time is deterministic and the effective service time, S_{e}, is the same as in (3.1), that is

$$\mathrm{E}(S_{\mathrm{e}}(S) \mid S = s) = \frac{\mathrm{E}(s \wedge T)}{\mathrm{P}(s \leq T)} = \frac{1 - \mathrm{e}^{-\lambda s}}{\lambda \mathrm{e}^{-\lambda s}} = \frac{1}{\lambda}[\mathrm{e}^{\lambda s} - 1].$$

Hence,

$$\mathrm{E}(S_{\mathrm{e}}) = \mathrm{E}(\mathrm{E}(S_{\mathrm{e}}(S) \mid S)) = \frac{1}{\lambda}[\mathrm{E}(\mathrm{e}^{\lambda S}) - 1]$$

and, for stability, we need $\mathrm{E}(\mathrm{e}^{\lambda S}) < (\mu + 1)/\mu$. If, for example, S is exponentially distributed with mean ν, then for stability we need $\nu > \lambda(\mu + 1)$. Note that this value is larger than for the repeat discipline with resampling. Intuitively, a large value of the service time has a large

https://doi.org/10.1239/jap/1127322037 Published online by Cambridge University Press

probability of being interrupted and having to restart, and each time it does so it will again have a large service time.

With X_0, S_0, and M defined as in the last subsection, and with T_1 defined to be the first interarrival time after the busy period starts, we now let $M(k, s) = (M \mid X_0 = k,\ S_0 = s)$ and $M(k) = M(k, S) = (M \mid X_0 = k)$. Let $I(s) = 1$ if $T_1 < s$ and let $I(s) = 0$ otherwise. We have the following lemma.

Lemma 3.3. *The maximum queue length, $M(k, s)$, for the $M^X/G/1$ queue under the LCFS-p-repeat discipline without resampling satisfies*

$$M(k, s) \stackrel{\mathrm{D}}{=} I(s) \max\{M + k, M(k, s)\} + (1 - I(s)) \max\{k, M(k - 1)\},$$

where $M(0) = 0$ and $I(s)$, M, $M(k, s)$, and $M(k - 1)$ are independent.

Proof. Given that $X_0 = k$ and $S_0 = s$, if an arrival occurs before the first service completion, a new (sub-)busy period starts, except that there are k additional customers in the queue. When this sub-busy period ends, the original busy period starts again (independently of T_1 and of M for the sub-busy period just ended) with $X_0 = k$ and $S_0 = s$. If the first service is completed before an arrival, then we may consider the remainder of the busy period to be a new independent busy period with $k - 1$ initial customers.

Let $P(k, b, s) = \mathrm{P}(M(k, s) \le b)$, $P(k, b) = \mathrm{P}(M(k) \le b) = \mathrm{E}(P(k, b, S_0))$, and $P(b) = \mathrm{P}(M < b) = \mathrm{E}(P(X_0, b, S_0))$. We then have the following corollary to Lemma 3.3.

Corollary 3.3. *For b and k such that $b \ge k \ge 1$, and for all s, we have*

$$P(k, b, s) = \frac{\mathrm{P}(T > s) P(k - 1, b)}{1 - \mathrm{P}(T < s) P(b - k)} = \frac{\mathrm{e}^{-\lambda s} P(k - 1, b)}{1 - (1 - \mathrm{e}^{-\lambda s}) P(b - k)}.$$

Using this corollary, we can prove the following result.

Theorem 3.3. *For $M^X/G/1$ queues operating under the LCFS-p-repeat discipline without resampling, if $S' \le_{\mathrm{icv}} S$ then $M' \le_{\mathrm{st}} M$. Hence, if $S' \ge_{\mathrm{cx}} S$ then $M' \le_{\mathrm{st}} M$.*

Proof. From Corollary 3.3, for b and k such that $b \ge k \ge 1$, and $s \ge 0$, we have

$$P(k, b, s) = \frac{P(k - 1, b)}{\mathrm{e}^{\lambda s}(1 - P(b - k)) + P(b - k)}.$$

It is easy to show that $f(s) := a/(c\mathrm{e}^{\lambda s} + d)$ is a decreasing convex function of s for all $a, c, d \ge 0$ such that $c + d > 0$ (meaning that $-f(s)$ is increasing and concave). Hence, if $S' \le_{\mathrm{icv}} S$ then $-\mathrm{E}(f(S')) \le -\mathrm{E}(f(S))$ and $\mathrm{E}(f(S')) \ge \mathrm{E}(f(S))$. Also note that $P(k, b, s)$ is increasing in $P(k - 1, b)$ and $P(b - k)$ for fixed s. The result now follows from an induction argument similar to the one in the proof of Theorem 3.1.

4. Other queueing disciplines

We were unable to obtain results for other queueing disciplines, though we make the following conjectures. By the foreground–background discipline (also known as 'LAS' or 'LAST' for 'least attained service time first') we mean the discipline that always serves the customer who has so far received the least amount of service. Thus, like the LCFS preemptive disciplines, new arrivals immediately preempt customers in service, but, once a new arrival

https://doi.org/10.1239/jap/1127322037 Published online by Cambridge University Press

attains as much service as other customers, the multiple customers with the least attained service are served in a processor-sharing fashion. Note that for nonpreemptive disciplines, such as the first-come–first-served and (nonpreemptive) LCFS disciplines, the number in the system and, hence, the maximum queue length in a busy period are independent of the particular discipline. We conjecture that, under the foreground–background discipline, if $S' \geq_{\mathrm{cx}} S$ then $M' \leq_{\mathrm{st}} M$. Under nonpreemptive service disciplines, if $S' \geq_{\mathrm{icx}} S$ then $M' \geq_{\mathrm{icx}} M$, where $X \geq_{\mathrm{icx}} Y$ means that $\mathrm{E}(f(X)) \geq \mathrm{E}(f(Y))$ for all increasing convex functions f.

Our second conjecture is contrary to the results for preemptive LCFS disciplines in this paper, in the sense that, when service times become more variable, the performance of nonpreemptive disciplines, as measured by the maximum queue length, deteriorates. This contrast may be explained as follows: when service times are highly variable, the preemptive LCFS and foreground–background disciplines permit customers with shorter services to be served first, which can improve performance relative to nonpreemptive disciplines. Furthermore, the effect of a larger variability in nonpreemptive systems is consistent with our intuition and results for other performance measures like the mean waiting time.

Acknowledgement

We would like to thank the referee for a careful review and excellent suggestions.

References

Adiri, I., Frostig, E. and Rinnooy Kan, A. H. G. (1991). Scheduling on a single machine with a single breakdown to minimize stochastically the number of tardy jobs. *Naval Res. Logistics* **38,** 261–271.

Birge, J., Frenk, J. B. G., Mittenthal, J. and Rinnooy Kan, A. H. G. (1990). Single-machine scheduling subject to stochastic breakdowns. *Naval Res. Logistics* **37,** 661–677.

Cai, X., Sun, X. Q. and Zhou, X. (2004). Stochastic scheduling subject to machine breakdowns: the preemptive-repeat model with discounted reward and other criteria. *Naval Res. Logistics* **51,** 800–817.

Cai, X., Wu, X. Y. and Zhou, X. (2005). Dynamically optimal policies for stochastic scheduling subject to preemptive-repeat machine breakdowns. To appear in *IEEE Trans. Automation Sci. Eng.*

Chang, C.-S., Chao, X., Pinedo, M. and Shanthikumar, J. G. (1991). Stochastic convexity for multidimensional processes and its applications. *IEEE Trans. Automatic Control* **36,** 1347–1355.

Hordijk, A. (2001). Comparison of queues with different discrete-time arrival processes. *Prob. Eng. Inf. Sci*. **15,** 1–14.

Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. John Wiley, New York.

Makowski, A. M. (1994). On an elementary characterization of the increasing convex ordering, with an application. *J. Appl. Prob*. **31,** 834–840.

Miyazawa, M. (1990). Complementary generating functions for $M^X/GI/1/k$ and $GI/M^Y/1/k$ queues and their application to the comparison of loss probabilities. *J. Appl. Prob.* **27,** 684–692.

Miyazawa, M. and Shanthikumar, J. G. (1991). Monotonicity of the loss probabilities of single server finite queues with respect to convex order of arrival or service processes. *Prob. Eng. Inf. Sci.* **5,** 43–52.

Shanthikumar, J. G. and Yao, D. D. (1994). Stochastic comparisons in closed Jackson networks. In *Stochastic Orders and Their Applications*, eds M. Shaked and J. G. Shanthikumar, Academic Press, New York, pp. 433–460.