Probability in the Engineering and Informational Sciences, **15**, 2001, 179–187. Printed in the U.S.A.

NOTE ON THE *GI/GI/*1 QUEUE WITH LCFS-PR OBSERVED AT ARBITRARY TIMES

RUDESINDO NÚÑEZ-QUEIJA
CWI, P.O. Box 94079
1090 GB Amsterdam, The Netherlands
E-mail: sindo@cwi.nl

Consider the *GI/GI/*1 queue with the Last-Come First-Served Preemptive-Resume service discipline. We give intuitive explanations for (1) the geometric nature of the stationary queue length distribution and (2) the mutual independence of the residual service requirements of the customers in the queue, both considered at arbitrary time points. These distributions have previously been established in the literature by either first considering the system at arrival instants or using balance equations. Our direct arguments provide further understanding of properties 1 and 2.

1. INTRODUCTION

The steady-state distributions of queue length and residual service requirements of the customers in the *GI/GI/*1 queue with the LCFS-PR (Last-Come First-Served Preemptive Resume) service discipline are well known. The distribution of the queue length considered only *at arrival* (or *departure*) *instants* is geometric and the remaining service requirements of the customers are independent and identically distributed (i.i.d.). At arbitrary time instants, the queue length distribution is geometric too—except for the probability of an empty queue—with the same parameter as that at arrival (and departure) instants. Furthermore, the remaining service requirements of all customers but the one in service are i.i.d. with the same distribution as before. The remaining service requirement of the customer in service has the forward recurrence distribution of the service requirements and is independent of the queue length and the other service requirements.

In the queueing literature (see the next paragraph for a short overview), the distributions at arbitrary times have been derived either from the steady-state dis-

 0269-9648/01 $12.50

tributions at arrival and/or departure instants or by solving the steady-state balance equations. Our purpose is to give direct arguments that lead to these time-average distributions and, at the same time, provide understanding of the results. Our arguments rely on basic renewal theory.

Let us briefly review the literature on the LCFS-PR discipline. For the case of Poisson arrivals, the joint distribution of the number of customers in the system and their residual service requirements was derived by Kelly [8]. Fakinos [4] extended the results to general interarrival time distributions by deriving the joint distribution of the queue length and the remaining service requirements *at arrival instants*. The proofs are based on the analysis of ascending ladder indices [10, p. 309]. Fakinos [4] further remarked that at departure instants, these distributions must be the same as at arrival instants, a fact that was proved by Yamazaki [19]. Direct and insightful arguments for these findings were provided later by Fakinos [5]. The corresponding distributions at arbitrary time instants were first derived by Yamazaki [20] and for a more general model, with queue-dependent services, by Fakinos [6] (both used balance equations). Shanthikumar and Sumita [16] considered generalizations in several directions (interarrival times not i.i.d. and queue-dependent acceptance probabilities, more general service disciplines). Using sample-path arguments and renewal theory, they related time averages and customer averages. Part of our approach relies on similar arguments. The analysis of the LCFS-PR discipline proved to be very useful for studying the workload distribution in queues. Fakinos [4] already observed that his results gave new insight into the workload distribution in the *GI/GI/*1 queue. Cooper and Niu [3] exploited the special case with Poisson arrivals to explain Beneš's inversion of the Pollaczek–Khintchine formula. Niu [13] gave representations for the workload in the *GI/GI/*1 queue.

The structure of the article is as follows. In Section 2, we specify the model and provide a preliminary analysis of the sojourn time of customers in the system. In Section 3, the geometric nature of the queue length distribution at arbitrary times is explained, and in Section 4, we extend the analysis to the residual service requirements of customers. In Section 5, we briefly comment on the special case of exponentially distributed service requirements. Section 6 concludes the article.

2. DESCRIPTION OF THE MODEL AND PRELIMINARY ANALYSIS

Let the cumulative distribution functions of the interarrival times and the service requirements be denoted by $A(x), x \geq 0$, and $B(x), x \geq 0$, respectively, with $A(0+) = B(0+) = 0$. We assume that the mean interarrival time a and the mean service requirement b are finite and that the queue is stable (i.e., $a > b$).

In the LCFS-PR discipline, a newly arriving customer is immediately taken into service. If, upon arrival of the new customer, there is a customer in service, then this service is interrupted, to be resumed at the moment that the new customer leaves the system. Note that the new customer's service can also be interrupted by subsequently arriving customers. The total sojourn time of a customer equals the time needed to

decrease the amount of work in the system by a random amount distributed according to $B(x)$ (the customer's service requirement), starting just after an arrival. Thus, the sojourn time of any customer is distributed as the busy period of the *GI/GI/1* queue and, moreover, it is independent of previous arrivals and service requirements. In particular, the sojourn time is independent of the number of customers found in the system.

To facilitate the presentation, it is convenient to decompose the sojourn time, as is done below. This decomposition is well known for the busy period of the *GI/GI/1* queue, but in order to set the notation, we give the decomposition in detail. In Figure 1, a typical sojourn time is depicted. Let B be the service requirement of an arriving customer (which we will indicate by $*$) and, for concreteness, let $n \geq 0$ be the number of customers present just previous to the arrival. Immediately upon arrival, customer $*$ is taken into service. Let the time until the next arrival be denoted by A_0; clearly, A_0 has cumulative distribution function $A(x)$. If $B \leq A_0$, then the service of customer $*$ is not interrupted and its sojourn time S equals B. An example of the case when $B > A_0$ is depicted in Figure 1. The arrow pointing upward (just after A_0 has elapsed) indicates that at that time, the number of customers in the system is increased from $n+1$ to $n+2$. The service of customer $*$ is interrupted at that moment and is resumed as soon as the number of customers decreases again from $n+2$ to $n+1$ (in Fig. 1, this is indicated by a downward arrow). The length of this interruption, which we denote by S_1, equals the sojourn time of the customer that entered after A_0. By the arguments given above, S_1 is distributed as the busy period of the *GI/GI/1* queue. At the end of S_1, the service of customer $*$ is resumed until the next arrival; this period of service is denoted by A_1. Note that, in general, A_1 is *not* distributed according to $A(x)$, since at the end of the busy period S_1, part of the current interarrival time has already elapsed. Instead, A_1 is distributed as the idle time between two busy periods in the *GI/GI/1* queue; see, for instance, Cohen [1, p. 283]. At the end of A_1, the service of customer $*$ is interrupted for a period S_2, which is again distributed as a busy period, followed by a period A_2 of service for customer $*$, which is distributed as an idle period, and so forth. In Figure 1, the service of customer $*$ is completed during A_3, indicated by a downward arrow marked by $*$ (the number of customers decreases from $n+1$ to n). If there was a customer in service when customer $*$ arrived (i.e., if $n \geq 1$), then this customer's service is resumed; otherwise, an idle period follows until the next arrival.

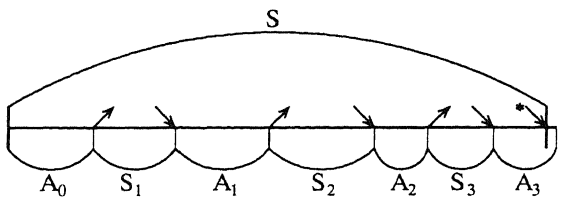

FIGURE 1. Customer's sojourn time.

Remark 2.1: In general, the random variables A_k and S_k, $k = 1,2,3,\ldots$, are not independent; however, the pairs $(S_1,A_1),(S_2,A_2),(S_3,A_3),\ldots$ form an i.i.d. sequence. The pair (S_k,A_k) constitutes a busy cycle with busy period S_k and idle period A_k, $k = 1,2,\ldots$ (cf. [1, p. 283]).

Let $\sigma_0 := 0$ and, for $k = 1,2,3,\ldots$,

$$\sigma_k := \sum_{j=1}^{k} A_{j-1}.$$

Furthermore, let

$$I(B) := \sup\{k : \sigma_k < B\}$$

be the number of times that the service of customer $*$ is interrupted. Then, the sojourn time of customer $*$ is given by

$$S = B + \sum_{k=1}^{I(B)} S_k, \tag{1}$$

where, by convention, we set the empty sum equal to 0. Note that S has the same marginal distribution as (but, clearly, is not independent of) the S_k.

3. GEOMETRIC QUEUE LENGTH DISTRIBUTION

We first argue that the distribution of the queue length at arbitrary time points is geometric (apart from the probability of an empty queue). The approach is similar to that used by several authors (e.g., Kleinrock [10, p. 247], Wolff [18, p. 396], Tijms [17, p. 128]) to determine the queue length distribution of the $GI/M/1$ queue at arrival instants. In Section 4, we show how the arguments can be extended to derive the distribution of the residual service requirement of the customers in the system.

Suppose we start at time $t = 0$ with less than $k \in \{1,2,\ldots\}$ customers in the system. Let T_1 be the first time that an arriving customer increases the number of customers in the system from $k-1$ to k and let T_2 be the first moment (thereafter) that the number of customers decreases again from k to $k-1$. Since $T_2 - T_1$ is equal to the sojourn time of the customer that arrived at time T_1, it is distributed as the busy period of the $GI/GI/1$ queue (cf. Sect. 2).

Let $N(t)$ be the queue length at time t and let N be distributed according to the stationary queue length distribution (here defined as the Césaro limit):

$$\mathbf{P}\{N = k\} = \lim_{t\to\infty} \frac{1}{t} \int_{u=0}^{t} \mathbf{P}\{N(u) = k\}\, du, \quad k \in \{0,1,2,\ldots\}.$$

Before giving a formal derivation, we provide the following intuitive argument for the distribution of $N|N > 0$ to be geometric. Note that immediately after time T_2, the queue length is less than k; therefore, we might define T_3 to be the next time instant at which the queue length is again equal to k. Necessarily, this must be immediately after an arrival (customers arrive one at a time since we

assumed that $A(0+) = 0$; see Sect. 6 when $A(0+) > 0$). Therefore, the processes $\{N(T_1 + t), t \geq 0\}$ and $\{N(T_3 + t), t \geq 0\}$ have the same distribution until the next visit to level $k - 1$ after time instants T_1 and T_3, respectively. Let $k \in \mathbb{N} := \{1,2,3,\ldots\}$ and $m \in \mathbb{N}_0 := \mathbb{N} \cup \{0\} = \{0,1,2,\ldots\}$. If we "delete" all periods of time during which the queue length is less than k and concatenate all periods with at least k customers, in the newly formed process the steady-state probability of $k + m$ customers in the system is equal to $\mathbf{P}\{N = k + m | N \geq k\}$, and by the arguments given earlier, this probability is independent of k. Therefore, $N|N > 0$ must have a geometrical distribution; see, for instance, Feller [7, Sect. XIII.9].

The queue length distribution may be formally derived along the following lines. Let C_k, $k \in \mathbb{N}$, be distributed as the amount of time between two consecutive moments at which the queue length increases from $k - 1$ to k. From Remark 2.1, we know that C_1 is distributed as the busy cycle in the *GI/GI/1* queue. Since we assumed the queue to be stable, we have $\mathbf{E}[C_1] < \infty$ (cf. Cohen [1, p. 286]). Let $\tau_{k,j}$, $k \in \mathbb{N}$ and $j \in \mathbb{N}_0$, be the expected amount of time spent with $k + j$ customers in the system during a period C_k. By the arguments given in Section 2, $\tau_{k,j+1} = \mathbf{E}[I(B)]\tau_{k+1,j}$ (see also Remark 3.1). Moreover, $\tau_{k,j}$ is independent of k: $\tau_{k,j} =: \tau_j$. Because of the Renewal–Reward theorem (see, for instance, Ross [14, Thm. 3.16, p. 52]),

$$\mathbf{P}\{N = j + 1\} = \frac{\tau_j}{\mathbf{E}[C_k]}.$$

Hence, for any $k \in \mathbb{N}$,

$$\frac{\mathbf{P}\{N = k + 1\}}{\mathbf{P}\{N = k\}} = \mathbf{E}[I(B)] =: \gamma,$$

so that $N|N > 0$ is geometrically distributed.

Remark 3.1: Note that γ is the expected number of "up-crossings" from k to $k + 1$ during C_k. After each such up-crossing, the expected amount of time spent with $k + j + 1$ customers until the next "down-crossing" to k is $\tau_{k+1,j}$, and so $\tau_{k,j+1} = \gamma\tau_{k+1,j}$. Using $\tau_{k,j} = \tau_j$ this directly implies $\tau_j = \gamma^j\tau_0$, $j \in \mathbb{N}_0$.

To find the complete queue length distribution, it suffices to note that $\mathbf{P}\{N = 0\} = 1 - \rho$, where $\rho := b/a$ is the traffic load. Thus,

$$\mathbf{P}\{N = k\} = \rho(1 - \gamma)\gamma^{k-1}, \quad k \in \mathbb{N}. \tag{2}$$

Using Little's law, the parameter γ can be expressed in terms of the mean busy period (which equals the mean sojourn time $\mathbf{E}[S]$):

$$\frac{\rho}{1 - \gamma} = \frac{1}{a}\mathbf{E}[S];$$

hence, $\gamma = 1 - b/\mathbf{E}[S]$.

Remark 3.2: Computing the parameter γ is therefore as difficult as computing the mean busy period. See [1, p. 286] for a formal expression of the latter.

4. RESIDUAL SERVICE REQUIREMENTS

We extend the results of the previous section, deriving the joint distribution of the queue length and the residual service requirements of the customers in the system. As earlier, let N denote the queue length in equilibrium and, given that $N = n \in \mathbb{N}$, let $X_k, k = 1,2,\ldots,n$, be the service requirement of the kth customer in the system. By convention, the kth customer in the system arrived later than the $(k-1)$st and prior to the $(k+1)$st.

OBSERVATION 4.1: *The distribution of X_1 given that $N = 1$ equals the excess distribution of the service requirements; that is,*

$$\mathbf{P}\{X_1 \leq x | N = 1\} = \tilde{B}(x) := \int_{u=0}^{x} \frac{1 - B(u)}{b}\, du.$$

To see this, suppose that we only monitor the queue length process $N(t)$ when there is exactly one customer in the system and we "delete" all periods during which $N(t) \neq 1$. What we observe is the concatenated sequence of service periods of customers that arrived to an empty system. (In Fig. 1, eliminate all periods S_k and the part of A_3 after the departure of customer $*$; what remains is exactly the service time of customer $*$.) The latter process is just a renewal process with renewal times drawn from $B(x)$.

$\tilde{B}(x)$ is the distribution of the residual service requirement of the first customer in a busy period, given that that customer is being served. In addition, it is convenient to introduce the distribution function of the residual service requirement of the first customer in the busy period when that customer's service has been interrupted:

$$B_I(x_1) = \mathbf{P}\{X_1 \leq x_1 | N \geq 2\}.$$

For $n \geq 2$, we write

$$\begin{aligned}&\mathbf{P}\{N = n, X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\}\\ &\quad = \mathbf{P}\{N = n, X_2 \leq x_2, \ldots, X_n \leq x_n | N \geq 2, X_1 \leq x_1\}\mathbf{P}\{N \geq 2\}B_I(x_1). \qquad (3)\end{aligned}$$

OBSERVATION 4.2: *For $n \geq 2$,*

$$\begin{aligned}&\mathbf{P}\{N = n, X_2 \leq x_2, \ldots, X_n \leq x_n | N \geq 2, X_1 \leq x_1\}\\ &\quad = \mathbf{P}\{N = n, X_2 \leq x_2, \ldots, X_n \leq x_n | N \geq 2\}\\ &\quad = \mathbf{P}\{N = n-1, X_1 \leq x_2, \ldots, X_{n-1} \leq x_n | N \geq 1\}.\end{aligned}$$

The first equality (independence of X_1) follows in the same way as the independence of the S_k from the state of the system in Eq. (1) and in Figure 1: The behavior of the queue length process above the level 1 (i.e., during a service interruption of the first customer) is independent of the residual service requirement of the first customer. A constructive proof may be given as earlier, by considering the system only at times when there are at least two customers in the system *and* the residual service requirement of the first customer is at most x_1. The stochastic evo-

lution of the processes (queue length and residual service requirements of all customers but the first) that result from this construction is independent of x_1. Similarly, the second equality (shift in level) follows from the fact that if we observe the system only at times that there are at least k customers, the number of additional customers (besides the first k) and their service requirements evolve stochastically in the same way for all k. Using Observation 4.2 repeatedly in Eq. (3), we have, for $n \geq 2$,

$$\begin{aligned}
&\mathbf{P}\{N = n, X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\} \\
&\quad = \mathbf{P}\{N = n-1, X_1 \leq x_2, \ldots, X_{n-1} \leq x_n | N \geq 1\} \mathbf{P}\{N \geq 2\}\, B_I(x_1) \\
&\quad = \mathbf{P}\{N = n-1, X_1 \leq x_2, \ldots, X_{n-1} \leq x_n\} \frac{\mathbf{P}\{N \geq 2\}}{\mathbf{P}\{N \geq 1\}} B_I(x_1) \\
&\quad \vdots \\
&\quad = \left(\frac{\mathbf{P}\{N \geq 2\}}{\mathbf{P}\{N \geq 1\}}\right)^{n-1} \mathbf{P}\{X_1 \leq x_n, N = 1\} \prod_{k=1}^{n-1} B_I(x_k). \qquad (4)
\end{aligned}$$

Setting the empty product equal to 1 and using Eq. (2) and Observation 4.1, we have, for $n \in \mathbb{N}$,

$$\mathbf{P}\{N = n, X_1 \leq x_1, \ldots, X_n \leq x_n\} = \rho(1-\gamma)\gamma^{n-1}\tilde{B}(x_n) \prod_{k=1}^{n-1} B_I(x_k). \qquad (5)$$

Remark 4.3: $B_I(x)$ can be shown to have the same distribution as the idle period in the dual queueing model (see Fakinos [6, Sect. 3]).

5. DISCUSSION OF THE *GI/M/1* QUEUE

When the service requirements have an exponential distribution, the evolution of the queue length is stochastically indistinguishable for all work-conserving service disciplines. Therefore, the geometric queue length distribution in the *GI/M/1* queue with FCFS (first-come first-served) services can be explained from the results for the LCFS-PR discipline. The *GI/M/1* queue was already extensively studied prior to Fakinos' [4] analysis of the *GI/GI/1* queue with LCFS-PR. The *GI/M/1* queue length distribution at arrival epochs was first obtained by Kendall [9], and from it, the distribution at arbitrary time points could be determined (see, e.g., Cohen [1, p. 208]). Alternative derivations using the Laplace transform with respect to time of the transient distribution, were given by Conolly [2], Saaty [15, p. 223], and Cohen [1, p. 222].

In this case, the parameter γ, which according to Remark 3.1 equals the expected number of service interruptions of an arbitrary customer, is the unique solution to the equation

$$\gamma = \alpha(\mu(1-\gamma)), \quad \gamma \in (0,1),$$

where $1/\mu\,(=b)$ is the mean service requirement and $\alpha(s)$, $\mathrm{Re}(s) \geq 0$, is the Laplace–Stieltjes transform of the interarrival time distribution $A(x)$.

As a consequence of the geometric nature of the steady-state queue length distribution—both at arrival instants and at arbitrary time points—and the exponential service times, various performance measures have the same exponential distribution, with a possible additional atom at 0. For the sojourn time S (under FCFS), the waiting time W (under FCFS), and the virtual waiting time V, we find, for $x > 0$,

$$\mathbf{P}\{S \leq x\} = \mathbf{P}\{W \leq x | W > 0\} = \mathbf{P}\{V \leq x | V > 0\} = 1 - e^{-(1-\gamma)\mu x}.$$

All three random variables (W and V conditioned on being positive) can be written as the sum of a geometric number (with mean $1/(1-\gamma)$) of independent and identical exponentially distributed terms (each with mean $1/\mu$).

Remark 5.1: Many results for the $GI/M/1$ queue easily generalize to the $GI/M/c$ queue with $c \geq 1$ parallel servers (see also Wolff [18, p. 398]). The queue length distribution at arbitrary times is geometric for queue sizes of c and larger. The parameter $\gamma_{(c)}$ of the geometric tail is determined by the equation $\gamma_{(c)} = \alpha(c\mu(1 - \gamma_{(c)}))$. Also, the steady-state waiting time $W_{(c)}$ (FCFS) and virtual waiting time $V_{(c)}$ are exponentially distributed (with an atom at 0):

$$\mathbf{P}\{W_{(c)} \leq x | W_{(c)} > 0\} = \mathbf{P}\{V_{(c)} \leq x | V_{(c)} > 0\} = 1 - e^{-(1-\gamma_{(c)})c\mu x}.$$

The sojourn time is *not* exponentially distributed. When the number of other customers upon arrival is less than c, the sojourn time equals a single service time (exponential with mean $1/\mu$); otherwise, it is the sum of a waiting time (exponential with mean $1/((1 - \gamma_{(c)})c\mu)$) and a service time.

6. CONCLUDING REMARKS

We have studied the steady-state distributions of queue length and residual service requirements at arbitrary times in the $GI/GI/1$ queue with LCFS-PR. In the queueing literature, these distributions have been obtained either through the steady-state distributions at arrival and departure times or through the balance equations. Our arguments apply directly to the system in continuous time, thus explaining the geometric nature of the queue length distribution as well as the fact that residual service requirements are independent and all but one are identically distributed.

The analysis can be extended in a straightforward manner to the case where customers arrive in batches having a geometric size distribution. This corresponds to allowing $A(0+) \in (0,1)$ (i.e., interarrival times may be equal to 0). In particular, in Section 3, a batch arrival of m customers when there are k already present must be counted as up-crossings of the levels $k, k+1, \ldots, k+m-1$. Another possible extension is to batch *services* with general batch size distributions. This is possible as long as the batch sizes do not depend on the queue length except for truncation of the batch when an attempt is made at servicing more customers than those present in the queue (see also Neuts [12, p. 183]). By similar arguments as those used in this article, a probabilistic treatment of the matrix-geometric theory developed by Neuts [12] can be given (see Latouche and Ramaswami [11]).

Acknowledgments

The author thanks Sem Borst, Richard Boucherie, Onno Boxma, and Jacques Resing for useful comments.

References

1. Cohen, J.W. (1982). *The single server queue*, rev. ed. Amsterdam: North-Holland.
2. Conolly, B. (1958). A difference equation technique applied to the simple queue with arbitrary arrival interval distribution. *Journal of the Royal Statistical Society Series B* 20: 168–175.
3. Cooper, R.B. & Niu, S.C. (1986). Beneš's formula for M/G/1-FIFO "explained" by preemptive-resume LIFO. *Journal of Applied Probability* 23: 550–554.
4. Fakinos, D. (1981). The G/G/1 queueing system with a particular queue discipline. *Journal of the Royal Statistical Society Series B* 43: 190–196.
5. Fakinos, D. (1986). On the single-server queue with the preemptive-resume last-come first-served queue discipline. *Journal of Applied Probability* 23: 243–248.
6. Fakinos, D. (1987). The single-server queue with service depending on queue size with the preemptive-resume last-come first-served queue discipline. *Journal of Applied Probability* 24: 758–767.
7. Feller, W. (1968). *An introduction to probability theory and its applications*, Vol. I, 3rd ed. New York: Wiley.
8. Kelly, F.P. (1976). The departure process from a queueing system. *Mathematical Proceedings of the Cambridge Philosophical Society* 80: 283–285.
9. Kendall, D.G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the embedded Markov chain. *Annals of Mathematical Statistics* 24: 338–354.
10. Kleinrock, L. (1975). *Queueing systems.* Vol. I: *Theory.* New York: Wiley.
11. Latouche, G. & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling.* Alexandria/Philadelphia: ASA/SIAM.
12. Neuts, M.F. (1981). *Matrix-geometric solutions in stochastic models—An algorithmic approach.* Baltimore: Johns Hopkins University Press.
13. Niu, S.C. (1988). Representing workloads in GI/G/1 queues through the preemptive-resume LIFO queue discipline. *Queueing Systems* 3: 157–178.
14. Ross, S.M. (1970). *Applied probability models with optimization applications.* San Francisco: Holden-Day.
15. Saaty, T.L. (1961). *Elements of queueing theory.* Chichester: McGraw-Hill.
16. Shanthikumar, J.G. & Sumita, U. (1986). On G/G/1 queues with LIFO-P service discipline. *Journal of the Operations Research Society of Japan* 29: 220–231.
17. Tijms, H.C. (1994). *Stochastic models: An algorithmic approach.* Chichester: Wiley.
18. Wolff, R.W. (1989). *Stochastic modeling and the theory of queues.* Englewood Cliffs, NJ: Prentice-Hall.
19. Yamazaki, G. (1982). The GI/G/1 queue with last-come-first-served. *Annals of the Institute of Statistical Mathematics* 34: 599–604.
20. Yamazaki, G. (1984). Invariance relations of GI/G/1 queueing systems with preemptive-resume last-come first-served queue discipline. *Journal of the Operations Research Society of Japan* 27: 338–347.