Representing Workloads in GI/G/1 Queues through the Preemptive-Resume LIFO Queue Discipline

Shun-Chen Niu
School of Management
The University of Texas at Dallas
P. O. Box 830688
Richardson, Texas 75083-0688

September 1986
Revision: December 1987[1]

[1] *Queueing Systems*, 3 (1988), pp. 157–178.

Abstract

We give in this paper a detailed sample-average analysis of $GI/G/1$ queues with the preemptive-resume LIFO (last-in-first-out) queue discipline: We study the long-run "state" behavior of the system by averaging over arrival epochs, departure epochs, as well as time, and obtain relations that express the resulting averages in terms of basic characteristics within busy cycles. These relations, together with the fact that the preemptive-resume LIFO queue discipline is work-conserving, imply new representations for both "actual" and "virtual" delays in standard $GI/G/1$ queues with the FIFO (first-in-first-out) queue discipline. The arguments by which our results are obtained unveil the underlying structural "explanations" for many classical and somewhat mysterious results relating to queue lengths and/or delays in standard $GI/G/1$ queues, including the well-known Beneš's formula for the delay distribution in $M/G/1$. We also discuss how to extend our results to settings more general than $GI/G/1$.

AMS 1980 subject classification. Primary: 90B22; Secondary: 60K25.
IAOR 1973 subject classification. Main: Queues.
OR/MS Index 1978 subject classification. Primary: 681 Queues.

Key words. $GI/G/1$ queues, actual and virtual delays, preemptive-resume LIFO queue discipline, Pollaczek-Khintchine formula.

1 Introduction

An obvious observation concerning workloads in $GI/G/1$ queues is that they are invariant with respect to queue disciplines that do not affect the length of the required service time of any customer, that is, work-conserving queue disciplines. Suppose we are given the task of analyzing the workload in a $GI/G/1$ queue. Among all work-conserving queue disciplines, which one should we select to facilitate analysis? The purpose of this paper is to show that a remarkably simple analysis is possible if we choose the preemptive-resume LIFO queue discipline. Under this queue discipline, each customer enters service immediately upon arrival, preempting the customer, if any, in service; the preempted customer backs up to the head of the queue and resumes his service from the point of interruption when he reenters service again.

The possibility of analyzing the workload through the preemptive-resume LIFO queue discipline has been noted in a number of recent papers. Interests in this model primarily grew out of the desire to provide probabilistic "explanations" to several classical and somewhat mysterious results concerning the queue-length and/or delay distributions in the standard FIFO model, particularly that (i) for the $GI/M/1$ queue, the equilibrium queue-length distribution for the Markov chain embedded at arrival epochs is geometric, and (ii) the well-known inversion (Beneš [1]) of the Pollaczek-Khintchine formula, which gives the Laplace-Stieltjes transform of the equilibrium delay distribution for the $M/G/1$ queue, takes the form of a geometric convolution of stationary excess service-time distributions.

Generalizing a corresponding result of Kelly [12], Fakinos [6] (and subsequently Yamazaki [24] and Fakinos [7]) proved that when the preemptive-resume LIFO model is observed exclusively at either arrival or departure epochs, the equilibrium queue-length distribution is geometric; and moreover, conditional on the queue length being fixed, the remaining service times of the customers waiting in queue are independently distributed as the idle period in the corresponding dual queue (obtained from the "original" by interchanging the interarrival-time and service-time distributions). This interesting result generalizes and helps to explain (i) (the queue-length distribution in $GI/M/1$ is the same under both FIFO and preemptive-resume LIFO queue disciplines—Fakinos [6], p. 196); it also generalizes but does not, however, readily specialize to (ii) (Fakinos [6], pp. 194–195), because the dual queue, being unstable, is not particularly easy to work with. Yamazaki [25] as well as Shanthikumar and Sumita [20] gave additional discussions on generalizations of (i). Cooper and Niu [5] supplied an intuitive argument for (ii).

As observed by Fakinos [6, 7] as well as by Cooper and Niu [5], a key property of the preemptive-resume LIFO model is that the experience of any customer is not affected in any way by those who arrive before him. By making constructive use of this property, we obtain in this paper a new relation that expresses, in a form similar to Fakinos's result, the long-run average "state" behavior of the system as observed exclusively at arrival epochs in terms of a "delayed" renewal function defined by the idle period in the "original" queue

(see Theorem 1, Section 3, for a precise statement). This relation, which we consider as our main result, provides additional insight into the preemptive-resume LIFO model, and in particular, specializes in interesting ways to both (i) and (ii); it also is, we believe, potentially useful for studying qualitative properties of delays in standard FIFO $GI/G/1$ queues.

Our method of analysis—relating sample averages—has been used by many authors to study queueing problems, particularly in connection with the fundamental relation "$L = \lambda W$" or "$H = \lambda G$"; see, for example, Brumelle [2], Stidham [21], Heyman and Stidham [14], Wolff [23], and Niu [17]. For the preemptive-resume LIFO model, Shanthikumar and Sumita [20] also followed a similar approach; and Yamazaki [24, 25] followed the "marked point process" approach.

The outline of the rest of this paper is as follows. We begin in Section 2 with a description of the necessary notation and assumptions. We then, in Section 3, establish the main result and discuss its immediate consequences, including similar results for long-run average "state" behaviors over departure epochs and over time, as well as the well-known Takács's relation between "actual" and "virtual" workloads in $GI/G/1$ queues. (A referee has pointed out that results similar to those given in Section 3 have also been obtained, by different methods, in a recent paper by Fakinos [8]). In Section 4, we obtain several alternative forms of the main result, clarifying in the process the connection between our main result and that of Fakinos. Finally, in Section 5, we indicate how to extend our results to settings with considerably weaker assumptions than that of $GI/G/1$, such as state-dependent interarrival and service times, (possibly) dependent interarrival times, or (possibly) dependent service times.

2 Notation and Assumptions

For $i = 1, 2, \cdots$, let T_i be the interarrival time between customers C_i and C_{i+1}, and S_i be the service time of customer C_i. We assume, without loss of generality, that the first customer arrives at time 0 finding an empty system, and that, unless explicitly stated otherwise, customers are served according to the preemptive-resume LIFO queue discipline. In addition, we make the following assumptions about the interarrival and service times:

(a) $\{T_i, i \geq 1\}$ and $\{S_i, i \geq 1\}$ are two sequences of iid (independent and identically distributed) random variables that are also independent of one another.

(b) The arrival rate $\lambda \equiv 1/E(T)$ $(0 < \lambda < \infty)$ is less than the service rate $\mu \equiv 1/E(S)$ $(0 < \mu < \infty)$, where T and S denote typical versions of interarrival and service times respectively.

Let $L(t)$ be the number of customers in the system at time t; and when $L(t) > 0$, let $\mathbf{R}(t) \equiv (R_1(t), R_2(t), \cdots, R_{L(t)}(t))$ be the remaining service times of these customers

arranged in increasing order of their arrival times, i.e., $R_{L(t)}$ is of the most recent arrival. Define

$$\mathbf{Z}(t) \equiv \begin{cases} 0\,, & \text{if } \; L(t) = 0\,, \\ \\ \mathbf{R}(t)\,, & \text{if } \; L(t) > 0\,. \end{cases}$$

Then, the process $\mathbf{Z} \equiv \{\mathbf{Z}(t),\, t \geq 0\}$ describes the evolution of the state of the system. To simplify later expressions, we shall adopt, for convenience, the convention that the process $\mathbf{Z}$ is left-continuous and right-continuous, respectively, at arrival and departure epochs. Also note that $\mathbf{Z}$ is, in general, not a Markov process.

For $j \geq 1$ and $x_1,\, x_2,\, \cdots,\, x_j \geq 0$, denote by $\{j;\, x_1, x_2, \cdots, x_j\}$ the set $(x_1, \infty) \times (x_2, \infty) \times \cdots \times (x_j, \infty)$ in the j-dimensional Euclidean space, and by $\mathbf{1}_{\{j;\, x_1, x_2, \cdots, x_j\}}(\cdot)$ the indicator function of such a set. When the dimension is clear from the context, we also sometimes write $\mathbf{x}$ in place of the vector $(x_1, x_2, \cdots, x_j)$. With these definitions, we shall study the following averages (or limiting proportions):

$$\alpha_j(\mathbf{x}) \equiv \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{j;\, \mathbf{x}\}} \left(\mathbf{Z}\left(A_i\right) \right)\,, \tag{1}$$

$$\delta_j(\mathbf{x}) \equiv \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{j;\, \mathbf{x}\}} \left(\mathbf{Z}\left(D_i\right) \right)\,, \tag{2}$$

and

$$\tau_j(\mathbf{x}) \equiv \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{1}_{\{j;\, \mathbf{x}\}}(\mathbf{Z}(y))\, dy\,, \tag{3}$$

where A_i and D_i denote the i^{th} arrival and, respectively, departure epochs; note that these limits (and others defined in the remainder of this section) converge to constants w.p.1 (with probability 1) because of assumption (b). Thus, $\alpha_j(\mathbf{x})$ and $\delta_j(\mathbf{x})$ are the proportions of arrivals finding and, respectively, of departures leaving the system in state $\{j;\, \mathbf{x}\}$; and $\tau_j(\mathbf{x})$ is the proportion of time the system spends in that state.

For $j \geq 1$, let $\alpha_j \equiv \alpha_j(\mathbf{0})$, $\delta_j \equiv \delta_j(\mathbf{0})$, and $\tau_j \equiv \tau_j(\mathbf{0})$, where $\mathbf{0}$ denotes a vector of zeros. Also, similar to (1), (2), and (3), define α_0 and δ_0 to be the proportion of arrivals finding and, respectively, of departures leaving the system in the empty state, and τ_0 to be the proportion of time the system spends in state 0.

Clearly, the total work $W(t)$ in system under the preemptive-resume LIFO queue discipline is, for any $t \geq 0$, given by

$$W(t) = \sum_{i=1}^{L(t)} R_i(t)\,; \tag{4}$$

ill-defined sums are interpreted as 0 throughout. It follows by definition that (4) also gives the total work in system at time t for corresponding $GI/G/1$ queues with any work-conserving queue discipline. If the queue discipline is FIFO without preemption *and* if a customer arrives at time t, then $W(t)$ equals the delay to be experienced by this customer as well.

For $x \geq 0$, define

$$\nu(x, \infty) \equiv \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{1;\, x\}} \left(W \left(A_i \right) \right) \tag{5}$$

and

$$\nu^*(x, \infty) \equiv \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{1}_{\{1;\, x\}}(W(y))\, dy \,; \tag{6}$$

that is, $\nu(x, \infty)$ is the proportion of arrivals finding the total work in system exceeding x, and $\nu^*(x, \infty)$ is the corresponding proportion of time. By varying x, (5) and (6) generate two probability measures ν and ν^*, respectively, on the Borel subsets of the nonnegative real line. We then define arrival-average workload W and time-average workload W^* to be *random variables* whose distributions are determined by ν and ν^* respectively. Note that both W and W^* are finite w.p.1, because of assumption (b). Traditional analyses of workloads focus on the limits

$$\lim_{i \to \infty} P \left\{ W \left(A_i \right) > x \right\} \tag{7}$$

and

$$\lim_{t \to \infty} P \left\{ W(t) > x \right\} \,, \tag{8}$$

where $x \geq 0$. Whenever the limits (7) and (8) exist, they must agree with $P\{W > x\}$ and $P\{W^* > x\}$ (or $\nu(x, \infty)$ and $\nu^*(x, \infty)$), respectively; see, for example, proof of Corollary 1 in Niu [17]. It is well known that while the limit (7) always exists (independently of the initial state), the limit (8) in general need not, unless one imposes additional assumptions (a typical one being non-lattice interarrival-time distribution). By considering W^*, we avoid making such technical assumptions. More substantially, since W and W^* are defined through sample averages, they inherit or reflect, in a direct manner, structural properties of the model; and statistically speaking, they *represent* the aggregated experiences of all "actual" (W) and, respectively, "virtual" (W^*) customers.

Finally, let $K = \min\{k \geq 1 : \ \Sigma_{i=1}^{k} T_i > \Sigma_{i=1}^{k} S_i\}$, $B = \Sigma_{i=1}^{K} S_i$, and $I = \Sigma_{i=1}^{K} \left(T_i - S_i \right)$, so that K, B, and I are, respectively, the number of customers served in, the duration of, and the idle period that follows the initial busy period; note that these variables *jointly* take the same values under all work-conserving queue disciplines.

3 The Main Result and Its Immediate Implications

For $t \geq 0$, denote by $m_D(t)$ the renewal function for a "delayed" renewal process (see, for example, Ross [18], p. 74) whose first interevent time is distributed as T and the others as I. Also, let $x^+ \equiv \max(0, x)$ for any real number x. We are now ready for the statement of our main result.

Theorem 1 *For $j \geq 1$ and $\mathbf{x} \geq \mathbf{0}$,*

$$\alpha_j(\mathbf{x}) = \{1 - E[m_D(S)]\} \{E[m_D(S)]\}^j \prod_{i=1}^{j} \frac{E[m_D((S - x_i)^+)]}{E[m_D(S)]} \tag{9}$$

where $E[m_D(S)] < 1$; and $\alpha_0 = 1 - E[m_D(S)]$.

Theorem 1 has the following important interpretation: The proportion of arrivals finding the system in state $\{j; \mathbf{0}\}$, $j \geq 1$, is given by the "geometric form"

$$\alpha_0 (1 - \alpha_0)^j ; \tag{10}$$

and *among* those arrival epochs where the system is in state $\{j; \mathbf{0}\}$, the proportion of epochs where the remaining service times of the j customers are, respectively, greater than x_1, x_2, $\cdots$, and x_j (that is, the ratio $\alpha_j(\mathbf{x})/\alpha_j(\mathbf{0})$) is given by the "product form"

$$\prod_{i=1}^{j} [1 - \Psi(x_i)] , \tag{11}$$

where, for $x \geq 0$,

$$\Psi(x) \equiv 1 - \frac{E[m_D((S - x)^+)]}{E[m_D(S)]} . \tag{12}$$

By varying j and $\mathbf{x}$, we could then interpret (10) as the *probability* of a "typical" arrival finding j customers in system, and (11) as the *conditional probability* that the remaining service times of the j customers found by this arrival are greater than x_1, x_2, $\cdots$, and x_j, respectively.

The proof of Theorem 1 depends on a key sample-path identity, given as Lemma 1 next, whose proof is based directly on the property that the experience of any customer is not affected in any way by those who arrive before him. A basic concept needed for the proof of Lemma 1 is that of *j-cycles*, $j \geq 0$. A j-cycle is defined as a time period that begins with an arrival finding the system in state $\{j; \mathbf{0}\}$ and ends when such an event occurs again for the next time.

Lemma 1 *For $j \geq 1$ and $x_1, \cdots, x_{j-1}, x_j \geq 0$,*

$$\frac{\alpha_j(x_1, \cdots, x_{j-1}, x_j)}{\alpha_{j-1}(x_1, \cdots, x_{j-1})} = E[m_D((S - x_j)^+)]\,, \tag{13}$$

where $\alpha_{j-1}(x_1, \cdots, x_{j-1})$ is defined to be α_0 when $j = 1$.

Proof For a given $j \geq 1$, consider successive $(j-1)$-cycles. Let n_k, $k \geq 1$, be the index of the arrival epoch at which the k^{th} $(j-1)$-cycle begins. (Note that $\{A_{n_k},\, k \geq 1\}$, a subsequence of $\{A_i,\, i \geq 1\}$, diverges to infinity w.p.1, as $k \to \infty$.) Then,

$$\begin{aligned}\frac{\alpha_j(x_1, \cdots, x_{j-1}, x_j)}{\alpha_{j-1}(x_1, \cdots, x_{j-1})} &= \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mathbf{1}_{\{j;\, x_1, \cdots, x_{j-1}, x_j\}}(\boldsymbol{Z}(A_i))/n}{\sum_{i=1}^{n} \mathbf{1}_{\{j-1;\, x_1, \cdots, x_{j-1}\}}(\boldsymbol{Z}(A_i))/n} \\ &= \lim_{m\to\infty} \frac{\sum_{k=1}^{m} \sum_{\ell=n_k}^{n_{k+1}-1} \mathbf{1}_{\{j;\, x_1, \cdots, x_{j-1}, x_j\}}(\boldsymbol{Z}(A_\ell))}{\sum_{k=1}^{m} \mathbf{1}_{\{j-1;\, x_1, \cdots, x_{j-1}\}}(\boldsymbol{Z}(A_{n_k}))}\,,\end{aligned}$$

by first considering m $(j-1)$-cycles and then letting $m \to \infty$. Now, observe that

$$\mathbf{1}_{\{j-1;\, x_1, \cdots, x_{j-1}\}}(\mathbf{Z}(A_{n_k})) = 0 \Rightarrow \mathbf{1}_{\{j;\, x_1, \cdots, x_{j-1}, x_j\}}(\mathbf{Z}(A_\ell)) = 0$$

for all $n_k \leq \ell < n_{k+1}$, since the status of those customers who are present immediately before time A_{n_k} remains unchanged as long as there are j or more customers in the system. Therefore, by ignoring $(j-1)$-cycles with $\mathbf{1}_{\{j-1;\, x_1, \cdots, x_{j-1}\}}(\mathbf{Z}(A_{n_k})) = 0$, the last expression further simplifies to

$$\frac{\alpha_j(x_1, \cdots, x_{j-1}, x_j)}{\alpha_{j-1}(x_1, \cdots, x_{j-1})} = \lim_{n\to\infty} \sum_{i=1}^{n} N_i(x_j \mid j-1;\, x_1, \cdots, x_{j-1})\,, \tag{14}$$

where $N_i(x_j \mid j-1;\, x_1, \cdots, x_{j-1})$ denotes the number of arrivals who find the system in state $\{j;\, x_1, \cdots, x_{j-1}, x_j\}$ *during* the i^{th} $(j-1)$-cycle that starts in state $\{j-1;\, x_1, \cdots, x_{j-1}\}$.

We next carefully examine the variables $N_i(x_j \mid j-1;\, x_1, \cdots, x_{j-1})$, $i \geq 1$. If we follow the experience of a "test" customer whose arrival initiates a $(j-1)$-cycle, then his remaining service requirement, when considered as a function of time, decreases from S to 0 at a constant rate of 1, except possibly for interruptions caused by other customers who find him in service upon their arrival. Now under the preemptive-resume LIFO queue discipline, each arrival finding the "test" customer in service generates a busy cycle consisting of a busy and an idle period (stochastically identical to those in the original queue), and the remaining service requirement of the "test" customer stays unchanged during each of these busy periods; see Figure 1 for a typical realization. Therefore, after deleting intervals with zero slope in Figure 1, we see from Figure 2 that for *each* i, $N_i(x_j \mid j-1;\, x_1, \cdots, x_{j-1})$

is, *independently of the values* of $j-1$, x_1, $\cdots$, and x_{j-1}, distributed as the number of renewals in the random interval $(0, (S-x_j)^+)$ in a "delayed" renewal process where the first interevent time is distributed as T and the others as I; furthermore, these random variables are *independent* of one another. Hence, (14) implies (13), by the strong law of large numbers; and the proof is complete. □

*** Figures 1 and 2 about here. ***

Proof of Theorem 1 Apply Lemma 1 iteratively (similar to Fakinos [7], p. 246). First, we have for $j=1$ and $x_1 \geq 0$,

$$\alpha_1(x_1) = \alpha_0 E[m_D((S-x_1)^+)]\,;$$

and hence by induction

$$\alpha_j(\mathbf{x}) = \alpha_0 \prod_{i=1}^{j} E[m_D((S-x_i)^+)] \tag{15}$$

for all $j \geq 1$ and $\mathbf{x} \geq \mathbf{0}$.

To determine α_0, note that with $\mathbf{x} = \mathbf{0}$ in (15), we have $\alpha_j = \alpha_0\{E[m_D(S)]\}^j$ for all $j \geq 1$. Hence,

$$\alpha_0 = 1 - E[m_D(S)]\,, \tag{16}$$

with $E[m_D(S)] < 1$, by virtue of the normalization condition $\Sigma_{j=0}^{\infty}\alpha_j = 1$, which is due to assumption (b). Finally, (9) follows from (15) by first substituting (16) and then creating the factors $\{E[m_D(S)]\}^j$; the proof is thus complete. □

We now give several immediate consequences of Theorem 1.

Corollary 1 *Let U, J, and H_i, $i \geq 1$, be independent random variables defined, respectively, by*

$$U = \begin{cases} 1 & \textit{with probability } \ 1-\alpha_0\,, \\ \\ 0 & \textit{with probability } \ \alpha_0\,, \end{cases}$$

$P\{J=j\} = \alpha_0(1-\alpha_0)^{j-1}$ for $j \geq 1$, and for $i \geq 1$, $P\{H_i \leq x\} = \Psi(x)$ (see (12)), $x \geq 0$. Then,

$$W =^d U \sum_{i=1}^{J} H_i\,, \tag{17}$$

where $=^d$ denotes equality in distribution.

Proof Apply Theorem 1, noting the relation (4). □

Remark 1 The relation (17) also "represents" the arrival-average delay in the corresponding $GI/G/1$ queue with FIFO queue discipline.

Corollary 2 *If the arrival process is Poisson, then* $\alpha_0 = 1 - \rho$ *where* $\rho \equiv \lambda E(S)$, *and*

$$\alpha_j(\mathbf{x}) = (1-\rho)\rho^j \prod_{i=1}^{j} \frac{E[(S-x_i)^+]}{E(S)} \tag{18}$$

for $j \geq 1$ *and* $\mathbf{x} \geq \mathbf{0}$*; consequently, for* $x \geq 0$,

$$P\{W \leq x\} = \sum_{i=0}^{\infty} (1-\rho)\rho^i \, G_e^{[i]}(x) \ , \tag{19}$$

where $G_e^{[i]}(\cdot)$ *denotes the* i*-fold, self-convolution of the stationary excess service-time distribution* $G_e(\cdot)$ *defined, for* $x \geq 0$, *by*

$$G_e(x) = \frac{1}{E(S)} \int_0^x P\{S > t\} \, dt \, . \tag{20}$$

Proof It follows from the memoryless property of the exponential distribution that $T =^d I$, and hence $m_D(t) = \lambda t$ for all $t \geq 0$. Therefore, from (16), $\alpha_0 = 1 - \rho$; and (9) simplifies to (18). Finally, since $G_e(x) = 1 - E[(S-x)^+]/E(S)$, (19) is an immediate consequence of (17) and (18). □

Remark 2 (19) is the well-known Beneš's formula for the stationary delay distribution in the $M/G/1$ queue. Thus, Corollary 2 clarifies the mysterious form of the Beneš's formula, particularly the appearance of the stationary excess service-time distribution $G_e(\cdot)$.

Corollary 3 *If the service-time distribution is exponential, then, for* $j \geq 1$ *and* $\mathbf{x} \geq \mathbf{0}$,

$$\alpha_j(\mathbf{x}) = \alpha_0(1-\alpha_0)^j \prod_{i=1}^{j} P\{S > x_i\} \, . \tag{21}$$

Proof Observe that, since $[(S-x)^+ \mid S > x] =^d S$ when S is exponentially distributed,

$$\begin{aligned} E[m_D((S-x)^+)] &= P\{S > x\} \, E[m_D((S-x)^+ \mid S > x)] \\ &= P\{S > x\} E[m_D(S)] \, . \end{aligned}$$

Therefore, (21) follows immediately from (9). □

The next proposition shows that explicit results for the departure averages $\delta_j(\cdot)$ and for the time averages $\tau_j(\cdot)$, $j \geq 1$, follow as direct consequences of Theorem 1.

Proposition 1 *(i) For $j \geq 1$ and $\mathbf{x} \geq \mathbf{0}$, $\delta_j(\mathbf{x}) = \alpha_j(\mathbf{x})$; and $\delta_0 = \alpha_0$. (ii) For $j \geq 1$ and $\mathbf{x} \geq \mathbf{0}$,*

$$\tau_j(\mathbf{x}) = \tau_j \left\{ \prod_{i=1}^{j-1} [1 - \Psi(x_i)] \right\} [1 - G_e(x_j)] \tag{22}$$

(ill-defined products are interpreted as 1), where $\tau_j = \rho \alpha_{j-1}$; and $\tau_0 = 1 - \rho$.

Proof (i) Since each departing customer leaves the system in the same state as found by him when he arrived, the result follows directly from definition (see (1) and (2)).

(ii) The idea is to apply the well-known queueing relation "$H = \lambda G$" (see, for example, Heyman and Stidham [14]) with "H" and "G", respectively, being interpreted as appropriately defined time-average and arrival-average "cost" as follows: For fixed $j \geq 1$ and $\mathbf{x} \geq \mathbf{0}$, let each customer contribute to "cost" at a rate of 1 per unit time whenever he is in service *and* the system is in state $\{j; \mathbf{x}\}$, and of 0 otherwise. Note that a customer contributes to "cost" only if he finds upon arrival the system in state $\{j-1; x_1, \cdots, x_{j-1}\}$.

Now, consider a "test" customer who, upon arrival, finds the system in state $\{j-1; x_1, \cdots, x_{j-1}\}$. Then, it is easily seen (cf. Figure 2) that under the preemptive-resume LIFO queue discipline, his total contribution to "cost" is equal to $(S - x_j)^+$, where S denotes his service time. Next, note that $\tau_j(\mathbf{x})$ is by definition (see (3)) the time-average "cost", and that the arrival rate of such "test" customers is equal to $\lambda\ \alpha_{j-1}(x_1, \cdots, x_{j-1})$. Hence, by "$H = \lambda G$" (it is easily seen that the stated conditions in Theorem 1 of Heyman and Stidham [14], p. 985, are satisfied w.p.1, since our model is regenerative with finite expected length of regeneration cycles),

$$\begin{aligned} \tau_j(x_1, \cdots, x_{j-1}, x_j) &= \lambda\ \alpha_{j-1}(x_1, \cdots, x_{j-1})\ \lim_{n \to \infty} (1/n) \sum_{i=1}^{n} (S_i - x_j)^+ \\ &= \lambda\ \alpha_{j-1}(x_1, \cdots, x_{j-1})\ E[(S - x_j)^+] \\ &= \lambda\ E(S)\ \alpha_{j-1}(x_1, \cdots, x_{j-1})\ \frac{E[(S - x_j)^+]}{E(S)}\ , \end{aligned} \tag{23}$$

where the second equality is due to the strong law of large numbers. In particular, by letting $\mathbf{x} = \mathbf{0}$ in (23), we have

$$\tau_j = \rho\, \alpha_{j-1} \tag{24}$$

for $j \geq 1$, and hence also $\tau_0 = 1 - \rho$. Finally, substituting first (9) and then (24) into the right-hand side of (23) yields (22); and the proof is complete. □

Remark 3 Part (i) of Proposition 1 was obtained first by Yamazaki [24] and later by Fakinos [7], using different methods; part (ii) appears to be new. The argument leading up to (23) is closely related to one given by Shanthikumar and Sumita [20] to prove their

Theorem 2.4, of which (23) is a generalization; while the present paper was under review, the author became aware (personal communication) of concurrent related work by Ghahramani and Wolff [10], who also use a similar argument to establish relation (26) below.

Corollary 4 *Let W, W^*, S_e, and V be independent random variables, where W and W^* are defined in Section 2, S_e follows the distribution given by (20), and V is Bernoulli with success probability ρ. Then,*

$$W^* =^d V\,(W + S_e)\ . \tag{25}$$

Proof From (22), (24), and (17), we see that for all $t \geq 0$,

$$\begin{aligned} P\{W^* > t\} &= \sum_{j=1}^{\infty} \tau_j\ P\left\{\sum_{i=1}^{j-1} H_i + S_e > t\right\} \\ &= \rho \sum_{j=1}^{\infty} \alpha_{j-1}\ P\left\{\sum_{i=1}^{j-1} H_i + S_e > t\right\} \qquad (26) \\ &= \rho\ P\left\{U \sum_{i=1}^{J} H_i + S_e > t\right\} \\ &= \rho\ P\{W + S_e > t\}\ , \end{aligned}$$

implying (25). This completes the proof. □

Remark 4 The relation (26) is well known and is originally due to Takács [22]; alternative derivations for it have since been given by other authors, particularly Harrison and Lemoine [13] and Cohen [3]. The derivation here is new (see also Ghahramani and Wolff [10], and Remark 3) and shows that (26) is really a relation among sample averages *under* the preemptive-resume LIFO queue discipline.

4 Alternative Representations

The usefulness of Theorem 1 critically depends on our ability to analyze $m_D(t)$, which in turn is related to the idle period. For the case of Poisson arrivals, we are, of course, lucky. In general, the explicit distribution of I is unknown; and even if we knew the distribution of I, evaluating $m_D(t)$ is still by no means an obvious task. In this section, we give several alternative representations for the arrival averages $\alpha_j(\cdot)$, $j \geq 1$, with each offering some additional insight.

Theorem 2 *For $j \geq 1$ and $\mathbf{x} \geq \mathbf{0}$,*

$$\alpha_j(\mathbf{x}) = \left[\frac{1}{E(K)}\right]\left[1 - \frac{1}{E(K)}\right]^j \prod_{i=1}^{j} \frac{E[K((S - x_i)^+)] - 1}{E(K) - 1}\,, \tag{27}$$

where $K((S-x)^+)$, $x \geq 0$, denotes the number of customers served in a $GI/G/1$ busy period with the (exceptional) first service time distributed as $(S-x)^+$, $x \geq 0$, and $K \equiv K(S)$; and $\alpha_0 = 1/E(K)$.

Proof Consider a $(j-1)$-cycle for some $j \geq 1$. Since the distribution of $N_i(x_j \mid j-1; x_1, \cdots, x_{j-1})$ as defined in the proof of Lemma 1 depends only on the *value* of x_j, we shall denote a typical version of such a random variable by $N(x)$, $x \geq 0$. In fact, we can and shall assume, without loss of generality, that $j = 1$. Under this assumption, we could interpret the time interval that begins with the arrival of a "test" customer finding the system empty and ends when this customer has expended $(S-x)^+$ amount of time *in service* as a busy period with the first service time distributed as $(S-x)^+$. (Figures 1 and 2 are again helpful here.)

Now, observe that each arrival finding the "test" customer in service generates a pair of *dependent* random variables K and I, and that the sequence of random vectors (K_i, I_i) for $i \geq 1$ generated this way is iid. It follows that

$$K\left((S-x)^+\right) =^d 1 + \sum_{i=1}^{N(x)} K_i\,, \tag{28}$$

where the constant 1 accounts for the "test" customer, and $N(x)$ is explicitly defined by

$$N(x) = \begin{cases} 0 & \text{if } T > (S-x)^+\,, \\ 1 + \max\left\{n \geq 0 : \sum_{i=1}^{n} I_i \leq (S-x)^+ - T\right\}\,, & \text{if } T \leq (S-x)^+\,. \end{cases} \tag{29}$$

Next, for any realizations of T and S for which $T < (S-x)^+$, the random variable $N(x)$, being positive, is a stopping time relative to the sequence (K_i, I_i), for $i \geq 1$. Therefore, by Wald's identity, we obtain from (28), for $x \geq 0$,

$$\begin{aligned} E[K((S-x)^+)] &= 1 + P\{T < (S-x)^+\}\; E\left[\sum_{i=1}^{N(x)} K_i \mid T < (S-x)^+\right] \\ &= 1 + P\{N(x) > 0\}\; E[N(x) \mid N(x) > 0]\; E(K) \\ &= 1 + E[N(x)]\; E(K) \end{aligned} \tag{30}$$

implying that

$$E\left[N\left(x\right)\right] = \frac{E[K((S-x)^+)] - 1}{E(K)}\,, \tag{31}$$

which is, by definition, also equal to $E[m_D((S-x)^+)]$. Substituting $x = 0$ into (31), we also have

$$E[m_D(S)] = E[N(0)] = \frac{E(K) - 1}{E(K)}\,, \tag{32}$$

which, together with (16), implies that $\alpha_0 = 1/E(K)$. Finally, substituting (31) and (32) into (12) reduces (9) to (27); and the proof is complete. □

Remark 5 That $\alpha_0 = 1/E(K)$ is expected: Consider successive busy cycles (or 0-cycles) consisting of K_i customers each, $i \geq 1$, and observe that exactly one customer finds the system in state 0 in each busy cycle. Hence, it follows by first considering n busy cycles and then letting $n \to \infty$ that, w.p.1,

$$\alpha_0 = \lim_{n\to\infty} \frac{n}{\sum_{i=1}^{n} K_i} = \frac{1}{\lim_{n\to\infty} \sum_{i=1}^{n} K_i/n} = \frac{1}{E(K)}\,,$$

where the last equality is due to the strong law of large numbers.

Consider a busy period, and let $M(x)$ be the number of "blocked" customers (that is, exclude the customer who initiates the busy period) who, upon arrival, find the remaining service time of the customer in service exceeding x, $x \geq 0$. (The distribution of $M(x)$ is, in general, queue discipline dependent.) The next result complements Theorem 2.

Theorem 3 *For $x \geq 0$,*

$$\frac{E[K((S-x)^+)] - 1}{E(K) - 1} = \frac{E\left[M\left(x\right)\right]}{E[M(0)]}\,, \tag{33}$$

where the right-hand side can be interpreted (by a standard "regenerative process" argument) as the long-run proportion of "blocked" customers who, upon arrival, find the remaining service time of the customer in service being greater than x.

Proof Consider a busy period that is initiated by the arrival of a "test" customer finding the system empty, and let $N(x)$, $x \geq 0$, be as defined in (29). Observe that each arrival finding the "test" customer in service generates a busy period, and that *inside* each of such busy periods, the number of "blocked" customers who, upon arrival, find the remaining service time of the customer in service being greater than x has the same distribution as $M(x)$. It follows that

$$M(x) =^d N(x) + \sum_{i=1}^{N(0)} M_i(x)\,,$$

where $M_i(x)$ for $i \geq 1$ are iid versions of $M(x)$. ($N(x)$ and $N(0)$ are dependent, being equal to the respective numbers of renewals in the nested intervals $(0,\ (S-x)^+)$ and $(0,\ S)$.) An argument similar to that leading to (30) then yields $E[M(x)] = E[N(x)] + E[N(0)]E[M(x)]$, implying that

$$E\left[M\left(x\right)\right] = \frac{E\left[N\left(x\right)\right]}{1 - E[N(0)]}\,, \tag{34}$$

and in particular,

$$E[M(0)] = \frac{E[N(0)]}{1 - E[N(0)]}\,. \tag{35}$$

Dividing (34) by (35) yields (33) (see (31) and (32)); and the proof is complete. □

For $i \geq 1$, let $X_i = S_i - T_i$; and define a random walk $\{Y_i,\, i \geq 0\}$ by letting $Y_0 = 0$, and, for $n \geq 1$, $Y_n = \Sigma_{i=1}^{n} X_i$. Also, denote by σ the probability that a strict ascending ladder epoch (see Feller [9], Chapter XII) occurs in this random walk. ($\sigma < 1$ if and only if $E(X_1) < 0$.) Then, it is well known that the arrival-average workload W has the representation

$$W =^d \hat{U} \sum_{i=1}^{\hat{J}} \hat{H}_i\,, \tag{36}$$

where $P\{\hat{J} = j\} = (1-\sigma)\sigma^{j-1}$ for $j \geq 1$, $\hat{U}$ is Bernoulli with success probability σ, and $\hat{H}_i$'s denote iid successive ladder heights that occur. The next result identifies (36) with (17).

Lemma 2 *(i) $U =^d \hat{U}$, (ii) $J =^d \hat{J}$, and (iii) $H_1 =^d \hat{H}_1$.*

Proof It is easily shown by a standard duality argument that $E(K)$ is equal to 1 plus the expected number of strict ascending ladder epochs in the random walk $\{Y_i,\, i \geq 0\}$; see, for example, Ross [18], p. 222. Thus, $E(K) = 1/(1-\sigma)$; but from Theorem 2 (or Remark 5), we also have $E(K) = 1/\alpha_0$. It follows that

$$\sigma = 1 - \alpha_0\,, \tag{37}$$

implying (i) and (ii). Comparison of (17) and (36) then shows that we must also have $H_1 =^d \hat{H}_1$, which is (iii), completing the proof. □

Remark 6 For the $GI/M/1$ queue, the probability $1 - \alpha_0$ can be determined through the classical embedded Markov chain analysis as the unique solution in the interval (0, 1) of the functional equation

$$\omega = \tilde{F}[(1-\mu)\omega]\,, \tag{38}$$

where $\tilde{F}(\cdot)$ is the Laplace-Stieltjes transform of the interarrival-time distribution F. Niu and Cooper [16] observe that the solution of (38) can also be interpreted as the probability of occurrence of a strict descending ladder epoch in the corresponding dual $M/G/1$ queue; here we see that this connection is due to the more general relation (37).

Denote by $\hat{I}$ the idle period in the corresponding dual $GI/G/1$ queue. (The idle period $\hat{I}$ may never occur, since the dual queue is unstable.) Then, Theorem 1, Lemma 2, and

the well-known observation that $\hat{I} =^d \hat{H}_1$ (see, for example, Kleinrock [15], p. 311) together yield the following result due to Fakinos:

Theorem 4 *For $j \geq 1$ and $\mathbf{x} \geq \mathbf{0}$,*

$$\alpha_j(\mathbf{x}) = (1-\sigma)\sigma^j \prod_{i=1}^{j} P\{\hat{I}_i > x_i\},$$

where $\hat{I}_i$'s are iid versions of $\hat{I}$; and $\alpha_0 = 1 - \sigma$.

5 Extensions

In order to preserve the simplicity of the arguments, we have chosen to work within the standard $GI/G/1$ framework. Most of our results remain valid under assumptions weaker than those given in (a) and (b). This is due to an important characteristic of sample-average based arguments: they bring out the underlying explanation of why the obtained results should hold true, and thus clarify what is essential. To indicate how to generalize our results, we give in this section several examples; these examples are for illustrative purposes only and serve as maps for generalizations, possibly substantial, in other directions.

(5.1) A critical step in the proof of Lemma 1 is that independently of the values of $j-1$, $x_1, \cdots$, and x_{j-1}, the limit on the right-hand side of (14) converges, w.p.1, to an expected value. The strong law of large numbers is applicable there because of assumption (a). This assumption can be replaced by any other set of assumptions which guarantees convergence in (14), possibly by invoking or proving other versions of strong laws. One simple example is to assume that:

(c) The sequence $\{(T_i, S_i), i \geq 1\}$ is iid.

That is, T_i and S_i for the same i are possibly dependent. Under assumption (c), the idle period I, being equal to the first strict descending ladder height of the random walk $\{Y_i, i \geq 0\}$, is still well defined. The right-hand side of (13), however, needs to be replaced by $E[N(x)]$, where $N(x)$ is as defined in (29).

With the exceptions of Corollaries 2 and 3, all other results in Sections 3 and 4 remain valid under assumption (c), after similar modifications.

(5.2) A further generalization is to let (T_i, S_i) be sampled from a possibly different joint distribution, depending on the number of customers present at an arrival epoch. More specifically, at the beginning of the i^{th} j-cycle ($i \geq 1$ and $j \geq 0$), let T_{ji} be the time to next arrival, and let S_{ji} be the service time of the arriving customer; and replace assumption (a) by

(d) The sequence $\{(T_{ji},\, S_{ji}),\, i \geq 1\}$ is iid for each $j \geq 0$, with the sequences for different j being independent.

(As in Section 5.1, T_{ji} and S_{ji} for the same j and i are possibly dependent.) We also assume, in place of (b), that the interarrival times and service times are such that the averages defined in Section 2 converge, w.p.1, to constants—our attitude, as in Wolff [23] and in Niu [17], is to obtain relations that hold *whenever* convergences occur.

Under assumption (d), (29) needs to be replaced by

$$N_j\,(x) = \begin{cases} 0 & \text{if } T_{j1} \geq (S_{j1} - x)^+, \\ 1 + \max\left\{n \geq 0 :\, \sum_{k=1}^{n} I_{jk} \leq (S_{j1} - x)^+ - T_{j1}\right\}, & \text{if } T_{j1} < (S_{j1} - x)^+, \end{cases}$$

where I_{jk} for $k \geq 1$ denote iid "idle periods" (defined in obvious ways) generated by arrivals finding the system in state $\{j;\, \mathbf{0}\}$, $j \geq 1$. Then it is easy to see that the following extension of Theorem 1 holds:

Theorem 5 *For $j \geq 1$ and $\mathbf{x} \geq \mathbf{0}$,*

$$\alpha_j\,(\mathbf{x}) = \left\{1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} E[N_i(0)]\right\}^{-1} \left\{\prod_{i=1}^{j} E[N_i(0)]\right\} \prod_{i=1}^{j} \frac{E[N_i(x_i)]}{E[N_i(0)]}\,; \tag{39}$$

and

$$\alpha_0 = \left\{1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} E[N_i(0)]\right\}^{-1}.$$

All other results in Sections 3 and 4 also admit similar extensions, with varying degrees of generality. We omit the details.

It is interesting to note the similarity of form between (39) (with $\mathbf{x} = \mathbf{0}$) and the standard solution of the set of balance equations for a birth-and-death process with state-dependent birth and death rates given, for example, in Cooper [4], p. 22, eq. (3.10); in fact, if we let the arrival rate be dependent on $L(t)$ for *all* t, then the proof of Theorem 1 can also be adapted to show, constructively, that the latter is valid even when the service times are not necessarily exponentially distributed.

(5.3) Being a consequence of "$H = \lambda G$", the first equation in (23), namely

$$\tau_j(x_1, \cdots, x_{j-1}, x_j) = \lambda\; \alpha_{j-1}(x_1, \cdots, x_{j-1})\; \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (S_i - x_j)^+, \tag{40}$$

is a *deterministic* relation, holding on each sample path for which the limits

$$\tau_j(x_1, \cdots, x_{j-1}, x_j)\,,$$

$$\lambda \equiv \lim_{t\to\infty} \frac{A\,(t)}{t}\,,$$

where $A\,(t) \equiv \max\{i \geq 1 :\ A_i \leq t\}$,

$$\alpha_{j-1}(x_1, \cdots, x_{j-1})\,,$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (S_i - x_j)^+$$

are well defined. In the $GI/G/1$ framework, these limits converge, w.p.1, to respective constants because of assumptions (a) and (b). In general, it is not critical (for the validity of (40)) that the sequences $\{T_i,\, i \geq 1\}$ and $\{S_i,\, i \geq 1\}$ be iid and that they be independent of one another; it is for the existence and for the *evaluation* (e.g., (9)) of the relevant limits that such probabilistic assumptions are imposed. For careful discussions of convergence issues related to "$H = \lambda G$", we refer the reader to Heyman and Stidham [14] and to Glynn and Whitt [11].

When the arrival process is Poisson *and* is independent of $\{S_i,\, i \geq 1\}$, where S_i's are *not* necessarily iid, (40) can be used to give a "direct" (i.e., without having to prove Lemma 1 first) proof of a more general version of (18): For $j \geq 1$ and $\mathbf{x} \geq \mathbf{0}$,

$$\alpha_j\,(\mathbf{x}) = \left[1 - \lambda \bar{S}(0)\right] \left[\lambda \bar{S}(0)\right]^j \prod_{i=1}^{j} \frac{\bar{S}(x_i)}{\bar{S}(0)}\,, \tag{41}$$

where $\bar{S}(x) \equiv \lim_{n\to\infty} \Sigma_{i=1}^{n} (S_i - x)^+/n$, $x \geq 0$ (assuming all relevant limits converge, w.p.1, to constants); when S_i's are iid, (41) specializes to (18), which, of course, in turn implies the Beneš's formula (19). The argument is as follows: First, note that for $j \geq 1$ and $\mathbf{x} \geq \mathbf{0}$,

$$\tau_j(x_1, \cdots, x_{j-1}, x_j) = \alpha_j(x_1, \cdots, x_{j-1}, x_j)\,, \tag{42}$$

by the property "Poisson arrivals see time averages"—Wolff [23] (notice that his "lack of anticipation" assumption holds for the situation here); next, substitute (42) into the left-hand side of (40) to get

$$\alpha_j(x_1, \cdots, x_{j-1}, x_j) = \lambda\ \alpha_{j-1}(x_1, \cdots, x_{j-1})\ \bar{S}(x_j)\,; \tag{43}$$

finally, apply (43) iteratively as in the proof of Theorem 1 to get (41), completing the argument.

(5.4) The validity of Corollary 4 does not critically depend on assumptions (a) and (b). (A similar observation is also noted in Ghahramani and Wolff [10]; see Remark 3.) The idea is to work directly with $\{W(t),\, t \geq 0\}$ and to preserve the independence of W and S_e. We could, for example, replace assumption (a) by the following:

(e) The sequence $\{S_i,\, i \geq 1\}$ is iid, with a finite service rate $\mu \equiv 1/E(S)$.

(f) The sequences $\{T_i,\, i \geq 1\}$ and $\{S_i,\, i \geq 1\}$, where T_i's are *not* necessarily iid, are independent of one another.

In place of (b), we assume that:

(g) The interarrival and service times are such that the embedded process $\{W(A_i),\, i \geq 1\}$ has a limiting distribution.

Again, we apply the relation "$H = \lambda G$". For a fixed $x \geq 0$, let each customer contribute to "cost" at a rate of 1 per unit time whenever he is in service *and* the total work in system exceeds x, and of 0 otherwise.

Under the preemptive-resume LIFO queue discipline, we claim that the total contribution to "cost" from customer C_i is given by

$$[S_i - (x - W(A_i))^+]^+ . \tag{44}$$

To prove this claim, first imagine stacking, at successive arrival epochs, pieces of "new" work on *top* of "old" work, and then interpret (44) as the contribution from customer C_i to the amount, if any, by which the total work immediately after time A_i exceeds x. Since work in system is worked off from the top of the stack, we see from Figure 3 that this interpretation implies that (44) also equals the total amount of time during which customer C_i is in service and the total work in system exceeds x; and the claim is proved.

*** Figure 3 about here. ***

Now, if the arrival rate is, w.p.1, equal to λ, $0 < \lambda < \infty$, then from "$H = \lambda G$",

$$\nu^*(x, \infty) = \lambda \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} [S_i - (x - W(A_i))^+]^+, \tag{45}$$

assuming that the (explicit) limit on the right-hand side converges, w.p.1, to a constant. We shall complete the argument by showing that under assumptions (e), (f), and (g), this constant can be evaluated to $E(S)P\{W + S_e > x\}$, and hence (26) follows from (45).

First, observe that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} [S_i - (x - W(A_i))^+]^+ \quad = \quad E\left\{ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} [S_i - (x - W(A_i))^+]^+ \right\}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} E\left\{\frac{1}{n}\sum_{i=1}^{n}[S_i - (x - W(A_i))^+]^+\right\} \\
&= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E[S_i - (x - W(A_i))^+]^+ \\
&= E[S - (x - W)^+]^+ ,
\end{aligned}
$$

where the second equality is due to Lebesgue's dominated convergence Theorem (see, e.g., Royden [19], p. 89, Theorem 16; note that $[S_i - (x - W(A_i))^+]^+ \leq S_i$ w.p.1, and that, by assumption (e), $\lim_{n\to\infty} \sum_{i=1}^{n} S_i/n \to 1/\mu < \infty$), and the fourth due to assumptions (e) and (g). Next, note that by assumption (f), W and S are independent; and hence

$$
\begin{aligned}
E[S - (x - W)^+]^+ &= \int_0^\infty E[S - (x - u)^+]^+ \, dP\{W \leq u\} \\
&= E(S) \int_0^\infty P\{S_e > (x - u)^+\} \, dP\{W \leq u\} \quad \text{(see (20))} \\
&= E(S)\, P\{W + S_e > x\} ,
\end{aligned}
$$

completing the argument.

References

[1] V. E. Beneš's, "On Queues with Poisson Arrivals," *The Annals of Mathematical Statistics*, 28 (1957), pp. 670–677.

[2] S. L. Brumelle, "On the Relation Between Customer and Time Averages in Queues," *Journal of Applied Probability*, 8 (1971), pp. 508–520.

[3] J. W. Cohen, "On Up- and Downcrossings," *Journal of Applied Probability*, 14 (1977), pp. 405–410.

[4] R. B. Cooper, *Introduction to Queueing Theory*, Second Edition, North Holland (Elsevier), New York, 1981.

[5] R. B. Cooper and S.-C. Niu, "Beneš's Formula for $M/G/1$-FIFO 'Explained' by Preemptive-Resume LIFO," *Journal of Applied Probability*, 23 (1986), pp. 550–554.

[6] D. Fakinos, "The $G/G/1$ Queueing System with a Particular Queue Discipline," *Journal of the Royal Statistical Society*, B 43 (1981), pp. 190–196.

[7] D. Fakinos, "On the Single-Server Queue with the Preemptive-Resume Last-Come-First-Served Queue Discipline," *Journal of Applied Probability*, 23 (1986), pp. 243–248.

[8] D. Fakinos, "The Single-Server Queue with Service Depending on Queue Size and with the Preemptive-Resume Last-Come-First-Served Queue Discipline," *Journal of Applied Probability*, 24 (1987), pp. 758–767.

[9] W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. II*, Second Edition, Wiley, New York, 1971.

[10] S. Ghahramani and R. W. Wolff, "Finite Moment Conditions for $GI/G/1$ Busy Periods," preprint (1986).

[11] P. W. Glynn and W. Whitt, "Extensions of the Queueing Relations $L = \lambda W$ and $H = \lambda G$," preprint (1986).

[12] F. P. Kelly, "The Departure Process from a Queueing System," *Math. Proc. Camb. Phil. Soc.*, 80 (1976), pp. 283–285.

[13] J. M. Harrison and A. J. Lemoine, "On the Virtual and Actual Waiting Time Distributions of a $GI/G/1$ Queue," *Journal of Applied Probability*, 13 (1976), pp. 833–836.

[14] D. P. Heyman and S. Stidham, Jr, "The Relation Between Customer and Time Averages in Queues," *Operations Research*, 28 (1980), pp. 983–994.

[15] L. Kleinrock, *Queueing Systems, Vol. I: Theory*, Wiley, New York, 1975.

[16] S.-C. Niu and R. B. Cooper, "Duality and Other Results for $M/G/1$ and $GI/M/1$ Queues, via a New Ballot Theorem," *Mathematics of Operations Research*, 14 (1989), pp. 281–293.

[17] S.-C. Niu, "Inequalities Between Arrival Averages and Time Averages in Stochastic Processes Arising from Queueing Theory," *Operations Research*, 32 (1984), pp. 785–795.

[18] S. M. Ross, *Stochastic Processes*, Wiley, New York, 1983.

[19] H. L. Royden, *Real Analysis*, Second Edition, Macmillan, New York, 1972.

[20] J. G. Shanthikumar and U. Sumita, "On $G/G/1$ Queues with LIFO-P Service Discipline," *Journal of the Operations Research Society of Japan*, 29 (1986), pp. 220–231.

[21] S. Stidham, Jr, "A Last Word on $L = \lambda W$," *Operations Research*, 22 (1974), pp. 417–421.

[22] L. Takács, "The Limiting Distribution of the Virtual Waiting Time and the Queue Size for a Single-Server Queue with Recurrent Input and General Service Times," *Sankhya*, A 25 (1963), pp. 91–100.

[23] R. W. Wolff, "Poisson Arrivals See Time Averages," *Operations Research*, 30 (1982), pp. 223–231.

[24] G. Yamazaki, "The $GI/G/1$ Queue with Last-Come-First-Served," *Ann. Inst. Statis. Math.*, A 34 (1982), pp. 599–604.

[25] G. Yamazaki, "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 (1984), pp. 338–346.