J. Appl. Prob. **42,** 223–234 (2005)
Printed in Israel

SOME TIME-DEPENDENT PROPERTIES OF SYMMETRIC M/G/1 QUEUES

OFFER KELLA,* *The Hebrew University of Jerusalem*
BERT ZWART,** *** *Eindhoven University of Technology and CWI*
ONNO BOXMA,** **** *EURANDOM, Eindhoven University of Technology and CWI*

Abstract

We consider an M/G/1 queue that is idle at time 0. The number of customers sampled at an independent exponential time is shown to have the same geometric distribution under the preemptive-resume last-in–first-out and the processor-sharing disciplines. Hence, the marginal distribution of the queue length at any time is identical for both disciplines. We then give a detailed analysis of the time until the first departure for *any* symmetric queueing discipline. We characterize its distribution and show that it is insensitive to the service discipline. Finally, we study the tail behavior of this distribution.

Keywords: Symmetric queue; time-dependent analysis; insensitivity; order statistic; random permutation; tail behavior

2000 Mathematics Subject Classification: Primary 60K25
Secondary 90B22

1. Introduction

One of the major success stories in applied probability has been the development of the theory of product-form queueing networks. Classic papers include [1], [4], [5], [8], and [10]; see [11] for a textbook treatment. Recent interesting papers are [3] and [15].

Important components of such product-form networks are M/G/1 queues operating under a *symmetric* queueing discipline. This class of disciplines, treated in Section 3.3 of [11], contains both the preemptive-resume last-in–first-out (LIFO) and the processor-sharing (PS) disciplines as special cases. A special feature of these symmetric queues is the fact that the steady-state distribution of the *queue length* (the number of customers in the system, including those in service) is geometric with probability of success $1 - \rho$, where $\rho < 1$ is the traffic intensity. In particular, the steady-state distribution of the queue length depends only on the mean of the service time and is otherwise insensitive to the service time distribution.

In this paper, a different approach to symmetric queues is taken. We focus on time-dependent, rather than steady-state, behavior, and also explore insensitivities with respect to the service discipline rather than the service time distribution.

We first investigate the queue length process $\{L(t), t \geq 0\}$ of the M/G/1 LIFO queue with $L(0) = 0$. Letting $\tau(q)$ be an independent exponential random variable with rate $q > 0$, we

Received 27 July 2004; revision received 15 September 2004.
* Postal address: Department of Statistics, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel. Email address: mskella@mscc.huji.ac.il. Supported in part by grant 819/03 from the Israel Science Foundation.
** Postal address: Department of Mathematics and Computer Science, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands.
*** Email address: zwart@win.tue.nl. Supported by an NWO VENI grant.
**** Email address: boxma@win.tue.nl. Work carried out within the Euro-NGI project.

show that $L(\tau(q))$ has a geometric distribution. We find this very pleasing, since exactly the same distribution was found earlier for the PS discipline [12]. This implies that, for any $t > 0$, $L(t)$ has the same distribution for both disciplines.

It would be a very nice result if the distribution of $L(t)$ were found to be the same for *all* symmetric disciplines. At present, however, this is beyond our reach and is left as an open question. Nevertheless, we do give a complete description of the distribution of the time D_1 until a first departure occurs. This distribution is shown to be insensitive to the particular symmetric discipline chosen. We prove this result by applying an insensitivity property of random permutations. As will become clear, the class of symmetric service disciplines is exactly the right class to consider in this setting.

The paper is organized as follows. Section 2 includes some preliminary notation and definitions. In Sections 3 and 4, the LIFO and PS queue length distributions are treated. In Section 5, we present a simple, but useful, insensitivity result for random permutations that is the basis for the analysis of Section 6. In Section 6, the Laplace-Stieltjes transform (LST) of D_1 is derived for an arbitrary symmetric queueing discipline. Section 7 is devoted to the distribution of D_1, which can be given in an explicit and intuitively appealing form. The tail behavior of the distribution of D_1 is derived in Section 8.

2. Preliminaries

We consider an M/G/1 queue. The Poisson process $\{N(t), t \geq 0\}$, with rate λ, represents the customer arrival process. The independent, identically distributed random variables B_i, $i \geq 1$, denote the service times of successively arriving customers, with distribution $B(\cdot)$. As usual, we write $\rho := \lambda \,\mathrm{E}\, B_1$. Let $\beta(\alpha) := \mathrm{E}\, \mathrm{e}^{-\alpha B_1}$ be the LST of B_1 for $\alpha \geq 0$, and define the net input process $X(t) = \sum_{i=1}^{N(t)} B_i - t$. This is a Lévy process with exponent $\phi(\alpha) = \alpha - \lambda(1 - \beta(\alpha))$, i.e. for $\mathrm{Re}\,\alpha \geq 0$,

$$\mathrm{E}\, \mathrm{e}^{-\alpha X(t)} = \mathrm{e}^{t\phi(\alpha)}.$$

Note that $\phi(\alpha)$ is strictly convex and continuous on $[0, \infty)$ and tends to infinity as $\alpha \to \infty$. In particular, $\phi(\alpha)$ is strictly increasing on the interval $[\alpha^*, \infty)$, where $\alpha^* = \inf\{\alpha : \phi(\alpha) > 0\}$. If $\rho \leq 1$ then $\alpha^* = 0$ and if $\rho > 1$ then $\alpha^* > 0$ since, in the first case, $\phi'(0) = 1 - \rho \geq 0$ and, in the second case, $\phi'(0) < 0$, where a prime denotes differentiation. Since ϕ is continuous and strictly increasing on $[\alpha^*, \infty)$, it has an inverse, which we denote by $\kappa(q)$, $q \geq 0$, when viewed as a function from $[\alpha^*, \infty)$ to $[0, \infty)$.

The service discipline is assumed to be *symmetric*. Recall (see Section 3.3 of [11]) that a symmetric queueing discipline is defined as follows. For each n, let $p_1^n, \ldots, p_n^n$ be nonnegative and sum to 1. If there are $n - 1$ customers in the system in positions $1, \ldots, n - 1$ upon the arrival of the kth customer, $k \geq n$, then this customer is put in position i with probability p_i^n. The customers who were in positions $1, \ldots, i - 1$ remain in their positions and the customers who were in positions $i, \ldots, n - 1$ move to positions $i + 1, \ldots, n$ respectively. After this repositioning, the customer in position j is allocated a service rate of p_j^n. Special cases of this discipline are the preemptive-resume LIFO discipline (for which $p_1^n = 1$) and the PS discipline (for which $p_i^n = 1/n, i = 1, \ldots, n$). The M/G/1 queue length process for these two disciplines is studied in the next two sections.

Throughout this paper, $\{L(t), t \geq 0\}$ denotes the queue length process (the number of customers in the system). When we want to distinguish between the queue length process of the LIFO and PS disciplines we will use the notation $L_{\mathrm{LIFO}}(t)$ and $L_{\mathrm{PS}}(t)$, respectively, but not otherwise.

https://doi.org/10.1239/jap/1110381382 Published online by Cambridge University Press

3. The LIFO queue length

In this section, we investigate the queue length process of the M/G/1 queue operating under the (preemptive-resume) LIFO discipline.

The first main step of our analysis is to observe that the queue length process $\{L(t), t \geq 0\}$ can be expressed, in terms of the net input process $\{X(t), t \geq 0\}$, as follows, where $X(s-) = \lim_{t \uparrow s} X(t)$. Henceforth, $\sharp\{s\colon S(s)\}$ denotes the number of s values for which statement $S(s)$ holds.

Lemma 1. *For any $t \geq 0$, we have*

$$L(t) = \sharp\Big\{s \in [0,t]\colon X(s-) = \inf_{r \in [s,t]} X(r)\Big\}.$$

This relation is explicitly stated in [13], where it is applied to derive a diffusion approximation for $\{L(t), t \geq 0\}$. It is also implicit in [17] and [18]. Furthermore, there is a close connection between LIFO queues and Galton–Watson processes: the process $\{L(t), t \geq 0\}$ can be seen as an encoding of a Galton–Watson tree. Such a connection also holds when the paths of $X(t)$ are almost surely of infinite variation. Then, a local-time analogue of $L(t)$, called the height process, can be used to encode the genealogy of a continuous-state branching process. We refer the reader to [7] for a recent study and the state of the art in this area.

We now give the main result of this section.

Theorem 1. *Let $\tau = \tau(q)$ be an independent, exponentially distributed random variable with rate $q > 0$. Then,*

$$\mathrm{P}[L(\tau(q)) = n] = \left(1 - \frac{q}{\kappa(q)}\right)^n \frac{q}{\kappa(q)}\,. \tag{1}$$

Proof. We apply Lemma 1, as follows. Let $X_t(s) = X(t) - X((t-s)-)$. Straightforward manipulations then show that

$$L(t) = \sharp\Big\{s \in [0,t]\colon X_t(s) = \sup_{r \in [0,s]} X_t(r)\Big\}.$$

Since $X(t)$ is reversible, we obtain

$$L(t) \stackrel{\mathrm{D}}{=} \sharp\Big\{s \in [0,t]\colon X(s) = \sup_{r \in [0,s]} X(r)\Big\}$$

for every $t > 0$, where '$\stackrel{\mathrm{D}}{=}$' denotes equality in distribution. From this, it follows that

$$L(\tau(q)) \stackrel{\mathrm{D}}{=} \sharp\Big\{s \in [0,\tau(q)]\colon X(s) = \sup_{r \in [0,s]} X(r)\Big\}.$$

Let τ_i, $i \geq 1$, denote the successive ladder epochs of the Lévy process $\{X(t), t \geq 0\}$. It is well known that $\{\tau_i, i \geq 1\}$ is a (possibly terminating) renewal process. It is clear that the number of renewals up to $\tau(q)$ (and, hence, also $L(\tau(q))$ must have a geometric distribution. Indeed, if r_i denote the renewal intervals, with $R_n = \sum_{i=1}^n r_i$, and if $K(t)$ denotes the number of renewals in $[0,t]$, then

$$\mathrm{P}[K(\tau(q)) \geq n] = \mathrm{P}[R_n \leq \tau(q)] = \mathrm{E}\,\mathrm{e}^{-qR_n} = (\mathrm{E}\,\mathrm{e}^{-qr_1})^n. \tag{2}$$

https://doi.org/10.1239/jap/1110381382 Published online by Cambridge University Press

To compute the probability of success for this geometric distribution, note that

$$\mathrm{P}[L(\tau(q)) = 0] = \mathrm{P}[\tau_1 > \tau(q)].$$

Let $S(t) = \sup_{0<s<t} X(s)$ and note that

$$\mathrm{P}[\tau_1 > \tau(q)] = \mathrm{P}[S(\tau(q)) = 0].$$

The LST of $S(\tau(q))$ is well known; see, e.g. Equation (3) on p. 192 of [2]: it is given by

$$\mathrm{E}\,\mathrm{e}^{-\alpha S(\tau(q))} = \frac{q(\kappa(q)-\alpha)}{\kappa(q)(q-\phi(\alpha))}.$$

Since $\phi(\alpha)/\alpha \to 1$ as $\alpha \to \infty$, we obtain

$$\mathrm{P}[S(\tau(q)) = 0] = \lim_{\alpha\to\infty} \mathrm{E}\,\mathrm{e}^{-\alpha S(\tau(q))} = \frac{q}{\kappa(q)}. \tag{3}$$

Equations (2) and (3) imply (1).

Remark 1. We note that $\mathrm{P}[L(\tau(q)) = 0 \mid L(0) = 0]$ is equivalent to the conditional probability that the workload is zero at time $\tau(q)$, starting from an empty system. Thus, this probability is $q/\kappa(q)$ for any work-conserving discipline and, in particular, for any symmetric discipline. An alternative derivation of this probability may be found on p. 260 of [6].

4. The PS queue length

Our starting point is the following formula, which is Equation (2.6) of [12]:

$$\int_0^\infty \mathrm{e}^{-qt}\,\mathrm{E}\,z^{L(t)}\mathrm{d}t = \frac{1}{q+(1-z)\lambda(1-\pi(q))}. \tag{4}$$

Here, $0 < z \le 1$ and $\pi(q)$ is the LST of the length of an M/G/1 busy period, i.e. $\pi(q)$ is the smallest root of the equation $\pi(q) = \beta(q+\lambda-\lambda\pi(q))$. However, we prefer to use the expression $\pi(q) = \beta(\kappa(q))$, which is easy to verify and can be found, for example, in [16]. From (4), observing that $\kappa(q) = q + \lambda(1-\pi(q))$, we obtain

$$\begin{aligned}\mathrm{E}\,z^{L(\tau(q))} &= \frac{q}{q+(1-z)\lambda(1-\pi(q))} \\ &= \frac{q}{\kappa(q)-z\lambda(1-\pi(q))} \\ &= \frac{q/\kappa(q)}{1-z(1-q/\kappa(q))}.\end{aligned}$$

This is the generating function of the right-hand side of (1), that is, of a geometric random variable with probability of success $q/\kappa(q)$. Thus, we arrive at the following interesting result.

Theorem 2. *Let* $L_{\mathrm{LIFO}}(0) = L_{\mathrm{PS}}(0) = 0$. *Then,*

$$L_{\mathrm{LIFO}}(\tau(q)) \overset{\mathrm{D}}{=} L_{\mathrm{PS}}(\tau(q)) \quad \textit{for all } q > 0 \tag{5}$$

and

$$L_{\mathrm{LIFO}}(t) \overset{\mathrm{D}}{=} L_{\mathrm{PS}}(t) \quad \textit{for all } t > 0. \tag{6}$$

https://doi.org/10.1239/jap/1110381382 Published online by Cambridge University Press

Proof. Equation (5) follows from Theorem 1 and the computations made above, and (6) follows from (5) by the uniqueness property of Laplace transforms, as sampling at an exponential time is equivalent to making a Laplace transform with respect to time.

Remark 2. Although, starting from an empty system, the queue length distribution at an exponential time is geometric for the LIFO and PS disciplines, the probability of success depends on the entire service time distribution. This is in contrast to the steady-state case, in which the distribution is also geometric but the probability of success is $1-\rho$ whenever $\rho < 1$, thus depending only on the mean.

5. An insensitivity property of random permutations

In the remainder of the paper, we focus on D_1, which is the time until a first departure occurs from an M/G/1 queue with an arbitrary symmetric service discipline. Our main results are given in the next two sections. In the present section, we derive a preliminary result, which could be of independent interest.

Lemma 2. *Let $\boldsymbol{U} = (U_1, \ldots, U_n)$ and $\boldsymbol{V} = (V_1, \ldots, V_n)$ be random variables and let $\boldsymbol{\Pi} = (\Pi_1, \ldots, \Pi_n)$ be a random permutation, such that the pair $(\boldsymbol{\Pi}, \boldsymbol{V})$ and $\boldsymbol{U}$ are independent and $U_1, \ldots, U_n$ are exchangeable. Then,*

$$\mathrm{P}[U_1 > V_{\Pi_1}, \ldots, U_n > V_{\Pi_n}] = \mathrm{P}[U_1 > V_1, \ldots, U_n > V_n].$$

When, in addition, $U_1, \ldots, U_n$ are independent and $V_1, \ldots, V_n$ are independent, identically distributed random variables, we have, in particular,

$$\mathrm{P}[U_1 > V_{\Pi_1}, \ldots, U_n > V_{\Pi_n}] = \mathrm{P}[U_1 > V_1]^n. \tag{7}$$

Proof. For any fixed permutation $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$,

$$\begin{aligned}\mathrm{P}[U_1 > V_{\Pi_1}, \ldots, U_n > V_{\Pi_n} \mid \boldsymbol{\Pi} = \boldsymbol{\pi}] &= \mathrm{P}[U_1 > V_{\pi_1}, \ldots, U_n > V_{\pi_n} \mid \boldsymbol{\Pi} = \boldsymbol{\pi}] \\ &= \mathrm{P}[U_{\pi_1} > V_{\pi_1}, \ldots, U_{\pi_n} > V_{\pi_n} \mid \boldsymbol{\Pi} = \boldsymbol{\pi}] \\ &= \mathrm{P}[U_1 > V_1, \ldots, U_n > V_n \mid \boldsymbol{\Pi} = \boldsymbol{\pi}].\end{aligned}$$

The second equality follows from the fact that $\boldsymbol{U}$ is independent of $(\boldsymbol{\Pi}, \boldsymbol{V})$ and because its components are exchangeable. Multiplying the left- and right-most expressions by $\mathrm{P}[\boldsymbol{\Pi} = \boldsymbol{\pi}]$ and summing over all possible permutations gives the result.

6. Insensitivity of the first departure time

Consider an M/G/1 queue with a symmetric queueing discipline, as described in Section 2. Let $\{Y_n, n \geq 1\}$ denote the interarrival times and recall that $\{B_n, n \geq 0\}$ are the service times. For the moment, we assume that customers arrive according to some Poisson process with rate 1 and never leave. The rate 1 is chosen without loss of generality and will later be replaced with another parameter. Thus, for now, $Y_n \sim \mathrm{Exp}(1)$ for $n \geq 1$.

We would like to show that the joint distribution of the times allocated to the first n customers up to the $(n+1)$th arrival epoch is identical to that of $Y_{\Pi_1}, \ldots, Y_{\Pi_n}$ for some random permutation $\boldsymbol{\Pi}$, which is a functional of $\boldsymbol{Y}$. If this can be achieved then, once we introduce the service times, $(\boldsymbol{\Pi}, \boldsymbol{Y})$ and $\boldsymbol{B}$ are independent and, thus, it follows from Lemma 2 that the probability that none of the first n arriving customers has departed by the $(n+1)$th arrival epoch is given by

$$\mathrm{P}[B_1 > Y_{\Pi_1}, \ldots, B_n > Y_{\Pi_n}] = \mathrm{P}[B_1 > Y_1]^n.$$

https://doi.org/10.1239/jap/1110381382 Published online by Cambridge University Press

This will allow us to study the distribution of the first departure time in a symmetric queue (see Theorems 3 and 4, below). For what follows, we define

$$\frac{y}{0} = \infty \quad \text{and} \quad 0 \times \infty = 0 \quad \text{for } y > 0.$$

This helps in avoiding the nuisance of separately considering those indices for which p_i^n is positive and those for which it is 0.

From $Y_1, \ldots, Y_n$, we will construct $X_1, \ldots, X_n, \Pi_1, \ldots, \Pi_n$, where $X_1, \ldots, X_n$ are independent of each other and of $\Pi_1, \ldots, \Pi_n$, and are Exp(1) distributed; and where

$$\mathrm{P}[\Pi_1 = \pi_1, \ldots, \Pi_n = \pi_n] = \prod_{k=1}^{n} p_{i_k}^k$$

for a unique choice of $i_1, \ldots, i_n$ that is compatible with the symmetric queueing discipline. Moreover, if $I_1, \ldots, I_n$ are the (unique) random indices that result in $\Pi_1, \ldots, \Pi_n$, then

$$Y_{\Pi_k} = p_{I_k}^k X_k + \cdots + p_{I_n}^n X_n, \tag{8}$$

where the right-hand side has the same distribution as the amount of work received by the kth arriving customer, provided that no one leaves.

We perform this construction recursively, starting with X_n and Π_n. Let

$$X_n = \min_{1 \le i \le n} \frac{Y_i}{p_i^n} \quad \text{and} \quad \Pi_n = \arg \min_{1 \le i \le n} \frac{Y_i}{p_i^n}.$$

In particular, due to our definition of $y/0$, only the indices for which p_i^n is positive are participating in this minimum. Since $Y_i/p_i^n \sim \mathrm{Exp}(p_i^n)$, it immediately follows that $X_n \sim \mathrm{Exp}(p_1^n + \cdots + p_n^n) = \mathrm{Exp}(1)$, that $\mathrm{P}[\Pi_n = i] = p_i^n$, and that X_n and Π_n are independent. The random variables Π_n and X_n have the following interpretation: Π_n is the position at which the nth arriving customer is inserted and $p_{\Pi_n}^n X_n$ is the amount of service received by that customer up to the next arrival epoch.

To construct X_{n-1} and Π_{n-1}, now consider

$$Y_j - p_j^n X_n = p_j^n \left(\frac{Y_j}{p_j^n} - \min_{1 \le i \le n} \frac{Y_i}{p_i^n} \right)$$

and denote by $J_1^{n-1}, \ldots, J_{n-1}^{n-1}$ the indices (in increasing order) for which $J_k^{n-1} \neq \Pi_n$. That is, if $\Pi_n = i$ for some $1 < i < n$, then

$$(J_1^{n-1}, \ldots, J_{n-1}^{n-1}) = (1, \ldots, i-1, i+1, \ldots, n);$$

if $\Pi_n = n$ then $(J_1^{n-1}, \ldots, J_{n-1}^{n-1}) = (1, \ldots, n-1)$; and if $\Pi_n = 1$ then $(J_1^{n-1}, \ldots, J_{n-1}^{n-1}) = (2, \ldots, n)$. It is easy to check that

$$Y_{J_1^{n-1}} - p_{J_1^{n-1}}^n X_n, \ldots, Y_{J_{n-1}^{n-1}} - p_{J_{n-1}^{n-1}}^n X_n, X_n, \Pi_n$$

are independent, with $Y_{J_k^{n-1}} - p_{J_k^{n-1}}^n X_n \sim \mathrm{Exp}(1)$.

https://doi.org/10.1239/jap/1110381382 Published online by Cambridge University Press

Next, we write $Y^{n-1}_{J_k} = Y_{J^{n-1}_k} - p^n_{J^{n-1}_k} X_n$,

$$X_{n-1} = \min_{1\le i\le n-1}\left(Y^{n-1}_{J^{n-1}_i}/p^{n-1}_i\right),$$

and

$$I_{n-1} = \arg\min_{1\le i\le n-1}\left(Y^{n-1}_{J^{n-1}_i}/p^{n-1}_i\right).$$

We set $\Pi_{n-1} = J^{n-1}_{I_{n-1}}$ and observe that X_{n-1}, X_n, and (Π_{n-1}, Π_n) are independent, with $X_{n-1} \sim \text{Exp}(1)$.

It is important to note that we associate p^{n-1}_i with the index J^{n-1}_i. If we were not careful to do this, we would get an ordering that is incompatible with the symmetric queueing discipline. For example, between the $(n-1)$th and nth arrival epochs, it is not possible for a customer to be in any position other than $i-1$ or i (depending on whether the newly arriving customer is placed in a position from i to n or from 1 to $i-1$, respectively) if he is to be in position i between the nth and $(n+1)$th arrival epochs. The above construction preserves this.

As before, we now let $J^{n-2}_1, \dots, J^{n-2}_{n-2}$ be the indices (in increasing order) that exclude Π_n and Π_{n-1}. It is again evident that

$$Y^{n-1}_{J^{n-2}_1} - p^{n-1}_{J^{n-2}_1} X_{n-1}, \dots, Y^{n-1}_{J^{n-2}_{n-2}} - p^{n-1}_{J^{n-2}_{n-2}} X_{n-1}, X_{n-1}, X_n, (\Pi_{n-1}, \Pi_n)$$

are independent, where

$$\mathrm{P}[\Pi_{n-1} = i, \Pi_n = j] = \begin{cases} p^{n-1}_i p^n_j & \text{if } i < j, \\ p^{n-1}_{i-1} p^n_j & \text{if } i > j, \end{cases}$$

and the other variables are Exp(1) distributed.

Letting $Y^{n-2}_{J_i} = Y^{n-1}_{J^{n-2}_i} - p^{n-1}_{J^{n-2}_i} X_{n-1}$ and associating p^{n-2}_i with $Y^{n-2}_{J_i}$, this process can be repeated, and eventually we obtain

$$X_1, \dots, X_n, \Pi_1, \dots, \Pi_n,$$

where $X_1, \dots, X_n$ are independent of each other, of $\Pi_1, \dots, \Pi_n$, and are Exp(1) distributed; and where

$$\mathrm{P}[\Pi_1 = \pi_1, \dots, \Pi_n = \pi_n] = \prod_{k=1}^{n} p^k_{i_k} \tag{9}$$

for an appropriate choice of $i_1, \dots, i_n$ that is compatible with the symmetric queueing discipline, as required. Here it should be noted that, for every permutation $\pi_1, \dots, \pi_n$, there is a unique choice of $i_1, \dots, i_n$ such that the right-hand side of (9) is equal to the left-hand side. Observe that i_k is the position at which the kth arriving customer is inserted.

With this construction, it can be checked that (8) holds, where $I_1, \dots, I_n$ are the unique insertion locations that result in $\Pi_1, \dots, \Pi_n$. Since $\boldsymbol{I} = (I_1, \dots, I_n)$ is a functional of $\Pi_1, \dots, \Pi_n$, we have that $X_1, \dots, X_n$, and $\boldsymbol{I}$ are independent and, thus,

$$\{p^k_{I_k} X_k + \cdots + p^n_{I_n} X_n, k = 1, \dots, n\}$$

have as their joint distribution that of the amount of service allocated to the arriving customers until the $(n+1)$th arrival epoch. Thus, $(Y_{\Pi_1}, \dots, Y_{\Pi_n})$ also has this distribution and we are done.

https://doi.org/10.1239/jap/1110381382 Published online by Cambridge University Press

Remark 3. We note that, in the special case of the PS discipline, $p_i^n = 1/n$ and our construction implies that $Y_{\Pi_1}, \ldots, Y_{\Pi_n}$ are the order statistics, that is, they are a reordering of $Y_1, \ldots, Y_n$ in decreasing order. For the special case of the LIFO discipline, $\Pi_i = i$ and, so, $\mathbf{\Pi} = (1, \ldots, n)$ with probability 1.

We now return to the original M/G/1 queue, that is, with a Poisson arrival process $N = \{N(t), t \geq 0\}$ with rate λ, and independent, identically distributed service times $B_1, B_2, \ldots$, that are independent of the arrival process. Let $D(t) = N(t) - L(t)$ be the number of departures by time t and let $D_1 = \inf\{t\colon D(t) = 1\}$ be the time until the first departure.

Theorem 3. *Let $\tau(q) \sim \text{Exp}(q)$ be independent of $(N, B_1, B_2, \ldots)$ and assume that, at time 0, the system is empty. Then, for any symmetric queueing discipline,*

$$\text{P}[D(\tau(q)) = 0 \mid N(\tau(q)) = n] = (1 - \text{E}\,\text{e}^{-(\lambda+q)B_1})^n. \tag{10}$$

Consequently,

$$\text{P}[D(\tau(q)) = 0] = \frac{q}{q + \lambda\,\text{E}\,\text{e}^{-(\lambda+q)B_1}} \tag{11}$$

and, hence,

$$\text{E}\,\text{e}^{-qD_1} = \frac{\lambda\,\text{E}\,\text{e}^{-(\lambda+q)B_1}}{q + \lambda\,\text{E}\,\text{e}^{-(\lambda+q)B_1}}. \tag{12}$$

Proof. It is well known that the number of arrivals until time $\tau(q)$ has a geometric distribution. That is,

$$\text{P}[N(\tau(q)) = n] = \left(\frac{\lambda}{q+\lambda}\right)^n \frac{q}{q+\lambda}. \tag{13}$$

Moreover, it is also well known and easy to check that if $S_1, \ldots, S_n$ are the first n arrival epochs of the Poisson process N, then the conditional distribution of $S_1, S_2 - S_1, \ldots, S_n - S_{n-1}, \tau(q) - S_n$, given that $N(\tau(q)) = n$, is that of $n+1$ independent random variables that are $\text{Exp}(q+\lambda)$ distributed. From (7) and the derivation that follows it, we have

$$\text{P}[D(\tau(q)) = 0 \mid N(\tau(q)) = n] = \text{P}[B_1 > Y_1(q+\lambda)]^n,$$

where $Y_1(q+\lambda) \sim \text{Exp}(q+\lambda)$ and is independent of B_1. Since

$$\text{P}[B_1 \leq Y_1(q+\lambda)] = \text{E}\,\text{e}^{-(q+\lambda)B_1},$$

we then obtain (10).

By multiplying (10) by (13), summing, and simplifying, we obtain the right-hand side of (11). Finally, we note that

$$\text{P}[D(\tau(q)) = 0] = \text{P}[D_1 > \tau(q)] = 1 - \text{P}[D_1 \leq \tau(q)] = 1 - \text{E}\,\text{e}^{-qD_1},$$

which gives (12).

7. The distribution of the first departure time

Theorem 3 will allow us to determine the distribution of the time until the first departure from the M/G/1 queue with symmetric service discipline. First, some notation. Recall from Section 2 that $B(\cdot)$ denotes the service time distribution, with LST $\beta(\cdot)$. Let

$$B_\lambda(\text{d}t) = \frac{\text{e}^{-\lambda t}}{\beta(\lambda)} B(\text{d}t)$$

https://doi.org/10.1239/jap/1110381382 Published online by Cambridge University Press

and

$$B_{\mathrm{e},\lambda}(\mathrm{d}t) = \frac{\mathrm{e}^{-\lambda t}(1 - B(t))\mathrm{d}t}{\int_0^\infty \mathrm{e}^{-\lambda u}(1 - B(u))\mathrm{d}u} = \frac{\lambda \mathrm{e}^{-\lambda t}(1 - B(t))\mathrm{d}t}{1 - \beta(\lambda)}.$$

In particular, note that

$$\int_0^t \mathrm{e}^{-\lambda u}(1 - B(u))\mathrm{d}u = \mathrm{E}\int_0^t \mathrm{e}^{-\lambda u}\,\mathbf{1}_{\{B_1 > u\}}\,\mathrm{d}u = \mathrm{E}\int_0^{B_1 \wedge t} \mathrm{e}^{-\lambda u}\mathrm{d}u,$$

where $a \wedge b = \min(a, b)$ and $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function of the event $\{\cdot\}$. Thus,

$$B_{\mathrm{e},\lambda}(t) = \frac{1 - \mathrm{E}\,\mathrm{e}^{-\lambda(B_1 \wedge t)}}{1 - \mathrm{E}\,\mathrm{e}^{-\lambda B_1}}.$$

From these definitions, it is clear that

$$\frac{\beta(q + \lambda)}{\beta(\lambda)} = \int_0^\infty \mathrm{e}^{-qt} B_\lambda(\mathrm{d}t) =: \beta_\lambda(q) \tag{14}$$

and that

$$\frac{(1 - \beta(\lambda + q))/(\lambda + q)}{(1 - \beta(\lambda))/\lambda} = \int_0^\infty \mathrm{e}^{-qt} B_{\mathrm{e},\lambda}(\mathrm{d}t) =: \beta_{e,\lambda}(q). \tag{15}$$

Also, it can be easily verified that

$$B_\lambda(t) = \mathrm{P}[B_1 \le t \mid B_1 \le Y_1(\lambda)]$$

and that

$$B_{\mathrm{e},\lambda}(t) = \mathrm{P}[Y_1(\lambda) \le t \mid B_1 > Y_1(\lambda)],$$

where $Y_1(\lambda) \sim \mathrm{Exp}(\lambda)$ and is independent of B_1.

With these definitions, we are now able to characterize the distribution of D_1, starting from an empty system. In what follows, by $R \sim G(p)$ we mean that $\mathrm{P}[R = n] = p(1 - p)^n$ for $n \ge 0$.

Theorem 4. *Let $Y \sim \mathrm{Exp}(\lambda)$, $X \sim B_\lambda$, $I \sim G(\beta(\lambda))$, and $Z_i \sim B_{\mathrm{e},\lambda}$, where Y, X, I, Z_1, $Z_2, \ldots$ are independent. Set $W_0 = 0$ and $W_n = \sum_{i=1}^n Z_i$ for $n \ge 1$. Then, under the conditions of Theorem 3,*

$$D_1 \sim Y + X + W_I. \tag{16}$$

Proof. From (12), (14), and (15), it is simple to verify that

$$\begin{aligned}
\mathrm{E}\,\mathrm{e}^{-qD_1} &= \frac{\lambda\,\mathrm{E}\,\mathrm{e}^{-(\lambda+q)B_1}}{q + \lambda\,\mathrm{E}\,\mathrm{e}^{-(\lambda+q)B_1}} = \frac{\lambda\beta(\lambda + q)}{\lambda + q - \lambda(1 - \beta(\lambda + q))} \\
&= \frac{\lambda}{\lambda + q}\,\frac{\beta(\lambda)\beta_\lambda(q)}{1 - (1 - \beta(\lambda))\beta_{e,\lambda}(q)} \\
&= \frac{\lambda}{\lambda + q}\beta_\lambda(q)\sum_{n=0}^\infty (1 - \beta(\lambda))^n \beta(\lambda)\beta_{e,\lambda}^n(q),
\end{aligned} \tag{17}$$

and the result follows.

https://doi.org/10.1239/jap/1110381382 Published online by Cambridge University Press

Remark 4. In the LIFO case, (16) has a simple interpretation. Indeed, D_1 then consists of the following three terms: (i) the first arrival interval Y; (ii) a service time X conditioned on being *smaller* than the next interarrival interval; and (iii) a number of service times (of newly arriving customers, who are immediately being taken into service), all conditioned on being larger than the next interarrival time – this is a $G(\beta(\lambda))$-distributed random variable.

We now recall that $\alpha^* = \inf\{\alpha\colon \phi(\alpha) > 0\}$, where $\phi(\alpha) = \alpha - \lambda(1 - \beta(\alpha))$ for $\alpha > 0$.

Corollary 1. *Let*

$$u^* = \sup\{u\colon u < \lambda, \lambda\beta(\lambda - u) > u\} = \lambda - \alpha^*.$$

Then, for each $u < u^*$,

$$\mathrm{E}\,\mathrm{e}^{uD_1} = \frac{\lambda\beta(\lambda - u)}{\lambda\beta(\lambda - u) - u} \tag{18}$$

is finite. Moreover,

$$\lim_{u\uparrow u^*} \mathrm{E}\,\mathrm{e}^{uD_1} = \infty.$$

Proof. Y, X, and Z_i have finite moment-generating functions for $u < \lambda$. Thus, if we show that

$$(1 - \beta(\lambda))\,\mathrm{E}\,\mathrm{e}^{uZ_1} = \frac{\lambda}{\lambda - u}(1 - \beta(\lambda - u)) \tag{19}$$

is strictly less than 1, then the form of $\mathrm{E}\,\mathrm{e}^{uD_1}$ follows from Theorem 4. The right-hand side of (19) is less than 1 if and only if $\lambda\beta(\lambda - u) - u$ is strictly positive, which is true because $u < u^*$. If $u^* = \lambda$ then, since $\mathrm{E}\,\mathrm{e}^{uY} = \lambda/(\lambda - u) \to \infty$ as $u \uparrow \lambda$, this must also hold for D_1. If $u^* < \lambda$ then the denominator of (18) converges to 0 from above and, hence, $\mathrm{E}\,\mathrm{e}^{uD_1}$ converges to infinity.

Remark 5. We recall that $\alpha^* = 0$ whenever $\rho := \lambda E B_1 \le 1$, and that $\alpha^* > 0$ whenever $\rho > 1$. In either case, $\alpha^* < \lambda$ since $\phi(\lambda) = \lambda\beta(\lambda) > 0$. Moreover, $\phi(\alpha) > 0$ for $\alpha > \alpha^*$ and, in particular, for $\alpha^* < \alpha \le \lambda$. This implies that if $\rho \le 1$ then $u^* = \lambda$, if $\rho > 1$ then $0 < u^* < \lambda$, and that $\lambda\beta(\lambda - u) - u > 0$ for $0 \le u < u^*$.

Remark 6. All moments of D_1 are finite, without the need for any moment conditions on the service times. In particular, it is not necessary to assume that the traffic intensity is less than 1 or even that the service time has a finite mean. This may seem surprising at first, as it is definitely false for, e.g. the first-come-first-served discipline. However, considering the preemptive-resume LIFO discipline, it becomes more plausible, since the first customer to depart is the first one whose service time is less than the exponential interarrival time that follows it.

We note that, with Y, X, I, and Z_i as in Theorem 4,

$$\mathrm{E}\,Y = \frac{1}{\lambda}, \qquad \mathrm{E}\,X = -\beta'_\lambda(0) = -\frac{\beta'(\lambda)}{\beta(\lambda)},$$
$$\mathrm{E}\,I = \frac{1 - \beta(\lambda)}{\beta(\lambda)}, \qquad \mathrm{E}\,Z_i = -\beta'_{e,\lambda}(0) = \frac{1}{\lambda} + \frac{\beta'(\lambda)}{1 - \beta(\lambda)}.$$

Since $\mathrm{E}\,D_1 = \mathrm{E}\,Y + \mathrm{E}\,X + \mathrm{E}\,I\,\mathrm{E}\,Z_1$, we can verify the following corollary.

https://doi.org/10.1239/jap/1110381382 Published online by Cambridge University Press

Corollary 2. *Under the conditions of Theorem 3,*

$$\mathrm{E}\, D_1 = \frac{1}{\lambda\beta(\lambda)}.$$

However, we note that this result is an immediate consequence of (12), which is obtained upon subtracting both sides of (12) from 1, dividing by q, and letting $q \downarrow 0$.

As for the variance, we observe that

$$\mathrm{var}(W_I) = \mathrm{E}\, I\, \mathrm{var}(Z_1) + \mathrm{var}(I)(\mathrm{E}\, Z_1)^2,$$

so that

$$\mathrm{var}(D_1) = \mathrm{var}(Y) + \mathrm{var}(X) + \mathrm{E}\, I\, \mathrm{var}(Z_1) + \mathrm{var}(I)(\mathrm{E}\, Z_1)^2.$$

Carrying out the computation, or directly from (12) via differentiation, we obtain a further corollary.

Corollary 3. *Under the conditions of Theorem 3,*

$$\mathrm{var}(D_1) = \frac{1 + 2\lambda\beta'(\lambda)}{(\lambda\beta^2(\lambda))^2}.$$

Since the function $f(x) = x\mathrm{e}^{-x}$ attains its maximum at $x = 1$,

$$-2\lambda\beta'(\lambda) = 2\,\mathrm{E}\,\lambda B\mathrm{e}^{-\lambda B} \leq 2\mathrm{e}^{-1} < 1,$$

meaning that the right-hand side of the formula for the variance is indeed positive.

8. The tail behavior of the first departure time

Here, we investigate the tail behavior of D_1, using Theorem 4. The logarithmic asymptotics follow from Corollary 1. If $\rho \neq 1$, it is also possible to derive exact asymptotics. We use the notation $f(x) \sim g(x)$ to indicate that $f(x) = g(x)(1 + \mathrm{o}(1))$ as $x \to \infty$. We first consider the case $\rho < 1$.

Proposition 1. *If* $\rho < 1$ *then*

$$\mathrm{P}[D_1 > x] \sim \frac{1}{1-\rho}\mathrm{e}^{-\lambda x}.$$

Proof. Proposition 5.1 of [14] implies the following. Let Y be exponential with rate λ and let A be such that $\mathrm{E}\,\mathrm{e}^{\lambda A} < \infty$. Then, $\mathrm{P}[Y + A > x] \sim \mathrm{E}\,\mathrm{e}^{\lambda A}\,\mathrm{P}[Y > x]$. We apply this result by choosing $A = X + W_I$ as defined in Theorem 4. From (17), it can easily be shown that $\mathrm{E}\,\mathrm{e}^{\lambda(X+W_I)} = 1/(1-\rho) < \infty$. This proves the assertion.

We now turn to the opposite case: $\rho > 1$. In this case, W_I will dominate the asymptotics. Recall the definition of u^* given in Corollary 1.

Proposition 2. *If* $\rho > 1$ *then*

$$\begin{aligned}\mathrm{P}[D_1 > x] &\sim \frac{(1-\beta(\lambda))\beta(\lambda-u^*)}{u^*(1-\beta'(\lambda-u^*) + (1-\beta(\lambda-u^*))/(\lambda-u^*))}\mathrm{e}^{-u^*x} \\ &= \frac{(1-\beta(\lambda))\beta(\alpha^*)}{(\lambda-\alpha^*)(1-\beta'(\alpha^*) + (1-\beta(\alpha^*))/\alpha^*)}\mathrm{e}^{-(\lambda-\alpha^*)x}. \qquad (20)\end{aligned}$$

https://doi.org/10.1239/jap/1110381382 Published online by Cambridge University Press

Proof. We first derive the tail behavior of $\mathrm{P}[W_I > x]$, using a general result on the tail behavior of geometric random sums. In particular, we use the version given as Theorem 2(ii) of [9], to obtain

$$\mathrm{P}[W_I > x] \sim \frac{\beta(\lambda)}{u^* \,\mathrm{E}[Z_1 \mathrm{e}^{u^* Z_1}]} \mathrm{e}^{-u^* x}.$$

The condition of that theorem is satisfied since $\mathrm{E}\,\mathrm{e}^{uZ_1}$ is finite for $u < \lambda$ and $u^* < \lambda$. Next, observe that $\mathrm{E}\,\mathrm{e}^{u^* Y} = \lambda/(\lambda - u^*)$ and $\mathrm{E}\,\mathrm{e}^{u^* X} = \beta_\lambda(-u^*)$ are finite. Applying Proposition 5.1 of [14] again, we obtain

$$\begin{aligned} \mathrm{P}[D_1 > x] &= \mathrm{P}[W_I + Y + X > x] \\ &\sim \frac{\lambda}{\lambda - u^*} \beta_\lambda(-u^*) \frac{\beta(\lambda)}{u^* \,\mathrm{E}[Z_1 \mathrm{e}^{u^* Z_1}]} \mathrm{e}^{-u^* x}. \end{aligned} \tag{21}$$

From (14),

$$\beta_\lambda(-u^*) = \frac{\beta(\lambda - u^*)}{\beta(\lambda)}$$

and, from (15),

$$\mathrm{E}[Z_1 \mathrm{e}^{u^* Z_1}] = -B'_{e,\lambda}(-u^*) = \frac{\lambda}{\lambda - u^*} \frac{1 - \beta'(\lambda - u^*) + (1 - \beta(\lambda - u^*))/(\lambda - u^*)}{1 - \beta(\lambda)}.$$

Hence, the right-hand side of (20) is equal to the right-hand side of (21).

If $\rho = 1$ and the service times have an exponentially bounded tail, then one can show that $\mathrm{P}[D_1 > x] \sim Cx\mathrm{e}^{-\lambda x}$ for some constant $C > 0$. We omit the details.

References

[1] Baskett, F., Chandy, K. M., Muntz, R. R. and Palacios, F. G. (1975). Open, closed, and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.* **22,** 248–260.

[2] Bertoin, J. (1995). *Lévy Processes.* Cambridge University Press.

[3] Bonald, T. and Proutière, A. (2002). Insensitivity in processor-sharing networks. *Performance Evaluation* **49,** 193–209.

[4] Burke, P. J. (1956). The output of a queueing system. *Operat. Res.* **4,** 699–704.

[5] Cohen, J. W. (1979). The multiple phase service network with generalized processor sharing. *Acta Informatica* **12,** 245–284.

[6] Cohen, J. W. (1982). *The Single Server Queue*, 2nd edn. North-Holland, Amsterdam.

[7] Duquesne, T. and Le Gall, J.-F. (2002). Random trees, Lévy processes and spatial branching processes. *Astérisque* **281,** vi + 147 pp.

[8] Jackson, J. R. (1963). Jobshop-like queueing systems. *Manag. Sci.* **10,** 131–142.

[9] Kalashnikov, V. and Tsitsiashvili, G. (1999). Tails of waiting times and their bounds. *Queueing Systems* **32,** 257–283.

[10] Kelly, F. P. (1976). Networks of queues. *Adv. Appl. Prob.* **8,** 416–432.

[11] Kelly, F. P. (1979). *Reversibility and Stochastic Networks.* John Wiley, Chichester.

[12] Kitaev, M. Yu. (1993). The M/G/1 processor-sharing model: transient behavior. *Queueing Systems* **14,** 239–273.

[13] Limic, V. (2001). A LIFO queue in heavy traffic. *Ann. Appl. Prob.* **11,** 301–331.

[14] Maulik, K. and Zwart, B. (2004). Tail asymptotics for exponential functionals of Lévy processes. EURANDOM Report 2004-036.

[15] O'Connell, N. and Yor, M. (2001). Brownian analogues of Burke's theorem. *Stoch. Process. Appl.* **96,** 285–304.

[16] Rosenkrantz, W. (1983). Calculation of the Laplace transform of the length of the busy period for the M/G/1 queue via martingales. *Ann. Prob.* **11,** 817–818.

[17] Shalmon, M. (1988). Analysis of the GI/G/1 queue and its variations via the LCFS preemptive resume discipline and its random walk interpretation. *Prob. Eng. Inf. Sci.* **2,** 215–230.

[18] Sigman, K. (1996). Queues under preemptive LIFO and ladder height distributions for risk processes: a duality. *Stoch. Models* **12,** 725–735.