Some Time-Dependent Properties of Symmetric M/G/1 Queues

Offer Kella[†]
Bert Zwart[‡]
Onno Boxma[§]

September 15, 2004

Abstract: Consider an $M/G/1$ queue which 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 (LIFO) and the processor sharing (PS) 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 queues, time-dependent analysis, insensitivity, order statistics, random permutations, tail behavior.
AMS Subject Classification: Primary 60K25, Secondary 90B22.

[†]Department of Statistics; The Hebrew University of Jerusalem; Mount Scopus, Jerusalem 91905; Israel (mskella@mscc.huji.ac.il). Supported in part by grant 819/03 from the Israel Science Foundation.

[‡]Department of Mathematics and Computer Science; Eindhoven University of Technology; P.O. Box 513; 5600 MB Eindhoven; The Netherlands and CWI; P.O. Box 94079; 1090 GB Amsterdam; The Netherlands (zwart@win.tue.nl). Supported by an NWO VENI grant.

[§]EURANDOM and Department of Mathematics and Computer Science; Eindhoven University of Technology; P.O. Box 513; 5600 MB Eindhoven; The Netherlands and CWI; P.O. Box 94079; 1090 GB Amsterdam; The Netherlands (boxma@win.tue.nl). Work carried out within the Euro-NGI project.

1 Introduction

One of the major success stories in applied probability has been the development of the theory of product-form queueing networks. Classical papers include [4], [9], [1], [11], and [5]; see [12] for a textbook treatment. Recent interesting papers are [3] and [7].

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 [12], 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 we 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 show that $L(\tau(q))$ has a geometric distribution. We find this very pleasing, since exactly the same distribution was found earlier by [13] for the PS discipline. This implies that for any $t > 0$, $L(t)$ has the same distribution for both disciplines.

It would be very nice if the distribution of $L(t)$ would be the same for *all* symmetric disciplines. At present, this is beyond our reach and 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 notations 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 i.i.d. random variables $B_i, i \geq 1$, denote the service times of successively arriving customers, with distribution $B(\cdot)$. As usual, denote $\rho \equiv \lambda E B_1$. Let $\beta(\alpha) \equiv Ee^{-\alpha B_1}$ be the LST of B_1 for $\alpha \geq 0$. Define the net input process $X(t) = \sum_{i=1}^{N(t)} B_i - t$. $X(t)$ is a Lévy process with exponent $\phi(\alpha) = \alpha - \lambda(1 - \beta(\alpha))$, i.e. for Re $\alpha \geq 0$,

$$E[\mathrm{e}^{-\alpha X(t)}] = \mathrm{e}^{t\phi(\alpha)}. \tag{1}$$

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 $\alpha^* = \inf\{\alpha : \phi(\alpha) > 0\}$. When $\rho \leq 1$ then $\alpha^* = 0$ and when $\rho > 1$ then $\alpha^* > 0$, since for the former $\phi'(0) = 1 - \rho \geq 0$ and for the latter $\phi'(0) < 0$. Since ϕ is continuous and strictly increasing on $[\alpha^*, \infty)$, then when viewed as a function from $[\alpha^*, \infty)$ to $[0, \infty)$, it has an inverse which we denote by $\kappa(q),\ q \geq 0$.

The service discipline is assumed to be *symmetric*. Recall (cf. Section 3.3 of [12]) 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 one. If there are $n-1$ customers in the system in positions $1, \ldots, n-1$ upon the arrival of a kth customer, $k \geq n$, then this arriving 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 (take $p_1^n = 1$) and the PS discipline (take $p_i^n = 1/n$ for $i = 1, \ldots, n$). The $M/G/1$ queue length process for the latter two disciplines is studied in the next two sections.

Throughout this paper $\{L(t),\ t \geq 0\}$ denotes the queue length process (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 notations $L_{\rm LIFO}(t)$ and $L_{\rm PS}(t)$, respectively, but not otherwise.

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 $X(s-) = \lim_{t \uparrow s} X(t)$. In the sequel, $\#\{s : S(s)\}$ denotes the number of s-values for which statement $S(s)$ holds.

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

$$L(t) = \#\{s \in [0,t] : X(s-) = \inf_{r \in [s,t]} X(r)\}. \tag{2}$$

This relation is explicitly stated in [14], where it is applied to derive a diffusion approximation for $\{L(t),\ t \geq 0\}$. It is also hidden 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 a.s. 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 to [8] 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*

$$P[L(\tau(q)) = n] = \left(1 - \frac{q}{\kappa(q)}\right)^n \frac{q}{\kappa(q)}\ . \tag{3}$$

Proof: We apply Lemma 2 as follows. Set $X_t(s) = X(t) - X((t-s)-)$. Then straightforward manipulations show that

$$L(t) = \#\{s \in [0,t] : X_t(s) = \sup_{r \in [0,s]} X_t(r)\}. \tag{4}$$

Since $X(t)$ is reversible, we obtain for every $t > 0$,

$$L(t) \stackrel{d}{=} \#\{s \in [0,t] : X(s) = \sup_{r\in[0,s]} X(r)\}. \tag{5}$$

From this it follows that

$$L(\tau(q)) \stackrel{d}{=} \#\{s \in [0,\tau(q)] : X(s) = \sup_{r\in[0,s]} X(r)\}. \tag{6}$$

Let $\tau_i, i \geq 1$, 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 easy to see 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

$$P[K(\tau(q)) \geq n] = P[R_n \leq \tau(q)] = E[\mathrm{e}^{-qR_n}] = (E[\mathrm{e}^{-qr_1}])^n. \tag{7}$$

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

$$P[L(\tau(q)) = 0] = P[\tau_1 > \tau(q)]. \tag{8}$$

Set $S(t) = \sup_{0<s<t} X(s)$. Note that

$$P[\tau_1 > \tau(q)] = P[S(\tau(q)) = 0]. \tag{9}$$

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

$$E[\mathrm{e}^{-\alpha S(\tau(q))}] = \frac{q(\kappa(q)-\alpha)}{\kappa(q)(q-\phi(\alpha))}. \tag{10}$$

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

$$P[S(\tau(q)) = 0] = \lim_{\alpha\to\infty} E[\mathrm{e}^{-\alpha S(\tau(q))}] = \frac{q}{\kappa(q)}. \tag{11}$$

(7) and (11) imply (3). ∎

Remark 1 *We note that $P[L(\tau(q)) = 0|L(0) = 0]$ is the same as 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 page 260 of [6].*

4 The PS queue length

Our starting point is the following formula, which is (2.6) in [13].

$$\int_0^\infty \mathrm{e}^{-qt} E[z^{L(t)}]dt = \frac{1}{q + (1-z)\lambda(1-\pi(q))}, \tag{12}$$

where $\pi(q)$ 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 (12) we obtain, observing that $\kappa(q) = q + \lambda(1 - \pi(q))$, that

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

This is the generating function of the right side of (3), that is, of a geometric random variable with probability of success $q/\kappa(q)$. We thus arrive at the following interesting result.

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

$$L_{\mathrm{LIFO}}(\tau(q)) \stackrel{d}{=} L_{\mathrm{PS}}(\tau(q)) \qquad \textit{for all } q > 0 \tag{13}$$

and

$$L_{\mathrm{LIFO}}(t) \stackrel{d}{=} L_{\mathrm{PS}}(t) \qquad \textit{for all } t > 0\ . \tag{14}$$

Proof: (13) follows from Theorem 1 and the computations made above. (14) is implied by (13) from the uniqueness property of Laplace transforms, as sampling at an exponential time is equivalent to taking 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, where 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* $\mathbf{U} = (U_1, \ldots, U_n)$, $\mathbf{V} = (V_1, \ldots, V_n)$ *be random variables and let* $\mathbf{\Pi} = (\Pi_1, \ldots, \Pi_n)$ *be a random permutation such that* $(\mathbf{\Pi}, \mathbf{V})$ *and* $\mathbf{U}$ *are independent and* $U_1, \ldots, U_n$ *are exchangeable. Then*

$$P[U_1 > V_{\Pi_1}, \ldots, U_n > V_{\Pi_n}] = P[U_1 > V_1, \ldots, U_n > V_n]. \tag{15}$$

When in addition $U_1, \ldots, U_n$ *are independent and* $V_1, \ldots, V_n$ *are i.i.d. r.v.'s then, in particular*

$$P[U_1 > V_{\Pi_1}, \ldots, U_n > V_{\Pi_n}] = P[U_1 > V_1]^n. \tag{16}$$

Proof: For any fixed permutation $\pi = (\pi_1, \ldots, \pi_n)$,

$$\begin{aligned} &P[U_1 > V_{\Pi_1}, \ldots, U_n > V_{\Pi_n} | \mathbf{\Pi} = \pi] \\ &\quad = P[U_1 > V_{\pi_1}, \ldots, U_n > V_{\pi_n} | \mathbf{\Pi} = \pi] \\ &\quad = P[U_{\pi_1} > V_{\pi_1}, \ldots, U_{\pi_n} > V_{\pi_n} | \mathbf{\Pi} = \pi] \\ &\quad = P[U_1 > V_1, \ldots, U_n > V_n | \mathbf{\Pi} = \pi]. \end{aligned} \tag{17}$$

The second equality follows from the fact that $\mathbf{U}$ is independent of $(\mathbf{\Pi}, \mathbf{V})$ and its components are exchangeable. Multiplying the leftmost and rightmost expressions by $P[\mathbf{\Pi} = \pi]$ 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 inter-arrival 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 be replaced later on with another parameter. Thus, for now, $Y_n \sim \exp(1)$ 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 $n+1$ arrival epoch is identical to that of $Y_{\Pi_1}, \ldots, Y_{\Pi_n}$ for some random permutation $\mathbf{\Pi}$ which is a functional of $\mathbf{Y}$. If this can be achieved, then once we introduce the service times, $(\mathbf{\Pi}, \mathbf{Y})$ and $\mathbf{B}$ are independent and thus it follows from Lemma 2 that the probability that none of the first n arriving customers has departed by the $n+1$ arrival epoch is given by

$$P[B_1 > Y_{\Pi_1}, \ldots, B_n > Y_{\Pi_n}] = P[B_1 > Y_1]^n. \tag{18}$$

This will allow us to study the distribution of the first departure time in a symmetric queue (cf. Theorems 3 and 4 below). For what follows we define for $y > 0$

$$\frac{y}{0} = \infty, \qquad 0 \cdot \infty = 0. \tag{19}$$

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

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 as well as independent from $\Pi_1, \ldots, \Pi_n$, are $\exp(1)$ distributed and

$$P[\Pi_1 = \pi_1, \ldots, \Pi_n = \pi_n] = \prod_{k=1}^{n} p_{i_k}^k \tag{20}$$

for a unique choice of $i_1, \ldots, i_n$ which 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

$$Y_{\Pi_k} = p_{I_k}^k X_k + \ldots + p_{I_n}^n X_n \ , \tag{21}$$

where the right side is distributed like the amount of work received by the kth arriving customer, provided that no one leaves.

We perform this construction recursively, starting with X_n, Π_n. Denote

$$X_n = \min_{1 \leq i \leq n} \frac{Y_i}{p_i^n}, \tag{22}$$

and

$$\Pi_n = \arg\min_{1\le i\le n} \frac{Y_i}{p_i^n}. \tag{23}$$

In particular, due to our definition of $y/0$ only the indices with positive p_i^n are participating in this minimum. Since $\frac{Y_i}{p_i^n} \sim \exp(p_i^n)$ it immediately follows that $X_n \sim \exp(p_1^n + \ldots + p_n^n) = \exp(1)$, that $P[\Pi_n = i] = p_i^n$ 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 where the n-th 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}, Π_{n-1}, consider now

$$Y_j - p_j^n X_n = p_j^n \left(\frac{Y_j}{p_j^n} - \min_{1\le i\le n} \frac{Y_i}{p_i^n} \right) \tag{24}$$

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

$$(J_1^{n-1}, \ldots, J_{n-1}^{n-1}) = (1, \ldots, i-1, i+1, \ldots, n) \ , \tag{25}$$

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

$$Y_{J_1^{n-1}} - p_{J_1^{n-1}}^n X_n, \ldots, Y_{J_{n-1}^{n-1}} - p_{J_{n-1}^{n-1}}^n X_n, X_n, \Pi_n \tag{26}$$

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

Next we denote $Y_{J_k}^{n-1} = Y_{J_k^{n-1}} - p_{J_k^{n-1}}^n X_n$, let

$$X_{n-1} = \min_{1\le i\le n-1} \frac{Y_{J_i^{n-1}}^{n-1}}{p_i^{n-1}} \tag{27}$$

and for

$$I_{n-1} = \arg\min_{1\le i\le n-1} \frac{Y_{J_i^{n-1}}^{n-1}}{p_i^{n-1}}. \tag{28}$$

We set $\Pi_{n-1} = J_{I_{n-1}}^{n-1}$ and observe that $X_{n-1}, X_n, (\Pi_{n-1}, \Pi_n)$ are independent with $X_{n-1} \sim \exp(1)$.

It is important to note that we associate p_i^{n-1} with the index J_i^{n-1}. If we would not be careful to do this we would get an ordering which is not compatible with the symmetric queueing discipline. For example, it is not possible that between the nth and $n+1$th arrival epochs a customer is in position i and between the $n-1$th and nth arrival epochs it is in any position other than $i-1$ or i, depending on whether the newly arriving customer is placed in position $i,\ldots,n$ or $1,\ldots,i-1$, respectively. The construction above preserves this.

Similarly as before, we now let $J_1^{n-2},\ldots,J_{n-2}^{n-2}$ be the indices in increasing order that exclude Π_n and Π_{n-1}. As before, it is evident that

$$Y_{J_1^{n-2}}^{n-1}-p_{J_1^{n-2}}^{n-1}X_{n-1},\ldots,Y_{J_{n-2}^{n-2}}^{n-1}-p_{J_{n-2}^{n-2}}^{n-1}X_{n-1},X_{n-1},X_n,(\Pi_{n-1},\Pi_n) \tag{29}$$

are independent, where

$$P[\Pi_{n-1}=i,\Pi_n=j]=\begin{cases} p_i^{n-1}p_j^n & \text{if } i<j,\\ \\ p_{i-1}^{n-1}p_j^n & \text{if } i>j,\end{cases} \tag{30}$$

and the other variables are exp(1) distributed.

Letting $Y_{J_i}^{n-2}=Y_{J_i^{n-2}}^{n-1}-p_{J_i^{n-2}}^{n-1}X_{n-1}$ and associating p_i^{n-2} with $Y_{J_i}^{n-2}$ this process can be repeated and eventually one obtains, as desired,

$$X_1,\ldots,X_n,\Pi_1,\ldots,\Pi_n \tag{31}$$

where $X_1,\ldots,X_n$ are independent as well as independent from $\Pi_1,\ldots,\Pi_n$, are exp(1) distributed and

$$P[\Pi_1=\pi_1,\ldots,\Pi_n=\pi_n]=\prod_{k=1}^{n}p_{i_k}^k \tag{32}$$

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

With this construction it can be checked that (21) holds, where $I_1,\ldots,I_n$ are the unique insertion locations that achieve

$\Pi_1, \ldots, \Pi_n$. Since $\mathbf{I} = (I_1, \ldots, I_n)$ is a functional of $\Pi_1, \ldots, \Pi_n$ we have that $X_1, \ldots, X_n, \mathbf{I}$ are independent and thus

$$\left\{p^k_{I_k} X_k + \ldots + p^n_{I_n} X_n |\ k = 1, \ldots, n\right\} \tag{33}$$

are jointly distributed like the amount of services allocated to the arriving customers until the $(n+1)$-st arrival epoch. Thus $(Y_{\Pi_1}, \ldots, Y_{\Pi_n})$ also has this distribution and we are done.

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

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

Theorem 3 *Let $\tau(q) \sim \exp(q)$ 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,*

$$P[D(\tau(q)) = 0 |\ N(\tau(q)) = n] = \left(1 - Ee^{-(\lambda+q)B_1}\right)^n \ . \tag{34}$$

Consequently,

$$P[D(\tau(q)) = 0] = \frac{q}{q + \lambda E e^{-(\lambda+q)B_1}} \ , \tag{35}$$

and hence

$$Ee^{-qD_1} = \frac{\lambda E e^{-(\lambda+q)B_1}}{q + \lambda E e^{-(\lambda+q)B_1}} \ . \tag{36}$$

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

$$P[N(\tau(q)) = n] = \left(\frac{\lambda}{q+\lambda}\right)^n \frac{q}{q+\lambda}. \tag{37}$$

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 which are distributed $\exp(q+\lambda)$. From (16) and the derivation that follows it we have that

$$P[D(\tau(q)) = 0|\ N(\tau(q)) = n] = P[B_1 > Y_1(q+\lambda)]^n \tag{38}$$

where $Y_1(q+\lambda) \sim \exp(q+\lambda)$ and is independent of B_1. Since

$$P[B_1 \leq Y_1(q+\lambda)] = Ee^{-(q+\lambda)B_1}\ , \tag{39}$$

(34) follows.

Multiplying (34) by (37), summing and simplifying, the right side of (35) is easily obtained. Finally, we note that

$$P[D(\tau(q)) = 0] = P[D_1 > \tau(q)] = 1 - P[D_1 \leq \tau(q)] = 1 - Ee^{-qD_1}\ , \tag{40}$$

which gives (36). ∎

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

$$B_\lambda(dt) = \frac{e^{-\lambda t}}{\beta(\lambda)} B(dt) \tag{41}$$

and

$$B_{e,\lambda}(dt) = \frac{e^{-\lambda t}(1 - B(t))dt}{\int_0^\infty e^{-\lambda u}(1 - B(u))du} = \frac{\lambda e^{-\lambda t}(1 - B(t))dt}{1 - \beta(\lambda)}. \tag{42}$$

In particular note that

$$\int_0^t e^{-\lambda u}(1 - B(u))du = E\int_0^t e^{-\lambda u} 1_{\{B_1 > u\}} du = E\int_0^{B_1 \wedge t} e^{-\lambda u} du\ , \tag{43}$$

where $a \wedge b = \min(a, b)$, and thus

$$B_{e,\lambda}(t) = \frac{1 - Ee^{-\lambda(B_1 \wedge t)}}{1 - Ee^{-\lambda B_1}}\ . \tag{44}$$

From these notations it is clear that

$$\frac{\beta(q+\lambda)}{\beta(\lambda)} = \int_0^\infty e^{-qt} B_\lambda(dt) \equiv \beta_\lambda(q) \tag{45}$$

and that

$$\frac{\dfrac{1-\beta(\lambda+q)}{\lambda+q}}{\dfrac{1-\beta(\lambda)}{\lambda}} = \int_0^\infty e^{-qt} B_{e,\lambda}(dt) \equiv \beta_{e,\lambda}(q) \tag{46}$$

Also, it can be easily verified that

$$B_\lambda(t) = P[B_1 \leq t |\ B_1 \leq Y_1(\lambda)] \tag{47}$$

and that

$$B_{e,\lambda}(t) = P[Y_1(\lambda) \leq t |\ B_1 > Y_1(\lambda)]\ , \tag{48}$$

where $Y_1(\lambda) \sim \exp(\lambda)$ 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. By $R \sim G(p)$ we mean that $P[R = n] = p(1-p)^n$ for $n \geq 0$.

Theorem 4 *Let $Y \sim \exp(\lambda)$, $X \sim B_\lambda$, $I \sim G\left(\beta(\lambda)\right)$ and $Z_i \sim B_{e,\lambda}$ 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 $n \geq 1$. Then, under the conditions of Theorem 3,*

$$D_1 \sim Y + X + W_I. \tag{49}$$

Proof: From (36), (45) and (46), it is simple to verify that

$$\begin{aligned} Ee^{-qD_1} &= \frac{\lambda E e^{-(\lambda+q)B_1}}{q + \lambda E e^{-(\lambda+q)B_1}} = \frac{\lambda\beta(\lambda+q)}{\lambda+q-\lambda(1-\beta(\lambda+q))} \\ &= \frac{\lambda}{\lambda+q} \cdot \frac{\beta(\lambda)\beta_\lambda(q)}{1-(1-\beta(\lambda))\beta_{e,\lambda}(q)} \\ &= \frac{\lambda}{\lambda+q} \cdot \beta_\lambda(q) \cdot \sum_{n=0}^{\infty} (1-\beta(\lambda))^n \beta(\lambda) \beta_{e,\lambda}^n(q) \end{aligned} \tag{50}$$

and the result follows. ∎

Remark 4 *In the LIFO case, Formula (49) can be easily interpreted. Indeed, D_1 then consists of the following three terms: (i) the first arrival interval Y; (ii) X, viz., a service time conditioned on being* smaller *than the next interarrival interval; (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 $\alpha^* = \inf\{\alpha \mid \ \phi(\alpha) > 0\}$ where, for $\alpha > 0$, $\phi(\alpha) = \alpha - \lambda(1 - \beta(\alpha))$.

Corollary 1 *Let*

$$u^* = \sup\{u \mid u < \lambda,\ \lambda\beta(\lambda - u) > u\} = \lambda - \alpha^* \ . \tag{51}$$

For each $u < u^*$,

$$Ee^{uD_1} = \frac{\lambda\beta(\lambda - u)}{\lambda\beta(\lambda - u) - u} \tag{52}$$

and is finite. Moreover,

$$\lim_{u \uparrow u^*} Ee^{uD_1} = \infty. \tag{53}$$

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

$$(1 - \beta(\lambda))Ee^{uZ_1} = \frac{\lambda}{\lambda - u}(1 - \beta(\lambda - u)) \tag{54}$$

is strictly less than one, then the form of Ee^{uD_1} follows from Theorem 4. The right side of (54) is less than one if and only if $\lambda\beta(\lambda - u) - u$ is strictly positive, which is true since $u < u^*$. If $u^* = \lambda$ then since $Ee^{uY} = \lambda/(\lambda - u) \to \infty$ as $u \uparrow \lambda$, this must also hold for D_1. If $u^* < \lambda$ then the denominator of (52) converges to zero from above and hence Ee^{uD_1} converges to infinity. ∎

Remark 5 *We recall that $\alpha^* = 0$ whenever $\rho \equiv \lambda E B_1 \leq 1$ 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 $\alpha^* < \alpha \leq \lambda$. This implies that if $\rho \leq 1$ then $u^* = \lambda$, if $\rho > 1$ then $0 < u^* < \lambda$ and that $\lambda\beta(\lambda - u) - u > 0$ for $0 \leq u < u^*$.*

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 one nor even that the service time has a finite mean. This may seem surprising at first sight, as this 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 one to depart is the first customer who has a service time which is less than the exponential inter-arrival time that follows it.*

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

$$\begin{aligned} EY &= \frac{1}{\lambda}, \\ EX &= -\beta'_\lambda(0) = -\frac{\beta'(\lambda)}{\beta(\lambda)}, \\ EI &= \frac{1-\beta(\lambda)}{\beta(\lambda)}, \\ EZ_i &= -\beta'_{e,\lambda}(0) = \frac{1}{\lambda} + \frac{\beta'(\lambda)}{1-\beta(\lambda)}. \end{aligned} \tag{55}$$

Since $ED_1 = EY + EX + EIEZ_1$, we can verify the following.

Corollary 2 *Under the conditions of Theorem 3*

$$ED_1 = \frac{1}{\lambda\beta(\lambda)}. \tag{56}$$

However, we note that this result is an immediate consequence of (36), which is obtained upon dividing the middle and right expressions of (36) by q and letting $q \downarrow 0$.

As for the variance we observe that

$$V(W_I) = EIV(Z_1) + V(I)(EZ_1)^2 \ , \tag{57}$$

so that

$$V(D_1) = V(Y) + V(X) + EIV(Z_1) + V(I)(EZ_1)^2. \tag{58}$$

Carrying out the computation or directly from (36) via differentiation one obtains:

Corollary 3 *Under the conditions of Theorem 3,*

$$V(D_1) = \frac{1 + 2\lambda\beta'(\lambda)}{(\lambda\beta^2(\lambda))^2}. \tag{59}$$

Since the function $f(x) = xe^{-x}$ attains its maximum at $x = 1$ then

$$-2\lambda\beta'(\lambda) = 2E\lambda Be^{-\lambda B} \leq 2e^{-1} < 1 \tag{60}$$

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

8 The tail behavior of the first departure time

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 denote that $f(x) = g(x)(1 + \mathrm{o}(1))$ as $x \to \infty$. We consider first the case $\rho < 1$.

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

$$P[D_1 > x] \sim \frac{1}{1-\rho} e^{-\lambda x}. \tag{61}$$

Proof: Proposition 5.1 of [15] implies the following: Let Y be exponential with rate λ and let A be such that $E[e^{\lambda A}] < \infty$. Then $P[Y + A > x] \sim E[e^{\lambda A}]P[Y > x]$. Apply this result by choosing $A = X + W_I$ as defined in Theorem 4. From (50) it can be easily be shown that $E[e^{\lambda(X+W_I)}] = 1/(1-\rho) < \infty$. This proves the assertion. ∎

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

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

$$\begin{aligned} P[D_1 > x] &\sim \frac{(1-\beta(\lambda))\beta(\lambda - u^*)}{u^* \left(1 - \beta'(\lambda - u^*) + \frac{1-\beta(\lambda-u^*)}{\lambda - u^*}\right)} e^{-u^* x} \\ &= \frac{(1-\beta(\lambda))\beta(\alpha^*)}{(\lambda - \alpha^*)\left(1 - \beta'(\alpha^*) + \frac{1-\beta(\alpha^*)}{\alpha^*}\right)} e^{-(\lambda-\alpha^*)x}. \end{aligned} \tag{62}$$

Proof: We first derive the tail behavior of $P[W_I > x]$ using a general result on the tail behavior of geometric random sums. In particular, we use the version given as Theorem 2(ii) in [10] to obtain

$$P[W_I > x] \sim \frac{\beta(\lambda)}{u^* E[Z_1 e^{u^* Z_1}]} e^{-u^* x}. \tag{63}$$

The condition of that theorem is satisfied since Ee^{uZ_1} is finite for $u < \lambda$ and $u^* < \lambda$. Next, observe that $E[e^{u^* Y}] = \lambda/(\lambda - u^*)$ and $Ee^{u^* X} = \beta_\lambda(-u^*)$ are finite. Applying again [15] we get

$$\begin{aligned} P[D_1 > x] &= P[W_I + Y + X > x] \\ &\sim \frac{\lambda}{\lambda - u^*} \beta_\lambda(-u^*) \frac{\beta(\lambda)}{u^* E[Z_1 e^{u^* Z_1}]} e^{-u^* x} \ . \end{aligned} \tag{64}$$

From (45),

$$\beta_\lambda(-u^*) = \frac{\beta(\lambda - u^*)}{\beta(\lambda)} \tag{65}$$

and from (46),

$$E[Z_1 e^{u^* Z_1}] = -B'_{e,\lambda}(-u^*) = \frac{\lambda}{\lambda - u^*} \frac{1 - \beta'(\lambda - u^*) + \frac{1-\beta(\lambda-u^*)}{\lambda - u^*}}{1 - \beta(\lambda)} \ . \tag{66}$$

Hence, the right side of (62) is equal to the right side of (64). ∎

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

References

[1] Baskett, F., Chandy, K.M., Muntz, R.R., Palacios-Gomez, F. (1975). Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM* **22**, 248–260.

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

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

[4] Burke, P.J. (1956). The output of a queueing system. *Operations Research* **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* (revised edition). North-Holland, Amsterdam.

[7] O'Connell, N., Yor, M. (2001). Brownian analogues of Burke's theorem. *Stochastic Processes and their Applications* **96**, 285–304.

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

[9] Jackson, J.R. (1963). Jobshop-like queueing systems. *Management Science* **10**, 131–142.

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

[11] Kelly, F.P. (1976). Networks of queues. *Advances in Applied Probability* **8**, 416–432.

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

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

[14] Limic, V. (2001). A LIFO queue in heavy traffic. *Annals of Applied Probability* **11**, 301–331.

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

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

[17] Shalmon, M. (1988). Analysis of the $GI/G/1$ queue and its variation via the LCFS preemptive resume discipline and its random walk interpretation. *Probability in the Engineering and Informational Sciences* **2**, 215–230.

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