

The Waiting-Time Distribution for the GI/G/1 Queue under the D-Policy

Jingwen Li
Department of Decision Sciences
National University of Singapore
10 Kent Ridge Crescent
Singapore 0511

Shun-Chen Niu¹
School of Management
University of Texas at Dallas
P. O. Box 830688
Richardson, Texas 75083-0688

July 1991

Revision: January 1992²

¹This research was supported in part by the National Science Foundation under grant DDM-9001751.

²*Probability in the Engineering and Informational Sciences*, Vol. 6 (1992), pp. 287–308.

Abstract

We study a generalization of the $GI/G/1$ queue in which the server is turned off at the end of each busy period and is reactivated only when the sum of the service times of all waiting customers exceeds a given threshold of size D . Using the concept of a “randomly-selected” arriving customer, we obtain as our main result a relation that expresses the waiting-time distribution of customers in this model in terms of characteristics associated with a corresponding standard $GI/G/1$ queue, obtained by setting $D = 0$. If either the arrival process is Poisson or the service times are exponentially distributed, then this representation of the waiting-time distribution can be specialized to yield explicit, transform-free formulas; we also derive, in both of these cases, the expected customer waiting times. Our results are potentially useful, for example, for studying optimization models in which the threshold D can be controlled.

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. $GI/G/1$ queue, $M/G/1$ queue, $GI/M/1$ queue, waiting-time distribution, D -policy, N -policy, T -policy, optimal control.

1 Introduction

We study in this paper a generalization of the $GI/G/1$ queue in which the server is turned off (i.e., becomes unavailable for service) at the end of each busy period and is activated again only when the sum of the service times of all waiting customers exceeds a threshold of size D , a fixed nonnegative real number. We shall call this model, of which the standard $GI/G/1$ queue is the special case with $D = 0$, the $GI/G/1$ queue under the D -policy.

The special case of our model in which the arrival process is Poisson, i.e., in the $M/G/1$ setting, was first formulated and studied by Balachandran [3] and Balachandran and Tijms [4]. Their primary interest was in optimal system control: Under the assumption that the system incurs a fixed “switching cost” each time the server is turned on and a “holding cost” at a constant rate per unit of unfinished work per unit time, they derived a formula for the long-run (time-)average cost to the system as a function of D . This formula was then used to determine the optimal value D^* that minimizes the long-run average cost, and to compare the resulting optimal long-run average cost to that of an $M/G/1$ queue under a related N -policy, which, originally introduced in Yadin and Naor [21] and further studied in Heyman [9], differs from the D -policy in that the server is activated when the *number* of waiting customers reaches level N . After studying several special classes of service-time distributions, it was conjectured that the D -policy is superior to the N -policy for all service-time distributions; this was subsequently confirmed by Boxma [5].

Our primary interest in this paper is to study the waiting-time distribution (in queue) of customers in the more general $GI/G/1$ queue under the D -policy. In addition to being of interest in its own right, the waiting-time distribution can be useful, for example, in the analysis of an optimization model in which the holding cost is incurred at a rate that is per unit of *customer* sojourn time (defined as the sum of waiting and service times), as opposed to per unit of unfinished work [2; 3, p. 1017, paragraph 1; 21]; in fact, knowledge of the waiting-time distribution makes possible, at least in principle, the analysis of more general models in which the holding cost is a *nonlinear* function of customer sojourn time. In the $M/G/1$ setting, such a model with linear holding cost has been studied in Rubin and Zhang [18] and, independently, in Li [11], where it is shown, among other results, that under this cost assumption the optimal D -policy is, perhaps surprisingly, not necessarily superior to the optimal N -policy; a result that can be attributed to the fact that, unlike the standard $M/G/1$ -FIFO queue where the virtual workload found by an arriving customer always equals his actual waiting time, the waiting time experienced by an arriving customer in the $M/G/1$ queue under the D -policy will typically be *longer* than the virtual workload if the server is unavailable for service at his arrival epoch.

The waiting-time distribution under the N -policy has been studied, in the $M/G/1$ setting, by Neuts [12,13] and by Shanthikumar [19]; in the same setting, Heyman [10] has studied the expected sojourn time under yet another related T -policy, according to which the server is activated T time units after it was last turned off (if there is no waiting

customer when the server is activated, it is turned off immediately). For the D -policy case, the waiting-time distribution, even in the $M/G/1$ setting, does not appear to have been studied in the literature before; and this is perhaps due to the difficulties that arise when one attempts to follow classical solution methods (e.g., the embedded Markov chain). Our main result, which is a relation that expresses the waiting-time distribution of customers in the $GI/G/1$ queue under the D -policy in terms of characteristics associated with the standard $GI/G/1$ queue, will be proved by constructive arguments that are based on the concept of a “randomly-selected” arriving customer. We believe our method of proof, which is closely related to that used earlier in Niu [14] and Niu and Cooper [16], is of some independent interest.

The outline of the rest of our paper is as follows. In Section 2, we formally define the concept of a randomly-selected arriving customer and state our main result, Theorem 1; we also summarize how to specialize Theorem 1 to the $M/G/1$ and the $GI/M/1$ settings, to obtain explicit, transform-free formulas for both the distribution and the expectation of waiting times. In Section 3, we provide proofs for the results stated in Section 2. Finally, in Section 4, we give some additional comments, and, in particular, show that our method of analysis can also be used to study the waiting-time distribution of customers in the $GI/G/1$ queue under the N -policy.

2 Main Results

We begin with some assumptions and notation. We assume that: Customers arrive according to a renewal process whose interarrival times follow distribution function $F(\cdot)$ with mean $1/\lambda$; the successive service times (brought in by arriving customers) are iid random variables, independent of the arrival process, following distribution function $G(\cdot)$ with mean $1/\mu$; the arrival rate λ is less than the service rate μ , so that $\rho \equiv \lambda/\mu < 1$ and the system is stable; and, finally, the first customer arrives at time zero, finding an empty system, and, when waiting times are considered, customers are served in the order of their arrival.

For $i \geq 1$, let W_i be the waiting time of the i^{th} customer before the commencement of his service. Then, with $\mathbf{1}_B(\cdot)$ denoting the indicator function of a given set B , the long-run (or limiting) proportion of customers with a waiting time less than or equal to x , $x \geq 0$, is given, w.p.1 (with probability 1), by

$$\nu[0, x] \equiv \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{[0, x]}(W_i) \quad (1)$$

(our model is regenerative, so that $\nu[0, x]$ is a constant w.p.1). As we vary x , Eq. (1) generates a probability measure ν on the Borel subsets of the nonnegative real line. The waiting time W of a randomly-selected arriving customer is, then, defined to be a *random variable* whose distribution is determined by ν ; i.e., let $P\{W \leq x\} = \nu[0, x]$. Notice that,

for any *given* sample path, the term $1/n$ in Eq. (1) can be interpreted as the “probability” of selecting any one of the first n customers, and $\mathbf{1}_{[0,x]}(W_i)$, $1 \leq i \leq n$, as the “conditional probability” for the i^{th} customer, if selected, to have a waiting time not exceeding x ; therefore, by letting $n \rightarrow \infty$, we are indeed taking the viewpoint of a randomly-selected arriving customer. (For related discussions, see Niu [14, Section 1] and Niu and Cooper [16].)

Under the D -policy, the server is activated as soon as the sum of the service times of all waiting customers exceeds D . We shall say that a busy period is of “type j ,” henceforth referred to as a j -busy period, if the number of waiting customers equals j when the server is activated. For $j \geq 1$, denote by b_j the probability that a busy period is of type j ; then, it is easily seen that

$$b_j = P \left\{ \sum_{i=1}^{j-1} S_i \leq D, \sum_{i=1}^j S_i > D \right\} = G^{[j-1]}(D) - G^{[j]}(D), \quad (2)$$

where S_i , $i \geq 1$, denotes the service time brought in by the i^{th} arriving customer in the busy period (a generic service time will be denoted by S), and $G^{[n]}(\cdot)$ denotes the n -fold self-convolution of $G(\cdot)$.

Denote by $E(K_j)$ the expectation of the number of customers K_j served in a j -busy period ($E(K_j)$ is finite if and only if $\rho < 1$), and let

$$a_j = \frac{b_j E(K_j)}{\sum_{i=1}^{\infty} b_i E(K_i)}, \quad (3)$$

which can be viewed as a “weighted version” of b_j (a more precise interpretation of this will be given in Section 3). Note that if we denote by K the number of customers served in a “typical” busy period (for a fixed D), then

$$E(K) = \sum_{i=1}^{\infty} b_i E(K_i), \quad (4)$$

the denominator of Eq. (3).

Let the notation $(X | B)$ denote a conditional random variable X given the occurrence of event B , and let $=^d$ denote equality in distribution. Define, for $1 \leq k \leq j$,

$$R_{kj} =^d \left(\sum_{i=1}^k S_i \mid \sum_{i=1}^{j-1} S_i \leq D, \sum_{i=1}^j S_i > D \right), \quad (5)$$

where $\sum_{i=1}^{j-1} S_i$ is interpreted as zero when $j = 1$; in particular, we note that, in a j -busy period, the total service requirement of all waiting customers immediately after server

activation is distributed as R_{jj} . The distribution function of R_{kj} , $1 \leq k \leq j$, will be denoted by $G_{kj}(\cdot)$.

Clearly, if $D = 0$ in our model, then we have a *standard GI/G/1* queue. We shall denote by $H(\cdot)$ the waiting-time distribution of a randomly-selected arriving customer in this corresponding *GI/G/1* queue. For $t \geq 0$, denote by $\mathbf{N} \equiv \{N(t), t \geq 0\}$ a “delayed” renewal process (e.g., Ross [17, p. 74]) whose first interevent time is distributed as T , a generic interarrival time in our model, and the subsequent ones as I , a generic idle period in the corresponding standard *GI/G/1* queue; i.e., let

$$N(t) = \begin{cases} 0, & \text{if } T > t, \\ 1 + \sup\{n \geq 0 : \sum_{i=1}^n I_i \leq t - T\}, & \text{if } T \leq t. \end{cases} \quad (6)$$

Also, denote by $m(t) \equiv E[N(t)]$, $t \geq 0$, the delayed renewal function associated with N , and, for a given random variable X , define an associated distribution function $\Psi_X(\cdot)$ by

$$\Psi_X(x) \equiv 1 - \frac{E[m((X-x)^+)]}{E[m(X)]}, \quad x \geq 0 \quad (7)$$

(see Niu [14, Eq. (12)], with $m_D(\cdot)$ replaced by $m(\cdot)$ here), where $y^+ \equiv \max(0, y)$ for any real number y . Then, it is shown in Niu [14, p. 165, Corollary 1] that

$$H(x) = \sum_{i=0}^{\infty} \{1 - E[m(S)]\}^i \{E[m(S)]\}^i \Psi_S^{[i]}(x), \quad x \geq 0. \quad (8)$$

We are now ready for the statement of our main result.

Theorem 1 *The waiting-time distribution of a randomly-selected arriving customer in the GI/G/1 queue under the D-policy is given, for $x \geq 0$, by*

$$P\{W \leq x\} = \sum_{j=1}^{\infty} a_j \left\{ \frac{j}{E(K_j)} \left[\frac{1}{j} \sum_{k=1}^j F^{[j-k]} * G_{k-1,j}(x) \right] + \left[1 - \frac{j}{E(K_j)} \right] \Psi_j * H(x) \right\}, \quad (9)$$

where $\Psi_j(\cdot) \equiv \Psi_{R_{jj}}(\cdot)$ and the symbol $*$ denotes a convolution.

Our proof of Theorem 1, which will be given in Section 3, is based on the following ideas: (i) classify customers according to the *type* of busy period in which they arrive; (ii) study the waiting times of customers in different types of busy periods; and (iii) obtain the waiting-time distribution of a randomly-selected arriving customer, whom we shall call the “test customer,” by “conditioning” on the type of busy period to which he belongs.

An important observation concerning Theorem 1 is that, apart from b_j and $G_{k-1,j}(\cdot)$, which are, from Eqs. (2) and (5), known in principle, the complete evaluation of the right-hand side of Eq. (9) requires as input $E(K_j)$, $\Psi_j(\cdot)$, and $H(\cdot)$, all of which can be linked eventually to knowledge of the distribution of I : From Eqs. (8) and (7), we have that both $H(\cdot)$ and $\Psi_j(\cdot)$ for $j \geq 1$ depend on $m(\cdot)$. In Section 3, we will prove that

$$E(K_j) = j + \frac{E[m(R_{jj})]}{1 - E[m(S)]}, \quad j \geq 1, \quad (10)$$

which implies that $E(K_j)$ for $j \geq 1$ also depend on $m(\cdot)$. Since $m(t) = \sum_{i=0}^{\infty} F * A^{[i]}(t)$ for $t \geq 0$ (see Eq. (6)), where $A(\cdot)$ denotes the distribution of I , we see that, indeed, $A(\cdot)$ is the fundamental piece of input information needed for the evaluation of $P\{W \leq x\}$; this, we believe, is a very interesting *structural* insight of Theorem 1.

In the remainder of this section, we summarize our results for two special cases, Poisson arrivals and exponential services, for which Eq. (9) can be evaluated explicitly; in particular, we give, for both cases, explicit formulas for the expected waiting time $E(W)$. Before proceeding, we note that, after substituting Eqs. (3) and (4), Eq. (9) can be rearranged into the following more compact form:

$$P\{W \leq x\} = \frac{1}{E(K)} \sum_{j=1}^{\infty} b_j \left\{ \sum_{k=1}^j F^{[j-k]} * G_{k-1,j}(x) + [E(K_j) - j] \Psi_j * H(x) \right\}. \quad (11)$$

2.1 Poisson Arrivals

We assume here that $F(\cdot)$ is the exponential distribution and $G(\cdot)$ is arbitrary. Under this assumption, we have $I =^d T$. Therefore, $m(t) = \lambda t$, implying that $E[m(S)] = \rho$ and $\Psi_S(\cdot) = G_e(\cdot)$ (see Eq. (7)), where

$$G_e(x) \equiv \mu \int_0^x [1 - G(y)] dy, \quad x \geq 0, \quad (12)$$

the forward-recurrence-time distribution of a service time. It follows that Eq. (8) simplifies to the well-known formula (e.g., Cooper [6, p. 217, Eq. (8.40)])

$$H(x) = \sum_{i=0}^{\infty} (1 - \rho) \rho^i G_e^{[i]}(x), \quad x \geq 0. \quad (13)$$

Next, from Eq. (10), we have

$$E(K_j) = j + \frac{\lambda E(R_{jj})}{1 - \rho}. \quad (14)$$

To calculate $E(R_{jj})$, we note that $P\{R_{jj} > x\} = 1$ for $0 \leq x < D$ and that, for $x \geq D$,

$$\begin{aligned}
P\{R_{jj} > x\} &= P\left\{\sum_{i=1}^j S_i > x \mid \sum_{i=1}^{j-1} S_i \leq D, \sum_{i=1}^j S_i > D\right\} \\
&= P\left\{\sum_{i=1}^j S_i > x, \sum_{i=1}^{j-1} S_i \leq D\right\} / P\left\{\sum_{i=1}^{j-1} S_i \leq D, \sum_{i=1}^j S_i > D\right\} \quad (15) \\
&= \frac{1}{b_j} \int_0^D [1 - G(x - y)] dG^{[j-1]}(y),
\end{aligned}$$

where the numerator of the second expression was evaluated by conditioning on $\sum_{i=1}^{j-1} S_i$; therefore,

$$\begin{aligned}
E(R_{jj}) &= \int_0^\infty P\{R_{jj} > y\} dy \\
&= D + \frac{1}{b_j} \int_D^\infty \int_0^D [1 - G(y - t)] dG^{[j-1]}(t) dy. \quad (16)
\end{aligned}$$

Substitution of Eq. (16) into the right-hand side of Eq. (14) now yields a formula for $E(K_j)$.

Finally, we have, similar to Eq. (12),

$$\Psi_j(x) = \frac{1}{E(R_{jj})} \int_0^x P\{R_{jj} > y\} dy, \quad x \geq 0, \quad (17)$$

which can be calculated using Eqs. (15) and (16); and Eqs. (13), (17), (14), and (4) together evaluate Eq. (11) to an explicit expression (which we omit) for the waiting-time distribution in the $M/G/1$ queue under the D -policy.

We can also derive an explicit formula for the expected waiting time: Let $\mathbf{N}_S \equiv \{N_S(t), t \geq 0\}$ be the renewal process associated with successive service times, defined by $N_S(t) = \sup\{n \geq 0 : \sum_{i=1}^n S_i \leq t\}$; and let $m_S(t)$ denote its corresponding renewal function. Then, we prove the following theorem in Section 3.

Theorem 2 *For the $M/G/1$ queue under the D -policy, we have*

$$E(W) = \frac{1}{m_S(D) + 1} \left\{ \frac{1 - \rho}{\lambda} \int_0^D [1 + m_S(D - t)] dm_S(t) + \int_0^D t dm_S(t) \right\} + E(W_M), \quad (18)$$

where $E(W_M)$, given by the well-known formula (e.g., Cooper [6, p. 189, Eq. (4.2)])

$$E(W_M) = \frac{\lambda E(S^2)}{2(1 - \rho)}, \quad (19)$$

is the expected waiting time of customers in the corresponding standard $M/G/1$ queue. (Equation (18) is closely related to a formula for the average number of customers in system, involving the second moment of $N(D)$, given in Rubin and Zhang [18, p. 333, Eq. (4.7)].)

2.2 Exponential Services

We assume here that $G(\cdot)$ is the exponential distribution and $F(\cdot)$ is now arbitrary. Under this assumption, it is easily seen that

$$R_{jj} = {}^d D + S \quad (20)$$

for every $j \geq 1$. Since the number of customers who arrive after server activation in a j -busy period is determined by R_{jj} , it follows from Eq. (20) that this number, distributed as $K_j - j$, does not depend on j ; therefore, we will denote its generic version by K_A . It also follows from Eq. (20) that the form of $\Psi_j(\cdot)$ (see Eq. (7)) is independent of j , and we will denote a generic $\Psi_j(\cdot)$ by $\hat{\Psi}(\cdot)$. With these observations, Eq. (11) simplifies to

$$P\{W \leq x\} = \frac{1}{E(K)} \left\{ \sum_{j=1}^{\infty} b_j \sum_{k=1}^j F^{[j-k]} * G_{k-1,j}(x) + E(K_A) \hat{\Psi} * H(x) \right\}, \quad (21)$$

where

$$E(K) = \mu D + 1 + E(K_A), \quad (22)$$

since $b_j = [(\mu D)^{j-1}/(j-1)!]e^{-\mu D}$ for $j \geq 1$.

It is well known (e.g., Cooper [6, pp. 270–272, Eqs. (14.22) and (14.11)], or from Eq. (8) here) that, for the $GI/M/1$ queue,

$$H(x) = 1 - \omega e^{-(1-\omega)\mu x}, \quad x \geq 0, \quad (23)$$

where ω is the smallest root of the equation

$$\omega = \int_0^{\infty} e^{-(1-\omega)\mu y} dF(y). \quad (24)$$

Therefore, in view of Eqs. (21) and (22), it remains for us to determine $E(K_A)$ and $\hat{\Psi}(\cdot)$ (we note in passing that in this case, $G_{k-1,j}(\cdot)$, $1 \leq k \leq j$, reduces to the Beta distribution).

For the standard $GI/M/1$ queue, it is known [15, p. 285, Eq. (16)] that

$$A(x) = \mu \int_0^{\infty} e^{-(1-\omega)\mu y} [F(y+x) - F(y)] dy, \quad x \geq 0. \quad (25)$$

Denote by $m_I(\cdot)$ the renewal function of the renewal process $\mathbf{N}_I \equiv \{N_I(t), t \geq 0\}$ associated with successive idle periods I_i for $i \geq 1$ (defined similar to \mathbf{N}_S). Then, we prove in Section 3 that

$$E(K_A) = \frac{\omega + m_I(D)}{1 - \omega} \quad (26)$$

and

$$\hat{\Psi}(x) = \begin{cases} 1 - [\omega + m_I(D-x)] [\omega + m_I(D)]^{-1}, & \text{if } 0 \leq x \leq D, \\ 1 - \omega e^{-\mu(x-D)} [\omega + m_I(D)]^{-1}, & \text{if } x > D. \end{cases} \quad (27)$$

Substitution of Eqs. (22), (23), (26), and (27) into the right-hand side of Eq. (21) now yields an explicit formula (which we omit) for the waiting-time distribution in the $GI/M/1$ queue under the D -policy.

Finally, we also prove the following theorem in Section 3.

Theorem 3 *For the $GI/M/1$ queue under the D -policy, we have*

$$E(W) = \frac{1}{E(K)} \left[\frac{D}{\rho} + \frac{1+\rho}{2\rho} \mu D^2 + \frac{1}{1-\omega} \int_0^D m_I(D-t) dt \right] + E(W_{GI}), \quad (28)$$

where $E(W_{GI})$, given by

$$E(W_{GI}) = \frac{\omega}{(1-\omega)\mu} \quad (29)$$

(see Eq. (23)), is the expected waiting time of customers in the standard $GI/M/1$ queue.

3 Proofs

We begin with Theorem 1. We will first argue that a_j , given by Eq. (3), is the “probability” for the test customer to belong to a j -busy period: Observe that (i) the distribution of the total number of customers served in a j -busy period depends explicitly on j , and (ii) it is more likely for the test customer to arrive in “longer” busy periods (the classical “inspection paradox”); i.e., the probability for the test customer to belong to a j -busy period does *not*, in general, equal b_j (see Eq. (2)). To account for this length-biasing effect, we shall interpret the *number* of customers served in a j -busy period as a “sojourn in state j ” in a *discrete*-time (or an ordinal) semi-Markov process. Then, it is easily seen that under this interpretation a_j is the proportion of “time epochs” (or indices) this semi-Markov process spends in “state j ,” and $E(K_j)$ is the expected sojourn time in state j . Hence, an application of, for example, Theorem 4.8.3 in Ross [17] yields Eq. (3).

For a *fixed* $j \geq 1$, we now study waiting times of arriving customers in j -busy periods. Our key idea is to further classify these customers into two types: Those who arrive when the server is inactive are said to be of type one; those when the server is active, type two. Let π_{j1} and π_{j2} be the “probabilities” for the test customer to be of type one and of type two, respectively. Then, since there are exactly j type-one customers in each j -busy period, we have, from a standard renewal reward argument (e.g., Ross [17, p. 78, Theorem 3.6.1]),

$$\pi_{j1} = \frac{j}{E(K_j)} = 1 - \pi_{j2} \quad (30)$$

We shall analyze in detail waiting times of these two types of customers separately.

Suppose first that the test customer is of type one and that he is the k^{th} arriving customer in a j -busy period, where $1 \leq k \leq j$. Then, since the server remains inactive until the arrival epoch of the j^{th} customer, the waiting time of the test customer equals the (independent) sum of (i) the elapsed time between his arrival epoch and that of the j^{th} customer and (ii) the total service requirement of the $k - 1$ customers who arrived before him. The random variable in (i) is, of course, the sum of $j - k$ (independent) random variables each distributed as $F(\cdot)$; therefore, it follows the distribution function $F^{[j-k]}(\cdot)$ (when $k = j$, $F^{[j-k]}(x)$ equals 1 for $x \geq 0$, and 0 otherwise). The random variable in (ii) is easily seen to have the representation $(\sum_{i=1}^{k-1} S_i | \sum_{i=1}^{j-1} S_i \leq D, \sum_{i=1}^j S_i > D)$; therefore, its distribution function is denoted by $G_{k-1,j}(\cdot)$ (see Eq. (5)). Finally, since a type-one customer is equally likely to be any one of the first j arrivals in the busy period, we conclude that the waiting-time distribution of a randomly-selected type-one customer is given, for $x \geq 0$, by

$$\frac{1}{j} \sum_{k=1}^j F^{[j-k]} * G_{k-1,j}(x). \quad (31)$$

We next focus on type-two customers. The idea here is to view the first j customers (all of whom are type one) in a j -busy period as a *single* “supercustomer,” with a (total) service requirement R_{jj} , and to imagine that the server is activated by the arrival of this supercustomer. With this view, it is easily seen that the waiting times of type-two customers are identical to those of customers who request *ordinary* services, with distribution function $G(\cdot)$, in a corresponding $GI/G/1$ queue with an *exceptional* service, distributed as R_{jj} , at the beginning of each busy period. This motivates the following key result, which, we believe, is of independent interest.

Proposition 1 *The waiting time W_o of a randomly-selected “ordinary customer” (i.e., those who do not initiate a busy period) in an exceptional-first-service $GI/G/1$ queue has distribution function*

$$P\{W_o \leq x\} = \Psi_{\hat{S}} * H(x), \quad x \geq 0, \quad (32)$$

where \hat{S} denotes an exceptional service time.

Proof Observe that, under the standard FIFO (first-in-first-out) queue discipline, the waiting time of a customer in the exceptional-first-service $GI/G/1$ model equals the workload found by him at his arrival epoch. Since the workload is, at any time epoch, invariant under all work-conserving queue disciplines, we will analyze the workload in this model, first at *all* customer arrival epochs and then at *ordinary*-customer arrival epochs, under the simplifying assumption that the queue discipline is preemptive-resume LIFO (last-in-first-out).

Our argument will be based on results in Niu [14]; in what follows, familiarity with that paper will be assumed.

Under the preemptive-resume LIFO queue discipline, let α_0 be the (long-run) proportion of arrivals finding the system empty; and, for $j \geq 1$ and \mathbf{x} , where \mathbf{x} is a nonnegative j -vector, let $\alpha_j(\mathbf{x})$ be the proportion of (all) customers who find, on arrival, j customers in the system with their respective remaining service requirements, arranged in the same order as their arrival times, greater than x_1, x_2, \dots, x_j , the components of \mathbf{x} . Since the workload at an arrival epoch is the sum of the remaining service times of all waiting customers, α_0 and $\alpha_j(\mathbf{x})$ for $j \geq 1$ completely determine the workload distribution as found by arrivals. We will, therefore, derive a set of formulas for α_0 and $\alpha_j(\mathbf{x})$ for $j \geq 1$. As in Niu [14], these formulas will be expressed in terms of the delayed renewal function $m(\cdot)$ defined in Section 2.

Observe that having an exceptional service at the beginning of each busy period is the same as sampling the service time from a different distribution whenever an arrival finds the system empty. Therefore, our model is a special case of the one discussed in Section 4.2 of Niu [14], where general *state-dependent* interarrival and service times are allowed; in particular, Theorem 5 there directly applies here. More specifically, we have

$$\alpha_0 = \{1 - E[m(S)]\} \{1 - E[m(S)] + E[m(\hat{S})]\}^{-1} \quad (33)$$

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

$$\alpha_j(\mathbf{x}) = \alpha_0 [Em(\hat{S})] \{E[m(S)]\}^{j-1} [1 - \Psi_{\hat{S}}(x_1)] \prod_{i=2}^j [1 - \Psi_S(x_i)] , \quad (34)$$

where S denotes an ordinary service time and an ill-defined product is interpreted as 1. (Equations (33) and (34) are derived recursively from a minor modification of Lemma 1 in Niu [14] that replaces S by \hat{S} on the right-hand side of Eq. (13) there when $j = 1$; see the proof of Theorem 1 there.)

We now focus on *ordinary* customers. For $j \geq 1$, let $\hat{\alpha}_j(\mathbf{x})$ be defined similar to $\alpha_j(\mathbf{x})$ but with the average taken over ordinary customers (an ordinary customer always finds *at least* one other customer in the system). Since $1 - \alpha_0$ is the proportion of arrivals that are ordinary, we clearly have

$$\hat{\alpha}_j(\mathbf{x}) = \frac{\alpha_j(\mathbf{x})}{1 - \alpha_0}, \quad (35)$$

a “relative proportion” (see Niu and Cooper [16, Section 3.2], for related discussions). Substitution of Eqs. (33) and (34) into Eq. (35) yields, after a little bit of algebra,

$$\hat{\alpha}_j(\mathbf{x}) = [1 - \Psi_{\hat{S}}(x_1)] \tilde{\alpha}_{j-1}(x_2, \dots, x_j), \quad (36)$$

where we have defined

$$\tilde{\alpha}_{j-1}(x_2, \dots, x_j) \equiv \{1 - E[m(S)]\} \{E[m(S)]\}^{j-1} [1 - \Psi_S(x_i)] \quad (37)$$

(the argument of $\tilde{\alpha}_{j-1}(x_2, \dots, x_j)$ is vacuous when $j = 1$). Comparison of the right-hand side of Eq. (37) with Theorem 1 in Niu [14] then shows that as we vary j (from 1 to ∞) and x_2, \dots, x_j , Eq. (37) describes precisely the workload distribution as observed by *all* arriving customers in a standard $GI/G/1$ queue. Hence, Eq. (32) is a consequence of Eq. (36). \square

We are finally in position to put the pieces together:

Proof of Theorem 1 By “conditioning” first on the type of busy period in which the test customer arrives and then on whether he is a type one or a type two, it is easily seen that Eq. (9) is a consequence of Eqs. (3), (30), (31), and (32) (with $\hat{S} =^d R_{jj}$). (See Niu and Cooper [16, Sections 3.2 and 4.1] for a rigorous justification of such conditioning arguments.) \square

The remainder of this section will be devoted to the proofs of Eq. (10), Theorem 2, Eqs. (26) and (27), and Theorem 3, in that order.

Proof of Eq. (10) Again, we shall view all type-one customers in a busy period as a single supercustomer (whose arrival activates the server). With this view, it is easily seen that

$$K_j =^d j + K_o(\hat{S}), \quad (38)$$

where $K_o(\hat{S})$ denotes the number of *ordinary* customers served in an exceptional-first-service $GI/G/1$ busy period in which the first service $\hat{S} =^d R_{jj}$. We will, therefore, consider such a busy period. Moreover, we will again assume, for convenience, that the queue discipline is preemptive-resume LIFO, and adapt related arguments given in Niu [14].

If we follow the experience of the supercustomer over time, then, from Figures 1 and 2, and associated discussions, in Niu [14, p. 164], we see that (i) each arrival finding the supercustomer *in service* generates an “interruption” of \hat{S} , during which the number of customers served is distributed as that of a *standard* $GI/G/1$ busy period, which we will denote by \tilde{K} , and (ii) the number of such interruptions is distributed as $N(\hat{S})$ (see Eq. (6)). Therefore (similar to Eq. (28) in Niu [14]),

$$K_o(\hat{S}) =^d \sum_{i=1}^{N(\hat{S})} \tilde{K}_i, \quad (39)$$

where the \tilde{K}_i 's denote i.i.d. versions of \tilde{K} . An application of Wald's identity to Eq. (39) (note that $N(\hat{S})$ and the \tilde{K}_i 's are *dependent*; see Eq. (30) in Niu [14]) then yields

$$E[K_o(\hat{S})] = E[N(\hat{S})] E[\tilde{K}]. \quad (40)$$

From Eq. (32) in Niu [14], we have

$$E(\tilde{K}) = \{1 - E[m(S)]\}^{-1}. \quad (41)$$

Therefore, Eq. (10) is a consequence of Eqs. (38), (40), and (41). \square

For the proofs of Theorems 2 and 3, we will need several preliminary lemmas, all of which are valid in the general $GI/G/1$ setting.

Lemma 1

$$\sum_{j=1}^{\infty} b_j \sum_{k=1}^j \int_0^{\infty} x dF^{[j-k]}(x) = \frac{1}{\lambda} \int_0^D [1 + m_S(D-t)] dm_S(t). \quad (42)$$

Proof Let T_i , $i \geq 1$, be the interarrival time between the i^{th} and the $(i+1)^{\text{th}}$ customer in a busy period; then, we obviously have $\int_0^{\infty} x dF^{[j-k]}(x) = E(\sum_{i=k}^{j-1} T_i)$ for $1 \leq k \leq j$, with the sum interpreted as zero when $k = j$. From this and the fact that $b_j = P\{N_S(D) = j-1\}$ for $j \geq 1$, we see that the left-hand side of Eq. (42) can be written as $E[\sum_{k=1}^{N_S(D)+1} \sum_{i=k}^{N_S(D)} T_i]$. Since $\sum_{i=k}^{N_S(D)} T_i = 0$ when $k = N_S(D) + 1$, we then have

$$\sum_{j=1}^{\infty} b_j \sum_{k=1}^j \int_0^{\infty} x dF^{[j-k]}(x) = E \left[\sum_{k=1}^{N_S(D)} \sum_{i=k}^{N_S(D)} T_i \right]. \quad (43)$$

We will, therefore, complete the proof by showing that the right-hand side of Eq. (43) evaluates to that of Eq. (42).

For $k \geq 1$, define the indicator random variable

$$Y_k = \begin{cases} 1 & \text{if } N_S(D) \geq k, \\ 0 & \text{if } N_S(D) < k; \end{cases} \quad (44)$$

then,

$$\begin{aligned} E \left[\sum_{k=1}^{N_S(D)} \sum_{i=k}^{N_S(D)} T_i \right] &= E \left[\sum_{k=1}^{\infty} Y_k \sum_{i=k}^{N_S(D)} T_i \right] \\ &= \sum_{k=1}^{\infty} E \left[Y_k \sum_{i=k}^{N_S(D)} T_i \right]. \end{aligned} \quad (45)$$

Next, by conditioning on $\sum_{i=1}^k S_i$, we have

$$\begin{aligned} E \left[Y_k \sum_{i=k}^{N_S(D)} T_i \right] &= \int_0^\infty E \left[Y_k \sum_{i=k}^{N_S(D)} T_i \mid \sum_{i=1}^k S_i = t \right] dP \left\{ \sum_{i=1}^k S_i \leq t \right\} \\ &= \int_0^D E \left[\sum_{i=k}^{N_S(D)} T_i \mid \sum_{i=1}^k S_i = t \right] dG^{[k]}(t), \end{aligned} \quad (46)$$

where the second equality is due to the fact that $Y_k = 1$ if and only if $\sum_{i=1}^k S_i \leq D$ (see Eq. (44)). Now, observe that if $\sum_{i=1}^k S_i = t$ with $t < D$, then the k^{th} renewal in the renewal process \mathbf{N}_S occurs at time t and the number of renewals in the interval $(t, D]$ is distributed as $N_S(D - t)$. It follows that

$$\left[\sum_{i=k}^{N_S(D)} T_i \mid \sum_{i=1}^k S_i = t \right] =^d \sum_{i=1}^{1+N_S(D-t)} T_i;$$

therefore

$$\begin{aligned} E \left[\sum_{i=k}^{N_S(D)} T_i \mid \sum_{i=1}^k S_i = t \right] &= E \left[\sum_{i=1}^{1+N_S(D-t)} T_i \right] \\ &= [1 + m_S(D - t)] \frac{1}{\lambda}, \end{aligned} \quad (47)$$

since $E(T_i) = 1/\lambda$ for all $i \geq 1$ and the T_i 's are independent of \mathbf{N}_S . Finally, from Eqs. (45), (46), (47), and the well-known fact that $m_S(t) = \sum_{k=1}^\infty G^{[k]}(t)$, we have

$$\begin{aligned} E \left[\sum_{k=1}^{N_S(D)} \sum_{i=k}^{N_S(D)} T_i \right] &= \sum_{k=1}^\infty \int_0^D [1 + m_S(D - t)] \frac{1}{\lambda} dG^{[k]}(t) \\ &= \frac{1}{\lambda} \int_0^D [1 + m_S(D - t)] dm_S(t), \end{aligned}$$

the right-hand side of Eq. (42). \square

Our proof of Lemma 1 actually justifies an interesting term-by-term interpretation for the right-hand side of Eq. (42): For $0 \leq t \leq D$, (i) $dm_S(t)$ is the ‘‘probability’’ of having a renewal at time t (e.g., Ross [17, p. 66, Remark (2)]), (ii) $1 + m_S(D - t)$ is the (conditional) expected number of renewals, including the one at time t , in the interval $[t, D]$, and (iii) $1/\lambda$, the expected length of one interarrival interval, is the (conditional) expected contribution *at time t* to the right-hand side of Eq. (42) from each of the renewals in the interval $[t, D]$. Variants of this interpretation, which is reminiscent of the Lebesgue integral, will be used again without further comment in the proofs of the next two lemmas, to shorten the arguments.

Lemma 2

$$\sum_{j=1}^{\infty} b_j \sum_{k=1}^j \int_0^{\infty} x dG_{k-1,j}(x) = \int_0^D t dm_S(t). \quad (48)$$

Proof From Eqs. (2) and (5), it is easily seen that the left-hand side of Eq. (48) can be written as $E[\sum_{k=1}^{N_S(D)+1} \sum_{i=1}^{k-1} S_i]$, which, since $\sum_{i=1}^{k-1} S_i = 0$ when $k = 1$, further simplifies to $E[\sum_{k=2}^{N_S(D)+1} \sum_{i=1}^{k-1} S_i]$. A change of summation index then leads to

$$\sum_{j=1}^{\infty} b_j \sum_{k=1}^j \int_0^{\infty} x dG_{k-1,j}(x) = E \left[\sum_{k=1}^{N_S(D)} \sum_{i=1}^k S_i \right]. \quad (49)$$

Now, for $0 \leq t \leq D$, if a renewal in the process \mathbf{N}_S occurs at time t (i.e., $\sum_{i=1}^k S_i = t$ for some k , which occurs with “probability” $dm_S(t)$), then the (conditional) expected contribution at time t to the sum $\sum_{k=1}^{N_S(D)} \sum_{i=1}^k S_i$ on the right-hand side of Eq. (49) is precisely t . Integrating $t dm_S(t)$ from 0 to D then leads to

$$E \left[\sum_{k=1}^{N_S(D)} \sum_{i=1}^k S_i \right] = \int_0^D t dm_S(t),$$

the right-hand side of Eq. (48). □

Let $R \equiv \sum_{i=1}^{N_S(D)+1} S_i$ be the sum of the required service times of all waiting customers immediately after server activation. Then, an application of Wald’s identity yields immediately that

$$E(R) = [m_S(D) + 1] \frac{1}{\mu} \quad (50)$$

In the next lemma, we evaluate the second moment of R .

Lemma 3

$$E(R^2) = [m_S(D) + 1]E(S^2) + \frac{2}{\mu} \int_0^D t dm_S(t). \quad (51)$$

Proof This formula was established in Balachandran and Tijms [4, p. 1074, Eq. (b)], by solving two renewal-type equations. We give an alternative proof that is somewhat more constructive.

From the definition of R , we have

$$E(R^2) = E \left[\left(\sum_{i=1}^{N_S(D)+1} S_i \right)^2 \right]$$

$$= E \left[\sum_{i=1}^{N_S(D)+1} S_i^2 \right] + 2 E \left[\sum_{1 \leq i < j \leq N_S(D)+1} S_i S_j \right] \quad (52)$$

Since $E[\sum_{i=1}^{N_S(D)+1} S_i^2] = E[m_S(D)+1] E(S^2)$ (similar to Eq. (50)), comparison of the right-hand sides of Eqs. (51) and (52) shows that the proof will be complete if we can establish that

$$E \left[\sum_{1 \leq i < j \leq N_S(D)+1} S_i S_j \right] = \frac{1}{\mu} \int_0^D t dm_S(t) . \quad (53)$$

To prove Eq. (53), observe that

$$\sum_{1 \leq i < j \leq N_S(D)+1} S_i S_j = \sum_{j=1}^{N_S(D)} S_{j+1} \sum_{i=1}^j S_i . \quad (54)$$

Therefore, if, for $0 \leq t \leq D$, a renewal in the process \mathbf{N}_S occurs at time t (i.e., $\sum_{i=1}^j S_i = t$ for some j , which occurs with “probability” $dm_S(t)$), then, since S_{j+1} , with mean $1/\mu$, is independent of $\sum_{i=1}^j S_i$ for every j , the (conditional) expected contribution at time t to the right-hand side of Eq. (54) equals $(1/\mu)t$. Integrating $(1/\mu)t dm_S(t)$ from 0 to D then yields the right-hand side of Eq. (53). \square

We are now in position to prove Theorem 2.

Proof of Theorem 2 From Eq. (11), we have

$$E(W) = \frac{1}{E(K)} \sum_{j=1}^{\infty} b_j \left\{ \sum_{k=1}^j \left[\int_0^{\infty} x dF^{[j-k]}(x) + \int_0^{\infty} x dG_{k-1,j}(x) \right] + [E(K_j) - j] \left[\int_0^{\infty} x d\Psi_j(x) + \int_0^{\infty} x dH(x) \right] \right\} . \quad (55)$$

After substituting Eq. (13), $\int_0^{\infty} x d\Psi_j(x) = E(R_{jj}^2)/[2E(R_{jj})]$ (see Eq. (17)), $\sum_{j=1}^{\infty} b_j E(R_{jj}^2) = E(R^2)$, and $\sum_{j=1}^{\infty} b_j E(R_{jj}) = E(R)$, Eq. (55) further simplifies to

$$E(W) = \frac{1}{E(K)} \left\{ \sum_{j=1}^{\infty} b_j \sum_{k=1}^j \int_0^{\infty} x dF^{[j-k]}(x) + \sum_{j=1}^{\infty} b_j \int_0^{\infty} x dG_{k-1,j}(x) + \frac{\lambda E(R^2)}{2(1-\rho)} + \frac{\lambda E(R)}{1-\rho} E(W_M) \right\} . \quad (56)$$

From Eqs. (4) and (10), $\sum_{j=1}^{\infty} j b_j = m_S(D) + 1$ (see Eq. (2)), $\sum_{j=1}^{\infty} b_j E(R_{jj}) = E(R)$, and Eq. (50), it is easily shown that

$$E(K) = \frac{[m_S(D) + 1]}{1 - \rho}. \quad (57)$$

Finally, from Eqs. (42), (48), (51), (19), (50), and (57), it is straightforward to verify that Eq. (56) reduces to Eq. (18). \square

Proof of Eq. (26) Observe that $K_A =^d K_o(\hat{S})$ (see Eq. (38)), where \hat{S} now consists of two (independent) “phases,” D and S (see Eq. (20)). We will, therefore, follow the argument given in the proof of Eq. (10).

From Remarks 5 and 6 in Niu [14], we have (and it is well known) that, for $GI/M/1$,

$$E(\tilde{K}) = \frac{1}{1 - \omega}. \quad (58)$$

Comparison of Eq. (26) to Eqs. (40) and (58) shows that we need to establish

$$E[N(\hat{S})] = \omega + m_I(D) \quad (59)$$

To prove Eq. (59), we shall, for convenience, let the two phases of \hat{S} go into service in the order S first, D second; and denote by C_S and C_D the respective numbers of interruptions during the two phases in that order (C_S and C_D are dependent, in general).

Clearly, we have $C_S =^d N(S)$ (similar to (ii) in the proof of Eq. (10)). Since the duration of S is distributed as an *ordinary* service time, we see that the stochastic law that governs in the time interval that begins with the arrival of the supercustomer and ends when the supercustomer has expended S amount of time *in service* (i.e., when the S -phase of \hat{S} is completed) is identical to that of a standard $GI/M/1$ busy period, for which we have, by substituting Eq. (58) into the right-hand side of Eq. (32) in Niu [14], $E[N(S)] = E[m(S)] = \omega$; therefore, $E(C_S) = \omega$. The same observation also shows that the elapsed time from the initiation of the D -phase of \hat{S} to the first arrival after that time epoch is stochastically identical to I , with distribution Eq. (25); and, since the subsequent (independent) elapsed times between interruptions, if any, during the D -phase of \hat{S} are also distributed as I , this implies that $C_D =^d N_I(D)$, the “nondelayed” version of $N(D)$ (see Eq. (6)). Hence, $E(C_D) = E[N_I(D)] = m_I(D)$; and Eq. (59) is now a consequence of $E[N(\hat{S})] = E(C_S) + E(C_D)$. \square

Proof of Eq. (27) From Eqs. (7) and (20), we have

$$\hat{\Psi}(x) = 1 - \frac{E[m((\hat{S} - x)^+)]}{E[m(\hat{S})]}, \quad x \geq 0, \quad (60)$$

where \hat{S} is as defined in the Proof of Eq. (26). Observe that $E[m((\hat{S} - x)^+)]$ specializes to $E[m(\hat{S})]$ if $x = 0$. We will, therefore, evaluate $E[m((\hat{S} - x)^+)]$ for $x \geq 0$. Consider two cases.

Case 1: $0 \leq x \leq D$. We have $(\hat{S} - x)^+ =^d (D - x) + S$, which is of the same form as \hat{S} . Therefore, Eq. (59) applies with $D - x$ replacing D on its right-hand side, yielding

$$E[m((\hat{S} - x)^+)] = \omega + m_I(D - x). \quad (61)$$

Case 2: $x > D$. Exploiting the memoryless property, we have

$$(\hat{S} - x)^+ =^d \begin{cases} S & \text{with probability } e^{-\mu(x-D)}, \\ 0 & \text{with probability } 1 - e^{-\mu(x-D)}. \end{cases}$$

It follows that

$$E[m((\hat{S} - x)^+)] = E[m(S)]e^{-\mu(x-D)} = \omega e^{-\mu(x-D)}. \quad (62)$$

Substitution of Eqs. (61) and (62) into Eq. (60) now yields Eq. (27). \square

Proof of Theorem 3 From Eq. (27), it is straightforward to show that

$$\begin{aligned} \int_0^\infty x d\hat{\Psi}(x) &= \int_0^\infty [1 - \hat{\Psi}(x)] dx \\ &= [\omega + m_I(D)]^{-1} \left[\omega \left(D + \frac{1}{\mu} \right) + \int_0^D m_I(D - x) dx \right]. \end{aligned} \quad (63)$$

Substitution of Eqs. (42), (48), (26) (recall that $E(K_j) - j = E(K_A)$), (63), and (29) into Eq. (55) yields, after a little bit of algebra,

$$\begin{aligned} E(W) &= \frac{1}{E(K)} \left\{ \frac{1}{\lambda} \int_0^D [1 + m_S(D - t)] dm_S(t) \right. \\ &\quad \left. + \int_0^D t dm_S(t) + \frac{1}{1 - \omega} \int_0^D m_I(D - t) dt \right\} + E(W_{GI}), \end{aligned}$$

which, since $m_S(t) = \mu t$, reduces straightforwardly to Eq. (28). \square

4 Comments and Additions

4.1 If $D = 0$, then the right-hand side of Eq. (11) reduces, as it should, to Eq. (8), the waiting-time distribution of customers in a standard $GI/G/1$ queue. This is verified as follows: Clearly, in this case, every busy period is a 1-busy period (i.e., the server is activated immediately after the arrival of the first customer). It follows that (i) $b_1 = 1$, (ii) $1/E(K_1) = 1/E(K) = 1 - E[m(S)]$ (see Eq. (41)), and (iii) $\Psi_1(\cdot) = \Psi_S(\cdot)$ (see Eq. (7)). Substitution of (i), (ii), and (iii) into the right-hand side of Eq. (11) now yields, after a little bit of algebra, Eq. (8).

4.2 The term

$$\frac{1}{E(K)} \sum_{j=1}^{\infty} b_j \sum_{k=1}^j F^{[j-k]} * G_{k-1,j}(x) \quad (64)$$

on the right-hand side of Eq. (11) can be given a direct interpretation: Observe first that the limiting proportion $P\{W \leq x\}$ (see Eq. (1)) can be decomposed into the sum of two proportions $\nu_1[0, x]$ and $\nu_2[0, x]$, representing respective contributions to the right-hand side of (1) from type-one and type-two customers; we then claim that Eq. (64) equals $\nu_1[0, x]$. To see this, let $M(x)$ be the number of type-one customers with a waiting time less than or equal to x in a typical busy cycle. Then, since successive busy cycles are iid, we have (e.g., Theorem 3.6.1 of Ross [17]) that, w.p.1, $\nu_1[0, x] = E[M(x)]/E(K)$. Now, if a busy period is of type j , then, from the discussion leading up to Eq. (31), we see that the k^{th} ($1 \leq k \leq j$) type-one customer in the busy period will have a waiting time less than or equal to x and hence be “counted” in $M(x)$ with probability $F^{[j-k]} * G_{k-1,j}(x)$. Therefore, by conditioning on the total number of type-one customers in a busy period, we have

$$E[M(x)] = \sum_{j=1}^{\infty} b_j \sum_{k=1}^j F^{[j-k]} * G_{k-1,j}(x),$$

establishing our claim.

That the other term in Eq. (11) equals $\nu_2[0, x]$, a fact that is more difficult to establish, is a consequence of our Proposition 1.

4.3 Proposition 1 resembles but is different from decomposition results for vacation models, as discussed in Doshi [7,8] and other references therein. Our constructive arguments can, in fact, be adapted to study vacation models as well; and this will be discussed in a future paper.

By conditioning on whether or not an arrival finds the system empty (see Eq. (33)), Proposition 1 also yields a representation for the waiting time of *all* customers in the exceptional-first-service $GI/G/1$ queue. If, for example, the arrival process is Poisson, then this representation can be specialized, as in Section 2.1, to obtain explicit results that are transform-free counterparts of (some of the) results obtained earlier by Welch [20] and by Avi-Itzhak, Maxwell, and Miller [1], among others; similar results for the exceptional-first-service $GI/M/1$ queue can also be obtained, using our arguments in Section 2.2.

4.4 Our method can also be used to derive the waiting-time distribution of customers in the $GI/G/1$ queue *under the N -policy*. To see this, notice first that under the N -policy, every busy period is an N -busy period; therefore, there is no length-biasing effect when a customer is selected at random. Next, we again classify customers as either type one or type two, according to whether or not the server is active at the time of their arrival. Lastly, for

$j \geq 1$, let $\tilde{R}_j \equiv \sum_{i=1}^j S_i$, and observe that the service requirement of the “supercustomer” in every busy period is distributed as \tilde{R}_N (as opposed to R_{NN} , defined in Eq. (5)). An argument similar to that leading to Theorem 1 then yields the following result.

Theorem 4 *The distribution of waiting time \tilde{W} of a randomly-selected arriving customer in the GI/G/1 queue under the N-policy is given, for $x \geq 0$, by*

$$P\{\tilde{W} \leq x\} = \frac{N}{E(\tilde{K}_N)} \frac{1}{N} \sum_{k=1}^N F^{[N-k]} * G^{[k-1]}(x) + \left[1 - \frac{N}{E(\tilde{K}_N)}\right] \tilde{\Psi} * H(x), \quad (65)$$

where $E(\tilde{K}_N)$ denotes the expected number of customers served in an N -busy period and $\tilde{\Psi}(\cdot) \equiv \Psi_{\tilde{R}_N}(\cdot)$.

If the arrival process is Poisson, then it can be shown that $E(\tilde{K}_N) = N/(1-\rho)$ (similar to Eq. (57)) and that $\tilde{\Psi}(\cdot)$ simplifies, as in Eq. (17), to the forward-recurrence-time distribution of \tilde{R}_N ; with these (and Eq. (13)), the representation (65) easily specializes to a transform-free formula. Corresponding transform formulas, derived by different methods, can be found, for example, in Neuts [13, p. 35, Eqs. (1.7.9) and (1.7.10)] (also, see Shanthikumar [19]); the explicit term-by-term interpretations of our formula appear to be new.

In the remainder of this subsection, we will consider the case of the GI/M/1 queue under the N -policy, for which our results appear to be new.

Again, we only need to determine $E(\tilde{K}_N)$ and $\tilde{\Psi}(\cdot)$, and we will adapt the proofs for Eqs. (10), (26), and (27). After replacing \hat{S} with \tilde{R}_N in Eq. (40), substituting Eq. (58), and accounting for the N initial customers, we see that

$$E(\tilde{K}_N) = N + \frac{E[N(\tilde{R}_N)]}{1 - \omega}. \quad (66)$$

Since \tilde{R}_N consists of N i.i.d. exponential phases, it is also easily seen that Eq. (62) generalizes, by conditioning on the number of Poisson events at rate μ in an interval of duration x , to

$$E[N((\tilde{R}_N - x)^+)] = \sum_{i=0}^{N-1} \frac{(\mu x)^i}{i!} e^{-\mu x} E[N(\tilde{R}_{N-i})], \quad x \geq 0, \quad (67)$$

which is needed in $\tilde{\Psi}(\cdot)$ (defined by replacing \hat{S} with \tilde{R}_N in Eq. (60)). We will, therefore, determine $E[N(\tilde{R}_j)]$ for $1 \leq j \leq N$.

For an arbitrary $j \geq 1$, we have, similar to Eq. (59),

$$E[N(\tilde{R}_j)] = \omega + E[N_I(\tilde{R}_{j-1})]. \quad (68)$$

Clearly, $E[N_I(\tilde{R}_0)] = 0$. We will evaluate $E[N_I(\tilde{R}_k)]$ for $k \geq 1$ recursively.

Denote by $C(I)$ the number of Poisson events at rate μ that occur during a generic idle period I . Our recursion will involve the distribution of $C(I)$, which we give in the next lemma.

Lemma 4 *For $i \geq 0$, we have*

$$P\{C(I) = i\} = \omega^{-(i+1)} \left[\omega - \sum_{n=0}^i \omega^n \delta_n \right], \quad (69)$$

where, for $n \geq 0$,

$$\delta_n \equiv \int_0^\infty \frac{(\mu y)^n}{n!} e^{-\mu y} dF(y). \quad (70)$$

Proof From Eq. (25), we have, after a change of variable,

$$dA(x) = \mu \int_x^\infty e^{-(1-\omega)\mu(z-x)} dF(z) dx.$$

Therefore,

$$P\{C(I) = i\} = \int_0^\infty \frac{(\mu x)^i}{i!} e^{-\mu x} \mu \int_x^\infty e^{-(1-\omega)\mu(z-x)} dF(z) dx,$$

which, after an interchange of the order of integration and some algebra (which we omit), leads to

$$P\{C(I) = i\} = \omega^{-(i+1)} \int_0^\infty e^{-(1-\omega)\mu z} \left[1 - \sum_{n=0}^i \frac{(\omega\mu z)^n}{n!} e^{-\omega\mu z} \right] dF(z).$$

Upon substitution of Eqs. (24) and (70), the last expression simplifies to Eq. (69). \square

We are now ready for the recursion.

Lemma 5 *For $k \geq 1$, we have*

$$E[N_I(\tilde{R}_k)] = \left(\frac{\omega}{\delta_0} - 1 \right) + \frac{\omega}{\delta_0} \sum_{i=1}^{k-1} P\{C(I) = i\} \{1 + E[N_I(\tilde{R}_{k-i})]\}. \quad (71)$$

Proof Initiate, simultaneously at time 0, two independent counting processes; and let one of these be Poisson at rate μ and the other N_I . Thus, the processes monitor, respectively, the numbers of “potential” phase-completions and interruptions (of and during \tilde{R}_k) over time.

Let Q be the smallest index i , $i \geq 1$, for which there is at least one Poisson event in the time interval $(\sum_{n=1}^{i-1} I_n, \sum_{n=1}^i I_n)$; then, from the independent-increments property of

the Poisson process, it is easily seen that Q has the geometric distribution with success probability $1 - P\{C(I) = 0\}$, which, from Eq. (69), equals δ_0/ω .

From the definition of Q , we see that the total number of interruptions, if any, *prior* to the first phase-completion of \tilde{R}_k , which occurs somewhere inside the time interval

$$\left(\sum_{n=1}^{Q-1} I_n, \sum_{n=1}^Q I_n \right)$$

is distributed as $Q - 1$. Denote by L the number of additional interruptions *after* the first phase-completion of \tilde{R}_k ; then, since $E(Q - 1) = (\omega/\delta_0) - 1$, we see from Eq. (71) that the proof will be complete if we can establish that

$$E(L) = \frac{\omega}{\delta_0} \sum_{i=1}^{k-1} P\{C(I) = i\} \{1 + E[N_I(\tilde{R}_{k-i})]\}. \quad (72)$$

Denote by J the number of Poisson events in the time interval $(\sum_{n=1}^{Q-1} I_n, \sum_{n=1}^Q I_n)$; then, we clearly have $J =^d (C(I) \mid C(I) > 0)$. Observe that additional interruptions after the first phase-completion of \tilde{R}_k can occur only if $1 \leq J \leq k - 1$. Now, if $1 \leq J \leq k - 1$, then a “genuine” interruption necessarily occurs at time $\sum_{n=1}^Q I_n$; moreover, from that time epoch onward, both the Poisson process and the process \mathbf{N}_I regenerate themselves. Therefore, by conditioning on J and noting that the number of remaining phases in \tilde{R}_k after time $\sum_{n=1}^Q I_n$ equals $k - J$, we obtain that $E(L) = \sum_{i=1}^{k-1} P\{C(I) = i \mid C(I) > 0\} \{1 + E[N_I(\tilde{R}_{k-i})]\}$, which, since $P\{C(I) = i \mid C(I) > 0\} = P\{C(I) = i\}(\delta_0/\omega)^{-1}$ for $i \geq 1$, reduces to Eq. (73). \square

Finally, combining Eqs. (67) and (68) and Lemmas 4 and 5 yields $\tilde{\Psi}(\cdot)$, which together with Eqs. (66) and (23) reduces Eq. (65) to an explicit formula (which we omit) for the waiting-time distribution in the $GI/M/1$ queue under the N -policy.

4.5 As noted in Section 1, our results in this paper can be applied to study, in the $GI/G/1$ setting, optimization models in which the threshold D is a controllable parameter. This is currently under investigation and will be reported in a future paper.

References

- [1] Avi-Itzhak, B., Maxwell, W. L., & Miller, L. W. (1965). Queueing with alternating priorities. *Operations Research* 13: 306–318.
- [2] Balachandran, K. R. (1971). Queue length dependent priority queues. *Management Science* 17: 463–471.

- [3] Balachandran, K. R. (1973). Control policies for a single server system. *Management Science* 19: 1013–1018.
- [4] Balachandran, K. R. & Tijms, H. (1975). On the D-policy for the M/G/1 queue. *Management Science* 21: 1073–1076.
- [5] Boxma, O. J. (1976). Note on a control problem of Balachandran and Tijms. *Management Science* 22: 916–917.
- [6] Cooper, R. B. (1981). *Introduction to queueing theory*, 2nd ed. Amsterdam: Elsevier North-Holland. (First edition, 1972, Macmillan. Republished, 1990, by CEEPress, The George Washington University, Washington, DC.)
- [7] Doshi, B. T. (1985). A note on stochastic decomposition in a GI/G/1 queue with vacations or set-up times. *Journal of Applied Probability* 22: 419–428.
- [8] Doshi, B. T. (1990). Single-server queues with vacations. In H. Takagi (ed.), *Stochastic analysis of computer and communication systems*. Amsterdam: Elsevier North-Holland.
- [9] Heyman, D. P. (1968). Optimal operating policies for M/G/1 queueing systems. *Operations Research* 16: 362–382.
- [10] Heyman, D. P. (1977). The T-policy for the M/G/1 queue. *Management Science* 23: 775–778.
- [11] Li, J. (1989). *Sample-average analyses of some generalizations of the M/G/1 queue*. Doctoral Dissertation, University of Texas at Dallas, Richardson.
- [12] Neuts, M. F. (1986). Generalizations of the Pollaczek-Khinchin integral equations in the theory of queues. *Advances in Applied Probability* 18: 952–990.
- [13] Neuts, M. F. (1989). *Structured stochastic matrices of M/G/1 type and their applications*. New York: Marcel Dekker.
- [14] Niu, S.-C. (1988). Representing workloads in GI/G/1 queues through the preemptive-resume LIFO queue discipline. *Queueing Systems* 3: 157–178.
- [15] Niu, S.-C. & Cooper, R. B. (1989). Duality and other results for M/G/1 and GI/M/1 queues, via a new ballot theorem. *Mathematics of Operations Research* 14: 281–293.
- [16] Niu, S.-C. & Cooper, R. B. (1989). Transform-free analysis of M/G/1/K and related queues. *Mathematics of Operations Research* (to appear, 1993).
- [17] Ross, S. M. (1983). *Stochastic processes*. New York: Wiley.

- [18] Rubin, I. & Zhang, Z. (1987). Switch on policies for communications and queueing systems. In *Proceedings of the Third International Conference on Data Communication*. Amsterdam: Elsevier North-Holland, pp. 329–339.
- [19] Shanthikumar, J. G. (1981). Optimal control of an M/G/1 priority queue via N-control. *American Journal of Mathematical Management Science* 1: 191–212.
- [20] Welch, P. D. (1964). On a generalized M/G/1 queueing process in which the first customer of each busy period receives exceptional service. *Operations Research* 12: 736–752.
- [21] Yadin, M. & Naor, P. (1963). Queueing systems with a removable service station. *Operational Research Quarterly* 14: 393–405.