Probability in the Engineering and Informational Sciences, **22**, 2008, 355–371. Printed in the U.S.A.
doi:10.1017/S0269964808000211

BALKING AND RENEGING IN *M/G/s* SYSTEMS EXACT ANALYSIS AND APPROXIMATIONS

Liqiang Liu and Vidyadhar G. Kulkarni
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599–3180
E-mail: vkulkarni@email.unc.edu

We consider the virtual queuing time (vqt, also known as work-in-system, or virtual-delay) process in an *M/G/s* queue with impatient customers. We focus on the vqt-based balking model and relate it to reneging behavior of impatient customers in terms of the steady-state distribution of the vqt process. We construct a single-server system, analyze its operating characteristics, and use them to approximate the multiserver system. We give both analytical results and numerical examples. We conduct simulation to assess the accuracy of the approximation.

1. INTRODUCTION

Motivated by analyzing the call center operations, we consider an *M/G/s* queuing system with impatient customers. The customers arrive according to a Poisson process with rate λ and request i.i.d. (independent and identically distributed) service times with a general distribution. There are $s \geq 1$ servers in the system available to serve the customers. All servers are identical and unit-rate; that is, each server is capable of processing one unit of service requirement per unit time.

An important aspect modeling call centers is the impatience behavior of the customers. Two common modes in which customers display their impatience are balking and reneging. A call-in customer who cannot be helped immediately by a human server might be told how long a wait he/she faces before an operator is available. Then the customer might hang up (i.e., balk) or decide to hold. This is the balking behavior: A customer refuses to enter the queue if the wait is too long. On the other hand, a customer who is waiting for an operator might hang up (i.e., renege) before getting

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

served if the wait in line becomes too long. This is the reneging behavior. Of course, there can be a combination of the two. It is acknowledged that customers' impatience is significant in practice and modeling call centers (cf. Koole and Mandelbaum [10], Garnett Mandelbaum, and Reiman [5], Whitt [21]).

To incorporate the customers' impatience in the queuing model, we use a balking rule corresponding to the balking example given earlier. Before stating the balking rule, we define the *virtual queuing time* (vqt) in the system. The vqt at time t in the system, denoted by $W(t)$, is the queuing time (i.e., time spent in the system before commencing service) that would be experienced if a customer joins the system at time t. The process $\{W(t), t \geq 0\}$ is referred to as the vqt process. The queuing system with balking based on the vqt works as follows. A customer arriving at time t joins the system if and only if $W(t-) \leq b$, where b is a fixed nonnegative constant. The balking customers (i.e., customers who do not join) are lost forever. The entering customers wait in an infinite-capacity FCFS (first come–first serve) queue until a server is available and leave when the service completes. The vqt process governed by such a vqt-based balking rule is the main focus of this article. Of course, such a model can also arise as a result of a threshold-type admission control policy.

The vqt process is also known as work-in-system or virtual-delay process, which is introduced in Beneš [1] and Takács [18]. See Heyman and Sobel [7, pp. 383–390] for details. Although the vqt process is introduced in the context of the vqt-based balking model described earlier, this process also plays an essential role in the analysis of the models with reneging customers. According to the definition of the vqt, if the system has customers who eventually will renege without being served, then the service times for those customers are not included in the vqt. The model with reneging customers can be analyzed via a closely related vqt-based balking model. Such an idea appears in Tijms [20, pp. 318–322]. We discuss the connection between balking and reneging in detail in Section 4.

The reneging version of the model we consider has been studied by Gnedenko and Kovalenko [6] under the name "systems with limited waiting time." They considered exponential service times to obtain a multidimensional Markov process for the number of busy servers and workload in each server. They derived a system of integro-differential equations for the limiting joint distribution and give explicit solution. They gave formulas for the loss probability and average queuing time. They also gave the limiting distribution of the vqt process. However, as Boots and Tijms [2] noted, the results in [6] are quite technical and not generally applicable. Instead, they gave an alternative formula for the loss probability as a function of the tail probability of the stationary vqt process in a corresponding queue with no impatience. They proved that their formula is exact in the *M/M/s* case and can be used as a heuristic for the *M/G/s* case. A severe restriction is that the formula is valid only when the traffic intensity is less than 1, which is not required for the reneging queue to be stable. The method we use in this article overcomes the preceding drawbacks and can be easily extended to the general case. Although we are unable to give the joint distribution for the workload and busy servers, we do not lose much since many common performance measures can be derived directly from the limiting distribution of the vqt process.

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

Even in the absence of the balking behavior, the *M/G/s* queuing system is notorious for its complexity, which forbids analytical solutions. Analytical results are available for only a few special cases, whereas a handful of approximations for the limiting analysis have been proposed in the past decades. In this article, we focus on *system approximations* (i.e., approximations that take the results from an exact analysis of a simpler system as approximations of the true operating characteristics of the original system). Although the approximate methods vary by motivations and the techniques that are used, it turns out that all results can be viewed as the so-called "systems interpolation" (i.e., some mixture of the known analytical results for a few special cases, such as *M/M/s*, $M/E_k/s$, *M/D/s*, and $M/G/\infty$). See Kimura [9] for details. We cannot find any system approximations of the *M/G/s* queuing system with impatient customers in the literature.

To develop a system approximation for the multiserver system with impatient customers, we borrow a simple idea used by Lee and Longton [12], Takács [18, p. 160], Newell [16, p. 86], Hokstad [8], Nozaki and Ross [17], Tijms [19], and Miyazawa [14]. In brief, the idea is to treat the s-server system as an $M/G/\infty$ system (or $M/G/s-1$ loss system) when some servers are idle and as an $M/G/1$ system when all servers are busy. Since balking and/or reneging can only happen when all servers are busy, we can easily extend this idea by introducing customer impatience to the $M/G/1$ system that approximates the original system during the period when all servers are busy. Using this idea, we construct a single-server system whose operating characteristics approximates those of the *M/G/s* queuing system with balking based on vqt. The approximation is exact when $G = M$, $b = 0$, or $s = 1$. The exact analysis of the approximate system follows the same line that of as Liu and Kulkarni [13], where the authors solve the $s = 1$ case. In the rest of this article, we begin with analyzing the $G = M$ case in Section 2. In Section 3 we consider general service times and propose an approximate system. We conclude with numerical results (Sec. 5).

2. THE *M/M/s* BALKING MODEL

In this section we consider the case in which the service requirement has an exponential distribution with mean $1/\mu$. The arrival process is Poisson with rate λ. This is an *M/M/s* FCFS system with balking based on vqt. At the arrival epoch, an arriving customer joins the queue if and only if he/she observes that the vqt is no more than a fixed amount b.

Let $N(t)$ be the number of customers in the system at time t. The definition of $W(t)$ implies that $W(t) = 0$ if and only if $N(t) \leq s - 1$. If the $N = \{N(t), t \geq 0\}$ process undergoes a transition from state $s - 1$ to s at time t, then there is a jump in the $W = \{W(t), t \geq 0\}$ process at the same time. This is illustrated in Figure 1, which shows the sample path of a two-server system. Note that in the sample path shown there, the customer who arrives at time T_3 balks. At time T_1 (and T_4), the number of customer in the system increases from one to two. At the same time, the vqt jumps

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

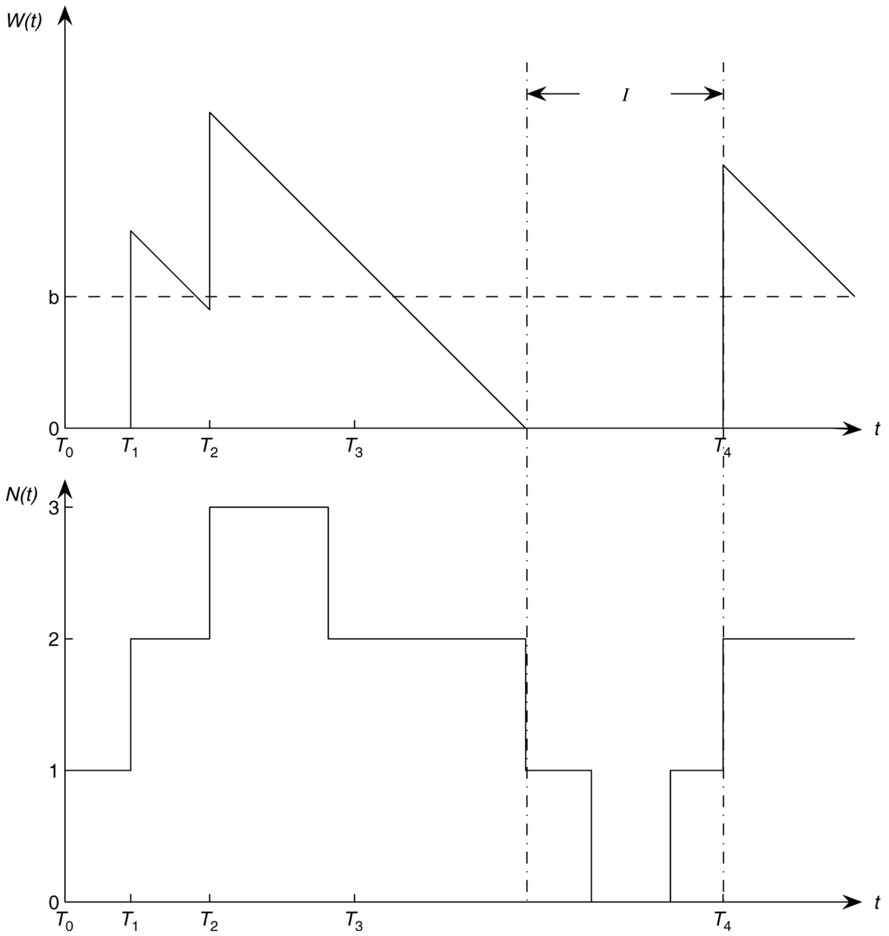

Figure 1. A sample path of $W(t)$ and $N(t)$.

from zero to a positive number. It is clear that the W process finishes a regenerative cycle from T_1 to T_4. It is easy to see that the size of the jumps are i.i.d. $\exp(s\mu)$.

Let I be a generic random variable representing the idle period defined as the interval of time during which $W(t) = 0$. In other words, let t_1 be the service completion epoch such that $N(t_1-) = s$. Then $I = \min_{t>0}\{t : W(t_1 + t) > 0\}$ or $I = \min_{t>0}\{t : N(t_1 + t) = s\}$.

Theorem 1: *The expected length of I is given by*

$$E(I) = \frac{1 - p_s}{s\mu p_s}, \tag{2.1}$$

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

where

$$p_s = \frac{(\lambda/\mu)^s}{s! \sum_{i=0}^{s}[(\lambda/\mu)^i/i!]}.$$

PROOF: Consider a standard *M/M/s/s* system with arrival rate λ and mean service times $1/\mu$. Obviously, the time between two consecutive periods when the system is full has the same distribution as I. The expected length of each system-full period is clearly $1/(s\mu)$. From the theory of alternating renewal process (ARP), we get

$$\frac{1/(s\mu)}{1/(s\mu) + \mathrm{E}(I)} = p_s,$$

where p_s is the probability that the *M/M/s/s* system is full (cf. [11]). The identity in the theorem follows. ■

Consider the regenerative cycle from T_1 to T_4 as shown in Figure 1. The cycle consists of a busy period (where $W(t) > 0$) and an idle period (where $W(t) = 0$). It is easy to see that the evolution of the W process during the busy period is stochastically identical to that in a single-server balking system (cf. [13]) with an arrival rate of λ and i.i.d. exp($s\mu$) service times. This observation motivates the following single-server system with i.i.d. exp($s\mu$) service times. Let $\tilde{W}(t)$ be the vqt and $\tilde{N}(t)$ be the number of customers at time t in this system. The same balking rule applies; that is, a customer arriving at time t enters iff $\tilde{W}(t) \le b$. Customers arrive as a Poisson process with arrival rate $\tilde{\lambda}(t)$ depending on $\tilde{N}(t)$ as follows:

$$\tilde{\lambda}(t) = \begin{cases} \gamma & \text{if } \tilde{N}(t) = 0 \\ \lambda & \text{otherwise,} \end{cases}$$

where γ is defined as $1/\mathrm{E}(I)$.

Denote the limiting cumulative distribution functions (c.d.f.) of the W process and $\tilde{W}$ process as follows:

$$F(x) = \lim_{t\to\infty} \Pr\{W(t) \le x\}, \qquad x \ge 0;$$

$$\tilde{F}(x) = \lim_{t\to\infty} \Pr\{\tilde{W}(t) \le x\}, \qquad x \ge 0.$$

Let

$$F(0) = c, \qquad \tilde{F}(0) = \tilde{c}.$$

Let

$$f(x) = \frac{dF(x)}{dx}, \qquad \tilde{f}(x) = \frac{d\tilde{F}(x)}{dx}, \qquad x > 0,$$

be the probability distribution functions (p.d.f.s).

Notice that for a single-server system, the vqt is equal to the workload of the system. The steady-state distribution of the $\tilde{W}$ process has been extensively studied.

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

We apply the same method used in Liu and Kulkarni [13] and the results in Theorem 2 and 3, omitting the proofs. Theorem 2 gives the balance equation and normalizing equation satisfied by $\tilde{f}(x)$. Theorem 3 gives the expression of $\tilde{f}(x)$ explicitly by solving the equations.

THEOREM 2: *The equilibrium p.d.f. $\tilde{f}(x)$ of the $\tilde{W}$ process satisfies*

$$\tilde{f}(x) = \lambda \int_0^{x\wedge b} \tilde{f}(u)e^{-s\mu(x-u)}\,du + \tilde{c}\gamma e^{-s\mu x}, \tag{2.2a}$$

$$\int_0^{\infty} \tilde{f}(x)\,dx + \tilde{c} = 1, \tag{2.2b}$$

where $x \wedge b = \min(x, b)$.

Let $\rho = \lambda/s\mu$ be the traffic intensity.

THEOREM 3: *The equilibrium p.d.f. of the $\tilde{W}$ process is*

$$\tilde{f}(x) = \begin{cases} \tilde{c}\gamma e^{-(s\mu-\lambda)x} & \text{if } 0 < x < b \\ \tilde{c}\gamma e^{\lambda b}e^{-s\mu x} & \text{if } x \geq b, \end{cases} \tag{2.3}$$

where

$$\tilde{c} = \begin{cases} \left[\gamma\left(\dfrac{1}{s\mu-\lambda} - e^{-(s\mu-\lambda)b}\dfrac{\lambda}{(s\mu-\lambda)s\mu}\right) + 1\right]^{-1} & \text{if } \rho \neq 1 \\ \dfrac{\lambda}{\lambda+\gamma+\lambda\gamma b} & \text{if } \rho = 1. \end{cases} \tag{2.4}$$

The following theorem gives the limiting distribution of the W process via the single-server system we construct.

THEOREM 4: *The W process and $\tilde{W}$ process have same limiting distribution; that is,*

$$F(x) = \tilde{F}(x), \quad x \geq 0.$$

PROOF: Let $B(t) = 1$ if $W(t) > 0$ and $B(t) = 0$ if $W(t) = 0$. Then $B = \{B(t), t \geq 0\}$ is an ARP. Define $\tilde{B}(t)$ associated with $\tilde{W}(t)$ in the same fashion; then $\tilde{B} = \{\tilde{B}(t), t \geq 0\}$ is also an ARP. Since the expected up and down times in the B process and the $\tilde{B}$ process are the same, we get

$$\lim_{t\to\infty} \Pr\{B(t) = i\} = \lim_{t\to\infty} \Pr\{\tilde{B}(t) = i\}, \qquad i = 0, 1.$$

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

It is easy to see that the sample paths of the W process and the $\tilde{W}$ process over the busy periods are stochastically identical. Hence, we get

$$\lim_{t\to\infty} \Pr\{W(t) \le x|B(t) = 1\} = \lim_{t\to\infty} \Pr\{\tilde{W}(t) \le x|\tilde{B}(t) = 1\}.$$

Now,

$$\begin{aligned} F(x) &= \lim_{t\to\infty} \Pr\{W(t) \le x\} \\ &= \lim_{t\to\infty} \Pr\{W(t) \le x|B(t) = 1\} \Pr\{B(t) = 1\} + \lim_{t\to\infty} \Pr\{B(t) = 0\} \\ &= \lim_{t\to\infty} \Pr\{\tilde{W}(t) \le x|\tilde{B}(t) = 1\} \Pr\{\tilde{B}(t) = 1\} + \lim_{t\to\infty} \Pr\{\tilde{B}(t) = 0\} \\ &= \lim_{t\to\infty} \Pr\{\tilde{W}(t) \le x\} \\ &= \tilde{F}(x). \end{aligned}$$

This proves the theorem. ■

Other performance measures of interest in the *M/M/s* balking system are computed directly based on the above results. Let r be the probability that a customer balks or the balking rate. Then

$$r = \int_b^\infty f(x)\,dx = c\gamma \frac{e^{-(s\mu-\lambda)b}}{s\mu}.$$

Define the queuing time of any arriving customer to be the time from the arrival epoch to the service starting epoch if he/she joins and zero if he/she balks. Let w be the long-run average queuing time for all customers. Then

$$w = \int_0^b x f(x)\,dx = \begin{cases} \dfrac{c\gamma[1-(s\mu-\lambda)be^{-(s\mu-\lambda)b} - e^{-(s\mu-\lambda)b}]}{(s\mu-\lambda)^2} & \text{if } \rho \ne 1 \\ \dfrac{c\gamma b^2}{2} & \text{if } \rho = 1. \end{cases}$$

It is clear that the long-run average queuing time or the expected queuing time for the entering customers is $w' = w/(1-r)$. It can be verified by straightforward algebra that the results given in this section are consistent with the corresponding ones in [6] and [2].[1] However, unlike in [2], we do not need to assume that $\rho < 1$, and our results are more explicit than those in [6].

3. THE *M/G/s* BALKING MODEL

In this section we extend the *M/M/s* balking model in Section 2 to an *M/G/s* balking model. All settings for the *M/M/s* balking model are unchanged except that we assume

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

the service times are i.i.d. with a general distribution with mean $1/\mu$ and complementary c.d.f. $G(x)$. We use the subscript G in our notations for the general service time case. The definitions correspond to those for the exponential case and are omitted. In order to follow the analysis in Section 2, we need $\mathrm{E}(I_G)$ and the distribution of the size of the jumps in the W_G process.

To compute $\mathrm{E}(I_G)$, we consider a standard *M/G/s/s* system with arrival rate λ and mean service time $1/\mu$. At each departure epoch of a customer who leaves behind $s-1$ customers in the system, the remaining service time in the busy servers might not have the same joint distribution as that of the *M/G/s* system with balking. Ignoring this fact, we use the expected length of the time between two consecutive periods when the *M/G/s/s* system is full as an approximation of $\mathrm{E}(I_G)$. We know that, in equilibrium, an arriving customer to the *M/G/s/s* system with $s-1$ busy servers sees the remaining service times in the busy servers as having an i.i.d. distribution with complementary c.d.f. $G_e(x)$, which is the associated complementary equilibrium distribution of $G(x)$ defined by

$$G_e(x) = \mu \int_x^\infty G(u)\, du.$$

Then the length of the period during which the *M/G/s/s* system is full is

$$\min\{R_1, R_2, \ldots, R_{s-1}, S\},$$

where $\{R_i, i = 1, 2, \ldots, s-1\}$ are i.i.d. random variables with complementary c.d.f. $G_e(x)$. Notice that

$$dG_e(x) = -\mu G(x)\, dx;$$

then the expected length of this period is

$$\int_0^\infty G_e^{s-1}(x) G(x)\, dx = \frac{1}{s\mu}. \qquad \textbf{(3.1)}$$

Using the same method in the proof of Theorem 1, we get that the expected duration of the interval during which the *M/G/s/s* system is *not* full is given by

$$\frac{1-p_s}{s\mu p_s}. \qquad \textbf{(3.2)}$$

This is the same as in the *M/M/s/s* system. We use this expression as an approximation for $\mathrm{E}(I_G)$.

The distribution of the size of the jumps is more complicated in the system with general service times. Suppose the kth jump in the W_G process occurs at time T_k. Let J_k be the size of this jump. Unfortunately, $\{J_i, i = 1, 2, \ldots\}$ are neither independent nor identically distributed in general and this makes the model intractable. In the next paragraph, we explain the source of this intractability and it can be skipped in first reading without affecting the flow of the material.

The jump size J_k in the W_G process at time T_k is the minimum of this customer's service time and the remaining service times of all other customers in

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

service at time $T_k + W_G(T_k)$ (when this customer begins the service). Thus, the distribution of the jump sizes are not even identical in general. Equivalently, let us regard the *M/G/s* system as s parallel single-server queues that operate as follows. Denote the workload at time t in the ith queue as $W'_i(t)$, $i = 1, 2, \ldots, s$, and let $W'(t) = (W'_1(t), W'_2(t), \ldots, W'_s(t))$ (cf. [3]). Every entering customer is routed to the queue with the least workload. Then

$$W_G(t) = \min\{W'_1(t), W'_2(t), \ldots, W'_s(t)\}. \tag{3.3}$$

Suppose the customer who arrives at time T_k with service time S is routed to the ith server, which has the least workload; then

$$\begin{aligned} J_k &= W_G(T_k+) - W_G(T_k-) \\ &= \min\{W'_1(T_k-), \ldots, W'_i(T_k-) + S, \ldots, W'_s(T_k-)\} - W'_i(T_k-). \end{aligned} \tag{3.4}$$

Clearly, the distribution of J_k is determined by the distribution of $W'(T_k)$ and the distribution of S. The dependence of J_k and $W'(T_k)$ causes great complexity in the analysis of the model and makes it intractable.

As an approximation, we assume that $\{J_i : W(T_i-) > 0\}$ are i.i.d., with $\dot{J}$ being the generic jump size, and $\{J_i : W(T_i-) = 0\}$ are i.i.d., with $\ddot{J}$ being the corresponding generic jump size. One principle of choosing the distribution for $\dot{J}$ and $\ddot{J}$ is to preserve the traffic intensity [i.e., $\rho = \lambda/(s\mu)$]; that is, keep the mean of $\dot{J}$ and $\ddot{J}$ to be $1/(s\mu)$. We assume $\mathrm{E}(\dot{J}) = \mathrm{E}(\ddot{J}) = 1/(s\mu)$ in the rest of this article. We consider two possibilities. The first choice is $\bar{S} = S/s$. It is easy to see that $\mathrm{E}(\bar{S}) = 1/(s\mu)$ in this case. The second choice is $\hat{S} = \min\{R_1, R_2, \ldots, R_{s-1}, S\}$. This is motivated by the renewal-theoretic result that in steady state, the remaining services times in the busy servers are independent random variables with common complementary c.d.f. G_e (cf. [18, p. 161]). From (3.2), we see that $\mathrm{E}(\hat{S}) = 1/(s\mu)$.

Analogous to the exponential case, we consider the $\tilde{W}_G$ process of the following single-server system. Customers arrive according to a Poisson process with arrival rate $\tilde{\lambda}(t)$ depending on $\tilde{N}_G(t)$ as follows:

$$\tilde{\lambda}(t) = \begin{cases} \gamma & \text{if } \tilde{N}_G(t) = 0 \\ \lambda & \text{otherwise,} \end{cases}$$

where γ is defined to be $s\mu p_s/(1 - p_s)$, which is the approximation for $1/\mathrm{E}(I_G)$. Let $G_{\dot{J}}(x)$ and $G_{\ddot{J}}(x)$ be the complementary c.d.f.'s of $\dot{J}$ and $\ddot{J}$, respectively. Service times of the customers who enter a nonempty system are i.i.d. with common complementary c.d.f. $G_{\dot{J}}(x)$. Service times of the customers who enter an empty system are i.i.d. with common complementary c.d.f. $G_{\ddot{J}}(x)$. A customer arriving at time t enters iff $\tilde{W}_G(t) \leq b$. We use the expression in (3.2) as an approximation of $\mathrm{E}(I_G)$. We approximate the distributions of $\{J_i\}$ by $\dot{J}$ and $\ddot{J}$. Moreover, the conditions for Theorem 4 do not hold in general. Therefore, the vqt process of the single-server model we construct approximates that of the *M/G/s* balking model [i.e., $F_G(x) \approx \tilde{F}_G(x), x \geq 0$]. It is worth noting that the approximation is exact in the following three cases: (1) The service times are exponential; (2) the balking threshold b is zero (when the system reduces to an *M/G/s/s* system); $\ddot{J} = \hat{S}$; (3) $s = 1$.

The following theorem is the general service time version of Theorem 2. It distinguishes the two appearances of the jump size distribution in the balance equation.

THEOREM 5: *The steady-state p.d.f.* $\tilde{f}_G(x)$ *of the* $\tilde{W}_G$ *process satisfies*

$$\tilde{f}_G(x) = \lambda \int_0^{x \wedge b} \tilde{f}_G(u) G_J(x-u)\, du + \tilde{c}_G \gamma G_{\ddot{J}}(x), \tag{3.5a}$$

$$\int_0^{\infty} \tilde{f}_G(x)\, dx + \tilde{c}_G = 1, \tag{3.5b}$$

where $x \wedge b = \min(x, b)$.

Notice that the first term on the right-hand side of (3.5a) is just the convolution of $\tilde{f}_G(x)$ and $G_J(x)$ multiplied by λ, when $x \wedge b$ is replaced by x. Let $f_1(x)$ be the solution to

$$f_1(x) = \lambda \int_0^x f_1(u) G_J(x-u)\, du + G_{\ddot{J}}(x), \qquad x \geq 0. \tag{3.6}$$

Let

$$f_2(x) = \lambda \int_0^b f_1(u) G_J(x-u)\, du + G_{\ddot{J}}(x), \qquad x \geq b. \tag{3.7}$$

The solution to (3.5a and 3.5b) is given in the following theorem.

THEOREM 6: *The solution to* (3.5) *is*

$$\tilde{f}_G(x) = \begin{cases} \tilde{c}_G \gamma f_1(x) & \text{if } x < b \\ \tilde{c}_G \gamma f_2(x) & \text{if } x \geq b, \end{cases} \tag{3.8}$$

where

$$\tilde{c}_G = \left[\gamma \int_0^b f_1(x)\, dx + \gamma \int_b^{\infty} f_2(x)\, dx + 1 \right]^{-1}. \tag{3.9}$$

PROOF: The solution is easy to verify by substitution. ■

From the above theorem, it is clear that a possible procedure to obtain $\tilde{f}_G(x)$ is to find $f_1(x)$ first, then compute $f_2(x)$ by using (3.7). By the normalizing equation (3.5b), after computing the integral we are able to compute $\tilde{c}_G$. This completes the computation of $\tilde{f}_G(x)$. Obviously, one main step is to solve (3.6) for $f_1(x)$. One method is to use the Laplace transform (LT).

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

Let $G^*_{\dot J}(\xi)$ and $G^*_{\ddot J}(\xi)$ be the LT of $G_{\dot J}(x)$ and $G_{\ddot J}(x)$, respectively; that is,

$$G^*_{\dot J}(\xi) = \int_0^\infty e^{-\xi x} G_{\dot J}(x)\,dx,$$
$$G^*_{\ddot J}(\xi) = \int_0^\infty e^{-\xi x} G_{\ddot J}(x)\,dx.$$

From (3.6), we get the LT of $f_1(x)$ (assuming its existence):

$$f_1^*(\xi) = \frac{G^*_{\ddot J}(\xi)}{1 - \lambda G^*_{\dot J}(\xi)}. \tag{3.10}$$

To continue our procedure, we need the inverse LT of $f_1^*(\xi)$. A closed-form inversion is possible if $f_1^*(\xi)$ is rational.

We can instantly obtain two interesting results from the above analysis when $b \to 0$ and $b \to \infty$. The first case is $b \to 0$. In this case, the system reduces to a normal *M/G/s/s* model. Our approximation is exact if $\ddot J = \hat S$. From Theorem 5, as $b \to 0$,

$$\tilde f_G(x) \to \tilde c_G \gamma G_{\ddot J}(x), \qquad x \ge 0,$$
$$\tilde c_G \to 1 - p_s.$$

Next, we compute $\tilde w_G$, the long-run average queuing time for all customers and $\tilde c_G$, as $b \to \infty$. The LT is convenient in this case. Using

$$\int_0^\infty f_1(x)\,dx = f_1^*(0),$$

$$G^*_{\dot J}(0) = \mathrm{E}(\dot J), \qquad G^*_{\ddot J}(0) = \mathrm{E}(\ddot J),$$

$$\tilde w_G = -\left.\frac{d\tilde f_G^*(\xi)}{d\xi}\right|_{\xi=0},$$

Theorem 6, and (3.10), we get

$$\tilde c_G \to \frac{1-\rho}{\nu + 1 - \rho}, \tag{3.11}$$

$$\tilde w_G \to \frac{\gamma[(1-\rho)\mathrm{E}(\ddot J^2) + \rho \mathrm{E}(\dot J^2)]}{2(1-\rho)(\nu + 1 - \rho)}, \tag{3.12}$$

where $\nu = \gamma/(s\mu) = p_s/(1-p_s)$.

When $b \to \infty$, the W_G process becomes the vqt process in a normal *M/G/s* system and $r_G \to 0$. For several nonexponential distributions (e.g., Erlang-k) of service time, the exact table of $E(W_G)$ (and other performance measures) is available and can be compared with $E(\tilde W_G)$ to assess the accuracy of our approximation.

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

Depending on the expression of the service time distribution and the choice of the distributions of $\dot{J}$ and $\ddot{J}$, the solution to (3.5a and 3.5b) can be obtained, in most cases, by numerical methods. We have two options for $\dot{J}$ or $\ddot{J}$, namely $\bar{S} = S/s$ and $\hat{S} = \min\{R_1, R_2, \ldots, R_{s-1}, S\}$. Let

$$G_{\bar{S}}(x) = \Pr\{\bar{S} \geq x\} = G(sx),$$
$$G^*_{\bar{S}}(\xi) = \int_0^\infty e^{-\xi x} G_{\bar{S}}(x)\, dx = \frac{1}{s} G^*(\xi/s),$$
$$G_{\hat{S}}(x) = \Pr\{\hat{S} \geq x\} = G_e^{s-1}(x) G(x),$$
$$G^*_{\hat{S}}(\xi) = \int_0^\infty e^{-\xi x} G_{\hat{S}}(x)\, dx = -\frac{1}{\mu} \int_0^\infty e^{-\xi x} G_e^{s-1}(x)\, d[G_e(x)],$$

where $G^*(\xi)$ is the LT for $G(x)$. The expressions of $G_{\bar{S}}(x)$ and $G^*_{\bar{S}}(\xi)$ are very easy to obtained once $G(x)$ and $G^*(\xi)$ are specified. However, it is hard to compute $G_{\hat{S}}(x)$ and $G^*_{\hat{S}}(\xi)$. Computing $G^*_{\hat{S}}(\xi)$ is hard even for phase-type (except exponential) service times. Since all jumps caused by the customers who see $s-1$ busy servers upon arrival are i.i.d. as $\hat{S}$, we introduce the complexity with the hope of improving accuracy. In the following subsections, we consider two possibilities in choosing $\dot{J}$ and $\ddot{J}$ and define Approximation I and Approximation II correspondingly.[2]

3.1. Approximation I: $\dot{J} = \ddot{J} = \bar{S}$

The equilibrium distribution of the W process is approximated by the solution to the following system of equations:

$$\tilde{f}_G(x) = \lambda \int_0^{x \wedge b} \tilde{f}_G(u) G_{\bar{S}}(x-u)\, du + \tilde{c}_G \gamma G_{\bar{S}}(x), \tag{3.13a}$$
$$\int_0^\infty \tilde{f}_G(x)\, dx + \tilde{c}_G = 1. \tag{3.13b}$$

When $b \to \infty$, this approximation becomes a classical *M/G/s* system approximation, which appears in [12]. It tells us that the system behaves as an *M/G/s/s* system when the vqt is zero. For vqt greater than zero, the system behaves like a busy *M/G/*1 system with service time S/s. See [6, 16–18] for details. Equation (3.12) becomes

$$\tilde{w}_G \to \frac{\gamma \mathrm{E}(S^2)}{2s^2(1-\rho)(\nu+1-\rho)}.$$

The most general case where an explicit closed-form solution to (3.13) is available is the phase-type service time. This is because $\bar{S}$ is PH(α, sM) if S is PH(α, M). The results in [13] can be easily adapted to the case we consider here. See [13] for details.

3.2. Approximation II: $\dot{J} = \bar{S}, \ddot{J} = \hat{S}$

The equilibrium distribution of the W process is approximated by the solution to the following system of equations:

$$\tilde{f}_G(x) = \lambda \int_0^{x \wedge b} \tilde{f}_G(u) G_{\bar{S}}(x-u)\, du + \tilde{c}_G \gamma G_{\hat{S}}(x), \qquad \textbf{(3.14a)}$$

$$\int_0^{\infty} \tilde{f}_G(x)\, dx + \tilde{c}_G = 1, \qquad \textbf{(3.14b)}$$

When $b \to \infty$, (3.12) becomes

$$\tilde{w}_G \to \frac{\gamma[(1-\rho)\mathrm{E}(\hat{S}^2) + \rho \mathrm{E}(\bar{S}^2)]}{2(1-\rho)(\nu+1-\rho)}. \qquad \textbf{(3.15)}$$

For phase-type service times, neither $\mathrm{E}(\hat{S}^2)$ nor the solution to (3.14a and 3.14b) can be obtained analytically. We use numerical methods. Particularly, we use quadrature method in solving (3.14), where numerically solving the Volterra integral equation (VIE) of the second kind plays a key role (note that the balance equation becomes a VIE of the second kind if $x \wedge b$ is replaced by x). Meanwhile, numerically solving VIE of the second kind alone is a deserving topic in applied mathematics (cf. [4,15][3]). Of course, the numerical method used in solving (3.14a and 3.14b) can also be used to solve Equation (3.13).

4. CONNECTION BETWEEN BALKING AND RENEGING

In this section we address the connection between the vqt-based balking model and the reneging model. Suppose the ith incoming customer has service time S_i and impatience time B_i. In the balking model, the customer leaves immediately in he/she sees the vqt is more than B_i. In the reneging model, the customer joins the queue and waits for service. If the service does not start before the impatience time B_i, the customer leaves. Let U_i be the interarrival times. Then the triplets $\{(U_i, S_i, B_i), i \geq 1\}$ determine the same vqt process in both balking and reneging models, since the entering of a customer who eventually renege does not cause a jump in the vqt process. From this observation, one can further consider the model that incorporates a mixture of balking and reneging as well.

The balking model and the reneging model do differ in the number of customers and workload in the system. The reneging behavior results in a greater number of customers and greater workload in the system than balking behavior. It is possible to derive the relationship between the performance measures of the reneging model and the vqt-equivalent balking model. For example, suppose the arrival is PP(λ). Let B be a generic random variable for i.i.d. impatience times. Let n_R and n_B be the long-run average number of customers in the reneging model and the vqt-equivalent balking model, respectively. Let w_R and w_B be the long-run average workload in the reneging

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

model and the vqt-equivalent balking model, respectively. Then, by PASTA, it can be shown that

$$n_R = n_B + p_r \lambda \mathrm{E}(B)$$

and

$$w_R = w_B + p_r \lambda \mathrm{E}(B)\mathrm{E}(S),$$

where

$$p_r = \lim_{t\to\infty} \Pr\{W_G(t) > B\}$$

is the fraction of balking customers. From the economic point of view, the balking rule saves system resources (waiting room, buffers, etc.). For the reneging rule, the reneging customers spend time waiting in the queue but do not get the desired service in the end.

It is clear that the balking interpretation is advantageous in analytical study. For example, the previous sections actually solve the corresponding problems for the reneging model with deterministic threshold, which is more difficult to analyze if we start from the reneging interpretation. On the other hand, we notice that the reneging interpretation is advantageous in simulation.

5. NUMERICAL RESULTS

In this section we illustrate our numerical results. We consider the *M/PH/s* model with vqt-dependent balking and three different service time distributions:

1. Exponential (exp): $\mu = 1(\tau = 1, \sigma^2 = 1)$;
2. 5-Erlang (erlang): $\mu = 5(\tau = 1, \sigma^2 = 0.2)$;
3. Hyperexponential (hyper): $\mu_1 = 4, \mu_2 = 2, \mu_3 = 1, \mu_4 = 0.8, \mu_5 = 0.5, \alpha_1 = \cdots = \alpha_5 = 0.2(\tau = 1, \sigma^2 = 1.75)$.

All of them have mean service time of 1. The variances are different, with 5-Erlang the smallest and hyperexponential the largest. We assume the balking threshold $b = 2$. For each of the three service time distributions, we use $s \in \{3, 10, 100\}$ and compute the long-run average queuing time for all served customers (w') and fraction of rejected customers (r) for different values of $\rho \in [0.1, 1.2]$ by using (1) Approximation I (using analytic formulas for the solution to (3.13a and 3.13b)), (2) Approximation II (using numerical method to solve (3.14a and 3.14b)), and (3) simulation methods. Notice that for exponential service time distribution, both Approximation I and Approximation II are exact. This gives us a method to verify the accuracy of simulation that turned out to be satisfactory in our experiments.

Figure 2 shows the long-run average queuing time for all served customers as a function of ρ. The w' values for exponential service times are exact. Others are from simulation. It can be seen that for almost all given values of ρ, there is an ordering of the queuing times for different service times according the order of the variances, either

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

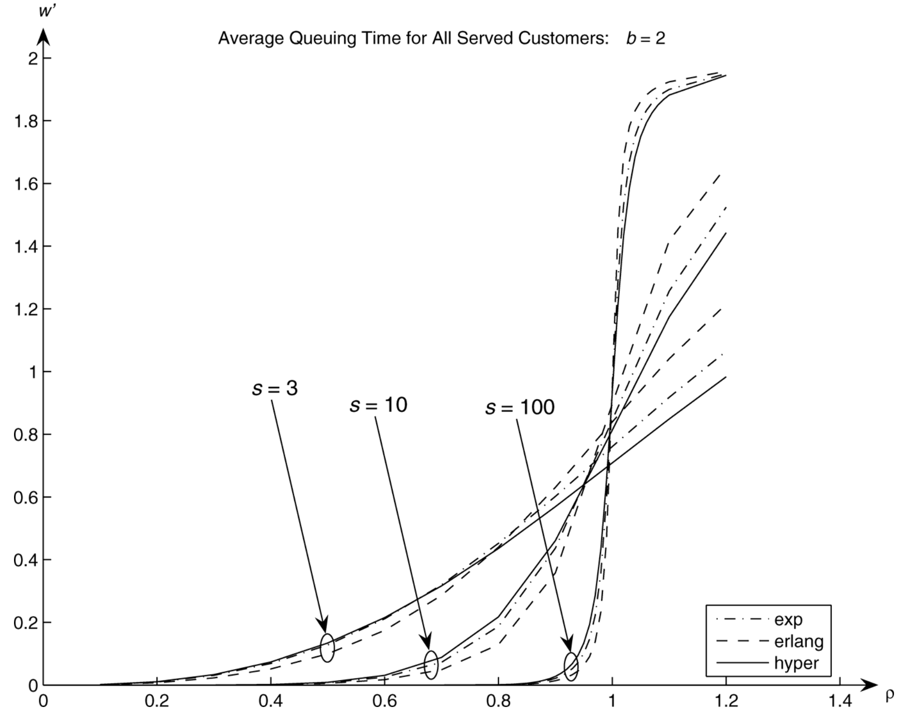

FIGURE 2. Long-run average queuing time for all served customers.

$w'_{\text{hyper}} > w'_{\text{exp}} > w'_{\text{erlang}}$ or $w'_{\text{hyper}} < w'_{\text{exp}} < w'_{\text{erlang}}$. The order reverses as ρ increases beyond a critical region. Intuitively, the value w' should converge to $b = 2$ as ρ approaches infinity. This trend is best illustrated by the set of curves for $s = 100$. The more the servers, the faster these curves approach b. Moreover, as s increases, the queuing time becomes more sensitive around the point $\rho = 1$ and the overall difference of the queuing time between these service time distributions diminishes.

Figure 3 is the same as Figure 2 except that it shows the fraction of rejected customers. We have similar observations as those for w'. Notice that for all given ρ, $r_{\text{hyper}} > r_{\text{exp}} > r_{\text{erlang}}$. The order reversion we observe from Figure 2 does not happen here. In addition, the fraction of rejected customers is almost linear in ρ when $\rho \geq 1$.

We also checked the accuracy of the approximations based on the relative error of w'. We verified the fact that as ρ increases, Approximation I and Approximation II become closer since the coefficient $\tilde{c}_G\gamma$ in (3.5a) approaches zero. We observed that both approximations have satisfactory accuracy over a wide rage of parameters, with no more than 20% deviation from the simulation (with 99% confidence intervals of width less than 1% of the estimate value), and the worse cases (error $> \pm 5\%$) occur where exact w' is small. Overall, Approximation II is more accurate than Approximation I, and the error of the former is less sensitive to ρ. However, the

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

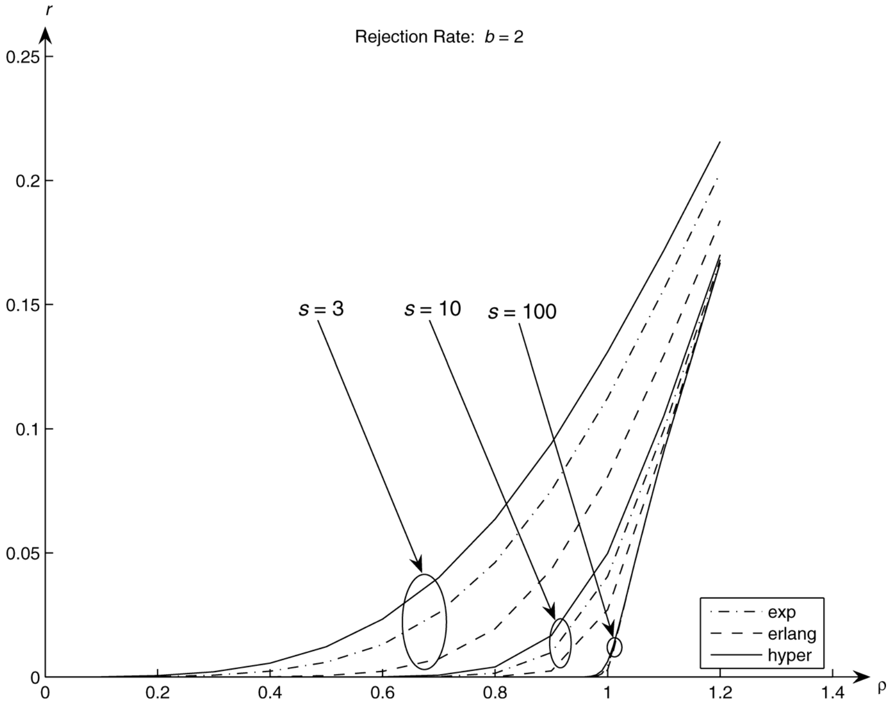

FIGURE 3. Fraction of rejected customers.

advantage is balanced by the fact that Approximation II needs numerical methods to solve (3.14a and 3.14b). Both approximations are, of course, far quicker than the simulation.

6. CONCLUSIONS

In this article we have obtained exact analytical results for the limiting behavior of an *M/M/s* system with vqt-dependent balking. These results also yield analytical results for the corresponding reneging case, which is more complicated if studied as a reneging system. Using these results, we have proposed two approximations for the *M/G/s* system with vqt-dependent balking. We have done extensive numerical and simulation experiments to conclude that Approximation I is easier to compute than Approximation II, but Approximation II is more accurate than Approximation I over a wide parameter space.

Notes

[1] The formula for $p_0^{(\tau)}$ in [2] consists of a term with inverted sign, which we believe is a typographical error.

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press

2. The approximation obtained by using $\dot{J} = \ddot{J} = \hat{S}$ is omitted since it is seen to be inferior according to the results of our numerical experiments.

3. We thank the anonymous referee who brought into our attention a fast and reliable method to solve VIE of the second kind developed in the [4].

References

1. Beneš, V.E. (1963). *General stochastic processes in the theory of queues*. Reading, MA: Addison-Wesley.
2. Boots, N.K. & Tijms, H.C. (1999). A multiserver queueing system with impatient customers. *Management Science* 45: 444– 448.
3. Daley, D.J. (1997). *Frontiers in queueing: Models and applications in science and engineering*. New York CRC Press, pp. 35–59.
4. den Iseger, P.W., Smith, M.A.J., and Dekker, R. (1997). Computing compound distributions faster! *Insurance: Mathematics and Economics* 20: 23–34.
5. Garnett, O., Mandelbaum, A., and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management* 4: 208–227.
6. Gnedenko, B.V. and Kovalenko, I.N. (1989). *Introduction to queueing theory*, 2nd ed. Boston: Birkhäuser, pp. 57–63.
7. Heyman, D.P. and Sobel, M.J. (1982). *Stochastic models in operations research: Stochastic processes and operating characteristics*. New York: McGraw-Hill.
8. Hokstad, P. (1978). Approximations for the *M/G/m* queue. *Operations Research* 26: 510–523.
9. Kimura, T. (1994). Approximations for multi-server queues: System interpolations. *Queueing Systems Theory and Applications* 17: 347–382.
10. Koole, G. & Mandelbaum, A. (2002). Queueing models of call centers: An introduction. *Annals of Operations Research* 113: 41–59.
11. Kulkarni, V.G. (1995). *Modeling and analysis of stochastic systems*. London: Chapman & Hall.
12. Lee, A.M. & Longton, P.A. (1959). Queueing processes associated with airline passenger check-in. *Operations Research Quarterly* 10: 56–71.
13. Liu, L.Q. & Kulkarni, V.G. (2006). Explicit solutions for the steady state distributions in *M/PH/*1 queues with workload dependent balking, *Queueing Systems Theory Applications* 52: 251–260.
14. Miyazawa, M. (1986). Approximation for the queue-length distribution of an *M/GI/s* queue by the basic equations. *Journal of Applied Probability* 23: 443–458.
15. Netravali, A.N. (1973). Spline approximation to the solution of the Volterra integral equation of the second kind. *Mathematics of Computation* 27: 99–106.
16. Newell, G. F. (1973). *Approximate stochastic behavior of n-server service systems with large n*. Lecture Notes in Economics and Mathematical Systems, Vol. 87, New York: Springer-Verlag.
17. Nozaki, A. & Ross, S.M. (1978). Approximations in finite-capacity multi-server queues with Poisson arrivals. *Journal of Applied Probability* 15: 826–834.
18. Takács, L. (1962). *Introduction to the theory of queues*. Oxford University Press.
19. Tijms, H.C. (1981). Approximations for the steady-state probabilities in the *M/G/c* queue. *Advances in Applied Probability* 13: 186 –206.
20. Tijms, H.C. (1986). *Stochastic modelling and analysis: A computational approach*. Chichester, UK: Wiley.
21. Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Science* 51: 221–235.

https://doi.org/10.1017/S0269964808000211 Published online by Cambridge University Press