JOURNAL OF INDUSTRIAL AND
MANAGEMENT OPTIMIZATION
Volume **7**, Number **3**, August **2011**

doi:10.3934/jimo.2011.7.593

pp. **593–606**

TAIL ASYMPTOTICS FOR WAITING TIME DISTRIBUTION OF AN M/M/S QUEUE WITH GENERAL IMPATIENT TIME

Yutaka Sakuma

Department of Distribution and Information Engineering
Hiroshima National College of Maritime Technology
Osakikamijima-Town, 725-0231, Japan

Atsushi Inoie

Department of Information Network and Communication
Kanagawa Institute of Technology
Atsugi-City, 243-0292, Japan

Ken'ichi Kawanishi

Department of Computer Science, Gunma University
Kiryu-City, 376-8515, Japan

Masakiyo Miyazawa

Department of Information Sciences, Tokyo University of Science
Noda-City, 278-8510, Japan

Abstract. In this paper, we consider an $M/M/s$ queueing model where customers may abandon waiting for service and leave the system without receiving their services. We assume that impatient time on waiting for each customer is an independent and identically distributed nonnegative random variable with a general distribution where the probability distribution is light-tailed and unbounded. The main objective of this paper is to provide an approximation for the waiting time distribution in an analytically tractable form. To this end, we obtain the tail asymptotics of the waiting time distributions of served and impatient customers. By using the tail asymptotics, we show that the fairly good approximations of the waiting time distributions can be obtained in asymptotic region with low numerical complexity.

1. **Introduction.** Suppose we have a multi-server queueing system where customers arrive according to Poisson process and the service time is exponentially distributed. The service discipline is first-come first-served and customers wait in the queue if no servers are idle upon arrivals. The waiting customers may leave the queue without receiving their services due to impatience. The time willing to wait in queue (or impatient time on waiting, or more simply 'impatient time') of each customer is independent and identically distributed (i.i.d.) nonnegative random variable. For such a queueing system, Brandt and Brandt [6] derived the stationary queue length distribution. Furthermore, by using the technique of stationary point

2000 *Mathematics Subject Classification.* Primary: 60K25, 60K25; Secondary: 60K25.

Key words and phrases. Queue, general impatient time, waiting time distribution, tail behavior.

This paper was presented at the QTNA2010 conference which was held in Beijing, China, during July 24-26, 2010. An earlier and brief version of this paper was published in the Conference Proceedings. The reviewing process of the paper was handled by Wuyi Yue and Yutaka Takahashi as Guest Editors.

processes, they also provided the formulas of the waiting time distributions of served and impatient customers in analytically tractable forms. Both of the formulas, however, are expressed in analytical forms that involve two consecutive integrals. Thus, it is not straightforward to find how the distributions behave in terms of the system parameters. It also is time-consuming and may require a fair amount of memory to compute the waiting time distributions numerically by using the formulas because we have to resort to numerical calculation to evaluate these integrals.

The main objective of this paper is to provide asymptotic approximations for the waiting time distributions. By exploiting the analytical forms derived by Brandt and Brandt [6], we obtain the asymptotic waiting time distributions that are analytically more tractable forms. To this end, we assume that the impatient time is unbounded and asymptotically light-tailed. More specifically, we consider the case where the tail distribution of the time willing to wait of customers in queue decays exponentially in asymptotic region. The light-tailed impatient time distribution allows us to obtain the tail behaviors of the waiting time distributions in analytically more tractable forms, i.e., the forms without the double integral. Those forms are more convenient to find the dependency of the system parameters on the distributions. By using the tail distributions of the waiting times, we also show that the good approximations of the waiting time distributions can be obtained in asymptotic region with less computing cost than the formulas in [6].

From the practical point of view, unbounded distribution of the impatient time is not unrealistic and deserves to investigate its effect on the performance measures. In fact, according to the statistical analysis of a call center by Brown *et al.* [8], it suggests that the unbounded distribution of the time willing to wait is not unreasonable for customers in a telephone call center. We, however, should notice that the bounded time willing to wait may appear because of the system-driven scheme that lead customers to abandon waiting. For example, the telephone call center plays a "Please wait" message for customers in queue after the specific elapsed times from arrival [8]. At those times, customers are likely to abandon waiting, and consequently the distribution of the time willing to wait observed by the system may become bounded if the system has such control mechanisms. Light-tailed distribution of the impatient time is also not unrealistic. The statistical analysis based on the same data set suggests that the estimated hazard rate is almost constant for sufficiently large value of the impatient time, although there are some peaks at the short time. Lastly, we point out that the class of phase-type distributions including the Erlang and hyper-exponential distributions possesses the property of the light-tailed asymptotics. It is well known that any probability distribution of positive random variable can be approximated by phase-type distribution with arbitrary accuracy because the set of phase-type distributions is dense in the field of distributions of positive random variable [17]. It implies that the light-tailed and unbounded assumptions are not restrictive.

We emphasize that our asymptotic analysis of the multi-server queue with impatient customers is on the tail distributions of the waiting times. Such an asymptotic analysis for the waiting time distributions is not included in the other asymptotic analyses. For example, Garnett *et al.* [15] or Zeltyn and Mandelbaum [22] deal with the asymptotic analysis of the multi-server queue in terms of the number of servers. By using the asymptotic analysis, the rules of thumb for the large call centers are provided. For comprehensive reviews on call center queueing analysis including effects on customers' patience time, see Gans *et al.* [14]. The asymptotic analysis,

however, is intended primarily for a large number of servers increasing together with arrival rate of customers. On the contrary, our asymptotic analysis is in terms of the tail of the waiting time for given number of servers and arrival rate, implying that it can be applied to not only a large number of servers but a small number of servers as well. The other sort of the asymptotic analysis can be found in Brandt and Brandt [7] based on their prior work [6]. However, the asymptotic results are for the intensity of customers leaving the queue due to impatience in terms of the queue length and hence different from our viewpoint.

Much effort has been given to the study of the queueing systems with impatient customers so far in the past. However, it has been revealed that the performance measures of the queueing systems with impatient customers are not easy to get in analytically tractable forms except for some cases. Prior work on queues with impatient customers goes back to Palm [19] who provided an approach to model the reneging in queue for telephone customers, and analyzed the $M/M/s+M$ queueing model, where customers may abandon waiting for their services if the exponentially distributed deadline expires. The notation $+M$ represents that the time willing to wait in queue is exponentially distributed and others are in the same way. The deterministic, not exponential, impatient time distribution also provides tractable queueing models, and was investigated by several authors. Barrer [4] analyzed the $M/M/1+D$ queueing model, where impatient times on waiting are constant, and derived the queue length distribution and the loss probability. Finch [13] showed the waiting time distribution of the single server queueing model with deterministic impatient time in the case where customers arrive by renewal processes. de Kok and Tijms [11] investigated the unfinished workload in the $M/G/1+D$ queueing model, and obtained approximations for the loss probability and the expectation of the waiting time. Recently, Xiong *et al.* [21] treated the $M/H_2/1+D$ queueing model and obtained the distribution of the unfinished workload by using the Volterra integral equation. Apart from analytical tractability, the $GI/G/1+G$ queueing model was also studied by Daley [10], Stanford [20], Baccelli and Hebuterne [3], Baccelli *et al.* [2].

Focusing on multi-server queueing systems with impatient customers, the $M/M/s+G$ queueing model and its variants seem the tractable queueing models. By choosing the state of the system to be the number of customers in the system, Barrer [5] studied the $M/M/s+D$ queueing model and obtained the steady-state queue length distribution. On the other hand, Baccelli and Hebuterne [3] studied the $M/M/s+G$ queueing model by using the state of the system to be the offered virtual waiting time, i.e., the waiting time of a virtual customer who is infinitely patient and never abandon waiting for service, and obtained the steady-state probabilities of the number of customers in the system less than s customers as well as the offered virtual waiting time distribution. Haugen and Skogan [16] studied the $M/M/s+G$ queueing model with several Poisson inputs of customers, and obtained the waiting time distribution in terms of integral equation. Movaghar [18] and Brandt and Brandt [6, 7] generalized the $M/M/s+G$ queueing model by extending state-dependent arrival and service rates. They obtained the steady-state probabilities of the number of customers in the system for not only less than s customers but also s and more customers as well. Choi and Kim [9] analyzed the $MAP/M/c+D$ queueing model, where the arrival process of customers is the Markovian arrival process, and obtained the loss probability and the waiting time distribution.

This paper is organized as follows. In Section 2, we describe the mathematical model for the queueing system that we analyze in this paper. In Section 3, we show our main results on tail behaviors of the waiting time distributions in the queueing system. The results indicate that the tail asymptotics of the waiting times decay exponentially and the decay rates are determined by the exponential service rate and the asymptotic decay rate of the time willing to wait. We illustrate some numerical examples of the tail behaviors in Section 4. Finally, we conclude this paper in Section 5.

2. **Queueing model with impatient customers.** We consider an s-server queueing model with an unlimited single waiting line. Customers arrive according to a Poisson process with rate λ, and are served with first-come first-served discipline. Service time at each server is exponentially distributed with mean μ^{-1}, and the offered load for this queueing system is denoted by $\rho \equiv \lambda/(s\mu)$. We assume that each customer has a deadline until the beginning of his/her service, and that the deadline is independently and identically distributed to a general and nonnegative random variable I. According to [6], this queueing model is referred to as the $M/M/s+G$ queueing model (see also Figure 1).

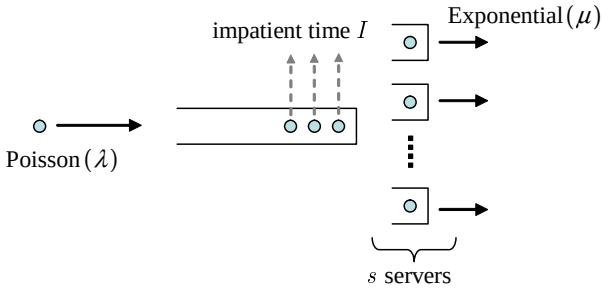

FIGURE 1. Queueing model with impatient customers

Let $\overline{C}$ be the tail distribution of the impatient time, that is, $\overline{C}(x) = \mathbb{P}(I > x)$ for $x \geq 0$. If we denote by $\overline{C}(\infty) \equiv \lim_{x\to\infty} \overline{C}(x)$, then the condition $\lambda\overline{C}(\infty) < s\mu$ ensures that the $M/M/s+G$ queueing model is stable [3]. In particular, it is clear that the stability condition holds if the tail distribution of the impatient time is not defective, i.e. $\overline{C}(\infty) = 0$.

We denote the tail distribution of the waiting time for a typical arriving customer under the condition that he/she will be served (resp. he/she will leave the system due to the end of his/her deadline) by $\overline{W}_{\mathrm{S}}$ (resp. $\overline{W}_{\mathrm{I}}$). From the results in [6], these waiting time distributions are known to have the following forms.

Proposition 1. *For $x \geq 0$, we have*

$$\overline{W}_{\mathrm{S}}(x) = \gamma_{\mathrm{S}} \int_{s\mu x}^{\infty} \lambda \exp(\lambda F(\xi)) F'(\xi) e^{-\xi} \mathrm{d}\xi, \tag{1}$$

$$\overline{W}_{\mathrm{I}}(x) = \gamma_{\mathrm{I}} \int_{s\mu x}^{\infty} \lambda \exp(\lambda F(\xi)) (F'(s\mu x) - F'(\xi)) e^{-\xi} \mathrm{d}\xi, \tag{2}$$

where $F(\xi) = \int_0^{\xi/(s\mu)} \overline{C}(\eta)\mathrm{d}\eta$ *and*

$$\gamma_{\mathrm{S}} = \left((s-1)! \sum_{k=0}^{s-1} \frac{(\lambda/\mu)^{k+1-s}}{k!} + \int_0^{\infty} e^{\lambda F(\xi)-\xi} \mathrm{d}\xi - 1 \right)^{-1},$$

$$\gamma_{\mathrm{I}} = \left(1 + (\lambda(s\mu)^{-1} - 1) \int_0^{\infty} e^{\lambda F(\xi)-\xi} \mathrm{d}\xi \right)^{-1}.$$

In spite of the nice appearance of the formulas in Proposition 1, we should note that the calculations of (1) and (2) involve two consecutive numerical integrations. Because these integrals must be evaluated numerically, the computer program may require a fair amount of time. Hence, we will give approximations for these distributions in analytically more tractable forms. For this, we consider the tail asymptotics of the waiting time distributions in the next section.

3. **Tail behaviors of waiting time distributions.** In the rest of this paper, we assume that the impatient time is asymptotically light-tailed and unbounded random variable, that is, there exist positive constants α and τ such that

$$\overline{C}(x) = \alpha e^{-\tau x} + o(e^{-\tau x}) \tag{3}$$

for $x \to \infty$. Then we obtain the tail asymptotics for the waiting time distributions of the served and the impatient customers, respectively.

Theorem 3.1. *The tail distributions of the waiting times exponentially decay as follows:*

$$\lim_{x\to\infty} e^{(\tau+s\mu)x} \overline{W}_{\mathrm{S}}(x) = \gamma_{\mathrm{S}} \frac{\lambda\alpha \exp(\lambda \mathbb{E}[I])}{\tau + s\mu}, \tag{4}$$

$$\lim_{x\to\infty} e^{(\tau+s\mu)x} \overline{W}_{\mathrm{I}}(x) = \gamma_{\mathrm{I}} \frac{\tau\lambda\alpha \exp(\lambda \mathbb{E}[I])}{s\mu(\tau + s\mu)}. \tag{5}$$

Remark 1. This theorem implies that we can evaluate the tail distributions of the waiting times without calculating numerical integrations other than those of γ_{S} and γ_{I}. Therefore, these asymptotic forms are quite useful to reduce numerical complexity. In contrast, the exact formulas (1) and (2) are more time-consuming than the asymptotic forms because the formulas are additionally required to carry out numerical integrations on the tail region.

Remark 2. In the end of this section, the assumption (3) is relaxed to a more general one (see, (18)). As you will see later, discussions under (3) and (18) are similar. Hence, the discussions in this section are given under (3) for notational ease.

We prove this theorem through the subsequent two lemmas.

Lemma 3.2. *For* $\xi \to \infty$*, we have*

$$F(\infty) - F(\xi) = \alpha\tau^{-1} e^{\frac{-\tau}{s\mu}\xi} + o(e^{\frac{-\tau}{s\mu}\xi}), \tag{6}$$

where $F(\infty) \equiv \lim_{\xi\to\infty} F(\xi)$.

Proof. We note that

$$F(\infty) - F(\xi) = \alpha\tau^{-1} e^{\frac{-\tau}{s\mu}x} + \int_{\xi/(s\mu)}^{\infty} (\overline{C}(x) - \alpha e^{-\tau x})\mathrm{d}x. \tag{7}$$

Under the assumption (3), for each $\epsilon > 0$ there exists $x_0 > 0$ such that

$$|\overline{C}(x) - \alpha e^{-\tau x}| < \epsilon e^{-\tau x} \tag{8}$$

for all $x > x_0$. Then we have

$$\begin{aligned} \left| \int_{\xi/(s\mu)}^{\infty} (\overline{C}(x) - \alpha e^{-\tau x}) \mathrm{d}x \right| &\leq \int_{\xi/(s\mu)}^{\infty} |\overline{C}(x) - \alpha e^{-\tau x}| \mathrm{d}x \\ &< \epsilon \tau^{-1} e^{\frac{-\tau}{s\mu}\xi} \end{aligned} \tag{9}$$

for $\xi/(s\mu) > x_0$, where the last inequality follows from (8). Then (9) shows that

$$\int_{\xi/(s\mu)}^{\infty} (\overline{C}(x) - \alpha e^{-\tau x}) \mathrm{d}x = o(e^{\frac{-\tau}{s\mu}\xi}). \tag{10}$$

Thus we obtain (6) from (7) and (10). □

Lemma 3.3. *For $x \to \infty$, we have*

$$\begin{aligned} \overline{W}_{\mathrm{S}}(x) &= \gamma_{\mathrm{S}} \Bigg((e^{\lambda F(\infty)} - e^{\lambda F(s\mu x)}) e^{-s\mu x} \\ &\quad - \frac{e^{\lambda F(\infty)} \lambda \alpha s \mu}{\tau(\tau + s\mu)} e^{(-\tau - s\mu)x} + o(e^{(-\tau - s\mu)x}) \Bigg). \end{aligned} \tag{11}$$

Proof. By integrating (1) by parts, we have

$$\begin{aligned} \overline{W}_{\mathrm{S}}(x) &= \gamma_{\mathrm{S}} \left(-e^{-s\mu x} e^{\lambda F(s\mu x)} + \int_{s\mu x}^{\infty} e^{-\xi} e^{\lambda F(\xi)} \mathrm{d}\xi \right) \\ &= \gamma_{\mathrm{S}} \Bigg(- e^{-s\mu x} e^{\lambda F(s\mu x)} \\ &\quad + e^{\lambda F(\infty)} \int_{s\mu x}^{\infty} e^{-\xi} \exp\left(-\lambda \alpha \tau^{-1} e^{\frac{-\tau}{s\mu}\xi} + o(e^{\frac{-\tau}{s\mu}\xi}) \right) \mathrm{d}\xi \Bigg), \end{aligned} \tag{12}$$

where the last equation follows from (6). Note that for a function $g(\xi)$ which converges to zero as ξ gets large, we have

$$\exp(g(\xi) + o(g(\xi))) = 1 + g(\xi) + o(g(\xi)) \tag{13}$$

for $\xi \to \infty$. Then we obtain (11) from (12) and (13) with $g(\xi) = -\lambda \alpha \tau^{-1} e^{\frac{-\tau}{s\mu}\xi}$. □

Proof of Theorem 3.1. From (6), we have

$$\begin{aligned} e^{\lambda F(\infty)} - e^{\lambda F(s\mu x)} &= e^{\lambda F(\infty)} (1 - \exp(-\lambda \alpha \tau^{-1} e^{-\tau x} + o(e^{-\tau x}))) \\ &= e^{\lambda F(\infty)} (\lambda \alpha \tau^{-1} e^{-\tau x} + o(e^{-\tau x})), \end{aligned} \tag{14}$$

where the last equation follows by (13). Then we obtain (4) from (11) and (14).

From (1) and (2), it is easy to see that

$$\overline{W}_{\mathrm{I}}(x) = \gamma_{\mathrm{I}} \left(\lambda F'(s\mu x) \int_{s\mu x}^{\infty} e^{-\xi} e^{\lambda F(\xi)} \mathrm{d}\xi - \gamma_{\mathrm{S}}^{-1} \overline{W}_{\mathrm{S}}(x) \right). \tag{15}$$

Since $F'(\xi) = \overline{C}(\xi/(s\mu))/(s\mu)$, (3) implies that

$$F'(s\mu x) = \frac{1}{s\mu} (\alpha e^{-\tau x} + o(e^{-\tau x})) \tag{16}$$

for $x \to \infty$. From the proof of Lemma 3.3 (see, (12) and (13)), we have

$$\int_{s\mu x}^{\infty} e^{-\xi} e^{\lambda F(\xi)} \mathrm{d}\xi = e^{\lambda F(\infty)} \left(e^{-s\mu x} - \frac{\lambda\alpha s\mu}{\tau(\tau + s\mu)} e^{(-\tau - s\mu)x} + o(e^{(-\tau - s\mu)x}) \right). \quad (17)$$

Then we obtain (5) from (15), (16) and (17). □

If the light-tailed assumption on the impatient time is relaxed as follows:

$$\overline{C}(x) = \alpha x^n e^{-\tau x} + o(x^n e^{-\tau x}) \quad (18)$$

for $x \to \infty$, where n is a nonnegative integer, then we obtain similar result as in Theorem 3.1. Although the proof of the following result is similar to Theorem 3.1, we give it in Appendix A for the convenience of readers.

Corollary 1. *Under the assumption (18), the waiting time distributions have the following tail asymptotics:*

$$\lim_{x\to\infty} x^{-n} e^{(\tau + s\mu)x} \overline{W}_{\mathrm{S}}(x) = \gamma_{\mathrm{S}} \frac{\lambda\alpha \exp(\lambda \mathbb{E}[I])}{\tau + s\mu}, \quad (19)$$

$$\lim_{x\to\infty} x^{-n} e^{(\tau + s\mu)x} \overline{W}_{\mathrm{I}}(x) = \gamma_{\mathrm{I}} \frac{\tau\lambda\alpha \exp(\lambda \mathbb{E}[I])}{s\mu(\tau + s\mu)}. \quad (20)$$

Remark 3. By Theorem 2.7.2 in [17], any continuous phase-type distribution has the tail behavior of (18). Furthermore, the continuous phase-type distribution is dense in the set of all distributions on $(0, \infty)$ (see, e.g. Theorem 4.2 in [1] and Theorem 2.7.1 in [17]). Hence the assumption (18) may be useful for our study because the impatient time is assumed to be light-tailed and unbounded random variable.

4. **Numerical examples.** We numerically confirm the accuracy of our analytical result of the tail asymptotics of waiting time distributions shown in (4) and (5) (also in (19) and (20)) compared with the result of Brandt and Brandt [6] (i.e., (1) and (2)). From (19) and (20), we approximate the tail distributions, $\overline{W}_{\mathrm{S}}$ and $\overline{W}_{\mathrm{I}}$, of waiting times as follows:

$$\overline{W}_{\mathrm{S}}(x) \approx \gamma_{\mathrm{S}} \frac{\lambda\alpha \exp(\lambda \mathbb{E}[I])}{\tau + s\mu} x^n e^{-(\tau + s\mu)x}, \quad (21)$$

$$\overline{W}_{\mathrm{I}}(x) \approx \gamma_{\mathrm{I}} \frac{\tau\lambda\alpha \exp(\lambda \mathbb{E}[I])}{s\mu(\tau + s\mu)} x^n e^{-(\tau + s\mu)x}. \quad (22)$$

In our numerical examples, we fix the parameter values as follows: $\mu = 0.025$. We further consider the following two tail distributions of impatient times.

$$\overline{C}_1(x) = 0.25e^{-0.01x} + 0.75e^{-0.05x},$$
$$\overline{C}_2(x) = e^{-0.05x} + 0.05xe^{-0.05x}.$$

Then the parameter values of α and τ in $\overline{C}_1$ (resp. $\overline{C}_2$) are given as 0.25 and 0.01 (resp. 0.05 and 0.05). Note that the expectations obtained by these impatient time distributions are the same and the value is 40. Actually, we cannot obtain the explicit expressions of the integrals involved in γ_{S} and γ_{I} of (1), (2), (21) and (22). We therefore try to calculate them numerically with a numerical integration method such as Newton-Cotes rules (see, e.g. [12]).

For investigating the accuracy of our analytical results, we calculate the quotient of the tail distribution, $\overline{W}_{\mathrm{I}}$ (resp. $\overline{W}_{\mathrm{S}}$), of the waiting time calculated from (1)

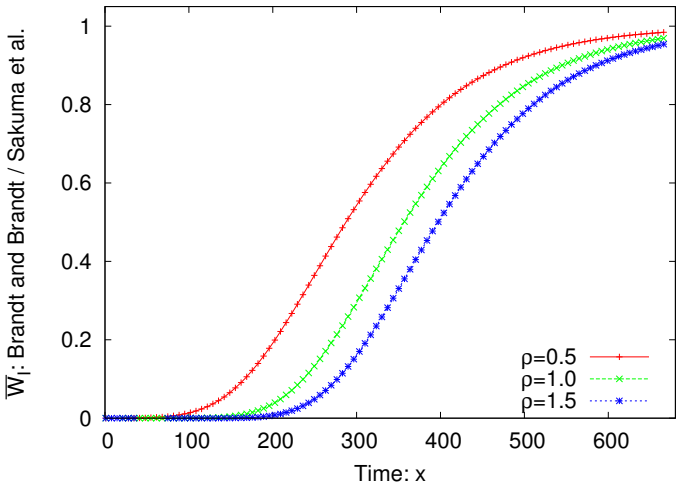

FIGURE 2. Convergence behavior of the tail distribution, $\overline{W}_{\rm I}$, of waiting time where $\overline{C}(x) = \overline{C}_1(x)$.

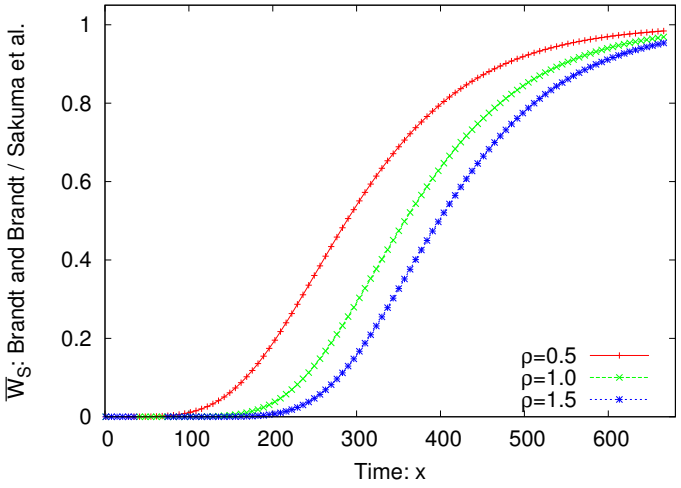

FIGURE 3. Convergence behavior of the tail distribution, $\overline{W}_{\rm S}$, of waiting time where $\overline{C}(x) = \overline{C}_1(x)$.

(resp. (2)) and that calculated from (21) (resp. (22)) for each x. Our analytical results converge to an exact value if the quotient achieves 1 for a larger x.

Tables 1-2 (resp. 3-4) show the tail distributions, $\overline{W}_{\rm I}$ and $\overline{W}_{\rm S}$, of the waiting times where $s = 2, 10, 40$ and $\overline{C}(x) = \overline{C}_1(x)$ (resp. $\overline{C}(x) = \overline{C}_2(x)$). Note that we fix the value of $\rho = \lambda/(s\mu)$ as 1 in the results shown in Tables 1-2 (resp. 3-4). We observe that our approximation improves the accuracy when s has a smaller value. The behaviors of the quotients of exact and approximate values seem non-monotonic. This may be caused by the non-convexity (or non-concavity) of the waiting time distributions of impatient queues.

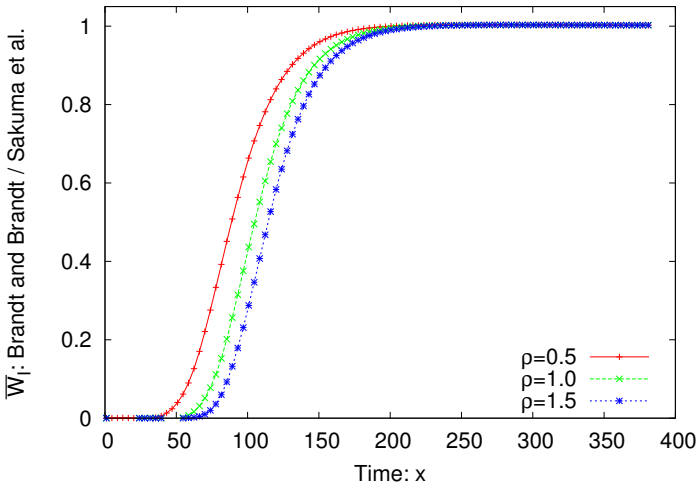

FIGURE 4. Convergence behavior of the tail distribution, $\overline{W}_{\mathrm{I}}$, of waiting time where $\overline{C}(x) = \overline{C}_2(x)$.

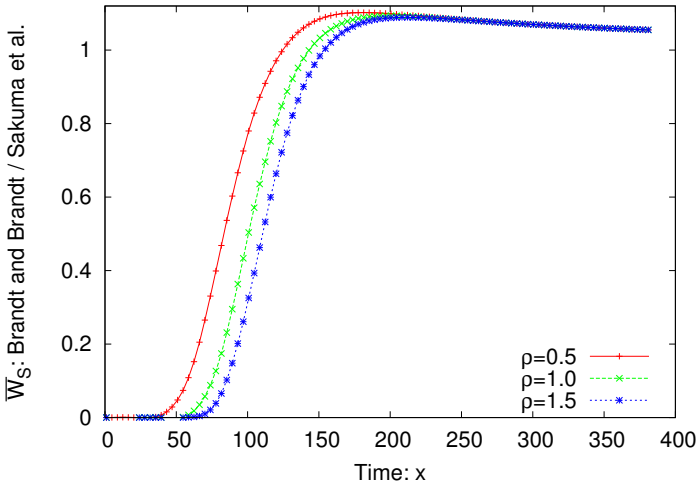

FIGURE 5. Convergence behavior of the tail distribution, $\overline{W}_S$, of waiting time where $\overline{C}(x) = \overline{C}_2(x)$.

Figures 2-3 (resp. 4-5) show the convergence behaviors of the tail distributions, $\overline{W}_{\mathrm{I}}$ and $\overline{W}_{\mathrm{S}}$, of the waiting times for some traffic intensities ρ where $\overline{C}(x) = \overline{C}_1(x)$ (resp. $\overline{C}(x) = \overline{C}_2(x)$). It is obvious that the values by our analytical results converge to exact values achieved by (1) and (2) when x has a larger value. In particular, the convergence is faster in the case where ρ is smaller.

TABLE 1. Tail behaviors of the waiting time distribution $\overline{W}_{\mathrm{I}}$ where $\overline{C}(x) = \overline{C}_1(x)$.

	$\lambda = 0.05,\ s = 2,\ \mu = 0.025$			$\lambda = 0.25,\ s = 10,\ \mu = 0.025$			$\lambda = 1,\ s = 40,\ \mu = 0.025$		
x	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.
100	0.00063913	0.00076315	0.83749	1.5738×10^{-10}	1.0821×10^{-9}	0.14544	1.1257×10^{-33}	7.9736×10^{-30}	0.00014118
150	3.1846×10^{-5}	3.7995×10^{-5}	0.83817	6.9465×10^{-16}	2.4458×10^{-15}	0.28401	4.0548×10^{-54}	9.3279×10^{-52}	0.0043469
200	1.6820×10^{-6}	1.8917×10^{-6}	0.88915	2.5391×10^{-21}	5.5284×10^{-21}	0.45929	3.9778×10^{-75}	1.0912×10^{-73}	0.036453
250	8.7572×10^{-8}	9.4180×10^{-8}	0.92983	7.7779×10^{-27}	1.2496×10^{-26}	0.62243	1.7089×10^{-96}	1.2766×10^{-95}	0.13387
300	4.4856×10^{-9}	4.6890×10^{-9}	0.95662	2.1179×10^{-32}	2.8245×10^{-32}	0.74982	4.4087×10^{-118}	1.4934×10^{-117}	0.29522
350	2.2725×10^{-10}	2.3345×10^{-10}	0.97342	5.3610×10^{-38}	6.3843×10^{-38}	0.83971	8.3347×10^{-140}	1.7470×10^{-139}	0.47708
400	1.1434×10^{-11}	1.1623×10^{-11}	0.98379	1.2980×10^{-43}	1.4431×10^{-43}	0.89945	1.3046×10^{-161}	2.0438×10^{-161}	0.63834
450	5.7296×10^{-13}	5.7866×10^{-13}	0.99013	3.0587×10^{-49}	3.2618×10^{-49}	0.93774	1.8210×10^{-183}	2.3909×10^{-183}	0.76165
500	2.8637×10^{-14}	2.8810×10^{-14}	0.99400	7.0908×10^{-55}	7.3727×10^{-55}	0.96176	2.3712×10^{-205}	2.7970×10^{-205}	0.84777
550	1.4291×10^{-15}	1.4344×10^{-15}	0.99636	1.6275×10^{-60}	1.6665×10^{-60}	0.97663	2.9601×10^{-227}	3.2720×10^{-227}	0.90469
600	7.1255×10^{-17}	7.1413×10^{-17}	0.99779	3.7132×10^{-66}	3.7668×10^{-66}	0.98576	3.6021×10^{-249}	3.8277×10^{-249}	0.94106
650	3.5507×10^{-18}	3.5554×10^{-18}	0.99866	8.4405×10^{-72}	8.5142×10^{-72}	0.99134	4.3159×10^{-271}	4.4779×10^{-271}	0.96382

TABLE 2. Tail behaviors of the waiting time distribution $\overline{W}_{\mathrm{S}}$ where $\overline{C}(x) = \overline{C}_1(x)$.

	$\lambda = 0.05,\ s = 2,\ \mu = 0.025$			$\lambda = 0.25,\ s = 10,\ \mu = 0.025$			$\lambda = 1,\ s = 40,\ \mu = 0.025$		
x	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.
100	0.0010883	0.0016171	0.67302	4.8661×10^{-10}	4.5345×10^{-9}	0.10732	6.3861×10^{-33}	6.3118×10^{-29}	0.00010118
150	6.1661×10^{-5}	7.8288×10^{-5}	0.78761	2.5695×10^{-15}	9.8465×10^{-15}	0.26096	2.8015×10^{-53}	7.0633×10^{-51}	0.0039662
200	3.3593×10^{-6}	3.8828×10^{-6}	0.86519	9.8048×10^{-21}	2.2133×10^{-20}	0.44299	2.8803×10^{-74}	8.2122×10^{-73}	0.035074
250	1.7696×10^{-7}	1.9321×10^{-7}	0.91588	3.0507×10^{-26}	4.9990×10^{-26}	0.61026	1.2584×10^{-95}	9.5990×10^{-95}	0.13109
300	9.1193×10^{-9}	9.6187×10^{-9}	0.94808	8.3735×10^{-32}	1.1298×10^{-31}	0.74113	3.2740×10^{-117}	1.1228×10^{-116}	0.29159
350	4.6364×10^{-10}	4.7888×10^{-10}	0.96817	2.1294×10^{-37}	2.5538×10^{-37}	0.83383	6.2200×10^{-139}	1.3135×10^{-138}	0.47355
400	2.3379×10^{-11}	2.3842×10^{-11}	0.98057	5.1699×10^{-43}	5.7723×10^{-43}	0.89563	9.7645×10^{-161}	1.5366×10^{-160}	0.63547
450	1.1730×10^{-12}	1.1870×10^{-12}	0.98817	1.2204×10^{-48}	1.3047×10^{-48}	0.93533	1.3654×10^{-182}	1.7976×10^{-182}	0.75957
500	5.8674×10^{-14}	5.9099×10^{-14}	0.99281	2.8319×10^{-54}	2.9491×10^{-54}	0.96026	1.7798×10^{-204}	2.1029×10^{-204}	0.84637
550	2.9295×10^{-15}	2.9424×10^{-15}	0.99563	6.5041×10^{-60}	6.6660×10^{-60}	0.97570	2.2233×10^{-226}	2.4600×10^{-226}	0.90378
600	1.4610×10^{-16}	1.4649×10^{-16}	0.99735	1.4844×10^{-65}	1.5067×10^{-65}	0.98519	2.7066×10^{-248}	2.8779×10^{-248}	0.94048
650	7.2816×10^{-18}	7.2934×10^{-18}	0.99839	3.3751×10^{-71}	3.4057×10^{-71}	0.99099	3.2437×10^{-270}	3.3667×10^{-270}	0.96347

TABLE 3. Tail behaviors of the waiting time distribution $\overline{W}_{\mathrm{I}}$ where $\overline{C}(x) = \overline{C}_2(x)$.

	$\lambda = 0.05,\ s = 2,\ \mu = 0.025$			$\lambda = 0.25,\ s = 10,\ \mu = 0.025$			$\lambda = 1,\ s = 40,\ \mu = 0.025$		
x	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.
10	0.80601	0.67957	1.1861	0.61349	91.386	0.0067131	0.33128	1.5433×10^{11}	2.1466×10^{-12}
20	0.50609	0.50000	1.0122	0.20465	9.0997	0.022490	0.020354	8.4992×10^{6}	2.3948×10^{-9}
30	0.27732	0.27591	1.0051	0.044239	0.67957	0.065099	0.00024610	351.06	7.0103×10^{-7}
40	0.13888	0.13534	1.0262	0.0067081	0.045112	0.14870	6.6419×10^{-7}	0.012889	5.1531×10^{-5}
50	0.065249	0.062234	1.0485	0.00076631	0.0028075	0.27295	5.1664×10^{-10}	4.4365×10^{-7}	0.0011645
60	0.029265	0.027474	1.0652	7.0454×10^{-5}	0.00016773	0.42004	1.5373×10^{-13}	1.4660×10^{-11}	0.010487
70	0.012684	0.011791	1.0757	5.5147×10^{-6}	9.7427×10^{-6}	0.56604	2.2517×10^{-17}	4.7096×10^{-16}	0.047811
80	0.0053586	0.0049575	1.0809	3.8422×10^{-7}	5.5435×10^{-7}	0.69310	1.9867×10^{-21}	1.4821×10^{-20}	0.13404
90	0.0022208	0.0020517	1.0824	2.4641×10^{-8}	3.1050×10^{-8}	0.79361	1.2294×10^{-25}	4.5914×10^{-25}	0.26776
100	0.00090693	0.00083866	1.0814	1.4906×10^{-9}	1.7176×10^{-9}	0.86781	5.9570×10^{-30}	1.4048×10^{-29}	0.42405
110	0.00036617	0.00033938	1.0789	8.6536×10^{-11}	9.4067×10^{-11}	0.91994	2.4431×10^{-34}	4.2551×10^{-34}	0.57416
120	0.00014650	0.00013620	1.0757	4.8804×10^{-12}	5.1091×10^{-12}	0.95524	8.9492×10^{-39}	1.2782×10^{-38}	0.70013
130	5.8190×10^{-5}	5.4281×10^{-5}	1.0720	2.6963×10^{-13}	2.7556×10^{-13}	0.97845	3.0374×10^{-43}	3.8131×10^{-43}	0.79658

TABLE 4. Tail behaviors of the waiting time distribution $\overline{W}_{\mathrm{S}}$ where $\overline{C}(x) = \overline{C}_2(x)$.

	$\lambda = 0.05,\ s = 2,\ \mu = 0.025$			$\lambda = 0.25,\ s = 10,\ \mu = 0.025$			$\lambda = 1,\ s = 40,\ \mu = 0.025$		
x	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.	Brandt & Brandt	Sakuma et al.	Brandt & Brandt /Sakuma et al.
10	0.33253	0.68667	0.48427	0.29262	160.82	0.0018196	0.17882	4.5948×10^{11}	3.8919×10^{-13}
20	0.20042	0.33682	0.59505	0.10370	10.676	0.0097136	0.013565	1.6870×10^{7}	8.0407×10^{-10}
30	0.10959	0.15489	0.70758	0.024586	0.66439	0.037015	0.00019949	580.67	3.4355×10^{-7}
40	0.055232	0.068375	0.80778	0.0040283	0.039694	0.10148	6.1811×10^{-7}	0.019188	3.2214×10^{-5}
50	0.026082	0.029346	0.88877	0.00048614	0.0023056	0.21085	5.2603×10^{-10}	6.1642×10^{-7}	0.00085337
60	0.011714	0.012338	0.94944	4.6298×10^{-5}	0.00013119	0.35292	1.6553×10^{-13}	1.9399×10^{-11}	0.0085332
70	0.0050662	0.0051063	0.99214	3.6988×10^{-6}	7.3479×10^{-6}	0.50339	2.5061×10^{-17}	6.0094×10^{-16}	0.041703
80	0.0021302	0.0020872	1.0206	2.6033×10^{-7}	4.0648×10^{-7}	0.64046	2.2513×10^{-21}	1.8387×10^{-20}	0.12244
90	0.00087717	0.00084463	1.0385	1.6750×10^{-8}	2.2261×10^{-8}	0.75245	1.4045×10^{-25}	5.5693×10^{-25}	0.25219
100	0.00035561	0.00033897	1.0491	1.0121×10^{-9}	1.2091×10^{-9}	0.83711	6.8194×10^{-30}	1.6730×10^{-29}	0.40761
110	0.00014248	0.00013509	1.0547	5.8538×10^{-11}	6.5212×10^{-11}	0.89766	2.7917×10^{-34}	4.9908×10^{-34}	0.55937
120	5.6575×10^{-5}	5.3520×10^{-5}	1.0571	3.2841×10^{-12}	3.4965×10^{-12}	0.93927	1.0185×10^{-38}	1.4800×10^{-38}	0.68817
130	2.2307×10^{-5}	2.1095×10^{-5}	1.0575	1.8036×10^{-13}	1.8651×10^{-13}	0.96703	3.4388×10^{-43}	4.3665×10^{-43}	0.78755

5. **Concluding remarks.** In this paper, we studied a multi-server queue with impatient customers, and gave approximations for the waiting time distributions of the served and the impatient customers, respectively (see, (21) and (22)). These results were obtained by considering the tail asymptotics of the waiting time distributions under the condition that the impatient time is unbounded and asymptotically light-tailed random variable. We demonstrated that our approximations work for some numerical examples.

For a queueing system with impatient customers, it may be interesting to consider the following conditional probability as its performance measure:

$$m(x) \equiv \mathbb{P}(V > I | I > x), \quad x > 0, \tag{23}$$

where V is the offered virtual waiting time which is the waiting time of a virtual arriving customer who has no limitation on his/her waiting time. That is, (23) is the probability that a typical arriving customer will not be served despite his/her patience time is greater than x. Since $\overline{W}_{\mathrm{I}}(x) = \mathbb{P}(I > x | V > I)$ by definition, (23) is rewritten by

$$m(x) = \frac{\overline{W}_{\mathrm{I}}(x)\mathbb{P}(V > I)}{\overline{C}(x)}, \tag{24}$$

where $\mathbb{P}(V > I)$ is the probability that a typical arriving customer will leave the system due to his/her impatience. From the results in Baccelli and Hebuterne [3] and Brandt and Brandt [6], we have

$$\mathbb{P}(V > I) = \frac{\tilde{\gamma}_{\mathrm{S}}}{\rho\gamma_{\mathrm{I}}}, \tag{25}$$

where $\tilde{\gamma}_{\mathrm{S}}$ is given by γ_{S} in Proposition 1 whose parameter s is replaced by $s+1$, i.e.,

$$\tilde{\gamma}_{\mathrm{S}} = \left(s! \sum_{k=0}^{s} \frac{(\lambda/\mu)^{k-s}}{k!} + \int_0^\infty e^{\lambda F(\xi) - \xi} \mathrm{d}\xi - 1 \right)^{-1}.$$

Then under the assumption (18), from (20), (24) and (25), we get

$$\lim_{x\to\infty} e^{s\mu x} m(x) = \tilde{\gamma}_{\mathrm{S}} \frac{\tau \exp(\lambda \mathbb{E}[I])}{\tau + s\mu}.$$

Hence, for large $x > 0$, the performance measure may be approximated as follows:

$$m(x) \approx \tilde{\gamma}_{\mathrm{S}} \frac{\tau \exp(\lambda \mathbb{E}[I])}{\tau + s\mu} e^{-s\mu x}.$$

That is, the probability of abandoning which is caused by a patient customer can be estimated by this formula.

Appendix A. **Proof of Corollary 1.** When the tail behavior of the impatient time distribution is given by (18), (7) is rewritten by

$$F(\infty) - F(\xi) = \alpha \int_{\xi/(s\mu)}^\infty x^n e^{-\tau x} \mathrm{d}x + \int_{\xi/(s\mu)}^\infty (\overline{C}(x) - \alpha x^n e^{-\tau x}) \mathrm{d}x. \tag{26}$$

By integrating by parts, it is easy to see that

$$\int_{\xi/(s\mu)}^\infty x^n e^{-\tau x} \mathrm{d}x = \sum_{k=0}^{n} \frac{n!}{(n-k)!} \tau^{-(k+1)} \left(\frac{\xi}{s\mu} \right)^{n-k} e^{\frac{-\tau\xi}{s\mu}}, \tag{27}$$

which implies that

$$\int_{\xi/(s\mu)}^{\infty} x^n e^{-\tau x}\mathrm{d}x = \tau^{-1}\left(\frac{\xi}{s\mu}\right)^n e^{\frac{-\tau\xi}{s\mu}} + o\left(\left(\frac{\xi}{s\mu}\right)^n e^{\frac{-\tau\xi}{s\mu}}\right). \tag{28}$$

Then the inequality (9) is rewritten as

$$\left|\int_{\xi/(s\mu)}^{\infty} (\overline{C}(x) - \alpha x^n e^{-\tau x})\mathrm{d}x\right| < \epsilon\left(\sum_{k=0}^{n} \frac{n!}{(n-k)!}\tau^{-(k+1)}\right)\left(\frac{\xi}{s\mu}\right)^n e^{\frac{-\tau\xi}{s\mu}}$$

by (18) and (27), which implies that

$$\int_{\xi/(s\mu)}^{\infty} (\overline{C}(x) - \alpha x^n e^{-\tau x})\mathrm{d}x = o\left(\left(\frac{\xi}{s\mu}\right)^n e^{\frac{-\tau\xi}{s\mu}}\right). \tag{29}$$

From (26), (28) and (29), we obtain

$$F(\infty) - F(\xi) = \alpha\tau^{-1}\left(\frac{\xi}{s\mu}\right)^n e^{\frac{-\tau\xi}{s\mu}} + o\left(\left(\frac{\xi}{s\mu}\right)^n e^{\frac{-\tau\xi}{s\mu}}\right), \tag{30}$$

which is a generalization of (6).

From (13) and (30), (12) is rewritten by

$$\begin{aligned}\overline{W}_{\mathrm{S}}(x) &= \gamma_{\mathrm{S}}\Bigg\{ - e^{-s\mu x}e^{\lambda F(s\mu x)} \\ &\quad + e^{\lambda F(\infty)}\int_{s\mu x}^{\infty} e^{-\xi}\left(1 - \alpha\tau^{-1}\lambda\left(\frac{\xi}{s\mu}\right)^n e^{\frac{-\tau\xi}{s\mu}} + o\left(\left(\frac{\xi}{s\mu}\right)^n e^{\frac{-\tau\xi}{s\mu}}\right)\right)\mathrm{d}\xi\Bigg\}.\end{aligned}$$

Similarly to (28), the second part of the above integral is estimated as follows:

$$\int_{s\mu x}^{\infty} e^{-\xi}\left(\frac{\xi}{s\mu}\right)^n e^{\frac{-\tau\xi}{s\mu}}\mathrm{d}\xi = \frac{s\mu}{\tau + s\mu}x^n e^{-(\tau+s\mu)x} + o(x^n e^{-(\tau+s\mu)x}).$$

Furthermore, the third integral is evaluated at $o(x^n e^{-(\tau+s\mu)x})$. Then we obtain

$$\begin{aligned}\overline{W}_{\mathrm{S}}(x) &= \gamma_{\mathrm{S}}\Bigg\{(e^{\lambda F(\infty)} - e^{\lambda F(s\mu x)})e^{-s\mu x} \\ &\quad - \frac{e^{\lambda F(\infty)}\alpha\lambda s\mu}{\tau(\tau + s\mu)}x^n e^{-(\tau+s\mu)x} + o(x^n e^{-(\tau+s\mu)x})\Bigg\},\end{aligned} \tag{31}$$

which is equivalent to (11) when $n = 0$. From (13), (30) and (31), we obtain (19). Similarly to (5), (20) is derived by (19).

REFERENCES

[1] S. Asmussen, "Applied Probability and Queues," 2nd ed., Applications of Mathematics (New York), **51**, Springer-Verlag, New York, 2003.

[2] F. Baccelli, P. Boyer and G. Hebuterne, *Single-server queues with impatient customers*, Advances in Applied Probability, **16** (1984), 887–905.

[3] F. Baccelli and G. Hebuterne, *On queues with impatient customers*, in "Performance '81" (ed. F. J. Kylstra), North-Holland, Amsterdam-New York, **32** (1981), 159–179.

[4] D. Y. Barrer, *Queueing with impatient customers and indifferent clerks*, Operations Research, **4** (1957), 644–649.

[5] D. Y. Barrer, *Queueing with impatient customers and ordered service*, Operations Research, **4** (1957), 650–656.

[6] A. Brandt and M. Brandt, *On the $M(n)/M(n)/s$ queue with impatient calls*, Performance Evaluation, **35** (1999), 1–18.

[7] A. Brandt and M. Brandt, *Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s+GI$ system*, Queueing Systems, **41** (2002), 73–94.

[8] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao, *Statistical analysis of a telephone call center: A queueing-science perspective*, Journal of the American Statistical Association, **100** (2005), 36–50.
[9] B. D. Choi and B. Kim, *$MAP/M/c$ queue with constant impatient time*, Mathematics of Operations Research, **29** (2004), 309–325.
[10] D. J. Daley, *General customer impatience in the queue $GI/G/1$*, Journal of Applied Probability, **2** (1965), 186–205.
[11] A. G. de Kok and H. C. Tijms, *A queueing system with impatient customers*, Journal of Applied Probability, **22** (1985), 688–696.
[12] G. Evans, "Practical Numerical Analysis," John Wiley & Sons, 1996.
[13] P. D. Finch, *Deterministic customer impatience in the queueing system $GI/M/1$*, Biometrika, **47** (1960), 45–52.
[14] N. Gans, G. Koole and A. Mandelbaum, *Telephone call centers: Tutorial, review, and research prospects*, Manufacturing and Service Operations Management, **5** (2003), 79–141.
[15] O. Garnett, A. Mandelbaum and M. Reiman, *Designing a call center with impatient customers*, Manufacturing & Service Operations Management, **4** (2002), 208–227.
[16] R. B. Haugen and E. Skogan, *Queueing systems with stochastic time out*, IEEE Transactions on Communications, **28** (1980), 1984–1989.
[17] G. Latouche and V. Ramaswami, "Introduction to Matrix Analytic Methods in Stochastic Modeling," American Statistical Association and the Society for Industrial and Applied Mathematics, Philadelphia, American Statistical Association, Alexandria, VA, 1999.
[18] A. Movaghar, *On queueing with customer impatience until the beginning of service*, Queueing Systems Theory Appl., **29** (1998), 337–350.
[19] C. Palm, *Methods of judging the annoyance caused by congestion*, Tele (English ed.), **2** (1953), 1–20.
[20] R. E. Stanford, *Reneging phenomena in single server queues*, Mathematics of Operations Research, **4** (1979), 162–178.
[21] W. Xiong, D. Jagerman and T. Altiok, *$M/G/1$ queue with deterministic reneging times*, Performance Evaluation, **65** (2008), 308–316.
[22] S. Zeltyn and A. Mandelbaum, *Call centers with impatient customers: Many-server asymptotics of the $M/M/n + G$ queue*, Queueing Systems, **51** (2005), 361–402.

Received September 2010; 1st revision December 2010; final revision May 2011.

E-mail address: sakuma@hiroshima-cmt.ac.jp
E-mail address: inoie@nw.kanagawa-it.ac.jp
E-mail address: kawanisi@cs.gunma-u.ac.jp
E-mail address: miyazawa@is.noda.tus.ac.jp