TUTORIALS IN
OPERATIONS RESEARCH
INFORMS 2011

 | ISBN 978-0-9843378-2-8
doi: http://dx.doi.org/10.1287/educ.1110.0083

Queues in Service Systems: Customer Abandonment and Diffusion Approximations

J. G. Dai
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, dai@gatech.edu

Shuangchi He
Department of Industrial and Systems Engineering, National University of Singapore, Singapore 117576, heshuangchi@gatech.edu

Abstract In a service system, the system performance is sensitive to customer abandonment. We focus on $G/GI/n+GI$ parallel-server queues, which serve as a building block to model service systems. Consistent with recent empirical findings, such a queue has a general arrival process (the G) that can be time nonhomogeneous, independent and identically distributed (iid) service times with a general distribution (the first GI), and iid patience times with a general distribution (the $+GI$). Following the square-root safety staffing rule, companies are able to operate such queues in the quality- and efficiency-driven (QED) regime that is characterized by large customer volume, the waiting times being a fraction of the service times, only a small fraction of customers abandoning the queue, and high server utilization. We survey recent results on many-server queues that operate in the QED regime. These results include the sensitivity of patience time distributions and diffusion models as a practical tool for performance analysis.

Keywords Kingman approximation; heavy traffic; square-root safety staffing; Halfin–Whitt regime; quality- and efficiency-driven regime; piecewise OU process; functional central limit theorem

1. Introduction

In a production system such as a semiconductor wafer fabrication line, parts may wait for a long time before they are processed at various stations. It is common for the long-run average waiting time to be several times longer than the average processing time; see, e.g., Lu et al. [35]. Customers in a *service system* such as a call center are human beings, however, and their patience of waiting is often limited. Thus, some of them may abandon the system before their service begins. The phenomenon of customer abandonment is ubiquitous because no one would wait for service indefinitely. As argued by Garnett et al. [22], customer abandonment is a key factor for call center operations. It can significantly impact the performance of a service system. The abandonment probability or the long-run fraction of customers who abandon the system is an important performance measure, at least for most revenue-generating service systems. One must model customer abandonment explicitly for an operational model to be relevant for decision making. See §5 for an example.

In this paper, we focus on a mathematical model that is denoted as a $G/GI/n+GI$ queue. In this queue, we model customer abandonment by assigning each customer a patience time. When a customer's waiting time for service exceeds his patience time, he abandons the queue without service. In the notation, the G refers to a general arrival process that can be time nonhomogeneous, the first GI refers to independent and identically distributed (iid) service times with a general service time distribution, n is the number of identical servers, and $+GI$

refers to iid patience times with a general distribution. We call a $G/GI/n+GI$ queue with a large number of parallel servers a *many-server queue*. Such a queue serves as a building block to model large-scale service systems. For service systems such as call centers, it is reasonable to assume that the patience times are iid, because the queue is usually invisible to waiting customers. As argued by Halfin and Whitt [27], the performance of many-server queues is qualitatively different from that of single-server queues or queues with a small number of servers. Therefore, single-server-based approximations cannot describe the operations of large-scale service systems.

In a single-server queue, the mean waiting time goes to infinity because of the stochastic variability in interarrival and service times as the server approaches 100% utilization. Unlike a single-server queue for which a manager has to make a painful choice between *quality* of service (short waiting times) and *efficiency* (high server utilization), a many-server queue can be operated in the quality- and efficiency-driven (QED) regime that is characterized by large customer volume, the waiting times being a fraction of the service times, only a small fraction of customers abandoning the queue, and high server utilization. Many-server queues can be operated in the QED regime because of the *pooling effect* achieved by a large number of servers working in parallel. Clearly, managers should strive to operate their systems in the QED regime, which is also called the *rationalized* regime by Gans et al. [21]. See §§2 and 3 for more discussions on the contrast between single-server queues and many-server queues.

An important decision for a manager is to decide how many servers should be used at different hours of a day. The square-root safety staffing rule is an important staffing principle that is both theoretically justified and widely practiced. Under the square-root safety staffing rule, when the customer volume is high, the system will be operated in the desired QED regime. See §4 for more discussion.

It was empirically reported that both the service time distribution and the patience time distribution are far from exponential; see, e.g., Brown et al. [14]. Therefore, one must use general distributions to model service times and patience times. Recent papers, e.g., those by He and Dai [28], Mandelbaum and Momčilović [36], and Zeltyn and Mandelbaum [52], show that a many-server queue in the QED regime is insensitive to the patience time distribution as long as the patience time density at the origin is fixed. See §6 for more discussion.

When the service time and the patience time distributions are general, except by computer simulation, there is no analytical or numerical method to evaluate the performance of such a queue. We survey diffusion approximations for a many-server queue in the QED regime. In our diffusion models, the service time distribution is modeled by a phase-type distribution, and the patience time distribution is general. In §7, we show that the diffusion models are accurate in predicting the system performance of a many-server queue.

2. Painful Choice: Quality or Efficiency in Single-Server Queues

For a typical service system such as a call center or a hospital emergency department, the customer arrival process approximately follows a Poisson process, possibly time nonhomogeneous; see, e.g., Brown et al. [14] and Shi et al. [45]. To evaluate the quality of service for such a system, the fraction of customers who have to wait before receiving service, also known as the delay probability, and the average customer waiting time are two important performance measures. Usually, these two measures should be maintained under certain levels to meet customer expectations.

If the system has only one server, an important model is an $M/GI/1$ queue, where the arrival process is assumed be a homogeneous Poisson process. Let λ be the arrival rate of the Poisson process, and let m be the mean service time. Then,

$$\rho = \lambda m$$

is the traffic intensity of the queue. When $\rho < 1$, it can also be interpreted as the server utilization. By the Poisson-arrivals-see-time-averages property (see, e.g., Wolff [51]), the delay probability is given by

$$P_w = 1 - \rho.$$

When ρ is close to one, almost all customers have to wait before receiving service.

Let c_s^2 be the squared coefficient of variation (SCV) of the service times. (The SCV of a positive random variable is defined as the variance divided by the squared mean.) By the Pollaczek–Khinchine formula (see, e.g., Gross and Harris [25]), the mean waiting time of the $M/GI/1$ queue is given by

$$w = m\left(\frac{\rho}{1-\rho}\right)\left(\frac{1+c_s^2}{2}\right). \tag{1}$$

Let

$$f = \left(\frac{\rho}{1-\rho}\right)\left(\frac{1+c_s^2}{2}\right).$$

We call f the *waiting time factor* of the queue. It is the ratio of the mean waiting time to the mean service time. Formula (1) shows that the waiting time factor is proportional to $\rho/(1-\rho)$ in a single-server queue. As a result, both the waiting time factor and the mean waiting time are large when the server utilization ρ is close to one.

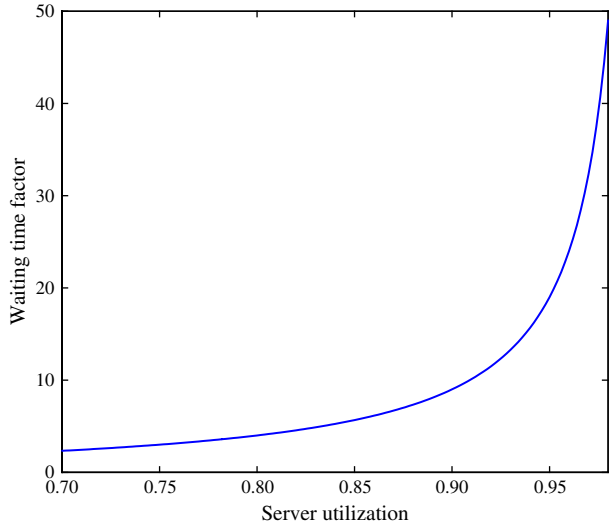
Consider a special case when the service times are exponentially distributed. The SCV of the service times is $c_s^2 = 1$, and the corresponding queue is denoted as an $M/M/1$ queue. Suppose that the mean service time is five minutes. If the server is kept busy 95% of time, then 95% of customers have to wait before receiving service. In the meantime, the waiting time factor is $f = 19$, giving an expected waiting time of more than one hour and a half. Such quality of service is unlikely acceptable if customers are human beings.

In the case when "customers" are jobs to be processed or messages to be transmitted, long waiting times can sometimes be acceptable to keep certain expensive bottleneck resources heavily utilized. For example, a modern semiconductor wafer fabrication line costs more than three billion dollars to build, and it is important to keep the parts in process waiting in queues to be processed at various stations. It is not uncommon for the waiting time factor to be larger than 19 in such a production system. In a service system, it is human beings, not parts, that wait. A customer would not be satisfied if his expected waiting time was 19 times his service time. In a service system, the quality of service is important. Although how to measure quality of service may differ, depending on managers and the application context, we choose to focus on the delay probability and the mean waiting time in this paper. The Pollaczek–Khinchine formula (1) dictates that, in the single-server setting, one cannot and should not maintain high server utilization to achieve good quality of service. In the setting of a single server or a small number of servers, the operational efficiency (high server utilization) must be sacrificed to maintain the service quality at a satisfactory level because of the variability in the arrival and service processes. However, managers do not have to face this painful trade-off between quality and efficiency when there are many servers working in parallel; see §3 below.

The Pollaczek–Khinchine formula (1) clearly shows the *nonlinear* effect of the server utilization on the mean waiting time. The waiting time factor increases rapidly as the server utilization approaches one. For example, for an $M/M/1$ queue as the server utilization grows from $\rho = 0.9$ to 0.95, the waiting time factor increases from $f = 9$ to 19. However, as the server utilization grows from $\rho = 0.8$ to 0.85, the waiting time factor increases only from $f = 4$ to 5.67. Therefore, when a single-server queue is already heavily loaded, even a slight increase in server utilization will degrade the quality of service significantly.

Figure 1 illustrates how the waiting time factor evolves with the server utilization in an $M/M/1$ queue.

FIGURE 1. Waiting time factor vs. server utilization in an $M/M/1$ queue.



3. Simultaneous Quality and Efficiency in Multiserver Queues

In contrast to a single-server system that has to compromise between service quality and operational efficiency, it is possible to achieve both of them simultaneously in a service system with many parallel servers. Let us consider an $M/M/n$ queue that has a Poisson arrival process with rate λ, exponentially distributed service times with mean $m = 1/\mu$, and n identical servers working in parallel. The traffic intensity is now defined as

$$\rho = \lambda/(n\mu).$$

Let $X(t)$ be the number of customers in the system at time t. Then $X = \{X(t)\colon t \geq 0\}$ is a continuous-time Markov chain. Assume that $\rho < 1$. The Markov chain has a unique stationary distribution $\pi = (\pi_i\colon i = 0, 1, 2, \ldots)$ that satisfies the following balance equations:

$$\begin{cases} \lambda\pi_i = (i+1)\mu\pi_{i+1} & \text{for } i = 0, 1, \ldots, n-1, \\ \lambda\pi_i = n\mu\pi_{i+1} & \text{for } i = n, n+1, \ldots, \\ \sum_{i=0}^{\infty} \pi_i = 1. \end{cases} \tag{2}$$

From (2), one has the Erlang-C formula for the delay probability, i.e.,

$$P_w = \sum_{i=0}^{\infty} \pi_{n+i} = \frac{(n\rho)^n}{n!}\left((1-\rho)\sum_{k=0}^{n-1}\frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!}\right)^{-1}. \tag{3}$$

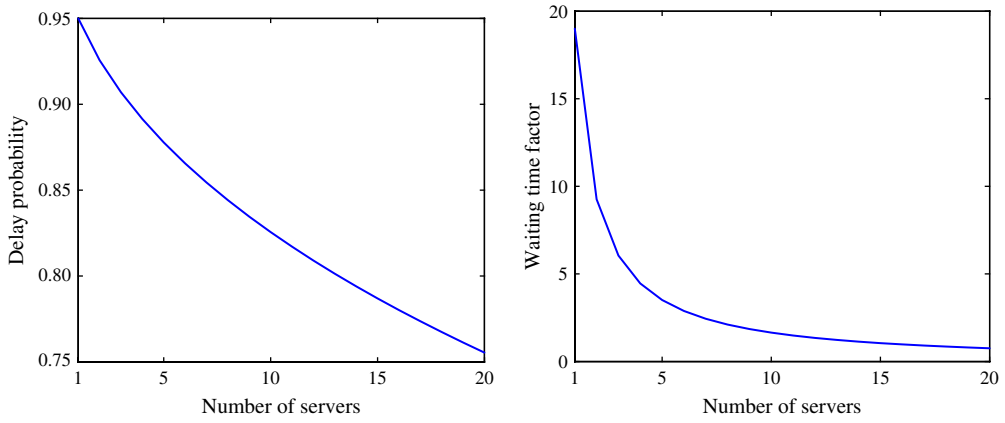
The mean waiting time is given by

$$w = \lambda^{-1}\sum_{i=0}^{\infty} i\pi_{n+i} = P_w\frac{1}{(1-\rho)n\mu}.$$

Hence, the waiting time factor for the $M/M/n$ queue is given by

$$f(n,\rho) = w\mu = P_w\frac{1}{(1-\rho)n}. \tag{4}$$

FIGURE 2. Delay probability and waiting time factor vs. number of servers for $M/M/n$ queues with $\rho = 0.95$.



Because $P_w \leq 1$, we have

$$f(n,\rho) \leq \frac{1}{(1-\rho)n}.$$

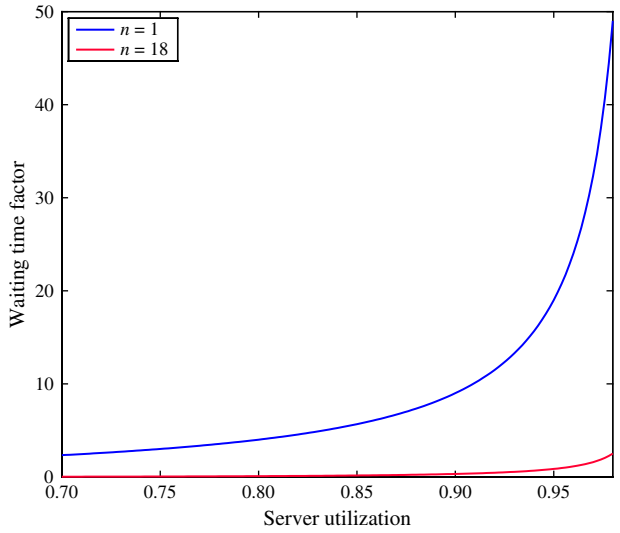
Thus, for each fixed $\rho < 1$, the waiting time factor $f(n,\rho)$ approaches zero as $n \to \infty$.

Set the average utilization per server to be $\rho = 0.95$. In Figure 2, we plot the delay probabilities and the waiting time factors as the number of servers n increases from one. The figure shows that the delay probability decreases gradually while the waiting time factor decreases rapidly. For example, when $n = 18$, the delay probability is 76.7%, and the waiting time factor is $f(18, 0.95) = 0.85 < 1$. Thus, when $n = 18$, the average utilization per server is 95%, 76.7% of customers are delayed before receiving service, and the mean waiting time is less than the mean service time; the quality of service is reasonable. If one further increases the number of servers to $n = 100$ and the average utilization per server is kept at 95%, then the delay probability decreases to 50.7%, and the waiting time factor is $f(100, 0.95) = 0.101$. In this case, nearly half of all customers are served immediately upon arrival without any delay, and the mean waiting time is only 10% of the mean service time. This level of service is highly attractive despite the fact that the servers are 95% utilized. Such a system achieves both high quality of service and operational efficiency. Therefore, it is operated in the QED parameter regime, a term coined by Atar et al. [3]. In the QED regime, the system has a large number of parallel servers, the arrival rate is high, and the arrival rate and the service capacity are approximately balanced so that the server utilization is close to one. The QED regime is also called a *rationalized* regime by Gans et al. [21] because, in most cases, a manager should operate his service system in this regime.

Even though the average server utilization is close to one, only a fraction of customers need to wait in a queue with enough servers working in parallel. This phenomenon is in sharp contrast to the one in which almost all customers have to wait in a single-server queue. To illustrate how the waiting time factor changes with the server utilization in a multiserver queue, we also plot the waiting time factor curve for a queue with $n = 18$ servers in Figure 3. Compared with the curve for the single-server queue that was first plotted in Figure 1 and is replotted in Figure 3, the waiting time factor for the 18-server queue increases much more slowly as the server utilization approaches one.

The Poisson arrival process assumption is equivalent to the assumption that the interarrival times are iid following an exponential distribution. For the Poisson arrival process, the

FIGURE 3. Waiting time factor vs. server utilization in an $M/M/1$ queue and an $M/M/18$ queue.



SCV of its interarrival times is $c_a^2 = 1$. Then, the Pollaczek–Khinchine formula (1) for the mean waiting time in an $M/GI/1$ queue can be written as

$$w = m\left(\frac{\rho}{1-\rho}\right)\left(\frac{c_a^2 + c_s^2}{2}\right). \tag{5}$$

Kingman [32] showed that when the server utilization is close to one, formula (5) approximately holds for a $GI/GI/1$ queue that has iid interarrival times following a general distribution (the first GI). When the queue has no variability so that $c_a^2 = c_s^2 = 0$, there is no waiting, and thus $w = 0$. In general, the variability in customer arrival and service processes contributes to the system congestion, degrading the quality of service, particularly when the system is heavily loaded. Figure 3, however, illustrates that the influence of the variability can be offset by pooling service facilities. The pooling principle has been widely used in resource management under uncertainty.

4. Square-Root Safety Staffing Rule

Consider a $GI/GI/n$ queue with arrival rate λ and mean service time $1/\mu$. Let $R = \lambda/\mu$ be the offered load of the queue. When the arrival rate λ is large, so is the offered load R. One expects that an appropriate staffing level n should be

$$n = R + \Delta,$$

where Δ is the excess service capacity against the system's stochastic variability. To keep the server utilization high, Δ should be much smaller than the offered load R. The *square-root safety staffing rule* recommends an amount of excess service capacity of

$$\Delta = \beta\sqrt{R}$$

for some $\beta > 0$. Thus, following the square-root safety staffing rule, the staffing level n satisfies

$$n = R + \beta\sqrt{R} \approx R + \beta\sqrt{n} \tag{6}$$

when the offered load R is high. Of course, the value of n given by (6) should be rounded to an integer. It turns out that, with a fixed $\beta > 0$, as the offered load increases, the corresponding

staffing level n in (6) stabilizes the delay probability and makes the waiting time factor on the order of $1/\sqrt{n}$. At the end of this section, we show the approximation

$$P_w \approx \frac{1}{\beta\Phi(\beta)/\phi(\beta)+1} \tag{7}$$

when the offered load R is high in the $M/M/n$ setting, where ϕ and Φ are the probability density and the cumulative distribution function, respectively, of the standard normal variable. An improvement, particularly when the offered load R is not necessarily high, over approximation (7) will also be discussed.

There are at least two usages for formula (7). The first one is for performance analysis with a given staffing level n, and the second one is to determine the staffing level that achieves a given delay probability. In the first usage, for a given staffing level n and a given utilization level $\rho < 1$, set

$$\beta = \sqrt{n}(1-\rho),$$

and use the right side of (7) to approximate the delay probability P_w. Once we have the delay probability, the waiting time factor $f(n,\rho)$ in (4) becomes

$$f(n,\rho) = \frac{P_w}{\sqrt{n}\beta}. \tag{8}$$

For example, when $n = 100$ and $\rho = 0.95$, one has $\beta = 0.5$, and the right side of (7) predicts P_w to be 0.505, compared with the exact value 0.507 from (3). The waiting time factor computed through (4) is 0.101 based on both exact and approximate values of P_w.

The second and more significant usage of (7) is that it leads to the following implementation of the square-root safety staffing rule in the $M/M/n$ setting. Suppose that the delay probability is required to be less than a target value $0 < \gamma < 1$. One needs to set the staffing level n (i.e., the number of servers) so that the delay probability is approximately γ. For this, one first solves for β using

$$\gamma = \frac{1}{\beta\Phi(\beta)/\phi(\beta)+1}. \tag{9}$$

Once we have β, for a given offered load R, one sets the staffing level n using the first part of (6).

The celebrated *Halfin–Whitt regime* refers to a parameter regime when the offered load $R \to \infty$ while the safety coefficient β remains fixed, and the staffing level n is set according to the square-root safety staffing rule (6). This regime was first analyzed in the seminal paper by Halfin and Whitt [27]. As the offered load R increases, as long as the staffing level n increases accordingly following (6) with the fixed β satisfying (9), the delay probability stabilizes at γ. Because $\rho = R/n$, following the square-root safety staffing rule (6), the server utilization gets close to one as the offered load increases to infinity. Also, it follows from (4) that the waiting time factor is $f = P_w/(\sqrt{n}\beta)$, which approaches zero at rate $1/\sqrt{n}$ as the offered load increases to infinity. Hence, in the $M/M/n$ setting, the square-root safety staffing rule automatically leads the system to the QED regime, achieving both quality and efficiency. The analysis in Reed [42] concludes that the square-root safety staffing rule also automatically leads the system to the QED regime in the $GI/GI/n$ setting.

The origin of the square-root safety staffing rule can be traced back to Erlang's paper written in 1923, which was collected in Brockmeyer et al. [13]. In the $M/M/n/n$ setting, which models a loss system such as a telephone system, Erlang derived the rule by marginal analysis of the benefit of adding a server. He also mentioned that such a rule had been practiced as early as in 1913. The square-root safety staffing rule has also been advocated by Grassmann [23, 24], Kolesar [34], Newell [40, 41], and Whitt [48]. Whitt [48] formally proposed and analyzed the rule.

Now we show approximation (7). In this derivation, we work directly with the continuous-time Markov chain, rather than with the diffusion approximations in Halfin and Whitt [27]. Recall that the delay probability P_w is given in (3). Using the fact that $R = n\rho$, we have

$$P_w = \frac{R^n/n!}{(1-\rho)\sum_{k=0}^{n-1} R^k/k! + R^n/n!} = \frac{(R^n/n!)e^{-R}}{(1-\rho)\sum_{k=0}^{n-1}(R^k/k!)e^{-R} + (R^n/n!)e^{-R}}.$$

The square-root safety staffing rule $n = R + \beta\sqrt{n}$ implies $R = n(1-\beta/\sqrt{n})$. Therefore,

$$\frac{R^n}{n!}e^{-R} = \frac{n^n(1-\beta/\sqrt{n})^n}{n!}e^{-n+\beta\sqrt{n}} = \frac{n^n}{n!}e^{-n+\beta\sqrt{n}+n\log(1-\beta/\sqrt{n})}. \tag{10}$$

Using the Stirling formula and the Taylor expansion for $\log(1-x)$, we have

$$n! \sim \sqrt{2\pi}n^{n+1/2}e^{-n} \quad \text{and} \quad n\log(1-\beta/\sqrt{n}) \sim -\beta^2/2 - \beta\sqrt{n} \tag{11}$$

as $n \to \infty$. (Here, $f(n) \sim g(n)$ means $\lim_{n\to\infty} f(n)/g(n) = 1$.) By (10) and (11), we have

$$\frac{R^n}{n!}e^{-R} \sim \frac{1}{\sqrt{n}}\frac{1}{\sqrt{2\pi}}e^{-\beta^2/2} = \frac{1}{\sqrt{n}}\phi(\beta).$$

Let N be a Poisson random variable with mean R. Then

$$\sum_{k=0}^{n-1}(R^k/k!)e^{-R} = \mathbb{P}[N < n].$$

When the offered load R is high, N is approximately normally distributed with both the mean and the variance being R. Thus,

$$\mathbb{P}[N < n] = \mathbb{P}\left[\frac{N-R}{\sqrt{R}} < \frac{n-R}{\sqrt{R}}\right] \approx \mathbb{P}\left[\frac{N-R}{\sqrt{R}} < \beta\right] \approx \Phi(\beta).$$

The above analysis leads to approximation (7).

When the offered load R is moderate, there are refined normal approximations for the Poisson probability $\mathbb{P}[N < n]$. For example, the Wilson–Hilferty approximation (see Johnson and Kotz [30]) gives

$$\mathbb{P}[N < n] \approx 1 - \Phi\left(\frac{c-\mu_0}{\sigma_0}\right),$$

where $c = (R/n)^{1/3}$, $\mu_0 = 1 - 1/(9n)$, and $\sigma_0 = 1/(3\sqrt{n})$; see also the approximations in Janssen et al. [29]. Once one has a refined normal approximation for $\mathbb{P}[N < n]$, one can estimate the delay probability by

$$P_w = \frac{1}{1 + \beta(1-\Phi((c-\mu_0)/\sigma_0))/\phi(\beta)}. \tag{12}$$

For example, when $n = 18$ and $\rho = 0.95$, (12) gives 76.8% as an approximation for P_w, compared with the exact value 76.7% computed from (3). The approximation (7) gives 75.9% for the delay probability.

5. Customer Abandonment

The phenomenon of customer abandonment is present in most service systems that serve human beings. For a service system with significant customer abandonment, any queueing model that ignores the abandonment phenomenon is likely irrelevant to operational decisions. To demonstrate the significant influence of customer abandonment to the system performance, let us consider an $M/M/n+M$ queue with $n = 50$ servers. The arrival process

Table 1. Comparison between queues with and without customer abandonment.

	$M/M/50+M$	$M/M/50$
Abandonment fraction	10.2%	N/A
Mean waiting time (in sec.)	12.5	87.7
Mean queue length	11.2	72.2
Server utilization (%)	98.8	98.8

is Poisson with rate $\lambda = 55$ customers per minute. The service times are exponentially distributed with mean one minute. Each customer has a patience time, and the patience times are iid following an exponential distribution with mean two minutes. Several performance measures of this queue are listed in Table 1.

In the same table, we also list the performance measures for a *modified* queue. The modified queue is an $M/M/n$ queue with the same mean service time, the same number of servers, and the same throughput as the original queue, but it has no customer abandonment. The arrival rate λ^* in the modified queue is set to be equal to the throughput of the original queue, i.e., $\lambda^* = 55 \times (1 - 0.102) = 49.39$. Table 1 shows that both the mean waiting time and the mean queue length in the original queue are much smaller than that in the modified queue. In other words, with the same service capacity and throughput, some key performance measures in a queue with abandonment can be much better than in a queue without abandonment. To meet a certain service requirement without considering customer abandonment, one tends to overestimate the staffing level. Of course, customer abandonment can be costly. One needs to find a trade-off between customer abandonment and staffing cost using a correct model.

The better performance on the waiting times in a queue with customer abandonment can be explained intuitively as follows. In the original queue with customer abandonment, when the system is in a congestion period, customers who experience long waiting times abandon the system. For these customers, their waiting times are capped by their patience times. In the corresponding queue without abandonment, these customers will experience extremely long delays, which degrades the overall waiting time statistics for the system.

The square-root safety staffing rule (6) also applies to a service system with high offered load in the presence of customer abandonment. Consider the staffing problem for $M/M/n+M$ queues. Let $1/\alpha$ be the mean patience time. As argued by Garnett et al. [22], to meet the target delay probability $0 < \gamma < 1$, one can set the staffing level following (6), but the value of β is now determined by solving

$$\gamma = \left(1 + \frac{h(\beta\sqrt{\mu/\alpha})}{\sqrt{\mu/\alpha}h(-\beta)}\right)^{-1} \tag{13}$$

with

$$h(x) = \frac{\phi(x)}{1 - \Phi(x)}$$

being the hazard rate function of the standard normal distribution. For a service system with customer abandonment, the fraction of abandoned customers is another important performance measure. Let β be a fixed real number that can be either positive or zero or negative. For each offered load R, set the staffing level n by (6). It follows from Garnett et al. [22] that the fraction of customers who abandon the $M/M/n+M$ queue is approximately given by

$$P_A \approx \frac{1}{\sqrt{R}}(\sqrt{\alpha/\mu}h(\beta\sqrt{\mu/\alpha}) - \beta)\left(1 + \frac{h(\beta\sqrt{\mu/\alpha})}{\sqrt{\mu/\alpha}h(-\beta)}\right)^{-1}.$$

Therefore, in the presence of customer abandonment, the square-root safety staffing rule can still lead the system to the QED regime and yield high server utilization, short waiting times, and a very small abandonment fraction.

Because abandoned customers do not receive any service, the traffic intensity $\rho = \lambda/(n\mu)$ should not be interpreted as the server utilization any longer. With customer abandonment, a service system can reach a steady state even if the customer arrival rate is larger than its service capacity. As more and more customers accumulate in the buffer, the customer abandonment rate keeps increasing until arrivals and departures (including both service completions and abandonments) reach an equilibrium. When solving (13) for β, it is possible to have a negative solution that results in a staffing level below the offered load. Because of the presence of customer abandonment, the service system can still achieve both quality and efficiency in this case.

There is a growing list of papers that study queueing models with customer abandonment. These papers include those by Baccelli et al. [4], Bassamboo et al. [6, 7], Bhattacharya and Ephremides [8], Boxma and de Waal [10], Brandt and Brandt [11, 12], Mandelbaum and Zeltyn [37], and Zeltyn and Mandelbaum [52]. Fluid models and related analysis of many-server queues with customer abandonment have been studied by Atar et al. [1], Bassamboo and Randhawa [5], Kang and Ramanan [31], Whitt [49, 50], and Zhang [53]. Diffusion models and related analysis have been studied in Atar et al. [2], Dai and He [17], Dai et al. [19], Dai and Tezcan [18], Garnett et al. [22], Gurvich and Whitt [26], He and Dai [28], Koçağa and Ward [33], Mandelbaum and Momčilović [36], Mandelbaum and Zeltyn [38], Reed and Tezcan [43], Reed and Ward [44], Talreja and Whitt [46], and Tezcan and Dai [47].

6. Performance Insensitivity to Patience Time Distributions

As we have demonstrated in the previous sections, service systems operating in the QED regime are characterized by short customer waiting times. For example, for $M/M/n$ queues operating in the QED regime with $\beta = \sqrt{n}(1-\rho)$ being fixed, it can be seen from (8) that the waiting time decreases to zero at rate $1/\sqrt{n}$ as the number of servers $n \to \infty$. The same decreasing rate of waiting times has been proved in Garnett et al. [22] for $M/M/n+M$ queues operating in the QED regime. The work of Dai and He [17] suggests that a similar result holds for $GI/GI/n+GI$ queues. Hence, the waiting times are relatively short for a service system in the QED regime. For example, if a service system has hundreds of servers working in parallel, and the service times are typically several minutes, then in the QED regime, the waiting times should be on the order of seconds. The above observation implies that when n is large, the patience time distribution, outside a small neighborhood of zero, barely has any influence on the system dynamics. Such a result can be confirmed by the following numerical example.

Consider an $M/M/n+GI$ queue. Let F be the patience time distribution that satisfies

$$F(0) = 0 \quad \text{and} \quad \alpha = F'(0+) = \lim_{x \downarrow 0} x^{-1} F(x) < \infty. \tag{14}$$

So α is the density of F at the origin. In particular, α is identical to the abandonment rate when the patience time distribution is exponential. If the waiting times are short, the abandonment process should depend on the patience time distribution mostly through its density at the origin. Suppose that the queue has $n = 100$ servers, the Poisson arrival process has rate $\lambda = 105$, and the service times are exponentially distributed with mean one. This system is slightly overloaded. A fraction of traffic, at least $(\lambda - 100)/\lambda = 4.8\%$ of the arrivals, has to abandon the system. For a general patience time distribution, there are no analytical tools or numerical methods to compute the abandonment fraction and the mean queue length. We consider three patience time distributions with the same densities at the origin: an exponential distribution (Exp) with rate α, a uniform distribution (Uniform) on the interval $[0, 1/\alpha]$, and a two-phase hyperexponential (H_2) distribution. A two-phase hyperexponential distribution can be determined by its initial distribution $p = (p_1, p_2)$ with $p_1 + p_2 = 1$ and its rate vector $\nu = (\nu_1, \nu_2)$. For such a hyperexponentially distributed random variable, with probability p_1 it is exponentially distributed with mean $1/\nu_1$, and with

probability p_2 it is exponentially distributed with mean $1/\nu_2$. In our example, the hyperexponential patience time distribution is set to have $p = (0.21, 0.79)$ and $\nu = (0.3\alpha, 79\alpha/30)$. Thus, 21% of customers have long service times with mean $10/(3\alpha)$, and 79% of customers have short service times with mean $30/(79\alpha)$. Equivalently, the density function of the hyperexponential patience time distribution is given by

$$f_{H_2}(x) = 0.21\alpha \exp(-0.3\alpha x) + 0.79\alpha \exp(-79\alpha x/30), \quad x \geq 0. \tag{15}$$

All three distributions have density α at the origin.

Computer simulation is conducted to estimate the abandonment fraction and the mean queue length for each case. The simulation estimates are averaged over 20 independent runs, and each run lasts 10^5 time units. Table 2 displays the results for different α values and different patience time distributions. For each row with a fixed α, the performance is very close for different patience time distributions.

The above simulation example indicates that in the QED regime, the system performance is generally invariant with the patience time distribution as long as its density at the origin is fixed and positive. This invariance also suggests that to obtain performance measures for a many-server queue with a general patience time distribution, it is generally accurate to replace the patience time distribution by an exponential distribution with the same density at the origin. An exponential patience time distribution is attractive in many aspects. For example, when the service time distribution is phase type, the matrix-analytic method sometimes can be effective to compute the performance of a queue with exponential patience time distribution. The computed performance is in turn used to approximate the original queue with a general patience time distribution. Section 7 will have more discussion on phase-type distributions and the matrix-analytic method.

Table 2 supports the replacement of an $M/M/100 + GI$ queue by an $M/M/100 + M$ queue. However, it is important that the two systems match the patience time density at the origin, not any other statistics such as the mean patience time. To highlight this point, suppose that a manager uses an $M/M/100+M$ system to replace an $M/M/100+GI$ system. But this time, the manager matches the *mean patience time*, a practice that is often used in industry. In Table 3, for a fixed mean patience time m_p, simulation estimates of mean queue lengths are given for different patience time distributions, including an exponential distribution with rate $\alpha = m_p^{-1}$, a uniform distribution in $[0, 2m_p]$ with $\alpha = m_p^{-1}/2$, and the hyperexponential distribution given by (15) with $\alpha = 2.447m_p^{-1}$. Table 3 shows that for each fixed m_p, the performance is drastically different as the patience time distribution changes. This example illustrates that the mean patience time is a wrong statistic to focus on and one should never use it to calibrate a customer abandonment model.

The phenomenon of performance insensitivity to patience time distributions was first studied by Zeltyn and Mandelbaum [52] for steady-state analysis and later elaborated on by Dai and He [17] for process-level analysis, where a deterministic relationship is established between the abandonment processes and the queue length processes for many-server queues.

Table 2. Performance insensitivity to patience time distributions.

	Abandonment fraction				Mean queue length			
	Exp	Uniform	H_2	Diffusion	Exp	Uniform	H_2	Diffusion
$\alpha = 0.1$	0.0497	0.0498	0.0496	0.0497	52.18	50.59	54.19	52.19
$\alpha = 0.5$	0.0603	0.0607	0.0599	0.0603	12.67	12.06	13.43	12.66
$\alpha = 1$	0.0670	0.0676	0.0662	0.0669	7.031	6.585	7.592	7.022
$\alpha = 2$	0.0739	0.0748	0.0730	0.0738	3.882	3.547	4.313	3.877
$\alpha = 10$	0.0886	0.0902	0.0869	0.0886	0.9301	0.7540	1.172	0.9302

TABLE 3. Mean patience time is a wrong statistic.

	Abandonment fraction			Mean queue length		
	Exp	Uniform	H_2	Exp	Uniform	H_2
$m_p = 0.1$	0.0886	0.0840	0.0926	0.9301	1.505	0.5840
$m_p = 0.5$	0.0739	0.0676	0.0794	3.882	6.585	2.455
$m_p = 1$	0.0670	0.0608	0.0730	7.031	12.06	4.313
$m_p = 2$	0.0603	0.0550	0.0682	12.67	22.10	6.438
$m_p = 10$	0.0497	0.0481	0.0543	52.18	98.07	24.52

This relationship says that for many-server queues in the QED regime, the cumulative number of customers who have abandoned the system is approximately equal to a constant multiple of the cumulative amount of waiting time among all customers. Clearly, the constant should be interpreted as the abandonment rate per unit of waiting time. It was proved in Dai and He [17] that this constant is equal to the patience time density at the origin when it is strictly positive. More specifically, if $A(t)$ is the number of abandonments by time t and $Q(t)$ is the queue length (i.e., the number of waiting customers) at time t, then $\int_0^t Q(s)\,\mathrm{d}s$ is the cumulative waiting time by time t among all customers, and the scaled difference

$$\frac{1}{\sqrt{n}}\left(A(t) - \alpha \int_0^t Q(s)\,\mathrm{d}s\right)$$

is close to zero for any time $t \geq 0$ when n is large. Hence, one may use

$$A(t) \approx \alpha \int_0^t Q(s)\,\mathrm{d}s \tag{16}$$

to approximate the abandonment process in a many-server queue.

7. Diffusion Model for Many-Server Queues

The exact analysis of a many-server queue has largely been limited to the $M/M/n+M$ model, also known as the Erlang-A model, which has a Poisson arrival process and exponential service and patience time distributions. However, as pointed out by Brown et al. [14], the service time distribution in a call center appears to follow a log-normal distribution. Such distributions were also observed by Shi et al. [45] for length of stay in a hospital. In addition, the patience time distribution in a call center was observed by Zeltyn and Mandelbaum [52] to be far from exponential. With a general service or patience time distribution, there is no finite-dimensional Markovian representation of the queue. Except computer simulations, no methods are available to analyze such a queue either analytically or numerically. Hence, much attention has been devoted to the approximate analysis of such a queue.

In our approximate analysis of a queue, we approximate a general service time distribution with a phase-type distribution. A *phase-type random variable* is defined to be the time until absorption of a transient, finite-state Markov chain. Any positive-valued distribution can be approximated by phase-type distributions. See Neuts [39] for more discussion on phase-type distributions. For an $M/Ph/n+GI$ queue with a phase-type service time distribution, two multidimensional diffusion processes were proposed by He and Dai [28] to approximate the dynamics of the queue.

In §7.1, we introduce Brownian motion and illustrate how an arrival process such as a Poisson process can be approximated by a Brownian motion model. In §7.2, we illustrate the diffusion approximation of $M/M/n+GI$ queues. Because the service time distribution is exponential in this case, we are able to spell out the details of every step clearly in deriving the diffusion approximation. The resulting diffusion process is a one-dimensional piecewise

Ornstein–Uhlenbeck (OU) process, whose stationary distribution has an explicit formula. In §7.3, the diffusion model for $M/H_2/n+GI$ queues is presented. The resulting diffusion process is two-dimensional, whose stationary distribution can be computed numerically using the algorithm developed in He and Dai [28]. Diffusion approximations are rooted in many-server heavy traffic limit theorems, which require that the number of servers go to infinity. Section 7.4 shows that the diffusion approximation is accurate, sometimes for as few as 20 servers. All these diffusion approximations use the patience time density at the origin only. When the patience time density is zero at the origin or changes rapidly around the origin, we present in §7.5 an alternative diffusion model that uses the hazard rate function of the patience time distribution. The hazard rate diffusion model is shown to be accurate when the previous diffusion model works poorly or fails.

7.1. Brownian Approximation

Let $E=\{E(t)\colon t\geq 0\}$ be a Poisson arrival process with rate $\lambda>0$. Assume that the arrival rate is $\lambda=100$ customers per minute. In Figure 4(a), we plot a sample path of the Poisson process in the first 10 minutes. One can see that $E(t)$ evolves around the straight line given by the expectation $\mathbb{E}[E(t)]=\lambda t$. To focus on the stochastic variability of the arrival process E, we plot the sample path of the corresponding centered process $\{E(t)-\lambda t\colon t\geq 0\}$ in Figure 4(b). The centered process records the fluctuation of the arrival process around its mean. In the plot, the x-axis represents the time, in a span of 10 minutes. The fluctuation represented by the y-axis is scaled automatically by the plotting software. To examine the effect of this scaling further, in Figure 5 we plot the centered process when the arrival rate is $\lambda=10{,}000$. It turns out that the magnitude of the centered process is on the order of $\sqrt{\lambda}$ as λ becomes large.

Let $\mu_B\in\mathbb{R}$ and $\sigma_B^2>0$ be given. A stochastic process $B=\{B(t)\colon t\geq 0\}$ is said to be a (μ_B,σ_B^2)-*Brownian motion* if (i) $B(0)=0$ and almost every sample path is continuous, (ii) $\{B(t)\colon t\geq 0\}$ has stationary, independent increments, and (iii) $B(t)$ is normally distributed with mean $\mu_B t$ and variance $\sigma_B^2 t$ for every $t>0$. The parameter μ_B is called the drift, and σ_B^2 is called the variance. Such a process is called a standard Brownian motion if $\mu_B=0$ and $\sigma_B^2=1$. By the well-known Donsker's theorem, (see, e.g., Billingsley [9]), $\tilde{E}_\lambda=\{\tilde{E}_\lambda(t)\colon t\geq 0\}$ converges in distribution to a standard Brownian motion as $\lambda\to\infty$, where the scaled, centered process $\tilde{E}_\lambda$ is defined by

$$\tilde{E}_\lambda(t)=\frac{E(t)-\lambda t}{\sqrt{\lambda}}. \tag{17}$$

For a Poisson process, Donsker's theorem suggests that one may replace its scaled fluctuation in (17) by the standard Brownian motion when the arrival rate λ is large. Donsker's theorem

FIGURE 4. Poisson process with rate $\lambda=100$.

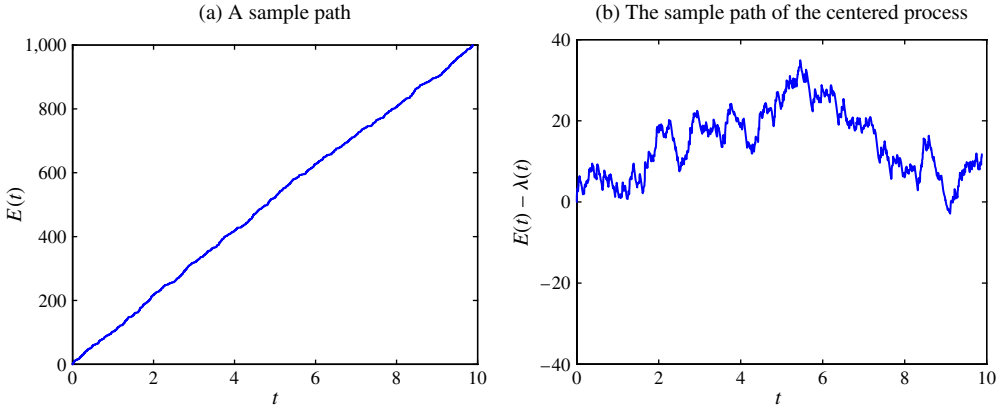
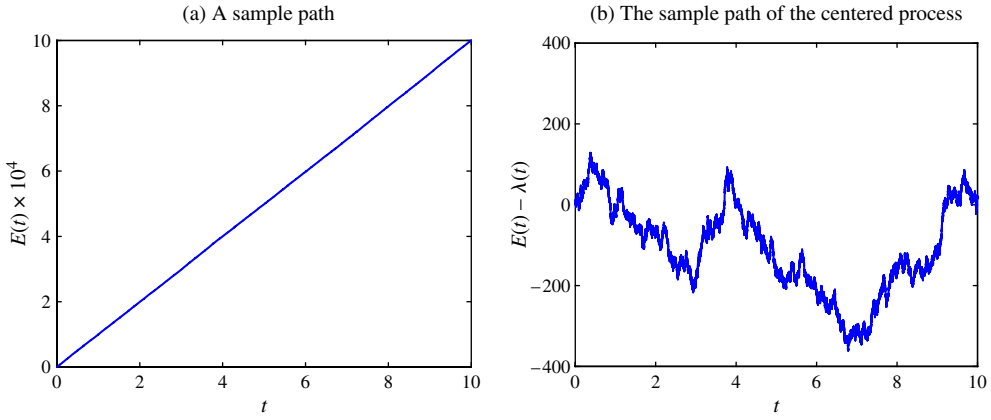


FIGURE 5. Poisson process with rate $\lambda = 10{,}000$.



is an example of a *functional central limit theorem*. Such a theorem holds for much more general arrival processes including renewal arrival processes.

For a general renewal arrival process E associated with a sequence of iid random variables that has mean m_0 and variance b_0^2, its scaled fluctuation process $\tilde{E}_\lambda$ in (17) converges to a Brownian motion with drift $\mu_B = 0$ and variance $\sigma_B^2 = b_0^2/m_0^3$. The central idea of *diffusion approximation* is to replace a scaled fluctuation process such as the one in (17) by an appropriate Brownian motion.

For a queue with a renewal arrival process and a certain service time distribution, Donsker's theorem implies that we may use Brownian motions to approximate stochastic variability in arrival and service. The diffusion model is obtained by replacing certain scaled renewal processes in system equations by Brownian motions.

7.2. Diffusion Model for $M/M/n+GI$ Queues

To illustrate the diffusion approximation of a queue, let us consider an $M/M/n+GI$ queue that has arrival rate λ, service rate μ, and the patience time distribution satisfying (14). Recall that $X(t)$ is the number of customers in the system at time t, including those in service and those waiting. Let

$$\tilde{X}(t) = \frac{1}{\sqrt{n}}(X(t) - n).$$

We call $\tilde{X} = \{\tilde{X}(t)\colon t \geq 0\}$ the *scaled customer-count* process. When the arrival rate λ is high and the square-root safety staffing rule is used so that

$$\beta = \sqrt{n}(1-\rho) \tag{18}$$

is a moderate number, we use a diffusion process Y to approximate $\tilde{X}$. To describe the diffusion process, for each function $u\colon \mathbb{R}_+ \to \mathbb{R}$, one can find a unique function y that satisfies

$$y(t) = u(t) + \mu \int_0^t y(s)^- \, \mathrm{d}s - \alpha \int_0^t y(s)^+ \, \mathrm{d}s, \quad t \geq 0,$$

where α is the patience time density at the origin defined in (14), $z^+ = \max\{z, 0\}$, and $z^- = \max\{-z, 0\}$ for any real number z. Thus, $\Psi\colon u \mapsto y$ defines a map from an arbitrary function u to another function y. Let

$$U(t) = \tilde{X}(0) - \beta\mu t + B(t),$$

where B is a $(0, \sigma_B^2)$-Brownian motion with variance

$$\sigma_B^2 = \mu(\rho + \rho \wedge 1). \tag{19}$$

Each sample path of U is a function. Thus, $Y = \Psi(U)$ is a well-defined function on each sample path. Note that Y satisfies the stochastic differential equation

$$Y(t) = \tilde{X}(0) - \beta\mu t + B(t) + \mu \int_0^t Y(s)^- \, \mathrm{d}s - \alpha \int_0^t Y(s)^+ \, \mathrm{d}s. \tag{20}$$

The stochastic differential equation (20) is the *diffusion model* for the $M/M/n+GI$ queue. Its solution $Y = \Psi(U)$ is the diffusion process that we use to approximate the scaled customer-count process $\tilde{X}$.

The drift coefficient of Y is piecewise linear, given by

$$b(z) = \begin{cases} -\beta\mu - \alpha z & \text{when } z \geq 0, \\ -\beta\mu - \mu z & \text{when } z < 0. \end{cases}$$

Suppose that $\alpha > 0$. At any time t, if $\beta \geq 0$, the drift is negative for $Y(t) > -\beta$ and is positive for $Y(t) < -\beta$; if $\beta < 0$, the drift is negative for $Y(t) > -\beta\mu/\alpha$ and is positive for $Y(t) < -\beta\mu/\alpha$. When $Y(t)$ is either too positive or too negative, this drift will "pull it back" to an equilibrium level. So over time, the process tends to evolve around its long-term mean. An OU process that has a linear drift also has the similar mean-reverting property. Because of its piecewise linear drift, Y is called *a piecewise OU process*. The piecewise OU process is analytically tractable. In particular, it admits a piecewise normal stationary distribution, which is given by

$$g(z) = \begin{cases} a_1 \exp\left(- \dfrac{\alpha(z + \alpha^{-1}\mu\beta)^2}{\sigma_B^2} \right) & \text{when } z \geq 0, \\ a_2 \exp\left(- \dfrac{\mu(z+\beta)^2}{\sigma_B^2} \right) & \text{when } z < 0, \end{cases} \tag{21}$$

where a_1 and a_2 are normalizing constants that make $g(z)$ continuous at zero. See Browne and Whitt [15] for more details. Recall that $Q(t)$ is the number of waiting customers at time t, excluding those in service, and let $Z(t)$ be the number of customers in service at time t. Clearly,

$$Q(t) = \sqrt{n}\tilde{X}(t)^+ \quad \text{and} \quad Z(t) = n - \sqrt{n}\tilde{X}(t)^-.$$

One can compute performance estimates such as the long-run average queue length $\bar{Q}$ and the long-run fraction of abandoned customers P_A for the queue. For that, let $Y(\infty)$ be a random variable that has the stationary distribution of Y. Using the stationary density (21), the long-run average queue length $\bar{Q}$ can be computed by

$$\bar{Q} \approx \sqrt{n}\mathbb{E}[Y(\infty)^+] = \sqrt{n} \int_0^\infty x g(x) \, \mathrm{d}x, \tag{22}$$

and the long-run average number of idle servers $\bar{Z}$ can be computed by

$$\bar{Z} \approx \sqrt{n}\mathbb{E}[Y(\infty)^-] = -\sqrt{n} \int_{-\infty}^0 x g(x) \, \mathrm{d}x.$$

Because $n - \bar{Z}$ is the mean number of busy servers, it follows that the abandonment fraction P_A is given by

$$P_A \approx 1 - \mu(n - \bar{Z})/\lambda. \tag{23}$$

We show the performance estimates computed by (22) and (23) from the diffusion model in Table 2 under "diffusion" columns. The table shows that the diffusion estimates agree well with the simulation results.

In the rest of this section, we give a detailed derivation of the diffusion model (20). Let $E(t)$ be the number of customer arrivals by time t, and let $S = \{S(t): t \geq 0\}$ be a Poisson process with rate one. We assume that $X(0)$, $E = \{E(t): t \geq 0\}$, and S are mutually independent. Let

$$T(t) = \int_0^t Z(s)\,\mathrm{d}s,$$

which is the cumulative service time received by *all* customers up to time t. Because μ is the service rate, $S(T(t))$ must be equal in distribution to the number of service completions. Recall that $A(t)$ is the cumulative number of abandoned customers by time t. One must have

$$X(t) = X(0) + E(t) - S(\mu T(t)) - A(t). \tag{24}$$

To derive Brownian approximations, we define several scaled processes by

$$\tilde{E}(t) = \frac{1}{\sqrt{n}}(E(t) - \lambda t), \quad \tilde{S}(t) = \frac{1}{\sqrt{n}}(S(nt) - nt),$$
$$\tilde{Z}(t) = \frac{1}{\sqrt{n}}(Z(t) - n), \quad \tilde{A}(t) = \frac{1}{\sqrt{n}}A(t).$$

Correspondingly, the dynamical equation (24) has a scaled version

$$\tilde{X}(t) = \tilde{X}(0) - \beta\mu t + \tilde{E}(t) - \tilde{S}(n^{-1}\mu T(t)) - \mu\int_0^t \tilde{Z}(s)\,\mathrm{d}s - \tilde{A}(t), \tag{25}$$

with β given in (18).

In the diffusion model, we replace the scaled primitive processes in (25) by certain Brownian motions. These approximations can be justified by Donsker's theorem. When the number of servers n is large, the corresponding diffusion process can be proved close to $\tilde{X}$. Please refer to Dai et al. [19] for related convergence results.

Because the arrival process E is a Poisson process with rate λ, the scaled process $\tilde{E} = \{\tilde{E}(t): t \geq 0\}$ is close to a Brownian motion. Note that $\tilde{E}(t)$ has mean zero and variance $\lambda t/n = \mu\rho t$. We use a Brownian motion $B_E = \{B_E(t): t \geq 0\}$ with variance $\mu\rho$ to replace $\tilde{E}$ in (25). Because S is a Poisson process with rate one, the scaled process $\tilde{S}$ can be replaced by a standard Brownian motion B_S. We assume that $X(0)$, B_E, and B_S are mutually independent. Because $T(t)$ is the cumulative service time for all customers up to t, $T(t)/(nt)$ should be close to the average utilization per server, i.e.,

$$\frac{1}{n}T(t) \approx (\rho \wedge 1)t.$$

Note that $\tilde{Z}(t)$ is the scaled number of idle servers at time t, and let $\tilde{Q}(t) = Q(t)/\sqrt{n}$ be the scaled queue length at time t. We have

$$\tilde{Q}(t) = \tilde{X}(t)^+ \quad \text{and} \quad \tilde{Z}(t) = \tilde{X}(t)^-.$$

By (16), we may approximate the scaled abandonment process by

$$\tilde{A}(t) \approx \alpha\int_0^t \tilde{X}(s)^+\,\mathrm{d}s.$$

It follows from (25) that

$$\tilde{X}(t) \approx \tilde{X}(0) - \mu\beta t + B_E(t) - B_S\big(\mu(\rho \wedge 1)t\big) + \mu\int_0^t \tilde{X}(s)^-\,ds - \alpha\int_0^t \tilde{X}(s)^+\,ds.$$

Let $B(t) = B_E(t) - B_S(\mu(\rho \wedge 1)t)$. Then $B = \{B(t): t \geq 0\}$ is a driftless Brownian motion with variance $\mu(\rho + \rho \wedge 1)$, the same one as in (19). Thus, $\tilde{X}$ is approximately a solution to

the stochastic differential equation (20). The proposed diffusion approximation is to use the solution Y to the stochastic differential equation (20) to replace $\tilde{X}$.

7.3. Diffusion Model for $M/H_2/n+GI$ Queues

By using a similar Brownian replacement procedure as in §7.2, a diffusion model was derived in He and Dai [28] for $GI/Ph/n+GI$ queues in the QED regime with phase-type service time distributions. A two-phase hyperexponential distribution, denoted as H_2, is a special case of phase-type distributions. In this section, we restrict ourselves to H_2 service time distributions and illustrate the diffusion approximation proposed by He and Dai [28].

When service times in a queue follow a two-phase hyperexponential distribution with initial distribution $p=(p_1,p_2)$ and rate $\nu=(\nu_1,\nu_2)$, one can envision two types of customers arriving at the queue. With probability p_1, a customer belongs to the first type, and his service time is exponentially distributed with mean $1/\nu_1$, and with probability p_2, he is of type two, and the service time is exponentially distributed with mean $1/\nu_2$. Then, the service rate is given by

$$\mu=\frac{1}{p_1/\nu_1+p_2/\nu_2}. \tag{26}$$

As before $\rho=\lambda/(n\mu)$ and β is given in (18).

In steady state, one expects that the customers in service are distributed between the two types following a distribution $\theta=(\theta_1,\theta_2)$, given by

$$\theta_1=\frac{p_1/\nu_1}{p_1/\nu_1+p_2/\nu_2} \quad \text{and} \quad \theta_2=\frac{p_2/\nu_2}{p_1/\nu_1+p_2/\nu_2}. \tag{27}$$

Let $X_1(t)$ and $X_2(t)$ be the respective numbers of customers of the first and the second type at time t. Because the customers in service are distributed following the distribution θ, we define its centered and scaled version by

$$\tilde{X}_j(t)=\frac{1}{\sqrt{n}}(X_j(t)-n\theta_j), \quad j=1,2.$$

In the diffusion model, we use a two-dimensional diffusion process (Y_1,Y_2) to approximate $(\tilde{X}_1,\tilde{X}_2)$, where (Y_1,Y_2) satisfies the following stochastic differential equation:

$$\begin{aligned} Y_j(t) &= Y_j(0)-\beta\mu p_j t+p_j B_E(t)+(-1)^{j-1}B_M(\rho\mu t)-B_j((\rho\wedge 1)\theta_j\nu_j t)\\ &\quad -\nu_j\int_0^t (Y_j(t)-p_j(Y_1(t)+Y_2(t))^+)\,\mathrm{d}s-p_j\alpha\int_0^t (Y_1(s)+Y_2(s))^+\,\mathrm{d}s, \end{aligned} \tag{28}$$

for $j=1,2$, where B_E is the same Brownian motion as in §7.2, B_1 and B_2 are two independent standard Brownian motions, and B_M is a Brownian motion with drift zero and variance p_1p_2. It has been proved by Dieker and Gao [20] that Y has a unique stationary distribution. The algorithm proposed by He and Dai [28] can be used to compute the stationary distribution numerically. Section 7.4 presents the performance estimates obtained from this diffusion approximation.

In the rest of this section, we derive the diffusion approximation that uses (Y_1,Y_2) to replace $(\tilde{X}_1,\tilde{X}_2)$. Let $C(i)=(C_1(i),C_2(i))$ be a two-dimensional random vector indicating the ith customer's type. The random vector takes a value of $(1,0)$ with probability p_1 and takes a value of $(0,1)$ with probability p_2. We assume that $C(1),C(2),\ldots$ are iid, so the random variable

$$M_j(k)=\sum_{i=1}^{k} C_j(i), \quad j=1,2,$$

is the number of type j customers among the first k arrivals. Let $M_j=\{M_j(k)\colon k=1,2,\ldots\}$, $M=(M_1,M_2)$, and $S_j=\{S_j(t)\colon t\ge 0\}$ be a Poisson process with rate one. We assume that $(X_1(0),X_2(0))$, E,S_1,S_2, and M are mutually independent.

Let $Z_j(t)$ denote the number of type j customers being served at time t. Then,

$$T_j(t) = \int_0^t Z_j(s)\,\mathrm{d}s \tag{29}$$

is the cumulative service time received by type j customers. Let $L_j(t)$ be the cumulative number of type j customers who have abandoned the system by time t. Then, the number of type j customers in the system must follow

$$X_j(t) = X_j(0) + M_j(E(t)) - S_j(\nu_j T_j(t)) - L_j(t) \tag{30}$$

for $j = 1, 2$. We define the scaled processes by

$$\tilde{S}_j(t) = \frac{1}{\sqrt{n}}(S_j(nt) - nt), \quad \tilde{Z}_j(t) = \frac{1}{\sqrt{n}}(Z(t) - n\theta_j),$$

$$\tilde{L}_j(t) = \frac{1}{\sqrt{n}}L_j(t), \quad \tilde{M}_j(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt \rfloor}(C_j(i) - p_j).$$

Then using (26)–(30), one can check that the scaled system equation is given by

$$\tilde{X}_j(t) = \tilde{X}_j(0) - \beta\mu p_j t + p_j\tilde{E}(t) + \tilde{M}_j(n^{-1}E(t)) - \tilde{S}_j(n^{-1}\nu_j T_j(t)) - \nu_j\int_0^t \tilde{Z}_j(s)\,\mathrm{d}s - \tilde{L}_j(t)$$

for $j = 1, 2$.

In the diffusion model for $M/H_2/n+GI$ queues, we replace $\tilde{E}$ with the Brownian motion B_E as in §7.2. The processes $\tilde{S}_1$ and $\tilde{S}_2$ are replaced by B_1 and B_2, two independent standard Brownian motions. Note that we always have $\tilde{M}_1(t) + \tilde{M}_2(t) = 0$. Hence, the process $\tilde{M}_1$ is replaced by a Brownian motion B_M with variance p_1p_2, and $\tilde{M}_2$ is replaced by $-B_M$. When the number of servers n is large, both the abandoned customers and the waiting customers in the queue are approximately distributed between the two types according to the distribution p. Hence,

$$\tilde{L}_j(t) \approx p_j\tilde{A}(t),$$

where $\tilde{A}(t)$ is the scaled number of total abandoned customers by time t as defined in §7.2. Recall that $Q(t)$ is the number of waiting customers at time t. Then,

$$Z_j(t) \approx X_j(t) - p_jQ(t).$$

Because $Q(t) = (X_1(t) + X_2(t) - n)^+$, this approximation has a scaled version,

$$\tilde{Z}_j(t) \approx \tilde{X}_j(t) - p_j(\tilde{X}_1(t) + \tilde{X}_2(t))^+.$$

We also exploit the approximations

$$\frac{E(t)}{n} \approx \frac{\lambda t}{n} = \rho\mu t, \quad \frac{T_j(t)}{n} \approx (\rho \wedge 1)\theta_j t,$$

as well as

$$\tilde{A}(t) \approx \alpha\int_0^t \tilde{Q}(s)^+\,\mathrm{d}s = \alpha\int_0^t (\tilde{X}_1(s) + \tilde{X}_2(s))^+\,\mathrm{d}s.$$

These replacements lead to the diffusion model (28) for $M/H_2/n+GI$ queues.

In our diffusion model, a two-dimensional diffusion process is used to approximate the scaled number of customers of each type. When this procedure applies to a general phase-type service time distribution with d phases, the corresponding diffusion model is a d-dimensional piecewise OU process.

7.4. Performance Estimation Using the Diffusion Model

To obtain the performance estimates of a queue using the diffusion model, one needs to know the stationary distribution of the multidimensional diffusion process. Except for the one-

dimensional case, the stationary distribution of a multidimensional piecewise OU process does not have an explicit formula. In He and Dai [28], the authors also developed a finite element algorithm computing the stationary distribution of a multidimensional diffusion process. Using the numerical results obtained by this algorithm, they demonstrated that the diffusion model is a good approximation of a many-server queue.

Consider an $M/H_2/n+M$ queue with $n=500$ servers. We set the arrival rate to be $\lambda=522.36$ customers per minute and the rate of the exponential patience time distribution to be $\alpha=0.5$. The hyperexponential service time distribution has parameters

$$p=(0.9351,0.0649) \quad \text{and} \quad \nu=(9.354,0.072).$$

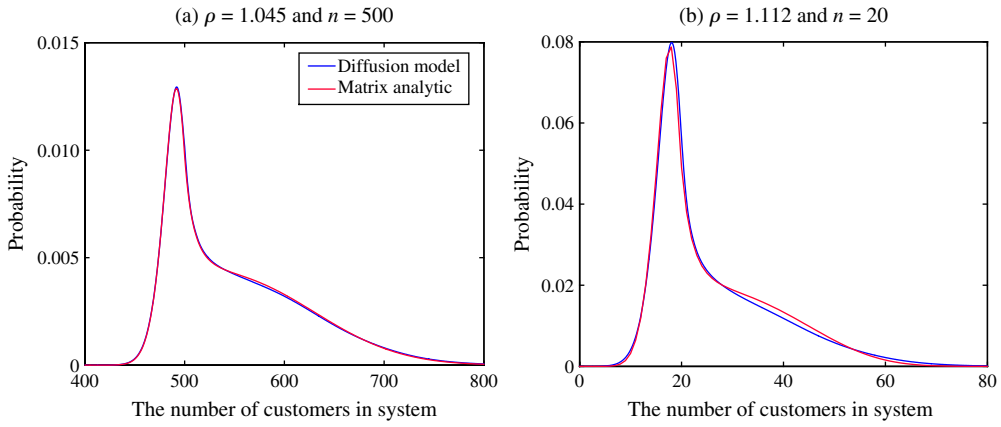
So the mean service time of the second-type customers is more than 100 times longer than that of the first type. Although over 90% of customers are of the first type, the fraction of its workload is merely 10%. One can check that the mean service time is one minute. Hence, the queue is a bit overloaded with $\rho=1.045$.

Recall that $X(t)$ is the number of customers in the system at time t. For this $M/H_2/n+M$ queue, the process X is a quasi-birth-and-death process. One can use the matrix-analytic method to solve the stationary distribution of X. See Neuts [39] for details on the matrix-analytic method. To evaluate the accuracy of the diffusion model, in Figure 6(a) we plot both the (approximate) stationary distribution of X obtained by the diffusion model and the stationary distribution produced by the matrix-analytic method. We see very good agreement between the two results.

When the number of servers is moderate, the diffusion model can still capture the dynamics of the queue. Next, we consider an $M/H_2/n+M$ queue with $n=20$ servers. Let the patience and service time distributions be the same as the previous scenario, and the arrival rate be $\lambda=22.24$. Thus, $\rho=1.112$. As illustrated by Figure 6(b), the diffusion model can still capture the exact stationary distribution for a queue with as few as 20 servers.

With an appropriate algorithm, performance estimation using the diffusion model can be much more computationally efficient than the matrix-analytic method. The computational complexity of the algorithm proposed by He and Dai [28], whether in computation time or in memory space, does not change with the number of servers n. In contrast, the matrix-analytic method becomes computationally expensive when n is large. In particular, the memory usage becomes a serious constraint when a huge number of iterations are required in the matrix-analytic method. For the $n=500$ scenario in this example, it took approximately two hours to finish the matrix-analytic computation, and the peak memory usage was nearly five gigabytes. Using the diffusion model and the proposed algorithm, it took less than one minute, and the peak memory usage was less than 200 megabytes on the same computer.

FIGURE 6. Stationary distribution of the customer number in the $M/H_2/n+M$ queue.



7.5. A Diffusion Model Using the Hazard Rate of Patience Times

In the above diffusion model (e.g., the one in (20) for an $M/M/n+GI$ queue), the patience time density α at the origin is the key parameter for modeling the abandonment process. This diffusion model, however, has its own limitations. First, one needs to estimate the patience time density at the origin using data from a service system. Estimating density at the origin is unreliable statistically. In addition, patience times are heavily censored data;– i.e., a customer's patience time can be observed only if he has abandoned the system. For a queue in the QED regime, only a small fraction of customers abandon the system. Although standard survival analysis tools, such as the Kaplan–Meier estimator (see, e.g., Cox and Oakes [16]), can be used to estimate this parameter, one has to record every customer's waiting or patience time, and a good estimate requires a large amount of data. Second, for a queue in the QED regime, no matter how small the waiting times are, the abandonment process still depends on the behavior of the patience time distribution in a neighborhood of the origin, not just at the origin. When the patience time density near the origin changes rapidly, using solely the density at the origin may not yield an adequate approximation for the abandonment process. Third, when $\alpha = 0$, the integral term corresponding to the abandonment process in the diffusion model, either (20) or (28), becomes zero. In this case, the diffusion model approximates a queue as if it has no customer abandonment. But in a queue with a zero patience time density at the origin, customer abandonment still occurs and may affect the system performance significantly. For example, if such a queue is overloaded (i.e., $\rho > 1$), it may still have a stationary distribution thanks to customer abandonment. However, the diffusion model, with $\alpha = 0$ and $\rho > 1$, does not have a stationary distribution and fails to provide any performance estimates for this queue.

A diffusion model using the entire patience time distribution was proposed by He and Dai [28]. This model exploits the idea of scaling the patience time hazard rate function, which was first proposed by Reed and Ward [44] for single-server queues and was extended to many-server queues by Reed and Tezcan [43]. This refined diffusion model provides a more accurate approximation for many-server queues.

Let us consider an $M/H_2/n+H_2$ queue in which the patience time density changes rapidly near the origin. In this queue, the hyperexponential service time distribution has

$$p = (0.5915, 0.4085) \quad \text{and} \quad \nu = (5.917, 0.454).$$

The resulting mean service time is still one minute. We assume that the patience times follow a two-phase hyperexponential distribution that has

$$p = (0.9, 0.1) \quad \text{and} \quad \nu = (1, 200).$$

Hence, 10% of customers are extremely impatient. Their mean patience time is only 0.005 minute. These customers would abandon the system right away if no servers were available upon their arrival.

Although the customer-count process X of this queue is a quasi-birth-and-death process, the extremely high computational complexity prevents the matrix-analytic method from producing the stationary distribution when n is moderate to large. See He and Dai [28] for more details. We have to simulate the queue to obtain adequate performance estimates. Two scenarios with $n = 50$ and 500 servers are investigated. The respective arrival rates are $\lambda = 57.071$ and 522.36. Thus, $\rho = 1.141$ and 1.045. Several performance estimates obtained by simulation, including the abandonment fraction, the mean queue length, and several tail probabilities, are listed in Table 4. We use $X(\infty)$ to denote the stationary number of customers in this system. In the same table, we also list the performance estimates from the diffusion model (28) with $\alpha = 20.9$. In this example, using solely the patience time density

Table 4. Performance measures of the $M/H_2/n+H_2$ queue.

	Simulation	Diffusion in (28)	Diffusion, hazard rate scaling
	(a) $\rho=1.141$ and $n=50$		
Mean queue length	4.845	0.4709	4.869
Abandonment fraction	0.1499	0.1714	0.1504
$\mathbb{P}[X(\infty)>40]$	0.9728	0.9578	0.9749
$\mathbb{P}[X(\infty)>50]$	0.6111	0.3158	0.6377
$\mathbb{P}[X(\infty)>60]$	0.1737	1.044×10^{-7}	0.1749
	(b) $\rho=1.045$ and $n=500$		
Mean queue length	6.413	1.475	6.359
Abandonment fraction	0.05512	0.05863	0.05517
$\mathbb{P}[X(\infty)>480]$	0.8881	0.8663	0.8929
$\mathbb{P}[X(\infty)>500]$	0.4720	0.3192	0.4822
$\mathbb{P}[X(\infty)>520]$	0.1050	9.274×10^{-5}	0.1074

at the origin cannot capture the behavior of the abandonment process. The diffusion model fails to produce proper performance estimates.

This issue can be fixed when the entire patience time distribution is built into the diffusion model. In the same table, we list the performance estimates obtained by the diffusion model using the hazard rate scaling (see He and Dai [28, §§4.3 and 6] for more details). This time, we see good agreement between the refined diffusion model and the simulation results.

Next, we consider an $M/H_2/n+E_3$ queue, where $+E_3$ signifies an Erlang-3 patience time distribution. In this queue, each patience time is the sum of three stages, and the stages are iid following an exponential distribution with a mean of one-third of a minute. So the mean patience time is one minute. The density at the origin of an Erlang-3 distribution is zero. The diffusion model (28) has $\alpha=0$ and hence does not have a stationary distribution when $\rho>1$. In this queue, the hyperexponential service time distribution is taken to be identical to that of the previous $M/H_2/n+H_2$ queue.

We study two scenarios, with $n=50$ and 500 servers, respectively. The arrival rates are $\lambda=57.071$ and 522.36. Then, $\rho=1.141$ and 1.045. We list performance estimates from simulation and from the diffusion model using the hazard rate scaling in Table 5. As in the previous example, the refined diffusion model produces adequate performance approximations.

Table 5. Performance measures of the $M/H_2/n+E_3$ queue.

	Simulation	Diffusion, hazard rate scaling
	(a) $\rho=1.141$ and $n=50$	
Mean queue length	19.31	19.44
Abandonment fraction	0.1305	0.1303
$\mathbb{P}[X(\infty)>45]$	0.9645	0.9704
$\mathbb{P}[X(\infty)>50]$	0.9066	0.9169
$\mathbb{P}[X(\infty)>70]$	0.4761	0.5037
	(b) $\rho=1.045$ and $n=500$	
Mean queue length	119.1	119.5
Abandonment fraction	0.04337	0.04340
$\mathbb{P}[X(\infty)>480]$	0.9940	0.9946
$\mathbb{P}[X(\infty)>500]$	0.9756	0.9770
$\mathbb{P}[X(\infty)>600]$	0.6645	0.6733

8. Summary

With the same server utilization, delays in single-server queues and in many-server queues are qualitatively different. As such, one should not apply a single-server-based approximation for the mean waiting time to study a service system that has at least a moderate number of servers working in parallel. For example, one should avoid using such an approximation to study the congestion of a hospital impatient department, in which hospital beds are modeled as servers. For a service system that can be modeled by a many-server queue, managers should staff the system following the square-root safety staffing rule so that the system operates in the QED regime, achieving both a high level of service quality and a high level of server utilization. When a system has a significant amount of abandonment, it is critical to model the customer abandonment explicitly. It is the behavior of the patience time distribution near the origin, not the mean patience time, that has the most impact on the performance of a system operating in the QED regime. Diffusion models can be a practical tool, sometimes the only tool besides computer simulation, to evaluate the performance of a many-server queue when the service times have a phase-type distribution and the patience time distribution is general.

References

[1] R. Atar, C. Giat, and N. Shimkin. The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research* 58(5):1427–1439, 2010.

[2] R. Atar, C. Giat, and N. Shimkin. On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost. *Queueing Systems* 67(2):127–144, 2011.

[3] R. Atar, A. Mandelbaum, M. I. Reiman. Scheduling a multiclass queue with many exponential servers: asymptotic optimality in heavy traffic. *Annals of Applied Probability* 14(3):1084–1134, 2004.

[4] F. Baccelli, P. Boyer, and G. Hebuterne. Single-server queues with impatient customers. *Advances in Applied Probability* 16(4):887–905, 1984.

[5] A. Bassamboo and R. S. Randhawa. On the accuracy of fluid models for capacity planning in queueing systems with impatient customers. *Operations Research* 58(5):1398–1413, 2010.

[6] A. Bassamboo, J. M. Harrison, and A. Zeevi. Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems* 51(3–4):249–285, 2005.

[7] A. Bassamboo, J. M. Harrison, and A. Zeevi. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Operations Research* 54(3):419–435, 2006.

[8] P. P. Bhattacharya and A. Ephremides. Stochastic monotonicity properties of multiserver queues with impatient customers. *Journal of Applied Probability* 28(3):673–682, 1991.

[9] P. Billingsley. *Convergence of Probability Measures*, 2nd ed. John Wiley & Sons, New York, 1999.

[10] O. J. Boxma and P. R. de Waal. Multiserver queues with impatient customers. J. Labetoulle and J. W. Roberts, eds. *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, Vol. 1. North Holland, Amsterdam, 743–756, 1994.

[11] A. Brandt and M. Brandt. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation* 35(1–2):1–18, 1999.

[12] A. Brandt and M. Brandt. Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s+GI$ system. *Queueing Systems* 41(1–2):73–94, 2002.

[13] E. Brockmeyer, H. L. Halstrøm, and A. Jensen. The life and works of A. K. Erlang. Transactions of the Danish Academy of Technical Sciences, Copenhagen, Denmark, 1948.

[14] L. Brown, A. Mandelbaum, A. Sakov, S. Zeltyn, L. Zhao, and H. Shen. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* 100(2):36–50, 2005.

[15] S. Browne and W. Whitt. Piecewise-linear diffusion processes. J. Dshalalow, ed. *Advances in Queueing*. CRC Press, Boca Raton, FL, 463–480, 1995.

[16] D. R. Cox and D. Oakes. Analysis of survival data. *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1984.

[17] J. G. Dai and S. He. Customer abandonment in many-server queues. *Mathematics of Operations Research* 35(2):347–362, 2010.

[18] J. G. Dai and T. Tezcan. Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems* 59:95–134, 2008.

[19] J. G. Dai, S. He, and T. Tezcan. Many-server diffusion limits for $G/Ph/n+GI$ queues. *Annals of Applied Probability* 20(5):1854–1890, 2010.

[20] A. B. Dieker and X. Gao. Positive recurrence of piecewise Ornstein-Uhlenbeck processes and common quadratic Lyapunov functions. Technical report, Georgia Institute of Technology, Atlanta, 2011.

[21] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2):79–141, 2003.

[22] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* 4(3):208–227, 2002.

[23] W. K. Grassmann. Is the fact that the emperor wears no clothes a subject worthy of publication? *Interfaces* 16(2):43–51, 1986.

[24] W. K. Grassmann. Finding the right number of servers in real-world queuing systems. *Interfaces* 18(2):94–104, 1988.

[25] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory.* Wiley Series in Probability and Statistics: Texts and References Section, 3rd ed. Wiley-Interscience, New York, 1998.

[26] I. Gurvich and W. Whitt. Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research* 34(2):363–396, 2009.

[27] S. Halfin and W. Whitt. Heavy-traffc limits for queues with many exponential servers. *Operations Research* 29:567–588, 1981.

[28] S. He and J. G. Dai. Many-server queues with customer abandonment: Numerical analysis of their diffusion models. Working paper, Georgia Institute of Technology, Atlanta, 2011.

[29] A. J. E. M. Janssen, J. S. H. van Leeuwaarden, and B. Zwart. Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. *Advances in Applied Probability* 40(1):122–143, 2008.

[30] N. L. Johnson and S. Kotz. *Discrete Distributions.* Houghton Mifflin, Boston, 1969.

[31] W. Kang and K. Ramanan. Fluid limits of many-server queues with reneging. *Annals of Applied Probability* 20(6):2204–2260, 2010.

[32] J. F. C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*) 57:902–904, 1961.

[33] Y. L. Koçağa and A. R. Ward. Admission control for a multi-server queue with abandonment. *Queueing Systems* 65(3):275–323, 2010.

[34] P. Kolesar. Comment on "Is the fact that the emperor wears no clothes a subject worthy of publication?" *Interfaces* 16(2):50–51, 1986.

[35] S. C. H. Lu, D. Ramaswamy, and P. R. Kumar. Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants. *IEEE Transactions on Semiconductor Manufacturing* 7(3):374–388, 1994.

[36] A. Mandelbaum and P. Momčilović. Queues with many servers and impatient customers. Working paper, Technion, Haifa, Israel, 2009.

[37] A. Mandelbaum and S. Zeltyn. The impact of customers' patience on delay and abandonment: some empirically-driven experiments with the $M/M/n+G$ queue. *OR Spektrum* 26:377–411, 2004.

[38] A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research* 57(5):1189–1205, 2009.

[39] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithm Approach.* John Hopkins University Press, Baltimore, 1981.

[40] G. F. Newell. *Approximate Stochastic Behavior of n-Server Service Systems with Large n.* Springer-Verlag, Berlin, 1973.

[41] G. F. Newell. *Applications of Queueing Theory.* Chapman-Hall, London, 1982.

[42] J. Reed. The $G/GI/N$ queue in the Halfin-Whitt regime. *Annals of Applied Probability* 19(6):2211–2269, 2009.

[43] J. Reed, T. Tezcan. Hazard rate scaling for the $GI/M/n+GI$ queue. Working paper, New York University, New York, 2009.

[44] J. E. Reed and A. R. Ward. Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Mathematics of Operations Research* 33(3):606–644, 2008.

[45] P. Shi, D. Ding, J. Ang, M. Chou, and J. G. Dai. NUH impatient operations: Empirical analsis and mathematical models. Working paper, Georgia Institute of Technology, Atlanta, GA, 2011.

[46] R. Talreja and W. Whitt. Heavy-traffic limits for waiting times in many-server queues with abandonment. *Annals of Applied Probability* 19(6):2137–2175, 2009.

[47] T. Tezcan and J. G. Dai. Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Operations Research* 58(1):94–110, 2010.

[48] W. Whitt. Understanding the efficiency of multi-server service systems. *Management Science* 38(5):708–723, 1992.

[49] W. Whitt. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* 50(10):1449–1461, 2004.

[50] W. Whitt. Fluid models for multiserver queues with abandonments. *Operations Research* 54(1):37–54, 2006.

[51] R. W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs, NJ, 1989.

[52] S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems* 51(3–4):361–402, 2005.

[53] J. Zhang. Fluid models of multi-server queues with abandonment. Technical report, Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Hong Kong, 2009.