J Syst Sci Syst Eng (Mar 2012) 21(1): 001-036 ISSN: 1004-3756 (Paper) 1861-9576 (Online)
DOI: 10.1007/s11518-012-5189-y CN11-2983/N

MANY-SERVER QUEUES WITH CUSTOMER ABANDONMENT: A SURVEY OF DIFFUSION AND FLUID APPROXIMATIONS*

J. G. DAI[1] Shuangchi HE[2]

[1]*H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA. dai@gatech.edu*(✉)

[2]*Department of Industrial and Systems Engineering, National University of Singapore, Singapore 11757 isehes@nus.edu.sg*

Abstract

The performance of a call center is sensitive to customer abandonment. In this survey paper, we focus on $G/GI/n+GI$ parallel-server queues that serve as a building block to model call center operations. Such a queue has a general arrival process (the G), 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, this queue can be operated in the quality- and efficiency-driven (QED) regime, which is characterized by large customer volume, the waiting times being a fraction of the service times, only a small fraction of customers abandoning the system, and high server utilization. Operational efficiency is the central target in a system whose staffing costs dominate other expenses. If a moderate fraction of customer abandonment is allowed, such a system should be operated in an overloaded regime known as the efficiency-driven (ED) regime. We survey recent results on the many-server queues that are operated in the QED and ED regimes. These results include the performance insensitivity to patience time distributions and diffusion and fluid approximate models as practical tools for performance analysis.

Keywords: Heavy traffic, square-root safety staffing, quality- and efficiency-driven regime, efficiency-driven regime, piecewise OU process

1. Introduction

Customer call centers have become an important part of the service economy in a modern society. To take advantage of the economy of scale, call centers with hundreds of agents are ubiquitous in many industries. These systems face a large amount of daily traffic that is intrinsically stochastic and has temporal variations. In a call center, a customer waiting for service may hang up the phone before being served. This is called *customer abandonment*. Such a phenomenon is common because customers usually have limited patience. Customer expectation demands that a proper

*Research supported in part by NSF grants CMMI-0825840 and CMMI-1030589.

staffing level be maintained in the call center so that most customers are served without waiting for a long time and only a small fraction of customers abandon the system. As pointed by Garnett et al. (2002), customer abandonment is a crucial factor for call center operations. It may significantly impact the system performance and must be modeled explicitly in order for an operational model to be relevant for decision making.

In this paper, we focus on a mathematical model that is denoted by 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 system without service. In the $G/GI/n+GI$ notation, the G refers to a general arrival process, the first GI refers to independent and identically distributed (iid) service times with a general distribution, n is the number of identical servers, and $+GI$ refers to iid patience times with a general distribution. As we do not assume iid interarrival times, the symbol for the arrival process is G, not GI. 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 call centers. For call center operations, it is reasonable to assume that the patience times are iid, as the queue is usually invisible to waiting customers.

As argued by Halfin & Whitt (1981), the performance of many-server queues is qualitatively different from that of single-server queues or queues with a small number of servers. Due to the stochastic variability in inter arrival and service times, the mean customer waiting time in a single-server queue goes to infinity as the server approaches 100% utilization. A manager has to make a painful choice between *quality* of service (short waiting times) and operational *efficiency* (high server utilization) in a single-server service system. In contrast, a many-server system can be operated in the *quality- and efficiency-driven* (QED) regime that is characterized by large customer volume, the mean waiting time being a fraction of the mean service time, only a small fraction of customers abandoning the queue, and high server utilization. This regime is also called the *rationalized* regime in Garnett et al. (2002) because in most cases, a manager *should* operate his service system in such a regime. To achieve both quality and efficiency, the manager can exploit the *pooling* effect by operating a large number of servers in parallel. More specifically, the manager can apply the *square-root safety staffing rule* to drive the system to the QED regime. This rule is an important staffing principle that is both theoretically justified and widely practiced. In certain service systems, the staffing costs dominate the costs of customer delay and abandonment. In such systems, a more reasonable operational regime is the *efficiency-driven* (ED) regime. In the ED regime, the service capacity is set below the customer arrival rate by a moderate fraction. In such a many-server system, the mean waiting time is comparable to the mean service time, a moderate fraction of customers abandon the system, and all servers are almost always busy. These two operational regimes are elaborated in Section 2.

It is empirically reported that in call centers, both the service time distributions and the patience time distributions are far from

exponential; see, e.g., Brown et al. (2005). Therefore, one must use general distributions to model service and patience times. Recent papers, such as Zeltyn & Mandelbaum (2005), Dai & He (2010), and Mandelbaum & Momčilović (2012), have demonstrated that the performance of 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. This phenomenon is discussed in Section 3.

When the service and patience time distributions are general, except by computer simulation, no analytical or numerical methods are available to evaluate the performance of such a queue. We survey approximate models for many-server queues in Sections 4 and 5. In Section 4, we study diffusion approximations for many-server queues in the QED regime. In these diffusion models, the service time distribution is modeled by a phase-type distribution and the patience time distribution is assumed to be general. We demonstrate that the diffusion models are accurate in predicting the system performance in the QED regime. For a many-server queue in the ED regime, fluid approximations have been shown to be useful. In Section 5, we survey a fluid model proposed by Whitt (2006). This fluid model is shown to be adequate in estimating the performance of a many-server queue in the ED regime.

2. Operational Regimes in the Presence of Customer Abandonment

As argued by Whitt (2006), most service systems can be classified into two types: revenue-generating systems and service-oriented systems. The former type aims to maintain high quality of service while the latter type focuses more on operational efficiency. As the scale of service systems goes large, the congestion due to variability in customer service demands can be offset by pooling service facilities. If the customer arrival rate and the service capacity are well balanced, it is possible to achieve both quality and efficiency in a service system with many servers. In this case, the service system is said to be working in the quality- and efficiency-driven (QED) regime. We demonstrate in Section 2.1 that one can follow the square-root safety staffing rule to operate a service system in the QED regime.

Customer abandonment may have a significant impact on the performance of a service system. For a service system whose customers are human beings, one must consider the abandonment phenomenon in decision making and operations. In Section 2.2, we demonstrate that in the presence of customer abandonment, the square-root safety staffing rule can still lead the system to the QED regime while keeping a small abandonment fraction.

To achieve the most operational efficiency in a service-oriented system, a certain percentage (say, 15% to 20%) of customer abandonment is usually allowed. In this case, the service system can be operated in an overloaded regime while still maintaining certain service levels. The key insight here is that customer abandonment could compensate for a slight excess in the arrival rate over the service capacity. Such a service system is highly efficient because all servers are busy almost all the time. This overloaded regime for a many-server system is called the efficiency-driven (ED) regime. We introduce the ED regime in Section 2.3. Usually, the arrival rate of

a service system is time-varying. It is worth mentioning that despite the best efforts, there would be certain time periods in which the service system has to be operated in the overloaded regime. These two regimes, QED and ED, were coined by Mandelbaum for many-server queues; see Gans et al. (2003) for more details.

2.1 The QED Regime and the Square-root Safety Staffing Rule

In a service system with a single or a small number of servers, the manager has to strike a compromise between service quality and operational efficiency. In contrast, it is possible to achieve both of them in a service system with many parallel servers. We use a numerical example to demonstrate this distinction.

To evaluate the quality of service in a service system, the fraction of customers who have to wait before receiving service (known as the delay probability) and the mean customer waiting time are two important performance measures. These two measures should be maintained under certain levels to meet customer expectations. Let us consider an $M/M/n$ queue that has a Poisson arrival process with rate λ, exponentially distributed service times with mean $1/\mu$, and n identical parallel servers. The traffic intensity of this queue is defined by

$$\rho = \frac{\lambda}{n\mu} \quad . \tag{1}$$

We assume that $\rho < 1$, which is also equal to the average utilization per server. The delay probability is given by the well-known Erlang-C formula (see, e.g., Gross & Harris (1985)),

$$P_{\mathrm{w}} = \frac{(n\rho)^n}{n!}\left((1-\rho)\sum_{k=0}^{n-1}\frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!}\right)^{-1} . \tag{2}$$

The mean waiting time is

$$\bar{W} = \frac{P_{\mathrm{w}}}{(1-\rho)n\mu} \quad . \tag{3}$$

and by Little's law, the mean queue length is

$$\bar{Q} = \lambda \bar{W} . \tag{4}$$

We evaluate the waiting time factor f_{w}, which is defined as the ratio of the mean waiting time to the mean service time, in the following numerical example. By this definition, the waiting time factor is

$$f_{\mathrm{w}} = \mu \bar{W} = \frac{P_{\mathrm{w}}}{(1-\rho)n} \quad . \tag{5}$$

for the $M/M/n$ queue.

In Figure 1, we plot the delay probability and the waiting time factor as the number of servers increases from 1. All these queues have the same traffic intensity $\rho = 0.95$. 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 0.85. Thus, 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. This level of quality of service is considered good for many service systems. If one further increases the number of servers to $n = 100$ and the average utilization per server is kept at 95%, the delay probability decreases to 50.7% and the waiting time factor is 0.101. In this case, nearly half of customers are served without delay and the mean waiting time is only around 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 is operated in the QED regime. In this 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 equal so that the server utilization is close to 1.

Even though the average server utilization is close to 1, only a fraction of customers need to wait in the queue with many parallel servers. This phenomenon is in sharp contrast to the observation that almost all customers have to wait in a single-server queue. To illustrate how the waiting time factor changes with the server utilization, we also plot the waiting time factor curves for a single-server queue and for a queue with 18 servers in Figure 2. Compared with the curve for the single-server queue, the waiting time factor for the multi-server queue increases much more slowly as the server utilization approaches 1.

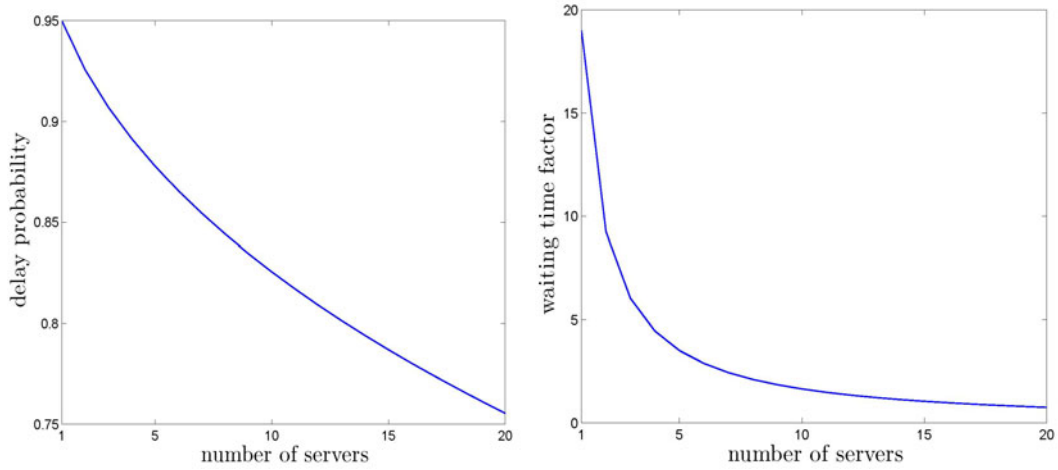


Figure 1 Delay probability and waiting time factor vs number of servers, for $M/M/n$ queues with $\rho = 0.95$

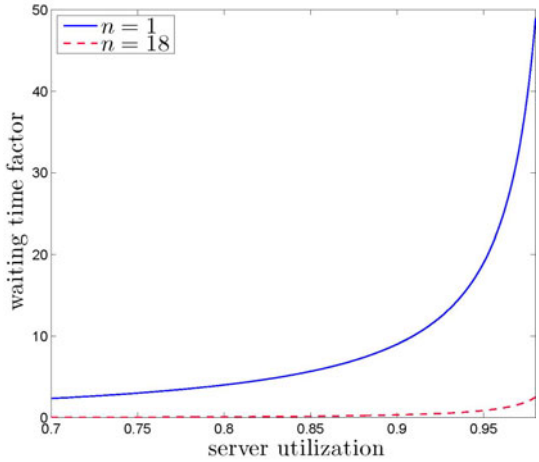


Figure 2 Waiting time factor vs server utilization, for an $M/M/1$ queue and for an $M/M/18$ queue

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. Figure2, 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.

The QED regime can be achieved by applying the square-root safety staffing rule. Let $R = \lambda / \mu$ be the offered load of the queue. One expects that an appropriate staffing level (i.e., the number of servers) 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 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 is

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

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 $O(n^{-1/2})$. It was proved by Halfin & Whitt (1981) that when the offered load R is high in the $M/M/n$ setting,

$$P_{\mathrm{w}} \approx \frac{1}{\beta\Phi(\beta)/\phi(\beta)+1} \quad . \tag{7}$$

Where ϕ and Φ are the probability density and the cumulative distribution function, respectively, of the standard normal distribution.

Formulas (6) and (7) can be used for performance analysis with a given staffing level, or to determine the staffing level that achieves a given delay probability. For a given staffing level n and a given utilization level $\rho < 1$, we can set

$$\beta = \sqrt{n}(1-\rho) \tag{8}$$

and use the right side of (7) to approximate the delay probability P_{w}. With the delay probability, the waiting time factor f_{w} in (5) becomes

$$f_{\mathrm{w}} = \frac{P_{\mathrm{w}}}{\sqrt{n}\beta}. \tag{9}$$

For example, when $n = 100$ and $\rho = 0.95$, one has $\beta = 0.5$, the right side of (7) predicts P_{w} to be 0.505, compared to the exact value 0.507 from (2). The waiting time factor computed through (5) is 0.101 based on both the exact and approximate values of P_{w}.

The second and more important usage of (6) is that it leads to the following staff provision in the $M/M/n$ setting. Suppose that the delay probability is required to be less than a target value $0 < \eta < 1$. One needs to set the staffing level so that the delay probability is approximately η. For this, one first solves for β from the following equation

$$\eta = \frac{1}{\beta\Phi(\beta)/\phi(\beta)+1},$$

and then set the staffing level n using (6) for a given offered load R.

2.2 Modeling Customer Abandonment

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. It has n identical servers, the arrival process is Poisson with rate λ, and the service times are iid following an exponential distribution with mean $1/\mu$. In this queue, each customer has a patience time and the patience times are iid following an exponential distribution with mean $1/\alpha$. This model is also known as the Erlang-A model. The traffic intensity of a queue with customer abandonment is still given by (1), but it should no longer be understood as the average server utilization since the customers who abandon the system do not receive any service. The $M/M/n+M$ queue is also a tractable model whose performance measures have explicit formulas. For example, the delay probability is

$$P_{\mathrm{w}} = \frac{\Gamma(\alpha^{-1}n\mu, \alpha^{-1}\lambda)E_{1,n}}{1+(\Gamma(\alpha^{-1}\mu n, \alpha^{-1}\lambda)-1)E_{1,n}}, \qquad (10)$$

where

$$\Gamma(x,y) = \frac{x\exp(y)}{x^y}\int_0^y t^{x-1}\exp(-t)\mathrm{d}t$$

for $x>0$ and $y\geq 0$, and

$$E_{1,n} = \frac{(n\rho)^n}{n!}\left(\sum_{k=0}^{n}\frac{(n\rho)^k}{k!}\right)^{-1}$$

denotes the blocking probability in the $M/M/n/n$ (Erlang-B) model. In addition, the fraction of customers who abandon the system is

$$P_{\mathrm{a}} = \left(\frac{1}{\rho\Gamma(\alpha^{-1}\mu n, \alpha^{-1}\lambda)}+1-\frac{1}{\rho}\right)P_{\mathrm{w}}, \qquad (11)$$

the mean waiting time (among all customers including those who have abandoned the system) is

$$\bar{W} = \frac{P_{\mathrm{a}}}{\alpha}, \qquad (12)$$

and the mean queue length is

$$\bar{Q} = \lambda\bar{W} \qquad (13)$$

by Little's law. Using the formula of the abandonment fraction P_{a}, the average utilization per server can be computed by

$$\frac{\lambda(1-P_{\mathrm{a}})}{n\mu} = \rho(1-P_{\mathrm{a}}). \qquad (14)$$

See Mandelbaum & Zeltyn (2007) for the complete details on the Erlang-A model.

We assume that the $M/M/n+M$ queue has $n=50$ servers, the arrival rate is $\lambda=55$ customers per minute, the mean service time is 1 minute, and the mean patience time is 2 minutes. Several performance measures of this queue, obtained via formulas (10)–(14), 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^* of the modified queue is equal to the throughput of the original queue, i.e., $\lambda^* = 55\times(1-0.102) = 49.39$. The corresponding performance measures can be obtained by formulas (2)–(4) and the fact that the average server utilization is equal to the traffic intensity. Table 1 shows that both the mean waiting time and the mean queue length in the original queue are much shorter than the corresponding quantities in the modified queue. In other words, with the same service capacity and throughput, some key performance measures in a queue with abandonment are much better than in a queue without abandonment. The performance

measures of the original queue indicate that the system is working in the QED regime, even though it is slightly overloaded (i.e., $\rho > 1$). 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 abandonment, when the system is in a congestion period, the customers who experience long waiting abandon the system. Their waiting times are capped by their patience times. In the modified queue without abandonment, these customers will experience extremely long delays, which degrades the overall waiting time statistics. With customer abandonment, a service system can reach a steady state even if the customer arrival rate is larger than its service capacity, i.e., $\rho > 1$. As more and more customers accumulate in the buffer, the abandonment rate keeps increasing until arrivals and departures (including both service completions and abandonments) reach an equilibrium.

The square-root safety staffing rule (6) also applies to a service system with high offered load in the presence of customer abandonment. It was proved by Garnett et al. (2002) that when the staffing level of the $M/M/n+M$ queue follows (6), the delay probability can be approximated by

$$P_{\mathrm{w}} \approx \left(1+\frac{h\left(\beta\sqrt{\alpha^{-1}\mu}\right)}{\sqrt{\alpha^{-1}\mu}h(-\beta)}\right)^{-1}, \tag{15}$$

where

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

is the hazard rate function of the standard normal distribution. Therefore, to meet a target delay probability $0<\eta<1$, one can set the staffing level using (6), with the value of β determined by solving

$$\eta=\left(1+\frac{h\left(\beta\sqrt{\alpha^{-1}\mu}\right)}{\sqrt{\alpha^{-1}\mu}h(-\beta)}\right)^{-1}. \tag{16}$$

Table 1 Comparison between queues with and without customer abandonment

	$M/M/50+M$	$M/M/50$
Delay probability (%)	86.6	89.8
Abandonment fraction (%)	10.2	N/A
Mean waiting time (in seconds)	12.5	87.7
Mean queue length	11.2	72.2
Server utilization (%)	98.8	98.8

When solving (16) for β, it is possible to have a negative solution that results in a staffing level below the offered load. Because the small fraction of customer abandonment reduces the excess service demands, the service system can still achieve a satisfactory quality of service when the system is slightly overloaded (as indicated by the performance of the $M/M/n+M$ queue in Table 1). Zeltyn & Mandelbaum (2005) proved that when the staffing level n follows (6) with a given β (which is *not* necessarily positive), the fraction of customers who abandon the $M/M/n+M$ queue is approximately

$$P_{\mathrm{a}} \approx \frac{1}{\sqrt{n}}\left(\sqrt{\mu^{-1}\alpha}h(\beta\sqrt{\alpha^{-1}\mu})-\beta\right) \times\left(1+\frac{h(\beta\sqrt{\alpha^{-1}\mu})}{\sqrt{\alpha^{-}\mu}h(-\beta)}\right)^{-1}. \quad (17)$$

We can see that P_{a} is on the order of $O(n^{-1/2})$ as $n\to\infty$. It follows from (12) that the mean waiting time is on the same order, and it follows from (14) that the average server utilization is close to one. Therefore, in the presence of customer abandonment, the square-root safety staffing rule still leads the system to the QED regime and yields high server utilization, short waiting times, and a very small abandonment fraction.

Diffusion models have been demonstrated to be accurate in estimating the performance of many-server queues in the QED regime. We will survey these diffusion models in Section 4.

2.3 The ED Regime

In certain service systems, the staffing costs dominate the expenses of customer delay and abandonment. For these systems, the rational operational regime is the ED regime that emphasizes server utilization over quality of service. In this regime, the arrival rate exceeds the service capacity by a moderate fraction (e.g., 20%). More precisely, the ED regime requires that the traffic intensity ρ be greater than 1 and that the order of $\rho-1$ be $O(1)$ as $n\to\infty$. Since the fraction of customers who cannot be served is at least $\rho-1$, the fraction of customer abandonment in the ED regime must also be on the order of $O(1)$ as $n\to\infty$. Note that if the square-root safety staffing rule (6) is applied with $\beta<0$, the resulting service capacity is also below the arrival rate. In this case, however, both $\rho-1$ and the fraction of customer abandonment are on the order of $O(n^{-1/2})$. The system is in the QED regime, not in the ED regime.

Because the system is overloaded, almost all customers are delayed in the buffer and all servers are busy nearly 100% of the time in the ED regime. Although it might be counterintuitive, a service system operated in the ED regime can still result in reasonable performance as measured by the mean waiting time and the fraction of customer abandonment. This is because the lost service demands of the abandoned customers compensate for the excess in the arrival rate over the service capacity. A fluid model proposed by Whitt (2006) has been shown to be useful in estimating the performance of a many-server queue in the ED regime. We will survey this model and discuss the performance of queues in the ED regime in Section 5.

A fluid model was studied by Bassamboo & Randhawa (2010) to solve the staffing problem

of an $M/M/n+GI$ queue. The goal is to balance the staffing costs and the costs from customer delay and abandonment. Since the exact optimization is not possible, they employed the fluid model to approximate the queue. They proved that if the patience time distribution has a non-decreasing hazard rate, it is asymptotically optimal for the system to operate in the QED regime with the service capacity approximately equal to the arrival rate. However, if the patience time distribution has a decreasing hazard rate, the operation costs are reduced when the system is overloaded. Hence, the optimized staffing level drives the queue to the ED regime. In the same paper, the authors also proved that in the steady state, the accuracy gaps of the fluid approximations for the mean queue length and the rate of customer abandonment do *not* increase with the arrival rate. This implies that the fluid approximations could be particularly accurate when the underlying system is operated in the ED regime.

3. Performance Insensitivity to Patience Time Distributions in the QED Regime

As we have demonstrated in the previous sections, service systems operated in the QED regime are characterized by short customer waiting times. For an $M/M/n+M$ queue in the QED regime with $\beta=\sqrt{n}(1-\rho)$ being fixed, it can be seen from (12) and (17) that the mean waiting time decreases to zero at rate $n^{-1/2}$ as the staffing level n goes to infinity. If a service system has hundreds of parallel servers 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, has little influence on the system dynamics. Such a result can be confirmed by the numerical example below.

Consider an $M/M/n+GI$ queue. Let F be the cumulative distribution function of the patience times that satisfies

$$F(0)=0 \quad \text{and} \quad \alpha=\lim_{x\downarrow 0} x^{-1}F(x)<\infty, \quad (18)$$

where 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 1. This system is slightly overloaded but still in the QED regime. A small fraction of traffic, at least $(\lambda-100)/\lambda=4.8\%$ of the arrivals, has to abandon the system. We consider three patience time distributions with the same density 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 distribution (H_2). 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 has

$p = (0.21,\ 0.79)$ and $\nu = (0.3\alpha, 79\alpha/30)$. Thus, 21% of customers have long patience times with mean $10/(3\alpha)$ and79% of customers have short patience times with mean $30/(79\alpha)$. Equivalently, the density function of the hyperexponential patience time distribution is given by

$$f(x) = 0.21\alpha\exp(-0.3\alpha x) + 0.79\alpha\exp(-79\alpha x/30), \quad x \geq 0. \quad (19)$$

The squared coefficient of variation of this distribution is 1.612. All three distributions have density α at the origin.

The exact formulas for several performance measures of the $M/M/n+GI$ model are summarized in Section 9 of Zeltyn & Mandelbaum (2005). We follow these formulas to obtain the abandonment fraction and the mean queue length. 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 "Diffusion" column in Table 2 will be explained in Section 4.2.)

This 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, sometimes the matrix-analytic method can be effective to compute the performance of a queue with an exponential patience time distribution. The computed performance is in turn used to approximate the original queue with a general patience time distribution. Section 4 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, the mean queue lengths are given for different patience time distributions, including an exponential distribution with rate $\alpha = 1/m$, a uniform distribution on $[0, 2m]$ with $\alpha = 1/(2m)$, and a hyperexponential distribution given by (19) with $\alpha = 2.447/m$. Table 3 shows that for each fixed m, 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 & Mandelbaum (2005) for the steady-state analysis of $M/M/n+GI$ queues and was later elaborated by Dai & He (2010) for the process level analysis under the $G/G/n+GI$ setting. In Dai & He (2010), a

Table 2 Performance insensitivity to patience time distributions

	Abandonment fractions (%)				Mean queue length			
	Exp	Uniform	H_2	Diffusion	Exp	Uniform	H_2	Diffusion
$\alpha = 0.1$	4.97	4.98	4.96	4.97	52.2	50.6	54.2	52.2
$\alpha = 0.5$	6.04	6.08	5.99	6.03	12.7	12.1	13.4	12.7
$\alpha = 1$	6.70	6.76	6.62	6.69	7.03	6.58	7.59	7.02
$\alpha = 2$	7.40	7.48	7.30	7.38	3.88	3.55	4.31	3.88
$\alpha = 10$	8.86	9.02	8.69	8.86	0.93	0.75	1.17	0.93

Table 3 Mean patience time is a wrong statistic

	Abandonment fractions (%)			Mean queue length		
	Exp	Uniform	H_2	Exp	Uniform	H_2
$m = 0.1$	8.86	8.40	9.27	0.93	1.50	0.58
$m = 0.5$	7.40	6.76	8.12	3.88	6.58	2.08
$m = 1$	6.70	6.08	7.49	7.03	12.1	3.66
$m = 2$	6.04	5.50	6.82	12.7	22.1	6.44
$m = 10$	4.97	4.81	5.43	52.2	98.1	24.5

deterministic relationship is established between the abandonment processes and the queue length processes for many-server queues. 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 this constant should be interpreted as the abandonment rate per unit of waiting time. It was proved by Dai & He (2010) 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. The relationship says that 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

$$A(t) \approx \alpha\int_0^t Q(s)\mathrm{d}s \tag{20}$$

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

4. Diffusion Models for Many-Server Queues in the QED Regime

The exact analysis of a many-server queue with customer abandonment has largely been limited to the $M/M/n+M$ model, which has a Poisson arrival process and exponential service and patience time distributions. However, as pointed out by Brown et al. (2005), the service time distribution in a call center appears to

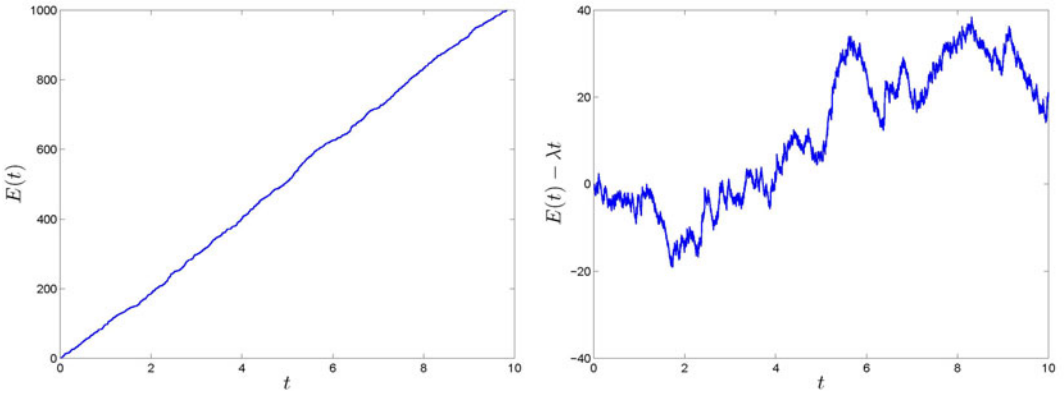
follow a log-normal distribution. In Zeltyn & Mandelbaum (2005), the patience time distribution in a call center has also been observed to be far from exponential. With general service and patience time distributions, there is no finite-dimensional Markovian representation of the queue. Except computer simulations, no methods are available to analyze such a queue either analytically or numerically. Much attention has been devoted to the approximate analysis of such a queue.

In our approximate analysis, 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 (1981) for more discussion on phase-type distributions. For a $GI/Ph/n+GI$ queue with a phase-type service time distribution, two multidimensional diffusion processes were proposed by He & Dai (2011) to approximate the dynamics of the queue.

In Section 4.1, we introduce Brownian motion and illustrate how an arrival process such as a Poisson process can be approximated by a Brownian motion model. In Section 4.2, we illustrate the diffusion approximation for $M/M/n+GI$ queues. Because the service time distribution is exponential, we are able to spell out the details of every step 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 Section 4.3, 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 & Dai (2011). Diffusion approximations are rooted in the limit theorems for many-server queues in heavy traffic. These theorems require that the number of servers go to infinity. Section 4.4 shows that the diffusion approximation is accurate, sometimes even for queues with as few as 20 servers. The patience time distribution is built into the above diffusion models only through its density at the origin. When the patience time density is zero at the origin or changes rapidly near the origin, we present in Section 4.5 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.

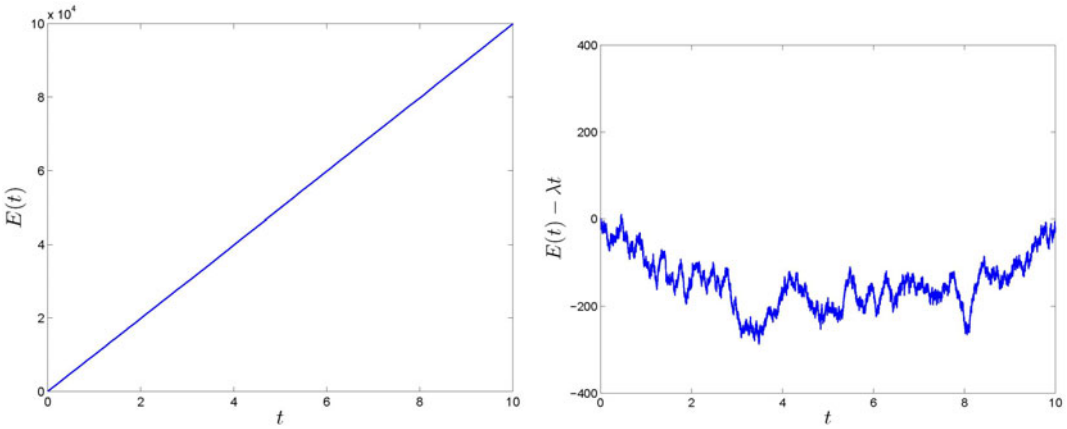
4.1 Brownian Approximation

Let $E=\{E(t):t\geq 0\}$ be a Poisson process with rate $\lambda=100$ arrivals per minute. In Figure 3a, 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 its expectation λt . To focus on the stochastic variability, we plot the sample path of the centered process $\{E(t)-\lambda t:t\geq 0\}$ in Figure 3b. The centered process records the fluctuation of the Poisson 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 further examine the effect of the scaling, in Figure 4 we plot the centered process when $\lambda=10,000$. It turns out that the magnitude of the centered process is on the order of $\sqrt{\lambda}$ as λ becomes large.



(a) A sample path (b) The sample path of the centered process

Figure 3 Poisson process with rate $\lambda = 100$



(a) A sample path (b) The sample path of the centered process

Figure 4 Poisson process with rate $\lambda = 10000$

Let $\gamma \in \mathbb{R}$ and $\sigma^2 > 0$ be given. A stochastic process $B = \{B(t) : t \geq 0\}$ is called a (γ, σ^2) -*Brownian motion* if (i) $B(0) = 0$ and almost every sample path is continuous, (ii) B has stationary, independent increments, and (iii) $B(t)$ is normally distributed with mean γt and variance $\sigma^2 t$ for every $t > 0$. The parameters γ and σ^2 are called the drift and the variance, respectively, of the Brownian motion. The process B is called a standard Brownian motion if $\gamma = 0$ and $\sigma^2 = 1$. By the well-known Donsker's theorem, (see, e.g., Billingsley (1999)), $\tilde{E}_\lambda = \{\tilde{E}_\lambda(t) : t \geq 0\}$ converges in distribution to a standard Brownian motion as $\lambda \to \infty$, where the scaled, centered process $\tilde{E}_\lambda$ is defined by

$$\tilde{E}_\lambda(t) = \frac{E(t) - \lambda t}{\sqrt{\lambda}}. \tag{21}$$

For a Poisson process, Donsker's theorem suggests that one may replace its scaled

fluctuation in (21) by the standard Brownian motion when λ is large. Donsker's theorem is a *functional central limit theorem*. Donsker's theorem holds for much more general processes including renewal processes.

For a general renewal process E associated with a sequence of iid random variables that has mean $1/\lambda$ and squared coefficient of variation c_a^2, its scaled fluctuation process $\tilde{E}_\lambda$ in (21) converges to a Brownian motion with drift $\gamma = 0$ and variance $\sigma^2 = c_a^2$ as $\lambda \to \infty$. The central idea of *diffusion approximations* is to replace a scaled fluctuation process such as the one in (21) by an appropriate Brownian motion.

For a many-server queue with a renewal arrival process (or an arrival process that satisfies the conditions of a functional central limit theorem) and a certain service time distribution, we may use Brownian motions to approximate the random variations in arrival and service. A diffusion model is obtained by replacing certain scaled processes in system equations by Brownian motions.

4.2 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 a patience time distribution satisfying (18). We use $X(t)$ to denote the number of customers in the system at time t, including those in service and those waiting. Let

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

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

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

is a moderate number, we can use a diffusion process Y to approximate $\tilde{X}$. The diffusion process may be described as follows. Let $\mathbb{D}$ be the space of functions $u : \mathbb{R}_+ \to \mathbb{R}$ that are right continuous on $[0,\infty)$ and has left limits on $(0,\infty)$. For each $u \in \mathbb{D}$, one can find a unique function $y \in \mathbb{D}$ 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 in (18), $x^+ = \max\{x, 0\}$, and $x^- = \max\{-x, 0\}$ for $x \in \mathbb{R}$. Thus, $u \mapsto y$ defines a map Ψ from an arbitrary function $u \in \mathbb{D}$ to another function $y \in \mathbb{D}$. Let

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

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

$$\sigma^2 = \mu\big(\rho + \rho \wedge 1\big). \tag{22}$$

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

$$Y(t) = \tilde{X}(0) - \beta\mu t + B(t) + \mu\int_0^t Y(s)^-\,\mathrm{d}s - \alpha\int_0^t Y(s)^+\,\mathrm{d}s. \tag{23}$$

The stochastic differential equation (23) 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

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

Suppose that $\alpha>0$. At any time t, the drift is negative if $Y(t)>-\beta\mu/\alpha$ and is positive if $Y(t)<-\beta$. When $Y(t)$ is either well above or well below zero, this drift will "pull it back" to an equilibrium level. The process tends to evolve around its long-term mean over time. An Ornstein–Uhlenbeck (OU) process that has a linear drift has the similar mean-reverting property. Because of its piecewise linear drift, Y is called a *piecewise Ornstein–Uhlenbeck (OU) process*. The piecewise OU process is analytically tractable. It admits a piecewise normal stationary distribution, whose density is

$$g(x)=\begin{cases}a_1\exp\left(-\dfrac{\alpha(x+\alpha^{-1}\mu\beta)^2}{\sigma^2}\right) & \text{when } x\ge 0,\\ a_2\exp\left(-\dfrac{\mu(x+\beta)^2}{\sigma^2}\right) & \text{when } x<0,\end{cases} \tag{24}$$

where a_1 and a_2 are normalizing constants that make $g(x)$ continuous at zero; see Browne & Whitt (1995). One may derive formula (15) for the delay probability in an $M/M/n+M$ queue by using (24) as well as the approximation

$$P_{\rm w}\approx\int_0^\infty g(x)\mathrm{d}x.$$

Because of the performance insensitivity to patience time distributions, formula (15) applies to the $M/M/n+GI$ model so long as α is taken to be the patience time density at the origin. Recall that $Q(t)$ is the number of customers waiting in the buffer at time t. 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 mean queue length $\bar{Q}$ and the fraction of customer abandonment $P_{\rm a}$ using the diffusion model. For that, let $Y(\infty)$ be a random variable that has the stationary distribution of Y. Using the stationary density in (24), the mean queue length $\bar{Q}$ can be computed by

$$\bar{Q}\approx\sqrt{n}\,\mathbb{E}[Y(\infty)^+]=\sqrt{n}\int_0^\infty xg(x)\mathrm{d}x \tag{25}$$

and the mean number of idle servers $\bar{I}$ can be computed by

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

Since $n-\bar{I}$ is the mean number of busy servers, the abandonment fraction $P_{\rm a}$ can be computed via

$$P_{\rm a}=1-\frac{\mu(n-\bar{I})}{\lambda}. \tag{26}$$

We show the performance estimates computed by (25) and (26) from the diffusion model in Table 2 under the "Diffusion" columns. The diffusion estimates agree well with the exact results.

In the rest of this section, we give a detailed derivation of the diffusion model (23). Let $E(t)$ be the number of customer arrivals by time t, and let $S=\{S(t):t\ge 0\}$ be a Poisson process with rate 1. We assume that $X(0)$, $E=\{E(t):t\ge 0\}$, 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. Since 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

$$X(t) = X(0) + E(t) - S(\mu T(t)) - A(t). \quad (27)$$

To derive Brownian approximations, we define several scaled processes by

$$\tilde{E}(t) = \frac{1}{\sqrt{n}}(E(t) - \lambda t),$$

$$\tilde{S}(t) = \frac{1}{\sqrt{n}}(S(nt) - nt),$$

$$\tilde{Q}(t) = \frac{1}{\sqrt{n}}Q(t), \quad \tilde{Z}(t) = \frac{1}{\sqrt{n}}(Z(t) - n),$$

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

Correspondingly, the dynamical equation (27) has a scaled version

$$\tilde{X}(t) = \tilde{X}(0) - \beta\mu t + \tilde{E}(t) - \tilde{S}(n^{-1}\mu T(t)) - \mu\int_0^t \tilde{Z}(s)\mathrm{d}s - \tilde{A}(t), \quad (28)$$

with β given in (8).

In the diffusion model, we replace the scaled primitive processes in (28) 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. (2010) for related convergence results.

Since 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 $\mu\rho t$. We use a Brownian motion $B_E = \{B_E(t) : t \geq 0\}$ with variance $\mu\rho$ to replace $\tilde{E}$ in (28). Because S is a Poisson process with rate 1, 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. Since $T(t)$ 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.,

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

Because $-\tilde{Z}(t)$ is the scaled number of idle servers and $\tilde{Q}(t)$ is the scaled queue length, we have

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

Because of (20), we may approximate the scaled abandonment process by

$$\tilde{A}(t) \approx \alpha\int_0^t \tilde{X}(s)^+ \mathrm{d}s. \quad (29)$$

It follows from (28) that

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

Let $B(t) = B_E(t) - B_S(\mu(\rho \wedge 1)t)$. Then B is a driftless Brownian motion with variance $\mu(\rho + \rho \wedge 1)$, the same one as in (22). Thus, $\tilde{X}$ is approximately a solution to the stochastic differential equation (23). In the proposed diffusion approximation, we use the solution Y to the stochastic differential equation (23) to replace $\tilde{X}$.

4.3 Diffusion Model for $M/H_2/n+GI$ Queues

Via a similar Brownian replacement procedure as in Section 4.2, a diffusion model has been derived by He & Dai (2011) for $GI/Ph/n+GI$ queues in the QED regime. A two-phase hyperexponential distribution (H_2), which has been discussed in Section 3, is a special case of phase-type distributions. In this section, we restrict our discussion to H_2 service time distributions, and illustrate the diffusion approximation proposed by He & Dai

(2011).

When the 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. 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

$$\mu=\frac{1}{p_1/\nu_1+p_2/\nu_2}. \tag{30}$$

In the 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

$$\theta_1=\frac{p_1/\nu_1}{p_1/\nu_1+p_2/\nu_2} \quad \text{and} \quad \theta_2=\frac{p_2/\nu_2}{p_1/\nu_1+p_2/\nu_2}. \tag{31}$$

Let $X_1(t)$ and $X_2(t)$ be the respective numbers of customers of these two types at time t. Since the customers in service are distributed following distribution θ, we define its scaled version after centering 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

$$\begin{aligned} Y_j(t)=&Y_j(0)-\beta\mu p_j t+p_j B_E(t)\\ &+(-1)^{j-1}B_M(\rho\mu t)-B_j((\rho\wedge 1)\theta_j\nu_j t)\\ &-\nu_j\int_0^t(Y_j(t)-p_j(Y_1(t)+Y_2(t))^+)\mathrm{d}s\\ &-p_j\alpha\int_0^t(Y_1(s)+Y_2(s))^+\mathrm{d}s \end{aligned} \tag{32}$$

for $j=1,2$. In (32), B_E is the same Brownian motion as in Section 4.2, 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 & Gao (2011) that Y has a unique stationary distribution. The algorithm proposed by He & Dai (2011) can be used to compute the stationary distribution numerically. Section 4.4 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 i th customer's type. The random vector takes $(1,0)$ with probability p_1 and takes $(0,1)$ with probability p_2. We assume that $C(1),C(2),\ldots$ are iid. Then,

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

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

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

$$T_j(t)=\int_0^t Z_j(s)\mathrm{d}s \tag{33}$$

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

$$X_j(t) = X_j(0) + M_j(E(t)) - S_j(\nu_j T_j(t)) - L_j(t). \quad (34)$$

We define the scaled processes by

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

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

$$\tilde{L}_j(t) = \frac{1}{\sqrt{n}}L_j(t),$$

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

Then using (30), (31), (33), and (34), we have the following scaled system equation

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

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 Section 4.2. 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 distribution p. Hence,

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

where $\tilde{A}(t)$ is the scaled number of abandoned customers by time t as defined in Section 4.2. Recall that $Q(t)$ is the queue length at time t. Then,

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

Since $Q(t) = (X_1(t) + X_2(t) - n)^+$, 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 (32) 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.

4.4 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 & Dai (2011), 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

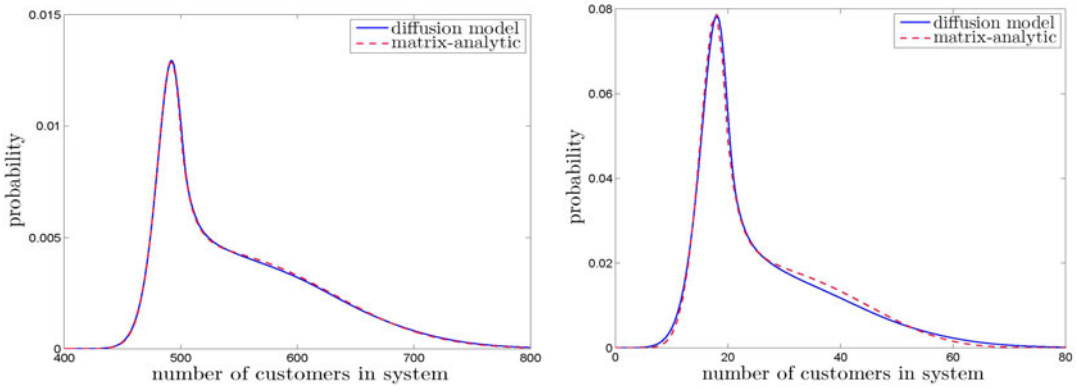
$$p=(0.9351,0.0649) \quad \text{and} \quad \nu=(9.354,0.072).$$

So the mean service time of the second-type customers is more than 100 times longer than that of the first type. Although over 93% of customers are of the first type, the fraction of its workload is merely 10%. Such a distribution has a large squared coefficient of variation $c_s^2=24$. One can check that the mean service time is 1 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-death process. One can use the matrix-analytic method to solve the stationary distribution of X. See Neuts (1981) for details on the matrix-analytic method. To evaluate the accuracy of the diffusion model, in Figure 5a 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 in the previous scenario, and the arrival rate be $\lambda=22.24$. Thus, $\rho=1.112$. As illustrated by Figure 5b, the diffusion model can still capture the exact stationary distribution for a queue with as few as20 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 & Dai (2011), 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



(a) $\rho=1.045$ and $n=500$ (b) $\rho=1.112$ and $n=20$

Figure 5 Stationary distribution of the customer number in the $M/H_2/n+M$ queue

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 around 2 hours to finish the matrix-analytic computation and the peak memory usage was nearly 5GB. Using the diffusion model and the proposed algorithm, it took less than 1 minute and the peak memory usage was less than 200MB on the same computer.

4.5 A Refined Diffusion Model Using the Hazard Rate of Patience Times

In the above diffusion model, the patience time density 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 the data collected from the service system. Estimating the density at the origin is statistically unreliable. The 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 & Oakes (1984)), can be used to estimate this parameter, one still has to record each 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 short 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 the density at the origin solely 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 (23) or (32), becomes zero. In this case, the diffusion model approximates a queue as if there is 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 slightly overloaded (i.e., $\rho > 1$), it still has a stationary distribution thanks to the customer abandonment that reduces service demands. 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.

Now we present a refined diffusion model using the entire patience time distribution. This model was proposed by He & Dai (2011). It exploits the idea of scaling the patience time hazard rate function, which was first proposed by Reed & Ward (2008) for single-server queues and was extended to many-server queues by Reed & Tezcan (2009). This refined diffusion model provides a more accurate approximation for many-server queues.

In this model, we assume that F, the cumulative distribution function of the patience times, satisfies $F(0) = 0$ and has a bounded hazard rate function h_F, given by

$$h_F(t) = \frac{f_F(t)}{1 - F(t)}, \quad t \geq 0, \tag{35}$$

where f_F is the density of F. With the hazard rate function, F can be written by

$$F(t) = 1 - \exp\left(-\int_0^t h_F(s)\mathrm{d}s\right), \quad t \geq 0.$$

In the refined diffusion model, the scaled abandonment process $\tilde{A}$ is approximated by

$$\tilde{A}(t) \approx \int_0^t \int_0^{\tilde{X}(s)^+} h_F\left(\frac{\sqrt{n}v}{\lambda}\right)\mathrm{d}v\mathrm{d}s, \quad t \geq 0. \quad (36)$$

The entire patience time distribution is built into the approximation through its hazard rate function. The intuition of the hazard rate scaling approximation was explained by Reed & Ward (2008): Consider the $Q(s)$ waiting customers in the buffer at time s. In general, only a small fraction of customers can abandon the system when the queue is working in the QED regime. Then by time s, the i th customer from the back of the queue has been waiting around i/λ time units. Approximately, this customer will abandon the queue during the next δ time units with probability $h_F(i/\lambda)\delta$. It follows that for the system, the instantaneous abandonment rate at time s is around $\sum_{i=1}^{Q(s)} h_F(i/\lambda)$. Hence, the scaled abandonment rate can be approximated by

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{Q(s)} h_F\left(\frac{i}{\lambda}\right) \approx \int_0^{\tilde{Q}(s)} h_F\left(\frac{\sqrt{n}v}{\lambda}\right)\mathrm{d}v = \int_0^{\tilde{X}(s)^+} h_F\left(\frac{\sqrt{n}v}{\lambda}\right)\mathrm{d}v, \quad (37)$$

from which (36) follows. Note that the arrival rate λ is on the order of $O(n)$ and $Q(s)$ is on the order of $O(n^{1/2})$. The patience time distribution in a small neighborhood of zero, not just its density at zero, is considered in the instantaneous abandonment rate in (37). Hence, the hazard rate scaling approximation in (36) is more accurate than that in (29).

Let k be a nonnegative integer. Suppose that the hazard rate function h_F is k times continuously differentiable in a neighborhood of zero. By Taylor's theorem,

$$h_F(x) \approx h_F(0) + \sum_{\ell=1}^{k} h_F^{(\ell)}(0)\frac{x^\ell}{\ell!}$$

for $x > 0$ small enough, where $h_F^{(\ell)}$ is the ℓ th order derivative of h_F. In this case, the approximation in (36) turns out to be

$$\tilde{A}(t) \approx h_F(0)\int_0^t \tilde{Q}(s)\mathrm{d}s + \sum_{\ell=1}^{k} \frac{n^{\ell/2} h_F^{(\ell)}(0)}{\lambda^\ell(\ell+1)!}\int_0^t \tilde{Q}(s)^{\ell+1}\mathrm{d}s.$$

Because $h_F(0)$ is identical to the patience time density at zero, the approximation in (29) can be regarded as the zeroth degree Taylor's approximation of (36). When the patience times are exponentially distributed, the hazard rate function is constant and the two approximations in (29) and (36) are identical.

Using the hazard rate scaling approximation and the Brownian motion replacements, we obtain another diffusion model for the $M/M/n+GI$ queue

$$Y(t) = \tilde{X}(0) - \beta\mu t + B(t) + \mu\int_0^t Y(s)^- \mathrm{d}s - \int_0^t \int_0^{Y(s)^+} h_F\left(\frac{\sqrt{n}v}{\lambda}\right)\mathrm{d}v\mathrm{d}s, \quad (38)$$

in which B is the driftless Brownian motion with variance given by (22). In (38), the diffusion process Y has the same diffusion coefficient as in (23), but its drift coefficient is

$$b(x) = \begin{cases} -\beta\mu - \int_0^x h_F\left(\frac{\sqrt{n}v}{\lambda}\right)\mathrm{d}v & \text{when } x \geq 0, \\ -\beta\mu - \mu x & \text{when } x < 0. \end{cases}$$

Comparing (23) and (38), one can see that the two models differ only in how the patience time distribution is incorporated. Because a more accurate approximation is used for the

abandonment process, the latter model can provide a better approximation for the queue.

Following the similar procedure as in Section 4.3, we can obtain the diffusion model for the $M/H_2/n+GI$ queue using the hazard rate scaling approximation, where the two-dimensional diffusion process (Y_1, Y_2) satisfies

$$\begin{aligned} Y_j(t) &= Y_j(0) - \beta\mu p_j t + p_j B_E(t) \\ &\quad + (-1)^{j-1} B_M(\rho\mu t) - B_j((\rho \wedge 1)\theta_j \nu_j t) \\ &\quad - \nu_j \int_0^t (Y_j(t) - p_j(Y_1(t) + Y_2(t))^+) \mathrm{d}s \\ &\quad - p_j \int_0^t \int_0^{(Y_1(s)+Y_2(s))^+} h_F\left(\frac{\sqrt{n}v}{\lambda}\right) \mathrm{d}v\mathrm{d}s \quad (39) \end{aligned}$$

for $j = 1, 2$.

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

$$p = (0.5915, 0.4085) \quad \text{and} \quad \nu = (5.917, 0.454). \quad (40)$$

The squared coefficient of variation of this distribution is $c_s^2 = 3$ and the resulting mean service time is still 1 minute. We assume that the patience times follow a two-phase hyperexponential distribution that has $p = (0.9, 0.1)$ and $\nu = (1, 200)$. Among the customers, 90% of them have exponentially distributed patience times with mean 1 minute, but the rest of customers are extremely impatient. Their patience times are exponentially distributed with mean 0.005 minute. These customers would abandon the system right away if no servers are available upon their arrival.

Although the customer-count process X of this queue is a quasi-birth-death process, the extremely high computational complexity prevents the matrix-analytic method from producing the stationary distribution when the number of servers n is moderate to large. See He & Dai (2011) for more discussion. 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. These simulation results are averaged over 20 independent runs and in each run, the system is simulated for 10^5 time units of operation. For each performance estimate, we list its 95% confidence interval (CI) generated from the 20 simulation runs. In the same table, we also list the performance estimates from the diffusion model (32) with $\alpha = 20.9$. In this example, using the patience time density at the origin solely cannot capture the behavior of the abandonment process. This 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 (39) that exploits the hazard rate scaling approximation. 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 mean $1/3$ minute. So the mean patience time is 1 minute. The density at the origin of the Erlang-3 distribution is zero. The diffusion model (32) 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 again. Then, $\rho = 1.141$ and 1.045. We list performance estimates from simulation (with 95% confidence intervals) and from the diffusion model using the hazard rate scaling in Table 5. As in the previous example, the refined diffusion model produces accurate performance approximations.

5. Fluid Model for Many-Server Queues in the ED Regime

In a many-server queue in the ED regime, the arrival rate exceeds the service capacity by a moderate fraction. As a result, almost all customers have to wait upon arrival and the queue length grows on the order of $O(n)$ as the number of servers $n \to \infty$. The fluid-scaled queue length, defined as the queue length divided by n, then converges to a non-zero deterministic limit under certain conditions. Such a limit is called a *fluid limit* and it can be used to build a fluid model for many-server queues. A fluid model, to be developed below, is appropriate for the analysis of a queue in the ED regime. In the QED regime, however, the queue

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

(a) $\rho = 1.141$ and $n = 50$

	Simulation (with 95% CI)	Diffusion in (32)	Refined diffusion
Mean queue length	4.845 ± 0.010	0.4709	4.869
Abandonment fraction (%)	14.99 ± 0.025	17.14	15.04
$\mathbb{P}[X(\infty) > 40]$	$0.9728 \pm 2.3\times10^{-4}$	0.9578	0.9749
$\mathbb{P}[X(\infty) > 50]$	$0.6111 \pm 7.4\times10^{-4}$	0.3158	0.6377
$\mathbb{P}[X(\infty) > 60]$	$0.1737 \pm 5.0\times10^{-4}$	1.044×10^{-7}	0.1749

(b) $\rho = 1.045$ and $n = 500$

	Simulation (with 95% CI)	Diffusion in (32)	Refined diffusion
Mean queue length	6.413 ± 0.015	1.475	6.359
Abandonment fraction (%)	5.512 ± 0.0088	5.863	5.517
$\mathbb{P}[X(\infty) > 480]$	$0.8881 \pm 6.3\times10^{-4}$	0.8663	0.8929
$\mathbb{P}[X(\infty) > 500]$	$0.4720 \pm 6.6\times10^{-4}$	0.3192	0.4822
$\mathbb{P}[X(\infty) > 520]$	$0.1050 \pm 3.7\times10^{-4}$	9.274×10^{-5}	0.1074

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

(a) $\rho = 1.141$ and $n = 50$

	Simulation (with 95% CI)	Refined diffusion
Mean queue length	19.31 ± 0.032	19.44
Abandonment fraction (%)	13.05 ± 0.031	13.03
$\mathbb{P}[X(\infty) > 45]$	$0.9645 \pm 4.1\times10^{-4}$	0.9704
$\mathbb{P}[X(\infty) > 50]$	$0.9066 \pm 7.4\times10^{-4}$	0.9169
$\mathbb{P}[X(\infty) > 70]$	0.4761 ± 0.0012	0.5037

(b) $\rho = 1.045$ and $n = 500$

	Simulation (with 95% CI)	Refined diffusion
Mean queue length	119.1 ± 0.22	119.5
Abandonment fraction (%)	4.337 ± 0.012	4.340
$\mathbb{P}[X(\infty) > 480]$	$0.9940 \pm 2.3\times10^{-4}$	0.9946
$\mathbb{P}[X(\infty) > 500]$	$0.9756 \pm 6.3\times10^{-4}$	0.9770
$\mathbb{P}[X(\infty) > 600]$	0.6645 ± 0.0018	0.6733

length is on the order of $O(n^{1/2})$. The fluid limit of the queue length is thus zero and the fluid model gives little insight to the dynamics of the queue. In this case, we should focus on a different scaling. Under certain conditions, the diffusion-scaled queue length, i.e., the queue length divided by $n^{1/2}$, converges to a diffusion process. Therefore, a diffusion model, as developed in Section 4, is more appropriate for the analysis of a queue in the QED regime.

In Whitt (2006), a fluid model was proposed by Ward Whitt and was shown to be useful in estimating the performance of a many-server queue in the ED regime. The system of interest is a $G/GI/n+GI$ queue, which has a general customer arrival process, iid service times, and iid patience times.

5.1 Whitt's Fluid Model for $G/GI/n+GI$ Queues

The fluid model is a deterministic approximation determined by a triple of parameters (ρ, H, F). Here, ρ is the traffic intensity, H is the cumulative distribution function of the service times, and F is the cumulative distribution function of the patience times. Recall that in the queue, $Z(t)$ and $Q(t)$ are the respective numbers of busy servers and waiting customers at time t. The fluid model is used to approximate the dynamics of the scaled processes Z/n and Q/n. Under this scaling, individual customers in the queue are approximated by "quanta" of fluid. The length of time that a quantity of fluid stays in the system is determined by the present fluid level and distributions H and F. If a quantity of

fluid enters service at time zero, then by time $t>0$, a proportion $H(t)$ of it will have finished service and left the system, while the remaining proportion $1-H(t)$ will be in service at time t. For a quantity of fluid that enters the buffer at time zero, if it has not entered service by time t, then a proportion $F(t)$ of it will have abandoned the system and the remaining proportion $1-F(t)$ will be waiting in the buffer at time t. Suppose that the fluid model has amount $z(t)$ of fluid in service at time t. Then in the corresponding queue, there are around $nz(t)$ customers in service. Similarly, if the fluid model has amount $q(t)$ of fluid in the buffer, then there are around $nq(t)$ customers waiting for service in the queue.

The service capacity of the fluid model is normalized to be 1. Accordingly, the rate of fluid input to the model is scaled to be ρ. The state of the fluid model is described by two functions $z(t,x)$ and $q(t,x)$ for $t\geq 0$ and $x\geq 0$. The function $z(t,x)$ is the amount of fluid in service at time t that has been served for no more than x time units, and $q(t,x)$ is the amount of fluid in the buffer at time t that has been waiting for no more than x time units. Clearly, the total amounts of fluid in service and in the buffer at time t are $z(t)=z(t,\infty)$ and $q(t)=q(t,\infty)$, respectively. Assume that f_H is the density of H. Then, its hazard rate function is

$$h_H(x)=\frac{f_H(x)}{1-H(x)},\quad x\geq 0.$$

Let $\delta>0$ be a small number. In the queue, a customer who has been in service for x time units would finish his service during the next δ time unit with probability $h_H(x)\delta$. Correspondingly, for a quantity of fluid that has been in service for x time units in the fluid model, a proportion $h_H(x)\delta$ of it will finish service during the next δ time unit. Hence, $h_H(x)$ is the conditional service rate for the fluid that has been in service for x time units. Assume that $z(t,x)$ has a density $z'(t,x)$ with respect to x, i.e.,

$$z(t,x)=\int_0^x z'(t,v)\mathrm{d}v. \tag{41}$$

Then, the total service rate at time t is given by

$$r(t)=\int_0^\infty z'(t,x)h_H(x)\mathrm{d}x. \tag{42}$$

Similarly, in the queue, a customer who has been waiting in the buffer for x time units would abandon the system during the next δ time unit with probability $h_F(x)\delta$, where h_F, defined in (35), is the hazard rate function of the patience times. Hence in the fluid model, for the quantity of fluid that has been waiting for x time units, a proportion $h_F(x)\delta$ of it will abandon the system during the next δ time unit if the fluid will not enter service. If we assume

$$q(t,x)=\int_0^x q'(t,v)\mathrm{d}v \tag{43}$$

where $q'(t,x)$ is the density of $q(t,x)$ with respect to x, the total abandonment rate at time t is

$$a(t)=\int_0^\infty q'(t,x)h_F(x)\mathrm{d}x. \tag{44}$$

The fluid model evolves according to the following dynamical equations. For the fluid that is in service at time t, the first equation describes the proportion of it that will be remaining in service at time $t+v$, i.e.,

$$z'(t+v,x+v)=z'(t,x)\frac{1-H(x+v)}{1-H(x)} \tag{45}$$

for $t\geq 0$, $x\geq 0$, and $v\geq 0$. If the buffer is not empty at time t, the fluid moves into service at rate $r(t)$; if there is idle service capacity available, the fluid enters service at rate ρ; if

the buffer is empty but all service capacity is used, this rate becomes $r(t)\wedge\rho$. Therefore, we have

$$z'(t,0)=\begin{cases} r(t) & \text{if } q(t)>0,\\ \rho & \text{if } z(t)<1,\\ r(t)\wedge\rho & \text{if } z(t)=1 \text{ and } q(t)=0. \end{cases} \tag{46}$$

Consider the small amount of fluid at the front of the buffer at time t. Let $w(t)$ be the length of time that it has been waiting by time t. Because of the first-in-first-out (FIFO) discipline, at time t there is no fluid that has been waiting for more than $w(t)$ time units. Hence, we must have

$$q'(t,x)=0 \tag{47}$$

for all $x>w(t)$. For the fluid that is in the buffer at time t, if it has not begun service by $t+v$, the proportion remaining in the buffer at time $t+v$ must satisfy

$$q'(t+v,x+v)=q'(t,x)\frac{1-F(x+v)}{1-F(x)} \tag{48}$$

for $t\geq 0$, $x\geq 0$, and $v\geq 0$. If the buffer is not empty, the new fluid input to the buffer arrives at rate ρ; if there is idle service capacity, no fluid input remains in the buffer; if the buffer is empty but all service capacity is used, then by (46), the fluid input to the buffer increases at rate $\rho-r(t)\wedge\rho$. Hence,

$$q'(t,0)=\begin{cases} \rho & \text{if } q(t)>0,\\ 0 & \text{if } z(t)<1,\\ \rho-r(t)\wedge\rho & \text{if } z(t)=1 \text{ and } q(t)=0. \end{cases} \tag{49}$$

The fluid model given by (41)–(49) depends on both the service time distribution and the patience time distribution through their entire distributions. It is in sharp contrast to most single-server fluid models in which the distributions appear only through their first moments.

5.2 The Fluid Model and the Fluid Limit

The fluid model is rooted in the stochastic-process limits for many-server queues. In the same paper, Whitt conjectured that under certain conditions, the two-parameter functions z and q, given by (41)–(49), are two limit processes for $G/GI/n+GI$ queues. To set up these limits, a sequence of $G/GI/n+GI$ queues, indexed by the number of servers n, is considered. These queues are assumed to be working in the many-server heavy-traffic regime, i.e., $\lambda_n\to\infty$ as $n\to\infty$ where λ_n is the arrival rate of the nth system. The service and the patience time distributions are assumed to be invariant with n. For simplicity, the mean service time is set to be 1 unit of time and all these queues have the same traffic intensity ρ. In the nth queue, let $Z_n(t,x)$ be the number of customers in service at time t that has been served for no more than x time units, and let $Q_n(t,x)$ be the number of customers in the buffer at time t that has been waiting for no more than x time units. Then, their respective fluid-scaled versions are defined by

$$\bar{Z}_n(t,x)=\frac{Z_n(t,x)}{n} \quad \text{and} \quad \bar{Q}_n(t,x)=\frac{Q_n(t,x)}{n}.$$

Whitt conjectured that as the number of servers n goes to infinity, the pair of processes $\{(\bar{Z}_n(t,x),\bar{Q}_n(t,x)):t\geq 0,x\geq 0\}$ weakly converges to the pair of deterministic functions $\{(z(t,x),q(t,x)):t\geq 0,x\geq 0\}$ in a functional space. Therefore, when n is large, the two-parameter functions z and q could serve as an approximation for the queue's dynamics.

Although the two-parameter fluid limits

remain as a conjecture in Whitt (2006), the fluid limits in terms of measure-valued processes were proved for $G/GI/n+GI$ queues by Kang & Ramanan (2010). In this work, the queue's dynamics are described by a pair of measure-valued processes, one keeping track of the customer waiting times in the buffer and the other keeping track of the amount of service each customer has received. Kang and Ramanan proved that in the many-server heavy-traffic regime and under certain assumptions, this pair of measure-valued processes weakly converges to the unique solution of a coupled pair of deterministic integral equations.

5.3 The Fluid Model in Steady State

Whitt (2006) also proved that the fluid model that satisfies (41)–(49) has a unique steady state. Since the performance of the fluid model does not change with time in the steady state, we delete the argument t and use z, q, z', q', r, a for the corresponding quantities in the steady state. Whitt discussed the steady state in two different cases: When $\rho \le 1$, i.e., the $G/GI/n+GI$ queue is operated in the underloaded or the QED regime, the fluid model has

$$r = z = \rho, \quad a = q = 0, \quad z'(x) = \rho(1-H(x)), \quad x \ge 0.$$

When $\rho > 1$, i.e., the queue is operated in the ED regime, the fluid model has

$$r = z = 1, \quad a = \rho - 1, \quad z'(x) = 1 - H(x), \quad x \ge 0;$$

in addition, by (48) and (49),

$$q'(x) = \begin{cases} \rho(1-F(x)) & \text{for } 0 \le x \le w \\ 0 & \text{for } x > w \end{cases},$$

where w is the solution of the equation

$$F(w) = \frac{\rho - 1}{\rho}, \tag{50}$$

and the total fluid in the buffer is

$$q = \int_0^w q'(x)\mathrm{d}x = \rho \int_0^w (1 - F(x))\mathrm{d}x.$$

The above steady-state quantities can be explained as follows. As we discussed earlier, for a queue with n servers that are operated in the QED regime, both the queue length and the number of abandonments are on the order of $O(n^{1/2})$. When n is large, one can expect that for any $t \ge 0$, both $Q(t)/n$ and $A(t)/n$ are very small. So in the fluid model, both the queue length q and the abandonment rate a must be zero when $\rho \le 1$. Because the abandonment rate is zero, the service rate must be equal to the arrival rate, i.e., $r = \rho$. It follows from (45) and (46) that $z'(0) = \rho$ and $z'(x) = \rho(1 - H(x))$. One can check that $z = \rho$ using (42). In the ED regime, since the queue is overloaded, we must have the service rate $r = 1$ and the abandonment rate $a = \rho - 1$. By similar arguments, one can deduce that $z'(x) = 1 - H(x)$ and $z = 1$. Using (49), we have $q'(0) = \rho$. Then it follows from (44) and (48) that

$$\begin{aligned} a &= \int_0^w q'(x) \frac{f_F(x)}{1 - F(x)} \mathrm{d}x \\ &= \int_0^w q'(0) f_F(x) \mathrm{d}x = \rho F(w), \end{aligned}$$

from which (50) follows.

Although the entire service time distribution is built in the diffusion model, the steady-state quantities, such as the abandonment fraction a and the mean waiting time w, does not depend on the service time distribution beyond its mean. However, these performance measures still depend on the entire patience time distribution. The current fluid model clearly demonstrates the efficiency of many-server queues when $\rho > 1$: As the fraction of abandonment compensates for

the excess arrival rate, the overloaded queue can still reach a steady state. In the steady state, all servers are working at 100% utilization while most customers need only wait for around w time units before receiving service. Hence, for a service system working in the ED regime, it is possible to meet a certain target service level (e.g., the average waiting time is less than 1 minute while the abandonment fraction is less than 20%) without sacrificing any utilization.

5.4 Performance Estimation Using the Fluid Model

Three queues are considered to evaluate the fluid model. All of them have $n=100$ servers and Poisson arrival processes with arrival rates $\lambda=120$, while both the mean service times and the mean patience times are 1 minute. Hence, the traffic intensities are all $\rho=1.2$. The first system is an $M/H_2/n+M$ queue, whose hyperexponential service time distribution is specified by (40) with variance 3 and whose patience time distribution is exponential. Several performance estimates by simulation and by the fluid model are listed in Table 6, where we can find a good agreement.

In the second queue, we change the patience time distribution to an Erlang-3 distribution with mean 1 minute. Table 7 compares the simulation results and the fluid model estimates. The third queue is an $M/LN/n+E_3$ queue, where LN stands for a log-normal service time distribution. We assume that the service time distribution has mean 1 and variance 8. The results for this queue can be found in Table 8. Again, the fluid model provides accurate performance estimates for those two queues.

Comparing Tables 6 and 7, one can see that the performance of a queue in the ED regime depends strongly upon the patience time distribution. Tables 7 and 8, however, indicate that these performance measures are not sensitive to the service time distribution as long as the mean service time is fixed.

6. Related Literature

We present a brief review of relevant research work at the end of this article. We would like to refer the readers who are interested in related topics to the original papers for full details.

The study of many-server queues has been mostly motivated by call center operations. Call centers have become a fertile ground for academic research due to the ever-growing size, complexity, and importance of the call center industry. A comprehensive tutorial and review for call center studies can be found in Gans et al. (2003). The paper covers both traditional operational models, such as multiple-server queues for performance analysis and control, and emerging multi-disciplinary topics, such as human resources problems, customer and agent behavior, and statistical analysis. Another important survey paper is Aksin et al. (2007). It is a valuable supplement to Gans et al. (2003). Brown et al. (2005) presented an extensive empirical study of historical operational data from an Israeli bank's call center. They performed a comprehensive statistical analysis and concluded that the arrival process of the call center follows an inhomogeneous Poisson process, the service times follow a log-normal distribution, and the patience time distribution appears to be non-exponential.

Table 6 Performance measures of the $M/H_2/100+M$ queue with $\rho=1.2$

	Simulation (with 95% CI)	Fluid model
Abandonment fraction (%)	16.95 ± 0.023	16.67
Mean waiting time (in minutes)	$0.1781\pm2.9\times10^{-4}$	0.1823
Mean queue length	20.33 ± 0.031	20.00
Server utilization (%)	99.660 ± 0.0027	100.00

Table 7 Performance measures of the $M/H_2/100+E_3$ queue with $\rho=1.2$

	Simulation (with 95% CI)	Fluid model
Abandonment fraction (%)	16.69 ± 0.026	16.67
Mean waiting time (in minutes)	$0.4324\pm4.1\times10^{-4}$	0.4669
Mean queue length	50.50 ± 0.043	53.15
Server utilization (%)	$99.973\pm7.4\times10^{-4}$	100.00

Table 8 Performance measures of the $M/LN/100+E_3$ queue with $\rho=1.2$

	Simulation (with 95% CI)	Fluid model
Abandonment fraction (%)	16.71 ± 0.030	16.67
Mean waiting time (in minutes)	$0.4328\pm4.7\times10^{-4}$	0.4669
Mean queue length	50.54 ± 0.049	53.15
Server utilization (%)	$99.968\pm9.1\times10^{-4}$	100.00

The mathematical study of customer abandonment in call centers can be traced back to the work of Palm (1937, 1946), where the author studied the $M/M/n+M$ (Erlang-A) model for the first time. Performance measures of the Erlang-A model are summarized in Mandelbaum & Zeltyn (2007). There is a growing list of papers that study queueing models with customer abandonment: The phenomenon of customer abandonment is studied for single-server queues in Baccelli et al. (1984) and Stanford (1979) under the $GI/GI/1+GI$ setting, and for multi-server queues in Boxma & de Waal (1994). Both the exact and asymptotic formulas of performance measures for the $M/M/n+GI$ model are summarized in Mandelbaum & Zeltyn (2004) and Zeltyn & Mandelbaum (2005). The queueing model with multiple servers and customer abandonment are also studied in Brandt & Brandt (1999, 2002), where the arrival and service rates are allowed to change with the respective numbers of customers in the system and in service. The stochastic monotonicity properties of multi-server queues with abandonment are investigated in Bhattacharya & Ephremides (1991) and Dai & He (2010). An asymptotic relationship between the abandonment processes and the queue length processes for many-server queues is proved in

Dai & He (2010). With respect to call center management in the presence of customer abandonment, the staffing and call routing problems were explored by Bassamboo et al. (2005, 2006), where asymptotic analysis is conducted to obtain the optimal policies.

The origin of the square-root safety staffing rule can be traced back to Erlang's paper written in 1923, which is collected in Brockmeyer et al. (1948). In the $M/M/n/n$ setting that models a loss system (e.g., a telephone system), Erlang derived this rule by the 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 and extended by later researchers in their research work, including Grassmann (1986, 1988), Kolesar (1986), Newell (1973, 1982), and Whitt (1992). Among them, Whitt (1992) formally proposed and analyzed this rule.

The QED regime was first introduced by Halfin & Whitt (1981) as an asymptotic regime in heavy traffic. In this regime, a sequence of queues indexed by their numbers of servers are considered. The traffic intensity of a queue in the sequence approaches 1 as its number of servers goes to infinity. The arrival rates of the queues increase with their numbers of servers, whereas the service rates of all queues remain the same. In this paper, Halfin and Whitt pioneered the study on diffusion approximations for many-server queues by establishing a diffusion limit for $GI/M/n$ queues. Ever since then, diffusion approximations have been demonstrated to be powerful in estimating the performance of many-server queues in the QED regime. Puhalskii & Reiman (2000) proved a diffusion limit for $GI/Ph/n$ queues. Garnett et al. (2002) proved a diffusion limit for the $M/M/n+M$ model that allows for customer abandonment. Whitt (2005) generalized this result to the $G/M/n+M$ model, and he set up a limit process for the $G/H_2^*/n$ model that has two customer classes in the same paper. For $G/Ph/n+GI$ queues in the QED regime, Dai et al. (2010) established a diffusion limit. This limit process is a multidimensional piecewise OU process. A numerical algorithm was developed by He & Dai (2011) for computing the stationary distribution of the piecewise OU process. This paper also demonstrates via numerical examples that a diffusion model can be a good approximation for a many-server queue. A recent work by Reed & Tezcan (2009) establishes a diffusion limit for the $GI/M/n+GI$ model in a new framework that scales the patience time hazard rate functions. As pointed out by He & Dai (2011), a refined diffusion model can be built in this framework and may produce more accurate performance estimates. Using the diffusion limits, the staffing problem for $M/M/n+GI$ queues is considered in Mandelbaum & Zeltyn (2009). Koçağa & Ward (2010) studied the admission control problem for $M/M/n+M$ queues. The optimal dynamic scheduling for a queue with multiple customer classes and a single agent pool in the QED regime is investigated in Harrison & Zeevi (2004). The optimal routing policies in a service system with multiple customer classes and multiple agent pools are considered in Dai & Tezcan (2008), Gurvich & Whitt (2009), and Tezcan & Dai (2010). The limit processes for the more general $G/GI/n+GI$ model in the QED regime can

be found in Mandelbaum & Momčilović (2012) and Talreja & Whitt (2009). These limit processes, however, are not diffusion processes.

For a call center with customer abandonment, the ED regime was introduced by Garnett et al. (2002). Whitt (2004) explored the key properties of queues in the ED regime by establishing and investigating the fluid limit for $M/M/n+M$ queues. The "efficiency" of the ED regime was reinforced by Bassamboo & Randhawa (2010): They proved that for $M/M/n+GI$ queues with certain patience time distributions and certain operational costs, the optimized staffing level leads the queues to the overloaded regime. For many-server queues with abandonment and multiple customer classes, a routing policy that is asymptotically optimal in the ED regime was studied by Atar et al. (2010, 2011). A fluid model proposed by Whitt (2006) is able to capture the dynamics of a $G/GI/n+GI$ queue in the ED regime. This fluid model was extended by Liu & Whitt (2011) to the $G_t/GI/n_t+GI_t$ model in which arrival rates, staffing levels, and patience time distributions are all allowed to change with time. For a many-server queue with a general service time distribution, measure-valued processes have been used to give a Markovian description of the system. Kaspi & Ramanan (2011) obtained a measure-valued fluid limit for the $G/GI/n$ model. Kang & Ramanan (2010) obtained a measure-valued fluid limit for the $G/GI/n+GI$ model with customer abandonment, and Zhang (2009) obtained a similar measure-valued fluid limit independently. Their work partially justifies the fluid model in Whitt (2006).

7. Summary

In service systems such as call centers, the system performance is sensitive to customer abandonment. When a system has a significant amount of abandonment, it is crucial to model the customer abandonment explicitly. According to the structure of operational costs, the manager of a call center may choose to operate his system in the QED or ED regime. Following the square-root safety staffing rule, he can drive the system to the QED regime, achieving both a high level of service quality and a high level of server utilization. When a moderate fraction of customer abandonment is allowed, the manager can even staff the system in an overloaded regime, known as the ED regime, to achieve higher efficiency while still maintaining satisfactory system performance. When the system is operated in the QED regime, it is the behavior of the patience time distribution near the origin, not the mean patience time, that has the most impact on the system performance. The diffusion approximations are useful in evaluating the performance of a many-server queue in the QED regime, while the fluid model is useful for a queue in the ED regime. These approximate models can be practical tools, sometimes the only tool besides computer simulation, to evaluate the performance of a many-server queue in heavy traffic.

References

[1] Aksin, Z., Armony, M. & Mehrotra, V. (2007). The modern call center: a multidisciplinary perspective on operations management research. Production and Operations Management, 16: 665-688

[2] Atar, R., Giat, C. & Shimkin, N. (2010). The

$c\mu/\theta$ rule for many-server queues with abandonment. Operations Research, 58: 1427-1439

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

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

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

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

[7] Bassamboo, A. & Randhawa, R. S. (2010). On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. Operations Research, 58: 1398-1413

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

[9] Billingsley, P. (1999). Convergence of Probability Measures, 2nd ed. Wiley, New York

[10] Boxma, O.J. & de Waal, P.R. (1994). Multiserver queues with impatient customers. In: Labetouille, J., Roberts, J.W. (eds.), The fundamental role of teletraffic in the evolution of telecommunications networks (Proc. ITC-14), pp. 743-756. North-Holland, Amsterdam

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

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

[13] Brockmeyer, E., Halstrøm, H.L. & Jensen, A. (1948). The life and works of A.K. Erlang. Transactions of the Danish Academy of Technical Sciences, 1948: 1-277

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

[15] Browne, S. & Whitt, W. (1995). Piecewise-linear diffusion processes. In: Dshalalow J. (ed.), Advances in queueing, pp. 463-480. CRC Press, Boca Raton, FL

[16] Cox, D.R. & Oakes, D. (1984). Analysis of Survival Data. Monographs on Statistics and Applied Probability, Chapman & Hall, London

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

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

[19] Dai, J.G. & Tezcan, T. (2008). Optimal control of parallel server systems with many servers in heavy traffic. Queueing Systems, 59: 95-134

[20] Dieker, A.B. & Gao, X. (2011). Positive recurrence of piecewise Ornstein-Uhlenbeck processes and common quadratic Lyapunov functions. Tech. rep., School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA

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

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

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

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

[25] Gross, D. & Harris, C.M. (1985). Fundamentals of Queueing Theory. Wiley, New York

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

[27] Haln, S. & Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. Operations Research, 29: 567-588

[28] Harrison, J.M. & Zeevi, A. (2004). Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. Operations Research, 52: 243-257

[29] He, S. & Dai, J.G. (2011). Many-server queues with customer abandonment: numerical analysis of their diffusion models. Tech. rep., School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA

[30] Kang, W. & Ramanan, K. (2010). Fluid limits of many-server queues with reneging. Annals of Applied Probability, 20: 2204-2260

[31] Kaspi, H. & Ramanan, K. (2011). Law of large numbers limits for many-server queues. Annals of Applied Probability, 21: 33-114

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

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

[34] Liu, Y. and Whitt, W. (2011). The $G_t / GI / s_t + GI_t$ many-server fluid queue. Preprint

[35] Mandelbaum, A. & Momčilović, P. (2012). Queues with many servers and impatient customers. Mathematics of Operations Research, 37: 41-65

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

[37] Mandelbaum, A. & Zeltyn, S. (2007). Service engineering in action: the Palm/Erlang-A queue with applications to call centers. In: Spath, D., Fähnrich, K.-P. (eds.), Advances in services innovations, pp. 17-45.Springer, Berlin

[38] Mandelbaum, A. & Zeltyn, S. (2009). Staffing many-server queues with impatient

customers: constraint satisfaction in call centers. Operations Research, 57: 1189-1205

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

[40] Newell, G.F. (1973). Approximate Stochastic Behavior of *n*-Server Service Systems with Large *n*. Lecture Notes in Economics and Mathematical Systems, Vol. 87, Springer-Verlag, Berlin

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

[42] Palm, C. (1937). Etude des delais d'attente. Erisson Technics, 5: 37-56

[43] Palm, C. (1946). Special issue of teletrafikteknik. Tekniska Meddelanden från Kungliga Telegrafstyrelsen, 4

[44] Puhalskii, A.A. & Reiman, M.I. (2000). The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. Advances in Applied Probability, 32: 564-595. Correction, (2004), 36: 971

[45] Reed, J. & Tezcan, T. (2009). Hazard rate scaling for the $GI/M/n+GI$ queue. Tech. rep., Stern School of Business, New York University, New York

[46] Reed, J.E. & Ward, A.R. (2008). Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion: hazard rate scaling in heavy traffic. Mathematics of Operations Research, 33: 606-644

[47] Stanford, R.E. (1979). Reneging phenomena in single channel queues. Mathematics of Operations Research, 4: 162-178

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

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

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

[51] Whitt, W. (2004). Eciency-driven heavy-traffic approximations for many-server queues with abandonments. Management Science, 50: 1449-1461

[52] Whitt, W. (2005). Heavy-traffic limits for the $G/H_2^*/n/m$ queue. Mathematics of Operations Research, 30: 1-27

[53] Whitt, W. (2006). Fluid models for multiserver queues with abandonments. Operations Research, 54: 37-54

[54] Zeltyn, S. & Mandelbaum, A. (2005). Call centers with impatient customers: many-server asymptotics of the $M/M/n+G$ queue. Queueing Systems, 51: 361-402

[55] Zhang, J. (2009). Fluid models of multi-server queues with abandonment. Tech. rep., Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Hong Kong

J. G. Dai is the Edeneld Professor of Industrial & Systems Engineering at Georgia Institute of Technology. He is a Special Term Professor at Tsinghua University and a Visiting Professor in Decision Sciences at National University of

Singapore. Over the last twenty years, he has worked on stochastic models arising from communications, manufacturing, and service systems that include data switches, semiconductor wafer fabrication lines, call centers, and healthcare-delivery systems. Jim Dai received B.A. and M.S. degrees from Nanjing University and a Ph.D. degree from Stanford University. He is an elected fellow of Institute of Mathematical Statistics and an elected fellow of Institute for Operations Research and the Management Sciences (INFORMS). He has received a number of awards for his research contributions including The Best Publication Award in 1997 and The Erlang Prize in 1998, both from the Applied Probability Society of INFORMS. He is currently an Area Editor for *Mathematics of Operations Research*, an Area Editor for *Operations Research*, and a past Series Editor for *Handbooks in Operations Research and Management Science*.

Shuangchi He is an assistant professor in the Department of Industrial and Systems Engineering at the National University of Singapore. He received his PhD degree from Georgia Institute of Technology in 2011. His research interests include applied probability, stochastic modeling, and statistical signal processing.