Economies-of-Scale in Many-Server Queueing Systems: Tutorial and Partial Review of the QED Halfin–Whitt Heavy-Traffic Regime*

Johan S. H. van Leeuwaarden[†]
Britt W. J. Mathijsen[†]
Bert Zwart[‡]

Abstract. Multiserver queueing systems describe situations in which users require service from multiple parallel servers. Examples include check-in lines at airports, waiting rooms in hospitals, queues in contact centers, data buffers in wireless networks, and delayed service in cloud data centers. These are all situations with jobs (clients, patients, tasks) and servers (agents, beds, processors) that have large capacity levels, ranging from the order of tens (checkouts) to thousands (processors). This survey investigates how to design such systems to exploit resource pooling and economies-of-scale. In particular, we review the mathematics behind the quality- and efficiency-driven (QED) regime, which lets the system operate close to full utilization, while the number of servers grows simultaneously large and delays remain manageable. Aimed at a broad audience, we describe in detail the mathematical concepts for the basic Markovian many-server system, and we provide only sketches or references for more advanced settings related to, e.g., load balancing, overdispersion, parameter uncertainty, general service requirements, and queueing networks. While serving as a partial survey of a massive body of work, the tutorial is not meant to be exhaustive.

Contents

†Department of Mathematics and Computer Science, Eindhoven University of Technology, 5600 MB Eindhoven, Netherlands (j.s.h.v.leeuwaarden@tue.nl, b.w.j.mathijsen@tue.nl).

‡Department of Stochastics, CWI, 1090 GB Amsterdam, Netherlands (Bert.Zwart@cwi.nl).

1. Introduction. Multiserver systems describe situations in which users require service from multiple parallel servers. Classical examples of such systems include call centers [37, 126, 157, 47, 21, 27, 166, 17, 95], health care delivery [5, 56, 164, 57], and communication systems [2, 92, 151, 147]. In all settings, one can think of such systems as being composed of *jobs* and *servers*. In call centers, jobs are customers' requests for help from one of the agents (servers). In communication networks, the data packets are the jobs and the communication channels are the servers. The system scale may refer to the size of the client base it caters to, or the magnitude of its capacity, or both.

Next to the central notions of jobs and servers, most multiserver systems are subject to uncertainty and hence give rise to stochastic systems. Although arrival volumes over a certain planning horizon can be anticipated to some extent, for instance, through historical data and forecasting methods, it is challenging to predict with certainty future arrival patterns. Moreover, job sizes are typically random as well, adding more uncertainty. This intrinsic stochastic variability is a predominant cause of delay experienced by jobs in the system, which is why stochastic models have proved instrumental in both quantifying and improving the operational performance of multiserver systems. Queueing theory provides the mathematical tools to analyze such stochastic models, and to evaluate and improve system performance. Queueing theory can also serve to reveal capacity-sizing rules that prescribe how to scale multiserver systems, in terms of matching capacity with demand, to meet certain performance targets. Often a trade-off exists between high system utilization and short delays.
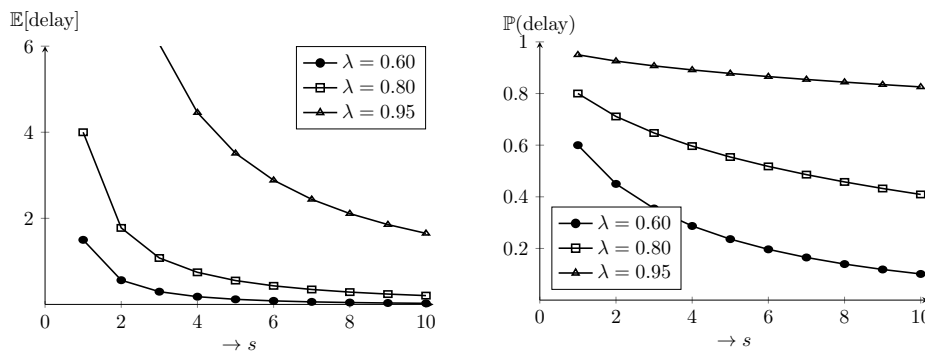
**Fig. I** *Effects of resource pooling in the M/M/s queue.*

Effects of Resource Pooling. Let us first demonstrate the effects of resource pooling for the most basic multiserver queueing model, the $M/M/s$ queue. This model assumes that jobs arrive according to a Poisson process, that their service times form an independent and identically distributed (i.i.d.) sequence of exponential random variables, and that jobs are processed in order of arrival by one of the s parallel servers. Delayed jobs are temporarily stored in an infinite-sized buffer. The three parameters that characterize this model are the arrival rate λ, the mean processing time $1/\mu$, and the number of servers s. We denote the number of jobs in the system at time t by $Q(t)$. The process $(Q(t))_{t \geq 0}$ is a continuous-time Markov chain with state space $\{0, 1, 2, \ldots\}$. The birth rate λ is constant, and the death rate is $\mu \cdot \min\{k, s\}$ when there are k jobs in the system. Observe now that we can change the time scale by considering the process $(Q(t\mu))_{t \geq 0}$, so that a busy server completes one job per unit of time. This allows us to consider the case $\mu = 1$ without loss of generality.

To illustrate the operational benefits of sharing resources, we compare a system of s separate $M/M/1$ queues, each serving a Poisson arrival stream with rate $\lambda < 1$, against one $M/M/s$ queue with arrival rate λs. The two systems thus face the same workload λ per server. We now fix the value of λ and vary s. Obviously, the delay and queue length distribution in the first scenario with parallel servers are unaffected by the parameter s, since there is no interaction between the single-server queues. This lack of coordination tolerates an event of having an idle server while the total number of jobs in the system exceeds s, therefore wasting resource capacity. Such an event cannot happen in the many-server scenario, due to the central queue. This central coordination improves the quality of service (QoS). Indeed, Figure 1 shows that the reduction in mean delay and delay probability can be substantial.

QED Regime. The quality- and efficiency-driven (QED) regime is a form of resource pooling that goes beyond the typical objective of improving performance by joining forces. For the $M/M/s$ queue, the QED regime is best explained in terms of the square-root rule

$$(1.1) \qquad s = \lambda + \beta\sqrt{\lambda}, \quad \beta > 0,$$

which prescribes how to size capacity as a function of the offered load. Notice that the number of servers s is taken equal to the sum of the mean load λ and an additional term $\beta\sqrt{\lambda}$ that is of the same order as the natural load fluctuations of the arrival process
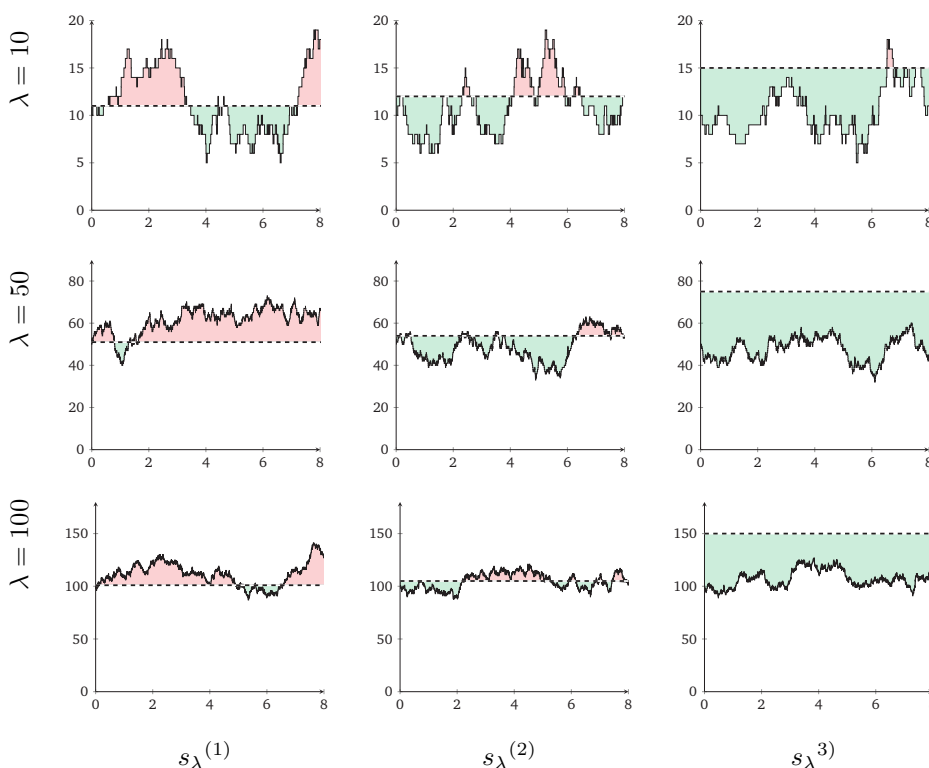
Fig. 2 *Sample paths of the $M/M/s$ queue with $\lambda = 10$, 50, and 100 and s set according to the three scaling rules in (1.2) with $\beta = 0.5$.*

(so of the order $\sqrt{\lambda}$). Observe that capacity increases with β, where we note that the free parameter β can take any positive value. The QED regime assumes the coupling between λ and s as in (1.1) and then lets both s and λ become large. This not only increases the scale of operation, but also lets the load per server $\rho = \lambda/s \sim 1 - \beta/\sqrt{\lambda}$ approach 1 as s (and λ) become(s) large. Now instead of diving immediately into the mathematical details, we shall first demonstrate the QED regime, or the capacity-sizing rule (1.1), by investigating typical sample paths of the queue length process $Q = (Q(t))_{t \geq 0}$ for increasing values of λ.

The upper middle panel of Figure 2 depicts a sample path for $\lambda = 10$ and s set according to (1.1) rounded to the nearest integer. The number of delayed jobs at time t is given by $(Q(t) - s)^+$ with $(\cdot)^+ := \max\{0, \cdot\}$. The number of idle servers is given by $(s - Q(t))^+$. In Figure 2, the upper and lower areas enclosed between the sample path and the horizontal dashed line s respectively represent the cumulative queue length and the cumulative number of idle servers over the given time period. Bearing in mind the dual goals of QoS and efficiency, we want to minimize both of these areas simultaneously.

We next show similar sample paths for increasing values of λ. Since $s > \lambda$ is required for stability, the value of s needs to be adjusted accordingly. We show three scaling rules,

$$(1.2) \qquad s_\lambda^{(1)} = [\lambda + \beta], \qquad s_\lambda^{(2)} = [\lambda + \beta\sqrt{\lambda}], \qquad s_\lambda^{(3)} = [\lambda + \beta\lambda],$$

with $\beta > 0$, where $[\cdot]$ denotes the rounding operator. Note that these three rules differ in terms of overcapacity $s_\lambda - \lambda$, and $s_\lambda^{(2)}$ is the (rounded) square-root rule introduced in (1.1). Figure 2 depicts typical sample paths of the queue length process for increasing values of λ for the three scaling rules with $\beta = 0.5$. Observe that for all scaling rules, the stochastic fluctuations of the queue length processes relative to s decrease with the system size. Moreover, the paths in Figure 2 appear to become smoother with increasing λ. Of course, the actual sample path always consists of upward and downward jumps of size 1, but we will show how proper centering and scaling of the queue length process indeed gives rise to a *diffusion process* in the limit as $\lambda \to \infty$ (section 2). Although the difference in performance of the three regimes is not yet evident for relatively small λ, clear distinctive behavior occurs for large λ.

ED and QD Regimes. With $s_\lambda^{(1)}$, most jobs are delayed and server idle time is low, since $\rho = (1 + \beta/\lambda)^{-1} \to 1$ as $\lambda \to \infty$. Systems scaled according to this rule value server efficiency over QoS, and therefore in the literature this regime is also known as the *efficiency-driven* (ED) regime [166]. In contrast, the third scaling rule $s_\lambda^{(3)}$ yields a constant utilization level $\rho = 1/(1 + \beta)$, which stays away from 1, even for large λ. Queues operating in this regime exhibit significant server idle times. Moreover, for the particular realization of the queueing processes for $\lambda = 50$ and $\lambda = 100$, none of the jobs is delayed. This is known as the *quality-driven* (QD) regime [166]. The scaling rule $s_\lambda^{(2)}$ is in some ways a combination of the other two regimes. First, we have $\rho = (1 + \beta/\sqrt{\lambda})^{-1} \to 1$ as $\lambda \to \infty$, which indicates efficient usage of resources as the system grows. The sample paths, however, indicate that only a fraction of all jobs is delayed, and only small queues arise, indications of good QoS. Figure 2 provides visual confirmation that the square-root rule $s_\lambda^{(2)}$, related to the QED regime, strikes the right balance between the two profound objectives of capacity allocation in multiserver systems: negligible delay and idling. We shall latter discuss the mathematical foundations of the QED regime and quantify the favorable properties revealed by Figure 2, including the nondegeneracy of the delay probability. To quote Halfin and Whitt [63], "The balance between service and economy usually dictates that the probability of delay be kept away from both zero and one, so that the number of jobs present fluctuates between the regions above and below the number of servers."

Central Limit Theorem. We will see that not only the $M/M/s$ queue but also a wide range of multiserver models will possess the same property (delay probability being strictly between zero and one). This is because the QED regime is intimately connected with the Central Limit Theorem (CLT). Let Φ denote the cumulative distribution function (cdf) of the standard normal distribution.

THEOREM 1.1 (Central Limit Theorem). *Let $X_1, X_2, \ldots$ be i.i.d. random variables of finite mean m and variance v. Then, for all $x \in \mathbb{R}$,*

$$(1.3) \qquad \lim_{n \to \infty} \mathbb{P}\Big(\sum_{i=1}^{n} X_i < mn + vx\sqrt{n} \Big) = \Phi(x).$$

Consider a $\mathrm{Pois}(\lambda)$ random variable, with λ integer valued, which is equal in distribution to the sum of λ independent Poisson random variables with unit mean and variance, i.e.,

$$\mathrm{Pois}(\lambda) \stackrel{d}{=} \sum_{i=1}^{\lambda} \mathrm{Pois}_i(1).$$

Direct application of the CLT hence implies that for $s = \lambda + \beta\sqrt{\lambda}$,

$$(1.4) \qquad \mathbb{P}(\text{Pois}(\lambda) \leq s) \to \Phi(\beta) \qquad \text{as } \lambda \to \infty.$$

Related Surveys. In this survey, we review the analysis of many-server systems operating in the QED regime, with special focus on various modeling assumptions that match well with the CLT. In recent years, several comprehensible surveys have appeared in the literature on topics related to queueing systems and their asymptotic analysis. We take the opportunity to mention a couple of them here. Some tutorial papers are devoted to specific applications. Telephone call centers are the main focus of survey papers by Gans, Koole, and Mandelbaum [47], Brown et al. [27], and Aksin, Armony, and Mehrotra [1]. Armony et al. [5] provide an extensive overview of queueing phenomena in health care environments. Focusing more on methodology, Pang, Talreja, and Whitt [127] discuss mathematical techniques to prove stochastic-process limits for queueing systems, and Ward [155] reviews queueing systems with abandonments in asymptotic regimes (including the QED regime). The survey paper by Dai and He [33] also concerns queueing systems with abandonments, particularly focusing on the ED and QED regimes. Whitt [162] provides an extensive bibliography of the literature on queueing models with time-varying demand, also covering the QED regime.

Organization. Section 2 introduces two classical queueing models that serve as a vehicle to convey the ideas behind the QED regime. We discuss in section 3 key properties that are common to these models under QED scaling and illustrate how these features stretch beyond these specific model settings. In section 4 we explain how asymptotic QED approximations of performance measures can be transformed into easy-to-use and robust capacity allocation principles. Furthermore, we illustrate how to adapt capacity allocation decisions to time-varying demand. Even though QED stochastic-process limits provide good first-order insight into the performance of large-scale systems, care needs to be taken with regard to the finiteness of the system. Therefore, we review in section 5 results that attempt to quantify the error made by asymptotic approximations, leading to both refinements and approximation bounds. We also consider the implication of approximation errors for capacity allocation decisions (so-called optimality gaps). Finally, in section 6 we review some model extensions that have received much attention due to practical applicability or theoretical challenges.

Notation. We conclude this section by introducing some notation that will be used throughout the paper. By $N(\mu, \sigma^2)$ we denote a normally distributed random variable with mean μ and variance σ^2. The probability density function (pdf) and cumulative distribution function (cdf) of the standard normal distribution are denoted by φ and Φ, respectively. The symbol $\overset{d}{=}$ means equal in distribution, and $\overset{d}{\Rightarrow}$ means convergence in distribution. The relation $u(\lambda) \sim v(\lambda)$ implies that $\lim_{\lambda \to \infty} u(\lambda)/v(\lambda) = 1$. By $u(\lambda) = O(v(\lambda))$ we mean that $\limsup_{\lambda \to \infty} u(\lambda)/v(\lambda) < \infty$, and $u(\lambda) = o(v(\lambda))$ implies that $\limsup_{\lambda \to \infty} u(\lambda)/v(\lambda) = 0$.

2. Example Models. This survey uses two running examples that are illustrative for both the model-specific and the universal features of the QED regime. The first example is the already introduced $M/M/s$ queue, a fully Markovian many-server system. The second example is the so-called bulk-service queue, a standard discrete-time model. Through these models, we shall describe in this section several easy ways of establishing QED limits that only require a standard application of the CLT.

2.1. Many Exponential Servers. Let us first consider an infinite-server system to which jobs arrive according to a Poisson process with rate λ. Each job requires an exponentially distributed service time with unit mean. The steady-state number of present jobs (or, equivalently, the steady-state number of busy servers) follows a Poisson distribution with mean λ. It is known that a Poisson distribution can be well approximated by a normal distribution for sufficiently large λ, so that it is approximately normally distributed with mean and variance λ. Therefore, the coefficient of variation (standard deviation divided by the mean) decreases as $1/\sqrt{\lambda}$, which makes the steady-state queue length become more concentrated around its mean with increasing λ.

If we now pretend, for a moment, that this infinite-server system serves as a good approximation for the $M/M/s$ queue, we could approximate the steady-state delay probability $\mathbb{P}(\text{delay})$ in the $M/M/s$ queue as

$$(2.1) \quad \mathbb{P}(\text{delay}) \approx \mathbb{P}(Q \geq s) = \mathbb{P}\left(\frac{Q-\lambda}{\sqrt{\lambda}} \geq \frac{s-\lambda}{\sqrt{\lambda}}\right) \approx 1 - \Phi\left(\frac{s-\lambda}{\sqrt{\lambda}}\right) = 1 - \Phi(\beta).$$

The use of this normal approximation in support of capacity allocation decisions was explored by Kolesar and Green [98]. Of course, the infinite-server system ignores the one thing that makes a queueing system unique: that a queue is formed when all servers are busy. During these periods of congestion, a system with a finite number of servers s will operate at a slower pace than its infinite-server counterpart, so the approximation in (2.1) is likely to underestimate $\mathbb{P}(\text{delay})$. Nevertheless, the infinite-server heuristic does suggest that, in large systems, the number of servers can be chosen close to the offered load as in (1.1).

We shall now make more precise statements about QED limits and use the intimate relation between the $M/M/s/s$ queue (Erlang loss model) and the $M/M/s$ queue (Erlang delay model). When $\rho = \lambda/s < 1$ the steady-state distribution of the $M/M/s$ queue exists and is given by

$$(2.2) \qquad \pi_k = \lim_{t\to\infty} \mathbb{P}(Q(t)=k) = \begin{cases} \pi_0 \frac{\lambda^k}{k!} & \text{if } k \leq s, \\ \pi_0 \frac{\lambda^s}{s!}\, \rho^{k-s} & \text{if } k > s, \end{cases}$$

where

$$\pi_0 = \left(\sum_{k=0}^{s} \frac{\lambda^k}{k!} + \frac{\rho}{1-\rho}\frac{\lambda^s}{s!}\right)^{-1}.$$

From Little's law and the PASTA (Poisson Arrivals See Time Averages) property [163], it follows that the delay probability, i.e., the probability that an arbitrary job needs to wait before taken into service, is given by the Erlang C formula

$$(2.3) \qquad C(s,\lambda) = \frac{\lambda^s}{s!}\left((1-\rho)\sum_{k=0}^{s-1}\frac{\lambda^k}{k!} + \frac{\lambda^s}{s!}\right)^{-1}.$$

The mean steady-state delay is given by

$$(2.4) \qquad \mathbb{E}[\text{delay}] = \frac{C(s,\lambda)}{(1-\rho)s}.$$

A closely related performance measure is the probability of blocking in the $M/M/s/s$ queue, also known as the Erlang loss formula, and is given by

$$(2.5) \qquad B(s,\lambda) = \frac{\frac{\lambda^s}{s!}}{\sum_{k=0}^{s} \frac{\lambda^k}{k!}} = \frac{\mathbb{P}(\mathrm{Pois}(\lambda)=s)}{\mathbb{P}(\mathrm{Pois}(\lambda)<s)},$$

where the latter probabilistic representation, with $\mathrm{Pois}(\lambda)$ denoting a Poisson random variable with mean λ, is convenient in light of the CLT. Note also that the Erlang B and C formulae are related by

$$(2.6) \qquad C(s,\lambda) = \left(\rho + \frac{1-\rho}{B(s,\lambda)}\right)^{-1}.$$

See [158] for an extensive overview of properties of the Erlang B and C formulae; see also [74, 82]. We now focus on how these formulae scale when λ and s both grow large.

Halfin and Whitt [63] showed that, just as with the tail probability in the infinite-server setting (2.1), the delay probability in the $M/M/s$ queue converges under scaling (1.1) to a value between 0 and 1. Moreover, they showed that this is in fact the only scaling regime in which such a nondegenerate limit exists and identified its value.

Let $\rho_\lambda := \lambda/s_\lambda$ denote the server utilization if capacity s_λ is scaled according to (1.1). The following result is obtained in [63].

PROPOSITION 2.1. *There is the nondegenerate limit*

$$(2.7) \qquad \lim_{\lambda \to \infty} C(s_\lambda, \lambda) = \left(1 + \frac{\beta \, \Phi(\beta)}{\varphi(\beta)}\right)^{-1} =: g(\beta) \in (0,1)$$

if and only if

$$(2.8) \qquad \lim_{\lambda \to \infty} (1 - \rho_\lambda)\sqrt{s_\lambda} \to \beta, \quad \beta > 0.$$

In this case

$$(2.9) \qquad \lim_{\lambda \to \infty} \sqrt{\lambda}\, B(s_\lambda, \lambda) = \frac{\varphi(\beta)}{\Phi(\beta)}.$$

Proof. Similar to (2.1), we find

$$\mathbb{P}(\mathrm{Pois}(\lambda) < s_\lambda) = \mathbb{P}\left(\frac{\mathrm{Pois}(\lambda)-\lambda}{\sqrt{\lambda}} < \frac{s_\lambda - \lambda}{\sqrt{\lambda}}\right) = \mathbb{P}\left(\frac{\mathrm{Pois}(\lambda)-\lambda}{\sqrt{\lambda}} < (1-\rho_\lambda)\frac{s_\lambda}{\sqrt{\lambda}}\right)$$

$$(2.10) \qquad = \mathbb{P}\left(\frac{\mathrm{Pois}(\lambda)-\lambda}{\sqrt{\lambda}} < (1-\rho_\lambda)\sqrt{s_\lambda}(1+o(1))\right) \to \Phi(\beta)$$

for $\lambda \to \infty$. Stirling's formula gives

$$(2.11)$$

$$\mathbb{P}(\mathrm{Pois}(\lambda) = s) = \mathrm{e}^{-\lambda}\frac{\lambda^{s_\lambda}}{s_\lambda!} \sim \mathrm{e}^{-\lambda}\lambda^{s_\lambda} \cdot \frac{1}{\sqrt{2\pi\, s_\lambda}}\left(\frac{\mathrm{e}}{s_\lambda}\right)^{s_\lambda} = \frac{1}{\sqrt{2\pi s_\lambda}}\,\mathrm{e}^{s_\lambda - \lambda - s_\lambda \ln(\rho_\lambda)}.$$

Since $\ln(\rho_\lambda) = -(1-\rho_\lambda) - \frac{1}{2}(1-\rho_\lambda)^2 + o((1-\rho_\lambda)^2)$ we find that

$$(2.12)$$

$$\frac{\mathbb{P}(\mathrm{Pois}(\lambda) = s_\lambda)}{1 - \rho_\lambda} = \frac{1}{(1-\rho_\lambda)\sqrt{s_\lambda}}\frac{\mathrm{e}^{-\frac{1}{2}(1-\rho_\lambda)^2 s_\lambda + o\left((1-\rho_\lambda)^2 s_\lambda\right)}}{\sqrt{2\pi}} \to \frac{1}{\beta}\frac{\mathrm{e}^{-\frac{1}{2}\beta^2}}{\sqrt{2\pi}} = \frac{\varphi(\beta)}{\beta}.$$

Substituting (2.10) and (2.12) into (2.6) gives (2.7), and as a by-product also (2.9). □
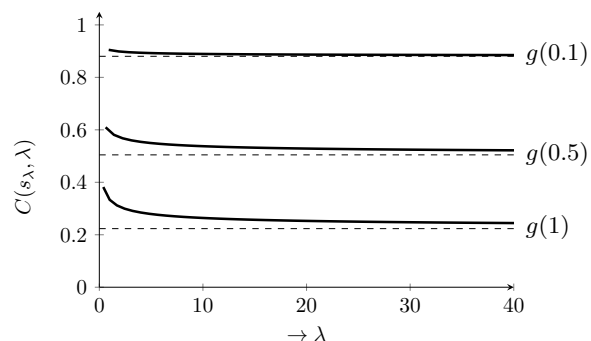
Fig. 3 *The delay probability $C(s_\lambda, \lambda)$ with $s_\lambda = [\lambda + \beta\sqrt{\lambda}]$ for $\beta = 0.1$, 0.5, and 1 as a function of λ. The limiting values $g(\beta)$ are plotted as dashed lines.*

Many of the subsequent results in this survey presented for the $M/M/s_\lambda$ queue can also be derived for the $M/M/s_\lambda/s_\lambda$ queue; we refer the reader to [82] for a detailed overview of these results. Observe that $g(\beta)$ is a strictly decreasing function on $(0, \infty)$ with $g(\beta) \to 1$ as $\beta \to 0$ and $g(\beta) \to 0$ for $\beta \to \infty$. Thus all possible delay probabilities are achievable in the QED regime, which will prove useful for the dimensioning of systems (see section 4). Although Proposition 2.1 is an asymptotic result for $\lambda \to \infty$, Figure 3 shows that $g(\beta)$ can serve as an accurate approximation for the delay probability for relatively small λ. From Proposition 2.1, it also follows that under (2.8), the limiting mean delay in (2.3) is given by

$$(2.13) \qquad \frac{C(s_\lambda, \lambda)}{(1-\rho_\lambda)\sqrt{s_\lambda}} \to \frac{g(\beta)}{\beta} =: h(\beta) \quad \text{as } \lambda \to \infty.$$

This implies that in the QED regime, the mean delay vanishes at rate $1/\sqrt{s_\lambda}$ as $\lambda \to \infty$. By Little's law this implies that the mean queue length is $O(\sqrt{s_\lambda})$. While these are all steady-state results, similar statements can be made for the entire queue-length process, as shown next.

Process-Level Convergence. QED scaling also gives rise to process-level limits, where the evolution of the system occupancy, properly centered around s_λ and normalized by $\sqrt{s_\lambda}$, converges to a diffusion process as $\lambda \to \infty$, which again is fully characterized by the single parameter β. This reflects that the system state typically hovers around the full-occupancy level s_λ, with natural fluctuations of the order $\sqrt{s_\lambda}$. Obtaining rigorous statements about stochastic-process limits poses considerable mathematical challenges. Rather than presenting the deep technical details of the convergence results, we give a heuristic explanation of how the limiting process arises and what it should look like.

The queue-length process $Q^{(s_\lambda)}(t)$ in Figure 2 with scaling rule $s_\lambda = [\lambda + \beta\sqrt{\lambda}]$ appears to concentrate around the level s_λ. As argued before, the stochastic fluctuations are of order $\sqrt{\lambda}$ or, equivalently, $\sqrt{s_\lambda}$. For that reason, we consider the centered and scaled process

$$(2.14) \qquad \bar{Q}^{(s_\lambda)}(t) := \frac{Q^{(s_\lambda)}(t) - s_\lambda}{\sqrt{s_\lambda}} \qquad \text{for all } t \geq 0$$

and ask what happens to this process as $\lambda \to \infty$. First, we consider the mean drift conditioned on $\bar{Q}^{(s_\lambda)}(t) = x$. When $x > 0$, this corresponds to a state in which

(a) $\lambda = 50$ (b) $\lambda = 100$ (c) $\lambda = 500$
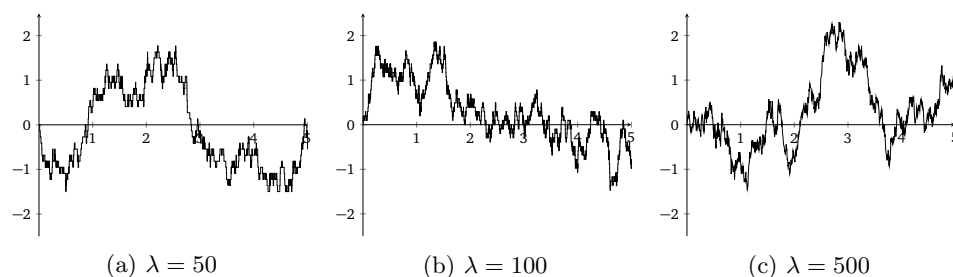
Fig. 4 *Sample paths of the normalized queue length process $\bar{Q}^{(s_\lambda)}(t)$ with $\lambda = 50$, $\lambda = 100$, and $\lambda = 500$ and $s_\lambda = [\lambda + 0.5\sqrt{\lambda}]$.*

$Q^{(s_\lambda)}(t) > s_\lambda$ and hence all servers are occupied. Therefore, the mean rate at which jobs leave the system is s_λ, while the arrival rate remains λ, so that the mean drift of $\bar{Q}^{(s_\lambda)}(t)$ in $x > 0$ satisfies

$$(2.15) \qquad \frac{\lambda - s_\lambda}{\sqrt{s_\lambda}} \to -\beta \qquad \text{as } \lambda \to \infty$$

under the scaling $\sqrt{s_\lambda}(1 - \rho_\lambda) \to \beta$ in (2.8). When $x \le 0$, only $s_\lambda + x\sqrt{s_\lambda}$ servers are working, so that the net drift is

$$(2.16) \qquad \frac{\lambda - (s_\lambda + x\sqrt{s_\lambda})}{\sqrt{s_\lambda}} \to -\beta - x \qquad \text{as } \lambda \to \infty.$$

Now, imagine what happens to the sample paths of $(\bar{Q}^{(s_\lambda)}(t))_{t \ge 0}$ as we increase λ. Within a fixed time interval, larger λ and s_λ will trigger more and more events, both arrivals and departures. Also, the jump size at each event epoch decreases as $1/\sqrt{s_\lambda}$ as a consequence of the scaling in (2.14). Hence, there will be more events, each with a smaller impact, and in the limit, as $\lambda \to \infty$, there will be infinitely many events of infinitesimally small impact. This heuristic explanation suggests that the process $\bar{Q}^{(s_\lambda)}(t)$ converges to a limit stochastic process that is continuous and has infinitesimal drift $-\beta$ above zero and $-\beta - x$ below zero. Figure 4 visualizes the emergence of the suggested scaling limit as λ and s_λ increase. The following theorem by Halfin and Whitt [63] characterizes this scaling limit formally.

THEOREM 2.2. *Let $\bar{Q}^{(s_\lambda)}(0) \overset{d}{\Rightarrow} D(0) \in \mathbb{R}$ and $\sqrt{s_\lambda}(1 - \rho_\lambda) \to \beta$. Then for all $t \ge 0$,*

$$(2.17) \qquad \bar{Q}^{(s_\lambda)}(t) \overset{d}{\Rightarrow} D(t) \qquad \text{as } \lambda \to \infty,$$

where $D(t)$ is the diffusion process with infinitesimal drift $m(x)$ given by

$$(2.18) \qquad m(x) = \begin{cases} -\beta & \text{if } x > 0, \\ -\beta - x & \text{if } x \le 0 \end{cases}$$

and infinitesimal variance $\sigma^2(x) = 2$.

The limiting diffusion process $(D(t))_{t \ge 0}$ in Theorem 2.2 is a combination of a negative-drift Brownian motion in the upper half plane and an Ornstein–Uhlenbeck process in the lower half plane. We refer to this hybrid diffusion process as the

Halfin–Whitt diffusion [153, 42, 28]. Studying this diffusion process provides valuable information for the system's performance.

The fact that the properly centered and scaled occupancy process $(\bar{Q}^{(s_\lambda)}(t))_{t\geq0}$ has the weak limit $(D(t))_{t\geq0}$, as stated in Theorem 2.2, has several important consequences. The boundary between the Brownian motion and the Ornstein–Uhlenbeck process can be thought of as the number of servers, and $(D(t))_{t\geq0}$ will keep fluctuating between these two regions. The process mimics a single-server queue above zero, and an infinite-server queue below zero, for which Brownian motion and the Ornstein–Uhlenbeck process are indeed the respective heavy-traffic limits. As β increases towards $+\infty$, capacity grows and the Halfin–Whitt diffusion will spend more time below zero.

The diffusion process $(D(t))_{t\geq0}$ can thus be employed to obtain simple approximations for the system behavior. Theorem 2.2 supports approximating the occupancy process in the $M/M/s_\lambda$ queue as

$$(2.19) \qquad Q^{(s_\lambda)}(\cdot) \stackrel{d}{\approx} s_\lambda + \sqrt{s_\lambda}D(\cdot)$$

when λ and s_λ are large. It is natural to expect that this carries over to approximations for the steady-state distribution of $(D(t))_{t\geq0}$. Let $D(\infty) := \lim_{t\to\infty} D(t)$ and $Q^{(s_\lambda)}(\infty) := \lim_{t\to\infty} Q^{(s_\lambda)}(t)$ denote the steady-state random variables. Then

$$(2.20) \qquad Q^{(s_\lambda)}(\infty) \stackrel{d}{\approx} s_\lambda + \sqrt{s_\lambda}D(\infty).$$

To rigorously justify the approximation (2.20) it is still required to show that the sequence of steady-state distributions associated with the queue-length process, when appropriately scaled, converge to the steady-state distribution associated with the diffusion process,

$$(2.21) \qquad \frac{Q^{(s_\lambda)}(\infty) - s_\lambda}{\sqrt{s_\lambda}} \stackrel{d}{\Rightarrow} D(\infty) \quad \text{as } \lambda \to \infty.$$

This has been done in [63].

The steady-state characteristics of the diffusion were studied in [63]. Since the diffusion process $(D(t))_{t\geq0}$ has piecewise-linear drift, the procedure developed in [28] to find the stationary distribution can be followed. This procedure consists of composing the density function as in (2.18) based on the density function of a Brownian motion with drift $-\beta$ for $x > 0$ and of an Ornstein–Uhlenbeck process with drift $-\beta - x$ for $x < 0$. The density function of the stationary distribution for $(D(t))_{t\geq0}$ is then proportional to $\varphi(x + \beta)/\Phi(\beta)$ for negative levels $x < 0$ and proportional to $\exp(\int_0^x m(u)\mathrm{d}u)$ for $x \geq 0$. Then, upon normalization, we find that

$$(2.22) \qquad \mathbb{P}(D(\infty) > 0) = g(\beta),$$

$$(2.23) \qquad \mathbb{P}(D(\infty) \geq x | D(\infty) > 0) = \mathrm{e}^{-\beta x} \quad \text{for } x > 0,$$

$$(2.24) \qquad \mathbb{P}(D(\infty) \leq x | D(\infty) \leq 0) = \frac{\Phi(\beta + x)}{\Phi(\beta)} \quad \text{for } x \leq 0.$$

This confirms the earlier result for the Erlang C formula in (2.7), i.e.,

$$(2.25) \qquad C(s_\lambda, \lambda) \to \mathbb{P}(D(\infty) > 0) = g(\beta) \quad \text{as } \lambda \to \infty,$$

and the scaled limiting mean delay in (2.13),

$$(2.26) \qquad \frac{\mathbb{E}[Q^{(s_\lambda)}]}{\sqrt{s_\lambda}} \to \mathbb{E}[D(\infty)] = \int_0^\infty g(\beta)\mathrm{e}^{-\beta x}\mathrm{d}x = \frac{g(\beta)}{\beta} \quad \text{as } \lambda \to \infty.$$

It is also of interest to study time-dependent characteristics like mixing times, time-dependent distributions, and first passage times to enhance our understanding of how the $M/M/s_\lambda$ queue behaves over various time and space scales. The mixing time is closely related to the spectral gap, which for the Halfin–Whitt diffusion $(D(t))_{t\geq0}$ has been identified by Gamarnik and Goldberg [43] building on the results of van Doorn [150] on the spectral gap of the $M/M/s_\lambda$ queue. An alternative derivation of this spectral gap was presented in [152, 153], along with expressions for the Laplace transform over time, and the large-time asymptotics for the time-dependent density. First passage times to large levels corresponding to highly congested states were obtained in [111, 42].

For obvious reasons, the QED regime is also referred to as the Halfin–Whitt regime, and both these names are used interchangeably in the literature.

2.2. Bulk-Service Queue. We next consider the bulk-service queue, a standard model for digital communication [29], but also for many more applications, including wireless networks, road traffic, reservation systems, and health care; see [151, Chap. 2] for an overview. Although the bulk-service queue gives rise to a plain reflected random walk and is not a multiserver queue, in the same sense as the $M/M/s$ queue, we explain below how these two models are connected.

We let jobs again arrive according to a Poisson process with rate λ, but now we discretize time, so the number of new arrivals per time period is given by a $\mathrm{Pois}(\lambda)$ random variable. Let $Q_k^{(s_\lambda)}$ denote the number of delayed jobs at the start of the kth period and assume that the system is able to process s_λ jobs at the end of each period. The queue length process can then be described by the Lindley-type recursion [101]

$$(2.27) \qquad Q_{k+1}^{(s_\lambda)} = \max\{0, Q_k^{(s_\lambda)} + \mathrm{Pois}_k(\lambda) - s_\lambda\},$$

with $Q_0^{(s_\lambda)} = 0$ and $(\mathrm{Pois}_k(\lambda))_{k\geq0}$ i.i.d. random variables. The queue length process is thus characterized by a random walk with i.i.d. steps of size $(\mathrm{Pois}(\lambda) - s_\lambda)$, with a reflecting barrier at zero. We can iterate the recursion in (2.27) to find

$$
\begin{aligned}
Q_{k+1}^{(s_\lambda)} &= \max\left\{0, Q_k^{(s_\lambda)} + \mathrm{Pois}_k(\lambda) - s_\lambda\right\} \\
&= \max\left\{0, \max\{0, Q_{k-1}^{(s_\lambda)} + (\mathrm{Pois}_{k-1}(\lambda) - s_\lambda)\} + (\mathrm{Pois}_k(\lambda) - s_\lambda)\}\right\} \\
&= \max\left\{0, (\mathrm{Pois}_k(\lambda) - s_\lambda), Q_{k-1}^{(s_\lambda)} + (\mathrm{Pois}_k(\lambda) - s_\lambda) + (\mathrm{Pois}_{k-1}(\lambda) - s_\lambda)\right\} \\
(2.28) \quad &= \max_{0\leq j\leq k}\left\{\sum_{i=1}^j (\mathrm{Pois}_{k-i}(\lambda) - s_\lambda)\right\} \overset{d}{=} \max_{0\leq j\leq k}\left\{\sum_{i=1}^j (\mathrm{Pois}_i(\lambda) - s_\lambda)\right\},
\end{aligned}
$$

where the last equality holds in distribution due to the duality principle for random walks; see, e.g., [138, sect. 7.1]. Stability requires that the mean step size satisfies $\mathbb{E}[\mathrm{Pois}(\lambda) - s_\lambda] = \lambda - s_\lambda < 0$. We use the shorthand notation for the partial sum $S_k := \sum_{i=1}^k (\mathrm{Pois}_i(\lambda) - s_\lambda)$. Let $Q^{(s_\lambda)} := \lim_{k\to\infty} Q_k^{(s_\lambda)}$ denote the stationary queue length. The probability generating function (pgf) of $Q^{(s_\lambda)}$ can then be expressed in

terms of the pgf of the positive parts of the partial sum:

$$(2.29) \qquad \mathbb{E}[z^{Q^{(s_\lambda)}}] = \exp\left\{ -\sum_{k=1}^{\infty} \frac{1}{k}\left(1 - \mathbb{E}[z^{S_k^+}]\right) \right\}, \qquad |z| \le 1.$$

From (2.29) we obtain for the mean queue length and empty-queue probability the expressions

$$\mathbb{E}[Q^{(s_\lambda)}] = \sum_{k=1}^{\infty} \frac{1}{k}\,\mathbb{E}[S_k^+],$$

$$(2.30) \qquad \mathbb{P}(Q^{(s_\lambda)} = 0) = \exp\left\{ -\sum_{k=1}^{\infty} \frac{1}{k}\,\mathbb{P}(S_k^+ > 0) \right\}.$$

There is a connection between the bulk-service queue and the $M/D/s$ queue. To see this, consider the number of queued jobs $Q^{(s_\lambda)}(k)$ at time epochs $k = 0, 1, 2, \ldots$. Then we set the period length equal to one service time. The number of new arrivals per time period is then given by the sequence of i.i.d. random variables $(\mathrm{Pois}_k(\lambda))_{k \ge 1}$. At the start of the kth period, $Q_k^{(s_\lambda)}$ customers are waiting. Since the service time of a customer is equal to the period length, all jobs that are in service at the beginning of the period will have left the system by time $k+1$. This implies that $\min\{Q_k^{(s_\lambda)}, s_\lambda\}$ of the jobs that were queued at time k are taken into service during period k. These, however, cannot possibly have departed before the end of the period, due to their deterministic service times. If $Q_k^{(s_\lambda)} < s_\lambda$, then additionally $\min\{\mathrm{Pois}_k(\lambda), s_\lambda - Q_k^{(s_\lambda)}\}$ of the new arrivals are taken into service. This yields a total of $\mathrm{Pois}_k(\lambda)$ arrivals, and $\min\{Q_k^{(s_\lambda)} + \mathrm{Pois}_k(\lambda), s_\lambda\}$ departures from the queueing system during period k. In total, this adds up to the Lindley recursion (2.27). Hence, although the bulk-service queue is technically not a multiserver queue, it gives rise to a recursive relation that describes the $M/D/s$ queue.

The reason why we choose to explain the QED regime through the bulk-service queue is that the elementary random walk perspective allows for a rather direct application of the CLT. To see this, let us ask ourselves what happens if λ grows large using the square-root rule (1.1). Since $\mathbb{E}[\mathrm{Pois}(\lambda) - s_\lambda] = \lambda - s_\lambda = -\beta\sqrt{\lambda} + o(\sqrt{\lambda})$, it makes sense to consider the scaled queue length process $\bar{Q}_k^{(s_\lambda)} := Q_k^{(s_\lambda)}/\sqrt{\lambda}$ for all $k \ge 0$, with scaled steps $Y_k^{(s_\lambda)} := (\mathrm{Pois}_k(\lambda) - s_\lambda)/\sqrt{\lambda}$. Dividing both sides of (2.28) by $\sqrt{\lambda}$ then gives

$$(2.31) \qquad \bar{Q}_{k+1}^{(s_\lambda)} = \max_{0 \le j \le k}\left\{ \sum_{i=1}^{j} Y_k^{(s_\lambda)} \right\}.$$

Hence by the CLT

$$Y_k^{(s_\lambda)} = \frac{A_k^{(\lambda)} - s_\lambda}{\sqrt{\lambda}} = \frac{A_k^{(\lambda)} - \lambda}{\sqrt{\lambda}} - \beta \overset{d}{\Rightarrow} Y_k \overset{d}{=} N(-\beta, 1)$$

for $\lambda \to \infty$. So we expect the scaled queue length process to converge in distribution to a reflected random walk with normally distributed increments, i.e., a reflected *Gaussian random walk*. Indeed, it is easily verified that [80]

$$(2.32) \qquad \bar{Q}_k^{(s_\lambda)} \overset{d}{\Rightarrow} M_{\beta,k} := \max_{0 \le j \le k}\left\{ \sum_{i=1}^{j} Y_i \right\} \quad \text{as } \lambda \to \infty.$$
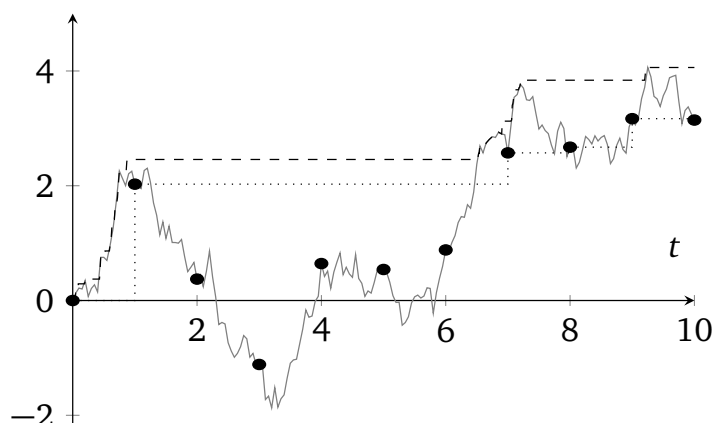
Fig. 5 *Brownian motion (gray) and embedded Gaussian random walk (marked) with their respective running maxima (dashed and dotted, respectively).*

Let $M_\beta := \lim_{k \to \infty} M_{\beta,k}$ denote the all-time maximum of a Gaussian random walk. It can be shown that M_β almost surely exists and that $\bar{Q}^{(s_\lambda)} := \lim_{k \to \infty} \bar{Q}^{(s_\lambda)} \stackrel{d}{\Rightarrow} M_\beta$, for instance, by [143, Prop. 19.2] and [9, Thm. X6.1]. The following theorem can be proved using an approach similar to that used in [83].

THEOREM 2.3. *If $(1 - \rho_\lambda)\sqrt{\lambda} \to \beta$ as $\lambda \to \infty$, then*
  (i) $\bar{Q}^{(s_\lambda)} \stackrel{d}{\Rightarrow} M_\beta$ *as $\lambda \to \infty$;*
 (ii) $\mathbb{P}(\bar{Q}^{(s_\lambda)} = 0) \to \mathbb{P}(M_\beta = 0)$ *as $\lambda \to \infty$;*
  (iii) $\mathbb{E}[\bar{Q}^{(s_\lambda)\,k}] \to \mathbb{E}[M_\beta^k]$ *as $\lambda \to \infty$ for any $k > 0$.*

Hence, Theorem 2.3 is the counterpart of Theorem 2.2, but for the bulk-service (or $M/D/s_\lambda$) queue, rather than for the $M/M/s_\lambda$ queue. Both theorems identify the stochastic-process limit in the QED regime: for the $M/M/s$ queue this is the Halfin–Whitt diffusion, and for the bulk-service queue this is the Gaussian random walk.

The Gaussian random walk is well studied [141, 30, 77, 19, 77], and there is an intimate connection with Brownian motion. The only difference, one could say, is that Brownian motion is a continuous-time process, whereas the Gaussian random walk only changes at discrete points in time. If $(B(t))_{t \geq 0}$ is a Brownian motion with drift $-\beta < 0$ and infinitesimal variance σ^2 and $(W(t))_{t \geq 0}$ is a random walk with $N(-\beta, \sigma^2)$ distributed steps and $B(0) = W(0)$, then W can be regarded as the process B embedded at equidistant time epochs. That is, $W(t) \stackrel{d}{=} B(t)$ for all $t \in \mathbb{N}^+$. For the maximum of both processes this coupling implies

$$(2.33) \qquad \max_{k \in \mathbb{N}^+} W(k) = \max_{k \in \mathbb{N}^+} B(k) \leq_{\text{st}} \max_{t \in \mathbb{R}^+} B(t),$$

where $\leq_{\text{st}}$ denotes stochastic dominance. This difference in maxima is visualized in Figure 5. It is known that the all-time maximum of Brownian motion with negative drift $-\mu$ and infinitesimal variance σ^2 has an exponential distribution with mean $\sigma/2\mu$ [64]. Hence, (2.33) implies that M_β is stochastically upper bounded by an exponential random variable with mean $1/2\beta$.

(a) Empty buffer probability

(b) Scaled queue length
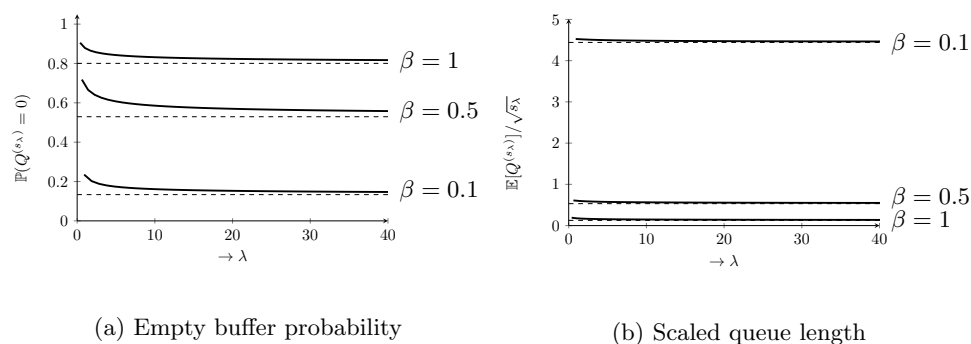
Fig. 6 *Delay probability and mean delay in the bulk-service queue with $s_\lambda = \lambda + \beta\sqrt{\lambda}$ and $\beta = 0.1$, 0.5, and 1 as a function of λ. The asymptotic approximations are plotted as dashed lines.*

Despite this easy bound, precise results for M_β are more involved. Let ζ denote the Riemann zeta function. In [30] and [77] it is shown that for $0 < \beta < 2\sqrt{\pi}$,

$$(2.34) \qquad \mathbb{P}(M_\beta = 0) = \sqrt{2}\beta \, \exp\left\{ \frac{\beta}{\sqrt{2\pi}} \sum_{l=0}^{\infty} \frac{\zeta(1/2 - l)}{l!(2l+1)} \left(\frac{-\beta^2}{2} \right)^l \right\}$$

and

$$(2.35) \qquad \mathbb{E}[M_\beta] = \frac{1}{2\beta} + \frac{\zeta(1/2)}{\sqrt{2\pi}} + \frac{\beta}{4} + \frac{\beta^2}{\sqrt{2\pi}} \sum_{l=0}^{\infty} \frac{\zeta(-1/2 - l)}{l!(2l+1)(2l+2)} \left(\frac{-\beta^2}{2} \right)^l.$$

In Figure 6, we have plotted the exact empty-buffer probability and scaled mean delay, together with their asymptotic approximations. We see that the performance measures associated with the Gaussian random walk serve as accurate approximations to performance measures describing the bulk-service queues of small to moderate size as well, just as we saw in Figure 3 for the $M/M/s_\lambda$ queue.

3. Key QED Properties. Now that we have seen how the square-root rule (1.1) yields nondegenerate limiting behavior in classical queueing models, we shall summarize the revealed QED properties and argue that these properties should hold for a more general class of models. The first property relates to the efficient usage of resources, expressed as

(Efficiency) $$\text{system load} \sim 1 - \frac{\text{constant}}{\sqrt{\text{system size}}}.$$

This property for the $M/M/s_\lambda$ queue and bulk-service queue is a direct consequence of the square-root rule. The second distinctive property is the balance between QoS and efficiency:

(Balance) $$\mathbb{P}(\text{delay}) \to \text{constant},$$

as the system size increases indefinitely. Indeed, we have shown, under (1.1) and by letting $\lambda, s_\lambda \to \infty$, that both limiting functions $g(\beta)$ in the $M/M/s_\lambda$ queue and $\mathbb{P}(M_\beta > 0)$ in the bulk-service queue can take all values in the interval $(0, 1)$ by tuning the parameter β. The third property relates to good QoS:

(QoS) $$\mathbb{E}[\text{delay}] = O(1/\sqrt{\text{system size}}).$$

Indeed, we have

$$(3.1) \quad \mathbb{E}[W^{(s_\lambda)}] = \frac{h(\beta)}{\sqrt{s_\lambda}} + o(1/\sqrt{s_\lambda}) \qquad \text{and} \qquad \mathbb{E}[Q^{(s_\lambda)}] = \sqrt{s_\lambda}\, \mathbb{E}[M_\beta] + o(1/\sqrt{s_\lambda})$$

in the $M/M/s_\lambda$ queue and the bulk-service queue, respectively. Hence the mean delay vanishes at rate $1/\sqrt{s_\lambda}$. Thus, an emerging appealing property of the QED regime is that the sojourn time of a customer is dominated by the magnitude of its service requirement. This contrasts with the so-called nondegenerate slowdown regime, where the slowdown (the ratio between the sojourn time and the service time) is strictly larger than one [11].

Since the mathematical underpinning of these properties comes from the CLT (as shown in section 2), we can expect the properties to hold for a larger class of models. We will illustrate this by discussing several extensions of the basic models discussed in section 2. The easiest way to do so seems to interpret the bulk-service queue as a many-sources model. Consider a stochastic system in which demand per period is given by some random variable A, with mean μ_A and variance $\sigma_A^2 < \infty$. For systems facing large demand we propose to set the capacity according to the more general rule

$$s = \mu_A + \beta \sigma_A,$$

which consists of a minimally required part μ_A and a variability hedge $\beta \sigma_A$. Assume that the demand is generated by n stochastically identical and independent sources. Each source i generates $A_{i,k}$ work in the kth period, with $\mathbb{E}[A_{i,k}] = \mu$ and $\mathrm{Var}\, A_{i,k} = \sigma^2$. Then the total amount of work arriving to the system during one period is $A_k^{(n)} = \sum_{i=1}^n A_{i,k}$ with mean $n\mu$ and variance $n\sigma^2$. Assume that the system is able to process a deterministic amount of work s_n per period and denote by $Q_k^{(n)}$ the amount of work left over at the end of period k. Then

$$(3.2) \qquad\qquad Q_{k+1}^{(n)} = \left(Q_k^{(n)} + A_k^{(n)} - s_n \right)^+.$$

Given that $s_n > \mathbb{E}[A_1^{(n)}] = n\mu$, the steady-state limit $Q^{(n)} := \lim_{t\to\infty} Q^{(n)}(t)$ exists and satisfies

$$(3.3) \qquad\qquad Q^{(n)} \stackrel{d}{=} \left(Q^{(n)} + A_k^{(n)} - s_n \right)^+.$$

With this many-sources interpretation [2, 76, 78], increasing the system size is done by increasing n, the number of sources. As we have seen before, it requires a rescaling of the process $Q^{(n)}$ by an increasing sequence c_n to obtain a nondegenerate scaling limit $Q := \lim_{n\to\infty} Q^{(n)}/c_n$. (We omit the technical details needed to justify the interchange of limits.) From (3.3) it becomes clear that the scaled increment

$$(3.4) \qquad \frac{A_k^{(n)} - s_n}{c_n} = \frac{\sum_{i=1}^n A_{i,k} - n\mu}{c_n} + \frac{n\mu - s_n}{c_n}$$

only admits a proper limit if c_n is of the form $c_n = O(\sqrt{n})$, by virtue of the CLT, and $(s_n - n\mu)/c_n \to \beta > 0$ as $n \to \infty$. Especially for $c_n = \sigma\sqrt{n}$, the standard deviation of the demand per period, this reveals that Q has a nondegenerate limit, which is equal in distribution to the maximum of a Gaussian random walk with drift $-\beta$ and variance 1, if

$$s_n = n\mu + \beta\sigma\sqrt{n} + o(\sqrt{n}).$$

Moreover, the results for the Gaussian random walk presented in section 2.2 are applicable to this model, and the key features of the QED scaling carry over to this more general setting. That is, for the bulk-service queue under the general assumptions above we get the QED approximation

$$(3.5) \qquad \mathbb{E}[Q^{(n)}] \approx \sigma\sqrt{n}\,\mathbb{E}[M_\beta] \approx \frac{\sigma\sqrt{n}}{2\beta}$$

for small β. Thus, the many-sources framework shows that the QED scaling finds much wider application than just queueing models with Poisson input.

Let us reflect on a key technical difference between the bulk-service queue and the $M/M/s$ queue. The bulk-service queue is and remains a one-dimensional reflected random walk, even under the QED scaling. Therefore, to establish the QED limits for the performance measure, one only needs to apply the CLT to the increments of the random walk, which readily shows that the queue converges to the Gaussian random walk. Analysis of multiserver queues is typically more challenging. Establishing QED limits for the elementary $M/M/s$ queue already contains some technically advanced steps. While we explained the high-level insights to argue the convergence of the birth-death process taking discrete steps to the continuous diffusion process, the formal proof in Halfin and Whitt [63] relies on Stone's theorem [146, 73, 99] for the weak convergence of birth-death processes to diffusion processes. However, for multiserver queues that cannot be viewed as a birth-death process, Stone's theorem cannot be applied and entirely different techniques are needed; see section 6.7.

4. Dimensioning. We adopt the term *dimensioning* used by Borst, Mandelbaum, and Reiman [21] to say that the capacity of a system is adapted to the load in order to reach certain performance levels. In [21] dimensioning refers to the staffing problem in a large-scale call center, and key ingredients are the square-root rule in (1.1) and the QED regime. We now revisit the results in [21] and its follow-up works to explain this connection to the QED regime. We also discuss the time-varying setting in which jobs arrive according to a nonhomogeneous Poisson process, and the capacity is dynamically adapted to the load.

4.1. Constraint Satisfaction. Consider the $M/M/s$ queue with arrival rate λ and service rate $\mu = 1$. A classical dimensioning problem is to determine the minimum number of servers s necessary to achieve a certain target level of service, say in terms of delay.

Suppose we want to determine the minimum number of servers such that the fraction of jobs that are delayed in the queue is at most $\varepsilon \in (0,1)$. Hence we should find

$$(4.1) \qquad s_\lambda^*(\varepsilon) := \min\left\{s > \lambda \,|\, C(s,\lambda) \leq \varepsilon\right\}.$$

But, alternatively, we can use the QED framework, which says that with s_λ as in (1.1), $\lim_{\lambda\to\infty} C(s_\lambda, \lambda) = g(\beta)$ (see Proposition 2.1). Then (4.1) can be replaced by

$$(4.2) \qquad s_\lambda^{\mathrm{QED}}(\varepsilon) = \lceil \lambda + \beta^*(\varepsilon)\sqrt{\lambda}\rceil,$$

where $\beta^*(\varepsilon)$ solves

$$(4.3) \qquad g(\beta^*) = \varepsilon.$$

In Figure 7 we plot the exact (optimal) capacity level $s_\lambda^*(\varepsilon)$ and the heuristically obtained capacity level $s_\lambda^{\mathrm{QED}}(\varepsilon)$ as functions of ε for several loads λ.

(a) $\lambda = 10$ (b) $\lambda = 100$ (c) $\lambda = 500$
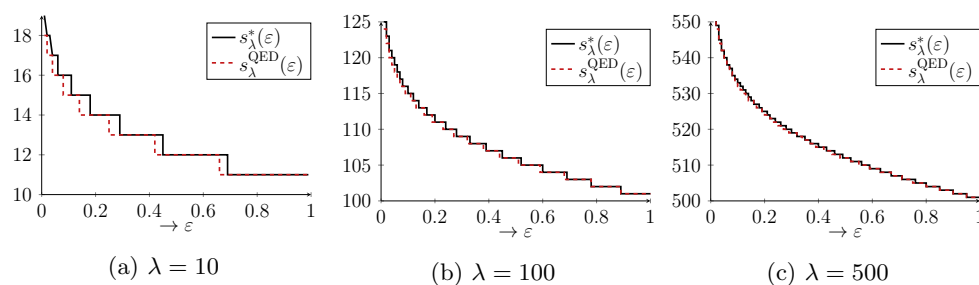
Fig. 7 *Capacity levels as a function of the delay probability targets ε.*

Observe that even for very small values of λ, the capacity function $s^{\mathrm{QED}}(\varepsilon)$ coincides with the exact solution for almost all $\varepsilon \in (0,1)$ and differs by no more than one server for all ε. Borst, Mandelbaum, and Reiman [21] recognized this in their numerical experiments too, and [81] later confirmed this theoretically (see section 4). One can easily formulate other constraint satisfaction problems and reformulate them in the QED regime—for instance, constraints on the mean delay or the tail probability of the duration of delay, e.g., $\mathbb{P}(\text{delay} > T)$, which are asymptotically approximated by $h(\beta)/\sqrt{\lambda}$ and $g(\beta)\mathrm{e}^{-\beta\sqrt{\lambda}T}$, respectively. See [21, 167, 139] for more examples.

4.2. Cost Minimization. Alternatively, one can consider optimization problems, for instance, to strike the right balance between the capacity allocation costs and delay costs incurred. More specifically, assume an allocation cost of a per server per unit time, and a penalty cost of q per delayed job per unit time, yielding the total cost function

$$\bar{K}(s,\lambda) := a\,s + q\,\lambda\mathbb{E}[\text{delay}] = a\,s + q\lambda\,\frac{C(s,\lambda)}{s-\lambda}$$

(see (2.4)), and then ask for the capacity level s that minimizes $\bar{K}(s,\lambda)$. Since $s > \lambda$, we have $\bar{K}(s,\lambda) > a\,\lambda$ for all feasible solutions s. Moreover, the minimizing value of $\bar{K}(s,\lambda)$ is invariant with respect to scalar multiplication of the objective function. Hence we equivalently seek to optimize

$$(4.4) \qquad K(s,\lambda) = r\,(s-\lambda) + \frac{\lambda}{s-\lambda}\,C(s,\lambda) \qquad \text{with } r = a/q.$$

Denote by $s^*_\lambda(r) := \arg\min_{s>\lambda} K(s,\lambda)$ the true optimal capacity level. With $s_\lambda = \lambda + \beta\sqrt{\lambda}$ and the QED limit in (2.13), we can replace (4.4) by its asymptotic counterpart:

$$(4.5) \qquad \frac{K(s_\lambda,\lambda)}{\sqrt{\lambda}} \to r\,\beta + \frac{g(\beta)}{\beta} =: K_*(\beta) \qquad \text{as } \lambda \to \infty.$$

We again obtain a limiting objective function that is easier to work with than its exact prelimit counterpart. Hence, in the spirit of the asymptotic resource allocation procedure in the previous subsection, we propose the following method to determine the capacity level that minimizes overall costs. First, (numerically) compute the value $\beta^*(r) = \arg\min_{\beta>0} K_*(\beta)$, which is well-defined, because the function $K_*(\beta)$ is strictly convex for $\beta > 0$. Then set $s^{\mathrm{QED}}_\lambda(r) = \lceil \lambda + \beta^*(r)\sqrt{\lambda} \rceil$. In Figure 8 we compare the outcomes of this asymptotic resource allocation procedure against the true optima as a function of $r \in (0,\infty)$ for several values of λ. The capacity levels $s^{\mathrm{QED}}_\lambda(r)$ and $s^*_\lambda(r)$ are aligned for almost all r and differ by no more than one server for all instances.

(a) $\lambda = 10$ (b) $\lambda = 100$ (c) $\lambda = 500$
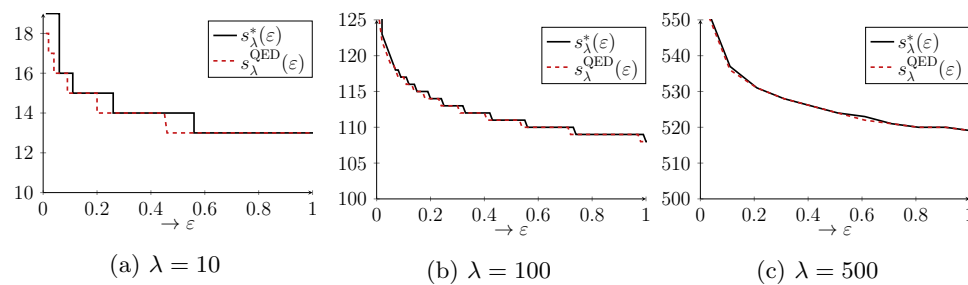
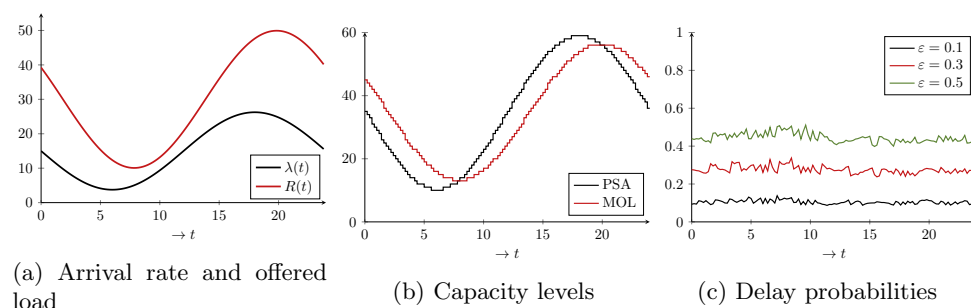**Fig. 8** *Optimal capacity levels as a function of $r = a/q$.*

4.3. Dynamic Rate Queues. We next discuss how the QED regime also finds application in systems facing a time-varying load. A time-varying arrival rate $\lambda(t)$ calls for a time-varying capacity rule $s(t)$. Again, we shall explain the main ideas through the $M/M/s$ queue, but now with a time-varying extension in which jobs arrive according to a nonhomogeneous Poisson process with rate function $\lambda(t)$, a setting typically referred to as the $M_t/M/s_t$ queue.

As in section 4.1 we want to set the capacity level $s(t)$ such that the delay probability is at most $\varepsilon \in (0,1)$ for all t. The analysis of this time-varying many-server queueing systems is cumbersome, and several approximative analysis have been proposed such as the *pointwise-stationary approximation* (PSA) [54], which evaluates the system at time t as if it were in steady state with instantaneous parameters $\lambda = \lambda(t)$, μ, and $s = s(t)$. PSA performs well in slowly varying environments with relatively short service times [54, 156], but the steady-state approximation becomes less accurate when $\lambda(t)$ displays significant fluctuations; see the numerical experiment at the end of this section. One reason for this lack of accuracy is that PSA does not account for the jobs that are actually present in the system (being in service or queued), an important piece of real-time information that should be taken into account in capacity allocation decisions. Jennings et al. [84] introduced an alternative to PSA that exploits the relation with infinite-server queues, facing a nonhomogeneous Poisson process with rate $\lambda(t)$, in which case the number of jobs at time t is Poisson distributed with mean

$$(4.6) \quad R(t) = \mathbb{E}\left[\lambda(t - B)\right]\mathbb{E}[B] = \int_0^\infty \lambda(t - u)\,\mathbb{P}(B > u)\,\mathrm{d}u = \int_0^\infty \lambda(t - u)\,\mathrm{e}^{-\mu u}\,\mathrm{d}u,$$

where B denotes the processing time of one job, in our case an exponentially distributed random variable. We remark that under general service time assumptions, we should replace $\mathbb{E}[\lambda(t - B)]$ in (4.6) with $\mathbb{E}[\lambda(t - B_e)]$, where B_e denotes the excess service time [36]. Recall that the mean delay in the QED regime is negligible; see (QoS). Hence, the total time in the system is roughly equal to its service time. Under these conditions, the many-server system can be approximated by the infinite-server approximation with offered load as in (4.6). Accordingly, we can determine the capacity levels $s(t)$ for each t based on steady-state $M/M/s$ measures with offered load $R = R(t)$. Jennings et al. [84] proceed by exploiting the heavy-traffic results of Halfin and Whitt (2.13). In conjunction with the dimensioning scheme in section 4.1, it is proposed in [84] to set

$$(4.7) \qquad\qquad s(t) = \left\lceil R(t) + \beta^*(\varepsilon)\sqrt{R(t)} \right\rceil,$$

(a) Arrival rate and offered load

(b) Capacity levels

(c) Delay probabilities

Fig. 9 *Time-varying parameters for the example with sinusoidal arrival rate.*

where $\beta^*(\varepsilon)$ solves $g(\beta^*(\varepsilon)) = \varepsilon$. Remark that the number of servers is rounded up to ensure that the achieved delay probability is indeed below ε. The time-dependent dimensioning rule in (4.7) was dubbed in [84, 119] the modified offered load (MOL) approximation. Let us now demonstrate how MOL works for an example with sinusoidal arrival rate function. Figure 9(a) shows an arrival rate pattern $\lambda(t)$ and corresponding offered load function $R(t)$ for $\mu = 1/2$. The resulting time-varying capacity levels based on the PSA and MOL approximations with $\varepsilon = 0.3$ are plotted in Figure 9(b). Through simulation, we evaluate the delay probability as a function of time for $\varepsilon = 0.1$, 0.3, and 0.5. While the PSA approach fails to stabilize the performance of the queue, the MOL method does stabilize around the target performance; see Figure 9(c). The slightly erratic nature of the delay probability as a function of time can be explained by rounding effects of the capacity level.

Because time-varying capacity allocation is an issue that recurs in many practical settings, this has been the topic of many works; see, e.g., [40, 35, 160, 161, 102, 103, 104, 67]. For an accessible overview of queueing-theoretical methods for determining capacity levels under time-varying demand, see Green, Kolesar, and Whitt [55] and references therein. Whitt [162] provides a review of queueing models with time-varying demand.

5. Convergence Rates. By now, it is clear that the QED paradigm is based on limit theorems that apply when systems become infinitely large. In practice, even large systems are finite, which makes it important to quantify the error made in approximating a finite system by a limiting object. As it turns out, QED approximations are in many cases highly accurate, already for relatively small or moderately sized systems. In this section we show how to quantify these errors by determining the rate of convergence of certain performance measures to their asymptotic limits. A first sign of this was seen through the accuracy of the asymptotic dimensioning schemes in section 4. These convergence rates are typically of order $1/\sqrt{s}$, with s the system size. This again confirms the deep connection with the CLT with a typically error also of order $1/\sqrt{s}$, but then with s the number of random variables in the sum.

5.1. Bounds. A convergence rate can also be interpreted as the (main) error made when using the QED limits as approximations for the real performance measures. Whenever we find ways to obtain explicit and precise descriptions of the convergence rates, this can also be used to correct the limiting expression for the finite size of the system. We will also show how such effective corrections can be obtained and applied directly in the QED framework.

Recall that when λ is a positive integer, $\mathrm{Pois}(\lambda)$ can be written as the sum of λ $\mathrm{Pois}(1)$ random variables. A more general version of the CLT in Theorem 1.1 related to the Berry–Esséen bound (see, e.g., [41, sect. XVI.5]) implies that

$$\mathbb{P}(\mathrm{Pois}(\lambda) \leq s_\lambda) = \Phi(\beta) + O(\lambda^{-1/2}) \tag{5.1}$$

as $\lambda \to \infty$ with s_λ as in (1.1). Comparing (5.1) with (1.4), (5.1) not only shows convergence of $\mathbb{P}(\mathrm{Pois}(\lambda) \leq s_\lambda)$ to $\Phi(\beta)$, but also quantifies (roughly) the convergence rate as $O(\lambda^{-1/2})$. To obtain better estimates for the error of order $1/\sqrt{\lambda}$, one can derive asymptotic expansions. There are various general theorems that yield asymptotic expansions for $\mathbb{P}(A_\lambda \leq s)$ in ascending positive powers of $\lambda^{-1/2}$; see, e.g., [15, 18, 41, 71, 86, 129]. One example would be the Edgeworth expansion, which for the Poisson distribution yields (see [15, eq. (4.18)])

$$\mathbb{P}(\mathrm{Pois}(\lambda) \leq s_\lambda) = \Phi(\beta) - \frac{\varphi(\beta)(\beta^2 - 1)}{6\sqrt{\lambda}} + O(1/\lambda). \tag{5.2}$$

The technical challenge in determining convergence rates is that we need to establish an asymptotic expansion rather than just the limit theorem. We shall demonstrate this for the $M/M/s_\lambda$ queue using the asymptotic evaluation of integrals through the Laplace method. The formula $C(s_\lambda, \lambda)$ in its basic form is only defined for integer values of s_λ. An extension of this formula that is well-defined for all real $s_\lambda > \lambda$ is given by (see, e.g., Jagers and Van Doorn [75])

$$C(s_\lambda, \lambda)^{-1} = \lambda \int_0^\infty t e^{-\lambda t} (1+t)^{s_\lambda - 1} \mathrm{d}t. \tag{5.3}$$

We introduce the following key parameters:

$$\alpha = \sqrt{-2s_\lambda(1 - \rho_\lambda + \ln \rho_l)}, \tag{5.4}$$

$$\beta = (s_\lambda - \lambda)/\sqrt{\lambda}, \tag{5.5}$$

$$\gamma = (s_\lambda - \lambda)/\sqrt{s} = (1 - \rho_\lambda)\sqrt{s_\lambda} = \beta\sqrt{\rho_\lambda}. \tag{5.6}$$

It has been shown in [82] that $\alpha < \beta$. By expanding $\frac{1}{2}\alpha^2$ in powers of $(1 - \rho_\lambda)$, it easily follows that $\gamma < \alpha$, so we have $\gamma < \alpha < \beta$.

THEOREM 5.1. *For $s > \lambda$,*

$$C(s_\lambda, \lambda) \leq \left[\rho_l + \gamma\left(\frac{\Phi(\alpha)}{\varphi(\alpha)} + \frac{2}{3}\frac{1}{\sqrt{s_\lambda}}\right)\right]^{-1} \tag{5.7}$$

and

$$C(s_\lambda, \lambda) \geq \left[\rho_\lambda + \gamma\left(\frac{\Phi(\alpha)}{\varphi(\alpha)} + \frac{2}{3}\frac{1}{\sqrt{s_\lambda}} + \frac{1}{\varphi(\alpha)}\frac{1}{12s_\lambda - 1}\right)\right]^{-1}. \tag{5.8}$$

Notice that the structure of the bounds (5.7) and (5.8) is quite similar to the Halfin–Whitt approximation $C(s_\lambda, \lambda) \approx g(\beta)$. Indeed, using $s_\lambda = \lambda + \beta\sqrt{\lambda}$ with β fixed and letting $\lambda \to \infty$, one can see that α and γ both converge to β. With the above theorem at hand, convergence of $C(s_\lambda, \lambda)$ towards the Halfin–Whitt function $g(\beta)$ follows, which provides an alternative proof and confirmation of Proposition 2.1. More importantly, the bounds (5.7)–(5.8) are sharp in the QED regime for small and

Table 1   *Results for the bounds on $C(s_\lambda, \lambda)$ for $\beta = 1$.*

s_λ	λ	α	(5.8)	$C(s_\lambda, \lambda)$	(5.7)	$\frac{(5.7)-(5.8)}{C(s_\lambda,\lambda)}$	(5.9)	$\frac{\lvert(5.9)-C(s_\lambda,\lambda)\rvert}{C(s_\lambda,\lambda)}$
1	0.382	0.830	0.36571	0.38197	0.39437	$7.504 \cdot 10^{-2}$	0.45085	$1.803 \cdot 10^{-1}$
2	1.000	0.879	0.32678	0.33333	0.33936	$3.772 \cdot 10^{-2}$	0.36395	$0.918 \cdot 10^{-2}$
5	3.209	0.924	0.28886	0.29097	0.29328	$1.518 \cdot 10^{-2}$	0.30185	$3.739 \cdot 10^{-2}$
10	7.298	0.946	0.26937	0.27030	0.27142	$7.616 \cdot 10^{-3}$	0.27540	$1.886 \cdot 10^{-2}$
20	16.000	0.962	0.25565	0.25608	0.25663	$3.818 \cdot 10^{-3}$	0.25851	$9.495 \cdot 10^{-3}$
50	43.411	0.976	0.24361	0.24377	0.24398	$1.531 \cdot 10^{-3}$	0.24470	$3.820 \cdot 10^{-3}$
100	90.488	0.983	0.23761	0.23769	0.23779	$7.665 \cdot 10^{-4}$	0.23814	$1.916 \cdot 10^{-4}$
200	186.349	0.988	0.23340	0.23344	0.23349	$3.836 \cdot 10^{-4}$	0.23366	$9.602 \cdot 10^{-4}$
500	478.134	0.993	0.22969	0.22970	0.22972	$1.536 \cdot 10^{-4}$	0.22979	$3.848 \cdot 10^{-4}$
1000	968.873	0.995	0.22783	0.22783	0.22784	$7.683 \cdot 10^{-5}$	0.22788	$1.926 \cdot 10^{-4}$

moderately sized systems. The difference between the lower and upper bounds is only $O(1/s_\lambda)$ In Table 1, we keep $\beta = 1$ fixed and vary s_λ. The load λ is chosen such that $s_\lambda = \lambda + \beta\sqrt{\lambda}$. The quality of the bounds is apparent, even for small systems, and certainly compared to the asymptotic approximation $g(1) = 0.22336$.

We by now know that $C(s_\lambda, \lambda) \to g(\beta)$, and D'Auria [34] proved that $C(s_\lambda, \lambda) \geq g(\beta)$ for all $\lambda, \beta > 0$. Using the bounds in (5.7) and (5.8), it was shown by Janssen, van Leeuwaarden, and Zwart [81] that as $\lambda \to \infty$,

$$(5.9) \qquad C(s_\lambda, \lambda) \approx g(\beta) + g_\bullet(\beta)\frac{\beta}{\sqrt{\lambda}},$$

with

$$(5.10) \qquad g_\bullet(\beta) = g(\beta)^2 \left[\frac{1}{3} + \frac{\beta^2}{6} + \frac{\Phi(\beta)}{\phi(\beta)}\left(\frac{\beta}{2} + \frac{\beta^3}{6}\right)\right].$$

This result can be interpreted as the counterpart of (5.2), but then not for the Poisson distribution in the CLT regime, but for the delay probability in the QED regime. In Table 1 we see that (5.9) leads to much sharper approximations than the original asymptotic approximation $g(1) = 0.22336$.

5.2. Optimality Gaps. Given these refinements to the asymptotic delay probability, we revisit the cost minimization problem discussed in section 4 and ask ourselves what can be said about the associated optimality gaps when dimensioning principles based on the asymptotic approximations are used.

Recall that under a linear cost structure, we aim to find the minimizing value s_λ^* of $K(s, \lambda)$ as in (4.4) (we omit the argument r in this section for brevity). Since $K(s_\lambda, \lambda) \to K_*(\beta)$ as $\lambda \to \infty$ with $s_\lambda = \lambda + \beta\sqrt{\lambda}$, we alternatively considered asymptotic minimizer $s_\lambda^{\mathrm{QED}} = [\lambda + \beta^*\sqrt{\lambda}]$ with β^* minimizing $K_*(\beta)$, and Figure 8 illustrated the accuracy of this asymptotic dimensioning scheme of systems of various sizes. Indeed, Borst, Mandelbaum, and Reiman [21] showed that s_λ^{QED} is asymptotically optimal in the sense that

$$(5.11) \qquad K\left(s_\lambda^{\mathrm{QED}}, \lambda\right) = K(s_\lambda^*, \lambda) + o(\sqrt{\lambda}).$$

The corrected approximation for the delay probability in (5.9), however, provides a means to improve the accuracy of s_λ^{QED}. Namely, by substituting (5.9) into (4.4), it

is clear that we can write

$$(5.12) \qquad \frac{K(s_\lambda, \lambda)}{\sqrt{\lambda}} \approx K_*(\beta) + \frac{g_\bullet(\beta)}{\sqrt{\lambda}} =: K_\bullet(\beta),$$

with an error that is of order $O(1/\lambda)$ for uniformly bounded β and $g_\bullet(\beta) := g_*(\beta)/\beta$. If we consider the approximated cost function $K_\bullet(\beta)$ in (5.12), and let β_λ^* be the associated minimizer, then we expect the refined square-root rule $s_\lambda^\bullet := [\lambda + \beta_\lambda^* \sqrt{\lambda}]$ to give a better approximation to the true optimizer s_λ^*. It is shown in Janssen, van Leeuwaarden, and Zwart [81], by invoking Taylor's theorem, that $\beta_\lambda^* = \beta^* + \beta_\bullet/\sqrt{\lambda} + O(1/\lambda)$ with

$$(5.13) \qquad \beta_\bullet = -\frac{\beta^* g_\bullet'(\beta^*)}{K_*''(\beta^*) + 2r}.$$

The resulting refined square-root rule $s_\lambda^\bullet = [s_\lambda^{\mathrm{QED}} + \beta_\bullet]$ indeed yields an improvement over the original square-root rule in terms of the optimality gap. Namely (see [81, Thm. 2]),

$$(5.14) \qquad K(s_\lambda^\bullet, \lambda) = K(s_\lambda^*, \lambda) + O(1/\sqrt{\lambda}).$$

Observe that the characterization of $s_\lambda^\bullet$ as an $O(1)$ correction to the original square-root rule (1.1) provides a rigorous mathematical underpinning for the exceptionally good performance of the QED dimensioning scheme observed in Figure 8.

In the context of $M/M/s + M$ queues, Zhang, van Leeuwaarden, and Zwart [167] obtain similar results on optimality gaps. Motivated by the results in [81, 167], Randhawa [132] takes a more abstract approach to quantify optimality gaps of asymptotic optimization problems. He shows under generally assumptions that when the approximation to the objective function is accurate up to $O(1)$, the prescriptions that are derived from this approximation are $o(1)$-optimal. The optimality gap thus asymptotically becomes zero. This general setup is shown in [132] to apply to the $M/M/s$ queues in the QED regime, which confirmed and sharpened the results on the optimality gaps in [81, 167]. The abstract framework in [132], however, can only be applied if refined approximations as discussed above are available. Optimality gaps in settings with admission control in the QED regime, based on a trade-off between revenue, costs, and service quality, have been studied in [140].

5.3. Refinements. A downside of heavy-traffic analysis is that the results are of an asymptotic nature, and therefore approximations. Obtaining corrections or refinements is one of the main goals of many research efforts, and the demonstration in the previous two subsections is only a small part of a richer and active line of research. In the QED context, this leads to the question of whether the three universal properties provide the correct insight if the system size is only moderate or if the efficiency hypothesis (Efficiency) is not exactly satisfied.

In the setting of a single server, Siegmund [142] proposes a corrected diffusion approximation for the waiting time. In heavy traffic, its distribution is approximately exponential. Siegmund gives a precise estimate of the correction term, nowadays a classical result and textbook material; cf. [9, p. 369]. Siegmund's first-order correction is extended by Blanchet and Glynn [19], who give a full series expansion for the $G/G/1$ waiting time distribution in heavy traffic. A result similar to [142] has been presented in the QED context for the $M/M/s$ queue in section 5 of this survey. A common

threat of these approaches is that detailed information on the prelimit distribution needs to be available.

In addition to corrected diffusion approximations, a number of other refinements exist in the literature that provide improved (with respect to the heavy-traffic limit) approximation of the invariant distribution. One class of such approximations is based on variations of Stein's method [23, 25]. Another class of approximations is based on the idea of considering the diffusion limit of a Markovian queue and replacing the drift and diffusion coefficients with terms that depend on the parameters in the prelimit. The goal is to improve the convergence rate in the QED regime and to make the approximations accurate in other scaling regimes, hence the term *universal approximations*. We refer the reader to [60, 58, 68] for a more in-depth discussion, and we explain the idea of modifying a diffusion in the context of the Halfin–Whitt diffusion, following an idea of Braverman and Dai [24].

Recall from Theorem 2.2 that the scaled queue length process in the QED regime converges to a diffusion process with infinitesimal drift $m(x) = -\beta - x\mathbb{1}_{\{x \leq 0\}}$ and infinitesimal variance $\sigma^2(x) = 2$. β can be expressed in terms of the prelimit characteristics by the expression $\beta = (s - \lambda)/\sqrt{\lambda}$. The idea in [24] is now to replace the diffusion coefficient and consider

$$(5.15) \qquad \sigma_\lambda^2(x) = 1 + \mathbb{1}_{\{x > -\sqrt{\lambda}\}} \left(1 - \frac{m(x)}{\sqrt{\lambda}} \right).$$

The resulting approximation for the steady-state density is explicit, and it is shown in [24] that the resulting distributional approximation has an error of the order $1/\lambda$, while the QED approximation has a much larger error of order $1/\sqrt{\lambda}$. Though the associated approximation for the delay probability is worse than the approximations and bounds presented in section 5 of this paper, the idea of modifying the limiting diffusion appropriately seems to be of high potential, and worthy of further investigation. The same can be said about Stein's method. Another line of research that we think deserves attention is the development of nonasymptotic bounds that are accurate in a QED setting. Recent work in this direction is [51].

6. Extensions. By now we should have developed a good understanding of why the mathematical theory that comes with the QED regime for many-server systems ranks among the most celebrated principles in applied probability. The goal of the present section is to provide a survey of results for models that are more elaborate than the basic models discussed so far. In particular, we shall consider more elaborate models that incorporate various forms of user behavior (such as abandonments and strategic behavior) and consider the impact of blocking in systems with finite waiting rooms, as well as loss networks. We also consider parameter uncertainty, systems with load balancing, and non-Markov $G/G/1$ systems. This list of extensions is far from exhaustive but gives the reader a taste of how the fundamental QED principles given in section 3 carry over or should be adapted in more elaborate settings. For multiclass job types we refer the reader to [65, 6, 13, 59, 62, 148, 111, 11, 12]. Extensions to heterogeneous servers can be found in [4, 8, 116, 145], and congestion control mechanisms are considered in [139, 97, 10, 12, 11, 20, 62, 79].

6.1. Abandonments. So far, we have surveyed standard systems in which all arriving jobs join the queue and stay until eventually being processed by one of the servers. One model extension that is featured prominently in the literature is abandonment caused by customer impatience, in which case customers leave the system without being served [47, 27, 126].

The canonical model for abandonments is the $M/M/s + M$ or Erlang A model [126, 48], with dynamics similar to the $M/M/s$ queue, with the additional feature that each job is assigned an i.i.d. patience time, which is exponentially distributed with mean $1/\theta$. If a job's patience time expires before reaching an available server, the job leaves (abandons) the system. As the number of jobs in the Erlang A queue remains a birth-death process, its stationary distribution and associated performance measures are fairly well understood, also in the QED regime [48, 166, 168]. Garnett, Mandelbaum, and Reiman [48] and Zeltyn and Mandelbaum [166] show that in the QED regime, with $s_\lambda = \lambda + \beta\sqrt{\lambda}$ and $\lambda \to \infty$,

$$(6.1) \qquad \mathbb{P}(\text{delay}) \to \left(1 + \sqrt{\theta}\,\frac{k(\beta/\sqrt{\theta})}{k(-\beta)}\right)^{-1}$$

and

$$(6.2) \qquad \sqrt{\lambda}\,\mathbb{P}(\text{abandon}) \to \frac{\sqrt{\theta}\,k(\beta/\sqrt{\theta}) - \beta}{1 + \sqrt{\theta}\,k(\beta/\sqrt{\theta})/k(-\beta)},$$

where $k(\beta) = \varphi(\beta)/\Phi(-\beta)$. Hence, the universal QED properties, discussed in section 3, remain intact when the model includes abandonments. Moreover, the probability that a job abandons the system vanishes at rate $O(1/\sqrt{\lambda})$ as $\lambda \to \infty$. In [166], the stationary QED limits for more generally distributed patience time are derived, for which similar limiting behavior is proved. More surprisingly, it is shown that the limit is insensitive to the patience time distribution as long as its density at 0, i.e., the probability of abandoning immediately upon arrival, is fixed. On the process level, the appropriately scaled queue length process of the $M/M/s_\lambda + M$ model in the QED regime can be shown to converge to a piecewise-linear Ornstein–Uhlenbeck process with drift terms

$$(6.3) \qquad m(x) = \begin{cases} -\beta - \theta x & \text{if } x > 0, \\ -\beta - x & \text{if } x \le 0 \end{cases}$$

and infinitesimal variance $\sigma^2(x) = 2$; see, e.g., [48]. Notice that for $\theta = 0$, we retrieve the Halfin–Whitt diffusion in Theorem 2.2. Under more general assumptions, [115] characterizes the QED limiting process for the $G/GI/s + GI$ queue. More specifically, they find that the QED limit of the $G/M/s + GI$ queue is still a piecewise-linear Ornstein–Uhlenbeck process.

The general $G/G/s + G$ queue under various modeling assumptions and its limiting process in the QED regime has been studied in [48, 47, 159, 117, 166, 115, 89, 32, 134, 85, 168]. These works also include the case where the system is balanced from the point of view of the abandonment probability, which relates to the efficiency driven regime in our setting. Surveys on systems with abandonments are Ward [155] and Dai and He [33].

6.2. Finite Waiting Space. We have assumed so far that systems have infinite buffers for storing delayed jobs. Systems in applications such as data centers and hospitals, however, are often limited in the number of jobs that can be held simultaneously. Depending on the practical setting and admission policy, if the maximum capacity, say k, is reached, newly arriving jobs can either leave the system immediately (blocking), reattempt getting access later (retrials), or queue outside the facility

(holding). In any case, expectations are that the queueing dynamics within the resource sharing facility are affected considerably in the presence of such additional capacity constraints.

We illustrate these implications through the $M/M/s/k$ queue, that is, the standard $M/M/s$ queue with the additional property that a job that finds upon arrival k jobs already present in the system is blocked/lost. To avoid trivialities, let $k \geq s$. Since the mean workload reaching the servers is less than in an finite buffer ($k = \infty$) scenario, one expects less congestion and resource utilization.

Consider the $M/M/s_\lambda/k_\lambda$ in the QED regime. So, let λ increase while s_λ scales as in (1.1). We then ask how k_λ should scale along with λ and s_λ to maintain the nondegenerate behavior as seen in section 2.1. We provide a heuristic answer. Let $Q^{(s_\lambda,k_\lambda)}$ and $W^{(s_\lambda,k_\lambda)}$ denote the number of jobs in the system and delay in the $M/M/s_\lambda/k_\lambda$ queue in steady state. If there were no finite-size constraints, then through (2.22)–(2.24), we find as $\lambda \to \infty$

$$(6.4) \qquad \mathbb{P}(Q^{(s_\lambda)} \geq k_\lambda) = \mathbb{P}\left(\frac{Q^{(s_\lambda)} - s_\lambda}{\sqrt{s_\lambda}} \geq \frac{k_\lambda - s_\lambda}{\sqrt{s_\lambda}}\right) \to g(\beta)\, e^{-\beta\gamma},$$

where $\gamma = \lim_{\lambda\to\infty}(k_\lambda - s_\lambda)/\sqrt{s_\lambda}$. Hence, asymptotically the finite-size effects only play a role if the extra variability hedge of k_λ is of order $\sqrt{s_\lambda}$ (or, equivalently, $o(\sqrt{\lambda})$). Furthermore, if the variability hedge is $o(\sqrt{\lambda})$, then we argue that, asymptotically, all jobs that do enter the system have probability of delay equal to zero. More formally, under the *twofold scaling rule*

$$(6.5) \qquad \begin{cases} s_\lambda = \lambda + \beta\sqrt{\lambda} + o(\sqrt{\lambda}), \\ k_\lambda = s_\lambda + \gamma\sqrt{s_\lambda} + o(\sqrt{\lambda}), \end{cases}$$

it is not difficult to deduce that (see, e.g., [118])

$$(6.6) \qquad \mathbb{P}(\text{delay}) \to \left(1 + \frac{\beta\,\Phi(\beta)}{(1 - e^{-\beta\gamma})\varphi(\beta)}\right)^{-1} \quad \text{as } \lambda \to \infty,$$

which is strictly smaller than $g(\beta)$ in (3), but still bounded away from both 0 and 1, and thus the balance property holds. Furthermore, the buffer size of the queue is $n_\lambda - s_\lambda = \gamma\sqrt{s_\lambda}$, so that by Little's law, the mean delay of an admitted job is $O(1/\sqrt{s_\lambda})$, implying that the QoS property holds. Even though resource utilization in the $M/M/s_\lambda/n_\lambda$ is less efficient than in the queue with unlimited waiting space, it can be shown that $\rho_\lambda \to 1$ as $\lambda \to \infty$. Hence, all three key characteristics of the QED regime are carried over to the finite-size setting if one uses (6.5).

On a process level, adding a capacity constraint translates to adding a reflection barrier to the normalized queue length process $\bar{Q}^{(s_\lambda,n_\lambda)} = (Q^{(s_\lambda,n_\lambda)} - s_\lambda)/\sqrt{s_\lambda}$, at γ, as is illustrated by the sample paths of $\bar{Q}^{s_\lambda,n_\lambda}$ for three values of λ in Figure 10. Indeed, nondegenerate limiting behavior can be expected when the additional space $\gamma\sqrt{s_\lambda}$ is of the same order as the natural fluctuations of the arrival process; see [118].

6.3. Strategic Behavior. The purpose of this section is to show that the universal QED properties may no longer hold when there is strategic interaction between the system operator and potential users. In several applications users have the option whether to join a certain congestion-dependent service or not, leading to a game theoretic setting where the provider of a service maximizes profit, and users decide to join a service depending on their utility, possibly involving the mean delay. If

(a) $\lambda = 50$ (b) $\lambda = 100$ (c) $\lambda = 500$
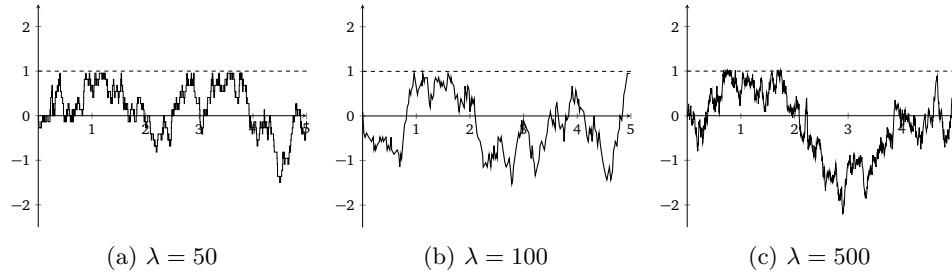
Fig. 10 *Sample paths of the normalized queue length process $\bar{Q}^{(s_\lambda, n_\lambda)}(t)$ with $\lambda = 50$, 100, and 500 under scaling (6.5) with $\beta = 0.5$ and $\gamma = 1$.*

the market size is large, the QED capacity allocation rule can emerge endogenously, though it is possible to obtain other scaling rules as well. Examples of such studies include [100, 52, 125]. For illustrative purposes, we briefly describe the model and results of Nair, Wierman, and Zwart[125] in more detail.

A user needs to decide whether or not to use a congestion-dependent service which is free for the user (and supported by advertisements—think of Google or Facebook). If the total user base that uses the service has magnitude λ, the user receives a utility $V(\lambda)$ (this may be increasing with λ in a social network context), and a congestion-dependent disutility $\xi(s, \lambda)$, chosen according to the mean delay in the $M/M/s$ queue, i.e., $\xi(s, \lambda) = C(s, \lambda)/(s - \lambda)$ for $\lambda < s$ and ∞ otherwise. Given the choice of a number of data processing units of the service provider, an infinitesimal user will join if and only if $V(\lambda) - \xi(s, \lambda)$ is nonnegative. The total market size of the user base is equal to Λ, which is assumed to be large. For illustrative purposes, we restrict our study to the case where the entire user population can cooperate, and therefore the total arrival rate becomes

$$(6.7) \qquad \hat{\lambda}_\Lambda(s) = \max\left\{\arg\max_{\lambda \in [0, \Lambda]} [\lambda V(\lambda) - \lambda \xi(s, \lambda)]\right\}.$$

The firm optimizes its revenue given this user behavior. The cost of each resource is scaled to 1, and the average advertisement revenue per unit of users is set to b_1. In this case the optimal number of services $k^*(\Lambda)$ becomes

$$(6.8) \qquad k^*_\Lambda = \max\left\{\arg\max_{k \geq 0} [b_1 \hat{\lambda}_\Lambda(k) - k]\right\}.$$

It is possible to determine how k^*_Λ scales with Λ. As is shown in Theorem 1 of [125], if $\alpha = \lim_{\lambda \to \infty} U'(\lambda) \in (0, \infty)$ (which is the case if V is converging to a constant, corresponding with an online service like Google), then there exists a strictly positive and decreasing function β of α such that $k^*_\Lambda = \Lambda + \sqrt{\beta(\alpha)\Lambda}(1 + o(1))$. In the case $V(\lambda) = \lambda^v$ for some $v > 0$, then $\alpha = \infty$ and users are more interested to join a service if other users are present (as is the case in a social network like Facebook). In this case, the firm can give less QoS: the number of spare servers becomes of the order $\sqrt{\Lambda^{1-c}}$. If users cannot collaborate, the firm only needs two spare servers to maximize its profit: the choice $k^*_\Lambda = \Lambda + 2$ makes the entire user population join the network. This is an example of what is called a *tragedy of the commons*. There are many additional opportunities for research in this domain; the recent monograph [66] on the interface of game theory and queueing provides an excellent starting point.

Another interesting research line is the analysis of the effect of delay announcements on consumer behavior in a many-server setting; see [7, 69]. A recent survey on this topic is [72].

6.4. Networks. The models shown so far are all single-station models. The analysis of networks in the QED regime is more challenging; see, e.g., [113] for a QED analysis of Markovian networks with time-varying rates and time-varying number of servers, using the technique of strong approximations to obtain functional CLTs. Parallel service systems with multiple service pools and multiple customer classes can also be viewed as network extensions of single stations; see [10] for a diffusion control problem and [62] for a strong approximation approach. In both papers the QED regime plays a key role. In this section, we restrict ourselves to explaining how the fundamental QED properties can be extended to a tractable class of loss networks.

A *loss network* is an extension of the Erlang B model and is especially relevant for the analysis of communication networks. Consider a network with J links, and suppose that link j, $j = 1, \ldots, J$, comprises C_j circuits (servers). There are R classes of calls called routes. A call on route r uses A_{jr} circuits from link j, where we take A_{jr} to be either 0 or 1. Calls of route r arrive according to a Poisson arrival process of rate λ_r, and a call is blocked if the appropriate servers are not available. Assuming unit exponential services on each route, it can be shown that the invariant distribution $\pi(n)$ can be written as a ratio of two Poisson probabilities. Specifically, let N be an R-dimensional vector of independent Poisson random variables where the rate of N_r, $r \in R$, equals λ_r. Now

$$(6.9) \qquad \pi(n) = \frac{P(N = n)}{\mathbb{P}(AN \leq C)} I(An \leq C),$$

with $C = (C_1, \ldots, C_J)$. Unfortunately the computation of the normalizing constant $\mathbb{P}(AN \leq C)$ is nontrivial for large systems. It is possible to develop a Gaussian approximation using a central limit approach which can be seen as an extension of our efficiency hypothesis (Efficiency). To have all links in the network critically loaded, one considers the case where $\lambda_r = \lambda \nu_r$ and replaces C with $\lambda C + \beta \sqrt{\lambda}$, with λ a scaling parameter, as before and β a vector. It is possible to show that all links in the network are critically loaded in this case if $A\nu = C$. For cases where only a subset of links in the network is critically loaded, one must proceed in a much more delicate manner; see [70].

The normalizing constant can, under our scaling hypothesis, be written as $\mathbb{P}((AN - \lambda C) \leq \sqrt{\lambda}\beta)$, which converges to a multivariate Gaussian distribution, as $A\lambda\nu = \lambda C$. The analogue of (QoS) can be seen as the fact that each user has a dedicated group of service once it is admitted, and the probability of blocking decays at a rate $O(1/\sqrt{\lambda})$, which is analogous to the blocking rate in a simple Erlang B system, relating to (Balance). For more details on these properties and more background, we refer the reader to [93], which is still a valuable source of information, and to the more recent [94]. For recent progress on computational procedures, we refer the reader to [88, 3].

Other network extensions of QED principles have been established, though it is typically hard to derive explicit results for the associated limiting distributions and/or processes. For work on fork–join networks in the QED regime, we refer the reader to [106, 107, 108], while bandwidth sharing networks in the QED regime have been

investigated in [135]. For network analogues of the twofold scaling rule presented in section 6.2 (in particular (6.5)), see [95, 164, 147].

6.5. Parameter Uncertainty. Models describing multiserver systems typically assume perfect knowledge of the model primitives, including the mean demand per time period. For large-scale systems, the dominant assumption in the literature is that demand arrives according to a nonhomogeneous Poisson process, just as in section 4, which translates to the assumption that arrival rates are known for each basic time period (second, hour, or day). In practice, however, estimates for mean demand typically rely on historical data and are therefore subject to uncertainty. This parameter uncertainty is likely to affect the effectiveness of capacity sizing rules. Examples of studies in staffing or resource allocation rules under parameter uncertainty are [112, 87, 61, 16, 159]; see also the references therein.

As an illustration, consider a resource allocation problem with Poisson λ arrivals and exponential (μ) servers. Suppose that $\mu = 1$, and λ is unknown. For instructive purposes, we make a resource allocation decision s based on the infinite server approximation $\mathbb{P}(\text{Pois}(\lambda) > s_\lambda) \leq \varepsilon$. In case λ is known and large, the choice $s_\lambda = \lambda + \beta\sqrt{\lambda}$, with $\beta = 1 - \Phi^{-1}(\varepsilon)$, would be natural; see (2.1). If λ is not known but needs to be estimated from data, it is instructive to see how the choice of s is affected. Suppose we have an estimator $\hat{\lambda}$ of λ which is approximate normally distributed with standard deviation σ. When would it be appropriate to simply take $s = \hat{\lambda} + \beta\sqrt{\hat{\lambda}}$? To obtain some insight, we use the approximation $\text{Pois}(\lambda) \sim \lambda + G\sqrt{\lambda}$ and assume $\hat{\lambda} = \lambda + G_0\sigma$, where G and G_0 are independent standard normal variables. Then we see the following: If λ is large, we need to pick s such that $P(\hat{\lambda} + \sigma G_0 + G\sqrt{\hat{\lambda}} > s) = \varepsilon$, yielding $s = \hat{\lambda} + \beta\sqrt{\sigma^2 + \hat{\lambda}}$. If σ^2 is of the order $\hat{\lambda}$, it follows that the naive rule $s = \hat{\lambda} + \beta\sqrt{\hat{\lambda}}$ leads to poor system performance. It would be valuable to develop similar quantitative insights for more realistic models.

A related yet fundamental difficulty arises when fluctuations in demand are larger than anticipated by the Poisson assumption. In this case, the Poisson assumption is wrong, and standard QED rules need to be modified, even when there is an infinite amount of data. Indeed, although natural and convenient from a mathematical viewpoint, the Poisson assumption often fails to be confirmed in practice. A deterministic arrival rate implies that the demand over any given period is a Poisson random variable, whose variance equals its expectation. A growing number of empirical studies of service systems shows that the variance of demand typically exceeds the mean significantly; see [14, 16, 17, 27, 31, 47, 61, 87, 96, 112, 121, 137, 144, 165]. The feature that variability is higher than one expects from the Poisson assumption is referred to as *overdispersion*.

Due to its inherent connection with the CLT, the square-root rule relies heavily on the premise that the variance of the number of jobs entering the system over a period of time is of the same order as the mean. Subsequently, when stochastic models do not take into account overdispersion, resulting performance estimates are likely to be overoptimistic. The system then ends up being underprovisioned, which possibly causes severe performance problems, particularly in critical loading. To deal with overdispersion, existing capacity sizing rules like the square-root rule need to be modified in order to incorporate a correct hedge against (increased) variability. Following our findings in section 3, the following adapted capacity allocation rule may be proposed:

$$(6.10) \qquad s = \mu_A + \beta\sigma_A,$$

where μ_A and σ_A are the mean and standard deviation of demand per period, respectively, and $\beta > 0$. This is similar to (1.1) in which the original variability hedge is replaced by an amount that is proportional to the square root of the variance of the arrival process. In [120], it is shown that this rule indeed leads to QED-type behavior in bulk-service queues as the system size grows. For the $M/M/s$ queue, this has been studied in [112] and [16], but more work in this area seems necessary.

6.6. Load Balancing. The analysis and design of load balancing schemes has attracted strong renewed interest in the last several years, mainly motivated by significant challenges involved in assigning tasks (e.g., file transfers, compute jobs, and database look-ups) to servers in large-scale data centers. Load balancing schemes provide an effective mechanism for improving QoS experienced by users while achieving high resource utilization levels, goals that are perfectly aligned with the QED regime. A distinguishing feature of such systems, however, is that there is no centralized queue, so that an incoming job should be forwarded instantaneously from the dispatcher to one of the servers. To achieve QED optimality, communication is needed between the dispatcher and servers. This can cause a prohibitive communication burden in large-scale deployments and asks for assessing load balancing schemes in terms of trade-offs between performance and implementation overhead.

A naive example of a load balancing scheme is Round Robin, a cyclic scheme that requires no communication, under which every sth job is assigned to the same server. For Poisson arrivals and service requirements equal to a constant, Round Robin achieves "perfect load balancing" among servers, and the delay distribution is the same as that of a single server serving every sth arrival of a Poisson input, or rather, Erlang input. In that case the delay distribution can be approximated by a Gaussian random walk, and all three structural properties are still justified. If deterministic job sizes are being replaced with general job sizes, the system still operates in heavy traffic, and the probability of delay converges to a value in the interval $(0, 1)$, but the mean delay will no longer be of the order $O(1/\sqrt{\lambda})$ but constant, so that the third structural QED property no longer holds.

A more involved example concerns the Join-the-Shortest-Queue (JSQ) scheme and several of its variations, such as versions where the shortest of $d = d(s)$ randomly chosen queues is selected [122, 154, 22, 109, 110, 53, 26, 39, 46]. In recent years several new results were discovered for JSQ$(d(s))$ multiserver systems that operate in the QED regime $(s - \lambda(s))/\sqrt{s} \to \beta > 0$ as $s \to \infty$. Eschenfeldt and Gamarnik [38] considered the JSQ scheme with $d(s) = s$ and introduced a properly centered and scaled version of the system occupancy processes. They showed that, as $s \to \infty$, the sequence of processes converges weakly to a system of coupled stochastic differential equations. Although this scaling limit differs from the diffusion limit obtained for the fully pooled $M/M/s$ queue, it shares similar favorable QED properties such as vanishing delay. The downside, however, is that a nominal implementation of JSQ comes with a large communication overhead. It was recently shown that for $d(s)$ such that $d(s)/(\sqrt{s}\log(s)) \to \infty$ as $s \to \infty$ the diffusion limit of JSQ$(d(s))$ corresponds to that for the JSQ policy [124, 123]. This indicates that the overhead of the JSQ policy can "almost" be reduced to $O(\sqrt{s}\log s)$ while retaining diffusion-level optimality. Many exciting problems in this area, which is still in its infancy, are still open; particular examples are scaling laws (in s) of the amount of memory used by the dispatcher, and the amount of communication overhead per packet. See [149] for a dedicated survey on this topic.

6.7. General Interarrival and Service Times. Now consider the $G/G/s$ queue, the natural extension of the $M/M/s$ queue to generally distributed interarrival times and service times. Establishing QED limits for the $G/G/s$ queue has led to a remarkable research effort, the majority of which having taken place over the last decade. When one moves beyond the exponential and deterministic assumptions, establishing QED behavior becomes mathematically more challenging, and most of the analysis of the $G/G/s$ queue in the QED regime has evolved around the characterization of the stochastic-process limit of the centered and scaled process, under various assumptions on the model primitives. We restrict our discussion to developments on the basic $G/G/s$ queue; a more extensive discussion, including work on abandonments, can be found in the surveys [127, 33]. Puhalskii and Reiman [131] analyze the multiclass queue with phase-type service times in the QED regime. Heavy-traffic limits for queues in which service time distributions are lattice-based and/or have finite support are studied by Mandelbaum and Momčilović [114] and Gamarnik and Momčilović [45]. The most general class of distributions is considered by Reed [133] and Puhalskii and Reed [130], who impose no assumptions on the service time distribution except for the existence of the first moment. Both of these papers focus on the queue length process. The paper by Reed [133] utilizes an ingenious connection with the infinite server queue, a connection which is developed further by Puhalskii and Reed [130], where results from modern empirical process theory (including the usage of outer measures to avoid measurability problems) are used to their full potential. Equally important steps forward concern the usage of measure-valued processes by Kang, Kaspi, and Ramanan [90, 89, 91]. Assuming minor additional regularity conditions on the service-time distribution (like a bounded density and sufficiently many finite moments), the paper [91] unravels the structure of the limit process that appears after scaling. A key insight from these works is that the limiting queue length process can be interpreted as a one-dimensional diffusion with a drift that depends on the entire history of the process, as opposed to the Halfin–Whitt diffusion that comes with exponential service times, where the drift depends on the current scaled queue length only. As a result, even after taking the limit, the resulting limit process for the $G/G/s$ queue still has a complicated steady-state distribution, and it is therefore not surprising that considerably less is known for the corresponding steady-state distribution of the $G/G/s$ queue in the QED regime. An exception is the work by Gamarnik and Goldberg [49, 44], who perform their analysis under the mild assumption that the service time distribution has finite $(2 + \varepsilon)$ moments and reveal suitable analogues of all three structural properties mentioned at the beginning of this section, and, in addition, explicit tail bounds for the distribution of the delay are developed. Without aiming to be exhaustive, we also point out the recent preprints focusing on infinite second moments [50] and mean waiting times [51]. This tutorial is not focused on heuristic approximations (for example, based on infinite-server models). A recent paper in this direction is [105].

Finally, we note that the impact of heavy tails is somewhat different than one might expect: the assumption of a service-time distribution with infinite variance is mainly made for technical purposes (as seen by the generality of the framework in [133, 130]), while an extension to interarrival times with infinite variance is changing the nature of the scaling procedure itself. For example, to achieve the balance property (Balance) in the $G/M/s$ queue where interarrival times have a power law tail with index $\alpha \in (1, 2)$, the proper dimensioning rule is $s_\lambda = \lambda + \beta \lambda^{(\alpha-1)^{-1}}$; see [127, 128, 136].

REFERENCES

[1] Z. Aksin, M. Armony, and V. Mehrotra, *The modern call center: A multi-disciplinary perspective on operations management research*, Production Oper. Manag., 16 (2007), pp. 665–688. (Cited on p. 408)

[2] D. Anick, D. Mitra, and M. Sondhi, *Stochastic theory of a data-handling system with multiple sources*, Bell System Tech. J., 61 (1982), pp. 1871–1894. (Cited on pp. 404, 418)

[3] J. Anselmi, Y. Lu, M. Sharma, and M. S. Squillante, *Improved approximations for the Erlang loss model*, Queueing Syst. Theory Appl., 63 (2009), pp. 217–239, https://doi.org/10.1007/s11134-009-9117-x. (Cited on p. 430)

[4] M. Armony, *Dynamic routing in large-scale service systems with heterogeneous servers*, Queueing Syst., 51 (2005), pp. 287–329. (Cited on p. 426)

[5] M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov, *On patient flow in hospitals: A data-based queueing-science perspective*, Stoch. Syst., 5 (2015), pp. 146–194. (Cited on pp. 404, 408)

[6] M. Armony and C. Maglaras, *On customer contact centers with a call-back option: Customer decisions, routing rules, and system design*, Oper. Res., 52 (2004), pp. 271–292. (Cited on p. 426)

[7] M. Armony, N. Shimkin, and W. Whitt, *The impact of delay announcements in many-server queues with abandonment*, Oper. Res., 57 (2009), pp. 66–81, https://doi.org/10.1287/opre.1080.0533. (Cited on p. 430)

[8] M. Armony and A. Ward, *Fair dynamic routing in large-scale heterogeneous-server systems*, Oper. Res., 58 (2010), pp. 624–637. (Cited on p. 426)

[9] S. Asmussen, *Applied Probability and Queues*, 2nd ed., Springer-Verlag, New York, 2003. (Cited on pp. 416, 425)

[10] R. Atar, *Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic*, Ann. Appl. Probab., 15 (2005), pp. 2606–2650. (Cited on pp. 426, 430)

[11] R. Atar, *A diffusion regime with nondegenerate slowdown*, Oper. Res., 60 (2012), pp. 490–500, https://doi.org/10.1287/opre.1110.1030. (Cited on pp. 418, 426)

[12] R. Atar and C. A., *A differential game for a multiclass queueing model in the moderate-deviation heavy-traffic regime*, Math. Oper. Res., 41 (2016), pp. 1354–1380. (Cited on p. 426)

[13] R. Atar, A. Mandelbaum, and M. Reiman, *Scheduling a multi class queue with many exponential servers: asymptotic optimality in heavy traffic*, Ann. Appl. Probab., 14 (2014), pp. 1084–1134. (Cited on p. 426)

[14] A. Avramidis, A. Deslauriers, and P. L'Ecuyer, *Rate-based daily arrival process models with application to call centers*, Manag. Sci., 50 (2004), pp. 893–908. (Cited on p. 431)

[15] O. Barnhoff-Nielsen and D. Cox, *Asymptotic Techniques for Use in Statistics*, Chapman & Hall, London, 1990. (Cited on p. 423)

[16] A. Bassamboo, R. Randhawa, and A. Zeevi, *Capacity sizing under parameter uncertainty: Safety staffing principles revisited*, Manag. Sci., 56 (2010), pp. 1668–1686. (Cited on pp. 431, 432)

[17] A. Bassamboo and A. Zeevi, *On a data-driven method for staffing large call centers*, Oper. Res., 57 (2009), pp. 714–726. (Cited on pp. 404, 431)

[18] R. Bhattaacharya and R. Rao, *Normal Approximations and Asymptotic Expansions*, Wiley, New York, 1976. (Cited on p. 423)

[19] J. Blanchet and P. Glynn, *Complete corrected diffusion approximations for the maximum of a random walk*, Ann. Appl. Probab., 16 (2006), pp. 951–983. (Cited on pp. 416, 425)

[20] C. Borgs, J. Chayes, S. Doroudi, M. Harchol-Balter, and K. Xu, *The optimal admission threshold in observable queues with state dependent pricing*, Probab. Engrg. Inform. Sci., 28 (2014), pp. 101–119. (Cited on p. 426)

[21] S. Borst, A. Mandelbaum, and M. Reiman, *Dimensioning large call centers*, Oper. Res., 52 (2004), pp. 17–34. (Cited on pp. 404, 419, 420, 424)

[22] M. Bramson, Y. Lu, and B. Prabhakar, *Asymptotic independence of queues under randomized load balancing*, Queueing Syst., 71 (2012), pp. 247–292, https://doi.org/10.1007/s11134-012-9311-0. (Cited on p. 432)

[23] A. Braverman and J. Dai, *Stein's method for steady-state diffusion approximations of $M/Ph/n + M$ systems*, Ann. Appl. Probab., 27 (2017), pp. 550–581. (Cited on p. 426)

[24] A. Braverman and J. Dai, *High Order Steady-State Diffusion Approximation of the Erlang-C System*, preprint, https://arxiv.org/abs/1602.02866, 2016. (Cited on p. 426)

[25] A. Braverman, J. Dai, and J. Feng, *Stein's method for steady-state diffusion approximations: An introduction through the Erlang-A and Erlang-C models*, Stoch. Syst., 6 (2016), pp. 301–366. (Cited on p. 426)

[26] G. Brightwell and M. Luczak, *The Supermarket Model with Arrival Rate Tending to One*, preprint, https://arxiv.org/abs/1201.5523, 2012. (Cited on p. 432)

[27] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, *Statistical analysis of a telephone call center: A queueing-science perspective*, J. Amer. Statist. Assoc., 100 (2005), pp. 36–50. (Cited on pp. 404, 408, 426, 431)

[28] S. Browne and W. Whitt, *Piecewise-linear diffusion processes*, in Advances in Queueing: Theory, Methods, and Open Problems, CRC Press, Boca Raton, FL, 1995, pp. 463–480. (Cited on p. 413)

[29] H. Bruneel and B. Kim, *Discrete-Time Models for Communication Systems Including ATM*, Kluwer Academic, Boston, 1993. (Cited on p. 414)

[30] J. Chang and Y. Peres, *Ladder heights, Gaussian random walks and the Riemann zeta function*, Ann. Probab., 25 (1997), pp. 787–802. (Cited on pp. 416, 417)

[31] B. Chen and S. G. Henderson, *Two issues in setting call center staffing levels*, Ann. Oper. Res., 108 (2001), pp. 175–192. (Cited on p. 431)

[32] J. Dai and S. He, *Customer abandonment in many-server queues*, Math. Oper. Res., 35 (2010), pp. 347–362. (Cited on p. 427)

[33] J. Dai and S. He, *Many-server queues with customer abandonment: A survey of diffusion and fluid approximations*, J. Syst. Sci. Syst. Engrg., 21 (2012), pp. 1–36. (Cited on pp. 408, 427, 433)

[34] B. D'Auria, *A short note on the monotonicity of the Erlang C formula in the Halfin-Whitt regime*, Queueing Syst., 71 (2012), pp. 469–472. (Cited on p. 424)

[35] M. Defraeye and I. van Niewenhuysen, *Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm*, Decision Support Syst., 54 (2013), pp. 1558–1567. (Cited on p. 422)

[36] S. Eick, W. Massey, and W. Whitt, *The physics of the $M_t/G/\infty$ queue*, Oper. Res., 41 (1993), pp. 731–742. (Cited on p. 421)

[37] A. Erlang, *Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges*, Elektroteknikeren, 13 (1917). (Cited on p. 404)

[38] P. Eschenfeldt and D. Gamarnik, *Join the shortest queue with many servers. The heavy traffic asymptotics*, Math. Oper. Res., 43 (2018), pp. 867–886. (Cited on p. 432)

[39] P. Eschenfeldt and D. Gamarnik, *Supermarket Queueing System in the Heavy Traffic Regime. Short Queue Dynamics*, preprint, https://arxiv.org/abs/1610.03522, 2016. (Cited on p. 432)

[40] Z. Feldman, A. Mandelbaum, W. Massey, and W. Whitt, *Staffing of time-varying queues to achieve time-stable performance*, Manag. Sci., 54 (2008), pp. 324–338. (Cited on p. 422)

[41] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, Wiley, 1971, https://books.google.nl/books?id=NFNQAAAAMAAJ. (Cited on p. 423)

[42] B. Fralix, C. Knessl, and J. van Leeuwaarden, *First passage times to congested states of many-server systems in the Halfin-Whitt regime*, Stoch. Models, 30 (2014), pp. 162–186. (Cited on pp. 413, 414)

[43] D. Gamarnik and D. Goldberg, *On the rate of convergence to stationarity of the $M/M/N$ queue in the Halfin-Whitt regime*, Ann. Appl. Probab., 23 (2013), pp. 1879–1912. (Cited on p. 414)

[44] D. Gamarnik and D. Goldberg, *Steady-state $GI/G/N$ queue in the Halfin-Whitt regime*, Ann. Appl. Probab., 23 (2013), pp. 2382–2419. (Cited on p. 433)

[45] D. Gamarnik and P. Momčilović, *Steady-state analysis of a multiserver queue in the Halfin-Whitt regime*, Adv. Appl. Probab., 40 (2008), pp. 548–577. (Cited on p. 433)

[46] D. Gamarnik, J. Tsitsiklis, and M. Zubeldia, *Delay, memory and messaging tradeoffs in distributed service systems*, in Proc. ACM SIGMETRICS, 2016, pp. 1–12, https://doi.org/10.1145/2896377.2901478. (Cited on p. 432)

[47] N. Gans, G. Koole, and A. Mandelbaum, *Telephone call centers: Tutorial, review, and research prospects*, Manufact. Serv. Oper. Manag., 5 (2003), pp. 79–141. (Cited on pp. 404, 408, 426, 431)

[48] O. Garnett, A. Mandelbaum, and M. Reiman, *Designing a call center with impatient customers*, Manufact. Serv. Oper. Manag., 4 (2002), pp. 208–227. (Cited on p. 427)

[49] D. Goldberg, *On the Steady-State Probability of Delay and Large Negative Deviations for the $GI/GI/n$ Queue in the Halfin-Whitt Regime*, preprint, https://arxiv.org/abs/1307.0241v2, 2013. (Cited on p. 433)

[50] D. GOLDBERG AND Y. LI, *Heavy-Tailed Queues in the Halfin-Whitt Regime*, preprint, https://arxiv.org/abs/1707.07775, 2017. (Cited on p. 433)

[51] D. GOLDBERG AND Y. LI, *Simple and Explicit Bounds for Multi-server Queues with Universal $1/(1 - \rho)$ Scaling*, preprint, https://arxiv.org/abs/1706.04628, 2017. (Cited on pp. 426, 433)

[52] R. GOPALAKRISHNAN, S. DOROUDI, A. R. WARD, AND A. WIERMAN, *Routing and staffing when servers are strategic*, Oper. Res., 64 (2016), pp. 1033–1050, https://doi.org/10.1287/opre.2016.1506. (Cited on p. 429)

[53] C. GRAHAM, *Functional central limit theorems for a large network in which customers join the shortest of several queues*, Probab. Theory Related Fields, 131 (2005), pp. 97–120, https://doi.org/10.1007/s00440-004-0372-9. (Cited on p. 432)

[54] L. GREEN AND P. KOLESAR, *The pointwise stationary approximation for queues with non-stationary arrivals*, Manag. Sci., 37 (1991), pp. 84–97. (Cited on p. 421)

[55] L. GREEN, P. KOLESAR, AND W. WHITT, *Coping with time-varying demand when setting staffing requirements for a service system*, Production Oper. Manag., 16 (2007), pp. 13–39. (Cited on p. 422)

[56] L. GREEN, S. SAVIN, AND M. MURRAY, *Providing timely access to care: What is the right patient panel size?*, Joint Commission J. Quality Patient Safety, 33 (2007), pp. 211–218. (Cited on p. 404)

[57] V. GUPTA, M. HARCHOL-BALTER, K. SIGMAN, AND W. WHITT, *Analysis of join-the-shortest-queue routing for web server farms*, Performance Eval., 46 (2007), pp. 1062–1081. (Cited on p. 404)

[58] I. GURVICH, *Diffusion models and steady-state approximations for exponentially ergodic Markovian queues*, Ann. Appl. Probab., 24 (2014), pp. 2527–2559. (Cited on p. 426)

[59] I. GURVICH, M. ARMONY, AND A. MANDELBAUM, *Service-level differentiation in call centers with fully flexible servers*, Manag. Sci., 54 (2008), pp. 279–294. (Cited on p. 426)

[60] I. GURVICH, J. HUANG, AND A. MANDELBAUM, *Excursion-based universal approximations for the Erlang-A queue in steady-state*, Math. Oper. Res., 39 (2014), pp. 325–373. (Cited on p. 426)

[61] I. GURVICH, J. LUEDTKE, AND T. TEZCAN, *Staffing call-centers with uncertain demand forecasts: A chance-constrained optimization approach*, Manag. Sci., 56 (2010), pp. 1093–1115. (Cited on p. 431)

[62] I. GURVICH AND W. WHITT, *Queue-and-idleness-ratio controls in many-server service systems*, Math. Oper. Res., 34 (2009), https://doi.org/10.1287/moor.1080.0366. (Cited on pp. 426, 430)

[63] S. HALFIN AND W. WHITT, *Heavy-traffic limits for queues with many exponential servers*, Oper. Res., 29 (1981), pp. 567–588. (Cited on pp. 407, 410, 412, 413, 419)

[64] J. HARRISON, *Brownian Motion and Stochastic Flow Systems*, John Wiley & Sons, 1985. (Cited on p. 416)

[65] J. HARRISON AND A. ZEEVI, *Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime*, Oper. Res., 52 (2004), pp. 243–257. (Cited on p. 426)

[66] R. HASSIN, *Rational Queueing*, Chapman & Hall, 2016. (Cited on p. 429)

[67] B. HE, Y. LIU, AND W. WHITT, *Staffing a service system with non-Poisson non-stationary arrivals*, Probab. Engrg. Inform. Sci., 30 (2016), pp. 593–621. (Cited on p. 422)

[68] J. HUANG AND I. GURVICH, *Beyond Heavy-Traffic Regimes: Universal Bounds and Controls for the Single-Server Queue*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2784752, 2016. (Cited on p. 426)

[69] J. HUANG, A. MANDELBAUM, H. ZHANG, AND J. ZHANG, *Refined models for efficiency-driven queues with applications to delay announcements and staffing*, Oper. Res., 65 (2017), pp. 1380–1397, https://doi.org/10.1287/opre.2017.1619. (Cited on p. 430)

[70] P. J. HUNT AND F. P. KELLY, *On critically loaded loss networks*, Adv. in Appl. Probab., 21 (1989), pp. 831–841, https://doi.org/10.2307/1427769. (Cited on p. 430)

[71] H. HWANG, *Asymptotic estimates of elementary probability distributions*, Stud. Appl. Math., 4 (1997), pp. 339–417. (Cited on p. 423)

[72] R. IBRAHIM, *Sharing delay information in service systems: a literature survey*, Queueing Syst., 89 (2018), pp. 49–79, https://doi.org/10.1007/s11134-018-9577-y. (Cited on p. 430)

[73] D. IGLEHART, *Weak convergence in applied probability*, Stoch. Process. Appl., 2 (1974), pp. 211–241. (Cited on p. 419)

[74] D. JAGERMAN, *Some properties of the Erlang loss function*, Bell System Tech. J., 53 (1974), pp. 525–551, https://doi.org/10.1002/j.1538-7305.1974.tb02756.x. (Cited on p. 410)

[75] A. A. JAGERS AND E. A. VAN DOORN, *On the continued Erlang loss function*, Oper. Res. Lett., 5 (1986), pp. 43–46. (Cited on p. 423)

[76] A. Janssen and J. van Leeuwaarden, *Analytic computation schemes for the discrete-time bulk service queue*, Queueing Syst., 50 (2005), pp. 141–163. (Cited on p. 418)

[77] A. Janssen and J. van Leeuwaarden, *On Lerch's transcendent and the Gaussian random walk*, Ann. Appl. Probab., 17 (2006), pp. 421–439. (Cited on pp. 416, 417)

[78] A. Janssen and J. van Leeuwaarden, *Back to the roots of the M/D/s queue and the works of Erlang, Crommelin, and Pollaczek*, Statist. Neerlandica, 62 (2008), pp. 299–313. (Cited on p. 418)

[79] A. Janssen, J. van Leeuwaarden, and J. Sanders, *Scaled control in the QED regime*, Performance Eval., 70 (2013), pp. 750–769. (Cited on p. 426)

[80] A. Janssen, J. van Leeuwaarden, and A. Zwart, *Corrected asymptotics for a multi-server queue in the Halfin-Whitt regime*, Queueing Syst., 58 (2008), pp. 261–301. (Cited on p. 415)

[81] A. Janssen, J. van Leeuwaarden, and A. Zwart, *Refining square-root safety staffing by expanding Erlang-C*, Oper. Res., 59 (2011), pp. 1512–1522. (Cited on pp. 420, 424, 425)

[82] A. Janssen, J. van Leeuwaarden, and B. Zwart, *Gaussian expansions and bounds for the Poisson distribution applied to the Erlang-B formula*, Adv. in Appl. Probab., 40 (2008), pp. 122–143. (Cited on pp. 410, 411, 423)

[83] P. Jelenkovic, A. Mandelbaum, and P. Momčilović, *Heavy traffic limits for queues with many deterministic servers*, Queueing Syst., 47 (2004), pp. 53–69. (Cited on p. 416)

[84] O. Jennings, A. Mandelbaum, W. Massey, and W. Whitt, *Server staffing to meet time-varying demand*, Manag. Sci., 42 (1996), pp. 1383–1394. (Cited on pp. 421, 422)

[85] O. Jennings and J. Reed, *An overloaded multiclass FIFO queue with abandonments*, Oper. Res., 60 (2012), pp. 1282–1295. (Cited on p. 427)

[86] N. Johnson, S. Kotz, and A. Kemp, *Univariate Discrete Distributions*, 2nd ed., Wiley, New York, 1992. (Cited on p. 423)

[87] G. Jongbloed and G. Koole, *Managing uncertainty in call centres using Poisson mixtures*, Appl. Stoch. Models Bus. Ind., 17 (2001), pp. 307–318. (Cited on p. 431)

[88] K. Jung, Y. Lu, D. Shah, M. Sharma, and M. Squillante, *Revisiting stochastic loss networks: Structures and algorithms*, SIGMETRICS Perform. Eval. Rev., 36 (2008), pp. 407–418, https://doi.org/10.1145/1384529.1375503. (Cited on p. 430)

[89] W. Kang and K. Ramanan, *Asymptotic approximations for stationary distributions of many-server queues with abandonment*, Ann. Appl. Probab., 22 (2012), pp. 477–521. (Cited on pp. 427, 433)

[90] H. Kaspi and K. Ramanan, *Law of large numbers limits for many-server queues*, Ann. Appl. Probab., 21 (2011), pp. 33–114. (Cited on p. 433)

[91] H. Kaspi and K. Ramanan, *SPDE limits of many-server queues*, Ann. Appl. Probab., 23 (2013), pp. 145–229. (Cited on p. 433)

[92] F. Kelly, *Stochastic models of computer communication systems*, J. Roy. Statist. Soc. Ser. B, 47 (1985), pp. 379–395. (Cited on p. 404)

[93] F. Kelly, *Loss networks*, Ann. Appl. Probab., 1 (1991), pp. 319–378. (Cited on p. 430)

[94] F. Kelly and E. Yudovina, *Stochastic Networks*, Cambridge University Press, 2014. (Cited on p. 430)

[95] P. Khudyakov, *Designing a Call Center with an IVR (Interactive Voice Response)*, Master's thesis, Technion, 2006. (Cited on pp. 404, 431)

[96] S.-H. Kim and W. Whitt, *Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes?*, Manufact. Serv. Oper. Manag., 16 (2014), pp. 464–480. (Cited on p. 431)

[97] Y. Koçaga, M. Armony, and A. Ward, *Staffing call centers with uncertain arrival rate and co-sourcing*, Production Oper. Manag., 24 (2015), pp. 1101–1117. (Cited on p. 426)

[98] P. Kolesar and L. Green, *Insights on service system design form a normal approximation to Erlang's delay formula*, Production Oper. Manag., 7 (1998), pp. 282–293. (Cited on p. 409)

[99] S. Kou and S. Kou, *A diffusion model for growth stocks*, Math. Oper. Res., 29 (2004), pp. 191–212. (Cited on p. 419)

[100] S. Kumar and R. Randhawa, *Exploiting market size in service systems*, Manufact. Serv. Oper. Manag., 12 (2010), pp. 511–526. (Cited on p. 429)

[101] D. Lindley, *The theory of queues with a single server*, Proc. Cambridge Philos. Soc., 48 (1952), pp. 277–289. (Cited on p. 414)

[102] Y. Liu and W. Whitt, *Stabilizing customer abandonment in many-server queues with time-varying arrivals*, Oper. Res., 60 (2012), pp. 1551–1564. (Cited on p. 422)

[103] Y. Liu and W. Whitt, *Stabilizing performance in nonstationary queues with non-Poisson arrivals*, Probab. Engrg. Inform. Sci., 28 (2014), pp. 419–449. (Cited on p. 422)

[104] Y. LIU AND W. WHITT, *Stabilizing performance in a service system with time-varying arrivals and customer feedback*, European J. Oper. Res., 256 (2017), pp. 473–486. (Cited on p. 422)

[105] Y. LIU, W. WHITT, AND Y. YU, *Approximations for heavily loaded $G/GI/n + GI$ queues*, Naval Res. Logist., 63 (2016), pp. 187–217, https://doi.org/10.1002/nav.21688. (Cited on p. 433)

[106] H. LU AND G. PANG, *Gaussian limits for a fork-join network with non-exchangeable synchronization in heavy traffic*, Math. Oper. Res., 41 (2015), pp. 560–595. (Cited on p. 430)

[107] H. LU AND G. PANG, *Heavy-traffic limits for a fork-join network in the Halfin-Whitt regime*, Stoch. Syst., 6 (2016), pp. 519–600. (Cited on p. 430)

[108] H. LU AND G. PANG, *Heavy-traffic limits for an infinite-server fork–join queueing system with dependent and disruptive services*, Queueing Syst., 85 (2017), pp. 67–115. (Cited on p. 430)

[109] M. LUCZAK AND C. MCDIARMID, *On the maximum queue length in the supermarket model*, Ann. Probab., 34 (2006), pp. 493–527, https://doi.org/10.1214/00911790500000710. (Cited on p. 432)

[110] M. LUCZAK AND J. NORRIS, *Strong approximation for the supermarket model*, Ann. Appl. Probab., 15 (2005), pp. 2038–2061, https://doi.org/10.1214/105051605000000368. (Cited on p. 432)

[111] C. MAGLARAS AND A. ZEEVI, *Diffusion approximations for a multiclass Markovian service system with "guaranteed" and "best-effort" service levels*, Math. Oper. Res., 29 (2004), pp. 786–813. (Cited on pp. 414, 426)

[112] S. MAMAN, *Uncertainty in the Demand for Service: The Case of Call Centers and Emergency Departments*, Master's thesis, Technion, 2009. (Cited on pp. 431, 432)

[113] A. MANDELBAUM, W. A. MASSEY, AND M. I. REIMAN, *Strong approximations for Markovian service networks*, Queueing Syst., 30 (1998), pp. 149–201. (Cited on p. 430)

[114] A. MANDELBAUM AND P. MOMČILOVIĆ, *Queues with many servers: The virtual waiting-time process in the QED regime*, Math. Oper. Res., 33 (2008), pp. 561–586. (Cited on p. 433)

[115] A. MANDELBAUM AND P. MOMČILOVIĆ, *Queues with many servers and impatient customers*, Math. Oper. Res., 37 (2012), pp. 41–65. (Cited on p. 427)

[116] A. MANDELBAUM, P. MOMČILOVIĆ, AND Y. TSEYTLIN, *On fair routing from emergency departments to hospital wards*, Manag. Sci., 58 (2012), pp. 1273–1291. (Cited on p. 426)

[117] A. MANDELBAUM AND S. ZELTYN, *Staffing many-server queues with impatient customers: Constraint satisfaction in call centers*, Oper. Res., 57 (2009), pp. 1189–1205. (Cited on p. 427)

[118] W. MASSEY AND R. WALLACE, *An Asymptotically Optimal Design of the $M/M/c/k$ Queue*, Unpublished report, 2004. (Cited on p. 428)

[119] W. MASSEY AND W. WHITT, *An analysis of the modified offered-load approximation for the non-stationary Erlang loss model*, Ann. Appl. Probab., 4 (1994), pp. 1145–1160. (Cited on p. 422)

[120] B. MATHIJSEN, A. JANSSEN, J. VAN LEEUWAARDEN, AND A. ZWART, *Robust Heavy-Traffic Approximations for Service Systems Facing Overdispersed Demand*, preprint, https://arxiv.org/abs/1512.05581, 2016. (Cited on p. 432)

[121] V. MEHROTRA, O. OZLÜK, AND R. SALTZMANN, *Intelligent procedures for intra-day updating of call center agent schedules*, Production Oper. Manag., 19 (2010), pp. 353–367. (Cited on p. 431)

[122] M. MITZENMACHER, *The power of two choices in randomized load balancing*, IEEE Trans. Parallel Distrib. Syst., 12 (2001), pp. 1094–1104, https://doi.org/10.1109/71.963420. (Cited on p. 432)

[123] D. MUKHERJEE, S. BORST, J. VAN LEEUWAARDEN, AND P. WHITING, *Asymptotic Optimality of Power-of-d Load Balancing in Large-Scale Systems*, preprint, https://arxiv.org/abs/1612.00722, 2016. (Cited on p. 432)

[124] D. MUKHERJEE, S. BORST, J. VAN LEEUWAARDEN, AND P. WHITING, *Universality of power-of-d load balancing in many-server systems*, Stoch. Syst., 8 (2018), pp. 265–292. (Cited on p. 432)

[125] J. NAIR, A. WIERMAN, AND A. ZWART, *Provisioning of large-scale systems: The interplay between network effects and strategic behavior in the user base*, Manag. Sci., 62 (2016), pp. 1830–1841. (Cited on p. 429)

[126] R. PALM, *Research on Telephone Traffic Carried by Full Availability Groups*, Tele, 1957. (Cited on pp. 404, 426, 427)

[127] G. PANG, R. TALREJA, AND W. WHITT, *Martingale proofs of many-server heavy-traffic limits for Markovian queues*, Probab. Surv., 4 (2007), pp. 193–267. (Cited on pp. 408, 433)

[128] G. PANG AND W. WHITT, *Heavy-traffic limits for many-server queues with service interruptions*, Queueing Syst., 61 (2009), pp. 167–202, https://doi.org/10.1007/s11134-009-9104-2. (Cited on p. 433)

[129] V. PETROV, *Limit Theorems of Probability Theory*, Clarendon Press, Oxford, 1995. (Cited on p. 423)

[130] A. PUHALSKII AND J. REED, *On many-server queues in heavy traffic*, Ann. Appl. Probab., 20 (2010), pp. 129–195. (Cited on p. 433)

[131] A. PUHALSKII AND M. REIMAN, *The multiclass GI/Ph/N queue in the Halfin-Whitt regime*, Adv. in Appl. Probab., 32 (2000), pp. 564–595. (Cited on p. 433)

[132] R. RANDHAWA, *Optimality gap of asymptotically derived prescriptions in queueing systems: o(1)-optimality*, Queueing Syst., 83 (2016), pp. 131–155. (Cited on p. 425)

[133] J. REED, *The G/GI/N queue in the Halfin-Whitt regime*, Ann. Appl. Probab., 19 (2009), pp. 2211–2269. (Cited on p. 433)

[134] J. REED AND T. TEZCAN, *Hazard rate scaling for the GI/M/n + GI queue*, Oper. Res., 70 (2012), pp. 1–34. (Cited on p. 427)

[135] J. REED AND A. ZWART, *Limit theorems for Markovian bandwidth sharing networks with rate constraints*, Oper. Res., 62 (2014), pp. 1453–1466. (Cited on p. 431)

[136] J. REED AND B. ZWART, *A piecewise linear stochastic differential equation driven by a Lévy process*, J. Appl. Probab., 48A (2011), pp. 109–119, https://doi.org/10.1239/jap/1318940459. (Cited on p. 433)

[137] T. ROBBINS, D. MEDEIROS, AND T. HARRISON, *Does the Erlang C model fit in real call centers?*, in Proceedings of the 2010 Winter Simulation Conference, 2010, pp. 2853–2864. (Cited on p. 431)

[138] S. ROSS, *Stochastic Processes*, John Wiley & Sons, 1996. (Cited on p. 414)

[139] J. SANDERS, S. BORST, A. JANSSEN, AND J. VAN LEEUWAARDEN, *Optimal admission control for many-server systems with QED-driven revenues*, Stoch. Syst., 7 (2017), pp. 315–341. (Cited on pp. 420, 426)

[140] J. SANDERS, S. BORST, A. JANSSEN, AND J. VAN LEEUWAARDEN, *Optimality gaps in asymptotic dimensioning of many-server systems*, Oper. Res. Lett., 44 (2016), pp. 369–365. (Cited on p. 425)

[141] D. SIEGMUND, *Corrected Diffusion Approximations in Certain Random Walk Problems*, Tech. report, Stanford University, 1978. (Cited on p. 416)

[142] D. SIEGMUND, *Sequential Analysis*, Springer Ser. Statist., Springer-Verlag, New York, 1985. (Cited on p. 425)

[143] F. SPITZER, *Principles of Random Walk*, D. van Nostrand, 1964. (Cited on p. 416)

[144] S. STECKLEY, S. HENDERSON, AND V. MEHROTRA, *Forecast errors in service systems*, Probab. Engrg. Inform. Sci., 23 (2009), pp. 305–332. (Cited on p. 431)

[145] A. STOLYAR AND T. TEZCAN, *Control of systems with flexible multi-server pools: A shadow routing approach*, Queueing Syst., 66 (2010), pp. 1–51. (Cited on p. 426)

[146] C. STONE, *Limit theorems for random walks, birth and death processes, and diffusion processes*, Illinois J. Math., 7 (1963), pp. 638–660, http://projecteuclid.org/euclid.ijm/1255645101. (Cited on p. 419)

[147] J. TAN, H. FENG, X. MENG, AND L. ZHANG, *Heavy-traffic analysis of cloud provisioning*, in Proceedings of the 24th International Teletraffic Congress, 2012, pp. 1–8. (Cited on pp. 404, 431)

[148] T. TEZCAN AND J. DAI, *Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic*, Oper. Res., 58 (2010), pp. 94–110. (Cited on p. 426)

[149] M. VAN DER BOOR, S. BORST, J. VAN LEEUWAARDEN, AND D. MUKHERJEE, *Scalable load balancing in networked systems: Universality properties and stochastic coupling methods*, in Proc. International Congress of Mathematicians (ICM), 2018; preprint available from https://arxiv.org/abs/1712.08555 (2017). (Cited on p. 432)

[150] E. VAN DOORN, *Conditions for exponential ergodicity and bounds for the decay parameter of a birth-death process*, Adv. in Appl. Probab., 17 (1985), pp. 514–530. (Cited on p. 414)

[151] J. VAN LEEUWAARDEN, *Queueing Models for Cable Access Networks*, Ph.D. thesis, Eindhoven University of Technology, 2005. (Cited on pp. 404, 414)

[152] J. VAN LEEUWAARDEN AND C. KNESSL, *Transient behavior of the Halfin-Whitt diffusion*, Stochastic Process. Appl., 21 (2011), pp. 1524–1545. (Cited on p. 414)

[153] J. VAN LEEUWAARDEN AND C. KNESSL, *Spectral gap of the Erlang-A model in the Halfin-Whitt regime*, Stoch. Syst., 2 (2012), pp. 149–207. (Cited on pp. 413, 414)

[154] N. VVEDENSKAYA, R. DOBRUSHIN, AND F. KARPELEVICH, *Queueing system with selection of the shortest of two queues: An asymptotic approach*, Problemy Peredachi Informatsii, 32 (1996), pp. 20–34. (Cited on p. 432)

[155] A. WARD, *Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO single class models.*, Surv. Oper. Res. Manag. Sci., 17 (2012), pp. 1–14. (Cited on pp. 408, 427)

[156] W. WHITT, *The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase*, Manag. Sci., 37 (1991), pp. 307–314. (Cited on p. 421)

[157] W. WHITT, *Dynamic staffing in a telephone call center aiming to immediately answer all calls*, Oper. Res. Lett., 24 (1999), pp. 205–212. (Cited on p. 404)

[158] W. WHITT, *The Erlang B and C formulas: Problems and solutions*, Lecture notes, http://www.columbia.edu/~ww2040/ErlangBandCFormulas.pdf, 2002. (Cited on p. 410)

[159] W. WHITT, *Staffing a call center with uncertain arrival rate and absenteeism*, Production Oper. Manag., 15 (2006), pp. 88–102. (Cited on pp. 427, 431)

[160] W. WHITT, *What you should know about queueing models to set staffing requirements in service systems*, Naval Res. Logist., 54 (2007), pp. 476–484. (Cited on p. 422)

[161] W. WHITT, *Offered load analysis for staffing*, Manufact. Serv. Oper. Manag., 15 (2013), pp. 166–169. (Cited on p. 422)

[162] W. WHITT, *Time-varying queues*, Queueing Models Serv. Manag., 1 (2018), pp. 79–1164. (Cited on pp. 408, 422)

[163] R. WOLFF, *Poisson arrivals see time averages*, Oper. Res., 30 (1982), pp. 223–231. (Cited on p. 409)

[164] G. YOM-TOV, *Queues in Hospitals: Queueing Networks with Re-entering Customers in the QED Regime*, Ph.D. thesis, Technion, 2010. (Cited on pp. 404, 431)

[165] J. ZAN, *Staffing Service Centers under Arrival-Rate Uncertainty*, Ph.D. thesis, University of Texas, 2012. (Cited on p. 431)

[166] S. ZELTYN AND A. MANDELBAUM, *Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue*, Queueing Syst., 51 (2005), pp. 361–402. (Cited on pp. 404, 407, 427)

[167] B. ZHANG, J. VAN LEEUWAARDEN, AND A. ZWART, *Staffing call centers with impatient customers: Refinements to many-server asymptotics*, Oper. Res., 60 (2012), pp. 461–474. (Cited on pp. 420, 425)

[168] J. ZHANG, *Fluid models of many-server queues with abandonment*, Queueing Syst., 73 (2013), pp. 147–193, https://doi.org/10.1007/s11134-012-9307-9; preprint available from https://arxiv.org/abs/0909.1671 (2009). (Cited on p. 427)