Queueing Syst (2018) 89:3–14
https://doi.org/10.1007/s11134-018-9580-3

A broad view of queueing theory through one issue

Ward Whitt[1]

Published online: 12 April 2018

Abstract This is an overview and appreciation of the contributions to this special issue.

Keywords Service systems · Sharing delay information · Heavy traffic · Time-varying arrival rates · Closure approximations · Reflected Le'vy processes

Mathematics Subject Classification 60K25 · 90B22

1 Introduction

I thank all involved with the conference and this special issue (SI). I was asked to provide some commentary to put queueing theory in perspective, as seen from this stage in my career, which I do by first discussing the papers in this SI and then by providing a few personal reflections.

2 The lifeblood of queueing theory

We all are captivated by the way probability can help us understand uncertainty. Thus, we are naturally attracted to queueing theory because it is a subfield of probability theory within mathematics. At the same time, queueing theory is also subfield of several more applied domains: of operations research within engineering, of operations management within business, and of performance analysis within computer science

✉ Ward Whitt
ww2040@columbia.edu

[1] Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA

and electrical engineering. Indeed, as emphasized by Uma Prabhu in his founding editorial in 1986 [42], the lifeblood of queueing theory lies in its applications.

Queueing theory has its origins in the work by A. K. Erlang on early telecommunication systems [4,12,35] and keeps being revitalized by new technology, including production, distribution, and computing systems. Today queueing theory is being revitalized by new applications in service systems, as illustrated by the papers by Zychlinski et al. [51] and Ibrahim [27] in this SI.

2.1 Improving healthcare: time-varying tandem queues with blocking

Zychlinski et al. [51] study tandem networks of many-server queues with time-varying arrivals. The models have both telecommunication blocking (loss upon arrival if there is no capacity) and production blocking (blocking after service, so prevented from leaving, when the next facility has no room), e.g., see [16,48]. The authors develop fluid models for several of these networks and justify them by establishing many-server heavy-traffic (MSHT) functional strong law of large numbers (FSLLNs). They show that the fluid models can be effectively analyzed, exploiting ordinary differential equations with discontinuous right-hand side [18].

The paper [51] has a compelling motivation to analyze the flow of elderly patients into beds in hospitals and then on to less costly beds in other geriatric institutions, as the authors describe in their companion paper [50]. The blocking after service (bed blocking in hospitals) occurs when there are no beds available in the geriatric institutions, forcing the patients to remain in the more expensive hospital beds, preventing new admissions to the hospital. Moreover, in [50] they validate their model and analysis by making comparisons with both 2 years of patient flow data and computer simulation experiments. Thus, [50] is part of the recent empirical emphasis within operations management. Avi has helped lead the way by developing the Service Enterprise Engineering SEELab at the Technion, where data are collected and data analysis tools are developed.

The basic Markov stochastic model studied in [51] has two stations, having N_1 and N_2 servers. There is a finite waiting room of capacity H before the first station and no waiting space in between the two stations. The content is represented by a vector $(X_1(t), X_2(t))$, where $X_1(t)$ is the number of customers at station 1 that have not completed service at station 1, while $X_2(t)$ is the number of customers in the entire system that have completed service at station 1, but not at station 2. The number of blocked servers at station 1 is then $B(t) \equiv (X_2(t) - N_2)^+$, while the number of unblocked servers at station 1 is $U(t) \equiv X_1(t) \wedge (N_1 - B(t))$. Customers arrive according to a nonhomogeneous Poisson process with rate $\lambda(t)$; the total service rate at station 1 is $\mu_1 U(t)$ and at station 2 is $\mu_2(X_2(t) \wedge N_2)$; a customer completing service at station 1 continues to station 2 with probability p and departs otherwise.

The authors show that the vector stochastic process $(X_1(t), X_2(t))$ can be analyzed conveniently (so that reflection occurs on the axes) by letting $R_1(t)$ represent the nonutilized capacity at station 1 (the blocked and idle servers plus the available waiting room), while $R_2(t)$ represents the available space in the entire system.

2.2 Sharing delay information in service systems

Ibrahim [27] surveys the rapidly growing literature on sharing delay information in service systems. This important direction is emphasized in Sections 3, 5 and 6 of the 2007 survey on call centers by Aksin et al. [1]. Recent research is responding to the complex costs and benefits of sharing delay information, involving economic and behavioral issues as well as probabilistic and statistical ones.

To see the growing interest in economic issues in queues, we observe that 320 of the 872 citations in Google Scholar to the classic 1969 paper in *Econometrica* by Naor [41] on "the regulation of queue size by levying tolls," have appeared in the last four years. Of course, this area has received attention, as can be seen from Hassin and Haviv [26] (updated in [25]) and Stidham [47], but there is a new level of excitement. Of particular interest is the number of empirical papers.

The fascinating new questions about sharing delay information can be seen in the following quote from a recent article in the New York Times [19]:

> LEONIA, N.J. - It is bumper to bumper as far as the eye can see, the kind of soul-sucking traffic jam that afflicts highways the way bad food afflicts rest stops. Suddenly, a path to hope presents itself: An alternate route, your smartphone suggests, can save time. Next thing you know, you're headed down an exit ramp, blithely following directions into the residential streets of some unsuspecting town, along with a slew of other frustrated motorists.
> With services like Google Maps, Waze and Apple Maps suggesting shortcuts for commuters through the hilly streets of Leonia, N.J., the borough has decided to fight back against congestion that its leaders say have reached crisis proportions. In mid-January, the borough's police force will close 60 streets to all drivers aside from residents and people employed in the borough during the morning and evening rush periods...

3 The great reach of the CLT

Probability theory can work wonders explaining complex phenomena associated with randomness. Indeed, much insight is provided by the central limit theorem (CLT). Partial sums of random variables have a complex exact distribution, even when they are independent and identically distributed, but in great generality they may be approximately Gaussian, a distribution that depends only on the mean and variance. Variants of the CLT also provide answers and insights into queueing problems. Gaussian approximations for complex many-server queueing systems are developed and exploited here by Aras et al. [2] and Massey and Pender [40].

3.1 Overloaded queues with abandonment and non-exponential service times

One way complexity can arise in queueing is by having a non-Markovian model. Aras et al. [2] show that tractable Gaussian approximations can be developed for the

overloaded general many-server $G/GI/n + GI$ model with customer abandonment (the $+GI$) by establishing a MSHT functional central limit theorem (FCLT).

First, service systems provide a good reason for focusing on many-server queues with customer abandonment, as explained by [21]. With abandonment, there is no issue about stability, so that the system might well be overloaded. In addition, data analysis has shown that service time and patience distributions are often not nearly exponential in practice [13], which is important because the system performance can depend significantly on the underlying service time and patience distributions beyond their means.

With or without abandonment, many-server queues with a non-exponential service distribution are hard to analyze, because the queue-length process is not Markov. Ways to address this problem have emerged over the last twenty years, including two-parameter and measure-valued processes, as in [2,31] and references therein.

The main FCLT in Theorem 4.2 of [2] exploits a stochastic integral with respect to a Gaussian process and the CLT for sums of equilibrium renewal processes. They identify the separate contributions of the arrival, service, and abandonment processes. Theorem 4.3 gives explicit expressions for the covariance functions of the components of the Gaussian limit process.

3.2 The dynamic-rate Erlang-A queue

Another way complexity can arise in queueing, even for a Markov model, is to have a time-varying arrival rate. A MSHT functional weak law of large numbers (FWLLN) and FCLT for many-server queues with a time-varying arrival rate were first established by Mandelbaum, Massey and Reiman [38], and subsequently analyzed elegantly in [44], but in general the limit is complicated, except for the special case when the system is almost never critically loaded, which occurs if the system is always underloaded, always overloaded or alternates between overloaded and underloaded regimes, instantaneously passing through critical loading in each transition.

However, recent research has shown that the exceptional special case can be exploited, because then the limit is a time-varying Gaussian approximation. Unfortunately, the accuracy of that Gaussian approximation can degrade significantly when the system is nearly critically loaded, the common design goal. In this SI, Massey and Pender [40] review and extend their closure approximations for improving the quality of the resulting Gaussian approximation for the $M_t/M/c_t + M$ (dynamic-rate Erlang-A) model, first developed in [39].

For a time-varying birth–death process with birth rates $\lambda_k(t)$ and death rates $\mu_k(t)$, the time-varying probability mass function (pmf) $p_k(t) \equiv P(X(t) = k)$ satisfies the forward Kolmogorov differential equations; i.e., with $\dot{p}_k(t)$ denoting the derivative with respect to time, the system of ordinary differential equations (ODE's) is

$$\dot{p}_k(t) = -(\lambda_k(t) + \mu_k(t))p_k(t) + \lambda_{k-1}(t)p_{k-1}(t) + \mu_{k+1}(t)p_{k+1}(t), \quad k \geq 1, \quad \text{and}$$
$$\dot{p}_0(t) = -\lambda_0(t)p_0(t) + \mu_1(t)p_1(t). \tag{3.1}$$

A closure approximation is a much smaller system of ODE's for a functional of the process, like a moment, which is obtained by assuming that $X(t)$ has a distribution

belonging to a convenient parametric family for all t. For a real-valued function f of the state, we define

$$m_f(t) \equiv E[f(X(t))] = p(t)f \equiv \sum_k p_k(t) f(k) \tag{3.2}$$

and examine the resulting ODE for $\dot{m}_f(t)$. If we let $f(k) = k^p$, $k \geq 0$, then $m_f(t) = E[X(t)^p]$.

For highly structured queueing models such as the $M_t/M/c_t + M$ model with service rate μ and abandonment rate θ for each customer, the functional Kolmogorov equations for the first few moments take a relatively tractable form. For the time-varying (TV) mean $m(t)$, we have $m_f(t)$ in (3.2) for $f(k) = k$, $k \geq 0$, so that

$$\dot{m}(t) = \lambda(t) - \mu E[(X(t) \wedge c_t)] - \theta E[(X(t) - c_t)^+]. \tag{3.3}$$

By appropriately combining the ODE's for the first two moments, we get the corresponding ODE for the TV variance, denoted by $v(t)$,

$$\begin{aligned} \dot{v}(t) = \lambda + \mu E[X(t) \wedge c_t] + \theta E[(X(t) - c_t)^+] \\ -2(\mu Cov(X(t), X(t) \wedge c_t) + \theta Cov(X(t), (X(t) - c_t)^+)). \end{aligned} \tag{3.4}$$

The first Gaussian closure approximation is obtained by assuming, as an approximation, that $X(t)$ has a Gaussian distribution, which of course is determined by its first two moments. Equivalently, $X(t)$ is distributed as

$$X(t) \approx m(t) + N(0, 1)\sqrt{v(t)} \quad \text{for all} \quad t, \tag{3.5}$$

where $N(0, 1)$ is a standard normal random variable.

Massey and Pender [40] also develop higher-order refinements based on the first k moments for $k > 2$. To do so, they exploiting the Hilbert space of Hermite orthogonal polynomials and a generalization of Stein's lemma [46]. The general framework represents the distribution at each time t as a polynomial function of a Gaussian random variable. Moreover, this general approach can be applied in other contexts.

Closure approximations to create reduced systems of ODE's are not always effective, even though the successive moment ODE's are exact. The procedure in [39,40] evidently is remarkably effective, because the closure step exploits the MSHT FCLT supporting the underlying Gaussian approximation. Thus, we see that there is potential to benefit more from an asymptotic result than can be gained by an immediate direct application.

4 New insights into classical reflected processes

To see the essential behavior of a single-server queueing system, we can consider a model having a net-input stochastic process with stationary and independent increments. Given that the associated queue content should be nonnegative, we impose a

reflecting lower barrier at the origin. To ensure stable long-run behavior, we assume that the net-input process has a negative drift. If we focus on the queue-length process, i.e., the number of "customers" or "jobs" in the system, then it is natural to assume that arrivals and departures occur one at a time. These requirements lead us to the classical $M/M/1$ queue.

4.1 Modeling the variability of the net-input process

A shortcoming of the $M/M/1$ model, like many of its Erlang (Markov) relatives, is that it provides no direct way to quantify the level of variability in the net-input process; it can be thought of as a "deterministic" stochastic model. The $M/M/1$ model can be fully specified by the deterministic arrival and service rates (as well as the initial conditions). Even though the behavior over time is of course stochastic, there is no separate quantification of the level of variability. That is a concern, because we are unlikely to understand the significance of features that we do not model.

If we drop the discreteness requirement, either by focusing on the remaining work in service time in the system or by approximating the queue-length process, then we are led to reflected Lévy processes, which allow the variability to be modeled directly. Variants of these Lévy queueing models are studied in this SI by Glynn and Wang [22] and by Asmussen and Ivanovs [6].

If we assume that the sample paths are continuous (which is quite realistic if the system is heavily loaded and we take a distant view), then the net-input process becomes Brownian motion (BM) and the $M/M/1$ queue is replaced by reflected Brownian motion (RBM), which is characterized by a negative drift $-r$, $r > 0$, and a positive volatility parameter or diffusion coefficient σ, which quantifies the level of variability in the RBM.

4.2 The role of scaling for RBM

If we think of RBM as an approximation for a discrete queueing process, then it can be helpful to exploit heavy-traffic limit theorems, which show that a family of queue-length processes indexed by the traffic intensity ρ, after appropriate scaling of space and time, converges to RBM as the traffic intensity ρ increases to the critical value. In particular, under regularity conditions, the scaled stochastic process

$$\{\hat{Q}_\rho(t) : t \geq 0\} \equiv \{(1-\rho)Q_\rho((1-\rho)^{-2}t) : t \geq 0\} \tag{4.1}$$

approaches RBM as $\rho \to 1$. That scaling of time and space can be understood by the way we scale random walks in the central limit theorem (CLT). Indeed, when we take a function space view, we see that the limit for the entire queueing process can be regarded as a consequence of Donsker's FCLT for random walks [9].

The asymptotic perspective described in the last paragraph is now widely accepted, but the situation was quite different 50 years ago. I can still recall being impressed by the new asymptotic ideas in Skorohod [45], Kolmogorov [36] and Prohorov [43] and the new approach to queues in Kingman [32–34], which were then pursued by my

advisor Iglehart [28], Borovkov [11], and others. I got caught up in this asymptotic fever, but these ideas were unconventional at the time, and were not always well received. However, the lesson is not always clear; the negative reaction to our work might mean that it is not interesting rather than breaking new ground.

When we take such an asymptotic perspective, we see that the scaling provides great insight into the transient behavior of the queue, but we should also carefully study the RBM, as was done by Harrison [23] (updated in [24]). Even for RBM, scaling provides important insight. In particular, we can express RBM with any negative drift and positive volatility in terms of *canonical RBM* with drift -1 and diffusion coefficient 1. Let $p(t, x, y; -r, \sigma)$ denote the probability transition density function, i.e., the probability density of being in state y after an elapsed time t, starting in state x, when the model parameters are $(-r, \sigma)$. With that definition, we can write

$$p(t, x, y : -r, \sigma) = \eta p(\nu t, \eta x, \eta y; -1, 1), \quad t \geq 0, \tag{4.2}$$

for $\eta \equiv r/\sigma^2$ and $\nu \equiv r^2/\sigma^2$, as shown in (2.4) of [22]. Hence, the impact of the volatility parameter is determined solely by the scaling. For further analysis, we can consider canonical RBM. Moreover, by the combination of the heavy-traffic scaling and the invariance scaling, the structure of general RBM is already captured by the $M/M/1$ queue.

Nevertheless, there is more to do, because there is a surprisingly rich structure underlying such a seemingly simple model [10,23] and Chapter 15 of [30]. And even richer structure is found when we go beyond BM net-input processes to consider Lévy net-input processes with discontinuous sample paths. Applications lead to important new questions. For example, the papers by Glynn and Wang [22] and Asmussen and Ivanovs [6] contribute significantly to stochastic simulation, which has important applications for queueing systems and beyond [5].

4.3 On the rate of convergence to equilibrium for RBM

Glynn and Wang [22] study the way RBM approaches equilibrium as time evolves. One source of motivation is the initial-transient or warm-up problem in stochastic simulation, which arises when we apply simulation to estimate unknown steady-state stochastic quantities, but are forced to start the system in some fixed state, commonly chosen to be the empty state. We ask if we should discard an initial portion of each run to allow the system to approach steady state, and if so, how much? Or should we start in a special initial state, and if so, which one? See [49] for discussion and references.

We are also interested in the relation between steady-state quantities and associated transient quantities when we employ steady-state analyses to approximate what are naturally transient quantities in a queueing system, such as an average cost. Studying the way RBM approaches equilibrium extends the literature on "the relaxation time" for queues, as for the $M/M/1$ queue in §III.7.3 of [15].

Glynn and Wang [22] start by taking a classical approach, as reviewed in Chapter 15 of [30], approaching the transient behavior of RBM via its transition probability function $p(t, x, y) \equiv p(t, x, y; -r, \sigma)$, as in (4.2). They show that several useful ways to measure the difference between transient and steady state can be expressed in

terms of the transition function and conveniently bounded. For example, Theorem 2.1 bounds the total variation distance between RBM at time t and its steady state limit via

$$d(t,x) \equiv \sup_A |P_x(X(t) \in A) - P(X(\infty) \in A)| = \frac{1}{2}\int_0^\infty |p(t,x,y) - p(y)|\,dy$$
$$\leq \sqrt{2/\pi}\,e^{-\nu t/2}(\nu t)^{-3/2}(1+\eta x)e^{\eta x} \quad \text{for} \quad t > 1/\nu, \tag{4.3}$$

where the supremum in line 1 is over all measurable sets A, $p(y) \equiv p(\infty, x, y)$ for any x, $\eta \equiv r/\sigma^2$ and $\nu \equiv r^2/\sigma^2$ as in (4.2). (The bound is 1 for $t \leq 1/\nu$, and so of no use near the origin.)

To establish the bound and related asymptotic results, [22] exploits the spectral representation of $p(t,x,y)$, as in §15.13 of [30]; see also [37]. In particular, [22] swiftly derives the representation

$$p(t,x,y) - p(y) = p(y)\int_{-\infty}^{-\nu/2} e^{\lambda t}u_\lambda(x)u_\lambda(y)\phi(\lambda)\,d\lambda, \tag{4.4}$$

where $u_\lambda(x)$ is a function of x and λ, ϕ is a function of λ, where λ is a real number in the open interval $(-\infty, -\nu/2)$ and ν is a positive scaling parameter depending on the drift and volatility. In other words, the difference $p(t,x,y) - p(y)$ can be represented as a continuous (not necessarily nonnegative) mixture of exponentials. They then exploit the trigonometric structure of the functions $u_\lambda(x)$ to obtain insightful expressions and bounds, e.g., for $E_x f(X(t)) - Ef(X(\infty))$, where $X \equiv \{X(t) : t \geq 0\}$ is the RBM and f is an appropriate real-valued function.

As discussed on p. 337 of [30], the spectrum for RBM is continuous. That explains why the rate of convergence is relatively complicated. That is in contrast with the exponential rate of convergence in an irreducible aperiodic finite-state discrete-time Markov chain (DTMC), as discussed in Chapter 10 of [30]. In that DTMC setting, the rate of convergence is determined by the modulus of the second largest eigenvalue.

The paper [22] considers rates of convergence to equilibrium for various functions and metrics. They elaborate on the underlying spectral theory for RBM and discuss the implications for simulation, in particular, to treat the initial-transient problem.

4.4 Discretization error for two-sided reflected Lévy processes

Asmussen and Ivanovs [6] study the discretization error when we approximate a two-sided reflected Lévy Processes by the associated two-sided reflection of the random walk obtained by sampling the net-input process, extending the corresponding study for RBM in [7]. First, for BM $B(t)$ with drift r and volatility σ, it is natural to simulate the BM by generating the associated Gaussian random walk with step size h, so that

$$B_h((k+1)h) = B_h(kh) + r\,B_h(kh)h + \sigma(B((k+1)h) - B(kh)), \quad k \geq 0, \tag{4.5}$$

on the lattice $h\mathbb{N}$ and $B_h(t) = B_h(\lfloor t/h \rfloor h)$ off the lattice, which corresponds to the Euler approximation of the SDE. As might be expected, this approximation scheme

performs well. The main focus of [7] is on reflected analogs. In particular, they start by approximating the RBM $\bar{B} \equiv \psi(B)$, where ψ is the standard one-dimensional reflection map, by $\bar{B}_h \equiv \psi(B_h)$, for B_h in (4.5). They show that the reflection map has a substantial impact on the quality of the approximation. In particular, they show that

$$h^{-1/2}(\bar{B}(t) - \bar{B}_h(t)) \Rightarrow \sqrt{\sigma^2 t}\, W \quad \text{as} \quad h \downarrow 0, \tag{4.6}$$

where W is a positive random variable with an explicit distribution related to the three-dimensional Bessel process. The obvious intuition for the positivity is that the continuous RBM experiences more upward pushing at its lower barrier. The order $1/2$ rate of convergence for RBM is slower than the associated order 1 rate of convergence for BM.

Asmussen and Ivanovs [6] study the corresponding problem for the two-sided reflection of a non-monotone Lévy process. Moreover, the authors suggest an improved simulation algorithm exploiting their asymptotic result. They treat two-sided reflection by noting that a two-sided reflection can be constructed by alternating one-sided reflection at the upper and lower barriers. They then exploit a generalization of the one-sided limit (actually the supremum) in (4.6) obtained in [29]. Useful background on Lévy processes and reflected versions can be found in Chapter IX of [3,8,17].

The key assumption for the limit in [6], which is carefully studied via self-similarity in [29], is an associated limit for the scaled local behavior of the Lévy process in their (2.4), i.e.,

$$X(h)/a(h) \Rightarrow \hat{X}(1) \quad \text{as} \quad h \downarrow 0, \tag{4.7}$$

where $a(h)$ is a scaling function. In the case of BM, the appropriate scaling function in (4.7) is $a(h) = h^{1/2}$ as in (4.6) and

$$h^{-1/2}B(h) \stackrel{\mathrm{d}}{=} N(0,1) \quad \text{for all} \quad h > 0 \tag{4.8}$$

where $N(0,1)$ is a standard normal random variable and $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution.

It is indeed evident that the difference between the sampled random walk approximations of the continuous-time RBM (and other Lévy processes) depend on the local behavior of these continuous-time processes. Moreover, it is well known that the local behavior is quite complex; e.g., the sample paths of BM are almost surely nowhere differentiable. For BM, the local behavior can be extracted from the classical time-inversion property, which says that $tB(1/t)$ is distributed as BM.

For queueing processes, RBM tends to be a good approximation over longer time intervals, but not over shorter time intervals (just as in the classical approximation for the location of particles suspended in a fluid by BM). It is thus natural to wonder, when our real interest is in a queueing process, if it might not be better to simulate a queueing process rather than to, first, approximate the queueing process by RBM and then simulate the RBM by generating the reflected random walk. Of course, the evolution of a general queueing process could be quite complicated. A natural alternative algorithm for the queue-length process in a general $G/G/1$ queue is to

simulate an $M/M/1$ queue that has parameters chosen to match the heavy-traffic limit of the given $G/G/1$ model. This might be done by exploiting strong approximations for queueing processes, as discussed in [14] and references there.

But one might ask: How can we capture the complex variability of the $G/G/1$ queue by the $M/M/1$ queue? That takes us back to §4.2: First, the heavy-traffic limit for the $G/G/1$ queue connects to RBM, which has two parameters, the drift and volatility parameters. Second, the scaling of RBM allows us to transform to canonical RBM, which connects directly to the heavy-traffic limit for the $M/M/1$ queue. Thus, we see that elementary theory acquires significant meaning in application.

5 Personal reflections

As in our probability models, life is full of random events, unexpected twists and turns. I have had good fortune in the hand I was dealt. I had supportive family and friends. I had good teachers, including my thesis advisor, Donald L. Iglehart. I lived at a good time and place, so that I could be a late bloomer without serious penalty. My path exposed me to many interesting gifted people, especially in the queueing community. I thank them all for their friendship and inspiration.

I thank Narahari Umanath (Uma) Prabhu for founding QUESTA in 1986, fostering a cohesive community in our subfield of applied probability and providing a home for much of our research. I also thank the subsequent editors Richard Serfozo (1996–2004), Onno Boxma (2004-2009) and Sergey Foss (2009-) for carrying the torch [20]. And I thank the authors, associate editors and referees for contributing interesting content and maintaining high scientific standards, including honesty, intellectual integrity and a sense of fairness. I am pleased to see that QUESTA is a diverse worldwide academic community, where individual human dignity is respected. I am also pleased to see that this diversity is well represented by the contributors to this SI.

Looking toward the future, the world seems to face more challenges than ever. Fortunately, I see among us many highly gifted researchers, so that the future of queueing seems bright. Hopefully we can do our part to make the world a better place.

References

1. Aksin, O.Z., Armony, M., Mehrotra, V.: The modern call center: a multi-disciplinary perspective on operations management research. Prod. Oper. Manag. **16**, 665–688 (2007)
2. Aras, A.K., Chen, X., Liu, Y.: Many-server Gaussian limits for overloaded queues with customer abandonment and nonexponential service times. Queueing Syst. (2018). https://doi.org/10.1007/s11134-018-9575-0
3. Asmussen, S.: Applied Probability and Queues, 2nd edn. Springer, New York (2003)
4. Asmussen, S., Boxma, O.J.: Editorial introduction: 100 years of queueing, the Erlang centennial. Queueing Syst. **62**, 1–2 (2009)
5. Asmussen, S., Glynn, P.W.: Stochastic Simulation. Springer, New York (2007)
6. Asmussen, S., Ivanovs, J.: Discretization error for a two-sided reflected Lévy process. Queueing Syst. (2018). https://doi.org/10.1007/s11134-018-9576-z
7. Asmussen, S., Glynn, P.W., Pitman, J.: Discretization error in simulation of one-dimensional reflecting Brownian motion. Ann. Appl. Probab. **5**(4), 875–896 (1995)

8. Asmussen, S., Anderson, L.N., Glynn, P.W., Pihlsgaard, M.: Lévy process with two sided reflection. In: Barndorff-Nielsen, O.E., Bertoin, J., Jacod, J., Klűppelberg, C. (eds.) Lévy Matters V, pp. 67–182. Springer, New York (2015)
9. Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley, New York (1999)
10. Borodin, A.N., Salminen, P.: A Handbook of Brownian Motion: Facts and Formulae, 2nd edn. Sppringer Basel, New York (2015)
11. Borovkov, A.A.: Some limit theorems in the theory of mass service, II. Theor. Prob. Appl. **10**, 375–500 (1965)
12. Brockmeyer, E., Halstrom, H.L., Jensen, A.: The Life and Works of A. K. Erlang. Academy of Technical Sciences, Copenhagen (1948)
13. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L.: Statistical analysis of a telephone call center: a queueing-science perspective. J. Am. Stat. Assoc. **100**, 36–50 (2005)
14. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks. Springer, New York (2001)
15. Cohen, J.W.: The Single Server Queue, 2nd edn. North-Holland, Amsterdam (1982)
16. Dallery, Y., Gershwin, B.: Manufacturing flow line systems: a review of models and analytical results. Queueing Syst. **12**, 3–94 (1992)
17. Debicki, K., Mandjes, M.: Queues and Lévy Fluctuation Theory. Springer, London (2015)
18. Filippov, A.F.: Differential Equations with Discontinuous Righthand Sides. Springer, New York (2013). reprinted from 1988
19. Foderaro, L.W.: Navigation apps are truning quiet neighborhoods into traffic nightmares. The New York Times, (December 24, 2017). New York Regional Section
20. Foss, S.: Editorial. Queueing Syst. **64**(1), 1–3 (2010)
21. Garnett, O., Mandelbaum, A., Reiman, M.I.: Designing a call center with impatient customers. Manuf. Serv. Oper. Manag. **4**(3), 208–227 (2002)
22. Glynn, P.W., Wang, R.J.: On the rate of convergence to equilibrium for reflected Brownian motion. Queueing Syst. (2018). https://doi.org/10.1007/s11134-018-9574-1
23. Harrison, J.M.: Brownian Motion and Stochastic Flow Systems. Wiley, New York (1985)
24. Harrison, J.M.: Brownian Models of Performance and Control. Cambridge University Press, New York (2013)
25. Hassin, R.: Rational Queueing. CRC Press, Boca Raton (2016)
26. Hassin, R., Haviv, M.: To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems. Springer, New York (2003)
27. Ibrahim, R.: Sharing delay information in service systems: a literature survey. Queueing Syst. (2018). https://doi.org/10.1007/s11134-018-9577-y
28. Iglehart, D.L.: Limit diffusion approximations for the many-server queue and the repairman problem. J. Appl. Probab. **2**, 429–441 (1965)
29. Ivanovs, J.: Zooming in on a Lévy process at its supremum. Ann. Appl. Prob. (2018) arXiv:1610.904471v3
30. Karlin, S., Taylor, H.M.: A Second Course in Stochastic Processes. Academic Press, New York (1981)
31. Kaspi, H., Ramanan, K.: SPDE limits of many-server queues. Ann. Appl. Probab. **23**, 145–229 (2013)
32. Kingman, J.F.C.: The single server queue in heavy traffic. Proc. Camb. Phil. Soc. **77**, 902–904 (1961)
33. Kingman, J.F.C.: On queues in heavy traffic. J. R. Stat. Soc. B **24**, 383–392 (1962)
34. Kingman, J.F.C.: The heavy-traffic approximation in the theory of queues. In: Smith, W.L., Wilkinson, W.E. (eds.) Proceedings of the Symposium on Congestion Theory, chapter 6, pp. 137–159. University of North Carolina Press, Chapel Hill, NC (1965)
35. Kingman, J.F.C.: The first Erlang century—and the next. Queueing Syst. **63**, 3–12 (2009)
36. Kolmogorov, A.N.: On Skorohod convergence. Theory Probab. Appl. **1**, 215–222 (1956)
37. Linetsky, V.: On the transition densities for reflected diffusions. Adv. Appl. Probab. **37**(2), 435–460 (2005)
38. Mandelbaum, A., Massey, W.A., Reiman, M.I.: Strong approximations for Markovian service networks. Queueing Syst. **30**, 149–201 (1998)
39. Massey, W.A., Pender, J.: Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. Queueing Syst. **75**, 243–277 (2013)
40. Massey, W.A., Pender, J.: Dynamic rate Erlang—a queues. Queueing Syst. (2018). https://doi.org/10.1007/s11134-018-9581-2
41. Naor, P.: The regulation of queue size by levying tolls. Econometrica **37**(1), 15–24 (1969)
42. Prabhu, N.U.: Editorial introduction. Queueing Syst. **1**(1), 1–4 (1986)

43. Prohorov, YuV: Convergence of random proccesses and limit theorems in probability. Theory Probab. Appl. **1**, 157–214 (1956)
44. Puhalskii, A.A.: On the $M_t/M_t/K_t+M_t$ queue in heavy traffic. Math. Methods Oper. Res. **78**, 119–148 (2013)
45. Skorohod, A.V.: Limit theorems for stochastic processes. Theory Probab. Appl. **1**, 261–290 (1956)
46. Stein, C.: Approximate Computation of Expectations. Institute of Mathematical Statistics, Hayward, California (1986). Lecture Notes - Monograph Series 7
47. Stidham, S.: The Optimal Design of Queues. CRC Press, Boca Raron, FL (2009)
48. van Vuuren, M., Adan, I.J.B.F., Resing-Sassen, S.A.E.: Performance analysis of multi-server tandem queues with finite buffers and blocking. OR Spectrum **27**, 315–338 (2005)
49. Wang, R., Glynn, P.W.: On the marginal standard error rule and testing of the initial transient deletion methods. ACM Trans Model Comput. Simul. **27**(1), 1–30 (2016)
50. Zychlinski, N., Mandelbaum, A., Momcilovic, P., Cohen, I.: Bed blocking in hospitals due to scarece capacity in geriatric institutions – cost minimization via fluid models. Working paper, the Technion, Haifa, Israel (2017)
51. Zychlinski, N., Mandelbaum, A., Momcilovic, P.: Time-varying tandem queues with blocking: modeling, analysis and operational insights for fluid models with reflection. Queueing Syst. (2018). https://doi.org/10.1007/s11134-018-9578-x