


Dynamic rate Erlang-A queues

William A. Massey¹ · Jamol Pender² 

Received: 4 August 2017 / Revised: 15 January 2018 / Published online: 3 May 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract The multi-server queue with non-homogeneous Poisson arrivals and customer abandonment is a fundamental dynamic rate queueing model for large-scale service systems such as call centers and hospitals. Scaling the arrival rates and number of servers arises naturally when a manager updates a staffing schedule in response to a forecast of increased customer demand. Mathematically, this type of scaling ultimately gives us the fluid and diffusion limits as found in Mandelbaum et al. (Queueing Syst 30(1):149–201, 1998) for Markovian service networks. These asymptotics were inspired by the Halfin and Whitt (Oper Res 29(3):567–588, 1981) scaling for multi-server queues. In this paper, we provide a review and an in-depth analysis of the Erlang-A queueing model. We prove new results about cumulant moments of the Erlang-A queue, the transient behavior of the Erlang-A limit cycle, new fluid limits for the delay time of a virtual customer, and optimal static staffing policies for healthcare systems. We combine tools from queueing theory, ordinary differential equations, complex analysis, cumulant moments, orthogonal polynomials, and dynamic optimization to obtain new insights about this fundamental queueing model.

This work is dedicated to Ward Whitt, on the occasion of his 75th birthday. We are eternally grateful for his friendship, mentorship, guidance, kindness, and infinite knowledge about stochastic processes.

✉ Jamol Pender
jjp274@cornell.edu

William A. Massey
wmassey@princeton.edu

- ¹ Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA
- ² School of Operations Research and Information Engineering, Cornell University, 228 Rhodes Hall, Ithaca, NY 14853, USA

Keywords Multi-server queues · Abandonment · Dynamical systems · Asymptotics · Time-varying rates · Time-inhomogeneous Markov processes · Hermite polynomials · Fluid and diffusion limits · Skewness · Cumulant moments

Mathematics Subject Classification 60K25



W. A. Massey, W. Whitt, and J. Pender at the 21st Applied Probability Day

1 Introduction

The year of 2017, when we are celebrating the 75th birthday of Ward Whitt, also marks the centennial of the Erlang blocking formula Erlang [8]. This was derived by Agner Krarup Erlang and published in 1917.

His innovation a century ago was to model the performance of telephone trunk line usage, commonly referred to as “all circuits are busy”, as a *constant* rate or *time-homogeneous* Markov process. In our telephony context, an average customer telephone calling rate is the constant rate. We assume that the mean calling times or abandonment items never have any time of day or hour effects. The transient distribution for such a discrete state Markov processes is the unique solution to a set of ordinary differential equations that form a *linear dynamical system*. The general theory for such processes and chains provides analysis for the long run distribution of these Markov models. There is also a unique stationary distribution. The analysis for determining this stationary limit reduces to solving a set of linear equations.

The Erlang blocking formula paved the way for 20th century *queueing theory* as an active branch of applied probability. Unfortunately, real life rates are *dynamic* in nature. Empirically, for example, average telephone call arrival rates can be a function of the time of day, the day of the week, and the week of the year. Hence, a queueing analysis is needed that focuses on *time-inhomogeneous* Markov processes. When this occurs, any closed-form solution for the distribution of a Markovian queueing model quickly becomes intractable.

Typical approaches for analyzing time and state dependent queueing models include asymptotic methods such as heavy traffic limit theory, strong approximation theory, and uniform acceleration; see, for example, Massey [32], Halfin and Whitt [11], Mandelbaum et al. [29], Pender and Phung-Duc [49], Puhalskii [54]. Uniform acceleration,

introduced in Puhalskii [32], and extended to time-inhomogeneous, finite state Markov processes in Massey and Whitt [36], is a dynamic generalization of steady state analysis. It has successfully captured the dynamic and transient behavior for the transition probabilities and moments such as the mean and variance for the dynamic rate analogues of many fundamental queueing models.

Moreover, the strong approximation methods are useful for extending the uniform acceleration analysis, from moments and transition probabilities, directly to the sample path behavior of the Markov process. This analysis can show that properly rescaled queueing sample paths converge to deterministic and Gaussian-like diffusion limits. In Hampshire et al. [14] for example, this sample path approach allows us to analyze the sojourn times for a processor sharing queue with non-homogeneous Poisson arrivals. However, one drawback of these methods is that they are asymptotic as we scale the rate functions upwards. Hence, the convergence of these methods depends on how large the rates are in the problem of interest. The inaccuracy for queueing processes with moderate to small rates has been noted by Massey and Pender [34].

Another method that is quite common in the queueing literature is *moment closure approximation*. Moment closure approximations are used to approximate the queueing process distribution with a lower-dimensional dynamical system. These deterministic systems are then used to compute approximate moments of the original queueing process. They can also be used to approximate unknown functions of the queue length to yield a set of equations that only depend on the moments that are being estimated.

One such method used by Massey and Pender [33, 34], Pender [41, 44, 46], Pender et al. [50], Pender and Ko [47], Engblom and Pender [7] is to use Hermite polynomials for approximating the distribution of the queue length process. In fact, they show that using a quadratic polynomial works quite well. Since the Hermite polynomials are orthogonal to the Gaussian distribution, which has support on the entire real line, these Hermite polynomial approximations do not take into account the discreteness of the queueing process and the fact that the queueing process is positive (including zero). However, they show that Hermite polynomials are natural to analyze since they are orthogonal with respect to the Gaussian distribution, and the heavy traffic limits of multi-server queues, which were pioneered by Ward Whitt, are Gaussian.

This paper is an in-depth analysis of the Erlang-A queue with dynamic rates. It is a fundamental and canonical queueing model for the dynamic behavior of customer abandonments, telecommunication system design, transportation system design and control, as well as emerging applications for resource sharing in healthcare and cloud computing. The lack of simple formulas for the transition probabilities of the Markovian version of the Erlang-A queue is circumvented by finding low-dimensional dynamical system approximations and representations. Our results are inspired by two papers co-authored by Ward Whitt. The first one, Halfin and Whitt [11], proves many-server heavy traffic limits for a multi-server queueing model, and the second one, Duffield et al. [4], exploits the cumulant generating function and cumulant moments to derive expressions for the mean and variance of packet networks. By combining new closure approximations, orthogonal polynomials, new types of cumulants, differential equations, and complex analysis, we are able to derive several new results that reveal the complex dynamics of the Erlang-A queue.

1.1 Organization and summary of results

This paper is organized both to review classic as well as recent results and to present new results for the Erlang-A queue. Section 2 describes the dynamic rate Erlang-A queue and its fluid and diffusion limits. Section 3 introduces cumulants and the new notion of a *functional* cumulant. Here, we derive several properties for both types of cumulants. We also rederive the classic result that all “analytic” distributions with a finite number of nonzero cumulant moments are Gaussian.

Section 4 uses the functional cumulant moments to express the forward equations for the cumulant moments of the Erlang-A queue. We use them to analyze these cumulant moments and obtain some new results for the Erlang-A departure process. We also show in this section an example of the phase space dynamics for an infinite server queue with a period arrival rate. It is an ellipse of a fixed shape with a moving center point. This center achieves a limit point which makes the resulting ellipse a limit cycle. Lastly, in this section, we review Hermite polynomials and introduce various closure approximations for the Erlang-A queue. They were referred to in Massey and Pender [34] as the deterministic mean approximation, the Gaussian variance approximation, and the Gaussian skewness approximation. These closure methods follow from the dynamics of the cumulant moments for the Erlang-A queue.

Section 5 reviews fluid limit approximations to the mean delay of the Erlang-A queue and develops new algorithms related to computing the mean delay using the Gaussian variance and skewness methods. We show how to use the closure methods to stabilize the probability of delay in the Erlang-A model.

Section 5.4 uses the Erlang-A fluid limit and dynamic optimization, as formulated in Hampshire [12], Hampshire and Massey [13], to derive the *static* optimal number of service agents and the dynamics of customer opportunity costs over a finite time period $(0, T]$. Problems involving the static optimization of dynamical systems are inspired by the type of healthcare issues arising over large time scales as found in nursing home management. Finally, we conclude in Sect. 6 and offer suggestions for future research.

2 Erlang-A queueing and Halfin–Whitt scaling

The Erlang-A queueing model is a fundamental queueing model for 21st century queueing. The work of Mandelbaum et al. [29] shows that this queueing system process $Q \equiv \{Q(t) | t \geq 0\}$ is represented by the following stochastic, time changed integral equation:

$$Q(t) = Q(0) + \Pi_+ \left(\int_0^t \lambda(s) ds \right) - \Pi_- \left(\int_0^t \delta(Q(s), c(s)) ds \right), \quad (1)$$

where $\Pi_+ \equiv \{\Pi_+(t) | t \geq 0\}$ and $\Pi_- \equiv \{\Pi_-(t) | t \geq 0\}$ are independent and identically distributed *standard* (rate 1) Poisson processes and

$$\delta(Q, c) \equiv \mu \cdot (Q \wedge c) + \beta \cdot (Q - c)^+. \quad (2)$$

Thus, we can write the sample path dynamics of the Erlang-A queueing process in terms of two independent unit rate Poisson processes. A deterministic time change for Π_+ transforms it into a non-homogeneous Poisson arrival process with rate function λ . This process counts the customer arrivals that occur over some time interval. A random time change for the Poisson process Π_- gives us a departure process that counts the number of both the serviced and abandoning customers. We implicitly assume that the number of servers is a deterministic function of time, $c(t)$, and that each server works at rate μ . We also assume that the abandonment distribution is exponential and the rate of abandonments is equal to β . When the mean number in the system $EQ(t)$ is less than the number of servers $c(t)$, or $EQ(t) < c(t)$, we say that the system is *underloaded*. Conversely, when $EQ(t) > c(t)$, we say that the system is *overloaded*. Finally, when $EQ(t) = c(t)$, we say that the system is *critically loaded*.

The Erlang-A queueing system, when underloaded for all t , behaves like a dynamic rate, infinite server queue. This is equivalent to setting the number of servers is equal to ∞ . When initialized by a Poisson distribution, the dynamic rate infinite server queue always has a Poisson transient distribution. Detailed explorations of infinite server queueing dynamics with non-homogeneous input can be found in the works of Palm [39], Khinchin et al. [20], and Eick et al. [5,6]. Moreover, under general conditions, the Poisson distribution is uniquely characterized by having all its cumulant moments equal to its mean [20].

The Erlang-A queue provides a unifying framework for the following three fundamental classical queueing models of the 20th century:

1. The $M(t)/M/c/c$ queue is a special case of the Erlang-A queue in the limit as the abandonment rate β approaches infinity. We liken this model to the extreme case of a multiple service agent system with highly *impatient* customers. Here, the output rate is $\mu \cdot Q$, but we always have the constraint $Q \leq c$.
2. The $M(t)/M/\infty$ queue is also special case of the Erlang-A queue, both for the case of an unlimited number of service agents and the case of a finite number of service agents, with the abandonment rate equaling the service rate, or $\beta = \mu$. For either case, the output rate is $\mu \cdot Q$.
3. The $M(t)/M/c/\infty$ queue is a special case of the Erlang-A queue when the abandonment rate β equals zero. We liken this model to the extreme case of a multiple service agent system with highly *patient* customers. Here, the output rate is $\mu \cdot \min(Q, c)$.

The enduring importance of the Halfin and Whitt [11] scaling is that for multiple service agent queueing systems, it is natural to scale up the arrival rate and the number of servers simultaneously. This is equivalent to applying the same scale factor to both the *demand* (customer arrival rate) and the *supply* (total number of available service agents) of a service enterprise. In the context of a telephone call center, this scaling of supply is called *resource pooling*. Since the Erlang-A queueing process is a special case of a single node *Markovian service network*, we can construct an associated, *uniformly accelerated* queueing process where both the new arrival rate $\eta \cdot \lambda(t)$ and the new number of servers $\eta \cdot c(t)$ are scaled by the same factor $\eta > 0$. Thus, the *Halfin–Whitt* scaling gives us the following sample path representation for the Erlang-A queue length process:

$$Q^\eta(t) = Q^\eta(0) + \Pi_+ \left(\int_0^t \eta \cdot \lambda(s) ds \right) - \Pi_- \left(\int_0^t \delta(Q^\eta(s), \eta \cdot c(s)) ds \right).$$

Observe that this scaling of supply and demand for a call center does *not* require the customers to speak to their service agents or decide to abandon the system more quickly. The Halfin–Whitt scaling is only for the aggregate or macroscopic phenomena of the queueing system.

Taking the following limits gives us the *fluid* and *diffusion* limit models of Mandelbaum et al. [29], i.e.,

$$\lim_{\eta \rightarrow \infty} \frac{Q^\eta(t)}{\eta} = q(t) \quad \text{a.s.} \quad \text{and} \quad \lim_{\eta \rightarrow \infty} \sqrt{\eta} \cdot \left(\frac{Q^\eta(t)}{\eta} - q(t) \right) \stackrel{d}{=} \hat{Q}(t), \quad (3)$$

where the deterministic process $q(t)$ or the *fluid limit* is governed by the one-dimensional dynamical system

$$\dot{q} = \lambda - \delta(q, c), \quad \text{where} \quad \dot{q}(t) \equiv \frac{dq}{dt}(t). \quad (4)$$

The latter statement is the “dot notation” of physics that we use to denote a time derivative when we are suppressing the time dependence for the given function of time.

As pointed out in Mandelbaum et al. [29], if the set of time points $\{t \mid q(t) = c(t)\}$ has measure zero, then $\hat{Q}(t)$ is a *Gaussian* diffusion process (with mean zero when $Q^\eta(0)$ is only a constant scaled by η) whose variance combines with the fluid mean to form a two-dimensional dynamical system given by (4) and

$$\frac{\dot{v} + \dot{q}}{2} = \lambda - (\mu \cdot \{q < c\} + \beta \cdot \{q \geq c\}) \cdot v, \quad \text{where} \quad \{q < c\} \equiv \begin{cases} 1 & \text{if } q < c, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

$v \equiv \text{Var} \hat{Q}$, and $\{q < c\}$ denotes an example of an *indicator function* for the event $q < c$.

In the case where the arrival rate function is a constant for all time, we can recover the result of Halfin and Whitt [11]. To bypass this asymptotic approach and apply closure approximation methods, we develop the notions of cumulant moments in the next section.

3 Cumulant moments and functional generalizations

3.1 Cumulant moments

We define a real-valued random variable X to be *analytic* if its moment generating function,

$$Ee^{zX} = \sum_{m=0}^{\infty} \frac{z^m}{m!} \cdot EX^m, \tag{6}$$

is an analytic function for all complex z in a neighborhood of zero. Thus, the distribution for an analytic random variable is uniquely characterized by all its existing moments.

The value of this function is close to one when z is close to zero. From this, it follows that the function $\log Ee^{zX}$ is also analytic in a neighborhood of zero. We can then define the *cumulant moments* of X or $\{C^{(k)}X \mid k \geq 1\}$ to be

$$\sum_{k=1}^{\infty} \frac{z^k}{k!} \cdot C^{(k)}X \equiv \log Ee^{zX}, \tag{7}$$

where the first four moments are

$$C^{(1)}X = EX, \tag{8}$$

$$C^{(2)}X = \text{Var}X \equiv E(X - EX)^2, \tag{9}$$

$$C^{(3)}X = E(X - EX)^3, \tag{10}$$

$$C^{(4)}X = E(X - EX)^4 - 3 \cdot (\text{Var}X)^2. \tag{11}$$

Now, define the *Skew* X and *Kurt* X , respectively, to be the (generalized) *skewness* and *kurtosis* of the random variable X , where

$$\text{Skew}X \equiv \frac{C^{(3)}X}{\sqrt{(\text{Var}X)^3}} \quad \text{and} \quad \text{Kurt}X \equiv \frac{C^{(4)}X}{(\text{Var}X)^2}. \tag{12}$$

These are the *normalized* versions of the third and fourth cumulant moments, in the same manner that the *correlation* of two random variables is the normalized version of their *covariance*.

Now, we summarize the queueing theoretic appeal and utility of cumulant moments for analytic random variables.

Theorem 1 *For all analytic random variables X and Y , we have:*

1. *For all constants a , we have*

$$C^{(k)}[aX] = a^k \cdot C^{(k)}X. \tag{13}$$

2. *The pair X and Y are independent if and only if*

$$C^{(k)}[a \cdot X + b \cdot Y] = a^k \cdot C^{(k)}X + b^k \cdot C^{(k)}Y \tag{14}$$

for all strictly positive integers k as well as all constants a and b .

3. *All cumulant moments for X of degree 2 or higher are zero if and only if X has the distribution of a constant (point mass distribution).*

4. All cumulant moments for X of degree 3 or higher are zero if and only if X has a Gaussian distribution.
5. All the cumulant moments of X are equal to the mean of X if and only if X is Poisson.

Proof See the Appendix of Hampshire [12] or the Appendix of Pender [53]. □

All cumulative moments are then additive over independent superpositions of random variables as pointed out in Whitt [61] and Choudhury and Whitt [1]. Hence, cumulative moments can give us a stochastic principal component analysis of our queueing processes.

Skewness and kurtosis are also invariant under both constant translations and scalings of analytic random variables. Since any Gaussian random variable is the translation and scaling of one that is *standard* (mean zero and variance one) Gaussian, the skewness and kurtosis of a random variable serve as intrinsic measures of how far an analytic random variable is from being Gaussian.

Cumulant moments give a simple characterization of Gaussian random variables but they can also identify Gaussian random variables with the following stronger result due to Marcinkiewicz [31].

Theorem 2 *Analytic random variables have a finite number of nonzero cumulant moments if and only if they are Gaussian.*

Proof For any complex number z , define the complex value $g(z)$, where

$$g(z) \equiv \log Ee^{z \cdot X}. \tag{15}$$

Notice that $g(r)$ is a real number whenever r is real. Moreover, we can apply Jensen’s inequality and obtain $g(r) \geq r \cdot EX$. This gives us

$$\log Ee^{r \cdot (X - EX)} = g(r) - r \cdot EX \geq 0. \tag{16}$$

Without any loss of generality, we can assume that $EX = 0$ and g is a *positive* real function when its domain is restricted to the real line.

For all complex z , we have

$$\operatorname{Re} g(z) = \operatorname{Re} \log Ee^{z \cdot X} = \log \left| Ee^{z \cdot X} \right| \leq \log E \left| e^{z \cdot X} \right| = \log Ee^{\operatorname{Re} z \cdot X} = g(\operatorname{Re} z). \tag{17}$$

Moreover, if X has only a finite number of nonzero cumulant moments, then g must be a polynomial of some degree n . The leading coefficient of this polynomial must be some strictly positive real number $\alpha > 0$, since

$$\alpha = \lim_{r \rightarrow +\infty} \frac{g(r)}{r^n}. \tag{18}$$

This means that positive scalings of functions like g gives us functions like z^n , since

$$z^n = \lim_{r \rightarrow +\infty} \frac{g(rz)}{\alpha \cdot r^n}. \tag{19}$$

Hence, the function z^n satisfies the same inequality as g , or

$$\operatorname{Re}(z^n) \leq (\operatorname{Re}z)^n \tag{20}$$

for all complex z .

Restricting ourselves to the unit circle, or $z = e^{i\theta}$, reduces this inequality to

$$\cos n\theta \leq (\cos \theta)^n \tag{21}$$

for all real θ . Now, the question is for which positive integers n does this hold? This inequality is true for the case of $n = 1$. It is also true for the case of $n = 2$ since $\cos 2\theta = (\cos \theta)^2 - (\sin \theta)^2 \leq (\cos \theta)^2$.

However, for $n \geq 3$, a contradiction occurs when we set $\theta \equiv 2\pi/n$. We have $0 < \theta < \pi$ but now $(\cos \theta)^n < 1 = \cos n\theta$. Hence, requiring that g be a polynomial forces its degree to be less than or equal to 2. \square

3.2 Functional cumulant moments

Consider a random variable X for which there exists some positive real number r with $Ee^{rX} < \infty$. Now, define an *exponential change of measure* to be an expectation E_r such that

$$E_r f(X) \equiv \frac{E[f(X) \cdot e^{rX}]}{Ee^{rX}}, \tag{22}$$

for all bounded continuous functions f .

If X is analytic, then, by using analytic continuation, we can define an analytic function E_z , when z is close to zero, which is

$$E_z X = \frac{E[X \cdot e^{zX}]}{Ee^{zX}} = \frac{d}{dz} \log Ee^{zX} = \sum_{k=0}^{\infty} \frac{z^k}{k!} \cdot C^{(k+1)} X. \tag{23}$$

From this, it follows that

$$C^{(k+1)} X = \left. \frac{d^k}{dz^k} \right|_{z=0} E_z X. \tag{24}$$

Moreover, these cumulant moments uniquely characterize the underlying distribution for X .

We now define the *functional cumulant* of X with respect to some bounded continuous function f to be equal to

$$C_z^{(k+1)} f[X] \equiv \left. \frac{d^k}{dz^k} \right|_{z=0} E_z [f(X)]. \tag{25}$$

The following result relates functional cumulants to exponential changes of measure.

Theorem 3 Whenever $E_z f(X)$ is analytic in z , then

$$\frac{d^k}{dz^k} E_z f(X) = C_z^{(k+1)} f[X], \quad (26)$$

for all positive integers k .

Proof Observe that

$$E_{z+w} f(X) = \frac{E[f(X)e^{(z+w)X}] / Ee^{zX}}{Ee^{(z+w)X} / Ee^{zX}} = \frac{E_z f(X)e^{wX}}{E_z e^{wX}} = (E_z)_w [f(X)]. \quad (27)$$

From this, it follows that

$$C_z^{(k+1)} f[X] = \left. \frac{\partial^k}{\partial w^k} E_{z+w} f(X) \right|_{w=0} = \frac{d^k}{dz^k} E_z f(X). \quad (28)$$

This completes the proof. \square

Corollary 1 The first four functional cumulant moment formulas are:

$$\begin{aligned} C^{(1)} f[X] &= Ef(X), \\ C^{(2)} f[X] &= \text{Cov}[f(X), X], \\ C^{(3)} f[X] &= \text{Cov}\left[f(X), (X - EX)^2\right], \\ C^{(4)} f[X] &= \text{Cov}\left[f(X), (X - EX)^3\right] - 3 \cdot \text{Var}X \cdot \text{Cov}[f(X), X]. \end{aligned}$$

Proof Without loss of generality, we can assume that $EX = 0$ and redefine f since

$$E_z f(X) = \frac{E[f(X) \cdot e^{zX}]}{Ee^{zX}} = \frac{E[f(X) \cdot e^{z \cdot (X-EX)}]}{Ee^{z \cdot (X-EX)}} = \frac{E[f(Y + EX) \cdot e^{zY}]}{Ee^{zY}}, \quad (29)$$

where $Y \equiv X - EX$. We can expand the numerator of the final ratio and obtain

$$\begin{aligned} E[f(Y) \cdot e^{zY}] &= Ef(Y) + z \cdot E[f(Y) \cdot Y] \\ &\quad + \frac{z^2}{2} \cdot E[f(Y) \cdot Y^2] + \frac{z^3}{6} \cdot E[f(Y) \cdot Y^3] + O(z^4). \end{aligned}$$

Expanding the denominator of the final ratio yields

$$E[e^{zY}] = 1 + \frac{z^2}{2} \cdot \text{Var}Y + \frac{z^3}{6} \cdot C^{(3)}Y + O(z^4), \quad (30)$$

which gives us

$$\frac{1}{E[e^{zY}]} = 1 - \frac{z^2}{2} \cdot \text{Var}Y - \frac{z^3}{6} \cdot C^{(3)}Y + O(z^4). \tag{31}$$

Finally, by combining these expansions, we have

$$\begin{aligned} E_z f(Y) &= Ef(Y) + z \cdot E[f(Y) \cdot Y] + \frac{z^2}{2} \cdot \left(E[f(Y) \cdot Y^2] - Ef(Y) \cdot \text{Var}Y \right) \\ &\quad + \frac{z^3}{6} \cdot \left(E[f(Y) \cdot Y^3] - E[f(Y)] \cdot C^{(3)}Y - \frac{6}{2} \cdot E[f(Y) \cdot Y] \cdot \text{Var}Y \right) \\ &\quad + O(z^4). \end{aligned}$$

□

When f is a scaling function, like $f(x) = \delta(x, c) = \mu \cdot x$ when $\mu = \beta$, then all the functional cumulant moments of X are equal to μ times the corresponding cumulant moment of X . In general, a functional moment of a random variable may *not* be equal or proportional to its cumulant moment or even a cumulant moment of the function applied to the random variable.

4 Closure approximations for the Erlang-A queue

In this section, we derive the functional forward equations for the Erlang-A queueing model and approximate the cumulant moments of the Erlang-A queue using Gaussian-based closure approximations. We show that these closure approximations are effective at approximating the cumulant moments of the Erlang-A queue. They also approximate other important performance measures like the probability of delay.

4.1 Functional forward equations

To gain a better understanding of the dynamics of the mean, variance, and third cumulant moment of the Erlang-A queueing process, we need to study their rates of change over time. Hence, we employ the *functional version* of the Kolmogorov forward equations for the Erlang-A queue, which is of the form

$$\dot{E}f(Q) = \lambda \cdot E[f(Q + 1) - f(Q)] + E[\delta(Q, c) \cdot (f(Q - 1) - f(Q))], \tag{32}$$

for all appropriate functions f . We always assume, for the remainder of this paper, that quantities such as β and μ are constant. To simplify our notation, time dependent quantities such as $Q(t)$, $\lambda(t)$, and $c(t)$ are denoted in this paper as Q , λ , and c , with their time dependence suppressed. We again use the dot notation of physics to denote a time derivative.

Theorem 4 *The dynamics for the cumulant moment generating function of an Erlang-A queueing process are*

$$\log \left(\dot{E} e^{zQ} \right) = (1 - e^{-z}) \cdot (\lambda \cdot e^z - E_z \delta(Q, c)). \tag{33}$$

Proof For any given function f , the functional forward equation for the Erlang-A queue is

$$\dot{E} [f(Q)] = \lambda \cdot E [f(Q + 1) - f(Q)] + E [\delta(Q, c) \cdot (f(Q - 1) - f(Q))]. \tag{34}$$

For the special case of $f(Q) \equiv e^{zQ}$, this reduces to

$$\dot{E} e^{zQ} = \lambda \cdot (e^z - 1) \cdot E e^{zQ} + (e^{-z} - 1) \cdot E [\delta(Q, c) \cdot e^{zQ}]. \tag{35}$$

Dividing both sides by $E e^{zQ}$ gives us

$$\begin{aligned} \log \left(\dot{E} e^{zQ} \right) &= \lambda \cdot (e^z - 1) + (e^{-z} - 1) \cdot E [\delta(Q, c) \cdot e^{zQ(t)}] / E e^{zQ} \\ &= (1 - e^{-z}) \cdot (\lambda \cdot e^z - E_z \delta(Q, c)). \end{aligned} \tag{36}$$

$$\tag{37}$$

□

Functional cumulant moments give us the language to express a new result for the steady state distribution of the Erlang-A queue.

Corollary 2 *For an Erlang-A queueing process with a constant arrival rate, we have, in steady state,*

$$\lambda = C^{(k)} \delta_c [Q], \tag{38}$$

for all strictly positive integers k , where $\delta_c(\cdot) \equiv \delta(\cdot, c)$.

Proof When λ is a constant and we are in steady state, then the time derivative of the logarithm term is zero. This gives us

$$0 = (1 - e^{-z}) \cdot (\lambda \cdot e^z - E_z \delta(Q, c)). \tag{39}$$

The rest follows by factoring out $(1 - e^{-z})$, differentiating with respect to z and setting $z = 0$. □

Corollary 3 *The dynamics of the cumulant moments for the dynamic rate Erlang-A queue are*

$$\dot{C}^{(k)} Q(t) = \lambda(t) - \sum_{j=1}^k \binom{k}{j-1} (-1)^{k-j} \cdot C^{(j)} \delta_c [Q(t)], \tag{40}$$

for all strictly positive integers k .

Proof If we differentiate this formula k times, with respect to z , then we have, for all strictly positive integers k ,

$$\dot{C}_z^{(k)} Q = \lambda \cdot e^{-z} + (e^{-z} - 1) \cdot C_z^{(k+1)} \delta_c [Q] + e^{-z} \cdot \sum_{j=0}^k \binom{k}{j} (-1)^{k-j} \cdot C_z^{(j+1)} \delta_c [Q]. \tag{41}$$

Setting $z = 0$ gives us the rest. □

Corollary 4 *We have the following dynamics for the first six cumulant moments:*

$$\dot{C}^{(1)} Q = \lambda - C^{(1)} \delta_c [Q], \tag{42}$$

$$\frac{\dot{C}^{(1)} Q}{2} + \frac{\dot{C}^{(2)} Q}{2} = \lambda - C^{(2)} \delta_c [Q], \tag{43}$$

$$\frac{\dot{C}^{(1)} Q}{6} + \frac{\dot{C}^{(2)} Q}{2} + \frac{\dot{C}^{(3)} Q}{3} = \lambda - C^{(3)} \delta_c [Q], \tag{44}$$

$$\frac{\dot{C}^{(2)} Q}{4} + \frac{\dot{C}^{(3)} Q}{2} + \frac{\dot{C}^{(4)} Q}{4} = \lambda - C^{(4)} \delta_c [Q], \tag{45}$$

$$-\frac{\dot{C}^{(1)} Q}{30} + \frac{\dot{C}^{(3)} Q}{3} + \frac{\dot{C}^{(4)} Q}{2} + \frac{\dot{C}^{(5)} Q}{5} = \lambda - C^{(5)} \delta_c [Q], \tag{46}$$

$$-\frac{\dot{C}^{(2)} Q}{12} + \frac{5\dot{C}^{(4)} Q}{12} + \frac{\dot{C}^{(5)} Q}{2} + \frac{\dot{C}^{(6)} Q}{6} = \lambda - C^{(6)} \delta_c [Q]. \tag{47}$$

Proof Rewriting cumulant moments dynamics in infinite matrix form gives us

$$\begin{bmatrix} \dot{C}^{(1)} Q \\ \dot{C}^{(2)} Q \\ \dot{C}^{(3)} Q \\ \dot{C}^{(4)} Q \\ \vdots \end{bmatrix} = \begin{bmatrix} \lambda \\ \lambda \\ \lambda \\ \lambda \\ \vdots \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ -1 & 2 & 0 & 0 & \cdots \\ 1 & -3 & 3 & 0 & \cdots \\ -1 & 4 & -6 & 4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \cdot \begin{bmatrix} C^{(1)} \delta_c [Q] \\ C^{(2)} \delta_c [Q] \\ C^{(3)} \delta_c [Q] \\ C^{(4)} \delta_c [Q] \\ \dots \end{bmatrix}. \tag{48}$$

The column vector where all entries equal λ is an eigenvector for this matrix, or

$$\begin{bmatrix} \lambda \\ \lambda \\ \lambda \\ \lambda \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ -1 & 2 & 0 & 0 & \cdots \\ 1 & -3 & 3 & 0 & \cdots \\ -1 & 4 & -6 & 4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \cdot \begin{bmatrix} \lambda \\ \lambda \\ \lambda \\ \lambda \\ \vdots \end{bmatrix}. \tag{49}$$

Combining this result with the inverse of this infinite matrix gives us

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots \\ -1 & 2 & 0 & 0 & \dots \\ 1 & -3 & 3 & 0 & \dots \\ -1 & 4 & -6 & 4 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}^{-1} \cdot \begin{bmatrix} \dot{C}^{(1)} \\ \dot{C}^{(2)} \\ \dot{C}^{(3)} \\ \dot{C}^{(4)} \\ \vdots \end{bmatrix} = \begin{bmatrix} \lambda \\ \lambda \\ \lambda \\ \lambda \\ \vdots \end{bmatrix} - \begin{bmatrix} C^{(1)}\delta_c [Q] \\ C^{(2)}\delta_c [Q] \\ C^{(3)}\delta_c [Q] \\ C^{(4)}\delta_c [Q] \\ \vdots \end{bmatrix}. \tag{50}$$

Observing that

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ -1 & 2 & 0 & 0 & 0 & 0 & \dots \\ 1 & -3 & 3 & 0 & 0 & 0 & \dots \\ -1 & 4 & -6 & 4 & 0 & 0 & \dots \\ 1 & -5 & 10 & -10 & 5 & 0 & \dots \\ -1 & 6 & -15 & 20 & -15 & 6 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & \dots \\ 1/6 & 1/2 & 1/3 & 0 & 0 & 0 & \dots \\ 0 & 1/4 & 1/2 & 1/4 & 0 & 0 & \dots \\ -1/30 & 0 & 1/3 & 1/2 & 1/5 & 0 & \dots \\ 0 & -1/12 & 0 & 5/12 & 1/2 & 1/6 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

gives us the equations. □

We can write the first three of these equations more explicitly as

$$\dot{E}Q = \lambda - E\delta(Q, c), \tag{51}$$

$$\frac{\dot{E}Q + \text{Var}\dot{Q}}{2} = \lambda - \text{Cov}[Q, \delta(Q, c)], \tag{52}$$

and

$$\frac{\dot{E}Q}{6} + \frac{\text{Var}\dot{Q}}{2} + \frac{C^{(3)}Q}{3} = \lambda - \text{Cov}[(Q - EQ)^2, \delta(Q, c)]. \tag{53}$$

From a computational perspective, we want the ensemble of formulas for the time derivatives of the mean, variance, and third cumulant moment, as summarized in (51)–(53), to be an *autonomous* set of differential equations. This means that their current

behavior should be some integral functional of their past behavior. We can achieve this by making a *closure approximation* in the same spirit as Rothkopf and Oren [56]. The philosophy that they give for this technique is as follows (see page 524 of Rothkopf and Oren [56]):

... The basic strategy of a closure technique is to reduce an infinite system of equations to a finite system by making a “closure assumption” in the form of a functional relationship between the variables of the system.

Similar techniques for approximating non-stationary (dynamic rate) queueing models are also used in Taaffe and Ong [60], Clark [2], Ingolfsson et al. [17], Taaffe and Clark [59], Schwarz et al. [57], Pender [46].

In general, we start by assuming that our underlying closure distribution for the queueing process is uniquely defined by a finite set of parameters. We assume that these parameters are uniquely defined by the same number of expectations of some distinct functions of the queueing process. The resulting approximation of the corresponding forward equations for these functional expectations now forms a finite-dimensional dynamical system for these parameters. Whereas Rothkopf and Oren [56] and Taaffe and Ong [60] assume an underlying discrete distribution for their closure assumptions, our underlying distribution is continuous and is based on polynomials of Gaussian random variables. Our choice of distributions is based on the diffusion limit that arises from the Halfin–Whitt asymptotic scalings.

4.2 Deterministic mean approximation is the fluid limit

Using our closure methodology, we can define a *deterministic mean approximation* (DMA) for our queueing model by assuming that some underlying deterministic process $q \equiv \{q(t) | t \geq 0\}$ approximates our Markovian queueing process, or $Q \approx q$. If we replace Q by q in the Kolmogorov forward equation for the mean of Q as given by (51), then q solves the resulting one-dimensional dynamical system:

$$\dot{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^+, \tag{54}$$

where we set $Q(0) = q(0)$. This method, however, takes us right back to the fluid limit given by (4). We can also write this fluid equation as

$$\dot{q} = \begin{cases} \lambda - \mu \cdot q & \text{when } q < c, \\ \lambda + (\beta - \mu) \cdot c - \beta \cdot q & \text{when } q \geq c. \end{cases} \tag{55}$$

The left-hand graph of Fig. 1 is the plot of both a *single* sample path of DMA or the fluid limit against an averaging of 10,000 sample paths from a stochastic simulation of the underlying Markovian queueing model. Given the trade-off in the number of simulation runs, DMA works well as an approximation. However, the gap between DMA and the simulation average shows that there is room for improvement. DMA consistently underestimates the average queue length behavior. This can be explained

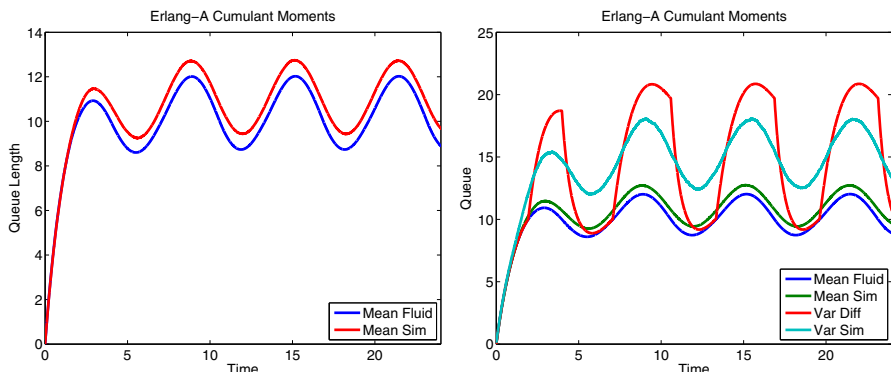


Fig. 1 $\lambda(t) = 10 + 2 \cdot \sin t, \mu = 1, \beta = 0.5, Q(0) = 100, c = 10$

by Jensen’s inequality in combination with the fact that the minimum function is concave and the service rate is larger than the abandonment rate.

The right-hand graph of Fig. 1 shows that the variance of the diffusion limit is not a good estimator of the variance of the queue length process. This inaccuracy of the fluid and diffusion limits motivated the method first introduced by Ko and Gautam [21] and extended by Massey and Pender [34]. This technique is discussed in a later subsection.

First, consider a visualization of the DMA dynamical system for the special case of $\lambda(t) = a + \hat{a} \cdot \cos bt$ and $\beta = \mu$. Our Erlang-A system now reduces to the case of an $M/M/\infty$ queue, which was extensively analyzed by Eick et al. [5,6], Massey and Whitt [35], McCalla and Whitt [37]. Note that here, the mean behavior of the infinite server queue is a one-dimensional dynamical system that equals its fluid limit. Moreover, like a Brownian bridge, the transient behavior of this system captures the dynamics of the corresponding Erlang-A system for the distinct cases of $q < c$ and $q \geq c$.

Theorem 5 *If $q(t)$ is the mean for an $M(t)/M/\infty$ queue, then the phase space pair $(q(t), \dot{q}(t))$ satisfies the quadratic relation*

$$b^2 \cdot \left(q(t) - \frac{a}{\mu} - x \cdot e^{-\mu t} \right)^2 + \left(\dot{q}(t) - y \cdot e^{-\mu t} \right)^2 = \frac{\hat{a}^2 b^2}{\mu^2 + b^2}. \tag{56}$$

This is an ellipse of extremal radii r^* and r_* , where

$$r^* = \frac{\hat{a}b}{\sqrt{\mu^2 + b^2}} \quad \text{and} \quad r_* = \frac{\hat{a}}{\sqrt{\mu^2 + b^2}}, \tag{57}$$

that is centered at $(a/\mu + x \cdot e^{-\mu t}, y \cdot e^{-\mu t})$, where

$$x = q(0) - \frac{a}{\mu} - \frac{\hat{a}\mu}{\mu^2 + b^2} \quad \text{and} \quad y = \dot{q}(0) - \frac{\hat{a}b^2}{\mu^2 + b^2}. \tag{58}$$

Note that for this example, the phase space dynamics of the mean behavior forms an ellipse of a fixed shape and size. Moreover, the dynamics of the centerpoint converge to a limit point. The resulting ellipse is then the limit cycle of the system as discussed in Horne et al. [16]. This cycle is a geometric summary of results first presented in Eick et al. [5].

Proof First, the solution for $\dot{q} = \lambda - \mu \cdot q$ is

$$q(t) = q(0) \cdot e^{-\mu t} + \int_0^t \lambda(s) \cdot e^{-\mu \cdot (t-s)} ds = q(0) \cdot e^{-\mu t} + \int_0^t \lambda(t-s) \cdot e^{-\mu s} ds. \tag{59}$$

Using complex numbers, we can rewrite the arrival rate function as

$$\lambda(t) = a + \hat{a} \cdot \cos bt = a + \langle e^{ibt}, \hat{a} \rangle, \tag{60}$$

where the angle bracket expression is the *inner product* for the complex numbers, or

$$\langle z, w \rangle \equiv \text{Re}(z \cdot \bar{w}) \tag{61}$$

for all complex z and w , where \bar{w} is a complex conjugate of w .

Now, we have

$$q(t) = q(0) \cdot e^{-\mu t} + \int_0^t \left(a + \hat{a} \cdot \langle e^{ibs}, 1 \rangle \right) \cdot e^{-\mu(t-s)} ds \tag{62}$$

$$= q(0) \cdot e^{-\mu t} + a \cdot \int_0^t e^{-\mu(t-s)} ds + \hat{a} \cdot e^{-\mu t} \left\langle \int_0^t e^{(\mu+ib)s} ds, 1 \right\rangle \tag{63}$$

$$= q(0) \cdot e^{-\mu t} + \frac{a}{\mu} \cdot (1 - e^{-\mu t}) + \hat{a} \cdot e^{-\mu t} \left\langle \frac{e^{(\mu+ib)t} - 1}{\mu + ib}, 1 \right\rangle \tag{64}$$

$$= \frac{a}{\mu} + \left(q(0) - \frac{a}{\mu} - \frac{\hat{a}\mu}{\mu^2 + b^2} \right) \cdot e^{-\mu t} + \frac{\hat{a}}{\mu^2 + b^2} \langle e^{ibt}, \mu + ib \rangle. \tag{65}$$

Now, if we take the time derivative of the second integral representation of $q(t)$, then we have

$$\dot{q}(t) = -\mu \cdot q(0) \cdot e^{-\mu t} + \lambda(0) \cdot e^{-\mu t} + \int_0^t \hat{a} b \cdot \langle i e^{ib \cdot (t-s)}, 1 \rangle \cdot e^{-\mu s} ds \tag{66}$$

$$= (\lambda(0) - \mu \cdot q(0)) \cdot e^{-\mu t} + \hat{a} b \cdot \left\langle i e^{ibt} \cdot \int_0^t e^{-(\mu+ib) \cdot s} ds, 1 \right\rangle \tag{67}$$

$$= \dot{q}(0) \cdot e^{-\mu t} + \hat{a} b \cdot \left\langle i \cdot \frac{e^{ibt} - e^{-\mu t}}{\mu^2 + b^2}, \mu + ib \right\rangle \tag{68}$$

$$= \left(\dot{q}(0) - \frac{\hat{a} b^2}{\mu^2 + b^2} \right) \cdot e^{-\mu t} + \frac{\hat{a} b}{\mu^2 + b^2} \langle i e^{ibt}, \mu + ib \rangle. \tag{69}$$

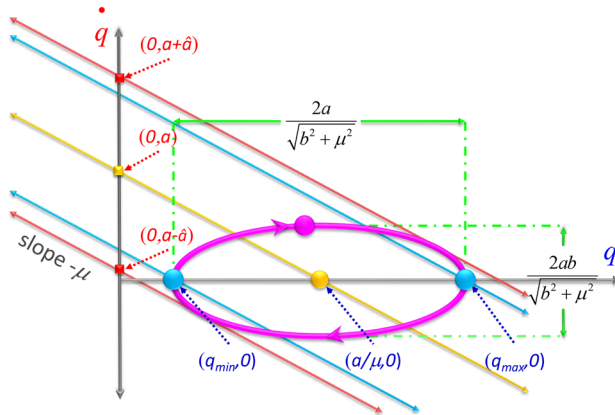


Fig. 2 Plot of $M(t)/M/\infty$ queue limiting ellipse

Finally, we obtain our quadratic relation by observing that

$$\langle e^{i\theta}, z \rangle^2 + \langle i \cdot e^{i\theta}, z \rangle^2 = |z|^2. \tag{70}$$

□

4.3 2D and 3D dynamic analytics

Figure 2 shows a two-dimensional picture of the limiting ellipse generated from the infinite server queue in phase space. The x -axis represents the queue length and the y -axis represents the time derivative of the queue length.

Figure 3 shows a three-dimensional plot of the Erlang-A queue in three different settings. The x -axis represents the queue length, the y -axis represents the time derivative of the queue length, and the z -axis represents time. We plot three different settings of the Erlang-A queue. The first setting in blue is the impatient case where the abandonment rate β is twice that of the service rate μ ; this represents the setting where customers are more impatient. The second setting is when the abandonment rate is equal to the service rate, thus yielding an infinite server queue. The third and last setting is where the abandonment rate is half that of the service rate; this represents the setting where customers are more patient. In Fig. 3, we observe an ordering of the queue length processes, where we see that when customers are more patient, the higher the queue length is going to be since more of them do not leave. Thus, we see the periodic behavior of an Erlang-A queue with more patient customers can have a larger amplitude.

Figure 3 plots the maximum and minimum queue lengths when $\dot{q}(t)$ equals zero. This reveals a recurring theme in the dynamics of queues with dynamic rates. It shows that the times of a locally peak queue length always lag behind the times of a locally peak arrival rate. A similar lag occurs for the times of local minimums.

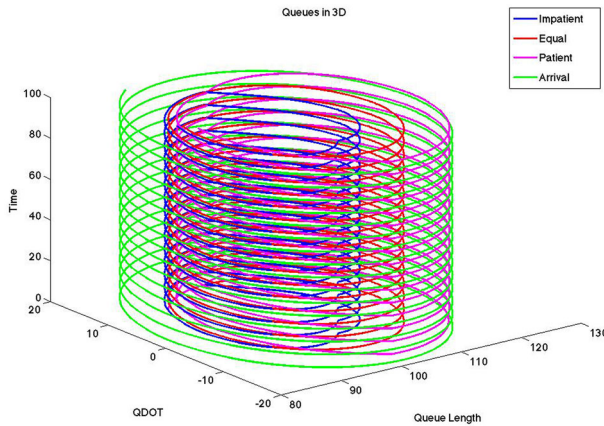


Fig. 3 Erlang-A queue 3D cylinder $(t, \frac{d}{dt}q(t), q(t))$. $\lambda(t) = 100 + 20 \cdot \sin t$, $\mu = 1$, $Q(0) = 80$, $\beta \in \{0.5, 1, 2\}$

Moreover, we observe a bit of symmetry in the case where the abandonment rate and the service rate are equal. In this case, the queue length in steady state is an ellipse. However, when the abandonment and the service rate are not equal, the dynamics are equal to an ellipse since there are different dynamics above and below the number of servers.

4.4 Gaussian variance approximation

Let us now extend this closure method to the case of a two-dimensional dynamical system. Inspired by our diffusion limit typically being Gaussian, consider a dynamical system $\{q(t), v(t) \mid t \geq 0\}$, where v is a strictly positive for all $t > 0$, such that

$$Q \approx \overset{d}{\approx} q + G \cdot \sqrt{v}, \tag{71}$$

where G is a *standard* (mean zero, unit variance) Gaussian random variable.

Such a condition would then give us both $q \approx EQ$ and $v \approx \text{Var}Q$. This two-dimensional closure approximation should give us some insight into the distribution of Q , or more formally

$$P\{Q \geq c\} \approx P\{G \geq \chi\}, \quad \text{where } \chi \equiv \frac{c - q}{\sqrt{v}} \tag{72}$$

for all $t \geq 0$. We also define φ and Φ to be the *density* and the *cumulative distribution* functions, for G , respectively, where

$$\begin{aligned} \varphi(x) &\equiv \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, & \Phi(x) &\equiv \int_{-\infty}^x \varphi(y) dy, \\ \text{and } \bar{\Phi}(x) &\equiv 1 - \Phi(x) = \int_x^{\infty} \varphi(y) dy. \end{aligned} \tag{73}$$

Making the substitutions of q , v , and G into the forward equations for the mean and the variance of Q , i.e., (51) and (52), the dynamical system for q and v is then the set of differential equations

$$\dot{q} = \lambda - \mu \cdot q - E\delta(G, \chi) \cdot \sqrt{v} \tag{74}$$

and

$$\frac{\dot{q} + \dot{v}}{2} = \lambda - \text{Cov}[G, \delta(G, \chi)] \cdot v. \tag{75}$$

We can call the resulting two-dimensional dynamical system the *Gaussian variance approximation (GVA)*.

We can solve these equations numerically if we can easily evaluate these expectation and covariance terms involving functions of the standard Gaussian random variable G . We can show that we can express these quantities in terms of *generic* functions such as the Gaussian density and lookup tables for the *error* function or the cumulative distribution function of a standard Gaussian distribution. If such functions are applied to χ , then they are generic functions of q and v .

We can simplify these Gaussian functional expectation terms by using the following lemma:

Lemma 1 (Stein [58]) *The random variable G is Gaussian(0, 1) if and only if*

$$E[G \cdot f(G)] = E\left[\frac{d}{dG} f(G)\right], \tag{76}$$

for all generalized functions f .

In this framework, the derivative of the indicator function $\{G \geq \chi\}$ (which is a unit step function of the value χ) is a generalized function and is the unit point mass measure at χ . As a result, Stein’s lemma gives us

$$E[G \cdot \{G \geq \chi\}] = \varphi(\chi). \tag{77}$$

Moreover, since $(G - \chi)^+ = G - G \wedge \chi$, we have

$$E[G \wedge \chi] = -E(G - \chi)^+ = \chi \cdot \bar{\Phi}(\chi) - \varphi(\chi). \tag{78}$$

Since $(G - \chi)^+ = (G - \chi) \cdot \{G \geq \chi\}$, then we also have

$$E(G - \chi)^+ = E[G \cdot \{G \geq \chi\}] - \chi \cdot P\{G \geq \chi\} = \varphi(\chi) - \chi \cdot \bar{\Phi}(\chi). \tag{79}$$

Since the generalized derivative of the function $x \vee \chi$ is the indicator function of $\{x \leq \chi\}$, then similar arguments give us

$$\text{Cov}[G, G \wedge \chi] = E[G \cdot (G \wedge \chi)] = P\{G \leq \chi\} = \Phi(\chi) \tag{80}$$

and

$$\text{Cov} [G, (G - \chi)^+] = 1 - \text{Cov} [G, G \wedge \chi] = \bar{\Phi}(\chi). \tag{81}$$

These positive covariances are in keeping with the FKG inequality by Fortuin et al. [10]. This theorem states that increasing functions of the same random variable are always positively correlated.

When we make the substitutions into (74) and (75), the GVA dynamical system reduces to

$$\dot{q} = \lambda - \mu \cdot q + (\beta - \mu) \cdot (\chi \cdot \bar{\Phi}(\chi) - \varphi(\chi)) \cdot \sqrt{v} \tag{82}$$

and

$$\frac{\dot{q} + \dot{v}}{2} = \lambda - (\mu \cdot \Phi(\chi) + \beta \cdot \bar{\Phi}(\chi)) \cdot v. \tag{83}$$

These equations are precisely the *g functions* used in Ko and Gautam [21]. The left-hand graph of Fig. 4 shows that GVA yields a better approximation to the simulated mean than DMA. In fact, it almost matches the simulation so well that the plot of three lines can easily be mistaken for a plot of two lines. One reason for the better accuracy is that the GVA method includes the stochastic fluctuations of deviations from the mean that the fluid limit or DMA does not.

The right-hand graph of Fig. 4 shows that the GVA plot is a significant improvement over the diffusion variance approximation. The gap does show that there is room for improvement. This motivated us to construct a new approximation called the *Gaussian Skewness Approximation* (GSA), which we review in the next subsection. Just as GVA is a refinement to the fluid and diffusion limits, GSA is a refinement to GVA.

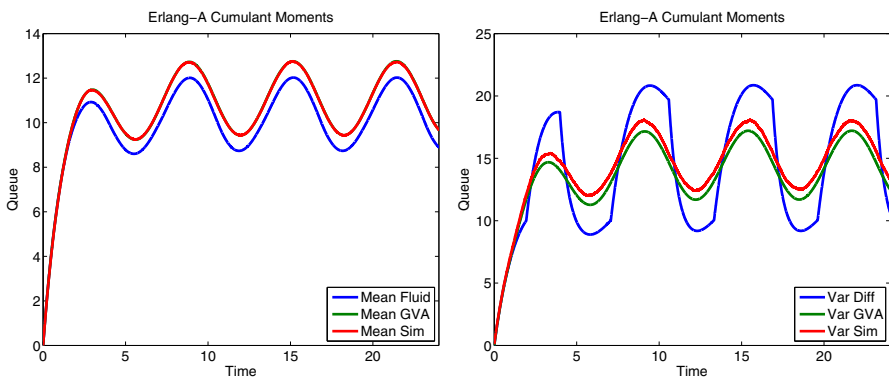


Fig. 4 $\lambda(t) = 10 + 2 \cdot \sin t, \mu = 1, \beta = 0.5, Q(0) = 100, c = 10$

4.5 Gaussian skewness approximation

Now, we extend our closure method to the case of a three-dimensional dynamical system. We want to construct a deterministic system $\{q(t), v(t), s_\theta(t) \mid t \geq 0\}$ that gives us

$$Q \stackrel{d}{\approx} q + H_\theta \cdot \sqrt{v}, \tag{84}$$

with

$$H_\theta \equiv G \cos \theta + \frac{G^2 - 1}{\sqrt{2}} \sin \theta. \tag{85}$$

This should give us $EQ \approx q$, $\text{Var} Q \approx v$, and

$$\text{Skew} Q \approx s_\theta = \text{Skew} H_\theta = \sqrt{2} \cdot (3 - \sin^2 \theta) \cdot \sin \theta, \tag{86}$$

or equivalently

$$\sin \theta = 2 \cdot \sin \left(\frac{1}{3} \sin^{-1} \left(\frac{s_\theta}{2\sqrt{2}} \right) \right), \tag{87}$$

and finally

$$P \{Q \geq c\} \approx P \{H_\theta \geq \chi\}. \tag{88}$$

Consider the Hilbert space of square integrable random variables whose inner product is the expectation of the product of the two random variables. In this space, Hermite polynomials applied to G form a *complete* orthogonal family of random variables. This is the inspiration for the choice of $(G^2 - 1)/\sqrt{2}$. This random variable along with 1 and G form an orthonormal basis for a three-dimensional subspace. Our representation of this subspace corresponds to *cylindrical coordinates*. Here, q parameterizes the z -“coordinate” which maps to the component vector 1. The remaining (x, y) -“plane” maps to the basis vectors G and $(G^2 - 1)/\sqrt{2}$. However, we parameterize this plane in *polar coordinates* where \sqrt{v} is the “radius” with “angle” θ . Finally, there is an invertible mapping between s_θ and $\sin \theta$. This gives us the G and $G^2 - 1$ components of $\sqrt{v} \cos \theta$ and $\sqrt{v} \sin \theta$, respectively.

Skewness is invariant with respect to deterministic translations and positive scalings of the underlying random variable. Hence, it captures an intrinsic property of the distribution. For example, any Gaussian random variable (regardless of mean or variance) has the same skewness (or kurtosis) as the standard Gaussian distribution for G , which is zero. Moreover, a Gaussian random variable is uniquely characterized among analytic random variables by having its third and all higher degree cumulant moments all equal to zero. Thus, we can use skewness informally as a metric that determines how *close* a random variable is to being Gaussian. The number of customers in a system is always a positive number. This ultimately limits the success of *any* Gaussian

approximation to a queueing process. Using H_θ , we can go beyond this restriction. Finally, we have an approximating distribution for the queueing process where q , v , and s_θ arise as independent variables and, respectively, as the *mean*, *variance*, and *skewness* of this distribution.

In Massey and Pender [34], we show that our three-dimensional closure approximation leads to the dynamical system

$$\dot{q} = \lambda - \mu \cdot q - E\delta(H_\theta, \chi) \cdot \sqrt{v}, \tag{89}$$

$$\frac{\dot{q} + \dot{v}}{2} = \lambda - \text{Cov}[H_\theta, \delta(H_\theta, \chi)] \cdot v, \tag{90}$$

and

$$\frac{\dot{q}}{6} + \frac{\dot{v}}{2} + \frac{(s_\theta \cdot \sqrt{v^3})}{3} = \lambda - \text{Cov}[H_\theta^2, \delta(H_\theta, \chi)] \cdot \sqrt{v^3}. \tag{91}$$

We call this the *Gaussian skewness approximation (GSA)*.

The next couple of figures explore the effectiveness of the GSA method, by comparing it to simulation, GVA, as well as to the fluid and diffusion limits. The left-hand graph of Figure 5 shows that GSA is as good an estimator of the mean as GVA. However, the right graph of Fig. 5 shows that GSA is a better estimator of the simulated variance for the queue length process. By incorporating the skewness of the queue length distribution, we are better able to capture the non-Gaussian dynamics of the queue length process. The left-hand graph of Figure 6 shows that the GSA method captures the simulated skewness of the queue length distribution. Note that we have *positive* skewness. This implies that extreme values of the queue length are more likely to be positive and the mean is larger than the median. The right-hand graph of Fig. 6 shows that the mean queue length is larger than the median queue length.

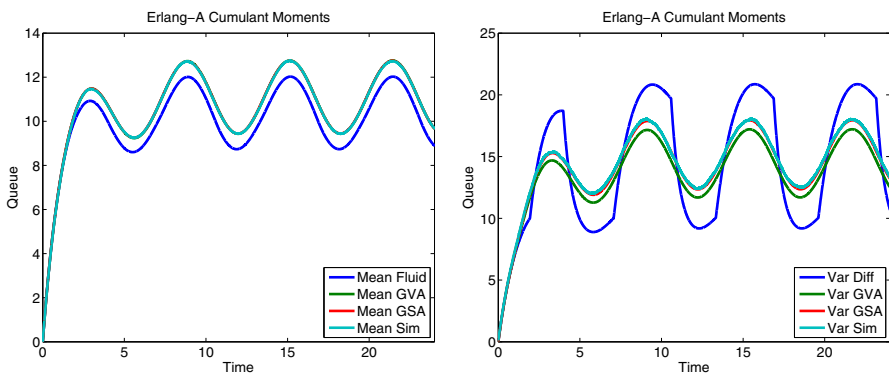


Fig. 5 $\lambda(t) = 10 + 2 \cdot \sin t, \mu = 1, \beta = 0.5, Q(0) = 100, c = 10$

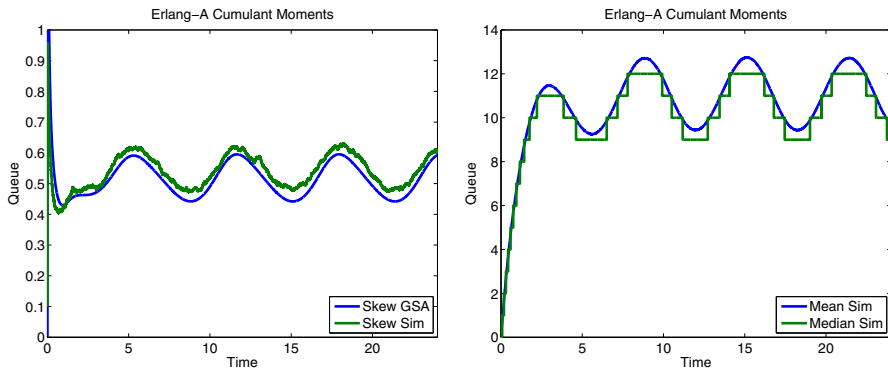


Fig. 6 $\lambda(t) = 10 + 2 \cdot \sin t$, $\mu = 1$, $\beta = 0.5$, $Q(0) = 100$, $c = 10$

4.6 Related closure approximations

Although we describe a hierarchy of successively better closure approximations for the Erlang-A queue, it is important to highlight other closure approximation methods that exist. The first method that we mention here is that of Pender [40]. The major insight of this method was to approximate the queue length density with a Gram-Charlier series expansion and a Gaussian surrogate distribution. Using this method, Pender [40] was able to demonstrate that one can view the Gram-Charlier series as a perturbation of the GVA method that incorporates higher cumulant moments linearly. This method was also applied successfully to dynamic rate Erlang-loss queues with abandonment in Pender [43].

The Gram-Charlier method works well in practice but it is not satisfying theoretically. The difficulty in proving error bounds for the methodology follows from the Gram-Charlier series (or any continuous distribution) approach not approximating the discreteness of the queue length process itself. However, an initial idea of Pender [42], inspired by the infinite server queue, approximates the queue length process with a Poisson distribution.

However, rigorous error bounds of this method in Pender [42] are difficult to obtain since it uses a process approximation idea instead of a density approach. Engblom and Pender [7], however, using Poisson-Charlier polynomial expansions and weighted Sobolev spaces, derive a new discrete closure approximation method based on a density approximation with rigorous error bounds. This is the first closure approximation method that guarantees error bounds for all moments as a function of the number of terms that are used in the approximation.

5 Performance analysis

In this section, we analyze various performance measures of the Erlang-A queue using the closure approximations we have developed. However, before we begin, we provide a 3D plot of the Erlang-A queue to understand its behavior in a variety of settings.

5.1 Delay analysis

The Erlang-A queue is a delay queueing model. This means arriving customers that discover all the service agents are busy must wait in the buffer until one of them becomes available. It follows that our Erlang-A queueing analysis should ultimately lead us to a queueing *delay* analysis, where we assume that the queueing service discipline is *first-come, first-serve* (FCFS). A dynamic queueing rate perspective provides us a simple way to make this transition.

Given a time τ , let $Q_\tau \equiv \{Q_\tau(t) | t \geq 0\}$ be an Erlang-A queueing process, where the only change to the original model is that the arrival process after time τ has been set to zero or *turned off*. This approach was used in Mandelbaum et al. [30] and corresponds to a *virtual* customer arriving at time τ . According to the FCFS service discipline, only the customer still in the queue right before this time influences the delay of this customer. Arrivals after time τ have *no* effect on the delay for this customer.

More precisely, our Poisson arrival rate process is now $\{\lambda_\tau(t) | t \geq 0\}$ and equals

$$\lambda_\tau(t) \equiv \lambda(t) \cdot \{t \leq \tau\}, \tag{92}$$

where the second factor denotes an indicator function that equals 1 if $t \leq \tau$ and 0 otherwise. The sample path construction of Q_τ reduces to

$$Q_\tau(t) = Q_\tau(\tau) - \Pi \left(\int_\tau^t \delta(Q_\tau(s), c) \, ds \right). \tag{93}$$

It follows that the process $Q_\tau(t)$ has *decreasing* sample paths. We can then define the *virtual delay process* to be $D_\tau \equiv \{D_\tau | \tau \geq 0\}$:

$$D_\tau \equiv \min \{t | Q_\tau(\tau + t) < c\}. \tag{94}$$

Assuming the FCFS discipline, D_τ is the time that an incoming customer arriving at time τ has to wait to get service from an agent. Note that, for all positive t , we have

$$\{Q_\tau(\tau + t) \geq c\} = \{D_\tau > t\}. \tag{95}$$

Note that the indicator event for the delay is a *strict* inequality, while on the left of Eq. (95) it is not. There is a nonzero probability that a customer arriving a time τ experiences no delay at all. Hence, the delay distribution always has a point mass at zero.

Taking expectations on both sides of Eq. (95) gives us the following probabilities:

$$P \{Q_\tau(\tau + t) \geq c\} = P \{D_\tau > t\}. \tag{96}$$

After a virtual customer arrives at time τ , we then set the arrival rate to zero. This reduces our notion of uniform acceleration to only scaling up the total number of

service agents. Hence, the *uniformly accelerated* version of virtual delay time process is defined to be

$$D_\tau[\eta c] = \min \{t \mid Q_\tau[\eta c](\tau + t) < \eta c \}. \tag{97}$$

Our key result below then shows how we can approximate the mean virtual customer delay by using the fluid limit for the queueing process. The final result we call the *fluid limit delay* for a virtual customer that arrives at some specific time.

Theorem 6 *The fluid limit delay equals*

$$\lim_{\eta \rightarrow \infty} D_\tau[\eta c] = d_\tau \text{ a.s.}, \tag{98}$$

where

$$d_\tau \equiv \min \{t \mid q_\tau(\tau + t) \leq c \}, \tag{99}$$

and

$$d_\tau = \frac{1}{\beta} \log \left(1 + \frac{\beta \cdot (q_\tau(\tau) - c)^+}{\mu \cdot c} \right). \tag{100}$$

Proof Observe that

$$D_\tau[\eta c] = \min \left\{ t \mid \frac{Q_\tau[\eta c](\tau + t)}{\eta} < c \right\}. \tag{101}$$

If $q_\tau(\tau)$ is smaller than the number of servers c , then the delay time is equal to zero. This agrees with our formula given in Eq. (100). Thus, it only remains to show the proof for the case when $q_\tau(\tau)$ is greater than the number of servers c . The queue length satisfies the following differential equation:

$$\dot{q}_\tau(\tau + t) = -\mu \cdot c - \beta \cdot (q_\tau(\tau + t) - c) = (\beta - \mu) \cdot c - \beta \cdot q_\tau(\tau + t). \tag{102}$$

Since the differential equation is linear, we can solve it explicitly. The solution to the differential equation is given by

$$q_\tau(\tau + t) = \frac{(\beta - \mu) \cdot c}{\beta} \cdot (1 - e^{-\beta t}) + q_\tau(\tau) \cdot e^{-\beta t}. \tag{103}$$

Now, we set the solution of the differential equation equal to c , where the customer has the opportunity to receive service

$$c = \frac{(\beta - \mu) \cdot c}{\beta} \cdot (1 - e^{-\beta \cdot d_\tau}) + q_\tau(\tau) \cdot e^{-\beta \cdot d_\tau}. \tag{104}$$

Table 1 Mean delay comparisons

$\lambda(t) = 10.0 + 2.0 \cdot \sin t$,
 $\mu = 1.0, \beta = 0.5, Q(0) = 0,$
 $c = 10$

τ	Fluid approx	GVA	GSA	Simulation
7.0	0.0	0.0590	0.0620	0.1298
8.0	0.1336	0.2140	0.2140	0.2102
9.0	0.1910	0.2680	0.2670	0.2456
10.0	0.1013	0.1790	0.1770	0.1956
11.0	0.0	0.0180	0.0170	0.1252

Solving for d_τ , we finally have the following solution to the delay time:

$$d_\tau = \frac{1}{\beta} \log \left(1 + \frac{\beta \cdot (q_\tau(\tau) - c)}{\mu \cdot c} \right), \tag{105}$$

when $q_\tau(\tau) > c$, otherwise $d_\tau = 0$. This completes the proof. □

Similar to the fluid delay approximation, we can also make a similar approximation for the GVA and GSA methods. In the case of GVA, the delay time can be approximated by

$$d_\tau^{\text{GVA}} \equiv \min \left\{ t \mid q_\tau^{\text{GVA}}(\tau + t) \leq c \right\}. \tag{106}$$

For GSA, we have

$$d_\tau^{\text{GSA}} \equiv \min \left\{ t \mid q_\tau^{\text{GSA}}(\tau + t) \leq c \right\}. \tag{107}$$

The fluid limit delay has a simple formula so it can easily be compared with many other estimators of the mean delay. Moreover, we have numerical examples showing that GVA and GSA are better approximations for the dynamics of the queueing mean and variance. This suggests that a delay approximation based on the GVA and GSA methods should be more accurate. The graphs of Table 1 show that GVA and GSA are better estimators of the mean delay time.

Moreover, Table 1 shows that the fluid limit delay consistently underestimates the mean delay time. This holds since the fluid mean underestimates the true mean queue length process for this example. However, in Table 2, the fluid delay approximation is better than the one in Table 1. This follows from both the demand (arrival rate) and supply (number of service agents) being increased ten fold. This larger scaling plays to the strength of the fluid approximation.

5.2 Computing the probability of delay with diffusion limits, GVA and GSA

Now, we compare the usefulness of approximating the probability of delay using the fluid and diffusion limits, GVA or GSA for the Erlang-A queue. This probability is defined as the probability that the queue length exceeds or is equal to the number of

Table 2 Mean delay comparisons

$\lambda(t) = 100.0 + 20.0 \cdot \sin t$,
 $\mu = 1.0, \beta = 0.5, Q(0) = 0,$
 $c = 100$

τ	Fluid approx	GVA	GSA	Simulation
7.0	0.0	0.0	0.0010	0.0426
8.0	0.1336	0.1480	0.1480	0.1377
9.0	0.1910	0.2020	0.2020	0.1938
10.0	0.1013	0.1120	0.1120	0.1133
11.0	0.0	0.0	0.0	0.0248

servers, i.e.,

$$P \{D_t > 0\} = P \{Q(t) \geq c\}. \tag{108}$$

For the fluid and diffusion limits and GVA, we can approximate the probability of delay with the Gaussian tail cdf. This implies that

$$P \{Q(t) \geq c\} \approx P \{G \geq \chi\} = \bar{\Phi}(\chi), \tag{109}$$

where the mean and variance pair q and v is due to either the fluid and diffusion limiting processes or GVA. However, for GSA, we have that

$$P \{Q(t) \geq c\} \approx P \{H_\theta \geq \chi\} = \bar{\Psi}_\theta(\chi), \tag{110}$$

where the mean, variance, and skewness triplet $q, v,$ and s_θ (which gives us $\sin \theta$) is due to GSA and $\bar{\Psi}_\theta$ is the tail distribution for the random variable H_θ . Its cumulative distribution function is denoted by Ψ_θ .

We can evaluate this distribution by solving for the roots of a specific quadratic equation. If $\sin \theta = 0$, then H_θ reduces to a Gaussian random variable. We now assume that $\sin \theta \neq 0$ and observe that

$$\{H_\theta \leq \chi\} = \{z_-(\theta, \chi) \leq G \leq z_+(\theta, \chi)\}, \tag{111}$$

where

$$z_+(\theta, \chi) = \frac{2\chi + \sqrt{2} \cdot \sin \theta}{\cos \theta + \sqrt{1 + 2\chi\sqrt{2} \cdot \sin \theta + \sin^2 \theta}} \tag{112}$$

and

$$z_-(\theta, \chi) = \frac{\cos \theta + \sqrt{1 + 2\chi\sqrt{2} \cdot \sin \theta + \sin^2 \theta}}{-\sqrt{2} \cdot \sin \theta}. \tag{113}$$

This follows from the quadratic formula, which gives us

$$H_\theta - \chi = \frac{\sin \theta}{\sqrt{2}} \cdot (G - z_+(\theta, \chi)) \cdot (G - z_-(\theta, \chi)). \tag{114}$$

Finally, we obtain from equating the two events that

$$\Psi_\theta(\chi) = P \{H_\theta \leq \chi\} = \Phi(z_+(\theta, \chi)) - \Phi(z_-(\theta, \chi)). \tag{115}$$

Since the parameters q , v , and c are independent of θ , then χ can range over all the real numbers.

In the next subsection, we show that we can construct delay stabilizing staffing schedules for the queue length process by inverting our delay probability approximations.

5.3 Stabilizing the probability of delay

We can use this closed-form expression for the probability of delay to construct a staffing algorithm that stabilizes the probability of delay. This notion of stabilizing performance measures was introduced by Ward Whitt and colleagues in the work of Jennings et al. [19]. They produced stable staffing schedules for the multi-server queue with general service and arrival distributions. A follow-up paper [9] constructed a simulation staffing algorithm to stabilize the probability of delay for arbitrary queueing networks. More work on the topic has been also pursued by Ward Whitt, his students and collaborators in Liu and Whitt [25–27], Li et al. [24], Pender [45], He et al. [15], Liu and Whitt [28], Whitt and Zhao [62], Pender and Massey [48].

Recall that both the diffusion limit coupled with the fluid limit and GVA approximate the probability of delay with the Gaussian tail distribution. We can invert this function and get

$$P \{Q(t) \geq c\} = \overline{\Phi}(\chi) = \epsilon \Leftrightarrow c = \left[q + \overline{\Phi}^{-1}(\epsilon) \cdot \sqrt{v} \right]. \tag{116}$$

For GSA, our stabilizing staffing schedule for a target delay probability ϵ is given by

$$P \{Q(t) \geq c\} = \overline{\Psi}_\theta(\chi) = \epsilon \Leftrightarrow c = \left[q + \overline{\Psi}_\theta^{-1}(\epsilon) \cdot \sqrt{v} \right]. \tag{117}$$

Although the first two Gaussian-based staffing procedures appear to be the same, the actual staffing functions are different because the dynamics for the mean and variance of the GVA and GSA methods are better estimates than the fluid and diffusion limits of the true mean and variance.

Our method of stabilization is deterministic. Hence, it bypasses any use of Monte Carlo simulation. Unlike Jennings et al. [19] and Feldman et al. [9], there is no need to actually simulate the queueing system in order to update the staffing schedule. Our method is completely scalable. It only requires the numerical solution of *two* ordinary differential equations for GVA or *three* such equations for GSA *regardless* of the actual

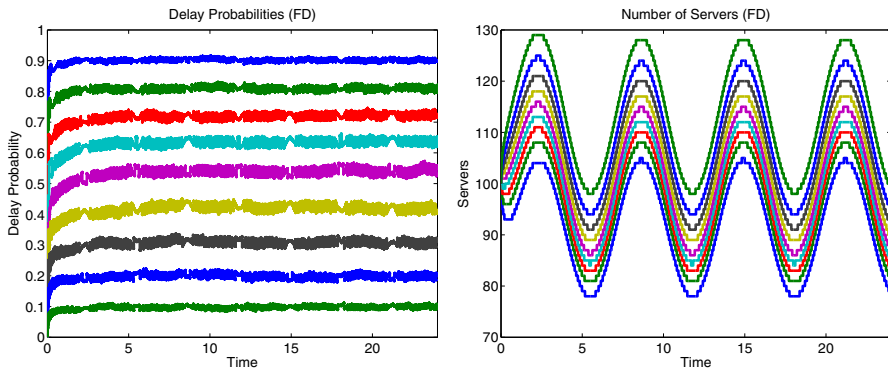


Fig. 7 Stabilizing the probability of delay with (FD) target levels: $\epsilon = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$: stabilized delay probabilities (left), the number of servers (right), $\lambda(t) = 100.0 + 20.0 \cdot \sin t$, $\mu = 1.0$, $\beta = 0.5$, $Q(0) = 100$

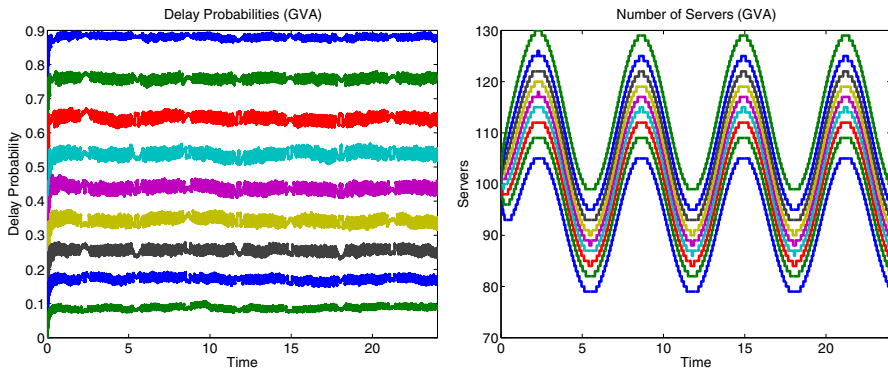


Fig. 8 Stabilizing the probability of delay with (GVA) target levels: $\epsilon = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$: stabilized delay probabilities (left), the number of servers (right), $\lambda(t) = 100.0 + 20.0 \cdot \sin t$, $\mu = 1.0$, $\beta = 0.5$, $Q(0) = 100$

number of servers being modeled. To demonstrate the effectiveness of our algorithms, we plot in Figs. 7, 8 and 9 the delay probabilities produced by our algorithms on the left and the number of servers on the right.

For this example, all three algorithms produce stable delay probabilities. However, GSA is the best at producing the most stable delay probabilities for each of the target values, especially near the value $\epsilon = 0.5$. This value produces the median queue length, and Figure 9 shows that the GSA does a better job of stabilization there (see the middle or magenta colored curve). The median value when the delay probability attains the value $\epsilon = 0.5$ is also where the fluid limit is more likely to “linger” near the number of servers. This is also when the queue length distribution is the *most* non-Gaussian.

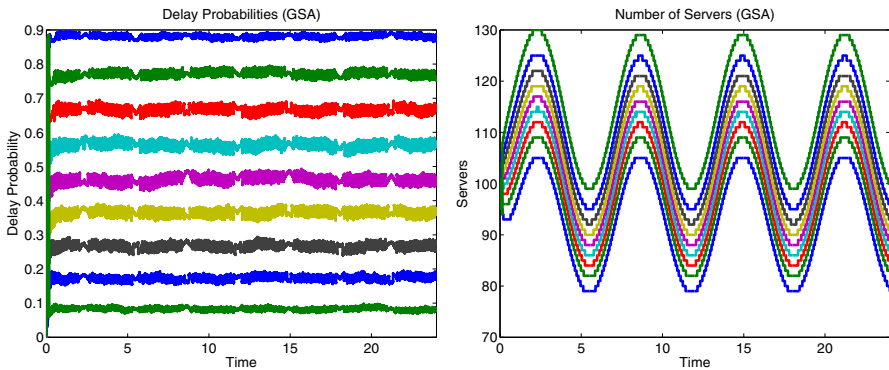


Fig. 9 Stabilizing the probability of delay with (GSA) target levels: $\epsilon = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$: stabilized delay probabilities (left), the number of servers (right), $\lambda(t) = 100.0 + 20.0 \cdot \sin t$, $\mu = 1.0$, $\beta = 0.5$, $Q(0) = 100$

5.4 Static staffing of the Erlang-A

We now discuss a problem that illustrates the utility of the Erlang-A queueing model for emerging healthcare applications. In the spirit of Jennings et al. [18] as well as McCalla and Whitt [37], we use our *dynamic* analysis methods to develop a *static* staffing algorithm for a nursing home during a time interval $(0, T]$. This is in contrast to the work of Niyirora and Pender [38] and Qin and Pender [55], where the optimal staffing algorithms are dynamic and varied over time. In the context of capacity planning in a nursing home, the time scale is larger and it is not practical to change the number of beds or rooms dynamically with time. Our goal is to find an algorithm that determines the optimal number of beds needed in a nursing home to minimize costs or achieve profit optimality.

5.4.1 Lagrangian algorithm to find the optimal number of beds

The fluid model for the mean behavior of an Erlang-A model is defined by the dynamical system $\{q(t) | t \geq 0\}$,

$$\dot{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^+, \tag{118}$$

where q and λ are implicitly time dependent. We use $q[c]$ with *square brackets* (not circular) to stress the fluid model dependency on a fixed total number of service agents c . When c equals zero or infinity, we have

$$\dot{q}[0] = \lambda - \beta \cdot q[0] \quad \text{and} \quad \dot{q}[\infty] = \lambda - \mu \cdot q[\infty]. \tag{119}$$

Now, let r equal the revenue obtained from each successfully served customer, i.e., one that departs as a service completion and not as a customer abandonment. If w equals the cost rate of each agent, then we define $\mathcal{P}[c]$ to equal the mean nursing

home profit over the time interval $(0, T]$, or

$$\mathcal{P}_T[c] \equiv \int_0^T (r\mu \cdot (q \wedge c) - wc) dt. \tag{120}$$

Since $\mathcal{P}_T[0] = 0$ and $\mathcal{P}_T[c] < 0$ as c becomes sufficiently large, then there exists some total number of service agents c_* such that

$$\mathcal{P}_T[c_*] = \max_{c \geq 0} \mathcal{P}_T[c]. \tag{121}$$

We construct an algorithm to find c_* by superimposing a Lagrangian structure, as formulated in Hampshire [12], Hampshire and Massey [13], onto the profit function for fixed c . Observe that, for any fixed c ,

$$\mathcal{P}_T[c] = \max_{q: \dot{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^+} \int_0^T (r\mu \cdot (q \wedge c) - wc) dt. \tag{122}$$

We can rewrite this as a *constrained* optimization problem, i.e.,

$$\mathcal{P}_T[c] = \max_{p, q: \dot{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^+} \int_0^T \mathcal{L}(c, p, q, \dot{q}) dt, \tag{123}$$

where our Lagrangian is

$$\mathcal{L}(c, p, q, \dot{q}) = r\mu \cdot (q \wedge c) - w \cdot c + p \cdot (\dot{q} - \lambda + \mu \cdot (q \wedge c) + \beta \cdot (q - c)^+). \tag{124}$$

The resulting Euler-Lagrange equations are

$$\begin{aligned} \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}}(c, p, q, \dot{q}) &= \frac{\partial \mathcal{L}}{\partial q}(c, p, q, \dot{q}) \Leftrightarrow \dot{p} = \mu \cdot (p + r) \cdot \{q < c\} + \beta \cdot p \cdot \{q \geq c\}, \\ \frac{\partial \mathcal{L}}{\partial p}(c, p, q, \dot{q}) &= 0 \Leftrightarrow \dot{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^+. \end{aligned}$$

We can now rewrite $\mathcal{P}_T[c]$ as

$$\mathcal{P}_T[c] = \int_0^T \mathcal{L}(p, q, \dot{q})[c] dt, \tag{125}$$

where

$$\mathcal{L}(p, q, \dot{q})[c](t) \equiv \mathcal{L}(c, p[c](t), q[c](t), \dot{q}[c](t), t), \tag{126}$$

and the Lagrange multiplier $p[c]$ solves the dynamical system

$$\dot{p}[c] = \mu \cdot (p[c] + r) \cdot \{q[c] < c\} + \beta \cdot p[c] \cdot \{q[c] \geq c\} \tag{127}$$

with terminal condition $p[c](T) = 0$. For the next result, we suppress the explicit dependence on c .

Theorem 7 *If p is the opportunity cost process for the Erlang-A fluid profit function, then we have*

$$-r < p(t) \leq 0 \tag{128}$$

for all $0 \leq t \leq T$, with $p(T) = 0$. Moreover, if $p(t) = 0$ for some $0 \leq t < T$ then $p = 0$ and $q \geq c$ over the entire closed interval $[t, T]$.

Proof If $p(t) \leq -r$ for some $0 \leq t \leq T$, then $\dot{p}(t) \leq 0$. This forces p to be a decreasing function on the interval $(0, T]$ bounded above by $-r$. This rules out the possibility of $p(T) = 0$; hence, the premise is false.

Similarly, the assumption of $p(t) > 0$ for some $0 \leq t \leq T$ also leads to a contradiction. Hence, $p(t) = 0$ for such a t holds only if $p = 0$ and $q \geq c$ on the entire interval $[t, T]$. □

Given some real-valued function f on $(0, T]$, we define the *decreasing rearrangement* of f to be the unique right continuous, decreasing function f^\downarrow on $(0, T]$ such that

$$\int_0^T \{f(t) > x\} dt = \int_0^T \{f^\downarrow(t) > x\} dt \tag{129}$$

for all real values x . For all decreasing functions g , we define its *generalized inverse* to be g^{-1} , where

$$g^{-1}(x) \equiv \inf \{y \mid g(y) \geq x\}. \tag{130}$$

For all x less than some number in the range of f , we have

$$\int_0^T \{f(t) > x\} dt = (f^\downarrow)^{-1}(x). \tag{131}$$

This method was first developed for queueing applications to analyze profit optimality in private telephone line services [18] using an $M/M/\infty$ queueing model. Below we create a new analysis for the Erlang-A queue that is a synthesis of this rearrangement approach with the Lagrangian and Hamiltonian techniques of Hampshire [12], Hampshire and Massey [13].

We now make the following two operational business assumptions that are relevant to nursing homes:

1. We assume that $w/\mu < r$, or $w < r\mu$. We are motivated here by wanting the average amount of money spent on an agent providing service to a customer to be *less* than the revenue received by the departing customer who did *not* abandon.
2. We assume that $1/\beta < 1/\mu$, or $\beta > \mu$. We are motivated here by wanting the average time that a customer wants to spend on a waiting list for service to be *less* than the average amount of time for the actual service.

This leads us to our fundamental result

Theorem 8 *Assume that $w < r\mu$, $\beta > \mu$, $q^\downarrow[c]$ is the decreasing rearrangement of q over the time interval $(0, T]$, and*

$$\bar{c} \equiv q^\downarrow[c] \left(\frac{w}{r\mu} T + \frac{\beta - \mu}{r\mu} \int_0^{q^\downarrow[c]^{-1}(c)} p^\downarrow[c] dt \right), \tag{132}$$

where $p^\downarrow[c]$ is defined to be the time-rearrangement of p that transforms q into its decreasing rearrangement q^\downarrow . We then have

$$\mathcal{P}'_T[c] \geq 0 \Rightarrow c \leq \bar{c} \text{ and } c \leq q^\downarrow[c] \left(\frac{w}{r\beta} T \right). \tag{133}$$

Similarly, we also have

$$\mathcal{P}'_T[c] \leq 0 \Rightarrow c \geq \bar{c} \geq q^\downarrow[c] \left(\frac{w}{r\mu} T \right). \tag{134}$$

Proof Using the sensitivity results for Lagrangian optimality, we have

$$\begin{aligned} \mathcal{P}'_T[c] &= \frac{d}{dc} \int_0^T \mathcal{L}(p, q, \dot{q})[c] dt \\ &= \int_0^T \frac{\partial \mathcal{L}}{\partial c}(c, p[c], q[c], \dot{q}[c])[c] dt \\ &= \int_0^T r\mu \cdot \{q[c] > c\} - w + p \cdot (\mu \cdot \{q[c] > c\} - \beta \cdot \{q[c] > c\}) dt \\ &= r\mu \cdot \int_0^T \{q[c] > c\} dt - \left(wT + (\beta - \mu) \cdot \int_0^T p[c] \cdot \{q[c] > c\} dt \right) \\ &= r\mu \cdot \int_0^T \{q^\downarrow[c] > c\} dt - \left(wT + (\beta - \mu) \cdot \int_0^T p[c] \cdot \{q^\downarrow[c] > c\} dt \right) \\ &= r\mu \cdot q^\downarrow[c]^{-1}(c) - \left(wT + (\beta - \mu) \cdot \int_0^{q^\downarrow[c]^{-1}(c)} p^\downarrow[c] dt \right). \end{aligned}$$

This completes the proof. □

Corollary 5 *For any Erlang-A system, the number of service agents c_* needed for profit optimality has the following properties:*

1. We have $c_* = \bar{c}_*$ or

$$c_* = q^\downarrow[c_*] \left(\frac{w}{r\mu} T + \frac{\beta - \mu}{r\mu} \int_0^{q^\downarrow[c_*]^{-1}(c_*)} p^\downarrow[c_*] dt \right). \tag{135}$$

2. We have the following upper and lower bounds for c_* :

$$q^\downarrow[c_*] \left(\frac{w}{r\mu} T \right) \leq c_* \leq q^\downarrow[c_*] \left(\frac{w}{r\beta} T \right). \tag{136}$$

3. Given our rearrangement $p^\downarrow[c_*]$ of $p[c_*]$, there exists a unique τ that solves the equation

$$\tau = \frac{w}{r\mu} T + \frac{\beta - \mu}{r\mu} \int_0^\tau p^\downarrow[c_*] dt. \tag{137}$$

We then have $c_* = q^\downarrow[c_*](\tau)$.

Proof Part 1 and Part 2 follow from the previous theorem. However, for Part 3, we have with $-r < p[c_*] \leq 0$ on $(0, T]$ and $\beta \geq \mu$ so that

$$0 < \frac{w}{r\mu} \cdot T + \frac{\beta - \mu}{r\mu} \cdot \int_0^0 p^\downarrow dt = \frac{w}{r\mu} \cdot T, \tag{138}$$

and

$$\frac{w}{r\mu} \cdot T + \frac{\beta - \mu}{r\mu} \cdot \int_0^T p^\downarrow dt \leq \frac{w}{r\mu} \cdot T < T. \tag{139}$$

This gives us the unique fixed point value τ , where $0 \leq \tau \leq T$, and completes the proof. □

6 Conclusion and final remarks

In this paper, we prove and review several important results for the dynamic rate Erlang-A queue. Although we provide many new results, there are many areas that are still ripe for new research. One interesting area of research would be to explore the possibilities for extending and replicating our closure approximations and cumulant moment results for non-Markovian queueing systems. For example, the many-server limit theorems in the spirit of Halfin and Whitt have been extended to phase-type, general distributions and self-exciting processes as in the work of Ko and Pender [22], Pender and Ko [47], Ko and Pender [23], Daw and Pender [3]. Future work should explore how well these closure approximations work for non-Markovian systems. Moreover, there is new research that explores the new aspect of adding delayed information to drive the arrival processes of queueing networks; see, for example, Pender et al. [50–52].

It would be interesting to extend this work to the dynamic rate Erlang-A queueing model and generalizations of it. We plan to pursue some of this work in the future.

Acknowledgements William A. Massey was partially supported by National Science Foundation Grant CMMI-1436334.

References

1. Choudhury, G.L., Whitt, W.: Heavy-traffic asymptotic expansions for the asymptotic decay rates in the BMAP/G/1 queue. *Stoch. Models* **10**(2), 453–498 (1994)
2. Clark, G.M.: Use of Polya distributions in approximate solutions to nonstationary M/M/s queues. *Commun. ACM* **24**(4), 206–217 (1981)
3. Daw, A., Pender, J.: Queues driven by Hawkes processes. *Stoch. Syst.* (2018) (to appear)
4. Duffield, N.G., Massey, W.A., Whitt, W.: A nonstationary offered-load model for packet networks. *Telecommun. Syst.* **16**(3–4), 271–296 (2001)
5. Eick, S.G., Massey, W.A., Whitt, W.: $M_T/G/\infty$ queues with sinusoidal arrival rates. *Manag. Sci.* **39**(2), 241–252 (1993)
6. Eick, S.G., Massey, W.A., Whitt, W.: The physics of the $M_T/G/\infty$ queue. *Oper. Res.* **41**(4), 731–742 (1993)
7. Engblom, S., Pender, J.: Approximations for the moments of nonstationary and state dependent birth-death queues. arXiv preprint [arXiv:14066164](https://arxiv.org/abs/14066164) (2014)
8. Erlang, A.K.: Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Trans. Dan. Acad. Tech. Sci.* **2**, 138–155 (1948)
9. Feldman, Z., Mandelbaum, A., Massey, W.A., Whitt, W.: Staffing of time-varying queues to achieve time-stable performance. *Manag. Sci.* **54**(2), 324–338 (2008)
10. Fortuin, C.M., Kasteleyn, P.W., Ginibre, J.: Correlation inequalities on some partially ordered sets. *Commun. Math. Phys.* **22**(2), 89–103 (1971)
11. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3), 567–588 (1981)
12. Hampshire, R.C.: Dynamic queueing models for the operations management of communication services. Ph.D. thesis, Princeton University (2007)
13. Hampshire, R.C., Massey, W.A.: Dynamic optimization with applications to dynamic rate queues. In: *Risk and Optimization in an Uncertain World*, INFORMS, pp. 208–247 (2010)
14. Hampshire, R.C., Harchol-Balter, M., Massey, W.A.: Fluid and diffusion limits for transient sojourn times of processor sharing queues with time varying rates. *Queueing Syst.* **53**(1), 19–30 (2006). <https://doi.org/10.1007/s11134-006-7584-x>
15. He, B., Liu, Y., Whitt, W.: Staffing a service system with non-poisson non-stationary arrivals. *Probab. Eng. Inf. Sci.* **30**(4), 593–621 (2016)
16. Horne, R.L., Mandelbaum, A., Massey, W.A.: Performance analysis of dynamic-rate many-server queues. Working notes (2011)
17. Ingolfsson, A., Akhmetshina, E., Budge, S., Li, Y., Wu, X.: A survey and experimental comparison of service-level-approximation methods for nonstationary M(t)/M(s(t)) queueing systems with exhaustive discipline. *INFORMS J. Comput.* **19**(2), 201–214 (2007)
18. Jennings, O., Massey, W., McCalla, C.: Optimal profit for leased lines services. In: *Proceedings of the 15th International Teletraffic Congress-ITC*, vol. 15, pp. 803–814 (1997)
19. Jennings, O.B., Mandelbaum, A., Massey, W.A., Whitt, W.: Server staffing to meet time-varying demand. *Manag. Sci.* **42**(10), 1383–1394 (1996)
20. Khinchin, A.Y., Andrews, D., Quenouille, M.H.: *Mathematical Methods in the Theory of Queueing*. Courier Corporation, North Chelmsford (2013)
21. Ko, Y.M., Gautam, N.: Critically loaded time-varying multiserver queues: computational challenges and approximations. *INFORMS J. Comput.* **25**(2), 285–301 (2013)
22. Ko, Y.M., Pender, J.: Diffusion limits for the $(MAP_t/Ph_t/\infty)^N$ queueing network. *Oper. Res. Lett.* **45**(3), 248–253 (2017)
23. Ko, Y.M., Pender, J.: Strong approximations for time-varying infinite-server queues with non-renewal arrival and service processes. *Stoch. Models* (2018). <https://doi.org/10.1080/15326349.2018.1425886>

24. Li, A., Whitt, W., Zhao, J.: Staffing to stabilize blocking in loss models with time-varying arrival rates. *Probab. Eng. Inf. Sci.* **30**(2), 185–211 (2016)
25. Liu, Y., Whitt, W.: Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* **60**(6), 1551–1564 (2012)
26. Liu, Y., Whitt, W.: Stabilizing performance in many-server queues with time-varying arrivals and customer feedback. Technical report, working paper (2014)
27. Liu, Y., Whitt, W.: Stabilizing performance in networks of queues with time-varying arrival rates. *Probab. Eng. Inf. Sci.* **28**(4), 419–449 (2014)
28. Liu, Y., Whitt, W.: Stabilizing performance in a service system with time-varying arrivals and customer feedback. *Eur. J. Oper. Res.* **256**(2), 473–486 (2017)
29. Mandelbaum, A., Massey, W.A., Reiman, M.I.: Strong approximations for Markovian service networks. *Queueing Syst.* **30**(1), 149–201 (1998)
30. Mandelbaum, A., Massey, W.A., Reiman, M.I., Stolyar, A., Rider, B.: Queue lengths and waiting times for multiserver queues with abandonment and retries. *Telecommun. Syst.* **21**(2), 149–171 (2002)
31. Marcinkiewicz, J.: Sur une propriete de la loi de Gauss. *Math. Z.* **44**, 612–618 (1939)
32. Massey, W.A.: Asymptotic analysis of the time dependent M/M/1 queue. *Math. Oper. Res.* **10**(2), 305–327 (1985)
33. Massey, W.A., Pender, J.: Poster: skewness variance approximation for dynamic rate multiserver queues with abandonment. *ACM SIGMETRICS Perform. Eval. Rev.* **39**(2), 74–74 (2011)
34. Massey, W.A., Pender, J.: Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Syst.* **75**(2–4), 243–277 (2013)
35. Massey, W.A., Whitt, W.: Networks of infinite-server queues with nonstationary poisson input. *Queueing Syst.* **13**(1), 183–250 (1993)
36. Massey, W. A., Whitt, W.: Uniform acceleration expansions for Markov chains with time-varying rates. *Ann. Appl. Probab.* **8**(4), 1130–1155 (1998)
37. McCalla, C., Whitt, W.: A time-dependent queueing-network model to describe the life-cycle dynamics of private-line telecommunication services. *Telecommun. Syst.* **19**(1), 9–38 (2002)
38. Niyirora, J., Pender, J.: Optimal staffing in nonstationary service centers with constraints. *Nav. Res. Logist. NRL* **63**(8), 615–630 (2016)
39. Palm, C.: Intensity variations in telephone traffic. *Ericsson Tech.* **44**, 1–189 (1988). (**English translation by North-Holland, Amsterdam**)
40. Pender, J.: Gram Charlier expansion for time varying multiserver queues with abandonment. *SIAM J. Appl. Math.* **74**(4), 1238–1265 (2014)
41. Pender, J.: Laguerre polynomial expansions for time varying multiserver queues with abandonment. Technical report (2014)
42. Pender, J.: A Poisson-Charlier approximation for nonstationary queues. *Oper. Res. Lett.* **42**(4), 293–298 (2014)
43. Pender, J.: Nonstationary loss queues via cumulant moment approximations. *Prob. Eng. Inf. Sci.* **29**(01), 27–49 (2015)
44. Pender, J.: The truncated normal distribution: applications to queues with impatient customers. *Oper. Res. Lett.* **43**(1), 40–45 (2015)
45. Pender, J.: Risk measures and their application to staffing nonstationary service systems. *Eur. J. Oper. Res.* **254**(1), 113–126 (2016)
46. Pender, J.: Sampling the functional Kolmogorov forward equations for nonstationary queueing networks. *INFORMS J. Comput.* **29**(1), 1–17 (2016)
47. Pender, J., Ko, Y.M.: Approximations for the queue length distributions of time-varying many-server queues. *INFORMS J. Comput.* **29**(4), 688–704 (2017)
48. Pender, J., Massey, W.A.: Approximating and stabilizing dynamic rate Jackson networks with abandonment. *Probab. Eng. Inf. Sci.* **31**(1), 1–42 (2017)
49. Pender, J., Phung-Duc, T.: A law of large numbers for M/M/c/Delayoff-Setup queues with nonstationary arrivals. In: *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, pp. 253–268. Springer (2016)
50. Pender, J., Rand, R.H., Wesson, E.: Queues with choice via delay differential equations. *Int. J. Bifurc. Chaos* **27**(04), 1730.016 (2017)
51. Pender, J., Rand, R.H., Wesson, E.: Strong approximations for queues with customer choice and constant delays (2017)

52. Pender, J., Rand, R.H., Wesson, E.: An analysis of queues with delayed information and time-varying arrival rates. *Nonlinear Dyn.* **91**(4), 2411–2427 (2018)
53. Pender, J.J.: Dynamic rate queues: estimation, stabilization, and control. Ph.D. thesis, Princeton University (2013)
54. Puhalskii, A.A.: On the $M_t/M_t/K_t + M_t$ queue in heavy traffic. *Math. Methods Oper. Res.* **78**(1), 119–148 (2013)
55. Qin, Z., Pender, J.: Dynamic control for nonstationary queueing networks. Working paper (2017)
56. Rothkopf, M.H., Oren, S.S.: A closure approximation for the nonstationary M/M/s queue. *Manag. Sci.* **25**(6), 522–534 (1979)
57. Schwarz, J.A., Selinka, G., Stolletz, R.: Performance analysis of time-dependent queueing systems: survey and classification. *Omega* **63**, 170–189 (2016)
58. Stein, C.: Approximate computation of expectations. *Lect. Notes Monogr. Ser.* **7**, i-164 (1986)
59. Taaffe, M.R., Clark, G.M.: Approximating nonstationary two-priority non-preemptive queueing systems. *Nav. Res. Logist. NRL* **35**(1), 125–145 (1988)
60. Taaffe, M.R., Ong, K.L.: Approximating nonstationary $Ph_t/M_t/s/c$ queueing systems. *Ann. Oper. Res.* **8**(1), 103–116 (1987)
61. Whitt, W.: Approximating a point process by a renewal process, I: two basic methods. *Oper. Res.* **30**(1), 125–147 (1982)
62. Whitt, W., Zhao, J.: Many-server loss models with non-Poisson time-varying arrivals. *Nav. Res. Logist. NRL* (2017). <https://doi.org/10.1002/nav.21741>