doi:10.1016/j.peva.2014.07.025 • www.stefancreemers.be • info@stefancreemers.be

G-RAND: A phase-type approximation for the nonstationary $\mathbf{G}(t)/\mathbf{G}(t)/s(t) + \mathbf{G}(t)$ queue

Stefan Creemers
Mieke Defraeye
Inneke Van Nieuwenhuyse

Abstract - **We present a Markov model to analyze the queueing behavior of the nonstationary $\mathbf{G}(t)/\mathbf{G}(t)/s(t) + \mathbf{G}(t)$ queue. We assume an exhaustive service discipline (where servers complete their current service before leaving) and use acyclic phase-type distributions to approximate the general interarrival, service, and abandonment time distributions. The time-varying performance measures of interest are: (1) the expected number of customers in queue, (2) the variance of the number of customers in queue, (3) the expected number of abandonments, and (4) the virtual waiting time distribution of a customer arriving at an arbitrary moment in time. We refer to our model as G-RAND since it analyzes a general queue using the randomization method. A computational experiment shows that our model allows the accurate analysis of small- to medium-sized problem instances.**

Keywords - **nonstationary arrivals, time-varying demand, Markov model, $\mathbf{G}(t)/\mathbf{G}(t)/s(t) + \mathbf{G}(t)$ queue, performance measurement**

1 Introduction

Many service systems exhibit a cyclic demand for service. For example, in call centers, emergency departments, banks, and retail stores, the number of arrivals typically displays a daily, weekly, or monthly recurring pattern. Figure 1, for instance, shows the daily fluctuations in arrival rate at the emergency department of a regional hospital in Belgium Defraeye and Van Nieuwenhuyse (2013); other examples can be found in Green et al. (2006), Brown et al. (2005), and Dietz (2011), among others. Apart from the time-varying nature of demand, additional complexities may arise because of (1) the presence of customer impatience, which causes customers to abandon before receiving service if their waiting time is too long and (2) the general distribution of service and abandonment times.

The Poisson assumption is commonly used in the literature for the arrival process, the service process, and the abandonment process (Kim and Whitt (2014), Ingolfsson et al. (2007), Whitt (1991), Garnett et al. (2002)). Kim and Whitt (2014) largely justify this assumption for the arrival process, whereas Zeltyn and Mandelbaum (2005) and Hueter and Swart (1998) use empirical data of an emergency department and a restaurant setting to justify the use of an exponential service time distribution. Yet, in many realistic settings, the exponential assumption does not hold for the service and/or abandonment processes. For instance, Brown et al. (2005) report lognormal distributions and Castillo et al. (2009)

report Erlang distributed service times and Mandelbaum and Zeltyn (2013) report good fits for the abandonment time distribution with with Log-Pearson III and generalized gamma distributions.

Moreover, many existing models in the literature implicitly assume a preemptive service discipline, such that service is interrupted and customers rejoin the queue when a server leaves. An exhaustive service policy, where servers complete their current service before leaving, is often more appropriate (especially in service systems with human customers and servers). This feature, however, is frequently overlooked in the literature (Ingolfsson et al., 2007; Chen and Henderson, 2001).

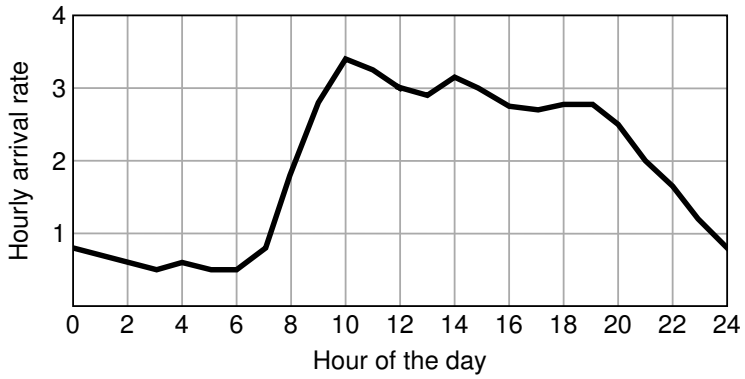


Figure 1: Hourly average arrival rates at the emergency department of a Belgian regional hospital

Capacity planning models rely on a performance evaluation method as a subroutine to assess the solution quality of any given capacity vector. Therefore, performance analysis for systems with time-varying arrivals is highly important when making capacity decisions. We refer to Green et al. (2007), Whitt (2007), and Defraeye and Van Nieuwenhuyse (2011) for extensive reviews on capacity planning in time-varying systems.

This article presents a Markov model that approximates the transient and periodic steady-state behavior of the $\mathrm{G}(t)/\mathrm{G}(t)/s(t) + \mathrm{G}(t)$ queue with exhaustive service discipline and time-varying arrival, service, and abandonment rates. The model evaluates the following time-varying performance metrics: (1) the expected queue length, (2) the variance of the queue length, (3) the expected number of abandonments, and (4) the virtual waiting time distribution of a customer arriving at an arbitrary moment in time. Our approach extends the work of Ingolfsson et al. (2007) and Ingolfsson (2005), who apply the randomization method introduced by Jensen (1953) and Grassmann (1977) to systems with nonstationary arrival rates. Ingolfsson et al. (2007) and Ingolfsson (2005) target $\mathrm{M}(t)/\mathrm{M}/s(t)$ queues with an exhaustive service policy (an outline on how to include customer impatience is provided, yet not implemented).

We refer to our model as G-RAND since it uses the randomization method to analyze a queue with general interarrival, service, and abandonment time distributions. To the best of our knowledge, this is the first analytical model that studies a queue with an exhaustive service policy, customer impatience, and generally distributed (time-varying) arrival, service,

and abandonment rates. Our model does not rely on heavy-traffic or many-server asymptotics that are commonly used in the literature, and is intended for small- to medium-sized systems with human servers (e.g., banks, retail stores, or small-scale call centers). Larger systems can be analyzed as well, albeit at a higher computational cost.

The remainder of the article is organized as follows: Section 2 starts with a brief overview of the literature on performance measurement in systems with time-varying arrivals. In Section 3, we present an in-depth description of the Markov model itself. Section 4 evaluates the accuracy of the model by means of a computational experiment. Finally, in Section 5, we highlight the main conclusions and suggest directions for further research.

2 Related literature

Previous work has mainly focused on systems with time-varying arrival rates. In this section, we provide a brief overview of the most popular performance evaluation methods for such systems.

Stationary approximations are by far the most widely adopted approach. The arrival rate that is fed into the stationary model can be, for instance, the instantaneous arrival rate (as in the Pointwise Stationary Approximation or PSA (Green et al., 1991; Green and Kolesar, 1991; Whitt, 1991)) or the average arrival rate over a given interval (Stationary Independent Period-by-Period or SIPP (Green et al., 2001; Whitt, 1991)). However, time-varying systems typically display a time lag (or congestion lag): peaks in actual offered load lag the arrival rate peaks, with an amount that is proportional to the expected service time (Green and Kolesar, 1995; Thompson, 1993). Accounting for this lag can greatly improve the accuracy of SIPP and PSA, particularly when service times are long (see the lagged variants of SIPP and PSA (Green and Kolesar, 1997, 1995; Green et al., 2001)). The Modified Offered Load (MOL) approximation accounts for the congestion lag by relying on analytically tractable results for infinite server queues, which can be found in Eick et al. (1993a,b). Further details on MOL can be found in Feldman et al. (2008), Jennings et al. (1996), Liu and Whitt (2009), Jagerman (1975), Massey and Whitt (1994, 1997), and Davis et al. (1995). Though stationary approximations are straightforward and generally applicable, additional challenges may arise in complex systems, for which the stationary model itself is intractable. For instance, the applicability of MOL to the $\mathrm{M}(t)/\mathrm{G}/s(t)+\mathrm{G}$ model necessarily relies on the availability and accuracy of approximations for the corresponding stationary $\mathrm{M}/\mathrm{G}/s+\mathrm{G}$ model (Whitt (2005) and Iravani and Balcioğlu (2008) provide approximations for this queue). We refer to Green et al. (2007), Whitt (2007), and Defraeye and Van Nieuwenhuyse (2011) for further references on the stationary approximations available in the literature.

For the $\mathrm{M}(t)/\mathrm{M}/s(t)$ system, performance can be evaluated by numerically integrating the Chapman-Kolmogorov forward equations, a set of Ordinary Differential Equations (ODEs) that describe the behavior of the system (see Gross et al. (2008) for general background; Ingolfsson et al. (2007) and Green and Soares (2007) provide a more thorough discussion). This can be achieved using an ODE-solver such as the Euler or Runge-Kutta ODE solver from the Matlab ODE Suite Shampine and Reichelt (1997). Ingolfsson et al. (2007) show that this approach requires substantial computational effort and suggest using the randomization approach instead: this enables a drastic reduction in computational effort, at the

cost of a slightly lower accuracy. The randomization (or uniformization) approach was originally developed for stationary systems (Jensen, 1953; Grassmann, 1977; Gross and Miller, 1984), but can be applied successfully to nonstationary queues (Ingolfsson, 2005; Ingolfsson et al., 2007). In general, methods that use randomization or that numerically solve ODEs, rely heavily on Markovian assumptions. The majority of these methods use an exponential distribution for the service and/or abandonment process. Izady (2010) describes how these methods can be extended to phase-type distributions, and concludes that the computational effort increases considerably (as is confirmed by the results of our computational experiment, see Section 4). Furthermore, these approaches currently do not take into account abandonments (though Ingolfsson (2005) provides an outline on how to accommodate abandonments in the randomization approach).

Closure approximations (Rothkopf and Oren, 1979; Clark, 1981; Taaffe and Ong, 1987) approximate the set of forward differential equations by a small number of differential equations. Rothkopf and Oren (1979), for instance, use one for the mean and one for the variance of the number in system at each time instant. However, as shown in Ingolfsson et al. (2007), the approach is cumbersome to implement and is dominated by other methods (such as MOL or randomization) in terms of both accuracy and computation speed.

Discrete-Time Modeling (DTM) is used for performance evaluation of systems with general service time distributions (Chassioti and Worthington, 2004; Brahimi, 1990; Brahimi and Worthington, 1991; Wall and Worthington, 1994, 2007). This approach approximates the general service process by means of a discrete process using a two-moment matching technique (Brahimi, 1990; Brahimi and Worthington, 1991). Wall and Worthington (2007) report distinct advantages over stationary approximations such as MOL and PSA, particularly when temporal overloading is present. The complexity and computational effort of DTM, however, increase drastically with the number of servers; Wall and Worthington (2007) propose an approximation method to mitigate this effect. Note that the current DTM articles all study the $\mathrm{M}(t)/\mathrm{G}/s$ system (i.e., they assume a constant number of servers and no abandonments).

Deterministic fluid models (intended for systems that do not display stochasticity) can be used as approximations to derive time-dependent performance in stochastic systems. These methods rely on so-called "fluid scaling": the system is scaled up (e.g., by multiplying the arrival rates and the number of servers by the same factor) such that the stochastic randomness decreases in importance relative to the system dynamics (see Helber and Henken (2010) for an example). Fluid approximations are particularly useful to assess performance in systems that are temporarily overloaded (Whitt, 2006a), but may fail to capture system dynamics accurately in underloaded systems (Aguir et al., 2004; Altman et al., 2001; Jiménez and Koole, 2004). Liu and Whitt (2010) suggest an approach that works for overloaded as well as underloaded systems (separate models are applied in the two situations). Additional literature on the use of fluid approximations for Markovian models, can be found in Mandelbaum et al. (1995, 1998, 1999a,b, 2002), Ridley et al. (2003), and Jiménez and Koole (2004). For systems with general service and/or abandonment time distributions, we refer to the more recent work of Whitt (2006a) on $\mathrm{G}(t)/\mathrm{GI}/s+\mathrm{GI}$ models (with state-dependent arrival rates), Liu and Whitt (2010, 2011b, 2012a,b) on the $\mathrm{G}(t)/\mathrm{GI}/s(t)+\mathrm{GI}$ queue, Liu and Whitt (2011a) for a network of $\mathrm{G}(t)/\mathrm{M}(t)/s(t)+\mathrm{GI}(t)$ queues, and references therein. A key characteristic of fluid models is that arrivals and departures are considered as continuous

flows, rather than discrete processes (an assumption that becomes more acceptable as the number of servers increases). Although Liu and Whitt (2010) report reasonably accurate results for a system with 20 servers, the assumption of fluid scaling renders these approximations less applicable to small-scale settings where the discreteness of capacity is an essential characteristic of the system.

Finally, discrete-event simulation is frequently used (see, e.g., (Law and Kelton, 2000) for a comprehensive textbook). The appeal of simulation lies in its inherent flexibility to evaluate the performance of virtually any given system. As such, simulation proves particularly useful in settings that are analytically intractable. On the downside, simulation tends to be rather time-consuming, both in terms of runtime and time required to build the model. The number of replications to ensure reliable accuracy may be extremely large; Koopman (1972) put forward this argument to highlight why numerically solving ODEs should be preferred over simulation. Although simulation models are commonly dedicated and context-specific (e.g., (McGuire, 1994; García et al., 1995; Evans et al., 1996; Takakuwa and Shiozaki, 2004; Hung et al., 2007; Ahmed and Alkhamis, 2009) describe simulation applications in emergency departments with time-varying arrivals), efforts are made to develop generic simulation models (e.g., (Pitt, 1997; Sinreich and Marmor, 2004; Fletcher et al., 2007a,b; Gunal and Pidd, 2009)).

3 Model

In this section, we develop a phase-type (PH) approximation for the $\mathrm{G}(t)/\mathrm{G}(t)/s(t)+\mathrm{G}(t)$ queue with exhaustive service discipline and abandonments. Analogous to the DTM models discussed in the previous section, we observe the state of the system at discrete moments in time. The main events that can take place at these observation moments are: arrivals, departures (service completions or abandonments), and capacity changes (these basic processes are defined in Section 3.1). Unlike the DTM models, however, we do not rely on discrete distributions, but use continuous-time PH distributions to match the continuous system processes. The PH distributions, described in Section 3.2, allow us to decompose a general distribution into a set of exponential building blocks (so-called "phases"): because each phase of a continuous-time PH distribution has an exponentially distributed visiting time, the system processes are approximated by mixtures of exponential distributions. A notable downside of DTM is that it requires keeping track of each server individually. In our approach, however, this is not the case: due to the memoryless property of the exponential distribution, it suffices to keep track of the number of active servers associated with a given phase of the service process.

Sections 3.3–3.6 detail how the system state is updated from one observation moment to the next. Our model requires a counting process to determine the number of arrivals in a given interval (Section 3.3), a procedure to determine the probability that a given number of customers advances a phase (Section 3.4), and a procedure to determine which customers have experienced the longest waiting time (Section 3.5). An in-depth discussion of the model logic is given in Section 3.6. Section 3.7 explains why G-RAND is an approximation.

Various time-varying performance metrics can be derived (i.e., the expected queue length, the variance of the queue length, the expected number of abandonments, and the virtual

waiting time distribution of a customer arriving at an arbitrary moment in time). They are discussed in Section 3.8. The Appendix provides an overview of the main notations, used throughout the article.

3.1 Basic Processes

We observe the state of the system at discrete, equidistant moments in time. The time between observation moments determines the granularity (and hence the precision) of the model and is denoted by Δ. Define $\mathbf{T} = \{1, \ldots, T\}$, the set of periods (where T is the last period; the period that marks the end of the time horizon). There are four basic processes: (I) the arrival process, (II) the service process, (III) the abandonment process, and (IV) the staffing process. In the remainder of this article, Roman numerals I, II, III, and IV are used to label these processes. At the start of any given period, the parameters of the arrival, service, abandonment, and staffing process are allowed to change. If such a change takes place, the start of the period corresponds to the start of a so-called "epoch". For each process, we thus partition the set of periods into a set of epochs, where each epoch is a set of consecutive periods during which the process parameters do not change. Let $\mathbf{D}^{(\cdot)} = \left\{1, 2, \ldots, D^{(\cdot)}\right\}$ denote the set of epochs for a process $(\cdot)$, where $D^{(\cdot)}$ is the total number of epochs over the time horizon. For each process $(\cdot)$, define t_d, the first period in epoch d, where $t_1 = 0$ and $t_i < t_j \leq t_{D^{(\cdot)}} \leq T$ for all $i, j : i < j \leq D^{(\cdot)}$. Function $\phi_t^{(\cdot)} = i$ maps a period t onto an epoch i, where i is the ongoing epoch at the start of period t (i.e., there exists no epoch j for which $t_i < t_j \leq t$). Figure 2 further illustrates the relation between periods and epochs for an arbitrary process $(\cdot)$.

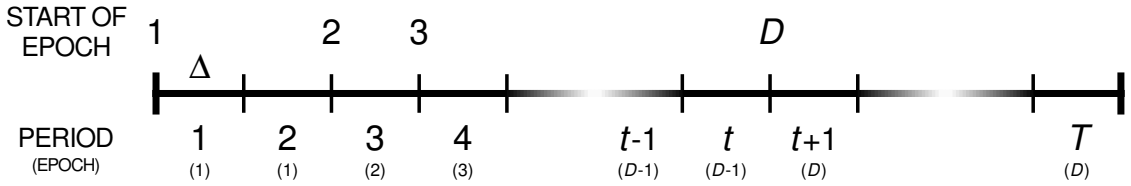


Figure 2: Relation between periods and epochs

Each epoch d of the arrival, service, and abandonment process is characterized by an independent distribution $G_d^{(\cdot)}$ that has mean $\mu_d^{(\cdot)}$ and standard deviation $\sigma_d^{(\cdot)}$. Each epoch of the staffing process represents a so-called "staffing interval" (during which staffing remains unchanged) and is associated with a number of servers $s_d : d \in \mathbf{D}^{(\mathrm{IV})}$. Note that Δ has to be chosen such that all staffing intervals are integer multiples of Δ. Figure 3 summarizes the multi-server service system with time-varying interarrival times, service times, abandonment times, and staffing levels.

3.2 Phase-type distributions

We adopt continuous-time PH distributions to approximate the general interarrival, service, and abandonment time distributions. Continuous-time PH distributions use exponentially-distributed building blocks to approximate any positive-valued continuous distribution with

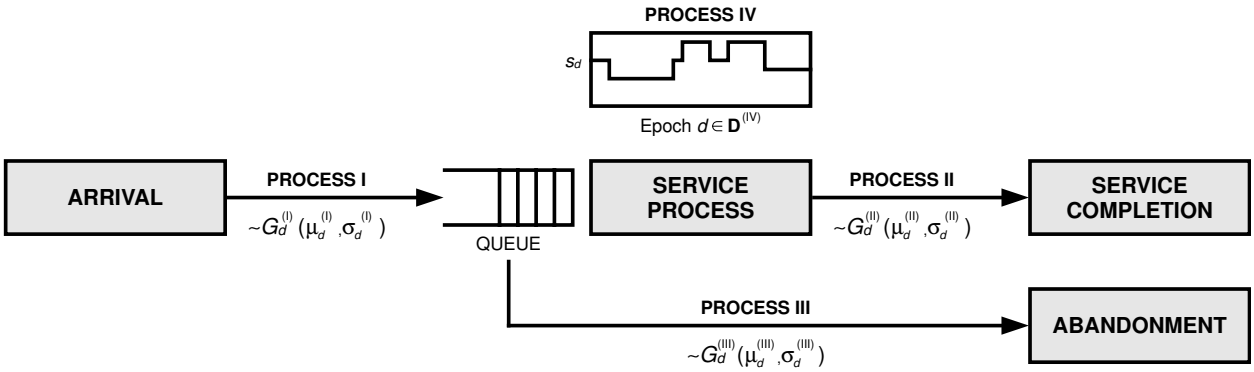


Figure 3: The $\mathrm{G}(t)/\mathrm{G}(t)/s(t) + \mathrm{G}(t)$ queueing system

arbitrary precision. More formally, the set of PH distributions is dense in the set of nonnegative distributions Nelson (1995) and, in theory, any nonnegative distribution can be approximated arbitrarily closely by a PH distribution Osogami (2005). For further details on PH distributions, refer to Neuts (1981), Nelson (1995), Latouche and Ramaswami (1999), and Osogami (2005).

A PH distribution may be seen as the distribution of time until absorption in a Markov chain with absorbing state 0 and state space $\{0, 1, \ldots, Z-1, Z\}$. It is fully characterized by parameters $\boldsymbol{\tau}$ and $\mathbf{R}$. $\boldsymbol{\tau}$ is the vector of probabilities to start the process in any of the Z transient states (i.e., phases) and $\mathbf{R}$ is the transient state transition matrix. The infinitesimal generator of the Markov chain representing the PH distribution is:

$$\mathbf{Q} = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{t} & \mathbf{R} \end{pmatrix},$$

where $\mathbf{0}$ is a zero matrix of appropriate dimension and $\mathbf{t} = -\mathbf{R}\mathbf{e}$ (with $\mathbf{e}$ a vector of ones of appropriate size).

Various techniques exist to approximate a given distribution by means of a PH distribution (for an overview, refer to Osogami (2005), Osogami (2006), and Gerhardt and Nelson (2009)). In this article, we adopt a two-moment matching procedure that uses a minimum number of phases. Let C^2 denote the squared coefficient of variation of the distribution we want to approximate:

$$C^2 = \sigma^2 \mu^{-2}. \tag{1}$$

We distinguish three cases: (1) $C^2 = 1$, (2) $C^2 > 1$, and (3) $C^2 < 1$. In the first case, we approximate the distribution by means of an exponential distribution with rate parameter $\lambda = \mu^{-1}$. The parameters of the corresponding PH distribution are:

$$\begin{aligned} \boldsymbol{\tau} &= 1, \\ \mathbf{R} &= (-\lambda). \end{aligned}$$

In the second case ($C^2 > 1$), we use a two-phase Coxian distribution where the rate parameter of the first phase is determined by means of a scaling factor κ:

$$\lambda_1 = \frac{1}{\mu\kappa}, \tag{2}$$

where $\kappa \in [0,1]$ and can be arbitrarily chosen. Unless mentioned otherwise, we assume $\kappa = 0.5$. The expected value of the two-phase Coxian distribution is:

$$\mu = \lambda_1^{-1} + \beta\lambda_2^{-1}, \tag{3}$$

where λ_2 is the exponential rate parameter of the second phase and β is the probability of visiting the second phase. The variance of the two-phase Coxian distribution is:

$$\sigma^2 = \lambda_1^{-2} + 2\beta\lambda_2^{-2} - \beta^2\lambda_2^{-2}. \tag{4}$$

When deriving parameters λ_2 and β as a function of parameters μ, C^2, and κ, we obtain:

$$\lambda_2 = \frac{2(\kappa-1)}{\mu(2\kappa-1-C^2)}, \tag{5}$$

$$\beta = \frac{2(\kappa-1)^2}{1+C^2-2\kappa}. \tag{6}$$

The parameters of the corresponding PH distribution are:

$$\begin{aligned} \boldsymbol{\tau} &= \mathbf{e}_1, \\ \mathbf{R} &= \begin{pmatrix} -\lambda_1 & \beta\lambda_1 \\ 0 & -\lambda_2 \end{pmatrix}, \end{aligned}$$

where $\mathbf{e}_1$ is the first unit vector.

In the third case ($C^2 < 1$), we use a hypo-exponential distribution (a convolution of exponential distributions whose parameters are allowed to differ; a generalization of the Erlang distribution). The number of required phases equals:

$$Z = \lceil C^{-2} \rceil. \tag{7}$$

We assume that the first $Z-1$ phases of the hypo-exponential distribution are exponentially distributed with rate parameter λ_1. The last phase is exponentially distributed with rate parameter λ_2. The expected value and variance of the hypo-exponential distribution are:

$$\mu = (Z-1)\lambda_1^{-1} + \lambda_2^{-1}, \tag{8}$$

$$\sigma^2 = (Z-1)\lambda_1^{-2} + \lambda_2^{-2}. \tag{9}$$

When deriving parameters λ_1 and λ_2 as a function of parameters μ, C^2, and Z, we obtain:

$$\lambda_1 = \frac{(Z-1) - \sqrt{(Z-1)(ZC^2-1)}}{\mu(1-C^2)}, \tag{10}$$

$$\lambda_2 = \frac{1 + \sqrt{(Z-1)(ZC^2-1)}}{\mu(1-ZC^2+C^2)}. \tag{11}$$

The parameters of the corresponding PH distribution are:

$$\begin{aligned} \boldsymbol{\tau} &= \mathbf{e}_1, \\ \mathbf{R} &= \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\lambda_1 & \lambda_1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_1 & \lambda_1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\lambda_1 & \lambda_1 \\ 0 & 0 & 0 & \cdots & 0 & 0 & -\lambda_2 \end{pmatrix}. \end{aligned}$$

For the three cases, Z equals 1, 2, and $\lceil C^{-2} \rceil$ respectively. Figure 4 provides an overview of the PH distributions that are used in this article.

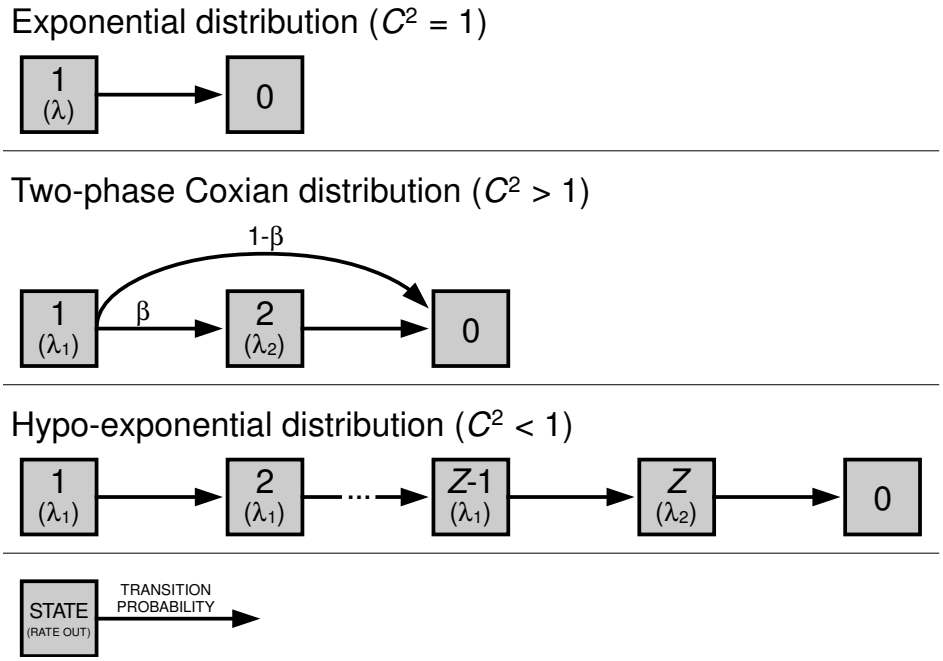


Figure 4: Overview of PH distributions

Many other two-moment matching procedures are available in the literature. These procedures typically rely on a mixture of Erlang distributions if C^2 is smaller than 1 (see Marie (1980) and Johnson and Taaffe (1989, 1990), for instance) and use hyperexponential distributions (e.g., Sauer and Chandy (1975) and Whitt (1982)) or two-phase Coxian distributions (e.g., Altiok (1985)) if C^2 is larger than 1. In Section 4.4 we further discuss the impact of our fitting procedures on the accuracy of our model.

Note that, although in this article we limit ourselves to the use of simple PH distributions, G-RAND can easily be extended to work with any acyclic, continuous-time PH distribution. Therefore, our model can also be used to assess the queueing behavior of systems where general processes are approximated by more complex PH distributions (albeit at a higher computational cost, if more phases are required).

3.3 Counting process

We use a counting process to obtain $\Pr(x, v|u, d)$, the probability of having x arrivals during an interval t (of length Δ) for which $\phi_t^{(\mathrm{I})} = d$, and an arrival process at final phase v given that the arrival process starts in phase u and is modeled using a PH distribution with parameters $\boldsymbol{\tau}_d^{(\mathrm{I})}$ and $\mathbf{R}_d^{(\mathrm{I})}$.

The counting process has continuous-time rate matrix (Ramaswami, 1988):

$$\mathbf{Q}_d = \begin{pmatrix} \mathbf{L}_d & \mathbf{F}_d & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{L}_d & \mathbf{F}_d & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{L}_d & \mathbf{F}_d & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{L}_d & \cdots \\ \cdots & \cdots & \cdots & \cdots & \ddots \end{pmatrix},$$

where $\mathbf{L}_d = \mathbf{R}_d^{(\mathrm{I})}$ and $\mathbf{F}_d = \mathbf{t}_d^{(\mathrm{I})} \left(\boldsymbol{\tau}_d^{(\mathrm{I})}\right)^\top$. $\mathbf{C}_d$ holds the transition probabilities of the counting process during an interval of length Δ during epoch d, and may be obtained using randomization (see for instance Grassmann (1977) and Van Moorsel (1994)):

$$\begin{aligned} \mathbf{C}_d &= e^{\Delta \mathbf{Q}_d}, && (12) \\ &= \sum_{i=0}^{\infty} \frac{\Delta^i}{i!} \mathbf{Q}_d^i, && (13) \\ &= e^{-\Delta \lambda_{d,\max}} \sum_{i=0}^{\infty} \frac{(\Delta \lambda_{d,\max})^i}{i!} \mathbf{P}_d^i, && (14) \end{aligned}$$

where $\lambda_{d,\max} = -\min\left(\mathrm{Diag}\left(\mathbf{R}_d\right)\right)$ and $\mathbf{P}_d$ is obtained as follows:

$$\mathbf{P}_d = \frac{\mathbf{Q}_d}{\lambda_{d,\max}} + \mathbf{I}, \tag{15}$$

where $\mathbf{I}$ is an identity matrix of appropriate dimension.

The first block row of $\mathbf{C}_d$ holds the distribution of the number of arrivals (i.e., probabilities $\Pr\left(x, v | u, d\right)$). In order to obtain the first block row of $\mathbf{C}_d$, it suffices to compute $\mathbf{P}_{d,1}^{(i)}$, the first block row of $\mathbf{P}_d^i$, for all $i \geq 0$. For $i = 0$, the first block row of $\mathbf{P}_d^i$ is defined as follows:

$$\mathbf{P}_{d,1}^{(0)} = \left(\mathbf{I} \quad \mathbf{0}_{d,1}\right), \tag{16}$$

where $\mathbf{0}_{d,1}$ is a zero-matrix with infinite number of columns and a number of rows equal to the number of phases in the PH distribution with parameters $\boldsymbol{\tau}_d^{(\mathrm{I})}$ and $\mathbf{R}_d^{(\mathrm{I})}$. For $i > 0$, $\mathbf{P}_{d,1}^{(i)}$ is obtained using the Chapman-Kolmogorov equations (see Latouche and Ramaswami (1999) and Tijms (2003) for instance):

$$\mathbf{P}_{d,1}^{(i)} = \left(\frac{\mathbf{L}_d}{\lambda_{d,\max}} + \mathbf{I}\right) \mathbf{P}_{d,1}^{(i-1)} + \left(\mathbf{0}_d \quad \frac{\mathbf{F}_d}{\lambda_{d,\max}} \mathbf{P}_{d,1}^{(i-1)}\right), \tag{17}$$

where $\mathbf{0}_d$ is a square zero-matrix with a number of columns/rows equal to the number of phases in the PH distribution with parameters $\boldsymbol{\tau}_d^{(\mathrm{I})}$ and $\mathbf{R}_d^{(\mathrm{I})}$.

3.4 Procedure to determine the probability of advancing a phase

The following procedure is used to determine the probability to advance a phase in the service and abandonment processes. Let $\Pr\left(y | x, u, d\right)^{(\cdot)}$ denote the probability that y customers

successfully complete phase u of process $(\cdot)$ during an interval of length Δ, given that x customers are present in phase u at the start of the interval and that the process is modeled using a PH distribution with parameters $\boldsymbol{\tau}_d^{(\cdot)}$ and $\mathbf{R}_d^{(\cdot)}$.

In order to compute $\Pr\left(y|x,u,d\right)^{(\cdot)}$, we use a Markov process that has infinitesimal generator:

$$\mathbf{Q}_{d,u}^{(\cdot)} = \begin{pmatrix} -y\lambda_{d,u}^{(\cdot)} & y\lambda_{d,u}^{(\cdot)} & \cdots & 0 & 0 & 0 \\ -(y-1)\lambda_{d,u}^{(\cdot)} & (y-1)\lambda_{d,u}^{(\cdot)} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -2\lambda_{d,u}^{(\cdot)} & 2\lambda_{d,u}^{(\cdot)} & 0 \\ 0 & 0 & \cdots & 0 & -\lambda_{d,u}^{(\cdot)} & \lambda_{d,u}^{(\cdot)} \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix},$$

where $\lambda_{d,u}^{(\cdot)}$ is the exponential rate that corresponds to the u-th phase of a PH distribution with parameters $\boldsymbol{\tau}_d^{(\cdot)}$ and $\mathbf{R}_d^{(\cdot)}$. $\mathbf{C}_{d,u}^{(\cdot)}$ holds the transition probabilities after an interval of length Δ during epoch d. The first row of $\mathbf{C}_{d,u}^{(\cdot)}$ holds the distribution of the number of successes (i.e., probabilities $\Pr\left(y|x,u,d\right)^{(\cdot)}$) and may be obtained using an approach similar to the one outlined in the Section 3.3.

3.5 Procedure to determine which customers have experienced the longest waiting time

In this section, we determine $\Pr\left(\mathbf{b}^{\mathrm{s}}|\mathbf{b},d\right)$, the probability that $\mathbf{b}^{\mathrm{s}}$ contains the distribution of customers who have experienced the longest waiting time, given that (1) $\mathbf{b}$ is the distribution of customers over the different phases of the abandonment process and (2) the abandonment process is modeled using a PH distribution with parameters $\boldsymbol{\tau}_d^{(\mathrm{III})}$ and $\mathbf{R}_d^{(\mathrm{III})}$. In order to obtain $\Pr\left(\mathbf{b}^{\mathrm{s}}|\mathbf{b},d\right)$, we first need to determine $\Pr\left(\mathbf{e}_u|\mathbf{b},d\right)$, the probability that a customer in abandonment phase u has waited the longest, given that the abandonment process is modeled using a PH distribution with parameters $\boldsymbol{\tau}_d^{(\mathrm{III})}$ and $\mathbf{R}_d^{(\mathrm{III})}$ (where $\mathbf{e}_u$ is the u-th unit vector).

If the abandonment process requires only a single phase (i.e., if $Z_d^{(\mathrm{III})} = 1$), $\Pr\left(\mathbf{e}_1|\mathbf{b},d\right) = 1$ for all $\mathbf{b} \in \mathbf{B}$. If $Z_d^{(\mathrm{III})} = 2$, a customer in the first phase has waited longer than any of the customers in the second phase if two criteria are met. First, the waiting time of the customer has to be larger than the maximum time that was spent in the first phase by any of the customers who are currently in the second phase. This occurs with probability:

$$\frac{b_1}{b_1+b_2},$$

where b_u is the u-th entry of vector $\mathbf{b}$. Second, the waiting time of the customer has to be larger than the maximum time that has already been spent in the second phase by any of the customers who are currently in the second phase. This occurs with probability:

$$\int_0^\infty g\left(x|b_1,\lambda_{d,1}^{(\mathrm{III})}\right)\int_0^x g\left(y|b_2,\lambda_{d,2}^{(\mathrm{III})}\right)\,\mathrm{d}y\,\mathrm{d}x,$$

where $g\left(x|n,\lambda\right)$ is the probability density function of the maximum of n i.i.d. exponential distributions with rate parameter λ. Note that if $\lambda_{d,1}^{(\mathrm{III})} = \lambda_{d,2}^{(\mathrm{III})}$:

$$\frac{b_1}{b_1+b_2} = \int_0^\infty g\left(x|b_1,\lambda_{d,1}^{(\mathrm{III})}\right)\int_0^x g\left(y|b_2,\lambda_{d,2}^{(\mathrm{III})}\right)\,\mathrm{d}y\,\mathrm{d}x.$$

If $\lambda_{d,1}^{(\mathrm{III})} \neq \lambda_{d,2}^{(\mathrm{III})}$, the probability can be evaluated numerically. Due to the memoryless property of the exponential distribution, both events (i.e., meeting the first and the second criterion) are independent and therefore, probabilities $\Pr\left(\mathbf{e}_u|\mathbf{b},d\right)$ can be obtained as follows:

$$\Pr\left(\mathbf{e}_1|\mathbf{b},d\right) = \frac{b_1}{b_1+b_2}\int_0^\infty g\left(x|b_1,\lambda_{d,1}^{(\mathrm{III})}\right)\int_0^x g\left(y|b_2,\lambda_{d,2}^{(\mathrm{III})}\right)\,\mathrm{d}y\,\mathrm{d}x, \tag{18}$$

$$\Pr\left(\mathbf{e}_2|\mathbf{b},d\right) = 1-\Pr\left(\mathbf{e}_1|\mathbf{b},d\right). \tag{19}$$

Note that:

- if $b_1 > 0$ and $b_2 = 0$, $\Pr\left(\mathbf{e}_1|\mathbf{b},d\right) = 1$ and $\Pr\left(\mathbf{e}_2|\mathbf{b},d\right) = 0$,
- if $b_1 = 0$ and $b_2 > 0$, $\Pr\left(\mathbf{e}_1|\mathbf{b},d\right) = 0$ and $\Pr\left(\mathbf{e}_2|\mathbf{b},d\right) = 1$,
- if $b_1 = 0$ and $b_2 = 0$, $\Pr\left(\mathbf{e}_1|\mathbf{b},d\right) = 0$ and $\Pr\left(\mathbf{e}_2|\mathbf{b},d\right) = 0$,
- if customers in the second phase do not visit the first phase, only the second criterion has to be met.

If $Z_d^{(\mathrm{III})} > 2$, a similar logic may be applied in order to obtain $\Pr\left(\mathbf{e}_u|\mathbf{b},d\right)$.

Given $\Pr\left(\mathbf{e}_u|\mathbf{b},d\right)$, $\Pr\left(\mathbf{b}^{\mathrm{s}}|\mathbf{b},d\right)$ can be computed recursively:

$$\Pr\left(\mathbf{b}^{\mathrm{s}}|\mathbf{b},d\right) = \sum_{u=1}^{Z_d^{(\mathrm{III})}}\Pr\left(\mathbf{b}^{\mathrm{s}}-\mathbf{e}_u|\mathbf{b},d\right)\Pr\left(\mathbf{e}_u|\mathbf{b},d\right). \tag{20}$$

3.6 Model building blocks

Let $\left(a,\mathbf{k},\mathbf{b}\right)_t$ denote the state of the system at the start of period t, where (1) a is the phase of the arrival process, (2) $\mathbf{k}$ is a vector that holds the number of customers in each service phase, and (3) $\mathbf{b}$ is a vector that holds the number of customers in each abandonment phase. $\mathbf{K}$ and $\mathbf{B}$ are the sets of all possible vectors $\mathbf{k}$ and $\mathbf{b}$ respectively. In addition, define $\pi\left(a,\mathbf{k},\mathbf{b}\right)_t$, the probability to visit state $\left(a,\mathbf{k},\mathbf{b}\right)_t$. The maximum dimension of the state space at the start of any period depends on (1) the maximum number of phases of the arrival process $Z_{\max}^{(\mathrm{I})}$, (2) the maximum number of phases of the service process $Z_{\max}^{(\mathrm{II})}$, (3) the maximum number of phases of the abandonment process $Z_{\max}^{(\mathrm{III})}$, (4) the maximum number of servers $s_{\max}$, and (5) the maximum number of customers allowed in queue $Q_{\max}$.

In order to determine the state of the system at the start of a period t, we propose a stepwise procedure in which the arrival, service, and abandonment process are decomposed and are processed independently. After each step, the state of the system is updated. The stepwise procedure executes the following steps in sequence:

1. Initialization.
2. Activate or deactivate servers if necessary.
3. Arrival of customers.
4. Service of customers.
5. Abandonment of customers.

To make a transition from a state $(a, \mathbf{k}, \mathbf{b})_t$ towards a state $(a, \mathbf{k}, \mathbf{b})_{t+1}$, we manipulate the statespace for each of these steps. We use temporary probability vectors $\pi(1-\delta, a, \mathbf{k}, \mathbf{b})$ and $\pi(\delta, a, \mathbf{k}, \mathbf{b})$ (where δ is a binary variable, $\pi(1-\delta, a, \mathbf{k}, \mathbf{b})$ is the probability vector that represents the state of the system before manipulation, and $\pi(\delta, a, \mathbf{k}, \mathbf{b})$ is the probability vector that represents the state of the system after manipulation). Our method requires the state of the system to be stored only before and after each manipulation, which enables significant memory savings. This is of critical importance, as it is often infeasible to store the state space over the entire time horizon (even for small instances). After each state space manipulation, the binary variable δ is updated as follows: $\delta = 1 - \delta$.

3.6.1 Initialization

During the initialization step, we initialize the temporary probability vectors. More formally, we let $\pi(1-\delta, a, \mathbf{s}, \mathbf{b}) = 0$ and $\pi(\delta, a, \mathbf{s}, \mathbf{b}) = \pi(a, \mathbf{s}, \mathbf{b})_t$ for all $\mathbf{s} \in \mathbf{S}$, $\mathbf{b} \in \mathbf{B}$, and $a \in \left\{1, \ldots, Z^{(\mathrm{I})}_{\phi_t}\right\}$.

3.6.2 Activate or deactivate servers

If the staffing process changes, two options arise: (1) new servers become available or (2) the number of servers decreases.

If new servers become available, waiting customers are selected according to a first-come first-serve (FCFS) policy (i.e., we select those customers who have experienced the longest waiting time). From Section 3.5, we obtain probabilities $\Pr(\mathbf{b}^{\mathrm{s}}|\mathbf{b}, d)$. Using these probabilities, we can determine the state of the system after new servers have become available (the transition probability is indicated above the arrow):

$$(1-\delta, a, \mathbf{k}, \mathbf{b}) \xrightarrow{\Pr(\mathbf{b}^{\mathrm{s}}|\mathbf{b})_{\phi_t^{(\mathrm{III})}}} (\delta, a, \mathbf{k} + n_{\mathbf{b}^{\mathrm{s}}}\mathbf{e}_1, \mathbf{b} - \mathbf{b}^{\mathrm{s}}),$$

with $n_{\mathbf{b}^{\mathrm{s}}}$ the sum of all entries in vector $\mathbf{b}^{\mathrm{s}}$:

$$n_{\mathbf{b}^{\mathrm{s}}} = \mathrm{tr}(\mathbf{b}^{\mathrm{s}}\mathbf{I}), \tag{21}$$

where tr is the matrix trace operator. We assume that customers who enter service, start in the first phase of the service process, however, it is easy to adapt the model to allow service to start in another phase as well.

In case of a decrease in capacity, we need to account for the exhaustive service policy: servers complete a customer's service, even if they are selected to leave. We adopt an

approach that is similar to the technique used by Ingolfsson (2005): since servers that work overtime no longer influence the performance of future customers, these are removed from the system (along with the customers they serve). Although in reality, these customers are still in the system, this modification is necessary to correctly calculate other performance measures (such as the distribution of the virtual waiting time, see Section 3.8). Whereas Ingolfsson (2005) randomly removes servers (which can be idle or busy), we accommodate a decrease of x servers by first removing idle servers (if any). If insufficient idle servers are available, $c_{(x,\mathbf{k},t)}$ active servers are removed:

$$c_{(x,\mathbf{k},t)} = \max\left(0, x - s_t + n_{\mathbf{k}}\right), \tag{22}$$

where $n_{\mathbf{k}} = \mathrm{tr}\left(\mathbf{kI}\right)$ and $s_t - n_{\mathbf{k}}$ represents the number of idle servers. Given a distribution of customers $\mathbf{k}$ over the different phases of the service process, the probability to remove a server that is processing a customer who is in phase u of his service process equals:

$$\Pr\left(u|\mathbf{k}\right) = \frac{k_u}{n_{\mathbf{k}}}, \tag{23}$$

where k_u is the u-th entry of vector $\mathbf{k}$. For each active server that is removed, the following state-space manipulation is performed:

$$\left(1-\delta, a, \mathbf{k}, \mathbf{b}\right) \xrightarrow{\Pr(u|\mathbf{k})} \left(\delta, a, \mathbf{k} - e_u, \mathbf{b}\right),$$

The exhaustive service policy can be implemented in other ways, depending on which servers are removed when capacity decreases: e.g., random selection, selecting idle servers, or selecting servers with the smallest remaining processing times first. G-RAND is not restricted to the implementation described above, and could be modified to accommodate alternative disciplines.

3.6.3 Arrival, service, and abandonment of customers

From the counting process discussed in Section 3.3, we obtain probabilities $\Pr\left(x, v|u, d\right)$. Using these probabilities, we can determine the state of the system after arrivals have taken place. Because the size of the queue is limited to $Q_{\max}$ customers, we impose a reflecting boundary (i.e., whenever x customers arrive, with $x \geq Q_{\max} - n_{\mathbf{b}}$, the resulting queue length equals $Q_{\max}$). More formally:

$$\left(1-\delta, u, \mathbf{k}, \mathbf{b}\right) \xrightarrow{\Pr\left(x,v|u,\phi_t^{(\mathrm{I})}\right)} \begin{cases} \left(\delta, v, \mathbf{k}, \mathbf{b} + x\mathbf{e}_1\right) & \text{if } Q_{\max} \geq n_{\mathbf{b}} + x, \\ \left(\delta, v, \mathbf{k}, \mathbf{b} + \left(Q_{\max} - n_{\mathbf{b}}\right)\mathbf{e}_1\right) & \text{otherwise.} \end{cases}$$

Customers in service are only allowed to advance a single phase during an interval of length Δ. The probabilities of advancing a phase (i.e., probabilities $\Pr\left(y|x, u, d\right)^{(\cdot)}$) are obtained from the procedure given in Section 3.4. For each phase, a state-space manipulation is performed and phases are processed in reverse order. Customers who are in the last phase of their service process complete service (note that $Z_{\phi_t}^{(\mathrm{II})}$ is the last phase of the service process):

$$\left(1-\delta, a, \mathbf{k}, \mathbf{b}\right) \xrightarrow{\Pr(x|k_u,u,\phi_t)^{(\mathrm{II})}} \begin{cases} \left(\delta, a, \mathbf{k} - x\mathbf{e}_u, \mathbf{b}\right) & \text{if } k_u > 0 \wedge u = Z_{\phi_t}^{(\mathrm{II})}, \\ \left(\delta, a, \mathbf{k}, \mathbf{b}\right) & \text{otherwise.} \end{cases}$$

If the service process is modeled using a hypo-exponential distribution, customers who are not in the last phase of their service process advance a phase:

$$(1-\delta, a, \mathbf{k}, \mathbf{b}) \xrightarrow{\Pr(x|k_u,u,\phi_t)^{(\mathrm{II})}} \begin{cases} (\delta, a, \mathbf{k} - x\mathbf{e}_u + x\mathbf{e}_{u+1}, \mathbf{b}) & \text{if } k_u > 0 \wedge 1 \leq u < Z_{\phi_t}^{(\mathrm{II})}, \\ (\delta, a, \mathbf{k}, \mathbf{b}) & \text{otherwise.} \end{cases}$$

If the service process is modeled using a two-phase Coxian distribution, there is a probability that customers in the first phase complete service instead of advancing a phase. The probability of completing service equals $1 - \beta_{\phi_t}^{(\mathrm{II})}$. The probability that y out of x customers complete service is binomially distributed and equals:

$$\Pr(y|x, \phi_t)^{(\mathrm{II})} = \frac{x!}{y!\,(x-y)!} \left(1 - \beta_{\phi_t}^{(\mathrm{II})}\right)^{y} \left(\beta_{\phi_t}^{(\mathrm{II})}\right)^{x-y}. \tag{24}$$

The state-space transitions are summarized as follows:

$$(1-\delta, a, \mathbf{k}, \mathbf{b}) \xrightarrow{\Pr(x|k_u,u,\phi_t)^{(\mathrm{II})}\Pr(y|x,\phi_t)^{(\mathrm{II})}} (\delta, a, \mathbf{k} - x\mathbf{e}_u + (x-y)\,\mathbf{e}_{u+1}, \mathbf{b}).$$

After service completion, waiting customers are taken into service (i.e., servers are activated; see Section 3.6.2).

With respect to the abandonment process, customers waiting in queue can only advance a single abandonment phase during an interval of length Δ. The state space manipulations are analogous to the ones described for the service process.

After the abandonment step, probabilities $\pi(a, \mathbf{k}, \mathbf{b})_{t+1}$ are readily available:

$$\pi(a, \mathbf{k}, \mathbf{b})_{t+1} = \pi(\delta, a, \mathbf{k}, \mathbf{b}). \tag{25}$$

3.7 Model discussion

We emphasize that the presented model is an approximation because of the following reasons:

- The general arrival, service, and abandonment processes are approximated by means of PH distributions. Within each period, the time-varying rates are assumed to remain constant.
- We assume a finite queue length (in heavily-loaded or in large-scale systems, the finite queue size may need to be very large to maintain accuracy).
- The arrival, service, and abandonment process are decomposed and are processed independently, using a stepwise procedure. As a result, any interaction between the different processes during an interval of length Δ is not taken into account.
- We assume that any phase in the service and abandonment process takes at least one interval to complete. Therefore, PH distributions that have short phases, require lower values of Δ in order to maintain accuracy.

Clearly, the error that is induced by the two last assumptions tends to zero as Δ approaches zero.

3.8 Performance measures

Because of the computational effort involved, performance metrics are not necessarily evaluated at the start of every period $t \in \mathbf{T}$. Instead, we measure the performance at intervals $w \in \mathbf{W}$, with $\mathbf{W} \subseteq \mathbf{T}$ the set of performance intervals. Define $\varphi_w^{(\cdot)} = i$, the function that maps a performance interval w onto an epoch i, where i is the ongoing epoch of process $(\cdot)$ at the start of performance interval w. We obtain the following performance measures: (1) the time-average expected queue length, (2) the expected queue length at the start of performance interval w, (3) the time-average variance of the expected queue length, (4) the variance of the queue length at the start of performance interval w, (5) the expected number of abandonments during performance interval w, and (6) the waiting time distribution of a virtual customer arriving at the start of performance interval w. The virtual waiting time at the start of period t is defined as the time a virtual customer spends in queue if he were to arrive at the start of period t (cf. Gross et al. (2008) and Campello and Ingolfsson (2011)). The expected queue length at the start of performance interval w equals:

$$\mathcal{Q}_w = \sum_{a=1}^{Z_{\varphi w}^{(\mathrm{I})}} \sum_{\mathbf{k} \in \mathbf{K}} \sum_{\mathbf{b} \in \mathbf{B}} \pi \left(a, \mathbf{k}, \mathbf{b}\right)_w n_{\mathbf{b}}. \tag{26}$$

The time-average expected queue length is approximated by:

$$\mathcal{Q} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{Q}_t, \tag{27}$$

where $\mathcal{Q}_t$ denotes the queue length at the start of period t.

The variance of the queue length at performance interval w equals:

$$\mathcal{V}_w = \sum_{a=1}^{Z_{\varphi w}^{(\mathrm{I})}} \sum_{\mathbf{k} \in \mathbf{K}} \sum_{\mathbf{b} \in \mathbf{B}} \pi \left(a, \mathbf{k}, \mathbf{b}\right)_w \left(n_{\mathbf{b}} - \mathcal{Q}_w\right)^2. \tag{28}$$

The time-average variance of the queue length is approximated by:

$$\mathcal{V} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{V}_t, \tag{29}$$

where $\mathcal{V}_t$ denotes the variance of the queue length at the start of period t.

Let $\mathcal{A}_w$ denote the expected number of abandonments during performance interval w. $\mathcal{A}_w$ can easily be computed during the abandonment step by keeping track of the transitions in which customers abandon the queue.

Define $\Pr\left(\mathcal{W}_w = h\right)$, the probability that a virtual customer who arrives at the start of performance interval w receives service during interval $w + h$ (i.e., the virtual customer receives service after waiting h intervals of length Δ). In addition, let $W_{\max}$ denote the user-defined maximum waiting time over which probabilities $\Pr\left(\mathcal{W}_w = h\right)$ are observed. In order to obtain $\Pr\left(\mathcal{W}_w = h\right)$, we use a quasi-death process and stop the arrival process at the start of performance interval w. The first period during which a server becomes idle

defines the waiting time of the virtual customer. More formally, the virtual waiting time equals $h\Delta$ where h is the first integer for which $\mathcal{N}_{w+h} < s_{w+h}$ (where $\mathcal{N}_{w+h}$ denotes the number of customers in system after $h\Delta$ time units if the arrival process is stopped at the start of performance interval w; note that $\mathcal{N}_{w+h}$ does not include customers serviced by servers working overtime). The analysis of the quasi-death process requires a significant computational effort, especially for large values of $W_{\max}$. Note, however, that the quasi-death process does not need to be analyzed in order to compute the delay probability (the delay probability is given by $1 - \Pr(\mathcal{W}_w = 0)$ and may be obtained by setting $W_{\max} = 0$).

G-RAND enables both the transient and the periodic steady-state analysis of the $\mathrm{G}(t)/\mathrm{G}(t)/s(t) + \mathrm{G}(t)$ queue. To reach steady state, the model may have to run for multiple consecutive "cycles" (each with a length equal to the time horizon $T\Delta$). Let $c_{\max}$ denote the number of cycles after which steady-state results are obtained. In addition, define ε_c, the relative difference in queue lengths for cycles $(c-1)$ and c:

$$\varepsilon_c = \sum_{t=1}^{T} \left| 1 - \frac{\mathcal{Q}_{t,c}}{\mathcal{Q}_{t,c-1}} \right|, \tag{30}$$

where $\mathcal{Q}_{t,c}$ denotes the expected queue length at the start of period t in cycle c. If ε_c is smaller than the (user-specified) parameter $\varepsilon_{\max}$, cycle c is the last cycle and steady-state results have been obtained. In other words, $c_{\max}$ is the smallest integer for which $\varepsilon_{c_{\max}} < \varepsilon_{\max}$. In the case of a transient analysis, the user can specify the number of cycles that need to be processed.

4 Results

We use a simulation study to assess the accuracy of the model over a set of 162 problem instances. Both the Markov model and the simulation model are implemented in Visual Studio C++. All tests are performed on an AMD Phenom II 3.40 GHz computer, with 4 GB RAM.

In what follows, we first describe the computational experiment (Section 4.1) and discuss the main drivers of model accuracy and computation speed (Section 4.2). Next, we evaluate the model and elaborate further on the trade-off between accuracy and computation times (Section 4.3). Finally, we discuss the impact of the PH matching procedure on the accuracy of the model (Section 4.4).

4.1 Experimental setting

Table 1 provides an overview of the parameter settings that are used to construct the test set. The parameters give rise to 162 problem instances that are representative of small- to medium-sized systems. Each instance covers a one-day time horizon (i.e., 1,440 minutes) which is divided into smaller periods of length Δ. In the experiment, Δ ranges from 0.0625 to 1 minute. The arrival rate is piecewise constant over 10-minute intervals and the staffing interval has a length of 30 minutes.

The time-varying arrival rate $\lambda_t^{(\mathrm{I})}$ is modeled as a discretized sine function with cycle equal to $T\Delta$. Let $\mathrm{RA}^{(\mathrm{I})} \equiv A/\bar{\lambda}^{(\mathrm{I})}$ denote the relative amplitude, with A the absolute amplitude and $\bar{\lambda}^{(\mathrm{I})}$ the average arrival rate over the time horizon. More formally:

$$\lambda_t^{(\mathrm{I})} = \frac{\bar{\lambda}^{(\mathrm{I})}}{2}\left(2 + \mathrm{RA}^{(\mathrm{I})} \sin\left(\frac{2\pi t}{T\Delta}\right) + \mathrm{RA}^{(\mathrm{I})} \sin\left(\frac{2\pi\,(t+1)}{T\Delta}\right)\right). \tag{31}$$

Note that $\bar{\lambda}^{(\mathrm{I})}$ is determined uniquely by the average capacity $\bar{c}$, the average service rate $\bar{\lambda}^{(\mathrm{II})}$, and the average traffic intensity $\bar{\rho} \equiv \bar{\lambda}^{(\mathrm{I})}/\left(\bar{c}\bar{\lambda}^{(\mathrm{II})}\right)$. Given the parameter settings in Table 1, it follows that $\bar{\lambda}^{(\mathrm{I})}$ ranges between 1 and 57 customers per hour. To limit the size of the test set, we assume that all processes have the same C^2 (i.e., 0.5, 1, or 2) and that the distribution parameters of the service and the abandonment process remain constant throughout the day. We emphasize that G-RAND is not limited to these C^2-values and that it is possible to analyze time-varying service and/or abandonment processes as well.

The staffing process is modeled as a discretized sine function with relative amplitude $\mathrm{RA}^{(\mathrm{IV})}$. As such:

$$c_t = \frac{\bar{c}}{2}\left(2 + \mathrm{RA}^{(\mathrm{IV})} \sin\left(\frac{2\pi t}{T\Delta}\right) + \mathrm{RA}^{(\mathrm{IV})} \sin\left(\frac{2\pi\,(t+1)}{T\Delta}\right)\right). \tag{32}$$

Note that the capacity function is not shifted compared to the arrival rate function (which could be done to account for the commonly observed congestion lag).

The size of the queue (i.e., $Q_{\max}$) is either a characteristic of the system itself (e.g., a limited number of phone lines in a call center) or it is a function of the desired level of accuracy (i.e., if $Q_{\max}$ is set too small, many of the arriving customers do not join the queue and therefore do not receive service; they are "reflected"). In the experiment, we set $Q_{\max} = 25$. Over all problem instances that we tested, the probability of an arrival being reflected is at most 0.00006 per cycle. Preliminary computational experiments may be required to determine an appropriate value for $Q_{\max}$ in other settings.

Parameter	Values
Time horizon $T\Delta$ (in min)	1440
Period length Δ (in min)	$\{0.0625, 0.125, 0.25, 0.5, 1\}$
Epoch length (arrival process, in min)	10
Epoch length (staffing process, in min)	30
Performance interval length (in min)	1 for $\mathcal{Q}_w$, $\mathcal{V}_w$, and $\mathcal{A}_w$; 30 for $\Pr(\mathcal{W}_w = h)$
Relative amplitude $\text{RA}^{(\text{I})}$	0.5
Average service rate $\bar{\lambda}^{(\text{II})}$ (customers/hr)	$\{1, 2, 6\}$
Average abandonment rate $\bar{\lambda}^{(\text{III})}$	$\{0.5\bar{\lambda}^{(\text{II})}, \bar{\lambda}^{(\text{II})}\}$
Average capacity $\bar{c}$	$\{2, 5, 10\}$
Relative amplitude $\text{RA}^{(\text{IV})}$	0.5
Average traffic intensity $\bar{\rho} \equiv \bar{\lambda}^{(\text{I})} / \left(\bar{c}\bar{\lambda}^{(\text{II})}\right)$	$\{0.5, 0.75, 0.95\}$
Squared coefficient of variation C^2	$\{0.5, 1, 2\}$
Maximum waiting time $W_{\max}$ (in min)	30
Maximum allowed deviation $\varepsilon_{\max}$	0.0001
Maximum queue length $Q_{\max}$	25

Table 1: Parameter settings used in the computational experiment

We assessed the accuracy of the following time-varying performance metrics: (1) the expected queue length $\mathcal{Q}_w$, (2) the variance of the queue length $\mathcal{V}_w$, (3) the expected number of abandonments $\mathcal{A}_w$, and (4) the delay probability $\Pr(\mathcal{W}_w > 0)$. The computation of the delay probabilities themselves is generally straightforward; in our experiment, however, they are calculated together with probabilities $\Pr(\mathcal{W}_w = h)$, which requires a computationally intensive quasi-death process. As such, we opt to measure the delay probabilities every 30 minutes. For all other performance measures (i.e., $\mathcal{Q}_w$, $\mathcal{V}_w$, and $\mathcal{A}_w$), we use a one-minute performance interval.

The results of our model are compared with the results of an accurate simulation model that uses 1,000,000 independent replications (the maximum halfwidth of the confidence interval on the time-varying expected queue length is 0.00666). As in the Markov model, the simulation starts with an empty system and continues until steady state is reached. Only the data in the last cycle is retained, the other cycles may be considered as a warm-up period. The simulation model uses the same distributions as G-RAND (i.e., hypo-exponential, exponential, and two-phase Coxian distributions). This allows us to evaluate the accuracy of the model without interference of the PH matching procedure (refer to Section 4.4 for a discussion of the impact of the PH matching procedure on model accuracy). We emphasize that G-RAND can easily be adapted to work with other PH distributions and that other moment-matching approaches can be applied.

Let $\mathcal{Q}_t^{\text{SIM}}$ denote the simulated expected queue length at the start of period t. The Relative Error (RE) of the expected queue length at the start of period t then can be

expressed as:

$$\mathrm{RE}_t = \frac{\left|\mathcal{Q}_t^{\mathrm{SIM}} - \mathcal{Q}_t\right|}{\mathcal{Q}_t^{\mathrm{SIM}}}. \tag{33}$$

To obtain an aggregate performance metric over the time horizon, RE_t is weighted with the queue length. As such, the Weighted Relative Error (WRE) of the queue length for a given problem instance is defined as follows:

$$\mathrm{WRE} = \sum_{t=1}^{T} \left(\frac{\mathcal{Q}_t^{\mathrm{SIM}}}{\sum_{t=1}^{T} \mathcal{Q}_t^{\mathrm{SIM}}} \mathrm{RE}_t \right) = \frac{\sum_{t=1}^{T} \left|\mathcal{Q}_t^{\mathrm{SIM}} - \mathcal{Q}_t\right|}{\sum_{t=1}^{T} \mathcal{Q}_t^{\mathrm{SIM}}}. \tag{34}$$

The weighted relative error of the other metrics can be derived analogously.

4.2 Drivers of accuracy and computation speed

We distinguish three main drivers of accuracy and computation speed:

1. The length of Δ.
2. The size of the state space.
3. The approximations used in the model.

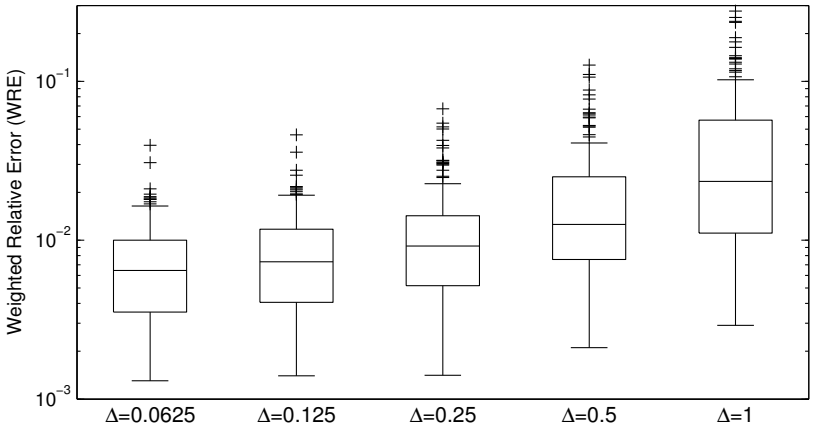
The choice of Δ determines the frequency at which the system is observed. Evidently, larger values of Δ lead to shorter computation times. Accurate results, however, can only be obtained if Δ is sufficiently small. Because service and abandonment processes are only allowed to advance a single phase during an interval of length Δ (see Section 3.7), accuracy will decrease if Δ is set too large. In addition, the arrival, service, and abandonment process are processed independently, using a stepwise procedure (see Section 3.6). As a result, the interaction between the different processes is not taken into account and the accuracy of the model decreases as more events are allowed to aggregate in between observation moments (i.e., if Δ is set too large and/or if the event frequency is too high).

The size of the state space only impacts the computation time. The state space grows linearly with the maximum capacity, the maximum queue length, and the required number of phases in the arrival, service, and abandonment processes.

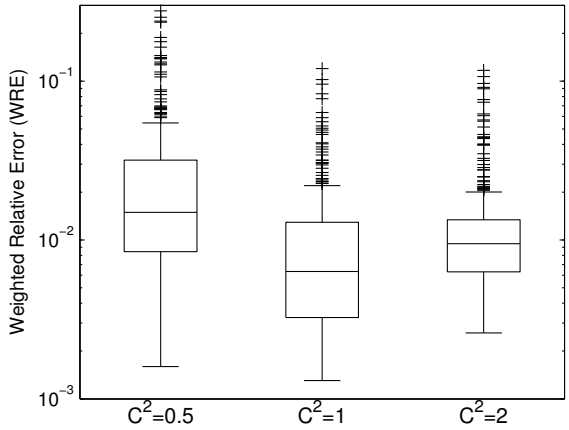
As the performance measures are calculated at each performance interval, an increase in the number of performance intervals will also increase the required computation time. This is particularly evident when calculating the virtual waiting time distribution as it involves the evaluation of a computationally intensive quasi-death process. Note that the computation times reported in this study include the computation of all aforementioned performance measures. Moreover, computation speed depends on the number of cycles needed to reach steady state. In our experiment, however, the model consistently terminates after 4 cycles.

4.3 Model accuracy and results

Figure 5(a) presents a box-and-whisker diagram of the WRE of the expected queue length, for different values of Δ (more detailed results can be found in Table 2). It is clear that the proposed method yields highly accurate results, provided that Δ is sufficiently small. Figure 6(a) shows the required CPU times in terms of Δ. We observe a distinct trade-off between accuracy and computational effort. In the remainder of this section, we further analyze this trade-off.



(a) WRE (Δ)



(b) WRE (C^2)

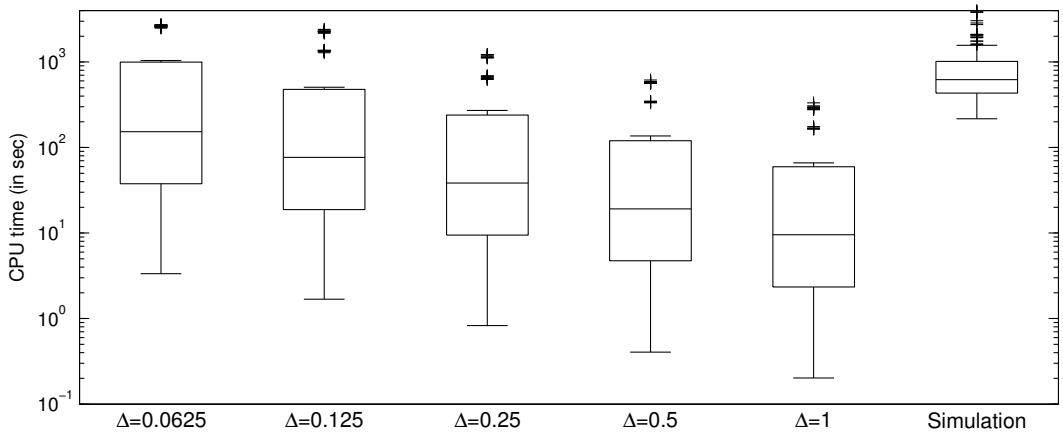
Figure 5: WRE of the expected queue length as a function of Δ and C^2

The lower quantiles of Figure 5(a) show that even for high values of Δ, the model can yield accurate results. From Table 2 and Figure 5(b) it is clear that the performance is

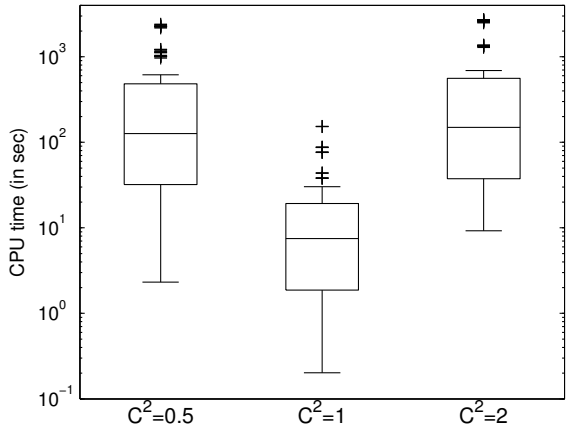
worst for the instances with $C^2 = 0.5$. If $C^2 = 0.5$, processes are modeled using a series of exponential distributions (see Section 3.2). The mean of these exponential distributions is smaller than the mean of the approximated distribution. As such, the event frequency increases and smaller values of Δ are required to obtain accurate results.

		$C^2 = 0.5$	$C^2 = 1$	$C^2 = 2$
$\Delta = 0.0625$	Min	0.002	0.001	0.003
	Avg	0.009	0.004	0.009
	Max	0.040	0.018	0.021
$\Delta = 0.125$	Min	0.002	0.001	0.003
	Avg	0.012	0.005	0.009
	Max	0.046	0.022	0.022
$\Delta = 0.25$	Min	0.003	0.001	0.003
	Avg	0.019	0.008	0.010
	Max	0.067	0.032	0.028
$\Delta = 0.5$	Min	0.007	0.002	0.003
	Avg	0.036	0.015	0.014
	Max	0.126	0.059	0.051
$\Delta = 1$	Min	0.012	0.004	0.003
	Avg	0.074	0.029	0.028
	Max	0.277	0.120	0.117

Table 2: WRE of the expected queue length, as a function of C^2



(a) CPU time (Δ)



(b) CPU time (C^2)

Figure 6: CPU times as a function of Δ and C^2

Figure 6(b) and Table 3 show that the CPU times increase drastically for non-exponential settings. This is no surprise, as the state space grows linearly with the number of phases. Figure 6(a), however, shows that the CPU times are still smaller than those of the simulation model.

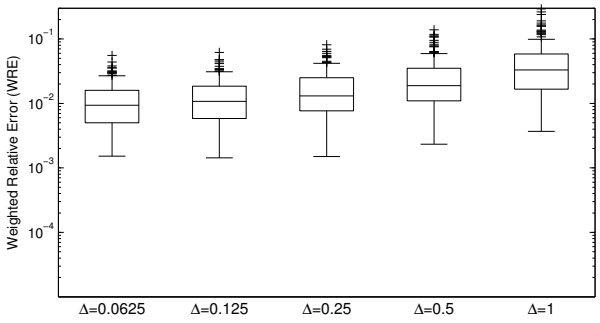
Figure 7 presents the WRE for the variance of the queue length, the expected number of abandonments, and the delay probability, as a function of Δ and C^2 (more detailed results are shown in Table 4). The results are similar to what we observed for the expected queue length (cf. Figures 5(a) and 5(b)): the WRE depends on the choice of Δ and the model is least accurate for $C^2 = 0.5$. The expected number of abandonments and the delay probability are markedly more accurate than the other metrics, hence, larger Δ-values may suffice to maintain an acceptable level of accuracy. The largest WREs are observed for the variance

		$C^2 = 0.5$	$C^2 = 1$	$C^2 = 2$
	Min	37.44	3.354	149.2
$\Delta = 0.0625$	Avg	1,336	53.71	1,115
	Max	4,664	153.8	2,720
	Min	18.67	1.685	74.66
$\Delta = 0.125$	Avg	677.7	26.82	565.6
	Max	2,411	76.94	1,372
	Min	9.407	0.827	37.21
$\Delta = 0.25$	Avg	346.7	13.43	279.5
	Max	1,224	38.42	692.0
	Min	4.695	0.405	18.52
$\Delta = 0.5$	Avg	172.2	6.722	144.1
	Max	614.6	19.28	347.5
	Min	2.324	0.202	9.235
$\Delta = 1$	Avg	86.78	3.348	70.44
	Max	333.7	9.579	176.2
	Min	266.5	216.6	261.7
Simulation	Avg	993.5	633.9	1,020
	Max	3,978	2,101	3,951

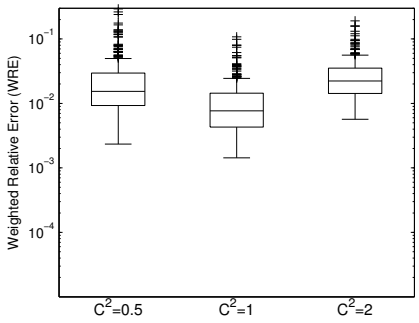
Table 3: CPU time (in sec), as a function of C^2

of the queue length.

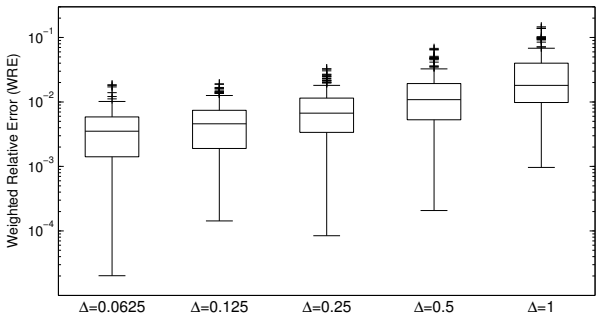
Next, we evaluate how the parameters in Table 1 affect the trade-off between accuracy and computation time. Figure 8 plots the trade-off between the accuracy and computation time of the expected queue length, for different values of the average utilization, the average service, the average capacity, and the average abandonment rate. In each plot, every observation point represents the combination of WRE and CPU time for a given value of Δ, averaged over all instances with a given parameter setting.



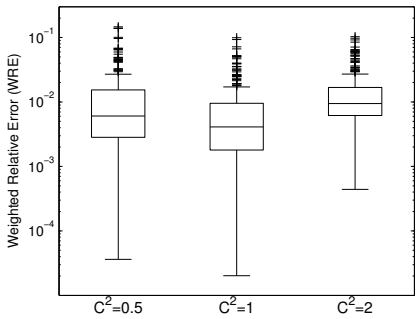
(a) Variance queue length



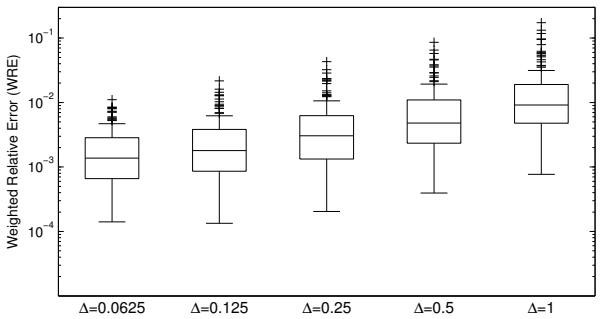
(b) Variance queue length



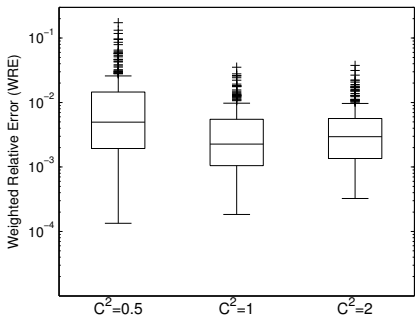
(c) Expected abandonments



(d) Expected abandonments



(e) Delay probability



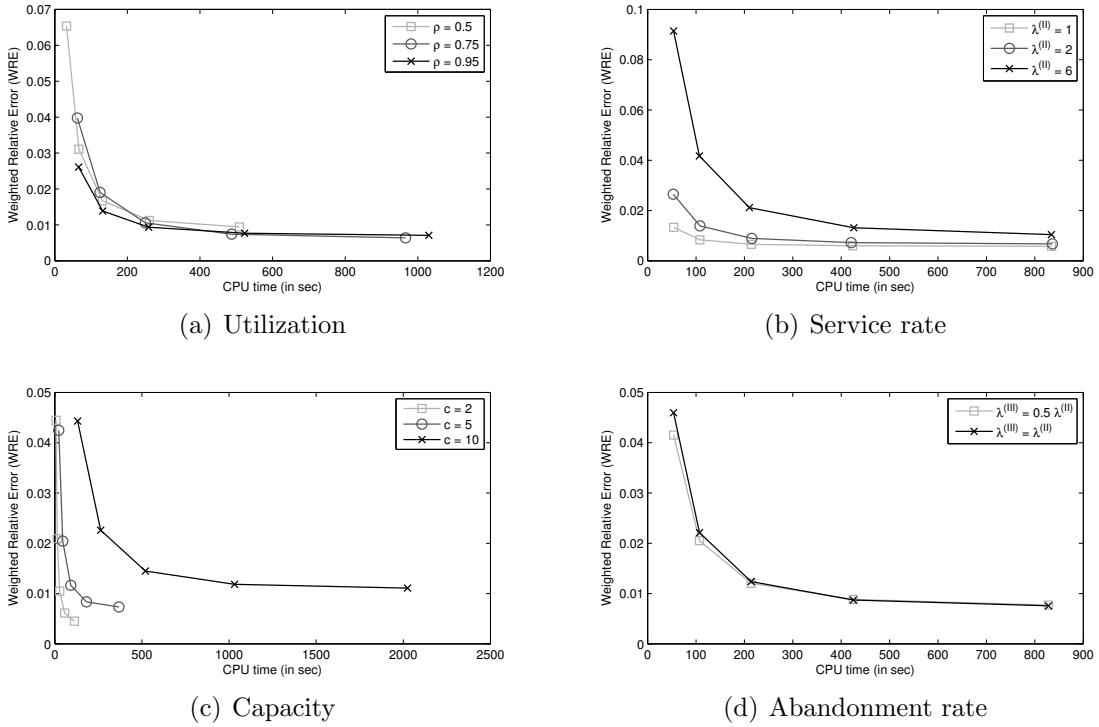
(f) Delay probability

Figure 7: WRE as a function of Δ,
and as a function of C^2 (for $\Delta = 0.0625$)

		Variance queue length (WRE)			Expected abandonments (WRE)			Delay probability (WRE)		
		$C^2 = 0.5$	$C^2 = 1$	$C^2 = 2$	$C^2 = 0.5$	$C^2 = 1$	$C^2 = 2$	$C^2 = 0.5$	$C^2 = 1$	$C^2 = 2$
$\Delta = 0.0625$	Min	0.002	0.002	0.006	0.000	0.000	0.000	0.000	0.000	0.000
	Avg	0.011	0.006	0.018	0.005	0.002	0.006	0.003	0.001	0.002
	Max	0.056	0.025	0.038	0.018	0.009	0.012	0.011	0.006	0.007
$\Delta = 0.125$	Min	0.002	0.001	0.006	0.000	0.000	0.001	0.000	0.000	0.000
	Avg	0.014	0.007	0.020	0.006	0.003	0.007	0.005	0.002	0.002
	Max	0.062	0.025	0.047	0.019	0.014	0.017	0.022	0.009	0.009
$\Delta = 0.25$	Min	0.003	0.001	0.007	0.000	0.000	0.002	0.000	0.000	0.000
	Avg	0.019	0.009	0.025	0.009	0.006	0.011	0.008	0.003	0.003
	Max	0.081	0.028	0.069	0.033	0.026	0.027	0.043	0.013	0.013
$\Delta = 0.5$	Min	0.005	0.002	0.008	0.000	0.001	0.005	0.000	0.001	0.001
	Avg	0.033	0.015	0.034	0.015	0.012	0.018	0.016	0.005	0.006
	Max	0.138	0.055	0.112	0.068	0.050	0.051	0.086	0.019	0.018
$\Delta = 1$	Min	0.008	0.004	0.009	0.001	0.003	0.009	0.001	0.001	0.001
	Avg	0.067	0.026	0.052	0.030	0.023	0.033	0.032	0.010	0.010
	Max	0.292	0.108	0.191	0.147	0.099	0.103	0.173	0.035	0.038

Table 4: WRE as a function of C^2

doi:10.1016/j.peva.2014.07.025 • www.stefancreemers.be • info@stefancreemers.be



(a) Utilization

(b) Service rate

(c) Capacity

(d) Abandonment rate

Figure 8: Trade-off between accuracy and computation time (expected queue length)

Figures 8(a) and 8(c) show that lower utilizations and/or capacity levels require less computational effort. This is not surprising, as lower utilizations and/or capacity levels also result in a smaller state space. The average service rate (see Figure 8(b)) and the average abandonment rate (see Figure 8(d)) do not impact the required computational effort. Figure 8(a) shows that lower utilizations result in a smaller accuracy for larger values of Δ. If utilization is low, the interaction between processes becomes more and more important and smaller values of Δ are required in order to maintain accuracy (see Section 4.2). Figure 8(b) and 8(c) show that smaller service rates and/or capacity yield a better accuracy. Again, this is not surprising, as a decrease in service rate and/or capacity results in a smaller event frequency. Figure 8(d) shows that the abandonment rate does not impact the computational effort required to obtain a given level of accuracy. On the one hand, small abandonment rates decrease the event frequency. They, however, also increase the utilization.

We can conclude that the trade-off between accuracy and computation time is mainly influenced by (1) the event frequency, (2) the C^2-values of the arrival, service, and abandonment process, and (3) the size of the state space. As a result, the model is most effective in settings with low service rate and/or low capacity (although other settings can also be accurately analyzed).

4.4 Impact of the PH matching procedure on model accuracy

In this section, we evaluate the impact of the PH matching procedures introduced in Section 3.2 on the accuracy of G-RAND. For this purpose, we replicate the experiment outlined in Section 4.1, using a lognormal distribution for the service and/or abandonment process. We use an exponential distribution to model the arrival process (as is common in the academic literature (Whitt, 1991; Garnett et al., 2002; Ingolfsson et al., 2007); Kim and Whitt (2014) show that this assumption is consistent with empirical arrival processes observed in call centers and emergency departments). In order to simulate the lognormal service and/or abandonment process, we adopt the following two-moment matching procedure:

$$\sigma_{\ln} = \sqrt{\ln(1+C^2)}, \tag{35}$$

$$\mu_{\ln} = \ln(\mu) - \frac{\sigma_{\ln}^2}{2}, \tag{36}$$

where $\sigma_{\ln}$ and $\mu_{\ln}$ are the shape and location parameter of the lognormal distribution respectively. Note that the skewness and excess kurtosis of the lognormal distribution only depend on $\sigma_{\ln}$ and hence, are defined by C^2 (i.e., no matter the mean, the skewness and excess kurtosis remain the same as long as C^2 does not change). The same holds for the PH distributions defined in Section 3.2. Table 5 compares the skewness and the excess kurtosis of the lognormal distribution (used in the simulation) and the PH distributions (used in G-RAND; for the two-phase Coxian distribution, we used a scaling factor $\kappa = 0.5$). It is clear that significant differences exist. In what follows, we analyze how these differences impact performance and explore how accuracy can be improved.

	Lognormal distribution		PH distribution	
C^2	Skewness	Excess kurtosis	Skewness	Excess kurtosis
0.5	2.475	12.56	1.414	3.000
1	4.000	38.00	2.000	6.000
2	7.071	156.0	3.359	16.50

Table 5: Skewness and excess kurtosis of the lognormal distribution and the PH distributions for various values of C^2

Each instance is simulated using 1,000,000 independent replications, such that the confidence interval halfwidths on the time-varying expected queue lengths are sufficiently small to conclude that the simulated metric closely approximates the "true" value (as is shown in Table 6 the largest confidence interval halfwidth is 0.00923).

Simulated queue		$C^2 = 0.5$	$C^2 = 1$	$C^2 = 2$
PH(t)/PH(t)/$s(t)$ + PH(t)	Min	0.00005	0.00016	0.00026
	Avg	0.00157	0.00180	0.00204
	Max	0.00663	0.00617	0.00666
M(t)/LN(t)/$s(t)$ + LN(t)	Min	0.00020	0.00017	0.00014
	Avg	0.00222	0.00203	0.00182
	Max	0.00921	0.00794	0.00664
M(t)/LN(t)/$s(t)$ + PH(t)	Min	0.00020	0.00017	0.00014
	Avg	0.00224	0.00208	0.00189
	Max	0.00923	0.00810	0.00698
M(t)/PH(t)/$s(t)$ + LN(t)	Min	0.00019	0.00015	0.00012
	Avg	0.00206	0.00174	0.00159
	Max	0.00790	0.00599	0.00535

Table 6: Halfwidth of the confidence interval on the time-varying expected queue lengths for different simulated queues

Table 7 reports the WRE of the expected queue length for different values of Δ and for different queues. The high WREs show that for the lognormal distribution, a simple two-moment matching procedure might not be sufficient to obtain accurate results. Moreover, the table reveals that the error introduced by the PH approximation cannot be compensated for by a decrease in the granularity parameter Δ. Table 7 also shows that the service process is least sensitive to the PH approximation. This seems to confirms the findings of Chassioti and Worthington (2004) and Chassioti et al. (2013), who suggest that, in systems with nonstationary demand and capacity, the second and higher moments of the service time distribution are relatively unimportant (note that (Chassioti and Worthington, 2004; Chassioti et al., 2013) study systems where customers balk rather than renege from the queue). In addition, our results suggest that the higher moments of the abandonment time distribution play an important role when determining the performance of a system with nonstationary demand and capacity.

To further explore the importance of the higher moments of the abandonment time distribution, we perform an additional experiment in which we vary the scale parameter (i.e., κ) of the two-phase Coxian distribution. Table 8 lists the skewness and excess kurtosis for various values of κ. A value equal to 0.9 yields the best fit with the lognormal distribution, as is confirmed in Figure 9, which plots the cumulative distribution functions of the lognormal distribution (with $C^2 = 2$) and matching two-phase Coxian distributions. Table 9 presents the WREs that result from the comparison of G-RAND (using different values of κ) and the simulated M(t)/LN(t)/$s(t)$ + LN(t) queue. It is clear that the accuracy can be increased through an adequate choice of κ (i.e., through a better matching of the higher moments of the abandonment time distribution).

doi:10.1016/j.peva.2014.07.025 • www.stefancreemers.be • info@stefancreemers.be

		M(t)/LN(t)/$s(t)$ + LN(t)			M(t)/LN(t)/$s(t)$ + PH(t)			M(t)/PH(t)/$s(t)$ + LN(t)		
		$C^2 = 0.5$	$C^2 = 1$	$C^2 = 2$	$C^2 = 0.5$	$C^2 = 1$	$C^2 = 2$	$C^2 = 0.5$	$C^2 = 1$	$C^2 = 2$
$\Delta = 0.0625$	Min	0.017	0.045	0.045	0.011	0.027	0.024	0.030	0.075	0.083
	Avg	0.068	0.145	0.126	0.022	0.044	0.050	0.079	0.173	0.163
	Max	0.168	0.306	0.254	0.044	0.077	0.110	0.174	0.319	0.274
$\Delta = 0.125$	Min	0.012	0.041	0.039	0.012	0.028	0.026	0.023	0.071	0.077
	Avg	0.065	0.143	0.124	0.025	0.045	0.052	0.076	0.171	0.162
	Max	0.166	0.305	0.253	0.054	0.083	0.110	0.173	0.318	0.273
$\Delta = 0.25$	Min	0.008	0.031	0.026	0.014	0.030	0.028	0.011	0.062	0.066
	Avg	0.061	0.140	0.120	0.032	0.049	0.057	0.071	0.168	0.158
	Max	0.164	0.303	0.250	0.075	0.096	0.111	0.171	0.316	0.271
$\Delta = 0.5$	Min	0.009	0.023	0.012	0.016	0.032	0.030	0.012	0.043	0.041
	Avg	0.057	0.134	0.112	0.045	0.056	0.066	0.064	0.162	0.150
	Max	0.162	0.301	0.246	0.122	0.121	0.129	0.168	0.315	0.268
$\Delta = 1$	Min	0.009	0.020	0.016	0.021	0.036	0.035	0.008	0.027	0.013
	Avg	0.061	0.124	0.100	0.074	0.071	0.087	0.064	0.151	0.134
	Max	0.157	0.299	0.241	0.234	0.184	0.210	0.163	0.312	0.265

Table 7: WRE of the expected queue length for lognormal service and/or abandonment process

	C^2	Skewness	Excess Kurtosis
$\kappa = 0.1$	2	2.463	8.506
$\kappa = 0.2$	2	2.643	9.767
$\kappa = 0.3$	2	2.844	11.38
$\kappa = 0.4$	2	3.076	13.51
$\kappa = 0.5$	2	3.359	16.50
$\kappa = 0.6$	2	3.730	21.07
$\kappa = 0.7$	2	4.278	29.15
$\kappa = 0.8$	2	5.268	47.82
$\kappa = 0.9^*$	2	8.026	127.4
$\kappa = 0.95$	2	13.38	397.8
Lognormal	2	7.071	156.0

Table 8: Distribution moments of the two-phase Coxian distribution with $C^2 = 2.0$ for various values of κ (*: best fit with lognormal distribution)

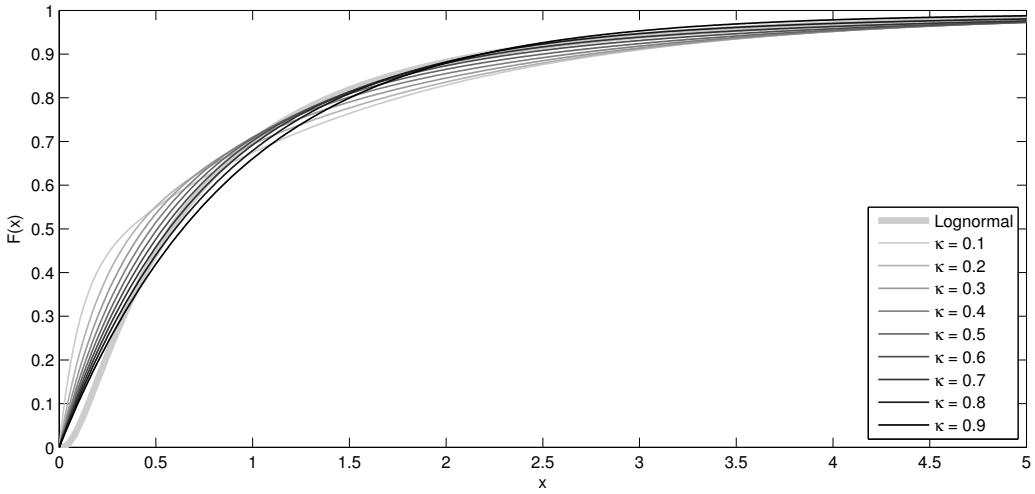


Figure 9: Cumulative distribution functions of the lognormal distribution (with $C^2 = 2$) and matching two-phase Coxian distributions (for various values of κ)

$\kappa^{(\mathrm{III})}$	WRE		
	Min	Avg	Max
$\kappa = 0.1$	0.065	0.259	0.500
$\kappa = 0.2$	0.056	0.213	0.414
$\kappa = 0.3$	0.046	0.176	0.347
$\kappa = 0.4$	0.036	0.146	0.293
$\kappa = 0.5$	0.026	0.120	0.250
$\kappa = 0.6$	0.017	0.098	0.214
$\kappa = 0.7$	0.009	0.077	0.183
$\kappa = 0.8$	0.010	0.059	0.153
$\kappa = 0.9$	0.012	0.048	0.127
$\kappa = 0.95$	0.017	0.052	0.147

Table 9: WRE of the expected queue length for various values of $\kappa^{(\mathrm{III})}$ and for the $\mathrm{M}(t)/\mathrm{LN}(t)/s(t) + \mathrm{LN}(t)$ queue where $C^2 = 2$ and $\Delta = 0.25$

We conclude that caution is advised when the higher moments of the abandonment time distribution are not adequately matched by the PH distributions. Note, however, that if the proper value of κ is selected, good results can still be obtained for $C^2 > 1$. Moreover, G-RAND can easily be extended to work with any acyclic, continuous-time PH distribution.

5 Conclusions and directions for further research

In this article, we have presented a Markov model that approximates the transient and periodic steady-state behavior of the $\mathrm{G}(t)/\mathrm{G}(t)/s(t) + \mathrm{G}(t)$ queue with exhaustive service policy. We refer to our model as G-RAND since it uses the randomization method to analyze a general queue. G-RAND yields the following time-varying performance measures: (1) the expected queue length, (2) the variance of the queue length, (3) the expected number of abandonments, and (4) the virtual waiting time distribution of a customer arriving at an arbitrary moment in time. Whereas most performance metrics can be computed with limited effort, the computation of the virtual waiting time distribution is more demanding because it requires the analysis of a death process.

A computational experiment has shown that results are highly accurate and that computational effort remains limited, especially for small- to medium-sized systems. Problem instances with a low service rate and/or a low average capacity typically required less computation time to achieve a given level of accuracy. Other problem instances can be analyzed as well, albeit at a higher computational cost. In contrast to most of the existing work, G-RAND does not rely on heavy-traffic or many-server asymptotics.

We use acyclic phase-type (PH) distributions to approximate the general interarrival, service, and abandonment time distributions. We adopt simple two-moment matching procedures, however, more complex PH distributions can be used as well (though this increases computational effort, in particular when the number of phases increases). The performance

of the model is best for settings that have moderate to high levels of process variability. Lower levels of variability require more phases and hence more computation time.

An additional experiment has shown that skewness and excess kurtosis are of crucial importance when modeling a system with nonstationary demand and capacity. Therefore, caution is advised when the skewness and excess kurtosis of the abandonment time distribution deviate from those of the PH distribution that is used to model the abandonment process. The experiment also revealed that the service process is least sensitive to the PH approximation (i.e., the higher moments of the service time distribution are of lesser importance).

Existing models are often incapable of accurately capturing the (time-varying) behavior of small- to medium-scaled systems. G-RAND is especially suited for these settings. Banks, retail stores, and emergency departments are just a few of the example systems that may benefit from our model. Our approach could, for instance, be used to evaluate the performance of alternative personnel schedules, or to determine the minimal required staffing levels. We intend to further explore G-RAND's applicability within the context of capacity planning in future research. Another avenue for future research is to study the trade-off between accuracy and CPU time. In our model, the Δ-parameter can be used to "tune" this trade-off. In a simulation model, the trade-off can also be tuned, through the number of replications. In order to identify the settings where our model offers a more favorable trade-off than simulation does, an experiment is required in which both Δ and the number of replications are varied.

Appendix: List of notation

Δ	:	Time in between two observation moments.
I	:	Arrival process.
II	:	Service process.
III	:	Abandonment process.
IV	:	Staffing process.
$G_d^{(\cdot)}$	:	Distribution of process $(\cdot)$ during epoch d.
$\mu_d^{(\cdot)}$	:	Mean process time for process $(\cdot)$ during epoch d.
$\sigma_d^{(\cdot)}$	:	Standard deviation of process times for process $(\cdot)$ during epoch d.
s_d	:	Number of servers during epoch d of the staffing process.
C^2	:	Squared coefficient of variation.
λ	:	Exponential rate parameter.
Z	:	Number of phases in the PH distribution.
β	:	Probability to visit the second phase of the two-phase Coxian distribution.
$\boldsymbol{\tau}$	:	Vector of starting probabilities of a PH distributions.
$\mathbf{R}$	:	Transient state transition matrix of a PH distribution.
$\mathbf{Q}$	:	Infinitesimal generator.
$\mathbf{t}$	:	Vector that holds the transition rates from transient states towards the absorbing state.
$\mathbf{P}$	:	Transition probability matrix.
$Z_{\max}^{(\cdot)}$	:	Maximum number of phases of process $(\cdot)$.
$s_{\max}$	:	Maximum number of servers.
$Q_{\max}$	:	Maximum queue size.
a	:	Phase of the arrival process.
$\mathbf{k}$	:	Distribution of customers over different phases of the service process.
$n_{\mathbf{k}}$	:	Sum of all entries in vector $\mathbf{k}$.
$\mathbf{K}$	:	Set of all vectors $\mathbf{k}$.
$\mathbf{b}$	:	Distribution of customers over different phases of the abandonment process.
$n_{\mathbf{b}}$	:	Sum of all entries in vector $\mathbf{b}$.
$\mathbf{B}$	:	Set of all vectors $\mathbf{b}$.
$\Pr(x, v\|u, d)$	:	Probability of having x arrivals and an arrival process at final phase v given that the arrival process starts in phase u.
$\Pr(y\|x, u, d)^{(\cdot)}$	:	Probability that y customers successfully complete phase u of process $(\cdot)$, given that x customers are present in phase u at the start.
$\Pr(\mathbf{b}^{\mathrm{s}}\|\mathbf{b}, d)$	:	Probability that $\mathbf{b}^{\mathrm{s}}$ contains the distribution of customers who have experienced the longest waiting time, conditional on $\mathbf{b}$.
$\Pr(u\|\mathbf{k})$	:	Probability to remove a server that is processing a customer who is in phase u.
$c_{(x,\mathbf{k},t)}$	:	Number of active servers removed upon a decrease of x servers.
$\Pr(y\|x, \phi_t)^{(\mathrm{II})}$	:	Probability that y out of x customers complete service.
$\pi(a, \mathbf{k}, \mathbf{b})_t$	:	Probability to visit state $(a, \mathbf{k}, \mathbf{b})_t$.
$\mathcal{Q}_w$	:	Expected queue length at the start of performance interval w.
$\mathcal{V}_w$	:	Variance of the expected queue length at performance interval w.
$\mathcal{A}_w$	:	Expected number of abandonments during performance interval w.
$\Pr(\mathcal{W}_w = h)$	:	Probability that a virtual customer who arrives at the start of performance interval w receives service after $h\Delta$ time units.

References

M. Defraeye, I. Van Nieuwenhuyse, Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm, Decision Support Systems, 54(4) (2013) 1558–1567.

L.V. Green, J. Soares, J.F. Giglio, R.A. Green, Using queueing theory to increase the effectiveness of emergency department provider staffing, Academic Emergency Medicine 13(1) (2006) 61–68.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao, Statistical analysis of a telephone call center: A queueing perspective, Journal of the American Statistical Association 100(469) (2005) 36–50.

D.C. Dietz, Practical scheduling for call center operations, Omega 39 (2011) 550–557.

S.-H. Kim, W. Whitt, Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? Manufacturing & Service Operations Management (2014), Published online in Articles in Advance 02 Jun 2014.

A. Ingolfsson, E. Akhmetshina, S. Budge, Y. Li, A survey and experimental comparison of service level approximation methods for non-stationary $\mathrm{M}(t)/\mathrm{M}/s(t)$ queueing systems with exhaustive discipline, INFORMS Journal on Computing 19(2) (2007) 201–214.

W. Whitt, The pointwise stationary approximation for $\mathrm{M}_t/\mathrm{M}_t/s$, Management Science 37(3) (1991) 307-314.

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

S. Zeltyn, A. Mandelbaum, Call centers with impatient customers: Many-server asymptotics of the $\mathrm{M}/\mathrm{M}/n + \mathrm{G}$ queue, Queueing Systems: Theory and Applications 51(3-4) (2005) 361–402.

J. Hueter, W. Swart, An integrated labor-management system for Taco Bell, Interfaces 28(1) (1998) 75–91.

I. Castillo, T. Joro, Y.Y. Li, Workforce scheduling with multiple objectives, European Journal of Operational Research 196(1) (2009) 162–170.

A. Mandelbaum, S. Zeltyn, Data-stories about (im)patient customers in tele-queues, Queueing Systems: Theory and Applications 75(2-4) (2013) 115–146.

B.K.P. Chen, S.G. Henderson, Two issues in setting call centre staffing levels, Annals of Operations Research 108(1-4) (2001) 175–192.

L.V. Green, P.J. Kolesar, W. Whitt, Coping with time-varying demand when setting staffing requirements for a service system, Production and Operations Management 16(1) (2007) 13–39.

W. Whitt, What you should know about queueing models to set staffing requirements in service systems, Naval Research Logistics 54(5) (2007) 476–484.

M. Defraeye, I. Van Nieuwenhuyse, Setting staffing levels in an emergency department: Opportunities and limitations of stationary queuing models, Review of Business and Economics 56(1) (2011) 73–100.

A. Ingolfsson, Modeling the M(t)/M/$s(t)$ queue with an exhaustive discipline, Working paper, University of Alberta, Canada (2005).

A. Jensen, Markov chains as an aid in the study of Markov processes, Skand. Aktuarietidskrift 3 (1953) 87–91.

W.K. Grassmann, Transient solutions in Markovian queueing systems, Computers & Operations Research 4(1) (1977) 47–53.

L.V. Green, P.J. Kolesar, A. Svoronos, Some effects of nonstationarity on multiserver Markovian queueing systems, Operations Research 39(3) (1991) 502–511.

L.V. Green, P.J. Kolesar, The pointwise stationary approximation for queues with nonstationary arrivals, Management Science 37(1) (1991) 84–97.

L.V. Green, P.J. Kolesar, J. Soares, Improving the SIPP approach for staffing service systems that have cyclic demands, Operations Research 49(4) (2001) 549–564.

L.V. Green, P.J. Kolesar, On the accuracy of the simple peak hour approximation for Markovian queues. Management Science 41(8) (1995) 1353–1370.

G.M. Thompson, Accounting for the multi-period impact of service when determining employee requirements for labor scheduling, Journal of Operations Management 11(3) (1993) 269–287.

L.V. Green, P.J. Kolesar, The lagged PSA for estimating peak congestion in multiserver Markovian queues with periodic arrival rates, Management Science 43(1) (1997) 80–87.

S.G. Eick, W.A. Massey, W. Whitt, The physics of the Mt/G/∞ queue, Operations Research 41(4) (1993a) 731–742.

S.G. Eick, W.A. Massey, W. Whitt, Mt/G/∞ queues with sinusoidal arrival rates, Management Science 39(2) (1993b) 241–252.

Z. Feldman, A. Mandelbaum, W.A. Massey, W. Whitt, Staffing of time-varying queues to achieve time-stable performance, Management Science 54(2) (2008) 324–338.

O.B. Jennings, A. Mandelbaum, W.A. Massey, W. Whitt, Server staffing to meet time-varying demand, Management Science 42(10) (1996) 1383–1394.

Y. Liu, W. Whitt, Stabilizing customer abandonment in many-server queues with time-varying arrivals, Working paper, Columbia University, New York, NY (2009).

D.L. Jagerman, Nonstationary blocking in telephone traffic, Bell Syst. Tech. 54 (1975) 625-661.

W.A. Massey, W. Whitt, An analysis of the modified offered-load approximation for the nonstationary Erlang loss model, The Annals of Applied Probability 4(4) (1994) 1145–1160.

W.A. Massey, W. Whitt, Peak congestion in multi-server service systems with slowly varying arrival rates, Queueing Systems 25(1) (1997) 157–172.

J.L. Davis, W.A. Massey, W. Whitt, Sensitivity to the service-time distribution in the nonstationary Erlang loss model, Management Science 41(6) (1995) 1107–1116.

W. Whitt, Engineering solution of a basic call-center model, Management Science 51(2) (2005) 221–235.

F. Iravani, B. Balcioğlu, Approximations for the M/GI/N + GI type call center, Queueing Systems 58(2) (2008) 137–153.

D. Gross, J.F. Shortle, J.M. Thompson, C.M. Harris, Fundamentals of queueing theory, 4th Edition, Wiley Series in Probability and Statistics, Wiley-Blackwell, 2008.

L.V. Green, J. Soares, Computing time-dependent waiting time probabilities in M(t)/M/$s(t)$ queueing systems, Manufacturing & Service Operations Management 9(1) (2007) 54-61.

L.F. Shampine, M.W. Reichelt, The MATLAB ODE suite, SIAM Journal on Scientific Computing 18(1) (1997) 1–22.

D. Gross, D.R. Miller, The randomization technique as a modeling tool and solution procedure for transient Markov processes, Operations Research 32(2) (1984) 343–361.

N. Izady, On queues with time-varying demand. PhD Thesis, University of Lancaster, Lancaster, UK (2010).

M.H. Rothkopf, S.S. Oren, A closure approximation for the nonstationary M/M/s Queue, Management Science 25(6) (1979) 522–534.

G.M. Clark, Use of Polya distributions in approximate solutions to nonstationary M/M/s queues, Commun. ACM 24(4) (1981) 206–217.

M. Taaffe, K. Ong, Approximating nonstationary Ph(t)/Ph(t)/l/c queueing systems, Annals of Operations Research 8(1) (1987) 103–116.

E. Chassioti, D.J. Worthington, A new model for call centre queue management, The Journal of the Operational Research Society 55(12) (2004) 1352–1357.

M. Brahimi, Approximating multi-server queues with inhomogeneous arrival rates and continuous service time distributions, PhD Dissertation, University of Lancaster, Lancaster, UK (1990).

M. Brahimi, D.J. Worthington, The finite capacity multi-server queue with inhomogeneous arrival rate and discrete service time distribution and its application to continuous service time problems, European Journal of Operational Research 50(3) (1991) 310–324.

A.D. Wall, D.J. Worthington, Using discrete distributions to approximate general service time distributions in queueing models, The Journal of the Operational Research Society 45(12) (1994) 1398–1404.

A.D. Wall, D.J. Worthington, Time-dependent analysis of virtual waiting time behaviour in discrete time queues, European Journal of Operational Research 178(2) (2007) 482–499.

S. Helber, K. Henken, Profit-oriented shift scheduling of inbound contact centers with skills-based routing, impatient customers, and retrials, OR Spectrum 32(1/4) (2010) 109–134.

W. Whitt, Fluid models for multiserver queues with abandonments, Operations Research 54(1) (2006a) 37–54.

S. Aguir, F. Karaesmen, O.Z. Akskin, F. Chauvet, The impact of retrials on call center performance, OR Spectrum 26(3) (2004) 353–376.

E. Altman, T. Jiménez, G. Koole, On the comparison of queueing systems with their fluid limits, Probability in the Engineering and Informational Sciences 15 (2001) 165–178.

T. Jiménez, G. Koole, Scaling and comparison of fluid limits of queues applied to call centers with time varying parameters, OR Spectrum 26(3) (2004) 413–422.

Y. Liu, W. Whitt, A fluid approximation for the GI(t)/GI/$s(t)$ + GI queue, Working paper, Columbia University, New York (2010).

A. Mandelbaum, W.A. Massey, Strong approximations for time-dependent queues, Mathematics of Operations Research 20(1) (1995) 33–64.

A. Mandelbaum, W.A. Massey, M. Reiman, Strong approximations for Markovian service networks, Queueing Systems 30(1) (1998) 149–201.

A. Mandelbaum, W.A. Massey, M.I. Reiman, R. Rider, Time varying multiserver queues with abandonments and retrials, Proceedings of the 16th International Teletraffic Conference 3 (1999a) 355–364.

A. Mandelbaum, W.A. Massey, M. I. Reiman, A. Stolyar, Waiting time asymptotics for time varying multiserver queues with abandonment and retrials, Proc. 37th Allerton Conf. Monticello, IL (1999b) 1095–1104.

A. Mandelbaum, W.A. Massey, M.I. Reiman, A. Stolyar, B. Rider, Queue lengths and waiting times for multiserver queues with abandonment and retrials, Telecommunication Systems 21(2-4) (2002) 149–171.

A.D. Ridley, M.C. Fu, W.A. Massey, Customer relations management: Call center operations: Fluid approximations for a priority call center with time-varying arrivals, Proceedings of the 35th Conference on Winter Simulation, New Orleans, LA, 2 (2003) 1817–1823.

Y. Liu, W. Whitt, Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid Queue with abandonment, Queueing systems 67(2) (2011b) 145–182.

Y. Liu, W. Whitt, The $G_t/GI/s_t + GI$ many-server fluid queue, Queueing Systems 71(4) (2012a) 405–444.

Y. Liu, W. Whitt, A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading, OR Letters 40 (2012b) 307–312.

Y. Liu, W. Whitt, A network of time-Varying many-server fluid queues with customer abandonment, Operations Research 59(4) (2011a) 835–846.

A.M. Law, W.D. Kelton, Simulation modeling and analysis, McGraw-Hill series in industrial engineering and management science, McGraw-Hill, Boston, 2000.

B.O. Koopman. Air-terminal queues under time-dependent conditions. Operations Research 20(6) (1972) 1089–1114.

F. McGuire, Using simulation to reduce length of stay in emergency departments, In Proceedings of the 26th conference on Winter simulation (WSC '94), M.S. Manivannan, J.D. Tew (Eds.). Society for Computer Simulation International, San Diego, CA, USA (1994) 861–867.

M.L García, M.A. Centeno, C. Rivera, N. DeCario, Reducing time in an emergency room via a fast-track, In Proceedings of the 27th conference on Winter simulation (WSC '95), C. Alexopoulos, K. Kang (Eds.). IEEE Computer Society, Washington, 1995, 1048–1053.

G.W. Evans, T.B. Gor, E. Unger, A simulation model for evaluating personnel schedules in a hospital emergency department, In Proceedings of the 28th conference on Winter simulation (WSC '96), J.M. Charnes, D.J. Morrice, D.T. Brunner, J.J. Swain (Eds.), IEEE Computer Society, Washington, 1996, 1205–1209.

S. Takakuwa, H. Shiozaki, Functional analysis for operating emergency department of a general hospital, In Proceedings of the 36th conference on Winter simulation(WSC '04). Winter Simulation Conference (2004) 2003–2011.

G.R. Hung, S.R. Whitehouse, C.B. O'Neill, A.P. Gray, N. Kissoon, Computer modeling of patient flow in a pediatric emergency department using discrete event simulation, Pediatric Emergency Care 23(1) (2007) 5–10.

M.A. Ahmed, T.M. Alkhamis. 2009. Simulation optimization for an emergency department healthcare unit in Kuwait, European Journal of Operational Research 198(3) (2009) 936–942.

M. Pitt, A generalised simulation system to support strategic resource planning in healthcare, In Proceedings of the 29th conference on Winter simulation (WSC '97), S. Andradottir, K.J. Healy, D.H. Withers, B.L. Nelson (Eds.). IEEE Computer Society, Washington, 1997, 1155–1162.

D. Sinreich, Y.N. Marmor, A simple and intuitive simulation tool for analyzing emergency department operations, In Proceedings of the 36th conference on Winter simulation(WSC '04). Winter Simulation Conference (2004) 1994–2002.

A. Fletcher, D. Halsall, S. Huxham, D. Worthington, The DH accident and emergency department model: A national generic model used locally, Journal of the Operational Research Society 58 (2007a) 1554–1562.

A. Fletcher, D.J. Worthington, What is a "generic" hospital model? Working Paper, Department of Management Science, Lancaster University, UK (2007b).

M.M. Gunal, M. Pidd, Understanding target-driven action in emergency department performance using simulation, Emergency medicine journal 26(10) (2009) 724–727.

R. Nelson, Probability, stochastic processes, and queueing theory: The mathematics of computer performance modeling, Springer Verlag New York, New York, 1995.

T. Osogami, Analysis of multiserver systems via dimensionality reduction of Markov chains, PhD thesis, School of Computer Science, Carnegie Mellon University (2005).

M.F. Neuts, Matrix-geometric solutions in stochastic models, Johns Hopkins University Press, Baltimore, 1981.

G. Latouche, V. Ramaswami, Introduction to matrix analytic methods in stochastic modeling. ASA-SIAM Series on Statistics and Applied Probability, Philadelphia, 1999.

T. Osogami, Closed form solutions for mapping general distributions to quasi-minimal PH distributions, Performance Evaluation 63(6) (2006) 524–552.

I. Gerhardt, B.L. Nelson, Transforming renewal processes for simulation of nonstationary arrival processes, INFORMS Journal on Computing, 21(4) (2009) 630–640.

R. Marie, Calculating equilibrium probabilities for $(n)/Ck/1/N$ queues, Proceedings of the 1980 international symposium on computer performance modelling, measurement and evaluation (1980), 117–125.

M.A. Johnson, M.R. Taaffe, Matching moments to phase distributions: Mixtures of Erlang distributions of common order, Stochastic Models 5(4) (1989) 711–743.

M.A. Johnson, M.R. Taaffe, Matching moments to phase distributions: Density function shapes, Stochastic Models 6(2) (1990) 283–306.

C.H. Sauer, K.M. Chandy, Approximate analysis of central server models, IBM Journal of Research and Development, 19(3) (1975) 301–313.

W. Whitt, Approximating a point process by a renewal process: Two basic methods, Operations Research, 30(1) (1982) 125–147.

T. Altiok, On the phase-type approximations of general distributions, IIE Transactions, 17(2) (1985) 110–116.

V. Ramaswami, A stable recursion for the steady state vector in Markov chains of M/G/1 type, Stochastic Models, 4(1) (1988) 183–189.

A.P.A. Van Moorsel, W.H. Sanders, Adaptive uniformization, Stochastic Models 10(3) (1994), 619–648.

H.C. Tijms, A first course in stochastic models, John Wiley & Sons, Chichester, England 2003.

F. Campello, A. Ingolfsson, Exact necessary staffing requirements based on stochastic comparisons with infinite-server models, Working paper, University of Alberta, Canada (2011).

E. Chassioti, D. Worthington, K. Glazebrook, Effects of state-dependent balking on multi-server non-stationary queueing systems. Journal of the Operational Research Society, 65 (2014) 278–290.