M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Articles in Advance, pp. 1–17
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2013.0474

Erlang-R: A Time-Varying Queue with Reentrant Customers, in Support of Healthcare Staffing

Galit B. Yom-Tov, Avishai Mandelbaum
Technion–Israel Institute of Technology, Technion City, Haifa 32000, Israel
{gality@tx.technion.ac.il, avim@ie.technion.ac.il}

We analyze a queueing model that we call Erlang-R, where the "R" stands for reentrant customers. Erlang-R accommodates customers who return to service several times during their sojourn within the system, and its modeling power is most pronounced in time-varying environments. Indeed, it was motivated by healthcare systems, in which offered-loads vary over time and patients often go through a repetitive service process. Erlang-R helps answer questions such as how many servers (physicians/nurses) are required to achieve predetermined service levels. Formally, it is merely a two-station open queueing network, which, in a steady state, evolves like an Erlang-C ($M/M/s$) model. In time-varying environments, on the other hand, the situation differs: here one must account for the reentrant nature of service to avoid excessive staffing costs or undesirable service levels. We validate Erlang-R against an emergency ward (EW) operating under normal conditions as well as during a mass casualty event (MCE). In both scenarios, we apply time-varying fluid and diffusion approximations: the EW is critically loaded and the MCE is overloaded. In particular, for the EW we propose a time-varying square-root staffing policy, based on the modified offered-load, which is proved to perform well over small-to-large systems.

Keywords: healthcare; queueing networks; modified offered-load; time-varying queues; Halfin–Whitt regime; QED regime; ED regime; emergency department staffing; mass casualty events; patient flow
History: Received: August 18, 2011; accepted: November 8, 2013. Published online in *Articles in Advance*.

1. Introduction: The Erlang-R Model

It is natural and customary to use queueing models in support of workforce management. Most common are the Erlang-C ($M/M/s$), Erlang-B ($M/M/s/s$), and Erlang-A ($M/M/s+M$) models, all used, for example, as models of call centers. But when considering healthcare environments, we find that these models lack a central prevalent feature, namely, that customers might return to service several times during their sojourn within the system. Therefore, the service offered has a discontinuous nature, because it is not provided at a single event. This has motivated our queueing model, (the time-varying) Erlang-R ("R" for reentrant customers or repetitive service), which accommodates the return-to-service phenomena.

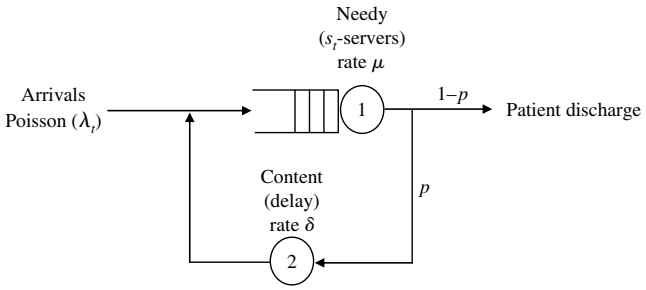
More explicitly, we consider a model where customers seek service from servers. After service is completed, with probability $1-p$ they exit the system and with probability p they return for further service, after a random delay time. We refer to the service phase as a *needy* state and to the delay phase as a *content* state (following Jennings and de Véricourt 2011). Thus, during their stay in the system, customers start in a needy state and then alternate between needy and content states. We assume that there are multiple servers in the system, and their number s_t can vary with time. When customers become needy and a server is idle, they are immediately treated by a server. Otherwise, customers wait in queue for an available server. The queueing policy is FCFS (first-come first-served). Needy service times are independent and identically distributed (i.i.d.), with general distributed G_1 and mean $1/\mu$, and content times are i.i.d. with general distribution G_2 and mean $1/\delta$. We also assume that the needy and content times are independent of each other and of the arrival process. The arrival process is a time-inhomogeneous Poisson process with rate function λ_t, $t \geq 0$; this is empirically justified, for example, in Maman (2009). Some of our results require that the needy and content times have concrete distributions (exponential, deterministic). We shall state specifically when this is the case. Figure 1 displays our system schematically.

1.1. Examples of Service Systems with Reentrant Customers

We now describe examples that underscore the practical relevance of Erlang-R: An emergency ward (EW) under normal conditions or during a mass casualty event (MCE), the radiology reviewing process, oncology bed management, and call centers.

The first example captures the complex medical service process, provided by EW physicians (or nurses) (Marmor and Sinreich 2005). We consider

Figure 1 The Erlang-R Queueing Model



separately normal and stressful EW conditions. For the first, the process starts by admitting patients and referring them to an EW physician. The physician examines them to decide between discharge versus hospitalization—a decision that could require a series of medical tests. Thus, the process that a patient experiences, from the physician's perspective, fits Erlang-R: a physician visit is a needy state; and between each visit, the patient is in a content state, which represents the delay caused by undergoing medical tests such as X-rays, blood tests, or examinations by specialists. After each visit to the physician, a decision is made to release the patient from the EW (home or hospitalized), or to direct the patient to additional tests. We shall verify later, in §6, that the *simple* Erlang-R model captures the essence of the *complete* EW process, enough to render the model useful for staffing applications.

EWs often accommodate MCEs, and these are inherently transient (Cohen et al. 2013). Based on data from an MCE drill, as described in §7.1, we demonstrate that our time-varying Erlang-R can accurately forecast MCE census and hence support its management. Ours is a chemical MCE, and these share treatment protocols that are especially amenable to Erlang-R modeling: every T minutes or so, each patient must be monitored and given an injection, where T depends on severity. (In our case, patients were triaged into four levels of severity: the most acute required treatment every 10 minutes, the second level every 30 minutes, etc.)

Our second example is the radiology reviewing process (Lahiri and Seidmann 2009). After a mammography test, the radiologist interprets the results. In some cases, part of the information on the patient is lacking: the radiologist starts the reviewing but the case must be put on hold. One then waits for this additional information to arrive, after which the reviewing process starts again. With radiologists being the servers, this can be modeled using our needy–content cycle.

The third example is the process of bed management in an oncology ward. In such a medical ward, patients return for hospitalization and treatment far more frequently than in regular wards. Here servers are the beds, the needy state models the times when a patient is in the hospital, and the content state corresponds to a patient being at home. A patient leaves the system when cured or unfortunately passes away. (A hospital colleague tells us that the same dynamics could possibly fit a geriatric ward during the flu season, when elderly patients transfer back and forth between their (nursing) home and the hospital.) Lessons from fitting Erlang-R to this and the above examples are summarized in §8.

Our prime motivation is healthcare, yet, Erlang-R is clearly relevant to other environments, for example, call center customers who return for additional services (Zhan and Ward 2012, Khudyakov et al. 2010). Note that our reentrant customers differ from what is traditionally referred to as retrial customers in queueing theory (redials in call centers) (e.g., Falin and Templeton 1997): these leave the system *prior* to service, in response to all lines being busy or after abandonment because of impatience, whereas our customers return *after* service and their returns are considered part of the service process.

1.2. Contributions

The contributions of our paper are both theoretical and practical. The main ones are as follows:

Theoretical understanding of the significance of reentrance, leading to practical insights for the above healthcare examples (§8). A central question is when must customer returns be acknowledged explicitly, as opposed to being absorbed within the service or arrival process. (This absorption has been common practice; see, e.g., Green et al. 2006.) Our important insight (§§3 and 4) is that returns become significant in time-varying systems (they are not so in a steady state)—roughly speaking, when the arrival rate varies noticeably during the sojourn-time of a customer within the system (§4.2). In particular, with periodic arrivals and exponential services, this significance is most pronounced when the period duration of the arrival process is around $\sqrt{\delta\mu(1-p)}$ (§4.3); another insight is that reentering customers smooth (reduce the amplitude of) staffing requirements over time (Theorem 5); the lessons are similar for deterministic service times but the story is then somewhat more complex (see §EC.1.5 of the Internet supplement, available at http://dx.doi.org/10.1287/msom.2013.0474).

Stabilizing performance of time-varying queueing networks via square-root staffing (SRS) rules (§5). Significantly, this has been so far proved feasible only for isolated queues (Jennings et al. 1996, Feldman et al. 2008, Whitt 2013). As explained below, the network for which performance is stabilized could be rather general—for example, the full-fledged EW network in §6. Our method requires explicit calculations of the time-varying *offered-load*, based on Massey and Whitt

(1993), as well as of key performance measures for Erlang-R (§§3 and 4).

Analytical approximations for the queue-length and number-of-busy-servers processes. These are derived separately for systems that are supercritical (e.g., EWs during MCEs as described in §7) by implementing methods from Mandelbaum et al. (1998), or systems that are well balanced, namely, quality and efficiency driven (QED; see the Internet supplement, §EC.3, which is a manifestation of the modified-offered-load (MOL) principle as in Massey and Whitt 1994).

Developing and implementing a complete framework for assessing the practical value of asymptotic queueing theory. This framework entails four network models: queueing, fluid, diffusion and simulation. To elaborate, asymptotic queueing models have been traditionally tested for *accuracy* against their mathematical origins: for example, our formulae for QED approximations (§EC.3) or transient fluid/diffusion models (§7) would have been compared, for numerical accuracy, against Erlang-R (Figure 1) steady-state formulae or transient simulation, respectively. In contrast, here we seek added value of asymptotic models rather than accuracy, which we test against a full-fledged proxy (simulation) of the complex EW reality. The added value comes about from

- stabilizing the performance of an EW in normal conditions, using staffing recommendations that are based on the QED Erlang-R (§6);
- capturing the dynamics of an EW during a chemical MCE via transient fluid and diffusion models—this utilizes radio-frequency identification (RFID)-based data from an MCE drill, which, interestingly, had to be uncensored (§7.1);
- validating the applicability (and understanding the limitations) of SRS to very small systems, e.g., with one to 10 servers (§5.2; this was first observed in Borst et al. 2004, then taken advantage of for healthcare systems in Jennings and de Véricourt 2011, and recently found theoretical explanations in Janssen et al. 2011).

Erlang-R can be viewed as a proxy for a general time-varying network from the viewpoint of a particular service station. To this end, one chooses the latter to be the needy station (e.g., physicians in our case) and the rest of the network is aggregated into the content station (the rest of the EW). The value of this approach, as discussed above, is the successful stabilization of EW performance via physician staffing that is Erlang-R generated.

2. Literature Review

The medical workforce of a hospital consists of nurses, physicians, and support staff, all jointly contributing as much as 70% to the hospital's operational budget (Israel Ministry of Health 2006). Thus, careful management of workforce capacity is called for, and here queueing models come naturally to the rescue. The first to consider the effect of returning patients in healthcare were Jennings and de Véricourt (2011). They used a closed queueing model to develop recommendations for nurse-to-patient ratios, which Yom-Tov (2010) then expanded to jointly accommodate bed allocations; both analyzed their system in a steady state. Green et al. (2006, 2007) and Zeltyn et al. (2011) consider explicitly time-varying queues in hospital staffing. They applied the Erlang-C model for staffing physicians in the EW: Green et al. (2006, 2007) using the lag-SIPP (stationary independent period-by-period) approach and Zeltyn et al. (2011) using the infinite-server approximation plus heuristics. One goal here is to demonstrate that Erlang-R is more appropriate for modeling the time-varying EW environment, which is due to the repetitive nature of service. We refer the reader to Green et al. (2007) for a comprehensive survey of time-varying queues and their applications in workforce management.

We focus on QED queues to balance patients' clinical needs for timely service against the economical preferences to operate at high efficiency. The QED regime is widely used in call centers (Gans et al. 2003). However, Jennings and de Véricourt (2011) discovered its relevance also for much smaller healthcare systems. QED queues adhere to some version of the *square-root staffing rule*, which was first analyzed by Halfin and Whitt (1981). For example, in an Erlang-C ($M/M/s$) model, the number of servers s is set to $s \approx R + \beta\sqrt{R}$; here R is the offered-load, given by $R = \lambda \cdot E[S] = \lambda/\mu$, and β is a quality-of-service parameter that is set to accommodate service-level constraints. Data from Zeltyn et al. (2011) suggest that EWs in fact use QED staffing with $0.4 < \beta < 1.6$.

When the arrival rate varies with time, it is natural to consider service-quality measures at *every moment in time*. Our goal, in this case, is to identify staffing procedures that maintain high levels of servers' utilization and, jointly, no matter what time of day customers enter the system, they will always encounter *the same* (high) service level. This goal has been addressed via two approaches. The first uses steady-state approximations, such as in PSA (piecewise stationary analysis), SIPP, or lag-SIPP (Jennings et al. 1996; Green et al. 2001, 2006). The approach works well if the system reaches a steady state quickly. The second approach includes the MOL in Jennings et al. (1996) or the infinite-server approximation of Feldman et al. (2008). Here one calculates or approximates the time-varying offered-load $R(\cdot)$, via a corresponding system with ample servers. For example, in the time-varying Erlang-C model $(M_t/M/s_t)$, $R(t) = E[\lambda(t - S_e)]E[S]$ (Eick et al. 1993b). Then one uses a time-varying adaptation of the SRS

formula: $s(t) = R(t) + \beta\sqrt{R(t)}$. This approach works very well for *single* queues, we shall apply it here to Erlang-R, which encapsulates a queueing network.

3. Steady-State Performance Measures

We start with a simple steady-state analysis of the Erlang-R model, when it is merely a two-state Jackson network. This provides the backbone for later analysis. We then present formulae for the standard quality measures of Erlang-R. We thus assume that the service times are exponentially distributed, and that the arrival rate is constant $\lambda(t) \equiv \lambda$. Let $Q = \{Q(t), t \geq 0\}$ be a two-dimensional stochastic queueing process, where $Q(t) = (Q_1(t), Q_2(t))$: $Q_1(t)$ represents the number of needy patients in the system at time t, and $Q_2(t)$ the number of content patients. Under our assumptions, the system is an open (product-form) Jackson network with the following steady-state distribution:

$$\pi_{ij} := P(Q_1(\infty) = i, Q_2(\infty) = j) = \frac{(R_1)^i}{\nu(i)} c_1 \frac{(R_2)^j}{j!} c_2,$$

where

$$\begin{aligned} c_1 &= \left[\frac{(R_1)^s}{s!(1 - R_1/s)} + \sum_{i=0}^{s-1} \frac{(R_1)^i}{i!} \right]^{-1}, \\ c_2 &= \left[\sum_{j=0}^{\infty} \frac{(R_2)^j}{j!} \right]^{-1} = e^{-R_2}, \end{aligned} \tag{1}$$

where $\nu(i)$ is defined as $\nu(i) := (i \wedge s)s^{(i-s)^+}$, and $R_1 = \lambda/((1-p)\mu)$, $R_2 = (p\lambda)/((1-p)\delta)$. We call R_1 and R_2 the *steady-state offered-load* of stations 1 and 2, respectively. Now let W_t be the waiting time for service of a (virtual) customer who becomes needy at time t (either upon first arrival or returning); let $W = \lim_{t\to\infty} W_t$ denote the corresponding steady-state waiting time (weak limit).

Theorem 1. *Assume that* $S_1 \stackrel{d}{=} \exp(\mu)$ *and* $S_2 \stackrel{d}{=} \exp(\delta)$, *and the arrival rate is constant* λ. *Then*

$$\begin{aligned} \alpha := P(W > 0) &= \left[\frac{(R_1)^s}{s!(1 - R_1/s)} \right] c_1, \\ E[W \mid W > 0] &= \frac{1}{\mu s(1-\rho)}, \\ (W \mid W > 0) &\stackrel{d}{=} \exp(E[W \mid W > 0]), \end{aligned}$$

where $\rho = R_1/s$, *and* c_1 *is defined in* (1). (*Here* $\stackrel{d}{=}$ *denotes equality in distribution*.)

Proof: Theorem 1 is a straightforward result of Erlang-R being a two-node Jackson network jointly with the arrival theorem for open Jackson networks.

In a steady state, node 1 is an *M/M/s* queue with parameters $(\lambda, \mu(1-p), s)$, and node 2 is an *M/M/*∞ queue with parameters $(\lambda, ((1-p)\delta)/p)$. It follows that, in a steady state, the appropriate QED staffing policy for our model sets $s = R_1 + \beta\sqrt{R_1}$, $\beta > 0$, where β is related to the desired α by

$$\alpha = \left[1 + \beta \frac{\Phi(\beta)}{\phi(-\beta)} \right]^{-1}; \tag{2}$$

here $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions, respectively, (Halfin and Whitt 1981). Hence, in a steady state, the staffing recommendations of Erlang-R and Erlang-C coincide.

For every Erlang-R with parameters $(\lambda, \mu, p, \delta)$, there are two naturally corresponding Erlang-C models: one with parameters $(\lambda, \mu(1-p))$, in which successive services are concatenated with no delay between them; the second has parameters $(\lambda/(1-p), \mu)$, in which the number of arrivals is amplified appropriately. Only the first option, with concatenated services, will be considered from now on; we refer to this model as *multiservice* Erlang-C. (The second option turns out to be an inferior fit over finite horizons, which was verified via simulations.)

4. The Offered-Load

As mentioned earlier, staffing levels that are based on the time-varying offered-load, do stabilize performance of nonstationary systems. Adopting this approach, we now introduce the offered-load function of our time-varying Erlang-R model, denoted $R = \{R(t), t \geq 0\}$. Here $R(t) = (R_1(t), R_2(t))$, where $R_i(t)$ is the offered-load of node i at time t. The function $R(\cdot)$ is defined in terms of a related system, with the same structure as ours, but in which the number of servers in node 1 is infinite, which results in an $(M_t/G/\infty)^2$ network: $R_i(t)$ is simply the average number of busy servers (served customers) in this latter network, in node i at time t; equivalently, $R_i(t)$ equals the average *least number* of servers that is required so that no arriving customer is delayed in queue prior to service.

We now calculate R under various scenarios:

4.1. The Offered-Load for General Arrivals and Exponential Services

Assume that S_i are exponentially distributed. The Erlang-R model is then a time- and state-dependent Markovian service network (Mandelbaum et al. 1998), for which the following holds:

Theorem 2. *Assume that* $S_1 \stackrel{d}{=} \exp(\mu)$ *and* $S_2 \stackrel{d}{=} \exp(\delta)$. *Then* $R(\cdot)$ *is given by the unique solution of the following ordinary differential equation* (*ODE*): *for* $t \geq 0$,

$$\begin{aligned} \frac{d}{dt} R_1(t) &= \lambda_t + \delta R_2(t) - \mu R_1(t), \\ \frac{d}{dt} R_2(t) &= p\mu R_1(t) - \delta R_2(t). \end{aligned} \tag{3}$$

The initial condition is determined by the originating system.

Proof. See the Internet supplement, §EC.1.1.

With general time-varying arrival rates, the ODE (3) is unlikely to be tractable analytically. Nevertheless, one can easily solve it numerically. We used this method for the experiments in §§5 and 6.

4.2. The Offered-Load for General Arrivals and General Services

Let J denote the number of returns to service, thus $J \stackrel{d}{=} Geom_{\geq 0}(1-p)$.

Theorem 3. *The offered-load $R(\cdot)$ is given by*

$$\begin{aligned} R_1(t) &= E\left[\sum_{j=0}^{\infty} p^j \lambda(t - S_1^{*j} - S_2^{*j} - S_{1,e})\right] E[S_1] \\ &= \frac{E[S_1]}{1-p} E[\lambda(t - S_1^{*J} - S_2^{*J} - S_{1,e})], \\ R_2(t) &= E\left[\sum_{j=1}^{\infty} p^j \lambda(t - S_1^{*j} - S_2^{*j-1} - S_{2,e})\right] E[S_2] \\ &= \frac{E[S_2]}{1-p} E[\lambda(t - S_1^{*J} - S_2^{*J-1} - S_{2,e})], \end{aligned} \tag{4}$$

where $S_{i,e}$ is a random variable representing the excess service time at node i, S_i^{*j} is the sum of j i.i.d random variables S_i (the j-convolution of S_i), and all these random variables are assumed independent.

Proof. This theorem follows from Massey and Whitt (1993). For completeness, we provide a proof in the Internet supplement, §EC.1.1.

Proposition 1. *A second-order Taylor-series approximation of $R_1(\cdot)$ is given by*

$$R_1(t) \approx \frac{E[S_1]}{1-p}\left[\lambda(t - E[S_{1,e} + S_1^{*J} + S_2^{*J}]) + \frac{1}{2}\lambda^{(2)}(t)\,\mathrm{Var}[S_{1,e} + S_1^{*J} + S_2^{*J}]\right]. \tag{5}$$

Proof. See the Internet supplement, §EC.1.1.

Approximation (5) reveals a fundamental difference between the offered-loads of Erlang-R and its corresponding Erlang-C. The multiservice Erlang-C second-order approximation is $R(t) \approx (E[S_1]/(1-p)) \cdot [\lambda(t - E[S_{1,e}^{*J}]) + \frac{1}{2}\lambda^{(2)}(t)\,\mathrm{Var}[S_{1,e}^{*J}]]$. This results from adjusting the Erlang-C formula in Whitt (2007) to the case where the service time is a random sum of i.i.d. (partial) service durations. We thus observe that Erlang-R corrects the time gap, relative to time t; it extends this gap further by S_2^{*J}, namely, the overall time spent in the content state during a customer's sojourn. It follows that time-varying approximations of the offered-load, which are based on Erlang-C, are potentially inaccurate in both time lag and magnitude—this will be confirmed in the sequel.

4.3. Analysis of Special Cases and Managerial Insights: Sinusoidal Arrival Rate

In this section, we analyze the offered-load for the special case of a sinusoidal arrival rate function. There are several reasons for using the sine function. First, any periodic time-varying arrival rate (hence the corresponding offered-load) can be approximated by a finite linear combination of sine functions, thus leading to a Fourier expansion of the offered-load. Second, sine functions yield closed-form solutions to the offered-load (in some special cases). This, in turn, reveals the role that the amplitude and frequency of the arrival rate, in conjunction with service and content time, play in our system evolution (§4.3.1). Specifically, all these parameters jointly specify the amplitude and phase of the offered-load function, which, in turn, determines magnitude changes in staffing levels and the timing of such changes. This explains and quantifies the gap and its magnitude between peak arrival rate and peak offered-load, hence consequent peak staffing. Finally, our closed forms enable a comparison between Erlang-R and the corresponding multiservice Erlang-C, thus highlighting the influence of returning customers and the circumstances under which Erlang-R is a modeling necessity—as opposed to absorbing returns into exogenous arrivals (§4.3.2).

Assume that

$$\lambda(t) = \bar{\lambda} + \bar{\lambda}\kappa \sin(2\pi t/f) = \bar{\lambda} + \bar{\lambda}\kappa \sin(\omega t), \quad t \geq 0, \tag{6}$$

where $\bar{\lambda}$ is the average arrival rate, κ is the relative amplitude, f is the period, and $\omega = 2\pi/f$ is the frequency. (We are assuming here, without loss, that the phase of the arrival rate is 0.) Substituting this arrival rate into (4) yields

$$\begin{aligned} R_1(t) = {} & \frac{\bar{\lambda}}{1-p} E[S_1] \\ & + E[S_1]\bar{\lambda}\kappa \sum_{j=0}^{\infty} p^j E[\sin(\omega(t - S_{1,e} - S_1^{*j} - S_2^{*j}))]. \end{aligned} \tag{7}$$

We now provide explicit solutions for $R(\cdot)$ in the case of exponential service times. (Deterministic service times are also amenable to the analysis; then the amplitude and phase behavior of $R(\cdot)$ is also interesting, but less realistic and, therefore, is only hinted at in the Internet supplement, §EC.1.5.)

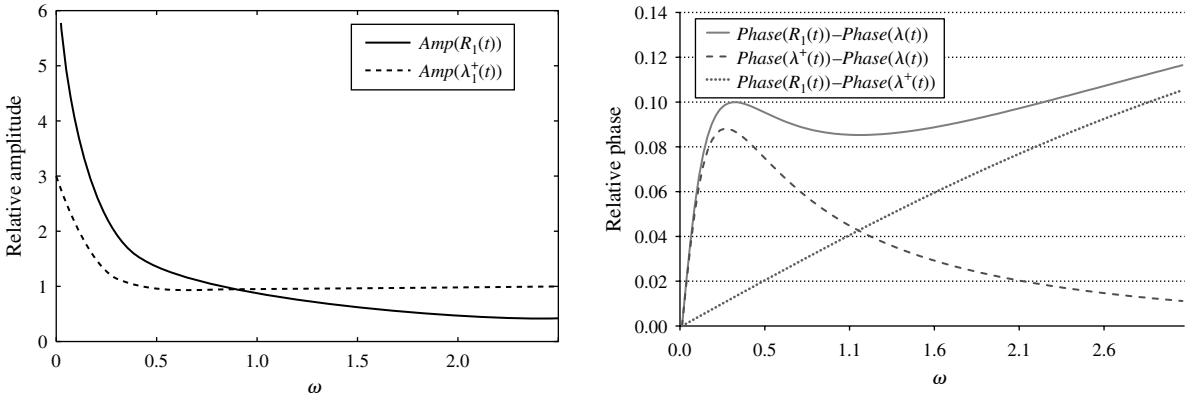
4.3.1. Exponential Service Times.

Theorem 4. *Assume that $\lambda(\cdot)$ is given in* (6)*, and $S_1 \stackrel{d}{=} \exp(\mu)$ and $S_2 \stackrel{d}{=} \exp(\delta)$. Then* (7) *has the following form*:

$$\begin{aligned} R_1(t) = {} & \frac{E[S_1]\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \\ & \cdot \sqrt{\frac{\delta - i\omega}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} \cdot \frac{\delta + i\omega}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}} \\ & \cdot \cos(\omega t + \pi + \tan^{-1}(\theta)), \end{aligned} \tag{8}$$

where $\theta = \mu(\delta^2 - p\delta^2 + \omega^2)/(\omega(\delta^2 + \omega^2 + p\mu\delta))$.

Figure 2 Relative Amplitude and Phase of $R_1(\cdot)$ and $\lambda_1^+(\cdot)$ as a Function of ω



Proof. The results follow from applying the characteristic function of the Exponential and Erlang distributions to (7). See the Internet supplement, §EC.1.2.

Therefore, the amplitude of $R_1(\cdot)$ is

$$Amp(R_1) = \bar{\lambda}\kappa\sqrt{\frac{\delta - i\omega}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} \cdot \frac{\delta + i\omega}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}} \quad (9)$$

and its phase is

$$Phase(R_1) = \frac{1}{2\pi}\cot^{-1}\left(\frac{\mu(\delta^2 - p\delta^2 + \omega^2)}{\omega(\delta^2 + \omega^2 + p\mu\delta)}\right).$$

A similar calculation for $\lambda_1^+(t)$ ($\lambda_i^+(\cdot)$ is the aggregated-arrival-rate function to node i) is provided in Theorem EC.1 of the Internet supplement, §EC.1.2. Theorem 4 yields a simple relation between the amplitudes of $R(\cdot)$ and $\lambda_1^+(\cdot)$: $Amp(R_1) = Amp(\lambda_1^+) \cdot \sqrt{\mu^2 + \omega^2}$, which separates two influences on the offered-load amplitude: $Amp(\lambda_1^+)$ is associated with returning customers and $\sqrt{\mu^2 + \omega^2}$ with the last service before departure. The right diagram of Figure 2 shows an analogous but additive relation between phases: the phase of $R_1(\cdot)$ is the sum of the phase shift between $\lambda_1^+(\cdot)$ and $\lambda(\cdot)$ (due to returning customers) with the phase shift between $R_1(\cdot)$ and $\lambda_1^+(\cdot)$ (last service). As indicated, phases determine timing of required staffing: a large phase corresponds to a long time lag between the peak of the arrival rate and the peak of staffing. We observe that the influence of the returning customers decreases and vanishes as $\omega \uparrow \infty$ (both in amplitude and phase).

In the the Internet supplement, §EC.1.4, we elaborate on the amplitude of $R_1(\cdot)$ and $\lambda_1^+(\cdot)$. We analyze limiting cases. We show that both amplitudes are decreasing functions of ω, and that the amplitude of $R_1(\cdot)$ is an increasing function of δ.

4.3.2. When Is Erlang-R Necessary? (Comparing to Erlang-C). We now compare amplitudes and phases of the offered-loads for Erlang-R with those of the multiservice Erlang-C model. The amplitude of the offered-load in Erlang-C, with arrival rates (6) and service rate $\mu_c = (1-p)\mu$, is given by $Amp(R_c) = \bar{\lambda}\kappa/\sqrt{\mu_c^2 + \omega^2}$, and its phase is $\theta_c = (1/(2\pi)) \cdot \cot^{-1}(\mu_c/\omega)$ (Eick et al. 1993a). The ratio between the amplitudes and phases are thus given by

$$AmpRatio = \frac{Amp(R_1)}{Amp(R_c)} = \frac{\bar{\lambda}\kappa\sqrt{\frac{\delta - i\omega}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta}\frac{\delta + i\omega}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}}}{\frac{\bar{\lambda}\kappa}{\sqrt{((1-p)\mu)^2 + \omega^2}}}, \quad (10a)$$

$$PhaseRatio = \frac{Phase(R_1)}{Phase(R_c)} = \frac{\cot^{-1}\left(\frac{\mu(\delta^2 - p\delta^2 + \omega^2)}{\omega(\delta^2 + \omega^2 + p\mu\delta)}\right)}{\cot^{-1}\left(\frac{(1-p)\mu}{\omega}\right)}. \quad (10b)$$

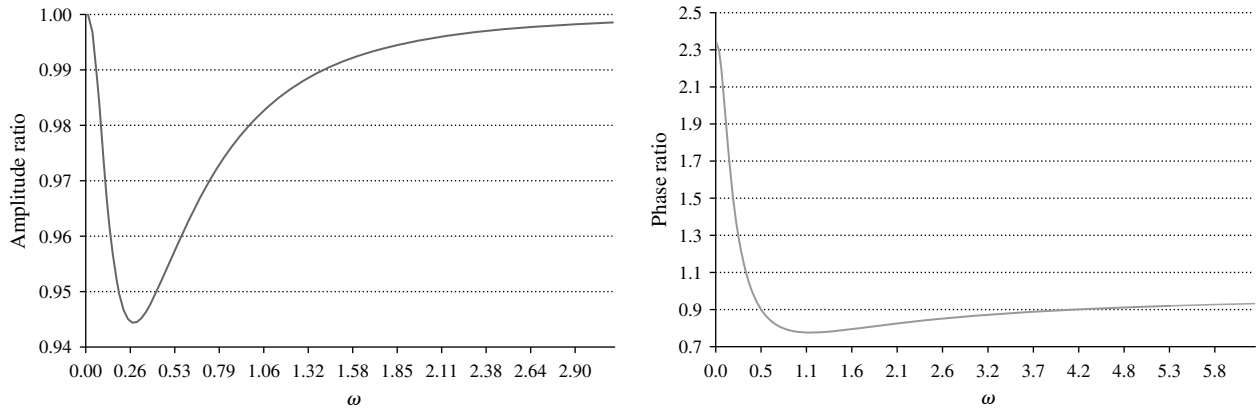
Theorem 5. *Assume that the arrival rate is sinusoidal and service times are exponential. Comparing Erlang-R with parameters* $(\lambda, \mu, p, \delta)$ *against the* (*multiservice*) *Erlang-C model with parameters* $(\lambda, (1-p)\mu)$:

1. *The amplitude of the offered-load in Erlang-R is always smaller than that of the multiservice Erlang-C.*
2. *The amplitude ratio attains its minimal value when* $\omega = \sqrt{\delta\mu(1-p)}$.
3. *Both amplitude and phase ratios approach one as* $\omega \uparrow \infty$ *or* $\delta \uparrow \infty$. *The amplitude ratio also approaches one as* $\omega \downarrow 0$.

Proof. All results follow from analyzing Equations (10a) and (10b); see the Internet supplement, §EC.1.3.

The first part of the theorem implies that returning customers have a *stabilizing* effect on the system. This means that the difference between high and low

Figure 3 Ratio of Amplitudes and Phases Between Erlang-R and Erlang-C as a Function of ω (Case Study 1, §5.1)



staffing levels is smaller when customers reenter service, which alleviates staffing scheduling decisions. An example of the difference between the amplitudes is given in the left diagram of Figure 3. Having a smaller amplitude means that for one part of the cycle, $R_1(\cdot)$ is higher, and in the other part $R_c(\cdot)$ will be higher (as we show later in Figure 5). The implication is that Erlang-C will both overstaff or understaff. The impact of this observation on the service level is further explored in §5; it shows that one must take into account the repetitive nature of service to avoid excessive staffing costs or undesirable service levels.

The second part of the theorem identifies the cases in which the difference between the amplitudes is maximal. In particular, for periodic arrivals, this difference is most pronounced when the period duration of the arrival process is a square-root order of the multiplication of needy service time, content time, and the average number of services. In such cases, the arrival rate varies significantly over the sojourn of a customer within the system.

The phase ratio, as a function of ω (see the right diagram of Figure 3), exceeds one up to $\omega = \sqrt{(2\delta^2 + p(1-p)\delta\mu)/p}$, and from that point on it is smaller than one. Therefore, for certain values of ω, the Erlang-C offered-load leads that of Erlang-R and for other values it lags behind.

From the last part of the theorem and Figure 3, we gain an understanding of when the influence of returning customers is not significant, and thus does not require the use of the Erlang-R model. We observe that if $\omega \uparrow \infty$, or $\delta \uparrow \infty$, the difference between the offered-load of Erlang-R and Erlang-C becomes negligible. An intuitive explanation for this finding is that when $\omega \uparrow \infty$, the arrival rate changes so rapidly that its changes are assimilated in the variance of the arrival process. In this case, the offered-load becomes constant; this is true for both Erlang-C and Erlang-R. As $\delta \uparrow \infty$, customers immediately return to the needy state; thus the system behaves as if the services were concatenated into a single exponential ($(1-p)\mu$) service. The limit $\omega \downarrow 0$ is interesting as well: here the amplitude ratio does indeed converge to one, but the phase ratio need not. (All the above observations will be used, in §8, to analyze the significance of Erlang-R in the healthcare examples of §1.1.)

5. Validation of MOL Staffing

We now propose a staffing procedure for the time-varying Erlang-R model, which we validate via several examples. We propose the use of the SRS with MOL approximation (e.g., Massey and Whitt 1994). We shall compare it to two other approaches: time-varying Erlang-C and PSA approximation. Importantly, MOL has been proven effective for staffing (time-varying) isolated queues. It has not been previously tested for time-varying queues within queueing networks, which is what we do here.

The MOL algorithm for Erlang-R runs simply as follows:

1. Calculate the time-varying offered-load $R(\cdot)$, generally by (4) or approximately via (3) or (5).

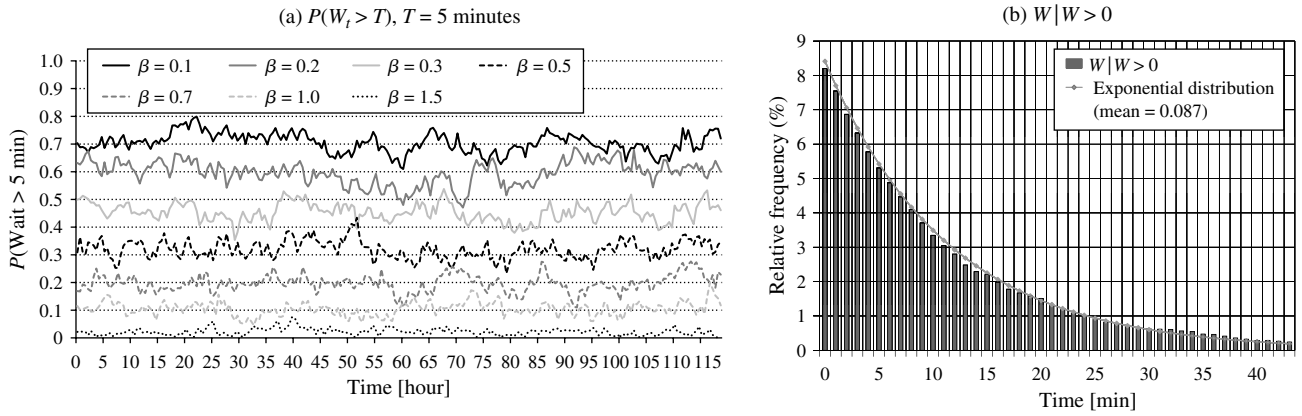
2. Staff the needy station according to the SRS formula: $s(t) = R_1(t) + \beta\sqrt{R_1(t)}$, $t \geq 0$, where β is chosen according to the steady-state Halfin–Whitt formula (2). (This follows from the needy part of Erlang-R having the same steady-state distribution of the multiservice Erlang-C.)

We use simulation to validate our approach. The first example (§5.1) serves as a proof of concept and does not mirror the hospital environment: it is too large of a system. The second (§5.2) is a small system with an arrival-rate shape that is taken from hospital data, and the third example (§6) is an actual EW.

5.1. Case Study 1—Large System

In this case study, we validate our assumption that the MOL algorithm stabilizes network performance over time, showing along the way that Erlang-R must be used in time-varying environments. We use a stylized

Figure 4 Case Study 1—Simulation Results of $P(W_t > T)$ for Various β Values and $W \mid W > 0$ in Large Systems



sinusoidal arrival rate (6). This example has a relatively large $\bar{\lambda}$ since we wish to start our validation process with a system where the asymptotic approximations are expected to work well. The parameters of this experiment are $\bar{\lambda} = 30$ customers per hour, $p = 2/3$, $\kappa = 0.2$, $f = 24$ hours, $\mu = 1$, $\delta = 0.5$, and $0.1 \leq \beta \leq 1.5$; 100 replications were generated for each β value.

We find that for a large enough system in the QED regime ($\beta > 0.3$), the MOL approach stabilizes all performance measures of the Erlang-R queueing network. Consequently, *any* prespecified QED service level can be achieved *stably over time*. For example, Figure 4(a) shows the empirical $P(W_t > T)$, the fraction of needy arrivals at time t, who are delayed in queue more than T units of time. This fraction was calculated over a five-day period, for various values of β. We note that $P(W_t > T)$ is relatively stable for all β tested. Figure 4(b) shows the conditional distribution of the waiting time given delay ($W \mid W > 0$), when $\beta = 0.5$. (It is calculated over all arrivals during the five-day period.) We compare it to the steady-state theoretical distribution, which is exponential with rate $s\mu(1-\rho)$ (as stated in Theorem 1). The simulation results depict the distribution of waiting times from all replications, over the entire time horizon. We observe a very good fit in the QED regime (here $\beta = 0.5$). Other performance measures are also considered in the Internet supplement, §EC.2. The reason for success appears to be that the time-varying SRS controls the system, at all times, in a state that is very close to a naturally corresponding *steady-state system*. This also explains why the constant β is calculated using steady-state formulae, and it need not vary in time.

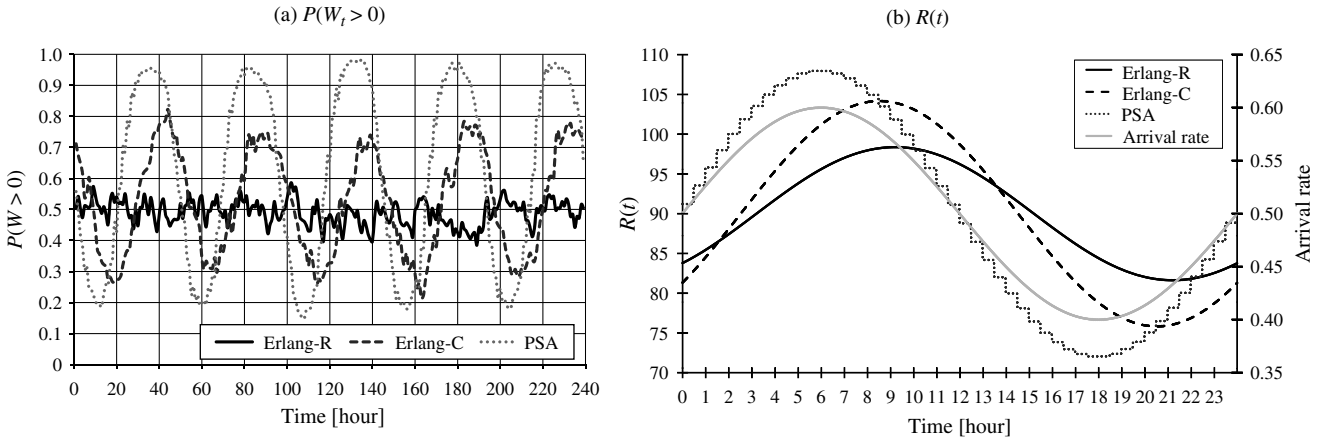
Remark. Although the above performance measures, under MOL QED staffing, are close to being constant over time, it is important to understand that the total number of customers in the system *does* vary over time. Specifically, the number of customers turns out to be accurately described by $E[Q_1(t)] = R_1(t) + \alpha(R_1(t)/s(t))(1 - (R_1(t)/s(t)))^{-1}$; see the Internet supplement, §EC.3, for more details.

Comparing Erlang-R, Erlang-C and PSA staffing. In applications, researchers have used Erlang-C to model systems in which customers return multiple times for service. For example, Green et al. (2001, 2007) used Lag-SIPP for staffing EW physicians. We now compare the outcome of using Erlang-R staffing against that of using Erlang-C staffing, the latter based on one of two methods: MOL and PSA. The performance measure we focus on is the delay probability, setting its target level to 0.5 (hence $\beta = 0.5$). Figure 5(a) shows that, whereas using Erlang-R stabilizes system performance around the prespecified target, using Erlang-C or PSA does not. PSA performs the worst (resulting in the least stable system), because PSA staffing does not take into account either the time lag or the reentrant effects. We explain the performance differences by considering the offered-load function $R(\cdot)$ (Figure 5(b)). We observe that for one half of the cycle, Erlang-C overestimates $R(\cdot)$, resulting in overstaffing, which, in turn, results in a better performance than the prespecified target. However, in the other half cycle, the opposite occurs, causing the performance to be worse than prespecified. Erlang-R, in contrast, stabilizes performance over the whole time horizon. (These observations also follow from our theoretical analysis in §4.3.2.) The conclusion again is that one must take into account the repetitive nature of service.

5.2. Case Study 2—Small System; Hospital Arrival Rates

In the second case study, we investigate the use of the MOL algorithm in small systems, specifically in setting staffing levels for EW physicians. To this end, we consider the actual arrival rate function of the emergency ward in Figure 6. The values for p, μ, and δ were inferred from that EW data.

Figure 5 Case Study 1—Comparing Erlang-R, Erlang-C, and PSA



There are obvious problems in applying our MOL approach to small systems: First, our approximations are expected to be less accurate, being limits as systems grow indefinitely. (In our simulation, the number of servers changes between one and eight.) Second, rounding up a "theoretical" need of say 1.5 servers to two servers means adding 30% excess capacity to the required capacity, which suggests difficulties in stabilizing performance around prespecified values. Related to this is the fact that the set of achievable performance measures is manifestly discrete for small systems: changing the staffing level of a small system by a single server could discontinuously change its performance. For example, if the offered-load is $R = 2.75$, the values that $P(W > 0)$ can have are shown in Table 1. Finally, one cannot have an EW operate with no physicians, and for small servers this lower bound of one plays a binding role. It is therefore unclear whether, under these circumstances, we shall still be able to stabilize system performance around a predetermined value. Nevertheless, we found that it is possible to stabilize even such small systems, given specific (though not all, as expected) target performance levels. The performance measures are relatively stable, and the four possible scenarios are visibly separable. (Because of space limitations, we have not included supporting graphs; furthermore, Figure 9(a) in §6 well demonstrates these phenomena in an even more complex environment.)

There is another important impact of system size that we observed in this case study. When verifying whether the relation between actual $P(W > 0)$ and β fits the Halfin–Whitt formula, we note a gap between the two (see the left diagram in Figure 7). The left plot in Figure 7 shows the relationship between these functions, when we consider the target β values used in the square-root formula. In most cases, the empirical function is shifted downward, and the gap between the two is reduced as β grows. This is mainly due to the rounding procedure. The right plot of Figure 7 shows the same graph, but as a function of the effective β values. We observe that the two functions have the same shape but the empirical function is shifted upward. The gap between them appears to be constant. As this seems to be the effect of using asymptotic approximations in such a small system, we also applied the refined approximations of Janssen et al. (2011). This caused the gap to narrow, but it is still noticeable.

The practical guideline that can be derived from these graphs is that, when targeting a specific $P(W > 0)$ value, one should use a smaller value of β,

Figure 6 Case Study 2—Plot of Arrival Rates in an Emergency Ward

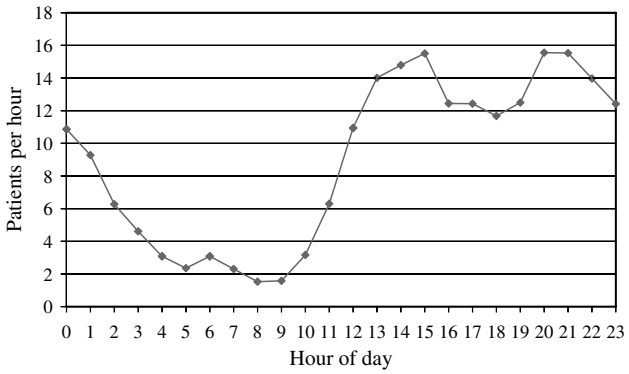
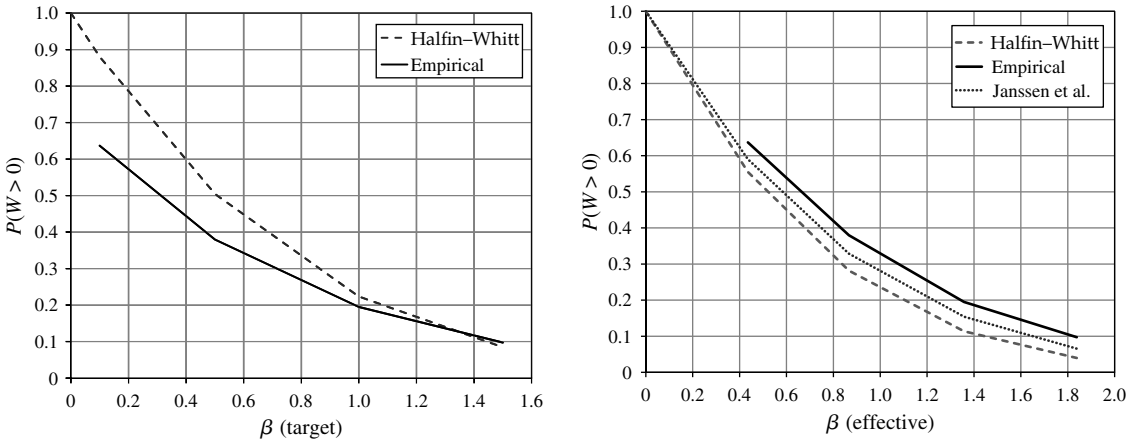


Table 1 Small Systems: An Example of a Discrete Range for $P(W > 0)$, as a Function of β

Target β range	Effective β	s	$P(W > 0)$ (%)
(0.474, 1.055]	1.055	4	34.0
(1.055, 1.658]	1.658	5	11.4
(1.658, 2.261]	2.261	6	3.0
1.658 and up	∞	7	0

Note. We distinguish between *target* β and *effective* β; the latter is the β actually used, calculated by $(\beta = (\lceil s \rceil - R_1)/\sqrt{R_1})$.

Figure 7 Case Study 2—Comparison of the Erlang-R Simulation to the Formulae in Halfin and Whitt (1981) and Janssen et al. (2011)



based on the left diagram of Figure 7. More research is also needed to understand the Halfin–Whitt (and Janssen et al. 2011) function for small systems while also considering the rounding effect. As a first step, one can develop graphs such as Figure 7, using a steady-state simulation of an Erlang-C model.

6. Using Erlang-R for Staffing EW Physicians: Fitting a Simple Model to a Complex Reality

In this last case study, we test Erlang-R as a support tool for planning a real system. Specifically, we demonstrate that it can be used to practically plan staffing of physicians in an EW, although the real system is far more complicated than our model. In passing, we show that applying Erlang-C to the real system is inferior to Erlang-R. The EW system was briefly described in our introduction; for a complete description see Marmor and Sinreich (2005). In our experiment, we use their accurate and detailed EW simulation model (it takes into account even walking distances), which is flexible in that it is easily adapted to a given EW. We fit the simulator to the EW of our partner Israeli hospital (Armony et al. 2011), and then use the simulator as an accurate portrait of the complex EW reality.

Clearly, many of our main assumptions do not hold in the EW environment. For example, service times are not exponentially distributed and could depend on the load in the EW, as follows from Armony et al. (2011). Moreover, there are seven types of patients that seek EW services, and each type goes through a different routing process during their sojourn. The physicians are divided into four groups, according to their expertise. There is an explicit connection between a patient type and a physician group. We now simplify this complex system into an Erlang-R by setting parameter values, for each physician type *separately*, as follows:

- Arrival rate: $\lambda(\cdot)$ is the average arrival rate for each hour of the day, for each physician group, as shown in Figure 8.
- Needy times: $E[S_1] = 1/\mu$ is estimated by averaging all services given by a specific physician group.
- Content times: $E[S_2] = 1/\delta$ is the average time between successive visits of a patient to the physician.
- Probability of returning to the physician for an additional service: p is deduced from the average number of visits of patients to their physician, which we take to be $1/(1-p)$ and solve for p.

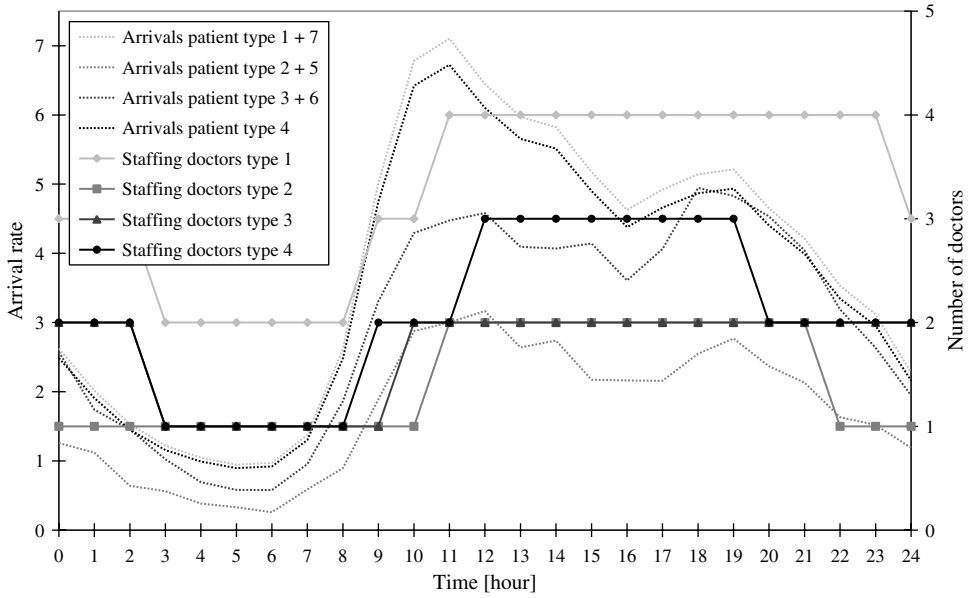
Table 2 specifies the estimated parameters according to physician type. We calculated (simply via a spreadsheet) the offered-load using the differential Equations (3), and ran the staffing recommendation with our EW simulation. We assumed that changes in staffing could be implemented in a one-hour resolution. For each interval, we calculated the average number of physicians needed and rounded *up* to the nearest integer. We used one replication of 100 weeks. (The first setup week was excluded.)

Figure 8 shows the arrival rate and the recommended number of physicians during the day, for each type of physician, with $\beta = 0.5$. The number of physicians varies between one and four. We observe that the staffing function lags behind the arrival rate function, with an approximate time lag of two hours. Note that the number of physicians does not change

Table 2 Emergency Ward Simulation Parameters

Physician type	Patient type	μ	$E[S_1]$ [hour]	δ	$E[S_2]$ [hour]	p
1	1, 7	8.91	0.112	0.953	1.049	0.7743
2	2, 5	8.86	0.113	0.969	1.031	0.6094
3	3, 6	10.33	0.097	0.572	1.749	0.6441
4	4	12.37	0.081	1.310	0.763	0.7268

Figure 8 Emergency Ward Case Study—Patient Arrivals and Physician Staffing for Each Physician Type in Emergency Ward Simulation ($\beta = 0.5$)



every hour, and natural shift schedules could be derived to fit this graph.

This EW system is small with merely a few "servers." Our results are summarized in Figure 9(a), which depicts the probability of waiting for four values of beta: 0.1, 0.5, 1.0, and 1.5; the four cases are clearly separable and become more stable as β increases. Figure 9(b) shows a comparison between the results of Erlang-R and Erlang-C for $\beta = 1.5$, which is the easiest case to stabilize since the number of physicians is the largest. We clearly observe the significant difference between the results of the two staffing procedures, where Erlang-R yields a much more stable performance. Table 3 completes the picture by presenting the residual mean square error (RMSE) and average percentage error (APE) for each β category and patient–physician combination. A smaller value of these measures indicates a more stable performance. We see that Erlang-R is superior across all β values and all physician types, but that the variability (when $\beta = 0.5$) is higher at the patient level than the aggregated one. This is mainly because some of the patient types have very small demand and therefore hit the staffing constraints more often than others. As β grows, this difference diminishes. (Supporting figures are omitted for lack of space.) We also observe that Erlang-R improves stability by 20%–350% (depending on β and patient-type), which could be very significant.

To conclude, despite the simplicity of the Erlang-R model, it does manage to capture the important aspects of patient visits in the EW, and hospital management can use it to calculate recommended staffing for physicians. The same outcome can be expected for nurse staffing. In fact, one would expect better results for nurse staffing since it gives rise to a higher number of servers, hence the MOL is likely to be more accurate.

Figure 9 Emergency Ward Case Study—$P(W_t > 0)$ for Various β Values

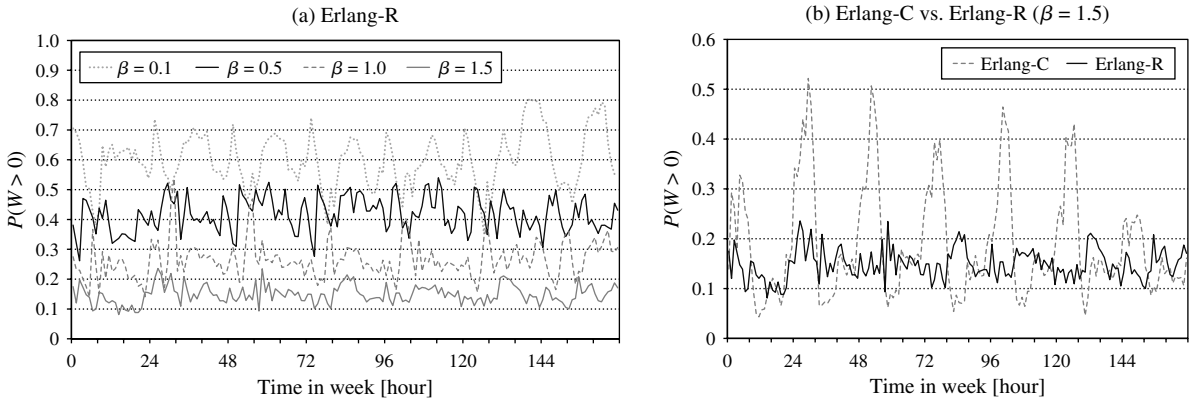


Table 3 Stability Comparison Between Erlang-R and Erlang-C Staffing in an Emergency Ward

(a) $P(W_t > 0)$ by β				(b) $P(W_t > 0)$ by physician type ($\beta = 0.5$)			
Model	β	RMSE	APE	Model	Physician type	RMSE	APE
Erlang-R	0.1	0.091	0.348	Erlang-R	1	0.105	0.217
	0.5	0.058	0.338		2	0.142	0.459
	1	0.061	0.410		3	0.109	0.259
	1.5	0.031	0.404		4	0.115	0.289
Erlang-C	0.1	0.113	0.397	Erlang-C	1	0.185	0.384
	0.5	0.131	0.499		2	0.139	0.480
	1	0.118	0.588		3	0.133	0.324
	1.5	0.111	0.688		4	0.162	0.436

Notes. RMSE $= \sqrt{(\sum_{t=1}^{n}(\alpha_s(t)-\alpha_e)^2)/n}$, APE $= (1/n)\sum_{t=1}^{n}|(\alpha_s(t)-\alpha_e)/\alpha_e|$, where $\alpha_s(t)$ is the simulated probability of waiting at time interval t and α_e is the stable theoretical value the system was designed to achieve. (Here the time interval is one hour, measured over a week, namely, $n = 167$.)

7. Fluid and Diffusion Models of the Number of Needy Customers, with Application to Mass Casualty Events

In this section, we develop fluid and diffusion limits for Erlang-R. We then use the resulting models/approximations to analyze an MCE, in which service demand fluctuates significantly and exceeds capacity, over a relatively short time period. Note that fluid models are naturally useful for analyzing time-varying systems, and they are also useful toward understanding the finite-horizon evolution of systems in a steady state. For example, one might seek to evaluate the probability that the number of customers (patients) in the system exceeds a certain threshold during a specific time horizon. This could support the design of alarm protocols such as when to commence special procedures: ambulance diversion or summoning additional medical staff. In designing such protocols, for example, toward avoiding excessive alarms, one would in fact require our diffusion refinements that determine confidence intervals around fluid sample paths; see Mandelbaum et al. (1999).

It was already noted that Erlang-R, both stationary and time varying, fits the mathematical framework of Markovian service networks in Mandelbaum et al. (1998). This framework justifies the existence and uniqueness of model solutions that accommodate time-varying arrivals and time-varying staffing policies. Specifically, Erlang-R is represented by $Q = \{Q(t), t \geq 0\}$, $Q(t) = (Q_1(t), Q_2(t))$: $Q_1(t)$ is the number of needy patients in the system at time t (i.e., those either waiting for service or being served), and $Q_2(t)$ is the number of content patients in the system. The process Q is characterized by the following sample-path equations, for $t \geq 0$:

$$\begin{aligned}
Q_1(t) &= Q_1(0) + A_1^a\left(\int_0^t \lambda_u\,du\right) - A_2^d\left(\int_0^t p\mu(Q_1(u)\wedge s_u)\,du\right)\\
&\quad - A_{12}\left(\int_0^t (1-p)\mu(Q_1(u)\wedge s_u)\,du\right)\\
&\quad + A_{21}\left(\int_0^t \delta Q_2(u)\,du\right),\\
Q_2(t) &= Q_2(0) + A_{12}\left(\int_0^t p\mu(Q_1(u)\wedge s_u)\,du\right)\\
&\quad - A_{21}\left(\int_0^t \delta Q_2(u)\,du\right),
\end{aligned}$$

where A_1^a, A_2^d, A_{12}, and A_{21} are four mutually independent time-homogeneous Poisson processes with rate 1. We now introduce a family of scaled queueing models, indexed by $\eta \nearrow \infty$, such that both the arrival rate and the number of physicians are scaled up by η, and the needy and content service rates remain unscaled:

$$\begin{aligned}
Q_1^\eta(t) &= Q_1^\eta(0) + A_1^a\left(\int_0^t \eta\lambda_u\,du\right)\\
&\quad - A_2^d\left(\int_0^t \eta p\mu\left(\frac{1}{\eta}Q_1^\eta(u)\wedge s_u\right)du\right)\\
&\quad - A_{12}\left(\int_0^t \eta(1-p)\mu\left(\frac{1}{\eta}Q_1^\eta(u)\wedge s_u\right)du\right)\\
&\quad + A_{21}\left(\int_0^t \eta\delta\left(\frac{1}{\eta}Q_2^\eta(u)\right)du\right),\\
Q_2^\eta(t) &= Q_2^\eta(0) + A_{12}\left(\int_0^t \eta p\mu\left(\frac{1}{\eta}Q_1^\eta(u)\wedge s_u\right)du\right)\\
&\quad - A_{21}\left(\int_0^t \eta\delta\left(\frac{1}{\eta}Q_2^\eta(u)\right)du\right).
\end{aligned} \tag{11}$$

Theorem 6. (*FSLLN*) *Through the scaling* (11), *we have*

$$\lim_{\eta\to\infty}\frac{Q^\eta(t)}{\eta} = Q^{(0)}(t), \quad t \geq 0,$$

where $Q^{(0)}(\cdot)$, *the fluid approximation/model, is the solution of the following ODE*:

$$\begin{aligned}
Q_1^{(0)}(t) &= Q_1^{(0)}(0)\\
&\quad + \int_0^t (\lambda_u - \mu(Q_1^{(0)}(u)\wedge s_u) + \delta Q_2^{(0)}(u))\,du,\\
Q_2^{(0)}(t) &= Q_2^{(0)}(0)\\
&\quad + \int_0^t (p\mu(Q_1^{(0)}(u)\wedge s_u) - \delta Q_2^{(0)}(u))\,du.
\end{aligned} \tag{12}$$

The convergence to $Q^{(0)}(\cdot)$ *is almost surely uniformly on compacts.*

The theorem follows from Theorem 2.2 in Mandelbaum et al. (1998). We continue by developing diffusion approximations for Erlang-R. These are used for

calculating variances and covariances, which, in turn, yield confidence intervals for the number of patients in the system.

THEOREM 7. (*FCLT*) *Through the scaling* (11) *and with the fluid limits* (12), *we have*

$$\lim_{\eta\to\infty}\sqrt{\eta}\left[\frac{Q^{\eta}(t)}{\eta}-Q^{(0)}(t)\right]\stackrel{d}{=}Q^{(1)}(t),\quad t\geq 0, \tag{13}$$

where $Q^{(1)}(\cdot)$, *the diffusion model/approximation, is the solution of a stochastic differential equation, as given by* (EC.9) *in the Internet supplement*, §EC.1.6. *The convergence to* $Q^{(1)}(\cdot)$ *is the standard Skorohod* J_1 *convergence in* $D[0,\infty)$.

The theorem is a consequence of Theorem 2.3 in Mandelbaum et al. (1998). Our fluid and diffusion models are easiest to apply when durations of critical loading are negligible (the zero-measure assumption in Mandelbaum et al. 2002). They are thus natural as models for MCEs, during which overloading constantly prevails. Formally, we have the following:

PROPOSITION 2. *Define* $\mathcal{S}$ *to be the set of times when the fluid number of physicians equals the number of patients in the needy state*: $\mathcal{S}=\{t>0 \mid Q_1^{(0)}(t)=s_t\}$. *Assume that this set of times* $\mathcal{S}$ *has measure zero. Then* (EC.9) *simplifies to*

$$\begin{aligned}
Q_1^{(1)}(t) &= Q_1^{(1)}(0)\\
&\quad+\int_0^t(-\mu 1_{\{Q_1^{(0)}(u)\leq s_u\}}Q_1^{(1)}(u)+\delta Q_2^{(1)}(u))\,du\\
&\quad+B_1^a\left(\int_0^t\lambda_u\,du\right)\\
&\quad-B_2^d\left(\int_0^t p\mu(Q_1^{(0)}(u)\wedge s_u)\,du\right)\\
&\quad-B_{12}\left(\int_0^t(1-p)\mu(Q_1^{(0)}(u)\wedge s_u)\,du\right)\\
&\quad+B_{21}\left(\int_0^t\delta Q_2^{(0)}(u)\,du\right),\\
Q_2^{(1)}(t) &= Q_2^{(1)}(0)\\
&\quad+\int_0^t(p\mu 1_{\{Q_1^{(0)}(u)\leq s_u\}}Q_1^{(1)}(u)-\delta Q_2^{(1)}(u))\,du\\
&\quad+B_{12}\left(\int_0^t p\mu(Q_1^{(0)}(u)\wedge s_u)\,du\right)\\
&\quad-B_{21}\left(\int_0^t\delta Q_2^{(0)}(u)\,du\right).
\end{aligned}\tag{14}$$

The mean vector for the diffusion approximation (EC.10) *is then*

$$\begin{aligned}
\frac{d}{dt}E[Q_1^{(1)}(t)] &= -\mu 1_{\{Q_1^{(0)}(t)\leq s_t\}}E[Q_1^{(1)}(t)]+\delta E[Q_2^{(1)}(t)],\\
\frac{d}{dt}E[Q_2^{(1)}(t)] &= p\mu 1_{\{Q_1^{(0)}(t)\leq s_t\}}E[Q_1^{(1)}(t)]-\delta E[Q_2^{(1)}(t)];
\end{aligned}$$

and the covariance matrix (EC.11) *is*

$$\begin{aligned}
&\frac{d}{dt}\mathrm{Var}[Q_1^{(1)}(t)]\\
&\quad=-2\mu 1_{\{Q_1^{(0)}(t)\leq s_t\}}\mathrm{Var}[Q_1^{(1)}(t)]+2\delta\mathrm{Cov}[Q_1^{(1)}(t),Q_2^{(1)}(t)]\\
&\qquad+\lambda_t+\delta Q_2^{(0)}(t)+\mu(Q_1^{(0)}(t)\wedge s_t),\\
&\frac{d}{dt}\mathrm{Var}[Q_2^{(1)}(t)]\\
&\quad=-2\delta\mathrm{Var}[Q_2^{(1)}(t)]+2p\mu\mathrm{Cov}[Q_1^{(1)}(t),Q_2^{(1)}(t)]\\
&\qquad+p\mu(Q_1^{(0)}(t)\wedge s_t)+\delta Q_2^{(0)}(t),\\
&\frac{d}{dt}\mathrm{Cov}[Q_1^{(1)}(t),Q_2^{(1)}(t)]\\
&\quad=-(\mu 1_{\{Q_1^{(0)}(t)\leq s_t\}}+\delta)\mathrm{Cov}[Q_1^{(1)}(t),Q_2^{(1)}(t)]\\
&\qquad+\delta\mathrm{Var}[Q_2^{(1)}(t)]+p\mu 1_{\{Q_1^{(0)}(t)\leq s_t\}}\mathrm{Var}[Q_1^{(1)}(t)]\\
&\qquad-p\mu(Q_1^{(0)}(t)\wedge s_t)-\delta Q_2^{(0)}(t).
\end{aligned}\tag{15}$$

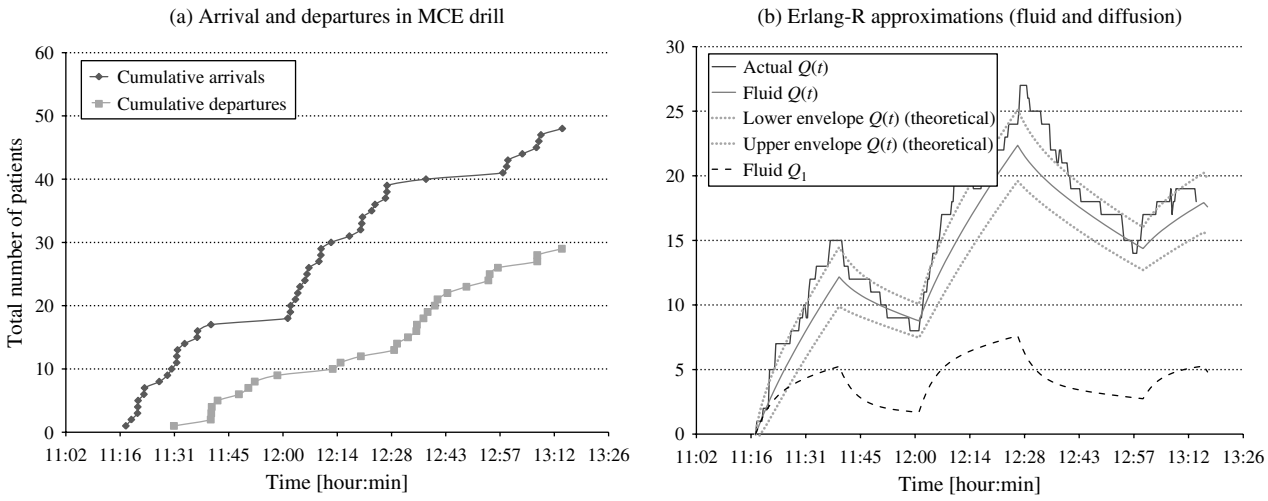
Proposition 2 supports MCE modeling and management, which we turn to next.

7.1. Mass Casualty Events

When an MCE is in progress, the EW must, over a short time period, attend to already admitted patients, release those who can be released, and most importantly, provide emergency care to new arrivals at overcapacity rates. We now demonstrate that our transient fluid and diffusion models, from the previous subsection, usefully capture the state of an EW during an MCE. This enables one to use Erlang-R for off-line *planning* of an MCE, *initial reaction* at its outset (customized to the MCE type, severity and scale), and subsequently online MCE *control* until the event winds up. We focus as before on staffing. To this end, we use data from a chemical MCE drill. The MCE took place in July 2010 at 11:00 A.M. and lasted till 13:15; its casualties were transported to an Israeli hospital where our data were collected. The short horizon of MCEs (here two hours) and the protocol of chemical events (periodic treatment of patients) renders the transient Erlang-R, with its recurrent service structure, naturally appropriate.

Our data is for the severely wounded nontrauma patients. Figure 10(a) depicts cumulative arrival and departure counts, collected roughly during 11:15–13:15. The arrival rate is clearly time varying: periods with no arrivals alternate with approximately constant arrival rates, with the rates decreasing as time progresses. (Our hospital partners, experienced in managing MCEs, inform us that this piecewise-constant pattern of arrival rate is typical of MCEs: it is attributed to the fact that casualties are transported from the MCE scene by a finite number of

Figure 10 Chemical Mass Casualty Event Drill: Arrivals, Departures, and Erlang-R Approximations



ambulances, who traverse back and forth.) The estimated arrival rate function (customers per minute) is as follows ($1_{[a,b)}$ is an indicator function):

$$\lambda_t = 0.773 \times 1_{[0,22)}(t) + 0.884 \times 1_{[44,69)}(t) + 0.5 \times 1_{[102,117)}(t), \quad 0 \le t \le 120. \tag{16}$$

Erlang-R parameters were estimated from medical specifications and the physics of Erlang-R, as we now explain. The severity level of the patients under consideration calls for medication every 30 minutes, in addition to treating their injuries. Staffing specs assigned every physician to four patients at a time. (In reality, and being a drill, there were ample physicians on site, which implies, no upper bound on the number of physicians ($s = \infty$). Such resource levels are unlikely to prevail in true-to-life MCEs, but they facilitate the estimation of parameter values—which are practice relevant.) One can now estimate μ, p, δ via the following three equations:

$$1/\mu + 1/\delta = 30; \quad 1/\mu + 30p/(1-p) = 62.4; \quad \mu/\delta = 3/p. \tag{17}$$

The first equation corresponds to the 30-minute cycle. The second represents length of stay (LOS) as the first service followed by a geometric number of cycles; the average LOS of 62.4 minutes is then the classical Kaplan–Meier estimator (Kaplan and Meier 1958) for censored data: indeed, patients that were still in treatment when the drill ended (about 20 out of 50) provided only *lower bounds* on their LOS. The last equation arises from the patients-to-physician ratio $(R_1 + R_2)/4 = R_1$, in which R_1, R_2 are the steady-state offered-loads from §3. Solving the equations in (17) yields average treatment time of 5.4 minutes ($\mu = 11.06$), average content time 24.6 minutes ($\delta = 2.44$) and $p = 0.662$.

We now compare, in Figure 10(b), Erlang-R estimators against MCE data. First we have fluid-based estimators for $Q = Q_1 + Q_2$, the total number of casualties, enveloped by a diffusion-based 95% confidence band. This is to be compared against the actual sample path, observed from our MCE data (the difference between cumulative arrivals and departures). Erlang-R clearly captures well the transient nature of the MCE: the data is essentially within its confidence band. Notably, a comparison (omitted for space constraints) of Erlang-R with Erlang-C demonstrated that the latter yields noticeably inferior path-estimators: an increase of about 45% in RMSE and APE measures, for the reasons that were explained in §4.3.2.

After validating Erlang-R against the observed Q, one can now trust it to infer the number of busy physicians—see the dashed function Q_1 in Figure 10(b). Its evolution was unobservable at the MCE drill, which is a state of affairs that is to be commonly expected. Yet Q_1 is essential for planning and control of MCEs, as discussed next.

7.1.1. Erlang-R in Support of MCE Staffing. Since Erlang-R reliably captures MCE dynamics, one can use it to support planning for an MCE, initial reaction to its severity and scale and, ultimately, controlling MCE evolution. For concreteness we consider staffing upon initial reaction. The procedure would be similar in planning, when applying Erlang-R for comparative analysis of plausible scenarios, and control, where parameter values are updated adaptively and then fed into Erlang-R over a rolling horizon. All these applications entail the following steps:

1. *Forecasting the arrival rate function* λ_t (e.g., (16)) for each severity group of patients. Any forecasting model should take into account the estimated number of casualties routed to the hospital, number of

ambulances available, and distance from the hospital (Jacobson et al. 2012).

2. *Estimating the offered-load* $R(\cdot)$ for each severity group, taking into account group-specific treatment protocols as demonstrated above.

3. *Calculating the staffing function* $s(\cdot)$ via $s(t) = [R(t) + \beta\sqrt{R(t)}]$, $t \geq 0$. We recommend a relatively high β, say $\beta \geq 2$, to account for the emergency situation at hand. One should then accommodate constraints such as the available number of physicians within the hospital and the availability and time-to-arrive of out-of-hospital physicians.

4. *Predicting EW evolution* via Erlang-R under the planned SRS.

Given our RFID-based data in Figure 10, we now demonstrate the above steps by planning for staffing an MCE. Being able to infer Q_1 (Figure 10) yields insights that exploit its special structure of three phases: a first surge of arrivals (11:00–12:00), peak period (12:00–13:00), and a closure phase from 13:00 till completion; each phase starts with an increase of load, which is immediately followed by a decrease because of ambulances returning to the MCE scene. As will be demonstrated, this allows one to initially divert physicians within the hospital to cater to the first surge while, in parallel, summon off-duty staff who would join (say from home) toward the second peak surge. Staffing remains constant within a phase, which gives rise to the following plan:

1. *Initial reaction*: Recall that the MCE occurred at 11:00. The first casualties arrived to the hospital at 11:15, thus starting a surge of demand (offered-load) that peaks at 11:40: $Q_1 = 5$. By SRS, this calls for $5 + 2\sqrt{5} \approx 9$ physicians, which are to arrive, conceivably from the hospital itself, until 11:15.

2. *Peak period*: From 12:07, demand for physicians increases to a peak $Q_1 = 7.5$ at 12:25. One needs now $7.5 + 2\sqrt{7.5} \approx 13$ physicians, or an additional group of four physicians that can join within hour from MCE start.

3. *Closure*: This last phase starts around 13:00, and arrivals cease at 13:15. A real MCE would continue at the hospital till all casualties are hospitalized, while gradually releasing physicians to their routine or reassigning them to help with already-hospitalized casualties. Similarly to the above (not pursued here), one can again use Erlang-R to plan for the release of physicians, which, interestingly, involves also the prediction of the MCE completion time.

As mentioned, Chemical MCEs naturally fit the recurrent service structure of Erlang-R. Other types of MCEs might need other models. For example, with relatively more trauma patients and during off-peak arrivals, physicians who perform initial lifesaving procedures could also accompany their patients through surgery. A corresponding model would then consist of two queues in tandem, as analyzed by Cohen et al. (2013).

8. Conclusions and Further Research

Motivated by staffing applications in healthcare, we have developed a simple-yet-not-too-simple service model, Erlang-R, which accommodates returning customers in a time-varying environment. The model valuably captures both normal operating conditions and MCEs. In the former, it gives rise to an explicit staffing recipe that matches service capacity with time-varying demand (the QED operational regime), which in turn stabilizes operational performance (service level, utilization). In MCEs the model can support planning for initial reaction to and control of such events.

We started, in the introduction, with four examples of returning customers/patients in healthcare systems. We can now conclude, based on the analysis in §§3, 4.3.2, and 7.1 and some additional hospital data, that Erlang-R better be used for modeling EWs (both in normal and MCE conditions) whereas, for oncology and radiology wards, Erlang-C suffices. To elaborate, for the EW under its normal conditions, the parameters $\omega = 0.2618$ (as $f = 24$, in hours) and $\sqrt{\mu\delta(1-p)} \approx 3.4$ are such that the EW fits the left part of Figure 3 (in both plots). The amplitude ratio is within $(0.93, 0.97)$ and the phase ratio is within $(1.7, 3)$, depending on patient type (see Table 2); hence, the significant difference is between phases rather than amplitudes, which means that using Erlang-C will be mostly wrong in timing—starting (and ending) shifts too soon. In the oncology ward, the corresponding values are $\omega = 6.283$ ($f = 1$, in days) and $\sqrt{\mu\delta(1-p)} = 0.0495$. This puts oncology on the right side of Figure 3, where we expect little if any difference between the two models. Indeed, the amplitude and phase ratios are 0.9987 and 0.9756, respectively, namely, very close to unity. Next, radiology operates in a steady-state environment, since the arrival rate is constant, and thus need not use Erlang-R. Finally, our last example, EW under MCE stress, must be modeled as Erlang-R since, in transient times (over a short time horizon), the difference between Erlang-R and Erlang-C is significant.

It is important to emphasize that, even in the case when Erlang-C suffices to capture overall performance, Erlang-R would still be preferable over a finite horizon, or for focusing on the performance of needy (content) patients. Erlang-R is also capable of capturing usefully, as in §6, the operational performance of a *full-scale* EW, from the point of view of its physicians: the model plainly aggregates the "world beyond physicians" into a single ample-server station. One could do the same with EW nurses.

One could also raise the more general question of approximating a general queueing network, from the point of a specific node, by an Erlang-R model (the specific node would be needy while the rest of the network is content)—when do such crude approximations work and, alternatively, when are their refinements necessary?

The healthcare environment suggests further extensions for Erlang-R. To name a few, Yom-Tov (2010) adds an upper bound on the overall number of customers in the system, which corresponds to finite bed capacity; Chan et al. (2014) consider state-dependent service times; Huang et al. (2012) trade off high priority to patients on their first visit versus, alternatively, to those who have been in the system for a long time; and, finally, customer abandonment can take place during a first waiting (left without being seen) or between services (left against medical advice). We conclude with an outstanding open theoretical problem, which is the analysis of the limiting time-varying diffusion process under SRS. This is a prerequisite for understanding the success of our time-varying MOL staffing.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2013.0474.

Acknowledgments

The authors thank Yariv Marmor for providing them access to his EW simulator. The work of Avishai Mandelbaum was partially supported by the United States–Israel Binational Science Foundation [Grants 2005175, 2008480]; the Israel Science Foundation [Grant 1357/08]; and by the Technion funds for promotion of research and sponsored research. Some of the research was funded by and carried out while Avishai Mandelbaum was visiting the Statistics and Applied Mathematical Sciences Institute of the National Science Foundation; the Department of Statistics and Operations Research, the University of North Carolina at Chapel Hill; the Department of Information, Operations and Management Sciences, Leonard N. Stern School of Business, New York University; and the Department of Statistics, the Wharton School, University of Pennsylvania—the wonderful hospitality of these institutions is gratefully acknowledged and truly appreciated. The work of Galit Yom-Tov was partially supported by the Israel National Institute for Health Policy and Health Services Research.

References

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2011) Patient flow in hospitals: A data-based queueing-science perspective. Working paper, Technion–Israel Institute of Technology, Technion City, Haifa. http://iew3.technion.ac.il/serveng/References/references.

Borst S, Mandelbaum A, Reiman M (2004) Dimensioning large call centers. *Oper. Res.* 52(1):17–34.

Chan CW, Yom-Tov GB, Escobar G (2014) When to use speedup: An examination of service systems with returns. *Oper. Res.* Forthcoming.

Cohen I, Mandelbaum A, Zychlinski N (2013) Minimizing mortality in a mass casualty event: Fluid networks in support of modeling and staffing. *IIE Trans.*, ePub ahead of print October 30, http://dx.doi.org/10.1080/0740817X.2013.855846.

Eick SG, Massey WA, Whitt W (1993a) $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* 39(2):241–252.

Eick SG, Massey WA, Whitt W (1993b) The physics of the $M_t/G/\infty$ queue. *Oper. Res.* 41(4):731–742.

Falin GI, Templeton JGC (1997) *Retrial Queues* (Chapman & Hall, London).

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

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: A tutorial and literature review. *Manufacturing Service Oper. Management* 5(2):79–141.

Green L, Kolesar PJ, Soares J (2001) Improving the SIPP approach for staffing service systems that have cyclic demands. *Oper. Res.* 49(4):549–564.

Green L, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* 16(1):13–39.

Green L, Soares J, Giglio JF, Green RA (2006) Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* 13(1):61–68.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29:567–587.

Huang J, Carmeli B, Mandelbaum A (2012) Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. Working paper, Technion–Israel Institute of Technology, Technion City, Haifa.

Israel Ministry of Health (2006) Financial report for years 2000–2005. Report, Israel Ministry of Health, Jerusalem. http://www.health.gov.il/publicationsFiles/finance2005.pdf.

Jacobson EU, Argon NT, Ziya S (2012) Priority assignment in emergency response. *Oper. Res.* 60(4):813–832.

Janssen AJEM, van Leeuwaarden JSH, Zwart B (2011) Refining square-root safety staffing by expanding Erlang C. *Oper. Res.* 59(6):1512–1522.

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

Jennings OB, de Véricourt F (2011) Nurse staffing in medical units: A queueing perspective. *Oper. Res.* 59(6):1320–1331.

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53(282):457–481.

Khudyakov P, Gorfine M, Feigin P (2010) Test for equality of baseline hazard functions for correlated survival data using frailty models. Working paper, Technion–Israel Institute of Technology, Technion City, Haifa.

Lahiri A, Seidmann A (2009) Analyzing the differential impact of radiology information systems across radiology modalities. *J. Amer. College of Radiology* 6(10):522–526.

Maman S (2009) Uncertainty in the demand for service: The case of call centers and emergency departments. Master's thesis, Technion–Israel Institute of Technology, Technion City, Haifa.

Mandelbaum A, Massey W, Reiman M (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30(1–2):149–201.

Mandelbaum A, Massey WA, Reiman M, Rider B (1999) Time varying multiserver queues with abandonment and retrials. Key P, Smith D, eds. *ITC-16, Teletraffic Engineering in a Competitive World* (Elsevier), 355–364.

Mandelbaum A, Massey WA, Reiman M, Stolyar A, Rider B (2002) Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecomm. Systems* 21(2–4):149–171.

Marmor Y, Sinreich DA (2005) Emergency department operations: The basis for developing a simulation tool. *IIE Trans.* 37(3): 233–245.

Massey WA, Whitt W (1993) Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* 13(1): 183–250.

Massey W, Whitt W (1994) An analysis of the modified offered-load approximation for the nonstationary Erlang loss model. *Ann. Appl. Probab.* 4(4):1145–1160.

Whitt W (2007) What you should know about queueing models to set staffing requirements in service systems. *Naval Res. Logist.* 55(5):476–484.

Whitt W (2013) Offered load analysis for staffing. *Manufacturing Service Oper. Management* 15(2):166–169.

Yom-Tov G (2010) Queues in hospitals: Queueing networks with reentering customers in the QED regime. Ph.D. thesis, Technion–Israel Institute of Technology, Technion City, Haifa.

Zeltyn S, Marmor YN, Greenshpan O, Mesika Y, Wasserkrug S, et al. (2011) Simulation-based models of emergency departments: Operational, tactical, and strategic staffing. *ACM Trans. Modeling Comput. Simulation (TOMACS)* 21(4):Article 24.

Zhan D, Ward AR (2014) Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing Service Oper. Management* 16(2).