Submitted to *Manufacturing & Service Operations Management*
manuscript (Please, provide the mansucript number!)

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

(Authors' names blinded for peer review)

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 in order to achieve predetermined service levels. Formally, it is merely a 2-station open queueing network which, in 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, in order 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 (QED) and the MCE is overloaded (ED). In particular, for the EW we propose a time-varying square-root staffing (SRS) policy, based on the modified-offered-load, which is proved to perform well over small-to-large systems.

Key words: Healthcare; Queueing Networks; Modified Offered-Load; Time Varying Queues; Halfin-Whitt Regime; QED Regime; ED Regime; Emergency Department Staffing; Mass Casualty Events; Patient Flow

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, as it is not provided at a single event. This has motivated our queueing model, (the time-varying) Erlang-R ("R" for Reentrant customers or Repetative service) which accommodates the Return-to-service phenomena.

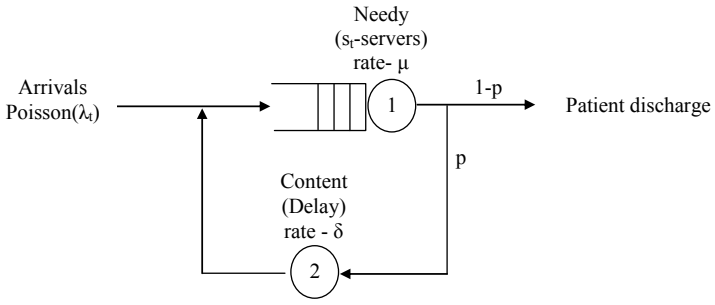
More explicitly, we consider a model where customers seek service from servers. After service is completed, with probability $1-p$ they exit the system and with probability p they return for further service, after a random delay time. We refer to the service phase as a *Needy* state, and to the delay 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 $\frac{1}{\mu}$, and Content times are i.i.d. with general distribution G_2 and mean $\frac{1}{\delta}$. 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 $\lambda_t,\ t \geq 0$; 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 separately normal and stressful EW conditions.

Figure 1 The Erlang-R Queueing Model.



For the first, the process starts by admitting patients and referring them to an EW physician. The physician examines them in order to decide between discharge vs. 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 Section 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 (Zychlinski et al. 2012). Based on data from an MCE drill, as described in Section 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 4 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 Section 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 due to impatience, while 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 *re-entrance*, 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 for example Green et al. (2006).) Our important insight (§3&4) is that returns become significant in time-varying systems (they are not so in 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 re-entering 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 (§A.5).

- 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 Section 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&4).

- Analytical approximations for the Queue-Length and Number-of-Busy-Servers processes. These are derived separately for systems that are super-critical (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 Internet Supplement §C, 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 4 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 (§C) 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 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 1 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) while the rest of the network is aggregated into the "Content" station (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 (IMH 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 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. using the Lag-SIPP (Stationary Independent Period by Period) approach and Zeltyn et al. using ISA (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 in order 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 $R = \lambda \cdot \mathrm{E}[S] = \frac{\lambda}{\mu}$, and β is a Quality-of-Service parameter that is set to accommodate service-level constraints. Data from Zeltyn et al. (2011) suggests 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 steady-state quickly. The second approach includes the MOL in Jennings et al. (1996) or ISA 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 $(M_t/M/s_t)$, $R(t) = E\left[\lambda(t - S_e)\right] E[S]$ (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:

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

where

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

$\nu(i)$ is defined as $\nu(i) := (i \wedge s)s^{(i-s)^+}$, and $R_1 = \frac{\lambda}{(1-p)\mu}$, $R_2 = \frac{p\lambda}{(1-p)\delta}$. 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*

$$\alpha := P(W>0) = \left[\frac{(R_1)^s}{s!(1-R_1/s)}\right] c_1,$$

$$E[W|W>0] = \frac{1}{\mu s(1-\rho)},$$

$$W|W>0 \stackrel{d}{=} \exp(E[W|W>0]),$$

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

Proof: Theorem 1 is a straightforward result of Erlang-R being a 2-node Jackson network, jointly with the arrival theorem for Open Jackson networks.

In 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 $(\lambda, \frac{(1-p)\delta}{p})$. It follows that, in 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 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 $(\frac{\lambda}{1-p}, \mu)$, 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 *multi-service* 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 non-stationary 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 ODE (Ordinary Differential Equation): 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: Internet Supplement, Section A.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 Sections 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:*

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

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 Internet Supplement, Section A.1.

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

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

Proof: Internet Supplement, Section A.1.

Approximation (5) reveals a fundamental difference between the offered-loads of Erlang-R and its corresponding Erlang-C. The multi-service Erlang-C second-order approximation is $R(t) \approx \frac{E[S_1]}{1-p}\left[\lambda(t - E\left[S_{1,e}^{*J}\right]) + \frac{1}{2}\lambda^{(2)}(t) Var[S_{1,e}^{*J}]\right]$. 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 (Section 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 multi-service 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 (Section 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, $\omega = \frac{2\pi}{f}$ 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

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

We now provide explicit solutions for $R(\cdot)$ in the case of exponential service times. (Deterministic service times are also amenable to analysis; then the amplitude and phase behavior of $R(\cdot)$ is also interesting, but less realistic and, therefore, is only hinted at in Internet Supplement, Section A.5.)

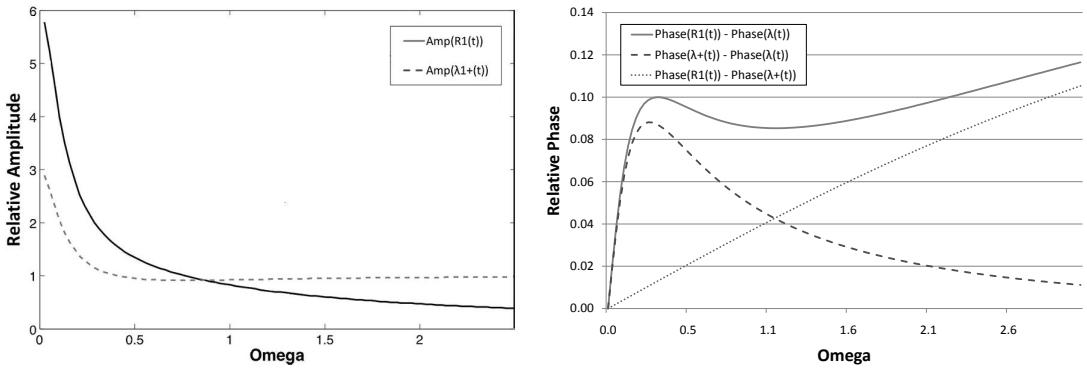
4.3.1. Exponential Service Times

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:*

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

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

Figure 2 **The 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 Internet Supplement, Section A.2.

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

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

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 8 (Internet Supplement, Section A.2). Theorem 4 yields a simple relation between the amplitudes of $R(\cdot)$ and $\lambda_1^+(\cdot)$: $Amp(R_1) = Amp(\lambda_1^+)\sqrt{\mu^2 + \omega^2}$, 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 Internet Supplement, Section A.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 multi-service 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 $Amp(R_c) = \frac{\bar{\lambda}\kappa}{\sqrt{\mu_c^2+\omega^2}}$, and its phase is $\theta_c = \frac{1}{2\pi} cot^{-1}(\mu_c/\omega)$ (Eick et al. 1993a). The ratio between the amplitudes and phases are thus given by

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

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

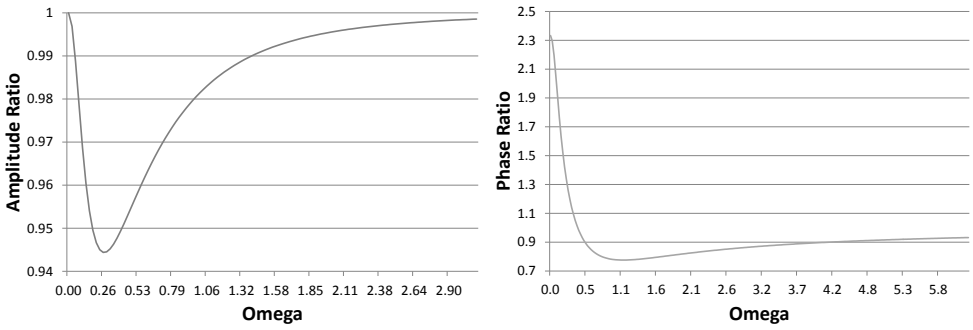
THEOREM 5. *Assume that the arrival rate is sinusoidal and service times are exponential. Comparing Erlang-R with parameters $(\lambda, \mu, p, \delta)$ against the (multi-service) Erlang-C model with parameters $(\lambda, (1-p)\mu)$:*

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

Proof: All results follow from analyzing Equations (10); see Internet Supplement, Section A.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 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 over- or under- staff. The impact of this observation on the service level is further explored in Section 5; it shows that one must take into account the repetitive nature of service, in order to avoid excessive staffing costs or undesirable service levels.

Figure 3 **Ratio of amplitudes & phases between Erlang-R and Erlang-C as a function of ω (Case Study 1, §5.1).**



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 1 up to $\omega = \sqrt{\frac{2\delta^2+p(1-p)\delta\mu}{p}}$, and from that point on it is smaller than 1. 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 1, but the phase ratio need not. (All the above observations will be used, in Section 8, to analyze the significance of Erlang-R in the healthcare examples of Section 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 multi-service 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.

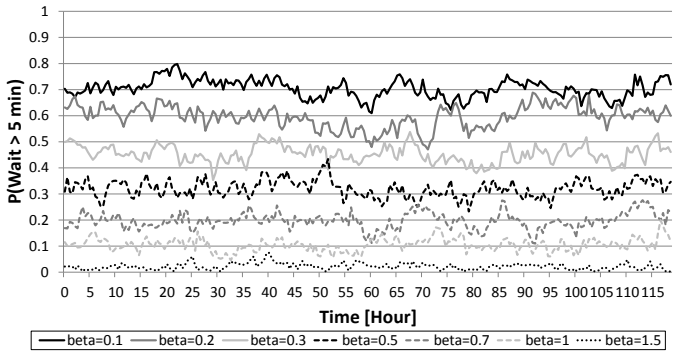
5.1. Case Study 1 - Large System

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 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.

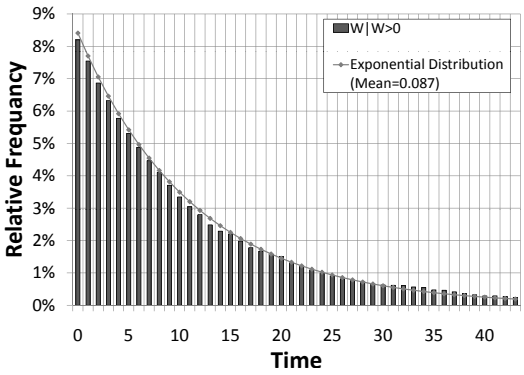
We find that for a large enough system in the QED regime ($\beta > 0.3$), the MOL approach stabilizes all performance measures of the Erlang-R queueing network. Consequently, *any* pre-specified QED service level can be achieved *stably over time*. For example, Figure 4a shows the empirical $P(W_t >$

Figure 4 **Case study 1 - Simulation results of $P(W_t > T)$ for various β values and $W|W>0$ in large systems.**

(a) $P(W_t > T)$, $T = 5$ minutes



(b) $W|W>0$



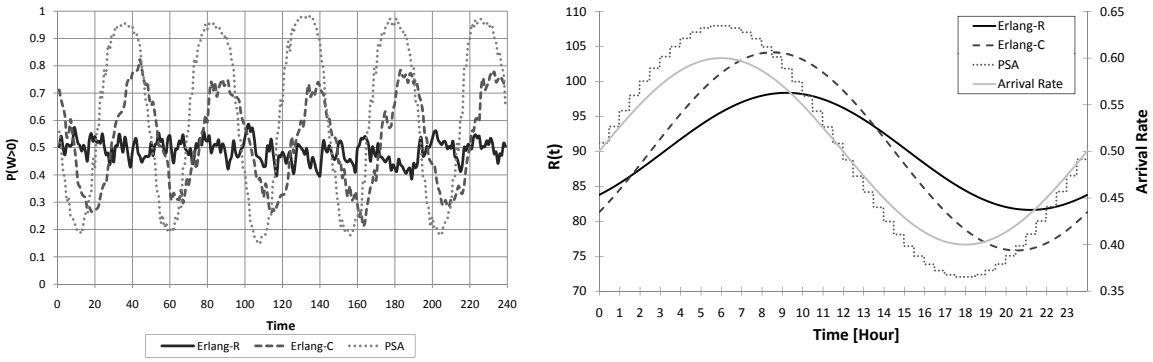
T), the fraction of Needy arrivals at time t, who are delayed in queue more than T units of time. This fraction was calculated over a 5-day period, for various values of β. We note that $P(W_t > T)$ is relatively stable for all β tested. Figure 4b shows the conditional distribution of the waiting time given delay $(W|W>0)$, when $\beta = 0.5$. (It is calculated over all arrivals during the 5-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 Appendix B. 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: While 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 $E[Q_1(t)] = R_1(t) + \alpha\frac{R_1(t)}{s(t)}\left(1 - \frac{R_1(t)}{s(t)}\right)^{-1}$; see Internet Supplement Section C, 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

Figure 5 Case study 1 - Comparing Erlang-R, Erlang-C, and PSA.

(a) $P(W_t > 0)$ (b) $R(t)$



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 5a shows that, while using Erlang-R stabilizes system performance around the pre-specified 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 5b). We observe that for one half of the cycle, Erlang-C over-estimates $R(\cdot)$, resulting in over-staffing which, in turn, results in a better performance than the pre-specified target. However, in the other half cycle, the opposite occurs, causing the performance to be worse than pre-specified. Erlang-R, in contrast, stabilizes performance over the whole time horizon. (These observations also follow from our theoretical analysis in Section 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.

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"

Figure 6 Case study 2 - Plot of arrival rates in an emergency ward.

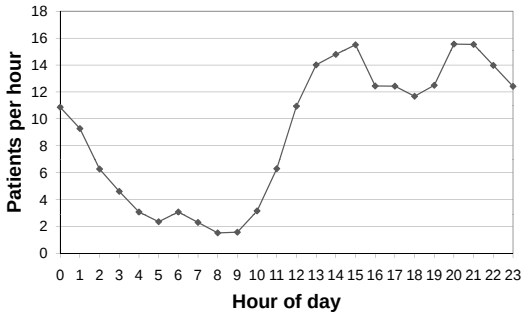


Table 1 Small Systems: An example of a discrete range for $P(W>0)$, as a function of β.

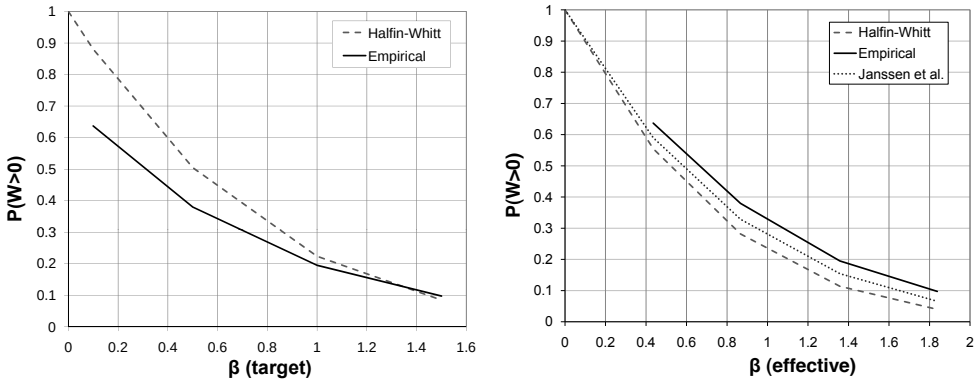
target β range	*effective* β	s	$P(W>0)$
$(0.474, 1.055]$	1.055	4	34.0%
$(1.055, 1.658]$	1.658	5	11.4%
$(1.658, 2.261]$	2.261	6	3.0%
1.658 and up	∞	7	0%

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

need of say 1.5 servers to 2 servers means adding 30% excess capacity to the required capacity, which suggests difficulties in stabilizing performance around pre-specified 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 1 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. (Due to space limitations, we have not included supporting graphs; furthermore, Figure 9a in Section 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

Figure 7 **Case study 2 - Comparison of the Halfin-Whitt and Janssen et al. formulae to simulation.**



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 downwards, 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 upwards. 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, based on the left diagram of Figure 7. More research is also needed to understand the Halfin-Whitt (and Janssen et al.) 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;

Table 2 EW 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

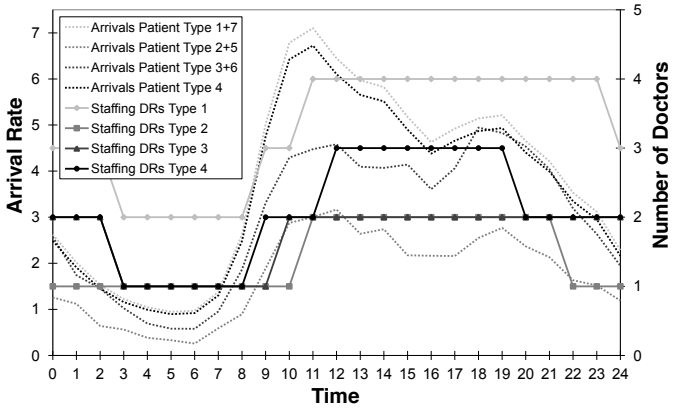
for a complete description see Marmor and Sinreich (2005). In our experiment, we use their accurate and detailed EW simulation model (it takes into account even walking distances), which is flexible in that it is easily adapted to a given EW. We fit the simulator to the EW of our partner Israeli hospital (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 7 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: $E[S_1] = \frac{1}{\mu}$ is estimated by averaging all services given by a specific physician group.
- Content times: $E[S_2] = \frac{1}{\delta}$ 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 $\frac{1}{1-p}$ 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

Figure 8 EW case study - Patient Arrivals and Physician Staffing for Each Physician Type in EW simulation ($\beta = 0.5$).



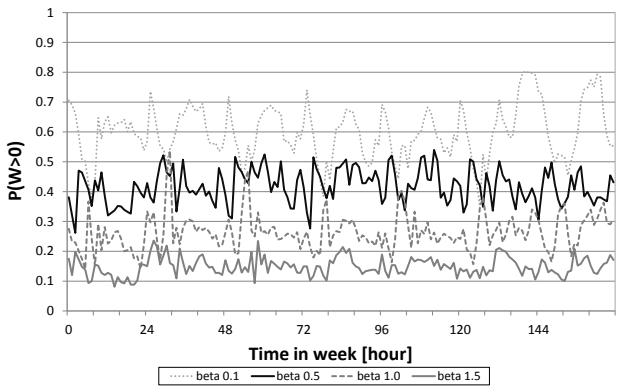
in a one-hour resolution. For each interval, we calculated the average number of physicians needed and rounded up to the nearest integer. We used one replication of one hundred weeks. (The first setup week was excluded.)

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

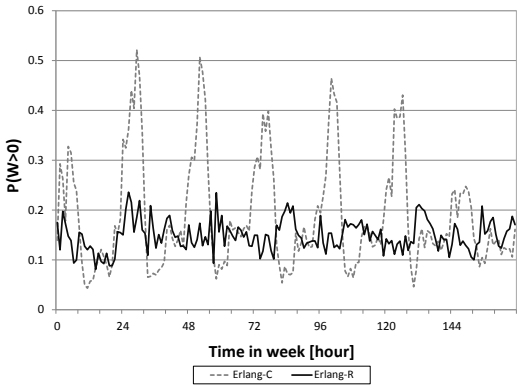
This EW system is small with merely a few "servers". Our results are summarized in Figure 9a, which depicts the probability of waiting for four values of beta: 0.1, 0.5, 1.0, and 1.5; the four cases are clearly separable and become more stable as β increases. Figure 9b shows a comparison between the 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

Figure 9 EW case study - $P(W_t > 0)$ for various β values.

(a) Erlang-R



(b) Erlang-C vs. Erlang-R ($\beta = 1.5$)



level than the aggregated one. This is mainly due to the fact that 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.

Table 3 Stability comparison between Erlang-R and Erlang-C staffing in EW.

(a) $P(W_t > 0)$ by β

Model	β	RMSE	APE
Erlang-R	0.1	0.091	0.348
	0.5	0.058	0.338
	1	0.061	0.410
	1.5	0.031	0.404
Erlang-C	0.1	0.113	0.397
	0.5	0.131	0.499
	1	0.118	0.588
	1.5	0.111	0.688

(b) $P(W_t > 0)$ by physician type ($\beta = 0.5$)

Model	Physician Type	RMSE	APE
Erlang-R	1	0.105	0.217
	2	0.142	0.459
	3	0.109	0.259
	4	0.115	0.289
Erlang-C	1	0.185	0.384
	2	0.139	0.480
	3	0.133	0.324
	4	0.162	0.436

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

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.

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 while fluid models are naturally useful for analyzing time-varying systems, they are also useful towards understanding the finite-horizon evolution of systems in 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 towards 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 $t\geq 0$:

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

where A_1^a, A_2^d, A_{12} and A_{21} are 4 mutually independent time-homogenous 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 while the Needy and Content service rates remain unscaled:

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

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:*

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

The convergence to $Q^{(0)}(\cdot)$ *is a.s. uniformly on compacts (u.o.c).*

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 an SDE (Stochastic Differential Equation), as given by* (26) *in the Internet Supplement, Section A.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:

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: $\mathcal{S}=\{t>0|Q_1^{(0)}(t)=s_t\}$. Assume that this set of times $\mathcal{S}$ has measure zero. Then (26) simplifies to*

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

The mean vector for the diffusion approximation (27) is then:

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

and the covariance matrix (28) is

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

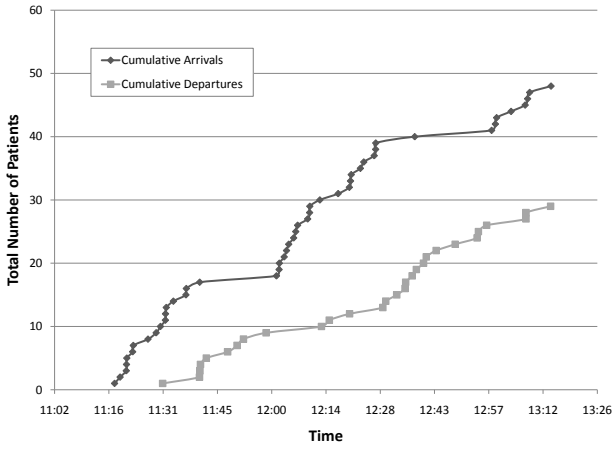
Proposition 2 supports MCE modeling and management, which we turn to next.

7.1. Mass-Casualty Events

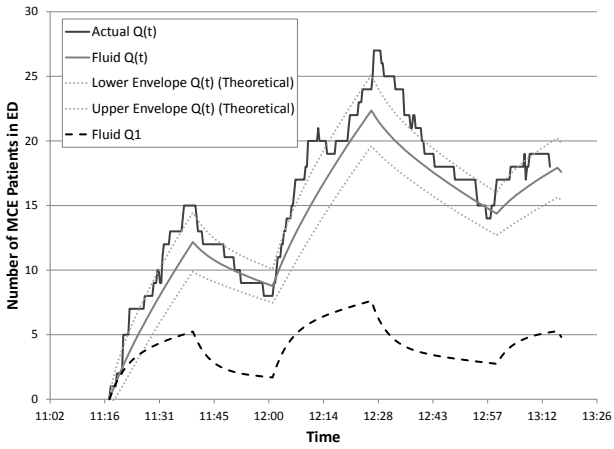
When an MCE is in progress, the EW must, over a short time period, attend to already admitted patients, release those who can be released and, most importantly, provide emergency care to new arrivals at over-capacity rates. We now demonstrate that our transient fluid and diffusion models, from the previous subsection, usefully capture the state of an EW during an MCE. This enables

Figure 10 Chemical MCE Drill: Arrivals, Departures, and Erlang-R Approximations.

(a) Arrival and Departures in MCE Drill



(b) Erlang-R Approximations (fluid and diffusion)



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 2 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 non-trauma patients. Figure 10a 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 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 4 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 3 equations:

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

The first equation corresponds to the 30-minute cycle. The second represents LOS as the first service followed by a geometric number of cycles; the average LOS of 64.2 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 Section 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 10b, 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 Section 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 10b. 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 $s(t) = \left[R(t) + \beta\sqrt{R(t)}\right]$, $t \geq 0$. 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 3 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, that is immediately followed by a decrease due to 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) towards 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 4 physicians that can join within 1 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 life-saving procedures could also accompany their patients through surgery. A corresponding model would then consist of 2 queues in tandem, as analyzed by Zychlinski et al. (2012).

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 Sections 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) while, 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 be still 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 Section 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 queuing 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. (2012) consider state-dependent service times; Huang et al.

(2012) trade-off high priority to patients on their first visit vs., 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.

References

Armony, M., S. Israelit, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, G.B. Yom-Tov. 2011. Patient flow in hospitals: A data-based queueing-science perspective. Working Paper, Technion. Available at http://iew3.technion.ac.il/serveng/References/references. 6

Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Operations Research* **52**(1) 17–34. 1.2

Chan, C.W., G. Escobar, G.B. Yom-Tov. 2012. When to use speedup: An examination of intensive care units with readmissions. Working Paper, Columbia University. 8

Eick, S.G., W.A. Massey, W. Whitt. 1993a. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science* **39**(2) 241–252. 4.3.2

Eick, S.G., W.A. Massey, W. Whitt. 1993b. The physics of the $M_t/G/\infty$ queue. *Operations Research* **41**(4) 731–742. 2

Falin, G.I., J.G.C. Templeton. 1997. *Retrial Queues*. Chapman & Hall. 1.1

Feldman, Z., A. Mandelbaum, W.A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Science* **54**(2) 324–338. 1.2, 2

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

Green, L., P.J. Kolesar, J. Soares. 2001. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research* **49**(4) 549–564. 2, 5.1

Green, L., P.J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16**(1) 13–39. 2, 5.1

Green, L., J. Soares, J.F. Giglio, R.A. Green. 2006. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13**(1) 61–68. 1.2, 2

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29** 567–587. 2, 3, C

Huang, J., B. Carmeli, A. Mandelbaum. 2012. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. Working Paper, Technion. 8

IMH. 2006. Israel ministry of health - financial report for years 2000–2005. 2

Jacobson, E.U., N.T. Argon, S. Ziya. 2012. Priority assignment in emergency response. *Operations Research* **60**(4) 813–832. 1

Janssen, A.J.E.M., J.S.H. van Leeuwaarden, B. Zwart. 2011. Refining square-root safety staffing by expanding Erlang C. *Operations Research* **59**(6) 1512–1522. 1.2, 5.2

Jennings, O., A. Mandelbaum, W. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Science* **42**(10) 1383–1394. 1.2, 2

Jennings, O.B., F. de Véricourt. 2011. Nurse staffing in medical units: A queueing perspective. *Operations Research* **59**(6) 1320–1331. 1, 1.2, 2

Kaplan, E.L., P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**(282) 457–481. 7.1

Khudyakov, P., M. Gorfine, P. Feigin. 2010. Test for equality of baseline hazard functions for correlated survival data using frailty models. Working Paper, Technion. 1.1

Lahiri, A., A. Seidmann. 2009. Analyzing the differential impact of radiology information systems across radiology modalities. *Journal of the American College of Radiology* **6**(10) 522–526. 1.1

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. 1

Mandelbaum, A., W. Massey, M. Reiman. 1998. Strong approximations for markovian service networks. *Queueing Systems* **30**(1-2) 149–201. 1.2, 4.1, 7, 7, 7, A.1

Mandelbaum, A., W.A. Massey, M. Reiman, B. Rider. 1999. Time varying multiserver queues with abandonment and retrials. P. Key, D. Smith, eds., *ITC-16, Teletraffic Engineering in a Competitive World*. Elsevier, 355–364. 7

Mandelbaum, A., W.A. Massey, M. Reiman, A. Stolyar, B. Rider. 2002. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecom. Systems* **21**(2-4) 149–171. 7

Marmor, Y., D.A. Sinreich. 2005. Emergency department operations: the basis for developing a simulation tool. *IIE Transactions* **37**(3) 233–245. 1.1, 6

Massey, W., W. Whitt. 1993. Networks of infinite-server queues with nonstationary poisson input. *Queueing Systems* **13** 183–250. 1.2, 4.2, A.1

Massey, W., W. Whitt. 1994. An analysis of the modified offered-load approximation for the nonstationary erlang loss model. *Annals of Applied Probability* **4**(4) 1145–1160. 1.2, 5

Whitt, W. 2007. What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics* **55**(5) 476–484. 4.2

Whitt, W. 2013. Offered load analysis for staffing. *Manufacturing and Service Operations Management* **15**(2) 166–169. 1.2

Yom-Tov, G. 2010. Queues in hospitals: Queueing networks with reentering customers in the QED regime. Ph.D. thesis, Technion - Israel Institute of Technology. 2, 8

Zeltyn, S., B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, Y.N. Marmor, A. Mandelbaum, A. Shtub, T. Lauterman, D. Schwartz, K. Moskovitch, S. Tzafrir, F. Basis. 2011. Simulation-based models of emergency departments: Operational, tactical and strategic staffing. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **21**(4) 24. 2

Zhan, D., A.R. Ward. 2012. Routing to minimize waiting and callbacks in large call centers. Working Paper, USC. 1.1

Zychlinski, N., A. Mandelbaum, I. Cohen. 2012. Minimizing mortality in a mass casualty event: Fluid networks in support of modeling and management. Working Paper, Technion. 1.1, 7.1.1

Appendix. Internet Supplement

A. Proofs of Theorems

A.1. The Offered-Load Measure

Proof of Theorem 2 in Section 4.1. Let $Q^\infty = \{Q^\infty(t), t \geq 0\}$ be a 2-dimensional stochastic process, where $Q^\infty(t) = (Q_1^\infty(t), Q_2^\infty(t))$: $Q_1^\infty(t)$ represents the number of *Needy* patients in the system at time t, and $Q_2^\infty(t)$ the number of *Content* patients, assuming we have an infinite number of servers in Node 1 (as well as Node 2).

The process $Q^\infty(t)$ is characterized by the following equations:

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

where A_1^a, A_2^d, A_{12} and A_{21} are four mutually independent, standard (mean rate 1), Poisson processes. We now introduce a family of scaled queues $Q^{\eta,\infty}(t)$, indexed by $\eta > 0$, so that the arrival rate grows to infinity, i.e. scaled up by η, but leaves the Needy and Content rates unscaled. By Theorem 2.2 (FSLLN) in Mandelbaum et al. (1998),

$$\lim_{\eta\to\infty} \frac{Q^{\eta,\infty}(t)}{\eta} = Q^{(0)}(t) \quad u.o.c.\ a.s.,$$

where $Q^{(0)}(\cdot)$ is called the *fluid approximation*, which is the solution to the following ODE:

$$\begin{aligned} Q_1^{(0),\infty}(t) &= Q_1^{(0),\infty}(0) + \int_0^t \left(\lambda_u - \mu Q_1^{(0),\infty}(u) + \delta Q_2^{(0),\infty}(u)\right) du \\ Q_2^{(0),\infty}(t) &= Q_2^{(0),\infty}(0) + \int_0^t \left(p\mu Q_1^{(0),\infty}(u) - \delta Q_2^{(0),\infty}(u)\right) du. \end{aligned}$$

Note that $R(\cdot) = Q^{(0),\infty}(\cdot)$ by definition.

Proof of Theorem 3 in Section 4.2. Following Massey and Whitt (1993), $\lambda_i^+(\cdot)$, which is the aggregated-arrival-rate function to Node i, is given by the minimal non-negative solution to the traffic equations

$$\lambda_1^+(t) = \lambda(t) + E[\lambda_2^+(t - S_2)], \quad \lambda_2^+(t) = pE[\lambda_1^+(t - S_1)], \tag{18}$$

for $t \geq 0$. Then

$$R_i(t) \equiv E[Q_i^\infty(t)] = E\left[\int_{t-S_i}^t \lambda_i^+(u)du\right] = E[\lambda_i^+(t - S_{i,e})]E[S_i], \tag{19}$$

where $S_{i,e}$ is a random variable representing the excess service time at Node i. Equations (18) constitute a variation of Fredholm's integral equation, which one can solve recursively (using the fact that S_1 and S_2 are independent) as follows:

$$\lambda_1^+(t) = \lambda(t) + pE[E[\lambda_1^+(t-S_2-S_1)]] = \cdots = \sum_{j=0}^{\infty} p^j E[\lambda(t-S_1^{*j}-S_2^{*j})],$$
$$\lambda_2^+(t) = pE[\lambda(t-S_1) + E[\lambda_2^+(t-S_1-S_2)]] = \cdots = \sum_{j=1}^{\infty} p^j E[\lambda(t-S_1^{*j}-S_2^{*j-1})].$$

Substituting $\lambda^+(t)$ into $R(t)$ yields

$$\begin{aligned} R_1(t) &= E[\lambda_1^+(t-S_{1,e})]E[S_1] = E\left[\sum_{j=0}^{\infty} p^j \lambda(t-S_1^{*j}-S_2^{*j}-S_{1,e})\right] E[S_1], \\ R_2(t) &= E[\lambda_2^+(t-S_{2,e})]E[S_2] = E\left[\sum_{j=1}^{\infty} p^j \lambda(t-S_1^{*j}-S_2^{*j-1}-S_{2,e})\right] E[S_2]. \end{aligned} \tag{20}$$

Since $J \stackrel{d}{=} Geom_{\geq 0}(1-p)$, $P(J=j) = (1-p)p^j$, which yields the final form of (4).

Proof of Proposition 1 in Section 4.2. Consider the following second-order Taylor-series approximation for the arrival-rate function $\lambda(\cdot)$: $\lambda(t-u) \approx \lambda(t) - \lambda^{(1)}(t)u + \lambda^{(2)}(t)\frac{u^2}{2}$, $u \geq 0$, where $\lambda^{(k)}(t)$ is the k^{th} derivative of $\lambda(\cdot)$ evaluated at time t. Then, from (4) we get an approximation for $R_1(t)$:

$$\begin{aligned} R_1(t) &= \frac{E[S_1]}{1-p} E\left[\lambda(t-S_{1,e}-S_1^{*J}-S_2^{*J})\right] \\ &\approx \frac{E[S_1]}{1-p} E\left[\lambda(t) - \lambda^{(1)}(t)\left(S_{1,e}+S_1^{*J}+S_2^{*J}\right) + \frac{1}{2}\lambda^{(2)}(t)\left(S_{1,e}+S_1^{*J}+S_2^{*J}\right)^2\right] \\ &= \frac{E[S_1]}{1-p}\left[\lambda\left(t - E\left[S_{1,e}+S_1^{*J}+S_2^{*J}\right]\right) + \frac{1}{2}\lambda^{(2)}(t) VAR[S_{1,e}+S_1^{*J}+S_2^{*J}]\right], \end{aligned}$$

where, by Wald's equation, $E\left[S_{1,e}+S_1^{*J}+S_2^{*J}\right] = E\left[S_{1,e}\right] + E[J]E\left[S_1+S_2\right]$, and $VAR\left[S_{1,e}+S_1^{*J}+S_2^{*J}\right] = VAR[S_{1,e}] + E[J]VAR[S_1+S_2] + VAR[J]E[S_1+S_2]$, in which $E[J] = \frac{p}{1-p}$ and $VAR[J] = \frac{p}{1-p^2}$.

A.2. The Offered-Load for Sinusoidal Arrival Rate

Proof of Theorem 4 in Section 4.3.1. Since S_i is exponentially distributed, $S_{i,e} \stackrel{d}{=} S_i$. Defining $X \equiv S_1^{*j_1} \stackrel{d}{=} Erlang(\mu, j_1)$, and $Y \equiv S_2^{*j_2} \stackrel{d}{=} Erlang(\delta, j_2)$:

$$\begin{aligned} E[e^{i\omega X}] &= \int_0^{\infty} e^{i\omega x} \frac{\mu^{j_1} x^{j_1-1} e^{-\mu x}}{(j_1-1)!} dx = \left(\frac{\mu}{\mu - i\omega}\right)^{j_1} := \left(\varphi_{S_1}(\omega)\right)^{j_1}, \\ E[e^{i\omega Y}] &= \left(\frac{\delta}{\delta - i\omega}\right)^{j_2} := \left(\varphi_{S_2}(\omega)\right)^{j_2}; \end{aligned} \tag{21}$$

$$\begin{aligned} &E[\cos(\omega(S_1^{*j_1}+S_2^{*j_2}))] = E[\cos(\omega(X+Y))] = \frac{1}{2}E\left[e^{i\omega(X+Y)} + e^{-i\omega(X+Y)}\right] \\ &= \frac{1}{2}E\left[e^{i\omega X}e^{i\omega Y} + e^{-i\omega X}e^{-i\omega Y}\right] = \frac{1}{2}\left[\left(\varphi_{S_1}(\omega)\right)^{j_1}\left(\varphi_{S_2}(\omega)\right)^{j_2} + \left(\varphi_{S_1}(-\omega)\right)^{j_1}\left(\varphi_{S_2}(-\omega)\right)^{j_2}\right], \end{aligned} \tag{22}$$

and similarly for

$$E[\sin(\omega(S_1^{*j_1}+S_2^{*j_2}))]=\frac{1}{2i}E\left[e^{i\omega(X+Y)}-e^{-i\omega(X+Y)}\right]=\frac{1}{2i}\left[(\varphi_{S_1}(\omega))^{j_1}(\varphi_{S_2}(\omega))^{j_2}-(\varphi_{S_1}(-\omega))^{j_1}(\varphi_{S_2}(-\omega))^{j_2}\right]. \tag{23}$$

Incorporating (22) and (23) into (7) and using $\sin(x-y)=\sin x\cos y-\sin y\cos x$, we get:

$$\begin{aligned}
R_1(t)&=\frac{E[S_1]\bar{\lambda}}{1-p}+E[S_1]\bar{\lambda}\kappa\sum_{j=0}^{\infty}p^jE\left[\sin(\omega t)\cos(\omega(S_1^{*j+1}+S_2^{*j}))-\sin(\omega(S_1^{*j+1}+S_2^{*j}))\cos(\omega t)\right]\\
&=\frac{E[S_1]\bar{\lambda}}{1-p}+E[S_1]\bar{\lambda}\kappa\left[\sin(\omega t)\sum_{j=0}^{\infty}p^j\frac{1}{2}\left[(\varphi_{S_1}(\omega))^{j+1}(\varphi_{S_2}(\omega))^{j}+(\varphi_{S_1}(-\omega))^{j+1}(\varphi_{S_2}(-\omega))^{j}\right]\right.\\
&\left.-\cos(\omega t)\sum_{j=0}^{\infty}p^j\frac{1}{2i}\left[(\varphi_{S_1}(\omega))^{j+1}(\varphi_{S_2}(\omega))^{j}-(\varphi_{S_1}(-\omega))^{j+1}(\varphi_{S_2}(-\omega))^{j}\right]\right]=\\
&=\frac{E[S_1]\bar{\lambda}}{1-p}+E[S_1]\bar{\lambda}\kappa\frac{1}{2}\left[\sin(\omega t)\left[\varphi_{S_1}(\omega)\sum_{j=0}^{\infty}(p\varphi_{S_1}(\omega)\varphi_{S_2}(\omega))^j+\varphi_{S_1}(-\omega)\sum_{j=0}^{\infty}(p\varphi_{S_1}(-\omega)\varphi_{S_2}(-\omega))^j\right]\right.\\
&\left.-\cos(\omega t)\frac{1}{i}\left[\varphi_{S_1}(\omega)\sum_{j=0}^{\infty}(p\varphi_{S_1}(\omega)\varphi_{S_2}(\omega))^j-\varphi_{S_1}(-\omega)\sum_{j=0}^{\infty}(p\varphi_{S_1}(-\omega)\varphi_{S_2}(-\omega))^j\right]\right]=\\
&=\frac{E[S_1]\bar{\lambda}}{1-p}+\frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\left[\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}+\frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]\\
&-\frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)\left[\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}-\frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]\\
&=\frac{E[S_1]\bar{\lambda}}{1-p}+\bar{\lambda}\kappa\sqrt{\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}\cdot\frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}}\cos(\omega t+\pi+tan^{-1}(\theta)),
\end{aligned}$$

where

$$\theta=i\cdot\frac{\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}+\frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}}{\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}-\frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}}=\frac{-\mu(-\delta^2+p\delta^2-\omega^2)}{\omega(\delta^2+\omega^2+p\mu\delta)}.$$

Similar calculations for $\lambda_1^+(t)$ yield the following theorem:

THEOREM 8. *Assuming that S_i are exponentially distributed, $\lambda_1^+(\cdot)$ has the following form:*

$$\lambda_1^+(t)=\frac{\bar{\lambda}}{1-p}+\bar{\lambda}\kappa\sqrt{\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}\cdot\frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}}\cos(\omega t+\pi+tan^{-1}(\theta)), \tag{24}$$

where

$$\theta=i\cdot\frac{\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}+\frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}}{\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}-\frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}}=\frac{\omega^2\delta^2+\omega^4+\omega^2p\mu\delta+\mu^2\delta^2-\mu^2p\delta^2+\mu^2\omega^2}{\mu\omega p\delta(\mu+\delta)}.$$

Therefore, the amplitude of $\lambda_1^+(\cdot)$ is given by

$$Amp(\lambda_1^+)=\bar{\lambda}\kappa\sqrt{\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}\cdot\frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}} \tag{25}$$

and the phase of $\lambda_1^+(\cdot)$ is given by

$$Phase(\lambda_1^+)=\frac{1}{2\pi}cot^{-1}\left(\frac{\omega^2\delta^2+\omega^4+\omega^2p\mu\delta+\mu^2\delta^2-\mu^2p\delta^2+\mu^2\omega^2}{\mu\omega p\delta(\mu+\delta)}\right).$$

A.3. Comparing to Erlang-C

Proof of Theorem 5 in Section 4.3.2. We must prove that $AmpRatio \leq 1$, which is given by:

$$AmpRatio = \sqrt{\frac{\delta^2+\omega^2}{((\mu-i\omega)(\delta-i\omega)-p\mu\delta)((\mu+i\omega)(\delta+i\omega)-p\mu\delta)}} \Big/ \frac{1}{\sqrt{((1-p)\mu)^2+\omega^2}}.$$

Thus, we shall prove that:

$$\frac{(\delta^2+\omega^2)((1-p)^2\mu^2+\omega^2)}{[(\mu-i\omega)(\delta-i\omega)-p\mu\delta][(\mu+i\omega)(\delta+i\omega)-p\mu\delta]} \overset{?}{<} 1$$

$$\frac{\delta^2(1-p)^2\mu^2+\omega^2(1-p)^2\mu^2+\delta^2\omega^2+\omega^4}{(\mu-i\omega)(\delta-i\omega)(\mu+i\omega)(\delta+i\omega)-p\mu\delta[(\mu+i\omega)(\delta+i\omega)+(\mu-i\omega)(\delta-i\omega)]+p^2\mu^2\delta^2} \overset{?}{<} 1$$

$$\frac{\delta^2(1-p)^2\mu^2+\omega^2(1-p)^2\mu^2+\delta^2\omega^2+\omega^4}{(\mu^2+\omega^2)(\delta^2+\omega^2)-p\mu\delta(2\mu\delta-2\omega^2)+p^2\mu^2\delta^2} \overset{?}{<} 1$$

$$\delta^2(1-p)^2\mu^2+\omega^2(1-p)^2\mu^2+\delta^2\omega^2+\omega^4 \overset{?}{<} \mu^2\delta^2+\omega^2\delta^2+\mu^2\omega^2+\omega^4+2p\mu\delta(\omega^2-\mu\delta)+p^2\mu^2\delta^2$$

$$\delta^2(1-p)^2\mu^2+\omega^2(1-p)^2\mu^2 \overset{?}{<} \mu^2\omega^2+\mu^2\delta^2(1-p)^2+2p\mu\delta\omega^2$$

$$\omega^2(1-p)^2\mu^2 \overset{?}{<} \mu^2\omega^2+2p\mu\delta\omega^2,$$

which is true for every μ, δ, ω, and $0 < p \leq 1$.

In the second part of the theorem, one must prove that *AmpRatio* reaches its minimum at $\omega = \sqrt{\delta\mu(1-p)}$. The derivative of *AmpRatio* with respect to ω is:

$$\frac{\partial AmpRatio}{\partial \omega} = \frac{2p\omega\mu(2\delta+(2-p)\mu)(\omega^2+(1-p)\mu\delta)(\omega^2-(1-p)\mu\delta)}{(\omega^4+(p-1)^2\delta^2\mu^2+\omega^2(\delta^2+2p\delta\mu+\mu^2))^2}.$$

This derivative vanishes when $\omega = 0$ or $\omega = \sqrt{\delta\mu(1-p)}$. For $\omega = 0$, the *AmpRatio* reaches its maximum which is 1, and at $\omega = \sqrt{\delta\mu(1-p)}$ it reaches its minimal value.

The third part of the theorem is a direct result of the limits of $R_1(t)$ as presented in Proposition 3 below.

A.4. Analysis of Limits of $R(\cdot)$ with Sinusoidal Arrivals and Exponential Services

We now further investigate the relative amplitudes of the offered-load $R_1(\cdot)$ and the aggregate arrival rate $\lambda_1^+(\cdot)$, when all service times are exponential. We state the following proposition that highlights some of the limits of $R_1(\cdot)$ and $\lambda_1^+(\cdot)$ with respect to ω and δ:

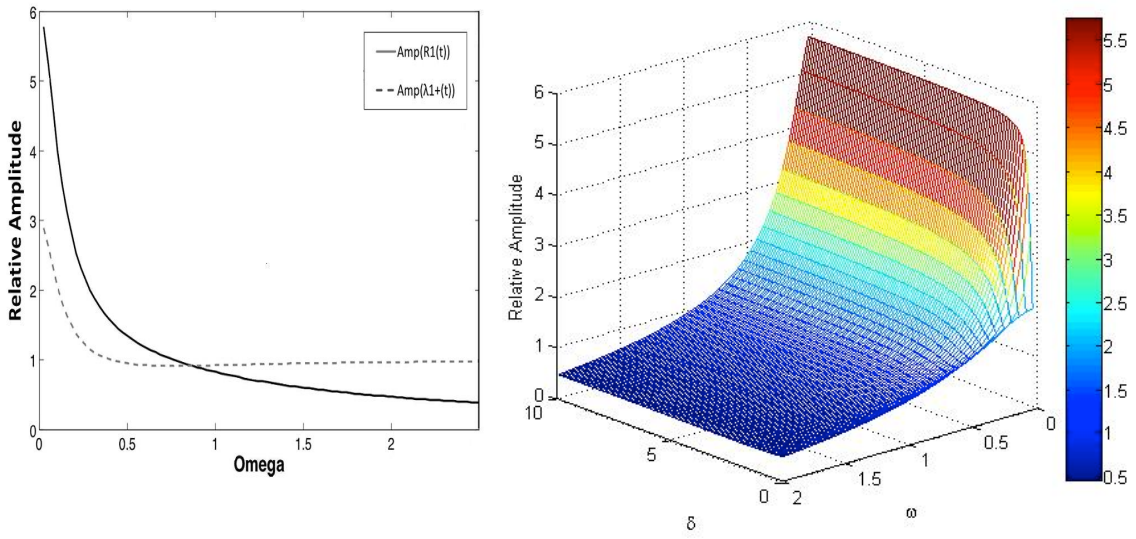
PROPOSITION 3. *In the case of sinusoidal arrival rates and exponential service times, with μ and δ being fixed:*

$$\lim_{\omega\downarrow 0} Amp(R_1(\cdot)) = \frac{\bar{\lambda}}{\mu(1-p)}\kappa, \quad \lim_{\omega\uparrow\infty} Amp(R_1(\cdot)) = 0,$$

$$\lim_{\omega\downarrow 0} Amp(\lambda_1^+(\cdot)) = \frac{\bar{\lambda}}{1-p}\kappa, \quad \lim_{\omega\uparrow\infty} Amp(\lambda_1^+(\cdot)) = \bar{\lambda}\kappa;$$

Figure 11 Plot of Relative Amplitude.

(a) $R_1(t)$ and $\lambda_1^+(t)$ with respect to ω (b) $R_1(t)$ with respect to δ and ω



if μ and ω are fixed:

$$\lim_{\delta\downarrow 0} R_1(t) = \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{\bar{\lambda}\kappa}{\mu^2+\omega^2}\left(\mu\sin(\omega t) - \omega\cos(\omega t)\right),$$
$$\lim_{\delta\uparrow\infty} R_1(t) = \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{\bar{\lambda}\kappa}{(1-p)^2\mu^2+\omega^2}\left((1-p)\mu\sin(\omega t) - \omega\cos(\omega t)\right).$$

Proof: The limits are obtained by straightforward calculations, based on (8), (9), and (25).

We would like to understand the changes in $R_1(\cdot)$ and $\lambda_1^+(\cdot)$ with respect to the external arrival rate $\lambda(\cdot)$. We call the ratio between the amplitudes *relative amplitude*. Figure 11a shows the relative amplitude of $R_1(\cdot)$ and $\lambda_1^+(\cdot)$, as a function of ω (μ and δ are fixed). We observe that the relative amplitude of $R_1(\cdot)$ is a decreasing function of ω, starting from the value $\frac{1}{\mu(1-p)}$, and decreasing to 0 as $\omega\to\infty$. On the other hand, $\lambda_1^+(\cdot)$ starts from the value $\frac{1}{1-p}$, and tends to 1 as $\omega\to\infty$. Figure 11b shows the relative amplitude of $R_1(\cdot)$ as a function of ω and δ (when $\mu=0.5$). We observe that the relative amplitude of $R_1(\cdot)$ is an increasing function of δ, starting from the value $\frac{1}{\sqrt{\mu^2+\omega^2}}$, and increasing to $\frac{E[S_1]\bar{\lambda}}{1-p} + \frac{\bar{\lambda}\kappa}{\sqrt{(1-p)^2\mu^2+\omega^2}}$, as $\delta\to\infty$. When $\delta\to 0$, the extreme values of $R_1(\cdot)$ are $\max_t(R_1(t)) = \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{\bar{\lambda}\kappa}{\sqrt{\mu^2+\omega^2}}$, and the relative amplitude is $\frac{1}{\sqrt{\mu^2+\omega^2}}$. When $\delta\to\infty$, the extreme values of $R_1(t)$ are $\max_t(R_1(t)) = \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{\bar{\lambda}\kappa}{\sqrt{(1-p)^2\mu^2+\omega^2}}$, and the relative amplitude is $\frac{1}{\sqrt{(1-p)^2\mu^2+\omega^2}}$.

A.5. Deterministic Service Times

We now discuss shortly deterministic service times. These are not usually found in healthcare systems, where exponential service times provide a good enough approximation for many applications. Nevertheless, they are common in manufacturing and communication and, moreover, they add insight here as well.

THEOREM 9. *Assume that S_i are deterministic, and the arrival rate is given by (6). Then, for $t \geq 0$,*

$$\lambda_1^+(t) = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa Re\left\{\frac{e^{i(\omega t - \frac{\pi}{2})}}{1-pe^{-i\omega(S_1+S_2)}}\right\}$$

and

$$R_1(t) = S_1\frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\left[Re\left\{\frac{\frac{1}{-i\omega}(e^{i(\omega(t-S_1)-\frac{\pi}{2})} - e^{i(\omega t-\frac{\pi}{2})})}{1-pe^{-i\omega(S_1+S_2)}}\right\}\right].$$

Proof We start with $\lambda_1^+(\cdot)$. In the deterministic case, $E[S_i^{*j}] = jS_i$. Consequently,

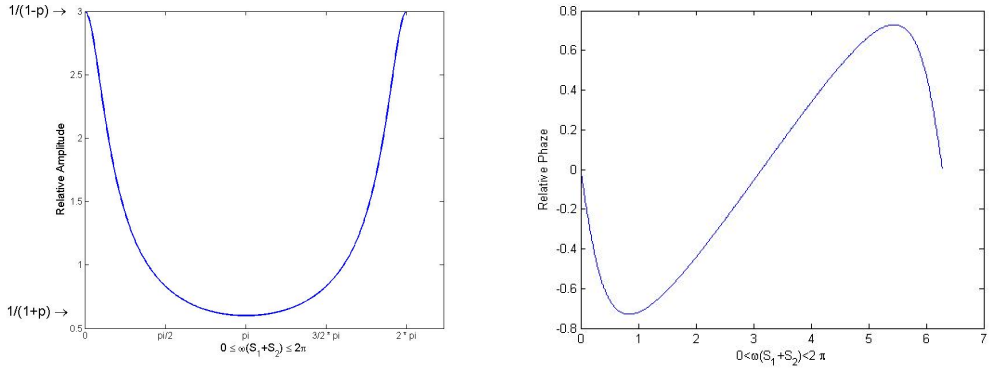
$$\lambda_1^+(t) = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\sum_{j=0}^{\infty}p^j E[\sin(\omega(t - S_1^{*j} + S_2^{*j}))] = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\sum_{j=0}^{\infty}p^j\sin(\omega(t - jS_1 + jS_2))$$
$$= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\sum_{j=0}^{\infty}p^j\cos\left(\omega(t-j(S_1+S_2)) - \frac{\pi}{2}\right) = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\sum_{j=0}^{\infty}p^j Re\left\{e^{i(\omega(t-j(S_1+S_2))-\frac{\pi}{2})}\right\}$$
$$= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa Re\left\{e^{i(\omega t-\frac{\pi}{2})}\sum_{j=0}^{\infty}p^j e^{-ij\omega(S_1+S_2)}\right\} = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa Re\left\{\frac{e^{i(\omega t-\frac{\pi}{2})}}{1-pe^{-i\omega(S_1+S_2)}}\right\}.$$

In order to calculate $R_1(t)$, we note that $S_{i,e}$ is uniformly distributed over $[0, S_i]$. Therefore:

$$R_1(t) = E[S_1]E[\lambda^+(t - S_{1,e})] = S_1\frac{\bar{\lambda}}{1-p} + S_1\bar{\lambda}\kappa E\left[Re\left\{\frac{e^{i(\omega(t-S_{1,e})-\frac{\pi}{2})}}{1-pe^{-i\omega(S_1+S_2)}}\right\}\right]$$
$$= S_1\frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\int_0^{S_1}\left[Re\left\{\frac{e^{i(\omega(t-x)-\frac{\pi}{2})}}{1-pe^{-i\omega(S_1+S_2)}}\right\}\right]dx = S_1\frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\left[Re\left\{\frac{\frac{1}{-i\omega}(e^{i(\omega(t-S_1)-\frac{\pi}{2})} - e^{i(\omega t-\frac{\pi}{2})})}{1-pe^{-i\omega(S_1+S_2)}}\right\}\right].$$

Figure 12 shows the changes in relative amplitude and phase as a function of $\omega\cdot(S_1+S_2)$. The deterministic case exhibits different characteristics from the exponential. First, the amplitude of $\lambda_1^+(\cdot)$ can reach as high as $\frac{\bar{\lambda}\kappa}{1-p}$ and as low as $\frac{\bar{\lambda}\kappa}{1+p}$; the former as in the exponential case, the latter in contrast to the exponential case where the minimal amplitude is $\bar{\lambda}\kappa$ (equals the arrival rate amplitude). Second, we now observe a cyclic behavior, where the amplitude is maximal when $\omega(S_1+S_2) = 2\pi j$ (for some integer j), and minimal when $\omega(S_1+S_2) = \pi j$; in the former case, the returning stream from Node 2 is fully synchronized with the external input stream $\lambda(\cdot)$ ($\frac{S_1+S_2}{f}$ is an integer), and in the latter the returning stream balances the external input stream. This is very different from the exponential case where we observed monotonicity and the amplitude decreases in ω. Finally, Erlang-R is most needed if $\omega(S_1+S_2) \approx 0.25\pi j$ or $\approx 1.75\pi j$, when both phase and amplitude are influenced by the reentering customers (patients). Note that, due to the cyclic shape of the

Figure 12 **Plot of relative amplitude and phase of $\lambda_1^+(t)$ as a function of ω.**



amplitude and phase functions, special care is required when optimizing the system. For example, reducing LOS (length-of-stay) is often attempted by reducing Needy and Content times (S_1 and S_2). However, if the system operates in the decreasing region of the left Figure 12, shortening S_1 or S_2 will increase the amplitude of $\lambda_1^+(\cdot)$, and therefore the amplitude of $R_1(\cdot)$ will also increase, which could destabilize the system. Indeed, a system in which staffing amplitude increases becomes more challenging to operate.

A.6. Time-Varying Diffusion Approximations

The Stochastic Differential Equations underlying Theorem 7 are:

$$\begin{aligned}
Q_1^{(1)}(t) &= Q_1^{(1)}(0) + \int_0^t \left(\mu 1_{\left\{Q_1^{(0)}(u)\le s_u\right\}} Q_1^{(1)}(u)^- - \mu 1_{\left\{Q_1^{(0)}(u)< s_u\right\}} Q_1^{(1)}(u)^+ + \delta Q_2^{(1)}(u)\right) du \\
&\quad + B_1^a\left(\int_0^t \lambda_u du\right) - B_2^d\left(\int_0^t p\mu\left(Q_1^{(0)}(u)\wedge s_u\right) du\right) - B_{12}\left(\int_0^t (1-p)\mu\left(Q_1^{(0)}(u)\wedge s_u\right) du\right) \\
&\quad + B_{21}\left(\int_0^t \delta Q_2^{(0)}(u) du\right), \qquad (26)\\
Q_2^{(1)}(t) &= Q_2^{(1)}(0) + \int_0^t \left(p\mu 1_{\left\{Q_1^{(0)}(u)< s_u\right\}} Q_1^{(1)}(u)^+ - p\mu 1_{\left\{Q_1^{(0)}(u)\le s_u\right\}} Q_1^{(1)}(u)^- - \delta Q_2^{(1)}(u)\right) du \\
&\quad + B_{12}\left(\int_0^t p\mu\left(Q_1^{(0)}(u)\wedge s_u\right) du\right) - B_{21}\left(\int_0^t \delta Q_2^{(0)}(u) du\right),
\end{aligned}$$

where B_1^a, B_2^d, B_{12} and B_{21} are four mutually independent, standard Brownian motions; $x^+ \equiv \max(x,0)$, and $x^- \equiv \max(-x,0) = -\min(x,0)$.

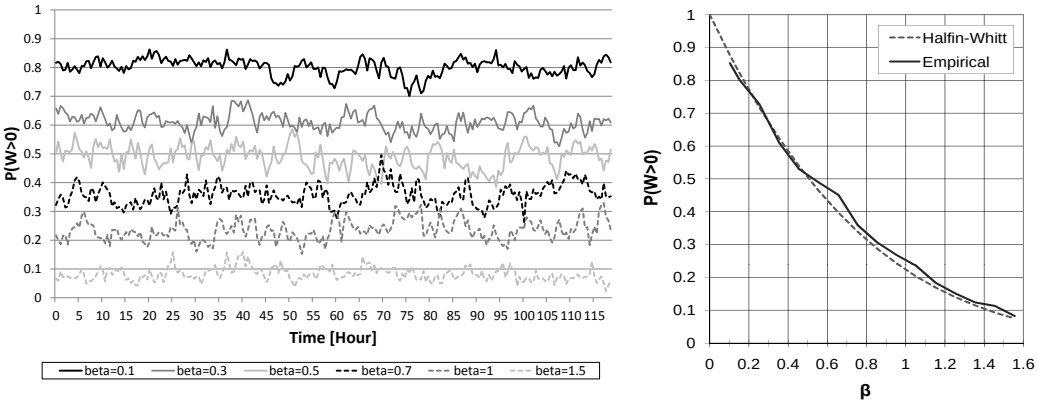
The following theorem presents the mean vector and the covariance matrix for the diffusion limit.

THEOREM 10. *Using the scaling (11), the mean vector for the diffusion limit (26) is the unique solution to the following two differential equations:*

$$\begin{aligned}
\frac{d}{dt}\mathrm{E}\left[Q_1^{(1)}(t)\right] &= \mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^-\right] - \mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^+\right] + \delta\mathrm{E}\left[Q_2^{(1)}(t)\right], \\
\frac{d}{dt}\mathrm{E}\left[Q_2^{(1)}(t)\right] &= p\mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^+\right] - p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^-\right] - \delta\mathrm{E}\left[Q_2^{(1)}(t)\right].
\end{aligned} \qquad (27)$$

Figure 13 **Case study 1 - $P(W_t>0)$ for various β values in large systems.**

(a) $P(W_t>0)$ (b) Halfin-Whitt formula vs. simulation



The covariance matrix for the diffusion limit solves:

$$\begin{aligned}\frac{d}{dt}\mathrm{Var}\left[Q_1^{(1)}(t)\right] &= 2\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{Cov}\left[Q_1^{(1)}(t),Q_1^{(1)}(t)^-\right]-2\mu 1_{\left\{Q_1^{(0)}(t)<s_t\right\}}\mathrm{Cov}\left[Q_1^{(1)}(t),Q_1^{(1)}(t)^+\right]\\ &\quad+2\delta\mathrm{Cov}\left[Q_1^{(1)}(t),Q_2^{(1)}(t)\right]+\lambda_t+\mu\left(Q_1^{(0)}(t)\wedge s_t\right)+\delta Q_2^{(0)}(t),\end{aligned} \tag{28}$$

$$\begin{aligned}\frac{d}{dt}\mathrm{Var}\left[Q_2^{(1)}(t)\right] &= 2p\mu 1_{\left\{Q_1^{(0)}(t)<s_t\right\}}\mathrm{Cov}\left[Q_2^{(1)}(t),Q_1^{(1)}(t)^+\right]\\ &\quad-2p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{Cov}\left[Q_2^{(1)}(t),Q_1^{(1)}(t)^-\right]-2\delta\mathrm{Var}\left[Q_2^{(1)}(t)\right]\\ &\quad+p\mu\left(Q_1^{(0)}(t)\wedge s_t\right)+\delta Q_2^{(0)}(t),\end{aligned}$$

$$\begin{aligned}\frac{d}{dt}\mathrm{Cov}\left[Q_1^{(1)}(t),Q_2^{(1)}(t)\right] &= \mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{Cov}\left[Q_2^{(1)}(t),Q_1^{(1)}(t)^-\right]-\mu 1_{\left\{Q_1^{(0)}(t)<s_t\right\}}\mathrm{Cov}\left[Q_2^{(1)}(t),Q_1^{(1)}(t)^+\right]\\ &\quad+\delta\left(\mathrm{Var}\left[Q_2^{(1)}(t)\right]-\mathrm{Cov}\left[Q_1^{(1)}(t),Q_2^{(1)}(t)\right]\right)+p\mu 1_{\left\{Q_1^{(0)}(t)<s_t\right\}}\mathrm{Cov}\left[Q_1^{(1)}(t),Q_1^{(1)}(t)^+\right]\\ &\quad-p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{Cov}\left[Q_1^{(1)}(t),Q_1^{(1)}(t)^-\right]-\delta Q_2^{(0)}(t)-p\mu\left(Q_1^{(0)}(t)\wedge s_t\right).\end{aligned}$$

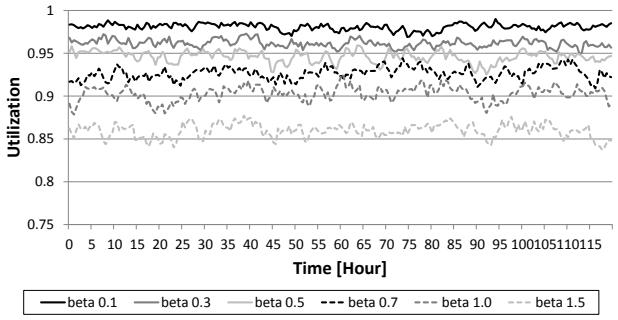
B. Stabilizing large Erlang-R network: Additional graphs for case study 1

In this appendix, we provide additional support that Erlang-R can stabilize various performance measures. Our testing ground is the large-scale Erlang-R queueing network, considered in Section 5.1.

Figure 13a depicts $P(W_t>0)$ over a 5-day period (120 hours), for six values of β. The performance measure is visibly stable, which indicates that the MOL algorithm works well. As mentioned before, we expect the relation between $P(W>0)$ and β to fit the Halfin-Whitt formula. We validated this by calculating the average waiting probability for the time-varying system, for each value of β, and comparing it to the steady-state Halfin-Whitt formula. In Figure 13b, the two are clearly very close to each other.

Figure 14 **Case study 1 - Simulation results of server utilization.**

(a) Utilization for various β values



(b) Time averages vs. theoretical values

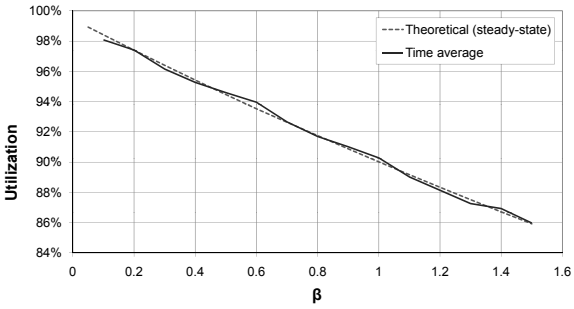


Figure 15 **Case study 1 - Simulation results of $E[W_t]$ for various β values in large systems.**

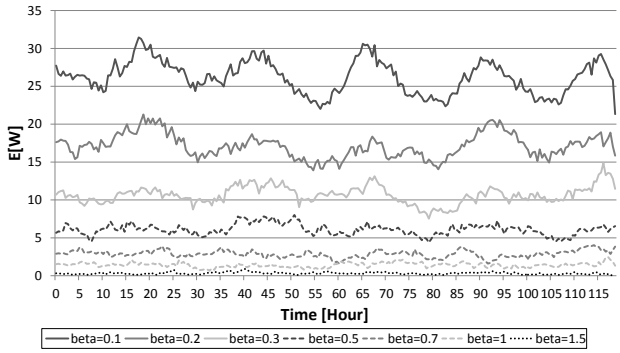


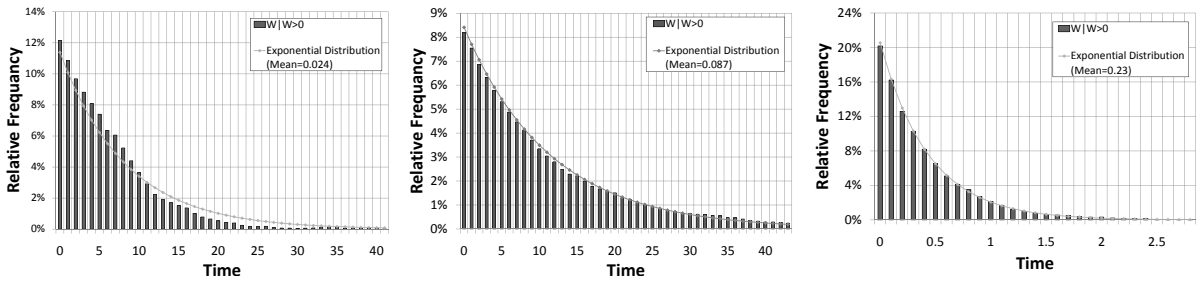
Figure 14a shows the evolution of servers utilization over time, for each value of β, which is also stable. Thus our staffing procedure stabilizes *both* service level and server utilization. In Figure 14b, we compare the average utilization over time with the theoretical values. The latter were calculated using the steady-state solution of our model, when given average values of λ and s. We observe that the two are almost identical.

Figure 15 depicts $E[W_t]$ over a 5-day period. We note that, as β grows, $E[W_t]$ becomes more stable and well ordered.

Figure 16 displays the conditional distribution of the waiting time given delay $(W|W>0)$, for three values of β (0.1,0.5, and 1.4). We compare them 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 for $\beta=0.5$ (QED) and $\beta=1.4$ (QD (Quality Driven)), but when β is 0.1 (ED (Efficiency Driven)), the quality of fit deteriorates visibly. This is in line with our observations for $E[W_t]$, where small values of β give rise to a performance that does vary in time and hence does not fit steady-state.

Figure 16 **Case study 1 - A comparison of the histogram of $W|W>0$ with the corresponding theoretical distribution.**

(a) $\beta = 0.1$ (b) $\beta = 0.5$ (c) $\beta = 1.4$



C. Approximating the Number of Needy Customers and Waiting Times in the QED Regime

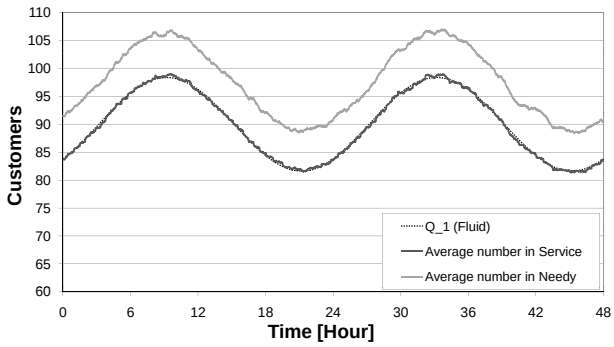
In this section, we derive QED approximations for the actual number of customers in the system and the virtual waiting time process. One could attempt to use the fluid and diffusion approximations developed in Section 7 for this purpose. However, these approximations work well under the zero-measure assumption, and when the system operates in the QED regime, the system is critical at **all** times. The problem when using these approximations under QED staffing is twofold: first, we have numerical difficulties in calculating the diffusion process itself since the diffusion approximation is non-autonomous. Second, the fluid process itself has a different interpretation under the QED regime: no longer does it represent the average behavior of its originating stochastic system.

To understand the interpretation problem, we use the following example from Case Study 1. Figure 17a shows the fluid solution of the process $Q_1^{(0)}(\cdot)$ (the number of Needy customers), as well as the following simulation results: the average number of customers in the Needy state, and the average number of customers in service. We note that the fluid model fits perfectly the number of customers *in service* and ignores the number of customers waiting in queue (for service). This is because our MOL staffing procedure keeps the staffing level always slightly above the average number of customers. Thus, the fluid approximation "sees" the system as if it had an infinite number of servers, and actually calculates the number of busy servers, without the queue.

In order to fill the gap and to estimate correctly the number of Needy customers (in queue and in service), recall the insight (§5.1) that, under MOL staffing, the system behaves as if the Needy state were a stationary M/M/s model (Erlang-C). Therefore, we attempt to use the stationary approximation of the Erlang-C

Figure 17 $Q_1(t)$ - **Fluid approximation vs. simulation results under QED staffing, for various β's.**

(a) Fluid vs. Simulation



(b) Corrected approximation vs. simulation

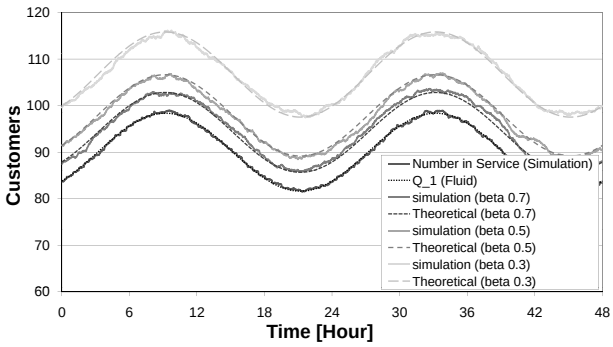
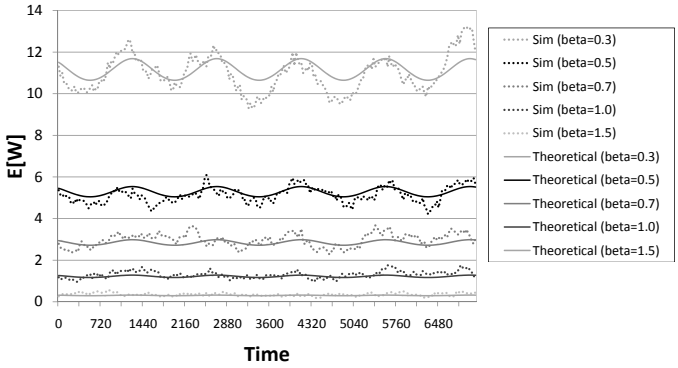


Figure 18 $E[W_t]$ - **Corrected Fluid approximation vs. simulation for various β's.**



model to estimate the number of customers in the queue. Halfin and Whitt (1981) approximated $E[Q(\infty)]$ by the following formula: $E[Q_1(\infty)] = \frac{\lambda}{\mu} + \alpha\frac{\lambda}{s\mu}\left(1-\frac{\lambda}{s\mu}\right)^{-1}$, with α in Theorem 1. We propose an MOL correction, adjusting this formula to time-varying environments, in the following manner: $E[Q_1(t)] = R(t) + \alpha\frac{R(t)}{s(t)}\left(1-\frac{R(t)}{s(t)}\right)^{-1}$. Figure 17b compares this corrected approximation to simulation results for various β values. We observe that the simulation and approximation are remarkably close.

One can also provide a correction to the $E[W_t]$ function in the QED regime, using the following expression: $E[W_t] = \frac{\alpha}{\mu s(t)}\left(1-\frac{R(t)}{s(t)}\right)^{-1}$. Experiments show that this correction works well for $\beta > 0.3$, as is apparent in Figure 18.