Internet Supplement

EC.1. Proofs of Theorems

EC.1.1. 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),

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

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

$$\lambda_1^+(t) = \lambda(t) + E[\lambda_2^+(t - S_2)], \quad \lambda_2^+(t) = pE[\lambda_1^+(t - S_1)], \tag{EC.1}$$

for $t \geq 0$. Then

$$R_i(t) \equiv E[Q_i^{\infty}(t)] = E\left[\int_{t-S_i}^{t} \lambda_i^+(u)du\right] = E[\lambda_i^+(t-S_{i,e})]E[S_i], \tag{EC.2}$$

where $S_{i,e}$ is a random variable representing the excess service time at Node i. Equations (EC.1) 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

$$\begin{aligned} R_1(t) &= E[\lambda_1^+(t-S_{1,e})]E[S_1] = E\left[\sum_{j=0}^{\infty} p^j \lambda(t-S_1^{*j}-S_2^{*j}-S_{1,e})\right] E[S_1], \\ R_2(t) &= E[\lambda_2^+(t-S_{2,e})]E[S_2] = E\left[\sum_{j=1}^{\infty} p^j \lambda(t-S_1^{*j}-S_2^{*j-1}-S_{2,e})\right] E[S_2]. \end{aligned} \tag{EC.3}$$

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 $VAR[J] = \frac{p}{1-p^2}$.

EC.1.2. 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 $Y \equiv S_2^{*j_2} \stackrel{d}{=} Erlang(\delta, j_2)$:

$$\begin{aligned} E[e^{i\omega X}] &= \int_0^\infty e^{i\omega x} \frac{\mu^{j_1} x^{j_1-1} e^{-\mu x}}{(j_1-1)!} dx = \left(\frac{\mu}{\mu - i\omega}\right)^{j_1} := (\varphi_{S_1}(\omega))^{j_1}, \\ E[e^{i\omega Y}] &= \left(\frac{\delta}{\delta - i\omega}\right)^{j_2} := (\varphi_{S_2}(\omega))^{j_2}; \end{aligned} \tag{EC.4}$$

$$\begin{aligned} &E[\cos(\omega(S_1^{*j_1} + S_2^{*j_2}))] = E[\cos(\omega(X+Y))] = \frac{1}{2} E\left[e^{i\omega(X+Y)} + e^{-i\omega(X+Y)}\right] \\ &= \frac{1}{2} E\left[e^{i\omega X} e^{i\omega Y} + e^{-i\omega X} e^{-i\omega Y}\right] = \frac{1}{2}\left[(\varphi_{S_1}(\omega))^{j_1} (\varphi_{S_2}(\omega))^{j_2} + (\varphi_{S_1}(-\omega))^{j_1} (\varphi_{S_2}(-\omega))^{j_2}\right], \end{aligned} \tag{EC.5}$$

and similarly for

$$E[\sin(\omega(S_1^{*j_1} + S_2^{*j_2}))] = \frac{1}{2i} E\left[e^{i\omega(X+Y)} - e^{-i\omega(X+Y)}\right] = \frac{1}{2i}\left[(\varphi_{S_1}(\omega))^{j_1} (\varphi_{S_2}(\omega))^{j_2} - (\varphi_{S_1}(-\omega))^{j_1} (\varphi_{S_2}(-\omega))^{j_2}\right]. \tag{EC.6}$$

Incorporating (EC.5) and (EC.6) into (7) and using $\sin(x-y) = \sin x \cos y - \sin y \cos x$, we get:

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

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 EC.1. *Assuming that S_i are exponentially distributed, $\lambda_1^+(\cdot)$ has the following form:*

$$\lambda_1^+(t) = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\sqrt{\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}\cdot\frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}}\cos\left(\omega t+\pi+tan^{-1}(\theta)\right), \tag{EC.7}$$

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

$$Amp(\lambda_1^+) = \bar{\lambda}\kappa\sqrt{\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}\cdot\frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}} \tag{EC.8}$$

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

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

EC.1.3. Comparing to Erlang-C

Proof of Theorem 5 in Section 4.3.2. We must prove that $AmpRatio \le 1$, 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 $0 < p \le 1$.

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 EC.1 below.

EC.1.4. 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 EC.1. *In the case of sinusoidal arrival rates and exponential service times, with μ and δ being fixed:*

$$\lim_{\omega\downarrow 0}Amp(R_1(\cdot))=\frac{\bar\lambda}{\mu(1-p)}\kappa,\quad \lim_{\omega\uparrow\infty}Amp(R_1(\cdot))=0,$$
$$\lim_{\omega\downarrow 0}Amp(\lambda_1^+(\cdot))=\frac{\bar\lambda}{1-p}\kappa,\quad \lim_{\omega\uparrow\infty}Amp(\lambda_1^+(\cdot))=\bar\lambda\kappa;$$

if μ and ω are fixed:

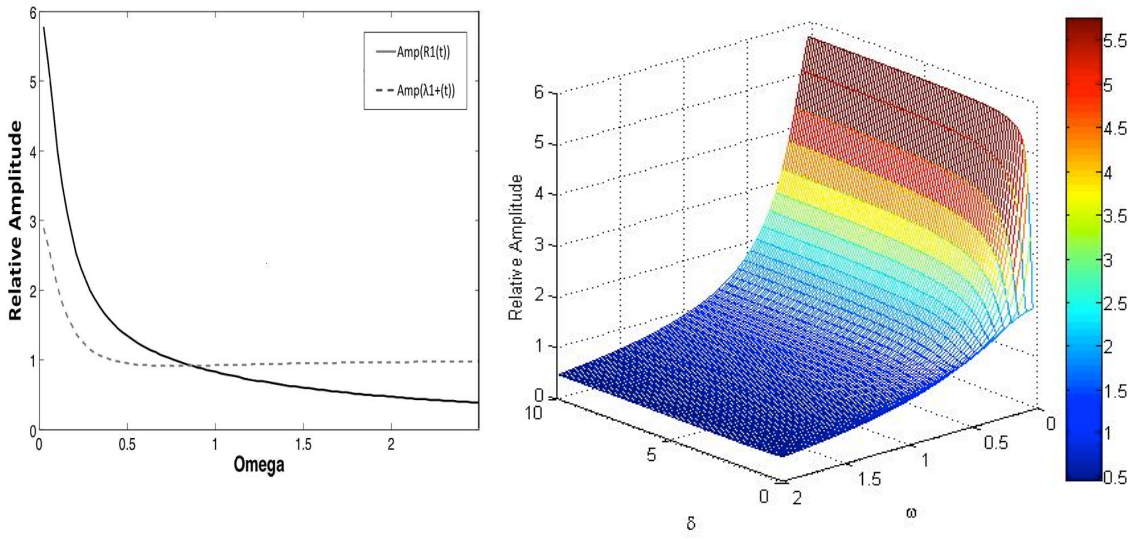
$$\lim_{\delta\downarrow 0}R_1(t)=\frac{E[S_1]\bar\lambda}{1-p}+\frac{\bar\lambda\kappa}{\mu^2+\omega^2}\left(\mu\sin(\omega t)-\omega\cos(\omega t)\right),$$
$$\lim_{\delta\uparrow\infty}R_1(t)=\frac{E[S_1]\bar\lambda}{1-p}+\frac{\bar\lambda\kappa}{(1-p)^2\mu^2+\omega^2}\left((1-p)\mu\sin(\omega t)-\omega\cos(\omega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

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

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 EC.1a 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

Figure EC.1 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



$\omega \to \infty$. Figure EC.1b 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 $\frac{1}{\sqrt{(1-p)^2\mu^2+\omega^2}}$.

EC.1.5. 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 EC.2. *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

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

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

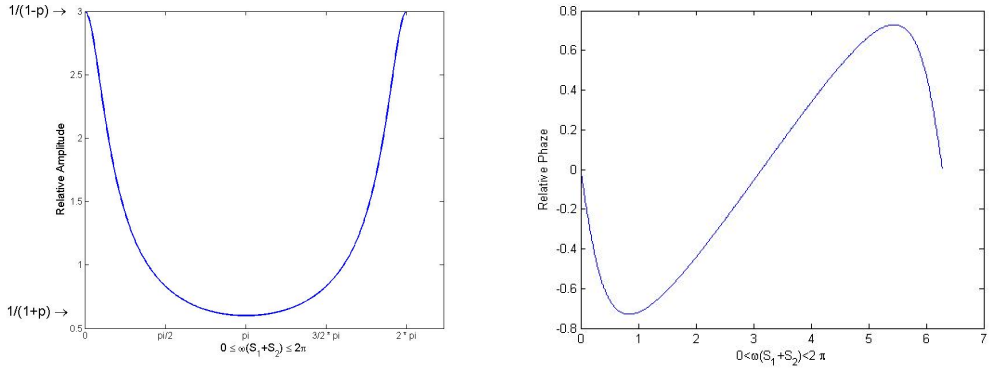
$$\lambda_1^+(t) = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \sum_{j=0}^{\infty} p^j E[\sin(\omega(t - S_1^{*j} + S_2^{*j}))] = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \sum_{j=0}^{\infty} p^j \sin(\omega(t - jS_1 + jS_2))$$
$$= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \sum_{j=0}^{\infty}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

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

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

Figure EC.2 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 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 EC.2, shortening S_1 or S_2 will increase the amplitude of $\lambda_1^+(\cdot)$, and therefore the amplitude of $R_1(\cdot)$ will

Figure EC.2 Plot of relative amplitude and phase of $\lambda_1^+(t)$ as a function of ω.



also increase, which could destabilize the system. Indeed, a system in which staffing amplitude increases becomes more challenging to operate.

EC.1.6. Time-Varying Diffusion Approximations

The Stochastic Differential Equations underlying Theorem 7 are:

$$\begin{aligned}
Q_1^{(1)}(t) &= Q_1^{(1)}(0) + \int_0^t \left(\mu 1_{\left\{Q_1^{(0)}(u) \le s_u\right\}} Q_1^{(1)}(u)^- - \mu 1_{\left\{Q_1^{(0)}(u) < s_u\right\}} Q_1^{(1)}(u)^+ + \delta Q_2^{(1)}(u) \right) du \\
&\quad + B_1^a \left(\int_0^t \lambda_u du \right) - B_2^d \left(\int_0^t p\mu \left(Q_1^{(0)}(u) \wedge s_u \right) du \right) - B_{12} \left(\int_0^t (1-p)\mu \left(Q_1^{(0)}(u) \wedge s_u \right) du \right) \\
&\quad + B_{21} \left(\int_0^t \delta Q_2^{(0)}(u) du \right), \\
Q_2^{(1)}(t) &= Q_2^{(1)}(0) + \int_0^t \left(p\mu 1_{\left\{Q_1^{(0)}(u) < s_u\right\}} Q_1^{(1)}(u)^+ - p\mu 1_{\left\{Q_1^{(0)}(u) \le s_u\right\}} Q_1^{(1)}(u)^- - \delta Q_2^{(1)}(u) \right) du \\
&\quad + B_{12} \left(\int_0^t p\mu \left(Q_1^{(0)}(u) \wedge s_u \right) du \right) - B_{21} \left(\int_0^t \delta Q_2^{(0)}(u) du \right),
\end{aligned} \tag{EC.9}$$

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 EC.3. *Using the scaling* (11)*, the mean vector for the diffusion limit* (EC.9) *is the unique solution to the following two differential equations:*

$$\begin{aligned}\frac{d}{dt}\mathrm{E}\left[Q_1^{(1)}(t)\right] &= \mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^-\right] - \mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^+\right] + \delta\mathrm{E}\left[Q_2^{(1)}(t)\right],\\ \frac{d}{dt}\mathrm{E}\left[Q_2^{(1)}(t)\right] &= p\mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^+\right] - p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^-\right] - \delta\mathrm{E}\left[Q_2^{(1)}(t)\right].\end{aligned} \tag{EC.10}$$

The covariance matrix for the diffusion limit solves:

$$\begin{aligned}\frac{d}{dt}\mathrm{Var}\left[Q_1^{(1)}(t)\right] &= 2\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{Cov}\left[Q_1^{(1)}(t),Q_1^{(1)}(t)^-\right] - 2\mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}\mathrm{Cov}\left[Q_1^{(1)}(t),Q_1^{(1)}(t)^+\right]\\ &\quad + 2\delta\mathrm{Cov}\left[Q_1^{(1)}(t),Q_2^{(1)}(t)\right] + \lambda_t + \mu\left(Q_1^{(0)}(t)\wedge s_t\right) + \delta Q_2^{(0)}(t),\end{aligned} \tag{EC.11}$$

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

EC.2. 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 EC.3a 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 EC.3b, the two are clearly very close to each other.

Figure EC.3 **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

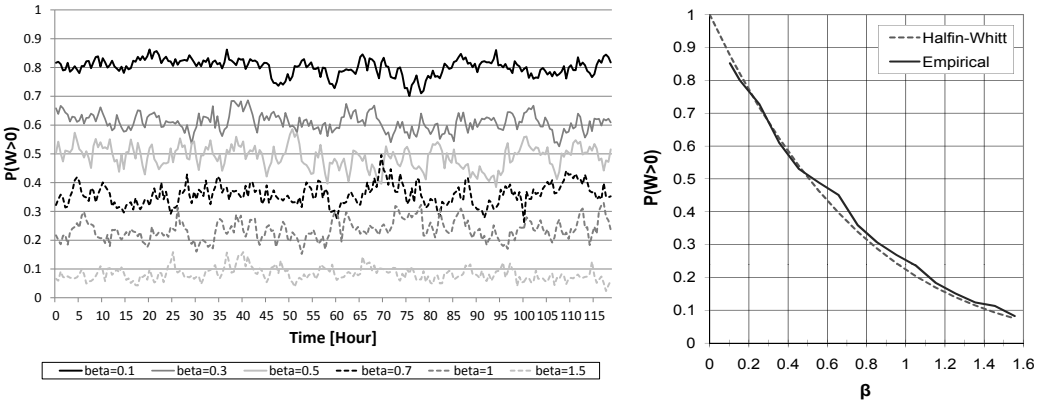


Figure EC.4 **Case study 1 - Simulation results of server utilization.**

(a) Utilization for various β values (b) Time averages vs. theoretical values

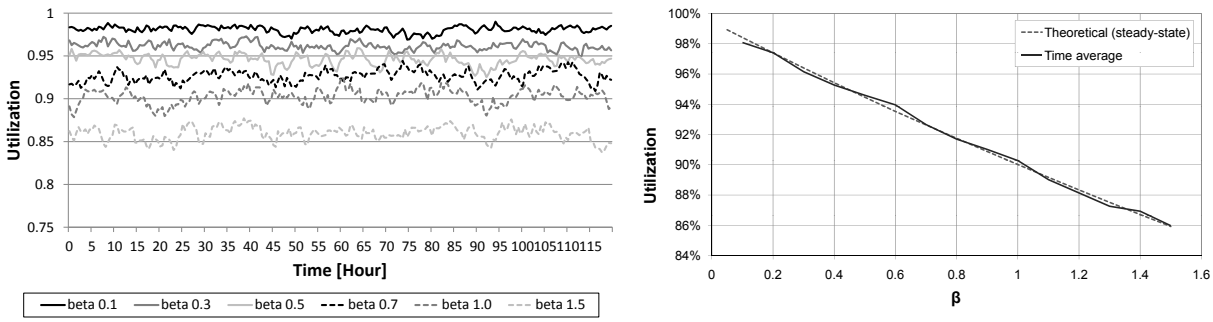


Figure EC.4a shows the evolution of servers utilization over time, for each value of β, which is also stable. Thus our staffing procedure stabilizes *both* service level and server utilization. In Figure EC.4b, we compare the average utilization over time with the theoretical values. The latter were 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 EC.5 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 EC.6 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

Figure EC.5 **Case study 1 - Simulation results of $E[W_t]$ for various β values in large systems.**

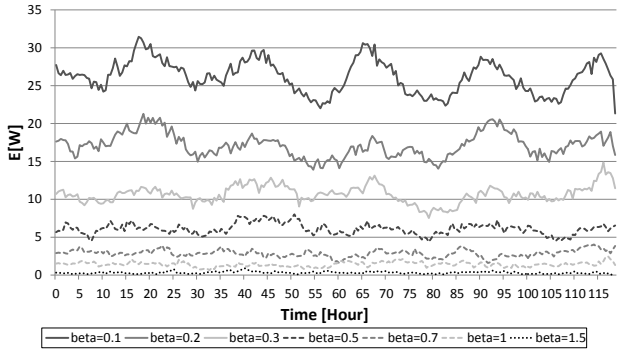
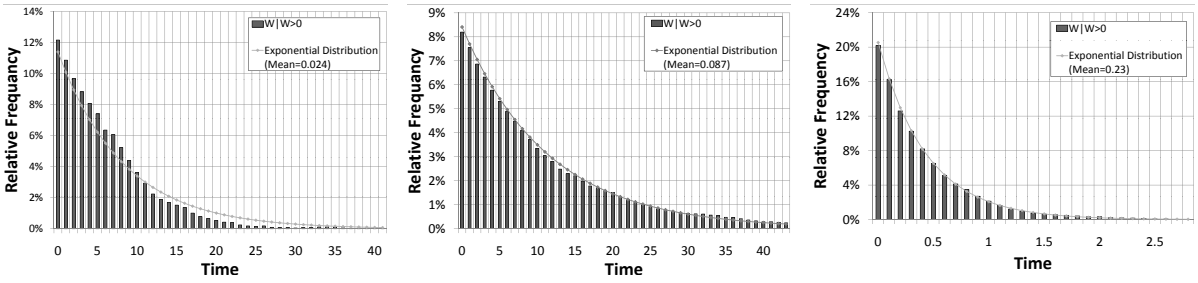


Figure EC.6 **Case study 1 - A comparison of the histogram of $W|W>0$ with the corresponding theoretical distribution.**

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



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.

EC.3. 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.

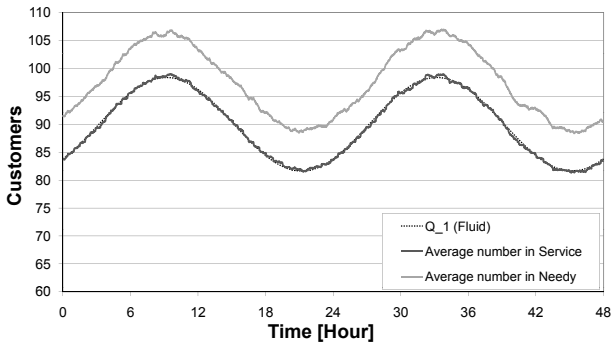
To understand the interpretation problem, we use the following example from Case Study 1. Figure EC.7a shows the fluid solution of the process $Q_1^{(0)}(\cdot)$ (the number of Needy customers), as well as the following simulation results: the average number of customers in the Needy state, and the average 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 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 EC.7b 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 EC.8.

Figure EC.7 $Q_1(t)$ - **Fluid approximation vs. simulation results under QED staffing, for various β's.**

(a) Fluid vs. Simulation



(b) Corrected approximation vs. simulation

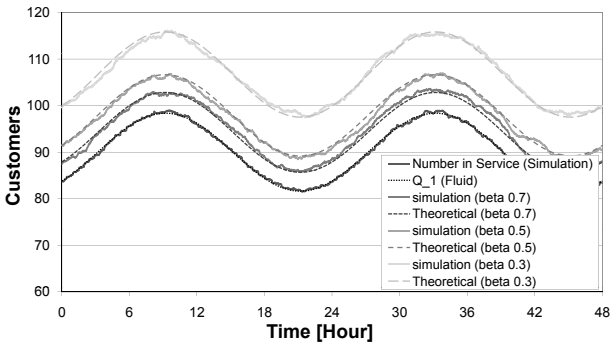


Figure EC.8 $E[W_t]$ - **Corrected Fluid approximation vs. simulation for various β's.**