The Erlang-R Queue: A Model Supporting Personnel Staffing in Emergency Departments

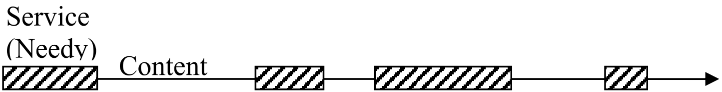
Galit Yom-Tov Avishai Mandelbaum

Industrial Engineering and Management
Technion

WITOR, September 2009

The Problem Studied

Standard assumption in service models: service time is continuous.
But we find systems in which: service is dis-continuous and customers re-enter service again and again.



Examples:

- Machine-Repairman (closed queueing network).
- Medical Unit model (semi-open queueing network).
- Emergency Department (open queueing network).

What is the appropriate staffing procedure? What is the significance of these cycles? Can one still use simple Erlang-C models for staffing?

Related Work

- Mandelbaum A., Massey W.A., Reiman M.
  Strong Approximations for Markovian Service Networks. 1998.
- Massey W.A., Whitt W.
  Networks of Infinite-Server Queues with Nonstationary Poisson Input. 1993.
- Green L., Kolesar P.J., Soares J.
  Improving the SIPP Approach for Staffing Service Systems that have Cyclic Demands. 2001.
- Jennings O.B., Mandelbaum A., Massey W.A., Whitt W.
  Server Staffing to Meet Time-Varying Demand. 1996.
- Feldman Z., Mandelbaum A., Massey W.A., Whitt W.
  Staffing of Time-Varying Queues to Achieve Time-Stable Performance. 2007.

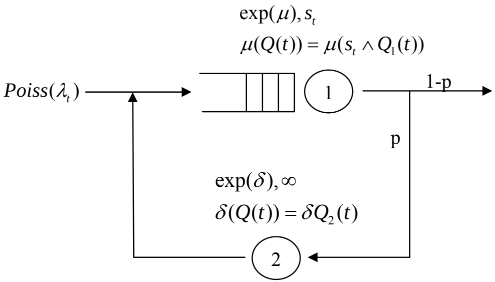
Research Outline

- Develop a queueing model that incorporates service and customers' **R**e-entrance: Erlang-**R**.
- Develop fluid and diffusion approximations.
- Develop staffing algorithm that attain pre-specified service levels.
- Validate our approach via simulation, based on hospital data.
- Understand when Erlang-R is needed.

Model Definition

The (Time-Varying) Erlang-R Queue:

- λ_t - Arrival rate of a time-varying Poisson arrival process.
- μ - Exponential service rate.
- δ - Delay rate ($1/\delta$ is the delay time between services).
- p - Probability of return to service.
- s_t - Number of servers at time t.
- $Q_i(t)$ - Number of customers in node i at time t, $i = 1, 2$.



Fluid and Diffusion Approximations for Erlang-R

We scale by η: $\lambda_t \to \eta\lambda_t$ and $s_t \to \eta s_t$.

Theorem: Fluid (FSSLN) and Diffusion (FCLT) Approximations

As $\eta \to \infty$,

$$\frac{Q^{\eta}(t)}{\eta} \to Q^{(0)}(t), \quad \textit{and} \quad \sqrt{\eta}\left[\frac{Q^{\eta}(t)}{\eta} - Q^{(0)}(t)\right] \stackrel{d}{=} Q^{(1)}(t),$$

where $Q^{(0)}(t)$ is the solution of the following ODE:

$$Q_1^{(0)}(t) = Q_1^{(0)}(0) + \int_0^t \left(\lambda_\tau - \mu_\tau \left(Q_1^{(0)}(\tau) \wedge s_\tau\right) + \delta_\tau Q_2^{(0)}(\tau)\right) d\tau$$

$$Q_2^{(0)}(t) = Q_2^{(0)}(0) + \int_0^t \left(p\mu_\tau \left(Q_1^{(0)}(\tau) \wedge s_\tau\right) - \delta_\tau Q_2^{(0)}(\tau)\right) d\tau,$$

and $Q^{(1)}(t)$ is the solution of a SDE (Stochastic Differential Equation).

Staffing: Determine s_t, $t \geq 0$

- Based on the QED-staffing formula:

$$s = R + \beta\sqrt{R}$$

- Two approaches:
  - PSA / SIPP (lag-SIPP) - divide the time-horizon to planning intervals, calculate average arrival rate and steady-state offered-load for each interval, then staff according to steady-state recommendation (i.e., $R(t) \approx \bar{\lambda}(t)E[S]$).
  - MOL/IS - assuming no constraints on number of servers, calculate time-varying offered load. For example in Erlang-C: $R(t) = E[\int_{t-S}^{t} \lambda(u)du] = E[\lambda(t - S_e)]E[S]$.

Staff according to the square-root formula where $R(t)$ replaces R: $s(t) = R(t) + \beta\sqrt{R(t)}$.

The Offered-Load

Offered-Load in Erlang-R queue = The number of busy servers (or the number of customers) in a corresponding $(M_t/M/\infty)^2$ network. $R(t)$ is denoted by the following two expressions:

$$R_1(t) = E[S_1]E\left[\lambda_1^+(t - S_{1,e})\right] = E\left[\int_{t-S_1}^{t} \lambda_u + \delta Q_2^{(0),\infty}(u)du\right]$$

$$R_2(t) = E[S_2]E\left[\lambda_2^+(t - S_{2,e})\right] = E\left[\int_{t-S_2}^{t} p\mu Q_1^{(0),\infty}(u)du\right]$$

where $Q^{(0),\infty}(t)$ is the solution of the following Fluid ODE:

$$\frac{d}{dt}Q_1^{(0),\infty}(t) = \lambda_t + \delta Q_2^{(0),\infty}(t) - \mu Q_1^{(0),\infty}(t),$$
$$\frac{d}{dt}Q_2^{(0),\infty}(t) = p\mu Q_1^{(0),\infty}(t) - \delta Q_2^{(0),\infty}(t).$$

Staffing: Determine s_t, $t \geq 0$

MOL Algorithm for Erlang-R:

- Solve differential equations for $Q^{\infty}(t)$.
- Calculate time-varying offered load - $R(t)$.
- Staff according to square-root formula: $s(t) = R(t) + \beta\sqrt{R(t)}$, where β is chosen according to the steady-state Halfin-Whitt formula.

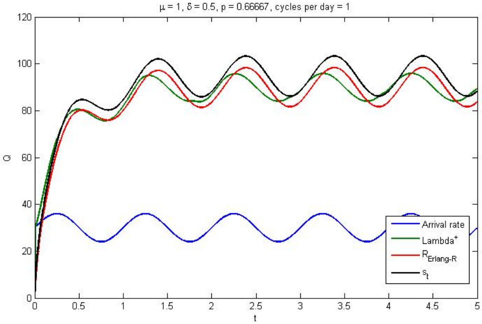
Case Study: Sinusoidal Arrival Rate

Examine the following periodic arrival rate:

$$\lambda_t = \bar{\lambda} + \bar{\lambda}\kappa sin(2\pi t/\psi) = \bar{\lambda} + \bar{\lambda}\kappa sin(\omega t)$$

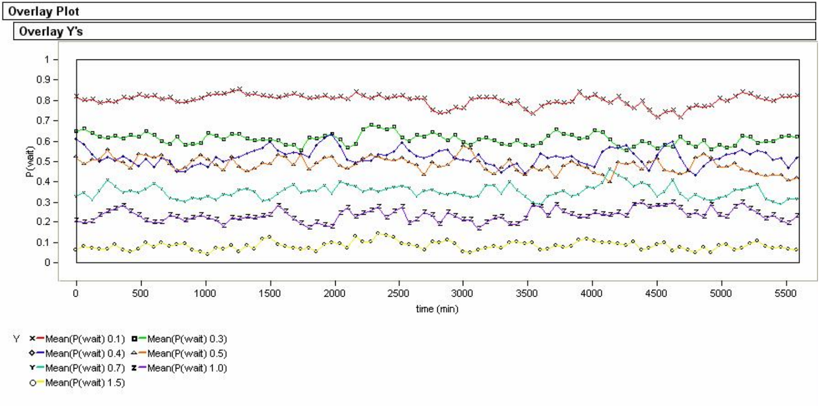
where $\bar{\lambda}$ is the average arrival rate, κ is the relative amplitude $(0 < \kappa < 1)$, and ψ is the period length (ω is the frequency).

Arrival rate, Offered load, and Staffing



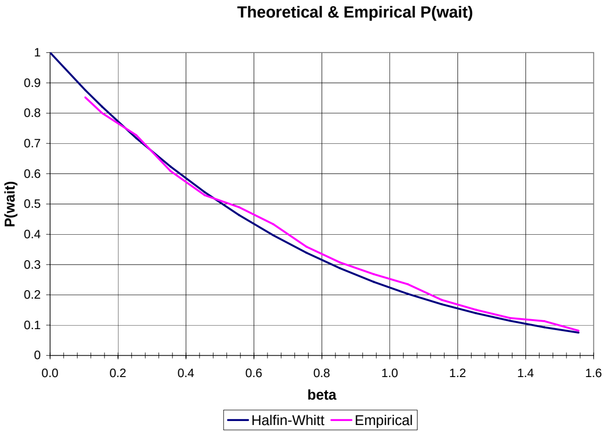
Case Study: Sinusoidal Arrival Rate

Simulation results of P(wait) for various β values



The performance measure is stable!

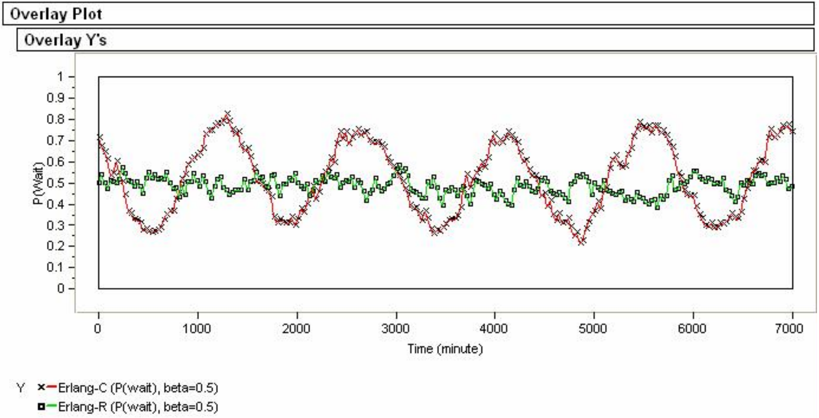
Case Study: Sinusoidal Arrival Rate



The relation between P(wait) and β fits the Halfin-Whitt formula

Can We Use Erlang-C?

Simulation results of P(wait): Erlang-C vs. Erlang-R



Using Erlang-C's R(t), does not stabilize systems' performance.

Why Erlang-C Does Not Fit Re-entrant Systems?

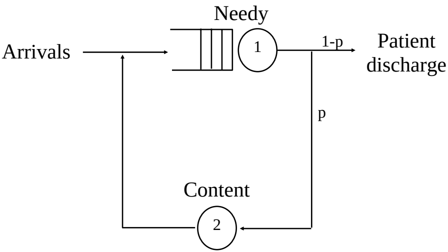
Compare R(t) of Erlang-C and Erlang-R:

Multi-service Erlang-C's offered load: $S_1 \leftrightarrow aS_1$;

$$R(t) = E\left[\lambda(t - aS_{1,e})\right] E\left[aS_1\right]$$

Erlang-R's offered load:

$$R_1(t) = ... = \sum_{i=1}^{\infty} p^i E[S_1] E_{(i+1)S_{1,e}} \left[E_{iS_{2,e}} \left[\lambda(t - (i+1)S_{1,e} - iS_{2,e}) \right] \right]$$

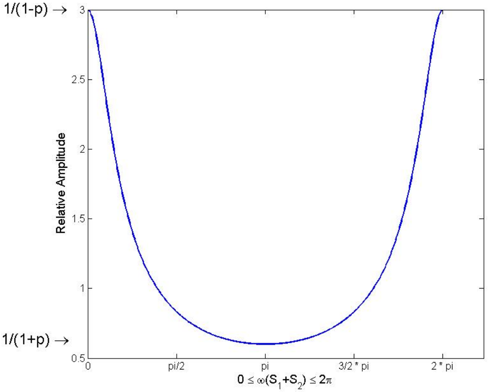


Deterministic Service Times

If S_1, S_2 are deterministic service time then the amplitude of $R(t)$ is given by

$$\left|\frac{e^{i\omega t S_1}}{1 - pe^{-i\omega t(S_1+S_2)}}\right|$$

Plot of relative amplitude of $R_1(t)$ with respect to ω

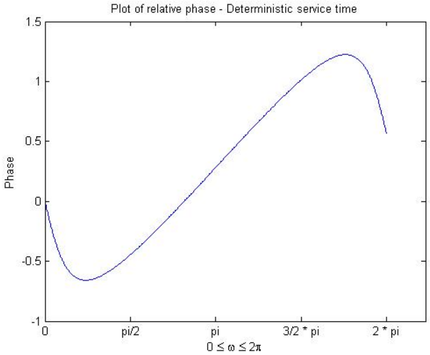


Deterministic Service Times

If S_1, S_2 are deterministic service time then the phase shift of $R(t)$ with respect to the entrance arrival rate is given by

$$Phase(R(t)) = Angle\left(\frac{e^{i\omega t S_1}}{1 - pe^{-i\omega t(S_1+S_2)}}\right)$$

Plot of relative phase of $R_1(t)$ with respect to ω



Exponential Service Times

Theorem:

If $S_1 \sim exp(\mu)$ and $S_1 \sim exp(\delta)$ then:
(1) The amplitude of $R(t)$ is given by

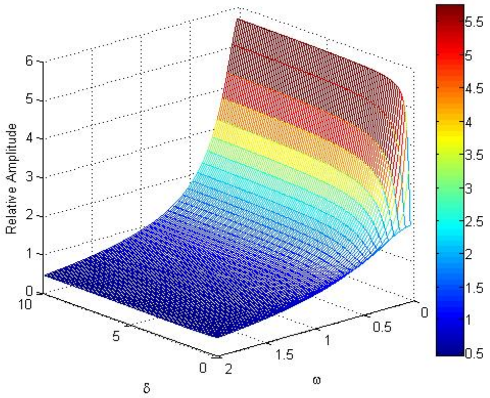
$$Amp(R(t)) = \bar{\lambda}\kappa\sqrt{\frac{(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} \cdot \frac{(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}}$$

and,
(2)the phase shift of $R(t)$ with respect to the entrance arrival rate is given by

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

Exponential Service Times

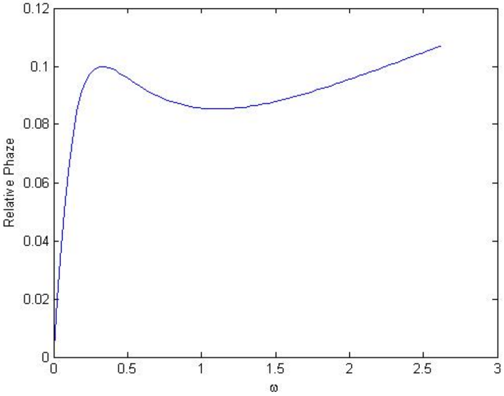
Plot of relative amplitude of $R(t)$ with respect to δ and ω



Relative amplitude of R(t) is a decreasing function of ω (from 1/(1-p) to 0) and a decreasing function of the Content time.

Exponential Service Times

Plot of relative phase of $R(t)$ with respect to ω

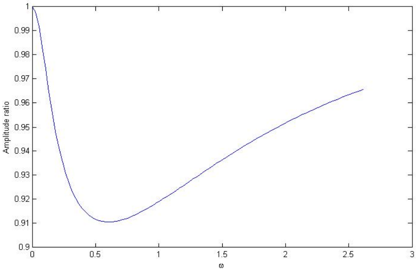


Comparison between Erlang-C to Erlang-R

The ratio of amplitudes between Erlang-R and Erlang-C is given by

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

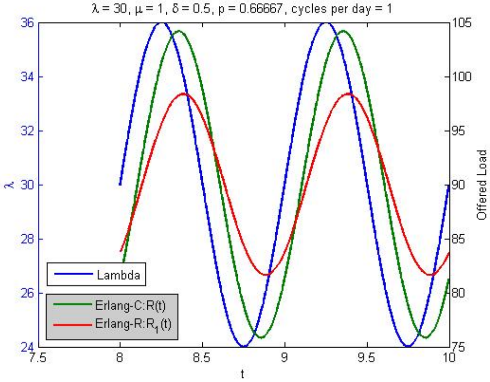
Plot of the ratio of amplitudes as a function of ω



Erlang-C over-estimate the amplitude of the offered-load.

Comparison between Erlang-C to Erlang-R

Erlang-C under- or over-estimates the offered load.

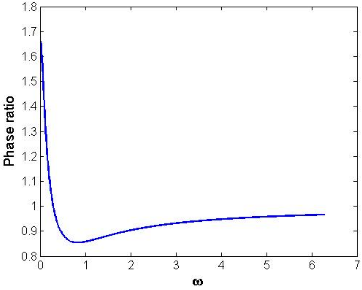


Comparison between Erlang-C to Erlang-R

The ratio between the phase shifts of Erlang-R and Erlang-C is given by

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

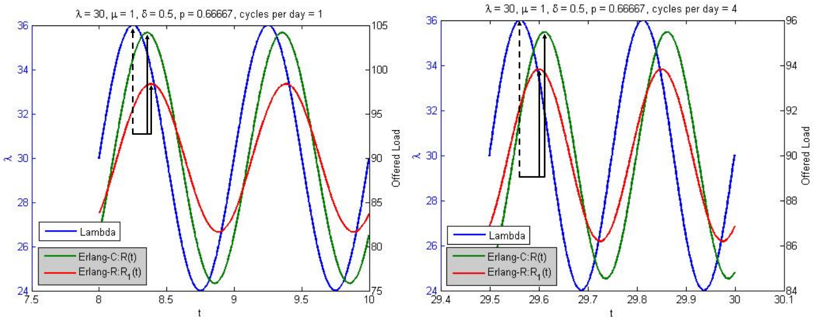
Plot of the ratio of amplitudes as a function of ω



Erlang-C under- or over-estimates the time-lag.

Comparison between Erlang-C to Erlang-R

Erlang-C under- or over-estimates this time-lag depending on the period's length.



Conclusion

In time-varying systems where patients return for multiple services:

1. Using the MOL algorithm for staffing stabilize system performance.
2. The re-entrant patients stabilize the system.
3. If Needy and Content times are deterministic, special care is needed in optimizing the system.
4. Using non re-entering model such as Erlang-C is problematic in most cases.

Thank You