Queues in Hospitals: Queueing Networks with ReEntrant Customers in the QED Regime

Galit Yom-Tov
Adviser: Prof. Avishai Mandelbaum

Industrial Engineering and Management
Technion

April 2010
Motivation

Why Healthcare?

- Total health expenditure as percentage of gross domestic product: Israel 8%, EU 10%, USA 14%.
- Human resource constitute 70% of hospital expenditure.
- Service-level effects health.

ReEntrant customers?

- Cycles of visits: Oncology Wards.
- Lack of Information: Radiology reviewing process.
Motivation

Why Healthcare?

- Total health expenditure as percentage of gross domestic product: Israel 8%, EU 10%, USA 14%.
- Human resource constitute 70% of hospital expenditure.
- Service-level effects health.

ReEntrant customers?

- Cycles of visits: Oncology Wards.
- Lack of Information: Radiology reviewing process.
The Problem Studied: Capacity Problem in Hospitals

Problems in Emergency Wards:

- Hospitals do not manage patients’ flow.
- Long waiting times in the EW for physicians, nurses, and tests.
  => Deterioration in medical state.
- Patients leave EW without being seen or abandon during process.
  => Patient return in severe state.

Service Engineering aim at reducing these effects.
The Problem Studied: Capacity Problem in Hospitals

Problems in **Emergency Wards**:

- Hospitals do not manage patients’ flow.
- Long waiting times in the EW for physicians, nurses, and tests.
  => Deterioration in medical state.
- Patients leave EW without being seen or abandon during process.
  => Patient return in severe state.

**Service Engineering** aim at reducing these effects.
Part I:

The Erlang-R Queue:

Time-Varying QED Queues with Reentrant Customers

in Support of Personnel Staffing
Can we determine the **number of physicians (and nurses)** needed to improve patients flow, and control the system in balance between service quality and efficiency?
Standard assumption in service models: service is continuously provided.

But we find systems in which: service is discontinuous and customers reenter service again and again.

What is the appropriate staffing procedure?
What is the significance of the reentering customers?
What is the implication of using simple Erlang-C models for staffing?
Standard assumption in service models: service is continuously provided. But we find systems in which: service is discontinuous and customers reenter service again and again.

What is the appropriate staffing procedure? What is the significance of the reentering customers? What is the implication of using simple Erlang-C models for staffing?
Related Work

Massey W.A., Whitt W.
*Networks of Infinite-Server Queues with Nonstationary Poisson Input.* 1993.

Feldman Z., Mandelbaum A., Massey W.A., Whitt W.

Jennings O.B., Mandelbaum A., Massey W.A., Whitt W.
*Server Staffing to Meet Time-Varying Demand.* 1996.

Green L., Kolesar P.J., Soares J.
*Improving the SIPI Approach for Staffing Service Systems that have Cyclic Demands.* 2001.
The (Time-Varying) Erlang-R Queue:

- $\lambda_t$ - Arrival rate of a time-varying Poisson process.
- $\mu$ - Service rate.
- $\delta$ - Delay rate ($1/\delta$ - Mean 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 $t$, $i = 1, 2$. 
Staffing: Determine $s_t$, $t \geq 0$

- Based on the QED-staffing formula:

$$s = R + \beta \sqrt{R}, \quad \text{where } R = \lambda E[S]$$

- 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 the time-varying offered-load. For example, in a single service system:

$$R(t) = E[\int_{t-S_e}^t \lambda(u)du] = E[\lambda(t-S_e)]E[S].$$

Staff according to the square-root formula:

$$s(t) = R(t) + \beta \sqrt{R(t)}, \quad \text{and } \beta \text{ is chosen according to the steady-state QED.}$$
Staffing: Determine $s_t, t \geq 0$

Based on the QED-staffing formula:

$$s = R + \beta \sqrt{R}, \quad \text{where } R = \lambda E[S]$$

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 the time-varying offered-load. For example, in a single service system:

$$R(t) = E[\int_{t-S_e}^{t} \lambda(u)du] = E[\lambda(t - S_e)]E[S].$$

Staff according to the square-root formula:

$$s(t) = R(t) + \beta \sqrt{R(t)}, \quad \text{and } \beta \text{ is chosen according to the steady-state QED.}$$
The Offered-Load

Offered-Load $R(t), t \geq 0$, in Erlang-R queue = The number of busy servers (or the number of customers) in a corresponding $(M_t/G/\infty)^2$ network.

$R(t) = (R_1(t), R_2(t))$ is determined by the following expression:

$$R_i(t) = E[\lambda_i^+(t - S_{i,e})]E[S_i]$$

where,

$$\lambda_1^+(t) = \lambda(t) + E[\lambda_2^+(t - S_2)]$$
$$\lambda_2^+(t) = pE[\lambda_1^+(t - S_1)]$$

When service times are exponential, $R(t)$ is the solution of the following Fluid ODE:

$$\frac{d}{dt} R_1(t) = \lambda_t + \delta R_2(t) - \mu R_1(t),$$
$$\frac{d}{dt} R_2(t) = p\mu R_1(t) - \delta R_2(t).$$
Hospital Arrival Rate

Arrivals and Offered-Load

-- Offered Load (Doctors) -- Arrivals

Galit Yom-Tov  Ph.D. Seminar
Case Study: Sinusoidal Arrival Rate

Periodic arrival rate: \( \lambda_t = \bar{\lambda} + \lambda \kappa \sin(\omega t) \).
\( \bar{\lambda} \) is the average arrival rate, \( \kappa \) is the relative amplitude, and \( \omega \) is the frequency.

External / Internal arrivals rate, Offered-load, and Staffing
Case Study: Sinusoidal Arrival Rate

Simulation of $P(\text{Wait})$ for various $\beta$ ($0.1 \leq \beta \leq 1.5$)

Performance measure is stable! ($0.15 \leq P(\text{Wait}) \leq 0.85$)
Case Study: Sinusoidal Arrival Rate

Theoretical & Empirical $P(\text{Wait})$

Relation between $P(\text{Wait})$ and $\beta$ fits steady-state theory!
Case Study: Sinusoidal Arrival Rate

Simulation of servers utilization for various $\beta$

Performance measure is stable! ($0.85 \leq Util \leq 0.98$)
Can We Use Erlang-C?

Simulation results of $P(\text{Wait})$: Erlang-R vs. Erlang-C and PSA

PSA and Erlang-C $R(t)$ do not stabilize performance.
Why Erlang-C Does Not Fit Re-entrant Systems?

Compare $R(t)$ of Erlang-C vs. Erlang-R:

Erlang-C offered-load (with concatenated services):

$$R(t) = E \left[ \lambda \left( t - \frac{1}{1-p} S_{1,e} \right) \right] E \left[ \frac{1}{1-p} S_1 \right]$$

Erlang-R offered-load:

$$R_1(t) = E \left[ \sum_{i=1}^{\infty} p^i \lambda \left( t - S_{1,i}^* - S_{2,i}^* - S_{1,e} \right) \right] E[S_1]$$
Comparison between Erlang-C and Erlang-R

Amplitudes ratio as a function of $\omega$

Erlang-C over-estimates the amplitude of the offered-load. The re-entrant patients stabilize the system. Minimum ratio achieved when $\omega = \sqrt{\delta \mu (1 - p)}$ (for example Emergency Ward).
Erlang-C under- or over-estimates the Erlang-R offered-load.
Small systems - Hospitals

Constraints:

- Staffing resolution: 1 hour
- Minimal staffing: 1 doctor per type
- Integer values: \( s(t) = \left[ R_1(t) + \beta \sqrt{R_1(t)} \right] \)

Small systems: Number of doctors ranges from 1 to 5

Example: \( R = 2.75 \)

<table>
<thead>
<tr>
<th>( \beta ) range</th>
<th>s</th>
<th>( P(W &gt; 0) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 0.474]</td>
<td>3</td>
<td>82.4%</td>
</tr>
<tr>
<td>(0.474, 1.055]</td>
<td>4</td>
<td>34.0%</td>
</tr>
<tr>
<td>(1.055, 1.658]</td>
<td>5</td>
<td>11.4%</td>
</tr>
<tr>
<td>(1.658, 2.261]</td>
<td>6</td>
<td>3.0%</td>
</tr>
<tr>
<td>1.658 and up</td>
<td>7</td>
<td>0%</td>
</tr>
</tbody>
</table>

=> Can not achieve all performance levels!
Small systems - Hospitals

P(Wait) is stable and separable!
Conclusion of Part I: Erlang-R

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

1. Using the MOL (IS) algorithm for staffing stabilizes performance.
2. Re-entrant patients stabilize the system.
3. Using single-service models, such as Erlang-C, is problematic in the re-entrant ED environment:
   - Time-varying arrivals
   - Transient behavior even with constant parameters
Part II:

The Semi-Open Model:

Semi-Open Queueing Networks with ReEntrant Customers in the QED Regime
Work-Force and Bed Capacity Planning:

1. There are 3M registered nurses in the U.S. but still a chronic shortage.

2. California law sets nurse-to-patient ratios such as 1:6 for pediatric care units.

3. O.B. Jennings and F. de Vericourt (2008) showed that fixed ratios do not account for economies of scale.

4. Management focuses average occupancy levels, while arrivals have seasonal patterns and stochastic variability (Green 2004).
Research Objectives:

1. Medical Unit with $s$ nurses and $n$ beds: semi-open queueing network with statistically identical customers and servers.

2. Questions addressed: How many servers (nurses) are required (**staffing**), and how many fixed resources (beds) are needed (**allocation**) needed?

3. Coping with time-variability
Related Work

Khudyakov P.
*Designing a Call Center with an IVR.* 2006.

Halfin S., Whitt W.
*Heavy-traffic Limits for Queues with many Exponential Servers.* 1981.

Mandelbaum A., Massey W.A., Reiman M.
*Strong Approximations for Markovian Service Networks.* 1998.

**Analytical models in HC:**

Jennings O.B., de Véricourt F.
*Dimensioning Large-Scale Membership Services.* 2008.

Yankovic N., Green L.

**Beds capacity:**

Green L.
The Internal Ward Queueing Network:

Service times are Exponential; Routing is Markovian
The MU Model as a Closed Jackson Network:

=> Product Form - $\pi_N(i, j, k)$ stationary distribution.
Service Level Objectives

- Blocking probability
- Delay probability
- Probability of timely service (wait more than t)
- Expected waiting time
- Average occupancy level of beds
- Average utilization level of nurses

Function of $\lambda$, $\mu$, $\delta$, $\gamma$, $p$, $s$, $n$
QED characteristics

- High service quality
- High resource efficiency
- Square-root staffing rule

\[
s = \frac{\lambda}{(1 - p)\mu} + \beta \sqrt{\frac{\lambda}{(1 - p)\mu}} + o(\sqrt{\lambda}), \quad -\infty < \beta < \infty
\]

\[
n - s = \frac{p\lambda}{(1 - p)\delta} + \frac{\lambda}{\gamma} + \eta \sqrt{\frac{p\lambda}{(1 - p)\delta}} + \frac{\lambda}{\gamma} + o(\sqrt{\lambda}) \quad -\infty < \eta < \infty
\]

where
\[
\frac{\lambda}{(1 - p)\mu}
\]
is the offered-load at service station 1 (Needy).
\[
\frac{p\lambda}{(1 - p)\delta}
\]
is the offered-load at non-service station 2+3 (Content + Cleaning).

- Many-server asymptotics

Galit Yom-Tov  Ph.D. Seminar
Theorem:

Let $\lambda$, $s$ and $n$ tend to $\infty$ simultaneously and satisfy the QED conditions. Then

$$\lim_{\lambda \to \infty} P(W > 0) = \left( 1 + \frac{\int_{-\infty}^{\beta} \Phi \left( \eta + (\beta - t)\sqrt{B} \right) d\Phi(t)}{\phi(\beta)\Phi(\eta_1) - \frac{\phi(\sqrt{\eta_1^2 + \beta})}{\beta} e^{\frac{1}{2} \eta_1^2} \Phi(\eta_1)} \right)^{-1},$$

where $B = \frac{\delta \gamma}{\mu(p \gamma + (1-p) \delta)}$ and $\eta_1 = \eta - \frac{\beta}{\sqrt{B}}$.

The delay probability is a function of three parameters: $\beta$ (servers), $\eta$ (beds), and $B$ (the offered-load-ratio).
Expected Waiting Time

Theorem:

Let $\lambda$, $s$ and $n$ tend to $\infty$ simultaneously and satisfy the QED conditions. Then

$$\lim_{\lambda \to \infty} \sqrt{s}E[W] = \mu \frac{1}{\beta^2} \phi(\beta)\Phi(\eta) + \phi(\sqrt{\eta^2 + \beta^2}) e^{\frac{1}{2} \eta^2} \Phi(\eta_1) \left( B^{-1}\beta^2 - \eta \beta \sqrt{B^{-1}} - 1 \right)$$

where $B = \delta \gamma \frac{\mu(p \gamma + (1-p) \delta)}{\mu(p \gamma + (1-p) \delta)}$ and $\eta_1 = \eta - \frac{\beta}{\sqrt{B}}$.

Waiting time is one order of magnitude less than service time.
Theorem:

Let $\lambda$, $s$ and $n$ tend to $\infty$ simultaneously and satisfy the QED conditions. Then

$$
\lim_{\lambda \to \infty} \sqrt{s}P(\text{block}) = \frac{\nu \phi(\nu_1)\Phi(\nu_2) + \phi(\sqrt{\eta^2 + \beta^2}) e^{\frac{\eta_1^2}{2}} \Phi(\eta_1)}{\int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{B}\right) d\Phi(t) + \frac{\phi(\beta)\phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2 + \beta^2})}{\beta} e^{\frac{1}{2} \eta_1^2} \Phi(\eta_1)}
$$

where $\eta_1 = \eta - \frac{\beta}{\sqrt{B}}$, $\nu_1 = \frac{\eta \sqrt{B^{-1} + \beta}}{\sqrt{1 + B^{-1}}}$, $\nu_2 = \frac{\beta \sqrt{B^{-1} - \eta}}{\sqrt{1 + B^{-1}}}$, $\nu = \frac{1}{\sqrt{1 + B^{-1}}}$.

$P(\text{Block}) \ll P(\text{Wait})$
The influence of $\eta$ and $\beta$

Reminder: $\beta$ sets the number of nurses, $\eta$ sets the number of beds:

\[
\begin{align*}
    s &= \frac{\lambda}{(1-p)\mu} + \beta \sqrt{\frac{\lambda}{(1-p)\mu}} + o(\sqrt{\lambda}), \\
    n - s &= \frac{\rho \lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} + \eta \sqrt{\frac{\rho \lambda}{(1-p)\delta}} + \frac{\lambda}{\gamma} + o(\sqrt{\lambda}),
\end{align*}
\]

$-\infty < \beta < \infty$  

$-\infty < \eta < \infty$
Numerical Example

- N=42 with 78% occupancy
- ALOS = 4.3 days
- Average service time = 15 min
- 0.4 requests per hour
- \( \lambda = 0.32, \mu = 4, \delta = 0.4, \gamma = 4, p = 0.975 \)
- \( \Rightarrow \) Ratio of offered-load = 0.1

Based on Lundgren and Segesten 2001 + Yankovic and Green 2007
How to find the required $\eta$ and $\beta$?

If $\beta = 0.5$ and $\eta = 0.5$ ($s = 4$, $n = 38$): $P(block) \approx 0.07$, $P(wait) \approx 0.4$

If $\beta = 1.5$ and $\eta \approx 0$ ($s = 6$, $n = 37$): $P(block) \approx 0.068$, $P(wait) \approx 0.084$

If $\beta = -0.1$ and $\eta \approx 0$ ($s = 3$, $n = 34$): $P(block) \approx 0.21$, $P(wait) \approx 0.70$
The Time-Varying Semi-Open Erlang-R Model:

- Arrivals
- N beds
- Patient is Needy
- Patient is Content
- Blocked patients
- 1-p
- p
- 1
- 2

Marginal compared to loss system: steady-state distributions differ.

=> Must consider Reentrant customers, even in steady-state!
Steady-State $P(\text{Wait}>0)$ - Semi-Open Erlang-R vs. Loss System
**MOL Algorithm for Semi-Open Erlang-R:**

- Calculate time-varying offered-load of *open* Erlang-R model - \( R(t) \).
- Staff nurses according to square-root formula:
  \[ s(t) = R_1(t) + \beta \sqrt{R_1(t)} \]
- Allocate \( n(t) \) beds according to square-root formula:
  \[ n(t) = s(t) + R_2(t) + \eta \sqrt{R_2(t)} \]

Here \( \beta \) and \( \eta \) are chosen according to the *steady-state* Semi-Open Erlang-R formula.
P(Wait) and P(Blocking) as a function of time, for semi-open Erlang-R:
Comparison to Steady-State Performance

Average $P(\text{Wait})$ and $P(\text{Blocking})$ as a function of $\beta$, for semi-open Erlang-R:
Conclusion of Part II: Semi-Open

- Steady-state QED approximations developed and tested.
- System performance is governed by the offered-load ratio.
- ReEntrant customers play important role in steady-state distribution, as well in transient time.
- MOL (IS) staffing stabilizes performance in time-varying semi-open networks.
Empirical Analysis of Patients Flow
Goals

1. Measurements / Data
2. Modeling, Analysis
3. Validation

Science

4. Maturity enables Deployment
5. Implementation
6. Improvement
7. Feedback
8. Novel needs, necessitating Science
EW’s vs. IW’s Arrival Rate

Emergency Ward

Average Arrival Rate per Hour, ED (Int), 2004-2008

Rate

Day of Week, Time

Internal Ward

Average Arrival Rate per Hour, Ward A, 2004-2008

Rate

Day of Week, Time

Have the same structure!

Time-lag between ED and IW: due to the ED LOS.
IW Departure Rate

Hourly Departure Rate, Ward A, 2008

=> The operational impact of release policy in IWs.
Arrival Rate, Departure Rate, and Number of Patients by Day and Hour, Ward A, 2004-2008

Number of patients changes dramatically over the day. Operational effect: The effect of flux in time-varying queues.
In most loaded hours - least nurses recommendation.
LOS is a mixture of Normally distributed Random variables.
Galit Yom-Tov
Ph.D. Seminar


LOS and Load

LOS as a function of Load

Number of Patients

LOS [days]

IW A
IW C
IW D
Trend in arrival rate explains part of the blocking. Beds capacity was reduced from 202 beds to 185. During 2006 War no blocking. Change in blocking policy.
## Returns to hospital

<table>
<thead>
<tr>
<th>Ward</th>
<th>No. of returns per patient (in 4 years)</th>
<th>Time between returns (days)</th>
<th>Probability of return within 3 month</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>1.76</td>
<td>208</td>
<td>22%</td>
</tr>
<tr>
<td>Oncology</td>
<td>5.76</td>
<td>29</td>
<td>76%</td>
</tr>
</tbody>
</table>
Conclusion

Queueing theory provides tools to model Healthcare operations. Data analysis provides the means to:

- Implement models.
- Characterize environments where the models are applicable.
- Identify where these models are bound to fail and need adjustments.
- Better understand the system and discover new phenomena.
Thank You