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 treatment: Emergency Wards, Internal Wards.
- 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 treatment: Emergency Wards, Internal Wards.
- 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.

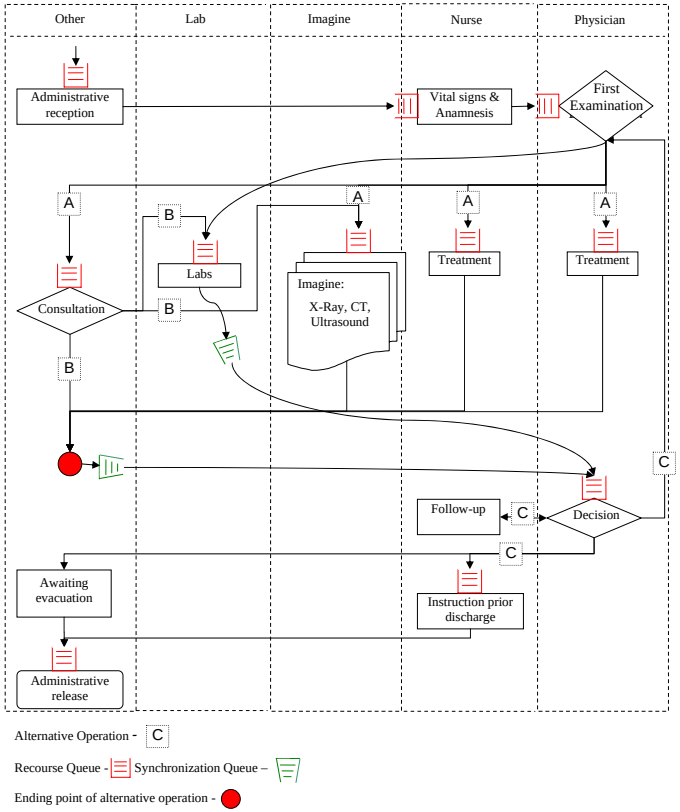
Service Engineering aim at reducing these effects.

Part I:

The Erlang-R Queue:

Time-Varying QED Queues with Reentrant Customers in Support of Personnel Staffing

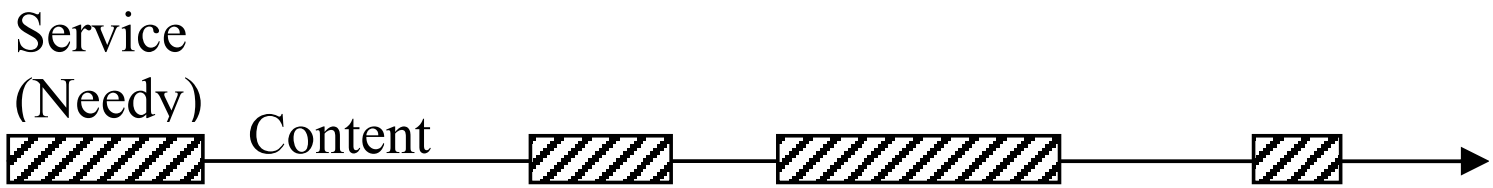
The Problem Studied



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?

The Problem Studied

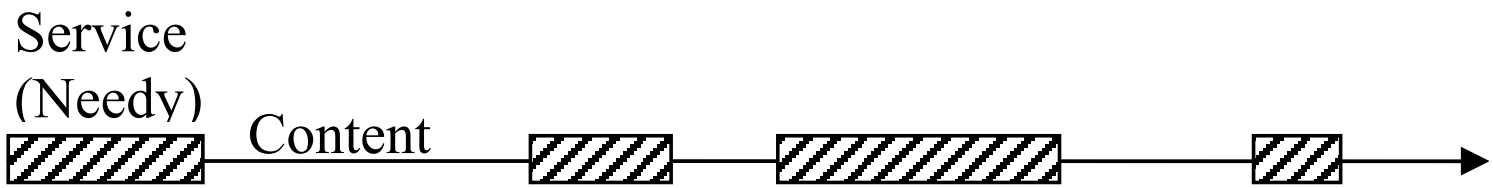
Standard assumption in service models: service is continuously provided.
But we find systems in which: service is discontinuous and customers reenter service again and again.



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?

The Problem Studied

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.

Staffing of Time-Varying Queues to Achieve Time-Stable Performance. 2007.

Jennings O.B., Mandelbaum A., Massey W.A., Whitt W.

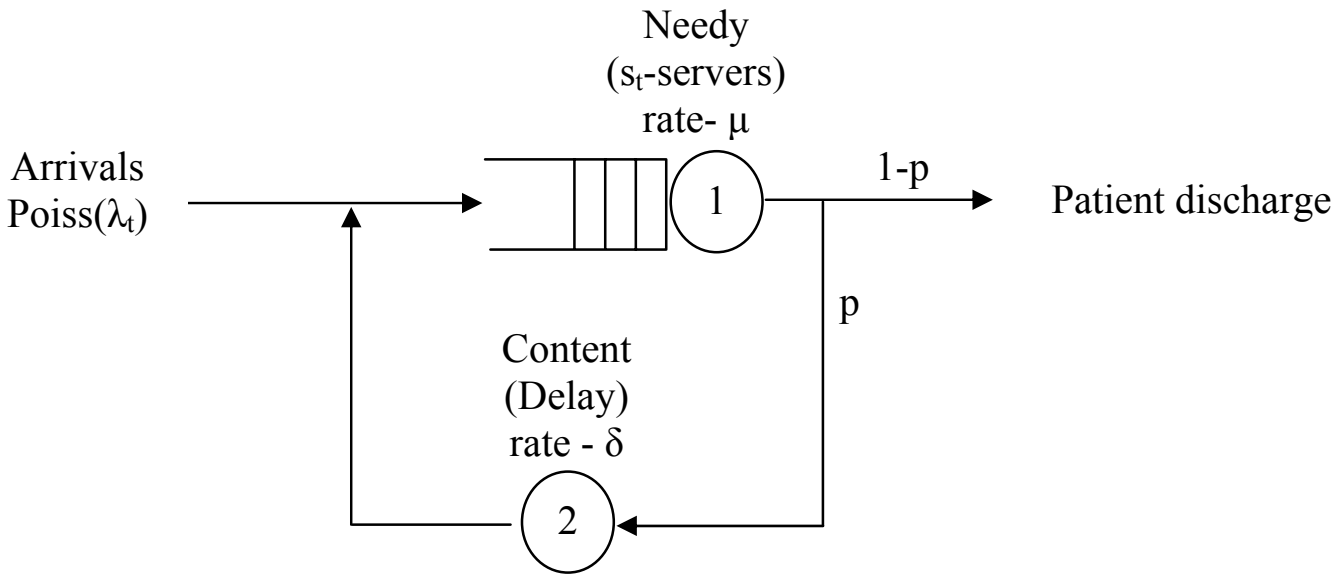
Server Staffing to Meet Time-Varying Demand. 1996.

Green L., Kolesar P.J., Soares J.

Improving the SIPP Approach for Staffing Service Systems that have Cyclic Demands. 2001.

Model Definition

The (Time-Varying) Erlang-R Queue:



- λ_t - Arrival rate of a time-varying Poisson process.
- μ - Service rate.
- δ - 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:

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

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

Staff according to the square-root formula:
$s(t) = R(t) + \beta\sqrt{R(t)}$, and β is chosen according to the steady-state QED.

Staffing: Determine s_t, $t \geq 0$

- Based on the QED-staffing formula:

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

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

Staff according to the square-root formula:
$s(t) = R(t) + \beta\sqrt{R(t)}$, and β 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:

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

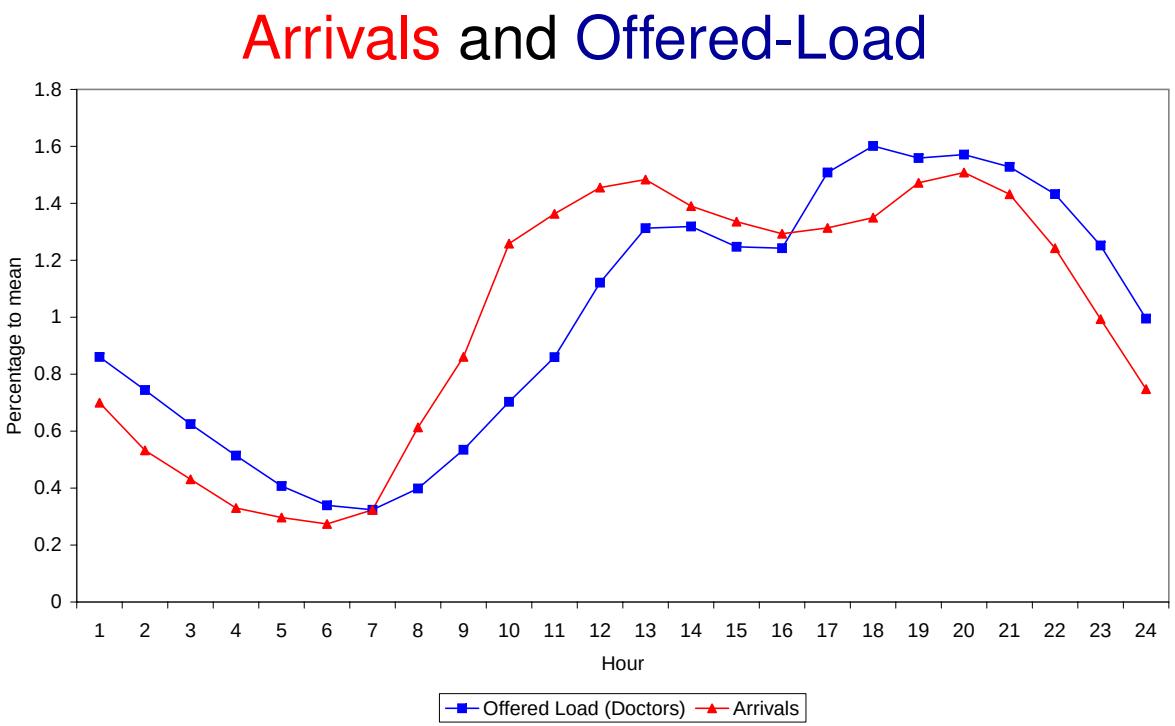
where,

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

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

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

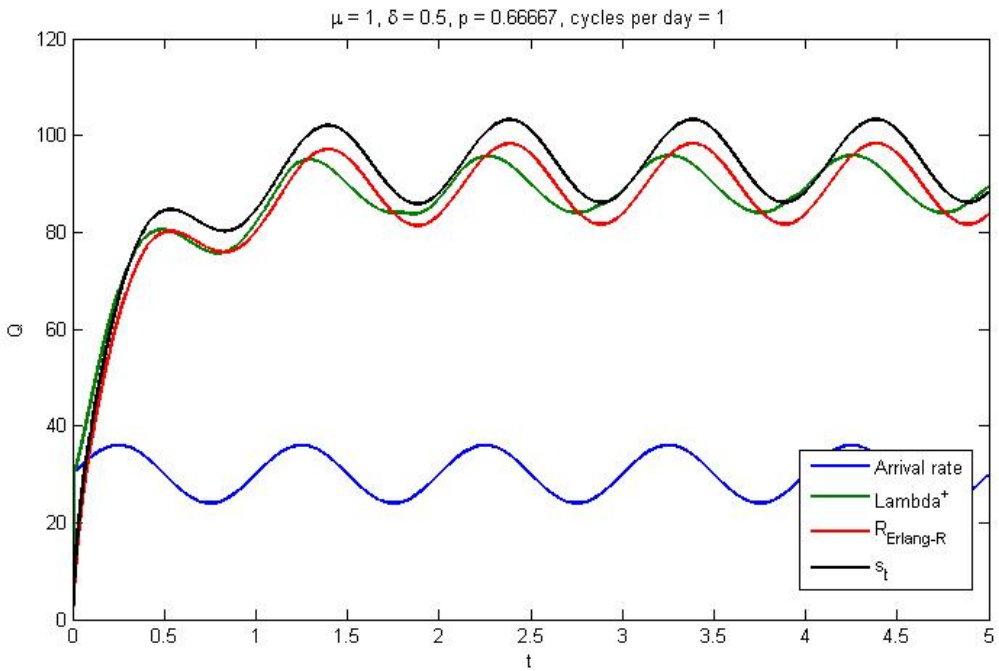
Hospital Arrival Rate



Case Study: Sinusoidal Arrival Rate

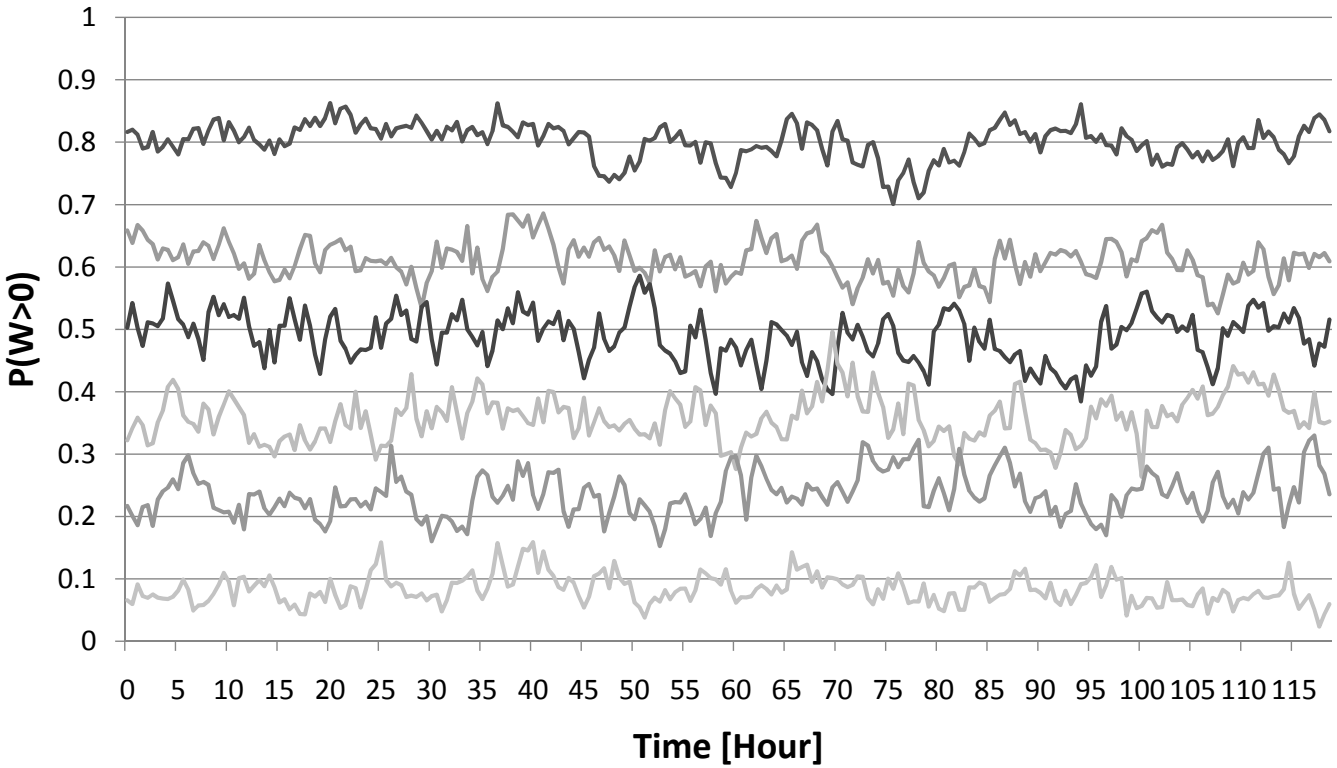
Periodic arrival rate: $\lambda_t = \bar{\lambda} + \bar{\lambda}\kappa sin(\omega t)$.
$\bar{\lambda}$ is the average arrival rate, κ is the relative amplitude, and ω is the frequency.

External / Internal arrivals rate, Offered-load, and Staffing



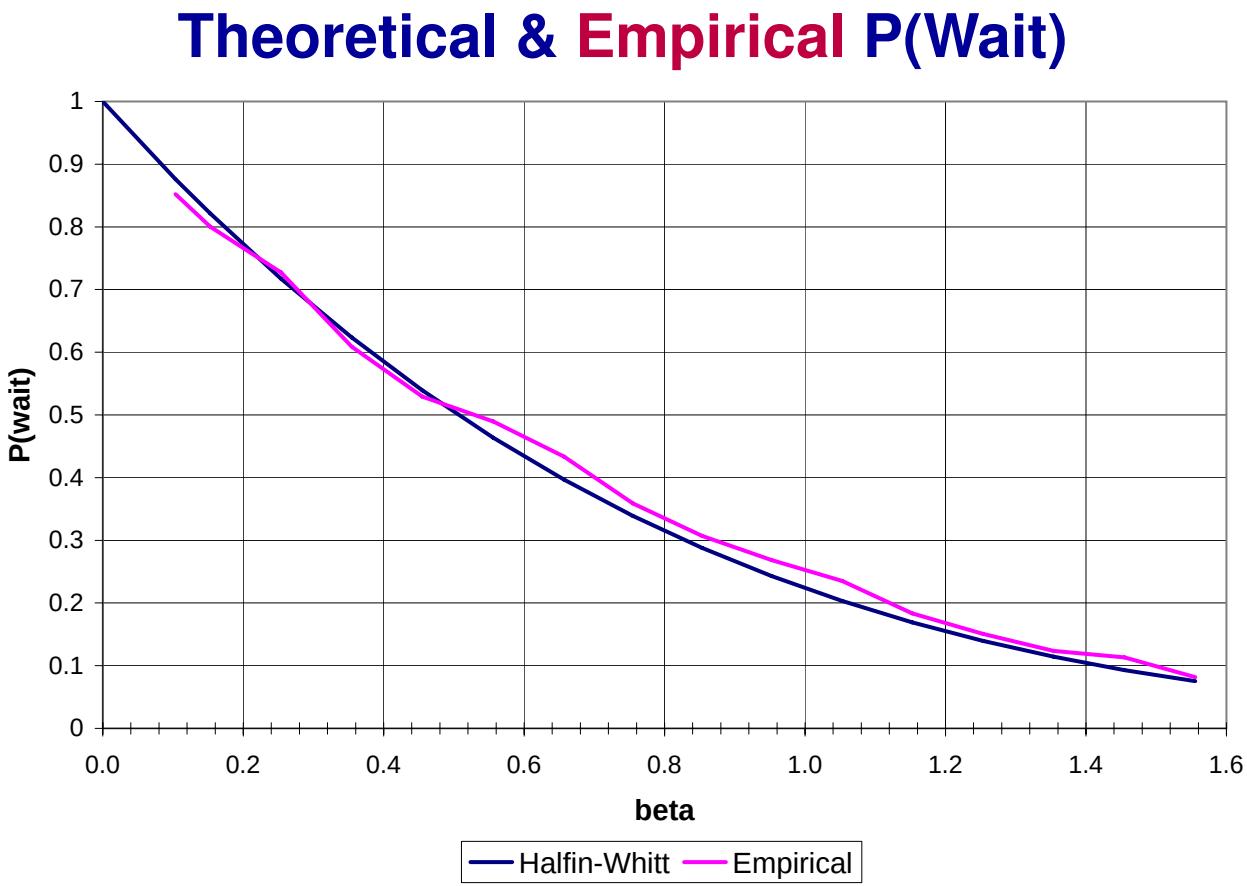
Case Study: Sinusoidal Arrival Rate

Simulation of P(Wait) for various β $(0.1 \leq \beta \leq 1.5)$



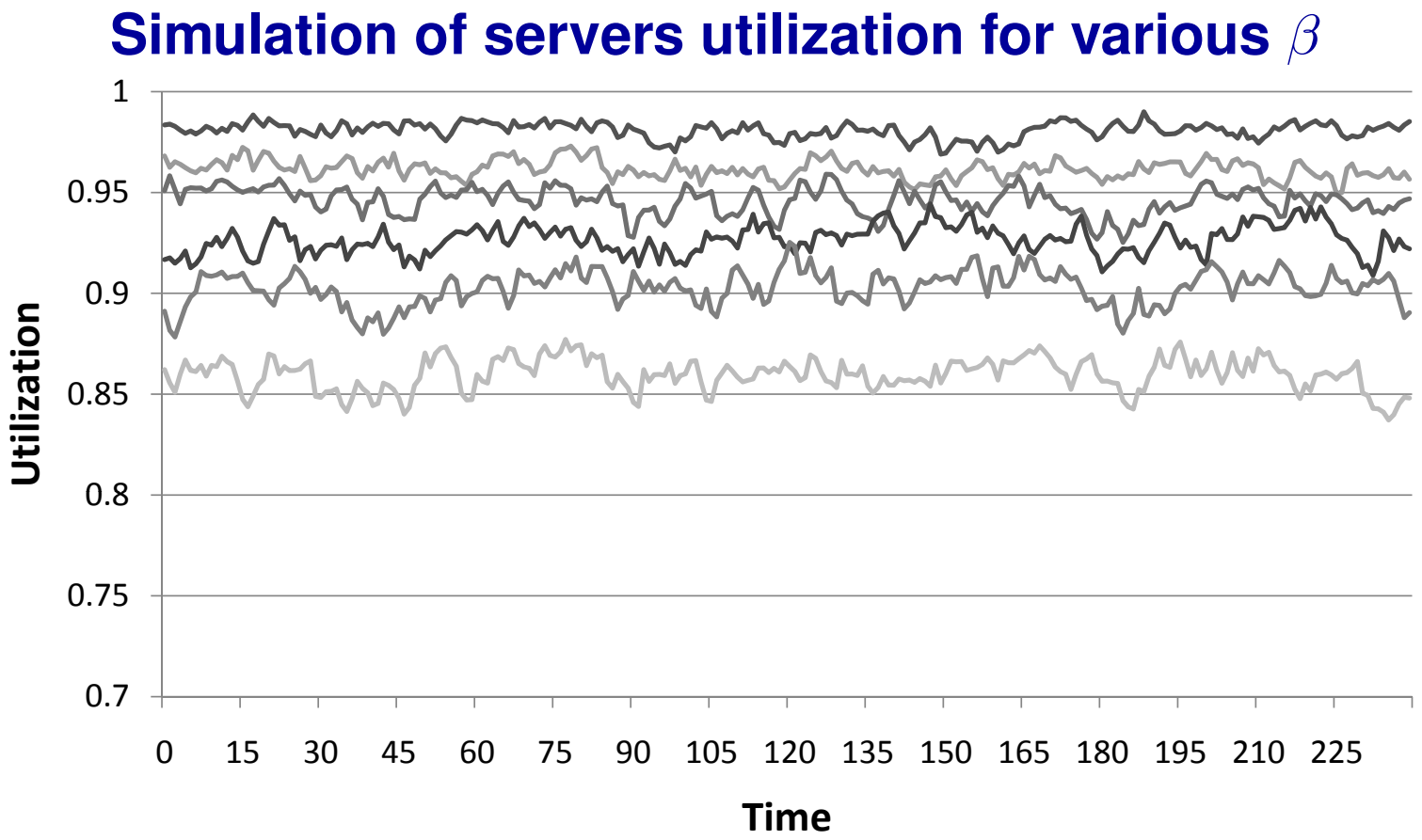
Performance measure is stable! $(0.15 \leq P(Wait) \leq 0.85)$

Case Study: Sinusoidal Arrival Rate



Relation between P(Wait) and β fits steady-state theory!

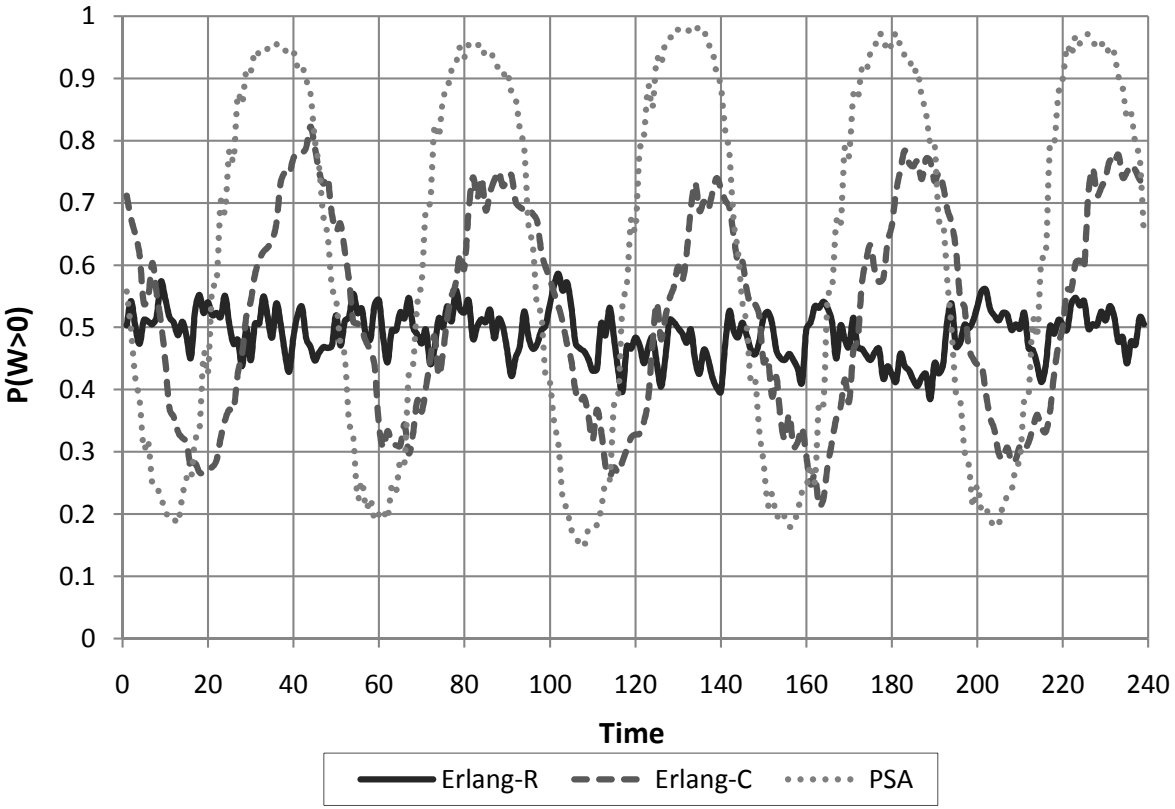
Case Study: Sinusoidal Arrival Rate



Performance measure is stable! $(0.85 \leq Util \leq 0.98)$

Can We Use Erlang-C?

Simulation results of P(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?

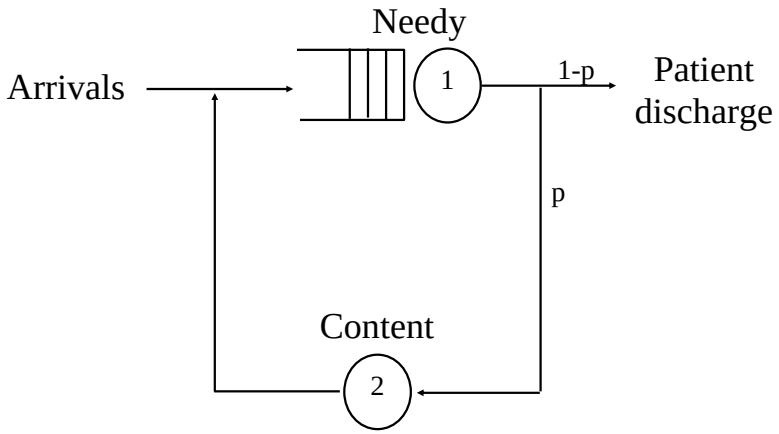
Compare R(t) of Erlang-C vs. Erlang-R:

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

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

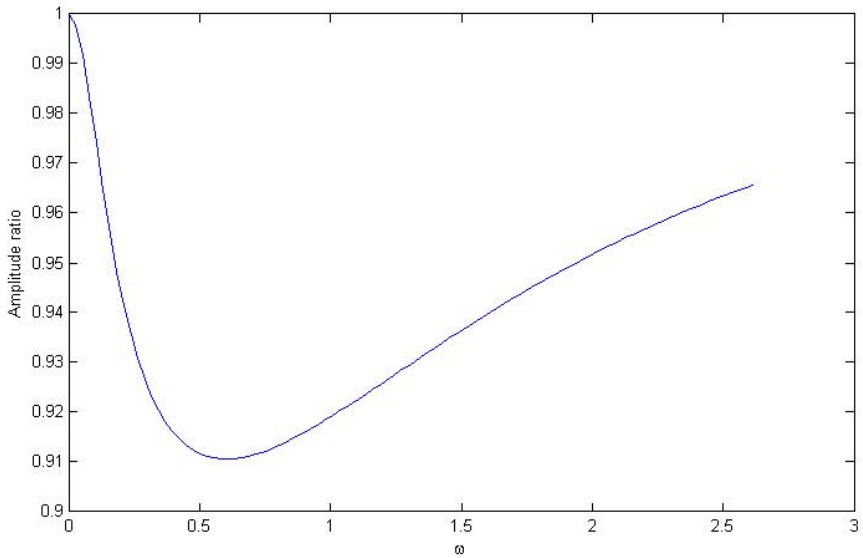
Erlang-R offered-load:

$$R_1(t) = E\left[\sum_{i=1}^{\infty} p^i \lambda\left(t - S_1^{*i} - S_2^{*i} - S_{1,e}\right)\right] E[S_1]$$



Comparison between Erlang-C and Erlang-R

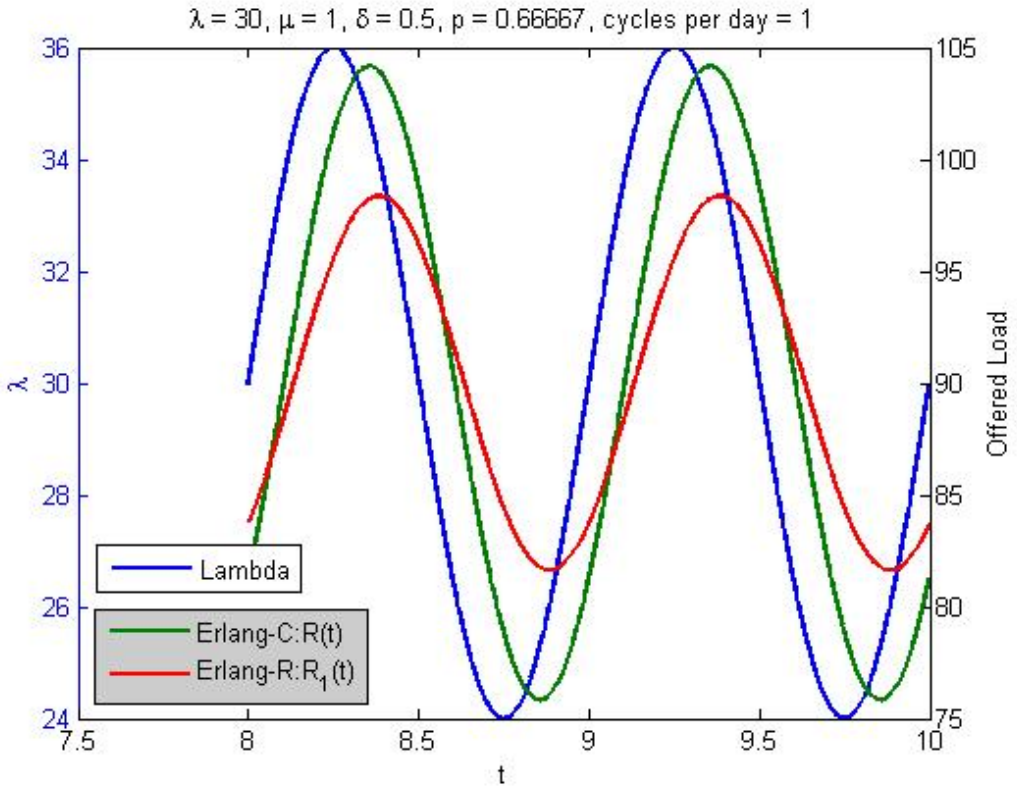
Amplitudes ratio as a function of ω



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).

Comparison between Erlang-C and Erlang-R

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: $s(t) = [R_1(t) + \beta\sqrt{R_1(t)}]$

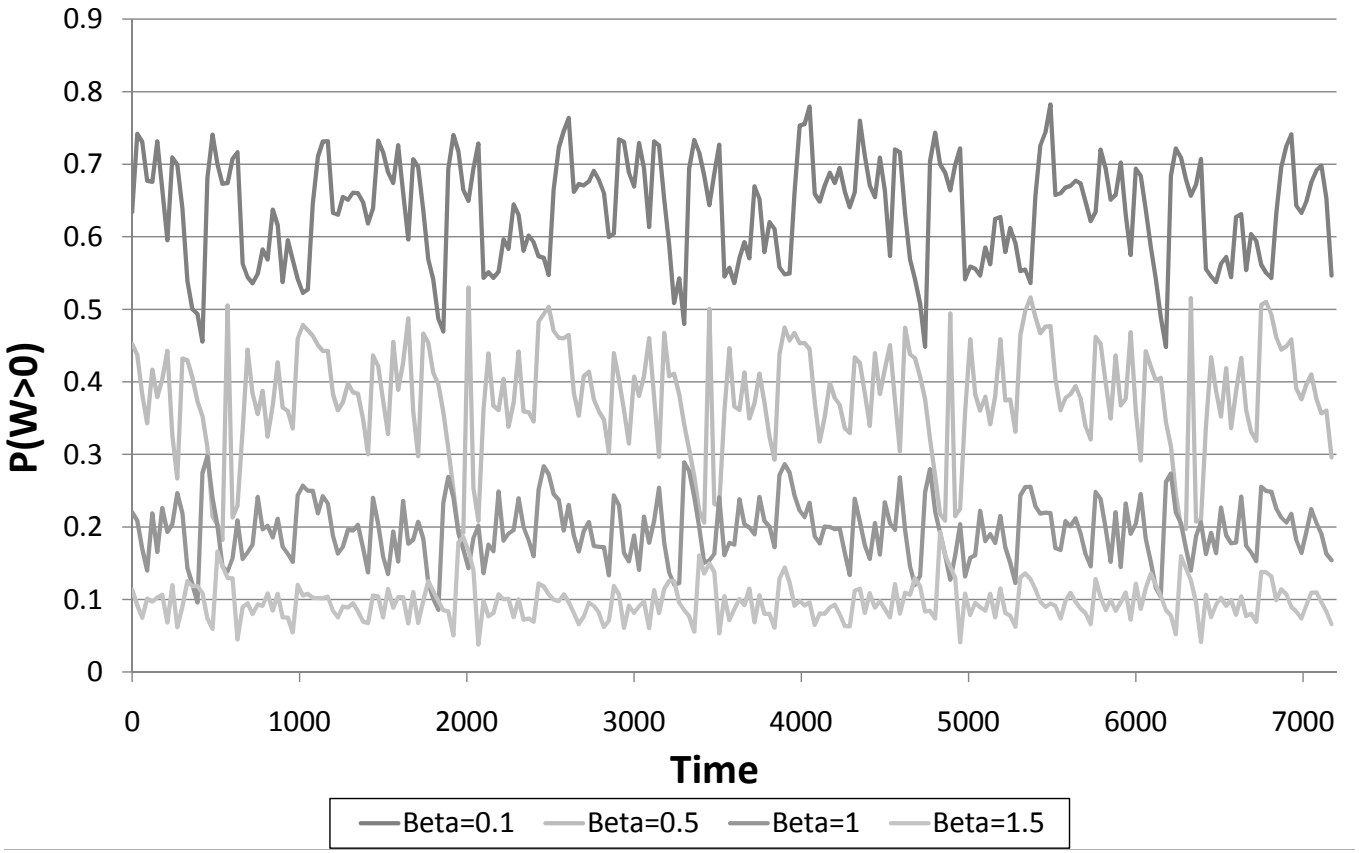
Small systems: Number of doctors ranges from 1 to 5

Example: $R = 2.75$

β range	s	$P(W > 0)$
$[0, 0.474]$	3	82.4%
$(0.474, 1.055]$	4	34.0%
$(1.055, 1.658]$	5	11.4%
$(1.658, 2.261]$	6	3.0%
1.658 and up	7	0%

=> 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

Motivation

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

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.

A Queueing Model for Nurse Staffing. 2007.

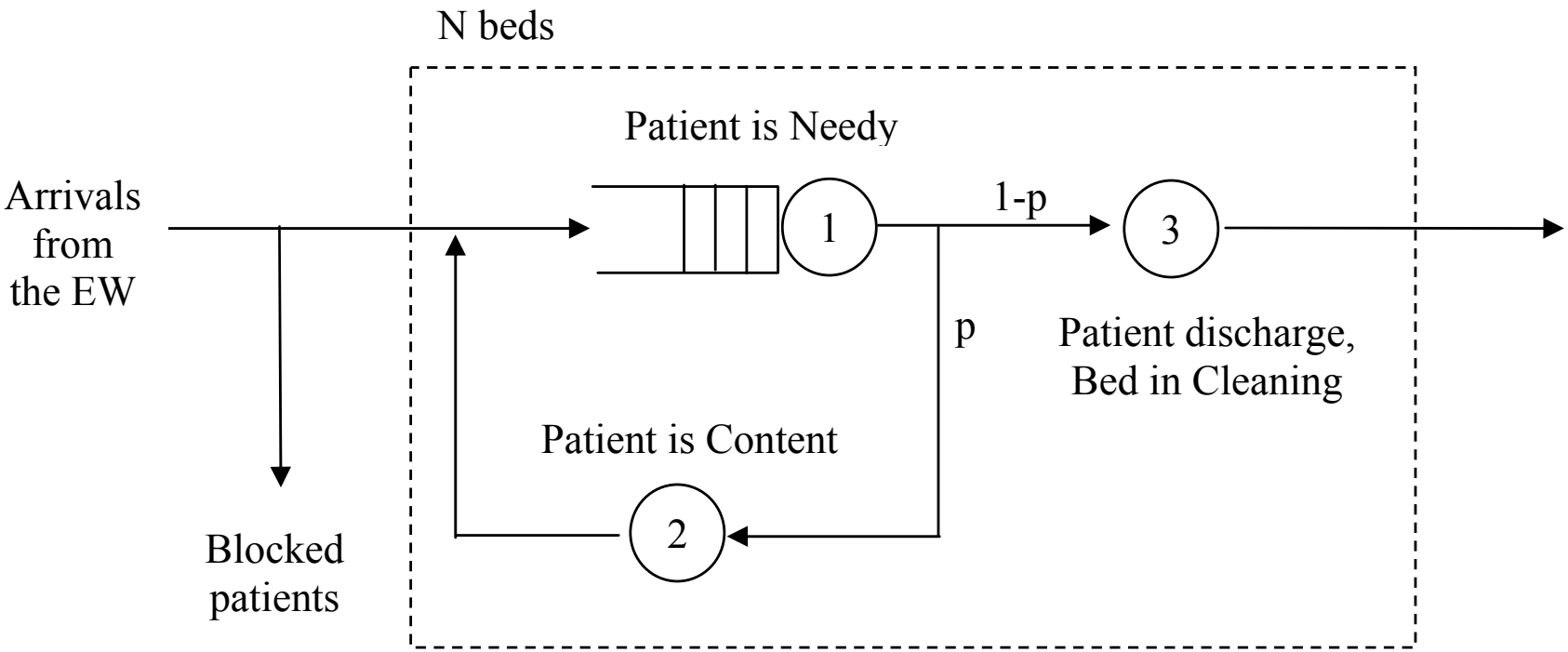
Beds capacity:

Green L.

How Many Hospital Beds?. 2003.

Model Definition

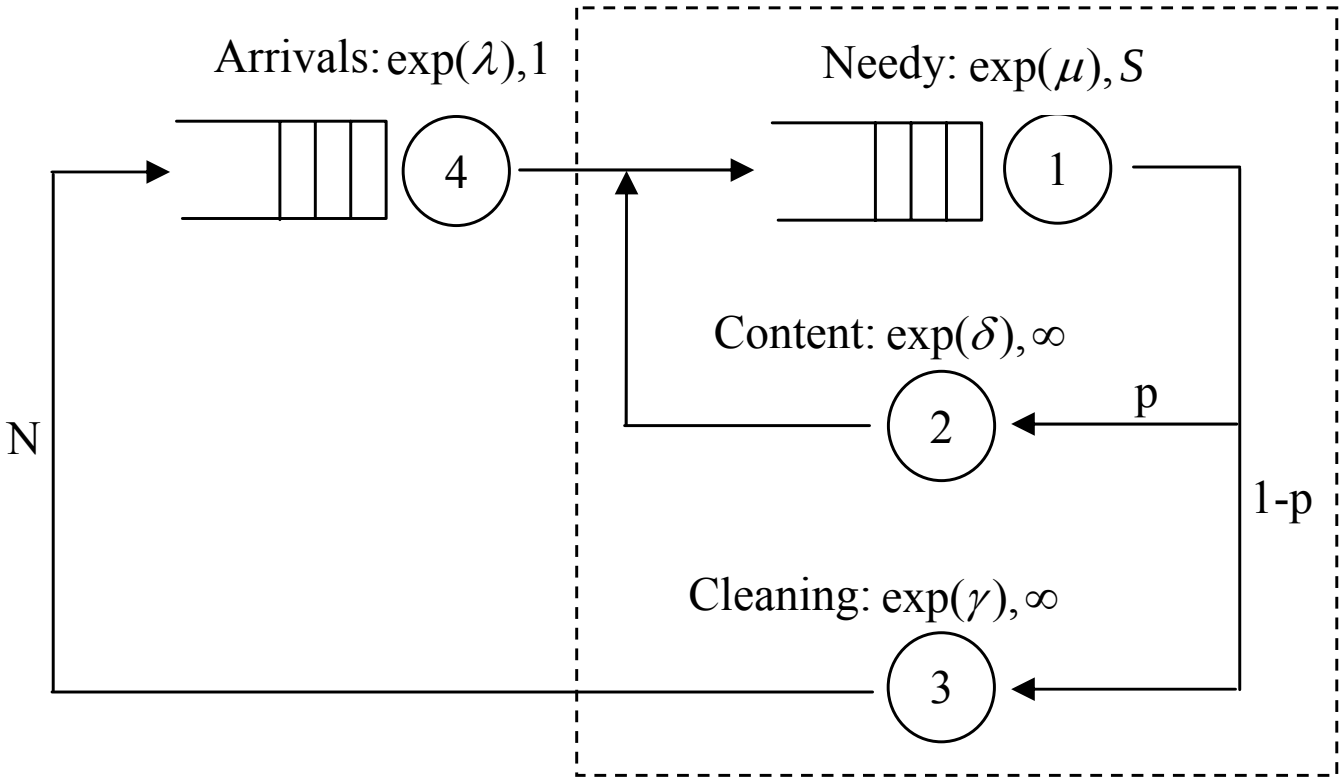
The Internal Ward Queueing Network:



Service times are Exponential; Routing is Markovian

Model Definition

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

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

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

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

Delay Probability

Theorem:

Let λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Then

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

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: β (servers), η (beds), and B (the offered-load-ratio).

Expected Waiting Time

Theorem:

Let λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Then

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

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

Waiting time is one order of magnitude less then service time.

Probability of Blocking

Theorem:

Let λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Then

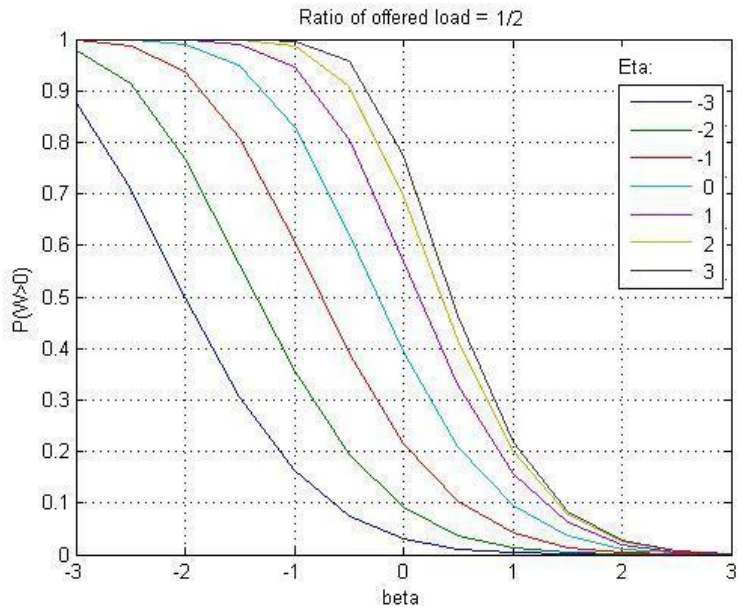
$$\lim_{\lambda\to\infty} \sqrt{s}P(block) = \frac{\nu\phi(\nu_1)\Phi(\nu_2) + \phi(\sqrt{\eta^2+\beta^2})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}{\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{B}\right)d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)}$$

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

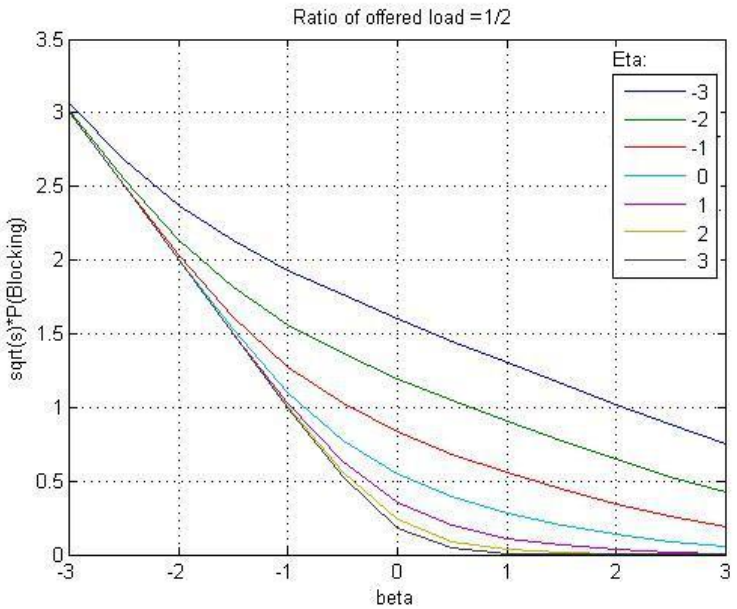
P(Block) $<<$ P(Wait)

The influence of η and β

Waiting



Blocking



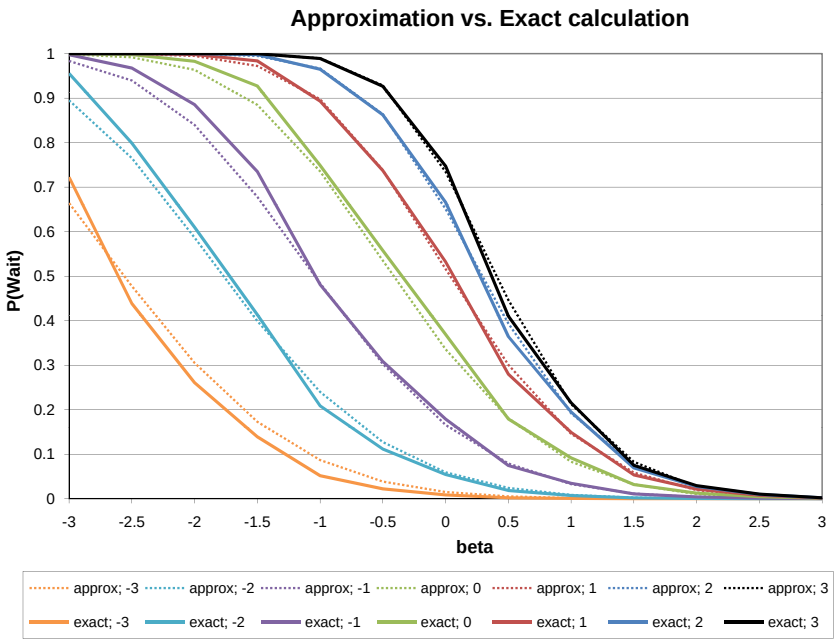
Reminder: β sets the number of nurses, η sets the number of beds:

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

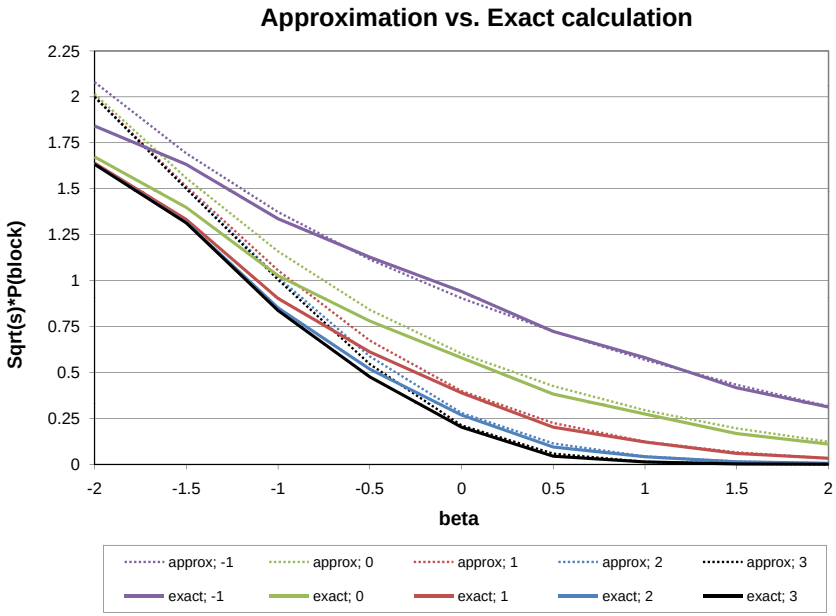
$$n - s = \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} + \eta\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}} + o(\sqrt{\lambda}) \qquad -\infty < \eta < \infty$$

Approximation vs. Exact Calculation

Waiting



Blocking

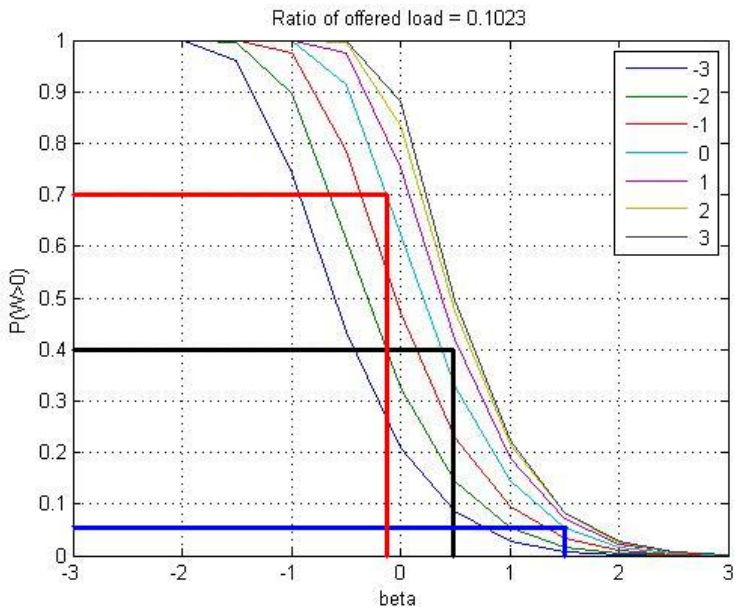


Numerical Example

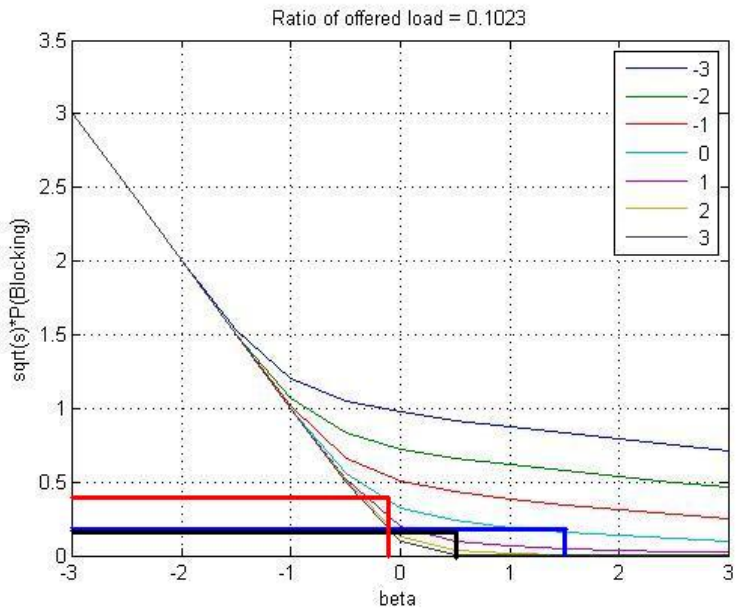
- N=42 with 78% occupancy
- ALOS = 4.3 days
- Average service time = 15 min
- 0.4 requests per hour
- => $\lambda = 0.32, \mu = 4, \delta = 0.4, \gamma = 4, p = 0.975$
- => Ratio of offered-load = 0.1

Based on Lundgren and Segesten 2001 + Yankovic and Green 2007

How to find the required η and β ?



P(W>0)



$\sqrt{s}$P(block)

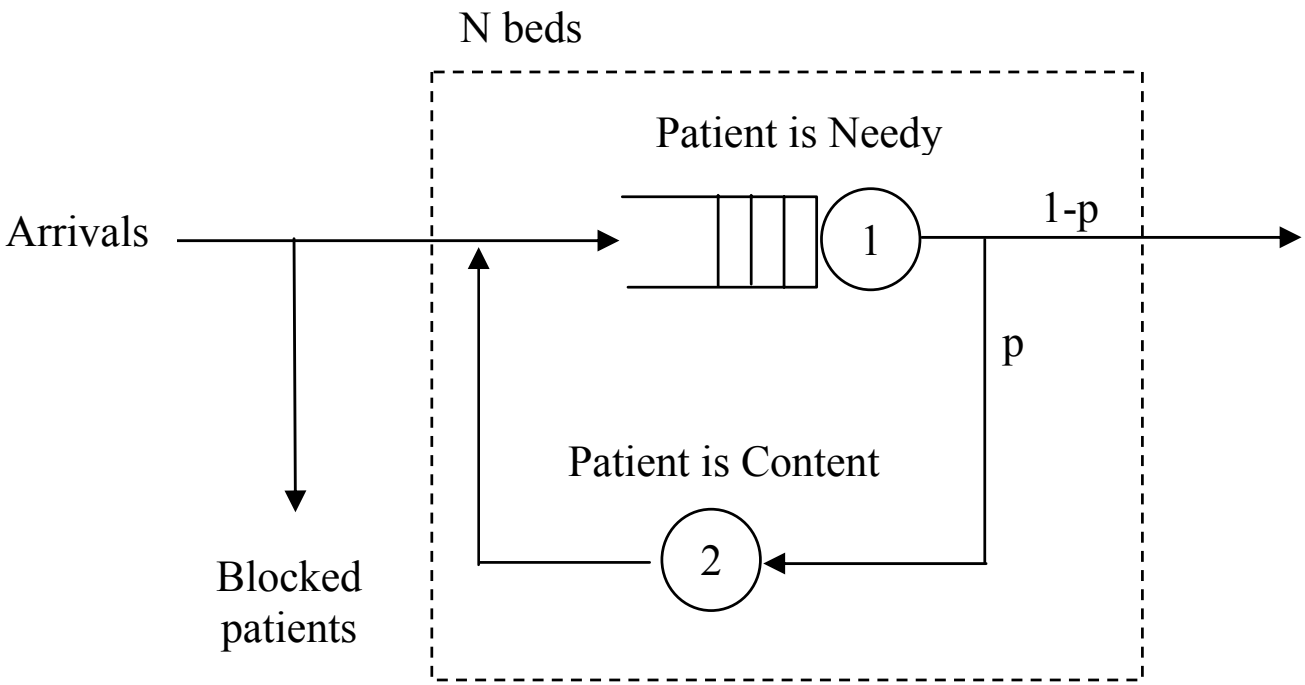
if $\beta = 0.5$ and $\eta = 0.5$ ($s = 4, n = 38$): $P(block) \approx 0.07, P(wait) \approx 0.4$

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 $\eta \approx 0$ ($s = 3, n = 34$): $P(block) \approx 0.21, P(wait) \approx 0.70$

Model Definition

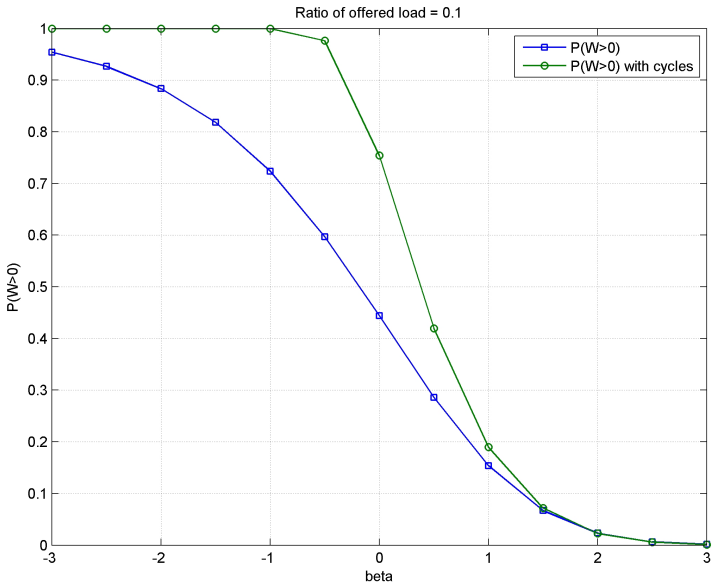
The Time-Varying Semi-Open Erlang-R Model:

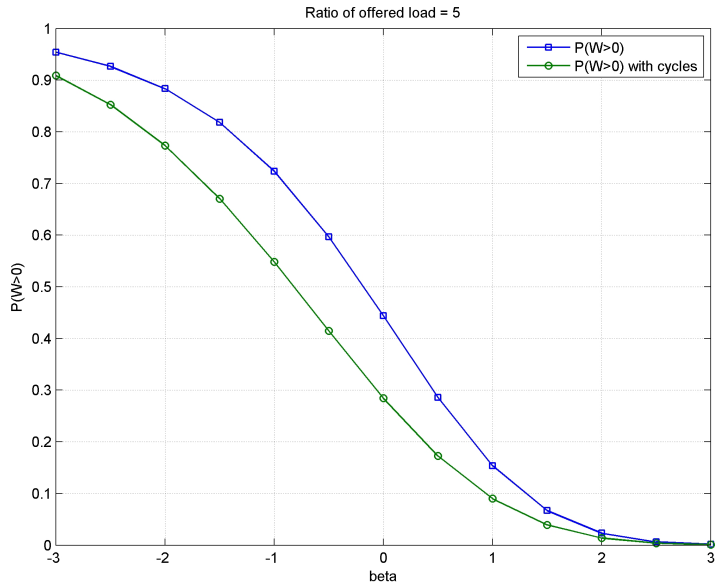


Marginal compared to loss system: steady-state distributions differ.

=> Must consider Reentrant customers, even in steady-state!

Steady-State P(Wait>0) - Semi-Open Erlang-R vs. Loss System





Staffing and Allocation: Determine $s(t)$ and $n(t)$, $t \geq 0$

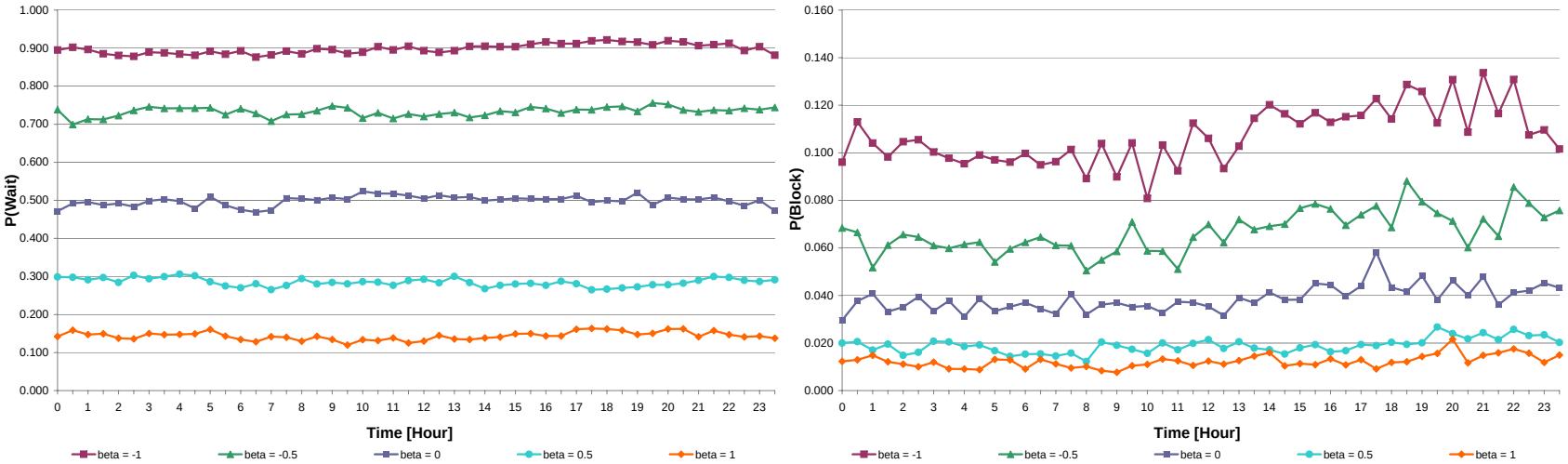
MOL Algorithm for Semi-Open Erlang-R:

- 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 β and η are chosen according to the *steady-state Semi-Open* Erlang-R formula.

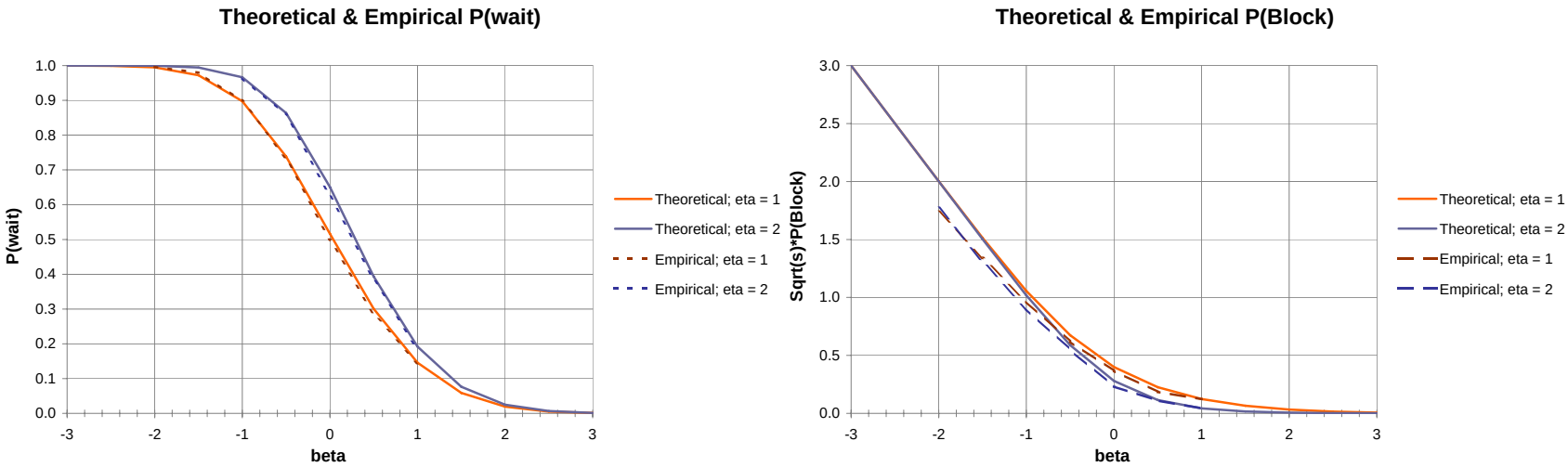
Time-varying Performance

P(Wait) and P(Blocking) as a function of time, for semi-open Erlang-R:



Comparison to Steady-State Performance

Average P(Wait) and P(Blocking) as a function of β, for semi-open Erlang-R:



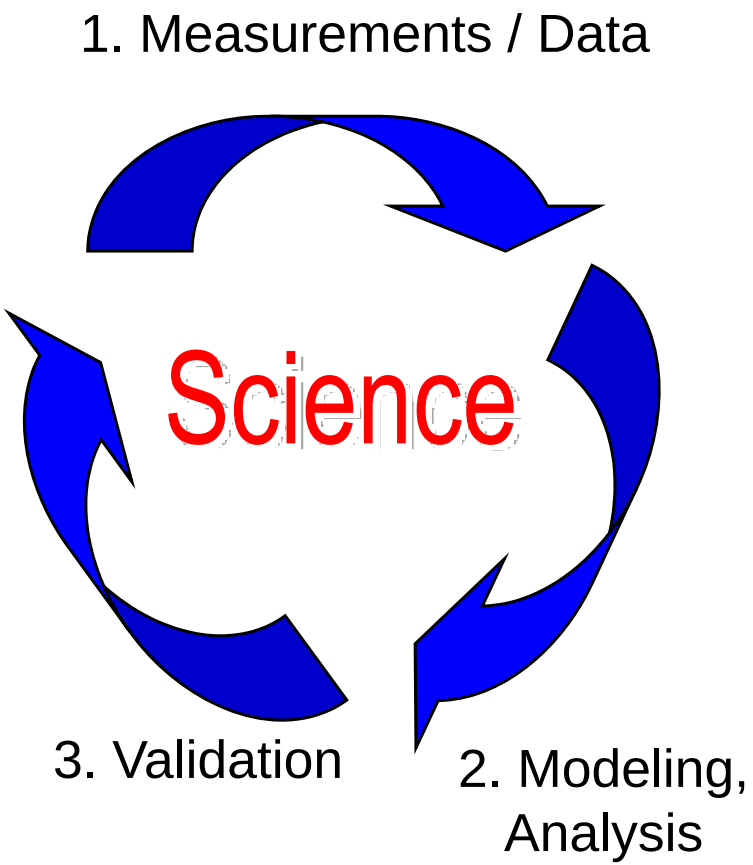
Conclusion of Part II: Semi-Open

- 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.

Part III:

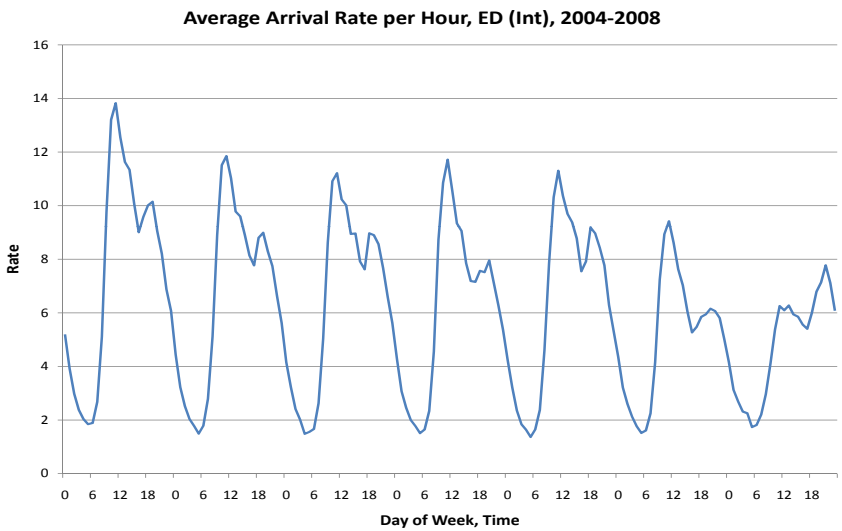
Empirical Analysis of Patients Flow

Goals

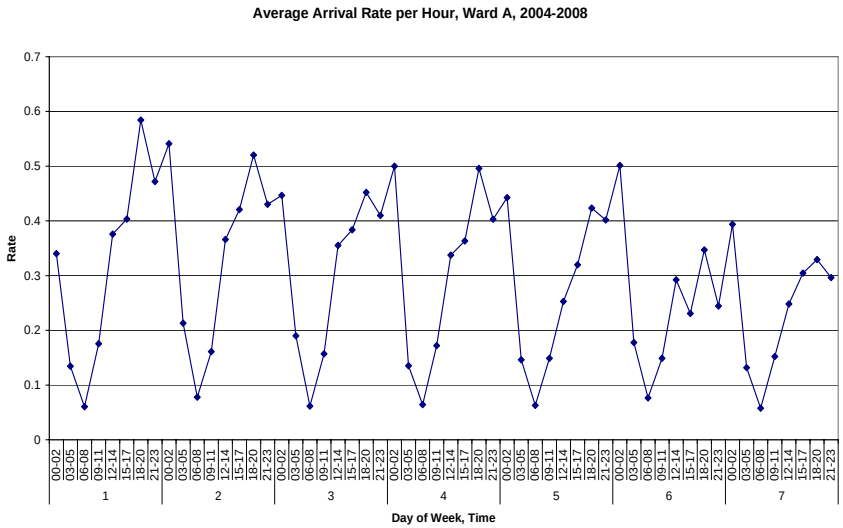

1. Measurements / Data
Science
3. Validation
2. Modeling, Analysis

EW's vs. IW's Arrival Rate

Emergency Ward



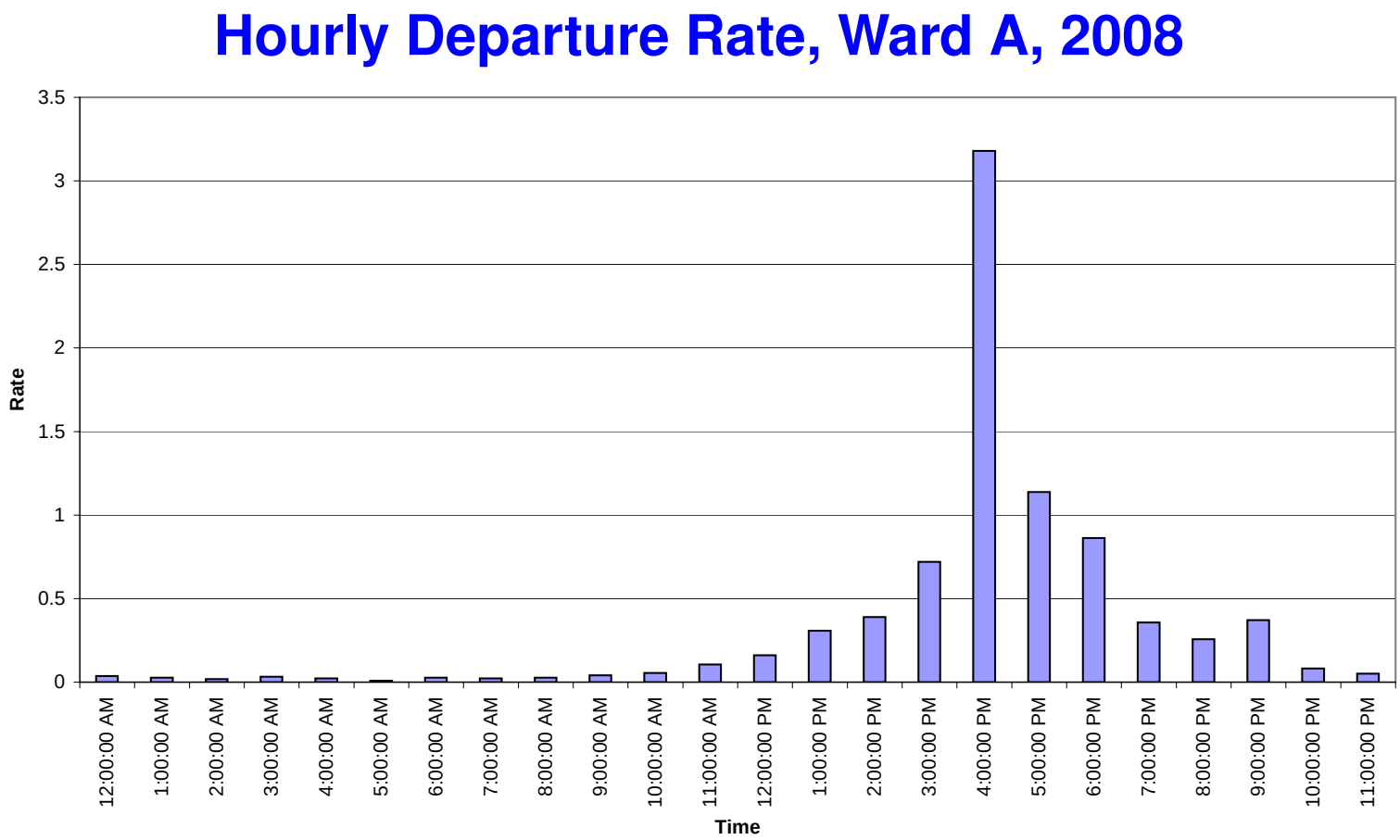
Internal Ward



Have the same structure!

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

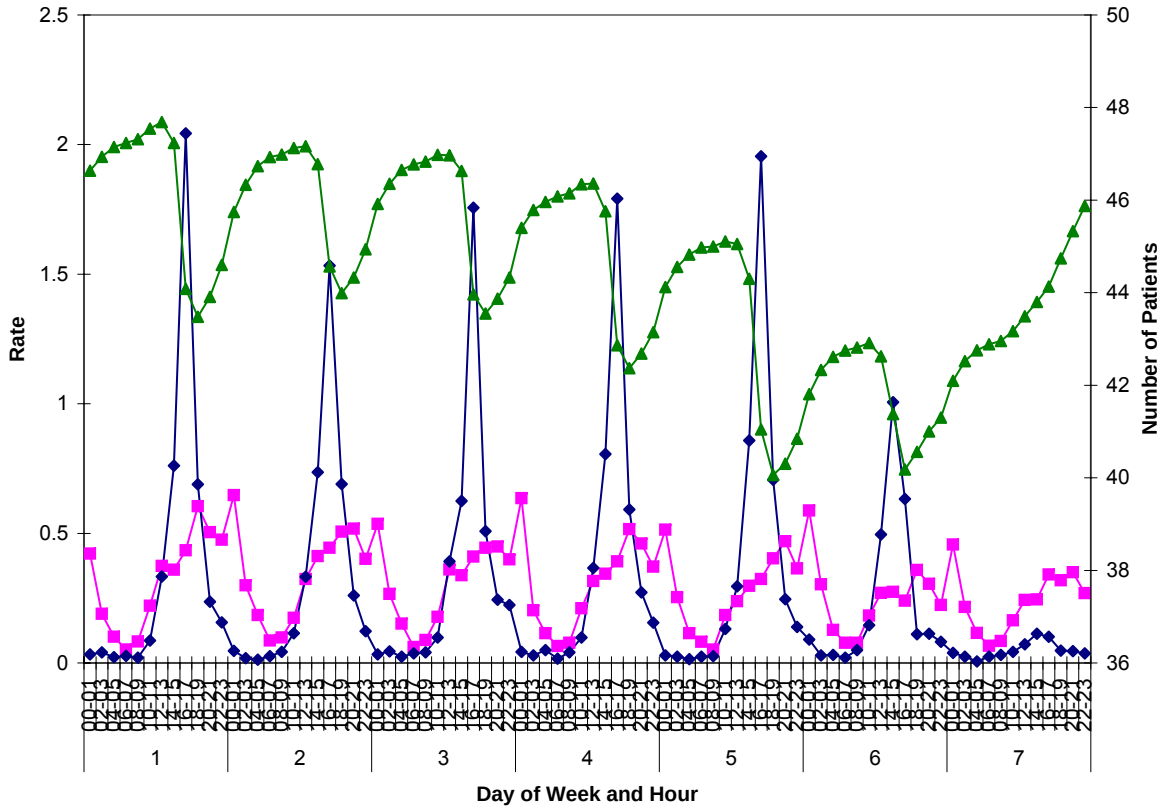
IW Departure Rate



=> The operational impact of release policy in IWs.

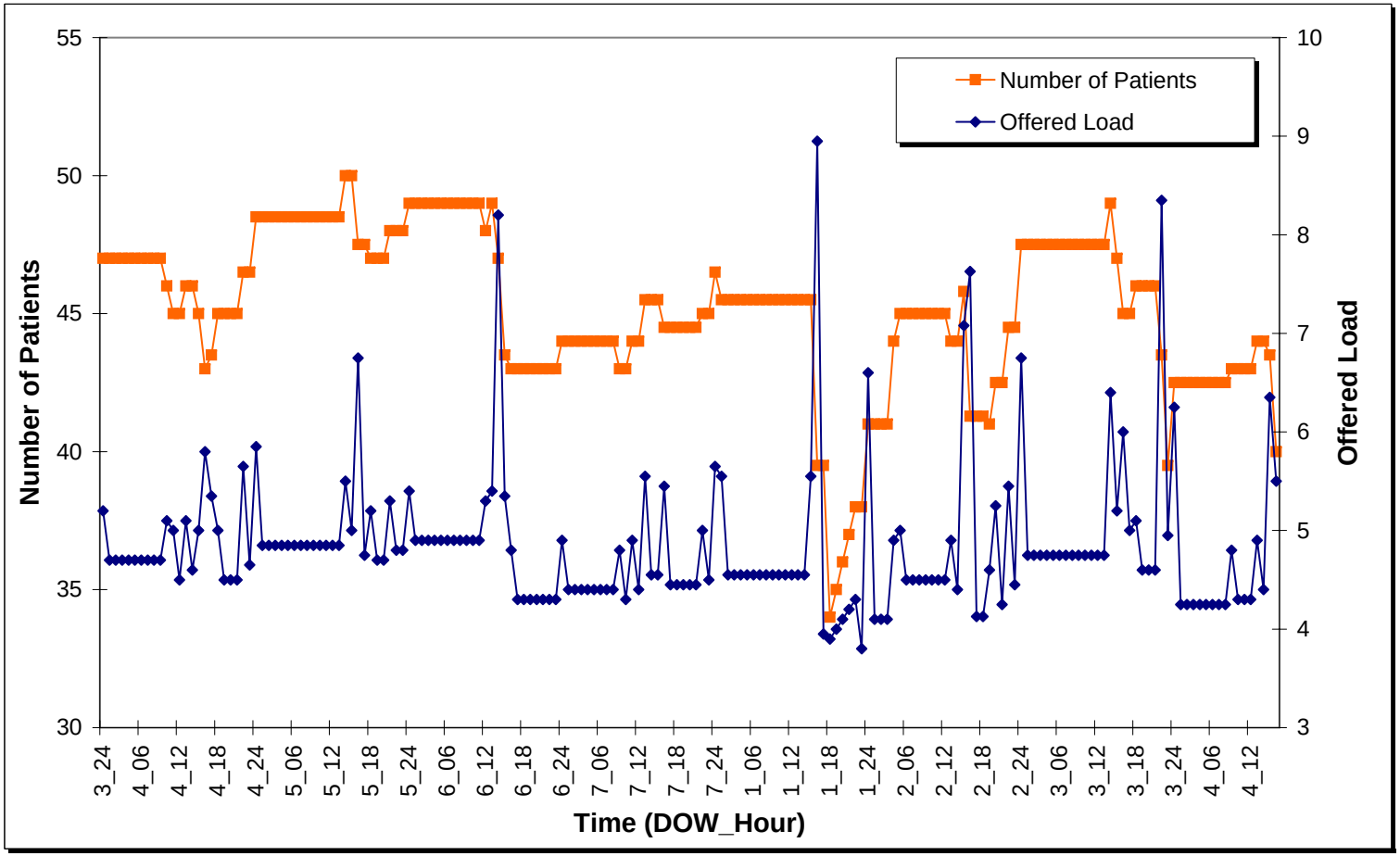
Patients, Arrival- and Departure-Rate

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.

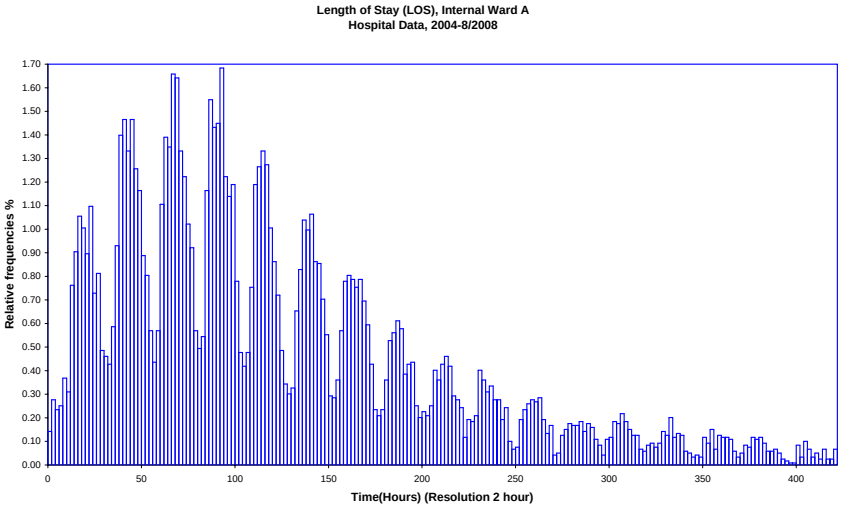
Patients and Load



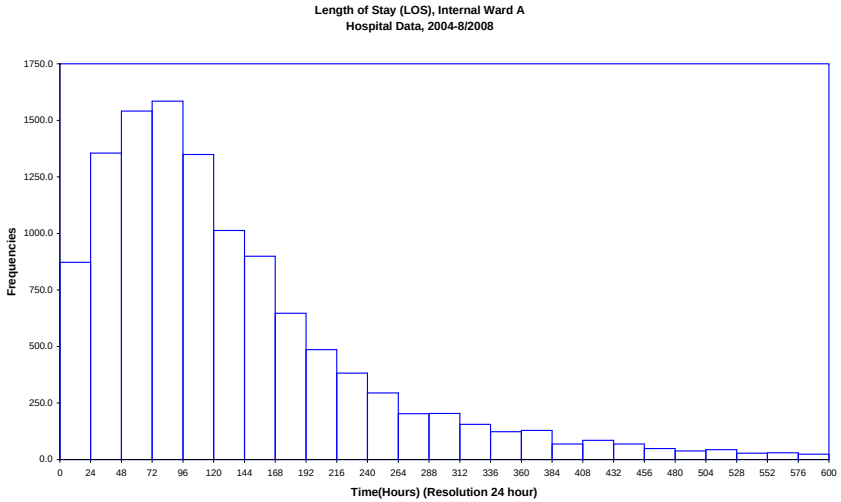
In most loaded hours - least nurses recommendation.

LOS distribution

Two-hours resolution

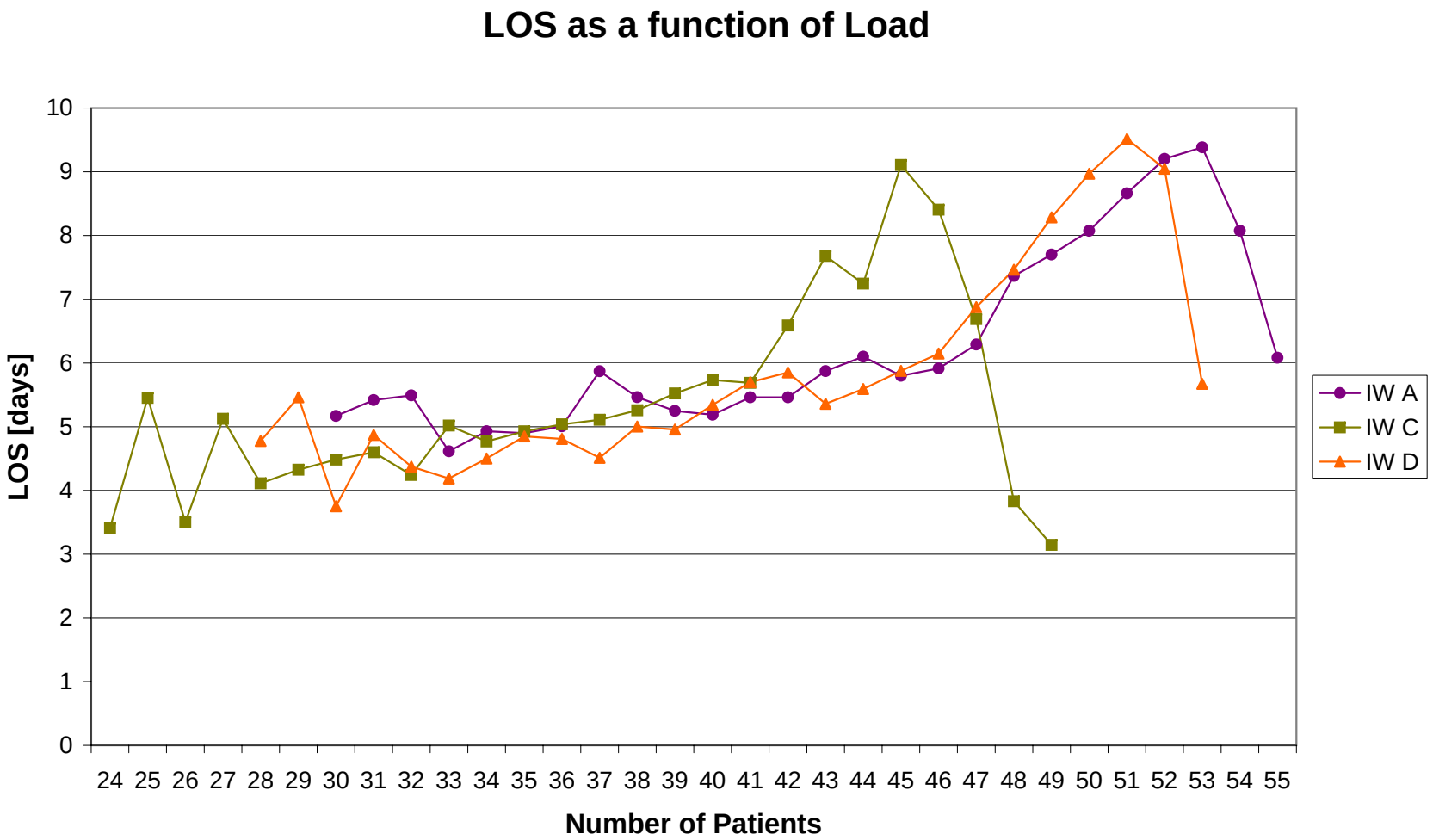


24-hours resolution



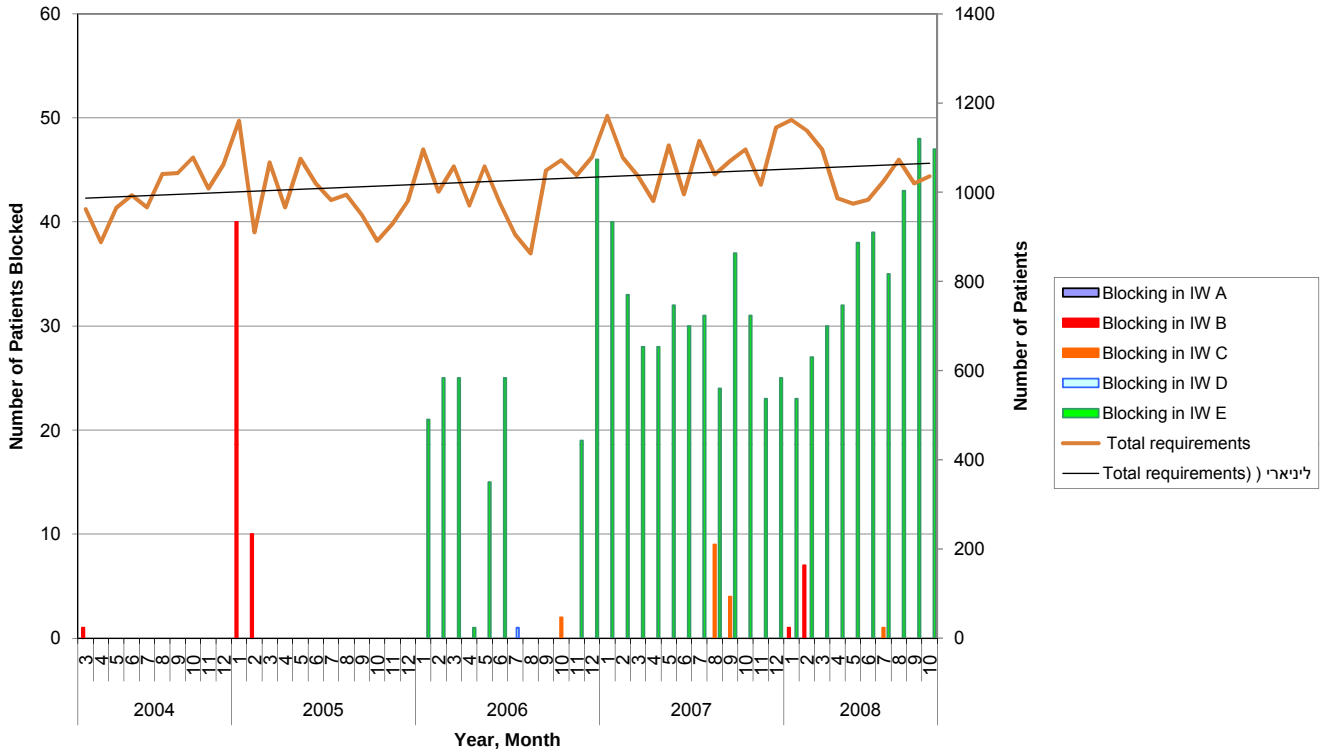
LOS is a mixture of Normally distributed Random variables.

LOS and Load


LOS as a function of Load
LOS [days]
0
1
2
3
4
5
6
7
8
9
10
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
Number of Patients
IW A
IW C
IW D

Blocking

Total Blocking Incidents per Year and Month, all IWs



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 in IW and Oncology

Returns to hospital

Ward	No. of returns per patient (in 4 years)	Time between returns (days)	Probability of return within 3 month
Internal	1.76	208	22%
Oncology	5.76	29	76%

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