Capacity Management in Hospitals: Semi-Open Queueing Networks in the QED Regime

Galit Yom-Tov

Joint work with Avishai Mandelbaum

Technion – Israel Institute of Technology

5/June/2008

Agenda

- Introduction
- Medical Unit Model
- Mathematical Results
- Numerical Example
- Time-varying Model
- Future Research

Work-Force and Bed Capacity Planning

- Total health expenditure as percentage of gross domestic product: Israel 8%, EU 10%, USA 14%.
- Human resource constitute 70% of hospital expenditure.
- There are 3M registered nurses in the U.S. but still a chronic shortage.
- California law set nurse-to-patient ratios such as 1:6 for pediatric care unit.
- O.B. Jennings and F. de Véricourt (2008) showed that fixed ratios do not account for economies of scale.
- Management measures average occupancy levels, while arrivals have seasonal patterns and stochastic variability (Green 2004).

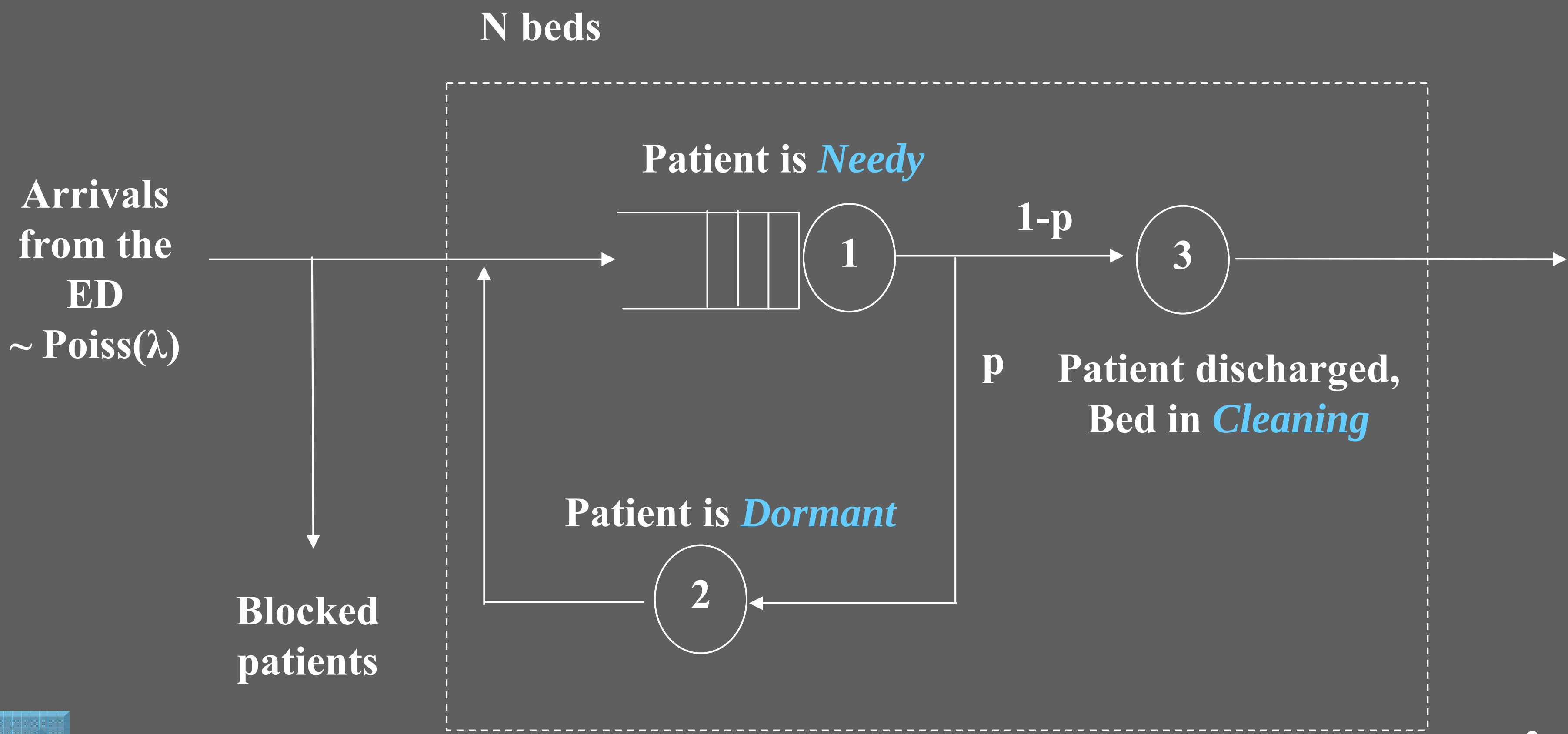
Research Objectives

- Analyzing model for a Medical Unit with *s* nurses and *n* beds, which are partly/fully occupied by patients: semi-open queueing network with multiple statistically identical customers and servers.
- Questions addressed: How many servers (nurses) are required (staffing), and how many fixed resources (beds) are needed (allocation) in order to minimize costs while sustaining a certain service level?
- Coping with time-variability

We Follow -

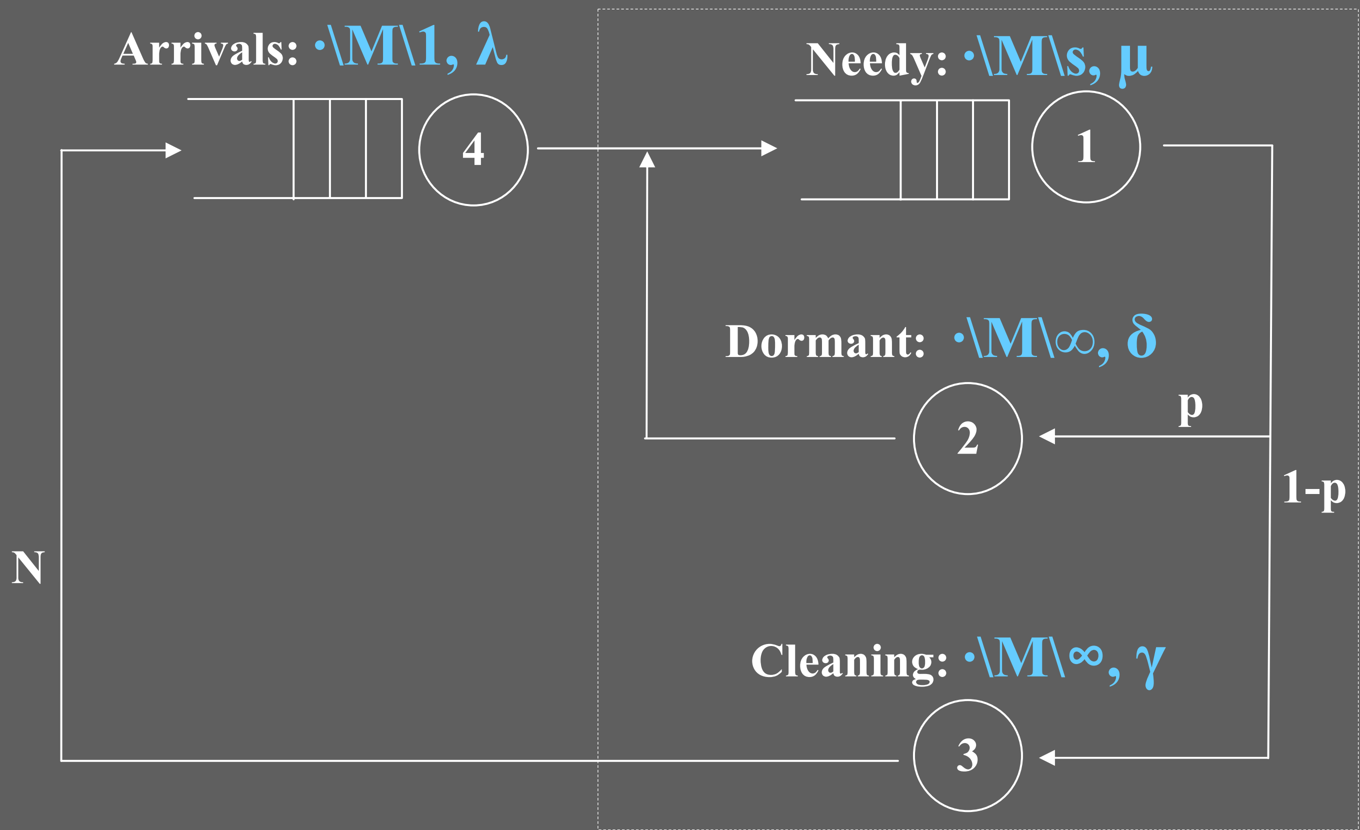
- Basic:
  - Halfin and Whitt (1981)
  - Mandelbaum, Massey and Reiman (1998)
  - Khudyakov (2006)
- Analytical models in HC:
  - Nurse staffing: Jennings and Véricourt (2007), Yankovic and Green (2007)
  - Beds capacity: Green (2002,2004)
- Service Engineering (mainly call centers):
  - Gans, Koole, Mandelbaum: “Telephone call centers: Tutorial, Review and Research prospects”

The MU Model as a Semi-Open Queueing Network



Service times are Exponential; Routing is Markovian

The MU Model as a Closed Jackson Network



→Product Form - $\pi_N(n,d,c)$ stationary dist.

Service Level Objectives

(Function of $\lambda,\mu,\delta,\gamma,p,s,n$)

- 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

QED Q's: Quality- and Efficiency-Driven Queues

- Traditional queueing theory predicts that service-quality and server's efficiency **must** trade off against each other.
- Yet, one can balance both requirements carefully (Example: in well-run call-centers, 50% served "immediately", along with over 90% agent's utilization, is not uncommon)
- This is achieved in a special asymptotic operational regime – the QED regime

QED Regime characteristics

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

The offered load at service station 1 (needy)

The offered load at non-service station 2+3 (dormant + cleaning)

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

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

- Many-server asymptotic

Probability of Delay

$$P(W > 0) = \sum_{i \geq s} \pi_{n-1}(i,j,k) = \sum_{l=s}^{n-1} \sum_{m=s}^{l} \sum_{i=s}^{m} \pi_{n-1}(i, m-i, l-m)$$

Theorem 2. *Let the variables 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 $B = \frac{R_N}{R_C + R_D} = \frac{\delta\gamma}{\mu(p\gamma + (1-p)\delta)}$, $\eta_1 = \eta - \beta\sqrt{B^{-1}}$.

- The probability is a function of three parameters: beta, eta, and offered-load-ratio

Expected Waiting Time

$$E[W] = \frac{1}{\mu s}\sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\pi_{n-1}(i, m-i, l-m)(i-s+1)$$

Theorem 4. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Then*

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

where $B = \frac{R_N}{R_C+R_D} = \frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}$, $\eta_1 = \eta - \beta\sqrt{B^{-1}}$.

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

Probability of Blocking

$$P_l = \pi_0 \left(\frac{1}{l!}\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)^l + I_{\{l>s\}} \sum_{i=s+1}^{l}\sum_{j=0}^{l-i}\left(\frac{1}{s!s^{i-s}}-\frac{1}{i!}\right)\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(l-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{l-i-j}\right)$$

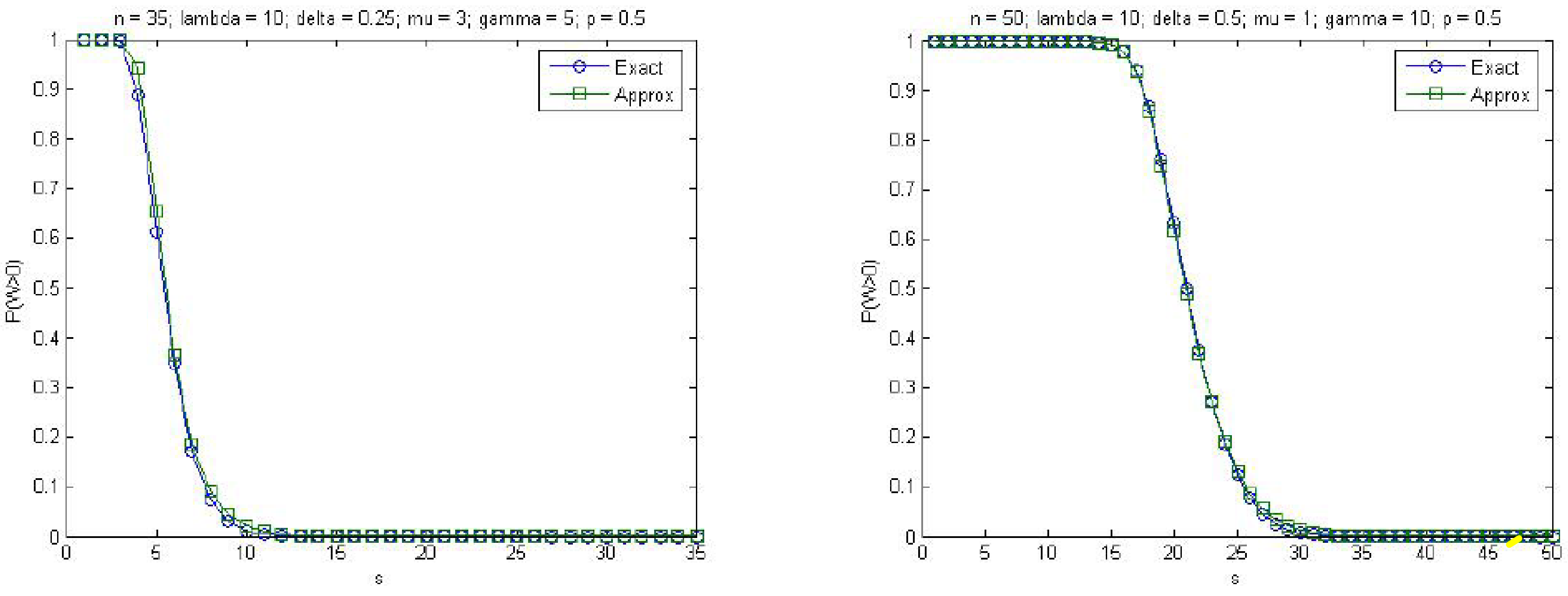
Theorem 6. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Define* $B = \frac{R_N}{R_C+R_D} = \frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}$, *then*

$$\lim_{\lambda\to\infty} \sqrt{s}P(block) = \frac{\nu\phi(\nu_1)\Phi(\nu_2)+\phi(\sqrt{\eta^2+\beta^2})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}{\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{B}\right)d\Phi(t)+\frac{\phi(\beta)\Phi(\eta)}{\beta}-\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)} \qquad (5.9)$$

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

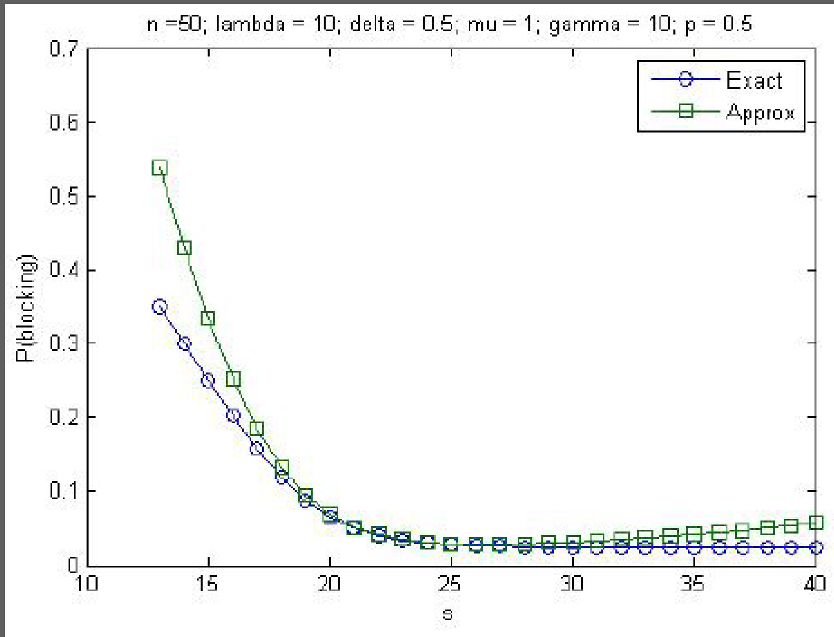
- P(Blocking) << P(Waiting)

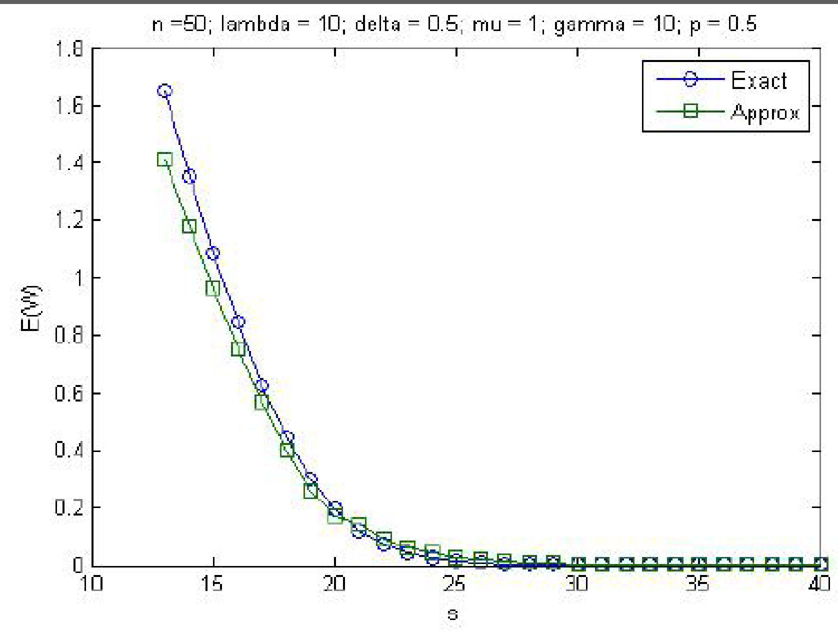
Approximation vs. Exact Calculation – Medium system (n=35,50), P(W>0)



(a) (b)

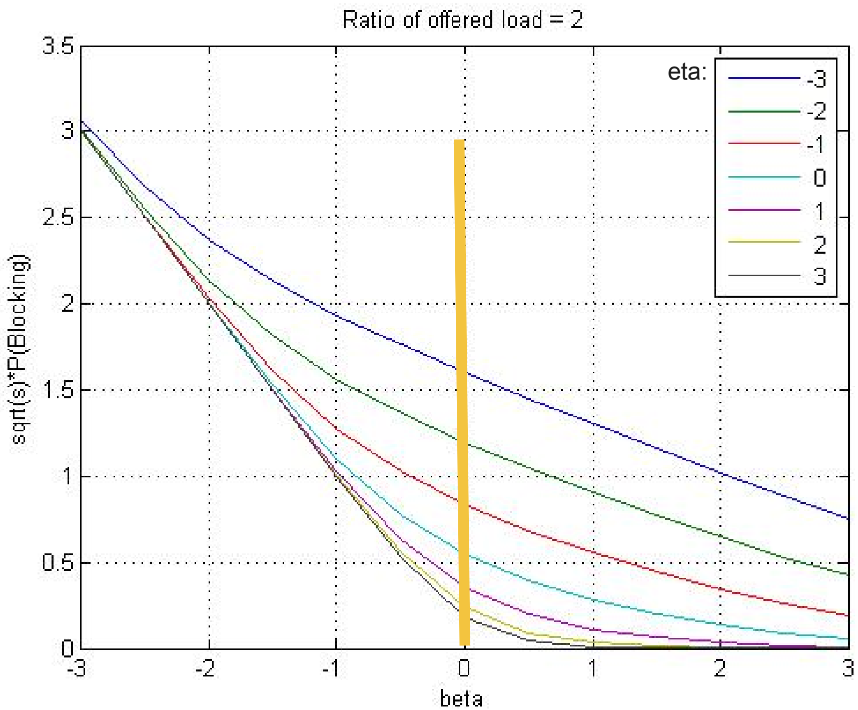
Approximation vs. Exact Calculation – P(blocking) and E[W]



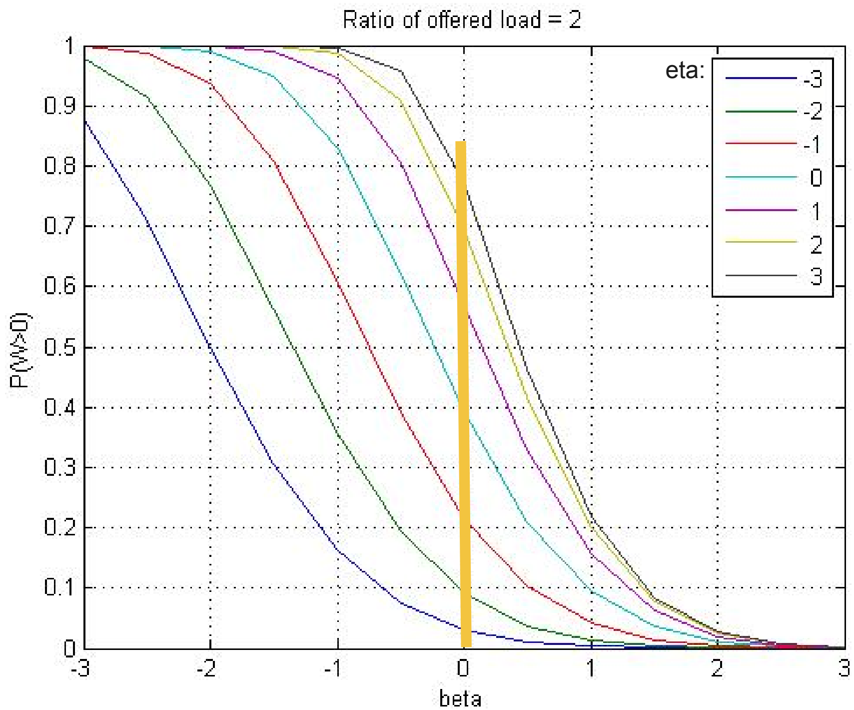


The influence of β and η?

Blocking



Waiting



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

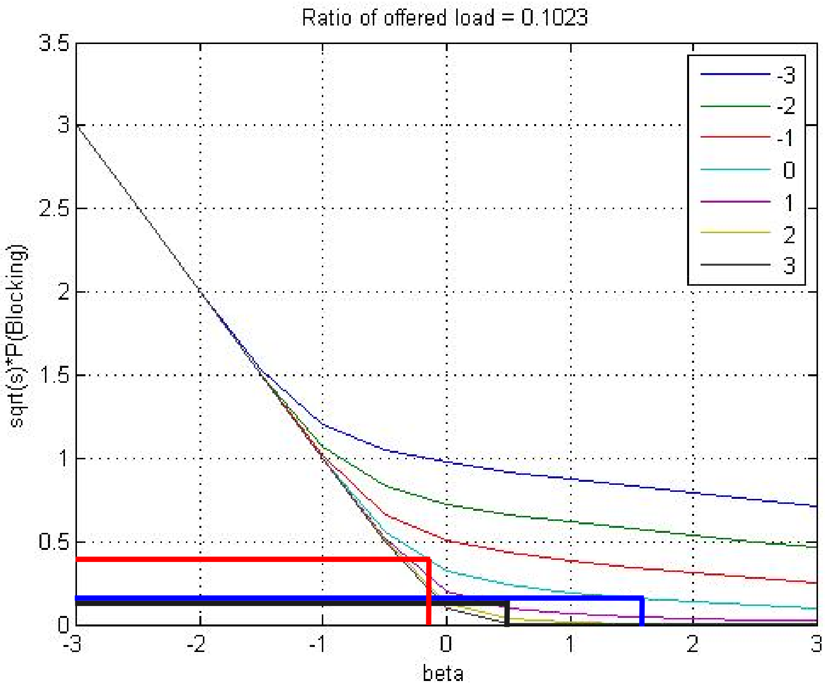
$$\text{(ii)}\ n - s = \eta\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} + o(\sqrt{\lambda}), \qquad -\infty < \eta < \infty$$

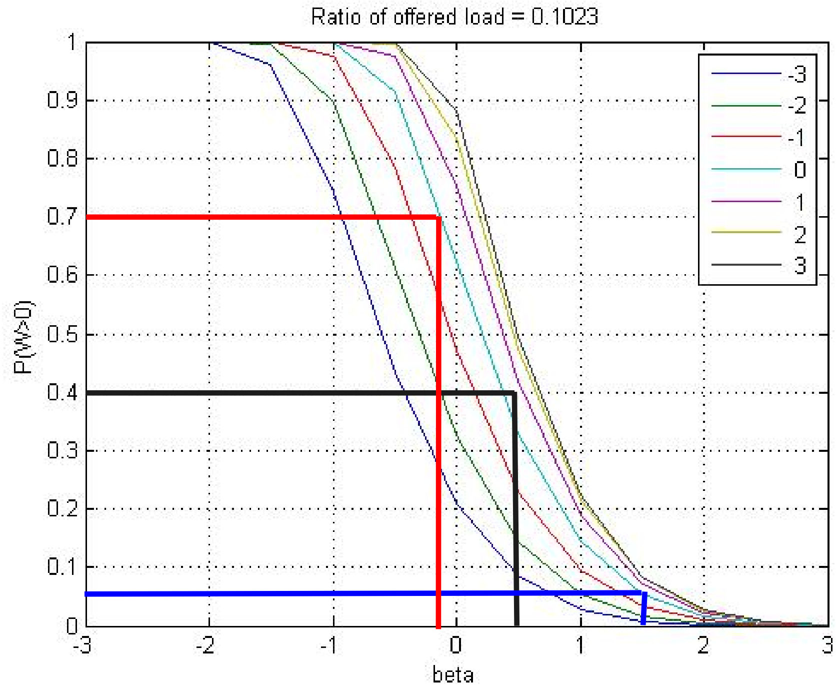
Numerical Example

(based on Lundgren and Segesten 2001 + Yankovic and Green 2007)

- 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

How to find the required β and η?

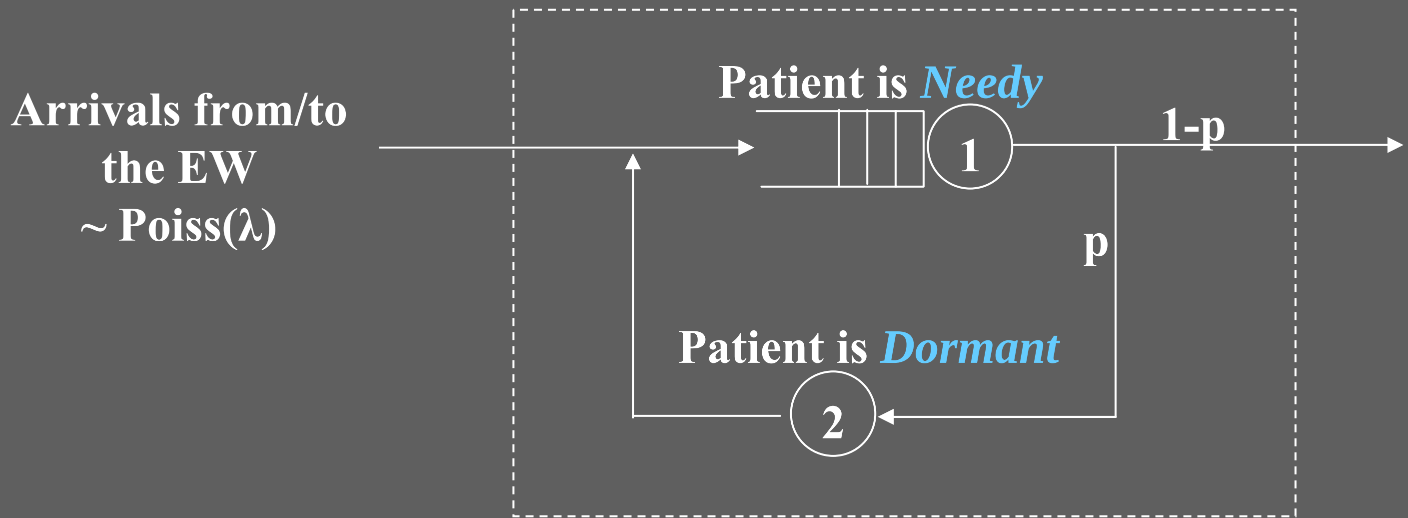




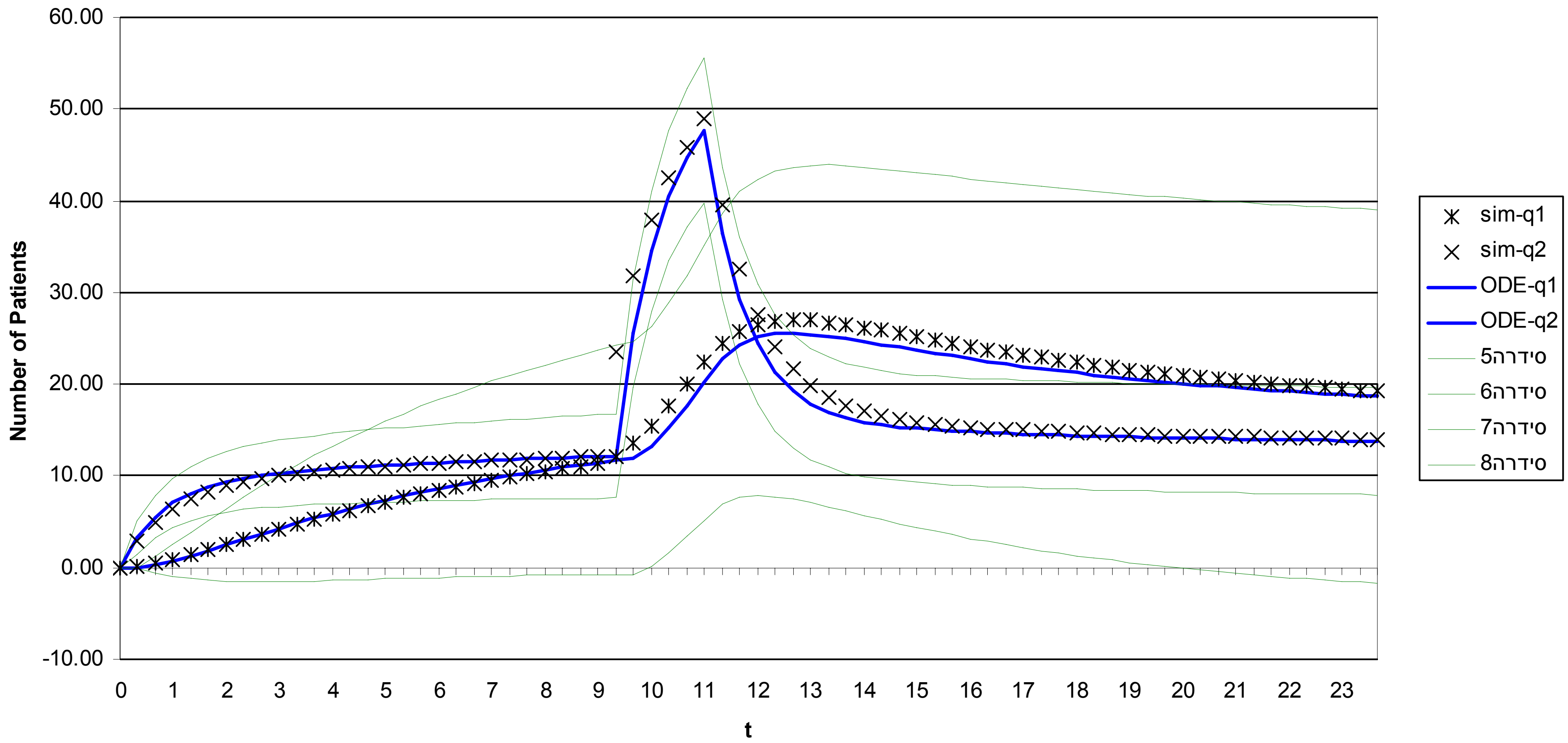
if β=0.5 and η=0.5 (s=4, n=38): P(block)≅0.07, P(wait) ≅ 0.4
if β=1.5 and η ≅ 0 (s=6, n=37): P(block)≅0.068, P(wait) ≅ 0.084
if β=-0.1 and η ≅ 0 (s=3, n=34): P(block)≅0.21, P(wait) ≅ 0.70

Modeling time-variability

- Procedures at mass-casualty event
- Blocking cancelled -> open system



Delta = 0.2; Mu = 1; p = 0.25; s = 50; Lambda=10 (t<9 or t>11), Lambda=50 (9<t<11)



Future Research

- Investigating approximation of closed system
  - From which *n* are the approximations accurate? (simulation vs. rates of convergence)
- Optimization
  - Solving the bed-nurse optimization problem
  - Difference between hierarchical and simultaneous planning methods
- Validation of model using RFID data
- Expanding the model (Heterogeneous patients; adding doctors)


Thank you