On Fair Routing from Emergency Departments to Hospital Wards: QED Queues with Heterogeneous Servers

Yulia Tseytlin

IBM, Technion

Applied Probability Society Conference, July 2011

Joint work with Avi Mandelbaum, Technion
and Petar Momčilović, University of Florida

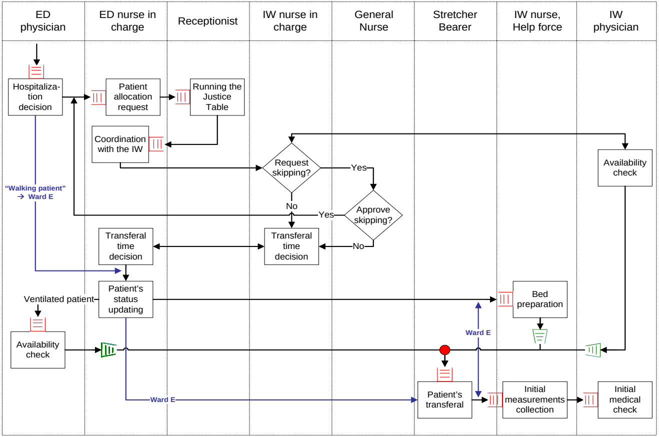
Hospital

- Anonymous Hospital – large Israeli hospital
  - 1000 beds
  - 45 medical units
  - $\sim$ 75,000 patients hospitalized yearly

- Variety of medical units
  - Emergency Department (ED):
    - average arrival rate = 240 patients/day
    - 50 beds
  - Internal Wards (IW):
    - A – D: the same medical capabilities

- ED-IW routing policy
  - current policy: cyclical

Flow Chart



Resource Queue - Synchronization Queue -

- Ending point of simultaneous processes

Fairness

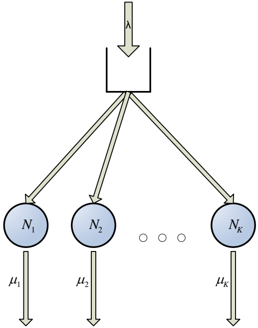
	Ward A	Ward B	Ward C	Ward D
Capacity (# beds)	45 (52)	**30** (35)	44 (46)	42 (44)
Average Length of Stay (days)	6.5	**4.5**	5.4	5.7
Return rate (within 3 months)	16.4%	17.4%	19.2%	17.6%
Mean occupancy level	97.8%	94.4%	86.8%	91.1%
Mean # patients per bed per month	4.58	**6.38**	4.89	4.86

- Nurses: Fixed nurse-to-bed ratio (1:5) + Salaried + Fixed assignment
- Load on Ward B staff is the highest
- Similar story in other hospitals

- Heterogeneity

Inverted-V Model

- Customer = Patient, Pool = Ward, **Server = Bed**
- Poisson arrivals with rate λ
- K server pools
- Pool j:
  - N_j exponential servers
  - server rate μ_j
  - service capacity $c_j = \mu_j N_j$
  - I_j idle servers



- Queue length Q
- $I = \sum_{j=1}^{K} I_j - Q$ $\quad$ ($(I)^+$ - total number of idle servers)
- Waiting line
  - infinite capacity
  - FCFS

Quality and Efficiency Driven Regime

- Informally...
  - A system with a large volume of arrivals and many servers
  - Waiting times are order of magnitude shorter than service times
  - Total service capacity equals the demand plus a safety capacity

- In Anonymous Hospital:
  - 30-50 servers (beds) in each pool (ward)
  - Waiting times vs. service times: hours vs. days
  - Servers utilization (beds occupancy) is above 85%

- Focus on:
  - Idleness ratios

$$\frac{1-\rho_i}{1-\rho_j} = \frac{\mathbb{E}I_i/N_i}{\mathbb{E}I_j/N_j}$$

  - Flux ratios

$$\frac{\gamma_i}{\gamma_j} = \frac{\mu_i\rho_i}{\mu_j\rho_j}$$

QED Regime

- **Rule 1: Pool capacities**

$$\frac{c_i^\lambda}{\sum c_j^\lambda} \to a_i > 0$$

- Traffic intensity: $\rho^\lambda = \lambda / \sum c_i$
- Arithmetic-mean service rate: $\hat{\mu} = \sum a_i \mu_i$
- System "size": $\nu^\lambda = \lambda / \hat{\mu}$
- **Rule 2: Square-root safety rule**

$$\sqrt{\nu^\lambda}(1 - \rho^\lambda) \to \delta > 0$$

Randomized Most-Idle (RMI) Routing

- At time t assign a customer to pool j with probability $I_j^\lambda(t)/(I^\lambda(t))^+$
  - "Blind", adaptive to changing capacity
  - Equivalent to LISF in QED: $I_j^\lambda \approx a_j(I^\lambda)^+$, or $I_i^\lambda/I_j^\lambda \approx a_i/a_j$
  - $\dfrac{1-\rho_i}{1-\rho_j} \approx \dfrac{\gamma_i}{\gamma_j} \approx \dfrac{\mu_i}{\mu_j}$
- Diffusion scale: $\hat{I}^\lambda = I^\lambda/\sqrt{\nu^\lambda}$
- **Dimensionality Reduction:** $I_j^\lambda \approx a_j(I^\lambda)^+$
- $I_j^\lambda - a_j(I^\lambda)^+ \approx ?$ $\Rightarrow$ **Sub-diffusion scale:**

$$\hat{I}_j^\lambda = \frac{1}{\sqrt{I^\lambda}}\left(I_j^\lambda - \frac{c_j^\lambda}{\sum c_i^\lambda} I^\lambda\right)$$

- $\sqrt[4]{\nu^\lambda} \ll \sqrt{\nu^\lambda}$ (wards' size = 30-50 beds)

Main Result

Theorem: Consider the inverted-V model in steady-state, under the RMI routing algorithm in the QED regime. Then, as $\lambda \to \infty$,

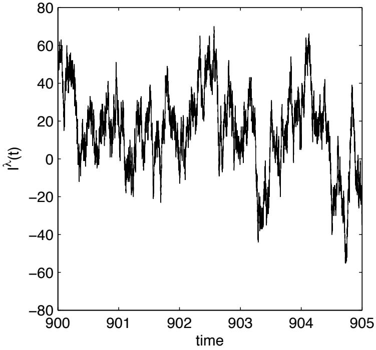
$$\left(\hat{I}^\lambda, (\hat{I}_1^\lambda, \ldots, \hat{I}_K^\lambda) 1_{\{\hat{I}^\lambda > 0\}}\right) \Rightarrow \left(\hat{I}, (\hat{I}_1, \ldots, \hat{I}_K) 1_{\{\hat{I} > 0\}}\right),$$

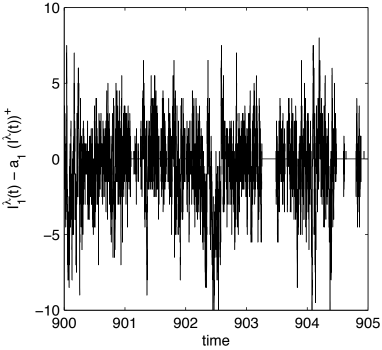
where

- $\hat{I}$ and $(\hat{I}_1, \ldots, \hat{I}_K)$ are independent
- $\mathbb{P}[\hat{I} \le 0] = \left(1 + \delta \, \frac{\Phi(\delta)}{\varphi(\delta)}\right)^{-1}$
- $\mathbb{P}[\hat{I} > x \,|\, \hat{I} > 0] = \Phi(\delta - x)/\Phi(\delta)$, $x \ge 0$
- $\mathbb{P}[\hat{I} \le x \,|\, \hat{I} \le 0] = e^{\delta x}$, $x \le 0$
- $(\hat{I}_1, \ldots, \hat{I}_K)$ is zero-mean multi-variate normal, with $\mathbb{E} \hat{I}_i \hat{I}_j = a_i 1_{\{i=j\}} - a_i a_j$

Dimensionality Reduction

- Example: $K = 2$, $N_1 = 138$, $N_2 = 276$ ...
- $\sqrt{\nu} \approx 18.7$ and $\sqrt[4]{\nu} \approx 4.3$





- Time scale

Three Scales

- Hospital data: $\lambda \approx 189.7$ patients/week, $\hat{\mu} \approx 1.18$ patients/week
- Thus $\nu^\lambda \approx 160.8$, $\sqrt{\nu^\lambda} \approx 12.7$ and $\sqrt[4]{\nu^\lambda} \approx 3.6$

- Finest scale: patient/hour
  - $1/\lambda \approx 0.86$ hours

- Coarsest scale: sub-ward/week
  - sub-ward $\approx 1/3$ or $1/4$ of a ward
  - $1/\hat{\mu} \approx 0.85$ weeks

- Intermediate scale: room/day
  - room $=$ 4 beds
  - $1/\sqrt{\lambda\hat{\mu}} \approx 0.87$ days
  - idleness ratios the same as under LISF
  - number of patients that need to be moved between wards

- Idleness Ratio (IR) policy: $\arg\max_j \left\{ I_j^\lambda(t-) - w_j(I^\lambda(t-))^+ \right\}$
- Diffusion scale: Equivalence of LISF, IR ($w_j = a_j$) and RMI
- Different information utilized
- Sub-diffusion scale: