Queueing Systems with Heterogeneous Servers: On Fair Routing from Emergency Departments to Internal Wards

A. Mandelbaum, P. Momčilović and Y. Tseytlin

Manufacturing and Service Operations Management Conference

June 29, 2010

Research Motivation

- Consider the process of patients' routing from an Emergency Department (ED) to Internal Wards (IW) in Anonymous Hospital.
- Patients' allocation to the wards does not appear to be fair and waiting times for a transfer to the IW are long.
- We model the "ED-to-IW process" as a queueing system with heterogeneous server pools.
- We analyze this system under various routing policies, in search for fairness and good operational performance, while accounting for availability of information.
- The analysis is in steady-state and in the QED (Quality and Efficiency Driven) regime.

Outline

Introduction

- Anonymous Hospital is a large Israeli hospital:
  - ⋆ 1000 beds
  - ⋆ 45 medical units
  - ⋆ about 75,000 patients hospitalized yearly.
- Among the variety of hospital's medical sections:
  - ⋆ Large ED *(Emergency Department)* with average arrival rate of 240 patients daily and capacity of 40 beds.
  - ⋆ Five IW *(Internal Wards)* which we denote from A to E.
- An internal patient to-be-hospitalized, is directed to one of the five IW according to a certain routing policy.

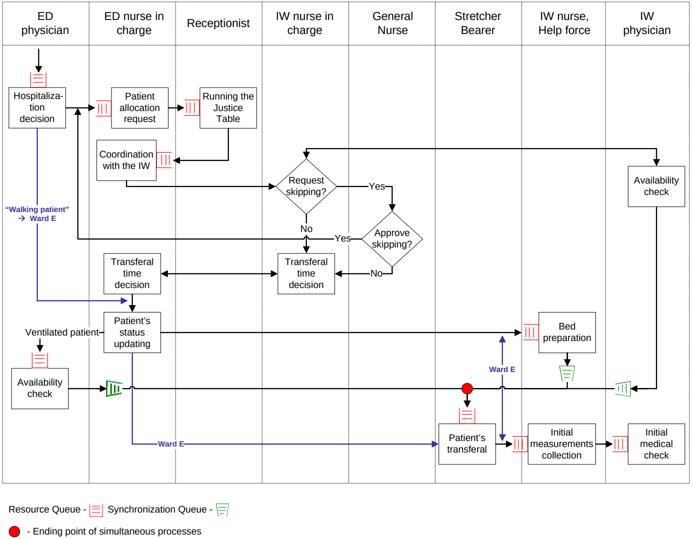
ED-to-IW Routing

- Wards A-D are more or less the same in their medical capabilities.
- Ward E treats only “walking” patients, and the routing to it from the ED is different.
- We focus on the routing process to wards A-D only.

Capacity (# beds) and ALOS:

	Ward A	Ward B	Ward C	Ward D
Capacity (# beds)	45	30	44	42
ALOS (days)	6.368	4.474	5.358	5.562

Integrated (Activities - Resources) Flow Chart

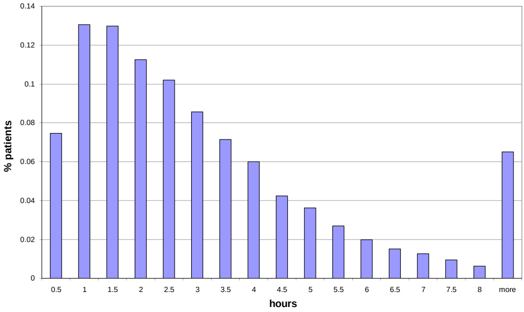


Problems in the ED-to-IW Process

- Waiting times in the ED for a transfer to the IWs could be long
- Patients' allocation to the IWs does not appear to be fair:
  - Staff - fairness:
    - Balance occupancy rates among the wards
    - Balance flux (number of patients per bed per time unit) among the wards
  - Patients - fairness:
    - Multi-queues vs. a single queue

Waiting Times

- Patients must often wait a long time in the ED until they are moved to their IW.
- From hospital database, average time from a decision of hospitalization till receiving a first treatment in a ward was 3.1 hours (for Wards A-D).



* Data refer to period: 1/05/06-30/10/08 (excluding 1-3/07).

IW Operational Measures

	Ward A	Ward B	Ward C	Ward D
ALOS (days)	6.368	4.474	5.358	5.562
Mean Occupancy Rate	97.8%	94.4%	86.8%	91.1%
Mean # Patients per Month	205.5	187.6	210.0	209.6
Standard capacity	45	30	44	42
Mean # Patients per Bed per Month	4.57	6.25	4.77	4.77
Return Rate (within 3 months)	16.4%	17.4%	19.2%	17.6%

* Data refer to period: 1/05/06-30/10/08 (excluding 1-3/07).

- The smallest + “fastest” ward is subject to the highest loads.
- The patients’ allocation appears unfair, as far as the wards are concerned.

Other Hospitals - Comparison Table

	Hosp.1	Hosp.2	Hosp.3	Hosp.4	Hosp.5	Anon.H
Number of IW	9	2	3	4	6	5
IW # beds	327	45	108	93	210	185
Average weekly # of transfers from ED to IW	525 (50%)	49 (14%)	266 (42%)	168 (26%)	469 (45%)	231 (22%)
Average weekly # of transfers per IW bed	1.606	1.089	2.463	1.806	2.233	1.249
IW Occupancy*	107.5%	118%	106.5%	116.4%	110%	93.8%
ED ALOS (hours)	2.2	6	2.83	6.8	2.5	4.2
IW ALOS (days)	3.9	3.9	3.5	6.1	3.5	5.2
Average waiting time in ED for IW (hours)	?	4	1	8	0.5	3
Wards differ?	yes	yes	no	yes	no	yes
Routing Policy	cyclical order	last digit of id	cyclical order	vacant bed	cyclical order**	cyclical order**

* Based on ynet article.

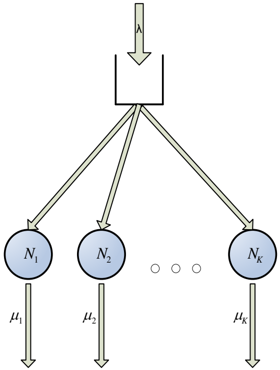
** Account for different patient types and ward capacities.

The ED-to-IW Process as a Queueing System

- Arrivals = patients to-be-hospitalized in the IW
- Pools = wards
- Service rates = 1/ALOS
- Servers in pool i = beds in ward i
- Arrivals to IW - Poisson process
- LOS in IW - exponentially distributed

Inverted-V Model ($\wedge$-model)

- Poisson arrivals with rate λ.
- K pools:
  - ⋆ Pool i consists of N_i i.i.d. exponential servers with service rates μ_i, $i=1,2,\ldots,K$;
  - ⋆ $\sum_{i=1}^{K} N_i = N$.
- One centralized waiting line:
  - ⋆ Infinite capacity;
  - ⋆ FCFS, non-preemptive, work-conserving.



The QED (Quality and Efficiency Driven) Asymptotic Regime

Definition (Informal):

- 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 (square root of the demand)

In our Hospital case:

- 30-50 servers (beds) in each pool (ward)
- Waiting times are order of magnitude shorter than service times: hours versus days
- Servers utilization (beds occupancy) is above 80%

Literature Review - "Slow Server Problem"

Rubinovitch M. - *The Slow Server Problem*
Journal of Applied Probability, vol. 22, pp. 205-213, 1983.

- System with two servers: fast and slow ($N = 2,\ \mu_1 > \mu_2$).
  - ⋆ *uninformed customers (Random Assignment - RA)*,
  - ⋆ *informed customers*,
  - ⋆ *partially informed customers*.
- For each case finds a critical number $\rho_c(\mu_1, \mu_2)$ such that if $\rho := \frac{\lambda}{\mu_1+\mu_2}$ is below ρ_c, the slow server should not be used, when one wishes to minimize the steady state *mean sojourn time* in the system.

Cabral F.B. - *The Slow Server Problem for Uninformed Customers*
Queueing Systems, vol. 50-4, pp. 353-370, 2005.

- Extends the analysis to N heterogeneous servers for the case with uninformed customers.

Literature Review - Dynamic Control

Armony M. - *Dynamic Routing in Large-Scale Service Systems with Heterogeneous Servers*

Queueing Systems, vol.51, pp. 287-329, 2005.

- **Fastest Servers First (FSF)** routing policy minimizes the steady state mean waiting time in the Quality and Efficiency Driven (QED) regime.

Atar R. - *Central Limit Theorem for a Many-Server Queue with Random Service Rates*

Ann. Appl. Probab., vol.18, no. 4, pp. 1548-1568, 2008.

- Analyzes FSF and **Longest-Idle Server First (LISF)** in a single-server pools model, where the number of servers and their service rates are random variables.

Literature Review - cont.

Armony M. and Ward A. - *Fair Dynamic Routing Policies in Large-Scale Systems with Heterogeneous Servers*
Operations Research, to appear.

- Propose a threshold policy that asymptotically achieves fixed server idleness ratios while minimizing the steady state mean waiting time.

Atar R., Shaki Y.Y. and Shwartz A. - *A Blind Policy for Equalizing Cumulative Idleness*
Manuscript under review, 2009.

- Propose **Longest Idle Pool First (LIPF)** routing policy that asymptotically balances cumulative idleness among the pools.

Gurvich I. and Whitt W. - *Queue-and-Idleness-Ratio Controls in Many-Server Service Systems*
Math. Oper. Res., vol.34, no.2, pp.363-396, 2009.

- For Parallel-Server Systems, propose *Queue-and-Idleness-Ratio* rules.

Randomized Most-Idle (RMI) Routing Policy

Define $\mathcal{I}_i(t)$ - number of idle servers in pool i at time t.

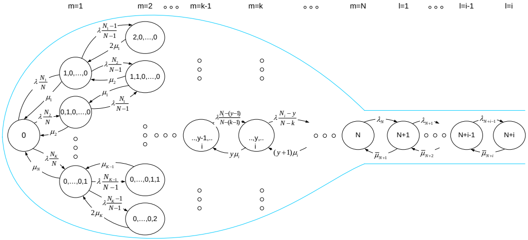
A customer arrives at time t.

- If $\exists i \in \{1, \ldots, K\} : \mathcal{I}_i(t) > 0$, the customer is routed to pool j with probability $\boxed{\frac{\mathcal{I}_j(t)}{\sum_{k=1}^{K} \mathcal{I}_k(t)}}$
  - ⋆ Equivalent to choosing a server out of all idle servers at random.
- Otherwise, the customer joins the queue (or leaves).

RMI is the only routing policy under which the $\wedge$-system forms a reversible MJP.

RMI Exact Analysis Summary

- General queue structure ("kite"):



- Steady-state performance measures calculation;
- Equivalence to a single-server-pools system under RA;
- Queue-length performance criterion - coupling proofs.
- Fast servers work less but serve more customers;

RMI Stationary Distribution

- $\mathcal{I}_i(t)$ - number of idle servers in pool i at time t.
- $\mathcal{I}(t)$ - total number of idle servers/customers awaiting service:
  - ⋆ $(\mathcal{I}(t))^+ = \sum_{i=1}^K \mathcal{I}_i(t)$
  - ⋆ $\{\mathcal{I}(t) = i\}$ for $i < 0$ - i customers awaiting service
- $\rho = \dfrac{\lambda}{\sum_{i=1}^K N_i \mu_i}$ - total traffic intensity

The process $\{(\mathcal{I}(t), \mathcal{I}_1(t)), \ldots, \mathcal{I}_K(t)),\ t \geq 0\}$ is a reversible continuous-time Markov chain with the stationary distribution π:

$$\pi(i, i_1, \ldots, i_K) = \begin{cases} \pi(0)\, i! \prod_{j=1}^{K} \binom{N_j}{i_j} (\mu_j/\lambda)^{i_j}, & i = \sum_{j=1}^{K} i_j \geq 0,\ 0 \leq i_j \leq N_j \\ \pi(0)\, (\rho)^{-i}, & i \leq 0,\ i_1 = \ldots = i_K = 0 \end{cases}$$

where

$$\pi(0) \equiv \pi(0, \ldots, 0) = \left(\frac{\rho}{1-\rho} + \sum_{i_1=0}^{N_1} \cdots \sum_{i_K=0}^{N_K} (i_1 + \cdots + i_K)! \prod_{j=1}^{K} \binom{N_j}{i_j} \left(\frac{\mu_j}{\lambda}\right)^{i_j} \right)^{-1}$$

RMI Exact Analysis - cont.

The $\wedge$-system under RMI routing policy is equivalent to a $\wedge$-system with N single-server pools:

- K server types:
  - N_i servers operate with rate μ_i ($\sum_{i=1}^{K} N_i = N$);
- *Random Assignment* routing policy.

Queue Length (Waiting Time) Criterion

- Under the optimality criterion of mean sojourn time in the system, sometimes it is better to discard the slow server.
- Alternative criterion: *mean waiting time* (*mean number of customers in queue*).
- Via an appropriate coupling, the queue length and waiting times in a system with N servers are path-wise dominated by the queue length and waiting times in a system with $N - 1$ servers.

Fast Servers vs. Slow Servers

- $\mathcal{I}_i$ - stationary number of idle servers in pool *i*.
- $\rho_i := 1 - \mathbb{E}\mathcal{I}_i/N_i$ - average steady-state occupancy rate in pool *i*.
- γ_i - average *flux* through pool *i* = average number of arrivals per server in pool *i* per time unit.
  - ⋆ $\gamma_i = \mu_i \rho_i$, by Little's law.

Theorem 1:

For any two pools i and j: if $\mu_i > \mu_j$, then

- $\rho_i < \rho_j$
- $\gamma_i > \gamma_j$

$\Rightarrow$ Faster servers work less but serve more than slower ones.

QED Scaling

Define:

- $c_i^\lambda = N_i^\lambda \mu_i$ - service capacity of pool *i*
- $c^\lambda = \sum_{i=1}^K c_i^\lambda$ - total service capacity

[Armony M., 2005]: Take $\lambda \to \infty$ such that:

$$\lim_{\lambda\to\infty} \frac{\sum_{i=1}^K c_i^\lambda - \lambda}{\sqrt{\lambda}} = \delta \quad (or\ c^\lambda = \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda}),\ \text{as } \lambda\to\infty)$$

$$\lim_{\lambda\to\infty} \frac{c_i^\lambda}{c^\lambda} = a_i \ (i=1,2,\ldots,K) \ - \ \text{prop. of service capacity of pool i}$$

Also define:

- $\mu := \left(\sum_{i=1}^K \frac{a_i}{\mu_i}\right)^{-1}, \quad \hat{\mu} := \sum_{i=1}^K a_i\mu_i$
- $\lim_{\lambda\to\infty} \frac{N_i^\lambda}{N^\lambda} = \frac{a_i}{\mu_i}\mu := q_i, \qquad i=1,2,\ldots,K$

RMI: QED Analysis

$\mathcal{I}^\lambda$ - stationary total number of idle servers/customers awaiting service in the system with arrival rate λ:

- $\star$ $(\mathcal{I}^\lambda)^+ = \sum_{i=1}^{K} \mathcal{I}_i^\lambda$
- $\star$ $\{\mathcal{I}^\lambda = i\}$ for $i < 0$ - i customers awaiting service

Theorem 2 (Informal):

- Approximation of performance measures (delay probability, etc)
- **Dimensionality Reduction (DR)**: $\mathcal{I}_i^\lambda \approx a_i(\mathcal{I}^\lambda)^+$ *as* $\lambda{\to}\infty$

$$\Rightarrow \frac{\mathcal{I}_i^\lambda}{\mathcal{I}_j^\lambda} \approx \frac{a_i}{a_j} \quad \textit{as } \lambda{\to}\infty$$

- Characterization of the system behavior on the *sub-diffusion* ($\sqrt[4]{\lambda}$) scale; $\sqrt[4]{\lambda}$-deviations of $\mathcal{I}_i^\lambda$ around $a_i(\mathcal{I}^\lambda)^+$

RMI: QED Analysis - cont.

Theorem 2:

Let $\hat{\mathcal{I}}^\lambda = \mathcal{I}^\lambda/\sqrt{\lambda}$ and $\hat{\mathcal{I}}_i^\lambda = \frac{1}{\sqrt{\mathcal{I}^\lambda}}\left(\mathcal{I}_i^\lambda - \frac{N_i^\lambda \mu_i}{\sum_{i=1}^K N_i^\lambda \mu_i}\mathcal{I}^\lambda\right)$, $i=1,\dots,K$.
Then, as $\lambda \to \infty$,

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

where:

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

Delay Probability Approximation

If $\mu_1 = \mu_2 = \ldots = \mu_K$:

Then $\mu = \hat{\mu} = \mu_1$, $\delta/\sqrt{\hat{\mu}} = \beta$ and $\mathbb{P}[\hat{\mathcal{I}} \leq 0] = \left(1 + \beta\,\frac{\Phi(\beta)}{\varphi(\beta)}\right)^{-1}$

$\Rightarrow$ Consistent with Erlang-C Approximation [S. Halfin and W. Whitt, 1981].

Example: exact values vs. QED approximations

$K = 2$

$q_2 = 2q_1 = 2/3$

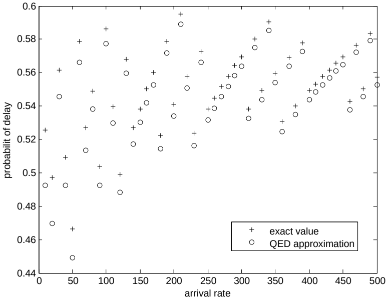
$\mu_1 = 2\mu_2 = 2$

$\delta = 0.5$

λ : $10 - 500$

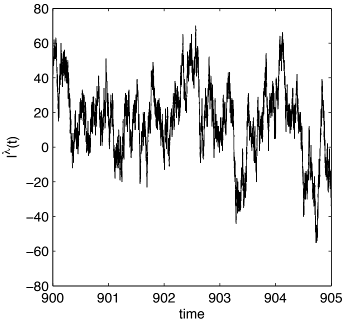
N_1^λ : $3 - 128$

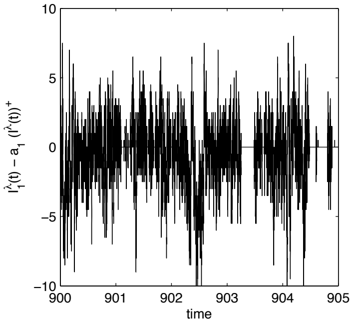
N_2^λ : $6 - 256$.



Dimensionality Reduction Illustration

- $K = 2,\ \lambda = 3950,\ \mu_1 = 15,\ \mu_2 = 7.5,\ N_1 = 138,\ N_2 = 276$ $(\delta = 3,\ a_1 = a_2 = 1/2)$
- $\{\mathcal{I}^\lambda(t),\ t \geq 0\}$ evolve on $\sqrt{\lambda}$-scale ($\sqrt{\lambda} \approx 62.8$)
- $\{\mathcal{I}_1^\lambda(t) - a_1(\mathcal{I}^\lambda(t))^+,\ t \geq 0\}$ evolve on $\sqrt[4]{\lambda}$-scale ($\sqrt[4]{\lambda} \approx 7.93$)





Fair Routing Criteria

Occupancy balancing

⋆ Idleness-criterion: compare the *idleness ratios* $\dfrac{1-\rho_i^\lambda}{1-\rho_j^\lambda}$

Flux balancing

⋆ Flux-criterion: compare the *flux ratios* $\dfrac{\gamma_i^\lambda}{\gamma_j^\lambda} = \dfrac{\mu_i \rho_i^\lambda}{\mu_j \rho_j^\lambda}$

In the QED regime $\lim\limits_{\lambda\to\infty} \dfrac{\gamma_i^\lambda}{\gamma_j^\lambda} = \dfrac{\mu_i}{\mu_j} \Rightarrow$ strive for $\rho_i^\lambda < \rho_j^\lambda$ if $\mu_i > \mu_j$.

In RMI - from Theorem 2:

- $\dfrac{\mathcal{I}_i^\lambda(t)}{\mathcal{I}_j^\lambda(t)} \to \dfrac{a_i(\mathcal{I}^\lambda(t))^+}{a_j(\mathcal{I}^\lambda(t))^+} = \dfrac{a_i}{a_j}$, thus
- $\dfrac{1-\rho_i^\lambda}{1-\rho_j^\lambda} = \dfrac{\mathbb{E}\mathcal{I}_i^\lambda}{N_i^\lambda} \dfrac{N_j^\lambda}{\mathbb{E}\mathcal{I}_j^\lambda} \to \dfrac{a_i q_j}{a_j q_i} = \dfrac{\mu_i}{\mu_j}$

Longest-Idle Server First (LISF) Routing Policy

- **LISF** policy routes a customer to the server that has been idle for the longest time, among all idle servers.
- Atar (2008), Armony and Ward (2008) show that, asymptotically (as $\lambda \to \infty$):
  - $\star$ $\dfrac{\mathcal{I}_i^\lambda(t)}{\mathcal{I}_j^\lambda(t)} \to \dfrac{a_i(\mathcal{I}^\lambda(t))^+}{a_j(\mathcal{I}^\lambda(t))^+} = \dfrac{a_i}{a_j}$, thus
  - $\star$ $\dfrac{1-\rho_i^\lambda}{1-\rho_j^\lambda} = \dfrac{\mathbb{E}\mathcal{I}_i^\lambda}{N_i^\lambda}\dfrac{N_j^\lambda}{\mathbb{E}\mathcal{I}_j^\lambda} \to \dfrac{a_i q_j}{a_j q_i} = \dfrac{\mu_i}{\mu_j}$

$\Rightarrow$ LISF and RMI are equivalent on the diffusion scale.

$\ast$ LISF requires more information than RMI.

Idleness-Ratio (IR) Routing Policy

IR policy, a special case of QIR policies (Gurvich and Whitt (2008)) , routes an arriving customer to the pool with the highest idleness imbalance:

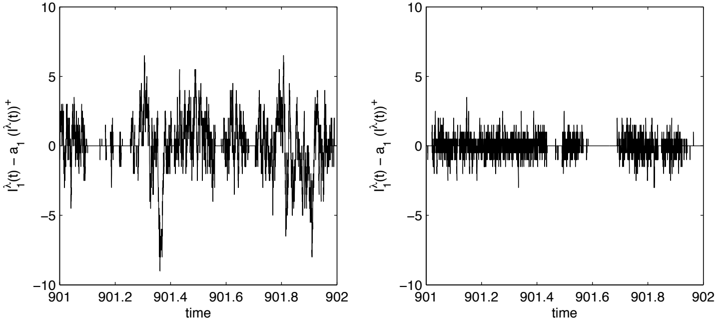
- Introduce a weight vector $(w_1, w_2, \ldots, w_K),\ w_i > 0,\ \sum_{i=1}^{K} w_i = 1$.
- A customer arriving at time t is routed to pool $\arg\max\{\mathcal{I}_i^\lambda(t-) - w_i(\mathcal{I}^\lambda(t-))^+\}$
- Asymptotically (as $\lambda \to \infty$): $\frac{1-\rho_i^\lambda}{1-\rho_j^\lambda} = \frac{\mathbb{E}\mathcal{I}_i^\lambda}{\mathbb{E}\mathcal{I}_j^\lambda}\frac{N_j^\lambda}{N_i^\lambda} \to \frac{w_i q_j}{w_j q_i}$

$\Rightarrow$ If $w_i = a_i$, IR and RMI are equivalent on the diffusion scale.

$*$ IR requires more information than RMI - for determining a_i's.

RMI versus IR: Sub-diffusion Scale

Typical sample paths of $\mathcal{I}_1^\lambda(t) - a_1(\mathcal{I}^\lambda(t))^+,\ t \geq 0$:



ED-to-IW: $\sqrt[4]{\lambda} \approx 2.3$

Partial-information Routing - Simulation Analysis

- RMI requires the information on the number of available beds at each ward at the moment of routing.
- The occupancy status in the IWs is not available on a real-time basis; instead, the ED relies on one bed census update per day.
- It is necessary to estimate the system state at the decision time, based on the system state at the last update time point.

Joint project with A. Zviran

- Create a computer simulation model of the ED-to-IW process in Anonymous Hospital.
- Examine various routing policies, while accounting for *availability of information* in the system.

Simulations

Summary of Results:

- *Weighted Algorithm* - minimizes at each decision point a convex combination of the two conflicting demands: balanced occupancy rates and balanced flux.
- Implementation in *partial information access systems* results in almost no worsening in performance.

Estimating occupancy:

- M_j - number of occupied beds in ward j; updated at time point T.
- Number of occupied beds in ward j at time t = $\max\{M_j - M_j \cdot \mu_j \cdot (t - T), 0\}, \forall j \in \{1, \ldots, 4\}$.
- $M_k = M_k + 1$, after routing to ward k.

Contribution

- Modeling ED-to-IW process: an important phase of patients' flow in hospitals
- Data analysis of the ED-to-IW process
- Quantify operational fairness
- Propose a practical routing algorithm - RMI
- Analyze RMI: in steady-state and in the QED regime (sub-diffusion insights)
- Compare RMI to LISF and IR: RMI results in the same server fairness but requires less information.

Future Research

- Extend theoretical analysis to several customer (patient) classes
- Include Ward E in the theoretical study
- Model hospital staff: two-scale (doctors/nurses and beds) model
- Attempt to capture possible dependency between the routing algorithm and service rates
- Psychological study: waiting time versus sojourn time criterion

Thank You!