Queueing Systems with Heterogeneous Servers: On Fair Patients’ Routing from the ED to IW

Yulia Tseytlin

Technion - Israel Institute of Technology

February 4, 2009

Advisor: Prof. Avishai Mandelbaum
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 queue-architectures and routing policies, in search for fairness and good operational performance.
Outline

Practical Background
   Hospital, ED and IW
   “ED-to-IW" Routing

RMI Routing Policy
   Introduction
   “Slow Server Problem"
   Exact Analysis
   Asymptotic Analysis

Additional Results
   Alternative Routing Policies
      MI
      WMI
   Distributed Finite Queues
   Joint Projects
   Summary and Future Research
The Process of Interest

- 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.
Internal Wards

- **Wards A-D** are more or less the same in their medical capabilities - each can treat multiple types of patients.
- **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.

**Standard and Maximal Capacity (# beds):**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard capacity</td>
<td>45</td>
<td>30</td>
<td>44</td>
<td>47</td>
<td>24</td>
</tr>
<tr>
<td>Maximal capacity</td>
<td>52</td>
<td>35</td>
<td>46</td>
<td>48</td>
<td>27</td>
</tr>
<tr>
<td>Max. to standard ratio</td>
<td>115%</td>
<td>116%</td>
<td>104%</td>
<td>102%</td>
<td>113%</td>
</tr>
</tbody>
</table>
Integrated (Activities - Resources) Flow Chart
The “Justice Table”

- The “Justice Table” is a computer program that determines routing.
- Its goal is to balance the load among the wards, thus making the patients’ allocation fair towards the wards.
- Prior to routing, patients are classified into three categories: ventilated, special-care and regular.
- For each patients’ category there are “fixed turns” among the wards, while accounting for standard capacities.
- The Justice Table does not take into account the actual number of occupied beds and patients’ discharge rate.
IW Operational Measures:

<table>
<thead>
<tr>
<th></th>
<th>Ward A</th>
<th>Ward B</th>
<th>Ward C</th>
<th>Ward D</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALOS (days)</td>
<td>6.318</td>
<td>4.574</td>
<td>5.446</td>
<td>5.642</td>
</tr>
<tr>
<td>Mean Occupancy Rate</td>
<td>98.7%</td>
<td>98.9%</td>
<td>87.9%</td>
<td>84.1%</td>
</tr>
<tr>
<td>Mean # Patients per Year</td>
<td>2,534</td>
<td>2,351</td>
<td>2,558</td>
<td>2,578</td>
</tr>
<tr>
<td>Standard capacity</td>
<td>45</td>
<td>30</td>
<td>44</td>
<td>47</td>
</tr>
<tr>
<td>Mean # Patients per Bed</td>
<td>56.3</td>
<td>78.4</td>
<td>58.1</td>
<td>54.9</td>
</tr>
<tr>
<td>Return Rate</td>
<td>15.4%</td>
<td>15.6%</td>
<td>16.2%</td>
<td>14.8%</td>
</tr>
</tbody>
</table>

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

- Patients must often wait a long time in the ED until they are moved to their IW.
- For 182 observations conducted in May 2007, average waiting time was 97 minutes.
Other Hospitals - Comparison Table

<table>
<thead>
<tr>
<th></th>
<th>Hosp.1</th>
<th>Hosp.2</th>
<th>Hosp.3</th>
<th>Hosp.4</th>
<th>Hosp.5</th>
<th>Anon.H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average daily no’ of arrivals to Internal ED</td>
<td>150</td>
<td>50</td>
<td>91</td>
<td>90</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>Average daily % of transfers from ED to IW</td>
<td>50%</td>
<td>14%</td>
<td>42%</td>
<td>26%</td>
<td>45%</td>
<td>20%</td>
</tr>
<tr>
<td>Number of IW</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Average waiting time in ED for IW (hours)</td>
<td>?</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Wards differ?</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Routing Policy</td>
<td>fixed turns</td>
<td>last digit of id</td>
<td>fixed turns</td>
<td>vacant bed</td>
<td>fixed turns*</td>
<td>fixed turns*</td>
</tr>
</tbody>
</table>

* Account for different patients’ types and ward capacities.
The ED-to-IW Process as a Queueing System

- Pools = wards;
- Service rates = 1/ALOS;
- Servers in pool $i$ = beds in ward $i$ (number of service providers is proportional to standard capacity);
- Arrivals to IW - Poisson process;
- LOS in IW - exponentially distributed.
Inverted-V Model ($\land$-model)

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

Armony M.

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


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

Armony M., Ward A.

*Fair Dynamic Routing Policies in Large-Scale Systems with Heterogeneous Servers*


- Propose a threshold policy that asymptotically achieves fixed server idleness ratios while minimizing the steady state mean waiting time.
Randomized Most-Idle (RMI) Routing Policy

Define $I_i(t)$ - number of idle servers in pool $i$ at time $t$.

A customer arrives at time $t$.

- If $\exists i \in \{1, \ldots, K\} : I_i(t) > 0$, the customer is routed to pool $i$ with probability
  $$\frac{I_i(t)}{\sum_{j=1}^{K} I_j(t)}$$

- Otherwise, the customer joins the queue (or leaves).

The $\land$-system presented before, under RMI routing policy, is equivalent to a $\land$-system with $N$ single-server pools:

- $K$ server types:
  - $N_i$ servers operate with rate $\mu_i$ ($\sum_{i=1}^{K} N_i = N$);
- Random Assignment routing policy.
∧-System with Single-Server Pools

- Poisson arrivals with rate $\lambda$.
- $N$ i.i.d. exponential servers with service rates $\mu_i$, $i=1,2,...,N$.
- One waiting line with infinite capacity.
“Slow Server Problem"

Find the best operating policy in order to minimize the steady state mean sojourn time of the customers in the system (or mean number of customers in the system).

Literature Review

Rubinovitch M.

*The Slow Server Problem*


Cabral F.B.

*The Slow Server Problem for Uninformed Customers*

Queueing Systems, vol. 50-4, pp. 353-370, 2005
Literature Review

Rubinovitch M., 1983

- System with two servers: fast and slow ($N = 2, \mu_1 > \mu_2$).
- Three different scenarios:
  - uninformed customers (Random Assignment),
  - 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 $\rho_c$, the slow server should not be used.

Cabral F.B., 2005

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

- We prove that, 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, when both systems operate under a Random Assignment policy.

- Hence, each server that we add to the system (even a very slow one) reduces queue length.
RMI Stationary Analysis

- RMI is the only routing policy under which the $\land$-system forms a reversible MJP.
  \[ \pi_i q_{ij} = \pi_j q_{ji} \quad \forall i,j \in S. \]
- We present here a Loss model (no queue possible); analysis of Delay models easily follows.

Stationary Distribution

- System states: \( y = (y_1, y_2, \ldots, y_K) \),
- \( y_i \) - number of busy servers in pool \( i \) (\( y_i \in \{0, 1, \ldots, N_i\} \))
- \( m_y = \sum_{i=1}^{K} y_i \) - total number of busy servers at state \( y \).
Stationary Distribution

$$\pi_y = \pi_0 \frac{\prod_{i=1}^{K} \binom{N_i}{y_i}}{\binom{N}{m_y}} \frac{\lambda^{m_y}}{m_y! \prod_{i=1}^{K} \mu_i^{y_i}}$$

$$y_i \in \{0, 1, ..., N_i\}, \ i \in \{1, 2, ..., K\}$$

$$\pi_0 = \left[ \sum_{y_1=0}^{N_1} \cdots \sum_{y_K=0}^{N_K} \frac{\prod_{i=1}^{K} \binom{N_i}{y_i}}{\binom{N}{m_y}} \frac{\lambda^{m_y}}{m_y! \prod_{i=1}^{K} \mu_i^{y_i}} \right]^{-1}$$
RMI Properties

Definitions:

- $\tilde{\rho}_i$ - stationary occupancy rate in pool $i$
- $\bar{\rho}_i$ - average occupancy rate in pool $i$
- $\gamma_i$ - average flux through pool $i$ = average number of arrivals per server in pool $i$ per time unit

$\gamma_i = \mu_i \bar{\rho}_i$, by Little’s law.

Proposition:

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

- $\bar{\rho}_i < \bar{\rho}_j$
- $\gamma_i > \gamma_j$
- Conjecture: $\tilde{\rho}_i \leq_{st} \tilde{\rho}_j \quad (\mathbb{P}(\tilde{\rho}_i > x) \leq \mathbb{P}(\tilde{\rho}_j > x) \forall x \in (0, 1))$
The QED (Quality and Efficiency Driven) Asymptotic Regime
Definition (Informal) [Armony M., 2005]:

- A system with a large volume of arrivals and many servers.
- The delay probability is neither near 0 nor near 1 (quality aspect).
- Total service capacity is equal to the demand plus a safety capacity, which is of the same order of magnitude as the square root of the demand (efficiency aspect).

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%.
- The probability that no server (bed) is available is neither near 0 nor near 1.
QED Limits

[Armony M., 2005]

We take $\lambda \to \infty$ such that the following limits hold:

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

$$\lim_{\lambda \to \infty} \frac{N_i^\lambda \mu_i}{\lambda} = a_i \quad (\text{or} \quad N_i = a_i \frac{\lambda}{\mu_i} + o(\lambda), \text{as } \lambda \to \infty), \quad i = 1, 2, \ldots, K$$

Define $\mu := \left( \sum_{i=1}^{K} \frac{a_i}{\mu_i} \right)^{-1}$. Then

$$\lim_{\lambda \to \infty} \frac{N_i^\lambda}{N^\lambda} = a_i \frac{\mu}{\mu_i} \mu := q_i \quad i = 1, 2, \ldots, K$$
Loss Probability: \( K = 2 \) Pools

Steady-state blocking probability:

\[
P_{\lambda}(\text{block}) = \pi_0^{\lambda} \cdot \frac{\lambda^N}{N! \mu_1^{N_1} \mu_2^{N_2}} = \frac{\lambda^N}{N! \mu_1^{N_1} \mu_2^{N_2}} \sum_{y_1=0}^{N_1} \sum_{y_2=0}^{N_2} \frac{(N_1)! (N_2)!}{(N_1+y_1)! (N_2+y_2)!} \frac{\lambda^{y_1+y_2}}{(\mu_1^{y_1} \mu_2^{y_2})(\mu_1 \mu_2)^{y_1+y_2}}
\]
Loss Probability Approximation

P. Momcilovic proved

\[ \lim_{\lambda \to \infty} \sqrt{\lambda} P_\lambda(block) = \sqrt{\hat{\mu}} \frac{\varphi(\delta / \sqrt{\hat{\mu}})}{\Phi(\delta / \sqrt{\hat{\mu}})} \]

where:

- \( \hat{\mu} := \mu_1 a_1 + \mu_2 a_2 \)
- \( \varphi(\cdot), \Phi(\cdot) \) - density and probability functions of \( \text{Norm}(0,1) \)

Using \( \lim_{\lambda \to \infty} \frac{\lambda}{N\mu} = 1 \), we deduce:

\[ \lim_{\lambda \to \infty} \sqrt{N} P_\lambda(block) = \sqrt{\hat{\mu}} \frac{\varphi(\delta / \sqrt{\hat{\mu}})}{\mu \Phi(\delta / \sqrt{\hat{\mu}})} \]
Loss Probability Approximation

If $\mu_1 = \mu_2$:
Then $\mu = \hat{\mu} = \mu_1 = \mu_2$

$$
\lim_{\lambda \to \infty} \sqrt{N} P_\lambda(block) = \frac{\varphi(\delta/\sqrt{\mu})}{\Phi(\delta/\sqrt{\mu})} = \frac{\varphi(\beta)}{\Phi(\beta)}
$$

where $\beta = \lim_{N \to \infty} \sqrt{N}(1 - \frac{\lambda}{N}\mu)$.

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

Insights:

- $\sqrt{\lambda} P_\lambda(block)$ is a function of three parameters: $\delta$, $\mu$ and $\hat{\mu}$:
  - As $\lambda \to \infty$, $a_i$ = proportion of customers served by pool $i$, $q_i$ = proportion of servers from pool $i$.
  - $\mu := \left(\frac{a_1}{\mu_1} + \frac{a_2}{\mu_2}\right)^{-1} = q_1\mu_1 + q_2\mu_2$
  - $\hat{\mu} := \mu_1 a_1 + \mu_2 a_2$

- $P_\lambda(block)$ is an order of magnitude of $1/\sqrt{\lambda}$. 
State-Space Collapse

P. Momcilovic finds:

Denote $I_i^\lambda$ - stationary number of idle servers in pool $i$, $i = 1, 2$. Given that $I_1^\lambda + I_2^\lambda = \gamma \sqrt{\lambda}$, $I_1^\lambda$ and $I_2^\lambda$ deviate from $a_1 \gamma \sqrt{\lambda}$ and $a_2 \gamma \sqrt{\lambda}$ by $\Xi \sqrt{\lambda}$, where $\Xi \Rightarrow \text{Norm}(0, \gamma a_1 a_2)$ as $\lambda \to \infty$.

Hence $a_2 I_1^\lambda \approx a_1 I_2^\lambda$ as $\lambda \to \infty$.

$\lambda = 3950$

$\downarrow$

$\sqrt[4]{\lambda} \approx 8$

$\mu_1 = 15$, $\mu_2 = 7.5$

$N_1 = 138$, $N_2 = 276$
RMI Routing Policy enjoys some desirable properties, but is problematic for a hospital environment due to its randomness.

The intuitive non-random equivalent to RMI is **MI (Most-Idle)** - routing an arriving customer to the most vacant pool (the one with maximal number of idle servers).

Asymptotically (as $N \to \infty$): $I_1 \approx I_2$.

Thus: $\tilde{\rho}_i = \frac{N_i - I_i}{N_i} = 1 - \frac{I_i}{N_i}$, i.e., larger pools (bigger $N_i$) have higher occupancy rates.
Comparison criteria

Fairness towards servers:

- **Idle-ratio** - ratio between proportion of idle servers in the pools:
  \[
  \frac{I_1/N_1}{I_2/N_2} = \frac{1 - \bar{\rho}_1}{1 - \bar{\rho}_2}.
  \]

- **Flux-ratio** - ratio between flux through the pools ("flux" - number of arrivals per server per time unit):
  \[
  \frac{\gamma_1}{\gamma_2} = \frac{\bar{\rho}_1 \mu_1}{\bar{\rho}_2 \mu_2}.
  \]

The closer the ratio is to 1, the more balanced the routing is.

Operational performance:

- Steady-state probability of loss, or \( P(\text{Block}) \).
General observations

- **RMI:** from Stace-Space Collapse follows that:

\[
\frac{1 - \bar{\rho}_1}{1 - \bar{\rho}_2} = \frac{I_1/N_1}{I_2/N_2} \approx \frac{N_2a_1}{N_1a_2} \approx \frac{q_2a_1}{q_1a_2} = \frac{\mu_1}{\mu_2}
\]

→ Idle-ratio depends only on service rates.

- **MI:**

\[
\frac{1 - \bar{\rho}_1}{1 - \bar{\rho}_2} = \frac{I_1/N_1}{I_2/N_2} \approx \frac{N_2}{N_1} \approx \frac{q_2}{q_1}
\]

→ Idle-ratio depends only on pool capacities.
Comparison: RMI versus MI

<table>
<thead>
<tr>
<th>$q_1 = q_2$</th>
<th>Idle-ratio</th>
<th>Flux-ratio</th>
<th>$\mathbb{P}(\text{Block})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI</td>
<td>RMI</td>
<td>MI</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$q_1 &gt; q_2$</th>
<th>$\frac{\mu_1}{\mu_2} &lt; \frac{q_1}{q_2}$</th>
<th>RMI</th>
<th>equal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{\mu_1}{\mu_2} = \frac{q_1}{q_2}$</td>
<td>equal</td>
<td>RMI</td>
<td>MI</td>
</tr>
<tr>
<td>$\frac{\mu_1}{\mu_2} &gt; \frac{q_1}{q_2}$</td>
<td>MI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$q_1 &lt; q_2$</th>
<th>$a_1 &lt; a_2$</th>
<th>RMI</th>
<th>MI</th>
<th>RMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1 = a_2$</td>
<td>equal</td>
<td>equal</td>
<td>equal</td>
<td>equal</td>
</tr>
<tr>
<td>$a_1 &gt; a_2$</td>
<td>MI</td>
<td>RMI</td>
<td>MI</td>
<td></td>
</tr>
</tbody>
</table>

- RMI and MI are not equivalent.
- For different sets of parameters and different target functions, a different policy is superior.
WMI Routing Policy

We propose **WMI** (*Weighted Most-Idle*) Routing Policy - routing an arriving customer to the pool where the number of idle servers multiplied by the pool’s weight is maximal.

Formally,

- Introduce a weight vector
  \[ (w_1, w_2), \; w_i \in (0, 1), \; w_1 + w_2 = 1. \]
  
- A customer arriving at time \( t \) is routed to pool
  \[ i = \arg\max \{w_1 I_1, w_2 I_2\}. \]

- Asymptotically (as \( N \to \infty \)): \( w_1 I_1 \approx w_2 I_2. \)
WMI Routing Policy

Interesting cases:

- $w_1 = w_2 = 1/2$
  - MI routing policy.

- $w_1 = a_2, w_2 = a_1$
  - Non-random Equivalent to RMI - NERMI routing policy.

- $w_1 = q_2, w_2 = q_1$
  - Idleness-Balancing - IB policy: routing an arriving customer to the least utilized pool (pool with the minimal occupancy rate).
## Comparison: WMI versus RMI

<table>
<thead>
<tr>
<th>w_1 \cdot q_1 = w_2 \cdot q_2</th>
<th>\text{Idle-ratio}</th>
<th>\text{Flux-ratio}</th>
<th>P(\text{Block})</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMI</td>
<td>RMI</td>
<td>WMI</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>w_1 \cdot q_1 &gt; w_2 \cdot q_2</th>
<th>w_1 \cdot a_1 &lt; w_2 \cdot a_2</th>
<th>\mu_1 &lt; \frac{w_1 \cdot q_1}{w_2 \cdot q_2}</th>
<th>RMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>w_1 \cdot q_1 &gt; w_2 \cdot q_2</td>
<td>w_1 \cdot a_1 = w_2 \cdot a_2</td>
<td>\mu_1 = \frac{w_1 \cdot q_1}{w_2 \cdot q_2}</td>
<td>equal</td>
</tr>
<tr>
<td>w_1 \cdot q_1 &gt; w_2 \cdot q_2</td>
<td>w_1 \cdot a_1 &gt; w_2 \cdot a_2</td>
<td>\mu_1 &gt; \frac{w_1 \cdot q_1}{w_2 \cdot q_2}</td>
<td>RMI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>w_1 \cdot q_1 &lt; w_2 \cdot q_2</th>
<th>\mu_1 &lt; \frac{w_1 \cdot q_1}{w_2 \cdot q_2}</th>
<th>\mu_1 = \frac{w_1 \cdot q_1}{w_2 \cdot q_2}</th>
<th>equal</th>
</tr>
</thead>
<tbody>
<tr>
<td>w_1 \cdot a_1 &lt; w_2 \cdot a_2</td>
<td>\mu_1 &gt; \frac{w_1 \cdot q_1}{w_2 \cdot q_2}</td>
<td>\mu_1 &gt; \frac{w_1 \cdot q_1}{w_2 \cdot q_2}</td>
<td>RMI</td>
</tr>
</tbody>
</table>

w_1 \cdot q_1 = w_2 \cdot q_2

\mu_1 \cdot w_1 \cdot q_1 = \mu_2 \cdot w_2 \cdot q_2
NERMI versus RMI

\[ I_1(t) - a_1 I(t) \]
Distributed Finite Queues

- Poisson arrivals with rate $\lambda$.
- $K$ pools: pool $i$ has
  - $N_i$ i.i.d. exponential servers with service rates $\mu_i$, $i=1,2,...,K$.
  - $\sum_{i=1}^{K} N_i = N$
  - Waiting line with finite capacity $b_i$, $\sum_{i=1}^{K} b_i = b$
RMI Routing Policy for Distributed Queues

Define

- $\mathcal{I}_i(t)$ - number of idle servers in pool $i$ at time $t$.
- $E_i(t)$ - number of empty places in buffer of pool $i$ at time $t$.
- $V_i(t) = \mathcal{I}_i(t) + E_i(t)$ - number of total vacant places in pool $i$ at time $t$.

A customer arrives at time $t$.

- If $\exists i \in \{1, \ldots, K\} : V_i(t) > 0$, the customer is routed to pool $i$ with probability $\frac{V_i(t)}{\sum_{j=1}^{K} V_j(t)}$.

- Otherwise, the customer leaves (or joins the centralized queue).
Stationary Analysis

- RMI is the only routing policy under which the distributed-finite-queues system forms a reversible MJP.

**Stationary Distribution: Case of K=2 Pools**

- $y_i$ - number of customers in pool $i$ ($y_i \in \{0, 1, \ldots, N_i + b_i\}$)
- $m_y = y_1 + y_2$ - total number of customers at state $y$.

\[
\pi(y_1, y_2) = \pi_0 \frac{\binom{N+b-m_y}{N_1+b_1-y_1} N_1^{-(y_1-N_1)^+} N_2^{-(y_2-N_2)^+}}{\binom{N+b}{N_1+b_1} (N_1^\wedge y_1)! (N_2^\wedge y_2)!} \frac{\lambda^m y}{\mu_1 y_1 \mu_2 y_2}
\]
Simulations

Joint project with A. Zviran in “System Analysis and Design" course

- Create a computer simulation model of the ED-to-IW process in Anonymous Hospital.
- Define various fairness and performance measures to form a single *integrated criterion of quality*.
- Examine various routing policies, while accounting for *availability of information* in the system.
- Evaluate the policies according to the optimality criteria.
Simulations

Summary of Results:

- **Occupancy Balancing Algorithm** - balances ward occupancies in each moment of routing.
- **Flow Balancing Algorithm** - keeps number of patients per bed per year equal among the wards.
- **Weighted Algorithm** - combines these two methods: achieves both fairness for the staff and good operational performance.
- Implementation in *partial information access systems* results in almost no worsening in performance.
Empirical Project

Joint project with Mandelbaum A., Marmor Y., Yom-Tov G.

- Analyze ED, IW and their interface, using simulations, empirical and theoretical models.
- Example of interesting research questions:
  - LOS analysis (both in the ED and in the IW):
    - Why is their distribution LogNormal?
    - Do LOS depend on “load”?
  - Is the real system QED?
  - Can we model waiting times as a function of the load on the wards?

Research is conducted within the OCR research project of Technion + IBM + Rambam, under the funding of IBM.
Summary

- Motivated by the process of patients’ routing from ED to IW’s, we study queueing systems with heterogeneous servers.
- For Inverted-V system we propose the RMI routing policy. We analyze the system in closed form and show its various properties.
- We compare the RMI policy to its non-random alternatives MI and WMI policies in the QED regime, with help of simulations.
- For distributed finite queues we propose the equivalent to the RMI policy and analyze the system in closed form.
Future Research

To be done:

• **Games Theory**: apply costs sharing approach in order to find to which extent each ward is “responsible" for some cost function (e.g., patients’ waiting time).

General Ideas:

• Extend the QED asymptotic analysis to more than 2 server pools.
• Find QED approximations for RMI in distributed queues.
• Psychological study: which criterion matters more for customers: waiting time or sojourn time?
Thank You!