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.

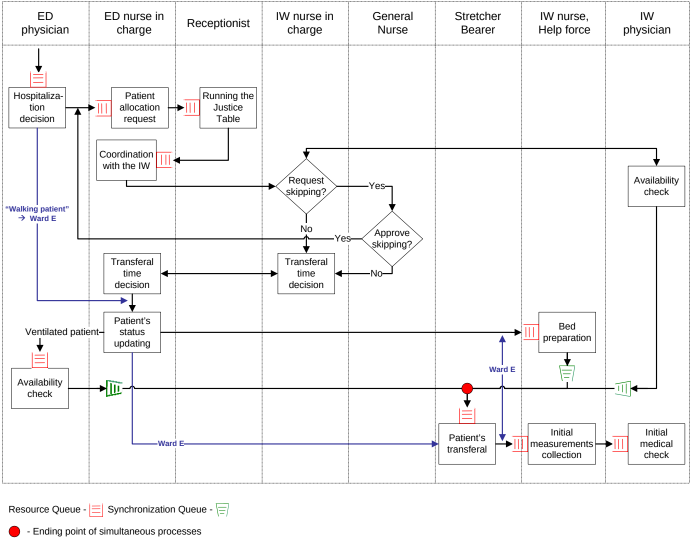
Internal Wards

- 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):

	Ward A	Ward B	Ward C	Ward D	Ward E
Standard capacity	45	30	44	47	24
Maximal capacity	52	35	46	48	27
Max. to standard ratio	115%	116%	104%	102%	113%

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.

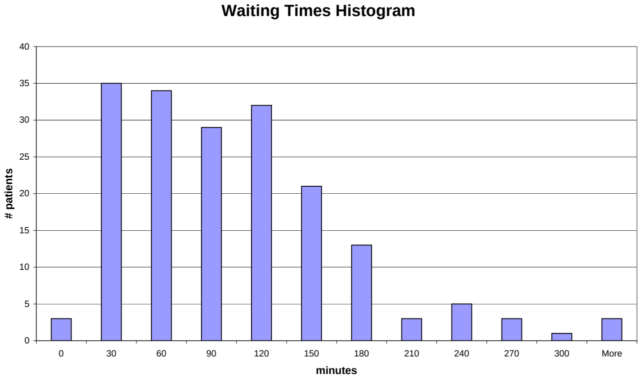
IW Operational Measures:

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

- **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

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

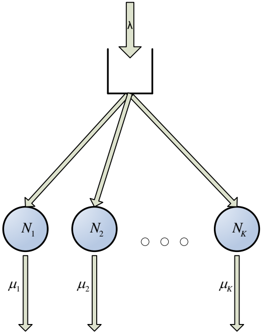
* Account for different patients' types and ward capacities.

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 ($\wedge$-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.



Literature Review

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.

Armony M., Ward A.
Fair Dynamic Routing Policies in Large-Scale Systems with Heterogeneous Servers
Manuscript under review, 2007.

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

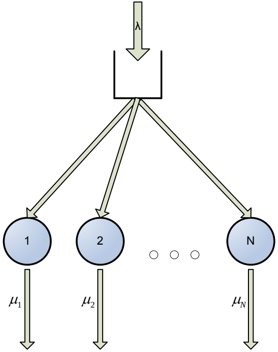
- If $\exists i \in \{1, \ldots, K\} : \mathcal{I}_i(t) > 0$, the customer is routed to pool i with probability $\boxed{\frac{\mathcal{I}_i(t)}{\sum_{j=1}^{K} \mathcal{I}_j(t)}}$
- Otherwise, the customer joins the queue (or leaves).

The $\wedge$-system presented before, 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.

$\wedge$-System with Single-Server Pools

- Poisson arrivals with rate λ.
- N i.i.d. exponential servers with service rates μ_i, $i=1,2,\ldots,N$.
- 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
Journal of Applied Probability, vol. 22, pp. 205-213, 1983.

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 ρ_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 $\wedge$-system forms a reversible MJP.
  - $\star$ $\pi_i q_{ij} = \pi_j q_{ji}$ $\forall i,j \in S$.
- 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

$$\pi_y = \pi_0 \frac{\prod_{i=1}^{K} \binom{N_i}{y_i}}{\binom{N}{m_y}} \frac{\lambda^{m_y}}{m_y! \prod_{i=1}^{K} \mu_i^{y_i}} \qquad y_i \in \{0,1,\ldots,N_i\},\ i \in \{1,2,\ldots,K\}$$

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

RMI Properties

Definitions:

- $\tilde{\rho}_i$ - stationary occupancy rate in pool *i*
- $\bar{\rho}_i$ - average occupancy rate in pool *i*
- γ_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:

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

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

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

$$\lim_{\lambda\to\infty} \frac{N_i^{\lambda}}{N^{\lambda}} = \frac{a_i}{\mu_i}\mu := q_i \qquad\qquad i = 1, 2, \ldots, K$$

Loss Probability: $K = 2$ Pools

Steady-state blocking probability:

$$P_\lambda(block) = \pi_\lambda^0 \cdot \frac{\lambda^N}{N!\mu_1^{N_1}\mu_2^{N_2}} = \frac{\frac{\lambda^N}{N!\mu_1^{N_1}\mu_2^{N_2}}}{\sum_{y_1=0}^{N_1}\sum_{y_2=0}^{N_2} \frac{\binom{N_1}{y_1}\binom{N_2}{y_2}}{\binom{N}{y_1+y_2}} \frac{\lambda^{y_1+y_2}}{(y_1+y_2)!\mu_1^{y_1}\mu_2^{y_2}}}$$

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 *Norm(0,1)*

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

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

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: δ, μ 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 $\mathcal{I}_i^\lambda$ - stationary number of idle servers in pool i, $i = 1, 2$.
Given that $\mathcal{I}_1^\lambda + \mathcal{I}_2^\lambda = \gamma\sqrt{\lambda}$, $\mathcal{I}_1^\lambda$ and $\mathcal{I}_2^\lambda$ deviate from $a_1\gamma\sqrt{\lambda}$ and $a_2\gamma\sqrt{\lambda}$ by $\Xi\sqrt[4]{\lambda}$, where $\Xi \Rightarrow Norm(0, \gamma a_1 a_2)$ *as* $\lambda \to \infty$.

Hence $a_2\mathcal{I}_1^\lambda \approx a_1\mathcal{I}_2^\lambda$ *as* $\lambda \to \infty$.

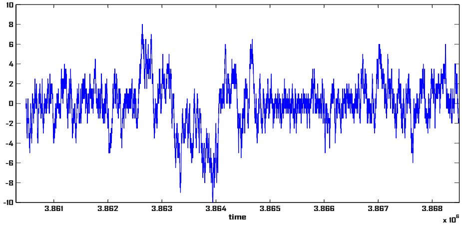
$\lambda = 3950$

$\downarrow$

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

$\mu_1 = 15, \quad \mu_2 = 7.5$
$N_1 = 138, N_2 = 276$

$\mathcal{I}_1(t) - a_1\mathcal{I}(t)$



Non-Random Equivalent to RMI

- 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 $N \to \infty$): $\mathcal{I}_1 \approx \mathcal{I}_2$.
- Thus: $\tilde{\rho}_i = \frac{N_i - \mathcal{I}_i}{N_i} = 1 - \frac{\mathcal{I}_i}{N_i}$, 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: $\dfrac{\mathcal{I}_1/N_1}{\mathcal{I}_2/N_2} = \dfrac{1-\bar{\rho}_1}{1-\bar{\rho}_2}$.
- *Flux-ratio* - ratio between flux through the pools ("flux" - number of arrivals per server per time unit): $\dfrac{\gamma_1}{\gamma_2} = \dfrac{\bar{\rho}_1\mu_1}{\bar{\rho}_2\mu_2}$.

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

Operational performance:

- Steady-state probability of loss, or $\mathbb{P}(\mathit{Block})$.

General observations

- RMI: from Stace-Space Collapse follows that:

$$\frac{1-\bar{\rho}_1}{1-\bar{\rho}_2} = \frac{\mathcal{I}_1/N_1}{\mathcal{I}_2/N_2} \approx \frac{N_2 a_1}{N_1 a_2} \approx \frac{q_2 a_1}{q_1 a_2} = \frac{\mu_1}{\mu_2}$$

$\rightarrow$ Idle-ratio depends only on service rates.

- MI:

$$\frac{1-\bar{\rho}_1}{1-\bar{\rho}_2} = \frac{\mathcal{I}_1/N_1}{\mathcal{I}_2/N_2} \approx \frac{N_2}{N_1} \approx \frac{q_2}{q_1}$$

$\rightarrow$ Idle-ratio depends only on pool capacities.

Comparison: RMI versus MI

		Idle-ratio	Flux-ratio	$\mathbb{P}(Block)$
$q_1 = q_2$		MI	RMI	MI
$q_1 > q_2$	$\frac{\mu_1}{\mu_2} < \frac{q_1}{q_2}$	RMI	RMI	MI
	$\frac{\mu_1}{\mu_2} = \frac{q_1}{q_2}$	equal		
	$\frac{\mu_1}{\mu_2} > \frac{q_1}{q_2}$	MI		
$q_1 < q_2$	$a_1 < a_2$	RMI	MI	RMI
	$a_1 = a_2$	equal	equal	equal
	$a_1 > a_2$	MI	RMI	MI

- 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

  $i = argmax\{w_1\mathcal{I}_1, w_2\mathcal{I}_2\}.$

- Asymptotically (as $N \to \infty$): $w_1\mathcal{I}_1 \approx w_2\mathcal{I}_2$.

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).

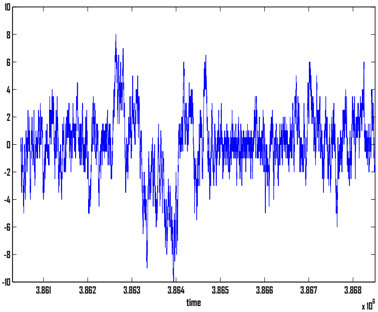
Comparison: WMI versus RMI

		Idle-ratio	Flux-ratio	$\mathbb{P}(Block)$
$w_1 q_1 = w_2 q_2$		WMI	RMI	WMI
$w_1 q_1 > w_2 q_2$	$\frac{\mu_1}{\mu_2} < \frac{w_1 q_1}{w_2 q_2}$	RMI	RMI	WMI
	$\frac{\mu_1}{\mu_2} = \frac{w_1 q_1}{w_2 q_2}$	equal		
	$\frac{\mu_1}{\mu_2} > \frac{w_1 q_1}{w_2 q_2}$	WMI		
$w_1 q_1 < w_2 q_2$	$w_1 a_1 < w_2 a_2$	RMI	WMI	RMI
	$w_1 a_1 = w_2 a_2$	equal	equal	equal
	$w_1 a_1 > w_2 a_2$	WMI	RMI	WMI

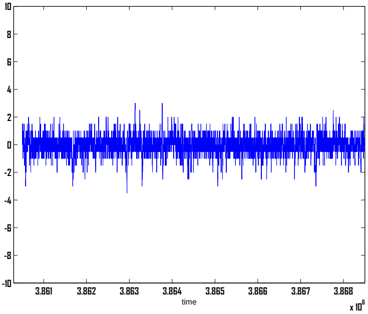
NERMI versus RMI

$$\mathcal{I}_1(t) - a_1\mathcal{I}(t)$$

RMI

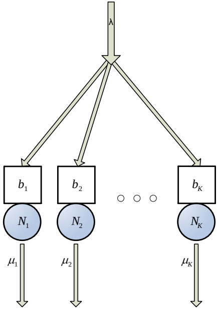


NERMI



Distributed Finite Queues

- Poisson arrivals with rate λ.
- K pools: pool i has
  - ⋆ N_i i.i.d. exponential servers with service rates μ_i, $i=1,2,\ldots,K$. $\sum_{i=1}^{K} N_i = N$
  - ⋆ 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 $\boxed{\frac{V_i(t)}{\sum_{j=1}^{K} V_j(t)}}$.
- 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!