

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

Yulia Tseytlin

Technion - Israel Institute of Technology

February 9, 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
Exact Analysis
Asymptotic Analysis

Additional Results

Alternative Routing Policies
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

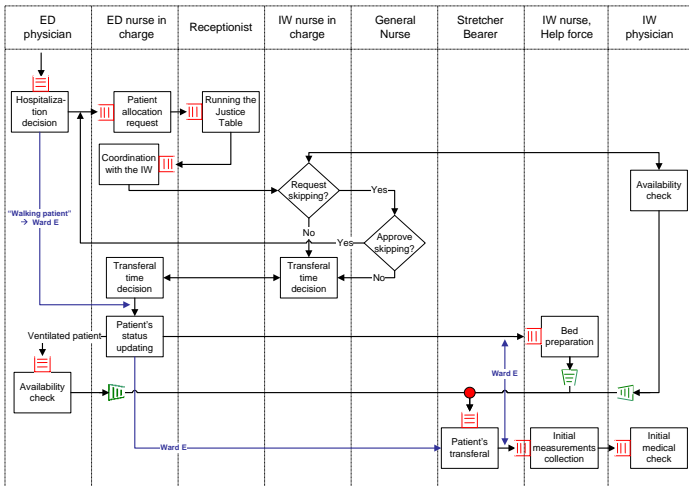
- **Wards A-D** are more or less the same in their medical capabilities - each can treat multiple types of patients.
- **Ward E** treats only “walking” patients, and the routing to it from the ED is different.
- We focus on the routing process to wards A-D only.

Standard and Maximal Capacity (# beds):

	Ward A	Ward B	Ward C	Ward D	Ward E
Standard capacity	45	30	44	42	24
Maximal capacity	52	35	46	44	27
Max. to standard ratio	115%	116%	104%	105%	113%



Integrated (Activities - Resources) Flow Chart



Resource Queue - [Icon] Synchronization Queue - [Icon]

● - Ending point of simultaneous processes



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 is **cyclical order** 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.



What is "Fair" Allocation?

Each nurse/doctor should have the same workload.

- Take care of an equal number of patients.
- Number of nurses/doctors is proportional to standard number of beds.

⇒ Balance **occupancy rates** among the wards.

- But then, by Little's law, wards with shorter ALOS will have a higher turnover rate.
- And the load on the wards' staff is not uniform during a patient's stay.

⇒ Balance number of patients per bed per time unit (*flux*) among the wards.



IW Operational Measures:

	Ward A	Ward B	Ward C	Ward D
ALOS (days)	6.318	4.574	5.446	5.642
Mean Occupancy Rate	97%	95%	86%	92%
Mean # Patients per Year	2,543	2,272	2,661	2,549
Standard capacity	45	30	44	42
Mean # Patients per Bed	56.5	75.7	60.5	60.7
Return Rate	15.4%	15.6%	16.2%	14.8%

* Data refer to period: 1/05/06-30/10/08

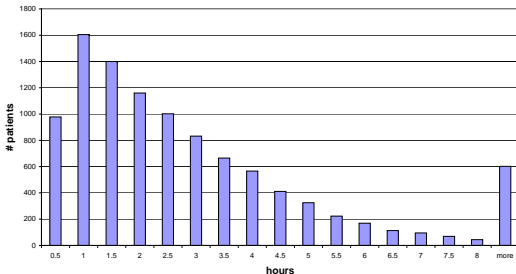
- **The smallest + “fastest” ward is subject to the highest loads.**
- **The patients’ routing appears unfair, as far as the wards are concerned.**



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.
- From hospital database, average time from a decision of hospitalization till receiving a first treatment in a ward was 2.9 hours (for Wards A-D).

Waiting Times Histogram



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 daily # of transfers from ED to IW	75 (50%)	7 (14%)	38 (42%)	24 (26%)	67 (45%)	33 (22%)
Average daily # of transfers per IW bed	0.229	0.156	0.352	0.258	0.319	0.178
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	1.5-3
Wards differ?	yes	yes	no	yes	no	yes
Routing Policy	cyclical order	last digit of id	cyclical order	vacant bed	cyclical order*	cyclical order*

* Account for different patient types and ward capacities.



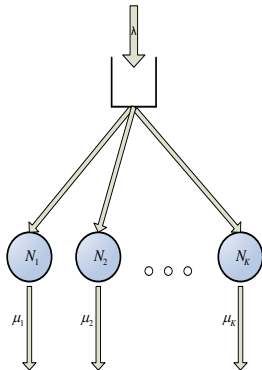
The ED-to-IW Process as a Queueing System

- Arrivals = patients to-be-hospitalized;
- Pools = wards;
- Service rates = $1/\text{ALOS}$;
- 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 , $i=1,2,\dots,K$;
 $\mu_1 > \mu_2 > \dots > \mu_K$.
 - ★ $\sum_{i=1}^K N_i = N$.
- 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.



Atar R.: Number of works on queueing systems with many servers.



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

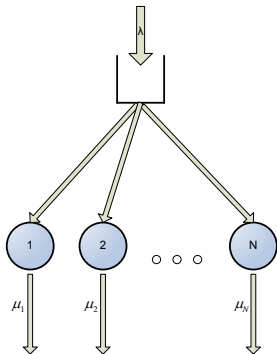
The \wedge -system presented before, under RMI routing policy, is equivalent to a \wedge -system with N single-server pools:

- 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 , $i=1,2,\dots,N$.
- One waiting line with infinite capacity.



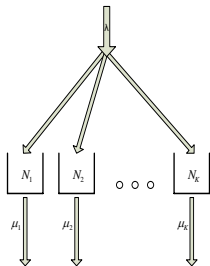
RMI Stationary Analysis

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

$$\star \pi_i q_{ij} = \pi_j q_{ji} \quad \forall i, j \in S.$$

- We present here a **Loss model** (*analysis of Delay models easily follows*).

- Poisson arrivals with rate λ .
- K pools:
 - Pool i consists of N_i i.i.d. exp. servers with rates μ_i , $i=1,2,\dots,K$;
$$\mu_1 > \mu_2 > \dots > \mu_K.$$
 - $\sum_{i=1}^K N_i = N$.
- No queueing possible.



Stationary Distribution

System states: $y = (y_1, y_2, \dots, y_K)$

- y_i - number of busy servers in pool i ($y_i \in \{0, 1, \dots, N_i\}$).
- $m_y = \sum_{i=1}^K y_i$ - total number of busy servers at state y .

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

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



RMI Properties

Definitions:

- $\tilde{\rho}_i$ - long-run occupancy in pool i
- $\bar{\rho}_i$ - average long-run occupancy 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: $\tilde{\rho}_i \leq_{st} \tilde{\rho}_j$ ($\mathbb{P}(\tilde{\rho}_i > x) \leq \mathbb{P}(\tilde{\rho}_j > x) \forall x \in (0, 1)$)



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 85%.



QED Scaling

[Armony M., 2005]

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

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

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

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

$$\lim_{\lambda \rightarrow \infty} \frac{N_i^\lambda}{N^\lambda} = \frac{a_i}{\mu_i} \mu := q_i \quad i = 1, 2, \dots, K$$



Loss Probability: $K = 2$ Pools

Steady-state blocking probability:

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



Loss Probability Approximation

P. Momcilovic proved:

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

where:

- $\hat{\mu} := \mu_1 \mathbf{a}_1 + \mu_2 \mathbf{a}_2$
- $\varphi(\cdot), \Phi(\cdot)$ - density and probability functions of $Norm(0, 1)$

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

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



Loss Probability Approximation

If $\mu_1 = \mu_2$:

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

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

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

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

Insights:

- $\sqrt{N} P_{\lambda}(block)$ is a function of three parameters: δ , μ and $\hat{\mu}$:
 - * As $\lambda \rightarrow \infty$, 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 $1/\sqrt{N}$.



State-Space Collapse

P. Momcilovic found:

Denote I_i^λ - stationary number of idle servers in pool $i, i = 1, 2$.
 Given that $I_1^\lambda + I_2^\lambda = \gamma\sqrt{\lambda}$, I_1^λ and I_2^λ deviate from $a_1\gamma\sqrt{\lambda}$ and $a_2\gamma\sqrt{\lambda}$ by $\Xi\sqrt[4]{\lambda}$, where $\Xi \Rightarrow \text{Norm}(0, \gamma a_1 a_2)$ as $\lambda \rightarrow \infty$.

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

$$\lambda = 3950$$

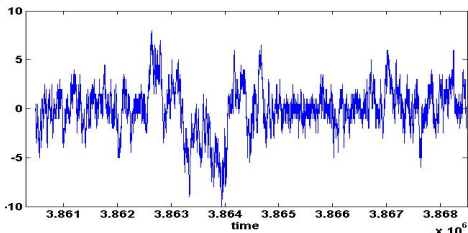


$$\sqrt[4]{\lambda} \approx 8$$

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

$$N_1 = 138, \quad N_2 = 276$$

$$I_1(t) - a_1 I(t)$$



Non-Random Equivalent to RMI

- RMI Routing Policy enjoys some desirable properties, but is problematic for a hospital environment due to its randomness.
- The naive 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* $N \rightarrow \infty$): $\mathcal{I}_1 \approx \mathcal{I}_2$.
- Thus: $\tilde{\rho}_i = \frac{N_i - \mathcal{I}_i}{N_i} = 1 - \frac{\mathcal{I}_i}{N_i}$, i.e., larger pool (bigger N_i) has higher occupancy.
- Asymptotic analysis shows that RMI and MI are not equivalent.



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

$$i = \operatorname{argmax}\{w_1 I_1, w_2 I_2\}.$$

- Asymptotically (as $N \rightarrow \infty$): $w_1 I_1 \approx w_2 I_2$.



WMI Routing Policy

Interesting cases:

- $w_1 = w_2 = 1/2$ $(\mathcal{I}_1 \approx \mathcal{I}_2)$
 - ★ **MI** routing policy.
- $w_1 = a_2, w_2 = a_1$ $(a_2 \mathcal{I}_1 \approx a_1 \mathcal{I}_2)$
 - ★ Non-random Equivalent to RMI - **NERMI** routing policy.
- $w_1 = q_2, w_2 = q_1$ $(q_2 \mathcal{I}_1 \approx q_1 \mathcal{I}_2)$
 - ★ *Occupancy-Balancing* - **OB** policy: routing an arriving customer to the least utilized pool (pool with the minimal occupancy).



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 $\mathbb{P}(\text{Block})$.



General observations

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

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

→ Idle-ratio depends only on service rates.

- **WMI:**

$$\frac{1 - \bar{\rho}_1}{1 - \bar{\rho}_2} = \frac{\mathcal{I}_1/N_1}{\mathcal{I}_2/N_2} \approx \frac{w_2 N_2}{w_1 N_1} \approx \frac{w_2 q_2}{w_1 q_1}$$

→ Idle-ratio depends on weights and pool capacities.



Comparison: WMI versus RMI

		Idle-ratio	Flux-ratio	$\mathbb{P}(\text{Block})$
$w_1 q_1 = w_2 q_2$		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

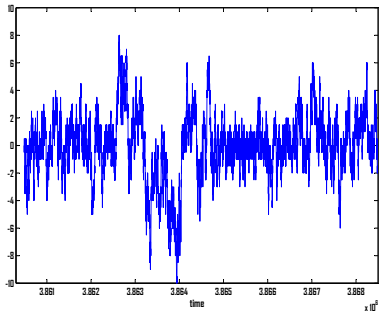
- For different sets of parameters and different target functions, a different policy is superior.



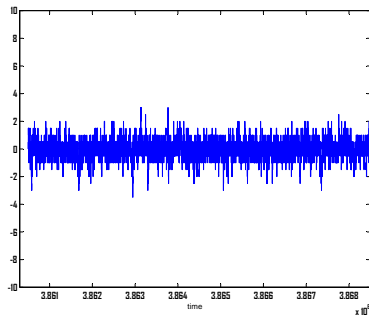
NERMI versus RMI

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

RMI



NERMI



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.
- *Flux 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?
 - ★ Are waiting times a function of 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 the 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.
- Extend the theoretical study to several customer (patient) classes.
- Find QED approximations for RMI in distributed queues.





Thank You!

