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
  Practical Background
  Theoretical Background
    Inverted-V System
    QED Asymptotic Regime

Routing Policies
  RMI - Exact Analysis
  RMI - QED Analysis
  RMI vs. LISF and IR Routing Policies

Additional Results
  Simulation Analysis
  Summary and Future Research
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:

<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>Capacity (# beds)</td>
<td>45</td>
<td>30</td>
<td>44</td>
<td>42</td>
</tr>
<tr>
<td>ALOS (days)</td>
<td>6.368</td>
<td>4.474</td>
<td>5.358</td>
<td>5.562</td>
</tr>
</tbody>
</table>
Integrated (Activities - Resources) Flow Chart

- ED physician
- ED nurse in charge
- Receptionist
- IW nurse in charge
- General Nurse
- Stretcher Bearer
- IW nurse, Help force
- IW physician

Hospitalization decision

Patient allocation request

Running the Justice Table

Coordination with the IW

Request skipping?

Yes

Approve skipping?

No

Yes

No

Transferal time decision

Transferal time decision

Bed preparation

Ward E

Patient's status updating

Patient's transferal

Initial measurements collection

Initial medical check

Ventilated patient

Availability check

“Walking patient” → Ward E

Availability check

Resource Queue - Resource Queue
Synchronization Queue - Resource Queue

 Ending point of simultaneous processes
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

<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>6.368</td>
<td>4.474</td>
<td>5.358</td>
<td>5.562</td>
</tr>
<tr>
<td>Mean Occupancy Rate</td>
<td>97.8%</td>
<td>94.4%</td>
<td>86.8%</td>
<td>91.1%</td>
</tr>
<tr>
<td>Mean # Patients per Month</td>
<td>205.5</td>
<td>187.6</td>
<td>210.0</td>
<td>209.6</td>
</tr>
<tr>
<td>Standard capacity</td>
<td>45</td>
<td>30</td>
<td>44</td>
<td>42</td>
</tr>
<tr>
<td>Mean # Patients per Bed per Month</td>
<td>4.57</td>
<td>6.25</td>
<td>4.77</td>
<td>4.77</td>
</tr>
<tr>
<td>Return Rate (within 3 months)</td>
<td>16.4%</td>
<td>17.4%</td>
<td>19.2%</td>
<td>17.6%</td>
</tr>
</tbody>
</table>

* 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

<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>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>IW # beds</td>
<td>327</td>
<td>45</td>
<td>108</td>
<td>93</td>
<td>210</td>
<td>185</td>
</tr>
<tr>
<td>Average weekly # of transfers</td>
<td>525 (50%)</td>
<td>49 (14%)</td>
<td>266 (42%)</td>
<td>168 (26%)</td>
<td>469 (45%)</td>
<td>231 (22%)</td>
</tr>
<tr>
<td>from ED to IW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average weekly # of transfers</td>
<td>1.606</td>
<td>1.089</td>
<td>2.463</td>
<td>1.806</td>
<td>2.233</td>
<td>1.249</td>
</tr>
<tr>
<td>per IW bed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IW Occupancy*</td>
<td>107.5%</td>
<td>118%</td>
<td>106.5%</td>
<td>116.4%</td>
<td>110%</td>
<td>93.8%</td>
</tr>
<tr>
<td>ED ALOS (hours)</td>
<td>2.2</td>
<td>6</td>
<td>2.83</td>
<td>6.8</td>
<td>2.5</td>
<td>4.2</td>
</tr>
<tr>
<td>IW ALOS (days)</td>
<td>3.9</td>
<td>3.9</td>
<td>3.5</td>
<td>6.1</td>
<td>3.5</td>
<td>5.2</td>
</tr>
<tr>
<td>Average waiting time in ED</td>
<td>?</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>0.5</td>
<td>3</td>
</tr>
<tr>
<td>for IW (hours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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>cyclical order</td>
<td>last digit of id</td>
<td>cyclical order</td>
<td>vacant bed</td>
<td>cyclical order**</td>
<td>cyclical order**</td>
</tr>
</tbody>
</table>

* 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 (⁻-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.
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*

- 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 $\rho_c$, 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*

- 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*


- **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*


- 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*

- 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*

- For Parallel-Server Systems, propose **Queue-and-Idleness-Ratio** rules.
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 $j$ with probability

  \[
  \frac{I_j(t)}{\sum_{k=1}^{K} I_k(t)}
  \]

  * 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
- $\rho = \frac{\lambda}{\sum_{i=1}^{K} N_i \mu_i}$ - 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 $\pi$:

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

where

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

- $I_i$ - stationary number of idle servers in pool $i$.
- $\rho_i := 1 - \mathbb{E}I_i/N_i$ - average steady-state occupancy rate in pool $i$.
- $\gamma_i$ - average flux through pool $i$ = average number of arrivals per server in pool $i$ per time unit.
  
  $\star \quad \gamma_i = \mu_i \rho_i$, 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:

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

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

Also define:

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

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

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

**Theorem 2 (Informal):**

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

$$\Rightarrow \frac{I_i^\lambda}{I_j^\lambda} \approx \frac{a_i}{a_j} \quad \text{as } \lambda \to \infty$$

- Characterization of the system behavior on the *sub-diffusion* $(\sqrt[4]{\lambda})$ scale; $\sqrt[4]{\lambda}$-deviations of $I_i^\lambda$ around $a_i(I^\lambda)^+$
RMI: QED Analysis - cont.

Theorem 2:
Let \( \hat{I}^\lambda = I^\lambda / \sqrt{\lambda} \) and \( \hat{I}_i^\lambda = \frac{1}{\sqrt{I^\lambda}} \left( I_i^\lambda - \frac{N_i^\lambda \mu_i}{\sum_{i=1}^K N_i^\lambda \mu_i} I^\lambda \right) \), \( i=1,\ldots,K \).

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

If $\mu_1 = \mu_2 = \ldots = \mu_K$:
Then $\mu = \hat{\mu} = \mu_1$, $\delta/\sqrt{\hat{\mu}} = \beta$ and $\Pr[\hat{T} \leq 0] = \left(1 + \beta \frac{\Phi(\beta)}{\varphi(\beta)}\right)^{-1}$
$\Rightarrow$ Consistent with Erlang-C Approximation [S. Halfin and W. Whitt, 1981].

Example: exact values vs. QED approximations

$K = 2$
$q_2 = 2q_1 = 2/3$
$\mu_1 = 2\mu_2 = 2$
$\delta = 0.5$
$\lambda : 10 - 500$
$N_1^\lambda : 3 - 128$
$N_2^\lambda : 6 - 256.$
Dimensionality Reduction Illustration

- $K = 2$, $\lambda = 3950$, $\mu_1 = 15$, $\mu_2 = 7.5$, $N_1 = 138$, $N_2 = 276$ ($\delta = 3$, $a_1 = a_2 = 1/2$)

- $\{I^\lambda(t), t \geq 0\}$ evolve on $\sqrt{\lambda}$-scale ($\sqrt{\lambda} \approx 62.8$)

- $\{I_1^\lambda(t) - a_1(I^\lambda(t))^+, t \geq 0\}$ evolve on $\sqrt[4]{\lambda}$-scale ($\sqrt[4]{\lambda} \approx 7.93$)
Fair Routing Criteria

Occupancy balancing

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

Flux balancing

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

In the QED regime \( \lim_{\lambda \to \infty} \frac{\gamma_i^\lambda}{\gamma_j^\lambda} = \frac{\mu_i}{\mu_j} \Rightarrow \text{strive for } \rho_i^\lambda < \rho_j^\lambda \text{ if } \mu_i > \mu_j \).

In RMI - from Theorem 2:

- $\frac{I_i^\lambda(t)}{I_j^\lambda(t)} \to \frac{a_i(I_i^\lambda(t))^+}{a_j(I_j^\lambda(t))^+} = \frac{a_i}{a_j}$, thus
- $\frac{1 - \rho_i^\lambda}{1 - \rho_j^\lambda} = \frac{E I_i^\lambda}{N_i^\lambda} \frac{N_j^\lambda}{E I_j^\lambda} \to \frac{a_i q_j}{a_j q_i} = \frac{\mu_i}{\mu_j}$
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 $\lambda \to \infty$):

  \[
  \frac{I_i^\lambda(t)}{I_j^\lambda(t)} \to \frac{a_i(I_i^\lambda(t))^+}{a_j(I_j^\lambda(t))^+} = \frac{a_i}{a_j}, \text{ thus}
  \]

  \[
  1 - \rho_i^\lambda = \frac{E[I_i^\lambda]}{N_i^\lambda} \to \frac{a_i q_j}{a_j q_i} = \frac{\mu_i}{\mu_j}
  \]

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

  $\star$ 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 \(\text{arg max}\{I_i^\lambda(t-) - w_i(I_j^\lambda(t-))\}^+\)
- Asymptotically (as \(\lambda \to \infty\)): \(\frac{1 - \rho_i^\lambda}{1 - \rho_j^\lambda} = \frac{E[I_i^\lambda]}{E[I_j^\lambda]} \frac{N_j^\lambda}{N_i^\lambda} \to \frac{w_i q_j}{w_j q_i}\)

⇒ 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 $I_1^\lambda(t) - a_1(I^\lambda(t))^+, \ t \geq 0$:

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!