# ON FAIR ROUTING FROM EMERGENCY DEPARTMENTS TO HOSPITAL WARDS: QED QUEUES WITH HETEROGENEOUS SERVERS

AVISHAI MANDELBAUM, PETAR MOMČILOVIĆ, AND YULIA TSEYTLIN

Abstract. The interface between an Emergency Department (ED) and Internal Wards (IWs) is often a hospital's bottleneck. Motivated by this interaction in Anonymous Hospital, we analyze queueing systems with heterogeneous server pools, where the pools represent the wards and servers are beds. Our queueing system, with a single centralized queue and several server pools, forms an inverted-V model. We introduce the Randomized Most-Idle (RMI) routing policy and analyze it in the QED (Quality and Efficiency Driven) regime, which is natural in our setting. The RMI policy results in the same server fairness (measured by idleness ratios) as the Longest-Idle Server First (LISF) policy, which is commonly used in call centers and considered fair. However, RMI utilizes only the information on the number of idle servers in different pools while LISF requires information that is unavailable in hospitals on a real-time basis.

## 1. Introduction

Operations research methodologies, and queueing theory in particular, have generated valuable insights into operational strategies and practices, thus leading to solutions of significant problems in healthcare systems. In concert with this state of affairs, we analyze patient flow in hospitals: specifically, we focus on the Emergency Department (ED) and its interface with four Internal Wards (IWs) in a large Israeli hospital – we refer to it as Anonymous Hospital. Two operational problems could arise in this process: patients' waiting times in the ED for a transfer to the IWs could be long; and patient routing to the wards need not be fair, as far as workload allocation among the wards is concerned. In contrast to the majority of studies that address the problem of long waiting times in EDs, we explore the process of patient allocation to wards from the fairness perspective. In search of an allocation protocol that is fairness-sensitive, we model the "ED-to-IW" process as a queueing system with heterogeneous server pools: the pools represent the wards and servers are beds. Within this modeling framework, we compare several routing strategies, thus identifying the one most appropriate for a hospital setting.

1.1. **Motivation.** Two conclusions can be drawn from Anonymous Hospital data (see Table 1 in Section 2.3). First, the fastest and smallest ward (Ward B) is subject to the highest load: it experiences the highest number of patients per bed per month, with bed occupancy that is comparable to the other wards. The reasons behind the high turnover rate of Ward B are superior managerial and staff practices, as well as (medically justified) varying policies for patients release, which results in significantly shorter Average Length of Stay (ALOS). Importantly, shorter ALOS does not come at the cost of inferior medical care. Indeed, the

*Date*: September 2, 2011.

*Key words and phrases.* Queueing systems; heterogeneous servers; healthcare; hospital routing policies; fairness; Quality and Efficiency Driven (QED) regime; asymptotic analysis.

A. Mandelbaum's research was supported in part by BSF (Binational Science Foundation) Grants 2005175/2008480, ISF (Israeli Science Foundation) Grant 1357/08, and by the Technion funds for the promotion of research and sponsored research.

P. Momčilović's research was supported in part by NSF (National Science Foundation) Grant CNS-0643213.

level of return rates (within 3 months) is comparable across wards. Hence, one could argue that the allocation process of patients from the ED to the IWs is not fair from the point of view of medical staff, a problem that has been observed in other Israeli hospitals as well.

The second empirical conclusion is that some key operational characteristics of the ED-to-IW process coincide with the features that characterize moderate-to-large-scale queueing systems in the Quality and Efficiency Driven (QED) regime [18, 24, 19]. Such a regime achieves, simultaneously, high levels of operational service quality (ED waiting times significantly shorter than service durations, specifically ALOS) and resource efficiency (high bed occupancy).

1.2. **Contributions.** Our paper focuses on modeling the ED-to-IW process, which is a key phase of patient flow in hospitals. More broadly, in 2005, the Joint Commission on the Accreditation of Healthcare Organizations (JCAHO) set a standard (LD.3.15) for patient flow leadership. The standard requires that healthcare leaders "develop and implement plans to identify and mitigate impediments to efficient patient flow throughout the hospital". It amplifies the need to identify the critical factors that impact patient flow, with the ultimate goal of designing and implementing policies, processes and procedures that track, monitor and improve patient flow throughout hospitals.

Although we concentrate on the impact of patient routing on staff fairness, our work should be viewed within the broader context of improving staff efficiency by creating an appropriate incentives structure. Operational policies that are not perceived to be fair could internalize inefficiencies, i.e., create situations where good work leads to more work. On the other hand, fair policies not only reward staff members with the best practices, but also promote the adoption of such practices. Thus, while fair policies come at a certain cost in the short run, in the long run they are likely to improve overall system efficiency.

Our contributions are as follows:

- Based on empirical data from Anonymous Hospital, we argue that an inverted-V queueing model in the QED regime is appropriate for describing the ED-to-IW process, which exhibits multi-scale behavior (e.g. IW lengths of stay being much longer than ED waiting times).
- We quantify operational fairness towards medical staff by means of ratios that take into account bed occupancy levels and bed *turnover rates* (the average number of admitted patients per bed per unit of time) – two main metrics as far as workload of medical staff is concerned. These metrics serve as operational proxies for fairness, which is an intricate concept. (Operational proxies relate to operations, are easy to measure, and they approximate notions that are hard to quantify.)
- Although routing algorithms that take into account fairness have been considered in the literature, we propose a practical routing algorithm (Randomized Most-Idle = RMI) which is suitable for hospitals where only limited and partial information is available for patient routing. That is, we consider the role of information in the performance of routing policies.
- The proposed algorithm, RMI, is analyzed and compared to known algorithms. Our results demonstrate that RMI achieves (long-run) level of fairness toward medical staff, that is the same as algorithms which require more information to operate. Moreover, within our narrow notion of fairness, based on occupancy and turnover rates, an extension of our algorithm, weighted RMI, is flexible enough to achieve any desired fairness.
- Furthermore, our results yield important insights on differences between various routing algorithms, differences that arise at the sub-diffusion (room-level) scale. In particular, these insights reveal that instantaneous imbalance of workload across the IWs is limited to just a few rooms of patients. Moreover, in the course of just a few days this imbalance

(if present) disappears due to time averaging, i.e., empirical workload time-averages converge to the desired long-run averages.

1.3. **Brief literature review.** There exists a vast amount of research on healthcare systems, in numerous scientific fields including operations research. We mention here the most relevant to the present work; additional related references can be found in [45, 7, 15, 27, 20, 16, 34], as well as in each of the papers mentioned below. Readers are also referred to Green [21, 22], who describes the general background and issues involved in hospital capacity planning.

Research papers (e.g. [32]) and popular articles (e.g. [43]) both recognize the importance of ED proper functioning and the consequences of its overcrowding. Patient flow from the ED to other medical units in the hospital (not just the IWs) has received special attention (becoming even the subject of a novel [49]). In fact, it has been acknowledged as a major trigger of ambulance diversion [32], which is of great concern. Ramakrishnan et al. [37] construct a two-time-scale model for a hospital system, where the wards operate on a time scale of days and are modeled by a discrete time Markov chain, and the ED operates on a much faster time scale and is modeled by a continuous time Markov chain. With the help of this model, [37] estimates expected occupancy of the wards and the probability of each ward reaching its capacity. The setup in [37] corresponds to ours in that the ED-to-IW process also operates in several scales. These scales arise naturally from our asymptotic analysis (QED regime). Specifically, length-of-say in wards, which correspond to our service times, are naturally measures in days, while waiting in the ED for transfer to the wards is naturally measured in hours, which corresponds to our waiting times.

Relevant analyses of queueing systems with heterogeneous servers date back to the slow server problem [39]. Initially, the focus was on finding the best operating policy in order to minimize the steady-state mean *sojourn* time of the customers in the system, which is equivalent to minimizing the long-run average number of customers in the system, due to Little's law. Under such criteria, it is preferable to use the faster servers more than the slower servers. In fact, under some circumstances, it is even advantageous to remove the very slow servers and thus reduce sojourn time. This slow server phenomenon is addressed, for the case of two heterogeneous servers and a single queue, in [29, 31, 39, 40, 42]; and for the general multi-servers case in Cabral [10] (under the Random Assignment (RA) policy, which routes a customer to one of the idle servers at random). Later, Cabral proved in [11] that, for any two servers, the fraction of time that the faster among them is busy is smaller than that of the slower one, and the effective service rate of the faster server is higher than that of the slower one. Except for [11], none of the studies mentioned above touches on the issue of *fairness* towards servers.

In the context of large-scale systems, Armony [1] analyzed the Fastest Servers First (FSF) routing policy that assigns customers to the fastest available pool. She shows that FSF is asymptotically optimal (within the set of all non-preemptive, non-anticipating First-Come First-Served (FCFS) policies), in the sense that it stochastically minimizes the stationary queue length and waiting time, as the arrival rate and number of servers grow large; Armony and Mandelbaum [2] extended this result to accommodate abandonments. Yet, under the FSF policy, asymptotically only the slowest servers have any idle time. This is obviously unfair towards the fast servers (which get "punished" for being fast by working more), and it gives them an incentive to slow down – an undesirable result for the system as a whole. Thus, there exists a trade-off between operational optimality for the system vs. fairness towards servers [3].

There is ample research aimed at achieving fairness to customers; see, for example, references in [45]. However, to our best knowledge, only recently Atar [4] was the first to deal with the operational fairness-towards-servers issue. He studied a single-server pools model in the QED regime, where the number of servers and their service rates are i.i.d. random variables, under

a policy that routes an arriving customer to the server that has been idle for the longest time among all idle servers (in a deterministic environment, this policy is called Longest-Idle Server First (LISF) in [1]). Armony and Ward [3] extended LISF routing to Longest-Weighted-Idle Server First (LWISF), and proposed a threshold policy that asymptotically (in the QED regime) outperforms LWISF while achieving the same target idleness ratios. Atar, Shaki and Shwartz [5] proposed the Longest-Idle Pool First (LIPF) policy, that routes a customer to the pool with the longest cumulative idleness among the available pools; in the QED regime, this policy is shown to balance cumulative idleness among the pools. Most of these papers examine transient system behavior by establishing weak convergence process-level results. In contrast, our work deals with steady-state behavior.

Fairness (or justice, or equity) is a well-researched area in the behavioral sciences [13]. We shall mention some relevant work later in Section 2.2.

1.4. **Organization.** The paper is organized as follows. In the next section we describe the process for routing patients from an ED to IWs, which reveals a fairness problem in this process. Our queueing model, as well as the QED regime, are introduced in Section 3. Fair routing algorithms are described in Section 4. Section 5 contains our theoretical results and a related discussion, including managerial implications. Concluding remarks appear in Section 6. The appendices contain a description of the routing algorithm currently implemented at Anonymous Hospital, the state-of-affairs in five other hospitals, technical proofs, and finally a table of notation and a list of acronyms.

## 2. Patient Routing

We study patient flow from the ED to the IWs in hospitals. Our research site is Anonymous Hospital, which is a large Israeli hospital with about 1000 beds, 45 medical units, and about 75,000 patients hospitalized yearly. Among its variety of units, it has a large ED with an average arrival rate of 200-300 patients daily, who occupy up to 50 beds; and five IWs, which we denote from A to E. The ED is divided into two major subunits: Internal and Trauma (the latter being surgical and orthopedic patients). An internal patient, whom the ED decides to hospitalize, is directed to one of the five IWs according to a certain routing policy – this routing process is the focus of our research.

Departments of Internal Medicine are responsible for catering to a wide range of internal disorders, providing inpatient medical care to thousands of patients each year. Wards A-D are more or less similar in their medical capabilities – each can treat multiple types of patients. Ward E, on the other hand, treats only "walking" patients, and routing to it differs from that to the other wards. In our study we thus concentrate on the routing process to Wards A-D only. The existence of multiple wards with similar medical capabilities is common in Israeli hospitals and can be attributed to various factors, for example: (i) there exist constraints in terms of physical space, e.g., wards can be located on different floors of a building; (ii) the existence of multiple wards implies the existence of multiple positions of ward managers – these positions are used by the hospital management to attract top-performing doctors; and (iii) informal research in Anonymous Hospital suggests that healthcare operations could exhibit diseconomies of scale – this would, of course, discourage the formation of a single super-ward.

2.1. **The Routing Process.** The decision of routing a to-be-hospitalized patient to one of Wards A-D is supported by a computer program, referred to as the "*Justice Table*" at the hospital. As its name suggests, the algorithm's goal is to make the patient allocation to the wards *fair*, by balancing the load among the wards. Prior to routing, patients are classified into three categories, according to the complexity of treatment: ventilated, special-care and regular. The program accepts a patient's category as an input parameter and returns a ward for the patient (A-D) as an output. For each patient category, there is round-robin order

among the wards, while accounting for the size of each ward by allocating fewer patients to a smaller ward: for example, 2 patients to B per 3 patients to A, reflecting 30:45 bed ratio. The Justice Table does not take into account the *actual* number of occupied beds and the patient discharge rate. In Appendix A, we provide a brief history of the Justice Table and a more detailed description of the ED-to-IW process.

We recognize a problem in the process of patient routing from the ID to the IWs: patient allocation to the wards does not appear to be fair. In what follows, we first examine the notion of fairness and then discuss the specifics present at Anonymous Hospital.

*Remark.* In addition to analyzing the ED-to-IW process in one specific hospital, we examined how this process is managed in five other Israeli hospitals (see Appendix A.4). In particular, we learned about the routing policies that are being used, and how successful they are in terms of delays and fair allocation. To this end, we used a questionnaire that included both qualitative (detailed description of the process, fairness considerations) and quantitative (operational measures of the ED and IWs: capacity, ALOS, waiting times) questions. Our study revealed that unbalanced loads on the wards due to heterogeneity of ALOS are common in all our surveyed hospitals.

2.2. **Fairness.** One can analyze fairness towards *patients* (customers) and fairness towards *wards* – the latter covering medical and nursing staff (servers). There is ample literature on measuring fairness in queues from the customer point of view (e.g., see [6, 30, 36]). Various aspects are investigated (for example, single queue vs. multi-queues, or FCFS vs. other queueing disciplines), but all agree that the FCFS policy is typically essential for justice perception. Consequently, customer satisfaction in a single queue is higher than in multi-queues [30]; and waiting in a multi-queue system produces a sense of lack of justice even when no objective discrimination exists [36]. The situation could be different in invisible queues (e.g. call centers) and healthcare queues (e.g. EDs). In the latter, clinical priority naturally dominates FCFS justice.

*Remark.* We note that patient "service" in a ward actually starts prior to the physical arrival of the patient to the ward. Indeed, we observed that the ward, once informed about a to-be-admitted patient, starts preparing for this *specific* patient: different patients, even if they fall under the same classification, might require different preparations. This sometimes leads to the following scenario. Suppose that a decision for hospitalization of patient X was made prior to a decision of hospitalization of patient Y (assuming both of them are clinically similar): say, patient X is directed to Ward A and patient Y to Ward B. Now, suppose that Ward B becomes ready to physically admit the patient earlier than Ward A – hence patient Y joins a ward *before* patient X, although Y "arrived" later [17]. In addition to this need for a ward to prepare for the patient, the hospital staff refrains from modifying patient-ward assignments also due to a psychological reason: a patient awaiting hospitalization (as well as accompanying individuals) experiences high levels of stress as is – one thus does not wish to aggravate stress by changing original ward assignments [17].

In our work, we focus on the process of patient *assignment* to wards, as opposed to the *physical* process of patient transfer from the ED to the IWs. We refer to the former as the ED-to-IW process. (In this process, deviations from FCFS do not raise fairness problems among patients since the assignment queue is typically invisible to them.)

The literature on justice from the server point of view is concerned with *Equity Theory* [26], according to which workers perceive their justice by comparing ratios of outputs from the job to inputs to the job. Specifically, if the output/input ratio of an individual is perceived to be unequal to others, then inequity exists. The larger the inequity the individuals perceive, the more uncomfortable they feel and the harder they work to restore equity [26]. In [8], it is

shown that in customer service centers, servers' equity perception has a positive influence on their performance and job satisfaction. References to additional studies on the importance of perceived justice among employees can be found in [3].

Our anchor point is the survey reported in [17], in which the staff (nurses, doctors and administration) were asked to grade the extent of fairness in different routing policies. When discussing fairness with wards staff, the consensus was that each nurse/doctor should have the same workload as others. Seemingly, this is the same as saying that each nurse/doctor should take care, at any given time, over an equal number of patients (assuming homogeneous customers, for simplicity). As the number of nurses and doctors is usually proportional to standard capacity, this criterion is equivalent to keeping *occupancy levels* of beds equal among the wards.

Note that, by Little's law, $\rho = \gamma \times \text{ALOS}$, where $\rho$ is the average occupancy level and $\gamma$ is the bed turnover rate. Thus, if one maintains ward occupancies equal then wards with shorter ALOS will have a higher turnover rate – admit more patients per bed – which gives rise to additional fairness concerns. Indeed, the load on staff is not spread uniformly over a patient's stay, as treatment during the first days of hospitalization requires much more time and effort from the staff than in the following days [17]; moreover, patient admissions and discharges significantly consume doctors' and nurses' time and effort as well. Thus, even if occupancy among wards is kept equal, the ward admitting more patients per bed ends up having a higher load on its staff. For these reasons, a natural alternative fairness criterion is balancing the turnover rate, or the *flux* – namely, the number of admitted patients per bed per unit of time (for example, per month), among the wards.

One can also combine utilization and flux to produce a single workload measure; fair routing would maintain this measure equal across wards. For example, load, experienced by medical staff in a ward, can be roughly divided into two parts: load associated with treating hospitalized patients (quantified by utilization) and load due to patient admissions/discharges (quantified by flux). For example, a single (objective) workload measure for a nurse, based on a linear combination of utilization and flux, was proposed in [45, Section 7.2]:

$$\frac{\gamma_i N_i T_i^{\gamma} + \rho_i N_i T_i^{\rho}}{n_i}, \tag{1}$$

where $\gamma_i$ is the flux, $\rho_i$ is the occupancy level, $N_i$ is the number of beds in the ward, $T_i^{\gamma}$ is the average amount of time required from a nurse to complete one admission plus discharge, $T_i^{\rho}$ is the average time of treatment required by a hospitalized patient per unit of time, and $n_i$ is the number of nurses in the ward; all quantities refer to a given ward $i$.

2.3. **The setting of Anonymous Hospital.** Although Wards A-D in Anonymous Hospital provide similar medical services, they do differ in their operational characteristics (see Table 1), which we now elaborate on. First of all, each medical unit is characterized by its *capacity*. Ward's capacity is measured by its number of beds (standard static capacity) and number of service providers – doctors, nurses, administrative staff and support staff (dynamic capacity). The "maximal" static capacity stands for the standard static capacity plus extra beds, which can be placed in corridors during overloaded periods. It is convenient to introduce notions of a sub-ward and a patient room; these will play a role in a discussion of our results. Wards consist of sub-wards, which, in turn, are made up of physically co-located rooms. There a few (2-3) sub-wards in a ward, and a sub-ward consists of several (3-4) rooms.

In our hospital IWs, the dynamic capacity during a particular shift is determined proportionally to the static capacity (see, however, discussions on the appropriateness of such "proportional" staffing in [16, 22, 51]). In particular, an IW beds-to-nurses ratio in morning shifts is 5:1 or 6:1 (depending on the number of "intensive care" beds in a ward); during night shifts, the ratio is 8:1 or 9:1. Note that the number of nurses is determined by the number of

beds in a ward rather than the number of occupied beds (patients). Hence a unit's operational capacity can be characterized by its number of beds only – denoted as its *standard* capacity.

TABLE 1. Internal Wards operational measures.

| | Ward A | Ward B | Ward C | Ward D |
|---|---|---|---|---|
| Average Length of Stay (ALOS)* (days) | 6.5 (± 0.19) | **4.5 (± 0.15)** | 5.4 (± 0.22) | 5.7 (± 0.18) |
| Mean occupancy level | 97.8% | 94.4% | 86.8% | 91.1% |
| Mean # patients per month | 205.5 | 187.6 | 210.0 | 209.6 |
| Standard (maximal) capacity (# beds) | 45 (52) | 30 (35) | 44 (46) | 42 (44) |
| Mean # patients per bed per month | 4.58 | **6.38** | 4.89 | 4.86 |
| Return rate (within 3 months) | 16.4% | 17.4% | 19.2% | 17.6% |

Data refer to period May 1, 2006 - October 30, 2008 (excluding the months 1-3/2007, when Ward B was in charge of an additional sub-ward). The data covers 16,947 admissions in total.
* The level of confidence for ALOS is 95%.

2.4. **Fairness at Anonymous Hospital.** Medical units are further characterized by various performance measures: *operational* – average bed occupancy level, ALOS, waiting times for various resources, number of patients admitted, or released, per bed per time unit (*flux*); and *quality* – patients' return rate, patients' satisfaction, mortality rate, etc. Note that occupancy levels and flux are calculated relatively to wards' standard capacities. (Thus, occupancy can exceed 100%.) Comparing the two basic measures, ward capacity and ALOS, we observe in Table 1 that the wards differ in both. Indeed, Ward B is significantly the smallest and the "fastest" (shortest ALOS) among Wards A-D. We observe that the mean occupancy rate in this ward is high (comparable to that in Wards A and D; higher than in Ward C). In addition, the number of patients hospitalized per month in this ward is about 90% of those hospitalized per month in the other wards, although its size is merely about 2/3 of the others. As a consequence, the flux in Ward B (6.38 patients per bed per month) is significantly the highest. Since nursers and doctors in Anonymous Hospital are assigned to particular wards and are salaried workers (as opposed to hourly workers), the load on the Ward B staff is hence the highest. Due to the fact that the staff-to-beds ratio is fixed, if ALOS is kept constant, the occupancy rate in a ward serves a measure of the load on the staff in that ward.

Short ALOS could be caused by multiple reasons. For example, it can result from a superior efficient clinical treatment, or a liberal (vs. conservative) release policy; a clinically too-early discharge of patients is clearly undesirable. One possible (and accessible) quality measure of clinical care is patients' *return rate* (proportion of patients who are re-hospitalized in the IWs within a certain period of time – in our case, three months). In Table 1 we observe that the return rate in Ward B does not differ significantly from the other wards. Also, patients' satisfaction level in surveys – another measure of care quality – does not differ in Ward B [17]. Furthermore, the difference in ALOS across wards could be due to different wards treating different types of patients and/or performing different types of procedures. However, our data do not support this hypothesis. In fact, according to Table 2, Ward B handles a disproportionate share of special-case and ventilated patients, patients that require longer ALOS on average. Rather, a shorter ALOS in Ward B can be attributed to superior staff practices. We conclude that the most efficient ward, instead of being rewarded, is exposed to the highest load; hence, patient allocation appears unfair, as far as the wards are concerned. Increasing fairness in the routing process is expected to increase staff satisfaction, as well as provide incentives for improved care and cooperation.

TABLE 2. Numbers of admitted patients to Wards A-D for each patient category

| IW\Patient Type | Regular | Special-care | Ventilated | Total |
|---|---|---|---|---|
| Ward A | 2,316 (50.3%) | 2,206 (47.9%) | 83 (1.8%) | 4,605 (25.2%) |
| Ward B | 1,676 (43.0%) | 2,135 (**54.7%**) | 90 (**2.3%**) | 3,901 (21.4%) |
| Ward C | 2,310 (49.9%) | 2,232 (48.2%) | 88 (1.9%) | 4,630 (25.4%) |
| Ward D | 2,737 (53.5%) | 2,291 (44.8%) | 89 (1.7%) | 5,117 (28.0%) |
| Total | 9,039 (49.5%) | 8,864 (48.6%) | 350 (1.9%) | 18,253 |

Data refer to period May 1, 2006 - September 1, 2008 (excluding the months 1-3/07, when Ward B was in charge of an additional sub-ward).

## 3. Model Formulation

We model the ED-to-IW process as a queueing system with heterogeneous pools of i.i.d. (independent and identically distributed) servers. Arrivals to the system are patients to-be-hospitalized in the IWs, pools represent the IWs, which indeed have different service rates (1/ALOS), and the number of servers in each pool corresponds to the number of beds in each ward. In order to create a tractable system, we assume that arrivals to the wards occur according to a Poisson process, and LOS (Lengths of Stay) in wards are exponentially distributed. (Both assumptions are important for analytical tractability, but they are inaccurate reality-wise – see [33] for empirical findings on arrivals and LOS). Although, as remarked in Section 2.1, patients to-be-hospitalized in the IWs are classified into several categories, we analyze here a single customer class model. This, as well as our distributional assumptions, certainly could present a limitation for application of our theoretical results, but we are still able to draw useful insights about fair routing in hospitals.

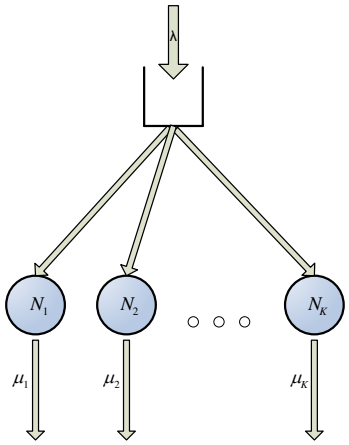
3.1. **The Inverted-V Model.** Consider the queueing system shown in Figure 1. This $\wedge$-model, or inverted-V model (in the terminology of [1]) consists of $K$ server pools: pool $i$ has $N_i$ i.i.d. servers, each with exponential service times of rate $\mu_i$ (namely service rates are equal within each pool but vary among the pools). The total number of servers in the system is $N = \sum_{i=1}^{K} N_i$. Upon arrival, a customer is routed to one of the available pools (if it has one or more idle servers), or joins a centralized queue of infinite capacity, if all the servers at all pools are busy. Homogeneous customers arrive according to a Poisson process with rate $\lambda > 0$. Each customer requires a service that can be provided by any of the servers, and each server can serve only one customer at a time. The queueing discipline is FCFS, non-preemptive (the service of a customer can not be interrupted once started), and work-conserving (there are no idle servers whenever there are customers awaiting service in the queue). In addition, we assume that all interarrival and service times are statistically independent.

*Remark.* Our model is centered around beds rather than personnel. In the setup of Anonymous Hospital (see Sections 2.3 and 2.4), there exists a direct connection between the number of patients (occupied beds) and staff workload in a ward, a key feature that is utilized in our model. It is possible to conceive of an alternative, more complex, model that focuses on the nursing staff directly; nevertheless, as explained in Section 6.2, our simpler model still captures the essentials of fair routing, if looked at through the appropriate "lenses".

3.2. **The QED Asymptotic Regime.** In this section, we formally introduce the Quality and Efficiency Driven (QED) regime, in the context of the inverted-V model. We provide some justification for its applicability in modeling the ED-to-IW process as well.

The QED regime was first discovered by Erlang [18]; it was mathematically formalized by Halfin and Whitt [24], hence it is often referred to as the Halfin-Whitt regime. The regime can

FIGURE 1. Inverted-V system.




be informally characterized in terms of any one of the following conditions: a (steady-state) system with a large volume of arrivals (demand) and many servers (supply, or capacity) is operating in the QED regime if (i) the delay probability is neither near 1 nor near 0, or (ii) its waiting-time is one order of magnitude shorter than service-time (e.g. seconds vs. minutes in call centers, hours vs. days in our case), or (iii) its total service capacity is equal to its demand up to a safety capacity, which is of the same order of magnitude as the square root of the demand. Characterizations (i) and (ii) relate to the quality aspect, and characterization (iii) points at high server efficiencies – thus the QED regime achieves high levels of both service quality and system efficiency, by carefully balancing between the two [1, 19]. The suitability of the QED regime to the ED-to-IW process was studied in detail in [33]. Here we emphasize the following relevant empirical facts:

- The number of servers (beds) in each pool (ward) is around 30-50 (Table 1) – the system is large enough for QED approximations to apply [9, 54, 52].
- Servers' utilization (bed occupancy) is above 85% (Table 1).
- Waiting times are indeed an order of magnitude shorter than service times: hours vs. days. Observe that a waiting time in our inverted-V model corresponds to a patient waiting in the ED due to only lack of available beds in IWs, i.e. bed allocation time – the amount of time from the decision to hospitalize the patient until a bed becomes available in one of the IWs. In particular, if a bed is available at the time of the hospitalization decision, the corresponding bed allocation time is zero. However, estimating bed allocation times is difficult since they are not tracked by Anonymous Hospital's information systems. Hence, we estimated the time between the hospitalization decision and the time of the first procedure in the IW (see Appendix A.3). This time serves as an upper bound on bed allocation time. Since the latter is a major component of the total waiting time in the ED, we conclude that both quantities are in the order of hours. A service time in the inverted-V model corresponds to the interval from the time when a patient occupies a bed until the time the patient releases the bed (or, more precisely, until the bed becomes available next).
- The probability of encountering an available bed in the designated ward, upon hospitalization decision (relatively to the wards' standard capacities), is estimated to be 43%, 48%, 76% and 55%, for Wards A-D respectively. The probability of encountering an available bed in any of the wards is approximately 84%; this estimate is based on a

long period of time, which includes lightly-loaded periods (the arrival process is non-stationary). If one considers only highly-loaded winter months (November to April), the probabilities of encountering an available bed in Wards A-D are 29%, 35%, 64% and 45%, respectively; the probability of encountering an available bed in any of the wards is 75% (this probability is the lowest in January – 59%).

*Remark.* The probability of being admitted to a ward immediately (or within a short time) after the hospitalization decision is much smaller than the probability of encountering an available bed (see Figure 7 in Appendix A.3). Indeed, during the period May 1, 2006 - October 30, 2008 (excluding the months 1-3/2007), only 2.7% of the patients were admitted to an IW within 15 minutes from their assignment to a ward. This fact is consistent with the *Efficiency Driven* regime. We further note that, in evening shifts (when most patients are admitted [33]), there usually are one or two doctors in each ward; hence, one expects that they operate under high loads. In that case, the probability of encountering an available doctor, which is a prerequisite for being admitted to a ward, should indeed be at ED levels (but we do not have any data on staff availability). We thus have two parallel queueing systems: beds, which are QED, and medical staff, who are ED. As already indicated, we focus on the former.

We use the following scaling, suitable for the inverted-V model. Consider a sequence of systems, indexed by $\lambda$ (to appear as a superscript), with increasing arrival rates $\lambda \to \infty$, and increasing total number of servers $N^\lambda$, but with fixed service rates $(\mu_1, \ldots, \mu_K)$. Then, the service capacity of pool $i$ is $c_i^\lambda = N_i^\lambda \mu_i$, the total service capacity is $c^\lambda = \sum_{i=1}^K c_i^\lambda$, and the total traffic intensity is $\rho^\lambda = \lambda / c^\lambda$. Both $\lambda$ and $N^\lambda$ tend to infinity simultaneously, in a way that two limiting relations are satisfied. First, for large $\lambda$, each pool has a non-negligible fraction of the total capacity:

$$c_i^\lambda / c^\lambda \to a_i, \quad \text{as } \lambda \to \infty, \tag{C1}$$

where $a_i > 0$ $(i = 1, 2, \ldots, K)$ and $\sum_{i=1}^K a_i = 1$. The scalar $a_i$ is the limiting proportion of the service capacity of pool $i$ $(i = 1, 2, \ldots, K)$, out of the total capacity. Second, it is convenient to define a scaling parameter:

$$\nu^\lambda := \lambda / \hat{\mu},$$

where $\hat{\mu}$ is the arithmetic-mean service rate:

$$\hat{\mu} := \sum_{i=1}^K a_i \mu_i;$$

it is appropriate to think of $\nu^\lambda$ as an effective system size. Then, the following condition is assumed to hold:

$$\sqrt{\nu^\lambda}\,(1 - \rho^\lambda) \to \delta > 0, \quad \text{as } \lambda \to \infty. \tag{C2}$$

This limit implies $\rho^\lambda \to 1$ (high utilization), as $\lambda \to \infty$, and is (asymptotically) equivalent to the classical square-root safety staffing rule: the total service capacity, $c^\lambda = \lambda + \delta\sqrt{\lambda\hat{\mu}} + o(\sqrt{\lambda})$, is equal to the arrival rate, $\lambda$, plus a square root safety capacity, $\delta\sqrt{\lambda\hat{\mu}}$, where $\delta$ is some quality-of-service parameter (the larger the value of $\delta$, the higher the service quality); here, $\delta$ is a unitless quantity, while $c^\lambda$, $\lambda$ and $\sqrt{\lambda\hat{\mu}}$ are measured in the same units (e.g. patients/week). Note that fluctuations of the arrival process from its mean ($\lambda$) are proportional to $\sqrt{\lambda}$. With $\mu$ being the harmonic-mean service-rate:

$$\mu^{-1} := \sum_{i=1}^K a_i \mu_i^{-1},$$

(C1) and (C2) then imply

$$\frac{\lambda}{N^\lambda} = \frac{\lambda}{\sum_{i=1}^K c_i^\lambda/\mu_i} \to \mu,$$

as $\lambda \to \infty$. In view of this, (C2) can be rewritten as

$$\sqrt{N^\lambda}(1-\rho^\lambda) \to \beta := \delta\sqrt{\hat{\mu}/\mu},$$

as $\lambda \to \infty$. Finally, we define $q_i$ to be the limiting fraction of pool $i$ servers, out of the total number of servers:

$$\frac{N_i^\lambda}{N^\lambda} \to \frac{a_i}{\mu_i}\mu := q_i, \tag{2}$$

as $\lambda \to \infty$; the limit is due to (C1) and (C2). Since $a_i > 0$, for all $i$, we also have that $q_i > 0$, for all $i$, i.e., the pools are of comparable sizes. Clearly, $\sum_{i=1}^K q_i = 1$ and $\sum_{i=1}^K q_i\mu_i = \mu$. Therefore, one can interpret $\mu$ as the (limiting) average service rate of a server in the system, whereas $\hat{\mu}$ is the (limiting) average service rate at which customers receive service. The two quantities differ because faster servers serve more customers. (Note that $\mu \le \hat{\mu}$, with equality if and only if all $\mu_i$'s are equal to each other, in which case $\beta = \delta$ and $\nu^\lambda = N^\lambda$.)

## 4. Fair Routing

Before introducing formally two criteria of fairness, within the context of the inverted-V model, we need some notation. Denote by $I_i^\lambda$ the long-run (steady-state) number of idle servers at pool $i$ ($i = 1, 2, \ldots, K$); each $I_i^\lambda$ is a random variable that attains values in $\{0, 1, \ldots, N_i^\lambda\}$ ($i = 1, 2, \ldots, K$). Let $\rho_i^\lambda := 1 - \mathbb{E}I_i^\lambda/N_i^\lambda$ be the mean long-run (steady-state) occupancy rate in pool $i$. As the servers within each pool are symmetric, $\rho_i^\lambda$ also stands for the utilization of servers in pool $i$ – the fraction of time that each server is busy in the long-run (steady-state). Finally, by $\gamma_i^\lambda$ we denote the average *flux* through pool $i$ (average number of service completions per pool $i$ server per time unit): $\gamma_i^\lambda = \mu_i\rho_i^\lambda$, by Little's law. Clearly, $\gamma_i$ stands also for the average effective service rate of a server in pool $i$.

When analyzing fairness towards servers, we consider the following two criteria: *occupancy* balancing and *flux* balancing. Server's utilization or, equivalently, pool's occupancy rate, is one of the prevalent measures of servers workload. As the occupancy rates at all pools tend to one in the QED regime (see (C2)), we thus compare the ratio between the proportions of *idle* servers in the pools $(1-\rho_i^\lambda)/(1-\rho_j^\lambda)$, referred to as the *idleness ratios*. The closer these ratios are to unity, the more balanced the routing is, according to the occupancy-balancing criterion. The second criterion takes into account the additional measure of workload – the average "flux" through the pools – namely, the number of customers served by a server per time unit. Hence, our second criterion is based on the *flux ratios* $\gamma_i/\gamma_j$. The closer this ratio is to unity, the more balanced the routing is, according to the flux-balancing criterion. Note that, in the QED regime, $\gamma_i/\gamma_j \to \mu_i/\mu_j$, as $\lambda \to \infty$, for any work-conserving routing algorithm. That is, pools with higher service rates experience higher flux. Hence, based on the discussion in Section 2.2, it is appropriate to have lower utilization for pools with higher service rates. Equivalently, if the flux ratio for two pools is greater than one (due to the difference in the service rates), then the idleness ratio should be greater than one as well. The algorithms we discuss in the next subsection all achieve this goal.

### 4.1. Routing Algorithms.

In this subsection, we describe three routing algorithms. All three are work conserving, a choice that is dictated by our goal to reduce queue length (or, equivalently, waiting time by Little's law) rather than the number-in-system (sojourn time). In the ED-to-IW setting, the objective is to minimize the "queue" for transfer to the wards, thus reducing the overload on the ED. Work-conservation is discussed further in the remark at the end of the present subsection.

Next, we describe the three routing algorithms and their implications on server fairness. Let $I_i^\lambda(t) \in \{0, \ldots, N_i^\lambda\}$ denote the number of idle servers in pool $i$ ($i = 1, 2, \ldots, K$) at time $t$. When there are no customers awaiting service at time $t$, the vector $(I_1^\lambda(t), \ldots, I_K^\lambda(t))$ specifies the state of the system. However, when the waiting queue is not empty, an additional variable is needed to specify the number of customers awaiting service; let $Q^\lambda(t) \in \{0, 1, \ldots\}$ be the number of customers awaiting service at time $t$. It is convenient to define $I^\lambda(t) \in \{\ldots, N^\lambda - 1, N^\lambda\}$ as (jointly) the total number of idle servers awaiting customers/number of customers awaiting servers, at time $t$:

$$I^\lambda(t) = \sum_{i=1}^{K} I_i^\lambda(t) - Q^\lambda(t). \tag{3}$$

Note that $I^\lambda(t)$ can take negative values: $\{I^\lambda(t) = -i\}$, $i \geq 1$, indicates that there are $i$ customers awaiting service at time $t$. Due to work-conservation, one has $\sum_{i=1}^K I_i^\lambda(t) = (I^\lambda(t))^+$ and $Q^\lambda(t) = (I^\lambda(t))^-$, where $x^+$ and $x^-$ denote the positive and negative part of $x$.

*LISF routing.* The Longest-Idle Server First (LISF) policy routes a customer to a server that has been idle for the longest time, among all idle servers. This policy is commonly used in call centers and considered to be fair [3]. It was first analyzed (in the QED regime) by Atar [4], in the context of a single pool system in a random environment (service rates were taken to be i.i.d. random variables). Armony and Ward [3] analyzed LISF routing in the inverted-V model with two ($K = 2$) pools. Informally, they show that, for large $\lambda > 0$, LISF maintains fixed ratios between the number of idle servers in different pools, whenever idle servers exists; that is, if $I^\lambda(t) > 0$ at time $t$, then

$$\frac{I_i^\lambda(t)}{I_j^\lambda(t)} \approx \frac{a_i (I^\lambda(t))^+}{a_j (I^\lambda(t))^+} = \frac{a_i}{a_j}.$$

Hence, up to some technical conditions (see the discussion prior to Corollary 4.2 in [23]), under LISF routing one expects

$$\frac{1-\rho_i^\lambda}{1-\rho_j^\lambda} = \frac{\mathbb{E}I_i^\lambda}{N_i^\lambda} \Big/ \frac{\mathbb{E}I_j^\lambda}{N_j^\lambda} \to \frac{a_i q_j}{a_j q_i} = \frac{\mu_i}{\mu_j},$$

and $\gamma_i^\lambda / \gamma_j^\lambda \to \mu_i / \mu_j$, as $\lambda \to \infty$, i.e., both the idleness and flux ratios tend to the ratios of server rates. The algorithm leads to a desirable outcome: fast servers work less (have lower utilization) but "produce" more (have higher flux).

Note that even though LISF is a "blind" policy (a policy that requires, at the time of routing decision, none or minimal information on the parameters of the system, or the system state [5]), implementing the LISF policy in the hospital setting is not straightforward. Namely, one must keep track not only on the number of idle servers (beds) in each pool (ward), but also the relative ordering of idle servers in terms of their idle times. The latter information is not currently available in hospitals. This fact motivates us to consider alternative routing policies that achieve the same (asymptotic) fairness towards servers but utilize less information for customer routing.

*IR routing.* The Idleness-Ratio (IR) routing policy is a way to achieve the same idleness ratio as LISF, but without the information on idleness times. This policy is a special case of Queues-and-Idleness-Ratio (QIR) policies, which were proposed and analyzed by Gurvich and Whitt [23]. They consider a generalization of the inverted-V model, a parallel-server system – a service system with multiple server pools and multiple customer classes. The problem of dynamic control of such systems is often referred to as "Skill-Based Routing" (borrowing the terminology from the world of call centers). Adopting the QIR policy to the inverted-V

model, IR routes a customer to the server pool with the highest idleness imbalance. The basic idea is to route each customer in such a way that the vector $(I_1^\lambda(t),\ldots,I_K^\lambda(t))$ is as close to $(w_1(I^\lambda(t))^+,\ldots,w_K(I^\lambda(t))^+)$ as possible, where the set of positive constants $\{w_i\}$ is a priori fixed, with $\sum_{i=1}^K w_i = 1$. For example, if $K = 2$ and $w_1 = w_2 = 1/2$, then the IR policy routes a customer to the pool with the higher number of idle servers. Specifically, at time $t$, a customer is routed to the pool with the index

$$\arg\max_i \left\{ I_i^\lambda(t-) - w_i(I^\lambda(t-))^+ \right\}.$$

Ties are broken in an arbitrary, but consistent fashion; the tie-breaking rule does not impact the results at the diffusion $(\sqrt{\nu^\lambda})$ scale. In [23] it was shown that (Q)IR controls drive the idleness process to the predetermined proportions $\{w_i\}$, in the QED regime. Following the same argument as in the LISF case, one expects

$$\frac{1-\rho_i^\lambda}{1-\rho_j^\lambda} = \frac{\mathbb{E}I_i^\lambda}{N_i^\lambda} \Big/ \frac{\mathbb{E}I_j^\lambda}{N_j^\lambda} \to \frac{w_i q_j}{w_j q_i},$$

and $\gamma_i^\lambda/\gamma_j^\lambda \to \mu_i/\mu_j$, as $\lambda\to\infty$. Therefore, with the IR algorithm, one can achieve the same ratio of the number of idle server (and idleness ratios) as under the LISF algorithm, by simply setting $w_i = a_i$, since $(w_i q_j)/(w_j q_i) = \mu_i/\mu_j$ (see (2)). Given its weights, the IR policy is a blind policy as only the information on the number of idle servers in each pool is needed. However, determining the values of $a_i$'s is far from straightforward in the hospital setting. In particular, note that $a_i$ represents the limiting ratio between the pool $i$ service capacity $(c_i^\lambda)$ and the total service capacity $(c^\lambda)$. Therefore, since one must estimate the service rates in order to evaluate the appropriate weights, the considered version of the policy is, effectively, not blind. Recall from Section 2 that the capacity (number of beds) of each ward can vary with time, e.g., during the first three months of 2007 Ward B was in charge of an additional sub-ward. These facts serve as a motivation for considering yet a third routing algorithm, based on further reduced information.

*RMI routing.* We now introduce the Randomized Most-Idle (RMI) routing policy. As our results in the next section indicate, the RMI policy achieves the same ratios of idleness between server pools as the LISF and IR (with $w_i = a_i$) policies on the diffusion scale, and yet it requires information on neither idleness times nor on pool capacities. Under RMI, a customer is assigned to one of the available pools, *with probability* that equals the fraction of idle servers in that pool out of the overall number of idle servers in the system (hence, the name of the policy). That is, a customer to be routed at time $t$ is assigned to pool $i$ with probability $I_i^\lambda(t-)/(I^\lambda(t-))^+$. In [45] it was observed that the RMI policy, in the inverted-V model, is equivalent to the Random Assignment (RA) policy in the single server pool model with $N^\lambda$ servers, where $N_i^\lambda$ servers have rate $\mu_i$ $(i = 1,\ldots,K)$. To illustrate this, consider the following example. Assume that a customer arrives to the system with two available pools: pool $i$ has 2 available servers and pool $j$ has 3 available servers. Thus, the customer is routed to pool $i$ with probability $2/5$ and to pool $j$ with probability $3/5$. As in pool $i$ there are two available i.i.d. servers, each one of them will serve this customer with probability $1/5$; similarly, each server in pool $j$ will serve the customer with probability $1/5$. As a consequence, the customer is assigned to any one of the five available servers with equal probabilities.

An analytically appealing feature of the RMI policy is that, when modeled as a Markov chain in continuous time, the system is reversible [28] (in [45] it is conjectured that this is the only routing policy under which the inverted-V system induces a reversible Markov chain). Therefore, one is able to derive its steady state probabilities in a straightforward manner and provide an exact analysis in steady state. (Due to their complexities, LISF and IR policies were analyzed only asymptotically.) Finally, we note that, even though RMI is a randomized

policy, it can be easily implemented in a hospital setting. For example, patient ID numbers can be utilized as sources of randomness (as stated in Appendix A.4, at least one hospital in our survey has implemented some randomized policy, based on patient ID numbers).

*Remark* (Why work-conserving?). The three considered algorithms are work-conserving. This assumption is not restrictive from the point of view of both hospital and patients. The goal is to reduce the waiting times in the ED so that the number of patients awaiting transfer from the ED to the IWs is minimized. Therefore, implementing a work-conserving policy is desirable since intentional server idling only increases the number of patients in the ED. In [45], it was shown that, under the waiting time criterion and RMI routing, it is not beneficial to discard slow servers, i.e., it is not desirable to intentionally idle servers. On the other hand, when the objective is to minimize the sojourn time (waiting plus service times), cases arise when it is beneficial to discard slow servers in a system with heterogeneous servers. Within the context of the well-known *slow-server* problem, this was shown for two serves [39] and many servers [10]. Namely, if some servers are slow enough, roughly speaking, it could turn out preferable for a waiting customer to wait for a fast server to become available rather than start service at a slow server. Thus, as far as sojourn time is concerned (but not waiting time), a non work-conserving policy could turn out preferable.

## 5. Theoretical Results

The following theorem characterizes RMI performance, in the non-asymptotic regime (finite arrival rate $\lambda$). The theorem states that, when comparing two pools, servers utilization in the faster pool is lower than that in the slower pool, but the flux in the faster is higher than in the slower. Due to the symmetry of servers within each pool, this implies that, when considering any two servers, the faster between the two will work less time than the slower one but, at the same time, the faster server will serve more customers than the slower. The result suggests, first of all, some form of fairness: faster servers are "rewarded" by working less time. In addition, operational preferences of the system are accommodated as well: more customers are served by faster servers than by slower servers. We note that Cabral [11] proved this result, for the single-server-pool system under the RA policy, independently.

**Theorem 1.** *In the inverted-V model under the RMI policy, for any two pools $i$ and $j$: if $\mu_i > \mu_j$, then $\rho_i^\lambda < \rho_j^\lambda$ and $\gamma_i^\lambda > \gamma_j^\lambda$.*

*Proof.* See Appendix B.1. □

The theorem provides an upper and lower bound on the ratio of server utilizations (assuming $\mu_i > \mu_j$): $\mu_j/\mu_i < \rho_i^\lambda/\rho_j^\lambda < 1$. This suggests that the difference in utilizations of any two servers is more significant the more their service rates differ: for $\mu_j \approx \mu_i$, one has $\rho_j^\lambda \approx \rho_i^\lambda$, but as $\mu_j$ grows smaller than $\mu_i$, the server-utilization ratio decreases. The bounds on the flux ratios are as follows (assuming $\mu_i > \mu_j$): $1 < \gamma_i^\lambda/\gamma_j^\lambda < \mu_i/\mu_j$, where the second inequality follows from $\rho_i^\lambda/\rho_j^\lambda < 1$. The latter upper bound is important – although the fact that faster servers serve more customers contributes to system performance, one should not forget that, in certain cases, higher flux actually implies higher workload. In particular, this is the case in our ED-to-IW process, because service admissions and releases impose workload that is plausibly proportional to flux. For the RMI policy, the servers' flux ratio $\gamma_i^\lambda/\gamma_j^\lambda$ is bounded by the ratio of their service rates: $\gamma_i^\lambda/\gamma_j^\lambda = (\rho_i^\lambda\mu_i)/(\rho_j^\lambda\mu_j) < \mu_i/\mu_j$ when $\mu_i > \mu_j$; thus, if server rates are comparable, a faster server's flux can not be much higher than that of a slower one.

*Remark.* The fact that utilization decreases the faster the server gets, provides an incentive for servers to work faster, which is positive on one side but, on the other, might harm service quality, if one starts serving customers *too* fast. For example, see Gans et al [19] who describe

telephone agents that intentionally hang up on customers in order to maintain low average service times. Another issue is that, if the slow server is not responsible for being slow (for example, a new server vs. a veteran), "punishing" the server for being slow appears quite unfair. However, as noted earlier, higher flux may be considered in certain cases a "punishment" as well. Hence, in order to decide which servers perceive themselves as better off: the faster or the slower; or alternatively, what are servers' incentives (to increase or decrease his/her service rate?), one must account for the servers utility functions (the ones they strive to optimize), which combine both criteria (utilization and flux) – for an example of such a utility function recall (1).

The next theorem characterizes the inverted-V model under RMI routing in the QED regime. It can serve as a means for evaluating performance measures (probability of wait, expected waiting time, etc.) of the inverted-V model in the QED regime. A summary of performance measures, both for finite $\lambda$ and in the limit, as $\lambda \to \infty$, can be found in Appendix B.4. Recalling (3), let $I^\lambda$ be the *stationary* total number of idle servers/customers awaiting service in the system with arrival rate $\lambda$; the random variable $I^\lambda$ takes values in $\{\dots, N^\lambda - 1, N^\lambda\}$, by definition $(I^\lambda)^+ = \sum_{i=1}^K I_i^\lambda$, and $\{I^\lambda = i\}$ for negative $i$ indicates that there are $|i|$ customers awaiting service. As the system size increases ($\lambda \to \infty$), the variability of demand increases as well, implying that the magnitude of $I^\lambda$ gets larger and larger. Hence, in order to gain understanding of the behavior of $I^\lambda$, we consider its scaled version:

$$\hat{I}^\lambda := I^\lambda / \sqrt{\nu^\lambda};$$

such a scaling is typical for the QED regime and is referred to as a diffusion scaling. Informally, the theorem states that, in stationarity, there exists a dimensionality reduction in the sense that the stationary number of idle servers is pool $i$ satisfies $I_i^\lambda \approx a_i (I^\lambda)^+$, for large $\lambda$, i.e., idle servers are distributed across the pools proportionally, according to the relative pool capacities. Moreover, the theorem turns out to provide an explicit estimate of how close $I_i^\lambda$ is to $a_i(I^\lambda)^+$. In particular, fluctuations of $I_i^\lambda$ around $a_i(I^\lambda)^+$ grow with $\lambda$ at a rate $\sqrt[4]{\nu^\lambda}$, and thus we consider

$$\hat{I}_i^\lambda := \frac{1}{\sqrt{I^\lambda}} \left( I_i^\lambda - \frac{c_i^\lambda}{c^\lambda} I^\lambda \right), \quad i = 1, \dots, K.$$

We refer to the $\sqrt[4]{\nu^\lambda}$ scale as a sub-diffusion scale since $\sqrt[4]{\nu^\lambda} \ll \sqrt{\nu^\lambda}$, for large $\lambda$. Characterization of RMI behavior on the sub-diffusion scale provides additional insights into the operation of the algorithm – we discuss this further in the next subsection. Denote by $\varphi(\cdot)$ and $\Phi(\cdot)$ the standard normal density and distribution functions, respectively. Let $\Rightarrow$ denote convergence in distribution.

**Theorem 2.** *Consider the inverted-V model in steady-state, under the RMI routing algorithm in the QED regime (C1-C2). Then, as $\lambda \to \infty$,*

$$\left( \hat{I}^\lambda, (\hat{I}_1^\lambda, \dots, \hat{I}_K^\lambda) 1_{\{\hat{I}^\lambda > 0\}} \right) \Rightarrow \left( \hat{I}, (\hat{I}_1, \dots, \hat{I}_K) 1_{\{\hat{I} > 0\}} \right), \tag{4}$$

*where $\hat{I}$ and $(\hat{I}_1, \dots, \hat{I}_K)$ are independent;*

$$\mathbb{P}[\hat{I} \le 0] = \left( 1 + \delta \frac{\Phi(\delta)}{\varphi(\delta)} \right)^{-1};$$

$\mathbb{P}[\hat{I} > x \,|\, \hat{I} > 0] = \Phi(\delta - x)/\Phi(\delta)$, $x \ge 0$; $\mathbb{P}[\hat{I} \le x \,|\, \hat{I} \le 0] = e^{\delta x}$, $x \le 0$; *and* $(\hat{I}_1, \dots, \hat{I}_K)$ *is zero-mean multi-variate normal, with* $\mathbb{E}\hat{I}_i \hat{I}_j = a_i 1_{\{i=j\}} - a_i a_j$.

*Proof.* See Appendix B.2. □

The contribution of the theorem to understanding system behavior under RMI is twofold. First, RMI achieves the same server fairness as LISF and IR (with appropriate weights) in the sense that idle servers are distributed across the pools according to the pools relative capacities ($a_i$'s). This is of significance since RMI requires less information for its operation than LISF and, unlike IR, RMI does not utilize information on pool capacities. Second, the "quality" of allocation of idle servers across the pools under RMI is revealed. Specifically, the number of idle servers in a pool deviates from $(c_i^\lambda/c^\lambda)I^\lambda$ (a number determined by pools relative capacities) by a random quantity (normally distributed) of the order $\sqrt[4]{\nu^\lambda}$. In view of the fact that the number of idle servers in a pool is proportional to $\sqrt{\nu^\lambda}$, fluctuations of order $\sqrt[4]{\nu^\lambda}$ are negligible, when $\lambda$ is not small.

*Remark* (Equal service rates). When the service rates are equal across the server pools, i.e., $\mu_1 = \mu_2 = \ldots = \mu_K$, then $\mu = \hat{\mu} = \mu_1$, $\delta = \beta$, $N^\lambda/\nu^\lambda \to 1$, as $\lambda \to \infty$, and one recovers the well-known Erlang-C QED approximation [24].

*Remark.* Note that $\sum_{i=1}^K \hat{I}_i^\lambda 1_{\{\hat{I}^\lambda > 0\}} = 0$ by definition. The limit (4) is consistent with this condition: $\mathbb{P}[\sum_{i=1}^K \hat{I}_i = 0] = 1$, since $\mathbb{E}(\sum_{i=1}^K \hat{I}_i)^2 = 0$.

*Remark.* The dimensionality reduction (on the diffusion scale) can be deduced from hydrodynamical equations in [14]. For example, consider the case when there are $K = 2$ pools. If the fraction of idle servers that are in the first pool exceeds $c_1^\lambda/c^\lambda$, then a disproportionate number of customers will be routed to the first pool, resulting in a lower number of idle servers in the first pool. As similar situation occurs when the fraction of idle servers that are in the first pool drops below $c_1^\lambda/c^\lambda$. Hence, the ratio of the number of idle servers in different pools should be equal to the ratio of the pool capacities.

*Example* 1 (Probability of delay). Consider an inverted-V system with the following parameters: $K = 2$, $q_2 = 2q_1 = 2/3$ and $\mu_1 = 2\mu_2 = 2$ (e.g. patients/week). The total number of servers (e.g. beds) is taken to be

$$N^\lambda = \left\lceil \frac{\lambda + 0.5\sqrt{\lambda}}{q_1\mu_1 + q_2\mu_2} \right\rceil,$$

while $N_1^\lambda = \text{round}(q_1 N^\lambda)$ and $N_2^\lambda = N^\lambda - N_1^\lambda$. We vary the arrival rate $\lambda$ from 10 to 500 – this corresponds to varying $N_1^\lambda$ and $N_2^\lambda$ from 3 to 128 and 6 to 256, respectively. In Figure 2, we plot the exact probability of delay along its QED approximation. Note that, due to the PASTA property, the probability of delay is equal to $\mathbb{P}[I^\lambda \leq 0]$ and, hence, Theorem 2 renders

$$\mathbb{P}[\text{delay}] \to \left(1 + \delta \frac{\Phi(\delta)}{\varphi(\delta)}\right)^{-1}, \tag{5}$$

as $\lambda \to \infty$. The preceding limit serves as the basis for calculating QED approximation for the finite system. To this end, the required QED parameter $\delta$ changes slightly with $\lambda$ due to the fact that the number of servers in each pool is a natural number. Thus, when evaluating the QED approximation, we compute $\delta$ for each value of $\lambda$:

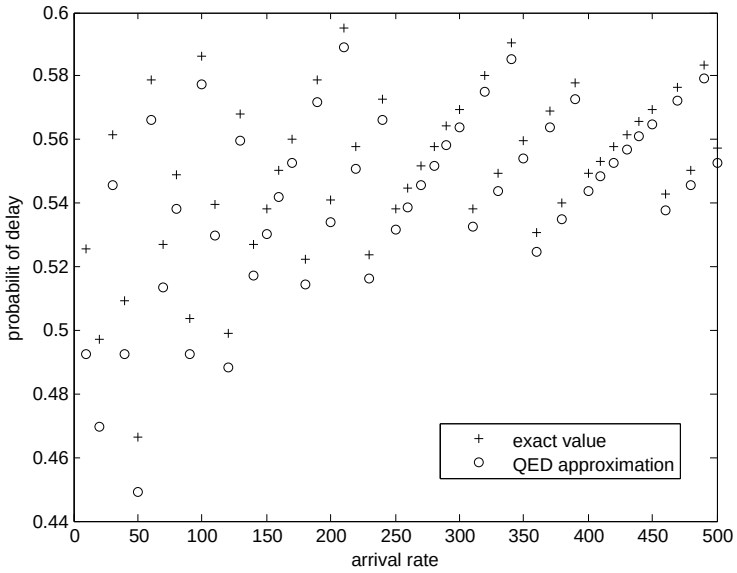
$$\delta^\lambda = \left(1 - \frac{\lambda}{N_1^\lambda\mu_1 + N_2^\lambda\mu_2}\right)\sqrt{\lambda/\hat{\mu}^\lambda},$$

where

$$\hat{\mu}^\lambda = \frac{N_1^\lambda\mu_1^2 + N_2^\lambda\mu_2^2}{N_1^\lambda\mu_1 + N_2^\lambda\mu_2}.$$

A QED approximation for the probability of delay is obtained by evaluating the right-hand side of (5) with $\delta = \delta^\lambda$. As can be seen in Figure 2, the QED approximation has a reasonable (useful)

FIGURE 2. Exact values and QED approximations of the probability of delay, for the sequence of systems described in Example 1.




accuracy when system sizes and service rates are similar to those in Anonymous Hospital (see Section 2).

*Example* 2 (Dimensionality reduction). To illustrate the dimensionality reduction that arises in Theorem 2, we simulated an inverted-V system under RMI, with parameters that adhere to the QED scaling (C1) and (C2): $K = 2$, $\lambda = 3950$, $\mu_1 = 15$, $\mu_2 = 7.5$, $N_1 = 138$, $N_2 = 276$ ($a_1 = a_2 = 1/2$, $\hat{\mu} = 11.25$, $\delta \approx 0.86$). In Figure 3, we plot typical realizations of the total number of idle servers, $\{I^\lambda(t),\, t \geq 0\}$, and the centered number of idle server in the first pool, $\{I_1^\lambda(t) - a_1(I^\lambda(t))^+,\, t \geq 0\}$. Initially, at time $t = 0$, there are no idle servers and no customers await service, i.e., $I^\lambda(0) = 0$. By time $t = 900$, the system is close to its stationary regime (there are more than $7 \cdot 10^6$ arrivals/departures in the time interval $[0, 900]$). One can observe that the processes $\{I^\lambda(t),\, t \geq 0\}$ and $\{I_1^\lambda(t) - a_1(I^\lambda(t))^+,\, t \geq 0\}$ evolve on two different *counting* scales – the first one on the $\sqrt{\nu^\lambda}$-scale ($\sqrt{\nu^\lambda} \approx 18.7$) and the second one on the $\sqrt[4]{\nu^\lambda}$-scale ($\sqrt[4]{\nu^\lambda} \approx 4.3$).

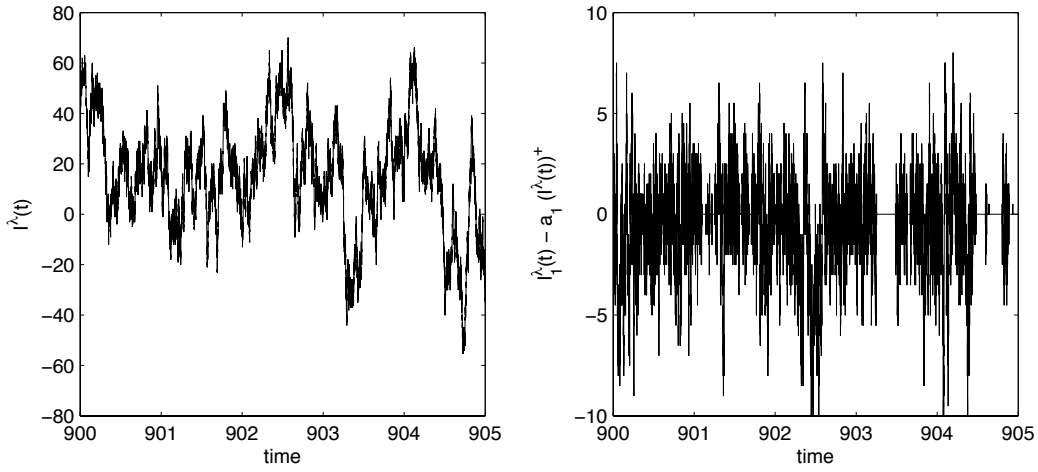
From the following corollary, it is immediate that, under RMI, the idleness ratios satisfy $(1-\rho_i^\lambda)/(1-\rho_j^\lambda) \to \mu_i/\mu_j$, as $\lambda \to \infty$. That is, RMI achieves the same idleness ratios as LISF.

**Corollary 1.** *Consider the inverted-V model in steady-state, under the RMI routing algorithm in the QED regime (C1-C2). Then, as* $\lambda \to \infty$, $\mathbb{E}(\hat{I}^\lambda)^- \to \mathbb{E}(\hat{I})^-$, $\mathbb{E}(\hat{I}^\lambda)^+ \to \mathbb{E}(\hat{I})^+$, *and* $\mathbb{E}I_i^\lambda/\sqrt{\nu^\lambda} \to a_i\mathbb{E}(\hat{I})^+$, *for* $i = 1, \ldots, K$, *where* $\hat{I}$ *is as in Theorem 2.*

*Proof.* See Appendix B.3. □

*Remark* (Loss of performance). As stated in the Introduction, Fastest Servers First (FSF) routing is asymptotically optimal (within the set of all non-preemptive, non-anticipating FCFS policies), in the sense that it stochastically minimizes the stationary queue length and waiting time, as the arrival rate and number of servers grow large in the considered inverted-V model. When analyzing the trade-off between fairness and performance (i..e., RMI vs. FSF), one must consider the following two aspects.

FIGURE 3. Illustration for Example 2.




On one hand, within our mathematical model, fairness comes at the cost of a decrease in system performance (e.g., an increase of the probability of delay). Theorem 2 (and Corollary 1), together with results on FSF routing in [1], provide means for quantifying the loss of performance under RMI (relative to FSF). In particular, results in [1] indicate that, under FSF,

$$\mathbb{P}[\hat{I} \le 0] = \left(1 + \delta_* \frac{\Phi(\delta_*)}{\varphi(\delta_*)}\right)^{-1};$$

$\mathbb{P}[\hat{I} > x \,|\, \hat{I} > 0] = \Phi(\delta_* - x\sqrt{\mu_\wedge/\hat{\mu}})/\Phi(\delta_*)$, $x \ge 0$; $\mathbb{P}[\hat{I} \le \,|\, \hat{I} \le 0] = e^{\delta x}$, $x \le 0$; $\mu_\wedge = \min \mu_i$ and $\delta_* = \delta\sqrt{\hat{\mu}/\mu_\wedge}$. Thus, the probability of delay under RMI is higher (since $\delta_* \ge \delta$) than the probability of delay under FSF by a factor of

$$\frac{1 + \delta_* \Phi(\delta_*)/\varphi(\delta_*)}{1 + \delta \Phi(\delta)/\varphi(\delta)}, \tag{6}$$

that depends on the spare capacity parameter $\delta$ and the ratio of the arithmetic-mean service rate $\hat{\mu}$ to the minimum service rate $\mu_\wedge$; recall that in heavy traffic, only servers with the smallest service rate are idled under FSF. The increase in the expected waiting time is given by the same ratio. This follows from the fact that the conditional expected wait, given that it is positive, is the same under the two routing policies. The decrease in performance should be carefully reviewed in order to determine if increased delays are clinically acceptable. For example, in Anonymous Hospital (see Table 1), $\hat{\mu} \approx 1.18$ patients/week, $\mu \approx 1.08$ patients/week and $\delta \approx 0.87$, leading to the ratio in (6) being equal to 1.07, i.e., the probability of delay increases by 7% when employing RMI instead of FSF.

On the other hand, as stated in Section 1, the issue of fairness needs to be examined beyond the mathematical model with fixed service rates. In particular, ensuring fairness towards medical staff should be viewed within the context of providing a right set of incentives for staff. While the waiting time of customers (patients) is stochastically minimized under FSF, medical staff working in wards with non-minimal service rates have no incentive to further reduce LOS – any improvement would lead to a higher load. Thus, although FSF nominally results in shorter waiting times, RMI can be more beneficial for customers in the long run since the implementation of RMI can eventually lead to shorter LOS and, as a result, shorter waiting times.

*Weighted RMI.* Finally, we note that the RMI algorithm can be generalized to a Weighted RMI (WRMI) as follows. Given a set of weights $w_i > 0$ $(i = 1, \ldots, K)$ such that $\sum_{i=1}^{K} w_i = 1$, the WRMI routes a customer to pool $i$, at time $t$, with probability $I_i^{w,\lambda}(t-)/\sum_{j=1}^{K} I_j^{w,\lambda}(t-)$, where $I_i^{w,\lambda}(t) = w_i I_i^{\lambda}(t)$. The weights can be used to adjust idleness ratios to a desired target (relative to the ratio of service rates) in the QED regime. Unless all weights are equal, the resulting system is not reversible, and thus our analysis of RMI can not be extended to WRMI. However, insights gained from RMI can be used to heuristically analyze WRMI as follows. Consider a time instance $t$ such that $I^{\lambda}(t)/\sqrt{\nu^{\lambda}} > 0$. Then, the departure rate of customers from pool $i$ is $c_i^{\lambda} - \mu_i I_i^{\lambda}(t)$ (where $c_i^{\lambda} \gg \mu_i I_i^{\lambda}(t)$); on the other hand, customers enter service at pool $i$ with rate $\lambda I_i^{w,\lambda}(t)/\sum_{j=1}^{K} I_j^{w,\lambda}(t)$. Therefore, for large $\lambda$,

$$\frac{a_i}{a_j} \approx \frac{c_i^{\lambda} - \mu_i I_i^{\lambda}(t)}{c_j^{\lambda} - \mu_j I_j^{\lambda}(t)} \approx \frac{I_i^{w,\lambda}(t)}{I_j^{w,\lambda}(t)} = \frac{w_i I_i^{\lambda}(t)}{w_j I_j^{\lambda}(t)}.$$

In view of the preceding, one expects that the idleness ratios satisfy, as $\lambda \to \infty$,

$$\frac{1-\rho_i^{\lambda}}{1-\rho_j^{\lambda}} \to \frac{w_j \mu_i}{w_i \mu_j}.$$

### 5.1. **Comments on the Sub-Diffusion Scale.**

*Relevance and Implications.* Based on empirical data from our Anonymous Hospital, we estimated that $\lambda \approx 189.7$ patients/week and $\hat{\mu} \approx 1.18$ patients/week (see Table 1); thus, $\nu^{\lambda} = \lambda/\hat{\mu} \approx 160.8$. These numbers and our theoretical analysis reveal that there exist 3 relevant counting scales (bed, room, sub-ward) and 3 corresponding time scales (hour, day, week). In order to describe a stochastic process of interest (e.g. the number of available beds), one needs to define not only an appropriate counting scale, but also appropriate time intervals (scale) over which relevant changes in the process occur. On time intervals that are too short, no changes in the value of the process can be observed on the counting scale; on the other hand, on time intervals that are too long, the process covers the whole counting scale, and time variability can not be studied.

The finest counting scale is at the level of an individual bed/patient (order-1 scale), while the corresponding time scale is at the level of an hour (order-$1/\lambda$). Indeed, when considering individual patient arrivals, the relevant time scale is based on hours since $1/\lambda \approx 0.86$ hours. The coarsest counting scale (order-$\sqrt{\nu^{\lambda}}$; diffusion scale) is used to describe the total number of available beds and patients awaiting hospitalization. For a large hospital, this scale is defined by a sub-ward, due to the fact that $\sqrt{\nu^{\lambda}} \approx 12.7$ (beds or patients) is roughly the size of 1/4 to 1/3 of a ward, and the number of available beds/patients waiting is proportional to $\sqrt{\nu^{\lambda}}$ (see Theorem 2). The corresponding time scale is defined by a "typical" LOS – a week in our case, since patients stay in a ward a bit less than a week on average (or $1/\hat{\mu} \approx 0.85$ weeks, see Table 1).

The preceding two pairs of scales are standard for systems operating in the QED regime. The third pair of scales is intermediate and due to RMI. The counting scale (sub-diffusion scale) is defined by a room (rooms in Anonymous Hospital have 4 beds) and is an order-$\sqrt[4]{\nu^{\lambda}}$ scale ($\sqrt[4]{\nu^{\lambda}} \approx 3.6$ (beds or patients)). This scale is relevant in describing the number of patients that need to be moved between wards in order to make the *instantaneous* (at a specific moment of time) idleness ratios equal to the long-run (average) idleness ratios introduced earlier. The latter ratios do not provide information on the system behavior at a specific moment of time.

For example, consider the following two scenarios, assuming two symmetric wards: (i) the number of available beds in the two wards is the same at all times; and (ii) the number of available beds in the first ward is higher than in the second one for a long period of time, and

then the situation is reversed for the same amount of time. Under both scenarios, the idleness ratio is unity. The room-level ($\sqrt[4]{\nu^\lambda}$) counting scale provides information on deviations of the numbers of available beds (in steady-state) from the numbers that ensure that the actual ratios of idle servers at a given moment of time are equal to the idleness ratios (average quantities). In particular, our results imply that fluctuations of instantaneous idleness ratios around their long-run averages are of the order $1/\sqrt[4]{\nu^\lambda}$. The associated time scale is based on a day (order-$1/\sqrt{\lambda\hat{\mu}}$; $1/\sqrt{\lambda\hat{\mu}} \approx 0.87$ days) and describes relevant time intervals over which these relatively minor fluctuations of idleness ratios average out. Recall the two scenarios mentioned above. In the second one, several cycles are needed for idleness ratios to converge to unity. Indeed, if one observes the system for a short period of time only, then the system appears unfair. Yet, over long periods of time the system is fair. Finally, we note that the intermediate counting and time scales are related, in the sense that the larger the fluctuations of the number of available beds, the longer time intervals one requires for convergence of the idleness ratios.

Informally, our results indicate that, under RMI, fluctuations of the numbers of idle beds in different wards, around values that ensure $\mu_i/\mu_j$ instantaneous idleness ratios are of the order $\sqrt[4]{\nu^\lambda} \approx 3.6$ (beds or patients), i.e., the room-based counting scale is relevant. Indeed, Theorem 2 implies

$$I_i^\lambda \approx \frac{c_i^\lambda}{c^\lambda} I^\lambda + \hat{I}_i^\lambda \sqrt{I^\lambda},$$

and, hence, the idleness rations obey

$$\frac{I_i^\lambda/N_i^\lambda}{I_j^\lambda/N_j^\lambda} \approx \frac{\mu_i}{\mu_j}\left(1 + \frac{1}{\sqrt{I^\lambda}}\left(\hat{I}_i^\lambda/a_i - I_j^\lambda/a_j\right)\right).$$

Moreover, when available beds exist, even the minor differences (order-$1/\sqrt{I^\lambda}$ or, equivalently, order-$1/\sqrt[4]{\nu^\lambda}$) between instantaneous idleness ratios and the corresponding long-run idleness ratios ($\mu_i/\mu_j$) are averaged out on time intervals of order $1/\sqrt{\lambda\hat{\mu}}$ (time intervals that contain $\sqrt{\nu^\lambda}$-order arrivals/departures are of interest here, since central limit theorem deviations are of the order $\sqrt[4]{\nu^\lambda}$ in that case; consequently, this corresponds to time intervals of length $\sqrt{\nu^\lambda}/\lambda = 1/\sqrt{\lambda\hat{\mu}}$). In Anonymous Hospital, this corresponds roughly to days ($1/\sqrt{\lambda\hat{\mu}} \approx 0.87$ days), and consequently one expects desired idleness ratios (with very high accuracy) on a weekly basis. Interestingly, the same behavior of intermediate scales occurs when the system operates under the LISF rule (see below).

*Technical Discussion.* Even though the three considered algorithms operate in different ways, they result in the same behavior on the diffusion scale. However, differences arise on the sub-diffusion scale. Informally, Theorem 2 states that, for large $\lambda$, RMI deviations of $I_i^\lambda$ around $c_i^\lambda/c^\lambda\,(I^\lambda)^+$ are on the order of $\sqrt[4]{\nu^\lambda}$; note that both $I_i^\lambda$ and $(I^\lambda)^+$ are $\sqrt{\nu^\lambda}$-order random variables. On the other hand, under IR policy, $I_i^\lambda(t) - a_i(I^\lambda)^+$ is an order-1 random variable, as $\lambda \to \infty$. Consequently, although implementing RMI in a hospital setting will not ensure that an instantaneous idleness ratio is equal exactly to the desired value $\mu_i/\mu_j$ at all times, the number of available beds in a ward will only differ from the desired one by a $\sqrt[4]{\nu^\lambda}$-quantity, which is negligible in comparison with the number of available beds in a ward.

Furthermore, it should be noted that, under both IR and RMI, there exists a separation of time scales. Namely, the processes $\{I^\lambda(t)/\sqrt{\lambda},\, t \geq 0\}$ and $\{I_i^\lambda(t)/\sqrt{\lambda},\, t \geq 0\}$ evolve on the order-1 time scale under both policies; this is typical in the QED regime – no time-speedup is needed, in contrast to the case of conventional heavy-traffic. However, the processes $\{I_i^\lambda(t) - a_i\,(I^\lambda(t))^+,\, t \geq 0\}$ and $\{(I_i^\lambda(t) - c_i^\lambda/c^\lambda\,(I^\lambda(t))^+)/\sqrt[4]{\nu^\lambda},\, t \geq 0\}$ evolve on the $\lambda^{-1}$- and $(\lambda\hat{\mu})^{-1/2}$-time scales, as $\lambda \to \infty$, under IR and RMI routing, respectively. Hence, a separation of time scales, since the latter processes evolve on much faster time scales (order-$\lambda^{-1}$ and

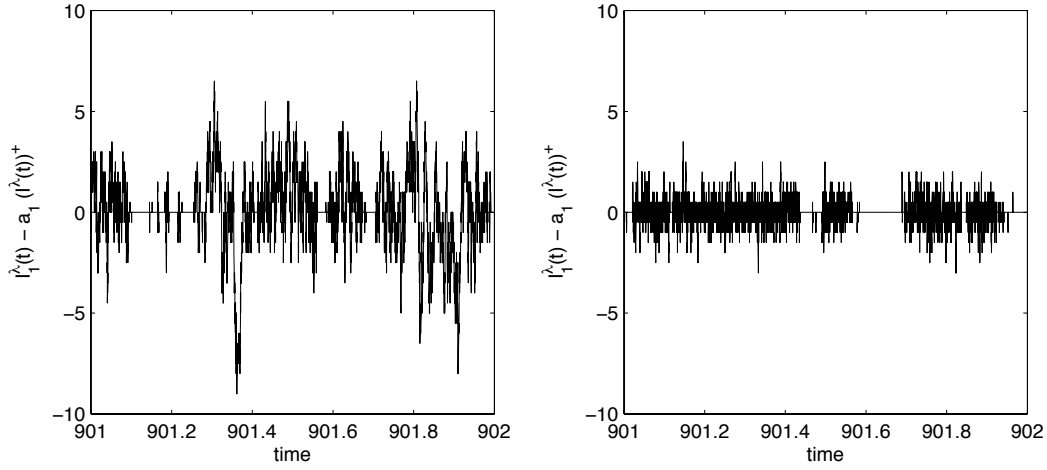
order-$(\lambda\hat{\mu})^{-1/2}$) than the former processes (order-1 time scale). In Figure 4, we plot typical sample paths of $\{I_1^\lambda - a_1(I^\lambda(t))^+,\, t \geq 0\}$ under RMI and IR routing, for the system described in Example 2 – the difference in counting and time scales is evident. Therefore, in the context of Anonymous Hospital, the desired idleness ratios are maintained not only in the long-run, but also on shorter time intervals. In particular, provided that idle servers exist ($I^\lambda(0)\sqrt{\nu^\lambda} > \varepsilon$ for some $\varepsilon > 0$), intervals of order $\lambda^{-1}$ and $(\lambda\hat{\mu})^{-1/2}$ are required under IR and RMI, respectively, for convergence of the empirical idleness ratios to the long-run averages $(1-\rho_i^\lambda)/(1-\rho_j^\lambda)$, in the sense that (for large $\lambda$)

$$\sqrt{\nu^\lambda}\left(\frac{\frac{1}{N_i^\lambda}\int_0^{t/\lambda} I_i^\lambda(u)\,du}{\frac{1}{N_j^\lambda}\int_0^{t/\lambda} I_j^\lambda(u)\,du} - \frac{1-\rho_i^\lambda}{1-\rho_j^\lambda}\right) \quad \text{and} \quad \sqrt[4]{\nu^\lambda}\left(\frac{\frac{1}{N_i^\lambda}\int_0^{t/\sqrt{\lambda\hat{\mu}}} I_i^\lambda(u)\,du}{\frac{1}{N_j^\lambda}\int_0^{t/\sqrt{\lambda\hat{\mu}}} I_j^\lambda(u)\,du} - \frac{1-\rho_i^\lambda}{1-\rho_j^\lambda}\right)$$

converge to 0, as $t$ increases ($t$ is an order-1 quantity here), for IR and RMI, respectively.

In certain cases, the time scale separation can be used to explicitly evaluate the sub-diffusion behavior of the system under IR routing, in the QED regime. As seen in the following example, the idea is to exploit the fact that the sub-diffusion process evolves on a faster time scale than diffusion processes. Recall that the sub-diffusion behavior under RMI is characterized in Theorem 2.

FIGURE 4. Typical sample paths of $\{I_1^\lambda(t) - a_1(I^\lambda(t))^+,\, t \geq 0\}$, in an inverted-V system, under the RMI (left) and IR (right) policies. Parameters of the system are the same as in Example 2.




*Example* 3 (Sub-diffusion scale under IR). Consider the inverted-V model in the QED regime, under IR routing, with $K = 2$, $w_1 = w_2 = a_1 = a_2 = 1/2$, and note that $(I_1^\lambda(t) - I^\lambda(t)/2, I_2^\lambda(t) - I^\lambda(t)/2) = ((I_1^\lambda(t) - I_2^\lambda(t))/2, (I_2^\lambda(t) - I_1^\lambda(t))/2)$. The heavy traffic averaging principle [12] states that, when considering the distribution of $(I_1^\lambda(t) - I_2^\lambda(t))$, as $\lambda \to \infty$, one can act as if the total number of (scaled) idle servers is fixed [48, p. 70]. In particular, on the event $\{I^\lambda(t)/\sqrt{\nu^\lambda} > \varepsilon\}$, $\varepsilon > 0$, the distribution of $(I_1^\lambda(t) - I_2^\lambda(t))$ converges, as $\lambda \to \infty$, to the stationary distribution of the birth-death continuous-time Markov chain, with transition rates $r_{i,i+1} = 1/2 + 1_{\{i<0\}} + (1-\chi)1_{\{i=0\}}$ and $r_{i,i-1} = 1/2 + 1_{\{i>0\}} + \chi 1_{\{i=0\}}$ (here we assume that, if the two pools have the same number of idle servers, then a customer is routed to the first pool

with probability $\chi \in [0,1]$). Indeed, if $(I_1^\lambda(t) - I_2^\lambda(t))$ is positive, then departures from pool 2 and new arrivals contribute to a decrease of this quantity; on the other hand, departures from pool 1 increase $(I_1^\lambda(t) - I_2^\lambda(t))$. This leads, for any $\varepsilon > 0$, to

$$\mathbb{P}[I_1^\lambda(t) - I_2^\lambda(t) = i \,|\, I^\lambda(t)/\sqrt{\nu^\lambda} > \varepsilon] \to \left(1 + 2(1-\chi)1_{\{i>0\}} + 2\chi 1_{\{i<0\}}\right) 3^{-|i|-1},$$

as $\lambda \to \infty$.

We conjecture that the sub-diffusion behavior of the system under the LISF algorithm is the same as the one under RMI. The conjecture is based on the following heuristic reasoning. A way to implement the LISF policy is to have servers completing service join a queue of idle servers. This queue operates in a FCFS fashion. Whenever a customer needs to be assigned to a server, it is routed to the server at the head of the queue. Observe that, under the described scheme, the server that has been idle for the longest time is assigned a customer before any other server. The state of the queue of idle servers at time $t$ is an ordered list that consists of pool labels $(1, 2, \ldots, K)$. Now, consider an inverted-V model in the QED regime ($\lambda \to \infty$), at a time instance $t$ such that $I^\lambda(t)/\sqrt{\nu^\lambda} > 0$. Then, all the servers in the idle queue joined the queue within a time interval that is on the order of $1/\sqrt{\lambda\hat{\mu}}$, i.e., $I^\lambda(t)/\sqrt{\nu^\lambda}$ remains approximately constant during this interval of time. The server pool labels in the idle queue are approximately independent, with a label being equal to $i$ with probability

$$\frac{c_i^\lambda - \mu_i I_i^\lambda(t)}{c^\lambda - \sum_{i=1}^K \mu_i I_i^\lambda(t)} \approx \frac{c_i^\lambda - a_i\mu_i I^\lambda(t)}{c^\lambda - \hat{\mu} I^\lambda(t)} = a_i + \Theta(1/\sqrt{\nu^\lambda}),$$

for large $\lambda$; the standard asymptotic notation $\Theta(1/\sqrt{\nu^\lambda})$ indicates that the second term is of the order $1/\sqrt{\nu^\lambda}$. The preceding equation is due to the fact that a given label is of type $i$ if a server in pool $i$ completes service before any other server in the other pools. As a consequence, the random variable $(I_1^\lambda(t) - a_1 I^\lambda(t))$ is of the order $\sqrt[4]{\nu^\lambda}$, due to the Central Limit Theorem and the fact that the total number of labels in the queue is $I^\lambda(t)$, a quantity proportional to $\sqrt{\nu^\lambda}$.

## 6. Concluding Remarks

We considered routing algorithms that are applicable to routing hospital patients, from the Emergency Department to Internal Wards. Given the heterogeneity of the wards, the objective is to achieve fairness from the point of view of hospital staff, while not hurting efficiency too severely. Wards are modeled as server pools, and two types of quantities are used to quantify fairness: *idleness* ratios and *flux* ratios. Under the LISF policy, which is considered to be "fair" and is commonly used in call centers, both ratios tend to the *ratios of service rates* in the respective server pools, when the system is in the QED regime; the appropriateness of the QED regime in modeling the ED-to-IW process is supported by empirical data, collected in Anonymous Hospital. In other words, LISF routing leads to a desirable outcome: faster servers work less, yet they produce more (serve more customers). However, the applicability of LISF in hospitals is limited since the algorithm requires information unavailable in hospitals on a real-time basis. The same idleness and flux ratios can be achieved by IR routing; yet, in that case, one must estimate (time-varying) ward (server pool) service capacities. We thus propose a randomized routing policy, RMI, that attains the same desired performance ratios as LISF, but requires only the number of idle servers in server pools for making decisions. The policy can be implemented in a hospital by using, for example, patient ID numbers as sources of randomness. A generalized version of RMI, WRMI, can be used to fine-tune the desired outcome.

The three algorithms (LISF, IR and RMI) share one feature in common. Namely, these algorithms achieve the desired idleness ratios (equal to service-rate ratios) by attempting to

maintain the ratios of the numbers of idle servers in different pools equal to these idleness ratios at all times. However, since idleness ratios represent an average performance measure, one can vary the instantaneous ratios of idle servers depending on the total number of idle servers and still achieve the target (average) idleness ratios. This approach has been proposed in [3]. The advantage of such an algorithm is that it delivers a lower average waiting time compared to LISF, while maintaining the same long-run idleness ratios as LISF. However, the gain in performance does not come for free. In particular, one must determine the optimal number of idle servers ratios as a function of the total number of servers and, therefore, the parameters of the system must be known (arrival rate, pool service capacities), i.e., the policy is effectively not blind.

6.1. **Partial information routing – Simulation analysis.** We mentioned above that availability of information is critical for determining an appropriate routing policy in hospitals. Our proposed routing, RMI, requires the information on the number of available beds at each ward at the moment of routing (information that is quite minimal, when compared to the other routing policies considered in the paper). However, the occupancy status in the IWs is not available on a real-time basis in Anonymous Hospital; instead, the ED relies on one bed census update per day (in the morning). Thus, in order to implement RMI routing, it is necessary to estimate the system state at decision times, based on the system state at the last update time point.

An example of such partial-information routing can be found in [46], where the authors created a computer simulation model of the ED-to-IW process in Anonymous Hospital. They used it to examine various routing policies, according to some fairness and performance criteria, while accounting for the scarcity of information in the system. Simulating the ED-to-IW process helped achieve additional practical insights, by accommodating some analytically intractable features (such as time-inhomogeneous Poisson arrivals), and allowed analysis of more complex routing algorithms. The best-performing algorithm (in terms of both staff fairness and operational performance) proposed by [46] was an algorithm which minimized at each decision point a convex combination of the two conflicting demands: balanced occupancy rates and balanced flux. The implementation under partial information resulted in almost no deterioration of performance.

To conclude, we now explain briefly the way that the lacking information is predicted. Denote by $M_j$ the number of occupied beds (busy servers) in ward (pool) $j$. This counter is updated at some time point $T$ and is estimated at other decision time points according to the patient routing and the ward service rate. Namely, we estimate the number of occupied beds in ward $j$ at time $t > T$ to be equal to $(M_j - M_j\mu_j(t-T))^+$, $j = 1, 2, 3, 4$. After a routing decision is made, we update $M_k = M_k + 1$ ($k$ denotes the ward chosen to admit the next patient).

6.2. **Routing at the Level of Individual Providers.** Instead of examining fairness via our bed-based model, one can study fairness be means of a more complex staff-based model. Such a more detailed model could potentially be used to study fairness at the level of individual care provides rather than wards. (It would thus be more relevant to the U.S. healthcare system, where typically nurses do not have fixed ward assignments and are paid on an hourly basis.) In that case, one still needs to keep track of the number of patients in wards. However, two additional aspects must be modeled explicitly: (i) patient "service" requests, i.e., when certain tasks need to be performed by nurses/doctors (for example, in Belgium, nurses work is broken down into 23 representative tasks for planning purposes [47]); and (ii) the process of assigning of these tasks to individual staff members. For example, a model can be constructed from the following components:

- Parallel-server systems with heterogeneous serves [4]: a system represents a ward, while servers represent medical staff; tasks in a ward are assigned to individual doctors/nurses according to a (ward) scheduling policy.
- Erlang-R ("R" for ReEntrant) model with a bounded number of customers [53]: the reentrant aspect of the model captures the fact that customers (patients) require service (tasks) multiple times during their sojourn times (LOS) in the system; the bounded number of customers is due to the finite number of beds in a ward.
- A (hospital) routing policy that assigns customers (patients) to one of parallel Erlang-R systems (wards) with heterogeneous servers.

Overall fairness can be achieved by enforcing fairness both at the inter- and intra-ward levels. There exist multiple time scales in this model: a task scale (minutes/tens of minutes), a "content" scale (tens of minutes/hours; time between tasks for a single patient), a shift scale (hours), and finally a length of stay (a week) scale. These scales can potentially be used to simplify the analysis of the system. Indeed, during a patient's stay in the hospital, not only many tasks are performed, but also many shifts are rotated. Hence, one expects that a patient experiences an "equivalent" service (task) rate on the LOS scale. This equivalent rate is a function of staff in the ward as well as the task assignment policy. As a consequence, when considering routing at the hospital level, one can plausibly replace the heterogeneous Erlang-R model with a parallel server system with an equivalent service rate, i.e., one obtains our inverted-V model.

6.3. **Future Research.** We propose a number of research directions motivated by our work. First, robustness of our insights against distributional assumptions on service times remains to be investigated. Second, the inverted-V model takes into account primarily the number of beds in each ward, while hospital staff (nurses and doctors) affect the model indirectly through server service rates. The model can be improved by explicitly modeling staff as well. In that case, two-scale (doctors and beds) models arise. Third, patients to be hospitalized in the IWs are classified into several categories. When arriving to the ED, patients are classified as "walking" or "lying"; in addition, prior to running the Justice Table, they are classified as "regular", "special care" or "ventilated". The load, imposed on the hospital by patients, varies significantly among different categories: in LOS, complexity of treatment, and waiting times. Thus, it is of prime interest to extend our model to accommodate multiple customer classes. Lastly, in the present work, service rates are taken to be exogenous quantities, i.e., there is no attempt to capture possible dependency between the routing algorithm and service rates of doctors and nurses. However, such dependency does exist since the hospital staff adapts to routing policies by increasing/decreasing their service rates and/or quality of care. Tools from Game Theory can be applicable in modeling such effects.

## References

[1] M. Armony. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Syst. Theory Appl.*, 51(3-4):287–329, 2005. 1.3, 3.1, 3.2, 5

[2] M. Armony and A. Mandelbaum. Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. *Oper. Res.*, to appear. 1.3

[3] M. Armony and A. Ward. Fair dynamic routing policies in large-scale systems with heterogeneous servers. *Oper. Res.*, 58(3):624–637, 2010. 1.3, 2.2, 4.1, 6

[4] R. Atar. Central limit theorem for a many-server queue with random service rates. *Ann. Appl. Probab.*, 18(4):1548–1568, 2008. 1.3, 4.1, 6.2

[5] R. Atar, Y.Y. Shaki, and A. Shwartz. A blind policy for equalizing cumulative idleness. Preprint. 1.3, 4.1

[6] B. Avi-Itzhak and H. Levy. On measuring fairness in queues. *Adv. Appl. Probab.*, 36(3):919–936, 2004. 2.2

[7] R. Bekker and A.M. de Bruin. Time-dependent analysis for refused admissions in clinical wards. *Ann. Oper. Res.*, 178(1):45–65, 2010. 1.3

[8] M. Ben-Zrihen, J. Borsher, A. Reiss, and Y. Tseytlin. Behavioral models in customer service centers. IE&M project, Technion, 2007. 2.2
[9] S. Borst, A. Mandelbaum, and M. Reiman. Dimensioning of large call centers. *Oper. Res.*, 52(1):17–34, 2004. 3.2
[10] F.B. Cabral. The slow server problem for uninformed customers. *Queueing Syst. Theory Appl.*, 50(4):353–370, 2005. 1.3, 4.1
[11] F.B. Cabral. Queues with heterogeneous servers and uninformed customers: Who works the most? arXiv:0706.0560vl, 2007. 1.3, 5, B.1
[12] E. Coffman, A. Puhalskii, and M. Reiman. Polling systems with zero switchover times: A heavy traffic averaging principle. *Ann. Appl. Probab.*, 5(3):681–719, 1995. 3
[13] J. Colquitt, D. Conlon, M. Wesson, C. Porter, and K. Y. Ng. Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *J. Appl. Psychology*, 86(3):425–445, 2001. 1.3
[14] J. Dai and T. Tezcan. State Space Collapse for Parallel Server Systems in Halfin-Whitt Regime. Preprint. 5
[15] A.M. de Bruin, R. Bekker, L. van Zanten, and G.M. Koole. Dimensioning hospital wards using the Erlang loss model. *Ann. Oper. Res.*, 178(1):23–43, 2010. 1.3
[16] F. de Véricourt and O.B. Jennings. Nurse-to-patient ratios in hospital staffing: A queuing perspective. Preprint, available at http://faculty.fuqua.duke.edu/∼fdv1/bio/ratios3.pdf. 1.3, 2.3
[17] K. Elkin and N. Rozenberg. Patients' flow from the emergency department to the internal wards. IE&M project, Technion, 2007. 2.2, 2.4, A.1
[18] A.K. Erlang. On the rational determination of the number of circuits. In E. Brockmeyer, H.L. Halstrom, and A. Jensen, editors, *The life and works of A.K. Erlang*. The Copenhagen Telephone Company, Copenhagen, 1948. 1.1, 3.2
[19] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5(2):79–141, 2003. 1.1, 3.2, 5
[20] P. González and C. Herrero. Optimal sharing of surgical costs in the presence of queues. *Math. Methods Oper. Res.*, 59(3):435–446, 2004. 1.3
[21] L.V. Green. Capacity planning and management in hospitals. In M.L. Brandeau, F. Sainfort, and W.P. Pierskalla, editors, *Operations Research and Health Care: A Handbook of Methods and Applications*, International Series in Operations Research & Management Science, pages 15–42. Kluwer, 2004. 1.3
[22] L.V. Green. Using operations research to reduce delays for healthcare. In Z.-L. Chen and S. Raghavan, editors, *Tutorials in Operations Research*, pages 1–16. Informs, 2008. 1.3, 2.3
[23] I. Gurvich and W. Whitt. Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.*, 34(2):363–396, 2009. 4.1, 4.1
[24] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.*, 29(3):567–588, 1981. 1.1, 3.2, 5
[25] Anonymous Hospital. Reducing waiting times in the emergency department and LOS in the internal wards. Report of the Quality Promotion Team, 1997. A.1
[26] R.C. Huseman, J.D. Hatfield, and E.W. Miles. A new perspective on equity theory: The equity sensitivity construct. *Academy of Management Review*, 12(2):222–234, 1987. 2.2
[27] D.S. Kc and C. Terwiesch. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498, 2009. 1.3
[28] F. Kelly. *Reversibility and Stochastic Networks*. Wiley, New York, 1979. 4.1, B.2
[29] R.L. Larsen and A.K. Agrawala. Control of a heterogeneous two-server exponential queueing system. *IEEE Trans. Software Engrg.*, 9(4):522–526, 1983. 1.3
[30] R.C. Larson. Perspectives on queues: Social justice and the psychology of queueing. *Oper. Res.*, 35(6):895–905, 1987. 2.2
[31] W. Lin and P.R. Kumar. Optimal control of a queueing system with two heterogeneous servers. *IEEE Trans. Automat. Control*, 29(8):696–703, 1984. 1.3
[32] E. Litvak, M.C. Long, A.B. Cooper, and M.L. McManus. Emergency department diversion: Causes and solutions. *Acad. Emerg. Med.*, 8(11):1108–1110, 2001. 1.3
[33] A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov. From the emergency department to hospitalization: Using theoretical models, empirical models and simulation for the operational analysis of the ED, IW, and their interface. Preprint. 3, 3.2
[34] M.L. McManus, M.C. Long, A. Cooper, and E. Litvak. Queuing theory accurately models the need for critical care resources. *Anesthesiology*, 100(5):1271–1276, 2004. 1.3
[35] O. Miro, M.T. Antonio, S. Jimenez, A. De Dios, M. Sanchez, A. Borras, and J. Milla. Decreased health care quality associated with emergency department overcrowding. *Eur. J. Emerg. Med.*, 6(2):105–107, 1999. A.3

[36] A. Rafaeli, G. Barron, and K. Haber. The effects of queue structure on attitudes. *J. Service Res.*, 5(2):125–139, 2002. 2.2

[37] M. Ramakrishnan, D. Sier, and P.G. Taylor. A two-time-scale model for hospital patient flow. *IMA J. Management Math.*, 16(3):197–215, 2005. 1.3

[38] D.B. Richardson. Increase in patient mortality at 10 days associated with emergency department overcrowding. *Med. J. Australia*, 184(5):213216, 2006. A.3

[39] M. Rubinovitch. The slow server problem. *J. Appl. Probab.*, 22(1):205–213, 1985. 1.3, 4.1

[40] M. Rubinovitch. The slow server problem: A queue with stalling. *J. Appl. Probab.*, 22(4):309–317, 1985. 1.3

[41] P.C. Sprivulis, J. Da Silva, I.G. Jacobs, A.R.L. Frazer, and G.A. Jelinek. The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *Med. J. Australia*, 184(5):208–212, 2006. A.3

[42] R.H. Stockbridge. A martingale approach to the slow server problem. *J. Appl. Probab.*, 28(2):480–486, 1991. 1.3

[43] New York Times. 1 in 3 hospitals say they divert ambulances, April 9, 2002. Available at http://www.nytimes.com/2002/04/09/us/1-in-3-hospitals-say-they-divert-ambulances.html. 1.3

[44] S. Trzeciak and E.P. Rivers. Emergency department overcrowding in the United States: An emerging threat to patient safety and public health. *Emerg. Med. J.*, 20:402–405, 2003. A.3

[45] Y. Tseytlin. Queueing systems with heterogeneous servers: On fair routing of patients in emergency departments. M.Sc. thesis, IE&M, Technion, 2009. Available at http://iew3.technion.ac.il/serveng/References/thesis-yulia.pdf. 1.3, 2.2, 4.1, 1, B.1

[46] Y. Tseytlin and A. Zviran. Simulation of patients routing from an emergency department to internal wards in Anonymous Hospital. IE&M project, Technion, 2009. Available at http://iew3.technion.ac.il/serveng/References/project_yulia_asaf.pdf. 6.1

[47] R. Williams. It all adds up. *Nurs. Stand.*, 14(31):12-13, 2000. 6.2

[48] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York, 2002. 3

[49] J. Wright and R. King. *We All Fall Down: Goldratt's Theory of Constraints for Healthcare Systems*. North River Press, 2006. 1.3

[50] ynet. What hospital is the most crowded in Israel?, 2009. Available at http://www.ynet.co.il/articles/0,7340,L-3656123,00.html. A.4, 3

[51] G. Yom-Tov. Queues in hospitals: Semi-open queueing networks in the QED regime. Technical report, IE&M, Technion, 2007. 2.3

[52] G. Yom-Tov. *Queues in Hospitals: Queueing Networks with ReEntering Customers in the QED Regime (QED = Quality- and Efficiency-Driven)*. PhD thesis, Technion, Haifa, Israel, 2010. 3.2

[53] G. Yom-Tov and A. Mandelbaum. Erlang-R: A time-varying queue with ReEntrant customers, in support of healthcare staffing. Preprint. 6.2

[54] B. Zhang, L.S.H. van Leeuwaarden, and B. Zwart. Staffing call centers with impatient customers: Refinements to many-server asymptotics. *Oper. Res.*, to appear. 3.2

## A. Additional Information on Patient Routing

A.1. **The "Justice Table" Algorithm.** *History:* Prior to 1997, patient allocation was decided according to a fixed "table of duty" of the wards (every day another ward was on duty and had to accommodate all incoming patients), but allocation was subject to wards' approval – each ward had the authority to refuse admitting a patient. Consequently, waiting times in the ED until transfer to the IWs were extremely long – 10.5 hours on average, with 12% of the patients forced to wait more than 24 hours (!) [25]. This was unbearable to patients, and caused a heavy overload on the ED that resulted in malfunctioning. In 1995, as part of a hospital quality program, a dedicated team was charged with the task of improving processes in the ED – its goal was, in particular, to reduce ED-to-IW delays. The team [25] proposed a change in the existing routing policy: a patient's placement would be determined by an algorithm, entitled the "Justice Table", and the authority for patient routing would be taken away from the wards. Implementation of this change-of-authority was not easy, for political reasons, but eventually prevailed.

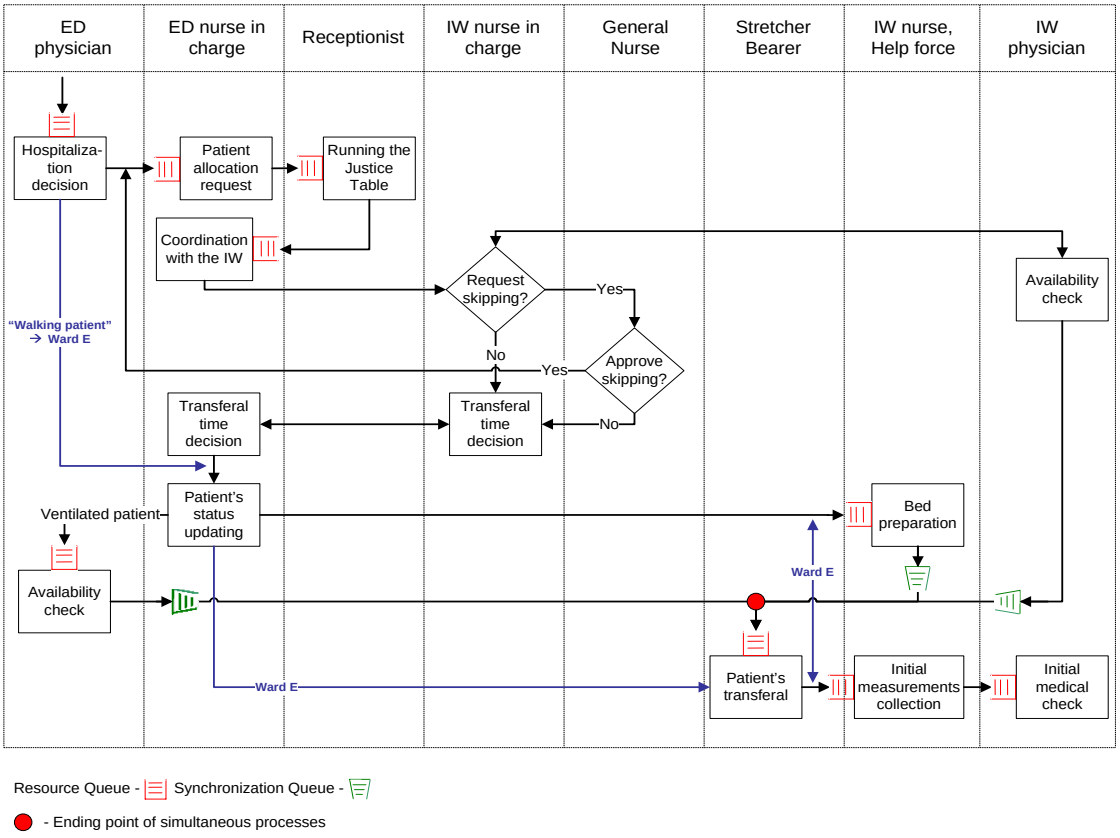
*Short description of the algorithm:* The purpose of the "Justice Table" was to balance load among the wards. It was decided to classify patients into three categories: *ventilated* – patients that required artificial respiration, *special care* – patients whose rate on the Norton scale (a table used to predict if a patient might develop a pressure ulceration) was below 14, and *regular* – all other patients. The algorithm treated each category independently in order to ensure a fair allocation, since Lengths of Stay (LOS) and complexities of treatment varied significantly among these patient categories. For each category, there was a *cyclical order* among the wards, i.e., each ward received one patient in its turn (round-robin). In addition to a patient's classification, the algorithm took into account the size of each ward, by allocating fewer patients to smaller wards. However, the algorithm took into account neither the actual number of occupied beds at the time the routing decision was made, nor the discharge rate at the wards.

*The current state:* The results of implementing the Justice Table routing policy were very impressive – the average waiting time from decision on hospitalization until moving the patient to a ward was reduced to 66 minutes (from over ten hours) [25]. In addition, significant improvements in other ED processes were measured as well (due to the overload reduction), along with a higher ward efficiency (more admissions, shorter LOS). In 2004, the use of the Justice Table was discontinued due to software changes in the hospital. In 2006, adapted to the new software, the Justice Table was reinstated with minor changes, but its influence grew smaller, as the medical staff had become used to making placement decisions without it. Moreover, as reported in [17], a significant number of the patients transferred from the ED to the IWs were, in fact, not routed via the Justice Table. Moreover, in 2007, the ED of Anonymous Hospital moved to a temporary location, and only recently has it returned to a renovated home. The present research is now helping the hospital in rejuvenating its Justice Table, towards an efficient and fair ED-to-IW process.

A.2. **Patient Routing.** One can appreciate the complexity of the ED-to-IW process through the Integrated (Activities - Resources) Flow Chart (Figure 5). We provide here a short description. A patient, whom a physician in charge of the ED decides to hospitalize in the IWs, is assigned to one of the five wards in the following manner: if this is an "independent" walking patient, usually s/he is assigned to Ward E – in this case usually s/he is transferred to the ward almost without any delay. Otherwise, a receptionist of the ED runs the Justice Table. S/he transfers the output (one of the Wards A-D) to the nurse in charge of the ED, who starts a negotiation process with the chosen ward. If the ward refuses to admit the patient (usually for reasons of overloading), the two sides appeal to a General Nurse, who is authorized to approve a so-called "skipping" – allowing a ward to skip its turn. If skipping is granted, the

receptionist runs the Justice Table again, and the process repeats itself until some ward agrees (or is forced) to admit the patient. The next stage of the negotiations is agreeing upon the

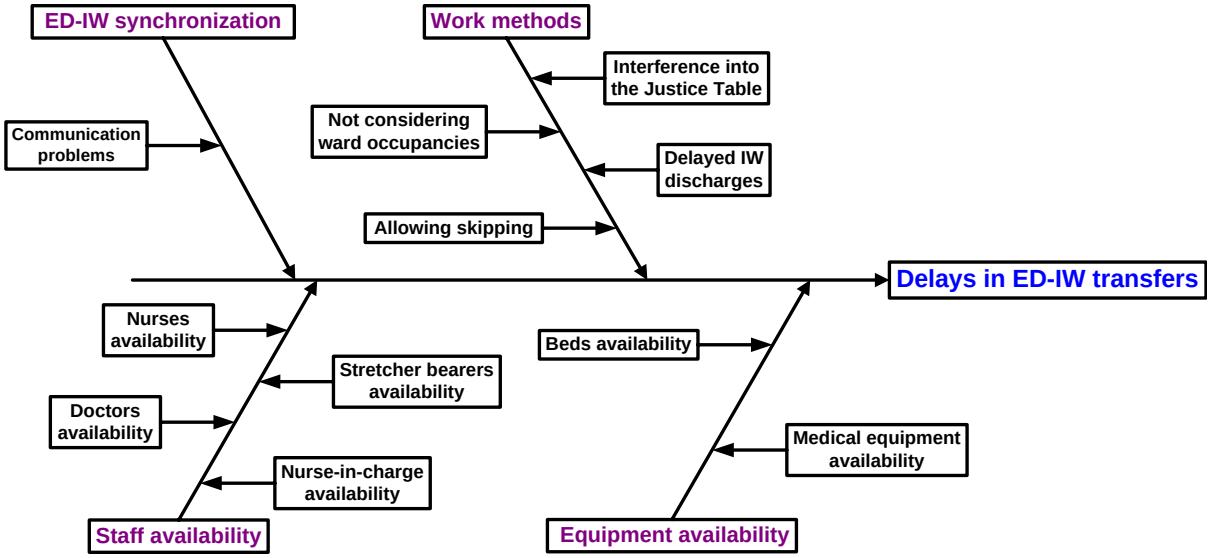
FIGURE 5. Integrated (activities - resources) flow chart.




time at which the patient will be transferred to her/his ward. Here interests are conflicting: the ED seeks to discharge the patient as soon as possible, in order to be able to accept new patients, and the IWs prefer to have the move carried out at a time convenient for them. From conversations with nurses from both sides we learned that, when deciding on a patient's transfer time, the main issue taken into account (assuming there is an available bed in the ward) is nurses' and doctors' availability (they might be unavailable because of treating other patients, shifts changing or meals, various staff meetings or resuscitation). Another parameter is the availability of necessary equipment and other logistic considerations. Patients to-be hospitalized wait in the ED till a transfer to their ward is carried out – sometimes these waiting times are extremely long. Through the Cause-and-Effect chart (fish-bone diagram) in Figure 6 one observes the various causes of these long delays. We emphasize that the delays are caused not only by beds unavailability: patients usually wait even when there are available beds (see the remark in Section 3.2).

A.3. **Long Delays.** Patients often wait long times in the ED until they are transferred to their IWs – the reasons for these long delays are summarized in the Cause-and-Effect chart (fish-bone diagram) in Figure 6. The main reasons are beds availability and staff availability (doctors

FIGURE 6. ED-to-IW Delays: Cause-and-Effect Chart.

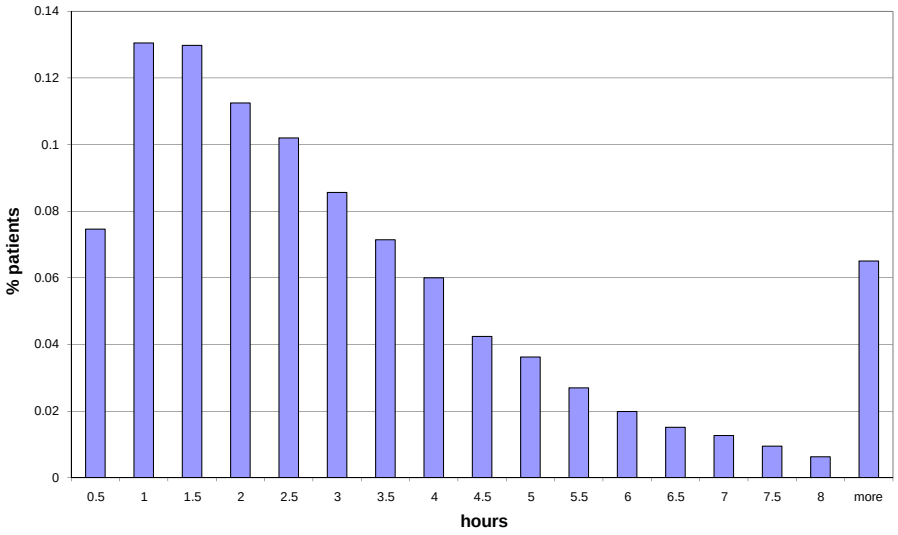



and nurses might be unavailable because of treating other patients, shifts changing, meals, various staff meetings or resuscitation). An additional reason is the unavailability of necessary equipment and other logistic considerations: for example, preparation for a "complicated" patient who requires special bed/equipment, or placement near a nursing station due to case severity, takes a longer time. Exact data on those waiting times are not kept in the hospital information systems. We thus estimated these delays by analyzing the time from a decision to hospitalize at a certain ward until receiving the first treatment in that ward (these data were acquired from the hospital database). The average delay in 2006-2008 was 3.1 hours (for Wards A-D); for 23% of the patients this time was longer than 4 hours, and for 6.5% longer than 8 hours. The waiting times histogram is shown in Figure 7.

Long waiting times cause an overload on the ED, as beds remain occupied while new patients continue to arrive. As mentioned, they cause ED blocking, which leads to ambulance diversion. They cause significant discomfort to the waiting patients as well: in the ED they suffer from noise, and lack of privacy and proper meals. In addition, patients do not enjoy the best professional medical treatment and dedicated attention as in the wards; hence, the longer patients wait in the ED, the lower their satisfaction and the higher the likelihood for clinical deterioration. Improving the efficiency of patients flow from the ED to the IWs, while shortening waiting times in the ED, will improve the service and care provided to patients. Indeed, reducing the load on the ED will lead to a better response to arriving patients, which has been shown to save lives [38, 35, 41, 44].

A.4. **Other Hospitals.** The hospitals we study differ in their functionality and geographical location. From Table 3 it is evident that they differ in *size* (number of IWs and beds in them; number of treated patients), in the *load* IWs are subjected to (average number of transfers to IWs per bed; average occupancy rate – as reported by [50]), and by their *efficiency* (ALOS in the ED and in the IWs). Despite these differences, we recognize the same problems in the ED-to-IW process at all hospitals. First of all, none of the hospitals (except for Hospital 5) measures waiting times of patients to-be hospitalized in the IWs – we were given just a rough

FIGURE 7. Waiting times histogram.




* Data refer to period May 1, 2006 - October 30, 2008 (excluding the months 1-3/2007, when Ward B was in charge of an additional sub-ward).

estimate. Those estimated waiting times are long (again, except for Hospital 5) – indeed, the situation in Anonymous Hospital is not the worst.

The routing policies are intuitive and simple – "cyclical order" policies prevail. Although in four out of the six hospitals the wards are heterogeneous (in terms of capacity and ALOS), this plays no role in routing: no hospital accounts for differences in ALOS, and only Anonymous Hospital accounts for ward capacities. In addition, none of the hospitals, besides our hospital and Hospital 5, allocates patients of different categories separately. Surely this cannot be fair, as load inflicted on the ward staff by patients of different categories varies significantly.

Hospital 2 has an original policy – routing is performed according to the one-before-last digit of a patient's identity number: if it is odd then the patient is assigned to Ward A, if even – to Ward B. This is equivalent to random assignment: each ward is chosen with probability 1/2. However, ward capacities differ: the size of one ward is 2/3 of the other's, hence this method can not be fair. Hospital 4 has an even "simpler" policy: a patient is assigned to a ward that has a vacant bed (we were told that the wards were *always* full). Indeed, the load on the IWs in this hospital (average occupancy and flux) is very high. Due to such a policy, the wards decide when to admit their patients, and they do not have any incentive to become more efficient and discharge patients faster – waiting times, ALOS in the ED and in the IWs are the longest in this hospital. Hospital 5 presents an exception – patients to be hospitalized in the IWs are transferred from the ED almost without any delay, there are separate cycles for different patient categories, and even the policy of cyclical routing appears to be fair as all the wards are the same (in terms of capacities and LOS). However, from a conversation with the ED receptionist during our visit to this hospital, we learned that the routing process was managed by her manually, and that she received frequent complaints from the wards' staff on unfairness of the allocations.

*Remark.* The discussion presented here is necessarily superficial, as it is based solely on questionnaires and interviews (if there were such) – no observations or data collection were performed at those hospitals. We are not familiar with any documentation on how the ED-to-IW process is managed in hospitals outside of Israel.

TABLE 3. Hospitals comparison.

| | Hospital 1 | Hospital 2 | Hospital 3 | Hospital 4 | Hospital 5 | Anon. H. |
|---|---|---|---|---|---|---|
| Number of IWs | 9 | 2 | 3 | 4 | 6 | 5 |
| IW # beds | 327 | 45 | 108 | 93 | 210 | 185 |
| Average weekly # of arrivals to Internal ED | 1050 | 350 | 637 | 630 | 1050 | 1050 |
| Average weekly # of transfers from ED to IWs | 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 | 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*** |

* Based on internet article [50].

** Differ in their capacities and LOS.

*** Accounting for different patient types and ward capacities.

## B. TECHNICAL APPENDIX

B.1. **Proof of Theorem 1.** The following lemma serves as a basis of our analysis. It provides an explicit characterization of the stationary probabilities for the inverted-V model. This characterization is used in Lemma 2 to obtain relative likelihoods that particular servers are busy. Finally, the inequalities from Lemma 2 are used to prove the theorem.

**Lemma 1** (Tseytlin [45])**.** *Consider the inverted-V model with $\lambda < c^\lambda$ under the RMI policy. The process $\{(I^\lambda(t), I_1^\lambda(t)), \dots, I_K^\lambda(t)), t \geq 0\}$ is a reversible continuous-time Markov chain with the stationary distribution $\pi^\lambda$:*

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

*where*

$$\pi^\lambda(0) \equiv \pi^\lambda(0, 0, \dots, 0) = \left( \frac{\rho^\lambda}{1-\rho^\lambda} + \sum_{i_1=0}^{N_1^\lambda} \cdots \sum_{i_K=0}^{N_K^\lambda} (i_1 + \cdots + i_K)! \prod_{j=1}^K \binom{N_j^\lambda}{i_j} \left(\frac{\mu_j}{\lambda}\right)^{i_j} \right)^{-1}. \quad (7)$$

For notational convenience, we prove the theorem for the single-server-pool model under the Random Assignment policy (note that Cabral [11] proved this result independently); this model is equivalent to the inverted-V model under the RMI policy by Theorem 4.3.1 in [45].

Thus, for the rest of the proof, we consider the case $K = N^\lambda$ and $N_1^\lambda = N_2^\lambda = \cdots = N_K^\lambda = 1$. For a set $\mathcal{X} \subset \{1, \ldots, N^\lambda\}$, let $\pi_{\mathcal{X}}^\lambda := \pi^\lambda(i, i_1, \ldots, i_K)$, where $i = \sum_{j=1}^K i_j$ and $i_j = 1_{\{j \notin \mathcal{X}\}}$, i.e., $\pi_{\mathcal{X}}^\lambda$ is the stationary probability that servers in $\mathcal{X}$ are busy while all other servers are idle; when $\mathcal{X} = \{1, \ldots, N^\lambda\}$, we let $\pi_{\mathcal{X}}^\lambda = \sum_{j=0}^\infty \pi^\lambda(-j, 0, \ldots, 0)$. Furthermore, define $p_m^i$ as the stationary probability that exactly $m \in \{1, \ldots, N^\lambda\}$ servers (out of $N^\lambda$) are busy, including the server with index $i \in \{1, \ldots, N^\lambda\}$, e.g., $p_1^i = \pi_i^\lambda$, $p_2^i = \sum_{j \neq i} \pi_{\{i,j\}}^\lambda$, etc. Note that $p_{N^\lambda}^i$ (for all $i$) is the probability that all servers are busy ($\pi_{\{1,\ldots,N^\lambda\}}^\lambda$).

**Lemma 2.** *If $\mu_j > \mu_k$, then $p_m^j < p_m^k$ and $\mu_j p_m^j > \mu_k p_m^k$, for any $m \in \{1, 2, \ldots, N-1\}$.*

*Proof.* The definition of $\pi_{\mathcal{X}}^\lambda$ and Lemma 1 yield

$$\pi_\emptyset^\lambda = \pi^\lambda(N^\lambda, 1, 1, \ldots, 1) = \pi^\lambda(0)\, N^\lambda! \prod_{j=1}^{N^\lambda} \frac{\mu_j}{\lambda}$$

and

$$\begin{aligned}\pi_{\mathcal{X}}^\lambda = \pi^\lambda(N^\lambda - |\mathcal{X}|, i_1, i_2, \ldots, i_{N^\lambda}) &= \pi^\lambda(0)\, (N^\lambda - |\mathcal{X}|)! \prod_{j \notin \mathcal{X}} \frac{\mu_j}{\lambda} \\ &= \pi_\emptyset^\lambda \frac{(N^\lambda - |\mathcal{X}|)!}{N^\lambda!} \frac{\lambda^{|\mathcal{X}|}}{\prod_{j \in \mathcal{X}} \mu_j},\end{aligned}$$

where $\mathcal{X} \subset \{1, \ldots, N^\lambda\}$ and $i_j = 1_{\{j \notin \mathcal{X}\}}$. From the preceding equality we obtain that, for every possible subset $\mathcal{X}$ of $\{1, 2, \ldots, N^\lambda\} \setminus \{j, k\}$:

$$\pi_{j \cup \mathcal{X}}^\lambda = \pi_\emptyset^\lambda \frac{(N^\lambda - (|\mathcal{X}| + 1))!}{N^\lambda!} \frac{\lambda^{|\mathcal{X}|+1}}{\mu_j \prod_{i \in \mathcal{X}} \mu_i} < \pi_\emptyset^\lambda \frac{(N^\lambda - (|\mathcal{X}| + 1))!}{N^\lambda!} \frac{\lambda^{|\mathcal{X}|+1}}{\mu_k \prod_{i \in \mathcal{X}} \mu_i} = \pi_{k \cup \mathcal{X}}^\lambda, \tag{8}$$

where the inequality follows from $\mu_j > \mu_k$. Next we note that:

$$p_m^j = \sum_{\substack{\mathcal{X}: |\mathcal{X}| = m-1, \\ j,k \notin \mathcal{X}}} \pi_{j \cup \mathcal{X}}^\lambda + \sum_{\substack{\mathcal{X}: |\mathcal{X}| = m-2, \\ j,k \notin \mathcal{X}}} \pi_{\{j,k\} \cup \mathcal{X}}^\lambda, \tag{9}$$

$$p_m^k = \sum_{\substack{\mathcal{X}: |\mathcal{X}| = m-1, \\ j,k \notin \mathcal{X}}} \pi_{k \cup \mathcal{X}}^\lambda + \sum_{\substack{\mathcal{X}: |\mathcal{X}| = m-2, \\ j,k \notin \mathcal{X}}} \pi_{\{j,k\} \cup \mathcal{X}}^\lambda. \tag{10}$$

The last sum is equal in both expressions, and (8) implies that the first sum in (9) is strictly smaller than the corresponding sum in (10). Hence, the first statement of the lemma: $p_m^j < p_m^k$ for $m \in \{1, 2, \ldots, N^\lambda - 1\}$.

Next, we consider the second statement of the lemma. Multiplying each side of (9) by $\mu_j$ and each side of (10) by $\mu_k$, yields, for $\mathcal{X} \subset \{1, \ldots, N^\lambda\} \setminus \{j, k\}$:

$$\mu_j\, p_m^j = \sum_{\mathcal{X}: |\mathcal{X}| = m-1} \mu_j\, \pi_{j \cup \mathcal{X}}^\lambda + \sum_{\mathcal{X}: |\mathcal{X}| = m-2} \mu_j\, \pi_{\{j,k\} \cup \mathcal{X}}^\lambda, \tag{11}$$

$$\mu_k\, p_m^k = \sum_{\mathcal{X}: |\mathcal{X}| = m-1} \mu_k\, \pi_{k \cup \mathcal{X}}^\lambda + \sum_{\mathcal{X}: |\mathcal{X}| = m-2} \mu_k\, \pi_{\{j,k\} \cup \mathcal{X}}^\lambda, \tag{12}$$

where

$$\mu_j\, \pi_{j \cup \mathcal{X}}^\lambda = \pi_\emptyset^\lambda \frac{(N - (|\mathcal{X}| + 1))!}{N!} \frac{\lambda^{|\mathcal{X}|+1}}{\prod_{i \in \mathcal{X}} \mu_i} = \mu_k\, \pi_{k \cup \mathcal{X}}^\lambda, \tag{13}$$

and

$$\mu_j\,\pi^{\lambda}_{\{j,k\}\cup\mathcal{X}} = \pi_\emptyset \frac{(N^\lambda - (|\mathcal{X}|+2))!}{N^\lambda!}\frac{\lambda^{|\mathcal{X}|+2}}{\mu_k\prod_{i\in\mathcal{X}}\mu_i} > \pi_\emptyset \frac{(N^\lambda - (|\mathcal{X}|+2))!}{N^\lambda!}\frac{\lambda^{|\mathcal{X}|+2}}{\mu_j\prod_{i\in\mathcal{X}}\mu_i} = \mu_k\,\pi^{\lambda}_{\{j,k\}\cup\mathcal{X}}, \tag{14}$$

where the inequality follows from $\mu_j > \mu_k$. Equations (11), (12), (13) and (14) imply $\mu_j\,p^j_m > \mu_k\,p^k_m$ for $m \in \{1,2,\dots,N-1\}$. □

Next, we complete the proof of Theorem 1.

*Proof of Theorem 1.* Consider two servers $j$ and $k$ such that $\mu_j > \mu_k$. Since server's utilization is equal to the steady-state probability that it is busy, $\rho^\lambda_i = \sum_{m=1}^{N^\lambda} p^i_m$ $(i = j,k)$; in the same manner, $\gamma^\lambda_i = \mu_i\rho^\lambda_i = \sum_{m=1}^{N^\lambda}\mu_i p^i_m$ $(i=j,k)$. Then, Lemma 2 implies

$$\rho^\lambda_j = \sum_{m=1}^{N} p^j_m < \sum_{m=1}^{N} p^k_m = \rho^\lambda_k$$

and

$$\gamma^\lambda_j = \mu_j\rho^\lambda_j = \sum_{m=1}^{N^\lambda}\mu_j p^j_m > \sum_{m=1}^{N^\lambda}\mu_k p^k_m = \mu_k\rho^\lambda_k = \gamma^\lambda_k.$$ □

B.2. **Proof of Theorem 2.** The proof of the theorem is based on the reversibility [28] of the process $\{(I^\lambda(t), I^\lambda_1(t)),\dots,I^\lambda_K(t)),\ t\ge 0\}$. First we provide some preliminary results.

**Lemma 3.** *Let $D$ be a $(K-1)\times(K-1)$ matrix with elements $D_{i,j} = a_i^{-1}1_{\{i=j\}} + a_K^{-1}$, where $a_i > 0$ $(i=1,\dots,K)$ and $\sum_{i=1}^K a_i = 1$. Then, $D^{-1}$ is a matrix with elements $(D^{-1})_{i,j} = a_i 1_{\{i=j\}} - a_i a_j$ and*

$$\det(D^{-1}) = (\det(D))^{-1} = \prod_{i=1}^{K} a_i. \tag{15}$$

*Proof.* First, it is straightforward to verify that $DD^{-1}$ is an identity matrix:

$$\begin{aligned}(DD^{-1})_{i,j} &= \sum_{n=1}^{K-1}(a_i^{-1}1_{\{i=n\}} + a_K^{-1})(a_n 1_{\{n=j\}} - a_n a_j)\\ &= 1_{\{i=j\}} - a_j + a_j a_K^{-1} - a_j a_K^{-1}(1-a_K) = 1_{\{i=j\}}.\end{aligned}$$

Second, $D^{-1}$ can be factorized: $D^{-1} = CB$, where $C_{i,j} = a_i 1_{\{i=j\}}$ and $B_{i,j} = 1_{\{i=j\}} - a_j$, and, hence,

$$\det(D^{-1}) = \det(B)\det(C) = \det(B)\prod_{i=1}^{K-1} a_i. \tag{16}$$

However, if $A$ is a matrix obtained by subtracting the first row of $B$ from all other rows, then $A_{i,j} = 1_{\{i=j\}} - a_j 1_{\{i=1\}} - 1_{\{j=1,\,i\neq 1\}}$ and

$$\det(B) = \det(A) = 1 - \sum_{i=1}^{K-1} a_i = a_K. \tag{17}$$

The statement (15) follows from (16) and (17). □

Given that the system is reversible (Lemma 1), there exists a straightforward relation between the stationary distributions of the delay model and the corresponding loss model. The following key proposition provides a characterization of the loss system.

**Proposition 1.** *Consider the inverted-V loss model under the RMI policy in the QED regime. Then* $(\hat{I}^\lambda, (\hat{I}_1^\lambda, \ldots, \hat{I}_K^\lambda)1_{\{\hat{I}^\lambda>0\}}) \Rightarrow (\hat{I}, (\hat{I}_1, \ldots, \hat{I}_K)1_{\{\hat{I}>0\}})$, *as* $\lambda \to \infty$, *where* $\hat{I}$ *and* $(\hat{I}_1, \ldots, \hat{I}_K)$ *are independent,* $\mathbb{P}[\hat{I} > x] = \Phi(\delta - x)/\Phi(\delta)$, $x \geq 0$, *and* $(\hat{I}_1, \ldots, \hat{I}_K)$ *is zero-mean multi-variate normal, with* $\mathbb{E}\hat{I}_i\hat{I}_j = a_i 1_{\{i=j\}} - a_i a_j$.

*Proof.* Let $\tilde{\pi}^\lambda(m)$ be the stationary probability of having $m_i$ idle servers in pool $i = 1, \ldots, K$ in the loss model. Here, $m = (m_1, \ldots, m_K)$; denote $m_\Sigma = \sum_{i=1}^K m_i$. The proof of the proposition proceeds along the following lines. First, we define states of the system that have non-negligible probabilities (see (19)) and establish the probabilities of those states relative to the probability that all servers are busy ($\tilde{\pi}^\lambda(0)$, see (21)). Second, the limiting probability of having exactly $\lfloor \psi\sqrt{\nu^\lambda} \rfloor$ idle servers is determined in terms of $\tilde{\pi}^\lambda(0)$ (see (22)). From this, the asymptotic probability of having no idle servers can be derived (see (24)). Finally, once the limiting value of $\tilde{\pi}^\lambda(0)$ is found, the rest of the proof follows.

Due to the fact that the system is time-reversible, the stationary distribution obeys (see Lemma 1)

$$\tilde{\pi}^\lambda(m) = m_\Sigma!\, \tilde{\pi}^\lambda(0) \prod_{i=1}^K \left(\frac{\mu_i}{\lambda}\right)^{m_i} \binom{N_i^\lambda}{m_i}, \tag{18}$$

where $m_i = 0, 1, \ldots, N_i^\lambda$, and $\tilde{\pi}^\lambda(0)$ is the probability that all servers are busy. Next, we introduce a vector $\dot{m}$ with elements

$$\dot{m}_i = \frac{c_i^\lambda}{c^\lambda}\sqrt{\psi^2\nu^\lambda} + \xi_i \sqrt[4]{\psi^2\nu^\lambda}, \tag{19}$$

where $\psi \geq 0$ and $\sum_{i=1}^K \xi_i = 0$; then $\dot{m}_\Sigma = \sum_{i=1}^K \dot{m}_i = \psi\sqrt{\nu^\lambda}$ and

$$\frac{c_i^\lambda \dot{m}_\Sigma}{c^\lambda \dot{m}_i} = \frac{\sqrt{\nu^\lambda}}{\sqrt{\nu^\lambda} + \xi_i \sqrt[4]{\nu^\lambda}\, c^\lambda/(\sqrt{\psi}\, c_i^\lambda)}. \tag{20}$$

Assuming that $\tilde{\pi}^\lambda(m)$ denotes $\tilde{\pi}^\lambda(\lfloor m_1 \rfloor, \ldots, \lfloor m_K \rfloor)$ whenever elements of $m$ are not integers, (18) and Stirling's approximation yield, as $\lambda \to \infty$,

$$\frac{(\psi^2\nu^\lambda)^{\frac{K-1}{4}}\tilde{\pi}^\lambda(\dot{m})}{\tilde{\pi}^\lambda(0)} = \frac{1+o(1)}{\sqrt{2\pi}^{K-1}} \sqrt{\frac{(\nu^\lambda)^{\frac{K-1}{2}}\dot{m}_\Sigma}{\prod_{i=1}^K \dot{m}_i}} \prod_{i=1}^K e^{-\dot{m}_i} \left(\frac{N_i^\lambda}{N_i^\lambda - \dot{m}_i}\right)^{N_i^\lambda + \frac{1}{2}} \left(\frac{c^\lambda}{\lambda}\right)^{\dot{m}_i} \left[\frac{c_i^\lambda \dot{m}_\Sigma}{c^\lambda \dot{m}_i}\left(1 - \frac{\dot{m}_i}{N_i^\lambda}\right)\right]^{\dot{m}_i}.$$

Now, by using (19), (20) and the Taylor expansion of the logarithmic function, we obtain limits (as $\lambda \to \infty$) for all terms on the right-hand side of the preceding equality:

$$\begin{aligned}
\sqrt{\frac{(\psi^2\nu^\lambda)^{\frac{K-1}{2}}\dot{m}_\Sigma}{\prod_{i=1}^K \dot{m}_i}} &\to \frac{1}{\sqrt{\prod_{i=1}^K a_i}}, \\
\prod_{i=1}^K e^{-\dot{m}_i}\left(\frac{N_i^\lambda}{N_i^\lambda - \dot{m}_i}\right)^{N_i^\lambda} &\to e^{\psi^2/2}, \\
\left(c^\lambda/\lambda\right)^{\dot{m}_\Sigma} &\to e^{\psi\delta}, \\
\prod_{i=1}^K \left[\frac{c_i^\lambda}{c^\lambda}\frac{\dot{m}_\Sigma}{\dot{m}_i}\left(1 - \frac{\dot{m}_i}{N_i^\lambda}\right)\right]^{\dot{m}_i} &\to e^{-\psi^2 - \frac{1}{2}\sum_{i=1}^K \xi_i^2/a_i}.
\end{aligned}$$

Therefore, we have, as $\lambda \to \infty$,

$$\begin{aligned}\frac{(\psi^2\nu^\lambda)^{\frac{K-1}{4}}\tilde{\pi}^\lambda(\dot{m})}{\tilde{\pi}^\lambda(0)} &\to \frac{1}{\sqrt{(2\pi)^{K-1}\prod_{i=1}^K a_i}}\, e^{\psi\delta-\frac{1}{2}\psi^2-\frac{1}{2}\sum_{i=1}^K \xi_i^2/a_i} \\ &= \frac{1}{\sqrt{(2\pi)^{K-1}\det(D)}} e^{-\frac{1}{2}(\xi_1^{K-1})' D^{-1}\xi_1^{K-1}} e^{-\frac{1}{2}(\psi-\delta)^2} e^{\frac{1}{2}\delta^2} \\ &= \varphi_D(\xi_1^{K-1})\,\frac{\varphi(\psi-\delta)}{\varphi(\delta)}, \end{aligned} \tag{21}$$

where $\varphi_D$ is the $(K-1)$-dimensional zero-mean multi-variate normal density function, defined by the covariance matrix $D$, $\xi_1^{K-1} = (\xi_1, \ldots, \xi_{K-1})'$, $D^{-1}$ is a $(K-1)\times(K-1)$ matrix with elements $(D^{-1})_{i,j} = a_i^{-1}1_{\{i=j\}} + a_K^{-1}$ (this stems from $\sum_{i=1}^K \xi_i = 0$, i.e., $\xi_K^2 = (\sum_{i=1}^{K-1}\xi_i)^2$); also, due to Lemma 3, we have $D_{i,j} = a_i 1_{\{i=j\}} - a_i a_j$ and

$$\det(D) = \prod_{i=1}^K a_i.$$

Next, for $\psi > 0$, (21) and the fact that the number of summands in the following sum is polynomial in $\nu^\lambda$ (bounded from above by $(\psi\sqrt{\nu^\lambda})^K$) result in, as $\lambda \to \infty$,

$$\begin{aligned}\sum_{m:\, m_\Sigma = \lfloor \psi\sqrt{\nu^\lambda}\rfloor} \frac{\tilde{\pi}^\lambda(m)}{\tilde{\pi}^\lambda(0)} &\to \frac{\varphi(\psi-\delta)}{\varphi(\delta)} \int_{-\infty}^{\infty}\cdots\int \varphi_D(\xi_1^{K-1})\, d\xi_1\cdots d\xi_{K-1} \\ &= \frac{\varphi(\psi-\delta)}{\varphi(\delta)}, \end{aligned} \tag{22}$$

since the integrand is a valid density function; note that the limit holds trivially for $\psi = 0$, i.e., $m_\Sigma = 0$. The preceding limit implies, as $\lambda \to \infty$,

$$\begin{aligned}\sum_{m_1=0}^{N_1^\lambda}\cdots\sum_{m_K=0}^{N_K^\lambda} \frac{\tilde{\pi}^\lambda(m)}{\sqrt{\nu^\lambda}\tilde{\pi}^\lambda(0)} &= \sum_{i=1}^{N^\lambda}\sum_{m:\, m_\Sigma = i} \frac{\tilde{\pi}^\lambda(m)}{\sqrt{\nu^\lambda}\tilde{\pi}^\lambda(0)} \\ &\to \frac{1}{\varphi(\delta)}\int_0^\infty \varphi(\psi-\delta)\, d\psi \\ &= \frac{\Phi(\delta)}{\varphi(\delta)}, \end{aligned} \tag{23}$$

and, hence, as $\lambda \to \infty$,

$$\sqrt{\nu^\lambda}\tilde{\pi}^\lambda(0) \to \frac{\varphi(\delta)}{\Phi(\delta)}. \tag{24}$$

Moreover, (22) and (23) result in, as $\lambda \to \infty$,

$$\sqrt{\nu^\lambda}\sum_{m:\, m_\Sigma = \lfloor \psi\sqrt{\nu^\lambda}\rfloor} \tilde{\pi}^\lambda(m) \to \frac{\varphi(\psi-\delta)}{\Phi(\delta)},$$

and, therefore,

$$\mathbb{P}[\hat{I}^\lambda \le x] = \sum_{j=1}^{\lfloor x\sqrt{\nu^\lambda}\rfloor} \sum_{m:\, m_\Sigma = j} \tilde{\pi}^\lambda(m)$$
$$= \frac{1}{\sqrt{\nu^\lambda}} \sum_{j=1}^{\lfloor x\sqrt{\nu^\lambda}\rfloor} \left( \sqrt{\nu^\lambda} \sum_{m:\, m_\Sigma = j} \tilde{\pi}^\lambda(m) \right) \to \frac{1}{\Phi(\delta)} \int_0^x \varphi(\psi - \delta)\, d\psi = \mathbb{P}[\hat{I} \le x], \quad (25)$$

as $\lambda \to \infty$, where $x \ge 0$, i.e., $\hat{I}^\lambda \Rightarrow \hat{I}$, as $\lambda \to \infty$. Finally, (21), (25) and $D_{i,j} = a_i 1_{\{i=j\}} - a_i a_j$ yield the statement of the proposition:

$$\mathbb{P}\left[\hat{I}^\lambda \le x,\, \hat{I}^\lambda_i 1_{\{\hat{I}^\lambda > 0\}} \le x_i,\, i \ne K\right] = \mathbb{P}\left[I^\lambda \le x\sqrt{\nu^\lambda},\, I^\lambda_i 1_{\{I^\lambda > 0\}} \le I^\lambda c^\lambda_i / c^\lambda + x_i \sqrt{I^\lambda},\, i \ne K\right]$$
$$= \sum_{j=1}^{\lfloor x\sqrt{\nu^\lambda}\rfloor} \sum_{\substack{m:\, m_\Sigma = j \\ m_i \le j c^\lambda_i / c^\lambda + x_i \sqrt{j},\, i \ne K}} \tilde{\pi}^\lambda(m)$$
$$= \frac{1}{\sqrt{\nu^\lambda}} \sum_{j=1}^{\lfloor x\sqrt{\nu^\lambda}\rfloor} \sqrt{\nu^\lambda} \tilde{\pi}^\lambda(0) \left( \frac{1}{\sqrt{j}^{K-1}} \sum_{\substack{m:\, m_\Sigma = j \\ m_i \le j c^\lambda_i / c^\lambda + x_i \sqrt{j},\, i \ne K}} \frac{\sqrt{j}^{K-1} \tilde{\pi}^\lambda(m)}{\tilde{\pi}^\lambda(0)} \right)$$
$$\to \int_0^x \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_{K-1}} \varphi_D(\xi_1^{K-1}) \frac{\varphi(\psi - \delta)}{\Phi(\delta)}\, d\xi_{K-1} \ldots d\xi_1 d\psi$$
$$= \mathbb{P}[\hat{I} \le x]\, \mathbb{P}\left[\hat{I}_i 1_{\{\hat{I} > 0\}} \le x_i,\, i \ne K\right],$$

as $\lambda \to \infty$. This concludes the proof of the proposition. □

Next, we present the proof of Theorem 2.

*Proof of Theorem 2.* Lemma 1 yields

$$\mathbb{P}[I^\lambda \le 0] = \mathbb{P}[\hat{I}^\lambda \le 0] = \sqrt{\lambda}\pi^\lambda(0) \frac{\sqrt{\lambda}}{c^\lambda - \lambda} = \sqrt{\nu^\lambda}\pi^\lambda(0) \frac{\rho^\lambda}{\sqrt{\nu^\lambda}(1 - \rho^\lambda)} \quad (26)$$

and (see (7))

$$\sqrt{\nu^\lambda}\pi^\lambda(0) = \left( \frac{\rho^\lambda}{\sqrt{\nu^\lambda}(1 - \rho^\lambda)} + \frac{\mathbb{P}[I^\lambda \le 0]}{\sqrt{\nu^\lambda}\pi^\lambda(0)} \right)^{-1}. \quad (27)$$

The relationship between the original system and the corresponding loss system implies that the second ratio in the preceding equality is equal to the same ratio for the loss system, i.e., the left-hand side of (23) in the proof of Proposition 1. Equations (27) and (23), together with the QED limit (C2) from Section 3.2, result, as $\lambda \to \infty$, in

$$\sqrt{\nu^\lambda}\pi^\lambda(0) \to \left( \frac{1}{\delta} + \frac{\Phi(\delta)}{\varphi(\delta)} \right)^{-1},$$

and, hence, due to (26) and (C2), as $\lambda \to \infty$,

$$\mathbb{P}[\hat{I}^\lambda \le 0] \to \left( 1 + \delta \frac{\Phi(\delta)}{\varphi(\delta)} \right)^{-1}.$$

Finally, from Proposition 1 we deduce $\mathbb{P}[\hat{I}^\lambda > x \,|\, \hat{I}^\lambda > 0] \to \Phi(\delta - x)/\Phi(\delta)$, $x \geq 0$, as $\lambda \to \infty$, the independence of $\hat{I}$ and $(\hat{I}_1, \dots, \hat{I}_K)$, as well as the distribution of $(\hat{I}_1, \dots, \hat{I}_K)$. The limit $\mathbb{P}[\hat{I}^\lambda \leq x \,|\, \hat{I}^\lambda \leq 0] \to e^{\delta x}$, $x \leq 0$, as $\lambda \to \infty$, follows from Lemma 1 and (C2). □

B.3. **Proof of Corollary 1.** The limit $\mathbb{E}(\hat{I}^\lambda)^- \to \mathbb{E}(\hat{I})^-$, as $\lambda \to \infty$, is immediate from the expression for $\mathbb{E}(\hat{I}^\lambda)^-$ (see Appendix B.4) and $\mathbb{E}(\hat{I})^- = -\mathbb{E}[\hat{I} \,|\, \hat{I} \leq 0]\,\mathbb{P}[\hat{I} \leq 0]$. The remaining two limits can be obtained by considering the loss system as in the proof of Theorem 2. By using the same argument as in (23) and recalling (24), we obtain

$$\mathbb{E}[\hat{I}^\lambda \,|\, \hat{I}^\lambda \geq 0] \to \frac{1}{\Phi(\delta)} \int_0^\infty \psi\, \varphi(\psi - \delta)\, d\psi = \mathbb{E}[\hat{I} \,|\, \hat{I} \geq 0]$$

and

$$\mathbb{E}[I^\lambda/\sqrt{\nu^\lambda} \,|\, \hat{I}^\lambda \geq 0] \to \frac{a_i}{\Phi(\delta)} \int_0^\infty \psi\, \varphi(\psi - \delta)\, d\psi = a_i \mathbb{E}[\hat{I} \,|\, \hat{I} \geq 0],$$

as $\lambda \to \infty$. □

B.4. **RMI Performance Measures.** Here we provide a summary of performance measure in the inverted-V model under RMI routing. The results are based on Lemma 1 and Theorem 2. Let $W^\lambda$ denote the stationary waiting time in the system with the arrival rate $\lambda$.

Finite-$\lambda$ case:

$$\begin{aligned}
\mathbb{P}[W^\lambda > 0] &= \mathbb{P}[I^\lambda \leq 0] = \pi^\lambda(0) \frac{1}{1-\rho^\lambda}, \\
\mathbb{P}[(I^\lambda)^- = i \,|\, W^\lambda > 0] &= (1-\rho^\lambda)(\rho^\lambda)^i, \quad i = 0, 1, \dots, \\
\mathbb{P}[(I^\lambda)^- = i] &= (1-\rho^\lambda)(\rho^\lambda)^i\, \mathbb{P}[W^\lambda > 0], \quad i = 0, 1, \dots, \\
\mathbb{E}(I^\lambda)^- &= \frac{\rho^\lambda}{1-\rho^\lambda}\, \mathbb{P}[W^\lambda > 0], \\
\mathbb{E}W^\lambda &= \frac{\mathbb{E}(I^\lambda)^-}{\lambda} = \frac{1}{c^\lambda(1-\rho^\lambda)}\, \mathbb{P}[W^\lambda > 0], \\
\mathbb{P}[W^\lambda > w] &= \mathbb{P}[W^\lambda > 0]\, e^{-(c^\lambda - \lambda)w}, \quad w \geq 0,
\end{aligned}$$

where $\pi^\lambda(0)$ is given in Lemma 1.

QED regime, as $\lambda \to \infty$:

$$\begin{aligned}
\sqrt{\nu^\lambda}\pi^\lambda(0) &\to \left(\frac{1}{\delta} + \frac{\Phi(\delta)}{\varphi(\delta)}\right)^{-1}, \\
\frac{1}{\sqrt{\nu^\lambda}} \mathbb{E}(I^\lambda)^- &\to \left(\delta + \delta^2 \frac{\Phi(\delta)}{\varphi(\delta)}\right)^{-1}, \\
\sqrt{\nu^\lambda}\hat{\mu}\, \mathbb{E}W^\lambda &\to \left(\delta + \delta^2 \frac{\Phi(\delta)}{\varphi(\delta)}\right)^{-1}, \\
\mathbb{P}[\sqrt{\nu^\lambda}\hat{\mu} W^\lambda > w \,|\, W^\lambda > 0] &\to e^{-\delta w}, \quad w \geq 0.
\end{aligned}$$

## C. Table of Main Notation

| | |
|---|---|
| $\lambda$ | arrival rate |
| $K$ | number of server pools |
| $\mu_i$ | service rate of a server in pool $i$ |
| $N_i^\lambda$ | number of servers in pool $i$ |
| $N^\lambda = \sum_{i=1}^K N_i^\lambda$ | total number of servers in the system |
| $c_i^\lambda = N_i^\lambda \mu_i$ | service capacity of pool $i$ |
| $c^\lambda = \sum_{i=1}^K c_i^\lambda$ | total service capacity |
| $I_i^\lambda(t)$ | number of idle servers in pool $i$ at time $t$ |
| $I_i^\lambda$ | stationary number of idle servers in pool $i$ |
| $I^\lambda(t)$ | number of idle servers / customers awaiting service at time $t$ |
| $I^\lambda$ | stationary number of idle servers / customers awaiting service |
| $W^\lambda$ | stationary waiting time |
| $\rho^\lambda = \lambda / c^\lambda$ | total traffic intensity |
| $\rho_i^\lambda = 1 - \mathbb{E} I_i^\lambda / N_i^\lambda$ | mean stationary occupancy rate in pool $i$ (servers' utilization in pool $i$) |
| $\gamma_i = \mu_i \rho_i$ | flux through pool $i$ (average number of arrivals per pool $i$ server per time unit) |
| $a_i$ $(c_i^\lambda / c^\lambda \to a_i)$ | limiting service capacity proportion of pool $i$ |
| $q_i$ $(N_i^\lambda / N^\lambda \to q_i)$ | limiting fraction of servers in pool $i$ out of the total number of servers |
| $\mu = (\sum_{i=1}^K a_i / \mu_i)^{-1}$ | mean (harmonic) service rate |
| $\hat{\mu} = \sum_{i=1}^K a_i \mu_i$ | mean (arithmetic) service rate |
| $\nu^\lambda = \lambda / \hat{\mu}$ | scaling parameter (system size) |
| $\delta$ $(\sqrt{\nu^\lambda}(1 - \rho^\lambda) \to \delta)$ | square root safety capacity coefficient |
| $\beta = \delta \sqrt{\hat{\mu} / \mu}$ | Quality-of-Service (QoS) parameter |
| $x^+ = \max\{x, 0\}$ | positive part |
| $x^- = \max\{-x, 0\}$ | negative part |
| $\mathbb{E}$ | expectation |
| $\mathbb{P}$ | probability |
| $\Phi(\cdot)$, $\varphi(\cdot)$ | standard normal distribution and density functions |
| $\Rightarrow$ | convergence in distribution |

## D. List of Acronyms

| | |
|---|---|
| ALOS | Average Length of Stay |
| ED | Emergency Department |
| FCFS | First-Come First-Served |
| FSF | Fastest Servers First |
| IR | Idleness-Ratio |
| IW | Internal Ward |
| LIPF | Longest-Idle Pool First |
| LISF | Longest-Idle Server First |
| LOS | Length of Stay |
| LWISF | Longest-Weighted-Idle Server First |
| PASTA | Poisson Arrivals See Time Averages |
| RA | Random Assignment |
| RMI | Randomized Most-Idle |
| QED | Quality and Efficiency Driven |
| QIR | Queue-and-Idleness-Ratio |
| QoS | Quality-of-Service |
| WRMI | Weighted Randomized Most-Idle |

Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa 3200, Israel

*E-mail address*: avim@tx.technion.ac.il

Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, U.S.A.

*E-mail address*: petar@ise.ufl.edu

Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa 3200, Israel

*E-mail address*: yulia@tx.technion.ac.il