ONLINE APPENDIX TO "ON FAIR ROUTING FROM EMERGENCY DEPARTMENTS TO HOSPITAL WARDS: QED QUEUES WITH HETEROGENEOUS SERVERS"

AVISHAI MANDELBAUM, PETAR MOMČILOVIĆ, AND YULIA TSEYTLIN

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 (!) Hospital [1997]. 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 Hospital [1997] 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.

Date: March 14, 2012.

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.

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) Hospital [1997]. 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 Elkin and Rozenberg [2007], 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 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 of Mandelbaum et al.).

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

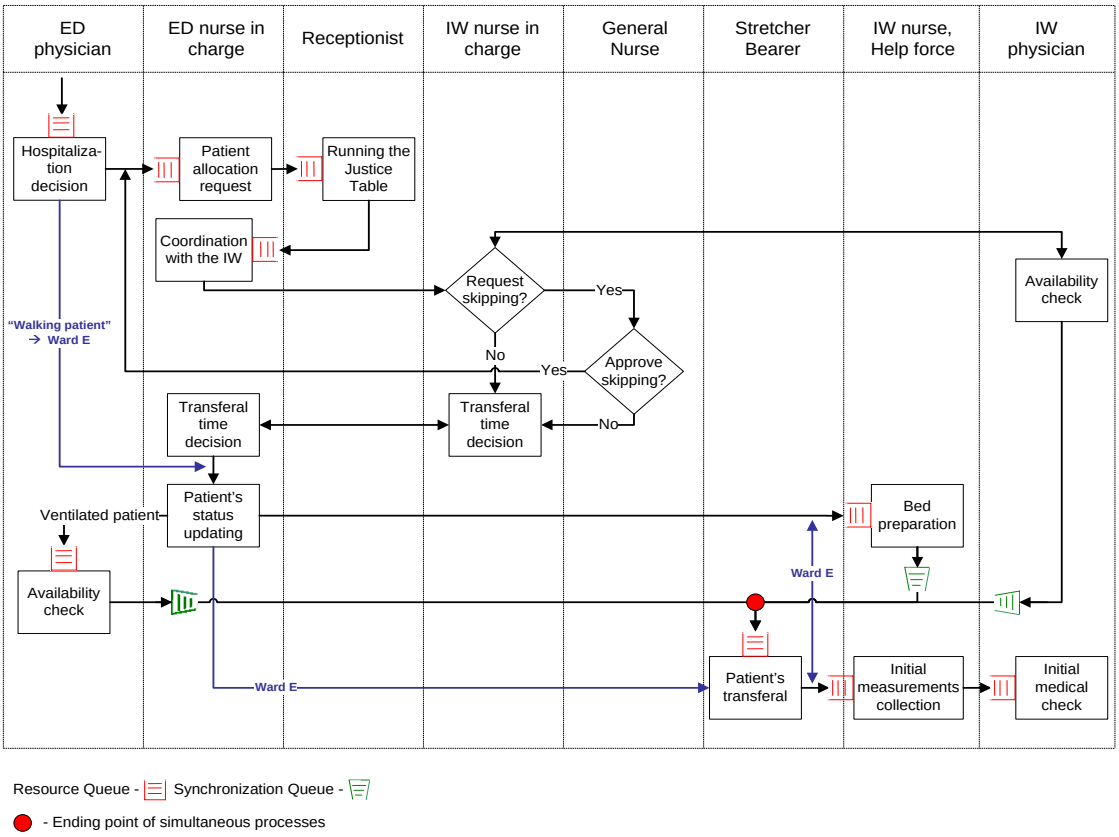
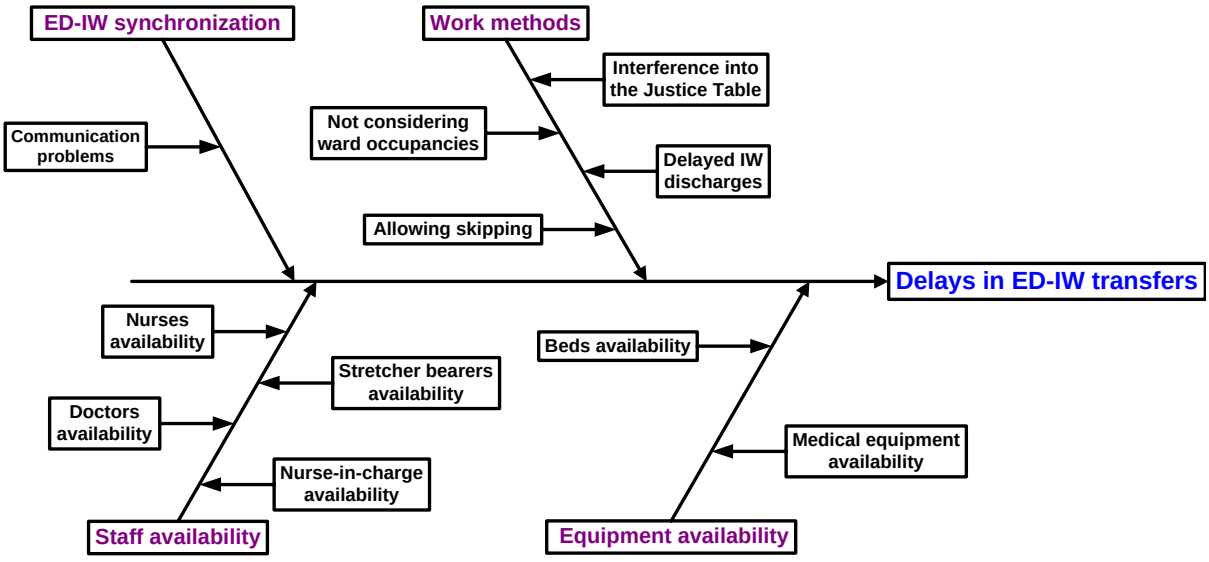


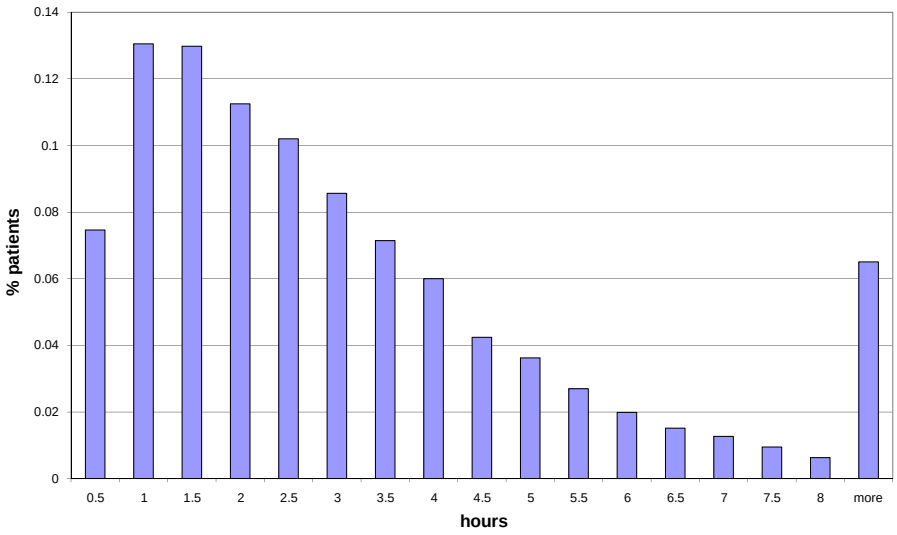
FIGURE 6. ED-to-IW Delays: Cause-and-Effect Chart.



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 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.

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).

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 Richardson [2006], Miro et al. [1999], Sprivulis et al. [2006], Trzeciak and Rivers [2003].

A.4. **Other Hospitals.** The hospitals we study differ in their functionality and geographical location. From Table 1 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 ynet [2009]), 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 estimate. Those estimated waiting times are long (again, except for Hospital 5) – indeed, the situation in Anonymous Hospital is not the worst.

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 ynet [2009].

** Differ in their capacities and LOS.

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

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.

B. Technical Appendix

B.1. Proof of Theorem 1 (Theorem 1 in Mandelbaum et al.)

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 $\gamma_i^\lambda > \gamma_j^\lambda$.*

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 Tseytlin [2009])**.** *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 π^λ*:*

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

where

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

For notational convenience, we prove the theorem for the single-server-pool model under the Random Assignment policy (note that Cabral Cabral [2007] proved this result independently); this model is equivalent to the inverted-V model under the RMI policy by Theorem 4.3.1 in Tseytlin [2009]. 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^λ) 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:

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

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

The last sum is equal in both expressions, and (2) implies that the first sum in (3) is strictly smaller than the corresponding sum in (4). Hence, the first statement of the lemma: $p^j_m < p^k_m$ for $m \in \{1, 2, \ldots, N^\lambda - 1\}$.

Next, we consider the second statement of the lemma. Multiplying each side of (3) by μ_j and each side of (4) by μ_k, yields, for $\mathcal{X} \subset \{1, \ldots, N^\lambda\} \setminus \{j, k\}$:

$$\mu_j \, p^j_m = \sum_{\mathcal{X}:|\mathcal{X}|=m-1} \mu_j \, \pi^\lambda_{j\cup\mathcal{X}} + \sum_{\mathcal{X}:|\mathcal{X}|=m-2} \mu_j \, \pi^\lambda_{\{j,k\}\cup\mathcal{X}}, \tag{11}$$

$$\mu_k \, p^k_m = \sum_{\mathcal{X}:|\mathcal{X}|=m-1} \mu_k \, \pi^\lambda_{k\cup\mathcal{X}} + \sum_{\mathcal{X}:|\mathcal{X}|=m-2} \mu_k \, \pi^\lambda_{\{j,k\}\cup\mathcal{X}}, \tag{12}$$

where

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

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 (5), (6), (7) and (8) imply $\mu_j \, p^j_m > \mu_k \, p^k_m$ for $m \in \{1, 2, \ldots, N - 1\}$. □

Next, we complete the proof of Theorem 1 (Theorem 1 in Mandelbaum et al.).

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 (Theorem 2 in Mandelbaum et al.)**

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

$$\left(\hat{I}^\lambda, (\hat{I}^\lambda_1, \ldots, \hat{I}^\lambda_K) 1_{\{\hat{I}^\lambda > 0\}}\right) \Rightarrow \left(\hat{I}, (\hat{I}_1, \ldots, \hat{I}_K) 1_{\{\hat{I} > 0\}}\right), \tag{15}$$

where $\hat{I}$ and $(\hat{I}_1, \ldots, \hat{I}_K)$ 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 $(\hat{I}_1, \ldots, \hat{I}_K)$ is zero-mean multi-variate normal, with $\mathbb{E}\hat{I}_i\hat{I}_j = a_i 1_{\{i=j\}} - a_i a_j$.*

The proof of the theorem is based on the reversibility Kelly [1979] of the process $\{(I^\lambda(t), I_1^\lambda(t)), \dots, I_K^\lambda(t)), t \geq 0\}$. 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*

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

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,

$$\det(D^{-1}) = \det(B)\det(C) = \det(B)\prod_{i=1}^{K-1} a_i. \tag{17}$$

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

$$\det(B) = \det(A) = 1 - \sum_{i=1}^{K-1} a_i = a_K. \tag{18}$$

The statement (10) follows from (11) and (12). □

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 $(\hat{I}^\lambda, (\hat{I}_1^\lambda, \dots, \hat{I}_K^\lambda)1_{\{\hat{I}^\lambda>0\}}) \Rightarrow (\hat{I}, (\hat{I}_1, \dots, \hat{I}_K)1_{\{\hat{I}>0\}})$, as $\lambda \to \infty$, where $\hat{I}$ and $(\hat{I}_1, \dots, \hat{I}_K)$ are independent, $\mathbb{P}[\hat{I} > x] = \Phi(\delta - x)/\Phi(\delta)$, $x \geq 0$, 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. Let $\tilde{\pi}^\lambda(m)$ be the stationary probability of having m_i idle servers in pool $i = 1, \dots, K$ in the loss model. Here, $m = (m_1, \dots, m_K)$; 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 (14)) and establish the probabilities of those states relative to the probability that all servers are busy ($\tilde{\pi}^\lambda(0)$, see (16)). 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 (17)). From this, the asymptotic probability of having no idle servers can be derived (see (19)). 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)

$$\tilde{\pi}^\lambda(m) = m_\Sigma!\,\tilde{\pi}^\lambda(0)\prod_{i=1}^K \left(\frac{\mu_i}{\lambda}\right)^{m_i}\binom{N_i^\lambda}{m_i}, \tag{19}$$

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

$$\dot{m}_i = \frac{c_i^\lambda}{c^\lambda}\sqrt{\psi^2\nu^\lambda} + \xi_i\sqrt[4]{\psi^2\nu^\lambda}, \tag{20}$$

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

$$\frac{c_i^\lambda \dot{m}_\Sigma}{c^\lambda\,\dot{m}_i} = \frac{\sqrt{\nu^\lambda}}{\sqrt{\nu^\lambda} + \xi_i\sqrt[4]{\nu^\lambda}\,c^\lambda/(\sqrt{\psi}\,c_i^\lambda)}. \tag{21}$$

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, (13) 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 (14), (15) 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:

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

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

$$\begin{aligned}
\frac{(\psi^2\nu^\lambda)^{\frac{K-1}{4}}\tilde{\pi}^\lambda(\dot{m})}{\tilde{\pi}^\lambda(0)} &\to \frac{1}{\sqrt{(2\pi)^{K-1}\prod_{i=1}^K a_i}}\,e^{\psi\delta-\frac{1}{2}\psi^2-\frac{1}{2}\sum_{i=1}^K \xi_i^2/a_i}\\
&= \frac{1}{\sqrt{(2\pi)^{K-1}\,\det(D)}}e^{-\frac{1}{2}(\xi_1^{K-1})'D^{-1}\xi_1^{K-1}}e^{-\frac{1}{2}(\psi-\delta)^2}e^{\frac{1}{2}\delta^2}\\
&= \varphi_D(\xi_1^{K-1})\,\frac{\varphi(\psi-\delta)}{\varphi(\delta)},
\end{aligned} \tag{22}$$

where φ_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 $\psi > 0$, (16) and the fact that the number of summands in the following sum is polynomial in ν^λ (bounded from above by $(\psi\sqrt{\nu^\lambda})^K$) result in, as $\lambda \to \infty$,

$$\begin{aligned}\sum_{m:\, m_\Sigma = \lfloor \psi\sqrt{\nu^\lambda}\rfloor} \frac{\tilde\pi^\lambda(m)}{\tilde\pi^\lambda(0)} &\to \frac{\varphi(\psi-\delta)}{\varphi(\delta)} \int_{-\infty}^{\infty}\cdots\int \varphi_D(\xi_1^{K-1})\, d\xi_1\cdots d\xi_{K-1} \\ &= \frac{\varphi(\psi-\delta)}{\varphi(\delta)},\end{aligned} \tag{23}$$

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 $\lambda \to \infty$,

$$\begin{aligned}\sum_{m_1=0}^{N_1^\lambda}\cdots\sum_{m_K=0}^{N_K^\lambda} \frac{\tilde\pi^\lambda(m)}{\sqrt{\nu^\lambda}\tilde\pi^\lambda(0)} &= \sum_{i=1}^{N^\lambda}\sum_{m:\, m_\Sigma = i} \frac{\tilde\pi^\lambda(m)}{\sqrt{\nu^\lambda}\tilde\pi^\lambda(0)} \\ &\to \frac{1}{\varphi(\delta)}\int_0^\infty \varphi(\psi-\delta)\, d\psi \\ &= \frac{\Phi(\delta)}{\varphi(\delta)},\end{aligned} \tag{24}$$

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

$$\sqrt{\nu^\lambda}\tilde\pi^\lambda(0) \to \frac{\varphi(\delta)}{\Phi(\delta)}. \tag{25}$$

Moreover, (17) and (18) 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,

$$\begin{aligned}\mathbb{P}[\hat I^\lambda \le x] &= \sum_{j=1}^{\lfloor x\sqrt{\nu^\lambda}\rfloor}\sum_{m:\, m_\Sigma = j}\tilde\pi^\lambda(m) \\ &= \frac{1}{\sqrt{\nu^\lambda}}\sum_{j=1}^{\lfloor x\sqrt{\nu^\lambda}\rfloor}\left(\sqrt{\nu^\lambda}\sum_{m:\, m_\Sigma = j}\tilde\pi^\lambda(m)\right) \to \frac{1}{\Phi(\delta)}\int_0^x \varphi(\psi-\delta)\, d\psi = \mathbb{P}[\hat I \le x],\end{aligned} \tag{26}$$

as $\lambda \to \infty$, where $x \geq 0$, i.e., $\hat{I}^\lambda \Rightarrow \hat{I}$, as $\lambda \to \infty$. Finally, (16), (20) and $D_{i,j} = a_i 1_{\{i=j\}} - a_i a_j$ yield the statement of the proposition:

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

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

Next, we present the proof of Theorem 2 (Theorem 2 in Mandelbaum et al.).

Proof of Theorem 2. Lemma 1 yields

$$\mathbb{P}[I^\lambda \leq 0] = \mathbb{P}[\hat{I}^\lambda \leq 0] = \sqrt{\lambda}\pi^\lambda(0)\frac{\sqrt{\lambda}}{c^\lambda - \lambda} = \sqrt{\nu^\lambda}\pi^\lambda(0)\frac{\rho^\lambda}{\sqrt{\nu^\lambda}(1-\rho^\lambda)} \tag{27}$$

and (see (1))

$$\sqrt{\nu^\lambda}\pi^\lambda(0) = \left(\frac{\rho^\lambda}{\sqrt{\nu^\lambda}(1-\rho^\lambda)} + \frac{\mathbb{P}[I^\lambda \leq 0]}{\sqrt{\nu^\lambda}\pi^\lambda(0)}\right)^{-1}. \tag{28}$$

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 (18) in the proof of Proposition 1. Equations (22) and (18), together with the QED limit (C2) from Section 3.2 in Mandelbaum et al., 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 (21) and (C2), as $\lambda \to \infty$,

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

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 $(\hat{I}_1, \ldots, \hat{I}_K)$, as well as the distribution of $(\hat{I}_1, \ldots, \hat{I}_K)$. 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 (Corollary 1 in Mandelbaum et al.)**

Corollary 1. *Consider the inverted-V model in steady-state, under the RMI routing algorithm in the QED regime ((C1)-(C2) in Mandelbaum et al.). 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* $i = 1, \dots, K$, *where* $\hat{I}$ *is as in Theorem 2.*

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 $\mathbb{E}(\hat{I})^- = -\mathbb{E}[\hat{I} \,|\, \hat{I} \le 0]\,\mathbb{P}[\hat{I} \le 0]$. 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 (18) and recalling (19), we obtain

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

and

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

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^λ denote the stationary waiting time in the system with the arrival rate λ.

Finite-λ case:

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

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

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

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

References

F.B. Cabral. Queues with heterogeneous servers and uninformed customers: Who works the most? arXiv:0706.0560vl, 2007. B.1

K. Elkin and N. Rozenberg. Patients' flow from the emergency department to the internal wards. IE&M project, Technion, 2007. A.1

Anonymous Hospital. Reducing waiting times in the emergency department and LOS in the internal wards. Report of the Quality Promotion Team, 1997. A.1

F. Kelly. *Reversibility and Stochastic Networks*. Wiley, New York, 1979. B.2

A. Mandelbaum, P. Momčilović, and Y. Tseytlin. On fair routing from Emergency Departments to hospital wards: QED queues with heterogeneous servers. to appear in Management Science. A.2, B.1, B.1, B.2, 2, B.2, B.2, B.3, 1

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

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

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

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

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, B.1

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, 1

C. Table of Main Notation

λ	arrival rate
K	number of server pools
μ_i	service rate of a server in pool i
N_i^λ	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^λ	stationary number of idle servers in pool i
$I^\lambda(t)$	number of idle servers / customers awaiting service at time t
I^λ	stationary number of idle servers / customers awaiting service
W^λ	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)
δ $(\sqrt{\nu^\lambda}(1-\rho^\lambda) \to \delta)$	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., IBM Research Lab, Haifa University, Haifa 31905, Israel

E-mail address: yuliatse@gmail.com