Submitted to *Operations Research*
manuscript (Please, provide the mansucript number!)

Control of Patient Flow in Emergency Departments, or Multiclass Queues with Deadlines and Feedback

Junfei Huang
The Chinese University of Hong Kong, junfeih@cuhk.edu.hk

Boaz Carmeli
IBM Haifa Research Lab, BOAZC@il.ibm.com

Avishai Mandelbaum
Technion – Israel Institute of Technology, avim@ie.technion.ac.il

We consider the control of patient flow through physicians in emergency departments (EDs). The physicians must choose between catering to patients right after triage, who are yet to be checked, and those that are in-process (IP), who are occasionally returning to be checked. Physician capacity is thus modeled as a queueing system with multiclass customers, where some of the classes face deadline constraints on their time-till-first-service, while the other classes feedback through service while incurring congestion costs. We consider two types of such costs: first, costs that are incurred at queue-dependent rates, and second, costs that are functions of IP sojourn time. The former is our base-model, which paves the way for the latter (perhaps more ED-realistic). In both cases, we propose and analyze scheduling policies that, asymptotically in conventional heavy-traffic, minimize congestion costs while adhering to all deadline constraints. Our policies have two parts: the first chooses between triage and IP patients; assuming triage patients are chosen, the physicians serve the one who is closest to violating the deadline; alternatively, IP patients are served according to a G$c\mu$ rule, in which μ is simply modified to account for feedbacks. For our proposed policies, we establish asymptotic optimality, and develop some congestion laws (snapshot principles) that support forecasting of waiting and sojourn times. Simulation then shows that these policies outperform some commonly-used ones. It also validates our laws and demonstrates that some ED features, the complexity of which reaches beyond our model (e.g., time-varying arrival rates, Leave-Without-Being-Seen (LWBS) or Leave-Against-Medical-Advice (LAMA)), do not lead to significant performance degradation.

Key words: Emergency Department, Patient Flow Triage, ED or ER Crowding, Heavy Traffic, Feedback Queues, Due Date Stochastic Control, ESI

1. Introduction

Control of patient flow is a major factor for improving hospital operations. Indeed, patient flow is a central driver of a hospital's operational performance, which is tightly coupled with the overall quality and cost of health care (Armony et al. (2013), Pitts et al. (2008), Niska et al. (2010)). In this work, we address the challenge of flow control at the main hospital "gate"—the Emergency Department (ED). The challenge stems from two flow characteristics: *deadlines and feedbacks.* First, arriving patients must be served within time-deadlines that are assigned after triage, based on clinical considerations (Farrohknia et al. (2011), Mace and Mayer (2008)). Second, ED flows have a significant feedback component that must be accounted for: in-process (IP) patients possibly return several times to physicians during their ED sojourn, before ultimately being either released or hospitalized (Yom-Tov and Mandelbaum (2014), Table 2).

IP patients present both *clinical* concerns (e.g. stabilizing their conditions) and *operational* concerns (e.g. they occupy beds). They should thus complete their treatments and leave the ED as soon as possible. On the other hand, *clinical* Triage constraints should be adhered to, so that arriving patients start their first treatment within pre-specified time windows. It is this Triage-IP friction that we focus on, doing so from the viewpoint of the *ED physician*: when becoming idle, what class should be served next—triage or in-process—after which one must decide on the specific patient to be examined.

To this end, ED dynamics are captured by a multiclass queueing system, with multiple servers (physicians), multiclasses of *triage* patients and multiclasses of *in-process* (IP) patients. (A patient class could embody information such as treatment type, emergency level or age; see Carmeli (2012).) Patients within each class are served on a First-Come-First-Served (FCFS) basis. The triage patients arrive to the system exogenously and are yet to be examined by a physician; each such patient must be served within a time-deadline from its arrival. After completing their first service, triage patients join the queue of IP patients, or exit the system. IP patients originate from either triage patients or from previous IP phases, and they require further treatment. While waiting, the IP patients incur queueing costs. Our objective is to minimize these cumulative costs, among all policies that satisfy the triage (deadline) constraints.

The objective can be achieved (asymptotically, see §5.2), if a physician that becomes idle adopts the following guidelines (a two-step policy):

- *First step – Triage or IP:* Triage patients have priority if any triage patient's deadline is close to being violated;
- *Second step (a) – Triage:* Given that a triage patient is to be served, the priority goes to the patient that is closest to violating a deadline;
- *Second step (b) – IP:* Given that an IP patient is to be served, a modified generalized $c\mu$-rule is used to decide which patient to serve.

Despite the apparent simplicity of its solution, the problem is not easy to solve. First, patients' waiting times are random while the deadlines are deterministic; hence consistently satisfying this deterministic constraint is too much to hope for, which calls for a rigorous formulation in an asymptotic sense (see §4). Second, multiclass queueing with feedback is in itself challenging to analyze, to which one adds deadline constraints.

Our mathematical framework is conventional heavy-traffic, where one analyzes a sequence of systems that approach critical loading. Within this framework, IP analysis follows the G$c\mu$-rule of van Mieghem (1995), after generalizing it to models with feedback. Triage analysis combines the due-date scheduling in van Mieghem (2003) with the formulation of Plambeck et al. (2001). The latter offers a rigorous meaning for adherence to (triage) time-constraints, by introducing “asymptotic compliance” as a relaxation for “feasibility”. Together, triage and in-process controls yield what we prove to be asymptotically optimal flow-control policies: they minimize IP congestion costs subject to triage compliance, that is, the above policy is asymptotically feasible and asymptotically optimal among all asymptotically feasible policies.

In our analysis, we assume that the deadlines for triage patients are not short. This is not necessarily true for all triage classes. Consider the Emergency Severity Index (ESI) for example (Mace and Mayer (2008)): patients are separated into 5 triage classes, and the physician response times for classes 1 and 2 patients should be within minutes. Our policy can be modified to systems with patients facing short deadlines by assigning them the highest priority so that they start treatment immediately upon their arrival. From State-Space-Collapse results (Bramson (1998)), their queue lengths and waiting times would be negligible in heavy traffic scaling. Thus, essentially without loss of generality, we focus on those patients whose deadlines are not short. For the ESI, these are patients in triage classes 3, 4 and 5.

In addition to queueing costs, we consider also models with waiting costs and sojourn time costs, and provide policies which minimize these costs while ensuring that triage deadline constraints are adhered to; see §6.

Why conventional heavy traffic? This is a relevant operational regime. Specifically, our experience suggests that, during peak load between late morning and late evening, the ED can be usefully viewed as a critically-loaded stationary system (Armony et al. (2013)). Moreover, simulation experiments (§7 and §EC.9) demonstrate that our proposed policies actually perform well over the whole (time-varying) day.

Beyond our ED models: Saghafian et al. (2012) remark that, due to the complexity of ED operations, it is challenging to capture prevalent ED features within a single tractable analytic model. While this is precisely what we do here, ours is by no means the final story. Additional ED features that seek modeling include time-varying arrival rates, treatment times between successive visits to the physician, limitation on the number of beds and ambulance diversion (admission control), “non-interchangeable” physician service, and patients who Leave-Without-Being-Seen (LWBS) or Leave-Against-Medical-Advice (LAMA). We comment on these features and offer related conjectures in Section 8. Moreover, we simulated systems incorporating these features and under various policies. The results, included in §7 and §EC.9, demonstrate that our proposed policy outperforms commonly-used ones, even with these additional features taken into account.

1.1. Literature review

There is ample medical literature about triage systems; we refer the reader to Farrohknia et al. (2011), Mace and Mayer (2008). Our research focus here is operational (Marmor et al. (2012), Wiler et al. (2010)) and, accordingly, so is the following literature review.

To the best of our knowledge, our paper is the first to analyze the control of patient flow in an ED from a queueing-theory perspective. (In contrast, there are practically hundreds of simulation-based studies; e.g., Brailsford et al. (2009).) Since starting this project, additional work has appeared on ED operations. The closest to ours are Saghafian et al. (2014, 2012): Saghafian et al. (2014) discuss a complexity-based triage system, based on the number of visits that patients pay to the ED physician (serving as an up-front proxy for complexity); and Saghafian et al. (2012) analyze the advantage of streaming patients (separating them into classes, e.g. by their admission vs. discharge status), comparing this practice against pooling and, what they call, "virtual-streaming". The latter supplements class-separation with dynamic resource allocation, and it is shown to dominate the other two. There are additional papers that cater to specific ED characteristics: Yom-Tov and Mandelbaum (2014) model the ED as a single-class time-varying queueing system with feedback (Erlang-R), operating in the QED regime, and in support of staffing physicians and nurses; Dobson et al. (2013) develop an overloaded queueing network to analyze the impact of interruptions on ED throughput; Zayas-Caban et al. (2013) considered a two-stage tandem queueing model for an ED triage and treatment process; and Atar et al. (2012) is relevant to synchronization of ED activities (e.g. interpretations of a blood-test and X-ray imaging must precede a visit to the ED physician), as it analyzes a fork-join queueing network in heavy-traffic. Finally, Zaied (2012) models an ED as a time-varying fork-join network; he then uses square-root staffing of physicians and/or nurses to stabilize ED performance.

Our ED models and analysis follow two main lines of research: formulation of the triage constraints is adapted from Plambeck et al. (2001), who analyze admission control; and our IP control generalizes van Mieghem (1995), who solves a cost minimization problem for a multiclass queue without feedback. The results in van Mieghem (1995) have been generalized by Mandelbaum and Stolyar (2004) to a feedforward network of parallel queues, and both papers establish asymptotic optimality of the generalized $c\mu$-rule. Here we generalize van Mieghem (1995) to a model with both feedback and deadlines, and prove asymptotic optimality of a routing rule in which a modified generalized $c\mu$-rule plays a central role.

Our model structure for IP patients resembles Klimov (1974, 1978), where the author considers a dynamic scheduling problem of a multiclass $M/GI/1$ queueing system with Markovian feedback. Unlike Klimov (1974, 1978), who minimizes a cost function that is linear in average queue lengths and proves the optimality of a static routing policy, here we consider a minimization problem with cumulative costs over a finite horizon, with cost rates that are convex functions of queue lengths (or waiting times), which gives rise to asymptotic optimality of a dynamic routing policy. Notably, our analysis of IP patients in fact covers Klimov: simply take the deadlines and means of service times for triage patients to be 0. We thus establish, indirectly, asymptotic optimality of the generalized $c\mu$-rule also for Klimov's model (with convex costs). Our method can also accommodate linear cost functions, for which a modified $c\mu$-type rule is asymptotically optimal; see Remark 3. A final related reference is Chen and Yao (1993), which concerns dynamic scheduling of a multiclass fluid network with feedback.

Diffusion approximations for queueing systems with multiclass customers and Markovian feedback have been analyzed in Reiman (1988) and Dai and Kurtz (1995), under the assumption of a global FCFS service discipline among all classes. Our analysis can also be adapted to prove convergence of the queue length processes there, as well as to other work-conserving disciplines. Indeed, our present results yield convergence of the weighted queue length to a reflected Brownian motion, under any work-conserving policy; then, proving convergence of individual queue lengths, for each class, amounts to establishing state-space collapse, which will follow from standard arguments (e.g. Bramson (1998)).

1.2. Contributions and outline

We view our main contributions to be the following:

- **Methodological.** We analyze multiclass queueing systems with feedback, in particular:

1. Identifying asymptotically optimal policies, which are simpler than the conjecture in Mandelbaum and Stolyar (2004) regarding feedback;
2. Solving Klimov's model with convex costs, for both queueing- and sojourn-costs;
3. Analyzing multiclass queueing systems with feedback, under any work-conserving policy;
4. Accommodating jointly delay constraints and congestion costs.

- **Practical.** We model and analyze the control of patient flow in EDs, from the point of view of ED physicians, which naturally gives rise to a queueing perspective:

1. Our models capture the tradeoff between clinical (triage) vs. operational (IP) concerns;
2. They yield scheduling policies that are insightful and implementable (minimizing IP-congestion subject to triage constraints);
3. They give rise to congestion laws that support forecasting of sojourn times.

With our method, one can also prove the conjecture in Mandelbaum and Stolyar (2004) and analyze the value of information embedded in the ED process (e.g. number of physician visits); see Huang (2013). Additional references will be provided in Section 8, where we propose generalizations to our main models, accompanied by corresponding conjectures.

Paper Outline: The rest of the paper is organized as follows. We end this introduction with a summary of notation. A detailed description of the basic ED model is given in §2. Heavy traffic conditions, asymptotic compliance and optimality are introduced in §3 and §4, respectively. The main results and some auxiliary propositions and extensions are presented in §5, with their discussions in §6. In §7 we describe simulation experiments that validate our analysis and proposed policies. We conclude with a discussion of future research directions in §8. The proofs for the main theorems, as well as additional proofs (of propositions) and complements are provided in the Appendix.

1.3. Notation

We use the standard notation $\mathbb{R}_+$ to denote the set of nonnegative real numbers. For a real number x, $\lfloor x \rfloor$ is the maximal integer less than or equal to x; $\mathbb{R}_+^J$ and $\mathbb{R}_+^K$ are the J-time and K-time products of $\mathbb{R}_+$, respectively; $\mathbb{Z}_+^K$ is the subset of $\mathbb{R}_+^K$ with all components being integers. Unless otherwise specified, all vectors are assumed to be column vectors. We reserve the notation $\{e_k\}$ for the standard basis of $\mathbb{R}^K$. The transposition of a vector or a matrix is indicated with a superscript T. Vector inequalities are understood to be componentwise; e.g., for $x, y \in \mathbb{R}^K$, $x < y$ if and only if $x_i < y_i$, for all $i = 1, 2, \ldots, K$. We use 0 to denote a column vector with all components being 0, with the dimension being clear from the context. For a matrix M, $M_{j\cdot}$ denotes the jth row, and $M_{\cdot k}$ the kth column of M. The function $1(\cdot)$ is the indicator function, the value of which is 1 when the event within $(\cdot)$ prevails, and 0 otherwise.

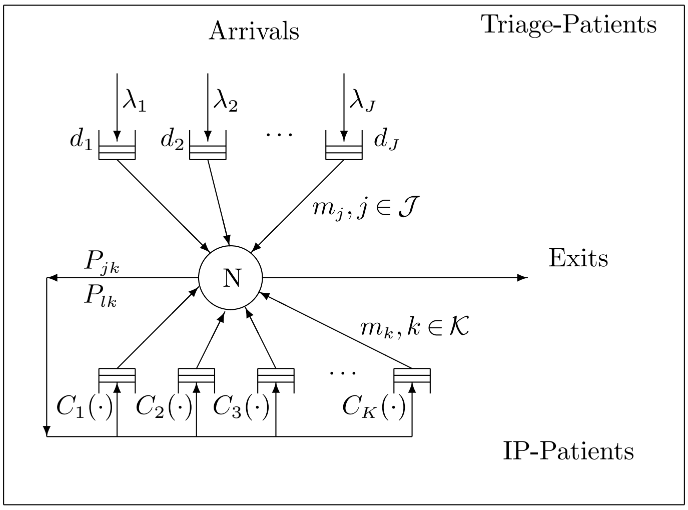
We assume that all random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Expectation with respect to $\mathbb{P}$ is $\mathbb{E}$. Let $\mathcal{D}[0,\infty)$ be the standard Skorohod space of right-continuous left-limit (RCLL) functions, defined on $[0,\infty)$ and equipped with the Skorohod J_1 topology. Similarly to $\mathcal{D}[0,\infty)$, $\mathcal{D}[0,t]$ is the space of functions on $[0,t]$. The symbol $\Rightarrow$ denotes weak convergence of stochastic processes, and $\rightarrow$ stands for convergence of non-random elements in $\mathcal{D}[0,\infty)$. The joint convergence of two or more processes will be understood implicitly from the context; it will be denoted by $j \in \mathcal{J}$ or $k \in \mathcal{K}$ or both (see e.g. (17)). Finally, $e(\cdot)$ is the 1-dimensional identity function on $\mathbb{R}_+$, where $e(t) = t$, $t \geq 0$.

2. The basic model

Consider the N-server queueing system in Figure 1: it has J classes of *triage* customers, who must adhere to deadline constraints, jointly with K classes of *in-process* (IP) customers who incur queueing costs. To highlight the application to EDs, we use "patient" interchangeably with "customer" and "physician" with "server". Let $\mathcal{J}$ and $\mathcal{K}$ denote the index sets of triage and IP patients, respectively: $j \in \mathcal{J}$ is an index for triage patients, and $l, k \in \mathcal{K}$ are indices for IP patients. It will be convenient to let $\mathcal{J} = \{1, 2, \ldots, J\}$ and $\mathcal{K} = \{1, 2, \ldots, K\}$, while keeping in mind that the indices $1, 2, \ldots$ in $\mathcal{J}$ differ from those in $\mathcal{K}$. To avoid ambiguity, we do write $j \in \mathcal{J}$ and $l, k \in \mathcal{K}$ as necessary.

Remark 1 *Due to our conventional heavy-traffic framework in §3, the N-physician system in Figure 1 is (asymptotically) equivalent to the same system but with $N = 1$, in which the single server is a "super" physician that is N-times faster than each of the original physicians (this can be proved as in Chen and Shanthikumar (1994), and supported by simulation in §EC.9.4); we thus assume hereafter that $N = 1$.*

Figure 1 Patient flow through the emergency department



2.1. Triage patients

For each triage class $j \in \mathcal{J}$ of patients, we are given two independent sequences of i.i.d. random variables, $\{u_j(i), i = 1, 2, \ldots\}$ and $\{v_j(i), i = 1, 2, \ldots\}$, as well as two real numbers λ_j and m_j. We assume $\mathbb{E}[u_j(1)] = 1$, $\mathbb{E}[v_j(1)] = 1$, and denote $a_j^2 = \text{var}(u_j(1))$, $b_j^2 = \text{var}(v_j(1))$. Among j-triage patients, the interarrival time between the $(i-1)$st and ith arrivals is $u_j(i)/\lambda_j$, and the service time required for the ith patient is $m_j v_j(i)$. As a result, λ_j is the arrival rate and m_j is the mean service time requirement of a j-triage patient. Assume $\lambda_j > 0$, for all $j \in \mathcal{J}$, then let $\Lambda_{\mathcal{J}}$ be the vector with components $\lambda_j, j \in \mathcal{J}$. Denote by $M_{\mathcal{J}}$ the vector with components $m_j, j \in \mathcal{J}$.

For $t \geq 0$ and $j \in \mathcal{J}$, let the renewal process

$$E_j(t) := \max\left\{ n \geq 0 : \sum_{i=1}^{n} u_j(i) \leq \lambda_j t \right\}$$

model the number of j-triage arrivals till time t, and the renewal process

$$S_j(t) := \max\left\{ n \geq 0 : \sum_{i=1}^{n} m_j v_j(i) \leq t \right\}$$

denote the number of service completions after the physician has devoted t time units to j-triage patients. Let $\mu_j = 1/m_j$, which is the *service rate* for j-triage patients.

Among each class of triage patients, the service discipline is First-Come-First-Served (FCFS). After completing service, a j-triage patient will join the queue of k-IP patients, with probability P_{jk}, or leave the system directly, with probability $1 - \sum_{k \in \mathcal{K}} P_{jk}$. Let the matrix $P_{\mathcal{J}\mathcal{K}} = (P_{jk})_{J \times K}$ be the triage-to-IP matrix. We use $\phi_j(n)$ to denote the indicator function recording the class that the nth j-triage patient will transfer to: this patient will transfer to the queue of k-IP patients if $\phi_j(n) = e_k$, or leave the system directly if $\phi_j(n) = 0$. Then $\{\phi_j(n), n \geq 1\}$ is a sequence of i.i.d. random vectors with $\mathbb{P}(\phi_j(n) = e_k) = P_{jk}$, and $\mathbb{P}(\phi_j(n) = 0) = 1 - \sum_{k \in \mathcal{K}} P_{jk}$.

2.2. IP patients

For IP classes, there are no external arrivals. All IP patients are transferred from either triage or IP patients. We use $E_k(t)$ to denote the number of k-IP arrivals till time t. As with triage patients, for each class $k \in \mathcal{K}$, one is given a sequence of i.i.d. random variables $\{v_k(i), i = 1, 2, \dots\}$ and a real number m_k. We assume $\mathbb{E}[v_k(1)] = 1$ and denote $b_k^2 = \text{var}(v_k(1))$. Among k-IP patients, the service time required for the ith patient receiving service is $m_k v_k(i)$. (Unless specified otherwise, we do not require the service discipline within each IP class to be FCFS.) Then, m_k is the mean service time requirement of a k-IP patient. Denote by M the vector with components m_k, $k \in \mathcal{K}$.

For $t \geq 0$ and $k \in \mathcal{K}$, use the renewal process

$$S_k(t) := \max\left\{ n \geq 0 : \sum_{i=1}^{n} m_k v_k(i) \leq t \right\}$$

to represent the number of service completions after the physician has devoted t time units to k-IP patients. Let $\mu_k = 1/m_k$, which is the *service rate* for k-IP patients.

After completing service, an l-IP patient will join the queue of k-IP patients with probability P_{lk}, or exit the system with probability $1 - \sum_{k\in\mathcal{K}} P_{lk}$. Let $P = (P_{lk})_{K\times K}$ denote the IP-to-IP transition matrix and assume that its spectral radius is strictly less than 1. (Equivalently, each IP patient eventually leaves the ED with probability one.) Let $\phi_l(n)$ be the indicator function, showing which class the nth served l-IP patient will transfer to; that is, the nth l-IP patient finishing service will join the queue of k-IP patients if $\phi_l(n) = e_k$, and leave the system if $\phi_k(n) = 0$. Then $\{\phi_l(n), n \geq 1\}$ is a sequence of i.i.d. random vectors with $\mathbb{P}(\phi_l(n) = e_k) = P_{lk}$ and $\mathbb{P}(\phi_l(n) = 0) = 1 - \sum_{k\in\mathcal{K}} P_{lk}$.

Remark 2 *Our main result, Theorem 1, does not require FCFS within each IP class. This is because only queue lengths are involved and the service order within an IP class does not affect the result of the theorem. In contrast, for results involving waiting times or sojourn times, FCFS will either appear in the assumptions (e.g. Propositions 2-4), or as part of the policy (e.g. Subsections 6.1 and 6.2). We shall then assume FCFS explicitly as needed.*

The arrivals of triage classes, services and transitions of triage and IP classes, are all assumed mutually independent. This is not necessary for our proofs, but it simplifies calculations and notation (as in Plambeck et al. (2001)). Indeed, our theory prevails when arrivals of triage classes are correlated with service times of triage and IP classes (Dai and Kurtz (1995)).

Introduce a K-dimensional vector $\Lambda = (\lambda_k)_{k\in\mathcal{K}}$, in which λ_k is interpreted as the *effective arrival rate* for k-IP patients, through the following equations:

$$\Lambda^T = (\Lambda_{\mathcal{J}})^T P_{\mathcal{J}\mathcal{K}} + \Lambda^T P. \tag{1}$$

Then Λ is given by

$$\Lambda^T = (\Lambda_{\mathcal{J}})^T P_{\mathcal{J}\mathcal{K}}(I-P)^{-1}. \tag{2}$$

Define $M^e_{\mathcal{J}} = (m^e_j)_{j\in\mathcal{J}}$ by

$$M^e_{\mathcal{J}} = M_{\mathcal{J}} + P_{\mathcal{J}\mathcal{K}}(I-P)^{-1}M, \tag{3}$$

and call m^e_j the *effective mean service time* of j-triage patients. Now let $M^e = (m^e_k)_{k\in\mathcal{K}}$ be

$$M^e = (I-P)^{-1}M, \tag{4}$$

where m^e_k is called the *effective mean service time* of k-IP patients. Then (3) can be written as

$$M^e_{\mathcal{J}} = M_{\mathcal{J}} + P_{\mathcal{J}\mathcal{K}}M^e. \tag{5}$$

We refer to m^e_j as "effective" because it is the expected *total* service requirement of a j-triage patient, accumulated up to leaving the system (and similarly for m^e_k).

2.3. An infeasible problem

Service goals for triage and IP patients are different:

- **Triage patients facing deadlines:** A j-triage patient must be served within a deadline of d_j time units from its arrival time; that is, a patient arriving to the system at time t must start service before time $t+d_j$. Formally, denote by $\tau_j(t)$ the age of the head-of-the-line j-triage patient at time t. Then a feasible policy must ensure $\tau_j(t) \le d_j$, for $j\in\mathcal{J}$ at all $t\ge 0$.
- **IP patients incurring costs:** Denote by $Q_k(t)$ the number of k-IP patients in the system at time t. Those k-IP patients incur cost at rate $C_k(Q_k(t))$, for some convex functions $C_k, k\in\mathcal{K}$. Consequently, the total cost will be incurred at rate $\sum_{k\in\mathcal{K}} C_k(Q_k(t))$.

A *control policy* is defined as $\pi = \{T_j,\ j\in\mathcal{J};\ T_k,\ k\in\mathcal{K}\}$, in which $T_j(t)$, $j\in\mathcal{J}$, and $T_k(t)$, $k\in\mathcal{K}$, are, respectively, the cumulative time allocated to j-triage patients and k-IP patients during the first t time units. Then our objective is to solve the following optimization problem, for any $T\ge 0$:

$$\begin{aligned} &\min_{\Pi} \quad \int_0^T \sum_{k\in\mathcal{K}} C_k(Q_k(s))ds \\ &\text{s.t.} \quad \tau_j(t)\le d_j, \quad \forall j\in\mathcal{J} \quad \text{and} \quad 0\le t\le T. \end{aligned} \tag{6}$$

Here π is implicit in the formulation, and $\pi\in\Pi$, the set of all candidate control policies (to be introduced later).

The above problem is infeasible as the age processes $\tau_j(\cdot), j\in\mathcal{J}$, are stochastic. Our first task is to assign to (6) a plausible meaning. To this end, we shall consider a converging sequence of systems with the same structure as above, and show that in the limit (conventional heavy traffic), there is a plausible generalization of "feasibility" for the triage constraints.

As for the optimal policy: if the physician always gives priority to triage patients, the queue length of the IP patients will become large and congestion cost high; on the other hand, if priority is always given to IP patients, this reduces cost but the triage patients are likely to violate their deadlines. We thus propose a threshold policy that determines the priority between triage and IP patients and we prove that this policy is asymptotically optimal in the following sense: it is asymptotically feasible, and it stochastically minimizes total congestion cost among all asymptotically feasible policies.

2.4. Our proposed policy

A physician that becomes idle at time t adopts the following guidelines:

- *Triage or IP:* Give priority to triage patients if there exists a $j \in \mathcal{J}$ such that $\tau_j(t) \geq d_j - \epsilon$, where ϵ is small relative to the smallest d_j. (Our theory suggests, and simulations confirm, that ϵ can be chosen one order of magnitude smaller than d_j. For example, with $\min_{j\in\mathcal{J}} d_j = 30$ minutes, one can use ϵ =3 or 4 minutes.)
- *Triage (Shortest-Deadline-First):* Given that a triage patient is to be served, choose the head-of-the-line patient from the class with index

$$j \in \underset{j\in\mathcal{J},\ Q_j(t)\neq 0}{\arg\min} \left[d_j - \tau_j(t)\right].$$

- *IP (Modified generalized cμ-rule):* Given that an IP patient is to be served, choose the head-of-the-line patient from the class with index

$$k \in \underset{k\in\mathcal{K}}{\arg\max} \frac{C'_k(Q_k(t))}{m^e_k}.$$

Here $C'_k(\cdot)$ is the derivative of $C_k(\cdot)$.

2.4.1. The intuition behind our proposed policy. The idea behind our proposed policy is first to maximize service effort for IP patients; given the fixed physician capacity, this is the same as minimizing effort for triage patients subject to adhering to their deadline constraints; then one allocates the physician capacity to IP patients to greedily minimize queueing cost rate. The approach is reasonable since physician capacity is assumed to be close to the arriving workload. As a result, in our critically loaded (heavy traffic) system, there is enough physician capacity for the triage patients to "see" the system in light-traffic, which implies that their needs can be accommodated essentially ad hoc. (From the simulation, most triage patients can meet their deadlines even in a time-varying environment, in which the system can be very crowded; see §EC.9 for further discussion.)

The driver of heavy-traffic dynamics is (total) *workload.* At time t, while conditioning on all queue lengths, its definition is

$$\sum_{j\in\mathcal{J}} m^e_j Q_j(t) + \sum_{k\in\mathcal{K}} m^e_k Q_k(t),$$

which can be interpreted as the average time that a single server would empty the system, assuming there are no new arrivals after time t. The significance of the workload is due to the fact that it is invariant to, and minimized by, any work-conserving policy (Proposition 1 and (EC.18)). Since most j-triage customers at time t arrived to the system during $(t-\tau_j(t),t]$, it must be that $Q_j(t)\approx\lambda_j\tau_j(t)$ and the workload equals approximately

$$\sum_{j\in\mathcal{J}} m_j^e\lambda_j\tau_j(t)+\sum_{k\in\mathcal{K}} m_k^e Q_k(t).$$

The invariance of the potential workload now implies that minimizing $\sum_{k\in\mathcal{K}} m_k^e Q_k(t)$ (which is in concert with minimizing IP congestion costs) is equivalent to maximizing $\sum_{j\in\mathcal{J}} m_j^e\lambda_j\tau_j(t)$.

Triage vs. IP: By the deadline constraints, an upper bound for $\sum_{j\in\mathcal{J}} m_j^e\lambda_j\tau_j(t)$ is $\omega=\sum_{j\in\mathcal{J}}\lambda_j d_j m_j^e$, and our policy should thrive to narrow their gap. It does so by assigning priority to triage patients when their deadlines are getting dangerously close.

Triage selection: The selection rule among triage classes is designed to ensure that their age processes are balanced so that one class of triage patients is about to violate its deadline constraint if and only if all other classes are close to their deadlines as well. Several balancing rules can achieve this goal. The Shortest-Deadline-First rule above is one example. Another example is to ensure $\frac{\tau_j(t)}{d_j}\approx\frac{\tau_{j'}(t)}{d_{j'}}$, for any $j,j'\in\mathcal{J}$, at all times t, which implies that the age of any one triage class reveals those of the others. (Such balancing rules are common in heavy traffic; see the age processes of Plambeck et al. (2001) in conventional heavy traffic, and the QIR controls of Gurvich and Whitt (2009) in the QED regime.) Simulations show that both rules perform well, and the one with the shortest-deadline-first rule is slightly better.

IP selection: After applying the threshold guideline and the triage selection rule, one expects that $\sum_{k\in\mathcal{K}} m_k^e Q_k(t)$ is minimized, thus invariant under any work conserving policy. To minimize cumulative queueing cost, it suffices to minimize cost rates greedily at each time. We are thus led to a convex optimization problem with linear constraints (10). The KKT condition now yields our generalized $c\mu$ rule, as in van Mieghem (1995) but with the μ's replaced by $1/m_k^e$ to account for feedbacks.

The above outline also guides the proofs of our main results, Theorems 1–3. These results are consequences of the parsimonious nature of heavy-traffic dynamics (developed in §3), which is also manifested through some congestion laws that will now be described.

The Snapshot principle: This is again a common feature of heavy traffic (Reiman (1982)) which, as explained on page 187 of Whitt (2002) and adopted here, tells us that during the sojourn time of a patient within the ED, the various queue lengths do not change significantly (or rather negligibly in diffusion scale). In a sense, the ED is temporarily in "steady state", which leads one to expect that some congestion laws in steady state, for example Little's Law, would also prevail temporarily. This snapshot principle then enables predictions of virtual waiting and sojourn times, as we proceed to explain.

Waiting times: When a patient of a particular class completes service, the queue length of that class approximately equals the number of arrivals during this patient's queueing time. (In heavy traffic, service duration is negligible relative to queueing time.) By the snapshot principle, the queue length Q_k and the virtual waiting time ω_k are then related via $Q_k(t) \approx \lambda_k \omega_k(t)$, with λ_k being the arrival rate to class k. On the other hand, $Q_k(t) \approx \lambda_k \tau_k(t)$, as those patients in the queue at time t arrived during the interval $(t - \tau_k(t), t]$. It follows that $\omega_k(t) \approx \tau_k(t)$, which suggests that an estimate of the virtual waiting time (or the waiting duration, predicted at an arrival time) is simply the age of the head-of-the-line patient (see §5.4, which is in the spirit of Ibrahim and Whitt (2009)).

Sojourn times: By the snapshot principle, the ED sojourn time of a patient arriving at time t constitutes the sum, over the patient's route, of all virtual waiting times at time t. Moreover, virtual waiting times remain unchanged during successive visits of the patient to a specific queue. It follows that, asymptotically, the ED sojourn time of a j-triage patient is $\omega_j(t) + \sum_{k\in\mathcal{K}} h_k \omega_k(t)$, given that the patient experiences h_k physician visits as a class k patient. Now replace waiting times on the route by the ages of the head-of-the-line patients at the time of arrival. One concludes that $\tau_j(t) + \sum_{k\in\mathcal{K}} h_k \tau_k(t)$ can serve as a forecast for the ED sojourn time, over a pre-specified route of an arrival at time t (§5.5).

3. Heavy traffic condition

From now on, we consider a sequence of systems, as discussed in Section 2. The sequence will be indexed by $r \uparrow \infty$, and r will be appended as a superscript to denote quantities associated with the rth system. Then, in the rth system, the arrival rate of j-triage class is λ_j^r and the effective arrival rate for k-IP class is λ_k^r. The deadline for j-triage patients is d_j^r, while the cost function C_k for k-IP patients will be specified in the next section. We assume that the service times and transition vectors are invariant with respect to r; hence there will be no superscript for terms relating to service times and transition vectors.

The *traffic intensity* for the rth system is defined to be

$$\rho^r := \sum_{j\in\mathcal{J}} \lambda_j^r m_j + \sum_{k\in\mathcal{K}} \lambda_k^r m_k.$$

By (2) and (3), it can also be represented as

$$\rho^r = \sum_{j\in\mathcal{J}} \lambda_j^r m_j^e.$$

This underscores the meaning of m_j^e being the effective mean service time for j-triage patients.

Assume that the sequence of our systems is in (conventional) *heavy-traffic*, that is,

$$\begin{aligned} &\lambda_j^r \to \lambda_j, \quad j \in \mathcal{J}, \quad \text{and} \\ &r(\rho^r - 1) \to \beta, \quad \text{as} \quad r \to \infty, \end{aligned} \tag{7}$$

for some given $\lambda_j > 0,\ j \in \mathcal{J}$, and $\beta \in \mathbb{R}$. Let $\Lambda = (\lambda_k)_{k\in\mathcal{K}}$ be the vector obtained from (2), with $\Lambda_{\mathcal{J}} = (\lambda_j)_{j\in\mathcal{J}}$ in (7).

Under condition (7), the queue lengths are expected to be $O(r)$, and similarly the ages of head-of-the-line triage patients. Hence, for each $j \in \mathcal{J}$, we assume the following convergence for the deadline of j-triage patients:

$$\frac{d_j^r}{r} \to \widehat{d}_j, \quad \text{as} \quad r \to \infty,$$

where $\widehat{d}_j > 0,\ j \in \mathcal{J}$, are given constants. We assume that the indices of triage classes are ordered such that $\widehat{d}_j$ is increasingly in j.

Denote by $Q_j^r(t)$ and $Q_k^r(t)$ the number of j-triage and k-IP patients in the rth system at time t, respectively. We assume that the following initial condition holds:

Assumption 1 *As* $r \to \infty$,

$$r^{-1}Q_j^r(0) \Rightarrow 0, \quad j \in \mathcal{J},$$
$$r^{-1}Q_k^r(0) \Rightarrow 0, \quad k \in \mathcal{K}.$$

4. Asymptotic compliance and optimality

A control policy $\pi^r = \{T_j^r,\ j \in \mathcal{J};\ T_k^r,\ k \in \mathcal{K}\}$ determines the age processes of the head-of-the-line patients in the rth system, $\tau^r(\cdot) = \{\tau_j^r(\cdot),\ j \in \mathcal{J}\}$. Introduce the diffusion-scaled age processes through

$$\widehat{\tau}_j^r(t) = r^{-1}\tau_j^r(r^2 t), \quad j \in \mathcal{J}.$$

We now consider policies that are asymptotically compliant, which is a generalization of "feasibility" for the optimization problem (6).

Definition 1 *A family of policies* $\{\pi^r\}$ *is said to be* asymptotically compliant *if, for any fixed* $T \geq 0$,

$$\sup_{0 \leq t \leq T} \left[\widehat{\tau}_j^r(t) - \widehat{d}_j\right]^+ \Rightarrow 0, \quad \textit{as} \quad r \to \infty, \quad \textit{for all} \quad j \in \mathcal{J}.$$

Introduce the diffusion-scaled number of k-IP patients in the system by

$$\widehat{Q}_k^r(t) = r^{-1}Q_k^r(r^2 t), \quad k \in \mathcal{K}. \tag{8}$$

We assume that, at time t (in diffusion scaling), k-IP patients incur a queueing cost at rate $C_k(\widehat{Q}_k^r(t))$, for some function C_k. (Concrete assumptions on C_k will be provided in Assumption 2.) Then the cumulative queueing cost is

$$\mathcal{U}^r(t) := \int_0^t \sum_{k \in \mathcal{K}} C_k\left(\widehat{Q}_k^r(s)\right) ds. \tag{9}$$

Our heavy-traffic adaptation of problem (6) is to stochastically minimize $\mathcal{U}^r(t)$, for $t > 0$, over all asymptotically compliant families of policies. Formally:

Definition 2 *A family of control policies* $\{\pi_*^r\}$ *is said to be* asymptotically optimal *if*

1. *it is asymptotically compliant and*
2. *for every* $t > 0$ *and every* $x > 0$,

$$\limsup_{r\to\infty} \mathbb{P}\left\{\mathcal{U}_*^r(t) > x\right\} \le \liminf_{r\to\infty} \mathbb{P}\left\{\mathcal{U}^r(t) > x\right\};$$

here $\{\mathcal{U}_*^r\}$ *is the family of cumulative queueing costs defined through* (9) *under the family of control policies* $\{\pi_*^r\}$, *and* $\{\mathcal{U}^r\}$ *is the sequence of queueing costs corresponding to any other asymptotically compliant family of policies* $\{\pi^r\}$.

5. Main results

5.1. Cost functions and an optimization problem

For any given $a \ge 0$, consider the following optimization problem over $x = (x_k)_{k\in\mathcal{K}}$:

$$\begin{aligned} \min_x \quad & \sum_{k\in\mathcal{K}} C_k(x_k) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{K}} m_k^e x_k = a, \\ & x \ge 0. \end{aligned} \tag{10}$$

We assume that the cost functions C_k, $k \in \mathcal{K}$, satisfy conditions that are analogous to van Mieghem (1995). Specifically:

Assumption 2 (Cost regularity) *The nondecreasing cost functions* $\{C_k,\ k \in \mathcal{K}\}$ *are strictly convex, continuously differentiable. In addition, for all* $a > 0$, *there is an optimal solution* x^* *to the optimization problem* (10) *such that* $x_k^* > 0$, $k \in \mathcal{K}$.

By this assumption and the KKT condition, a sufficient condition for a nonnegative vector $x^* = (x_k^*)_{k\in\mathcal{K}}$ to be optimal is the existence of $\alpha_0 \in \mathbb{R}$ such that

$$\begin{aligned} C_k'(x_k^*) - \alpha_0 m_k^e = 0, \\ \sum_{k\in\mathcal{K}} m_k^e x_k^* = a. \end{aligned}$$

This optimal vector x^* satisfies $C_l'(x_l^*)/m_l^e = C_k'(x_k^*)/m_k^e$, for all $l, k \in \mathcal{K}$. Then the proof of the following is elementary:

Lemma 5.1 *Denote the optimal solution to* (10) *by*

$$x^* = \Delta_{\mathcal{K}}(a).$$

Then the function $\Delta_{\mathcal{K}}(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+^K$ *is well defined, and* $\Delta_k(a)$ *is nondecreasing in* a, *for each* $k \in \mathcal{K}$.

The mapping $\Delta_{\mathcal{K}}$ is part of the lifting mapping used in our state-space collapse result; see Theorem 3.

5.2. An asymptotically optimal policy

We propose the following sequence of scheduling policies, which we denote by $\{\pi_*^r\}$:

- Fix a sequence of ϵ^r such that $\frac{\epsilon^r}{r} \to 0$, as $r \to \infty$.
- When becoming idle, the physician deploys a threshold rule to determine which type of patient classes to serve next—a triage-type patient or an IP-type patient.

  —If there exists a $j \in \mathcal{J}$ such that $\tau_j^r(t) \geq d_j^r - \epsilon^r$, priority is given to triage-type patients;

  —Otherwise, priority is given to IP-type patients.
- If triage patients are chosen to be served at time t, the physician chooses the head-of-the-line patient from the class with index

$$j \in \underset{j\in\mathcal{J},\ Q_j^r(t)\neq 0}{\arg\min}\ [d_j^r - \tau_j^r(t)]. \tag{11}$$

- If IP patients are chosen to be served at time t, the physician chooses the head-of-the-line patient from the class with index

$$k \in \underset{k\in\mathcal{K}}{\arg\max} \frac{C_k'(\widehat{Q}_k^r(t))}{m_k^e}. \tag{12}$$

Our main result is the following theorem, which we prove in §EC.5.

Theorem 1 (Asymptotic Optimality) *The family of control policies* $\{\pi_*^r\}$ *is asymptotically optimal.*

Remark 3 *Though in the current work we assume that the cost functions are strictly convex (Assumption 2), our analysis still applies to linear cost functions. In that case,* (10) *becomes a linear optimization problem. Then the optimal policy can be modified to one using a* $c\mu$*-type rule, which is a static priority rule that gives higher priority to the class with larger* $\frac{c_k}{m_k^e}$*; here* c_k *is the cost rate parameter.*

5.3. A roadmap to prove Theorem 1

The proof takes two steps. First in Theorem 2 we prove that under any asymptotically "feasible" policy, queueing costs can be stochastically bounded from below. Then we show that, under the proposed policy, the lower bound can be achieved. This entails the "state-space-collapse" result in Theorem 3 which, together with Step 1, establishes the asymptotic optimality of our proposed policies.

For $j \in \mathcal{J}$ and $k \in \mathcal{K}$, introduce $K \times K$ matrices $\Gamma^j = (\Gamma_{ll'}^j)$ and $\Gamma^k = (\Gamma_{ll'}^k)$ through

$$\Gamma_{ll'}^j = \begin{cases} P_{jl}(1-P_{jl'}), & \text{if } l = l' \\ -P_{jl}P_{jl'}, & \text{if } l \neq l' \end{cases} \quad \text{and} \quad \Gamma_{ll'}^k = \begin{cases} P_{kl}(1-P_{kl'}), & \text{if } l = l' \\ -P_{kl}P_{kl'}, & \text{if } l \neq l' \end{cases}.$$

Define $\widehat{Q}_w = \Phi(\widehat{X})$; here Φ is the 1-dimensional Skorohod mapping (Theorem 6.1 in Chen and Yao (2001)), and $\widehat{X}$ is a Brownian motion with drift rate β and variance

$$\sum_{j\in\mathcal{J}}(m_j^e)^2\lambda_j a_j^2 + \sum_{j\in\mathcal{J}}\left(\sum_{k\in\mathcal{K}} m_k^e P_{jk} - m_j^e\right)^2 \lambda_j b_j^2 + \sum_{k\in\mathcal{K}}\left(\sum_{l\in\mathcal{K}} P_{kl} m_l^e - m_k^e\right)^2 \lambda_k b_k^2$$
$$+ \sum_{j\in\mathcal{J}} \lambda_j (M^e)^T \Gamma^j M^e + \sum_{k\in\mathcal{K}} \lambda_k (M^e)^T \Gamma^k M^e.$$

Finally let $\widehat{\omega} = \sum_{j\in\mathcal{J}} \lambda_j \widehat{d}_j m_j^e$.

Theorem 2 (Lower Bound) *Fix any asymptotically compliant family of policies, with the corresponding cumulative costs $\mathcal{U}^r$ defined in* (9). *Then for any $t, x > 0$,*

$$\liminf_{r\to\infty} \mathbb{P}\{\mathcal{U}^r(t) > x\} \geq \mathbb{P}\left\{\int_0^t \sum_{k\in\mathcal{K}} C_k\left(\Delta_k\left((\widehat{Q}_w(s) - \widehat{\omega})^+\right)\right) ds > x\right\}.$$

This theorem is proved in §EC.2.

In proving Theorem 1, we show that the proposed policy tames the system in the sense that the weighted queue length converges (Proposition 1), and there is state-space collapse for the queue length processes (Theorem 3).

Proposition 1 in fact holds under any family of work-conserving policies. To state it, recall $\widehat{Q}_k^r$ from (8) and define similarly the diffusion-scaled queue length processes for triage classes: $\widehat{Q}_j^r(t) = r^{-1} Q_j^r(r^2 t)$, $j \in \mathcal{J}$. The diffusion-scaled weighted queue length processes is given by

$$\widehat{Q}_w^r(t) = \sum_{j\in\mathcal{J}} m_j^e \widehat{Q}_j^r(t) + \sum_{k\in\mathcal{K}} m_k^e \widehat{Q}_k^r(t). \tag{13}$$

Proposition 1 (Invariance principle for work-conserving policies) *Under any family of work-conserving policies,*

$$\widehat{Q}_w^r \Rightarrow \widehat{Q}_w, \quad as \quad r \to \infty. \tag{14}$$

This proposition is proved in §EC.3.

For any $a \in \mathbb{R}_+$, let $\Delta_{\mathcal{J}}(a) = (\Delta_j(a))_{j\in\mathcal{J}}$ be defined as follows ($\widehat{d}_0 = 0$): if $\sum_{j\in\mathcal{J}} \lambda_j m_j^e(\widehat{d}_j - \widehat{d}_{j'})^+ \leq a < \sum_{j\in\mathcal{J}} \lambda_j m_j^e(\widehat{d}_j - \widehat{d}_{j'-1})^+$, then

$$\Delta_{j_1}(a) = \begin{cases} \lambda_{j_1}\left(\widehat{d}_{j_1} - \widehat{d}_{j'} + \left(a - \sum_{j\in\mathcal{J}} \lambda_j m_j^e(\widehat{d}_j - \widehat{d}_{j'})^+\right)/(\sum_{j\geq j'} \lambda_j m_j^e)\right), & \text{for } j_1 \geq j', \\ 0, & \text{for } j_1 < j'. \end{cases} \tag{15}$$

The function pair $(\Delta_{\mathcal{J}},\ \Delta_{\mathcal{K}})$ is the lifting mapping in the state-space collapse result. Let $\widehat{Q}^r = (\widehat{Q}_j^r,\ j \in \mathcal{J};\ \widehat{Q}_k^r,\ k \in \mathcal{K})$ and recall that $\widehat{\omega} = \sum_{j\in\mathcal{J}} \lambda_j \widehat{d}_j m_j^e$.

Theorem 3 (State-Space Collapse) *Under the family of control policies $\{\pi_*^r\}$, $\widehat{Q}^r \Rightarrow \widehat{Q}$, where $\widehat{Q} = (\widehat{Q}_j,\ j \in \mathcal{J};\ \widehat{Q}_k,\ k \in \mathcal{K})$ is specified by*

$$\begin{aligned} \widehat{Q}_j(t) &= \Delta_j\left(\min\left(\widehat{Q}_w(t), \widehat{\omega}\right)\right), \quad j \in \mathcal{J}, \\ \widehat{Q}_k(t) &= \Delta_k\left((\widehat{Q}_w(t) - \widehat{\omega})^+\right), \quad k \in \mathcal{K}. \end{aligned}$$

This theorem is proved in §EC.4.

Remark 4 *One can verify that $\Delta_j(\cdot)$ is increasing for all $j \in \mathcal{J}$, and $\Delta_j(\widehat{\omega}) = \lambda_j \widehat{d}_j$. As a result, $\widehat{Q}_j(t) \le \lambda_j \widehat{d}_j$, which can be translated into "asymptotic compliance" of the family of control policies $\{\pi_*^r\}$; on the other hand, those limits $\widehat{Q}_k$ appear in the lower bound of Theorem 2, which shows that the family of control policies $\{\pi_*^r\}$ achieves the lower bound asymptotically.*

5.4. Virtual waiting times

In this and the next subsection, we analyze our family of control policies $\{\pi_*^r\}$. For its complete characterization, assume that the service order within each IP class is FCFS.

Define the virtual waiting time of a patient class at time t as the time that a virtual patient of this class, arriving at t, would have to wait till completing the current phase of service. This definition is notationally convenient in our case, but it is slightly different from the traditional one, which is the waiting time till service starts. As the service time is negligible in heavy traffic scaling, these two definitions yield the same result. Denote by $\omega_j^r(t)$ and $\omega_k^r(t)$ the virtual waiting times for j-triage class and k-IP class respectively, and define the diffusion-scaled virtual waiting time processes by

$$\widehat{\omega}_j^r(t) = r^{-1}\omega_j^r(r^2 t), \quad j \in \mathcal{J}, \quad \text{and} \quad \widehat{\omega}_k^r(t) = r^{-1}\omega_k^r(r^2 t), \quad k \in \mathcal{K}. \tag{16}$$

Proposition 2 (Asymptotic Sample-Path Little's Law) *Consider the family of control policies $\{\pi_*^r\}$, with FCFS service discipline among each IP patient class. As $r \to \infty$, we have*

$$\begin{aligned} \widehat{\omega}_j^r - \widehat{Q}_j^r/\lambda_j^r &\Rightarrow 0, \quad j \in \mathcal{J}, \\ \widehat{\omega}_k^r - \widehat{Q}_k^r/\lambda_k^r &\Rightarrow 0, \quad k \in \mathcal{K}. \end{aligned} \tag{17}$$

This proposition is proved in §EC.7.1.

Remark 5 *From the convergence of $\widehat{Q}^r$ in Theorem 3, one deduces the convergence of the vector of virtual waiting times under the family of control policies $\{\pi_*^r\}$.*

Recall that $\tau_j^r(t)$ is defined as the age of the head-of-the-line j-triage patient in the rth system. Similarly, let $\tau_k^r(t)$ be the age of the head-of-the-line k-IP patient in the rth system, with its diffusion scaling $\widehat{\tau}_k^r(t) = r^{-1}\tau_k^r(r^2 t)$, $k \in \mathcal{K}$. Our next proposition establishes a connection between virtual waiting time and age. Thus patients, arriving at a queue, can estimate their waiting time to be the age of the head-of-the-line patient at that queue. This kind of result is often referred to as a *snapshot principle:* during the stay of a patient in the system, the state of the system remains essentially unchanged.

Proposition 3 (Snapshot Principle—Virtual Waiting Time and Age) *Consider the family of control policies $\{\pi_*^r\}$, with FCFS among each IP patient class. As $r \to \infty$, we then have*

$$\widehat{\omega}_j^r - \widehat{\tau}_j^r \Rightarrow 0, \quad j \in \mathcal{J},$$
$$\widehat{\omega}_k^r - \widehat{\tau}_k^r \Rightarrow 0, \quad k \in \mathcal{K}.$$

This proposition is proved in §EC.7.2.

5.5. Sojourn times

We now consider sojourn times associated with specific routes through the system, as in Reiman (1984). To this end, one associates a route vector $h \in \mathbb{Z}_+^K$ with each patient going through the system, where h_k denotes the number of times that the patient visits the physician as a k-IP patient before leaving the system. A vector $h \in \mathbb{Z}_+^K$ is called *j-feasible* if it is possible (there is a positive probability) that a patient entering the system as a j-triage patient has a route vector h. Denote by $W_{jh}^r(t)$ the *sojourn time* of the first j-triage patient that arrives after time t with route vector h. This gives rise to the diffusion-scaled processes

$$\widehat{W}_{jh}^r(t) = r^{-1} W_{jh}^r\left(r^2 t\right), \quad j \in \mathcal{J}.$$

Proposition 4 (Snapshot Principle—Sojourn Time and Queue Lengths) *Under the family of control policies $\{\pi_*^r\}$, with FCFS among each IP patient class, if a route vector h is j-feasible, then as $r \to \infty$,*

$$\widehat{W}_{jh}^r - \frac{\widehat{Q}_j^r}{\lambda_j^r} - \sum_{k\in\mathcal{K}} \frac{h_k}{\lambda_k^r} \widehat{Q}_k^r \Rightarrow 0, \quad j \in \mathcal{J}.$$

This proposition is proved in §EC.7.3.

Remark 6 *From Theorem 3, as $r \to \infty$, we have*

$$\frac{\widehat{Q}_j^r}{\lambda_j} + \sum_{k\in\mathcal{K}} \frac{h_k}{\lambda_k} \widehat{Q}_k^r \Rightarrow \Delta_j\left(\min\left(\widehat{Q}_w, \widehat{\omega}\right)\right) + \sum_{k\in\mathcal{K}} \frac{h_k}{\lambda_k} \Delta_k\left((\widehat{Q}_w - \widehat{\omega})^+\right).$$

Then Proposition 4 yields an estimator for the distribution of $\widehat{W}_{jh}^r(\cdot)$:

$$\Delta_j\left(\min\left(\widehat{Q}_w(\cdot), \widehat{\omega}\right)\right) + \sum_{k\in\mathcal{K}} \frac{h_k}{\lambda_k} \Delta_k\left((\widehat{Q}_w(\cdot) - \widehat{\omega})^+\right).$$

The following is a direct corollary of Propositions 2, 3 and 4.

Corollary 1 (Snapshot Principle—Sojourn Time and Ages) *Under the family of control policies $\{\pi_*^r\}$, with FCFS among each IP patient class, if a route vector h is j-feasible, then as $r \to \infty$,*

$$\widehat{W}_{jh}^r - \widehat{\tau}_j^r - \sum_{k\in\mathcal{K}} h_k \widehat{\tau}_k^r \Rightarrow 0, \quad j \in \mathcal{J}.$$

Remark 7 *This corollary suggests that, upon arrival, patients can estimate their sojourn time by using the current age of the head-of-the-line patients on their routes (assuming the route is known apriori). As in Reiman (1984), the diffusion limit does not depend on the specific order in which physician-queues are visited.*

6. Extensions and further discussion

6.1. Waiting costs

Consider now waiting costs, instead of queueing costs. To this end, assume that the service discipline among each IP class is FCFS. This is without loss of generality, since every policy has another policy that is at least as good and which serves FCFS within each IP class (van Mieghem (1995)). Recall that $\omega_k^r(t)$ is the virtual waiting time of a k-IP patient at time t, and its diffusion scaling $\widehat{\omega}_k^r(t)$ is defined in (16). We seek to stochastically minimize the cost

$$\widetilde{\mathcal{U}}^r(t) := \sum_{k\in\mathcal{K}} \int_0^t C_k\left(\widehat{\omega}_k^r(s)\right) d\bar{\bar{E}}_k^r(s), \tag{18}$$

among all asymptotically compliant families of control policies. Here $\bar{\bar{E}}_k^r(t) = r^{-2}E_k^r(r^2 t)$.

We now slightly modify the control policy $\{\pi_*^r\}$ in Section 5. The first step, using a threshold rule to determine priority between triage classes vs. IP classes, and the step using (11) to determine priorities among triage patients, do not change. The service principle among each class is FCFS. The step to determine priority among IP classes changes as follows:

- If the IP classes are chosen to be served at time t, the physician chooses the head-of-the-line patient from the class with index

$$k \in \arg\max_{k\in\mathcal{K}} \frac{C_k'\left(\frac{\widehat{Q}_k^r(t)}{\lambda_k^r}\right)}{m_k^e}.$$

Denote this family of modified policies by $\{\widetilde{\pi}_*^r\}$.

Proposition 5 (Waiting Time Cost) *The family of control policies $\{\widetilde{\pi}_*^r\}$ is asymptotically compliant. It is also asymptotically optimal among all asymptotically compliant families of work-conserving control policies, in the sense that for any fixed $t > 0$ and $x > 0$,*

$$\limsup_{r\to\infty} \mathbb{P}\left\{\widetilde{\mathcal{U}}_*^r(t) > x\right\} \le \liminf_{r\to\infty} \mathbb{P}\left\{\widetilde{\mathcal{U}}^r(t) > x\right\},$$

where $\{\widetilde{\mathcal{U}}_*^r\}$ is the family of cumulative cost, defined through (18) under the family of control policies $\{\widetilde{\pi}_*^r\}$, and $\{\widetilde{\mathcal{U}}^r\}$ is the corresponding cost under any other asymptotically compliant family of work-conserving policies $\{\pi^r\}$.

The outline of the proof can be found in §EC.7.4.

6.2. An alternative criterion: IP sojourn time

In this subsection, we consider an alternative model. The structure is identical to Figure 1, except that congestion cost is associated with each patient's *sojourn time* in the IP stage (as opposed to individual queueing and waiting costs previously). We now add the assumption that the routing matrix P is upper-triangular. Then, by enlarging the number of IP classes, the routing behavior in the IP stage can be assumed to be deterministic; that is, the routing is not random now. With the upper-triangular assumption, the number of routing vectors is finite. Thus, without loss of generality, we assume that each patient follows a deterministic routing vector and there is a finite number of routing vectors. We use $\mathcal{C}_0$ to denote the set of *starting IP classes* of routes. For $k \in \mathcal{C}_0$, let $\mathcal{C}_k$ denote all the classes on the route that starts at k, and call any class in $\bigcup_{k\in\mathcal{C}_0} \mathcal{C}_k \backslash \{k\}$ a *subsequent class.* If a patient with starting class k waits $\omega_{k'}$, as a k'-IP patient ($k' \in \mathcal{C}_k$), then the sojourn time of this patient is $\sum_{k'\in\mathcal{C}_k} \omega_{k'}$. Our problem is to stochastically minimize the cost

$$\widetilde{\mathcal{S}}^r(t) = \sum_{k\in\mathcal{C}_0} \int_0^t C_k \left(\sum_{k'\in\mathcal{C}_k} \widehat{\omega}^r_{k'}(s) \right) d\bar{\bar{E}}^r_k(s), \tag{19}$$

among all asymptotically compliant families of control policies, for all $t > 0$.

We propose the following routing policy: The first step, using a threshold rule to determine priority between triage classes and IP classes, and the step using (11) to determine priorities among triage patients, do not change. The service principle among each class is FCFS. The step determining the priority among IP classes changes as follows:

- Assign priority to all *subsequent* classes, while allocating the remaining physician capacity to all starting classes by choosing the head-of-the-line patient from the class with index

$$k \in \arg\max_{k\in\mathcal{C}_0} \frac{C'_k \left(\widehat{Q}^r_k(t)/\lambda^r_k \right)}{m^e_k}. \tag{20}$$

Here $\widehat{Q}^r_k$ is the diffusion-scaled queue length of the starting classes $k \in \mathcal{C}_0$, and m^e_k is the corresponding effective mean service time.
We denote this family of policies by $\{\widetilde{\pi}^r_{**}\}$.

Proposition 6 (Sojourn Time Cost) *The family of control policies $\{\widetilde{\pi}^r_{**}\}$ is asymptotically compliant. It is asymptotically optimal among all asymptotically compliant families of control policies in the sense that, for any fixed $t > 0$ and $x > 0$,*

$$\limsup_{r\to\infty} \mathbb{P}\left\{ \widetilde{\mathcal{S}}^r_{**}(t) > x \right\} \le \limsup_{r\to\infty} \mathbb{P}\left\{ \widetilde{\mathcal{S}}^r(t) > x \right\};$$

here $\{\widetilde{\mathcal{S}}^r_{**}\}$ is the family of cumulative cost defined through (19) *under the family of control policies $\{\widetilde{\pi}^r_{**}\}$, and $\{\widetilde{\mathcal{S}}^r\}$ is the corresponding cost under any other asymptotically compliant family of policies $\{\pi^r\}$.*

The outline of the proof can be found in §EC.7.5.

Giving priority to all subsequent classes when serving IP classes is consistent with the observation in Saghafian et al. (2012), where it is referred to as 'Prioritize Old' policy.

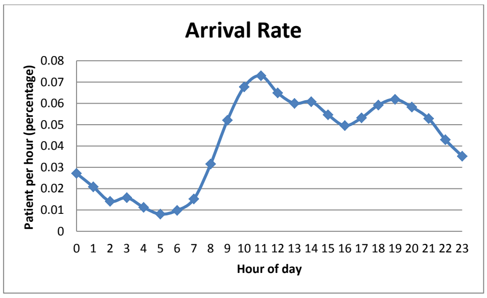
7. Numerical experiments

We use simulation to assess the relevance of our theory and the performance of our proposed policy (§5.2). We simulate several systems. One system has stationary arrival rates plus only the features analyzed in our paper. The other systems have time-varying arrival rates, delays between successive visits to physicians, finite ED capacity, multiple servers and abandonment (LWBS+LAMA), separately or jointly. These are features that were not assumed in our model. As observed in the simulations, our proposed policy performs very well, and it outperforms commonly-used alternatives in *all* systems. In §7.1–7.3, we present the parameters and simulation results for the stationary model. We check the robustness of the proposed policy in §7.4 and §EC.9: §7.4 includes a model with several features, and §EC.9 provides more detailed simulations.

7.1. Parameters

The empirical characteristics of our models are taken from 4 sources: the ED data at the Technion SEELab (see SEELab Link), Carmeli (2012), Yom-Tov and Mandelbaum (2014) and Armony et al. (2013). In our ED, there are 5 triage classes. We do not consider triage classes 1 and 2; they correspond to patients in critical condition and, hence, are treated separately and with the highest priority. We thus focus on triage classes 3, 4 and 5, which we index by 1, 2, 3. This means that our j-triage patients are triage class $j+2$ patients in practice, $j = 1, 2, 3$. The deadlines for those three classes are 30, 60 and 120 minutes. In the ED we use, on average, there were 302 patients (from all 5 triage classes) arriving at the ED each weekday in January 2004. Figure 2 depicts the shape (percentage) of daily arrivals per hour.

Figure 2 Hourly % of arrivals to an Israeli ED (January 2004)



The arrival rates will be approximated here as stationary from 9:00 to 22:00 (Armony et al. (2013)). During these hours, the average arrival rates of all patients to the ED is 17.82 per hour. As Triage 1 and 2 patients are excluded, we assume that the average arrival rate of Triage 3, 4 and 5 patients (hence 1-triage, 2-triage and 3-triage patients) is 14 per hour, that is, 14/60 per minute. The relative weights of those three triage classes are 10%, 40% and 50% (see Carmeli (2012)). The overall arrival process is taken to be Poisson, with Markovian split into the 3 triage classes.

IP patients may be classified into several classes, according to several factors such as treatment type, emergency level and age; see Carmeli (2012). In our ED, patients experience 1-5 IP phases (doctor visits): 28% go through 1 phase only and are then released, 30% have 2 phases, 28% - 3, 11% - 4, and 3% go through 5 IP phases. For the sake of illustration, we classify the IP patients into classes according to their number of IP visits, where we combine phases 3, 4 and 5 into a single phase. As a result, we have 3 IP classes, with the following transition matrices (that is, after phase 1, $100-28\% = 0.72$ of the patients move on to class 2; after phase 2, $(72\%-30\%)/72\% = 0.58$ of the patients switch to class 3):

$$P_{\mathcal{J}\mathcal{K}} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} 0 & 0.72 & 0 \\ 0 & 0 & 0.58 \\ 0 & 0 & 0 \end{pmatrix}.$$

Average service times depend on the class of the patients. Generally, it varies from 5.8 minutes (or 4.8 minutes if we include the trauma class) to 6.7 minutes (see Table 2 in Yom-Tov and Mandelbaum (2014)). Service times are assumed to follow exponential distributions with mean 6.5 minutes. (This is without loss since only expectations determine our policy.) There are typically 4–6 physicians working simultaneously in the ED, and sometimes this number reaches 8 physicians. In our model we assume that there are 5 physicians. This induces a service time of a corresponding "super" physician (single-server), which is taken to be $6.5/5 = 1.3$ minutes. It follows that the traffic intensity is 0.9517.

Cost functions are generally difficult to estimate. Discussions with the director of our partner ED suggested quadratic cost functions: $C_k(x) = c_k x^2$ (see Carmeli (2012)). We assume that the parameters c_k are 1, 1.5, 2 for the 3 IP classes respectively.

7.2. Guideline on choosing ϵ in our proposed policy

Our recommendation for the threshold part is as follows: Assign priority to triage classes if, for some $j \in \mathcal{J}$, $\tau_j(t) \geq d_j - \epsilon$. Here ϵ is one order of magnitude smaller than the deadlines, and its specific value depends on the *target* percentage of patients who violate the deadlines. From our simulation experiments, when the minimum deadline is about 20 times longer than the single-server's service time, ϵ is to be chosen 2 or 3 times the service time so that less than 5% of the patients violate their deadlines. In our stationary model, the minimum deadline is 30 minutes. The average service time of the "super" single-server is 1.3 minutes. This gives rise to $\epsilon = 3 \approx 2.3 \times 1.3$ minutes. Note that $\epsilon = 3$ is about 1/2 of 6.5 minutes, the real physician service time. If the ratio between the deadlines and the service time is larger, we can choose an even larger ϵ. In systems with time-varying arrival rates (when there is a long period during which the system is overloaded), we propose to use a different ϵ for different classes, which may be somewhat larger than in the stationary case. (For example, in §EC.9.3, where there is a long period with traffic intensity that exceeds 1.2, we use $\epsilon = 4, 6, 8$ for the three triage classes with deadlines 30, 60, 120, respectively.) In systems with finite ED capacity and patient abandonment, ϵ can be chosen slightly smaller (§7.4).

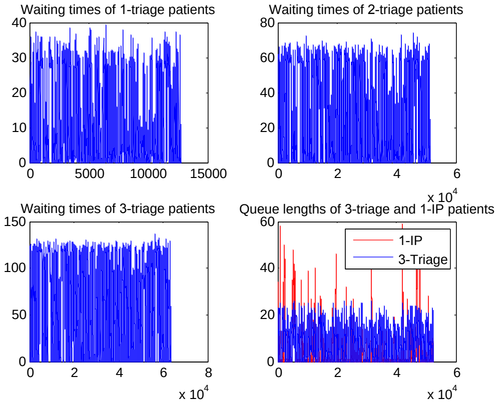
7.3. Simulation outcomes

We first simulate the ED under our recommended policy (denoted by TGcμ) to gain insight into system performance. Then we compare this policy to three alternatives: global FCFS (denoted FCFS), IP-patients-First (denoted IPF) and Triage-patients-First (denoted TrF). See Appendix (§EC.9.1) for more detailed descriptions of these three policies.

We ran the system over 380 days (with time unit of 1 minute, the duration is $60 * 24 * 380 = 547,200$ minutes). The initial period of 15 days is a warm-up period and hence excluded from our output analysis. In particular, we excluded the initial triage patients: $14 * 24 * 15 * (0.1, 0.4, 0.5)$, who are roughly those arriving during the first 15 days. For each policy, we simulated 160 sample paths. Here we present the results for the stationary model with no other features. (Other models, which may have time-varying arrivals, delays between physician visits, and other features, separately or jointly, are described in §EC.9 and §7.4.)

Figure 3 displays a typical sample path under our proposed policy. (Corresponding histograms will be shown and explained momentarily.) We plot the waiting times for all three triage classes (in the figures except the low-right one: X-axes count the number of patients and Y-axes represent waiting times in minutes), as well as the queue lengths of 3-triage patients and 1-IP patients (in the low-right figure: X-axis represents time (in minutes) and Y-axis represents queue sizes). The reason we chose these two classes is because 3-triage has the longest queue length among the 3 triage classes (as expected) and 1-IP has the longest among the 3 IP classes. Moreover, our simulated sample paths exhibit state-space collapse, hence the evolution of these two classes determines that of the others.

Figure 3 A typical sample path of the system under our proposed policy (waiting time in minutes)



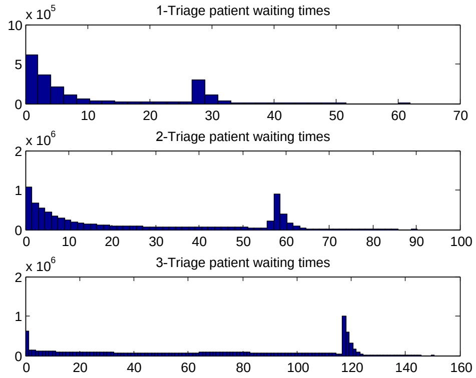
From Figure 3 we make the following observations:

1. Most triage patients meet their deadlines: The ages of most triage patients are bounded by their corresponding deadlines. Even if some violate the deadlines, these violations are very small: Over the 160 sample paths, the fractions of violations are 4.61%, 4.57%, 4.57%, for 1-triage, 2-triage and 3-triage patients respectively; the fractions of triage deadline violations by more than 10% of their corresponding deadlines are negligible (less than 1%).

2. The queue lengths of IP classes are away from 0 only when the triage patients are close to violating their deadlines. Alternatively, if triage deadlines are not tight then IP queues are close to 0, which is expected from our theory.

Figure 4 depicts three histograms of waiting times for the three triage classes: the top one is for 1-triage patients, the middle for 2-triage and the third for 3-triage (X-axes represent time in minutes). Those histograms include data of all 160 sample paths, for a visualization of how the triage patients violate their deadlines.

Figure 4 Histograms of patient waiting times under our proposed policy



We see ample 1-triage patients whose waiting times are short, which is what we seek to achieve since 1-triage patients are likely to be in a more serious medical state. This is due to the fact that the deadline of 1-triage patients is much shorter than the other two deadlines: 1-triage patients are more likely to enjoy high priority because their $d_1 - \tau_1(t)$ is conceivably the shortest (recall that $d_1 = 30, d_2 = 60$ and $d_3 = 120$).

We also compare our policy with the above-mentioned three policies. We summarize our findings in the following table, where $P_j, j = 1, 2, 3,$ is the fraction of j-triage patients who violate their corresponding deadline. "Cost Rate" is IP-cost per time-unit, averaged over samples. (The numbers in parenthesis are half-length 95% confidence intervals.)

Our proposed policy (TGcμ) outperforms the global FCFS policy. The cost rate for IP-patients-First (IPF) is small, but a large fraction of triage patients violate the deadlines. Triage-patients-First (TrF) has patients that satisfy the deadline constraints, while its cost rate exceeds 3 times that of TGcμ. In summary, our proposed policy (TGcμ) clearly dominates the other three alternatives.

Table 1 Comparison of the four policies

Policy	P_1	P_2	P_3	Cost Rate
TGcμ	4.61% (0.10%)	4.57% (0.09%)	4.57% (0.09%)	125.21 (10.36)
FCFS	31.27% (0.49%)	10.16% (0.38%)	1.15% (0.16%)	187.46 (7.20)
IPF	21.26% (0.48%)	21.28% (0.48%)	21.26% (0.48%)	0.88 (0.07)
TrF	0.00% (0.00%)	0.00% (0.00%)	0.00% (0.00%)	523.69 (20.11)

7.4. Robustness of the proposed policy

As will be elaborated on in §8, there are several ED features that have been theoretically left out in our model. We now analyze the robustness of the proposed policy with respect to such features, focusing on the ones that are most significant. To this end, we simulate a queueing system incorporating these features *jointly*, which can be viewed as a proxy for a real ED. More simulation results can be found in the Appendix (§EC.9).

Customers arrive according to a time-varying arrival rate, as in Figure 2. The average total arrivals of 1-triage, 2-triage and 3-triage patients per day is $14 * 24 = 336$. (Notice that this is even more than the actual arrival rate including triage 1 and 2 patients.) We further assume constant arrival rates *per hour*, which are then given by 9.13, 7.00, 4.72, 5.31, 3.77, 2.71, 3.29, 5.09, 10.61, 17.51, 22.76, 24.51, 21.81, 20.16, 20.43, 18.36, 16.66, 17.88, 19.90, 20.80, 19.58, 17.77, 14.43, 11.83. We assume that there are 5 physicians and the mean service time is 6.5 minutes. Then the traffic intensity varies from 0.1839 to 1.6663. We modify the TGcμ policy to the following: give priority to triage classes if $\tau_1(t) > d_1 - 4$, or $\tau_2(t) > d_2 - 4$, or $\tau_3(t) > d_3 - 5$. The ED capacity is fixed as 70. This number is chosen as a compromise between Armony et al. (2013) and Bolandifar et al. (2014). The ED in Armony et al. (2013) is an Israeli one with 40 beds, and patients can be served after waiting on chairs; the ED in Bolandifar et al. (2014) is a US one with 70 beds. The length of delays between transfers follows an exponential distribution with mean 60 minutes. When patients stay in the delayed queues, they still occupy beds; thus no IP patient is blocked, i.e., only triage patients can be blocked. Let PB_j be the blocking probability of j-triage patients, $j = 1, 2, 3$. Finally, patients may leave the system without completing all treatments. To be specific, each patient has a patience time when they join a queue, and a new patience time starts after transfer to another queue. If the patience time expires when a patient waits in the queue, that patient leaves the ED without further treatments and will never return. The patience times are assumed to be exponential with the same rate $\theta = 0.001$ for all classes (mean patience time is about 16 hours). Let $\mathbb{P}(Ab)$ denote the fraction of abandoning patients, among all patients who join the system; recall that it is the sum of LWBS and LAMA. We simulate the systems under the four policies and the results are summarized in Table 2.

From Table 2, our proposed (TG$c\mu$) policy still outperforms the other three policies. Most of the triage patients (more than 96%) meet the deadline constraints, and the cost rate is much smaller than those under the FCFS policy and the IPF policy. The FCFS policy can neither meet the deadline constraint nor minimize cost rate. The IPF policy minimizes cost rate, but all three triage classes violate their deadlines (about 35% of the patients). The TrF policy incurs the highest cost among all four policies.

Table 2 System performances under the four policies

Policy	P_1	P_2	P_3	$\mathbb{P}(Ab)$	PB_1	PB_2	PB_3	Cost Rate
TGcμ	3.90%	3.76%	3.02%	6.80%	7.27%	7.30%	7.27%	43.27
FCFS	39.60%	0.90%	0.00%	7.18%	9.53%	9.53%	9.53%	139.47
IPF	37.42%	35.03%	34.24%	6.80%	6.11%	6.10%	6.10%	15.94
TrF	0.00%	0.00%	0.00%	7.21%	10.93%	10.93%	10.93%	261.26

8. Some future research directions

We considered the control problem of a multiclass queueing system with feedback and deadlines, motivated by its application to EDs. While our model, as is, captures usefully the dynamics of ED patient flow, it does leave out several noticeable ED characteristics, for example, delays between physician visits, time-varying arrivals, finite ED capacity and patient abandonment; all these have been incorporated in our simulation (§7.4 and §EC.9). The simulation results suggest that our proposed policy still works better than its competitors. These, and other ED features, are research worthy and will now be discussed.

8.1. Adding delays between physician visits

ED patients experience delays between successive visits to physicians. In Yom-Tov and Mandelbaum (2014), the delay phases are modeled as infinite-server queues (*content* phases). One would expect that, if the delays are short, those delays will have no impact asymptotically; at the other extreme, if the delays are long, then those patients experiencing long delays can be regarded as new arrivals and the system's performance will change accordingly. The question is how to make precise the meaning of "short" and "long", which we now formalize as a conjecture on the duration of delays.

Consider the basic (queue-length) model as an example. Following Yom-Tov and Mandelbaum (2014), we model delays between visits to physicians as infinite-server queues with exponential service times—these include the service time as well as the waiting time in say lab tests or for X-ray results. The individual service rate for the infinite-server queue between j-triage patients and k-IP patients is assumed to be $r^{\alpha_{jk}}\mu_{jk}$, and the one between l-IP patients and k-IP patients is $r^{\alpha_{lk}}\mu_{lk}$; here μ_{jk} and μ_{lk} are fixed positive constants. The magnitude of the α's will determine "short" delays (large α) vs. "long" (small). Specifically, we conjecture that when $\alpha > -2$ (for all α's), the delays are then short enough to leave our results intact. Conversely, $\alpha_{jk} < -2$ (for all j,k) decouples the triage from IP—both can be controlled separately; and $\alpha_{lk} < -2$ (for all l,k) pushes the IP feedback far enough into the future so that the IP sub-system can be analyzed as a queueing system without feedback. All other cases require further thought and plausibly a more delicate analysis. We provide an additional brief discussion in §EC.8.

Simulations in §EC.9.2 show that, even with relatively long delays, our proposed policy still outperforms its competitors. Moreover, the queue lengths of IP classes are away from 0 only if the triage patients are close to violating the deadlines, which suggests that our proposed policy is still asymptotically optimal. However, the latter is yet to be proved.

8.2. Time-varying arrival rates

Emergency departments, like many other service systems, must cope with arrival rates that are significantly time-varying (Figure 2). In the present paper, we have focused our attention on the ED afternoon-to-evening peak, which renders relevant a stationary critically-loaded model. Nevertheless, it is still of interest, and theoretically challenging, to view the ED as a time-varying queueing system. This is especially true when physician capacity cannot be matched well with demand—an unfortunate recurring scene in EDs—in which case the system could alternate between underloaded and overloaded periods of a day (Mandelbaum and Massey (1995), Liu and Whitt (2012)). The triage part of the time-varying ED flow control is analyzed in Carmeli (2012), where the following problem is solved, in a fluid framework and for a single triage-class: Minimize physician capacity for triage patients subject to adhering to their triage constraints. A corresponding IP part is carried out in Bäuerle and Stidham (2001). Combining these two results could provide the starting point for solving the flow control problem for a time-varying ED, within a fluid framework.

On the practical side, in §EC.9.3 we simulated an ED with time-varying arrival rates, using parameters collected from a real ED. In this simulation, the traffic intensity exceeds 1.2 for a long period of the day. Yet it shows that, under the proposed policy, most of the triage patients meet their corresponding deadlines. The proposed policy also outperforms the three commonly-used alternatives. One is thus left to theoretically explain the success of our proposed policy in the face of time-varying arrivals.

8.3. Finite ED capacity

In our current work, we assume that ED capacity is infinite. This is true in many Israeli EDs (including the ED of our partner hospital), as well as other EDs around the world (in which patients can stay not only on beds, but also say in chairs). We showed that (theoretically and via simulation), our proposed policy can keep the number of IP patients under control, which may ameliorate the need for ample ED capacity. However, finite ED capacity is also one of the major reasons for ED congestion; see for example Batt and Terwiecsch (2012), Hoot and Aronsky (2008). As a result, it is of interest to understand the impact of finite ED capacity, which would give rise to an admission control problem, as in Plambeck et al. (2001) and Ward and Kumar (2008). Interestingly, admission control problems, with costs incurred by blocked customers, in fact motivated Plambeck et al. (2001) and Ward and Kumar (2008).

We simulated EDs with finite ED capacity in §EC.9.5. Under realistic parameters, we varied the ED capacity from 10 to 200 beds. The simulation results show that systems with large ED capacity (larger than 100) are almost the same as systems with infinite ED capacity; hence our results can be applied to systems with large ED capacity. For systems with moderate to small ED capacity, a new theory is called for, but simulations still show that our proposed policy outperforms its competitors.

8.4. Length-of-Stay constraints

Many EDs implement, or at least strive for, an upper bound on patients' overall Length of Stay (LoS). The goal of our ED-Partner, for example, is to release a patient within at most 4 hours. Note, however, that if there are too many patients within the ED, LoS constraints could simply become infeasible. As in §8.3, one could or, perhaps, should apply a rationalized admission control—a rare protocol in our ED-Partner, but relatively prevalent in U.S. EDs in the form of ambulance diversion (Deo and Gurvich (2011), Allon et al. (2013)).

8.5. On "non-interchangeable" physicians

In the current paper, we assume that the N-physicians are interchangeable, which is then asymptotically equivalent to a system with a single "super" physician. In reality, ED physicians are often "non-interchangeable": a patient that starts service with a physician must remain with that physician through all successive visits. This "non-interchangeable" system is not work-conserving. However, we conjecture that it is still asymptotically equivalent to the system analyzed in the present paper. Here is a brief discussion to justify such a conjecture.

When physicians are "non-interchangeable," a physician cannot handle the patients being assigned to other physicians; thus they can be viewed as N parallel service stations. Consequently, the system has an inverted-V structure with N service stations, with each station having a queue in which patients can wait for treatment. Since servers are i.i.d., we conjecture that there is a "state-space-collapse" between the workload of those stations (Bramson (1998)). Denote by $\widehat{W}_n^r(t)$ the diffusion-scaled workload at station n. Then, for every $T > 0$, we expect to find a sequence $\delta^r \downarrow 0$ with $\mathbb{P}(\sup_{0\le t\le T} \sup_{m,n} |\widehat{W}_n^r(t) - \widehat{W}_m^r(t)| \ge \delta^r) \le \delta^r$. Then, if there is one physician whose diffusion-scaled workload exceeds δ^r, other physicians cannot be idle (with probability $1-\delta^r$). With such a sequence of δ^r's, one can apply Theorem 4.1 of Williams (1998), which would imply that the diffusion limit of all servers' workload is equivalent to one that arises from "interchangeable" physicians.

8.6. Adding abandonment to triage or IP patients

Empirical evidence shows that the fraction of registered emergency patients who 'Leave Without Being Seen' (LWBS) is non-negligible (Armony et al. (2013), Green et al. (2006)). This has become a growing concern in overcrowded EDs, as those LWBS patients may miss their necessary care and be exposed to unnecessary medical risk (see for example Batt and Terwiecsch (2013), Bolandifar et al. (2014)). The 'LWBS' phenomenon corresponds to adding abandonment in our model. Customer abandonment has been analyzed in service systems such as call centers, and has proved significant in affecting system performance and optimal decisions: see Ward and Glynn (2005), Reed and Ward (2008) for single-server systems; Garnett et al. (2002), Mandelbaum and Zeltyn (2009) for many-server systems; and Ward (2011) for a comprehensive summary.

Abandonment also significantly impact the structure of optimal policies. For systems without feedback, Kim and Ward (2013) considered linear cost, with hazard rate scaling of patience time distributions, and Ata and Tongarlak (2012) covered general cost functions with exponential patience time distributions. Both works analyze the corresponding Brownian control problem, and then interpret the results back to the original queueing system. They show that the $c\mu$ (or the generalized $c\mu$) is no longer an optimal policy. As a result, for systems with feedback, it is also natural to conjecture that the generalized $c\mu$ rule is not optimal. More fundamentally, a theoretical understanding of the impact of abandonment on systems with feedback is still lacking.

Being practical, we simulated EDs with patient abandonment over a widely varied level of impatience (§EC.9.6). Patient abandonment is costly, but it reduces both violation probabilities and cost rates. The simulations show that our proposed policy outperforms its competitors across all abandonment rates (though a new theory is required for rate exceeding 1%).

Acknowledgments

We are grateful to the Area Editor, Associate Editor and the referees for their insightful comments, which helped us significantly improve the quality and clarity of the paper.

This research grew out of challenges and insights from Dr. Shlomi Israelit, ED Director at the Rambam Hospital, Israel. The authors would thus like to thank Shlomi, as well as Jim Dai, Itai Gurvich, Martin Reiman, Nahum Shimkin, Melvyn Sim and Hanqin Zhang, for discussions that helped shape our research and improve the present paper. We also gratefully acknowledge the help of Galit Yom-Tov with our simulation.

The joint research of J.H. and A.M. was partially supported by grants from the National University of Singapore (NUS); the funds for promotion of research and sponsored research at the Technion, Haifa, Israel; and the Statistics and Applied Mathematical Sciences Institute (SAMSI) of the NSF, North Carolina, U.S.A. - the hospitality and support of these institutions is thankfully acknowledged. The work of J.H. has been partially supported by the Hong Kong RGC grants (24500314, T32-102/14N). The work of A.M. has been partially supported by BSF Grants 2005175 and 2008480, as well as ISF Grant 1357/08.

References

Allon, G., S. Deo, W. Lin. 2013. Impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research.* **61** 544–562.

Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, G. B. Yom-Tov. 2013. Patient flow in hospitals: A data-based queueing-science perspective. *Working Paper.* Technion – Israel Institute of Technology.

Ata, B., H. M. Tongarlak. 2013. On scheduling a multiclass queue with abandonments under general delay costs. *Queueing Systems.* **74**(1) 65–104.

Atar, R., A. Mandelbaum, A. Zviran. 2012. Control of Fork-Join networks in heavy traffic. *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing.*

Batt, R.J., C. Terwiesch. 2012. Doctors under load: An empirical study of state-dependent service times in emergency care. *Working Paper.* The Wharton School.

Batt, R.J., C. Terwiesch. 2013. Waiting patiently: Queue abandonment in an emergency department. *Working Paper.* The Wharton School.

Bäuerle, N., S. Stidham. 2001. Conservation laws for single-server fluid networks. *Queueing Systems.* **38**(2) 185–194.

Bolandifar, E., DeHoratius, N., Olsen, T., Wiler, J. L. Modeling the behavior of patients who leave the emergency department without being seen by a physician. *Working Paper.*

Brailsford, S. C., P. R. Harper, B. Patel, M. Pitt. 2009. An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation.* **3**(3) 130–140.

Bramson, M. 1998. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems.* **30**(1-2) 89–140.

Carmeli, B. 2012. *Real Time Optimization of Patient Flow in Emergency Departments*. M.Sc. Thesis. Technion – Israel Institute of Technology.

Chen, H., J. G. Shanthikumar. 1994. Fluid limits and diffusion approximations for networks of multi-server queues in heavy traffic. *Discrete Event Dynamic Systems.* **4**(3) 269–291.

Chen, H., D. D. Yao. 1993. Dynamic scheduling of a multiclass fluid network. *Operations Research.* **41**(6) 1104–1115.

Chen, H., D. D. Yao. 2001. *Fundamentals of queuing networks: Performance, asymptotics, and optimization.* Springer-Verlag.

Dai, J. G., T. G. Kurtz. 1995. A multiclass station with Markovian feedback in heavy traffic. *Mathematics of Operations Research.* **20**(3) 721–742.

Deo, S., I. Gurvich. 2011. Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science.* **57**(7) 1300–1319.

Dobson, G., T. Tezcan, V. Tilson. 2013. Optimal workflow decisions for investigators in systems with interruptions. *Management Science.* **59**(5) 1125–1141.

Farrohknia, N., M. Castrén, A. Ehrenberg, L. Lind, S. Oredsson, H. Jonsson, K. Asplund, K. E. Göransson. 2011. Emergency department triage scales and their components: A systematic review of the scientific evidence. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine.* **19:**42.

Garnett, O., A. Mandelbaum, M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing & Service Operations Management.* **4**(3) 208–227.

Green, L. V., J. Soares, J. F. Giglio, R. A. Green. 2006. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine.* **13** 61–68.

Gurvich, I., W. Whitt. 2009. Queue-and-Idleness-Ratio controls in many-server service systems. *Mathematics of Operations Research.* **34**(2) 363–396.

Hoot, N., D. Aronsky. 2008. Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine.* **52**(2) 126–136.

Huang, J. 2013. Patient flow management in emergency departments. Ph.D. Thesis. Available at http://ie.technion.ac.il/serveng/References/Thesis_Junfei_Huang.pdf.

Ibrahim, R., W. Whitt. 2009. Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management.* **11**(3) 397–415.

Kim, J., A. R. Ward. 2013. Dynamic scheduling of a GI/GI/1+ GI queue with multiple customer classes. *Queueing Systems.* **75**(2-4) 339–384.

Klimov, G. P. 1974. Time-sharing service systems. I. *Theory of Probability and its Applications.* **19**(3) 532–551.

Klimov, G. P. 1978. Time-sharing service systems. II. *Theory of Probability and its Applications.* **23**(2) 314–321.

Liu, Y., W. Whitt. 2012. The $G_t/GI/s_t+GI$ many-server fluid queue. *Queueing Systems.* **71**(4) 405–444.

Mace, S. E., T. A. Mayer. 2008. Triage. Chapter 155 in *Pediatric Emergency Medicine.* Baren, J. M., Rothrock, S. G., Brennan, J. A. and Brown, L. (eds.), Philadelphia: Saunders, Elsevier. 1087–1096.

Mandelbaum, A., W. A. Massey. 1995. Strong approximations for time-dependent queues. *Mathematics of Operations Research.* **20**(1) 33–64.

Mandelbaum, A., A. L. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research.* **52**(6) 836–855.

Mandelbaum, A., S. Zeltyn. 2009. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research.* **57**(5) 1189–1205.

Marmor, Y. N., B. Golany, S. Israelit, A. Mandelbaum. 2012. Designing patient flow in emergency departments. *Working Paper.* Technion – Israel Institute of Technology.

Niska, R., F. Bhuiya, J. Xu. August 6, 2010. National hospital ambulatory medical care survey: 2007 emergency department summary. *National Health Statistics Reports.* **26**.

Pitts, S. R., E. W. Nawar, J. Xu, C. W. Burt. August 6, 2008. National hospital ambulatory medical care survey: 2006 emergency department summary. *National Health Statistics Reports.* **7**.

Plambeck, E., S. Kumar, J. M. Harrison. 2001. A multiclass queue in heavy traffic with throughput time constraints: Asymptotically optimal dynamic controls. *Queueing Systems.* **39**(1) 23–54.

Reed, J. E., A. R. Ward. 2008 Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic *Mathematics of Operations Research.* **33** 606-644.

Reiman, M. I. 1982. The heavy traffic diffusion approximation for sojourn times in Jackson networks. *Applied Probability and Computer Science – The Interface* Volumne 2, R. L. Disney and T. J. Ott (eds.), Boston: Birkhauser. 409–422.

Reiman, M. I. 1984. Open queueing networks in heavy traffic. *Mathematics of Operations Research.* **9**(3) 441–458.

Reiman, M. I. 1988. A multiclass feedback queue in heavy traffic. *Advances in Applied Probability.* **20**(1) 179–207.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L. Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research.* **60**(5), 1080–1097.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L. Kronick. 2014. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management.* **16**(3), 329–345.

SEELab, Technion. http://ie.technion.ac.il/Labs/Serveng/.

van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Annals of Applied Probability.* **5**(3) 809–833.

van Mieghem, J. A. 2003. Due-date scheduling: Asymptotic optimality of generalized longest queue and generalized largest delay rules. *Operations Research.* **51**(1) 113–122.

Ward, A. 2011. Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys in Operations Research and Management Science.* **16** 1-14.

Ward, A., P. Glynn. 2005. A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing Systems.* **50** 371-400.

Ward, A. R., S. Kumar. 2008. Asymptotically optimal admission control of a queue with impatient customers. *Mathematics of Operations Research.* **33** 167–202.

Whitt, W. 2002. *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*. Springer-Verlag.

Wiler, J. L., C. Gentle, J. M. Halfpenny, A. Heins, A. Mehrotra, M. G. Mikhail, D. Fite. 2002. Optimizing emergency department Front-End operations. *Annals of Emergency Medicine.* **55**(2) 142–160.

Williams, R. J. 1998. An invariance principle for semimartingale reflecting Brownian motion. *Queueing Systems.* **30**(1-2) 5-25.

Yom-Tov, G. B., A. Mandelbaum. 2014. Erlang-R: A time-varying queue with ReEntrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management.* **16**(2) 283-299.

Zaied, I. 2012. The offered load in Fork-Join networks: Calculations and applications to service engineering of emergency department. M.Sc. Thesis. Available at http://ie.technion.ac.il/serveng/References/Thesis_Itamar_Zaied.pdf.

Zayas-Caban, G., J. Xie, L. V. Green, M. E. Lewis. 2013. Optimal control of an emergency room triage and treatment process. *Working Paper.*

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Proofs

EC.1. Preliminary analysis

In this section, we derive some consequences of the asymptotically compliant assumption. We also set up system dynamic equations that apply to all policies. These results will be used in subsequent sections.

We start with an analysis that covers any asymptotically compliant family of control policies. An implicit corollary from asymptotic compliance is that $\{\widehat{\tau}_j^r, j \in \mathcal{J}\}$ are stochastically bounded, which gives rise to many useful stochastic boundedness results on other processes.

For any j-triage class, $j \in \mathcal{J}$, introduce diffusion-scaled processes

$$\widehat{E}_j^r(t) = r^{-1}\left(E_j^r(r^2 t) - \lambda_j^r r^2 t\right),$$

$$\widehat{S}_j^r(t) = r^{-1}(S_j(\lfloor r^2 t \rfloor) - \mu_j r^2 t), \qquad \widehat{T}_j^r(t) = r^{-1}\left(T_j^r(r^2 t) - \lambda_j^r m_j r^2 t\right),$$

and fluid-scaled processes

$$\bar{\bar{Q}}_j^r(t) = r^{-2} Q_j^r(r^2 t), \qquad \bar{\bar{E}}_j^r(t) = r^{-2} E_j^r(r^2 t),$$

$$\bar{\bar{T}}_j^r(t) = r^{-2} T_j^r(r^2 t), \qquad \bar{\bar{S}}_j^r(t) = r^{-2} S_j(r^2 t).$$

From Donsker's Theorem, as $r \to \infty$,

$$(\widehat{E}_j^r,\ \widehat{S}_j^r,\ j \in \mathcal{J}) \Rightarrow (\widehat{E}_j,\ \widehat{S}_j,\ j \in \mathcal{J}); \tag{EC.1}$$

here $(\widehat{E}_j,\ j \in \mathcal{J})$ and $(\widehat{S}_j,\ j \in \mathcal{J})$ are independent driftless Brownian motions, with the corresponding covariance matrices

$$\operatorname{diag}(\lambda_j a_j^2), \qquad \operatorname{diag}(\mu_j b_j^2).$$

The following lemma follows from the fact that the customers in queue at time t are those customers arriving during the waiting time of the head-of-the-line customer.

Lemma EC.1.1 *Under any asymptotically compliant family of control policies, and for all* $T \geq 0$*,*

$$\max_{j \in \mathcal{J}} \sup_{0 \le t \le T} \left| \widehat{Q}_j^r(t) - \lambda_j \widehat{\tau}_j^r(t) \right| \Rightarrow 0, \quad \textit{as} \quad r \to \infty. \tag{EC.2}$$

Proof: For each triage class $j \in \mathcal{J}$, the patients in queue at time t are those patients arriving between $[t - \tau_j^r(t), t]$, thus

$$\left| Q_j^r(t) - \left(E_j^r(t) - E_j^r\left((t - \tau_j^r(t))-\right)\right)\right| \le 1.$$

Then

$$\left| \widehat{Q}_j^r(t) - \lambda_j^r \widehat{\tau}_j^r(t) \right| \le \left| \widehat{E}_j^r(t) - \widehat{E}_j^r\left((t - \bar{\bar{\tau}}_j^r(t))-\right)\right| + \frac{1}{r}, \quad j \in \mathcal{J}, \tag{EC.3}$$

where $\bar{\bar{\tau}}_j^r(t) = r^{-2}\tau_j^r(r^2 t)$. From the definition of asymptotic compliance, $\bar{\bar{\tau}}_j^r \Rightarrow 0$ and $\widehat{\tau}_j^r$ are stochastically bounded for all $j \in \mathcal{J}$. Together with (EC.1) and (7), (EC.2) is easily proved from (EC.3), in view of the Random-Time-Change theorem. □

The following is a direct corollary, which translates the asymptotic compliance condition to the language of queue length processes. As a result, the queue lengths of the triage patients have upper bounds.

Corollary 2 *Under any asymptotically compliant family of control policies, as $r \to \infty$,*

$$\sup_{0 \le t \le T} \left[\widehat{Q}_j^r(t)/\lambda_j - \widehat{d}_j \right]^+ \Rightarrow 0, \quad j \in \mathcal{J}.$$

In the following lemma, we analyze the fluid busy time for triage patients, under any asymptotically optimal policy. We also prove that $\widehat{Q}_j^r(\cdot) + \mu_j \widehat{T}_j^r(\cdot)$ converge, though we cannot (and need not) prove that each of the summands converges individually. An important corollary is stochastic boundedness, which will help us in choosing the appropriate scaling in the lower bound proof.

Lemma EC.1.2 *Under any asymptotically compliant family of control policies, as $r \to \infty$,*

$$\bar{\bar{T}}_j^r(\cdot) \Rightarrow \lambda_j m_j e(\cdot), \tag{EC.4}$$

$$\widehat{Q}_j^r(\cdot) + \mu_j \widehat{T}_j^r(\cdot) \Rightarrow \widehat{E}_j(\cdot) - \widehat{S}_j\left(\lambda_j m_j e(\cdot)\right). \tag{EC.5}$$

Consequently, $\widehat{Q}_j^r$ and $\widehat{T}_j^r$ are stochastically bounded.

Proof: For $j \in \mathcal{J}$, as

$$Q_j^r(t) = Q_j^r(0) + E_j^r(t) - S_j(T_j^r(t)),$$

then

$$\bar{\bar{Q}}_j^r(t) = \bar{\bar{Q}}_j^r(0) + \bar{\bar{E}}_j^r(t) - \lambda_j^r t - \left[\bar{\bar{S}}_j^r\left(\bar{\bar{T}}_j^r(t)\right) - \mu_j \bar{\bar{T}}_j^r(t)\right] + \mu_j\left[\lambda_j^r m_j t - \bar{\bar{T}}_j^r(t)\right] \tag{EC.6}$$

and

$$\widehat{Q}_j^r(t) = \widehat{Q}_j^r(0) + \widehat{E}_j^r(t) - \widehat{S}_j^r(\bar{\bar{T}}_j^r(t)) - \mu_j \widehat{T}_j^r(t). \tag{EC.7}$$

From Corollary 2 and the Functional Law of Large Numbers, for any $T \ge 0$, as $r \to \infty$,

$$\sup_{0 \le t \le T} \bar{\bar{Q}}_j^r(t) \Rightarrow 0, \qquad \sup_{0 \le t \le T} \left|\bar{\bar{E}}_j^r(t) - \lambda_j^r t\right| \Rightarrow 0, \tag{EC.8}$$

$$\sup_{0 \le t \le T} \left|\bar{\bar{S}}_j^r\left(\bar{\bar{T}}_j^r(t)\right) - \mu_j \bar{\bar{T}}_j^r(t)\right| \le \sup_{0 \le t \le T} \left|\bar{\bar{S}}_j^r(t) - \mu_j t\right| \Rightarrow 0, \tag{EC.9}$$

and (EC.4) can be easily obtained from (EC.6). Then (EC.1) and (EC.7), together with the Random-Time-Change theorem, imply (EC.5). □

We next discuss system dynamics, without assuming a specific policy. Thus the following discussion (till the end of this subsection) can be applied to all policies.

Recall that $\phi_j(n)$ is the indicator function recording the class to which the nth j-triage patient transfers (§2.1), and $\phi_l(n)$ the indicator function recording the class to which the nth

l-IP patient transfers (§2.2). We use $\phi_{jk}(n)$ to denote $(\phi_j(n))_k$, the kth element of $\phi_j(n)$, and introduce

$$\Phi_{jk}(n) := \sum_{i=1}^{n} \phi_{jk}(i),$$

to record the transition to k-IP patients from the first n j-triage patients. Similarly we use $\phi_{lk}(n)$ to denote $(\phi_l(n))_k$, the kth element of $\phi_l(n)$ and then

$$\Phi_{lk}(n) := \sum_{i=1}^{n} \phi_{lk}(i),$$

records the transition to k-IP patients from the first n served l-IP patients. Since the transition vectors are assumed invariant with respect to r, there is no superscript to Φ_{jk} and Φ_{lk}.

Define the diffusion-scaled processes for $j \in \mathcal{J}, l, k \in \mathcal{K}$:

$$\begin{aligned}
&\widehat{E}_k^r(t) = r^{-1}(E_k^r(r^2t) - \lambda_k^r r^2 t), && \\
&\widehat{S}_k^r(t) = r^{-1}(S_k(r^2t) - \mu_k r^2 t), && \widehat{T}_k^r(t) = r^{-1}(T_k^r(r^2t) - \lambda_k^r m_k r^2 t), \\
&\widehat{\Phi}_{jk}^r(t) = r^{-1}\left(\Phi_{jk}(\lfloor r^2 t\rfloor) - P_{jk} r^2 t\right), && \widehat{\Phi}_{lk}^r(t) = r^{-1}\left(\Phi_{lk}(\lfloor r^2 t\rfloor) - P_{lk} r^2 t\right).
\end{aligned}$$

Then from Donsker's Theorem, as $r \to \infty$,

$$\begin{aligned}
&\left(\widehat{\Phi}_{jk}^r(\cdot), \widehat{\Phi}_{lk}^r(\cdot), \widehat{S}_k^r(\cdot);\ j \in \mathcal{J}, l, k \in \mathcal{K}\right) \\
\Rightarrow &\left(\widehat{\Phi}_{jk}(\cdot), \widehat{\Phi}_{lk}(\cdot), \widehat{S}_k(\cdot);\ j \in \mathcal{J}, l, k \in \mathcal{K}\right);
\end{aligned} \tag{EC.10}$$

here $(\widehat{\Phi}_{jk}(\cdot), k \in \mathcal{K}),\ j \in \mathcal{J}$, $(\widehat{\Phi}_{kl}(\cdot), l \in \mathcal{K}),\ k \in \mathcal{K}$, $(\widehat{S}_k(\cdot), k \in \mathcal{K})$ are independent driftless Brownian motions, with covariance matrices

$$\Gamma^j,\ j \in \mathcal{J}, \quad \Gamma^k,\ k \in \mathcal{K}, \quad \text{and} \quad \text{diag}(b_k^2),$$

respectively.

Recall that $E_k^r(t)$ is the arrival process for k-IP patients, $k \in \mathcal{K}$. Then

$$Q_k^r(t) = Q_k^r(0) + E_k^r(t) - S_k(T_k^r(t)), \tag{EC.11}$$

and

$$E_k^r(t) = \sum_{j \in \mathcal{J}} \Phi_{jk}^r\left(S_j\left(T_j^r(t)\right)\right) + \sum_{l \in \mathcal{K}} \Phi_{lk}^r\left(S_l\left(T_l^r(t)\right)\right).$$

From this and (1), similarly to (EC.7),

$$\begin{aligned}
\widehat{Q}_k^r(t) &= \widehat{Q}_k^r(0) + \widehat{E}_k^r(t) - \widehat{S}_k^r(\bar{\bar{T}}_k^r(t)) - \mu_j \widehat{T}_k^r(t) \\
&= \widehat{Q}_k^r(0) + \widehat{\mathcal{E}}_k^r(t) - \widehat{S}_k^r(\bar{\bar{T}}_k^r(t)) + \sum_{j \in \mathcal{J}} P_{jk}\mu_j \widehat{T}_j^r(t) + \sum_{l \in \mathcal{K}} P_{lk}\mu_l \widehat{T}_l^r(t) - \mu_k \widehat{T}_k^r(t);
\end{aligned} \tag{EC.12}$$

here

$$\widehat{\mathcal{E}}_k^r(t) = \sum_{j \in \mathcal{J}} \widehat{\Phi}_{jk}^r\left(\bar{\bar{S}}_j^r\left(\bar{\bar{T}}_j^r(t)\right)\right) + \sum_{l \in \mathcal{K}} \widehat{\Phi}_{lk}^r\left(\bar{\bar{S}}_l^r\left(\bar{\bar{T}}_l^r(t)\right)\right) + \sum_{j \in \mathcal{J}} P_{jk}\widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t)\right) + \sum_{l \in \mathcal{K}} P_{lk}\widehat{S}_l^r\left(\bar{\bar{T}}_l^r(t)\right). \tag{EC.13}$$

Now introduce the following processes ($\widehat{Q}^r_w(t)$ is recalled from (13) for convenience):

$$\begin{aligned}
\widehat{Q}^r_w(t) =& \sum_{j\in\mathcal{J}} m^e_j \widehat{Q}^r_j(t) + \sum_{k\in\mathcal{K}} m^e_k \widehat{Q}^r_k(t),\\
\widehat{X}^r_w(t) =& \widehat{Q}^r_w(0) + r(\rho^r - 1)t + \sum_{j\in\mathcal{J}} m^e_j \left[\widehat{E}^r_j(t) - \widehat{S}^r_j\left(\bar{\bar{T}}^r_j(t)\right)\right] + \sum_{k\in\mathcal{K}} m^e_k \left[\widehat{\mathcal{E}}^r_k(t) - \widehat{S}^r_k\left(\bar{\bar{T}}^r_k(t)\right)\right],\\
\widehat{T}^r_+(t) =& r^{-1}\left(r^2 t - \sum_{j\in\mathcal{J}} T^r_j(r^2 t) - \sum_{k\in\mathcal{K}} T^r_k(r^2 t)\right).
\end{aligned} \tag{EC.14}$$

From (5) and (4), one can verify that

$$-m^e_j\mu_j + \sum_{k\in\mathcal{K}} P_{jk}\mu_j m^e_k = -1, \tag{EC.15}$$

$$-m^e_k\mu_k + \sum_{l\in\mathcal{K}} P_{kl}\mu_k m^e_l = -1. \tag{EC.16}$$

Multiplying (EC.7) by m^e_j, (EC.12) by m^e_k, and summing them up, one has for all $t \ge 0$:

$$\begin{aligned}
&\widehat{Q}^r_w(t) = \widehat{X}^r_w(t) + \widehat{T}^r_+(t),\\
&\widehat{Q}^r_w(t) \ge 0,\\
&\widehat{T}^r_+(\cdot) \text{ is nondecreasing with } \widehat{T}^r_+(0) = 0.
\end{aligned} \tag{EC.17}$$

Note that the policy at hand needs not be work-conserving, thus it is possible for $\widehat{T}^r_+$ to increase at t while still $\widehat{Q}^r_w(t) > 0$. Hence

$$\widehat{Q}^r_w(t) \ge \Phi(\widehat{X}^r_w)(t), \tag{EC.18}$$

for all $t \ge 0$, here Φ is the 1-dimensional Skorohod mapping; see for example, Theorem 6.1 in Chen and Yao (2001). Equality in (EC.18) holds when the system operates under any work-conserving policy.

EC.2. Proof of Theorem 2: Lower Bound

We prove Theorem 2, the lower bound, in this section. We relate the event in the probability to three events. For the first, we can establish the desired lower bound; the second enables flexibility to construct a new converging sequence with the desired lower bound; and the third is negligible in probability.

Proof of Theorem 2: Fix an arbitrary family of control policies $\{\pi^r\}$, which is asymptotically compliant. Define

$$\begin{aligned}
\Gamma^r_1(t) &= \left\{\mathcal{U}^r(t) > x,\ \max_{k\in\mathcal{K}} \sup_{0\le s\le t} \bar{\bar{Q}}^r_k(s) \le \frac{1}{r^{1/4}}\right\},\\
\Gamma^r_2(t) &= \left\{\max_{k\in\mathcal{K}} \sup_{0\le s\le t} \bar{\bar{Q}}^r_k(s) > \frac{1}{r^{1/4}}\right\},\\
\Gamma^r_3(t) &= \left\{\mathcal{U}^r(t) \le x,\ \max_{k\in\mathcal{K}} \sup_{0\le s\le t} \bar{\bar{Q}}^r_k(s) > \frac{1}{r^{1/4}}\right\}.
\end{aligned}$$

Here $\bar{\bar{Q}}_k^r$ is the fluid-scaled number of k-IP patients in the system, defined via

$$\bar{\bar{Q}}_k^r(t) = r^{-2} Q_k^r(r^2 t), \quad k \in \mathcal{K}.$$

Then

$$\{\mathcal{U}^r(t) > x\} = (\Gamma_1^r(t) \cup \Gamma_2^r(t)) \setminus \Gamma_3^r(t). \tag{EC.19}$$

First we prove

$$\lim_{r\to\infty} \mathbb{P}\{\Gamma_3^r(t)\} = 0. \tag{EC.20}$$

For notational simplicity, denote $I^r(s,\vartheta) = [s, s + \frac{1}{\vartheta r^{1/4}}]$ and $\vartheta_0 = 4\max_{k\in\mathcal{K}} \mu_k$. For $s < u$, let $S_k^r(s,u) = S_k(T^r(r^2 s) + r^2(u-s)) - S_k(T^r(r^2 s))$, and $\bar{\bar{S}}_k^r(s,u) = r^{-2} S_k^r(s,u)$. By the strong law of large numbers for renewal processes, one can prove that

$$\lim_{r\to\infty} \mathbb{P}\left\{ \max_{k\in\mathcal{K}} \sup_{0\le s\le t} \sup_{u\in I^r(s,\vartheta_0)} \bar{\bar{S}}_k^r(s,u) > \frac{1}{2r^{1/4}} \right\} = 0.$$

Note that, for all $k \in \mathcal{K}$ and $u > s$, $Q_k^r(r^2 s) \le Q_k^r(r^2 u) + S_k^r(s,u)$, because $S_k^r(s,u)$ is the number of departures of k-IP patients during $[r^2 s, r^2 u]$, if the physician allocates all the capacity to k-IP patients during this period. Thus $\bar{\bar{Q}}_k^r(s) - \bar{\bar{Q}}_k^r(u) \le \bar{\bar{S}}_k^r(s,u)$ and

$$\lim_{r\to\infty} \mathbb{P}\left\{ \max_{k\in\mathcal{K}} \sup_{0\le s\le t} \sup_{u\in I^r(s,\vartheta_0)} \left[\bar{\bar{Q}}_k^r(s) - \bar{\bar{Q}}_k^r(u)\right] > \frac{1}{2r^{1/4}} \right\} = 0.$$

It follows that

$$\begin{aligned}
\lim_{r\to\infty} \mathbb{P}\{\Gamma_3^r(t)\} &\le \limsup_{r\to\infty} \mathbb{P}\left\{ \mathcal{U}^r(t) \le x, \max_{k\in\mathcal{K}} \sup_{0\le s\le t} \inf_{u\in I^r(s,\vartheta_0)} \bar{\bar{Q}}_k^r(u) > \frac{1}{2r^{1/4}} \right\} \\
&\le \limsup_{r\to\infty} \mathbb{P}\left\{ \min_{k\in\mathcal{K}} \frac{2}{\vartheta_0 r^{1/4}} C_k\left(\frac{1}{2} r^{3/4}\right) \le x, \max_{k\in\mathcal{K}} \sup_{0\le s\le t} \inf_{u\in I^r(s,\vartheta_0)} \bar{\bar{Q}}_k^r(u) > \frac{1}{2r^{1/4}} \right\} \\
&\le \limsup_{r\to\infty} \mathbb{P}\left\{ \frac{r^{1/2}}{\vartheta_0} \min_{k\in\mathcal{K}} \frac{2}{r^{3/4}} C_k\left(\frac{1}{2} r^{3/4}\right) \le x \right\} = 0.
\end{aligned}$$

This completes the proof of (EC.20).

We conclude from (EC.19) and (EC.20) that,

$$\liminf_{r\to\infty} \mathbb{P}\{\mathcal{U}^r(t) > x\} = \liminf_{r\to\infty} \mathbb{P}\{\Gamma_1^r(t) \cup \Gamma_2^r(t)\}. \tag{EC.21}$$

Next we derive a lower bound for the latter term.

Denote

$$\Gamma_0^r(t) = \left\{ \max_{k\in\mathcal{K}} \sup_{0\le s\le t} \bar{\bar{Q}}_k^r(s) \le r^{-1/4} \right\}.$$

We first prove that, on the sets $\Gamma_0^r(t)$, the following is true in $\mathcal{D}[0,t]$:

$$\bar{\bar{T}}_k^r(\cdot) \Rightarrow \lambda_k m_k e(\cdot), \quad k \in \mathcal{K}. \tag{EC.22}$$

This is similar to (EC.4), but for IP patients. It basically shows that, in fluid scaling, the physician allocates the desired amount of time to k-IP patients.

For $s \le t$, define $\widetilde{T}_j^r(s) = r^{-1}\widehat{T}_j^r(s)$, for $j \in \mathcal{J}$, and

$$\begin{aligned} \widetilde{Q}_k^r(s) &= r^{-1}\widehat{Q}_k^r(s), & \widetilde{\mathcal{E}}_k^r(s) &= r^{-1}\widehat{\mathcal{E}}_k^r(s), \\ \widetilde{S}_k^r(s) &= r^{-1}\widehat{S}_k^r(s), & \widetilde{T}_k^r(s) &= r^{-1}\widehat{T}_k^r(s), \\ \widetilde{\Phi}_{jk}^r(s) &= r^{-1}\widehat{\Phi}_{jk}^r(s), & \widetilde{\Phi}_{lk}^r(s) &= r^{-1}\widehat{\Phi}_{lk}^r(s), \end{aligned}$$

for $j \in \mathcal{J}$, $l,k \in \mathcal{K}$. Then from (EC.12),

$$\sum_{l \in \mathcal{K}} P_{lk}\mu_l \widetilde{T}_l^r(s) - \mu_k \widetilde{T}_k^r(s) \; = \widetilde{Q}_k^r(s) - \widetilde{Q}_k^r(0) - \widetilde{\mathcal{E}}_k^r(s) + \widetilde{S}_k^r\left(\bar{\bar{T}}_k^r(s)\right) - \sum_{j \in \mathcal{J}} P_{jk}\mu_j \widetilde{T}_j^r(s). \tag{EC.23}$$

On $\Gamma_0^r(t)$, we know that $\sup_{0\le s\le t}\widetilde{Q}_k^r(s) \Rightarrow 0$. Together with (EC.4), the expression of $\widetilde{\mathcal{E}}_k^r$ in (EC.13), and $\bar{\bar{T}}_k^r(s) \le s$ for all $k \in \mathcal{K}$ (those hold for all asymptotic compliant policies). We deduce that the terms on the right-hand side of (EC.23) converge to 0. Then, on $\Gamma_0^r(t)$,

$$\sum_{l \in \mathcal{K}} P_{lk}\mu_l \widetilde{T}_l^r(\cdot) - \mu_k \widetilde{T}_k^r(\cdot) \Rightarrow 0, \quad \text{in} \quad \mathcal{D}[0,t].$$

Introducing the K-dimensional process $\widetilde{T}_\mu^r(s) = (\mu_k \widetilde{T}_k^r(s))_{k\in\mathcal{K}}$ in $\mathcal{D}[0,t]$, the above is then

$$(P^T - I)\widetilde{T}_\mu^r(\cdot) \Rightarrow 0, \quad \text{on} \quad \Gamma_0^r(t).$$

Since $P^T - I$ is invertible, and all μ_k, $k \in \mathcal{K}$, are nonzero, we have

$$\widetilde{T}_k^r(\cdot) \Rightarrow 0, \quad k \in \mathcal{K} \;\; \text{in} \;\; \mathcal{D}[0,t],$$

which is equivalent to (EC.22).

For $s \le t$, define $\widehat{\mathcal{X}}_0^r(s) = \widehat{X}_w^r(s)$ on $\Gamma_0^r(t)$, and otherwise,

$$\begin{aligned} \widehat{\mathcal{X}}_0^r(s) \; &= \sum_{j\in\mathcal{J}} m_j^e \widehat{Q}_j^r(0) + \sum_{k\in\mathcal{K}} m_k^e \widehat{Q}_k^r(0) + r(\rho^r - 1)s \\ &\quad + \sum_{j\in\mathcal{J}} m_j^e \left[\widehat{E}_j^r(s) - \widehat{S}_j^r\left(\lambda_j^r m_j s\right)\right] + \sum_{k\in\mathcal{K}} m_k^e \left[\check{\mathcal{E}}_k^r(s) - \widehat{S}_k^r\left(\lambda_k^r m_k s\right)\right]; \end{aligned}$$

here for $k \in \mathcal{K}$,

$$\check{\mathcal{E}}_k^r(s) = \sum_{j\in\mathcal{J}} \widehat{\Phi}_{jk}^r\left(\lambda_j^r s\right) + \sum_{l\in\mathcal{K}} \widehat{\Phi}_{lk}^r\left(\lambda_l^r s\right) + \sum_{j\in\mathcal{J}} P_{jk}\widehat{S}_j^r\left(\lambda_j^r m_j s\right) + \sum_{l\in\mathcal{K}} P_{lk}\widehat{S}_l^r\left(\lambda_l^r m_l s\right).$$

From (EC.22) on $\Gamma_0^r(t)$, (EC.4) and $\lambda_k^r \to \lambda_k$, $k \in \mathcal{K}$, one deduces that

$$\widehat{\mathcal{X}}_0^r \Rightarrow \widehat{X},$$

in $\mathcal{D}[0,t]$, as $r \to \infty$. Here $\widehat{X}$ is the Brownian motion defined in §5.3. For $s \le t$, introduce

$$\widehat{\mathcal{Z}}_+^r(s) = \left(\Phi(\widehat{\mathcal{X}}_0^r)(s) - \sum_{j\in\mathcal{J}} m_j^e(\widehat{Q}_j^r(s) - \lambda_j^r \widehat{d}_j)^+ - \sum_{j\in\mathcal{J}} m_j^e \lambda_j^r \widehat{d}_j\right)^+.$$

Then, by the continuity of Φ and the definition of asymptotic compliance, in $\mathcal{D}[0,t]$ as $r \to \infty$,

$$\widehat{\mathcal{Z}}_+^r(\cdot) \Rightarrow \left(\widehat{Q}_w(\cdot) - \widehat{\omega}\right)^+.$$

From (EC.18), on $\Gamma_0^r(t)$,

$$\sum_{k\in\mathcal{K}} m^e \widehat{Q}_k^r(s) \geq \widehat{\mathcal{Z}}_+^r(s), \quad s \leq t.$$

By the definition of $\Delta_{\mathcal{K}}$ and the nondecreasing property of Δ_k, for all $k \in \mathcal{K}$, we have

$$\begin{aligned}\Gamma_1^r(t) \cup \Gamma_2^r(t) \ &\supseteq \left\{ \int_0^t \sum_{k\in\mathcal{K}} C_k \left(\Delta_k(\widehat{\mathcal{Z}}_+^r(s)) \right) ds > x, \max_{k\in\mathcal{K}} \sup_{0\leq s\leq t} \bar{\bar{Q}}_k^r(s) \leq r^{-1/4} \right\} \cup \Gamma_2^r(t) \\ &\supseteq \left\{ \int_0^t \sum_{k\in\mathcal{K}} C_k \left(\Delta_k(\widehat{\mathcal{Z}}_+^r(s)) \right) ds > x \right\}.\end{aligned}$$

Combined with (EC.21),

$$\liminf_{r\to\infty} \mathbb{P}\left\{ \mathcal{U}^r(t) > x \right\} \geq \liminf_{r\to\infty} \mathbb{P}\left\{ \int_0^t \sum_{k\in\mathcal{K}} C_k \left(\Delta_k(\widehat{\mathcal{Z}}_+^r(s)) \right) ds > x \right\}.$$

From the convergence of $\widehat{\mathcal{Z}}_+^r$, the right-hand side is exactly the lower bound in Theorem 2. This completes the proof. □

EC.3. Proof of Proposition 1: Invariance principle for work-conserving policies

In this section we prove Proposition 1, which is the invariance principle for all work-conserving policies. From the discussion after (EC.18), one has the expression $\widehat{Q}_w^r(t) = \Phi(\widehat{X}_w^r)(t)$. As a result, it is enough to prove the convergence of $\widehat{X}_\omega^r$. The challenge is to establish the fluid limits needed for the Random-time-change theorem: these are in the form of (EC.4) and (EC.22), and they can be derived using the stochastic boundedness of the queue lengths. Then Proposition 1 can be proved using the Continuous mapping theorem together with the Random-time-change theorem. We now carry out these steps.

Proof of Proposition 1: For any family of work-conserving policies, in addition to (EC.17), the following holds as well:

$$\widehat{T}_+^r \text{ increases at } t \text{ only when } \widehat{Q}_w^r(t) = 0.$$

As a result, equality holds in (EC.18).

From (EC.10), (EC.1) and the fact that $\bar{\bar{T}}_j^r(s) \leq s$, $j \in \mathcal{J}$, and $\bar{\bar{T}}_k^r(s) \leq s$, $k \in \mathcal{K}$, one can see that $\widehat{X}_w^r$ in (EC.14) is stochastically bounded. By the Lipschitz continuity of Φ (Theorem 6.1 in Chen and Yao (2001)), $\widehat{Q}_w^r$ is stochastically bounded, which implies the stochastic boundedness of $\widehat{Q}_j^r$, $j \in \mathcal{J}$, and $\widehat{Q}_k^r$, $k \in \mathcal{K}$. Then $\bar{\bar{Q}}_j^r \Rightarrow 0$, for $j \in \mathcal{J}$. Note that (EC.6) is still true (under work-conserving policies). One then has

$$\bar{\bar{T}}_j^r(\cdot) \Rightarrow \lambda_j m_j e(\cdot), \quad j \in \mathcal{J}. \tag{EC.24}$$

For $k \in \mathcal{K}$, following the procedure in proving (EC.22) in the proof of Theorem 2, one also has

$$\bar{\bar{T}}_k^r(\cdot) \Rightarrow \lambda_k m_k e(\cdot), \quad k \in \mathcal{K}. \tag{EC.25}$$

Now (EC.24) and (EC.25), together with (EC.10), (EC.1) and the Random-Time-Change theorem, imply that, as $r \to \infty$,

$$\widehat{X}_w^r \Rightarrow \widehat{X}. \tag{EC.26}$$

By the continuity of the mapping Φ, (14) follows. □

EC.4. Proof of Theorem 3: State-Space Collapse

We now analyze the family of control policies $\{\pi_*^r\}$ and prove Theorem 3. We follow the standard framework in Bramson (1998) to prove State-space collapse.

EC.4.1. Hydrodynamic limit

Under the policies $\{\pi_*^r\}$, the following dynamic equations hold:

$$\begin{aligned}
&Q_j^r(t) = Q_j^r(0) + E_j^r(t) - D_j^r(t), \quad j \in \mathcal{J}, \\
&D_j^r(t) = S_j\left(T_j^r(t)\right), \quad j \in \mathcal{J}, \\
&Q_k^r(t) = Q_k^r(0) + E_k^r(t) - D_k^r(t), \quad k \in \mathcal{K}, \\
&E_k^r(t) = \sum_{j\in\mathcal{J}} \Phi_{jk}^r\left(S_j\left(T_j^r(t)\right)\right) + \sum_{l\in\mathcal{K}} \Phi_{lk}^r\left(S_l\left(T_l^r(t)\right)\right), \quad k \in \mathcal{K}, \\
&D_k^r(t) = S_k\left(T_k^r(t)\right), \quad k \in \mathcal{K}, \\
&\sum_{j\in\mathcal{J}} \left[T_j^r(t) - T_j^r(s)\right] + \sum_{k\in\mathcal{K}} \left[T_k^r(t) - T_k^r(s)\right] \le t - s, \quad \text{for} \quad s < t, \\
&Y^r(t) = t - \left(\sum_{j\in\mathcal{J}} T_j^r(t) + \sum_{k\in\mathcal{K}} T_k^r(t)\right), \\
&\int_0^\infty \left(\left(d_j^r - \tau_j^r(t) - \min_{j'\in\mathcal{J}, Q_{j'}(t)\neq 0} \left\{d_{j'}^r - \tau_{j'}^r(t)\right\}\right)^+ \wedge 1\right) dT_{j'}^r(t) = 0, \quad j' \in \mathcal{J}, \\
&\int_0^\infty 1\left(\max_{j\in\mathcal{J}}(\tau_j^r(t) - d_j^r) > -\epsilon^r\right) d\sum_{k\in\mathcal{K}} T_k^r(t) = 0, \\
&\int_0^\infty \left(\max_{k'\in\mathcal{K}} \frac{C_{k'}'(\bar{Q}_{k'}(t))}{m_{k'}^e} - \frac{C_k'(\bar{Q}_k(t))}{m_k^e}\right)^+ d\bar{T}_k(t) = 0, \quad k \in \mathcal{K}, \\
&\int_0^\infty 1\left(\max_{j\in\mathcal{J}}(\tau_j^r(t) - d_j^r) \le -\epsilon^r, \ \sum_{k\in\mathcal{K}} Q_k^r(t) > 0\right) d\sum_{j\in\mathcal{J}} T_j^r(t) = 0, \\
&\int_0^\infty 1\left(\sum_{j\in\mathcal{J}} m_j^e Q_j^r(t) + \sum_{k\in\mathcal{K}} m_k^e Q_k^r(t) > 0\right) dY^r(t) = 0.
\end{aligned}$$

Introduce the hydrodynamic scaled processes for j-triage classes, $j \in \mathcal{J}$, by

$$\begin{aligned}
&\bar{E}_j^r(t) = r^{-1}E_j^r(rt), \qquad \bar{S}_j^r(t) = r^{-1}S_j(rt), \qquad \bar{\tau}_j^r(t) = r^{-1}\tau_j^r(rt), \\
&\bar{T}_j^r(t) = r^{-1}T_j^r(rt), \qquad \bar{Q}_j^r(t) = r^{-1}Q_j^r(rt), \qquad \bar{D}_j^r(t) = r^{-1}D_j^r(rt),
\end{aligned}$$

and for k-IP classes, $k \in \mathcal{K}$,

$$\begin{aligned}
&\bar{E}_k^r(t) = r^{-1}E_k^r(rt), \qquad \bar{S}_k^r(t) = r^{-1}S_k(rt), \\
&\bar{T}_k^r(t) = r^{-1}T_k^r(rt), \qquad \bar{Q}_k^r(t) = r^{-1}Q_k^r(rt), \qquad \bar{D}_k^r(t) = r^{-1}D_k^r(rt).
\end{aligned}$$

First we prove the following lemma, which is similar to Lemma EC.1.1. This lemma helps to express age processes of triage patients in terms of their queue lengths.

Lemma EC.4.1 *For any $T > 0$, $\sup_{0\le t\le T} \left|\lambda_j^r \bar{\tau}_j^r(t) - \bar{Q}_j^r(t)\right| \Rightarrow 0$.*

Proof: For each triage class $j \in \mathcal{J}$, the patients in queue at time t are those patients arriving between $[t - \tau_j^r(t), t]$. Thus

$$\left|Q_j^r(t) - \left(E_j^r(t) - E_j^r\left((t - \tau_j^r(t))-\right)\right)\right| \le 1,$$

which implies

$$\left|\bar{Q}_j^r(t) - \left(\bar{E}_j^r(t) - \bar{E}_j^r\left((t - \bar{\tau}_j^r(t))-\right)\right)\right| \le \frac{1}{r}, \quad j \in \mathcal{J}. \tag{EC.27}$$

The lemma now follows from the functional law of large numbers, $\sup_{0\le t\le T} |\bar{E}_j^r(t) - \lambda_j^r t| \Rightarrow 0$, and (EC.27). □

With Lemma EC.4.1, similarly to Plambeck et al. (2001), we have the following

Proposition 7 *Assume $\bar{Q}_j^r(0) \Rightarrow \bar{Q}_j(0), j \in \mathcal{J}$, and $\bar{Q}_k^r(0) \Rightarrow \bar{Q}_k(0), k \in \mathcal{K}$, as $r \to \infty$. Then under the proposed family of control policies, almost surely, every sequence $\{r\}$ contains a subsequence $\{r_n\}$ such that, the hydrodynamic scaled processes $(\bar{E}_j^{r_n}, \bar{S}_j^{r_n}, \bar{\tau}_j^{r_n}, \bar{T}_j^{r_n}, \bar{Q}_j^{r_n}, \bar{D}_j^{r_n}, j \in \mathcal{J}; \bar{E}_k^{r_n}, \bar{S}_k^{r_n}, \bar{T}_k^{r_n}, \bar{Q}_k^{r_n}, \bar{D}_k^{r_n}, k \in \mathcal{K})$, converge uniformly on compact time sets to limit processes $(\bar{E}_j, \bar{S}_j, \bar{\tau}_j, \bar{T}_j, \bar{Q}_j, \bar{D}_j, j \in \mathcal{J}; \bar{E}_k, \bar{S}_k, \bar{T}_k, \bar{Q}_k, \bar{D}_k, k \in \mathcal{K})$, which satisfy the following equations:*

$$\bar{Q}_j(t) = \bar{Q}_j(0) + \lambda_j t - \bar{D}_j(t), \quad j \in \mathcal{J}, \tag{EC.28}$$

$$\bar{D}_j(t) = \mu_j \bar{T}_j(t), \quad j \in \mathcal{J}, \tag{EC.29}$$

$$\bar{Q}_k(t) = \bar{Q}_k(0) + \bar{E}_k(t) - \bar{D}_k(t), \quad k \in \mathcal{K}, \tag{EC.30}$$

$$\bar{E}_k(t) = \sum_{j\in\mathcal{J}} \mu_j P_{jk} \bar{T}_j(t) + \sum_{l\in\mathcal{K}} \mu_l P_{lk} \bar{T}_l(t), \quad k \in \mathcal{K}, \tag{EC.31}$$

$$\bar{D}_k(t) = \mu_k \bar{T}_k(t), \quad k \in \mathcal{K}, \tag{EC.32}$$

$$\lambda_j \bar{\tau}_j(t) = \bar{Q}_j(t), \quad j \in \mathcal{J}, \tag{EC.33}$$

$$\sum_{j\in\mathcal{J}} [\bar{T}_j(t) - \bar{T}_j(s)] + \sum_{k\in\mathcal{K}} [\bar{T}_k(t) - \bar{T}_k(s)] \le t - s, \quad \textit{for} \quad s < t, \tag{EC.34}$$

$$\bar{Y}(t) = t - \left(\sum_{j\in\mathcal{J}} \bar{T}_j(t) + \sum_{k\in\mathcal{K}} \bar{T}_k(t)\right), \tag{EC.35}$$

$$\int_0^\infty \left(\left(d_j - \frac{\bar{Q}_j(t)}{\lambda_j} - \min_{j'\in\mathcal{J}, \bar{Q}_{j'}(t)\neq 0} \left\{d_{j'} - \frac{\bar{Q}_{j'}(t)}{\lambda_{j'}}\right\}\right)^+ \wedge 1\right) d\bar{T}_j(t) = 0, \quad j \in \mathcal{J}, \tag{EC.36}$$

$$\int_0^\infty 1\left(\max_{j\in\mathcal{J}} (\bar{Q}_j(t) - \lambda_j \widehat{d}_j) > 0\right) d\sum_{k\in\mathcal{K}} \bar{T}_k(t) = 0, \tag{EC.37}$$

$$\int_0^\infty \left(\max_{k'\in\mathcal{K}} \frac{C'_{k'}(\bar{Q}_{k'}(t))}{m_{k'}^e} - \frac{C'_k(\bar{Q}_k(t))}{m_k^e}\right)^+ d\bar{T}_k(t) = 0, \quad k \in \mathcal{K}, \tag{EC.38}$$

$$\int_0^\infty 1\left(\max_{j\in\mathcal{J}}(\bar{Q}_j(t)-\lambda_j\widehat{d}_j)<0, \sum_{k\in\mathcal{K}}\bar{Q}_k(t)>0\right) d\sum_{j\in\mathcal{J}}\bar{T}_j(t)=0, \tag{EC.39}$$

$$\int_0^\infty 1\left(\sum_{j\in\mathcal{J}}m_j^e\bar{Q}_j(t)+\sum_{k\in\mathcal{K}}m_k^e\bar{Q}_k(t)>0\right) d\bar{Y}(t)=0. \tag{EC.40}$$

Remark 8 *We call any* $\bar{\mathcal{S}}=(\bar{E}_j,\bar{S}_j,\bar{\tau}_j,\bar{T}_j,\bar{Q}_j,\bar{D}_j, j\in\mathcal{J};\bar{E}_k,\bar{S}_k,\bar{T}_k,\bar{Q}_k,\bar{D}_k,k\in\mathcal{K})$ *satisfying* (EC.28)–(EC.40) *a* hydrodynamic model solution. *One can prove that any hydrodynamic model solution is Lipschitz, hence absolutely continuous and differentiable almost everywhere.*

Proposition 8 *Any hydrodynamic model solution satisfies*

$$\sum_{j\in\mathcal{J}}m_j^e\bar{Q}_j(t)+\sum_{k\in\mathcal{K}}m_k^e\bar{Q}_k(t)=\sum_{j\in\mathcal{J}}m_j^e\bar{Q}_j(0)+\sum_{k\in\mathcal{K}}m_k^e\bar{Q}_k(0).$$

Proof: From the fact that $\sum_{j\in\mathcal{J}}\lambda_j m_j^e=1$, (EC.15)–(EC.16) and (EC.28)–(EC.32), one gets

$$\sum_{j\in\mathcal{J}}m_j^e\bar{Q}_j(t)+\sum_{k\in\mathcal{K}}m_k^e\bar{Q}_k(t)=\sum_{j\in\mathcal{J}}m_j^e\bar{Q}_j(0)+\sum_{k\in\mathcal{K}}m_k^e\bar{Q}_k(0)+\bar{Y}(t).$$

From (EC.40), (EC.34) and (EC.35), we deduce that $\bar{Y}(\cdot)=0$. This completes the proof. □

EC.4.2. State-space collapse

First we prove a state-space collapse result for any hydrodynamic model solution. Here is the idea: the queue length of a class cannot be too small, otherwise that class will not receive service and its queue length will increase; conversely, queue length of a class cannot be too large, otherwise high priority will be assigned to that class and the queue length will decrease.

Proposition 9 (State-space collapse for hydrodynamic model solutions) *Fix* $C>0$. *For any hydrodynamic model solution with* $\sum_{j\in\mathcal{J}}m_j^e\bar{Q}_j(0)+\sum_{k\in\mathcal{K}}m_k^e\bar{Q}_k(0)<C$, *there exists a constant* T_0 *such that, for all* $t\ge T_0$,

$$\begin{aligned}\bar{Q}_{\mathcal{J}}(t) &= \Delta_{\mathcal{J}}\left(\min\left(\sum_{j\in\mathcal{J}}m_j^e\bar{Q}_j(t)+\sum_{k\in\mathcal{K}}m_k^e\bar{Q}_k(t),\ \widehat{\omega}\right)\right),\\ \bar{Q}_{\mathcal{K}}(t) &= \Delta_{\mathcal{K}}\left(\left(\sum_{j\in\mathcal{J}}m_j^e\bar{Q}_j(t)+\sum_{k\in\mathcal{K}}m_k^e\bar{Q}_k(t)-\widehat{\omega}\right)^+\right).\end{aligned}$$

Furthermore, if

$$\begin{aligned}\bar{Q}_{\mathcal{J}}(0) &= \Delta_{\mathcal{J}}\left(\min\left(\sum_{j\in\mathcal{J}}m_j^e\bar{Q}_j(0)+\sum_{k\in\mathcal{K}}m_k^e\bar{Q}_k(0),\ \widehat{\omega}\right)\right),\\ \bar{Q}_{\mathcal{K}}(0) &= \Delta_{\mathcal{K}}\left(\left(\sum_{j\in\mathcal{J}}m_j^e\bar{Q}_j(0)+\sum_{k\in\mathcal{K}}m_k^e\bar{Q}_k(0)-\widehat{\omega}\right)^+\right),\end{aligned}$$

then $\bar{Q}_{\mathcal{J}}(t)\equiv\bar{Q}_{\mathcal{J}}(0)$ *and* $\bar{Q}_{\mathcal{K}}(t)\equiv\bar{Q}_{\mathcal{K}}(0)$, *for all* $t\ge 0$.

Proof: We first prove the results for triage patients. For $j \in \mathcal{J}$, define

$$f_j(t) = \frac{1}{\lambda_j \widehat{d}_j}\left(\bar{Q}_j(t) - \Delta_j\left(\min\left(\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j(0) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(0), \widehat{\omega}\right)\right)\right)^-, \quad t \geq 0.$$

If $f_j(t) > 0$ and f_j is differentiable at t, then we claim

$$f_j'(t) = -\frac{1}{\widehat{d}_j} < 0.$$

Indeed, if this is not the case, then $\bar{T}_j'(t) > 0$ and from (EC.36), one has $\widehat{d}_j - \bar{Q}_j(t)/\lambda_j = \min_{j'\in\mathcal{J},\bar{Q}_{j'}(t)\neq 0}\{\widehat{d}_{j'} - \bar{Q}_{j'}(t)/\lambda_{j'}\}$. Together with $f_j(t) > 0$, one can prove by contradiction that $\bar{Q}_j(t) < \lambda_j \widehat{d}_j$, which then implies $\max_{j'\in\mathcal{J}}(\bar{Q}_{j'} - \lambda_{j'}\widehat{d}_{j'}) < 0$. Then from (EC.39), one has $\bar{Q}_k(t) = 0$, for all $k \in \mathcal{K}$. This, together with $f_j(t) > 0$ and $\widehat{d}_j - \bar{Q}_j(t)/\lambda_j = \min_{j'\in\mathcal{J},\bar{Q}_{j'}(t)\neq 0}\{\widehat{d}_{j'} - \bar{Q}_{j'}(t)/\lambda_{j'}\}$, contradict the definition of Δ_j.

As a result, f_j will decrease to 0 in a finite time (denoted by T_1) and once becoming 0, it will never be positive again. Now there are a finite number of triage classes, hence, after a finite time (denoted by $T_2 \geq T_1$), all f_j will be 0 and will never become positive again.

For $t \geq T_2$, we have $f_j(t) = 0$, for all $j \in \mathcal{J}$. Define

$$g_j(t) = \frac{1}{\lambda_j \widehat{d}_j}\left(\bar{Q}_j(t) - \Delta_j\left(\min\left(\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j(0) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(0), \widehat{\omega}\right)\right)\right)^+, \quad t \geq 0. \tag{EC.41}$$

We can assume $g_1(t) > 0$ whenever $\sum_{j\in\mathcal{J}} \lambda_j \widehat{d}_j m_j g_j(t) > 0$. Otherwise, if $g_1(t) = 0$ and there is another $j \in \mathcal{J}$ such that $g_j(t) > 0$, then from the definition of $\Delta_{\mathcal{J}}$, $\widehat{d}_1 - \bar{Q}_1(t)/\lambda_1 > \min_{j\in\mathcal{J}}[\widehat{d}_j - \bar{Q}_j(t)/\lambda_j]$, and from (EC.36), $\bar{T}_1'(t) = 0$ and $g_1'(t) = \frac{1}{\widehat{d}_1} > 0$. Hence right after t, $g_1(\cdot) > 0$ holds.

Now, as $f_j(t) = 0$, for all $j \in \mathcal{J}$ and over $t \geq T_2$, together with $g_1(t) > 0$ and the definition of Δ_j, we have $\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j(t) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(t) > \widehat{\omega}$, $\sum_{k\in\mathcal{K}} \bar{Q}_k(t) > 0$, and for $1 \in \mathcal{J}$, $\bar{Q}_1(t) > \lambda_1 \widehat{d}_1$. Then from (EC.37), $\sum_{k\in\mathcal{K}} \bar{T}_k'(t) = 0$. From (EC.40), $\sum_{j\in\mathcal{J}} \bar{T}_j'(t) = 1$. As a result, the derivative of $\sum_{j\in\mathcal{J}} \lambda_j \widehat{d}_j m_j g_j(t)$ is

$$\sum_{j\in\mathcal{J}} \lambda_j m_j - 1 \ < \ 0.$$

Thus in finite time (denoted by $T_3 \geq T_2$), $\sum_{j\in\mathcal{J}} \lambda_j \widehat{d}_j m_j g_j(t)$ will hit 0. It follows that, for all $t \geq T_3$, $f_j(t) = g_j(t) = 0$, $j \in \mathcal{J}$. Finally, from Proposition 8,

$$\begin{aligned}\bar{Q}_{\mathcal{J}}(t) =& \Delta_{\mathcal{J}}\left(\min\left(\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j(0) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(0), \widehat{\omega}\right)\right)\\ =& \Delta_{\mathcal{J}}\left(\min\left(\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j(t) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(t), \widehat{\omega}\right)\right), \quad \text{for} \quad t \geq T_3.\end{aligned}$$

Now we turn to the IP patients. From the above discussion, for all $t \geq T_3$,

$$\sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(t) = \left(\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j(0) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(0) - \widehat{\omega}\right)^+.$$

Recall the definition of Δ_k in Lemma 5.1 and let

$$\bar{Q}_0 = \Delta_{\mathcal{K}}\left(\left(\sum_{j\in\mathcal{J}} m^e_j \bar{Q}_j(0) + \sum_{k\in\mathcal{K}} m^e_k \bar{Q}_k(0) - \widehat{\omega}\right)^+\right).$$

Then for any $k,l\in\mathcal{K}$, $\frac{C'_k(\bar{Q}_{0k})}{m^e_k} = \frac{C'_l(\bar{Q}_{0l})}{m^e_l}$. We now prove that there exists a constant T_0 such that, for all $t \geq T_0$,

$$\bar{Q}_{\mathcal{K}}(t) = \bar{Q}_0. \tag{EC.42}$$

If (EC.42) does not hold at a certain t, then there exists a k_+ such that, for all $k_- \in \left\{k\in\mathcal{K}, \frac{C'_k(\bar{Q}_k(t))}{m^e_k} - \frac{C'_k(\bar{Q}_{0k})}{m^e_k} < 0\right\}$,

$$\frac{C'_{k+}(\bar{Q}_{k+}(t))}{m^e_{k+}} > \frac{C'_{k+}(\bar{Q}_{0k+})}{m^e_{k+}} = \frac{C'_{k-}(\bar{Q}_{0k-})}{m^e_{k-}} > \frac{C'_{k-}(\bar{Q}_{k-}(t))}{m^e_{k-}}.$$

From (EC.38) we then have $\bar{T}'_{k_-}(t) = 0$, which implies $\bar{Q}'_{k_-}(t) = \lambda_k > 0$. As a result, $\frac{C'_{k-}(\bar{Q}_{k-}(t))}{m^e_{k-}}$ is increasing in t. As there is a finite number of IP classes, there must be a finite time T_0 such that, for all $t \geq T_0$, $\frac{C'_k(\bar{Q}_k(t))}{m^e_k} \geq \frac{C'_k(\bar{Q}_{0k})}{m^e_k}$, for all $k\in\mathcal{K}$, which is equivalent to $\bar{Q}_k(t) \geq \bar{Q}_{0k}$. However, we have

$$\sum_{k\in\mathcal{K}} m^e_k \bar{Q}_k(t) = \sum_{k\in\mathcal{K}} m^e_k \bar{Q}_{0k},$$

hence $\bar{Q}_k(t) = \bar{Q}_{0k}$, for all $k\in\mathcal{K}$ and $t \geq T_0$. This completes the proof. □

Our main result in this subsection is the following proposition, which establishes state-space collapse for triage patients. The proof follows from Proposition 9 and the framework of Bramson (1998). For completeness, we include it here.

Proposition 10 *Under Assumption 1 and the proposed family of control policies, as $r\to\infty$,*

$$\sup_{0\leq t\leq T}\left|\widehat{Q}^r_j(t) - \Delta_j\left(\min\left(\widehat{Q}^r_w(t),\ \widehat{\omega}\right)\right)\right| \Rightarrow 0,$$
$$\sup_{0\leq t\leq T}\left|\widehat{Q}^r_k(t) - \Delta_k\left(\left(\widehat{Q}^r_w - \widehat{\omega}\right)^+\right)\right| \Rightarrow 0.$$

Proof: Lemma 9 implies Assumption 3.2 of Bramson (1998). Then from Theorem 5 of Bramson (1998), we deduce "multiplicative state-space collapse" (Equation (3.41) there):

$$\frac{\sup_{0\leq t\leq T}\left|\widehat{Q}^r_j(t) - \Delta_j\left(\min\left(\widehat{Q}^r_w(t),\ \widehat{\omega}\right)\right)\right|}{\sup_{0\leq t\leq T}\widehat{Q}^r_w(t)\vee 1} \Rightarrow 0,$$
$$\frac{\sup_{0\leq t\leq T}\left|\widehat{Q}^r_k(t) - \Delta_k\left(\left(\widehat{Q}^r_w - \widehat{\omega}\right)^+\right)\right|}{\sup_{0\leq t\leq T}\widehat{Q}^r_w(t)\vee 1} \Rightarrow 0.$$

Note that here $\widehat{Q}^r_w(t)$ plays the role of $\widehat{W}^r$ in Theorem 5 of Bramson (1998).

Next, our Proposition 1 implies that $\sup_{0\le t\le T}\widehat{Q}^r_w(t)\vee 1$ is stochastically bounded. As a result,

$$\sup_{0\le t\le T}\left|\widehat{Q}^r_j(t)-\Delta_j\left(\min\left(\widehat{Q}^r_w(t),\ \widehat{\omega}\right)\right)\right|\Rightarrow 0,$$
$$\sup_{0\le t\le T}\left|\widehat{Q}^r_k(t)-\Delta_k\left(\left(\widehat{Q}^r_w-\widehat{\omega}\right)^+\right)\right|\Rightarrow 0,$$

which proves the proposition. □

Proof of Theorem 3: This can be deduced from Propositions 1 and 10. □

EC.5. Proof of Theorem 1: Asymptotic Optimality

Proof of Theorem 1: First, it can be verified that $\Delta_j(\min(x,\widehat{\omega}))\le\lambda_j\widehat{d}_j$, for any x and $j\in\mathcal{J}$. Then from Theorem 3, under the proposed policies $\{\pi^r_*\}$, $\widehat{Q}^r_j\Rightarrow\widehat{Q}_j\le\lambda_j\widehat{d}_j$. An analysis of work-conserving policies (Lemma EC.6.2) shows that (EC.2) still holds for any work-conserving policy. This implies that $\widehat{Q}^r_j\Rightarrow\widehat{Q}_j\le\lambda_j\widehat{d}_j$ is equivalent to “asymptotic compliance” for work-conserving policies. As a result, the family of policies $\{\pi^r_*\}$ is asymptotically compliant.

By Theorem 3, together with the continuity of the cost functions, we also have

$$\int_0^t\sum_{k\in\mathcal{K}}C_k\left(\widehat{Q}^r_k(s)\right)ds\quad\Rightarrow\quad\int_0^t\sum_{k\in\mathcal{K}}C_k\left(\widehat{Q}_k(s)\right)ds=\int_0^t\sum_{k\in\mathcal{K}}C_k\left(\Delta_k\left((\widehat{Q}_w(s)-\widehat{\omega})^+\right)\right)ds.$$

Hence, under the family of the proposed policies, the lower bound in Theorem 2 is attained. As a result, the family of the proposed policies is asymptotically optimal. □

EC.6. Additional results for work-conserving policies

We now establish some additional results for work-conserving policies which, in particular, apply to our proposed family of control policies $\{\pi^r_*\}$. We first prove stochastic boundedness of the arrival processes for IP classes, and the busy time processes of triage and IP classes. This stochastic boundedness will be then used to prove that the fluid virtual waiting times converge to 0. Finally, we prove that the queue length and the age process for triage patients are close in diffusion scale. Notice that we are now considering work-conserving policies, instead of asymptotically compliant policies as in Lemma EC.1.1. Consequently, we do not have at our disposal the stochastic boundedness of $\widehat{\tau}^r_j$ for work-conserving policies, until we justify such boundedness independently.

From the discussion in the proof of Proposition 1, $\widehat{Q}^r_j$, $j\in\mathcal{J}$, are stochastically bounded and (EC.24) holds for any work-conserving policies. With these facts, notice that (EC.7) still prevails under work-conserving policies; hence we can verify the convergence (EC.5). As $\widehat{Q}^r_j$, $j\in\mathcal{J}$, are stochastically bounded, $\widehat{T}^r_j$, $j\in\mathcal{J}$, are also stochastically bounded.

Next consider IP patients. Define $\widehat{\mathcal{Y}}^r_{\mathcal{K}}=(\widehat{\mathcal{Y}}^r_k)_{k\in\mathcal{K}}$, $k\in\mathcal{K}$, by

$$\widehat{\mathcal{Y}}^r_k(t)=\widehat{Q}^r_k(t)-\widehat{Q}^r_k(0)-\widehat{\mathcal{E}}^r_k(t)+\widehat{S}^r_k(\bar{\bar{T}}^r_k(t))-\sum_{j\in\mathcal{J}}P_{jk}\mu_j\widehat{T}^r_j(t),$$

and recall that $\widehat{\mathcal{E}}_k^r$ is defined in (EC.13). Denote $\widehat{T}_\mu^r = (\mu_k \widehat{T}_k^r)_{k\in\mathcal{K}}$. Then from (EC.12),

$$\widehat{T}_\mu^r = (P^T - I)^{-1}\widehat{\mathcal{Y}}_{\mathcal{K}}^r. \tag{EC.43}$$

We can easily verify the stochastic boundedness of $\widehat{\mathcal{Y}}_{\mathcal{K}}^r$ from the facts $\bar{\bar{T}}_j^r(s) \le s$ and $\bar{\bar{T}}_k^r(s) \le s$, for all $j \in \mathcal{J}$, $k \in \mathcal{K}$ and $s \ge 0$. This implies the stochastic boundedness of $\widehat{T}_\mu^r$, and consequently the stochastic boundedness of $\widehat{T}_{\mathcal{K}}^r = (\widehat{T}_k^r)_{k\in\mathcal{K}}$.

Note that, for all $k \in \mathcal{K}$,

$$\widehat{E}_k^r(t) = \widehat{\mathcal{E}}_k^r(t) + \sum_{j\in\mathcal{J}} P_{jk}\mu_j \widehat{T}_j^r(t) + \sum_{l\in\mathcal{K}} P_{lk}\mu_l \widehat{T}_l^r(t). \tag{EC.44}$$

The stochastic boundedness of $\widehat{E}_k^r$ can be then obtained from the stochastic boundedness of $\widehat{\mathcal{E}}_k^r$, $\widehat{T}_j^r$ and $\widehat{T}_l^r$ $(j \in \mathcal{J},\ k, l \in \mathcal{K})$.

Define the fluid-scaled virtual waiting time processes as

$$\bar{\bar{\omega}}_j^r(t) = r^{-2}\omega_j^r\left(r^2 t\right), \quad j \in \mathcal{J}, \qquad \bar{\bar{\omega}}_k^r(t) = r^{-2}\omega_k^r\left(r^2 t\right), \quad k \in \mathcal{K}.$$

First we prove the following:

Lemma EC.6.1 *Under any family of work-conserving policies, with FCFS among each IP class, as $r \to \infty$,*

$$\bar{\bar{\omega}}_j^r \Rightarrow 0, \quad j \in \mathcal{J}, \tag{EC.45}$$

$$\bar{\bar{\omega}}_k^r \Rightarrow 0, \quad k \in \mathcal{K}. \tag{EC.46}$$

Proof: We only prove the results for $j \in \mathcal{J}$, as the proof for $k \in \mathcal{K}$ is the same. First note that, for any $\epsilon > 0$, if $\omega_j^r(t) \ge \epsilon$,

$$S_j\left(T_j^r(t+\epsilon)\right) \le Q_j^r(0) + E_j^r(t).$$

Then $\bar{\bar{\omega}}_j^r(t) \ge \epsilon$ ensures

$$\widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t+\epsilon)\right) + \mu_j \widehat{T}_j^r(t+\epsilon) + \lambda_j^r r\epsilon \le \widehat{Q}_j^r(0) + \widehat{E}_j^r(t).$$

Hence, for any fixed $T > 0$ and $\epsilon > 0$, we have

$$\mathbb{P}\left\{\sup_{0\le t\le T} \bar{\bar{\omega}}_j^r(t) \ge \epsilon\right\} \le \mathbb{P}\left\{\lambda_j^r r\epsilon \le \sup_{0\le t\le T}\left|\widehat{Q}_j^r(0) + \widehat{E}_j^r(t) - \widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t+\epsilon)\right) - \mu_j \widehat{T}_j^r(t+\epsilon)\right|\right\}.$$

However, the stochastic boundedness of $\sup_{0\le t\le T}\left|\widehat{Q}_j^r(0) + \widehat{E}_j^r(t) - \widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t+\epsilon)\right) - \mu_j \widehat{T}_j^r(t+\epsilon)\right|$, together with the fact that $\lambda_j^r r\epsilon \to \infty$, implies that the probability on the right-hand side above converges to 0. Hence

$$\lim_{r\to\infty} \mathbb{P}\left\{\sup_{0\le t\le T} \bar{\bar{w}}_j^r(t) \ge \epsilon\right\} = 0.$$

This completes the proof. □

Lemma EC.6.2 *Under any family of work-conserving policies, for any given $T>0$, as $n\to\infty$, we have*

$$\sup_{0\le t\le T}\left|\lambda_j^r\widehat{\tau}_j^r(t)-\widehat{Q}_j^r(t)\right|\Rightarrow 0,\quad j\in\mathcal{J}.$$

Proof: The proof follows exactly that of Lemma EC.1.1, by starting with $\sup_{0\le s\le t}\bar{\bar{\tau}}_j^r(s)\Rightarrow 0$; the latter is a consequence of Lemma EC.6.1 and $\sup_{s\le t}\tau_j^r(s)\le\sup_{s\le t}\omega_j^r(s)$, for all t and j. Note that the assumptions here slightly differ from Lemma EC.1.1. In the latter, the stochastic boundedness of $\widehat{\tau}_j^r, j\in\mathcal{J}$, follows from asymptotic compliance, while here we do not have the stochastic boundedness of $\widehat{\tau}_j^r, j\in\mathcal{J}$, in advance. □

EC.7. Proofs of Propositions 2–6

EC.7.1. Proof of Proposition 2: Asymptotic Sample-Path Little's Law

Lemma EC.7.1 *Under the family of control policies $\{\pi_*^r\}$, as $r\to\infty$,*

$$\left(\widehat{T}_j^r, j\in\mathcal{J};\ \widehat{E}_k^r,\widehat{T}_k^r,k\in\mathcal{K}\right)\Rightarrow\left(\widehat{T}_j,j\in\mathcal{J};\ \widehat{E}_k,\widehat{T}_k,k\in\mathcal{K}\right),$$

for some continuous processes $\left(\widehat{T}_j,j\in\mathcal{J};\ \widehat{E}_k,\widehat{T}_k,k\in\mathcal{K}\right)$ satisfying

$$\mu_j\widehat{T}_j(t)=-\widehat{Q}_j(t)+\widehat{E}_j(t)-\widehat{S}_j\left(\lambda_j m_j t\right),\tag{EC.47}$$

$$\widehat{E}_k(t)=\widehat{\mathcal{E}}_k(t)+\sum_{j\in\mathcal{J}}P_{jk}\mu_j\widehat{T}_j(t)+\sum_{l\in\mathcal{K}}P_{lk}\mu_l\widehat{T}_l(t),\tag{EC.48}$$

$$(P^T-I)\left(\mu_k\widehat{T}_k\right)_{k\in\mathcal{K}}=\widehat{\mathcal{Y}}_{\mathcal{K}}.\tag{EC.49}$$

Here

$$\widehat{\mathcal{E}}_k(t)=\sum_{j\in\mathcal{J}}\widehat{\Phi}_{jk}\left(\lambda_j t\right)+\sum_{l\in\mathcal{K}}\widehat{\Phi}_{lk}\left(\lambda_l t\right)+\sum_{j\in\mathcal{J}}P_{jk}\widehat{S}_j\left(\lambda_j m_j t\right)+\sum_{l\in\mathcal{K}}P_{lk}\widehat{S}_l\left(\lambda_l m_l t\right),$$
$$\widehat{\mathcal{Y}}_k(t)=\widehat{Q}_k(t)-\widehat{\mathcal{E}}_k(t)+\widehat{S}_k(\lambda_k m_k t)-\sum_{j\in\mathcal{J}}P_{jk}\mu_j\widehat{T}_j(t).$$

Proof: From (EC.7), (EC.44) and (EC.43), we have ($\widehat{T}_\mu^r=(\mu_k\widehat{T}_k^r)_{k\in\mathcal{K}}$)

$$\widehat{T}_j^r(t)=\left[\widehat{Q}_j^r(0)-\widehat{Q}_j^r(t)+\widehat{E}_j^r(t)-\widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t)\right)\right]/\mu_j,\tag{EC.50}$$

$$\widehat{E}_k^r(t)=\widehat{\mathcal{E}}_k^r(t)+\sum_{j\in\mathcal{J}}P_{jk}\mu_j\widehat{T}_j^r(t)+\sum_{l\in\mathcal{K}}P_{lk}\mu_l\widehat{T}_l^r(t),\tag{EC.51}$$

$$\widehat{T}_\mu^r(t)=(P^T-I)^{-1}\widehat{\mathcal{Y}}_{\mathcal{K}}^r(t),\tag{EC.52}$$

where

$$\widehat{\mathcal{E}}_k^r(t)=\sum_{j\in\mathcal{J}}\widehat{\Phi}_{jk}^r\left(\bar{\bar{S}}_j^r\left(\bar{\bar{T}}_j^r(t)\right)\right)+\sum_{l\in\mathcal{K}}\widehat{\Phi}_{lk}^r\left(\bar{\bar{S}}_l^r\left(\bar{\bar{T}}_l^r(t)\right)\right)+\sum_{j\in\mathcal{J}}P_{jk}\widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t)\right)+\sum_{l\in\mathcal{K}}P_{lk}\widehat{S}_l^r\left(\bar{\bar{T}}_l^r(t)\right),$$
$$\widehat{\mathcal{Y}}_k^r(t)=\widehat{Q}_k^r(t)-\widehat{Q}_k^r(0)-\widehat{\mathcal{E}}_k^r(t)+\widehat{S}_k^r(\bar{\bar{T}}_k^r(t))-\sum_{j\in\mathcal{J}}P_{jk}\mu_j\widehat{T}_j^r(t).$$

As a result, $\left(\widehat{T}_j^r, j\in\mathcal{J};\ \widehat{E}_k^r, \widehat{T}_k^r, k\in\mathcal{K}\right)$ can be represented as a continuous mapping from $\left(\widehat{Q}_j^r, \widehat{E}_j^r, \widehat{S}_j^r, \bar{\bar{T}}_j^r, \widehat{\Phi}_{jk}^r, \widehat{\Phi}_{lk}^r, \widehat{Q}_k^r, \widehat{S}_k^r, \bar{\bar{T}}_k^r, j\in\mathcal{J}, l,k\in\mathcal{K}\right)$, the convergence of which can be obtained from the assumptions and Theorem 3. The expressions (EC.47)–(EC.49) in the lemma can be verified from (EC.50)–(EC.52). This completes the proof. □

Proof of Proposition 2: We prove the result only for j-triage patients. For k-IP patients, the proof is similar. The convergence of $\widehat{Q}_j^r$, together with Lemma EC.6.1, ensure that for any $T>0$,

$$\sup_{0\le t\le T}\left|\widehat{Q}_j^r(t)-\widehat{Q}_j^r\left(t+\bar{\bar{\omega}}_j^r(t)\right)\right|\Rightarrow 0,\quad \text{as}\quad r\to\infty.$$

Thus it is enough to prove

$$\sup_{0\le t\le T}\left|\lambda_j^r\widehat{\omega}_j^r(t)-\widehat{Q}_j^r\left(t+\bar{\bar{\omega}}_j^r(t)\right)\right|\Rightarrow 0,\quad \text{as}\quad r\to\infty.$$

Note that the j-triage patients that are present at time $t+\omega_j^r(t)$ arrived during the time interval $(t,t+\omega_j^r(t)]$, and those j-triage patients arriving during this interval will remain in this class, or finish this stage of service at $t+\omega_j^r(t)$. Hence

$$Q_j^r\left(t+\omega_j^r(t)\right)\le E_j^r(t+\omega_j^r(t))-E_j^r(t)\le Q_j^r\left(t+\omega_j^r(t)\right)+\Delta S_j^r\left(t+\omega_j^r(t)\right); \tag{EC.53}$$

here, with some abuse of notation, $\Delta S_j^r\left(t+\omega_j^r(t)\right)=S_j\left(T^r(t+\omega_j^r(t))\right)-S_j\left(T^r(t+\omega_j^r(t)-)\right)$. From this last relationship, we deduce the following for the diffusion-scaled processes:

$$\left|\lambda_j^r\widehat{\omega}_j^r(t)-\widehat{Q}_j^r\left(t+\bar{\bar{\omega}}_j^r(t)\right)\right|\le\left|\widehat{E}_j^r\left(t+\bar{\bar{\omega}}_j^r(t)\right)-\widehat{E}_j^r(t)\right|+\triangle\widehat{S}_j^r(t+\bar{\bar{\omega}}_j^r(t))+\mu_j\triangle\widehat{T}_j^r(t+\bar{\bar{\omega}}_j^r(t)). \tag{EC.54}$$

Here $\triangle\widehat{S}_j^r(t+\bar{\bar{\omega}}_j^r(t))=\widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t+\bar{\bar{\omega}}_j^r(t))\right)-\widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t+\bar{\bar{\omega}}_j^r(t)-)\right)$ and $\triangle\widehat{T}_j^r\left(t+\bar{\bar{\omega}}_j^r(t)\right)=\widehat{T}_j^r(t+\bar{\bar{\omega}}_j^r(t))-\widehat{T}_j^r(t+\bar{\bar{\omega}}_j^r(t)-)$. From the convergence of $\widehat{S}_j^r(\bar{\bar{T}}_j^r(\cdot))$ and $\widehat{T}_j^r(\cdot)$, both $\triangle\widehat{S}_j^r(\cdot+\bar{\bar{\omega}}_j^r(\cdot))$ and $\triangle\widehat{T}_j^r\left(\cdot+\bar{\bar{\omega}}_j^r(\cdot)\right)$ converge to 0. Together with Lemma EC.6.1 and the convergence of $\widehat{E}_j^r, j\in\mathcal{J}$, the processes on the right-hand side in (EC.54) will converge to 0; thus the process on the left-hand side will also converge to 0, which completes the proof. □

EC.7.2. Proof of Proposition 3: Snapshot Principle—Virtual Waiting Time and Age

Lemma EC.7.2 *Under the family of control policies $\{\pi_*^r\}$, for any given $T>0$, as $r\to\infty$,*

$$\sup_{0\le t\le T}\left|\lambda_k^r\widehat{\tau}_k^r(t)-\widehat{Q}_k^r(t)\right|\Rightarrow 0,\quad k\in\mathcal{K}.$$

Proof: The proof follows exactly the one for Lemma EC.1.1. For $k\in\mathcal{K}$, note that the convergence of $\widehat{E}_k^r$ has been proved in Lemma EC.7.1. On the other hand, $\sup_{s\le t}\tau_k^r(s)\le\sup_{s\le t}\omega_k^r(s)$, for all t and k; hence, from Lemma EC.6.1 we have $\sup_{0\le s\le t}\bar{\bar{\tau}}_k^r(s)\Rightarrow 0$. □

Proof of Proposition 3: This can be deduced from Proposition 2, Lemmas EC.6.2 and EC.7.2. □

EC.7.3. Proof of Proposition 4: Snapshot Principle—Sojourn Time and Queue Lengths

The argument here is adapted from Reiman (1984). Introduce the following notation: $\tau^r_{jh}(t)$ is the time at which the patient of interest to us arrives to the system, and $\zeta^r_{jki}(t)$ is the time at which this patient becomes a k-IP patient for the ith time (it is also related to h, but we omit h to simplify notation). Then

$$t \le \zeta^r_{jki}(t) \le \tau^r_{jh}(t) + W^r_{jh}(t). \tag{EC.55}$$

Define the fluid-scaled processes

$$\bar{\bar{\zeta}}^r_{jki}(t) = r^{-2}\zeta^r_{jki}(r^2 t), \quad \bar{\bar{W}}^r_{jh}(t) = r^{-2}W^r_{jh}(r^2 t), \quad \bar{\bar{\tau}}^r_{jh}(t) = r^{-2}\tau^r_{jh}(r^2 t).$$

Lemma EC.7.3 *Under the family of control policies $\{\pi^r_*\}$, with FCFS among each IP class, if h is j-feasible, then for any $T \ge 0$, as $r \to \infty$,*

$$\sup_{0\le t\le T} \bar{\bar{W}}^r_{jh}(t) \Rightarrow 0, \tag{EC.56}$$

$$\sup_{0\le t\le T} \left[\bar{\bar{\tau}}^r_{jh}(t) - t\right] \Rightarrow 0. \tag{EC.57}$$

Consequently, as $r \to \infty$,

$$\sup_{0\le t\le T} \left[\bar{\bar{\zeta}}^r_{jki}(t) - t\right] \Rightarrow 0. \tag{EC.58}$$

We first assume that this last lemma prevails and prove Proposition 4.

Proof of Proposition 4: The sojourn time $W^r_{jh}(t)$ can be represented as

$$W^r_{jh}(t) = \omega^r_j(\tau^r_{jh}(t)) + \sum_{k\in\mathcal{K}}\sum_{i=1}^{h_k} \omega^r_k\left(\zeta^r_{jki}(t)\right).$$

From this we get

$$\begin{aligned}
&\widehat{W}^r_{jh}(t) - \left[\frac{\widehat{Q}^r_j(t)}{\lambda^r_j} + \sum_{k\in\mathcal{K}} \frac{h_k}{\lambda^r_k}\widehat{Q}^r_k(t)\right] \\
=\;& \widehat{\omega}^r_j(\bar{\bar{\tau}}^r_{jh}(t)) + \sum_{k\in\mathcal{K}}\sum_{i=1}^{h_k} \widehat{\omega}^r_k\left(\bar{\bar{\zeta}}^r_{jki}(t)\right) - \left[\frac{\widehat{Q}^r_j(t)}{\lambda^r_j} + \sum_{k\in\mathcal{K}} \frac{h_k}{\lambda^r_k}\widehat{Q}^r_k(t)\right] \\
=\;& \left[\widehat{\omega}^r_j(t) - \frac{\widehat{Q}^r_j(t)}{\lambda^r_j}\right] + \sum_{k\in\mathcal{K}} h_k \left[\widehat{\omega}^r_k(t) - \frac{\widehat{Q}^r_k(t)}{\lambda_k}\right] \\
&+ \left[\widehat{\omega}^r_j\left(\bar{\bar{\tau}}^r_{jh}(t)\right) - \widehat{\omega}^r_j(t)\right] + \sum_{k\in\mathcal{K}}\sum_{i=1}^{h_k} \left[\widehat{\omega}^r_k\left(\bar{\bar{\zeta}}^r_{jki}(t)\right) - \widehat{\omega}^r_k(t)\right].
\end{aligned}$$

From Lemma EC.7.3 and the convergence of $\widehat{\omega}^r_j, j \in \mathcal{J}$ and $\widehat{\omega}^r_k, k \in \mathcal{K}$,

$$\left[\widehat{\omega}^r_j\left(\bar{\bar{\tau}}^r_{jh}(t)\right) - \widehat{\omega}^r_j(t)\right] + \sum_{k\in\mathcal{K}}\sum_{i=1}^{h_k} \left[\widehat{\omega}^r_k\left(\bar{\bar{\zeta}}^r_{jki}(t)\right) - \widehat{\omega}^r_k(t)\right] \Rightarrow 0.$$

Together with Proposition 2, the conclusion is immediate. □

Proof of Lemma EC.7.3: We first prove (EC.56). It is enough to show that, for any $\epsilon > 0$, there exists an $N < \infty$ such that, for all $r \geq N$,

$$\mathbb{P}\left\{\sup_{0\leq t\leq T} \bar{\bar{W}}^r_{jh}(t) \geq \epsilon\right\} \leq \epsilon.$$

Similarly to Reiman (1984), denote $\|h\| = \sum_{k=1}^K h_k$. Then we have

$$\mathbb{P}\left\{\sup_{0\leq t\leq T} \bar{\bar{W}}^r_{jh}(t) \geq \epsilon\right\} \leq \max_{k\in\mathcal{K}} \mathbb{P}\left\{\sup_{0\leq t\leq T+\epsilon} \bar{\bar{\omega}}^r_k(t) \geq \frac{\epsilon}{\|h\|+1}\right\} + \mathbb{P}\left\{\sup_{0\leq t\leq T+\epsilon} \bar{\bar{\omega}}^r_j(t) \geq \frac{\epsilon}{\|h\|+1}\right\}. \tag{EC.59}$$

From Lemma EC.6.1, the right-hand side of (EC.59) converges to 0, hence (EC.56) holds.

Let $L^r_{i,j,h} = \min\{n > i; h^r(j,n) = h\}$, where $h^r(j,n)$ is the visit vector associated with the nth j-triage patient. We can write

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{0\leq t\leq T} [\bar{\bar{\tau}}^r_{jh}(t) - t] \geq \epsilon\right\} \\
\leq\ & \mathbb{P}\left\{\inf_{0\leq t\leq T} [E^r_j(r^2 t + r^2\epsilon) - E^r_j(r^2 t)] < \frac{1}{2}\lambda_j r^2 \epsilon\right\} \\
& + \mathbb{P}\left\{E^r_j(r^2 T) > 2\lambda_j r^2\right\} + \mathbb{P}\left\{\sup_{1\leq i\leq 2\lambda_j r^2} [L^r_{i,j,h} - i] > \frac{1}{2}\lambda_j r^2 \epsilon\right\}.
\end{aligned}$$

The first two terms on the right-hand side converge to zero by the strong law of large numbers. The j-triage patients have i.i.d. paths and hence i.i.d. visit vectors. Let the probability of a particular j-triage patient, having visit vector h, be g_h, where $g_h > 0$ since h is j-feasible. Define $\hat{g}_h = 1 - g_h$. Then

$$\mathbb{P}\left\{\sup_{1\leq i\leq 2\lambda_j r^2} [L^r_{i,j,h} - i] > \frac{1}{2}\lambda_j r^2 \epsilon\right\} \leq 1 - \left[1 - \hat{g}_h^{\frac{1}{2}\lambda_k r^2 \epsilon}\right]^{2\lambda_k r^2} = 1 - \left[1 - \frac{r^2 \hat{g}_h^{\frac{1}{2}\lambda_k r^2 \epsilon}}{r^2}\right]^{2\lambda_k r^2}.$$

The same reasoning as in Reiman (1984) implies that the latter expression vanishes, as $r \to \infty$. This establishes (EC.57).

Combining (EC.56), (EC.57) with (EC.55), now yields (EC.58). □

Proof of Corollary 1: This is implied by Propositions 4, 2 and 3. □

EC.7.4. Outline of the proof for Proposition 5: Waiting Time Cost

We outline the proof of the lower bound, which is similar to Theorem 2. Then one can prove that the family of modified policies $\{\tilde{\pi}^r_*\}$ attains the lower bound, following the discussion in §EC.4; in particular, one requires similar state-space collapse results. Thus, $\{\tilde{\pi}^r_*\}$ is asymptotically optimal.

For all work-conserving policies, Proposition 1 and Lemma EC.6.1 hold. Then, similarly to the proof of Proposition 4 in van Mieghem (1995), we can prove that for any $0 \leq a < b \leq T$,

$$\frac{1}{\bar{\bar{E}}^r_k(b) - \bar{\bar{E}}^r_k(a)} \left(\int_a^b \widehat{\tau}^r_k d\bar{\bar{E}}^r_k - \int_a^b \widehat{Q}^r_k(s) ds\right) \Rightarrow 0.$$

Next, as in Proposition 6 and the discussion prior to Proposition 8 of van Mieghem (1995), the following is true:

$$\liminf_{r\to\infty} \mathbb{P}\left\{\widetilde{\mathcal{U}}^r(t) > x\right\} \geq \mathbb{P}\left\{\int_0^t \sum_{k\in\mathcal{K}} \lambda_k C_k\left(\widehat{\Delta}_k\left((\widehat{Q}_w(s)-\widehat{\omega})^+\right)/\lambda_k\right) ds > x\right\}.$$

Here $\widehat{\Delta}_{\mathcal{K}} = (\widehat{\Delta}_k)_{k\in\mathcal{K}}$ is defined, for any $a \geq 0$, as the solution $x^* = \widehat{\Delta}_{\mathcal{K}}(a)$ to the following:

$$\begin{aligned} \min_x \quad & \sum_{k\in\mathcal{K}} \lambda_k C_k(x_k/\lambda_k) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{K}} m_k^e x_k = a, \\ & x \geq 0. \end{aligned} \tag{EC.60}$$

EC.7.5. Proof of Proposition 6: Sojourn Time Cost

We first provide an outline for proving an asymptotic lower bound for all asymptotically compliant policies. Whenever there are IP patients in the ED, the physician should not be idle, as the physician can always serve an IP patient to reduce that patient's sojourn cost. Thus, we restrict our discussion to asymptotically compliant policies, in which the physician does not idle if there are IP patients. Then, for any asymptotically compliant family of control policies, one can prove that the family $\{\widehat{Q}^r_\omega\}$ is stochastically bounded, in particular the diffusion-scaled queue length processes of IP patients are stochastically bounded. Then (EC.45) and (EC.46) hold under asymptotically compliant policies, assuming that physicians are non-idle if IP patients are presented. Similarly to the proof of Proposition 4 in van Mieghem (1995), we can prove that, for any $0 \leq a < b \leq T$,

$$\frac{1}{\bar{\bar{E}}_k^r(b) - \bar{\bar{E}}_k^r(a)}\left(\int_a^b \widehat{\tau}_k^r d\bar{\bar{E}}_k^r - \int_a^b \widehat{Q}_k^r(s)ds\right) \Rightarrow 0.$$

Now, following Proposition 6 and the discussion prior to Proposition 8 of van Mieghem (1995), we can prove that

$$\lim_{r\to\infty} \mathbb{P}\left(\widetilde{S}^r(t) > x\right) \geq \mathbb{P}\left(\int_0^t \sum_{k\in\mathcal{K}} \lambda_k C_k\left(\widehat{\Delta}^*_k\left((\widehat{Q}_w(s)-\widehat{\omega})^+\right)/\lambda_k\right) ds > x\right).$$

Here $\widehat{\Delta}^*_{\mathcal{K}} = (\widehat{\Delta}^*_k)_{k\in\mathcal{K}}$ is defined, for any $a \geq 0$, via the solution to the following:

$$\begin{aligned} \min_x \quad & \sum_{k\in\mathcal{C}_0} \lambda_k C_k\left(\sum_{j\in\mathcal{C}_k} x_j/\lambda_k\right) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{C}_0}\sum_{k'\in\mathcal{C}_k} m_{k'}^e x_{k'} = a, \\ & x \geq 0. \end{aligned} \tag{EC.61}$$

One can prove that the proposed family of control policies $\{\widetilde{\pi}^r_{**}\}$ attains the lower bound by showing the corresponding state-space collapse. Here we give some structural insights into the optimal solution of (EC.61). For classes in $\mathcal{C}_k$, we know that if $\sum_{k'\in\mathcal{C}_k} m_{k'}^e x_{k'}$ is fixed, then the solution minimizing $C_k(\sum_{j\in\mathcal{C}_k} x_j/\lambda_k)$ has necessarily x_k as non-zero, while all other x_j with $j\in\mathcal{C}_k\backslash\{k\}$ are 0 (this is because $m_k^e > m_j^e$, for all $j\in\mathcal{C}_k\backslash\{k\}$). As a result, if the problem has an optimal solution with some $k'\in\mathcal{C}_k\backslash\{k\}$, for some k, then one can always find a better solution, which is a contradiction. The problem has been thus reduced to the following one:

$$\begin{aligned} \min_x \quad & \sum_{k\in\mathcal{C}_0} \lambda_k C_k\left(x_k/\lambda_k\right) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{C}_0} m_k^e x_k = a, \\ & x\geq 0. \end{aligned} \tag{EC.62}$$

Following the solution of (10) (using the KKT conditions), we can define a new function, in analogy to $\widehat{\Delta}_{\mathcal{K}}(\cdot)$ from (EC.60) (but now with subscript $\mathcal{C}_0$), and under $\{\tilde{\pi}_{**}^r\}$, this function plays the role of a lifting mapping in the corresponding state-space collapse.

EC.8. Discussing the conjecture in §8.1: Adding delays between physician visits

In this section, we briefly discuss our conjecture on the duration of delays. An analysis of the infinite-server queue with fast service rate will be useful, which we provide at the end of the present section.

The ED system with delays between physician visits: Let $Q_{jk}^r(t)$ denote the number of patients in the delayed system between j-triage and k-IP patients at time t; similarly, $Q_{kl}^r(t)$ is the number of patients in the delayed system between the k-IP and l-IP patients at time t.

The number of k-IP patients at time t is

$$\begin{aligned} Q_k^r(t) &= Q_k^r(0)+\sum_{j\in\mathcal{J}}\left(\Phi_{jk}\left(S_j\left(T_j^r(t)\right)\right)+Q_{jk}^r(0)-Q_{jk}^r(t)\right) \\ &\quad +\sum_{l\in\mathcal{K}}\left(\Phi_{lk}\left(S_l\left(T_l^r(t)\right)\right)+Q_{lk}^r(0)-Q_{lk}^r(t)\right)-S_k\left(T_k^r(t)\right) \\ &= Q_k^r(0)+\sum_{j\in\mathcal{J}}\Phi_{jk}^r\left(S_j\left(T_j^r(t)\right)\right)+\sum_{l\in\mathcal{K}}\Phi_{lk}^r\left(S_l\left(T_l^r(t)\right)\right)-S_k\left(T_k^r(t)\right) \\ &\quad -\sum_{j\in\mathcal{J}}\left(Q_{jk}^r(t)-Q_{jk}^r(0)\right)-\sum_{l\in\mathcal{K}}\left(Q_{lk}^r(t)-Q_{lk}^r(0)\right), \qquad k\in\mathcal{K}. \end{aligned} \tag{EC.63}$$

Ignoring the changes of $T_j^r, j\in\mathcal{J}$ and $T_k^r, k\in\mathcal{K}$, the difference between (EC.63) and (EC.11) is $\sum_{j\in\mathcal{J}}\left(Q_{jk}^r(t)-Q_{jk}^r(0)\right)+\sum_{l\in\mathcal{K}}\left(Q_{lk}^r(t)-Q_{lk}^r(0)\right)$, which is the total change in the number of patients within the infinite-server queues that experience delays between services.

First we argue that the fluid limits of those $T_j^r, j\in\mathcal{J}$ and $T_k^r, k\in\mathcal{K}$, are the same as the fluid limits in the system without delays between physician visits. This, together with Random-time-change, ensures that the diffusion approximation of $Q_k^r(0)+\sum_{j\in\mathcal{J}}\Phi_{jk}^r\left(S_j\left(T_j^r(t)\right)\right)+\sum_{l\in\mathcal{K}}\Phi_{lk}^r\left(S_l\left(T_l^r(t)\right)\right)-S_k\left(T_k^r(t)\right)$ is the same as in the system without delays. It is enough to prove that the fluid limit of $\sum_{j\in\mathcal{J}}\left(Q_{jk}^r(t)-Q_{jk}^r(0)\right)+\sum_{l\in\mathcal{K}}\left(Q_{lk}^r(t)-Q_{lk}^r(0)\right)$ is 0. Indeed, if we have the latter fact, we can first argue that the fluid limit of $\sum_{j\in\mathcal{J}} m_j^e\bar{\bar{Q}}_j^r+\sum_{k\in\mathcal{K}} m_k^e\bar{\bar{Q}}_k^r$ equals that in the system without delays, and then follow the steps in §EC.2 to prove that the fluid limit for the busy time processes is also the same, namely these limits are $\lambda_j m_j t$, for $j \in \mathcal{J}$, and $\lambda_k m_k t$, for $k \in \mathcal{K}$.

We now prove that the fluid limit of $\sum_{j\in\mathcal{J}} \left(Q^r_{jk}(t) - Q^r_{jk}(0)\right) + \sum_{l\in\mathcal{K}} \left(Q^r_{lk}(t) - Q^r_{lk}(0)\right)$ is 0. Notice that the delayed queues are infinite-server queues and the arrival processes for these queueing systems are part of the departure process from the physician. We can then verify that the requirements for the fluid approximation of the $G/M/\infty$ with fast service rates (in §EC.8.1) hold, in particular the sequence of the fluid-scaled arrival processes is tight. As a result, those delayed queues remain constant in fluid scaling, meaning that the delays will have no impact on the fluid limit of the ED model. Hence the fluid limits of $T^r_j, j \in \mathcal{J}$, and $T^r_k, k \in \mathcal{K}$, remain constant.

Next we discuss the diffusion-scaled processes. From the differences between (EC.63) and (EC.11), to prove that $\sum_{j\in\mathcal{J}} m^e_j \widehat{Q}^r_j + \sum_{k\in\mathcal{K}} m^e_k \widehat{Q}^r_k$ is invariant to all work-conserving policies, it is enough to argue that the following holds, for each $k \in \mathcal{K}$:

$$\frac{1}{r}\left[\sum_{j\in\mathcal{J}} \left(Q^r_{jk}(r^2t) - Q^r_{jk}(0)\right) + \sum_{l\in\mathcal{K}} \left(Q^r_{lk}(r^2t) - Q^r_{lk}(0)\right)\right] \Rightarrow 0.$$

As those are infinite-server queues with fast service rates, from the discussion at the end of this section, it is enough to prove that the diffusion scaled arrival processes to the delayed queues are tight. This is a gap that we are leaving for future research.

EC.8.1. Infinite-server queues with fast service rates

Here we develop the fluid and diffusion approximation for a sequence of infinite-server queues with fast server rates, which are used in our conjecture on the duration of the delays. We will use the following analytical result.

From Lemma 3.4 of Atar and Solomon (2011), we know that for any given sequence of $x^n \in \mathcal{D}$, there are $y^n \in \mathcal{D}$ satisfying the following equation:

$$y^n(t) = x^n(t) - \mu^n \int_0^t y^n(s)ds. \tag{EC.64}$$

Furthermore, if $\mu^n \to \infty$ and the sequence of $\{x^n\}$ is tight with $x^n(0) \to 0$, then $y^n \to 0$. We shall use this result in the following discussion, to gain insight into infinite-server queues.

Consider a sequence of infinite-server queueing systems $G/M/\infty$. In the rth system, the arrival process is $E^r(\cdot)$, with individual service rate $\mu^r = \mu r^\alpha$, in which $\alpha > -2$.

We first establish fluid approximation. Assume that the fluid-scaled arrival processes $\bar{\bar{E}}^r$ are tight. Here

$$\bar{\bar{E}}^r(t) = r^{-2}E^r(r^2t).$$

Denote by S a unit rate Poisson process, with its fluid scaling $\bar{\bar{S}}^r(t) = r^{-2}(S(r^2t) - r^2t)$. Then the fluid-scaled queue length process $\bar{\bar{X}}^r = r^{-2}X^r(r^2t)$ can be represented as

$$\bar{\bar{X}}^r(t) = \bar{\bar{X}}^r(0) + \bar{\bar{E}}^r(t) - \bar{\bar{S}}^r\left(\mu r^{2+\alpha}\int_0^t \bar{\bar{X}}^r(s)ds\right) - \mu r^{2+\alpha}\int_0^t \bar{\bar{X}}^r(s)ds.$$

Fix a $T>0$, and assume that there is $M>0$ such that $\limsup_{r\to\infty}\bar{\bar{E}}^r(T)<M/2$. Define a sequence of stopping times (indexed by r) via

$$\sigma^r=\inf\left\{t>0,\ \mu r^{2+\alpha}\int_0^t \bar{\bar{X}}^r(s)ds>M\right\}\wedge T.$$

Using (EC.64), if $\bar{\bar{X}}^r(0)\Rightarrow 0$, then one can show that $\bar{\bar{X}}^r(\sigma^r\wedge\cdot)\Rightarrow 0$. Following the proof of (39) in Atar and Solomon (2011), we can also prove $\sigma^r\Rightarrow T$. As a result, $\bar{\bar{X}}^r\Rightarrow 0$ on $[0,T]$. As this T is arbitrary, we have $\bar{\bar{X}}^r\Rightarrow 0$ on $[0,\infty)$.

Now we develop the diffusion approximation. For the above sequence of $G/M/\infty$ systems, fix a sequence of $\{\lambda^r\}$, and denote $\widehat{X}^r(t)=r^{-1}(X^r(r^2t)-\lambda^r/\mu^r)$, as well as

$$\widehat{E}^r(t)=r^{-1}(E^r(r^2t)-\lambda^r r^2 t),\quad\text{and}\quad \widehat{S}^r(t)=r^{-1}(S(r^2t)-r^2t).$$

We then have

$$\widehat{X}^r(t)=\widehat{X}^r(0)+\widehat{E}^r(t)-\widehat{S}^r\left(\mu r^{2+\alpha}\int_0^t\bar{\bar{X}}^r(s)ds\right)-\mu r^{2+\alpha}\int_0^t\widehat{X}^r(s)ds.$$

Suppose that there is a sequence of $\{\lambda^r\}$ with (i) $\lambda^r\to\lambda$, for some $\lambda>0$, (ii) $\widehat{X}^r(0)\Rightarrow 0$, and (iii) making $\{\widehat{E}^r\}$ tight. Then, from the fluid limit argument, we can prove that $\widehat{S}^r\left(\mu r^{2+\alpha}\int_0^t\bar{\bar{X}}^r(s)ds\right)$ converge to a driftless Brownian motion with variance λ; using (EC.64), we can now deduce that $\widehat{X}^r(\cdot)\Rightarrow 0$.

EC.9. More simulation outputs

We provide here more simulation results that complement §7.

EC.9.1. Descriptions of the other three policies

We start with descriptions of the alternative three policies, used for comparison in §7.

- FCFS: The patients are served on a global First-Come-First-Served basis, as in Dai and Kurtz (1995) and Reiman (1988);
- IP-patients-First (IPF): Priority is always given to IP patients if there are any. Among all triage classes, one determines the priority according to the Shortest-Deadline-First rule (11), while among all IP classes, the priority is according to the modified generalized $c\mu$ rule;
- Triage-patients-First (TrF): Priority is always given to triage patients if there are any. Among all triage classes, one determines the priority according to the Shortest-Deadline-First rule (11), while among all IP classes, the priority is according to the modified generalized $c\mu$ rule.

Table EC.1 EDs under different durations of delays

Duration	P_1	P_2	P_3	Cost Rate	LoS
No delay	4.61% (0.10%)	4.57% (0.09%)	4.57% (0.09%)	125.21 (10.36)	68.96
1 minute	4.46% (0.11%)	4.45% (0.10%)	4.43% (0.10%)	133.38 (10.96)	73.17
10 minutes	4.62% (0.11%)	4.47% (0.10%)	4.46% (0.11%)	132.80 (10.55)	93.59
60 minutes	5.44% (0.09%)	5.00% (0.09%)	4.89% (0.09%)	138.46 (9.73)	204.42
120 minutes	5.80% (0.10%)	5.35% (0.10%)	5.15% (0.10%)	141.60 (11.79)	335.23

EC.9.2. On the duration of delays between physician visits

We now incorporate in the simulations delays between successive visits to physicians. These delays model services and waiting times beyond physicians (e.g., X-ray, lab tests). We consider the following delays (all in minutes): 0, 1, 10, 60 and 120. (Delays are assumed exponentially distributed). Other parameters are identical to those in §7.1. The performance measures of our TG$c\mu$ policy are shown in Table EC.1.

In the table, we observe small changes in P_j and cost, over delays between visits that range from the very short up to 120 minutes. For a better grasp of the effects of delays, we also exhibit Length of Stay (LoS, or sojourn time), which predictably increase as the delays increase.

EC.9.3. Simulating a time-varying ED with delays

In this subsection, we present simulation results for an ED with time-varying arrival rates, as well as delays between successive visits to physicians.

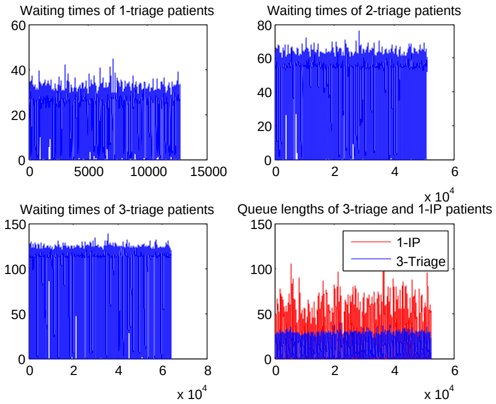
Daily arrival rates are time-varying, as in Figure 2, such that the average total arrivals of 1-triage, 2-triage and 3-triage patients per day is $14 * 24 = 336$. We further assume constant arrival rates *per hour*, which are then given by 9.13, 7.00, 4.72, 5.31, 3.77, 2.71, 3.29, 5.09, 10.61, 17.51, 22.76, 24.51, 21.81, 20.16, 20.43, 18.36, 16.66, 17.88, 19.90, 20.80, 19.58, 17.77, 14.43, 11.83. Service rates do not vary with time. Then the traffic intensity varies from 0.1839 to 1.6663.

Assume also constant transition probabilities, with delays between successive physician visits. Delay duration may depend on the class. Table 2 in Yom-Tov and Mandelbaum (2014) summarizes the duration of delays between physician visits for different classes. From the table, we conclude that 60 minutes is reasonable for the average duration of delays. We thus model the delays as infinite-server queues (with exponential service times), all with 60 minutes as their average service times.

In time-varying environments, we recommend a slight change in ϵ of our proposed policy. Specifically, our simulations gave rise to the following rule: assign priority to triage classes if $\tau_1(t) > d_1 - 4$, or $\tau_2(t) > d_2 - 6$, or $\tau_3(t) > d_3 - 8$ (in order to achieve at most 5% violations of delay). Our theory can be easily modified to accommodate different ϵ_j's, and all remains intact. Other parameters, such as distributions and cost rates, are assumed equal to those in the system without delays and with constant arrival rates, as in §7.1.

Similarly to the stationary model in §7, we plot a representative sample path of the system under the proposed policy, with histograms of the triage waiting times for service. We then compare our proposed policy against the 3 alternatives FCFS, IPF and TrF (Table EC.2).

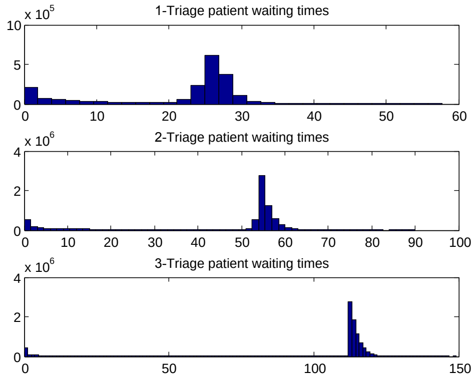
Figure EC.1 A sample path of the time-varying system under the proposed policy



In Figure EC.1, we observe a phenomenon similar to Figure 3. After summarizing all 160 sample paths, the fraction of triage patients who violate their corresponding deadlines are 4.54%, 3.14% and 2.64%, respectively; the fractions who violate their deadlines by more than 10% of their corresponding deadlines are negligible (less than 1%). As the system is overloaded most times (load is often above 1.2 and can be as high as 1.6663), this suggests that our proposed policy would also work well in overloaded systems.

Figure EC.2 includes the histograms of triage waiting times.

Figure EC.2 Histogram of triage waiting times in the time-varying system under our proposed policy



We now compare our policy with the three alternatives, as done in the stationary case (§7). The performances of these four policies appear in Table EC.2.

From Table EC.2, we note that the TGcμ policy performs reasonably well. The cost rate in the IP-patients-First (IPF) is very small, but a very large fraction of triage patients violate their deadlines. The same problem exists for FCFS policy. Triage-patients-First (TrF) policy does ensure that triage patients adhere to their deadlines, but its cost rate is about 2 times the TGcμ policy. In summary, our proposed policy TGcμ clearly outperforms its competitors.

Table EC.2 Comparison of the four policies

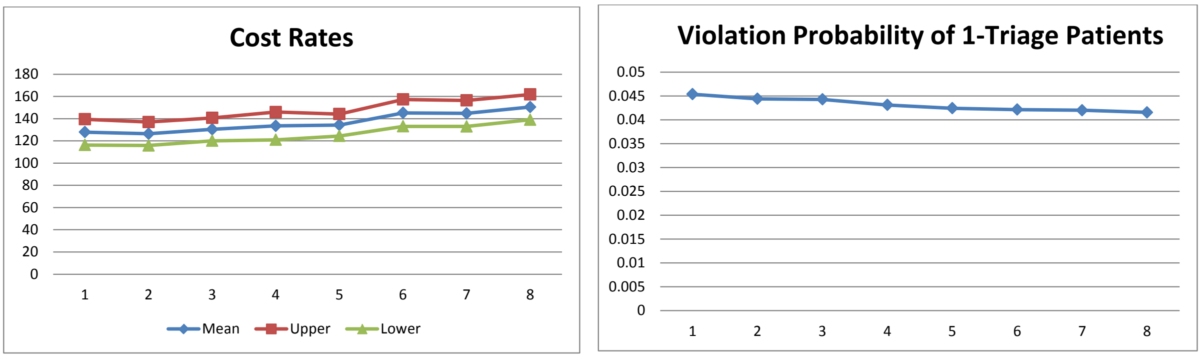
Policy	P_1	P_2	P_3	Cost Rate	LoS
TG$c\mu$	4.44% (0.04%)	3.21% (0.02%)	2.75% (0.02%)	1561.89 (40.02)	368.64
FCFS	76.29% (0.19%)	56.81% (0.28%)	17.63% (0.31%)	1160.82 (15.72)	371.93
IPF	68.95% (0.21%)	69.82% (0.21%)	74.04% (0.20%)	7.33 (0.01)	305.42
TrF	0.00% (0.00%)	0.00% (0.00%)	0.00% (0.00%)	3251.88 (44.39)	412.86

EC.9.4. Multiple physicians

In our theoretical analysis, we argued that a multiple-server system is asymptotically equivalent to a single-server system. This is theoretically true due to conventional heavy-traffic theory; see, e.g., Chen and Shanthikumar (1994). In this subsection, we use simulation to support this claim.

We keep the arrival rates, transition probabilities and cost functions the same as in §7.1. We vary the number of servers from 1 to 8 and denote the number of servers by i. As discussed in §7.1, the single super-server, with mean service time 1.3 minutes, is a combination of the i physicians. Then the mean service time in a system with i physicians is $1.3i$ minutes. We simulate the systems under our proposed policy. The performance metrics are plotted in Figure EC.3.

Figure EC.3 System performances under our proposed policy, but with multiple-servers (one to eight)



The X-axes in both figures of Figure EC.3 represent the number of physicians. The left figure shows cost rates (mean and the upper and lower boundaries of the 95% confidence interval), while the right one shows the fraction of 1-triage patients who violate their triage deadlines. The violation probabilities of the other two triage classes are close to those of 1-triage patients, hence we omit them here.

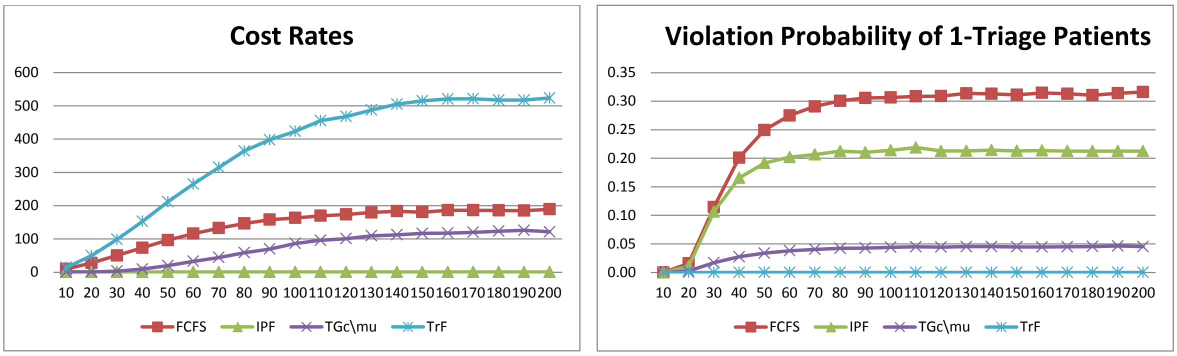
In Figure EC.3, as the number of physicians increases, the cost rate increases, while the violation probability decreases. This can be explained as follows: with more physicians, patients have a greater likelihood to start their treatments earlier (hence the violation probability of triage patients decreases), but are less likely to complete their treatments as the service time increases and more physician capacity is shared by other classes (hence the cost rate of IP patients increases). The changes are not significant, which confirms the claim that in conventional heavy traffic, the number of physicians does not matter (in diffusion scaling).

EC.9.5. Finite ED capacities

Here we investigate the impact of finite ED capacity on system performances, under different policies, as discussed in §8.3.

The parameters are identical to those in §7.1, except that the ED capacity is now finite and varies from 10 to 200. That is, the emergency department is modeled as a system with finite waiting capacity. When a triage patient arrives at the ED and finds that the waiting room is full, the patient is blocked. For those patients already entering the system, their occupied beds are not released between class transfers. As a result, no IP-patient is blocked. We simulate the system under the four policies, and the performance metrics are plotted in Figure EC.4.

Figure EC.4 System performances under different policies and different ED capacities

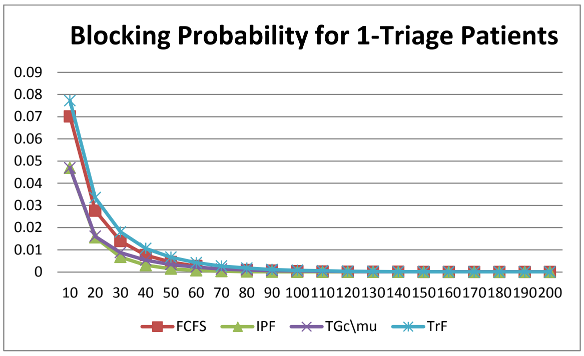


The X-axes in both figures of Figure EC.4 represent the ED capacities. The left figure shows the cost rates under different policies, while the right one shows the violation probability of 1-triage patients. Here we make the following observations:

- For systems with moderate to large ED capacities (30 to 200), our proposed (TG$c\mu$) policy and the Triage-patients-First (TrF) policy can ensure that most triage patients (more than 95%) adhere to their deadlines, but the TrF policy incurs much larger cost rates. A large proportion of 1-triage patients under the FCFS and IP-patients-First (IPF) policies cannot meet their deadline constraints. As a result, our proposed policy outperforms the other three alternatives for systems with moderate to large sizes.
- For systems with small ED capacities (10 to 20), all 4 policies can ensure that most triage patients (more than 95%) adhere to their deadline constraints. The cost rates incurred by our proposed policy and the IPF policy are the smallest.
- As ED capacity decreases, the cost rates and the violation probabilities decrease. The changes in violation probabilities are non-negligible when ED capacity is smaller than 60. Due to the quadratic cost rate functions, the cost rates decrease a little faster. This is because long queues have a significant impact on quadratic costs, and finite ED capacity can reduce the probability of such long queues.

We plot the blocking probabilities for 1-triage patients in Figure EC.5. The blocking probabilities for the other two triage classes are almost the same; thus Figure EC.5 can also be viewed as approximations for the total blocking probabilities.

Figure EC.5 Blocking probabilities for 1-triage patients under different policies and different ED capacities



From Figure EC.5, IPF policy incurs the smallest blocking probabilities, and the blocking probabilities of our proposed policy (TG$c\mu$) are almost the same as those of the IPF policy. This suggests that our proposed policy may also be able to reduce blocking probability for systems with finite ED capacities.

EC.9.6. Adding abandonment (LWBS and LAMA)

As discussed in §8.6, patients in EDs may leave before completing all desired treatments (LWBS or LAMA). This can be modeled as customer abandonment. In this subsection, we use simulation to investigate the impact of abandonment on system performance under the four policies.

ED parameters are kept the same as those in §7.1, except that patients waiting for service, at any phase of their ED process, may abandon. To be specific, each patient has a patience time when they join a queue, and a new patience time starts after transfer to another queue. If the patience time expires when a patient waits in the queue, that patient leaves the ED without further treatments and will never return. The patience times are assumed to be exponential with the same means for all classes. We denote this common mean by $1/\theta$ and call θ the *individual* abandonment rate. For concreteness, we control the probability of abandonment (LWBS+LAMA) below 4%, which corresponds to abandonment rate θ varying from 10^{-3} to 10^{-6}; this is the same as average patience varying between 16 hours and infinity (and it is in concert with Wiler et al. (2013), who report that 4.1% LWBS, having 10.68 (±7.76) hours of average patience). (Note that the LWBS+LAMA in §7.4 is around 7%. This is higher than the 4% here since, in §7.4, one is facing a higher control challenge: arrival rates are time-varying, and the system is overloaded during a significant part of the day.)

We simulate the systems under the four policies, and their performance metrics are plotted in Figure EC.6 (congestion cost rates and violation probability) and Figure EC.7 (probability of abandonment). The X-axes in both figures of Figure EC.6 and the one in Figure EC.7

represent abandonment rates. The left figure in Figure EC.6 shows the cost rates under the four policies, while the right one shows the violation probability of 1-triage patients. Here are some observations from the figures:

Figure EC.6 System performances under different policies and different abandonment rates

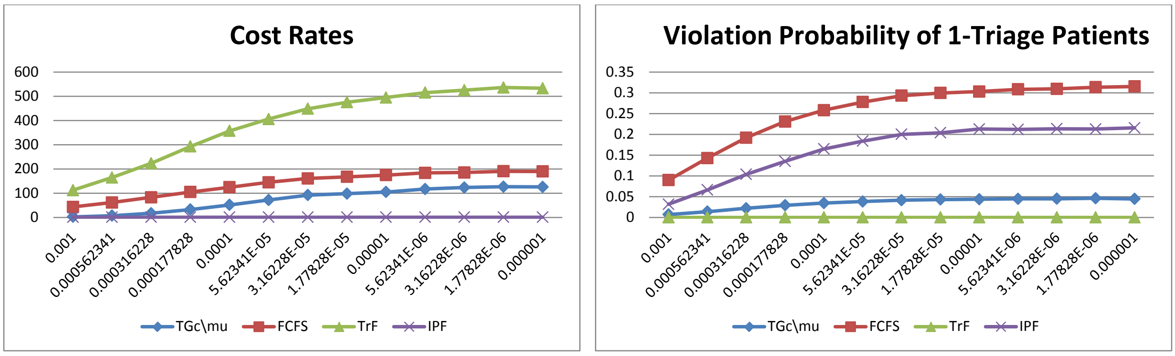
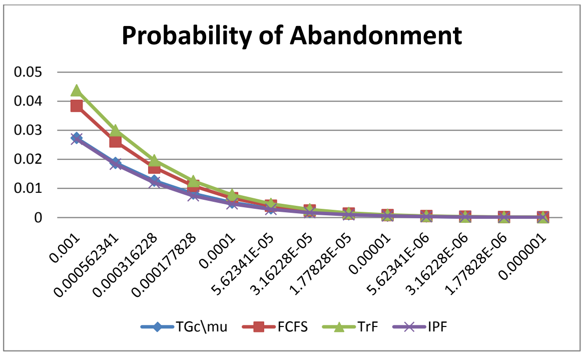


Figure EC.7 Probabilities of abandonment under different policies and different abandonment rates



- Our proposed (TG$c\mu$) policy and the Triage-patients-First (TrF) policy can ensure that most triage patients (more than 95%) meet their deadlines, but the TrF policy incurs a much higher cost. Generally, a significant fraction of 1-triage patients under the FCFS and the IPF policies cannot adhere to their deadline constraints. Consequently, our proposed policy outperforms the other three policies.

- As the abandonment rate increases, the cost rates and the violation probabilities decrease. Due to the quadratic cost rate functions, the cost rates decrease somewhat faster. The explanation is the same as for blocking probabilities when ED capacity is finite: long queues have a significant impact on quadratic costs, and customer abandonment can reduce the probability of such long queues.

- An essentially infinite patience corresponds to LWBS+LAMA less than 1%. This is not implausible for an ED reality where patients do need emergency care (excluding perhaps triage class 5 patients). Alternatively, and as already mentioned, average patience less than 16 hours gives rise to abandonment that exceeds 4%. In this case, 3 out of the 4 policies perform well, in terms of both violation probability (less than 5%) and congestion costs (negligible). The cost rate of the IPF policy is the smallest, and more than 95% patients adhere to their deadlines. Our proposed policy ensures that more than 99% of the triage patients meet their deadlines, and the cost rate is comparable to the one under IPF.

- Systems under IPF policy and our proposed (TG$c\mu$) policy enjoy the smallest probabilities of abandonment. (IPF has a slightly smaller abandonment probability, but the differences with our policy are negligible.) This suggests that our proposed policy would fare well also against LWBS+LAMA.

References

Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, G. B. Yom-Tov. 2013. Patient flow in hospitals: A data-based queueing-science perspective. *Working Paper.*

Atar, R., N. Solomon. 2011. Asymptotically optimal interruptible service policies for scheduling jobs in a diffusion regime with non-degenerate slowdown. *Queueing Systems.* **69**(3) 217–235.

Bramson, M. 1998. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems.* **30**(1-2) 89–140.

Chen, H., J. G. Shanthikumar. 1994. Fluid limits and diffusion approximations for networks of multi-server queues in heavy traffic. *Discrete Event Dynamic Systems.* **4**(3) 269–291.

Chen, H., D. D. Yao. 2001. *Fundamentals of queuing networks: Performance, asymptotics, and optimization.* Springer-Verlag.

Dai, J. G., T. G. Kurtz. 1995. A multiclass station with Markovian feedback in heavy traffic. *Mathematics of Operations Research.* **20**(3) 721–742.

Plambeck, E., S. Kumar, J. M. Harrison. 2001. A multiclass queue in heavy traffic with throughput time constraints: Asymptotically optimal dynamic controls. *Queueing Systems.* **39**(1) 23–54.

Reiman, M. I. 1984. Open queueing networks in heavy traffic. *Mathematics of Operations Research.* **9**(3) 441–458.

Reiman, M. I. 1988. A multiclass feedback queue in heavy traffic. *Advances in Applied Probability.* **20**(1) 179–207.

van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Annals of Applied Probability.* **5**(3) 809–833.

Wiler, J. L., E. Bolandifar, R. T. Griffey, R. F. Poirier, T. Olsen. 2013. An Emergency Department Patient Flow Model Based on Queueing Theory Principles. *Academic Emergency Medicine.* **20**(9) 939–946.

Yom-Tov, G. B., A. Mandelbaum. 2014. Erlang-R: A time-varying queue with ReEntrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management.* **16**(2) 283-299.