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 multi-class 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 violate the deadline; alternatively, IP patients are served according to a $G_{c\mu}$ rule, in which $\mu$ 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) 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

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 (2013), Table 2).

Thus, IP patients impose operational congestion (e.g. they occupy beds), which must be controlled while adhering to clinical triage constraints (e.g. stabilizing patient conditions). It is this operational-clinical friction that we focus on, 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, we propose a flow control policy that minimizes congestion costs subject to deadline constraints, doing so under the prevalent conditions of ED heavy-traffic.

We consider two models in this paper, which differ by their congestion costs. In the first (basic) model (§1.1–1.2), a waiting IP patient incurs cost at a rate that depends on the present queue length (see (6)). In the second model (§1.3), the cost is a function of present sojourn time (measured since becoming an IP patient). In the spirit of Baron et. al. (2013), the latter model takes the traditional view of overall experience (sojourn time). The basic model, on the other hand, accounts for individual waiting experience.

Our mathematical framework is conventional heavy-traffic, in which one analyzes a sequence of systems that converge to critical loading. This is a relevant operational regime, despite the fact that EDs are inherently time-varying. Specifically, our experience suggests that, during regular peak shifts between late morning till late evening, the ED can be usefully viewed as a critically-loaded stationary system (Armony et al. (2013)). Moreover, simulation experiments (§7 and EC.16.5) demonstrate that our proposed policies actually perform well over the whole (time-varying) day.

Within our asymptotic framework, the in-process analysis follows the Gcµ-rule of van Mieghem (1995), after generalizing it to models with feedback. The 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. We now continue with describing such policies for our two IP models—individual queueing and cumulative sojourn time costs.

1.1. A basic ED model and its flow control—queue-dependent cost rates

ED dynamics are captured by a multiclass queueing system, with $N$ servers (physicians), $J$ classes of triage patients and $K$ classes of in-process (IP) patients. Triage patients are yet to be examined by a physician, and in-process (IP) patients require further treatment. (A patient class could embody information such as treatment type, emergency level or age; see Carmeli (2012).) The system is depicted in the following figure:
The $J$ classes of triage patients are subject to deadline constraints, and the $K$ classes of IP patients incur queueing costs. Patients within each class are served on a First-Come-First-Served (FCFS) basis. Denote by $j \in J$ and $l, k \in K$ the class indices for triage patients and IP patients, respectively.

We now introduce the minimal notation that suffices for describing our flow control problem and its solution. The model specification is completed later in §2. The $j$-triage patients arrive to the system exogenously; each such patient must be served within a deadline of $d_j$ time units from its arrival time. Formally, a $j$-triage patient arriving to the system at time $t$ must start service before time $t + d_j$; equivalently, $\tau_j(t) \leq d_j$, for all $j \in J$ and $t \geq 0$, where $\tau_j(t)$ is the age of the head-of-the-line $j$-triage patient at time $t$. Note that $\tau_j(t)$ is random and $d_j$ is deterministic; hence consistently satisfying this last deadline constraint is too much to hope for, which calls for a rigorous formulation in an asymptotic sense (later in §4).

After completing their first service, $j$-triage patients join the queue of $k$-IP patients, or exit the system. Turning attention to IP patients, they originate from either triage patients or from other IP patients. While waiting, the $k$-IP patients incur queueing costs at rate $C_k(Q_k(t))$; here $C_k$ is an increasing convex function ($C_k(0) = 0$) and $Q_k(t)$ is the queue length of $k$-IP patients at time $t$. (We also offer an alternative formulation of costs that depend on waiting times, in §6.3.) Our objective is to minimize the cumulative queueing costs (alternatively waiting costs) incurred by IP patients, among all policies that satisfy the deadline constraints.

To formulate our proposed flow control, let $m^e_j$ denote the mean effective service time of $j$-triage patients: this is the expected total service time, required by $j$-triage patients throughout their ED stay (see (3) for its calculation). The traffic intensity is then

$$\rho = \frac{1}{N} \sum_{j \in J} \lambda_j m^e_j,$$

and we think in terms of $\rho \approx 1$ (ED in heavy-traffic). Now let $m^e_k$ denote the mean effective service time of $k$-IP patients: this is the expected total remaining service time, accumulated from becoming a $k$-IP patient till departing from the ED (see (4) for its calculation).
The notation has been now set for describing our flow control policies. A physician that becomes idle at time $t$ adopts the following guidelines:

- **Triage or IP:** Give priority to triage patients if there exists a $j \in J$ such that $	au_j(t) \geq d_j - \epsilon$, where $\epsilon$ is small relative to the smallest $d_j$. (Our theory suggests, and our simulations confirm, that $\epsilon$ can be chosen one order of magnitude smaller than $d_j$. For example, with $\min_{j \in J} d_j = 30$ minutes, one can use $\epsilon = 3$ 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

$$j \in \arg\min_{j \in J} [d_j - \tau_j(t)].$$

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

$$k \in \arg\max_{k \in K} \frac{C'_k(Q_k(t))}{m_k^e}.$$

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

Within a suitable heavy traffic framework (Section 3), the above policy is asymptotically “feasible” and asymptotically optimal among all asymptotically “feasible” policies. The simplicity of our asymptotically optimal policies, as well as state-space collapse and snapshot properties that it enjoys (Theorem 3 and Proposition 3), are all due to the fact that heavy-traffic analysis exposes macroscopic and mesoscopic essentials, which is formalized by fluid and diffusion approximations (§EC.4). An example of such a property is that the following three systems are (asymptotically) equivalent in heavy traffic: the $N$-physician system in Figure 1; 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 (proved as in Chen and Shanthikumar (1994)); and a system of $N$ i.i.d. physicians, who are non-interchangeable (conjectured and discussed in §8.5) in the sense that IP patients must remain with the same physician all throughout their ED visit. We thus assume hereafter that $N = 1$.

**Non-unique optima:** Under the relative crudeness of heavy-traffic dynamics, there are other policies that emerge as asymptotically optimal (Section 6). For example, the decision of triage vs. IP can be formulated in terms of only one class, say class $1 \in J$; that is, give priority to triage patients if $\tau_1(t) \geq d_1 - \epsilon$. It can also be formulated in terms of a threshold $\omega = \sum_{j \in J} \lambda_j d_j m_j^e$; denote by $Q_j(t)$ the queue length of $j$-triage patients; then if $\sum_{j \in J} m_j^e Q_j(t) \geq \omega$, a server just becoming idle caters to triage patients, otherwise to IP patients. Furthermore, triage classes can be alternatively chosen, similarly to Plambeck et al. (2001) and van Mieghem (2003), that is, serve $j \in \arg\max_{j \in J} \frac{\tau_j(t)}{d_j};$ and the selection criterion of IP-classes can also be any rule that satisfies (12), in particular the one conjectured on page 853 of Mandelbaum and Stolyar (2004). While all the above options are asymptotically optimal, they do differ in their rate of convergence. This is manifested in simulation experiments that we carried out, which we describe in §7 and §EC.16. The above recommended policy performed best in these experiments.
1.2. Intuition

The idea is first to maximize service effort for IP patients which, given the server’s fixed capacity, is the same as minimizing it for triage patients subject to adhering to their deadline constraints; then one allocates the service capacity to IP patients to greedily minimize the queueing cost rate. This is a reasonable approach since service capacity is assumed to be close to the arriving workload. As a result, in our critically loaded (heavy traffic) system, there is enough 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.16 for further discussion.)

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

$$\sum_{j \in J} m_j^e Q_j(t) + \sum_{k \in K} m_k^e Q_k(t),$$

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 $\left(t - \tau_j(t), t\right]$, it must be that $Q_j(t) \approx \lambda_j \tau_j(t)$ and the workload equals approximately

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

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

**Triage vs. IP:** By the deadline constraints, an upper bound for $\sum_{j \in J} m_j^e \lambda_j \tau_j(t)$ is $\omega = \sum_{j \in J} \lambda_j d_j m_j^e$, and our policy should strive 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. An example is $\frac{\tau_j(t)}{d_j} \approx \frac{\tau_{j'}(t)}{d_j'}$, for any $j, j' \in J$, at all times $t$, which implies that the age of any one triage class tells 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.) We discuss this rule in Theorem 1. The Shortest-Deadline-First rule, with more complicated notation (§EC.11), is discussed in §6.1. Simulations show that both rules perform well, and the one with the shortest-deadline-first rule slightly better.
**IP selection:** After applying the threshold guideline and the triage selection rule, one expects that \( \sum_{k \in \mathcal{K}} m_k Q_k(t) \) 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 \( \mu \)'s replaced by \( 1/m_k^\epsilon \) 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, which is also manifested through some congestion laws that will now be described.

**A 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 now 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 \( \omega_k \) are then related via \( Q_k(t) \approx \lambda_k \omega_k(t) \), with \( \lambda_k \) 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).

### 1.3. An alternative ED model—cost of IP sojourn times

Our alternative ED model differs in its IP congestion costs. To be specific, the model is the same as in Figure 1, except that the cost now depends on the total time spent within the ED.
The problem is to minimize sojourn time costs, incurred by all patients who arrived to the ED within a finite horizon, while again adhering to triage constraints.

For our analysis of sojourn time costs, we make the additional assumption that the transition matrix $P$ is upper-triangular; this is needed for our method of proof but it is practically unrestrictive, at the possible cost of some class proliferation. Formally, a triage patient, turning first into a $k$-IP patient and ultimately spending $W$ time units in the ED, incurs congestion cost $C_k(W)$; here $C_k(\cdot)$ is a convex increasing function (which differs from those in the previous section).

For choosing between triage vs. IP patients, and selecting a specific triage patient to be served, our proposed asymptotically optimal policies are the same as before. The rule for choosing which IP class to serve is modified, however: one assigns priority to those patients who have already received at least one IP treatment; any remaining service capacity is then allocated to the new IP patients, according to our modified $Gc\mu$ rule—see (22). Note that such a service policy is not FCFS within classes: indeed, if patients in an IP class can originate from both triage and IP patients, priority must be given to the latter. It follows that, even under Markovian routing, one must record the class-history of each patient. We do so by further enlarging the set of classes, having IP patients follow one of finitely-many disjoint, deterministic IP routes (disjoint means different routes consist of different classes); this, in practice, might require predictions of class designations—more on that in the following discussions.

**Congestion laws:** Similarly to our queue-dependent cost model, and assuming the above class designation, the snapshot principle also prevails for the model with IP cost per sojourn time, under our proposed policy. The snapshot principle then implies a sample-path version of Little’s law (Reed and Ward (2008)): the relation between waiting time and queue length for any starting class, where the former is asymptotically identical to the age of head-of-the-line patient in that class. Moreover, the overall IP sojourn time is approximately the waiting time in the corresponding starting class, since higher priority is given to the subsequent classes. Thus, a predictor for the sojourn time of a patient, who is starting the IP process in class $k$, would be simply the age of the head-of-the-line patient in class $k$.

**The value of information:** In the Appendix (§EC.15), we apply our sojourn time framework, with the expert-elicited sojourn time costs from Carmeli (2012), to support analysis of the value of information in ED flow-control. Specifically, we show that accurate prediction of both the number of visits to a physician and whether a patient will be hospitalized or discharged, reduces IP congestion cost by as much as 27%. From our ED sources (Barak-Corren et al. (2013)), and supported by Saghaian et al. (2011, 2012), we learn that such predictions can be accurately made and, hence, are worth taking into account.

**Beyond our two ED models:** Saghaian 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 Against-Medical-Advice (LAMA). We comment on these features, and offer related conjectures, in Section 8.

1.4. Literature review and contributions

There is ample medical literature about triage systems, to which we refer the reader through Farrohnia et al. (2011), Mace and Mayer (2008). Our research focus here is operational (Mar- mor et al. (2012)) and, accordingly, so is the following literature review.

To the best of our knowledge, our paper is the first to analyze control of patient flow in an ED from a queueing-theory perspective. (In contrast, there are practically hundreds of simulation-based studies; see Brailsford et al. (2009).) Since starting this project, additional work has appeared on ED operations. The closest to ours are Saghafian et al. (2011, 2012): Saghafian et al. (2011) 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. We return to Saghafian et al. (2011, 2012) in §EC.15, where we analyze the value of the information that virtual-streaming requires.

There are additional papers that cater to specific ED characteristics: Yom-Tov and Mandelbaum (2013) 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. (2012) develop an overloaded queueing network to analyze the impact of interruptions on ED throughput; and Atar et al. (2012) address synchronization of ED activities (e.g. interpretations of a blood-test and X-ray imaging must precede a visit to the ED physician), by analyzing a fork-join queueing network in heavy-traffic.

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 multi-class 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 multi-class fluid network with feedbacks.

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 be also 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)).

To summarize, we view our main contributions to be the following:

- **Methodological.** We analyze multiclass queueing systems with feedback, in particular:
  1. Proving the conjecture in Mandelbaum and Stolyar (2004) regarding feedback, and improving upon it by identifying simpler asymptotically optimal policies;
  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 and analysis of the value of information.

Additional references are provided in Section 8, where we propose generalizations to our main models, accompanied with 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 in our basic model. We conclude with a discussion of future research directions in §8. The proofs for the main theorems, as well as additional proofs (for propositions) and complements are provided in the Appendix. Our alternative ED model, with sojourn time costs, is applied in §EC.15 to analyze the value of information, using data from an Israeli ED, and expert-elicited costs.

1.5. Notation

We use the standard notation $\mathbb{R}_+$ to denote the set of nonnegative real numbers. For a real number $x$, $[x]$ 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' 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 $M$, $M_j$ denotes the $j$th row, and $M_k$ the $k$th 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 $(\Omega, \mathcal{F}, P)$. Expectation with respect to $P$ is $E$. 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. Similar 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 context; most times it will be denoted by $j \in J$ or $k \in K$ or both. Finally, $e(\cdot)$ is the 1-dimensional identity function on $\mathbb{R}_+$, where $e(t) = t$, $t \geq 0$.

2. The basic model

Consider a single-server queueing system: It constitutes $J$ classes of triage customers subject 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 $J$ and $K$ denote the index sets of triage and IP patients, respectively: $j \in J$ is an index for triage patients, and $l, k \in K$ are indices for IP patients. It will be convenient to let $J = \{1, 2, \ldots, J\}$ and $K = \{1, 2, \ldots, K\}$, while keeping in mind that the indices 1, 2, \ldots in $J$ differ from those in $K$. To avoid ambiguity, we do write $j \in J$ and $l, k \in K$ as necessary.
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 \( \lambda_j \) 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 \( i \)th arrivals is \( u_j(i)/\lambda_j \) and the service time required for the \( i \)th patient is \( m_j v_j(i) \). As a result, \( \lambda_j \) is the arrival rate and \( m_j \) is the mean service time requirement of a \( j \)-triage patient. We assume \( \lambda_j > 0 \) for all \( j \in \mathcal{J} \) and use \( \Lambda_\mathcal{J} \) to denote 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 if the physician has devoted \( t \) time units to \( j \)-triage patients. Denote \( \mu_j = 1/m_j \), 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 \( P_{JK} = (P_{jk})_{J \times K} \) be the triage-to-IP matrix. We use \( \phi_j(n) \) to denote the indicator function recording the class that the \( n \)th \( j \)-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 \( t \). Just like triage patients, for each class \( k \in \mathcal{K} \), we are given a sequence of i.i.d. random variables \( \{v_k(i), i = 1, 2, \ldots\} \) 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 \( i \)th patient receiving service is \( m_k v_k(i) \). (Unless specified, 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 if the physician has devoted $t$ time units to $k$-IP patients. Denote $\mu_k = 1/m_k$, 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 $1 - \sum_{k \in K} P_{lk}$. 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. Let $\phi_l(n)$ be the indicator function, showing which class the $n$th served $l$-IP patient will transfer to; that is, the $n$th $l$-IP patient finishing service will go to 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 $\mathbb{P}(\phi_l(n) = 0) = 1 - \sum_{k \in K} P_{lk}$.

**Remark 1** 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 ages, 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 6.3 and 6.4). We shall then assume FCFS explicitly as needed.

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

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

$$\Lambda^T = (\Lambda_{J^T}) P_{JK} + \Lambda^T P.$$ \hspace{1cm} (1)

Then $\Lambda$ is given by

$$\Lambda^T = (\Lambda_{J^T}) P_{JK} (I - P)^{-1}.$$ \hspace{1cm} (2)

Define $M_{\mathcal{J}} = (m_j)_{j \in \mathcal{J}}$ by

$$M_{\mathcal{J}} = M_{\mathcal{J}} + P_{JK} (I - P)^{-1} M,$$ \hspace{1cm} (3)

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

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

where $m_k$ is called the effective mean service time of $k$-IP patients. Then (3) can be written as

$$M_{\mathcal{J}} = M_{\mathcal{J}} + P_{JK} M^e.$$ \hspace{1cm} (5)

We refer to $m_j$ as “effective” because it is the expected total service requirement of a $j$-triage patient, accumulated up to leaving the system (and similarly for $m_k$).
2.3. An infeasible problem

Service goals for triage and IP patients are different:

- **Triage patients facing deadlines**: 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 $\tau_j(t) \leq d_j$, for $j \in J$ at all $t \geq 0$.

- **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 functions $C_k, k \in K$. Consequently, the total cost will be incurred at rate $\sum_{k \in K} C_k(Q_k(t))$.

A control policy is defined as $\pi = \{T_j, j \in J; T_k, k \in K\}$, in which $T_j(t), j \in J$, and $T_k(t), k \in K$, 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 $T \geq 0$:

$$\min_\Pi \int_0^T \sum_{k \in K} C_k(Q_k(s)) ds$$

s.t. $\tau_j(t) \leq d_j$, $\forall j \in J$ and $0 \leq t \leq T$.  \hspace{1cm} (6)

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

The problem above is clearly infeasible, as the age processes $\tau_j(\cdot), j \in J$, 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 the physician always gives priority to IP patients, this reduces the 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.

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 $r$th system. Then, in the $r$th system, the arrival rate of $j$-triage class is $\lambda_{jr}$ and the effective arrival rate for $k$-IP class is $\lambda_{rk}$. The deadline for $j$-triage patients is $d_j$, 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 $r$th system is defined to be
\[ \rho^r := \sum_{j \in J} \lambda^r_j m_j + \sum_{k \in K} \lambda^r_k m_k. \]

By (2) and (3), it can also be represented as
\[ \rho^r = \sum_{j \in J} \lambda^r_j m^e_j. \]

This underscores the meaning of $m^e_j$ being the effective mean service time for $j$-triage patients.

Assume that the sequence of our systems is in (conventional) heavy-traffic, that is,
\[ \lambda^r_j \to \lambda_j, \quad j \in J, \quad \text{and} \quad r(\rho^r - 1) \to \beta, \quad \text{as} \quad r \to \infty, \]
for some given $\lambda_j > 0, \quad j \in J,$ and $\beta \in \mathbb{R}$. Let $\Lambda = (\lambda_k)_{k \in K}$ be the vector obtained from (2), with $\Lambda_J = (\lambda_j)_{j \in J}$ 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 $j \in J$, we assume the following convergence for the deadline of $j$-triage patients:
\[ \frac{d^r_j}{r} \to \hat{d}_j, \quad \text{as} \quad r \to \infty, \]
where $\hat{d}_j > 0, \quad j \in J$, are given constants.

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

**Assumption 1** When $r \to \infty$,
\[ r^{-1}Q^r_j(0) \Rightarrow 0, \quad j \in J, \]
\[ r^{-1}Q^r_k(0) \Rightarrow 0, \quad k \in K. \]

**4. Asymptotic compliance and optimality**

A control policy $\pi^r = \{T^r_j, \quad j \in J, \quad T^r_k, \quad k \in K\}$ determines the age processes of the head-of-the-line patients in the $r$th system, $\tau^r(\cdot) = \{\tau^r_j(\cdot), \quad j \in J\}$. We define the diffusion-scaled age processes through
\[ \hat{\tau}^r_j(t) = r^{-1}\tau^r_j(r^2t), \quad j \in J. \]

We 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 $T \geq 0$,
\[ \sup_{0 \leq t \leq T} \left[ \hat{\tau}^r_j(t) - \hat{d}_j \right]^+ \Rightarrow 0, \quad \text{as} \quad r \to \infty, \quad \text{for all} \quad j \in J. \]
Define the diffusion-scaled number of $k$-IP patients in the system by
\[ \hat{Q}_k(t) = r^{-1}Q_k(t^2), \quad k \in \mathcal{K}. \]  
(8)

We assume that, at time $t$ (in the diffusion scaling), $k$-IP patients incur a queueing cost at rate $C_k(\hat{Q}_k(t))$, for some function $C_k$. (Concrete assumptions on $C_k$ will be provided in Assumption 2.) Then the cumulative queueing cost is
\[ U_r(t) := \int_0^t \sum_{k \in \mathcal{K}} C_k \left( \hat{Q}_k(s) \right) ds. \]  
(9)

Our heavy-traffic adaptation of problem (6) is to stochastically minimize $U_r(t)$, 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 $x > 0$,
\[ \limsup_{r \to \infty} \mathbb{P} \{ U_r^*(t) > x \} \leq \liminf_{r \to \infty} \mathbb{P} \{ U^*(t) > x \}; \]

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

### 5. Main results

#### 5.1. Cost functions and an optimization problem

For any given $a \geq 0$, consider the optimization problem over $x = (x_k)_{k \in \mathcal{K}}$:
\[
\begin{align*}
\min_x & \quad \sum_{k \in \mathcal{K}} C_k(x_k) \\
\text{s.t.} & \quad \sum_{k \in \mathcal{K}} m_k^e x_k = a, \\
& \quad x \geq 0.
\end{align*}
\]  
(10)

We assume that the cost functions $C_k, k \in \mathcal{K}$, satisfy the following, in analogy to van Mieghem (1995).

**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
\[
\begin{align*}
C'_k(x_k^*) - \alpha_0 m_k^e &= 0, \\
\sum_{k \in \mathcal{K}} m_k^e x_k^* &= a.
\end{align*}
\]

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

\[ x^* = \Delta_{K}(a). \]

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

The mapping \( \Delta_{K} \) 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 \( \epsilon^r \) such that \( \frac{\epsilon^r}{r} \rightarrow 0 \) as \( r \rightarrow \infty \).
- 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 \( j \in J \) 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

\[ j \in \arg \max_{j \in J} \frac{\tau_j^r(t)}{d_j^r}. \tag{11} \]

- If IP patients are chosen to be served at time \( t \), the physician uses a rule ensuring (for any \( T > 0 \))

\[ \max_{l,k \in K} \sup_{0 \leq t \leq T} \left| \frac{C'_l(\hat{Q}_r^l(t))}{m_l^r} - \frac{C'_k(\hat{Q}_r^k(t))}{m_k^r} \right| = 0. \tag{12} \]

An example of such a rule is to choose \( k \in \arg \max_{l \in K} \frac{C'_l(\hat{Q}_r^l(t))}{m_l^r} \), which is a modified generalized \( c \mu \)-rule. (More examples of rules ensuring (12) are presented in §6.2.)

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 2 Note that (11) is different from the Shortest-Deadline-First rule from the introduction. (The latter is introduced formally in (17) below.) Both policies, using either (11) or (17), are asymptotically optimal, as discussed in §6.1 and proved in §EC.11. The reason for using them as we did is that the proof for (11) is notation-wise easier while (17) performs slightly better in simulations (§EC.16.3). A formal comparison of the two would involve rates of convergence, which is beyond the scope of the present paper.
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 $\frac{c_k}{m_k}$; 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 $j \in J$ and $k \in K$, introduce $K \times K$ matrices $\Gamma^j_{ll'} = (\Gamma^j_{ll'})$ and $\Gamma^k_{ll'} = (\Gamma^k_{ll'})$ through

$$
\Gamma^j_{ll'} = \begin{cases} 
P_{jl}(1 - P_{jl'}), & \text{if } l = l' \\
-P_{jl}P_{jl'}, & \text{if } l \neq l' 
\end{cases}
$$

and

$$
\Gamma^k_{ll'} = \begin{cases} 
P_{kl}(1 - P_{kl'}), & \text{if } l = l' \\
-P_{kl}P_{kl'}, & \text{if } l \neq l' 
\end{cases}
$$

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

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

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

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

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

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 $\hat{Q}_k^r$ from (8) and define similarly the diffusion-scaled queue length processes for triage classes: $\hat{Q}_j^r(t) = r^{-1} \hat{Q}_j^r(r^2 t)$, $j \in J$. The diffusion-scaled weighted queue length processes is given by

$$
\hat{Q}_w^r(t) = \sum_{j \in J} m_j^e \hat{Q}_j^r(t) + \sum_{k \in K} m_k^e \hat{Q}_k^r(t).
$$
Proposition 1 (Invariance principle for work-conserving policies) Under any family of work-conserving policies,

\[ \hat{Q}_w \Rightarrow \hat{Q}_w, \quad \text{as} \quad r \to \infty. \]  

(14)

This proposition is proved in §EC.3.

For any \( a \in \mathbb{R}_+ \), let \( x = \Delta \mathcal{J}(a) \in \mathbb{R}_+^J \) be the solution to the following linear equation:

\[
\sum_{j \in \mathcal{J}} m_j^e x_j = a,
\]

\[
\frac{x_j}{\lambda_j \hat{d}_j} = \frac{x_{j'}}{\lambda_{j'} \hat{d}_{j'}}, \quad \text{for} \quad j, j' \in \mathcal{J}.
\]  

(15)

As in Lemma 5.1, we can also prove that \( \Delta \mathcal{J}(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+^J \) is well-defined, and \( \Delta_j \) is nondecreasing for each \( j \in \mathcal{J} \). In fact, \( \Delta \mathcal{J}(a) = (\Delta_j(a))_{j \in \mathcal{J}} \) with \( \Delta_j = \frac{\lambda_j \hat{d}_j}{\hat{\omega}} \). The function pair \( (\Delta \mathcal{J}, \Delta \mathcal{K}) \) is the lifting mapping in the state-space collapse result. Let \( \hat{Q} = (\hat{Q}_j, j \in \mathcal{J}; \hat{Q}_k, k \in \mathcal{K}) \) and recall that \( \hat{\omega} = \sum_{j \in \mathcal{J}} \lambda_j \hat{d}_j m_j^e \).

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

\[
\hat{Q}_j(t) = \Delta_j \min \left( \hat{Q}_w(t), \hat{\omega} \right), \quad j \in \mathcal{J},
\]

\[
\hat{Q}_k(t) = \Delta_k \left( (\hat{Q}_w(t) - \hat{\omega})^+ \right), \quad k \in \mathcal{K}.
\]

This theorem is proved in §EC.4.

Remark 4 The expression of \( \Delta_j \) yields \( \hat{Q}_j(t) \leq \lambda_j \hat{d}_j \), which can be translated into “asymptotic compliance” of the family of control policies \( \{\pi^*_r\} \); on the other hand, those limits \( \hat{Q}_k \) appear in the lower bound of Theorem 2, which shows 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 brings notational convenience in our case but 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 \( \omega^*_j(t) \) and \( \omega^*_k(t) \) the virtual waiting times for \( j \)-triage class and \( k \)-IP class respectively, and define the diffusion-scaled virtual waiting time processes by

\[
\hat{\omega}^*_j(t) = r^{-1} \omega^*_j(r^2 t), \quad j \in \mathcal{J}, \quad \text{and} \quad \hat{\omega}^*_k(t) = r^{-1} \omega^*_k(r^2 t), \quad k \in \mathcal{K}.
\]  

(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
\[
\hat{\omega}_r^j - \hat{Q}_r^j/\lambda_j^r \Rightarrow 0, \quad j \in J,
\]
\[
\hat{\omega}_r^k - \hat{Q}_r^k/\lambda_k^r \Rightarrow 0, \quad k \in K.
\]
This proposition is proved in §EC.7.

Remark 5 From the convergence of \( \hat{Q}_r^r \) in Theorem 3, one deduces the convergence of the vector of virtual waiting times under the family of control policies \( \{ \pi^*_r \} \).

Recall that \( \tau_r^j(t) \) is defined as the age of the head-of-the-line \( j \)-triage patient in the \( r \)th system. Similarly, let \( \tau_r^k(t) \) be the age of the head-of-the-line \( k \)-IP patient in the \( r \)th system, with its diffusion scaling \( \hat{\tau}_r^k(t) = r^{-1} \tau_k^r(r^2 t) \), \( k \in K \). 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 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 have
\[
\hat{\omega}_r^j - \hat{\tau}_r^j \Rightarrow 0, \quad j \in J,
\]
\[
\hat{\omega}_r^k - \hat{\tau}_r^k \Rightarrow 0, \quad k \in K.
\]
This proposition is proved in §EC.8.

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
\[
\hat{W}_{jh}^r(t) = r^{-1} W_{jh}^r(r^2 t), \quad j \in J.
\]

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 \( r \to \infty \),
\[
\hat{W}_{jh}^r - \frac{\hat{Q}_r^j}{\lambda_j^r} - \sum_{k \in K} \frac{h_k}{\lambda_k^r} \hat{Q}_r^k \Rightarrow 0, \quad j \in J.
\]
This proposition is proved in §EC.9.

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

\[
\frac{\hat{Q}_r}{\lambda_j} + \sum_{k \in K} \frac{h_k}{\lambda_k} \hat{Q}_{rk} = \Delta_j \min \left( \hat{Q}_w, \tilde{\omega} \right) + \sum_{k \in K} \frac{h_k}{\lambda_k} \Delta_k \left( (\hat{Q}_w - \tilde{\omega})^+ \right).
\]

Then Proposition 4 yields an estimator for the distribution of \( \hat{W}^r_{jh}(\cdot) \):

\[
\Delta_j \min \left( \hat{Q}_w(\cdot), \tilde{\omega} \right) + \sum_{k \in K} \frac{h_k}{\lambda_k} \Delta_k \left( (\hat{Q}_w(\cdot) - \tilde{\omega})^+ \right).
\]

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 \( r \to \infty \),

\[
\hat{W}^r_{jh} - \hat{\tau}^r_j - \sum_{k \in K} h_k \hat{\tau}^r_k \Rightarrow 0, \quad j \in J.
\]

**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. Alternative triage rules to (11)

The recipe in (11), as part of an asymptotically optimal policy, is not unique. One alternative control rule, assuming that triage classes were chosen to be served at time \( t \), is having the physician attend to the head-of-the-line patient from the class with index

\[
j \in \arg \max_{j \in J} \frac{Q^r_j(t)}{\lambda_j d_j^r};
\]

this can be proved asymptotically equivalent to (11).

Next we consider the Shortest-Deadline-First rule: When triage classes are chosen to be served at time \( t \), the physician chooses the head-of-the-line patient from the class with index

\[
j \in \arg \min_{j \in J} (d_j^r - \tau_j^r(t)). \tag{17}
\]

From Lemma EC.6.2, the above is asymptotically equivalent to choosing the head-of-the-line patient from the class with index

\[
j \in \arg \min_{j \in J} \left( d_j^r - Q^r_j(t)/\lambda_j^r \right).
\]

One can prove that the family of control policies, with (17) replacing (11), is also asymptotically compliant and asymptotically optimal (see §EC.11 for a comprehensive discussion). Formally, Theorems 1–3 still prevail, but the lifting mapping \( \Delta_J \) (15) in Theorem 3 is changed to another \( \tilde{\Delta}_J \), the expression of which is much more complicated (see (EC.62)). We also note that the results in §5.4 and 5.5 hold under the policy with (11) replaced by (17).
6.2. IP-Rules that imply (12)

In this subsection we describe a collection of rules that imply (12).

Consider any $K \times K$-dimensional invertible matrix $G$, with $G_{kk} > 0$, $G_{kl} < 0$ for $l \neq k, l \in K$ while $\sum_l G_{kl} \geq 0$. We also assume that all terms in $G^{-1}$ are non-negative, and all components of $GM^e$ are positive ($M^e$ is defined in (3)). Let $H$ denote the $K$-dimensional vector with the $k$th component $1/(GM^e)_k$. When IP classes are chosen to be served at time $t$, the physician chooses a patient from the class with index

$$k \in \arg \max_{k \in K} H_k \cdot \left(GC'\left(\hat{Q}'(t)\right)\right)_k; \quad (18)$$

here $C'(\hat{Q}'(t))$ is a $K$-dimensional column vector with $C'_k(\hat{Q}'_k(t))$ being its $k$th component.

**Lemma 6.1** For any $T \geq 0$, as $r \to \infty$,

$$\sup_{0 \leq t \leq T} \left| \frac{C'_k(\hat{Q}'_k(t))}{m^e_k} - \frac{C'_l(\hat{Q}'_l(t))}{m^e_l} \right| \Rightarrow 0,$$

for all $l, k \in K$. As a result, (12) holds.

This lemma will be proved in §EC.12.

There are two special choices of $G$ which are especially interesting:

1. $G = I$: then $H_k = 1/m^e_k$; hence (18) becomes

$$k \in \arg \max_{k \in K} \frac{C'_k(\hat{Q}'_k(t))}{m^e_k}. \quad (19)$$

This is a generalized $c\mu$ rule, modified from van Mieghem (1995) and Mandelbaum and Stolyar (2004) to account for feedbacks.

2. $G = I - P$: noticing that $M^e = (I - P)^{-1}M$, then $H$ is a vector with $\mu_k$ being the $k$th component; hence (18) is now

$$k \in \arg \max_{k \in K} \left[ C'_k(\hat{Q}'_k(t)) - \sum_{l \in K} P_{kl} C'_l(\hat{Q}'_l(t)) \right] \mu_k.$$  

Note that this is the rule conjectured in Mandelbaum and Stolyar (2004).

The expression in (12) is similar to equation (51) in van Mieghem (1995), with the waiting times there replaced by queue lengths, and the mean service times there replaced by effective mean service times. As the effective mean service time is in fact the expected total service time of a patient, accumulated over all visits, the following exhaustive rule is also expected to satisfy (12): when the IP classes are chosen to be served, the physician chooses a patient from the class with index $k \in \arg \max_{k \in K} C'_k(\hat{Q}'_k(t))/m^e_k$, and serves this patient continuously until completing all services—the current one as well as successive feedbacks. This exhaustive rule, which is not FCFS within each IP class, is not plausible in an ED setup. Indeed, patient flow in the ED must allow for delays between successive visits to physician (see §8.1). Moreover, even without such delays, some additional structural conditions must hold for the exhaustive rule to be asymptotically optimal (specifically, each IP class must be the starting class of a route); hence we do not pursue it further here.
6.3. Waiting costs

We now consider waiting costs, instead of queueing costs. To this end, we 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 $\hat{\omega}_k^r(t)$ is defined in (16). We seek to stochastically minimize the cost

$$\tilde{U}_r^*(t) := \sum_{k \in K} \int_0^t C_k (\hat{\omega}_k^r(s)) \, d\bar{E}_k^r(s),$$

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

We now slightly modify the control policy $\{\pi_r^*\}$ in Section 5. The first step, using a threshold
rule to determine 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 uses a rule ensuring that, for any $T \geq 0$,

$$\max \sup_{l, k \in K} \left| \frac{C_l^r \left( \bar{Q}_l^r(t) \lambda_k^r \right)}{m_l^r} - \frac{C_k^r \left( \bar{Q}_k^r(t) \lambda_k^r \right)}{m_k^r} \right| \to 0.$$ 

An example of such a rule is to choose $k \in \arg \max_{k \in K} \frac{C_k^r \left( \bar{Q}_k^r(t) \lambda_k^r \right)}{m_k^r}$. Other examples of rules satisfying the above can be deduced from the rules in §6.2.

Denote this family of modified policies by $\{\tilde{\pi}_r^*\}$.

**Proposition 5 (Waiting Time Cost)** The family of control policies $\{\tilde{\pi}_r^*\}$ 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 $x > 0$,

$$\limsup_{r \to \infty} \mathbb{P} \left\{ \tilde{U}_r^*(t) > x \right\} \leq \liminf_{r \to \infty} \mathbb{P} \left\{ \tilde{U}_r^*(t) > x \right\},$$

where $\{\tilde{U}_r^*\}$ is the family of cumulative cost, defined through (20) under the family of control
policies $\{\tilde{\pi}_r^*\}$, and $\{\tilde{U}_r^*\}$ 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.10.

6.4. An alternative criterion: IP sojourn time

In this subsection, we consider the alternative model discussed in §1.3. 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
Huang, Carmeli, and Mandelbaum: Patient Flow Control in ED
Article submitted to Operations Research; manuscript no. (Please, provide the manuscript number!)

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 $C_0$ to denote the set of starting IP classes of routes. For $k \in C_0$, let $C_k$ denote all the classes on the route that starts at $k$, and call any class in $\bigcup_{k \in C_0} C_k \setminus \{k\}$ a subsequent class. If a patient with starting class $k$ waits $\omega_{k'}$, as a $k'$-IP patient ($k' \in C_k$), then the sojourn time of this patient is $\sum_{k' \in C_k} \omega_{k'}$. Our problem is to stochastically minimize the cost

$$\tilde{S}^r_t = \sum_{k \in C_0} \int_0^t C_k \left( \sum_{k' \in C_k} \hat{Q}_{k'}(s) \right) d\tilde{E}^r_k(s),$$

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 the 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:

- Give priority to all subsequent classes, while allocating the remaining service capacity to all starting classes to ensure that

$$\max_{l,k \in C_0} \sup_{0 \leq t \leq T} \left| \frac{C_l' \left( \frac{\hat{Q}_l(t)}{\lambda_l} \right)}{m_l^c} - \frac{C_k' \left( \frac{\hat{Q}_k(t)}{\lambda_k} \right)}{m_k^c} \right| \Rightarrow 0. \quad (22)$$

Here $\hat{Q}_l, \hat{Q}_k$ are the diffusion-scaled queue lengths of the starting classes $l, k \in C_0$, and $m_l^c, m_k^c$ are the corresponding effective mean service times. An example of such a rule is to choose $k \in \arg \max_{k \in C_0} \frac{C_k' \left( \frac{\hat{Q}_k(t)}{\lambda_k} \right)}{m_k^c}$. Other examples of rules satisfying the above can be modified from the rules in §6.2.

We denote this family of policies by $\{\tilde{\pi}^r_{**}\}$.

**Proposition 6 (Sojourn Time Cost)** The family of control policies $\{\tilde{\pi}^r_{**}\}$ 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 $x > 0$,

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

here $\{\tilde{S}^r_{**}\}$ is the family of cumulative cost defined through (21) under the family of control policies $\{\tilde{\pi}^r_{**}\}$, and $\{\tilde{S}^r\}$ is the corresponding cost under any other asymptotically compliant family of policies $\{\pi^r\}$.

The outline of the proof can be found in §EC.13.

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 two systems. One system has stationary arrival rates and no delays between successive visits to physicians, as analyzed in the paper. The second system has time-varying arrival rates as well as delays between successive visits to physicians. These are features that were assumed out in our model. As observed in our simulations, the proposed policy performs very well, and it outperforms commonly-used alternatives in both systems. After justifying its relevance empirically, we present the results for the stationary model here. The results for the time-varying model can be found in the Appendix (§EC.16.5).

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 (2013) 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. We thus focus on triage classes 3, 4 and 5, which we index by 1, 2, 3. The deadlines for those three classes are 30, 60 and 120 minutes. Empirical analysis shows that, on average, there were 302 patients arriving at the ED each weekday in January 2004. Figure 2 depicts the shape (percentage) of daily arrivals per hour.

![Figure 2](image)

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 is 14 per hour, that is, 14/60 per minute. The proportion 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.

In our ED, patients experience 1-5 IP phases (doctor visits): 28% go through 1 phase only and then released, 30% have 2 phases, 28% - 3, 11% - 4, and 3% go through 5 IP phases. We classify the IP patients into different classes according to their number of IP visits, where we
combine phases 3, 4 and 5. 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):

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

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 (2013)). Service times are assumed (harmlessly) to follow exponential distributions. 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 ED partner 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. Our proposed policy

We use simulation to shed light on which (asymptotically optimal) policy to use in practice.

- Our recommendation for the threshold part is as follows: Assign priority to triage classes if, for some $j \in J$, $\tau_j(t) \geq d_j - \epsilon$. Here $\epsilon$ 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, $\epsilon$ 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 for each physician is around 6.5 minutes and there are 5 physicians. Thus the corresponding average service time by the “super” single-server is $6.5/5 = 1.3$. 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 $\epsilon$. 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 $\epsilon$ for different classes, which may be somewhat larger than in the stationary case. (For example, in §EC.16.5, 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.)

- Rule for triage patients: As in the threshold part, we recommend to use ages of triage patients. We compared the rule (11) and the Shortest-Deadline-First rule (17) via simulation (see §EC.16.3 and §EC.16.5). Both rules perform well, with (17) performing slightly better. As a result, our recommendation is the Shortest-Deadline-First rule (17).
• Rule for IP patients: We recommend the modified $Gc\mu$ rule (19). It has a very simple form and hence it is easy to implement. More importantly, our simulations confirm that it performs well.

We first simulate the ED under our recommended policy (denoted by $TGc\mu$) to gain an insight on 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.16.1) for more detailed descriptions of these three policies.

7.3. Simulation outcomes

We ran the system over 380 days (with time unit of 1 minute, the duration is $60 \times 24 \times 380 = 547,200$ 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: $14 \times 24 \times 15 \times (0.1, 0.4, 0.5)$, 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. (The time-varying model, which also has delays between physician visits, is described in §EC.16.5.)

Figure 3 displays a typical sample path under our proposed policy. We plot the waiting times for all three triage classes, as well as the queue lengths of class 3 triage patients and class 1 IP patients. The reason we choose these two classes is because class 3 triage has the longest queue length among the 3 triage classes (as expected) and class 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:

From the figure we make the following observations:

1. The 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 classes 3, 4 and 5 respectively; the fractions of triage deadline violations by more than 10% of
their corresponding deadlines are negligible (less than 1%). In the Appendix, we support these
statements through the histograms of patient waiting times.

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

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” is IP-cost per time-unit, averaged over samples.
(The numbers in brackets are half-length of 95% confidence intervals.)

<table>
<thead>
<tr>
<th>Policy</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGcµ</td>
<td>4.61% (0.10%)</td>
<td>4.57% (0.09%)</td>
<td>4.57% (0.09%)</td>
<td>125.21 (10.36)</td>
</tr>
<tr>
<td>FCFS</td>
<td>31.27% (0.49%)</td>
<td>10.16% (0.38%)</td>
<td>1.15% (0.16%)</td>
<td>187.46 (7.20)</td>
</tr>
<tr>
<td>IPF</td>
<td>21.26% (0.48%)</td>
<td>21.28% (0.48%)</td>
<td>21.26% (0.48%)</td>
<td>0.88 (0.07)</td>
</tr>
<tr>
<td>TrF</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>552.96 (23.71)</td>
</tr>
</tbody>
</table>

Our proposed policy (TGcµ) outperforms the global FCFS policy. The cost rate for IP-
patients-First (IPF) is small, but a large proportion of triage patients violate the deadlines.
(In the time-varying model, this IPF policy is much worse; see §EC.16.5 in the Appendix.)
Triage-patients-First (TrF) has patients that satisfy the deadline constraints, while its cost rate
exceeds 4 times that of TGcµ. In summary, our proposed policy (TGcµ) clearly dominates the
other three alternatives.

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. The examples are
delays between physician visits and time-varying arrivals, which have been incorporated in our
simulation (§EC.16.5, without significant effect on performance). 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 Man-
delbaum (2013), 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. We now
formulate our conjecture on the length of delays.
Consider the basic (queue-length) model as an example. Following Yom-Tov and Mandelbaum (2013), we model the 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 $r^αj^kμ_{jk}$, and the one between $l$-IP patients and $k$-IP patients is $r^αl^kμ_{lk}$. 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 $α > -2$ (for all $α$‘s), the delays are then short enough to leave our results intact. Conversely, $α_{jk} < -2$ (for all $j,k$) decouples the triage from IP—both can be controlled separately; and $α_{lk} < -2$ (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.14.

Simulations 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 must, like many other service systems, cope with arrival rates that are significantly time-varying (Yom-Tov and Mandelbaum 2013, Figure 10). 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 staffing 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 service 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, 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. 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. Limitation on the number of beds

In our current work, we assume that there is no limitation on the number of beds. This is true in many Israeli EDs (including the ED of our partner hospital, in which essentially all patients are admitted to the ED), as well as other EDs around the world (for example, those that do not allow ambulance diversion). Furthermore, we showed (theoretically and via simulation) that our proposed policy can keep the number of IP patients under control, which may ameliorate the need for ample beds. Nevertheless, it is of interest to understand the impact of a finite number of beds, which would give rise to an admission control problem, as in Plambeck et al. (2001). Interestingly, admission control problems, with costs incurred by blocked customers, in fact motivated Plambeck et al. (2001). However, we opted for the analysis of triage-constraints first, in the belief (and support of our doctor partner) that they play a higher order (clinical) role.

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 turn 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 the U.S. EDs in the form of ambulance diversion (Deo and Gurvich (2011), Allon et al. (2013), Armony 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”, the system is an inverted-V model in which the $N$ physicians can be viewed as $N$ service stations. 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 $\bar{W}_n(t)$ the diffusion-scaled workload at station $n$. Then, for every $T > 0$, we expect to find a sequence $\delta^* \downarrow 0$ with $\mathbb{P}(\sup_{0 \leq t \leq T} \sup_{m,n} |\bar{W}_n^*(t) - \bar{W}_m^*(t)| \geq \delta^*) \leq \delta^*$. Then, if there is one physician whose diffusion-scaled workload exceeds $\delta^*$, other physicians cannot be idle (with probability $1 - \delta^*$). With such a sequence of $\delta^*$‘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 around 5% (Armony et al. (2013)). 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. 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. But more fundamentally, understanding of the impact of abandonment on systems with feedback is still lacking.

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 A.M. has been partially supported by BSF Grants 2005175 and 2008480, as well as ISF Grant 1357/08.

References


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


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 \( \{\tau_j, j \in J\} \) are stochastically bounded, which gives rise to many useful stochastic boundedness results on other processes.

For any \( j \)-triage class, \( j \in J \), introduce diffusion-scaled processes

\[
\hat{E}_j(t) = r^{-1} (E_j(r^2t) - \lambda_j r^2 t), \\
\hat{S}_j(t) = r^{-1} (S_j([r^2t]) - \mu_j r^2 t), \\
\hat{T}_j(t) = r^{-1} (T_j(r^2t) - \lambda_j m_j r^2 t),
\]

and fluid-scaled processes

\[
\tilde{Q}_j(t) = r^{-2} Q_j(r^2t), \\
\tilde{E}_j(t) = r^{-2} E_j(r^2t), \\
\tilde{T}_j(t) = r^{-2} T_j(r^2t), \\
\tilde{S}_j(t) = r^{-2} S_j(r^2t).
\]

From Donsker’s Theorem, as \( r \to \infty \),

\[
(\hat{E}_j, \hat{S}_j, j \in J) \Rightarrow (\tilde{E}_j, \tilde{S}_j, j \in J); \tag{EC.1}
\]

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

\[
\text{diag}(\lambda_j a_j^2), \quad \text{diag}(\mu_j b_j^2).
\]

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 \( T \geq 0 \),

\[
\max_j \sup_{0 \leq t \leq T} \left| \tilde{Q}_j(t) - \lambda_j \tilde{\tau}_j(t) \right| \Rightarrow 0, \quad \text{as} \quad r \to \infty. \tag{EC.2}
\]

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

\[
\left| Q_j(t) - \left( E_j(t) - E_j\left((t - \tau_j(t))\right)\right) \right| \leq 1.
\]

Then

\[
\left| \tilde{Q}_j(t) - \lambda_j \tilde{\tau}_j(t) \right| \leq \left| \hat{E}_j(t) - \hat{E}_j\left((t - \tilde{\tau}_j(t))\right)\right| + \frac{1}{r}, \quad j \in J. \tag{EC.3}
\]

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

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

In the following lemma, we prove the fluid busy time for triage patients, under all asymptotically optimal policies. We also prove that \( \hat{Q}_j^r(\cdot) + \mu_j \hat{T}_j^r(\cdot) \) converge, though we cannot 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 \( r \to \infty \),

\[
\hat{T}_j^r(\cdot) \Rightarrow \lambda_j m_j e(\cdot),
\]

\[
\hat{Q}_j^r(\cdot) + \mu_j \hat{T}_j^r(\cdot) \Rightarrow \bar{E}_j(\cdot) - \bar{S}_j(\lambda_j m_j e(\cdot)).
\]

Consequently, \( \hat{Q}_j^r \) and \( \hat{T}_j^r \) are stochastically bounded.

**Proof:** For \( j \in J \), as

\[
Q_j^r(t) = Q_j^r(0) + E_j^r(t) - S_j(T_j^r(t)),
\]

then

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

and

\[
\hat{Q}_j^r(t) = \hat{Q}_j^r(0) + \hat{E}_j^r(t) - \hat{S}_j^r(\hat{T}_j^r(t)) - \mu_j \hat{T}_j^r(t). \tag{EC.7}
\]

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

\[
\sup_{0 \leq t \leq T} \hat{Q}_j^r(t) \Rightarrow 0,
\]

\[
\sup_{0 \leq t \leq T} \left| \hat{E}_j^r(t) - \lambda_j t \right| \Rightarrow 0, \tag{EC.8}
\]

\[
\sup_{0 \leq t \leq T} \left| \hat{S}_j^r \left( \hat{T}_j^r(t) \right) - \mu_j \hat{T}_j^r(t) \right| \leq \sup_{0 \leq t \leq T} \left| \hat{S}_j^r(t) - \mu_j t \right| \Rightarrow 0, \tag{EC.9}
\]

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). \( \square \)

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 \( n \)-th \( j \)-triage patient transfers (§2.1), and \( \phi_i(n) \) the indicator function recording the class to which the \( n \)-th
l-IP patient transfers (§2.2). We use \( \phi_{jk}(n) \) to denote \( (\phi_j(n))_k \), the \( k \)th 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 \( k \)th 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 \( \Phi_{jk} \) and \( \Phi_{lk} \).

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

\[
\begin{align*}
\hat{E}_k^r(t) &= r^{-1}(E_k^r(t) - \lambda_k^r r^2 t), \\
\hat{S}_k^r(t) &= r^{-1}(S_k^r(t) - \mu_k^r r^2 t), \\
\hat{T}_k^r(t) &= r^{-1}(T_k^r(t) - \lambda_k m_k r^2 t), \\
\hat{\phi}_{jk}^r(t) &= r^{-1} (\phi_{jk}([r^2 t]) - P_{jk} r^2 t), \\
\hat{\phi}_{lk}^r(t) &= r^{-1} (\phi_{lk}([r^2 t]) - P_{lk} r^2 t).
\end{align*}
\]

Then from Donsker’s Theorem, as \( r \to \infty \),

\[
\left( \hat{\phi}_{jk}^r(\cdot), \hat{\phi}_{lk}^r(\cdot), \hat{S}_k^r(\cdot); \ j \in \mathcal{J}, l, k \in \mathcal{K} \right) \Rightarrow \left( \hat{\phi}_{jk}(\cdot), \hat{\phi}_{lk}(\cdot), \hat{S}_k(\cdot); \ j \in \mathcal{J}, l, k \in \mathcal{K} \right); \tag{EC.10}
\]

here \( \hat{\phi}_{jk}(\cdot), k \in \mathcal{K} \), \( j \in \mathcal{J} \), \( \hat{\phi}_{lk}(\cdot), l \in \mathcal{K} \), \( k \in \mathcal{K} \), \( \hat{S}_k(\cdot), k \in \mathcal{K} \) are independent driftless Brownian motions, with covariance matrices

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

respectively.

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

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

and

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

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

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

here

\[
\begin{align*}
\hat{E}_k^r(t) &= \sum_{j \in \mathcal{J}} \hat{\phi}_{jk}^r(S_j(T_j^r(t))) + \sum_{l \in \mathcal{K}} \hat{\phi}_{lk}^r(S_l(T_l^r(t))) + \sum_{j \in \mathcal{J}} P_{jk} \hat{S}_j^r(T_j^r(t)) + \sum_{l \in \mathcal{K}} P_{lk} \hat{S}_l^r(T_l^r(t)). \tag{EC.13}
\end{align*}
\]
Denote \( \hat{Q}_w(t) \) is recalled from (13) for convenience:

\[
\hat{Q}_w(t) = \sum_{j \in J} m_j \hat{Q}_j(t) + \sum_{k \in K} m_k \hat{Q}_k(t),
\]

\[
\hat{X}_w(t) = \hat{Q}_w(0) + r(\rho^r - 1)t + \sum_{j \in J} m_j \left[ \hat{E}_j(t) - \hat{S}_j \left( \hat{T}_j(t) \right) \right] + \sum_{k \in K} m_k \left[ \hat{E}_k(t) - \hat{S}_k \left( \hat{T}_k(t) \right) \right],
\]

\[
\hat{T}_+^r(t) = r^{-1} \left( r^2 t - \sum_{j \in J} T_j^r(r^2 t) - \sum_{k \in K} T_k^r(r^2 t) \right),
\]

(EC.14)

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

\[
-m_j \mu_j + \sum_{k \in K} P_{jk} \mu_j m_k^r = -1,
\]

(EC.15)

\[
-m_k \mu_k + \sum_{l \in K} P_{kl} \mu_k m_l^r = -1.
\]

(EC.16)

Multiplying (EC.7) by \( m_j^r \), (EC.12) by \( m_k^r \), and summing them up, one has

\[
\hat{Q}_w(t) = \hat{X}_w^r(t) + \hat{T}_+^r(t),
\]

(ΕC.17)

\[
\hat{Q}_w(t) \geq 0,
\]

\[
\hat{T}_+^r(\cdot) \text{ is nondecreasing with } \hat{T}_+^r(0) = 0.
\]

Note that the policy at hand needs not be work-conserving, thus it is possible for \( \hat{T}_+^r \) to increase at \( t \) when \( \hat{Q}_w(t) \neq 0 \). Hence

\[
\hat{Q}_w^r(t) \geq \Phi(\hat{X}_w^r(t)),
\]

(EC.18)

where \( \Phi \) 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 one enables flexibility to construct a new converging sequence with the desired lower bound; the third one is negligible in probability.

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

\[
\Gamma_1^r(t) = \left\{ \mathcal{U}^r(t) > x, \ \max_{k \in K} \sup_{0 \leq s \leq t} \bar{Q}_k(s) \leq \frac{1}{r^{1/4}} \right\},
\]

\[
\Gamma_2^r(t) = \left\{ \max_{k \in K} \sup_{0 \leq s \leq t} \bar{Q}_k(s) > \frac{1}{r^{1/4}} \right\},
\]

\[
\Gamma_3^r(t) = \left\{ \mathcal{U}^r(t) \leq x, \ \max_{k \in K} \sup_{0 \leq s \leq t} \bar{Q}_k(s) \geq \frac{1}{r^{1/4}} \right\}.
\]
Here \( \bar{\bar{Q}}_k^x \) is the fluid-scaled number of \( k \)-IP patients in the system, defined via

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

Then

\[
\{ \mathcal{U}^r(t) > x \} = (\Gamma_1^r(t) \cup \Gamma_2^r(t)) \setminus \Gamma_3^r(t).
\]

We first prove

\[
\lim_{r \to \infty} \mathbb{P} \{ \Gamma_3^r(t) \} = 0.
\]

For notational simplicity, denote \( r'(s, \vartheta) = [s, s + \frac{1}{\vartheta r^2}] \) and \( \vartheta_0 = 4 \max_{k \in \mathcal{K}} \mu_k \). For \( s < u \), denote \( S_k^r(s, u) = S_k(r(T^r(r^2s) + r^2(u - s)) - S_k(r(T^r(r^2s))) \) and \( \bar{\bar{S}}_k^r(s, u) = r^{-2}S_k^r(s, u) \). One can prove that

\[
\lim_{r \to \infty} \mathbb{P} \left\{ \max_{k \in \mathcal{K}} \sup_{0 \leq s \leq t \leq u \in r'(s, \vartheta_0)} \bar{\bar{S}}_k^r(s, u) > \frac{1}{2r^{1/2}} \right\} = 0.
\]

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

\[
\lim_{r \to \infty} \mathbb{P} \left\{ \max_{k \in \mathcal{K}} \sup_{0 \leq s \leq t \leq u \in r'(s, \vartheta_0)} \bar{\bar{Q}}_k^x(s) - \bar{\bar{Q}}_k^x(u) \right\} = 0.
\]

It follows that

\[
\lim_{r \to \infty} \mathbb{P} \{ \Gamma_3^r(t) \} \leq \limsup_{r \to \infty} \mathbb{P} \left\{ \mathcal{U}^r(t) \leq x, \max_{k \in \mathcal{K}} \sup_{0 \leq s \leq t \leq u \in r'(s, \vartheta_0)} \bar{\bar{Q}}_k^x(u) > \frac{1}{2r^{1/2}} \right\}
\]

\[
\leq \limsup_{r \to \infty} \mathbb{P} \left\{ \min_{k \in \mathcal{K}} \frac{2}{\vartheta_0 r^{1/4}} C_k \left( \frac{1}{2} \right)^{3/4} \leq x, \max_{k \in \mathcal{K}} \sup_{0 \leq s \leq t \leq u \in r'(s, \vartheta_0)} \bar{\bar{Q}}_k^x(u) > \frac{1}{2r^{1/2}} \right\}
\]

\[
\leq \limsup_{r \to \infty} \mathbb{P} \left\{ \frac{r^{1/2}}{\vartheta_0} \min_{k \in \mathcal{K}} \frac{2}{\vartheta_0 r^{1/4}} C_k \left( \frac{1}{2} \right)^{3/4} \leq x \right\} = 0.
\]

This completes the proof of (EC.20).

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

\[
\liminf_{r \to \infty} \mathbb{P} \{ \mathcal{U}^r(t) > x \} = \liminf_{r \to \infty} \mathbb{P} \{ \Gamma_1^r(t) \cup \Gamma_2^r(t) \}.
\]

Next we derive a lower bound for the latter term.

Denote

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

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

\[
\bar{T}_k^x(\cdot) = \lambda_k m_k e(\cdot), \quad k \in \mathcal{K}.
\]

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 $s \leq t$, define $\hat{T}_j^r(s) = r^{-1} \hat{T}_j^r(s)$, for $j \in J$, and 
\[
\begin{align*}
\hat{Q}_k^r(s) &= r^{-1} \hat{Q}_k^r(s), \quad \hat{E}_k^r(s) = r^{-1} \hat{E}_k^r(s), \\
\hat{S}_k^r(s) &= r^{-1} \hat{S}_k^r(s), \quad \hat{T}_k^r(s) = r^{-1} \hat{T}_k^r(s), \\
\hat{\Phi}_{jk}^r(s) &= r^{-1} \hat{\Phi}_{jk}^r(s), \quad \hat{\Phi}_{lk}^r(s) = r^{-1} \hat{\Phi}_{lk}^r(s),
\end{align*}
\]
for $j \in J$, $l, k \in K$. Then from (EC.12),
\[
\sum_{l \in K} P_{lk} \mu_l \hat{T}_l^r(s) - \mu_k \hat{T}_k^r(s) = \hat{Q}_k^r(s) - \hat{Q}_k^r(0) - \hat{E}_k^r(s) + \hat{S}_k^r(\hat{T}_k^r(s)) - \sum_{j \in J} P_{jk} \mu_j \hat{T}_j^r(s). \tag{EC.23}
\]
On $\Gamma_0^r(t)$, we know that $\sup_{0 \leq s \leq t} \hat{Q}_k^r(s) \Rightarrow 0$. Together with (EC.4), the expression of $\hat{E}_k^r$ in (EC.13), and $\hat{T}_k^r(s) \leq s$ for all $k \in K$ (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 $\Gamma_0^r(t)$,
\[
\sum_{l \in K} P_{lk} \mu_l \hat{T}_l^r(\cdot) - \mu_k \hat{T}_k^r(\cdot) \Rightarrow 0, \quad \text{in } D[0, t].
\]
Introducing the $K$-dimensional process $\tilde{T}_\mu^r(s) = (\mu_k \hat{T}_k^r(s))_{k \in K}$ in $D[0, t]$, the above is then
\[
(PT - I) \tilde{T}_\mu^r(\cdot) \Rightarrow 0, \quad \text{on } \Gamma_0^r(t).
\]
Since $PT - I$ is invertible, and all $\mu_k$, $k \in K$, are nonzero, we have
\[
\tilde{T}_k^r(\cdot) \Rightarrow 0, \quad k \in K \quad \text{in } D[0, t],
\]
which is equivalent to (EC.22).

For $s \leq t$, define $\hat{X}_0^r(s) = \hat{X}_w^r(s)$ on $\Gamma_0^r(t)$, and otherwise,
\[
\begin{align*}
\hat{X}_0^r(s) &= \sum_{j \in J} m_j^r \hat{Q}_j^r(0) + \sum_{k \in K} m_k^r \hat{Q}_k^r(0) + r(\rho^r - 1)s \\
&\quad + \sum_{j \in J} m_j^r \left[ \hat{E}_j^r(s) - \hat{S}_j^r(\lambda_j^r m_j s) \right] + \sum_{k \in K} m_k^r \left[ \hat{E}_k^r(s) - \hat{S}_k^r(\lambda_k^r m_k s) \right],
\end{align*}
\]
here for $k \in K$,
\[
\hat{E}_k^r(s) = \sum_{j \in J} \hat{\Phi}_{jk}^r(\lambda_j^r s) + \sum_{l \in K} \hat{\Phi}_{lk}^r(\lambda_l^r s) + \sum_{j \in J} P_{jk} \hat{S}_j^r(\lambda_j^r m_j s) + \sum_{l \in K} P_{lk} \hat{S}_l^r(\lambda_l^r m_l s).
\]
From (EC.22) on $\Gamma_0^r(t)$, (EC.4) and $\lambda_k^r \rightarrow \lambda_k$, $k \in K$, one deduces that
\[
\hat{X}_0^r \Rightarrow \hat{X}
\]
in $D[0, t]$, as $r \rightarrow \infty$. Here $\hat{X}$ is the Brownian motion defined in §5.3. For $s \leq t$, denote
\[
\hat{Z}_0^r(s) = \left( \Phi(\hat{X}_0^r(s)) - \sum_{j \in J} m_j^r (\hat{Q}_j^r(s) - \lambda_j^r \hat{d}_j)^+ - \sum_{j \in J} m_j^r \lambda_j^r \hat{d}_j \right)^+.
\]
Then by the continuity of $\Phi$ and the definition of asymptotic compliance, in $D[0, t]$ as $r \rightarrow \infty$,
\[
\hat{Z}_0^r(\cdot) \Rightarrow \left( \hat{Q}_w(\cdot) - \hat{\omega} \right)^+.
\]
From (EC.18), on $\Gamma^r_0(t)$,
$$\sum_{k \in \mathcal{K}} m^r \hat{Q}^r_k(s) \geq \hat{Z}^r_+(s), \quad s \leq t.$$

By the definition of $\Delta_{\mathcal{K}}$ and the nondecreasing property of $\Delta_k$, for all $k \in \mathcal{K}$, we have
$$\Gamma^r_1(t) \cup \Gamma^r_2(t) \supseteq \left\{ \int_0^t \sum_{k \in \mathcal{K}} C_k \left( \Delta_k \left( \hat{Z}^r_+(s) \right) \right) ds > x, \max_{k \in \mathcal{K}} \sup_{0 \leq s \leq t} \hat{Q}^r_k(s) \leq r^{-1/4} \right\} \cup \Gamma^r_2(t)$$
$$\supseteq \left\{ \int_0^t \sum_{k \in \mathcal{K}} C_k \left( \Delta_k \left( \hat{Z}^r_+(s) \right) \right) ds > x \right\}.$$ 

Combined with (EC.21),
$$\liminf_{r \to \infty} \mathbb{P} \{ U^r(t) > x \} \geq \liminf_{r \to \infty} \mathbb{P} \left\{ \int_0^t \sum_{k \in \mathcal{K}} C_k \left( \Delta_k \left( \hat{Z}^r_+(s) \right) \right) ds > x \right\}.$$ 

From the convergence of $\hat{Z}^r_+$, the right-hand side is exactly the lower bound in Theorem 2. This completes the proof. \[\square\]

**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 $\hat{Q}^r_w(t) = \Phi(\hat{X}^r_w(t))$. As a result, it is enough to prove the convergence of $\hat{X}^r_w$. 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.

**Proof of Proposition 1:** For any family of work-conserving policies, in addition to (EC.17),
the following is also true:
$$\hat{X}^r_+ \text{ increases at } t \text{ only when } \hat{Q}^r_w(t) = 0.$$ 

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

From (EC.10), (EC.1) and the fact that $\hat{T}^r_j(s) \leq s, j \in \mathcal{J}$, and $\hat{T}^r_k(s) \leq s, k \in \mathcal{K}$, it is easy to see that $\hat{X}^r_w$ in (EC.14) is stochastically bounded. By the Lipschitz continuity of $\Phi$ (Theorem 6.1 in Chen and Yao (2001)), $\hat{Q}^r_w$ is stochastically bounded, which implies the stochastic boundedness of $\hat{Q}^r_j, j \in \mathcal{J}$, and $\hat{Q}^r_k, k \in \mathcal{K}$. Then $\hat{Q}^r_j \Rightarrow 0$, for $j \in \mathcal{J}$. Note that (EC.6) is still true (under work-conserving policies). One then has
$$\hat{T}^r_j(\cdot) \Rightarrow \lambda_j m_j e(\cdot), \quad j \in \mathcal{J}. \quad (EC.24)$$

For $k \in \mathcal{K}$, following the procedure in proving (EC.22) in the proof of Theorem 2, one also has
$$\hat{T}^r_k(\cdot) \Rightarrow \lambda_k m_k e(\cdot), \quad k \in \mathcal{K}. \quad (EC.25)$$
Now (EC.24) and (EC.25), together with (EC.10), (EC.1) and the Random-Time-Change theorem, imply that, as \( r \to \infty \),
\[
\hat{X}_n^r \Rightarrow \hat{X}.
\]  

(EC.26)

By the continuity of the mapping \( \Phi \), (14) follows.

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

We now analyze the family of control policies \( \{\pi_i^r\} \) and prove Theorem 3. The proof is separated into two parts: in the first part (§EC.4.1 and §EC.4.2) we follow the standard framework in Bramson (1998) to handle triage patients; in the second part (§EC.4.3) we follow the method in van Mieghem (1995) to handle IP patients.

**EC.4.1. Hydrodynamic limit**

In the present subsection, we focus on triage patients.

Under the policies \( \{\pi_i^r\} \), the following dynamic equations hold:

\[
Q_j^r(t) = Q_j^r(0) + E_j^r(t) - D_j^r(t), \quad j \in \mathcal{J},
\]

\[
D_j^r(t) = S_j(T_j^r(t)), \quad j \in \mathcal{J},
\]

\[
Q_k^r(t) = Q_k^r(0) + E_k^r(t) - D_k^r(t), \quad k \in \mathcal{K},
\]

\[
E_k^r(t) = \sum_{j \in \mathcal{J}} \Phi_{jk}^r(S_j(T_j^r(t))) + \sum_{l \in \mathcal{K}} \Phi_{lk}^r(S_l(T_l^r(t))), \quad k \in \mathcal{K},
\]

\[
D_k^r(t) = S_k(T_k^r(t)), \quad k \in \mathcal{K},
\]

\[
\sum_{j \in \mathcal{J}} [T_j^r(t) - T_j^r(s)] + \sum_{k \in \mathcal{K}} [T_k^r(t) - T_k^r(s)] \leq t - s, \quad \text{for } s < t,
\]

\[
Y^r(t) = t - \left( \sum_{j \in \mathcal{J}} T_j^r(t) + \sum_{k \in \mathcal{K}} T_k^r(t) \right),
\]

\[
\int_0^\infty \left( \max_{j \in \mathcal{J}} \frac{\tau_j^r(t)}{d_j^r} - \frac{\tau_{j'}^r(t)}{d_{j'}^r} \right)^+ \wedge 1 \, dT_{j'}^r(t) = 0, \quad j' \in \mathcal{J},
\]

\[
\int_0^\infty 1 \left( \max_{j \in \mathcal{J}} (\tau_j^r(t) - d_j^r) > -\epsilon \right) d \sum_{k \in \mathcal{K}} T_k^r(t) = 0,
\]

\[
\int_0^\infty 1 \left( \max_{j \in \mathcal{J}} (\tau_j^r(t) - d_j^r) \leq -\epsilon, \sum_{k \in \mathcal{K}} Q_k^r(t) > 0 \right) d \sum_{j \in \mathcal{J}} T_j^r(t) = 0,
\]

\[
\int_0^\infty 1 \left( \sum_{j \in \mathcal{J}} m_j^r Q_j^r(t) + \sum_{k \in \mathcal{K}} m_k^r Q_k^r(t) > 0 \right) dY^r(t) = 0.
\]

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

\[
\bar{E}_j^r(t) = r^{-1} E_j^r(rt), \quad \bar{S}_j^r(t) = r^{-1} S_j(rt), \quad \bar{\tau}_j^r(t) = r^{-1} \tau_j^r(rt),
\]

\[
\bar{T}_j^r(t) = r^{-1} T_j^r(rt), \quad \bar{Q}_j^r(t) = r^{-1} Q_j^r(rt), \quad \bar{D}_j^r(t) = r^{-1} D_j^r(rt),
\]
and for $k$-IP classes, $k \in \mathcal{K}$,
\[
\begin{align*}
\bar{E}_k(t) &= r^{-1} E_k'(rt), \\
\bar{S}_k(t) &= r^{-1} S_k(rt), \\
\bar{T}_k(t) &= r^{-1} T_k'(rt), \\
\bar{Q}_k(t) &= r^{-1} Q_k'(rt), \\
\bar{D}_k(t) &= r^{-1} D_k'(rt).
\end{align*}
\]

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

**Lemma EC.4.1** For any $T > 0$, $\sup_{0 \leq t \leq T} |\lambda_j \tau_j(t) - \bar{Q}_j(t)| \Rightarrow 0$.

**Proof:** For each triage class $j \in \mathcal{J}$, the patients in queue at time $t$ are those patients arriving between $[t - \tau_j(t), t]$, thus
\[
|Q_j'(t) - (E_j'(t) - E_j'((t - \tau_j(t)))| \leq 1.
\]
Then
\[
|\bar{Q}_j(t) - (\bar{E}_j'(t) - E_j'((t - \tau_j(t))))| \leq \frac{1}{r}, \quad j \in \mathcal{J}.
\] (EC.27)

The lemma now follows from the functional law of large numbers, $\sup_{0 \leq t \leq T} |\bar{E}_j'(t) - \lambda_j t| \Rightarrow 0$, and (EC.27).

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

**Proposition 7** Assume $Q_j'(0) \Rightarrow \bar{Q}_j(0), j \in \mathcal{J}$, and $Q_k'(0) \Rightarrow \bar{Q}_k(0), k \in \mathcal{K}$, as $r \rightarrow \infty$. Then under the proposed family of control policies, almost surely, every sequence contains a subsequence $\{r_n\}$ such that, the hydrodynamic scaled processes $\bar{E}_j', \bar{S}_j', \bar{\tau}_j, \bar{T}_j', \bar{Q}_j', \bar{D}_j', j \in \mathcal{J}$, $\bar{E}_k', \bar{S}_k', \bar{T}_k', \bar{Q}_k', \bar{D}_k', k \in \mathcal{K}$, 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

\[
\begin{align*}
\bar{Q}_j(t) &= \bar{Q}_j(0) + \lambda_j t - \bar{D}_j(t), \quad j \in \mathcal{J}, \\
\bar{D}_j(t) &= \mu_j \bar{T}_j(t), \quad j \in \mathcal{J}, \\
\bar{Q}_k(t) &= \bar{Q}_k(0) + \bar{E}_k(t) - \bar{D}_k(t), \quad k \in \mathcal{K}, \\
\bar{E}_k(t) &= \sum_{j \in \mathcal{J}} \mu_j P_{jk} \bar{T}_j(t) + \sum_{l \in \mathcal{K}} \mu_l P_{lk} \bar{T}_l(t), \quad k \in \mathcal{K}, \\
\bar{D}_k(t) &= \mu_k \bar{T}_k(t), \quad k \in \mathcal{K}, \\
\lambda_j \bar{\tau}_j(t) &= \bar{Q}_j(t), \quad j \in \mathcal{J}, \\
\sum_{j \in \mathcal{J}} [\bar{T}_j(t) - \bar{T}_j(s)] + \sum_{k \in \mathcal{K}} [\bar{T}_k(t) - \bar{T}_k(s)] &\leq t - s, \quad \text{for } s < t, \\
\bar{Y}(t) &= t - \left( \sum_{j \in \mathcal{J}} \bar{T}_j(t) + \sum_{k \in \mathcal{K}} \bar{T}_k(t) \right), \\
\int_0^\infty \left( \max_{j \in \mathcal{J}} \frac{\bar{Q}_j(t)}{\lambda_j \bar{d}_j} - \frac{\bar{Q}_j'(t)}{\lambda_j' \bar{d}_j'} \right) + 1 \, d\bar{T}_j(t) &= 0, \quad j' \in \mathcal{J}, \quad \text{(EC.36)}
\end{align*}
\]
\[
\int_0^\infty 1 \left( \max_{j\in J} (\bar{Q}_j(t) - \lambda_j \hat{d}_j) > 0 \right) d' \sum_{k \in K} \bar{T}_k(t) = 0, \quad \text{(EC.37)}
\]
\[
\int_0^\infty 1 \left( \max_{j\in J} (\bar{Q}_j(t) - \lambda_j \hat{d}_j) < 0, \sum_{k \in K} \bar{Q}_k(t) > 0 \right) d \sum_{j \in J} \bar{T}_j(t) = 0, \quad \text{(EC.38)}
\]
\[
\int_0^\infty 1 \left( \sum_{j \in J} m_j^e \bar{Q}_j(t) + \sum_{k \in K} m_k^e \bar{Q}_k(t) > 0 \right) d\bar{Y}(t) = 0. \quad \text{(EC.39)}
\]

**Remark 8** We call any \( \bar{S} = (\bar{E}_j, \bar{S}_j, \bar{\tau}_j, \bar{T}_j, \bar{Q}_j, \bar{D}_j, j \in J, \bar{E}_k, \bar{S}_k, \bar{T}_k, \bar{Q}_k, \bar{D}_k, k \in K) \) satisfying (EC.28)-(EC.39) 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
\[
\sum_{j \in J} m_j^e \bar{Q}_j(t) + \sum_{k \in K} m_k^e \bar{Q}_k(t) = \sum_{j \in J} m_j^e \bar{Q}_j(0) + \sum_{k \in K} m_k^e \bar{Q}_k(0).
\]

**Proof:** From the fact that \( \sum_{j \in J} \lambda_j m_j^e = 1 \), (EC.15)-(EC.16) and (EC.28)-(EC.32), one gets
\[
\sum_{j \in J} m_j^e \bar{Q}_j(t) + \sum_{k \in K} m_k^e \bar{Q}_k(t) = \sum_{j \in J} m_j^e \bar{Q}_j(0) + \sum_{k \in K} m_k^e \bar{Q}_k(0) + \bar{Y}(t).
\]
From (EC.39), (EC.34) and (EC.35), we deduce that \( \bar{Y}(\cdot) = 0 \). This completes the proof. \( \square \)

**EC.4.2. State-space collapse for triage patients**

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

**Proposition 9 (State-space collapse for hydrodynamic model solutions)** Fix \( C > 0 \). For any hydrodynamic model solution with \( \sum_{j \in J} m_j^e \bar{Q}_j(0) + \sum_{k \in K} m_k^e \bar{Q}_k(0) < C \), there exists a constant \( T_0 \) such that, for all \( t \geq T_0 \),
\[
\bar{Q}_J(t) = \Delta_J \min \left( \sum_{j \in J} m_j^e \bar{Q}_j(t) + \sum_{k \in K} m_k^e \bar{Q}_k(t), \bar{\omega} \right).
\]
Furthermore, if
\[
\bar{Q}_J(0) = \Delta_J \min \left( \sum_{j \in J} m_j^e \bar{Q}_j(0) + \sum_{k \in K} m_k^e \bar{Q}_k(0), \bar{\omega} \right),
\]
then \( \bar{Q}_J(t) = \bar{Q}_J(0) \), for all \( t \geq 0 \).
Proof: For \( j \in J \), we define

\[
    f_j(t) = \frac{1}{\lambda_j d_j} \left( \bar{Q}_j(t) - \Delta_j \min \left( \sum_{j \in J} \sum_{k \in K} m_j^e \bar{Q}_j(0) + \sum_{k \in K} m_k^e \bar{Q}_k(0), \bar{\omega} \right) \right) .
\]

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

\[
    f_j'(t) = -\frac{1}{d_j} < 0.
\]

Indeed, if this is not true, then \( \bar{T}_j(t) \neq 0 \) and from (EC.36), one has \( \frac{\bar{Q}_j(t)}{\lambda_j d_j} = \max_{j' \in J} \frac{\bar{Q}_{j'}(t)}{\lambda_j d_j} \). Together with \( f_j(t) > 0 \), one can prove by contradiction that \( \bar{Q}_j(t) < \lambda_j \hat{d}_j \), which then implies \( \max_{j' \in J} (\bar{Q}_{j'}(t) - \lambda_j \hat{d}_j') < 0 \). Then from (EC.38), one has \( \bar{Q}_k(t) = 0 \), for all \( k \in K \). This, together with \( f_j(t) > 0 \), contradict the definition of \( \Delta_j \).

As a result, \( f_j \) will decrease to 0 in a finite time (denote it by \( T_j \)) and once becoming 0, it will never be positive again. As there are finite number of triage classes, consequently, after a finite time (denote it 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 \( j \in J \). Define

\[
    g_j(t) = \frac{1}{\lambda_j d_j} \left( \bar{Q}_j(t) - \Delta_j \min \left( \sum_{j \in J} \sum_{k \in K} m_j^e \bar{Q}_j(0) + \sum_{k \in K} m_k^e \bar{Q}_k(0), \bar{\omega} \right) \right)^+ .
\] (EC.40)

We can assume \( g_1(t) > 0 \) whenever \( \sum_{j \in J} \lambda_j \hat{d}_j m_j g_j(t) > 0 \). Otherwise, if \( g_1(t) = 0 \) and there is another \( j \in J \) such that \( g_j(t) > 0 \), then from the definition of \( \Delta_j \), \( \bar{Q}_j(t)/\lambda_j \hat{d}_j < \max_{j' \in J} (\bar{Q}_{j'}(t)/\lambda_j \hat{d}_j') \) and from (EC.36), \( \bar{T}_1(t) = 0 \) and \( g_1(t) = \frac{1}{d_1} > 0 \). Hence right after \( t \), \( g_1(\cdot) \) will be positive.

Now, as we have proved that \( f_j(t) = 0 \), for all \( j \in J \) and over \( t \geq T_2 \), together with \( g_1(t) > 0 \) and the definition of \( \Delta_j \), we have \( \sum_{j \in J} \sum_{k \in K} m_j^e \bar{Q}_j(t) + \sum_{k \in K} m_k^e \bar{Q}_k(t) > \bar{\omega}, \sum_{k \in K} \bar{Q}_k(t) > 0 \), and for \( 1 \in J \), \( \bar{Q}_1(t) > \lambda_1 \hat{d}_1 \). Then from (EC.37), \( \sum_{k \in K} \bar{T}_k(t) = 0 \). From (EC.39), we know \( \sum_{j \in J} \bar{T}_j(t) = 1 \). As a result, the derivative of \( \sum_{j \in J} \lambda_j \hat{d}_j m_j g_j(t) \) is

\[
    \sum_{j \in J} \lambda_j m_j - 1 < 0.
\]

Thus in finite time (denote it by \( T_0 \geq T_2 \)), \( \sum_{j \in J} \lambda_j \hat{d}_j m_j g_j(t) \) will hit 0. It follows that, for all \( t \geq T_0 \), \( f_j(t) = g_j(t) = 0 \), \( j \in J \). Finally, from Proposition 8, \( \bar{Q}_J(t) = \Delta_J \min \left( \sum_{j \in J} m_j^e \bar{Q}_j(0) + \sum_{k \in K} m_k^e \bar{Q}_k(0), \bar{\omega} \right) = \Delta_J \min \left( \sum_{j \in J} m_j^e \bar{Q}_j(t) + \sum_{k \in K} m_k^e \bar{Q}_k(t), \bar{\omega} \right) \)

for \( t \geq T_0 \).

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 Bramson (1998)'s framework. For completeness, we include it here.
**Proposition 10** Under Assumption 1 and the proposed family of control policies, as \( r \to \infty \),

\[
\sup_{0 \leq t \leq T} \left| \hat{Q}_j^r(t) - \Delta_j \min \left( \hat{Q}_w^r(t), \hat{\omega} \right) \right| \to 0.
\]

**Proof:** From Lemma 9, we know that Assumption 3.2 of Bramson (1998) holds. Then from Theorem 5 of Bramson (1998), we deduce “multiplicative state-space collapse” (Bramson (1998)’s equation (3.41)):

\[
\sup_{0 \leq t \leq T} \left| \hat{Q}_j^r(t) - \Delta_j \min \left( \hat{Q}_w^r(t), \hat{\omega} \right) \right| \to 0,
\]

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

Next, our Proposition 1 implies that \( \sup_{0 \leq t \leq T} \hat{Q}_w^r(t) \vee 1 \) is stochastically bounded. As a result,

\[
\sup_{0 \leq t \leq T} \left| \hat{Q}_j^r(t) - \Delta_j \min \left( \hat{Q}_w^r(t), \hat{\omega} \right) \right| \to 0,
\]

which proves the proposition. \( \square \)

**EC.4.3. State-space collapse for IP patients**

From Propositions 1 and 10, as \( r \to \infty \), one has

\[
\sum_{k \in K} m_k^e \hat{Q}_k^r \Rightarrow \left( \hat{Q}_w - \hat{\omega} \right)^+.
\]  

(EC.41)

Recall that the proposed rule for IP patients ensures

\[
\max_{l, k \in K} \sup_{0 \leq t \leq T} \left| \frac{C_l \left( \hat{Q}_l^r(t) \right)}{m_l^e} - \frac{C_k \left( \hat{Q}_k^r(t) \right)}{m_k^e} \right| \Rightarrow 0. \quad (EC.42)
\]

**Proposition 11** Under the family of control policies \( \{ \pi_r^* \} \), we have \( \left( \hat{Q}_k^r, k \in K \right) \Rightarrow \left( \hat{Q}_k, k \in K \right) \). Here

\[
\hat{Q}_k = \Delta_k \left( \left( \hat{Q}_w - \hat{\omega} \right)^+ \right), \quad k \in K.
\]  

(EC.43)

**Proof:** The proof is similar to van Mieghem (1995); for completeness, we include it here. From (EC.42), for any given \( T > 0 \),

\[
\max_{l, k \in K} \sup_{0 \leq t \leq T} \left| C_l^\nu \left( \frac{m_l^e}{m_k^e} C_k^\nu \left( \hat{Q}_k^r(t) \right) \right) - \hat{Q}_l(t) \right| \Rightarrow 0.
\]

(EC.44)

From the assumption on \( C_k^\nu, k \in K \), \( C_l^\nu \left( \frac{m_l^e}{m_k^e} C_k^\nu \left( \cdot \right) \right) \) is a nondecreasing function.

From (EC.44) and (EC.41), we have

\[
\sum_{l \in K} m_l^e C_l^\nu \left( \frac{m_l^e}{m_k^e} C_k^\nu \left( \hat{Q}_k^r \right) \right) \Rightarrow \left( \hat{Q}_w - \hat{\omega} \right)^+.
\]  

(EC.44)
As the function on the left-hand of the above equation has a continuous inverse, \( \hat{Q}_r \) converges. From (EC.44), \((\hat{Q}_r, l \in K) \Rightarrow (\hat{Q}_l, l \in K)\). We also have,

\[
\frac{C'_l(\hat{Q}_l)}{m'_l} = \frac{C'_k(\hat{Q}_k)}{m'_k}, \quad l, k \in K.
\]

We have thus proved (EC.43).

Proof of Theorem 3: This can be deduced from Propositions 1, 10 and 11.

**EC.5. Proof of Theorem 1: Asymptotic Optimality**

Proof of Theorem 1: First, it can be verified that \( \Delta_j \min (x, \hat{\omega}) \leq \lambda_j \hat{d}_j \), for any \( x \) and \( j \in J \). Then from Theorem 3, under the proposed policies \( \{\pi^*_r\} \), \( \hat{Q}_j \Rightarrow \hat{Q}_j \leq \lambda_j \hat{d}_j \). An analysis of work-conserving policies will show that (EC.2) is still true (see Lemma EC.6.2), hence \( \hat{Q}_j \Rightarrow \hat{Q}_j \leq \lambda_j \hat{d}_j \) 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

\[
\int_0^t \sum_{k \in K} C_k \left( \hat{Q}_k(s) \right) ds \Rightarrow \int_0^t \sum_{k \in K} C_k \left( \hat{Q}_k(s) \right) ds = \int_0^t \sum_{k \in K} C_k \left( \Delta_k \left( (\hat{Q}_w(s) - \hat{\omega})^+ \right) \right) ds.
\]

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. These stochastic boundedness results 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 \( \hat{\tau}_r^j \) for work-conserving policies, until we justify such boundedness independently.

From the discussion in the proof of Proposition 1, \( \hat{Q}_r^j, j \in J \), 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 \( \hat{Q}_r^j, j \in J \), are stochastically bounded, \( \hat{\tau}_r^j, j \in J \), are also stochastically bounded.

Next consider IP patients. Define \( \hat{Y}_k^r = (\hat{Y}_k^r)_{k \in K}, k \in K \), by

\[
\hat{Y}_k^r(t) = \hat{Q}_k^r(t) - \hat{Q}_k^r(0) - \hat{\xi}_k^r(t) + \hat{S}_k^r(\hat{T}_k^r(t)) - \sum_{j \in J} P_{j,k} \hat{\tau}_j^r(t),
\]
and recall that $\hat{E}^r_k$ is defined in (EC.13). Denote $\hat{T}^r_\mu = (\mu_k \hat{T}^r_k)_{k \in \mathcal{K}}$. Then from (EC.12),

$$\hat{T}^r_\mu = (P^T - I)^{-1} \hat{\gamma}^r_K.$$  \hfill (EC.45)

We can easily verify the stochastic boundedness of $\hat{\gamma}^r_K$ from the facts $\hat{T}^r_j(s) \leq s$ and $\hat{T}^r_k(s) \leq s$, for all $j \in \mathcal{J}$ and $k \in \mathcal{K}$. This implies the stochastic boundedness of $\hat{T}^r_\mu$, then $\hat{T}^r_K = (\hat{T}^r_k)_{k \in \mathcal{K}}$.

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

$$\hat{E}^r_k(t) = \hat{E}^r_k(0) + \sum_{j \in \mathcal{J}} P_{jk} \mu_j \hat{T}^r_j(t) + \sum_{l \in \mathcal{K}} P_{kl} \mu_l \hat{T}^r_l(t).$$  \hfill (EC.46)

The stochastic boundedness of $\hat{E}^r_k$ can be then obtained from the stochastic boundedness of $\hat{E}^r_k$, $\hat{T}^r_j$ and $\hat{T}^r_l$ ($j \in \mathcal{J}$, $k, l \in \mathcal{K}$).

Define the fluid-scaled virtual waiting time processes as

$$\bar{\omega}^r_j(t) = r^{-2} \omega^r_j(r^2 t), \quad j \in \mathcal{J}, \quad \bar{\omega}^r_k(t) = r^{-2} \omega^r_k(r^2 t), \quad k \in \mathcal{K}.$$

First we prove the following:

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

$$\bar{\omega}^r_j \Rightarrow 0, \quad j \in \mathcal{J},$$  \hfill (EC.47)

$$\bar{\omega}^r_k \Rightarrow 0, \quad k \in \mathcal{K}. $$  \hfill (EC.48)

**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 $\omega^r_j(t) \geq \epsilon$, then

$$S_j(T^r_j(t + \epsilon)) \leq Q^r_j(0) + E^r_j(t).$$

Then $\bar{\omega}^r_j(t) \geq \epsilon$ ensures

$$\bar{S}^r_j(\hat{T}^r_j(t + \epsilon)) + \mu_j \bar{T}^r_j(t + \epsilon) + \lambda_j r \epsilon \leq \bar{Q}^r_j(0) + \bar{E}^r_j(t).$$

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

$$\mathbb{P} \left\{ \sup_{0 \leq t \leq T} \bar{\omega}^r_j(t) \geq \epsilon \right\} \leq \mathbb{P} \left\{ \lambda_j r \epsilon \leq \sup_{0 \leq t \leq T} |\bar{Q}^r_j(0) + \bar{E}^r_j(t) - \bar{S}^r_j(\hat{T}^r_j(t + \epsilon)) - \mu_j \hat{T}^r_j(t + \epsilon)| \right\}.$$

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

$$\lim_{r \to \infty} \mathbb{P} \left\{ \sup_{0 \leq t \leq T} \bar{\omega}^r_j(t) \geq \epsilon \right\} = 0.$$

This completes the proof. \hfill \square
Lemma EC.6.2 Under any family of work-conserving policies, for any given $T > 0$, as $n \to \infty$, we have

$$\sup_{0 \leq t \leq T} \left| \lambda^*_j \hat{\tau}^*_j(t) - \hat{Q}^*_j(t) \right| \Rightarrow 0, \quad j \in J.$$

Proof: The proof follows exactly that of Lemma EC.1.1, by starting with $\sup_{0 \leq t \leq T} \hat{\tau}^*_j(s) \Rightarrow 0$, which is a consequence of Lemma EC.6.1 and $\sup_{s \leq t} \hat{\tau}^*_j(s) \leq \sup_{s \leq t} \omega^*_j(s)$, for all $t$ and $j$. Note that the assumptions here slightly differ from Lemma EC.1.1. In the latter, the stochastic boundedness of $\hat{\tau}^*_j, j \in J$, follows from asymptotic compliance, while here we do not have the stochastic boundedness of $\hat{\tau}^*_j, j \in J$, in advance.

EC.7. Proof of Proposition 2: Asymptotic Sample-Path Little’s Law

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

$$\left( \hat{T}^*_j, j \in J, \hat{E}^*_k, \hat{T}^*_k, k \in K \right) \Rightarrow \left( \bar{T}^*_j, j \in J, \hat{E}^*_k, \hat{T}^*_k, k \in K \right),$$

for some continuous processes $\left( \hat{T}^*_j, j \in J, \hat{E}^*_k, \hat{T}^*_k, k \in K \right)$ satisfying

$$\mu_j \hat{T}^*_j(t) = -\hat{Q}^*_j(t) + \hat{E}^*_j(t) - \hat{S}^*_j (\lambda_j m_j t), \quad (EC.49)$$

$$\hat{E}^*_k(t) = \bar{E}^*_k(t) + \sum_{j \in J} P_{jk} \mu_j \hat{T}^*_j(t) + \sum_{l \in K} P_{lk} \mu_l \hat{T}^*_l(t), \quad (EC.50)$$

$$\left( P^T - I \right) \left( \mu_k \hat{T}^*_k \right)_{k \in K} = \hat{Y}^*_k. \quad (EC.51)$$

Here

$$\bar{E}^*_k(t) = \sum_{j \in J} \hat{E}_{jk}(\lambda_j t) + \sum_{l \in K} \hat{E}_{lk}(\lambda_l t) + \sum_{j \in J} P_{jk} \hat{S}^*_j (\lambda_j m_j t) + \sum_{l \in K} P_{lk} \hat{S}^*_l (\lambda_l m_l t),$$

$$\hat{Y}^*_k(t) = \bar{Q}^*_k(t) - \bar{E}^*_k(t) + \bar{S}^*_k (\lambda_k m_k t) - \sum_{j \in J} P_{jk} \mu_j \hat{T}^*_j(t).$$

Proof: From (EC.7), (EC.46) and (EC.45), we have $(\hat{T}^*_\mu = (\mu_k \hat{T}^*_k)_{k \in K})$

$$\hat{T}^*_j(t) = \left[ \hat{Q}^*_j(0) - \hat{Q}^*_j(t) + \hat{E}^*_j(t) - \hat{S}^*_j (\hat{T}^*_j(t)) \right] / \mu_j, \quad (EC.52)$$

$$\bar{E}^*_k(t) = \bar{E}^*_k(t) + \sum_{j \in J} P_{jk} \mu_j \hat{T}^*_j(t) + \sum_{l \in K} P_{lk} \mu_l \hat{T}^*_l(t), \quad (EC.53)$$

$$\hat{T}^*_\mu(t) = \left( P^T - I \right)^{-1} \hat{Y}^*_k(t), \quad (EC.54)$$

where

$$\hat{E}^*_k(t) = \sum_{j \in J} \hat{E}_{jk} \left( \hat{S}^*_j \left( \hat{T}^*_j(t) \right) \right) + \sum_{l \in K} \hat{E}_{lk} \left( \hat{S}^*_l \left( \hat{T}^*_l(t) \right) \right) + \sum_{j \in J} P_{jk} \hat{S}^*_j \left( \hat{T}^*_j(t) \right) + \sum_{l \in K} P_{lk} \hat{S}^*_l \left( \hat{T}^*_l(t) \right),$$

$$\hat{Y}^*_k(t) = \bar{Q}^*_k(t) - \bar{E}^*_k(t) + \bar{S}^*_k (\hat{T}^*_k(t)) - \sum_{j \in J} P_{jk} \mu_j \hat{T}^*_j(t).$$

As a result, $\left( \hat{T}^*_j, j \in J, \hat{E}^*_k, \hat{T}^*_k, k \in K \right)$ can be represented as a continuous mapping from $\left( \hat{Q}^*_j, \hat{E}^*_j, \hat{S}^*_j, \bar{E}^*_j, \hat{E}_{jk}, \bar{E}_{jk}, \hat{S}^*_k, \bar{S}^*_k, \bar{T}^*_j, j \in J, l, k \in K \right)$, the convergence of which can be obtained
Proof of Proposition 2: We prove the result only for \( T > 0 \), verified from (EC.52)-(EC.54). This completes the proof.

Proof of Proposition 3: We prove the result only for \( j \)-triage patients. For \( k \)-IP patients, the proof is similar. The convergence of \( \hat{Q}_j^r \), together with Lemma EC.6.1, ensure that, for any \( T > 0 \),

\[
\sup_{0 \leq t \leq T} \left| \hat{Q}_j^r(t) - \hat{Q}_j^r(t + \hat{\omega}_j^r(t)) \right| \Rightarrow 0, \quad \text{as} \quad r \to \infty.
\]

Thus it is enough to prove

\[
\sup_{0 \leq t \leq T} \left| \lambda_j^r \hat{\omega}_j^r(t) - \hat{Q}_j^r(t + \hat{\omega}_j^r(t)) \right| \Rightarrow 0, \quad \text{as} \quad r \to \infty.
\]

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

\[
Q_j^r(t + \omega_j^r(t)) \leq E_j^r(t + \omega_j^r(t)) - E_j^r(t) \leq Q_j^r(t + \omega_j^r(t)) + \Delta S_j^r(t + \omega_j^r(t)); \quad (EC.55)
\]

here, with some abuse of notation, \( \Delta S_j^r(t + \omega_j^r(t)) = S_j(T^r(t + \omega_j^r(t))) - S_j(T^r(t + \omega_j^r(t)-)) \).

From this relationship, we deduce the following for the diffusion-scaled processes:

\[
\left| \lambda_j^r \hat{\omega}_j^r(t) - \hat{Q}_j^r(t + \hat{\omega}_j^r(t)) \right| \leq \left| \hat{E}_j^r(t + \hat{\omega}_j^r(t)) - \hat{E}_j^r(t) \right| + \Delta S_j^r(t + \hat{\omega}_j^r(t)) + \mu_j \Delta \hat{T}_j^r(t + \hat{\omega}_j^r(t)).
\]

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

EC.8. Proof of Proposition 3: Snapshot Principle—Virtual Waiting Time and Age

Lemma EC.8.1 Under the family of control policies \( \{\pi_k^r\} \), for any given \( T > 0 \), as \( r \to \infty \),

\[
\sup_{0 \leq t \leq T} \left| \lambda_k^r \hat{\tau}_k^r(t) - \hat{Q}_k^r(t) \right| \Rightarrow 0, \quad k \in K.
\]

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

Proof of Proposition 3: This can be deduced from Proposition 2, Lemmas EC.6.2 and EC.8.1.
EC.9. Proof of Proposition 4: Snapshot Principle—Sojourn Time and Queue Lengths

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

\[
t \leq \zeta_{jki}(t) \leq \tau_{jk}(t) + W_{jk}(t). \tag{EC.56}
\]

Define the fluid-scaled processes

\[
\tilde{\zeta}_{jki}(t) = r^{-2}\zeta_{jki}(r^2t), \quad \tilde{W}_{jk}(t) = r^{-2}W_{jk}(r^2t), \quad \tilde{\tau}_{jk}(t) = r^{-2}\tau_{jk}(r^2t).
\]

Lemma EC.9.1 Under the family of control policies \( \{\pi^*_i\} \), with FCFS among each IP class, if \( h \) is \( j \)-feasible, then for any \( T \geq 0 \), as \( r \to \infty \),

\[
\sup_{0 \leq t \leq T} \tilde{W}_{jk}(t) \Rightarrow 0, \tag{EC.57}
\]

\[
\sup_{0 \leq t \leq T} [\tilde{\tau}_{jk}(t) - t] \Rightarrow 0. \tag{EC.58}
\]

Consequently, as \( r \to \infty \),

\[
\sup_{0 \leq t \leq T} [\tilde{\zeta}_{jki}(t) - t] \Rightarrow 0. \tag{EC.59}
\]

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

Proof of Proposition 4: The sojourn time \( W_{jk}(t) \) can be represented as

\[
W_{jk}(t) = \omega^*_j(\tau_{jk}(t)) + \sum_{k \in \mathcal{K}} h_k \sum_{i=1}^{h_k} \omega^*_k(\zeta_{jki}(t)).
\]

From this we get

\[
\tilde{W}_{jk}(t) - \left[ \frac{\tilde{Q}_j(t)}{\lambda_j} + \sum_{k \in \mathcal{K}} h_k \frac{\tilde{Q}_k(t)}{\lambda_k} \right] \\
= \tilde{\omega}^*_j(\tilde{\tau}_{jk}(t)) + \sum_{k \in \mathcal{K}} h_k \sum_{i=1}^{h_k} \tilde{\omega}^*_k(\tilde{\zeta}_{jki}(t)) - \left[ \frac{\tilde{Q}_j(t)}{\lambda_j} + \sum_{k \in \mathcal{K}} h_k \frac{\tilde{Q}_k(t)}{\lambda_k} \right] \\
= \left[ \tilde{\omega}^*_j(t) - \frac{\tilde{Q}_j(t)}{\lambda_j} \right] + \sum_{k \in \mathcal{K}} h_k \left[ \tilde{\omega}^*_k(t) - \frac{\tilde{Q}_k(t)}{\lambda_k} \right] \\
+ \left[ \tilde{\omega}^*_j(\tilde{\tau}_{jk}(t)) - \tilde{\omega}^*_j(t) \right] + \sum_{k \in \mathcal{K}} h_k \left[ \tilde{\omega}^*_k(\tilde{\zeta}_{jki}(t)) - \tilde{\omega}^*_k(t) \right].
\]

From Lemma EC.9.1 and the convergence of \( \tilde{\omega}^*_j, j \in \mathcal{J} \) and \( \tilde{\omega}^*_k, k \in \mathcal{K} \),

\[
\left[ \tilde{\omega}^*_j(\tilde{\tau}_{jk}(t)) - \tilde{\omega}^*_j(t) \right] + \sum_{k \in \mathcal{K}} h_k \left[ \tilde{\omega}^*_k(\tilde{\zeta}_{jki}(t)) - \tilde{\omega}^*_k(t) \right] \Rightarrow 0.
\]
Similarly to Reiman (1984), denote \( \|h\| = \sum_{k=1}^{K} h_k \). Then we have
\[
P \left\{ \sup_{0 \leq t \leq T} \tilde{W}^r_{jh}(t) \geq \epsilon \right\} \leq \max_{k \in K} P \left\{ \sup_{0 \leq t \leq T+\epsilon} \tilde{g}^r_k(t) \geq \frac{\epsilon}{\|h\| + 1} \right\} + P \left\{ \sup_{0 \leq t \leq T+\epsilon} \tilde{g}^r_j(t) \geq \frac{\epsilon}{\|h\| + 1} \right\}.
\]

From Lemma EC.6.1, the right-hand side of (EC.60) converges to 0, hence (EC.57) holds.

Let \( L_{i,j,h} = \min\{n > i; \hat{h}^r(j,n) = h\} \), where \( \hat{h}^r(j,n) \) is the visit vector associated with the \( n \)-th \( j \)-triage patient. We can write
\[
P \left\{ \sup_{0 \leq t \leq T} [\hat{\tau}^r_{j,h}(t) - t] \geq \epsilon \right\}
\leq \inf_{0 \leq t \leq T} \left\{ E^r_j(r^2t + r^2\epsilon) - E^r_j(r^2t) \right\} < \frac{1}{2} \lambda_j r^2 \epsilon
\]
\[+ P \left\{ E^r_j(r^2T) > 2\lambda_j r^2 \right\} + P \left\{ \sup_{1 \leq i \leq 2\lambda_j r^2} [L_{i,j,h} - i] > \frac{1}{2} \lambda_j r^2 \epsilon \right\}.
\]
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 \( \bar{g}_h = 1 - g_h \). Then
\[
P \left\{ \sup_{1 \leq i \leq 2\lambda_j r^2} [L_{i,j,h} - i] > \frac{1}{2} \lambda_j r^2 \epsilon \right\} \leq 1 - \left[ 1 - \bar{g}_h^{\frac{1}{2\lambda_j r^2} - 2\lambda_j r^2} \right] 2\lambda_j r^2 = 1 - \left[ 1 - \frac{\bar{g}_h^{\frac{1}{2\lambda_j r^2} - 2\lambda_j r^2} \epsilon}{r^2} \right] 2\lambda_j r^2.
\]
The same reason as in Reiman (1984) implies that the latter expression vanishes, as \( r \to \infty \). This establishes (EC.58).

Combining (EC.57), (EC.58) with (EC.56), now yields (EC.59).

\[\square\]

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

\[\square\]

**EC.10. 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 \( \{\hat{\pi}^r_i\} \) attains the lower bound, following the discussion in §EC.4, in particular, one requires similar state-space collapse results. As a result, \( \{\hat{\pi}^r_i\} \) 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 \( 0 \leq a < b \leq T \),
\[
\frac{1}{E^r_k(b) - E^r_k(a)} \left( \int_a^b \hat{\tau}^r_k d\hat{E}^r_k - \int_a^b \hat{Q}^r_k(s) ds \right) \Rightarrow 0.
\]
Next, as in Proposition 6 and the discussion prior to Proposition 8 of van Mieghem (1995), the following is true:

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

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

$$\begin{align*}
\min_x & \quad \sum_{k \in K} \lambda_k C_k (x_k / \lambda_k) \\
\text{s.t.} & \quad \sum_{k \in K} m_k^j x_k = a, \\
& \quad x \geq 0.
\end{align*}$$

(EC.61)

### EC.11. On the Shortest-Deadline-First rule (17)

For the policies using the Shortest-Deadline-First rule (17), we have the following lemma:

**Lemma EC.11.1** For any $T \geq 0$, as $r \to \infty$,

$$\sup_{0 \leq t \leq T} \left| \tilde{Q}_j(t) - \hat{\Delta}_j \left( \min \left\{ \tilde{Q}_j^*(t), \hat{\omega} \right\} \right) \right| = 0.$$ 

Here $\hat{\Delta}_j(a) = (\hat{\Delta}_j(a))_{j \in J}$ is defined as follows (where we assume that the indices of triage classes are ordered such that $\hat{\Delta}_j$ is increasingly in $j$, with $\hat{\Delta}_0 = 0$): if $\sum_{j \in J} \lambda_j m_j^i (\hat{d}_j - \hat{d}_{j'})^+ \leq a < \sum_{j \in J} \lambda_j m_j^i (\hat{d}_j - \hat{d}_{j'-1})^+$, then

$$\begin{align*}
\hat{\Delta}_{j_1}(a) &= \left\{ \begin{array}{ll}
\lambda_{j_1} (\hat{d}_{j_1} - \hat{d}_{j'} + (a - \sum_{j \in J} \lambda_j m_j^i (\hat{d}_j - \hat{d}_{j'})^+) / (\sum_{j \geq j'} \lambda_j m_j^i) ) , & \text{for } j_1 \geq j', \\
\lambda_{j_1} (\hat{d}_{j_1} - \hat{d}_{j'} + (a - \sum_{j \in J} \lambda_j m_j^i (\hat{d}_j - \hat{d}_{j'})^+) / (\sum_{j \geq j'} \lambda_j m_j^i) ) , & \text{for } j_1 < j'.
\end{array} \right.
\end{align*}$$

(EC.62)

**Proof:** The proof follows the framework in §EC.4.2. First, the argument in proving Proposition 7 still works for the new scheduling policies, except for equation (EC.36), which we replace by

$$\int_0^\infty \left( d_j - \frac{\tilde{Q}_j(t)}{\lambda_j} - \min_{j' \in J, \tilde{Q}_{j'}(t) \neq 0} \left\{ d_{j'} - \frac{\tilde{Q}_{j'}(t)}{\lambda_{j'}} \right\} \right)^+ 1 \text{d}T_j(t) = 0, \quad j \in J.$$ 

(EC.63)

The proof of Lemma 8 does not use (EC.36), thus it still applies to the new policies.

Next we prove that, for any fixed $C > 0$ and a hydrodynamic model solution with $\sum_{j \in J} m_j^\omega \tilde{Q}_j(0) + \sum_{k \in K} m_k^\omega \tilde{Q}_k(0) < C$, there exists a constant $\tilde{T}_0$ such that, for all $t \geq \tilde{T}_0$,

$$\tilde{Q}_J(t) = \hat{\Delta}_J \left( \min \left( \sum_{j \in J} m_j^\omega \tilde{Q}_j(t) + \sum_{k \in K} m_k^\omega \tilde{Q}_k(t), \hat{\omega} \right) \right).$$

The proof is similar to the one in §EC.4.2. We point out the differences in the following. For $j \in J$, define

$$\tilde{f}_j(t) = \frac{1}{\lambda_j d_j} \left( \tilde{Q}_j(t) - \hat{\Delta}_j \left( \min \left( \sum_{j \in J} m_j^\omega \tilde{Q}_j(t) + \sum_{k \in K} m_k^\omega \tilde{Q}_k(t), \hat{\omega} \right) \right) \right)^-. $$
If $\tilde{f}_j(t) > 0$ and is differentiable at $t$, then one can claim

$$\tilde{f}_j(t) = -\frac{1}{d_j} < 0.$$  

The proof is similar to §EC.4.2, except that we now use

$$\tilde{d}_j - \frac{Q_j(t)}{\lambda_j} = \min_{j' \in J, Q_{j'}(t) \neq 0} \left\{ \tilde{d}_{j'} - \frac{Q_{j'}(t)}{\lambda_{j'}} \right\}$$

to replace $\frac{Q_j(t)}{\lambda_j d_j}$ in the proof of Lemma 8.

Other steps follow exactly the same as in §EC.4.2. We omit the details. \hfill \square

With the above lemma, we can prove that $(\tilde{Q}_w$ and $\tilde{\omega}$ were defined in §5.3)

$$\sum_{j \in J} m_{\bar{r}} \tilde{Q}_j(\cdot) \Rightarrow \min \left( \tilde{Q}_w(\cdot), \tilde{\omega} \right), \text{ as } r \to \infty.$$  

The rest of the asymptotic optimality follows the steps in §EC.4.3.

**EC.12. Proof for Lemma 6.1**

**Proof**: We follow the framework in §EC.4, but now for IP patients. We first prove that under the proposed policies with selection rule (18), any limit of the hydrodynamic scaled processes $\bar{E}_j, \bar{S}_j, \bar{\tau}_j, \bar{T}_j, \bar{Q}_j, \bar{D}_j, j \in J, \bar{E}_k, \bar{S}_k, \bar{T}_k, \bar{Q}_k, \bar{D}_k, k \in K$ should satisfy (EC.28)-(EC.39), as well as the following:

$$\int_0^\infty \left( \max_{k \in K} H_k \cdot (GC'(\bar{Q}(t)))_k - H_k \cdot (GC'(\bar{Q}(t)))_k \right) d\bar{T}_k(t) = 0. \tag{EC.64}$$

The proof is by contradiction. Assume the above is not true. Then there is $t_0$ and $\delta > 0$ such that $\max_{k' \in K} H_{k'} \cdot (GC'(\bar{Q}(t_0)))_{k'} > H_k \cdot (GC'(\bar{Q}(t_0)))_k$ and $\bar{T}_k(t_0 + \delta) - \bar{T}_k(t_0 - \delta) > 0$. We can also assume that this $\delta$ is chosen so that $\max_{k' \in K} H_{k'} \cdot (GC'(\bar{Q}(t)))_{k'} > H_k \cdot (GC'(\bar{Q}(t)))_k$, for all $t \in [t_0 - \delta, t_0 + \delta]$. Then for $n$ large enough, $\max_{k' \in K} H_{k'} \cdot (GC'(\bar{Q}^n(t)))_{k'} > H_k \cdot (GC'(\bar{Q}^n(t)))_k$, for all $t \in [t_0 - \delta, t_0 + \delta]$, and $\bar{T}_k^n(t_0 + \delta) - \bar{T}_k^n(t_0 - \delta) > 0$. However, this contradicts the selection rule (18) for IP patients. As a result, (EC.64) should be true.

Without loss of generality (from Propositions 8 and 9), we assume that for all $t \geq 0$,

$$\sum_{k \in K} m_{\bar{r}} \bar{Q}_k(t) = \left( \sum_{j \in J} m_{\bar{r}} \bar{Q}_j(0) + \sum_{k \in K} m_{\bar{r}} \bar{Q}_k(0) - \bar{\omega} \right).$$

For any fixed $k \in K$, define a $K \times K$ matrix $B_k = \Upsilon + \Theta_k$, where $\Upsilon$ is a $K \times K$ diagonal matrix with component $-H_i$ in the $i$th place, and $\Theta_k$ is a $K \times K$ matrix with its $k$th column being $H_k$ while all others are 0; that is,

$$\Upsilon = \begin{pmatrix} -H_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & -H_i & \cdots & 0 \\ \vdots & \cdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & -H_K \end{pmatrix} \quad \text{and} \quad \Theta_k = \begin{pmatrix} 0 \cdots H_k \cdots 0 \\ \vdots \vdots \vdots \vdots \vdots \\ 0 \cdots H_k \cdots 0 \\ \vdots \vdots \vdots \vdots \vdots \\ 0 \cdots H_k \cdots 0 \end{pmatrix}.$$
One can verify that the vector $M^c$ is the only column vector (up to scaling) that satisfies

$$B_kGM^c = 0.$$  

Recall the definition of $\Delta_k$ in Lemma 5.1 and define

$$Q_0 = \Delta_K \left( \left( \sum_{j \in J} m_j^c Q_j(0) + \sum_{k \in K} m_k^c Q_k(0) - \tilde{\omega} \right)^+ \right).$$

Then

$$B_kGC'(Q_0) = 0.$$  

Here $C'(Q_0)$ is a $K$-dimensional vector with $C_k'(Q_{0k})$ being its $k$th component. This means that, for any $k, l \in K$,

$$H_k \cdot (GC'(Q_0))_k = H_l \cdot (GC'(Q_0))_l.$$  

Next we prove that, for any fixed $C > 0$ and a hydrodynamic model solution with $\sum_{j \in J} m_j^c Q_j(0) + \sum_{k \in K} m_k^c Q_k(0) < C$, there exists a constant $\tilde{T}_0$ such that, for all $t \geq \tilde{T}_0$,

$$Q_K(t) = \Delta_K \left( \left( \sum_{j \in J} m_j^c Q_j(t) + \sum_{k \in K} m_k^c Q_k(t) - \tilde{\omega} \right)^+ \right).$$

We first prove that, if $\dot{Q}(t) \neq Q_0$, then there is a $k_+$ such that

$$H_{k_+} \cdot (G \cdot (C'(\dot{Q}(t)) - C'(\bar{Q}_0)))_{k_+} > 0.$$  

It is enough to prove $(G \cdot (C'(\dot{Q}(t)) - C'(\bar{Q}_0)))_{k_+} > 0$. This holds because $G^{-1} G \cdot (C'(\dot{Q}(t)) - C'(\bar{Q}_0)) = (C'(\dot{Q}(t)) - C'(\bar{Q}_0))$, and $(C'(\dot{Q}(t)) - C'(\bar{Q}_0))$ is a vector with positive component(s), together with the assumption that all components of $G^{-1}$ are nonnegative, which implies that there must be at least one positive term in $G \cdot (C'(\dot{Q}(t)) - C'(\bar{Q}_0))$.

Choose any $k_- \in \arg \min_{k \in K} \left\{ \frac{C_k'(Q_0(t))}{m_k^c} - \frac{C_k'(Q_{0k})}{m_k^c} \right\}$; then $C_{k_-}'(\bar{Q}_{k_-}(t)) - C_{k_-}'(\bar{Q}_{0k_-}) \leq 0$. Now we prove that

$$H_{k_-} \cdot (G \cdot (C'(\dot{Q}(t)) - C'(\bar{Q}_0)))_{k_-} \leq 0.$$  

It is enough to prove $(G \cdot (C'(\dot{Q}(t)) - C'(\bar{Q}_0)))_{k_-} \leq 0$. As $G_{k_- k'} \leq 0$, for all $k' \neq k_-$ and $k_- \in \arg \min_{k \in K} \left\{ \frac{C_k'(Q_0(t))}{m_k^c} - \frac{C_k'(Q_{0k})}{m_k^c} \right\}$, we have

$$(G \cdot (C'(\dot{Q}(t)) - C'(\bar{Q}_0)))_{k_-} \leq \sum_{k' \in K} G_{k_- k'} \cdot \left( C_{k_-}'(\bar{Q}_{k_-}(t)) - C_{k_-}'(\bar{Q}_{0k_-}) \right).$$

From the assumption of $\sum_{k' \in K} G_{k_- k'} \geq 0$, and the fact that $C_{k_-}'(\bar{Q}_{k_-}(t)) - C_{k_-}'(\bar{Q}_{0k_-}) \leq 0$, we get $(G \cdot (C'(\dot{Q}(t)) - C'(\bar{Q}_0)))_{k_-} \leq 0$.

Now $H_{k_+} \cdot (GC'(Q_0))_{k_-} = H_{k_+} \cdot (GC'(\bar{Q}_0))_{k_-}$, thus $H_{k_-} \cdot (GC'(\dot{Q}(t)))_{k_-} < H_{k_+} \cdot (GC'(\dot{Q}(t)))_{k_+}$. From (EC64) we then have $T_{k_-}(t) = 0$, which implies $\dot{Q}_{k_-}(t) = \lambda_k > 0$.  

As there is a finite number of IP classes, there must be a finite time $\bar{\tau}$ such that, for all $t \geq \bar{\tau}$, $\sum_{k \in \mathcal{K}} m_k \tilde{Q}_k(t) = \sum_{k \in \mathcal{K}} m_k Q_0$, hence $\tilde{Q}_k(t) = Q_0$, for all $k \in \mathcal{K}$ and $t \geq \bar{\tau}$.

One can now follow the argument in Bramson (1998), first proving “multiplicative state-space collapse” (Bramson (1998)’s equation (3.41)):

$$\sup_{0 \leq t \leq T} \left| \tilde{Q}_k(t) - \Delta_k \min \left( \tilde{Q}_w(t), \tilde{\omega} \right) \right| \Rightarrow 0,$$

Then from our Proposition 1, we get

$$\sup_{0 \leq t \leq T} \left| \tilde{Q}_k(t) - \Delta_k \min \left( \tilde{Q}_w(t), \tilde{\omega} \right) \right| \Rightarrow 0,$$

which completes the proof of Lemma 6.1.

**EC.13. 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 $\{\tilde{Q}_w\}$ is stochastically bounded, in particular the diffusion-scaled queue length processes of IP patients are stochastically bounded. Then (EC.47) and (EC.48) 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 $0 \leq a < b \leq T$,

$$\frac{1}{E_k(b) - E_k(a)} \left( \int_a^b \tau_k d\tilde{E}_k - \int_a^b \tilde{Q}_k(s) ds \right) \Rightarrow 0.$$

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

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

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

$$\begin{align*}
\min_x & \sum_{k \in \mathcal{K}_0} \lambda_k \tilde{C}_k \left( \sum_{j \in \mathcal{K}_k} x_j / \lambda_k \right) \\
\text{s.t.} & \sum_{k \in \mathcal{K}_0} \sum_{k' \in \mathcal{K}_k} m_{kk'} x_{kk'} = a, \\
& x \geq 0.
\end{align*}$$

(EC.65)
One can prove that the proposed family of control policies \( \{ \tilde{\pi}^*_k \} \) attains the lower bound by showing the corresponding state-space collapse. Here we give some structural insights on the optimal solution to the problem (EC.65). For classes in \( C_k \), we know that, if \( \sum_{k' \in C_k} m^r_{k'} x_{k'} \) is fixed, then the solution minimizing \( C_k(\sum_{j \in C_k} x_j / \lambda_k) \) has necessarily \( x_k \) non-zero, while all other \( x_j \) with \( j \in C_k \setminus \{ k \} \) are 0 (this is because \( m^r_j > m^c_j \), for all \( j \in C_k \setminus \{ k \} \)). As a result, if the problem has an optimal solution with some \( k' \in C_k \setminus \{ k \} \), for some \( k \), then one can always find a better solution, which is a contradiction. The problem has been thus reduced to the following problem:

\[
\begin{align*}
\min_x & \quad \sum_{k \in C_0} \lambda_k C_k(x_k / \lambda_k) \\
\text{s.t.} & \quad \sum_{k \in C_0} m^r_k x_k = a, \\
& \quad x \geq 0,
\end{align*}
\]

(EC.66)

Following the solution of (10) (using the KKT conditions), we can define a new function, in analogy to \( \tilde{\Delta}_k(\cdot) \) from (EC.61) (but now with subscript \( C_0 \)), and under \( \{ \tilde{\pi}^*_k \} \), this function plays the role of a lifting mapping in the corresponding state-space collapse.

**EC.14. 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 this section.

**The ED system with delays between physician visits:** Let \( Q^r_{jk}(t) \) denote the number of patients in the delayed system between \( j \)-triage and \( k \)-IP patients at time \( t \), and \( Q^r_{kl}(t) \) 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

\[
Q^r_k(t) = Q^r_k(0) + \sum_{j \in \mathcal{J}} \left( \Phi^r_{jk}(S_j(T^r_j(t))) + Q^r_{jk}(0) - Q^r_{jk}(t) \right)
+ \sum_{l \in \mathcal{K}} \left( \Phi^r_{lk}(S_l(T^r_l(t))) + Q^r_{lk}(0) - Q^r_{lk}(t) \right) - S_k(T^r_k(t))
\]

\[= Q^r_k(0) + \sum_{j \in \mathcal{J}} \Phi^r_{jk}(S_j(T^r_j(t))) + \sum_{l \in \mathcal{K}} \Phi^r_{lk}(S_l(T^r_l(t))) - S_k(T^r_k(t))\]

\[- \sum_{j \in \mathcal{J}} (Q^r_{jk}(t) - Q^r_{jk}(0)) - \sum_{l \in \mathcal{K}} (Q^r_{lk}(t) - Q^r_{lk}(0)), \quad k \in \mathcal{K}.
\]

(EC.67)

Ignoring the changes of \( T^r_j, j \in \mathcal{J} \) and \( T^r_k, k \in \mathcal{K} \), the difference between (EC.67) and (EC.11) is \( \sum_{j \in \mathcal{J}} (Q^r_{jk}(t) - Q^r_{jk}(0)) + \sum_{l \in \mathcal{K}} (Q^r_{lk}(t) - Q^r_{lk}(0)) \), 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 \( T^r_j, j \in \mathcal{J} \) and \( T^r_k, k \in \mathcal{K} \) is 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 \( Q^r_k(0) + \sum_{j \in \mathcal{J}} \Phi^r_{jk}(S_j(T^r_j(t))) + \)
\[ \sum_{l \in K} \Phi_{ik}(S_l(T_r^i(t))) - S_k(T_r^i(t)) \] is the same as in the system without delays. It is enough to prove that the fluid limit of \[ \sum_{j \in J} (Q_{jk}(t) - Q_{jk}(0)) + \sum_{l \in K} (Q_{lk}(t) - Q_{lk}(0)) \] is 0. Indeed, if we have the latter fact, we can first argue that the fluid limit of \[ \sum_{j \in J} m_j \hat{Q}_j(t) + \sum_{k \in K} m_k \hat{Q}_k(t) \] 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 \( j \in J \) and \( \lambda_k m_k t \) for \( k \in K \).

We now prove that the fluid limit of \[ \sum_{j \in J} (Q_{jk}(t) - Q_{jk}(0)) + \sum_{l \in K} (Q_{lk}(t) - Q_{lk}(0)) \] 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/∞ with fast service rates (at the end of this subsection) 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 \( T_r^j, j \in J \) and \( T_r^k, k \in K \) remain constant.

Next we discuss the diffusion-scaled processes. From the differences between (EC.67) and (EC.11), to prove that \[ \sum_{j \in J} m_j \hat{Q}_j(t) + \sum_{k \in K} m_k \hat{Q}_k(t) \] is invariant to all work-conserving policies, it is enough to argue that the following prevails for each \( k \in K \):

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

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.

**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 \( x^n \in D \), there are \( y^n \in D \) satisfying the following equation:

\[ y^n(t) = x^n(t) - \mu^n \int_0^t y^n(s) ds. \]  \hspace{1cm} \text{(EC.68)}

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 get some insights from infinite-server queues.

Consider a sequence of infinite-server queueing systems G/M/∞. In the \( r \)th 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 \( \bar{E}^r \) are tight. Here

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

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

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

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

Using (EC.68), if $\hat{X}^r(0) \Rightarrow 0$, then one can show that $\hat{X}^r(\sigma^r \wedge \cdot) \Rightarrow 0$. Following the discussion of proving (39) in Atar and Solomon (2011), we can also prove $\sigma^r \Rightarrow T$. As a result, $\hat{X}^r \Rightarrow 0$ on $[0,T]$. As this $T$ is arbitrary, we have $\hat{X}^r \Rightarrow 0$ on $[0,\infty)$.

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

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

We then have

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

Suppose that there is a sequence of $\{\lambda^r\}$ with (i) $\lambda^r \to \lambda$, for some $\lambda > 0$, (ii) $\hat{X}^r(0) \Rightarrow 0$, and (iii) making $\{\hat{E}^r\}$ tight. Then, from the fluid limit argument, we can prove that $\hat{S}^r \left( \mu r^{2+\alpha} \int_0^t \hat{X}^r(s) ds \right)$ converge to a driftless Brownian motion with variance $\lambda$; using (EC.68), we can now deduce that $\hat{X}^r(\cdot) \Rightarrow 0$.

**EC.15. An ED case study: the value of information & imputed costs**

Most ED triage systems are based on 5 severity levels (Farrohknia et al. (2011), Mace and Mayer (2008)). This granularity is typically too lean to account for patient characteristics that are relevant for decision making - clinical and operational. For example, our ED-Partner (Carmeli (2012)), which uses the Canadian Triage and Acuity Scale (CTAS), attempts to also take into account age and predicted A&D status (will the patient be Admitted, Discharged or transferred to another facility); other EDs, for example those implementing the U.S. Emergency Severity Index (ESI), consider the number of ED resources used by the patient, a proxy for which could be the number of visits to an ED physician that a patient experiences. Note that A&D status and the number of IP phases are unknown at the triage state, but our hospital partners tell us that experienced ED physicians or nurses can predict them accurately; see Barak-Corren et al. (2013), Saghafian et al. (2011, 2012). In this subsection, based on data from our ED-Partner, we use our models to assess the operational benefits of such predictions.
For simplicity and insight, we analyze only the IP part of the ED patient flow, and we focus on A&D status and the number of IP visits to an ED physician (which we refer to as IP phases: each such phase will be regarded as a separate class in our formal model.). Patients among each class are served on a FCFS basis. In ED-Partner, patients experience 1-5 IP phases: 28% go through 1 phase only, 30% have 2 phases, 28% - 3 phases, 11% - 4 phases, and 3% go through 5 IP phases. The fractions of patients who are Discharged is close to 60%; the others are admitted or transferred elsewhere - both referred to as Admitted. We assume that A&D status and the number of IP phases are independent; hence, for example, the fraction of patients who will be admitted after 3 IP phases is 40% × 28% = 11.2%. Expert-solicitation in Carmeli (2012) revealed that sojourn time costs can be assumed quadratic. Specifically, the cost function for admitted patients is $c_a(t) = Ct^2$ for some constant $C$; the specific value of $C$ turns out unimportant for the comparisons that we shall perform - we thus assume $C = 1$. For discharged patients, the cost is twice that of the admitted ones, hence it is $c_d(t) = 2t^2$. Assume that the external arrival rate is 1, and the mean service time for IP patients is equal across all phases (this is not unreasonable from our experience); we denote this common value by $m$, which is determined so that the ED operates in heavy traffic (traffic intensity $\rho \approx 1$).

We now compare three scenarios: no-information, where the ED controller is aware of neither A&D status nor the number of IP phases; partial-information, where only the number of IP phases is known, which will be shown to lead to a reduction of 18% in congestion costs; and full-information, where both are known, which results in about 27% reduction relative to the no-information cost.

**No information:** Each patient goes (stochastically) through 1 to 5 phases; e.g. the probability of continuing to phase 3 after a 2nd physician visit is $P_{23} = (1 - 0.28 - 0.3)/(1 - 0.28) \approx 0.583$. The individual sojourn cost function is
\[
c(t) = 0.4c_a(t) + 0.6c_d(t) = 1.6t^2.
\] (EC.69)

In §EC.15.1, we analyze a system with only two phases. From the analysis there, with the above cost functions and means of service times, an asymptotically optimal policy is to give priority to the second phase. This argument can be generalized to multi-phases: for example, in our 5-phase problem, we first consider the last two phases. It can be argued, similarly to §EC.15.1, that an asymptotically optimal policy assigns priority to the last phase. Then the 2-phase system is reduced to a system with only one phase and, in turn, our 5-phase to a 4-phase system. Continuing this way, an asymptotically optimal policy assigns priority to phases $2 - 5$ over phase 1, and only the queue length of the latter remains non-negligible asymptotically. From the argument in the Appendix, the minimal queueing costs, corresponding to the above policy, accrues approximately at rate $1.6 \left( \frac{\hat{Q}_w - \omega}{m_1} \right)^2 = 1.6 \times 0.1874 \left( \frac{\hat{Q}_w - \omega}{m_2} \right)^2 = 0.2998 \left( \frac{\hat{Q}_w - \omega}{m_2} \right)^2$. Here $m_1 = (28\% + 30\% \times 2 + 28\% \times 3 + 11\% \times 4 + 3\% \times 5) \times m = 2.31m$, hence $\frac{1}{(m_1)^2} \approx \frac{1}{3.31} \times \frac{1}{m} = 0.1874 \frac{1}{m^2}$. As a reminder, here $\hat{Q}_w$ is a reflected Brownian motion, $\hat{\omega}$ is a weighted summation of the triage deadlines, and both can be calculated via the formulae in §5.3.
Partial information: Now assume that the ED controller knows, for individual patients, their number of IP phases (1-5). Then the cost function is still as in (EC.69). The patients are initially classified into 5 IP classes; e.g. Class 3 returns 3 times to the physician, giving rise to 2 additional classes along the way and ultimately being either admitted or discharged. (There is a total of 15 classes.) From our sojourn time analysis in the previous section, an asymptotically optimal policy assigns priority to all non-starting IP classes, while allocating the remaining service capacity to the 5 starting phases as follows: serve a class with index 
\[
k \in \max_{i \in K} \frac{Q_i(t)}{l \times p_i}.
\]

(EC.70)

Here \(Q_i\) is the queue length of class \(l\) IP patients, and \(p_i\) is the fraction of patients that visit the physician \(l\) times, \(l = 1, \ldots, 5\). From the argument in the Appendix (especially (EC.66) and the paragraph above it), the minimal cost rate will be the value of the following problem:
\[
\begin{align*}
\min & \quad 0.28c(\frac{Q_1}{0.28}) + 0.30c(\frac{Q_2}{0.30}) + 0.28c(\frac{Q_3}{0.28}) + 0.11c(\frac{Q_4}{0.11}) + 0.03c(\frac{Q_5}{0.03}) \\
\text{s.t.} & \quad m(Q_1 + 2Q_2 + 3Q_3 + 4Q_4 + 5Q_5) = (Q_w - \tilde{w})^+.
\end{align*}
\]

with \(Q_i\) being the queue length of starting class \(i\) (\(i\) phases). Then the optimal solution satisfies \(Q_5^* = 0.15\frac{Q_1}{0.28}, Q_4^* = 0.44\frac{Q_1}{0.28}, Q_3^* = 0.84\frac{Q_1}{0.28}, Q_2^* = 0.9\frac{Q_1}{0.28}, \) with \(Q_1^* = \frac{m(0.28 + 1.2 + 2.52 + 1.76 + 0.75)}{m^2} (Q_w - \tilde{w})^+\). Simple algebra leads to the asymptotically minimal cost rate of
\[
(0.28 + 0.3 \times 4 + 0.28 \times 9 + 0.11 \times 16 + 0.03 \times 25) \times 1.6 \times \left(\frac{Q_1}{0.28}\right)^2
\]
\[
= \frac{1.6 \times (Q_w - \tilde{w})^2}{m^2(0.28 + 1.2 + 2.52 + 1.76 + 0.75)} = 1.6 \times 0.1536 (Q_w - \tilde{w})^2
\]
\[
= \frac{0.2458 (Q_w - \tilde{w})^2}{m^2}.
\]

Calculating \(\frac{0.2998 - 0.2458}{0.2998} = 0.1801\), it follows that having the information on the number of IP visits will reduce 18.01% of the no-information cost. This is consistent with Saghafian et al. (2011), in which this number of visits (complexity) is identified as an important factor for improving ED operations.

Complete information: Now assume, at the controller’s disposal, an accurate prediction of both the number of IP phases and the A&D status. By the assumed independence of these two pieces of information, one can first analyze the unilateral impact of A&D status, then multiply the two impacts together. For completeness, we present an analysis that accounts jointly for both factors.

Denote by \(Q_{ai}\) and \(Q_{di}\) the queue length of \(i\)-phase patients who will be admitted and discharged, respectively. From our analysis in the Appendix (especially (EC.66) and the paragraph above it), and now having 10 initial classes (the rest, due to their high-priority, enjoy negligible queueing), the minimal cost rate is approximately the optimal value of the following optimization problem:
\[
\begin{align*}
\min & \quad \frac{1}{0.6} \left(0.28c_a(\frac{Q_{a1}}{0.28}) + 0.30c_a(\frac{Q_{a2}}{0.30}) + 0.28c_a(\frac{Q_{a3}}{0.28}) + 0.11c_a(\frac{Q_{a4}}{0.11}) + 0.03c_a(\frac{Q_{a5}}{0.03})\right) \\
& \quad + \frac{1}{0.4} \left(0.28c_d(\frac{Q_{d1}}{0.28}) + 0.30c_d(\frac{Q_{d2}}{0.30}) + 0.28c_d(\frac{Q_{d3}}{0.28}) + 0.11c_d(\frac{Q_{d4}}{0.11}) + 0.03c_d(\frac{Q_{d5}}{0.03})\right) \\
\text{s.t.} & \quad m(Q_{a1} + 2Q_{a2} + 3Q_{a3} + 4Q_{a4} + 5Q_{a5} + Q_{d1} + 2Q_{d2} + 3Q_{d3} + 4Q_{d4} + 5Q_{d5}) = (Q_w - \tilde{w})^+.
\end{align*}
\]
(In the above, we use the fact that $c_a$ and $c_d$ are quadratic functions, and $b(\frac{x}{2})^2 = \frac{1}{8}x^2$.) Similarly to the partial information case, our problem can be further reduced to the following:

$$\min \left( 0.28 + 0.3 \times 4 + 0.28 \times 9 + 0.11 \times 16 + 0.03 \times 25 \right) \times \left( \frac{2}{0.6} \times \left( \frac{Q_{a1}}{0.28} \right)^2 + \frac{1}{0.4} \times \left( \frac{Q_{d1}}{0.28} \right)^2 \right)$$

s.t. \[ \frac{Q_{a1} + Q_{d1}}{0.28} = \frac{(\hat{Q}_w - \hat{\omega})^+}{m(0.28 + 1.2 + 2.52 + 1.76 + 0.75)}. \]

The optimal value, namely the minimal cost rate, is $10.7 \times 0.1536 \frac{(\hat{Q}_w - \hat{\omega})^2}{m^2} = 0.2194 \frac{(\hat{Q}_w - \hat{\omega})^2}{m^2}$. As $0.2458 - 0.2194 = 0.1074$ and $0.2998 - 0.2194 = 0.2682$, we conclude that the information of A&D status unilaterally reduces 10.7% cost; this is consistent with Saghafian et al. (2012), who showed that A&D status contributes to improving ED operations. Furthermore, having jointly the A&D status and the number of IP phases reduces congestion costs by 26.8%.

**EC.15.1. Incomplete information**

We consider a two phase problem, in which we assume that each patient in the ED will need at most two phases of treatment. After the first phase, some of patients will leave the ED directly, while others will go to phase 2. Assume that the mean service times at both phases are 1, and the fraction of patients continuing to the second phase is $p$. Note that this system is different from the system in the sojourn time cost section (§6.4): when a patient arrives now at the ED, we cannot determine whether this patient will visit a physician once or twice, hence we cannot determine the class. After the patient receives one service, we will have our information updated.

The physician in the ED does not have the complete information. That is, when a new patient arrives at the ED, the physician does not know how many phases will this patient go through in the ED. While arriving at the second phase, the physician naturally knows that this is the second visit. Assume that the cost function of a patient is $ax^2$ if his/her sojourn time is $x$. (As $a$ is not important in the following analysis, we fix it to be 1.)

The physician seeks a routing policy which asymptotically minimizes the following cost:

$$\tilde{S}^*(t) = \int_0^t (\tilde{\tau}_{11}(s))^2 d\tilde{E}^*_1(s) + \int_0^t (\tilde{\tau}_{12}(s) + \tilde{\tau}_2(s))^2 d\tilde{E}^*_2(s), \quad (EC.71)$$

in which $\tilde{\tau}_{11}(s)$ represents the waiting time of a patient arriving at time epoch $s$ and will go through only phase 1, $\tilde{\tau}_{12}(s)$ represents the waiting time in phase 1 of a patient arriving at time epoch $s$ and going through both phases, and $\tilde{\tau}_2(s)$ represents the waiting time in phase 2 of that patient; $E^*_1$ is the arrival process for patients with 1 visit only, and $E^*_2$ is the arrival process for patients with 2 phases. Denote by $\lambda$ the external arrival rate to this 2-phase system.

Following the discussion in the previous section, one can prove that

$$\lim_{t \to \infty} \tilde{S}^*(t) \geq \frac{1}{\lambda} \left[ (1 - p) \int_0^t \left( \tilde{\Delta}_1 \left( \tilde{Q}_w(s) - \tilde{\omega} \right)^+ \right)^2 ds + p \int_0^t \left( \tilde{\Delta}_1 \left( \tilde{Q}_w(s) - \tilde{\omega} \right)^+ + \tilde{\Delta}_2 \left( \tilde{Q}_w(s) - \tilde{\omega} \right)^+/p \right)^2 ds \right], \quad (EC.72)$$
where \((\tilde{\Delta}_1(a), \tilde{\Delta}_2(a))\) is the solution to the following optimization problem:

\[
\begin{align*}
\min_x & \quad (1 - p)x_1^2 + p(x_1 + x_2/p)^2 \\
\text{s.t.} & \quad (1 + p)x_1 + x_2 = a, \\
& \quad x_1, x_2 \geq 0.
\end{align*}
\]

It is easy to see that the optimal solution to this problem is \(x_1 = \frac{a}{1 + p}\) and \(x_2 = 0\). As a result, in this two phase problem, an asymptotically optimal policy is to give priority to the second phase.

Here we briefly discuss how to establish (EC.72) and (EC.73), starting with an intuitive argument. As patients in the first stage are served on a FCFS basis, \(\tau_1^{-}(s) \approx \tilde{\tau}_1^-(s)\), and similarly to Lemma EC.1.1, they are close to \(\frac{\tilde{Q}_1^-(s)}{\lambda}\). Here \(\tilde{Q}_1^-(s)\) is the queue length of the patients in the first phase at time \(s\). Similarly, \(\tau_2^-(s) \approx \tilde{\tau}_2^-(s)\), noticing that \(\lambda p\) is the arrival rate to the second phase. On the other hand, the arrival rate for patients visiting phase 1 only is \(\lambda(1 - p)\), and the arrival rate for patients visiting both phases is \(\lambda p\); thus the cost can be approximated by

\[
\frac{1}{\lambda} \left[ (1 - p) \int_0^s (\tilde{Q}_1^-(s))^2 ds + \int_0^s (\tilde{Q}_1^-(s) + \tilde{Q}_2^-(s))/p \right] ds.
\]

The effective mean service time at phase 1 is \(1 + p\) and the effective mean service time at phase 2 is 1. We thus deduce that \((1 + p)\tilde{Q}_1^-(s) + \tilde{Q}_2^-(s) \approx \left( \tilde{Q}^r(s) - \tilde{\omega} \right)\), which supports (EC.72) and (EC.73).

Now we briefly mention what are the necessary steps to prove the above results (However, we omit the details.). Following the discussion in van Mieghem (1995), one can prove that (this is like a sample path Little’s Law)

\[
\frac{1}{E_1^r(a) + E_2^r(a) - E_1^r(a) - E_2^r(a)} \left( \int_a^b \tilde{\tau}_1^r(s) d\tilde{E}_1^r(s) + \int_a^b \tilde{\tau}_2^r(s) d\tilde{E}_2^r(s) - \int_a^b \tilde{Q}_1^r(s) ds \right) \Rightarrow 0.
\]

As the service discipline in the first phase is FCFS, one can prove that

\[
\frac{1}{E_1^r(a) - E_1^r(a)} \left( \int_a^b \tilde{\tau}_1^r(s) d\tilde{E}_1^r(s) \right) = \frac{1}{E_2^r(b) - E_2^r(a)} \left( \int_a^b \tilde{\tau}_2^r(s) d\tilde{E}_2^r(s) \right).
\]

By using (EC.75), together with (EC.74), then following the discussion in van Mieghem (1995), we deduce (EC.72).

**EC.15.2. Imputed cost**

Our ED case study was based on expert estimates of costs in an Israeli hospital. Generally, such cost parameters are unavailable or, at the least, difficult to assess. This raises a natural question: assume that an ED, after accumulating ample experience, operates close to optimally; can one then infer the relative costs associated with patient classes? The answer will shed light on the implicit understanding of these costs by ED physicians. As an example, assume that patients are classified into two classes: admitted and discharged, with the same means of service
times; assume further that sojourn time costs are quadratic, but the parameters are unknown. Our results suggest that, if the proportion of the queue lengths of the admitted class to the discharged class are roughly a constant (state-space collapse), then the inverse of this constant is an estimator of the ratio of the cost parameters. This is because, under our assumptions on mean service times, we expect that

\[ c_a Q_a(t) \approx c_d Q_d(t) \]

from our state-space collapse results; here \( c_a, c_d \) are the cost parameters of patients admitted and discharged, respectively, and \( Q_a, Q_d \) are the corresponding rate. Then one has, as discussed above,

\[ \frac{c_a}{c_d} \approx \frac{Q_d(t)}{Q_a(t)}. \]

**EC.16. More simulation outputs**

We provide here more simulation results that complement §7.

**EC.16.1. 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 (17), 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 (17), while among all IP classes, the priority is according to the modified generalized \( c\mu \) rule.

**EC.16.2. How many triage patients violate the deadlines**

As mentioned in §7, the fractions of patients who violate the deadlines are less than 5%, and the fractions who violate the deadlines by more than 10% of their corresponding deadlines are negligible (less than 1%). Here we visualize these statements though histograms and summaries of all 160 sample paths.

Figure EC.1 depicts three histograms for the three triage classes: the top one is for 1-triage patients, the middle for 2-triage and the third for 3-triage.

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 more serious. 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 receive high priority because their \( d_1 - \tau_1(t) \) is conceivably the shortest (recall that \( d_1 = 30, d_2 = 60 \) and \( d_3 = 120 \)).
Figure EC.1  Histograms of patient waiting times under our proposed policy

EC.16.3. A comparison between (11) and (17)

In Theorem 1 we use (11) to determine the triage class to be served next, while we recommend the Shortest-Deadline-First rule (17) in practice. We simulated the systems under these two policies and compared them. Denote by “Ratio” the policy using (11), while “SDF” is the policy using Shortest-Deadline-First rule (17). Other parts of the policies remain the same (see §7.2).

Again, we use $P_j$, $j = 1, 2, 3$, for the fraction of $j$-triage patients who violate their corresponding deadline. The following table summarizes the performances of the two policies:

<table>
<thead>
<tr>
<th>Policy</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td>7.62% (0.16%)</td>
<td>5.23% (0.11%)</td>
<td>3.64% (0.08%)</td>
<td>127.64 (11.11)</td>
</tr>
<tr>
<td>SDF</td>
<td>4.61% (0.10%)</td>
<td>4.57% (0.09%)</td>
<td>4.57% (0.09%)</td>
<td>125.21 (10.36)</td>
</tr>
</tbody>
</table>

From the table, we observe that both policies perform well, but the “SDF” policy performs slightly better than the “Ratio” policy.

In Figure EC.2, we visualize the fractions of patients violating the deadlines under the “Ratio” policy through histograms. The histograms differ from those in Figure EC.1, which is due to the difference between $\tilde{\Delta}_j$ and $\Delta_j$.

Figure EC.2  Histograms of patient waiting times under the “Ratio” policy with rule (11)
EC.16.4. On the duration of delays between physician visits

We now incorporate in the simulations delays between successive visits to physicians. We consider the following delays (all in minutes): 0, 1, 10, 60 and 120. (Delays are assumed exponentially distributed). The performance measures are summarized in the following table.

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

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.16.5. 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. These delays model services and waiting times beyond physicians (e.g., X-ray, lab tests).

Daily arrival rates are time-varying, as in Figure 2, such that the average total arrivals of triage classes 3, 4 and 5 per day is $14 \times 24 = 336$. 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. Their duration may depend on the class. Table 2 in Yom-Tov and Mandelbaum (2013) 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 $\epsilon$ of our proposed policy. Specifically, our simulation gives rise to the following rule: give 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 $\epsilon_j$'s. Other parameters, such as distributions and cost rates, are assumed equal to those in the system without delays and with constant arrival rates.

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
make the following comparisons: our proposed policy with rule (17) vs. rule (11) (Table EC.3) and our proposed policy against the 3 alternatives FCFS, IPF and TrF (Table EC.4).

**Figure EC.3** A sample path of the time-varying system under the proposed policy

From Figure EC.3, we see a phenomenon similar to Figure 3. After summarizing all 160 sample paths, the fraction of triage patients who violate their correspond 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 over most times (load is often above 1.2 and can be as high as 1.6663), this suggests that our proposed policy may also work well in overloaded systems.

Here are the histograms of triage waiting times:

**Figure EC.4** Histogram of triage waiting times in the time-varying system under our proposed policy

We now compare between policies using rule (11) and the Shortest-Deadline-First rule (17), in a time-varying environment.

**Table EC.3** Comparison between “Ratio” (11) and “SDF” (17)

<table>
<thead>
<tr>
<th>Policy</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>Cost</th>
<th>LoS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td>12.07%</td>
<td>4.75%</td>
<td>1.95%</td>
<td>1553.82</td>
<td>369.02</td>
</tr>
<tr>
<td>SDF</td>
<td>4.44%</td>
<td>3.21%</td>
<td>2.75%</td>
<td>1561.89</td>
<td>368.64</td>
</tr>
</tbody>
</table>
From Table EC.3, we observe that both policies perform reasonably well, with “SDF” performing better than the “Ratio” policy.

We now compare our policy with the three alternatives, as in the stationary case (§7). We summarize the performances of these four policies in Table EC.4.

<table>
<thead>
<tr>
<th>Policy</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>Cost</th>
<th>LoS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGcµ</td>
<td>4.44% (0.04%)</td>
<td>3.21% (0.02%)</td>
<td>2.75% (0.02%)</td>
<td>1561.89 (40.02)</td>
<td>368.64</td>
</tr>
<tr>
<td>FCFS</td>
<td>76.29% (0.19%)</td>
<td>56.81% (0.28%)</td>
<td>17.63% (0.31%)</td>
<td>1160.82 (15.72)</td>
<td>371.93</td>
</tr>
<tr>
<td>IPF</td>
<td>68.95% (0.21%)</td>
<td>69.82% (0.21%)</td>
<td>74.04% (0.20%)</td>
<td>7.33 (0.01)</td>
<td>305.42</td>
</tr>
<tr>
<td>TrF</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>3251.88 (44.39)</td>
<td>412.86</td>
</tr>
</tbody>
</table>

From Table EC.4, 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.

**EC.16.6. Using queue lengths for determining thresholds**

From our proof, one can also use queue lengths in the threshold part: if there exists a $j \in J$ such that $Q_j(t) \geq \lambda_j(d_j - \epsilon)$, then assign priority to triage patients. However, we do not recommend this rule. There are several reasons: 1) it involves external arrival rates which, furthermore, can be time-varying; 2) when the arrival rate $\lambda_j$ is small, $\lambda_j d_j$ is not large; thus triage patients could be given priority even if uncalled for.

There is an additional reason which we now illustrate via simulation. We use the parameter set in §7, and assign priority to triage patients if there exists a $j$ such that $Q_j \geq \lambda_j(d_j - \epsilon)$, with $\epsilon = 3$. In this case, $\lambda_1 (d_1 - 3)$ is less than 1 ($\lambda_1 = \frac{14}{60} \times 0.1$); thus triage patients will receive priority if a 1-triage patient is present. We plot, in Figure EC.5, a typical sample path of the system under this policy:

**Figure EC.5** A sample path of the stationary system, controlled by a queue-length based policy
We observe that there are many triage patients violating their corresponding deadlines, especially 2-triage and 3-triage patients. A detailed analysis of the simulated data shows that no 1-triage patients violate the deadline 30, 11.009% 2-triage patients violate the deadline 60, and 11.853% 3-triage patients violate the deadline 120.

Another interesting phenomenon is that the queue lengths of triage classes are now bounded hence well controlled (relative to our proposed policy, under which ages are bounded while the queue lengths fluctuate). This then explains the empirical observations: even with bounded queue lengths and the sample-path Little’s Law, the waiting time can still fluctuate (if the arrival rate is large, then deviations can be large), which would lead to waiting times that violate the deadlines. As a result, we recommend to monitor the ages directly. A formal comparison between our proposed policy and the policy using queue lengths would involve rates of convergence, hence we do not pursue it here.

**EC.16.7. Using one triage class in the threshold rule**

In Theorem 1, we prioritize between triage patients and IP patients according to the following rule: if there exists a \( j \in J \) such that \( \tau^r_j(t) \geq d^r_j - \epsilon^r \), then assign priority to triage patients. In fact, we can prove the same asymptotic optimality result under a much weaker condition: fix a triage class, for example triage class 1; then assign priority to triage patients if \( \tau^r_1(t) > d^r_1 - \epsilon^r \). However, this rule is not recommended, in view of our simulation. In Figure EC.6, we plot a typical sample path (in which we control 1-triage patients only, with \( \epsilon = 6 \)).

**Figure EC.6** A sample path of the stationary system, controlled via 1-triage only

The simulation shows that the ages of all three triage class patients have similar shapes, which suggests “state-space-collapse”, that is, with only a single triage class, one can indeed control the ages of the other two classes. However, this control needs not be good enough. A detailed analysis of the simulated data shows that on the above sample path, 13.339% triage patients violate their deadline 30, 21.048% 2-triage patients violate 60, and 20.993% 3-triage violate 120. This demonstrates that, even though we have state-space collapse, using only triage
class 1 in the threshold fails to control age processes, hence these fluctuate and violate their deadlines. Again, a formal comparison between our proposed policy and the policy that uses only triage class 1 in its threshold, would involve rates of convergence – hence it is beyond the scope of the present paper.

References


