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

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

Junfei Huang
National University of Singapore, junfeih@gmail.com

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: per individual visit to a server or cumulative over all visits. The former is our base-model, which paves the way for the latter (more ED-realistic) one. 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 with the largest delay relative to deadline; alternatively, IP patients are served according to some G$c\mu$ policy, in which μ is simply modified to account for feedbacks. For our proposed policies, we establish asymptotic optimality, and develop some congestion laws that support forecasting of waiting and sojourn times. Finally, via data from the complex ED reality, we use our models to quantify the value of refined individual information, for example whether an ED patient will be admitted to the hospital as opposed to being discharged.

Key words: Emergency Department, Triage, ED Congestion, Heavy Traffic, Feedback Queues, 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. (2011), 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 Departments (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 (2011), Table 2).

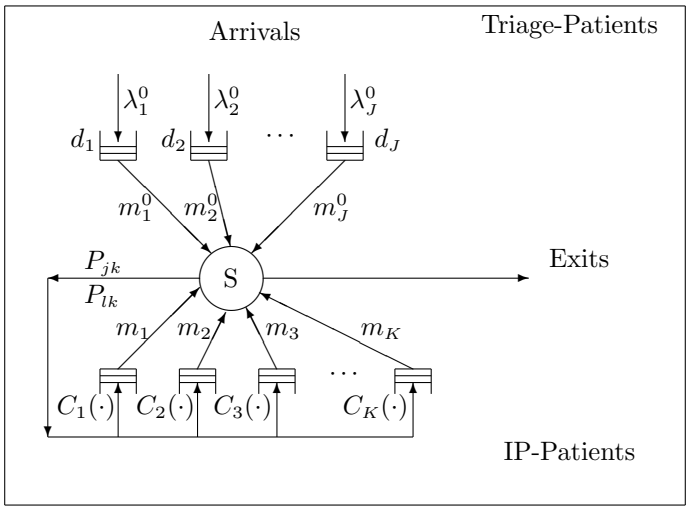
Thus, IP patients impose *operational* congestion (e.g. they occupy beds), which must be controlled while adhering to *clinical* triage constraints (e.g. stabilizing patient conditions). It is this operational-clinical friction that we focus on, from the viewpoint of the ED physician: when becoming idle, 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: the first is a basic ED model where queueing costs are incurred per individual doctor visits; in the second, congestion costs accumulate over all visits during patient sojourn-times. The basic model is introduced in §1.1 and 1.2, and the alternative model, together with a realistic ED example, in §1.3.

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. (2011)). Within this 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 congestions 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 - cost per individual visits

ED dynamics is captured by a multiclass queueing system, with S 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 $j \in \mathcal{J}$ and $l, k \in \mathcal{K}$ the class indices for triage patients and IP patients, respectively.

The j-triage patients arrive to the system exogenously at rate λ_j^0; each such patient requires mean service (triage) time of m_j^0, and 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, $\tau_j(t) \le d_j$, for all $j \in \mathcal{J}$ and $t \ge 0$, 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). Here we only introduce the minimal notation that suffices for describing our flow control problem and its solution.

After completing their first service, j-triage patients join the queue of k-IP patients with probability P_{jk}, $k \in \mathcal{K}$, or exit the system with probability $1 - \sum_{k\in\mathcal{K}} P_{jk}$. Turning attention to IP patients, they originate from either triage patients or from other IP patients: P_{lk} is the probability of switching from IP-class l to k, and $1 - \sum_{k\in\mathcal{K}} P_{lk}$ is the probability that an l-IP patient exits the ED, after service. The k-IP patients have mean service requirements of m_k and, while waiting, they 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_j^e denote the mean *effective service time* of j-triage patients. Vector-wise, $M_{\mathcal{J}}^e = M_{\mathcal{J}} + P_{\mathcal{J}\mathcal{K}}[I - P]^{-1}M$, where $M_{\mathcal{J}}^e = (m_j^e)$, $M_{\mathcal{J}} = (m_j)$, $P_{\mathcal{J}\mathcal{K}} = [P_{jk}]$ is the triage-to-IP transition matrix, $P = [P_{lk}]$ is the IP-to-IP transition matrix, and

$M=(m_k)$. Thus, m_j^e is the expected *total* service time, required by j-triage patients throughout their ED stay. The *traffic intensity* is then

$$\rho=\frac{1}{S}\sum_{j\in\mathcal{J}}\lambda_j^0 m_j^e,$$

and we think in terms of $\rho\approx 1$ (ED in heavy-traffic). Finally, let m_k^e denote the mean *effective service time* of k-IP patients. Vector-wise, $M^e=[I-P]^{-1}M$, where $M^e=(m_k^e)$.

Notation has been now set for describing flow control policies. Choose any *one* of the triage classes (conceivably the least d_j, say d_1). Then a physician that becomes idle at time t adopts the following guidelines:

- Serve triage patients if $\tau_1(t)\geq d_1-\epsilon$, where ϵ is small relative to d_1 (e.g. $d_1=30$ minutes while $\epsilon=3$ minutes);
- Given that a triage patient is to be served, choose the head-of-the-line patient from the class with index

$$j\in\arg\max_{j\in\mathcal{J}}\frac{\tau_j(t)}{d_j};$$

- Given that an IP patient is to be served, choose the head-of-the-line patient from the class with index

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

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 snap-shot 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). For example, our S-server system behaves as a single-server one, in which this virtual server is S-times faster than each of the original servers; accordingly, and without loss of generality by Chen and Shanthikumar (1994), our subsequent analysis will assume $S=1$.

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 a threshold $\omega=\sum_{j\in\mathcal{J}}\lambda_j^0 d_j m_j^e$: if $\sum_{j\in\mathcal{J}} m_j^e Q_j(t)\geq\omega$, a server just becoming idle caters to triage patients, otherwise to IP patients. Furthermore, triage classes can be alternatively prioritized according to shortest-deadline-first, that is, serve $j\in\arg\min_{j\in\mathcal{J}}[d_j-\tau_j(t)]$; and the selection criterion of IP-classes can also be any rule that satisfies (13), in particular the one conjectured in page 853 of Mandelbaum and Stolyar (2004).

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, 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. (The situation could be very different in a significantly time-varying environment, in contrast to our assumed stationarity. An example is a mass-casualty event during which triage patients overload the system; see Section 8 for further discussion.)

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

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

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 $(t-\tau_j(t), t]$, it must be that $Q_j(t) \approx \lambda_j^0 \tau_j(t)$ and the workload equals approximately

$$\sum_{j\in\mathcal{J}} m_j^e \lambda_j^0 \tau_j(t) + \sum_{k\in\mathcal{K}} m_k^e Q_k(t).$$

The invariance of the potential workload now implies that minimizing $\sum_{k\in\mathcal{K}} m_k^e Q_k(t)$ (which is in concert with minimizing IP congestion costs) is equivalent to maximizing $\sum_{j\in\mathcal{J}} m_j^e \lambda_j^0 \tau_j(t)$.

Triage vs. IP: By the deadline constraints, an upper bound for $\sum_{j\in\mathcal{J}} m_j^e \lambda_j^0 \tau_j(t)$ is $\omega = \sum_{j\in\mathcal{J}} \lambda_j^0 d_j m_j^e$, and our policy should thrive to narrow their gap. From the light-traffic view of triage patients, this can be achieved by serving triage patients only as their deadline in getting "dangerously" close - a "threat" that can be monitored through the status of (any) single triage class, as we explain next.

Triage selection: The selection rule among triage classes is designed to ensure that their age processes are so balanced 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. In fact, $\frac{\tau_j(t)}{d_j} \approx \frac{\tau_{j'}(t)}{d_{j'}}$, for any $j, j' \in \mathcal{J}$, at all times t, which implies that the age of any one triage class 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.) Alternative selection rules could also achieve the desired balance, as described in §6.1.

IP selection: After applying the threshold guideline and the triage selection rule, one expects that $\sum_{k\in\mathcal{K}} m_k^e Q_k(t)$ is minimized, thus invariant under any work conserving policy. To minimize cumulative queueing cost, it suffices to minimize cost rates greedily at each time. We are thus led to a convex optimization problem with linear constraints (9). The KKT condition now yields our generalized $c\mu$ rule, as in van Mieghem (1995) but with the μ's replaced by $1/m_k^e$ to account for feedbacks.

The above outline also guides the proofs of our main results, Theorems 2 and 1. These results are consequences of the parsimonious nature of heavy-traffic dynamics, which is also manifested through some congestion laws that will be now described.

A Snapshot principle: This is again a common feature of heavy traffic (Reiman (1982)) which, as explained in page 187 of Whitt (2002) and adopted here, 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 some 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 or ASTA, 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. (The service duration is negligible relative to queueing time.) By the snapshot principle, the queue length Q_k and the virtual waiting time ω_k are then related via $Q_k(t) \approx \lambda_k \omega_k(t)$, with λ_k being the arrival rate to class k. On the other hand, $Q_k(t) \approx \lambda_k \tau_k(t)$, as those patients in the queue at time t arrive 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 of all virtual waiting times at time t over the patient's route. 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 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 be taken as a forecast for the ED sojourn time, over a pre-specified route of a patient that arrives at time t (§5.5).

1.3. An alternative ED model - cost per 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, 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; one then allocates any remaining service capacity to the *new* IP patients, according to our modified G$c\mu$ rule - see (24). 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, it is necessary to 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; this, in practice, might require predictions of class designations - more on that in the following discussions.

Congestion laws: Similarly to our cost-per-visit model, and assuming the above class designation, the snapshot principle also prevails for the IP cost per sojourn time model, under our proposed policy. The snapshot principle then implies the sample-path version of Little's law: 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: 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 (§7). 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, and supported by Saghafian et al. (2011, 2012), we learn that such predictions can be accurately made and, hence, are worth being accounted for.

Beyond our two ED models: Saghafian et al. (2012) remark that, due to the complexity of ED operations, it is challenging to capture prevalent ED features within a single tractable analytic model. While this is precisely what we do here, ours is by no means the final story.

Additional ED features that seek modeling include time-varying arrival rates, treatment times between successive visits to the physician, ambulance diversion (admission control) 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 Farrohknia et al. (2011), Mace and Mayer (2008). Our research focus here is operational (Marmor 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).) After 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 systems, 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 vs. 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 Section 7, where we analyze the value of the information they require. There are additional papers that cater to specific ED characteristics: Yom-Tov and Mandelbaum (2011) 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). A final related references 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 feedback have been analyzed in Reiman (1988), Dai and Kurtz (1995), restricting to a global FCFS service discipline among all classes. Our analysis can be also adapted to the FCFS discipline, as well as to other work-conserving disciplines. Indeed, we provide a detailed analysis of all work-conserving disciplines; then, the additional work required for a specific discipline entails proving state-space collapse, which can follow the framework in Bramson (1998).

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

- **Methodological:** we analysis multiclass queueing systems with feedback, particularly,

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 individual waiting times and cumulative sojourn times;
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 catering to triage- vs. IP-patients;
2. They give rise to scheduling policies that are insightful and implementable;
3. They enable analysis of the value of information in a real ED setup.

Additional references are provided in Section 8, where we propose generalizations and offer conjectures to our main models.

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. Our alternative ED model, with sojourn time costs, is applied in §7, using data from an Israeli ED, and expert-elicited costs. 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.

1.5. Notation

We use the standard notation $\mathbb{R}_+$ to denote the set of nonnegative real numbers. For a real number x, $\lceil x \rceil$ is the maximal integer less than or equal to x; $\mathbb{R}_+^J$ and $\mathbb{R}_+^K$ are the J-times and K-times 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 $x, y \in \mathbb{R}^N$, $x < y$ if and only if $x_i < y_i$, for all $i = 1, 2, \ldots, N$. We also use 0 to denote a column vector with all components being 0, with the dimension being clear from the context. For a matrix M, we use $M_{j\cdot}$ to denote the jth row, and $M_{\cdot k}$ the kth column of M. The function $1(\cdot)$ is the indicator function, the value of which is 1 when the event within $(\cdot)$ prevails, and 0 otherwise.

We assume that all random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Expectation with respect to $\mathbb{P}$ is $\mathbb{E}$. Let $\mathcal{D}[0,\infty)$ be the standard Skorohod space of right-continuous left-limit (RCLL) functions defined on $[0,\infty)$ and equipped with the Skorohod J_1 topology. 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 $\to$ stands for convergence of non-random elements in $\mathcal{D}[0,\infty)$. 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 $\mathcal{J}$ and $\mathcal{K}$ denote the index sets of triage and IP patients, respectively: $j \in \mathcal{J}$ is an index for triage patients, and $l, k \in \mathcal{K}$ are indices for IP patients. It will be convenient to let $\mathcal{J} = \{1, 2, \ldots, J\}$ and $\mathcal{K} = \{1, 2, \ldots, K\}$, while keeping in mind that the indices $1, 2, \ldots$ in $\mathcal{J}$ differ from those in $\mathcal{K}$. To simplify notation, we shall omit the superscript 0 from the arrival rates and mean service times of triage patients: their index $j \in \mathcal{J}$ suffices for their characterization.

2.1. Triage patients

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

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

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

indicate 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 $P_{\mathcal{J}\mathcal{K}} = (P_{jk})_{J\times K}$ be the triage-to-IP matrix. We use $\phi_j(n)$ to denote the indicator function recording to which class the nth j-triage patient will transfer: 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 $\mathbb{P}(\phi_j(n) = 0) = 1 - \sum_{k\in\mathcal{K}} P_{jk}$. We use $\phi_{jk}(n)$ to denote $(\phi_j(n))_k$, the kth element of $\phi_j(n)$, and use

$$\Phi_j(n) := \sum_{i=1}^{n} \phi_j(i),$$

to record the transition of the first n j-triage patients.

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 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 $b_k^2 = \mathrm{var}(v_k(1))$. Among k-IP patients, the service time required for the ith patient receiving service is $m_k v_k(i)$. (Unless specified, we do not require the service discipline among 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

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

represent the number of service completions if the physician has devoted t time units to k-IP patients. Denote $\mu_k = 1/m_k$; then this is the *service rate* for k-IP patients.

After completing service, an l-IP patient will join the queue of k-IP patients, with probability P_{lk}, or exit the system with probability $1 - \sum_{k\in\mathcal{K}} P_{lk}$. Denote the matrix $P = (P_{lk})_{K\times K}$ to be 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 nth served l-IP patient will transfer to; that is, the nth 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 $\mathbb{P}(\phi_l(n) = 0) = 1 - \sum_{k\in\mathcal{K}} P_{lk}$. We use $\phi_{lk}(n)$ to denote $(\phi_l(n))_k$, the kth element of $\phi_l(n)$ and, as before, use

$$\Phi_l(n) := \sum_{i=1}^{n} \phi_l(i),$$

to record the transition of the first n served l-IP patients.

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

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

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

Then

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

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

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

in which m^e_j is called the *effective mean service time* of j-triage patients, and define $M^e = (m^e_k)_{k\in\mathcal{K}}$ to be

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

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

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

The reason we call m^e_j "effective" is because it is the expected total service requirement of a j-triage patient, accumulated up to leaving the system. The reason for m^e_k to be "effective" is similar.

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 $j \in \mathcal{J}$ and $t \geq 0$.

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

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

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

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

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

However, even if one can generalize the problem (6) to a reasonable one, the optimal policy could not be a trivial one: if the physician always gives priority to triage patients, the queue length of the IP patients will get large and the 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 not start their service before their deadlines. Indeed, we propose a threshold policy that determines between triage patients and IP patients and we shall 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 rth system. Then, in the rth system, the arrival rate of j-triage class is λ_j^r and the effective arrival rate for k-IP class is λ_k^r. The deadline for j-triage patients is d_j^r, while the cost function C_k for k-IP patients will be specified in the next section. We assume that the service times and transition vectors are invariant with respect to r, hence there will be no superscript for terms relating to the service times and transition vectors.

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

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

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

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

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

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

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

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

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

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

where $\widehat{d}_j,\ j\in\mathcal{J}$, are strictly positive constants.

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

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

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

4. Asymptotic compliance and optimality

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

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

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

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

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

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

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

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

$$\mathcal{U}^r(t) := \int_0^t \sum_{k\in\mathcal{K}} C_k\left(\widehat{Q}_k^r(s)\right) ds. \tag{8}$$

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

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

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

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

here $\{\mathcal{U}_*^r\}$ is the family of cumulative queueing costs defined through (8) *under the family of control policies $\{\pi_*^r\}$, and $\{\mathcal{U}^r\}$ is the sequence of queueing costs corresponding to any other asymptotically compliant family of policies $\{\pi^r\}$.*

5. Main results

5.1. Cost functions and an optimization problem

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

$$\begin{aligned} \min_x \quad & \sum_{k\in\mathcal{K}} C_k(x_k) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{K}} m_k^e x_k = a, \\ & x \ge 0. \end{aligned} \tag{9}$$

We denote the optimal solution as

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

The mapping $\Delta_{\mathcal{K}} : \mathbb{R}_+ \to \mathbb{R}_+^K$ is part of the lifting mapping used in our state-space collapse result; see Theorem 3.

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* (9) *such that $x_k^* > 0$, $k \in \mathcal{K}$.*

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

$$\begin{aligned} C_k'(x_k^*) - \alpha_0 m_k^e &= 0, \\ \sum_{k\in\mathcal{K}} m_k^e x_k^* &= a. \end{aligned} \tag{10}$$

It is easy to see that this optimal vector x^* satisfies $C_l'(x_l^*)/m_l^e = C_k'(x_k^*)/m_k^e$, for all $l, k \in \mathcal{K}$. Using this fact, the proof of the following is elementary:

Lemma 5.1 *The function $\Delta_{\mathcal{K}}(\cdot)$ is well defined, and $\Delta_k(a)$ is nondecreasing in a, for each $k \in \mathcal{K}$.*

5.2. A lower bound

Our first result gives a lower bound for the costs, among all asymptotically compliant families of policies.

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

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

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

$$\begin{aligned} &\sum_{j\in\mathcal{J}} (m_j^e)^2 \lambda_j a_j^2 + \sum_{j\in\mathcal{J}} \left(\sum_{k\in\mathcal{K}} m_k^e P_{jk} - m_j^e \right)^2 \lambda_j b_j^2 + \sum_{k\in\mathcal{K}} \left(\sum_{l\in\mathcal{K}} P_{kl} m_l^e - m_k^e \right)^2 \lambda_k b_k^2 \\ &+ \sum_{j\in\mathcal{J}} \lambda_j (M^e)^T \Gamma^j M^e + \sum_{k\in\mathcal{K}} \lambda_k (M^e)^T \Gamma^k M^e. \end{aligned} \tag{11}$$

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

Theorem 1 (Lower Bound) *Fix any asymptotically compliant family of policies, with the corresponding cumulative costs $\mathcal{U}^r$ defined in* (8)*. Then for any $t, x > 0$,*

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

This theorem is proved in §EC.2.

5.3. The proposed policy and its asymptotic optimality

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

- When becoming idle, the physician deploys a threshold policy to determine which type of patient classes to serve next – a triage-type patient or an IP-type patient. Fix any $j \in \mathcal{J}$, for example, $1 \in \mathcal{J}$:

  —If $Q_1^r(t) \geq \lambda_1^r d_1^r$, priority is given to triage-type patients;

— Otherwise, priority is given to IP-type patients.

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

$$j \in \arg\max_{j\in\mathcal{J}} \frac{\tau_j^r(t)}{d_j^r}. \tag{12}$$

• If the IP classes are chosen to be served at time t, the physician uses a policy ensuring (for any $T>0$)

$$\max_{l,k\in\mathcal{K}} \sup_{0\le t\le T} \left| \frac{C_l'(\widehat{Q}_l^r(t))}{m_l^e} - \frac{C_k'(\widehat{Q}_k^r(t))}{m_k^e} \right| \Rightarrow 0. \tag{13}$$

An example of such a policy is to choose $k \in \arg\max_{k\in\mathcal{K}} \frac{C_k'(\widehat{Q}_k^r(t))}{m_k^e}$, which is a modified generalized $c\mu$-rule. (More examples of policies ensuring (13) can be found in §6.2.)

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

Theorem 2 (Asymptotic Optimality) *The family of control policies $\{\pi_*^r\}$ is asymptotically optimal.*

In proving Theorem 2, we show that the proposed policy makes the system "well behaved", in the sense that the weighted queue length converges, and there is state-space collapse for the queue length processes; see Proposition 1 and Theorem 3 below.

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

$$\widehat{Q}_w^r(t) = \sum_{j\in\mathcal{J}} m_j^e \widehat{Q}_j^r(t) + \sum_{k\in\mathcal{K}} m_k^e \widehat{Q}_k^r(t). \tag{14}$$

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

$$\widehat{Q}_w^r \Rightarrow \widehat{Q}_w, \quad as \quad r\to\infty. \tag{15}$$

This proposition is proved in §EC.3.

To state the state-space collapse result, define the lifting vector $\Delta_{\mathcal{J}}:\mathbb{R}_+\to\mathbb{R}_+^J$ as the J-dimensional vector $x=\Delta_{\mathcal{J}}a$, which is the solution to the following equation

$$\begin{aligned} &\sum_{j\in\mathcal{J}} m_j^e x_j = a, \\ &\frac{x_j}{\lambda_j \widehat{d}_j} = \frac{x_{j'}}{\lambda_{j'}\widehat{d}_{j'}}, \quad \text{for} \quad j,\ j'\in\mathcal{J}. \end{aligned}$$

As in Lemma 5.1, we can also prove $\Delta_{\mathcal{J}}(\cdot)$ is well-defined, and Δ_j is nondecreasing for each $j\in\mathcal{J}$. Unlike $\Delta_{\mathcal{K}}$, the mapping $\Delta_{\mathcal{J}}$ is linear. The function pair $(\Delta_{\mathcal{J}},\Delta_{\mathcal{K}})$ is the lifting mapping in the state-space collapse result. Let $\widehat{Q}^r=\{\widehat{Q}_j^r,\ j\in\mathcal{J},\ \widehat{Q}_k^r,\ k\in\mathcal{K}\}$ and recall $\widehat{\omega}=\sum_{j\in\mathcal{J}}\lambda_j\widehat{d}_j m_j^e$.

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

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

This theorem is proved in §EC.4.

5.4. Virtual waiting times

In this and the next subsection, we analyze our family of control policies $\{\pi_*^r\}$. In addition, we assume that the service discipline among 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 service. (Note that this definition 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 $\omega_j^r(t)$ and $\omega_k^r(t)$ as the virtual waiting times for j-triage class and k-IP class respectively, and define the diffusion scaled virtual waiting time processes by

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

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

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

This proposition is proved in §EC.7.

Remark 1 *From the convergence of* $\widehat{Q}^r$ *in Theorem 3, one can obtain the convergence of the vector of virtual waiting times under the family of control policies* $\{\pi_*^r\}$.

Recall that $\tau_j^r(t)$ is defined as the age of the head-of-the-line j-triage patient in the rth system. Now, define $\tau_k^r(t)$ as the age of the head-of-the-line k-IP patient in the rth system, and similarly its diffusion scaling $\widehat{\tau}_k^r(t) = r^{-1}\tau_k^r(r^2 t)$, $k \in \mathcal{K}$. Our next proposition establishes connections between the virtual waiting time processes and the age processes. This kind of result is often referred to as a *snapshot principle*.

Proposition 3 (Snapshot Principle – Virtual Waiting Time and Age) *Under the family of control policies* $\{\pi_*^r\}$, *with FCFS among each IP patient class, when* $r \to \infty$,

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

This proposition is proved in §EC.8.

5.5. Sojourn times

We consider sojourn times associated with specific routes through the system, as in Reiman (1984). We associate a route vector $h \in \mathbb{Z}_+^K$ with each patient who goes 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 that a patient entering the system as a j-triage patient has a route vector h. Denote $W_{jh}^r(t)$ as the *sojourn time* of the next j-triage patient, arriving after t, with route vector h, and the diffusion scaled processes

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

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

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

This proposition is proved in §EC.9.

Remark 2 *From Theorem 3, when $r \to \infty$,*

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

Then Proposition 4 gives rise to

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

being a good candidate for estimating the distribution of $\widehat{W}_{jh}^r(\cdot)$.

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

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

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

Remark 3 *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 they know their route). As in Reiman (1984), the diffusion limit does not depend on the specific order in which the physician is visited.*

6. Extensions and further discussion

6.1. Alternative triage policies to (12)

The recipe in (12), as part of an asymptotically optimal policy, is not unique. From our proof in §EC.4, it will be seen that any asymptotically compliant family of control policies ensuring

$$\sum_{j\in\mathcal{J}} m_j^e \widehat{Q}_j^r(\cdot) \Rightarrow \min\left(\widehat{Q}_w(\cdot), \widehat{\omega}\right), \quad \text{as} \quad r\to\infty, \tag{17}$$

is asymptotically optimal (recall that $\widehat{Q}_w$ and $\widehat{\omega}$ are defined in §5.2). One such control policy, assuming that triage classes are chosen to be served at time t, is having the physician cater to the head-of-the-line patient from the class with index

$$j \in \arg\max_{j\in\mathcal{J}} \frac{Q_j^r(t)}{\lambda_j^r d_j^r};$$

the latter can be easily proved asymptotically equivalent to (12).

Next we consider the *Shortest-Deadline-First* policy: when the triage classes are chosen to be served at time t, the physician chooses the head-of-the-line patient from the class with index

$$j \in \arg\min_{j\in\mathcal{J}} \left(d_j^r - \tau_j^r(t)\right). \tag{18}$$

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

$$j \in \arg\min_{j\in\mathcal{J}} \left(d_j^r - Q_j^r(t)/\lambda_j^r\right).$$

Following the same framework in §EC.4.1 and §EC.4.2, one can prove that, for any $T\geq 0$, as $r\to\infty$,

$$\sup_{0\leq t\leq T}\left|\widehat{Q}_j^r(t) - \widetilde{\Delta}_j \min\left(\sum_{j\in\mathcal{J}} m_j^e \widehat{Q}_j^r(t) + \sum_{k\in\mathcal{K}} m_k^e \widehat{Q}_k^r(t),\ \widehat{\omega}\right)\right| \Rightarrow 0.$$

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

$$\widetilde{\Delta}_{j_1}(a) = \begin{cases} \lambda_{j_1}\left(\widehat{d}_{j_1} - \widehat{d}_{j'} + \left(a - \sum_{j\in\mathcal{J}}\lambda_j m_j^e(\widehat{d}_j - \widehat{d}_{j'})^+\right)/j'\right), & \text{for} \quad j_1 \leq j', \\ 0, & \text{for} \quad j_1 > j'. \end{cases}$$

One can now prove that the family of control policies, with (18) replacing (12), is asymptotically compliant, and satisfies (17) – it is thus asymptotically optimal.

The expression of $\widetilde{\Delta}_{\mathcal{J}}$ is more complicated than $\Delta_{\mathcal{J}}$. On the other hand, a discussion in Plambeck et al. (2001) suggests that the policy in (12) is a more natural one, as it uses a 'relative' term. As a result, we choose (12) for elaboration. The comparison of (12) to (18) may involve rates of convergence, which is beyond the scope of the present paper.

6.2. IP-Policies that imply (13)

For any $K \times K$-dimensional invertible matrix G, with all components of GM^e nonzero, let H denote the K-dimensional vector with the kth component $1/(GM^e)_k$. When the IP classes are chosen to be served at time t, the physician chooses a patient from the class with index

$$k \in \arg\max_{k \in \mathcal{K}} H_k \left(GC' \left(\widehat{Q}^r(t) \right) \right)_k ; \tag{19}$$

here $C'(\widehat{Q}^r(t))$ is a K-dimensional column vector with $C'_k(\widehat{Q}^r_k(t))$ being its kth component.

For any $l, k \in \mathcal{K}$, denote

$$\mathcal{H}^r_{kl}(t) = H_k \left(GC' \left(\widehat{Q}^r(t) \right) \right)_k - H_l \left(GC' \left(\widehat{Q}^r(t) \right) \right)_l .$$

Similarly to the proof of Proposition 1 in van Mieghem (2003), one can prove that, for any $T \geq 0$, as $r \to \infty$,

$$\sup_{l,k \in \mathcal{K}} \sup_{0 \leq t \leq T} |\mathcal{H}^r_{kl}(t)| \Rightarrow 0. \tag{20}$$

For any fixed $k \in \mathcal{K}$, note that $\mathcal{H}^r_{kl}(t)$ is the lth component of the following

$$\mathcal{H}^r_k(t) := \mathcal{B}_k GC'(\widehat{Q}^r(t)); \tag{21}$$

here $\mathcal{B}_k$ is a $K \times K$ matrix defined as $\mathcal{B}_k = \Upsilon + \Theta_k$, where Υ is a $K \times K$ diagonal matrix with component $-H_l$ in the lth place, and Θ_k is a $K \times K$ matrix with its kth column being H_k while all others are 0, that is,

$$\Upsilon = \begin{pmatrix} -H_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \vdots \\ 0 & \cdots & \ddots & \cdots & 0 \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & -H_K \end{pmatrix} \quad \text{and} \quad \Theta_k = \begin{pmatrix} 0 & \cdots & H_k & \cdots & 0 \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ 0 & \cdots & H_k & \cdots & 0 \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & H_k & \cdots & 0 \end{pmatrix}.$$

It is easy to verify that the vector M^e is the only column vector (up to scaling) satisfying

$$\mathcal{B}_k GM^e = 0.$$

From this uniqueness and $m^e_k \neq 0$, we deduce that, after deleting the kth column of $\mathcal{B}_k G$, the remaining matrix has rank $K-1$. Denote this new matrix by $\mathcal{W}_k$ and fix any $l \in \mathcal{K} \backslash \{k\}$. Then we can always find a vector D_{kl} ensuring that all the elements of $D^T_{kl}\mathcal{W}_k$ are 0, except the lth components, which we denote by d_l (for example, we require $d_l = 1$). Denote by $-d_k$ the term in the kth places of $D^T_{kl}\mathcal{B}_k G$. From $D^T_{kl}\mathcal{B}_k GM^e = 0$ and $M^e > 0$, we deduce that $d_k m^e_k = d_l m^e_l$. As a result, $d_k \neq 0$ and $d_k/d_l = m^e_l/m^e_k$.

According to (20) and (21), for any $T > 0$ and $l, k \in \mathcal{K}$,

$$\sup_{0 \leq t \leq T} \left| d_k C'_k \left(\widehat{Q}^r_k(t) \right) - d_l C'_l \left(\widehat{Q}^r_l(t) \right) \right| \Rightarrow 0.$$

Using $d_k/d_l = m_l^e/m_k^e$, we have

$$\sup_{0\le t\le T}\left|\frac{C_k'\left(\widehat{Q}_k^r(t)\right)}{m_k^e}-\frac{C_l'\left(\widehat{Q}_l^r(t)\right)}{m_l^e}\right|\Rightarrow 0,$$

for all $l,k\in\mathcal{K}$, (13) holds.

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

1. $G=I$: then $H_k=1/m_k^e$; hence (19) is

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

This is a generalized $c\mu$ policy, 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 μ_k being the kth component; hence (19) is

$$k\in\arg\max_{k\in\mathcal{K}}\left[C_k'\left(\widehat{Q}_k^r(t)\right)-\sum_{l\in\mathcal{K}}P_{kl}C_l'\left(\widehat{Q}_l^r(t)\right)\right]\mu_k.$$

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

Our expression in (13) is similar to equation (51) in van Mieghem (1995), with the waiting times there replaced by the queue lengths, and the mean service times there replaced by the 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 policy is also expected to satisfy (13): when the IP classes are chosen to be served, the physician chooses a patient from the class with index $k\in\arg\max_{k\in\mathcal{K}}C_k'(\widehat{Q}_k^r(t))/m_k^e$, and serves this patient continuously until completing all services – the current one as well as feedbacks. This exhaustive policy is not FCFS within each IP class. Alternatively, this system can be viewed as a new one with no feedback, but with the service times for k-IP patients being now the cumulative service requirement – with mean m_k^e. To have this system enjoy asymptotically the queueing-cost lower bound in Theorem 1, there must exist at least one triage class for each IP class, such that after the triage service, this class of triage patients will transfer directly to the IP class with positive probability – that is, for each column in $P_{\mathcal{J}\mathcal{K}}$, there must be at least one positive element. Needless to say, such is not plausible in an ED setup.

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. Recall that $\omega_k^r(t)$ is the virtual waiting time of a k-IP patient at time t, and its diffusion scaling $\widehat{\omega}_k^r(t)$ is defined in (16). We seek to stochastically minimize the following cost:

$$\widetilde{\mathcal{U}}^r(t):=\sum_{k\in\mathcal{K}}\int_0^t C_k\left(\widehat{\omega}_k^r(s)\right)d\bar{\bar{E}}_k^r(s), \tag{22}$$

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

We now slightly modify the control policy $\{\pi_*^r\}$ in Section 5. The first step, using a threshold policy to determine between triage classes and IP classes, and the step using (12) to determine priorities among triage patients, do not change. The step determining the priority among IP classes changes as follows:

- If the IP classes are chosen to be served at time t, the physician uses a policy ensuring that, for any $T \geq 0$,

$$\max_{l,k\in\mathcal{K}} \sup_{0\leq t\leq T} \left| \frac{C_l'\left(\frac{\widehat{Q}_l^r(t)}{\lambda_l^r}\right)}{m_l^e} - \frac{C_k'\left(\frac{\widehat{Q}_k^r(t)}{\lambda_k^r}\right)}{m_k^e} \right| \Rightarrow 0.$$

An example of such a policy is to choose $k \in \arg\max_{k\in\mathcal{K}} \frac{C_k'\left(\widehat{Q}_k^r(t)/\lambda_k^r\right)}{m_k^e}$. Other examples of policies satisfying the above can be deduced from the policies in §6.2.
Denote this family of modified policies by $\{\widetilde{\pi}_*^r\}$.

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

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

where $\{\widetilde{\mathcal{U}}_*^r\}$ is the family of cumulative cost, defined through (22) *under the family of control policies $\{\widetilde{\pi}_*^r\}$, and $\{\widetilde{\mathcal{U}}^r\}$ is the corresponding cost under any other asymptotically compliant family of work-conserving policies $\{\pi^r\}$.*

The outline of the proof can be found in §EC.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 the figure in §1.1, except that congestion cost is associated with each patient's sojourn time in the IP stage (as opposed to queueing and waiting costs previously). We now add the assumption that the routing matrix P is upper-triangular. 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 has a deterministic routing vector and there are finite number of routing vectors. We use $\mathcal{C}_0$ to denote the set of *starting classes* of routes, for $k \in \mathcal{C}_0$, let $\mathcal{C}_k$ denote all the classes on the route that starts at k. If the waiting time of a patient with starting class k waits $\omega_{k'}$, as a k'-IP patient ($k' \in \mathcal{C}_k$), then the sojourn time of this patient is $\sum_{k'\in\mathcal{C}_k} \omega_{k'}$. We call the class in $\bigcup_{k\in\mathcal{C}_0} \mathcal{C}_k \backslash \{k\}$ a *subsequent class.*

Our problem is to stochastically minimize the cost

$$\widetilde{\mathcal{S}}^r(t) = \sum_{k\in\mathcal{C}_0} \int_0^t C_k \left(\sum_{k'\in\mathcal{C}_k} \widehat{\omega}^r_{k'}(s) \right) d\bar{\bar{E}}^r_k(s), \tag{23}$$

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

We propose the following routing policy: the first step, using a threshold policy to determine the priority between triage classes and IP classes, and the step using (12) to determine priorities among triage patients, do not change. The step determining the priority among IP classes will change as follows:

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

$$\max_{l,k\in\mathcal{C}_0} \sup_{0\le t\le T} \left| \frac{C_l'\left(\frac{\widehat{Q}^r_l(t)}{\lambda^r_l}\right)}{m^e_l} - \frac{C_k'\left(\frac{\widehat{Q}^r_k(t)}{\lambda^r_k}\right)}{m^e_k} \right| \Rightarrow 0. \tag{24}$$

Here Q_l, Q_k are the queue lengths of the starting classes $j,k\in\mathcal{C}_0$, and m^e_l, m^e_k are the corresponding effective means of service times. An example of such a policy is to choose $k\in \arg\max_{k\in\mathcal{C}_0} \frac{C_k'\left(\widehat{Q}^r_k(t)/\lambda^r_k\right)}{m^e_k}$. Other examples of policies satisfying the above can be modified from the policies in §6.2.

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

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

$$\limsup_{r\to\infty} \mathbb{P}\left\{\widetilde{\mathcal{S}}^r_{**}(t) > x\right\} \le \limsup_{r\to\infty} \mathbb{P}\left\{\widetilde{\mathcal{S}}^r(t) > x\right\}; \tag{25}$$

here $\{\widetilde{\mathcal{S}}^r_{**}\}$ is the family of cumulative cost defined through (23) *under the family of control policies $\{\widetilde{\pi}^r_{**}\}$, and $\{\widetilde{\mathcal{S}}^r\}$ is the corresponding cost under any other asymptotically compliant family of policies $\{\pi^r\}$.*

The outline of the proof can be found in §EC.11.

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.

6.5. Remark on FCFS multiclass queues with feedback

As mentioned already, Dai and Kurtz (1995) analyzed a multiclass queueing network with Markovian feedback, under the FCFS policy across all classes. We remark that our analysis can be also applied to prove convergence of the queue length processes there. Indeed, our present results yield convergence of the weighted queue length to a reflected Brownian motion, under any work-conserving policy. 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)) - details are omitted.

7. An ED case study: the value of information & imputed costs

Most triage indices 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 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.) 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 $40\% \times 28\% = 11.2\%$. 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

$$c(t) = 0.4c_a(t) + 0.6c_d(t) = 1.6t^2. \tag{26}$$

In §EC.12 of the Appendix, 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.12, that an 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 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 $1.6(\frac{(\widetilde{Q}_w-\widehat{\omega})^+}{m_1^e})^2 = 1.6\times 0.1874\frac{(\widetilde{Q}_w-\widehat{\omega})^2}{m^2} = 0.2998\frac{(\widetilde{Q}_w-\widehat{\omega})^2}{m^2}$. As a reminder, here $\widetilde{Q}_w$ is a reflected Brownian motion, $\widehat{\omega}$ is a weighted summation of the triage deadlines, and both can be calculated via the formulae in §5.2.

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 (26). 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

$$k \in \max_{l\in\mathcal{K}} \frac{Q_l(t)}{l\times p_l}. \tag{27}$$

Here Q_l is the queue length of class l IP patients, and p_l is the fraction of patients that visit the physician l times, $l=1,\cdots,5$. From the argument in the Appendix (especially (EC.61) and the paragraph above it), the minimal cost rate will be the value of the following problem:

$$\begin{aligned}\min\ & 0.28c(\frac{Q_1}{0.28})+0.30c(\frac{Q_2}{0.30})+0.28c(\frac{Q_3}{0.28})+0.11c(\frac{Q_4}{0.11})+0.03c(\frac{Q_5}{0.03})\\ \text{s.t.}\ & m(Q_1+2Q_2+3Q_3+4Q_4+5Q_5)=(\widetilde{Q}_w-\widehat{\omega})^+.\end{aligned}$$

with Q_i being the queue length of starting class i (i phases). Then the optimal solution satisfies $Q_5^*=\frac{0.15}{0.28}Q_1^*, Q_4^*=\frac{0.44}{0.28}Q_1^*, Q_3^*=\frac{0.84}{0.28}Q_1^*, Q_2^*=\frac{0.6}{0.28}Q_1^*$, with $\frac{Q_1^*}{0.28}=\frac{(\widetilde{Q}_w-\widehat{\omega})^+}{m(0.28+1.2+2.52+1.76+0.75)}$. Simple algebra leads to the asymptotically minimal cost rate of

$$\begin{aligned}&(0.28+0.3\times 4+0.28\times 9+0.11\times 16+0.03\times 25)\times 1.6\times\left(\frac{Q_1^*}{0.28}\right)^2\\ =&\frac{1.6\times(\widetilde{Q}_w-\widehat{\omega})^2}{m^2(0.28+1.2+2.52+1.76+0.75)}=1.6\times 0.1536\frac{(\widetilde{Q}_w-\widehat{\omega})^2}{m^2}=0.2458\frac{(\widetilde{Q}_w-\widehat{\omega})^2}{m^2}.\end{aligned}$$

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.61) 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:

$$\begin{aligned}\min\ & \frac{1}{0.6}\left(0.28c_a(\frac{Q_{a1}}{0.28})+0.30c_a(\frac{Q_{a2}}{0.30})+0.28c_a(\frac{Q_{a3}}{0.28})+0.11c_a(\frac{Q_{a4}}{0.11})+0.03c_a(\frac{Q_{a5}}{0.03})\right)\\ &+\frac{1}{0.4}\left(0.28c_d(\frac{Q_{d1}}{0.28})+0.30c_d(\frac{Q_{d2}}{0.30})+0.28c_d(\frac{Q_{d3}}{0.28})+0.11c_d(\frac{Q_{d4}}{0.11})+0.03c_d(\frac{Q_{d5}}{0.03})\right)\\ \text{s.t.}\ & m(Q_{a1}+2Q_{a2}+3Q_{a3}+4Q_{a4}+5Q_{a5}+Q_{d1}+2Q_{d2}+3Q_{d3}+4Q_{d4}+5Q_{d5})=(\widetilde{Q}_w-\widehat{\omega})^+.\end{aligned}$$

(In the above, we use the fact that c_a and c_d are quadratic functions, and $b(\frac{x}{b})^2=\frac{1}{b}x^2$).) Similarly to the partial information case, our problem can be further reduced to the following:

$$\begin{aligned}\min\ & (0.28+0.3\times4+0.28\times9+0.11\times16+0.03\times25)\times\left(\frac{2}{0.6}\times\left(\frac{Q_{a1}}{0.28}\right)^2+\frac{1}{0.4}\times\left(\frac{Q_{d1}}{0.28}\right)^2\right)\\ \text{s.t.}\ & \frac{Q_{a1}+Q_{d1}}{0.28}=\frac{(\widetilde{Q}_w-\widehat{\omega})^+}{m(0.28+1.2+2.52+1.76+0.75)}.\end{aligned}$$

The optimal value, namely the minimal cost rate, is $\frac{10}{7}\times0.1536\frac{(\widetilde{Q}_w-\widehat{\omega})^2}{m^2}=0.2194\frac{(\widetilde{Q}_w-\widehat{\omega})^2}{m^2}$. As $\frac{0.2458-0.2194}{0.2458}=0.1074$ and $\frac{0.2998-0.2194}{0.2998}=0.2682$, 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%.

7.1. Imputed cost

Our ED case study was based on expert estimates of costs in an Israeli hospital. Generally, such cost parameters are unavailable, which 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,

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

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, already as is, captures usefully the control of ED patient flow, it does leave out several noticeable ED characteristics. In the following, we list some of the features that, we believe, are research worthy.

8.1. Adding delays between transfers

In emergency department, there are delays between successive patient visits to physicians. In Yom-Tov and Mandelbaum (2011), 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. The question is the precise meaning of "short" and "long", which we now formalize.

We consider the basic model as an example. Similarly to Yom-Tov and Mandelbaum (2011), we model the delays as infinite-server queues with exponential service times. The individual service rate for the infinite-server queue between j-triage patients and k-IP patients is $r^{\alpha_{jk}}\mu_{jk}$, and the one between l-IP patients and k-IP patients is $r^{\alpha_{lk}}\mu_{lk}$. Here μ_{jk} and μ_{lk} are fixed positive constants. The magnitude of the α's will determine "short" delays (large α) vs. "long" (small). Specifically, we conjecture that when $\alpha > -2$ (for all α's), the delays are then short enough to leave our results intact. Conversely, $\alpha_{jk} < -2$ (for all j, k) decouples the triage from IP - both can be controlled separately; and $\alpha_{lk} < -2$ (for all l, k) pushes the IP feedback far enough into the future so that the IP sub-system can be analyzed as a queueing system without feedback. All other cases require further thought and plausibly a more delicate analysis. We further provide a brief discussion in §EC.13.

8.2. Time-varying arrival rates

Emergency departments, like many other service systems, must cope with arrival rates that are significantly time-varying (Yom-Tov and Mandelbaum 2011, Figure 10). In the present paper, we have focused our attention on the ED afternoon-evening peak, which rendered 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 can not 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.

8.3. Length-of-Stay constraints

Many EDs implement, or at least strive for, an upper bound on patients' overall Length of Stay (LOS). In our ED-Partner, for example, the goal 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. To this end, one could, perhaps should apply a rationalized admission control - a rare protocol in our ED-Partner, but relatively prevalent in U.S. EDs in the form of ambulance diversion (Deo and Gurvich (2011), Allon et al. (2008), Armony et al. (2011)). Interestingly, admission control problems, with costs incurred by blocked customers, in fact motivated Plambeck et al. (2001). But we opted for the analysis of triage-constraints first, in the belief that they play a higher order (clinical) role. Nevertheless, accommodating LOS and Triage constraints simultaneously is of interest and significance - we thus leave it for future research.

8.4. Adding abandonment to triage or IP patients

Statistical evidence shows that the fraction of registered emergency patients who 'Leave Without Being Seen' (LWBS) is around 5% (Armony et al. (2011)). This has become a growing concern in overcrowded EDs, as those LWBS patients may miss out 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 Garnett et al. (2002), Mandelbaum and Zeltyn (2009).

Indeed, abandonment could significantly impact the structure of optimal policies. For systems without feedback, Kim and Ward (2012) 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 the works analyze the corresponding Brownian control problem, and then interpret the results to the original queueing systems. Both works 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

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.

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

References

Allon, G., S. Deo, W. Lin. 2008. Impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Working Paper.* Northwestern University.

Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, G. B. Yom-Tov. 2011. Patient flow in hospitals: A data-based queueing-science perspective. *Working Paper.* Technion – Israel Institute of Technology.

Ata, B., H. M. Tongarlak. 2012. On scheduling a multiclass queue with abandonments under general delay costs. *Queueing Systems.* Forthcoming.

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

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

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

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

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

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

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

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

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

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

Dobson, G., T. Tezcan, V. Tilson. 2012. Optimal workflow decisions for investigators in systems with interruptions. *Management Science.* Forthcoming.

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

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

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

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

Kim, J., A. R. Ward. 2012. Dynamic scheduling of a GI/GI/1+ GI queue with multiple customer classes. *Queueing Systems.* Forthcoming.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L. Kronick. 2011. Complexity-based triage: A tool for improving patient safety and operational efficiency. *Working Paper.* Arizona State University and University of Michigan.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L. Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research.* Forthcoming.

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

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

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

Yom-Tov, G. B., A. Mandelbaum. 2011. Erlang-R: A time-varying queue with ReEntrant customers, in support of healthcare staffing. *Working Paper.* Technion – Israel Institute of Technology.

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

We start with an analysis that covers any asymptotically compliant family of control policies.

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

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

and fluid scaled processes

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

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

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

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

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

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

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

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

$$Q_j^r(t) = E_j^r(t) - E_j^r\left((t - \tau_j^r(t))-\right).$$

Then

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

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

The following is a direct corollary, which translates the asymptotic compliance condition to the language of queue length processes.

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

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

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

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

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

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

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

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

then

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

and

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

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

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

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

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

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

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

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

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

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

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

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

respectively.

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

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

and

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

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

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

here

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

Denote ($\widehat{Q}_w^r(t)$ is defined in the paper, but we would like to repeat it here)

$$\begin{aligned}\widehat{Q}_w^r(t) =& \sum_{j\in\mathcal{J}} m_j^e \widehat{Q}_j^r(t) + \sum_{k\in\mathcal{K}} m_k^e \widehat{Q}_k^r(t),\\ \widehat{X}_w^r(t) =& \widehat{Q}_w^r(0) + r(\rho^r - 1)t + \sum_{j\in\mathcal{J}} m_j^e\left[\widehat{E}_j^r(t) - \widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t)\right)\right] + \sum_{k\in\mathcal{K}} m_k^e\left[\widehat{\mathcal{E}}_k^r(t) - \widehat{S}_k^r\left(\bar{\bar{T}}_k^r(t)\right)\right],\\ \widehat{T}_+^r(t) =& r^{-1}\left(r^2 t - \sum_{j\in\mathcal{J}} T_j^r(r^2 t) - \sum_{k\in\mathcal{K}} T_k^r(r^2 t)\right).\end{aligned} \tag{EC.14}$$

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

$$-m_j^e\mu_j + \sum_{k\in\mathcal{K}} P_{jk}\mu_j m_k^e = -1, \tag{EC.15}$$

$$-m_k^e\mu_k + \sum_{l\in\mathcal{K}} P_{kl}\mu_k m_l^e = -1. \tag{EC.16}$$

Multiply (EC.7) by m_j^e, (EC.12) by m_k^e, and summing them together, one has

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

Note that the policy may not be work-conserving, thus it is possible that $\widehat{T}_+^r$ increases at t when $\widehat{Q}_w^r(t) \neq 0$. Hence

$$\widehat{Q}_w^r(t) \geq \Phi(\widehat{X}_w^r)(t); \tag{EC.18}$$

here Φ is the 1-dimensional Skorohod mapping; see for example, Mandelbaum and Stolyar (2004). Equality in (EC.18) holds when the system operates under any work-conserving policy.

EC.2. Proof of Theorem 1: Lower Bound

Proof of Theorem 1: Fix an arbitrary family of control policies $\{\pi^r\}$ which is asymptotically compliant. Define

$$\Gamma_1^r(t) = \left\{\mathcal{U}^r(t) > x,\ \max_{k\in\mathcal{K}} \sup_{0\le s\le t} \bar{\bar{Q}}_k^r(s) \le \frac{1}{r^{1/4}}\right\},$$
$$\Gamma_2^r(t) = \left\{\max_{k\in\mathcal{K}} \sup_{0\le s\le t} \bar{\bar{Q}}_k^r(s) > \frac{1}{r^{1/4}}\right\},$$
$$\Gamma_3^r(t) = \left\{\mathcal{U}^r(t) \le x,\ \max_{k\in\mathcal{K}} \sup_{0\le s\le t} \bar{\bar{Q}}_k^r(s) > \frac{1}{r^{1/4}}\right\}.$$

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

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

Then

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

First we prove

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

For notation simplicity, denote $I^r(s,\vartheta) = [s, s+\frac{1}{\vartheta r^{1/4}}]$ and $\vartheta_0 = 4\max_{k\in\mathcal{K}}\mu_k$. For $s<u$, denote $S_k^r(s,u) = S_k\left(T^r(r^2 s) + r^2(u-s)\right) - S_k\left(T^r(r^2 s)\right)$ and $\bar{\bar{S}}_k^r(s,u) = r^{-2} S_k^r(s,u)$. One can prove that

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

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

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

It follows that

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

This completes the proof of (EC.20).

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

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

Next we derive a lower bound for the latter term.

Denote

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

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

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

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

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

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

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

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

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

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

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

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

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

which is equivalent to (EC.22).

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

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

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

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

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

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

on $\mathcal{D}[0,t]$. Here $\widehat{X}$ is the Brownian motion defined in §5.2. For $s \le t$, denote

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

then by the continuity of Φ and the definition of asymptotic compliance, on $\mathcal{D}[0,t]$, when $r\to\infty$,

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

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

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

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

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

Combined with (EC.21),

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

From the convergence of $\widehat{\mathcal{Z}}_+^r$, the right-hand side is exactly the lower bound in Theorem 1. This completes the proof. □

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

Proof of Proposition 1: For any family of work-conserving policies, besides (EC.17), the following is also true:

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

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

From (EC.10), (EC.1) and the fact that $\bar{\bar{T}}_j^r(s) \le s$, $j\in\mathcal{J}$ and $\bar{\bar{T}}_k^r(s) \le s$, $k\in\mathcal{K}$, it is easy to see that $\widehat{X}_w^r$ in (EC.14) is stochastically bounded. By the Lipschitz continuity of Φ (Mandelbaum and Stolyar (2004)), $\widehat{Q}_w^r$ is stochastically bounded, which implies the stochastic boundedness of $\widehat{Q}_j^r$, $j\in\mathcal{J}$, and $\widehat{Q}_k^r$, $k\in\mathcal{K}$. Then $\bar{\bar{Q}}_j^r \Rightarrow 0$ for $j\in\mathcal{J}$. Note that (EC.6) is still true. One then has

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

For $k\in\mathcal{K}$, following the procedure in proving (EC.22) in the proof of Theorem 1, one also has

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

Together with (EC.24), (EC.10), (EC.1) and the Random-Time-Change theorem, when $r\to\infty$,

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

By the continuity of the mapping Φ, (15) follows. □

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

EC.4.1. Hydrodynamic limit

We now start to analyze the family of control policies $\{\pi_*^r\}$. In the present subsection, we focus only on the triage part.

Under the policies $\{\pi_*^r\}$, one has the following dynamic equations of the system:

$$
\begin{aligned}
&Q_j^r(t) = Q_j^r(0) + E_j^r(t) - D_j^r(t), \quad j \in \mathcal{J},\\
&D_j^r(t) = S_j\left(T_j^r(t)\right), \quad j \in \mathcal{J},\\
&Q_k^r(t) = Q_k^r(0) + E_k^r(t) - D_k^r(t), \quad k \in \mathcal{K},\\
&E_k^r(t) = \sum_{j\in\mathcal{J}} \Phi_{jk}^r\left(S_j\left(T_j^r(t)\right)\right) + \sum_{l\in\mathcal{K}} \Phi_{lk}^r\left(S_l\left(T_l^r(t)\right)\right), \quad k \in \mathcal{K},\\
&D_k^r(t) = S_k\left(T_k^r(t)\right), \quad k \in \mathcal{K},\\
&\sum_{j\in\mathcal{J}}\left[T_j^r(t) - T_j^r(s)\right] + \sum_{k\in\mathcal{K}}\left[T_k^r(t) - T_k^r(s)\right] \le t - s, \quad \text{for} \quad s < t,\\
&Y^r(t) = t - \left(\sum_{j\in\mathcal{J}} T_j^r(t) + \sum_{k\in\mathcal{K}} T_k^r(t)\right),\\
&\int_0^\infty \left(\max_{j\in\mathcal{J}} \frac{\tau_j^r(t)}{d_j^r} - \frac{\tau_{j'}^r(t)}{d_{j'}^r}\right)^+ \wedge 1 dT_{j'}^r(t) = 0, \quad j' \in \mathcal{J},\\
&\int_0^\infty 1\left(Q_1^r(t) > \lambda_1^r d_1^r\right) d\sum_{k\in\mathcal{K}} T_k^r(t) = 0,\\
&\int_0^\infty 1\left(Q_1^r(t) < \lambda_1^r d_1^r,\ \sum_{k\in\mathcal{K}} Q_k^r(t) > 0\right) d\sum_{j\in\mathcal{J}} T_j^r(t) = 0,\\
&\int_0^\infty 1\left(\sum_{j\in\mathcal{J}} m_j^e Q_j^r(t) + \sum_{k\in\mathcal{K}} m_k^e Q_k^r(t) > 0\right) dY^r(t) = 0.
\end{aligned}
$$

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

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

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

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

First we can prove the following lemma, which is similar to Lemma EC.1.1.

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

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

$$
Q_j^r(t) = E_j^r(t) - E_j^r\left((t - \tau_j^r(t))-\right).
$$

Then

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

By the functional law of large numbers, $\sup_{0\le t\le T}|\bar{E}_j^r(t) - \lambda_j^r t| \Rightarrow 0$, together with (EC.27), the conclusion can be easily proved. □

Similar to Plambeck et al. (2001), we have the following

Proposition 7 *Almost surely, every sequence contains a subsequence* $\{r_n\}$ *such that, the hydrodynamic scaled processes* $\bar{E}_j^{r_n}, \bar{S}_j^{r_n}, \bar{\tau}_j^{r_n}, \bar{T}_j^{r_n}, \bar{Q}_j^{r_n}, \bar{D}_j^{r_n}, j \in \mathcal{J}$, $\bar{E}_k^{r_n}, \bar{S}_k^{r_n}, \bar{T}_k^{r_n}, \bar{Q}_k^{r_n}, \bar{D}_k^{r_n}, k \in \mathcal{K}$, *converge uniformly on compact time sets to limit processes* $\bar{E}_j, \bar{S}_j, \bar{\tau}_j, \bar{T}_j, \bar{Q}_j, \bar{D}_j, j \in \mathcal{J}$, $\bar{E}_k, \bar{S}_k, \bar{T}_k, \bar{Q}_k, \bar{D}_k, k \in \mathcal{K}$ *which satisfy the following equations*

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

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

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

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

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

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

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

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

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

$$\int_0^\infty 1\left(\bar{Q}_1(t) > \lambda_1 \widehat{d}_1\right) d\sum_{k\in\mathcal{K}} \bar{T}_k(t) = 0, \tag{EC.37}$$

$$\int_0^\infty 1\left(\bar{Q}_1(t) < \lambda_1 \widehat{d}_1, \sum_{k\in\mathcal{K}} \bar{Q}_k(t) > 0\right) d\sum_{j\in\mathcal{J}} \bar{T}_j(t) = 0, \tag{EC.38}$$

$$\int_0^\infty 1\left(\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j(t) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(t) > 0\right) d\bar{Y}(t) = 0. \tag{EC.39}$$

Remark 4 *We call any* $\bar{\mathcal{S}} = (\bar{E}_j, \bar{S}_j, \bar{\tau}_j, \bar{T}_j, \bar{Q}_j, \bar{D}_j, j \in \mathcal{J}, \bar{E}_k, \bar{S}_k, \bar{T}_k, \bar{Q}_k, \bar{D}_k, k \in \mathcal{K})$ *satisfying* (EC.28)-(EC.39) *a* hydrodynamic model solution*, and we can prove that, any hydrodynamic model solution is Lipschitz, hence absolutely continuous and differentiable almost everywhere.*

Proposition 8 *Any hydrodynamic model solution satisfies*

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

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

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

From (EC.39), (EC.34) and (EC.35), $\bar{Y}(\cdot)=0$. This completes the proof. □

EC.4.2. State-space collapse for triage patients

First we prove a state-space collapse result for the hydrodynamic model solution.

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

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

Furthermore, if

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

then $\bar{Q}_{\mathcal{J}}(t)=\bar{Q}_{\mathcal{J}}(0)$.

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

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

If $f_1(t)>0$ and is differentiable, then we claim

$$f_1'(t)=-\frac{1}{\widehat{d}_1}<0.$$

Indeed, if this is not true, then $\bar{T}_1'(t)\neq 0$ and from (EC.36), one has $\frac{\bar{Q}_1(t)}{\lambda_1\widehat{d}_1}=\max_{j\in\mathcal{J}}\frac{\bar{Q}_j(t)}{\lambda_j\widehat{d}_j}$. Together with $f_1(t)>0$, one can prove by contradiction that $\bar{Q}_1(t)<\lambda_1\widehat{d}_1$. Then from (EC.38), one has $\bar{Q}_k(t)=0$ for all $k\in\mathcal{K}$. This, together with $f_1(t)>0$, will contradict the definition of Δ_j.

As a result, f_1 will decrease to 0 in a finite time (denote it as T_1) and once becoming 0, it will never be positive again. Then for each $j\in\mathcal{J}$, if $f_j(t)>0$ for some $t\geq T_1$, then $\bar{T}_j'(t)=0$ from (EC.36), hence

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

Consequently, after a finite time (denote it by $T_2\geq T_1$), all f_j will be 0 and will never be positive again.

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

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

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

Now, as we have proved that $f_j(t) = 0$ for all $j \in \mathcal{J}$ and over $t \geq T_2$, together with $g_1(t) > 0$ and the definition of Δ_j, we have $\sum_{j \in \mathcal{J}} m_j^e \bar{Q}_j(t) + \sum_{k \in \mathcal{K}} m_k^e \bar{Q}_k(t) > \widehat{\omega}$, $\sum_{k \in \mathcal{K}} \bar{Q}_k(t) > 0$ and for $1 \in \mathcal{J}$, $\bar{Q}_1(t) > \lambda_1 \widehat{d}_1$. Then from (EC.37), $\sum_{k \in \mathcal{K}} \bar{T}_k'(t) = 0$. From (EC.39), we know $\sum_{j \in \mathcal{J}} \bar{T}_j'(t) = 1$. As a result, the derivative of $\sum_{j \in \mathcal{J}} \lambda_j \widehat{d}_j m_j g_j(t)$ is

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

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

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

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

$$\sup_{0 \leq t \leq T} \left| \widehat{Q}_j^r(t) - \Delta_j \min \left(\widehat{Q}_w^r(t), \ \widehat{\omega} \right) \right| \Rightarrow 0.$$

EC.4.3. State-space collapse for IP patients

From Propositions 1 and 10, when $r \to \infty$, one has

$$\sum_{k \in \mathcal{K}} m_k^e \widehat{Q}_k^r \Rightarrow \left(\widehat{Q}_w - \widehat{\omega} \right)^+. \tag{EC.41}$$

Recall that the proposed policy for IP patients is to ensure

$$\max_{l,k \in \mathcal{K}} \sup_{0 \leq t \leq T} \left| \frac{C_l'(\widehat{Q}_l^r(t))}{m_l^e} - \frac{C_k'(\widehat{Q}_k^r(t))}{m_k^e} \right| \Rightarrow 0. \tag{EC.42}$$

Proposition 11 *Under the family of control policies* $\{\pi_*^r\}$, *we have* $(\widehat{Q}_k^r, k \in \mathcal{K}) \Rightarrow (\widehat{Q}_k, k \in \mathcal{K})$. *Here*

$$\widehat{Q}_k = \Delta_k \left((\widehat{Q}_w - \widehat{\omega})^+ \right), \quad k \in \mathcal{K}. \tag{EC.43}$$

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

$$\max_{l,k\in\mathcal{K}} \sup_{0\le t\le T} \left| C_l'^{-1}\left(\frac{m_l^e}{m_k^e} C_k'\left(\widehat{Q}_k^r(t)\right)\right) - \widehat{Q}_l^r(t)\right| \Rightarrow 0. \tag{EC.44}$$

From the assumption on $C_k', k\in\mathcal{K}$, $C_l'^{-1}\left(\frac{m_l^e}{m_k^e} C_k'(\cdot)\right)$ is a nondecreasing function.

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

$$\sum_{l\in\mathcal{K}} m_l^e C_l'^{-1}\left(\frac{m_l^e}{m_k^e} C_k'\left(\widehat{Q}_k^r\right)\right) \Rightarrow \left(\widehat{Q}_w - \widehat{\omega}\right)^+.$$

As the function on the left-hand of the above equation has a continuous inverse, $\widehat{Q}_k^r$ converges. From (EC.44), we know $(\widehat{Q}_l^r, l\in\mathcal{K}) \Rightarrow (\widehat{Q}_l, l\in\mathcal{K})$. We also have,

$$\frac{C_l'(\widehat{Q}_l)}{m_l^e} = \frac{C_k'(\widehat{Q}_k)}{m_k^e}, \quad l,k\in\mathcal{K}.$$

This proves (EC.43). □

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

EC.5. Proof of Theorem 2: Asymptotic Optimality

Proof of Theorem 2: First, it can be verified that $\Delta_j \min(x,\widehat{\omega}) \le \lambda_j \widehat{d}_j$ for any x and $j\in\mathcal{J}$. Then from Theorem 3, under the proposed policies $\{\pi_*^r\}$, $\widehat{Q}_j^r \Rightarrow \widehat{Q}_j \le \lambda_j \widehat{d}_j$. An analysis of work-conserving policies will show that (EC.2) is equivalent to "asymptotic compliance" for work-conserving policies (see Lemma EC.6.2); hence the family of the policies $\{\pi_*^r\}$ is asymptotically compliant.

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

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

Hence, under the family of the proposed policies, the lower bound in Theorem 1 is attained. As a result, the family of the proposed policies is asymptotically optimal. □

EC.6. Additional results for work-conserving policies

In this section, we prove some additional results for work-conserving policies; in particular, they apply to $\{\pi_*^r\}$. From the discussion in proving Proposition 1, $\widehat{Q}_j^r$, $j\in\mathcal{J}$, are stochastically bounded and (EC.24) holds for any work-conserving policies. With these, notice that (EC.7) is still true, hence we can verify the convergence (EC.5). As $\widehat{Q}_j^r$, $j\in\mathcal{J}$, are stochastically bounded, $\widehat{T}_j^r$, $j\in\mathcal{J}$, are also stochastically bounded.

Next consider IP patients. Define $\widehat{\mathcal{Y}}_{\mathcal{K}}^r = (\widehat{\mathcal{Y}}_k^r)_{k\in\mathcal{K}}$ with each $k\in\mathcal{K}$,

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

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

$$\widehat{T}_\mu^r = (P^T - I)^{-1} \widehat{\mathcal{Y}}_\mathcal{K}^r. \tag{EC.45}$$

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

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

$$\widehat{E}_k^r(t) = \widehat{\mathcal{E}}_k^r(t) + \sum_{j\in\mathcal{J}} P_{jk}\mu_j \widehat{T}_j^r(t) + \sum_{l\in\mathcal{K}} P_{lk}\mu_l \widehat{T}_l^r(t). \tag{EC.46}$$

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

Define the fluid scaled virtual waiting time processes as

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

First we prove the following:

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

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

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

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

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

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

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

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

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

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

This completes the proof. □

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

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

Proof: The proof follows exactly that of Lemma EC.1.1, by noticing $\sup_{0\le s\le t}\bar{\bar{\tau}}_j^r(s)\Rightarrow 0$ which follows from Lemma EC.6.1 and the fact $\sup_{s\le t}\tau_j^r(s)\le\sup_{s\le t}\omega_j^r(s)$, for all t and j. Note that the result here is slightly different from Lemma EC.1.1, as only after proving this lemma, can we have the stochastic boundedness of $\widehat{\tau}_j^r, j\in\mathcal{J}$, for all work-conserving policies. □

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

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

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

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

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

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

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

Here

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

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

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

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

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

where

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

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

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

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

Thus it is enough to prove

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

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

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

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

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

Here $\triangle\widehat{S}_j^r(t+\bar{\bar{\omega}}_j^r(t))=\widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t+\bar{\bar{\omega}}_j^r(t))\right)-\widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t+\bar{\bar{\omega}}_j^r(t)-)\right)$ and $\triangle\widehat{T}_j^r\left(t+\bar{\bar{\omega}}_j^r(t)\right)=\widehat{T}_j^r(t+\bar{\bar{\omega}}_j^r(t))-\widehat{T}_j^r(t+\bar{\bar{\omega}}_j^r(t)-)$. From the convergence of $\widehat{S}_j^r(\bar{\bar{T}}_j^r(\cdot))$ and $\widehat{T}_j^r(\cdot)$, both $\triangle\widehat{S}_j^r(\cdot+\bar{\bar{\omega}}_j^r(\cdot))$ and $\triangle\widehat{T}_j^r\left(\cdot+\bar{\bar{\omega}}_j^r(\cdot)\right)$ converge to 0. Together with Lemma EC.6.1 and the convergence of $\widehat{E}_j^r, j\in\mathcal{J}$, the processes on the right-hand side 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 $\{\pi_*^r\}$, for any given $T>0$, when $r\to\infty$,*

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

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

Proof of Proposition 3: This can be easily 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 follows the framework in Reiman (1984). Introduce the following notation: $\tau^r_{jh}(t)$ is the time at which the patient of interest to us arrives to the system, and $\zeta^r_{jki}(t)$ is the time at which this patient becomes a k-IP patient for the ith time (it is also related to h, but we omit h to simplify the notation). Then

$$t \le \zeta^r_{jki}(t) \le \tau^r_{jh}(t) + W^r_{jh}(t). \tag{EC.54}$$

Define the fluid scaled processes

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

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

$$\sup_{0\le t\le T} \bar{\bar{W}}^r_{jh}(t) \Rightarrow 0, \tag{EC.55}$$

$$\sup_{0\le t\le T} \left[\bar{\bar{\tau}}^r_{jh}(t) - t\right] \Rightarrow 0. \tag{EC.56}$$

As a result, when $r \to \infty$,

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

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

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

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

From this we then have

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

From Lemma EC.9.1 and the convergence of $\widehat{\omega}^r_j, j\in\mathcal{J}$ and $\widehat{\omega}^r_k, k\in\mathcal{K}$,

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

Together with Proposition 2, the conclusion is immediate. □

Proof of Lemma EC.9.1: We first prove (EC.55). It is enough to show that, for any $\epsilon > 0$, there exists an $N < \infty$ such that, for all $r \geq N$,

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

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

$$\mathbb{P}\left\{\sup_{0\leq t\leq T} \bar{\bar{W}}^r_{jh}(t) \geq \epsilon\right\} \leq \max_{k\in\mathcal{K}} \mathbb{P}\left\{\sup_{0\leq t\leq T+\epsilon} \bar{\bar{\omega}}^r_k(t) \geq \frac{\epsilon}{\|h\|+1}\right\} + \mathbb{P}\left\{\sup_{0\leq t\leq T+\epsilon} \bar{\bar{\omega}}^r_j(t) \geq \frac{\epsilon}{\|h\|+1}\right\}. \tag{EC.58}$$

From Lemma EC.6.1, the right-hand side of (EC.58) converges to 0, hence (EC.55) holds.

The proof of (EC.56) follows the one in Reiman (1984). Let $L^r_{i,j,h} = \min\{n > i; h^r(j,n) = h\}$, where $h^r(j,n)$ is the visit vector associated with the nth j-triage patient. We can write

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

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

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

The same reason as in Reiman (1984) then implies that the latter expression vanishes, as $r \to \infty$. This establishes (EC.56).

Combining (EC.55), (EC.56) with (EC.54), now yields (EC.57). □

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

EC.10. Outline of the proof for Proposition 5: Waiting Time Cost

We only give an outline for the proof of a lower bound, which is similar to Theorem 1. One can prove that the family of modified policies $\{\tilde{\pi}^r_*\}$ reaches the lower bound following the discussion in §EC.4, especially, there are similar state-space collapse results. As a result, $\{\tilde{\pi}^r_*\}$ is asymptotically optimal.

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

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

Similarly to the discussion in proving Proposition 6 of van Mieghem (1995), we can prove that the following is true in stochastic sense:

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

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

$$\begin{aligned}\min_x\quad&\sum_{k\in\mathcal{K}}\lambda_k C_k(x_k/\lambda_k)\\ \text{s.t.}\quad&\sum_{k\in\mathcal{K}}m_k^e x_k=a,\\ &x\geq 0.\end{aligned}\tag{EC.59}$$

This yields

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

EC.11. Proof of Proposition 6: Sojourn Time Cost

We first provide an outline for proving an asymptotic lower bound for all asymptotically compliant policies. Note that, when an IP patient transfers to a next stage, the cost accumulates and the cost function does not change. As a result, 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 sojourn cost. Then for any asymptotically compliant family of control policies, one can prove that the family $\{\widehat{Q}^r_\omega\}$ is stochastically bounded, in particular the diffusion scaled queue length processes of IP patients are stochastically bounded. We now restrict our discussion to asymptotically compliant policies, in which the physician can not be idle if there are IP patients. Then one can prove Lemma EC.6.1. Similarly to the discussion in the proof of Proposition 4 in van Mieghem (1995), we can prove that, for any $0\leq a<b\leq T$,

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

Now, following the discussion in proving Proposition 6 of van Mieghem (1995), we can prove that the following is true in stochastic sense:

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

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

$$\begin{aligned}\min_x\quad&\sum_{k\in\mathcal{C}_0}\lambda_k C_k\left(\sum_{j\in\mathcal{C}_k}x_j/\lambda_k\right)\\ \text{s.t.}\quad&\sum_{k\in\mathcal{C}_0}\sum_{k'\in\mathcal{C}_k}m^e_{k'}x_{k'}=a,\\ &x\geq 0.\end{aligned}\tag{EC.60}$$

Then, following a discussion similar to the proof in Theorem 1, we get

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

The fact that the proposed family of control policies $\{\widetilde{\pi}^r_{**}\}$ reaches the lower bound can be proved easily, by showing the corresponding state-space collapse result. Here we just give some structural insights on the optimal solution to the problem (EC.60). For classes in $\mathcal{C}_k$, we know that, if $\sum_{k'\in\mathcal{C}_k}m^e_{k'}x_{k'}$ is fixed, then the solution minimizing $C_k(\sum_{j\in\mathcal{C}_k}x_j/\lambda_k)$ is making x_k non-zero, while all other x_j with $j\in\mathcal{C}_k\backslash\{k\}$ are 0 (this is because $m^e_k>m^e_j$, for all $j\in\mathcal{C}_k\backslash\{k\}$). As a result, if the problem has an optimal solution with some $k'\in\mathcal{C}_k\backslash\{k\}$ for some k, then one can always find a better solution, which is a contradiction. Now the problem is reduced to the following problem:

$$\begin{aligned}\min_x\quad&\sum_{k\in\mathcal{C}_0}\lambda_k C_k\left(x_k/\lambda_k\right)\\ \text{s.t.}\quad&\sum_{k\in\mathcal{C}_0}m^e_k x_k=a,\\ &x\geq 0,\end{aligned}\tag{EC.61}$$

Following the discussion in solving (9) (using the KKT conditions), we can define a new function, in analogy to $\widehat{\Delta}_{\mathcal{K}}(\cdot)$ from (EC.59) (but now with subscript $\mathcal{C}_0$), and under $\{\tilde{\pi}^r_{**}\}$, this function plays the role of a lifting mapping in state-space collapse results.

EC.12. Incomplete information

We consider a two phase problem as outlined in §7. 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 means service times at both phases are 1, and the fraction of patients continuing to the second phase is p.

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, when the 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:

$$\widetilde{\mathcal{S}}^r(t)=\int_0^t\left(\widehat{\tau}^r_{11}(s)\right)^2d\bar{\bar{E}}^r_1(s)+\int_0^t\left(\widehat{\tau}^r_{12}(s)+\widehat{\tau}^r_2(s)\right)^2d\bar{\bar{E}}^r_2(s),\tag{EC.62}$$

in which $\tau^r_{11}(s)$ represents the waiting time of a patient arriving at time epoch s and will go through only phase 1, $\tau^r_{12}(s)$ represents the waiting time in phase 1 of a patient arriving at time epoch s and going through both phases, and τ^r_2 represents the waiting time in phase 2 of that patient; E^r_1 is the arrival process for patients with 1 visit only, and E^r_2 is the arrival process for patients with 2 phases.

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

$$\begin{aligned}\lim_{r\to\infty}\widetilde{\mathcal{S}}^r(t)\geq&(1-p)\int_0^t\left(\widetilde{\Delta}_1\left(\widetilde{Q}_w(s)-\widehat{\omega}\right)^+\right)^2ds\\&+p\int_0^t\left(\widetilde{\Delta}_1\left(\widetilde{Q}_w(s)-\widehat{\omega}\right)^+/p+\widetilde{\Delta}_2\left(\widetilde{Q}_w(s)-\widehat{\omega}\right)^+/p\right)^2ds,\end{aligned}\tag{EC.63}$$

where $(\widetilde{\Delta}_1(a),\widetilde{\Delta}_2(a))$ is the solution to the following optimization problem:

$$\begin{aligned}\min_x\ &(1-p)x_1^2+p(x_1+x_2/p)^2\\\text{s.t.}\ &(1+p)x_1+x_2=a,\\&x_1,x_2\geq 0.\end{aligned}\tag{EC.64}$$

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.

(Note that, there is some secretly trick in getting the lower bound above. Following the discussion in van Mieghem (1995), one can only prove that

$$\frac{1}{\bar{\bar{E}}_1^r(b)+\bar{\bar{E}}_2^r(b)-\bar{\bar{E}}_1^r(a)-\bar{\bar{E}}_2^r(a)}\left(\int_a^b\widehat{\tau}_{11}^r(s)d\bar{\bar{E}}_1^r(s)+\int_a^b\widehat{\tau}_{12}^r(s)d\bar{\bar{E}}_2^r(s)-\int_a^b\widetilde{Q}_1^r(s)ds\right)\Rightarrow 0.\tag{EC.65}$$

Here $\widetilde{Q}_1(t)$ is the queue length of those patients in the first phase at time t. But this is not enough for proving (EC.63). Indeed, the service discipline in the first phase is FCFS. One can thus expect that

$$\begin{aligned}&\frac{1}{\bar{\bar{E}}_1^r(b)-\bar{\bar{E}}_1^r(a)}\left(\int_a^b\widehat{\tau}_{11}^r(s)d\bar{\bar{E}}_1^r(s)\right)=\frac{1}{\bar{\bar{E}}_2^r(b)-\bar{\bar{E}}_2^r(a)}\left(\int_a^b\widehat{\tau}_{12}^r(s)d\bar{\bar{E}}_2^r(s)\right)\\=&\frac{1}{\bar{\bar{E}}_1^r(b)+\bar{\bar{E}}_2^r(b)-\bar{\bar{E}}_1^r(a)-\bar{\bar{E}}_2^r(a)}\left(\int_a^b\widehat{\tau}_{11}^r(s)d\bar{\bar{E}}_1^r(s)+\int_a^b\widehat{\tau}_{12}^r(s)d\bar{\bar{E}}_2^r(s)\right).\end{aligned}\tag{EC.66}$$

By using (EC.66), together with (EC.65), then following the discussion in van Mieghem (1995), we deduce (EC.63).)

EC.13. Discussion for the conjecture in §8.1: Adding delays after service

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

$$y^n(t)=x^n(t)-\mu^n\int_0^t y^n(s)ds;\tag{EC.67}$$

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.

We use $Q_{jk}^r(t)$ to denote the number of patients in the delayed system between j-triage and k-IP patients at time t, and $Q_{kl}^r(t)$ 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

$$\begin{aligned} Q_k^r(t) &= Q_k^r(0)+\sum_{j\in\mathcal{J}}\left(\Phi_{jk}\left(S_j\left(T_j^r(t)\right)\right)+Q_{jk}^r(0)-Q_{jk}^r(t)\right) \\ &\quad +\sum_{l\in\mathcal{K}}\left(\Phi_{lk}\left(S_l\left(T_l^r(t)\right)\right)+Q_{lk}^r(0)-Q_{lk}^r(t)\right)-S_k\left(T_k^r(t)\right) \\ &= Q_k^r(0)+\sum_{j\in\mathcal{J}}\Phi_{jk}^r\left(S_j\left(T_j^r(t)\right)\right)+\sum_{l\in\mathcal{K}}\Phi_{lk}^r\left(S_l\left(T_l^r(t)\right)\right)-S_k\left(T_k^r(t)\right) \\ &\quad -\sum_{j\in\mathcal{J}}\left(Q_{jk}^r(t)-Q_{jk}^r(0)\right)-\sum_{l\in\mathcal{K}}\left(Q_{lk}^r(t)-Q_{lk}^r(0)\right), \qquad k\in\mathcal{K}. \end{aligned} \tag{EC.68}$$

If we ignore the changes of $T_j^r, j\in\mathcal{J}$ and $T_k^r, k\in\mathcal{K}$, then the difference between (EC.68) and (EC.11) is $\sum_{j\in\mathcal{J}}\left(Q_{jk}^r(t)-Q_{jk}^r(0)\right)+\sum_{l\in\mathcal{K}}\left(Q_{lk}^r(t)-Q_{lk}^r(0)\right)$, which is the total change in the numbers of patients within the infinite-server queues that would delays between services. As a result, we first describe an analysis for infinite-server queues.

Consider a sequence of infinite-server queueing systems $G/M/\infty$. In the rth system, the arrival process is $E^r(\cdot)$, with individual service rate $\mu^r=\mu r^\alpha$, in which $\alpha>-2$. Assume that the fluid scaled arrival processes $\bar{\bar{E}}^r$ are tight. Here

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

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

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

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

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

Using (EC.67), if $\bar{\bar{X}}^r(0)\Rightarrow 0$, then one can show that $\bar{\bar{X}}^r(\sigma^r\wedge\cdot)\Rightarrow 0$. And following the discussion in proving (39) in Atar and Solomon (2011), we can also prove $\sigma^r\Rightarrow T$. As a result, $\bar{\bar{X}}^r\Rightarrow 0$ on $[0,T]$. As this T is arbitrary, we have $\bar{\bar{X}}^r\Rightarrow 0$ on $[0,\infty)$.

Now return to our queueing systems with delays. Note that the arrival processes for the infinite-server queueing systems are parts of the departure processes from the physician. We can then easily verify that the requirements for the analysis of the above $G/M/\infty$ hold, in particular the sequence of the fluid scaled arrival processes is tight. As a result, the $G/M/\infty$ system will not change in fluid scaling, meaning that the delays will have no impact on the fluid limit of the ED model. (For a rigorous discussion, we can first argue that the fluid limit of $\sum_{j\in\mathcal{J}}m_j^e\bar{\bar{Q}}_j^r+\sum_{k\in\mathcal{K}}m_k^e\bar{\bar{Q}}_k^r$ will not change, and then follow the steps in §EC.2 to prove that the fluid limit for the busy time processes do not change, namely they are $\lambda_j m_j t$ for $j\in\mathcal{J}$ and $\lambda_k m_k t$ for $k\in\mathcal{K}$.)

Finally we discuss the diffusion scaled processes. From the differences between (EC.68) and (EC.11), to prove that $\sum_{j\in\mathcal{J}} m_j^e \widehat{Q}_j^r + \sum_{k\in\mathcal{K}} m_k^e \widehat{Q}_k^r$ is invariant to all work-conserving policies, it is enough to argue that the following is true for each $k \in \mathcal{K}$:

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

This again brings us to the analysis of $G/M/\infty$ systems. Now for a sequence of $G/M/\infty$ systems, fix a sequence of $\{\lambda^r\}$, and denote $\widehat{X}^r(t) = r^{-1}(X^r(r^2t) - \lambda^r/\mu^r)$ as well as

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

We then have

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

Suppose that there is a sequence of $\{\lambda^r\}$ with (i) $\lambda^r \to \lambda$ for some $\lambda > 0$, (ii) $\widehat{X}^r(0) \Rightarrow 0$, and (iii) making $\{\widehat{E}^r\}$ tight. Then from the fluid limit argument, we can prove that $\widehat{S}^r\left(\mu r^{2+\alpha}\int_0^t \bar{\bar{X}}^r(s)ds\right)$ converge to a driftless Brownian motion with variance λ; using (EC.67), we can now deduce that $\widehat{X}^r(\cdot) \Rightarrow 0$.

Finally, return to the queueing systems with delays. From the above discussion, 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 our future research.

References

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

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

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

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

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

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