PATIENT FLOW MANAGEMENT IN EMERGENCY DEPARTMENTS

JUNFEI HUANG

NATIONAL UNIVERSITY OF SINGAPORE

2013

PATIENT FLOW MANAGEMENT IN EMERGENCY DEPARTMENTS

JUNFEI HUANG

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF DECISION SCIENCES

NATIONAL UNIVERSITY OF SINGAPORE

2013

ACKNOWLEDGMENT

In my opinion, the most enjoyable thing in writing a thesis is to thank those people who have helped and inspired me during my PhD study.

First and foremost, I would like to express my sincerest gratitude to Professors Itai Gurvich, Avi Mandelbaum and Hanqin Zhang, for their innumerous encouragement, help, supervision and confidence in me. The experience of working with them in different places, from different perspectives, not only improves my ability to do research, but also broadens my view on life. Their knowledge, experience, insights and rigorous attitude, besides many other things, are great gifts to me, from which I will benefit for my whole life.

My special thanks go to Professors James Ang, Jim Dai and Melvyn Sim. Without them, my study at NUS would not be possible. I am grateful for their continuous encouragement, support and guidance. I also want to thank Jiheng Zhang for many discussions and guidance, which are very enjoyable and will definitely benefit me in the future.

I would like to thank my thesis committee members, Professors Shuangchi He, Jeannette Song, Heng-Qing Ye and Xueming Yuan, for their invaluable suggestions.

I would like to thank those people at Technion, especially Professors

Rami Atar, Haya Kaspi, Nahum Shimkin, Nitzan Yuvilar and the people in SEELab, for their help, support and courses.

I am grateful to Professors Mor Armony and Tolga Tezcan for their kindly support and encouragement during my job-hunting.

I also want to thank those professors who have taught me at NUS, among them are: Lucy Chen, Fee Seng Chou, Mabel Chou, Jie Sun, Chung-Piaw Teo, Tong Wang and Yaozhong Wu.

I would also like to thank Chwee Ming Lee, Cheow Loo Lim, Hamidah Bte Rabu and Dorothy Tan for their assistance during my study at NUS.

The wonderful times and life-long memories at NUS should be attributed to those friends at NUS, among them are: Qingxia Kong, Vinit Kumar, Lijian Lu, Zhuoyu Long, Jin Qi, Nishant Rohit, Li Xiao, Yunchao Xu, Dacheng Yao, Xuchun Yuan, Meilin Zhang, Su Zhang, Zhichao Zheng and Yuanguang Zhong.

I would like to thank the friends at Northwestern University, among them are: Ruomeng Cui, Qun Li, Xu Liu, Xiaoshan Peng and Mu Zhao.

During my PhD study, I have been supported by several universities and institutes: NUS, Technion, Kellogg and SAMSI. I appreciate their support.

I would like to thank my parents and my brother, for their unconditional love and support. I would like to dedicate this thesis to them.

Singapore Junfei Huang

May, 2013

ABSTRACT

In this thesis, we consider the control of patient flow through physicians in emergency departments (EDs), which have attracted many researchers' attention. Our work here seems to be the first model to quantitatively analyze the control of patient flow in an emergency department from a queueing theory perspective.

Problem: In emergency departments, the physicians must choose between catering to patients right after triage, who are yet to be checked, and those that are work-in-process (WIP), who are occasionally returning to be checked. The service requirements for the two kinds of patients are different: for the patients right after triage, they must see a doctor within targeted time windows (that may depend on the patients' severity and other parameters); while the WIP patients, on the other hand, impose congestion costs. The physicians in the emergency departments have to balance between triage and WIP patients so as to minimize costs, while meeting the constraints on the time-till-first-service.

Model: We model this prioritization problem as a queueing system with multi-class customers, combining deadline constraints, feedback and congestion costs together. We consider two types of congestion costs: per individual

visit to a server or cumulative over all visits. The former is the base-model, which paves the way for the latter (more ED-realistic) one.

Method: The method we use is conventional heavy-traffic analysis in queueing theory, based on the empirical evidence that the emergency departments can be viewed as critically-loaded stationary systems between late morning till late evening. We propose and analyze scheduling policies that asymptotically minimize congestion costs while adhering to all deadline constraints.

Solution: The policies have two parts: the first chooses between triage and WIP patients using a simple threshold policy; assuming triage patients are chosen, the physicians serve the one with the largest delay relative to deadline; alternatively, WIP patients are served according to some generalized $c\mu$ policy, in which μ is simply modified to account for feedbacks. The policies that we propose are easy to implement and, from an implementation point of view, has the appealing property that all information required is indeed typically available in emergency departments. For the proposed policies, asymptotic optimality, as well as some congestion laws that support forecasting of waiting and sojourn times, are established.

Application: 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. This is an illustration on how our recommendations can improve the operational efficiency and service quality.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Very few things can be more important than health care to our lives. Health care is such a holistic topic that each part of it deserves a lot of efforts for understanding. In this thesis, we will focus on hospitals, especially, the "gate" of hospitals – emergency departments. Emergency departments are *service* systems, which are so crucial in the sense that the quality of service there is closely related to people's lives. Emergency departments without good quality of service (long delays) can result into unnecessary death ([28]).

It is the job of physicians to provide good treatments to patients, while it is the system managers' job to manage the emergency departments well and eliminate unnecessary delay of treatments. The problem of long delays in emergency departments has been observed in many places, and has attracted system managers' attentions.

In 2011, the Ministry of Health (MOH) of China proposed to use a triage system to manage the emergency departments. The objective is to improve the quality of care (the safety of patients). By using this triage system, patients are classified into 4 classes according to their severities. Patients will be scheduled to see the physicians according their severities, and a natural problem after this triage system is, what is the best scheduling, so that physicians can provide the best quality of care to the patients.

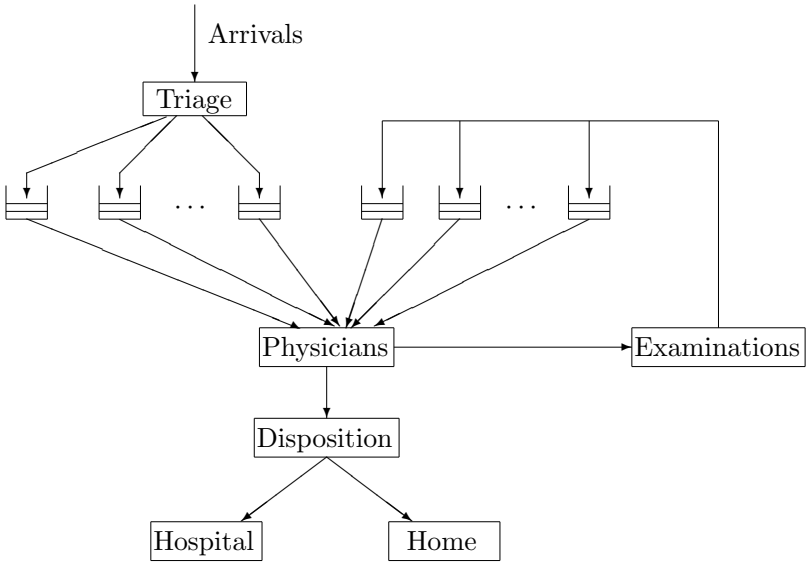
In developed countries such as the United States, triage systems are widely used in emergency departments, but similar scheduling problem still exists. From 1991 to 2009 in the United States, the number of emergency departments decreased by 10%, while the number of visits to the emergency departments increased by more than 20% ([22]). As a result, the ED environment in the United States has become more crowded. Indeed, from 2003 to 2009, the waiting time in most of the emergence departments increased by 25% (from 46.5 to 58.1 mins, see [21]).

All these show the needs to manage emergency departments, that is, to improve the performance of emergency departments. To do this, it is necessary to understand the operations in the emergency departments. As it has been observed, 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 ([2, 34, 33]). This is also true in emergency departments. This brings the research problem of patient flow management in emergency departments.

Here is how patients go through the emergency departments: a new patient arriving at an emergency department is first triaged by a nurse, then waits in the waiting area for the first examination by the physicians; after the first examination, the patient may leave the emergency department directly, or go to the other parts of the emergency department to do further examinations, such as doing CT scan or blood test; after getting the report of the examination, the patient returns to the physicians to get guidelines for further treatments; the patient may have to do several examinations to

complete the treatments; after all treatments are done, the patient leaves the emergency department, either by being admitted to the hospital, or discharged to go home. This is illustrated in the following:

Fig. 1.1: Patient flow in emergency departments



There are two kinds of patients in an emergency department: one is new patients arriving from the outside of the system, the other is the work-in-process (WIP) patients who have stayed in the system for a while and have received some treatments. Different from the new-arriving patients, there is no external arrivals for those WIP patients and they are all transferred from new-arriving or other WIP patients. In addition, each patient may visit the physicians for several times, this corresponds to the several examinations took place in the emergency department, as described above. The following table is calculated using the numbers from Table 2 of [44], in which the authors did empirical analysis with data from an emergency department in Israel.

The patients in this emergency department are classified into 7 classes, and the physicians are also classified into 4 classes, according to their expertise. From this table, it can be seen that, each patient visits the physician for at least 3 times.

Tab. 1.1: Number of visits in an Israeli emergency department

Physician type	Patient type	Average number of visits
1	1, 7	3.9698
2	2, 5	2.9904
3	3, 6	2.9700
4	4	2.9904

The service requirements for those new and WIP patients are different:

- New patients: when arriving at the emergency department, the new patients are generally classified into different classes via a pre-specified triage system, for example, Canadian Emergency Department Triage & Acuity Scale (CTAS, [7]). The triage system classifies the patients into different classes according to the severity of those patients (by using emergency severity index, ESI), and puts requirements on the Time-Till-First-Examination (TTFE) for patients in different classes – that is, a patient must start the first examination in some pre-specified time-window. For example, in the CTAS, patients are classified into 5 classes according to the clinical conditions, and the corresponding deadlines are as follows:

Tab. 1.2: Deadlines specified in CTAS

Severity	Resuscitation	Emergent	Urgent	Less urgent	Non-urgent
Deadlines	Immediate	15 mins	30 mins	60 mins	120 mins

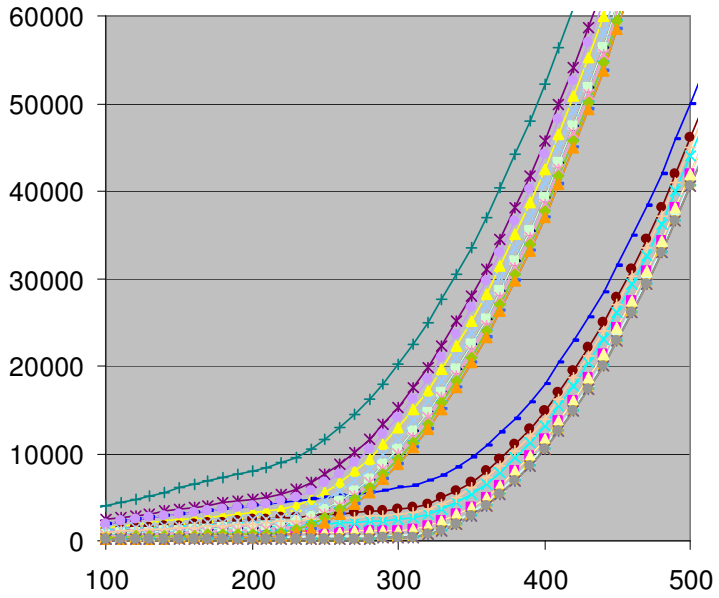
Patients in level 1 (Resuscitation) are with very serious conditions, and they must start their first examination immediately. Generally those patients are treated in a separate area so those patients are not considered in this thesis. For the patients in the other four levels, they may wait for a while, but with different deadlines on waiting times. For example, a patient whose health condition is identified as "Urgent" can wait for as long as 30 minutes, which is the length of the safe period for this patient before receiving the first treatment.

- WIP patients: work-in-process patients use the resources in the emergency department, and bring congestions to the emergency department. The directors of the emergency department use cost which is called congestion cost to measure the congestion incurred by those WIP patients – the congestion cost can represent several costs, such as waiting costs, clinical costs, emotional costs, psychological costs and others. One example of what congestion cost can measure is the impact of long waiting time: if a patient waits in the emergency department for a long time, then that patient may face high risk of suffering additional disease (either caused by the existing disease or infected from other patients – the emergency department is indeed not a safe place to stay.)

  However, how to identify the cost functions by WIP patients are generally difficult. Here is an example of how the director in an Israeli

emergency department do. In that Israeli emergency department, the director identified the cost functions by the patients' triage class, age and the decisions after treatments, see the following Fig. 1.2 (cited from [10]):

Fig. 1.2: Cost functions in an Israeli emergency department



Generally, a patient with more serious condition needs more care from the emergency department, and will bring more congestion, hence will result more cost to the emergency department. The patients are classified into different age groups, and the cost functions for patients in different age groups are different, for example, the cost from the patients over 75 is higher than patients from other groups. Also the decisions after treatments have impact on the cost functions: for those patients who will be discharged to go home, their cost is twice as much

as the cost by those patients who will be admitted to the hospitals (as it can be seen from the figure, the cost functions are grouped into two groups. The functions in the upper group correspond to the patients who will be admitted to the hospitals, while the functions in the lower group correspond to the patients who will be discharged to go home).

This system has attracted many researchers' attention. However, there are several complexities in analyzing emergency departments. As described above, there are several classes of new patients, as well as several classes of WIP patients. The service requirements for different classes of patients are different. An optimal scheduling policy should not be a (simple) static one: if the priority is always given to the new patients, then the queue lengths of the WIP patients will be long – this makes the emergency department blocked and all patients have to experience long sojourn times; on the other hand, if the priority is always given to the WIP patients, the waiting time for the new patients will increase – this is dangerous for the new patients with serious health conditions.

Till now, there are hundreds of simulation-based studies for emergency departments. Simulation models are useful to compare different policies and get insights on the importance of features, however, they are not suitable to find the optimal policy among all reasonable policies. As a result, analytical models are needed. However, because of the complexities mentioned above, building analytical models is indeed a challenging problem ([40]). In this thesis, we will build two analytical models, and propose the corresponding asymptotically optimal policies to manage the emergency departments.

The complexities, as well as the challenge to manage patient flow, stem

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 ([18, 29]). Second, patient flow has a significant feedback component that must be accounted for: WIP patients possibly return several times to physicians during their stay in the emergency department, before ultimately being either released or hospitalized, see Tab. 1.1. Another challenge is that the new and the WIP patients have different service requirements: the service requirements for those WIP patients is to minimize the *congestion cost* (recall that the service requirements for triage patients is to meet pre-specified deadlines).

In a summary, WIP 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-and-clinical friction that makes the problem interesting and complicated, from the viewpoint of the physician: when becoming idle, what class should be served next - triage or WIP - after which one must decide on the specific patient to be examined.

In this thesis, we will model and analyze the emergency department by using queueing models. We will ignore some unimportant features to make the models tractable. Other possible important features will be discussed in Chapter 4. The first feature we ignore is the time for the examination steps. This means that feedbacks in the model are immediate. If the times for the examination steps are very short, then this is a reasonable assumption. We will give a conjecture in §4.1 on how long can these examination steps be. Another feature is the triage-steps. This thesis will not focus on any

triage system here, but only considers the steps after the triage stage. It is worth to mention that, though triage systems are not the main subject of this thesis, the results in this thesis can give us suggestions on how to design triage systems (for example, the case study in subsection 3.2). Finally, the decisions after the treatments (whether to be discharged to go home or be accepted by hospital)are not the main focus here either, though they can also be used to help managers improve the operations of the emergency departments, as shown in §3.2.

Instead, this thesis will focus on the features which we regard as the most important ones in emergency departments, they are: feedback, deadlines on the time-till-first-treatment, and congestion cost incurred by those work-in-process patients. As a reminder, the deadline constraints are clinical constraints while the congestion costs are operational consideration. The problem faced by the director is how to balance between the clinical and operational considerations. To this end, we model and analyze the emergency departments by using queueing theory, and propose flow control policies that minimize congestion costs while subject to deadline constraints.

In this thesis, we consider two models, which differ by their congestion costs: the first is a basic (queue length) model in which the congestion costs are incurred per individual doctor visits; in the second, congestion costs accumulate over all visits during patient sojourn-times. The mathematical framework used here 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 emergency departments are inherently time-varying. Specifically, empirical evidence suggests that, during regular

peak shifts between late morning till late evening, the emergency departments can be usefully viewed as critically-loaded stationary systems ([2]). Within this asymptotic framework, the work-in-process analysis follows the generalized cμ-rule of [41], after generalizing it to models with feedback. The triage analysis combines the due-date scheduling in [42] with the formulation of [35]. The latter offers a rigorous meaning for adherence to (triage) time-constraints, by introducing "asymptotic compliance" as a relaxation for "feasibility". Together, triage and work-in-process controls yield what we prove to be asymptotically optimal flow-control policies: they minimize WIP congestions costs subject to triage compliance.

The proposed policies for both models have the same structure: they are two-stage policies. At the first stage, the physicians first determine the priority between the triage classes and the WIP classes. Then at the second stage, the physicians determine the specific patient to be served next. The details of the policies are slightly different.

In the basic (queue length) model, the proposed policy is as follows: first fix any one triage class, for example, the triage class with index 1. If a physician becomes idle at time t, then:

- At the first stage, use a threshold policy to determine the priority between the triage classes and the WIP classes: the threshold policy only uses the information of triage class with index 1;
- At the second stage:
  - If the physician decides to serve a patient from triage classes, the physician chooses the head-of-the-line patient from the triage class

with the largest relative age;

– If the physician decides to serve a patient from WIP classes, the physician uses a policy which is similar to the generalized $c\mu$ rule ([41]), and we call this policy the modified generalized $c\mu$ rule.

For the second (sojourn time) model, the system first classifies the WIP patients into starting classes and subsequent classes. Then the physicians use the following guidelines: the threshold policy between triage and WIP patients, as well as the policy part to determine the priorities among triage patients, do not change; while if a physician decides to serve a patient from the WIP classes, the physician first serves a patient in the subsequent classes if there is any, then uses a policy which is similar to the generalized cμ rule for the starting classes.

Under the proposed policies, some congestion laws are established. Those congestion laws are based on a principle which is known as the *snapshot principle*, that is, under the heavy traffic scaling, the duration of a patient staying in the emergency department is very short, and the status of the emergency department will not change too much. The congestion laws can help us estimate the waiting time of a new arriving patient, the age of the patient at the head-of-the-line, and the sojourn time of a new arriving patient if the routing vector is known.

Finally, we apply the sojourn time framework, with the expert-elicited sojourn time costs from [10], to support analysis of the value of information in ED flow-control. Specifically, we show that accurate prediction of both the number of visits to a physician and whether a patient will be hospitalized

or discharged, reduces WIP congestion cost by as much as 27%. From our ED sources, and supported by [39, 40], such predictions can be accurately made and, hence, are worth being accounted for.

Literature review and contributions: To the best of our knowledge, this thesis is the first research analyzing the control of patient flow in an emergency department, from a queueing-theory perspective. (As mentioned before, there are hundreds of simulation-based studies; [8].) After starting this project, additional work has appeared on the operations of emergency departments. The closest to this thesis are [39, 40]: [39] discusses a complexity-based triage systems, based on the number of visits that patients pay to the emergency department physician (serving as an up-front proxy for complexity); and [40] analyzes 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 will return to [39, 40] in §3.2, where we analyze the value of the information they require. There are additional papers that cater to specific emergency department characteristics: [44] models the emergency department 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; [17] develops an overloaded queueing network to analyze the impact of interruptions on throughput of emergency department; and [4] addresses synchronization of activities in emergency department (e.g. interpretations of a blood-test and x-ray imaging must precede a visit to the

physician), by analyzing a fork-join queueing network in heavy-traffic.

The models and analysis in this thesis follow two main lines of research: formulation of the triage constraints is adapted from [35], which analyzes admission control; and our IP control generalizes [41], which solves a cost minimization problem for a multi-class queue without feedback. The results in [41] have been generalized by [31] to a feedforward network of parallel queues, and both papers establish asymptotic optimality of the generalized $c\mu$-rule. Here we generalize [41] 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.

The model structure for IP patients here resembles [25, 26], where the author considers a dynamic scheduling problem of a multiclass $M/GI/1$ queueing system with Markovian feedback. Unlike [25, 26], which minimize a cost function that is linear in average queue lengths and proves the optimality of a static routing policy (and the model is known as Klimov's model), 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, the analysis of WIP patients here in fact can be applied to cover Klimov's model: 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 [12], 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 [15, 38], restricting to a global FCFS service discipline among all classes. The analysis here can be also adapted to the FCFS discipline, as well as to other work-conserving disciplines. Indeed, we prove the convergence of a weighted queue length to a reflected Brownian motion, under any work-conserving policy (Proposition 2.4.1), in which the global FCFS policy is a special case. Proving convergence of individual queue lengths, for each class, amounts to establishing state-space collapse, which will follow from standard arguments (e.g. [9]).

The main contributions of this thesis can be summarized as follows:

- **Methodological:** We analysis multiclass queueing systems with feedback, particularly,
    1. Proving the conjecture in [31] 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. The models here capture the tradeoff between catering to triage- vs. WIP-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.

Structure of the thesis: This thesis is organized as follows. A basic ED model is introduced in Chapter 2: a detailed description of the model, policy and insights is given in §2.1; heavy traffic conditions, asymptotic compliance and optimality are introduced in §2.2 and §2.3, respectively. The main results and some auxiliary propositions and extensions are presented in §2.4, with their discussions in §2.5. The proofs for the main theorems are in §2.6, and the proofs for propositions and complements are provided in §2.7. Our alternative ED model, with sojourn time costs, is discussed in §3.1, with an application in §3.2, using data from an Israeli ED, and expert-elicited costs. The technical discussions are in §3.4-3.5. We conclude with a discussion of future research directions in Chapter 4.

Notation: Firstly, from now on, we will use abbreviation "ED" for "emergency department". 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-dimensional and K-dimensional nonnegative orthant, 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. The notation $\{e_k\}$ is reserved 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 any given 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. A BASIC MODEL

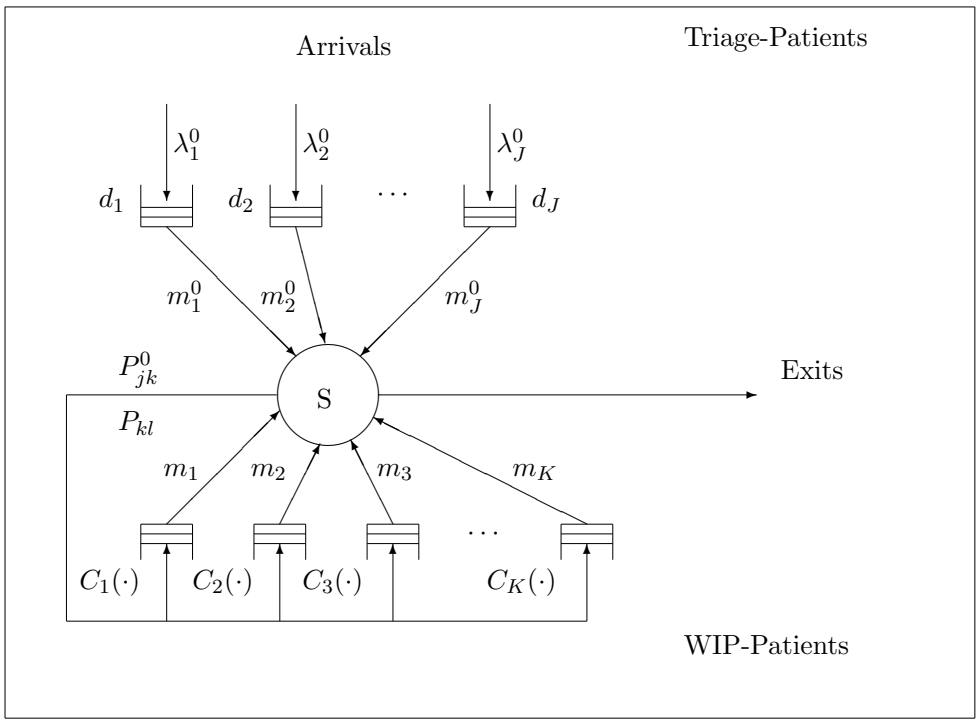
This section will build a basic quantitative model for emergency departments (EDs). The congestion cost in this model is based on the queue length of each class, thus the model here is also called as *queue length model*.

2.1 The model, policy and intuitions

In this basic model, ED dynamics is captured by a multiclass queueing system, with S servers (physicians), J classes of *triage* patients and K classes of *work-in-process* (WIP) patients. Triage patients are yet to be examined by a physician, and work-in-process (WIP) patients require further treatment. (A patient class could embody information such as treatment type, emergency level or age; see [10].) Triage customers subject to deadline constraints, while WIP customers incur queueing costs. To highlight the application to EDs, "patient" is used interchangeably with "customer" and "physician" with "server". Let $\mathcal{J}$ and $\mathcal{K}$ denote the index sets of triage and WIP patients, respectively: $j \in \mathcal{J}$ is an index for triage patients, and $l, k \in \mathcal{K}$ are indices for WIP 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 $1, 2, \ldots, J$ in $\mathcal{J}$ differ from those in $\mathcal{K}$.

The system is depicted in the following Fig. 2.1. In this figure, we use superscript 0 for the terms of the triage patients. For example, we denote by λ_1^0 and m_1^0 the arrival rate and the mean service time of class 1 triage patients, while λ_1 and m_1 the arrival rate and the mean service time of class 1 WIP patients.

Fig. 2.1: Patient flow in emergency department (queue cost)



From now on, we will not use the specific index number and instead, will use $j \in \mathcal{J}$ and $k, l \in \mathcal{K}$ to represent the indices of triage and WIP patients. For example, we use $\lambda_j^0, j \in \mathcal{J}$ to denote the arrival rates of triage patients. As index $j \in \mathcal{J}$ suffices for their characterization, we shall omit the superscript 0, that is, we will use $\lambda_j, m_j, j \in \mathcal{J}$ to denote the arrival rates

and mean service times for triage patients.

Triage patients: For each triage patient class $j \in \mathcal{J}$, there are 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. Assume $\mathbb{E}[u_j(1)] = 1$, $\mathbb{E}[v_j(1)] = 1$ and denote $a_j^2 = \text{var}(u_j(1))$, $b_j^2 = \text{var}(v_j(1))$. Among j-triage patients, the interarrival time between the $(i-1)$st and ith arrivals is $u_j(i)/\lambda_j$ and the service time required for the ith patient is $m_j v_j(i)$. As a result, λ_j is the arrival rate and m_j is the mean service time requirement of a j-triage patient. Assume $\lambda_j > 0$ for all $j \in \mathcal{J}$ 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-WIP patients, with probability P_{jk} (again in the figure, we

denote it by P^0_{jk}), or leave the system directly, with probability $1-\sum_{k\in\mathcal{K}} P_{jk}$. Let the matrix $P_{\mathcal{JK}} = (P_{jk})_{J\times K}$ be the triage-to-WIP matrix. 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-WIP patients if $\phi_j(n) = e_k$ (recall that $\{e_k\}$ is reserved for the standard basis of $\mathbb{R}^K$), 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 $\mathbb{P}(\phi_j(n) = 0) = 1-\sum_{k\in\mathcal{K}} P_{jk}$. 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.

WIP patients: For WIP classes, there are no external arrivals. All WIP patients are transferred from either triage or WIP patients. Denote the number of k-WIP arrivals till time t by $E_k(t)$. Just like triage patients, for each class $k \in \mathcal{K}$, there are a sequence of random variables $\{v_k(i), i = 1, 2, \ldots\}$ and a real number m_k. Assume $\mathbb{E}[v_k(1)] = 1$ and denote $b_k^2 = \mathrm{var}(v_k(1))$. Among k-WIP patients, the service time required for the ith patient receiving service is $m_k v_k(i)$. When we discuss queue lengths, service order among each WIP class will not affect the result, thus we do not assume it to be FCFS. When there is a need (such as when discussing waiting times and sojourn times), we will put the FCFS discipline explicitly. Then, m_k is the mean service time requirement of a k-WIP 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-WIP patients. Denote $\mu_k = 1/m_k$; then this is the *service rate* for k-WIP patients.

After completing service, an l-WIP patient will join the queue of k-WIP 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 WIP-to-WIP 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-WIP patient will transfer to; that is, the nth l-WIP patient finishing service will go to the queue of k-WIP 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 $\mathbb{P}(\phi_l(n) = 0) = 1 - \sum_{k\in\mathcal{K}} P_{lk}$. 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-WIP patients.

Assume that all the arrivals of triage classes, services and transitions of all triage and WIP classes, are mutually independent. This assumption is not necessary for the proofs, but it simplifies calculations and saves the notation (as in [35]). (Practically, arrivals of triage classes can be correlated

with service times of triage and WIP classes, as in [15].)

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-WIP patients, through the following equation:

$$\Lambda^T = (\Lambda_{\mathcal{J}})^T P_{\mathcal{J}\mathcal{K}} + \Lambda^T P. \tag{2.1}$$

Then

$$\Lambda^T = (\Lambda_{\mathcal{J}})^T P_{\mathcal{J}\mathcal{K}}(I-P)^{-1}. \tag{2.2}$$

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

$$M^e_{\mathcal{J}} = M_{\mathcal{J}} + P_{\mathcal{J}\mathcal{K}}(I-P)^{-1}M, \tag{2.3}$$

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

$$M^e = (I-P)^{-1}M, \tag{2.4}$$

in which m^e_k is called the *effective mean service time* of k-WIP patients. Then (2.3) can be written as

$$M^e_{\mathcal{J}} = M_{\mathcal{J}} + P_{\mathcal{J}\mathcal{K}}M^e. \tag{2.5}$$

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.

An infeasible problem: Service goals for triage and WIP 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 $t \geq 0$.

- **WIP patients incurring costs:** Denote by $Q_k(t)$ the number of k-WIP patients in the system at time t. Those k-WIP 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-WIP patients during the first t time units. Then the objective is to solve the following optimization problem for any $T \geq 0$,

$$\begin{aligned} \min_{\Pi} \quad & \int_0^T \sum_{k\in\mathcal{K}} C_k(Q_k(s))ds \\ \text{s.t.} \quad & \tau_j(t) \leq d_j, \quad \forall j \in \mathcal{J} \quad \text{and} \quad 0 \leq t \leq T. \end{aligned} \tag{2.6}$$

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

The problem above is clearly infeasible, as the age processes $\tau_j(\cdot), j \in \mathcal{J}$, are stochastic. The first task of this thesis is to generalize (2.6) 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 (2.6) 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 WIP patients will get long and the cost high; on the other hand, if the physician always gives priority to WIP 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 WIP patients which we describe in the following. 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.

The Proposed policy: 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);
- If a physician decides to serve a patient from triage classes, the physician chooses the head-of-the-line patient from the class with index

$$j \in \text{argmax}_{j \in \mathcal{J}} \frac{\tau_j(t)}{d_j};$$

- If a physician decides to serve a patient from WIP classes, the physician

chooses the head-of-the-line patient from the class with index

$$k \in \text{argmax}_{k\in\mathcal{K}} \frac{C_k'(Q_k(t))}{m_k^e}.$$

Within a suitable heavy traffic framework (Section 2.2), the above policy is asymptotically "feasible" and asymptotically optimal among all asymptotically "feasible" policies. The simplicity of the asymptotically optimal policies, as well as state-space collapse and snap-shot properties that it enjoys (Theorem 2.4.3 and Proposition 2.4.3), are all due to the fact that heavy-traffic analysis exposes macroscopic and mesoscopic essentials, which is formalized by fluid and diffusion approximations (§2.6.4). For example, the S-server system here behaves as one with a super single-server, in which this virtual server is S-times faster than each of the original servers, see for example, [11]; accordingly and without loss of generality, 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 2.5). For example, the decision of triage vs. WIP 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 physician just becoming idle caters to triage patients, otherwise to WIP patients. Furthermore, triage classes can be alternatively prioritized according to shortest-deadline-first, that is, serve $j \in \text{argmin}_{j\in\mathcal{J}}[d_j - \tau_j(t)]$; and the selection cri-

terion of WIP-classes can also be any rule that makes

$$\max_{l,k\in\mathcal{K}} \sup_{0\le t\le T} \left| \frac{C_l'(Q_l(t))}{m_l^e} - \frac{C_k'(Q_k(t))}{m_k^e} \right| \approx 0, \tag{2.7}$$

in particular the one conjectured in page 853 of [31].

Intuition of the policy: The idea is first to maximize service effort for WIP 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 WIP patients to greedily minimize the queueing cost rate. This is a reasonable approach since, in a 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 the stationary case assumed here. An example is a mass-casualty event during which triage patients overload the system; see Section 4 for further discussion.)

The driver of heavy-traffic dynamics is the (total) *workload* in the 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 2.4.1 and (2.41)). 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 the weighted queue lengths of WIP patients $\sum_{k\in\mathcal{K}} m_k^e Q_k(t)$ (which is in concert with minimizing WIP congestion costs) is equivalent to maximizing the weighted age processes of triage patients $\sum_{j\in\mathcal{J}} m_j^e \lambda_j^0 \tau_j(t)$.

Triage vs. WIP patients: 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 a "good" 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 [35] in conventional heavy traffic, and the QIR controls of [20] in the QED regime.) Alternative selection rules could also achieve the desired balance, as described in §2.5.1.

WIP 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 (2.10). The KKT condition now yields the generalized $c\mu$ rule in this thesis, as in [41] but with the μ's replaced by $1/m_k^e$ to account for feedbacks.

The above outline also guides the proofs of the main results, Theorems 2.4.1 and 2.4.2. 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.

Performance analysis: Under the proposed policy, we can also do system's performance analysis. A tool is known as *snapshot principle*.

A Snapshot principle: This is a common feature of heavy traffic ([36]) which, as explained in page 187 of [43] 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 Arrival See Time Average (ASTA), would also prevail temporarily. This snapshot principle then enables predictions of virtual waiting and sojourn times (when the service discipline among each WIP class is FCFS), 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, if we denote by $\tau_k(t)$ the age of the head-of-the-line k-WIP patient at time t, then $Q_k(t) \approx \lambda_k \tau_k(t)$, 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 §2.4.4, which is in the spirit of [23]).

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 (§2.4.5).

2.2 Heavy traffic condition

As noticed before, the problem (2.6) may not have any feasible solution in conventional meaning. In this thesis, we will analyze and solve this problem

in an asymptotical framework, which is known as the *conventional heavy traffic framework*.

Consider a sequence of systems, as discussed in Section 2.1. 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-WIP class is λ_k^r. The deadline for j-triage patients is d_j^r, while the cost function C_k for k-WIP 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 (2.2) and (2.3), 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,

$$\begin{aligned} &\lambda_j^r \to \lambda_j, \quad j \in \mathcal{J}, \quad \text{and} \\ &r(\rho^r - 1) \to \beta, \quad \text{as} \quad r \to \infty, \end{aligned} \tag{2.8}$$

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 (2.2), with $\Lambda_{\mathcal{J}} = (\lambda_j)_{j \in \mathcal{J}}$ in (2.8).

Under condition (2.8), the queue lengths are expected to be $O(r)$, and similarly the ages of head-of-the-line triage patients. Hence, for each $j \in \mathcal{J}$, assume the deadline of j-triage patients satisfies the following convergence:

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

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-WIP patients in the rth system at time t, respectively. Assume that the following initial condition holds:

Assumption 2.2.1: When $r \to \infty$,

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

2.3 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, $\tau^r(\cdot) = \{\tau_j^r(\cdot),\ j \in \mathcal{J}\}$. Define the diffusion scaled age processes through

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

The following concept of "asymptotically compliant" family of control policies, is a generalization of "feasibility" for the optimization problem (2.6).

Definition 2.3.1: A family of policies $\{\pi^r\}$ is said to be *asymptotically compliant* if, for any fixed $T \geq 0$,

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

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

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

Assume that, at time t, k-WIP 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 2.4.1.) Then the cumulative queueing cost is

$$\mathcal{U}^r(t) := \int_0^t \sum_{k \in \mathcal{K}} C_k\left(\widehat{Q}_k^r(s)\right) ds. \tag{2.9}$$

The heavy-traffic adaptation of problem (2.6) is to stochastically minimize $\mathcal{U}^r(t)$, for each t, over all asymptotically compliant families of policies. Formally:

Definition 2.3.2: 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 $x > 0$,

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

here $\{\mathcal{U}^r_*\}$ is the family of cumulative queueing costs defined through (2.9) under the family of control policies $\{\pi^r_*\}$, and $\{\mathcal{U}^r\}$ is the sequence of queueing costs corresponding to any other asymptotically compliant family of policies $\{\pi^r\}$.

2.4 Main results

2.4.1 Cost functions and an optimization problem

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

$$\begin{aligned} \min_{x} \quad & \sum_{k\in\mathcal{K}} C_k(x_k) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{K}} m_k^e x_k = a, \\ & x \geq 0. \end{aligned} \tag{2.10}$$

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 the state-space collapse result; see Theorem 2.4.3.

Assume that the cost functions C_k, $k \in \mathcal{K}$, satisfy the following, in analogy to [41]:

Assumption 2.4.1 (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 (2.10) such that $x_k^* > 0$, $k \in \mathcal{K}$.

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

such that

$$\begin{aligned} C'_k(x^*_k) - \alpha_0 m^e_k &= 0, \\ \sum_{k\in\mathcal{K}} m^e_k x^*_k &= a. \end{aligned} \tag{2.11}$$

It is easy to see that this optimal vector x^* satisfies $C'_l(x^*_l)/m^e_l = C'_k(x^*_k)/m^e_k$, for all $l, k \in \mathcal{K}$. Using this fact, the proof of the following is elementary:

Lemma 2.4.1: The function $\Delta_{\mathcal{K}}(\cdot)$ is well defined, and $\Delta_k(a)$ is nondecreasing in a, for each $k \in \mathcal{K}$.

Proof: From the assumption on cost functions, we can write $x^*_l = C'^{-1}_l\left(\frac{m^e_l}{m^e_k} C'_k(x^*_k)\right)$ for all $l, k \in \mathcal{K}$. Then for any fixed $k \in \mathcal{K}$, from (2.11),

$$\sum_{l\in\mathcal{K}} m^e_l C'^{-1}_l\left(\frac{m^e_l}{m^e_k} C'_k(x^*_k)\right) = a.$$

From Assumption 2.4.1, $\sum_{l\in\mathcal{K}} m^e_l C'^{-1}_l\left(\frac{m^e_l}{m^e_k} C'_k(\cdot)\right)$ is a strictly increasing function, as a result, for any a there will be a unique solution x^*_k, and if a increases, x^*_k will also increase. These prove that Δ_k is well defined and is nondecreasing. □

2.4.2 A lower bound

The first result in this section 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 ([13]), and $\widehat{X}$ is a Brownian motion with drift rate β and variance

$$\sum_{j\in\mathcal{J}}(m^e_j)^2\lambda_j a_j^2 + \sum_{j\in\mathcal{J}}\left(\sum_{k\in\mathcal{K}} m^e_k P_{jk} - m^e_j\right)^2 \lambda_j b_j^2 + \sum_{k\in\mathcal{K}}\left(\sum_{l\in\mathcal{K}} P_{kl} m^e_l - m^e_k\right)^2 \lambda_k b_k^2$$
$$+ \sum_{j\in\mathcal{J}} \lambda_j (M^e)^T \Gamma^j M^e + \sum_{k\in\mathcal{K}} \lambda_k (M^e)^T \Gamma^k M^e. \tag{2.12}$$

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

Theorem 2.4.1 (**Lower Bound**): Fix any asymptotically compliant family of policies, with the corresponding cumulative costs $\mathcal{U}^r$ defined in (2.9). Then for any $t, x > 0$,

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

This theorem is proved in §2.6.2.

2.4.3 The proposed policy and its asymptotic optimality

Now we will modify the proposed policy in Section 2.1 to 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 WIP-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 WIP-type patients.
- If the physician chooses to serve a patient from the triage classes at time t, the physician chooses the head-of-the-line patient from the class with index
$$j \in \text{argmax}_{j\in\mathcal{J}} \frac{\tau_j^r(t)}{d_j^r}. \tag{2.13}$$
- If the physician chooses to serve a patient from WIP classes at time t, the physician uses a policy ensuring (for any $T > 0$)
$$\max_{l,k\in\mathcal{K}} \sup_{0\leq t\leq T} \left| \frac{C_l'(\widehat{Q}_l^r(t))}{m_l^e} - \frac{C_k'(\widehat{Q}_k^r(t))}{m_k^e} \right| \;\Rightarrow\; 0. \tag{2.14}$$
An example of such a policy is to choose $k \in \text{argmax}_{k\in\mathcal{K}} \frac{C_k'(\widehat{Q}_k^r(t))}{m_k^e}$, which is a modified generalized $c\mu$-rule. (More examples of policies ensuring (2.14) can be found in §2.5.2.)

The main result for this basic (queue length) model is the following theorem, which we prove in §2.6.5.

Theorem 2.4.2 (**Asymptotic Optimality**): The family of control policies $\{\pi^r_*\}$ is asymptotically optimal.

In proving Theorem 2.4.2, we will 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 2.4.1 and Theorem 2.4.3 below.

Proposition 2.4.1 indeed holds under any family of work-conserving policies. To state it, define the diffusion scaled queue length processes for triage classes: $\widehat{Q}^r_j(t) = r^{-1}Q^r_j(r^2t)$, $j \in \mathcal{J}$, and diffusion scaled weighted queue length processes

$$\widehat{Q}^r_w(t) = \sum_{j\in\mathcal{J}} m^e_j \widehat{Q}^r_j(t) + \sum_{k\in\mathcal{K}} m^e_k \widehat{Q}^r_k(t). \tag{2.15}$$

Proposition 2.4.1 (**Invariance principle for work-conserving policies**): Under any family of work-conserving policies,

$$\widehat{Q}^r_w \;\Rightarrow\; \widehat{Q}_w, \quad \text{as} \quad r \to \infty. \tag{2.16}$$

This proposition is proved in §2.6.3.

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

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

As in Lemma 2.4.1, one 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 $\widehat{Q}^r = \{\widehat{Q}^r_j,\ j \in \mathcal{J},\ \widehat{Q}^r_k,\ k \in \mathcal{K}\}$ and recall $\widehat{\omega} = \sum_{j\in\mathcal{J}} \lambda_j \widehat{d}_j m_j^e$.

Theorem 2.4.3 (**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

$$\widehat{Q}_j(t) = \Delta_j \min\left(\widehat{Q}_w(t), \widehat{\omega}\right), \quad j \in \mathcal{J}, \tag{2.17}$$
$$\widehat{Q}_k(t) = \Delta_k\left((\widehat{Q}_w(t) - \widehat{\omega})^+\right), \quad k \in \mathcal{K}. \tag{2.18}$$

This theorem is proved in §2.6.4.

2.4.4 *Virtual waiting times*

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

$$\widehat{\omega}_j^r(t) = r^{-1}\omega_j^r(r^2t), \quad j \in \mathcal{J}, \quad \text{and} \quad \widehat{\omega}_k^r(t) = r^{-1}\omega_k^r(r^2t), \quad k \in \mathcal{K}. \quad (2.19)$$

Proposition 2.4.2 (**Asymptotic Sample-Path Little's Law**): Under the family of control policies $\{\pi_*^r\}$, with FCFS service discipline among each WIP patient class, when $r \to \infty$,

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

This proposition is proved in §2.7.2.

Remark 2.4.1: From the convergence of $\widehat{Q}^r$ in Theorem 2.4.3, 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-WIP patient in the rth system, and similarly its diffusion scaling $\widehat{\tau}_k^r(t) = r^{-1}\tau_k^r(r^2t)$, $k \in \mathcal{K}$. The 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 2.4.3 (**Snapshot Principle – Virtual Waiting Time and Age**): Under the family of control policies $\{\pi_*^r\}$, with FCFS among each WIP 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 §2.7.3.

2.4.5 Sojourn times

This section considers sojourn times associated with specific routes through the system, as in [37]. Each patient is associated with a route vector $h \in \mathbb{Z}_+^K$, where h_k denotes the number of times that the patient visits the physician as a k-WIP 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 by $W_{jh}^r(t)$ the *sojourn time* of the next j-triage patient, arriving after t, with route vector h, and the diffusion scaled processes

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

Proposition 2.4.4 (**Snapshot Principle – Sojourn Time and Queue Lengths**): Under the family of control policies $\{\pi_*^r\}$, with FCFS among each WIP 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 §2.7.4.

Remark 2.4.2: From Theorem 2.4.3, 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 2.4.4 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 2.4.2, 2.4.3 and 2.4.4.

Corollary 2.4.1 (**Snapshot Principle – Sojourn Time and Ages**): Under the family of control policies $\{\pi_*^r\}$ with FCFS among each WIP patient class,

if a route vector h is j-feasible, then as $r \to \infty$,

$$\widehat{W}^r_{jh} - \widehat{\tau}^r_j - \sum_{k \in \mathcal{K}} h_k \widehat{\tau}^r_k \;\; \Rightarrow \;\; 0, \quad j \in \mathcal{J}.$$

Remark 2.4.3: 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 [37], the diffusion limit does not depend on the specific order in which the physician is visited.

2.5 Further discussion

2.5.1 Alternative triage policies to (2.13)

The recipe in (2.13), as part of an asymptotically optimal policy, is not unique. From the proof in §2.6.4, it will be seen that any asymptotically compliant family of control policies ensuring

$$\sum_{j \in \mathcal{J}} m^e_j \widehat{Q}^r_j(\cdot) \;\; \Rightarrow \;\; \min\left(\widehat{Q}_w(\cdot), \widehat{\omega}\right), \quad \text{as} \quad r \to \infty, \tag{2.20}$$

is asymptotically optimal (recall that $\widehat{Q}_w$ and $\widehat{\omega}$ are defined in §2.4.2). 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

$$j \in \text{argmax}_{j\in\mathcal{J}} \frac{Q_j^r(t)}{\lambda_j^r d_j^r};$$

the latter can be easily proved asymptotically equivalent to (2.13).

Next 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

$$j \in \text{argmin}_{j\in\mathcal{J}} \left(d_j^r - \tau_j^r(t)\right). \tag{2.21}$$

From Lemma 2.7.2, the above is asymptotically equivalent to choosing the head-of-the-line patient from the class with index

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

Lemma 2.5.1: 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

$\sum_{j\in\mathcal{J}} \lambda_j m_j^e (\widehat{d}_j - \widehat{d}_{j'})^+ \le a < \sum_{j\in\mathcal{J}} \lambda_j m_j^e (\widehat{d}_j - \widehat{d}_{j'+1})^+$, then

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

This lemma will be proved in §2.7.5.

One can now prove that the family of control policies, with (2.21) replacing (2.13), is asymptotically compliant, and satisfies (2.20) – 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 [35] suggests that the policy in (2.13) is a more natural one, as it uses a 'relative' term. As a result, we choose (2.13) for elaboration. The comparison of (2.13) and (2.21) may involve rates of convergence, which is beyond the scope of the present paper.

2.5.2 WIP-Policies that imply (2.14)

For any $K \times K$-dimensional invertible matrix G, with $G_{kk} > 0$, $G_{kk'} < 0$ for $k \neq k' \in \mathcal{K}$ while $\sum_l G_{kl} \ge 0$. We also assume that all terms in G^{-1} are non-negative, and all components of GM^e being positive. Let H denote the K-dimensional vector with the kth component $1/(GM^e)_k$. When the WIP classes are chosen to be served at time t, the physician chooses a patient from the class with index

$$k \in \text{argmax}_{k\in\mathcal{K}} H_k \left(GC' \left(\widehat{Q}^r(t) \right) \right)_k ; \tag{2.22}$$

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

Lemma 2.5.2: For any $T \geq 0$, as $r \to \infty$,

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

for all $l, k \in \mathcal{K}$. As a result, (2.14) holds.

This lemma will proved in §2.7.6.

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

1. $G = I$: then $H_k = 1/m^e_k$; hence (2.22) is

$$k \in \text{argmax}_{k\in\mathcal{K}}\frac{C'_k(\widehat{Q}^r_k(t))}{m^e_k}.$$

This is a generalized $c\mu$ policy, modified from [41] and [31] 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 (2.22) is

$$k \in \text{argmax}_{k\in\mathcal{K}}\left[C'_k\left(\widehat{Q}^r_k(t)\right) - \sum_{l\in\mathcal{K}} P_{kl}C'_l\left(\widehat{Q}^r_l(t)\right)\right]\mu_k.$$

Note that this is the policy conjectured in [31].

The expression in (2.14) is similar to equation (51) in [41], 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 (2.14): when the WIP classes are chosen to be served, the physician chooses a patient from the class with index $k \in \text{argmax}_{k\in\mathcal{K}} C'_k(\widehat{Q}^r_k(t))/m^e_k$, and serves this patient continuously until completing all services – the current one as well as feedbacks. This exhaustive policy is not FCFS within each WIP class. Alternatively, this system can be viewed as a new one with no feedback, but with the service times for k-WIP patients being now the cumulative service requirement – with mean m^e_k. To have this system enjoy asymptotically the queueing-cost lower bound in Theorem 2.4.1, there must exist at least one triage class for each WIP class, such that after the triage service, this class of triage patients will transfer directly to the WIP 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.

2.5.3 Waiting costs

This section considers waiting costs, instead of queueing costs. To this end, assume that the service discipline among each WIP class is FCFS. Recall that $\omega^r_k(t)$ is the virtual waiting time of a k-WIP patient at time t, and its diffusion scaling $\widehat{\omega}^r_k(t)$ is defined in (2.19). Define $\bar{\bar{E}}^r_k(t) = r^{-2}E^r_k(r^2t)$ for

$k \in \mathcal{K}$. One seeks to stochastically minimize the following cost:

$$\widetilde{\mathcal{U}}^r(t) := \sum_{k\in\mathcal{K}} \int_0^t C_k\left(\widehat{\omega}_k^r(s)\right) d\bar{\bar{E}}_k^r(s), \tag{2.23}$$

among all asymptotically compliant families of control policies.

The control policy $\{\pi_*^r\}$ in Section 2.4 is slightly modified as follows. The first step, using a threshold policy to determine between triage classes and WIP classes, and the step using (2.13) to determine priorities among triage patients, do not change. The step determining the priority among WIP classes changes as follows:

- If the physician decides to serve a patient from the WIP classes, 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 $k \in \mathrm{argmax}_{k\in\mathcal{K}} \frac{C_k'\left(\widehat{Q}_k^r(t)/\lambda_k^r\right)}{m_k^e}$. Other examples of policies satisfying the above can be deduced from the policies in §2.5.2: assume G and H are $K \times K$-dimensional invertible matrix and K-dimensional vectors in §2.5.2, the physician chooses a patients from the class with index

$$k \in \mathrm{argmax}_{k\in\mathcal{K}} H_k \left(GC'\left(\widehat{Q}^r(t)/\lambda^r\right)\right)_k,$$

  here $C'\left(\frac{\widehat{Q}^r(t)}{\lambda^r}\right)$ is a K-dimensional column vector with $C_k'\left(\frac{\widehat{Q}_k^r(t)}{\lambda_k^r}\right)$ being

its kth component.

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

Proposition 2.5.1 (**Waiting Time Cost**): The family of control policies $\{\widetilde{\pi}^r_*\}$ is asymptotically compliant. It is also asymptotically optimal among all asymptotically compliant families of work-conserving control policies, in the sense that for any fixed $t > 0$ and $x > 0$,

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

where $\{\widetilde{\mathcal{U}}^r_*\}$ is the family of cumulative cost, defined through (2.23) 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 proof for this proposition can be found in §2.7.7.

2.6 Proofs for theorems

2.6.1 Preliminary analysis

This section starts with an analysis that covers any asymptotically compliant family of control policies.

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

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

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 $r \to \infty$,

$$(\widehat{E}_j^r,\ \widehat{S}_j^r,\ j \in \mathcal{J}) \ \Rightarrow\ (\widehat{E}_j,\ \widehat{S}_j,\ j \in \mathcal{J}); \tag{2.24}$$

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 2.6.1: Under any asymptotically compliant family of control policies, and for all $T \geq 0$,

$$\max_{j\in\mathcal{J}} \sup_{0\leq t\leq T} \left|\widehat{Q}_j^r(t) - \lambda_j \widehat{\tau}_j^r(t)\right| \ \Rightarrow\ 0, \quad \text{as} \quad r \to \infty. \tag{2.25}$$

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

$$\widehat{Q}_j^r(t) - \lambda_j^r \widehat{\tau}_j^r(t) = \widehat{E}_j^r(t) - \widehat{E}_j^r\left((t - \bar{\bar{\tau}}_j^r(t))-\right), \quad j \in \mathcal{J}. \tag{2.26}$$

Here $\bar{\bar{\tau}}_j^r(t) = r^{-2}\tau_j^r(r^2 t)$ is the fluid scaled age process of class j triage patients. 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 (2.24) and (2.8), (2.25) is easily proved from (2.26), 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 2.6.1: Under any asymptotically compliant family of control policies, when $r \to \infty$,

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

Proof: As $(x+y)^+ \le x^+ + y^+$ for any $x, y \in \mathbb{R}$, we have

$$\sup_{0 \le t \le T} \left[\widehat{Q}_j^r(t)/\lambda_j - \widehat{d}_j\right]^+ \;\le\; \sup_{0 \le t \le T} \left[\widehat{Q}_j^r(t)/\lambda_j - \widehat{\tau}_j^r\right]^+ + \sup_{0 \le t \le T} \left[\widehat{\tau}_j^r(t) - \widehat{d}_j\right]^+.$$

From (2.25) and the definition of asymptotic compliance, both terms on the right-hand side of the above equation converge to 0 in probability, as a result, the term on the left-hand side should also converge to 0 in probability. This proves the conclusion. □

Lemma 2.6.2: Under any asymptotically compliant family of control policies, when $r \to \infty$,

$$\bar{\bar{T}}_j^r(\cdot) \Rightarrow \lambda_j m_j e(\cdot), \tag{2.27}$$

$$\widehat{Q}_j^r(\cdot) + \mu_j \widehat{T}_j^r(\cdot) \Rightarrow \widehat{E}_j(\cdot) - \widehat{S}_j\left(\lambda_j m_j e(\cdot)\right). \tag{2.28}$$

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

$$\bar{\bar{Q}}_j^r(t) = \bar{\bar{Q}}_j^r(0) + \bar{\bar{E}}_j^r(t) - \lambda_j^r t - \left[\bar{\bar{S}}_j^r\left(\bar{\bar{T}}_j^r(t)\right) - \mu_j \bar{\bar{T}}_j^r(t)\right] + \mu_j\left[\lambda_j^r m_j t - \bar{\bar{T}}_j^r(t)\right] \tag{2.29}$$

and

$$\widehat{Q}_j^r(t) = \widehat{Q}_j^r(0) + \widehat{E}_j^r(t) - \widehat{S}_j^r(\bar{\bar{T}}_j^r(t)) - \mu_j \widehat{T}_j^r(t). \tag{2.30}$$

From Corollary 2.6.1 and the Functional Law of Large Numbers, for any $T \geq 0$, when $r \to \infty$,

$$\sup_{0\leq t\leq T} \bar{\bar{Q}}_j^r(t) \Rightarrow 0, \qquad \sup_{0\leq t\leq T} \left|\bar{\bar{E}}_j^r(t) - \lambda_j^r t\right| \Rightarrow 0, \tag{2.31}$$

$$\sup_{0\leq t\leq T} \left|\bar{\bar{S}}_j^r\left(\bar{\bar{T}}_j^r(t)\right) - \mu_j \bar{\bar{T}}_j^r(t)\right| \leq \sup_{0\leq t\leq T} \left|\bar{\bar{S}}_j^r(t) - \mu_j t\right| \Rightarrow 0, \tag{2.32}$$

and (2.27) can be easily obtained from (2.29). Then (2.24) and (2.30), together with the Random-Time-Change theorem, imply (2.28). □

The rest of this subsection discusses system dynamics, without assuming a specific policy. Thus the following discussion can be applied to all policies.

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

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

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

$$\begin{aligned}
&\left(\widehat{\Phi}_{jk}^r(\cdot), \widehat{\Phi}_{lk}^r(\cdot), \widehat{S}_k^r(\cdot);\ j \in \mathcal{J}, l, k \in \mathcal{K}\right)\\
&\quad \Rightarrow \left(\widehat{\Phi}_{jk}(\cdot), \widehat{\Phi}_{lk}(\cdot), \widehat{S}_k(\cdot);\ j \in \mathcal{J}, l, k \in \mathcal{K}\right);
\end{aligned} \tag{2.33}$$

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

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

respectively.

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

$$Q_k^r(t)=Q_k^r(0)+E_k^r(t)-S_k(T_k^r(t)), \tag{2.34}$$

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 (2.1), similar to (2.30),

$$\begin{aligned}\widehat{Q}_k^r(t)=&\widehat{Q}_k^r(0)+\widehat{E}_k^r(t)-\widehat{S}_k^r(\bar{\bar{T}}_k^r(t))-\mu_j\widehat{T}_k^r(t)\\ =&\widehat{Q}_k^r(0)+\widehat{\mathcal{E}}_k^r(t)-\widehat{S}_k^r(\bar{\bar{T}}_k^r(t))+\sum_{j\in\mathcal{J}}P_{jk}\mu_j\widehat{T}_j^r(t)\\ &+\sum_{l\in\mathcal{K}}P_{lk}\mu_l\widehat{T}_l^r(t)-\mu_k\widehat{T}_k^r(t);\end{aligned} \tag{2.35}$$

here

$$\begin{aligned}\widehat{\mathcal{E}}_k^r(t)=&\sum_{j\in\mathcal{J}}\widehat{\Phi}_{jk}^r\left(\bar{\bar{S}}_j^r\left(\bar{\bar{T}}_j^r(t)\right)\right)+\sum_{l\in\mathcal{K}}\widehat{\Phi}_{lk}^r\left(\bar{\bar{S}}_l^r\left(\bar{\bar{T}}_l^r(t)\right)\right)\\ &+\sum_{j\in\mathcal{J}}P_{jk}\widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t)\right)+\sum_{l\in\mathcal{K}}P_{lk}\widehat{S}_l^r\left(\bar{\bar{T}}_l^r(t)\right).\end{aligned} \tag{2.36}$$

Denote ($\widehat{Q}^r_w(t)$ is defined in (2.15), but we would like to repeat it here)

$$\begin{aligned}
\widehat{Q}^r_w(t) =& \sum_{j\in\mathcal{J}} m^e_j \widehat{Q}^r_j(t) + \sum_{k\in\mathcal{K}} m^e_k \widehat{Q}^r_k(t),\\
\widehat{X}^r_w(t) =& \widehat{Q}^r_w(0) + r(\rho^r - 1)t + \sum_{j\in\mathcal{J}} m^e_j \left[\widehat{E}^r_j(t) - \widehat{S}^r_j\left(\bar{\bar{T}}^r_j(t)\right)\right]\\
&+ \sum_{k\in\mathcal{K}} m^e_k \left[\widehat{\mathcal{E}}^r_k(t) - \widehat{S}^r_k\left(\bar{\bar{T}}^r_k(t)\right)\right],\\
\widehat{T}^r_+(t) =& r^{-1}\left(r^2 t - \sum_{j\in\mathcal{J}} T^r_j(r^2 t) - \sum_{k\in\mathcal{K}} T^r_k(r^2 t)\right).
\end{aligned} \tag{2.37}$$

From (2.5) and (2.4), one can verify that

$$-m^e_j\mu_j + \sum_{k\in\mathcal{K}} P_{jk}\mu_j m^e_k = -1, \tag{2.38}$$

$$-m^e_k\mu_k + \sum_{l\in\mathcal{K}} P_{kl}\mu_k m^e_l = -1. \tag{2.39}$$

Multiply (2.30) by m^e_j, (2.35) by m^e_k, and summing them together, one has

$$\begin{aligned}
&\widehat{Q}^r_w(t) = \widehat{X}^r_w(t) + \widehat{T}^r_+(t),\\
&\widehat{Q}^r_w(t) \geq 0,\\
&\widehat{T}^r_+(\cdot) \text{ is nondecreasing with } \widehat{T}^r_+(0) = 0.
\end{aligned} \tag{2.40}$$

Note that the policy may not be work-conserving, thus it is possible that $\widehat{T}^r_+$

increases at t when $\widehat{Q}^r_w(t) \neq 0$. Hence

$$\widehat{Q}^r_w(t) \geq \Phi(\widehat{X}^r_w)(t); \tag{2.41}$$

here Φ is the 1-dimensional Skorohod mapping; see for example, [31]. Equality in (2.41) holds when the system operates under any work-conserving policy.

2.6.2 Proof of Theorem 2.4.1: Lower bound

Proof of Theorem 2.4.1: Fix an arbitrary family of control policies $\{\pi^r\}$ which is asymptotically compliant. Define

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

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

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

Then

$$\{\mathcal{U}^r(t) > x\} = (\Gamma^r_1(t) \cup \Gamma^r_2(t)) \setminus \Gamma^r_3(t). \tag{2.42}$$

First we prove

$$\lim_{r\to\infty} \mathbb{P}\left\{\Gamma_3^r(t)\right\} = 0. \tag{2.43}$$

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-WIP patients during $[r^2 s, r^2 u]$ if the physician allocates all the capacity to k-WIP 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

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

This completes the proof of (2.43).

From (2.42) and (2.43) one can conclude that,

$$\liminf_{r\to\infty} \mathbb{P}\left\{\mathcal{U}^r(t) > x\right\} = \liminf_{r\to\infty} \mathbb{P}\left\{\Gamma_1^r(t) \cup \Gamma_2^r(t)\right\}. \tag{2.44}$$

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 $\mathcal{D}[0,t]$:

$$\bar{\bar{T}}_k^r(\cdot) \;\Rightarrow\; \lambda_k m_k e(\cdot), \quad k\in\mathcal{K}. \tag{2.45}$$

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

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

for $j \in \mathcal{J}$, $l, k \in \mathcal{K}$. Then from (2.35),

$$\begin{aligned}
&\sum_{l\in\mathcal{K}} P_{lk}\mu_l \widetilde{T}_l^r(s) - \mu_k \widetilde{T}_k^r(s)\\
&= \widetilde{Q}_k^r(s) - \widetilde{Q}_k^r(0) - \widetilde{\mathcal{E}}_k^r(s) + \widetilde{S}_k^r\left(\bar{\bar{T}}_k^r(s)\right) - \sum_{j\in\mathcal{J}} P_{jk}\mu_j \widetilde{T}_j^r(s).
\end{aligned} \tag{2.46}$$

On $\Gamma_0^r(t)$, $\sup_{0\le s\le t} \widetilde{Q}_k^r(s) \Rightarrow 0$. Together with (2.27), the expression of $\widetilde{\mathcal{E}}_k^r$ in (2.36), and $\bar{\bar{T}}_k^r(s) \le s$ for all $k \in \mathcal{K}$ (those hold for all asymptotic compliant

policies), one can deduce that the terms on the right-hand side of (2.46) 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 (2.45).

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

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

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

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

From (2.45) on $\Gamma_0^r(t)$, (2.27) 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 §2.4.2. 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 (2.41), 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

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

Combined with (2.44),

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

From the convergence of $\widehat{\mathcal{Z}}_+^r$, the right-hand side is exactly the lower bound in Theorem 2.4.1. This completes the proof. □

2.6.3 *Proof of Proposition 2.4.1: Invariant principle for work-conserving policies*

Proof of Proposition 2.4.1: For any family of work-conserving policies, besides (2.40), 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 (2.41).

From (2.33), (2.24) 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 (2.37) is stochastically bounded. By the

Lipschitz continuity of Φ ([31]), $\widehat{Q}_w^r$ 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 (2.29) is still true. We then have

$$\bar{\bar{T}}_j^r(\cdot) \;\Rightarrow\; \lambda_j m_j e(\cdot), \quad j \in \mathcal{J}. \tag{2.47}$$

For $k \in \mathcal{K}$, following the procedure in proving (2.45) in the proof of Theorem 2.4.1, one also has

$$\bar{\bar{T}}_k^r(\cdot) \;\Rightarrow\; \lambda_k m_k e(\cdot), \quad k \in \mathcal{K}. \tag{2.48}$$

Together with (2.47), (2.33), (2.24) and the Random-Time-Change theorem, when $r \to \infty$,

$$\widehat{X}_w^r \Rightarrow \widehat{X}. \tag{2.49}$$

By the continuity of the mapping Φ, (2.16) follows. □

2.6.4 *Proof of Theorem 2.4.3: State-space collapse*

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

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

$$Q_j^r(t) = Q_j^r(0) + E_j^r(t) - D_j^r(t), \quad j \in \mathcal{J}, \tag{2.50}$$

$$D_j^r(t) = S_j\left(T_j^r(t)\right), \quad j \in \mathcal{J}, \tag{2.51}$$

$$Q_k^r(t) = Q_k^r(0) + E_k^r(t) - D_k^r(t), \quad k \in \mathcal{K}, \tag{2.52}$$

$$E_k^r(t) = \sum_{j\in\mathcal{J}} \Phi_{jk}^r\left(S_j\left(T_j^r(t)\right)\right) + \sum_{l\in\mathcal{K}} \Phi_{lk}^r\left(S_l\left(T_l^r(t)\right)\right), \quad k \in \mathcal{K}, \tag{2.53}$$

$$D_k^r(t) = S_k\left(T_k^r(t)\right), \quad k \in \mathcal{K}, \tag{2.54}$$

$$\sum_{j\in\mathcal{J}} \left[T_j^r(t) - T_j^r(s)\right] + \sum_{k\in\mathcal{K}} \left[T_k^r(t) - T_k^r(s)\right] \le t - s, \quad \text{for} \quad s < t, \tag{2.55}$$

$$Y^r(t) = t - \left(\sum_{j\in\mathcal{J}} T_j^r(t) + \sum_{k\in\mathcal{K}} T_k^r(t)\right), \tag{2.56}$$

$$\int_0^\infty \left(\max_{j\in\mathcal{J}} \frac{\tau_j^r(t)}{d_j^r} - \frac{\tau_{j'}^r(t)}{d_{j'}^r}\right)^+ \wedge 1 dT_{j'}^r(t) = 0, \quad j' \in \mathcal{J}, \tag{2.57}$$

$$\int_0^\infty 1\left(Q_1^r(t) > \lambda_1^r d_1^r\right) d\sum_{k\in\mathcal{K}} T_k^r(t) = 0, \tag{2.58}$$

$$\int_0^\infty 1\left(Q_1^r(t) < \lambda_1^r d_1^r,\ \sum_{k\in\mathcal{K}} Q_k^r(t) > 0\right) d\sum_{j\in\mathcal{J}} T_j^r(t) = 0, \tag{2.59}$$

$$\int_0^\infty 1\left(\sum_{j\in\mathcal{J}} m_j^e Q_j^r(t) + \sum_{k\in\mathcal{K}} m_k^e Q_k^r(t) > 0\right) dY^r(t) = 0. \tag{2.60}$$

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

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

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

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

First we can prove the following lemma, which is similar to Lemma 2.6.1.

Lemma 2.6.3: 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

$$\bar{Q}_j^r(t) = \bar{E}_j^r(t) - \bar{E}_j^r\left((t - \bar{\tau}_j^r(t))-\right), \quad j \in \mathcal{J}. \tag{2.61}$$

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 (2.61), the conclusion can be easily proved. □

Similar to [35], there is

Lemma 2.6.4: 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

$$\bar{Q}_j(t) = \bar{Q}_j(0) + \lambda_j t - \bar{D}_j(t), \quad j \in \mathcal{J}, \tag{2.62}$$

$$\bar{D}_j(t) = \mu_j \bar{T}_j(t), \quad j \in \mathcal{J}, \tag{2.63}$$

$$\bar{Q}_k(t) = \bar{Q}_k(0) + \bar{E}_k(t) - \bar{D}_k(t), \quad k \in \mathcal{K}, \tag{2.64}$$

$$\bar{E}_k(t) = \sum_{j\in\mathcal{J}} \mu_j P_{jk} \bar{T}_j(t) + \sum_{l\in\mathcal{K}} \mu_l P_{lk} \bar{T}_l(t), \quad k \in \mathcal{K}, \tag{2.65}$$

$$\bar{D}_k(t) = \mu_k \bar{T}_k(t), \quad k \in \mathcal{K}, \tag{2.66}$$

$$\lambda_j \bar{\tau}_j(t) = \bar{Q}_j(t), \quad j \in \mathcal{J}, \tag{2.67}$$

$$\sum_{j\in\mathcal{J}} [\bar{T}_j(t) - \bar{T}_j(s)] + \sum_{k\in\mathcal{K}} [\bar{T}_k(t) - \bar{T}_k(s)] \le t - s, \quad \text{for} \quad s < t, \tag{2.68}$$

$$\bar{Y}(t) = t - \left(\sum_{j\in\mathcal{J}} \bar{T}_j(t) + \sum_{k\in\mathcal{K}} \bar{T}_k(t) \right), \tag{2.69}$$

$$\int_0^\infty \left(\max_{j\in\mathcal{J}} \frac{\bar{Q}_j(t)}{\lambda_j \widehat{d}_j} - \frac{\bar{Q}_{j'}(t)}{\lambda_{j'} \widehat{d}_{j'}} \right)^+ \wedge 1 d\bar{T}_{j'}(t) = 0, \quad j' \in \mathcal{J}, \tag{2.70}$$

$$\int_0^\infty 1\left(\bar{Q}_1(t) > \lambda_1 \widehat{d}_1 \right) d \sum_{k\in\mathcal{K}} \bar{T}_k(t) = 0, \tag{2.71}$$

$$\int_0^\infty 1\left(\bar{Q}_1(t) < \lambda_1 \widehat{d}_1, \sum_{k\in\mathcal{K}} \bar{Q}_k(t) > 0 \right) d \sum_{j\in\mathcal{J}} \bar{T}_j(t) = 0, \tag{2.72}$$

$$\int_0^\infty 1\left(\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j(t) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(t) > 0 \right) d\bar{Y}(t) = 0. \tag{2.73}$$

Remark 2.6.1: 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 (2.62)-(2.73) is called a *hydrodynamic model solution*, and one can prove that, any hydrodynamic model solution is Lipschitz, hence absolutely continuous and differentiable almost everywhere.

Proof: We follow the argument for Theorem 4.1 in [14]. Almost surely, we have the following convergence ([13]):

$$\bar{E}_j^r(t) \to \lambda_j t, \quad \text{u.o.c.}, \quad j \in \mathcal{J}, \tag{2.74}$$

$$\bar{S}_j^r \to \mu_j t, \quad \text{u.o.c.}, \quad j \in \mathcal{J}, \tag{2.75}$$

$$\bar{S}_k^r \to \mu_k t, \quad \text{u.o.c.}, \quad k \in \mathcal{K}, \tag{2.76}$$

$$\frac{1}{r}\Phi_{jk}^r(\lfloor rt \rfloor) \to P_{jk} t, \quad \text{u.o.c.}, \quad j \in \mathcal{J}, k \in \mathcal{K}, \tag{2.77}$$

$$\frac{1}{r}\Phi_{lk}^r(\lfloor rt \rfloor) \to P_{lk} t, \quad \text{u.o.c.}, \quad l, k \in \mathcal{K}. \tag{2.78}$$

Fix such a sample path and notice that on this sample path, we always have

$$\frac{1}{r}\left(\sum_{k\in\mathcal{K}}\left(T_k^r(rt) - T_k^r(rs)\right) + \sum_{j\in\mathcal{J}}\left(T_j^r(rt) - T_j^r(rs)\right)\right) \le t - s, \quad \text{for} \quad t > s.$$

As a result, there exists a subsequence $\{r_n\}$ such that as $n \to \infty$,

$$\frac{1}{r_n}T_j^{r_n}(r_n t) \to \bar{T}_j(t), \quad \text{u.o.c.}, \quad j \in \mathcal{J}, \tag{2.79}$$

$$\frac{1}{r_n}T_k^{r_n}(r_n t) \to \bar{T}_k(t), \quad \text{u.o.c.}, \quad k \in \mathcal{K}. \tag{2.80}$$

Here $\bar{T}_j, j \in \mathcal{J}, \bar{T}_k, k \in \mathcal{K}$ are Lipschitz continuous processes. Then (2.62)-(2.69) (except (2.67), which is from Lemma 2.6.3) follow from (2.50)-(2.56), (2.74)-(2.78) and (2.79)-(2.80). From the Lipschitz continuity of $T_j, j \in \mathcal{J}, T_k, k \in \mathcal{K}$, it is easy to see that $(\bar{\tau}_j, \bar{Q}_j, \bar{D}_j, j \in \mathcal{J}, \bar{E}_k, \bar{Q}_k, \bar{D}_k, k \in \mathcal{K})$ is also Lipschitz continuous.

The proofs for (2.70)-(2.73) are similar. Here we give a proof for (2.73). If (2.73) is not true, then there is a t_0 and $\delta > 0$ such that $\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j(t_0) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(t_0) > 0$ and $\bar{Y}(t_0+\delta) - \bar{Y}(t_0-\delta) \ge 0$. As $\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j(t_0) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(t_0)$ is Lipschitz continuous, we can also assume that this δ is chosen so that $\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j(t) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k(t) > 0$ for all $t \in [t_0 - \delta, t_0 + \delta]$.

Then for n large enough, $\sum_{j\in\mathcal{J}} m_j^e \bar{Q}_j^{r_n}(t) + \sum_{k\in\mathcal{K}} m_k^e \bar{Q}_k^{r_n}(t) > 0$ for all $t \in [t_0-\delta, t_0+\delta]$, and $\bar{Y}^{r_n}(t_0+\delta) - \bar{Y}^{r_n}(t_0-\delta) \geq 0$. However, this contradicts with our work-conserving assumption. As a result, (2.73) should be true. □

Lemma 2.6.5: 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 $\sum_{j\in\mathcal{J}} \lambda_j m_j^e = 1$, (2.38)-(2.39) and (2.62)-(2.66), 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 (2.73), (2.68) and (2.69), $\bar{Y}(\cdot) = 0$. This completes the proof. □

State-space collapse for triage patients: First we prove a state-space collapse result for the hydrodynamic model solution.

Lemma 2.6.6 (**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 $j \in \mathcal{J}$, 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 one can 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 (2.70), 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 (2.72), 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 (2.70), 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.

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

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

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 (2.70), $\bar{T}_1'(t) = 0$ 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 (2.71), $\sum_{k\in\mathcal{K}} \bar{T}_k'(t) = 0$. From (2.73), $\sum_{j\in\mathcal{J}} \bar{T}_j'(t) = 1$. 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 Lemma 2.6.5,

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

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

for $t \geq T_0$. □

The main result in this subsection is the following lemma, which proves the state-space collapse result for triage patients.

Lemma 2.6.7: Under Assumption 2.2.1 and the proposed family of control policies, when $r \to \infty$,

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

Proof: The basic argument is similar to the arguments in [9]. For completeness, we include it here. From Lemma 2.6.6, we know that Assumption 3.2 of Bramson holds, then from Theorem 5 of [9], we can obtain what the terms "multiplicative state space collapse" (equation (3.41) of [9]):

$$\frac{\sup_{0\leq t\leq T}\left|\widehat{Q}_j^r(t) - \Delta_j \min\left(\widehat{Q}_w^r(t),\ \widehat{\omega}\right)\right|}{\sup_{0\leq t\leq T} \widehat{Q}_w^r(t) \wedge 1} \ \Rightarrow\ 0.$$

Notice that here $\widehat{Q}_w^r(t)$ plays the role of $\widehat{W}^r$ in Theorem 5 of [9] there.

Next from our Proposition 2.4.1, we know that $\sup_{0\leq t\leq T} \widehat{Q}_w^r(t) \wedge 1$ is stochastically bounded. As a result,

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

This proves the result. □

State-space collapse for WIP patients: From Propositions 2.4.1 and Lemma 2.6.7, when $r \to \infty$, one has

$$\sum_{k\in\mathcal{K}} m_k^e \widehat{Q}_k^r \;\; \Rightarrow \;\; \left(\widehat{Q}_w - \widehat{\omega}\right)^+. \tag{2.81}$$

Recall that the proposed policy for WIP patients is to ensure

$$\max_{l,k\in\mathcal{K}} \sup_{0\le t\le T} \left| \frac{C_l'(\widehat{Q}_l^r(t))}{m_l^e} - \frac{C_k'(\widehat{Q}_k^r(t))}{m_k^e} \right| \;\; \Rightarrow \;\; 0. \tag{2.82}$$

Lemma 2.6.8: Under the family of control policies $\{\pi_*^r\}$, one has $(\widehat{Q}_k^r, k \in \mathcal{K}) \Rightarrow (\widehat{Q}_k, k \in \mathcal{K})$. Here

$$\widehat{Q}_k \;\; = \;\; \Delta_k \left((\widehat{Q}_w - \widehat{\omega})^+ \right), \quad k \in \mathcal{K}. \tag{2.83}$$

Proof: The proof is similar to [41]; for completeness, we include it here. From (2.82), for any given $T > 0$,

$$\max_{l,k\in\mathcal{K}} \sup_{0\le t\le T} \left| C_l'^{-1} \left(\frac{m_l^e}{m_k^e} C_k' \left(\widehat{Q}_k^r(t) \right) \right) - \widehat{Q}_l^r(t) \right| \;\; \Rightarrow \;\; 0. \tag{2.84}$$

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 (2.84) and (2.81), 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 (2.84), $(\widehat{Q}_l^r, l\in\mathcal{K}) \Rightarrow (\widehat{Q}_l, l\in\mathcal{K})$. Also

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

This proves (2.83). □

Proof of Theorem 2.4.3: Lemma 2.6.8 proved (2.18). From Proposition 2.4.1, Lemma 2.6.7 and the continuity of the function $\psi(x) = \min(x,\widehat{\omega})$, with the application of the continuous mapping theorem and the Convergence-together Theorem (Theorem 11.4.7 in [43]), we get (2.17). □

2.6.5 Proof of Theorem 2.4.2: Asymptotic optimality

Proof of Theorem 2.4.2: 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 2.4.3, 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 (2.25) is equivalent to "asymptotic compliance" for work-conserving policies (see Lemma 2.7.2); hence the family of the policies $\{\pi_*^r\}$ is asymptotically compliant.

By Theorem 2.4.3, together with the continuity of the cost functions,

one also has

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

Hence, under the family of the proposed policies, the lower bound in Theorem 2.4.1 is attained. As a result, the family of the proposed policies is asymptotically optimal. □

2.7 Additional proofs

2.7.1 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 2.4.1, $\widehat{Q}_j^r$, $j\in\mathcal{J}$, are stochastically bounded and (2.47) holds for any work-conserving policies. With these, notice that (2.30) is still true, hence we can verify the convergence (2.28). 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 WIP 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 (2.36). Denote $\widehat{T}_\mu^r=(\mu_k\widehat{T}_k^r)_{k\in\mathcal{K}}$. Then from

(2.35),

$$\widehat{T}_{\mu}^{r} = (P^T - I)^{-1}\widehat{\mathcal{Y}}_{\mathcal{K}}^{r}. \tag{2.85}$$

We can easily verify the stochastic boundedness of $\widehat{\mathcal{Y}}_{\mathcal{K}}^{r}$ from the facts $\bar{\bar{T}}_{j}^{r}(s) \leq s$ and $\bar{\bar{T}}_{k}^{r}(s) \leq s$, 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 $k \in \mathcal{K}$,

$$\widehat{E}_{k}^{r}(t) = \widehat{\mathcal{E}}_{k}^{r}(t) + \sum_{j\in\mathcal{J}} P_{jk}\mu_j \widehat{T}_{j}^{r}(t) + \sum_{l\in\mathcal{K}} P_{lk}\mu_l \widehat{T}_{l}^{r}(t). \tag{2.86}$$

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 2.7.1: Under any family of work-conserving policies, with FCFS among each WIP 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 $\omega_j^r(t) \geq \epsilon$, then

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

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

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

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

$$\begin{aligned} & \mathbb{P}\left\{\sup_{0\leq t\leq T} \bar{\bar{\omega}}_j^r(t) \geq \epsilon\right\} \\ \leq \quad & \mathbb{P}\left\{\lambda_j^r r\epsilon \leq \sup_{0\leq t\leq T}\left|\widehat{Q}_j^r(0) + \widehat{E}_j^r(t) - \widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t+\epsilon)\right) - \mu_j \widehat{T}_j^r(t+\epsilon)\right|\right\}. \end{aligned}$$

However, noticing that $\sup_{0\leq t\leq T}\left|\widehat{Q}_j^r(0) + \widehat{E}_j^r(t) - \widehat{S}_j^r\left(\bar{\bar{T}}_j^r(t+\epsilon)\right) - \mu_j \widehat{T}_j^r(t+\epsilon)\right|$ is stochastically bounded, 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

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

This completes the proof. □

Lemma 2.7.2: 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 2.6.1, by noticing that from Lemma 2.7.1 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, we have $\sup_{0\le s\le t}\bar{\tau}_j^r(s) \Rightarrow 0$. Note that the result here is slightly different from Lemma 2.6.1, 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. □

2.7.2 *Proof of Proposition 2.4.2: Asymptotic sample-path Little's law*

Lemma 2.7.3: 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

$$\mu_j\widehat{T}_j(t) = -\widehat{Q}_j(t) + \widehat{E}_j(t) - \widehat{S}_j\left(\lambda_j m_j t\right), \tag{2.87}$$

$$\widehat{E}_k(t) = \widehat{\mathcal{E}}_k(t) + \sum_{j\in\mathcal{J}} P_{jk}\mu_j\widehat{T}_j(t) + \sum_{l\in\mathcal{K}} P_{lk}\mu_l\widehat{T}_l(t), \tag{2.88}$$

$$(P^T - I)\left(\mu_k\widehat{T}_k\right)_{k\in\mathcal{K}} = \widehat{\mathcal{Y}}_{\mathcal{K}}. \tag{2.89}$$

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 (2.30), (2.86) and (2.85), we have $(\widehat{T}^r_\mu = (\mu_k \widehat{T}^r_k)_{k\in\mathcal{K}})$

$$\widehat{T}^r_j(t) = \left[\widehat{Q}^r_j(0) - \widehat{Q}^r_j(t) + \widehat{E}^r_j(t) - \widehat{S}^r_j\left(\bar{\bar{T}}^r_j(t)\right)\right]/\mu_j, \tag{2.90}$$

$$\widehat{E}^r_k(t) = \widehat{\mathcal{E}}^r_k(t) + \sum_{j\in\mathcal{J}} P_{jk}\mu_j \widehat{T}^r_j(t) + \sum_{l\in\mathcal{K}} P_{lk}\mu_l \widehat{T}^r_l(t), \tag{2.91}$$

$$\widehat{T}^r_\mu(t) = (P^T - I)^{-1}\widehat{\mathcal{Y}}^r_{\mathcal{K}}(t), \tag{2.92}$$

where

$$\begin{aligned}\widehat{\mathcal{E}}^r_k(t) &= \sum_{j\in\mathcal{J}} \widehat{\Phi}^r_{jk}\left(\bar{\bar{S}}^r_j\left(\bar{\bar{T}}^r_j(t)\right)\right) + \sum_{l\in\mathcal{K}} \widehat{\Phi}^r_{lk}\left(\bar{\bar{S}}^r_l\left(\bar{\bar{T}}^r_l(t)\right)\right) \\ &\qquad + \sum_{j\in\mathcal{J}} P_{jk}\widehat{S}^r_j\left(\bar{\bar{T}}^r_j(t)\right) + \sum_{l\in\mathcal{K}} P_{lk}\widehat{S}^r_l\left(\bar{\bar{T}}^r_l(t)\right), \\ \widehat{\mathcal{Y}}^r_k(t) &= \widehat{Q}^r_k(t) - \widehat{Q}^r_k(0) - \widehat{\mathcal{E}}^r_k(t) + \widehat{S}^r_k(\bar{\bar{T}}^r_k(t)) - \sum_{j\in\mathcal{J}} P_{jk}\mu_j \widehat{T}^r_j(t).\end{aligned}$$

As a result, $\left(\widehat{T}^r_j, j\in\mathcal{J},\ \widehat{E}^r_k, \widehat{T}^r_k, k\in\mathcal{K}\right)$ can be represented as a continuous mapping from $\left(\widehat{Q}^r_j, \widehat{E}^r_j, \widehat{S}^r_j, \bar{\bar{T}}^r_j, \widehat{\Phi}^r_{jk}, \widehat{\Phi}^r_{lk}, \widehat{Q}^r_k, \widehat{S}^r_k, \bar{\bar{T}}^r_k, j\in\mathcal{J}, l, k\in\mathcal{K}\right)$, whose convergence can be obtained from the assumptions and Theorem 2.4.3. The expressions (2.87)-(2.89) in the lemma can be easily verified from (2.90)-(2.92). This completes the proof. □

Proof of Proposition 2.4.2: We prove the result for j-triage patients. For k-WIP patients, the proof is similar. The convergence of $\widehat{Q}_j^r$, together with Lemma 2.7.1, ensure that, for any $T > 0$,

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

Thus it is enough to prove

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

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

$$\begin{aligned} & Q_j^r\left(t+\omega_j^r(t)\right) \\ \le\; & E_j^r(t+\omega_j^r(t)) - E_j^r(t) \le Q_j^r\left(t+\omega_j^r(t)\right) + \Delta S_j^r\left(t+\omega_j^r(t)\right); \end{aligned} \tag{2.93}$$

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:

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

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 $\widehat{T}_j^r(\cdot)$, then 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 2.7.1 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. □

2.7.3 Proof of Proposition 2.4.3: Snapshot principle – virtual waiting time and age

Lemma 2.7.4: 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 2.6.1. For $k\in\mathcal{K}$, note that the convergence of $\widehat{E}_k^r$ has been proved in Lemma 2.7.3. 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 2.7.1 we have $\sup_{0\le s\le t}\bar{\bar{\tau}}_k^r(s) \Rightarrow 0$. □

Proof of Proposition 2.4.3: This can be easily deduced from Proposition 2.4.2, Lemmas 2.7.2 and 2.7.4. □

2.7.4 Proof of Proposition 2.4.4: Snapshot principle – sojourn time and queue lengths

The argument here follows the framework in [37]. Introduce the following notation: $\tau_{jh}^r(t)$ is the time at which the patient of interest to us arrives to the system, and $\zeta_{jki}^r(t)$ is the time at which this patient becomes a k-WIP patient for the ith time (it is also related to h, but we omit h to simplify the notation). Then

$$t \le \zeta_{jki}^r(t) \le \tau_{jh}^r(t) + W_{jh}^r(t). \tag{2.94}$$

Define the fluid scaled processes

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

Lemma 2.7.5: Under the family of control policies $\{\pi_*^r\}$ with FCFS among each WIP class, if h is j-feasible, then for any $T \ge 0$, as $r \to \infty$,

$$\sup_{0\le t\le T} \bar{\bar{W}}_{jh}^r(t) \Rightarrow 0, \tag{2.95}$$

$$\sup_{0\le t\le T} \left[\bar{\bar{\tau}}_{jh}^r(t) - t\right] \Rightarrow 0. \tag{2.96}$$

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

$$\sup_{0\le t\le T}\left[\bar{\bar{\zeta}}^r_{jki}(t) - t\right] \;\Rightarrow\; 0. \tag{2.97}$$

We first assume this last lemma is true and prove Proposition 2.4.4.

Proof of Proposition 2.4.4: 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 2.7.5 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 2.4.2, the conclusion is immediate. □

Proof of Lemma 2.7.5: We first prove (2.95). 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 [37], denote $\|h\| = \sum_{k=1}^{K} h_k$. Then we have

$$\begin{aligned}\mathbb{P}\left\{\sup_{0\leq t\leq T} \bar{\bar{W}}^r_{jh}(t) \geq \epsilon\right\} \quad &\leq \quad \max_{k\in\mathcal{K}} \mathbb{P}\left\{\sup_{0\leq t\leq T+\epsilon} \bar{\bar{\omega}}^r_k(t) \geq \frac{\epsilon}{\|h\|+1}\right\} \\ &\quad + \mathbb{P}\left\{\sup_{0\leq t\leq T+\epsilon} \bar{\bar{\omega}}^r_j(t) \geq \frac{\epsilon}{\|h\|+1}\right\}.\end{aligned} \tag{2.98}$$

From Lemma 2.7.1, the right-hand side of (2.98) converges to 0, hence (2.95) holds.

The proof of (2.96) follows the one in [37]. 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

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

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

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

The same reason as in [37] then implies that the latter expression vanishes, as $r \to \infty$. This establishes (2.96).

Combining (2.95), (2.96) with (2.94), now yields (2.97). □

Proof of Corollary 2.4.1: This is implied by Propositions 2.4.4, 2.4.2 and 2.4.3. □

2.7.5 Proof for Lemma 2.5.1

Proof: The proof follows the framework in §2.6.4. For completeness, we include the steps here.

Firstly, the argument in proving Lemma 2.6.4 still works for the new scheduling policy, except for the equation (2.70), which we replace by the following:

$$\int_0^\infty \left(d_j - \frac{\bar{Q}_j(t)}{\lambda_j} - \min_{j'\in\mathcal{J}}\left\{d_{j'} - \frac{\bar{Q}_{j'}(t)}{\lambda_{j'}}\right\}\right)^+ \wedge 1 d\bar{T}_j(t) = 0, \quad j \in \mathcal{J}. \quad (2.99)$$

Because the proof of Lemma 2.6.5 does not use (2.70), thus it is still true for the new policy.

Next we prove that, for any fixed $C > 0$ and a 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 $\widetilde{T}_0$ such that, for all $t \geq \widetilde{T}_0$,

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

To prove this, for $j \in \mathcal{J}$, define

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

If $\widetilde{f}_1(t) > 0$ and is differentiable, then one can claim

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

Indeed, if this is not true, then $\bar{T}_1'(t) \neq 0$ and from (2.99), one has $\widehat{d}_1 - \frac{\bar{Q}_1(t)}{\lambda_1} = \min_{j'\in\mathcal{J}}\left\{\widehat{d}_{j'} - \frac{\bar{Q}_{j'}(t)}{\lambda_{j'}}\right\}$. Together with $\tilde{f}_1(t) > 0$, one can prove by contradiction that $\bar{Q}_1(t) < \lambda_1 \widehat{d}_1$. Then from (2.72), one has $\bar{Q}_k(t) = 0$ for all $k \in \mathcal{K}$. This, together with $\widetilde{f}_1(t) > 0$, will contradict the definition of $\widetilde{\Delta}_j$.

As a result, $\widetilde{f}_1$ will decrease to 0 in a finite time (denote it as $\widetilde{T}_1$) and once becoming 0, it will never be positive again. Then for each $j \in \mathcal{J}$, if $\widetilde{f}_j(t) > 0$ for some $t \geq \widetilde{T}_1$, then $\bar{T}_j'(t) = 0$ from (2.70), hence

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

Consequently, after a finite time (denote it by $\widetilde{T}_2 \geq \widetilde{T}_1$), all $\widetilde{f}_j$ will be 0 and

will never be positive again.

Now for any $t \geq \widetilde{T}_2$, $\widetilde{f}_j(t) = 0$ for all $j \in \mathcal{J}$. Define

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

We can assume $\widetilde{g}_1(t) > 0$ whenever $\sum_{j\in\mathcal{J}} \lambda_j \widehat{d}_j m_j \widetilde{g}_j(t) > 0$. Otherwise, if $\widetilde{g}_1(t) = 0$ and there is another $j \in \mathcal{J}$ such that $\widetilde{g}_j(t) > 0$, then from the definition of $\widetilde{\Delta}_{\mathcal{J}}$, $\widehat{d}_1 - \bar{Q}_1(t)/\lambda_1 > \min_{j\in\mathcal{J}}(\widehat{d}_j - \bar{Q}_j(t)/\lambda_j)$, and from (2.99), $\bar{T}_1'(t) = 0$ and $\widetilde{g}_1'(t) = \frac{1}{\widehat{d}_1} > 0$. Hence right after t, $\widetilde{g}_1(\cdot)$ will be positive.

Then the discussion in the last paragraph in the proof of Lemma 2.6.6, we can prove that in finite time (denote it by $\widetilde{T}_0 \geq \widetilde{T}_2$), $\sum_{j\in\mathcal{J}} \lambda_j \widehat{d}_j m_j \widetilde{g}_j(t)$ will converge to 0. It follows that, for all $t \geq \widetilde{T}_0$, $\widetilde{f}_j(t) = \widetilde{g}_j(t) = 0, j \in \mathcal{J}$. This proves that

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

for $t \geq T_0$.

Then we can follow the argument in [9], first proving "multiplicative state space collapse" (equation (3.41) of [9]):

$$\frac{\sup_{0\leq t\leq T}\left|\widehat{Q}_j^r(t) - \widetilde{\Delta}_j \min\left(\widehat{Q}_w^r(t),\ \widehat{\omega}\right)\right|}{\sup_{0\leq t\leq T} \widehat{Q}_w^r(t) \wedge 1} \;\Rightarrow\; 0.$$

Then from our Proposition 2.4.1, we know

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

This proves the result. □

2.7.6 Proof for Lemma 2.5.2

Proof: We follow the framework in §2.6.4, but for WIP patients now. We first prove that under the proposed policy, any limit of the hydrodynamic scaled processes $\bar{E}_j, \bar{S}_j, \bar{\tau}_j, \bar{T}_j, \bar{Q}_j, \bar{D}_j, j\in\mathcal{J}$, $\bar{E}_k, \bar{S}_k, \bar{T}_k, \bar{Q}_k, \bar{D}_k, k\in\mathcal{K}$ should satisfy (2.62)-(2.73), as well as the following:

$$\int_0^\infty\left(\max_{k'\in\mathcal{K}}H_{k'}\left(GC'\left(\bar{Q}(t)\right)\right)_{k'}-H_k\left(GC'\left(\bar{Q}(t)\right)\right)_k\right)^+d\bar{T}_k(t)=0. \quad (2.100)$$

This is because, if the above is not true, then there is a t_0 and $\delta>0$ such that $\max_{k'\in\mathcal{K}}H_{k'}\left(GC'\left(\bar{Q}(t_0)\right)\right)_{k'} > H_k\left(GC'\left(\bar{Q}(t_0)\right)\right)_k$ and $\bar{T}_k(t_0+\delta)-\bar{Y}_k(t_0-\delta)\geq 0$. We can also assume that this δ is chosen so that $\max_{k'\in\mathcal{K}}H_{k'}\left(GC'\left(\bar{Q}(t)\right)\right)_{k'} > H_k\left(GC'\left(\bar{Q}(t)\right)\right)_k$ for all $t\in[t_0-\delta,t_0+\delta]$. Then for n large enough, $\max_{k'\in\mathcal{K}}H_{k'}\left(GC'\left(\bar{Q}^{r_n}(t)\right)\right)_{k'} > H_k\left(GC'\left(\bar{Q}^{r_n}(t)\right)\right)_k$ for all $t\in[t_0-\delta,t_0+\delta]$, and $\bar{T}_k^{r_n}(t_0+\delta)-\bar{T}_k^{r_n}(t_0-\delta)\geq 0$. However, this contradicts with the work principle for WIP patients. As a result, (2.100) should be true.

Without loss of generality (from Lemma 2.6.5 and 2.6.6), we assume

that for all $t \geq 0$,

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

For any fixed $k \in \mathcal{K}$, define a $K \times K$ matrix $\mathcal{B}_k = \Upsilon + \Theta_k$, 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

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

Recall the definition of Δ_k in Lemma 2.4.1 and define

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

Then

$$\mathcal{B}_k G C'(\bar{Q}_0) = 0.$$

Here $C'(\bar{Q}_0)$ is a K-dimensional vector with $C'_k(\bar{Q}_{0k})$ being its kth component.

This means that for any $k, l \in \mathcal{K}$,

$$H_k \left(GC'(\bar{Q}_0)\right)_k = H_l \left(GC'(\bar{Q}_0)\right)_l .$$

Next we prove that for any fixed $C > 0$ and a 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 $\widetilde{T}_0$ such that, for all $t \geq \widetilde{T}_0$,

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

We first prove that, if $\bar{Q}(t) \neq \bar{Q}_0$, then there is a k_+ such that

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

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

Now choose any $k_- \in \text{argmin}_{k\in\mathcal{K}} \{ \frac{C_k'(\bar{Q}_k(t))}{m_k^e} - \frac{C_k'(\bar{Q}_{0k})}{m_k^e} \}$, then $C_{k_-}'(\bar{Q}_{k_-}(t)) - C_{k_-}'(\bar{Q}_{0k_-}) \leq 0$. Now we will prove that

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

It is enough to prove that $\left(G\left(C'(\bar{Q}(t))-C'(\bar{Q}_0)\right)\right)_{k_-} \leq 0$. As $G_{k_-k'} \leq 0$ for all $k' \neq k_-$ and $k_- \in \text{argmin}_{k\in\mathcal{K}}\{\frac{C'_k(\bar{Q}_k(t))}{m^e_k} - \frac{C'_k(\bar{Q}_{0k})}{m^e_k}\}$, we have

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

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

As $H_{k_-}\left(GC'(\bar{Q}_0)\right)_{k_-} = H_{k_0}\left(GC'(\bar{Q}_0)\right)_{k_0}$, thus $H_{k_-}\left(GC'(\bar{Q}(t))\right)_{k_-} < H_{k_0}\left(GC'(\bar{Q}(t))\right)_{k_0}$. From (2.100) we then have $\bar{T}'_{k_-}(t) = 0$. Then $\frac{\bar{Q}'_{k_-}(t)}{m^e_{k_-}} = \frac{\lambda_k}{m^e_{k_-}} > 0$. As a result, there will be a finite time $\bar{T}_0$ such that for all $t \geq \bar{T}_0$, $\frac{C'_k(\bar{Q}_k(t))}{m^e_k} \geq \frac{C'_k(\bar{Q}_{0k})}{m^e_k}$, which is equivalent to $\bar{Q}_k(t) \geq \bar{Q}_{0k}$. However, we have

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

as a result, $\bar{Q}_k(t) = \bar{Q}_{0k}$ for all k and $t \geq \bar{T}_0$.

Then we can follow the argument in [9], first proving "multiplicative state space collapse" (equation (3.41) of [9]):

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

Then from our Proposition 2.4.1, we know

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

This proves the result. □

2.7.7 Proof for Proposition 2.5.1: Waiting time cost

We first provide the proof of a lower bound, which is similar to the proof of Theorem 2.4.1.

For any fixed $k \in \mathcal{K}$, define C_a^r, C_b^r, C_c^r as follows:

$$
\begin{aligned}
C_{k1}^n &= \frac{1}{E_k^r(r^2b) - E_k^r(r^2a)} \sum_{i=E_k^r(r^2a)}^{E_k^r(r^2b)} \frac{1}{r}\tau_{k,i}^r, \\
C_{k2}^n &= \frac{1}{E_k^r(r^2b) - E_k^r(r^2a)} \int_{na}^{nb} \frac{1}{r} Q_k^r(t)dt, \\
C_{k3}^n &= \frac{1}{E_k^r(r^2b) - E_k^r(r^2a)} \sum_{i=E_k^r(r^2a)}^{E_k^r(r^2b)-Q_k^r(b)} \frac{1}{r}\tau_{k,i}^r.
\end{aligned}
$$

Then $C_{k3}^r \le C_{k2}^r \le C_{k1}^r$. Further,

$$
\begin{aligned}
C_{k1}^r - C_{k3}^r &= \frac{1}{E_k^r(r^2b) - E_k^r(r^2a)} \sum_{i=E_k^r(r^2b)-Q_k^r(r^2b)+1}^{E_k^r(r^2b)} \frac{1}{r}\tau_{k,i}^r \\
&\le \frac{r^{-1}}{\bar{\bar{E}}_k^r(b) - \bar{\bar{E}}_k^r(a)} \widetilde{Q}_k^r(b) \max_{E_k^r(r^2b)-Q_k^r(r^2b)+1 \le i \le E_k^r(r^2b)} \frac{1}{r}\tau_{k,i}^r.
\end{aligned}
$$

As the proposed policy is work-conserving, Proposition 2.4.1 and Lemma 2.7.1 hold. Then we have $C_{k1}^r - C_{k3}^r \Rightarrow 0$ as $r \to \infty$. As a result, $C_{k1}^r - C_{k2}^r \Rightarrow 0$ as $r \to \infty$, that is, for any $0 \le a < b \le T$,

$$
\frac{1}{\bar{\bar{E}}_k^r(b) - \bar{\bar{E}}_k^r(a)} \left(\int_a^b \widehat{\tau}_k^r d\bar{\bar{E}}_k^r - \int_a^b \widehat{Q}_k^r(s)ds \right) \Rightarrow 0. \tag{2.101}
$$

With this, we can prove the following

$$\liminf_{r\to\infty} \mathbb{P}\left\{\widetilde{\mathcal{U}}^r(t) > x\right\} \geq \mathbb{P}\left\{\int_0^t \sum_{k\in\mathcal{K}} \lambda_k C_k \left(\widehat{\Delta}_k \left((\widehat{Q}_w(s) - \widehat{\omega})^+\right)/\lambda_k\right) ds > x\right\}. \tag{2.102}$$

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:

$$\begin{aligned} \min_x \quad & \sum_{k\in\mathcal{K}} \lambda_k C_k(x_k/\lambda_k) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{K}} m_k^e x_k = a, \\ & x \geq 0. \end{aligned} \tag{2.103}$$

The argument is modified slightly from the discussion in proving Proposition 6 of [41]. For completeness, we provide the several key steps here. Fix $\epsilon > 0$ and for any $t \geq 0$, consider a sequence of stopping times of $\tilde{Q}_\omega$ defined as

$$\begin{aligned} t_1 =& \min\{1, \inf\{0 < t \leq 1 : |(\widetilde{Q}_\omega(t) - \widehat{\omega})^+ - \lfloor(\widetilde{Q}_\omega(0) - \widehat{\omega})^+/\epsilon\rfloor\epsilon| \geq \epsilon\}\}, \\ t_{i+1} =& \min\{1, \inf\{t_i < t : |(\widetilde{Q}_\omega(t) - \widehat{\omega})^+ - \lfloor(\widetilde{Q}_\omega(t_i) - \widehat{\omega})^+/\epsilon\rfloor\epsilon| \geq \epsilon\}\}. \end{aligned}$$

Thus t_{i+1} is the first time $(\widetilde{Q}_\omega - \widehat{\omega})^+$ changes by ϵ starting from $(\widetilde{Q}_\omega(t_i) - \widehat{\omega})^+$ at time t_i. Because $(\widetilde{Q}_\omega - \widehat{\omega})^+$ is continuous, $\sup_i(t_{i+1} - t_i) \to 0$ as $\epsilon \to 0$, so that $\sup_i(t_{i+1} - t_i) = O(\epsilon)$. Via Jensen's inequality and (2.101), we can

prove that

$$
\begin{aligned}
&\widetilde{\mathcal{U}}^r(t)\\
&\geq \sum_{k\in\mathcal{K}}\sum_i \lambda_k(t_{i+1}-t_i)C_k\left(\lambda_k^{-1}(t_{i+1}-t_i)^{-1}\int_{t_i}^{t_{i+1}}(\widetilde{Q}_k^r(t)-\widehat{\omega})^+dt\right)+o_r(1)+O(\epsilon)\\
&\geq \sum_{k\in\mathcal{K}}\sum_i \lambda_k(t_{i+1}-t_i)C_k\left((t_{i+1}-t_i)^{-1}\int_{t_i}^{t_{i+1}}\widehat{\Delta}_k\left((\widetilde{Q}_\omega^r(t)-\widehat{\omega})^+\right)dt\right)+o_r(1)+O(\epsilon).
\end{aligned}
$$

Using the fact that $(\widetilde{Q}_\omega^r-\widehat{\omega})^+\Rightarrow(\widetilde{Q}_\omega-\widehat{\omega})^+$ and the way defining t_i, we have

$$
(t_{i+1}-t_i)\int_{t_i}^{t_{i+1}}\widetilde{Q}_\omega^r(t)dt=(\widetilde{Q}_\omega(t_i)-\widehat{\omega})^+ + O(\epsilon)+o_r(1).
$$

We can arrive at the following

$$
\liminf_{r\to\infty}\widetilde{\mathcal{U}}^r(t)\geq\sum_{k\in\mathcal{K}}\sum_i\lambda_k(t_{i+1}-t_i)C_k\left(\lambda_k^{-1}\widehat{\Delta}_k\left((\widetilde{Q}_\omega(t_i)-\widehat{\omega})^+\right)\right)+O(\epsilon).
$$

Letting $\epsilon\to 0$, then we have (2.102).

Finally, following the discussion in §2.6.4 exactly, especially the steps to get the state-space collapse results, one can prove that the family of modified policies $\{\tilde{\pi}_*^r\}$ reaches the lower bound. As a result, $\{\tilde{\pi}_*^r\}$ is asymptotically optimal.

3. AN ALTERNATIVE MODEL

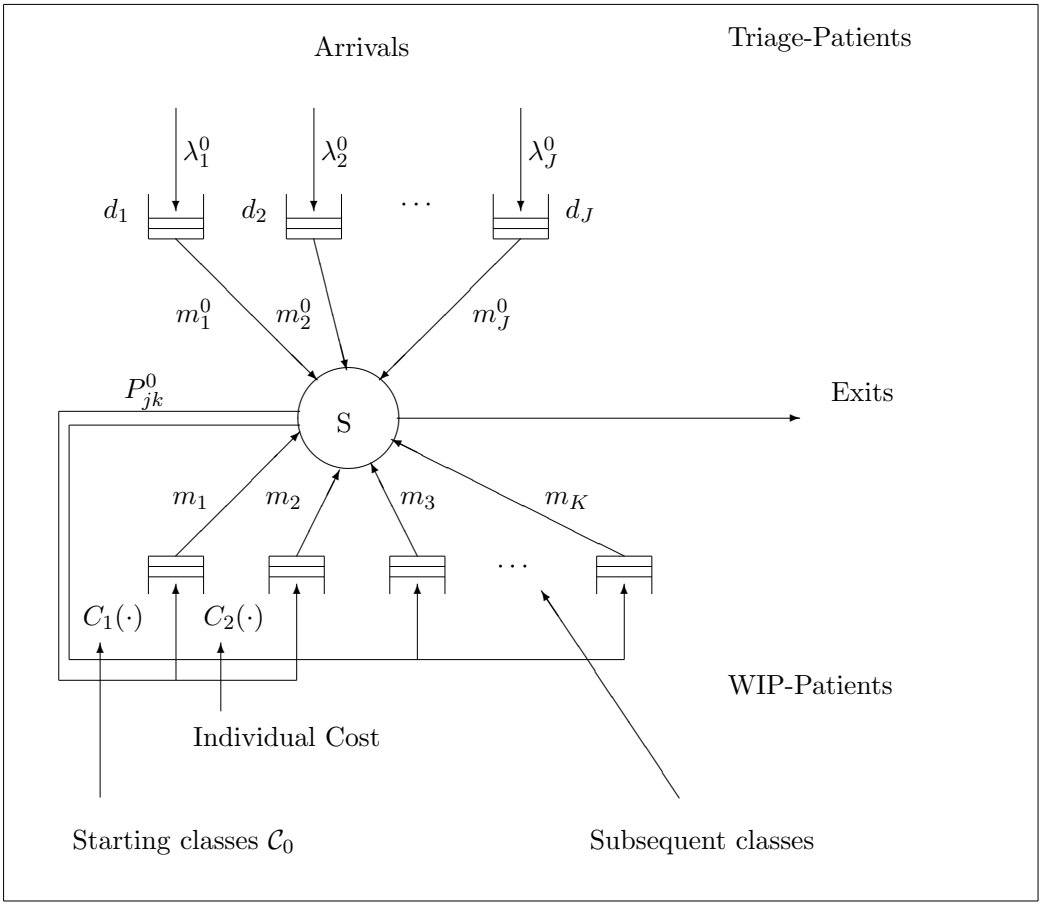
3.1 Sojourn time model and its policy

In some emergency departments, the costs are not based on the queue lengths, but on individual's sojourn time ([10]). In this section, we consider an alternative model, which is called *sojourn time model*. The structure of this model is identical to the figure in §2, except that congestion cost is associated with each patient's sojourn time in the WIP stage (as opposed to queueing and waiting costs previously).

With an assumption that the routing matrix P is upper-triangular, the number of routing vectors is finite. Thus, without loss of generality, one can assume that in the WIP stage 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 $k \in \mathcal{C}_0$, and let $\mathcal{C}_k$ denote all the classes on the route that starts at k. If the waiting time for a specific patient, say patient i, with starting class k waits $\omega_{k'}(i)$, as a k'-WIP patient ($k' \in \mathcal{C}_k$), then the sojourn time of this patient is $\sum_{k' \in \mathcal{C}_k} \omega_{k'}(i)$. Any class in $\bigcup_{k \in \mathcal{C}_0} \mathcal{C}_k \backslash \{k\}$ is called a *subsequent class*. The structure of the model is as in the following figure.

Now the cumulative cost incurred by those WIP patients till time t is

Fig. 3.1: Patient flow in emergency department (sojourn time cost)



as follow:

$$\mathcal{S}^r(t) = \sum_{k \in \mathcal{C}_0} \sum_{i=1}^{E_k(t)} C_k \left(\sum_{k' \in \mathcal{C}_k} \omega_{k'}(i) \right). \tag{3.1}$$

Here $C_k(\cdot)$ is a convex increasing function (which differs from those in the previous section).

There are still deadline constraints on triage patients. As a result, the problem is still to minimize the cumulative cost above, while subject to the

deadline constraints on those triage patients. Similarly to the queue length model, we will solve this problem in the conventional heavy traffic framework.

The heavy traffic framework is similar to the one for the queue length model. The definition of *asymptotic compliance* is also identical to the one in the queue length model.

Define the diffusion scaled waiting time of a WIP patient at stage k' with starting class k by

$$\widetilde{\omega}^r_{k'}(t) = r^{-1}\omega^r_{k'}(r^2 t).$$

Assume that, a WIP patient starting with class k incur a sojourn time cost $C_k\left(\sum_{k'\in\mathcal{C}_k}\widetilde{\omega}^r_{k'}\right)$. Then the cumulative queueing cost till time t is

$$\widetilde{\mathcal{S}}^r(t) = \sum_{k\in\mathcal{C}_0}\int_0^t C_k\left(\sum_{k'\in\mathcal{C}_k}\widehat{\omega}^r_{k'}(s)\right)d\bar{\bar{E}}^r_k(s). \tag{3.2}$$

Definition 3.1.1: 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 $x > 0$,

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

here $\{\widetilde{\mathcal{S}}^r_{**}\}$ is the family of cumulative queueing costs defined through

(2.9) under the family of control policies $\{\pi^r_{**}\}$, and $\{\widetilde{\mathcal{S}}^r\}$ is the sequence of queueing costs corresponding to any other asymptotically compliant family of policies $\{\pi^r\}$.

We propose the following routing policy: the first step, using a threshold policy to determine the priority between triage classes and WIP classes, and the step using (2.13) to determine priorities among triage patients, do not change. The step determining the priority among WIP classes will change as follows:

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

$$\max_{l,k\in\mathcal{C}_0}\sup_{0\le t\le T}\left|\frac{C_l'\left(\frac{\widehat{Q}_l^r(t)}{\lambda_l^r}\right)}{m_l^e}-\frac{C_k'\left(\frac{\widehat{Q}_k^r(t)}{\lambda_k^r}\right)}{m_k^e}\right| \;\Rightarrow\; 0. \tag{3.3}$$

Here Q_l, Q_k are the queue lengths of the starting classes $j,k\in\mathcal{C}_0$, and m_l^e, m_k^e are the corresponding effective means of service times. An example of such a policy is to choose $k\in\operatorname{argmax}_{k\in\mathcal{C}_0}\frac{C_k'\left(\widehat{Q}_k^r(t)/\lambda_k^r\right)}{m_k^e}$. Other examples of policies satisfying the above can be modified from the policies in §2.5.2: assume G and H are $K\times K$-dimensional invertible matrix and K-dimensional vectors in §2.5.2, the physician chooses a patients from the class with index

$$k\in\operatorname{argmax}_{k\in\mathcal{C}_0}H_k\left(GC'\left(\frac{\widehat{Q}^r(t)}{\lambda^r}\right)\right)_k,$$

here $C'\left(\frac{\widehat{Q}^r(t)}{\lambda^r}\right)$ is a $|\mathcal{C}_0|$-dimensional column vector with $C'_k\left(\frac{\widehat{Q}^r_k(t)}{\lambda^r_k}\right)$ being its kth component (here $|\mathcal{C}_0|$ is the number of terms in $\mathcal{C}_0$).

This family of policies is denoted by $\{\widetilde{\pi}^r_{**}\}$.

Giving priority to all subsequent classes when serving WIP classes is consistent with the observation in [40], where it is referred to as "Prioritize Old" policy.

Theorem 3.1.1 (**Sojourn Time Cost**): The family of control policies $\{\widetilde{\pi}^r_{**}\}$ is asymptotically optimal.

The proof for this Theorem can be found in §3.4.

Remark 3.1.1: 1. A different feature from the queue length case in choosing which WIP class to serve is as follows: one assigns priority to those patients who have already received at least one WIP treatment. Then when applying the results to the case with upper WIP-to-WIP transition matrix, such a service policy is not FCFS within classes: indeed, if patients in an WIP class can originate from both triage and WIP 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.

2. *Congestion laws*: Similarly to our cost-per-visit model, and assuming the above class designation, the snapshot principle also prevails for the

WIP 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 WIP 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 WIP process in class k, would be simply the age of the head-of-the-line patient in class k.

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

Most triage indices are based on 5 severity levels ([18, 29]). This granularity is typically too lean to account for patient characteristics that are relevant for decision making - clinical and operational. For example, the ED in Israel ([10]), 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 WIP phases are unknown at the triage state, but existing report, as well as ED directors, tell us that experienced ED physicians or nurses can predict them accurately; see [39, 40]. In this subsection, based on data from our partner, we use our models to assess the operational benefits of such predictions.

For simplicity and insight, we analyze only the WIP part of the ED patient flow, and we focus on A&D status and the number of WIP visits to an ED physician (which we refer to as WIP phases: each such phase will be regarded as a separate class in our formal model.)

In ED-Partner, patients experience 1-5 WIP phases: 28% go through 1 phase only, 30% have 2 phases, 28% - 3 phases, 11% - 4 phases, and 3% go through 5 WIP phases.

Tab. 3.1: Number of WIP visits

# WIP visits	1	2	3	4	5
Proportion	0.28	0.30	0.28	0.11	0.03

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 WIP phases are independent; hence, for example, the fraction of patients who will be admitted after 3 WIP phases is $40\% \times 28\% = 11.2\%$. Expert-solicitation in [10] 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 $c_d(t) = 2t^2$.

Tab. 3.2: Cost functions

A & D Status	Admitted	Discharged
Proportion/Cost function	0.40, t^2	0.60, $2t^2$

Assume that the external arrival rate is 1, and the mean service time for WIP patients is equal across all phases (this is not unreasonable from our experience); denote this common value by m, which is determined so that the ED operates in heavy traffic (traffic intensity $\rho \approx 1$).

Now we compare three scenarios: no-information, where the ED controller is aware of neither A&D status nor the number of WIP phases; partial-information, where only the number of WIP 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. The results of the three scenarios are summarized in the following table (here "Y" means the information can be estimated when a patient arriving at the ED, "N" means can not; "$\Downarrow$ 18.01%" means by comparing to the "Benchmark" in Case 1, the congestion cost can be reduced by 18.01%, similarly to "$\Downarrow$ 26.8%".):

Tab. 3.3: Comparison of results

	Case 1	Case 2	Case 3
# WIP visits	N	Y	Y
A & D Status	N	N	Y
Congestion Cost	Benchmark	$\Downarrow$18.01%	$\Downarrow$26.8%

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

$$c(t) = 0.4c_a(t) + 0.6c_d(t) = 1.6t^2. \tag{3.4}$$

In §3.5 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 the 5 phase problem, one can first consider the last two phases. It can be argued, similarly to §3.5, 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, the 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 §2.4.2.

Partial information: Now assume that the ED director knows, for individual patients, their number of WIP phases (1-5). Then the cost function is still as in (3.4). The patients are initially classified into 5 WIP 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 the sojourn time analysis in the previous section, an asymptotically optimal policy assigns priority to all non-starting WIP classes, while allocating the remaining service capacity to the 5 starting phases as follows: serve a class with index

$$k \in \max_{l \in \mathcal{K}} \frac{Q_l(t)}{l \times p_l}. \tag{3.5}$$

Here Q_l is the queue length of class l WIP 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 (3.7) 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 WIP visits will reduce 18.01% of the no-information cost. This is consistent with [39], in which this number of visits (complexity) is identified as an important factor for improving ED operations.

Complete information: Now assume, at the director's disposal, an accurate prediction of both the number of WIP 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 the analysis in the Appendix (especially (3.7) 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 opti-

mization problem:

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

(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, the 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 [40], who showed that A&D status contributes to improving ED operations. Furthermore, having jointly the A&D status and the number of WIP phases reduces congestion costs by 26.8%.

3.3 Imputed cost

The 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. The results in this thesis 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 the assumptions on mean service times, one can expect that

$$c_a Q_a(t) \approx c_d Q_d(t)$$

from the 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,

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

3.4 Proof of Theorem 3.1.1: Sojourn time cost

We first provide the proof of an asymptotic lower bound for all asymptotically compliant policies. Note that, when an WIP patient transfers to a next stage, the cost accumulates and the cost function does not change. As a result, whenever there are WIP patients in the ED, the physician should not be idle, as the physician can always serve an WIP 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 WIP patients are stochastically bounded. We now restrict our discussion to asymptotically compliant policies, in which the physician can not be idle if there are WIP patients. Then one can prove Lemma 2.7.1. Following exactly the discussion in §2.7.7, we can prove that, for any $0 \le a < b \le T$,

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

Now, following the steps in §2.7.7 (also in [41]), we can prove that for all $x \ge 0$,

$$\liminf_{r\to\infty} \mathbb{P}\left\{ \widetilde{\mathcal{S}}^r(t) > x \right\} \ge \mathbb{P}\left\{ \int_0^t \sum_{k\in\mathcal{C}_0} \lambda_k C_k \left(\widehat{\Delta}^*_k \left((\widehat{Q}_w(s) - \widehat{\omega})^+ \right) / \lambda_k \right) ds > x \right\}.$$

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

$$\begin{aligned} \min_x \quad & \sum_{k\in\mathcal{C}_0} \lambda_k C_k \left(\sum_{j\in\mathcal{C}_k} x_j/\lambda_k \right) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{C}_0} \sum_{k'\in\mathcal{C}_k} m^e_{k'} x_{k'} = a, \\ & x \geq 0. \end{aligned} \tag{3.6}$$

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 (3.6). 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 $C_k(\sum_{k'\in\mathcal{C}_k} x_{k'}/\lambda_k)$ is making x_k non-zero, while all other $x_{k'}$ with $k' \in \mathcal{C}_k\backslash\{k\}$ are 0 (this is because $m^e_k > m^e_{k'}$, for all $k' \in \mathcal{C}_k\backslash\{k\}$). 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:

$$\begin{aligned} \min_x \quad & \sum_{k\in\mathcal{C}_0} \lambda_k C_k \left(x_k/\lambda_k \right) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{C}_0} m^e_k x_k = a, \\ & x \geq 0, \end{aligned} \tag{3.7}$$

Following the discussion in solving (2.10) (using the KKT conditions), we can define a new function, in analogy to $\widehat{\Delta}_{\mathcal{K}}(\cdot)$ from (2.103) (but now with sub-

script $\mathcal{C}_0$), and under $\{\tilde{\pi}^r_{**}\}$, this function plays the role of a lifting mapping in state-space collapse results.

3.5 Incomplete information

We consider a two phase problem as outlined in §3.2. Assume that each patient in the ED will need at most two phases of treatment. After the first phase, some of patients will leave the ED directly, while others will go to phase 2. Assume that the mean service times at both phases are 1, and the fraction of patients continuing to the second phase is p.

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:

$$\widetilde{\mathcal{S}}^r(t) = \int_0^t \left(\widehat{\tau}^r_{11}(s)\right)^2 d\bar{\bar{E}}^r_1(s) + \int_0^t \left(\widehat{\tau}^r_{12}(s) + \widehat{\tau}^r_2(s)\right)^2 d\bar{\bar{E}}^r_2(s), \tag{3.8}$$

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_2^r is the arrival process for patients with 2 phases.

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

$$\begin{aligned}\lim_{r\to\infty}\widetilde{\mathcal{S}}^r(t) \geq &(1-p)\int_0^t \left(\widetilde{\Delta}_1\left(\widetilde{Q}_w(s)-\widehat{\omega}\right)^+\right)^2 ds \\ &+ p\int_0^t \left(\widetilde{\Delta}_1\left(\widetilde{Q}_w(s)-\widehat{\omega}\right)^+ + \widetilde{\Delta}_2\left(\widetilde{Q}_w(s)-\widehat{\omega}\right)^+/p\right)^2 ds,\end{aligned} \tag{3.9}$$

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

$$\begin{aligned}\min_x \quad & (1-p)x_1^2 + p(x_1 + x_2/p)^2 \\ \text{s.t.} \quad & (1+p)x_1 + x_2 = a, \\ & x_1, x_2 \geq 0.\end{aligned} \tag{3.10}$$

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 §2.7.7, one can only prove that

$$\begin{aligned}&\frac{1}{\bar{\bar{E}}_1^r(b) + \bar{\bar{E}}_2^r(b) - \bar{\bar{E}}_1^r(a) - \bar{\bar{E}}_2^r(a)} \times \\ &\left(\int_a^b \widehat{\tau}_{11}^r(s) d\bar{\bar{E}}_1^r(s) + \int_a^b \widehat{\tau}_{12}^r(s) d\bar{\bar{E}}_2^r(s) - \int_a^b \widetilde{Q}_1^r(s) ds\right) \Rightarrow 0.\end{aligned} \tag{3.11}$$

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 (3.9). Indeed, the service discipline in the

first phase is FCFS. One can thus expect that

$$\begin{aligned}&\frac{1}{\bar{\bar{E}}_1^r(b)-\bar{\bar{E}}_1^r(a)}\left(\int_a^b \widehat{\tau}_{11}^r(s)d\bar{\bar{E}}_1^r(s)\right)=\frac{1}{\bar{\bar{E}}_2^r(b)-\bar{\bar{E}}_2^r(a)}\left(\int_a^b \widehat{\tau}_{12}^r(s)d\bar{\bar{E}}_2^r(s)\right)\\ =&\frac{1}{\bar{\bar{E}}_1^r(b)+\bar{\bar{E}}_2^r(b)-\bar{\bar{E}}_1^r(a)-\bar{\bar{E}}_2^r(a)}\left(\int_a^b \widehat{\tau}_{11}^r(s)d\bar{\bar{E}}_1^r(s)+\int_a^b \widehat{\tau}_{12}^r(s)d\bar{\bar{E}}_2^r(s)\right).\end{aligned} \tag{3.12}$$

By using (3.12), together with (3.11), then following the discussion in §2.7.7 (also in [41]), we deduce (3.9).

4. SOME FUTURE RESEARCH DIRECTIONS

This thesis modeled and analyzed the patient flow in EDs by using queueing theory. Two ED models are built, and asymptotic optimality of proposed policies are also established. The differences of these two models (as shown in the following table) are the assumptions (cost structure and the routing behavior) and the information used in the policies (queue lengthes or ages):

Tab. 4.1: Comparison of two models

	Queue Length Model	Sojourn Time Model
Congestion Cost	Queue Length	Sojourn Time
blueWIP Transition	Markovian	Deterministic
WIP Policy	Queue Length	Age

The models considered in this thesis capture usefully the control of ED patient flow, however, they are by no means the final story. Several noticeable ED characteristics are left out. 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). Those features, we believe, are worthy of future researchs.

4.1 Adding delays between transfers

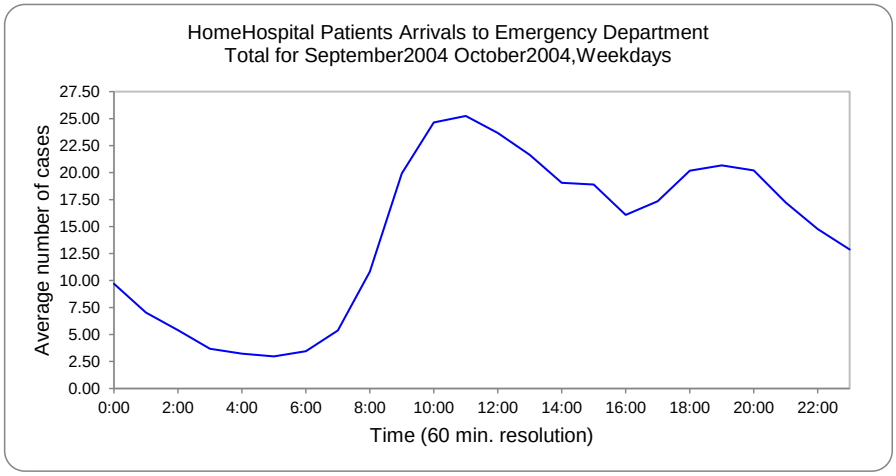
In emergency department, there are delays between successive patient visits to physicians. In [44], the delay phases are modeled as infinite-server queues (*content* phases). One would expect that, if the delays are short, those delays will have no impact asymptotically; at the other extreme, if the 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.

Consider the basic model as an example. Similarly to [44], 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-WIP patients is $r^{\alpha_{jk}}\mu_{jk}$, and the one between l-WIP patients and k-WIP 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 the results in this thesis intact. Conversely, $\alpha_{jk} < -2$ (for all j, k) decouples the triage from WIP - both can be controlled separately; and $\alpha_{lk} < -2$ (for all l, k) pushes the WIP feedback far enough into the future so that the WIP sub-system can be analyzed as a queueing system without feedback. All other cases require further thought and plausibly a more delicate analysis. A brief discussion is provided in §A.

4.2 Time-varying arrival rates

Emergency departments, like many other service systems, must cope with arrival rates that are significantly time-varying. The following figure, plotted using SEEStat developed in SEELab at Technion, elaborate the arrival rate to the emergency department of a hospital in Israel based on data on all workdays in September-October 2004:

Fig. 4.1: Arrival rate in an Israeli ED



As it can be seen from this figure, the arrival rate on the whole day is time-dependent. 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 (to analyze the whole day), 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 ([30], [27]). The triage part of the time-varying ED flow control is analyzed in [10], 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 WIP part is carried out in [6]. Combining these two results could provide the starting point for solving the flow control problem for a time-varying ED, within a fluid framework.

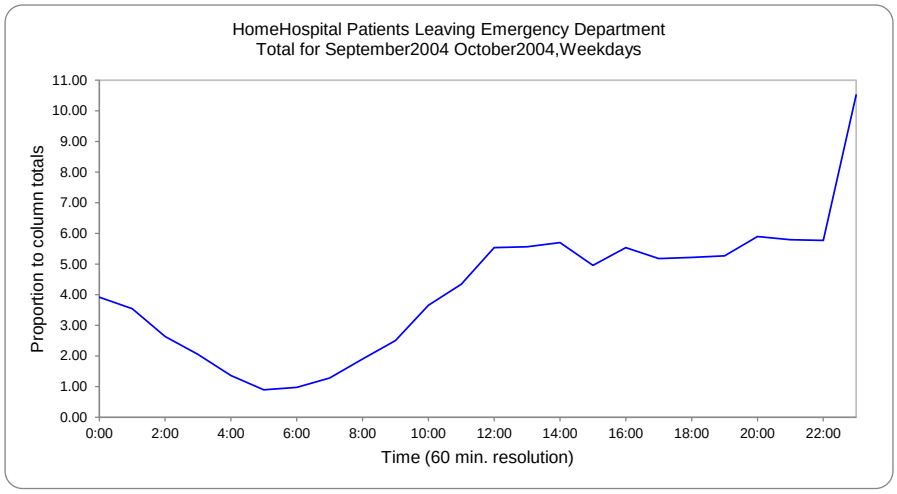
4.3 Length-of-Stay constraints

Many EDs implement, or at least strive for, an upper bound on patients' overall Length of Stay (LOS). In an Israeli ED ([10]), 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 the Israeli ED, but relatively prevalent in U.S. EDs in the form of ambulance diversion ([16, 1, 2]). Interestingly, admission control problems, with costs incurred by blocked customers, in fact motivated [35]. 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.

4.4 Adding abandonment to triage or WIP patients

The following figure, plotted using SEEStat again, elaborate the proportion of patients leaving the ED in the Israeli hospital based on all workdays in September-October 2004: From the figure, it can be seen that during the afternoon-evening peak, the fraction of patients abandoning the ED is around 5%. Similar proportion is also observed in [2].

Fig. 4.2: Abandon proportion in an Israeli ED



Abandonment phenomenon has become a growing concern in overcrowded EDs. There are two kinds of abandonment: Left-Without-Been-Seen (LWBS) and Against-Medical-Advice (AMA), in which the former represents the phenomenon that the triage patients leave the ED before receiving any treatments, while the latter represents the phenomenon that the WIP patients leave the ED before finishing all treatments. For those LWBS patients may miss out their necessary care and be exposed to unnecessary medical risk. Similarly for those AMA patients. Thus it is necessary to analyze a model

with customer abandonment. Queueing models with customer abandonment has been analyzed in service systems such as call centers, and has proved significant in affecting system performance and optimal decisions; see [19, 32].

Indeed, abandonment could significantly impact the structure of optimal policies. For systems without feedback, [24] considered linear cost, with hazard rate scaling of patience time distributions, and [3] 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.

APPENDIX

A. DISCUSSION FOR THE CONJECTURE IN §4.1: ADDING DELAYS AFTER SERVICE

From Lemma 3.4 of [5], we know that, for any given sequence of $x^n \in \mathcal{D}$, there are $y^n \in \mathcal{D}$ satisfying the following equation:

$$y^n(t) = x^n(t) - \mu^n \int_0^t y^n(s)ds; \tag{A.1}$$

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 $Q^r_{jk}(t)$ to denote the number of patients in the delayed system between j-triage and k-WIP patients at time t, and $Q^r_{kl}(t)$ the number of patients in the delayed system between the k-WIP and l-WIP patients at time t.

The number of k-WIP patients at time t is

$$\begin{aligned} Q_k^r(t) &= Q_k^r(0) + \sum_{j\in\mathcal{J}} \left(\Phi_{jk}\left(S_j\left(T_j^r(t)\right)\right) + Q_{jk}^r(0) - Q_{jk}^r(t)\right) \\ &\quad + \sum_{l\in\mathcal{K}} \left(\Phi_{lk}\left(S_l\left(T_l^r(t)\right)\right) + Q_{lk}^r(0) - Q_{lk}^r(t)\right) - S_k\left(T_k^r(t)\right) \\ &= Q_k^r(0) + \sum_{j\in\mathcal{J}} \Phi_{jk}^r\left(S_j\left(T_j^r(t)\right)\right) + \sum_{l\in\mathcal{K}} \Phi_{lk}^r\left(S_l\left(T_l^r(t)\right)\right) - S_k\left(T_k^r(t)\right) \\ &\quad - \sum_{j\in\mathcal{J}} \left(Q_{jk}^r(t) - Q_{jk}^r(0)\right) - \sum_{l\in\mathcal{K}} \left(Q_{lk}^r(t) - Q_{lk}^r(0)\right), \qquad k\in\mathcal{K}. \end{aligned} \tag{A.2}$$

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 (A.2) and (2.34) 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 (A.1), 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 [5], 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 §2.6.2 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 (A.2) and (2.34), 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 $k \in \mathcal{K}$:

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

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

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

We then have

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

Suppose that there is a sequence of $\{\lambda^r\}$ with (i) $\lambda^r \to \lambda$ for some $\lambda > 0$, (ii) $\widehat{X}^r(0) \Rightarrow 0$, and (iii) making $\{\widehat{E}^r\}$ tight. Then from the fluid limit argument, we can prove that $\widehat{S}^r\left(\mu r^{2+\alpha}\int_0^t \bar{\bar{X}}^r(s)ds\right)$ converge to a driftless Brownian motion with variance λ; using (A.1), 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.

BIBLIOGRAPHY

[1] G. Allon, S. Deo, and W. Lin. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research*, Forthcoming, 2013.

[2] M. Armony, S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, and G. B. Yom-Tov. Patient flow in hospitals: A data-based queueing-science perspective. *Submitted*, 2011.

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

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

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

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

[7] R. Beveridge, B. Clarke, L. Janes, N. Savage, J. Thompson, G. Dodd, M. Murray, C. N. Jordan, D. Warren, and A. Vadeboncoeur. Implementation guidelines for the Canadian Emergency Department Triage & Acuity Scale (CTAS), 1998.

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

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

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

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

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

[13] H. Chen and D. D. Yao. *Fundamentals of Queuing Networks: Performance, Asymptotics, and Optimization*. Springer-Valeg, 2001.

[14] J. G. Dai. On positive Harris recurrence of multiclass queueing networks:

a unified approach via fluid limit models. *Annals of Applied Probability*, 5(1):49–77, 1995.

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

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

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

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

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

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

[21] E. Hing and F. Bhuiya. Wait time for treatment in hospital emergency departments: 2009. *NCHS data brief*, (102):1–8, 2012.

[22] W. J. Hopp and W. S. Lovejoy. *Hospital Operations: Principles of High Efficiency Health Care*. FT Press, 2012.

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

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

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

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

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

[28] J. Maa. The waits that matter. *The New England Jornal of Medicine*, June 16:2279–2281, 2011.

[29] S. E. Mace and T. A. Mayer. 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, 2008.

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

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

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

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

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

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

[36] M. I. Reiman. 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, 1982.

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

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

[39] S. Saghafian, W. J. Hopp, M. P. Van Oyen, J. S. Desmond, and S. L. Kronick. Complexity-based triage: A tool for improving patient safety and operational efficiency. *Submitted*, 2011.

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

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

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

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

[44] G. B. Yom-Tov and A. Mandelbaum. Erlang-R: A time-varying queue with ReEntrant customers, in support of healthcare staffing. *Submitted*, 2011.